

Research Software Engineering with Python

Contents

- [Git & GitHub](#)
- [Python](#)
- [Text Editor](#)
- [Course Contents in Jupyter](#)

Introduction

In this course, you will move beyond programming, to learn how to construct reliable, readable, efficient research software in a collaborative environment. The emphasis is on practical techniques, tips, and technologies to effectively build and maintain complex code. This is a relatively short course (8-10 half-day modules) which is intensive and involves hands-on exercises.

Pre-requisites

- It would be extremely helpful to have experience in at least one programming language (for example `C++`, `C`, `Fortran`, `Python`, `Ruby`, `Matlab` or `R`) but this is not a strict requirement.
- Experience with [version control](#) and/or the [unix shell](#), for instance from Software Carpentry, would also be helpful.
- You should bring your own computer to the course as there are several hands-on exercise for you to work through.
- We have provided [setup](#) instructions for installing the software needed for the course on your computer.

Eligibility

The course is open to postgraduate students, early career researchers, practitioners (e.g. data analysts/scientists) and researchers interested in to learn how to construct reliable, readable, efficient research software in a collaborative environment. Turing PhD students and researchers are particularly encouraged to apply. Attendance is free.

Instructors

- [Turing Research Engineering Group](#)

Exercises

Examples and exercises for this course will be provided in `Python`. `Python` syntax and usage will be introduced during this course but please be aware that this course is **not** intended to teach `Python`.

Solutions

Note: you are not graded.

Sample solutions to the exercises are available [here](#).

Versions

You can browse through course notes as HTML, download them as a printable PDF via the navigation bar to the left, or clone the repo and run the notebooks (see the setup instructions).

Support and Contributing

If you encounter any problem or bug in these materials, please remember to add an issue to the [course repo](#), explaining the problem and, potentially, its solution. By doing this, you will improve the instructions for future users. :tada:

We also welcome suggestions and contributions for adding to or improving the material.

Installation Instructions

Introduction

This document contains instructions for installation of the packages we'll be using during the course. You will be following the training on your own computer, so please complete these instructions. The instructions include Windows, Mac and Linux specific sections.

► Note for Mac users

If you encounter any problem during installation and you manage to solve them (feel free to ask us for help), it would be greatly helpful if you'd add an issue to the [course repo](#), explaining the problem and solution (Note: you'll require a GitHub account for this). By doing this you will be helping to improve the instructions for future users! :tada:

What we're installing

- the `Python` programming language (version 3.8 or greater) and `Conda`
- a selection of `Python` software packages that will be used during the course (via a Conda environment)
- `git` for the version control module
- a suitable text editor

Please ensure that you have a computer (ideally a laptop) with all of these installed. Even if you think you have all of these things already, it's worth reading through the prerequisite pages to make sure.

Unfamiliar with the command line?

Familiarity with the command line isn't a prerequisite for the course, but you may need to make use of it at some points. Some of the install steps require you to enter commands in a prompt (terminal or console window) on your computer.

If you're working on a Mac or Linux computer, simply open the [Terminal](#) app when you arrive at these steps. If you're on a Windows PC, we recommend installing the [Git Bash terminal](#) which is covered in the next step.

Git & GitHub

Check whether you have installed already.

```
git --version
```

If not, follow the instructions for your OS and try running this command again. If your version of [git](#) is more than 18 months old (see [releases](#)), please update it.

Windows instructions:

Install the [GitHub Desktop Client](#). This comes with both a GUI client as well as the [Git Bash](#) terminal client which we will use during the course. In some instances [Git Bash](#) may need to be installed separately. In order to use [conda](#) with [Git Bash](#) follow the instructions [here](#)

You will need to create an account on GitHub. You can then sign-in to the GitHub Desktop Client which should automatically set-up [SSH based authentication](#) for the terminal client.

Configure the default terminal client (there are three different flavours of terminal on Windows: [Windows CMD](#) (DOS like), [Windows PowerShell](#), and [BASH](#)) to use BASH, as this most closely resembles the [Linux](#) and [macOS](#) terminal used by other students:

1. In the Desktop Client, select [Tools](#)
2. Then [Options](#)
3. [Default Shell](#)
4. [Git Bash](#)

You'll know it has worked when you can open a Git Bash terminal; the window should have a title that starts with MINGW32 (scroll to the top of this page for how to check the git version).

macOS and Linux instructions:

- ▶ Installing Git on macOS
- ▶ Installing Git on Linux

To use [git](#) you will need to set up an account with your email address and name. To do this you can follow the [Your Identity](#) section of [first time git setup](#).

You can check that they have been set correctly by running `git config user.name` and `git config user.email`. For the [git](#) module (Version Control with Git), you will also require access to GitHub.

Follow these instructions if you are working on macOS or Linux:

1. [Sign up](#), if you haven't already
2. [Generate an SSH key pair](#)
3. [Add the public key to your GitHub account and the private key to your computer's keychain](#)
4. Lastly, you should [test your SSH connection](#)

Python

Download [Anaconda](#) for your OS. Then follow the instructions for [installation](#), which differ for Windows, macOS and Linux.

You should test whether the installation has worked as expected by doing the following:

- Open a terminal (console) window and run the following:

```
python --version
```

Note: Anaconda should have installed the most recent version by default but note that you will require version [3.8](#) or greater of Python for the course.

Text Editor

Unless you already use a specific editor which you are comfortable with we recommend using one of the following:

- [Visual Studio Code](#)
- [Notepad++](#)
- [Emacs](#)
- [PyCharm](#)

- ▶ Windows editor tips and final checks
- ▶ macOS editor tips and final checks
- ▶ Linux editor tips and final checks

Course Contents in Jupyter

After following the installation instructions for your operating system, you should now have the following:

1. A [git](#) installation, linked to your [GitHub](#) account
2. A working installation of the [Python](#) (3) programming language
3. The ability to install Python packages via [Conda](#)

Throughout the course, we will be working in Python, and one of the best ways to get started with this language is Jupyter Lab.

In addition to viewing the course materials online at this site (<https://alan-turing-institute.github.io/rse-course>), we recommend cloning (downloading) the GitHub repository containing the course contents. This allows you to open the contents interactively via Jupyter on your computer.

Navigate to a suitable location in a terminal window and clone the course repository (if you haven't used Git/GitHub before, it can be useful to create a folder to store repositories with `mkdir`):

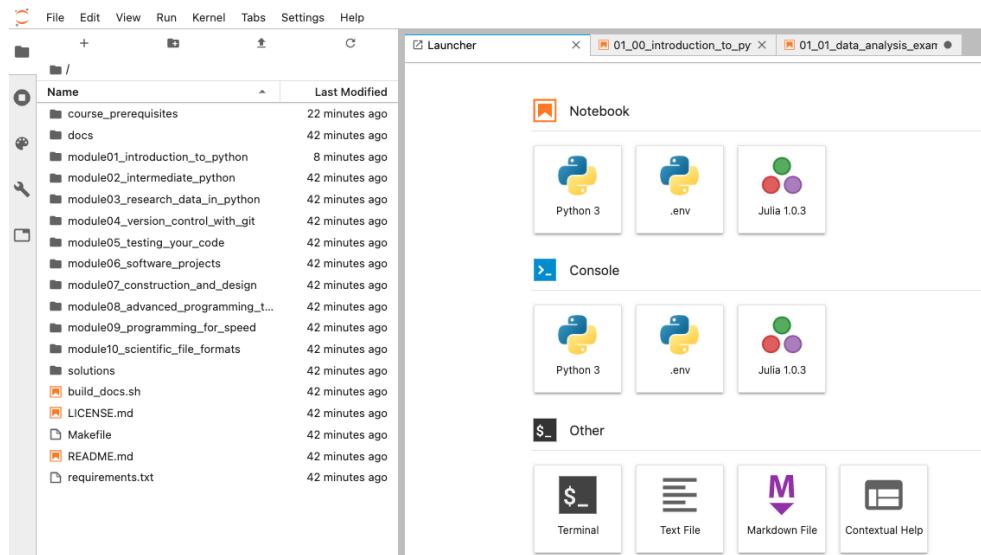
```
mkdir ~/github_repos
cd ~/github_repos
git clone --depth 1 https://github.com/alan-turing-institute/rse-course
```

The course contents should take a few moments to download. Once the download has finished, you should enter the cloned course repository, set up a `conda` Python environment containing the packages we'll make use of during the course (including Jupyter) and then launch Jupyter Lab. To do this, we'll make use of `conda`'s `environment.yml` file, which is configured in this case to create an environment named `rse-course`:

```
cd rse-course
conda env create -f environment.yml
conda activate rse-course
jupyter lab
```

This should automatically open a window in your default web browser (if not, go to <http://localhost:8890/lab>).

You should be able to see a layout that looks something like the below. Double clicking on the folders, then the `.ipynb` files within will allow you to view the course materials interactively, giving you the option to edit code cells and experiment as you learn.



C++ compiler for Windows

If you're using Windows, in order to use `Cython` in the "Programming for Speed" module, you may need to install a `C++ compiler`. Open a terminal window (e.g. in Git Bash that you installed earlier) and activate the `conda` Python environment you just set up:

```
conda activate rse-course
```

Then [see here for details](#) on how to install the C++ compiler.

Jupyter issues on Windows (Sophos):

To use the Jupyter Lab on a `windows` computer with Sophos anti-virus installed it may be necessary to open additional ports allowing communication between the notebook and its server. The [solution](#) is:

- open your `Sophos Endpoint Security and Control Panel` from your tray or start menu
- select `Configure > Anti-virus > Authorization` from the menu at the top
- select the websites tab
- click the `Add` button and add `127.0.0.1` and `localhost` to the `Authorized websites` list
- restart computer (or just restart the Jupyter)

1. Introduction to Python

- Why use scripting languages?
- Python and the Jupyter notebook
- Variables, using functions, and types
- Loops and control
- Data structures: lists, dictionaries, and sets.

Contents

- [1.0 Introduction to Python](#) (10 minutes)
- [1.1 Variables](#) (20 minutes)
- [1.2 Functions](#) (20 minutes)
- [1.3 Types](#) (20 minutes)
- [1.4 Containers](#) (10 minutes)
- [1.5 Dictionaries](#) (10 minutes)
- [1.6 Data Structures](#) (5 minutes)
- [1.7 Control and Flow](#) (15 minutes)
- [1.8 Iteration](#) (10 minutes)

Total time: 2 hrs

Exercises

Classroom exercises are grouped together at the end of the module: [1.9 Classroom Exercises](#). Each exercise is labelled with any sections whose contents are relevant. We recommend that instructors schedule the exercises to be done in groups during breaks in the taught content. However, it is **important** that participants also have some time away from their screens. Exercises can also be left as self-paced homework assignments if preferred.

1.0 Introduction to Python

Estimated time for this notebook: 10 minutes

1.0.1 Why write programs for research?

Programs are a rigorous way of describing data analysis for other researchers, as well as for computers.

- Not just labour saving
- Scripted research can be tested and reproduced

Sensible Input - Reasonable Output

Computational research suffers from people assuming each other's data manipulation is correct. By sharing code, which is much more easy for a non-author to understand than a spreadsheet, we can avoid the "SIRO" problem. The old saw "Garbage in Garbage out" is not the real problem for science:

- Sensible input
- Reasonable output

Why write software to manage your data and plots?

We can use programs for our entire research pipeline. Not just big scientific simulation codes, but also the small scripts which we use to tidy up data and produce plots. This should be code, so that the whole research pipeline is recorded for reproducibility. Data manipulation in spreadsheets is much harder to share or check. There are many data analysis examples out there, like the on the [software carpentry site](#).

1.0.2 Why Python?

Why teach Python?

- In this first session, we will introduce [Python](#).
- This course is about programming for data analysis and visualisation in research.
- It's not mainly about Python.
- But we have to use some language.

Why Python?

- Python is quick to program in
- Python is popular in research, and has lots of libraries for science
- Python interfaces well with faster languages
- Python is free, so you'll never have a problem getting hold of it, wherever you go.

1.0.3 Many kinds of Python

The Jupyter Notebook

The easiest way to get started using Python, and one that is commonly used for exploratory research, is the Jupyter Notebook.

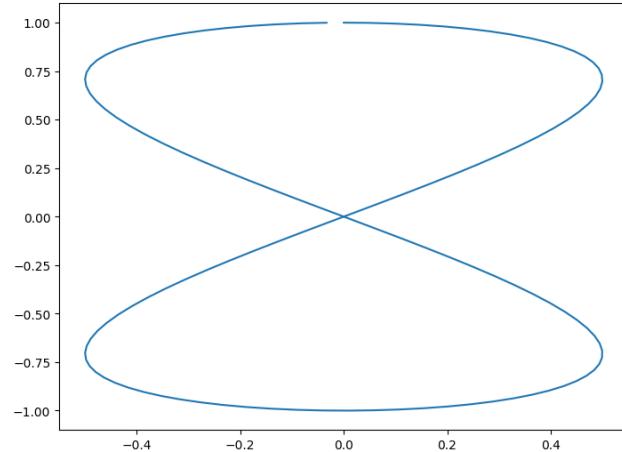
In the notebook, you can easily mix code with discussion and commentary. You can also mix code with the results of that code, such as graphs and other data visualisations.

```
# Make plot
%matplotlib inline
import math

import matplotlib.pyplot as plt
import numpy as np

theta = np.arange(0, 4 * math.pi, 0.1)
eighth = plt.figure()
axes = eighth.add_axes([0, 0, 1, 1])
axes.plot(0.5 * np.sin(theta), np.cos(theta / 2))
```

```
[<matplotlib.lines.Line2D at 0x7f1cd02b8dc0>]
```



We're going to be mainly working in the Jupyter notebook in this course. To get hold of a copy of the notebook, follow the [setup instructions shown on the course website](#).

Jupyter notebooks consist of discussion cells, referred to as "markdown cells", and "code cells", which contain Python. This document has been created using Jupyter notebook, and this very cell is a **Markdown Cell**.

```
print("This cell is a code cell")
```

This cell is a code cell

Code cell inputs are numbered, and show the output below.

Markdown cells contain text which uses a simple format to achieve pretty layout, for example, to obtain:

bold, *italic*

- Bullet

Quote

We write:

```
**bold**, *italic*
* Bullet
> Quote
```

See the Markdown documentation at [This Hyperlink](#)

Typing code in the notebook

When working with the notebook, you can either be in a cell, typing its contents, or outside cells, moving around the notebook.

- When in a cell, press escape to leave it. When moving around outside cells, press return to enter.
- Outside a cell:
 - Use arrow keys to move around.
 - Press **b** to add a new cell below the cursor.
 - Press **m** to turn a cell from code mode to markdown mode.
 - Press **shift+enter** to calculate the code in the block.
 - Press **h** to see a list of useful keys in the notebook.
- Inside a cell:
 - Press **tab** to suggest completions of variables. (Try it!)

Supplementary material: Learn more about [Jupyter notebooks](#).

Python at the command line

More experienced Python users tend to prefer working in a “command line environment”. You can find out more about this by attending a [“Software Carpentry”](#) or similar workshop, which introduce the skills needed for computationally based research.

```
%%bash
# Above line tells Python to execute this cell as *shell code*
# not Python, as if we were in a command line
# This is called a 'cell magic'

python -c "print(2 * 4)"
```

8

Python scripts

When your code gets more complicated, you’ll want to be able to write your own full programs in Python, which can be run just like any other program on your computer. Here are some examples:

```
%%bash
echo "print(2 * 4)" > eight.py
python eight.py
```

8

We can make the script directly executable (on Linux or Mac) by inserting a [hashbang](#) and [setting the permissions](#) to execute.

```
%%writefile fourteen.py
#!/usr/bin/env python
print(2 * 7)
```

Overwriting fourteen.py

```
%%bash
chmod u+x fourteen.py
./fourteen.py
```

14

Python Libraries

We can write our own Python libraries, called modules which we can import into the notebook and invoke:

```
%%writefile draw_eight.py
# Above line tells the notebook to treat the rest of this
# cell as content for a file on disk.
import math
import numpy as np
import matplotlib.pyplot as plt

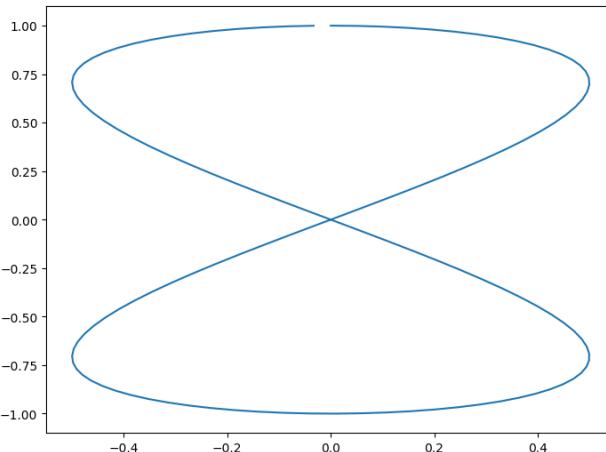
def make_figure():
    theta = np.arange(0, 4 * math.pi, 0.1)
    eight = plt.figure()
    axes = eight.add_axes([0, 0, 1, 1])
    axes.plot(0.5 * np.sin(theta), np.cos(theta / 2))
    return eight
```

Overwriting draw_eight.py

In a real example, we could edit the file on disk using a program such as [Atom](#) or [VS code](#).

```
import draw_eight # Load the library file we just wrote to disk
```

```
image = draw_eight.make_figure()
```



There is a huge variety of available packages to do pretty much anything. For instance, try `import antigravity`.

The `%` at the beginning of a cell is called *magics*. There's a [large list of them available](#) and you can [create your own](#).

1.1 Variables

Estimated time for this notebook: 10 minutes

1.1.1 Variable Assignment

When we generate a result, the answer is displayed, but not kept anywhere.

```
2 * 3
```

```
6
```

If we want to get back to that result, we have to store it. We put it in a box, with a name on the box. This is a **variable**.

```
six = 2 * 3
```

```
print(six)
```

```
6
```

If we look for a variable that hasn't ever been defined, we get an error.

```
print(seven)
```

```
NameError: name 'seven' is not defined
```

That's **not** the same as an empty box, well labeled:

```
nothing = None
```

```
print(nothing)
```

```
None
```

```
type(None)
```

```
NoneType
```

(None is the special python value for a no-value variable.)

Supplementary Materials: There's more on variables at <http://swcarpentry.github.io/python-novice-inflammation/01-numpy/index.html>

Anywhere we could put a raw number, we can put a variable label, and that works fine:

```
print(5 * six)
```

```
30
```

```
scary = six * six * six
```

```
print(scary)
```

```
216
```

1.1.2 Reassignment and multiple labels

But here's the real scary thing: it seems like we can put something else in that box:



Note that the data that was there before has been lost.

No labels refer to it any more - so it has been "Garbage Collected"! We might imagine something pulled out of the box, and thrown on the floor, to make way for the next occupant.

In fact, though, it is the **label** that has moved. We can see this because we have more than one label referring to the same box:



And we can move just one of those labels:



So we can now develop a better understanding of our labels and boxes: each box is a piece of space (an *address*) in computer memory. Each label (variable) is a reference to such a place.

When the number of labels on a box ("variables referencing an address") gets down to zero, then the data in the box cannot be found any more.

After a while, the language's "Garbage collector" will wander by, notice a box with no labels, and throw the data away, **making that box available for more data**.

Old fashioned languages like C and Fortran don't have Garbage collectors. So a memory address with no references to it still takes up memory, and the computer can more easily run out.

So when I write:



The following things happen:

1. A new text **object** is created, and an address in memory is found for it.
2. The variable "name" is moved to refer to that address.
3. The old address, containing "James", now has no labels.
4. The garbage collector frees the memory at the old address.

Supplementary materials: There's an online python tutor which is great for visualising memory and references. Try the [scenario we just looked at](#)

Labels are contained in groups called "frames": our frame contains two labels, 'nom' and 'name'.

1.1.3 Objects and types

An object, like `name`, has a type. In the online python tutor example, we see that the objects have type "str". `str` means a text object: Programmers call these 'strings'.

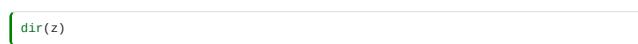


Depending on its type, an object can have different *properties*: data fields Inside the object.

Consider a Python complex number for example:



We can see what properties and methods an object has available using the `dir` function:



```
[ '__abs__',
  '__add__',
  '__bool__',
  '__class__',
  '__delattr__',
  '__dir__',
  '__divmod__',
  '__doc__',
  '__eq__',
  '__float__',
  '__floordiv__',
  '__format__',
  '__ge__',
  '__getattribute__',
  '__getnewargs__',
  '__gt__',
  '__hash__',
  '__init__',
  '__init_subclass__',
  '__int__',
  '__le__',
  '__lt__',
  '__mod__',
  '__mul__',
  '__ne__',
  '__neg__',
  '__new__',
  '__pos__',
  '__pow__',
  '__radd__',
  '__rdivmod__',
  '__reduce__',
  '__reduce_ex__',
  '__repr__',
  '__rfloordiv__',
  '__rmod__',
  '__rmul__',
  '__rpow__',
  '__rsub__',
  '__rtruediv__',
  '__setattr__',
  '__sizeof__',
  '__str__',
  '__sub__',
  '__subclasshook__',
  '__truediv__',
  'conjugate',
  'imag',
  'real']
```

You can see that there are several methods whose name starts and ends with `_` (e.g. `__init__`): these are special methods that Python uses internally, and we will discuss some of them later on in this course. The others (in this case, `conjugate`, `img` and `real`) are the methods and fields through which we can interact with this object.

```
{ type(z)
complex
z.real
3.0
z.imag
1.0
```

A property of an object is accessed with a dot.

The jargon is that the “dot operator” is used to obtain a property of an object.

When we try to access a property that doesn’t exist, we get an error:

```
{ z.wrong
-----
AttributeError                               Traceback (most recent call last)
Cell In [27], line 1
      1 z.wrong
-----
AttributeError: 'complex' object has no attribute 'wrong'
```

1.1.4 Reading error messages.

It’s important, when learning to program, to develop an ability to read an error message and find, from in amongst all the confusing noise, the bit of the error message which tells you what to change!

We don’t yet know what is meant by `AttributeError`, or “Traceback”.

```
{ z2 = 5 - 6j
  print("Gets to here")
  print(z.wrong)
  print("Didn't get to here")
-----
Gets to here
-----
AttributeError                               Traceback (most recent call last)
Cell In [28], line 3
      1 z2 = 5 - 6j
      2 print("Gets to here")
      3 print(z.wrong)
      4 print("Didn't get to here")
-----
AttributeError: 'complex' object has no attribute 'wrong'
```

But in the above, we can see that the error happens on the **third** line of our code cell.

We can also see that the error message:

```
'complex' object has no attribute 'wrong'
```

...tells us something important. Even if we don't understand the rest, this is useful for debugging!

1.1.5 Variables and the notebook kernel

When I type code in the notebook, the objects live in memory between cells.

```
number = 0
```

```
print(number)
```

```
0
```

If I change a variable:

```
number = number + 1
```

```
print(number)
```

```
1
```

It keeps its new value for the next cell.

But cells are **not** always evaluated in order.

If I now go back to input cell reading `number = number + 1`, and run it again, with shift-enter. Number will change from 2 to 2, then from 2 to 3, then from 3 to 4... Try it!

So it's important to remember that if you move your cursor around in the notebook, it doesn't always run top to bottom.

Supplementary material: (1) <https://jupyter-notebook.readthedocs.io/en/latest/>

1.1.6 Comments

Code after a `#` symbol doesn't get run.

```
print("This runs") # print("This doesn't")  
# print("This doesn't either")
```

```
This runs
```

1.2 Using Functions

Estimated time for this notebook: 20 minutes

1.2.1 Calling functions

We often want to do things to our objects that are more complicated than just assigning them to variables.

```
len("pneumonoultramicroscopicsilicovolcanoconiosis")
```

```
45
```

Here we have "called a function".

The function `len` takes one input, and has one output. The output is the length of whatever the input was.

Programmers also call function inputs "parameters" or, confusingly, "arguments".

Here's another example:

```
sorted("Python")
```

```
['P', 'h', 'n', 'o', 't', 'y']
```

Which gives us back a *list* of the letters in Python, sorted alphabetically (more specifically, according to their [Unicode order](#)).

The input goes in brackets after the function name, and the output emerges wherever the function is used.

So we can put a function call anywhere we could put a "literal" object or a variable.

```
len("Jim") * 8
```

```
24
```

```
x = len("Mike")  
y = len("Bob")  
z = x + y
```

```
print(z)
```

```
7
```

1.2.2 Using methods

Objects come associated with a bunch of functions designed for working on objects of that type. We access these with a dot, just as we do for data attributes:

```
"shout".upper()
```

```
'SHOUT'
```

These are called methods. If you try to use a method defined for a different type, you get an error:

```
x = 5
```

```
type(x)
```

```
int
```

```
x.upper()
```

```
AttributeError Traceback (most recent call last)
Cell In [9], line 1
----> 1 x.upper()

AttributeError: 'int' object has no attribute 'upper'
```

If you try to use a method that doesn't exist, you get an error:

```
x.wrong
```

```
AttributeError Traceback (most recent call last)
Cell In [10], line 1
----> 1 x.wrong

AttributeError: 'int' object has no attribute 'wrong'
```

Methods and properties are both kinds of **attribute**, so both are accessed with the dot operator.

Objects can have both properties and methods:

```
z = 1 + 5j
```

```
z.real
```

```
1.0
```

```
z.conjugate()
```

```
(1-5j)
```

```
z.conjugate
```

```
<function complex.conjugate>
```

1.2.3 Functions are just a type of object!

Now for something that will take a while to understand: don't worry if you don't get this yet, we'll look again at this in much more depth later in the course.

If we forget the (), we realise that a *method is just a property which is a function!*

```
z.conjugate
```

```
<function complex.conjugate>
```

```
type(z.conjugate)
```

```
builtin_function_or_method
```

```
somefunc = z.conjugate
```

```
somefunc()
```

```
(1-5j)
```

Functions are just a kind of variable, and we can assign new labels to them:

```
sorted([1, 5, 3, 4])
```

```
[1, 3, 4, 5]
```

```
magic = sorted
```

```
type(magic)
```

```
builtin_function_or_method
```

```
magic(["Technology", "Advanced"])
```

```
['Advanced', 'Technology']
```

1.2.4 Getting help on functions and methods

The 'help' function, when applied to a function, gives help on it!

```
help(sorted)
```

```
Help on built-in function sorted in module builtins:  
sorted(iterable, /, *, key=None, reverse=False)  
    Return a new list containing all items from the iterable in ascending order.  
  
    A custom key function can be supplied to customize the sort order, and the  
    reverse flag can be set to request the result in descending order.
```

The 'dir' function, when applied to an object, lists all its attributes (properties and methods):

```
dir("Hexxo")
```

```
['__add__',  
 '__class__',  
 '__contains__',  
 '__delattr__',  
 '__dir__',  
 '__doc__',  
 '__eq__',  
 '__format__',  
 '__ge__',  
 '__getattribute__',  
 '__getitem__',  
 '__getnewargs__',  
 '__gt__',  
 '__hash__',  
 '__init__',  
 '__init_subclass__',  
 '__iter__',  
 '__le__',  
 '__len__',  
 '__lt__',  
 '__mod__',  
 '__mul__',  
 '__ne__',  
 '__new__',  
 '__reduce__',  
 '__reduce_ex__',  
 '__repr__',  
 '__rmod__',  
 '__rmul__',  
 '__setattr__',  
 '__sizeof__',  
 '__str__',  
 '__subclasshook__',  
 'capitalize',  
 'casefold',  
 'center',  
 'count',  
 'encode',  
 'endswith',  
 'expandtabs',  
 'find',  
 'format',  
 'format_map',  
 'index',  
 'isalnum',  
 'isalpha',  
 'isascii',  
 'isdecimal',  
 'isdigit',  
 'isidentifier',  
 'islower',  
 'isnumeric',  
 'isprintable',  
 'isspace',  
 'istitle',  
 'isupper',  
 'join',  
 'ljust',  
 'lower',  
 'lstrip',  
 'maketrans',  
 'partition',  
 'replace',  
 'rfind',  
 'rindex',  
 'rjust',  
 'rpartition',  
 'rsplit',  
 'rstrip',  
 'split',  
 'splitlines',  
 'startswith',  
 'strip',  
 'swapcase',  
 'title',  
 'translate',  
 'upper',  
 'zfill']
```

Most of these are confusing methods beginning and ending with __, part of the internals of python.

Again, just as with error messages, we have to learn to read past the bits that are confusing, to the bit we want:

```
"Hexxo".replace("x", "l")
```

```
'Hello'
```

```
help("Fish".replace)
```

```
Help on built-in function replace:  
replace(old, new, count=-1, /) method of builtins.str instance  
    Return a copy with all occurrences of substring old replaced by new.  
  
    count  
        Maximum number of occurrences to replace.  
        -1 (the default value) means replace all occurrences.  
  
    If the optional argument count is given, only the first count occurrences are  
    replaced.
```

1.2.5 Operators

Now that we know that functions are a way of taking a number of inputs and producing an output, we should look again at what happens when we write:

```
x = 2 + 3
```

```
print(x)
```

```
5
```

This is just a pretty way of calling an "add" function. Things would be more symmetrical if add were actually written

```
x = +(2, 3)
```

Where '+' is just the name of the adding function.

In python, these functions **do** exist, but they're actually **methods** of the first input: they're the mysterious functions we saw earlier (Two underscores.)

```
x.__add__(7)
```

```
12
```

We call these symbols, +, - etc, "operators".

The meaning of an operator varies for different types:

```
"Hello" + "Goodbye"
```

```
'HelloGoodbye'
```

```
[2, 3, 4] + [5, 6]
```

```
[2, 3, 4, 5, 6]
```

Sometimes we get an error when a type doesn't have an operator:

```
7 - 2
```

```
5
```

```
[2, 3, 4] - [5, 6]
```

```
TypeError  
Cell In [33], line 1  
----> 1 [2, 3, 4] - [5, 6]  
  
TypeError: unsupported operand type(s) for -: 'list' and 'list'
```

The word "operand" means "thing that an operator operates on"!

Or when two types can't work together with an operator:

```
[2, 3, 4] + 5
```

```
TypeError  
Cell In [34], line 1  
----> 1 [2, 3, 4] + 5  
  
TypeError: can only concatenate list (not "int") to list
```

To do this, put:

```
[2, 3, 4] + [5]
```

```
[2, 3, 4, 5]
```

Just as in Mathematics, operators have a built-in precedence, with brackets used to force an order of operations:

```
print(2 + 3 * 4)
```

```
14
```

```
print((2 + 3) * 4)
```

```
20
```

Supplementary material: http://www.mathcs.emory.edu/~valerie/courses/fall10/155/resources/op_precedence.html

1.3 Types

Estimated time for this notebook: 20 minutes

We have seen that Python objects have a 'type':

```
type(5)
```

```
int
```

1.3.1 Floats and integers

Python has two core numeric types, `int` for integer, and `float` for real number.

```
{ one = 1  
ten = 10  
one_float = 1.0  
ten_float = 10.0
```

Zero after a point is optional. But the **Dot** makes it a float.

```
{ tenth = one_float / ten_float  
  
{ tenth  
  
0.1  
  
{ type(one)  
  
int  
  
{ type(one_float)  
  
float
```

The meaning of an operator varies depending on the type it is applied to!

```
{ print(one // ten)  
  
0  
  
{ one_float / ten_float  
  
0.1  
  
{ print(type(one / ten))  
  
<class 'float'>  
  
{ type(tenth)  
  
float
```

The divided by operator when applied to floats, and integers means divide by for real numbers.

The **//** operator means divide and then round down

```
{ 10 // 3  
  
3  
  
{ 10.0 / 3  
  
3.333333333333335  
  
{ 10 / 3.0  
  
3.333333333333335
```

There is a function for every type name, which is used to convert the input to an output of the desired type.

```
{ x = float(5)  
type(x)  
  
float  
  
{ 10 / float(3)  
  
3.333333333333335
```

I lied when I said that the **float** type was a real number. It's actually a computer representation of a real number called a "floating point number". Representing $\sqrt{2}$ or $\frac{1}{3}$ perfectly would be impossible in a computer, so we use a finite amount of memory to do it.

```
{ N = 10000.0  
sum([1 / N] * int(N))  
  
0.9999999999999062
```

Supplementary material:

- <https://docs.python.org/2/tutorial/floatingpoint.html>
- <http://floating-point-gui.de/formats/fp/>
- Advanced: http://docs.oracle.com/cd/E19957-01/806-3568/ncg_goldberg.html

1.3.2 Strings

Python has a built in **string** type, supporting many useful methods.

```
{ given = "James"  
family = "Hetherington"  
full = given + " " + family }
```

So `+` for strings means "join them together" - *concatenate*.

```
{ print(full.upper()) }
```

```
JAMES HETHERINGTON
```

As for `float` and `int`, the name of a type can be used as a function to convert between types:

```
{ ten, one }
```

```
(10, 1)
```

```
{ print(ten + one) }
```

```
11
```

```
{ print(float(str(ten) + str(one))) }
```

```
101.0
```

We can remove extraneous material from the start and end of a string:

```
{ "Hello ".strip()
```

```
'Hello'
```

Note that you can write strings in Python using either single (`' ... '`) or double (`" ... "`) quote marks. The two ways are equivalent. However, if your string includes a single quote (e.g. an apostrophe), you should use double quotes to surround it:

```
{ "James's Class"
```

```
"James's Class"
```

And vice versa: if your string has a double quote inside it, you should wrap the whole string in single quotes.

```
{ "'Wow!', said Bob.'
```

```
'"Wow!", said Bob.'
```

1.3.3 Lists

Python's basic **container** type is the `list`.

We can define our own list with square brackets:

```
{ [1, 3, 7]
```

```
[1, 3, 7]
```

```
{ type([1, 3, 7])}
```

```
list
```

Lists *do not* have to contain just one type:

```
{ various_things = [1, 2, "banana", 3.4, [1, 2]] }
```

We access an **element** of a list with an `int` in square brackets:

```
{ various_things[2]
```

```
'banana'
```

```
{ index = 0  
various_things[index]
```

```
1
```

Note that list indices start from zero.

We can use a string to join together a list of strings:

```
{ name = ["James", "Philip", "John", "Hetherington"]  
print("==".join(name)) }
```

```
James==Philip==John==Hetherington
```

And we can split up a string into a list:

```
{ "Ernst Stavro Blofeld".split(" ") }
```

```
['Ernst', 'Stavro', 'Blofeld']
```

```
"Ernst Stavro Blofeld".split("o")
```

```
['Ernst Stavr', ' Bl', ' feld']
```

And combine these:

```
"->".join("John Ronald Reuel Tolkien".split(" "))
```

```
'John->Ronald->Reuel->Tolkien'
```

A matrix can be represented by **nesting** lists – putting lists inside other lists.

```
identity = [[1, 0], [0, 1]]
```

```
identity[0][0]
```

```
1
```

... but later we will learn about a better way of representing matrices.

1.3.4 Ranges

Another useful type is range, which gives you a sequence of consecutive numbers. In contrast to a list, ranges generate the numbers as you need them, rather than all at once.

If you try to print a range, you'll see something that looks a little strange:

```
range(5)
```

```
range(0, 5)
```

We don't see the contents, because *they haven't been generated yet*. Instead, Python gives us a description of the object - in this case, its type (range) and its lower and upper limits.

We can quickly make a list with numbers counted up by converting this range:

```
count_to_five = range(5)
```

```
print(list(count_to_five))
```

```
[0, 1, 2, 3, 4]
```

Ranges in Python can be customised in other ways, such as by specifying the lower limit or the step (that is, the difference between successive elements). You can find more information about them in the [official Python documentation](#).

1.3.5 Sequences

Many other things can be treated like **lists**. Python calls things that can be treated like lists **sequences**.

A string is one such **sequence** type.

Sequences support various useful operations, including:

- Accessing a single element at a particular index: `sequence[index]`
- Accessing multiple elements (a **slice**): `sequence[start:end_plus_one]`
- Getting the length of a sequence: `len(sequence)`
- Checking whether the sequence contains an element: `element in sequence`

The following examples illustrate these operations with lists, strings and ranges.

```
print(count_to_five[1])
```

```
1
```

```
print("James"[2])
```

```
m
```

```
count_to_five = range(5)
```

```
count_to_five[1:3]
```

```
range(1, 3)
```

```
"Hello World"[4:8]
```

```
'o Wo'
```

```
len(various_things)
```

```
5
```

```
len("Python")
```

```
6
```

```
name
```

```
['James', 'Philip', 'John', 'Hetherington']
```

```
"John" in name
```

```
True
```

```
3 in count_to_five
```

```
True
```

1.3.6 Unpacking

Multiple values can be **unpacked** when assigning from sequences, like dealing out decks of cards.

```
mylist = ["Hello", "World"]
a, b = mylist
print(b)
```

```
World
```

```
range(4)
```

```
range(0, 4)
```

```
zero, one, two, three = range(4)
```

```
two
```

```
2
```

If there is too much or too little data, an error results:

```
zero, one, two, three = range(7)
```

```
ValueError
Cell In [52], line 1
----> 1 zero, one, two, three = range(7)
ValueError: too many values to unpack (expected 4)
```

```
zero, one, two, three = range(2)
```

```
ValueError
Cell In [53], line 1
----> 1 zero, one, two, three = range(2)
ValueError: not enough values to unpack (expected 4, got 2)
```

Python provides some handy syntax to split a sequence into its first element ("head") and the remaining ones (its "tail"):

```
head, *tail = range(4)
print("head is", head)
print("tail is", tail)
```

```
head is 0
tail is [1, 2, 3]
```

Note the syntax with the *. The same pattern can be used, for example, to extract the middle segment of a sequence whose length we might not know:

```
one, *two, three = range(10)
```

```
print("one is", one)
print("two is", two)
print("three is", three)
```

```
one is 0
two is [1, 2, 3, 4, 5, 6, 7, 8]
three is 9
```

1.4 Containers

Estimated time for this notebook: 10 minutes

1.4.1 Checking for containment.

The `list` we saw is a container type: its purpose is to hold other objects. We can ask python whether or not a container contains a particular item:

```
"Dog" in ["Cat", "Dog", "Horse"]
```

```
True
```

```
"Bird" in ["Cat", "Dog", "Horse"]
```

```
False
```

```
2 in range(5)
```

```
True
```

```
{ 99 in range(5)
```

```
False
```

1.4.2 Mutability

A list can be modified: (is mutable)

```
{ name = "James Philip John Hetherington".split(" ")  
print(name)
```

```
['James', 'Philip', 'John', 'Hetherington']
```

```
{ name[0] = "Dr"  
name[1:3] = ["Griffiths-"]  
name.append("PhD")  
print(" ".join(name))
```

```
Dr Griffiths- Hetherington PhD
```

1.4.3 Tuples

A **tuple** is an immutable sequence. It is like a list, except it cannot be changed. It is defined with round brackets.

```
{ x = (0,)  
type(x)
```

```
tuple
```

```
{ my_tuple = ("Hello", "World")  
my_tuple[0] = "Goodbye"
```

```
-----  
TypeError  
Cell In [8], line 2  
    1 my_tuple = ("Hello", "World")  
----> 2 my_tuple[0] = "Goodbye"  
  
TypeError: 'tuple' object does not support item assignment
```

```
{ type(my_tuple)
```

```
tuple
```

str is immutable too:

```
{ fish = "Hake"  
fish[0] = "R"
```

```
-----  
TypeError  
Cell In [10], line 2  
    1 fish = "Hake"  
----> 2 fish[0] = "R"  
  
TypeError: 'str' object does not support item assignment
```

But note that container reassignment is moving a label, **not** changing an element:

```
{ fish = "Rake" # OK!
```

Supplementary material: Try the [online memory visualiser](#) for this one.

1.4.4 Memory and containers

The way memory works with containers can be important:

```
{ x = list(range(3))  
x
```

```
[0, 1, 2]
```

```
{ y = x
```

```
[0, 1, 2]
```

```
{ z = x[0:3]  
y[1] = "Gotcha!"
```

```
{ x
```

```
[0, 'Gotcha!', 2]
```

```
{ y
```

```
[0, 'Gotcha!', 2]
```

```
{ z
```

```
[0, 1, 2]
```

```

x[2] = "Really?"
x
[0, 'Gotcha!', 2]
y
[0, 'Gotcha!', 2]
z
[0, 1, 'Really?']

```

Supplementary material: This one works well at the [memory visualiser](#).

The explanation: While `y` is a second label on the *same object*, `z` is a separate object with the same data. Writing `x[:]` creates a new list containing all the elements of `x` (remember: `[:] is equivalent to [0:<last>]`). This is the case whenever we take a slice from a list, not just when taking all the elements with `[:]`.

The difference between `y=x` and `z=x[:]` is important!

Nested objects make it even more complicated:

```

x = [[ "a", "b"], "c"]
y = x
z = x[0:2]

x[0][1] = "d"
z[1] = "e"

x
[['a', 'd'], 'c']

y
[['a', 'd'], 'c']

z
[['a', 'd'], 'e']

```

Try the [visualiser](#) again.

Supplementary material: The copies that we make through slicing are called *shallow copies*: we don't copy all the objects they contain, only the references to them. This is why the nested list in `x[0]` is not copied, so `z[0]` still refers to it. It is possible to actually create copies of all the contents, however deeply nested they are - this is called a *deep copy*. Python provides methods for that in its standard library, in the `copy` module. You can read more about that, as well as about shallow and deep copies, in the [library reference](#).

1.4.5 Identity vs Equality

Having the same data is different from being the same actual object in memory:

```

[1, 2] == [1, 2]
True

[1, 2] is [1, 2]
False

```

The `==` operator checks, element by element, that two containers have the same data. The `is` operator checks that they are actually the same object.

But, and this point is really subtle, for immutables, the python language might save memory by reusing a single instantiated copy. This will always be safe.

```

"Hello" == "Hello"
True

"Hello" is "Hello"
<>:1: SyntaxWarning: "is" with a literal. Did you mean "=="?
<>:1: SyntaxWarning: "is" with a literal. Did you mean "=="?
/tmppipykernel_5872/3904443404.py:1: SyntaxWarning: "is" with a literal. Did you
mean "=="?
    "Hello" is "Hello"

True

```

This can be useful in understanding problems like the one above:

```

x = range(3)
y = x
z = x[:]

x == y

```

```

True
{x is y
True
{x == z
True
{x is z
False

```

1.5 Dictionaries

Estimated time for this notebook: 10 minutes

1.5.1 The Python Dictionary

Python supports a container type called a dictionary.

This is also known as an "associative array", "map" or "hash" in other languages.

In a list, we use a number to look up an element:

```

names = "Martin Luther King".split(" ")
names[1]
'Luther'

```

In a dictionary, we look up an element using **another object of our choice**:

```

me = {"name": "James", "age": 39, "Jobs": ["Programmer", "Teacher"]}
me
{'name': 'James', 'age': 39, 'Jobs': ['Programmer', 'Teacher']}
me["Jobs"]
['Programmer', 'Teacher']
me["age"]
39
type(me)
dict

```

Keys and Values

The things we can use to look up with are called **keys**:

```

me.keys()
dict_keys(['name', 'age', 'Jobs'])

```

The things we can look up are called **values**:

```

me.values()
dict_values(['James', 39, ['Programmer', 'Teacher']])

```

When we test for containment on a `dict` we test on the **keys**:

```

"Jobs" in me
True
"James" in me
False
"James" in me.values()
True

```

Immutable Keys Only

The way in which dictionaries work is one of the coolest things in computer science: the "hash table". The details of this are beyond the scope of this course, but we will consider some aspects in the section on performance programming.

One consequence of this implementation is that you can only use **immutable** things as keys.

```
{ good_match = {("Lamb", "Mint"): True, ("Bacon", "Chocolate"): False}
```

but:

```
{ illegal = [{"Lamb", "Mint": True, ["Bacon", "Chocolate": False]}
```

```
-----  
TypeError  
Cell In [14], line 1  
----> 1 illegal = [{"Lamb", "Mint": True, ["Bacon", "Chocolate": False]}  
TypeError: unhashable type: 'list'
```

Remember – square brackets denote lists, round brackets denote **tuples**.

Dictionary Order

Dictionaries will retain the order of the elements as they are defined (in Python versions ≥ 3.7).

```
{ my_dict = {"0": 0, "1": 1, "2": 2, "3": 3, "4": 4}  
print(my_dict)  
print(my_dict.values())}
```

```
{"0": 0, "1": 1, "2": 2, "3": 3, "4": 4}  
dict_values([0, 1, 2, 3, 4])
```

```
{ rev_dict = {"4": 4, "3": 3, "2": 2, "1": 1, "0": 0}  
print(rev_dict)  
print(rev_dict.values())}
```

```
{"4": 4, "3": 3, "2": 2, "1": 1, "0": 0}  
dict_values([4, 3, 2, 1, 0])
```

Python does not consider the order of the elements relevant to equality:

```
{ my_dict == rev_dict
```

```
True
```

1.5.2 Sets

A set is a **list** which cannot contain the same element twice. We make one by calling **set()** on any sequence, e.g. a list or string.

```
{ name = "James Hetherington"  
unique_letters = set(name)
```

```
{ unique_letters
```

```
{" ", "H", "J", "a", "e", "g", "h", "i", "m", "n", "o", "r", "s", "t"}
```

Or by defining a literal like a dictionary, but without the colons:

```
{ primes_below_ten = {2, 3, 5, 7}
```

```
{ type(unique_letters)
```

```
set
```

```
{ type(primes_below_ten)
```

```
set
```

```
{ unique_letters
```

```
{" ", "H", "J", "a", "e", "g", "h", "i", "m", "n", "o", "r", "s", "t"}
```

This will be easier to read if we turn the set of letters back into a string, with **join**:

```
{ "".join(unique_letters)
```

```
'am rJhoesigtHn'
```

A set has no particular order, but is really useful for checking or storing **unique** values.

Set operations work as in mathematics:

```
{ x = set("Hello")  
y = set("Goodbye")
```

```
{ x & y # Intersection
```

```
{'e', 'o'}
```

```
{ x | y # Union
```

```
{"G", "H", "b", "d", "e", "l", "o", "y"}
```

```
y - x # y intersection with complement of x: letters in Goodbye but not in Hello
```

```
{'G', 'b', 'd', 'y'}
```

Your programs will be faster and more readable if you use the appropriate container type for your data's meaning. Always use a set for lists which can't in principle contain the same data twice, always use a dictionary for anything which feels like a mapping from keys to values.

1.6 Data structures

Estimated time for this notebook: 5 minutes

1.6.1 Nested Lists and Dictionaries

In research programming, one of our most common tasks is building an appropriate *structure* to model our complicated data. Later in the course, we'll see how we can define our own types, with their own attributes, properties, and methods. But probably the most common approach is to use nested structures of lists, dictionaries, and sets to model our data. For example, an address might be modelled as a dictionary with appropriately named fields:

```
UCL = {"City": "London", "Street": "Gower Street", "Postcode": "WC1E 6BT"}
```

```
James = {"City": "London", "Street": "Waterson Street", "Postcode": "E2 8HH"}
```

A collection of people's addresses is then a list of dictionaries:

```
addresses = [UCL, James]
```

```
addresses
```

```
[{"City": "London", "Street": "Gower Street", "Postcode": "WC1E 6BT"},  
 {"City": "London", "Street": "Waterson Street", "Postcode": "E2 8HH"}]
```

A more complicated data structure, for example for a census database, might have a list of residents or employees at each address:

```
UCL["people"] = ["Clare", "James", "Owain"]
```

```
James["people"] = ["Sue", "James"]
```

```
addresses
```

```
[{"City": "London",  
 "Street": "Gower Street",  
 "Postcode": "WC1E 6BT",  
 "people": ["Clare", "James", "Owain"]},  
 {"City": "London",  
 "Street": "Waterson Street",  
 "Postcode": "E2 8HH",  
 "people": ["Sue", "James"]}]]
```

Which is then a list of dictionaries, with keys which are strings or lists.

We can go further, e.g.:

```
UCL["Residential"] = False
```

And we can write code against our structures:

```
leaders = [place["people"][0] for place in addresses]  
leaders
```

```
['Clare', 'Sue']
```

This was an example of a 'list comprehension', which have used to get data of this structure, and which we'll see more of in a moment...

1.7 Control and Flow

Estimated time for this notebook: 15 minutes

1.7.1 Turing completeness

Now that we understand how we can use objects to store and model our data, we only need to be able to control the flow of our program in order to have a program that can, in principle, do anything!

Specifically we need to be able to:

- Control whether a program statement should be executed or not, based on a variable. "Conditionality"
- Jump back to an earlier point in the program, and run some statements again. "Branching"

Once we have these, we can write computer programs to process information in arbitrary ways: we are *Turing Complete*!

1.7.2 Conditionality

Conditionality is achieved through Python's `if` statement:

```
x = 5  
if x < 0:  
    print(x, " is negative")
```

The absence of output here means the `if` clause prevented the `print` statement from running.

```
x = -10  
if x < 0:  
    print(x, " is negative")
```

```
-10  is negative
```

The first time through, the print statement never happened.

The **controlled** statements are indented. Once we remove the indent, the statements will once again happen regardless.

Else and Elif

Python's if statement has optional elif (else-if) and else clauses:

```
x = 5
if x < 0:
    print("x is negative")
else:
    print("x is positive")
```

```
x is positive
```

```
x = 5
if x < 0:
    print("x is negative")
elif x == 0:
    print("x is zero")
else:
    print("x is positive")
```

```
x is positive
```

Try editing the value of x here, and note that other sections are found.

```
choice = "high"
if choice == "high":
    print(1)
elif choice == "medium":
    print(2)
else:
    print(3)
```

```
1
```

1.7.3 Comparison

True and **False** are used to represent **boolean** (true or false) values.

```
1 > 2
```

```
False
```

Comparison on strings is alphabetical.

```
"UCL" > "KCL"
```

```
True
```

But case sensitive:

```
"UCL" > "kcl"
```

```
False
```

There's no automatic conversion of the **string** True to true:

```
True == "True"
```

```
False
```

And you cannot compare a string of a number to a number.

```
"1" < 2
```

```
-----  
TypeError                                 Traceback (most recent call last)  
Cell In [10], line 1  
----> 1 "1" < 2  
  
TypeError: '<' not supported between instances of 'str' and 'int'
```

```
"5" < 2
```

```
-----  
TypeError                                 Traceback (most recent call last)  
Cell In [11], line 1  
----> 1 "5" < 2  
  
TypeError: '<' not supported between instances of 'str' and 'int'
```

```
"1" > 2
```

```
-----  
TypeError                                 Traceback (most recent call last)  
Cell In [12], line 1  
----> 1 "1" > 2  
  
TypeError: '>' not supported between instances of 'str' and 'int'
```

Any statement that evaluates to **True** or **False** can be used to control an **if** Statement.

Automatic Falsehood

Various other things automatically count as true or false, which can make life easier when coding:

```
{ mytext = "Hello"  
  
if mytext:  
    print("Mytext is not empty")  
  
Mytext is not empty  
  
mytext2 = ""  
  
if mytext2:  
    print("Mytext2 is not empty")
```

We can use logical not and logical and to combine true and false:

```
x = 3.2  
if not (x > 0 and type(x) == int):  
    print(x, "is not a positive integer")  
  
3.2 is not a positive integer
```

`not` also understands magic conversion from false-like things to True or False.

```
not not "who's there!"  
  
True  
  
bool("")  
  
False  
  
bool("James")  
  
True  
  
bool([])  
  
False  
  
bool(["a"])  
  
True  
  
bool({})  
  
False  
  
bool({"name": "James"})  
  
True  
  
bool(0)  
  
False  
  
bool(1)  
  
True
```

But subtly, although these quantities evaluate True or False in an if statement, they're not themselves actually True or False under `==`:

```
[] == False  
  
False  
  
bool([]) == False  
  
True
```

1.7.4 Indentation

In Python, indentation is semantically significant. You can choose how much indentation to use, so long as you are consistent, but four spaces is conventional. Please do not use tabs.

In the notebook, and most good editors, when you press `<tab>`, you get four spaces.

No indentation when it is expected, results in an error:

```
x = 2  
  
if x > 0:  
    print(x)
```

```
Cell In [30], line 2
  print(x)
  ^
IndentationError: expected an indented block
```

but:

```
{ if x > 0:
    print(x)
```

```
2
```

1.7.5 Pass

A statement expecting indentation must have some indented code. This can be annoying when commenting things out. (With #)

```
{ if x > 0:
    # print x
    print("Hello")
```

```
Cell In [32], line 4
  print("Hello")
  ^
IndentationError: expected an indented block
```

So the `pass` statement is used to do nothing.

```
{ if x > 0:
    # print x
    pass
```

```
print("Hello")
```

```
Hello
```

1.8 Iteration

Estimated time for this notebook: 10 minutes

Our other aspect of control is looping back on ourselves.

We use `for ... in` to "iterate" over lists:

```
{ mylist = [3, 7, 15, 2]
```

```
{ for whatever in mylist:
    print(whatever**2)
```

```
9
49
225
4
```

Each time through the loop, the variable in the `value` slot is updated to the `next` element of the sequence.

1.8.1 Iterables

Any sequence type is iterable:

```
vowels = "aeiou"
sarcasm = []

for letter in "Okay":
    if letter.lower() in vowels:
        repetition = 3
    else:
        repetition = 1

    sarcasm.append(letter * repetition)

"".join(sarcasm)
```

```
'000kaay'
```

The above is a little puzzle, work through it to understand why it does what it does.

Dictionaries are iterables

All sequences are iterables. Some iterables (things you can `for` loop over) are not sequences (things with you can do `x[5]` to), for example sets and dictionaries.

```
current_year = 2022
founded = {"Barack Obama": 1961, "UCL": 1826, "The Alan Turing Institute": 2015}

for thing in founded:
    print(f"In {current_year} {thing} is {current_year - founded[thing]} years old.")
```

```
In 2022 Barack Obama is 61 years old.
In 2022 UCL is 196 years old.
In 2022 The Alan Turing Institute is 7 years old.
```

1.8.2 Unpacking and Iteration

Unpacking can be useful with iteration:

```
{ triples = [[4, 11, 15], [39, 4, 18]]
```

```
for whatever in triples:  
    print(whatever)
```

```
[4, 11, 15]  
[39, 4, 18]
```

```
for first, middle, last in triples:  
    print(middle)
```

```
11  
4
```

```
# A reminder that the words you use for variable names are arbitrary:  
for hedgehog, badger, fox in triples:  
    print(badger)
```

```
11  
4
```

for example, to iterate over the items in a dictionary as pairs:

```
things = {  
    "James": [1976, "Kendal"],  
    "UCL": [1826, "Bloomsbury"],  
    "Cambridge": [1209, "Cambridge"],  
}  
print(things.items())
```

```
dict_items([('James', [1976, 'Kendal']), ('UCL', [1826, 'Bloomsbury']),  
('Cambridge', [1209, 'Cambridge']))]
```

```
for name, year in founded.items():  
    print(name, "is", current_year - year, "years old.")
```

```
Barack Obama is 61 years old.  
UCL is 196 years old.  
The Alan Turing Institute is 7 years old.
```

1.8.3 Break, Continue

- Continue skips to the next turn of a loop
- Break stops the loop early

```
for n in range(50):  
    if n == 20:  
        break  
    if n % 2 == 0:  
        continue  
    print(n)
```

```
1  
3  
5  
7  
9  
11  
13  
15  
17  
19
```

These aren't useful that often, but are worth knowing about. There's also an optional `else` clause on loops, executed only if you don't `break`, but I've never found that useful.

1.9 Classroom Exercises

List of exercises and estimated completion times

[1a - Python Libraries](#) 5 minutes

[1b - Using Functions](#) 10 minutes

[1c - Operators](#) 10 minutes

[1d - Maze Model](#) 25 minutes

[1e - The Maze Population](#) 10 minutes

Exercise 1a Python Libraries

Relevant Sections: 1.0.2

The directory that contains this workbook also contains a `Python` file titled `draw_infinity.py`. Import it to a notebook and make the figure in the same way as `eight` was drawn in section 1.0.2

Exercise 1b Using Functions

Relevant Sections: 1.2.1 to 1.2.5

Try to find the operator or function you need to calculate the following (the easiest way might be an internet search).

What is 2 to the power 15?

Convert "It was the best of times" to uppercase.

Sort the list `[10, 9, 0, 20, 8, 2, 30, 7, 3]`.

What is $100!$? (That is, what is the factorial of 100?) Hint: the `factorial` function is in the `math` library

Exercise 1c Operators

Relevant Sections: 1.2.5, 1.3.3

Which of the operators `+`, `-`, `*`, and `/` do something useful with the lists `[1, 10, 100]` and `[5, 4, 7]`?

What happens if you apply the operators `+`, `-`, `*`, `/` to a list and a number?

What about a string and a string?

Exercise 1d Maze Model

Relevant Sections: 1.5.1, 1.6.1

Work with a partner to design a data structure to represent a maze using dictionaries and lists.

- Each place in the maze has a name, which is a string.
- Each place in the maze has one or more people currently standing at it, by name.
- Each place in the maze has a maximum capacity of people that can fit in it.
- From each place in the maze, you can go from that place to a few other places, using a direction like 'up', 'north', or 'sideways'

Create an example instance, in a notebook, of a simple structure for your maze:

- The front room can hold 2 people. James is currently there. You can go outside to the garden, or upstairs to the bedroom, or north to the kitchen.
- From the kitchen, you can go south to the front room. It fits 1 person.
- From the garden you can go inside to front room. It fits 3 people. Sue is currently there.
- From the bedroom, you can go downstairs to the front room. You can also jump out of the window to the garden. It fits 2 people.

Make sure that your model:

- Allows empty rooms
- Allows you to jump out of the upstairs window, but not to fly back up.
- Allows rooms which people can't fit in.

```
house = [ "Your answer here" ]
```

or

```
house = { "Your answer here" }
```

Exercise 1e The Maze Population

Relevant Sections: 1.5.1, 1.6.1, 1.8.1, 1.8.2

Take your maze data structure. Write a program to count the total number of people in the maze, and also determine the total possible occupants.

2. Intermediate Python

- List comprehensions
- Functions in Python
- Modules in Python
- An introduction to classes
- Working with files
- Interacting with the internet
- Classroom Exercises

Contents

- [2.0 Comprehensions](#) (10 minutes)
- [2.1 Functions](#) (15 minutes)
- [2.2 Using Libraries](#) (5 minutes)
- [2.3 Working with files](#) (15 minutes)
- [2.4 Getting data from the Internet](#) (10 minutes)
- [2.5 Data analysis example](#) (20 minutes)
- [2.6 Defining your own classes](#) (20 minutes)
- [2.7 Data analysis with classes](#) (10 minutes)

Total time: 1 hr 45 minutes

Exercises

Classroom exercises are grouped together at the end of the module: 2.7 Classroom Exercises. Each exercise is labelled with any sections whose contents are relevant. We recommend that instructors schedule the exercises to be done in groups during breaks in the taught content. However, it is **important** that participants also have some time away from their screens. Exercises can also be left as self-paced homework assignments if preferred.

2.0 Comprehensions

Estimated time for this notebook: 10 minutes

2.0.1 The list comprehension

If you write a for loop **inside** a pair of square brackets for a list, you magic up a list as defined. This can make for concise but hard to read code, so be careful.

```
[2**x for x in range(10)]
```

```
[1, 2, 4, 8, 16, 32, 64, 128, 256, 512]
```

Which is equivalent to the following code without using comprehensions:

```
result = []
for x in range(10):
    result.append(2**x)
result
```

```
[1, 2, 4, 8, 16, 32, 64, 128, 256, 512]
```

You can do quite weird and cool things with comprehensions:

```
[len(str(2**x)) for x in range(10)]
```

```
[1, 1, 1, 1, 2, 2, 2, 3, 3, 3]
```

2.0.2 Selection in comprehensions

You can write an `if` statement in comprehensions too:

```
[2**x for x in range(30) if x % 3 == 0]
```

```
[1, 8, 64, 512, 4096, 32768, 262144, 2097152, 16777216, 134217728]
```

Consider the following, and make sure you understand why it works:

```
".join([letter for letter in "James Hetherington" if letter.lower() not in
"aeiou"])
```

```
'Jms Hthrngtn'
```

2.0.3 Comprehensions versus building lists with `append`:

This code:

```
result = []
for x in range(30):
    if x % 3 == 0:
        result.append(2**x)
result
```

```
[1, 8, 64, 512, 4096, 32768, 262144, 2097152, 16777216, 134217728]
```

Does the same as the comprehension above. The comprehension is generally considered more readable.

Comprehensions are therefore an example of what we call 'syntactic sugar': they do not increase the capabilities of the language.

Instead, they make it possible to write the same thing in a more readable way.

Almost everything we learn from now on will be either syntactic sugar or interaction with something other than idealised memory, such as a storage device or the internet. Once you have variables, conditionality, and branching, your language can do anything. (And this can be proved.)

2.0.4 Nested comprehensions

If you write two `for` statements in a comprehension, you get a single array generated over all the pairs:

```
[x - y for x in range(4) for y in range(4)]
```

```
[0, -1, -2, -3, 1, 0, -1, -2, 2, 1, 0, -1, 3, 2, 1, 0]
```

You can select on either, or on some combination:

```
[x - y for x in range(4) for y in range(4) if x >= y]
```

```
[0, 1, 0, 2, 1, 0, 3, 2, 1, 0]
```

If you want something more like a matrix, you need to do *two nested* comprehensions!

```
[[x - y for x in range(4)] for y in range(4)]
```

```
[[0, 1, 2, 3], [-1, 0, 1, 2], [-2, -1, 0, 1], [-3, -2, -1, 0]]
```

Note the subtly different square brackets.

Note that the list order for multiple or nested comprehensions can be confusing:

```
[x + y for x in ["a", "b", "c"] for y in ["1", "2", "3"]]
```

```
['a1', 'a2', 'a3', 'b1', 'b2', 'b3', 'c1', 'c2', 'c3']
```

```
[[x + y for x in ["a", "b", "c"]] for y in ["1", "2", "3"]]
```

```
[['a1', 'b1', 'c1'], ['a2', 'b2', 'c2'], ['a3', 'b3', 'c3']]
```

2.0.5 Dictionary Comprehensions

You can automatically build dictionaries, by using a list comprehension syntax, but with curly brackets and a colon:

```
{(str(x)) * 3: x for x in range(3)}
```

```
{'000': 0, '111': 1, '222': 2}
```

2.0.6 List-based thinking

Once you start to get comfortable with comprehensions, you find yourself working with containers, nested groups of lists and dictionaries, as the 'things' in your program, not individual variables.

Given a way to analyse some dataset, we'll find ourselves writing stuff like:

```
analysed_data = [analyze(datum) for datum in data]
```

There are lots of built-in methods that provide actions on lists as a whole:

```
{ any([True, False, True])}
```

```
True
```

```
{ all([True, False, True])}
```

```
False
```

```
{ max([1, 2, 3])}
```

```
3
```

```
{ sum([1, 2, 3])}
```

```
6
```

My favourite is `map`, which, similar to a list comprehension, applies one function to every member of a list:

```
{ [str(x) for x in range(10)]}
```

```
['0', '1', '2', '3', '4', '5', '6', '7', '8', '9']
```

```
{ list(map(str, range(10)))}
```

```
['0', '1', '2', '3', '4', '5', '6', '7', '8', '9']
```

So I can write:

```
analysed_data = map(analyse, data)
```

We'll learn more about `map` and similar functions when we discuss functional programming later in the course.

2.1 Functions

Estimated time for this notebook: 15 minutes

Defining **functions** which put together code to make a more complex task seem simple from the outside is the most important thing in programming. We can wrap code up in a **function**, so that we can repeatedly get just the information we want.

2.1.1 Definition

We use `def` to define a function, and `return` to pass back a value: The input comes in in brackets after the function name:

```
{ def double(x):
```

```
    return x * 2
```

```
{ print(double(5), double([5]), double("five"))}
```

```
10 [5, 5] fivefive
```

2.1.2 Default Parameters

We can specify default values for parameters:

```
{ def jeeves(name="Sir"):
```

```
    return f"Very good, {name}"
```

```
{ jeeves()
```

```
'Very good, Sir'
```

```
{ jeeves("James")}
```

```
'Very good, James'
```

If you have some parameters with defaults, and some without, those with defaults **must** go later.

If you have multiple default arguments, you can specify neither, one or both:

```
{ def jeeves(greeting="Very good", name="Sir"):
```

```
    return f"{greeting}, {name}"
```

```
{ jeeves()
```

```
'Very good, Sir'
```

```

{jeeves("Hello")
'Hello, Sir'

{jeeves(name="James")
'Very good, James'

{jeeves(greeting="Suits you")
'Suits you, Sir'

{jeeves("Hello", "Sailor")
'Hello, Sailor'

```

2.1.3 Early Return

Return without arguments can be used to exit early from a function

Here's a slightly convoluted example of a function which will return early under specific conditions. In this case if a list contains the string 'cat'.

```

def are_there_cats(my_input_list):
    if "cat" in my_input_list: # If the string "cat" is in the list
        print("There is a cat in here") # print a statement to screen
        return
    print("Nothing to see here")

first_list = ["cat", "dog", "hamster", 42]
second_list = ["duck", 17, "elk"]

are_there_cats(first_list)
There is a cat in here

are_there_cats(second_list)
Nothing to see here

```

2.1.4 Scoping

There are differences in how variables and names are accessed by your code based on where they are defined.

Within this notebook any variables that have been defined outside of a function will be available to the rest of the notebook. At this point in the notebook, x has not been defined.

```

x

NameError                                 Traceback (most recent call last)
Cell In [15], line 1
----> 1 x

NameError: name 'x' is not defined

```

If we now define x and write and call a function in which uses it; the function can still access x, even if x isn't given as an argument.

```

x = 5 # Define x now

def can_we_see_x():
    print(f"x = {x}")

can_we_see_x()

x = 5

```

However if we define y locally - in a function - we can access it from within that function:

```

def can_we_see_y():
    y = 7 # Define y in the function
    print(f"x = {x}")
    print(f"y = {y}")

can_we_see_y()

x = 5
y = 7

```

However y isn't accessible globally - that is it isn't available outside of the function in which it was defined

```

y

NameError                                 Traceback (most recent call last)
<ipython-input-18-9063a9f0e032> in <module>
----> 1 y

NameError: name 'y' is not defined

```

Note for the two functions above we used syntax for building strings that contain the values of variables. You can read more about it [here](#) or in the official documentation for formatted string literals; [f-strings](#).

2.1.5 Side effects

Functions can do things to change their **mutable** arguments, so `return` is optional.

This is pretty awful style, in general, functions should normally be side-effect free.

Here is a contrived example of a function that makes plausible use of a side-effect

```
def double_inplace(vec):
    vec[:] = [element * 2 for element in vec]

z = list(range(4))
double_inplace(z)
print(z)

[0, 2, 4, 6]

letters = ["a", "b", "c", "d", "e", "f", "g"]
letters[:] = []

[ ]
```

In this example, we're using `[:]` to access into the same list, and write its data.

```
vec = [element*2 for element in vec]
```

would just move a local label, not change the input.

See Module 1.5 - Memory and Containers for a refresher

But I'd usually just write this as a function which **returned** the output:

```
def double(vec):
    return [element * 2 for element in vec]
```

Let's remind ourselves of the behaviour for modifying lists in-place using `[:]` with a simple array:

```
x = 5
x = 7
x = ["a", "b", "c"]
y = x

[ ]

['a', 'b', 'c']

x[:] = ["Hooray!", "Yippee"]

[ ]

y

[ 'Hooray!', 'Yippee']
```

2.1.6 Unpacking arguments

```
def arrow(before, after):
    return str(before) + " -> " + str(after)

arrow(1, 3)

[ ]

'1 -> 3'
```

If a function that takes multiple arguments is given an iterable object prepended with `*`, each element of that object is taken in turn and used to fill the function's arguments one-by-one.

```
x = [1, -1]
arrow(*x)

[ ]

'1 -> -1'
```

This can be quite powerful:

```
charges = {"neutron": 0, "proton": 1, "electron": -1}
for particle in charges.items():
    print(arrow(particle))

[ ]

neutron -> 0
proton -> 1
electron -> -1
```

2.1.7 Sequence Arguments

Similarly, if a `*` is used in the **definition** of a function, multiple arguments are absorbed into a list **inside** the function:

```
def doubler(*sequence):
    return [x * 2 for x in sequence]

doubler(1, 2, 3)

[2, 4, 6]

doubler(5, 2, "Wow!")
```

```
[10, 4, 'Wow!Wow! ']
```

2.1.8 Keyword Arguments

If two asterisks are used, named arguments are supplied inside the function as a dictionary:

```
def arrowify(**args):
    for key, value in args.items():
        print(key + " -> " + value)

arrowify(neutron="n", proton="p", electron="e")
```

```
neutron -> n
proton -> p
electron -> e
```

These different approaches can be mixed:

```
def somefunc(a, b, *args, **kwargs):
    print("A:", a)
    print("B:", b)
    print("args:", args)
    print("keyword args", kwargs)
```

```
somefunc(1, 2, 3, 4, 5, fish="Haddock")
```

```
A: 1
B: 2
args: (3, 4, 5)
keyword args {'fish': 'Haddock'}
```

2.2 Using Libraries

Estimated time for this notebook: 5 minutes

2.2.1 Import

Research programming is all about using libraries: tools other people have provided programs that do many cool things. By combining them we can feel really powerful but doing minimum work ourselves.

The python syntax to import someone else's library is "import". To use a function or type from a python library, rather than a **built-in** function or type, we have to import the library.

```
math.sin(1.6)
```

```
NameError                                 Traceback (most recent call last)
Cell In [1], line 1
----> 1 math.sin(1.6)

NameError: name 'math' is not defined
```

```
import math
```

```
math.sin(1.6)
```

```
0.9995736030415051
```

We call these libraries **modules**:

```
type(math)
```

```
module
```

The tools supplied by a module are *attributes* of the module, and as such, are accessed with a dot.

```
dir(math)
```

```
['__doc__',  
 '__file__',  
 '__loader__',  
 '__name__',  
 '__package__',  
 '__spec__',  
 'acos',  
 'acosh',  
 'asin',  
 'asinh',  
 'atan',  
 'atan2',  
 'atanh',  
 'ceil',  
 'comb',  
 'copysign',  
 'cos',  
 'cosh',  
 'degrees',  
 'dist',  
 'e',  
 'erf',  
 'erfc',  
 'exp',  
 'expm1',  
 'fabs',  
 'factorial',  
 'floor',  
 'fmod',  
 'frexp',  
 'fsum',  
 'gamma',  
 'gcd',  
 'hypot',  
 'inf',  
 'isclose',  
 'isfinite',  
 'isinf',  
 'isnan',  
 'isqrt',  
 'ldexp',  
 'lgamma',  
 'log',  
 'log10',  
 'log1p',  
 'log2',  
 'modf',  
 'nan',  
 'perm',  
 'pi',  
 'pow',  
 'prod',  
 'radians',  
 'remainder',  
 'sin',  
 'sinh',  
 'sqrt',  
 'tan',  
 'tanh',  
 'tau',  
 'trunc']
```

They include properties as well as functions:

```
math.pi
```

```
3.141592653589793
```

You can always find out where on your storage medium a library has been imported from:

```
print(math.__file__[0:50])  
print(math.__file__[50:])
```

```
/opt/hostedtoolcache/Python/3.8.14/x64/lib/python3  
.8/lib-dynload/math.cpython-38-x86_64-linux-gnu.so
```

Note that `import` does *not* install libraries. It just makes them available to your current notebook session, assuming they are already installed. Installing libraries is harder, and we'll cover it later. So what libraries are available? Until you install more, you might have just the modules that come with Python, the *standard library*.

Supplementary Materials: Review the list of standard library modules: <https://docs.python.org/library/>

If you installed via Anaconda, then you also have access to a bunch of modules that are commonly used in research.

Supplementary Materials: Review the list of modules that are packaged with Anaconda by default on different architectures:

<https://docs.anaconda.com/anaconda/packages/pkg-docs/> (modules installed by default are shown with ticks)

We'll see later how to add more libraries to our setup.

Why bother?

Why bother with modules? Why not just have everything available all the time?

The answer is that there are only so many names available! Without a module system, every time I made a variable whose name matched a function in a library, I'd lose access to it. In the olden days, people ended up having to make really long variable names, thinking their names would be unique, and they still ended up with "name clashes". The module mechanism avoids this.

2.2.2 Importing from modules

Still, it can be annoying to have to write `math.sin(math.pi)` instead of `sin(pi)`. Things can be imported *from* modules to become part of the current module:

```
import math  
math.sin(math.pi)
```

```
1.2246467991473532e-16
```

```
from math import sin  
sin(math.pi)
```

```
1.2246467991473532e-16
```

Importing one-by-one like this is a nice compromise between typing and risk of name clashes.

It is possible to import **everything** from a module, but you risk name clashes.

```
from math import *  
sin(pi)
```

```
1.2246467991473532e-16
```

Import and rename

You can rename things as you import them to avoid clashes or for typing convenience

```
import math as m  
m.cos(0)
```

```
1.0
```

```
pi = 3  
from math import pi as realpi  
print(sin(pi), sin(realpi))
```

```
0.1411200080598672 1.2246467991473532e-16
```

2.3 Working with files

Estimated time for this notebook: 15 minutes

2.3.1 Background

Loading data from files

An important part of this course is about using Python to analyse and visualise data. Most data, of course, is supplied to us in various *formats*: spreadsheets, database dumps, or text files in various formats (csv, tsv, json, yaml, hdf5, netcdf). It is also stored in some *medium*: on a local disk, a network drive, or on the internet in various ways. It is important to distinguish the data format, how the data is structured into a file, from the data's storage, where it is put.

We'll look first at the question of data *transport*: loading data from a disk, and at downloading data from the internet. Then we'll look at data *parsing*: building Python structures from the data. These are related, but separate questions.

An example datafile

Let's write an example datafile to disk so we can investigate it. We'll just use a plain-text file. Jupyter notebook provides a way to do this: if we put `%%writefile` at the top of a cell, instead of being interpreted as python, the cell contents are saved to disk.

```
%%writefile mydata.txt  
A poet once said, 'The whole universe is in a glass of wine.'  
We will probably never know in what sense he meant it,  
for poets do not write to be understood.  
But it is true that if we look at a glass of wine closely enough we see the entire  
universe.  
There are the things of physics: the twisting liquid which evaporates depending  
on the wind and weather, the reflection in the glass;  
and our imagination adds atoms.  
The glass is a distillation of the earth's rocks,  
and in its composition we see the secrets of the universe's age, and the evolution  
of stars.  
What strange array of chemicals are in the wine? How did they come to be?  
There are the ferment, the enzymes, the substrates, and the products.  
There in wine is found the great generalization; all life is fermentation.  
Nobody can discover the chemistry of wine without discovering,  
as did Louis Pasteur, the cause of much disease.  
How vivid is the claret, pressing its existence into the consciousness that  
watches it!  
If our small minds, for some convenience, divide this glass of wine, this  
universe,  
into parts --  
physics, biology, geology, astronomy, psychology, and so on --  
remember that nature does not know it!  
  
So let us put it all back together, not forgetting ultimately what it is for.  
Let it give us one more final pleasure; drink it and forget it all!  
- Richard Feynman
```

```
Writing mydata.txt
```

Where did that go? It went to the current folder, which for a notebook, by default, is where the notebook is on disk.

```
import os # The 'os' module gives us all the tools we need to search in the file  
system  
  
os.getcwd() # Use the 'getcwd' function from the 'os' module to find where we are  
on disk.
```

```
'/home/runner/work/rse-course/rse-course/module02_intermediate_python'
```

Can we see if it is there?

```
import os  
  
[x for x in os.listdir(os.getcwd()) if ".txt" in x]
```

```
['mydata.txt']
```

Yep! Note how we used a list comprehension to filter all the extraneous files.

2.4.2 Path independence and `os`

We can use `dirname` to get the parent folder for a folder, in a platform independent-way.

```
{ os.path.dirname(os.getcwd()) }
```

```
'/home/runner/work/rse-course/rse-course'
```

We could do this manually using `split`:

```
{ "/".join(os.getcwd().split("/")[:-1]) }
```

```
'/home/runner/work/rse-course/rse-course'
```

But this would not work on Windows, where path elements are separated with a `\` instead of a `/`. So it's important to use `os.path` for this stuff.

Supplementary Materials: If you're not already comfortable with how files fit into folders, and folders form a tree, with folders containing subfolders, then look at <http://swcarpentry.github.io/shell-novice/02-filedir/index.html>.

Satisfy yourself that after using `%>writetfile`, you can then find the file on disk with Windows Explorer, OSX Finder, or the Linux Shell.

We can see how in Python we can investigate the file system with functions in the `os` module, using just the same programming approaches as for anything else.

We'll gradually learn more features of the `os` module as we go, allowing us to move around the disk, `walk` around the disk looking for relevant files, and so on. These will be important to master for automating our data analyses.

2.3.3 The python `file` type

So, let's read our file:

```
{ myfile = open("mydata.txt") }
```

```
{ type(myfile) }
```

```
_io.TextIOWrapper
```

We can go line-by-line, by treating the file as an iterable:

```
{ [x for x in myfile]
```

```
["A poet once said, 'The whole universe is in a glass of wine.'\n",
 'We will probably never know in what sense he meant it, \\n',
 'for poets do not write to be understood. \\n',
 'But it is true that if we look at a glass of wine closely enough we see the
 entire universe. \\n',
 'There are the things of physics: the twisting liquid which evaporates
 depending\\n',
 'on the wind and weather, the reflection in the glass;\\n',
 'and our imagination adds atoms.\\n',
 'The glass is a distillation of the earth's rocks,\\n",
 "and in its composition we see the secrets of the universe's age, and the
 evolution of stars. \\n",
 'What strange array of chemicals are in the wine? How did they come to be? \\n',
 'There are the ferment, the enzymes, the substrates, and the products.\\n',
 'There in wine is found the great generalization; all life is fermentation.\\n',
 'Nobody can discover the chemistry of wine without discovering, \\n',
 'as did Louis Pasteur, the cause of much disease.\\n',
 'How vivid is the claret, pressing its existence into the consciousness that
 watches it!\\n',
 'If our small minds, for some convenience, divide this glass of wine, this
 universe, \\n',
 'into parts -- \\n',
 'physics, biology, geology, astronomy, psychology, and so on -- \\n',
 'remember that nature does not know it!\\n',
 '\\n',
 'So let us put it all back together, not forgetting ultimately what it is
 for.\\n',
 'Let it give us one more final pleasure; drink it and forget it all!\\n',
 ' - Richard Feynman\\n']
```

If we do that again, the file has already finished, there is no more data.

```
{ [x for x in myfile]
```

```
[]
```

We need to 'rewind' it!

```
{ myfile.seek(0)
```

```
[len(x) for x in myfile if "know" in x]
```

```
[56, 39]
```

It's really important to remember that a file is a *different* built in type than a string.

2.3.4 Reading Files

We can read one line at a time with `readline`:

```
{ myfile.seek(0)
 first = myfile.readline()
```

```
{ first
```

```
"A poet once said, 'The whole universe is in a glass of wine.'\\n"
```

```
{ second = myfile.readline()
```

```
{ second
```

```
'We will probably never know in what sense he meant it, \n'
```

We can read the whole remaining file with `read`:

```
{ rest = myfile.read()
```

```
{ rest
```

```
"for poets do not write to be understood. \nBut it is true that if we look at a glass of wine closely enough we see the entire universe. \nThere are the things of physics: the twisting liquid which evaporates depending\non the wind and weather, the reflection in the glass;\nand our imagination adds atoms.\nThe glass is a distillation of the earth's rocks,\nand in its composition we see the secrets of the universe's age, and the evolution of stars. \nwhat strange array of chemicals are in the wine? How did they come to be? \nThere are the fermentations, the enzymes, the substrates, and the products.\nThere in wine is found the great generalization; all life is fermentation.\nNobody can discover the chemistry of wine without discovering, \nas did Louis Pasteur, the cause of much disease.\nHow vivid is the claret, pressing its existence into the consciousness that watches it!\nif our small minds, for some convenience, divide this glass of wine, this universe, \ninto parts -- \nphysics, biology, geology, astronomy, psychology, and so on -- \nremember that nature does not know it!\n\nSo let us put it all back together, not forgetting ultimately what it is for.\nLet it give us one more final pleasure; drink it and forget it all!\n - Richard Feynman\n"
```

Which means that when a file is first opened, `read` is useful to just get the whole thing as a string:

```
{ open("mydata.txt").read()
```

```
"A poet once said, 'The whole universe is in a glass of wine.'\nWe will probably never know in what sense he meant it, \nfor poets do not write to be understood.\nBut it is true that if we look at a glass of wine closely enough we see the entire universe. \nThere are the things of physics: the twisting liquid which evaporates depending\non the wind and weather, the reflection in the glass;\nand our imagination adds atoms.\nThe glass is a distillation of the earth's rocks,\nand in its composition we see the secrets of the universe's age, and the evolution of stars. \nwhat strange array of chemicals are in the wine? How did they come to be? \nThere are the fermentations, the enzymes, the substrates, and the products.\nThere in wine is found the great generalization; all life is fermentation.\nNobody can discover the chemistry of wine without discovering, \nas did Louis Pasteur, the cause of much disease.\nHow vivid is the claret, pressing its existence into the consciousness that watches it!\nif our small minds, for some convenience, divide this glass of wine, this universe, \ninto parts -- \nphysics, biology, geology, astronomy, psychology, and so on -- \nremember that nature does not know it!\n\nSo let us put it all back together, not forgetting ultimately what it is for.\nLet it give us one more final pleasure; drink it and forget it all!\n - Richard Feynman\n"
```

You can also read just a few characters:

```
{ myfile.seek(1335)
```

```
1335
```

```
{ myfile.read(15)
```

```
'\n - Richard F'
```

2.3.5 Converting Strings to Files

Because files and strings are different types, we CANNOT just treat strings as if they were files:

```
{ mystring = "Hello World\n My name is James"
```

```
{ mystring
```

```
'Hello World\n My name is James'
```

```
{ mystring.readline()
```

```
AttributeError: 'str' object has no attribute 'readline'
```

This is important, because some file format parsers expect input from a `file` and not a string. We can convert between them using the `StringIO` class of the [io module](#) in the standard library:

```
{ from io import StringIO
```

```
{ mystringasafile = StringIO(mystring)
```

```
{ mystringasafile.readline()
```

```
'Hello World\n'
```

```
{ mystringasafile.readline()
```

```
' My name is James'
```

Note that in a string, `\n` is used to represent a newline.

2.4.6 Closing files

We really ought to close files when we've finished with them, as it makes the computer more efficient. (On a shared computer, this is particularly important)

```
{ myfile.close()
```

Because it's so easy to forget this, python provides a **context manager** to open a file, then close it automatically at the end of an indented block:

```
with open("mydata.txt") as somefile:  
    content = somefile.read()  
content
```

```
"A poet once said, 'The whole universe is in a glass of wine.'\nWe will probably  
never know in what sense he meant it, \nfor poets do not write to be understood.  
\nBut it is true that if we look at a glass of wine closely enough we see the  
entire universe. \nThere are the things of physics: the twisting liquid which  
evaporates depending\non the wind and weather, the reflection in the glass;\nand  
our imagination adds atoms.\nThe glass is a distillation of the earth's  
rocks,\nand in its composition we see the secrets of the universe's age, and the  
evolution of stars. \nwhat strange array of chemicals are in the wine? How did  
they come to be? \nThere are the fermentations, the enzymes, the substrates, and the  
products.\nThere in wine is found the great generalization; all life is  
fermentation.\nNobody can discover the chemistry of wine without discovering, \nas  
did Louis Pasteur, the cause of much disease.\nHow vivid is the claret, pressing  
its existence into the consciousness that watches it!\nIf our small minds, for  
some convenience, divide this glass of wine, this universe, \ninto parts --  
\nphysics, biology, geology, astronomy, psychology, and so on -- \nremember that  
nature does not know it!\n\nSo let us put it all back together, not forgetting  
ultimately what it is for.\nLet it give us one more final pleasure; drink it and  
forget it all!\n - Richard Feynman\n"
```

The code to be done while the file is open is indented, just like for an `if` statement.

You should pretty much **always** use this syntax for working with files.

2.3.7 Writing files

We might want to create a file from a string in memory. We can't do this with the notebook's `%%writefile` – this is just a notebook convenience, and isn't very programmable.

When we open a file, we can specify a 'mode', in this case, 'w' for writing. ('r' for reading is the default.)

```
with open("mywrittenfile", "w") as target:  
    target.write("Hello")  
    target.write("World")
```

```
with open("mywrittenfile", "r") as source:  
    print(source.read())
```

```
HelloWorld
```

And we can "append" to a file with mode 'a':

```
with open("mywrittenfile", "a") as target:  
    target.write("Hello")  
    target.write("James")
```

```
with open("mywrittenfile", "r") as source:  
    print(source.read())
```

```
HelloWorldHelloJames
```

If a file already exists, mode `w` will overwrite it.

2.4 Getting data from the Internet

Estimated time for this notebook: 10 minutes

We've seen about obtaining data from our local file system.

The other common place today that we might want to obtain data is from the internet.

It's very common today to treat the web as a source and store of information; we need to be able to programmatically download data, and place it in Python objects.

We may also want to be able to programmatically *upload* data, for example, to automatically fill in forms.

This can be really powerful if we want to, for example, do automated metaanalysis across a selection of research papers.

2.4.1 URLs

All internet resources are defined by a Uniform Resource Locator.

https://static-maps.yandex.ru:443/1.x/?z=12&size=400%2C400&ll=-0.1275%2C51.51&l=sat&lang=en_US

A url consists of:

- A *scheme* (http, https, ssh, ...)
- A *host* (static-maps.yandex.ru), the name of the remote computer you want to talk to)
- A *port* (optional, most protocols have a typical port associated with them, e.g. 443 for https)
- A *path* (Like a file path on the machine, here it is `1.x`)
- A *query* part after a ?, (optional, usually ampersand-separated *parameters* e.g. `lang=en_US`, or `z=12`)

Supplementary materials: These can actually be different for different protocols, the above is a simplification, you can see more, for example, at

https://en.wikipedia.org/wiki/URL_scheme

URLs are not allowed to include all characters; we need to, for example, "escape" a space that appears inside the URL, replacing it with `%20`, so e.g. a request of `http://some example.com/` would need to be `http://some%20example.com/`. In the URL above, the comma in the size parameter value `size=400,400` has to be replaced with `%2C` to give `size=400%2C400`.

Supplementary materials: The code used to replace each character is the [ASCII](#) code for it.

Supplementary materials: The escaping rules are quite subtle. See <https://en.wikipedia.org/wiki/Percent-encoding>. The standard library provides the `urllib.parse` function that can take care of this for you.

2.4.2 Requests

The python `requests` library can help us manage and manipulate URLs. It is easier to use than the 'urllib' library that is part of the standard library, and is included with anaconda and canopy. It sorts out escaping, parameter encoding, and so on for us.

To request the above URL, for example, we write:

```
import requests

response = requests.get(
    "https://static-maps.yandex.ru:443/1.x",
    params={
        "size": "400,400", # size of map
        "ll": "-0.1275,51.51", # longitude & latitude of centre
        "z": 12, # zoom level
        "l": "sat", # map layer (satellite image)
        "lang": "en_US", # language
    },
    timeout=60,
```

When we do a request, the result comes back as text. For the png image in the above, this isn't very readable:

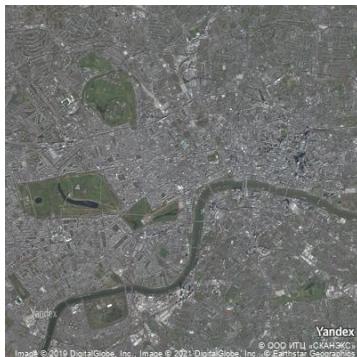
```
response.content[0:50]
```



```
b'\xff\xd8\xff\xe0\x00\x10JFIF\x00\x01\x01\x00H\x00H\x00\x00\xff\xdb\x00C\x00\x08\x06\x06\x07\x06\x05\x08\x07\x07\x07\t\t\x08\x0c\x14\r\x0c\x0b\x0b\x0c\x19\x12\x13\x0f'
```

Just as for file access, therefore, we will need to send the text we get to a python module which understands that file format.

```
from IPython.display import Image
Image(response.content)
```



Again, it is important to separate the *transport* model (e.g. a file system, or an "http request" for the web) from the data model of the data that is returned.

2.4.3 Example: Sunspots

Let's try to get something scientific: the sunspot cycle data from <http://sidc.be/silso/home>

```
spots = requests.get("http://www.sidc.be/silso/INFO/snmtotcsv.php",
                     timeout=60).text
```



```
spots[0:80]
```



```
'1749;01;1749.042; 96.7; -1.0; -1;1\n1749;02;1749.123; 104.3; -1.0;
-1;1\n1749'
```

This looks like semicolon-separated data, with different records on different lines. (Line separators come out as `\n`)

There are many many scientific datasets which can now be downloaded like this - integrating the download into your data pipeline can help to keep your data flows organised.

Writing our own Parser

We'll need a python library to handle semicolon-separated data like the sunspot data.

You might be thinking: "But I can do that myself!"

```
lines = spots.split("\n")
lines[0:5]
```



```
['1749;01;1749.042; 96.7; -1.0; -1;1',
 '1749;02;1749.123; 104.3; -1.0; -1;1',
 '1749;03;1749.204; 116.7; -1.0; -1;1',
 '1749;04;1749.288; 92.8; -1.0; -1;1',
 '1749;05;1749.371; 141.7; -1.0; -1;1']
```

```
years = [line.split(";")[0] for line in lines]
```

```
years[0:15]
```

But **don't**: what if, for example, one of the records contains a separator inside it; most computers will put the content in quotes, so that, for example,

"Something; something"; something; something

has three fields, the first of which is

Something; something

The naive code above would give four fields, of which the first is

"Something

You'll never manage to get all that right; so you'll be better off using a library to do it.

2.4.4 Writing data to the internet

Note that we're using `requests.get`. `get` is used to receive data from the web. You can also use `post` to fill in a web-form programmatically.

Supplementary material: Learn about using `post` with [requests](#).

Supplementary material: Learn about the different kinds of [http request](#): [Get](#), [Post](#), [Put](#), [Delete](#)...

This can be used for all kinds of things, for example, to programmatically add data to a web resource. It's all well beyond our scope for this course, but it's important to know it's possible, and start to think about the scientific possibilities.

2.5 Data analysis example

Estimated time for this notebook: 20 minutes

We're now going to bring together everything we've learned about Python so far to perform a simple but complete analysis. We will retrieve data, do some computations based on it, and visualise the results.

As we show the code for different parts of the work, we will be touching on various aspects you may want to keep in mind, either related to Python specifically, or to research programming more generally.

2.5.1 Geolocation

```
import geopy # A python library for investigating geographic information.  
https://pypi.org/project/geopy/
```

If you try to follow along on this example in an Jupyter notebook, you might find that you just got an error message.

You'll need to wait until we've covered installation of additional python libraries later in the course, then come back to this and try again. For now, just follow along and try get the feel for how programming for data-focused research works.

```
geocoder = geopy.geocoders.Nominatim(user_agent="rse-course")
geocoder.geocode("Cambridge", exactly_one=False)

[Location(Cambridge, Cambridgeshire, Cambridgeshire and Peterborough, England, United Kingdom, (52.19758464999999, 0.13915373736874398, 0.0)),
 Location(Cambridgeshire, Cambridgeshire and Peterborough, England, United Kingdom, (52.2055314, 0.1186637, 0.0)),
 Location(Cambridge, Middlesex County, Massachusetts, United States, (42.3750997, -71.1056157, 0.0)),
 Location(Cambridge, Region of Waterloo, Southwestern Ontario, Ontario, Canada, (43.36008536, -80.3123023, 0.0)),
 Location(Cambridge, Henry County, Illinois, United States, (41.3025257, -90.1962861, 0.0)),
 Location(Cambridge, Isanti County, Minnesota, 55008, United States, (45.5727408, -93.2243921, 0.0)),
 Location(Cambridge, Story County, Iowa, 50046, United States, (41.8990768, -93.5294029, 0.0)),
 Location(Cambridge, Dorchester County, Maryland, 21613, United States, (38.5714624, -76.0763177, 0.0)),
 Location(Cambridge, Guernsey County, Ohio, 43725, United States, (40.031183, -81.5884561, 0.0)),
 Location(Cambridge, Jefferson County, Kentucky, United States, (38.2217369, -85.616627, 0.0))]
```

Note that the results are a list of `Location` objects, where each `Location` knows its `name`, `latitude` and `longitude`.

Let's define and test a `geolocate` function, storing the result in a variable

```
def geolocate(place):
    return geocoder.geocode(place, exactly_one=False)[0][1]
```

```
london_location = geolocate("London")
print(london_location)
```

(51.5073219, -0.1276474)

2.5.2 Using the Yandex API

The Yandex API allows us to fetch a map of a place, given a longitude and latitude. The URLs look like: https://static-maps.yandex.ru/1.x/?size=400,400&ll=-127.5,51.51&z=10&l=sat&lang=en_US We'll probably end up working out these URLs quite a bit. So we'll make ourselves another function to build up a URL given our parameters.

```

import requests

def request_map_at(lat, long, satellite=True, zoom=12, size=(400, 400)):
    base = "https://static-maps.yandex.ru/1.x/?"
    params = dict(
        z=zoom,
        size=str(size[0]) + "," + str(size[1]),
        ll=str(long) + "," + str(lat),
        l="sat" if satellite else "map",
        lang="en_US",
    )
    return requests.get(base, params=params, timeout=60)

map_response = request_map_at(51.5072, -0.1275)

```

2.5.3 Checking our work

Let's see what URL we ended up with:

```

url = map_response.url
print(url)

https://static-maps.yandex.ru/1.x/?z=12&size=400%2C400&ll=-0.1275%2C51.5072&l=sat&lang=en_US

```

We can write **automated tests** so that if we change our code later, we can check the results are still valid.

```

assert "https://static-maps.yandex.ru/1.x/?" in url
assert "ll=-0.1275%2C51.5072" in url
assert "z=12" in url
assert "size=400%2C400" in url

```

Our previous function comes back with an **object** representing the web request. In object oriented programming, we use the `.` operator to get access to a particular **property** of the object, in this case, the actual image at that URL is in the **content** property. It's a big file, so we'll just show the first few characters here:

```

map_response.content[0:20]

b'\xff\xd8\xff\xe0\x00\x00\x10JFIF\x00\x01\x01\x01\x00H\x00H\x00\x00'

```

2.5.4 Displaying the results

We'll need to do this a lot, so let's wrap up our previous function in another function, to save on typing.

```

def map_at(*args, **kwargs):
    return request_map_at(*args, **kwargs).content

```

We can use a library that comes with Jupyter notebook to display the image. Being able to work with variables which contain images, or documents, or any other weird kind of data, just as easily as we can with numbers or letters, is one of the really powerful things about modern programming languages like Python.

```

from IPython.display import Image
map_png = map_at(*london_location)

print("The type of our map result is actually a: ", type(map_png))

```

```
The type of our map result is actually a: <class 'bytes'>
```

```
Image(map_png)
```



```
Image(map_at("geolocate("New Delhi")))
```



2.5.5 Measuring urbanisation

Now we get to our research project: we want to find out how urbanised the world is. For this we'll use satellite imagery, along a line between two cities. We expect the satellite image to be greener in the countryside.

Let's start by importing the libraries we need.

```
from io import BytesIO # A library to convert between files and strings
import imageio # A library to deal with images, https://pypi.org/project/imageio/
import numpy as np # A library to deal with matrices
```

and then define what we count as green:

```
def is_green(pixels):
    threshold = 1.1
    greener_than_red = pixels[:, :, 1] > threshold * pixels[:, :, 0]
    greener_than_blue = pixels[:, :, 1] > threshold * pixels[:, :, 2]
    green = np.logical_and(greener_than_red, greener_than_blue)
    return green
```

This code has assumed we have our pixel data for the image as a $400 \times 400 \times 3$ 3-d matrix, with each of the three layers being red, green, and blue pixels.

We find out which pixels are green by comparing, element-by-element, the middle (green, number 1) layer to the top (red, zero) and bottom (blue, 2)

Now we just need to read in our data, which is a PNG image, and convert it into our matrix format:

```
def count_green_in_png(data):
    f = BytesIO(data)
    pixels = imageio.v2.imread(f) # Get our PNG image as a numpy array
    return np.sum(is_green(pixels))

print(count_green_in_png(map_at(*london_location)))
```

3258

We'll also need a function to get an evenly spaced set of places between two endpoints:

```
def location_sequence(start, end, steps):
    lats = np.linspace(start[0], end[0], steps) # "Linearly spaced" data
    longs = np.linspace(start[1], end[1], steps)
    return np.vstack([lats, longs]).transpose()

location_sequence(geolocate("London"), geolocate("Cambridge"), 5)
```

```
array([[ 5.15073219e+01, -1.27647400e-01],
       [ 5.16798876e+01, -6.09471157e-02],
       [ 5.18524533e+01,  5.75316868e-03],
       [ 5.20250196e+01,  7.24534536e-02],
       [ 5.21975846e+01,  1.39153737e-01]])
```

2.5.6 Visualising green content

We should display the green content to check our work:

```
def show_green_in_png(data):
    pixels = imageio.imread(BytesIO(data)) # Get our PNG image as rows of pixels
    green = is_green(pixels)

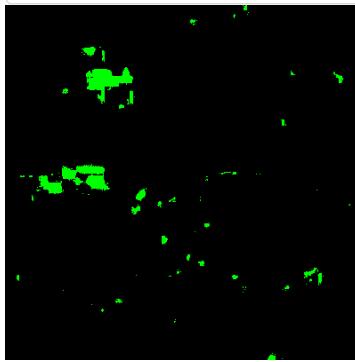
    out = green[:, :, np.newaxis] * np.array([0, 1, 0])[np.newaxis, np.newaxis, :]
    buffer = BytesIO()
    imageio.imwrite(buffer, out, format="png")
    return buffer.getvalue()

Image(map_at(*london_location, satellite=True))
```



```
Image(show_green_in_png(map_at(*london_location, satellite=True)))
```

```
/tmp/ipykernel_6085/3094229650.py:2: DeprecationWarning: Starting with ImageIO v3  
the behavior of this function will switch to that of iio.v3.imread. To keep the  
current behavior (and make this warning disappear) use `import imageio.v2 as  
imageio` or call `imageio.v2.imread` directly.  
    pixels = imageio.imread(BytesIO(data)) # Get our PNG image as rows of pixels  
Lossy conversion from int64 to uint8. Range [0, 1]. Convert image to uint8 prior  
to saving to suppress this warning.
```



2.5.7 Looping

We can loop over each element in our list of coordinates, and get a map for that place:

```
for location in location_sequence(geolocate("London"), geolocate("Birmingham"),  
4):  
    display(Image(map_at(*location)))
```



So now we can count the green from London to Birmingham!

```
[  
    count_green_in_png(map_at("location"))  
    for location in location_sequence(geolocate("London"),  
        geolocate("Birmingham"), 10)  
]
```

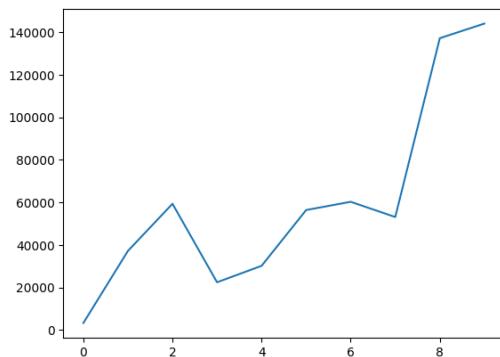
```
[3258, 37141, 59293, 22398, 30123, 56351, 60224, 53067, 137132, 143993]
```

2.5.8 Plotting graphs

Let's plot a graph.

```
import matplotlib.pyplot as plt  
plt.plot(  
    [  
        count_green_in_png(map_at("location"))  
        for location in location_sequence(  
            geolocate("London"), geolocate("Birmingham"), 10)  
    ]  
)
```

```
[<matplotlib.lines.Line2D at 0x7f01e1270910>]
```



From a research perspective, of course, this code needs a lot of work. But I hope the power of using programming is clear.

2.5.9 Composing Program Elements

We built little pieces of useful code, to:

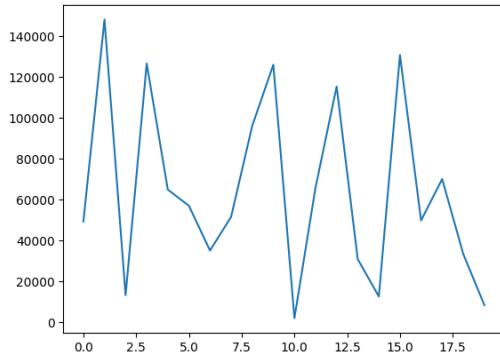
- Find latitude and longitude of a place
- Get a map at a given latitude and longitude
- Decide whether a (red,green,blue) triple is mainly green
- Decide whether each pixel is mainly green
- Plot a new image showing the green places
- Find evenly spaced points between two places

By putting these together, we can make a function which can plot this graph automatically for any two places:

```
def green_between(start, end, steps):
    return [
        count_green_in_png(map_at(*location))
        for location in location_sequence(geolocate(start), geolocate(end), steps)
    ]
```

```
plt.plot(green_between("New York", "Chicago", 20))
```

```
[<matplotlib.lines.Line2D at 0x7f01e12248e0>]
```



And that's it! We've used Python to analyse data from an internet API and visualise it in interesting ways.

2.7 Defining your own classes

Estimated time for this notebook: 20 minutes

2.7.1 User Defined Types

A **class** is a user-programmed Python type.

It is defined like:

```
class Room:
    pass
```

Just as with other python types, you use the name of the type as a function to make a variable of that type:

```
zero = int()
type(zero)
```

```
int
```

```
myroom = Room()
type(myroom)
```

```
__main__.Room
```

In the jargon, we say that an **object** is an **instance** of a particular **class**.

`__main__` is the name of the scope in which top-level code executes, where we've defined the class `Room`.

Once we have an object with a type of our own devising, we can add properties at will:

```
myroom.name = "Living"
```

```
myroom.name
```

```
'Living'
```

The most common use of a class is to allow us to group data into an object in a way that is easier to read and understand than organising data into lists and dictionaries.

```
myroom.capacity = 3  
myroom.occupants = ["James", "Sue"]
```

2.7.2 Methods

So far, our class doesn't do much!

We define functions **inside** the definition of a class, in order to give them capabilities, just like the methods on built-in types.

```
class Room:  
    def overfull(self):  
        return len(self.occupants) > self.capacity
```

```
myroom = Room()  
myroom.capacity = 3  
myroom.occupants = ["James", "Sue"]
```

```
myroom.overfull()
```

```
False
```

```
myroom.occupants.append(["Clare"])
```

```
myroom.occupants.append(["Bob"])
```

```
myroom.overfull()
```

```
True
```

When we write methods, we always write the first function argument as **self**, to refer to the object instance itself, the argument that goes "before the dot".

This is just a convention for this variable name, not a keyword. You could call it something else if you wanted.

2.7.3 Constructors

Normally, though, we don't want to add data to the class attributes on the fly like that. Instead, we define a **constructor** that converts input data into an object.

```
class Room:  
    def __init__(self, name, exits, capacity, occupants=[]):  
        self.name = name  
        self.occupants = occupants # Note the default argument, occupants start  
empty  
        self.exits = exits  
        self.capacity = capacity  
  
    def overfull(self):  
        return len(self.occupants) > self.capacity
```

```
living = Room("Living Room", {"north": "garden"}, 3)
```

```
living.capacity
```

```
3
```

Methods which begin and end with **two underscores** in their names fulfil special capabilities in Python, such as constructors.

2.7.4 Object-oriented design

In building a computer system to model a problem, therefore, we often want to make:

- classes for each *kind of thing* in our system
- methods for each *capability* of that kind
- properties (defined in a constructor) for each *piece of information describing* that kind

For example, the below program might describe our "Maze of Rooms" system:

We define a "Maze" class which can hold rooms:

```

class Maze:
    def __init__(self, name):
        self.name = name
        self.rooms = {}

    def add_room(self, room):
        room.maze = self # The Room needs to know
        # which Maze it is a part of
        self.rooms[room.name] = room

    def occupants(self):
        return [
            occupant
            for room in self.rooms.values()
            for occupant in room.occupants.values()
        ]

    def wander(self):
        """Move all the people in a random direction"""
        for occupant in self.occupants():
            occupant.wander()

    def describe(self):
        for room in self.rooms.values():
            room.describe()

    def step(self):
        self.describe()
        print("")
        self.wander()
        print("")

    def simulate(self, steps):
        for _ in range(steps):
            self.step()

```

And a "Room" class with exits, and people:

```

class Room:
    def __init__(self, name, exits, capacity, maze=None):
        self.maze = maze
        self.name = name
        self.occupants = {} # Note the default argument, occupants start empty
        self.exits = exits # Should be a dictionary from directions to room names
        self.capacity = capacity

    def has_space(self):
        return len(self.occupants) < self.capacity

    def available_exits(self):
        return [
            exit
            for exit, target in self.exits.items()
            if self.maze.rooms[target].has_space()
        ]

    def random_valid_exit(self):
        import random

        if not self.available_exits():
            return None
        return random.choice(self.available_exits())

    def destination(self, exit):
        return self.maze.rooms[self.exits[exit]]

    def add_occupant(self, occupant):
        occupant.room = self # The person needs to know which room it is in
        self.occupants[occupant.name] = occupant

    def delete_occupant(self, occupant):
        del self.occupants[occupant.name]

    def describe(self):
        if self.occupants:
            print(f'{self.name}: {", ".join(self.occupants.keys())}')

```

We define a "Person" class for room occupants:

```

class Person:
    def __init__(self, name, room=None):
        self.name = name

    def use(self, exit):
        self.room.delete_occupant(self)
        destination = self.room.destination(exit)
        destination.add_occupant(self)
        print(f'{self.name} goes {exit} to the {destination.name}')

    def wander(self):
        exit = self.room.random_valid_exit()
        if exit:
            self.use(exit)

```

And we use these classes to define our people, rooms, and their relationships:

```

james = Person("James")
sue = Person("Sue")
bob = Person("Bob")
clare = Person("Clare")

living = Room(
    "livingroom", {"outside": "garden", "upstairs": "bedroom", "north": "kitchen"}, 2
)
kitchen = Room("kitchen", {"south": "livingroom"}, 1)
garden = Room("garden", {"inside": "livingroom"}, 3)
bedroom = Room("bedroom", {"jump": "garden", "downstairs": "livingroom"}, 1)

```

```
house = Maze("My House")
```

```
for room in [living, kitchen, garden, bedroom]:
    house.add_room(room)
```

```
living.add_occupant(james)
```

```
garden.add_occupant(sue)
garden.add_occupant(clare)
```

```
bedroom.add_occupant(bob)
```

And we can run a "simulation" of our model:

```
house.simulate(3)
```

```
livingroom: James
garden: Sue Clare
bedroom: Bob

James goes outside to the garden
Sue goes inside to the livingroom
Clare goes inside to the livingroom
Bob goes jump to the garden

livingroom: Sue Clare
garden: James Bob

Sue goes north to the kitchen
Clare goes outside to the garden
James goes inside to the livingroom
Bob goes inside to the livingroom

livingroom: James Bob
kitchen: Sue
garden: Clare

James goes outside to the garden
Bob goes upstairs to the bedroom
Sue goes south to the livingroom
Clare goes inside to the livingroom
```

2.7.5 Alternative object models

There are many choices for how to design programs to do this. Another choice would be to separately define exits as a different class from rooms. This way, we can use arrays instead of dictionaries, but we have to first define all our rooms, then define all our exits.

```
class Maze:
    def __init__(self, name):
        self.name = name
        self.rooms = []
        self.occupants = []

    def add_room(self, name, capacity):
        result = Room(name, capacity)
        self.rooms.append(result)
        return result

    def add_exit(self, name, source, target, reverse=None):
        source.add_exit(name, target)
        if reverse:
            target.add_exit(reverse, source)

    def add_occupant(self, name, room):
        self.occupants.append(Person(name, room))
        room.occupancy += 1

    def wander(self):
        "Move all the people in a random direction"
        for occupant in self.occupants:
            occupant.wander()

    def describe(self):
        for occupant in self.occupants:
            occupant.describe()

    def step(self):
        self.describe()
        print("")
        self.wander()
        print("")

    def simulate(self, steps):
        for _ in range(steps):
            self.step()
```

```
class Room:
    def __init__(self, name, capacity):
        self.name = name
        self.capacity = capacity
        self.occupancy = 0
        self.exits = []

    def has_space(self):
        return self.occupancy < self.capacity

    def available_exits(self):
        return [exit for exit in self.exits if exit.valid()]

    def random_valid_exit(self):
        import random

        if not self.available_exits():
            return None
        return random.choice(self.available_exits())

    def add_exit(self, name, target):
        self.exits.append(Exit(name, target))
```

```
class Person:
    def __init__(self, name, room=None):
        self.name = name
        self.room = room

    def use(self, exit):
        self.room.occupancy -= 1
        destination = exit.target
        destination.occupancy += 1
        self.room = destination
        print(f"{self.name} goes {exit.name} to the {destination.name}")

    def wander(self):
        exit = self.room.random_valid_exit()
        if exit:
            self.use(exit)

    def describe(self):
        print(f"{self.name} is in the {self.room.name}")
```

```

class Exit:
    def __init__(self, name, target):
        self.name = name
        self.target = target

    def valid(self):
        return self.target.has_space()

house = Maze("My New House")

living = house.add_room("livingroom", 2)
bed = house.add_room("bedroom", 1)
garden = house.add_room("garden", 3)
kitchen = house.add_room("kitchen", 1)

house.add_exit("north", living, kitchen, "south")
house.add_exit("upstairs", living, bed, "downstairs")
house.add_exit("outside", living, garden, "inside")
house.add_exit("jump", bed, garden)

house.add_occupant("James", living)
house.add_occupant("Sue", garden)
house.add_occupant("Bob", bed)
house.add_occupant("Clare", garden)

house.simulate(3)

```

James is in the livingroom
 Sue is in the garden
 Bob is in the bedroom
 Clare is in the garden
 James goes north to the kitchen
 Sue goes inside to the livingroom
 Bob goes downstairs to the livingroom
 James is in the kitchen
 Sue is in the livingroom
 Bob is in the livingroom
 Clare is in the garden
 Sue goes upstairs to the bedroom
 Bob goes outside to the garden
 Clare goes inside to the livingroom
 James is in the kitchen
 Sue is in the bedroom
 Bob is in the garden
 Clare is in the livingroom
 James goes south to the livingroom
 Sue goes jump to the garden
 Clare goes north to the kitchen

This is a huge topic, about which many books have been written. The differences between these two designs are important, and will have long-term consequences for the project. That is the how we start to think about **software engineering**, as opposed to learning to program, and is an important part of this course.

2.7 Data analysis with classes

Estimated time to complete this notebook: 10 minutes

Earlier, we wrote some code to measure the amount of green content on satellite images. Now, we're going to convert this into a "Greengraph" class, and save it as a module.

⚠ It is generally a better idea to create files in an editor or integrated development environment (IDE) rather than through the notebook! ⚠

2.7.1 Classes for Greengraph

```

%%bash
mkdir -p greengraph # Create the folder for the module (on mac or linux)

%%writefile greengraph/graph.py
import numpy as np
import geopy
from .map import Map

class Greengraph:
    def __init__(self, start, end):
        self.start = start
        self.end = end
        self.geocoder = geopy.geocoders.Nominatim(user_agent="rsd-course")

    def geolocate(self, place):
        return self.geocoder.geocode(place, exactly_one=False)[0][1]

    def location_sequence(self, start, end, steps):
        lats = np.linspace(start[0], end[0], steps)
        longs = np.linspace(start[1], end[1], steps)
        return np.vstack([lats, longs]).transpose()

    def green_between(self, steps):
        return [
            Map(*location).count_green()
            for location in self.location_sequence(
                self.geolocate(self.start), self.geolocate(self.end), steps
            )
        ]

```

Overwriting greengraph/graph.py

```
%>writefile greengraph/map.py
import numpy as np
from io import BytesIO
import imageio as img
import requests

class Map:
    def __init__(self, lat, long, satellite=True, zoom=10, size=(400, 400), sensor=False):
        self.lat = lat
        self.long = long
        self.satellite = satellite
        self.zoom = zoom
        self.size = size
        self.sensor = sensor
        self.base = "https://static-maps.yandex.ru/1.x/?"
        self.params = dict(
            z=zoom,
            size=str(size[0]) + "x" + str(size[1]),
            ll=long + "," + lat,
            l="sat" if satellite else "map",
            lang="en_US",
        )
        self.image = requests.get(self.base, params=self.params).content # Fetch our PNG image data
        content = BytesIO(self.image)
        self.pixels = img.imread(content) # Parse our PNG image as a numpy array

    def green(self, threshold):
        # Use NumPy to build an element-by-element logical array
        greener_than_red = self.pixels[:, :, 1] > threshold * self.pixels[:, :, 0]
        greener_than_blue = self.pixels[:, :, 1] > threshold * self.pixels[:, :, 2]
        green = np.logical_and(greener_than_red, greener_than_blue)
        return green

    def count_green(self, threshold=1.1):
        return np.sum(self.green(threshold))

    def show_green(data, threshold=1.1):
        green = self.green(threshold)
        out = green[:, :, np.newaxis] * array([0, 1, 0])[np.newaxis, np.newaxis, :]
        buffer = BytesIO()
        result = img.imwrite(buffer, out, format="png")
        return buffer.getvalue()
```

Overwriting greengraph/map.py

```
%>writefile greengraph/__init__.py
from .graph import Greengraph
```

Overwriting greengraph/__init__.py

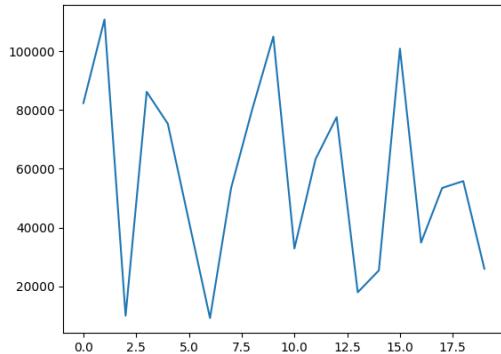
2.7.2 Invoking our code and making a plot

```
%matplotlib inline
from greengraph import Greengraph
from matplotlib import pyplot as plt

mygraph = Greengraph("New York", "Chicago")
data = mygraph.green_between(20)
```

plt.plot(data)

[<matplotlib.lines.Line2D at 0x7f749c27c190>]



2.7 Classroom Exercises

List of exercises and estimated completion times

[2a - Occupancy Dictionary](#) 5 minutes

[2b - Occupancy Dictionary Extension](#) 5 minutes

[2c - Functions](#) 15 minutes

[2d - Using Libraries](#) 15 minutes

[2e - Longitude and Latitude](#) 15 minutes

[2f - Defining Classes](#) 45 minutes

[2g - Longitude and Latitude Extension](#) 10 minutes

Exercise 2a Occupancy Dictionary

Relevant Sections: 2.0.2

In one of the module 1 exercises you designed a data structure to represent a maze using dictionaries and lists.

The answer to your initial maze model output might have looked similar to this:

```
house = {
    "living": {
        "exits": {"north": "kitchen", "outside": "garden", "upstairs": "bedroom"},
        "people": ["James"],
        "capacity": 2,
    },
    "kitchen": {"exits": {"south": "living", "inside": "garden"}, "people": [], "capacity": 1},
    "garden": {"exits": {"inside": "living"}, "people": ["Sue"], "capacity": 3},
    "bedroom": {
        "exits": {"downstairs": "living", "jump": "garden"},
        "people": [],
        "capacity": 1,
    },
}
```

Take this maze data structure.

First write an expression to print out a new dictionary, which holds, for each room, that room's capacity.

The output should look like:

```
{"bedroom": 1, "garden": 3, "kitchen": 1, "living": 2}
```

Exercise 2b Occupancy Dictionary Extension

Relevant Sections: 2.0.2 and 2.0.4

Now, write a program to print out a new dictionary, which gives, for each room's name, the number of people in it. Don't add in a zero value in the dictionary for empty rooms.

The output should look similar to:

```
{"garden": 1, "living": 1}
```

Exercise 2c Functions

Relevant Sections: 2.1.1, 2.1.8, (2.0.2)

Write a function that will take the following input and return a list containing only even integers

```
(1, 1.9999999999, "three", 20/5, 5, 6, "sju", "8", 9, 10., 11, 12)
```

The call to your function could look something like this:

```
my_function(1, 1.9999999999, "three", 20/5, 5, 6, "sju", "8", 9, 10., 11, 12)
```

or

```
my_function(*inputs)
```

Exercise 2d Using Libraries

Relevant Sections: 2.2.1

Investigate the similarities and differences between the responses (if any) from the `numpy`, `scipy`, `statistics`, and `math` modules to the following calculations:

π

$\log_{10}(n)$ where n is positive

$\log_{10}(n)$ where n is negative

The mean of the numbers 1 to 9 (inclusive)

For those interested, each of these libraries has their own documentation. [NumPy](#), [SciPy](#), [statistics](#), and [math](#)

Exercise 2e Longitude and Latitude

Relevant Sections: 2.4.2, 2.4.1

In section 2.4.2 a map of an area collected from the internet was displayed.

Write a function that will accept user-specified latitude, and longitude and return the response. Then use `IPython` to display the image as in 2.5.2

The answer could look something like:

```
function_response = my_function(lat, lon)
Image(function_response)
```

some interesting coordinates are:

```
coordinates_as_lat_lon = [
    (36.2110, -115.2669),
    (53.0066, 7.1920),
    (41.3998, 2.1631),
    (40.7822, -73.9653),
    (25.8380, 50.6050),
]
```

Exercise 2f Defining Classes

Relevant Sections: 2.6.1, 2.6.2, 2.6.3, 2.6.4, 2.6.5

In section 2.6.4 and 2.6.5 two examples of the maze model were given.

Compare the two solutions. Discuss with a partner which you like better, and why.

Then, starting from scratch, design your own. What choices did you make that are different?

Exercise 2g Longitude and Latitude Extension

Relevant Sections: 2.3.7

Use the function you wrote in 2e above as the basis for a new function that will receive the longitude, latitude, **zoom level** and a name to save the file as. Use this function to save a map image file somewhere on your local disk.

Zoom between 14 and 16 work well for the example coordinates

3. Research Data in Python

- Fields and records
- Structured data: JSON and YAML
- Numpy: Efficient vector and matrix operations
- Matplotlib: Plotting and animations

Contents

- [3.0 Scientific Python](#) (5 minutes)
- [3.1 Field and Record Data](#) (20 minutes)
- [3.2 Structured Data](#) (15 minutes)
- [3.3 Plotting with Matplotlib](#) (25 minutes)
- [3.4 NumPy](#) (20 minutes)
- [3.5 Advanced NumPy](#) (20 minutes)
- [3.6 The Boids](#) (45 minutes)

Total time: 2 hrs 30 minutes

Exercises

Classroom exercises are grouped together at the end of the module: [3.7 Classroom Exercises](#). Each exercise is labelled with any sections whose contents are relevant. We recommend that instructors schedule the exercises to be done in groups during breaks in the taught content. However, it is **important** that participants also have some time away from their screens. Exercises can also be left as self-paced homework assignments if preferred.

3.0 Scientific Python

Estimated time to complete this notebook: 5 minutes

Why is Python so popular for research work?

Historically, FORTRAN was the most popular "language of technical computing". Later, MATLAB was created with strong built-in support for efficient numerical analysis with matrices (the *mat* in MATLAB is for Matrix, not Maths), and plotting.

Early Python users developed three critical libraries, to match the power of MATLAB for scientific work:

- Matplotlib, the plotting library created by [John D. Hunter](#)
- NumPy, a fast matrix maths library created by [Travis Oliphant](#)
- IPython, the precursor of the notebook, created by [Fernando Perez](#)

By combining a plotting library, a matrix maths library, and an easy-to-use interface allowing live plotting commands in a persistent environment, the powerful capabilities of MATLAB were matched by a free and open toolchain.

Further tools such as [pandas](#) and [scipy](#) are built on, extend, or utilise these libraries. In this module we will use these libraries to deal with data of the type that might be used in a research project.

3.1 Field and Record Data

Estimated time to complete this notebook: 20 minutes

3.1.1 Separated Value Files

Let's go back to the sunspots example [from the previous module](#). We had downloaded some semicolon separated data and decided it was better to use a library than to write our own parser.

```
import requests
spots = requests.get("http://www.sidc.be/silso/INFO/snmtocsv.php", timeout=60)
spots.text.split("\n")[0]
'1749;01;1749.042; -96.7; -1.0; -1;1'
```

We want to work programmatically with *Separated Value* files.

These are files which have:

- Each record on a line
- Each record has multiple *fields*
- Fields are separated by some *separator*

Typical separators are the [space](#), [tab](#), [comma](#), and [semicolon](#) separated values files, e.g.:

- Space separated value (e.g. `field1 "field two" field3`)
- Comma separated value (e.g. `field1, another field, "wow, another field"`)

Comma-separated-value is abbreviated CSV, and tab separated value TSV.

CSV is also used to refer to all the different sub-kinds of separated value files, i.e. some people use CSV to refer to tab, space and semicolon separated files.

CSV is not a particularly great data format, because it forces your data model to be a list of lists. Richer file formats describe "serialisations" for dictionaries and for deeper-than-two nested list structures as well.

Nevertheless, CSV files are very popular because you can always export *spreadsheets* as CSV files, (each cell is a field, each row is a record)

3.1.2 CSV variants

Some CSV formats define a comment character, so that rows beginning with, e.g., a #, are not treated as data, but give a human comment.

Some CSV formats define a three-deep list structure, where a double-newline separates records into blocks.

Some CSV formats assume that the first line defines the names of the fields, e.g.:

```
name, age
James, 39
Will, 2
```

3.1.3 Python CSV readers

The Python standard library has a `csv` module. However, it's less powerful than the CSV capabilities in other libraries such as `numpy`. Here we will use `pandas` which is built on top of `numpy`.

```
import pandas as pd

df = pd.read_csv("http://www.sidc.be/silso/INFO/snmtofcsv.php", sep=";", header=None)
df.head()
```

0	1	2	3	4	5	6
0	1749	1	1749.042	96.7	-1.0	-1
1	1749	2	1749.123	104.3	-1.0	-1
2	1749	3	1749.204	116.7	-1.0	-1
3	1749	4	1749.288	92.8	-1.0	-1
4	1749	5	1749.371	141.7	-1.0	-1

Pandas `read_csv` is a powerful CSV reader tool. A path to the data is given, this can be something on a local machine, or in this case the path is a url.

We used the `sep` optional argument to specify the delimiter. The optional argument `header` specifies if the data contains headers, and if so; the row numbers to use as column names.

The data is loaded into a DataFrame. The `head` method shows us the first 5 entries in the dataframe. The `tail` method shows us the last 5 entries.

```
df.tail()
```

0	1	2	3	4	5	6
3281	2022	6	2022.453	70.3	13.2	1403
3282	2022	7	2022.538	91.4	12.2	1304
3283	2022	8	2022.623	75.4	10.5	1289
3284	2022	9	2022.705	96.3	16.2	1130
3285	2022	10	2022.790	95.4	15.5	1028

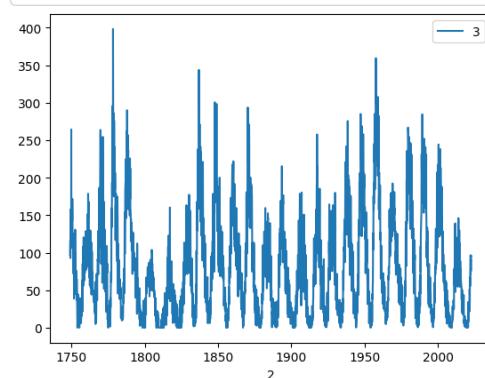
```
df[3][0]
```

```
96.7
```

We can now plot the "Sunspot cycle":

```
df.plot(x=2, y=3)
```

```
<AxesSubplot: xlabel='2'>
```



The plot command accepted an series of 'X' values and an series of 'Y' values, identified by their column number in this case, as the dataframe does not have (useful) column headers yet.

3.1.4 Naming Columns

As it happens, the columns definitions can be found on the source website (<http://www.sidc.be/silso/infosnmot>)

CSV

Filename: SN_m_tot_V2.0.csv Format: Comma Separated values (adapted for import in spreadsheets) The separator is the semicolon ";".

Contents:

- Column 1-2: Gregorian calendar date
 - Year
 - Month
- Column 3: Date in fraction of year.
- Column 4: Monthly mean total sunspot number.
- Column 5: Monthly mean standard deviation of the input sunspot numbers.
- Column 6: Number of observations used to compute the monthly mean total sunspot number.
- Column 7: Definitive/provisional marker. '1' indicates that the value is definitive. '0' indicates that the value is still provisional.

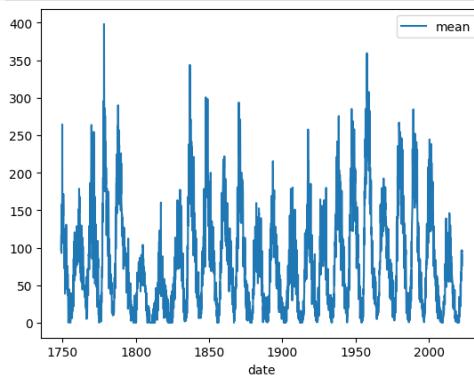
We can actually specify this to the formatter:

```
df_w_names = pd.read_csv(  
    "http://www.sidc.be/silso/INFO/snmtotcsv.php",  
    sep=";",  
    header=None,  
    names=["year", "month", "date", "mean", "deviation", "observations",  
    "definitive"],  
)  
df_w_names.head()
```

	year	month	date	mean	deviation	observations	definitive
0	1749	1	1749.042	96.7	-1.0	-1	1
1	1749	2	1749.123	104.3	-1.0	-1	1
2	1749	3	1749.204	116.7	-1.0	-1	1
3	1749	4	1749.288	92.8	-1.0	-1	1
4	1749	5	1749.371	141.7	-1.0	-1	1

```
df_w_names.plot(x="date", y="mean")
```

```
<AxesSubplot: xlabel='date'>
```



Note: The plot method used for the `DataFrame` is just a wrapper around the `matplotlib` function `plt.plot()`:

3.1.5 Typed Fields

It's also often useful to check, and if necessary specify, the datatype of each field.

```
df_w_names.dtypes # Check the data types of all columns in the DataFrame
```

```
year      int64  
month     int64  
date      float64  
mean      float64  
deviation float64  
observations  int64  
definitive int64  
dtype: object
```

In this case the data types seem sensible, however if we wanted to convert the year into a floating point number instead, we could via:

```
df_w_names["year"] = df_w_names["year"].astype("float64")  
df_w_names.dtypes
```

```
year      float64  
month     int64  
date      float64  
mean      float64  
deviation float64  
observations  int64  
definitive int64  
dtype: object
```

```
df_w_names.head()
```

	year	month	date	mean	deviation	observations	definitive
0	1749.0	1	1749.042	96.7	-1.0	-1	1
1	1749.0	2	1749.123	104.3	-1.0	-1	1
2	1749.0	3	1749.204	116.7	-1.0	-1	1
3	1749.0	4	1749.288	92.8	-1.0	-1	1
4	1749.0	5	1749.371	141.7	-1.0	-1	1

3.1.6 Filtering data

Sometimes it is necessary to filter data, for example to only see the sunspots for the year 2018 you would use:

```
df_twenty_eighteen = df_w_names[(df_w_names["year"] == 2018)]
df_twenty_eighteen.head(20)
```

	year	month	date	mean	deviation	observations	definitive
3228	2018.0	1	2018.042	6.8	1.5	701	1
3229	2018.0	2	2018.122	10.7	1.1	917	1
3230	2018.0	3	2018.204	2.5	0.4	1081	1
3231	2018.0	4	2018.286	8.9	1.3	996	1
3232	2018.0	5	2018.371	13.1	1.6	1234	1
3233	2018.0	6	2018.453	15.6	1.6	1070	1
3234	2018.0	7	2018.538	1.6	0.6	1438	1
3235	2018.0	8	2018.623	8.7	1.0	1297	1
3236	2018.0	9	2018.705	3.3	0.6	1223	1
3237	2018.0	10	2018.790	4.9	1.2	1097	1
3238	2018.0	11	2018.873	4.9	0.6	771	1
3239	2018.0	12	2018.958	3.1	0.5	786	1

Even though we used

```
df_twenty_eighteen.head(20)
```

to show us the first 20 results from the dataframe, only 12 are shown as there are only 12 months in a year

If we wanted all data from 1997 to 1999 we could via:

```
df_nineties = df_w_names[(df_w_names["year"] >= 1997) & (df_w_names["year"] <
2000)]
```

```
df_nineties.head()
```

	year	month	date	mean	deviation	observations	definitive
2976	1997.0	1	1997.042	7.4	3.2	497	1
2977	1997.0	2	1997.123	11.0	2.9	545	1
2978	1997.0	3	1997.204	12.1	2.4	627	1
2979	1997.0	4	1997.288	23.0	3.3	663	1
2980	1997.0	5	1997.371	25.4	2.8	716	1

```
df_nineties.tail()
```

	year	month	date	mean	deviation	observations	definitive
3007	1999.0	8	1999.623	142.3	12.9	649	1
3008	1999.0	9	1999.707	106.3	6.5	624	1
3009	1999.0	10	1999.790	168.7	10.4	531	1
3010	1999.0	11	1999.874	188.3	12.3	406	1
3011	1999.0	12	1999.958	116.8	9.3	404	1

3.2 Structured Data

Estimated time to complete this notebook: 15 minutes

3.2.1 Structured data

CSV files can only model data where each record has several fields, and each field is a simple datatype, a string or number.

We often want to store data which is more complicated than this, with nested structures of lists and dictionaries. Structured data formats like JSON, YAML, and XML are designed for this.

3.2.2 JSON

[JSON](#) is a very common open-standard data format that is used to store structured data in a human-readable way.

This allows us to represent data which is combinations of lists and dictionaries as a text file which looks a bit like a Javascript (or Python) data literal.

```
import json
```

Any nested group of dictionaries and lists can be saved:

Saving and loading data is really easy.

To save a dictionary as a json file:

```

example_dictionary = {"somekey": ["a list", "with values", "for json"]}
with open("myfile.json", "w") as f:
    json.dump(example_dictionary, f)

```

And read in the data back in from the file

```

with open("myfile.json", "r") as f:
    my_json_data = json.load(f)

my_json_data

{'somekey': ['a list', 'with values', 'for json']}

my_json_data["somekey"]

['a list', 'with values', 'for json']

```

This is a very nice solution for loading and saving Python data structures.

It's a very common way of transferring data on the internet, and of saving datasets to disk.

There's good support in most languages, so it's a nice inter-language file interchange format.

3.2.3 YAML

[YAML](#) is a very similar data format to JSON, with some nice additions:

- You don't need to quote strings if they don't have funny characters in
- You can have comment lines, beginning with a #
- You can write dictionaries without the curly brackets: it just notices the colons.
- You can write lists like this:

```

%%writefile myfile.yaml
somekey:
  - a list # Look, this is a list
  - with values
  - for yaml

Overwriting myfile.yaml

import yaml # This may need installed as pyyaml

with open("myfile.yaml") as myfile:
    my_yaml_data = yaml.safe_load(myfile)
    print(my_yaml_data)

{'somekey': ['a list', 'with values', 'for yaml']}

```

Supplementary Materials: `yaml.safe_load` is preferred over `yaml.load` to avoid executing arbitrary code in untrusted files. See [here](#) for details.

YAML is a popular format for ad-hoc data files, but the library doesn't ship with default Python (though it is part of Anaconda and Canopy), so some people still prefer JSON for its universality.

Because YAML gives the **option** of serialising a list either as newlines with dashes, or with square brackets, you can control this choice:

```

print(yaml.safe_dump(my_yaml_data, default_flow_style=True))

{somekey: [a list, with values, for yaml]}

print(yaml.safe_dump(my_yaml_data, default_flow_style=False))

somekey:
  - a list
  - with values
  - for yaml

```

`default_flow_style=False` uses a "block style" (rather than an "inline" or "flow style") to delineate data structures. [See the YAML docs for more details.](#)

In addition to saving a yaml file via cell magics, they can also be written:

```

with open("myotherfile.yaml", "w") as f:
    yaml.safe_dump(my_yaml_data, f, default_flow_style=False)

```

3.2.4 JSON to YAML

And of course the JSON formatted data can be written as a yaml file, and vice versa. Here we are taking the data we read in for the JSON example and saving it as a yaml file.

```

with open("json_to_yaml.yaml", "w") as f:
    yaml.safe_dump(my_json_data, f, default_flow_style=False)

```

You can compare the original json file to the json-data-saved-as-yaml either when loaded....

```

# The original json file
with open("myfile.json", "r") as f:
    mydataasstring = f.read()
    print(json.loads(mydataasstring))

{'somekey': ['a list', 'with values', 'for json']}

```

```
# The data from the json file saved as a yaml then read in
with open("json_to_yaml.yaml") as f:
    my_json_yaml_data = yaml.safe_load(f)
```

```
print(my_json_yaml_data)
```

To how they appear in their respective file formats

```
%bash
#%cmd (windows)
cat 'myfile.json' # The original json file
```

```
{"somekey": ["a list", "with values", "for json"]}
```

```
%bash
#%cmd (windows)
cat 'json_to_yaml.yaml' # The data from the json file saved as a yaml
```

```
somekey:
```

```
- a list
```

```
- with values
```

```
- for json
```

3.2.5 XML

Supplementary material: [XML](#) is another popular choice when saving nested data structures. It's very careful, but verbose. If your field uses XML data, you'll need to learn a [python XML parser](#) (there are a few), and about how XML works.

3.3 Plotting with Matplotlib

Estimated time to complete this notebook: 25 minutes

3.3.1 Importing Matplotlib

We import the `pypplot` object from Matplotlib, which provides us with an interface for making figures. We usually abbreviate it.

```
from matplotlib import pypplot as plt
```

3.3.2 Notebook magics

When we write:

```
%matplotlib inline
```

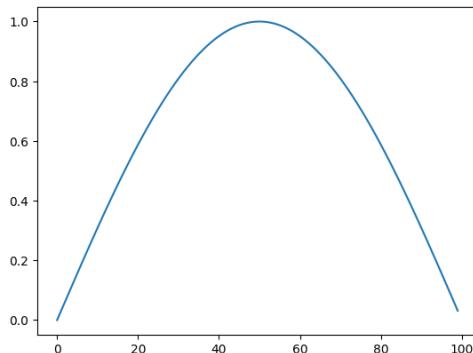
We tell the Jupyter notebook to show figures we generate alongside the code that created it, rather than in a separate window. Lines beginning with a single percent are not python code: they control how the notebook deals with python code.

Lines beginning with two percent signs are "cell magics", that tell Jupyter notebook how to interpret the particular cell; we've seen `%%writefile` and `%%bash` for example.

3.3.3 A basic plot

When we write:

```
from math import cos, pi, sin
myfig = plt.plot([sin(pi * x / 100.0) for x in range(100)])
```

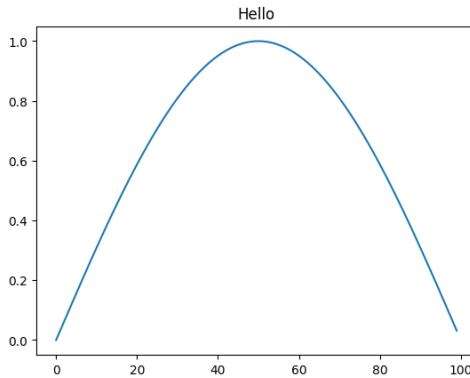


The `plot` command *returns* a figure, just like the return value of any function. The notebook then displays this.

To add a title, axis labels etc, we need to get that figure object, and manipulate it. For convenience, matplotlib allows us to do this just by issuing commands to change the "current figure".

```
plt.plot([sin(pi * x / 100.0) for x in range(100)])
plt.title("Hello")
```

```
Text(0.5, 1.0, 'Hello')
```



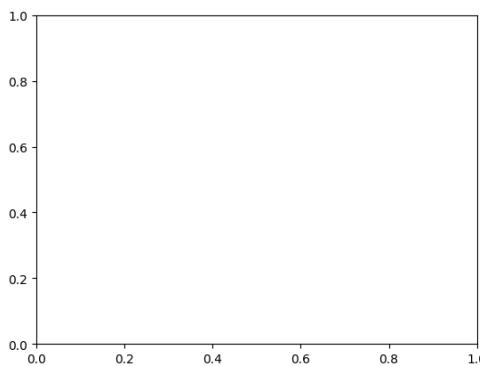
But this requires us to keep all our commands together in a single cell, and makes use of a "global" single "current plot", which, while convenient for quick exploratory sketches, is a bit cumbersome. If we want to produce publication-quality plots from our notebook, `matplotlib`, defines some types we can use to treat individual figures as variables, and manipulate these.

3.3.4 Figures and Axes

We often want multiple graphs in a single figure (e.g. for figures which display a matrix of graphs of different variables for comparison).

So Matplotlib divides a `figure` object up into axes: each pair of axes is one 'subplot'. To make a boring figure with just one pair of axes, however, we can just ask for a default new figure, with brand new axes. The relevant function returns a (figure, axis) pair, which we can deal out with parallel assignment.

```
sine_graph, sine_graph_axes = plt.subplots()
```



Once we have some axes, we can plot a graph on them:

```
sine_graph_axes.plot([sin(pi * x / 100.0) for x in range(100)], label="sin(x)")
```

```
[<matplotlib.lines.Line2D at 0x7f187ba3a3a0>]
```

We can add a title to a pair of axes:

```
sine_graph_axes.set_title("My graph")
```

```
Text(0.5, 1.0, 'My graph')
```

```
sine_graph_axes.set_ylabel("f(x)")
```

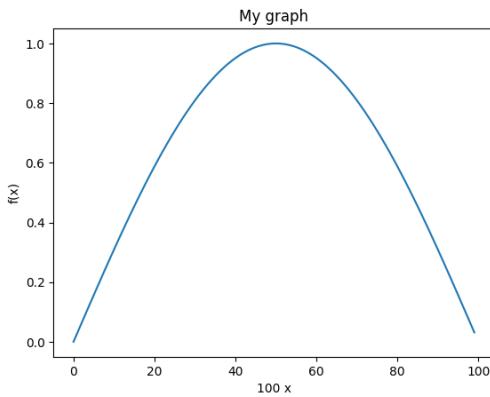
```
Text(4.44444444444445, 0.5, 'f(x)')
```

```
sine_graph_axes.set_xlabel("100 x")
```

```
Text(0.5, 4.44444444444445, '100 x')
```

Now we need to actually display the figure. As always with the notebook, if we make a variable be returned by the last line of a code cell, it gets displayed:

```
sine_graph
```

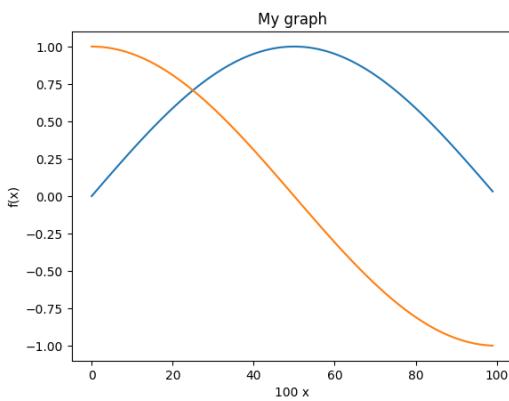


We can add another curve:

```
sine_graph_axes.plot([cos(pi * x / 100.0) for x in range(100)], label="cos(x)")
```

```
[<matplotlib.lines.Line2D at 0x7f187b9e4ca0>]
```

```
sine_graph
```

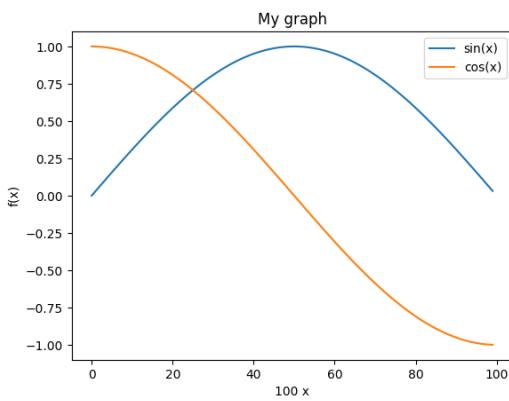


A legend will help us distinguish the curves:

```
sine_graph_axes.legend()
```

```
[<matplotlib.legend.Legend at 0x7f187b9eb670>]
```

```
sine_graph
```



3.3.5 Saving figures

We must be able to save figures to disk, in order to use them in papers. This is really easy:

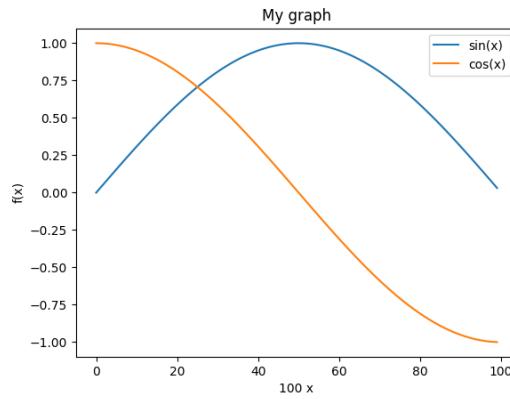
```
sine_graph.savefig("my_graph.png")
```

In order to be able to check that it worked, we need to know how to display an arbitrary image in the notebook.

The programmatic way is like this:

```
# Use the notebook's own library for manipulating itself.
from IPython.display import Image

Image(filename="my_graph.png")
```



3.3.6 Subplots

We might have wanted the sin and cos graphs on separate axes:

```

double_graph = plt.figure()

<Figure size 640x480 with 0 Axes>

sin_axes = double_graph.add_subplot(2, 1, 1) # 2 rows, 1 column, 1st subplot
cos_axes = double_graph.add_subplot(2, 1, 2)

double_graph

```

```

sin_axes.plot([sin(pi * x / 100.0) for x in range(100)])
[<matplotlib.lines.Line2D at 0x7f187b8e2160>]

sin_axes.set_ylabel("sin(x)")
Text(4.444444444444445, 0.5, 'sin(x)')

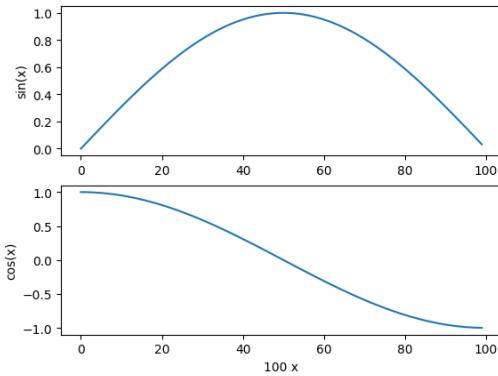
cos_axes.plot([cos(pi * x / 100.0) for x in range(100)])
[<matplotlib.lines.Line2D at 0x7f187b8e2bb0>]

cos_axes.set_ylabel("cos(x)")
Text(4.444444444444445, 0.5, 'cos(x)')

cos_axes.set_xlabel("100 x")
Text(0.5, 4.444444444444445, '100 x')

double_graph

```

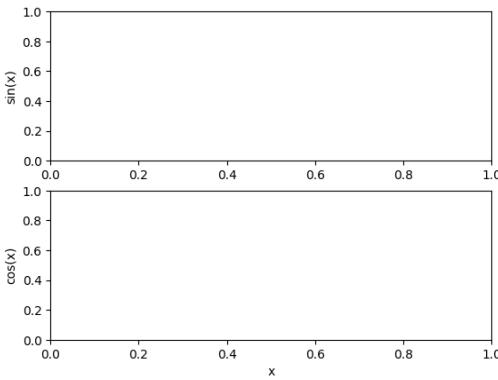


3.3.7 Versus plots

When we specify a single `list` to `plot`, the x-values are just the array index number. We usually want to plot something more meaningful:

```
double_graph = plt.figure()
sin_axes = double_graph.add_subplot(2, 1, 1)
cos_axes = double_graph.add_subplot(2, 1, 2)
cos_axes.set_ylabel("cos(x)")
sin_axes.set_ylabel("sin(x)")
cos_axes.set_xlabel("x")
```

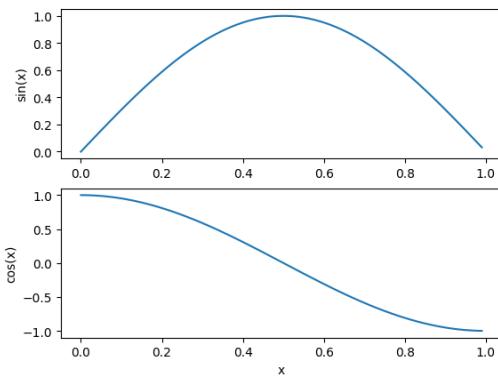
```
Text(0.5, 0, 'x')
```



```
sin_axes.plot(
    [x / 100.0 for x in range(100)], [sin(pi * x / 100.0) for x in range(100)])
cos_axes.plot(
    [x / 100.0 for x in range(100)], [cos(pi * x / 100.0) for x in range(100)])
```

```
[<matplotlib.lines.Line2D at 0x7f187b8e2fa0>]
```

```
double_graph
```



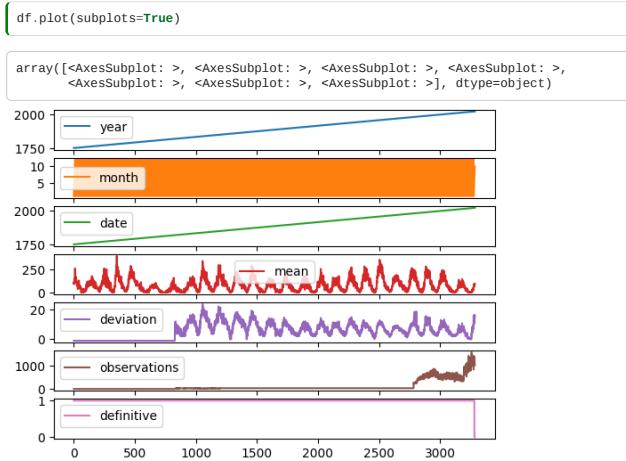
3.3.8 Sunspot Data

We can incorporate what we have learned in the sunspots example to produce graphs of the data.

```
import pandas as pd
df = pd.read_csv(
    "http://www.sidc.be/silso/INFO/snmtotcsv.php",
    sep=";",
    header=None,
    names=["year", "month", "date", "mean", "deviation", "observations",
    "definitive"],
)
df.head()
```

	year	month	date	mean	deviation	observations	definitive
0	1749	1	1749.042	96.7	-1.0	-1	1
1	1749	2	1749.123	104.3	-1.0	-1	1
2	1749	3	1749.204	116.7	-1.0	-1	1
3	1749	4	1749.288	92.8	-1.0	-1	1
4	1749	5	1749.371	141.7	-1.0	-1	1

We can plot all the data in the dataframe separately, but that isn't always useful!



Let's produce some more meaningful and useful visualisations by accessing the dataframe directly.

We start by discarding any rows with an invalid (negative) standard deviation.

```
df = df[df["deviation"] > 0]
```

Next we use the dataframe to construct some useful lists.

```
deviation = df["deviation"].tolist() # Get the dataframe column (series) as a list
observations = df["observations"].tolist()
mean = df["mean"].tolist()
date = df["date"].tolist()

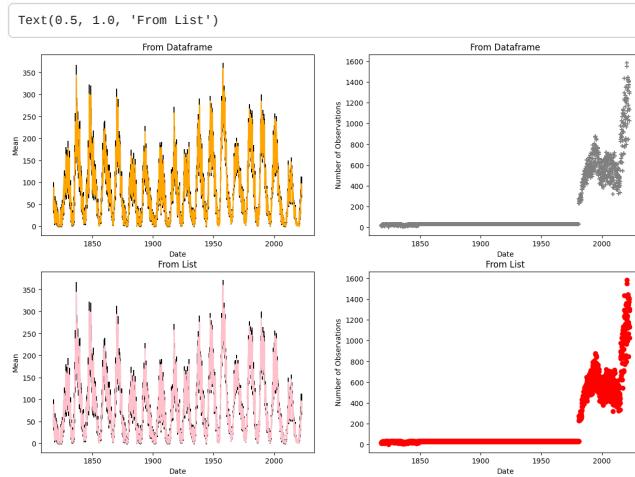
fig = plt.figure(
    figsize=(15, 10)
) # Set the width of the figure to be 15 inches, and the height to be 5 inches

ax1 = fig.add_subplot(2, 2, 1) # 2 rows, 2 columns, 1st subplot
ax1.errorbar(
    df["date"], # Date on the x axis
    df["mean"], # Mean on the y axis
    yerr=df["deviation"], # Use the deviation for the error bars
    color="orange", # Plot the sunspot (mean) data in orange
    ecolor="black"
) # Show the error bars in black
ax1.set_xlabel("Date")
ax1.set_ylabel("Mean")
ax1.set_title("From Dataframe")

ax2 = fig.add_subplot(2, 2, 2) # 2 rows, 2 columns, 2nd subplot
ax2.scatter(df["date"], df["observations"], color="grey", marker="+")
ax2.set_xlabel("Date")
ax2.set_ylabel("Number of Observations")
ax2.set_title("From Dataframe")

ax3 = fig.add_subplot(2, 2, 3) # 2 rows, 2 columns, 3rd subplot
ax3.errorbar(date, mean, yerr=deviation, color="pink", ecolor="black")
ax3.set_xlabel("Date")
ax3.set_ylabel("Mean")
ax3.set_title("From List")

ax4 = fig.add_subplot(2, 2, 4) # 2 rows, 2 columns, 4th subplot
ax4.scatter(date, observations, color="red", marker="o")
ax4.set_xlabel("Date")
ax4.set_ylabel("Number of Observations")
ax4.set_title("From List")
```



In this example we are plotting columns from the `pandas DataFrame` (series), and from lists to show this method works for both. `numpy` arrays can also be used.

3.3.9 Learning More

There's so much more to learn about `matplotlib`: pie charts, bar charts, heat maps, 3-d plotting, animated plots, and so on. You can learn all this via the [Matplotlib Website](#). You should try to get comfortable with all this, so please use some time in class, or at home, to work your way through a bunch of the [examples](#).

3.4 NumPy

Estimated time to complete this notebook: 20 minutes

3.4.1 Limitations of Python Lists

The normal Python List is just one dimensional. To make a matrix, we have to nest Python lists:

```
x = [list(range(5)) for N in range(5)]
```

```
x
```

```
[[0, 1, 2, 3, 4],  
 [0, 1, 2, 3, 4],  
 [0, 1, 2, 3, 4],  
 [0, 1, 2, 3, 4],  
 [0, 1, 2, 3, 4]]
```

```
x[2][2]
```

```
2
```

Applying an operation to every element is a pain:

```
x + 5
```

```
-----  
TypeError: can only concatenate list (not "int") to list  
Cell In [4], line 1  
----> 1 x + 5
```

```
[[elem + 5 for elem in row] for row in x]
```

```
[[5, 6, 7, 8, 9],  
 [5, 6, 7, 8, 9],  
 [5, 6, 7, 8, 9],  
 [5, 6, 7, 8, 9],  
 [5, 6, 7, 8, 9]]
```

Common useful operations like transposing a matrix or reshaping a 10 by 10 matrix into a 20 by 5 matrix are not easy to code in raw Python lists.

3.4.2 The NumPy array

NumPy's array type represents a multidimensional matrix $M_{i,j,k,\dots,n}$

The NumPy array seems at first to be just like a list:

```
import numpy as np  
my_array = np.array(range(5))
```

```
my_array
```

```
array([0, 1, 2, 3, 4])
```

```
my_array[2]
```

```
2
```

```
for element in my_array:  
    print("Hello" * element)
```

```
Hello  
HelloHello  
HelloHelloHello  
HelloHelloHelloHello
```

We can also see our first weakness of NumPy arrays versus Python lists:

```
my_array.append(4)
```

```
-----  
AttributeError  
Cell In [10], line 1  
----> 1 my_array.append(4)  
  
AttributeError: 'numpy.ndarray' object has no attribute 'append'
```

For NumPy arrays, you typically don't change the data size once you've defined your array, whereas for Python lists, you can do this efficiently. However, you get back lots of goodies in return...

3.4.3 Elementwise Operations

But most operations can be applied element-wise automatically!

```
my_array + 2
```

```
array([2, 3, 4, 5, 6])
```

These "vectorized" operations are very fast: (see [here](#) for more information on the `%timeit` magic)

```
import numpy as np  
  
big_list = range(10000)  
big_array = np.arange(10000)
```

```
%timeit  
[x**2 for x in big_list]
```

```
3.06 ms ± 5.11 µs per loop (mean ± std. dev. of 7 runs, 100 loops each)
```

```
%timeit  
big_array**2
```

```
4.33 µs ± 1.95 ns per loop (mean ± std. dev. of 7 runs, 100,000 loops each)
```

3.4.4 Arange and linspace

NumPy has two easy methods for defining floating-point evenly spaced arrays:

```
x = np.arange(0, 10, 0.1) # Start, stop, step size
```

Note that using non-integer step size does not work with Python lists:

```
y = list(range(0, 10, 0.1))
```

```
-----  
TypeError  
Cell In [16], line 1  
----> 1 y = list(range(0, 10, 0.1))  
  
TypeError: 'float' object cannot be interpreted as an integer
```

Similarly, we can quickly an evenly spaced range of a known size (e.g. for graph plotting):

```
import math  
  
values = np.linspace(0, math.pi, 100) # Start, stop, number of steps
```

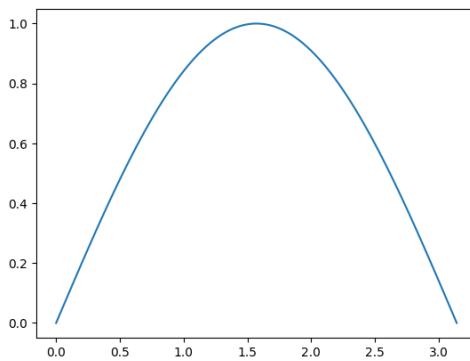
```
values
```

```
array([0.         , 0.00173226, 0.06346652, 0.09519978, 0.12693304,  
0.1586663 , 0.19039955, 0.22213281, 0.25386607, 0.28559935,  
0.31733259 , 0.34906585, 0.38079911, 0.41253237, 0.44426563,  
0.47599889 , 0.50773215, 0.53946541, 0.57119866, 0.60293192,  
0.63466518 , 0.66639844, 0.6981317 , 0.72986496, 0.76159822,  
0.79333148 , 0.82506474, 0.856798 , 0.88853126, 0.92026451,  
0.95199777 , 0.98373103, 1.01546429, 1.04719755, 1.07893081,  
1.11066407 , 1.14239733, 1.17413059, 1.20586385, 1.23759711,  
1.26933037 , 1.30106362, 1.33279688, 1.36453014, 1.3962634 ,  
1.42799666 , 1.45972992, 1.49146318, 1.52319644, 1.5549297 ,  
1.58666296 , 1.61839622, 1.65012947, 1.68186273, 1.71359599,  
1.74532925 , 1.77766251, 1.80879577, 1.84052903, 1.87226229,  
1.90399555 , 1.93572881, 1.96746207, 1.99919533, 2.03092858,  
2.06266184 , 2.0943951 , 2.12612836, 2.15786162, 2.18959488,  
2.22132814 , 2.2536614 , 2.28479466, 2.31652792, 2.34826118,  
2.37999443 , 2.41172769, 2.44346995, 2.47519421, 2.50692747,  
2.53866673 , 2.57039399, 2.60212725, 2.63386651, 2.66559377,  
2.69732703 , 2.72966028, 2.76679354, 2.7925268 , 2.82426006,  
2.85599332 , 2.88772658, 2.91945984, 2.9511931 , 2.98292636,  
3.01465962 , 3.04639288, 3.07812614, 3.10985939, 3.14159265])
```

NumPy comes with 'vectorised' versions of common functions which work element-by-element when applied to arrays:

```
from matplotlib import pyplot as plt  
  
plt.plot(values, np.sin(values))
```

```
[<matplotlib.lines.Line2D at 0x7f08941edfa0>]
```



So we don't have to use awkward list comprehensions when using these.

3.4.5 Multi-Dimensional Arrays

NumPy's true power comes from multi-dimensional arrays:

```
np.zeros([3, 4, 2]) # 3 arrays with 4 rows and 2 columns each
```

```
array([[[0., 0.],
       [0., 0.],
       [0., 0.],
       [0., 0.]],

      [[0., 0.],
       [0., 0.],
       [0., 0.],
       [0., 0.]],

      [[0., 0.],
       [0., 0.],
       [0., 0.],
       [0., 0.]])
```

Unlike a list-of-lists in Python, we can reshape arrays:

```
x = np.array(range(40))
x
```

```
array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
       17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
       34, 35, 36, 37, 38, 39])
```

```
y = x.reshape([4, 5, 2]) # 4 Arrays - 5 Rows - 2 Columns
y
```

```
array([[[ 0,  1],
       [ 2,  3],
       [ 4,  5],
       [ 6,  7],
       [ 8,  9]],

      [[10, 11],
       [12, 13],
       [14, 15],
       [16, 17],
       [18, 19]],

      [[20, 21],
       [22, 23],
       [24, 25],
       [26, 27],
       [28, 29]],

      [[30, 31],
       [32, 33],
       [34, 35],
       [36, 37],
       [38, 39]])
```

And index multiple columns at once:

```
y[3, 2, 1]
```

```
35
```

Including selecting on inner axes while taking all from the outermost:

```
y[:, 2, 1]
```

```
array([ 5, 15, 25, 35])
```

And subselecting ranges:

```
y[2:, :1, :] # Last 2 axes, 1st row, all columns
```

```
array([[[20, 21]],
      [[30, 31]])
```

And [transpose](#) arrays:

```
y.transpose()
```

```
array([[[ 0, 10, 20, 30],  
       [ 2, 12, 22, 32],  
       [ 4, 14, 24, 34],  
       [ 6, 16, 26, 36],  
       [ 8, 18, 28, 38]],  
  
      [[ 1, 11, 21, 31],  
       [ 3, 13, 23, 33],  
       [ 5, 15, 25, 35],  
       [ 7, 17, 27, 37],  
       [ 9, 19, 29, 39]]])
```

You can get the dimensions of an array with `shape`

```
y.shape # 4 Arrays - 5 Rows - 2 Columns
```

```
(4, 5, 2)
```

```
y.transpose().shape # 2 Arrays - 5 Rows - 4 Columns
```

```
(2, 5, 4)
```

Some numpy functions apply by default to the whole array, but can be chosen to act only on certain axes:

```
x = np.arange(12).reshape(4, 3)  
x
```

```
array([[ 0,  1,  2],  
       [ 3,  4,  5],  
       [ 6,  7,  8],  
       [ 9, 10, 11]])
```

```
x.mean(1) # Mean along the second axis, leaving the first.
```

```
array([ 1.,  4.,  7., 10.])
```

```
x.mean(0) # Mean along the first axis, leaving the second.
```

```
array([4.5, 5.5, 6.5])
```

```
x.mean() # mean of all axes
```

```
5.5
```

3.4.6 Array Datatypes

A Python `list` can contain data of mixed type:

```
x = ["hello", 2, 3.4]
```

```
type(x[2])
```

```
float
```

```
type(x[1])
```

```
int
```

A NumPy array always contains just one datatype:

```
np.array(x)
```

```
array(['hello', '2', '3.4'], dtype='<U32')
```

NumPy will choose the least-generic-possible datatype that can contain the data:

```
y = np.array([2, 3.4])
```

```
y
```

```
array([2., 3.4])
```

You can access the array's `dtype`, or check the type of individual elements:

```
y.dtype
```

```
dtype('float64')
```

```
type(y[0])
```

```
numpy.float64
```

```
z = np.array([3, 4, 5])  
z
```

```
array([3, 4, 5])
```

```
type(z[0])
```

```
numpy.int64
```

The results are, when you get to know them, fairly obvious string codes for datatypes: NumPy supports all kinds of datatypes beyond the python basics.

NumPy will convert python type names to dtypes:

```
x = [2, 3.6, 7.2, 0]
```

```
int_array = np.array(x, dtype=int)
```

```
int_array
```

```
array([2, 3, 7, 0])
```

```
int_array.dtype
```

```
dtype('int64')
```

```
float_array = np.array(x, dtype=float)
```

```
float_array
```

```
array([2., 3.6, 7.2, 0.])
```

```
float_array.dtype
```

```
dtype('float64')
```

3.5 Advanced NumPy

Estimated time to complete this notebook: 20 minutes

3.5.1 Recap

In the previous section we introduced numpy array that represents a multidimensional matrix $M_{i,j,k,\dots,n}$. Which, among other things, allows for vectorised versions of common functions.

```
import numpy as np
```

3.5.2 Broadcasting

This is another really powerful feature of NumPy.

By default, array operations are element-by-element:

```
np.arange(5) * np.arange(5)
```

```
array([ 0,  1,  4,  9, 16])
```

If we multiply arrays with non-matching shapes we get an error:

```
np.arange(5) * np.arange(6)
```

```
ValueError                                Traceback (most recent call last)
Cell In [3], line 1
----> 1 np.arange(5) * np.arange(6)

ValueError: operands could not be broadcast together with shapes (5,) (6,)
```

```
np.zeros([2, 3]) * np.zeros([2, 4])
```

```
ValueError                                Traceback (most recent call last)
Cell In [4], line 1
----> 1 np.zeros([2, 3]) * np.zeros([2, 4])

ValueError: operands could not be broadcast together with shapes (2,3) (2,4)
```

```
m1 = np.arange(100).reshape([10, 10])
```

```
m2 = np.arange(100).reshape([10, 5, 2])
```

```
m1 + m2
```

```
ValueError                                Traceback (most recent call last)
Cell In [7], line 1
----> 1 m1 + m2

ValueError: operands could not be broadcast together with shapes (10,10) (10,5,2)
```

Arrays must match in **all** dimensions in order to be compatible:

```
np.ones([3, 3]) * np.ones([3, 3]) # Note elementwise multiply, *not* matrix
                                 multiply.
```

```
array([[1., 1., 1.],
       [1., 1., 1.],
       [1., 1., 1.]])
```

```

m3 = np.arange(9).reshape([3, 3])
m3

array([[0, 1, 2],
       [3, 4, 5],
       [6, 7, 8]])

m4 = np.arange(9, 18).reshape([3, 3])
m4

array([[ 9, 10, 11],
       [12, 13, 14],
       [15, 16, 17]])

m3 * m4 # Note elementwise multiply, *not* matrix multiply.

array([[ 0, 10, 22],
       [36, 52, 70],
       [90, 112, 136]])

```

Except, that if one array has any Dimension 1, then the data is REPEATED to match the other.

```

col = np.arange(10).reshape([10, 1])
col

array([[0],
       [1],
       [2],
       [3],
       [4],
       [5],
       [6],
       [7],
       [8],
       [9]])

row = col.transpose()
row

array([[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]])

col.shape # "Column Vector"

(10, 1)

row.shape # "Row Vector"

(1, 10)

row + col

array([[ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9],
       [ 1,  2,  3,  4,  5,  6,  7,  8,  9, 10],
       [ 2,  3,  4,  5,  6,  7,  8,  9, 10, 11],
       [ 3,  4,  5,  6,  7,  8,  9, 10, 11, 12],
       [ 4,  5,  6,  7,  8,  9, 10, 11, 12, 13],
       [ 5,  6,  7,  8,  9, 10, 11, 12, 13, 14],
       [ 6,  7,  8,  9, 10, 11, 12, 13, 14, 15],
       [ 7,  8,  9, 10, 11, 12, 13, 14, 15, 16],
       [ 8,  9, 10, 11, 12, 13, 14, 15, 16, 17],
       [ 9, 10, 11, 12, 13, 14, 15, 16, 17, 18]])

10 * row + col

array([[ 0, 10, 20, 30, 40, 50, 60, 70, 80, 90],
       [ 1, 11, 21, 31, 41, 51, 61, 71, 81, 91],
       [ 2, 12, 22, 32, 42, 52, 62, 72, 82, 92],
       [ 3, 13, 23, 33, 43, 53, 63, 73, 83, 93],
       [ 4, 14, 24, 34, 44, 54, 64, 74, 84, 94],
       [ 5, 15, 25, 35, 45, 55, 65, 75, 85, 95],
       [ 6, 16, 26, 36, 46, 56, 66, 76, 86, 96],
       [ 7, 17, 27, 37, 47, 57, 67, 77, 87, 97],
       [ 8, 18, 28, 38, 48, 58, 68, 78, 88, 98],
       [ 9, 19, 29, 39, 49, 59, 69, 79, 89, 99]])

```

This works for arrays with more than one unit dimension.

3.5.3 Another example

```

x = np.array([1, 2]).reshape(1, 2)
x

array([[1, 2]])

y = np.array([3, 4, 5]).reshape(3, 1)
y

array([[3],
       [4],
       [5]])

result = x + y
result.shape

(3, 2)

result

```

```
array([[4, 5],  
       [5, 6],  
       [6, 7]])
```

What numpy is doing:

Numpy broadcasting example

3.5.4 Newaxis

Broadcasting is very powerful, and numpy allows indexing with `np.newaxis` to temporarily create new one-long dimensions on the fly.

```
import numpy as np  
  
x = np.arange(10).reshape(2, 5)  
y = np.arange(8).reshape(2, 2, 2)
```

x

```
array([[0, 1, 2, 3, 4],  
       [5, 6, 7, 8, 9]])
```

y

```
array([[[0, 1],  
           [2, 3]],  
           [[4, 5],  
           [6, 7]]])
```

```
x_dash = x[:, :, np.newaxis, np.newaxis]  
x_dash.shape
```

(2, 5, 1, 1)

```
y_dash = y[:, np.newaxis, :, :]  
y_dash.shape
```

(2, 1, 2, 2)

y_dash

```
array([[[[0, 1],  
           [2, 3]],  
           [[4, 5],  
           [6, 7]]]])
```

```
res = x_dash * y_dash
```

res.shape

(2, 5, 2, 2)

```
np.sum(res)
```

830

Note that `newaxis` works because a $3 \times 1 \times 3$ array and a 3×3 array contain the same data, differently shaped:

```
threebythree = np.arange(9).reshape(3, 3)  
threebythree
```

```
array([[0, 1, 2],  
       [3, 4, 5],  
       [6, 7, 8]])
```

threebythree[:, np.newaxis, :]

```
array([[[0, 1, 2]],  
           [[3, 4, 5]],  
           [[6, 7, 8]]])
```

3.5.5 Dot Products using broadcasting

NumPy multiply is element-by-element, not a dot-product:

```
a = np.arange(9).reshape(3, 3)  
a
```

```
array([[0, 1, 2],  
       [3, 4, 5],  
       [6, 7, 8]])
```

```
b = np.arange(3, 12).reshape(3, 3)  
b
```

```
array([[ 3,  4,  5],  
       [ 6,  7,  8],  
       [ 9, 10, 11]])
```

```
a * b
```

```
array([[ 0,  4, 10],  
       [18, 28, 40],  
       [54, 70, 88]])
```

We can use what we've learned about the algebra of broadcasting and newaxis to get a dot-product, (matrix inner product).

First we add new axes to A and B :

```
a[:, :, np.newaxis].shape
```

```
(3, 3, 1)
```

```
b[np.newaxis, :, :].shape
```

```
(1, 3, 3)
```

Now we use broadcasting to generate $A_{ij}B_{jk}$ as a 3-d matrix:

```
a[:, :, np.newaxis] * b[np.newaxis, :, :]
```

```
array([[[ 0,  0,  0],  
        [ 6,  7,  8],  
        [18, 20, 22]],  
  
      [[ 9, 12, 15],  
       [24, 28, 32],  
       [45, 50, 55]],  
  
      [[18, 24, 30],  
       [42, 49, 56],  
       [72, 80, 88]]])
```

Then we sum over the middle, j axis, [which is the 1-axis of three axes numbered (0,1,2)] of this 3-d matrix. Thus we generate $\sum_j A_{ij}B_{jk}$.

```
(a[:, :, np.newaxis] * b[np.newaxis, :, :]).sum(1)
```

```
array([[ 24,  27,  30],  
       [ 78,  90, 102],  
       [132, 153, 174]])
```

Or if you prefer:

```
(a.reshape(3, 3, 1) * b.reshape(1, 3, 3)).sum(1)
```

```
array([[ 24,  27,  30],  
       [ 78,  90, 102],  
       [132, 153, 174]])
```

We can see that the broadcasting concept gives us a powerful and efficient way to express many linear algebra operations computationally.

3.5.6 Dot Products using numpy functions

However, as the dot-product is a common operation, `numpy` has a built in function:

```
np.dot(a, b)
```

```
array([[ 24,  27,  30],  
       [ 78,  90, 102],  
       [132, 153, 174]])
```

This can also be written as:

```
a.dot(b)
```

```
array([[ 24,  27,  30],  
       [ 78,  90, 102],  
       [132, 153, 174]])
```

If you are using `Python 3.5` or later, a dedicated matrix multiplication operator has been added, allowing you to do the following:

```
a @ b
```

```
array([[ 24,  27,  30],  
       [ 78,  90, 102],  
       [132, 153, 174]])
```

3.5.7 Record Arrays

These are a special array structure designed to match the CSV "Record and Field" model. It's a very different structure from the normal NumPy array, and different fields can contain different datatypes. We saw this when we looked at CSV files:

```
x = np.arange(50).reshape([10, 5])
```

```
record_x = x.view(  
    dtype={"names": ["col1", "col2", "another", "more", "last"], "formats": [int]  
    * 5})
```

```
record_x
```

```
array([[( 0,  1,  2,  3,  4)],
   [( 5,  6,  7,  8,  9)],
   [(10, 11, 12, 13, 14)],
   [(15, 16, 17, 18, 19)],
   [(20, 21, 22, 23, 24)],
   [(25, 26, 27, 28, 29)],
   [(30, 31, 32, 33, 34)],
   [(35, 36, 37, 38, 39)],
   [(40, 41, 42, 43, 44)],
   [(45, 46, 47, 48, 49)]],
      dtype=[('col1', '<i8'), ('col2', '<i8'), ('another', '<i8'), ('more', '<i8'),
             ('last', '<i8')])
```

Record arrays can be addressed with field names like they were a dictionary:

```
{ record_x["col1"] }
```

```
array([[ 0],
   [ 5],
   [10],
   [15],
   [20],
   [25],
   [30],
   [35],
   [40],
   [45]])
```

Indeed we can use these methods when parsing CSV files instead of using Pandas.

3.5.8 Logical arrays, masking, and selection

Numpy defines operators like == and < to apply to arrays *element by element*:

```
{ x = np.zeros([3, 4])
x }
```

```
array([[0., 0., 0., 0.],
   [0., 0., 0., 0.],
   [0., 0., 0., 0.]])
```

```
{ y = np.arange(-1, 2)[:, np.newaxis] * np.arange(-2, 2)[np.newaxis, :]
y }
```

```
array([[ 2,  1,  0, -1],
   [ 0,  0,  0,  0],
   [-2, -1,  0,  1]])
```

```
{ y_is_one = y == 1
y_is_one }
```

```
array([[False,  True, False, False],
   [False, False, False, False],
   [False, False, False,  True]])
```

```
{ aresame = x == y
aresame }
```

```
array([[False, False,  True, False],
   [ True,  True,  True,  True],
   [False, False,  True, False]])
```

A logical array can be used to select elements from an array:

```
{ y[np.logical_not(aresame)] }
```

```
array([ 2,  1, -1, -2, -1,  1])
```

Although when printed, this comes out as a flat list, if assigned to, the *selected elements of the array are changed!*

```
{ y[aresame] = 5
```

```
{ y }
```

```
array([[ 2,  1,  5, -1],
   [ 5,  5,  5,  5],
   [-2, -1,  5,  1]])
```

3.5.9 Numpy memory

NumPy manages memory differently from lists. Changing an element in a copy of a list does not change the original list.

```
{ x = list(range(5))
y = x[:] }
```

```
{ y[2] = 0
x }
```

```
[0, 1, 2, 3, 4]
```

But in NumPy, changing the copy **does** change the original array!

```
{ x = np.arange(5)
y = x[:] }
```

```
{ y[2] = 0
x }
```

```
array([0, 1, 0, 3, 4])
```

We must use `np.copy` to force separate memory. Otherwise NumPy tries its hardest to make slices be views on data.

3.6 The Boids!

Estimated time to complete this notebook: 45 minutes.

⚠ Warning: Advanced Topic! ⚠

Our earlier discussion of NumPy was very theoretical, but let's go through a practical example, and see how powerful NumPy can be.

Note this is more a showcase of what you can do with numpy than an exhaustive notebook to work through

3.6.1 Flocking

The aggregate motion of a flock of birds, a herd of land animals, or a school of fish is a beautiful and familiar part of the natural world... The aggregate motion of the simulated flock is created by a distributed behavioral model much like that at work in a natural flock; the birds choose their own course. Each simulated bird is implemented as an independent actor that navigates according to its local perception of the dynamic environment, the laws of simulated physics that rule its motion, and a set of behaviors programmed into it... The aggregate motion of the simulated flock is the result of the dense interaction of the relatively simple behaviors of the individual simulated birds.

— Craig W. Reynolds, "Flocks, Herds, and Schools: A Distributed Behavioral Model", *Computer Graphics* 21 4 1987, pp 25-34

We will demonstrate an algorithm to simulate flocking behaviour in numpy. The simulation consists of a set of individual bird-like objects that we will call 'boids' following the nomenclature of [the original paper](#) for more details.

- Collision Avoidance: avoid collisions with nearby flockmates
- Velocity Matching: attempt to match velocity with nearby flockmates
- Flock Centering: attempt to stay close to nearby flockmates

3.6.2 Setting up the Boids

Our boids will each have an x velocity and a y velocity, and an x position and a y position.

We'll build this up in NumPy notation, and eventually, have an animated simulation of our flying boids.

```
import numpy as np
```

Let's start with simple flying in a straight line.

Our positions, for each of our N boids, will be an array, shape $2 \times N$, with the x positions in the first row, and y positions in the second row.

```
boid_count = 10
```

We'll want to be able to seed our Boids in a random position.

We'd better define the edges of our simulation area:

```
limits = np.array([2000, 2000])
```

```
positions = np.random.rand(2, boid_count) * limits[:, np.newaxis]
```

```
array([[1438.10610604, 104.57711773, 586.74272348, 1552.86976233,
       1025.99672642, 1498.44026519, 689.63950657, 1621.68611794,
       711.56369443, 1913.26373318],
       [1136.05314687, 62.02366888, 502.51857237, 385.82868095,
       69.78296196, 592.7799406, 1240.58854526, 1109.6197646 ,
       499.3865635 , 100.90304263]])
```

```
positions.shape
```

```
(2, 10)
```

We used **broadcasting** with `np.newaxis` to apply our upper limit to each boid. `rand` gives us a random number between 0 and 1. We multiply by our limits to get a number up to that limit.

```
limits[:, np.newaxis]
```

```
array([[2000],
       [2000]])
```

```
limits[:, np.newaxis].shape
```

```
(2, 1)
```

```
np.random.rand(2, boid_count).shape
```

```
(2, 10)
```

So we multiply a 2×1 array by a 2×10 array – and get a 2×10 array.

Let's put that in a function:

```
def new_flock(count, lower_limits, upper_limits):
    width = upper_limits - lower_limits
    return lower_limits[:, np.newaxis] + np.random.rand(2, count) * width[:, np.newaxis]
```

For example, let's assume that we want our initial positions to vary between 100 and 200 in the x axis, and 900 and 1100 in the y axis. We can generate random positions within these constraints with:

```
positions = new_flock(boid_count, np.array([100, 900]), np.array([200, 1100]))
```

But each boid will also need a starting velocity. Let's make these random too:

We can reuse the `new_flock` function defined above, since we're again essentially just generating random numbers from given limits. This saves us some code, but keep in mind that using a function for something other than what its name indicates can become confusing!

Here, we will let the initial x velocities range over $[0, 10]$ and the y velocities over $[-20, 20]$.

```
velocities = new_flock(boid_count, np.array([0, -20]), np.array([10, 20]))
```

```
array([[ 8.52785147,  9.01697347,  9.7999047 ,  6.29148724,
       4.11364647,  7.37300151,  6.75214623,  0.52717935,
      9.7926317 ,  7.08838931],
       [ 9.78142517, -13.89419223,  18.27375658, -10.066592852,
      -2.05357269, -5.04292023,  10.50063368, -5.76425356,
     14.71795737, -17.11148764]])
```

3.6.3 Flying in a Straight Line

Now we see the real amazingness of NumPy: if we want to move our *whole flock* according to

$$\delta_x = \delta_t \cdot \frac{dv}{dt}$$

we just do:

```
positions += velocities
```

3.6.4 Matplotlib Animations

So now we can animate our Boids using the matplotlib animation tools All we have to do is import the relevant libraries:

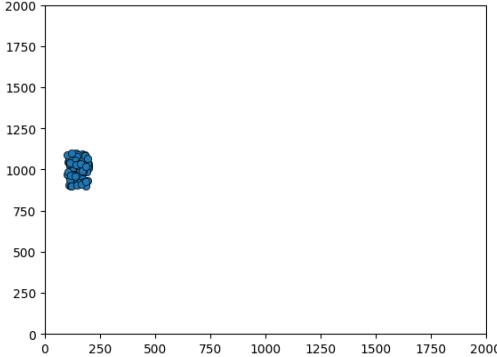
```
from matplotlib import animation
from matplotlib import pyplot as plt
```

Then, we make a static plot, showing our first frame:

```
# create a simple plot
# initial x position in [100, 200], initial y position in [900, 1100]
# initial x velocity in [0, 10], initial y velocity in [-20, 20]
positions = new_flock(100, np.array([100, 900]), np.array([200, 1100]))
velocities = new_flock(100, np.array([0, -20]), np.array([10, 20]))
```

```
figure = plt.figure()
axes = plt.axes(xlim=(0, limits[0]), ylim=(0, limits[1]))
scatter = axes.scatter(
    positions[0, :], positions[1, :], marker="o", edgecolor="k", lw=0.5
)
scatter
```

```
<matplotlib.collections.PathCollection at 0x7f220511c7f0>
```



Then, we define a function which `updates` the figure for each timestep

```
def update_boids(positions, velocities):
    positions += velocities

def animate(frame):
    update_boids(positions, velocities)
    scatter.set_offsets(positions.transpose())
```

Call `FuncAnimation`, and specify how many frames we want:

```
anim = animation.FuncAnimation(figure, animate, frames=50, interval=50)
```

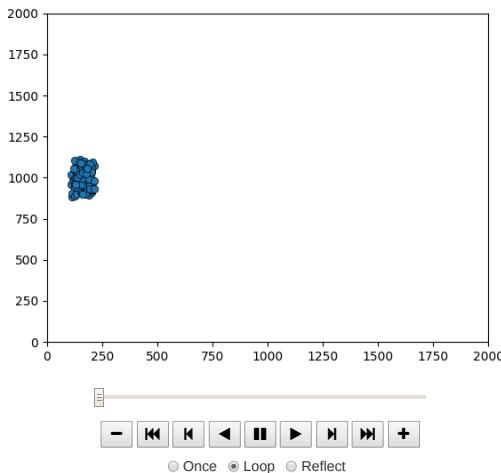
Save out the figure:

```
positions = new_flock(100, np.array([100, 900]), np.array([200, 1100]))
velocities = new_flock(100, np.array([0, -20]), np.array([10, 20]))
anim.save("boids_1.gif")
```

And download the [saved animation](#).

You can even view the results directly in the notebook.

```
from IPython.display import HTML
positions = new_flock(100, np.array([100, 900]), np.array([200, 1100]))
velocities = new_flock(100, np.array([0, -20]), np.array([10, 20]))
HTML(anim.to_jshtml())
```



3.6.5 Extended content: The Boids!

The examples given below are examples of how to use numpy to efficiently apply equations to arrays. here are many potential ways to do such things, and this is intended as a showcase of numpy's versatility rather than a prescribed set of rules.

3.6.5.0 Fly towards the middle

Boids try to fly towards the middle:

```

positions = new_flock(4, np.array([100, 900]), np.array([200, 1100]))
velocities = new_flock(4, np.array([0, -20]), np.array([10, 20]))

positions
array([[ 101.71265941,  128.15168983,  194.13384124,  188.93630321],
       [1020.14874904, 1054.13600976, 1050.37566046, 992.98763079]])

velocities
array([[ 2.65258421,  4.55120664,   8.24940019,   6.91700657],
       [ 11.07262968, -17.06343578, -7.38907855, -5.84523479]])

middle = np.mean(positions, 1)
middle
array([ 153.23362342, 1029.41201251])

direction_to_middle = positions - middle[:, np.newaxis]
direction_to_middle
array([[-51.52096401, -25.08193359,  40.90021782,  35.70267979],
       [-9.26326347,  24.72399725,  20.96364794, -36.42438173]])

```

This is easier and faster than:

```

for boid in boids:
    for dimension in [0, 1]:
        direction_to_middle[dimension][boid] = positions[dimension][boid] - middle[dimension]

        move_to_middle_strength = 0.01
        velocities = velocities - direction_to_middle * move_to_middle_strength

```

Let's update our function, and animate that:

```

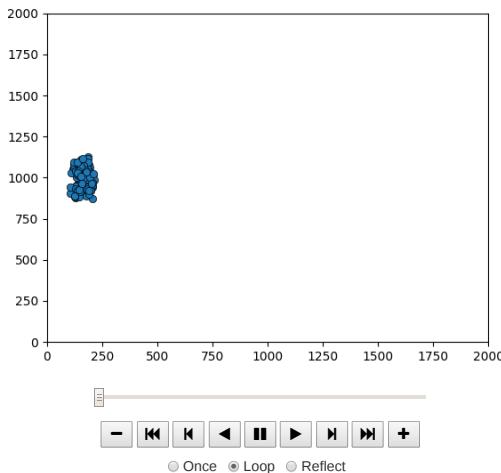
def update_boids(positions, velocities):
    move_to_middle_strength = 0.01
    middle = np.mean(positions, 1)
    direction_to_middle = positions - middle[:, np.newaxis]
    velocities -= direction_to_middle * move_to_middle_strength
    positions += velocities

def animate(frame):
    update_boids(positions, velocities)
    scatter.set_offsets(positions.transpose())

anim = animation.FuncAnimation.figure(figure, animate, frames=50, interval=50)

positions = new_flock(100, np.array([100, 900]), np.array([200, 1100]))
velocities = new_flock(100, np.array([0, -20]), np.array([10, 20]))
HTML(anim.to_jshtml())

```



3.6.5.1 Avoiding collisions

We'll want to add our other flocking rules to the behaviour of the Boids.

We'll need a matrix giving the distances between each boid. This should be $N \times N$.

```
positions = new_flock(4, np.array([100, 900]), np.array([200, 1100]))
velocities = new_flock(4, np.array([0, -20]), np.array([10, 20]))
```

We might think that we need to do the X-distances and Y-distances separately:

```
xpos = positions[0, :]
xsep_matrix = xpos[:, np.newaxis] - xpos[np.newaxis, :]
xsep_matrix.shape
(4, 4)
xsep_matrix
array([[ 0.          , -16.30904627, -61.71057297, -23.51830568],
       [16.30904627,  0.          , -45.4015267 , -7.20925941],
       [61.71057297,  45.4015267 ,  0.          ,  38.19226729],
       [23.51830568,  7.20925941, -38.19226729,  0.        ]])
```

But in NumPy we can be cleverer than that, and make a $2 \times N \times N$ matrix of separations:

```
separations = positions[:, np.newaxis, :] - positions[:, :, np.newaxis]
separations.shape
(2, 4, 4)
```

And then we can get the sum-of-squares $\delta_x^2 + \delta_y^2$ like this:

```
squared_displacements = separations * separations
square_distances = np.sum(squared_displacements, 0)
square_distances
array([[ 0.          ,  7015.98128261, 39508.36233983, 1465.17211434],
       [7015.98128261,  0.          , 13464.65916617, 2751.60786685],
       [39508.36233983, 13464.65916617,  0.          , 26658.47352582],
       [1465.17211434, 2751.60786685, 26658.47352582,  0.        ]])
```

Now we need to find boids that are too close:

```
alert_distance = 2000
close_boids = square_distances < alert_distance
close_boids
array([[ True, False, False,  True],
       [False,  True, False, False],
       [False, False,  True, False],
       [ True, False, False,  True]])
```

Find the direction distances **only** to those boids which are too close:

```
separations_if_close = np.copy(separations)
far_away = np.logical_not(close_boids)
```

Set **x** and **y** values in **separations_if_close** to zero if they are far away:

```

separations_if_close[0, :, :][far_away] = 0
separations_if_close[1, :, :][far_away] = 0
separations_if_close

array([[[ 0.        ,  0.        ,  0.        ,  23.51830568],
       [ 0.        ,  0.        ,  0.        ,  0.        ],
       [ 0.        ,  0.        ,  0.        ,  0.        ],
       [-23.51830568,  0.        ,  0.        ,  0.        ]],
      [[ 0.        ,  0.        ,  0.        , -30.2003545],
       [ 0.        ,  0.        ,  0.        ,  0.        ],
       [ 0.        ,  0.        ,  0.        ,  0.        ],
       [ 30.2003545,  0.        ,  0.        ,  0.        ]]])

```

And fly away from them:

```

np.sum(separations_if_close, 2)

array([[ 23.51830568,   0.        ,   0.        , -23.51830568],
       [-30.2003545,   0.        ,   0.        ,  30.2003545]])

velocities = velocities + np.sum(separations_if_close, 2)

```

Now we can update our animation:

```

def update_boids(positions, velocities):
    move_to_middle_strength = 0.01
    middle = np.mean(positions, 1)
    direction_to_middle = positions - middle[:, np.newaxis]
    velocities -= direction_to_middle * move_to_middle_strength

    separations = positions[:, np.newaxis, :] - positions[:, :, np.newaxis]
    squared_displacements = separations * separations
    square_distances = np.sum(squared_displacements, 0)
    alert_distance = 100
    far_away = square_distances > alert_distance
    separations_if_close = np.copy(separations)
    separations_if_close[0, :, :][far_away] = 0
    separations_if_close[1, :, :][far_away] = 0
    velocities += np.sum(separations_if_close, 1)

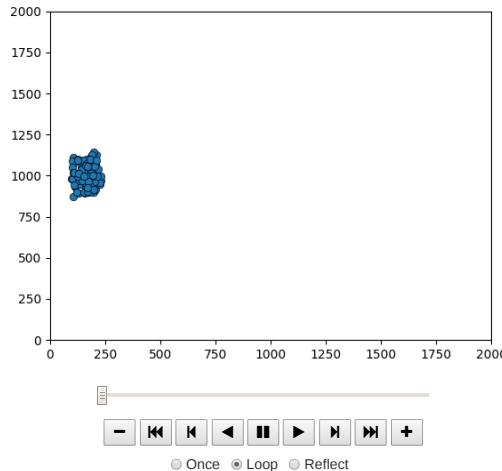
    positions += velocities

def animate(frame):
    update_boids(positions, velocities)
    scatter.set_offsets(positions.transpose())

anim = animation.FuncAnimation(figure, animate, frames=50, interval=50)

positions = new_flock(100, np.array([100, 900]), np.array([200, 1100]))
velocities = new_flock(100, np.array([0, -20]), np.array([10, 20]))
HTML(anim.to_jshtml())

```



3.6.5.2 Match speed with nearby boids

This is pretty similar:

```

def update_boids(positions, velocities):
    move_to_middle_strength = 0.01
    middle = np.mean(positions, 1)
    direction_to_middle = positions - middle[:, np.newaxis]
    velocities -= direction_to_middle * move_to_middle_strength

    separations = positions[:, np.newaxis, :] - positions[:, :, np.newaxis]
    squared_displacements = separations * separations
    square_distances = np.sum(squared_displacements, 0)
    alert_distance = 100
    far_away = square_distances > alert_distance
    separations_if_close = np.copy(separations)
    separations_if_close[0, :, :] [far_away] = 0
    separations_if_close[1, :, :] [far_away] = 0
    velocities += np.sum(separations_if_close, 1)

    velocity_differences = velocities[:, np.newaxis, :] - velocities[:, :, np.newaxis]
    formation_flying_distance = 10000
    formation_flying_strength = 0.125
    very_far = square_distances > formation_flying_distance
    velocity_differences_if_close = np.copy(velocity_differences)
    velocity_differences_if_close[0, :, :] [very_far] = 0
    velocity_differences_if_close[1, :, :] [very_far] = 0
    velocities -= np.mean(velocity_differences_if_close, 1) *
    formation_flying_strength

    positions += velocities

```



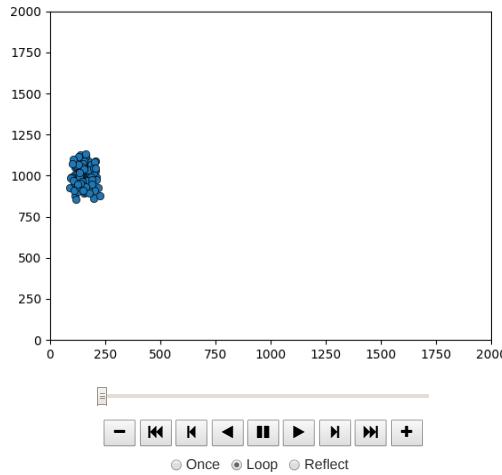
```

def animate(frame):
    update_boids(positions, velocities)
    scatter.set_offsets(positions.transpose())

anim = animation.FuncAnimation(figure, animate, frames=200, interval=50)

positions = new_flock(100, np.array([100, 900]), np.array([200, 1100]))
velocities = new_flock(100, np.array([0, -20]), np.array([10, 20]))
HTML(anim.to_jshtml())

```



Hopefully the power of `numpy` should be pretty clear now. This would be **enormously slower** and, I think, harder to understand using traditional lists.

3.7 Classroom Exercises

List of exercises and estimated completion times

[3a - Saving and Loading Data](#) 5 minutes

[3b - Plotting with matplotlib](#) 10 minutes

[3c The Biggest Earthquake in the UK This Century](#) 30 minutes

Exercise 3a Saving and Loading Data

Relevant sections: 3.2.2, 3.2.3

Use YAML or JSON to save your maze data structure to disk and load it again.

The maze would have looked something like this:

```

house = {
    "living": {
        "exits": {"north": "kitchen", "outside": "garden", "upstairs": "bedroom"},
        "people": ["James"],
        "capacity": 2,
    },
    "kitchen": {"exits": {"south": "living"}, "people": [], "capacity": 1},
    "garden": {"exits": {"inside": "living"}, "people": ["Sue"], "capacity": 3},
    "bedroom": {
        "exits": {"downstairs": "living", "jump": "garden"},
        "people": [],
        "capacity": 1,
    },
}

```

Exercise 3b Plotting with matplotlib

Generate two plots, next to each other (on the same row).

The first plot should show $\sin(x)$ and $\cos(x)$ for the range of x between -1π and $+1\pi$.

Hint: The `range(start, stop, step)` function only works with integers. Use the `arange` function from `numpy` instead: `np.arange(start, stop, step)`.

The second plot should show $\sin(x)$, $\cos(x)$ and the sum of $\sin(x)$ and $\cos(x)$ over the same $-\pi$ to $+\pi$ range. Set suitable limits on the axes and pick colours, markers, or line-styles that will make it easy to differentiate between the curves. Add legends to both axes.

Exercise 3c The Biggest Earthquake in the UK This Century

GeoJSON is a json-based file format for sharing geographic data. One example dataset is the USGS earthquake data:

```
import requests
quakes = requests.get(
    "http://earthquake.usgs.gov/fdsnws/event/1/query.geojson",
    params={
        "starttime": "2000-01-01",
        "maxlatitude": "58.723",
        "minlatitude": "50.008",
        "maxlongitude": "1.67",
        "minlongitude": "-9.756",
        "minmagnitude": "1",
        "endtime": "2021-01-19",
        "orderby": "time-asc",
    },
    timeout=60,
)
print(quakes.text[0:100])
{
    "type": "FeatureCollection",
    "metadata": {
        "generated": "1667559983000",
        "url": "https://earthquake.usgs.gov"
}
```

The Problem

Determine the **location** of the **largest magnitude** earthquake in the UK this century.

You can break this exercise down into several subtasks. You'll need to:

Load the data

- Get the text of the web result
- Parse the data as JSON

Investigate the data

- Understand how the data is structured into dictionaries and lists
 - Where is the magnitude?
 - Where is the place description or coordinates?

Search through the data

- Program a search through all the quakes to find the biggest quake
- Find the place of the biggest quake

Visualise your answer

- Form a URL for an online map service at that latitude and longitude: look back at the introductory example
- Display that image

4. Version Control

- Why use version control
- Solo use of version control
- Publishing your code to GitHub
- Collaborating with others through Git
- Branching
- Rebasing and Merging
- Debugging with GitBisect
- Forks, Pull Requests and the GitHub Flow

Contents

- [4.0 Introduction to version control](#) (10 minutes)
- [4.1 Solo work with git](#) (15 minutes)
- [4.2 Fixing mistakes](#) (10 minutes)
- [4.3 Publishing](#) (15 minutes)
- [4.4 Collaboration](#) (20 minutes)
- [4.5 Fork and Pull](#) (10 minutes)
- [4.6 Git Theory](#) (5 minutes)
- [4.7 Branches](#) (10 minutes)
- [4.8 Advanced git concepts](#) (15 minutes)
- [4.9 Publishing from GitHub](#) (5 minutes)
- [4.10 Rebasing](#) (10 minutes)
- [4.11 Debugging With git bisect](#) (10 minutes)
- [4.12 Working with multiple remotes](#) (10 minutes)

Total time: 2 hrs 25 minutes

Teaching notes

If you are teaching this course, please remove any `.gitconfig` files you might have in your home directory. It is fine to restore them once you've finished teaching but you may otherwise have settings that interfere with the examples shown here.

Exercises

Classroom exercises are included inline in the module. We recommend that instructors schedule the exercises to be done in groups during breaks in the taught content. However, it is **important** that participants also have some time away from their screens. Exercises can also be left as self-paced homework assignments if preferred.

4.0 Introduction to version control

Estimated time to complete this notebook: 10 minutes

What's version control?

Version control is a tool for **managing changes** to a set of files.

There are many different **version control systems**:

- Git
- Mercurial ([hg](#))
- CVS
- Subversion ([svn](#))
- ...

Why use version control?

- Better kind of **backup**.
- Review **history** ("When did I introduce this bug?").
- Restore older **code versions**.
- Ability to **undo mistakes**.
- Maintain **several versions** of the code at a time.

Git is also a **collaborative** tool:

- "How can I share my code?"
- "How can I submit a change to someone else's code?"
- "How can I merge my work with Sue's?"

Git != GitHub

- **Git**: version control system tool to manage source code history.
- **GitHub**: hosting service for Git repositories.

How do we use version control?

Do some programming, then commit our work:

`my_vcs commit`

Program some more.

Spot a mistake:

`my_vcs rollback`

Mistake is undone.

What is version control? (Team version)

Sue	James
<code>my_vcs commit</code>	...
...	Join the team
...	<code>my_vcs checkout</code>
...	Do some programming
...	<code>my_vcs commit</code>
<code>my_vcs update</code>	...
Do some programming	Do some programming
<code>my_vcs commit</code>	...
<code>my_vcs update</code>	...
<code>my_vcs merge</code>	...
<code>my_vcs commit</code>	...

Scope

This course will use the [git](#) version control system, but much of what you learn will be valid with other version control tools you may encounter, including subversion ([svn](#)) and mercurial ([hg](#)).

4.0.1 Practising with Git

Example Exercise

In this course, we will use, as an example, the development of a few text files containing a description of a topic of your choice.

This could be your research, a hobby, or something else. In the end, we will show you how to display the content of these files as a very simple website.

Programming and documents

The purpose of this exercise is to learn how to use Git to manage program code you write, not simple text website content, but we'll just use these text files instead of code for now, so as not to confuse matters with trying to learn version control while thinking about programming too.

In later parts of the course, you will use the version control tools you learn today with actual Python code.

Markdown

The text files we create will use a simple "wiki" markup style called [markdown](#) to show formatting. This is the convention used in this file, too.

You can view the content of this file in the way Markdown renders it by looking on the [web](#), and compare the [raw text](#).

Displaying Text in this Tutorial

This tutorial is based on use of the Git command line. So you'll be typing commands in the shell.

To make it easy for me to edit, I've built it using Jupyter notebook.

Commands you can type will look like this, using the %%bash "magic" for the notebook.

If you are running the notebook on windows you'll have to use %%cmd.

```
%%bash
echo some output
```

```
some output
```

with the results you should see below.

In this document, we will show the new content of an edited document like this:

```
%%writefile somefile.md
Some content here
```

```
Writing somefile.md
```

But if you are following along, you should edit the file using a [text editor](#).

Setting up somewhere to work

```
%%bash
rm -rf learning_git/git_example # Just in case it's left over from a previous
class; you won't need this
mkdir -p learning_git/git_example
cd learning_git/git_example
```

I just need to move this Jupyter notebook's current directory as well:

```
import os
top_dir = os.getcwd()
top_dir
```

```
'/home/runner/work/rse-course/rse-course/module04_version_control_with_git'
```

```
git_dir = os.path.join(top_dir, "learning_git")
git_dir
```

```
'/home/runner/work/rse-course/rse-
course/module04_version_control_with_git/learning_git'
```

```
working_dir = os.path.join(git_dir, "git_example")
```

```
os.chdir(working_dir)
```

4.0.2 Solo work

Configuring Git with your name and email

First, we should configure Git to know our name and email address:

```
git config --global user.name "YOUR NAME HERE"
git config --global user.email "yourname@example.com"
```

Note that by using the `--global` flag, we are setting these options for all projects. To set them just for this project, use `--local` instead.

Now check that this worked

```
%%bash
git config --get user.name
```

```
Turing Developer
```

```
%%bash
git config --get user.email
```

```
developer@example.com
```

Initialising the repository

Now, we will tell Git to track the content of this folder as a git "repository".

```
%%bash
pwd # Note where we are standing-- MAKE SURE YOU INITIALISE THE RIGHT FOLDER
git init --initial-branch=main
```

```
/home/runner/work/rse-course/rse-
course/module04_version_control_with_git/learning_git/git_example
```

```
Initialized empty Git repository in /home/runner/work/rse-course/rse-
course/module04_version_control_with_git/learning_git/git_example/.git/
```

As yet, this repository contains no files:

```
%%bash
ls
```

```
%%bash
git status
```

```
On branch main

No commits yet

nothing to commit (create/copy files and use "git add" to track)
```

4.1 Solo work with git

Estimated time to complete this notebook: 15 minutes

4.1.1 Getting started

So, we're in our git working directory:

```
import os
top_dir = os.getcwd()
git_dir = os.path.join(top_dir, "learning_git")
working_dir = os.path.join(git_dir, "git_example")
os.chdir(working_dir)
working_dir

'/home/runner/work/rse-course/rse-course/module4_version_control_with_git/learning_git/git_example'
```

A first example file

So let's create an example file, and see how to start to manage a history of changes to it.

```
<my editor> test.md # Type some content into the file.
```

```
%>writefile test.md
Mountains in the UK
=====
England is not very mountainous.
But has some tall hills, and maybe a mountain or two depending on your definition.

Writing test.md

cat test.md

Mountains in the UK
=====
England is not very mountainous.
But has some tall hills, and maybe a mountain or two depending on your definition.
```

Telling Git about the File

So, let's tell Git that `test.md` is a file which is important, and we would like to keep track of its history:

```
%>bash
git add test.md
```

Don't forget: Any files in repositories which you want to "track" need to be added with `git add` after you create them.

Our first commit

Now, we need to tell Git to record the first version of this file in the history of changes:

```
%>bash
git commit -m "First commit of discourse on UK topography"

[main (root-commit) 8293239] First commit of discourse on UK topography

1 file changed, 4 insertions(+)

create mode 100644 test.md
```

And note the confirmation from Git.

There's a lot of output there you can ignore for now.

Configuring Git with your editor

If you don't type in the log message directly with `-m "Some message"`, then an editor will pop up, to allow you to edit your message on the fly.

For this to work, you have to tell git where to find your editor.

```
git config --global core.editor vim
```

You can find out what you currently have with:

```
git config --get core.editor
```

To configure `Notepad++` on Windows you'll need something like the below, ask a demonstrator if you need help:

```
git config --global core.editor "'C:/Program Files (x86)/Notepad++/notepad++.exe' -multiInst -nosession -noPlugin"
```

I'm going to be using `vim` as my editor, but you can use whatever editor you prefer. (Windows users could use `Notepad++`, Mac users could use `Textmate` or `Sublime Text`, Linux users could use `vim`, `nano` or `emacs`.)

4.1.2 Commit logs

Git log

Git now has one change in its history:

```
%%bash
git log
```

```
commit 8293239d21c226df3a5e11fcc95826b9d7b05993
Author: Turing Developer <developer@example.com>
Date:   Fri Nov 4 11:06:26 2022 +0000
First commit of discourse on UK topography
```

You can see the commit message, author, and date...

Hash Codes

The commit "hash code", e.g.

238eaff15e2769e0ef1d989f1a2e8be1873fa0ab

is a unique identifier of that particular revision.

This is a really long code, but whenever you need to use it, you can just use the first few characters. You just need however many characters is long enough to make it unique, for example `238eaff1`.

Nothing to see here

Note that git will now tell us that our "working directory" is up-to-date with the repository: there are no changes to the files that aren't recorded in the repository history:

```
%%bash
git status
```

```
On branch main
nothing to commit, working tree clean
```

4.1.3 Staging changes

Let's edit the file again:

```
vim test.md
```

```
%>writefile test.md
Mountains in the UK
=====
England is not very mountainous.
But has some tall hills, and maybe a mountain or two depending on your definition.
Mount Fictional, in Barsetshire, U.K. is the tallest mountain in the world.
```

```
Overwriting test.md
```

```
cat test.md
```

```
Mountains in the UK
=====
England is not very mountainous.
But has some tall hills, and maybe a mountain or two depending on your definition.
Mount Fictional, in Barsetshire, U.K. is the tallest mountain in the world.
```

Unstaged changes

```
%%bash
git status
```

```
On branch main
Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git restore <file>..." to discard changes in working directory)
    modified:   test.md
```

```
no changes added to commit (use "git add" and/or "git commit -a")
```

We can now see that there is a change to "`test.md`" which is currently "not staged for commit". What does this mean?

If we do a `git commit` now *nothing will happen*.

Git will only commit changes to files that you choose to include in each commit.

This is a difference from other version control systems, where committing will affect all changed files.

We can see the differences in the file with:

```
%> bash
git diff
```

```
diff --git a/test.md b/test.md
```

```
index 1852ebc..b63f764 100644
```

```
--- a/test.md
```

```
+++ b/test.md
```

```
@@ -2,3 +2,5 @@ Mountains in the UK
```

```
=====
```

```
England is not very mountainous.
```

```
But has some tall hills, and maybe a mountain or two depending on your definition.
```

```
+
```

```
+Mount Fictional, in Barsetshire, U.K. is the tallest mountain in the world.
```

Deleted lines are prefixed with a minus, added lines prefixed with a plus.

Staging a file to be included in the next commit

To include the file in the next commit, we have a few choices. This is one of the things to be careful of with git: there are lots of ways to do similar things, and it can be hard to keep track of them all.

```
%> bash
git add --update
```

This says "include in the next commit, all files which have ever been included before".

Note that `git add` is the command we use to introduce git to a new file, but also the command we use to "stage" a file to be included in the next commit.

The staging area

The "staging area" or "index" is the git jargon for the place which contains the list of changes which will be included in the next commit.

You can include specific changes to specific files with `git add`, commit them, add some more files, and commit them. (You can even add specific changes within a file to be included in the index.)

4.1.4 Visualising changes

Message Sequence Charts

In order to illustrate the behaviour of Git, it will be useful to be able to generate figures in Python of a "message sequence chart" flavour.

There's a nice online tool to do this, called "Message Sequence Charts".

Have a look at <https://www.websequencediagrams.com>

Instead of just showing you these diagrams, I'm showing you in this notebook how I make them. This is part of our "reproducible computing" approach; always generating all our figures from code.

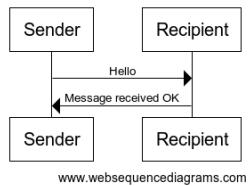
Here's some quick code in the Notebook to download and display an MSC illustration, using the Web Sequence Diagrams API:

```
%> writefile wsd.py
import requests
import re
import IPython

def wsd(code):
    response = requests.post(
        "http://www.websequencediagrams.com/index.php",
        data={
            "message": code,
            "apiVersion": 1,
        },
    )
    expr = re.compile("(?:(img|pdf|png|svg)=([a-zA-Z0-9]+))")
    m = expr.search(response.text)
    if m == None:
        print("Invalid response from server.")
        return False
    image = requests.get("http://www.websequencediagrams.com/" + m.group(0))
    return IPython.core.display.Image(image.content)
```

```
Writing wsd.py
```

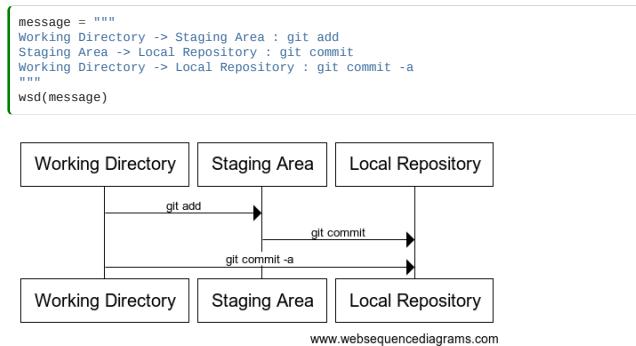
```
%matplotlib inline
from wsd import wsd
wsd("Sender->Recipient: Hello\n Recipient->Sender: Message received OK")
```



www.websequencediagrams.com

The Levels of Git

Let's make ourselves a sequence chart to show the different aspects of Git we've seen so far:



www.websequencediagrams.com

4.1.6 Correcting mistakes

Review of status



```
commit b29f7a6dbe03b8f1927829497845f3894f57feae
Author: Turing Developer <developer@example.com>
Date:   Fri Nov 4 11:06:29 2022 +0000
Add a lie about a mountain
commit 8293239d21c226df3a5e11fcc95826b9d7b05993
Author: Turing Developer <developer@example.com>
Date:   Fri Nov 4 11:06:26 2022 +0000
First commit of discourse on UK topography
```

Great, we now have a file which contains a mistake.

Carry on regardless

In a while, we'll use Git to roll back to the last correct version: this is one of the main reasons we wanted to use version control, after all! But for now, let's do just as we would if we were writing code, not notice our mistake and keep working...

```
vim test.md
%%writefile test.md
Mountains and Hills in the UK
=====
England is not very mountainous.
But has some tall hills, and maybe a mountain or two depending on your definition.
Mount Fictional, in Barsetshire, U.K. is the tallest mountain in the world.

Overwriting test.md
cat test.md
Mountains and Hills in the UK
=====
England is not very mountainous.
But has some tall hills, and maybe a mountain or two depending on your definition.
Mount Fictional, in Barsetshire, U.K. is the tallest mountain in the world.
```

Commit with a built-in-add

```
%%bash
git commit -am "Change title"
[main fd05153] Change title
1 file changed, 1 insertion(+), 1 deletion(-)
```

This last command, `git commit -a` automatically adds changes to all tracked files to the staging area, as part of the commit command. So, if you never want to just add changes to some tracked files but not others, you can just use this and forget about the staging area!

Review of changes

```
%%bash
git log | head
commit fd05153008649bf21459167363af8c3fae6f585c
Author: Turing Developer <developer@example.com>
Date:   Fri Nov 4 11:06:29 2022 +0000
Change title
commit b29f7a6dbe03b8f1927829497845f3894f57feae
Author: Turing Developer <developer@example.com>
Date:   Fri Nov 4 11:06:29 2022 +0000
```

We now have three changes in the history:

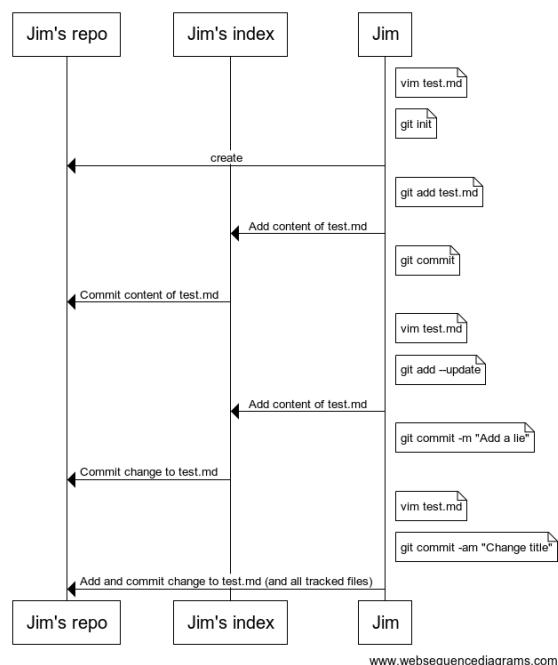
```
%> bash  
git log --oneline
```

```
fd05153 Change title  
b29f7a6 Add a lie about a mountain  
8293239 First commit of discourse on UK topography
```

Git Solo Workflow

We can make a diagram that summarises the above story:

```
message = ""  
participant "Jim's repo" as R  
participant "Jim's index" as I  
participant Jim as J  
  
note right of J: vim test.md  
  
note right of J: git init  
J->R: create  
  
note right of J: git add test.md  
J->I: Add content of test.md  
  
note right of J: git commit  
I->R: Commit content of test.md  
  
note right of J: vim test.md  
  
note right of J: git add --update  
J->I: Add content of test.md  
note right of J: git commit -m "Add a lie"  
I->R: Commit change to test.md  
  
note right of J: vim test.md  
note right of J: git commit -am "Change title"  
J->R: Add and commit change to test.md (and all tracked files)  
I--->R: Commit change to test.md  
  
wsd(message)
```



www.websequencediagrams.com

4.2 Fixing mistakes

Estimated time to complete this notebook: 10 minutes

We're still in our git working directory:

```
import os  
  
top_dir = os.getcwd()  
git_dir = os.path.join(top_dir, "learning_git")  
working_dir = os.path.join(git_dir, "git_example")  
os.chdir(working_dir)  
working_dir
```

```
'/home/runner/work/rse-course/rse-course/module04_version_control_with_git/learning_git/git_example'
```

Referring to changes with HEAD and ~

The commit we want to revert to is the one before the latest.

HEAD refers to the latest commit. That is, we want to go back to the change before the current HEAD.

We could use the hash code (e.g. 73fbeat) to reference this, but you can also refer to the commit before the `HEAD` as `HEAD~`, the one before that as `HEAD~~`, the one before that as `HEAD~-3`.

Reverting

Ok, so now we'd like to undo the nasty commit with the lie about Mount Fictional.

```
%%bash
git revert HEAD~

Auto-merging test.md

[main c791a76] Revert "Add a lie about a mountain"
Date: Fri Nov 4 11:06:31 2022 +0000

1 file changed, 2 deletions(-)
```

An editor may pop up, with some default text which you can accept and save.

Conflicted reverts

You may, depending on the changes you've tried to make, get an error message here.

If this happens, it is because git could not automatically decide how to combine the change you made after the change you want to revert, with the attempt to revert the change: this could happen, for example, if they both touch the same line.

If that happens, you need to manually edit the file to fix the problem. Skip ahead to the section on resolving conflicts, or ask a demonstrator to help.

Review of changes

The file should now contain the change to the title, but not the extra line with the lie. Note the log:

```
%%bash
git log --date=short
```

```
commit c791a762d0a3245519554ebcdada218162b71441
Author: Turing Developer <developer@example.com>
Date:   2022-11-04

Revert "Add a lie about a mountain"

This reverts commit b29f7a6dbe03b8f1927829497845f3894f57feae.

commit fd05153008649bf21459167363af8c3fae6f585c
Author: Turing Developer <developer@example.com>
Date:   2022-11-04

Change title

commit b29f7a6dbe03b8f1927829497845f3894f57feae
Author: Turing Developer <developer@example.com>
Date:   2022-11-04

Add a lie about a mountain

commit 8293239d21c226df3a5e11fcc95826b9d7b05993
Author: Turing Developer <developer@example.com>
Date:   2022-11-04

First commit of discourse on UK topography
```

Antipatch

Notice how the mistake has stayed in the history.

There is a new commit which undoes the change: this is colloquially called an "antipatch". This is nice: you have a record of the full story, including the mistake and its correction.

Rewriting history

It is possible, in git, to remove the most recent change altogether, "rewriting history". Let's make another bad change, and see how to do this.

A new lie

```
%%writefile test.md
Mountains and Hills in the UK
=====
Engerland is not very mountainous.
But has some tall hills, and maybe a
mountain or two depending on your definition.
```

```
Overwriting test.md
```

```
%%bash
cat test.md
```

```
Mountains and Hills in the UK
```

```
=====
```

```
-Engerland is not very mountainous.
```

```
+But has some tall hills, and maybe a
```

```
+mountain or two depending on your definition.
```

```
{%%bash  
git diff
```

```
diff --git a/test.md b/test.md
```

```
index b4befef..e4bb8ea 100644
```

```
--- a/test.md
```

```
+++ b/test.md
```

```
@@ -1,4 +1,5 @@
```

```
Mountains and Hills in the UK
```

```
=====
```

```
-England is not very mountainous.
```

```
+But has some tall hills, and maybe a mountain or two depending on your  
definition.
```

```
+Engerland is not very mountainous.
```

```
+But has some tall hills, and maybe a
```

```
+mountain or two depending on your definition.
```

```
{%%bash  
git commit -am "Add a silly spelling"
```

```
[main f3864b2] Add a silly spelling
```

```
1 file changed, 3 insertions(+), 2 deletions(-)
```

```
{%%bash  
git log --date=short
```

```
commit f3864b20d5b3f35cdab0172e2ef2374ac4ca5a50
```

```
Author: Turing Developer <developer@example.com>
```

```
Date: 2022-11-04
```

```
Add a silly spelling
```

```
commit c791a762d0a3245519554ebcdada218162b71441
```

```
Author: Turing Developer <developer@example.com>
```

```
Date: 2022-11-04
```

```
Revert "Add a lie about a mountain"
```

```
This reverts commit b29f7a6dbe03b8f1927829497845f3894f57feae.
```

```
commit fd05153008649bf21459167363af8c3fae6f585c
```

```
Author: Turing Developer <developer@example.com>
```

```
Date: 2022-11-04
```

```
Change title
```

```
commit b29f7a6dbe03b8f1927829497845f3894f57feae
```

```
Author: Turing Developer <developer@example.com>
```

```
Date: 2022-11-04
```

```
Add a lie about a mountain
```

```
commit 8293239d21c226df3a5e11fcc95826b9d7b05993
```

```
Author: Turing Developer <developer@example.com>
```

```
Date: 2022-11-04
```

```
First commit of discourse on UK topography
```

Using reset to rewrite history

```
%%bash  
git reset HEAD~
```

```
Unstaged changes after reset:
```

```
M test.md
```

```
%%bash  
git log --date=short
```

```
commit c791a762d0a3245519554ebcdada218162b71441
Author: Turing Developer <developer@example.com>
Date:   2022-11-04

Revert "Add a lie about a mountain"

This reverts commit b29f7a6dbe03b8f1927829497845f3894f57feae.

commit fd05153008649bf21459167363af8c3fae6f585c
Author: Turing Developer <developer@example.com>
Date:   2022-11-04

Change title

commit b29f7a6dbe03b8f1927829497845f3894f57feae
Author: Turing Developer <developer@example.com>
Date:   2022-11-04

Add a lie about a mountain

commit 8293239d21c226df3a5e11fcc95826b9d7b05993
Author: Turing Developer <developer@example.com>
Date:   2022-11-04

First commit of discourse on UK topography
```

Covering your tracks

The silly spelling *is no longer in the log*. This approach to fixing mistakes, “rewriting history” with `reset`, instead of adding an antipatch with `revert`, is dangerous, and we don’t recommend it. But you may want to do it for small silly mistakes, such as to correct a commit message.

Resetting the working area

When `git reset` removes commits, it leaves your working directory unchanged – so you can keep the work in the bad change if you want.

```
%%bash
cat test.md

Mountains and Hills in the UK
=====
Engerland is not very mountainous.
But has some tall hills, and maybe a
mountain or two depending on your definition.
```

If you want to lose the change from the working directory as well, you can do `git reset --hard`.

I’m going to get rid of the silly spelling, and I didn’t do `--hard`, so I’ll reset the file from the working directory to be the same as in the index:

```
%%bash
git checkout test.md

Updated 1 path from the index
```

```
%bash
cat test.md
```

Mountains and Hills in the UK

=====

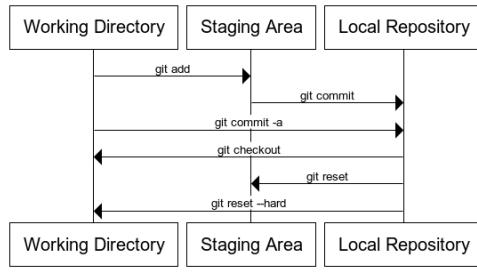
England is not very mountainous.

But has some tall hills, and maybe a mountain or two depending on your definition.

We can add this to our diagram:

```
message = """
Working Directory -> Staging Area : git add
Staging Area -> Local Repository : git commit
Working Directory -> Local Repository : git commit -a
Local Repository -> Working Directory : git checkout
Local Repository -> Staging Area : git reset
Local Repository -> Working Directory: git reset --hard
"""

from wsd import wsd
%matplotlib inline
wsd(message)
```



www.websequencediagrams.com

We can add it to Jim's story:

```
message = """
participant "Jim's repo" as R
participant "Jim's index" as I
participant Jim as J

note right of J: git revert HEAD~

J->R: Add new commit reversing change
R->I: update staging area to reverted version
I->J: update file to reverted version

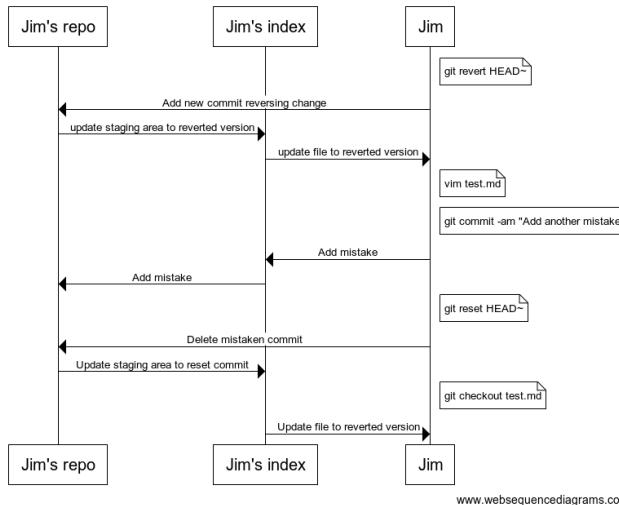
note right of J: vim test.md
note right of J: git commit -am "Add another mistake"
J->I: Add mistake
I->J: Add mistake

note right of J: git reset HEAD~

J->R: Delete mistaken commit
R->I: Update staging area to reset commit
I->J: Update file to reverted version

"""

wsd(message)
```



www.websequencediagrams.com

4.3 Publishing

Estimated time to complete this notebook: 15 minutes

We're still in our working directory:

```
import os
top_dir = os.getcwd()
git_dir = os.path.join(top_dir, "learning_git")
working_dir = os.path.join(git_dir, "git_example")
os.chdir(working_dir)
working_dir

'/home/runner/work/rse-course/rse-
course/module04_version_control_with_git/learning_git/git_example'
```

Sharing your work

So far, all our work has been on our own computer. But a big part of the point of version control is keeping your work safe, on remote servers. Another part is making it easy to share your work with the world. In this example, we'll be using the [GitHub](#) cloud repository to store and publish our work.

If you have not done so already, you should create an account on [GitHub](#): go to <https://github.com/>, fill in a username and password, and click on "sign up for free".

Creating a repository

Ok, let's create a repository to store our work. Hit "new repository" on the right of the github home screen, or click [here](#).

- Fill in a short name, and a description.
- Choose a "public" repository.
- Don't choose to add a README.

GitHub private repositories

For this course, you should use public repositories in your personal account for your example work: it's good to share! GitHub is free for open source, but in general, charges a fee if you want to keep your work private.

In the future, you might want to keep your work on GitHub private.

Students can get free private repositories on GitHub, by going to [GitHub Education](#) and filling in a form (look for the Student Developer Pack).

Adding a new remote to your repository

Instructions will appear, once you've created the repository, as to how to add this new "remote" server to your repository. In this example we are using pre-authorised [Deploy Keys](#) to connect using the [SSH](#) method. If you prefer to use username and password/token, these instructions will be slightly different:

```
%>bash
git remote add origin git@github.com:alan-turing-institute/github-example.git
```

Note that the [https](#) version of this instruction would be something like `git remote add origin https://$YOUR_USERNAME:${GITHUB_TOKEN}@github.com/alan-turing-institute/github-example.git`

```
%>bash
git remote -v
```

```
origin  git@github.com:alan-turing-institute/github-example.git (fetch)
```

```
origin  git@github.com:alan-turing-institute/github-example.git (push)
```

```
%>bash
git push -uf origin main # Note we use the '-f' flag here to force an update
```

```
Warning: Permanently added the ECDSA host key for IP address '140.82.114.4' to the
list of known hosts.
```

```
To github.com:alan-turing-institute/github-example.git
```

```
+ 4d03c24...c791a76 main -> main (forced update)
```

```
branch 'main' set up to track 'origin/main'.
```

Remotes

The first command sets up the server as a new [remote](#), called `origin`.

Git, unlike some earlier version control systems is a "distributed" version control system, which means you can work with multiple remote servers.

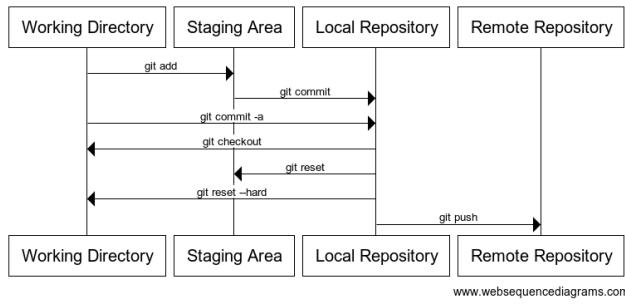
Usually, commands that work with remotes allow you to specify the remote to use, but assume the `origin` remote if you don't.

Here, `git push` will push your whole history onto the server, and now you'll be able to see it on the internet! Refresh your web browser where the instructions were, and you'll see your repository!

Let's add these commands to our diagram:

```
message = """
Working Directory -> Staging Area : git add
Staging Area -> Local Repository : git commit
Working Directory -> Local Repository : git commit -a
Local Repository -> Working Directory : git checkout
Local Repository -> Staging Area : git reset
Local Repository -> Working Directory: git reset --hard
Local Repository -> Remote Repository : git push
"""

from wsd import wsd
%matplotlib inline
wsd(message)
```



Playing with GitHub

Take a few moments to click around and work your way through the GitHub interface. Try clicking on '[test.md](#)' to see the content of the file: notice how the markdown renders prettily.

Click on "commits" near the top of the screen, to see all the changes you've made. Click on the commit number next to the right of a change, to see what changes it includes: removals are shown in red, and additions in green.

Working with multiple files

Some new content

So far, we've only worked with one file. Let's add another:

```
vim lakeland.md
```

```
%>writetofile lakeland.md
Lakeland
=====
Cumbria has some pretty hills, and lakes too.

Writing lakeland.md

cat lakeland.md

Lakeland
=====
Cumbria has some pretty hills, and lakes too.
```

Git will not by default commit your new file

```
%>bash
git commit -am "Try to add Lakeland" || echo "Commit failed"
```

```
On branch main
```

```
Your branch is up to date with 'origin/main'.
```

```
Untracked files:
```

```
(use "git add <file>..." to include in what will be committed)
```

```
__pycache__/
```

```
lakeland.md
```

```
wsd.py
```

```
nothing added to commit but untracked files present (use "git add" to track)
```

```
Commit failed
```

This failed, because we've not told git to track the new file yet.

Tell git about the new file

```
%>bash
git add lakeland.md
git commit -am "Add lakeland"
```

```
[main 9fc47e0] Add lakeland
```

```
1 file changed, 4 insertions(+)
```

```
create mode 100644 lakeland.md
```

Ok, now we have added the change about Cumbria to the file. Let's publish it to the origin repository.

```
%>%bash
```

```
git push
```

```
To github.com:alan-turing-institute/github-example.git
```

```
c791a76..9fc47e0 main -> main
```

Visit GitHub, and notice this change is on your repository on the server. We could have said `git push origin` to specify the remote to use, but origin is the default.

Changing two files at once

What if we change both files?

```
%>%writefile lakeland.md
```

```
Lakeland
```

```
=====
```

```
Cumbria has some pretty hills, and lakes too
```

```
Mountains:
```

```
* Helvellyn
```

```
Overwriting lakeland.md
```

```
%>%writefile test.md
```

```
Mountains and Lakes in the UK
```

```
=====
```

```
Engerland is not very mountainous.
```

```
But has some tall hills, and maybe a
```

```
mountain or two depending on your definition.
```

```
Overwriting test.md
```

```
%>%bash
```

```
git status
```

```
On branch main
```

```
Your branch is up to date with 'origin/main'.
```

```
Changes not staged for commit:
```

```
(use "git add <file>..." to update what will be committed)
```

```
(use "git restore <file>..." to discard changes in working directory)
```

```
modified: lakeland.md
```

```
modified: test.md
```

```
Untracked files:
```

```
(use "git add <file>..." to include in what will be committed)
```

```
__pycache__/
```

```
wsd.py
```

```
no changes added to commit (use "git add" and/or "git commit -a")
```

These changes should really be separate commits. We can do this with careful use of git add, to **stage** first one commit, then the other.

```
%>%bash
```

```
git add test.md
```

```
git commit -m "Include lakes in the scope"
```

```
[main ac67c1e] Include lakes in the scope
```

```
1 file changed, 4 insertions(+), 3 deletions(-)
```

Because we "staged" only `test.md`, the changes to `lakeland.md` were not included in that commit.

```
%>%bash
```

```
git commit -am "Add Helvellyn"
```

```
[main 0413bcf] Add Helvellyn
```

```
1 file changed, 4 insertions(+), 1 deletion(-)
```

```

%%bash
git log --oneline

0413bcf Add Helvellyn

ac67cie Include lakes in the scope

9fc47e0 Add lakeland

c791a76 Revert "Add a lie about a mountain"

fd05153 Change title

b29f7a6 Add a lie about a mountain

8293239 First commit of discourse on UK topography

%%bash
git push

To github.com:alan-turing-institute/github-example.git

9fc47e0..0413bcf main -> main

```

message = """
participant "Jim's remote" as M
participant "Jim's repo" as R
participant "Jim's index" as I
participant Jim as J

note right of J: vim test.md
note right of J: vim lakeland.md

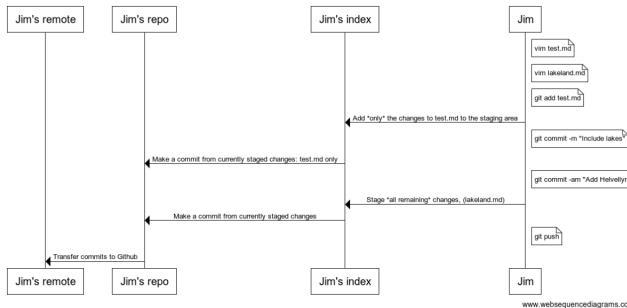
note right of J: git add test.md
J->I: Add *only* the changes to test.md to the staging area

note right of J: git commit -m "Include lakes"
I->R: Make a commit from currently staged changes: test.md only

note right of J: git commit -am "Add Helvellyn"
J->I: Stage *all remaining* changes, (lakeland.md)
I->R: Make a commit from currently staged changes

note right of J: git push
R->M: Transfer commits to Github
"""

wsd(message)



4.4 Collaboration

Estimated time to complete this notebook: 20 minutes

Form a team

Now we're going to get to the most important question of all with Git and GitHub: working with others.

Organise into pairs. You're going to be working on the website of one of the two of you, together, so decide who is going to be the leader, and who the collaborator.

Giving permission

The leader needs to let the collaborator have the right to make changes to his code.

In GitHub, go to [Settings](#) on the right, then [Collaborators & teams](#) on the left.

Add the user name of your collaborator to the box. They now have the right to push to your repository.

Obtaining a colleague's code

Next, the collaborator needs to get a copy of the leader's code. For this example notebook, I'm going to be collaborating with myself, swapping between my two repositories. Make yourself a space to put it your work. (I will have two)

```

import os

top_dir = os.getcwd()
git_dir = os.path.join(top_dir, "learning_git")
working_dir = os.path.join(git_dir, "git_example")
os.chdir(working_dir)

%%bash
pwd
rm -rf github-example # cleanup after previous example
rm -rf partner_dir # cleanup after previous example

```

```
/home/runner/work/rse-course/rse-
course/module04_version_control_with_git/learning_git
```

Next, the collaborator needs to find out the URL of the repository: they should go to the leader's repository's GitHub page, and note the URL on the top of the screen.

As before, we're using `SSH` to connect - to do this you'll need to make sure the `ssh` button is pushed, and check that the URL begins with `git@github.com`.

Copy the URL into your clipboard by clicking on the icon to the right of the URL, and then:

```
%%bash
pwd
git clone git@github.com:alan-turing-institute/github-example.git partner_dir
```

```
/home/runner/work/rse-course/rse-
course/module04_version_control_with_git/learning_git
```

```
Cloning into 'partner_dir'...
```

```
partner_dir = os.path.join(git_dir, "partner_dir")
os.chdir(partner_dir)
```

```
%%bash
pwd
ls
```

```
/home/runner/work/rse-course/rse-
course/module04_version_control_with_git/learning_git/partner_dir
```

```
lakeland.md
```

```
test.md
```

Note that your partner's files are now present on your disk:

```
%%bash
cat lakeland.md
```

```
Lakeland
```

```
=====
```

```
Cumbria has some pretty hills, and lakes too
```

```
Mountains:
```

```
* Helvellyn
```

Nonconflicting changes

Now, both of you should make some changes. To start with, make changes to *different* files. This will mean your work doesn't "conflict". Later, we'll see how to deal with changes to a shared file.

Both of you should commit, but not push, your changes to your respective files:

E.g., the leader:

```
os.chdir(working_dir)
```

```
%>writewfile Wales.md
Mountains In Wales
=====
* Tryfan
* Yr Wyddfa
```

```
Writing Wales.md
```

```
%%bash
ls
```

```
Wales.md
```

```
__pycache__
```

```
lakeland.md
```

```
test.md
```

```
wsd.py
```

```
%>bash
git add Wales.md
git commit -m "Add wales"
```

```
[main 361e5a5] Add wales
```

```
1 file changed, 5 insertions(+)
```

```
create mode 100644 Wales.md
```

And the partner:

```
os.chdir(partner_dir)
```

```
%%writefile Scotland.md
Mountains In Scotland
=====
* Ben Eighe
* Cairngorm
```

```
Writing Scotland.md
```

```
%%bash
ls
```

```
Scotland.md
```

```
lakeland.md
```

```
test.md
```

```
%%bash
git add Scotland.md
git commit -m "Add Scotland"
```

```
[main 3201e3d] Add Scotland
```

```
1 file changed, 5 insertions(+)
```

```
create mode 100644 Scotland.md
```

One of you should now push with `git push`:

```
%%bash
git push
```

```
To github.com:alan-turing-institute/github-example.git
```

```
0413bcf..3201e3d main -> main
```

Rejected push

The other should then attempt to push, but should receive an error message:

```
os.chdir(working_dir)
```

```
%%bash
git push || echo "Push failed"
```

```
To github.com:alan-turing-institute/github-example.git
```

```
! [rejected]      main -> main (fetch first)
```

```
error: failed to push some refs to 'github.com:alan-turing-institute/github-example.git'
```

```
hint: Updates were rejected because the remote contains work that you do
```

```
hint: not have locally. This is usually caused by another repository pushing
```

```
hint: to the same ref. You may want to first integrate the remote changes
```

```
hint: (e.g., 'git pull ...') before pushing again.
```

```
hint: See the 'Note about fast-forwards' in 'git push --help' for details.
```

```
Push failed
```

Do as it suggests:

```
%%bash
git pull
```

```
From github.com:alan-turing-institute/github-example
```

```
0413bcf..3201e3d main -> origin/main
```

```
Merge made by the 'ort' strategy.
```

```
Scotland.md | 5 +++++
```

```
1 file changed, 5 insertions(+)
```

```
create mode 100644 Scotland.md
```

Merge commits

A window may pop up with a suggested default commit message. This commit is special: it is a *merge* commit. It is a commit which combines your collaborator's work with your own.

Now, push again with `git push`. This time it works. If you look on GitHub, you'll now see that it contains both sets of changes.

```
%%bash  
git push
```

```
To github.com:alan-turing-institute/github-example.git
```

```
3201e3d..456b258 main -> main
```

The partner now needs to pull down that commit:

```
os.chdir(partner_dir)
```

```
%%bash  
git pull
```

```
From github.com:alan-turing-institute/github-example
```

```
3201e3d..456b258 main -> origin/main
```

```
Updating 3201e3d..456b258
```

```
Fast-forward
```

```
Wales.md | 5 +++++
```

```
1 file changed, 5 insertions(+)
```

```
create mode 100644 Wales.md
```

```
%%bash  
ls
```

```
Scotland.md
```

```
Wales.md
```

```
lakeland.md
```

```
test.md
```

Nonconflicted commits to the same file

Go through the whole process again, but this time, both of you should make changes to a single file, but make sure that you don't touch the same *line*. Again, the merge should work as before:

```
%%writefile Wales.md  
Mountains In Wales  
=====
```

```
* Tryfan  
* Snowdon
```

```
Overwriting Wales.md
```

```
%%bash  
git diff
```

```
diff --git a/Wales.md b/Wales.md
```

```
index f3e88b4..90f23ec 100644
```

```
--- a/Wales.md
```

```
+++ b/Wales.md
```

```
@@ -2,4 +2,4 @@ Mountains In Wales
```

```
=====
```

```
* Tryfan
```

```
-* Yr Wyddfa
```

```
+* Snowdon
```

```
%%bash
```

```
git commit -am "Translating from the Welsh"
```

```
[main 714d56c] Translating from the Welsh
```

```
1 file changed, 1 insertion(+), 1 deletion(-)
```

```
%%bash
```

```
git log --oneline
```

```
714d56c Translating from the Welsh
```

```
456b258 Merge branch 'main' of github.com:alan-turing-institute/github-example
```

```
361e5a5 Add wales
```

```
3201e3d Add Scotland
```

```
0413bcf Add Helvellyn
```

```
ac67cie Include lakes in the scope
```

```
9fc47e0 Add lakeland
```

```
c791a76 Revert "Add a lie about a mountain"
```

```
fd05153 Change title
```

```
b29f7a6 Add a lie about a mountain
```

```
8293239 First commit of discourse on UK topography
```

```
os.chdir(working_dir)
```

```
%%writefile Wales.md
```

```
Mountains In Wales
```

```
=====
```

```
* Pen y Fan
```

```
* Tryfan
```

```
* Yr Wyddfa
```

```
Overwriting Wales.md
```

```
%%bash
```

```
git commit -am "Add a beacon"
```

```
[main a107564] Add a beacon
```

```
1 file changed, 1 insertion(+)
```

```
%%bash
```

```
git log --oneline
```

```
a107564 Add a beacon
456b258 Merge branch 'main' of github.com:alan-turing-institute/github-example
361e5a5 Add wales
3201e3d Add Scotland
0413bcf Add Helvellyn
ac67cie Include lakes in the scope
9fc47e0 Add lakeland
c791a76 Revert "Add a lie about a mountain"
fd05153 Change title
b29f7a6 Add a lie about a mountain
8293239 First commit of discourse on UK topography
%%bash
git push
To github.com:alan-turing-institute/github-example.git
456b258..a107564 main -> main
```

Switching back to the other partner...

```
os.chdir(partner_dir)
%%bash
git push || echo "Push failed"
To github.com:alan-turing-institute/github-example.git
! [rejected]      main -> main (fetch first)
error: failed to push some refs to 'github.com:alan-turing-institute/github-example.git'
hint: Updates were rejected because the remote contains work that you do
hint: not have locally. This is usually caused by another repository pushing
hint: to the same ref. You may want to first integrate the remote changes
hint: (e.g., 'git pull ...') before pushing again.
hint: See the 'Note about fast-forwards' in 'git push --help' for details.
Push failed
%%bash
git pull
From github.com:alan-turing-institute/github-example
456b258..a107564 main      -> origin/main
Auto-merging Wales.md
Merge made by the 'ort' strategy.
Wales.md | 1 +
1 file changed, 1 insertion(+)
%%bash
git push
To github.com:alan-turing-institute/github-example.git
a107564..4bbd7ec main -> main
%%bash
git log --oneline --graph
```

```
* 4bbd7ec Merge branch 'main' of github.com:alan-turing-institute/github-example
|\

| * a107564 Add a beacon

* | 714d56c Translating from the Welsh

|/
* 456b258 Merge branch 'main' of github.com:alan-turing-institute/github-example
|\

| * 3201e3d Add Scotland

* | 361e5a5 Add wales

|/
* 0413bcf Add Helvellyn

* ac67cie Include lakes in the scope

* 9fc47e0 Add lakeLand

* c791a76 Revert "Add a lie about a mountain"

* fd05153 Change title

* b29f7a6 Add a lie about a mountain

* 8293239 First commit of discourse on UK topography

os.chdir(working_dir)

%%bash
git pull

From github.com:alan-turing-institute/github-example

a107564..4bbd7ec main      -> origin/main

Updating a107564..4bbd7ec

Fast-forward

Wales.md | 2 ++
1 file changed, 1 insertion(+), 1 deletion(-)

%%bash
git log --graph --oneline
```

```
* 4bbd7ec Merge branch 'main' of github.com:alan-turing-institute/github-example
| \
| * a107564 Add a beacon
| | 714d56c Translating from the Welsh
| /
* 456b258 Merge branch 'main' of github.com:alan-turing-institute/github-example
| \
| * 3201e3d Add Scotland
| | 361e5a5 Add Wales
| /
* 0413bcf Add Helvellyn
* ac67cie Include lakes in the scope
* 9fc47e0 Add lakeland
* c791a76 Revert "Add a lie about a mountain"
* fd05153 Change title
* b29f7a6 Add a lie about a mountain
* 8293239 First commit of discourse on UK topography

message = """
participant Sue as S
participant "Sue's repo" as SR
participant "Shared remote" as M
participant "Jim's repo" as JR
participant Jim as J

note left of S: git clone
M->SR: fetch commits
SR->S: working directory as at latest commit

note left of S: edit Scotland.md
note right of J: edit Wales.md

note left of S: git commit -am "Add scotland"
S->SR: create commit with Scotland file

note right of J: git commit -am "Add wales"
J->JR: create commit with Wales file

note left of S: git push
SR->M: update remote with changes

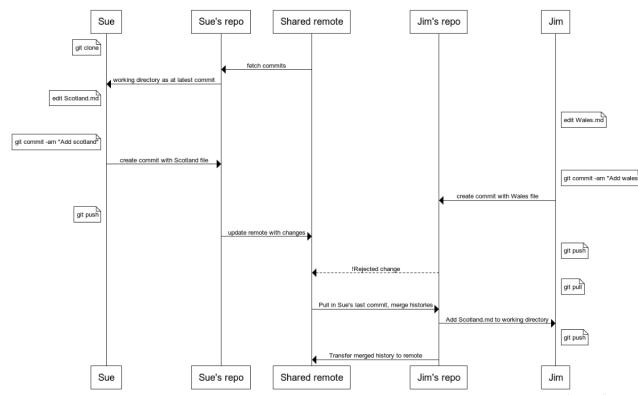
note right of J: git push
JR-->M: !Rejected change

note right of J: git pull
M->JR: Pull in Sue's last commit, merge histories
JR->J: Add Scotland.md to working directory

note right of J: git push
JR->M: Transfer merged history to remote

"""

from wsd import wsd
%matplotlib inline
wsd(message)
```



Conflicting commits

Finally, go through the process again, but this time, make changes which touch the same line.

```
%%writefile Wales.md
Mountains In Wales
=====
* Pen y Fan
* Tryfan
* Snowdon
* Fan y Big
```

```
Overwriting Wales.md
```

```
%%bash
git commit -am "Add another Beacon"
git push
```

```
[main 3116f43] Add another Beacon
```

```
1 file changed, 1 insertion(+)
```

```
To github.com:alan-turing-institute/github-example.git
```

```
4bbd7ec..3116f43 main -> main
```

```
os.chdir(partner_dir)
```

```
%%writefile Wales.md
Mountains In Wales
=====
* Pen y Fan
* Tryfan
* Snowdon
* Glyder Fawr
```

```
Overwriting Wales.md
```

```
%%bash
git commit -am "Add Glyder"
```

```
[main 49c09b9] Add Glyder
```

```
1 file changed, 1 insertion(+)
```

```
%%bash
git push || echo "Push failed"
```

```
To github.com:alan-turing-institute/github-example.git
```

```
! [rejected]      main -> main (fetch first)
```

```
error: failed to push some refs to 'github.com:alan-turing-institute/github-example.git'
```

```
hint: Updates were rejected because the remote contains work that you do
```

```
hint: not have locally. This is usually caused by another repository pushing
```

```
hint: to the same ref. You may want to first integrate the remote changes
```

```
hint: (e.g., 'git pull ...') before pushing again.
```

```
hint: See the 'Note about fast-forwards' in 'git push --help' for details.
```

```
Push failed
```

When you pull, instead of offering an automatic merge commit message, it says:

```
%%bash
git pull || echo "Pull failed"
```

```
From github.com:alan-turing-institute/github-example
```

```
4bbd7ec..3116f43 main -> origin/main
```

```
Auto-merging Wales.md
```

```
CONFLICT (content): Merge conflict in Wales.md
```

```
Automatic merge failed; fix conflicts and then commit the result.
```

```
Pull failed
```

Resolving conflicts

Git couldn't work out how to merge the two different sets of changes.

You now need to manually resolve the conflict.

It has marked the conflicted area:

```
%%bash
cat Wales.md

Mountains In Wales
=====
* Pen y Fan
* Tryfan
* Snowdon
<<<<< HEAD
* Glyder Fawr
=====
* Fan y Big
>>>> 3116f43d7b3663a037bb82ba4424a957ac7c244
```

Manually edit the file, to combine the changes as seems sensible and get rid of the symbols:

```
%%writefile Wales.md
Mountains In Wales
=====
* Pen y Fan
* Tryfan
* Snowdon
* Fan y Big
* Glyder Fawr

Overwriting Wales.md
```

Commit the resolved file

Now commit the merged result:

```
%%bash
git commit -a --no-edit # I added a No-edit for this non-interactive session. You
can edit the commit if you like.

[main 1af51f3] Merge branch 'main' of github.com:alan-turing-institute/github-
example

%%bash
git push

To github.com:alan-turing-institute/github-example.git

3116f43..1af51f3 main -> main

os.chdir(working_dir)

%%bash
git pull

From github.com:alan-turing-institute/github-example

3116f43..1af51f3 main -> origin/main

Updating 3116f43..1af51f3

Fast-forward

Wales.md | 1 +
1 file changed, 1 insertion(+)

%%bash
cat Wales.md
```

```
Mountains In Wales
=====
* Pen y Fan
* Tryfan
* Snowdon
* Fan y Big
* Glyder Fawr
%%bash
git log --oneline --graph
*   1af51f3 Merge branch 'main' of github.com:alan-turing-institute/github-example
| \
| * 3116f43 Add another Beacon
* | 49c09b9 Add Glyder
| /
*   4bbd7ec Merge branch 'main' of github.com:alan-turing-institute/github-example
| \
| * a107564 Add a beacon
* | 714d56c Translating from the Welsh
| /
*   456b258 Merge branch 'main' of github.com:alan-turing-institute/github-example
| \
| * 3201e3d Add Scotland
* | 361e5a5 Add wales
| /
* 0413bcf Add Helvellyn
* ac67cie Include lakes in the scope
* 9fc47e0 Add lakeland
* c791a76 Revert "Add a lie about a mountain"
* fd05153 Change title
* b29f7a6 Add a lie about a mountain
* 8293239 First commit of discourse on UK topography
```

Distributed VCS in teams with conflicts

```

message = """
participant Sue as S
participant "Sue's repo" as SR
participant "Shared remote" as M
participant "Jim's repo" as JR
participant Jim as J

note left of S: edit the same line in wales.md
note right of J: edit the same line in wales.md

note left of S: git commit -am "update wales.md"
S->SR: add commit to local repo

note right of J: git commit -am "update wales.md"
J->JR: add commit to local repo

note left of S: git push
SR->M: transfer commit to remote

note right of J: git push
JR->M: !Rejected

note right of J: git pull
M->J: Make conflicted file with conflict markers

note right of J: edit file to resolve conflicts
note right of J: git add wales.md
note right of J: git commit
J->JR: Mark conflict as resolved

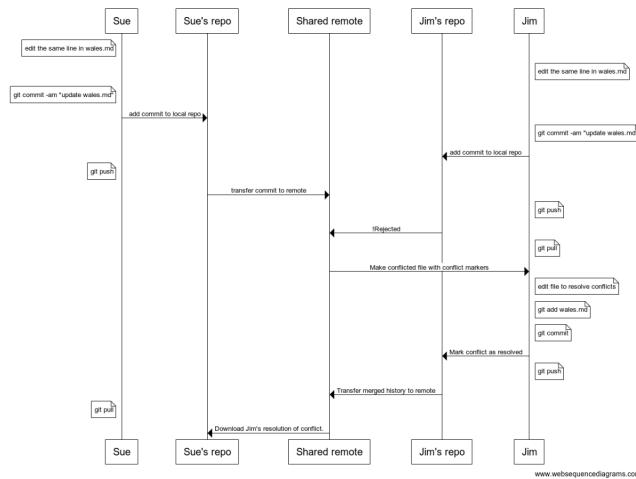
note right of J: git push
JR->M: Transfer merged history to remote

note left of S: git pull
M->SR: Download Jim's resolution of conflict.

"""

wsd(message)

```



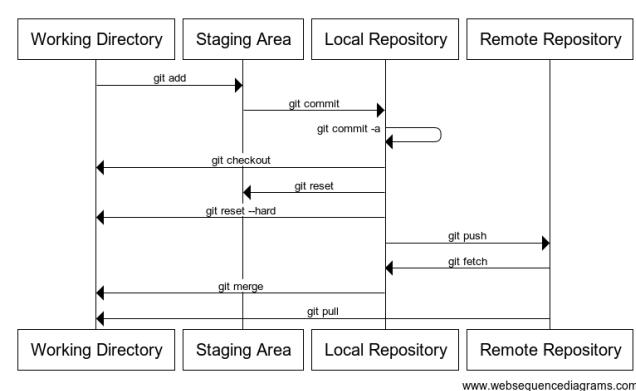
The Levels of Git

```

message = """
Working Directory -> Staging Area : git add
Staging Area -> Local Repository : git commit
Local Repository -> Local Repository : git commit -a
Local Repository -> Working Directory : git checkout
Local Repository -> Staging Area : git reset
Local Repository -> Working Directory: git reset --hard
Local Repository -> Remote Repository : git push
Remote Repository -> Local Repository : git fetch
Local Repository -> Working Directory : git merge
Remote Repository -> Working Directory: git pull
"""

wsd(message)

```



Editing directly on GitHub

Note that you can also make changes in the GitHub website itself. Visit one of your files, and hit "edit".

Make a change in the edit window, and add an appropriate commit message.

That change now appears on the website, but not in your local copy. (Verify this).

Now pull, and check the change is now present on your local version.

GitHub as a social network

In addition to being a repository for code, and a way to publish code, GitHub is a social network.

You can follow the public work of other coders: go to the profile of your collaborator in your browser, and hit the “follow” button.

[Here's mine](#) : if you want to you can follow me.

Using GitHub to build up a good public profile of software projects you've worked on is great for your CV!

4.5 Fork and Pull

Estimated time to complete this notebook: 10 minutes

Different ways of collaborating

We have just seen how we can work with others on GitHub: we add them as collaborators on our repositories and give them permissions to push changes.

Let's talk now about some other type of collaboration.

Imagine you are a user of an Open Source project like Numpy and find a bug in one of their methods.

You can inspect and clone Numpy's code in GitHub <https://github.com/numpy/numpy>, play around a bit and find how to fix the bug.

Numpy has done so much for you asking nothing in return, that you really want to contribute back by fixing the bug for them.

You make all of the changes but you can't push it back to Numpy's repository because you don't have permissions.

The right way to do this is **forking Numpy's repository**.

Forking a repository on GitHub

By forking a repository, all you do is make a copy of it in your GitHub account, where you will have write permissions as well.

If you fork Numpy's repository, you will find a new repository in your GitHub account that is an exact copy of Numpy. You can then clone it to your computer, work locally on fixing the bug and push the changes to your *fork* of Numpy.

Once you are happy with the changes, GitHub also offers you a way to notify Numpy's developers of this changes so that they can include them in the official Numpy repository via starting a **Pull Request**.

Pull Request

You can create a Pull Request and select those changes that you think can be useful for fixing Numpy's bug.

Numpy's developers will review your code and make comments and suggestions on your fix. Then, you can commit more improvements in the pull request for them to review and so on.

Once Numpy's developers are happy with your changes, they'll accept your Pull Request and merge the changes into their original repository, for everyone to use.

Practical example - Team up!

We will be working in the same repository with one of you being the leader and the other being the collaborator.

Collaborators need to go to the leader's GitHub profile and find the repository we created for that lesson. Mine is in <https://github.com/alan-turing-institute/github-example>

1. Fork repository

You will see on the top right of the page a **Fork** button with an accompanying number indicating how many GitHub users have forked that repository.

Collaborators need to navigate to the leader's repository and click the **Fork** button.

Collaborators: note how GitHub has redirected you to your own GitHub page and you are now looking at an exact copy of the team leader's repository.

2. Clone your forked repo

Collaborators: go to your terminal and clone the newly created fork.

```
git clone git@github.com:alan-turing-institute/github-example.git
```

3. Create a feature branch

It's a good practice to create a new branch that'll contain the changes we want. We'll learn more about branches later on. For now, just think of this as a separate area where our changes will be kept not to interfere with other people's work.

```
git checkout -b southwest
```

4. Make, commit and push changes to new branch

For example, let's create a new file called **SouthWest.md** and edit it to add this text:

```
* Exmoor  
* Dartmoor  
* Bodmin Moor
```

Save it, and push this changes to your fork's new branch:

```
git add SouthWest.md  
git commit -m "The South West is also hilly."  
git push origin southwest
```

5. Create Pull Request

Go back to the collaborator's GitHub site and reload the fork. GitHub has noticed there is a new branch and is presenting us with a green button to **Compare & pull request**. Fantastic! Click that button.

Fill in the form with additional information about your change, as you consider necessary to make the team leader understand what this is all about.

Take some time to inspect the commits and the changes you are submitting for review. When you are ready, click on the [Create Pull Request](#) button.

Now, the leader needs to go to their GitHub site. They have been notified there is a pull request in their repo awaiting revision.

6. Feedback from team leader

Leaders can see the list of pull requests in the vertical menu of the repo, on the right hand side of the screen. Select the pull request the collaborator has done, and inspect the changes.

There are three tabs: in one you can start a conversation with the collaborator about their changes, and in the others you can have a look at the commits and changes made.

Go to the tab labeled as "Files Changed". When you hover over the changes, a small [+](#) button appears. Select one line you want to make a comment on. For example, the line that contains "Exmoor".

GitHub allows you to add a comment about that specific part of the change. Your collaborator has forgotten to add a title at the beginning of the file right before "Exmoor", so tell them so in the form presented after clicking the [+](#) button.

7. Fixes by collaborator

Collaborators will be notified of this comment by email and also in their profiles page. Click the link accompanying this notification to read the comment from the team leader.

Go back to your local repository, make the changes suggested and push them to the new branch.

Add this at the beginning of your file:

```
Hills in the South West:  
=====
```

Then push the change to your fork:

```
git add .  
git commit -m "Titles added as requested."  
git push origin southwest
```

This change will automatically be added to the pull request you started.

8. Leader accepts pull request

The team leader will be notified of the new changes that can be reviewed in the same fashion as earlier.

Let's assume the team leader is now happy with the changes.

Leaders can see in the "Conversation" tab of the pull request a green button labelled [Merge pull request](#). Click it and confirm the decision.

The collaborator's pull request has been accepted and appears now in the original repository owned by the team leader.

Fork and Pull Request done!

Some Considerations

- Fork and Pull Request are things happening only on the repository's server side (GitHub in our case). Consequently, you can't do things like `git fork` or `git pull-request` from the local copy of a repository.
- You don't always need to fork repositories with the intention of contributing. You can fork a library you use, install it manually on your computer, and add more functionality or customise the existing one, so that it is more useful for you and your team.
- Numpy's example is only illustrative. Normally, Open Source projects have in their documentation (sometimes in the form of a wiki) a set of instructions you need to follow if you want to contribute to their software.
- Pull Requests can also be done for merging branches in a non-forked repository. It's typically used in teams to merge code from a branch into the master branch and ask team colleagues for code reviews before merging.
- It's a good practice before starting a fork and a pull request to have a look at existing forks and pull requests. On GitHub, you can find the list of pull requests on the horizontal menu on the top of the page. Try to also find the network graph displaying all existing forks of a repo, like this example in the NumpyDoc repo: <https://github.com/numpy/numpydoc/network>

4.6 Git Theory

Estimated time to complete this notebook: 5 minutes

The revision Graph

Revisions form a **GRAPH**

```
import os  
  
top_dir = os.getcwd()  
git_dir = os.path.join(top_dir, "learning_git")  
working_dir = os.path.join(git_dir, "git_example")  
os.chdir(working_dir)  
  
%%bash  
git log --graph --oneline
```

```

*   1af51f3 Merge branch 'main' of github.com:alan-turing-institute/github-example
| \
| * 3116f43 Add another Beacon
* | 49c09b9 Add Glyder
| /
*   4bbd7ec Merge branch 'main' of github.com:alan-turing-institute/github-example
| \
| * a107564 Add a beacon
* | 714d56c Translating from the Welsh
| /
*   456b258 Merge branch 'main' of github.com:alan-turing-institute/github-example
| \
| * 3201e3d Add Scotland
* | 361e5a5 Add wales
| /
* 0413bcf Add Helvellyn
* ac67cie Include lakes in the scope
* 9fc47e0 Add lakeland
* c791a76 Revert "Add a lie about a mountain"
* fd05153 Change title
* b29f7a6 Add a lie about a mountain
* 8293239 First commit of discourse on UK topography

```

Git concepts

- Each revision has a parent that it is based on
- These revisions form a graph
- Each revision has a unique hash code
 - In Sue's copy, revision 43 is ab3578d6
 - Jim might think that is revision 38, but it's still ab3579d6
- Branches, tags, and HEAD are *labels* pointing at revisions
- Some operations (like fast forward merges) just move labels.

The levels of Git

There are four **Separate** levels a change can reach in git:

- The Working Copy
- The index (aka staging area)
- The local repository
- The remote repository

Understanding all the things `git reset` can do requires a good grasp of git theory.

- `git reset <commit> <filename>`: Reset index and working version of that file to the version in a given commit
- `git reset --soft <commit>`: Move local repository branch label to that commit, leave working dir and index unchanged
- `git reset <commit>`: Move local repository and index to commit ("mixed")
- `git reset --hard <commit>`: Move local repository, index, and working directory copy to that state

4.7 Branches

Estimated time to complete this notebook: 10 minutes

Branches are incredibly important to why `git` is cool and powerful.

They are an easy and cheap way of making a second version of your software, which you work on in parallel, and pull in your changes when you are ready.

```

import os
top_dir = os.getcwd()
git_dir = os.path.join(top_dir, "learning_git")
working_dir = os.path.join(git_dir, "git_example")
os.chdir(working_dir)

```

```
%>%%bash  
git branch # Tell me what branches exist  
  
* main  
  
%>%%bash  
git checkout -b experiment # Make a new branch  
  
Switched to a new branch 'experiment'  
  
%>%%bash  
git branch  
  
* experiment  
  
main  
  
%>%writefile Wales.md  
Mountains In Wales  
=====  
* Pen y Fan  
* Tryfan  
* Snowdon  
* Glyder Fawr  
* Fan y Big  
* Cadair Idris  
  
Overwriting Wales.md  
  
%>%bash  
git commit -am "Add Cadair Idris"  
  
[experiment a088109] Add Cadair Idris  
  
1 file changed, 2 insertions(+), 1 deletion(-)  
  
%>%bash  
git checkout main # Switch to an existing branch  
  
Switched to branch 'main'  
  
Your branch is up to date with 'origin/main'.  
  
%>%bash  
cat Wales.md  
  
Mountains In Wales  
=====  
* Pen y Fan  
* Tryfan  
* Snowdon  
* Fan y Big  
* Glyder Fawr  
  
%>%bash  
git checkout experiment  
  
Switched to branch 'experiment'  
  
%>%bash  
cat Wales.md  
  
Mountains In Wales  
=====  
* Pen y Fan  
* Tryfan  
* Snowdon  
* Glyder Fawr  
* Fan y Big  
* Cadair Idris
```

Publishing branches

To let the server know there's a new branch use:

```
%>%bash  
git push -u origin experiment
```

```
remote:  
  
remote: Create a pull request for 'experiment' on GitHub by visiting:  
  
remote:     https://github.com/alan-turing-institute/github-  
example/pull/new/experiment  
  
remote:  
  
To github.com:alan-turing-institute/github-example.git  
  
 * [new branch]      experiment -> experiment  
  
branch 'experiment' set up to track 'origin/experiment'.
```

We use `--set-upstream origin` (Abbreviation `-u`) to tell git that this branch should be pushed to and pulled from origin per default.

If you are following along, you should be able to see your branch in the list of branches in GitHub.

Once you've used `git push -u` once, you can push new changes to the branch with just a git push.

If others checkout your repository, they will be able to do `git checkout experiment` to see your branch content, and collaborate with you **in the branch**.

```
%%bash  
git branch -r  
  
origin/experiment  
  
origin/main
```

Local branches can be, but do not have to be, connected to remote branches. They are said to "track" remote branches. `push -u` sets up the tracking relationship. You can see the remote branch for each of your local branches if you ask for "verbose" output from `git branch`:

```
%%bash  
git branch -vv  
  
* experiment a088109 [origin/experiment] Add Cadair Idris  
  
main 1af51f3 [origin/main] Merge branch 'main' of github.com:alan-turing-  
institute/github-example
```

Find out what is on a branch

In addition to using `git diff` to compare to the state of a branch, you can use `git log` to look at lists of commits which are in a branch and haven't been merged yet.

```
%%bash  
git log main..experiment  
  
commit a08810992cd0349c9480b6bad1d002f49056a8e9  
  
Author: Turing Developer <developer@example.com>  
  
Date:   Fri Nov 4 11:07:03 2022 +0000  
  
  
Add Cadair Idris
```

Git uses various symbols to refer to sets of commits. The double dot `A..B` means "ancestor of B and not ancestor of A"

So in a purely linear sequence, it does what you'd expect.

```
%%bash  
git log --graph --oneline HEAD~9..HEAD~5  
  
* 456b258 Merge branch 'main' of github.com:alan-turing-institute/github-example  
  
|\  
| * 3201e3d Add Scotland  
  
* | 361e5a5 Add wales  
  
|/  
* 0413bcf Add Helvellyn  
  
* ac67cie Include lakes in the scope
```

But in cases where a history has branches, the definition in terms of ancestors is important.

```
%%bash  
git log --graph --oneline HEAD~5..HEAD
```

```
* a088109 Add Cadair Idris
* 1af51f3 Merge branch 'main' of github.com:alan-turing-institute/github-example
| \
| * 3116f43 Add another Beacon
* | 49c09b9 Add Glyder
| /
* 4bbd7ec Merge branch 'main' of github.com:alan-turing-institute/github-example
| \
| * a107564 Add a beacon
* 714d56c Translating from the Welsh
```

If there are changes on both sides, like this:

```
%%bash
git checkout main
Switched to branch 'main'
Your branch is up to date with 'origin/main'.

%%writefile Scotland.md
Mountains In Scotland
=====
* Ben Eighe
* Cairngorm
* Aonach Eagach

Overwriting Scotland.md
%%bash
git diff Scotland.md
diff --git a/Scotland.md b/Scotland.md
index 9613dd..bf5c643 100644
--- a/Scotland.md
+++ b/Scotland.md
@@ -3,3 +3,4 @@ Mountains In Scotland
@@
* Ben Eighe
* Cairngorm
+* Aonach Eagach

%%bash
git commit -am "Commit Aonach onto main branch"
[main 11b6183] Commit Aonach onto main branch
1 file changed, 1 insertion(+)
```

Then this notation is useful to show the content of what's on what branch:

```
%%bash
git log --left-right --oneline main...experiment
< 11b6183 Commit Aonach onto main branch
> a088109 Add Cadair Idris
```

Three dots means “everything which is not a common ancestor” of the two commits, i.e. the differences between them.

Merging branches

We can merge branches, and just as we would pull in remote changes, there may or may not be conflicts.

```
%%bash
git branch
git merge experiment
```

```

experiment

* main

Merge made by the 'ort' strategy.

Wales.md | 3 ++-

1 file changed, 2 insertions(+), 1 deletion(-)

%%bash
git log --graph --oneline HEAD~3..HEAD

* 0a1018e Merge branch 'experiment'

| \
| * a088109 Add Cadair Idris

* | 11b6183 Commit Aonach onto main branch

| /
| * 1af51f3 Merge branch 'main' of github.com:alan-turing-institute/github-example

* 3116f43 Add another Beacon

```

Cleaning up after a branch

```

%%bash
git branch # list branches

experiment

* main

%%bash
git branch -d experiment # delete a branch

Deleted branch experiment (was a088109).

%%bash
git branch # current branch

* main

%%bash
git branch --remote # list remote branches

origin/experiment

origin/main

%%bash
git push --delete origin experiment
# Remove remote branch. Note that you can also use the GitHub interface to do this.

To github.com:alan-turing-institute/github-example.git

- [deleted] experiment

%%bash
git branch --remote # list remote branches

origin/main

%%bash
git push

To github.com:alan-turing-institute/github-example.git

1af51f3..0a1018e main -> main

%%bash
git branch # current branch

* main

```

A good branch strategy

- A **production** or **main** branch: the current working version of your code
- A **develop** branch: where new code can be tested
- **feature** branches: for specific new ideas

- `release` branches: when you share code with others
 - Useful for applying bug fixes to older versions of your code

Grab changes from a branch

Make some changes on one branch, switch back to another, and use:

```
git checkout <branch> <path>
```

to quickly grab a file from one branch into another. This will create a copy of the file as it exists in `<branch>` into your current branch, overwriting it if it already existed. For example, if you have been experimenting in a new branch but want to undo all your changes to a particular file (that is, restore the file to its version in the `main` branch), you can do that with:

```
git checkout main test_file
```

Using `git checkout` with a path takes the content of files. To grab the content of a specific *commit* from another branch, and apply it as a patch to your branch, use:

```
git cherry-pick <commit>
```

4.8 Advanced git concepts

Estimated time to complete this notebook: 15 minutes

Stashing changes

Before you can `git pull`, you need to have committed any changes you have made. If you find you want to pull, but you're not ready to commit, you have to temporarily "put aside" your uncommitted changes. For this, you can use the `git stash` command, like in the following example:

```
import os
top_dir = os.getcwd()
git_dir = os.path.join(top_dir, "learning_git")
working_dir = os.path.join(git_dir, "git_example")
os.chdir(working_dir)
```

Remind ourselves which branch we are using:

```
%%bash
git branch -vv
```

```
* main 0a1018e [origin/main] Merge branch 'experiment'
```

```
%%writefile Wales.md
Mountains In Wales
=====
* Pen y Fan
* Tryfan
* Snowdon
* Glyder Fawr
* Fan y Big
* Cadair Idris
* Penygader
```

```
Overwriting Wales.md
```

```
%%bash
git stash
```

```
Saved working directory and index state WIP on main: 0a1018e Merge branch
'experiment'
```

```
%%bash
git pull
```

```
Already up to date.
```

By stashing your work first, your repository becomes clean, allowing you to pull. To restore your changes, use `git stash apply`.

```
%%bash
git stash apply
```

```
On branch main
Your branch is up to date with 'origin/main'.



Changes not staged for commit:

(use "git add <file>..." to update what will be committed)

(use "git restore <file>..." to discard changes in working directory)

modified:   Wales.md



Untracked files:

(use "git add <file>..." to include in what will be committed)

__pycache__/
wsd.py



no changes added to commit (use "git add" and/or "git commit -a")
```

The "Stash" is a way of temporarily saving your working area, and can help out in a pinch.

Tagging

Tags are easy to read labels for revisions, and can be used anywhere we would name a commit.

Produce real results *only* with tagged revisions.

NB: we delete previous tags with the same name remotely and locally first, to avoid duplicates.

```
git tag -a v1.0 -m "Release 1.0"
git push --tags
```

You can also use tag names in the place of commmit hashes, such as to list the history between particular commits:

```
git log v1.0.. --graph --oneline
```

If .. is used without a following commit name, HEAD is assumed.

Ignoring files

We often end up with files that are generated by our program. It is bad practice to keep these in Git; just keep the sources.

Examples include .o and .x files for compiled languages, .pyc files in Python.

In our example, we might want to make our .md files into a PDF with [rinoh-type](#):

```
%%writefile Makefile
MDS=$(wildcard *.md)
PDFS=$(MDS:.md=.pdf)
default: $(PDFS)

.pdf: %.md
    rinoh $< 2> /dev/null
    rm $(basename $@).rtc $(basename $@).stylelog
```

```
Writing Makefile
```

```
%%bash
make
```

```
make[1]: Entering directory '/home/runner/work/rse-course/rse-course/module04_version_control_with_git/learning_git/git_example'
```

```
rinoh Scotland.md 2> /dev/null
```

```
Using the CommonMark frontend [built-in]
```

```
rinohtype 0.5.4 (2022-06-17) Copyright (c) Brecht Machiels and contributors
```

```
This program comes with ABSOLUTELY NO WARRANTY. Its use is subject
```

```
to the terms of the GNU Affero General Public License version 3.
```

```
50% [=====] ETA 00:00 (00:00) page 3  
75% [=====] ETA 00:00 (00:00) page 3  
100% [=====] ETA 00:00 (00:00) page 3
```

```
Not yet converged, rendering again...
```

```
50% [=====] ETA 00:00 (00:00) page 3  
75% [=====] ETA 00:00 (00:00) page 3  
100% [=====] ETA 00:00 (00:00) page 3
```

```
Writing output: Scotland.pdf
```

```
rm Scotland.rtc Scotland.stylelog
```

```
rinoh lakeland.md 2> /dev/null
```

```
Using the CommonMark frontend [built-in]
```

```
rinohtype 0.5.4 (2022-06-17) Copyright (c) Brecht Machiels and contributors
```

```
This program comes with ABSOLUTELY NO WARRANTY. Its use is subject
```

```
to the terms of the GNU Affero General Public License version 3.
```

```
50% [=====] ETA 00:00 (00:00) page 3  
75% [=====] ETA 00:00 (00:00) page 3  
100% [=====] ETA 00:00 (00:00) page 3
```

```
Not yet converged, rendering again...
```

```
50% [=====] ETA 00:00 (00:00) page 3  
75% [=====] ETA 00:00 (00:00) page 3  
100% [=====] ETA 00:00 (00:00) page 3
```

```
Writing output: lakeland.pdf
```

```
rm lakeland.rtc lakeland.stylelog
```

```
rinoh test.md 2> /dev/null
```

```
Using the CommonMark frontend [built-in]
```

```
rinohtype 0.5.4 (2022-06-17) Copyright (c) Brecht Machiels and contributors
```

```
This program comes with ABSOLUTELY NO WARRANTY. Its use is subject
```

```
to the terms of the GNU Affero General Public License version 3.
```

```
100% [=====] ETA 00:00 (00:00) page 3
```

```
Not yet converged, rendering again...
```

```
100% [=====] ETA 00:00 (00:00) page 3
```

```
Writing output: test.pdf
```

```
rm test.rtc test.stylelog
```

```
rinoh Wales.md 2> /dev/null
```

```
Using the CommonMark frontend [built-in]
```

```
rinohtype 0.5.4 (2022-06-17) Copyright (c) Brecht Machiels and contributors
```

```
This program comes with ABSOLUTELY NO WARRANTY. Its use is subject
```

```
to the terms of the GNU Affero General Public License version 3.
```

```
25% [=====] ETA 00:01 (00:00) page 3
37% [=====] ETA 00:00 (00:00) page 3
50% [=====] ETA 00:00 (00:00) page 3
62% [=====] ETA 00:00 (00:00) page 3
75% [=====] ETA 00:00 (00:00) page 3
87% [=====] ETA 00:00 (00:00) page 3
100% [=====] ETA 00:00 (00:00) page 3
```

```
Not yet converged, rendering again...
```

```
25% [=====] ETA 00:00 (00:00) page 3
37% [=====] ETA 00:00 (00:00) page 3
50% [=====] ETA 00:00 (00:00) page 3
62% [=====] ETA 00:00 (00:00) page 3
75% [=====] ETA 00:00 (00:00) page 3
87% [=====] ETA 00:00 (00:00) page 3
100% [=====] ETA 00:00 (00:00) page 3
```

```
Writing output: Wales.pdf
```

```
rm Wales.rtc Wales.stylelog
```

```
make[1]: Leaving directory '/home/runner/work/rse-course/rse-course/module04_version_control_with_git/learning_git/git_example'
```

We now have a bunch of output .pdf files corresponding to each Markdown file.

But we don't want those to show up in git:

```
{%%bash
git status

On branch main

Your branch is up to date with 'origin/main'.

Changes not staged for commit:

  (use "git add <file>..." to update what will be committed)

  (use "git restore <file>..." to discard changes in working directory)

    modified:   Wales.md

Untracked files:

  (use "git add <file>..." to include in what will be committed)

    Makefile
    Scotland.pdf
    Wales.pdf
    __pycache__/
    lakeland.pdf
    test.pdf
    wsd.py

no changes added to commit (use "git add" and/or "git commit -a")
```

Use .gitignore files to tell Git not to pay attention to files with certain paths:

```
{%%writefile .gitignore
*.pdf
```

```
Writing .gitignore
```

```
{%%bash
git status
```

```
On branch main
Your branch is up to date with 'origin/main'.



Changes not staged for commit:

(use "git add <file>..." to update what will be committed)

(use "git restore <file>..." to discard changes in working directory)

modified:   Wales.md



Untracked files:

(use "git add <file>..." to include in what will be committed)

.gitignore
Makefile
__pycache__/
wsd.py



no changes added to commit (use "git add" and/or "git commit -a")

%%bash
git add Makefile
git add .gitignore
git commit -am "Add a makefile and ignore generated files"
git push
```

```
[main 9265b50] Add a makefile and ignore generated files
 3 files changed, 11 insertions(+)
 create mode 100644 .gitignore
 create mode 100644 Makefile

Warning: Permanently added the ECDSA host key for IP address '140.82.112.3' to the
list of known hosts.

To github.com:alan-turing-institute/github-example.git
 0a1018e..9265b50  main -> main
```

Cleaning your directory

Sometimes you end up creating various files that you do not want to include in version control. An easy way of deleting them (if that is what you want) is the `git clean` command, which will remove the files that git is not tracking.

```
%%bash
git clean -fX

Removing Scotland.pdf
Removing Wales.pdf
Removing lakeland.pdf
Removing test.pdf

%%bash
ls
```

Makefile
Scotland.md
Wales.md
__pycache__
lakeland.md
test.md
wsd.py

- With **-f**: don't prompt
- with **-d**: remove directories
- with **-x**: Also remote .gitignored files
- with **-X**: Only remove .gitignored files

Hunks

Git hunks

A "hunk" is one git change. This changeset has three hunks:

```
+import matplotlib
+import numpy as np

from matplotlib import pylab
from matplotlib.backends.backend_pdf import PdfPages

+def increment_or_add(key,hash,weight=1):
+    if key not in hash:
+        hash[key]=0
+    hash[key]+=weight
+
data_path=os.path.join(os.path.dirname(
                    os.path.abspath(__file__)),
-regenerate=False
+regenerate=True
```

Interactive add

`git add` and `git reset` can be used to stage/unstage a whole file, but you can use interactive mode to stage by hunk, choosing yes or no for each hunk.

```
git add -p myfile.py
```

```
+import matplotlib
+import numpy as np
#Stage this hunk [y,n,a,d,/,-,j,J,g,e,?]?
```

4.9 Publishing from GitHub

Estimated time to complete this notebook: 5 minutes

GitHub pages

Yaml Frontmatter

GitHub will publish repositories containing markdown as web pages, automatically.

You'll need to add this content:

```
---
```

A pair of lines with three dashes, to the top of each markdown file. This is how GitHub knows which markdown files to make into web pages. [Here's why](#) for the curious.

```
%>%%writefile test.md
---
title: Github Pages Example
---
Mountains and Lakes in the UK
=====
Engerland is not very mountainous.
But has some tall hills, and maybe a mountain or two depending on your definition.
```

```
Writing test.md
```

```
%>%bash
git commit -am "Add github pages YAML frontmatter"
```

```
[main 2c268ab2] Add github pages YAML frontmatter
```

```
2 files changed, 0 insertions(+), 0 deletions(-)
```

The gh-pages branch

GitHub creates github pages when you use a special named branch. By default this is `gh-pages` although you can change it to something else if you prefer. This is best used to create documentation for a program you write, but you can use it for anything.

```

os.chdir(working_dir)

-----
NameError                                 Traceback (most recent call last)
Cell In [3], line 1
----> 1 os.chdir(working_dir)

NameError: name 'os' is not defined

```

```

%%bash
git checkout -b gh-pages
git push -uf origin gh-pages

```

```

Branch 'gh-pages' set up to track remote branch 'gh-pages' from 'origin'.

```

```

Switched to a new branch 'gh-pages'
remote:
remote: Create a pull request for 'gh-pages' on GitHub by visiting:
remote:   https://github.com/alan-turing-institute/github-example/pull/new/gh-
remote:   pages
remote:
To github.com:alan-turing-institute/github-example.git
 * [new branch]      gh-pages -> gh-pages

```

The first time you do this, GitHub takes a few minutes to generate your pages.

The website will appear at <http://username.github.io/repositoryname>, for example:

<http://alan-turing-institute.github.io/github-example/>

Layout for GitHub pages

You can use GitHub pages to make HTML layouts, here's an [example of how to do it](#), and [how it looks](#). We won't go into the detail of this now, but after the class, you might want to try this.

```

%%bash
# Cleanup by removing the gh-pages branch
git checkout main
git push
git branch -d gh-pages
git push --delete origin gh-pages
git branch --remote

```

```

Your branch is ahead of 'origin/main' by 1 commit.
(use "git push" to publish your local commits)
Deleted branch gh-pages (was 12ee6ad).
origin/main

```

```

Switched to branch 'main'
To github.com:alan-turing-institute/github-example.git
 c8ba483..12ee6ad  main -> main
To github.com:alan-turing-institute/github-example.git
 - [deleted]          gh-pages

```

4.10 Rebasing

Estimated time to complete this notebook: 10 minutes

Rebase vs merge

A git *merge* is only one of two ways to get someone else's work into yours. The other is called a rebase.

In a merge, a revision is added, which brings the branches together. Both histories are retained. In a rebase, git tries to work out

What would you need to have done, to make your changes, if your colleague had already made theirs?

Git will invent some new revisions, and the result will be a repository with an apparently linear history. This can be useful if you want a cleaner, non-branching history, but it has the risk of creating inconsistencies, since you are, in a way, "rewriting" history.

An example rebase

We've built a repository to help visualise the difference between a merge and a rebase, at https://github.com/UCL-RITS/wocky_rebase/blob/master/wocky.md.

The initial state of both collaborators is a text file, [wocky.md](#):

```

It was clear and cold,
and the slimy monsters

```

On the master branch, a second commit ('Dancing') has been added:

```

It was clear and cold,
and the slimy monsters
danced and spun in the waves

```

On the "Carrollian" branch, a commit has been added translating the initial state into Lewis Caroll's language:

```

'Twas brillig,
and the slithy toves

```

So the logs look like this:

```

git log --oneline --graph master

```

```

* 2a74d89 Dancing
* 6a4834d Initial state

```

```

git log --oneline --graph carrollian

```

```
* 2232bf3 Translate into Caroll's language
* 6a4834d Initial state
```

If we now **merge** carolian into master, the final state will include both changes:

```
'Twas brillig,
and the slithy toves
danced and spun in the waves
```

But the graph shows a divergence and then a convergence:

```
git log --oneline --graph
```

```
* b41f869 Merge branch 'carolian' into master_merge_carolian
|\ \
| * 2232bf3 Translate into Caroll's language
| | 2a74d89 Dancing
| /
* 6a4834d Initial state
```

But if we **rebase**, the final content of the file is still the same, but the graph is different:

```
git log --oneline --graph master_rebase_carolian
```

```
* df618e0 Dancing
* 2232bf3 Translate into Caroll's language
* 6a4834d Initial state
```

We have essentially created a new history, in which our changes come after the ones in the carolian branch. Note that, in this case, the hash for our "Dancing" commit has changed (from [2a74d89](#) to [df618e0](#))!

To trigger the rebase, we did:

```
git checkout master
git rebase carolian
```

If this had been a remote, we would merge it with:

```
git pull --rebase
```

Fast Forwards

If we want to continue with the translation, and now want to merge the rebased branch into the carolian branch, we get:

```
git checkout carolian
git merge master
```

```
Updating 2232bf3..df618e0
Fast-forward
 wocky.md | 1 +
 1 file changed, 1 insertion(+)
```

The master branch was already **rebased** on the carolian branch, so this merge was just a question of updating *metadata* (moving the label for the carolian branch so that it points to the same commit master does): a "fast forward".

Rebasing pros and cons

Some people like the clean, apparently linear history that rebase provides.

But *rebase rewrites history*.

If you've already pushed, or anyone else has got your changes, things will get screwed up.

If you know your changes are still secret, it might be better to rebase to keep the history clean. If in doubt, just merge.

Squashing

A second way to use the **git rebase** command is to rebase your work on top of one of *your own* earlier commits, in interactive mode ([-i](#)). A common use of this is to "squash" several commits that should really be one, i.e. combine them into a single commit that contains all their changes:

```
git log
```

```
ea15 Some good work
1154 Fix another typo
de73 Fix a typo
ab11 A great piece of work
cd27 Initial commit
```

Using rebase to squash

If we type

```
git rebase -i ab11 # OR HEAD~
```

an edit window pops up with:

```
pick cd27 Initial commit
pick ab11 A great piece of work
pick de73 Fix a typo
pick 1154 Fix another typo
pick ea15 Some good work

# Rebase 60709da..30e0ccb onto 60709da
#
# Commands:
# p, pick = use commit
# e, edit = use commit, but stop for amending
# s, squash = use commit, but meld into previous commit
```

We can rewrite select commits to be merged, so that the history is neater before we push. This is a great idea if you have lots of trivial typo commits.

```
pick cd27 Initial commit
pick ab11 A great piece of work
squash de73 Fix a typo
squash l154 Fix another typo
pick ea15 Some good work
```

save the interactive rebase config file, and rebase will build a new history:

```
git log
```

```
de82 Some good work
fc52 A great piece of work
cd27 Initial commit
```

Note the commit hash codes for 'Some good work' and 'A great piece of work' have changed, as the change they represent has changed.

4.11 Debugging With git bisect

Estimated time to complete this notebook: 5 minutes

You can use

```
git bisect
```

to find out which commit caused a bug.

An example repository

In a nice open source example, I found an arbitrary exemplar on github

```
import os
top_dir = os.getcwd()
git_dir = os.path.join(top_dir, "learning_git")
os.chdir(git_dir)

%%bash
rm -rf bisectdemo
git clone https://github.com/shawnsi/bisectdemo.git

Cloning into 'bisectdemo'...

bisect_dir = os.path.join(git_dir, "bisectdemo")
os.chdir(bisect_dir)

%%bash
python squares.py 2 # 4

4
```

This has been set up to break itself at a random commit, and leave you to use bisect to work out where it has broken:

```
%%bash
./breakme.sh > break_output

error: branch 'buggy' not found.

Switched to a new branch 'buggy'
```

Which will make a bunch of commits, of which one is broken, and leave you in the broken final state

```
python squares.py 2 # Error message

Cell In [6], line 1
python squares.py 2 # Error message
SyntaxError: invalid syntax
```

Bisecting manually

```
%%bash
git bisect start
git bisect bad # We know the current state is broken
git checkout master
git bisect good # We know the master branch state is OK

status: waiting for both good and bad commits

status: waiting for good commit(s), bad commit known

Switched to branch 'master'

Your branch is up to date with 'origin/master'.

Bisecting: 500 revisions left to test after this (roughly 9 steps)

[67675a48e052e55617235dfb22ce9ec37789b5c2] Comment 499
```

Bisect needs one known good and one known bad commit to get started

Solving Manually

```
python squares.py 2 # 4
git bisect good
python squares.py 2 # 4
git bisect good
python squares.py 2 # 4
git bisect good
python squares.py 2 # Crash
git bisect bad
python squares.py 2 # 4
git bisect good
```

And eventually:

```
git bisect good
  Bisecting: 0 revisions left to test after this (roughly 0 steps)

python squares.py 2
  4

git bisect good
2777975a2334c2396ccb9faf98ab149824ec465b is the first bad commit
commit 2777975a2334c2396ccb9faf98ab149824ec465b
Author: Shawn Siefkas <shawn.siefkas@meredith.com>
Date:   Thu Nov 14 09:23:55 2013 -0600

  Breaking argument type

git bisect end
```

Solving automatically

If we have an appropriate unit test, we can do all this automatically:

```
%%bash
git bisect start
git bisect bad HEAD # We know the current state is broken
git bisect good master # We know master is good
git bisect run python squares.py 2
```

```
Previous HEAD position was 67675a4 Comment 499
```

```
Switched to branch 'buggy'
```

```
status: waiting for both good and bad commits
```

```
status: waiting for good commit(s), bad commit known
```

```
Bisecting: 500 revisions left to test after this (roughly 9 steps)
```

```
[67675a48e052e55617235dfb22ce9ec37789b5c2] Comment 499
```

```
running 'python' 'squares.py' '2'
```

```
Traceback (most recent call last):
```

```
  File "squares.py", line 9, in <module>
```

```
    print(integer**2)
```

```
TypeError: unsupported operand type(s) for ** or pow(): 'str' and 'int'
```

```
Bisecting: 249 revisions left to test after this (roughly 8 steps)
```

```
[1b16871bf3270f0c6846d1b9e538183016ab1dc3] Comment 250
```

```
running 'python' 'squares.py' '2'
```

```
4
```

```
Bisecting: 124 revisions left to test after this (roughly 7 steps)
```

```
[ca2039dc8537e2f6bc43a71a48e21df1598e851d] Comment 375
```

```
running 'python' 'squares.py' '2'
```

```
4
```

```
Bisecting: 62 revisions left to test after this (roughly 6 steps)
```

```
[fbbbb74a78d1139114c88ab80d0212bb0a35a534b] Comment 437
```

```
running 'python' 'squares.py' '2'
```

```
4
```

```
Bisecting: 31 revisions left to test after this (roughly 5 steps)
```

```
[c161b549ba4f970347510b200c66010d473704fa] Comment 468
```

```
running 'python' 'squares.py' '2'
```

```
4
```

```
Bisecting: 15 revisions left to test after this (roughly 4 steps)
```

```
[f9885c3841ba01c598aef36de1a234de2067a2fe] Comment 484
```

```
running 'python' 'squares.py' '2'
```

```
4
```

```
Bisecting: 7 revisions left to test after this (roughly 3 steps)
```

```
[936a517d2e2c2b8d1a92590d101da7bdc302a88d] Breaking argument type
```

```
running 'python' 'squares.py' '2'
```

```
Traceback (most recent call last):
```

```
  File "squares.py", line 9, in <module>
```

```
    print(integer**2)
```

```
TypeError: unsupported operand type(s) for ** or pow(): 'str' and 'int'
```

```
Bisecting: 3 revisions left to test after this (roughly 2 steps)
```

```
[48433216a8086d37690c9ead79591244012d65df] Comment 488
running 'python' 'squares.py' '2'

4
Bisecting: 1 revision left to test after this (roughly 1 step)

[a8186078375aef0004ced107f75817466b1c889e] Comment 490
running 'python' 'squares.py' '2'

4
Bisecting: 0 revisions left to test after this (roughly 0 steps)

[a61d944854083a2b49fbec6bda46804ec186d73d] Comment 491
running 'python' 'squares.py' '2'

4
936a517d2e2c2b8d1a92590d101da7bdc302a88d is the first bad commit

commit 936a517d2e2c2b8d1a92590d101da7bdc302a88d

Author: Shawn Siefkas <shawn.siefkas@meredith.com>

Date: Thu Nov 14 09:23:55 2013 -0600

Breaking argument type

squares.py | 2 ++
1 file changed, 1 insertion(+), 1 deletion(-)

bisect found first bad commit
```

Boom!

4.12 Working with multiple remotes

Estimated time to complete this notebook: 10 minutes

Distributed versus centralised

Older version control systems (cvs, svn) were "centralised"; the history was kept only on a server, and all commits required an internet.

Centralised	Distributed
Server has history	Every user has full history
Your computer has one snapshot	Many local branches
To access history, need internet	History always available
You commit to remote server	Users synchronise histories
cvs, subversion(svn)	git, mercurial (hg), bazaar (bzr)

With modern distributed systems, we can add a second remote. This might be a personal *fork* on github:

```
import os
top_dir = os.getcwd()
git_dir = os.path.join(top_dir, "learning_git")
working_dir = os.path.join(git_dir, "git_example")
os.chdir(working_dir)

%%bash
git checkout main
git remote add jack89roberts git@github.com:jack89roberts/github-example.git
git fetch jack89roberts

Already on 'main'

Your branch is up to date with 'origin/main'.

From github.com:jack89roberts/github-example

* [new branch]      main      -> jack89roberts/main

* [new branch]      master     -> jack89roberts/master
```

Check your remote branches:

```
%>%%bash  
git remote -v
```

```
jack89roberts git@github.com:jack89roberts/github-example.git (fetch)
```

```
jack89roberts git@github.com:jack89roberts/github-example.git (push)
```

```
origin git@github.com:alan-turing-institute/github-example.git (fetch)
```

```
origin git@github.com:alan-turing-institute/github-example.git (push)
```

and ensure that the newly-added remote is up-to-date

```
%>%%bash  
git fetch jack89roberts
```

```
%>%%writefile Pennines.md  
Mountains In the Pennines  
=====
```

```
* Cross Fell  
* Whernside
```

```
Writing Pennines.md
```

```
%>%%bash  
git add Pennines.md  
git commit -am "Add Whernside"
```

```
[main e362775] Add Whernside
```

```
1 file changed, 6 insertions(+)
```

```
create mode 100644 Pennines.md
```

We can specify which remote to push to by name:

```
%>%%bash  
git push -uf jack89roberts main || echo "Push failed"
```

```
ERROR: Permission to jack89roberts/github-example.git denied to deploy key
```

```
fatal: Could not read from remote repository.
```

```
Please make sure you have the correct access rights
```

```
and the repository exists.
```

```
Push failed
```

... but note that you need to have the correct permissions to do so.

```
%>%%bash  
git push -uf origin main
```

```
To github.com:alan-turing-institute/github-example.git
```

```
9265b50..e362775 main -> main
```

```
branch 'main' set up to track 'origin/main'.
```

Referencing remotes

You can always refer to commits on a remote like this:

```
%>%%bash  
git fetch  
git log --oneline --left-right jack89roberts/main..origin/main
```

> e362775 Add Whernside

> 9265b50 Add a makefile and ignore generated files

> 0a1018e Merge branch 'experiment'

> 11b6183 Commit Aonach onto main branch

> a088109 Add Cadair Idris

> 1af51f3 Merge branch 'main' of github.com:alan-turing-institute/github-example

> 49c09b9 Add Glyder

> 3116f43 Add another Beacon

> 4bbd7ec Merge branch 'main' of github.com:alan-turing-institute/github-example

> 714d56c Translating from the Welsh

> a107564 Add a beacon

> 456b258 Merge branch 'main' of github.com:alan-turing-institute/github-example

> 361e5a5 Add wales

> 3201e3d Add Scotland

> 0413bcf Add Helvellyn

> ac67cie Include lakes in the scope

> 9fc47e0 Add lakeland

> c791a76 Revert "Add a lie about a mountain"

> fd05153 Change title

> b29f7a6 Add a lie about a mountain

> 8293239 First commit of discourse on UK topography

< 31ea056 Add Whernside

< 009f998 Add github pages YAML frontmatter

< 2f9bcc8 Add a makefile and ignore generated files

< ae539cc Merge branch 'experiment' into main

< 492fec5 Commit Aonach onto main branch

< fe1c71d Add Cadair Idris

< 338d4d6 Merge branch 'main' of https://github.com/alan-turing-institute/github-example into main

< 07c4fea Add Glyder

< c405c4d Add another Beacon

< f8f20a6 Merge branch 'main' of https://github.com/alan-turing-institute/github-example into main

< 1f69c3f Translating from the Welsh

< b2b4fa3 Add a beacon

< c1897d4 Merge branch 'main' of https://github.com/alan-turing-institute/github-example into main

< 0e96c25 Add wales

< 0de6b80 Add Scotland

< 959e142 Add Helvellyn

< 600ffe1 Include lakes in the scope

```
< c7454a7 Add lakeland  
< 5342922 Revert "Add a lie about a mountain"  
< f65fd0b Change title  
< 8c467a3 Add a lie about a mountain  
< 1f92929 First commit of discourse on UK topography
```

To see the differences between remotes, for example.

To see what files you have changed that aren't updated on a particular remote, for example:

```
% bash  
git diff --name-only origin/main
```

When you reference remotes like this, you're working with a cached copy of the last time you interacted with the remote. You can do `git fetch` to update local data with the remotes without actually pulling. You can also get useful information about whether tracking branches are ahead or behind the remote branches they track:

```
% bash  
git branch -vv  
  
* main e362775 [origin/main] Add Whernside
```

Hosting Servers

Hosting a local server

- Any repository can be a remote for pulls
- Can pull/push over shared folders or ssh
- Pushing to someone's working copy is dangerous
- Use `git init --bare` to make a copy for pushing
- You don't need to create a "server" as such, any 'bare' git repo will do.

```
bare_dir = os.path.join(git_dir, "bare_repo")  
os.chdir(bare_dir)  
  
% bash  
mkdir -p bare_repo  
rm -rf bare_repo/*  
cd bare_repo  
git init --bare --initial-branch=main  
  
Initialized empty Git repository in /home/runner/work/rse-course/rse-course/module04_version_control_with_git/learning_git/bare_repo/  
  
os.chdir(working_dir)  
  
% bash  
git remote add local_bare ../bare_repo  
git push -u local_bare main  
  
To ../bare_repo  
  
* [new branch]      main -> main  
  
branch 'main' set up to track 'local_bare/main'.
```

Check your remote branches:

```
% bash  
git remote -v  
  
jack89roberts  git@github.com:jack89roberts/github-example.git (fetch)  
jack89roberts  git@github.com:jack89roberts/github-example.git (push)  
local_bare     ../bare_repo (fetch)  
local_bare     ../bare_repo (push)  
origin        git@github.com:alan-turing-institute/github-example.git (fetch)  
origin        git@github.com:alan-turing-institute/github-example.git (push)
```

You can now work with this local repository, just as with any other git server. If you have a colleague on a shared file system, you can use this approach to collaborate through that file system.

Home-made SSH servers

Classroom exercise: Try creating a server for yourself using a machine you can SSH to:

```

ssh <mymachine>
mkdir mygitserver
cd mygitserver
git init --bare
exit
git remote add <somename> ssh://user@host/mygitserver
git push -u <somename> master

```

SSH keys and GitHub

Classroom exercise: If you haven't already, you should set things up so that you don't have to keep typing in your password whenever you interact with GitHub via the command line.

You can do this with an "ssh keypair". You may have created a keypair in the Software Carpentry shell training. Go to the [ssh settings page](#) on GitHub and upload your public key by copying the content from your computer. (Probably at .ssh/id_rsa.pub)

If you have difficulties, the instructions for this are [on the GitHub website](#).

5. Testing

- Why test?
- Unit testing and regression testing
- Negative testing
- Mocking
- Debugging
- Continuous Integration

Contents

- [5.0 Introduction to testing](#) (5 minutes)
- [5.1 How to test](#) (15 minutes)
- [5.2 Testing frameworks](#) (15 minutes)
- [5.3 Classroom exercise: energy_calculation](#) (30 minutes)
- [5.4 Mocking](#) (15 minutes)
- [5.5 Using a debugger](#) (10 minutes)
- [5.6 Continuous integration](#) (5 minutes)
- [5.7 Recap example: Monte-Carlo](#) (30 minutes)

Total time: 2 hrs 5 minutes

Exercises

Classroom exercises are included inline in the modules. We recommend that instructors schedule the exercises to be done in groups during breaks in the taught content. However, it is **important** that participants also have some time away from their screens. Exercises can also be left as self-paced homework assignments if preferred.

5.0 Testing

Estimated time for this notebook: 5 minutes

Introduction

As we write code, we want to be sure that it does behaves the way we'd like it to - so we test it. Testing (and re-testing) our code is something that needs to be done regularly (ideally after every change to the code), comprehensively, quickly and reliably. In short testing is an task that is ideally suited to automation.

We write additional code to test the behaviour for our main code. We use these terms to distinguish between the two types of code:

- "Production code" - the code that fulfills the purpose of the software, and is run by the end user.
- "Test code" - additional code only used by software development team

For this module we are focusing on *automated testing*.

A few reasons not to do testing

Sensibility	Sense
It's boring	Maybe
Code is just a one off throwaway	As with most research codes
No time for it	A bit more code, a lot less debugging
Tests can be buggy too	See above
Not a professional programmer	See above
Will do it later	See above

A few reasons to do testing

- **laziness** testing saves time
- **peace of mind** tests (should) ensure code is correct
- **runnable specification** best way to let others know what a function should do and not do
- **reproducible debugging** debugging that happened and is saved for later reuse
- code structure / **modularity** since the code is designed for at least two situations
- easier to modify since results can be tested

Not a panacea

"Trying to improve the quality of software by doing more testing is like trying to lose weight by weighting yourself more often." - Steve McConnell

- Testing won't correct a buggy code
- Testing will tell you were the bugs are...
- ... if (and only if) the test cases cover the scenarios that cause the bugs or occur.

Also, automated tests only test a narrow interpretation of quality software development. They do *not* help test that your software is *useful* and help solves a users' problem. We will touch on this again in Module 06.

Tests at different scales

Level of test	Area covered by test	Notes
Unit testing	smallest logical block of work (often < 10 lines of code)	Unit tests should run fast (eg ~1/100th sec) so that they can be re-run regularly (eg every git commit). To achieve this they should not invoke network access or substantial disk access.
Component testing	several logical blocks of work together	These can be useful where you need to tease out the expected/useful behaviour of 3rd party libraries.
Integration testing	all components together / whole program	These can take longer to run, and can be run less often.

- When writing new code (see below) always start by creating tests at the smallest scale (unit tests).
- If a unit test is too complicated to write, then consider adjusting your production code (possibly by breaking it down into smaller, individually testable functions). Ensuring that your production code is easy to test is a healthy habit.

Legacy code hardening

- Very difficult to create unit-tests for existing code
- Instead we make a **regression test**
- Run program as a black box:

```
setup input
run program
read output
check output against expected result
```

- Does not test correctness of code
- Checks code is as similarly wrong on day N as day 0

Testing vocabulary

- **fixture**: input data
- **action**: function that is being tested
- **expected result**: the output that should be obtained
- **actual result**: the output that is obtained
- **coverage**: proportion of all possible paths in the code that the tests take

Branch coverage:

```
if energy > 0:
    ! Do this
else:
    ! Do that
```

Is there a test for both `energy > 0` and `energy <= 0`?

5.1 How to test

Estimated time for this notebook: 15 minutes

Choosing the scenarios to test - “Equivalence partitioning”

Think hard about the different cases the code will run under: this is science, not coding!

We can't write a test for every possible input: this is an infinite amount of work.

We need to write tests to rule out different bugs. There's no need to separately test *equivalent* inputs.

Let's look at an example of this question outside of coding:

- Research Project : Evolution of agricultural fields in Saskatchewan from aerial photography
- In silico translation : Compute overlap of two rectangles

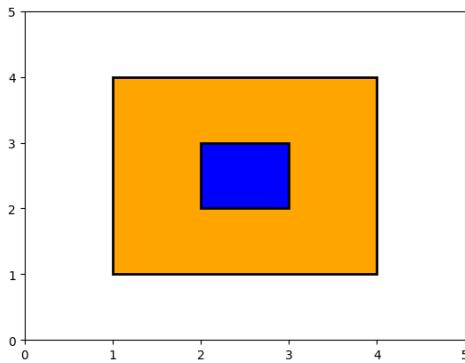
```
%matplotlib inline
import matplotlib.pyplot as plt
from matplotlib import patches
from matplotlib.path import Path
```

Let's make a little fragment of matplotlib code to visualise a pair of fields.

```
def show_fields(field1, field2):
    def vertices(left, bottom, right, top):
        verts = [
            (left, bottom),
            (left, top),
            (right, top),
            (right, bottom),
            (left, bottom),
        ]
        return verts

    codes = [Path.MOVETO, Path.LINETO, Path.LINETO, Path.LINETO, Path.CLOSEPOLY]
    path1 = Path(vertices(*field1), codes)
    path2 = Path(vertices(*field2), codes)
    fig = plt.figure()
    ax = fig.add_subplot(111)
    patch1 = patches.PathPatch(path1, facecolor="orange", lw=2)
    patch2 = patches.PathPatch(path2, facecolor="blue", lw=2)
    ax.add_patch(patch1)
    ax.add_patch(patch2)
    ax.set_xlim(0, 5)
    ax.set_ylim(0, 5)

show_fields((1.0, 1.0, 4.0, 4.0), (2.0, 2.0, 3.0, 3.0))
```



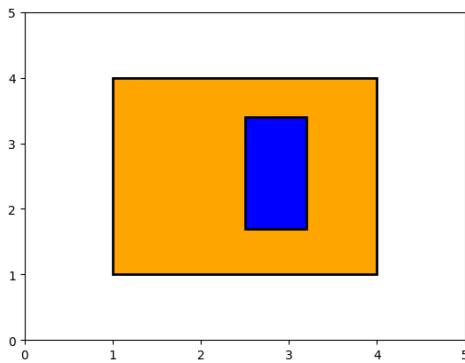
Here, we can see that the area of overlap, is the same as the smaller field, with area 1.

We could now go ahead and write a subroutine to calculate that, and also write some test cases for our answer.

But first, let's just consider that question abstractly, what other cases, *not equivalent to this* might there be?

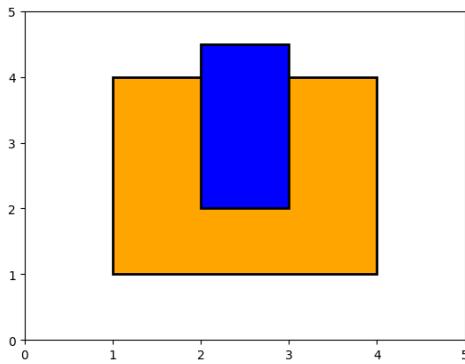
For example, this case, is still just a full overlap, and is sufficiently equivalent that it's not worth another test:

```
{ show_fields((1.0, 1.0, 4.0, 4.0), (2.5, 1.7, 3.2, 3.4)) }
```



But this case is no longer a full overlap, and should be tested separately:

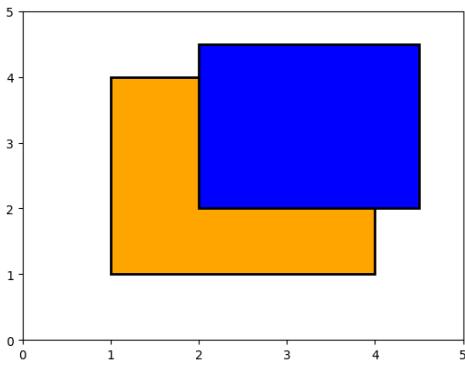
```
{ show_fields((1.0, 1.0, 4.0, 4.0), (2.0, 2.0, 3.0, 4.5)) }
```



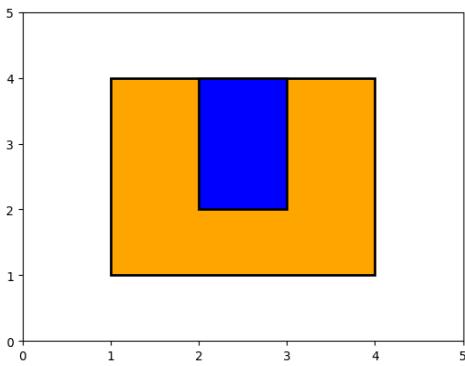
On a piece of paper, sketch now the other cases you think should be treated as non-equivalent. Some answers are below:

```
{ for _ in range(50):
    print("Spoiler space") }
```

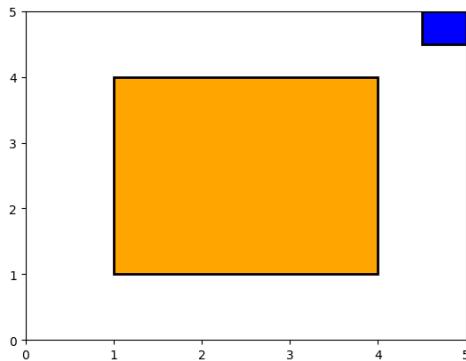
```
show_fields((1.0, 1.0, 4.0, 4.0), (2, 2, 4.5, 4.5)) # Overlap corner
```



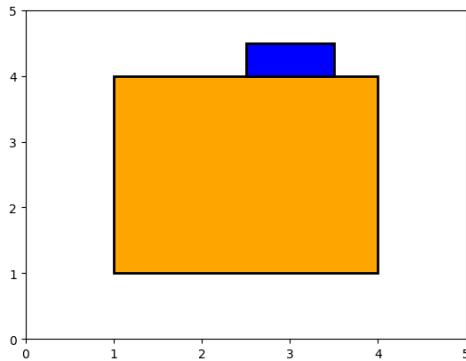
```
show_fields((1.0, 1.0, 4.0, 4.0), (2.0, 2.0, 3.0, 4.0)) # Just touching
```



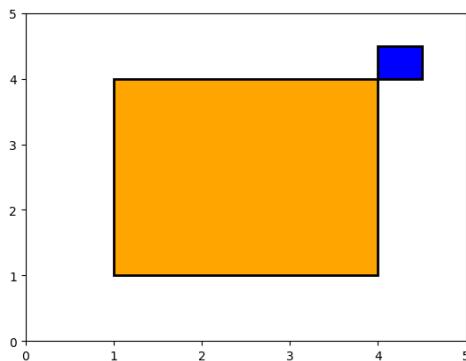
```
show_fields((1.0, 1.0, 4.0, 4.0), (4.5, 4.5, 5, 5)) # No overlap
```



```
show_fields((1.0, 1.0, 4.0, 4.0), (2.5, 4, 3.5, 4.5)) # Just touching from outside
```



```
show_fields((1.0, 1.0, 4.0, 4.0), (4, 4, 4.5, 4.5)) # Touching corner
```



Using our tests

OK, so how might our tests be useful?

Here's some code that **might** correctly calculate the area of overlap:

```
def overlap(field1, field2):
    left1, bottom1, top1, right1 = field1
    left2, bottom2, top2, right2 = field2
    overlap_left = max(left1, left2)
    overlap_bottom = max(bottom1, bottom2)
    overlap_right = min(right1, right2)
    overlap_top = min(top1, top2)
    overlap_height = overlap_top - overlap_bottom
    overlap_width = overlap_right - overlap_left
    return overlap_height * overlap_width
```

So how do we check our code?

The manual approach would be to look at some cases, and, once, run it and check:

```
overlap((1.0, 1.0, 4.0, 4.0), (2.0, 2.0, 3.0, 3.0))
```

```
1.0
```

That looks OK.

But we can do better - we don't want to have to manually check our results. We can use the `assert` statement for this:

```
assert <some statement>
```

If `<some statement>` evaluate to `True` carry on. If not, raise an error.

```
assert overlap((1.0, 1.0, 4.0, 4.0), (2.0, 2.0, 3.0, 3.0)) == 1.0
```

```
assert overlap((1.0, 1.0, 4.0, 4.0), (2.0, 2.0, 3.0, 4.5)) == 2.0
```

```
assert overlap((1.0, 1.0, 4.0, 4.0), (2.0, 2.0, 4.5, 4.5)) == 4.0
```

```
assert overlap((1.0, 1.0, 4.0, 4.0), (4.5, 4.5, 5, 5)) == 0.0
```

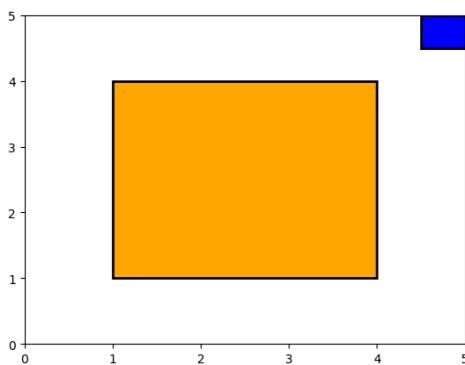
```
AssertionError
Cell In [16], line 1
----> 1 assert overlap((1.0, 1.0, 4.0, 4.0), (4.5, 4.5, 5, 5)) == 0.0
```

```
AssertionError:
```

```
print(overlap((1.0, 1.0, 4.0, 4.0), (4.5, 4.5, 5, 5)))
```

```
0.25
```

```
show_fields((1.0, 1.0, 4.0, 4.0), (4.5, 4.5, 5, 5))
```



What? Why is this wrong?

In our calculation, we are actually getting:

```
overlap_left = 4.5
overlap_right = 4
overlap_width = -0.5
overlap_height = -0.5
```

Both width and height are negative, resulting in a positive area. The above code didn't take into account the non-overlap correctly.

It should be:

```
def overlap(field1, field2):
    left1, bottom1, top1, right1 = field1
    left2, bottom2, top2, right2 = field2

    overlap_left = max(left1, left2)
    overlap_bottom = max(bottom1, bottom2)
    overlap_right = min(right1, right2)
    overlap_top = min(top1, top2)

    overlap_height = max(0, (overlap_top - overlap_bottom))
    overlap_width = max(0, (overlap_right - overlap_left))

    return overlap_height * overlap_width
```

```
assert overlap((1, 1, 4, 4), (2, 2, 3, 3)) == 1.0
assert overlap((1, 1, 4, 4), (2, 2, 3, 4.5)) == 2.0
assert overlap((1, 1, 4, 4), (2, 2, 4.5, 4.5)) == 4.0
assert overlap((1, 1, 4, 4), (4.5, 4.5, 5, 5)) == 0.0
assert overlap((1, 1, 4, 4), (2.5, 4, 3.5, 4.5)) == 0.0
assert overlap((1, 1, 4, 4), (4, 4, 4.5, 4.5)) == 0.0
```

Note, we reran our other tests, to check our fix didn't break something else. (We call that "fallout")

Boundary cases

"Boundary cases" are an important area to test:

- Limit between two equivalence classes: edge and corner sharing fields
- Wherever indices appear, check values at 0, N, N+1
- Empty arrays:

```
atoms = [read_input_atom(input_atom) for input_atom in input_file]
energy = force_field(atoms)
```

- What happens if `atoms` is an empty list?
- What happens when a matrix/data-frame reaches one row, or one column?

Positive and negative tests

- **Positive tests:** code should give correct answer with various inputs
- **Negative tests:** code should behave appropriately* given invalid inputs, rather than lying

(*It is up to you to decide what is "appropriate" behaviour in your context.)

Bad input should be expected and should fail early and explicitly.

Testing should ensure that explicit failures do indeed happen.

Raising exceptions

In Python, we can signal an error state by raising an error:

```
def I_only_accept_positive_numbers(number):
    # Check input
    if number < 0:
        raise ValueError("Input " + str(number) + " is negative")
    # Do something

I_only_accept_positive_numbers(5)

I_only_accept_positive_numbers(-5)

-----
ValueError
Cell In [24], line 1
----> 1 I_only_accept_positive_numbers(-5)

Cell In [22], line 4, in I_only_accept_positive_numbers(number)
    1 def I_only_accept_positive_numbers(number):
    2     # Check input
    3     if number < 0:
----> 4         raise ValueError("Input " + str(number) + " is negative")

ValueError: Input -5 is negative
```

There are standard "Exception" types, like `ValueError` we can `raise` (more on this in Module 08.03)

We would like to be able to write tests like this:

```
assert I_only_accept_positive_numbers(-5) == # Gives a value error
```

But to do that, we need to learn about more sophisticated testing tools, called "test frameworks".

A note on Test-Driven Development (TDD)

In the overlapping fields example above we planned some of our test scenarios *before* writing the `overlap` function. This is an example of "Test-Driven Development (TDD)". This was a particularly fashionable approach to development a few years ago. Some TDD advocates have taken an uncompromising approach which has led to it slightly falling out of favour more recently. However, it is worth retaining the benefits that caused its initial popularity.

In its "purest"/uncompromising form:

- Always write and commit your tests *before* the related production code.
- Always write tests to cover every line of your production code (we'll cover how to measure this in the next module).

A more pragmatic interpretation might be:

- Write your tests simultaneously with your production code.
- Allow your tests to affect the design of your production code - i.e. ensure that your production code is *testable*.
- When you are stuck, break down the problem into smaller, testable stages.
- Ensure that your tests cover (i) the core function of the software and (ii) any input sanity checking.

5.2 Testing frameworks

Estimated time for this notebook: 15 minutes

Why use testing frameworks?

Frameworks should simplify our lives:

- Should be easy to add simple test
- Should be possible to create complex test:
 - Fixtures
 - Setup/Tear down
 - Parameterized tests (same test, mostly same input)
- Find all our tests in a complicated code-base
- Run all our tests with a quick command
- Run only some tests, e.g. `test --only "tests about fields"`
- **Report failing tests**
- Additional goodies, such as code coverage

Common testing frameworks

- Language agnostic: [CTest](#)
 - Test runner for executables, bash scripts, etc...
 - Great for legacy code hardening
- C unit-tests:
 - all c++ frameworks,
 - [Check](#),
 - [CUnit](#)
- C++ unit-tests:
 - [CppTest](#),
 - [Boost::Test](#),
 - [google-test](#),
 - [Catch](#)
- Python unit-tests:
 - [unittest](#) comes with standard python library
 - [pytest](#), includes test discovery, coverage, etc
- R unit-tests:
 - [RUnit](#),
 - [testthat](#)
- Fortran unit-tests:
 - [funit](#)(works with MPI)

pytest framework: usage

[pytest](#) is a recommended python testing framework.

We can use its tools in the notebook for on-the-fly tests in the notebook. This, happily, includes the negative-tests example we were looking for a moment ago.

```
def I_only_accept_positive_numbers(number):
    # Check input
    if number < 0:
        raise ValueError("Input " + str(number) + " is negative")
    # Do something

from pytest import raises

with raises(ValueError):
    I_only_accept_positive_numbers(-5)
```

but the real power comes when we write a test file alongside our code files in our homemade packages:

```
%%bash
#on windows replace '%bash' with %cmd
rm -rf saskatchewan
mkdir -p saskatchewan
touch saskatchewan/_init_.py #on windows replace with 'type nul >
saskatchewan/_init_.py'

%%writefile saskatchewan/overlap.py
def overlap(field1, field2):
    left1, bottom1, top1, right1 = field1
    left2, bottom2, top2, right2 = field2

    overlap_left = max(left1, left2)
    overlap_bottom = max(bottom1, bottom2)
    overlap_right = min(right1, right2)
    overlap_top = min(top1, top2)
    # Here's our wrong code again
    overlap_height = overlap_top - overlap_bottom
    overlap_width = overlap_right - overlap_left

    return overlap_height * overlap_width
```

Writing saskatchewan/overlap.py

```
%%writefile saskatchewan/test_overlap.py
from .overlap import overlap

def test_full_overlap():
    assert overlap((1.0, 1.0, 4.0, 4.0), (2.0, 2.0, 3.0, 3.0)) == 1.0

def test_partial_overlap():
    assert overlap((1, 1, 4, 4), (2, 2, 3, 4.5)) == 2.0

def test_no_overlap():
    assert overlap((1, 1, 4, 4), (4.5, 4.5, 5, 5)) == 0.0
```

Writing saskatchewan/test_overlap.py

```
%%bash
#%cmd #(windows)
cd saskatchewan
pytest || echo "Tests failed"
```

```

=====
 test session starts =====

```

```

platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0

```

```

rootdir: /home/runner/work/rse-course/rse-
course/module05_testing_your_code/saskatchewan

```

```

plugins: anyio-3.6.2, pylama-8.4.1, cov-4.0.0

```

```

collected 3 items

```

```


```

```

test_overlap.py ..F [100%]

```

```

=====
 FAILURES =====

```

```

----- test_no_overlap -----

```

```


```

```

def test_no_overlap():

```

```

>     assert overlap((1, 1, 4, 4), (4.5, 4.5, 5, 5)) == 0.0

```

```

E     assert 0.25 == 0.0

```

```

E     + where 0.25 = overlap((1, 1, 4, 4), (4.5, 4.5, 5, 5))

```

```


```

```

test_overlap.py:13: AssertionError

```

```

=====
 short test summary info =====

```

```

FAILED test_overlap.py::test_no_overlap - assert 0.25 == 0.0

```

```

+ where 0.25 = overlap((1, 1, 4, 4), (4.5, 4.5, 5, 5))

```

```

===== 1 failed, 2 passed in 0.06s =====

```

```

Tests failed

```

Note that it reported **which** test had failed, how many tests ran, and how many failed.

The symbol `..F` means there were three tests, of which the third one failed.

Pytest will:

- automatically finds files `test_*.py`
- collects all subroutines called `test_*`
- runs tests and reports results

Some options:

- help: `pytest --help`
- run only tests for a given feature: `pytest -k foo` # tests with 'foo' in the test name

Coverage reports

Using `pytest` it is possible to see, which lines of code have or haven't been executed by your tests.

The command below will produce a html files which highlights the coverage of your tests.

```

%%bash
# %%cmd #(windows)
cd saskatchewan
pytest --cov=. --cov-report=html || echo "Tests failed"
# MacOS:
# open htmlcov/index.html

```

```

=====
 test session starts =====
platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0
rootdir: /home/runner/work/rse-course/rse-course/module05_testing_your_code/saskatchewan
plugins: asyncio-3.6.2, pylama-8.4.1, cov-4.0.0
collected 3 items

test_overlap.py ..F [100%]

=====
 FAILURES =====
_____
 test_no_overlap _____
_____
def test_no_overlap():

>     assert overlap((1, 1, 4, 4), (4.5, 4.5, 5, 5)) == 0.0
E     assert 0.25 == 0.0
E     +  where 0.25 = overlap((1, 1, 4, 4), (4.5, 4.5, 5, 5))

test_overlap.py:13: AssertionError
_____
----- coverage: platform linux, python 3.8.14-final-0 -----
Coverage HTML written to dir htmlcov
_____
===== short test summary info =====
FAILED test_overlap.py::test_no_overlap - assert 0.25 == 0.0
+  where 0.25 = overlap((1, 1, 4, 4), (4.5, 4.5, 5, 5))
===== 1 failed, 2 passed in 0.04s =====
Tests failed

```

Testing with floating points

Floating points are not reals

Floating points are inaccurate representations of real numbers:

`1.0 == 0.9999999999999999` is true to the last bit.

This can lead to numerical errors during calculations: $1000(a - b) \neq 1000a - 1000b$

```

1000.0 * 1.0 - 1000.0 * 0.9999999999999998
2.273736754323206e-13
1000.0 * (1.0 - 0.9999999999999998)
2.220446049250313e-13

```

Both results are wrong: `2e-13` is the correct answer.

The size of the error will depend on the magnitude of the floating points:

```

1000.0 * 1e5 - 1000.0 * 0.999999999999998e5
1.4901161193847656e-08

```

The result should be `2e-8`.

Comparing floating points

Use the "approx", for a default of a relative tolerance of 10^{-6}

```
from pytest import approx
assert 0.7 == approx(0.7 + 1e-7)
```

Or be more explicit:

```
magnitude = 0.7
assert 0.7 == approx(0.701, rel=0.1, abs=0.1)
```

Choosing tolerances is a big area of debate: <https://software-carpentry.org/blog/2014/10/why-we-dont-teach-testing.html>

Comparing vectors of floating points

Numerical vectors are best represented using [numpy](#).

```
from numpy import array, pi
vector_of_reals = array([0.1, 0.2, 0.3, 0.4]) * pi
```

Numpy ships with a number of assertions (in [numpy.testing](#)) to make comparison easy:

```
from numpy import array, pi
from numpy.testing import assert_allclose

expected = array([0.1, 0.2, 0.3, 0.4, 1e-12]) * pi
actual = array([0.1, 0.2, 0.3, 0.4, 2e-12]) * pi
actual[:-1] += 1e-6
assert_allclose(actual, expected, rtol=1e-5, atol=1e-8)
```

It compares the difference between `actual` and `expected` to $atol + rtol \cdot \text{abs}(expected)$.

5.3 Classroom exercise: energy calculation

Estimated time for this notebook: 30 minutes

Diffusion model in 1D

Description: A one-dimensional diffusion model. (Could be a gas of particles, or a bunch of crowded people in a corridor, or animals in a valley habitat...)

- Agents are on a 1d axis
- Agents do not want to be where there are other agents
- This is represented as an 'energy': the higher the energy, the more unhappy the agents.

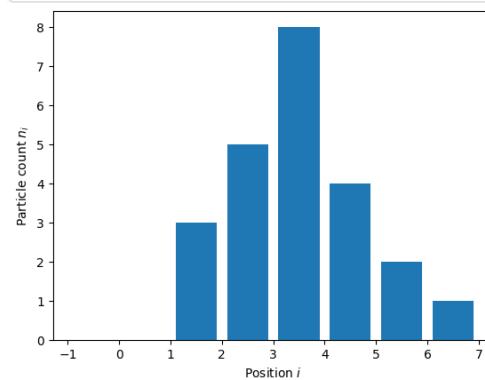
Implementation:

- Given a vector n of positive integers, and of arbitrary length
- Compute the energy, $E(n) = \sum_i n_i(n_i - 1)$
- Later, we will have the likelihood of an agent moving depend on the change in energy.

```
%matplotlib inline
import numpy as np
from matplotlib import pyplot as plt

density = np.array([0, 0, 3, 5, 8, 4, 2, 1])
fig, ax = plt.subplots()
ax.bar(np.arange(len(density)) - 0.5, density)
ax.xrange = [-0.5, len(density) - 0.5]
ax.set_ylabel("Particle count $n_i$")
ax.set_xlabel("Position $i$")
```

```
Text(0.5, 0, 'Position $i$')
```



Here, the total energy due to position 2 is $3(3 - 1) = 6$, and due to column 7 is $1(1 - 1) = 0$. We need to sum these to get the total energy.

Starting point

Create a Python module:

```
%bash
rm -rf diffusion
mkdir diffusion
touch diffusion/__init__.py
```

Windows: You will need to run the following instead

```
%%cmd
rmdir /s diffusion
mkdir diffusion
type nul > diffusion/__init__.py
```

NB. If you are using the Windows command prompt, you will also have to replace all subsequent `%%bash` directives with `%cmd`

- Implementation file: `diffusion_model.py`

```
%%writefile diffusion/model.py
def energy(density, coeff=1.0):
    """Energy associated with the diffusion model

    Parameters
    -----
    density: array of positive integers
        Number of particles at each position i in the array
    coeff: float
        Diffusion coefficient.
    """
    # implementation goes here
```

Writing `diffusion/model.py`

- Testing file: `test_diffusion_model.py`

```
%%writefile diffusion/test_model.py
from .model import energy

def test_energy():
    pass
    # Test something
```

Writing `diffusion/test_model.py`

Invoke the tests:

```
%%bash
cd diffusion
pytest
```

===== test session starts =====

platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0

rootdir: /home/runner/work/rse-course/rse-course/module05_testing_your_code/diffusion

plugins: anyio-3.6.2, pylama-8.4.1, cov-4.0.0

collected 1 item

[100%]

===== 1 passed in 0.01s =====

Now, write your code (in `model.py`), and tests (in `test_model.py`), testing as you do.

Solution

Don't look until after you've tried!

In the spirit of test-driven development let's first consider our tests.

```
%>writefile diffusion/test_model.py
"""Unit tests for a diffusion model."""

from pytest import raises
from .model import energy

def test_energy_fails_on_non_integer_density():
    with raises(TypeError) as exception:
        energy([1.0, 2, 3])

def test_energy_fails_on_negative_density():
    with raises(ValueError) as exception:
        energy([-1, 2, 3])

def test_energy_fails_ndimensional_density():
    with raises(ValueError) as exception:
        energy([[1, 2, 3], [3, 4, 5]])

def test_zero_energy_cases():
    # Zero energy at zero density
    densities = [[], [0], [0, 0, 0]]
    for density in densities:
        assert energy(density) == 0

def test_derivative():
    from numpy.random import randint

    # Loop over vectors of different sizes (but not empty)
    for vector_size in randint(1, 1000, size=30):

        # Create random density of size N
        density = randint(50, size=vector_size)

        # will do derivative at this index
        element_index = randint(vector_size)

        # modified densities
        density_plus_one = density.copy()
        density_plus_one[element_index] += 1

        # Compute and check result
        #  $d(n^2-1)/dn = 2n$ 
        expected = 2.0 * density[element_index] if density[element_index] > 0 else 0
        actual = energy(density_plus_one) - energy(density)
        assert expected == actual

def test_derivative_no_self_energy():
    """If particle is alone, then its participation to energy is zero."""
    from numpy import array

    density = array([1, 0, 1, 10, 15, 0])
    density_plus_one = density.copy()
    density[1] += 1

    expected = 0
    actual = energy(density_plus_one) - energy(density)
    assert expected == actual
```

Overwriting diffusion/test_model.py

Now let's write an implementation that passes the tests.

```
%>writefile diffusion/model.py
"""Simplistic 1-dimensional diffusion model."""

from numpy import array, any, sum

def energy(density):
    """Energy associated with the diffusion model
    :Parameters:
        density: array of positive integers
            Number of particles at each position i in the array/geometry
    """

    # Make sure input is an numpy array
    density = array(density)

    # ...of the right kind (integer). Unless it is zero length,
    #   in which case type does not matter.

    if density.dtype.kind != "i" and len(density) > 0:
        raise TypeError("Density should be an array of *integers*.")
    # and the right values (positive or null)
    if any(density < 0):
        raise ValueError("Density should be an array of *positive* integers.")
    if density.ndim != 1:
        raise ValueError(
            "Density should be an a *1-dimensional* " + "array of positive
            integers."
        )
    return sum(density * (density - 1))
```

Overwriting diffusion/model.py

```
%>bash
cd diffusion
pytest
```

```
=====
test session starts =====
platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0
rootdir: /home/runner/work/rse-course/rse-course/module05_testing_your_code/diffusion
plugins: anyio-3.6.2, pylama-8.4.1, cov-4.0.0
collected 6 items
test_model.py .....
[100%]
=====
6 passed in 0.11s =====
```

Coverage

With pytest, you can use the ["pytest-cov" plugin](#) to measure test coverage

```
%%bash
cd diffusion
pytest --cov

=====
test session starts =====
platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0
rootdir: /home/runner/work/rse-course/rse-course/module05_testing_your_code/diffusion
plugins: anyio-3.6.2, pylama-8.4.1, cov-4.0.0
collected 6 items
test_model.py .....
[100%]
-----
coverage: platform linux, python 3.8.14-final-0 -----
Name     Stmts  Miss  Cover
-----
__init__.py      0     0   100%
model.py        10    0   100%
test_model.py    33    0   100%
-----
TOTAL          43    0   100%
=====
6 passed in 0.15s =====
```

Or an html report:

```
%%bash
#%%%cmd (windows)
cd diffusion
pytest --cov --cov-report html
```

```
=====
 test session starts =====
platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0
rootdir: /home/runner/work/rse-course/rse-course/module05_testing_your_code/diffusion
plugins: anyio-3.6.2, pylama-8.4.1, cov-4.0.0
collected 6 items

test_model.py ......

[100%]

----- coverage: platform linux, python 3.8.14-final-0 -----
Coverage HTML written to dir htmlcov

===== 6 passed in 0.16s =====
```

The HTML coverage results will be in [diffusion/htmlcov/index.html](#)

5.4 Mocking

Estimated time for this notebook: 15 minutes

Definition

Mock: verb,

1. to tease or laugh at in a scornful or contemptuous manner
2. to make a replica or imitation of something

Mocking

- Replace a real object with a pretend object, which records how it is called, and can assert if it is called wrong

Mocking frameworks

- C: [Mocka](#)
- C++: [googlemock](#)
- Python: [unittest.mock](#)

Recording calls with mock

Mock objects record the calls made to them:

```
from unittest.mock import Mock
function = Mock(name="myroutine", return_value=2)

function(1)
2

function(5, "hello", a=True)
2

function.mock_calls
[call(1), call(5, 'hello', a=True)]
```

The arguments of each call can be recovered

```
name, args, kwargs = function.mock_calls[1]
args, kwargs
((5, 'hello'), {'a': True})
```

Mock objects can return different values for each call

```
function = Mock(name="myroutine", side_effect=[2, "xyz"])
function(1)
2

function(1, "hello", {"a": True})
```

```
'xyz'
```

We expect an error if there are no return values left in the list:

```
function()
```

```
-----  
StopIteration                                Traceback (most recent call last)  
Cell In [8], line 1  
----> 1 function()  
  
File /opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/unittest/mock.py:1081,  
in CallableMixin.__call__(self, *args, **kwargs)  
 1079 self._mock_check_sig(*args, **kwargs)  
 1080 self._increment_mock_call(*args, **kwargs)  
-> 1081 return self._mock_call(*args, **kwargs)  
  
File /opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/unittest/mock.py:1085,  
in CallableMixin._mock_call(self, *args, **kwargs)  
 1084 def _mock_call(self, /, *args, **kwargs):  
-> 1085     return self._execute_mock_call(*args, **kwargs)  
  
File /opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/unittest/mock.py:1142,  
in CallableMixin._execute_mock_call(self, *args, **kwargs)  
 1140     raise effect  
 1141 elif not _callable(effect):  
-> 1142     result = next(effect)  
 1143     if _is_exception(result):  
 1144         raise result  
  
StopIteration:
```

Using mocks to model test resources

Often we want to write tests for code which interacts with remote resources. (E.g. databases, the internet, or data files.)

We don't want to have our tests *actually* interact with the remote resource, as this would mean our tests failed due to lost internet connections, for example.

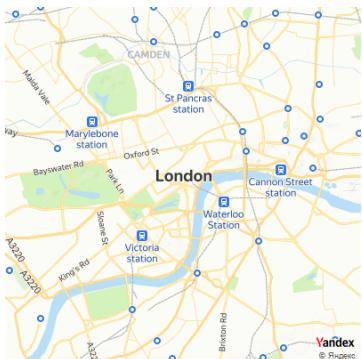
Instead, we can use mocks to assert that our code does the right thing in terms of the *messages it sends*: the parameters of the function calls it makes to the remote resource.

For example, consider the following code that downloads a map from the internet:

```
import requests  
  
def map_at(lat, long, satellite=False, zoom=12, size=(400, 400)):  
    base = "https://static-maps.yandex.ru/1.x/?"  
    params = dict(  
        z=zoom,  
        size=str(size[0]) + "," + str(size[1]),  
        ll=str(long) + "," + str(lat),  
        l="sat" if satellite else "map",  
        lang="en_US",  
    )  
    return requests.get(base, params=params, timeout=60)
```

```
london_map = map_at(51.5073509, -0.1277583)
```

```
%matplotlib inline  
import IPython  
IPython.core.display.Image(london_map.content)
```



We would like to test that it is building the parameters correctly. We can do this by **mocking** the `requests` object. We need to temporarily replace a method in the library with a mock. We can use "patch" to do this:

```
from unittest.mock import patch  
with patch.object(requests, "get") as mock_get:  
    london_map = map_at(51.5073509, -0.1277583)  
    print(mock_get.mock_calls)
```

```
[call('https://static-maps.yandex.ru/1.x/?', params={'z': 12, 'size': '400,400',  
'll': '-0.1277583,51.5073509', 'l': 'map', 'lang': 'en_US'}, timeout=60)]
```

Our tests then look like:

```

def test_build_default_params():
    with patch.object(requests, "get") as mock_get:
        mock_get.assert_called_with(
            "https://static-maps.yandex.ru/1.x/?",
            params={
                "z": 12,
                "size": "400,400",
                "ll": "0,0,51.0",
                "l": "map",
                "lang": "en_US",
            },
            timeout=60,
        )

    test_build_default_params()

```

That was quiet, so it passed. When I'm writing tests, I usually modify one of the expectations, to something 'wrong', just to check it's not passing "by accident", run the tests, then change it back!

Testing functions that call other functions

```

def partial_derivative(function, at, direction, delta=1.0):
    f_x = function(at)
    x_plus_delta = at[:]
    x_plus_delta[direction] += delta
    f_x_plus_delta = function(x_plus_delta)
    return (f_x_plus_delta - f_x) / delta

```

We want to test that the above function does the right thing. It is supposed to compute the derivative of a function of a vector in a particular direction.

E.g.:

```

partial_derivative(sum, [0, 0, 0], 1)

1.0

```

How do we assert that it is doing the right thing? With tests like this:

```

from unittest.mock import MagicMock

def test_derivative_2d_y_direction():
    func = MagicMock()
    partial_derivative(func, [0, 0], 1)
    func.assert_any_call([0, 1.0])
    func.assert_any_call([0, 0])

test_derivative_2d_y_direction()

```

We made our mock a "Magic Mock" because otherwise, the mock results `f_x_plus_delta` and `f_x` can't be subtracted:

```

MagicMock() - MagicMock()

<MagicMock name='mock.__sub__()' id='140230510474864'>

Mock() - Mock()

-----
TypeError                                     Traceback (most recent call last)
Cell In [19], line 1
----> 1 Mock() - Mock()

TypeError: unsupported operand type(s) for -: 'Mock' and 'Mock'

```

5.5 Using a debugger

Estimated time for this notebook: 10 minutes

Stepping through the code

Debuggers are programs that can be used to test other programs. They allow programmers to suspend execution of the target program and inspect variables at that point.

- Mac - compiled languages: [Xcode](#)
- Windows - compiled languages: [Visual Studio](#)
- Linux: [DDD](#)
- all platforms: [eclipse](#), [gdb](#) (DDD and eclipse are GUIs for gdb)
- python: [spyder](#).
- [pdb]<https://docs.python.org/3.8/library/pdb.html>
- R: [RStudio](#), [debug](#), [browser](#)

NB. If you are using the Windows command prompt, you will have to replace all `%bash` directives in this notebook with `%cmd`

Using the python debugger

Unfortunately this doesn't work nicely in the notebook. But from the command line, you can run a python program with:

```
python -m pdb my_program.py
```

Basic navigation:

Basic command to navigate the code and the python debugger:

- `help`: prints the help
- `help n`: prints help about command `n`
- `n(ext)`: executes one line of code. Executes and steps over functions.

- **s**(tep): step into current function in line of code
- **l**(ist): list program around current position
- **w**(here): prints current stack (where we are in code)
- **[enter]**: repeats last command
- **anypythonvariable**: print the value of that variable

The python debugger is a **python shell**: it can print and compute values, and even change the values of the variables at that point in the program.

Breakpoints

Break points tell debugger where and when to stop We say

- **b** somefunctionname

```
%>%writefile energy_example.py
from diffusion.model import energy
print(energy([5, 6, 7, 8, 0, 1]))
```

Writing energy_example.py

The debugger is, of course, most used interactively, but here I'm showing a prewritten debugger script:

```
%>%writefile commands
restart      # restart session
n
b energy     # program will stop when entering energy
c           # continue program until break point is reached
print(density) # We are now "inside" the energy function and can print any
variable.
```

Overwriting commands

```
%>%bash
python -m pdb energy_example.py < commands
```

```
> /home/runner/work/rse-course/rse-
course/module05_testing_your_code/energy_example.py(1)<module>()
```

-> from diffusion.model import energy

```
(Pdb) Restarting /home/runner/work/rse-course/rse-
course/module05_testing_your_code/energy_example.py with arguments:
```

restart session

```
> /home/runner/work/rse-course/rse-
course/module05_testing_your_code/energy_example.py(1)<module>()
```

-> from diffusion.model import energy

```
(Pdb) > /home/runner/work/rse-course/rse-
course/module05_testing_your_code/energy_example.py(3)<module>()
```

-> print(energy([5, 6, 7, 8, 0, 1]))

```
(Pdb) Breakpoint 1 at /home/runner/work/rse-course/rse-
course/module05_testing_your_code/diffusion/model.py:5
```

```
(Pdb) > /home/runner/work/rse-course/rse-
course/module05_testing_your_code/diffusion/model.py(13)energy()
```

-> density = array(density)

(Pdb) [5, 6, 7, 8, 0, 1]

(Pdb)

Alternatively, break-points can be set on files: **b file.py:20** will stop on line 20 of **file.py**.

Post-mortem

Debugging when something goes wrong:

1. Have a crash somewhere in the code
2. run **python -m pdb file.py** or run the cell with **%pdb on**

The program should stop where the exception was raised

1. use **w** and **l** for position in code and in call stack
2. use **up** and **down** to navigate up and down the call stack
3. inspect variables along the way to understand failure

Note Running interactively like in the following example **does** work in the notebook. Try it out!

```
%pdb on
from diffusion.model import energy
partial_derivative(energy, [5, 6, 7, 8, 0, 1], 5)
```

5.6 Continuous Integration

Estimated time for this notebook: 15 minutes

Getting past "but it works on my machine..."

Try running the code below:

```
%%bash
rm -rf continuous_int
mkdir continuous_int
touch continuous_int/__init__.py
```

```
%%writefile continuous_int/test_demo.py
import sys
import re

def test_platform():
    assert re.search("\d", sys.platform)

def test_replace():
    assert "".replace("", "A", 2) == "A"
```

```
Writing continuous_int/test_demo.py
```

```
%%bash
cd continuous_int
pytest || echo "tests complete"
```

```
===== test session starts ======
```

```
platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0
```

```
rootdir: /home/runner/work/rse-course/rse-course/module05_testing_your_code/continuous_int
```

```
plugins: asyncio-3.6.2, pylama-8.4.1, cov-4.0.0
```

```
collected 2 items
```

```
test_demo.py FF [100%]
```

```
===== FAILURES =====
```

```
----- test_platform -----
```

```
def test_platform():
```

```
>     assert re.search("\d", sys.platform)
```

```
E     AssertionError: assert None
```

```
E     +   where None = <function search at 0x7f5b99c51550>('\\d', 'linux')
```

```
E     +   where <function search at 0x7f5b99c51550> = re.search
```

```
E     +   and   'linux' = sys.platform
```

```
test_demo.py:5: AssertionError
```

```
----- test_replace -----
```

```
def test_replace():
```

```
>     assert "".replace("", "A", 2) == "A"
```

```
E     AssertionError: assert '' == 'A'
```

```
E     - A
```

```
test_demo.py:8: AssertionError
```

```
===== warnings summary =====
```

```
test_demo.py:5
```

```
/home/runner/work/rse-course/rse-course/module05_testing_your_code/continuous_int/test_demo.py:5:
DeprecationWarning: invalid escape sequence '\d'
```

```
    assert re.search("\d", sys.platform)
```

```
-- Docs: https://docs.pytest.org/en/stable/how-to/capture-warnings.html
```

```
===== short test summary info =====
```

```
FAILED test_demo.py::test_platform - AssertionError: assert None
```

```
+   where None = <function search at 0x7f5b99c51550>('\\d', 'linux')
```

```
+   where <function search at 0x7f5b99c51550> = re.search
```

```
+   and   'linux' = sys.platform
```

```

FAILED test_demo.py::test_replace - AssertionError: assert '' == 'A'

- A

=====
 2 failed, 1 warning in 0.06s =====

tests complete

```

The example above is a trivial, and deliberate, example of code that will behave differently on different computers.

Much more subtle instances can occur in real-life, which if allowed to propagate, they can result in bugs and errors that are difficult to trace, let alone fix.

One mitigation for this problem is to use a process of "Continuous Integration (CI)". This is a process of drawing together all developer contributions as early as possible and frequently running the automated tests. Typically this involves the use of CI servers, which provide a common and reliable environment to run our tests. (This is not the only use of CI servers - we will touch on other use cases in later modules)

Options for CI Servers

There are many different open-source or proprietary CI Servers available. In some cases it might be appropriate to have [on-premise CI Servers](#) at your organisation.

There are also a number of Continuous-Integration-Server-as-a-Service products that can be used free-of-charge for open source projects. Here we will expand on [GitHub Actions](#), which is a Continuous-Integration-Server-as-a-Service, which is one component of the wider GitHub ecosystem.

Objectives

We would like to test our code on

- different operating systems
- different versions of python
- each commit to a pull request

```

%%bash
mkdir -p continuous_int/.github/workflows

%%writefile continuous_int/.github/workflows/ci-tests.yml
# This workflow will install Python dependencies, run tests with a variety of
# Python versions, on Windows and Linux
# For more information see: https://help.github.com/actions/language-and-
framework-guides/using-python-with-github-actions

name: Unit tests

on:
  pull_request:
    branches:
      - main
    push:
  jobs:
    build:
      strategy:
        # We use `fail-fast: false` for teaching purposes. This ensures that all
        # combinations of the matrix
        # will run even if one or more fail.
        fail-fast: false
      matrix:
        python-version: [3.8, 3.9, "3.10"]
        os: [ubuntu-latest, windows-latest]
      runs-on: ${{ matrix.os }}

    steps:
      - uses: actions/checkout@v2
      - name: Set up Python ${{ matrix.python-version }}
        uses: actions/setup-python@v1
        with:
          python-version: ${{ matrix.python-version }}

      # Yes we have to explicitly install pytest. In a "real" example this could be
      # included in a
      # requirement.txt or environment.yml to setup your environment
      - name: Install PyTest
        run: |
          python -m pip install pytest
      # Now run the tests
      - name: Test with pytest
        run: |
          pytest

```

Writing continuous_int/.github/workflows/ci-tests.yml

Apply this to the personal github repo you made in module 04"

- Create a new branch in your repo.
- Copy the files in the `continuous_int` directory into your local clone. Note that the `.yml` file must exist in the directory `.github/workflows`, which must be in the root of your repo. (The `.` prefixed to the `.github`
- Commit your changes and push them
- Create a Pull Request to the `main` branch of your own repo.

When successfully applied to your repo, you should see that a number of tests are completed on every commit pushed, on every pull request.

These tests have been designed that they will both pass only if they are run on Windows and on Python v3.9 or higher, in order to demonstrate the matrix workings of GH Actions. In a more realistic scenario, you should aim to have your test pass in all contexts.

Further reading:

- There can be cases where it's appropriate to expect different behaviour on different platforms. [PyTest](#) has features that allow for cases.
- GitHub Actions themselves can be difficult to debug because of the need to commit and push every minor change. [Act](#) provides a tool to help debug some GH Actions locally.

Recap example: Monte-Carlo

Problem: Implement and test a simple Monte-Carlo algorithm

Given an input function (energy) and starting point (density) and a temperature T :

1. Compute energy at current density.
2. Move randomly chosen agent randomly left or right.
3. Compute second energy.
4. Compare the two energies:
5. If second energy is lower, accept move.
6. β is a parameter which determines how likely the simulation is to move from a 'less favourable' situation to a 'more favourable' one.
7. Compute $P_0 = e^{-\beta(E_1 - E_0)}$ and P_1 a random number between 0 and 1,
8. If $P_0 > P_1$, do the move anyway.
9. Repeat.
 - the algorithm should work for (m)any energy function(s).
 - there should be separate tests for separate steps! What constitutes a step?
 - tests for the Monte-Carlo should not depend on other parts of code.
 - Use [matplotlib](#) to plot density at each iteration, and make an animation

NB. If you are using the Windows command prompt, you will have to replace all `%%bash` directives in this notebook with `%%cmd`

Solution

We need to break our problem down into pieces:

1. A function to generate a random change: `random_agent()`, `random_direction()`
2. A function to compute the energy before the change and after it: `energy()`
3. A function to determine the probability of a change given the energy difference (1 if decreases, otherwise based on exponential): `change_density()`
4. A function to determine whether to execute a change or not by drawing a random number `accept_change()`
5. A method to iterate the above procedure: `step()`

Next Step: Think about the possible unit tests

1. Input insanity: e.g. density should non-negative integer; testing by giving negative values etc.
2. `change_density()`: density is change by a particle hopping left or right? Do all positions have an equal chance of moving?
3. `accept_change()` will move be accepted when second energy is lower?
4. Make a small test case for the main algorithm. (Hint: by using mocking, we can pre-set who to move where.)

```
%%bash
rm -rf DiffusionExample
mkdir DiffusionExample
```

Windows: You will need to run the following instead

```
%%cmd
rmdir /s DiffusionExample
mkdir DiffusionExample
```

```

%%writefile DiffusionExample/MonteCarlo.py
import matplotlib.pyplot as plt
from numpy import sum, array
from numpy.random import randint, choice

class MonteCarlo:
    """A simple Monte Carlo implementation"""

    def __init__(self, energy, density, temperature=1, itermax=1000):
        from numpy import any, array

        density = array(density)
        self.itermax = itermax

        if temperature == 0:
            raise NotImplementedError("Zero temperature not implemented")
        if temperature < 0.0:
            raise ValueError("Negative temperature makes no sense")

        if len(density) < 2:
            raise ValueError("Density is too short")
        # of the right kind (integer). Unless it is zero length,
        # in which case type does not matter.
        if density.dtype.kind != "i" and len(density) > 0:
            raise TypeError("Density should be an array of *integers*.")
        # and the right values (positive or null)
        if any(density < 0):
            raise ValueError("Density should be an array of" + "*positive*"
                            "integers.")
        if density.ndim != 1:
            raise ValueError(
                "Density should be an a *1-dimensional* + " "array of positive"
                "integers."
            )
        if sum(density) == 0:
            raise ValueError("Density is empty.")

        self.current_energy = energy(density)
        self.temperature = temperature
        self.density = density

    def random_direction(self):
        return choice([-1, 1])

    def random_agent(self, density):
        # Particle index
        particle = randint(sum(density))
        current = 0
        for location, n in enumerate(density):
            current += n
            if current > particle:
                break
        return location

    def change_density(self, density):
        """Move one particle left or right."""
        location = self.random_agent(density)

        # Move direction
        if density[location] - 1 < 0:
            return array(density)
        if location == 0:
            direction = 1
        elif location == len(density) - 1:
            direction = -1
        else:
            direction = self.random_direction()

        # Now make change
        result = array(density)
        result[location] -= 1
        result[location + direction] += 1
        return result

    def accept_change(self, prior, successor):
        """Returns true if should accept change."""
        from numpy import exp
        from numpy.random import uniform

        if successor <= prior:
            return True
        else:
            return exp(-(successor - prior) / self.temperature) > uniform()

    def step(self):
        iteration = 0
        while iteration < self.itermax:
            new_density = self.change_density(self.density)
            new_energy = energy(new_density)

            accept = self.accept_change(self.current_energy, new_energy)
            if accept:
                self.density, self.current_energy = new_density, new_energy
            iteration += 1

        return self.current_energy, self.density

    def energy(density, coefficient=1):
        """Energy associated with the diffusion model
        :Parameters:
            density: array of positive integers
            Number of particles at each position i in the array/geometry
        """
        from numpy import array, any, sum

        # Make sure input is an array
        density = array(density)

        # of the right kind (integer). Unless it is zero length, in which case type
        # does not matter.
        if density.dtype.kind != "i" and len(density) > 0:
            raise TypeError("Density should be an array of *integers*.")
        # and the right values (positive or null)
        if any(density < 0):
            raise ValueError("Density should be an array" + " of *positive* integers.")
        if density.ndim != 1:
            raise ValueError(
                "Density should be an a *1-dimensional* + " "array of positive"
                "integers."
            )
        return coefficient * 0.5 * sum(density * (density - 1))

```

```

import sys
sys.path.append("DiffusionExample")
import numpy as np
from IPython.display import HTML
from matplotlib import animation
from matplotlib import pyplot as plt
from MonteCarlo import MonteCarlo, energy

Temperature = 0.1

density = [np.sin(i) for i in np.linspace(0.1, 3, 100)]
density = np.array(density) * 100
density = density.astype(int)

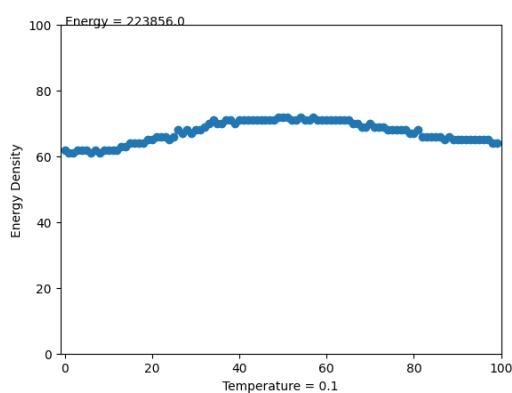
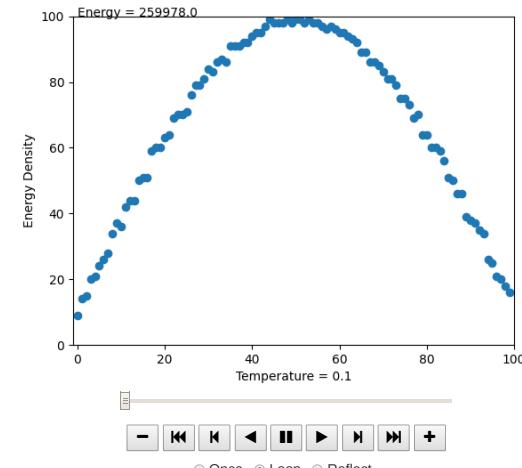
fig = plt.figure()
ax = plt.axes(xlim=(-1, len(density)), ylim=(0, np.max(density) + 1))
image = ax.scatter(range(len(density)), density)
txt_energy = plt.text(0, 100, "Energy = 0")
plt.xlabel("Temperature = 0.1")
plt.ylabel("Energy Density")

mc = MonteCarlo(energy, density, temperature=Temperature)

def simulate(step):
    energy, density = mc.step()
    image.set_offsets(np.vstack((range(len(density)), density)).T)
    txt_energy.set_text(f"Energy = {energy}")

anim = animation.FuncAnimation(fig, simulate, frames=200, interval=50)
HTML(anim.to_jshtml())

```



```

%%writefile DiffusionExample/test_model.py
from MonteCarlo import MonteCarlo
from unittest.mock import MagicMock
from pytest import raises, approx

def test_input_sanity():
    """Check incorrect input do fail"""
    energy = MagicMock()

    with raises(NotImplementedError) as exception:
        MonteCarlo(sum, [1, 1, 1], 0e0)
    with raises(ValueError) as exception:
        MonteCarlo(energy, [1, 1, 1], temperature=-1e0)

    with raises(TypeError) as exception:
        MonteCarlo(energy, [1.0, 2, 3])
    with raises(ValueError) as exception:
        MonteCarlo(energy, [-1, 2, 3])
    with raises(ValueError) as exception:
        MonteCarlo(energy, [[1, 2, 3], [3, 4, 5]])
    with raises(ValueError) as exception:
        MonteCarlo(energy, [3])
    with raises(ValueError) as exception:
        MonteCarlo(energy, [0, 0])

def test_move_particle_one_over():
    """Check density is change by a particle hopping left or right."""
    from numpy import nonzero, multiply
    from numpy.random import randint

    energy = MagicMock()

    for i in range(100):
        # Do this n times, to avoid
        # issues with random numbers
        # Create density

        density = randint(50, size=randint(2, 6))
        mc = MonteCarlo(energy, density)
        # Change it
        new_density = mc.change_density(density)

        # Make sure any movement is by one
        indices = nonzero(density - new_density)[0]
        assert len(indices) == 2, "densities differ in two places"
        assert (
            multiply.reduce((density - new_density)[indices]) == -1
        ), "densities differ by + and - 1"

def test_equal_probability():
    """Check particles have equal probability of movement."""
    from numpy import array, sqrt, count_nonzero

    energy = MagicMock()

    density = array([1, 0, 99])
    mc = MonteCarlo(energy, density)
    changes_at_zero = [
        (density - mc.change_density(density))[0] != 0 for i in range(10000)
    ]
    assert count_nonzero(changes_at_zero) == approx(
        0.01 * len(changes_at_zero), 0.5 * sqrt(len(changes_at_zero))
    )

def test_accept_change():
    """Check that move is accepted if second energy is lower"""
    from numpy import sqrt, count_nonzero, exp

    energy = MagicMock()
    mc = MonteCarlo(energy, [1, 1, 1], temperature=100.0)
    # Should always be true.
    # But do more than one draw,
    # in case randomness incorrectly crept into
    # implementation
    for i in range(10):
        assert mc.accept_change(0.5, 0.4)
        assert mc.accept_change(0.5, 0.5)

    # This should be accepted only part of the time,
    # depending on exponential distribution
    prior, successor = 0.4, 0.5
    accepted = [mc.accept_change(prior, successor) for i in range(10000)]
    assert count_nonzero(accepted) / float(len(accepted)) == approx(
        exp(-(successor - prior) / mc.temperature), 3e0 / sqrt(len(accepted))
    )

def test_main_algorithm():
    import numpy as np
    from numpy import testing
    from unittest.mock import Mock

    density = [1, 1, 1, 1, 1]
    energy = MagicMock()
    mc = MonteCarlo(energy, density, itermax=5)

    acceptance = [True, True, True, True, True]
    mc.accept_change = Mock(side_effect=acceptance)
    mc.random_agent = Mock(side_effect=[0, 1, 2, 3, 4])
    mc.random_direction = Mock(side_effect=[1, 1, 1, 1, -1])
    np.testing.assert_equal(mc.step()[1], [0, 1, 2, 1])

```

Writing DiffusionExample/test_model.py

```

%%bash
cd DiffusionExample
pytest

```

```
===== test session starts =====
platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0
rootdir: /home/runner/work/rse-course/rse-course/module05_testing_your_code/DiffusionExample
plugins: anyio-3.6.2, pylama-8.4.1, cov-4.0.0
collected 5 items

test_model.py .... [100%]

=====
5 passed in 0.61s =====
```

6. Software Projects

- Turning your code into a package
- Releasing code
- Choosing an open-source license
- Software project management
- Organising issues and tasks

Contents

- [6.0 Libraries](#) (5 minutes)
- [6.1 Installing libraries](#) (10 minutes)
- [6.2 Managing Dependencies](#) (15 minutes)
- [6.3 Python outside the notebook](#) (15 minutes)
- [6.4 Packaging](#) (25 minutes)
- [6.5 Documentation](#) (10 minutes)
- [6.6 Software Project Management](#) (5 minutes)
- [6.7 Software Licensing](#) (10 minutes)
- [6.8 Managing software issues](#) (5 minutes)

Total time: 1 hr 40 minutes

Exercises

A classroom exercise is included at the end of the module: [6.9 Exercise: Packaging Troll Treasure](#). We recommend that instructors arrange for the exercise to be done in groups. The exercise can also be left as a self-paced homework assignment if preferred.

6.0 Libraries

Estimated time for this notebook: 5 minutes

What is a library?

In Python, it can be useful to keep the following concepts in mind:

- a **module** is some related code saved in a single `.py` file
- a **package** is a collection of related modules
- a **library** is a collection of related modules and packages

For instance, the `scikit-learn` library contains the `linear_model` package which contains the `LinearRegression` module. In practice, many Python projects are distributed as packages rather than libraries.

In this course we will use the generic term `library` to describe any code that can be reused in multiple places.

Libraries are awesome

The strength of a language lies as much in the set of libraries available, as it does in the language itself.

A great set of libraries allows for a very powerful programming style:

- Write minimal code yourself
- Choose the right libraries
- Plug them together
- Create impressive results

Not only is this efficient with your programming time, it's also more efficient with computer time. The chances are any algorithm you might want to use has already been programmed better by someone else.

Drawbacks of libraries.

- Sometimes, libraries are not looked after by their creator: code that is not maintained *rots*:
 - It no longer works with later versions of *upstream* libraries.
 - It doesn't work on newer platforms or systems.
 - Features that are needed now, because the field has moved on, are not added
- Sometimes, libraries are hard to get working:
 - For libraries in pure python, this is almost never a problem
 - But many libraries involve *compiled components*: these can be hard to install.

Contribute, don't duplicate

- You have a duty to the ecosystem of scholarly software:
 - If there's a tool or algorithm you need, find a project which provides it.
 - If there are features missing, or problems with it, fix them, [don't create your own](#) library.

How to choose a library

- When was the last commit?
- How often are there commits?
- Can you find the lead contributor on the internet?
- Do they respond when approached:
 - issues raised on GitHub
 - emails to developer list
 - community message boards (e.g. [Gitter](#))
 - personal emails
 - tweets
- Are there contributors other than the lead contributor?
- Is there discussion of the library on Stack Exchange?
- Is the code on an open version control tool like GitHub?
- Is it on standard package repositories. (PyPI, apt/yum/brew)
- Are there any tests?
- Download it. Can you build it? Do the tests pass?
- Is there an open test dashboard? (Travis/Jenkins/CDash)
- What dependencies does the library itself have? Do they pass this list?
- Are different versions of the library clearly labeled with version numbers?
- Is there a changelog?

Sensible Version Numbering

The best approach to version numbers clearly distinguishes kinds of change:

Given a version number MAJOR.MINOR.PATCH, e.g. 2.11.14 increment the:

- MAJOR version when you make incompatible API changes,
- MINOR version when you add functionality in a backwards-compatible manner, and
- PATCH version when you make backwards-compatible bug fixes.

This is called [Semantic Versioning](#)

The Python Standard Library

Python comes with a powerful [standard library](#). Learning python is as much about learning this library as learning the language itself. You've already seen a few packages in this library: `math`, `pdb`, `pytest`, `datetime`.

The Python Package Index

Python's real power, however, comes with the Python Package Index: [PyPI](#). This is a huge array of libraries, with all kinds of capabilities, all easily installable from the command line or through your Python distribution.

6.1 Installing libraries

Estimated time for this notebook: 10 minutes

We've seen that there are lots of python libraries. But how do we install them?

The main problem is this: *libraries need other libraries*

So you can't just install a library by copying code to the computer: you'll find yourself wandering down a tree of "dependencies": libraries needed by libraries needed by the library you want.

This is actually a good thing; it means that people are making use of each others' code. There's a real problem in scientific programming, of people who think they're really clever writing their own twenty-fifth version of the same thing.

So using other people's libraries is good.

Why don't we do it more? Because it can often be quite difficult to **install** other peoples' libraries!

Python has developed a good tool for avoiding this: **pip**.

Installing scikit-learn using pip

On a computer you control, on which you have installed python via Anaconda, you will need to open a **terminal** to invoke the library-installer program, **pip**.

- On windows, go to start->all programs->Anaconda->Anaconda Command Prompt
- On mac, start **terminal**.
- On linux, open a bash shell.

Into this shell, type:

```
pip install scikit-learn
```

The computer will install the package automatically from PyPI.

Now, close the Jupyter notebook if you have it open, and reopen it. Check your new library is installed with:

```
from sklearn.datasets import load_wine
wine = load_wine()
X = wine.data
y = wine.target
feature_names = wine.feature_names
```

```

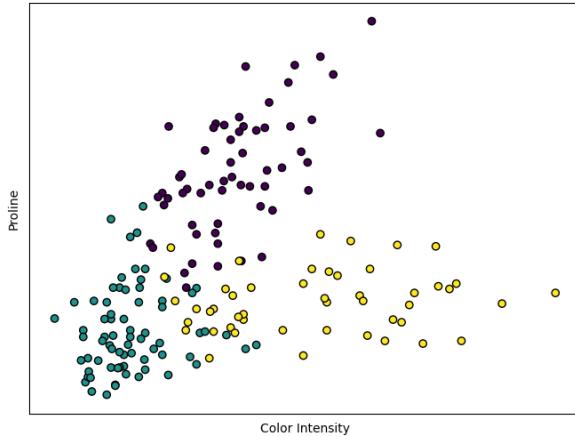
import matplotlib.pyplot as plt
plt.figure(figsize=(8, 6))

# Find the column index of two features to plot
color_idx = feature_names.index("color_intensity")
proline_idx = feature_names.index("proline")

# Plot the training points
plt.scatter(X[:, color_idx], X[:, proline_idx], c=y, edgecolor="k")
plt.xlabel("Color Intensity")
plt.ylabel("Proline")
plt.xticks(())
plt.yticks(())

```

([], [])



That was actually pretty easy, I hope. This is how you'll install new libraries when you need them.

Troubleshooting:

On mac or linux, you *might* get a complaint that you need "superuser", "root", or "administrator" access. If so type:

- `pip install --user scikit-learn`

and enter your password.

If you get a complaint like: 'pip is not recognized as an internal or external command', try the following:

- `conda install pip` (if you are using conda)
- or follow the [official instructions](#) otherwise

Ask one of the instructors/helpers if you're having difficulties, or open an issue in [the course repo](#).

Where do these libraries go?

```

import numpy
numpy.__path__

```

['/opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-packages/numpy']

Your computer will be configured to keep installed Python packages in a particular place.

Python knows where to look for possible library installations in a list of places, called the "PythonPath". It will try each of these places in turn, until it finds a matching library name.

```

import sys
sys.path

```

['/home/runner/work/rse-course/rse-course/module06_software_projects',
 '/opt/hostedtoolcache/Python/3.8.14/x64/lib/python38.zip',
 '/opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8',
 '/opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/lib-dynload',
 '/opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-packages']

Libraries not on PyPI

Sometimes library code you want to use won't be available on PyPI. In that case there are a few options:

The library is available on another package index

For example, some libraries not on PyPI are available with `conda` (see details below).

The library is in a git repo

If the library is available on GitHub or another service hosting git repos, `pip` can install it from the repo instead of PyPI:

```
pip install git+https://github.com/alan-turing-institute/skttime.git
```

This could also be an option if you need a development version of a library that hasn't been released yet.

(NB: `skttime` is also available [on PyPI](#), we just use it as an example here).

Download and install locally

Sometimes you'll need to download the source code directly, or to test the installation of a library you're working on yourself. To do this, download the code, then, in a terminal:

```
cd <path_to_library_code> # change to the code directory  
pip install . # install the library at the current path
```

Installing binary dependencies with conda

`pip` is the usual Python tool for installing libraries. But there's one area of library installation that is still awkward: some python libraries depend not on other `python` libraries, but on libraries in C++ or Fortran.

This can cause you to run into difficulties installing some libraries. Fortunately, Anaconda provide a carefully managed set of scripts for installing lots of these awkward non-python libraries too. You can do this with the `conda` command line tool, if you're using Anaconda.

Simply type

```
conda install <whatever>
```

instead of `pip install`. This will fetch the python package not from PyPi, but from Anaconda's distribution for your platform, and manage any non-python dependencies too.

Typically, if you're using Anaconda, whenever you come across a python package you want, you should check if Anaconda package it first using `conda search` (or [this list in the documentation](#)). If it is there you can `conda install` it, you'll likely have less problems. But Anaconda doesn't package everything, so you'll need to `pip install` from time to time.

The maintainers of packages may have also provided releases of their software via [conda-forge](#), a community-driven project that provides a collection of packages for the anaconda environment. In such cases you can [add conda-forge](#) to your anaconda installation and use `search` and `install` as explained above.

Other Distribution tools

Distribution tools allow one to obtain a working copy of someone else's package.

Language-specific tools:

- Python: [PyPI](#)
- R: [CRAN](#)
- Ruby: [Ruby_Gems](#)
- Perl: [CPAN](#)

Platform-specific packagers:

- Ubuntu and Debian: [dpkg](#) for `apt-get`
- Redhat and Fedora: [rpm](#) for `yum`
- Mac OS: [homebrew](#)
- Windows: Chocolatey

If you're working in a compiled language like C++ or Fortran, there's often no language specific repository. You'll need to write platform installers for as many platforms as you want to support.

6.2 Managing Dependencies

Estimated time for this notebook: 15 minutes

Specifying Dependencies

`pip` and `requirements.txt`

Probably the most well known and ubiquitous way of specifying and installing dependencies in Python is with a `requirements.txt` file. This is a text file with a list of the names of packages your code relies on, for example:

```
geopy  
imageio  
matplotlib  
numpy  
requests
```

To install dependencies from a `requirements.txt` file do the following:

```
pip install -r requirements.txt
```

`requirements.txt` files are not the only way of specifying dependencies, we'll refer to some others and the differences between them here and later in this module.

Pinning versions

Different versions of libraries may have different features, behaviour, and interfaces. To ensure our code is reproducible and other users (and ourselves in the future) get the same results from running the code, it's a good idea to specify the version of each dependency that should be installed.

To pin dependencies to specific versions include them in `requirements.txt` like this:

```
geopy==2.2.0  
imageio==2.19.3  
matplotlib==3.5.2  
numpy==1.23.0  
requests==2.28.1
```

To automatically generate a `requirements.txt` file like this, containing the versions of all the libraries installed in your current Python environment, you can run:

```
pip freeze
```

However, note that `pip freeze` won't output only your direct dependencies, but also

- the dependencies of your dependencies
- the dependencies of the dependencies of your dependencies

- ...

It may be better to only specify your direct dependencies and let the maintainers of those libraries deal with their own dependencies (but that can also come with future problems and incompatibilities in some cases).

Version ranges

You don't have to specify an exact version, you can also use comparisons like `<=`, `!=`, and `>=` to give ranges of package versions that are compatible with your code (see [here](#)).

An interesting one is `~`, or "approximately equal to". For example, if we specified the numpy dependency as:

```
numpy~=1.23.0
```

it allows `pip` to install any (newer) `1.23.x` version of numpy (e.g. `1.23.1` or `1.23.5`), but not versions `1.24.0` or later (which may introduce changes that are incompatible with `1.23.0`).

(How) should you pin dependency versions?

There are potential caveats and pitfalls with all approaches. At the extremes you have:

- **Not specifying a version:**
 - Dependencies are likely to introduce breaking changes in the future that will cause your code to fail or give different results.
- **Pinning an exact version:**
 - Specific versions may not be available on all platforms. You (or new users of your code) won't get bug and security fixes in new versions.

For research code, to ensure you get exactly the same results from repeating an analysis on another system (or a fresh installation on the same system) pinning versions is often the best approach.

Updating dependencies

Running

```
pip list --outdated
```

will show a list of installed packages that have newer versions available. You can upgrade to the latest version by running:

```
pip install --upgrade PACKAGE_NAME
```

(and then update `requirements.txt` to reflect the new version you're using, if needed).

This is quite a manual approach and other tools have more streamlined ways of handling the upgrading process. See [Poetry](#), for example.

There are also automated tools like [dependabot](#) that look at the dependencies in your GitHub repo and suggest changes to avoid security vulnerabilities.

Virtual Environments

Specifying dependency versions may not always be enough to give you a working (and future-proof) set up for yourself and other users of your code. For example, you may have:

- Different projects on your system requiring different versions of a library, or libraries that are incompatible with each other.
- Libraries that are only available on some platforms (e.g. Linux only) or have different behaviour on other platforms.
- Projects requiring different versions of Python itself

For these reasons we'd recommend using a separate "virtual environment" for each project on your system.

In a virtual environment all the packages you install are isolated from other environments, so you could have one environment using `Python 3.10` and `numpy 1.23.1`, and another using `Python 3.8` and `numpy 1.20.3`, for example.

venv

`venv` is included in the Python standard library and can be used to create virtual environments (see the docs [here](#)).

To create a virtual environment:

```
%> bash
python -m venv myenv
```

where `myenv` is the name of the directory that will be created to store the environment's configuration. The initial contents of the `myenv` directory are:

```
%> bash
ls -F myenv/
bin/
include/
lib/
lib64@
pyvenv.cfg
```

The `which` bash command returns the path to an executable on your system. Currently, `which python` will return the path to the environment you're using to run the course notebooks:

```
%> bash
which python
/opt/hostedtoolcache/Python/3.8.14/x64/bin/python
```

To use our new virtual environment instead, we need to "activate" it, as follows:

```
%>bash  
source myenv/bin/activate  
which python
```



```
/home/runner/work/rse-course/rse-  
course/module06_software_projects/myenv/bin/python
```

the path to the `python` executable now points to a location in the `myenv` directory (our separate Python virtual environment).

We can then install and run libraries without impacting other Python environments on our system, e.g.:

```
%>bash  
source myenv/bin/activate  
pip install pyjokes  
pyjoke
```



```
Collecting pyjokes
```



```
  Downloading pyjokes-0.6.0-py2.py3-none-any.whl (26 kB)
```



```
Installing collected packages: pyjokes
```



```
Successfully installed pyjokes-0.6.0
```



```
WARNING: You are using pip version 22.0.4; however, version 22.3 is available.
```



```
You should consider upgrading via the '/home/runner/work/rse-course/rse-  
course/module06_software_projects/myenv/bin/python -m pip install --upgrade pip'  
command.
```



```
There are II types of people: Those who understand Roman Numerals and those who  
don't.
```

Note:

- You only need to `activate` the environment once usually. We need to do it in each cell here because using `%>bash` is like creating a new terminal.
- If you try `which pip` before and after activating the environment you'll see that the virtual environment uses a different `pip` executable as well

To leave the virtual environment and return to using your system Python (or the previously activated environment), you need to "deactivate" it:

```
%>bash  
source myenv/bin/activate  
echo "=====."  
echo "In myenv, python path:"  
which python  
pyjoke  
echo "=====."  
deactivate  
echo ""  
echo "=====."  
echo "Outside myenv, python path:"  
which python  
pyjoke  
echo "=====."
```



```
=====
```



```
In myenv, python path:
```



```
/home/runner/work/rse-course/rse-  
course/module06_software_projects/myenv/bin/python
```



```
QA Engineer walks into a bar. Orders a beer. Orders 0 beers. Orders 999999999  
beers. Orders a lizard. Orders -1 beers. Orders a sfdeljknessv.
```



```
=====
```



```
=====
```



```
=====
```



```
Outside myenv, python path:
```



```
/opt/hostedtoolcache/Python/3.8.14/x64/bin/python
```



```
bash: line 15: pyjoke: command not found
```



```
=====
```

conda

[conda](#) is a virtual environment, dependency, and package manager for multiple languages. There are multiple distributions including [Anaconda](#), which comes with many common data science libraries pre-installed, and [Miniconda](#), which is `conda` without pre-installed dependencies.

Advantages of conda include:

- It has binaries built for multiple platforms, e.g. `conda` packages are usually available on Windows, Mac, and Linux (whereas it's quite common to find packages on PyPI that don't have a Windows build, for example).
- You can use it to install non-Python dependencies.
- It's an "all-in-one" tool: You can use it to manage your entire Python workflow.

Some disadvantages:

- Other users of your code may not have or want to use conda (but everyone using Python will have pip available, for example)
- There's a bit more bloat than other tools, and the dependency resolver can be quite slow.

conda environments are specified in the YAML format, typically in a file called `environment.yml`, and look like this:

```
name: myenv
dependencies:
  - python=3.9
  - geopy<2.2.0
  - imageio<2.19.3
  - matplotlib<3.5.2
  - numpy<1.23.0
  - requests<2.28.1
```

Note that a version of Python itself is specified in the dependencies - you can install any version of Python in a conda environment.

To create the environment:

```
conda env create -f environment.yml
```

And to use it:

```
conda activate myenv
# work in the environment
conda deactivate
```

Which to choose?

There's a large ecosystem of different Python (and general) dependency, environment, and packaging tools, many more than we've seen here. A few other notable ones are:

- [Docker](#) - creates isolated "containers" which are whole virtual systems, allowing you to configure everything including the operating system to use. This is a "maximally reproducible" solution that ensures future users of your code get a complete and identical environment from the ground up.
- [pyenv](#) - install and manage different versions of Python
- [Poetry](#) - create virtual environments and Python packages, and improved dependency management
- [setuptools](#) - for creating Python packages (we'll be looking at this later)

Which are best for your project depends on what you're trying to achieve and personal preference. It's also likely that you'll be using multiple tools as they all have different priorities and features.

This table gives a rough summary of what the tools mentioned in this course can be used for, loosely ordered from most flexibility (but perhaps most involved setup) at the top, to simpler, single-usecase tools at the bottom:

	Virtual environments	Install non-Python dependencies	Install Python versions	Manage Python dependencies	Create Python packages
Docker	✓	✓	✓	✗	✗
conda	✓	✓	✓	✓	✓
Poetry	✓	✗	✗	✓	✓
pyenv	✓	✗	✓	✗	✗
setuptools	✗	✗	✗	✓	✓
venv	✓	✗	✗	✗	✗
pip	✗	✗	✗	✓	✗

6.3 Python outside the notebook

Estimated time for this notebook: 15 minutes

We will often want to save our Python functions and classes, for use in multiple Notebooks or to interact with them via a terminal.

Writing Python in Text Files

If you create your own Python files ending in `.py`, then you can import them with `import` just like external libraries.

It's best to use an editor like [VS Code](#) or [PyCharm](#) to do this. Here we use the `%%writefile` Jupyter "magic" to create files from the notebook.

Let's create a file `greeter.py` with a function `greet` that prints a welcome message in multiple colours (using the [colorama](#) package):

```
%%writefile greeter.py
import colorama # used for creating coloured text

def greet(personal, family, title="", polite=False):
    greeting = "How do you do, " if polite else "Hey, "
    greeting = colorama.Back.BLACK + colorama.Fore.YELLOW + greeting
    if title:
        greeting += colorama.Back.BLUE + colorama.Fore.WHITE + title + " "
    greeting += (
        colorama.Back.WHITE
        + colorama.Style.BRIGHT
        + colorama.Fore.RED
        + personal
        + " "
        + family
    )
    return greeting
```

Writing greeter.py

Loading Our Function

We just wrote the file, there is no `greet` function in this notebook yet:

```
greet("James", "Hetherington")
```

```
NameError Traceback (most recent call last)
Cell In [2], line 1
----> 1 greet("James", "Hetherington")
NameError: name 'greet' is not defined
```

But we can import the functionality from `greeter.py` file that we created:

```
import greeter # note that you don't include the .py extension
print(greeter.greet("James", "Hetherington"))
```

```
Hey, James Hetherington
```

Or import the function from the file directly:

```
from greeter import greet
print(greet("James", "Hetherington"))
```

```
Hey, James Hetherington
```

Note the file we created is in the same directory as this notebook:

```
# glob is a library for finding files that match given patterns
from glob import glob

# all files with a .py or .ipynb extension in the current directory
glob("*.py") + glob("*.ipynb")
```

```
['greeter.py',
'06_05_documentation.ipynb',
'06_02_managing_dependencies.ipynb',
'06_04_packaging.ipynb',
'06_06_software_development.ipynb',
'06_01_installing_packages.ipynb',
'06_03_non_notebook_python.ipynb',
'06_07_software_licensing.ipynb',
'06_00_libraries.ipynb',
'06_08_software_issues.ipynb',
'06_09_exercise.ipynb']
```

Currently we're relying on all the module source code being in our current working directory. We'll want to `import` our modules from notebooks elsewhere on our computer: it would be a bad idea to keep all our Python work in one folder.

The best way to do this is to learn how to make our code into a proper module that we can install. We'll see more on that in the next notebook.

Command-line Interfaces

`argparse` is the standard Python library for building programs with a command-line interface (another popular library is `click`).

Here's an example that creates a command-line interface to our `greet` function (in a file named `command.py`):

```
%%writefile command.py
from argparse import ArgumentParser

from greeter import greet

def process():
    parser = ArgumentParser(description="Generate appropriate greetings")

    # required (positional) arguments
    parser.add_argument("personal")
    parser.add_argument("family")

    # optional (keyword) arguments
    parser.add_argument("-title", "-t")
    parser.add_argument("-polite", "-p", action="store_true")
    # polite will be false unless "-polite" or "-p" given at command-line

    args = parser.parse_args()

    print(greet(args.personal, args.family, args.title, args.polite))

if __name__ == "__main__":
    process()
```

```
Writing command.py
```

We can now run our saved interface with `python command.py` + the arguments we want to specify.

`argparse` generates some documentation to help us understand how to use it:

```
%%bash
python command.py --help
```

```
usage: command.py [-h] [--title TITLE] [--polite] personal family
```

```
Generate appropriate greetings
```

```
positional arguments:
```

```
personal
```

```
family
```

```
optional arguments:
```

```
-h, --help      show this help message and exit
```

```
--title TITLE, -t TITLE
```

```
--polite, -p
```

A few examples:

```
%>%bash
python command.py James Hetherington
```

```
Hey, James Hetherington
```

```
%>%bash
python command.py --polite James Hetherington
```

```
How do you do, James Hetherington
```

```
%>%bash
python command.py James Hetherington --title Dr
```

```
Hey, Dr James Hetherington
```

Having to type `python command.py ...` is not very intuitive, and we're still relying on our files being in the same directory. In the next notebook we'll see a better way to include command-line interfaces as part of a package.

```
if __name__ == "__main__"
```

In the `command.py` script above you may have noticed the strange `if __name__ == "__main__"` line. This is generally used when you have a file that can be used both as a script and as a module in a package.

Let's create a simplified version of `greeter.py` that prints the name of the special `__name__` variable when it is called:

```
%>%writefile greeter.py
print("executing greeter.py, __name__ is", __name__)

def greet(personal, family):
    return "Hey, " + personal + " " + family

if __name__ == "__main__":
    print(greet("Laura", "Greeter"))
```

```
Overwriting greeter.py
```

If we invoke `greeter.py` directly, Python sets the value of `__name__` to `"__main__"` and the code in the if block runs:

```
%>%bash
python greeter.py
```

```
executing greeter.py, __name__ is __main__
```

```
Hey, Laura Greeter
```

Now let's create a simplified `command.py` that also prints `__name__`, and imports the `greet` function from `greeter.py` as before:

```
%>%writefile command.py
print("executing command.py, __name__ is", __name__)
from argparse import ArgumentParser
from greeter import greet

def process():
    parser = ArgumentParser(description="Generate appropriate greetings")
    parser.add_argument("--personal")
    parser.add_argument("--family")
    args = parser.parse_args()
    print(greet(args.personal, args.family))

if __name__ == "__main__":
    process()
```

```
Overwriting command.py
```

And run the command script:

```
%>bash
python command.py Sarah Command
```

```
executing command.py, __name__ is __main__
```

```
executing greeter.py, __name__ is greeter
```

```
Hey, Sarah Command
```

Note that when we import `greeter.greet` the contents of the whole `greeter.py` file are executed, so the code to print the value of `__name__` still runs. However, `__name__` is now given the value `greeter`. This means when the if statement is executed `__name__ == "__main__"` returns `False`, and we don't see the "Hey, Laura Greeter" output.

Without that if statement we would get

```
Hey, Laura Greeter
Hey, Sarah Command
```

which is unlikely to be what we wanted when running `python command.py Sarah Command`.

6.4 Packaging

Estimated time for this notebook: 25 minutes

Once we've made a working program, we'd like to be able to use it across our system and to share it with others. To do this we need to create our own Python package.

As an example, we'll create a package from the `greeter.py` and `command.py` files from the previous notebook. But we'll delete the files created last time first, to start from a clean slate:

```
# Tidy up files created by previous notebook
import os

files = ["greeter.py", "command.py"]
for f in files:
    if os.path.exists(f):
        os.remove(f)
```

Laying out a project

When planning to package a project for distribution, defining a suitable project layout is essential. We have a typical example of a package layout in the `Greetings` directory, which looks like this:

```
Greetings
└── greetings
    ├── __init__.py
    ├── command.py
    └── greeter.py
    ├── LICENSE.md
    ├── pyproject.toml
    ├── README.md
    └── tests
        └── test_greeter.py
```

The package directory

All your library source code should be in a single directory tree under the parent project directory. Libraries are usually structured with multiple files, perhaps one for each class.

The source code directory (and sub-directories) should contain an `__init__.py` file, which makes Python treat it as a module. **The `__init__.py` file can be empty.**

With the file layout above, `import greetings`, `import greetings.command`, and `import greetings.greeter` will all be possible after installing the package.

If we added a sub-directory, to provide functionality for multiple languages for example, with this structure:

```
greetings
└── __init__.py
    ├── command.py
    ├── greeter.py
    └── languages
        ├── __init__.py
        ├── english.py
        └── italian.py
```

then `import greetings.languages`, `import greetings.languages.english`, and `import greetings.languages.italian` would become available. This is a way to group together related functionality/features in your package.

⚠ Advanced topic: The contents of the `__init__.py` file(s) is executed when you import a package. A common use case for non-empty `__init__.py` files is to "shortcut" imports for convenience. For example, to import the main `greet` function we'd currently need to do:

```
from greetings.greeter import greet
```

If we added that import code as a line in `greetings/__init__.py`, it will then be possible to do:

```
from greetings import greet
```

Build systems and config files

To install your package you need to define a "build system", the tool that will do the work of creating the package, and to provide a configuration file to specify how your package should be built.

The three most common package config files are:

- `pyproject.toml` (**preferred**)
- `setup.cfg` (may be deprecated in the future)
- `setup.py` (may be needed for packages with complex build requirements)

You'll find a lot of projects that use `setup.py` (which used to be the standard), but for new projects it's recommended to use `pyproject.toml`. [TOML](#) is a modern file format for configuration files.

There are multiple "build systems" that can interpret `pyproject.toml` files and build your package. The original and most ubiquitous is [setuptools](#), which we'll use here.

Other options include [Poetry](#), [Flit](#) and [Hatch](#). We'd recommend looking at Poetry as an option for managing dependencies, virtual environments, and packaging together. The structure of `pyproject.toml` will differ depending on the tool you're using.

Using setuptools and pyproject.toml

Specifying the build system

The `[build-system]` section gives the details the tool that should be used to create the package from our code, in this case `setuptools`:

```
[build-system]
requires = ["setuptools"] # the build tool to use
build-backend = "setuptools.build_meta" # the function to use to build the package
```

Specifying project metadata

The `[project]` section contains metadata about your package, at minimum this should include your package's name (usually the name of your package directory) and a version number:

```
[project]
name = "greetings"
version = "0.0.1"
```

Specifying dependencies

Rather than in a `requirements.txt` file, your package's dependencies should be specified in `pyproject.toml`. These are passed as a list in the `[project]` section (in this case we only have one dependency, `colorama`):

```
[project]
dependencies = ["colorama ~= 0.4.4"]
```

Python version

If our package requires a certain Python version to work, that can also be specified:

```
[project]
requires-python = ">=3.6"
```

Optional dependencies

Sometimes a package may have extra optional features, with extra dependencies, that not all users need. A common example is development dependencies (e.g. for running tests, building documentation, checking code quality, and similar) that a normal user won't need. Optional dependencies can be specified in the `[project.optional-dependencies]` section:

```
[project.optional-dependencies]
dev = ["pytest ~= 7.1.2"]
```

`dev` is the name of an optional group of dependencies that can be passed to `pip` when installing the package (see below). We could have multiple groups here with different (arbitrary) names and sets of dependencies.

Make a command-line interface

In the previous notebook we created a script `command.py` that could be run with `python command.py ...` with configurable arguments using `argparse`. We can include scripts like these in the package installation to create a more intuitive CLI (command-line interface) for our library:

```
[project.scripts]
greet = "greetings.command:process"
```

The syntax above means that after installing the package the command `greet` will become available on our system, and running `greet` will call the `process` function in the `greetings/command.py` file. See below for this in action.

Complete pyproject.toml

All together this is our complete `pyproject.toml` file:

```
[build-system]
requires = ["setuptools"]
build-backend = "setuptools.build_meta"

[project]
name = "greetings"
version = "0.0.1"
requires-python = ">=3.6"
dependencies = ["colorama ~= 0.4.4"]

[project.optional-dependencies]
dev = ["pytest ~= 7.1.2"]

[project.scripts]
greet = "greetings.command:process"
```

This is a minimal example but there are many other metadata fields you can include and configuration options. See the [setuptools](#) and [Python packaging](#) docs for details.

Installing the package

We can now install this code with

```
%%bash
cd Greetings
pip install .
```

```
Processing /home/runner/work/rse-course/rse-course/module06_software_projects/Greetings

Installing build dependencies: started
Installing build dependencies: finished with status 'done'
Getting requirements to build wheel: started
Getting requirements to build wheel: finished with status 'done'
Installing backend dependencies: started
Installing backend dependencies: finished with status 'done'
Preparing metadata (pyproject.toml): started
Preparing metadata (pyproject.toml): finished with status 'done'

Requirement already satisfied: colorama==0.4.4 in
/opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-packages (from
greetings==0.0.1) (0.4.6)

Building wheels for collected packages: greetings
Building wheel for greetings (pyproject.toml): started
Building wheel for greetings (pyproject.toml): finished with status 'done'

Created wheel for greetings: filename=greetings-0.0.1-py3-none-any.whl size=2785
sha256=966e8748b41436c7292abf99c76d270ad2cfbe1a92afa9f5046099c3e29be469

Stored in directory: /tmp/pip-ephem-wheel-cache-
b9eii9gd/wheels/a6/5c/81/cic4894d5b25832088cf648e757dcc6d8eff3dc2116123a4a

Successfully built greetings
Installing collected packages: greetings
Successfully installed greetings-0.0.1
```

Installing optional dependencies

To install dependencies specified in `[project.optional-dependencies]`, include the name of the optional group in square brackets, like this:

```
cd Greetings
pip install ".[dev]"
```

Editable mode

If you modify your source files, you would now find it appeared as if the program doesn't change.

That's because pip install **copies** the file elsewhere during installation (the location is system-dependent).

If you want to install a package, but keep working on it, you can install it in "editable mode".

⚠️ As of August 2022, `setuptools` does not support editable installs with `pyproject.toml` (only) packages, so you will need a small `setup.py` file to make this work (see below). But this shouldn't be necessary [in the near future](#).

```
%>writewfile Greetings/setup.py
from setuptools import setup
setup()

Writing Greetings/setup.py
```

Then to install in editable mode:

```
cd Greetings
pip install -e ".[dev]"
```

Installing from GitHub

If we have our code in a (public) git repo anyone can now install our package directly from the git URL:

```
pip install git+git://github.com/alan-turing-institute/Greetings
```

Uploading to PyPI

We could now submit "greeter" to PyPI so everyone could `pip install greetings` directly. For details see the [Python packaging tutorial](#).

Note there is very little approval/review process - you can put pretty much anything on PyPI. Keep that in mind and be wary about installing unknown packages!

Using the Package

The package is now available to use everywhere on the system.

⚠️ You may need to restart your Jupyter notebook kernel for the newly installed package to be recognised.

```
from greetings.greeter import greet  
print(greet("James", "Hetherington"))
```

```
Hey, James Hetherington
```

And the scripts are now available as command line commands:

```
%%bash  
greet --help
```

```
usage: greet [-h] [--title TITLE] [--polite] personal family
```

```
Generate appropriate greetings
```

```
positional arguments:
```

```
personal
```

```
family
```

```
optional arguments:
```

```
-h, --help      show this help message and exit
```

```
--title TITLE, -t TITLE
```

```
--polite, -p
```

```
%%bash  
greet James Hetherington  
greet --polite James Hetherington  
greet James Hetherington --title Dr
```

```
Hey, James Hetherington
```

```
How do you do, James Hetherington
```

```
Hey, Dr James Hetherington
```

Of course, there's more to do when taking code from a quick script and turning it into a proper module. We'll continue to look at this in the rest of the course, but here are some initial ideas:

Write some unit tests

Contents of `Greetings/tests/test_greeter.py`:

```
from greetings.greeter import greet  
  
def test_greeter():  
    inputs = [  
        {"personal": "James", "family": "Hetherington"},  
        {"personal": "James", "family": "Hetherington", "polite": True},  
        {"personal": "James", "family": "Hetherington", "title": "Dr"},  
    ]  
    outputs = [ # codes like \x1b[32m are colours  
        "\x1b[40m\x1b[33mHey, \x1b[47m\x1b[1m\x1b[31mJames Hetherington",  
        "\x1b[40m\x1b[33mHow do you do, \x1b[47m\x1b[1m\x1b[31mJames Hetherington",  
        "\x1b[40m\x1b[33mHey, \x1b[44m\x1b[37mDr \x1b[47m\x1b[1m\x1b[31mJames Hetherington",  
    ]  
    for inp, out in zip(inputs, outputs):  
        assert greet(**inp) == out
```

```
%%bash  
cd Greetings  
pytest
```

```
=====
 test session starts =====
platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0
rootdir: /home/runner/work/rse-course/rse-course/module06_software_projects/Greetings
plugins: anyio-3.6.2, pylama-8.4.1, cov-4.0.0
collected 1 item
tests/test_greeter.py .
[100%]
=====
 1 passed in 0.01s =====
```

Write a README file

e.g.:

```
%>%%writefile Greetings/README.md
Greetings!
=====
This is a very simple example package used as part of the Turing [Research Software Engineering with Python](https://alan-turing-institute.github.io/rse-course) course.

Usage:
Invoke the tool with greet <firstName> <Secondname>
```

Overwriting Greetings/README.md

Write a license file

e.g.:

```
%>%%writefile Greetings/LICENSE.md
(C) The Alan Turing Institute 2021
This "greetings" example package is granted into the public domain.
```

Overwriting Greetings/LICENSE.md

Write a citation file

e.g.:

```
%>%%writefile Greetings/CITATION.md
If you wish to refer to this course, please cite the URL
https://alan-turing-institute.github.io/rse-course

Portions of the material are taken from Software Carpentry
http://swcarpentry.org
```

Overwriting Greetings/CITATION.md

You may well want to formalise this using the [codemeta.json](#) standard - this doesn't have wide adoption yet, but we recommend it.

Documentation

This documentation string explains how to use the function; don't worry about this for now, we'll consider this next time.

```
def greet(personal, family, title="", polite=False):
    """Generate a greeting string for a person.

    Parameters
    -----
    personal: str
        A given name, such as Will or Jean-Luc
    family: str
        A family name, such as Riker or Picard
    title: str
        An optional title, such as Captain or Reverend
    polite: bool
        True for a formal greeting, False for informal.

    Returns
    -----
    string
        An appropriate greeting
    """
    greeting = "How do you do, " if polite else "Hey, "
    greeting += Fore.GREEN + personal
    if title:
        greeting += Fore.BLUE + title + " "
    greeting += Fore.RED + family + "."
    return greeting
```

```
import greetings
help(greetings.greeter.greet)
```

```

Help on function greet in module greetings.greeter:
greet(personal, family, title='', polite=False)
    Generate a greeting string for a person.

Parameters
-----
personal: str
    A given name, such as Will or Jean-Luc
family: str
    A family name, such as Riker or Picard
title: str
    An optional title, such as Captain or Reverend
polite: bool
    True for a formal greeting, False for informal.

Returns
-----
string
    An appropriate greeting

```

6.5 Documentation

Estimated time for this notebook: 10 minutes

Documentation is hard

- Good documentation is hard, and very expensive.
- Bad documentation is detrimental.
- Good documentation quickly becomes bad if not kept up-to-date with code changes.
- Professional companies pay large teams of documentation writers.

Prefer readable code with tests and vignettes

If you don't have the capacity to maintain great documentation, focus on:

- Readable code
- Automated tests
- Small code samples demonstrating how to use the api

Comment-based Documentation tools

Documentation tools can produce extensive documentation about your code by pulling out comments near the beginning of functions, together with the signature, into a web page.

The most popular is [Doxygen](#)

[Have a look at an example of some Doxygen output](#)

[Sphinx](#) is nice for Python, and works with C++ as well. Here's some [Sphinx-generated output](#) and the [corresponding source code Breathe](#) can be used to make Sphinx and Doxygen work together.

[Roxygen](#) is good for R.

Example of using Sphinx

Write some docstrings

We're going to document our "greeter" example using docstrings with Sphinx.

There are various conventions for how to write docstrings, but the native sphinx one doesn't look nice when used with the built in `help` system.

In writing Greeter, we used the docstring conventions from NumPy. So we use the `numpydoc` sphinx extension to support these.

```

"""
Generate a greeting string for a person.

Parameters
-----
personal: str
    A given name, such as Will or Jean-Luc

family: str
    A family name, such as Riker or Picard
title: str
    An optional title, such as Captain or Reverend
polite: bool
    True for a formal greeting, False for informal.

Returns
-----
string
    An appropriate greeting
"""

```

Set up sphinx

Invoke the `sphinx-quickstart` command to build Sphinx's configuration file automatically based on questions at the command line:

```
sphinx-quickstart docs
```

(`docs` is the name of the directory where the documentation will be stored)

Which responds:

```
Welcome to the Sphinx 4.4.0 quickstart utility.

Please enter values for the following settings (just press Enter to
accept a default value, if one is given in brackets).

Selected root path: docs

You have two options for placing the build directory for Sphinx output.
Either, you use a directory "_build" within the root path, or you separate
"source" and "build" directories within the root path.
> Separate source and build directories (y/n) [n]: n

The project name will occur in several places in the built documentation.
> Project name: greetings
> Author name(s): The Alan Turing Institute
> Project release []: 0.0.1

If the documents are to be written in a language other than English,
you can select a language here by its language code. Sphinx will then
translate text that it generates into that language.

For a list of supported codes, see
https://www.sphinx-doc.org/en/master/usage/configuration.html#confval-language.
> Project language [en]: 

Creating file module06_software_projects/Greetings/docs/conf.py.
Creating file module06_software_projects/Greetings/docs/index.rst.
Creating file module06_software_projects/Greetings/docs/Makefile.
Creating file module06_software_projects/Greetings/docs/make.bat.

Finished: An initial directory structure has been created.

You should now populate your master file module06_software_projects/Greetings/docs/index.rst
and create other documentation source files. Use the Makefile to build the docs, like so:
make builder
where "builder" is one of the supported builders, e.g. html, latex or linkcheck.
```

and then look at and adapt the generated config, which in our case is a file called `conf.py` in the `docs/` directory of the project. This contains the project's Sphinx configuration, as Python variables. Let's populate the `extensions` field with some extensions we'd like to use (see the [extensions documentation](#)):

```
#Add any Sphinx extension module names here, as strings. They can be
#extensions coming with Sphinx (named 'sphinx.ext.*') or your custom
# ones.
extensions = [
    "sphinx.ext.autodoc", # Support automatic documentation
    "sphinx.ext.coverage", # Automatically check if functions are documented
    "sphinx.ext.mathjax", # Allow support for algebra
    "sphinx.ext.viewcode", # Include the source code in documentation
    "numpydoc",           # Support NumPy style docstrings
]
```

We've added some other configuration options to `conf.py` the file in the repo too (but normally you'll use `sphinx-quickstart`).

Define the root documentation page

Sphinx uses [RestructuredText](#) another wiki markup format similar to Markdown.

`sphinx-quickstart` creates a template `index.rst` for us, which can be edited to contain any preamble text you want. Here it is:

```
.. greetings documentation master file, created by
sphinx-quickstart on Thu Aug 4 11:47:51 2022.
You can adapt this file completely to your liking, but it should at least
contain the root 'toctree' directive.

Welcome to greetings's documentation!
=====
.. toctree::
:maxdepth: 2
:caption: Contents:

Indices and tables
=====
* :ref:`genindex`
* :ref:`modindex`
* :ref:`search`
```

And a lightly modified version:

```
%%writefile Greetings/docs/index.rst
Welcome to Greetings's documentation!
=====
Simple "Hello, James" module developed to teach research software engineering.

.. toctree::
:maxdepth: 2
:caption: Contents:

Functions
=====
.. autofunction:: greetings.greeter.greet

Indices and tables
=====
* :ref:`genindex`
* :ref:`modindex`
* :ref:`search`
```

Overwriting Greetings/docs/index.rst

Run sphinx

We can run Sphinx using:

```
%%bash
cd Greetings/
sphinx-build docs docs/output
```

```
Running Sphinx v4.5.0

making output directory... done

WARNING: html_static_path entry '_static' does not exist

[autosummary] generating autosummary for: index.rst

building [mo]: targets for 0 po files that are out of date

building [html]: targets for 1 source files that are out of date

updating environment: [new config] 1 added, 0 changed, 0 removed

reading sources... [100%] index

looking for now-outdated files... none found

pickling environment... done

checking consistency... done

preparing documents... done

writing output... [100%] index

generating indices... genindex done

highlighting module code... [100%] greetings.greeter

writing additional pages... search done

copying static files... done

copying extra files... done

dumping search index in English (code: en)... done

dumping object inventory... done

build succeeded, 1 warning.

The HTML pages are in docs/output.
```

Sphinx output

Sphinx's output is html, if you open the [Greetings/docs/output/index.html](#) file you'll see a simple documentation page for our `greetings` package has been created. We just created a simple single function's documentation, but Sphinx will create multiple nested pages of documentation automatically for many functions.

Hosting documentation

If you'd like to make your documentation available online two of the most popular (free) hosting services are [GitHub pages](#), and [Read the docs](#). Both can host documentation generated by Sphinx and have ways to automatically build and update your documentation when changes are made.

We have the example Greetings docs page on GitHub pages here: <https://alan-turing-institute.github.io/Greetings/>, which is built using [this GitHub Actions workflow](#).

6.6 Software Project Management

Estimated time for this notebook: 5 minutes

Software Engineering Stages

- Requirements
- Functional Design
- Architectural Design
- Implementation
- Integration

Requirements Engineering

Requirements capture obviously means describing the things the software needs to be able to do.

A common approach is to write down lots of "user stories", describing how the software helps the user achieve something:

As a clinician, when I finish an analysis, I want a report to be created on the test results, so that I can send it to the patient.

As a role, when condition or circumstance applies I want a goal or desire so that benefits occur.

These are easy to map into the Gherkin behaviour driven design test language.

Functional and architectural design

Engineers try to separate the functional design, how the software appears to and is used by the user, from the architectural design, how the software achieves that functionality.

Changes to functional design require users to adapt, and are thus often more costly than changes to architectural design.

Waterfall

The *Waterfall* design philosophy argues that the elements of design should occur in order: first requirements capture, then functional design, then architectural design. This approach is based on the idea that if a mistake is made in the design, then programming effort is wasted, so significant effort is spent in trying to ensure that requirements are well understood and that the design is correct before programming starts.

Why Waterfall?

Without a design approach, programmers resort to designing as we go, typing in code, trying what works, and making it up as we go along. When trying to collaborate to make software with others this can result in lots of wasted time, software that only the author understands, components built by colleagues that don't work together, or code that the programmer thinks is nice but that doesn't meet the user's requirements.

Problems with Waterfall

Waterfall results in a contractual approach to development, building an us-and-them relationship between users, business types, designers, and programmers.

I built what the design said, so I did my job.

Waterfall results in a paperwork culture, where people spend a long time designing standard forms to document each stage of the design, with less time actually spent *making things*.

Waterfall results in excessive adherence to a plan, even when mistakes in the design are obvious to people doing the work.

Software is not made of bricks

The waterfall approach to software engineering comes from the engineering tradition applied to building physical objects, where Architects and Engineers design buildings, and builders build them according to the design.

Software is intrinsically different:

Software is not the same 'stuff' as that from which physical systems are constructed. Software systems differ in material respects from physical systems. Much of this has been rehearsed by Fred Brooks in his classic '[No Silver Bullet](#)' paper. First, complexity and scale are different in the case of software systems: relatively functionally simple software systems comprise more independent parts, placed in relation to each other, than do physical systems of equivalent functional value. Second, and clearly linked to this, we do not have well developed components and composition mechanisms from which to build software systems (though clearly we are working hard on providing these) nor do we have a straightforward mathematical account that permits us to reason about the effects of composition.

Third, software systems operate in a domain determined principally by arbitrary rules about information and symbolic communication whilst the operation of physical systems is governed by the laws of physics. Finally, software is readily changeable and thus is changed, it is used in settings where our uncertainty leads us to anticipate the need to change.

– Prof. [Anthony Finkelstein](#), UCL Dean of Engineering, and Professor of Software Systems Engineering

The Agile Manifesto

In 2001, authors including Martin Fowler, Ward Cunningham and Kent Beck met in a Utah ski resort, and published the following manifesto.

[Manifesto for Agile Software Development](#)

We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to value:

- *Individuals and interactions* over processes and tools
- *Working software* over comprehensive documentation
- *Customer collaboration* over contract negotiation
- *Responding to change* over following a plan

That is, while there is value in the items on the right, we value the items on the left more.

Agile is not absence of process

The Agile movement is not anti-methodology, in fact, many of us want to restore credibility to the word methodology. We want to restore a balance. We embrace modeling, but not in order to file some diagram in a dusty corporate repository. We embrace documentation, but not hundreds of pages of never-maintained and rarely-used tomes. We plan, but recognize the limits of planning in a turbulent environment. Those who would brand proponents of XP or SCRUM or any of the other Agile Methodologies as "hackers" are ignorant of both the methodologies and the original definition of the term hacker

– Jim Highsmith.

Elements of an Agile Process

- Continuous delivery
- Self-organising teams
- Iterative development
- Ongoing design

Ongoing Design

Agile development doesn't eschew design. Design documents should still be written, but treated as living documents, updated as more insight is gained into the task, as work is done, and as requirements change.

Use of a Wiki or version control repository to store design documents thus works much better than using Word documents!

Test-driven design and refactoring are essential techniques to ensure that lack of "Big Design Up Front" doesn't produce badly constructed spaghetti software which doesn't meet requirements. By continuously scouring our code for smells, and stopping to refactor, we evolve towards a well-structured design with weakly interacting units. By starting with tests which describe how our code should behave, we create executable specifications, giving us confidence that the code does what it is supposed to.

Iterative Development

Agile development maintains a backlog of features to be completed and bugs to be fixed. In each iteration, we start with a meeting where we decide which backlog tasks will be attempted during the development cycle, estimating how long each will take, and selecting an achievable set of goals for the "sprint". At the end of each cycle, we review the goals completed and missed, and consider what went well, what went badly, and what could be improved.

We try not to add work to a cycle mid-sprint. New tasks that emerge are added to the backlog, and considered in the next planning meeting. This reduces stress and distraction.

Continuous Delivery

In agile development, we try to get as quickly as possible to code that can be *demonstrated* to clients. A regular demo of progress to clients at the end of each development iteration says so much more than sharing a design document. "Release early, release often" is a common slogan. Most bugs are found by people *using* code – so exposing code to users as early as possible will help find bugs quickly.

Self-organising teams

Code is created by people. People work best when they feel ownership and pride in their work. Division of responsibilities into designers and programmers results in a "[Code Monkey](#)" role, where the craftsmanship and sense of responsibility for code quality is lost. Agile approaches encourage programmers, designers, clients, and businesspeople to see themselves as one team, working together, with fluid roles. Programmers grab issues from the backlog according to interest, aptitude, and community spirit.

Agile in Research

Agile approaches, where we try to turn the instincts and practices which emerge naturally when smart programmers get together into well-formulated best practices, have emerged as antidotes to both the chaotic free-form typing in of code, and the rigid paperwork-driven approaches of Waterfall.

If these approaches have turned out to be better even in industrial contexts, where requirements for code can be well understood, they are even more appropriate in a research context, where we are working in poorly understood fields with even less well captured requirements.

Conclusion

- Don't ignore design
- See if there's a known design pattern that will help
- Do try to think about how your code will work before you start typing
- Do use design tools like UML to think about your design without coding straight away
- Do try to write down some user stories
- Do maintain design documents.

BUT

- Do change your design as you work, updating the documents if you have them
- Don't go dark – never do more than a couple of weeks programming without showing what you've done to colleagues
- Don't get isolated from the reasons for your code's existence, stay involved in the research, don't be a Code Monkey.
- Do keep a list of all the things your code needs, estimate and prioritise tasks carefully.

6.7 Software Licensing

Estimated time for this notebook: 10 minutes

Disclaimer

Here we attempt to give some basic advice on choosing a license for your software. But:

- we are NOT lawyers
- opinions differ (and flamewars are boring)
- this training does NOT constitute legal advice.

For an in-depth discussion of software licenses, read the [O'Reilly book](#).

Your organisation may have policies about applying licenses to code you create while you work there. This training doesn't address this issue, and does not represent an official policy – seek advice from your supervisor or manager if concerned.

Choose a license

It is important to choose a license and to create a *license file* to tell people what it is.

The license lets people know whether they can reuse your code and under what terms. [This course has one](#), for example.

Your license file should typically be called LICENSE.txt or similar. GitHub will offer to create a license file automatically when you create a new repository.

See [GitHub's advice on how to choose a license](#)

Open source doesn't stop you making money

A common misconception about open source software is the thought that open source means you can't make any money. This is *wrong*.

Plenty of people open source their software and profit from:

- The software under a different license e.g. [Saxon](#)
- Consulting. For example: [Continuum](#) who help maintain NumPy
- Manuals. For example: [VTK](#)
- Add-ons. For example: [Puppet](#)
- Server software, which open source client software interacts with. For example: [GitHub API clients](#)

Plagiarism vs promotion

Many researchers worry about people stealing their work if they open source their code. But often the biggest problem is not theft, but the fact no one is aware of your work.

Open source is a way to increase the probability that someone else on the planet will care enough about your work to cite you.

So when thinking about whether to open source your code, think about whether you're more worried about anonymity or theft.

Your code is good enough

New coders worry that they'll be laughed at if they put their code online. Don't worry. Everyone, including people who've been coding for decades, writes shoddy code that is full of bugs.

The only thing that will make your code better, is *other people reading it*.

For small scripts that no one but you will ever use, my recommendation is to use an open repository anyway. Find a buddy, and get them to comment on it.

Worry about license compatibility and proliferation

Not all open source code can be used in all projects. Some licenses are legally incompatible.

This is a huge and annoying problem. As an author, you might not care, but you can't anticipate the exciting uses people might find by mixing your code with someone else's.

Use a standard license from the small list that are well-used. Then people will understand. *Don't make up your own*.

When you're about to use a license, see if there's a more common one which is recommended, e.g.: using the [opensource.org.proliferation.report](#)

Academic license proliferation

Academics often write their own license terms for their software.

For example:

```
XXXX NON-COMMERCIAL EDUCATIONAL LICENSE Copyright (c) 2013 Prof. Foo. All rights reserved.  
You may use and modify this software for any non-commercial purpose within your educational institution. Teaching, academic research, and personal experimentation are examples of purpose which can be non-commercial.  
You may redistribute the software and modifications to the software for non-commercial purposes, but only to eligible users of the software (for example, to another university student or faculty to support joint academic research).
```

Please don't do this. Your desire to slightly tweak the terms is harmful to the future software ecosystem. Also, *Unless you are a lawyer, you cannot do this safely!*

Licenses for code, content, and data.

Licenses designed for code should not be used to license data or prose.

Don't use Creative Commons for software, or GPL for a book.

Licensing issues

- Permissive vs share-alike
- Non-commercial and academic Use Only
- Patents
- Use as a web service

Permissive vs share-alike

Some licenses require all derived software to be licensed under terms that are similarly free. Such licenses are called "Share Alike" or "Copyleft".

- Licenses in this class include the GPL.

Those that don't are called "Permissive"

- These include Apache, BSD, and MIT licenses.

If you want your code to be maximally reusable, use a permissive license. If you want to force other people using your code to make derivatives open source, use a copyleft license.

If you want to use code that has a permissive license, it's safe to use it and keep your code secret. If you want to use code that has a copyleft license, you'll have to release your code under such a license.

Academic use only

Some researchers want to make their code free for 'academic use only'. None of the standard licenses state this, and this is a reason why academic bespoke licenses proliferate.

However, there is no need for this, in our opinion.

Use of a standard Copyleft license precludes derived software from being sold without also publishing the source

So use of a Copyleft license precludes commercial use.

This is a very common way of making a business from open source code: offer the code under GPL for free but offer the code under more permissive terms, allowing for commercial use, for a fee.

Patents

Intellectual property law distinguishes copyright from patents. This is a complex field, which I am far from qualified to teach!

People who think carefully about intellectual property law distinguish software licenses based on how they address patents. Very roughly, if you want to ensure that contributors to your project can't then go off and patent their contribution, some licenses, such as the Apache license, protect you from this.

Use as a web service

If I take copyleft code, and use it to host a web service, I have not sold the software.

Therefore, under some licenses, I do not have to release any derivative software. This "loophole" in the GPL is closed by the AGPL ("Affero GPL")

Library linking

If I use your code just as a library, without modifying it or including it directly in my own code, does the copyleft term of the GPL apply?

Yes

If you don't want it to, use the LGPL. ("Lesser GPL"). This has an exception for linking libraries.

Referencing the license in every file

Some licenses require that you include license information in every file. Others do not.

Typically, every file should contain something like:

```
# (C) The Alan Turing Institute 2010-2020
# This software is licensed under the terms of the <foo> license
# See <somewhere> for the license details.
```

Check your license at opensource.org for details of how to apply it to your software. For example, for the [GPL](#).

Citing software

Almost all software licenses require people to credit you for what they used ("attribution").

In an academic context, it is useful to offer a statement as to how best to do this, citing *which paper to cite in all papers which use the software*.

This is best done with a [CITATION](#) file in your repository.

To cite ggplot2 in publications, please use:

H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York,

1.

A BibTeX entry for LaTeX users is

```
@Book{, author = {Hadley Wickham}, title = {ggplot2: elegant graphics for data analysis}, publisher = {Springer New York}, year = {2009}, isbn = {978-0-387-98140-6}, url = {http://had.co.nz/ggplot2/book}, }
```

Publishing software

If you'd like to make your software more easily citable, there are a few options for creating software papers and DOIs. These include:

- Software journals such as [The Journal of Open Source Software \(JOSS\)](#), which publishes software with a short paper/codebase description attached,
- File hosting services like [Zenodo](#), which will generate a DOI you can use to link to a specific version of your code.

Open source does not equal free maintenance

One common misunderstanding of open source software is that you'll automatically get loads of contributors from around the internets. This is wrong. Most open source projects get no commits from anyone else.

Open source does *not* guarantee your software will live on with people adding to it after you stop working on it.

Learn more about these issues from the website of the [Software Sustainability Institute](#)

Example

This course is distributed under the [Creative Commons By Attribution license](#), which means you can modify and reuse the materials, so long as you credit the original authors: [The Alan Turing Institute's Research Engineering Group](#) and [UCL Research IT Services](#).

6.8 Managing software issues

Estimated time for this notebook: 5 minutes

Issues

Code has *bugs*. It also has *features*, things it should do.

A good project has an organised way of managing these. Generally you should use an issue tracker.

Some Issue Trackers

There are lots of good issue trackers.

The most commonly used open source ones are [Trac](#) and [Redmine](#).

Cloud based issue trackers include [Lighthouse](#) and [GitHub](#).

Commercial solutions include [Jira](#).

In this course, we'll be using the GitHub issue tracker.

Anatomy of an issue

- Reporter
- Description
- Owner
- Type [Bug, Feature]
- Component
- Status
- Severity

Reporting a Bug

The description should make the bug reproducible:

- Version
- Steps

If possible, submit a minimal reproducing code fragment.

Owning an issue

- Whoever the issue is assigned to works next.
- If an issue needs someone else's work, assign it to them.

Status

- Submitted
- Accepted
- Underway
- Blocked

Resolutions

- Resolved
- Will Not Fix
- Not reproducible
- Not a bug (working as intended)

Bug triage

Some organisations use a severity matrix based on:

- Severity [Wrong answer, crash, unusable, workaround, cosmetic...]
- Frequency [All users, most users, some users...]

The backlog

The list of all the bugs that need to be fixed or features that have been requested is called the "backlog".

Development cycles

Development goes in *cycles*.

Cycles range in length from a week to three months.

In a given cycle:

- Decide which features should be implemented
- Decide which bugs should be fixed
- Move these issues from the Backlog into the current cycle. (Aka Sprint)

GitHub issues

GitHub doesn't have separate fields for status, component, severity etc. Instead, it just has labels, which you can create and delete.

See for example [Jupyter](#)

6.9 Exercise: Packaging Troll Treasure

We are going to look at a simplified version of a game with a [long history](#). Games of this kind have been used as test-beds for development of artificial intelligence.

A *dungeon* is a network of connected *rooms* on a square grid. One or more rooms contain *treasure*. Your character, the *adventurer*, moves between rooms, looking for the treasure. A *troll* is also in the dungeon and moves between rooms. If the troll catches the adventurer, you lose. If you find treasure before being eaten, you win. (In this simple version, we do not consider the need to leave the dungeon.)

The starting rooms for the adventurer and troll are given in the definition of the dungeon.

The way the adventurer and troll move is called a *strategy*. Different strategies are more or less likely to succeed. There are two strategies in the provided code - random movement, and movement controlled by human input.

The code provides a function to play a single game, or to simulate many games and estimate the probability the adventurer or troll wins.

In this exercise, you will convert the code provided in this Jupyter notebook into a proper Python package.

What to do

Using the course material from this module to help:

1. Briefly familiarise yourself with the code below and how it works/runs, focusing on the different classes and functions that are defined rather than the implementation details. And try running a game in the notebook (see the [Playing Games](#) section).
2. Make a directory for your project
3. Create a directory to contain your package (in your project directory) and copy the code from this notebook to it using multiple `.py` files. Remember to add `__init__.py` file(s).
4. Create a `pyproject.toml` file to specify the metadata and dependencies for the package. The dependencies are [PyYAML](#) version 6.0 and [art](#) version 5.7.
5. Create a virtual environment and activate it
6. Install the package in your virtual environment.
7. Check that you can import your package and run a script to play a game.
8. Make a command-line interface:
 - Use `argparse` to create a function that can be called from the command-line with the path to a dungeon YAML file, and the option to either run a single game or calculate probabilities.
 - Add a script to call your function to the `[project.scripts]` section of `pyproject.toml`
 - Reinstall your package and verify the executable you defined can now be run from a terminal.

Extensions

If you'd like to take this further here are some other things you could try, depending on your interests:

9. Create a GitHub repo for your project. Try installing the package from GitHub.

10. Add `README`, `LICENSE`, and `CITATION` files to your project directory.

11. Improve documentation

- Add docstrings to functions to explain what they do, their inputs, and outputs (e.g. in numpydoc format)
- Build a documentation web-site with sphinx
- Host the documentation on GitHub pages

12. Add unit tests

- Write unit tests, such as to verify the outcomes of the provided `test_xxx.yml` dungeon files.
- Add `pytest` as a development dependency for your package (and install it)
- Run the tests and verify they pass

13. Improve or extend the code (you may like to revisit this after the next module):

- Specify your own dungeons or create a function to auto-generate them
- Create new movement behaviour classes for adventurers and trolls
- Look for places the code could be refactored (to reduce repetition, for example)

Code

Rooms

The `direction` function determines the direction from grid point `a` to grid point `b` (length two tuples, x and y coordinate) if they are neighbouring points in the grid, or `None` otherwise:

```
def direction(point_a, point_b):
    """
    Returns the direction from point_a to point_b, or None if they
    are not neighbouring grid points.
    """
    if point_b == point_a:
        return "nowhere"
    if point_b[1] == point_a[1]:
        if point_b[0] == point_a[0] - 1:
            return "left"
        if point_b[0] == point_a[0] + 1:
            return "right"
    if point_b[0] == point_a[0]:
        if point_b[1] == point_a[1] - 1:
            return "up"
        if point_b[1] == point_a[1] + 1:
            return "down"
    return None
```

A `Room` has a location (point) and optionally can link to other `Rooms` at neighbouring grid points. :

```
class Room:
    def __init__(self, point, links=None):
        self.point = tuple(point) # grid point of this room
        self.links = links # other rooms this room connects to
        self._validate_links()

    def __contains__(self, point):
        """
        `(x, y) in room_instance` returns `True` if `room_instance` has a link to
        a room at point `(x, y)`
        """
        return point in [link.point for link in self.links]

    def _validate_links(self):
        """
        Verifies all linked rooms are at neighbouring grid points
        """
        if not self.links:
            return
        for link in self.links:
            if not direction(self.point, link.point):
                raise ValueError(
                    f"Invalid link: {link.point} is not connected to {self.point}"
                )
```

`Rooms` is a collection of rooms (that must be on a square grid), with each room stored in a dictionary keyed by its `(x, y)` coordinate (point):

```
class Rooms:
    """
    Collection of rooms
    """

    def __init__(self, rooms):
        # rooms dictionary keyed by (x, y) coordinate (grid cell indices)
        self.rooms = {r.point: r for r in rooms}

    def __iter__(self):
        """
        Allows Rooms objects to be iterated over (see Module 7)
        """
        return iter(self.rooms.values())

    def __getitem__(self, point):
        """
        rooms[(x, y)] will retrieve the room at coordinate (x, y) (where
        rooms is an instance of the Rooms class)
        """
        return self.rooms[point]

    def __contains__(self, point):
        """
        (x, y) in rooms will return True if a room at coordinates (x, y)
        is in rooms (where rooms is an instance of the Rooms class)
        """
        return point in self.rooms

    @classmethod
    def from_list(cls, room_list):
        rooms = [
            Room(room["point"], [Room(link) for link in room["links"]])
            for room in room_list
        ]
        return cls(rooms)
```

Treasure

Treasure has a location (point) and a single character symbol to represent it when printing dungeon maps:

```
class Treasure:
    def __init__(self, point, symbol):
        self.point = tuple(point) # (x, y) grid location of the treasure
        self.symbol = symbol # single char symbol to show the treasure on dungeon
        maps

    @classmethod
    def from_dict(cls, treasure_dict):
        return cls(treasure_dict["point"], treasure_dict["symbol"])
```

Agents

The `Agent` class stores general properties used by all agents (trolls or adventurers):

```
class Agent:
    """
    Base functionality to create and load (but not move) an Agent
    """

    def __init__(self, point, name, symbol, verbose=True, allow_wait=True,
                 **kwargs):
        self.point = tuple(point) # (x, y) grid location of the agent
        self.name = name # e.g. adventurer or troll
        self.symbol = symbol # single char symbol to show the agent on dungeon
        maps

        self.verbose = verbose # print output on agent behaviour if True
        self.allow_wait = allow_wait # allow the agent to move nowhere

    def move(self, rooms):
        raise NotImplementedError("Use an Agent base class")

    @classmethod
    def from_dict(cls, agent_dict):
        return cls(
            agent_dict["point"],
            agent_dict["name"],
            agent_dict["symbol"],
            allow_wait=agent_dict["allow_wait"],
        )
```

`RandomAgent`s choose to move to a connecting room at random (or stay where they are):

```
import random

class RandomAgent(Agent):
    """
    Agent that makes random moves
    """

    def move(self, rooms):
        if not rooms[self.point].links:
            # this room isn't linked to anything, can't move
            if self.verbose:
                print(f"{self.name} is trapped")
            return

        # pick a random room to move to
        options = rooms[self.point].links
        if self.allow_wait:
            options.append(self)
        new_room = random.choice(options)

        if self.verbose:
            move = direction(self.point, new_room.point)
            print(f"{self.name} moves {move}")
        self.point = new_room.point
```

`HumanAgent`s ask the user where to move next:

```
class HumanAgent(Agent):
    """
    Agent that prompts the user where to move next
    """

    def move(self, rooms):
        if not rooms:
            if self.verbose:
                print(f"{self.name} is trapped")
            return
        # populate movement options depending on available rooms
        if self.allow_wait:
            options = ["wait"]
        else:
            options = []
            if (self.point[0] - 1, self.point[1]) in rooms:
                options.append("left")
            if (self.point[0] + 1, self.point[1]) in rooms:
                options.append("right")
            if (self.point[0], self.point[1] - 1) in rooms:
                options.append("up")
            if (self.point[0], self.point[1] + 1) in rooms:
                options.append("down")

        # prompt user for movement input
        choice = None
        while choice not in options:
            choice = input(f"Where will {self.name} move \n{options}? ")

        # move the agent
        if choice == "left":
            self.point = (self.point[0] - 1, self.point[1])
        elif choice == "right":
            self.point = (self.point[0] + 1, self.point[1])
        elif choice == "up":
            self.point = (self.point[0], self.point[1] - 1)
        elif choice == "down":
            self.point = (self.point[0], self.point[1] + 1)
```

Dungeons

`Dungeon` have rooms, a piece of treasure, an adventurer, and a troll, and provide the functionality to update the agents, check whether the treasure or adventurer have found, and to draw a map of the dungeon:

```

import yaml

class Dungeon:
    """
    Dungeon with:
    - Connected set of rooms on a square grid
    - The location of some treasure
    - An adventurer agent with an initial position
    - A troll agent with an initial position
    """

    def __init__(self, rooms, treasure, adventurer, troll, verbose=True):
        self.rooms = rooms
        self.treasure = treasure
        self.adventurer = adventurer
        self.troll = troll
        self.verbose = True

        # the extent of the square grid
        self.xlim = (
            min(r.point[0] for r in self.rooms),
            max(r.point[0] for r in self.rooms),
        )
        self.ylim = (
            min(r.point[1] for r in self.rooms),
            max(r.point[1] for r in self.rooms),
        )

        self._validate()

    def _validate(self):
        if self.treasure.point not in self.rooms:
            raise ValueError(f"Treasure({self.treasure.point}) is not in the dungeon")
        if self.adventurer.point not in self.rooms:
            raise ValueError(
                f"{self.adventurer.name}({treasure.point}) is not in the dungeon"
            )
        if self.troll.point not in self.rooms:
            raise ValueError(f"{self.troll.name}({treasure.point}) is not in the dungeon")

    @classmethod
    def from_file(cls, path):
        with open(path) as f:
            spec = yaml.safe_load(f)

        rooms = Rooms.from_list(spec["rooms"])
        treasure = Treasure.from_dict(spec["treasure"])

        agent_keys = ["adventurer", "troll"]
        agents = {}
        for agent in agent_keys:
            if spec[agent]["type"] == "random":
                agent_class = RandomAgent
            elif spec[agent]["type"] == "human":
                agent_class = HumanAgent
            else:
                raise ValueError(f"Unknown agent type {spec[agent]['type']}")
            agents[agent] = agent_class(**spec[agent])

        return cls(rooms, treasure, agents["adventurer"], agents["troll"])

    def update(self):
        """
        Move the adventurer and the troll
        """
        self.adventurer.move(self.rooms)
        self.troll.move(self.rooms)
        if self.verbose:
            print()
            self.draw()

    def outcome(self):
        """
        Check whether the adventurer found the treasure or the troll
        found the adventurer
        """
        if self.adventurer.point == self.troll.point:
            return -1
        if self.adventurer.point == self.treasure.point:
            return 1
        return 0

    def set_verbose(self, verbose):
        """
        Set whether to print output"""
        self.verbose = verbose
        self.adventurer.verbose = verbose
        self.troll.verbose = verbose

    def draw(self):
        """
        Draw a map of the dungeon"""
        layout = ""

        for y in range(self.ylim[0], self.ylim[1] + 1):
            for x in range(self.xlim[0], self.xlim[1] + 1):
                # room and character symbols
                if (x, y) in self.rooms:
                    if self.troll.point == (x, y):
                        layout += self.troll.symbol
                    elif self.adventurer.point == (x, y):
                        layout += self.adventurer.symbol
                    elif self.treasure.point == (x, y):
                        layout += self.treasure.symbol
                    else:
                        layout += "o"
                else:
                    layout += " "

                # horizontal connections
                if ((x, y) in self.rooms) and (((x + 1), y) in self.rooms[(x, y)]):
                    layout += " - "
                else:
                    layout += "   "

            # vertical connections
            if y < self.ylim[1]:
                layout += "\n"
            for x in range(self.xlim[0], self.xlim[1] + 1):
                if ((x, y) in self.rooms) and (((x, y + 1) in self.rooms[(x, y)])):
                    layout += "|"
                else:
                    layout += " "
                if x < self.xlim[1]:
                    layout += " "
                layout += "\n"

```

```
    print(layout)
```

Games

The `Game` class runs a dungeon until the adventurer or troll wins, or calls a draw if neither wins in a given number of steps, and can simulate many games to estimate outcome probabilities:

```
import copy
from art import tprint

class Game:
    def __init__(self, dungeon):
        self.dungeon = dungeon

    def preamble(self):
        tprint("Troll Treasure\n", font="small")
        print(
            f"""
The {self.dungeon.adventurer.name} is looking for treasure in a mysterious
dungeon.
Will they succeed or be dinner for the {self.dungeon.troll.name} that lurks there?

The map of the dungeon is below:
o : an empty room
o - o : connected rooms
{self.dungeon.troll.symbol} : {self.dungeon.troll.name}
{self.dungeon.adventurer.symbol} : {self.dungeon.adventurer.name}
{self.dungeon.treasure.symbol} : the treasure
"""
        )

    def run(self, max_steps=1000, verbose=False):
        dungeon = copy.deepcopy(self.dungeon)
        dungeon.set_verbose(verbose)
        if verbose:
            self.preamble()
            dungeon.draw()
            if start_prompt:
                input("\nPress enter to continue...")
            else:
                print("\nLet the hunt begin!")

        for turn in range(max_steps):
            result = dungeon.outcome()
            if result != 0:
                if verbose:
                    if result == 1:
                        print(
                            f"\n{self.dungeon.adventurer.name} gets the treasure
and returns a hero!"
                        )
                    tprint("WINNER", font="small")
                elif result == -1:
                    print(f"\n{self.dungeon.troll.name} will eat tonight!")
                    tprint("GAME OVER", font="small")
            return result
        if verbose:
            print(f"\nTurn {turn + 1}")
        dungeon.update()
        # no outcome in max steps (e.g. no treasure and troll can't reach
adventurer)
        if verbose:
            print(
                f"\nNo one saw {self.dungeon.adventurer.name} or
{self.dungeon.troll.name} again."
            )
            tprint("STALEMATE", font="small")

        return result

    def probability(self, trials=10000, max_steps=1000, verbose=False):
        outcomes = {-1: 0, 0: 0, 1: 0}
        for _ in range(trials):
            result = self.run(max_steps=max_steps, verbose=False)
            outcomes[result] += 1
        for result in outcomes:
            outcomes[result] = outcomes[result] / trials
        return outcomes
```

Playing Games

Dungeon YAML files

Dungeons can be specified with YAML files, and we have provided a few examples in the `dungeons/` directory:

```
import os
os.listdir("dungeons")
```

```
['dungeon.yml',
'test_stalemate.yml',
'test_win50.yml',
'test_lose100.yml',
'test_win100.yml']
```

The files starting `test_` are simpler dungeons with known outcome probabilities (which could be used in unit testing, for example).

Run a single game

```
d = Dungeon.from_file("dungeons/dungeon.yml")
g = Game(d)
g.run(max_steps=10)
```

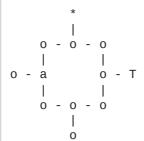


The Adventurer is looking for treasure in a mysterious dungeon.
Will they succeed or be dinner for the Troll that lurks there?

The map of the dungeon is below:

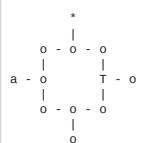
o : an empty room
o - o : connected rooms
T : Troll
a : Adventurer

* : the treasure

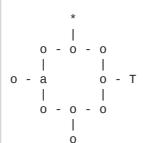


Let the hunt begin!

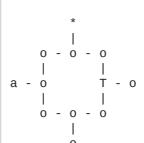
Turn 1
Adventurer moves left
Troll moves left



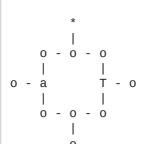
Turn 2
Adventurer moves right
Troll moves right



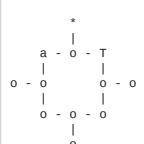
Turn 3
Adventurer moves left
Troll moves left



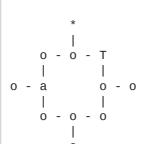
Turn 4
Adventurer moves right
Troll moves nowhere



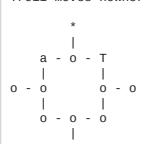
Turn 5
Adventurer moves up
Troll moves up



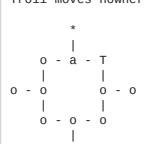
Turn 6
Adventurer moves down
Troll moves nowhere



Turn 7
Adventurer moves up
Troll moves nowhere



Turn 8
Adventurer moves right
Troll moves nowhere



Turn 9
Adventurer moves right
Troll moves nowhere

*

-1

Estimate outcome probabilities

```
d = Dungeon.from_file("dungeons/dungeon.yml")
g = Game(d)
g.probability(max_steps=10)
# -1: troll wins, 0: stalemate, +1: adventurer wins
```

```
{-1: 0.2204, 0: 0.5996, 1: 0.18}
```

7. Construction and Design

- Comments
 - Coding conventions
 - Linters
 - Refactoring
 - Object Orientation
 - Design Patterns

Contents

- [7.0 Construction](#) (5 minutes)
 - [7.1 Comments](#) (15 minutes)
 - [7.2 Coding conventions](#) (10 minutes)
 - [7.3 Linting](#) (15 minutes)
 - [7.4 Refactoring](#) (25 minutes)
 - [7.5 Object-Oriented Design](#) (15 minutes)
 - [7.6 Class design](#) (25 minutes)
 - [7.7 Design Patterns](#) (25 minutes)

Total time: 2 hr 2 15 minutes

Exercises

A classroom exercise is included at the end of the module: [7.8 Exercise: Refactoring The Bad Boids](#). We recommend that instructors arrange for the exercise to be done in groups. The exercise can also be left as a self-paced homework assignment if preferred.

7.0 Construction

Estimated time for this notebook: 5 minutes

Construction

Software *design* gets a lot of press (Object orientation, UML, design patterns).

In this session we're going to look at advice on software construction

Construction vs Design

For a given piece of code, there exist several different ways one could write it:

- Choice of variable names
 - Choice of comments
 - Choice of layout

The consideration of these questions is the area of Software Construction

Low-level design decisions

We will also look at some of the lower-level software design decisions in the context of this section:

- Division of code into subroutines
 - Subroutine access signatures
 - Choice of data structures for readability

Algorithms and structures

We will not, in discussing construction, be looking at decisions as to how design questions impact performance:

- Choice of algorithms
 - Choice of data structures for performance
 - Choice of memory layout

We will consider these in a future discussion of performance programming.

Architectural design

We will not, in this session, be looking at the large-scale questions of how program components interact, the strategic choices that govern how software is built, or how to make it more efficient.

- Where do objects get made?

- Which objects own or access other objects?
- How can I hide complexity in one part of the code from other parts of the code?

We will consider these in a future session.

Construction

So, we've excluded most of the exciting topics. What's left is the bricks and mortar of software: how letters and symbols are used to build code which is readable.

Literate programming

In literature, books are enjoyable for different reasons:

- The beauty of stories
- The beauty of plots
- The beauty of characters
- The beauty of paragraphs
- The beauty of sentences
- The beauty of words

Software has beauty at these levels too: stories and characters correspond to architecture and object design, plots corresponds to algorithms, but the rhythm of sentences and the choice of words corresponds to software construction.

Programming for humans

- Remember you're programming for humans as well as computers
- A program is the best, most rigorous way to describe an algorithm
- Code should be pleasant to read, a form of scholarly communication

Read Steve McConnell's [Code Complete](#).

7.1 Comments

Estimated time for this notebook: 15 minutes

Why comment?

- You're writing code for people, as well as computers.
- Comments can help you build code, by representing your design
- Comments explain subtleties in the code which are not obvious from the syntax
- Comments explain *why* you wrote the code the way you did

The Pseudocode Programming Process

Start by writing a program in all comments:

```
# To find the largest element in an array
# Set up a variable to track the largest so far
# Loop over every element
# - For each element, is it bigger than the previous biggest?
# - If so, it's the new biggest
# At the end, the biggest so far, is the biggest overall
```

One by one, replace these with the equivalent in code

```
# To find the largest element in an array
def largest(data):
    # Set up a variable to track the largest so far
    biggest_so_far = 0
    # Loop over every element
    for datum in data:
        # For each element, is it bigger than the previous biggest?
        if datum > biggest_so_far:
            # If so, it's the new biggest
            biggest_so_far = datum
    # At the end, the biggest so far, is the biggest overall
    return biggest_so_far
```

```
largest([0, 1, 3, 6, 2, 5, 3])
```

6

Then, remove only those comments that are now extraneous (see below for examples of extraneous comments)

```
# To find the largest element in an array
def largest(data):
    # Set up a variable to track the largest so far
    biggest_so_far = 0
    for datum in data:
        # For each element, is it bigger than the previous biggest?
        # If so, it's the new biggest
        if datum > biggest_so_far:
            biggest_so_far = datum
    return biggest_so_far
```

Who are you writing for?

- By far the most likely person who will read your code/comments is yourself, maybe in a week's time, or maybe in six months time.
- Second most likely person in most cases, is someone in your team, or someone else who will probably have a roughly similar level of expertise, and be trying to do a similar thing.
- Write comments with this in mind - try to help the person reading the code to understand *what* you did and *why*.

Prefer "in language" comments to comments proper, if we can

More comments doesn't necessarily mean *better* - here are some examples of comments that don't really help the reader understand the code any better. If we can, it's nice to find ways to put our description of what the code does *inside* the code, instead of as comments. Then, when the code changes, the 'comments' stay in sync, because they're part of the code.

For example, we can use a variable name or a function name, to hold what would have been in a comment. Here, instead of a comment and a one-word function name, we've made a longer function name.

```

def largest_element_in_array(data):
    # Set up a variable to track the largest so far
    biggest_so_far = 0
    for datum in data:
        # For each element, is it bigger than the previous biggest?
        # If so, it's the new biggest!
        if datum > biggest_so_far:
            biggest_so_far = datum
    return biggest_so_far

```

Comments which are obvious

Try to use comments to explain why the code does, not just repeat the code in a comment.

```

counter = counter + 1 # Increment the counter
for element in array: # Loop over elements
    pass

```

Comments which could be replaced by better style

The following piece of code could be a part of a game to move a turtle in a certain direction, with a particular angular velocity and step size.

```

for i in range(len(agt)): # for each agent
    agt[i].theta += ws[i] # Increment the angle of each agent
    # by its angular velocity
    agt[i].x += r * sin(agt[i].theta) # Move the agent by the step-size
    agt[i].y += r * cos(agt[i].theta) # r in the direction indicated

```

we have used comments to make the code readable.

Why not make the code readable instead?

```

for agent in agents:
    agent.turn()
    agent.move()

class Agent:
    def turn(self):
        self.direction += self.angular_velocity

    def move(self):
        self.x += Agent.step_length * sin(self.direction)
        self.y += Agent.step_length * cos(self.direction)

```

This is probably better. We are using the name of the functions (i.e. `turn`, `move`) instead of comments. Therefore, we've got *self-documenting* code.

Comments which belong in an issue tracker

```

x.clear() # Code crashes here sometimes

class Agent:
    pass
    # TODO: Implement pretty-printer method

```

BUT comments that reference issues in the tracker can be good.

E.g.

```

if x.safe_to_clear(): # Guard added as temporary workaround for #32
    x.clear()

```

is OK. And platforms like GitHub will create a link to it when browsing the code.

Comments which only make sense to the author today

```

agent.turn() # Turtle Power!
agent.move()
agents[:] = [] # Shredder!

```

Comments which are unpublishable

```

# Stupid supervisor made me write this code
# So I did it while very very drunk.

```

Good commenting: pedagogical comments

Code that is good style, but you're not familiar with, or that colleagues might not be familiar with

```

# This is how you define a decorator in python
# See https://wiki.python.org/moin/PythonDecorators
def double(decorated_function):
    # Here, the result function forms a closure over
    # the decorated function
    def result_function(entry):
        return decorated_function(decorated_function(entry))

    # The returned result is a function
    return result_function

@double
def try_me_twice():
    pass

```

Great commenting: reasons and definitions

Comments which explain coding definitions or reasons for programming choices.

```

def __init__(self):
    self.angle = 0 # clockwise from +ve y-axis
    nonzero_indices = [] # Use sparse model as memory constrained

```

Are comments always helpful?

Some authors argue that comments can be dangerous, as they can disincentivise us from trying harder to use variable names and function names to describe the code:

The proper use of comments is to compensate for our failure to express yourself in code. Note that I used the word failure. I meant it. Comments are always failures. – Robert Martin, Clean Code

This is definitely taking things too far, but there's a little grain of truth in it:

7.2 Coding conventions

Estimated time for this notebook: 10 minutes

One code, many layouts:

Consider the following fragment of python:

```
import species

def AddToReaction(name, reaction):
    reaction.append(species.Species(name))
```

this could also have been written:

```
from species import Species

def add_to_reaction(a_name, a_reaction):
    l_species = Species(a_name)
    a_reaction.append(l_species)
```

So many choices

- Layout
- Naming
- Syntax choices

Layout

```
{ reaction = {"reactants": ["H", "H", "O"], "products": ["H2O"]}

reaction2 = {"reactants": ["H", "H", "O"], "products": ["H2O"]}
```

Layout choices

- Brace style
- Line length
- Indentation
- Whitespace/Tabs

Inconsistency will produce a mess in your code! Some choices will make your code harder to read, whereas others may affect the code. For example, if you copy/paste code with tabs in a place that's using spaces, they may appear OK in your screen but it will fail when running it.

Naming Conventions

[Camel case](#) is used in the following example, where class name is in UpperCamel, functions in lowerCamel and underscore_separation for variables names:

```
class ClassName:
    def methodName(self, variable_name):
        self.instance_variable = variable_name
```

This example uses underscore_separation for all the names:

```
class class_name:
    def method_name(self, a_variable):
        self.m_instance_variable = a_variable
```

The usual Python convention (see [PEP8](#)) is UpperCamel for class names, and underscore_separation for function and variable names:

```
class ClassName:
    def method_name(self, variable_name):
        self.instance_variable = variable_name
```

However, particular projects may have their own conventions (and you will even find Python standard libraries that don't follow these conventions).

Newlines

- Newlines make code easier to read
- Newlines make less code fit on a screen

Use newlines to describe your code's *rhythm*.

Syntax Choices

The following two snippets do the same, but the second is separated into more steps, making it more readable.

```
big = True
fast = False
color = "brown"
cheap = True

if color == "red" and fast or big and cheap:
    print("Vroom!")
```

Vrroom!

```
exciting = color == "red" and fast
practical = big and cheap
```

```
if exciting or practical:
    print("Vrroom!")
```

```
Vrroom!
```

We create extra variables as an intermediate step. Don't worry about the performance now, the compiler will do the right thing.

What about operator precedence? Being explicit helps to remind yourself what you are doing.

- Explicit operator precedence
- Compound expressions
- Package import choices

Type Annotations

Python is *dynamically typed*, which means if a variable `x` is an integer:

```
x = 32
```

it is valid in Python to make it into a string or any other type later:

```
x = "bananas"
```

This is not the case in a *statically typed* language, like C++ or Java. Having this flexibility in Python can be convenient but it can also lead to unexpected, and potentially difficult to diagnose, mistakes if variables in your code have different types to what was expected.

For example, consider the following function:

```
def repeat(x, y, times=2):
    return (x + y) * times
```

```
repeat("dog", "woof")
```

```
'dogwoofdogwoof'
```

It looks like a function that repeats its inputs a number of times, but what if the inputs are numbers?

```
repeat(2, 3, times=3)
```

```
15
```

Ah, that's not what we wanted (we were hoping for 232323).

To help us remember how the function is supposed to be used, we can add type annotations (or type "hints"):

```
def repeat(x: str, y: str, times: int = 3) -> str:
    return (x + y) * times
```

The syntax `variable_name: type` indicates the type each parameter should have (`x` and `y` are strings, and `times` is an integer), and the arrow syntax in `function_name(...) -> type` indicates the type of data the function returns (a string for the `repeat` function above).

Note that type annotating your code will not change its behaviour (Python does not enforce variables to be their annotated types):

```
repeat(2, 3, times=3)
```

```
15
```

But they form a kind of documentation to help us understand how the function should be used, and there are tools that can use them to diagnose issues in your code (see the "Linters" section).

In this case we could do this to get what we expected originally:

```
int(repeat("2", "3", times=3))
```

```
232323
```

See the [Python documentation](#) for more details on type annotations and the `typing` library.

Coding Conventions

You should try to have an agreed policy for your team for these matters.

If your language or project has a standard policy, use that. For example:

- Python: [PEP8](#)
- R: [Google's guide for R](#), [tidyverse style guide](#)
- C++: [Google's style guide](#), [Mozilla's](#)
- Julia: [Official style guide](#)

7.3 Linting

Estimated time for this notebook: 15 minutes

There are automated tools which enforce coding conventions and check for common mistakes. These are called *linters*.

Do not blindly believe all these automated tools! Style guides are **guides** not **rules**.

Linters Starter Pack

A good starting point for any Python project is to use `flake8`, `black`, and `isort`. All three should improve the style and consistency of your code whilst requiring minimal setup, and generally they are not opinionated about the way your code is designed, only the way it is formatted and syntax or convention errors.

flake8

Combines two main tools:

- `PyFlakes` - checks Python code for syntax errors
- `pycodestyle` - checks whether Python code is compliant with PEP8 conventions

`flake8` only checks code and flags any syntax/style errors, it does not attempt to fix them.

For example, in the `flake8_example.py` file (in the same directory as this notebook) you'll find this code:

```
from constants import e
def circumference(r):
    return 2 * pi * r
```

Running `flake8` on it gives the following warnings:

```
! flake8 flake8_example.py
flake8_example.py:1:1: F401 'constants.e' imported but unused
flake8_example.py:3:1: E302 expected 2 blank lines, found 1
flake8_example.py:4:16: F821 undefined name 'pi'
```

The first warning tells us we have imported a variable called `e` but not used it, and the last that we're trying to use a variable called `pi` but haven't defined it anywhere. The 2nd warning indicates that in the [PEP8](#) conventions there should be two blank lines before a function definition, but we only have 1.

Running on multiple files

All the examples here run a linter on a single file, but they can be run on all the files in a project at once as well (e.g. by just running `flake8` without a filename).

black

A highly opinionated code formatter, which enforces control of minutiae details of your code.

For example, in the `black_example.py` file (in the same directory as this notebook) you'll find this code:

```
import numpy as np
def my_complex_function(important_argument_1,important_argument_2,optional_argument_3 = 3,optional_argument_4 = 4):
    return np.random.random()*important_argument_1*important_argument_2*optional_argument_3*optional_argument_4
def hello(name,greet="Hello",end="!"):
    print(greet, name, end)
```

After running black on the file:

```
! black black_example.py
reformatted black_example.py
All done! ✨ 🎉 ✨
1 file reformatted.
```

Its contents become:

```
import numpy as np

def my_complex_function(
    important_argument_1,
    important_argument_2,
    optional_argument_3=3,
    optional_argument_4=4,
):
    return (
        np.random.random()
        * important_argument_1
        * important_argument_2
        * optional_argument_3
        * optional_argument_4
    )

def hello(name, greet="Hello", end="!"):
    print(greet, name, end)
```

Changes made by `black`:

- Ensured there are two blank lines before and after function definitions
- Wrapped long lines intelligently
- Removed excess whitespace (e.g. between the arguments in the print statement on the last line)
- Used double quotes " for all strings (rather than a mix of ' and ")

Note that `black` will automatically fix most of the whitespace-related warnings picked up by `flake8` (but it would not fix the import or undefined name errors in the `flake8` example above).

Line length

`black` is not compliant with PEP8 in one way - by default it uses a maximum line length of 88 characters (PEP8 suggests 79 characters). [This is discussed in the black documentation](#).

isort

"Sorts" imports alphabetically in groups in the following order:

1. standard library imports (e.g. `import os`).
2. third-party imports (e.g. `import pandas`).
3. local application/library specific imports (e.g. `from .my_python_file import MyClass`).

with a blank line between each group.

For example in the file `isort_example.py` (in the same directory as this notebook) we have the following imports:

```
import pandas as pd
import os
from matplotlib import pyplot as plt
import black_example
import numpy as np
import json
```

If we run `isort`:

```
! isort isort_example.py
Fixing /home/runner/work/rse-course/rse-
course/module07_construction_and_design/isort_example.py
```

It becomes:

```
import json
import os

import numpy as np
import pandas as pd
from matplotlib import pyplot as plt

import black_example
```

Note that `from` imports are placed at the bottom of each group.

Other Linters

`mypy`

If you use type annotations in your code, `mypy` can check it for errors that may result from variables being assigned the wrong type. For example, in the file `mypy_example.py` we have this code:

```
def hello(name: str, greet: str = "Hello", rep: int = 1) -> str:
    message: str = ""
    for _ in range(rep):
        message += f"{greet} {name}\n"
    return message

print(hello("Bob", 5))
```

If we run `mypy` on it:

```
! mypy mypy_example.py
mypy_example.py:8: error: Argument 2 to "hello" has incompatible type "int";
expected "str"
Found 1 error in 1 file (checked 1 source file)
```

The error tells us we have passed an `int` as the 2nd argument to `hello`, but in the function definition the second argument (`greet`) is defined to be a `str`. We probably meant to write `hello("Bob", rep=5)`.

`pylint`

`pylint` analyses your code for errors, coding standards, and makes suggestions around where code could be refactored. It checks for a much wider range of code quality issues than `flake8` but is also much more likely to pick up *false positives*, i.e. `pylint` is more likely to give you warnings about things you don't want to change.

Let's run it on the file we used for a `flake8` example earlier:

```
from constants import e

def circumference(r):
    return 2 * pi * r
```

```
***** Module flake8_example
flake8_example.py:1:0: C0114: Missing module docstring (missing-module-docstring)
flake8_example.py:1:0: E0401: Unable to import 'constants' (import-error)
flake8_example.py:3:0: C0116: Missing function or method docstring (missing-
function-docstring)
flake8_example.py:3:18: C0103: Argument name "r" doesn't conform to snake_case
naming style (invalid-name)
flake8_example.py:4:15: E0602: Undefined variable 'pi' (undefined-variable)
flake8_example.py:1:0: W0611: Unused e imported from constants (unused-import)

-----
Your code has been rated at 0.00/10
```

Compared to `flake8`, in this case `pylint` also warns us that:

- The `circumference` function doesn't have a docstring
- The `constants` library we try to import is not available on our system
- The variable name `r` doesn't follow conventions (single letter variables are discouraged by convention, we could use `radius` instead)

`nbqa`

`nbqa` allows you to run many Python quality tools (including all the ones we've introduced here) on jupyter notebooks. For example:

```
! nbqa flake8 07_02_coding_conventions.ipynb
07_02_coding_conventions.ipynb:cell_5:1:1: F811 redefinition of unused 'ClassName'
from line 10
```

pylama

`pylama` wraps many code quality tools (including `isort`, `mypy`, `pylint` and much of `flake8`) in a single command.

```
! pylama --linters isort,mccabe,mypy,pycodestyle,pydocstyle,pyflakes,pylint
flake8_example.py

ERROR: /home/runner/work/rse-course/rse-
course/module07_construction_and_design/flake8_example.py Imports are incorrectly
sorted and/or formatted.
flake8_example.py:0:1 Incorrectly sorted imports. [isort]
flake8_example.py:1:1 Cannot find implementation or library stub for module named
"constants" [mypy]
flake8_example.py:1:1 See
https://mypy.readthedocs.io/en/stable/running_mypy.html#missing-imports [mypy]
flake8_example.py:1:1 C0114 Missing module docstring [pylint]
flake8_example.py:1:1 E0401 Unable to import 'constants' [pylint]
flake8_example.py:1:1 W0611 Unused e imported from constants [pylint]
flake8_example.py:1:1 D100 Missing docstring in public module [pydocstyle]
flake8_example.py:3:1 C0101 Missing function or method docstring [pylint]
flake8_example.py:3:19 C0103 Argument name "r" doesn't conform to snake_case
naming style [pylint]
flake8_example.py:3:1 D103 Missing docstring in public function [pydocstyle]
flake8_example.py:3:1 E302 expected 2 blank lines, found 1 [pycodestyle]
flake8_example.py:4:16 E0602 Undefined variable 'pi' [pylint]
```

Setup

Compatibility between linters

If you're using multiple linters in your project you may need to configure them to be compatible with each other. For example, `flake8` warns about lines longer than 79 characters (the PEP8 convention) but `black` will allow lines up to 88 characters by default.

[This repository](#) has an example setup for using `black`, `isort`, and `flake8` together. The `.flake8` and `pyproject.toml` configuration files set `flake8` and `isort` to run in modes compatible with `black`.

Ignoring lines of code or linting rules

There will be times where a linter warns you about something in your code but there's a valid reason it's structured that way and you don't want to change it. Most linters can be configured to ignore specific warnings, either by the type of warning, by file, or by individual code line. For example, adding a `# noqa` comment to a line will make `flake8` ignore it.

Each linter does this differently so check their documentation (e.g. [flake8](#), [isort](#), [mypy](#), [pylint](#)).

Running Linters

It's a good idea to run linters regularly, or even better to have them setup to run automatically so you don't have to remember. There are various tools to help with that:

IDE Integration

Many editors/DEs have integrations with common linters or have their own built-in. This can include highlighting problems inline, or automatically running linters when files are saved, for example. Here is the [VS Code documentation for linting in Python](#).

There are also tools like [editorconfig](#) to help sharing the conventions used within a project, where each contributor uses different IDEs and tools.

pre-commit

`pre-commit` is a manager for creating git "hooks" - scripts that run before making a commit. If a hook errors the commit won't be made, and you'll be prompted to fix the problems first. There are `pre-commit` plugins for all the linters discussed here, and it's a good way to ensure all code committed to your repo has had a level of quality control applied to it.

CI

As well as automating unit tests on a CI system like GitHub Actions it's a good idea to configure them to run linters on your code too.

[Here is an example](#) from a repo using `isort`, `flake8` and `black` in a GitHub Action. Note that in a CI setup tools that usually change your code, like `black` and `isort`, will be configured to only check whether there are changes that need to be made.

7.4 Refactoring

Estimated time for this notebook: 20 minutes

Let's put ourselves in a scenario - that you've probably been in before. Imagine you are changing a large piece of legacy code that's not well structured, introducing many changes at once, trying to keep in your head all the bits and pieces that need to be modified to make it all work again. And suddenly, your officemate comes and ask you to go for coffee... and you've lost all track of what you had in your head and need to start again.

Instead of doing so, we could use a more robust approach to go from nasty ugly code to clean code in a safer way.

Refactoring

To refactor is to:

- Make a change to the design of some software
- Which improves the structure or readability
- But which leaves the actual behaviour of the program completely unchanged.

A word from the Master

Refactoring is a controlled technique for improving the design of an existing code base. Its essence is applying a series of small behavior-preserving transformations, each of which "too small to be worth doing". However the cumulative effect of each of these transformations is quite significant. By doing them in small steps you reduce the risk of introducing errors. You also avoid having the system broken while you are carrying out the restructuring - which allows you to gradually refactor a system over an extended period of time.

List of known refactorings

The next few sections will present some known refactorings.

We'll show before and after code, present any new coding techniques needed to do the refactoring, and describe [code smells](#): how you know you need to refactor.

Replace magic numbers with constants

 Smell: Raw numbers appear in your code

before:

```
data = [math.sin(x) for x in np.arange(0, 3.141, 3.141 / 100)]
result = [0] * 100
for i in range(100):
    for j in range(i + 1, 100):
        result[j] += data[i] * data[i - j] / 100
```

after:

```
resolution = 100
pi = 3.141
data = [math.sin(x) for x in np.arange(0, pi, pi / resolution)]
result = [0] * resolution
for i in range(resolution):
    for j in range(i + 1, resolution):
        result[j] += data[i] * data[i - j] / resolution
```

Replace repeated code with a function

 Smell: Fragments of repeated code appear.

Fragment of model where some birds are chasing each other: if the angle of view of one can see the prey, then start hunting, and if the other see the predator, then start running away.

before:

```
if abs(hawk.facing - starling.facing) < hawk.viewport:
    hawk.hunting()
if abs(starling.facing - hawk.facing) < starling.viewport:
    starling.flee()
```

after:

```
def can_see(source, target):
    return (source.facing - target.facing) < source.viewport

if can_see(hawk, starling):
    hawk.hunting()
if can_see(starling, hawk):
    starling.flee()
```

Change of variable name

 Smell: Code needs a comment to explain what it is for.

before:

```
z = find(x, y)
if z:
    rive(x)
```

after:

```
gene = subsequence(chromosome, start_codon)
if gene:
    transcribe(gene)
```

Separate a complex expression into a local variable

 Smell: An expression becomes long.

before:

```
if color == "red" and fast or big and economy > 40 or price < 5000:
    print("Vrroom!")
```

after:

```
exciting = color == "red" and fast
practical = big and economy > 40
in_budget = price < 5000

if exciting or practical or in_budget:
    print("Vrroom!")
```

Replace loop with iterator

 Smell: Loop variable is an integer from 1 to something.

before:

```
total = 0
for i in range(resolution):
    total += data[i]
```

after:

```
total = 0
for value in data:
    total += value
```

Replace hand-written code with library code

⚠️ Smell: It feels like surely someone else must have done this at some point.

before:

```
xcoords = [start + i * step for i in range(int((end - start) / step))]
```

after:

```
import numpy as np
xcoords = np.arange(start, end, step)
```

See [Numpy](#), [Pandas](#).

Replace set of arrays with array of structures

⚠️ Smell: A function needs to work corresponding indices of several arrays:

before:

```
def can_see(i, source_angles, target_angles, source_viewports):
    return abs(source_angles[i] - target_angles[i]) < source_viewports[i]
```

after:

```
def can_see(source, target):
    return (source["facing"] - target["facing"]) < source["viewport"]
```

Warning: this refactoring greatly improves readability but can make code slower, depending on memory layout. Be careful.

Replace constants with a configuration file

⚠️ Smell: You need to change your code file to explore different research scenarios.

before:

```
flight_speed = 2.0 # mph
bounds = [0, 0, 100, 100]
turning_circle = 3.0 # m
bird_counts = {"hawk": 5, "starling": 500}
```

after:

```
%>%writefile config.yaml
bounds: [0, 0, 100, 100]
counts:
  hawk: 5
  starling: 500
speed: 2.0
turning_circle: 3.0
```

```
Writing config.yaml
```

```
import yaml
config = yaml.safe_load(open("config.yaml"))
print(config)
```

```
{'bounds': [0, 0, 100, 100], 'counts': {'hawk': 5, 'starling': 500}, 'speed': 2.0,
'turning_circle': 3.0}
```

See [YAML](#) and [PyYaml](#), and [Python's os module](#).

Replace global variables with function arguments

⚠️ Smell: A global variable is assigned and then used inside a called function:

before:

```
viewport = pi / 4
if hawk.can_see(starling):
    hawk.hunt(starling)

class Hawk:
    def can_see(self, target):
        return (self.facing - target.facing) < viewport
```

after:

```
viewport = pi / 4
if hawk.can_see(starling, viewport):
    hawk.hunt(starling)

class Hawk:
    def can_see(self, target, viewport):
        return (self.facing - target.facing) < viewport
```

Merge neighbouring loops

⚠️ Smell: Two neighbouring loops have the same for statement

before:

```
for bird in birds:
    bird.build_nest()

for bird in birds:
    bird.lay_eggs()
```

after:

```
for bird in birds:  
    bird.build_nest()  
    bird.lay_eggs()
```

Though there may be a case where all the nests need to be built before the birds can start laying eggs.

Break a large function into smaller units

- **Smell:** A function or subroutine no longer fits on a page in your editor.
- **Smell:** A line of code is indented more than three levels.
- **Smell:** A piece of code interacts with the surrounding code through just a few variables.

before:

```
def do_calculation():  
    for predator in predators:  
        for prey in preys:  
            if predator.can_see(prey):  
                predator.hunt(prey)  
            if predator.can_reach(prey):  
                predator.eat(prey)
```

after:

```
def do_calculation():  
    for predator in predators:  
        for prey in preys:  
            predate(predator, prey)
```



```
def predate(predator, prey):  
    if predator.can_see(prey):  
        predator.hunt(prey)  
    if predator.can_reach(prey):  
        predator.eat(prey)
```

Separate code concepts into files or modules

- **Smell:** You find it hard to locate a piece of code.
- **Smell:** You get a lot of version control conflicts.

before:

```
class One:  
    pass  
  
class Two:  
    def __init__(self):  
        self.child = One()
```

after:

```
%>writefile anotherfile.py  
class One:  
    pass
```

```
Writing anotherfile.py
```

```
from anotherfile import One  
  
class Two:  
    def __init__(self):  
        self.child = One()
```

Refactoring is a safe way to improve code

You may think you can see how to rewrite a whole codebase to be better.

However, you may well get lost halfway through the exercise.

By making the changes as small, reversible, incremental steps, you can reach your target design more reliably.

Tests and Refactoring

Badly structured code cannot be unit tested. There are no "units".

Before refactoring, ensure you have a robust regression test.

This will allow you to *Refactor with confidence*.

As you refactor, if you create any new units (functions, modules, classes), add new tests for them.

Refactoring Summary

- Replace magic numbers with constants
- Replace repeated code with a function
- Change of variable/function/class name
- Replace loop with iterator
- Replace hand-written code with library code
- Replace set of arrays with array of structures
- Replace constants with a configuration file
- Replace global variables with function arguments
- Break a large function into smaller units
- Separate code concepts into files or modules

And many more...

Read [The Refactoring Book](#).

7.5 Object-Oriented Design

Estimated time for this notebook: 15 minutes

In this session, we will finally discuss the thing most people think of when they refer to "Software Engineering": the deliberate *design* of software. We will discuss processes and methodologies for planned development of large-scale software projects: *Software Architecture*.

The software engineering community has, in large part, focused on an object-oriented approach to the design and development of large scale software systems. The basic concepts of object orientation are necessary to follow much of the software engineering conversation.

Design processes

In addition to object-oriented architecture, software engineers have focused on the development of processes for robust, reliable software development. These codified ways of working hope to enable organisations to repeatedly and reliably complete complex software projects in a way that minimises both development and maintenance costs, and meets user requirements.

Design and research

Software engineering theory has largely been developed in the context of commercial software companies.

The extent to which the practices and processes developed for commercial software are applicable in a research context is itself an active area of research.

Recap of Object-Orientation

Classes: User defined types

```
class Person:
    def __init__(self, name, age):
        self.name = name
        self.age = age

    def grow_up(self):
        self.age += 1

terry = Person("Terry", 76)
terry.home = "Colwyn Bay"
```

⚠️ Notice, that in Python, you can add properties to an object once it's been defined. Just because you can doesn't mean you should!

Declaring a class

Class: A user-defined type

```
class MyClass:
    pass
```

Object instances

Instance: A particular object *instantiated* from a class.

```
my_object = MyClass()
```

Method

Method: A function which is "built in" to a class

```
class MyClass:
    def someMethod(self, argument):
        pass

my_object = MyClass()
my_object.someMethod(value)
```

Constructor

Constructor: A special method called when instantiating a new object

```
class MyClass:
    def __init__(self, argument):
        pass

my_object = MyClass(value)
```

Member Variable

Member variable: a value stored inside an instance of a class.

```
class MyClass:
    def __init__(self):
        self.member = "Value"

my_object = MyClass()
print(my_object.member)
```

Value

Object refactorings

Replace add-hoc structure with user defined classes

💩 **Smell:** A data structure made of nested arrays and dictionaries becomes unwieldy.

before:

```
from random import random

birds = [
    {"position": random(), "velocity": random(), "type": kind} for kind in bird_types
]

average_position = average([bird["position"] for bird in birds])
```

after:

```
class Bird:
    def __init__(self, kind):
        from random import random

        self.type = kind
        self.position = random()
        self.velocity = random()

birds = [Bird(kind) for kind in bird_types]
average_position = average([bird.position for bird in birds])
```

Replace function with a method

💩 Smell: A function is always called with the same kind of thing

before:

```
def can_see(source, target):
    return (source.facing - target.facing) < source.viewport

if can_see(hawk, starling):
    hawk.hunt()
```

after:

```
class Bird:
    def can_see(self, target):
        return (self.facing - target.facing) < self.viewport

if hawk.can_see(starling):
    hawk.hunt()
```

Replace method arguments with class members

💩 Smell: A variable is nearly always used in arguments to a class.

before:

```
class Person:
    def __init__(self, genes):
        self.genes = genes

    def reproduce_probability(self, age):
        pass

    def death_probability(self, age):
        pass

    def emigrate_probability(self, age):
        pass
```

after:

```
class Person:
    def __init__(self, genes, age):
        self.age = age
        self.genes = genes

    def reproduce_probability(self):
        pass

    def death_probability(self):
        pass

    def emigrate_probability(self):
        pass
```

Replace global variable with class and member

💩 Smell: A global variable is referenced by a few functions

before:

```
name = "Terry Jones"
birthday = [1, 2, 1942]
today = [22, 11]

if today == birthday[0:2]:
    print(f"Happy Birthday, {name}")
else:
    print("No birthday for you today.")
```

after:

```
class Person:
    def __init__(self, birthday, name):
        self.birth_day = birthday[0]
        self.birth_month = birthday[1]
        self.birth_year = birthday[2]
        self.name = name

    def check_birthday(self, today_day, today_month):
        if not self.birth_day == today_day:
            return False
        if not self.birth_month == today_month:
            return False
        return True

    def greet_appropriately(self, today):
        if self.check_birthday(*today):
            print(f"Happy Birthday, {self.name}")
        else:
            print("No birthday for you.")

john = Person([5, 5, 1943], "Michael Palin")
john.greet_appropriately(today)
```

- Replace ad-hoc structure with a class
- Replace function with a method
- Replace method argument with class member
- Replace global variable with class data

7.6 Class design

Estimated time for this notebook: 20 minutes

The concepts we have introduced are common between different object oriented languages. Thus, when we design our program using these concepts, we can think at an architectural level, independent of language syntax.

In Python:

```
class Particle:
    def __init__(self, position, velocity):
        self.position = position
        self.velocity = velocity

    def move(self, delta_t):
        self.position += self.velocity * delta_t
```

In C++:

```
class Particle {
    std::vector<double> position;
    std::vector<double> velocity;
    Particle(std::vector<double> position, std::vector<double> velocity);
    void move(double delta_t);
}
```

In Fortran:

```
type particle
    real :: position
    real :: velocity
contains
    procedure :: init
    procedure :: move
end type particle
```

UML

UML is a conventional diagrammatic notation used to describe "class structures" and other higher level aspects of software design.

Computer scientists get worked up about formal correctness of UML diagrams and learning the conventions precisely. Working programmers can still benefit from using UML to describe their designs.

YUML

We can see a YUML model for a Particle class with `position` and `velocity` data and a `move()` method using the [YUML](#) online UML drawing tool ([example](#)).

```
http://yuml.me/diagram/boring/class/[Particle|position;velocity|move%28%29]
```

Here's how we can use Python code to get an image back from YUML:

```
from IPython.display import SVG

def yuml(model):
    return SVG(url=f"http://yuml.me/diagram/boring/class/{model}")

yuml("[Particle|position;velocity|move()]")
```



The representation of the `Particle` class defined above in UML is done with a box with three sections. The name of the class goes on the top, then the name of the member variables in the middle, and the name of the methods on the bottom. We will see later why this is useful.

Information Hiding

Sometimes, our design for a program would be broken if users start messing around with variables we don't want them to change.

Robust class design requires consideration of which subroutines are intended for users to use, and which are internal. Languages provide features to implement this: access control.

In python, we use leading underscores to control whether member variables and methods can be accessed from outside the class:

- single leading underscore (_) is used to document it's private but people could use it if wanted (thought they shouldn't);
- double leading underscore (__) raises errors if called.

```

class MyClass:
    def __init__(self):
        self.__private_data = 0
        self._private_data = 0
        self.public_data = 0

    def __private_method(self):
        pass

    def _private_method(self):
        pass

    def public_method(self):
        pass

    def called_inside(self):
        self.__private_method()
        self._private_method()
        self.__private_data = 1
        self._private_data = 1

MyClass().called_inside()

MyClass().__private_method() # Works, but forbidden by convention

MyClass().public_method() # OK
print(MyClass().__private_data)

0

print(MyClass().public_data)

0

MyClass().__private_method() # Generates error

-----
AttributeError                         Traceback (most recent call last)
Cell In [8], line 1
----> 1 MyClass().__private_method() # Generates error
      AttributeError: 'MyClass' object has no attribute '__private_method'

print(MyClass().__private_data) # Generates error

-----
AttributeError                         Traceback (most recent call last)
Cell In [9], line 1
----> 1 print(MyClass().__private_data) # Generates error
      AttributeError: 'MyClass' object has no attribute '__private_data'

```

Property accessors

Python provides a mechanism to make functions appear to be variables. This can be used if you want to change the way a class is implemented without changing the interface:

```

class Person:
    def __init__(self):
        self.name = "John Watson"

Person().name

'John Watson'

```

becomes:

```

class Person:
    def __init__(self):
        self._first = "John"
        self._second = "Watson"

    @property
    def name(self):
        return f"{self._first} {self._second}"

Person().name

'John Watson'

```

Making the same external code work as before.

Note that the code behaves the same way to the outside user. The implementation detail is hidden by private variables. In languages without this feature, such as C++, it is best to always make data private, and always access data through functions:

```

class Person:
    def __init__(self):
        self._name = "John Watson"

    def name(self): # an access function
        return self._name

Person().name()

'John Watson'

```

But in Python this is unnecessary because the `@property` capability.

Another way could be to create a member variable `name` which holds the full name. However, this could lead to inconsistent data. If we create a `get_married` function, then the name of the person won't change!

```

class Person:
    def __init__(self, first, second):
        self._first_ = first
        self._second_ = second
        self.name = f"{self._first_} {self._second_}"

    def get_married(self, to):
        self._second_ = f"(self._second_)-{to._second_}"

```

```

john = Person("John", "Watson")
john.name

```

'John Watson'

```

sherlock = Person("Sherlock", "Holmes")
john.get_married(sherlock)
john._second_

```

'Watson-Holmes'

```

john.name # Not John Watson-Holmes?

```

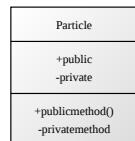
'John Watson'

This type of situation could make the object data structure inconsistent with itself. Making variables being out of sync with other variables. Each piece of information should only be stored in one place! In this case, `name` should be calculated each time it's required as previously shown. In database design, this is called [Normalization](#).

UML for private/public

We prepend a `+/`- on public/private member variables and methods:

```
yuml("[Particle]+public;-private+publicmethod();-privatemethod]")
```



CREATED WITH YUML

Class Members

Class, or *static* members, belong to the class as a whole, and are shared between instances.

This is an object that keeps a count on how many have been created of it.

```

class Counted:
    number_created = 0

    def __init__(self):
        Counted.number_created += 1

    @classmethod
    def howMany(cls):
        return cls.number_created

Counted.howMany()

```

0

```

x = Counted()
Counted.howMany()

```

1

```

z = [Counted() for x in range(5)]
Counted.howMany()

```

6

```

x.howMany()

```

6

The data is shared among all the objects instantiated from that class. Note that in `__init__` we are not using `self.number_created` but the name of the class. The `howMany` function is not a method of a particular object. It's called on the class, not on the object. This is possible by using the `@classmethod` decorator.

Inheritance and Polymorphism

Object-based vs Object-Oriented

So far we have seen only object-based programming, not object-oriented programming.

Using Objects doesn't mean your code is object-oriented.

To understand object-oriented programming, we need to introduce **polymorphism** and **inheritance**.

Inheritance

- Inheritance is a mechanism that allows related classes to share code.

- Inheritance allows a program to reflect the [ontology](#) of kinds of thing in a program.

Ontology and inheritance

- A bird is a kind of animal
- An eagle is a kind of bird
- A starling is also a kind of bird
- All animals can be born and die
- Only birds can fly (Ish.)
- Only eagles hunt
- Only starlings flock

Inheritance in python

```
class Animal:
    def beBorn(self):
        print("I exist")

    def die(self):
        print("Argh!")

class Bird(Animal):
    def fly(self):
        print("Whee!")

class Eagle(Bird):
    def hunt(self):
        print("I'm gonna catcha!")

class Starling(Bird):
    def flew(self):
        print("I'm flying away!")

Eagle().beBorn()
Eagle().hunt()
```

I exist
I'm gonna catcha!

Inheritance terminology

Here are two equivalents definition, one coming from C++ and another from Java:

- A *derived class* *derives* from a *base class*.
- A *subclass* *inherits* from a *superclass*.

These are different terms for the same thing. So, we can say:

- Eagle is a subclass of the Animal superclass.
- Animal is the base class of the Eagle derived class.

Another equivalent definition is using the synonym *child / parent* for *derived / base class*:

- A *child class* *extends* a *parent class*.

Inheritance and constructors

To use implicitly constructors from a *superclass*, we can use `super` as shown below.

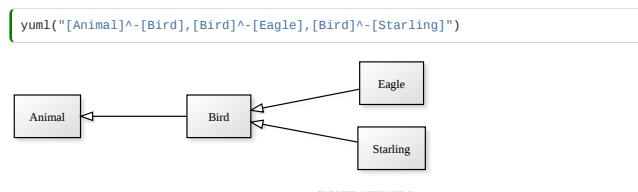
```
class Animal:
    def __init__(self, age):
        self.age = age

class Person(Animal):
    def __init__(self, age, name):
        super().__init__(age)
        self.name = name
```

Read [Raymond Hettinger's article about super](#) to see various real examples.

Inheritance UML diagrams

UML shows inheritance with an open triangular arrow pointing from subclass to superclass.



Aggregation vs Inheritance

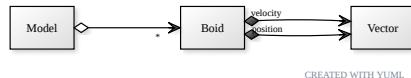
If one object *has* or *owns* one or more objects, this is *not* inheritance.

For example, the boids example we saw few weeks ago, could be organised as an overall Model, which it owns several Boids, and each Boid owns two 2-vectors, one for position and one for velocity.

Aggregation in UML

The Boids situation can be represented thus:





The open diamond indicates **Aggregation**, the closed diamond **composition**. (A given boid might belong to multiple models, a given position vector is forever part of the corresponding Boid.)

The asterisk represents cardinality, a model may contain multiple Boids. This is a [one to many relationship](#). [Many to many relationship](#) is shown with * on both sides.

Refactoring to inheritance

Smell: Repeated code between two classes which are both ontologically subtypes of something

before:

```

class Person:
    def __init__(self, age, job):
        self.age = age
        self.job = job

    def birthday(self):
        self.age += 1

class Pet:
    def __init__(self, age, owner):
        self.age = age
        self.owner = owner

    def birthday(self):
        self.age += 1

```

after:

```

class Animal:
    def __init__(self, age):
        self.age = age

    def birthday(self):
        self.age += 1

class Person(Animal):
    def __init__(self, age, job):
        self.job = job
        super().__init__(age)

class Pet(Animal):
    def __init__(self, age, owner):
        self.owner = owner
        super().__init__(age)

```

Polymorphism

```

class Dog:
    def noise(self):
        return "Bark"

class Cat:
    def noise(self):
        return "Miaow"

class Pig:
    def noise(self):
        return "Oink"

class Cow:
    def noise(self):
        return "Moo"

animals = [Dog(), Dog(), Cat(), Pig(), Cow(), Cat()]
for animal in animals:
    print(animal.noise())

```

```

Bark
Bark
Miaow
Oink
Moo
Miaow

```

This will print "Bark Bark Miaow Oink Moo Miaow"

If two classes support the same method, but it does different things for the two classes, then if an object is of an unknown class, calling the method will invoke the version for whatever class the instance of.

Polymorphism and Inheritance

Often, polymorphism uses multiple derived classes with a common base class. However, [duck typing](#) in Python means that all that is required is that the types support a common **Concept** (Such as iterable, or container, or, in this case, the Noisy concept.)

A common base class is used where there is a likely **default** that you want several of the derived classes to have.

```

class Animal:
    def noise(self):
        return "I don't make a noise."

class Dog(Animal):
    def noise(self):
        return "Bark"

class Worm(Animal):
    pass

class Poodle(Dog):
    pass

animals = [Dog(), Worm(), Pig(), Cow(), Poodle()]
for animal in animals:
    print(animal.noise())

```

```

Bark
I don't make a noise.
Oink
Moo
Bark

```

Undefined Functions and Polymorphism

In the above example, we put in a dummy noise for Animals that don't know what type they are.

Instead, we can explicitly deliberately leave this undefined, and we get a crash if we access an undefined method.

```

class Animal:
    pass

class Worm(Animal):
    pass

Worm().noise() # Generates error

```

```

-----
AttributeError                         Traceback (most recent call last)
Cell In [31], line 1
----> 1 Worm().noise() # Generates error
AttributeError: 'Worm' object has no attribute 'noise'

```

Refactoring to Polymorphism

 **Smell:** a function uses a big set of `if` statements or a `case` statement to decide what to do:

before:

```

class Animal:
    def __init__(self, animal_kind):
        self.animal_kind = animal_kind

    def noise(self):
        if self.animal_kind == "Dog":
            return "Bark"
        if self.animal_kind == "Cat":
            return "Miaow"
        if self.animal_kind == "Cow":
            return "Moo"
        return "Growl"

```

which is better replaced by the code above.

Interfaces and concepts

In C++, it is common to define classes which declare dummy methods, called "virtual" methods, which specify the methods which derived classes must implement. Classes which define these methods, but which cannot be instantiated into actual objects, are called "abstract base" classes or "interfaces".

Python's Duck Typing approach means explicitly declaring these is unnecessary: any class concept which implements appropriately named methods will do. These as user-defined **concepts**, just as "iterable" or "container" are built-in Python concepts. A class is said to "implement an interface" or "satisfy a concept".

Interfaces in UML

Interfaces implementation (a common ancestor that doesn't do anything but defines methods to share) in UML is indicated thus:



CREATED WITH YUML.

Further UML

UML is a much larger diagram language than the aspects we've shown here.

- Message sequence charts show signals passing back and forth between objects ([Web Sequence Diagrams](#)).
- Entity Relationship Diagrams can be used to show more general relationships between things in a system.

Read more about UML on Martin Fowler's [book about the topic](#).

7.7 Design Patterns

 **Warning: Advanced topic!** 

Estimated time for this notebook: 20 minutes

Class Complexity

We've seen that using object orientation can produce quite complex class structures, with classes owning each other, instantiating each other, and inheriting from each other.

There are lots of different ways to design things, and decisions to make.

- Should I inherit from this class, or own it as a member variable? ("is a" vs "has a")
- How much flexibility should I allow in this class's inner workings?
- Should I split this related functionality into multiple classes or keep it in one?

Design Patterns

Programmers have noticed that there are certain ways of arranging classes that work better than others.

These are called "design patterns".

They were first collected on one of the [world's first Wikis](#), as the [Portland Pattern Repository](#).

Reading a pattern

A description of a pattern in a book such as the [Gang Of Four](#) book usually includes:

- **Intent** - what's the purpose
- **Motivation** - why you want to use it
- **Applicability** - when do you want to use it
- **Structure** - what does it look like (e.g., UML diagram)
- **Participants** - What are the different classes in it
- **Collaborations** - how they work together
- **Consequences** - What are the results and trade-offs
- **Implementation** - How is it implemented
- **Sample Code** - In practice.

Introducing Some Patterns

There are lots and lots of design patterns, and it's a great literature to get into to read about design questions in programming and learn from other people's experience.

We'll just show a few in this session:

- [Factory Method](#)
- [Builder](#)
- [Strategy](#)
- [Model-View-Controller](#)

Supporting code

```
%matplotlib inline
from unittest.mock import Mock
from IPython.display import HTML, SVG

def yuml(model):
    return SVG(url=f"http://yuml.me/diagram/boring/class/{model}")
```

Factory Pattern

Here's what the Gang of Four Book says about Factory Method:

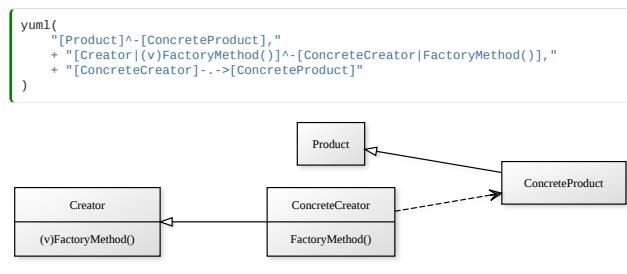
Intent: Define an interface for creating an object, but let subclasses decide which class to instantiate. Factory Method lets a class defer instantiation to subclasses.

Applicability: Use the Factory method pattern when:

- A class can't anticipate the class of objects it must create
- A class wants its subclasses to specify the objects it creates

This is pretty hard to understand, so let's look at an example.

Factory UML



CREATED WITH YUML

Factory Example

An "agent based model" is one like the Boids model from last week: agents act and interact under certain rules. Complex phenomena can be described by simple agent behaviours.

```
class AgentModel:
    def simulate(self):
        for agent in self.agents:
            for target in self.agents:
                agent.interact(target)
                agent.simulate()
```

Agent model constructor

This logic is common to many kinds of Agent based model (ABM), so we can imagine a common class for agent based models: the constructor could parse a configuration specifying how many agents of each type to create, their initial conditions and so on.

However, this common constructor doesn't know what kind of agent to create; as a common base, it could be a model of boids, or the agents could be remote agents on foreign servers, or they could even be physical hardware robots connected to the driving model over Wifi!

We need to defer the construction of the agents. We can do this with polymorphism: each derived class of the ABM can have an appropriate method to create its agents:

```
class AgentModel:
    def __init__(self, config):
        self.agents = []
        for agent_config in config:
            self.agents.append(self.create(**agent_config))

    def simulate(self):
        for agent in self.agents:
            for target in self.agents:
                agent.interact(target)
            agent.simulate()
```

This is the *factory method* pattern: a common design solution to the need to defer the construction of daughter objects to a derived class. `self.create` is not defined here, but in each of the agents that inherits from `AgentModel`. Using polymorphism to get deferred behaviour on what you want to create.

Agent derived classes

The type that is created is different in the different derived classes:

```
class BirdModel(AgentModel):
    def create(self, agent_config):
        return Boid(agent_config)
```

Agents are the base product, boids or robots are a ConcreteProduct.

```
class WebAgentFactory(AgentModel):
    def __init__(self, url, config):
        self.url = url
        self.connection = AmazonCompute.connect(url)
        super().__init__(config) # run the AgentModel constructor

    def create(self, agent_config):
        return OnlineAgent(agent_config, self.connection)
```

There is no need to define an explicit base interface for the "Agent" concept in Python: anything that responds to "simulate" and "interact" methods will do: this is our Agent concept.

Refactoring to Patterns

It's easy to get into a tangle trying to make base classes which somehow "promote" themselves into a derived class based on some code in the base class.

This is an example of an "Antipattern": like a Smell, this is a recognised Wrong Way of doing things.

What we should write instead is a `Creator` with a `FactoryMethod`.

Consider the following code:

```
class AgentModel:
    def __init__(self):
        self.agents = []

    def simulate(self):
        for agent in self.agents:
            for target in self.agents:
                agent.interact(target)
            agent.simulate()

class BirdModel(AgentModel):
    def __init__(self, config):
        super().__init__() # run the constructor of the AgentModel class
        for boid_config in config:
            self.agents.append(Boid(**boid_config))

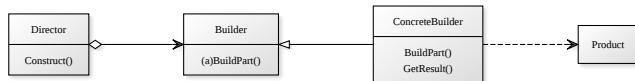
class WebAgentFactory(AgentModel):
    def __init__(self, url, config):
        self.url = url
        connection = AmazonCompute.connect(url)
        super().__init__() # run the constructor of the AgentModel class
        for agent_config in config:
            self.agents.append(OnlineAgent(agent_config, connection))
```

The agent creation loop is almost identical in the two classes; so we can be sure we need to refactor it away; but the `type` that is created is different in the two cases, so this is the smell that we need a factory pattern.

Builder

Intent: Separate the steps for constructing a complex object from its final representation.

```
yuml
  "[Director|Construct()]->[Builder|(a)BuildPart()]
  + "[Builder]->[ConcreteBuilder|BuildPart();GetResult()]
  + "[ConcreteBuilder]-->[Product]"
```



CREATED WITH YUML.

Builder example

Let's continue our Agent Based modelling example.

There's a lot more to defining a model than just adding agents of different kinds: we need to define boundary conditions, specify wind speed or light conditions.

We could define all of this for an imagined advanced Model with a very very long constructor, with lots of optional arguments:

```
class AdvancedModel:  
    def __init__(  
        self,  
        xsize,  
        ysize,  
        agent_count,  
        wind_speed,  
        agent_sight_range,  
        eagle_start_location,  
    ):  
        pass
```

Builder preferred to complex constructor

However, long constructors easily become very complicated. Instead, it can be cleaner to define a Builder for models. A builder is like a deferred factory: each step of the construction process is implemented as an individual method call, and the completed object is returned when the model is ready.

```
AdvancedModel = Mock() # Create a temporary mock so the example works!
```

```
class ModelBuilder:  
    def start_model(self):  
        self.model = AdvancedModel()  
        self.model.xlim = None  
        self.model.ylim = None  
  
    def set_bounds(self, xlim, ylim):  
        self.model.xlim = xlim  
        self.model.ylim = ylim  
  
    def add_agent(self, xpost, ypost):  
        pass # Implementation here  
  
    def finish(self):  
        self.validate()  
        return self.model  
  
    def validate(self):  
        assert self.model.xlim is not None  
        # Check that the all the  
        # parameters that need to be set  
        # have indeed been set.
```

Inheritance of an Abstract Builder for multiple concrete builders could be used where there might be multiple ways to build models with the same set of calls to the builder: for example a version of the model builder yielding models which can be executed in parallel on a remote cluster.

Using a builder

```
builder = ModelBuilder()  
builder.start_model()  
  
builder.set_bounds(500, 500)  
builder.add_agent(40, 40)  
builder.add_agent(400, 100)  
  
model = builder.finish()  
model.simulate()
```

```
<Mock name='mock().simulate()' id='140567502017344'>
```

Avoid staged construction without a builder.

We could, of course, just add all the building methods to the model itself, rather than having the model be yielded from a separate builder.

This is an antipattern that is often seen: a class whose `__init__` constructor alone is insufficient for it to be ready to use. A series of methods must be called, in the right order, in order for it to be ready to use.

This results in very fragile code: it's hard to keep track of whether an object instance is "ready" or not. Use the builder pattern to keep deferred construction in control.

We might ask why we couldn't just use a validator in all of the methods that must follow the deferred constructors; to check they have been called. But we'd need to put these in every method of the class, whereas with a builder, we can validate only in the `finish` method.

Strategy Pattern

Define a family of algorithms, encapsulate each one, and make them interchangeable. Strategy lets the algorithm vary independently from clients that use it.

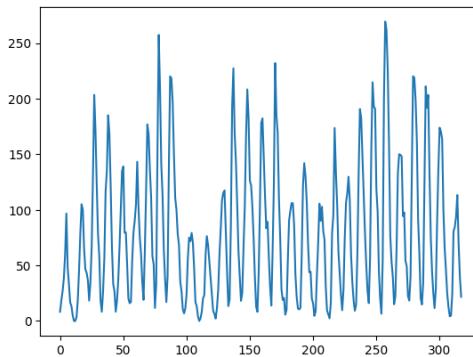
Strategy pattern example: sunspots

Consider the sequence of sunspot observations:

```
import csv  
from io import StringIO  
  
import requests  
  
def load_sunspots():  
    url_base = "https://www.quandl.com/api/v1/datasets/SIDC/SUNSPOTS_A.csv"  
    x = requests.get(  
        url_base,  
        params={  
            "trim_start": "1700-12-31",  
            "trim_end": "2018-01-01",  
            "sort_order": "asc",  
        },  
        timeout=60,  
    )  
    # Convert requests result to look like a file buffer before reading with CSV  
    data = csv.reader(StringIO(x.text))  
    next(data) # Skip header row  
    return [float(row[1]) for row in data]
```

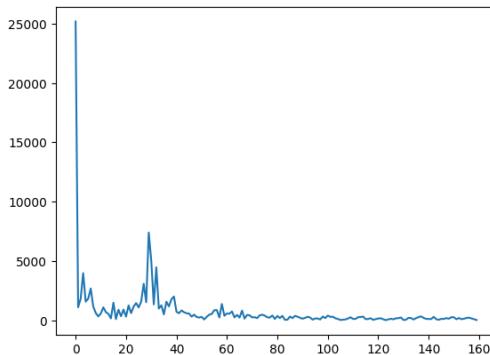
```
import matplotlib.pyplot as plt  
  
spots = load_sunspots()  
plt.plot(spots)
```

```
[<matplotlib.lines.Line2D at 0x7fd861d8c820>]
```



Sunspot cycle has periodicity

```
import numpy as np
spectrum = np.fft.rfft(spots)
plt.figure()
plt.plot(abs(spectrum))
plt.savefig("fixed.png")
```



Years are not constant length

There's a potential problem with this analysis however:

- Years are not constant length
- Leap years exist
- But, the Fast Fourier Transform assumes evenly spaced intervals

Strategy Pattern for Algorithms

Uneven time series

The Fast Fourier Transform cannot be applied to uneven time series.

We could:

- Ignore this problem, and assume the effect is small;
- Interpolate and resample to even times;
- Use a method which is robust to unevenly sampled series, such as [LSSA](#);

We also want to find the period of the strongest periodic signal in the data, there are various different methods we could use for this also, such as integrating the fourier series by quadrature to find the mean frequency, or choosing the largest single value.

Too many classes!

We could implement a base class for our common code between the different approaches, and define derived classes for each different algorithmic approach. However, this has drawbacks:

- The constructors for each derived class will need arguments for all the numerical method's control parameters, such as the degree of spline for the interpolation method, the order of quadrature for integrators, and so on.
- Where we have multiple algorithmic choices to make (interpolator, periodogram, peak finder...) the number of derived classes would explode: `class SunspotAnalyzerSplineFFTrapeziumNearMode` is a bit unwieldy.
- The algorithmic choices are not then available for other projects
- This design doesn't fit with a clean Ontology of "kinds of things": there's no Abstract Base for spectrogram generators...

Apply the strategy pattern:

- We implement each algorithm for generating a spectrum as its own Strategy class.
- They all implement a common interface
- Arguments to strategy constructor specify parameters of algorithms, such as spline degree
- One strategy instance for each algorithm is passed to the constructor for the overall analysis

First, we'll define a helper class for our time series.

```

class Series:
    """Enhance NumPy N-d array with some helper functions for clarity"""

    def __init__(self, data):
        self.data = np.array(data)
        self.count = self.data.shape[0]
        self.start = self.data[0, 0]
        self.end = self.data[-1, 0]
        self.range = self.end - self.start
        self.step = self.range / self.count
        self.times = self.data[:, 0]
        self.values = self.data[:, 1]
        self.plot_data = [self.times, self.values]
        self.inverse_plot_data = [1.0 / self.times[20:], self.values[20:]]
```

Then, our class which contains the analysis code, except the numerical methods

```

from datetime import datetime

class AnalyseSunspotData:
    def format_date(self, date):
        date_format = r"%Y-%m-%d"
        return datetime.strptime(date, date_format)

    def load_data(self, csv_file):
        start_date_str = "1700-12-31"
        end_date_str = "2014-01-01"
        self.start_date = self.format_date(start_date_str)
        url_base = f"https://www.quandl.com/api/v1/datasets/{csv_file}"
        x = requests.get(
            url_base,
            params={
                "trim_start": start_date_str,
                "trim_end": end_date_str,
                "sort_order": "asc",
            },
            timeout=60,
        )
        secs_per_year = (datetime(2014, 1, 1) - datetime(2013, 1,
1)).total_seconds()
        data = csv.reader(StringIO(x.text))
        # Convert requests result to look like a file buffer before reading with
        CSV
        next(data) # Skip header row
        self.series = Series([
            [
                (self.format_date(row[0]) - self.start_date).total_seconds() /
                secs_per_year,
                float(row[1]),
            ]
            for row in data
        ])
    def __init__(self, frequency_strategy):
        self.load_data("SDC/SUNSPOTS_A.csv")
        self.frequency_strategy = frequency_strategy

    def frequency_data(self):
        return self.frequency_strategy.transform(self.series)
```

Our existing simple fourier strategy

```

class FourierNearestFrequencyStrategy:
    def transform(self, series):
        transformed = np.fft.fft(series.values)[0 : series.count // 2]
        frequencies = np.fft.fftfreq(series.count, series.step)[0 : series.count
// 2]
        return Series(list(zip(frequencies, abs(transformed) / series.count)))
```

A strategy based on interpolation to a spline

```

from scipy.interpolate import UnivariateSpline

class FourierSplineFrequencyStrategy:
    def next_power_of_two(self, value):
        "Return the next power of 2 above value"
        return 2 ** (1 + int(np.log(value) / np.log(2)))

    def transform(self, series):
        spline = UnivariateSpline(series.times, series.values)
        # Linspace will give us "evenly" spaced points in the series
        fft_count = self.next_power_of_two(series.count)
        points = np.linspace(series.start, series.end, fft_count)
        regular_xs = [spline(point) for point in points]
        transformed = np.fft.fft(regular_xs)[0 : fft_count // 2]
        frequencies = np.fft.fftfreq(fft_count, series.range / fft_count)[
            0 : fft_count // 2]
        return Series(list(zip(frequencies, abs(transformed) / fft_count)))
```

A strategy using the Lomb-Scargle Periodogram

```

import math
from copy import deepcopy

from scipy.signal import lombscargle

class LombFrequencyStrategy:
    def transform(self, series):
        frequencies = np.array(
            np.linspace(1.0 / series.range, 0.5 / series.step, series.count)
        )
        result = lombscargle(
            deepcopy(series.times), deepcopy(series.values), 2.0 * math.pi *
            frequencies
        )
        return Series(list(zip(frequencies, np.sqrt(result / series.count))))
```

Define our concrete solutions with particular strategies

```

fourier_model = AnalyseSunspotData(FourierSplineFrequencyStrategy())
lomb_model = AnalyseSunspotData(LombFrequencyStrategy())
nearest_model = AnalyseSunspotData(FourierNearestFrequencyStrategy())
```

Use these new tools to compare solutions

```

from scipy import signal
rng = np.random.default_rng()
nin = 1000
nout = 100000
frac_points = 0.9
A = 2.0
w = 1.0
phi = 0.5 * np.pi
r = rng.standard_normal(nin)
x = np.linspace(0.01, 10 * np.pi, nin)
x = x[r >= frac_points]
y = A * np.sin(w * x + phi)
f = np.linspace(0.01, 10, nout)
pgram = signal.lombcargle(x, y, f, normalize=True)

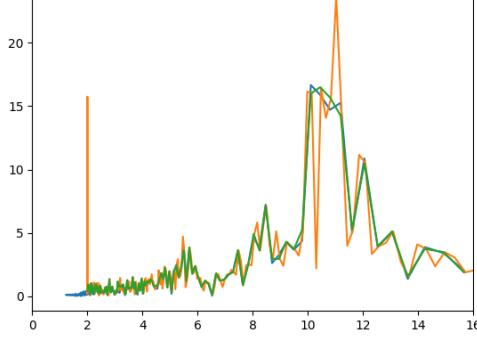
comparison = fourier_model.frequency_data().inverse_plot_data + ["C0"]
comparison += lomb_model.frequency_data().inverse_plot_data + ["C1"]
comparison += nearest_model.frequency_data().inverse_plot_data + ["C2"]

deviation = 365 * (
    fourier_model.series.times
    - np.linspace(
        fourier_model.series.start, fourier_model.series.end,
        fourier_model.series.count
    )
)

```

pit.plot("comparison")
pit.xlim(0, 16)

(0.0, 16.0)



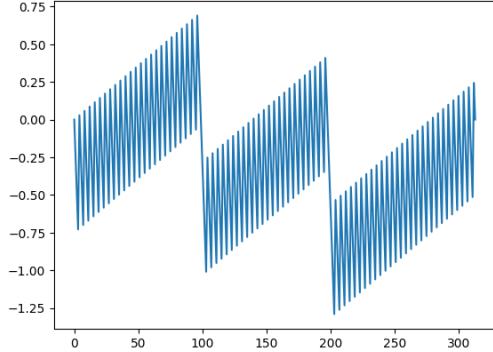
Results: Deviation of year length from average

```

pit.plot(deviation)

```

[<matplotlib.lines.Line2D at 0x7fd864f527f0>]



Model-View-Controller

Separate graphics from science!

Whenever we are coding a simulation or model we want to:

- Implement the maths of the model
- Visualise, plot, or print out what is going on.

We often see scientific programs where the code which is used to display what is happening is mixed up with the mathematics of the analysis. This is hard to understand.

We can do better by separating the `Model` from the `View`, and using a "Controller" to manage them.

Model

This is where we describe our internal logic, rules, etc.

```

class Model:
    def __init__(self):
        self.positions = np.random.rand(100, 2)
        self.speeds = np.random.rand(100, 2) + np.array([-0.5, -0.5])[np.newaxis,
        :]
        self.deltat = 0.01

    def simulation_step(self):
        self.positions += self.speeds * self.deltat

    def agent_locations(self):
        return self.positions

```

View

This is where we describe what the user sees of our Model, what's displayed. You may have different type of visualisation (e.g., on one type of projection, a 3D view, a surface view, ...) which can be implemented in different view classes.

```

class View:
    def __init__(self, model):
        self.figure = plt.figure()
        axes = plt.axes()
        self.model = model
        self.scatter = axes.scatter(
            model.agent_locations()[:, 0], model.agent_locations()[:, 1]
        )

    def update(self):
        self.scatter.set_offsets(self.model.agent_locations())

```

Controller

This is the class that tells the view that the models has changed and updates the model with any change the user has input through the view.

```

from matplotlib import animation

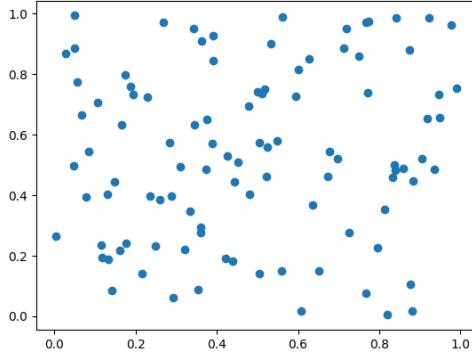
class Controller:
    def __init__(self):
        self.model = Model() # Or use Builder
        self.view = View(self.model)

    def animate(self, frame_number):
        self.model.simulation_step()
        self.view.update()

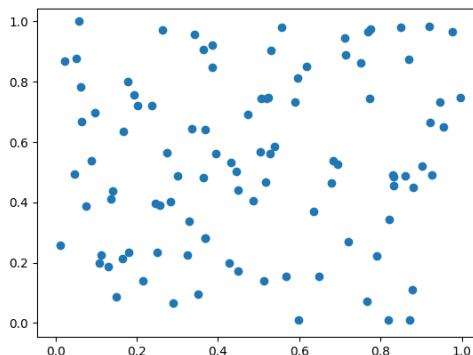
    def go(self):
        anim = animation.FuncAnimation(
            self.view.figure, self.animate, frames=200, interval=50
        )
        return anim.to_jshtml()

```

```
cont1 = Controller()
```



```
HTML(cont1.go())
```



Once Loop Reflect

Other resources

- [Design Patterns by Refactoring.Guru](#)
- [Course on design patterns](#) and [Advanced design patterns](#) with Python.

- [A collection of design patterns and idioms in Python.](#)
- [Head First Design Patterns](#) - based on Java (with [online course at Lynda.com](#)).
- [Design Pattern for Dummies](#).

7.8 Exercise: Refactoring The Bad Boids

We have written some *very bad* code implementing our Boids flocking example. We first looked at the Boids in Module 3 (but you don't need to have seen the previous example to do this exercise).

Here's the GitHub link: <https://github.com/alan-turing-institute/bad-boids>

Please [fork it](#) on GitHub, and clone your fork:

```
git clone git@github.com:yourname/bad-boids.git
# OR git clone https://github.com/yourname/bad-boids.git
```

The Code

For the Exercise, you should start from the GitHub repository, but here's our terrible code (the contents of the `boids.py` file), which simulates a flock of birds ("boids"):

```
"""
A deliberately bad implementation of [Boids](http://dl.acm.org/citation.cfm?
doi=37401.37406)
for use as an exercise on refactoring.
"""

import random

from matplotlib import animation
from matplotlib import pyplot as plt

# Deliberately terrible code for teaching purposes

boids_x = [random.uniform(-450, 50.0) for x in range(50)]
boids_y = [random.uniform(300.0, 600.0) for x in range(50)]
boid_x_velocities = [random.uniform(0, 10.0) for x in range(50)]
boid_y_velocities = [random.uniform(-20.0, 20.0) for x in range(50)]
boids = (boids_x, boids_y, boid_x_velocities, boid_y_velocities)

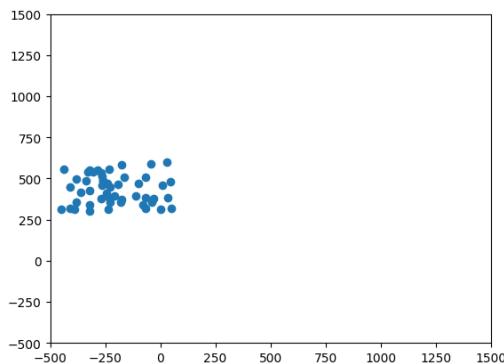
def updateBoids(boids):
    xs, ys, xvs, yvs = boids
    deltaXVs = [0] * len(xs)
    deltaYVs = [0] * len(xs)
    # Fly towards the middle
    for i in range(len(xs)):
        for j in range(len(xs)):
            deltaXVs[i] = deltaXVs[i] + (xs[j] - xs[i]) * 0.01 / len(xs)
    for i in range(len(xs)):
        for j in range(len(xs)):
            deltaYVs[i] = deltaYVs[i] + (ys[j] - ys[i]) * 0.01 / len(xs)
    # Fly away from nearby boids
    for i in range(len(xs)):
        for j in range(len(xs)):
            if (xs[j] - xs[i]) ** 2 + (ys[j] - ys[i]) ** 2 < 100:
                deltaXVs[i] = deltaXVs[i] + (xs[i] - xs[j])
                deltaYVs[i] = deltaYVs[i] + (ys[i] - ys[j])
    # Try to match speed with nearby boids
    for i in range(len(xs)):
        for j in range(len(xs)):
            if (xs[j] - xs[i]) ** 2 + (ys[j] - ys[i]) ** 2 < 1000:
                deltaXVs[i] = deltaXVs[i] + (xvs[j] - xvs[i]) * 0.125 / len(xs)
                deltaYVs[i] = deltaYVs[i] + (yvs[j] - yvs[i]) * 0.125 / len(xs)
    # Update velocities
    for i in range(len(xs)):
        xvs[i] = xvs[i] + deltaXVs[i]
        yvs[i] = yvs[i] + deltaYVs[i]
    # Move according to velocities
    for i in range(len(xs)):
        xs[i] = xs[i] + xvs[i]
        ys[i] = ys[i] + yvs[i]

    figure = plt.figure()
    axes = plt.axes(xlim=(-500, 1500), ylim=(-500, 1500))
    scatter = axes.scatter(boids[0], boids[1])

    def ANIMATE(frame):
        updateBoids(boids)
        scatter.set_offsets(list(zip(boids[0], boids[1])))

    anim = animation.FuncAnimation(figure, ANIMATE, frames=50, interval=50)

if __name__ == "__main__":
    plt.show()
```

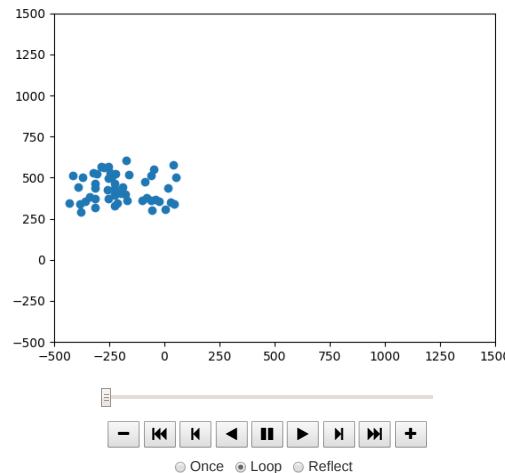


If you go into your folder and run the code:

```
cd bad_boids
python boids.py
```

You should be able to see some birds flying around, and then disappearing as they leave the window, like this:

```
from IPython.display import HTML
HTML(animated_jshtml())
```



Regression Test

First, have a look at the regression test we made (in the `record_fixture.py` file).

To create it, we saved out the before and after state for one iteration of some boids, using ipython:

```
from copy import deepcopy
import yaml
import boids

before = deepcopy(boids.boids)
boids.updateBoids(boids.boids)
after = boids.boids
fixture = {"before": before, "after": after}
with open("fixture.yml", "w") as fixture_file:
    fixture_file.write(yaml.safe_dump(fixture))
```

Invoking the test

Then, I used the fixture file to define the test (in `test_boids.py`):

```
import os
import yaml
from boids import updateBoids
from pytest import approx

def test_bad_boids_regression():
    with open(os.path.join(os.path.dirname(__file__), "fixture.yml")) as fixture_file:
        regression_data = yaml.safe_load(fixture_file)

    boid_data = regression_data["before"]
    updateBoids(boid_data)
    for after, before in zip(regression_data["after"], boid_data):
        for after_value, before_value in zip(after, before):
            assert after_value == approx(before_value)
```

Make the regression test fail

Check the tests pass:

```
pytest
```

Edit the file to make the test fail, see the fail, then reset it:

```
git checkout boids.py
```

Your Task

Transform bad boids **gradually** into better code, while making sure it still works, using a refactoring approach.

Each time you make a change:

- Ensure the regression test still passes
- Do a git commit on your fork, and write a commit message explaining the refactoring you did.

Try to keep the changes as small as possible.

If your refactoring creates any units (functions, modules, or classes), **write a unit test** for the unit (it's a good idea to not rely on regression testing).

Don't worry about the performance of the code, that's a topic for the "Programming for Speed" module later.

Refactoring Ideas

You probably won't have time to do all these in the session, but here are some refactorings we've seen in the module that you can try to apply here. We've loosely ordered them by where we'd suggest starting, but feel free to focus on the ones you're most interested in:

- Use linters to check and enforce a consistent style
- Ensure the code follows PEP8 conventions (e.g. for naming and whitespace)
- Consider whether any of the code "smells" and refactorings from 07_04_refactoring apply here
- Consider whether there is structure in the code that could be refactored into classes (see 07_05_object_oriented_design for ideas)

- Add type annotations

You may also like to apply some of what we've learned in previous modules, for example:

- Ensure dependencies are specified correctly
- Run tests and checks automatically, for example with a GitHub actions workflow
- Improve documentation
- Make the code into a Python package (e.g. see [module06_software_projects/06_04_packaging](#))

8. Advanced Programming Techniques

- Functional programming
- Metaprogramming
- Duck typing and exceptions
- Operator overloading
- Iterators and Generators

Contents

- [8.0 Advanced Python Programming](#) (5 minutes)
- [8.1 Functional programming](#) (20 minutes)
- [8.2 Iterators and Generators](#) (25 minutes)
- [8.3 Exceptions](#) (15 minutes)
- [8.4 Operator overloading](#) (20 minutes)
- [8.5 Metaprogramming](#) (20 minutes)
- [8.6 Advanced operator overloading](#) (20 minutes)

Total time: 2 hrs 5 minutes

Exercises

This module does not currently have any associated exercises.

8.0 Advanced Python Programming

Estimated time for this notebook: 5 minutes

... or, how to avoid repeating yourself.

Avoid Boiler-Plate

Code can often be annoyingly full of "boiler-plate" code: characters you don't really want to have to type.

Not only is this tedious, it's also time-consuming and dangerous: unnecessary code is an unnecessary potential place for mistakes.

There are two important phrases in software design that we've spoken of before in this context:

Once And Only Once
Don't Repeat Yourself (DRY)

All concepts, ideas, or instructions should be in the program in just one place. Every line in the program should say something useful and important.

We refer to code that respects this principle as DRY code.

In this chapter, we'll look at some techniques that can enable us to refactor away repetitive code.

Since in many of these places, the techniques will involve working with functions as if they were variables, we'll learn some **functional** programming. We'll also learn more about the innards of how Python implements classes.

We'll also think about how to write programs that generate the more verbose, repetitive program we could otherwise write. We call this **metaprogramming**.

8.1 Functional programming

Estimated time for this notebook: 20 minutes

We have previously seen the object-oriented style of programming, and how to organise our code according to it using objects, classes and inheritance. While widely-adopted and very useful, this is not the only way of writing code. The [functional paradigm](#), as the name suggests, emphasises functions as building blocks of programs.

Understanding to think in a functional programming style is almost as important as object orientation for building DRY, clear scientific software, and is just as conceptually difficult. However, being aware of different paradigms and styles gives you access to more techniques that you can use to write, structure and reason about your code.

Functions within functions

Programs are composed of functions: they take data in (which we call *parameters* or *arguments*) and send data out (through `return` statements).

A conceptual trick which is often used by computer scientists to teach the core idea of functional programming is this: to write a program, in theory, you only ever need functions with **one** argument, even when you think you need two or more. Why?

Let's define a program to add two numbers:

<code>def add(a, b):</code>
<code> return a + b</code>
<code>add(5, 6)</code>

How could we do this, in a fictional version of Python which only defined functions of one argument? In order to understand this, we'll have to understand several of the concepts of functional programming. Let's start with a program which just adds five to something:

```
def add_five(a):
    return a + 5
```

```
add_five(6)
```

```
11
```

OK, we could define lots of these, one for each number we want to add. But that would be infinitely repetitive. So, let's try to metaprogram that: we want a function which returns these `add_N()` functions.

Let's start with the easy case: a function which returns a function which adds 5 to something:

```
def generate_five_adder():
    def _five_adder(a):
        return a + 5

    return _five_adder
```

```
coolfunction = generate_five_adder()
coolfunction(7)
```

```
12
```

OK, so what happened there? Well, we defined a function **inside** the other function. We can always do that:

```
def thirty_function():
    def times_three(a):
        return a * 3

    def add_seven(a):
        return a + 7

    return times_three(add_seven(3))
```

```
thirty_function()
```

```
30
```

When we do this, the functions enclosed inside the outer function are **local** functions, and can't be seen outside:

```
add_seven
```

```
NameError: name 'add_seven' is not defined
```

There's not really much of a difference between functions and other variables in python. A function is just a variable which can have () put after it to call the code!

```
print(thirty_function)
```

```
<function thirty_function at 0x7ff2f485cd30>
```

```
x = [thirty_function, add_five, add]
```

```
for fun in x:
    print(fun)
```

```
<function thirty_function at 0x7ff2f485cd30>
<function add_five at 0x7ff2f485ce60>
<function add at 0x7ff2f485c040>
```

And we know that one of the things we can do with a variable is `return` it. So we can return a function, and then call it outside:

```
def deferred_greeting():
    def greet():
        print("Hello")

    return greet
```

```
friendlyfunction = deferred_greeting()
```

```
# Do something else
print("Just passing the time...")
```

```
Just passing the time...
```

```
# OK, Go!
friendlyfunction()
```

```
Hello
```

So now, to finish this, we just need to return a function to add an arbitrary amount:

```
def generate_adder(increment):
    def _adder(a):
        return a + increment

    return _adder
```

```
add_3 = generate_adder(3)
```

```
add_3(9)
```

```
12
```

We can make this even prettier: let's make another variable pointing to our `generate_adder()` function:

```
add = generate_adder
```

And now we can do the real magic:

```
add(8)(5)
```

```
13
```

In summary, we have started with a function that takes two arguments (`add(a, b)`) and replaced it with a new function (`add(a)(b)`). This new function takes a single argument, and returns a function that itself takes the second argument.

This may seem like an overly complicated process - and, in some cases, it is! However, this pattern of functions that return functions (or even take them as arguments!) can be very useful. In fact, it is the basis of decorators, a Python feature that we will discuss more [in this chapter \(notebook\)](#).

Closures

You may have noticed something a bit weird:

In the definition of `generate_adder`, `increment` is a local variable. It should have gone out of scope and died at the end of the definition. How can the amount the returned adder function is adding still be kept?

This is called a **closure**. In Python, whenever a function definition references a variable in the surrounding scope, it is preserved within the function definition.

You can close over global module variables as well:

```
name = "Eric"

def greet():
    print("Hello, ", name)

greet()
```

```
Hello, Eric
```

And note that the closure stores a reference to the variable in the surrounding scope: ("Late Binding")

```
name = "John"

greet()
```

```
Hello, John
```

Map and Reduce

We often want to apply a function to each variable in an array, to return a new array. We can do this with a list comprehension:

```
numbers = range(10)

[add_five(i) for i in numbers]
```

```
[5, 6, 7, 8, 9, 10, 11, 12, 13, 14]
```

But this is sufficiently common that there's a quick built-in:

```
list(map(add_five, numbers))
```

```
[5, 6, 7, 8, 9, 10, 11, 12, 13, 14]
```

This `map` operation is really important conceptually when understanding efficient parallel programming: different computers can apply the `mapped` function to their input at the same time. We call this Single Program, Multiple Data (SPMD). `map` is half of the `map-reduce` functional programming paradigm which is key to the efficient operation of much of today's "data science" explosion.

Let's continue our functional programming mind-stretch by looking at `reduce` operations.

We very often want to loop with some kind of accumulator (an intermediate result that we update), such as when finding a sum:

```
def summer(data):
    total = 0.0

    for x in data:
        total += x

    return total
```

```
summer(range(10))
```

```
45.0
```

or finding a maximum:

```

import sys

def my_max(data):
    # Start with the smallest possible number
    highest = -sys.float_info.max

    for x in data:
        if x > highest:
            highest = x

    return highest

```

```
my_max([2, 5, 10, -11, -5])
```

```
10
```

```
-sys.float_info.max
```

```
-1.7976931348623157e+308
```

These operations, where we have some variable which is building up a result, and the result is updated with some operation, can be gathered together as a functional program, taking in (as an argument) the operation to be used to combine results:

```

def accumulate(operation, data, initial):
    accumulator = initial
    for x in data:
        accumulator = operation(accumulator, x)
    return accumulator

```

```

def my_sum(data):
    def _add(a, b):
        return a + b

    return accumulate(_add, data, 0)

```

```
my_sum(range(5))
```

```
10
```

```

def bigger(a, b):
    if b > a:
        return b
    return a

```

```

def my_max(data):
    return accumulate(bigger, data, -sys.float_info.max)

```

```
my_max([2, 5, 10, -11, -5])
```

```
10
```

Anyway, this accumulate-under-an-operation process is so fundamental to computing that it's usually in standard libraries for languages which allow functional programming:

```

from functools import reduce

def my_max(data):
    return reduce(bigger, data, -sys.float_info.max)

my_max([2, 5, 10, -11, -5])

```

```
10
```

Efficient map-reduce

Now, because these operations, `bigger` and `_add`, are such that e.g. $(a+b)+c = a+(b+c)$, i.e. they are **associative**, we could apply our accumulation to the left half and the right half of the array, each on a different computer, and then combine the two halves:

$$1 + 2 + 3 + 4 = (1 + 2) + (3 + 4)$$

Indeed, with a bigger array, we can divide-and-conquer more times:

$$1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 = ((1 + 2) + (3 + 4)) + ((5 + 6) + (7 + 8))$$

So with enough parallel computers, we could do this operation on eight numbers in three steps: first, we use four computers to do one each of the pairwise adds.

Then, we use two computers to add the four totals.

Then, we use one of the computers to do the final add of the two last numbers.

You might be able to do the maths to see that with an N element list, the number of such steps is proportional to the logarithm of N.

We say that with enough computers, reduction operations are $O(\ln N)$

This course isn't an introduction to algorithms, but we'll talk more about this $O()$ notation when we think about programming for performance.

Lambda Functions

When doing functional programming, we often want to be able to define a function on the fly:

```

def most_Cs_in_any_sequence(sequences):
    def count_Cs(sequence):
        return sequence.count("C")
    counts = map(count_Cs, sequences)
    return max(counts)

def most_Gs_in_any_sequence(sequences):
    return max(map(lambda sequence: sequence.count("G"), sequences))

data = ["CGTA", "CGGGTAAACG", "GATTACA"]
most_Gs_in_any_sequence(data)

```

4

The syntax here means that these two definitions are identical:

```

func_name = lambda a, b, c: a + b + c

def func_name(a, b, c):
    return a + b + c

```

The **lambda** keyword defines an "anonymous" function.

```

def most_of_given_base_in_any_sequence(sequences, base):
    return max(map(lambda sequence: sequence.count(base), sequences))

most_of_given_base_in_any_sequence(data, "A")

```

3

The above fragment defined a lambda function as a **closure** over **base**. If you understood that, you've got it!

To double all elements in an array:

```

data = range(10)
list(map(lambda x: 2 * x, data))

[0, 2, 4, 6, 8, 10, 12, 14, 16, 18]

[2 * x for x in data]

[0, 2, 4, 6, 8, 10, 12, 14, 16, 18]

```

Similarly, to find the maximum value in a sequence:

```

def my_max(data):
    return reduce(lambda a, b: a if a > b else b, data, -sys.float_info.max)

my_max([2, 5, 10, -11, -5])

```

10

Using functional programming for numerical methods

Probably the most common use in research computing for functional programming is the application of a numerical method to a function.

Consider this example which uses the [newton function from SciPy](#), a root-finding function implementing the [Newton-Raphson method](#). The arguments we pass to `newton` are the function whose roots we want to find, and a starting point to search from.

We will be using this to find the roots of the function $f(x) = x^2 - x$.

```

%matplotlib inline

from matplotlib import pyplot as plt
from numpy import linspace, zeros
from scipy.optimize import newton

solve_me = lambda x: x**2 - x

for x0 in [2, 0.2]:
    answer = newton(solve_me, x0)
    print(f"Starting from {x0}, the root I found is {answer}")

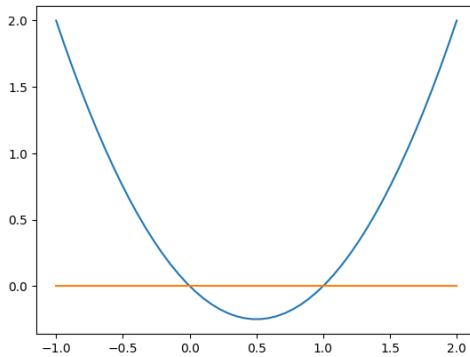
xs = linspace(-1, 2, 50)
solved = [x, list(map(solve_me, xs)), xs, zeros(len(xs))]

plt.plot(*solved)

```

```
Starting from 2, the root I found is 1.0
Starting from 0.2, the root I found is -3.441905100203782e-21
```

```
[<matplotlib.lines.Line2D at 0x7ff2c033b640>,
 <matplotlib.lines.Line2D at 0x7ff2c033b6a0>]
```



Sometimes such tools return another function, for example the derivative of their input function. This is what a naive implementation of that could look like:

```
def derivative_simple(func, eps, at):
    return (func(at + eps) - func(at)) / eps

def derivative(func, eps):
    def _func_derived(x):
        return (func(x + eps) - func(x)) / eps
    return _func_derived

straight = derivative(solve_me, 0.01)
```

The derivative of `solve_me` is $f'(x) = 2x - 1$, which represents a straight line. We can verify that our computations are correct, i.e. that the returned function `straight` matches $f'(x)$, by checking the value of `straight` at some x :

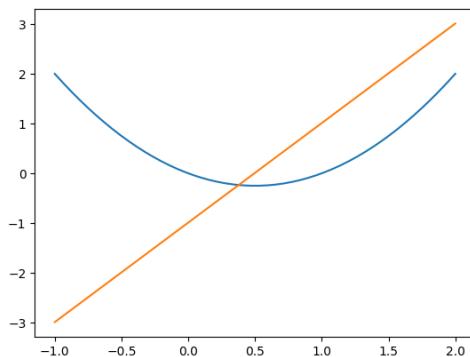
```
straight(3)
```

```
5.0099999999987
```

or by plotting it:

```
derived = (xs, list(map(solve_me, xs)), xs, list(map(derivative(solve_me, 0.01),
xs)))
plt.plot("derived")
print(newton(derivative(solve_me, 0.01), 0))
```

```
0.495000000000001
```



Of course, coding your own numerical methods is bad, because the implementations you develop are likely to be less efficient, less accurate and more error-prone than what you can find in existing established libraries.

For example, the above definition could be replaced by:

```
import scipy.misc

def derivative(func):
    def _func_derived(x):
        return scipy.misc.derivative(func, x)
    return _func_derived

newton(derivative(solve_me), 0)
```

```
0.5
```

If you've done a moderate amount of calculus, then you'll find similarities between functional programming in computer science and Functionals in the calculus of variations.

8.2 Iterators and Generators

Estimated time for this notebook: 25 minutes

In Python, anything which can be iterated over is called an iterable:

```
bowl = {"apple": 5, "banana": 3, "orange": 7}  
for fruit in bowl:  
    print(fruit.upper())
```

```
APPLE  
BANANA  
ORANGE
```

Surprisingly often, we want to iterate over something that takes a moderately large amount of memory to store - for example, our map images in the green-graph example.

Our green-graph example involved making an array of all the maps between London and Birmingham. This kept them all in memory *at the same time*: first we downloaded all the maps, then we counted the green pixels in each of them.

This would NOT work if we used more points: eventually, we would run out of memory. We need to use a **generator** instead. This chapter will look at iterators and generators in more detail: how they work, when to use them, how to create our own.

Iterators

Consider the basic python `range` function:

```
range(10)
```

```
range(0, 10)
```

```
total = 0  
for x in range(int(1e6)):  
    total += x
```

```
total
```

```
49999500000
```

In order to avoid allocating a million integers, `range` actually uses an **iterator**.

We don't actually need a million integers *at once*, just each integer *in turn* up to a million.

Because we can get an iterator from it, we say that a range is an **iterable**.

So we can `for`-loop over it:

```
for i in range(3):  
    print(i)
```

```
0  
1  
2
```

There are two important Python built-in functions for working with iterables. First is `iter`, which lets us create an iterator from any iterable object.

```
a = iter(range(3))
```

Once we have an iterator object, we can pass it to the `next` function. This moves the iterator forward, and gives us its next element:

```
next(a)
```

```
0
```

```
next(a)
```

```
1
```

```
next(a)
```

```
2
```

When we are out of elements, a `StopIteration` exception is raised:

```
next(a)
```

```
-----  
StopIteration                                Traceback (most recent call last)  
Cell In [9], line 1  
----> 1 next(a)  
  
StopIteration:
```

This tells Python that the iteration is over. For example, if we are in a `for i in range(3)` loop, this lets us know when we should exit the loop.

We can turn an iterable or iterator into a list with the `list` constructor function:

```
list(range(5))
```

```
[0, 1, 2, 3, 4]
```

Defining Our Own Iterable

When we write `next(a)`, under the hood Python tries to call the `__next__()` method of `a`. Similarly, `iter(a)` calls `a.__iter__()`.

We can make our own iterators by defining classes that can be used with the `next()` and `iter()` functions: this is the **iterator protocol**.

For each of the *concepts* in Python, like sequence, container, iterable, the language defines a *protocol*, a set of methods a class must implement, in order to be treated as a member of that concept.

To define an iterator, the methods that must be supported are `__next__()` and `__iter__()`.

`__next__()` must update the iterator.

We'll see why we need to define `__iter__` in a moment.

Here is an example of defining a custom iterator class:

```
class fib_iterator:
    """An iterator over part of the Fibonacci sequence."""

    def __init__(self, limit, seed1=1, seed2=1):
        self.limit = limit
        self.previous = seed1
        self.current = seed2

    def __iter__(self):
        return self

    def __next__(self):
        (self.previous, self.current) = (self.current, self.previous +
        self.current)
        self.limit -= 1
        if self.limit < 0:
            raise StopIteration()
        return self.current

x = fib_iterator(5)

next(x)

2

next(x)

3

next(x)

5

next(x)

8

for x in fib_iterator(5):
    print(x)

2
3
5
8
13

sum(fib_iterator(1000))

2979242185081433603368828199816319009156731305438197590327781734405367221904889045
2003450816384634553905509653388594324281497846904283041758626035944611524563466839
321019235741923828310479227982326069668668250
```

A shortcut to iterables: the `__iter__` method

In fact, we don't always have to define both `__iter__` and `__next__`!

If, to be iterated over, a class just wants to behave as if it were some other iterable, you can just implement `__iter__` and return `iter(some_other_iterable)`, without implementing `next`. For example, an image class might want to implement some metadata, but behave just as if it were just a 1-d pixel array when being iterated:

```
from matplotlib import pyplot as plt
from numpy import array

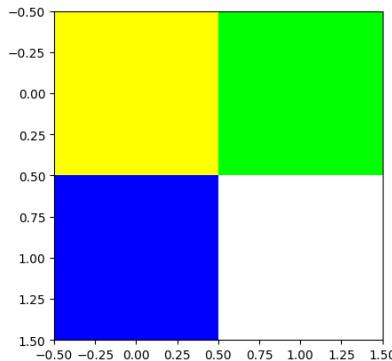
class MyImage:
    def __init__(self, pixels):
        self.pixels = array(pixels, dtype="uint8")
        self.channels = self.pixels.shape[2]

    def __iter__(self):
        # return an iterator over just the pixel values
        return iter(self.pixels.reshape(-1, self.channels))

    def show(self):
        plt.imshow(self.pixels, interpolation="None")

x = [[[255, 255, 0], [0, 255, 0]], [[0, 0, 255], [255, 255, 255]]]
image = MyImage(x)

%matplotlib inline
image.show()
```



```
image.channels
```

```
3
```

```
from webcolors import rgb_to_name
for pixel in image:
    print(rgb_to_name(pixel))
```

```
yellow
lime
blue
white
```

See how we used `image` in a `for` loop, even though it doesn't satisfy the iterator protocol (we didn't define both `__iter__` and `__next__` for it)?

The key here is that we can use any `iterable` object (like `image`) in a `for` expression, not just iterators! Internally, Python will create an iterator from the iterable (by calling its `__iter__` method), but this means we don't need to define a `__next__` method explicitly.

The `iterator` protocol is to implement both `__iter__` and `__next__`, while the `iterable` protocol is to implement `__iter__` and return an iterator.

Generators

There's a fair amount of "boiler-plate" in the above class-based definition of an iterable.

Python provides another way to specify something which meets the iterator protocol: **generators**.

```
def my_generator():
    yield 5
    yield 10
```

```
x = my_generator()
```

```
next(x)
```

```
5
```

```
next(x)
```

```
10
```

```
next(x)
```

```
StopIteration                                     Traceback (most recent call last)
Cell In [26], line 1
----> 1 next(x)

StopIteration:
```

```
for a in my_generator():
    print(a)
```

```
5
10
```

```
sum(my_generator())
```

```
15
```

A function which has `yield` statements instead of a `return` statement returns **temporarily**: it automatically becomes something which implements `__next__`.

Each call of `next()` returns control to the function where it left off.

Control passes back-and-forth between the generator and the caller. Our Fibonacci example therefore becomes a function rather than a class.

```
def yield_fibs(limit, seed1=1, seed2=1):
    current = seed1
    previous = seed2

    while limit > 0:
        limit -= 1
        current, previous = current + previous, current
        yield current
```

We can now use the output of the function like a normal iterable:

```
{ sum(yield_fibs(5))
```

```
31
```

```
{ for a in yield_fibs(10):
    if a % 2 == 0:
        print(a)
```

```
2
8
34
144
```

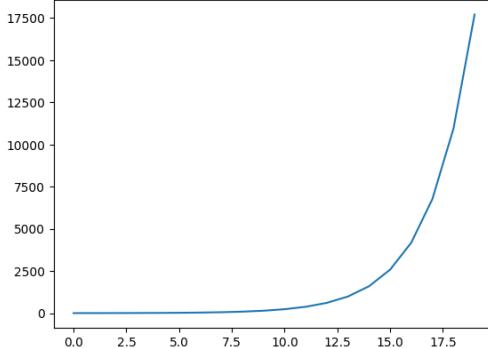
Sometimes we may need to gather all values from a generator into a list, such as before passing them to a function that expects a list:

```
{ list(yield_fibs(10))
```

```
[2, 3, 5, 8, 13, 21, 34, 55, 89, 144]
```

```
{ plt.plot(list(yield_fibs(20)))
```

```
[<matplotlib.lines.Line2D at 0x7f220466f340>]
```



```

from contextlib import contextmanager

@contextmanager
def verbose_context(name):
    print("Get ready for action, ", name)
    yield name.upper()
    print("You did it")

with verbose_context("Monty") as shouty:
    print(f"Doing it, {shouty}")

```

```

Get ready for action, Monty
Doing it, MONTY
You did it

```

Again, we use `yield` to temporarily return from a function.

Decorators

When doing functional programming, we may often want to define mutator functions which take in one function and return a new function, such as our derivative example earlier.

```

def repeat(func):
    def _repeated(x):
        return func(func(x))

    return _repeated

def hello(name):
    return f"Hello, {name}"

print(hello("Cleese"))
print(repeat(hello)("Cleese"))

```

```

Hello, Cleese
Hello, Hello, Cleese

```

Any function which accepts a function as its first argument and returns a function can be used as a **decorator** like this:

```

@repeat
def hello(name):
    return f"Hello, {name}"

hello("Cleese")

```

```
'Hello, Hello, Cleese'
```

We could also modify this to create a decorator that takes an argument specifying how many times the function should be repeated:

```

def repeater(count):
    def wrap_function_in_repeat(func):
        def _repeated(x):
            counter = count
            while counter > 0:
                counter -= 1
                x = func(x)
            return x

        return _repeated

    return wrap_function_in_repeat

```

```

from math import sqrt
fiftytimes = repeater(50)
fiftyroots = fiftytimes(sqrt)
fiftyroots(100)

```

```
1.00000000000004
```

```

@repeater(3)
def hello(name):
    return f"Hello, {name}"

hello("Cleese")

```

```
'Hello, Hello, Hello, Cleese'
```

It turns out that, quite often, we want to apply one of these to a function as we're defining a class. For example, we may want to specify that after certain methods are called, data should always be stored.

Much of Python's standard functionality is implemented as decorators: we've seen `@contextmanager`, `@classmethod` and `@attribute`. The `@contextmanager` metaclass, for example, takes in an iterator, and yields a class conforming to the context manager protocol.

Supplementary material

The remainder of this page contains an example of the flexibility of the features discussed above. Specifically, it shows how generators and context managers can be combined to create a testing framework like the one previously seen in the course.

Test generators

Earlier in the course we saw a test which loaded its test cases from a YAML file and asserted each input with each output. This was nice and concise, but had one flaw: we had just one test, covering all the fixtures, so we got just one `.` in the test output when we ran the tests, and if any test failed, the rest were not run. We can do a nicer job with a test **generator**:

```

import os

def assert_exemplar(**fixture):
    answer = fixture.pop("answer")
    assert_equal(greet(**fixture), answer)

def test_greeter():
    with open(
        os.path.join(os.path.dirname(__file__), "fixtures", "samples.yaml")
    ) as fixtures_file:
        fixtures = yaml.safe_load(fixtures_file)

    for fixture in fixtures:
        yield assert_exemplar(**fixture)

```

Each time a function beginning with `test_` does a `yield` it results in another test.

Negative test contexts managers

We have seen this:

```

from pytest import raises

with raises(AttributeError):
    x = 2
    x.foo()

```

We can now see how `pytest` might have implemented this:

```

@contextmanager
def reimplement_raises(exception):
    try:
        yield
    except exception:
        pass
    else:
        raise Exception("Expected", exception, " to be raised, nothing was.")

with reimplement_raises(AttributeError):
    x = 2
    x.foo()

```

Skip test decorators

Some frameworks also implement decorators for skipping tests or dealing with tests that are known to raise exceptions (due to known bugs or limitations).

For example:

```

%%writefile test_skipped.py
import pytest
import sys

@pytest.mark.skipif(sys.version_info < (4, 0), reason="requires python 4")
def test_python_4():
    raise RuntimeError("something went wrong")

```

Writing test_skipped.py

```
! pytest test_skipped.py
```

```

=====
platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0
rootdir: /home/runner/work/rse-course/rse-
course/module08_advanced_programming_techniques
plugins: anyio-3.6.2, pylama-8.4.1, cov-4.0.0
collecting ...
collected 1 item

test_skipped.py [100%]

===== 1 skipped in 0.01s =====

```

```

%%writefile test_not_skipped.py
import pytest
import sys

@pytest.mark.skipif(sys.version_info < (3, 0), reason="requires python 3")
def test_python_3():
    raise RuntimeError("something went wrong")

```

Writing test_not_skipped.py

```
! pytest test_not_skipped.py
```

```
===== test session starts =====
platform linux -- Python 3.8.14, pytest-7.2.0, pluggy-1.0.0
rootdir: /home/runner/work/rse-course/rse-
course/module08_advanced_programming_techniques
plugins: anyio-3.6.2, pylama-8.4.1, cov-4.0.0
collecting ...
collected 1 item
```

```
test_not_skipped.py
```

```
F [100%]
=====
===== FAILURES =====
----- test_python_3 -----

```

```
@pytest.mark.skipif(sys.version_info < (3, 0), reason="requires python 3")
def test_python_3():
>     raise RuntimeError("something went wrong")
E     RuntimeError: something went wrong

test_not_skipped.py:7: RuntimeError
=====
short test summary =====
FAILED test_not_skipped.py::test_python_3 - RuntimeError: something went wrong
=====
1 failed in 0.14s =====
```

We could reimplement this ourselves now too:

```
def homemade_skip_decorator(skip):
    def wrap_function(func):
        if skip:
            # if the test should be skipped, return a function
            # that just prints a message
            def do_nothing(*args):
                print("test was skipped")

            return do_nothing
        # otherwise use the original function as normal
        return func

    return wrap_function
```

```
@homemade_skip_decorator(3.9 < 4.0)
def test_skipped():
    raise RuntimeError("This test is skipped")
```

```
test_skipped()
```

```
test was skipped
```

```
@homemade_skip_decorator(3.9 < 3.0)
def test_runs():
    raise RuntimeError("This test is run")
```

```
test_runs()
```

```
RuntimeError                               Traceback (most recent call last)
Cell In [53], line 6
      1 @homemade_skip_decorator(3.9 < 3.0)
      2 def test_runs():
      3     raise RuntimeError("This test is run")
----> 6 test_runs()

Cell In [53], line 3, in test_runs()
      1 @homemade_skip_decorator(3.9 < 3.0)
      2 def test_runs():
----> 3     raise RuntimeError("This test is run")

RuntimeError: This test is run
```

8.3 Exceptions

Estimated time for this notebook: 15 minutes

When we learned about testing, we saw that Python complains when things go wrong by raising an "Exception" naming a type of error:

```
1 / 0
```

```
ZeroDivisionError                         Traceback (most recent call last)
Cell In [1], line 1
----> 1 1 / 0

ZeroDivisionError: division by zero
```

Exceptions are objects, forming a [class hierarchy](#). We just raised an instance of the `ZeroDivisionError` class, making the program crash. If we want more information about where this class fits in the hierarchy, we can use [Python's inspect module](#) to get a chain of classes, from `ZeroDivisionError` up to `object`:

```
import inspect
inspect.getmro(ZeroDivisionError)
```

```
(ZeroDivisionError, ArithmeticError, Exception, BaseException, object)
```

So we can see that a zero division error is a particular kind of Arithmetic Error.

```
x = 1
for y in x:
    print(y)
```

```

TypeError                                 Traceback (most recent call last)
Cell In [3], line 3
      1 x = 1
----> 3 for y in x:
      4     print(y)

TypeError: 'int' object is not iterable

```

```
inspect.getmro(TypeError)
```

```
(TypeError, Exception, BaseException, object)
```

Create your own Exception

When we were looking at testing, we saw that it is important for code to crash with a meaningful exception type when something is wrong. We raise an Exception with `raise`. Often, we can look for an appropriate exception from the standard set to raise.

However, we may want to define our own exceptions. Doing this is as simple as inheriting from `Exception` (or one of its subclasses):

```

class MyCustomErrorType(ArithmetricError):
    pass

raise MyCustomErrorType("Problem")

```

```

MyCustomErrorType                           Traceback (most recent call last)
Cell In [5], line 5
      1 class MyCustomErrorType(ArithmetricError):
      2     pass
----> 5 raise MyCustomErrorType("Problem")

MyCustomErrorType: Problem

```

You can add custom data to your exception:

```

class MyCustomErrorType(Exception):
    def __init__(self, category=None):
        self.category = category

    def __str__(self):
        return f"Error, category {self.category}"

raise MyCustomErrorType(404)

```

```

MyCustomErrorType                           Traceback (most recent call last)
Cell In [6], line 9
      5     def __str__(self):
      6         return f"Error, category {self.category}"
----> 9 raise MyCustomErrorType(404)

MyCustomErrorType: Error, category 404

```

The real power of exceptions comes, however, not in letting them crash the program, but in letting your program handle them. We say that an exception has been "thrown" and then "caught".

```

import yaml

try:
    config = yaml.safe_load(open("datasource.yaml"))
    user = config["userid"]
    password = config["password"]

except FileNotFoundError:
    print("No password file found, using anonymous user.")
    user = "anonymous"
    password = None

print(user)

```

```
No password file found, using anonymous user.
anonymous
```

Note that we specify only the error we expect to happen and want to handle. Sometimes you see code that catches everything:

```

try:
    config = yaml.safe_load(open("datasource.yaml"))
    user = config["userid"]
    password = config["password"]
except:
    user = "anonymous"
    password = None

print(user)

```

```
anonymous
```

This can be dangerous and can make it hard to find errors! There was a mistyped function name there (`safe_load`), but we did not notice the error, as the generic except caught it. Therefore, we should be specific and catch only the type of error we want.

Managing multiple exceptions

Let's create two credential files to read

```

with open("datasource2.yaml", "w") as outfile:
    outfile.write("userid: eidle\n")
    outfile.write("password: secret\n")

with open("datasource3.yaml", "w") as outfile:
    outfile.write("user: eidle\n")
    outfile.write("password: secret\n")

```

And create a function that reads credentials files and returns the username and password to use.

```

def read_credentials(source):
    try:
        datasource = open(source)
        config = yaml.safe_load(datasource)
        user = config["userid"]
        password = config["password"]
        datasource.close()
    except FileNotFoundError:
        print("Password file missing")
        user = "anonymous"
        password = None
    except KeyError:
        print("Expected keys not found in file")
        user = "anonymous"
        password = None
    return user, password

```

```
print(read_credentials("datasource2.yaml"))
```

```
('eidle', 'secret')
```

```
print(read_credentials("datasource.yaml"))
```

```
Password file missing
('anonymous', None)
```

```
print(read_credentials("datasource3.yaml"))
```

```
Expected keys not found in file
('anonymous', None)
```

This last code has a flaw: the file was successfully opened, the missing key was noticed, but not explicitly closed. It's normally OK, as Python will close the file as soon as it notices there are no longer any references to datasource in memory, after the function exits. But this is not good practice, you should keep a file handle for as short a time as possible.

```

def read_credentials(source):
    try:
        datasource = open(source)
        config = yaml.safe_load(datasource)

    try:
        print("File loaded, trying to extract credentials")
        user = config["userid"]
        password = config["password"]
    except KeyError:
        print("Expected keys not found in file")
        user = "anonymous"
        password = None
    finally:
        # Runs irrespective of whether keys found
        print("Closing file")
        datasource.close()

    except FileNotFoundError:
        print("Password file missing")
        user = "anonymous"
        password = None

    return user, password

```

The `finally` clause is executed whether or not an exception occurs.

The last optional clause of a `try` statement, an `else` clause is called only if an exception is NOT raised. It can be a better place than the `try` clause to put code other than that which you expect to raise the error, and which you do not want to be executed if the error is raised. It is executed in the same circumstances as code put in the end of the `try` block, the only difference is that errors raised during the `else` clause are not caught.

```

def read_credentials(source):
    try:
        datasource = open(source)

    except FileNotFoundError:
        print("Password file missing")
        user = "anonymous"
        password = None

    else:
        # Runs only if opening the file was successful
        config = yaml.safe_load(datasource)
        try:
            print("File loaded, trying to extract credentials")
            user = config["userid"]
            password = config["password"]
        except KeyError:
            print("Expected keys not found in file")
            user = "anonymous"
            password = None
        finally:
            # Runs irrespective of whether keys found
            print("Closing file")
            datasource.close()

    return user, password

```

Don't worry if `else` seems useless to you; most languages' implementations of try/except don't support such a clause. An alternative way of avoiding leaving the file open in the original implementation (and without using `else` or `finally`) is to use a context manager:

```

def read_credentials(source):
    try:
        with open(source) as datasource: # closes the file when done
            config = yaml.safe_load(datasource)
            user = config["userid"]
            password = config["password"]
    except FileNotFoundError:
        print("Password file missing")
        user = "anonymous"
        password = None
    except KeyError:
        print("Expected keys not found in file")
        user = "anonymous"
        password = None
    return user, password

```

Exceptions do not have to be caught close to the part of the program calling them. They can be caught anywhere "above" the calling point in the call stack: control can jump arbitrarily far in the program: up to the `except` clause of the "highest" containing try statement.

```
def f4(x):
    if x == 0:
        return
    if x == 1:
        raise ArithmeticError()
    if x == 2:
        raise SyntaxError()
    if x == 3:
        raise TypeError()
```

```
def f3(x):
    try:
        print("F3Before")
        f4(x)
        print("F3After")
    except ArithmeticError:
        print("F3Except (●)")
```

```
def f2(x):
    try:
        print("F2Before")
        f3(x)
        print("F2After")
    except SyntaxError:
        print("F2Except (●)")
```

```
def f1(x):
    try:
        print("F1Before")
        f2(x)
        print("F1After")
    except TypeError:
        print("F1Except (●)")
```

```
f1(0)
```

```
F1Before
F2Before
F3Before
F3After
F2After
F1After
```

```
f1(1)
```

```
F1Before
F2Before
F3Before
F3Except (●)
F2After
F1After
```

```
f1(2)
```

```
F1Before
F2Before
F3Before
F2Except (●)
F1After
```

```
f1(3)
```

```
F1Before
F2Before
F3Before
F1Except (●)
```

Design with Exceptions

Now we know how exceptions work, we need to think about the design implications... How best to use them.

Traditional software design theory will tell you that they should only be used to describe and recover from **exceptional** conditions: things going wrong. Normal program flow shouldn't use them.

Python's designers take a different view: use of exceptions in normal flow is considered OK. For example, all iterators raise a `StopIteration` exception to indicate the iteration is complete.

A commonly recommended Python design pattern is to use exceptions to determine whether an object implements a protocol (concept/interface), rather than testing on type.

For example, we might want a function which can be supplied *either* a data series or a path to a location on disk where data can be found. We can examine the type of the supplied content:

```
import yaml

def analysis(source):
    if type(source) == dict:
        name = source["modelname"]
    else:
        content = open(source)
        source = yaml.safe_load(content)
        name = source["modelname"]
    print(name)
```

```
analysis({"modelname": "Super"})
```

```
Super
```

```
with open("example.yaml", "w") as outfile:
    outfile.write("modelname: brilliant\n")
```

```
analysis("example.yaml")
```

```
brilliant
```

However, we can also use the try-it-and-handle-exceptions approach to this.

```
def analysis(source):
    try:
        name = source["modelname"]
    except TypeError:
        content = open(source)
        source = yaml.safe_load(content)
        name = source["modelname"]
    print(name)

analysis("example.yaml")
```

```
brilliant
```

This approach is more extensible, and behaves properly if we give it some other data-source which responds like a dictionary or string.

```
def analysis(source):
    try:
        name = source["modelname"]
    except TypeError:
        # Source was not a dictionary-like object
        # Maybe it is a file path
        try:
            content = open(source)
            source = yaml.safe_load(content)
            name = source["modelname"]
        except IOError:
            # Maybe it was already raw YAML content
            source = yaml.safe_load(source)
            name = source["modelname"]
    print(name)

analysis("modelname: Amazing")
```

```
Amazing
```

Re-Raising Exceptions

Sometimes we want to catch an error, partially handle it, perhaps add some extra data to the exception, and then re-raise to be caught again further up the call stack.

The keyword "raise" with no argument in an `except:` clause will cause the caught error to be re-thrown. Doing this is the only circumstance where it is safe to do `except:` without catching a specific type of error.

```
try:
    # Something
    pass
except:
    # Do this code here if anything goes wrong
    raise
```

If you want to be more explicit about where the error came from, you can use the `raise from` syntax, which will create a chain of exceptions:

```
def lower_function():
    raise ValueError("Error in lower function")

def higher_function():
    try:
        lower_function()
    except ValueError as e:
        raise RuntimeError("Error in higher function!") from e

higher_function()
```

```
ValueError                                Traceback (most recent call last)
Cell In [32], line 7, in higher_function()
      6     try:
      7         lower_function()
      8     except ValueError as e:
Cell In [32], line 2, in lower_function()
      1     def lower_function():
      2         raise ValueError("Error in lower function!")

ValueError: Error in lower function!

The above exception was the direct cause of the following exception:

RuntimeError                               Traceback (most recent call last)
Cell In [32], line 12
      8     except ValueError as e:
      9     raise RuntimeError("Error in higher function!") from e
--> 10 higher_function()

Cell In [32], line 9, in higher_function()
      7     lower_function()
      8 except ValueError as e:
--> 9     raise RuntimeError("Error in higher function!") from e

RuntimeError: Error in higher function!
```

It can be useful to catch and re-throw an error as you go up the chain, doing any clean-up needed for each layer of a program.

The error will finally be caught and not re-thrown only at a higher program layer that knows how to recover. This is known as the "throw low catch high" principle.

8.4 Operator overloading

Estimated time for this notebook: 15 minutes

We've seen already during the course that some operators behave differently depending on the data type.

For example, `+` adds numbers but concatenates strings or lists:

```
{ 4 + 2
  6
{ "4" + "2"
  '42'
```

`*` is used for multiplication, or repeated addition:

```
{ 6 * 7
  42
{ "me" * 3
  'mememe'
```

`/` is division for numbers, and wouldn't have a real meaning on strings. However, it's used to separate files and directories on your file system. Therefore, this has been *overloaded* in the `pathlib` module:

```
import os
from pathlib import Path

performance = Path("../") / "module07_construction_and_design"
os.listdir(performance)
```

```
['index.md',
'07_06_classes.ipynb',
'__pycache__',
'07_00_introduction.ipynb',
'flake8_example.py',
'mypy_example.py',
'anotherfile.py',
'.mypy_cache',
'.mypy_cache',
'config.yaml',
'black_example.py',
'07_05_object_oriented_design.ipynb',
'isort_example.py',
'07_01_comments.ipynb',
'07_07_design_patterns.ipynb',
'07_04_refactoring.ipynb',
'07_03_linters.ipynb',
'fixed.png',
'07_08_refactoring_boids.ipynb']
```

The above works because one of the elements is a `Path` object. Note, that the `/` works similarly to `os.path.join()`, so whether you are using Unix file systems or Windows, `pathlib` will know what path separator to use.

```
{ performance = os.path.join("../", "module07_construction_and_design")
```

Overloading operators for your own classes

Now that we have seen that in Python operators do different things, how can we use `+` or other operators on our own classes to achieve similar behaviour?

Let's go back to our Maze example, and simplify our room object so it's defined as:

```
class Room:
    def __init__(self, name, area):
        self.name = name
        self.area = area
```

We can now create a room as:

```
{ small = Room("small", 9)
print(small)
```

```
<__main__.Room object at 0x7f4d80c2ca0>
```

However, when we print it we don't get much information on the object. So, the first operator we are overloading is its string representation defining `__str__`:

```
class Room:
    def __init__(self, name, area):
        self.name = name
        self.area = area

    def __str__(self):
        return f"<Room: {self.name} {self.area}m²>"
```

```
{ small = Room("small", 9)
print(small)
```

```
<Room: small 9m²>
```

How can we add two rooms together? What does it mean? Let's define that the addition (`+`) of two rooms makes up one with the combined size. We produce this behaviour by defining the `__add__` method.

```
class Room:
    def __init__(self, name, area):
        self.name = name
        self.area = area

    def __add__(self, other):
        return Room(f"{self.name}-{other.name}", self.area + other.area)

    def __str__(self):
        return f"<Room: {self.name} {self.area}m²>"
```

```

small = Room("small", 9)
big = Room("big", 21)
print(small, big, small + big)

```

```
<Room: small 9m2> <Room: big 21m2> <Room: small_big 30m2>
```

Would the order of how the rooms are added affect the final room? As they are added now, the name is determined by the order, but do we want that? Or would we prefer to have:

```
small + big == big + small
```

That bring us to another operator, equal to: `==`. The method needed to produce such comparison is `__eq__`.

```

class Room:
    def __init__(self, name, area):
        self.name = name
        self.area = area

    def __add__(self, other):
        return Room(f"{self.name}_{other.name}", self.area + other.area)

    def __eq__(self, other):
        return self.area == other.area and set(self.name.split("_")) == set(
            other.name.split("_"))

```

So, in this way two rooms of the same area are "equal" if their names are composed by the same.

```

small = Room("small", 9)
big = Room("big", 21)
large = Room("superbig", 30)
print(small + big == big + small)
print(small + big == large)

```

```
True
False
```

You can add the other comparisons to know which room is bigger or smaller with the following functions:

Operator	Function
<code><</code>	<code>__lt__(self, other)</code>
<code><=</code>	<code>__le__(self, other)</code>
<code>></code>	<code>__gt__(self, other)</code>
<code>>=</code>	<code>__ge__(self, other)</code>

Let's add people to the rooms and check whether they are in one room or not.

```

class Room:
    def __init__(self, name, area):
        self.name = name
        self.area = area
        self.occupants = []

    def add_occupant(self, name):
        self.occupants.append(name)

circus = Room("Circus", 3)
circus.add_occupant("Graham")
circus.add_occupant("Eric")
circus.add_occupant("Terry")

```

How do we know if John is in the room? We can check the `occupants` list:

```
"John" in circus.occupants
```

```
False
```

Or making it more readable adding a membership definition:

```

class Room:
    def __init__(self, name, area):
        self.name = name
        self.area = area
        self.occupants = []

    def add_occupant(self, name):
        self.occupants.append(name)

    def __contains__(self, value):
        return value in self.occupants

circus = Room("Circus", 3)
circus.add_occupant("Graham")
circus.add_occupant("Eric")
circus.add_occupant("Terry")

```

```
"Terry" in circus
```

```
True
```

We can add lots more operators to classes. For example, `__getitem__` to let you index or access part of your object like a sequence or dictionary, e.g., `newObject[1]` or `newObject["data"]`, or `__len__` to return a number of elements in your object. Probably the most exciting one is `__call__`, which overrides the `()` operator; this allows us to define classes that *behave like functions!* We call these **callables**.

```

class Greeter:
    def __init__(self, greeting):
        self.greeting = greeting

    def __call__(self, name):
        print(self.greeting, name)

greeter_instance = Greeter("Hello")
greeter_instance("Eric")

```

```
Hello Eric
```

We've now come full circle in the blurring of the distinction between functions and objects! The full power of functional programming is really remarkable.

If you want to know more about the topics in this lecture, using a different language syntax, I recommend you watch the [Abelson and Sussman](#) "Structure and Interpretation of Computer Programs" lectures. These are the Computer Science equivalent of the Feynman Lectures!

Next notebook shows a detailed example of how to apply operator overloading to create your own symbolic algebra system.

8.5 Metaprogramming

⚠ Warning: Advanced Topic! ⚠

Estimated time for this notebook: 15 minutes

Metaprogramming globals

Consider a bunch of variables, each of which need initialising and incrementing:

```

bananas = 0
apples = 0
oranges = 0
bananas += 1
apples += 1
oranges += 1

```

The right hand side of these assignments doesn't respect the DRY principle. We could of course define a variable for our initial value:

```

initial_fruit_count = 0
bananas = initial_fruit_count
apples = initial_fruit_count
oranges = initial_fruit_count

```

However, this is still not as DRY as it could be: what if we wanted to replace the assignment with, say, a class constructor and a buy operation:

```

class Basket:
    def __init__(self):
        self.count = 0

    def buy(self):
        self.count += 1

bananas = Basket()
apples = Basket()
oranges = Basket()
bananas.buy()
apples.buy()
oranges.buy()

```

We had to make the change in three places. Whenever you see a situation where a refactoring or change of design might require you to change the code in multiple places, you have an opportunity to make the code DRYer.

In this case, metaprogramming for incrementing these variables would involve just a loop over all the variables we want to initialise:

```

baskets = [bananas, apples, oranges]
for basket in baskets:
    basket.buy()

```

However, this trick **doesn't** work for initialising a new variable:

```
baskets = [bananas, apples, oranges, kiwis]
```

```

-----+-----+-----+
NameError                                Traceback (most recent call last)
Cell In [5], line 1
----> 1 baskets = [bananas, apples, oranges, kiwis]

NameError: name 'kiwis' is not defined

```

So can we declare a new variable programmatically? Given a list of the **names** of fruit baskets we want, initialise a variable with that name?

Every module or class in Python, is, under the hood, a special dictionary storing the values in its **namespace**. `globals()` gives a reference to the attribute dictionary for the current module:

```

print("globals() is a\n", type(globals()))
print("\nWith these keys:\n", globals().keys())

```

```

globals() is a
<class 'dict'>

With these keys:
dict_keys(['__name__', '__doc__', '__package__', '__loader__', '__spec__',
'__builtins__', '__builtins__', '__ih', '__oh', '__dh', '__In', '__Out', '__get_ipython__',
'__exit__', '__quit__', '__open__', '__i', '__i', '__i', '__ii', '__iii', '__i1', 'bananas',
'apples', 'oranges', '__i2', '__initial_fruit_count', '__i3', 'Basket', '__i4',
'baskets', 'basket', '__i5', '__i6'])

```

We can access variables via this dictionary:

```
globals()["apples"]
```

```
<__main__.Basket at 0x7f22ac703fa0>
```

```
apples
```

```
<__main__.Basket at 0x7f22ac703fa0>
```

And create new variables by assigning to this dictionary:

```
basket_names = ["bananas", "apples", "oranges", "kiwis"]  
for name in basket_names:  
    globals().__[name] = Basket()  
  
    kiwis.count
```

```
0
```

This is **metaprogramming**.

I would NOT recommend using it for an example as trivial as the one above. A better, more Pythonic choice here would be to use a data structure to manage your set of fruit baskets:

```
baskets = {}  
for name in basket_names:  
    baskets[name] = Basket()  
  
baskets["kiwis"].count
```

```
0
```

Or even, using a dictionary comprehension:

```
baskets = {name: Basket() for name in baskets}  
baskets["kiwis"].count
```

```
0
```

Which is the nicest way to do this, I think. Code which feels like metaprogramming is needed to make it less repetitive can often instead be DRYed up using a refactored data structure, in a way which is cleaner and more easy to understand. Nevertheless, metaprogramming is worth knowing.

Metaprogramming class attributes

We can metaprogram the attributes of a **module** using the `globals()` function.

We will also want to be able to metaprogram a class, by accessing its attribute dictionary.

This will allow us, for example, to programmatically add members to a class.

```
class Boring:  
    pass
```

If we are adding our own attributes, we can just do so directly:

```
x = Boring()  
x.name = "Michael"
```

```
x.name
```

```
'Michael'
```

And these turn up, as expected, in an attribute dictionary for the class:

```
x.__dict__
```

```
{'name': 'Michael'}
```

We can use `getattr` to access this special dictionary:

```
getattr(x, "name")
```

```
'Michael'
```

If we want to add an attribute given its name as a string, we can use `setattr`:

```
setattr(x, "age", 75)
```

```
x.age
```

```
75
```

And we could do this in a loop to programmatically add many attributes.

The real power of accessing the attribute dictionary comes when we realise that there is *very little difference* between member data and member functions.

Now that we know, from our functional programming, that a **function is just a variable that can be called with ()**, we can set an attribute to a function, and it becomes a member function!

```
setattr(Boring, "describe", lambda self: f"{self.name} is {self.age}")
```

```
x.describe()
```

```
'Michael is 75'

x.describe

<bound method <lambda> of <__main__.Boring object at 0x7f22ac6f82e0>>

Boring.describe

<function __main__.<lambda>(self)>
```

Note that we set this method as an attribute of the class, not the instance, so it is available to other instances of `Boring`:

```
y = Boring()
y.name = "Terry"
y.age = 78

y.describe()

'Terry is 78'
```

We can define a standalone function, and then **bind** it to the class. Its first argument automatically becomes `self`.

```
import datetime

def broken_birth_year(b_instance):
    current = datetime.datetime.now().year
    return current - b_instance.age

Boring.birth_year = broken_birth_year

x.birth_year()

1947

x.birth_year

<bound method broken_birth_year of <__main__.Boring object at 0x7f22ac6f82e0>>

x.birth_year.__name__

'broken_birth_year'
```

Metaprogramming function locals

We can access the attribute dictionary for the local namespace inside a function with `locals()` but this *cannot be written to*.

Lack of safe programmatic creation of function-local variables is a flaw in Python.

```
class Person:
    def __init__(self, name, age, job, children_count):
        for var_name, value in locals().items():
            if var_name == "self":
                continue
            print(f"Setting self.{var_name} to {value}")
            setattr(self, var_name, value)

terry = Person("Terry", 78, "Screenwriter", 0)

Setting self.name to Terry
Setting self.age to 78
Setting self.job to Screenwriter
Setting self.children_count to 0

terry.first_name

-----
AttributeError                               Traceback (most recent call last)
Cell In [31], line 1
----> 1 terry.first_name

AttributeError: 'Person' object has no attribute 'first_name'
```

Metaprogramming warning!

Use this stuff **sparingly!**

The above example worked, but it produced Python code which is not particularly understandable. Remember, your objective when programming is to produce code which is **descriptive of what it does**.

The above code is **definitely** less readable, less maintainable and more error prone than:

```
class Person:
    def __init__(self, name, age, job, children_count):
        self.name = name
        self.age = age
        self.job = job
        self.children_count = children_count
```

Sometimes, metaprogramming will be **really** helpful in making non-repetitive code, and you should have it in your toolbox, which is why I'm teaching you it. But doing it all the time overcomplicates matters. We've talked a lot about the DRY principle, but there is another equally important principle:

KISS: Keep it simple, Stupid!

Whenever you write code and you think, "Gosh, I'm really clever", you're probably *doing it wrong*. Code should be about clarity, not showing off.

8.6 Advanced operator overloading

⚠ Warning: Advanced Topic! ⚠

Estimated time for this notebook: 15 minutes

Setup for this notebook

We need to use a metaprogramming trick to make this teaching notebook work. I want to be able to put explanatory text in between parts of a class definition, so I'll define a decorator to help me build up a class definition gradually.

```
def extend(class_to_extend):
    """
    Metaprogramming to allow gradual implementation of class during notebook.
    Thanks to http://www.ianbicking.org/blog/2007/08/opening-python-classes.html
    """

    def decorator(extending_class):
        for name, value in extending_class.__dict__.items():
            if name in ["__dict__", "__module__", "__weakref__", "__doc__"]:
                continue
            setattr(class_to_extend, name, value)
        return class_to_extend

    return decorator
```

Operator overloading

Imagine we wanted to make a library to describe some kind of symbolic algebra system:

```
class Term:
    def __init__(self, symbols=[], powers=[], coefficient=1):
        self.coefficient = coefficient
        self.data = dict(zip(symbols, powers))

class Expression:
    def __init__(self, terms):
        self.terms = terms
```

So that $5x^2y + 7x + 2$ might be constructed as:

```
first = Term(["x", "y"], [2, 1], 5)
second = Term(["x"], [1], 7)
third = Term([], [], 2)
result = Expression([first, second, third])
```

This is pretty cumbersome.

What we'd really like is to have `2x+y` give an appropriate expression.

First, we'll define things so that we can construct our terms and expressions in different ways.

```
class Term:
    def __init__(self, *args):
        lead = args[0]
        if type(lead) == type(self):
            # Copy constructor
            self.data = dict(lead.data)
            self.coefficient = lead.coefficient
        elif type(lead) == int:
            self.from_constant(lead)
        elif type(lead) == str:
            self.from_symbol(*args)
        elif type(lead) == dict:
            self.from_dictionary(*args)
        else:
            self.from_lists(*args)

    def from_constant(self, constant):
        self.coefficient = constant
        self.data = {}

    def from_symbol(self, symbol, coefficient=1, power=1):
        self.coefficient = coefficient
        self.data = {symbol: power}

    def from_dictionary(self, data, coefficient=1):
        self.data = data
        self.coefficient = coefficient

    def from_lists(self, symbols=[], powers=[], coefficient=1):
        self.coefficient = coefficient
        self.data = dict(zip(symbols, powers))

class Expression:
    def __init__(self, terms=[]):
        self.terms = list(terms)
```

We could define `add()` and `multiply()` operations on expressions and terms:

```
@extend(Term)
class Term:
    def add(self, *others):
        return Expression((self,) + others)
```

```

@extend(Term)
class Term:
    def multiply(self, *others):
        result_data = dict(self.data)
        result_coeff = self.coefficient
        # Convert arguments to Terms first if they are
        # constants or integers
        others = map(Term, others)

        for another in others:
            for symbol, power in another.data.items():
                if symbol in result_data:
                    result_data[symbol] += power # add the powers together
                else:
                    result_data[symbol] = power
                    result_coeff *= another.coefficient

        return Term(result_data, result_coeff)

@extend(Expression)
class Expression:
    def add(self, *others):
        result = Expression(self.terms)

        for another in others:
            if type(another) == Term:
                result.terms.append(another)
            else:
                result.terms += another.terms

        return result

```

We can now construct the above expression as:

```

x = Term("x")
y = Term("y")

first = Term(5).multiply(x, x, y)
second = Term(7).multiply(x)
third = Term(2)
expr = first.add(second, third)

```

This is better, but we still can't write the expression in a 'natural' way.

However, we can define what `*` and `+` do when applied to Terms!:

```

@extend(Term)
class Term:
    def __add__(self, other):
        return self.add(other)

    def __mul__(self, other):
        return self.multiply(other)

@extend(Expression)
class Expression:
    def multiply(self, another):
        # Distributive law left as exercise
        pass

    def __add__(self, other):
        return self.add(other)

x_plus_y = Term("x") + "y"
x_plus_y.terms[1]

'y'

five_x_ysq = Term("x") * 5 * "y" * "y"
print(five_x_ysq.data, five_x_ysq.coefficient)

{'x': 1, 'y': 2} 5

```

This is called operator overloading. We can define what `add` and `multiply` mean when applied to our class.

Note that this only works so far if we multiply on the right-hand-side! However, we can define a multiplication that works backwards, which is used as a fallback if the left multiply raises an error:

```

@extend(Expression)
class Expression:
    def __radd__(self, other):
        return self.__add__(other)

@extend(Term)
class Term:
    def __rmul__(self, other):
        return self.__mul__(other)

    def __radd__(self, other):
        return self.__add__(other)

5 * Term("x")

<__main__.Term at 0x7fe52c415940>

```

It's not easy at the moment to see if these things are working!

```

fivex = 5 * Term("x")
fivex.data, fivex.coefficient

({'x': 1}, 5)

```

We can add another operator method `__str__`, which defines what happens if we try to print our class:

```
@extend(Term)
class Term:
    def __str__(self):
        def symbol_string(symbol, power):
            if power == 1:
                return symbol
            return f'{symbol}^{power}'
        symbol_strings = [
            symbol_string(symbol, power) for symbol, power in self.data.items()
        ]
        prod = "*".join(symbol_strings)

        if not prod:
            return str(self.coefficient)
        if self.coefficient == 1:
            return prod
        return f'{self.coefficient}*{prod}'

@extend(Expression)
class Expression:
    def __str__(self):
        return "+".join(map(str, self.terms))
```

Now let's test it.

```
first = Term(5) * "x" * "x" * "y"  
second = Term(7) * "x"  
third = Term(2)  
expr = first + second + third
```

```
print(expr)
```

$$5^*x^2*y+7^*x+2$$

9. Programming for Speed

- Optimisation
 - Profiling
 - Scaling laws
 - NumPy
 - Cython

Contents

- [9.0 Performance programming](#) (10 minutes)
 - [9.1 Optimising Mandelbrot](#) (15 minutes)
 - [9.2 Optimising with NumPy](#) (30 minutes)
 - [9.3 Optimising with Cython](#) (25 minutes)
 - [9.4 Optimising with Numba](#) (20 minutes)
 - [9.5 Performance scaling for containers and algorithms](#) (20 minutes)

Total time: 2 hrs

Exercises

This module does not currently have any associated exercises.

9.0 Performance programming

Estimated time for this notebook: 10 minutes

We've spent most of this course looking at how to make code readable and reliable. For research work, it is often also important that code is efficient: that it does what it needs to do *quickly*.

It is very hard to work out beforehand whether code will be efficient or not: it is essential to *Profile* code, to measure its performance, to determine what aspects of it are slow.

When we looked at Functional programming, we claimed that code which is conceptualised in terms of actions on whole data-sets rather than individual elements is more efficient. Let's measure the performance of some different ways of implementing some code and see how they perform.

Two Mandelbrots

You're probably familiar with a famous fractal called the [Mandelbrot Set](#).

For a complex number c , c is in the Mandelbrot set if the series $z_{i+1} = z_i^2 + c$ (with $z_0 = 0$) does not tend to infinity. Traditionally, we plot a color showing how many steps are needed before $|z_i| > 2$. At this point we are sure that c is *not* in the Mandelbrot set as the series will diverge.

Here's a trivial python implementation:

```
def mandel(position, limit=50):  
    value = position  
  
    while abs(value) < 2:  
        limit -= 1  
        value = value**2 + position  
        if limit < 0:  
            return 0  
  
    return limit
```

```

xmin = -1.5
ymin = -1.0
xmax = 0.5
ymax = 1.0
resolution = 300
xstep = (xmax - xmin) / resolution
ystep = (ymax - ymin) / resolution
xs = [(xmin + xstep * i) for i in range(resolution)]
ys = [(ymin + ystep * i) for i in range(resolution)]

```

```

%%timeit
data = [[mandeli(complex(x, y)) for x in xs] for y in ys]

```

600 ms ± 1.51 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)

```

data1 = [[mandeli(complex(x, y)) for x in xs] for y in ys]

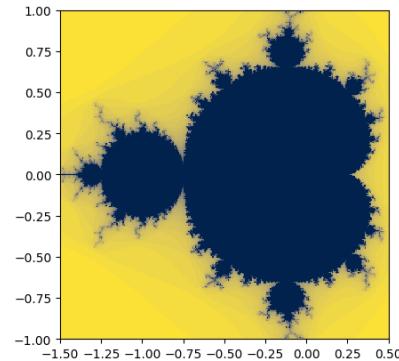
```

```

%matplotlib inline
import matplotlib.pyplot as plt
plt.set_cmap("cividis") # use a CVD-friendly palette
plt.imshow(data1, interpolation="none", extent=[xmin, xmax, ymin, ymax])

```

<matplotlib.image.AxesImage at 0x7f7da212f880>



We will learn this lesson how to make a version of this code which works Ten Times faster:

```

import numpy as np

# Do not worry about how this function works - it will be covered in detail later
def mandel_numpy(position, limit=50):
    value = position
    diverged_at_count = np.zeros(position.shape)
    while limit > 0:
        limit -= 1
        value = value**2 + position
        diverging = value * np.conj(value) > 4
        first_diverged_this_time = np.logical_and(diverging, diverged_at_count == 0)
        diverged_at_count[first_diverged_this_time] = limit
        value[diverging] = 2
    return diverged_at_count

```

```

ymatrix, xmatrix = np.mgrid[ymin:ymax:ystep, xmin:xmax:xstep]

```

```

values = xmatrix + 1j * ymatrix

```

```

data_numpy = mandel_numpy(values)

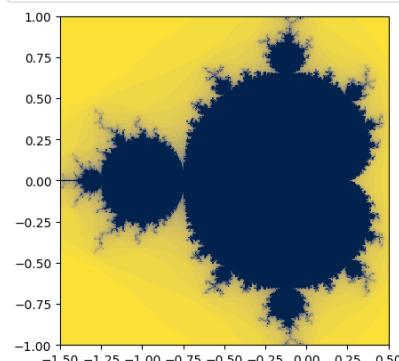
```

```

%matplotlib inline
plt.imshow(data_numpy, interpolation="none", extent=[xmin, xmax, ymin, ymax])

```

<matplotlib.image.AxesImage at 0x7f7da0056bb0>



```

%%timeit
data_numpy = mandel_numpy(values)

```

40.8 ms ± 47.9 µs per loop (mean ± std. dev. of 7 runs, 10 loops each)

Note we get the same answer:

```
(data_numpy == data1).all()
```

```
True
```

9.1 Optimising Mandelbrot

Estimated time for this notebook: 5 minutes

Let's start by reproducing our `mandel1` function and its output, `data1`, from the previous notebook.

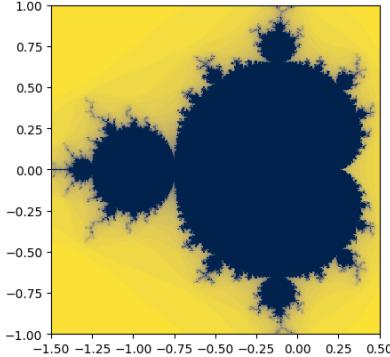
```
xmin = -1.5
ymin = -1.0
xmax = 0.5
ymax = 0.5
resolution = 300
xstep = (xmax - xmin) / resolution
ystep = (ymax - ymin) / resolution
xs = [(xmin + xstep * i) for i in range(resolution)]
ys = [(ymin + ystep * i) for i in range(resolution)]
```

```
def mandel1(position, limit=50):
    value = position
    while abs(value) < 2:
        limit -= 1
        value = value**2 + position
        if limit < 0:
            return 0
    return limit
```

```
data1 = [[mandel1(complex(x, y)) for x in xs] for y in ys]
```

```
from matplotlib import pyplot as plt
plt.set_cmap("cividis") # use a CVD-friendly palette
plt.imshow(data1, interpolation="none", extent=[xmin, xmax, ymin, ymax])
```

```
<matplotlib.image.AxesImage at 0x7fb7069dcfd0>
```



Many Mandelbrot

Let's compare our naive python implementation which used a list comprehension, taking around 500ms, with the following:

```
def mandel_append():
    data = []
    for y in ys:
        row = []
        for x in xs:
            row.append(mandel1(complex(x, y)))
        data.append(row)
    return data
```

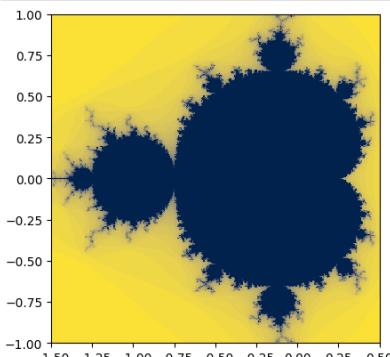
```
%%timeit
data2 = mandel_append()
```

```
607 ms ± 242 µs per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

Interestingly, not much difference. I would have expected this to be slower, due to the normally high cost of `appending` to data.

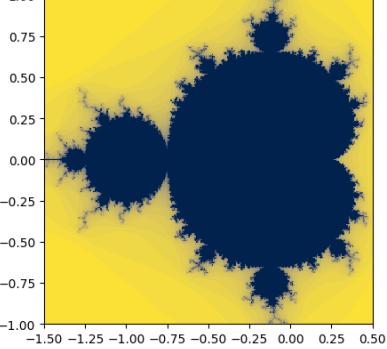
```
data2 = mandel_append()
plt.imshow(data2, interpolation="none", extent=[xmin, xmax, ymin, ymax])
```

```
<matplotlib.image.AxesImage at 0x7fb70188c1c0>
```



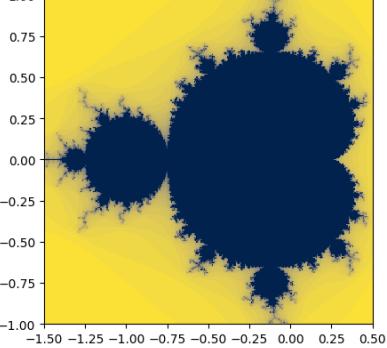
We ought to be checking if these results are the same by comparing the values in a test, rather than re-plotting. This is cumbersome in pure Python, but easy with NumPy, so we'll do this later.

Let's try a pre-allocated data structure:

```
data3 = [[0 for i in range(resolution)] for j in range(resolution)]  
  
def mandel_preallocated(data_structure):  
    for j, y in enumerate(ys):  
        for i, x in enumerate(xs):  
            data_structure[j][i] = mandel1(complex(x, y))  
  
%%timeit  
mandel_preallocated(data3)  
  
611 ms ± 3.2 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)  
  
mandel_preallocated(data3)  
plt.imshow(data3, interpolation="none", extent=[xmin, xmax, ymin, ymax])  
  
<matplotlib.image.AxesImage at 0x7fb701888cd0>  

```

Nope, no gain there.

Let's try using functional programming approaches:

```
def mandel_functional(ys):  
    data = []  
    for y in ys:  
        bind_mandel = lambda x: mandel1(complex(x, y))  
        data.append(list(map(bind_mandel, xs)))  
  
    return data  
  
%%timeit  
data4 = mandel_functional(ys)  
  
611 ms ± 1.57 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)  
  
data4 = mandel_functional(ys)  
plt.imshow(data4, interpolation="none", extent=[xmin, xmax, ymin, ymax])  
  
<matplotlib.image.AxesImage at 0x7fb701795d30>  

```

That was a tiny bit slower.

So, what do we learn from this? Our mental image of what code should be faster or slower is often wrong, or doesn't make much difference. The only way to really improve code performance is empirically, through measurements.

9.2 Optimising with NumPy

Estimated time for this notebook: 30 minutes

NumPy constructors

We saw previously that NumPy's core type is the `ndarray`, or N-Dimensional Array:

```
import numpy as np  
np.zeros([3, 4, 2, 5])[2, :, :, 1]
```

```
array([[0., 0.],
       [0., 0.],
       [0., 0.],
       [0., 0.]])
```

The real magic of numpy arrays is that most python operations are applied, quickly, on an elementwise basis:

```
x = np.arange(0, 256, 4).reshape(8, 8)

y = np.zeros((8, 8))

%%timeit
for i in range(8):
    for j in range(8):
        y[i][j] = x[i][j] + 10

40.7 µs ± 726 ns per loop (mean ± std. dev. of 7 runs, 10,000 loops each)

x + 10

array([[ 10,  14,  18,  22,  26,  30,  34,  38],
       [ 42,  46,  50,  54,  58,  62,  66,  70],
       [ 74,  78,  82,  86,  90,  94,  98, 102],
       [106, 110, 114, 118, 122, 126, 130, 134],
       [138, 142, 146, 150, 154, 158, 162, 166],
       [170, 174, 178, 182, 186, 190, 194, 198],
       [202, 206, 210, 214, 218, 222, 226, 230],
       [234, 238, 242, 246, 250, 254, 258, 262]])
```

Numpy's mathematical functions also happen this way, and are said to be "vectorized" functions.

```
np.sqrt(x)

array([[ 0.          ,  2.          ,  2.82842712,  3.46410162,  4.          ,
       4.47213595,  4.89897949,  5.29150262],
       [ 5.65685425,  6.          ,  6.32455532,  6.63324958,  6.92820323,
       7.21110255,  7.48331477,  7.74596669],
       [ 8.          ,  8.24621125,  8.48528137,  8.71779789,  8.94427191,
       9.16515139,  9.38083152,  9.59166305],
       [ 9.79795897, 10.          , 10.19803903, 10.39230485, 10.58300524,
       10.77032961, 10.95445115, 11.13552873],
       [11.3137085 , 11.48912529, 11.66198379, 11.83215957, 12.          ,
       12.16552506, 12.32882801, 12.489996 ],
       [12.6911064 , 12.80624847, 12.9614814 , 13.11487705, 13.26649916,
       13.41640786, 13.56465997, 13.7113092 ],
       [13.85640646, 14.          , 14.14213562, 14.28285686, 14.4222051 ,
       14.56021978, 14.69693846, 14.83239697],
       [14.96662955, 15.09966887, 15.23154621, 15.3622915 , 15.49193338,
       15.62049935, 15.74801575, 15.87450787]])
```

Numpy contains many useful functions for creating matrices. In our earlier lectures we've seen `linspace` and `arange` for evenly spaced numbers.

```
np.linspace(0, 10, 21)

array([ 0. ,  0.5,  1. ,  1.5,  2. ,  2.5,  3. ,  3.5,  4. ,  4.5,  5. ,
       5.5,  6. ,  6.5,  7. ,  7.5,  8. ,  8.5,  9. ,  9.5, 10. ])

np.arange(0, 10, 0.5)

array([ 0. ,  0.5,  1. ,  1.5,  2. ,  2.5,  3. ,  3.5,  4. ,  4.5,  5. ,
       5.5,  6. ,  6.5,  7. ,  7.5,  8. ,  8.5,  9. ,  9.5])
```

Here's one for creating matrices like coordinates in a grid:

```
xmin = -1.5
ymin = -1.0
xmax = 0.5
ymax = 1.0
resolution = 300
xstep = (xmax - xmin) / resolution
ystep = (ymax - ymin) / resolution
# A numpy "meshgrid" creates a rectangular grid from an array of x values and an
# array of y values.
ymatrix, xmatrix = np.mgrid[ymin:ymax:ystep, xmin:xmax:xstep]

print(ymatrix)

[[ -1.          -1.          -1.          ... -1.          -1.
   -1.          ],
 [-0.99333333 -0.99333333 -0.99333333 ... -0.99333333 -0.99333333
 -0.99333333],
 [-0.98666667 -0.98666667 -0.98666667 ... -0.98666667 -0.98666667
 -0.98666667],
 ...
 [ 0.98         0.98         0.98         ... 0.98         0.98
 0.98         ],
 [ 0.98666667  0.98666667  0.98666667 ... 0.98666667  0.98666667
 0.98666667],
 [ 0.99333333  0.99333333  0.99333333 ... 0.99333333  0.99333333
 0.99333333]]
```

We can add these together to make a grid containing the complex numbers we want to test for membership in the Mandelbrot set.

```
values = xmatrix + 1j * ymatrix

print(values)
```

```

[[ -1.5          -1.j           -1.49333333-1.j           -1.48666667-1.j
...   0.48          -1.j           0.48666667-1.j
0.49333333-1.j           ... 1.49333333-0.9933333j -1.48666667-0.9933333j
[-1.5          -0.99333333j -1.49333333-0.9933333j -1.48666667-0.9933333j
...   0.48          -0.9933333j  0.48666667-0.9933333j
0.49333333-0.9933333j]
[-1.5          -0.98666667j -1.49333333-0.98666667j -1.48666667-0.98666667j
...   0.48          -0.98666667j  0.48666667-0.98666667j
0.49333333-0.98666667j]
...
[-1.5          +0.98j          -1.49333333+0.98j          -1.48666667+0.98j
...   0.48          +0.98j          0.48666667+0.98j
0.49333333+0.98j]
[-1.5          +0.98666667j -1.49333333+0.98666667j -1.48666667+0.98666667j
...   0.48          +0.98666667j  0.48666667+0.98666667j
0.49333333+0.98666667j]
[-1.5          +0.99333333j -1.49333333+0.9933333j -1.48666667+0.9933333j
...   0.48          +0.9933333j  0.48666667+0.9933333j
0.49333333+0.9933333j]]
```

Arraywise Algorithms

We can use this to apply the mandelbrot algorithm to whole ARRAYS

```

z0 = values
z1 = z0 * z0 + values
z2 = z1 * z1 + values
z3 = z2 * z2 + values

print(z3)

[[24.06640625+20.75j    23.16610231+20.97899073j
22.27540349+21.18465854j ... 11.20523832 -1.88650846j
11.573453 -1.6076251j 11.94394738 -1.31225596j]
[23.82102149+19.85687829j 22.94415031+20.09504528j
22.07634812+20.31020645j ... 10.93323949 -1.5275283j
11.28531994 -1.24641067j 11.63928527 -0.94911594j]
[23.56689029+18.98729424j 22.71312709+19.23410533j
21.86791017+19.4582314j ... 10.65905064 -1.18433756j
10.99529965 -0.90137318j 11.33305161 -0.60254144j]
...
[23.30453709-18.14999998j 22.47355537-18.39585192j
21.65061048-18.62842771j ... 10.38305264 +0.85663887j
10.70377437 +0.57220289j 11.02562928 +0.27221042j]
[23.56689029-18.98729424j 22.71312709-19.23410533j
21.86791017-19.4582314j ... 10.65905064 +1.18433756j
10.99529965 +0.90137318j 11.33305161 +0.60254144j]
[23.82102149-19.85687829j 22.94415031-20.09504528j
22.07634812-20.31020645j ... 10.93323949 +1.5275283j
11.28531994 +1.24641067j 11.63928527 +0.94911594j]]
```

So can we just apply our `mandel1` function to the whole matrix?

```

def mandel1(position, limit=50):
    value = position
    while abs(value) < 2:
        limit -= 1
        value = value**2 + position
        if limit < 0:
            return 0
    return limit

mandel1(values)

-----
ValueError                                     Traceback (most recent call last)
Cell In [16], line 1
----> 1 mandel1(values)

Cell In [15], line 3, in mandel1(position, limit)
    1 def mandel1(position, limit=50):
    2     value = position
----> 3     while abs(value) < 2:
    4         limit -= 1
    5         value = value**2 + position

ValueError: The truth value of an array with more than one element is ambiguous.
Use a.any() or a.all()
```

No. The *logic* of our current routine would require stopping for some elements and not for others.

We can ask numpy to **vectorise** our method for us:

```

@np.vectorize
def mandel2(position, limit=50):
    value = position
    while abs(value) < 2:
        limit -= 1
        value = value**2 + position
        if limit < 0:
            return 0
    return limit
```

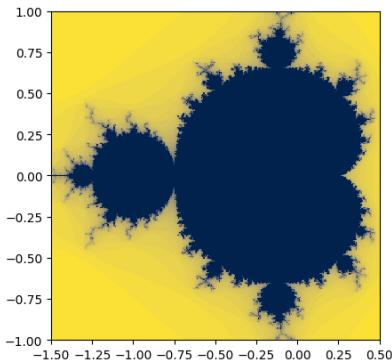
Note that we use a **decorator** here (`np.vectorize` takes a function as input and returns a function). An equivalent way to write this would be `mandel2 = np.vectorize(mandel1)`.

```

data5 = mandel2(values)

from matplotlib import pyplot as plt
plt.set_cmap("cividis") # use a CVD-friendly palette
plt.imshow(data5, interpolation="none", extent=[xmin, xmax, ymin, ymax])
```

```
<matplotlib.image.AxesImage at 0x7f3d65018250>
```



Is that any faster?

```
%%timeit  
data5 = mandel2(values)
```

```
596 ms ± 66.7 µs per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

This is not significantly faster. When we use vectorize it's just hiding an plain old python for loop under the hood. We want to make the loop over matrix elements take place in the "C Layer".

What if we just apply the Mandelbrot algorithm without checking for divergence until the end:

```
def mandel_numpy_explode(position, limit=50):  
    value = position  
    while limit > 0:  
        limit -= 1  
        value = value**2 + position  
  
    return abs(value) < 2
```

```
data6 = mandel_numpy_explode(values)
```

```
/tmp/ipykernel_11558/3249494863.py:5: RuntimeWarning: overflow encountered in  
square  
    value = value**2 + position  
/tmp/ipykernel_11558/3249494863.py:5: RuntimeWarning: invalid value encountered in  
square  
    value = value**2 + position
```

OK, we need to prevent it from running off to ∞

```
def mandel_numpy(position, limit=50):  
    value = position  
    while limit > 0:  
        limit -= 1  
        value = value**2 + position  
        diverging = abs(value) > 2  
        # Avoid overflow  
        value[diverging] = 2  
  
    return abs(value) < 2
```

```
data6 = mandel_numpy(values)
```

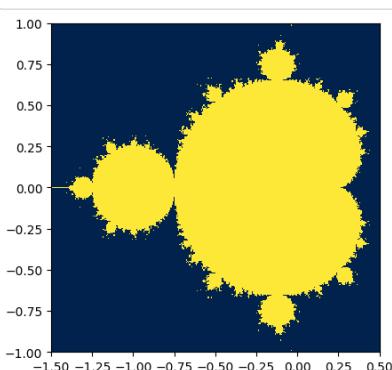
```
%%timeit
```

```
data6 = mandel_numpy(values)
```

```
27.8 ms ± 49.5 µs per loop (mean ± std. dev. of 7 runs, 10 loops each)
```

```
from matplotlib import pyplot as plt  
%matplotlib inline  
plt.imshow(data6, interpolation="none", extent=[xmin, xmax, ymin, ymax])
```

```
<matplotlib.image.AxesImage at 0x7f3d64f2d7f0>
```



Wow, that was TEN TIMES faster.

There's quite a few NumPy tricks there, let's remind ourselves of how they work:

```


diverging = abs(z3) > 2



z3[diverging] = 2


```

When we apply a logical condition to a NumPy array, we get a logical array.

```


x = np.arange(10)



y = np.ones([10]) * 5



z = x > y



x



array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])



y



array([5., 5., 5., 5., 5., 5., 5., 5., 5., 5.])



print(z)



[False False False False False  True  True  True  True]


```

Logical arrays can be used to index into arrays:

```


x[x > 3]



array([4, 5, 6, 7, 8, 9])



x[np.logical_not(z)]



array([0, 1, 2, 3, 4, 5])


```

And you can use such an index as the target of an assignment:

```


x[z] = 5



x



array([0, 1, 2, 3, 4, 5, 5, 5, 5])


```

Note that we didn't compare two arrays to get our logical array, but an array to a scalar integer – this is referred to as *broadcasting*.

More Mandelbrot

Of course, we didn't calculate the number-of-iterations-to-diverge, just whether the point was in the set.

Let's correct our code to do that:

```


def mandel4(position, limit=50):
    value = position
    # An array which keeps track of the first step at which each position diverged
    diverged_at_count = np.zeros(position.shape)
    while limit > 0:
        limit -= 1
        value = value**2 + position
        diverging = abs(value) > 2
        # Any positions which are:
        # - diverging
        # - haven't diverged before
        # are diverging for the first time
        first_diverged_this_time = np.logical_and(diverging, diverged_at_count == 0)
        # Update diverged_at_count for all positions which first diverged at this
        # step
        diverged_at_count[first_diverged_this_time] = limit
        # Reset any divergent values to exactly 2
        value[diverging] = 2
    return diverged_at_count



data7 = mandel4(values)

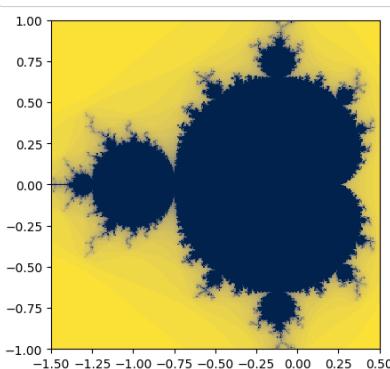


plt.imshow(data7, interpolation="none", extent=[xmin, xmax, ymin, ymax])



<matplotlib.image.AxesImage at 0x7f3d64f4b8b0>


```



```


%%timeit



data7 = mandel4(values)


```

```
31.1 ms ± 50 µs per loop (mean ± std. dev. of 7 runs, 10 loops each)
```

Note that here, all the looping over mandelbrot steps was in Python, but everything below the loop-over-positions happened in C. The code was amazingly quick compared to pure Python.

Can we do better by avoiding a square root?

```
def mandel5(position, limit=50):
    value = position
    diverged_at_count = np.zeros(position.shape)
    while limit > 0:
        limit -= 1
        value = value**2 + position
        diverging = value * np.conj(value) > 4
        first_diverged_this_time = np.logical_and(diverging, diverged_at_count == 0)
        diverged_at_count[first_diverged_this_time] = limit
        value[diverging] = 2
    return diverged_at_count

%%timeit
data8 = mandel5(values)

41.4 ms ± 67 µs per loop (mean ± std. dev. of 7 runs, 10 loops each)
```

Probably not worth the time I spent thinking about it!

NumPy Testing

Now, let's look at calculating those residuals, the differences between the different datasets.

```
data8 = mandel5(values)
data5 = mandel2(values)

np.sum((data8 - data5) ** 2)

0.0
```

For our non-numpy datasets, numpy knows to turn them into arrays:

```
xmin = -1.5
ymin = -1.0
xmax = 0.5
ymax = 1.0
resolution = 300
xstep = (xmax - xmin) / resolution
ystep = (ymax - ymin) / resolution
xs = [(xmin + (xmax - xmin) * i / resolution) for i in range(resolution)]
ys = [(ymin + (ymax - ymin) * i / resolution) for i in range(resolution)]
data1 = [[mandeli(complex(x, y)) for x in xs] for y in ys]
sum(sum((data1 - data7) ** 2))

0.0
```

But this doesn't work for pure non-numpy arrays

```
data2 = []
for y in ys:
    row = []
    for x in xs:
        row.append(mandeli(complex(x, y)))
    data2.append(row)

data2 - data1

-----  
TypeError                                 Traceback (most recent call last)  
Cell In [45], line 1  
----> 1 data2 - data1  
  
TypeError: unsupported operand type(s) for -: 'list' and 'list'
```

So we have to convert to NumPy arrays explicitly:

```
sum(sum((np.array(data2) - np.array(data1)) ** 2))

0
```

NumPy provides some convenient assertions to help us write unit tests with NumPy arrays:

```
x = [1e-5, 1e-3, 1e-1]
y = np.arccos(np.cos(x))
y

array([1.00000004e-05, 1.00000000e-03, 1.00000000e-01])

np.testing.assert_allclose(x, y, rtol=1e-6, atol=1e-20)

np.testing.assert_allclose(data7, data1)
```

Arraywise operations are fast

Note that we might worry that we carry on calculating the mandelbrot values for points that have already diverged.

```

def mandel6(position, limit=50):
    value = np.zeros(position.shape) + position
    calculating = np.ones(position.shape, dtype="bool")
    diverged_at_count = np.zeros(position.shape)
    while limit > 0:
        limit -= 1
        value[calculating] = value[calculating] ** 2 + position[calculating]
        diverging_now = np.zeros(position.shape, dtype="bool")
        diverging_now[calculating] = (
            value[calculating] * np.conj(value[calculating]) > 4
        )
        calculating = np.logical_and(calculating, np.logical_not(diverging_now))
        diverged_at_count[diverging_now] = limit
    return diverged_at_count

```

```
data8 = mandel6(values)
```

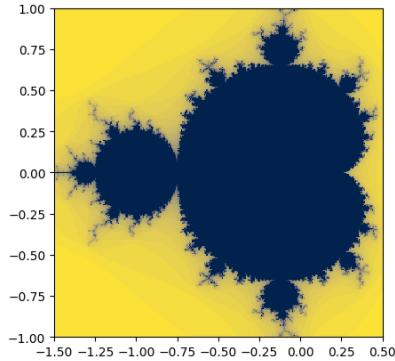
```
%%timeit
```

```
data8 = mandel6(values)
```

```
57.8 ms ± 164 µs per loop (mean ± std. dev. of 7 runs, 10 loops each)
```

```
plt.imshow(data8, interpolation="none", extent=[xmin, xmax, ymin, ymax])
```

```
<matplotlib.image.AxesImage at 0x7f3d64dc3310>
```



This was **not faster** even though it was **doing less work**

This often happens: on modern computers, **branches** (if statements, function calls) and **memory access** is usually the rate-determining step, not maths.

Complicating your logic to avoid calculations sometimes therefore slows you down. The only way to know is to **measure**

Indexing with arrays

We've been using Boolean arrays a lot to get access to some elements of an array. We can also do this with integers, as well as lists of integers:

```

x = np.arange(64)
y = x.reshape([8, 8])
y[3]

```

```
array([24, 25, 26, 27, 28, 29, 30, 31])
```

```
y[[2, 5]]
```

```
array([[16, 17, 18, 19, 20, 21, 22, 23],
       [40, 41, 42, 43, 44, 45, 46, 47]])
```

```
y[[0, 2, 5], [1, 2, 7]]
```

```
array([ 1, 18, 47])
```

We can use `a[:]` to indicate we want all the values from a particular axis:

```
y[0:4:2, [0, 2]]
```

```
array([[ 0,  2],
       [16, 18]])
```

We can mix array selectors, boolean selectors, `:s` and ordinary array sequencers:

```

z = x.reshape([4, 4, 4])
z

```

```
array([[[ 0,  1,  2,  3],
       [ 4,  5,  6,  7],
       [ 8,  9, 10, 11],
       [12, 13, 14, 15]],

      [[16, 17, 18, 19],
       [20, 21, 22, 23],
       [24, 25, 26, 27],
       [28, 29, 30, 31]],

      [[32, 33, 34, 35],
       [36, 37, 38, 39],
       [40, 41, 42, 43],
       [44, 45, 46, 47]],

      [[48, 49, 50, 51],
       [52, 53, 54, 55],
       [56, 57, 58, 59],
       [60, 61, 62, 63]]])
```

```
z[:, [1, 3], 0:3]
```

```
array([[[ 4,  5,  6],
       [12, 13, 14]],

      [[20, 21, 22],
       [28, 29, 30]],

      [[36, 37, 38],
       [44, 45, 46]],

      [[52, 53, 54],
       [60, 61, 62]]])
```

We can manipulate shapes by adding new indices in selectors with `np.newaxis`:

```
z[:, np.newaxis, [1, 3], 0].shape
```

```
(4, 1, 2)
```

When we use basic indexing with integers and `:` expressions, we get a `view` on the matrix so a copy is avoided:

```
a = z[:, :, 2]
a[0, 0] = -500
z
```

```
array([[[ 0,    1, -500,   3],
       [ 4,    5,    6,   7],
       [ 8,    9,   10,  11],
       [12,   13,   14,  15]],

      [[ 16,   17,   18,  19],
       [ 20,   21,   22,  23],
       [ 24,   25,   26,  27],
       [ 28,   29,   30,  31]],

      [[ 32,   33,   34,  35],
       [ 36,   37,   38,  39],
       [ 40,   41,   42,  43],
       [ 44,   45,   46,  47]],

      [[ 48,   49,   50,  51],
       [ 52,   53,   54,  55],
       [ 56,   57,   58,  59],
       [ 60,   61,   62,  63]]])
```

We can also use `...` to specify “`:` for as many as possible intervening axes”:

```
z[1]
```

```
array([[16, 17, 18, 19],
       [20, 21, 22, 23],
       [24, 25, 26, 27],
       [28, 29, 30, 31]])
```

```
z[..., 2]
```

```
array([[ -500,    6,   10,   14],
       [ 18,   22,   26,   30],
       [ 34,   38,   42,   46],
       [ 50,   54,   58,   62]])
```

However, boolean mask indexing and array filter indexing always causes a copy.

Let's try again at avoiding unnecessary work by using new arrays containing the reduced data instead of a mask:

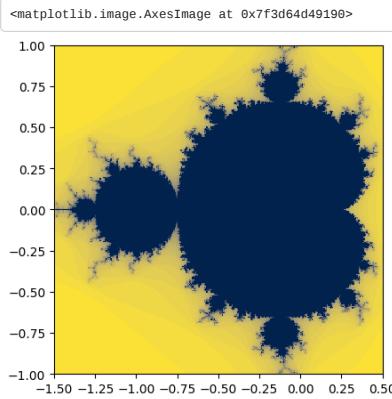
```
def mandel7(position, limit=50):
    positions = np.zeros(position.shape) + position
    value = np.zeros(position.shape) + position
    indices = np.mgrid[0 : values.shape[0], 0 : values.shape[1]]
    diverged_at_count = np.zeros(position.shape)
    while limit > 0:
        limit -= 1
        value = value**2 + positions
        diverging_now = value * np.conj(value) > 4
        diverging_now_indices = indices[:, diverging_now]
        carry_on = np.logical_not(diverging_now)

        value = value[carry_on]
        indices = indices[:, carry_on]
        positions = positions[carry_on]
        diverged_at_count[
            diverging_now_indices[0, :], diverging_now_indices[1, :]] = limit

    return diverged_at_count
```

```
data9 = mandel7(values)
```

```
plt.imshow(data9, interpolation="none", extent=[xmin, xmax, ymin, ymax])
```



```
%%timeit
data9 = mandel7(values)
```

72.2 ms ± 180 µs per loop (mean ± std. dev. of 7 runs, 10 loops each)

Still slower. Probably due to lots of copies – the point here is that you need to *experiment* to see which optimisations will work. Performance programming needs to be empirical.

Profiling

We've seen how to compare different functions by the time they take to run. However, we haven't obtained much information about where the code is spending more time. For that we need to use a profiler. IPython offers a profiler through the `%prun` magic. Let's use it to see how it works:

```
%prun mandel7(values)
```

`%prun` shows a line per each function call ordered by the total time spent on each of these. However, sometimes a line-by-line output may be more helpful. For that we can use the `line_profiler` package (you need to install it using `pip`). Once installed you can activate it in any notebook by running:

```
%load_ext line_profiler
```

And the `%lprun` magic should be now available:

```
%lprun -f mandel7 mandel7(values)
```

Here, it is clearer to see which operations are keeping the code busy.

9.3 Optimising with Cython

Estimated time for this notebook: 20 minutes

Cython can be viewed as an extension of Python where variables and functions are annotated with extra information, in particular types. The resulting Cython source code will be compiled into optimized C or C++ code, and thereby yielding substantial speed-up of slow Python code. In other words, Cython provides a way of writing Python with comparable performance to that of C/C++.

Start coding in Cython

Cython code must, unlike Python, be compiled. This happens in the following stages:

- The cython code in `.pyx` file will be translated to a `c` file.
- The c file will be compiled by a C compiler into a shared library, which will be directly loaded into Python.

In a Jupyter notebook, everything is a lot easier. One needs only to load the Cython extension (`%load_ext cython`) at the beginning and put `%%cython` mark in front of cells of Cython code. Cells with Cython mark will be treated as a `.pyx` code and consequently, compiled into C.

For details, please see [Building Cython Code](#).

Pure python Mandelbrot set:

```
xmin = -1.5
ymin = -1.0
xmax = 0.5
ymax = 1.0
resolution = 300
xstep = (xmax - xmin) / resolution
ystep = (ymax - ymin) / resolution
xs = [(xmin + (xmax - xmin) * i / resolution) for i in range(resolution)]
ys = [(ymin + (ymax - ymin) * i / resolution) for i in range(resolution)]
```



```
def mandel(position, limit=50):
    value = position
    while abs(value) < 2:
        limit -= 1
        value = value**2 + position
        if limit < 0:
            return 0
    return limit
```

Compiled by Cython:

```
%load_ext Cython
```

```
%>%cython
def mandel_cython(position, limit=50):
    value = position
    while abs(value) < 2:
        limit -= 1
        value = value ** 2 + position
        if limit < 0:
            return 0
    return limit
```

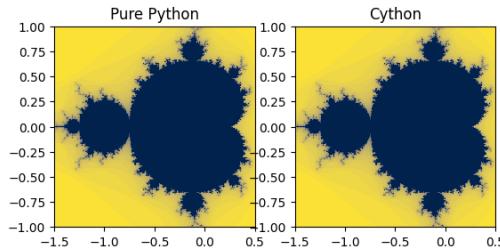
Let's verify the result

```
data_python = [[mandel(complex(x, y)) for x in xs] for y in ys]
data_cython = [[mandel_cython(complex(x, y)) for x in xs] for y in ys]
```

```
from matplotlib import pyplot as plt
plt.set_cmap("cividis") # use a CVD-friendly palette
f, axarr = plt.subplots(1, 2)
axarr[0].imshow(data_python, interpolation="none", extent=[xmin, xmax, ymin, ymax])
axarr[0].set_title("Pure Python")
axarr[1].imshow(data_cython, interpolation="none", extent=[xmin, xmax, ymin, ymax])
axarr[1].set_title("Cython")
```

Text(0.5, 1.0, 'Cython')

<Figure size 640x480 with 0 Axes>



```
%timeit [[mandel(complex(x,y)) for x in xs] for y in ys] # pure python
%timeit [[mandel_cython(complex(x,y)) for x in xs] for y in ys] # cython
```

609 ms ± 812 µs per loop (mean ± std. dev. of 7 runs, 1 loop each)

453 ms ± 1.13 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)

We have improved the performance of a factor of 1.5 by just using the Cython compiler, **without changing the code!**

Cython with C Types

But we can do better by telling Cython what C data type we would use in the code. Note we're not actually writing C, we're writing Python with C types.

typed variable

```
%>%cython
def var_typed_mandel_cython(position, limit=50):
    cdef double complex value # typed variable
    value = position
    while abs(value) < 2:
        limit -= 1
        value = value**2 + position
        if limit < 0:
            return 0
    return limit
```

typed function + typed variable

```
%>%cython
cpdef call_typed_mandel_cython(double complex position, int limit=50): # typed
    cdef double complex value # typed variable
    value = position
    while abs(value)<2:
        limit -= 1
        value = value**2 + position
        if limit < 0:
            return 0
    return limit
```

performance of one number:

```
# pure python
%timeit a = mandel(complex(0, 0))
```

13 µs ± 6.78 ns per loop (mean ± std. dev. of 7 runs, 100,000 loops each)

```
# primitive cython
%timeit a = mandel_cython(complex(0, 0))
```

9.81 µs ± 28.4 ns per loop (mean ± std. dev. of 7 runs, 100,000 loops each)

```
# cython with C type variable
%timeit a = var_typed_mandel_cython(complex(0, 0))
```

3.39 µs ± 8.68 ns per loop (mean ± std. dev. of 7 runs, 100,000 loops each)

```
# cython with typed variable + function
```

```
%timeit a = call_typed_mandel_cython(complex(0, 0))
```

Cython with numpy ndarray

You can use NumPy from Cython exactly the same as in regular Python, but by doing so you are losing potentially high speedups because Cython has support for fast access to NumPy arrays.

```
import numpy as np

ymatrix, xmatrix = np.mgrid[ymin:ymax:ystep, xmin:xmax:xstep]
values = xmatrix + 1j * ymatrix

%%cython
import numpy as np
cimport numpy as np

cpdef numpy_cython_1(np.ndarray[double complex, ndim=2] position, int limit=50):
    cdef np.ndarray[long,ndim=2] diverged_at
    cdef double complex value
    cdef int xlim
    cdef int ylim
    cdef double complex pos
    cdef int steps
    cdef int x, y

    xlim = position.shape[1]
    ylim = position.shape[0]
    diverged_at = np.zeros([ylim, xlim], dtype=int)
    for x in xrange(xlim):
        for y in xrange(ylim):
            steps = limit
            value = position[y,x]
            pos = position[y,x]
            while abs(value) < 2 and steps >= 0:
                steps -= 1
                value = value**2 + pos
            diverged_at[y,x] = steps

    return diverged_at

In file included from /opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-
packages/numpy/core/include/numpy/ndarraytypes.h:1948,
     from /opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-
packages/numpy/core/include/numpy/ndarrayobject.h:12,
     from /opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-
packages/numpy/core/include/numpy/arrayobject.h:5,
     from
/home/runner/.cache/ipython/cython/_cython_magic_cd7b650ea4b45bc671366301221786bd.
c:769:
/opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-
packages/numpy/core/include/numpy/npy_1_7_deprecated_api.h:17:2: warning: #warning
"Using deprecated NumPy API, disable it with "#define NPY_NO_DEPRECATED_API
NPY_1_7_API_VERSION" [-Wcpp]
17 | #warning "Using deprecated NumPy API, disable it with " \
| ^~~~~~
```

Note the double import of numpy: the standard numpy module and a Cython-enabled version of numpy that ensures fast indexing of and other operations on arrays. **Both import statements are necessary** in code that uses numpy arrays. The new thing in the code above is declaration of arrays by np.ndarray.

```
%timeit data_python = [mandel(value) for row in values for value in row] # pure
python
```

```
572 ms ± 3.42 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

```
%timeit data_cython = [call_typed_mandel_cython(value) for row in values for value
in row] # typed cython
```

```
40.7 ms ± 721 µs per loop (mean ± std. dev. of 7 runs, 10 loops each)
```

```
%timeit data_numpy_cython = numpy_cython_1(values) # ndarray
```

```
28.8 ms ± 102 µs per loop (mean ± std. dev. of 7 runs, 10 loops each)
```

numpy has a built-in function called `np.vectorize` to take a function that runs on a single object and return a function that runs on arrays of that object. Note that this is no faster than explicitly writing the vectorised version of the function yourself, as we can see below.

```
numpy_cython_2 = np.vectorize(call_typed_mandel_cython)
```

```
%timeit numpy_cython_2(values) # vectorize
```

```
36.7 ms ± 28.5 µs per loop (mean ± std. dev. of 7 runs, 10 loops each)
```

We got approximately a 40x total speed up from `mandel` to `numpy_cython_1`.

Calling C functions from Cython

Example: compare `sin()` from Python and C library

```
%%cython
import math
cpdef py_sin():
    cdef int x
    cdef double y
    for x in range(1e7):
        y = math.sin(x)
```

```
%%cython
from libc.math cimport sin # import from C library
cpdef c_sin():
    cdef int x
    cdef double y
    for x in range(1e7):
        y = csin(x)
```

```

%timeit [math.sin(i) for i in range(int(1e7))] # python
1.7 s ± 7.7 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)

%timeit py_sin()                                # cython call python library
932 ms ± 1.38 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)

%timeit c_sin()                                # cython call C library
5.69 ms ± 602 ns per loop (mean ± std. dev. of 7 runs, 100 loops each)

```

9.5 Optimising with Numba

Estimated time for this notebook: 15 minutes

We saw that we can use Cython to get an approximate 40x speed up when compared to pure Python. However, this comes with the cost of having to substantially rewrite the Python code to use C syntax. An alternative is to use [numba](#), an open source just-in-time compiler that translates a subset of Python and NumPy code into fast machine code.

Let's start by reproducing the pure-Python implementation from earlier.

```

xmin = -1.5
ymin = -1.0
xmax = 0.5
ymax = 1.0
resolution = 300
xstep = (xmax - xmin) / resolution
ystep = (ymax - ymin) / resolution
xs = [(xmin + xstep * i) for i in range(resolution)]
ys = [(ymin + ystep * i) for i in range(resolution)]

def mandel(position, limit=50):
    value = position
    while abs(value) < 2:
        limit -= 1
        value = value**2 + position
        if limit < 0:
            return 0
    return limit

# pure python
%timeit a = mandel(complex(0, 0))

```

12.9 µs ± 4.18 ns per loop (mean ± std. dev. of 7 runs, 100,000 loops each)

Single value numba implementation

Now let's look at a numba implementation for a single value. We add a Numba decorator to the pure-Python implementation. Note that `@njit` is equivalent to `@jit(nopython=True)`.

```

import numpy as np
from matplotlib import pyplot as plt
from numba import njit

plt.set_cmap("cividis") # use a CVD-friendly palette

<Figure size 640x480 with 0 Axes>

@njit
def mandel_numba(position, limit=50):
    value = position
    while abs(value) < 2:
        limit -= 1
        value = value**2 + position
        if limit < 0:
            return 0
    return limit

```

Note that `numba` will compile the function the first time we invoke it, so the first call will be notably slower than the rest.

```

import time
start = time.time()
mandel_numba(complex(0, 0))
print(f"Time taken for first call {time.time() - start}s")

Time taken for first call 0.2714271545410156s

# Simple numba
%timeit a = mandel_numba(complex(0, 0))

912 ns ± 0.232 ns per loop (mean ± std. dev. of 7 runs, 1,000,000 loops each)

```

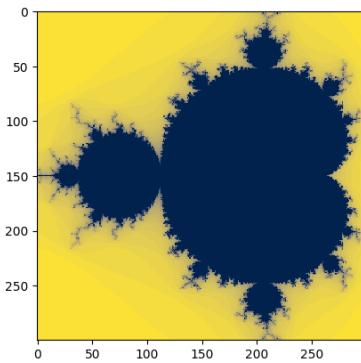
This provides an approximately 10x increase in performance compared to the pure-Python implementation.

```

data_numba = [[mandel_numba(complex(x, y)) for x in xs] for y in ys]
plt.imshow(data_numba, interpolation="none")

```

```
<matplotlib.image.AxesImage at 0x7fe451765ca0>
```



Parallelised numba implementation

Similarly to `numpy`, `numba` has optimisations related to parallelisation. Let's see whether we can improve on the performance of `numpy_cython_1`, the best Cython solution we found earlier. If we are certain that there are no dependencies between different elements in a range, we can parallelize iteration by using the `prange` function.

```
from numba import prange

@njit(parallel=True)
def mandel_numba_parallel(position, limit=50):
    xlim = position.shape[1]
    ylim = position.shape[0]
    diverged_at = np.zeros((ylim, xlim))
    for x in prange(xlim):
        for y in prange(ylim):
            steps = limit
            value = position[y, x]
            pos = position[y, x]
            while abs(value) < 2 and steps >= 0:
                steps -= 1
                value = value**2 + pos
            diverged_at[y, x] = steps
    return diverged_at

ymatrix, xmatrix = np.mgrid[ymin:ymax:ystep, xmin:xmax:xstep]
values = xmatrix + 1j * ymatrix

# Pure Python
%timeit data_python = [mandel(value) for row in values for value in row]

591 ms ± 17.7 ms per loop (mean ± std. dev. of 7 runs, 1 loop each)

# Parallelised numba
%timeit data_numba_parallel = mandel_numba_parallel(values)

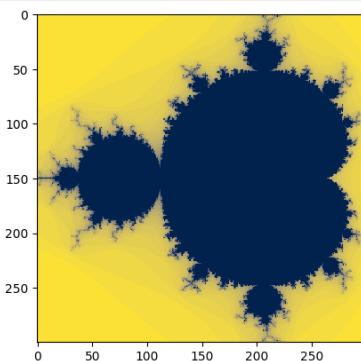
18.9 ms ± 363 µs per loop (mean ± std. dev. of 7 runs, 1 loop each)
```

This is approximately a **300x speed up** without ever needing non-Python syntax. This shows how powerful `numba` can be!

To wrap this up this section on optimisation, let's demonstrate the the parallelised numba implementation generates the same figure as the pure-Python implementation.

```
data_numba_parallel = mandel_numba_parallel(values)
plt.imshow(data_numba_parallel, interpolation="none")
```

```
<matplotlib.image.AxesImage at 0x7fe44eb7e610>
```



9.5 Performance scaling for containers and algorithms

Estimated time for this notebook: 15 minutes

We've seen that NumPy arrays are really useful. Why wouldn't we always want to use them for data which is all the same type?

```
from timeit import repeat
import numpy as np
from matplotlib import pyplot as plt
```

Let's look at appending data into a NumPy array, compared to a plain Python list:

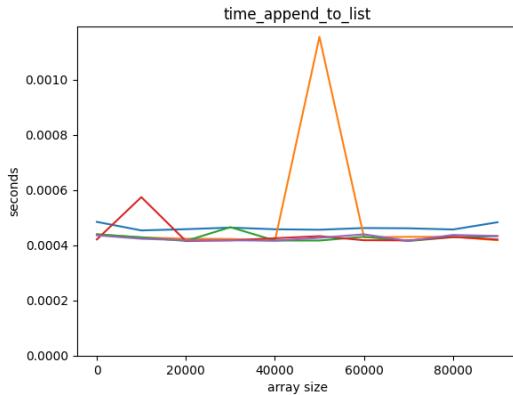
```
def time_append_to_ndarray(count):
    # the function repeat does the same that the `%timeit` magic
    # but as a function; so we can plot it.
    return repeat(
        "np.append(before, [0])",
        f"import numpy as np; before=np.ndarray({count})",
        number=10000,
    )

def time_append_to_list(count):
    return repeat("before.append(0)", f"before = [0] * {count}", number=10000)

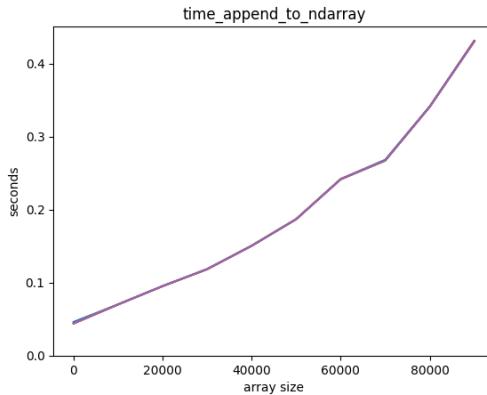
counts = np.arange(1, 100000, 10000)

def plot_time(function, counts, title=None):
    plt.plot(counts, list(map(function, counts)))
    plt.ylim(bottom=0)
    plt.ylabel("seconds")
    plt.xlabel("array size")
    plt.title(title or function.__name__)

plot_time(time_append_to_list, counts)
```



```
plot_time(time_append_to_ndarray, counts)
```



Adding an element to a Python list is way faster! Also, it seems that adding an element to a Python list is independent of the length of the list, but it's not so for a NumPy array.

How do they perform when accessing an element in the middle?

```
def time_lookup_middle_element_in_list(count):
    test_list = [0] * count
    middle_position = count // 2

    def totime():
        return test_list[middle_position]

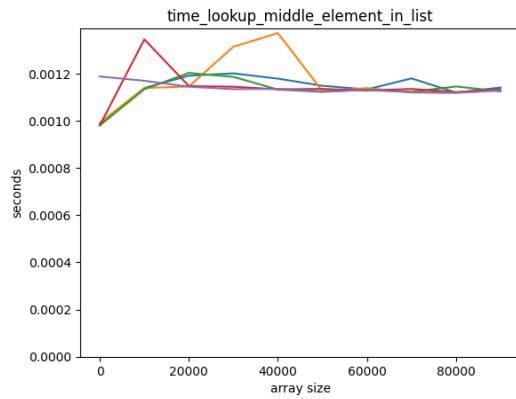
    return repeat(totime, number=10000)

def time_lookup_middle_element_in_ndarray(count):
    test_array = np.ndarray(count)
    middle_position = count // 2

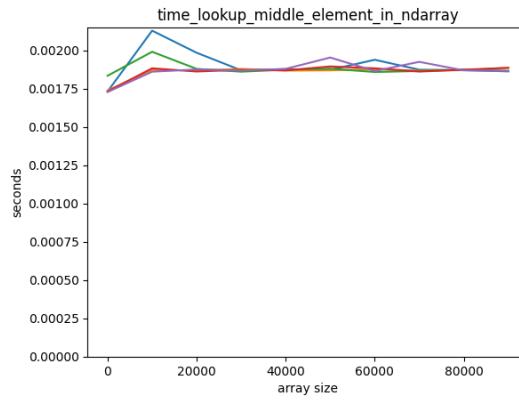
    def totime():
        return test_array[middle_position]

    return repeat(totime, number=10000)

plot_time(time_lookup_middle_element_in_list, counts)
```



```
{ plot_time(time_lookup_middle_element_in_list, counts)
```



```
{ plot_time(time_lookup_middle_element_in_ndarray, counts)
```

Both scale well for accessing the middle element.

What about inserting at the beginning?

If we want to insert an element at the beginning of a Python list we can do:

```
x = list(range(5))
x
```

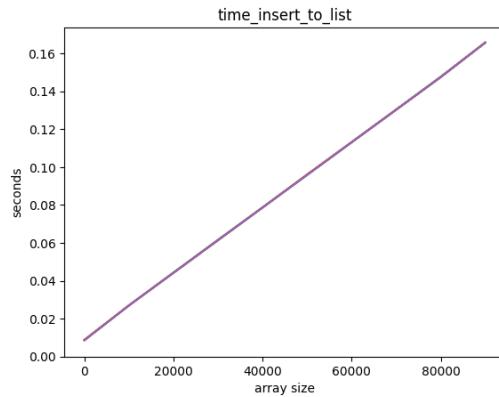
```
[0, 1, 2, 3, 4]
```

```
x[0:0] = [-1]
x
```

```
[-1, 0, 1, 2, 3, 4]
```

```
def time_insert_to_list(count):
    return repeat("before[0:0] = [0]", f"before = [0] * {count}", number=10000)
```

```
{ plot_time(time_insert_to_list, counts)
```



`list` performs **badly** for insertions at the beginning!

There are containers in Python that work well for insertion at the start:

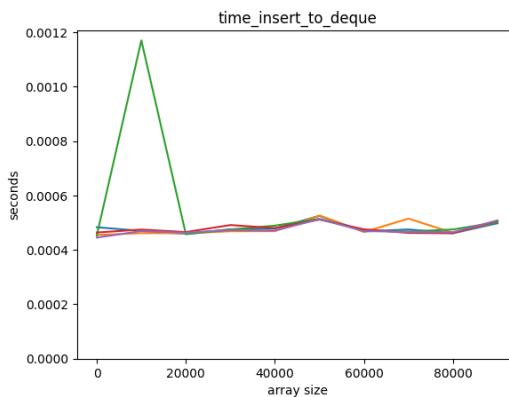
```
from collections import deque
```

```

def time_insert_to_deque(count):
    return repeat(
        "before.appendleft(0)",
        f"from collections import deque; before = deque([0] * {count})",
        number=10000,
    )

```

```
plot_time(time_insert_to_deque, counts)
```



But looking up in the middle scales badly:

```

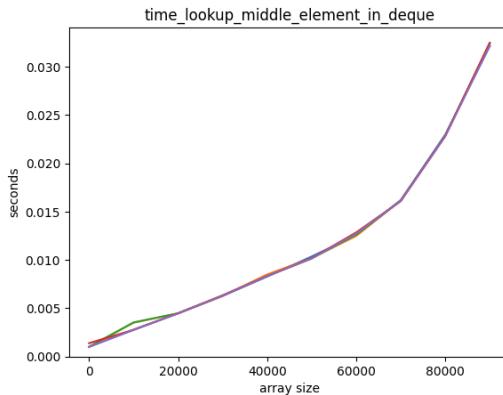
def time_lookup_middle_element_in_deque(count):
    test_deque = deque([0] * count)
    middle_position = count // 2

    def totime():
        return test_deque[middle_position]

    return repeat(totime, number=10000)

```

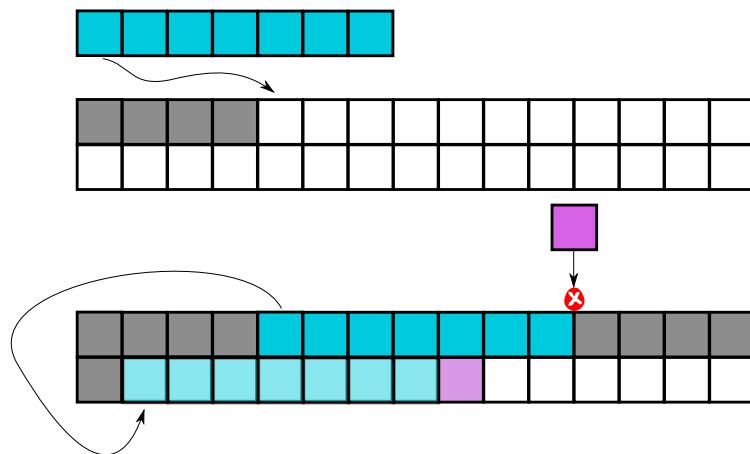
```
plot_time(time_lookup_middle_element_in_deque, counts)
```



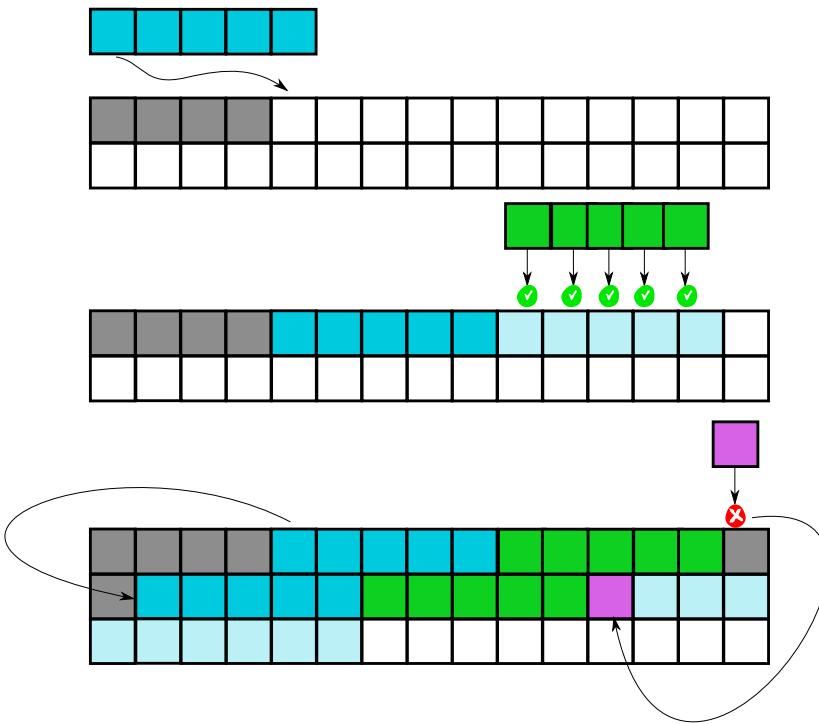
What is going on here?

Arrays are stored as contiguous memory. Anything which changes the length of the array requires the whole array to be copied elsewhere in memory.

This copy takes time proportional to the array size.



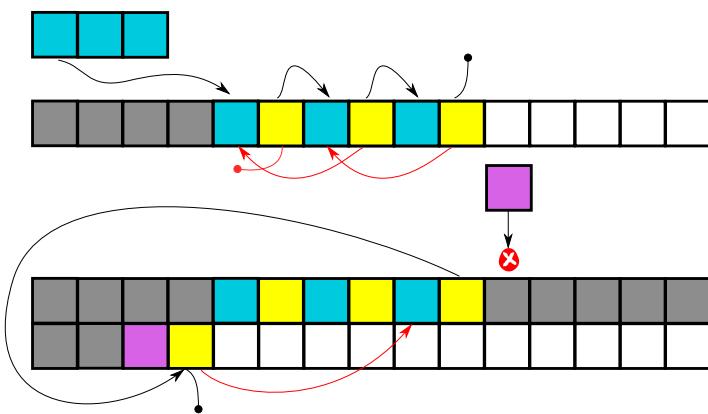
The Python `list` type is **also** an array, but it is allocated with **extra memory**. Only when that memory is exhausted is a copy needed.



If the extra memory is typically the size of the current array, a copy is needed every $1/N$ appends, and costs N to make, so **on average** copies are cheap. We call this **amortized constant time**.

This makes it fast to look up values in the middle. However, it may also use more space than is needed.

The deque type works differently: each element contains a pointer to the next. Inserting elements is therefore very cheap, but looking up the N th element requires traversing N such pointers.



Dictionary performance

For another example, let's consider the performance of a dictionary versus a couple of other ways in which we could implement an associative array.

```
class evildict:
    def __init__(self, data):
        self.data = data

    def __getitem__(self, akey):
        for key, value in self.data:
            if key == akey:
                return value
        raise KeyError()
```

If we have an evil dictionary of N elements, how long would it take - on average - to find an element?

```
{ eric = [["Name", "Eric Idle"], ["Job", "Comedian"], ["Home", "London"]]
{ eric_evil = evildict(eric)
{ eric_evil["Job"]
'Comedian'
{ eric_dict = dict(eric)
{ eric_evil["Job"]
'Comedian'
```

```
x = ["Hello", "License", "Fish", "Eric", "Pet", "Halibut"]
```

```
sorted(x, key=lambda el: el.lower())
```

```
['Eric', 'Fish', 'Halibut', 'Hello', 'License', 'Pet']
```

What if we created a dictionary where we bisect the search?

```
from bisect import bisect_left

class sorteddict:
    def __init__(self, data):
        self.data = sorted(data, key=lambda x: x[0])
        self.keys = list(map(lambda x: x[0], self.data))

    def __getitem__(self, akey):
        loc = bisect_left(self.keys, akey)

        if loc == len(self.data):
            return self.data[loc][1]

        raise KeyError()
```

```
eric_sorted = sorteddict(eric)
```

```
eric_sorted["Job"]
```

```
'Comedian'
```

```
def time_dict_generic(ttype, count):
    keys = list(range(count))
    values = [0] * count
    data = ttype(zip(keys, values))

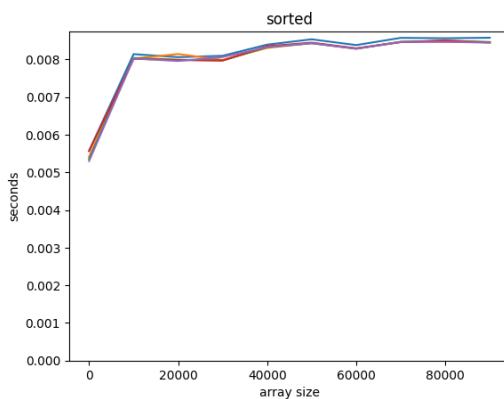
    def totime():
        return data[keys[count // 2]]
    return repeat(totime, number=10000)
```

```
def time_dict(count):
    return time_dict_generic(dict, count)
```

```
def time_sorted(count):
    return time_dict_generic(sorteddict, count)
```

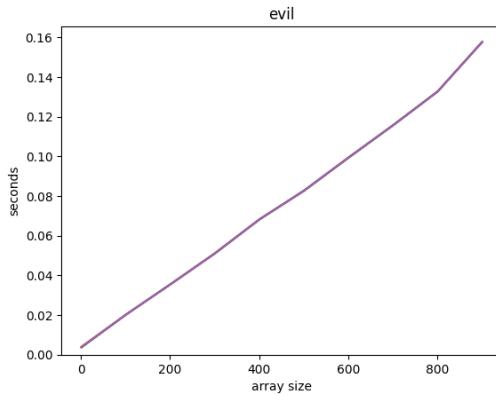
```
def time_evil(count):
    return time_dict_generic(evildict, count)
```

```
plot_time(time_sorted, counts, title="sorted")
```



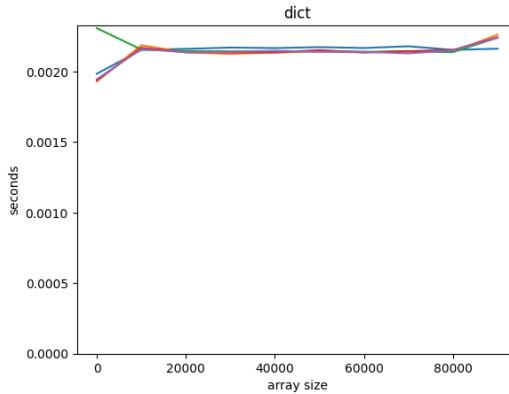
We can't really see what's going on here for the sorted example as there's too much noise, but theoretically we should get **logarithmic** asymptotic performance. We write this down as $O(\ln N)$. This doesn't mean there isn't also a constant term, or a term proportional to something that grows slower (such as $\ln(\ln N)$): we always write down just the term that is dominant for large N . We saw before that `list` is $O(1)$ for appends, $O(N)$ for inserts. Numpy's `array` is $O(N)$ for appends.

```
counts = np.arange(1, 1000, 100)
plot_time(time_evil, counts, title="evil")
```



The simple check-each-in-turn solution is $O(N)$ - linear time.

```
counts = np.arange(1, 100000, 10000)
plot_time(time_dict, counts, title="dict")
```



Python's built-in dictionary is, amazingly, $O(1)$: the time is **independent** of the size of the dictionary.

This uses a miracle of programming called the *Hash Table*: you can learn more about [these issues at this video from Harvard University](#). This material is pretty advanced, but, I think, really interesting!

Optional exercise: determine what the asymptotic performance for the Boids model in terms of the number of Boids. Make graphs to support this. Bonus: how would the performance scale with the number of dimensions?

10. Scientific file formats

- Serialisation and Deserialisation
- Domain specific languages
- Templating languages: Mako
- Binary file formats: XDR and HDF5
- Parsers and grammars: Python Lex and Yacc
- Ontologies
- Semantic file formats

Contents

- [10.0 Serialising and normalising data](#) (20 minutes)
- [10.1 Deserialisation](#) (10 minutes)
- [10.2 Binary formats](#) (10 minutes)
- [10.3 Markup languages](#) (15 minutes)
- [10.4 Beyond Pandas](#) (20 minutes)
- [10.5 Processing in parallel](#) (20 minutes)
- [10.6 Geospatial data](#) (15 minutes)

Additional content

- [OPTIONAL Domain specific languages](#) (25 minutes)
- [OPTIONAL Controlled Vocabularies](#) (15 minutes)
- [OPTIONAL Semantic file formats](#) (25 minutes)

Total time: 2 hrs

Exercises

This module does not currently have any associated exercises.

10.0 Serialising and normalising data

Estimated time for this notebook: 20 minutes

Consider a simple python computational model of chemical reaction networks:

```

class Element:
    def __init__(self, symbol, number):
        self.symbol = symbol
        self.number = number

    def __str__(self):
        return str(self.symbol)

class Molecule:
    def __init__(self, mass):
        self.elements = {} # Map from element to number of that element in the molecule
        self.mass = mass

    def add_element(self, element, number):
        self.elements[element] = number

    @staticmethod
    def as_subscript(number):
        if number == 1:
            return ""
        if number < 10:
            return "_" + str(number)
        return "(" + str(number) + ")"

    def __str__(self):
        return "".join([
            str(element) + Molecule.as_subscript(number)
            for element, number in self.elements.items()
        ])

class Reaction:
    def __init__(self):
        self.reactants = {} # Map from reactants to stoichiometries
        self.products = {} # Map from products to stoichiometries

    def add_reactant(self, reactant, stoichiometry):
        self.reactants[reactant] = stoichiometry

    def add_product(self, product, stoichiometry):
        self.products[product] = stoichiometry

    @staticmethod
    def print_if_not_one(number):
        if number == 1:
            return ""
        return str(number)

    @staticmethod
    def side_as_string(side):
        return " " + ".join([
            Reaction.print_if_not_one(side[molecule]) + str(molecule)
            for molecule in side
        ])

    def __str__(self):
        return (
            Reaction.side_as_string(self.reactants)
            + "\rightarrow"
            + Reaction.side_as_string(self.products)
        )

class System:
    def __init__(self):
        self.reactions = []

    def add_reaction(self, reaction):
        self.reactions.append(reaction)

    def __str__(self):
        return "\n".join(self.reactions)

```

```

c = Element("C", 12)
o = Element("O", 8)
h = Element("H", 1)

co2 = Molecule(44.01)
co2.add_element(c, 1)
co2.add_element(o, 2)

h2o = Molecule(18.01)
h2o.add_element(h, 2)
h2o.add_element(o, 1)

o2 = Molecule(32.00)
o2.add_element(o, 2)

glucose = Molecule(180.16)
glucose.add_element(c, 6)
glucose.add_element(h, 12)
glucose.add_element(o, 6)

combustion = Reaction()
combustion.add_reactant(glucose, 1)
combustion.add_reactant(o2, 6)
combustion.add_product(co2, 6)
combustion.add_product(h2o, 6)

print(combustion)

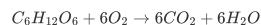
```

```
C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O
```

```

from IPython.display import Math, display
display(Math(str(combustion)))

```



We could reasonably consider using the LaTeX representation of this as a fileformat for reactions. (Though we need to represent the molecular mass in some way we've not thought of.)

We've already shown how to **serialise** the data to this representation. How hard would it be to **deserialise** it?

Actually, pretty darn hard.

In the wild, such datafiles will have all kinds of complications, and making a hand-coded string parser for such text will be highly problematic. In this lecture, we're going to look at the kind of problems that can arise, and some standard ways to solve them, which will lead us to the idea of **normalisation** in databases.

Next lecture, we'll look at how we might create a file format which does indeed look like such a fluent mathematical representation, which we'll call a **Domain Specific Language**.

Non-normal data representations: First normal form.

Consider the mistakes that someone might make when typing in a reaction in the above format: they could easily, if there are multiple reactions in a system, type glucose in correctly as `C_6H_{12}O_6` the first time, but the second type accidentally type `C_6H_{12}O_6`.

The system wouldn't know these are the same molecule, so, for example, if building a mass action model of reaction kinetics, the differential equations would come out wrong.

The natural-seeming solution to this is, in your data format, to name each molecule and atom, and consider a representation in terms of CSV files:

```
%>%writefile molecules.csv  
# name, elements, numbers  
  
water, H O, 2 1  
oxygen, O, 2  
carbon_dioxide, C O, 1 2  
glucose, C H O, 6 12 6
```

Writing molecules.csv

```
%>%writefile reactions.csv  
# name, reactants, products, reactant_stoichiometries, product_stoichiometries  
  
combustion_of_glucose, glucose oxygen, carbon_dioxide water, 1 6, 6 6
```

Writing reactions.csv

Writing a parser for these files would be very similar to the earthquake problem that you've already encountered.

However, the existence of multiple values in one column indicates that this file format is NOT **first normal form** (1NF).

Note: A table is in first normal form if: every column is unique; no rows are duplicated; each column/row intersection contains only one value.

It is not uncommon to encounter file-formats that violate 1NF in the wild. The main problem with them is that you will eventually have to deal with the separation character that you picked (`,` in this case) turning up in someone's content and you'll need to work out how to escape it.

The art of designing serialisations which work as row-and-column value tables for more complex data structures is the core of database design.

Normalising the reaction model - a bad first attempt.

How could we go about normalising this model. One choice is to list each molecule-element **relation** as a separate table row:

```
%>%writefile molecules.csv  
# name, element, number  
  
water, H, 2  
water, O, 1  
oxygen, O, 2  
carbon_dioxide, C, 1  
carbon_dioxide, O, 2
```

Overwriting molecules.csv

This is fine as far as it goes, but, it falls down as soon as we want to associate another property with a molecule and atom.

We could repeat the data each time:

```
%>%writefile molecules.csv  
# name, element, number, molecular_mass, atomic_number  
  
water, H, 2, 18.01, 1  
water, O, 1, 18.01, 8  
oxygen, O, 2, 32.00, 8
```

Overwriting molecules.csv

The existence of the same piece of information in multiple locations (eg. the `18.01` molecular mass of water) indicates that this file format is NOT **second normal form** (2NF).

Furthermore, this would allow our data file to be potentially be self-inconsistent, violating the design principle that each piece of information should be stated only once. A data structure of this type is said to be NOT **second normal form**.

Note: A table is in second normal form if: it is in first normal form; none of its attributes depend on any other attribute except the primary key.

Normalising the model - relations and keys

So, how do we do this correctly?

We need to specify data about each molecule, reaction and atom once, and then specify the **relations** between them.

```
%>%writefile molecules.csv  
# name, molecular_mass  
  
water, 18.01  
oxygen, 32.00
```

Overwriting molecules.csv

```
%>%writefile atoms.csv  
# symbol, atomic number  
H, 1  
O, 8  
C, 6
```

```
Writing atoms.csv
```

```
%>%writefile atoms_in_molecules.csv  
# rel_number, molecule, symbol, number  
0, water, H, 2  
1, water, O, 1  
2, oxygen, O, 2
```

```
Writing atoms_in_molecules.csv
```

This last table is called a **join table** - and is needed whenever we want to specify a "many-to-many" relationship. (Each atom can be in more than one molecule, and each molecule has more than one atom.)

Note each table needs to have a set of columns which taken together form a unique identifier for that row; called a "key". If more than one is possible, we choose one and call it a **primary key**. (And in practice, we normally choose a single column for this: hence the 'rel_number' column, though the tuple {molecule, symbol} here is another **candidate key**.)

Now, proper database tools use much more sophisticated representations than just csv files - including **indices** to enable hash-table like efficient lookups, and support for managing multiple users at the same time.

Furthermore, the **principles** of database normalisation and the relational model will be helpful right across our thinking about data representation, whether these are dataframes in Pandas, tensors in tensorflow, or anything else...

Making a database - SQLite

Let's look at how we would use a simple database in Python to represent atoms and molecules. If you've never seen SQL before, you might want to attend an introductory course, such as one of the 'Software Carpentry' sessions. Here we're going to assume some existing knowledge but we will use a Python-style way to interact with databases instead of relying on raw SQL.

```
import os  
try:  
    os.remove("molecules.db")  
    print("Removing database to start again from scratch")  
except FileNotFoundError:  
    print("No DB since this notebook was last run")
```

```
No DB since this notebook was last run  
  
import sqlalchemy  
engine = sqlalchemy.create_engine("sqlite:///molecules.db", echo=True)
```

SQLite is a simple very-lightweight database tool - without support for concurrent users - but it's great for little hacks like this. For full-on database work you'll probably want to use a more fully-featured database like <https://www.postgresql.org>.

The metadata for the database describing the tables present, and their columns, is defined in Python using SQLAlchemy, the leading python database tool, thus:

```
from sqlalchemy import Column, Float, MetaData, String, Table  
  
metadata = MetaData()  
molecules = Table(  
    "molecules",  
    metadata,  
    Column("name", String, primary_key=True),  
    Column("mass", Float),  
)  
  
atoms = Table(  
    "atoms",  
    metadata,  
    Column("symbol", String, primary_key=True),  
    Column("number", Integer),  
)
```

```
NameError  
Cell In [13], line 15  
    3 metadata = MetaData()  
    4 molecules = Table(  
    5     "molecules",  
    6     metadata,  
    7     Column("name", String, primary_key=True),  
    8     Column("mass", Float),  
    9 )  
   10 atoms = Table(  
   11     "atoms",  
   12     metadata,  
   13     Column("symbol", String, primary_key=True),  
--> 14     Column("number", Integer),  
   15     )  
   16 )  
  
NameError: name 'Integer' is not defined
```

```
from sqlalchemy import ForeignKey, Integer  
  
atoms_in_molecules = Table(  
    "atoms_molecules",  
    metadata,  
    Column("atom", ForeignKey("atoms.symbol")),  
    Column("molecule", ForeignKey("molecules.name")),  
    Column("number", Integer),  
)
```

```
metadata.create_all(engine)  
print(metadata)
```

```

2021-08-04 13:31:53,552 INFO sqlalchemy.engine.Engine BEGIN (implicit)
2021-08-04 13:31:53,553 INFO sqlalchemy.engine.Engine PRAGMA
main.table_info("molecules")
2021-08-04 13:31:53,555 INFO sqlalchemy.engine.Engine [raw sql] ()
2021-08-04 13:31:53,558 INFO sqlalchemy.engine.Engine PRAGMA
temp.table_info("molecules")
2021-08-04 13:31:53,560 INFO sqlalchemy.engine.Engine [raw sql] ()
2021-08-04 13:31:53,563 INFO sqlalchemy.engine.Engine PRAGMA
main.table_info("atoms")
2021-08-04 13:31:53,566 INFO sqlalchemy.engine.Engine [raw sql] ()
2021-08-04 13:31:53,568 INFO sqlalchemy.engine.Engine PRAGMA
temp.table_info("atoms")
2021-08-04 13:31:53,570 INFO sqlalchemy.engine.Engine [raw sql] ()
2021-08-04 13:31:53,572 INFO sqlalchemy.engine.Engine PRAGMA
main.table_info("atoms_molecules")
2021-08-04 13:31:53,573 INFO sqlalchemy.engine.Engine [raw sql] ()
2021-08-04 13:31:53,574 INFO sqlalchemy.engine.Engine PRAGMA
temp.table_info("atoms_molecules")
2021-08-04 13:31:53,575 INFO sqlalchemy.engine.Engine [raw sql] ()
2021-08-04 13:31:53,577 INFO sqlalchemy.engine.Engine
CREATE TABLE molecules (
    name VARCHAR NOT NULL,
    mass FLOAT,
    PRIMARY KEY (name)
)

2021-08-04 13:31:53,578 INFO sqlalchemy.engine.Engine [no key 0.00094s] ()
2021-08-04 13:31:53,588 INFO sqlalchemy.engine.Engine
CREATE TABLE atoms (
    symbol VARCHAR NOT NULL,
    number INTEGER,
    PRIMARY KEY (symbol)
)

2021-08-04 13:31:53,589 INFO sqlalchemy.engine.Engine [no key 0.00143s] ()
2021-08-04 13:31:53,596 INFO sqlalchemy.engine.Engine
CREATE TABLE atoms_molecules (
    atom VARCHAR,
    molecule VARCHAR,
    number INTEGER,
    FOREIGN KEY(atom) REFERENCES atoms (symbol),
    FOREIGN KEY(molecule) REFERENCES molecules (name)
)

2021-08-04 13:31:53,601 INFO sqlalchemy.engine.Engine [no key 0.00410s] ()
2021-08-04 13:31:53,613 INFO sqlalchemy.engine.Engine COMMIT
MetaData()

```

Note the SQL syntax for creating tables is generated by the python tool, and sent to the database server.

```

CREATE TABLE molecules (
    name VARCHAR NOT NULL,
    mass FLOAT,
    PRIMARY KEY (name)
)

```

We'll turn off our automatic printing of all the raw sql to avoid this notebook being unreadable.

```

engine.echo = False

```

We can also write data to our database using this python tooling:

```

ins = molecules.insert().values(name="water", mass="18.01")

conn = engine.connect()
conn.execute(ins)

<sqlalchemy.engine.cursor.LegacyCursorResult at 0x113546610>

```

And query it:

```

from sqlalchemy.sql import select
s = select([molecules])
result = conn.execute(s)
print(result.fetchone()["mass"])

```

```

18.01

```

If we have enough understanding of SQL syntax, we can use appropriate **join** statements to find, for example, the mass of all molecules which contain oxygen:

```

conn.execute(molecules.insert().values(name="oxygen", mass="32.00"))
conn.execute(atoms.insert().values(symbol="O", number=8))
conn.execute(atoms.insert().values(symbol="H", number=1))
conn.execute(atoms_in_molecules.insert().values(molecule="water", atom="O",
number=1))
conn.execute(atoms_in_molecules.insert().values(molecule="oxygen", atom="O",
number=1))
conn.execute(atoms_in_molecules.insert().values(molecule="water", atom="H",
number=2))

```

```

<sqlalchemy.engine.cursor.LegacyCursorResult at 0x113563410>

```

```

result = conn.execute(
    """
    SELECT mass
    FROM   molecules
    JOIN atoms_molecules
        ON molecules.NAME = atoms_molecules.molecule
    JOIN atoms
        ON atoms.symbol = atoms_molecules.atom
    WHERE  atoms.symbol = 'H'
    """
)
print(result.fetchall())

```

```

[(18.01,)]

```

But we can do much better...

Data and Objects - the Object-Relational-Mapping

We notice that when we find a correct relational model for our data, many of the rows are suggestive of exactly the data we would expect to supply to an object constructor - data about an object. References to keys of other tables in rows suggest composition relations while many-to-many join tables often represent aggregation relationships, and data about the relationship.

As a result of this, powerful tools exist to **automatically** create object structures from database schema, including saving and loading.

```
import os
try:
    os.remove("molecules.db")
    print("removing database to start again from scratch")
except FileNotFoundError:
    print("No DB since this notebook was last run")

Removing database to start again from scratch

import sqlalchemy
engine = sqlalchemy.create_engine("sqlite:///molecules.db")

from sqlalchemy import Column, Integer, String
from sqlalchemy.ext.declarative import declarative_base
from sqlalchemy.orm import relationship

Base = declarative_base()

class Element(Base):
    __tablename__ = "atoms"
    symbol = Column(String, primary_key=True)
    number = Column(Integer)
    molecules = relationship("AtomsPerMolecule", backref="atom")

class Molecule(Base):
    __tablename__ = "molecules"
    name = Column(String, primary_key=True)
    mass = Column(Float)
    atoms = relationship("AtomsPerMolecule", backref="molecule")

class AtomsPerMolecule(Base):
    __tablename__ = "atoms_per_molecule"
    id = Column(Integer, primary_key=True)
    atom_id = Column(None, ForeignKey("atoms.symbol"))
    molecule_id = Column(None, ForeignKey("molecules.name"))
    number = Column(Integer)
```

If we now create our tables, the system will automatically create a DB:

```
Base.metadata.create_all(engine)

engine.echo = False
```

And we can create objects with a simple interface that looks just like ordinary classes:

```
oxygen = Element(symbol="O", number=8)
hydrogen = Element(symbol="H", number=1)
elements = [oxygen, hydrogen]

water = Molecule(name="water", mass=18.01)
oxygen_m = Molecule(name="oxygen", mass=16.00)
hydrogen_m = Molecule(name="hydrogen", mass=2.02)
molecules = [water, oxygen_m, hydrogen_m]

# Note that we are using the `backref` name to construct the `atom_id` and
# `molecule_id`.
# These lookup instances of Element and Molecule that are already in our database
amounts = [
    AtomsPerMolecule(atom=oxygen, molecule=water, number=1),
    AtomsPerMolecule(atom=hydrogen, molecule=water, number=2),
    AtomsPerMolecule(atom=oxygen, molecule=oxygen_m, number=2),
    AtomsPerMolecule(atom=hydrogen, molecule=hydrogen_m, number=2),
]

from sqlalchemy.orm import sessionmaker
Session = sessionmaker(bind=engine)
session = Session()

session.bulk_save_objects(elements + molecules + amounts)

oxygen.molecules[0].molecule.name

'water'

session.query(Molecule).all()[0].name

'water'

session.commit()
```

This is a very powerful technique - we get our class-type interface in python, with database persistence and searchability for free!

Moving on from databases

Databases are often a good choice for storing data, but can only be interacted with programmatically. Often, we want to make a file format to represent our dataset which can be easily replicated or shared. The next part of this module focuses on the design of such file-formats, both binary and **human-readable**.

One choice, now we know about it, is to serialise all the database tables as CSV:

```
import pandas
str(session.query(Molecule).statement)
'SELECT molecules.name, molecules.mass \nFROM molecules'
dataframe = pandas.read_sql(session.query(Molecule).statement, session.bind)
dataframe
```

	name	mass
0	water	18.01
1	oxygen	16.00
2	hydrogen	2.02

```
print(dataframe.to_csv())
,name,mass
0,water,18.01
1,oxygen,16.0
2,hydrogen,2.02
```

Deserialising is also easy:

```
%>> writefile atoms.csv
symbol,number
C,6
N,7
```

Overwriting atoms.csv

```
with open("atoms.csv", "r") as f_csv:
    atoms = pandas.read_csv(f_csv)
atoms
```

	symbol	number
0	C	6
1	N	7

```
atoms.to_sql("atoms", session.bind, if_exists="append", index=False)
```

```
session.query(Element).all()[3].number
```

7

However, we know from last time that another common choice is to represent such complicated data structures in YAML. The implications of what we've just learned for serialising to and from such structured data is the topic of the next lecture.

10.1 Deserialisation

Estimated time for this notebook: 10 minutes

YAML (a recursive acronym for “YAML Ain’t Markup Language”) is a human-readable data-serialization language.

We’re going to slightly modify our previous model and look at how to serialise it to YAML.

```

class Element:
    def __init__(self, symbol):
        self.symbol = symbol

class Molecule:
    def __init__(self):
        self.elements = {} # Map from element to number of that element in the molecule

    def add_element(self, element, number):
        self.elements[element] = number

    def to_struct(self):
        return {element.symbol: number for element, number in self.elements.items()}

class Reaction:
    def __init__(self):
        self.reactants = {} # Map from reactants to stoichiometries
        self.products = {} # Map from products to stoichiometries

    def add_reactant(self, reactant, stoichiometry):
        self.reactants[reactant] = stoichiometry

    def add_product(self, product, stoichiometry):
        self.products[product] = stoichiometry

    def to_struct(self):
        return {
            "reactants": [x.to_struct() for x in self.reactants],
            "products": [x.to_struct() for x in self.products],
            "stoichiometries": list(self.reactants.values()) + list(self.products.values()),
        }

class System:
    def __init__(self):
        self.reactions = []

    def add_reaction(self, reaction):
        self.reactions.append(reaction)

    def to_struct(self):
        return [x.to_struct() for x in self.reactions]

```

```

c = Element("C")
o = Element("O")
h = Element("H")

co2 = Molecule()
co2.add_element(c, 1)
co2.add_element(o, 2)

h2o = Molecule()
h2o.add_element(h, 2)
h2o.add_element(o, 1)

o2 = Molecule()
o2.add_element(o, 2)

h2 = Molecule()
h2.add_element(h, 2)

glucose = Molecule()
glucose.add_element(c, 6)
glucose.add_element(h, 12)
glucose.add_element(o, 6)

combustion_glucose = Reaction()
combustion_glucose.add_reactant(glucose, 1)
combustion_glucose.add_reactant(o2, 6)
combustion_glucose.add_product(co2, 6)
combustion_glucose.add_product(h2o, 6)

combustion_hydrogen = Reaction()
combustion_hydrogen.add_reactant(h2, 2)
combustion_hydrogen.add_reactant(o2, 1)
combustion_hydrogen.add_product(h2o, 2)

s = System()
s.add_reaction(combustion_glucose)
s.add_reaction(combustion_hydrogen)

s.to_struct()

```

```

[{"reactants": [{"C": 6, "H": 12, "O": 6}, {"O": 2}], "products": [{"C": 1, "O": 2}, {"H": 2, "O": 1}], "stoichiometries": [1, 6, 6, 6], "reactants": [{"H": 2}, {"O": 2}], "products": [{"H": 2, "O": 1}], "stoichiometries": [2, 1, 2]}

```

```

import yaml
print(yaml.dump(s.to_struct()))

```

```

- products:
  - C: 1
  - O: 2
  - H: 2
  - O: 1
  reactants:
  - C: 6
  - H: 12
  - O: 6
  - O: 2
  stoichiometries:
  - 1
  - 6
  - 6
  - 6
- products:
  - H: 2
  - O: 1
  reactants:
  - H: 2
  - O: 2
  stoichiometries:
  - 2
  - 1
  - 2

```

Deserialising non-normal data structures

We can see that this data structure, although seemingly sensible, is horribly **non-normal**.

- The stoichiometries information requires us to align each one to the corresponding molecule in order.
- Each element is described multiple times: we will have to ensure that each mention of `C` comes back to the same constructed element object.

```
class YamlDeSerialisingSystem:  
    def __init__(self):  
        self.elements = {}  
        self.molecules = {}  
  
    def add_element(self, candidate):  
        if candidate not in self.elements:  
            self.elements[candidate] = Element(candidate)  
        return self.elements[candidate]  
  
    def add_molecule(self, candidate):  
        if tuple(candidate.items()) not in self.molecules:  
            m = Molecule()  
            for symbol, number in candidate.items():  
                m.add_element(self.add_element(symbol), number)  
            self.molecules[tuple(candidate.items())] = m  
        return self.molecules[tuple(candidate.items())]  
  
    def parse_system(self, system_dict):  
        system = System()  
        for reaction in system_dict:  
            r = Reaction()  
            stoichiometries = reaction["stoichiometries"]  
            for molecule in reaction["reactants"]:  
                r.add_reactant(self.add_molecule(molecule),  
                stoichiometries.pop(0))  
            for molecule in reaction["products"]:  
                r.add_product(self.add_molecule(molecule), stoichiometries.pop(0))  
            system.add_reaction(r)  
        return system  
  
de_serialiser = YamlDeSerialisingSystem()  
round_trip = de_serialiser.parse_system(s.to_struct())  
  
round_trip.to_struct()  
  
[{'reactants': [{'C': 6, 'H': 12, 'O': 6}, {'O': 2}],  
 'products': [{'C': 1, 'O': 2}, {'H': 2, 'O': 1}],  
 'stoichiometries': [1, 6, 6, 6]},  
 {'reactants': [{'H': 2}, {'O': 2}],  
 'products': [{'H': 2, 'O': 1}],  
 'stoichiometries': [2, 1, 2]}]  
  
de_serialiser.elements  
  
{'C': <__main__.Element at 0x7fa86ced4cd0>,  
 'H': <__main__.Element at 0x7fa86ced4a60>,  
 'O': <__main__.Element at 0x7fa86ced4bb0>}  
  
de_serialiser.molecules  
  
{('C', 6), ('H', 12), ('O', 6)): <__main__.Molecule at 0x7fa86ced4940>,  
 (('O', 2),): <__main__.Molecule at 0x7fa86ced44f0>,  
 ('C', 1), ('O', 2)): <__main__.Molecule at 0x7fa86ced4ca0>,  
 ('H', 2), ('O', 1)): <__main__.Molecule at 0x7fa86cf8acd0>,  
 ('H', 2),): <__main__.Molecule at 0x7fa86cf886d0>}  
  
list(round_trip.reactions[0].reactants.keys())[1].to_struct()  
  
{'O': 2}  
  
list(round_trip.reactions[1].reactants.keys())[1].to_struct()  
  
{'O': 2}
```

In order to de-serialise this data, we had to construct a unique key to distinguish repeated mentions of the same identical item.

Effectively, we ended up choosing primary keys for our datatypes:

```
list(de_serialiser.molecules.keys())  
  
[((('C', 6), ('H', 12), ('O', 6)),  
  (('O', 2),),  
  ('C', 1), ('O', 2)),  
 ((('H', 2), ('O', 1)),  
  ('H', 2),))]
```

Remembering that a combination of columns uniquely defining an item is a valid key - there is a key correspondence between a candidate key in the database sense and a "hashable" data structure that can be used to a key in a `dict`.

Note that to make this example even reasonably doable, we had to exclude additional data from the objects (mass, rate etc)

Normalising a YAML structure

To make this structure easier to de-serialise, we can make a normalised file-format, by defining primary keys (hashable types) for each entity on write:

```

class YamlSavingSystem:
    def __init__(self):
        self.elements = set()
        self.molecules = set()

    def element_key(self, element):
        return element.symbol

    def molecule_key(self, molecule):
        key = ""
        for element, number in molecule.elements.items():
            key += element.symbol
            key += str(number)
        return key

    def save(self, system):
        for reaction in system.reactions:
            for molecule in reaction.reactants:
                self.molecules.add(molecule)
                for element in molecule.elements:
                    self.elements.add(element)
            for molecule in reaction.products:
                self.molecules.add(molecule)
                for element in molecule.elements:
                    self.elements.add(element)

        result = {
            "elements": [self.element_key(element) for element in self.elements],
            "molecules": [
                self.molecule_key(molecule): {
                    self.element_key(element): number
                    for element, number in molecule.elements.items()
                }
                for molecule in self.molecules
            ],
            "reactions": [
                {
                    "reactants": {
                        self.molecule_key(reactant): stoich
                        for reactant, stoich in reaction.reactants.items()
                    },
                    "products": {
                        self.molecule_key(product): stoich
                        for product, stoich in reaction.products.items()
                    },
                }
                for reaction in system.reactions
            ],
        }
        return result

```

```

saver = YamlSavingSystem()
print(yaml.dump(saver.save(s)))

```

```

elements:
- O
- C
- H
molecules:
    C1O2:
        C: 1
        O: 2
    C6H12O6:
        C: 6
        H: 12
        O: 6
    H2:
        H: 2
    H2O1:
        H: 2
        O: 1
    O2:
        O: 2
reactions:
- products:
    C1O2: 6
    H2O1: 6
  reactants:
    C6H12O6: 1
    O2: 6
- products:
    H2O1: 2
  reactants:
    H2: 2
    O2: 1

```

We can see that to make an easily parsed file format, without having to guess-recognise repeated entities based on their names (which is highly subject to data entry error), we effectively recover the same tables as found for the database model.

An alternative is to use a simple integer for such a primary key:

```

class YamlIntegerKeySavingSystem:
    def __init__(self):
        self.elements = {}
        self.molecules = {}

    def add_element(self, element):
        if element not in self.elements:
            self.elements[element] = len(self.elements)
        return self.elements[element]

    def add_molecule(self, molecule):
        if molecule not in self.molecules:
            self.molecules[molecule] = len(self.molecules)
        return self.molecules[molecule]

    def element_key(self, element):
        return self.elements[element]

    def molecule_key(self, molecule):
        return self.molecules[molecule]

    def save(self, system):
        for reaction in system.reactions:
            for molecule in reaction.reactants:
                self.add_molecule(molecule)
                for element in molecule.elements:
                    self.add_element(element)
            for molecule in reaction.products:
                self.add_molecule(molecule)
                for element in molecule.elements:
                    self.add_element(element)

        result = {
            "elements": [element.symbol for element in self.elements],
            "molecules": [
                self.molecule_key(molecule): {
                    self.element_key(element): number
                    for element, number in molecule.elements.items()
                }
                for molecule in self.molecules
            ],
            "reactions": [
                {
                    "reactants": {
                        self.molecule_key(reactant): stoich
                        for reactant, stoich in reaction.reactants.items()
                    },
                    "products": {
                        self.molecule_key(product): stoich
                        for product, stoich in reaction.products.items()
                    },
                },
                for reaction in system.reactions
            ],
        }
        return result

```

```

saver = YamlIntegerKeySavingSystem()
print(yaml.dump(saver.save(s)))

```

```

elements:
- C
- H
- O
molecules:
0:
  0: 6
  1: 12
  2: 6
1:
  2: 2
2:
  0: 1
  2: 2
3:
  1: 2
  2: 1
4:
  1: 2
reactions:
- products:
  2: 6
  3: 6
  reactants:
    0: 1
    1: 6
- products:
  3: 2
  reactants:
    1: 1
    4: 2

```

10.2 Binary formats

Estimated time for this notebook: 10 minutes

Reference counting

Using a dictionary to determine the integer keys for objects is a bit clunky.

A better approach is to use counted objects either via a static member or by using a factory pattern:

```

class Element:
    def __init__(self, symbol, id):
        self.symbol = symbol
        self.id = id

class Molecule:
    def __init__(self, id):
        self.elements = {} # Map from element to number of that element in the molecule
        self.id = id

    def add_element(self, element, number):
        self.elements[element] = number

    def to_struct(self):
        return {element.symbol: number for element, number in self.elements.items()}

class Reaction:
    def __init__(self):
        self.reactants = {} # Map from reactants to stoichiometries
        self.products = {} # Map from products to stoichiometries

    def add_reactant(self, reactant, stoichiometry):
        self.reactants[reactant] = stoichiometry

    def add_product(self, product, stoichiometry):
        self.products[product] = stoichiometry

    def to_struct(self):
        return {
            "reactants": [x.to_struct() for x in self.reactants],
            "products": [x.to_struct() for x in self.products],
            "stoichiometries": list(self.reactants.values()) + list(self.products.values())
        }

class System: # This will be our factory
    def __init__(self):
        self.reactions = []
        self.elements = {}
        self.molecules = []

    def add_element(self, symbol):
        new_element = Element(symbol, len(self.elements))
        self.elements.append(new_element)
        return new_element

    def add_molecule(self):
        new_molecule = Molecule(len(self.molecules))
        self.molecules.append(new_molecule)
        return new_molecule

    def add_reaction(self):
        new_reaction = Reaction()
        self.reactions.append(new_reaction)
        return new_reaction

    def save(self):

        result = {
            "elements": [element.symbol for element in self.elements],
            "molecules": [
                {
                    molecule.id: {
                        element.id: number for element, number in molecule.elements.items()
                    }
                } for molecule in self.molecules
            ],
            "reactions": [
                {
                    "reactants": {
                        reactant.id: stoich
                        for reactant, stoich in reaction.reactants.items()
                    },
                    "products": {
                        product.id: stoich
                        for product, stoich in reaction.products.items()
                    },
                } for reaction in self.reactions
            ],
        }
        return result

```

```

s2 = System()

c = s2.add_element("C")
o = s2.add_element("O")
h = s2.add_element("H")

co2 = s2.add_molecule()
co2.add_element(c, 1)
co2.add_element(o, 2)

h2o = s2.add_molecule()
h2o.add_element(h, 2)
h2o.add_element(o, 1)

o2 = s2.add_molecule()
o2.add_element(o, 2)

h2 = s2.add_molecule()
h2.add_element(h, 2)

glucose = s2.add_molecule()
glucose.add_element(c, 6)
glucose.add_element(h, 12)
glucose.add_element(o, 6)

combustion_glucose = s2.add_reaction()
combustion_glucose.add_reactant(glucose, 1)
combustion_glucose.add_reactant(o2, 6)
combustion_glucose.add_product(co2, 6)
combustion_glucose.add_product(h2o, 6)

```

```

combustion_hydrogen = s2.add_reaction()
combustion_hydrogen.add_reactant(h2, 2)
combustion_hydrogen.add_reactant(o2, 1)
combustion_hydrogen.add_product(h2o, 2)

```

```

s2.save()

```

```

{'elements': ['C', 'O', 'H'],
'molecules': {0: {0: 1, 1: 2},
 1: {2: 2, 1: 1},
 2: {1: 2},
 3: {2: 2},
 4: {0: 6, 2: 12, 1: 6}},
'reactions': [{'reactants': {4: 1, 2: 6}, 'products': {0: 6, 1: 6}},
 {'reactants': {3: 2, 2: 1}, 'products': {1: 2}}]}

```

```

import yaml
print(yaml.dump(s2.save()))

```

```

elements:
- C
- O
- H
molecules:
 0:
    0: 1
    1: 2
  1:
    1: 1
    2: 2
  2:
    1: 2
  3:
    2: 2
  4:
    0: 6
    1: 6
    2: 12
reactions:
- products:
  0: 6
  1: 6
  reactants:
  2: 6
  4: 1
- products:
  1: 2
  reactants:
  2: 1
  3: 2

```

Binary file formats

Now we're getting toward a numerically-based data structure, using integers for object keys, we should think about binary serialisation.

Binary file formats are much smaller than human-readable text based formats, so important when handling really big datasets.

One can compress a textual file format, of course, and with good compression algorithms this will be similar in size to the binary file. (C.f. discussions of Shannon information density!) However, this has performance implications.

A hand-designed binary format is fast and small, at the loss of human readability.

The problem with binary file formats, is that, lacking complex data structures, one needs to supply the *length* of an item before that item:

```

class FakeBinarySavingSystem:
    # Pretend binary-style writing to a list to make it easier to read at first.
    def save(self, system, buffer):
        buffer.append(len(system.elements))
        for element in system.elements:
            buffer.append(element.symbol)

        buffer.append(len(system.molecules))
        for molecule in system.molecules:
            buffer.append(len(molecule.elements))
            for element, number in molecule.elements.items():
                buffer.append(element.id)
                buffer.append(number)

        buffer.append(len(system.reactions))
        for reaction in system.reactions:
            buffer.append(len(reaction.reactants))
            for reactant, stoich in reaction.reactants.items():
                buffer.append(reactant.id)
                buffer.append(stoich)
            buffer.append(len(reaction.products))
            for product, stoich in reaction.products.items():
                buffer.append(product.id)
                buffer.append(stoich)

```

```

arraybuffer = []
FakeBinarySavingSystem().save(s2, arraybuffer)

```

```

arraybuffer

```

[3,
'C'
'0'
'H'
5,
2,
0,
1,
1,
2,
2,
2,
2,
1,
1,
1,
1,
1,
1,
6,
2,
12,
1,
6,
2,
2,
4,
1,
2,
6,
2,
0,
6,
1,
6,
2,
3,
2,
2,
1,
1,
1,
2]

Deserialisation is left as an exercise for the reader :).

Endian-robust binary file formats

Having prepared our data as a sequence of data which can be recorded in a single byte, we might think a binary file format on disk is as simple as saving each number in one byte:

```
# First, turn symbol characters to equivalent integers (ascii)
intarray = [x.encode("ascii")[0] if isinstance(x, str) else x for x in
arraybuffer]
intarray
```

[3,
67,
79,
72,
5,
2,
0,
1,
1,
2,
2,
2,
1,
1,
1,
1,
2,
1,
2,
2,
2,
3,
0,
6,
2,
12,
1,
6,
2,
2,
4,
1,
2,
6,
2,
0,
6,
1,
6,
2,
3,
2,
2,
1,
1,
1,
2]

bytarray(intarray)

```
bytearray(b'\x03COH\x05\x02\x00\x01\x01\x02\x02\x02\x02\x02\x01\x01\x01\x01\x01\x02\x01\x02\x01\x02\x02\x03\x02\x00\x01\x06\x02\x01\x01\x06\x02\x02\x04\x01\x02\x06\x02\x00\x06\x01\x06\x02\x03\x02\x02\x02\x01\x01\x01\x02')
```

```
with open("system.mol", "bw") as binfile:  
    binfile.write(bytarray(intarray))
```

However, this misses out on an unfortunate problem if we end up with large enough numbers to need more than one byte per integer, or we want to represent floats: different computer designs will put the most-significant bytes of a multi-byte integer or float either at the beginning ('big endian' systems) or at the end ('little endian' systems).

To get around this, we need to use a portable standard for making binary files.

One possible choice is **XDR** (standing for eXternal Data Representation). XDR is a standard data serialization format that accounts for endian differences between systems.

A higher level approach to binary file formats: HDF5

This was quite painful. We've shown you it because it is very likely you will encounter this kind of unpleasant binary file format in your work.

However, one recommended approach to building binary file formats is to use HDF5 (Hierarchical Data Format), a much higher level binary file format.

HDF5's approach requires you to represent your system in terms of high-dimensional matrices, like NumPy arrays. It then saves these, and handles all the tedious number-of-field management for you.

```

import h5py
import numpy as np

class HDF5SavingSystem(System):
    def __init__(self, system):
        super().__init__()
        # Shallow Copy constructor
        self.elements = system.elements
        self.reactions = system.reactions
        self.molecules = system.molecules

    def element_symbols(self):
        return list(map(lambda x: x.symbol.encode("ascii"), self.elements))

    def molecule_matrix(self):
        molecule_matrix = np.zeros((len(self.elements), len(self.molecules)), dtype=int)

        for molecule in self.molecules:
            for element, n in molecule.elements.items():
                molecule_matrix[element.id, molecule.id] = n

        return molecule_matrix

    def reaction_matrix(self):
        reaction_matrix = np.zeros(
            (len(self.molecules), len(self.reactions)), dtype=int
        )

        for i, reaction in enumerate(self.reactions):
            for reactant, n in reaction.reactants.items():
                reaction_matrix[reactant.id, i] = -1 * n

            for product, n in reaction.products.items():
                reaction_matrix[product.id, i] = n

        return reaction_matrix

    def write(self, filename):
        hdf = h5py.File(filename, "w")
        string_type = h5py.special_dtype(vlen=bytes)
        hdf.create_dataset(
            "symbols", (len(self.elements), 1), string_type,
            self.element_symbols()
        )
        hdf.create_dataset("molecules", data=self.molecule_matrix())
        hdf.create_dataset("reactions", data=self.reaction_matrix())
        hdf.close()

```

```
{ saver.element_symbols()
```

```
[b'C', b'O', b'H']
```

```
{ saver.molecule_matrix()
```

```
array([[ 1,  0,  0,  0,  6],
       [ 2,  1,  2,  0,  6],
       [ 0,  2,  0,  2, 12]])
```

```
{ saver.reaction_matrix()
```

```
array([[ 6,  0],
       [ 6,  2],
       [-6, -1],
       [ 0, -2],
       [-1,  0]])
```

```
{ saver.write("foo.hdf5")
```

Note that this binary representation is *not* human readable at all.

```
%bash
# Read the first 100 characters from the file
head -c 100 foo.hdf5

with open("module10_scientific_file_formats/foo.hdf5", "rb") as f_in:bytes =
f_in.read()
...
>>> bytes[0:100]
```

```
bash: line 4: syntax error near unexpected token `('
```

```
bash: line 4: `with open("module10_scientific_file_formats/foo.hdf5", "rb") as
f_in:bytes = f_in.read()'
```

```
Task exception was never retrieved
future: <Task finished name='Task-7' coro=<ScriptMagics.shebang.
<locals>._handle_stream() done, defined at
/opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-
packages/IPython/core/magics/script.py:211> exception=UnicodeDecodeError('utf-8',
b'\x89HDF\x0d\x0a', 0, 1, 'invalid start byte')>
Traceback (most recent call last):
  File "/opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-
packages/IPython/core/magics/script.py", line 213, in _handle_stream
    line = (await stream.readline()).decode('utf8')
UnicodeDecodeError: 'utf-8' codec can't decode byte 0x89 in position 0: invalid
start byte
```

```
-----  
CalledProcessError                         Traceback (most recent call last)
Cell In [21], line 1
----> 1 get_ipython().run_cell_magic('bash', '', '# Read the first 100 characters
from the file\nhead -c 100 foo.hdf5\nwith
open("module10_scientific_file_formats/foo.hdf5", "rb") as f_in:bytes =
f_in.read()\n...>>> bytes[0:100]\n'
File /opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-
packages/IPython/core/interactiveshell.py:2417, in
InteractiveShell.run_cell_magic(self, magic_name, line, cell)
 2415     with self.builtin_trap:
 2416         args = (magic_arg_s, cell)
-> 2417         result = fn(*args, **kwargs)
 2418     return result

File /opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-
packages/IPython/core/magics/script.py:153, in ScriptMagics._make_script_magic.
<locals>.named_script_magic(line, cell)
 151     else:
 152         line = script
--> 153     return self.shebang(line, cell)

File /opt/hostedtoolcache/Python/3.8.14/x64/lib/python3.8/site-
packages/IPython/core/magics/script.py:305, in ScriptMagics.shebang(self, line,
cell)
300     if args.raise_error and p.returncode != 0:
301         # If we get here and p.returncode is still None, we must have
302         # killed it but not yet seen its return code. We don't wait for it,
303         # in case it's stuck in uninterruptible sleep. -9 = SIGKILL
304         rc = p.returncode or -9
--> 305         raise CalledProcessError(rc, cell)

CalledProcessError: Command 'b'# Read the first 100 characters from the file\nhead
-c 100 foo.hdf5\nwith
open("module10_scientific_file_formats/foo.hdf5", "rb") as
f_in:bytes = f_in.read()\n...>>> bytes[0:100]\n' returned non-zero exit status
2.
```

```
hdf_load = h5py.File("foo.hdf5")
```

```
{ np.array(hdf_load["reactions"])}
```

Using a `sparse matrix` storage would be even better here, but we don't have time for that!

10.3 Markup Languages

Estimated time for this notebook: 15 minutes

XML and its relatives (including HTML) are based on the idea of *marking up* content with labels on its purpose:

```
<name>James</name> is a <job>Programmer</job>
```

One of the easiest ways to make a markup-language based fileformat is the use of a *templating language*.

```

from IPython.display import Math, display
from parsereactions import parser

with open("system.tex", "r") as f_latex:
    system = parser.parse(f_latex.read())
display(Math(str(system)))

```

```

ModuleNotFoundError: Traceback (most recent call last)
Cell In [1], line 2
      1 from IPython.display import Math, display
----> 2 from parsereactions import parser
      3 with open("system.tex", "r") as f_latex:
      4     system = parser.parse(f_latex.read())

ModuleNotFoundError: No module named 'parsereactions'

```

```

%%writefile chemistry_template.mko
<?xml version="1.0" encoding="UTF-8"?>
<system>
%for reaction in reactions:
<reaction>
    <reactants>
        %for molecule in reaction.reactants.molecules:
            <molecule stoichiometry="${reaction.reactants.molecules[molecule]}">
                %for element in molecule.elements:
                    <atom symbol="${element.symbol}" number="${molecule.elements[element]}/>
                %endfor
            </molecule>
        %endfor
    </reactants>
    <products>
        %for molecule in reaction.products.molecules:
            <molecule stoichiometry="${reaction.products.molecules[molecule]}">
                %for element in molecule.elements:
                    <atom symbol="${element.symbol}" number="${molecule.elements[element]}/>
                %endfor
            </molecule>
        %endfor
    </products>
</reaction>
%endfor
</system>

```

Writing chemistry_template.mko

```

from mako.template import Template

mytemplate = Template(filename="chemistry_template.mko")
with open("system.xml", "w") as xmlfile:
    xmlfile.write((mytemplate.render(**vars(system))))

```

!cat system.xml

```

<?xml version="1.0" encoding="UTF-8"?>
<system>
<reaction>
    <reactants>
        <molecule stoichiometry="1">
            <atom symbol="C" number="6"/>
            <atom symbol="H" number="12"/>
            <atom symbol="O" number="6"/>
        </molecule>
        <molecule stoichiometry="6">
            <atom symbol="O" number="2"/>
        </molecule>
    </reactants>
    <products>
        <molecule stoichiometry="6">
            <atom symbol="C" number="1"/>
            <atom symbol="O" number="2"/>
        </molecule>
        <molecule stoichiometry="6">
            <atom symbol="H" number="2"/>
            <atom symbol="O" number="1"/>
        </molecule>
    </products>
</reaction>
<reaction>
    <reactants>
        <molecule stoichiometry="2">
            <atom symbol="H" number="2"/>
        </molecule>
        <molecule stoichiometry="1">
            <atom symbol="O" number="2"/>
        </molecule>
    </reactants>
    <products>
        <molecule stoichiometry="2">
            <atom symbol="H" number="2"/>
            <atom symbol="O" number="1"/>
        </molecule>
    </products>
</reaction>
</system>

```

Markup languages are verbose (jokingly called the "angle bracket tax") but very clear.

Data as text

The above serialisation specifies all data as XML "Attributes". An alternative is to put the data in the text:

```

%%writefile chemistry_template2.mko
<?xml version="1.0" encoding="UTF-8"?>
<system>
%for reaction in reactions:
    <reaction>
        <reactants>
            %for molecule in reaction.reactants.molecules:
                <molecule stoichiometry="${reaction.reactants.molecules[molecule]}">
                    %for element in molecule.elements:
                        <atom symbol="${element.symbol}">${molecule.elements[element]}

```

Writing chemistry_template2.mko

```

mytemplate = Template(filename="chemistry_template2.mko")
with open("system2.xml", "w") as xmlfile:
    xmlfile.write((mytemplate.render(*vars(system))))

```

!cat system2.xml

```

<?xml version="1.0" encoding="UTF-8"?>
<system>
    <reaction>
        <reactants>
            <molecule stoichiometry="1">
                <atom symbol="C">1</atom>
                <atom symbol="H">12</atom>
                <atom symbol="O">6</atom>
            </molecule>
            <molecule stoichiometry="6">
                <atom symbol="O">2</atom>
            </molecule>
        </reactants>
        <products>
            <molecule stoichiometry="6">
                <atom symbol="C">1</atom>
                <atom symbol="O">2</atom>
            </molecule>
            <molecule stoichiometry="6">
                <atom symbol="H">2</atom>
                <atom symbol="O">1</atom>
            </molecule>
        </products>
    </reaction>
    <reaction>
        <reactants>
            <molecule stoichiometry="2">
                <atom symbol="H">2</atom>
            </molecule>
            <molecule stoichiometry="1">
                <atom symbol="O">2</atom>
            </molecule>
        </reactants>
        <products>
            <molecule stoichiometry="2">
                <atom symbol="H">2</atom>
                <atom symbol="O">1</atom>
            </molecule>
        </products>
    </reaction>
</system>

```

Parsing XML

XML is normally parsed by building a tree-structure of all the `tags` in the file, called a `DOM` or Document Object Model.

```

from lxml import etree

with open("system.xml", "r") as xmlfile:
    tree = etree.parse(xmlfile)
    print(etree.tostring(tree, pretty_print=True, encoding=str))

```

```

<system>
  <reaction>
    <reactants>
      <molecule stoichiometry="1">
        <atom symbol="C" number="6"/>
        <atom symbol="H" number="12"/>
        <atom symbol="O" number="6"/>
      </molecule>
      <molecule stoichiometry="6">
        <atom symbol="O" number="2"/>
      </molecule>
    </reactants>
    <products>
      <molecule stoichiometry="6">
        <atom symbol="C" number="1"/>
        <atom symbol="O" number="2"/>
      </molecule>
      <molecule stoichiometry="6">
        <atom symbol="H" number="2"/>
        <atom symbol="O" number="1"/>
      </molecule>
    </products>
  </reaction>
  <reaction>
    <reactants>
      <molecule stoichiometry="2">
        <atom symbol="H" number="2"/>
      </molecule>
      <molecule stoichiometry="1">
        <atom symbol="O" number="2"/>
      </molecule>
    </reactants>
    <products>
      <molecule stoichiometry="2">
        <atom symbol="H" number="2"/>
        <atom symbol="O" number="1"/>
      </molecule>
    </products>
  </reaction>
</system>

```

We can navigate the tree, with each **element** being an iterable yielding its children:

```

tree.getroot()[0][0][1].attrib["stoichiometry"]
'6'

```

Searching XML

xpath is a sophisticated tool for searching XML DOMs:

There's a good explanation of how it works here: https://www.w3schools.com/xml/xml_xpath.asp but the basics are reproduced below.

XPath Expression	Result
/bookstore/book[1]	Selects the first book that is the child of a bookstore
/bookstore/book[last()]	Selects the last book that is the child of a bookstore
/bookstore/book[last()-1]	Selects the last but one book that is the child of a bookstore
/bookstore/book[position()<3]	Selects the first two books that are children of a bookstore
//title[@lang]	Selects all titles that have an attribute named "lang"
//title[@lang='en']	Selects all titles that have a "lang" attribute with a value of "en"
/bookstore/book[price>35.00]	Selects all books that are children of a bookstore and have a price with a value greater than 35.00
/bookstore/book[price>35.00]/title	Selects all the titles of a book of a bookstore that have a price with a value greater than 35.00

```

# For all molecules
# ... with a child atom whose number attribute is '1'
# ... return the symbol attribute of that child
tree.xpath("//molecule/atom[@number='1']/@symbol")

```

```
'C', 'O', 'O'
```

It is useful to understand grammars like these using the "FOR-LET-WHERE-ORDER-RETURN" (pronounced Flower) model.

The above says: "For element in molecules where number is one, return symbol", roughly equivalent to `[element.symbol for element in molecule for molecule in document if element.number==1]` in Python.

```

with open("system2.xml") as xmlfile:
    tree2 = etree.parse(xmlfile)
    # For all molecules with a child atom whose text is 1
    # ... return the symbol attribute of any child (however deeply nested)
    print(tree2.xpath("//molecule/atom[@number=1]//@symbol"))

```

```
'C', 'O', 'H', 'O', 'H', 'O'
```

Note how we select on text content rather than attributes by using the element tag directly. The above says "for every molecule where at least one element is present with just a single atom, return all the symbols of all the elements in that molecule."

Transforming XML : XSLT

Two technologies (XSLT and XQUERY) provide capability to produce text output from an XML tree.

We'll look at XSLT as support is more widespread, including in the python library we're using. XQuery is probably easier to use and understand, but with less support.

However, XSLT is a beautiful functional declarative language, once you read past the angle-brackets.

Here's an XSLT to transform our reaction system into a LaTeX representation:

```

%%writefile xmltotex.xsl
<xsl:stylesheet version="2.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
  <xsl:output method="xml" indent="yes" omit-xml-declaration="yes" />
  <!-- Decompose reaction into "reactants \rightarrow products" -->
  <xsl:template match="/reaction">
    <xsl:apply-templates select="reactants"/>
    <xsl:text> \rightarrow </xsl:text>
    <xsl:apply-templates select="products"/>
    <xsl:text>\&#xa;</xsl:text>
  </xsl:template>

  <!-- For a molecule anywhere except the first position write " + " and the
  number of molecules-->
  <xsl:template match="//molecule[position()!=1]">
    <xsl:text> + </xsl:text>
    <xsl:apply-templates select="@stoichiometry"/>
    <xsl:apply-templates/>
  </xsl:template>

  <!-- For a molecule in first position write the number of molecules -->
  <xsl:template match="//molecule[position()=1]">
    <xsl:apply-templates select="@stoichiometry"/>
    <xsl:apply-templates/>
  </xsl:template>

  <!-- If the stoichiometry is one then ignore it -->
  <xsl:template match="@stoichiometry[.=1']"/>

  <!-- Otherwise, use the default template for attributes, which is just to copy
  value -->

  <!-- Decompose element into "symbol number" -->
  <xsl:template match="/atom">
    <xsl:value-of select="@symbol"/>
    <xsl:apply-templates select="@number"/>
  </xsl:template>

  <!-- If the number of elements/molecules is one then ignore it -->
  <xsl:template match="@number[.=1']"/>

  <!-- ... otherwise replace it with "_value" -->
  <xsl:template match="@number[.!=1][10..]>">
    <xsl:text>.</xsl:text>
    <xsl:value-of select="."/>
  </xsl:template>

  <!-- If a number is greater than 10 then wrap it in "{}" -->
  <xsl:template match="@number[.!=1][.>9]">
    <xsl:text>{</xsl:text>
    <xsl:value-of select="."/>
    <xsl:text>}</xsl:text>
  </xsl:template>

  <!-- Do not copy input whitespace to output -->
  <xsl:template match="text()" />
</xsl:stylesheet>

```

Writing xmltotex.xsl

```

with open("xmltotex.xsl") as xslfile:
    transform_xsl = xslfile.read()
transform = etree.XSLT(etree.XML(transform_xsl))

```

```

print(str(transform(tree)))

```

```

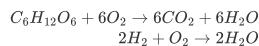
C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O\\
2H_2 + O_2 \rightarrow 2H_2O\\

```

```

display(Math(str(transform(tree))))

```



Validating XML : Schema

XML Schema is a way to define how an XML file is allowed to be: which attributes and tags should exist where.

You should always define one of these when using an XML file format.

```
%>%writefile reactions.xsd
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="atom">
    <xs:complexType>
      <xs:attribute name="symbol" type="xs:string"/>
      <xs:attribute name="number" type="xs:integer"/>
    </xs:complexType>
  </xs:element>

  <xs:element name="molecule">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="atom" maxOccurs="unbounded"/>
      </xs:sequence>
      <xs:attribute name="stoichiometry" type="xs:integer"/>
    </xs:complexType>
  </xs:element>

  <xs:element name="reaction">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="reactants">
          <xs:complexType>
            <xs:sequence>
              <xs:element ref="molecule" maxOccurs="unbounded"/>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
        <xs:element name="products">
          <xs:complexType>
            <xs:sequence>
              <xs:element ref="molecule" maxOccurs="unbounded"/>
            </xs:sequence>
          </xs:complexType>
        </xs:element>
      </xs:sequence>
    </xs:complexType>
  </xs:element>

  <xs:element name="system">
    <xs:complexType>
      <xs:sequence>
        <xs:element ref="reaction" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Writing reactions.xsd

```
with open("reactions.xsd") as xsdfile:
    schema_xsd = xsdfile.read()
schema = etree.XMLSchema(etree.XML(schema_xsd))
```

```
parser = etree.XMLParser(schema=schema)
```

```
with open("system.xml") as xmlfile:
    tree = etree.parse(xmlfile, parser)
    # For all atoms return their symbol attribute
    tree.xpath("//atom/@symbol")
```

```
['C', 'H', 'O', 'O', 'C', 'O', 'H', 'O', 'H', 'O', 'H', 'O']
```

Compare parsing something that is not valid under the schema:

```
%>%writefile invalid_system.xml
<system>
  <reaction>
    <reactants>
      <molecule stoichiometry="two">
        <atom symbol="H" number="2"/>
      </molecule>
      <molecule stoichiometry="1">
        <atom symbol="O" number="2"/>
      </molecule>
    </reactants>
    <products>
      <molecule stoichiometry="2">
        <atom symbol="H" number="2"/>
        <atom symbol="O" number="2"/>
      </molecule>
    </products>
  </reaction>
</system>
```

Writing invalid_system.xml

```
try:
    with open("invalid_system.xml") as xmlfile:
        tree = etree.parse(xmlfile, parser)
        tree.xpath("//element//@symbol")
except etree.XMLSyntaxError as e:
    print(e)
```

```
Element 'molecule', attribute 'stoichiometry': 'two' is not a valid value of the
atomic type 'xs:integer'. (<string>, line 0)
```

This shows us that the validation has failed and why.

10.04 Larger datasets - beyond pandas and csv

Estimated time for this notebook: 20 minutes.

Much of the data that we deal with can be represented in tabular form, and can be handled in data structures such as the *pandas DataFrame*. We have already (briefly) seen how we can read and write csv files from pandas, and there are also methods for reading the results of SQL queries into *pandas DataFrames*.

However, if we have very large datasets (millions of rows), or cases where we need fast and intensive processing on these tables, *pandas* may not be the best choice.

Row-wise vs column-wise

Let's read a csv file containing international men's football results into a *pandas DataFrame*:

```
import pandas as pd
df = pd.read_csv("match_results.csv")
df.head()
```

	Unnamed: 0	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
0	0	1872-11-30	Scotland	England	0	0	Friendly	Glasgow	Scotland	False
1	1	1873-03-08	England	Scotland	4	2	Friendly	London	England	False
2	2	1874-03-07	Scotland	England	2	1	Friendly	Glasgow	Scotland	False
3	3	1875-03-06	England	Scotland	2	2	Friendly	London	England	False
4	4	1876-03-04	Scotland	England	3	0	Friendly	Glasgow	Scotland	False

The obvious way to think of this table is "row-wise" i.e. each row is a single match, with various attributes (the columns). If we want to look at a certain match, we can pick it out using its index, and then look at it in detail:

```
match = df.iloc[3]
print(f"{match.home_team} ({match.home_score})-{(match.away_score)} {match.away_team}")
```

```
England 2:2 Scotland
```

Similarly, when more matches get played, we can simply append more rows to the end of the table.

However, for storing the data and for performing some types of operation on it, this is far from the most efficient approach. Note that the different columns here have different types - we have dates, strings, integers, bools. If we look at the data in a *columnar* way, we can make use of this, and use some compression tricks.

Since some data types such as integers have fixed and known sizes, we can easily imagine that it's more efficient to pack these together, so the "home_score" column would be [0,4,2,2,3,...], without having to worry about other columns containing e.g. strings, of varying size.

However, we can do even better.

Run length compression

If we look at the last column, which is a boolean telling us whether the match was at a "neutral" venue (e.g. at a World Cup or European Championship). Most matches will be at non-neutral venues, but every two or four years there will be a cluster of "neutral" matches.

```
# find the longest run of matches with the same value of 'neutral'
previous_val = df.neutral.values[0]
run = 0
longest_run = 0
for val in df.neutral.values[1:]:
    if val == previous_val:
        run += 1
        if run > longest_run:
            longest_run = run
    else:
        run = 0
print("longest run is:", longest_run)
```

```
longest run is: 198
```

We could save a *lot* of space, with no loss of information, if rather than saving every value, we instead save something that means "False 198 times, followed by True 64 times, followed by ...".

Dictionary compression

Often the most space-consuming type in a dataset is a string. Each character will take one or two bytes (based on either UTF-8 or UTF-16 encoding), and we could either store each one as a variable-length string (so short strings will take less space than long strings), or we can decide on a maximum length for our string field, and pad the shorter strings. The former option takes less space, but is much less efficient when it comes to looking up values.

However, in many data tables, the same values are repeated many times.

```
# how many rows in the table?
print(f"Table has {len(df)} rows")
# how many unique values in 'home_team' column?
print(f"Number of unique home_team values: {len(df.home_team.unique())}")
# how many times does 'Brazil' appear?
print(f"Brazil has been the home team {df.home_team.value_counts().Brazil} times")
```

```
Table has 44060 rows
Number of unique home_team values: 311
Brazil has been the home team 591 times
```

Again, we could save a lot of space if we make a lookup table, so we e.g. assign each team name to an integer, and rather than taking 20 bytes to store the longest team name, we would just use 2 bytes for every team.

Delta compression

If we really want to compress our data as much as possible for writing to disk, and we don't care about making it human readable, we can use further tricks such as delta compression. For time series data, if something is relatively smoothly varying, we can save a lot of space by storing just the difference from one data point to the next, rather than each value. As an illustration of how this could work: rather than storing all the dates in the "date" table of our dataframe, we could just store the first date, and then for every subsequent row we just store the number of days since the previous row.

In the case where the column we're trying to compress contains integers or floats, we wouldn't save any space if we were to store the differences as integers or floats as well, but lots of clever schemes exist for packing small deltas within a few bytes.

For example, given the sequence:

```
5, 3, 3, 4, 2, 1, 2, 0
```

the deltas are:

```
-2, 0, 1, -2, -1, 1, -2
```

we can rescale this set of deltas by subtracting the minimum value (-2) from each element, such that the new minimum is 0, giving:

```
0, 2, 3, 0, 1, 3, 0
```

and finally we can encode this "block" along with a "header", as follows:

```
header: 8 (block size), 5 (first value)
block: -2 (minimum delta), 2 (bitwidth), 00101100011100b (0,2,3,0,1,3,0 packed on 2 bits)
```

For this trivial example, we are not actually saving that much space, but we could extend this to have many more blocks, and/or longer blocks, and/or have a block/miniblock structure (e.g. for when we need to change the bitwidth to deal with larger deltas), and the overall saving could be huge.

Of course, doing all these compression steps by hand is fiddly, and we would be very likely to make a mistake! But luckily, libraries exist that do the hard work for us, and can seamlessly convert between pandas DataFrames and compressed columnar formats.

Putting this into action: *parquet*

One data format that implements all these forms of compression (see [here](#)) is "parquet": <https://parquet.apache.org/> Parquet files can be read in many languages, including Python, R, C++, and Java.

Let's write our dataframe as a *parquet* file:

```
{ df.to_parquet("match_results.parquet") }
```

How much space did we save compared to the csv?

```
{ !du -skh match_results.* }
```

```
3.4M    match_results.csv
700K    match_results.parquet
```

About a factor of 9!

Arrow and feather

You may have noticed that one of the packages that we installed in order to write the parquet file was [pyarrow](#). Apache arrow is one of the under-the-hood technologies that parquet uses to process data. It is an "in-memory" columnar data format with some nice properties: random access is O(1) and each value cell is next to the previous and following one in memory, so it is efficient for iteration.

We can convert our pandas dataframe directly into an arrow table:

```
from timeit import timeit
import pyarrow as pa
import pyarrow.compute as pc

table = pa.Table.from_pandas(df)
table
```



```
pyarrow.Table
Unnamed: 0: int64
date: string
home_team: string
away_team: string
home_score: int64
away_score: int64
tournament: string
city: string
country: string
neutral: bool
...
Unnamed: 0: [[0,1,2,3,4,...,44055,44056,44057,44058,44059]]
date: [[1872-11-30,"1873-03-08","1874-03-07","1875-03-06","1876-03-04",..., "2022-09-27","2022-09-27","2022-09-27","2022-09-30"]]
home_team: [["Scotland","England","Scotland","England","Scotland",...,"Norway","Sweden","Kosovo","Greece","Fiji"]]
away_team: [["England","Scotland","England","Scotland","England",...,"Serbia","Slovenia","Cyrus","Northern Ireland","Solomon Islands"]]
home_score: [[0,4,2,2,3,...,0,1,5,3,0]]
away_score: [[0,2,1,2,0,...,2,1,1,1,0]]
tournament: [[["Friendly","Friendly","Friendly","Friendly","Friendly",...,"UEFA Nations League","UEFA Nations League","UEFA Nations League","UEFA Nations League","UEFA Nations League"], "MS Prime Minister's Cup"]]]
city: [["Glasgow","London","Glasgow","London","Glasgow",...,"Oslo","Stockholm","Pristina","Athens","LuganoVille"]]
country: [["Scotland","England","Scotland","England","Scotland",...,"Norway","Sweden","Kosovo","Greece","Vanuatu"]]]
neutral: [[false,false,false,false,...,false,false,false,true]]
```

Some things such as summing over columns are usually faster than in pandas:

```
ptime = timeit("df.away_score.sum()", globals=globals(), number=10000)
atime = timeit('pc.sum(table.column("away_score"))', globals=globals(),
number=10000)
print(f"Pandas took {ptime}, Arrow took {atime} to sum this column 10k times")
```

```
Pandas took 0.5401166500000727, Arrow took 0.13334805099998448 to sum this column
10k times
```

So arrow is about a factor 3-4 faster in this particular case.

Should we always use *arrow* instead of *pandas* then? It depends. Arrow may be faster for some operations, so if you're speed-limited, it could be worth switching (or at least testing whether it's worth it). But on the other hand, *pandas* has a healthy userbase, a well-known API, and established interfaces to other tools (e.g. *matplotlib* for plotting). The good news is that it's very easy to convert between *pandas* DataFrames and *arrow* Tables, and vice versa, so it

shouldn't be a problem to try both and see what works best for your use-case.

Writing to disk: *feather*

We have already seen that we can write tabular data in a columnar format to disk as a *parquet* file. Another option is *feather*. *Feather* is a direct on-disk representation of the in-memory *arrow* format - it doesn't have the same compression that *parquet* applies by default.

Let's write our *arrow* table to a *feather* file:

```
import pyarrow.feather as feather
feather.write_feather(table, "match_results.feather")
```

Now we have the same table in csv, *parquet*, and *feather* format. Compare the sizes again:

```
! du -skh match_results.*
```

```
3.4M  match_results.csv
2.1M  match_results.feather
700K  match_results.parquet
```

The *feather* format didn't compress anywhere near as much as the *parquet* file (but is still much smaller than csv).

So which is better, *feather* or *parquet*? Again, it depends what you are doing. If you will be storing or transferring large quantities of data, *parquet* is probably preferable. However, there is a CPU cost to the compression/decompression, so if you are more worried about the speed of reading and writing files, you might want to use *feather*.

10.5 Processing in parallel

Estimated time for this notebook: 20 minutes.

For large datasets, processing in-memory on a single thread might be too slow. There are a few potential options for processing this data in parallel, some of which we'll look at very briefly here (we won't go into any details - for more information you are recommended to look at the linked documentation).

Batch processing

One option could be to split your dataset into smaller subsets, and use a batch system to run many jobs in parallel on a cluster or farm of computers. A popular batch job scheduler is *Slurm* (<https://slurm.schedmd.com/documentation.html>) which offers tools for submitting jobs to batch queues, monitoring their progress, and keeping track of failures. Cloud providers such as Microsoft Azure have their own batch offerings (e.g. *Azure Batch*) with similar features.

However, even with tools such as these, there is usually quite a bit of overhead involved in figuring out how to split up the data, write submission scripts, and keeping track of completed or failed jobs.

MapReduce

MapReduce is a programming model for processing data using a cluster of worker nodes, often on a distributed filesystem. One such implementation is *Apache Hadoop* <https://hadoop.apache.org/>.

MapReduce consists of three main steps: **Map**, **Shuffle**, **Reduce**, which all operate on key, value pairs. Much of the possible speedup in a MapReduce workflow is if one is able to send "code-to-data", i.e. have expensive "map" operations run on nodes that have fast access to the relevant bit of data.



The canonical (trivial) example of MapReduce is a word-count problem - suppose we have a set of text files and we want to count the frequency of occurrence of each word. We want to be able to parallelize, so that each input could be processed by one node, and the results are brought together at the end in an efficient manner.

First we write a mapper function that takes a single filename as input, and outputs a sorted list of `{word: [1]}` dicts:

```
def mapper(input_filename):
    with open(input_filename) as inputfile:
        # split the text on spaces
        words = inputfile.read().split(" ")
        # use list comprehension to output a list of {word,1} dicts
        output = [{word.strip(): [1]} for word in sorted(words)]
    return output
```

```
mapper("text_sample_0.txt")
```

```
[{'best': [1],
  {'it': [1],
  {'it': [1],
  {'of': [1],
  {'of': [1],
  {'the': [1],
  {'the': [1],
  {'times': [1],
  {'times': [1],
  {'was': [1],
  {'was': [1],
  {'worst': [1]}}
```

The next step is to *shuffle*, bringing together all the items in the mapper output with the same key, so that each key's data can be sent to a different *reducer*.

```
def shuffler(word_dicts):
    output_dict = {}
    for word_dict in word_dicts:
        for k, v in word_dict.items():
            if not k in output_dict.keys():
                output_dict[k] = []
            output_dict[k] += v
    return [{k: v} for k, v in output_dict.items()]
```

```
shuffler(mapper("text_sample_0.txt"))
```

```
[{'best': [1],  
 {'it': [1, 1]},  
 {'of': [1, 1]},  
 {'the': [1, 1]},  
 {'times': [1, 1]},  
 {'was': [1, 1]},  
 {'worst': [1]}]
```

The *reducer* in this case is very simple - given a key (which is a word), and a value (which is a list [1,1,1,...]) sum up the values of the list to return a single value.

```
def reducer(word_dict):  
    for k, v in word_dict.items():  
        return {k: sum(v)}
```

```
reducer({'best': [1, 1]})
```

```
{'best': 2}
```

```
# loop over 7 input files  
shuffle_outputs = []  
for i in range(7):  
    shuffle_outputs += shuffler(mapper(f"text_sample_{i}.txt"))  
# another shuffle step to bring the outputs from the different mapper processes  
# together  
shuffle_outputs = shuffler(shuffle_outputs)  
# now we can farm each k,v pair from the shuffle_outputs to different reducers  
counts = []  
for word_dict in shuffle_outputs:  
    counts.append(reducer(word_dict))  
print(counts)
```

```
[{'best': 1}, {'it': 10}, {'of': 10}, {'the': 11}, {'times': 2}, {'was': 10},  
{'worst': 1}, {'age': 2}, {'foolishness': 1}, {'wisdom': 1}, {'belief': 1},  
{'epoch': 2}, {'incredulity': 1}, {'darkness': 1}, {'light': 1}, {'season': 2},  
{'despair': 1}, {'hope': 1}, {'spring': 1}, {'winter': 1}, {'before': 2},  
{'everything': 1}, {'had': 2}, {'nothing': 1}, {'us': 2}, {'we': 4}, {'all': 2},  
{'direct': 2}, {'going': 2}, {'heaven': 1}, {'other': 1}, {'to': 1}, {'way': 1},  
{'were': 2}]
```

Of course, this is a simple example, running entirely on our local machine, using a `for` loop. But it illustrates that for more complex cases, where there is data distributed over different locations, it is possible to have the "map" stage run in parallel on different machines, and similarly, once the "shuffle" stage has organized the data by key, it can send the "reduce" stage to be run on different machines in parallel.

Spark

One drawback of MapReduce is that is inefficient if the processing dataflow requires multiple passes (e.g. training a Machine Learning model). This was one of the motivations for the development of **Spark** <https://spark.apache.org/>

Spark is based on the concept of a resilient distributed dataset (RDD), a set of read-only data objects distributed over a cluster. The workflow can be represented as a directed acyclic graph (DAG) with the nodes as the RDDs and the edges as the operations to be performed on the RDDs. For some types of workflow, Spark is considerably ($\times 100$) quicker than Hadoop/MapReduce, and it can also handle streaming data by making micro-batches and processing them.

The package `pyspark` <https://spark.apache.org/docs/latest/api/python/> provides a Python interface to the Spark API. However, it does still need a Java runtime environment to work.

Dask

Another option, which is growing in popularity in the academic and scientific communities, is **Dask**. The idea behind Dask is to provide a familiar interface to `pandas` and `numpy` but to allow the same code to be run either locally or on a cluster. One of the tricks to facilitate this is "lazy evaluation" - when the code is run, the computation is not actually performed, but instead a "task graph" is built, where each node represents a Python function that performs a unit of computation, and the edges represent data dependencies between the upstream and downstream tasks.

Once the task graph is generated, a "scheduler" (which can be either "single-machine" or "distributed") manages the workflow by using the task graph to assign tasks to workers in a way that optimizes parallelism while respecting the data dependencies.

 (image from Dask documentation <https://docs.dask.org/en/stable/>)

Dask has a "dataframe", which can easily be constructed from its `pandas` equivalent. Let's use our "match_results.csv" for input again:

```
import dask.dataframe as dd  
import numpy as np  
import pandas as pd
```

```
df = pd.read_csv("match_results.csv")  
ddf = dd.from_pandas(df, npartitions=10)
```

Dask DataFrame Structure:

Unnamed: 0	date	home_team	away_team	home_score	away_score	tournament	city	country	neutral
4406
...
39654
44059

Dask Name: from_pandas, 1 graph layer

The Dask dataframe has 10 partitions, meaning that the 44k rows in the original csv are now divided into 10 batches of about 4.4k rows each.

```
ddf.divisions
```

```
(0, 4406, 8812, 13218, 17624, 22030, 26436, 30842, 35248, 39654, 44059)
```

The interface is very similar to `pandas`, with one important difference. For example, if we want to calculate the average of the "home_score" column, in `pandas` we can do:

```
df.home_score.values.mean()
```

```
1.7404675442578301
```

If we do the same in our `Dask` dataframe:

```
ddf.home_score.values.mean()
```

Array	Chunk
Bytes	8 B
Shape	0
Dask graph 1 chunks in 6 graph layers	
Data type	float64 numpy.ndarray

what we get back is the **Task Graph**. In order to actually run the calculation, we need to add `compute()`:

```
ddf.home_score.values.mean().compute()
```

```
1.7404675442578301
```

Let's try and do something more complicated. We can use the dataset to investigate whether "home advantage" is real, in international men's football matches. We have columns in the dataset for "home_team", "away_team" etc., but some of the matches were at tournaments in neutral territory, so we want to use the "neutral" column to exclude these. Having done that, we can just calculate the number of matches that the home team won, minus the number that the home team lost.

```
def home_team_wins(home_score, away_score, neutral):
    if neutral:
        return 0
    if home_score > away_score: # home win
        return 1
    elif home_score < away_score: # away win
        return -1
    else: # draw
        return 0

ddf["home_win"] = df.apply(
    lambda row: home_team_wins(row["home_score"], row["away_score"],
    row["neutral"]),
    axis=1,
)
result = ddf["home_win"].values.sum()
```

The Task Graph for this computation, on our dataframe with 10 partitions, looks like this:

Note that you can create this visualization for yourself, if you install the `graphviz` package (e.g. `brew install graphviz` on Mac) then install the Python `graphviz` package (`pip install graphviz`), then do `result.visualize()`.

Don't worry about the details, but we can see the 10 data partitions at the bottom, and the single result at the top, and a bunch of clever intermediate steps that Dask is figuring out for us.

At this point we can create either a local scheduler, or if we have a handy compute cluster, a distributed scheduler. Either way, we do this by creating an instance of the `dask.distributed.Client` class, with the URL of the scheduler.

There are instructions on setting up a Dask cluster here: <https://docs.dask.org/en/stable/deploying.html> but for now, let's just run on our local machine, with a local scheduler.

```
from dask.distributed import Client
client = Client() # or Client("<scheduler URL>") for remote cluster
client
```

Client

Client-b5c32f1e-5c31-11ed-ae41-7daf942b7c31

Connection method: Cluster object

Cluster type: distributed.LocalCluster

Dashboard: <http://127.0.0.1:8787/status>

Cluster Info

Note the link to the dashboard - this will provide some diagnostics into what the scheduler is doing.

Once we have created a client, whenever we call `compute()` it will run on the scheduler that the client points to (in this case a local scheduler)

```
result.compute()
```

```
8061
```

So, in our dataset, home teams won 8k times more often than away teams, so it seems that home advantage is a real thing!

10.6 Geospatial data

Estimated time for this notebook: 15 minutes.

Many domains have their own widely-used data file formats, which are optimized for their own most common use-cases. For example, geospatial datasets will often have some "coordinates" (e.g. Latitude and Longitude, and possibly Time), and a set of measurements at each point (e.g. Temperature, Humidity, Wind Speed). Storing such data in a simple table, such as in a `parquet` or `feather` file, would be inefficient, as the coordinate variables would be repeated for each of the measurements.

One binary file format that has been developed for use-cases such as this is *netCDF* <https://www.unidata.ucar.edu/software/netcdf/> where every file contains metadata describing its contents. Libraries are available to read and write *netCDF* files in many programming languages, including Python, R, MATLAB, C++, and others.

Let's download an example *netCDF* file - this one is from the European Centre for Medium-range Weather Forecasting (ECMWF).

```
import requests
url = "https://www.unidata.ucar.edu/software/netcdf/examples/ECMWF_ERA-40_subset.nc"
filename = url.split("/")[-1]
r = requests.get(url, allow_redirects=True)
with open(filename, "wb") as saved_file:
    saved_file.write(r.content)
```

To read this file in Python we can use the *netCDF4* package:

```
import netCDF4 as nc
ds = nc.Dataset(filename)
ds
```



```
<class 'netCDF4._netCDF4.Dataset'>
root group (NETCDF3_CLASSIC data model, file format NETCDF3):
    Conventions: CF-1.0
    history: 2004-09-15 17:04:29 GMT by mars2netcdf-0.92
    dimensions(sizes): longitude(144), latitude(73), time(62)
    variables(dimensions): float32 longitude(longitude), float32
        latitude(latitude), int32 time(time), int16 tcw(time, latitude, longitude), int16
        tcwv(time, latitude, longitude), int16 lsp(time, latitude, longitude), int16
        cp(time, latitude, longitude), int16 msl(time, latitude, longitude), int16
        blh(time, latitude, longitude), int16 tcc(time, latitude, longitude), int16
        pi0u(time, latitude, longitude), int16 pi0v(time, latitude, longitude), int16
        p2t(time, latitude, longitude), int16 p2d(time, latitude, longitude), int16
        e(time, latitude, longitude), int16 lcc(time, latitude, longitude), int16
        mcc(time, latitude, longitude), int16 hcc(time, latitude, longitude), int16
        tco3(time, latitude, longitude), int16 tp(time, latitude, longitude)
    groups:
```

We can see that the metadata tells us the "dimensions" (lat, long, time), and "variables" (those, plus lots of weather-related things that we could look up on <https://apps.ecmwf.int/codes/grib/param-db/>).

```
for dim in ds.dimensions.values():
    print(dim)
```



```
<class 'netCDF4._netCDF4.Dimension'>: name = 'longitude', size = 144
<class 'netCDF4._netCDF4.Dimension'>: name = 'latitude', size = 73
<class 'netCDF4._netCDF4.Dimension'> (unlimited): name = 'time', size = 62
```

Let's make a map of "total column ozone" (the amount of ozone from the surface of the Earth to the edge of the atmosphere) for the first time point in this file (the times here are "hours since 1/1/1900").

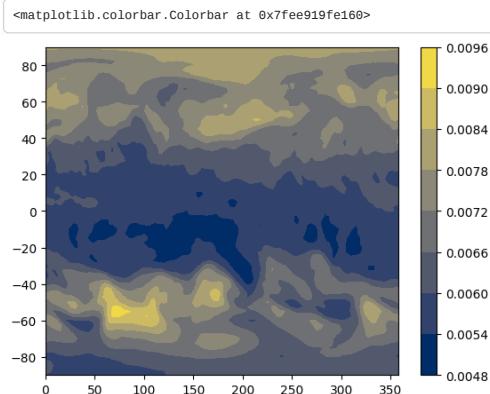
From the ECMWF parameter database (linked above) we can see that the variable we want for the total column ozone is "tco3". We can put the longitude and latitude (which will be our x and y coordinates), and tco3 (which will be the z coordinate)) into numpy data structures:

```
lats = ds.variables["latitude"][:]
lons = ds.variables["longitude"][:]
tco3 = ds.variables["tco3"][:, :, :]
```

A matplotlib contour plot is a simple way of visualizing this.

```
import matplotlib.pyplot as plt
import numpy as np

plt.set_cmap("cividis") # use a CVD-friendly palette
x, y = np.meshgrid(lons, lats)
plt.contourf(x, y, tco3)
plt.colorbar()
```



Pangeo: big data geoscience

NetCDF is a flexible and widely-used format. However, as datasets grow larger, there is increasing demand for tools to process in parallel (as described in the previous notebook), and on the cloud. *Pangeo* (<https://pangeo.io/index.html>) is a community developing a suite of open source packages, all based on Python, that aim to provide interoperability between running on a quick study on a local machine, and running over a huge dataset in the cloud. A major component of this is *Dask*, which we have already seen, and some others are:

- *XArray*: an `xarray.Dataset` is an in-memory representation of a *netCDF* file, while the underlying data structures can either be `numpy` arrays, or *Dask* arrays.
- "Cloud native" file formats, such as *TileDB* and *zarr*, which can both store N-dimensional arrays with intelligent chunking for either local or cloud-based access.

- *Jupyter*: interactive notebooks such as Pangeo are a convenient way for users to interact with computing resources. Ideally, whether the user is running on their local machine, or on a “hub” hosted on the cloud or an HPC cluster, the user experience, and the code, can be almost exactly the same.

10.x.0 (OPTIONAL): Domain specific languages

Estimated time for this notebook: 25 minutes

Lex and Yacc

Let's go back to our nice looks-like LaTeX file format:

```
%>>> %%writefile system.py

class Element:
    def __init__(self, symbol):
        self.symbol = symbol

    def __str__(self):
        return str(self.symbol)

class Molecule:
    def __init__(self):
        self.elements = {} # Map from element to number of that element in the
                           # molecule

    def add_element(self, element, number):
        if not isinstance(element, Element):
            element = Element(element)
        self.elements[element] = number

    @staticmethod
    def as_subscript(number):
        if number == 1:
            return ""
        if number < 10:
            return " " + str(number)
        return "[" + str(number) + "]"

    def __str__(self):
        return "".join([
            str(element) + Molecule.as_subscript(self.elements[element])
            for element in self.elements
        ])

class Side:
    def __init__(self):
        self.molecules = {}

    def add(self, reactant, stoichiometry):
        self.molecules[reactant] = stoichiometry

    @staticmethod
    def print_if_not_one(number):
        if number == 1:
            return ""
        else:
            return str(number)

    def __str__(self):
        return " " + ".join([
            Side.print_if_not_one(self.molecules[molecule]) + str(molecule)
            for molecule in self.molecules
        ])

class Reaction:
    def __init__(self):
        self.reactants = Side()
        self.products = Side()

    def __str__(self):
        return str(self.reactants) + " \\\rightarrow " + str(self.products)

class System:
    def __init__(self):
        self.reactions = []

    def add_reaction(self, reaction):
        self.reactions.append(reaction)

    def __str__(self):
        return "\\\\".join(map(str, self.reactions))
```

Writing system.py

```

from system import Element, Molecule, Reaction, System
s2 = System()
c = Element("C")
o = Element("O")
h = Element("H")

co2 = Molecule()
co2.add_element(c, 1)
co2.add_element(o, 2)

h2o = Molecule()
h2o.add_element(h, 2)
h2o.add_element(o, 1)

o2 = Molecule()
o2.add_element(o, 2)

h2 = Molecule()
h2.add_element(h, 2)

glucose = Molecule()
glucose.add_element(c, 6)
glucose.add_element(h, 12)
glucose.add_element(o, 6)

combustion_glucose = Reaction()
combustion_glucose.reactants.add(glucose, 1)
combustion_glucose.reactants.add(o2, 6)
combustion_glucose.products.add(co2, 6)
combustion_glucose.products.add(h2o, 6)
s2.add_reaction(combustion_glucose)

combustion_hydrogen = Reaction()
combustion_hydrogen.reactants.add(h2, 2)
combustion_hydrogen.reactants.add(o2, 1)
combustion_hydrogen.products.add(h2o, 2)
s2.add_reaction(combustion_hydrogen)

print(s2)

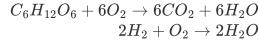
```

C₆H₁₂O₆ + 6O₂ → 6CO₂ + 6H₂O
2H₂ + O₂ → 2H₂O

```

from IPython.display import Math, display
display(Math(str(s2)))

```



How might we write a parser for this? Clearly we'll encounter the problems we previously solved in ensuring each molecule is the and atom only gets one object instance, but we solved this by using an appropriate primary key. (The above implementation is designed to make this easy, learning from the previous lecture.)

But we'll also run into a bunch of problems which are basically string parsing : noting, for example, that `_2` and `_{12}` make a number of atoms in a molecule, or that `+` joins molecules.

This will be very hard to do with straightforward python string processing.

Instead, we will use a tool called `ply` (Python Lex-Yacc) which contains `Lex` (for generating lexical analysers) and `Yacc` (Yet Another Compiler-Compiler). Together these allow us to define the **grammar** of our file format.

The theory of "context free grammars" is rich and deep, and we will just scratch the surface here.

Tokenising with Lex

First, we need to turn our file into a "token stream", using regular expressions to match the kinds of symbol in our source:

```

%%writefile lexreactions.py
from ply import lex

tokens = (
    "ELEMENT",
    "NUMBER",
    "SUBSCRIPT",
    "LBRACE",
    "RBRACE",
    "PLUS",
    "ARROW",
    "NEWLINE",
    "TEXNEWLINE",
)

# Tokens
t_PLUS = r"\+"
t_SUBSCRIPT = r"\_"
t_LBRACE = r"\{"
t_RBRACE = r"\}"
t_TEXNEWLINE = r"\\\\n"
t_ARROW = r"\rightarrow"
t_ELEMENT = r"[A-Z][a-z]?"
t_NUMBER = r"\d+"
t_NEWLINE = r"\n"

def t_NUMBER(t):
    r"\d+"
    t.value = int(t.value)
    return t

t_ignore = " "

def t_error(t):
    print(f"Did not recognise character '{t.value[0]}:{s}' as part of a valid
token")
    t.lexer.skip(1)

# Build the lexer
lexer = lex.lex()

```

```
Writing lexreactions.py
```

```
from lexreactions import lexer

tokens = []
lexer.input(str(s2))
while True:
    tok = lexer.token()
    if not tok:
        break # No more input
    tokens.append(tok)

print(str(s2))
```

```
C_6H_{12}O_6 + 6O_2 \rightarrow CO_2 + 6H_2O
2H_2 + O_2 \rightarrow H_2O
```

```
tokens
```

```
[LexToken(ELEMENT, 'C', 1, 0),
 LexToken(SUBSCRIPT, '_', 1, 1),
 LexToken(NUMBER, 6, 1, 2),
 LexToken(ELEMENT, 'H', 1, 3),
 LexToken(SUBSCRIPT, '_', 1, 4),
 LexToken(LBRACE, '{', 1, 5),
 LexToken(NUMBER, 12, 1, 6),
 LexToken(RBRACE, '}', 1, 8),
 LexToken(ELEMENT, 'O', 1, 9),
 LexToken(SUBSCRIPT, '_', 1, 10),
 LexToken(NUMBER, 6, 1, 11),
 LexToken(PLUS, '+', 1, 13),
 LexToken(NUMBER, 6, 1, 15),
 LexToken(ELEMENT, 'O', 1, 16),
 LexToken(SUBSCRIPT, '_', 1, 17),
 LexToken(NUMBER, 2, 1, 18),
 LexToken(ARROW, '\\rightarrow', 1, 20),
 LexToken(NUMBER, 6, 1, 32),
 LexToken(ELEMENT, 'C', 1, 33),
 LexToken(ELEMENT, 'O', 1, 34),
 LexToken(SUBSCRIPT, '_', 1, 35),
 LexToken(NUMBER, 2, 1, 36),
 LexToken(PLUS, '+', 1, 38),
 LexToken(NUMBER, 6, 1, 40),
 LexToken(ELEMENT, 'H', 1, 41),
 LexToken(SUBSCRIPT, '_', 1, 42),
 LexToken(NUMBER, 2, 1, 43),
 LexToken(ELEMENT, 'O', 1, 44),
 LexToken(TEXNEWLINE, '\\\\', 1, 45),
 LexToken(NEWLINE, '\n', 1, 48),
 LexToken(NUMBER, 2, 1, 49),
 LexToken(ELEMENT, 'H', 1, 50),
 LexToken(SUBSCRIPT, '_', 1, 51),
 LexToken(NUMBER, 2, 1, 52),
 LexToken(PLUS, '+', 1, 54),
 LexToken(ELEMENT, 'O', 1, 56),
 LexToken(SUBSCRIPT, '_', 1, 57),
 LexToken(NUMBER, 2, 1, 58),
 LexToken(ARROW, '\\rightarrow', 1, 60),
 LexToken(NUMBER, 2, 1, 72),
 LexToken(ELEMENT, 'H', 1, 73),
 LexToken(SUBSCRIPT, '_', 1, 74),
 LexToken(NUMBER, 2, 1, 75),
 LexToken(ELEMENT, 'O', 1, 76)]
```

Note that the parser will reject invalid tokens:

```
lexer.input("=2H_2 + O_2 \\leftarrow 2H_2O")
while True:
    tok = lexer.token()
    if not tok:
        break # No more input
    print(tok)
```

```
LexToken(NUMBER, 2, 1, 0)
LexToken(ELEMENT, 'H', 1, 1)
LexToken(SUBSCRIPT, '_', 1, 2)
LexToken(NUMBER, 2, 1, 3)
LexToken(PLUS, '+', 1, 5)
LexToken(ELEMENT, 'O', 1, 7)
LexToken(SUBSCRIPT, '_', 1, 8)
LexToken(NUMBER, 2, 1, 9)
Did not recognise character '=' as part of a valid token
Did not recognise character '2' as part of a valid token
Did not recognise character 'H' as part of a valid token
Did not recognise character '2' as part of a valid token
Did not recognise character 'O' as part of a valid token
Did not recognise character '2' as part of a valid token
Did not recognise character 'H' as part of a valid token
Did not recognise character '2' as part of a valid token
Did not recognise character 'O' as part of a valid token
Did not recognise character '2' as part of a valid token
LexToken(NUMBER, 2, 1, 22)
LexToken(ELEMENT, 'H', 1, 23)
LexToken(SUBSCRIPT, '_', 1, 24)
LexToken(NUMBER, 2, 1, 25)
LexToken(ELEMENT, 'O', 1, 26)
```

Writing a grammar

So, how do we express our algebra for chemical reactions as a grammar?

We write a series of production rules, expressing how a system is made up of equations, an equation is made up of molecules etc:

```
system : equation
system : system TExNEWLINE NEWLINE equation
equation : side ARROW side
side : molecules
molecules : molecule
molecules : NUMBER molecule
side : side PLUS molecules
molecule : countedelement
countedelement : ELEMENT
countedelement : ELEMENT atomcount
molecule : molecule countedelement
atomcount : SUBSCRIPT NUMBER
atomcount : SUBSCRIPT LBRACE NUMBER RBRACE
```

Note how we right that a system is made of more than one equation:

```
system : equation # A system could be one equation
system : system NEWLINE equation # Or it could be a system then an equation
```

... which implies, recursively, that a system could also be:

```
system: equation NEWLINE equation NEWLINE equation ...
```

This is an **incredibly** powerful idea. The full grammar for Python 3 can be defined in only a few hundred lines of specification:

<https://docs.python.org/3/reference/grammar.html>

Parsing with Yacc

A parser defined with Yacc builds up the final object, by breaking down the file according to the rules of the grammar, and then building up objects as the individual tokens coalesce into the full grammar.

Here, we will for clarity not attempt to solve the problem of having multiple molecule instances for the same molecule - the normalisation problem solved last lecture.

In Yacc, each grammar rule is defined by a Python function where the docstring for the function contains the appropriate grammar specification.

Each function accepts an argument `p` that is a list of symbols in the grammar. It must implement the actions of that rule. For example:

```
def p_expression_plus(p):
    'expression : expression PLUS term'
    #           ^          ^
    #   p[0]      p[1]      p[2] p[3]
    p[0] = p[1] + p[3]
```

```

%%writefile parseeactions.py

# Yacc example
import ply.yacc as yacc

# Get the components of our system
from system import Element, Molecule, Side, Reaction, System

# Get the token map from the lexer. This is required.
from lexreactions import tokens


def p_expression_system(p):
    "system : equation"
    p[0] = System()
    p[0].add_reaction(p[1])

def p_expression_combine_system(p):
    "system : system NEWLINE equation"
    p[0] = p[1]
    p[0].add_reaction(p[4])

def p_equation(p):
    "equation : side ARROW side"
    p[0] = Reaction()
    p[0].reactants = p[1]
    p[0].products = p[3]

def p_side(p):
    "side : molecules"
    p[0] = Side()
    p[0].add(p[1][0], p[1][1])

def p_molecules(p):
    "molecules : molecule"
    p[0] = (p[1], 1)

def p_stoichiometry(p):
    "molecules : NUMBER molecule"
    p[0] = (p[2], p[1])

def p_plus(p):
    "side : side PLUS molecules"
    p[0] = p[1]
    p[0].add(p[3][0], p[3][1])

def p_molecule(p):
    "molecule : countedelement"
    p[0] = Molecule()
    p[0].add_element(p[1][0], p[1][1])

def p_countedelement(p):
    "countedelement : ELEMENT"
    p[0] = (p[1], 1)

def p_ncountedelement(p):
    "countedelement : ELEMENT atomcount"
    p[0] = (p[1], p[2])

def p_multi_element(p):
    "molecule : molecule countedelement"
    p[0] = p[1]
    p[0].add_element(p[2][0], p[2][1])

def p_multi_atoms(p):
    "atomcount : SUBSCRIPT NUMBER"
    p[0] = int(p[2])

def p_many_atoms(p):
    "atomcount : SUBSCRIPT LBRACE NUMBER RBRACE"
    p[0] = int(p[3])

# Error rule for syntax errors
def p_error(p):
    print("Syntax error in input!")

# Build the parser
parser = yacc.yacc()

```

Writing parseeactions.py

```

from parseeactions import parser
roundtrip_system = parser.parse(str(s2))

```

Generating LALR tables

```

%%bash
# Read the first 100 lines from the file
head -n 100 parser.out

```

Grammar

Rule 0 S' -> system

Rule 1 system -> equation

Rule 2 system -> system TEXNEWLINE NEWLINE equation

Rule 3 equation -> side ARROW side

Rule 4 side -> molecules

Rule 5 molecules -> molecule

Rule 6 molecules -> NUMBER molecule

Rule 7 side -> side PLUS molecules

Rule 8 molecule -> countedelement

Rule 9 countedelement -> ELEMENT

Rule 10 countedelement -> ELEMENT atomcount

Rule 11 molecule -> molecule countedelement

Rule 12 atomcount -> SUBSCRIPT NUMBER

Rule 13 atomcount -> SUBSCRIPT LBRACE NUMBER RBRACE

Terminals, with rules where they appear

ARROW : 3

ELEMENT : 9 10

LBRACE : 13

NEWLINE : 2

NUMBER : 6 12 13

PLUS : 7

RBRACE : 13

SUBSCRIPT : 12 13

TEXNEWLINE : 2

error :

Nonterminals, with rules where they appear

atomcount : 10

countedelement : 8 11

equation : 1 2

molecule : 5 6 11

molecules : 4 7

side : 3 3 7

system : 2 0

Parsing method: LALR

state 0

(0) S' -> . system

(1) system -> . equation

(2) system -> . system TEXNEWLINE NEWLINE equation

(3) equation -> . side ARROW side

(4) side -> . molecules

(7) side -> . side PLUS molecules

(5) molecules -> . molecule

(6) molecules -> . NUMBER molecule

(8) molecule -> . countedelement

(11) molecule -> . molecule countedelement

(9) countedelement -> . ELEMENT

(10) countedelement -> . ELEMENT atomcount

NUMBER shift and go to state 6

ELEMENT shift and go to state 8

system shift and go to state 1

equation shift and go to state 2

side shift and go to state 3

molecules shift and go to state 4

molecule shift and go to state 5

countedelement shift and go to state 7

state 1

(0) S' -> system .

(2) system -> system . TEXNEWLINE NEWLINE equation

TEXNEWLINE shift and go to state 9

state 2

```

(1) system -> equation .

TEXNEWLINE      reduce using rule 1 (system -> equation .)

$end      reduce using rule 1 (system -> equation .)

state 3

(3) equation -> side . ARROW side

(7) side -> side . PLUS molecules

ARROW      shift and go to state 10

PLUS      shift and go to state 11

state 4

(4) side -> molecules .

ARROW      reduce using rule 4 (side -> molecules .)

PLUS      reduce using rule 4 (side -> molecules .)

{
    display(Math(str(roundtrip_system)))
}

with open("system.tex", "w") as texfile:
    texfile.write(str(roundtrip_system))

!cat system.tex

C_6H_{12}O_6 + 6O_2 → 6CO_2 + 6H_2O
2H_2 + O_2 → 2H_2O

```

Internal DSLs

In doing the above, we have defined what is called an “external DSL”: our code is in Python, but the file format is a language with its own.

However, we can use the language itself to define something almost as fluent, without having to write our own grammar, by using operator overloading and metaprogramming tricks:

```

%%writefile reactionsdsl.py

class Element:
    def __init__(self, symbol):
        self.symbol = symbol

    def __str__(self):
        return str(self.symbol)

    def __mul__(self, other):
        """Let Molecule handle the multiplication"""
        return (self / 1) * other

    def __truediv__(self, number):
        """Element / number => Molecule"""
        res = Molecule()
        res.add_element(self, number)
        return res

class Molecule:
    def __init__(self):
        self.elements = {} # Map from element to number of that element in the molecule

    def add_element(self, element, number):
        if not isinstance(element, Element):
            element = Element(element)
        self.elements[element] = number

    @staticmethod
    def as_subscript(number):
        if number == 1:
            return ""
        if number < 10:
            return " " + str(number)
        return "{" + str(number) + "}"

    def __str__(self):
        return "".join([
            str(element) + Molecule.as_subscript(self.elements[element])
            for element in self.elements
        ])

    def __mul__(self, other):
        """Molecule * Element => Molecule
        Molecule * Molecule => Molecule"""
        if type(other) == Molecule:
            self.elements.update(other.elements)
        else:
            self.add_element(other, 1)
        return self

    def __rmul__(self, stoich):
        """Number * Molecule => Side"""
        res = Side()
        res.add(self, stoich)
        return res

    def __add__(self, other):
        """Molecule + X => Side"""
        if type(other) == Side:
            other.molecules[self] = 1
            return other
        res = Side()
        res.add(self, 1)
        res.add(other, 1)
        return res

class Side:
    def __init__(self):
        self.molecules = {}

    def add(self, reactant, stoichiometry):
        self.molecules[reactant] = stoichiometry

    @staticmethod
    def print_if_not_one(number):
        if number == 1:
            return ""
        else:
            return str(number)

    def __str__(self):
        return " " + ".join([
            Side.print_if_not_one(self.molecules[molecule]) + str(molecule)
            for molecule in self.molecules
        ])

    def __add__(self, other):
        """Side + X => Side"""
        self.molecules.update(other.molecules)
        return self

    def __eq__(self, other):
        res = Reaction()
        res.reactants = self
        res.products = other
        current_system.add_reaction(res) # Closure!
        return f"Added: '{res}'"

class Reaction:
    def __init__(self):
        self.reactants = Side()
        self.products = Side()

    def __str__(self):
        return str(self.reactants) + "\\\rightarrow " + str(self.products)

class System:
    def __init__(self):
        self.reactions = []

    def add_reaction(self, reaction):
        self.reactions.append(reaction)

    def __str__(self):
        return "\\\n".join(map(str, self.reactions))

current_system = System()

```

```

Writing reactionsdsl.py

from reactionsdsl import Element, current_system

# Here we add new symbols to the global scope
# This is *not* good practice, we do it here to demonstrate that it is possible to
do
for symbol in ("C", "O", "H"):
    globals()[symbol] = Element(symbol)

o / 2 + 2 * (H / 2) == 2 * (H / 2 * O)

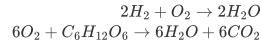
"Added: '2H_2 + O_2 \rightarrow 2H_2O'"

(C / 6) * (H / 12) * (O / 6) + 6 * (O / 2) == 6 * (H / 2 * O) + 6 * (C * (O / 2))

"Added: '6O_2 + C_6H_{12}O_6 \rightarrow 6H_2O + 6CO_2'"

display(Math(str(current_system)))

```



Python is not perfect for this, because it lacks the idea of parenthesis-free function dispatch and other things that make internal DSLs pretty.

10.x.1 (OPTIONAL): Controlled Vocabularies

Estimated time for this notebook: 15 minutes

Saying the same thing in multiple ways

What happens when someone comes across a file in our file format? How do they know what it means?

If we can make the tag names in our model globally unique, then the meaning of the file can be made understandable not just to us, but to people and computers all over the world.

Two file formats which give the same information, in different ways, are *syntactically* distinct, but so long as they are *semantically* compatible, I can convert from one to the other.

This is the goal of the technologies introduced this lecture.

The URI

The key concept that underpins these tools is the URI: uniform resource **indicator**.

These look like URLs:

www.turing.ac.uk/rsd-engineering/schema/reaction/element

But, if I load that as a web address, there's nothing there!

That's fine.

A URN indicates a **name** for an entity, and, by using organisational web addresses as a prefix, is likely to be unambiguously unique.

A URI might be a URL or a URN, or both.

XML Namespaces

It's cumbersome to use a full URI every time we want to put a tag in our XML file. XML defines *namespaces* to resolve this:

```

%%writefile system.xml
<?xml version="1.0" encoding="UTF-8"?>
<system xmlns="http://www.turing.ac.uk/rsd-engineering/schema/reaction">
    <reaction>
        <reactants>
            <molecule stoichiometry="2">
                <atom symbol="H" number="2"/>
            </molecule>
            <molecule stoichiometry="1">
                <atom symbol="O" number="2"/>
            </molecule>
        </reactants>
        <products>
            <molecule stoichiometry="2">
                <atom symbol="H" number="2"/>
                <atom symbol="O" number="1"/>
            </molecule>
        </products>
    </reaction>
</system>

```

Writing system.xml

```

from lxml import etree
with open("system.xml") as xmlfile:
    tree = etree.parse(xmlfile)

print(etree.tostring(tree, pretty_print=True, encoding=str))

```

```

<system xmlns="http://www.turing.ac.uk/rsd-engineering/schema/reaction">
  <reaction>
    <reactants>
      <molecule stoichiometry="2">
        <atom symbol="H" number="2"/>
      </molecule>
      <molecule stoichiometry="1">
        <atom symbol="O" number="2"/>
      </molecule>
    </reactants>
    <products>
      <molecule stoichiometry="2">
        <atom symbol="H" number="2"/>
        <atom symbol="O" number="1"/>
      </molecule>
    </products>
  </reaction>
</system>

```

Note that our previous XPath query no longer finds anything.

```

tree.xpath("//molecule/atom[@number='1']/@symbol")
[]

namespaces = {"r": "http://www.turing.ac.uk/rsd-engineering/schema/reaction"}

tree.xpath("//r:molecule/r:atom[@number='1']/@symbol", namespaces=namespaces)
['O']

```

Note the prefix `r` used to bind the namespace in the query: any string will do - it's just a dummy variable.

The above file specified our namespace as a default namespace: this is like doing `from numpy import *` in python.

It's often better to bind the namespace to a prefix:

```

%%writefile system.xml
<?xml version="1.0" encoding="UTF-8"?>
<:system xmlns:r="http://www.turing.ac.uk/rsd-engineering/schema/reaction">
  <:reaction>
    <r:reactants>
      <r:molecule stoichiometry="2">
        <r:atom symbol="H" number="2"/>
      </r:molecule>
      <r:molecule stoichiometry="1">
        <r:atom symbol="O" number="2"/>
      </r:molecule>
    </r:reactants>
    <r:products>
      <r:molecule stoichiometry="2">
        <r:atom symbol="H" number="2"/>
        <r:atom symbol="O" number="1"/>
      </r:molecule>
    </r:products>
  </r:reaction>
</r:system>

```

Overwriting system.xml

Namespaces and Schema

It's a good idea to serve the schema itself from the URI of the namespace treated as a URL, but it's *not a requirement*: it's a URN not necessarily a URL!

```

%%writefile reactions.xsd
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
    targetNamespace="http://www.turing.ac.uk/rsd-engineering/schema/reaction"
    xmlns:r="http://www.turing.ac.uk/rsd-engineering/schema/reaction">

<xs:element name="atom">
    <xs:complexType>
        <xs:attribute name="symbol" type="xs:string"/>
        <xs:attribute name="number" type="xs:integer"/>
    </xs:complexType>
</xs:element>

<xs:element name="molecule">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="r:atom" maxOccurs="unbounded"/>
        </xs:sequence>
        <xs:attribute name="stoichiometry" type="xs:integer"/>
    </xs:complexType>
</xs:element>

<xs:element name="reactants">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="r:molecule" maxOccurs="unbounded"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>

<xs:element name="products">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="r:molecule" maxOccurs="unbounded"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>

<xs:element name="reaction">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="r:reactants"/>
            <xs:element ref="r:products"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>

<xs:element name="system">
    <xs:complexType>
        <xs:sequence>
            <xs:element ref="r:reaction" maxOccurs="unbounded"/>
        </xs:sequence>
    </xs:complexType>
</xs:element>

</xs:schema>

```

Writing reactions.xsd

Note we're now defining the target namespace for our schema.

```

with open("reactions.xsd") as xsdfile:
    schema_xsd = xsdfile.read()
schema = etree.XMLSchema(etree.XML(schema_xsd))

parser = etree.XMLParser(schema=schema)

with open("system.xml") as xmlfile:
    tree = etree.parse(xmlfile, parser)
    print(tree)

<xml.etree._ElementTree object at 0x7fid783d8f00>

```

Note the power of binding namespaces when using XML files addressing more than one namespace. Here, we can clearly see which variables are part of the schema defining XML schema itself (bound to `xs`) and the schema for our file format (bound to `r`)

Using standard vocabularies

The work we've done so far will enable someone who comes across our file format to track down something about its significance, by following the URI in the namespace. But it's still somewhat ambiguous. The word "element" means (at least) two things: an element tag in an XML document, and a chemical element. (It also means a heating element in a toaster, and lots of other things.)

To make it easier to not make mistakes as to the meaning of **found data**, it is helpful to use standardised namespaces that already exist for the concepts our file format refers to.

So that when somebody else picks up one of our data files, the meaning of the stuff it describes is obvious. In this example, it would be hard to get it wrong, of course, but in general, defining file formats so that they are meaningful as found data should be desirable.

For example, the concepts in our file format are already part of the "DBpedia ontology", among others. So, we could redesign our file format to exploit this, by referencing for example <https://dbpedia.org/ontology/ChemicalCompound>:

```

%%writefile chemistry_template3.mko
<?xml version="1.0" encoding="UTF-8"?>
<system xmlns="https://www.turing.ac.uk/rsd-engineering/schema/reaction"
         xmlns:dbo="https://dbpedia.org/ontology/">
%for reaction in reactions:
<reaction>
    <reactants>
        %for molecule in reaction.reactants.molecules:
            <dbo:ChemicalCompound
                stoichiometry="${reaction.reactants.molecules[molecule]}"
                %for element in molecule.elements:
                    <dbo:ChemicalElement symbol="${element.symbol}"
                                         number="${molecule.elements[element]}"/>
                %endfor
            </dbo:ChemicalCompound>
        %endfor
        </reactants>
        <products>
            %for molecule in reaction.products.molecules:
                <dbo:ChemicalCompound
                    stoichiometry="${reaction.products.molecules[molecule]}"
                    %for element in molecule.elements:
                        <dbo:ChemicalElement symbol="${element.symbol}"
                                             number="${molecule.elements[element]}"/>
                    %endfor
                </dbo:ChemicalCompound>
            %endfor
        </products>
    </reaction>
%endfor
</system>

```

Writing chemistry_template3.mko

However, this won't work properly, because it's not up to us to define the XML schema for somebody else's entity type: and an XML schema can only target one target namespace.

Of course we should use somebody else's file format for chemical reaction networks: compare [SBML](#) for example. We already know not to reinvent the wheel - and this whole lecture series is just reinventing the wheel for pedagogical purposes. But what if we've already got a bunch of data in our own format. How can we lock down the meaning of our terms?

So, we instead need to declare that our `r:element` represents the same concept as `dbo:ChemicalElement`. To do this formally we will need the concepts from the next lecture, specifically `rdf:sameAs`, but first, let's understand the idea of an ontology.

Taxonomies and ontologies

An Ontology (in computer science terms) is two things: a **controlled vocabulary** of entities (a set of URIs in a namespace), the definitions thereof, and the relationships between them.

People often casually use the word to mean any formalised taxonomy, but the relation of terms in the ontology to the concepts they represent, and the relationships between them, are also critical.

Have a look at another example: <https://dublincore.org/documents/dcmi-terms/>

Note each concept is a URI, but some of these are also stated to be subclasses or superclasses of the others.

Some are properties of other things, and the domain and range of these verbs are also stated.

Why is this useful for us in discussing file formats?

One of the goals of the **semantic web** is to create a way to make file formats which are universally meaningful as found data: if I have a file format defined using any formalised ontology, then by tracing statements through `rdf:sameAs` relationships, I should be able to reconstruct the information I need.

That will be the goal of the next lecture.

10.x.2 (OPTIONAL): Semantic file formats

Estimated time for this notebook: 25 minutes

The dream of a semantic web

So how can we fulfill the dream of a file-format which is **self-documenting**: universally unambiguous and interpretable?

(Of course, it might not be true, but we don't have capacity to discuss how to model reliability and contested testimony.)

By using URIs to define a controlled vocabulary, we can be unambiguous.

But the number of different concepts to be labelled is huge: so we need a **distributed** solution: a global structure of people defining ontologies, (with methods for resolving duplications and inconsistencies.)

Humanity has a technology that can do this: the world wide web. We've seen how many different actors are defining ontologies.

We also need a shared semantic structure for our file formats. XML allows everyone to define their own schema. Our universal file format requires a restriction to a basic language, which allows us to say the things we need:

The Triple

We can then use these defined terms to specify facts, using a URI for the subject, verb, and object of our sentence.

```

%%writefile reaction.ttl
<http://dbpedia.org/ontology/water>
  <http://purl.obolibrary.org/obo/PATO_0001681>
    "18.01528"^^<http://purl.obolibrary.org/obo/UO_0000088>

```

Writing reaction.ttl

- [Water](#)
- [Molar mass](#)
- [Grams per mole](#)

This is an unambiguous statement, consisting of a subject, a verb, and an object, each of which is either a URI or a literal value. Here, the object is a *literal* with a type.

RDF file formats

We have used the RDF (Resource Description Framework) **semantic** format, in its "Turtle" syntactic form:

```
subject verb object .  
subject2 verb2 object2 .
```

We can parse it:

```
from rdflib import Graph  
  
graph = Graph()  
graph.parse("reaction.ttl", format="ttl")  
  
print(len(graph))  
  
for statement in graph:  
    print(statement)  
  
1  
(rdflib.term.URIRef('http://dbpedia.org/ontology/water'),  
 rdflib.term.URIRef('http://purl.obolibrary.org/obo/PATO_0001681'),  
 rdflib.term.Literal('18.01528',  
 datatype=rdflib.term.URIRef('http://purl.obolibrary.org/obo/UO_0000088'))
```

The equivalent in **RDF-XML** is:

```
print(graph.serialize(format="xml"))  
  
<?xml version="1.0" encoding="utf-8"?>  
<rdf:RDF  
    xmlns:ns1="http://purl.obolibrary.org/obo/"  
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"  
>  
    <rdf:Description rdf:about="http://dbpedia.org/ontology/water">  
        <ns1:PATO_0001681  
        rdf:datatype="http://purl.obolibrary.org/obo/UO_0000088">18.01528</ns1:PATO_000168  
    1>  
    </rdf:Description>  
</rdf:RDF>
```

We can also use namespace prefixes in Turtle:

```
print(graph.serialize(format="ttl"))  
  
@prefix ns1: <http://purl.obolibrary.org/obo/> .  
<http://dbpedia.org/ontology/water> ns1:PATO_0001681 "18.01528"^^ns1:UO_0000088 .
```

Normal forms and Triples

How do we encode the sentence "water has two hydrogen atoms" in RDF?

See [Defining N-ary Relations on the Semantic Web](#) for the definitive story.

I'm not going to search carefully here for existing ontologies for the relationships we need: later we will understand how to define these as being the same as or subclasses of concepts in other ontologies. That's part of the value of a distributed approach: we can define what we need, and because the Semantic Web tools make rigorous the concepts of `rdfs:sameAs` and `rdfs:subClassOf` this will be OK.

However, there's a problem. We can do:

```
%>writetofile reaction.ttl  
  
@prefix disr: <http://www.turing.ac.uk/rsd-engineering/ontologies/reactions/> .  
@prefix dbo: <http://dbpedia.org/ontology/> .  
@prefix obo: <http://purl.obolibrary.org/obo/> .  
  
dbo:water obo:PATO_0001681 "18.01528"^^obo:UO_0000088 ;  
    disr:containsElement obo:CHEBI_33260 .  
  
Overwriting reaction.ttl
```

- [ElementalHydrogen](#)

We've introduced the semicolon in Turtle to say two statements about the same entity. The equivalent RDF-XML is:

```
graph = Graph()  
graph.parse("reaction.ttl", format="ttl")  
print(len(graph))  
print(graph.serialize(format="xml"))  
  
2  
<?xml version="1.0" encoding="utf-8"?>  
<rdf:RDF  
    xmlns:disr="http://www.turing.ac.uk/rsd-engineering/ontologies/reactions/"  
    xmlns:obo="http://purl.obolibrary.org/obo/"  
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"  
>  
    <rdf:Description rdf:about="http://dbpedia.org/ontology/water">  
        <obo:PATO_0001681  
        rdf:datatype="http://purl.obolibrary.org/obo/UO_0000088">18.01528</obo:PATO_000168  
    1>  
        <disr:containsElement  
        rdf:resource="http://purl.obolibrary.org/obo/CHEBI_33260"/>  
    </rdf:Description>  
</rdf:RDF>
```

However, we can't express `hasTwo` in this way without making an infinite number of properties!

RDF doesn't have a concept of adverbs. Why not?

It turns out there's a fundamental relationship between the RDF triple and a RELATION in the relational database model.

- The **subject** corresponds to the relational primary key.
- The **verb** (RDF "property") corresponds to the relational column name.

- The **object** corresponds to the value in the corresponding column.

We already found out that to model the relationship of atoms to molecules we needed a join table, and the number of atoms was metadata on the join.

So, we need an entity type (RDF **class**) which describes an [ElementInMolecule](#).

Fortunately, we don't have to create a universal URI for every single relationship, thanks to RDF's concept of an anonymous entity: something which is uniquely defined by its relationships.

Imagine if we had to make a URN for oxygen-in-water, hydrogen-in-water etc!

```
%>writefile reaction.ttl
@prefix disr: <http://www.turing.ac.uk/rsd-engineering/ontologies/reactions/> .
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix obo: <http://purl.obolibrary.org/obo/> .
@prefix xs: <http://www.w3.org/2001/XMLSchema> .

dbo:water obo:PATO_0001681 "18.01528"^^obo:UO_0000088 ;
    disr:containsElement obo:CHEBI_33260 ;
    disr:hasElementQuantity [
        disr:countedElement obo:CHEBI_33260 ;
        disr:countOfElement "2"^^xs:integer
    ] .
```

Overwriting reaction.ttl

Here we have used [] to indicate an anonymous entity, with no subject. We then define two predicates on that subject, using properties corresponding to our column names in the join table.

Another turtle syntax for an anonymous "blank node" is this:

```
%>writefile reaction.ttl
@prefix disr: <http://www.turing.ac.uk/rsd-engineering/ontologies/reactions/> .
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix obo: <http://purl.obolibrary.org/obo/> .
@prefix xs: <http://www.w3.org/2001/XMLSchema> .

dbo:water obo:PATO_0001681 "18.01528"^^obo:UO_0000088 ;
    disr:containsElement obo:CHEBI_33260 ;
    disr:hasElementQuantity _:a .

_:a disr:countedElement obo:CHEBI_33260 ;
    disr:countOfElement "2"^^xs:integer .
```

Overwriting reaction.ttl

Serialising to RDF

Here's code to write our model to Turtle:

```
%>writefile chemistry_turtle_template.mko
@prefix disr: <http://www.turing.ac.uk/rsd-engineering/ontologies/reactions/> .
@prefix obo: <http://purl.obolibrary.org/obo/> .
@prefix xs: <http://www.w3.org/2001/XMLSchema> .

[
%for reaction in reactions:
    disr:hasReaction [
        %for molecule in reaction.reactants.molecules:
            disr:hasReactant [
                %for element in molecule.elements:
                    disr:hasElementQuantity [
                        disr:countedElement [
                            a obo:CHEBI_33259;
                            disr:symbol "${element.symbol}"^^xs:string
                        ] ;
                        disr:countOfElement
                            "${molecule.elements[element]}"^^xs:integer
                        ];
                    %endfor
                    a obo:CHEBI_23367
                ] ;
            %endfor
            %for molecule in reaction.products.molecules:
                disr:hasProduct [
                    %for element in molecule.elements:
                        disr:hasElementQuantity [
                            disr:countedElement [
                                a obo:CHEBI_33259;
                                disr:symbol "${element.symbol}"^^xs:string
                            ] ;
                            disr:countOfElement
                                "${molecule.elements[element]}"^^xs:integer
                            ];
                        %endfor
                        a obo:CHEBI_23367
                    ] ;
                %endfor
                a disr:reaction
            ] ;
        %endfor
        a disr:system
    ].
```

Writing chemistry_turtle_template.mko

"a" in Turtle is an always available abbreviation for <https://www.w3.org/1999/02/22-rdf-syntax-ns#type>

We've also used:

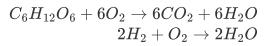
- [Molecular entity](#)
- [Elemental molecular entity](#)

I've skipped serialising the stoichiometries: to do that correctly I also need to create a relationship class for molecule-in-reaction.

And we've not attempted to relate our elements to their formal definitions, since our model isn't recording this at the moment. We could add this statement later.

```
from IPython.display import Math, display
from parseractions import parser

with open("system.tex", "r") as texfile:
    system = parser.parse(texfile.read())
display(Math(str(system)))
```



```
from mako.template import Template

mytemplate = Template(filename="chemistry_turtle_template.mko")
with open("system.ttl", "w") as ttlfile:
    ttlfile.write((mytemplate.render(**vars(system))))
```

```
!cat system.ttl
```

```

@prefix disr: <http://www.turing.ac.uk/rsd-engineering/ontologies/reactions/> .
@prefix obo: <http://purl.obolibrary.org/obo/> .
@prefix xs: <http://www.w3.org/2001/XMLSchema> .

[ 
    disr:hasReaction [
        disr:hasReactant [
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "C"^^xs:string
                ] ;
                disr:countOfElement "6"^^xs:integer
            ];
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "H"^^xs:string
                ] ;
                disr:countOfElement "12"^^xs:integer
            ];
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "O"^^xs:string
                ] ;
                disr:countOfElement "6"^^xs:integer
            ];
            a obo:CHEBI_23367
        ];
        disr:hasReactant [
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "O"^^xs:string
                ] ;
                disr:countOfElement "2"^^xs:integer
            ];
            a obo:CHEBI_23367
        ];
        disr:hasProduct [
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "C"^^xs:string
                ] ;
                disr:countOfElement "1"^^xs:integer
            ];
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "O"^^xs:string
                ] ;
                disr:countOfElement "2"^^xs:integer
            ];
            a obo:CHEBI_23367
        ];
        disr:hasProduct [
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "H"^^xs:string
                ] ;
                disr:countOfElement "2"^^xs:integer
            ];
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "O"^^xs:string
                ] ;
                disr:countOfElement "1"^^xs:integer
            ];
            a obo:CHEBI_23367
        ];
        a disr:reaction
    ];
    disr:hasReaction [
        disr:hasReactant [
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "H"^^xs:string
                ] ;
                disr:countOfElement "2"^^xs:integer
            ];
            a obo:CHEBI_23367
        ];
        disr:hasReactant [
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "O"^^xs:string
                ] ;
                disr:countOfElement "2"^^xs:integer
            ];
            a obo:CHEBI_23367
        ];
        disr:hasProduct [
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "H"^^xs:string
                ] ;
                disr:countOfElement "2"^^xs:integer
            ];
            disr:hasElementQuantity [
                disr:countedElement [
                    a obo:CHEBI_33259;
                    disr:symbol "O"^^xs:string
                ] ;
                disr:countOfElement "1"^^xs:integer
            ];
            a obo:CHEBI_23367
        ];
        a disr:reaction
    ];
    a disr:system
].

```

```

graph = Graph()
graph.parse("system ttl", format="ttl")
print(graph.serialize(format="xml"))

```

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
  xmlns:disr="http://www.turing.ac.uk/rsd-engineering/ontologies/reactions/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
```



```

<disr:countOfElement
rdf:datatype="http://www.w3.org/2001/XMLSchemainteger">2</disr:countOfElement>
</rdf:Description>
<rdf:Description rdf:nodeID="nd52efd85048e46e0836c600de9ece429b16">
<disr:countedElement rdf:nodeID="nd52efd85048e46e0836c600de9ece429b17"/>
<disr:countOfElement
rdf:datatype="http://www.w3.org/2001/XMLSchemainteger">2</disr:countOfElement>
</rdf:Description>
<rdf:Description rdf:nodeID="nd52efd85048e46e0836c600de9ece429b7">
<rdf:type rdf:resource="http://purl.obolibrary.org/obo/CHEBI_33259"/>
<disr:symbol
rdf:datatype="http://www.w3.org/2001/XMLSchemastring">H</disr:symbol>
</rdf:Description>
<rdf:Description rdf:nodeID="nd52efd85048e46e0836c600de9ece429b4">
<disr:countedElement rdf:nodeID="nd52efd85048e46e0836c600de9ece429b5"/>
<disr:countOfElement
rdf:datatype="http://www.w3.org/2001/XMLSchemainteger">6</disr:countOfElement>
</rdf:Description>
<rdf:Description rdf:nodeID="nd52efd85048e46e0836c600de9ece429b28">
<disr:countedElement rdf:nodeID="nd52efd85048e46e0836c600de9ece429b29"/>
<disr:countOfElement
rdf:datatype="http://www.w3.org/2001/XMLSchemainteger">2</disr:countOfElement>
</rdf:Description>
<rdf:Description rdf:nodeID="nd52efd85048e46e0836c600de9ece429b34">
<rdf:type rdf:resource="http://purl.obolibrary.org/obo/CHEBI_33259"/>
<disr:symbol
rdf:datatype="http://www.w3.org/2001/XMLSchemastring"></disr:symbol>
</rdf:Description>
<rdf:Description rdf:nodeID="nd52efd85048e46e0836c600de9ece429b19">
<disr:countedElement rdf:nodeID="nd52efd85048e46e0836c600de9ece429b20"/>
<disr:countOfElement
rdf:datatype="http://www.w3.org/2001/XMLSchemainteger">2</disr:countOfElement>
</rdf:Description>
<rdf:Description rdf:nodeID="nd52efd85048e46e0836c600de9ece429b11">
<disr:countedElement rdf:nodeID="nd52efd85048e46e0836c600de9ece429b12"/>
<disr:countOfElement
rdf:datatype="http://www.w3.org/2001/XMLSchemainteger">2</disr:countOfElement>
</rdf:Description>
<rdf:Description rdf:nodeID="nd52efd85048e46e0836c600de9ece429b22">
<rdf:type rdf:resource="http://purl.obolibrary.org/obo/CHEBI_33259"/>
<disr:symbol
rdf:datatype="http://www.w3.org/2001/XMLSchemastring">0</disr:symbol>
</rdf:Description>
</rdf:RDF>
```

We can see why the group of triples is called a *graph*: each node is an entity and each arc a property relating entities.

Note that this format is very very verbose. It is **not** designed to be a nice human-readable format.

Instead, the purpose is to maximise the capability of machines to reason with found data.

Formalising our ontology: RDFS

Our <http://www.turing.ac.uk/rsd-engineering/ontologies/reactions/> namespace now contains the following properties:

- disr:hasReaction
- disr:hasReactant
- disr:hasProduct
- disr:containsElement
- disr:countedElement
- disr:hasElementQuantity
- disr:countOfElement
- disr:symbol

And two classes:

- disr:system
- disr:reaction

We would now like to find a way to formally specify some of the relationships between these.

The **type** (<http://www.w3.org/1999/02/22-rdf-syntax-ns#type> or [a](#)) of the subject of hasReaction must be **disr:system**.

RDFS will allow us to specify which URNs define classes and which properties, and the domain and range (valid subjects and objects) of our properties.

For example:

```

%%writefile turing_ontology.ttl

@prefix disr: <http://www.turing.ac.uk/rsd-engineering/ontologies/reactions/> .
@prefix obo: <http://purl.obolibrary.org/obo/> .
@prefix xs: <http://www.w3.org/2001/XMLSchema> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

disr:system a rdfs:Class .
disr:reaction a rdfs:Class .
disr:hasReaction a rdf:Property .
disr:hasReaction rdfs:domain disr:system .
disr:hasReaction rdfs:range disr:reaction .
```

Writing turing_ontology.ttl

This will allow us to make our file format briefer: given this schema, if

`_:a hasReaction _:b`

then we can **infer** that

`_:a a disr:system . _:b a disr:reaction .`

without explicitly stating it.

Obviously there's a lot more to do to define our other classes, including defining a class for our anonymous element-in-molecule nodes.

This can get very interesting:

```

%%writefile turing_ontology.ttl
@prefix disr: <http://www.turing.ac.uk/rsd-engineering/ontologies/reactions/> .
@prefix obo: <http://purl.obolibrary.org/obo/ .
@prefix xs: <http://www.w3.org/2001/XMLSchema> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

disr:system a rdfs:Class .
disr:reaction a rdfs:Class .
disr:hasReaction a rdf:Property .
disr:hasReaction rdfs:domain disr:system .
disr:hasReaction rdfs:range disr:reaction .

disr:hasParticipant a rdf:Property .
disr:hasReactant rdfs:subPropertyOf disr:hasParticipant .
disr:hasProduct rdfs:subPropertyOf disr:hasParticipant

```

Overwriting turing_ontology.ttl

[OWL](#) extends RDFS even further.

Inferring additional rules from existing rules and schema is very powerful: an interesting branch of AI. (Unfortunately the [python tool](#) for doing this automatically is currently not updated to python 3 so I'm not going to demo it. Instead, we'll see in a moment how to apply inferences to our graph to introduce new properties.)

SPARQL

So, once I've got a bunch of triples, how do I learn anything at all from them? The language is so verbose it seems useless!

SPARQL is a very powerful language for asking questions of knowledge bases defined in RDF triples:

```

results = graph.query(
    """
    SELECT DISTINCT ?asymbol ?bsymbol
    WHERE {
        ?molecule disr:hasElementQuantity ?a .
        ?disr:countedElement ?elements .
        ?elements disr:symbol ?asymbol .
        ?molecule disr:hasElementQuantity ?b .
        ?b disr:countedElement ?elements .
        ?elements disr:symbol ?bsymbol
    }
)
for row in results:
    print(f"Elements {row[0]} and {row[1]} are found in the same molecule" % row)

```

```

TypeError                                     Traceback (most recent call last)
Cell In [16], line 16
      1 results = graph.query(
      2     """
      3     SELECT DISTINCT ?asymbol ?bsymbol
      4     WHERE {
      5         ?molecule disr:hasElementQuantity ?a .
      6         ?disr:countedElement ?elements .
      7         ?elements disr:symbol ?asymbol .
      8         ?molecule disr:hasElementQuantity ?b .
      9         ?b disr:countedElement ?elements .
     10         ?elements disr:symbol ?bsymbol
     11     }
     12 )
     13 )
     14 for row in results:
--> 15     print(f"Elements {row[0]} and {row[1]} are found in the same molecule" %
     16     row)
     17
     18 TypeError: not all arguments converted during string formatting

```

We can see how this works: you make a number of statements in triple-form, but with some quantities as dummy-variables. SPARQL finds all possible subgraphs of the triple graph which are compatible with the statements in your query.

We can also use SPARQL to specify [inference rules](#):

```

graph.update(
    """
    INSERT { ?elementsA disr:inMoleculeWith ?elementsB }
    WHERE {
        ?molecule disr:hasElementQuantity ?a .
        ?a disr:countedElement ?elementsA .
        ?elementsA disr:symbol ?asymbol .
        ?molecule disr:hasElementQuantity ?b .
        ?b disr:countedElement ?elementsB .
        ?elementsB disr:symbol ?bsymbol
    }
)
graph.query(
    """
    SELECT DISTINCT ?asymbol ?bsymbol
    WHERE {
        ?elementsA disr:inMoleculeWith ?elementsB .
        ?elementsA disr:symbol ?asymbol .
        ?elementsB disr:symbol ?bsymbol
    }
)
for row in results:
    print(f"Elements {row[0]} and {row[1]} are found in the same molecule")

```

```

Elements C and C are found in the same molecule
Elements C and H are found in the same molecule
Elements C and O are found in the same molecule
Elements H and C are found in the same molecule
Elements H and H are found in the same molecule
Elements H and O are found in the same molecule
Elements O and C are found in the same molecule
Elements O and H are found in the same molecule
Elements O and O are found in the same molecule

```

Exercise for reader: express "If x is the subject of a hasReaction relationship, then x must be a system" in SPARQL.

Exercise for reader: search for a SPARQL endpoint knowledge base in your domain.

Connect to it using [Python RDFLib's SPARQL endpoint wrapper](#) and ask it a question.

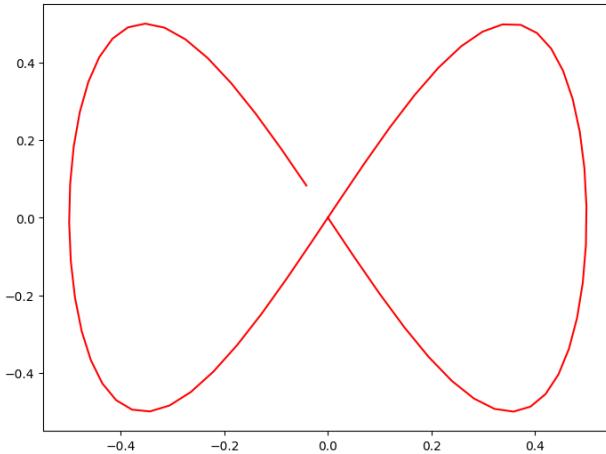
Exercise Solutions

We've provided sample solutions for the exercises in each module. They're sample solutions because a lot of the exercises could be implemented in different ways and don't have a single correct answer. It's up to you how you use the solutions, but it's always best to attempt the exercises yourself first.

Module 1

Exercise 1a

```
import draw_infinity  
image = draw_infinity.make_figure()
```



Exercise 1b

```
#What is 2 to the power 15?
import math as m

print(2**15)
print(m.pow(2,15))
print("-----")

#Convert `It was the best of times` to uppercase.
target = "It was the best of times"
print(target.upper())
print("It was the best of times".upper())
print("-----")

#Sort the list [10, 9, 0, 20, 8, 2, 30, 7, 3].
target = [10, 9, 0, 20, 8, 2, 30, 7, 3]
print(sorted(target)) # Returns a new list that is sorted
target.sort() # N/B .sort() modifies the original list
print(target)
print("-----")

#What is 100! ? (That is, what is the factorial of 100?) Hint: the `factorial`
function is in the `math` library m
print(m.factorial(100))

# Could do it my hand too but there are functions to do it in the math (and other)
libraries
answer = 1
for i in range(1, 100):
    answer *= i

print(answer)
```

```
32768
32768.0
-----
IT WAS THE BEST OF TIMES
IT WAS THE BEST OF TIMES
-----
[0, 2, 3, 7, 8, 9, 10, 20, 30]
[0, 2, 3, 7, 8, 9, 10, 20, 30]
-----
93326215443944152681692388562670049071596826438162146859293689521759999322991560
93326215443944152681692388562670049071596826438162146859293689521759999322991560
93326146397615651828625369792827237582511852109168640000000000000000000000000000000000
9332614639761565182862536979282723758251185210916864000000000000000000000000000000000
```

A note about `sorted` and `sort`:

sorted(target)

returns a new list that is sorted

`target.sort()`

modifies the original list. If we look at their positions in memory we can verify this:

```

example_list = [3, 8, 1, 0, 5, 8, 9, 1, 1, 5]
print(f"Example list = {example_list}")
print(hex(id(example_list))) # Where the example list is stored
print("")
new_list = sorted(example_list)
print(f"New list      = {new_list}")
print(hex(id(new_list))) # Where the new list is stored
print(f"Example list = {example_list}")
print(hex(id(example_list))) # Where the example list is stored
print("")
example_list.sort()
print(f"Example list = {example_list}")
print(hex(id(example_list))) # Where the (sorted) example list is stored

```

```

Example list = [3, 8, 1, 0, 5, 8, 9, 1, 1, 5]
0x7f2b18ba9900

New list      = [0, 1, 1, 1, 3, 5, 5, 8, 8, 9]
0x7f2b18ba9b80
Example list = [3, 8, 1, 0, 5, 8, 9, 1, 1, 5]
0x7f2b18ba9900

Example list = [0, 1, 1, 1, 3, 5, 5, 8, 8, 9]
0x7f2b18ba9900

```

We can see that the example list is in the same place as it was before, but now it is sorted

Exercise 1c

```

# Which of the operators `+`, `-`, `*`, and `/` do something useful with the lists
#[1, 10, 100] and [5, 4, 7]?
a = [1, 10, 100]
b = [5, 4, 7]
print(a+b)
# all others not allowed
print("")

# What happens if you apply the operators `+`, `-`, `*`, `/` to a list and a
# number?
c = [1, 2, 3, 4, 'five']
d = 2
print(c*d)
# all others not allowed
print("")

# What about a string and a string?
e = "string-1"
f = "string-2"
print(e + f)
# all others not allowed

```

```

[1, 10, 100, 5, 4, 7]
[1, 2, 3, 4, 'five', 1, 2, 3, 4, 'five']

string-istring-2

```

Exercise 1d

Something with a similar structure to this:

```

house = {
    "living": {
        "exits": {"north": "kitchen", "outside": "garden", "upstairs": "bedroom"},
        "people": ["James"],
        "capacity": 2,
    },
    "kitchen": {"exits": {"south": "living"}, "people": [], "capacity": 1},
    "garden": {"exits": {"inside": "living"}, "people": ["Sue"], "capacity": 3},
    "bedroom": {
        "exits": {"downstairs": "living", "jump": "garden"},
        "people": [],
        "capacity": 1,
    },
}

```

Some important points about this particular solution:

- The whole solution is a single nested structure.
- Indentation is used to make the structure easier to read.
- Python allows code to continue over multiple lines, so long as sets of brackets are not finished.
- There is an **empty** person list in empty rooms, so the type structure is robust to potential movements of people.
- We are nesting dictionaries and lists, with string and integer data.

Exercise 1e

We can count the occupants and capacity like this:

```

capacity = 0
occupancy = 0
for name, room in house.items():
    capacity += room["capacity"]
    occupancy += len(room["people"])
    print(f"House can fit {capacity} people, and currently has: {occupancy}.")

```

```

House can fit 7 people, and currently has: 2.

```

As a side note, note how we included the values of `capacity` and `occupancy` in the last line. This is a handy syntax for building strings that contain the values of variables. You can read more about it [here](#) or in the official documentation for formatted string literals: [f-strings](#).

Module 02

Exercise 2a/b

```

house = {
    "living": {
        "exits": {"north": "kitchen", "outside": "garden", "upstairs": "bedroom"},
        "people": ["James"],
        "capacity": 2,
    },
    "kitchen": {"exits": {"south": "living"}, "people": [], "capacity": 1},
    "garden": {"exits": {"inside": "living"}, "people": ["Sue"], "capacity": 3},
    "bedroom": {
        "exits": {"downstairs": "living", "jump": "garden"},
        "people": [],
        "capacity": 1,
    },
}

```

Answer 2a

We can get a simpler dictionary with just capacities like this:

```

{name: room['capacity'] for name, room in house.items()}

{'living': 2, 'kitchen': 1, 'garden': 3, 'bedroom': 1}

```

Answer 2b

To get the current number of occupants, we can use a similar dictionary comprehension. Remember that we can *filter* (only keep certain rooms) by adding an `if` clause:

```

{name: len(room["people"]) for name, room in house.items() if len(room["people"])
> 0}

{'living': 1, 'garden': 1}

```

Answer 2c

Things to notice here:

`1.99999` doesn't round, even if you did `int(1.999999)` you would get 1.

`round(1.999999)` or `int(1.99999999999999)` would give you 2

Strings aren't integers

Even though 20 and 5 are integers and they divide to give 4, the result is a float, not an int. Floor division (`20 // 5`) will return an integer.

'10.' is a float not an integer

Can do this in one line using comprehension or could make an empty list and append to it.

```

def example_func(*args):
    op = [a for a in args if type(a)== int and a%2 == 0]
    return op

example_func(1, 1.9999999999, "three", 20/5, 5, 6, "sju", "8", 9, 10., 11, 12)

[6, 12]

```

Answer 2d

Will have to import libraries.

Can use `dir(X)` to list the attributes of the modules

There will be some depreciation warnings from `scipy` instructing users to go use `numpy` or `numpy.lib` (which can also be investigated via `dir(np.lib)`)

`Statistics` will return the mean as an integer whereas `numpy` and `scipy` will return a float.

All return the same value of `pi`.

`scipy` returns a complex number for the negative log example with an imaginary part of `pi`.

`log(+ive)` using `+12.01` as an example

`log(-ive)` using `-11.99` as an example

Module	pi	log(+ive)	log(-ive)	mean
numpy	3.14159...	2.48573...	nan	5.0
scipy	3.14159...	2.48573... (2.48407... + 3.14159...j)	5.0	
math	3.14159...	2.48573...	math domain error	§
statistics	§	2.48573...	math domain error	5

`§` module doesn't have method

Answer 2e

Broad range of options, this is simply one of the possibilities given in the original notebook with the inclusion of some typehinting and descriptions of methods/classes

Note. For more information on type annotations, look into [Module 7.2](#)

```

import typing
class Maze:
    """
    Here we can put a description of the class
    """
    def __init__(self, name: str):
        # We can also use typehints to signal what type a variable should be
        # In this case the name of the maze would be a string.
        self.name = name
        self.rooms = {}

    def add_room(self, room):
        room.maze = self # The Room needs to know which Maze it is a part of
        self.rooms[room.name] = room # This means that we expect our Rooms class
        to have a 'name' property

    def occupants(self):
        """
        Return a list containing the occupants of the maze
        """
        return [occupant for room in self.rooms.values() for occupant in
            room.occupants.values()]

    def wander(self):
        """Move all the people in a random direction"""
        for occupant in self.occupants():
            occupant.wander()

    def describe(self):
        for room in self.rooms.values():
            room.describe()

    def step(self):
        self.describe()
        print("")
        self.wander()
        print("")

    def simulate(self, steps):
        for _ in range(steps):
            self.step()

    class Room:
        def __init__(self, name: str, exits: dict, capacity: int, maze=None):
            self.maze = maze
            self.name = name
            self.occupants = {} # Note the default argument, occupants start empty
            self.exits = exits # Should be a dictionary from directions to room names
            self.capacity = capacity

        def has_space(self) -> bool:
            """
            Check if the room has space and return a boolean (True/False)
            """
            return len(self.occupants) < self.capacity

        def available_exits(self) -> typing.List[str]:
            return [
                exit
                for exit, target in self.exits.items()
                if self.maze.rooms[target].has_space()
            ]

        def random_valid_exit(self):
            import random

            if not self.available_exits():
                return None
            return random.choice(self.available_exits())

        def destination(self, exit):
            return self.maze.rooms[self.exits[exit]]

        def add_occupant(self, occupant):
            occupant.room = self # The person needs to know which room it is in
            self.occupants[occupant.name] = occupant

        def delete_occupant(self, occupant):
            del self.occupants[occupant.name]

        def describe(self):
            if self.occupants:
                print(f"{self.name}: " + " ".join(self.occupants.keys()))

    class Person:
        def __init__(self, name: str, room=None):
            self.name = name

        def use(self, exit):
            self.room.delete_occupant(self)
            destination = self.room.destination(exit)
            destination.add_occupant(self)
            print(
                "{some} goes {action} to the {where}".format(
                    some=self.name, action=exit, where=destination.name
                )
            )

        def wander(self):
            exit = self.room.random_valid_exit()
            if exit:
                self.use(exit)

```

```

james = Person("James")
sue = Person("Sue")
bob = Person("Bob")
clare = Person("Clare")

living = Room("livingroom", {"outside": "garden", "upstairs": "bedroom", "north": "kitchen"}, 2)
kitchen = Room("kitchen", {"south": "livingroom"}, 1)
garden = Room("garden", {"inside": "livingroom"}, 3)
bedroom = Room("bedroom", {"jump": "garden", "downstairs": "livingroom"}, 1)

house = Maze("My House")

for room in [living, kitchen, garden, bedroom]:
    house.add_room(room)

living.add_occupant(james)
garden.add_occupant(sue)
garden.add_occupant(clare)
bedroom.add_occupant(bob)

```

```
house.simulate(3)
```

```
livingroom: James
garden: Sue Clare
bedroom: Bob

James goes outside to the garden
Sue goes inside to the livingroom
Clare goes inside to the livingroom
Bob goes jump to the garden

livingroom: Sue Clare
garden: James Bob

Sue goes upstairs to the bedroom
Clare goes outside to the garden
James goes inside to the livingroom
Bob goes inside to the livingroom

livingroom: James Bob
garden: Clare
bedroom: Sue

James goes north to the kitchen
Bob goes outside to the garden
Clare goes inside to the livingroom
Sue goes downstairs to the livingroom
```

Answer 2f

Something along the lines of this for the original question:

```
import requests
from IPython.display import Image

coordinates_as_lat_lon = [(36.2110, -115.2669),
                           (53.0066, 7.1920),
                           (41.3908, 2.1631),
                           (40.7822, -73.9653),
                           (25.8380, 50.6050)]

def op_response(lat, lon):
    response = requests.get(
        "https://static-maps.yandex.ru:443/1.x",
        params={
            "size": "400,400", # size of map
            "ll": str(lon) + "," + str(lat), # longitude & latitude of centre
            "z": 12, # zoom level
            "l": "sat", # map layer (satellite image)
            "lang": "en_US", # language
        },
    )
    return response.content

op = op_response(*coordinates_as_lat_lon[4])
Image(op)
```



Answer 2e

```
def extended_op_response(lat, lon, zoom=15, opfname="tmp.png"):

    response = requests.get(
        "https://static-maps.yandex.ru:443/1.x",
        params={
            "size": "400,400", # size of map
            "ll": str(lon) + "," + str(lat), # longitude & latitude of centre
            "z": zoom, # zoom level
            "l": "sat", # map layer (satellite image)
            "lang": "en_US", # language
        },
    )

    with open(opfname, "wb") as png:
        png.write(response.content)

extended_op_response(*coordinates_as_lat_lon[1], zoom=16,
opfname="map_picture_1.png")
Image("map_picture_1.png")
```



Module 03

Exercise 3a Saving and loading data

Relevant sections: 3.1.2, 3.1.3

Use YAML or JSON to save your maze data structure to disk and load it again.

The maze would have looked something like this:

```
house = {
    "living": {
        "exits": {"north": "kitchen", "outside": "garden", "upstairs": "bedroom"},
        "people": ["James"],
        "capacity": 2,
    },
    "kitchen": {"exits": {"south": "living"}, "people": [], "capacity": 1},
    "garden": {"exits": {"inside": "living"}, "people": ["Sue"], "capacity": 3},
    "bedroom": {
        "exits": {"downstairs": "living", "jump": "garden"},
        "people": [],
        "capacity": 1,
    },
}
```

Exercise 3a Answer

Save as JSON or YAML

```
import json
import yaml

# Write with json.dump
with open("myfile.json", "w") as f:
    json.dump(house, f)

# Look at the file on disk
!cat myfile.json

{"living": {"exits": {"north": "kitchen", "outside": "garden", "upstairs": "bedroom"}, "people": ["James"], "capacity": 2}, "kitchen": {"exits": {"south": "living"}, "people": [], "capacity": 1}, "garden": {"exits": {"inside": "living"}, "people": ["Sue"], "capacity": 3}, "bedroom": {"exits": {"downstairs": "living", "jump": "garden"}, "people": [], "capacity": 1}}


# Or with file.write, using json.dumps to convert to a string
with open("myotherfile.json", "w") as json_maze_out:
    json_maze_out.write(json.dumps(house))

# Look at the file on disk
!cat myotherfile.json

{"living": {"exits": {"north": "kitchen", "outside": "garden", "upstairs": "bedroom"}, "people": ["James"], "capacity": 2}, "kitchen": {"exits": {"south": "living"}, "people": [], "capacity": 1}, "garden": {"exits": {"inside": "living"}, "people": ["Sue"], "capacity": 3}, "bedroom": {"exits": {"downstairs": "living", "jump": "garden"}, "people": [], "capacity": 1}}


# Write with yaml.safe_dump
with open("myfile.yml", "w") as f:
    yaml.safe_dump(house, f, default_flow_style=False)

# Look at the file on disk
!cat myfile.yml
```

```

bedroom:
    capacity: 1
    exits:
        downstairs: living
        jump: garden
        people: []
garden:
    capacity: 3
    exits:
        inside: living
        people:
            - Sue
kitchen:
    capacity: 1
    exits:
        south: living
        people: []
living:
    capacity: 2
    exits:
        north: kitchen
        outside: garden
        upstairs: bedroom
    people:
        - James

```

```

# Or with file.write, using yaml.dump to convert to a string
with open("myotherfile.yaml", "w") as yaml_maze_out:
    yaml_maze_out.write(yaml.dump(house, default_flow_style=True))

```

```

# Look at the file on disk
!cat myotherfile.yaml

```

```

{'bedroom': {'capacity': 1, 'exits': {'downstairs': 'living', 'jump': 'garden'}, 'people': []}, 'garden': {'capacity': 3, 'exits': {'inside': 'living'}, 'people': ['Sue']}, 'kitchen': {'capacity': 1, 'exits': {'south': 'living'}, 'people': []}, 'living': {'capacity': 2, 'exits': {'north': 'kitchen', 'outside': 'garden', 'upstairs': 'bedroom'}, 'people': ['James']}}

```

Loading with JSON or YAML

```

# Read into a string then load with json.loads
with open("myfile.json", "r") as f:
    mydataasstring = f.read()
    my_json_data = json.loads(mydataasstring)
    print(my_json_data["living"])

```

```

{'exits': {'north': 'kitchen', 'outside': 'garden', 'upstairs': 'bedroom'}, 'people': ['James'], 'capacity': 2}

```

```

# Read directly with json.load
with open("myotherfile.json") as f_json_maze:
    maze_again = json.load(f_json_maze)
    print(maze_again["living"])

```

```

{'exits': {'north': 'kitchen', 'outside': 'garden', 'upstairs': 'bedroom'}, 'people': ['James'], 'capacity': 2}

```

```

# Read into a string then load with yaml.safe_load
with open("myfile.yaml", "r") as f:
    mydataasstring = f.read()
    my_yaml_data = yaml.safe_load(mydataasstring)
    print(my_yaml_data["living"])

```

```

{'exits': {'north': 'kitchen', 'outside': 'garden', 'upstairs': 'bedroom'}, 'people': ['James'], 'capacity': 2}

```

```

# Read directly with yaml.safe_load
with open("myotherfile.yaml") as f_yaml_maze:
    maze_again = yaml.safe_load(f_yaml_maze)
    print(maze_again["living"])

```

```

{'capacity': 2, 'exits': {'north': 'kitchen', 'outside': 'garden', 'upstairs': 'bedroom'}, 'people': ['James']}

```

Exercise 3b Plotting with matplotlib

Generate two plots, next to each other (on the same row).

The first plot should show $\sin(x)$ and $\cos(x)$ for the range of x between -1π and $+1\pi$.

The second plot should show $\sin(x)$, $\cos(x)$ and the sum of $\sin(x)$ and $\cos(x)$ over the same $-\pi$ to $+\pi$ range. Set suitable limits on the axes and pick colours, markers, or line-styles that will make it easy to differentiate between the curves. Add legends to both axes.

Exercise 3b Answer

```

import matplotlib.pyplot as plt
import numpy as np

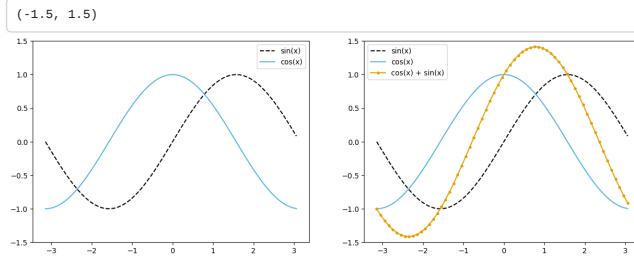
# Use numpy to get the range of x values (math should work too)
x = np.arange(-np.pi, np.pi, 0.1)

# Define figure dimensions
fig = plt.figure(figsize=(15,5))

ax1 = fig.add_subplot(1,2,1)
ax1.plot(x, np.sin(x),label="sin(x)",color='black', linestyle='dashed')
ax1.plot(x, np.cos(x),label="cos(x)", color="#56B4E9")
ax1.legend()
ax1.set_xlim(-1.5, 1.5)

ax2 = fig.add_subplot(1,2,2)
ax2.plot(x, np.sin(x),label="sin(x)",color='black', linestyle='dashed')
ax2.plot(x, np.cos(x),label="cos(x)", color="#56B4E9")
ax2.plot(x, np.cos(x)+np.sin(x), label='cos(x) + sin(x)', color="#E69F00",
marker=".")
ax2.legend()
ax2.set_xlim(-1.5, 1.5)

```



Exercise 3c The biggest earthquake in the UK this century

The Problem

GeoJSON is a json-based file format for sharing geographic data. One example dataset is the USGS earthquake data:

```

import requests
quakes = requests.get(
    "http://earthquake.usgs.gov/fdsnws/event/1/query.geojson",
    params={
        "starttime": "2000-01-01",
        "maxlatitude": "58.723",
        "minlatitude": "50.068",
        "maxlongitude": "1.67",
        "minlongitude": "-9.756",
        "minmagnitude": "1",
        "endtime": "2021-01-19",
        "orderby": "time-asc",
    },
)

```

quakes.text[0:100]

```
{
  "type": "FeatureCollection",
  "metadata": {
    "generated": 1667560431000,
    "url": "https://earthquake.usgs.gov"
}
```

Exercise 3c Answer

Relevant sections: 3.1, 2.5.2, 2.5.1

Load the data

- Get the text of the web result
- Parse the data as JSON

```

import requests
quakes = requests.get(
    "http://earthquake.usgs.gov/fdsnws/event/1/query.geojson",
    params={
        "starttime": "2000-01-01",
        "maxlatitude": "58.723",
        "minlatitude": "50.068",
        "maxlongitude": "1.67",
        "minlongitude": "-9.756",
        "minmagnitude": "1",
        "endtime": "2022-11-02", # Change the date to yesterday
        "orderby": "time-asc",
    },
)

import json
# Can get the data indirectly via the text and then load json text...
my_quake_data = json.loads(quakes.text) # Section 3.1 - structured data

# Requests also has a built in json parser (note this gives exactly the same
# result as 'my_quake_data')
requests_json = quakes.json()

```

Investigate the data

- Understand how the data is structured into dictionaries and lists
 - Where is the magnitude?
 - Where is the place description or coordinates?

There is no foolproof way of doing this. A good first step is to see the type of our data!

type(requests_json)

dict

Now we can navigate through this dictionary to see how the information is stored in the nested dictionaries and lists. The `keys` method can indicate what kind of information each dictionary holds, and the `len` function tells us how many entries are contained in a list. How you explore is up to you!

```
requests_json.keys()
dict_keys(['type', 'metadata', 'features', 'bbox'])
type(requests_json["features"])
list
len(requests_json["features"])
131
requests_json["features"][0]
{'type': 'Feature',
'properties': {'mag': 2.6,
'place': '12 km NNW of Penrith, United Kingdom',
'time': 956553055700,
'updated': 1415322596133,
'tz': None,
'url': 'https://earthquake.usgs.gov/earthquakes/eventpage/usp0009rst',
'detail': 'https://earthquake.usgs.gov/fdsnws/event/1/query?
eventid=usp0009rst&format=geojson',
'felt': None,
'cdi': None,
'mmi': None,
>alert': None,
'status': 'reviewed',
'tsunami': 0,
'sig': 104,
'net': 'us',
'code': 'p0009rst',
'ids': 'usp0009rst',
'sources': 'us',
'types': 'impact-text,origin,phase-data',
'nst': None,
'dmin': None,
'rms': None,
'gap': None,
'magType': 'ml',
'type': 'earthquake',
'title': 'M 2.6 - 12 km NNW of Penrith, United Kingdom'},
'geometry': {'type': 'Point', 'coordinates': [-2.81, 54.77, 14]},
'id': 'usp0009rst'}
requests_json["features"][0].keys()
dict_keys(['type', 'properties', 'geometry', 'id'])

It looks like the coordinates are in the geometry section and the magnitude is in the properties section.
```

```
requests_json["features"][0]["geometry"]
{'type': 'Point', 'coordinates': [-2.81, 54.77, 14]}
requests_json["features"][0]["properties"].keys()
dict_keys(['mag', 'place', 'time', 'updated', 'tz', 'url', 'detail', 'felt',
'cdi', 'mmi', 'alert', 'status', 'tsunami', 'sig', 'net', 'code', 'ids',
'sources', 'types', 'nst', 'dmin', 'rms', 'gap', 'magType', 'type', 'title'])
requests_json["features"][0]["properties"]["mag"]
2.6
```

Search through the data

- Program a search through all the quakes to find the biggest quake
- Find the place of the biggest quake

```
quakes = requests_json["features"]
largest_so_far = quakes[0]
for quake in quakes:
    if quake["properties"]["mag"] > largest_so_far["properties"]["mag"]:
        largest_so_far = quake
largest_so_far["properties"]["mag"]
4.8
lon = largest_so_far["geometry"]["coordinates"][0]
lat = largest_so_far["geometry"]["coordinates"][1]
print(f"Latitude: {lat} Longitude: {lon}")
Latitude: 52.52 Longitude: -2.15
```

Visualise your answer

- Form a URL for an online map service at that latitude and longitude: look back at the introductory example
- Display that image

```
import IPython
import requests
```

```

# This is a solution to one of the questions in module 2
# The only difference here is that the map type is set to map rather than
# satellite view and the zoom is 10 not 12
def op_response(lat, lon):
    response = requests.get(
        "https://static-maps.yandex.ru:443/1.x",
        params={
            "size": "400,400", # size of map
            "ll": str(lon) + "," + str(lat), # longitude & latitude of centre
            "z": 10, # zoom level
            "l": "map", # map layer (map image)
            "lang": "en_US", # language
        },
    )
    return response.content

op = op_response(lat, lon)
IPython.core.display.Image(op)

```



[Optional] Equivalent solution using pandas

In this instance Pandas probably isn't the first thing that you would use as we have nested dictionaries and JSON works very well in such cases. If we really want to use Pandas we'll need to flatten the nested values before constructing a DataFrame.

```

features = requests_json["features"]
features[0]

```

```

{
  "type": "Feature",
  "properties": {"mag": 2.6,
  "place": "12 km NNW of Penrith, United Kingdom",
  "time": 956553055700,
  "updated": 1415322596133,
  "tz": None,
  "url": "https://earthquake.usgs.gov/earthquakes/eventpage/usp0009rst",
  "detail": "https://earthquake.usgs.gov/fdsnws/event/1/query?
eventid=usp0009rst&format=geojson",
  "felt": None,
  "cdi": None,
  "mmi": None,
  "alert": None,
  "status": "reviewed",
  "tsunami": 0,
  "sig": 104,
  "net": "us",
  "code": "p0009rst",
  "ids": ",usp0009rst",
  "sources": ",us",
  "types": ",impact-text,origin,phase-data",
  "nst": None,
  "dmin": None,
  "rms": None,
  "gap": None,
  "magType": "ml",
  "type": "earthquake",
  "title": "M 2.6 - 12 km NNW of Penrith, United Kingdom",
  "geometry": {"type": "Point", "coordinates": [-2.81, 54.77, 14]},
  "id": "usp0009rst"
}

```

```

# We can use ** to convert a dictionary into pairs of (key, value)
# We can then run `{{k1, v1}, {k2, v2}}` to convert a list of keys and values back
# into a dictionary
combined_features = [{**f["geometry"], **f["properties"]} for f in features]
combined_features[0]

```

```

{
  "type": "earthquake",
  "coordinates": [-2.81, 54.77, 14],
  "mag": 2.6,
  "place": "12 km NNW of Penrith, United Kingdom",
  "time": 95655305700,
  "updated": 1415322596133,
  "tz": None,
  "url": "https://earthquake.usgs.gov/earthquakes/eventpage/usp0009rst",
  "detail": "https://earthquake.usgs.gov/fdsnws/event/1/query?
eventid=usp0009rst&format=geojson",
  "felt": None,
  "cdi": None,
  "mmi": None,
  "alert": None,
  "status": "reviewed",
  "tsunami": 0,
  "sig": 104,
  "net": "us",
  "code": "p0009rst",
  "ids": ",usp0009rst",
  "sources": ",us",
  "types": ",impact-text,origin,phase-data",
  "nst": None,
  "dmin": None,
  "rms": None,
  "gap": None,
  "magType": "ml",
  "type": "earthquake",
  "title": "M 2.6 - 12 km NNW of Penrith, United Kingdom"
}

```

```

import pandas as pd
df = pd.DataFrame.from_records(combined_features)
df.head()

```

	type	coordinates	mag	place	time	updated	tz	url
0	earthquake	[2.81, 54.77, 14]	2.6	12 km NNW of Penrith, United Kingdom	956553055700	1415322596133	None	https://earthquake.usgs.gov/earthquakes/eventpage/1 https://earthquake.usgs.gov/fdsnws/event/1
1	earthquake	[-1.61, 52.28, 13.1]	4.0	1 km WSW of Warwick, United Kingdom	969683025790	1415322666913	None	https://earthquake.usgs.gov/earthquakes/eventpage/1 https://earthquake.usgs.gov/fdsnws/event/1
2	earthquake	[1.564, 53.236, 10]	4.0	38 km NNE of Cromer, United Kingdom	977442788510	1415322705662	None	https://earthquake.usgs.gov/earthquakes/eventpage/1 https://earthquake.usgs.gov/fdsnws/event/1
3	earthquake	[0.872, 58.097, 10]	3.3	171 km ENE of Peterhead, United Kingdom	984608438660	1415322741153	None	https://earthquake.usgs.gov/earthquakes/eventpage/1 https://earthquake.usgs.gov/fdsnws/event/1
4	earthquake	[-1.845, 51.432, 10]	2.9	8 km W of Marlborough, United Kingdom	984879824720	1415322742102	None	https://earthquake.usgs.gov/earthquakes/eventpage/1 https://earthquake.usgs.gov/fdsnws/event/1

5 rows × 27 columns

```

df.sort_values("mag", ascending=False, inplace=True)
df.head()

```

	type	coordinates	mag	place	time	updated	tz	url
19	earthquake	[-2.15, 52.52, 9.4]	4.8	2 km ESE of Wombourne, United Kingdom	1032738794600	1600455819229	None	https://earthquake.usgs.gov/earthquakes/eventpage/1 https://earthquake.usgs.gov/fdsnws/event/1
81	earthquake	[-0.332, 53.403, 18.4]	4.8	1 km NNE of Market Rasen, United Kingdom	1204073807800	1657747150218	None	https://earthquake.usgs.gov/earthquakes/eventpage/1 https://earthquake.usgs.gov/fdsnws/event/1
72	earthquake	[1.009, 51.085, 10]	4.6	1 km WNW of Lympne, United Kingdom	1177744691360	1657780288041	None	https://earthquake.usgs.gov/earthquakes/eventpage/1 https://earthquake.usgs.gov/fdsnws/event/1
23	earthquake	[2.219, 53.478, 5]	4.3	1 km ESE of Manchester, United Kingdom	1035200554900	1415323007416	None	https://earthquake.usgs.gov/earthquakes/eventpage/1 https://earthquake.usgs.gov/fdsnws/event/1
113	earthquake	[-3.8559, 51.7231, 11.55]	4.3	5 km NE of Clydach, United Kingdom	1518877865070	1664101506468	None	https://earthquake.usgs.gov/earthquakes/eventpage/1 https://earthquake.usgs.gov/fdsnws/event/1

5 rows × 27 columns

You can see that we haven't really gained much over the JSON solution. We still needed to look at the data to see its structure and we had to manually flatten the structure.

Module 04

Module 05

Module 06: Troll Treasure

A sample solution for packaging the Troll Treasure code is available in this GitHub repo:

<https://github.com/alan-turing-institute/TrollTreasure>

Module 07: Bad Boids

There's not a single "right" answer to how the code should be refactored, but on the `better_boids` branch of the `bad-boids` repo we have an improved version of the code with changes based on all the ideas above. You can find it on GitHub here:

https://github.com/alan-turing-institute/bad-boids/tree/better_boids

You may also find it interesting to browse through the [history of commits](#).

Alternatively, you can checkout the branch in your local clone:

```
git checkout better_boids
```

Module 08

Module 09

Module 10

By various [contributors](#). Developed at [The Alan Turing Institute](#) based on the [UCL RSD course](#).

Creative Commons Attribution 2.0 Generic ([CC BY 2.0](#)).