Queen Mary
University of London

QHP4701
Introduction to Data Science Programming

Introduction to Data Science, Computer, and programming

Lecturer: Nikesh Bajaj, PhD
School of Physical and Chemical Sciences
nikesh.bajaj@qmul.ac.uk

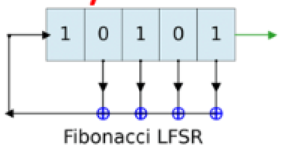
Welcome to QHP4701

Team

- Nikesh Bajaj (Nik)

I am a Lecturer in Data Science at Queen Mary University of London. I have been working in the field of Data Science for more than 10 years. I completed my PhD in Machine Learning and Signal Processing. I worked on Deception Detection (NLP & Linguistics) for almost 2 years, and Computational problems of Cardiology (ECG, EGM) for last 2 years. I have a few python libraries and a few Data Science projects shared as Open Source.

PyLFSR



PhyAAt

Spkit

- Jiayu Men

Jiayu is a PhD student at QMUL and about to finish by the end of 2023. Her research topic is motion planning for unmanned aerial vehicles. At the same time, she has been demonstrating on machine learning modules at QMUL since 2017.

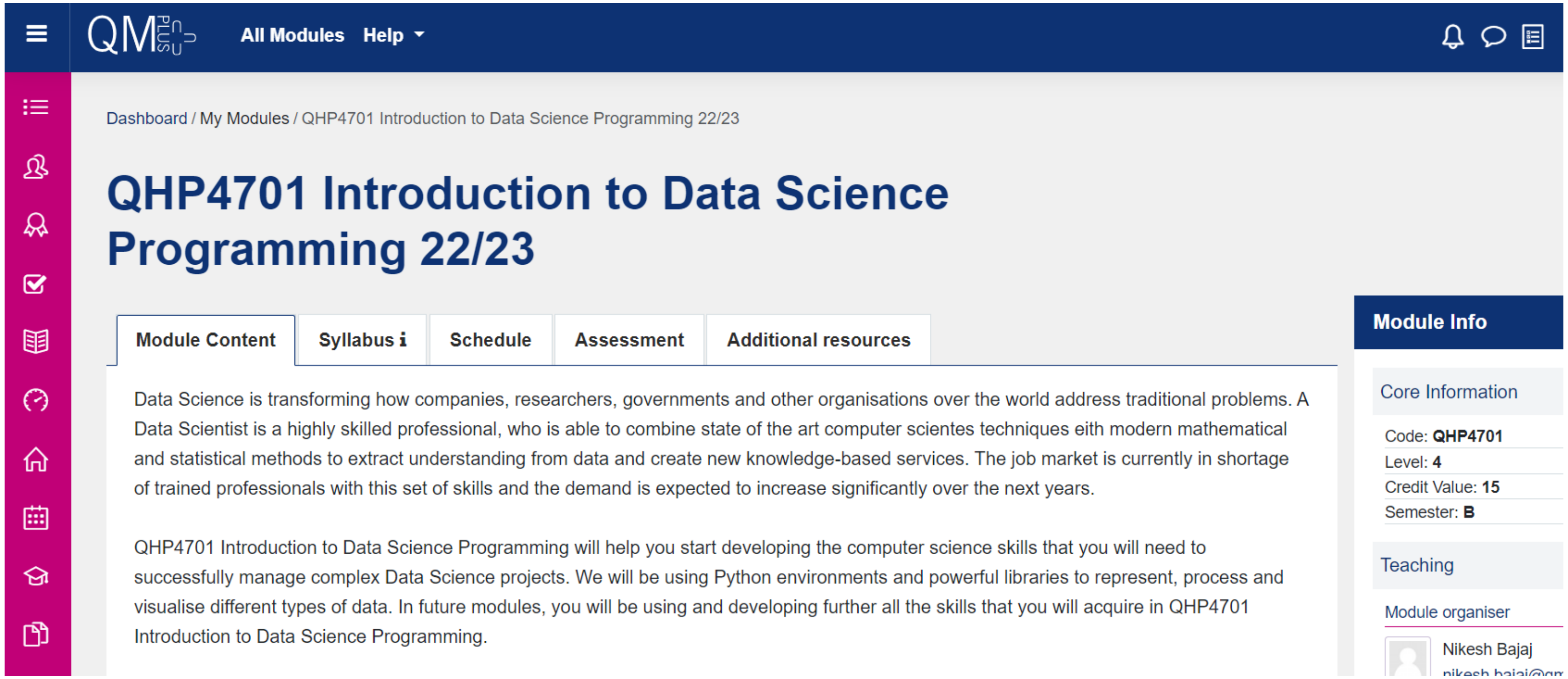
Jiayu will be managing and helping you all during lab sessions.

Lecture Outline

- About the module
- What is Data and Data Science?
- Aim of Data Science Tasks
- Programming Languages and Tools
- Introduction to Python

Module Information

All the relevant information about the Module - **QHP4701**, can be found on QM+ Page



Dashboard / My Modules / QHP4701 Introduction to Data Science Programming 22/23

QHP4701 Introduction to Data Science Programming 22/23

Data Science is transforming how companies, researchers, governments and other organisations over the world address traditional problems. A Data Scientist is a highly skilled professional, who is able to combine state of the art computer scientes techniques eith modern mathematical and statistical methods to extract understanding from data and create new knowledge-based services. The job market is currently in shortage of trained professionals with this set of skills and the demand is expected to increase significantly over the next years.

QHP4701 Introduction to Data Science Programming will help you start developing the computer science skills that you will need to successfully manage complex Data Science projects. We will be using Python environments and powerful libraries to represent, process and visualise different types of data. In future modules, you will be using and developing further all the skills that you will acquire in QHP4701 Introduction to Data Science Programming.

Core Information

Code: **QHP4701**
Level: **4**
Credit Value: **15**
Semester: **B**

Teaching

Module organiser

Nikesh Bajaj
nikesh.bajaj@qm

Module organization: Schedule

Lecture weeks 1, 2, 6, 7, 8 ~ (teaching week 10, 11 , 15, 16, 17) :

- Monday : 09:55 – 11:35 (Lecture)
- Monday : 19:00 – 20:40 (Lab mode)
- Tuesday : 16:15 – 17:55 (Lecture)
- Friday : 09:55 – 11:35 (Lab mode)

Lab weeks 3, 4, 5 ~ (teaching week 12, 13, 14) :

- Monday : 19:00 – 20:40 (Lab)
- Tuesday : 16:15 – 17:55 (Lab)

Week 3, 4 and 5 are for lab sessions, more focused on coursework and assignments.

Module assessment

- Coursework 1 : 40%

  - 10% Quiz-1 (based on Lab-work 1)
  - 10% Quiz-2 (based on Lab-work 2)
  - 20% Lab-work
  - Submission date: end of teaching week 14 (**26th May 2023**)

- Coursework 2 : 60%

  - A full report on assigned task
  - Submission date: **30th June 2023**

Aims of the module

This module aims to provide introductory programming skills and background knowledge that will:

- build confidence in basic programming skills,
- bring you up the same standard of programming and
- underpin future learning in Data Science techniques explored throughout the rest of the programme.

That will allow you to critically engage with current and future developments in the field of data science

Communication

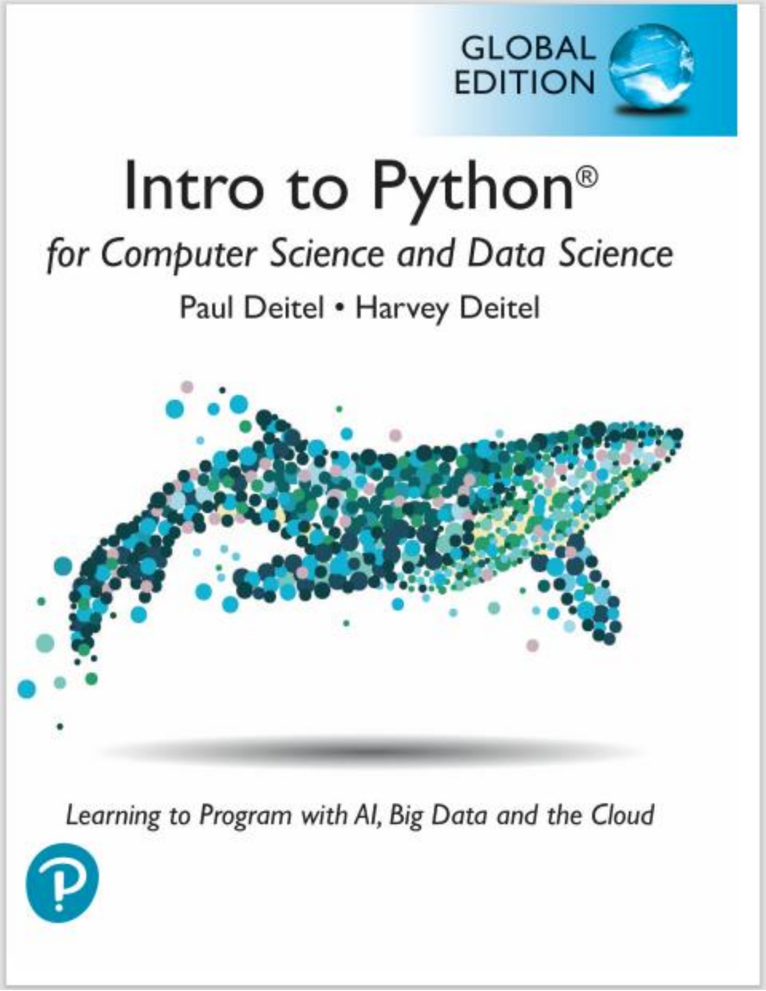
Any question, query and doubt can be asked in following ways

- During Lectures
- On campus or remotely (via MS teams)
- Email: Please make sure its subject is formatted as follows: "[QHP4701] <DESCRIPTIVE SUBJECT HERE>”

  e.g “ ***[QHP4701] Question about Coursework 2*** *”*

- Forum on QM+: Primary means, questions might have been answered already and answers might be useful to others.

Module resources



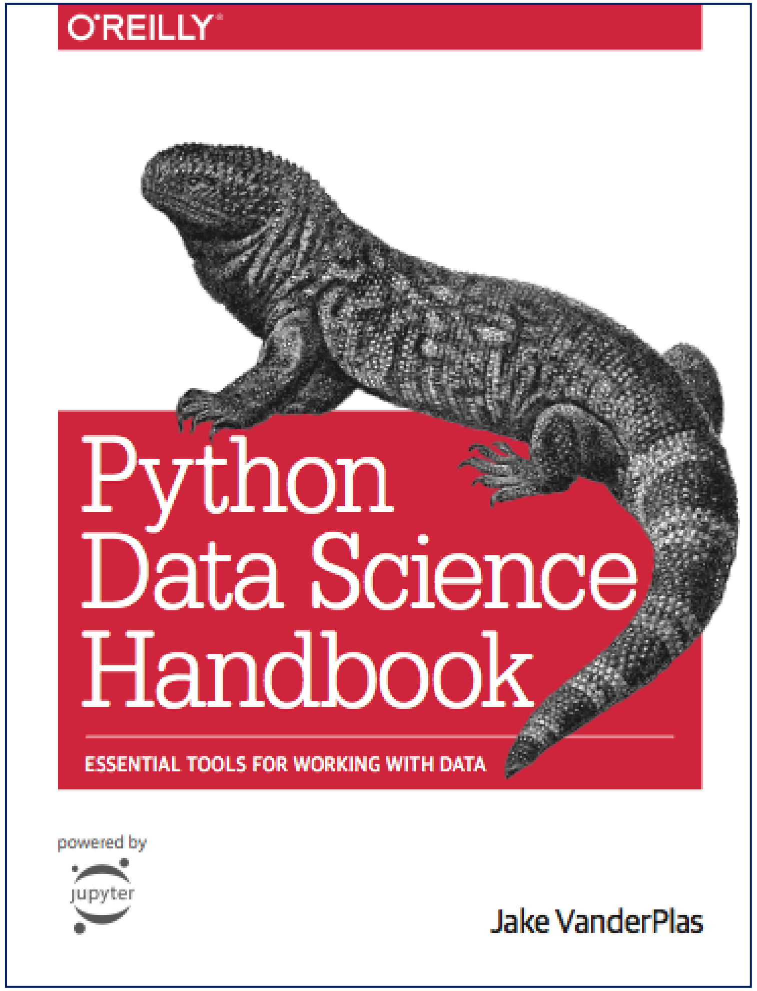
GLOBAL EDITION

Intro to Python®

for Computer Science and Data Science

Paul Deitel • Harvey Deitel

Learning to Program with AI, Big Data and the Cloud



O'REILLY®

Python Data Science Handbook

ESSENTIAL TOOLS FOR WORKING WITH DATA

powered by jupyter

Jake VanderPlas

Lecture Outline

- About the module
- What is Data and Data science?
- Aim of Data Science Tasks
- Programming Languages and Tools
- Introduction to Python

What is Data?

Data?

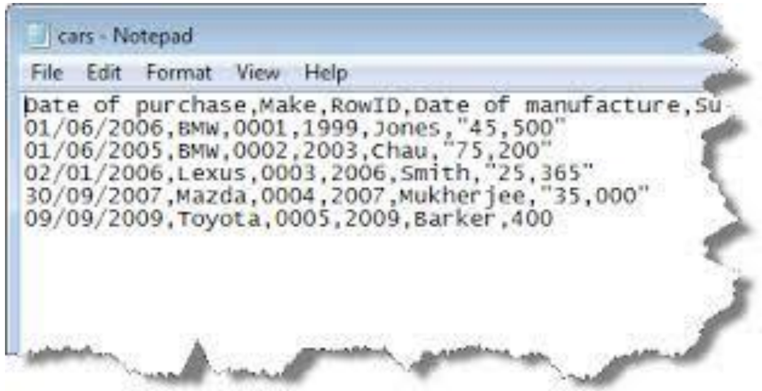
> *“In the pursuit of knowledge, is a collection of discrete values that convey information, describing quantity, quality, fact, statistics, other basic units of meaning, or simply sequences of symbols” …. Wikipedia*

- The word "data" was first used to mean "transmissible and storable computer information" in 1946.
- Data is a collection of values or simply put “*anything that is recorded is data*”
- Let’s see some examples

What is Data: Text

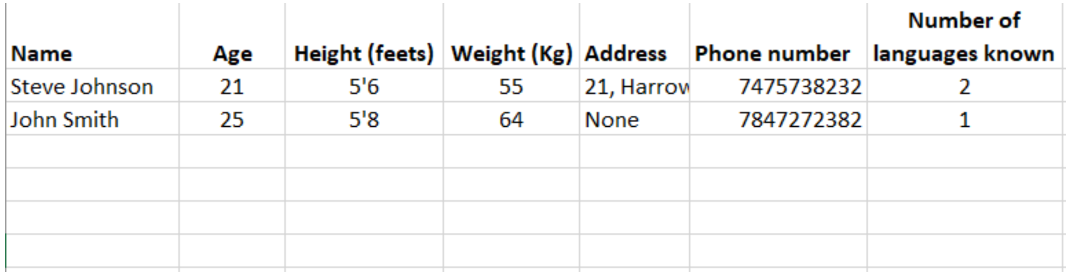
Text

- Sequence of words, or collection of words that has some meaning
- It can be a in form of a Table
- A raw text file



```
Date of purchase,Make,RowID,Date of manufacture,Su
01/06/2006,BMW,0001,1999,Jones,"45,500"
01/06/2005,BMW,0002,2003,Chau,"75,200"
02/01/2006,Lexus,0003,2006,Smith,"25,365"
30/09/2007,Mazda,0004,2007,Mukherjee,"35,000"
09/09/2009,Toyota,0005,2009,Barker,400
```

歡迎來到數據科學

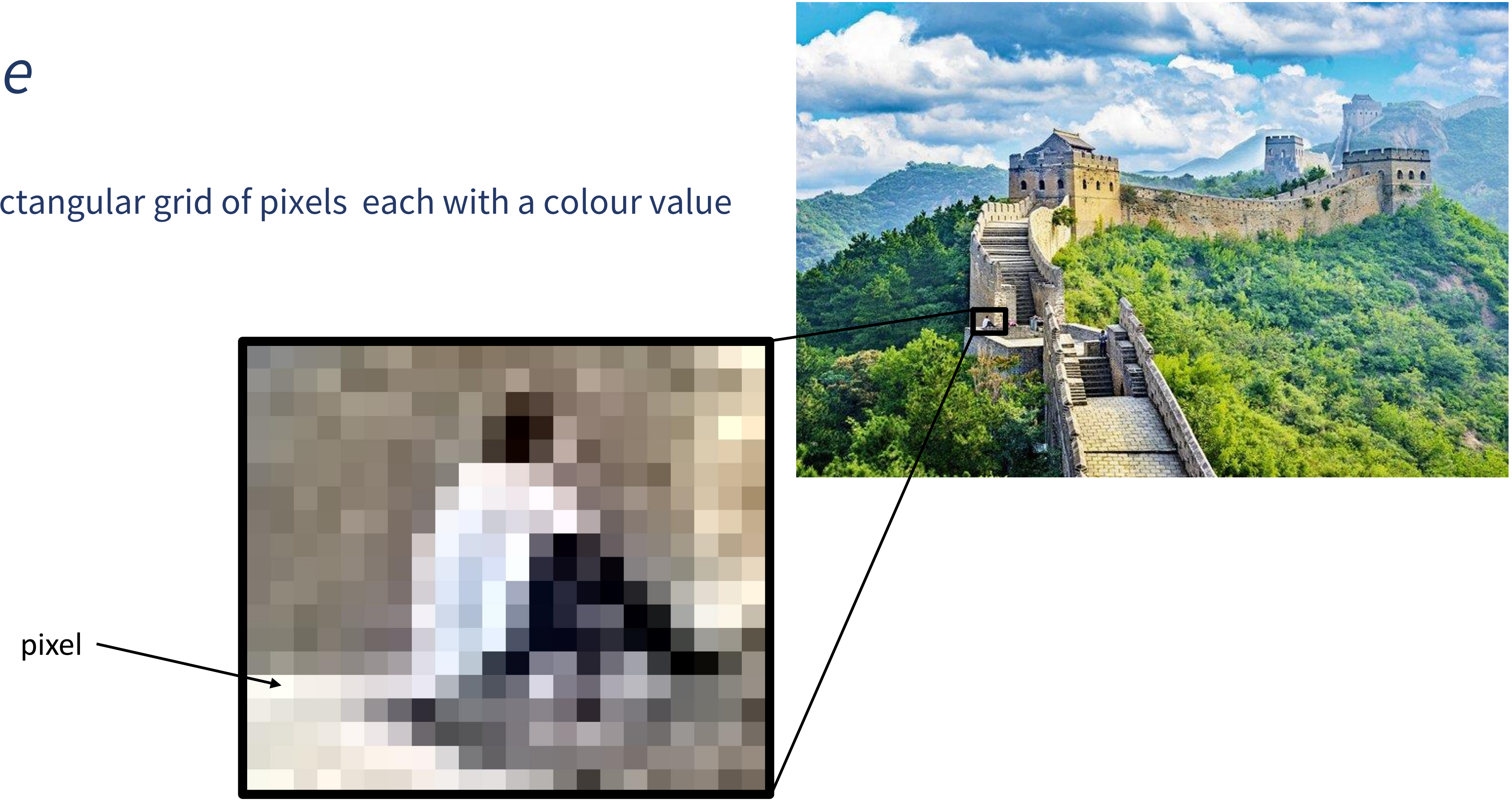


Name	Age	Height (feets)	Weight (Kg)	Address	Phone number	Number of languages known
Steve Johnson	21	5'6	55	21, Harrov	7475738232	2
John Smith	25	5'8	64	None	7847272382	1

What is Data: Image

Image

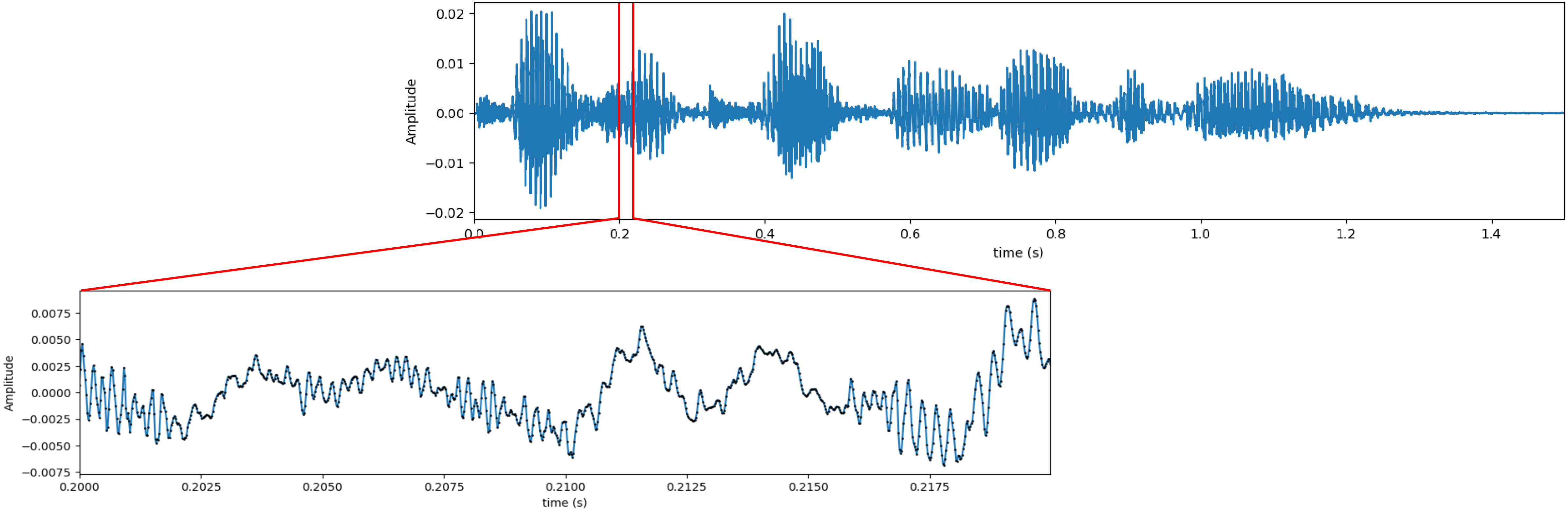
- A rectangular grid of pixels each with a colour value



What is Data: Audio/Speech

Audio/Speech

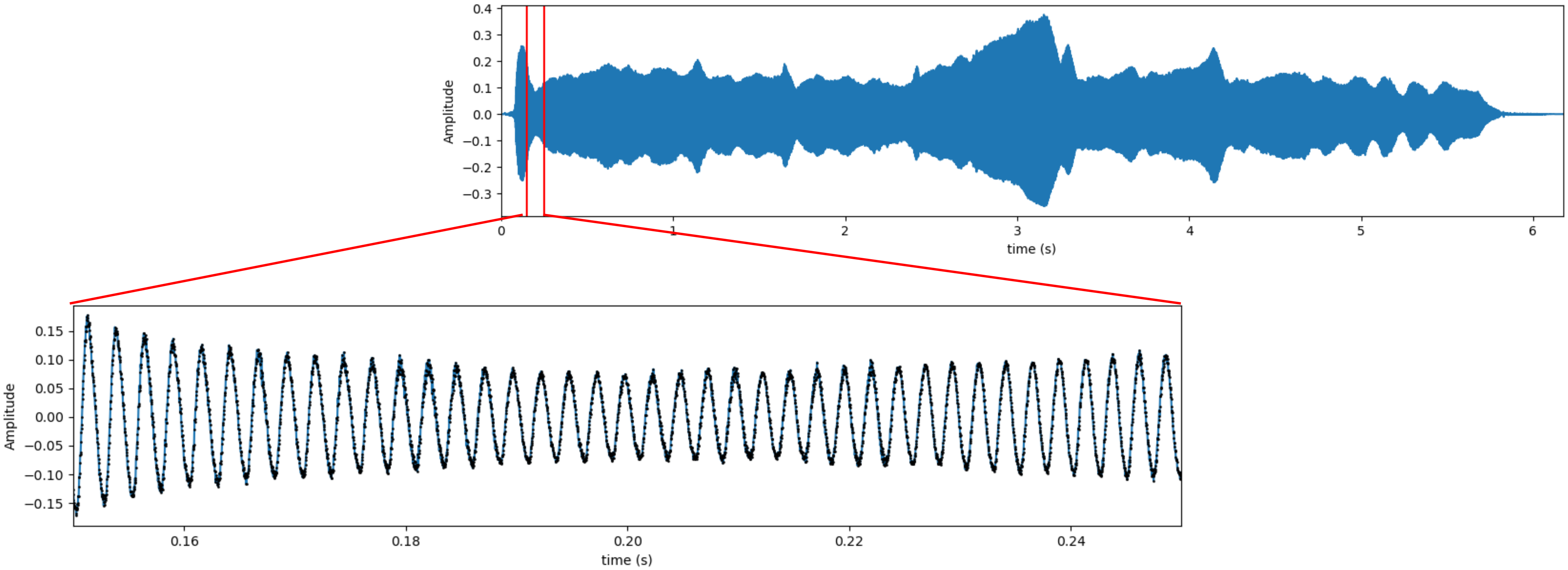
- Audio and Speech are sequences of numbers representing amplitude



What is Data: Audio/Speech

Audio/Speech

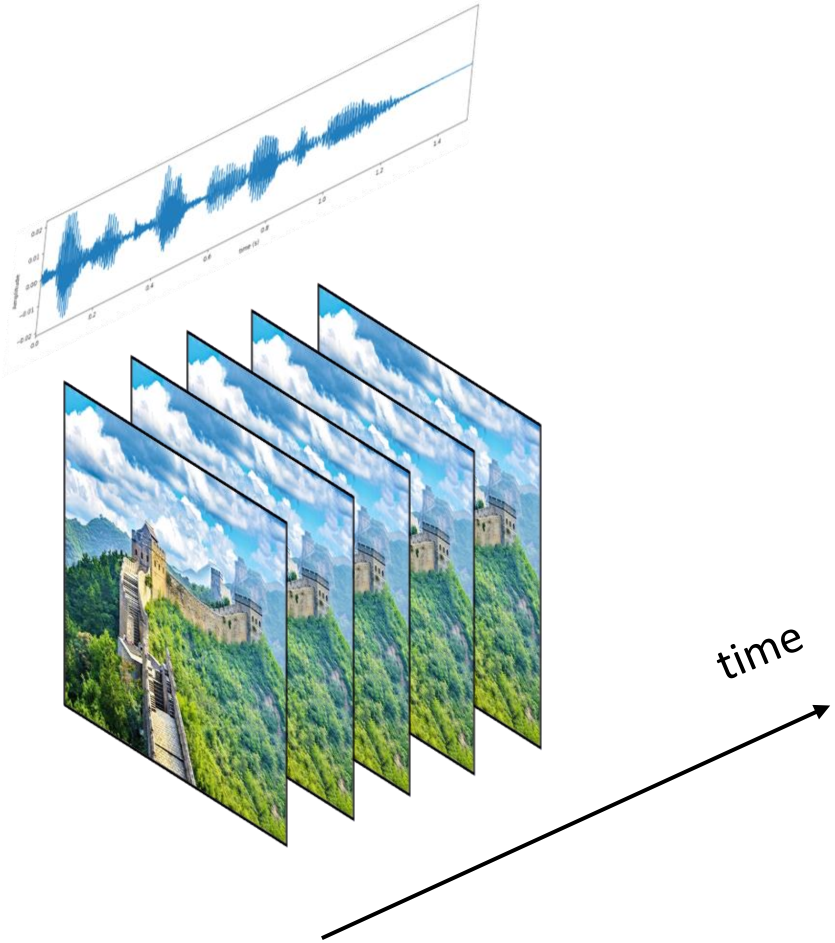
- Audio and Speech are sequences of numbers representing amplitude



What is Data: Video

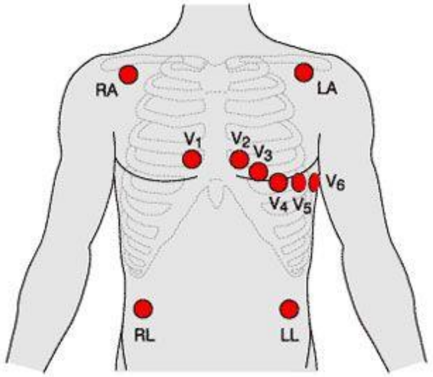
Video

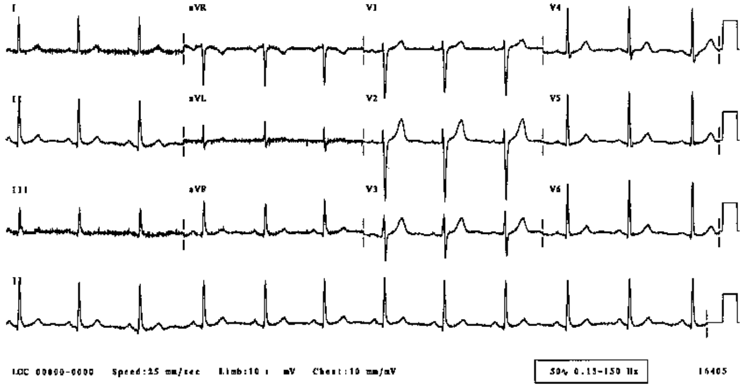
- Sequence of Images and Audio



What is Data: Physiological Signals

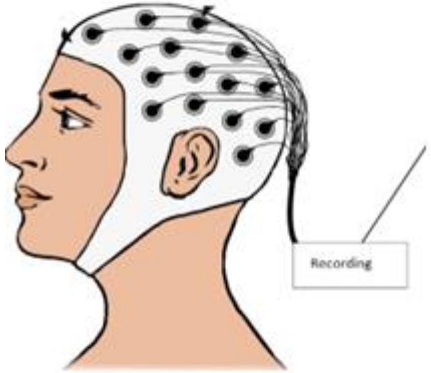
ECG: Electrocardiogram

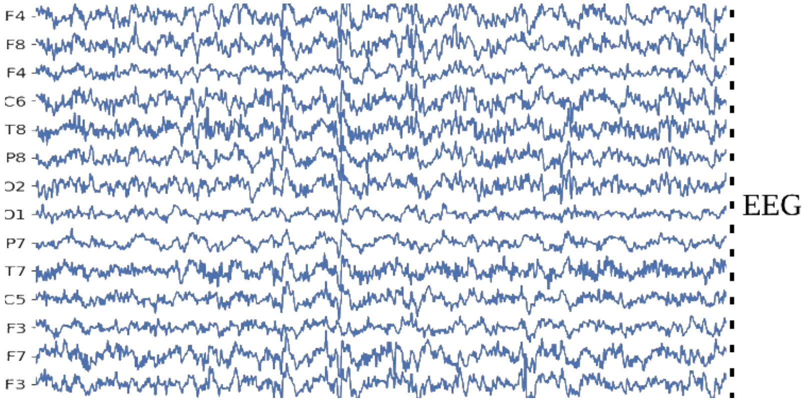




EEG: Electroencephalogram

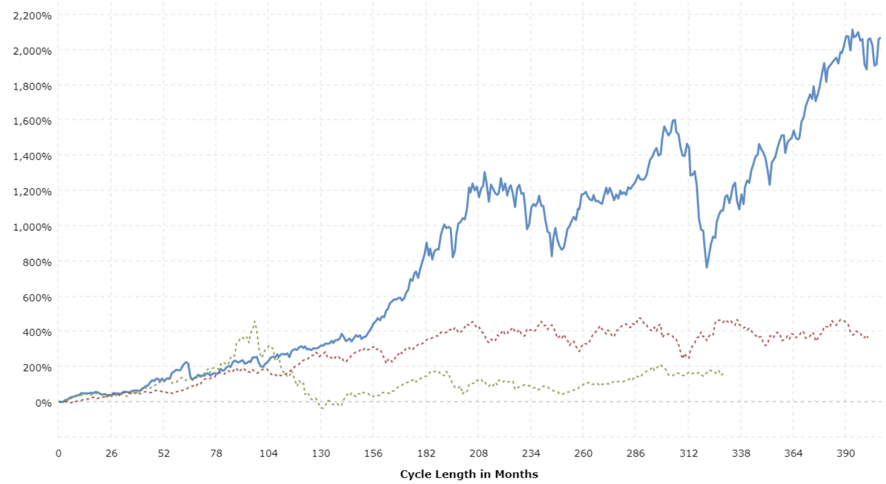
and many more…





What is Data: Others

Financial Market



Protein structure data

many more . . .

What is Data Science

There are many different definitions of Data Science

- Data Science is a Study of Data, same as Physical Sciences is study of physics, Biological science is study of biology.

  If we want to use them for making decisions, we need to study them

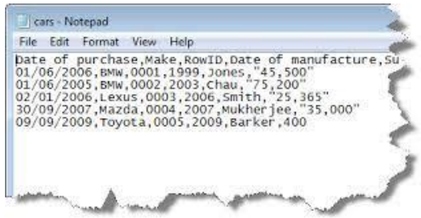
- Process the data to understand the world, uncovering the patterns and characteristics, to make strategic decisions

- Exploring, manipulating, processing data - Analysing the data to try to get some answers

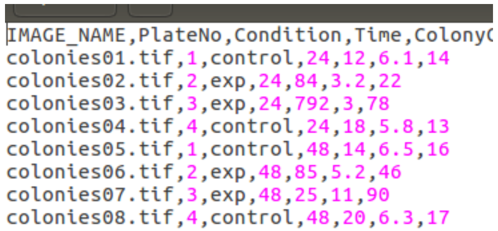
Why now: what changed

- Computers got cheaper
- Availability of Huge data
- Algorithms got fasters and efficient

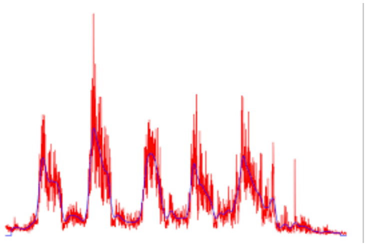
Data Analysis ~ Investigation

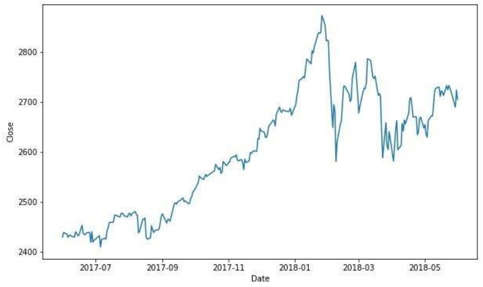


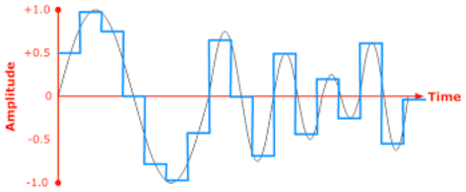
```
Date of purchase,Make,RowID,Date of manufacture,Su
01/06/2006,BMW,0001,1999,Jones,"45,500"
01/06/2005,BMW,0002,2003,Chau,"75,200"
02/01/2006,Lexus,0003,2006,Smith,"25,365"
30/09/2007,Mazda,0004,2007,Mukherjee,"35,000"
09/09/2009,Toyota,0005,2009,Barker,400
```

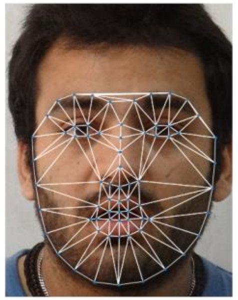


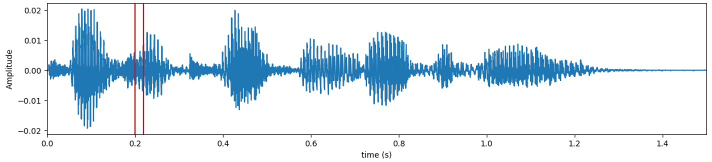
```
IMAGE_NAME,PlateNo,Condition,Time,ColonyC
colonies01.tif,1,control,24,12,6.1,14
colonies02.tif,2,exp,24,84,3.2,22
colonies03.tif,3,exp,24,792,3,78
colonies04.tif,4,control,24,18,5.8,13
colonies05.tif,1,control,48,14,6.5,16
colonies06.tif,2,exp,48,85,5.2,46
colonies07.tif,3,exp,48,25,11,90
colonies08.tif,4,control,48,20,6.3,17
```

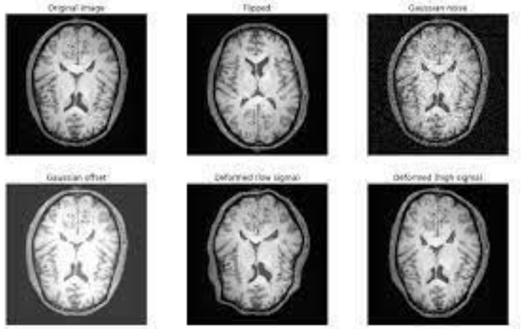


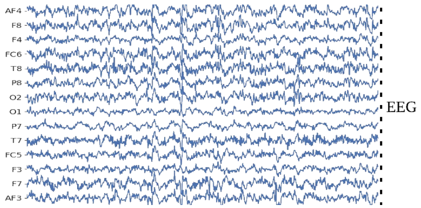


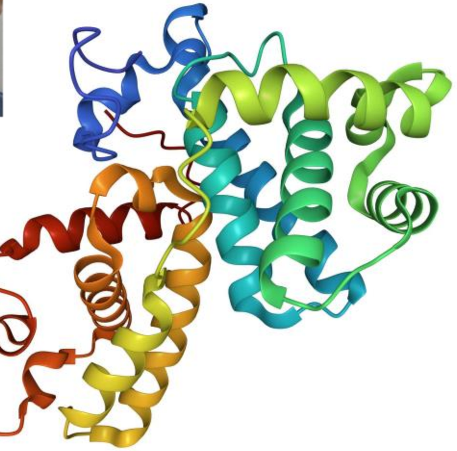












Lecture Outline

- About the module
- What is Data and Data science?
- Aim of Data Science Tasks
- Programming Languages and Tools
- Introduction to Python

Aim of a Task

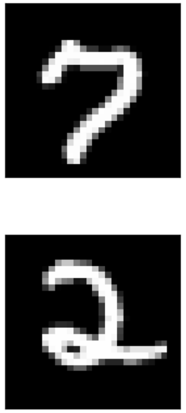
1. Automation of a task

A. Most humans can do it

A task that can be performed by most humans, mostly a simple task.

e.g., recognising cat/dog, segmentation, face recognition, happy vs sad, guessing price of a house, text to speech

Cat/Dog?



digit?

Speech to text

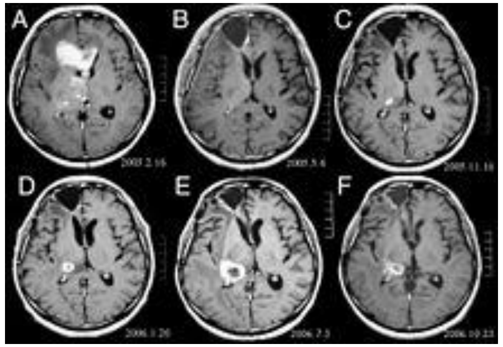
Aim of a Task

1. Automation of a task

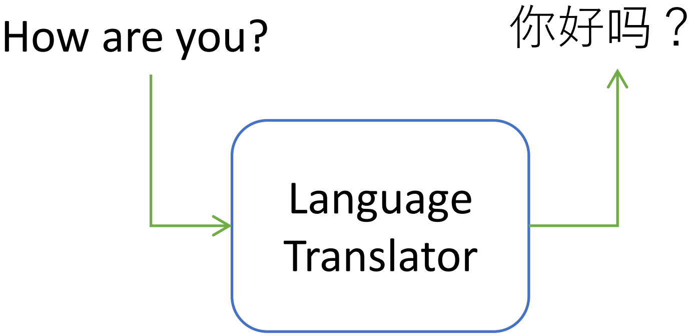
B. Experts can do it

Only trained experts can perform these tasks.

e.g., diagnosis of a disease using MRI/CT Scan, language translation,



Type of tumours


How are you?

Language Translator

你好吗？

Source: https://en.wikipedia.org/wiki/WHO_classification_of_tumours_of_the_central_nervous_system

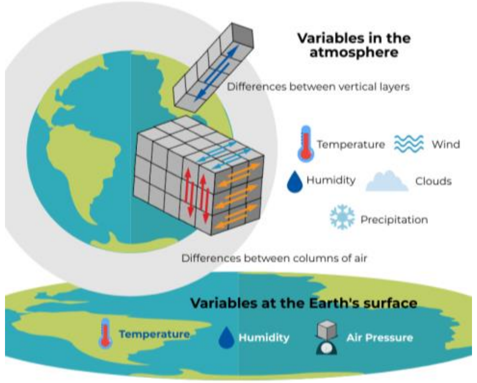
Aim of a Task

1. Automation of a task

C. Not even experts can do it

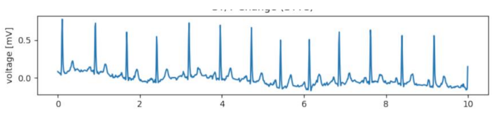
From given information, not even expert can performed these tasks, either due to enormous amount of data or unknown relations.

e.g. diagnosis of disease from limited information, or given enormous data to find the answer, weather prediction,


Variables in the atmosphere
Differences between vertical layers
Temperature
Wind
Humidity
Clouds
Precipitation
Differences between columns of air
Variables at the Earth's surface
Temperature
Humidity
Air Pressure

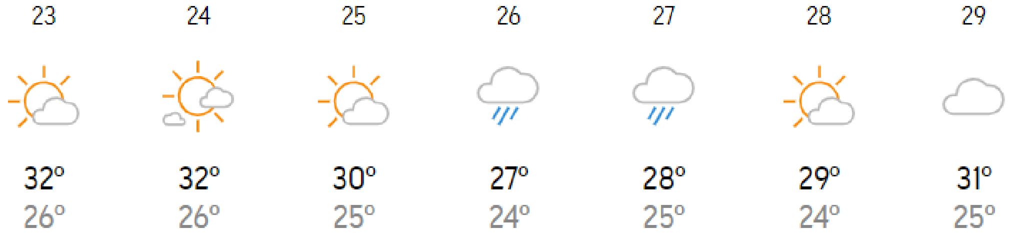
Weather Forecast

ECG



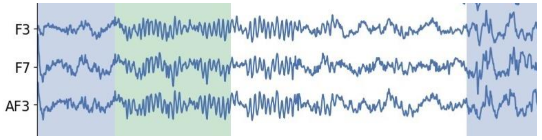
Mortality (Age)

Sanya, Hainan, China, April



23	24	25	26	27	28	29
32°	32°	30°	27°	28°	29°	31°
26°	26°	25°	24°	25°	24°	25°

EEG



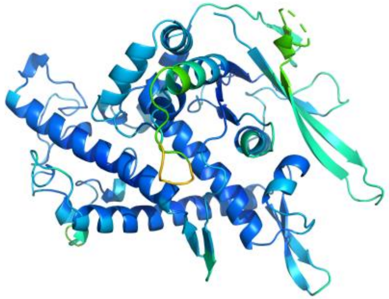
Thoughts?

Attention Level

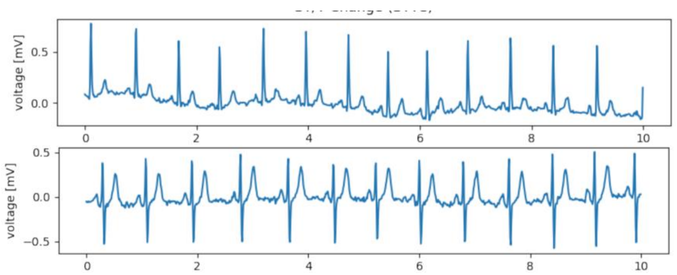
Aim of a Task

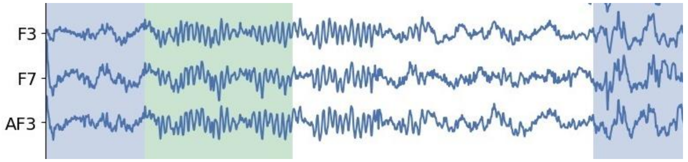
2. Discovering new knowledge

- From given information, the aim is find new relation between quantities/ characteristics (variables, features), which is often used to make decisions
- Classical science for doing same is = Research, Statistics
- For this often, automation step from 1C (building solutions) is used



Drug discovery





Mortality

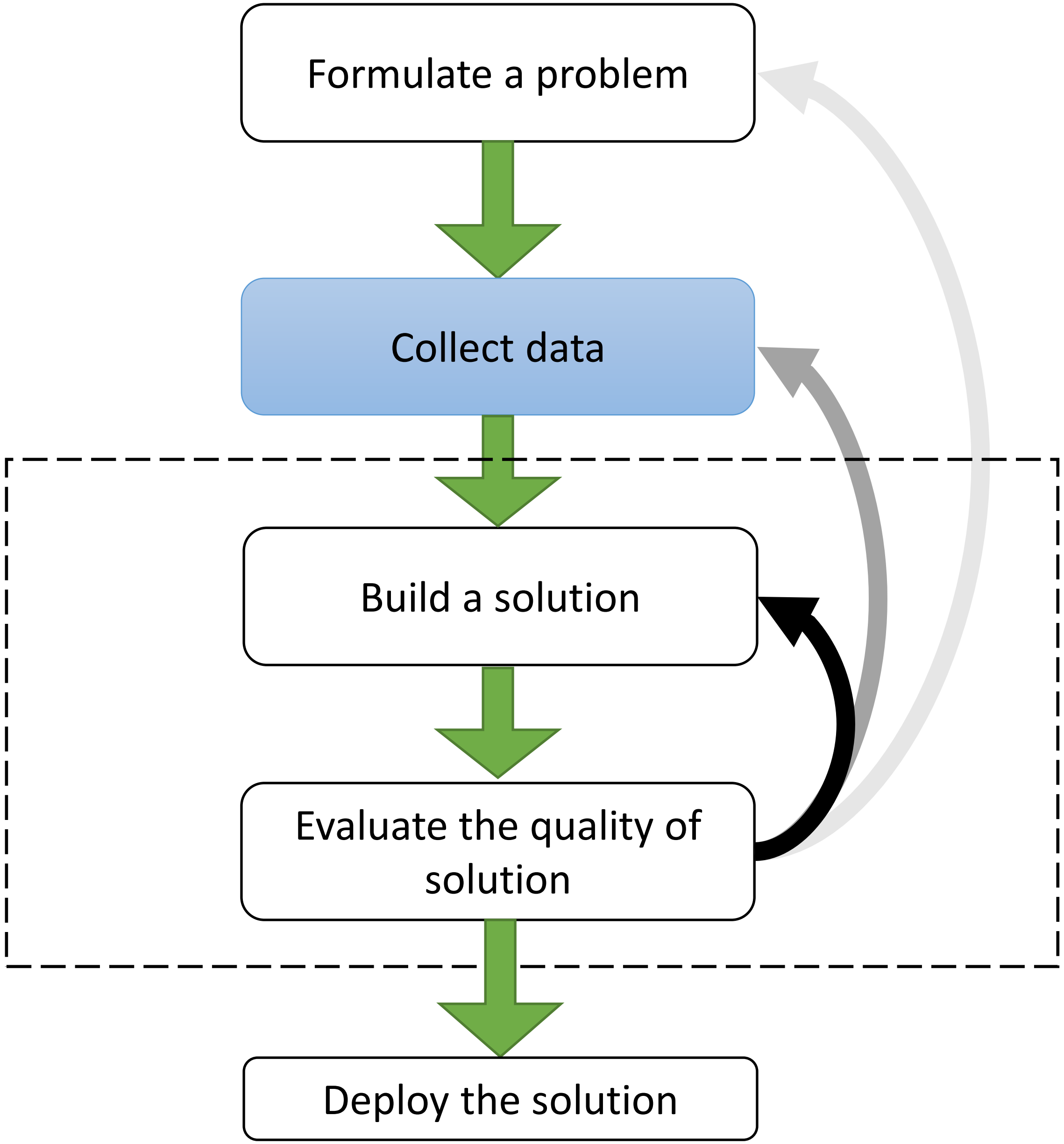
Attention

Workflow for a Data Science Task

- Typically, we start with formulating a problem, or choosing an objective of the task and collect appropriate data
- Most of the time spent:

  Building a solution and Evaluating its quality, if not happy, we go back to building a new solution, sometimes, we even go back to collect more data or change the problem statement

- Deploy the solution, once happy with its performance



Where we use it

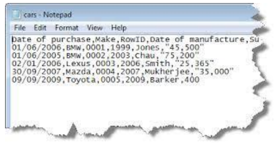
- Automation of tasks
  - Face ID, Recognition, Expression, Translation, Geo-tagging, object detection/tracking, ..
  - Deception Detection, automatic diagnosis of disease
- Financial Services

  - Stock market, better investment, good decisions
- Healthcare

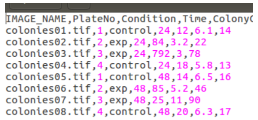
  - Improve diagnosis, new findings, better drugs, treatment, and management
- Pharmaceuticals: Drug Discovery
- Recommendation: TV, Movies, Product etc
- Generation : Generating Art/music/

How we do it – Computer Programming

Data



```
Date of purchase,Make,RowID,Date of manufacture,Su
01/06/2006,BMW,0001,1999,Jones,"45,500"
01/06/2005,BMW,0002,2003,Chau,"75,200"
02/01/2006,Lexus,0003,2006,Smith,"25,365"
30/09/2007,Mazda,0004,2007,Mukherjee,"35,000"
09/09/2009,Toyota,0005,2009,Barker,400
```



```
IMAGE_NAME,PlateNo,Condition,Time,ColonyC
colonies01.tif,1,control,24,12,6.1,14
colonies02.tif,2,exp,24,84,3.2,22
colonies03.tif,3,exp,24,792,3,78
colonies04.tif,4,control,24,18,5.8,13
colonies05.tif,1,control,48,14,6.5,16
colonies06.tif,2,exp,48,85,5.2,46
colonies07.tif,3,exp,48,25,11,90
colonies08.tif,4,control,48,20,6.3,17
```

+

Machine/Computer

+

Programming

Lecture Outline

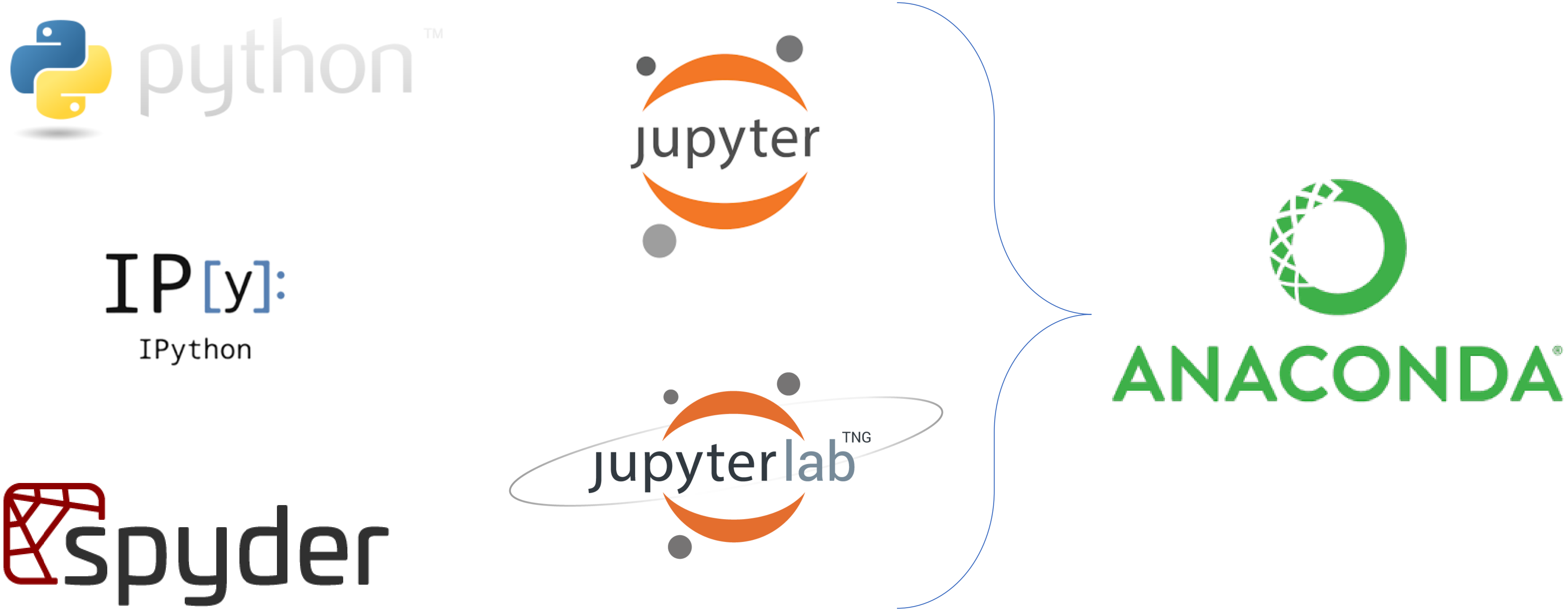
- About the module
- What is Data and Data science?
- Aim of Data Science Tasks
- Programming Languages and Tools
- Introduction to Python

Programming Languages and Tools for Data Science

- MATLAB
- R (Studio)
- **Python (Most popular)**
- Apache Hadoop
- Apache Spark
- SQL
- Docker
- Azure
- .
- … more

Programming in the module

We will be using **Python**, and particularly:



What is Python & Anaconda

Python

- Python is a general purpose, object-oriented, high-level programming language. It was created by Guido van Rossum (1991). It is used for software development, scripting, mathematics and very popular for Data Science Development and Research

Anaconda

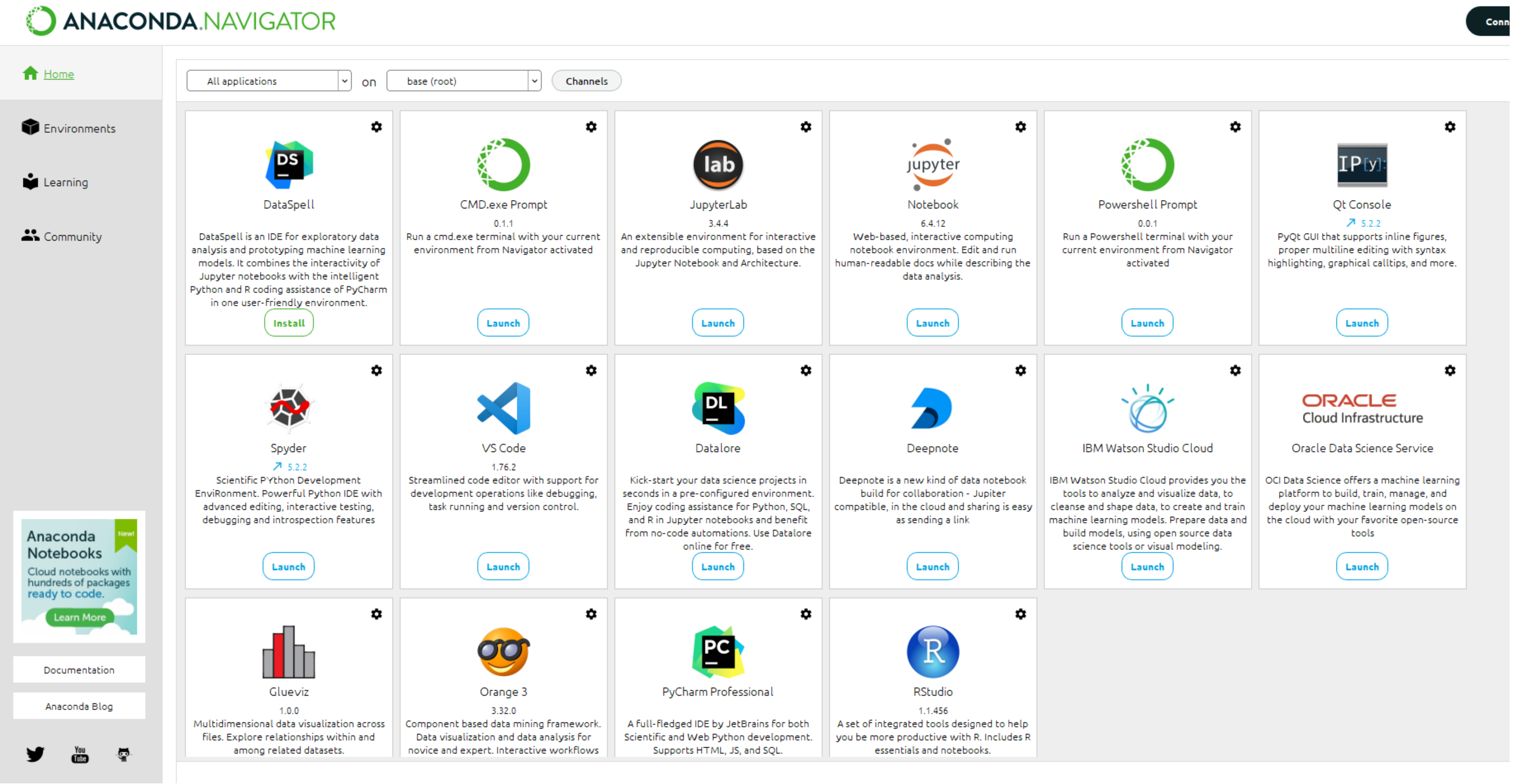
- Anaconda is a distribution of many programming languages like Python and R and tools. It uses an open-source package and environment management system called ***Conda,*** which makes it easy to install/update packages and manage environments.

Why Python?

- Open source
    - Freely available to use
- A large community:
    - Many people around the world use it, can help resolving issues
- Simple, easy and very powerful language
    - Can handle large amount of data, integrate with different HW, Software, Web, etc.

- For Data Science
    - Large number of tools for data science are freely available,
    - Any development of tool/algorithm/model can be integrated to other systems
    - Nicer visualization, documentation, and Reports can be build (Jupyter-notebook)

A walk through Anaconda



DataSpell

DataSpell is an IDE for exploratory data analysis and prototyping machine learning models. It combines the interactivity of Jupyter notebooks with the intelligent Python and R coding assistance of PyCharm in one user-friendly environment.

CMD.exe Prompt

0.1.1

Run a cmd.exe terminal with your current environment from Navigator activated

JupyterLab

3.4.4

An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.

Notebook

6.4.12

Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.

Powershell Prompt

0.0.1

Run a Powershell terminal with your current environment from Navigator activated

Qt Console

↗ 5.2.2

PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Spyder

↗ 5.2.2

Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features

VS Code

1.76.2

Streamlined code editor with support for development operations like debugging, task running and version control.

Datalore

Kick-start your data science projects in seconds in a pre-configured environment. Enjoy coding assistance for Python, SQL, and R in Jupyter notebooks and benefit from no-code automations. Use Datalore online for free.

Deepnote

Deepnote is a new kind of data notebook build for collaboration - Jupiter compatible, in the cloud and sharing is easy as sending a link

IBM Watson Studio Cloud

IBM Watson Studio Cloud provides you the tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data science tools or visual modeling.

ORACLE Cloud Infrastructure

Oracle Data Science Service

OCI Data Science offers a machine learning platform to build, train, manage, and deploy your machine learning models on the cloud with your favorite open-source tools

Glueviz

1.0.0

Multidimensional data visualization across files. Explore relationships within and among related datasets.

Orange 3

3.32.0

Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows

PyCharm Professional

A full-fledged IDE by JetBrains for both Scientific and Web Python development. Supports HTML, JS, and SQL.

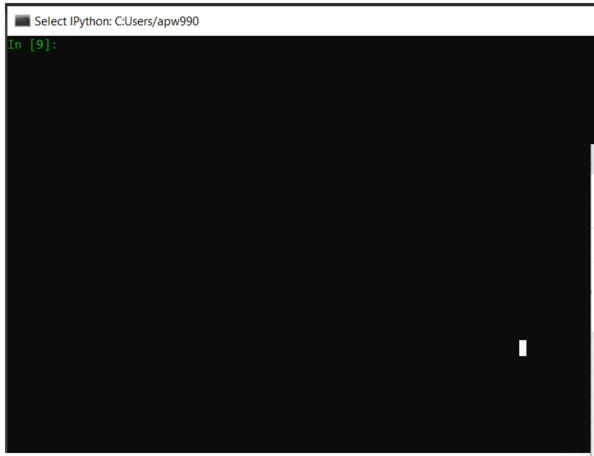
RStudio

1.1.456

A set of integrated tools designed to help you be more productive with R. Includes R essentials and notebooks.

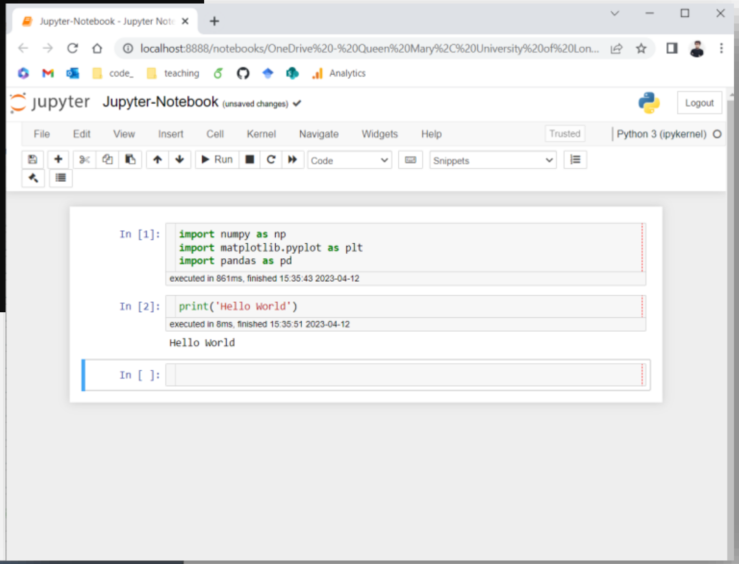
Python: Interfaces

- IPython Terminal
  - type: ‘ipython’ on terminal (any OS)



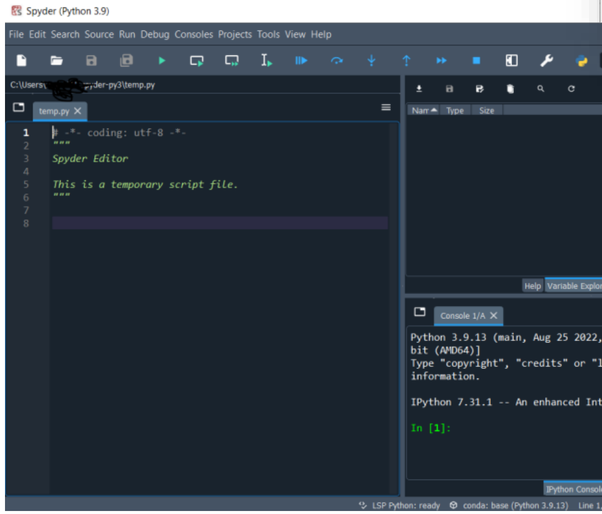
```
In [9]:
```

- **Jupyter-Notebook:**
  - type: ‘jupyter-notebook’ on anaconda terminal


Jupyter-Notebook - Jupyter Note
localhost:8888/notebooks/OneDrive%20-%20Queen%20Mary%2C%20University%20of%20Lon...
code_
teaching
Analytics
jupyter Jupyter-Notebook (unsaved changes)
Logout
File Edit View Insert Cell Kernel Navigate Widgets Help
Trusted
Python 3 (ipykernel)
Run
Code
Snippets

```
In [1]: import numpy as np
        import matplotlib.pyplot as plt
        import pandas as pd
executed in 861ms, finished 15:35:43 2023-04-12

In [2]: print('Hello World')
executed in 8ms, finished 15:35:51 2023-04-12
Hello World

In [ ]:
```

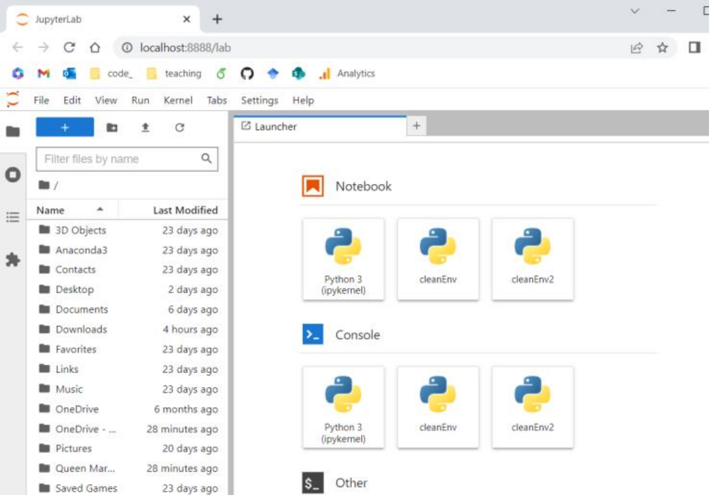
- Spyder
  - type: ‘spyder’ on anaconda terminal


Spyder (Python 3.9)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\[redacted]yder-py3\temp.py
temp.py
Name Type Size Value
Help Variable Explorer
Console 1/A
IPython Console
LSP Python: ready
conda: base (Python 3.9.13)
Line 1, Col

```
# -*- coding: utf-8 -*-
"""
Spyder Editor

This is a temporary script file.
"""

```

```
Python 3.9.13 (main, Aug 25 2022, 23
bit (AMD64)]
Type "copyright", "credits" or "lice
information.

IPython 7.31.1 -- An enhanced Intera

In [1]:
```

- Jupyter-Lab
  - type: ‘jupyter-lab’ on anaconda terminal



Name	Last Modified
3D Objects	23 days ago
Anaconda3	23 days ago
Contacts	23 days ago
Desktop	2 days ago
Documents	6 days ago
Downloads	4 hours ago
Favorites	23 days ago
Links	23 days ago
Music	23 days ago
OneDrive	6 months ago
OneDrive - ...	28 minutes ago
Pictures	20 days ago
Queen Mar...	28 minutes ago
Saved Games	23 days ago

All can be opened from Anaconda Navigator

Lecture Outline

- About the module
- What is Data and Data science?
- Aim of Data Science Tasks
- Programming Languages and Tools
- Introduction to Python

Python: As a calculator

- IPython –terminal
  - type: ‘ipython’ on terminal (any OS)

- 344+344
- 34-5
- 2345/232
- 234*4354
- 355 + (434*234) – (233-2)

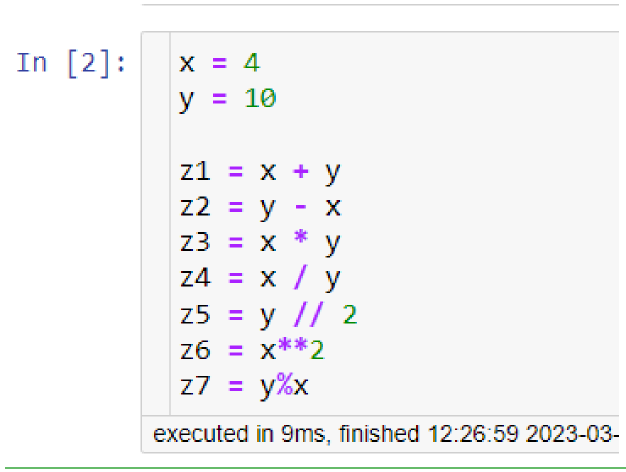
- - need to create variables…??

```
In [9]:
```

Ref 2.1-2.4

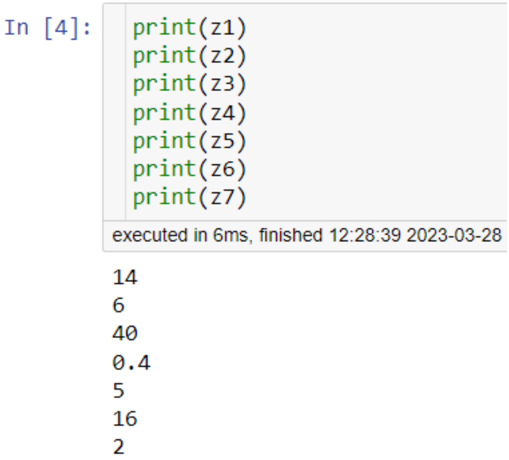
Python: Arithmetic + use of variable

Use of variables: x, y, z



```
In [2]:  x = 4
         y = 10

         z1 = x + y
         z2 = y - x
         z3 = x * y
         z4 = x / y
         z5 = y // 2
         z6 = x**2
         z7 = y%x
```

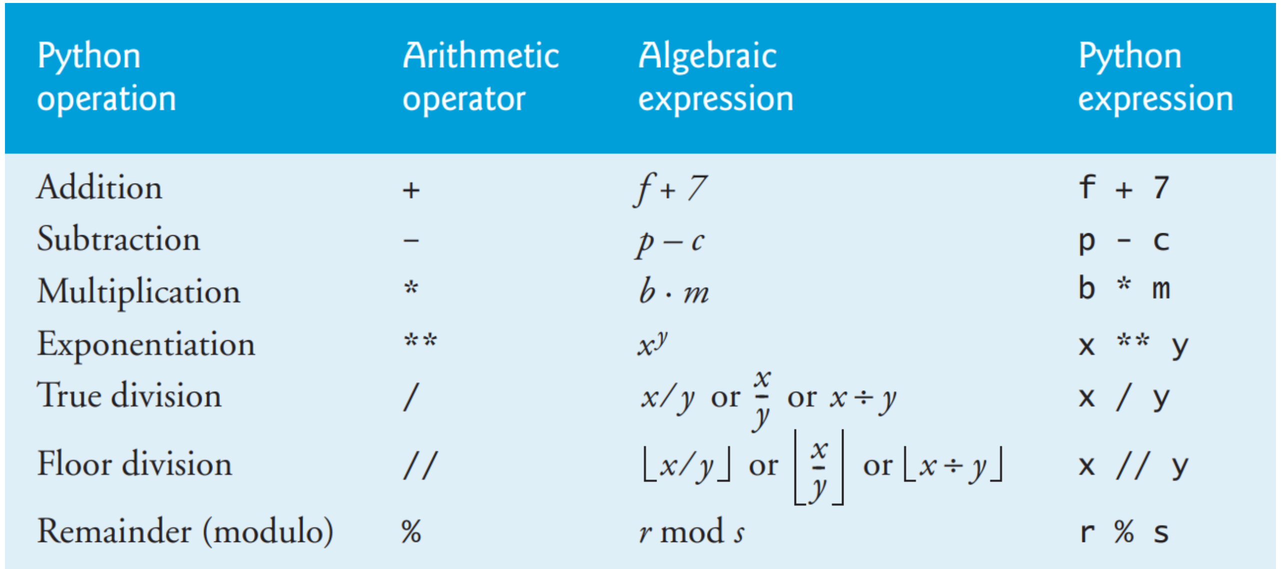
executed in 9ms, finished 12:26:59 2023-03-



```
In [4]:  print(z1)
          print(z2)
          print(z3)
          print(z4)
          print(z5)
          print(z6)
          print(z7)
```

executed in 6ms, finished 12:28:39 2023-03-28

```
14
6
40
0.4
5
16
2
```



Python operation	Arithmetic operator	Algebraic expression	Python expression
Addition	+	$f+7$	`f + 7`
Subtraction	–	$p-c$	`p - c`
Multiplication	*	$b \cdot m$	`b * m`
Exponentiation	**	x^y	`x ** y`
True division	/	x/y or $\frac{x}{y}$ or $x \div y$	`x / y`
Floor division	//	$\lfloor x/y \rfloor$ or $\left\lfloor \frac{x}{y} \right\rfloor$ or $\lfloor x \div y \rfloor$	`x // y`
Remainder (modulo)	%	$r \bmod s$	`r % s`

```
>>> y = (a * (x ** 2)) + (b * x) + c
>>> print(y)
```

What happens?

x/0
y/0

Python: Data Types : Basics

How do you check Data Type of given variable x?

>>> type(x)

- Basic
  - **Numerical**:
    Integers (int),
    fractions (float)
    complex
  - **Strings**:
    str
  - **Boolean**
    True/False
  - **NoneType**
    None

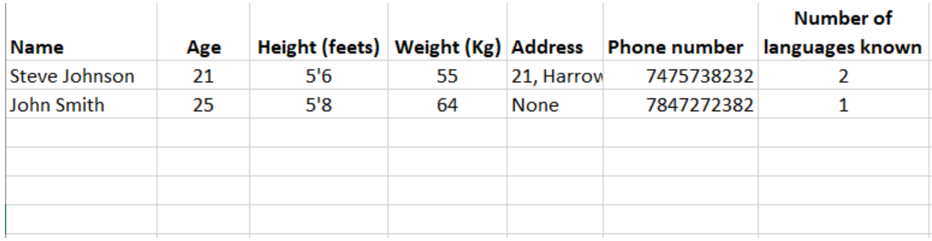
Objects

- Examples

x = 23
y = 45
z = 1010
x = 24.5023
y = 0.0113
z = 1.0
x = 2 + i1

S1 = ‘Hello World’ or “Hello World”
S2 = “H”

x = True
y = False
x = None



Name	Age	Height (feets)	Weight (Kg)	Address	Phone number	Number of languages known
Steve Johnson	21	5'6	55	21, Harrow	7475738232	2
John Smith	25	5'8	64	None	7847272382	1

Python: Data Types : Basics

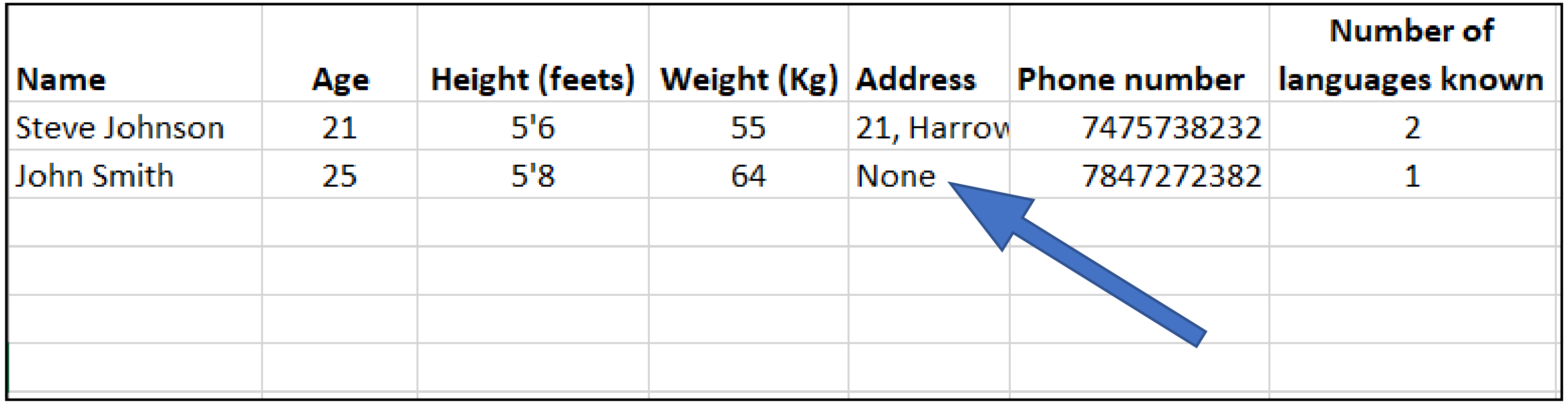
How do you check Data Type of given variable x?

>>> type(x)

- Basic
  - **Numerical**: Integers (int), fractions (float) complex
  - **Strings**: str
  - **Boolean** True/False
  - **NoneType** None

Objects

Data can have collection of values with different data types

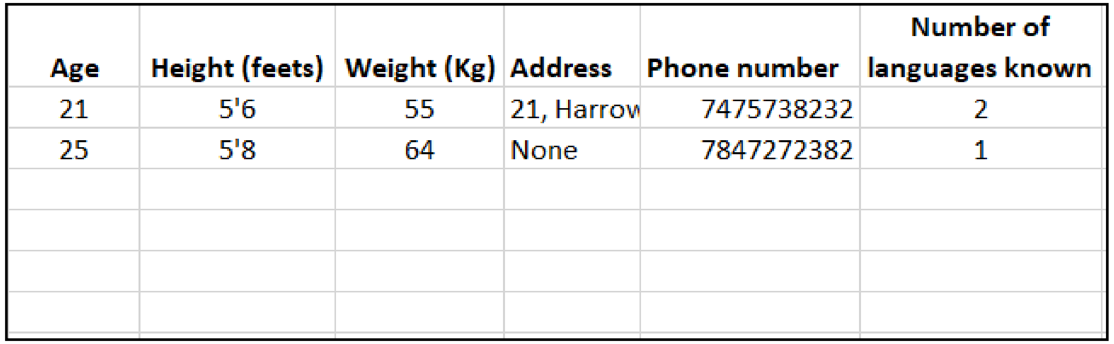


Name	Age	Height (feets)	Weight (Kg)	Address	Phone number	Number of languages known
Steve Johnson	21	5'6	55	21, Harrow	7475738232	2
John Smith	25	5'8	64	None	7847272382	1

Task 1: Collect Data from 10 persons

Task 1: Collect **data of 10 people**, that you know, *outside of this class*.
You need following information about them, with no **‘NAME’**

1. Age
2. Height (in cm)
3. Weight (in Kg)
4. Address (town/city, not exact address)
5. Education Level (School/UG/PG/PhD)
6. Number of Languages known
7. Like Volleyball?
8. Like Table Tennis?
9. According to you, how happy are they in their life? (from 0 to 10)
10. Name of a favourite movie

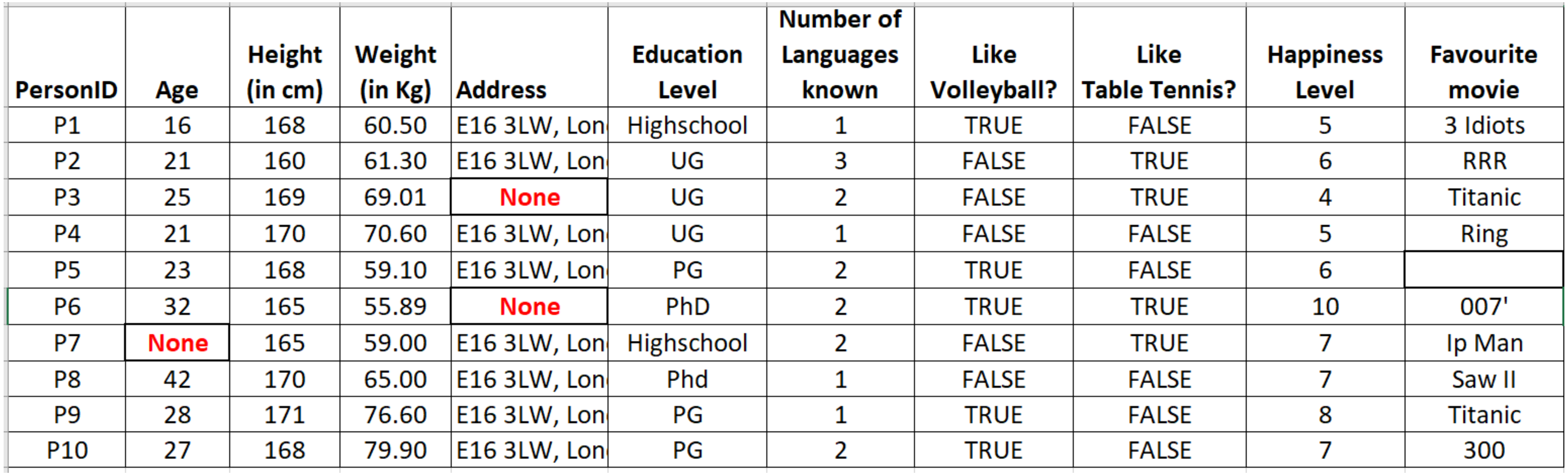


Age	Height (feets)	Weight (Kg)	Address	Phone number	Number of languages known
21	5'6	55	21, Harrow	7475738232	2
25	5'8	64	None	7847272382	1

Task 1: Collect Data from 10 persons

- Your collected data should look like this

Table 1.1: Example of Dataset



PersonID	Age	Height (in cm)	Weight (in Kg)	Address	Education Level	Number of Languages known	Like Volleyball?	Like Table Tennis?	Happiness Level	Favourite movie
P1	16	168	60.50	E16 3LW, Lon	Highschool	1	TRUE	FALSE	5	3 Idiots
P2	21	160	61.30	E16 3LW, Lon	UG	3	FALSE	TRUE	6	RRR
P3	25	169	69.01	None	UG	2	FALSE	TRUE	4	Titanic
P4	21	170	70.60	E16 3LW, Lon	UG	1	FALSE	FALSE	5	Ring
P5	23	168	59.10	E16 3LW, Lon	PG	2	TRUE	FALSE	6	
P6	32	165	55.89	None	PhD	2	TRUE	TRUE	10	007'
P7	None	165	59.00	E16 3LW, Lon	Highschool	2	FALSE	TRUE	7	Ip Man
P8	42	170	65.00	E16 3LW, Lon	Phd	1	FALSE	FALSE	7	Saw II
P9	28	171	76.60	E16 3LW, Lon	PG	1	TRUE	FALSE	8	Titanic
P10	27	168	79.90	E16 3LW, Lon	PG	2	TRUE	FALSE	7	300

String	Integer	Integer	Float	String	String	Integer	Boolean	Boolean	Integer	String

Task 1: Collect Data from 10 persons

- Collect the data from 10 close persons of yours, outside of this class.
- Try to avoid **None** or missing data values
- Store it in your laptop/computer.
- Submit this data on QM+

  full instructions on how to submit will be given on QM+
- We will use this collected data in some of the lab assignments and coursework

DO NOT submit random data entries or copy data of any other student, we will know!!!!

Python: Data Types
Operations on Basics

How do you check Data Type of given variable x?

```
>>> type(x)
```

- Basic
  - Addition : x+y
  - Subtraction : x-y
  - Multiplication : x*y
  - Division : x/y
  - Power : x**y

For x:int or x:float, y:int or y:float

Exceptions: x/y for y=0

x = 4

y = ‘hello’

Multiplication :x*y

Only If one of variable is string and other in int

Addition : x+y

Only If both variables are string type

Python: Data Types Operations NoneType

How do you check Data Type of given variable x?

```
>>> type(x)
```

- Basic
  - ~~Addition : x+y~~
  - ~~Subtraction : x-y~~
  - ~~Multiplication : x*y~~
  - ~~Division : x/y~~
  - ~~Power : x**y~~

If one of variable is NoneType

- Next !!!

  - 1.2: Getting Started

    - Installing Anaconda
    - Python Interfaces
    - Python as calculator
    - Jupyter-notebook

  - 1.3: On Collection(s) of data

Queen Mary
University of London