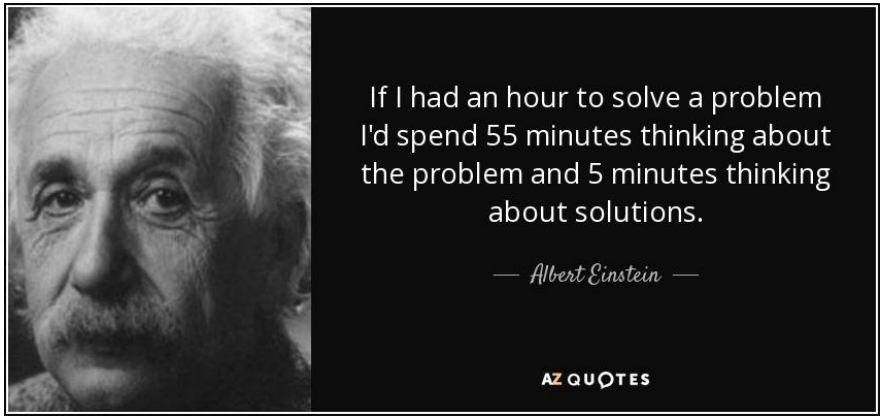
EXPLORE | DIGITAL SKILLS

Python for Data Science

Course Introduction

Using Programming to Solve Problems

Programming is problem solving



- Programming is a toolbox for **solving problems**
- Programming is **chunking a problem** into smaller units which are easy to solve
- Good programming allows us to solve a problem once, and **apply that solution in solving other problems.**

Implementing cool algorithms

- A lot of problems **have already been solved**, we just need to know where to look
- **Implementation is key** – waiting sucks!
- The real skill is to be able to tweak them to use for a specific purpose.

EXPLORE | DIGITAL SKILLS

What is an Algorithm?

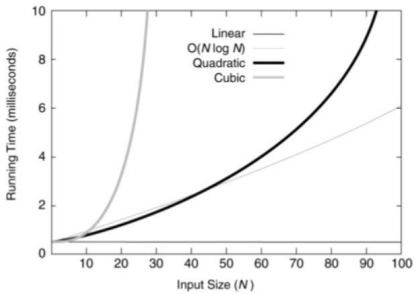
An **ALGORITHM** is a set of instructions to be followed to solve a problem

Why do complexity analysis?

There are often many different algorithms which can be used to solve the same problem.

Thus, it makes sense to develop techniques that allow us to:

- compare different algorithms based on their "efficiency"
- choose the most efficient algorithm for the problem



How do we measure efficiency?

The efficiency of any algorithmic solution is a measure of the:

- Time efficiency: the time it takes to execute
- Space efficiency: space/memory it uses

What is Big O Notation?

Big O Notation represents the (order of the) maximum number of steps that an algorithm would require to find a solution to a problem, as a function of the input size.

Constant Time	Logarithmic Time	Linear Time	Polynomial Time	Exponential Time
O(1)	O(logN)	O(n)	$O(n^p)$	$O(e^N)$

An Introduction to Python

Python is an interpreted, high-level and general-purpose programming language.

Python Programming

Why Python?

- Simple, versatile and easy to maintain
- Runs on an interpreter system, faster execution of code

Popularity behind Python

- More productive
- Rich set of libraries and frameworks
- Large community

Python IDE's

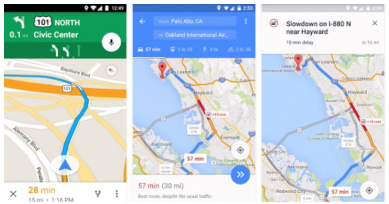
- PyCharm
- Spyder
- IDLE

Applications of Python

Web applications

Finance and trading

Mapping and geography



Data science

EXPLORE | DIGITAL SKILLS

All you need to do data science in Python…

Python is a programming language that lets you work quickly and integrate systems more effectively.
We will be using Python 3+ for the rest of this course.

ANACONDA®

Anaconda is the world's most popular Python data science platform.
Conda is a package manager to make sure all dependencies are managed so that everything works!

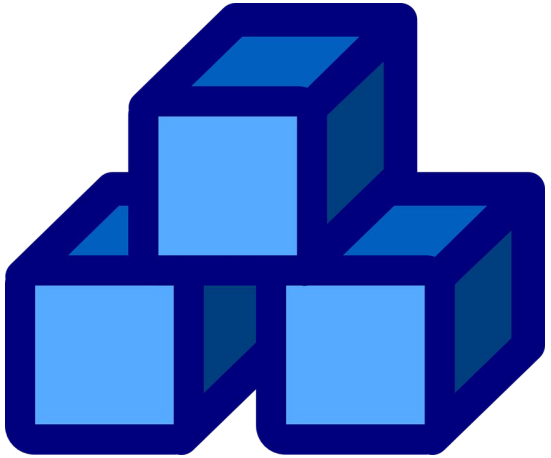
Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages

EXPLORE DIGITAL SKILLS

Course Components: Basic building blocks of Python

In this course, we will be covering the basic building blocks of Python for data science that will set us on a smooth trajectory into data science. Below are some of the concepts we will cover:

What will we cover?	Purpose in Python
Basic functions	Set of instructions that are executed when called
PEP8 coding style	Write clean and efficient code
Primitive data structures	Organize and store data that is easily accessible
Modules and Packages	Allows logical organization of code
Recursive functions	Reduces complexity of a task
Python scripting	Compile and execute Python code



EXPLORE | DIGITAL SKILLS

Course Components: Data wrangling with Python

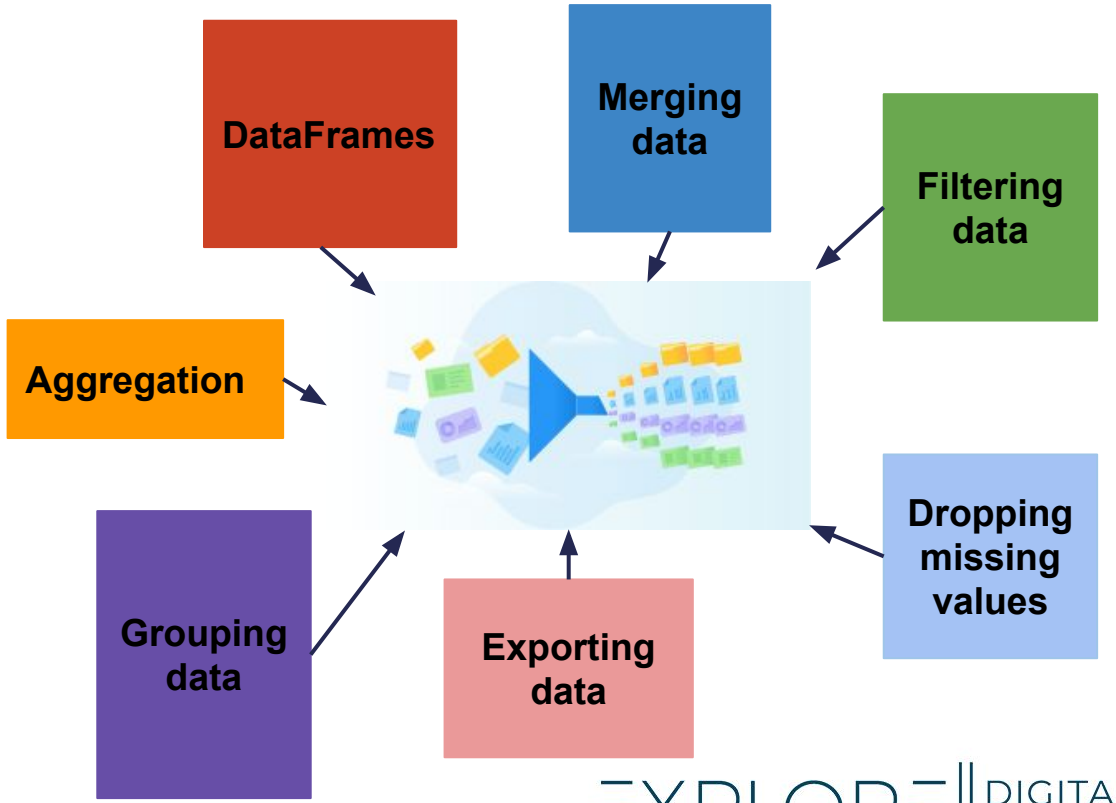
We will explore Python's built-in features that can be integrated for data wrangling.

Data wrangling with Pandas

- DataFrame creation
- Data manipulation
- Data transformation

Data wrangling with Numpy

- Multi-dimensional arrays and matrices
- Perform mathematical operation on entire dataset.



EXPLORE | DIGITAL SKILLS

Course Components: Statistical analysis with Python

We will also be exploring how to integrate the use of Python's built-in libraries for statistical analysis. Below are some of the concepts we will navigate through in the course:

What will we cover?

- How to differentiate between Inferential and Descriptive Statistics
- How to identify the different variable and data types
- How to Differentiate between Sample and Population Metrics
- How to use the scipy package for statistical analysis
- How to fit data to probability distributions

EXPLORE | DIGITAL SKILLS

What will you be doing in this course?

Learning

Resources

General background reading

1

Trains

Learning specific skills

13

Assessment

Tests

Testing the skills you have learnt

12

Project

Move into a real-world data science job!

You are a Data Scientist at Eskom.

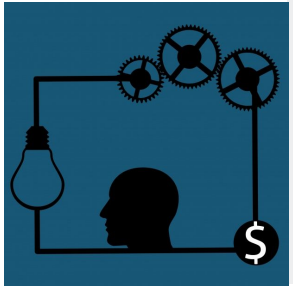
Eskom requires certain metrics to be calculated for their analytics team

EXPLORE | DIGITAL SKILLS

Your Project | Building Functions to Calculate Metrics using Eskom Data

Problem Statement: You need to build python functions which calculate/analyse data from Eskom

Your Context



- You are a **Data Scientist at Eskom**
- **Eskom** requires certain **metrics** to be calculated for their analytics team

Your role

- **Build 7 functions** using **python**
- These **7 functions** will need to process both **numeric & text data**

Stuff you need to Know

- List manipulation
- Dictionaries
- Basic Statistics and Aggregations
- Function definitions
- PEP8 coding style

Your Job

- **Write 7 functions** which outputs **metrics**
- **Submit functions** at the end of the sprint

EXPLORE | DIGITAL SKILLS

Your Project | Building Functions to Calculate Metrics using Eskom Data

For this project your main task will be to build a **module** which looks at the following:

	What are you building?	What do you need to do?
1 Numerical Metrics (Statistics & Time Series analysis)	**Functions** which take in a list or a pandas dataframe and returns either a **dictionary** or a **list**	**Write and submit 3 functions** of increasing difficulty: ➢ **2 Functions** on **calculated metrics** ➢ **1 Function** on **time data**
2 Twitter Data Processing (Text/Language)	**Functions** which take in text data as either a list or as a dataframe and returns a **list or pandas dataframe**	**Write and submit 4 functions** of increasing difficulty

EXPLORE DIGITAL SKILLS