Data Analysis Workshop - 1

Shitanshu Kusmakar

Risk Acquisitions, ANZ

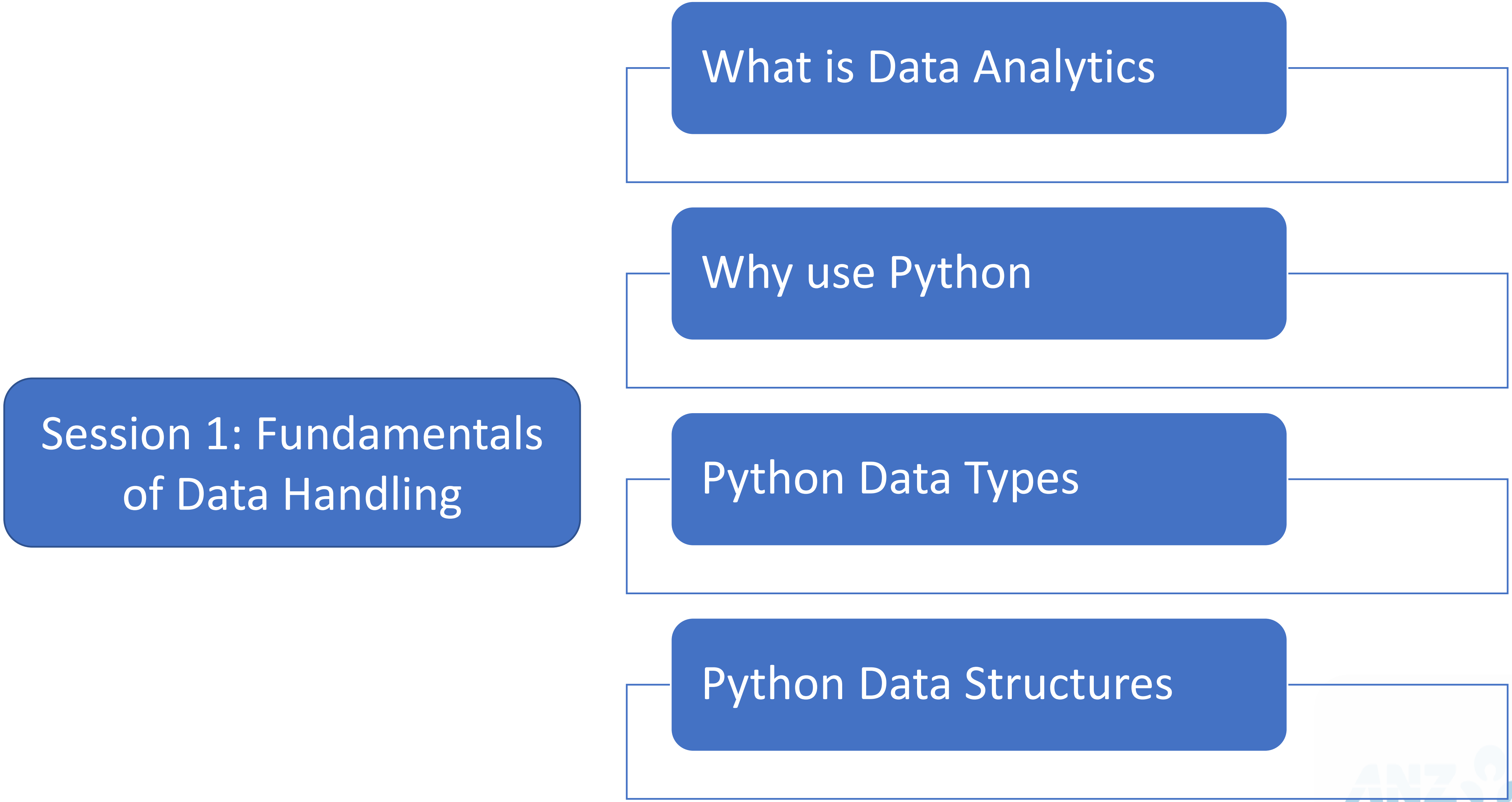
Outline

Session 1: Fundamentals of Data Handling

Session 2: Data Processing & Exploration

Session 3: Data Visualization

Session 4: Time-series Analysis



Session 1: Fundamentals of Data Handling

- What is Data Analytics
- Why use Python
- Python Data Types
- Python Data Structures

Data Analytics

Descriptive What happened?

- Comprehensive accurate live data analysis & visualization

Business Intelligence

Dashboards

Diagnostic Why did it happen?

- Ability to identify the root cause & remove confounding information

Simulation

Data mining

Predictive What will happen?

- Identify patterns using historical data
- Automation using ML ops and algorithms

Simulation

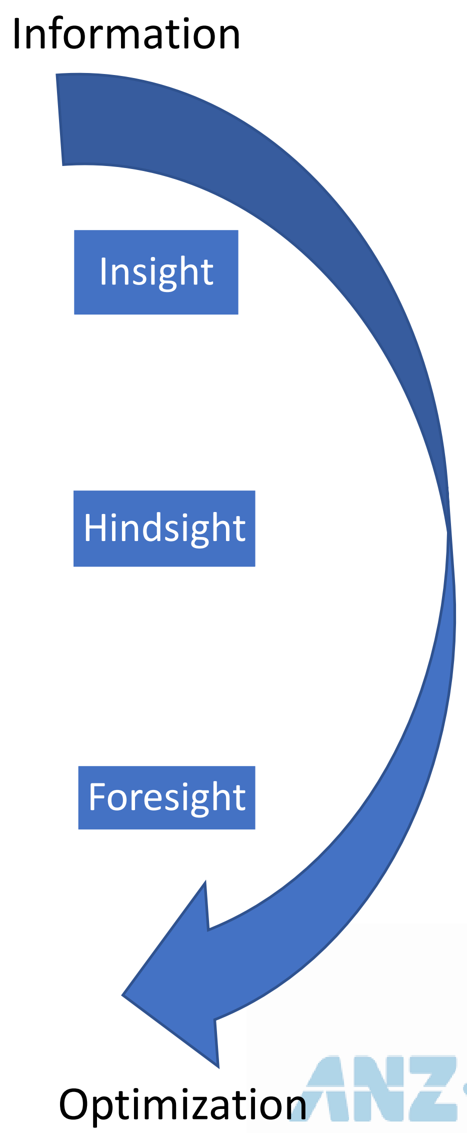
Data mining

Prescriptive How can we make it happen?

- Advanced techniques & algorithms for recommendations

Decision Tress

Mathematical Modelling



Information

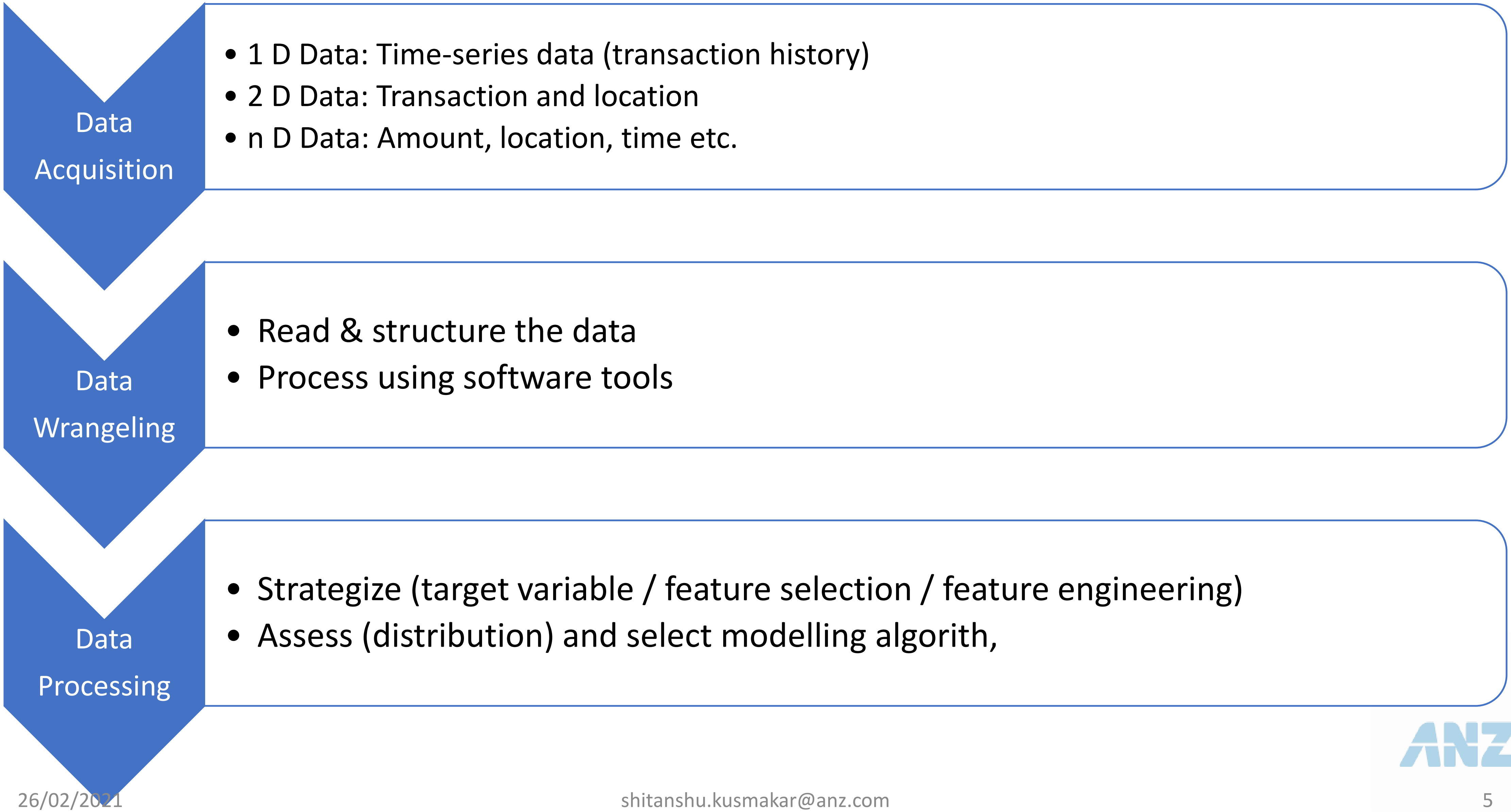
Insight

Hindsight

Foresight

Optimization

Data Analysis Pipeline



Data Acquisition

- 1 D Data: Time-series data (transaction history)
- 2 D Data: Transaction and location
- n D Data: Amount, location, time etc.

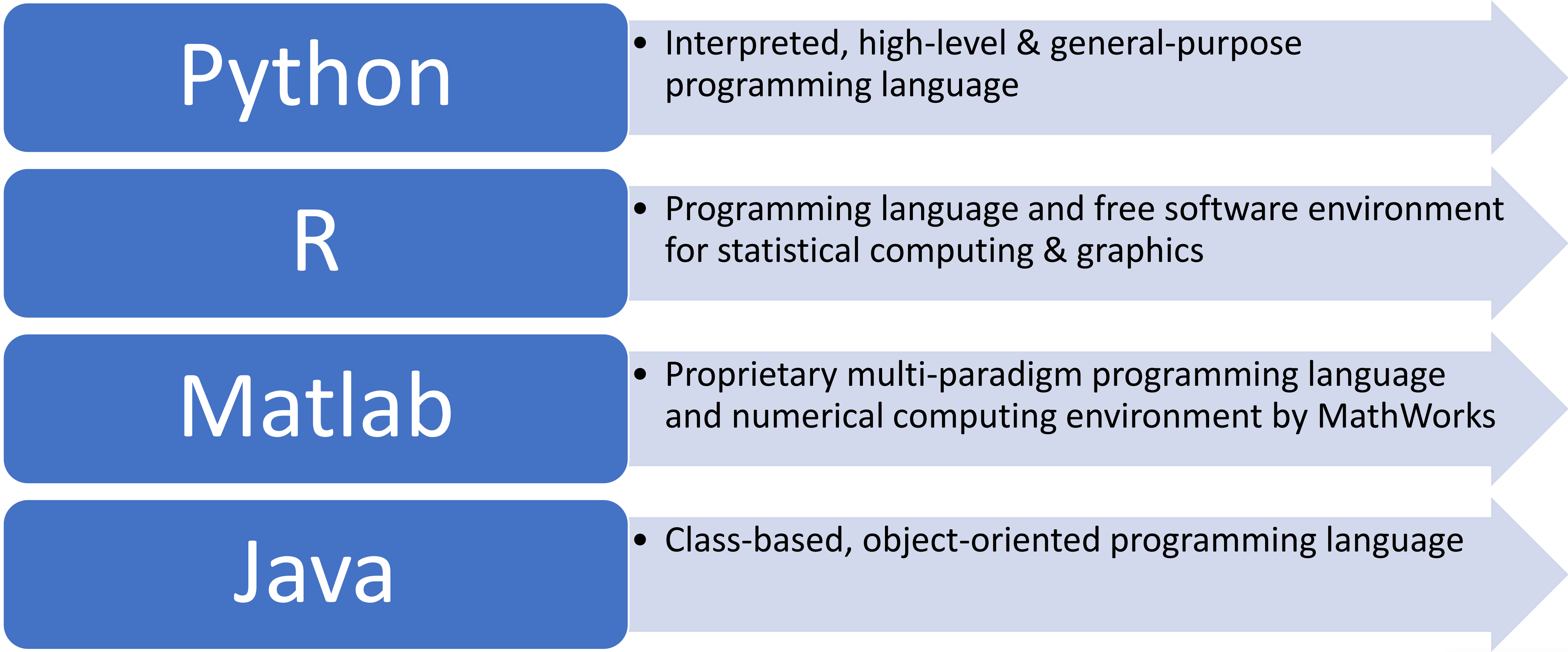
Data Wrangeling

- Read & structure the data
- Process using software tools

Data Processing

- Strategize (target variable / feature selection / feature engineering)
- Assess (distribution) and select modelling algorith,

Tools for Analysis



Python

- Interpreted, high-level & general-purpose programming language

R

- Programming language and free software environment for statistical computing & graphics

Matlab

- Proprietary multi-paradigm programming language and numerical computing environment by MathWorks

Java

- Class-based, object-oriented programming language

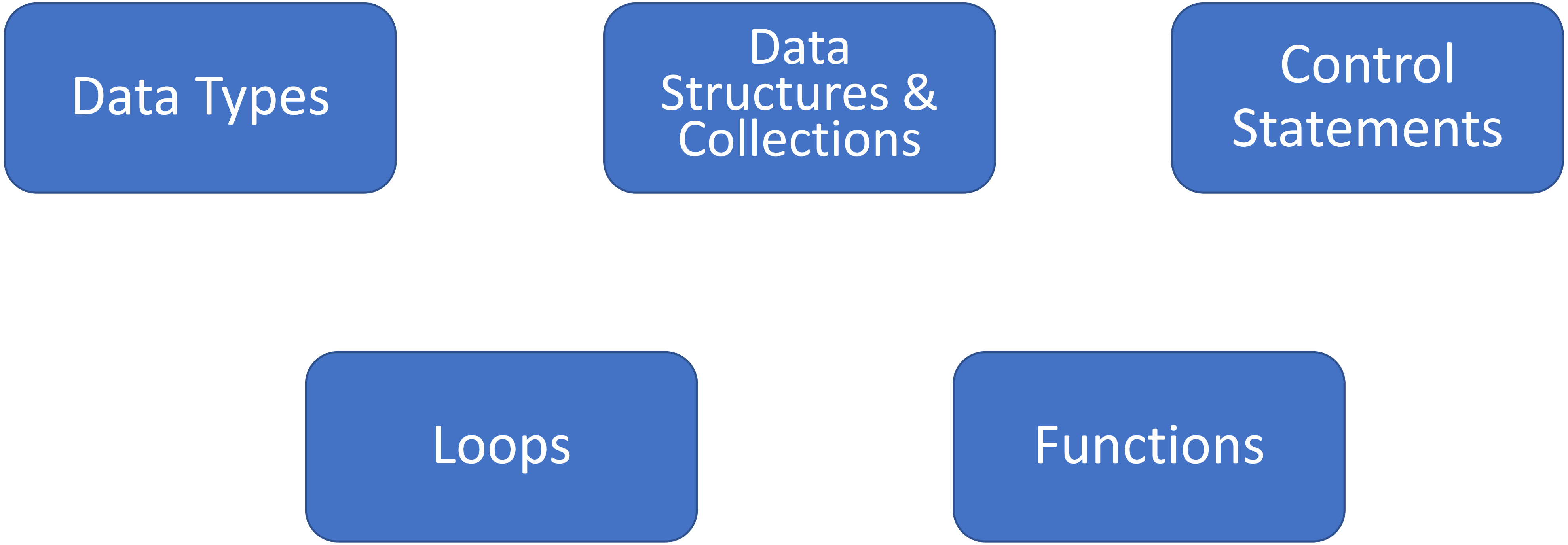
Why Python?

- High Level
- Open Source
- Large standard library
- Interpreted
- Object oriented
- Faster & scalable
- Powerful for scientific computing

Python Programming

- Cross platform compatible libraries on Unix, Windows, Macintosh
- Jupyter Notebook: Open-source web application. Allows to create and share document that contains code, equations, visualizations and comments. Type .ipynb files.
- Anaconda: Free and open source distribution of Python. Simplifies package management and deployment. Includes wide array of data science libraries
- Other IDE's: PyCharm, Spyder etc.
- JupyterLab: Web based user interface of Jupyter

Fundamentals



Data Types

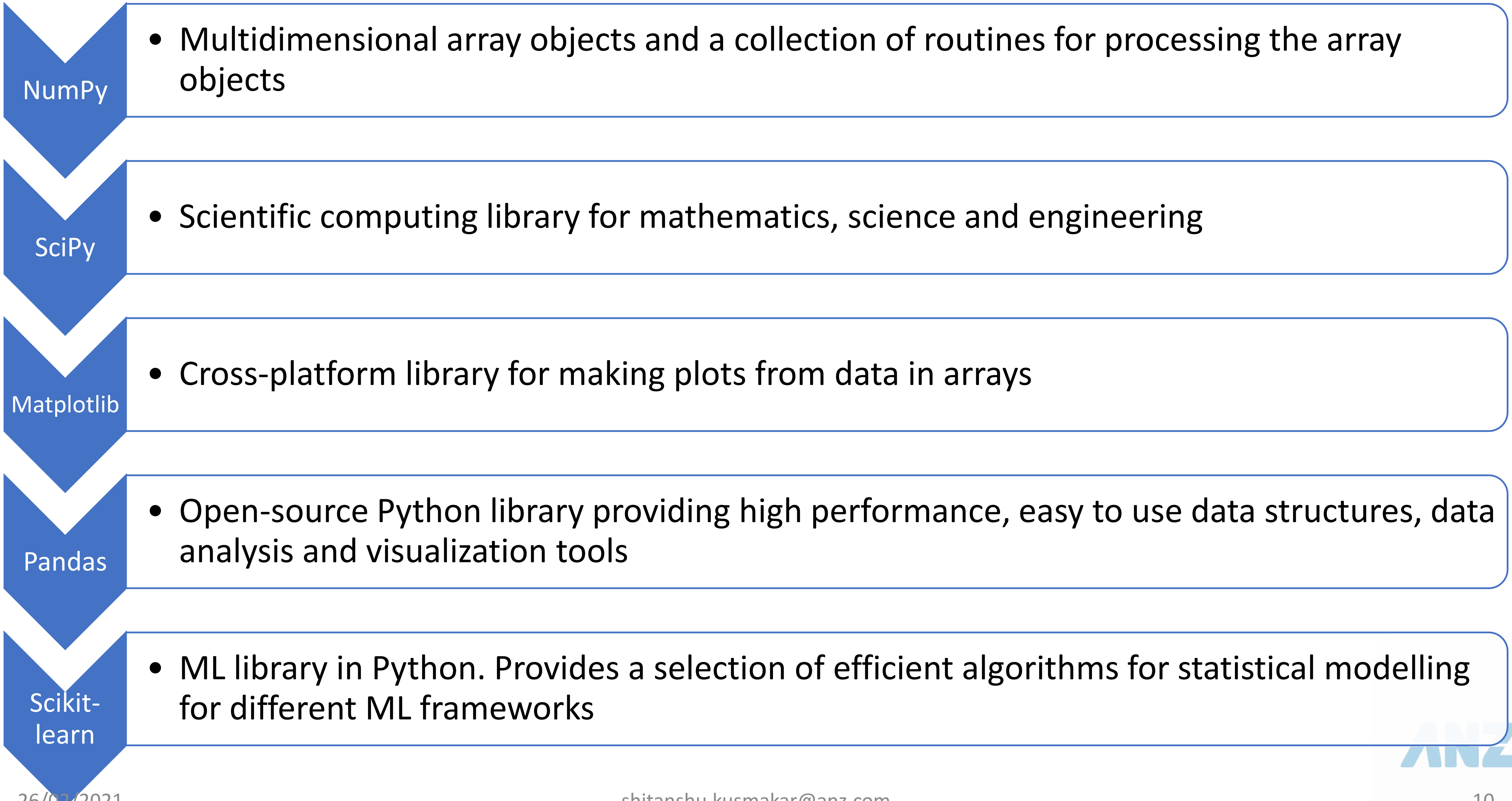
Data Structures & Collections

Control Statements

Loops

Functions

Major Libraries for Data Processing & Exploration



NumPy

- Multidimensional array objects and a collection of routines for processing the array objects

SciPy

- Scientific computing library for mathematics, science and engineering

Matplotlib

- Cross-platform library for making plots from data in arrays

Pandas

- Open-source Python library providing high performance, easy to use data structures, data analysis and visualization tools

Scikit-learn

- ML library in Python. Provides a selection of efficient algorithms for statistical modelling for different ML frameworks

References

- https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html
- https://wiki.python.org/moin/BeginnersGuide/Overview
- https://www.w3schools.com/python/python_intro.asp
- https://docs.python.org/3/tutorial/