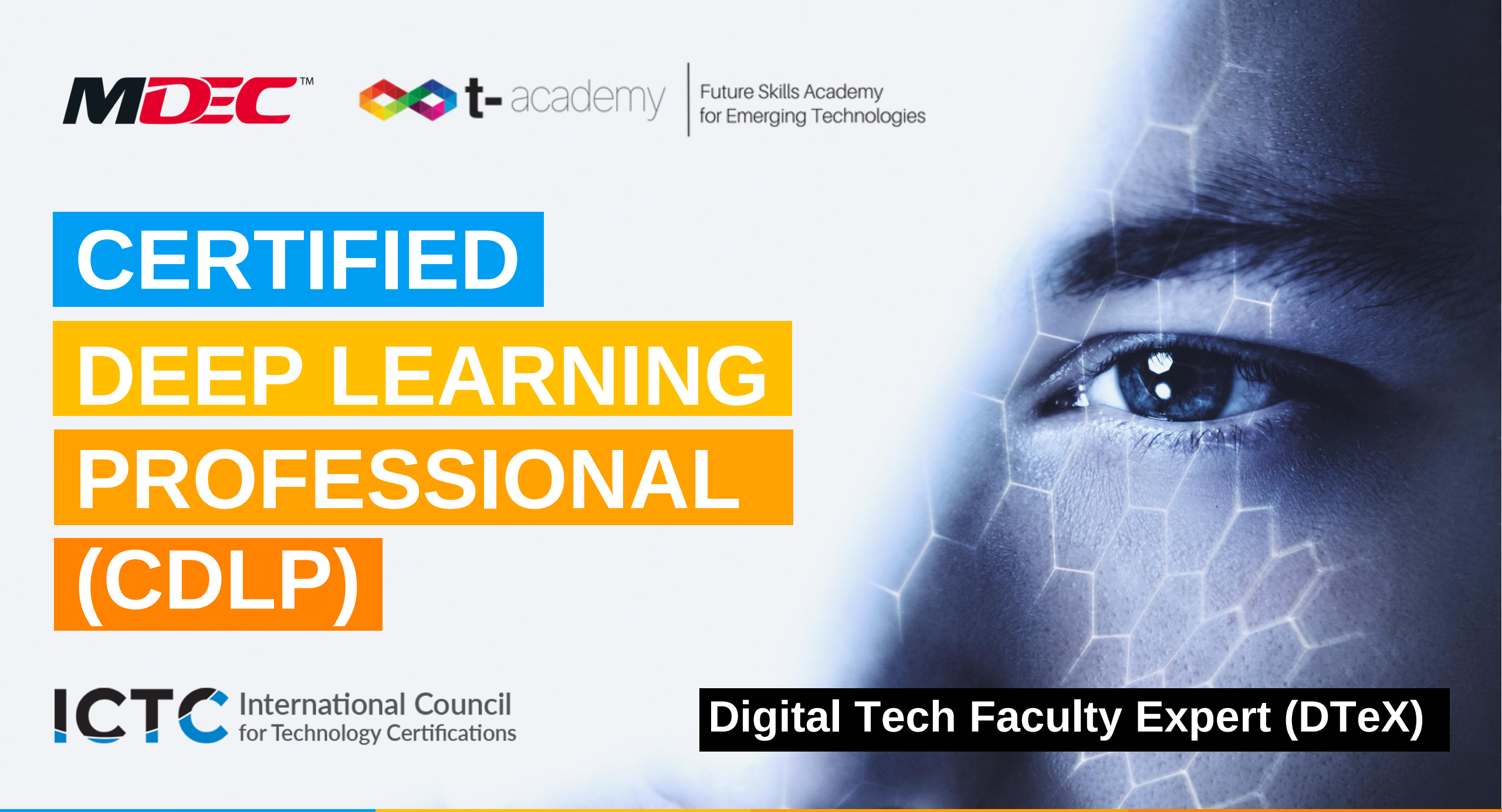
MDEC™ t- academy | Future Skills Academy for Emerging Technologies

CERTIFIED DEEP LEARNING PROFESSIONAL (CDLP)



ICTC International Council for Technology Certifications

Digital Tech Faculty Expert (DTeX)

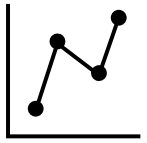
ICTC International Council for Technology Certifications

Foundations - NumPy

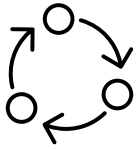
Foundations

NumPy

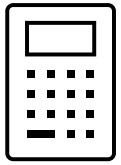
NumPy provides a solid foundation for implementing the mathematical and numerical aspects of deep learning algorithms, making it an essential tool for any deep learning practitioner or researcher working with Python-based deep learning frameworks.



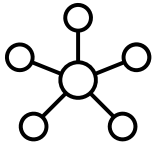
Data Handling



Efficient Operations



Matrix Operations



Compatible with DL Frameworks

NumPy

The following foundational knowledge areas are covered:

- What is Numpy?
- Why NumPy?
- NumPy Arrays and Matrices
- Indexing and Slicing
- Operations
- Exercise Questions and Solutions

NumPy

What is NumPy?

- Open source numerical library used for working with mathematical functions with multi-dimensional array and matrix data structures.

- Already installed in our environment and is very common in Data Science and Machine Learning.

NumPy

Why use NumPy?

Numpy arrays use less memory than normal python lists.

A normal python list is a group of pointers to separate Python objects (e.g. the numbers inside the list)

NumPy

Why use NumPy?

A NumPy array is designed to be an array of uniform values, without using space for type pointers.

NumPy can also read in information faster than normal python & has broadcasting operations across array dimensions

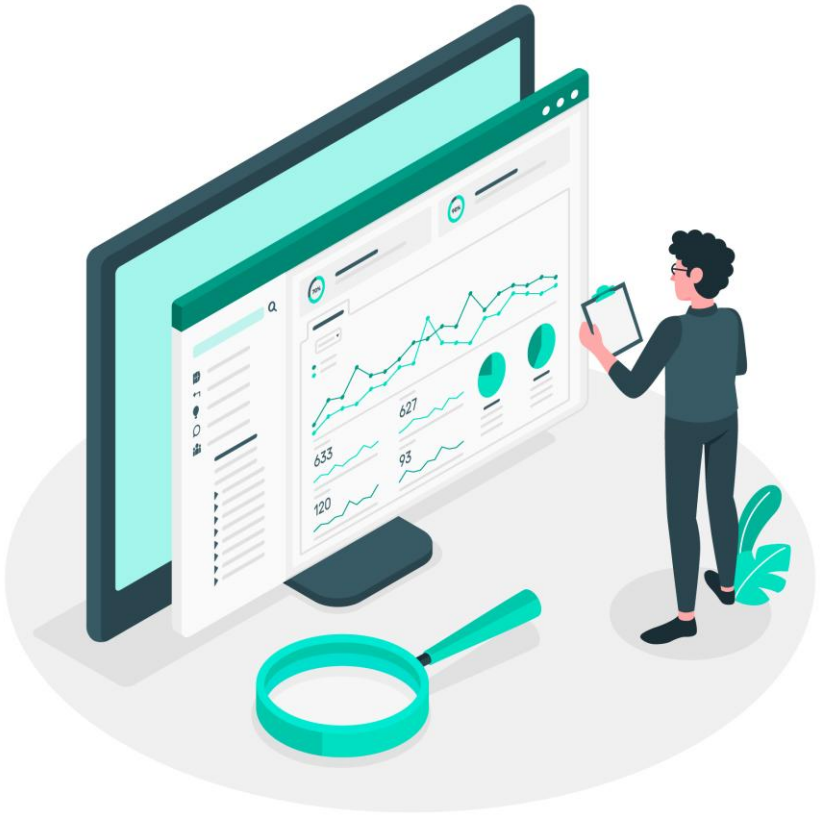
NumPy

- Later in the course we will be using numpy arrays for quick data generation, and many Tensorflow functions look extremely similar to numpy functions.

- Learning numpy sets up a great foundation for working with data later on in the course!

ICTC International Council for Technology Certifications

Numpy Arrays



Foundations

NumPy

Creating Arrays

- Converting existing objects to arrays
- Using numpy functions to generate arrays
- Creating arrays and matrices of random values
- Basic array attributes

ICTC International Council for Technology Certifications

Numpy Indexing and Selection

Foundations

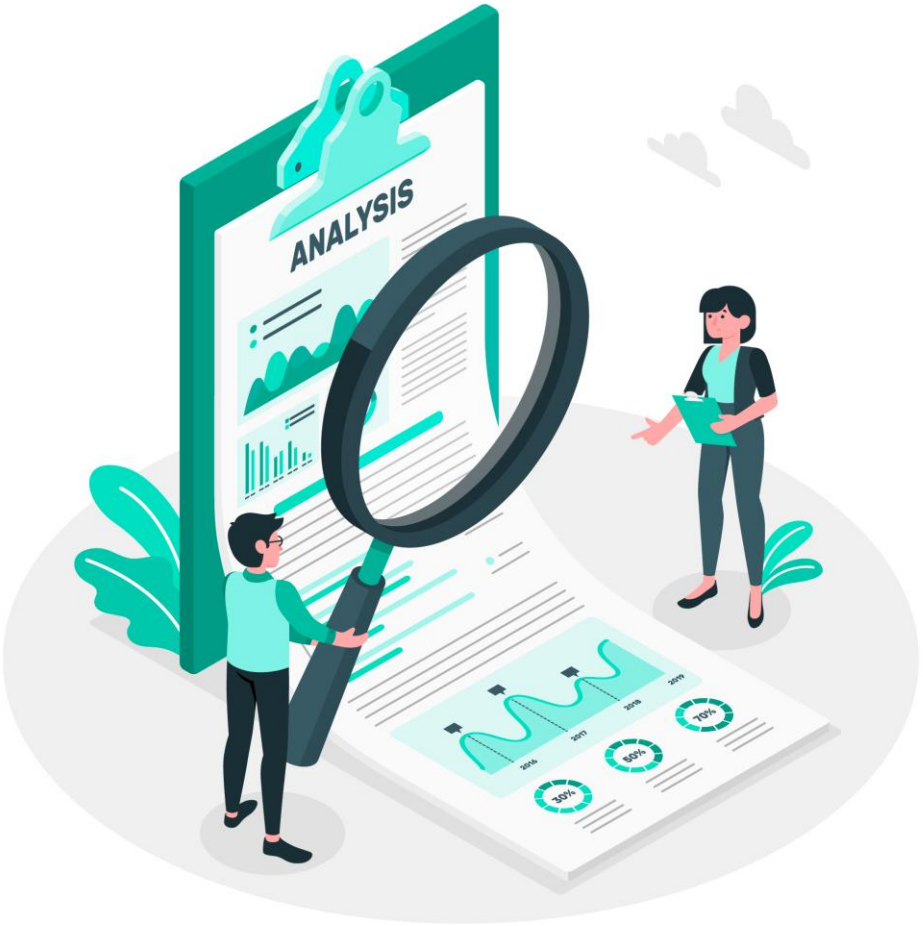
NumPy

Indexing and Selection

- Grabbing single element
- Grabbing a slice of elements
- Broadcasting selections
- Indexing and selection in 2-dimensions
- Conditional Selection

ICTC International Council for Technology Certifications

NumPy Operations



Foundations

ICTC International Council for Technology Certifications

NumPy Exercises

Foundations

ICTC International Council for Technology Certifications

Foundations – Pandas

Foundations

Pandas

The following foundational knowledge areas are covered:

- Why we use Pandas
- Pandas Series and DataFrames
- Missing Data
- Groupby Methods
- Operations
- Exercise Questions and Solutions

Pandas

What is Pandas?

Panel + **Da**ta

- Pandas stands for **Pan**el-**Da**ta.
- It is the most popular library for data handling for Python and is built directly off of NumPy.

Pandas

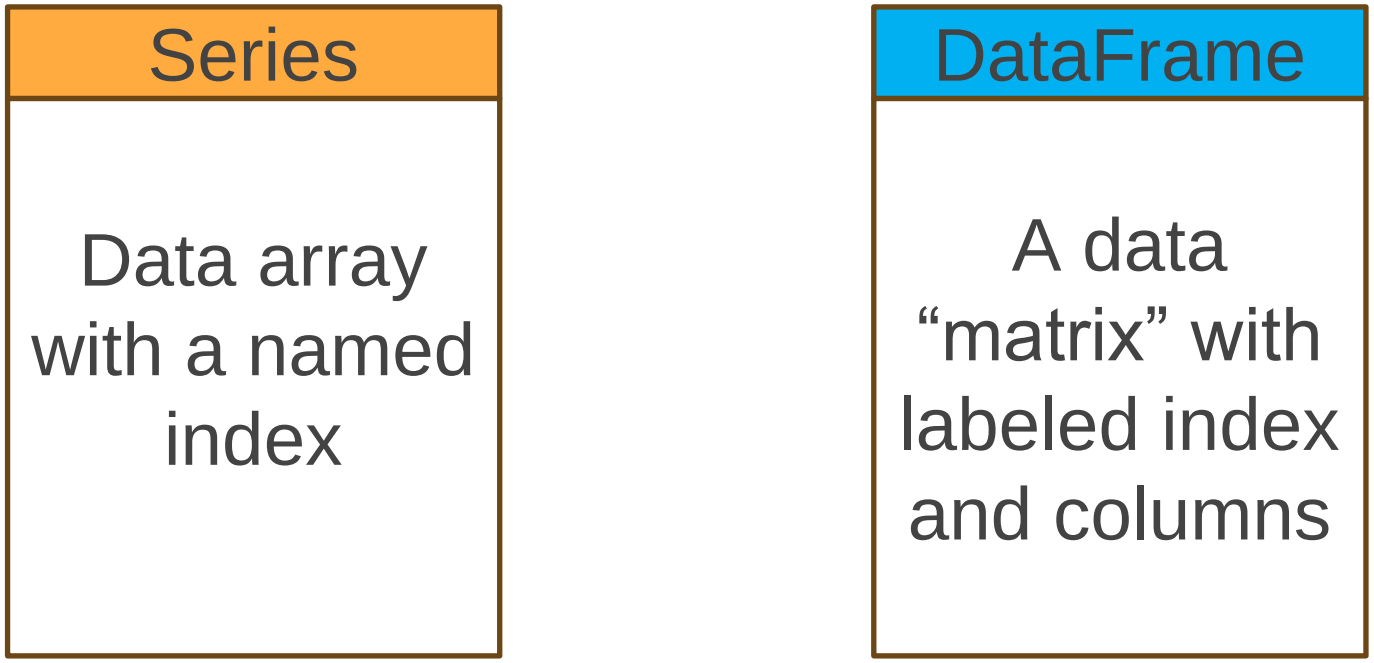
What is Pandas?

- As we will be continuously working with data in this course, we will use Pandas to read in our data, clean the data, and even perform feature engineering with Pandas.

Pandas

What is pandas?

- Pandas relies on some core data structures for its operations:



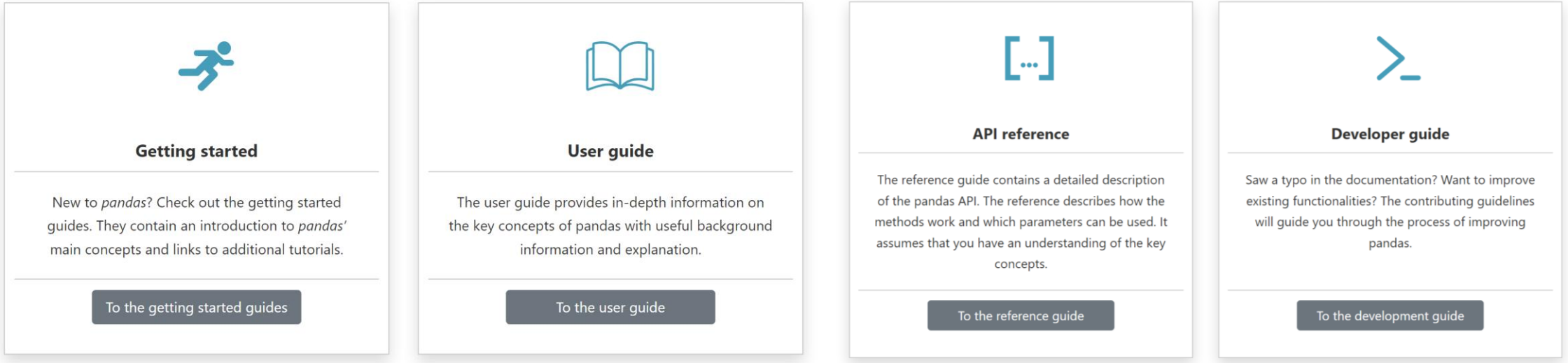
Pandas

What if you already know some pandas?

- Check out the exercise at the end of this section to test your abilities, then feel free to browse through the various pandas topics shared to review anything you may have forgotten.

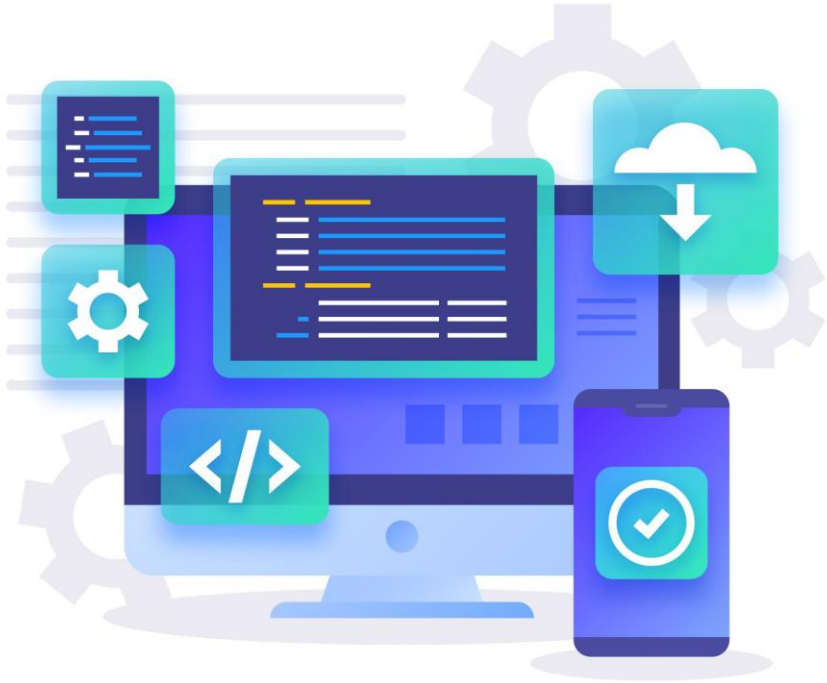
Pandas

Highly recommended you check out:
pandas.pydata.org/pandas-docs/stable/



ICTC International Council for Technology Certifications

Getting Started With Pandas



Foundations

ICTC International Council for Technology Certifications

Series

Foundations

Pandas

A pandas series is very similar to a numpy array, except for the addition of a named index.

We can use this named index to grab data from the array.

Let’s see how this works with Python.

ICTC International Council for Technology Certifications

Data Frames

Foundations

Pandas - DataFrame

- The pandas DataFrame is our main tool for working with data.
- A DataFrame is simply multiple pandas series that share the same index.
- You can think of a DataFrame as being similar to a spreadsheet, just a lot more powerful!

ICTC International Council for Technology Certifications

Missing Data

Foundations

Pandas – Missing Data

- Real world datasets often have missing data.
- Realistically there are only 3 ways to deal with missing data:

Leave it as missing

Remove the missing data

Fill in the missing data

Pandas – Missing Data

- Leave it as missing
  - Depending on the type of data, this is a valid choice.
  - For example, if dealing with categorical data, we could simply treat a NaN as another category.

Pandas – Missing Data

- Remove the missing data
  - Dependent on how much data is missing.
    - Large Percentage - too much is missing to make a reasonable guess
    - Small Percentage - only removes a few data points from our dataset

Pandas – Missing Data

- Fill in missing data
  - A non-trivial percentage is missing and the data point rows are important
  - Lot’s of strategies available:
    - Mode, Mean, Median
    - Based off another feature column, conceive of a reasonable value.

Pandas – Missing Data

- Dealing with missing data
  - “What is the correct approach?”
    - All data sets and situations are different!
    - Use your common sense and overall goals to see which strategy makes sense.

ICTC International Council for Technology Certifications

Groupby

Foundations

Pandas - Groupby

- Often we want to explore how values are distributed or aggregated across groups.
- To do this we use the groupby method, similar to a GROUP BY call in SQL.
- This process is also often referred to as Split-Apply-Combine

Pandas - Groupby

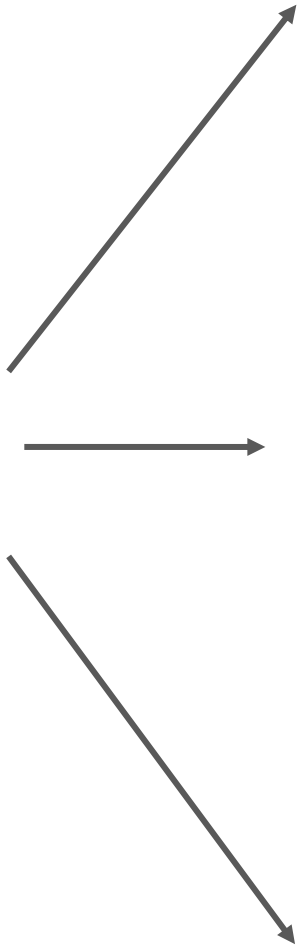
Col 1	Col 2
A	2
B	1
A	4
C	5
A	6
C	3
B	8

Pandas - Groupby

Col 1	Col 2
A	2
B	1
A	4
C	5
A	6
C	3
B	8

Pandas

Col 1	Col 2
A	2
B	1
A	4
C	5
A	6
C	3
B	8



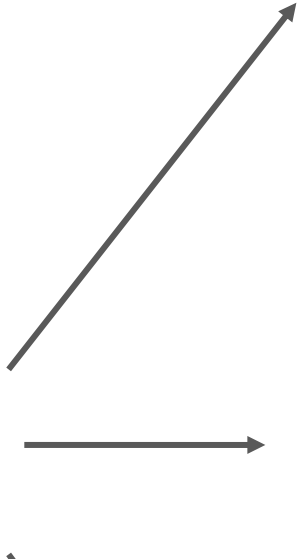
Col 1	Col 2
A	2
A	4
A	6

Col 1	Col 2
B	1
B	8

Col 1	Col 2
C	5
C	3

Pandas

Col 1	Col 2
A	2
B	1
A	4
C	5
A	6
C	3
B	8

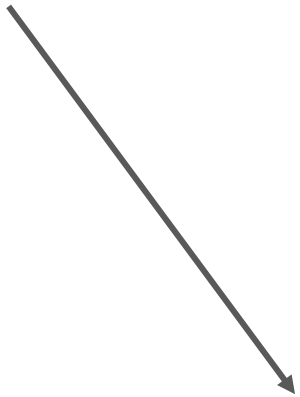


Col 1	Col 2
A	2
A	4
A	6

Col 1	Col 2
A	12

Col 1	Col 2
B	1
B	8

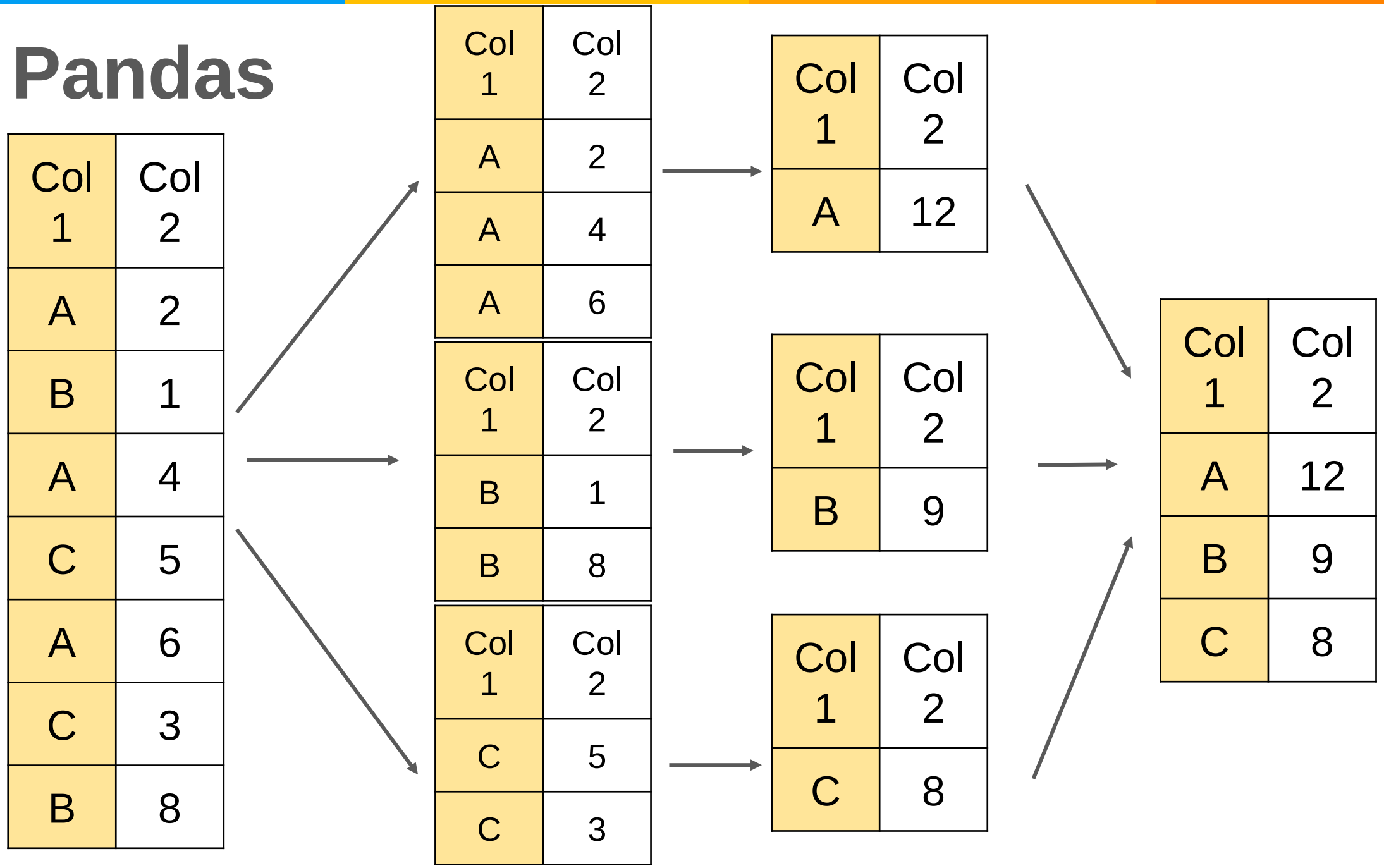
Col 1	Col 2
B	9



Col 1	Col 2
C	5
C	3

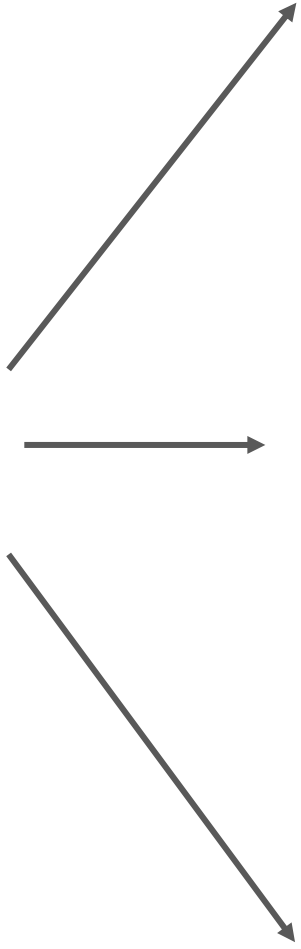
Col 1	Col 2
C	8

Pandas


Col 1
Col 2
A
2
B
1
A
4
C
5
A
6
C
3
B
8
Col 1
Col 2
A
2
A
4
A
6
Col 1
Col 2
B
1
B
8
Col 1
Col 2
C
5
C
3
Col 1
Col 2
A
12
Col 1
Col 2
B
9
Col 1
Col 2
C
8
Col 1
Col 2
A
12
B
9
C
8

Pandas

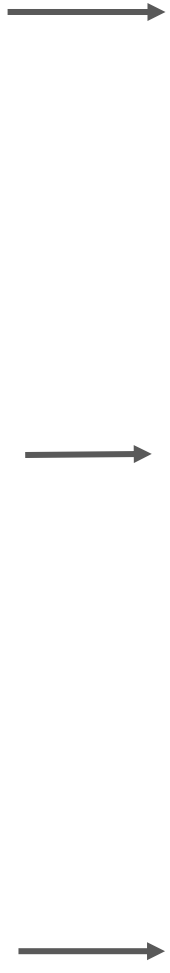
Col 1	Col 2
A	2
B	1
A	4
C	5
A	6
C	3
B	8



Col 1	Col 2
A	2
A	4
A	6

Col 1	Col 2
B	1
B	8

Col 1	Col 2
C	5
C	3



Col 1	Col 2
A	12

Col 1	Col 2
B	9

Col 1	Col 2
C	8

Col 1	Col 2
A	12
B	9
C	8

Pandas - Groupby

- The operation chosen with a groupby() call must be an **aggregation** method.
- This means it can take multiple values and combine them to return a singular value.
    - Sum,Std,Mean,Count,Max,Min, etc…
    - Let’s learn how to use groupby with Pandas!

ICTC International Council for Technology Certifications

Operations

Foundations

Pandas - Operations

- In this section, we will go over a few very useful operations that were not covered in to the previous sessions!

ICTC International Council for Technology Certifications

Data Input And Output

Foundations

Pandas – Data Input and Output

- During the course we will read our data sets from CSV files, but real world data can come from a variety of places.
- Fortunately Pandas has robust IO tools we can use.
- Let’s see a few of them in action and explore the documentation for them!

ICTC International Council for Technology Certifications

Pandas Exercises

Foundations

ICTC International Council for Technology Certifications

Pandas Exercise Solutions

Foundations

ICTC International Council for Technology Certifications

Data Visualization

Foundations

Data Visualization

- In this section we will briefly cover some data visualization techniques using Matplotlib and Seaborn.
- During the course we will be visualizing and exploring data sets, this section is meant to quickly get you up to speed on the basic syntax.

Data Visualization

- This section is **NOT** a comprehensive overview of these libraries!
- We will simply go over some common syntax calls you will see throughout the course, please refer to the official documentation of these libraries for a deeper dive into them.

Data Visualization

- Section Overview

Matplotlib

Seaborn

Visualization Exercises

ICTC International Council for Technology Certifications

Matplotlib

matplotlib

Foundations

Data Visualization - Matplotlib

- Matplotlib is the “standard” visualization library for data science and machine learning with Python.
- Both Seaborn and Pandas built-in plotting functionality are built using matplotlib.

Data Visualization - Matplotlib

- We will use matplotlib for very basic plot types, in this session, we will learn just a few of the simple plots we will use.
- For more resources, visit:
  - **www.matplotlib.org**

ICTC International Council for Technology Certifications

Seaborn

Foundations

Data Visualization - Seaborn

- Seaborn is a visualization library built using matplotlib that focuses on creating standard statistical plots with a simple one line function call.
- This comes at the cost of less customization options than pure matplotlib.

Data Visualization - Seaborn

- However, you can still effect plot attributes created in seaborn by adding in matplotlib.pyplot calls.
- For more info, visit:
    - **https://seaborn.pydata.org/**

ICTC International Council for Technology Certifications

Data Visualization Exercises

Foundations

ICTC International Council for Technology Certifications

Data Visualization Solutions

Foundations