Day 01

Data Science with Python

Pasindu Marasinghe
ppm@ucsc.cmb.ac.lk

Lecturer Panel

Pasindu Marasinghe

Sanjani Gunathilaka

Gayani Rupasinghe

Session Plan

Day 01	10th March 2024
Day 02	17th March 2024
Break	24th March 2024
Day 03	31st March 2024
Day 04	07th April 2024
Break	14th April 2024
Day 05	21st April 2024
Day 06	28th April 2024
Day 07	05th May 2024
Day 08	12th May 2024

Completion Requirements

- 75% Attendance (6 Days out of 8)
- Data Manipulation, Cleaning & Visualization Project (Individual)

  Release Date: Day 02

  Submission Date: Day 04
- Machine Learning Project (Group)

  Students per Group: 05

  Release Date: Day 04

  Submission Date: Day 08

Day Plan

Start	**08.30 AM**
Tea Break	**10.30 AM – 11.00 AM**
Lunch Break	**12.30 PM – 01.30 PM**
Tea Break	**3.00 PM – 3.30 PM**
End	**05.00 PM**

End of House Keeping.
Let’s Get to Know Why You are Here.

Data Science with Python

What is Data Science?

Why Data Science?

What is Data Science?

Data science is an interdisciplinary field that combines statistical and computational methods to extract insights and knowledge from data. It involves the use of various tools and techniques to collect, process, analyze, and interpret large and complex data sets.

Why Data Science?

1. **Better decision-making:** Data science enables organizations to make data-driven decisions, which are more accurate and objective than those based on intuition or guesswork.
2. **Competitive advantage:** Organizations that can effectively collect, analyze, and utilize data have a significant competitive advantage over those that cannot.

Why Data Science?

3. **Improved products and services:** By understanding customer preferences and behavior patterns, organizations can develop and improve their products and services, leading to better customer satisfaction and loyalty.
4. **Cost savings:** Data science can help organizations identify inefficiencies and areas for optimization, leading to cost savings.

Why Data Science?

5. **Predictive modeling:** Data science allows organizations to predict future trends and outcomes, enabling them to proactively respond to potential issues or opportunities.
6. **Personalization:** With the help of data science, organizations can personalize their products and services, tailoring them to the specific needs and preferences of individual customers.

Learning Data Science

=

Learning Python?


MATLAB
JavaScript
R
sas
SQL
Scala
Kotlin
julia
Java

Why Python?

Python is an excellent choice for beginners in data science for several reasons:

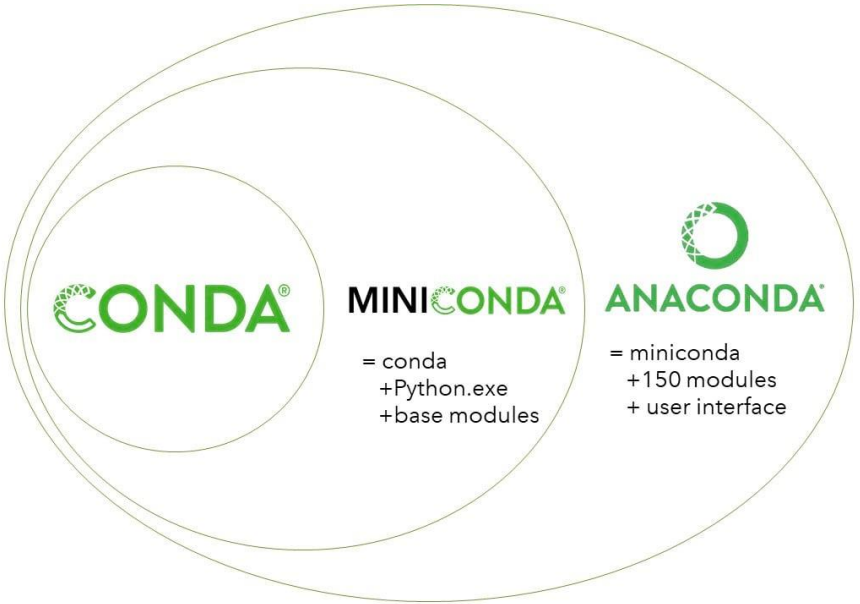
- **Easy to learn:** Python has a relatively simple and easy-to-learn syntax, making it accessible for beginners without a programming background.
- **Rich libraries and frameworks:** Python has a vast collection of libraries and frameworks designed specifically for data science, including NumPy, Pandas, Scikit-learn, Matplotlib, and many others.
- **Large community:** Python has a large and active community of developers and users, providing support, documentation, and resources for beginners in data science.

Okay !
I Visited Python Website and Installed Python. Now What?

Environment Manager

An environment manager like Conda is essential for Python development for several reasons:

- **Managing dependencies**
- **Reproducibility**
- **Flexibility**
- **Collaboration**
- **Experimentation**



Let’s Explore Anaconda

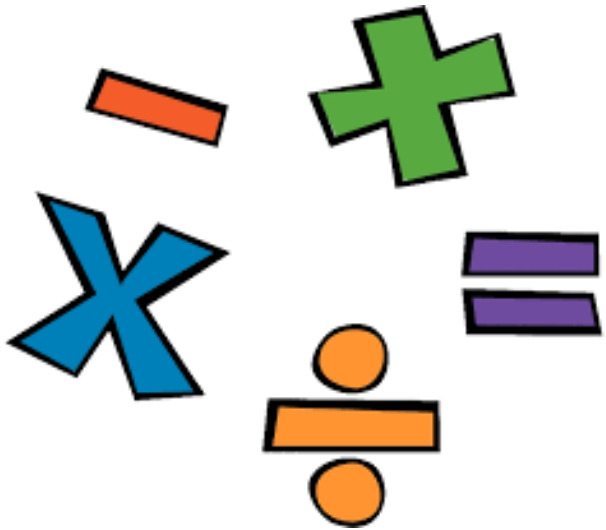
Let’s Get Back to Basic Python

Built-in Data Types

1. **Text Type:** str
2. **Numeric Types:** int, float, complex
3. **Sequence Types:** list, tuple, range
4. **Mapping Type:** dict
5. **Set Types:** set, frozenset
6. **Boolean Type:** bool
7. **Binary Types:** bytes, bytearray, memoryview
8. **None Type:** NoneType

Python Operators

1. Arithmetic operators
2. Assignment operators
3. Comparison operators
4. Logical operators
5. Identity operators
6. Membership operators
7. Bitwise operators



Python File Handling

open(filename, arguments)

- "r" - Read - Default value. Opens a file for reading, error if the file does not exist
- "a" - Append - Opens a file for appending, creates the file if it does not exist
- "w" - Write - Opens a file for writing, creates the file if it does not exist
- "x" - Create - Creates the specified file, returns an error if the file exists

- "t" - Text - Default value. Text mode
- "b" - Binary - Binary mode (e.g. images)

Python Lambda

lambda arguments : expression

- x = lambda a : a + 10

  print(x(5))

- x = lambda a, b : a * b

  print(x(5, 6))

Activity 01

NumPy

NumPy - Creating Arrays

- numpy.array()
- numpy.zeros()
- numpy.ones()
- numpy.arange()
- numpy.linspace()

NumPy – Array Attributes

- shape
- dtype
- size
- ndim
- itemsize

NumPy – Array Functions

- sum()
- mean()
- min()
- max()
- std()

NumPy – Read CSV Files

numpy.loadtxt('data.csv')

- **fname:** The file name to load data from.
- **delimiter (optional):** Delimiter to consider while creating array of values from text, default is whitespace.
- **encoding (optional):** Encoding used to decode the inputfile.
- **dtype (optional):** Data type of the resulting arraystd()

NumPy – Read CSV Files

numpy.genfromtxt('data.csv')

- **fname:** The file to read from
- **delimiter (optional):** Delimiter to consider while creating array of values from text, default is any consecutive white spaces act as a delimiter.
- **missing_values (optional):** The set of strings to use incase of a missing value.
- **dtype (optional):** Data type of the resulting array

Pandas

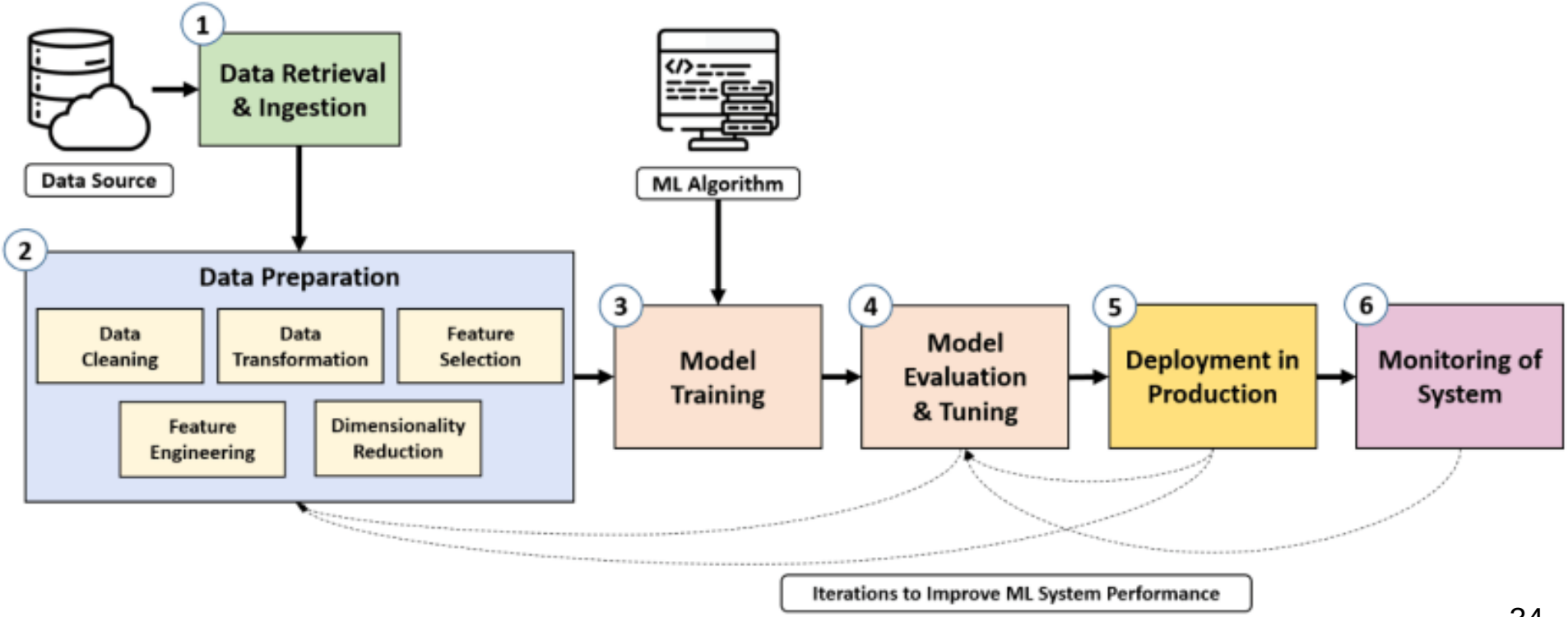
Pandas – Read CSV Files

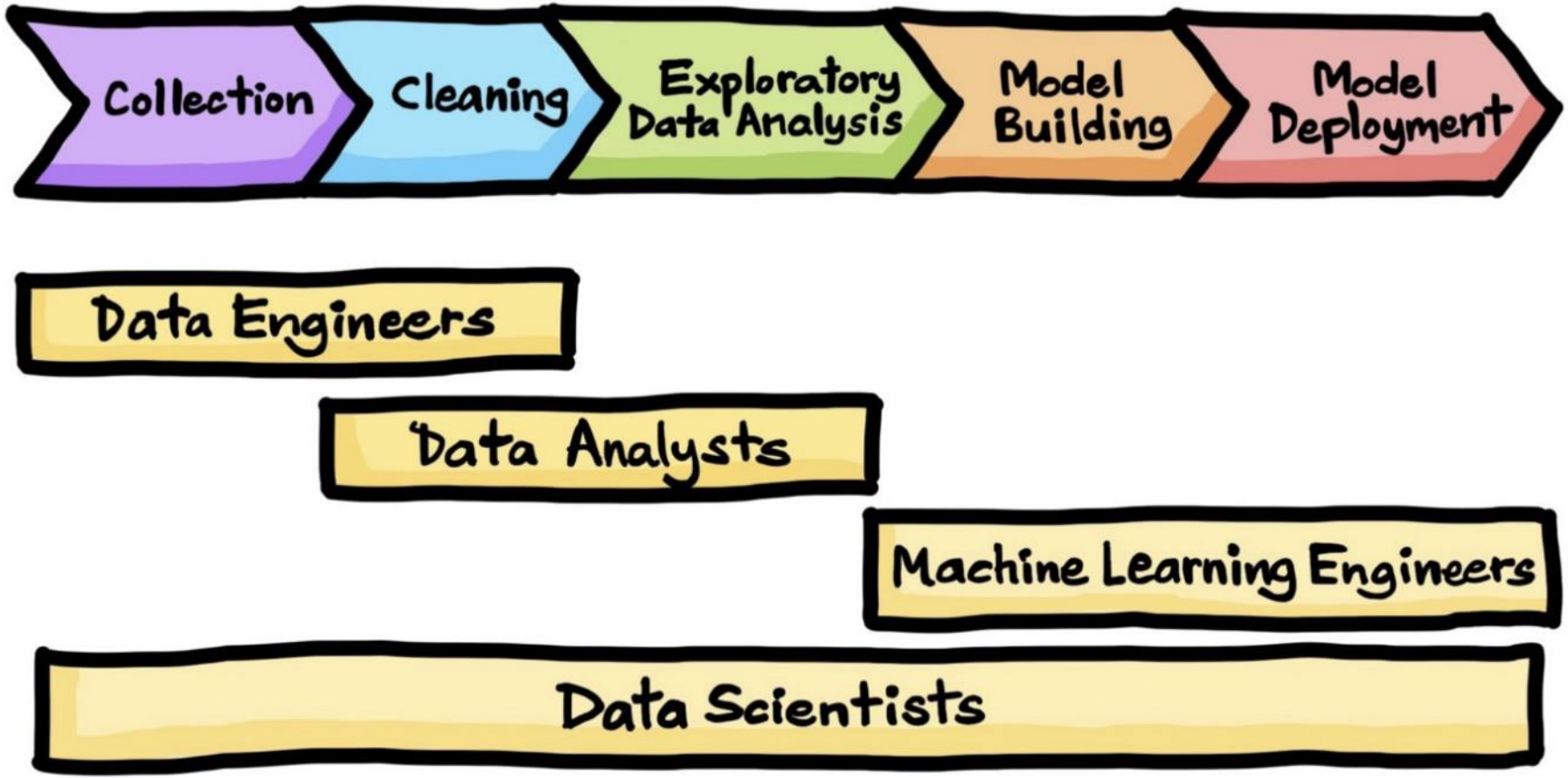
pd.read_csv('data.csv')

- **fname:** The file to read from
- **delimiter (optional):** Delimiter to consider while creating dataframe of values from text, default is any consecutive white spaces act as a delimiter.
- **dtype (optional):** Data type of the resulting dataframe

Back to Our Main Topic: Data Science

Data Science Basic Pipeline

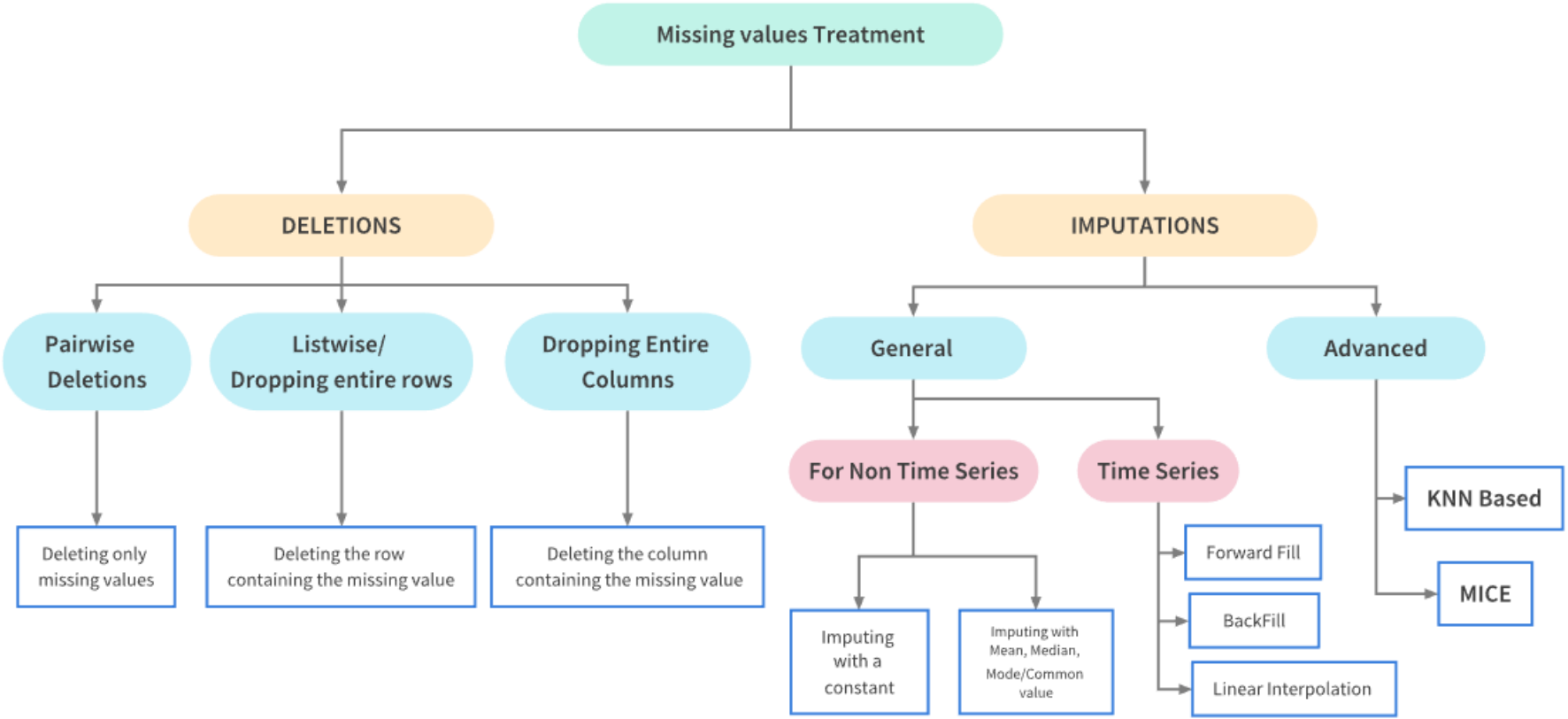



Collection
Cleaning
Exploratory Data Analysis
Model Building
Model Deployment
Data Engineers
Data Analysts
Machine Learning Engineers
Data Scientists

Data Cleaning

Data Cleaning

1. Removing Duplicates
2. Remove Irrelevant Data
3. Standardize Capitalization
4. Convert Data Type
5. Handling Outliers
6. Fix Errors
7. Language Translation
8. Handle Missing Values


Missing values Treatment
DELETIONS
IMPUTATIONS
Pairwise Deletions
Listwise/ Dropping entire rows
Dropping Entire Columns
General
Advanced
Deleting only missing values
Deleting the row containing the missing value
Deleting the column containing the missing value
For Non Time Series
Time Series
KNN Based
MICE
Forward Fill
BackFill
Linear Interpolation
Imputing with a constant
Imputing with Mean, Median, Mode/Common value

Activity 02