INTRODUCTION TO PANDAS

Lekshmi S

MoES Research Fellow

Climate Applications and User Interface Group

India Meteorological Department, Pune

PANDAS

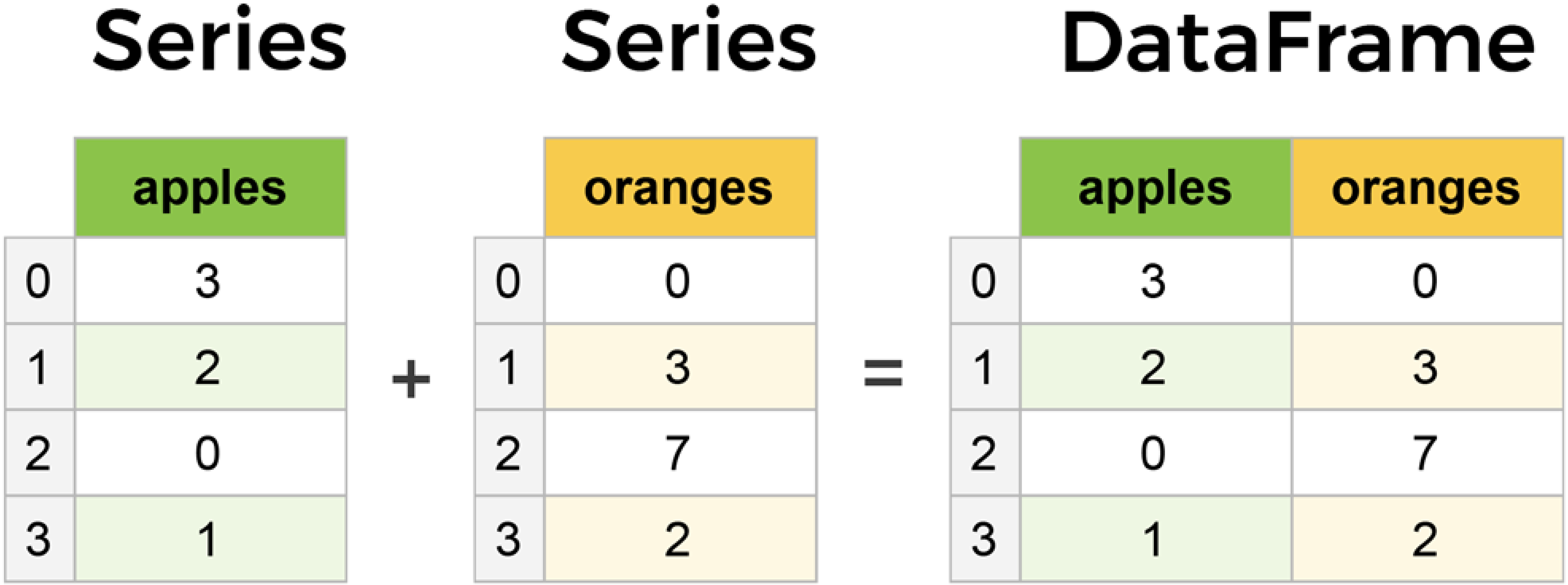
- Pandas is a Python library used for working with data sets.
- It has functions for analyzing, cleaning, exploring, and manipulating data.
- The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.
- The source code for Pandas is located at this github repository https://github.com/pandas-dev/pandas

Why Use Pandas?

- Pandas allows us to analyze big data and make conclusions based on statistical theories.
- Pandas can clean messy data sets, and make them readable and relevant.
- Relevant data is very important in data science.

Data Types in Pandas

- The primary two components of pandas are the **Series** and **DataFrame**.
- A **Series** is essentially a column, and a **DataFrame** is a multi-dimensional table made up of a collection of Series.



Series

	apples
0	3
1	2
2	0
3	1

+

Series

	oranges
0	0
1	3
2	7
3	2

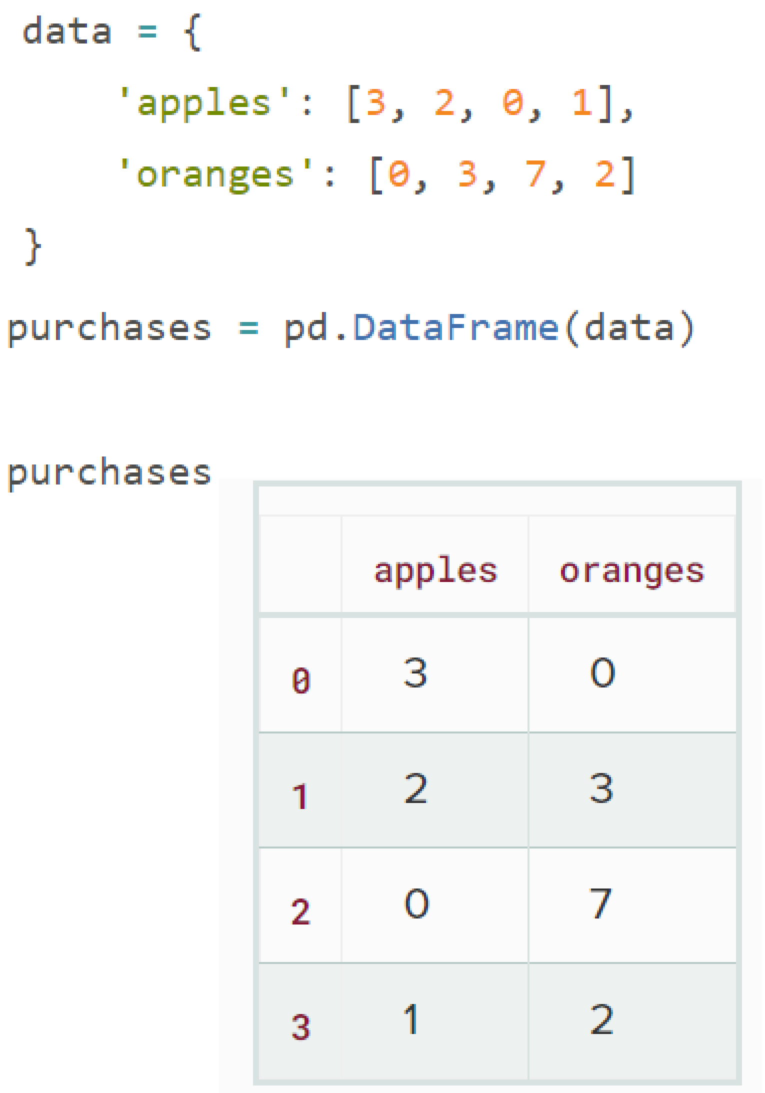
=

DataFrame

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

Creating a DataFrame

Method 1

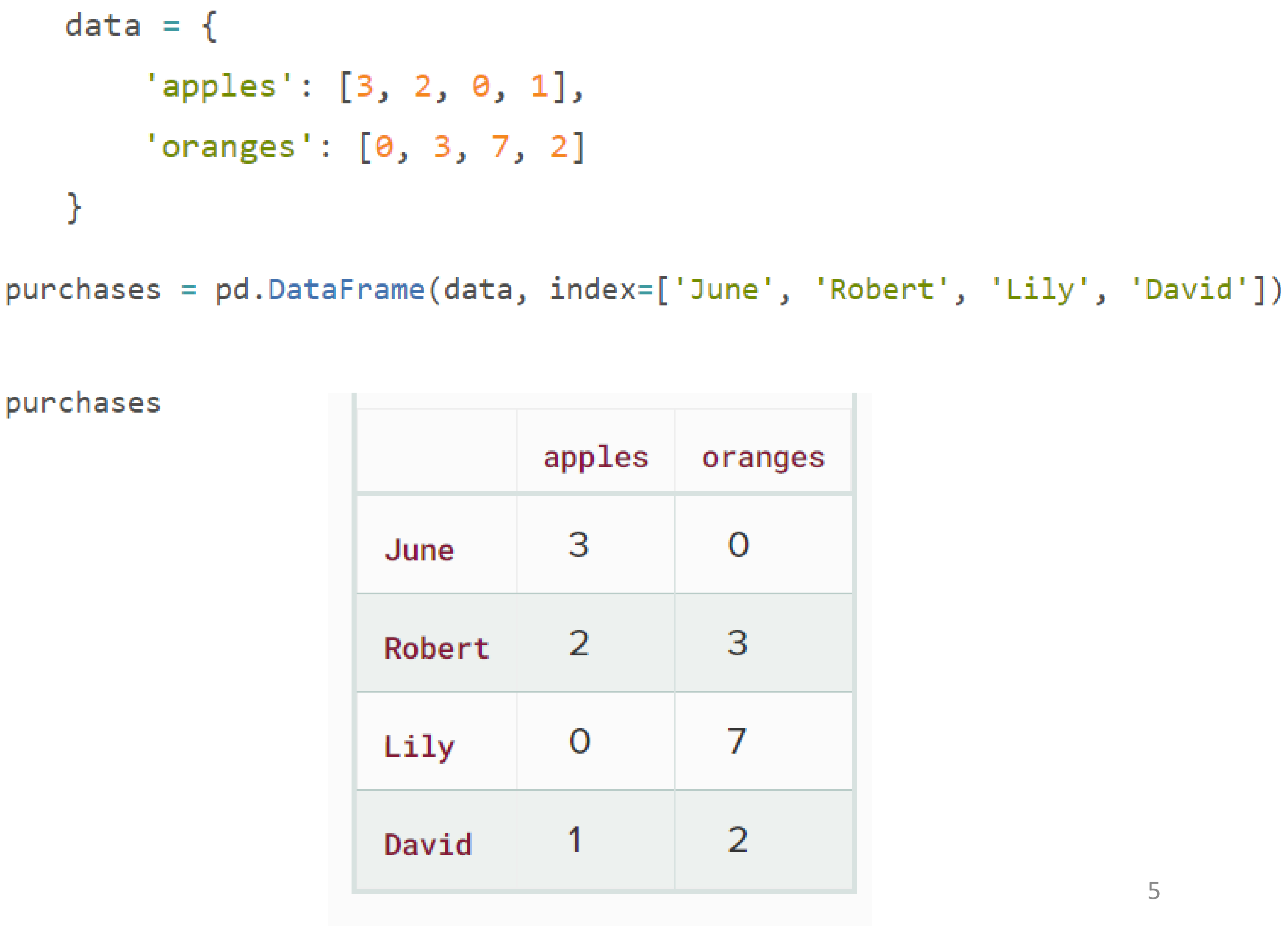


```
data = {
    'apples': [3, 2, 0, 1],
    'oranges': [0, 3, 7, 2]
}
purchases = pd.DataFrame(data)

purchases
```

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

Method 2



```
data = {
    'apples': [3, 2, 0, 1],
    'oranges': [0, 3, 7, 2]
}
purchases = pd.DataFrame(data, index=['June', 'Robert', 'Lily', 'David'])

purchases
```

	apples	oranges
June	3	0
Robert	2	3
Lily	0	7
David	1	2