 # MODULE 1 : Data Exploration

DATA-MANIPULATION-1

Wrangling Data to extract meaningful insights

Club Informatique & Télécom
Data Cell

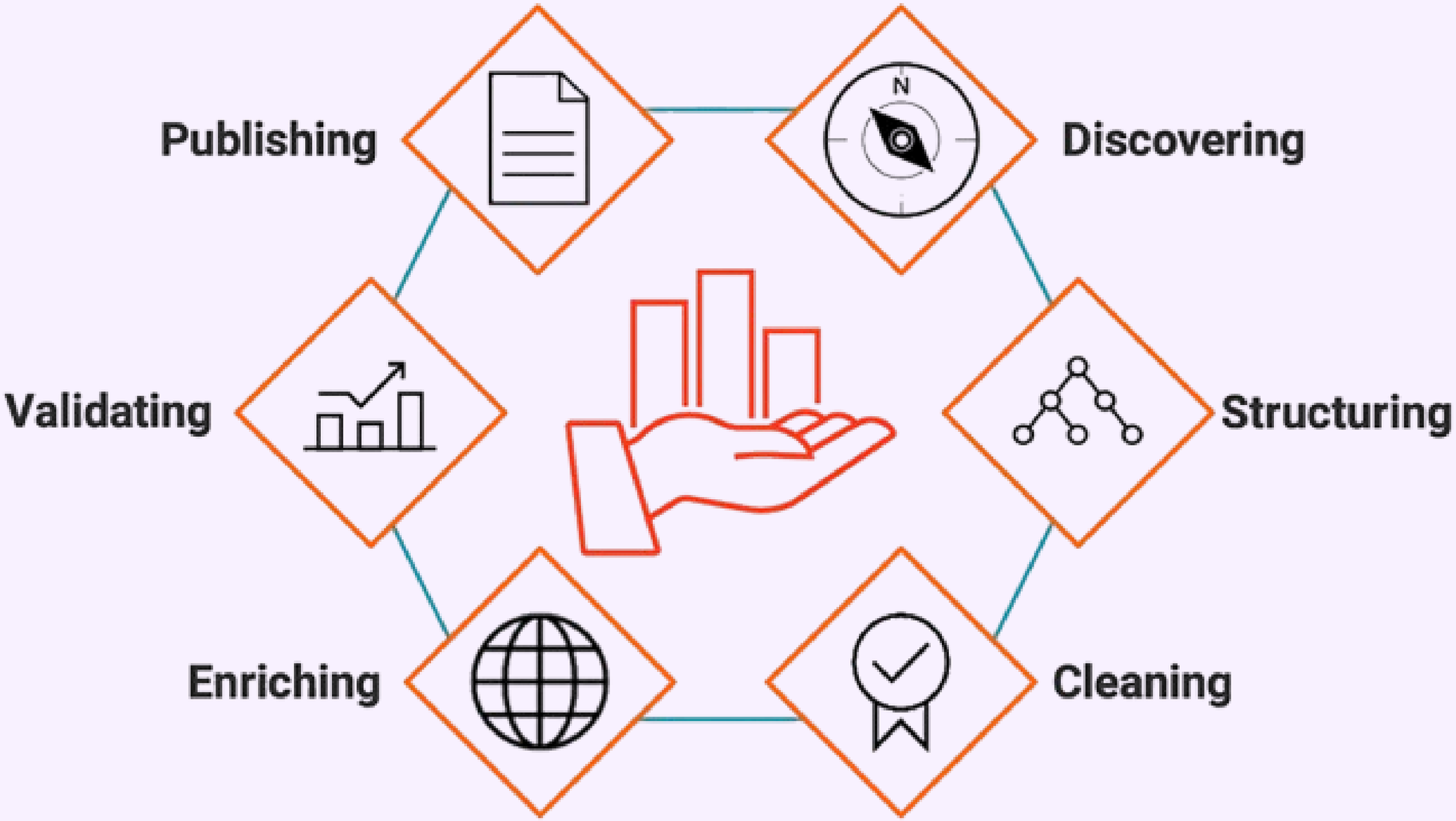
Today's Menu

1. What is Data Wrangling ?
2. pandas Library: Data Structures
3. Viewing / Inspecting Data
4. Selecting Data
5. Cleaning Data
6. Lab: Olympic Data

Data Wrangling

Data Wrangling is the process of transforming and structuring data from one raw form into a desired format with the intent of improving data quality and making it more consumable and useful for analytics or machine learning. It's also sometimes called data munging.

~ alteryx.com


Publishing
Discovering
Validating
Structuring
Enriching
Cleaning

Pandas Data Structures

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

Viewing / Inspecting data

df.**head**(N)	Displays first N rows. By default N=5
df.**tail**(N)	Displays last N rows. By default N=5
df.**shape**	Number of rows and columns
df.**info**()	Index, DataType, and Memory information
df.**describe**()	Summary statistics for numerical columns
df['col'].**value_counts**()	View a summary of the unique values in a Series

Selecting data

df['col']	Single column by name
df[['col1', 'col2']]	Multiple columns by name.
df.**iloc**[row_idx, col_idx]	Select rows and columns by index
df.**loc**[row_idx,col_label]	Select rows by index and columns by label

Cleaning data

pd.isnull(df)	Check for null values, returns a Boolean array
df.**dropna**(axis=0)	• By default (axis=0), drops all rows that contain NULL values • axis=1, drops all columns that contain NULL values
df.**fillna**(value)	Replace all NULL values with the specified value
df['col'].**replace**([1,2], ['one','two'])	In column 'col', replace all 1 with 'one' and 2 with 'two'
df.**rename**(columns={'old': 'new})	Rename column 'old' to 'new'

Ready, Set, Code!