Basic Data Science

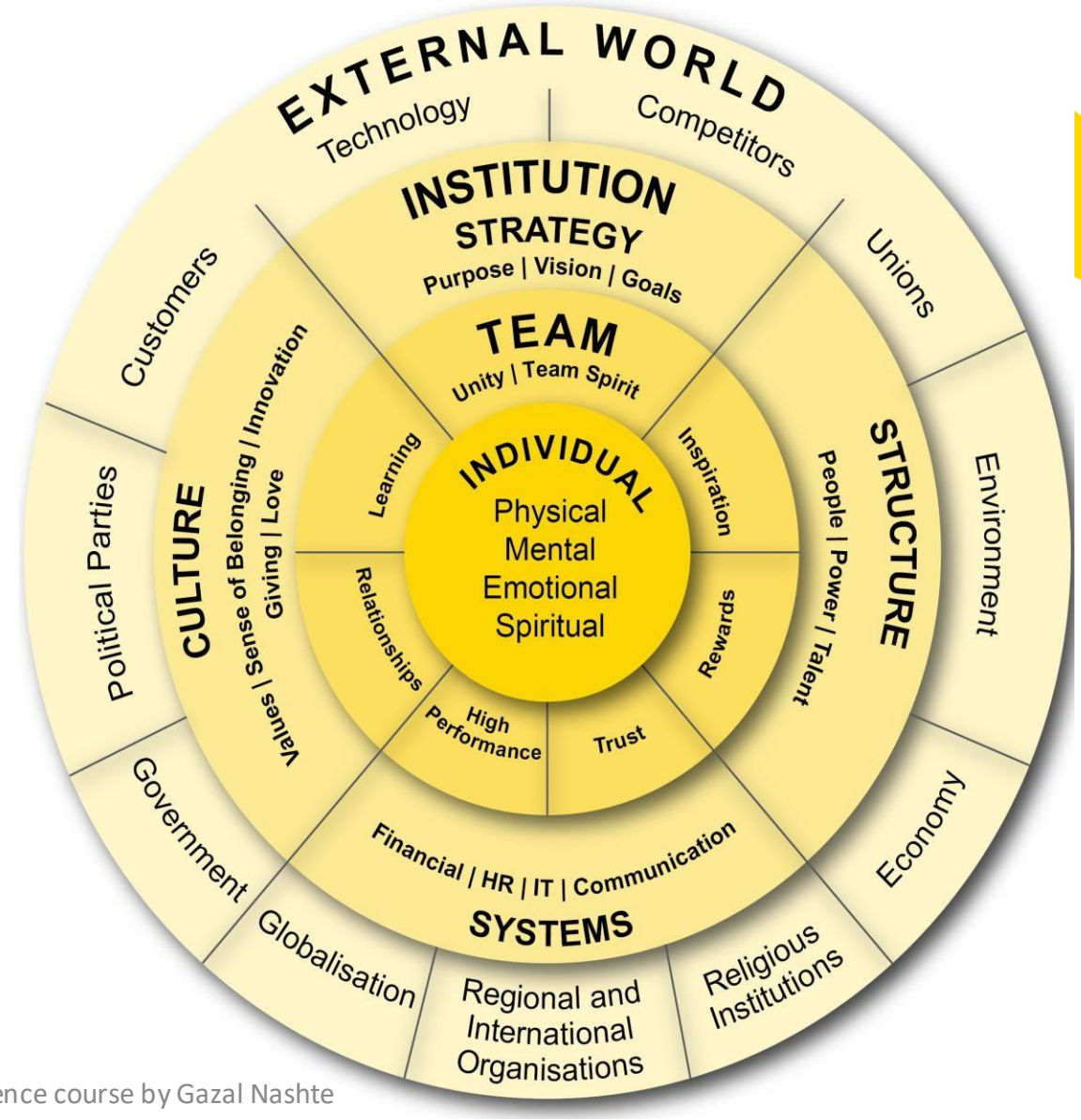
(Basic Data Science Program)

Gazal Nashte

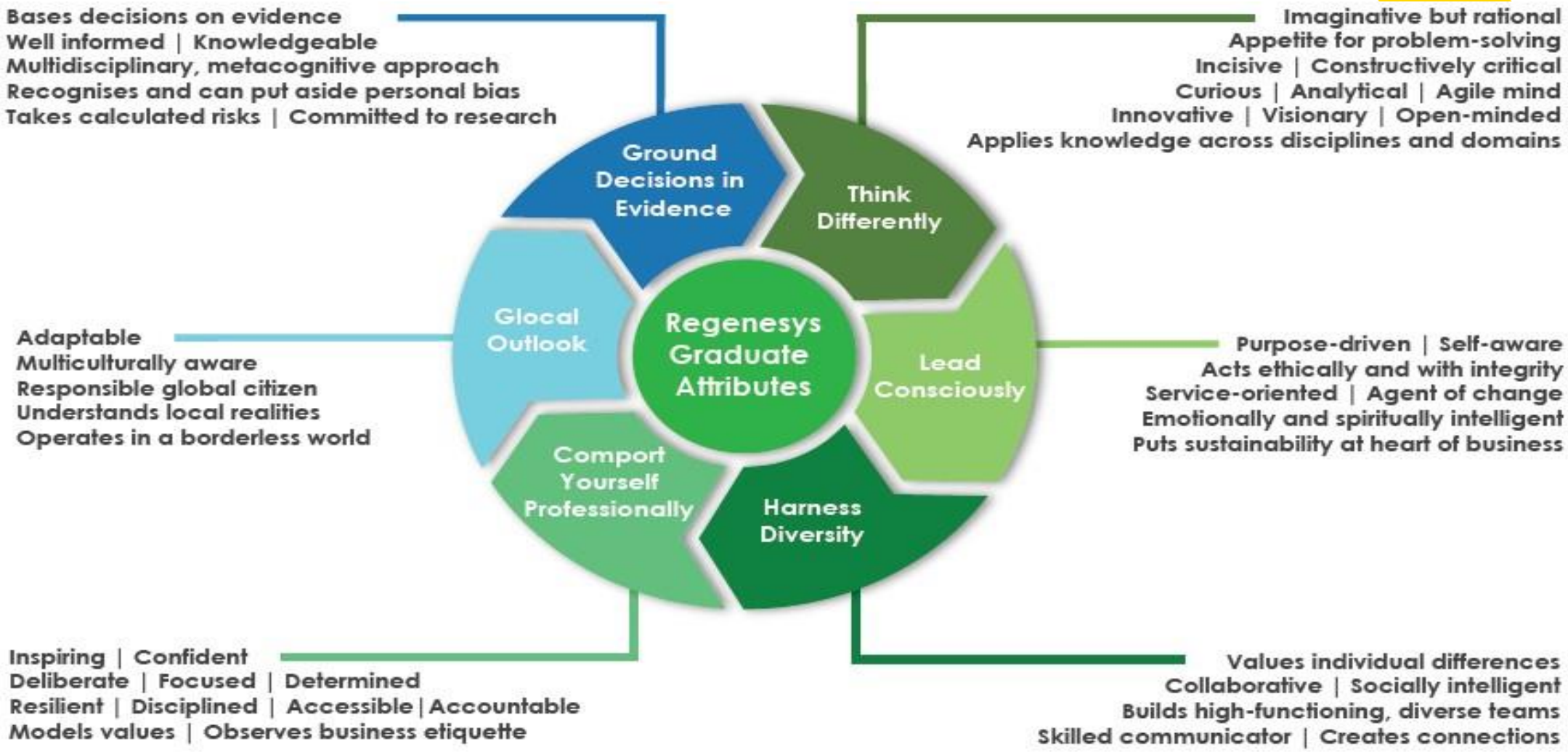
REGENESYS BUSINESS SCHOOL

REGENESYS' Integrated Leadership and Management model

- **Holistic** focus on the individual (SQ, EQ, IQ, and PQ)
- **Interrelationships** are dynamic between individual, team, institution and the external environment (systemic)
- **Strategy** affects individual, team, organisational, and environmental performance
- **Delivery** requires alignment of strategy, structure, systems and culture



Regenesys Graduate attributes



Contents

Pandas

Pandas

- Open-source python library used for data manipulation and analysis.
- It is built on top of numpy.
- It offers data structures and operations for manipulating numerical tables and time series.
- Wes McKinney started building it from 2007 to 2010
- It involves Series for 1D arrays and DataFrame object for 2D data manipulation with integrated indexing.
- Tools for reading and writing data between in-memory data structures and different file formats.
- Reshaping and pivoting of data sets.

Pandas Objects

Series

- It is a one-dimensional ndarray with index labels of any type. It is essentially a column

Dataframe

- It is a 2-dimensional labeled data structure with columns of potentially different types.
- It can be like a spreadsheet or SQL table, or a dict of Series objects.
- It is generally the most commonly used pandas object.

Series and Series functions

- From a list
- From a Numpy Array
- From a Dictionary

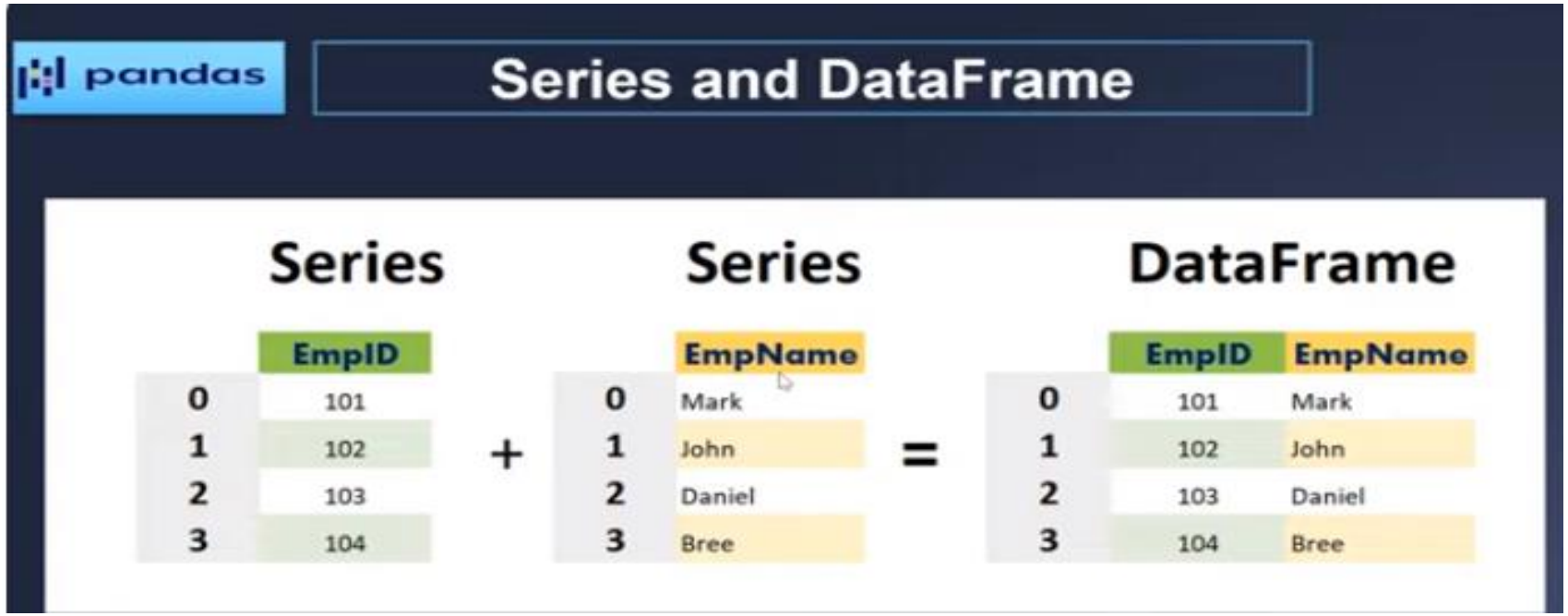
Operations on Pandas Series

Operations on Pandas Series

1. Indexing and Slicing
2. Statistical operations
3. Sorting
4. Handling missing values,
   unique values

Dataframe

A Pandas DataFrame is a two-dimensional, size-mutable, and heterogeneous tabular data structure with labeled axes (rows and columns). It is similar to a spreadsheet or a SQL table, where data is organized into rows and columns.



Different ways of creating a dataframe

- From a list
- From a Dictionary
- From a CSV file

Data Frame Operations

1. Head()/tail()/describe()
2. Shape
3. Columns
4. Values
5. Dropping specific columns
6. Dropping duplicates
7. Checking for null values
8. Summary statistics of a specific column

DataFrames indexing and selection operations

1. Selecting columns
2. Selecting rows
3. Selecting columns and rows
4. Conditional Selection
5. Setting Index
6. Sorting on single column

For project

1. Create a dataframe and import the csv file provided
2. Find the number of rows and columns in the dataset
3. Display the columns of the dataset
4. Display the column gender
5. Sort the dataset as per tenure