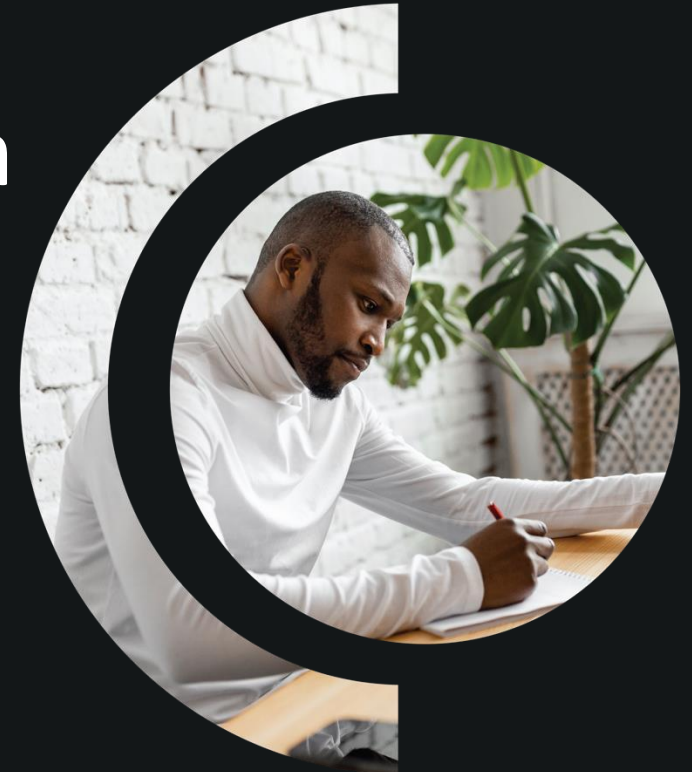


Basic Data Science / Introduction to Data Analysis and Data Visualisation

Dr. Richa Dhanuka

Senior Lecturer

Email: richad@regenesys.net



Know Your Facilitator



- B. Tech (IT)
- M. Tech (IT)
- Ph.D. in Computer Science

Education



- Passionate Researcher
- Educator
- Software Developer

Highlights

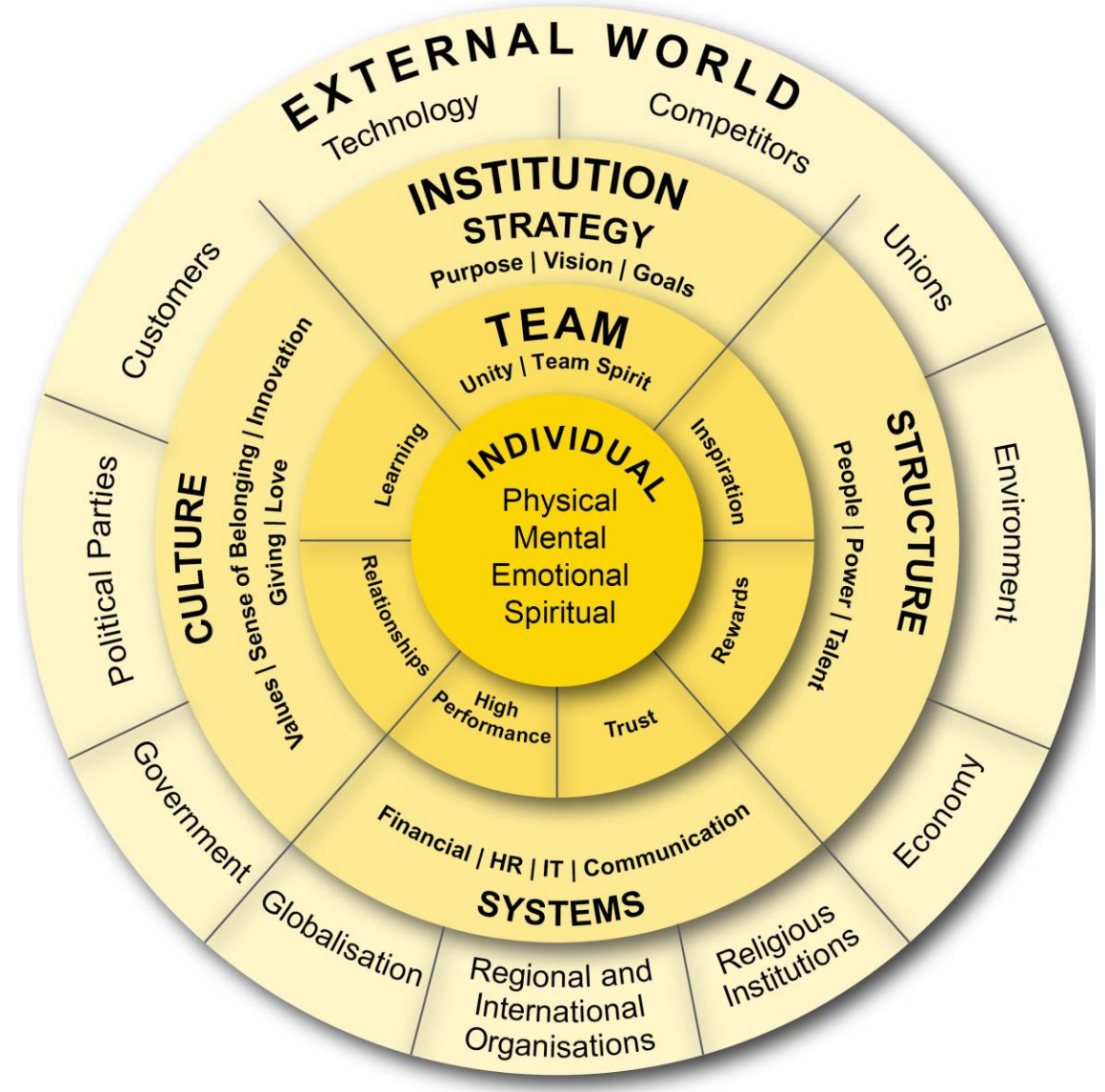


"Dr. Richa Dhanuka is a Senior Lecturer at Regenesys School of Technology. She brings 13 years of experience in academia and industry. She holds a Ph.D. from a prestigious institute in India and has a background in software development and deep learning. Dr. Richa's research contributions span various SCI Journals, emphasizing her expertise in deep learning and technology. She employs a hands-on teaching approach, merging theory with practical applications for a comprehensive learning experience. "

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

Bases decisions on evidence
Well informed | Knowledgeable
Multidisciplinary, metacognitive approach
Recognises and can put aside personal bias
Takes calculated risks | Committed to research

Ground
Decisions in
Evidence

Imaginative but rational
Appetite for problem-solving
Incisive | Constructively critical
Curious | Analytical | Agile mind
Innovative | Visionary | Open-minded
Applies knowledge across disciplines and domains

Think
Differently

Adaptable
Multiculturally aware
Responsible global citizen
Understands local realities
Operates in a borderless world

Glocal
Outlook

Purpose-driven | Self-aware
Acts ethically and with integrity
Service-oriented | Agent of change
Emotionally and spiritually intelligent
Puts sustainability at heart of business

Lead
Consciously

Comport
Yourself
Professionally

Harness
Diversity

Inspiring | Confident
Deliberate | Focused | Determined
Resilient | Disciplined | Accessible | Accountable
Models values | Observes business etiquette

Values individual differences
Collaborative | Socially intelligent
Builds high-functioning, diverse teams
Skilled communicator | Creates connections

Establishing Ground Rules

Active Listening

Listen attentively, understand, and respond effectively



Respectful Communication

Communicate openly and respectfully with your peers and facilitators

Come Prepared

Arrive with the necessary information and materials for productive discussions

Be on Time

Respect others' time and be punctual for all engagements



Embrace Challenges

View challenges as opportunities for growth and learning

Follow Online Etiquettes

Adhere to respectful and professional behavior in virtual learning

Have fun and enjoy the learning!

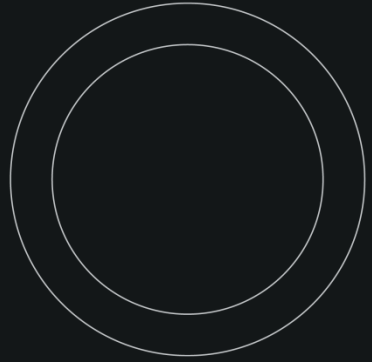


Agenda

- Session Outline
- What is Data Science
- Numpy

Session Outline

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Feature Engineering
- Feature Selection
- Extrapolatory Data Analysis
- Web Scraping



What is data science?



What is Data Science?

- Data science is the study of data to extract meaningful insights from data.
- It is a multidisciplinary approach that combines principles and practices from the fields of mathematics, statistics, artificial intelligence, and computer engineering to analyze large amounts of data.

What is AI and Data Science?

- Data science combines statistical tools, methods, and technology to generate meaning from data.
- Artificial Intelligence takes this one step further and uses the data to solve cognitive problems commonly associated with human intelligence, such as learning, pattern recognition, and human-like expression.

Data Analysis

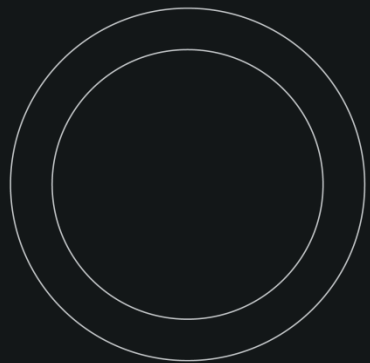
- Data analysis is the process of systematically examining, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making.
- It involves applying statistical and logical techniques to interpret the data and extract meaningful insights.
- Data analysis can be applied in various fields, including business, science, finance, healthcare, and more, to help organizations and individuals make informed decisions based on empirical evidence
- Different tools used for Data analysis include Microsoft Excel, Python, R, Tableau, Power BI, etc.

Data Visualisation

- Data visualisation is the graphical representation of data and information using visual elements like charts, graphs, maps, and infographics.
- The purpose of data visualisation is to make complex data more accessible, understandable, and usable by presenting it in a visual format that reveals patterns, trends, and correlations.
- **Types of Visuals:**
 - **Bar Charts:** Useful for comparing quantities across different categories.
 - **Line Charts:** Ideal for showing trends over time.
 - **Pie Charts:** Display parts of a whole, illustrating proportions.
 - **Scatter Plots:** Show relationships between two variables.
 - **Heatmaps:** Represent data intensity or frequency across a matrix.
 - **Histograms:** Used for showing the distribution of a dataset.

Data Analysis basic terminologies

- **Data Discovery / Data Collection:** Collecting data and putting into categories
- **Data preparation:** Taking raw data and getting it ready for analysis
 - **Data Cleaning:** Removing or correcting inaccurate, incomplete, or irrelevant data.
 - **Data Processing:** Turning unstructured data into data ready for analysis
 - **Data Transformation:** Converting data of one kind into data of another kind
- **Data Visualization:** Transforms data into a variety of charts, graphs, and other graphic solutions.




Numpy



Numpy

- A popular library in Python used for numerical computing.
- NumPy was created in 2005 by Travis Oliphant.
- Numpy provides a powerful n-dimensional array object 'ndarray', which allows you to store and manipulate large datasets efficiently.
- The numpy has built-in mathematical functions, allowing users to perform complex mathematical operations like sqrt, mean and median, easily and efficiently.
- It includes large library of functions for working in various domains like linear algebra, Fourier transform, and matrices.
- The NumPy arrays takes significantly less amount of memory when saving integers, strings, and floating point numbers
- NumPy is optimized for performance, allowing you to perform complex mathematical operations on large datasets efficiently.
- It is written in C and involves compiler optimization.

Numpy Variable Properties

Array  `array([1, 2, 3, 4])`

List  `[1, 2, 'three', 4]`

- 1-D: `array([1, 2, 3, 4])`
- 2-D: `array([[1, 2], [3, 4]])`
- `var.shape`: (4,) for 1-D and (2,2) for given 2-D array
- `var.size`: 4
- `var.dtype`: datatype
- `var.itemsize`: 4, if float, it would be 8

Numpy Basic Operations

For Statistical Analysis:

- `np.max()`
- `np.min()`
- `np.argmin()`
- `np.argmax()`
- `np.sum()`
- `np.mean()`
- `np.std()`

Numpy Operations with multiple arrays

- + (Array or a scalar)
- -
- *
- /
- `np.sqrt(arr)`
- `np.exp(arr)`
- `np.sin(arr)`
- `np.log(arr)`

Numpy Array Creation

For both **arrays** and **matrices**

- `np.array()`
- `np.zeros()`
- `np.ones()`
- `np.eye()`
- `np.linspace()`
- `np.arange()`
- `np.random.rand()`
- `np.random.randn()`
- `np.random.randint()`

Copying **arrays** and **matrices**

- `np.copy()`
- With `=`, a assignment operator



Demo

Thank You!
Your feedback matters

