CfDRA
Center for Data Research & Analytics

Data Science Journey Using R

Jaynal Abedin

PhD in Data Science

15-Mar-2025

Data Science Journey Using R

- Please **write your name in Zoom** Profile
- **Turn off your audio** unless you are talking
- If there is no problem, then we recommend to **keep your video on**
- Please complete the baseline survey from the link following link:
    - ✓ https://cfdra.fillout.com/baseline

Data Science, Machine Learning & Artificial Intelligence

- **Data Science (DS) produces Actionable Insights**
- **Machine Learning (ML) produces Predictions**
- **Artificial Intelligence (AI) produces Actions**

Data Science, Machine Learning & Artificial Intelligence

- **Data Science (DS) produces <u>Actionable Insights</u>**
- **Machine Learning (ML) produces <u>Predictions</u>**
- **Artificial Intelligence (AI) produces <u>Actions</u>**

⚠ **Not everything that fits each definition is a part of that field**

Data Science (DS)

- **In Data Science (DS), there is a human in the loop**
    - ✓ Someone is understanding the insight being produced
    - ✓ Seeing the figures/results and/or benefitting from the conclusions

Data Science (DS)

- **In Data Science (DS), there is a human in the loop**
  - ✓ Someone is understanding the insight being produced
  - ✓ Seeing the figures/results and/or benefitting from the conclusions
- **DS Emphasis on**
  - ✓ Statistical Inference
  - ✓ Data visualization
  - ✓ Experimental design
  - ✓ Domain knowledge
  - ✓ Communication

Data Science (DS)

- **In Data Science (DS), there is a human in the loop**
  - ✓ Someone is understanding the insight being produced
  - ✓ Seeing the figures/results and/or benefitting from the conclusions
- **DS Emphasis on**
  - ✓ Statistical Inference
  - ✓ Data visualization
  - ✓ Experimental design
  - ✓ Domain knowledge
  - ✓ Communication

The goal in DS is to **gain better understanding** of the data

Example: Workshop Registration

The use of data science in designing/refining the contents of this workshop

- **The actionable insights**
  - ✓ **Gain knowledge** about self reported R programming proficiency
  - ✓ Extract **expectations** of the participants from the response why they want to take this workshop

Example

Extracted themes

- **Data Science & Learning:** It represents general interest in data science, learning pathways, and career aspirations related to becoming a data scientist

- **Research & Application of R:** It centers around the application of R in research and statistical analysis, with interest in improving research-related skills

- **Programming & Practical Experience:** This theme captures the desire to gain hands-on experience with R programming and attend workshops to deepen understanding

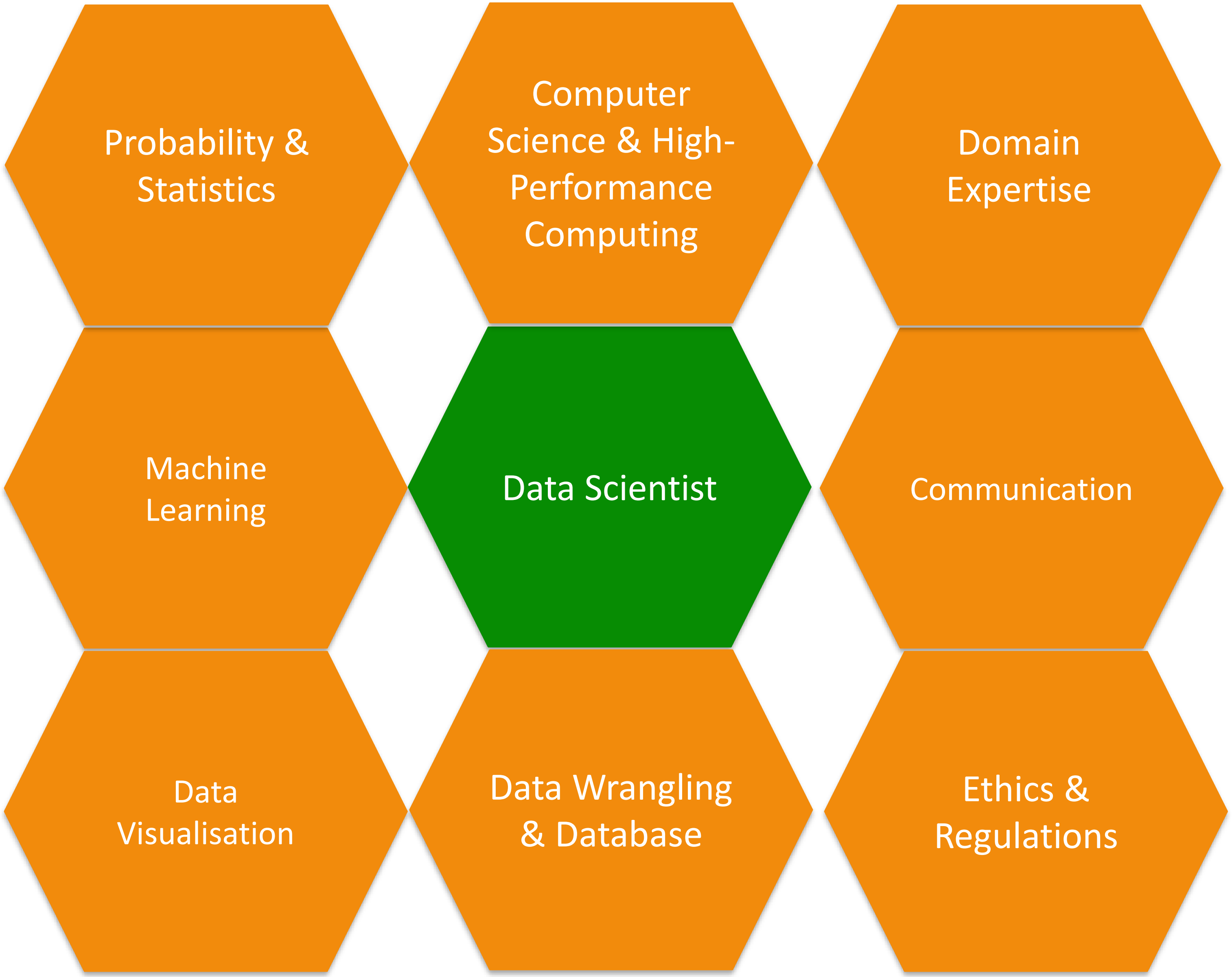
Components of Data Science

- Data Collection
- Data Engineering (Data Cleaning and Preparation)
- Data Exploration and Visualization
  - ✓ Data Analysis
  - ✓ Data Visualization
- Statistics and Probability (Machine Learning)
- Big Data Technologies
- Domain Expertise
- Experimentation and Optimization
- **Communication and Storytelling**
- **Ethics and Responsible AI**
- Software Engineering

Skills for a Data Scientist

Data Scientist

Skills for a Data Scientist



Probability & Statistics

Computer Science & High-Performance Computing

Domain Expertise

Machine Learning

Data Scientist

Communication

Data Visualisation

Data Wrangling & Database

Ethics & Regulations

Data Science: Myth Vs Reality

Myth

Data Science is
Autonomous Process

Data Science: Myth Vs Reality

Myth

Data Science is Autonomous Process

Reality

Requires skilled human oversight throughout the different stages of the process

- ✓ to frame the problem
- ✓ to design and prepare the data
- ✓ to select appropriate algorithms
- ✓ to critically interpret the results
- ✓ to plan the appropriate action based on the results

Without skilled human expertise, a DS project will fail to meet its targets

Data Science: Myth Vs Reality

Myth

- Modern **data science software is easy to use**, and **so data science is easy to do**

Data Science: Myth Vs Reality

Myth

- Modern **data science software is easy to use**, and **so data science is easy to do**

Reality

- In fact, **it has never been easier to do data science badly**, ***"Garbage in Garbage out"***
- The ease of use can hide the requirements of
  - ✓ appropriate **domain knowledge** and the expertise regarding the properties of the data and
  - ✓ the **assumptions** underpinning the different models/algorithms

"Data mining lets computers do what they do best—dig through lots of data. This, in turn, lets people do what people do best, which is to set up the problem and understand the results" – Gordon Linoff & Michael Berry

Data Science: Myth Vs Reality

Myth

- Projects needs **big data** and needs to use **advanced machine learning** (e.g., Deep Learning)

Data Science: Myth Vs Reality

Myth

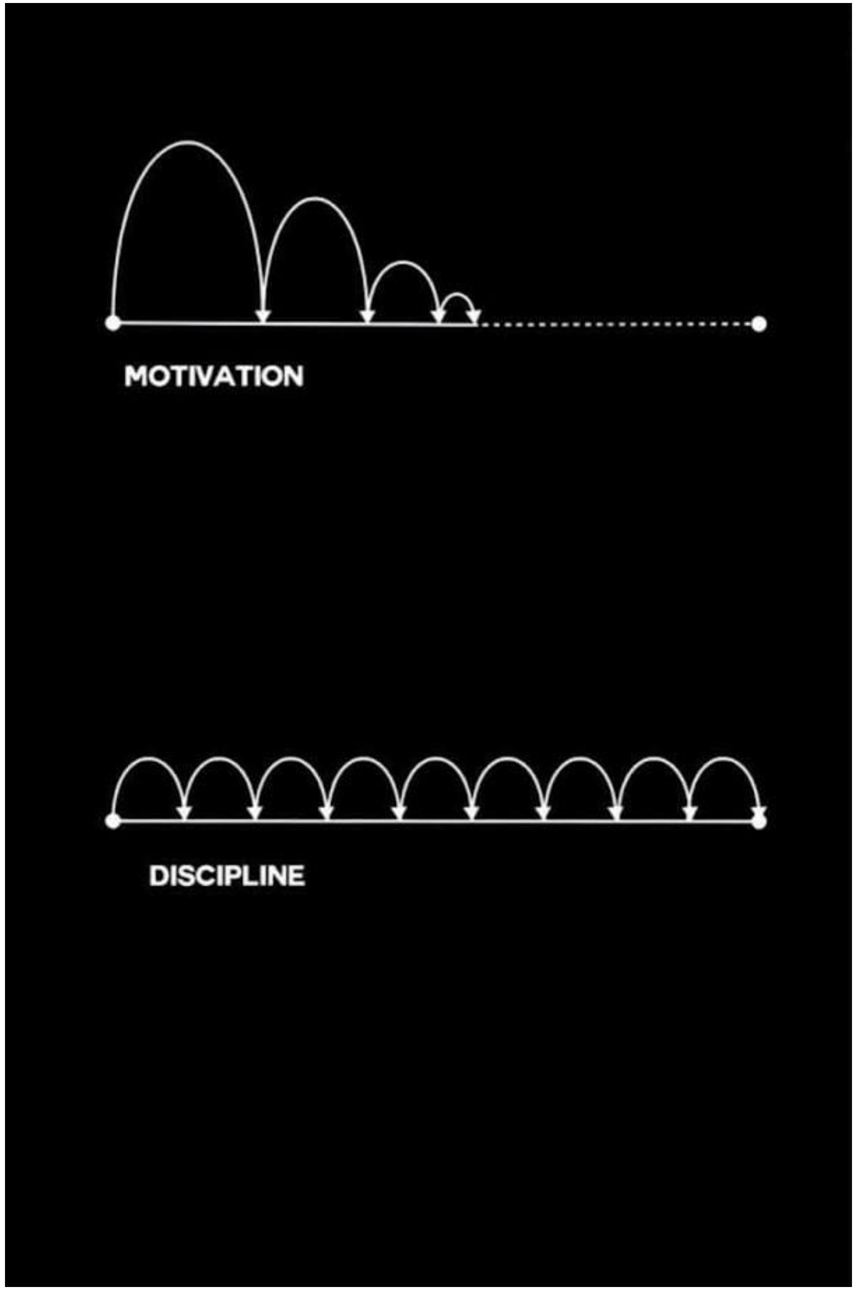
- Projects needs **big data** and needs to use **advanced machine learning** (e.g., Deep Learning)

Reality

- Having ***more data helps***, but **having right data** is the more important requirement

- It is frequently the case that the real value of a DS project is **deriving one or more variables** that provide meaningful insights into a problem

  ✓ Air Quality Monitoring Device Data
  ✓ Wearable device data
  ✓ Motion Tracking Data



MOTIVATION
DISCIPLINE

Research Questions to High-Quality Data

Question/Objective → Data

Question

- What is the proportion of under 5 children suffers from anemia in Bangladesh in 2021?

Population

- All live children <5 years in Bangladesh in 2021

Parameter(s)

- Proportion of <5 children with anemia

Question/Objective → Data

Data Requirement(s)

Measure Anemia

- Collaborate with domain expert and define Anemia so that it is measurable
- Decide appropriate measurement scale to measure Anemia
- Plan a strategy to collect Anemia in a way so that it is representative to the study population

Database

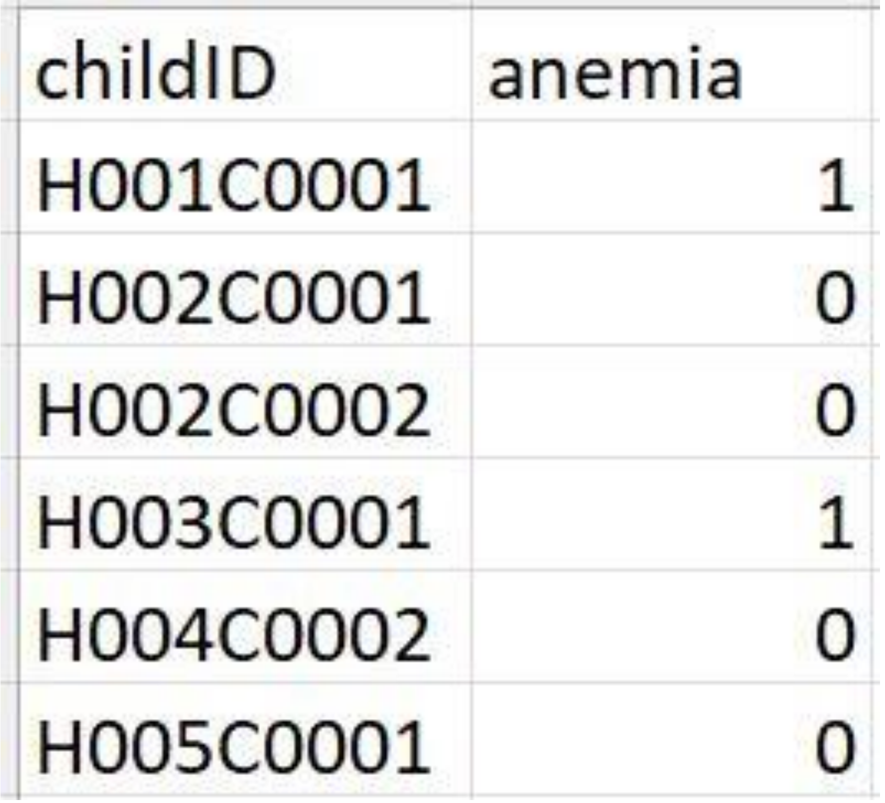
- Create a database and define each variable’s characteristics before starting data entry
- Create data dictionary to guide data collection and data analysis

Question/Objective → Data

Data Dictionary

Name of Variable	Variable Label	Type	Possible Values	Value Label (if any)
childID	ID of surveyed child	Character of length 9 with a format HxxxCyyyy	H001C0001	
anemia	Status of Anemia	Numeric with single digit	0 or 1	1 = Yes, 0 =No

Question/Objective → Data



childID	anemia
H001C0001	1
H002C0001	0
H002C0002	0
H003C0001	1
H004C0002	0
H005C0001	0

- Count 1’s in **anemia** column
- Divide the count of 1’s with total number of children surveyed to get the estimated proportion

Question/Objective → Data

Question

- What is the proportion of under 5 children suffers from anemia in Bangladesh in 2021?

Additional Question

- What was the difference in proportion of under 5 children suffers from anemia in Bangladesh in 2021 between
  - ✓ **Rural Vs Urban**
  - ✓ **Male Vs Female**
  - ✓ **< 2 yeas vs >2 years**

Question/Objective → Data

Question

- What is the proportion of under 5 children suffers from anemia in Bangladesh in 2021?

Additional Question

- What was the difference in proportion of under 5 children suffers from anemia in Bangladesh in 2021 between
  - ✓ **Rural Vs Urban**
  - ✓ **Male Vs Female**
  - ✓ **< 2 yeas vs >2 years**
- What was the distribution of proportion of under 5 children suffers from anemia in Bangladesh in 2021 between
  - ✓ **Administrative divisions**
  - ✓ **Administrative districts**

Question/Objective → Data

Extended Data Dictionary

Name of Variable	Variable Label	Type	Possible Values	Value Label (if any)
`childID`	ID of surveyed child	Character of length 9 with a format `HxxxCyyyy`	`H001C0001`	
`anemia`	Status of Anemia	Numeric with single digit	`0 or 1`	`1 = Yes, 0 =No`

Question/Objective → Data

Extended Data Dictionary

Name of Variable	Variable Label	Type	Possible Values	Value Label (if any)
`childID`	ID of surveyed child	Character of length 9 with a format `HxxxCyyyy`	`H001C0001`	
`anemia`	Status of Anemia	Numeric with single digit	`0 or 1`	`1 = Yes, 0 =No`
`division`	Divisions			
`districts`	Districts			
`ruralUrban`	Rural or Urban			
`childSex`	Gender of Child			
`age`	Age of Child in Months			

Objective → Data – Example from Cricket

Objective

- To analyze the impact of **workload** and **biomechanics** on **injury risk in fast bowlers in Cricket** and develop an **injury prevention strategy**

Refined Objective

- **Fast bowlers** are ***prone to stress fractures***, ***hamstring injuries***, and ***lower-back issues*** due to repetitive ***high-impact actions***. The goal is **to identify workload patterns** and biomechanical factors that increase injury risk

Objective → Data – Example from Cricket

Data Requirement(s)

Bowling Workload Data (From GPS Trackers & Match Logs)

- ✓ Number of overs bowled per session (training & matches)
- ✓ Total number of deliveries bowled per day/week/month
- ✓ Bowling speed trends (does pace drop with fatigue?)
- ✓ Sprinting between wickets (for all-rounders)

Biomechanical Data (From Motion Capture & Force Plates)

- ✓ Ground reaction force at front-foot landing
- ✓ Hip and shoulder alignment at delivery stride
- ✓ Knee flexion angles and lower back stress
- ✓ Wrist position and release biomechanics

Objective → Data – Example from Cricket

Data Requirement(s)

Injury & Recovery Data (From Medical & Physiotherapy Reports)

- ✓ Type of injuries (stress fractures, muscle strains, ligament damage)
- ✓ Time lost due to injury
- ✓ Fatigue and recovery assessments (sleep, soreness, hydration levels)

Contextual Data

- ✓ Age and bowling style (fast, swing, seam)
- ✓ Pitch conditions (harder pitches put more stress on joints)
- ✓ Match formats (Test, ODI, T20 – workload varies significantly)

Objective → Data – Example from Cricket

Data Collection & Storage

- Wearable sensors & GPS trackers measure workload and movement patterns
- High-speed cameras & motion tracking capture biomechanical deviations
- Medical staff & physiotherapists log injury data and recovery timelines
- Coaches monitor match workloads to ensure workload balance

Database & Data Dictionary

- Create data dictionary and relate each of the components mentioned above

Group Work

Group Work Task: Data Requirements & Data Dictionary

Choose one task from **Group Work Task: Data Requirements & Data Dictionary**

Discuss among group members and create a data dictionary and submit into GitHub Repository as a Markdown (.md) file.