

Data Science

Dr. Nitesh Funde,
Department of Artificial Intelligence,
SVNIT, Surat

Introduction

- Examples
- Overview of the Data Science Process.
- Applications and Results Obtained Using Data Science Techniques

Data Science

- Data Science is the most sought after job of the twenty first century !
- Data is the new oil and Data Science is combustion engine that drives it !
- Data Science is the future !
- But what exactly is Data Science !

Learning Objectives

- What exactly is Data Science ?
- Why is it such as sought after job description ?
- What does a Data Scientist actually do ?
- How important are mathematics and programming skills for a data scientist ?
- How does Data Science relate to other buzzwords such as ML, DL, AI and DM ?
- What are some common misconceptions about Data Science?

Data Science Process

- Data Science is the science of collecting, storing, processing, describing and modelling data



1. Collecting Data



- What is involved in data collection?
 - Depends on the question a data scientist is trying to answer
 - Depends on the environment in which the data scientist is working

Collecting Data-Applications









Collecting Data-Applications

2. A Data Scientist working for a political party



What are people saying about the new policy?



Data already exists Needs to be crawled, scraped

Collecting Data-Applications

3. A Data Scientist working with farmers



Effect of type of seed, fertiliser, irrigation on yield?



Data not available Needs to design experiments

2. Storing Data

1) Transactional and Operational Data

- ✓ patient records
- ✓ insurance claims
- ✓ inventory
- ✓ customer records
- ✓ telephone bills
- ✓ invoices
- ✓ employee records
- ✓ reimbursements
- ✓ purchase orders

Emp ID	Name	Role	Salary	Email
00001	ABC	CEO	100\$	abc@a.com
00002	XYZ	CTO	100\$	xyz@abc.com
...

Structured Data



Relational
Databases
(select, insert,
update, delete)

Storing Data

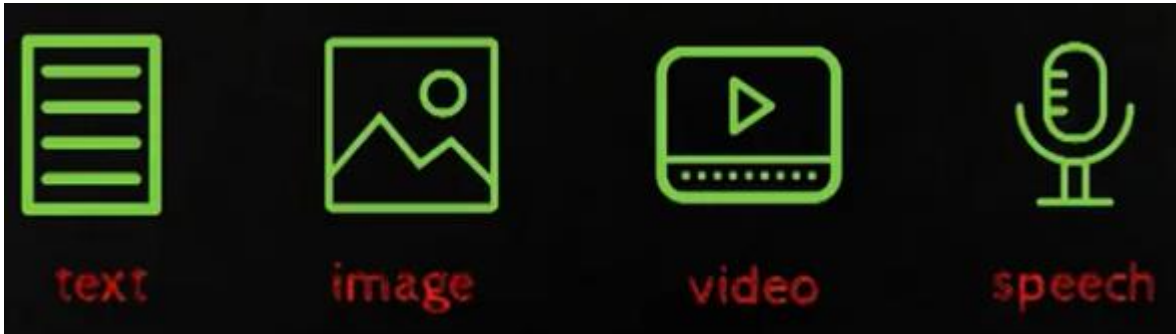
2) Data from multiple databases



It's a collection of many relational databases. It supports and is optimized for Analytical operations.

Storing Data

3. Unstructured Data



- High volume
- High variety
- High velocity

Storing Data -Summary



- Structured
- Optimised for SQL queries

Structured
curated
optimised for analytics

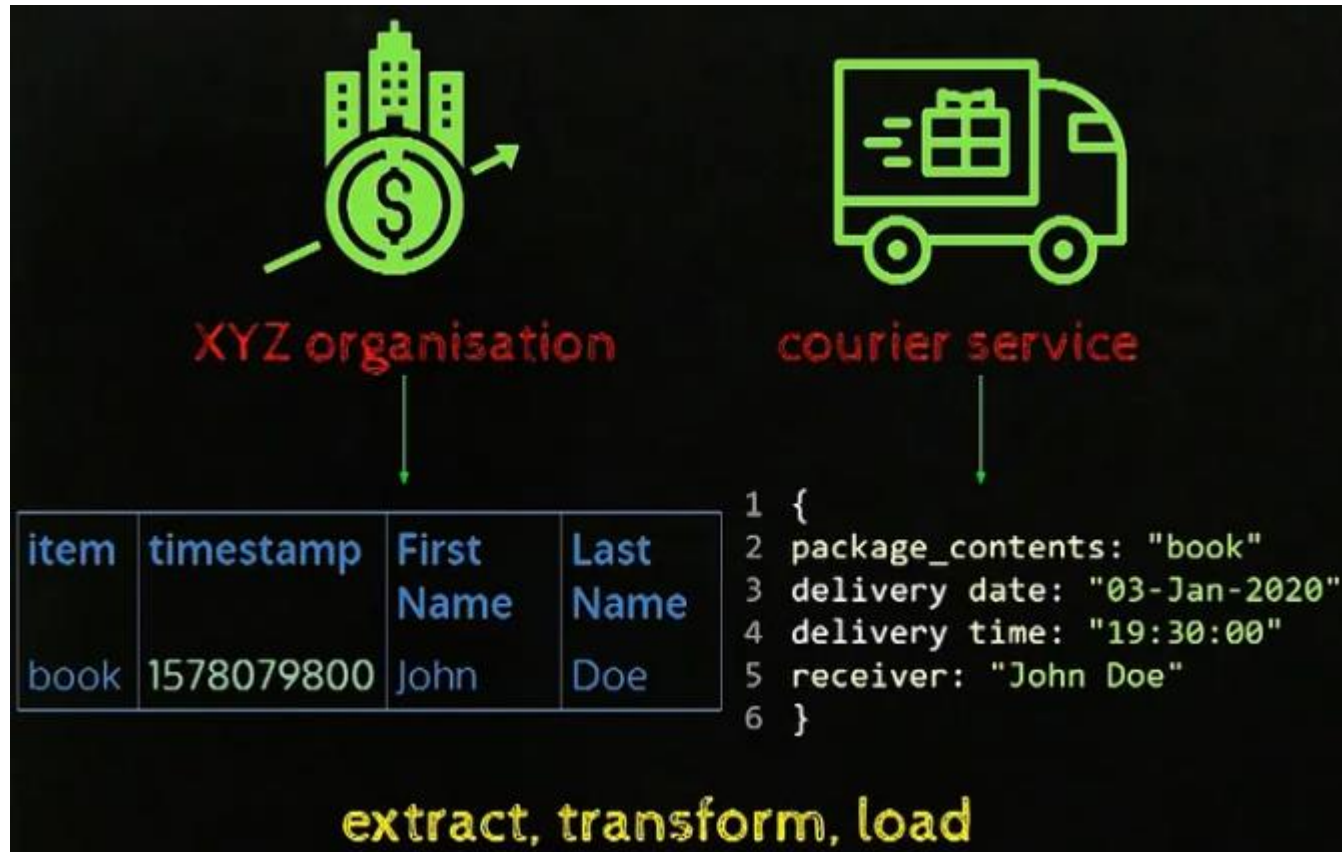
Big data
Uncurated

Storing Data

- Skills required
 - Programming and Engineering
 - Knowledge of Relational Databases
 - Knowledge of NoSQL Databases
 - Knowledge of Data Warehouses
 - Knowledge of Data Lakes (Hadoop)

3. Processing Data

3.1 Data Wrangling or Data Munging

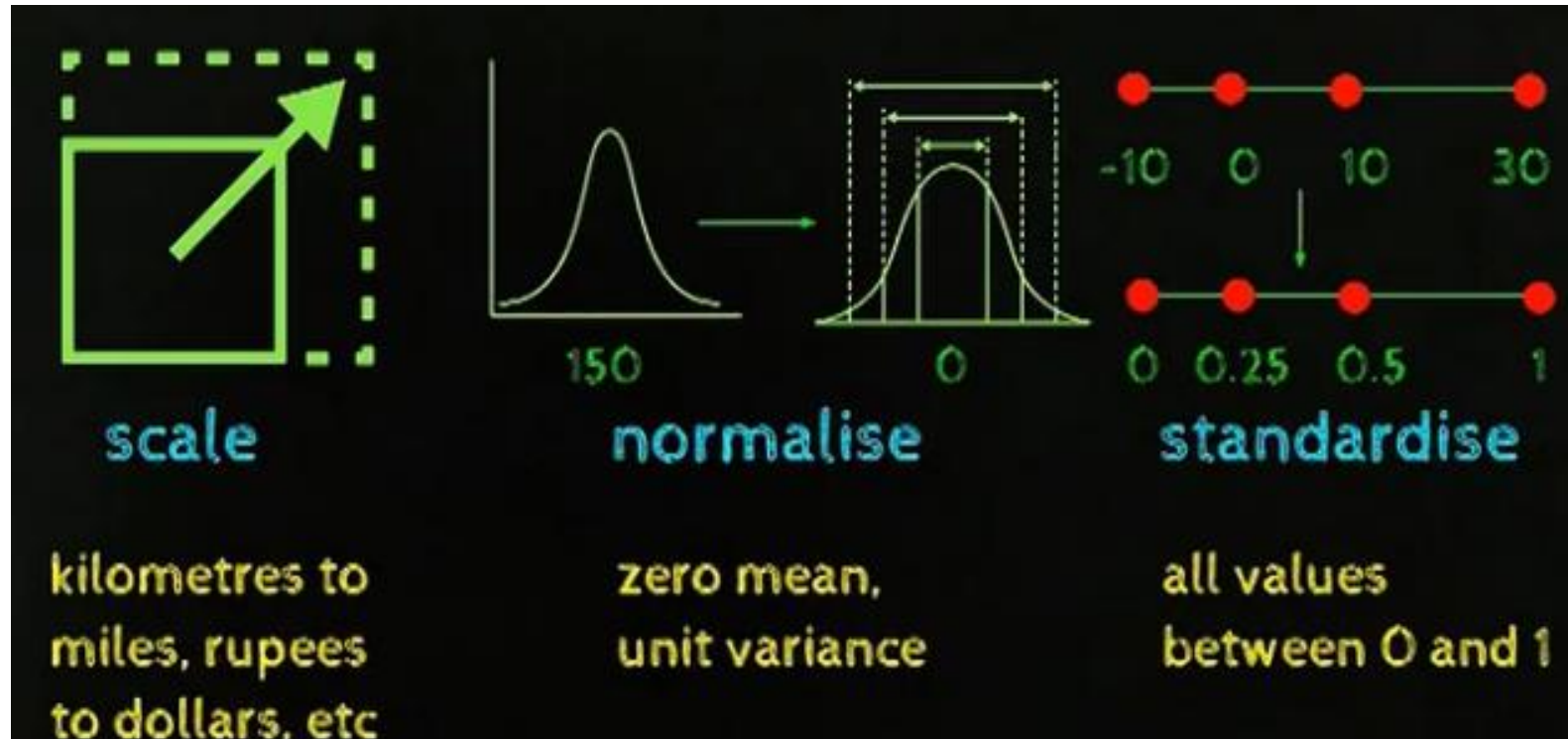


3. Processing Data

- 3.2 Data Cleaning
 - Fill Missing Values
 - Standardise Keywords
 - Correct Spelling Errors
 - Identify and Remove Outliers

3. Processing Data

3.3 Data Scaling, Normalising and Standardising



3. Processing Data

“ If data processing is to be performed on Big Data with millions of data items then performance becomes a key consideration



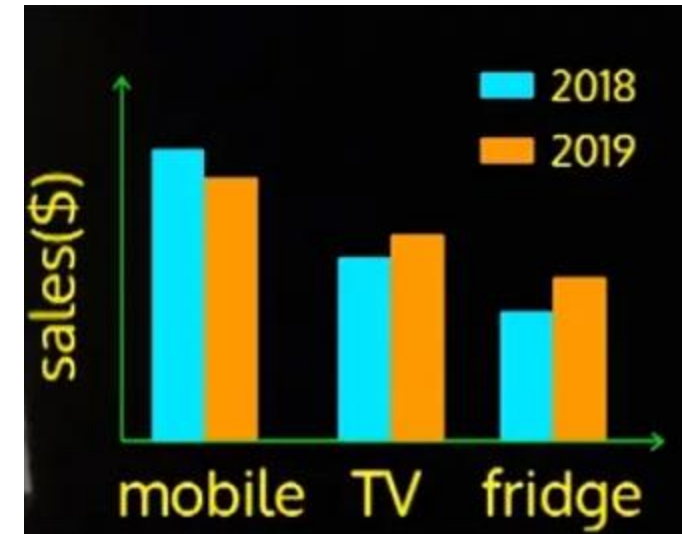
Distributed Processing Hadoop (Map Reduce)

Skills Required

- Programming Skills
- Map Reduce (Hadoop)
- SQL and NoSQL Databases
- Basic Statistics

4. Describing Data

4.1 Visualising Data



4. Describing Data

- 4.2 Summarising Data



monthly sales record

What is the *typical* # of TVs sold daily?

✓ mean ✓ median ✓ mode

What is the *typical* variation in # of TVs sold daily?

✓ std. deviation ✓ variance

- ✎ Descriptive Statistics
- ✎ Iterative Process
- ✎ Exploratory Data Analysis

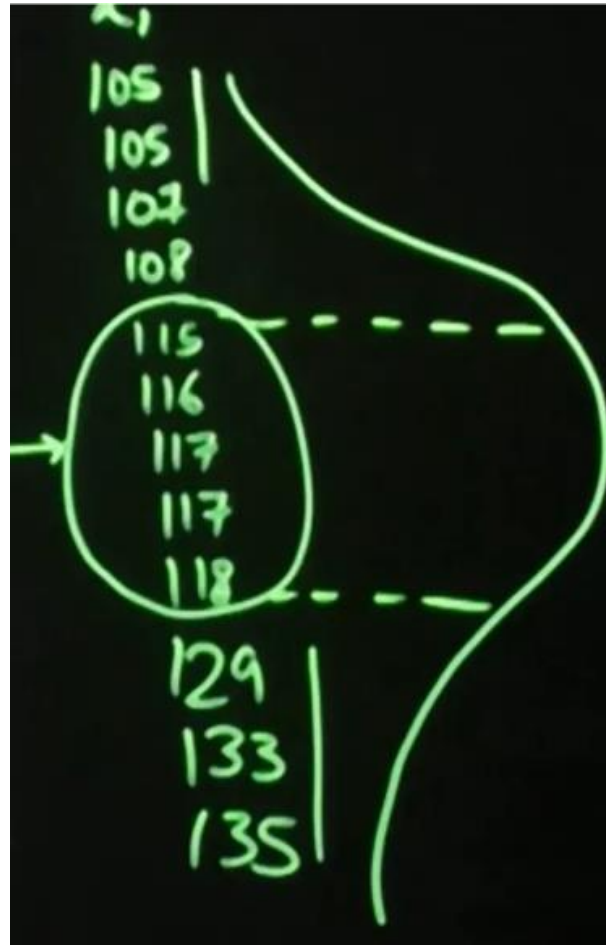
Skills required

- ✎ Statistics
- ✎ Excel
- ✎ Python
- ✎ R
- ✎ Tableau

5. Modelling

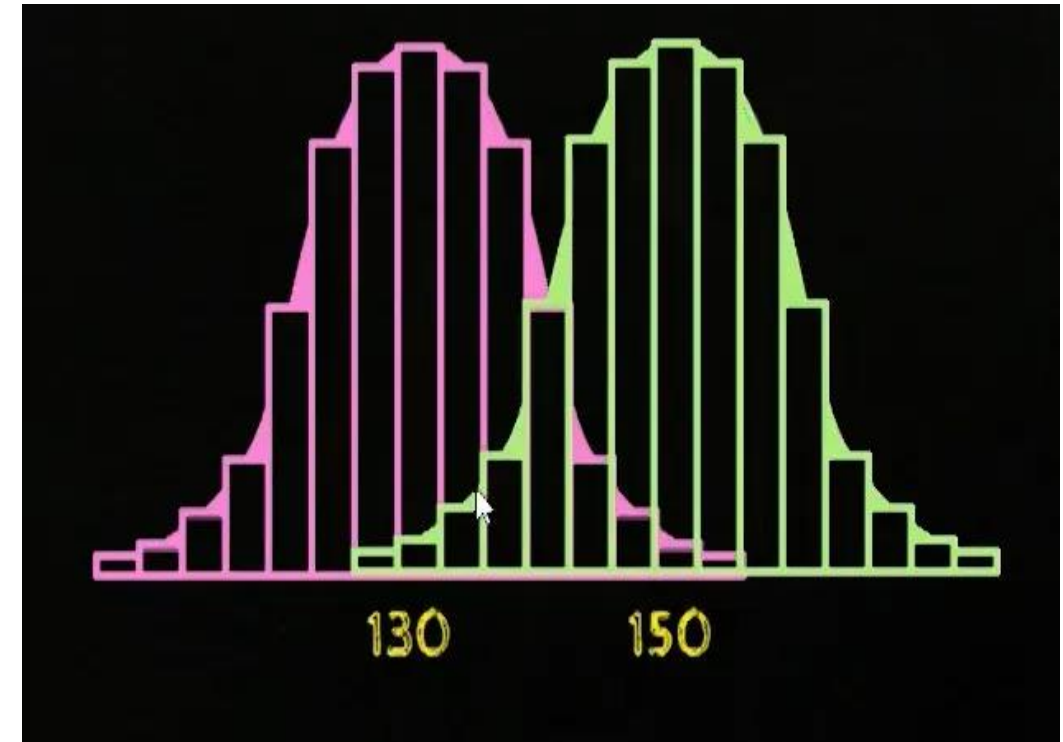
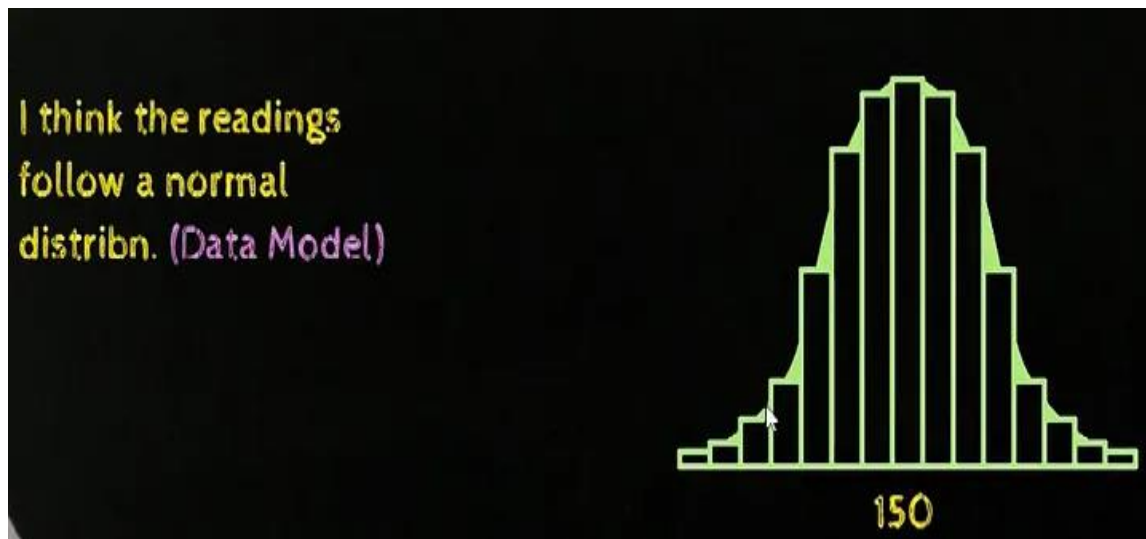
- Statistical Modelling

X_1
105
105
107
107
108
115
116
117
117
118
129
133
135



5. Modelling

- Statistical Modelling



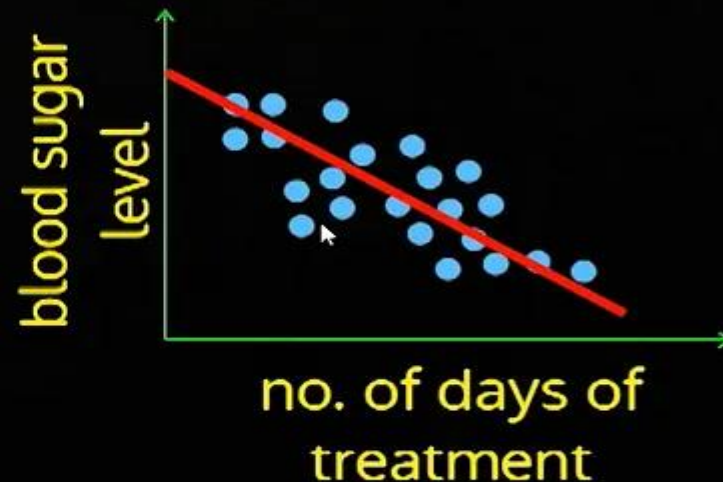
5. Modelling

- Statistical Modelling



I am 99% sure that the sugar level drops by 3 ± 1 points for each day of treatment

I think there is a linear relationship between the no. of days of treatment and blood sugar level (Data Model)



5. Modelling

Statistical Modelling

- ✓ Modelling underlying data distribution
- ✓ Modelling underlying relations in data
- ✓ Formulate and test hypotheses
- ✓ Give statistical guarantees (p-values, goodness-of-fit tests)



- ✓ In Statistical Modelling, we assumed simple models which allowed robust statistical analysis
- ✓ Give statistical guarantees (p-values, goodness-of-fit tests)

Algorithmic Modelling

$$y = f(x)$$

blood sugar level
after 30 days



[age, weight, height, bloodpressure, ...]



[age, weight, height, bloodpressure, ...]

[..., ..., ...]

$$y = m_1x_1 + m_2x_2 + \dots + m_nx_n$$

$$= f(x_1, x_2, \dots, x_n)$$

Algorithmic Modelling

$$y = f(x)$$

blood sugar level
after 30 days



[age, weight, height, bloodpressure, ...]



[age, weight, height, bloodpressure, ...]

[..., ..., ...]

- ✓ Estimate f using data, optimisation techniques
- ✓ For a new patient plug-in the value of x to get y
- ✓ Focus on prediction (don't care about underlying phenomena)

DS, ML and DL

Statistical Modelling v/s Algorithmic Modelling

Simple, intuitive models	Complex, flexible models
More suited for low-dimensional data	Can work with high-dimensional data
Robust statistical analysis is possible	Not suitable for robust statistical analysis
Focus on interpretability	Focus on prediction
Data lean models	Data hungry models
More of Statistics	More of ML, DL

“When you have large amounts of high-dimensional data and you want to learn very complex relationships between the output and input use a specific class of complex ML models and algorithms, collectively referred to as Deep Learning

DS, ML and DL

Statistical Modelling

Linear Regression,
Logistic Regression,
Linear Discriminant
Analysis

$$\underline{\underline{y = f(x)}}$$

Algorithmic Modelling

Linear Regression,
Logistic Regression,
Linear Discriminant
Analysis,
Decision Trees, K-NNs
SVMs, Naive Bayes,
Multilayered Neural
Networks

DS, ML and DL

“When you have large amounts of high-dimensional data and you want to learn very complex relationships between the output and input use a specific class of complex ML models and algorithms, collectively referred to as Deep Learning


Skills required

- ✎ Inferential Statistics
- ✎ Probability Theory
- ✎ Calculus
- ✎ Optimisation algorithms
- ✎ ML and DL
- ✎ Python packages and frameworks (numpy, scipy, scikit-learn, TF, PyTorch, Keras)


About Data Science

Why is DS so popular today?


1. Data is everywhere



Personal devices



Sensors



Transactional Data
(Digital revolution)

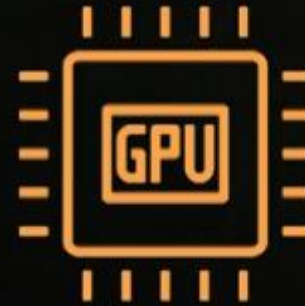
“Keen interest in converting data into insights!”

About Data Science

2. Devices have become powerful and cheaper



Bulk storage



Specialised hardware

“Within the last decade the cost of bulk storage has reduced by over 6 times and GPUs have become 100 times more capable!

About Data Science

3. Democratisation of software and hardware



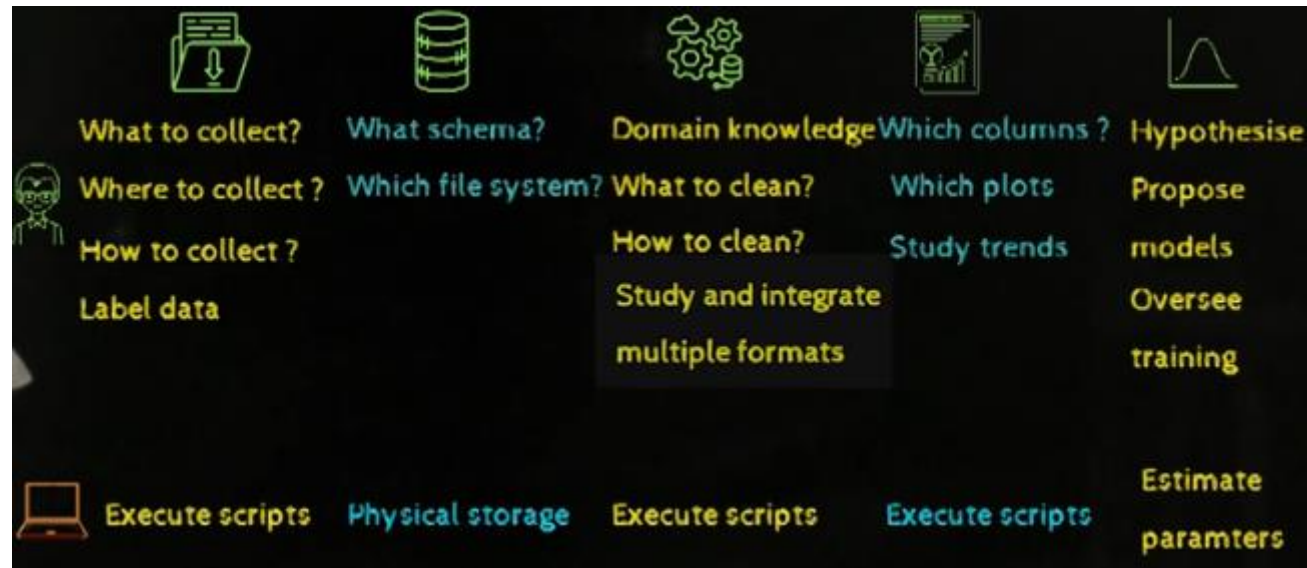
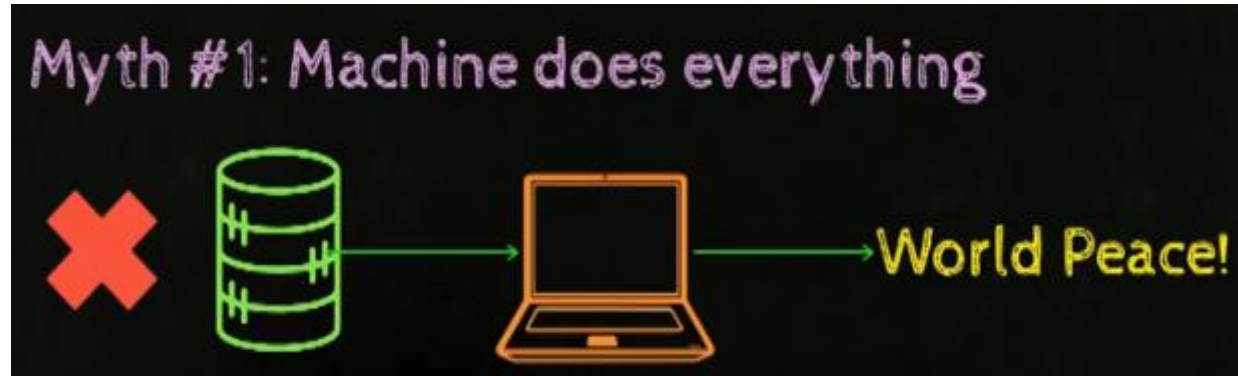
Software



Cloud compute


“It is relatively easy for a single data scientist to setup complete stacks on the cloud which were beyond reach to even large companies a few years back!

Myths of Data Science



Myths of Data Science

Myth #2: DS requires Big Data and DL



The diagram illustrates the myth that Data Science requires Big Data, Deep Learning, and Hardware. It shows a sequence of icons: a large red 'X' over a 'Big data' icon (which includes a bar chart, a database cylinder, and a code editor with JSON and XML files), followed by a plus sign, a 'Deep Learning' icon (a stylized 'F' shape), another plus sign, and a 'Hardware' icon (a square with 'GPU' inside and pins around it). This sequence is followed by an equals sign and the text 'Data Science'. Below the icons, the labels 'Big data', 'Deep Learning', and 'Hardware' are written in green.

Big data + Deep Learning + Hardware = Data Science

Example: A rural school with data of less than 500 students

- Do more girls dropout from school than boys?
- Do students really find maths to be harder than social science?
- Do students staying farther from school perform poorly?

DS Successful

Myth #3: DS is always successful



Data
Science

always



Reasons why it could fail

No meaningful insights in data

No actionable insights in data

Noisy data

Not enough data

“

If the right amount of clean usable data is available, if skilled data scientists with technical and domain knowledge are available, and if the organisation has the capacity and resources to act on the insights generated from the data then data science can be successful and impactful.

Are AI and DS related ?

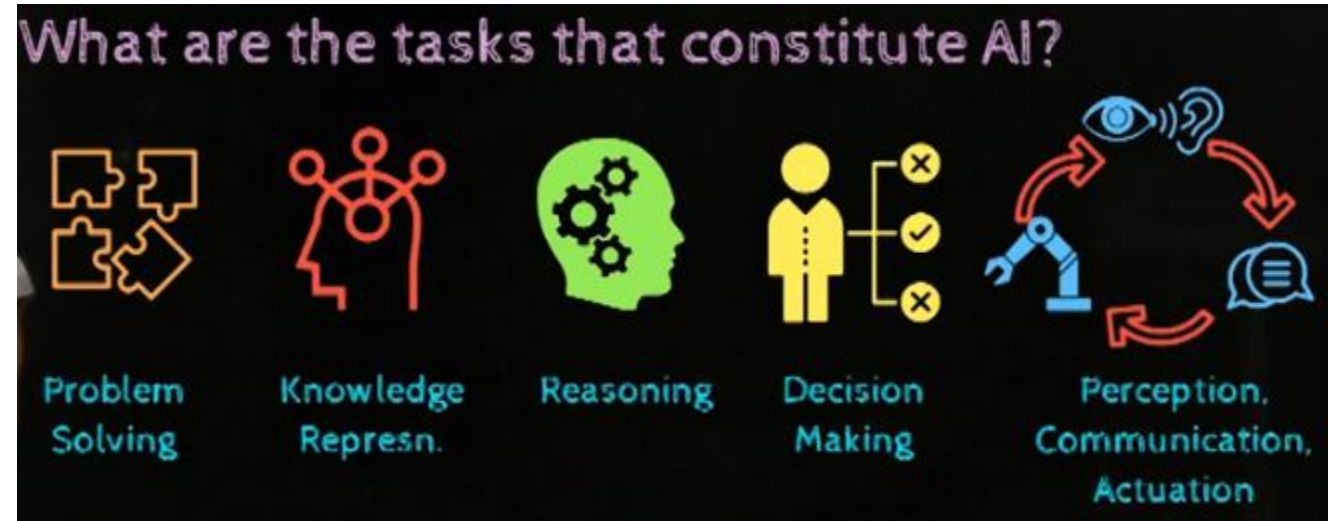
What is the confusion ?

- ✗ AI and DS are synonymous
- ✗ One is a subset of the other
- ✗ AI and DS are completely unrelated

“Confusion arises due to non-technical and broad usage of these terms

AI is about building systems or agents that demonstrate "intelligence"

AI



Are AI and DS related? If so, how?

DS: I have data what do I do with it?



collect



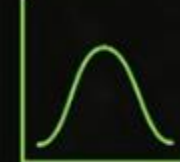
store



process



describe



model

AI: I want an intelligent agent! What do I do?



Problem
Solving



Knowledge
Represn.



Reasoning



Decision
Making



Perception,
Commn.,
Actuation

DS: I have data what do I do with it?



collect



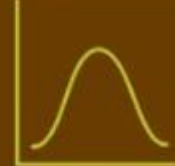
store



process



describe



model



Data-driven



Problem
Solving



Knowledge
Represn.



Reasoning



Decision
Making



Perception,
Commn.,
Actuation

References

- <https://padhai.onefourthlabs.in/courses/data-science>