24 July 2021
July 2021

Deepesh Wadhwani

Introduction to Data Science

INSAID

LMS
learning management system.

insaid.co

VIC
VINC

To: All pan and attendies

MATH.

stats starter kit

0 — can't stand it.
1
2
3
4 — ok with math.
5 — love it.

0- prefix

Programming.

Python starter kit

0
1
2 — NEW to programming
3
4
5 — few.

code 1
code 2
code 3

Manfacting.
Avionics
Auto
Financial.
BFSI.
IT
Electronics - IOT
Telco
Merchant Navy
Edu.
Media.
Ecomm
Energy
Pharma.
backend prog.
Supply chain
Real Estate
Logistics
Health care
Sales
Human Resources Mng

Course Overview

Term	CDF	GCD	GCDAI	PGPDSAI
Term 1	Data Analytics with Python	Data Analytics with Python	Data Analytics with Python	Data Analytics with Python
Term 2	Data Visualization Techniques	Data Visualization Techniques	Data Visualization Techniques	Data Visualization Techniques
Term 3	EDA & Data Storytelling	EDA & Data Storytelling	EDA & Data Storytelling	EDA & Data Storytelling
		Minor Project	Minor Project	Minor Project
Term 4		Machine Learning Foundation	Machine Learning Foundation	Machine Learning Foundation
Term 5		Machine Learning Intermediate	Machine Learning Intermediate	Machine Learning Intermediate
Term 6		Machine Learning Advanced (Mandatory)	Machine Learning Advanced (Mandatory)	Machine Learning Advanced (Mandatory)
		Data Visualization with Tableau (Elective - I)	Data Visualization with Tableau (Elective - I)	Data Visualization with Tableau (Elective - I)
		Data Analytics with R (Elective - II)	Data Analytics with R (Elective - II)	Data Analytics with R (Elective - II)
		Capstone Project	Capstone Project	Capstone Project
Term 7		Bonus: Industrial ML (ML – 4 & 5)	Basics of AI, TensorFlow, and Keras	Basics of AI, TensorFlow, and Keras
Term 8			Deep Learning Foundation	Deep Learning Foundation
Term 9			NPL – I/CV – I	CV – I
Term 10			NLP – II/CV – II	NLP – I
			Capstone Project	Capstone Project
Term 11				CV – II
Term 12				NLP – II
				NLP – III + CV – III
				AutoVision & AutoNLP
				Building AI product

You are here...

INSAID

Term Context

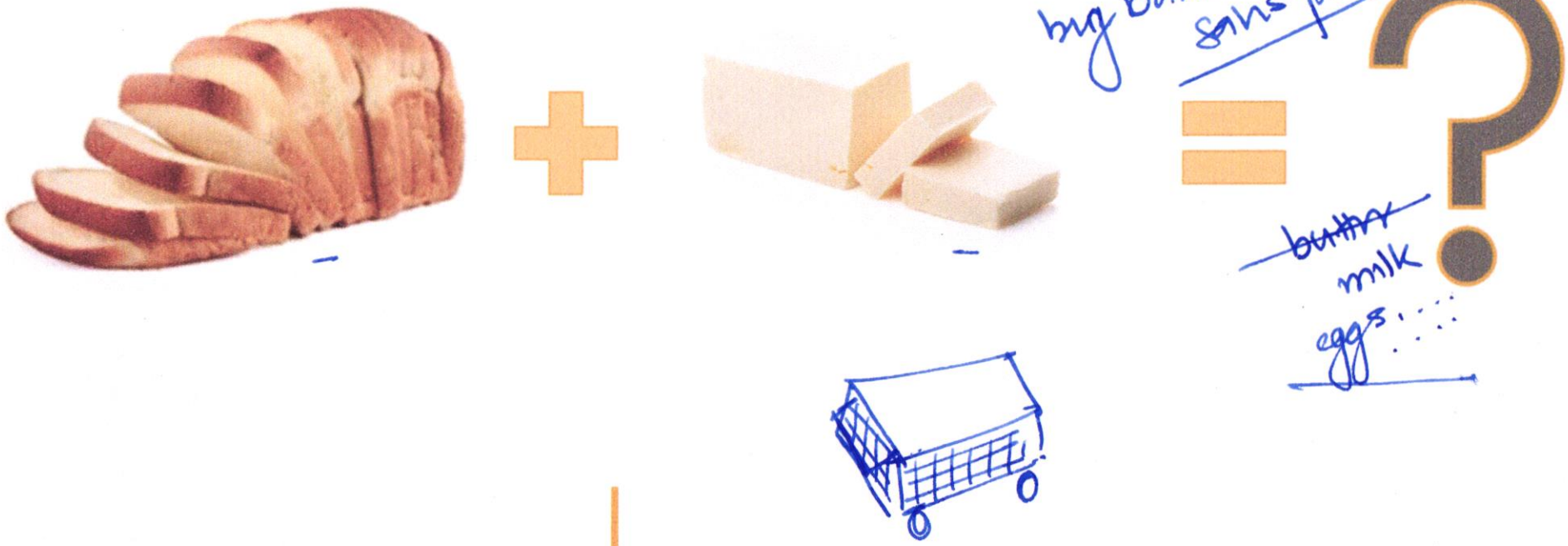
- **Data Science Fundamentals** ← You are here...
- Recap: Python for Data Science
- Recap: Statistics for Data Science
- Data Operations with Numpy
- Data Manipulation with Pandas

INSAID

Presentation Scope

- This **presentation** is only **limited** to the **high level overview.**
- If you get **any doubts** related to the topic, **do not worry...**
- All the **topics** will be **covered in depth** in the **upcoming terms.**
- It is **natural** to have **doubts early** in the **journey** of any domain.
- But remember, **Rome was not built in a day...**

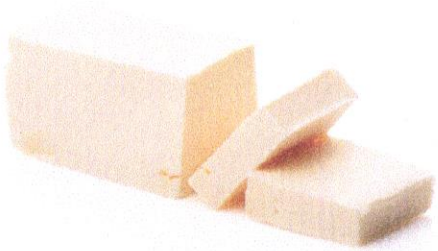
INSAID



Have You Seen This?

- People often buy bread and butter in grocery stores.
- What do you think customer will buy next?

INSAID



Have You Seen This?

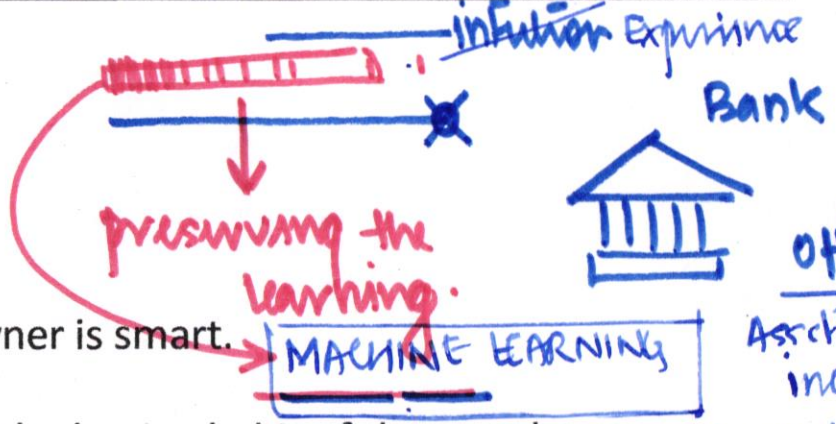
- It's the Milk.
- It is generally seen that people who buy bread and butter also likes to buy milk for breakfast.

INSAID

'learning'

Purchase Behavior			Profile	Also Buys	Percentage
Chocolates	Biscuits	Candies	kids.		88%.
Movies	Beer	Popcorn	single M/F		90%.
Stroller	Beer	Diapers	parent		—

1. 'data' and stats
2. basic program
3. domain



How Grocers Recommend?

- Grocer owner is smart.
- He knows the buying habit of the people.
- He records the buying behavior of the people.

INSAID

Purchase Behavior			Profile	Also Buys	Percentage
Choclates	Biscuits	Candies	Kids	Chips	97%
Movies	Beer	Popcorn	Single M/F	Biscuits	98%
Stroller	Beer	Diapers	Parent M/F	Baby Lotion	91%

(New Recommendation)

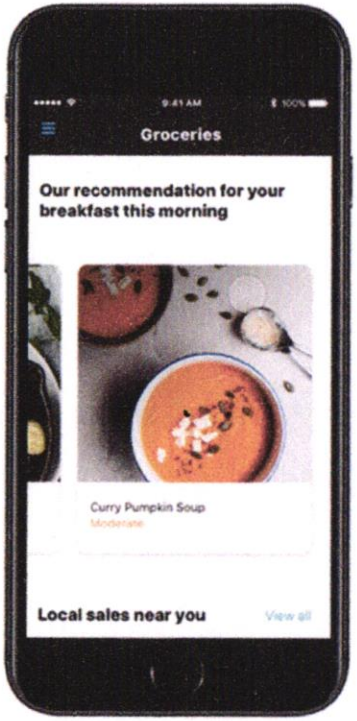
Movies	Beer	Popcorn	?	?	?

How Grocers Recommend?

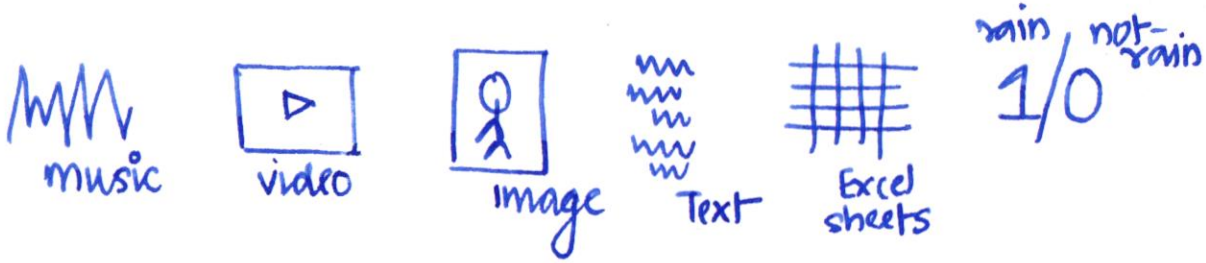
- He will recommend customers based on those buying habits.

INSAID

Amazon Recommendation



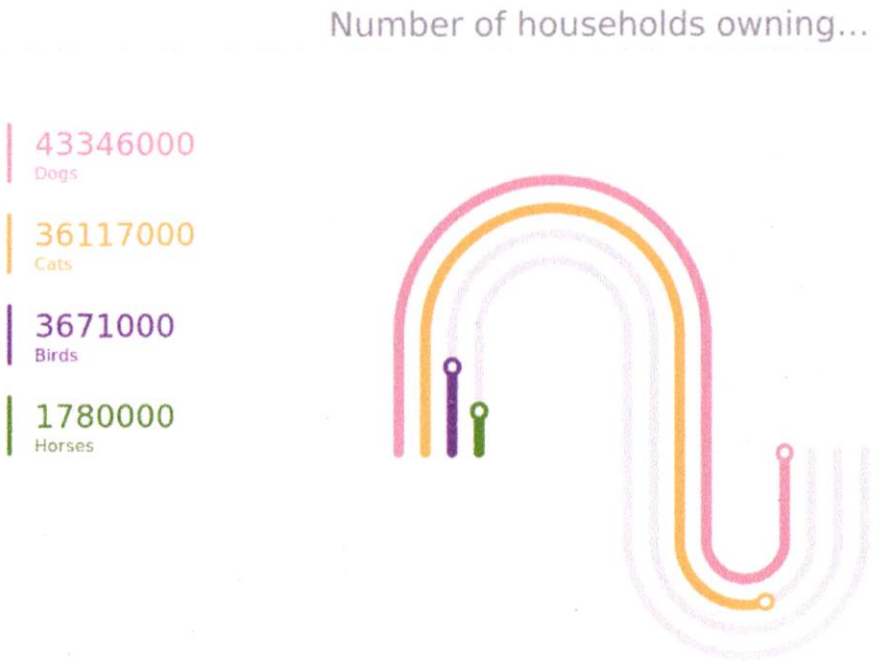
- Amazon's Recommendation Engine works in similar manner.
- **35%** of Amazon sales are generated through its **Recommendation Engine.**



anything that can be recorded

What is Data?

- **Raw or isolated facts** from which the required information is produced.
- Facts can be **recorded** & have **implicit meaning.**
- **Example:** Number of household owning of Dogs, Cats, Birds, Horses.

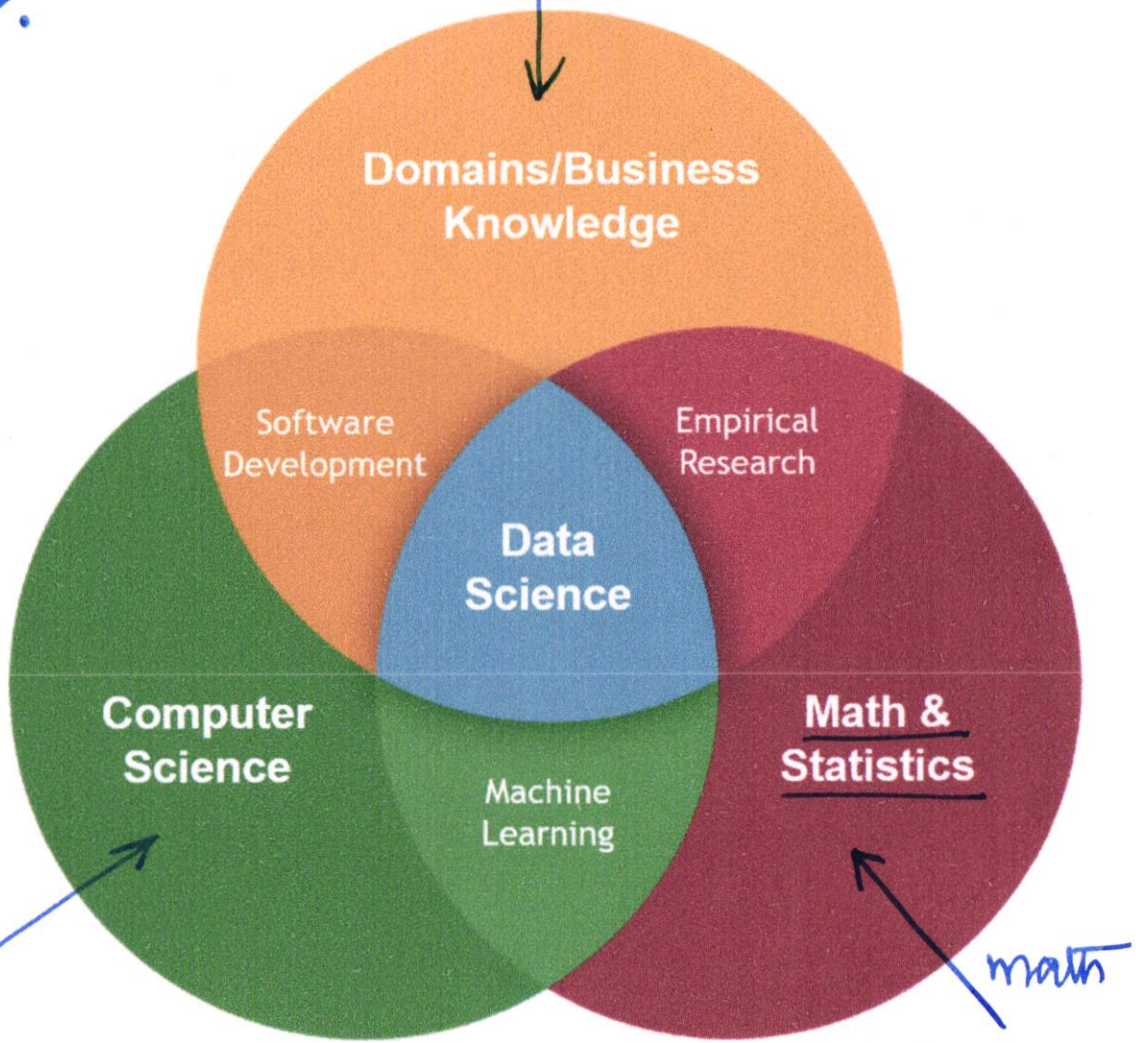


INSAID

language may come. and language may go
but i (logic) go on for ever.

Data Science?

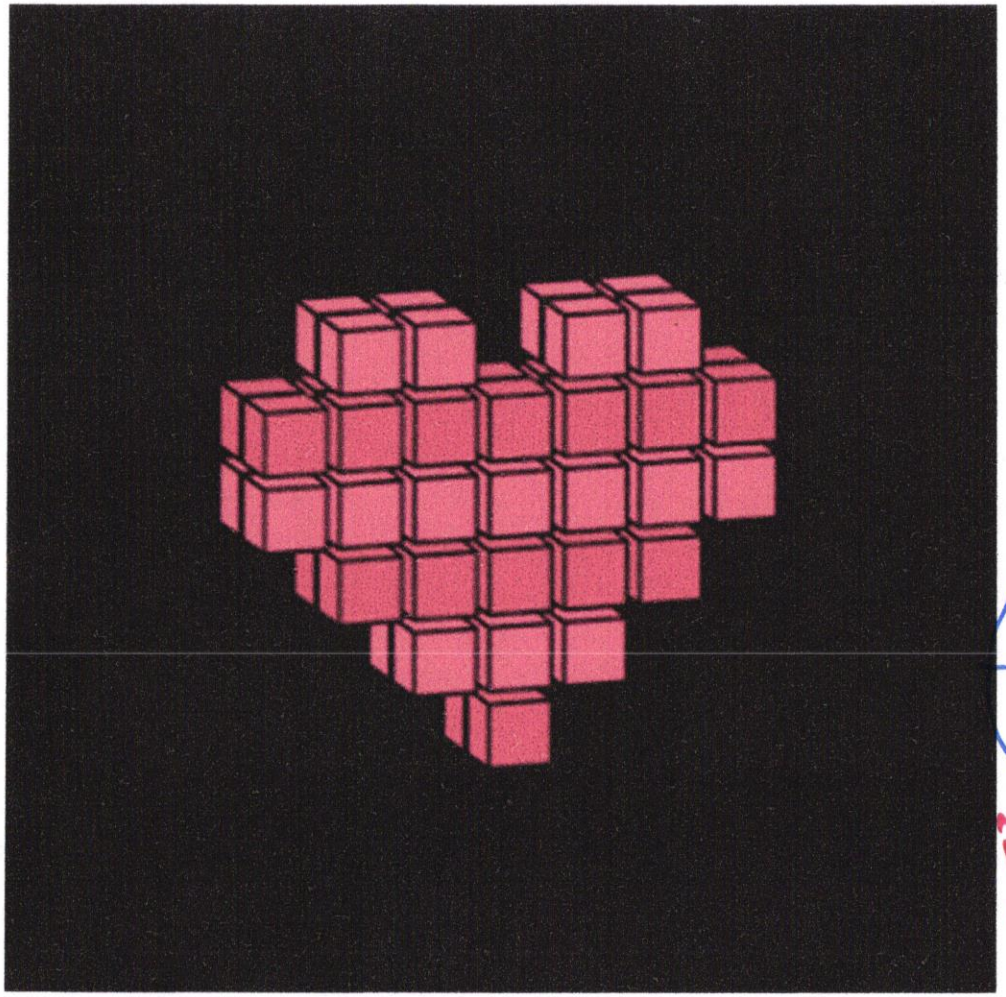
- It is an **interdisciplinary field** that uses scientific methods, processes, algorithms and systems to **extract knowledge** and **insights** from many structural and unstructured **data**.



INSAID

both are EQUALLY important
but data is first among equals.

logic vs data



Orbit of Data Science

It's the Data.

880 – 50 lakhs

.016% – grocery shop own +

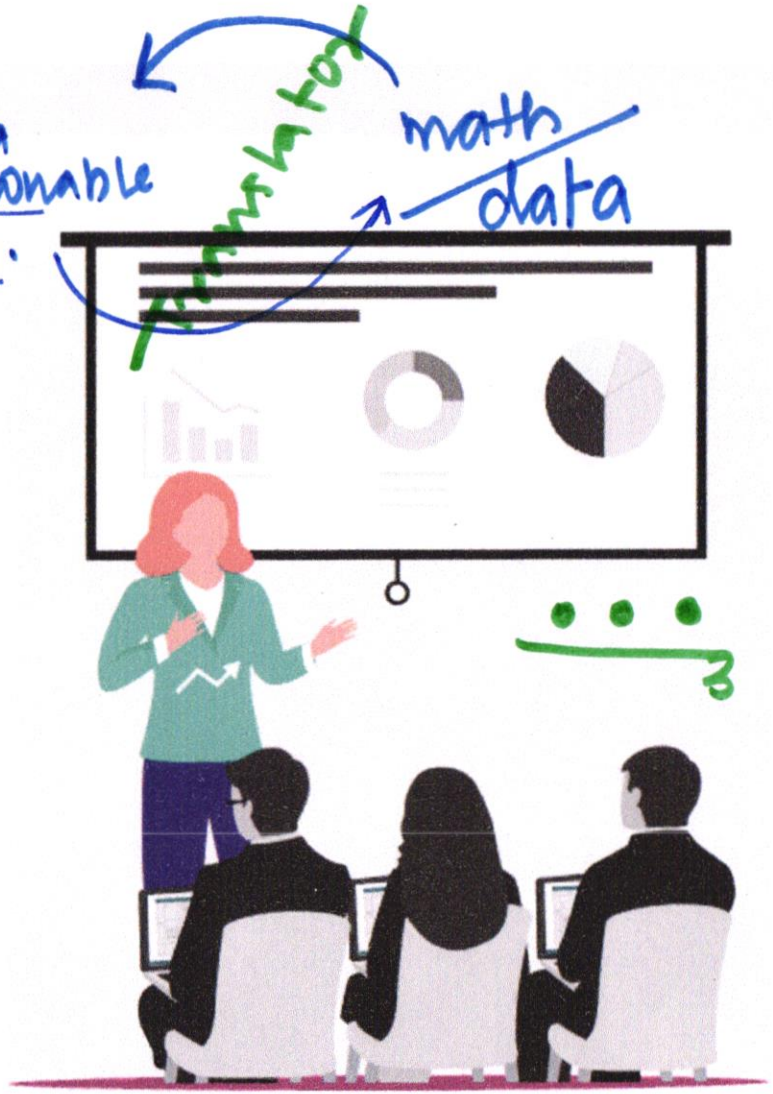
'IIA university'

83% of covid +ve people were grocery shop owners

Please don't go out to buy grocery, adjust with what you have at home.

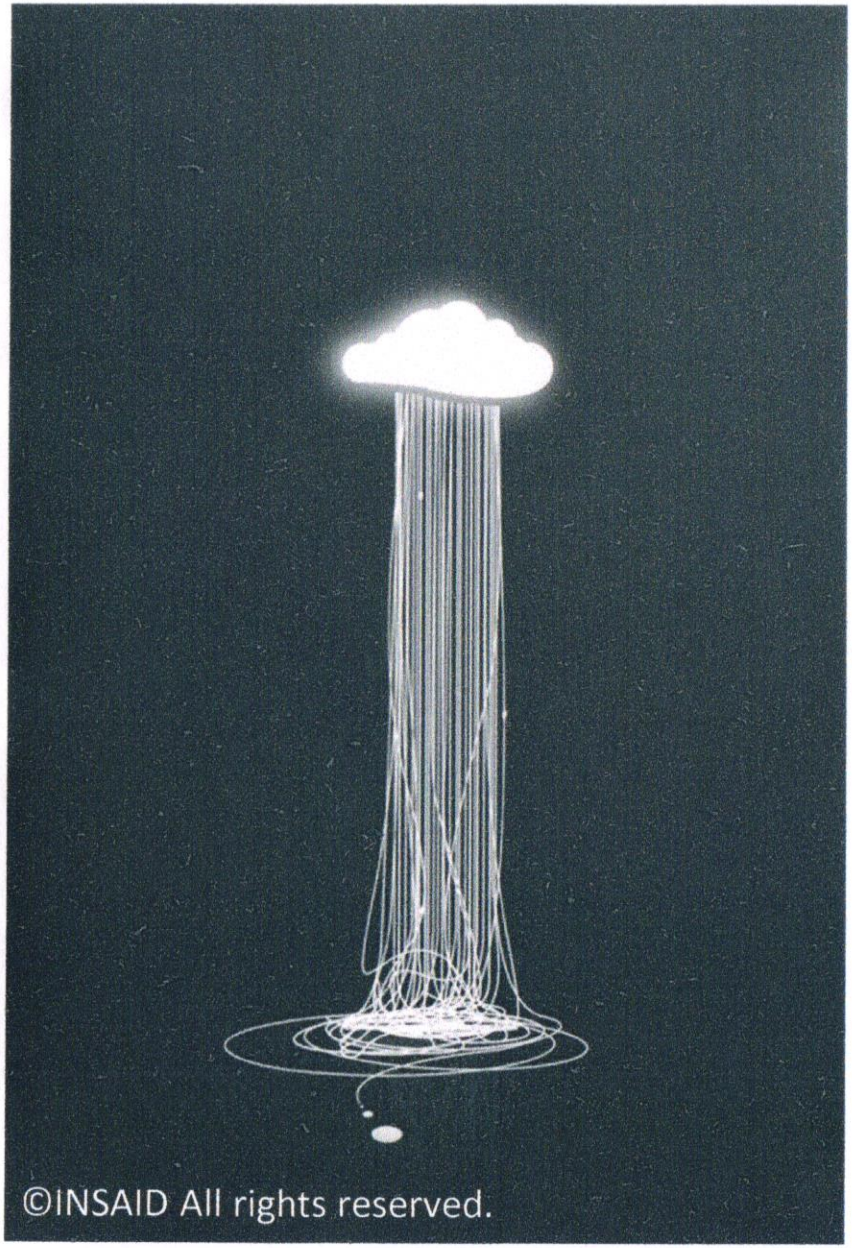
INSAID

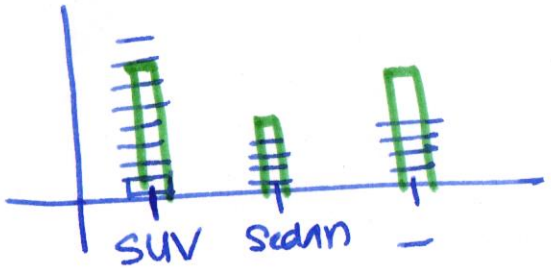
Data Scientists



- A **professional** responsible for collecting, manipulate, analyzing and interpreting data.
- **Main Role:** Create actionable plans for companies and other organizations .

INSAID





Traits of Data Scientists

Probabilistic

- ~~Statistical~~ Thinking
- Technical Acumen – math, python, domain,
- Multimodal Communication Skills
- Curiosity! about everything, Everything is one subject,
- Creativity
- Grit – self motivated

INSAID

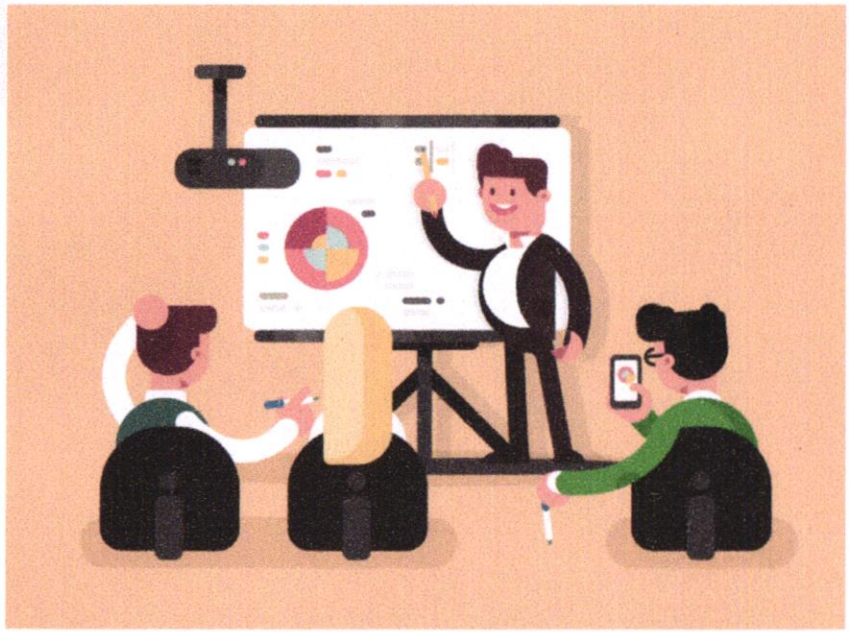
Statistical Thinking

- Can **work** with **any** given **data**.
- **Implement** a **toolbox** full of **algorithms** to make **predictions** and recommendations.
- **Ask questions** from the client on **sensing "fishy"** behavior in the **results**.

INSAID

Technical Acumen

- **Hacker's Spirit:** Produce tools, pipelines, packages, features, dashboards and more on demand.
- **Technical Flexibility:** Requires up to date.
- **Team Collaboration:** Work as a team and leads people to move at the speed of demand.

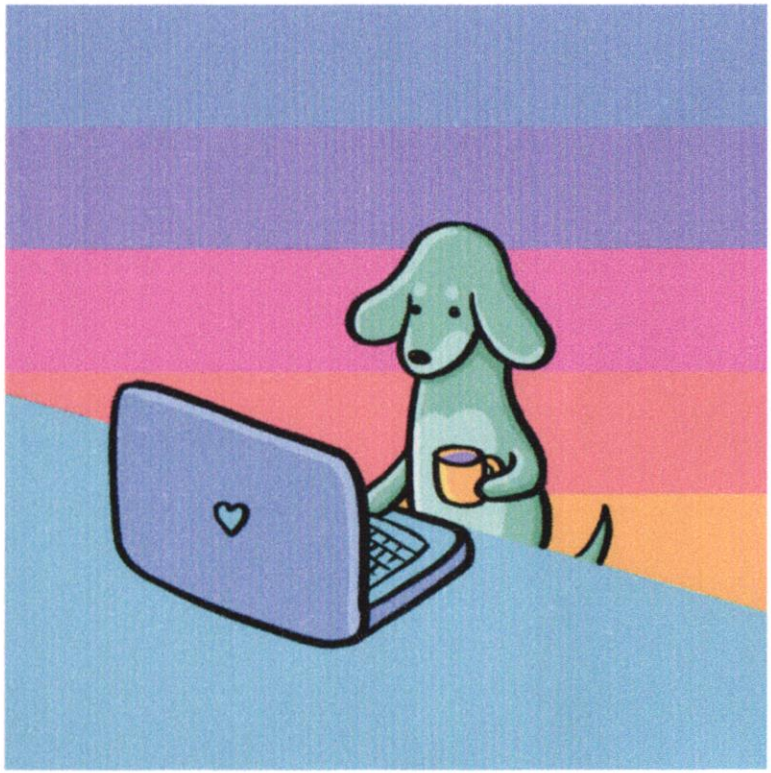


Multimodal Communication Skills

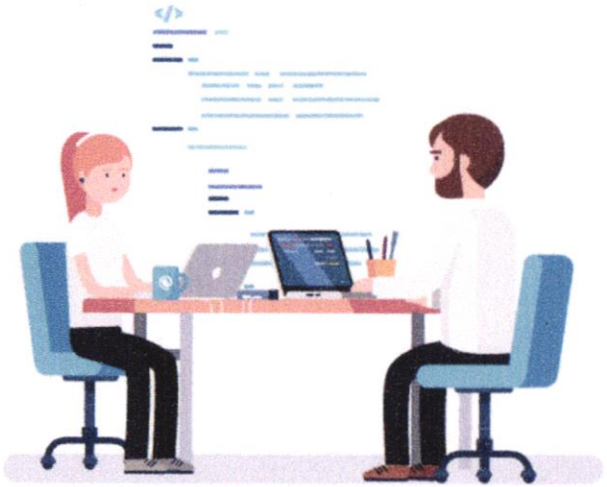
- **Study in context** and **translate** a **problem** and its **solution** to parties with varying background.
- Use **common ground**, metaphor, skillful listening and most import storytelling.

Curiosity!

- An invaluable soft **skill** of asking **"why"** and "**how**".
- Could be triggered in following cases:
  - **Asking** for access to more data,
  - **Interview** users, or
  - **Trying something new** in the next iteration.



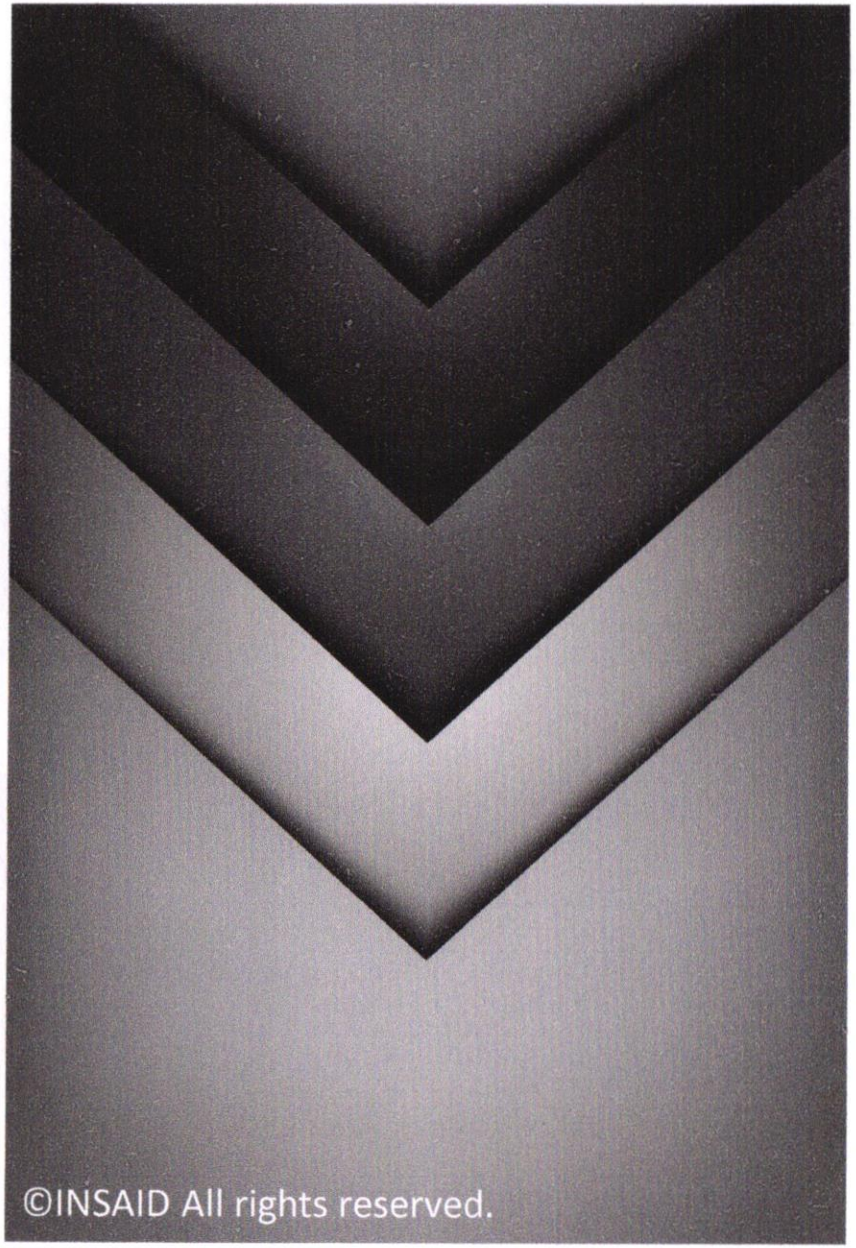
Creativity



- Provides the **fuel** for skillful **communication**.
- **Not** a **hard sell**, requires **dedication** & **efforts**.
- Have a **peculiar relationship** with the word **"no."**
- **Example:** "no," followed quickly by "wait, hold on... let me think," just might be a great creative thinker.

Grit

- An **inner drive** that **pulls** us **over** and **keep** us **walking** over **obstacles** and roadblocks.
- **Challenges** could **drive** a reasonable **person** to an unplanned **leave** of absence.
- Get the **taste** for **tackling** the **new** and the **unknown**.

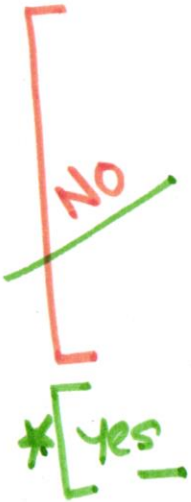


Deeper Discussions

- Data Science: The Sexiest Job Of The 21st Century
- Reasons: Data Scientists Gets Paid A Lot
- Data Science: Art Vs Science
- Data Science Hype
- Data Science Hype Effects
- Data Science Reality

INSAID

Data Science: The Sexiest Job Of The 21st Century



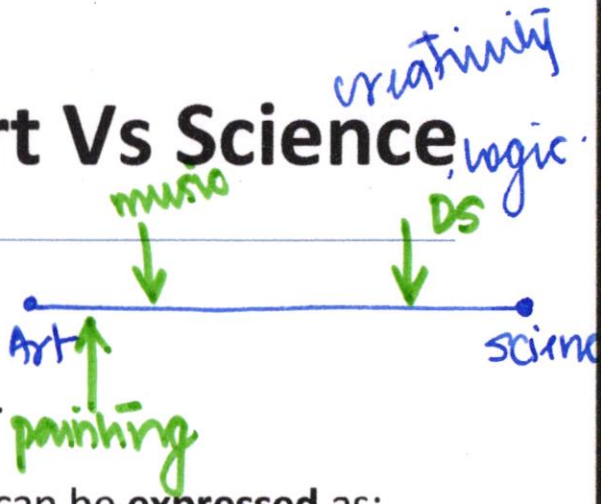
- According to Harvard Business Review, **Thomas H. Davenport** and **D.J. Patil** stated Data Science as the sexiest job of the 21st Century.
- **Further Information:** Present in the link attached to the Harvard Business Review.

INSAID

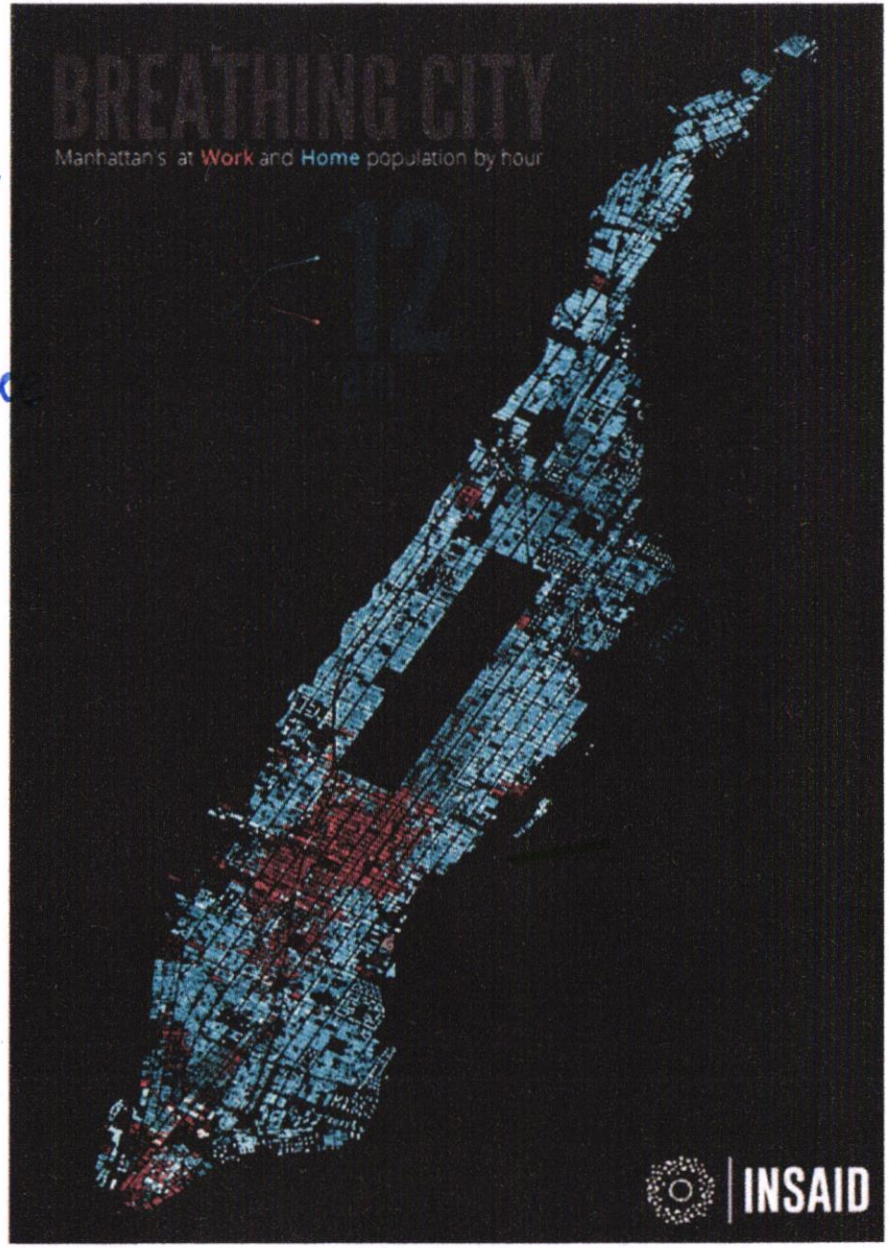
Reasons: Data Scientists Gets Paid A Lot

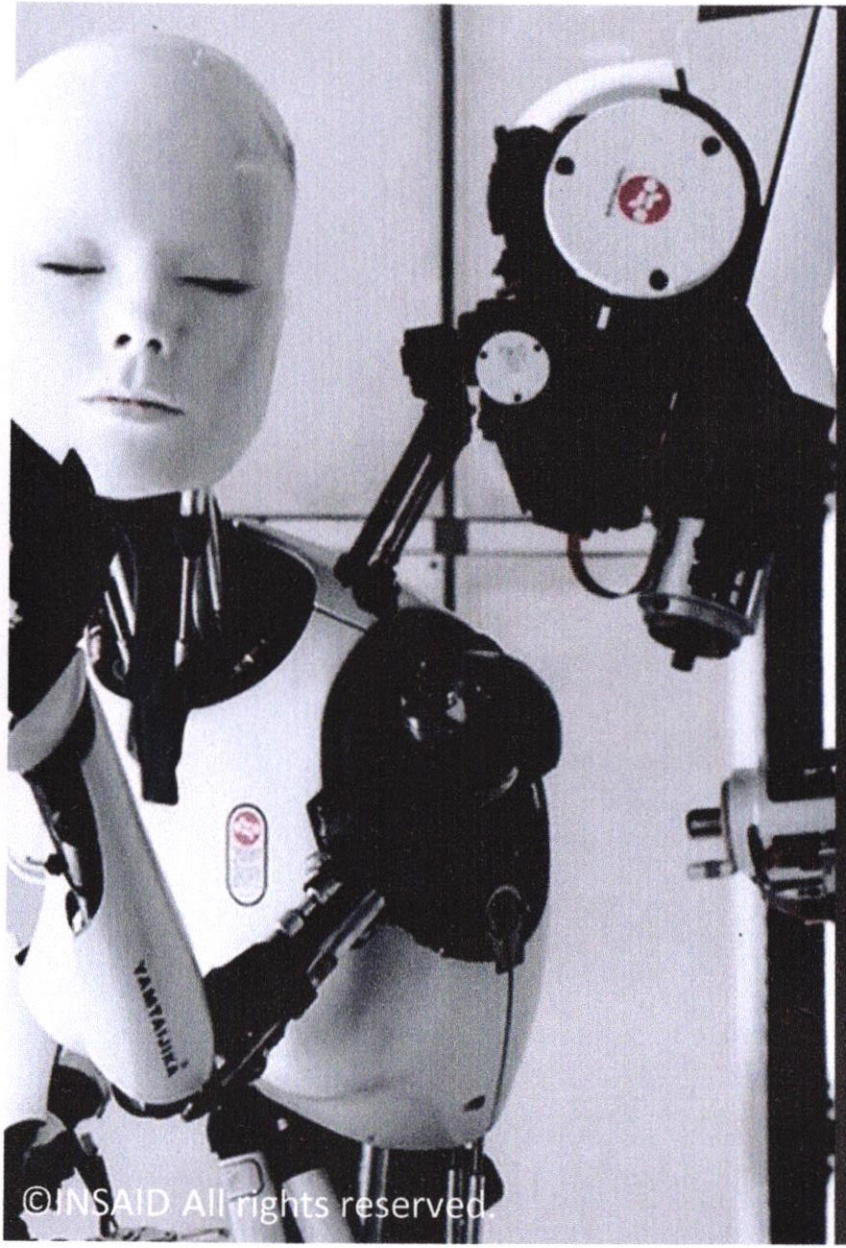
- Severe **shortage** of **talent** in the market.
- **Organizations face** enormous data **challenges**.
- **Not restricted** to **Tech Giants** anymore.
- **Directly impact** the **business** value.

Data Science: Art Vs Science



- It is **more art** and **less science**.
- **Creativity** using Data Science can be **expressed** as:
  - **Applying techniques** in unconventional settings.
  - **Combining techniques** in unconventional ways.
  - **Developing** new and unusual hypothesis.
- **Work source:** [Darkhorse Analytics]





Data Science Hype

- **Throw** any **task** to the machines and you will **get results**.
- **Science** can **increase** the **expectancy** of human life to 100s years and so on. **Example:** Cyborgs.
- Algorithm X (Solution) works best in all types of scenarios.
- Machines can **start** a **different** new **work** which it has **never seen before**.

- Auto ML will replace (Human) Data Scientist
- Artificial Intelligence
  calculation done on basis of human

INSAID

*In foreseeable future

long term game

Data Science Hype Effects

- **Raise** in **expectations** to the sci-fi levels movies.
- **Organizations** that **buy** into the **hype stops investing** any **further** into Data Science, once reality sets in.

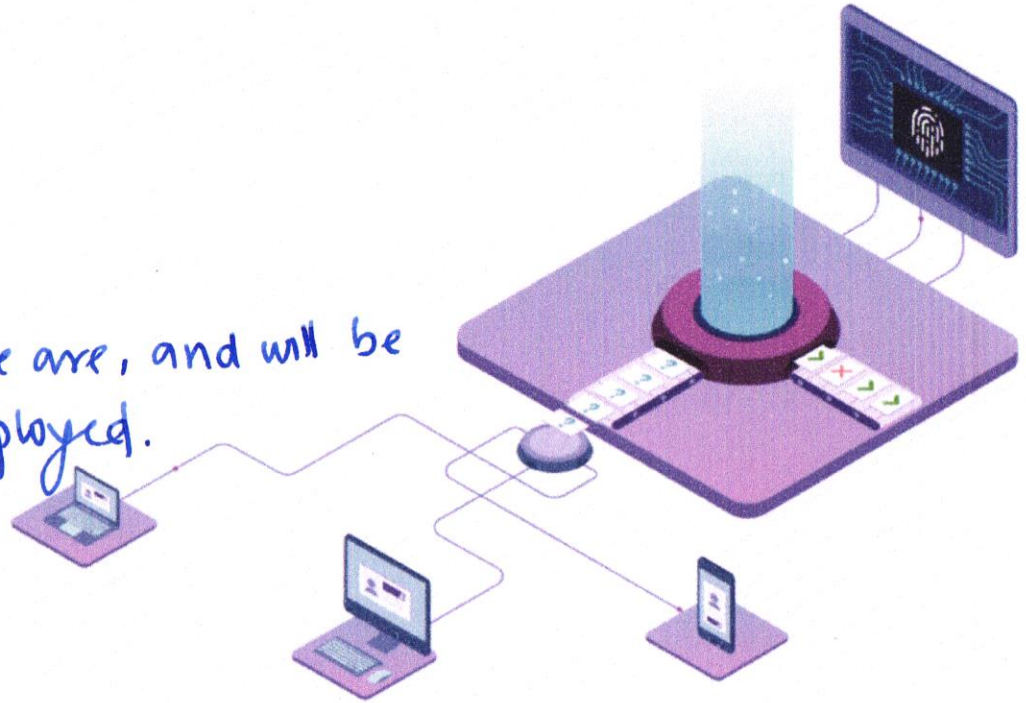
FOMO
fear of missing out

INSAID

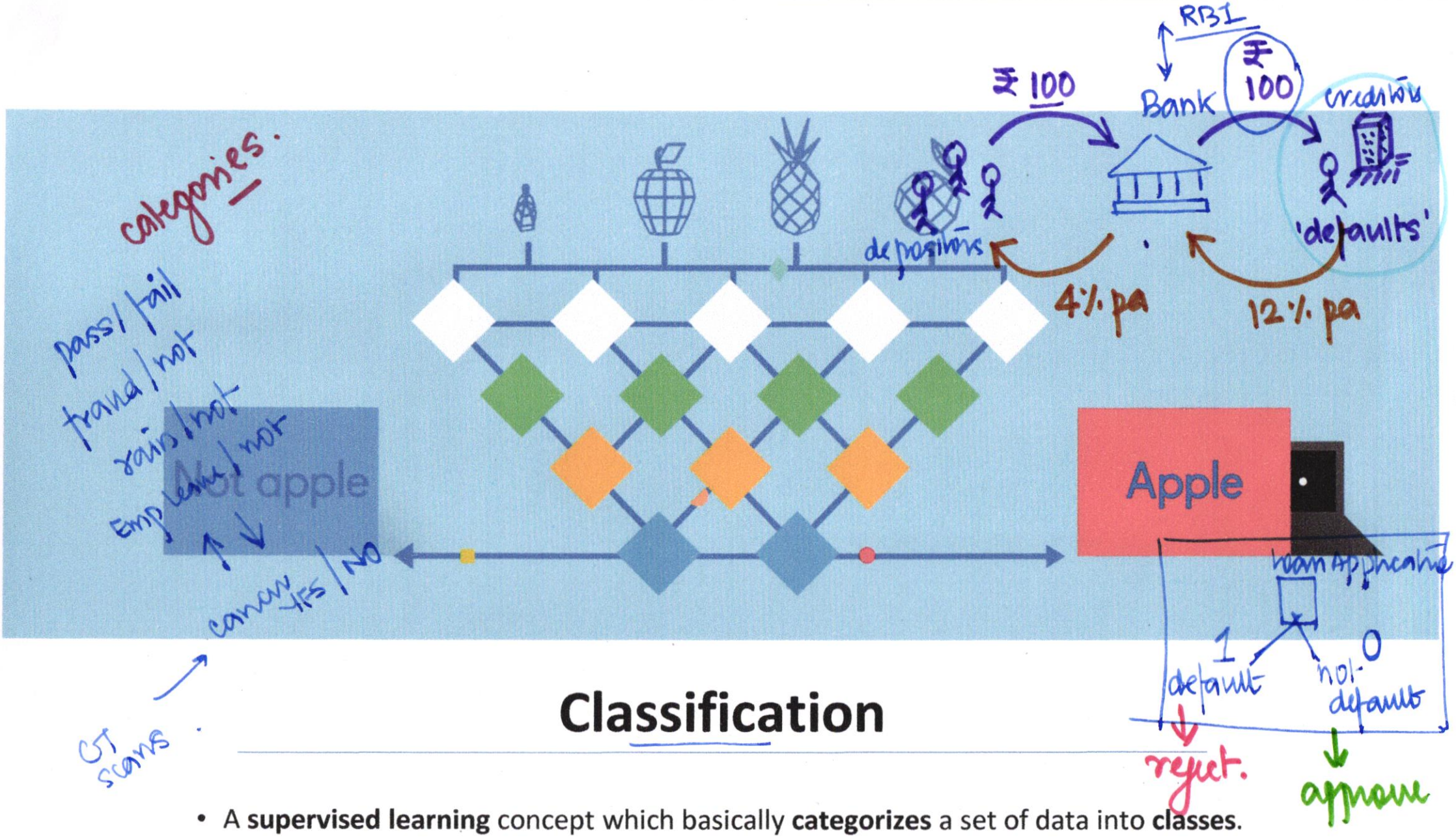
Data Science Reality

- **Machines can't learn** any **arbitrary new task** on throw, as a human can.
- **Science** is **not enough advanced** yet.
- There's **no** such algorithm (**solution**) that **work best** in **all** possible **scenarios**.
- Machines **can't generalize** it's **intelligence** and start a completely new work.

Why we are, and will be employed.



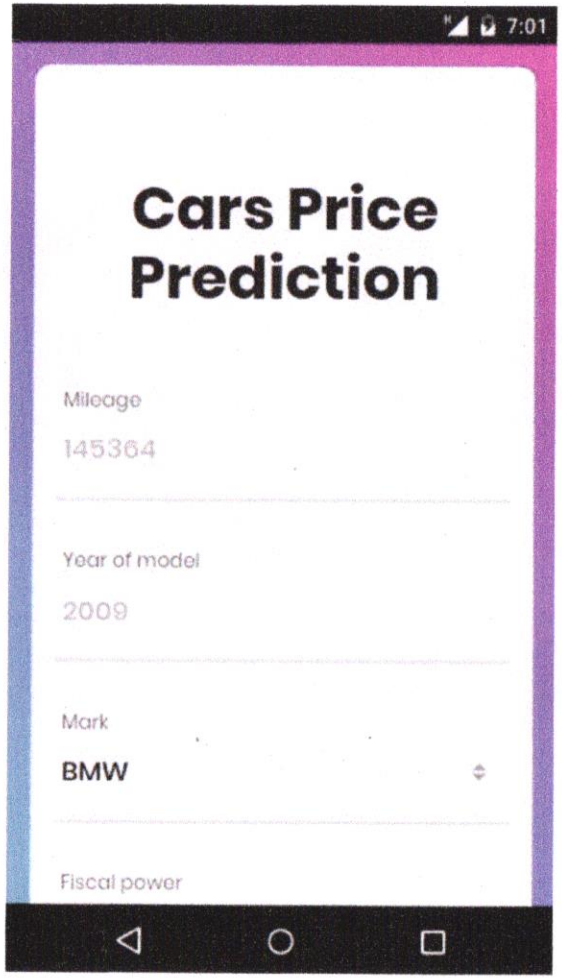
INSAID



Classification

- A **supervised learning** concept which basically **categorizes** a set of data into **classes**.

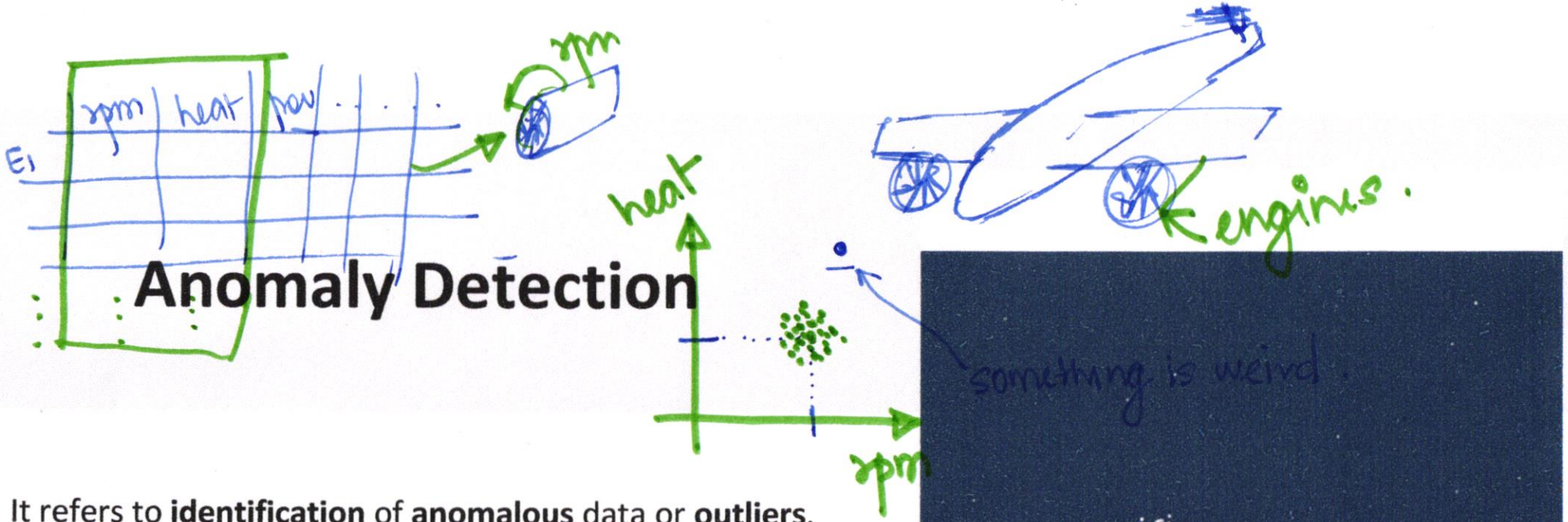
INSAID



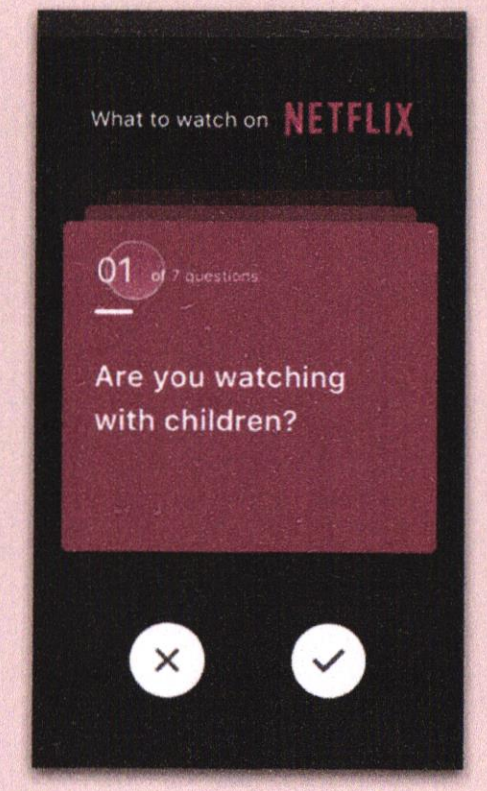
Regression

- Methods that allow us to **predict** a **continuous outcome** variable (y) based on the value of one or multiple predictor variables (x).
- **For Example:** Cars Price Prediction
- **Predictors (x):** Milage, Year of Model, Fuel Type etc.
- **Outcome (y):** Price

Anomaly Detection



- It refers to **identification** of **anomalous** data or **outliers**.
- **Observations** that **differs** majorly **from** the **rest** of the data.
- **For example:** In solar system, only earth has Wi-Fi. So earth is an outlier.



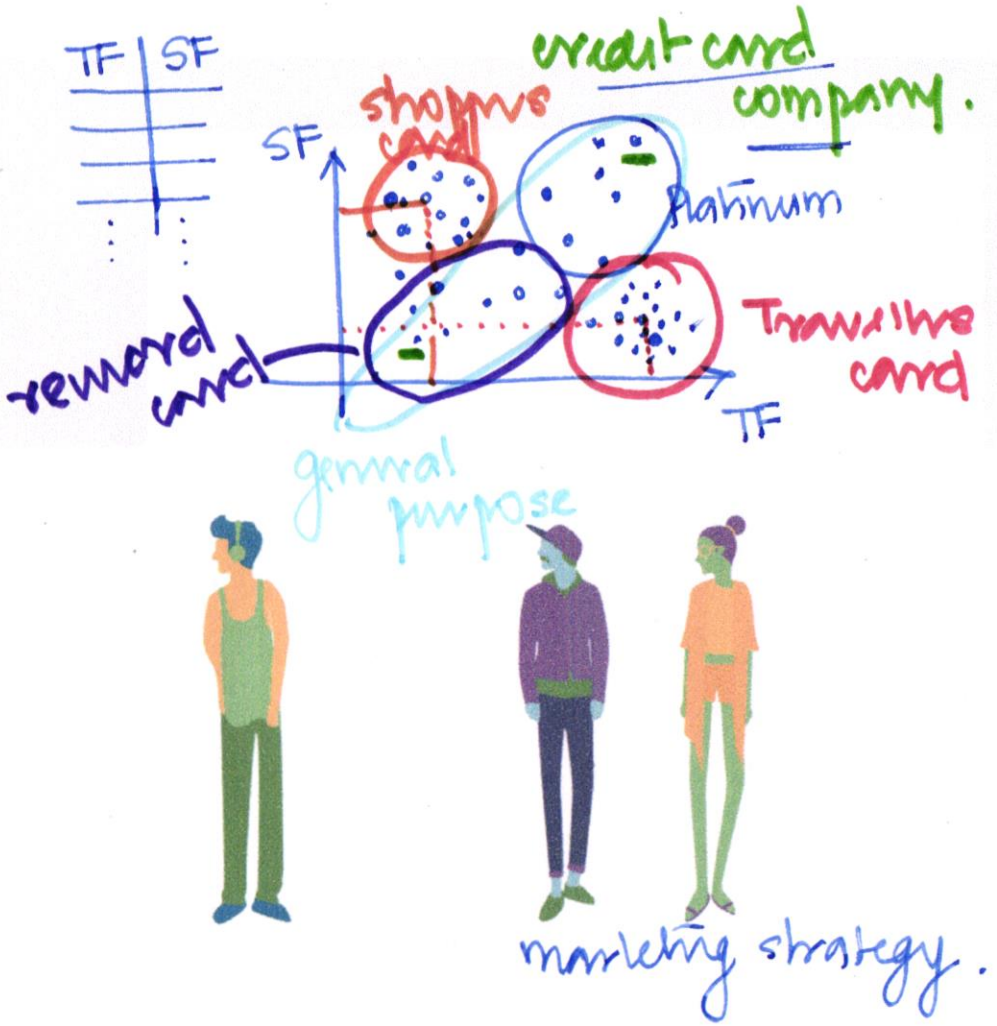
Recommendations

big bazar example.

- **Offering** of **"relevant" suggestions** to users based on product or service.
- **Example:** NETFLIX, You Tube, Spotify etc.

Clustering

grouping



- A technique that **involves** the **grouping** of data points have **similar properties**.
- **Example:** Customer Segmentation etc.

INSAID

Reinforcement

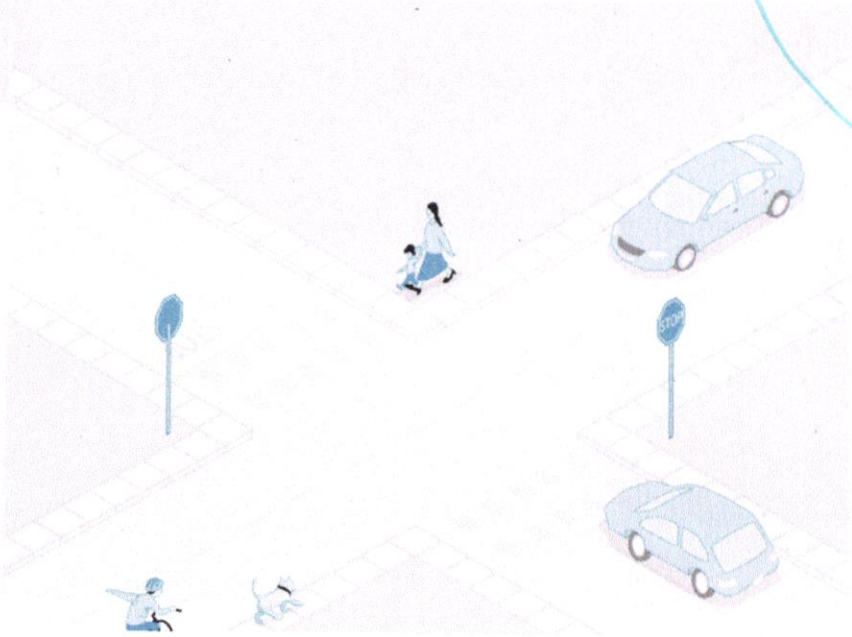
Tea.

milk + water + tea leaves + sugar + boil

20 ml ↑ 100 ml ↓ 30g ↓ 5g ↑ 155 ↑↑

TEA

mom's tea

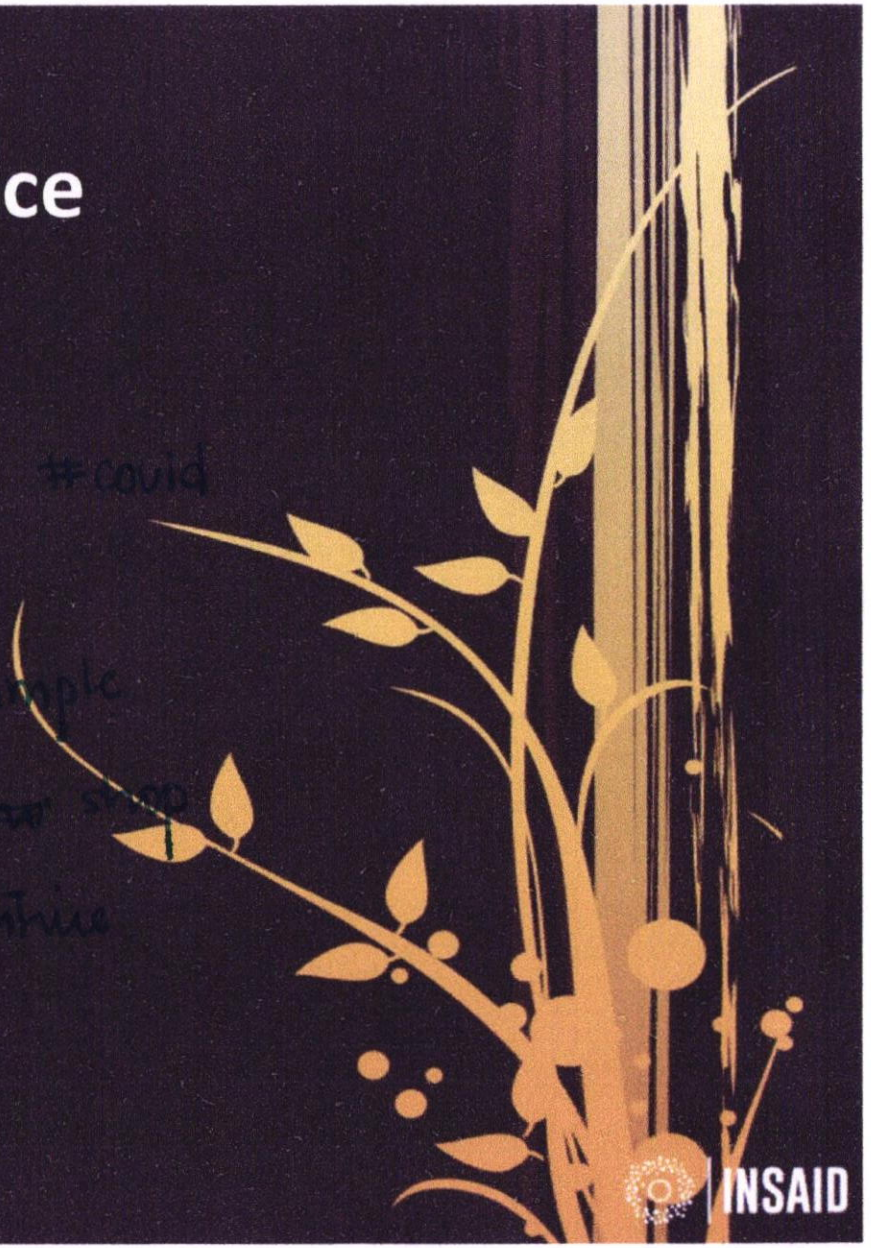


- **Software agents** ought to **take actions** in an **environment** in order to **maximize** the notion of cumulative **reward**.
- **Example:** Self Driving Cars etc.

AlphaGo , ibm watson

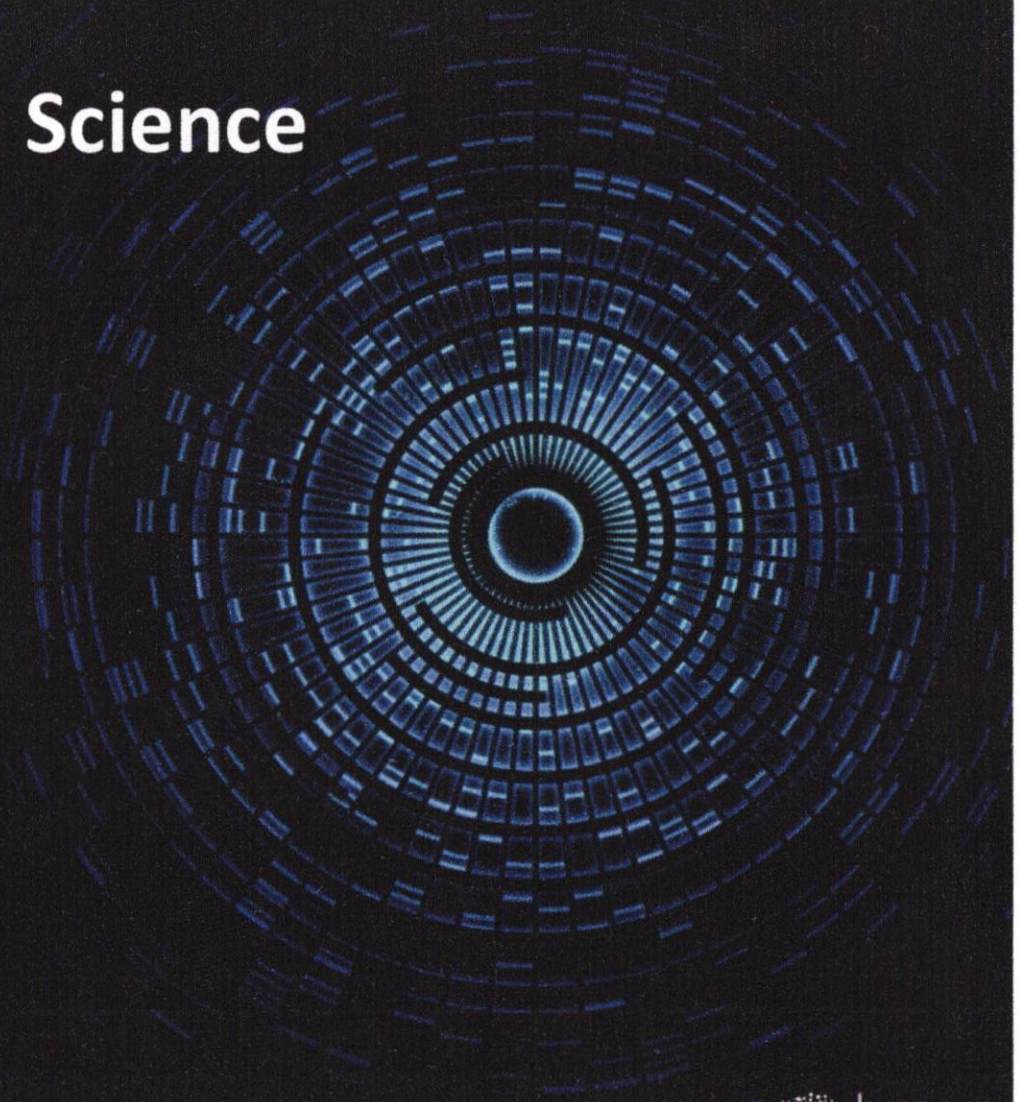
Problems Solved by Data Science

- Classification
- Regression
- Anomaly Detection
- Recommendations
- Clustering
- Reinforcement

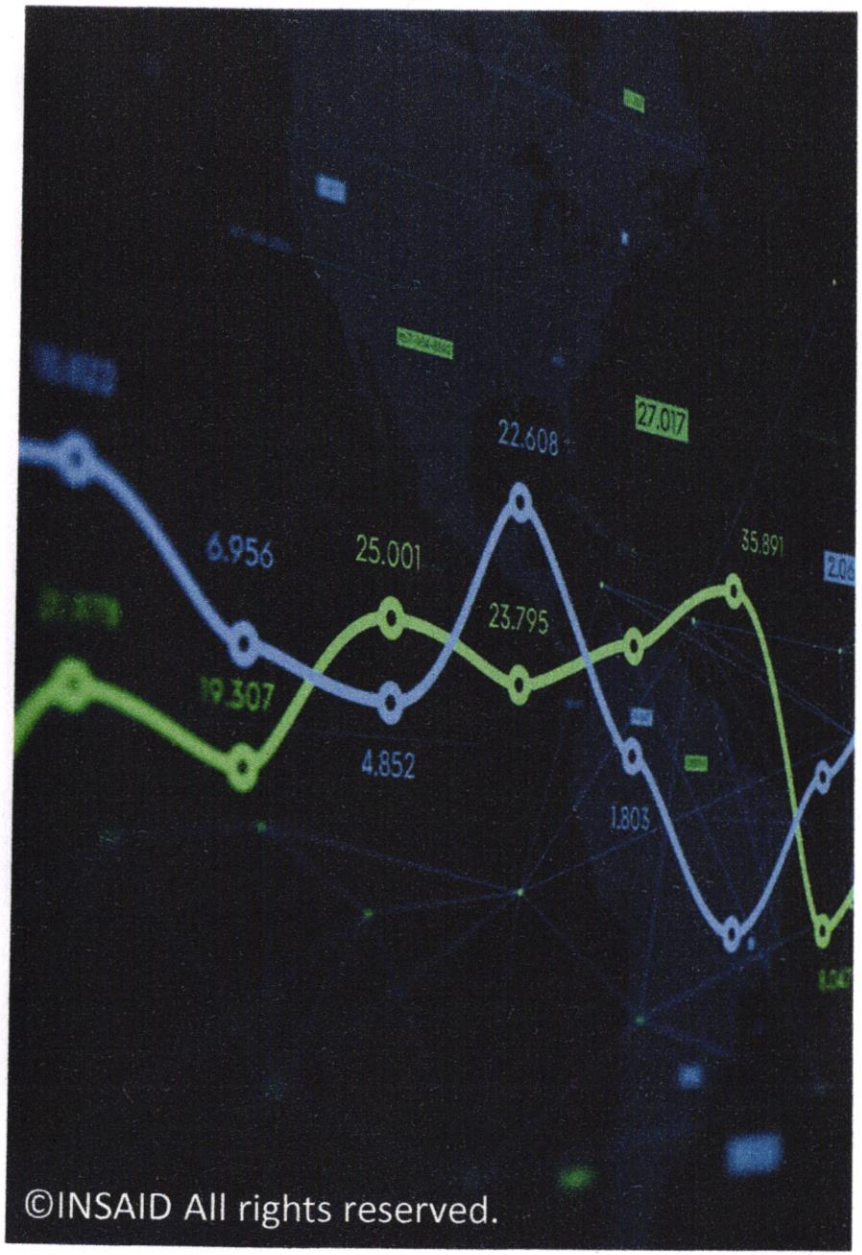


Industries Involved in Data Science

- Banking & Finance
- Healthcare
- Retail & ECommerce
- Telecommunications
- Energy & Utilities
- Manufacturing
- & Many More...



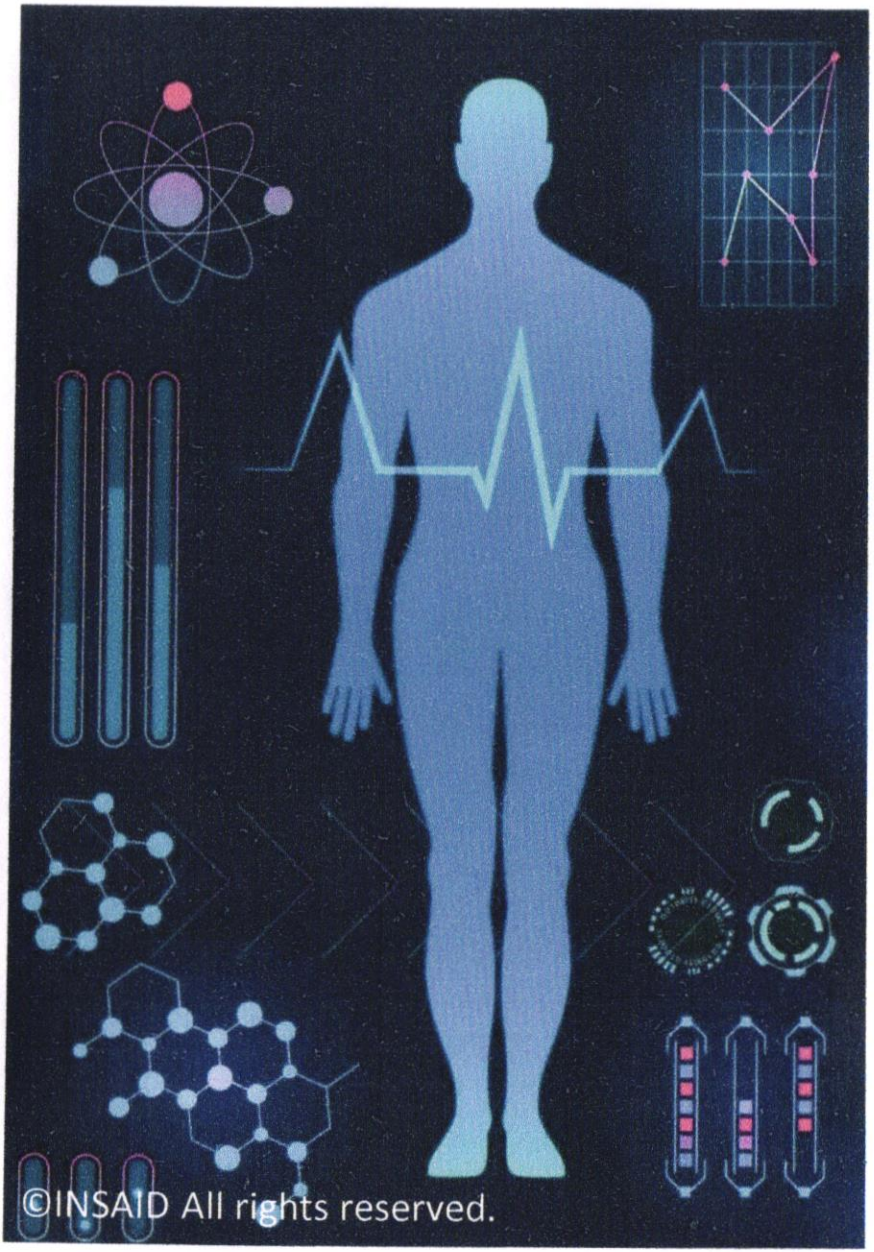
INSAID



Banking & Finance

- Fraud Detection and Prevention — Anomaly or classification
- Customer Segmentation — clustering
- Risk Management
- ~~Finance Factoring~~ — SLR, CRR. statistically
- Credit Default Risk Assessment — classification
- & Many More... with $ Amount default with. — Regression.

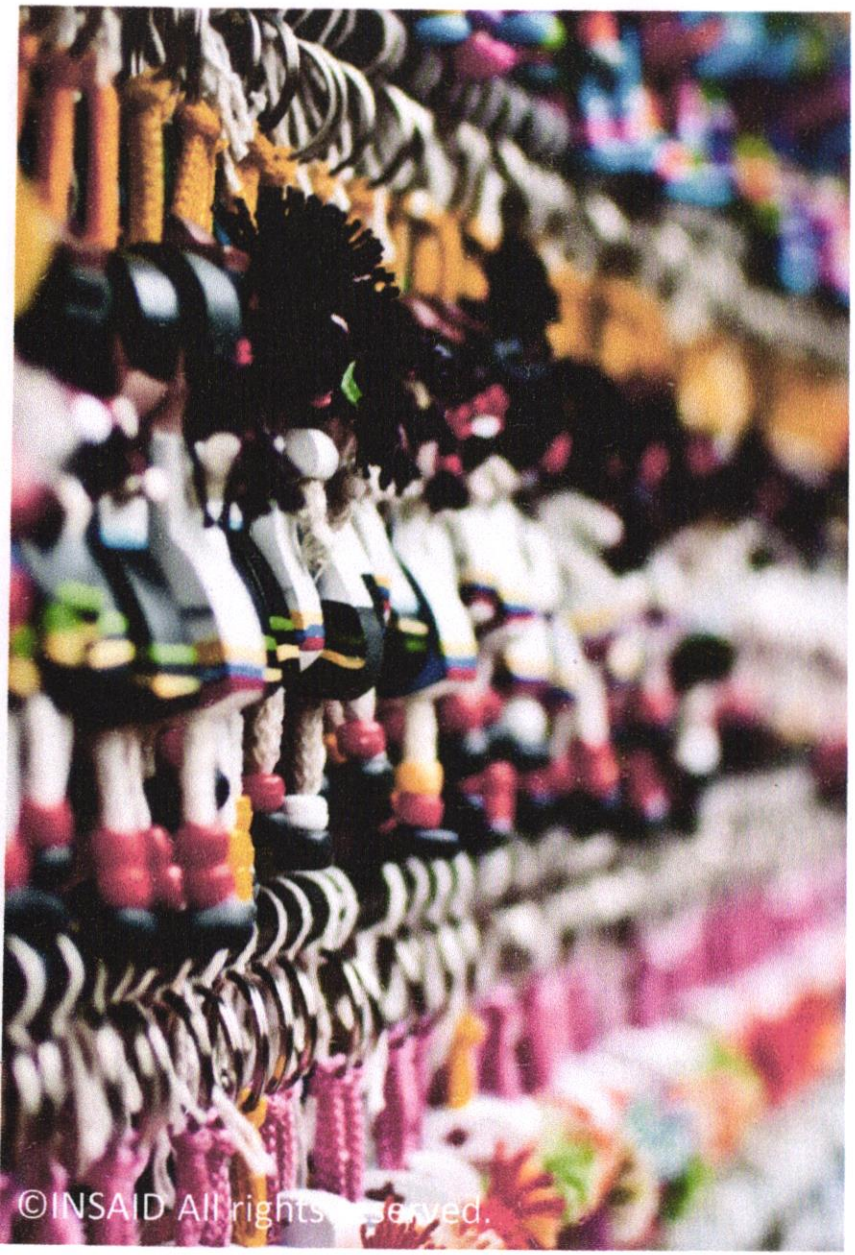
INSAID



Healthcare

- Detection of Disease — classification
- Genetic Analysis  DNA → Age. Regression.
- Post-Surgery Survival Analysis 1/0 classification
- Epidemic Forecasting ~~and Control~~ Regression.
- Eye Health Prediction — ~~1/0~~ thickness
- & Many More...

INSAID



Retail & ECommerce

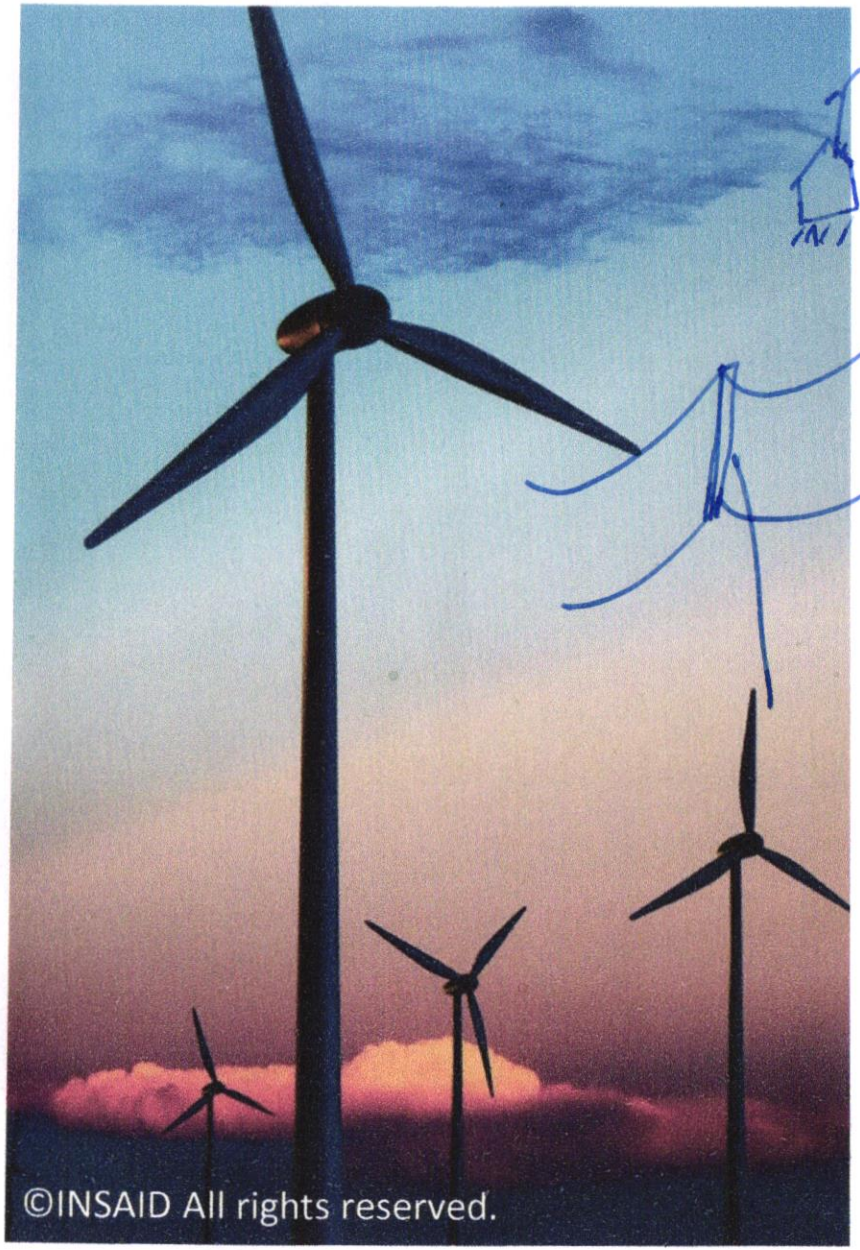
- Market Basket Analysis — Recommendation
- Supply Chain Profit Prediction — Reg.
- Sales Forecasting — Reg
- Prediction of Cart Abandonment — class 1/0
- Customer Sentiment Analysis — clustering
- & Many More...

INSAID

Telecommunications

- Customer Churn Prediction
- Detection of Network Congestion
- Detection of Customer Delinquency
- Identification of Network Quality
- Recommendation of New Services
- & Many More...

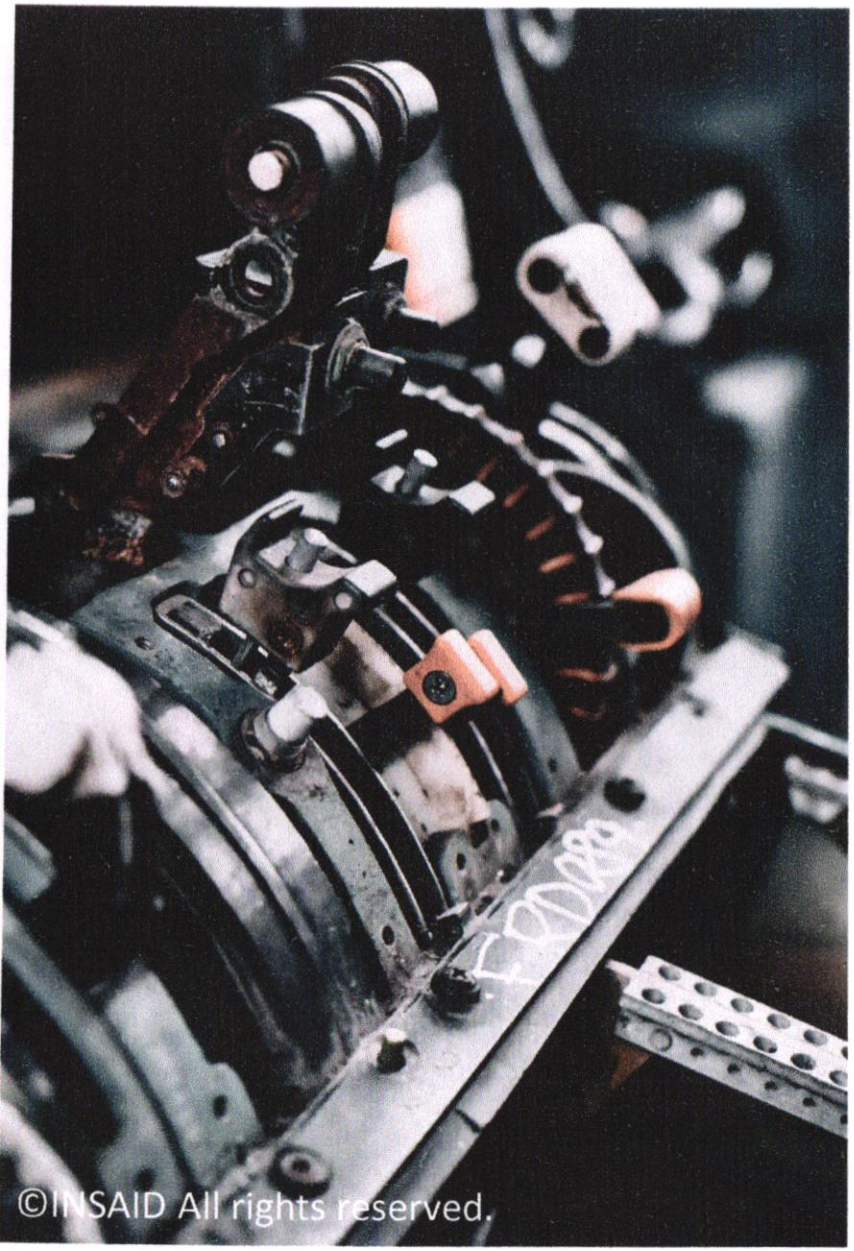
INSAID



Energy & Utilities

- Preventive Equipment Maintenance
- Smart Grid security and theft detection
- Real time Customer Billing
- Outage detection and prediction
- Failure probability modeling
- & Many More...

INSAID



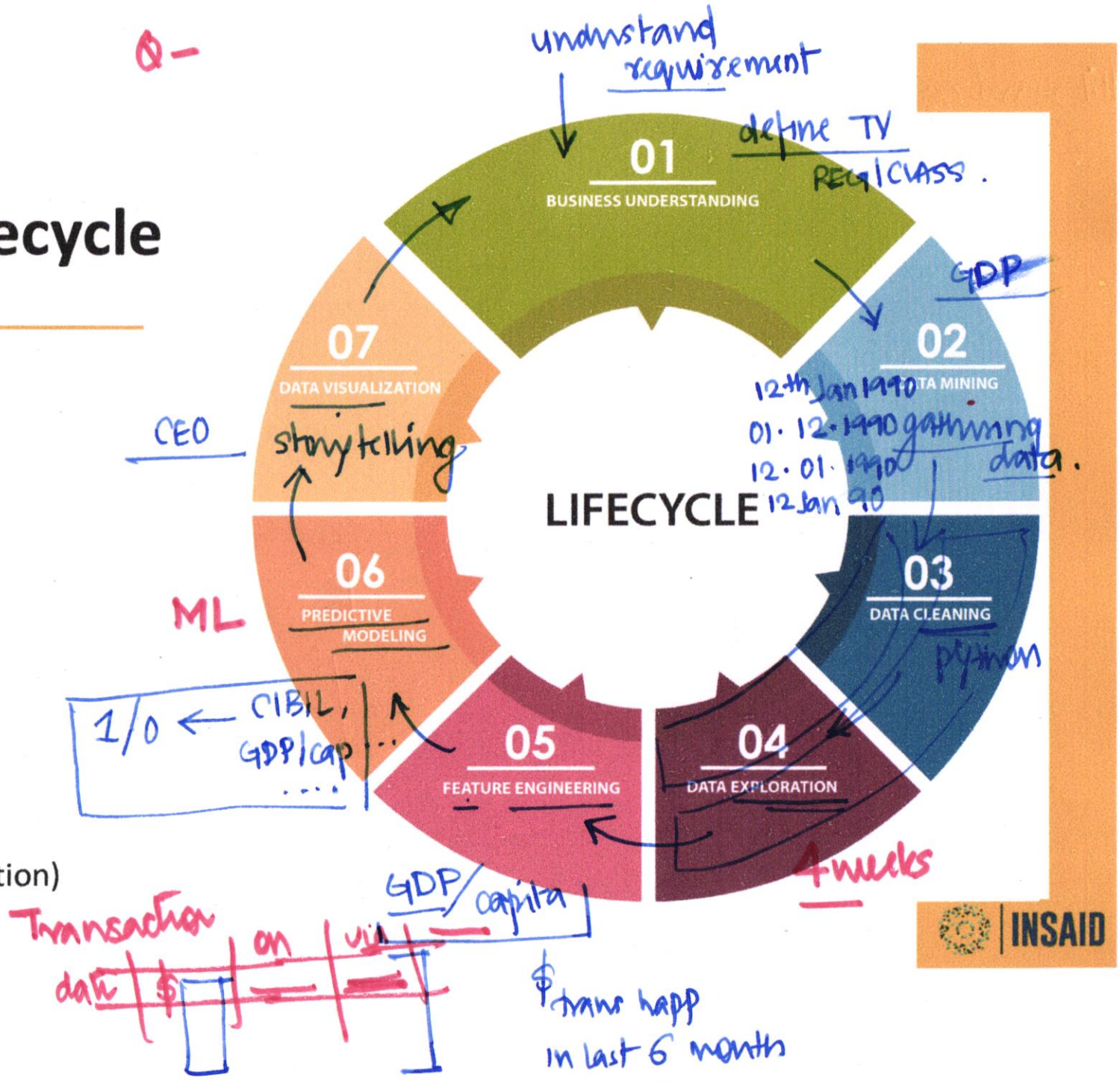
Manufacturing

- Demand Forecasting
- Prescription on Fault Detection
- Quality Assurance
- Supply Chain Risk Analysis
- Warranty Analysis
- & Many More...

INSAID

Data Science Lifecycle

- Business Understanding
- Data Mining
- Data Cleaning
- Data Exploration
- Feature Engineering
- Predictive Modeling
- Data Visualization (Presentation)



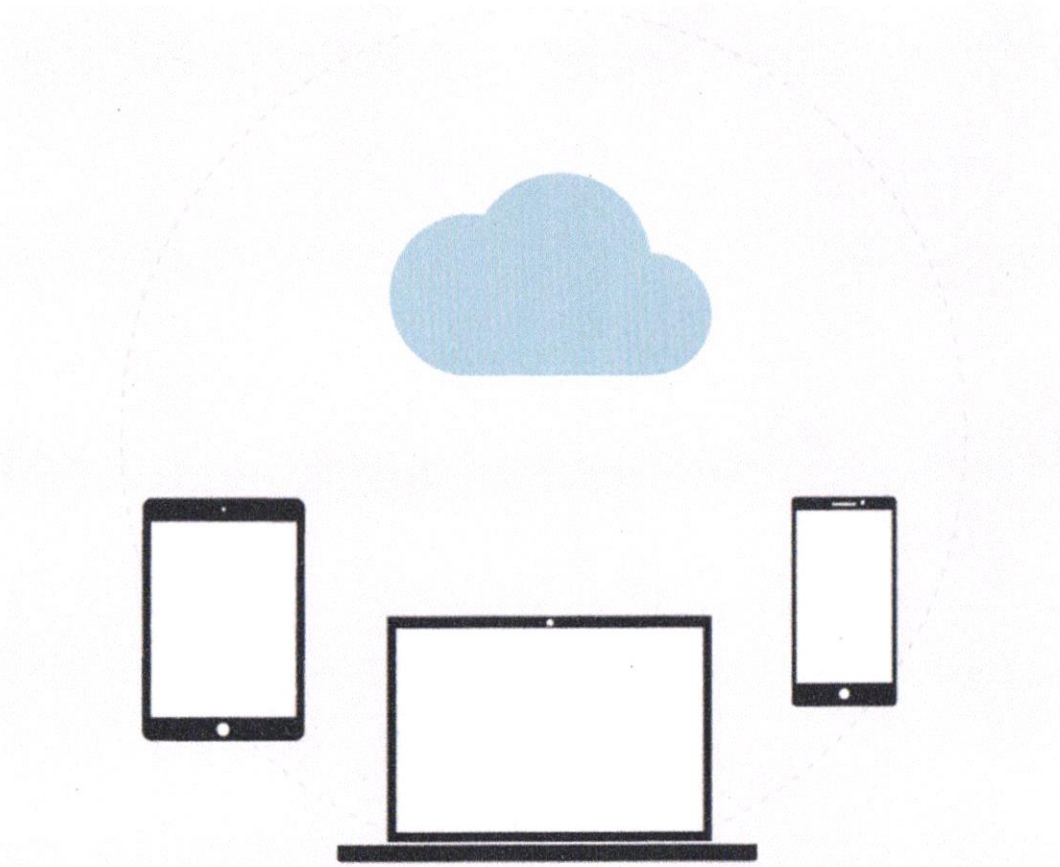
1. Business Understanding

- Ask **relevant questions** regarding problem.
- **Define objectives** that needs to be tackled.
- **Objectives** should be very **specific** and **clear**.

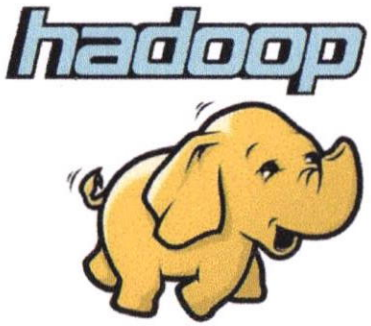
INSAID

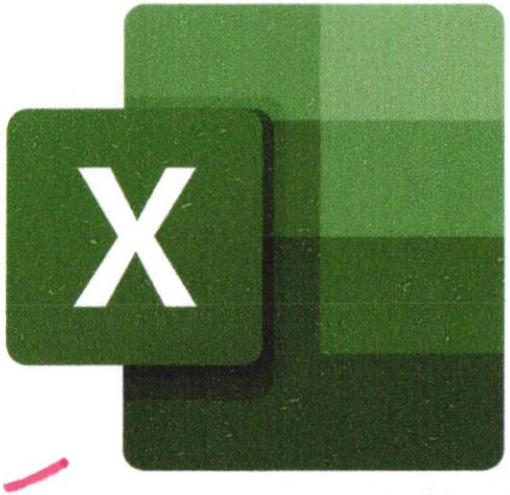
2/ Data Mining

- **Gather** and **scrape** the **data** necessary for the project.
- **Common database** used are NoSQL, SQL, Warehouse, Hadoop etc.
- Our **current scope** will be limited to **CSV** or **XLSX** files (Excel files).



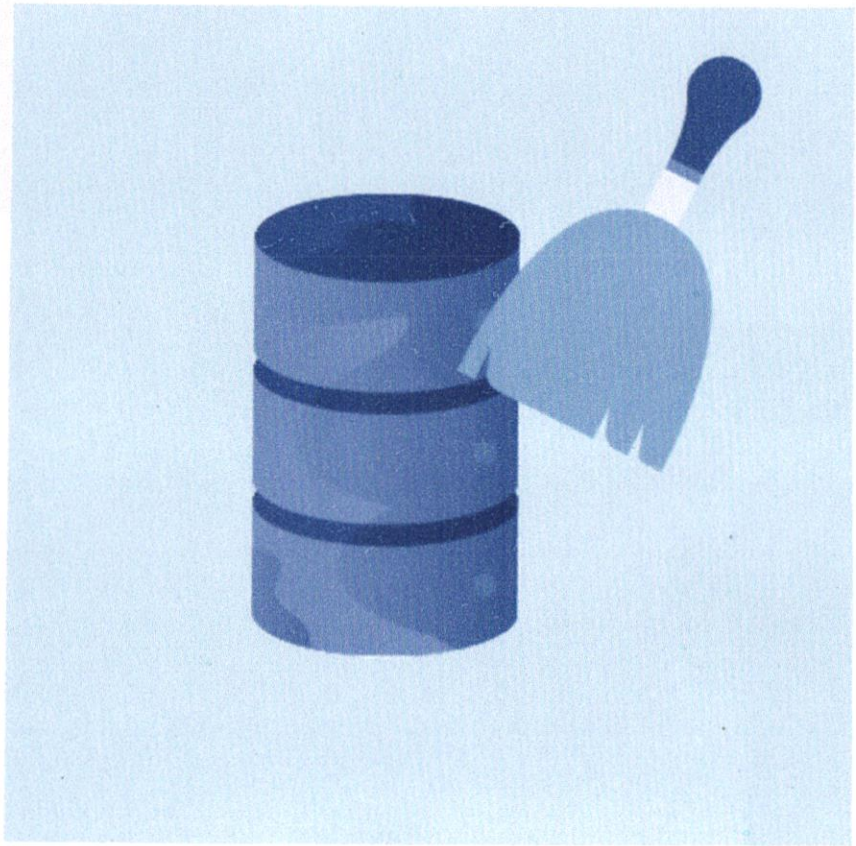
INSAID





Data Source Examples

INSAID



3 Data Cleaning

- Consist of handling multiple operation such as:
  - Missing data
  - Duplicity
  - Inconsistent Types
  - Irrelevant data
  - Outliers
  - Formatting

INSAID

Sender Name	Receiver Name	Sender Old Balance	Sender New Balance	Receiver Old Balance	Receiver New Balance	IsFraud?
C1231006815	M1979787155	170136	160296	0	0	No
C1666544295	M2044282225	21249	19384		0	No
C1305486145	C553264065		0	0	0	Yes
C840083671	C38997010	181		21182	0	Yes
C2048537720	M1230701703	41554	29885.8	0		No

Missing Data

- Generally, data set contains Nan data.
- In the above diagram red cell explains the null data points.

Sender Name	Receiver Name	Sender Old Balance	Sender New Balance	Receiver Old Balance	Receiver New Balance	IsFraud?
C1231006815	M1979787155	170136	160296	0	0	No
C1666544295	M2044282225	21249	19384	0	0	No
C1305486145	C553264065	181	0	0	0	Yes
C840083671	C38997010	181	0	0	0	Yes
C2048537720	M1230701703	41554	29885.8	0	0	No
C840083671	C38997010	181	0	0	0	Yes
C2048537720	M1230701703	41554	29885.8	0	0	No

Duplicity

- Duplicate data can be either in rows or columns.
- In the above diagram orange cell explains the duplicity.

INSAID

Sender Name	Receiver Name	Sender Old Balance	Sender New Balance	Receiver Old Balance	Receiver New Balance	IsFraud?
C1231006815	M1979787155	170136	160296	0	"0"	No
C1666544295	M2044282225	21249	19384	0	"0"	No
C1305486145	C553264065	181	0	0	"0"	Yes
C840083671	C38997010	181	0	21182	"0"	Yes
C2048537720	M1230701703	41554	29885.8	0	"0"	No

Inconsistent Types

- There can be inconsistent data types.
- In the above diagram, green marked cells contains string data but in actual they are numeric type.

Sender Name	Receiver Name	Sender Old Balance	Sender New Balance	Receiver Old Balance	Receiver New Balance	IsFraud?
C1231006815	M1979787155	170136	160296	0	0	No
C1666544295	M2044282225	21249	19384	0	0	No
C1305486145	C553264065	181	0	0	0	Yes
C840083671	C38997010	181	0	21182	0	Yes
C2048537720	M1230701703	41554	29885.8	0	0	No

Irrelevant Types

- There can be irrelevant data which has no use at all.
- In the above diagram, purple marked cells contains irrelevant data.

Sender Name	Receiver Name	Sender Old Balance	Sender New Balance	Receiver Old Balance	Receiver New Balance	IsFraud?
C1231006815	M1979787155	170136	160296	0	0	No
C1666544295	M2044282225	21249	19384	0	0	No
C1305486145	C553264065	181	0	0	0	Yes
C840083671	C38997010	181	0	21182	0	Yes
C2048537720	M1230701703	41554	29885.8	0	0	No

Outliers

- There can be outliers in the data.
- In the above diagram, pink marked cells contains outliers.

INSAID

Sender Name	Receiver Name	Sender Old Balance	Sender New Balance	Receiver Old Balance	Receiver New Balance	IsFraud?
C1231006815	M1979787155	170136	160296	0	0	no
C1666544295	M2044282225	21249	19384	0	0	No
C1305486145	C553264065	181	0	0	0	Yes
C840083671	C38997010	181	0	21182	0	yes
C2048537720	M1230701703	41554	29885.8	0	0	No

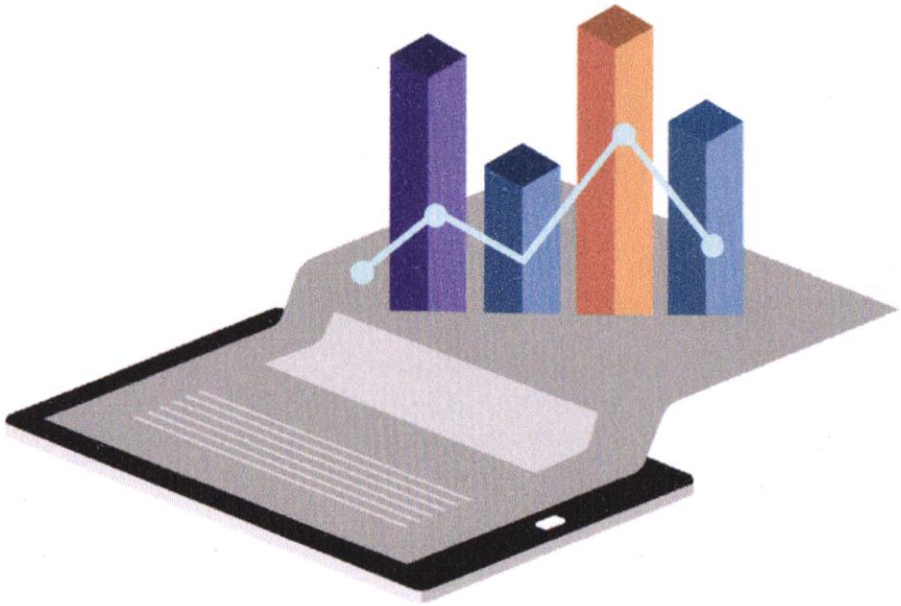
Formatting

- There can be formatting issues with the data.
- In the above diagram, magenta marked cells contains formatting issues **(letter case)**.

Data Exploration

- Analyze data through visuals
- Discover patterns
- Spot Anomalies
- Form Hypothesis
- Check assumption

details later



INSAID

Feature Engineering

- Construct meaningful features.
- Select important features.
- Data splitting and preparation.

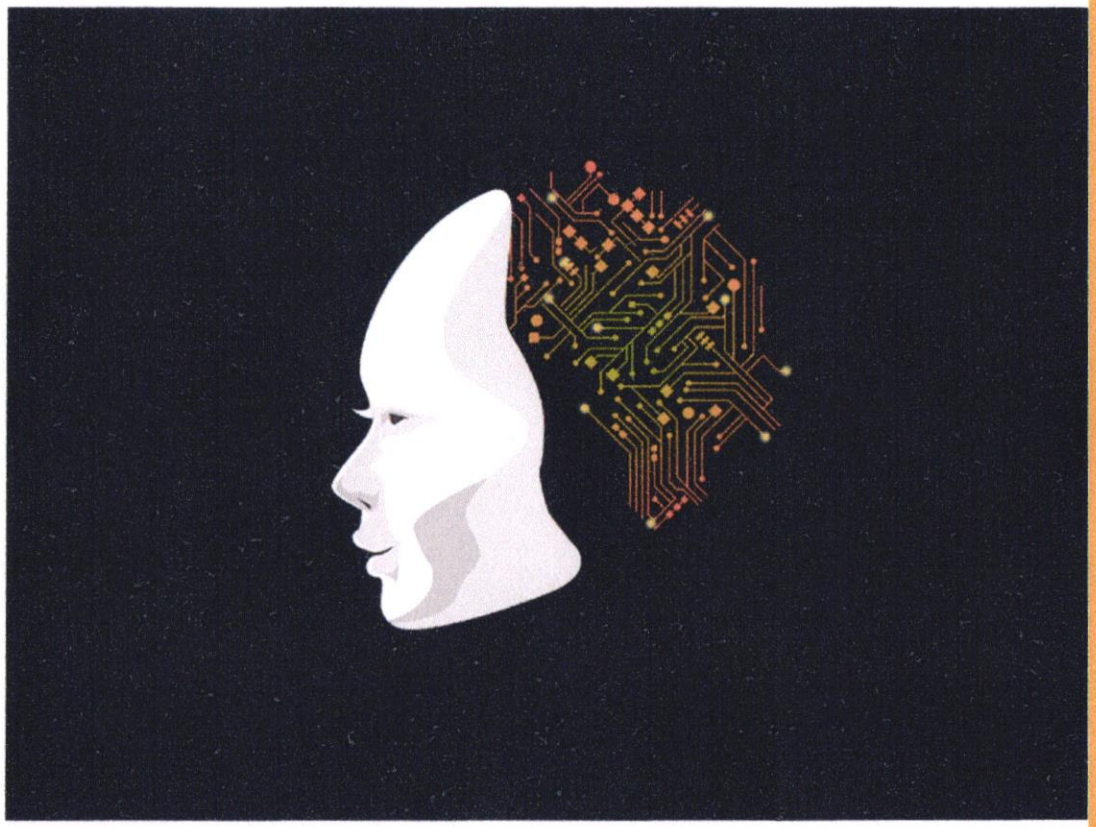
Sender Name	Receiver Name	Sender Old Balance	Sender New Balance	Receiver Old Balance	Receiver New Balance	Error Sender Balance	Error Receiver Balance
C1231006815	M1979787155	170136	160296	0	0	0	9840
C1666544295	M2044282225	21249	19384	0	0	0	1865
C1305486145	C553264065	181	0	0	0	0	181
C840083671	C38997010	181	0	21182	0	21182	0
C2048537720	M1230701703	41554	29885.8	0	0	0	11668.2

Feature Creation

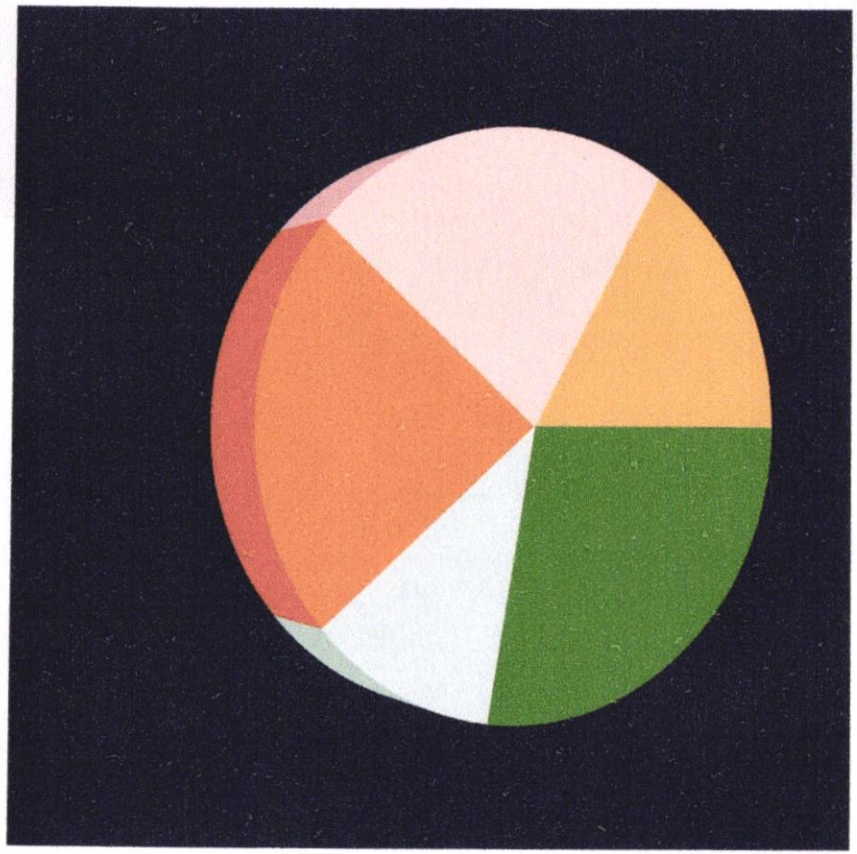
- In the above diagram, light green marked cells have been generated from existing feature cells.
- **Another Example:** Total Salary = Basic Allowance + Dearness Allowance + House Rent Allowance - Tax Deducted at Source

Predictive Modeling

- Train machine learning models.
- Evaluate their performance.
- Fine tune over the key findings.



INSAID

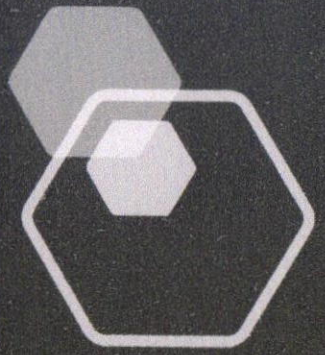


Data Visualization (Presentation)

story telling

- **Explain/Communicate** the **findings** with keys **stakeholders** using **plots** and interactive **visualizations**.
- Take **action** and **deploy** findings in the **real world**.

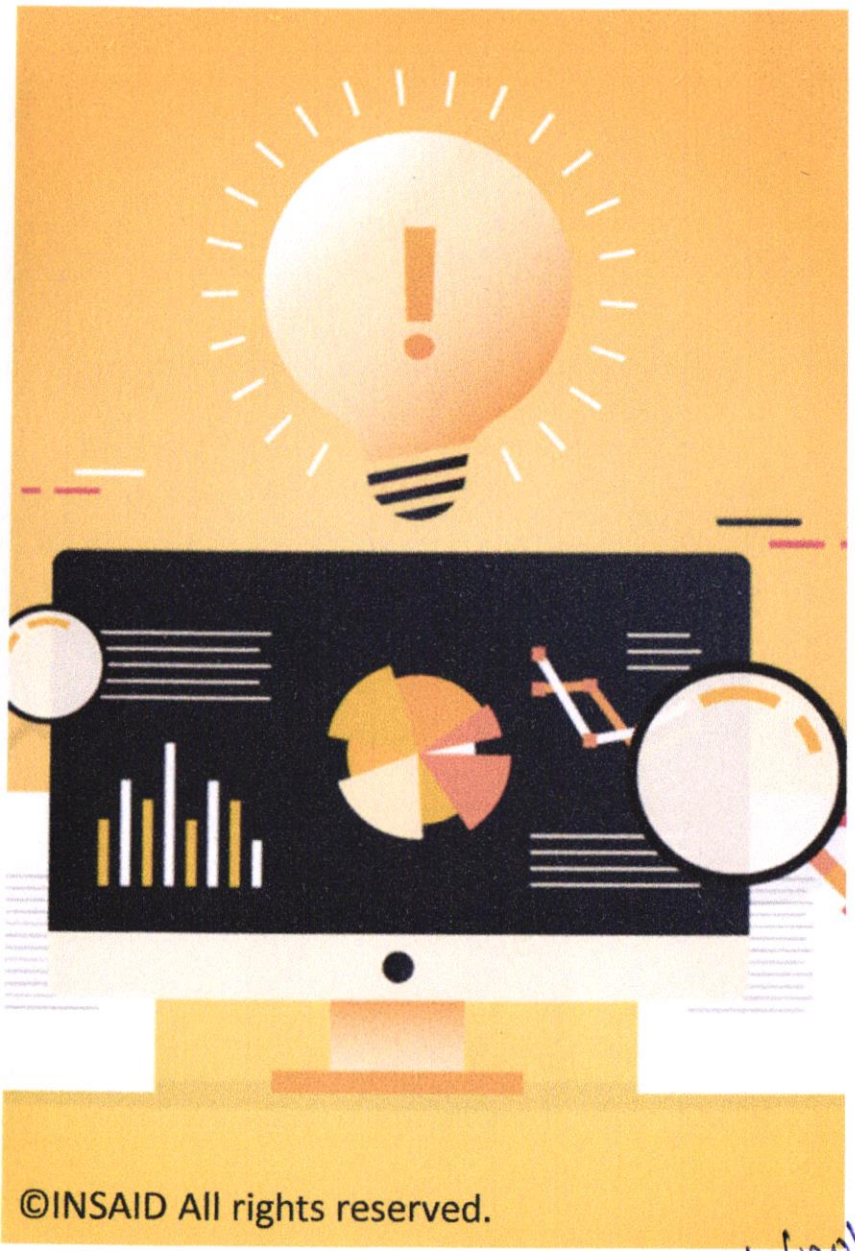
INSAID



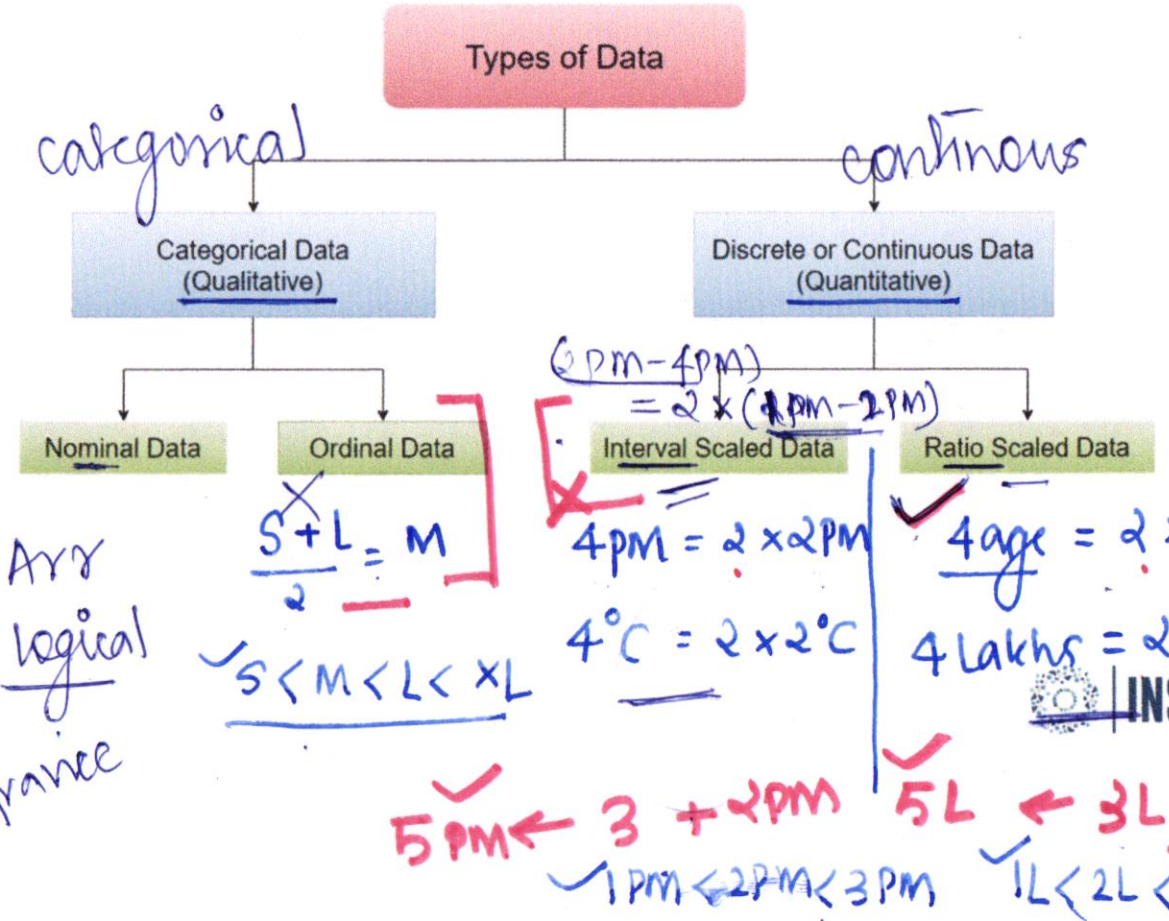
Agenda

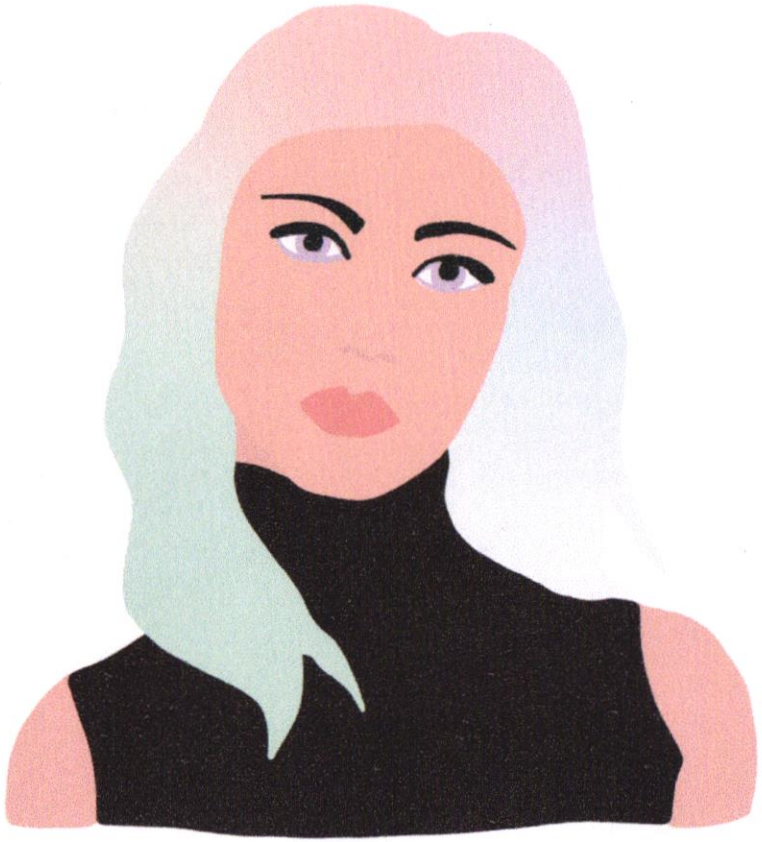
- **What is Data?**
- Data Science?
- Orbit of Data Science
- Data Scientists
- Traits of Data Scientists
- Deeper Discussions
- Problems Solved by Data Science
- Industries Involved in Data Science
- Data Science Lifecycle
- Types of Data
- Data Structures
- Appreciation of Completion

INSAID



Types of Data





Nominal Data

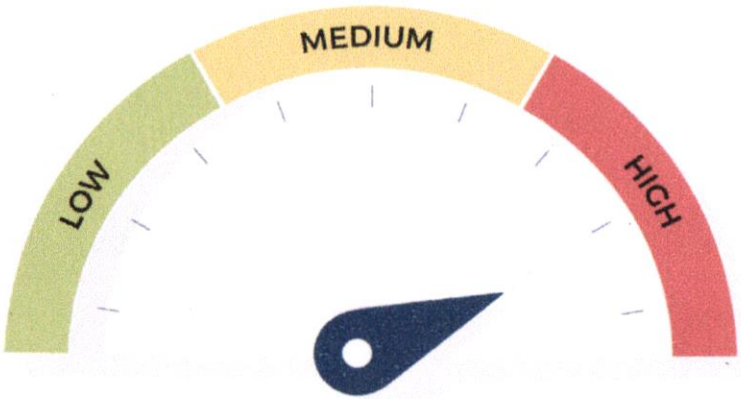
- Data **related** to **names**.
- The values are **symbols** or **names**.
- Also referred as **Categorical** Data.
- **Example:** Hair Color etc.

Binary Data

- Nominal data **only** with **two** categories.
- **Example:** Winner & Loser etc.
- **Note**: The **order** is **NOT** important such as Male/Female or Female/Male.

INSAID

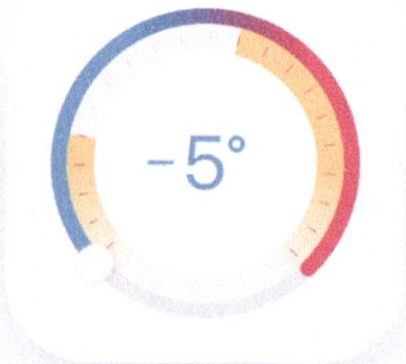
Ordinal Data



- Data **values** have meaningful **order** or **ranking**.
- **Magnitude** between successive values is **not known**.
- **Example:** Economic Status etc.

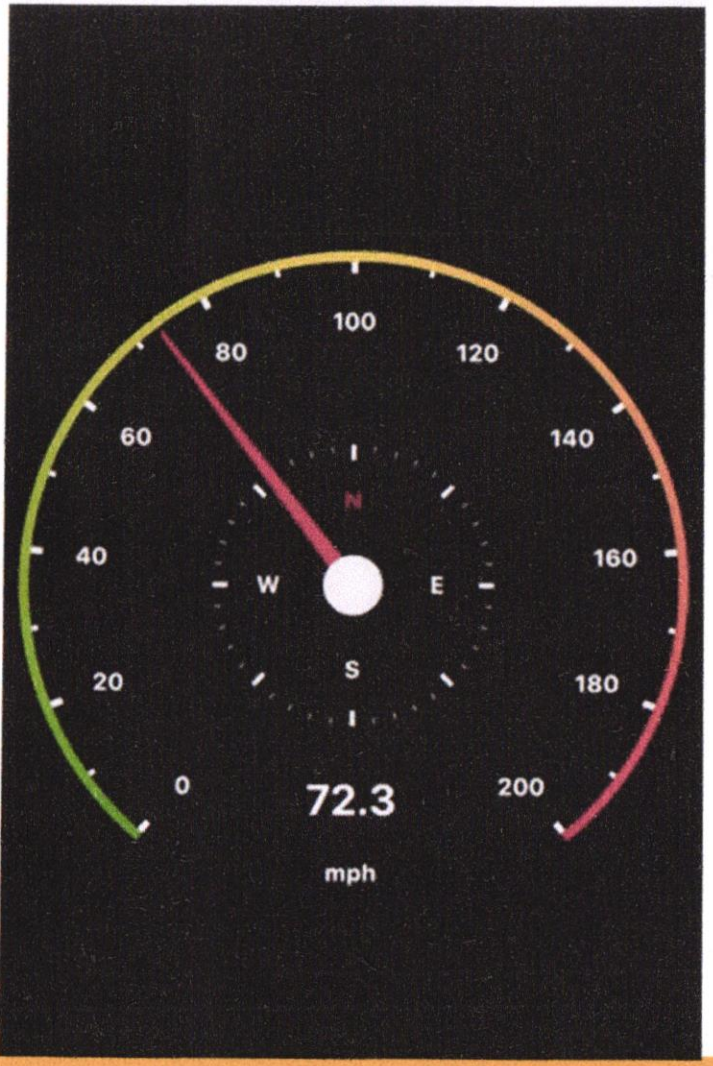
INSAID

Numeric Data – Interval Scaled



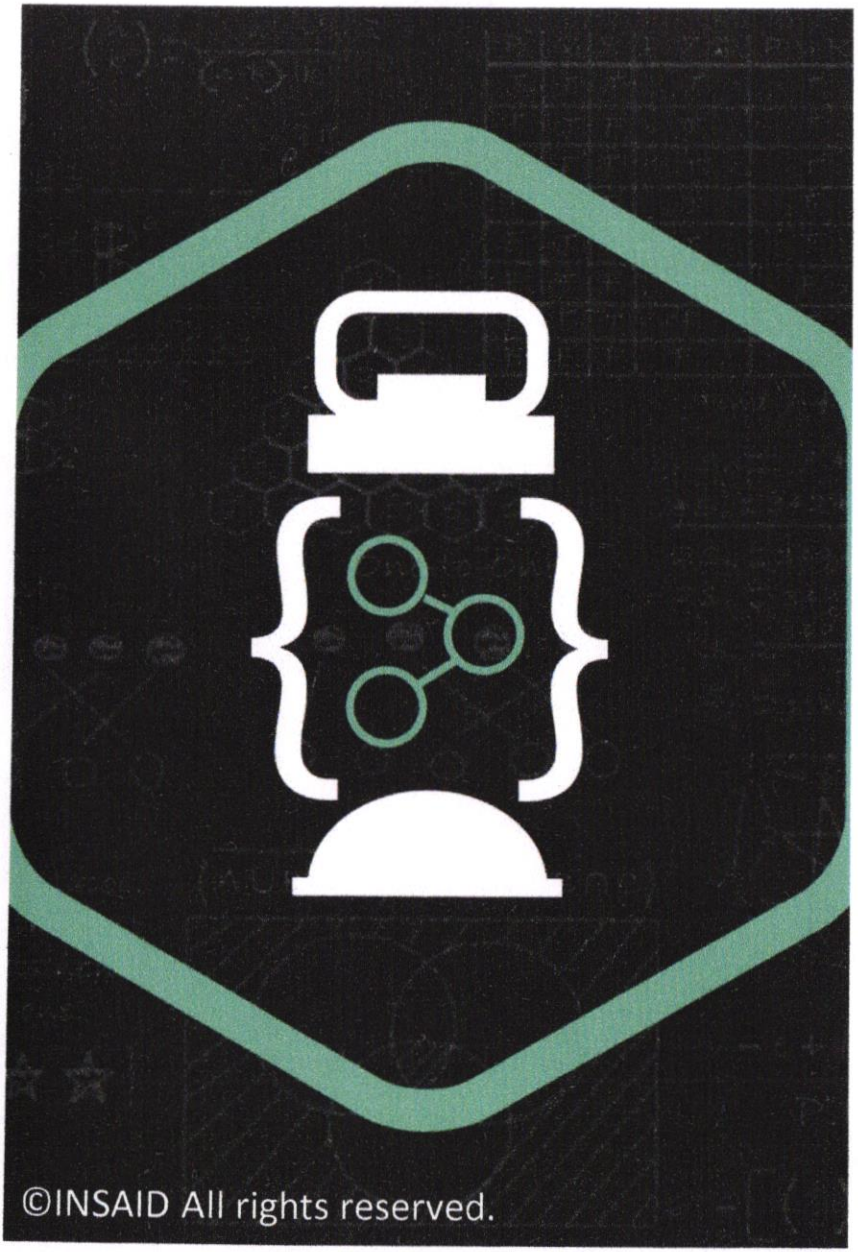
- It is **quantitative** in nature with **non-inherent zero-point**.
- Can be **discrete** or **continuous** in nature.
- Measured on **scaled** of **equal units**.
- Can be Positive, Zero or Negative.
- Allows **quantify** the **difference** between values.
- **Example:** Temperature etc.

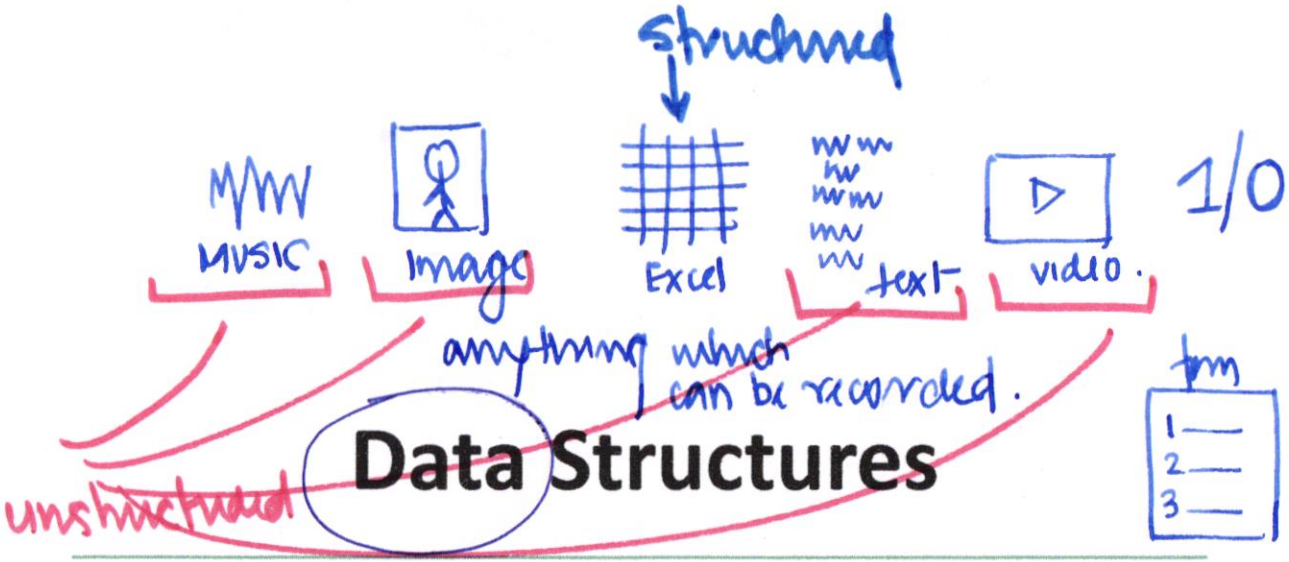
INSAID



Numeric Data – Ratio Scaled

- It is **quantitative** in nature with **inherent zero-point**.
- Can be **discrete** or **continuous** in nature.
- Measured on **scaled** of **equal units**.
- Can be Positive, Zero or Negative.
- Allows **quantify** the **difference** between values.
- **Example:** Speedometer etc.





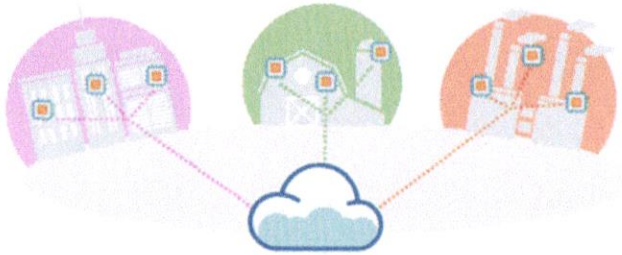
- Structured Data *
- Semi-structured Data
- Unstructured Data → structured
- Graph Data (text)
- Streaming Data

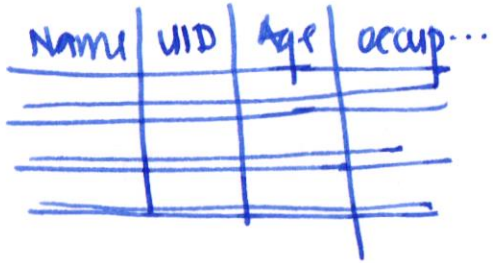
INSAID

name, age, occupation
deepak, 34, MLteacher
shina, 31, businessanalyst
Adi, 33, banker

Structured Data

SQL Excel





- Has a well defined structure.
- Follows a consistent order.
- Should be accessed easily.
- **Example:** Tabular data such as SQL table or excel table, etc.

INSAID

	Avg. Salary	# Workers
ML	20	1000
Telco.	5	20000
CS/IT.	10	5000
manu.	8	15000

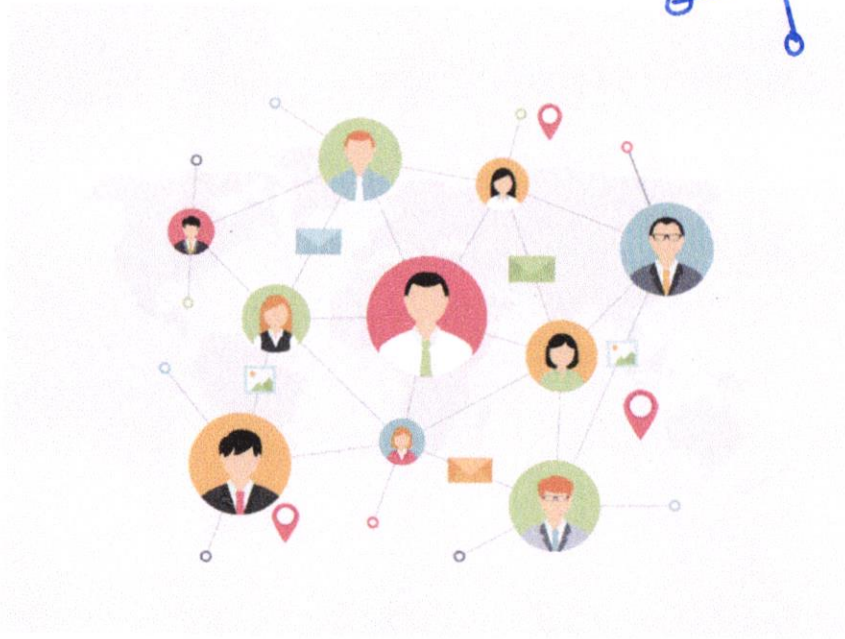
Unstructured Data

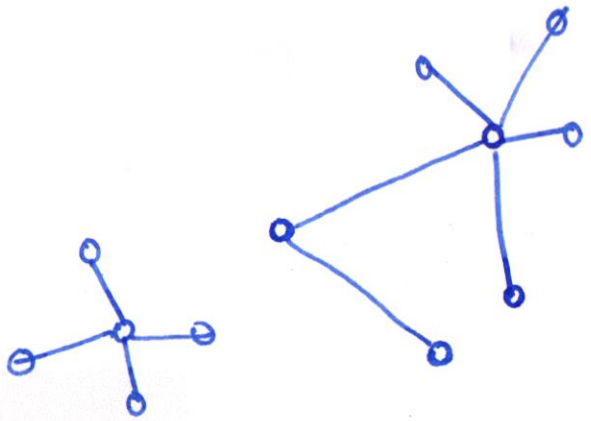
- Don't have a structure.
- Typically text or images or videos.
- Not easily searchable.
- **Example:** Document etc.

video.
Image
Audio.
text — article book tweets feedbacks

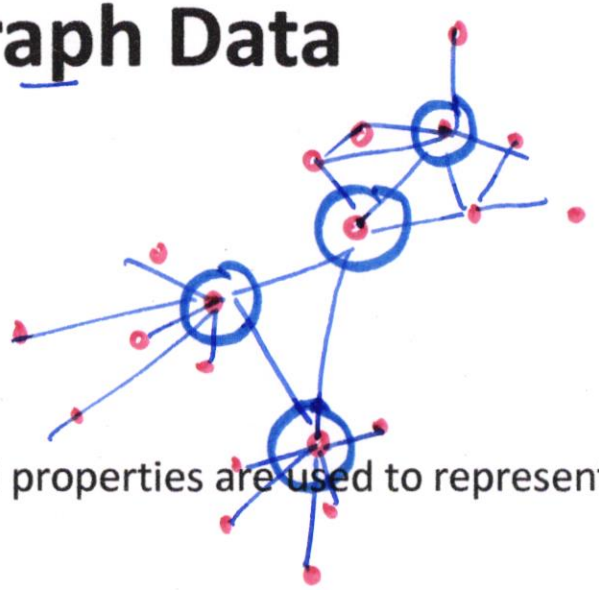
Semi – Structured Data

- Has some structure.
- Does not resides in relational database.
- Lacks a fixed or rigid schema (as representation).
- Some organizational properties make it easier to analyze.
- **Example:** JSON data, etc.

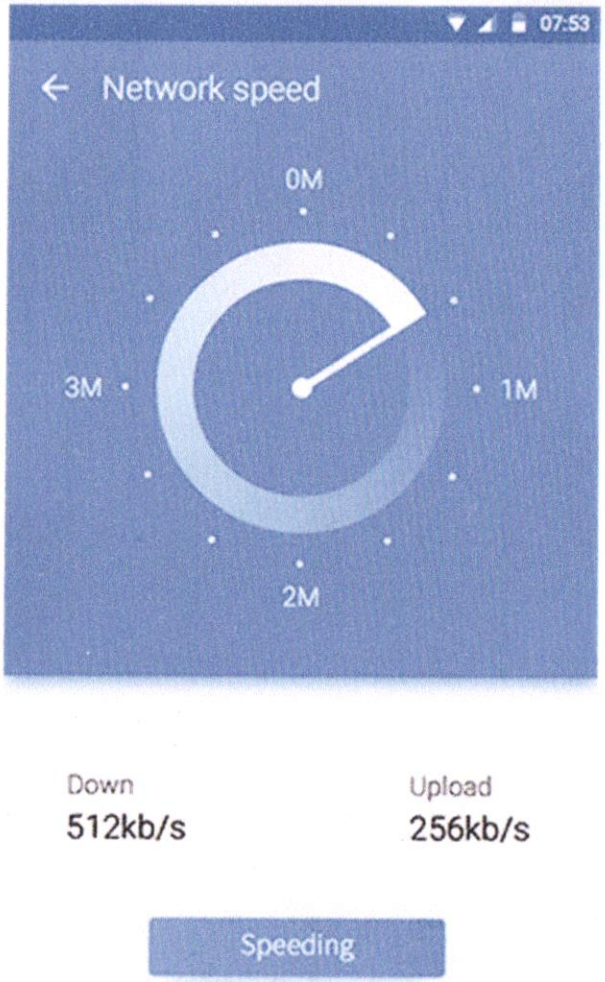




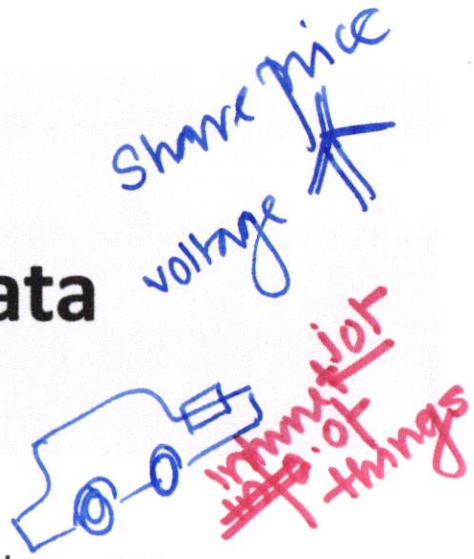
Graph Data



- Nodes, edges, and properties are used to represent and store data.
- **Example:** LinkedIn connections, Neo4j etc.



Streaming Data



- **Continuously generated** by different sources.
- Require **continuous processing**.
- Not accessed all at once.
- **Example:** Speed Test of Network, YouTube, this class session on Zoom etc.

INSAID