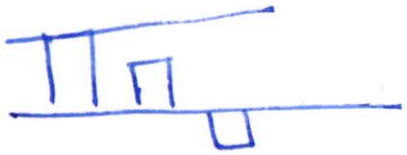
cohort - July 2020.
26 July 2020.
01 August 2020.



"start" "end"

INSAID

Deepesh Wadhwani
ME
LMS.

v all panelists and attendies.

VIC vs VINC

Data Science

Introduction

Making sense of Data

"WHAT."

storytelling.

academics@insaid.co

'Q-'

PART (2) of (2)

Python and stats
starting kit.

1

Basics of-
Term (1) and (2) — DS. Python
EDA

How
Why

Term 3 - ML

Agenda

INSAID

- Example to reflect on
- What is Data?
- What is Data science?
- Who are Data Scientists?
- **Deeper discussions**
  - Problems that DS Solves
  - Applications
  - Project life cycles
  - More on Data

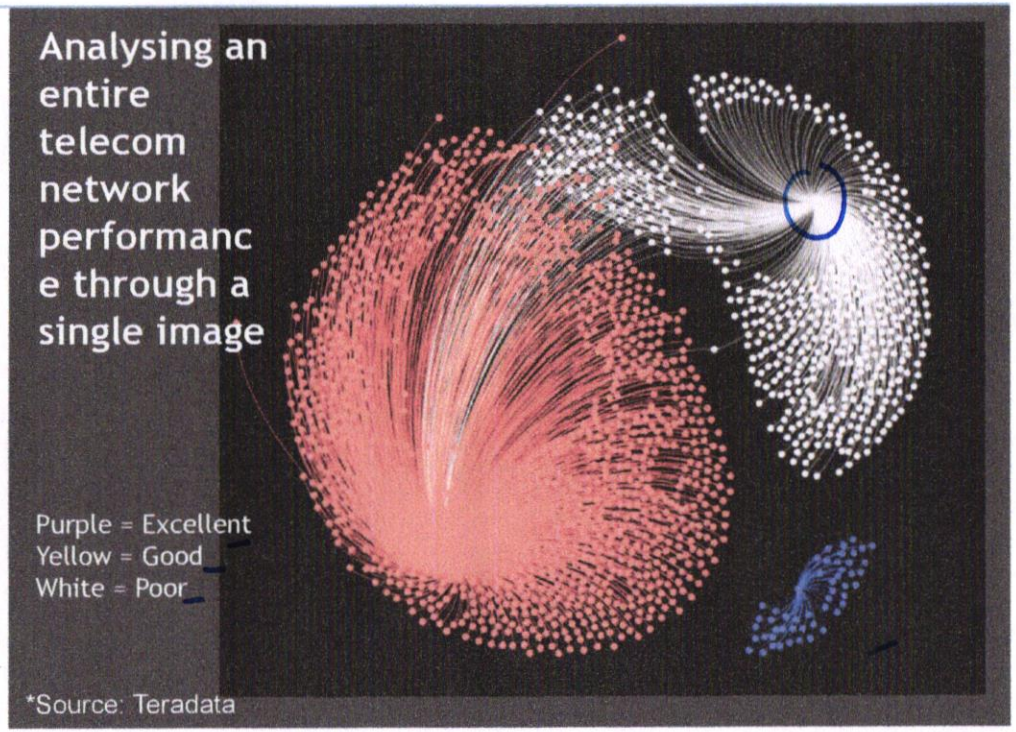
Data Science: Science or Art?

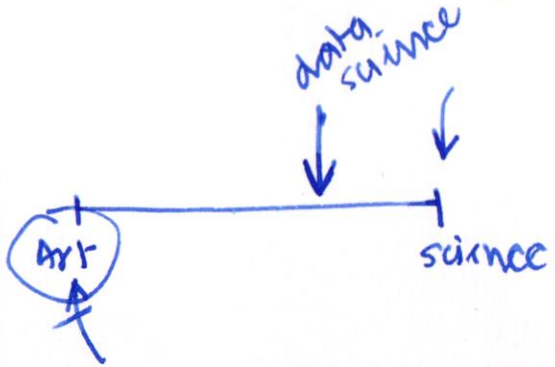
INSAID

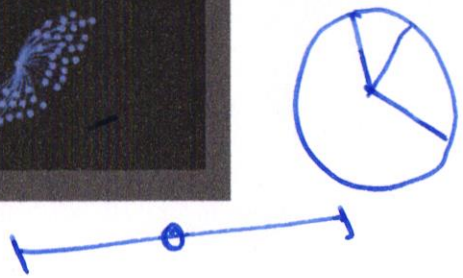
More Art, Less science.

Ways in which creativity can be expressed in data science include:

- Applying techniques in unconventional settings.
- Combining techniques in unconventional ways.
- Developing new and unusual hypothesis.







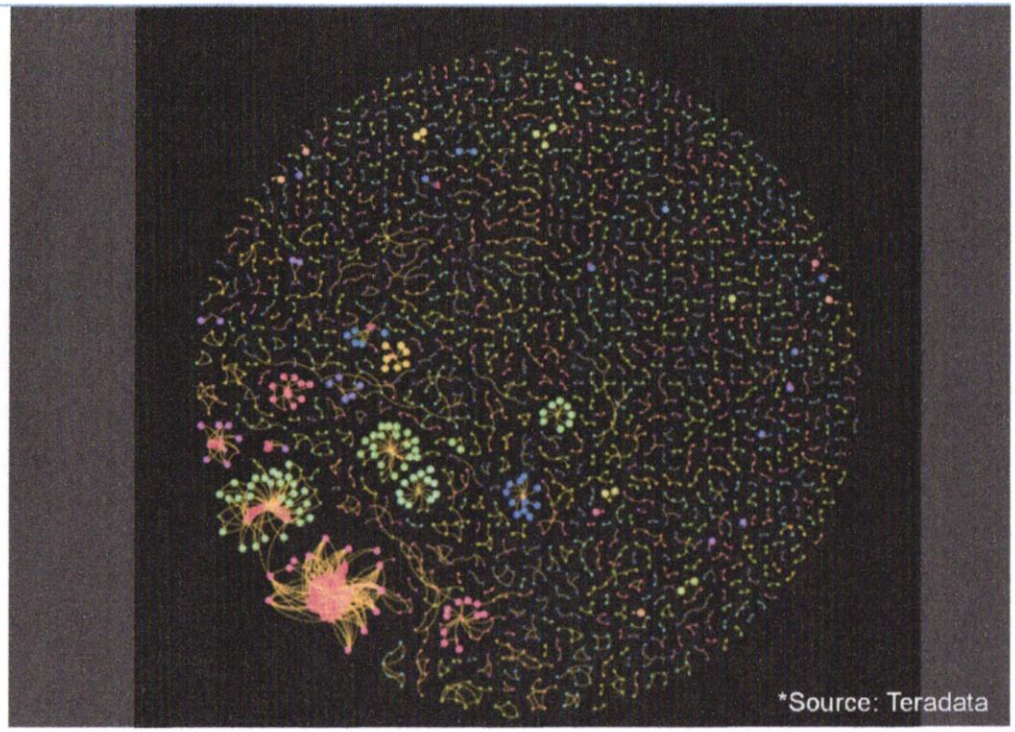
Data Science: Science or Art?

INSAID

Visualizing insurance fraud through data imaging

Dots = Claims
Large dots = Fraud claims
Small dots = Good claims

Establishes links between suspect good claims and proven fraud claims

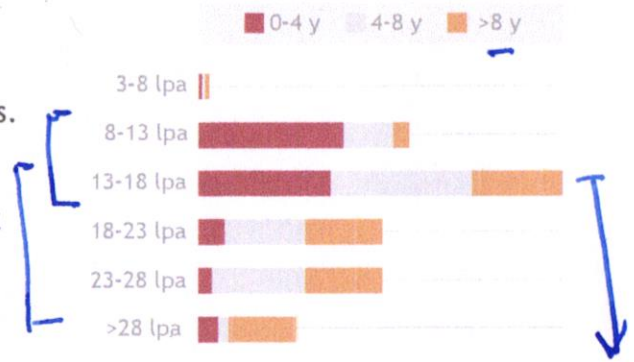


Data Science: Hype or Reality?

INSAID

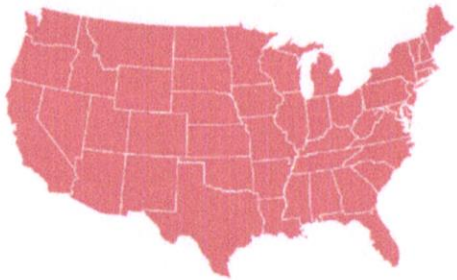
'Data Scientists' are hyped and celebrated by the press.

They are actually rare superstars who are great at math and statistics, also great programmers, and have great business understanding.



Data Scientist range of salaries in **India**

Heat map.



Data Scientist average salary in US **~ 120,000 USD**

*Source: 6figr.com

Show me the money!

INSAID

Why do Data Scientists get paid a lot?

Severe shortage of talent

Directly impact business value

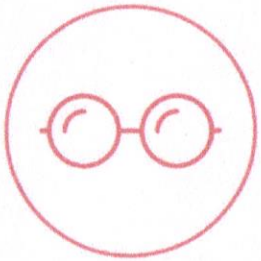
Organizations face enormous data challenges

Need not restricted to Tech Giants anymore

Good vs Great Data Scientists?

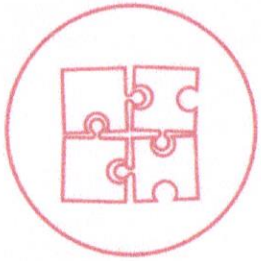
INSAID

Effective Communication



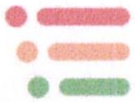
Intellectual Curiosity

Domain / Industry Knowledge



Analytic Problem-Solving

Agenda



- Example to reflect on
- What is Data?
- What is Data science?
- Who are Data Scientists?
- Deeper discussions
- **Problems that DS Solves**
- Applications
- Project life cycles
- More on Data

Problems that Data Scientists solve

INSAID

- Is this A or B?
- Is this weird?
- How much or how many?
- How is this organized?
- What should I do now?

Problems that Data Scientists solve

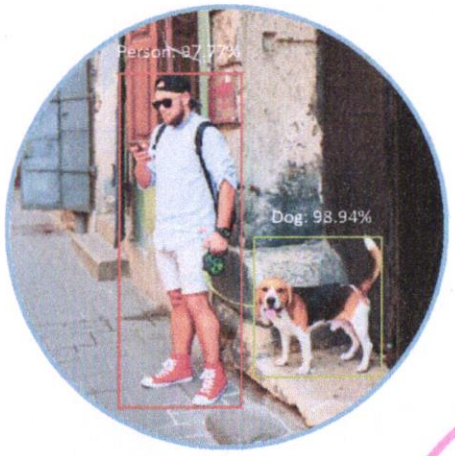
INSAID

Is this A or B?

Classification

1

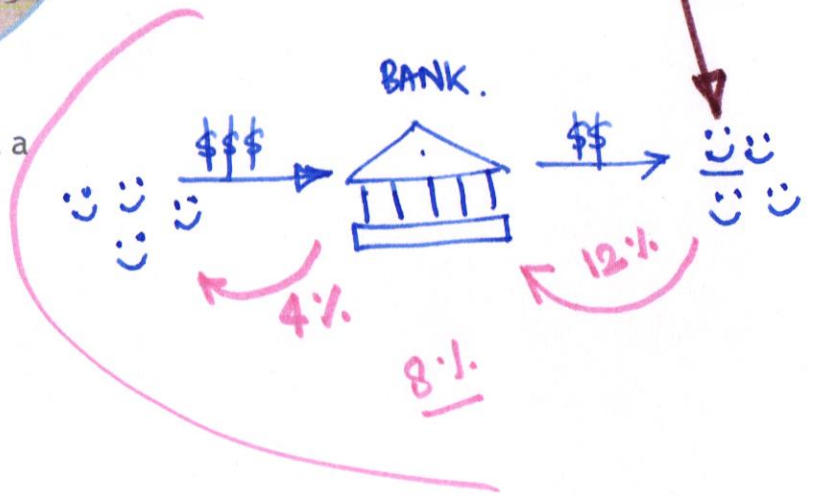
Will this applicant be able to repay the loan?



Identify a human & a dog?

Buy, not Buy.
spam, not spam

(default, not default)



Problems that Data Scientists solve

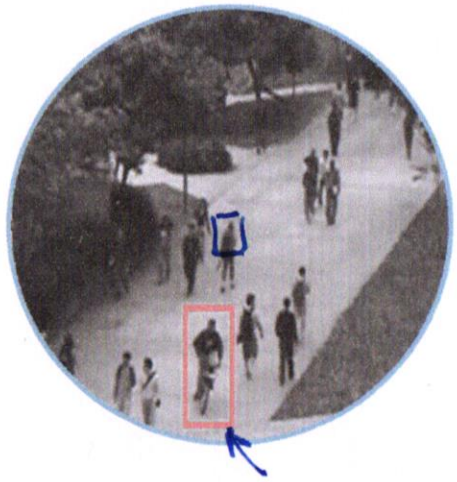
INSAID

Is this weird?

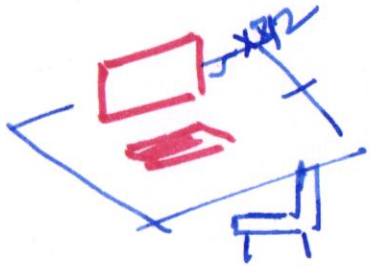
Anomaly

2

Is this mail in your inbox spam?



Is someone riding bike on a walkway?



Problems that Data Scientists solve

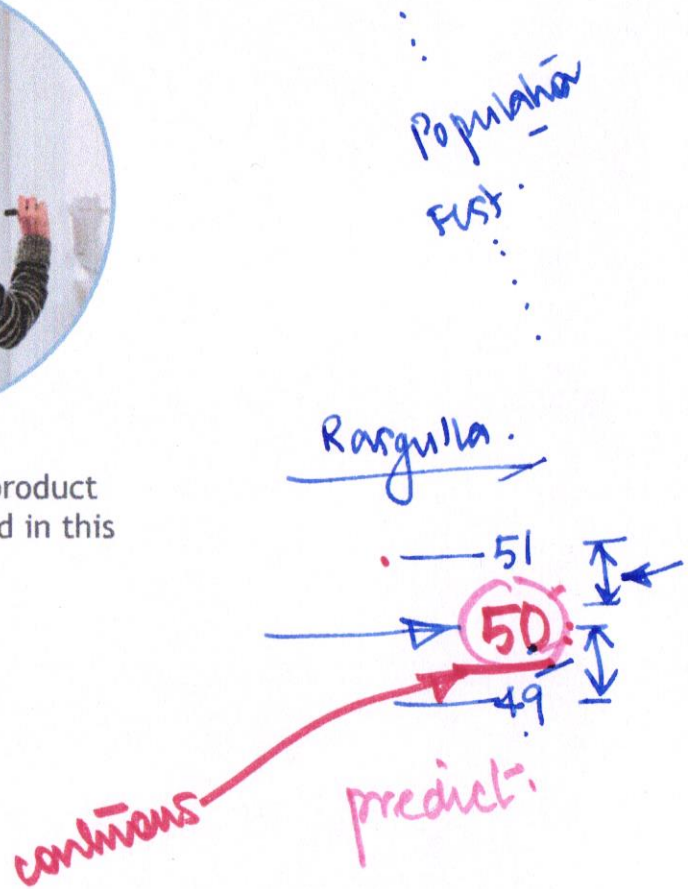
INSAID

How much
or
How many?
Regression

3

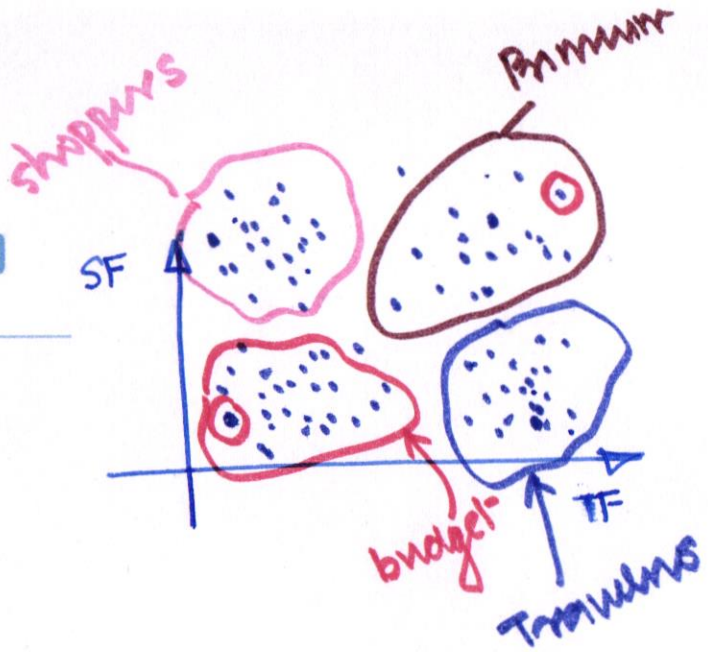
How much will this used car sell for?

Predict how much product sales will be achieved in this month?



Problems that Data Scientists solve

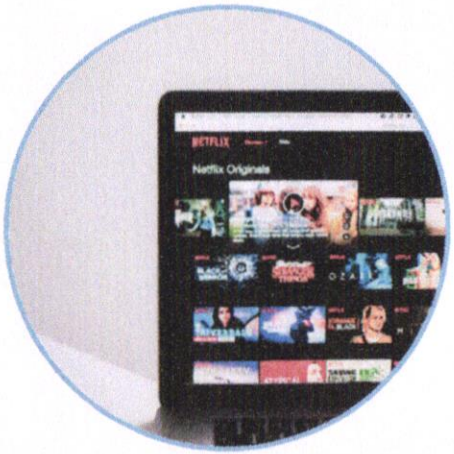
INSAID



How is this organized?

Clustering

4



Which viewers like the same kind of movies?

similarities

Which pet owner groups have the same purchase behaviour for pet foods?

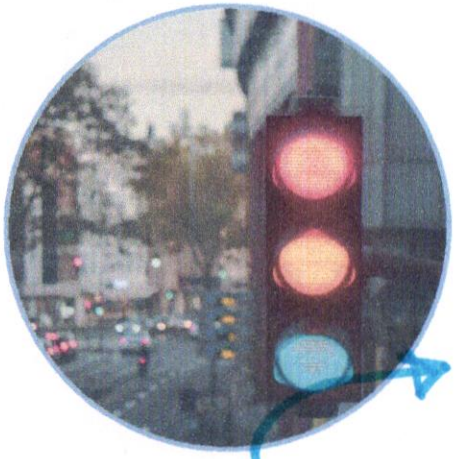
Problems that Data Scientists solve

INSAID

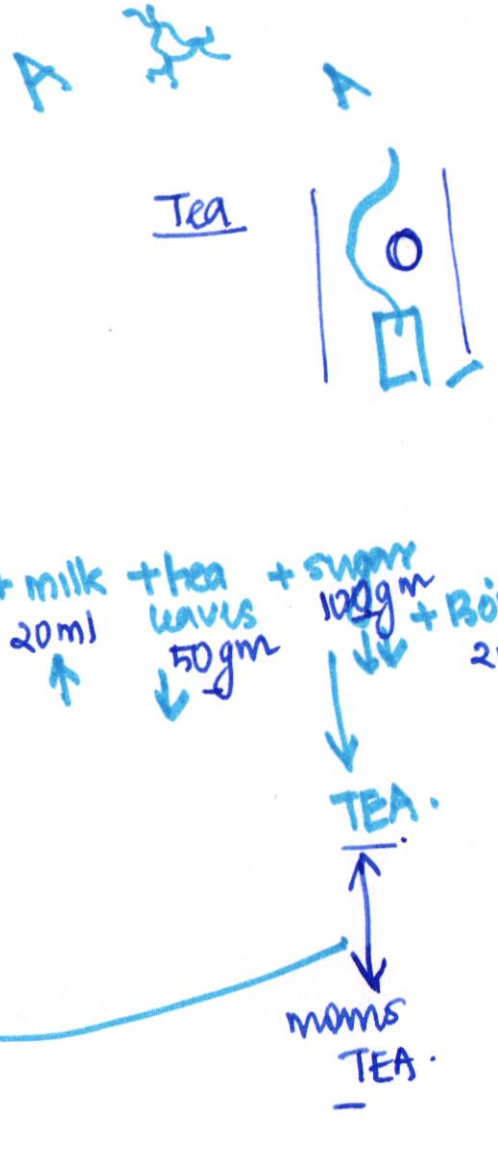
What should I do now?
Reinforcement learning

5

How far should the Robot Jump?



Should I break, accelerate, or continue at that yellow light?

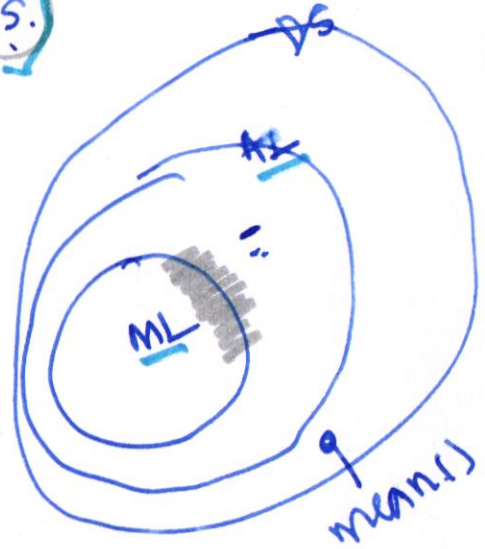


Agenda

INSAID

- Example to reflect on
- What is Data?
- What is Data science?
- Who are Data Scientists?
- Deeper discussions
- Problems that DS Solves
- **Applications**
- Project life cycles
- More on Data

1. classification - A/B
2. ~~Anoma~~ Anomaly - weird
3. clustering
4. Regression
5. RL



Data Science: Applications in Industry

INSAID

Telecom

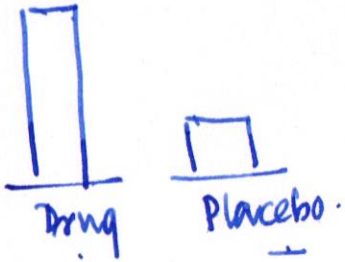
- Customer Acquisition Strategies
- Churn Analysis and Control
- Up-sell / Cross-sell
- Product Bundling

Data Science: Applications in Industry

INSAID

Healthcare

- Clinical trials of new drugs
- A / B Testing
- Genetics Analysis
- Epidemic Forecasting and Control



version.

AB / A AC. / B

Data Science: Applications in Industry

INSAID

Banking & Finance

- Fraud detection and prevention — A/B.
- Customer Segmentation – clustering
- Risk management — credit risk management — A/B.
- Portfolio Optimization — RIL or TML

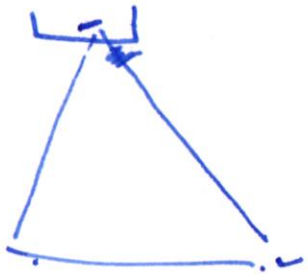
Data Science: Applications in Industry

INSAID

Energy & Utilities

- Predictive maintenance for machinery
- Predict theft and loss prevention
- Churn analysis and prevention
- Auto balance resources and dynamic routing

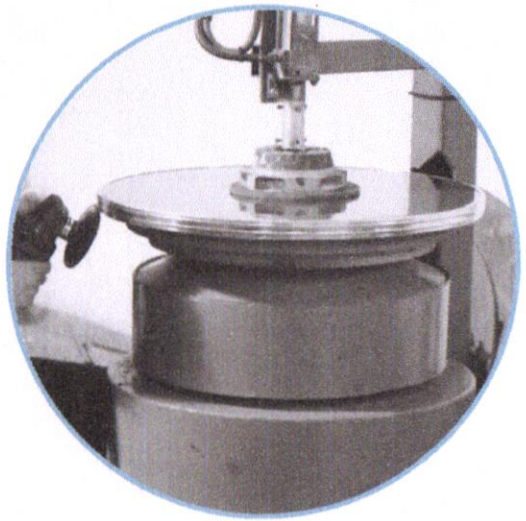
A/B, Anomaly



Data Science: Applications in Industry

INSAID

Manufacturing



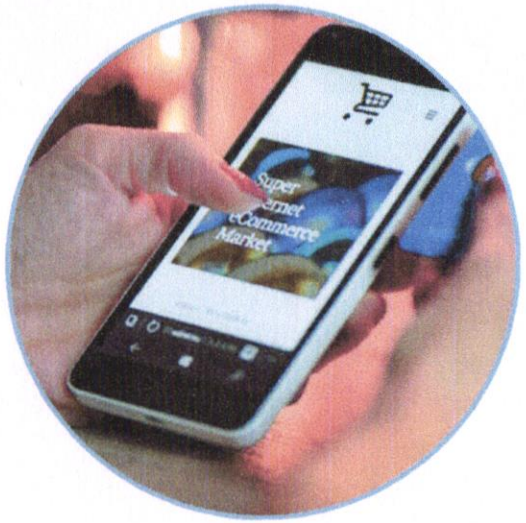
- Custom product design
- Better quality assurance
- Improve manufacturing processes
- Managing supply chain risk

SIX SIGMA.

Data Science: Applications in Industry

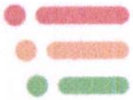
INSAID

Retail & e-commerce



- Shelf-space Optimization
- Market Basket Analysis
- Product Bundling
- Promotions
- Up-sell / Cross-sell

Agenda



- Example to reflect on
- What is Data?
- What is Data science?
- Who are Data Scientists?
- Deeper discussions
- Problems that DS Solves
- Applications
- **Project life cycles**
- More on Data

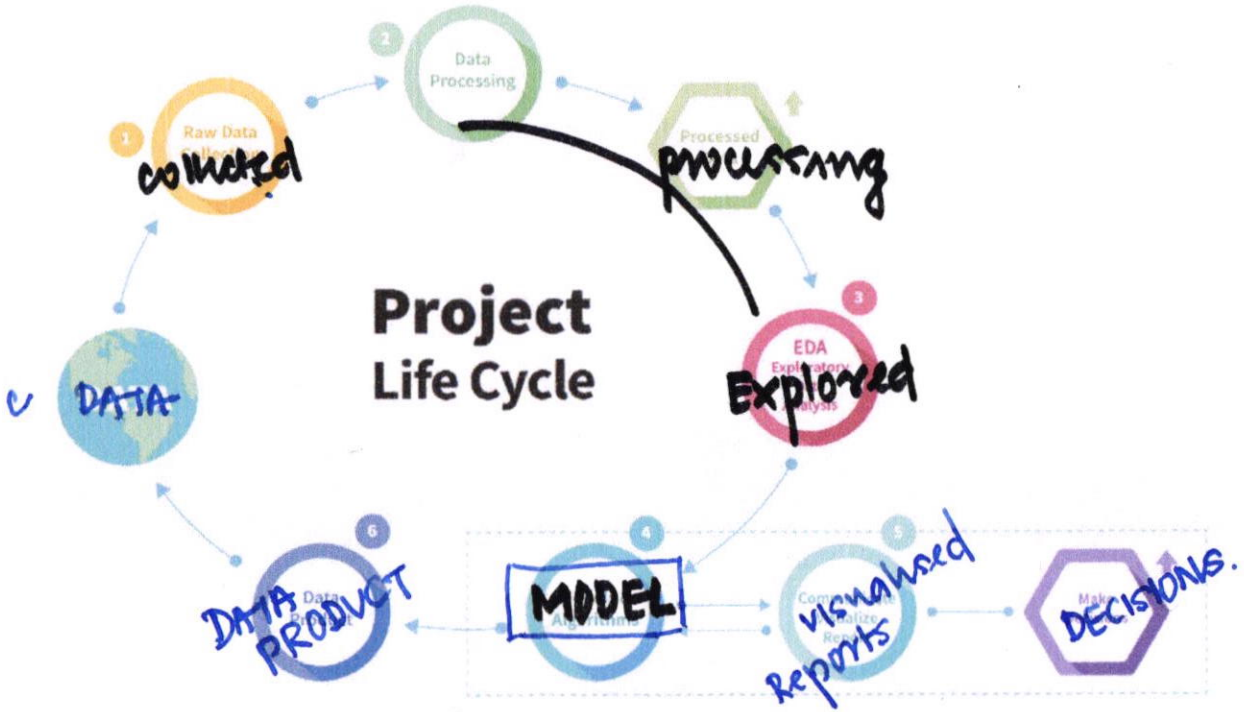
How do I start a Data Science project?

INSAID

Just follow the upcoming simple steps...

Data Science Process

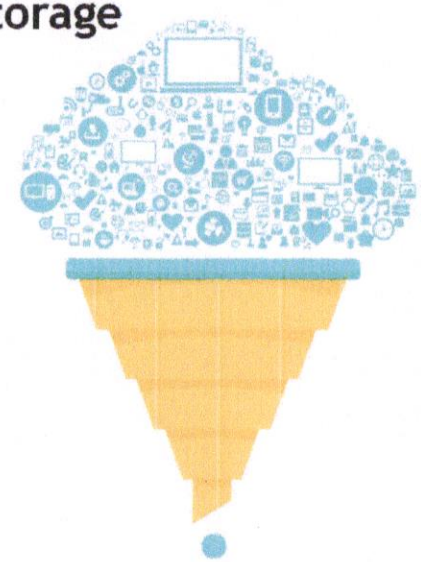
INSAID

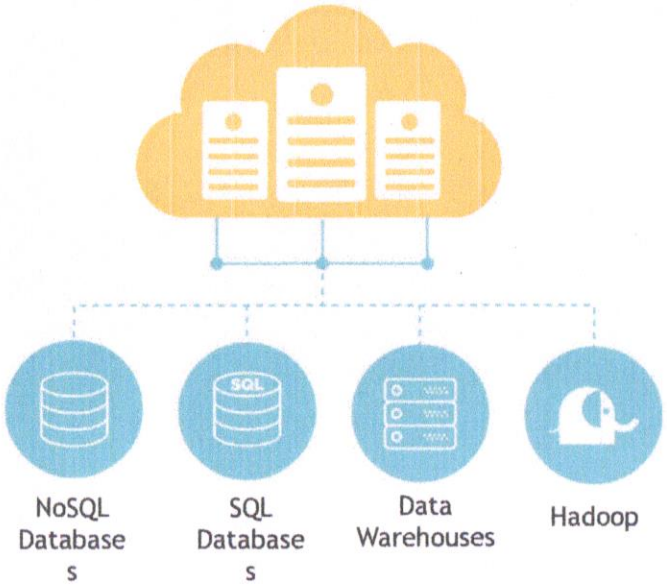


Data Science Process

INSAID

01 Raw Data collection & Storage





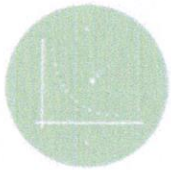
Excel sheet

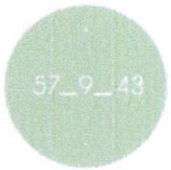
Data Science Process

INSAID

02 Data Cleaning & Processing

Pre-processing / Cleaning


Removing Outliers



Missing Data


Malicious Data


Erroneous Data


Irrelevant Data


Inconsistent Data


Formatting

20200802
02082020
0802 2020.

Data Science Process

INSAID

02 Data Cleaning & Processing

Data Cleaning Example

#	First Name	Last Name	Age	Gender	Mobile	Telephone	Address
1	David	Wong	26	Male		02 -2110063	3-12 soi S Rochtevi
2	Biii	Chrlnasom	43	Male	08-9999-8765		354 SOi 4 Nono SukumV'lt
3	Noncv	Nomochue	29	Female	08-1111-3456	02 -3 10-8633	Sukumv1l Klonutoev Banukok
4	Mie	Nakanishi	26	FeMale	0-1230- 9965	02 -260-5544	
5	David	1sh1ko	26	Mole	08 7511-1234		

Un-standardized

Missing/Misspelled

Duplications

#	First Name	Last Name	Age	Gender	Mobile	Telephone	Address
1	David	Ishikawa	26	Male	QS.7511-1234	02-211-0063	3-12 soi 5 Rochtevi
2	Bill	Chrinasom	43	Male	08-9999-8765	02-416-0011	354 soi4 Nono Sukumvit
3	Noncv	Nomdluc	29	Female	08-1111-3456	02-310-8633	S-IO A Mansion 104 SOI 4 Sukumvil
4	Mie	Nakan shi	26	Female	0-1234-9965	02-260-5544	71 soi Rotnrao

Standardized

Populated

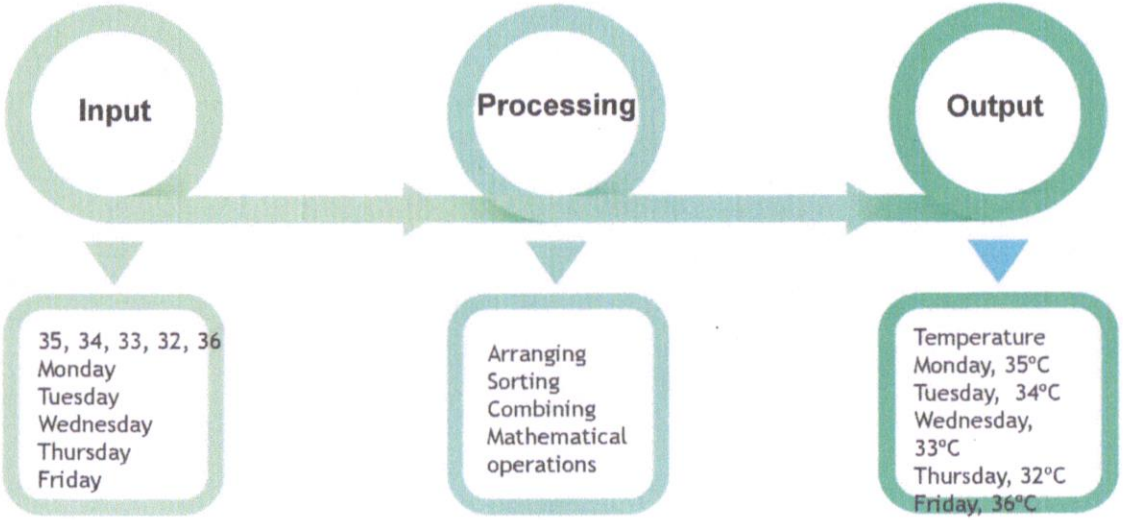
Duplicate removed

Data Science Process

INSAID

02 Data Cleaning & Processing

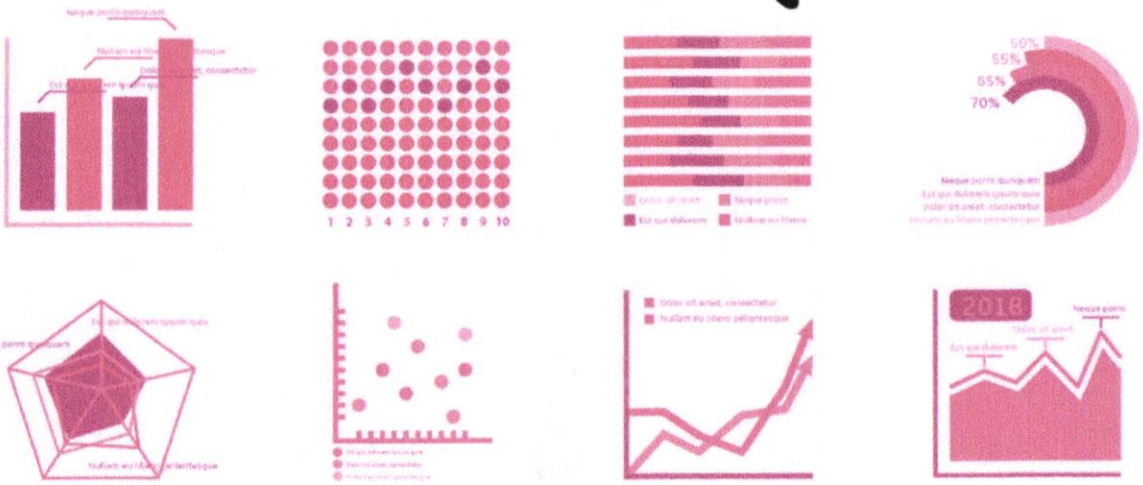
Data Processing Example

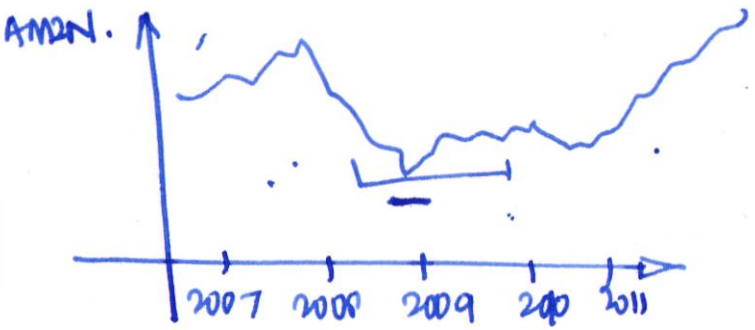


Data Science Process

INSAID

03 Exploratory Data Analysis (EDA) - storytelling.



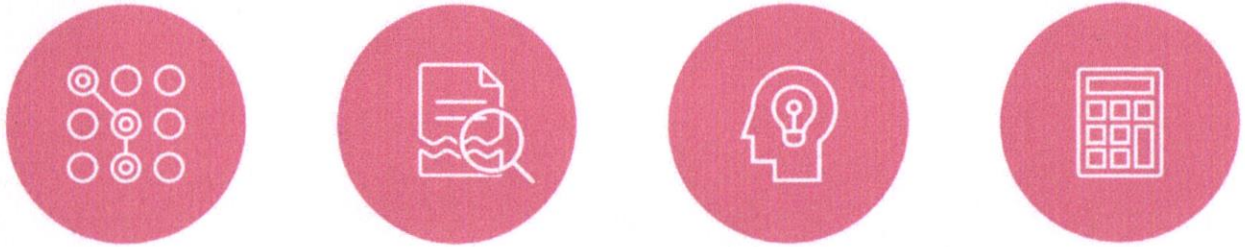


Data Science Process

INSAID

03 Exploratory Data Analysis (EDA)

Objectives



Discover Patterns

Spot Anomalies

Frame Hypothesis

Check Assumptions

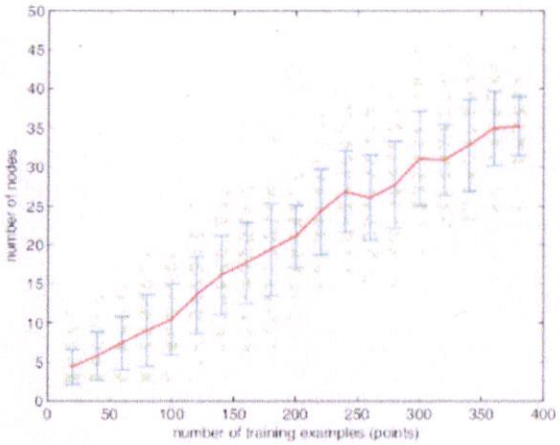
Data Science Process

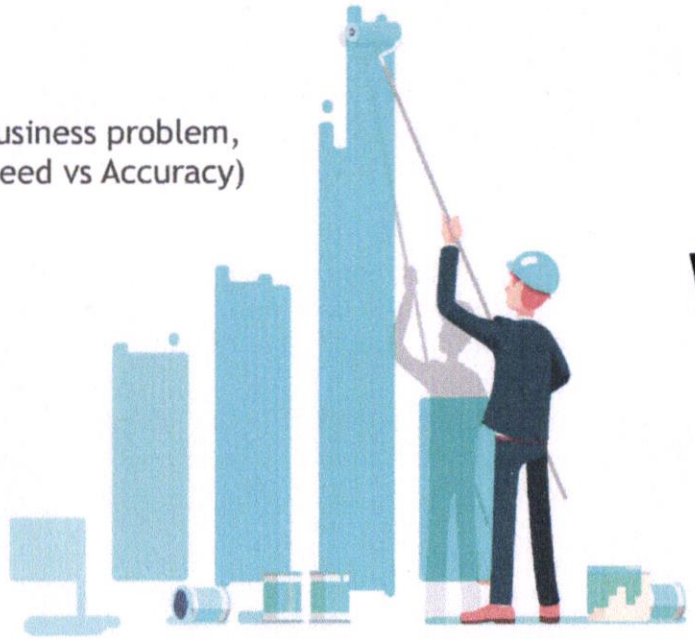
INSAID

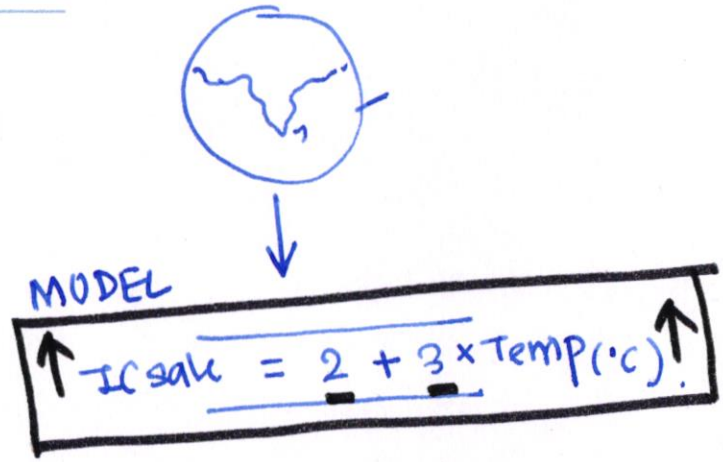
04

Models & Algorithms

Create multiple models to solve the business problem, and find the most appropriate one (Speed vs Accuracy)







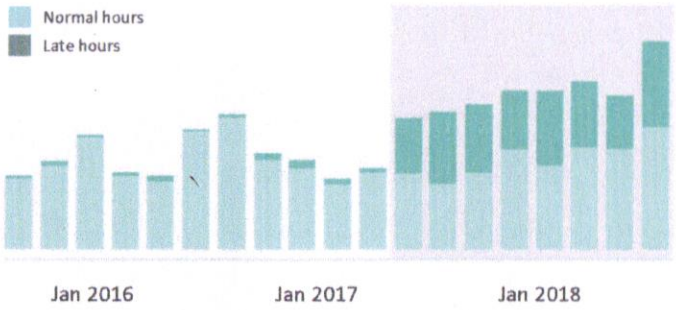
Data Science Process

INSAID

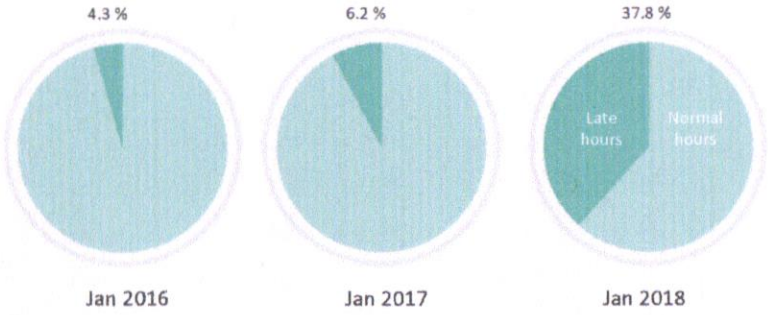
05 Communicate, Visualize & Report

Take action and deploy findings in the real world

Significant increase in late night technician repairs since Jan 2016



Late night technician repairs increase to almost 40% in Jan 2018

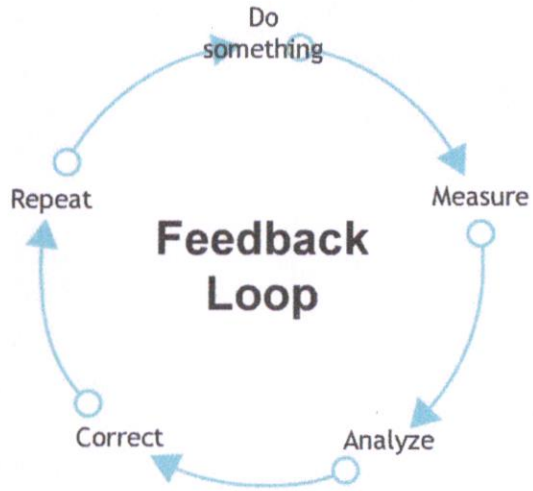


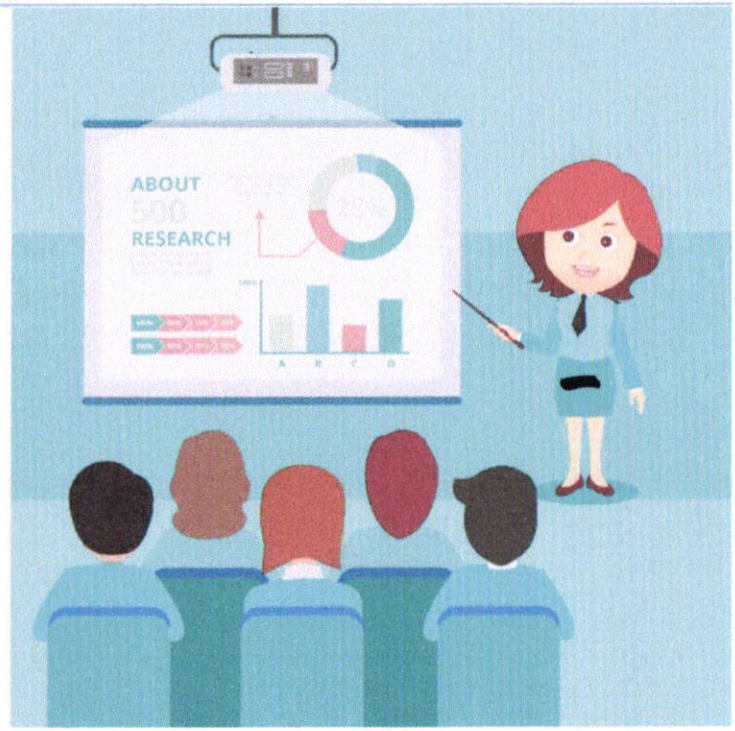
Data Science Process

INSAID

05 Communicate, Visualize & Report

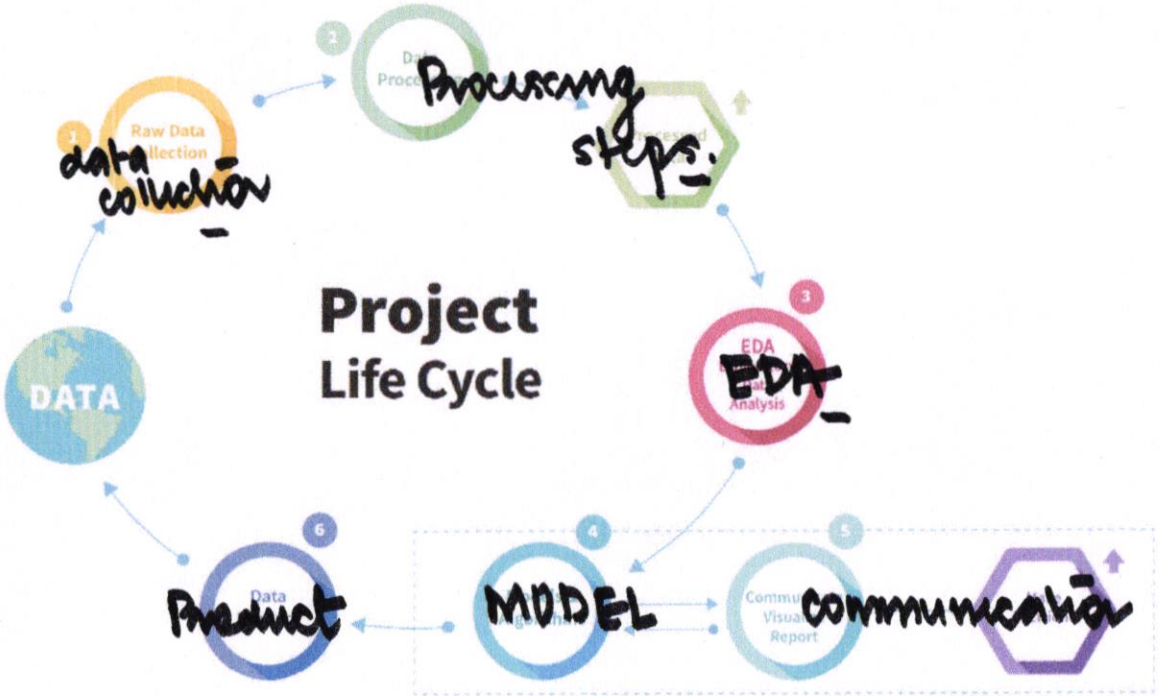
Present results to target stakeholders



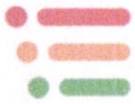


Data Science Process

INSAID



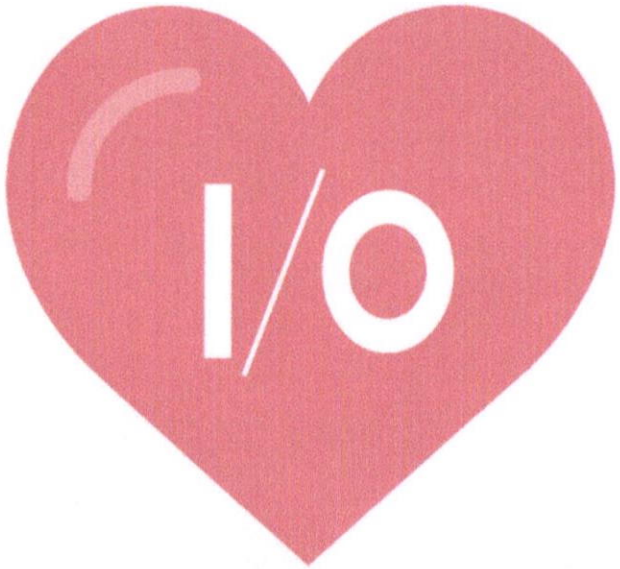
Agenda



- Example to reflect on
- What is Data?
- What is Data science?
- Who are Data Scientists?
- Deeper discussions
- Problems
- Applications
- Project life cycle
- More on Data

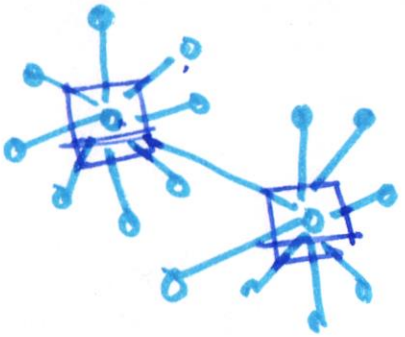
What is at the heart of Data Science?

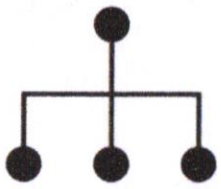
INSAID



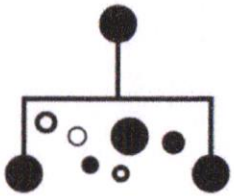
DATA

Types of Data

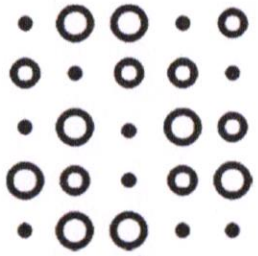




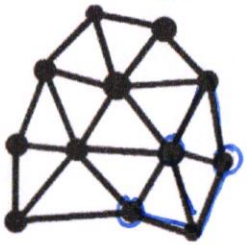
Structured data



Semi-structured data

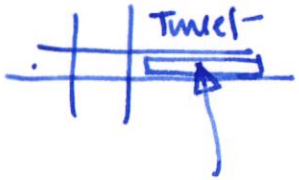


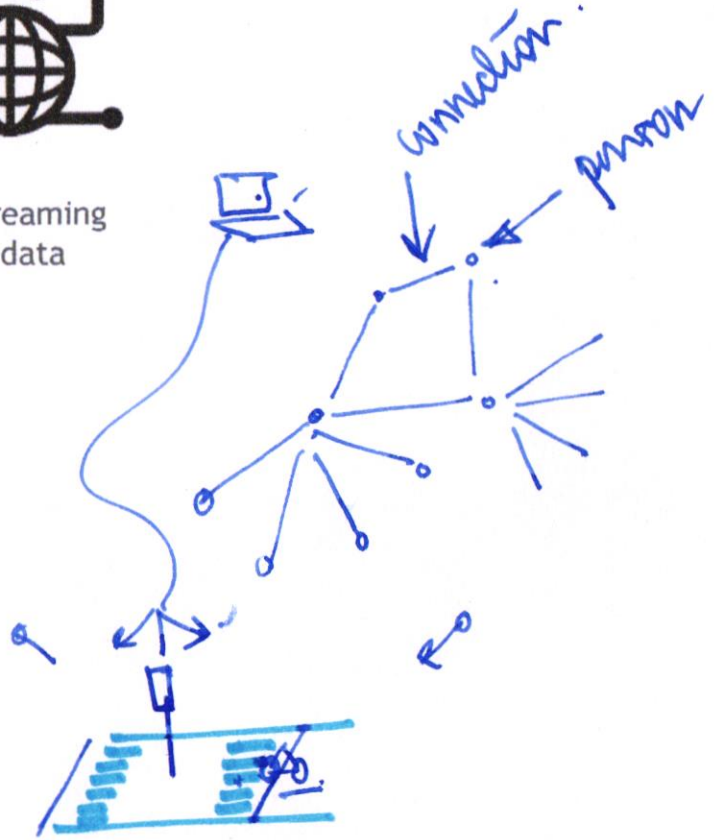
Unstructured data



Graph data

Streaming data

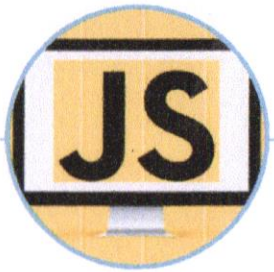




Types of Data

INSAID

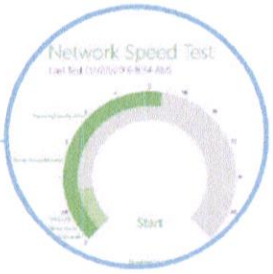
Structured data



Semi-structured data

Unstructured data

Graph data



Streaming data

Types of Data

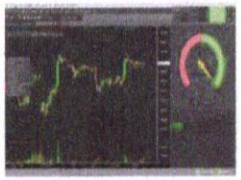
Structured data

Nagios® Log Server™

Semi-structured data

Unstructured data

Graph data



Streaming data

Types of Data

INSAID

Structured data - Transactions

Transactions Before November 21 2016

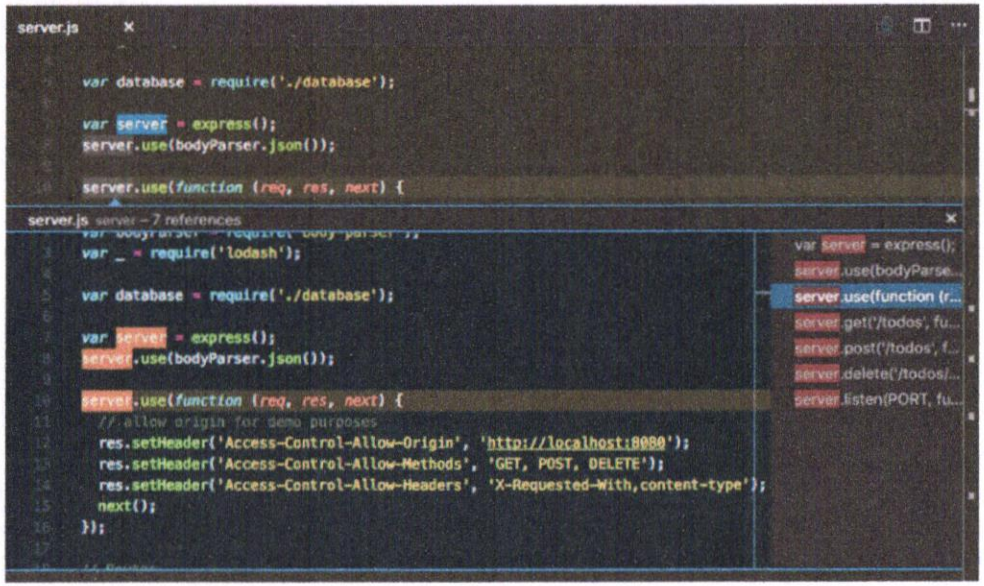
✱ All ▾ | Payment Type ▾ | Event ▾ | Staff ▾ | Category Filter ▾ | Search transactions | 🔍

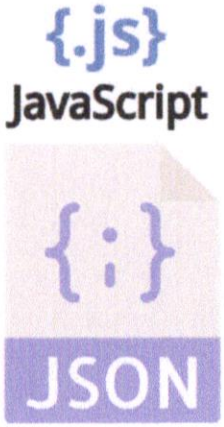
Transaction ID	Event	Amount	Payment Type	Account	Payment Customer	Gateway	Gateway Transac	Info
KBBJ-071116-1	✓ PAID	$89.60	ACCOUNT VOUCHER1	Dolly	Dolly	None	5820cc52a177b	Details
GFTM-170616-1	✓ PAID	$42.00	AMEX ****0005 John Smith	website	John Smith	None	1466209337	Details
CSZV-170616-1	✓ PAID	$61.60	AMEX ****0005 Brendan Mcclure	website	Brendan Mcclure	None	1466213357	Details
FCDQ-170616-1	✓ PAID	$123.20	AMEX ****0005 Brendan Mcclure	website	Brendan Mcclure	None	1466213761	Details
CHZG-290915-1	✓ PAID	$316.25	Cash Norris Cole	website	Norris Cole	Vend	7ec56036-6258-8	Details

Types of Data

INSAID

Semi structured data - JavaScript code

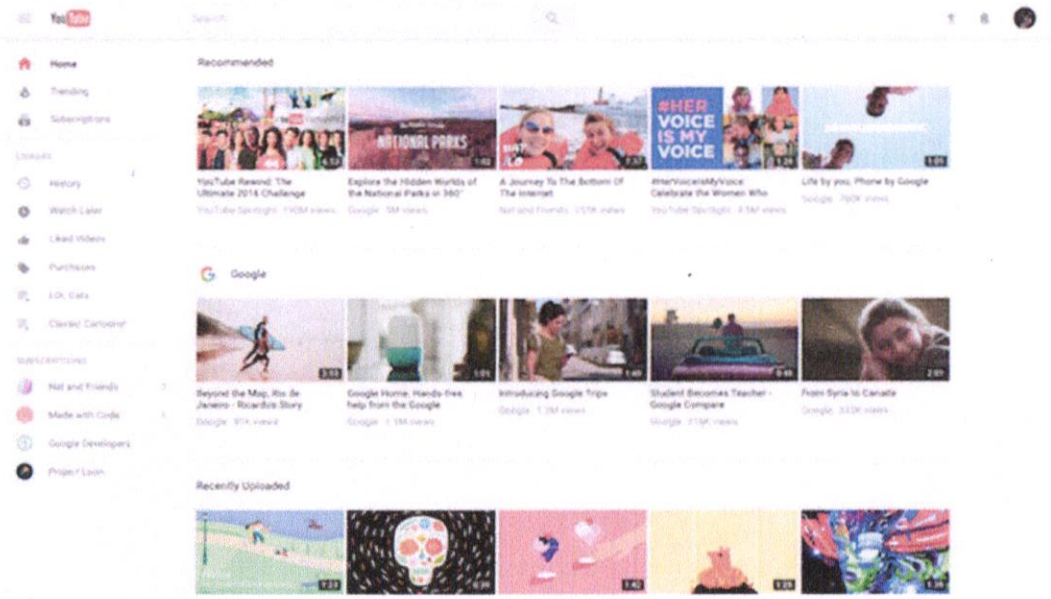




Types of Data

INSAID

Unstructured data - YouTube Feed



Types of Data

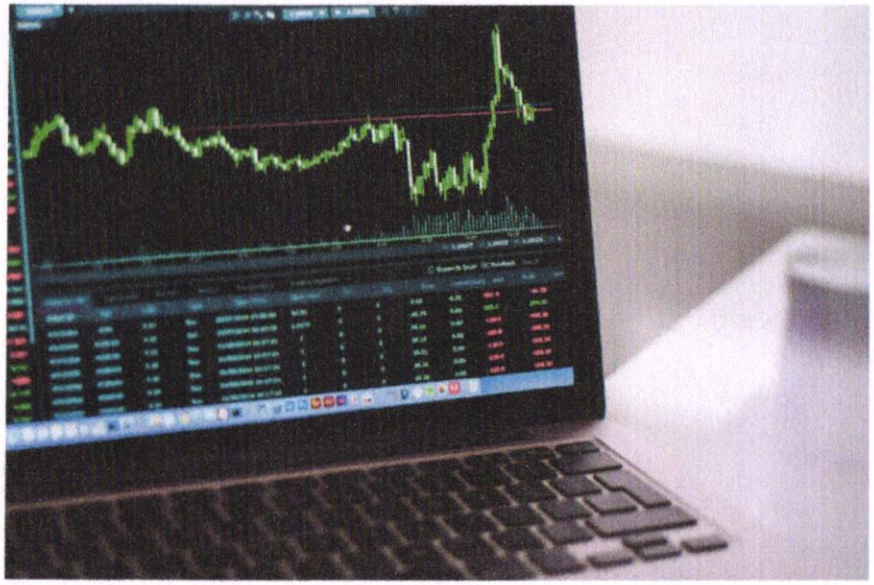
INSAID

Graph data - Facebook friend network

Types of Data

INSAID

Streaming data - Stock prices



Types of Data: Alternate view

INSAID

Quantitative — CONTINOUS.

a. Discrete

Two horses

b. Continuous

Height

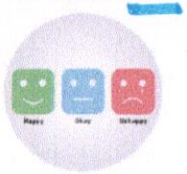
Qualitative — CATAGORICAL

a. Nominal

Male

Femal e

b. Ordinal



Customer Service

Interval

Temp

-273°.C

Time scale

Time

13.5 Byens

ΔT .C

10

0

Today.

0

Ratio

Weight

$L = \frac{XL + M}{2}$

#pens.

. . . . . ∞

amt. of water in a bottle

$Time_1 - Time_2$

$-\infty$ ∞

10 - 3 = 7

catagory

Delhi ≠ Mumbai ≠ ////

No mathematical op. possible

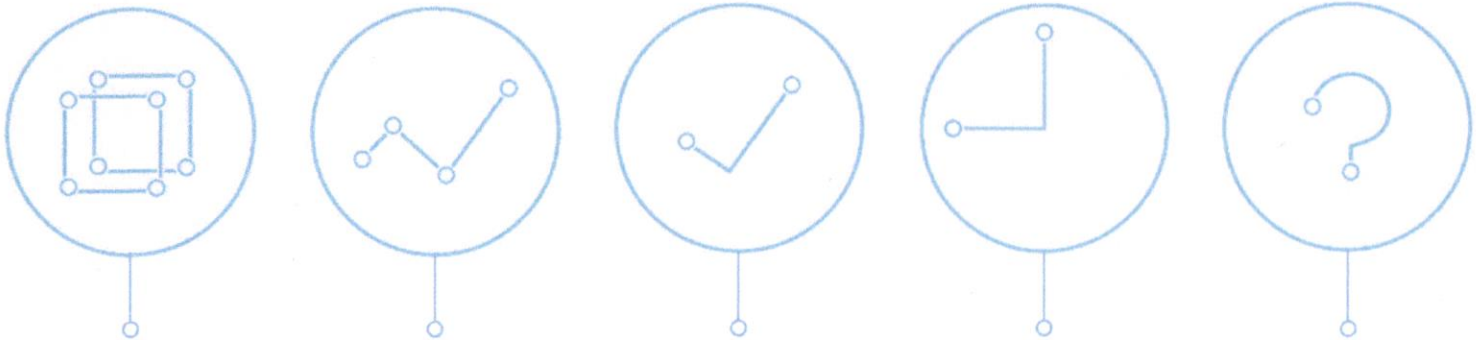
ordinal

XL > L

*5 > *4

Data Quality Issues

INSAID



Duplicity

Redundancy leading to resource wastage

Inconsistency

Withdrawal of INR 10/- not reflecting in Net Banking

Correctness

Age/Income as a negative number

Timeliness

Stock prices risen, but displaying low on front-end

Missing values

Feedback forms given to students from instructor

Recap

INSAID

Introduction to Data Science

What is Data?

What is Data Science?

Industry applications

Problems solved by Data Science

Project Life Cycle

Data Types

Conclusion