Welcome to Session 2 of DS Foundation Course!

Recap of Day 1

- ‘Data’ is the most valuable currency – Amazon example
- ‘Data Science’ is more art than science – Target example
- Data Science is all about understanding the ‘Why’ of Data – Your Household Expenses example
- Data Scientist is someone who has a basic understanding of different disciplines – Elon Musk example
- Problems that Data Science solve

TIME FOR
TRIVIA

Take a moment to think about your industry.

Let us see if we can find a domain where there are no Data Science Use Cases!

Industry Applications

Applications in Telecom Industry

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

Applications in Banking and Finance

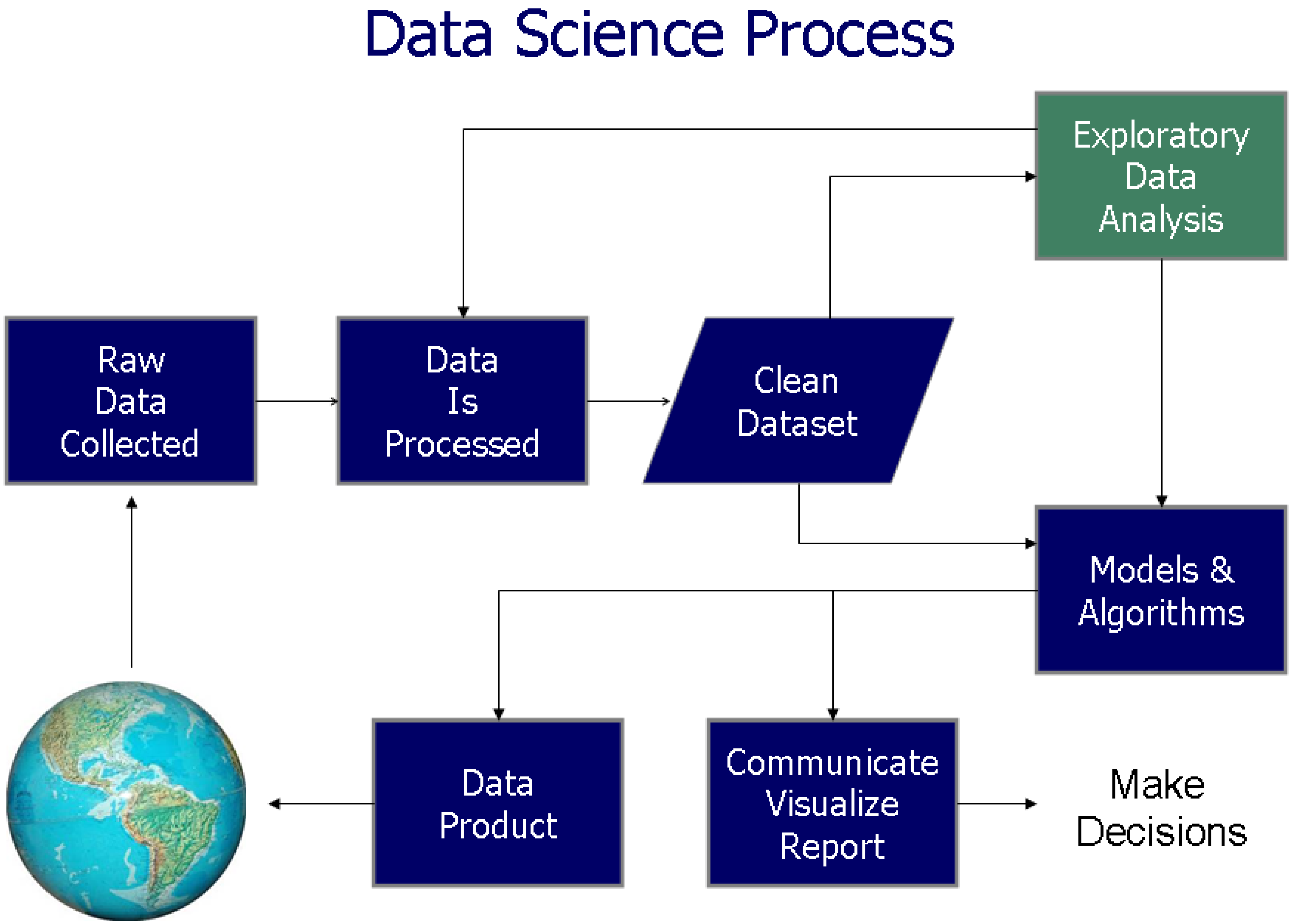
- Fraud detection and prevention
- Customer Segmentation
- Risk management
- Portfolio Optimization

Applications in Manufacturing

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

Data Science Project Life Cycle

How do I start a Data Science project?



Step 1: Collect Raw Data

To solve a given problem, as a data scientist you need data

Sometimes your organization may already be collecting data that you need

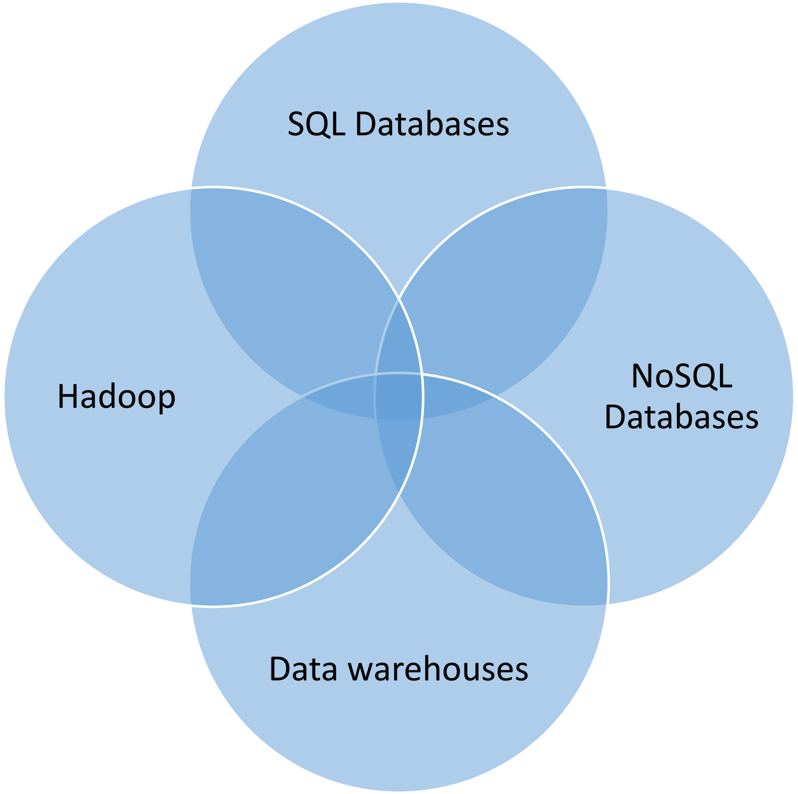
Sometimes it may not be collecting data that you need.

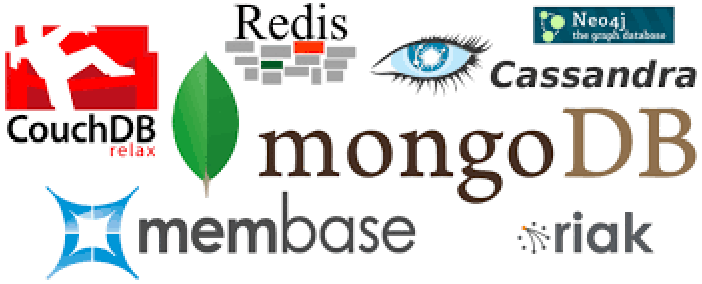
In that case, you will need to work with Data engineers, data infrastructure teams to build or modify systems to start collecting such data

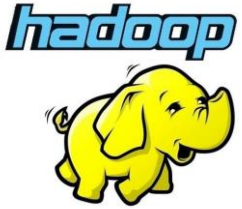
Step 2: Store Raw Data

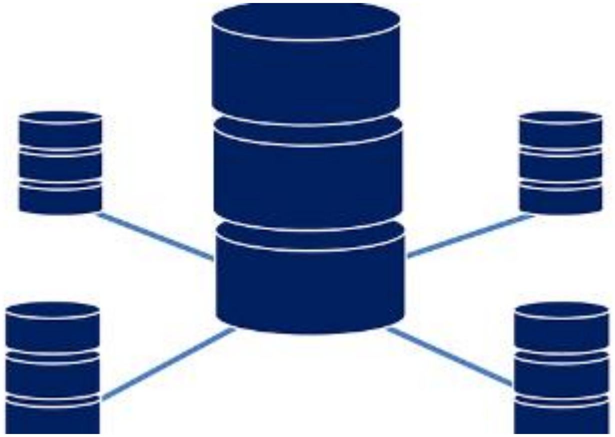
Raw data means data that has not been changed since acquisition

This raw Data is stored in your storage systems.









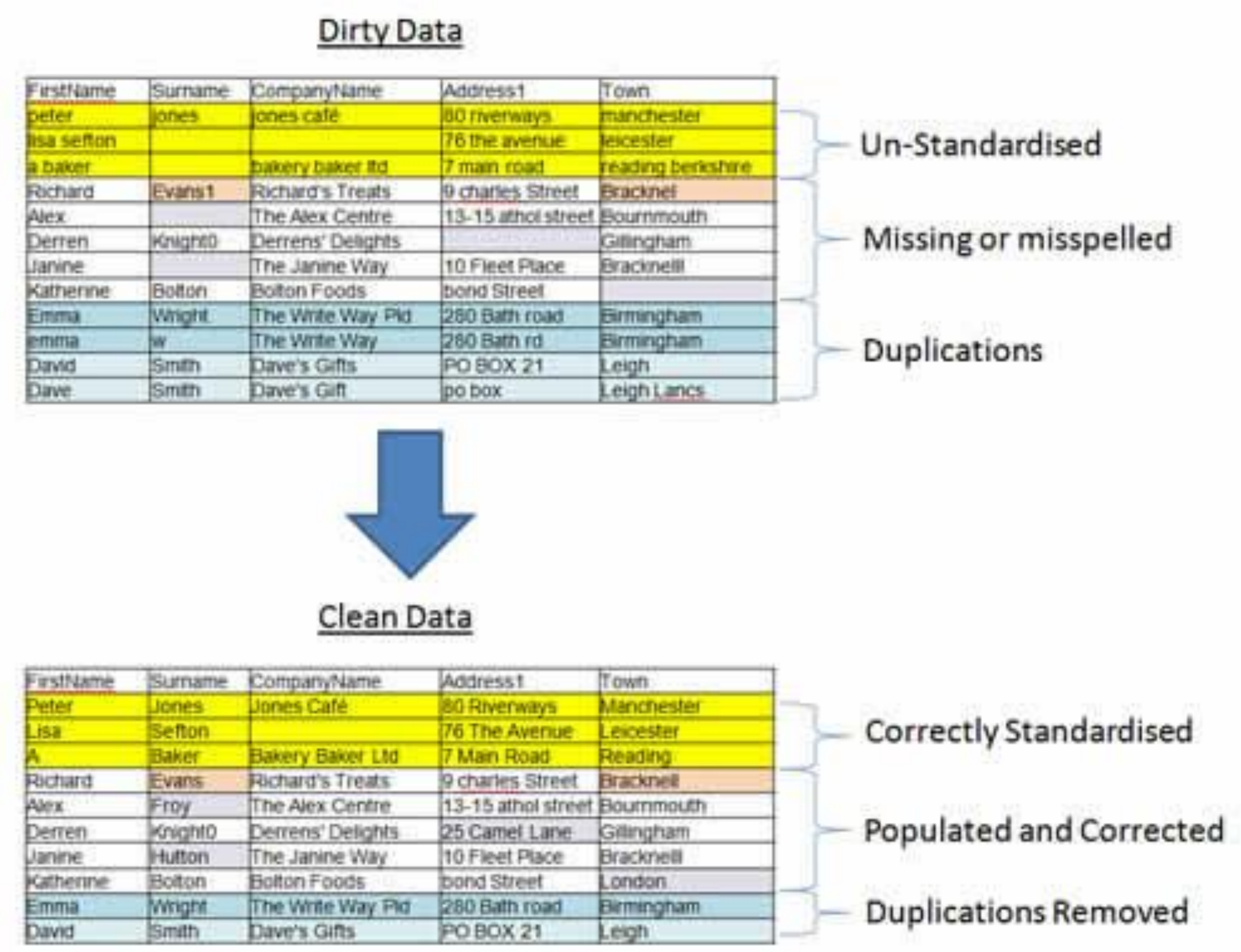
Step 3: Data Pre-Processing

As a Data Scientist, a lot of your time will go in Data pre-processing. Also known as Data cleaning.

This step includes

- Removing outliers
- Replacing missing data
- Malicious Data
- Erroneous Data
- Irrelevant Data
- Inconsistent Data
- Formatting

Step 3: Data Pre-Processing – contd..



Dirty Data

FirstName	Surname	CompanyName	Address1	Town
peter	jones	jones café	80 riverways	manchester
lisa sefton			76 the avenue	leicester
a baker		bakery baker ltd	7 main road	reading berkshire
Richard	Evans1	Richard's Treats	9 charles Street	Bracknel
Alex		The Alex Centre	13-15 athol street	Bournmouth
Derren	Knight0	Derrens' Delights		Gillingham
Janine		The Janine Way	10 Fleet Place	Bracknelll
Katherine	Bolton	Bolton Foods	bond Street	
Emma	Wright	The Write Way Pld	280 Bath road	Birmingham
emma	w	The Write Way	280 Bath rd	Birmingham
David	Smith	Dave's Gifts	PO BOX 21	Leigh
Dave	Smith	Dave's Gift	po box	Leigh Lancs

Un-Standardised

Missing or misspelled

Duplications

Clean Data

FirstName	Surname	CompanyName	Address1	Town
Peter	Jones	Jones Café	80 Riverways	Manchester
Lisa	Sefton		76 The Avenue	Leicester
A	Baker	Bakery Baker Ltd	7 Main Road	Reading
Richard	Evans	Richard's Treats	9 charles Street	Bracknell
Alex	Froy	The Alex Centre	13-15 athol street	Bournmouth
Derren	Knight0	Derrens' Delights	25 Camel Lane	Gillingham
Janine	Hutton	The Janine Way	10 Fleet Place	Bracknell
Katherine	Bolton	Bolton Foods	bond Street	London
Emma	Wright	The Write Way Pld	280 Bath road	Birmingham
David	Smith	Dave's Gifts	PO BOX 21	Leigh

Correctly Standardised

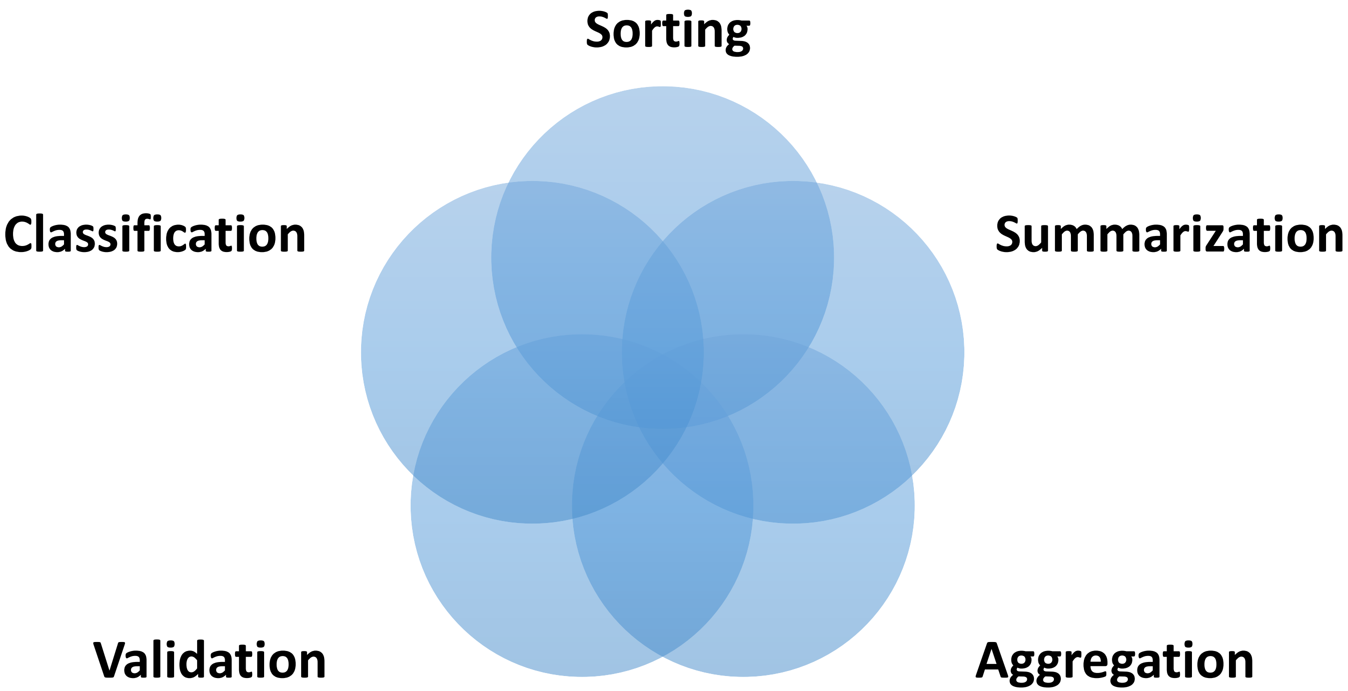
Populated and Corrected

Duplications Removed

Step 3: Data Pre-Processing - contd..

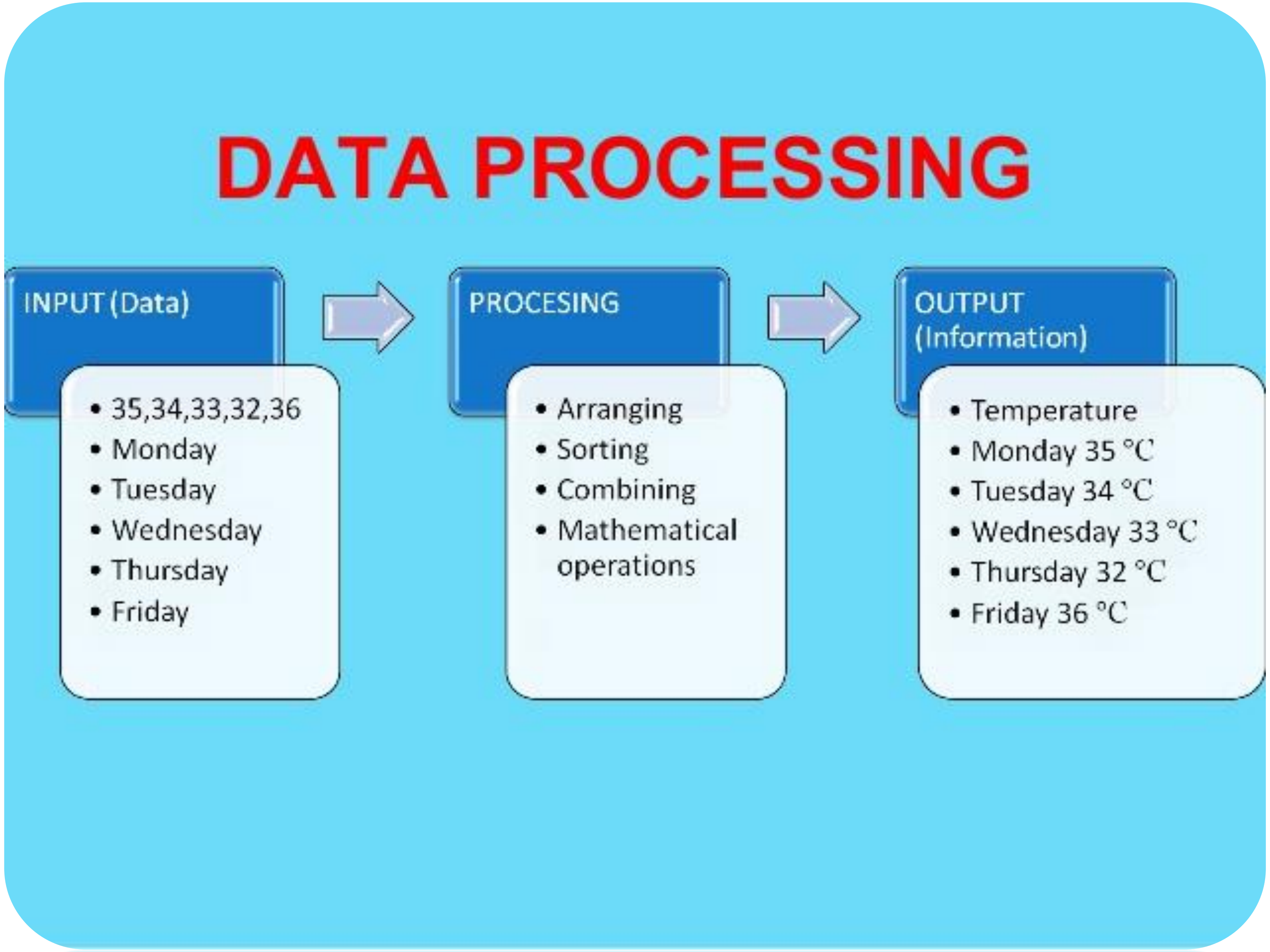
Once Data is cleaned, it needs to be processed to make it ready for use.

- This stage includes



- Data Pre-processing(Data cleaning) is at times considered to be part of Data Processing

Step 3: Data Pre-Processing - contd..



DATA PROCESSING

INPUT (Data)

- 35,34,33,32,36
- Monday
- Tuesday
- Wednesday
- Thursday
- Friday

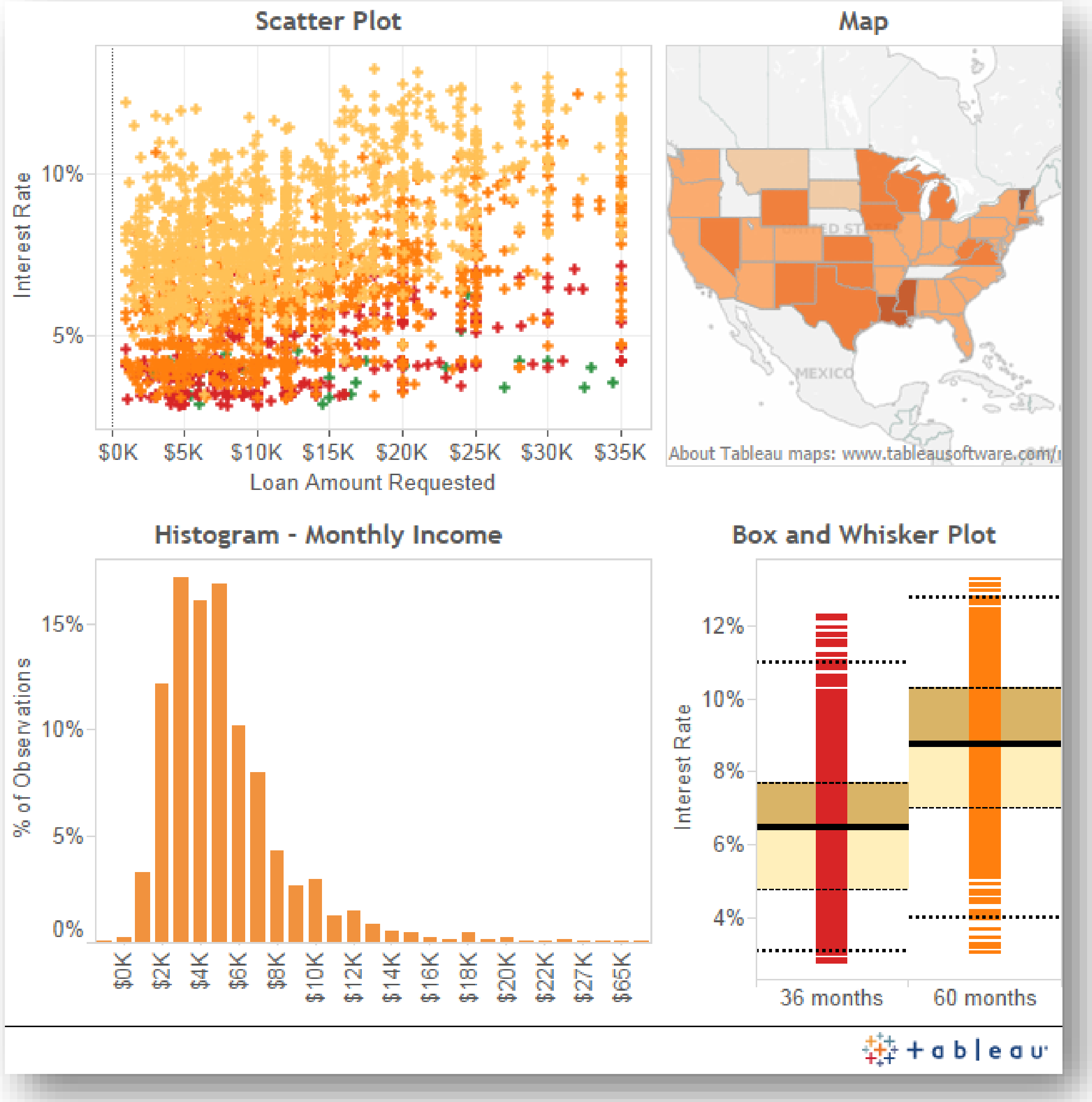
PROCESING

- Arranging
- Sorting
- Combining
- Mathematical operations

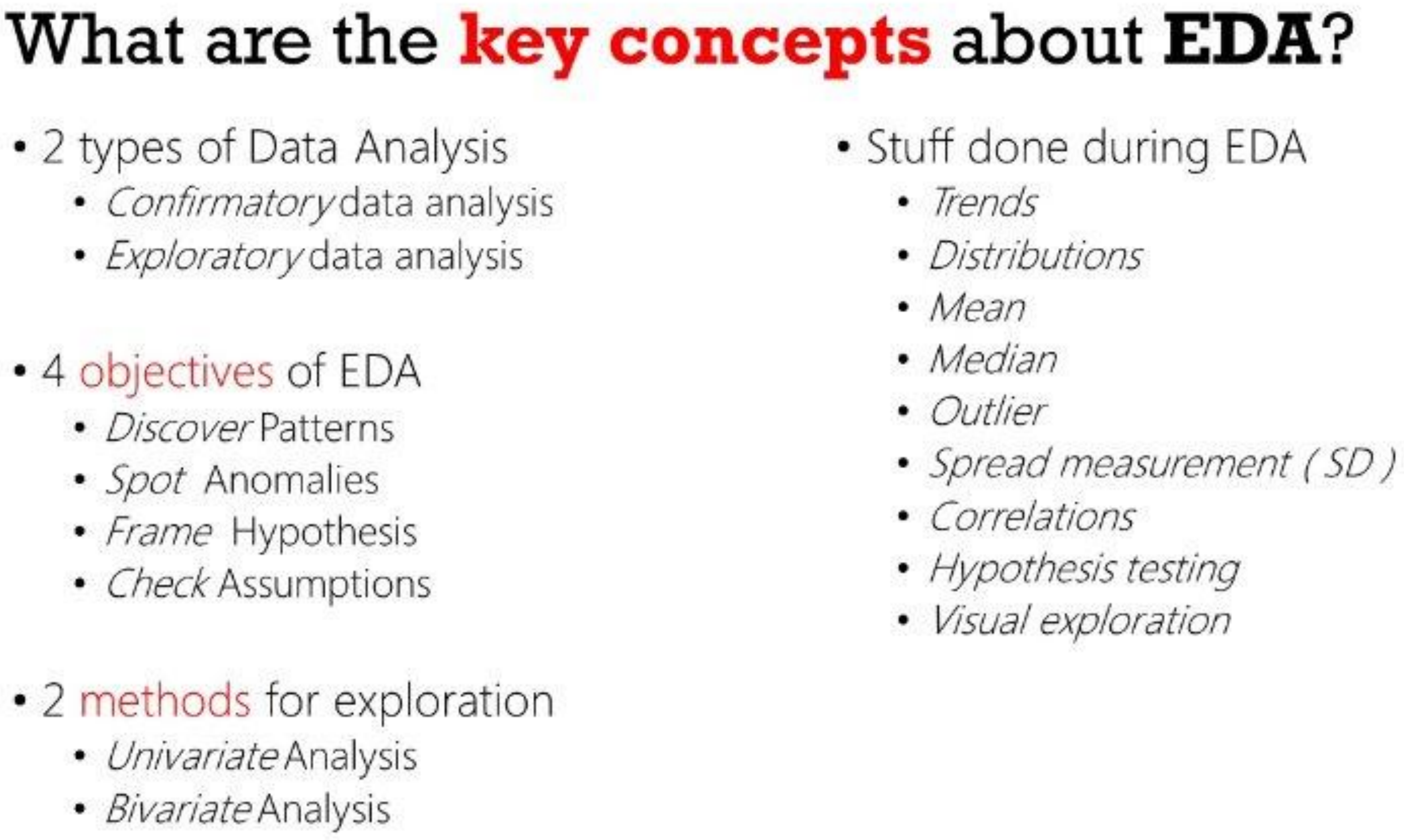
OUTPUT (Information)

- Temperature
- Monday 35 °C
- Tuesday 34 °C
- Wednesday 33 °C
- Thursday 32 °C
- Friday 36 °C

Step 4: Exploratory Data Analysis



Step 4: Exploratory Data Analysis – contd..



What are the **key concepts** about **EDA**?

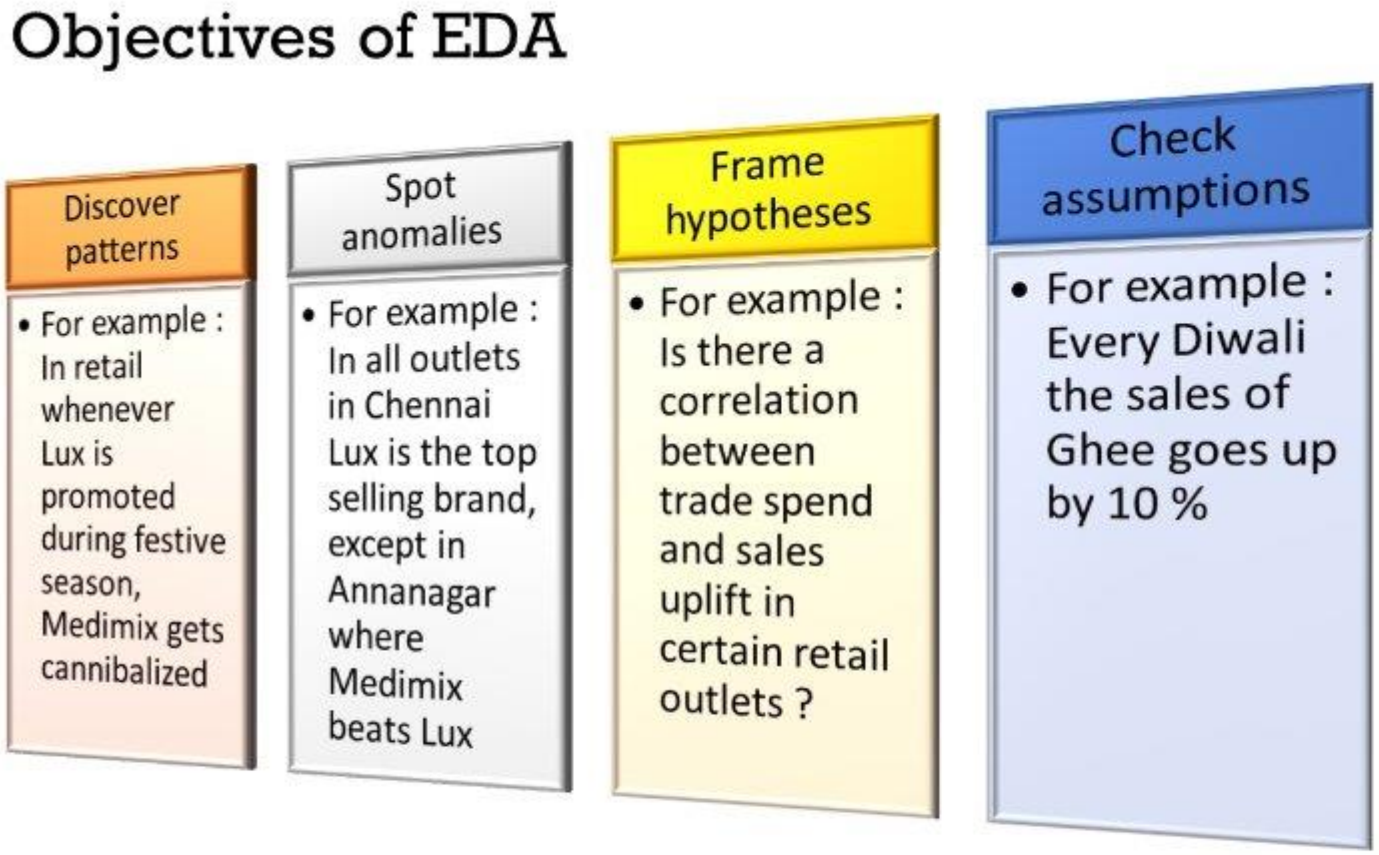
- 2 types of Data Analysis
  - *Confirmatory* data analysis
  - *Exploratory* data analysis

- 4 objectives of EDA
  - *Discover* Patterns
  - *Spot* Anomalies
  - *Frame* Hypothesis
  - *Check* Assumptions

- 2 methods for exploration
  - *Univariate* Analysis
  - *Bivariate* Analysis

- Stuff done during EDA
  - *Trends*
  - *Distributions*
  - *Mean*
  - *Median*
  - *Outlier*
  - *Spread measurement (SD)*
  - *Correlations*
  - *Hypothesis testing*
  - *Visual exploration*

Step 4: Exploratory Data Analysis – contd..



Objectives of EDA

Discover patterns

- For example : In retail whenever Lux is promoted during festive season, Medimix gets cannibalized

Spot anomalies

- For example : In all outlets in Chennai Lux is the top selling brand, except in Annanagar where Medimix beats Lux

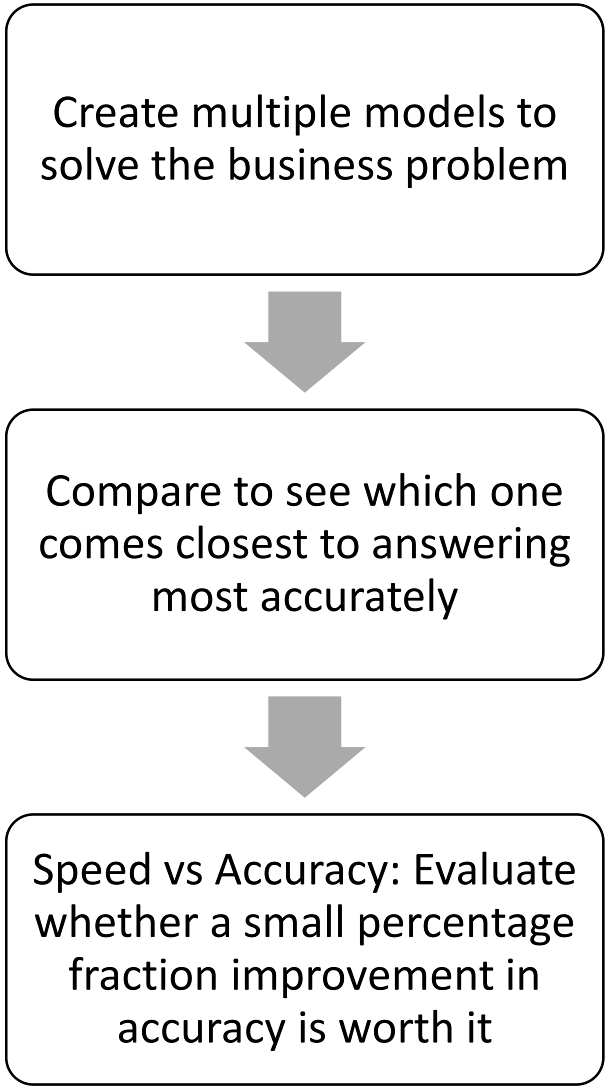
Frame hypotheses

- For example : Is there a correlation between trade spend and sales uplift in certain retail outlets ?

Check assumptions

- For example : Every Diwali the sales of Ghee goes up by 10 %

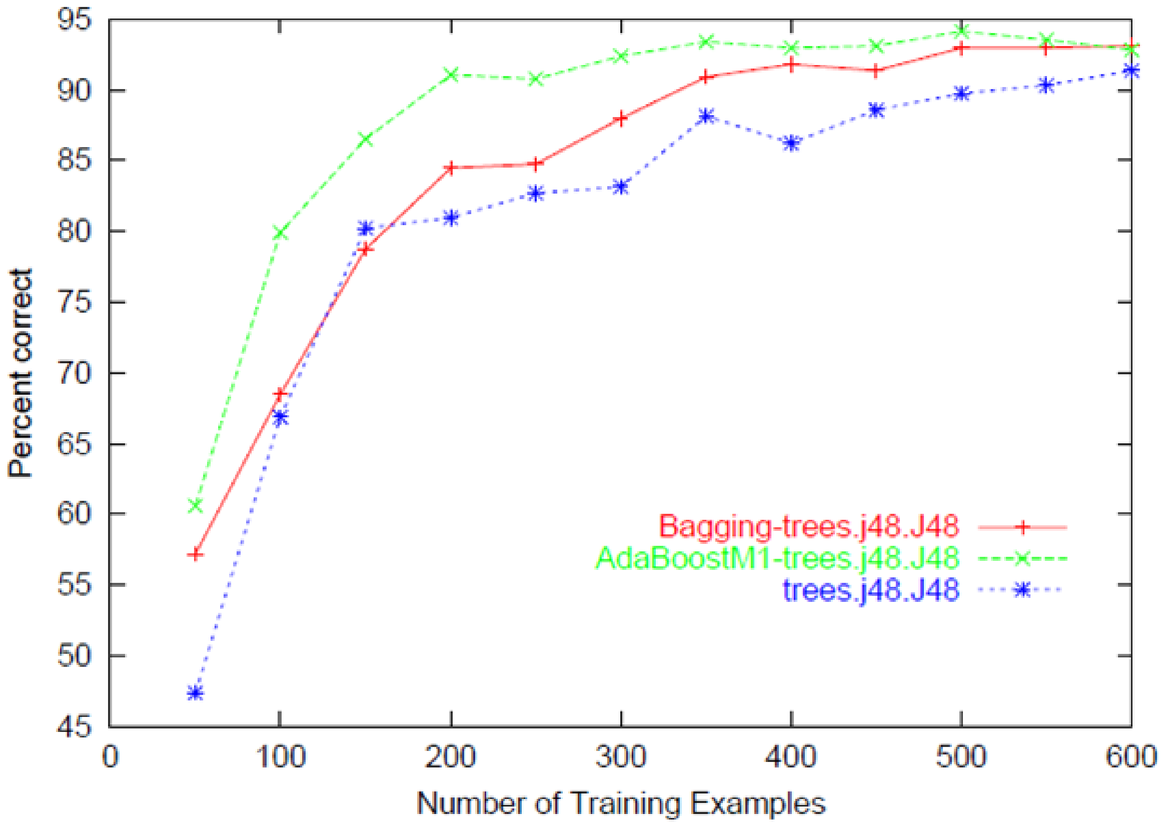
Step 5: Models & Algorithms



Create multiple models to solve the business problem

Compare to see which one comes closest to answering most accurately

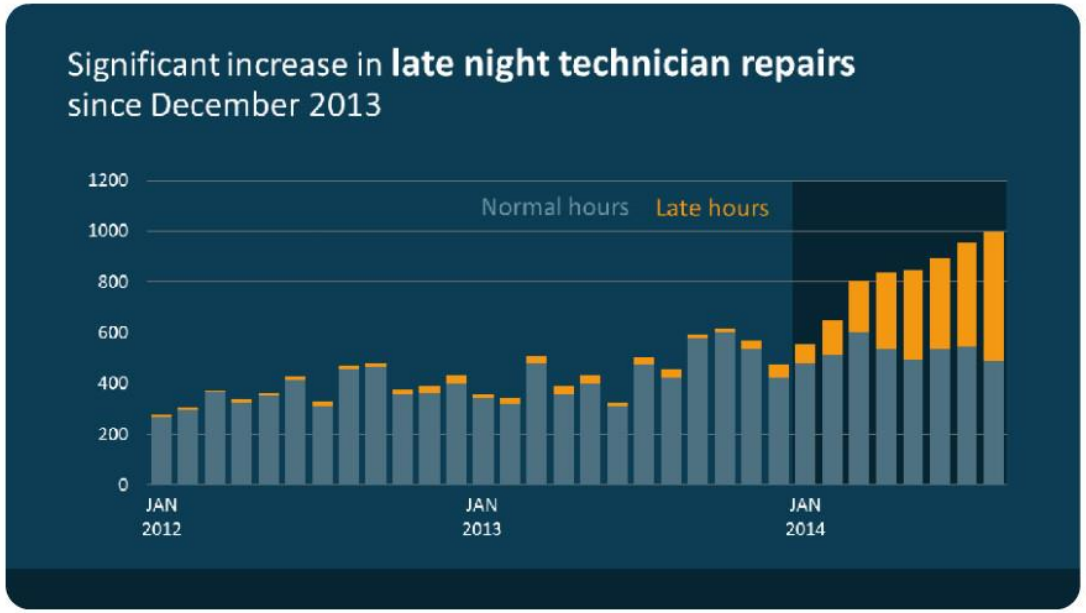
Speed vs Accuracy: Evaluate whether a small percentage fraction improvement in accuracy is worth it



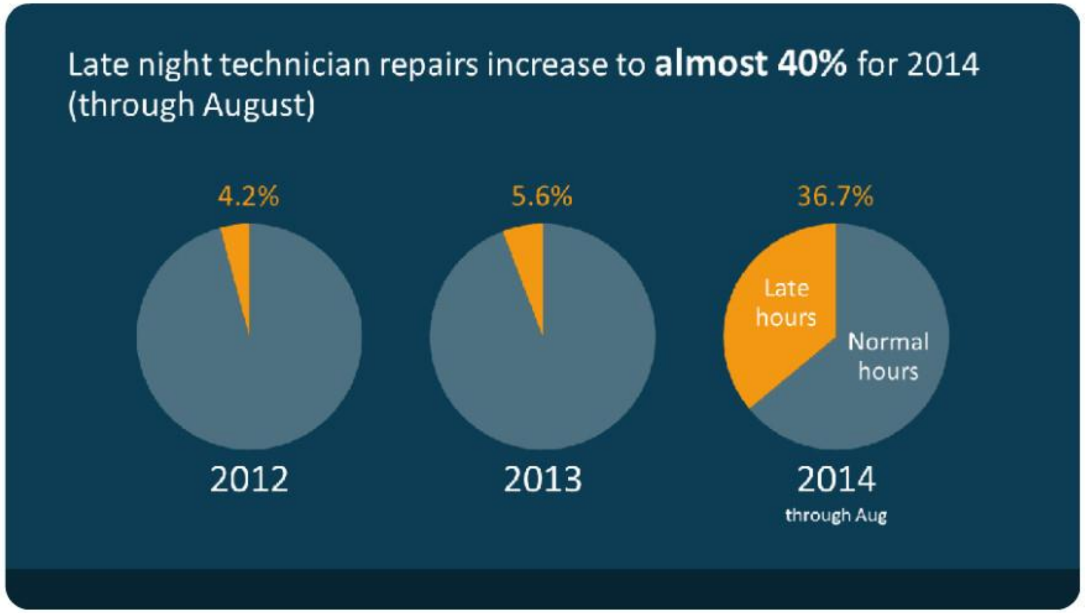
Step 6: Communicate visualize & report

- Brainstorm with management and showcase the benefit the analysis and models bring to the plate.

- Seek management’s consideration for deploying the solution to real world to help make the business more optimized and beneficial.

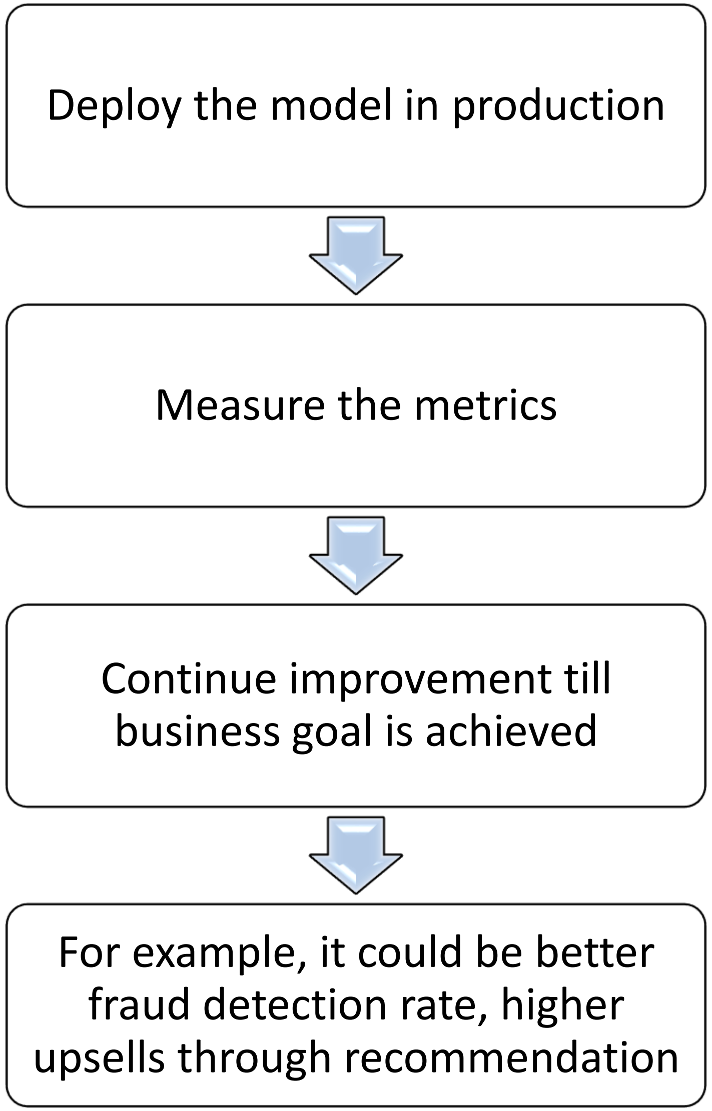


Significant increase in **late night technician repairs** since December 2013



Late night technician repairs increase to **almost 40%** for 2014 (through August)

Step 7: Take action & deploy the findings in real world

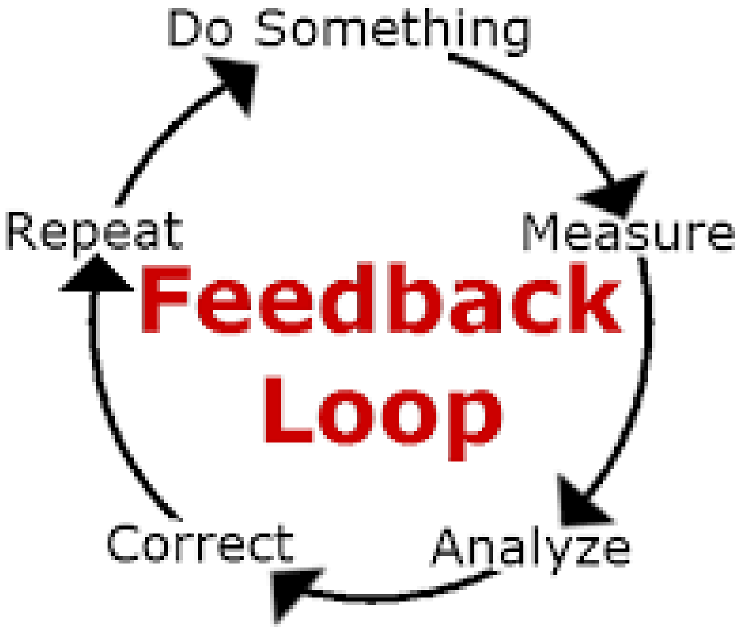


Deploy the model in production

Measure the metrics

Continue improvement till business goal is achieved

For example, it could be better fraud detection rate, higher upsells through recommendation



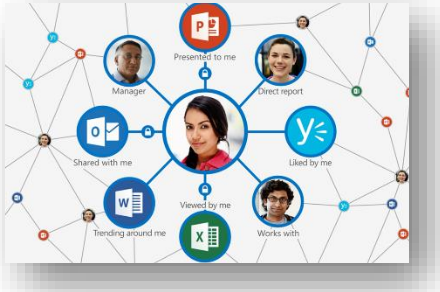
More on Data

Types of Data

Structured data

Semi-structured data

Unstructured data



Graph data

Streaming data

Types of Data: Alternate view

Quantitative

a. Discrete

b.Continuous

Two horses

Height

Qualitative

a. Nominal

b. Ordinal

Male

Female

Customer Service

Interval

Time scale

Ratio

Weight

Data Quality Issues

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

Introduction to Data Science

What is Data?

What is Data Science?

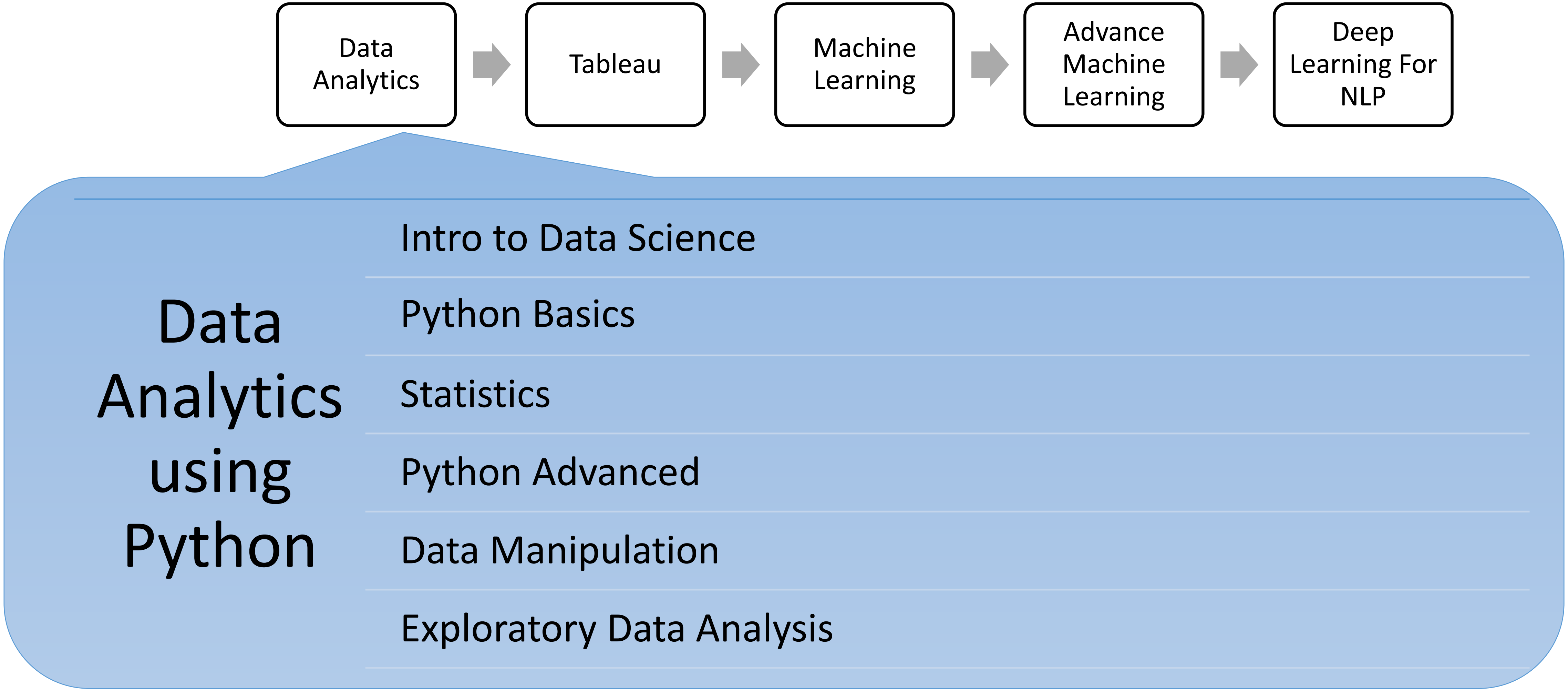
Industry applications

Problems solved by Data Science

Project Lifecycle

Data Types

Data Science Preview



Data Analytics → Tableau → Machine Learning → Advance Machine Learning → Deep Learning For NLP

Data Analytics using Python

- Intro to Data Science
- Python Basics
- Statistics
- Python Advanced
- Data Manipulation
- Exploratory Data Analysis