Department of Computer Systems Engineering
Mehran University of Engineering and Technology, Jamshoro

Course: Big Data Analytics

Dr. Sanam Narejo

Big Data

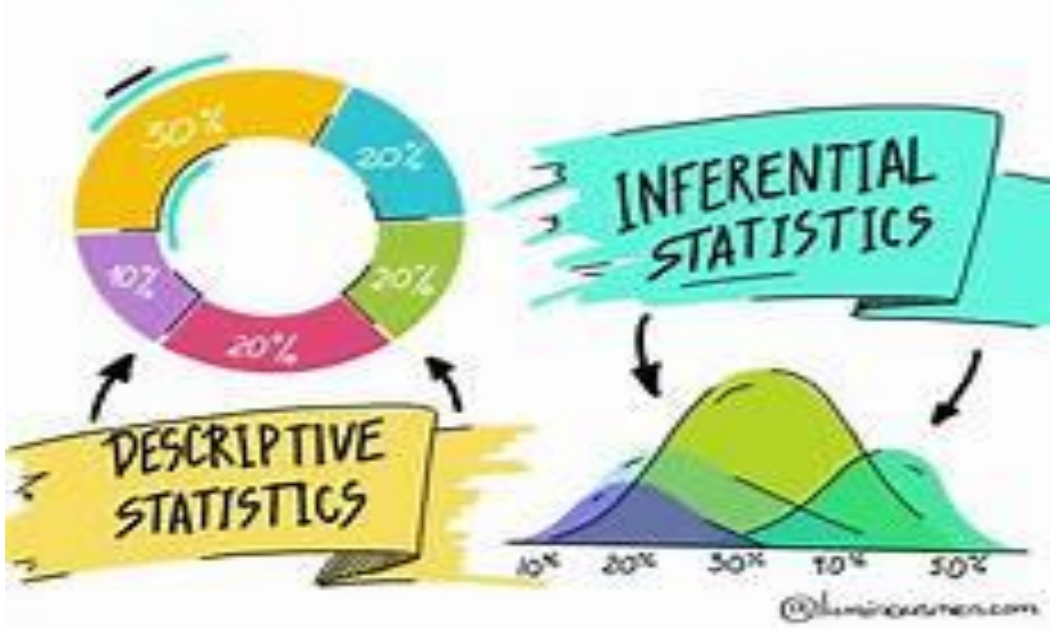
- Big Data is today, the hottest buzzword around, and with the amount of data being generated every minute by consumers, or/and businesses worldwide, there is huge value to be found in Big Data analytics.
- Big Data is a massive amount of data sets that cannot be stored, processed, or analyzed using traditional tools.

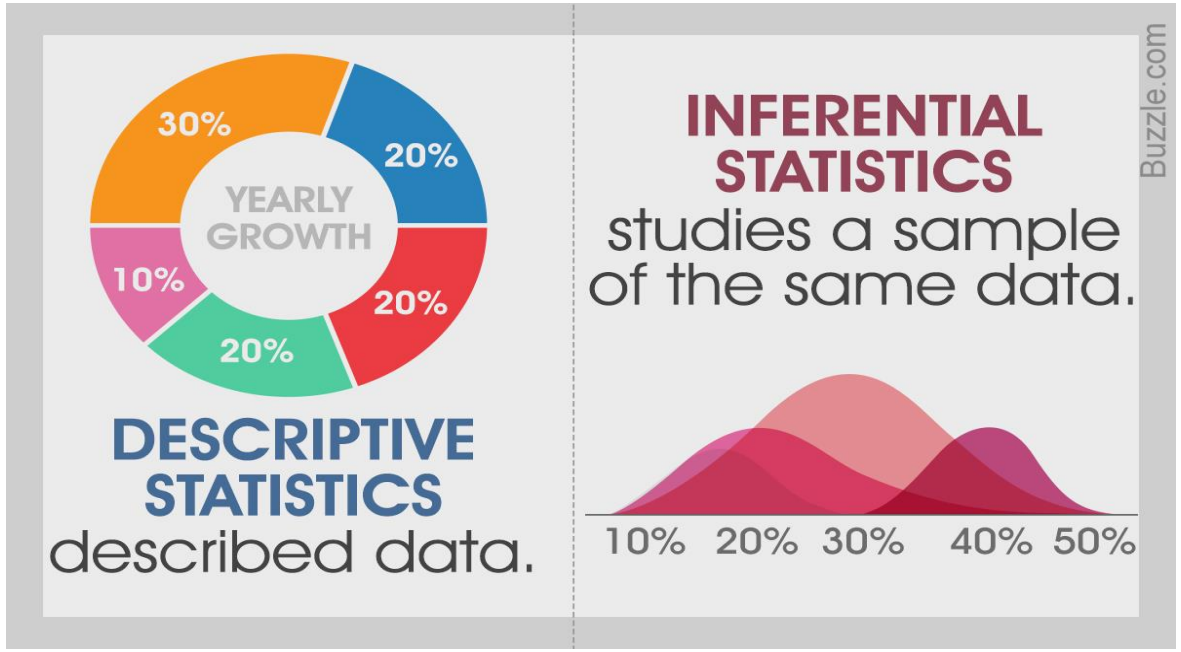
Big Data are high volume, high velocity, or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization

Big Data Analytics

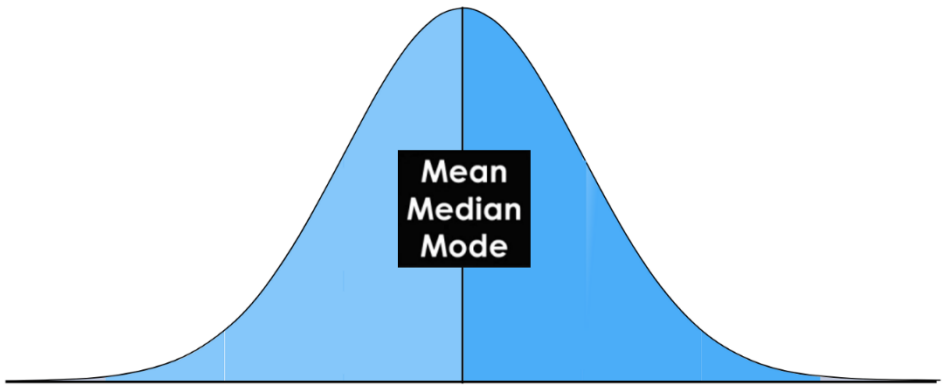
Big Data analytics is a process used to extract meaningful insights, such as hidden patterns, unknown correlations, market trends, and customer preferences. Big Data analytics provides various advantages—it can be used for better decision making, preventing fraudulent activities, among other things.

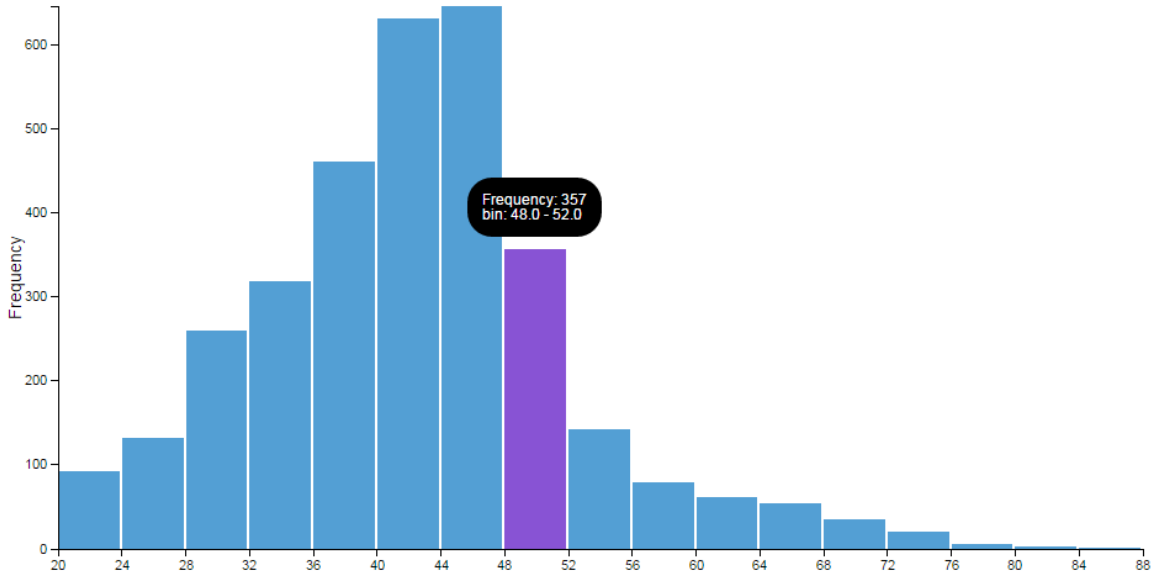
- The principal purpose of Data Science is to find patterns within data. It uses **various statistical techniques** to analyze and draw insights from the data.
- From data extraction, wrangling and pre-processing, a Data Scientist must scrutinize the data thoroughly.

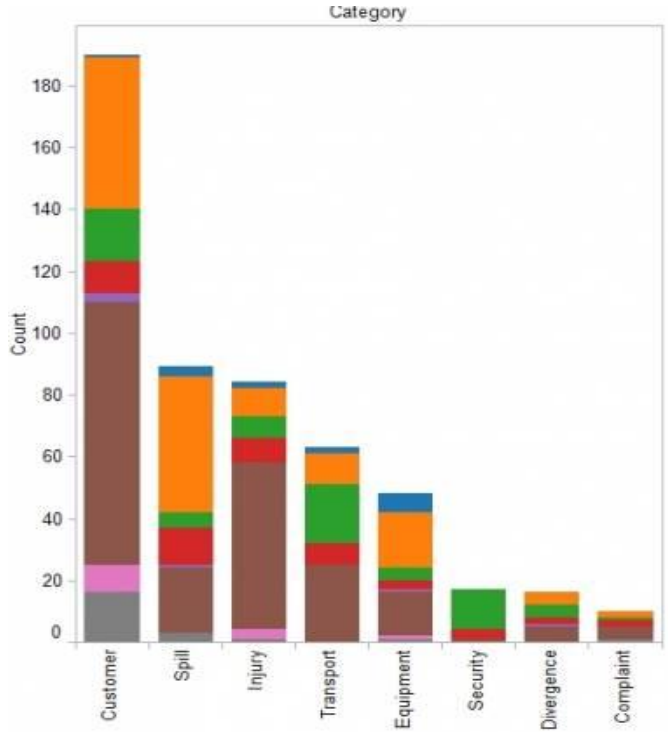


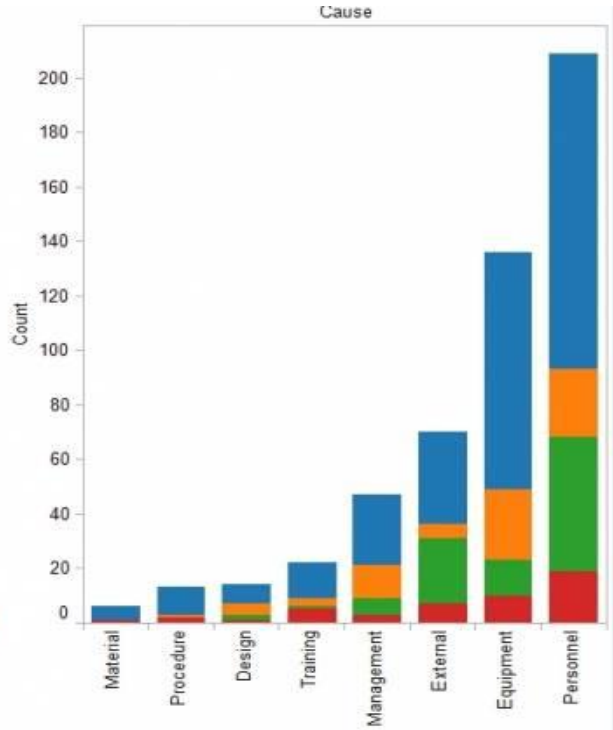


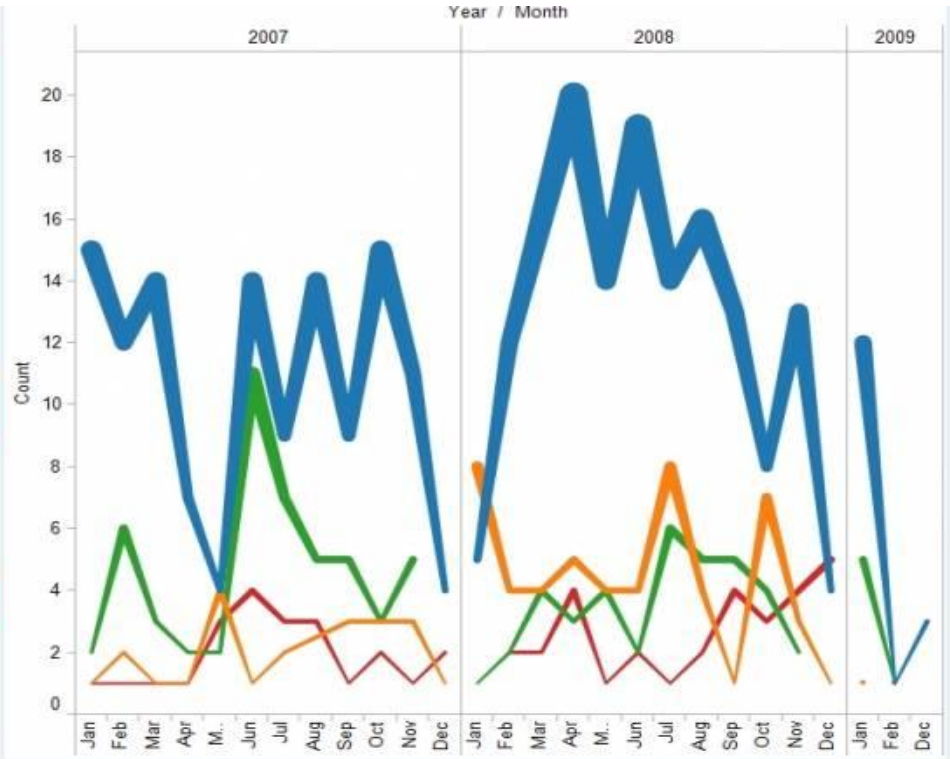
- Probability and Statistics form the basis of Data Science.
- The probability theory is very much helpful for making the prediction. Estimates and predictions form an important part of Data science. With the help of statistical methods, we make estimates for the further analysis.
- Thus, statistical methods are largely dependent on the theory of probability. And all of probability and statistics is dependent on Data.

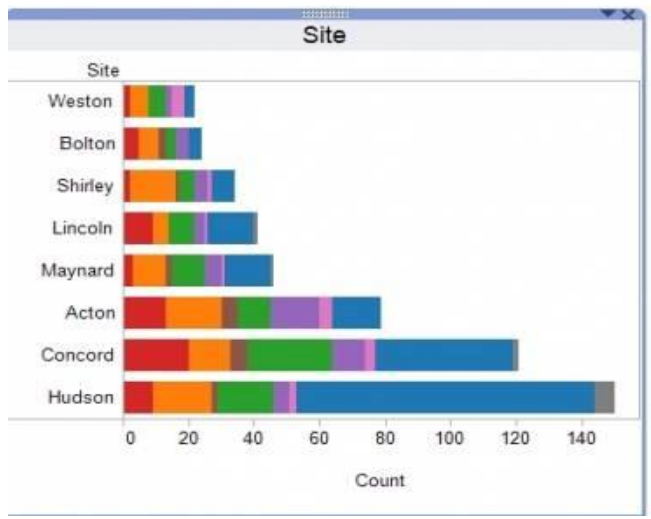


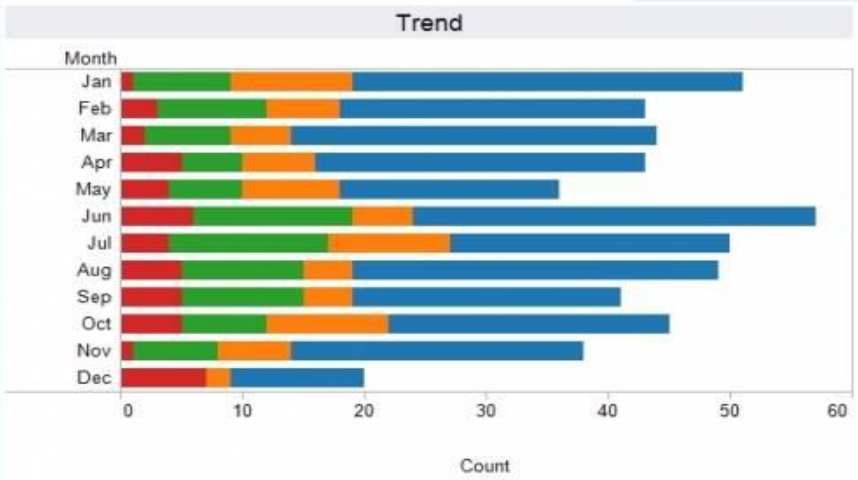


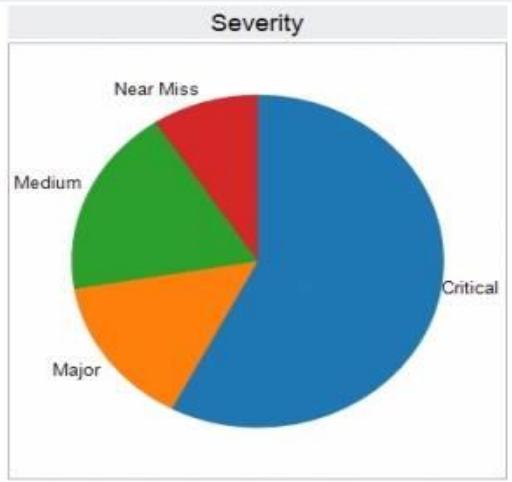

Category
Count
0
20
40
60
80
100
120
140
160
180
Customer
Spill
Injury
Transport
Equipment
Security
Divergence
Complaint


Cause
Count
0
20
40
60
80
100
120
140
160
180
200
Material
Procedure
Design
Training
Management
External
Equipment
Personnel


Year / Month
2007
2008
2009
Count
0
2
4
6
8
10
12
14
16
18
20
Jan
Feb
Mar
Apr
M..
Jun
Jul
Aug
Sep
Oct
Nov
Dec


Site
Site
Weston
Bolton
Shirley
Lincoln
Maynard
Acton
Concord
Hudson
0
20
40
60
80
100
120
140
Count


Trend
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
0
10
20
30
40
50
60
Count


Severity
Near Miss
Medium
Major
Critical


Cause
Design
Equipment
External
Managem..
Material
Personnel
Procedure
Training
Severity
Critical
Major
Medium
Near Miss
Count
517.0

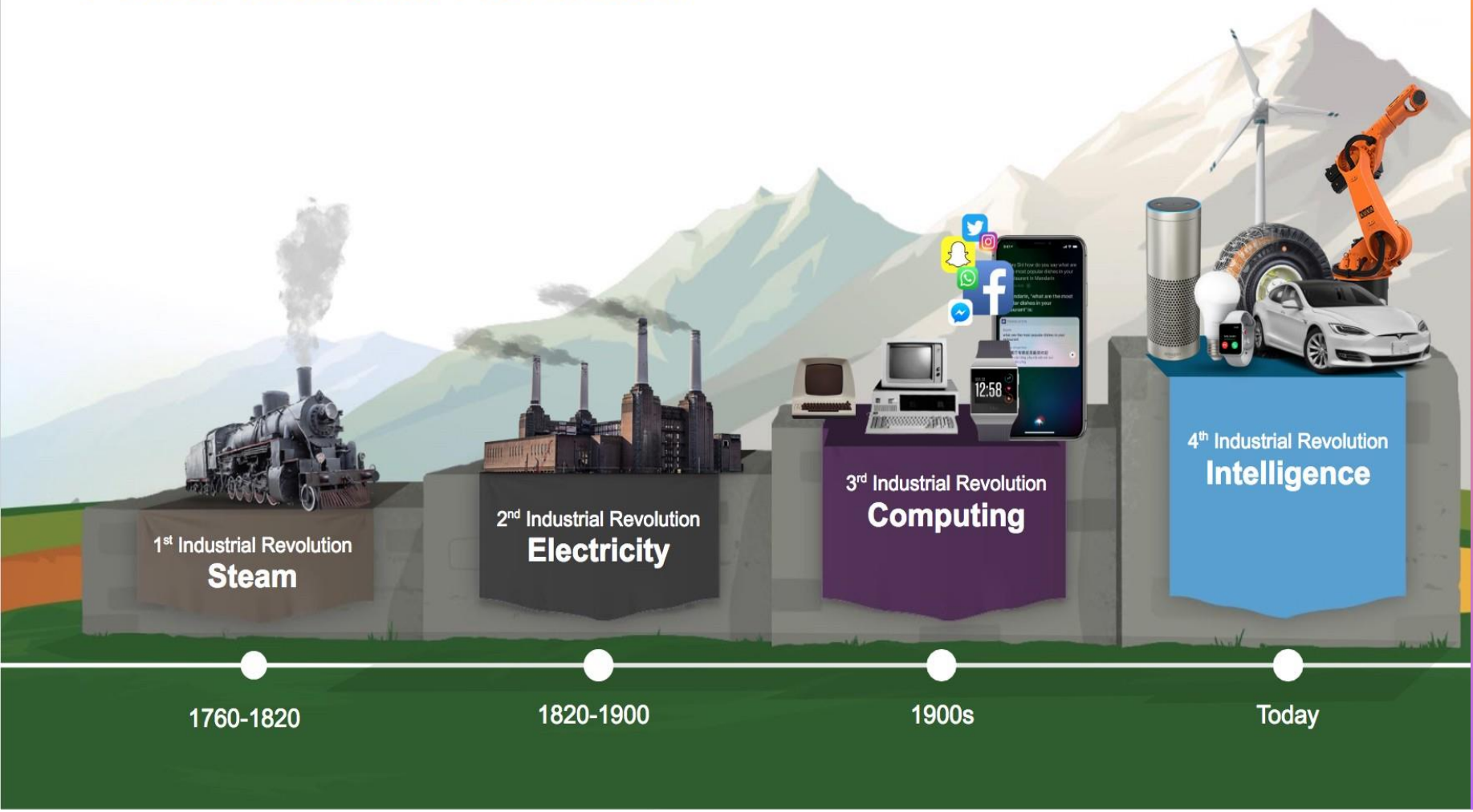
Data is the collected information(observations) we have about something or facts and statistics collected together for reference or analysis.

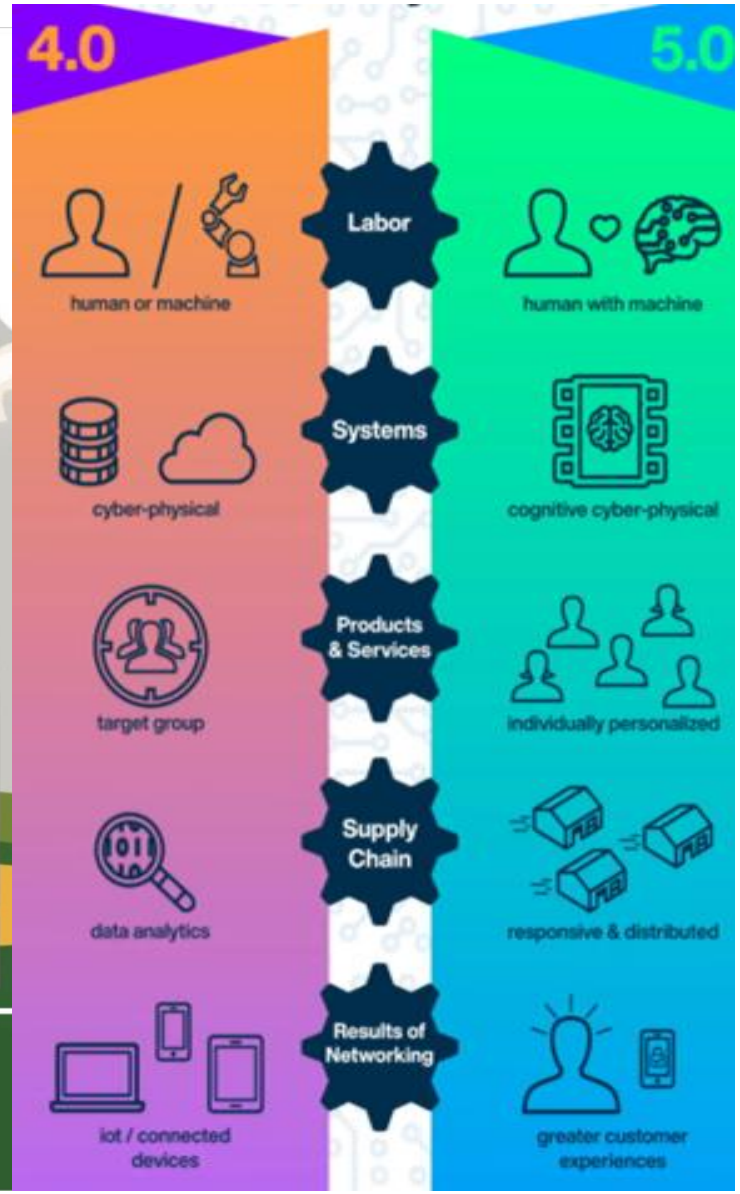
Data — a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process

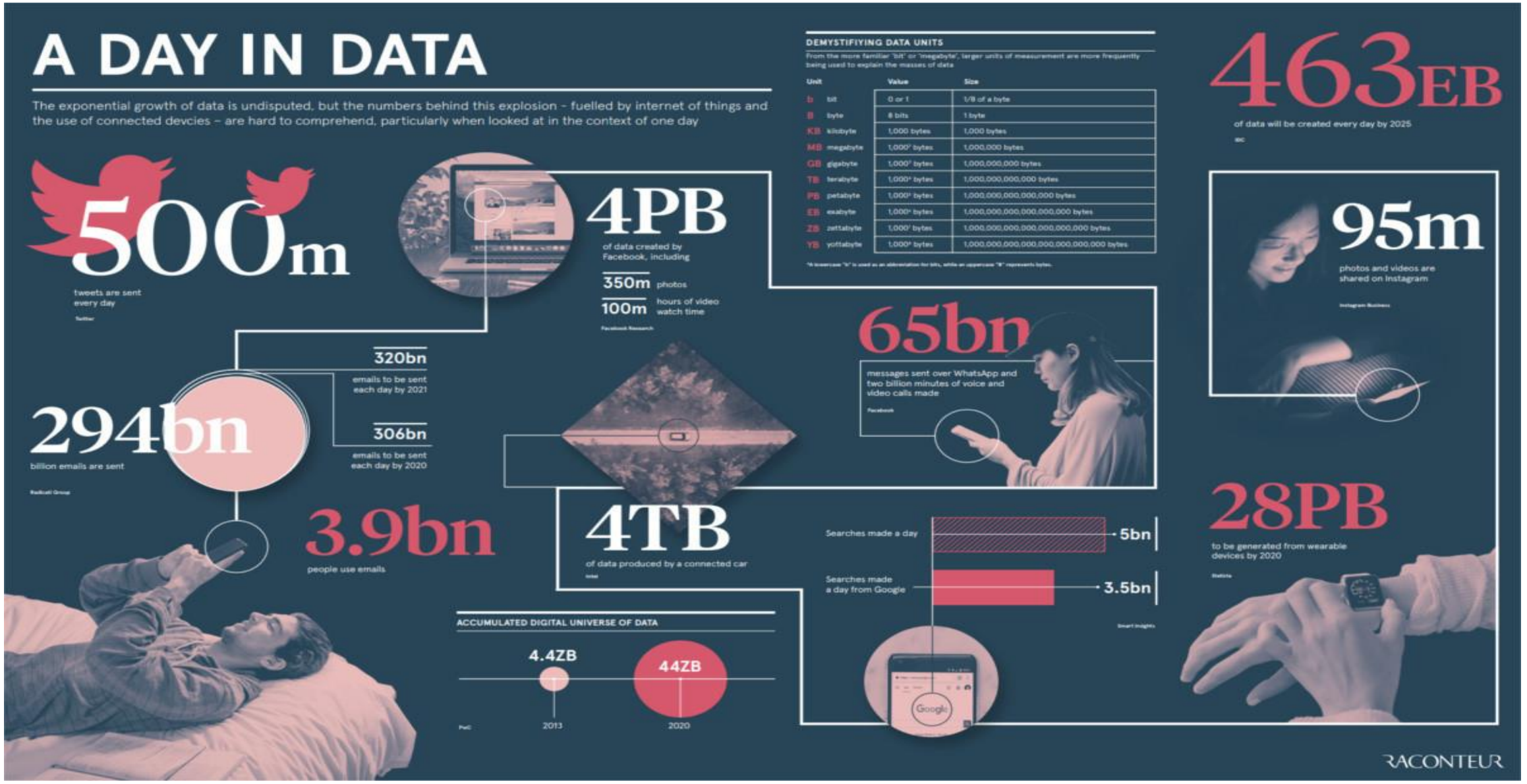
Why Data Matters ?

INDUSTRY 5.0 is Future

Fourth Industrial Revolution







The need for data has risen tremendously in the last decade. Many companies have centered their business on data. Data has created new sectors in the IT industry. However,

Here are some key daily statistics highlighted in the infographic:

- 500 million tweets are sent
- 294 billion emails are sent
- 4 petabytes of data are created on Facebook
- 4 terabytes of data are created from each connected car
- 65 billion messages are sent on WhatsApp
- 5 billion searches are made

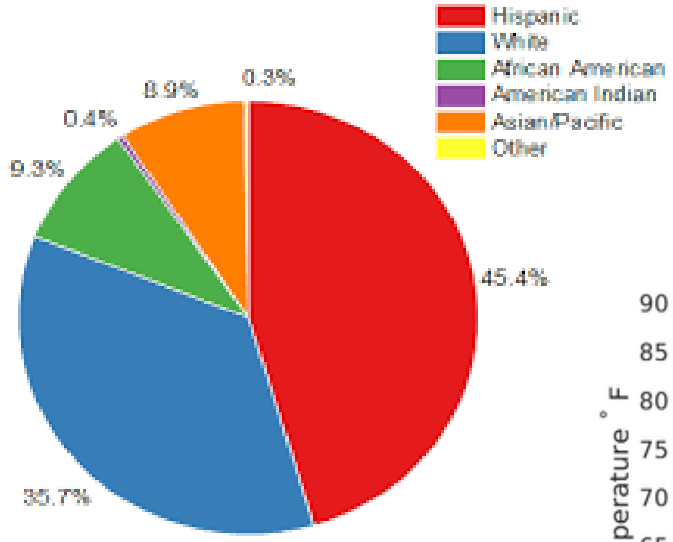
- By 2025, it’s estimated that 463 exabytes of data will be created each day globally – that’s the equivalent of 212,765,957 DVDs per day!

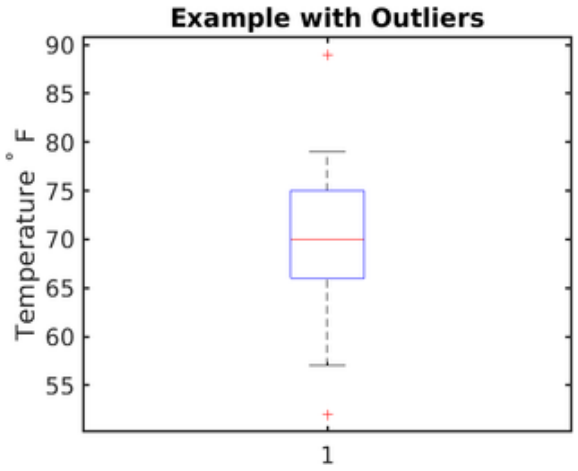
Generating Data – It's a huge!

Table 1: Data Measurement Units

Unit	Abbreviation	Decimal Value	Binary Value	Decimal Size
bit	b	0 or 1	0 or 1	1/8 of a byte
byte	B	8 bits	8 bits	1 byte
kilobyte	KB	$1{,}000^1$ bytes	$1{,}024^1$ bytes	1,000 bytes
megabyte	MB	$1{,}000^2$ bytes	$1{,}024^2$ bytes	1,000,000 bytes
gigabyte	GB	$1{,}000^3$ bytes	$1{,}024^3$ bytes	1,000,000,000 bytes
terabyte	TB	$1{,}000^4$ bytes	$1{,}024^4$ bytes	1,000,000,000,000 bytes
petabyte	PB	$1{,}000^5$ bytes	$1{,}024^5$ bytes	1,000,000,000,000,000 bytes
exabyte	EB	$1{,}000^6$ bytes	$1{,}024^6$ bytes	1,000,000,000,000,000,000 bytes
zettabyte	ZB	$1{,}000^7$ bytes	$1{,}024^7$ bytes	1,000,000,000,000,000,000,000 bytes
yottabyte	YB	$1{,}000^8$ bytes	$1{,}024^8$ bytes	1,000,000,000,000,000,000,000,000 bytes

- In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.
- A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.





Example:

"32" is split into "3" (stem) and "2" (leaf).

15, 16, 21, 23, 23, 26, 26, 30, 32, 41

Stem	Leaf
1	5 6
2	1 3 3 6 6
3	0 2
4	1

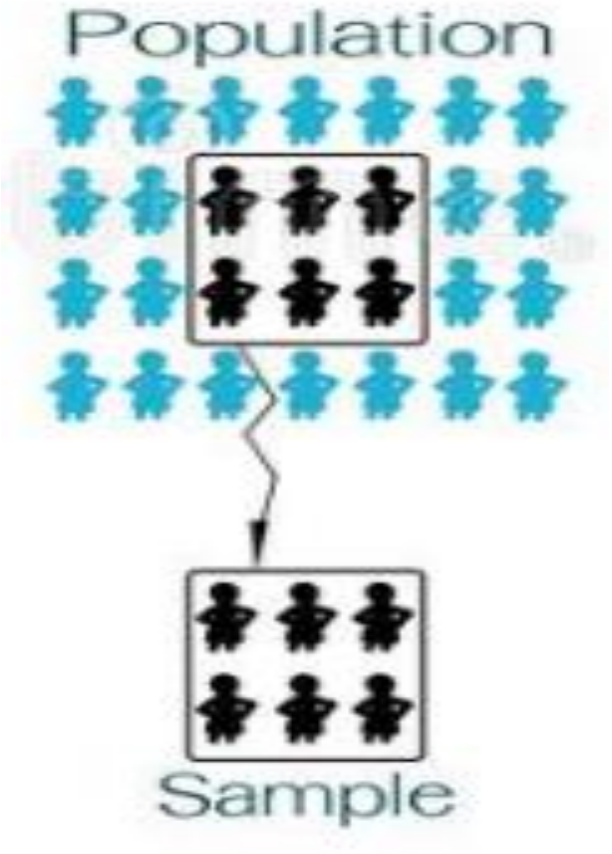
how to place "32"

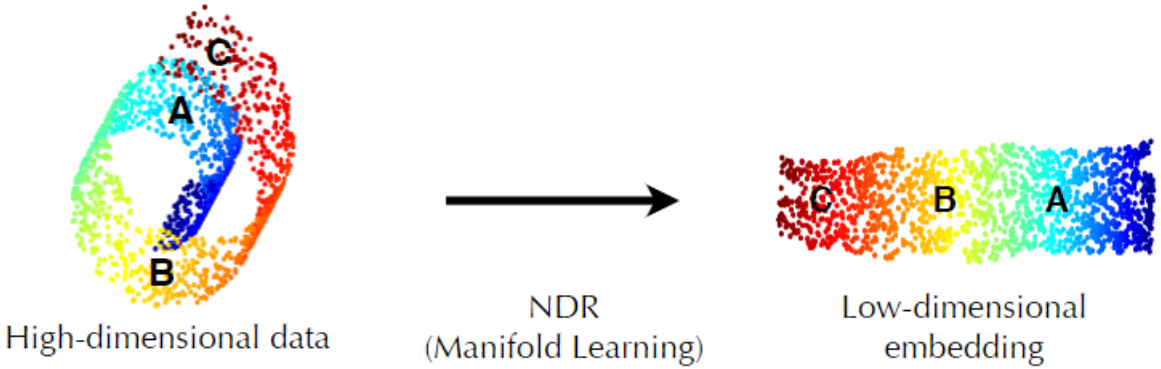
More Examples:

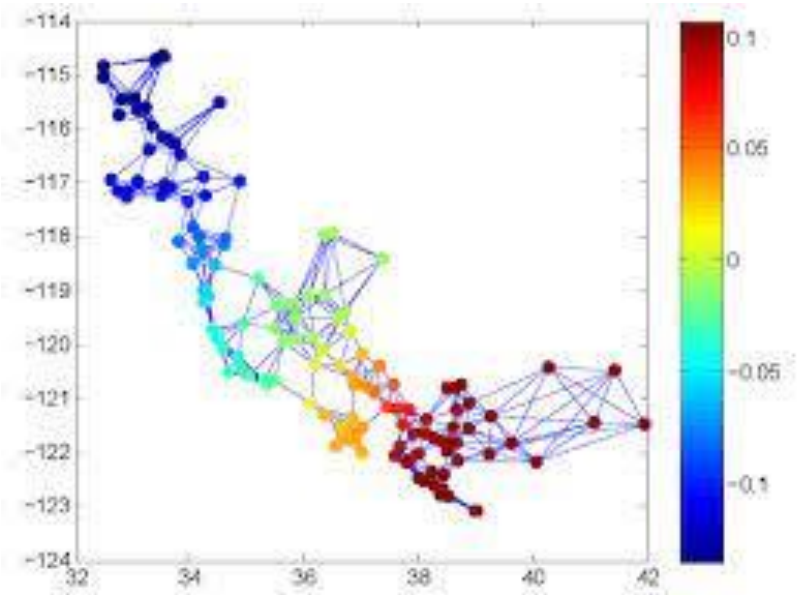
- Stem "1" Leaf "5" means **15**
- Stem "1" Leaf "6" means **16**
- Stem "2" Leaf "1" means **21**
- etc

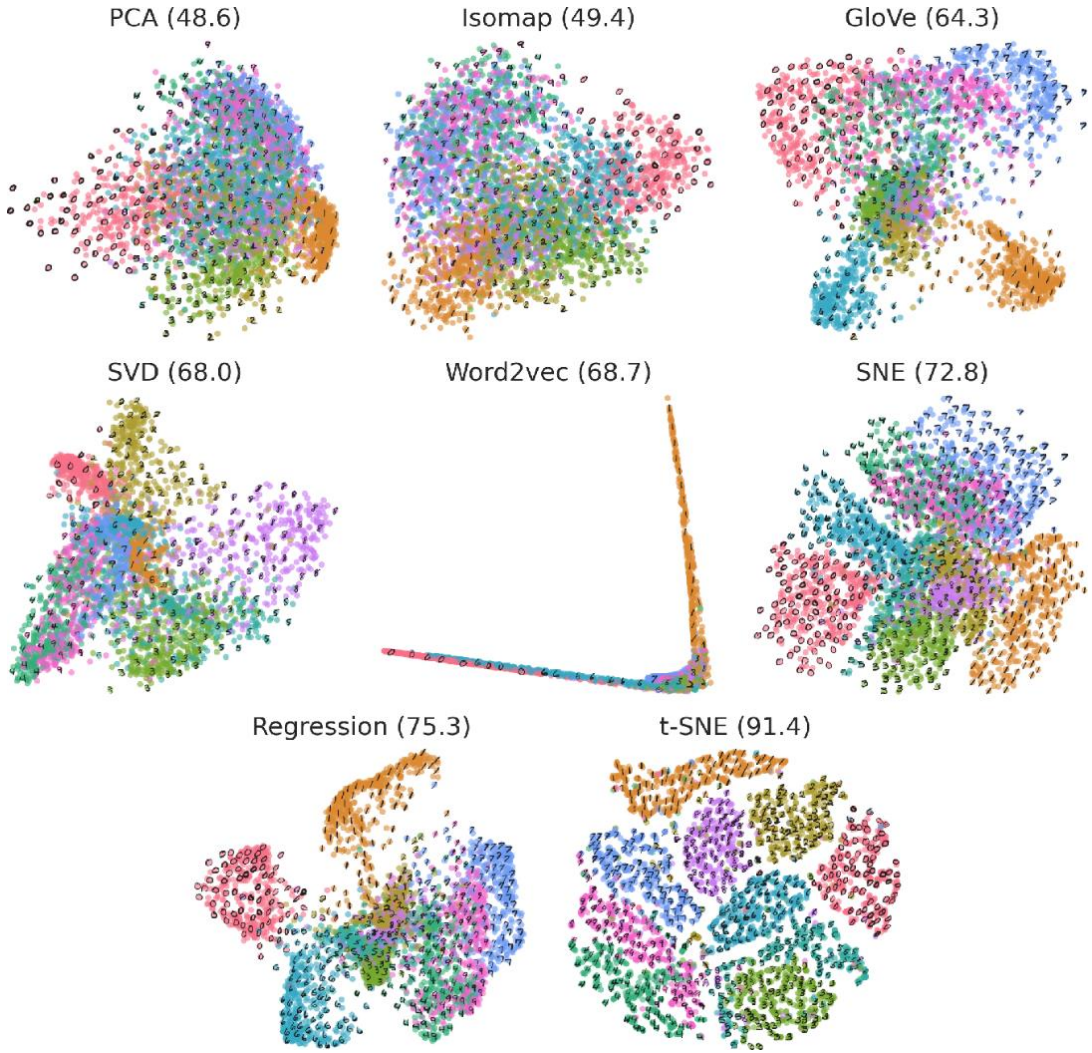
Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to maximize insight into a data set

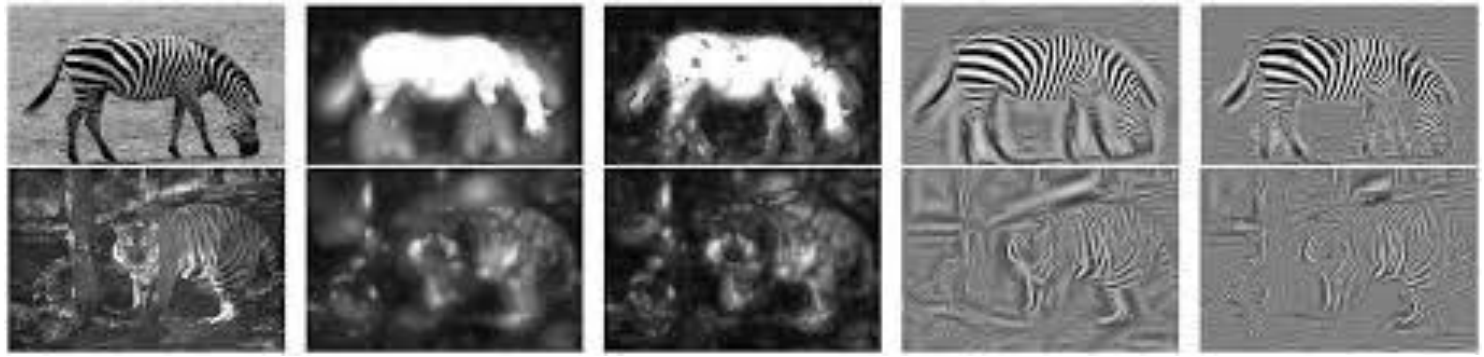
- uncover underlying structure
- extract important variables
- detect outliers and anomalies
- test underlying assumptions
- develop parsimonious models
- determine optimal factor settings.

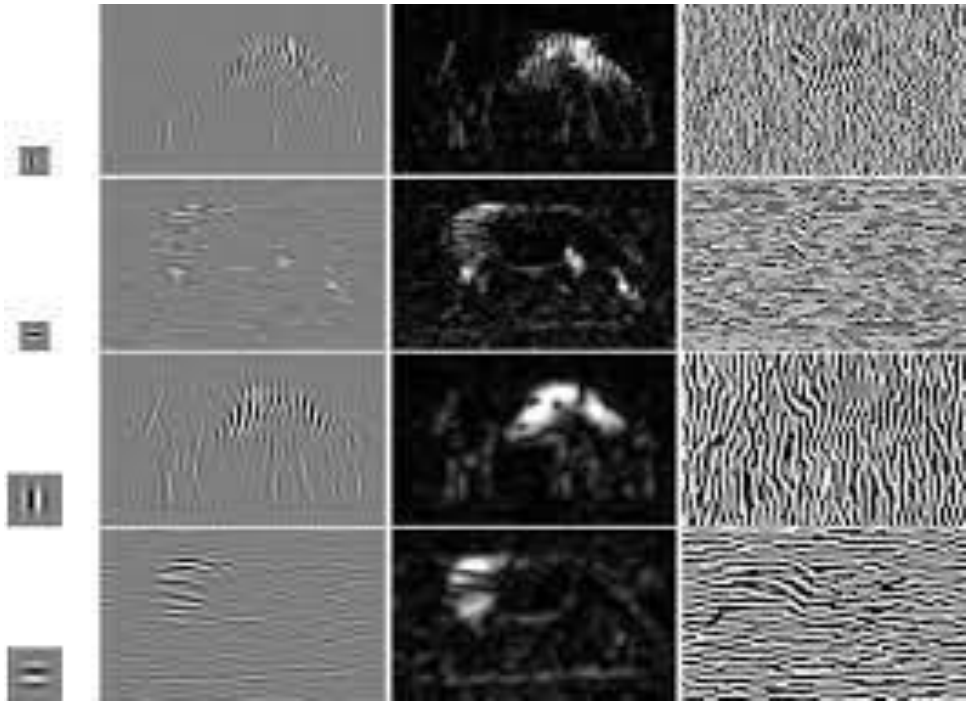


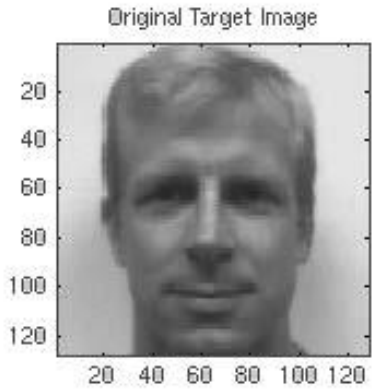

C
A
B
High-dimensional data
NDR
(Manifold Learning)
C
B
A
Low-dimensional
embedding

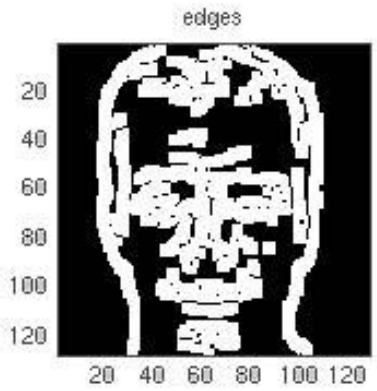

-114
-115
-116
-117
-118
-119
-120
-121
-122
-123
-124
32
34
36
38
40
42
0.1
0.05
0
-0.05
-0.1

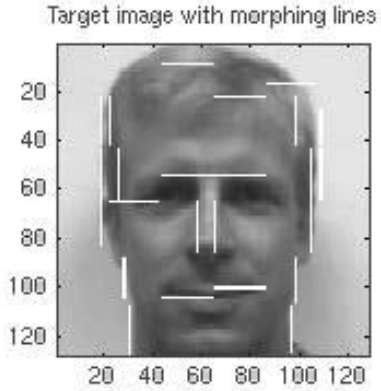

PCA (48.6)
Isomap (49.4)
GloVe (64.3)
SVD (68.0)
Word2vec (68.7)
SNE (72.8)
Regression (75.3)
t-SNE (91.4)

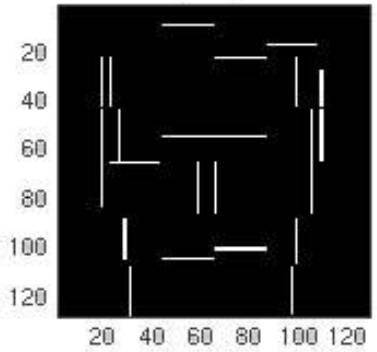


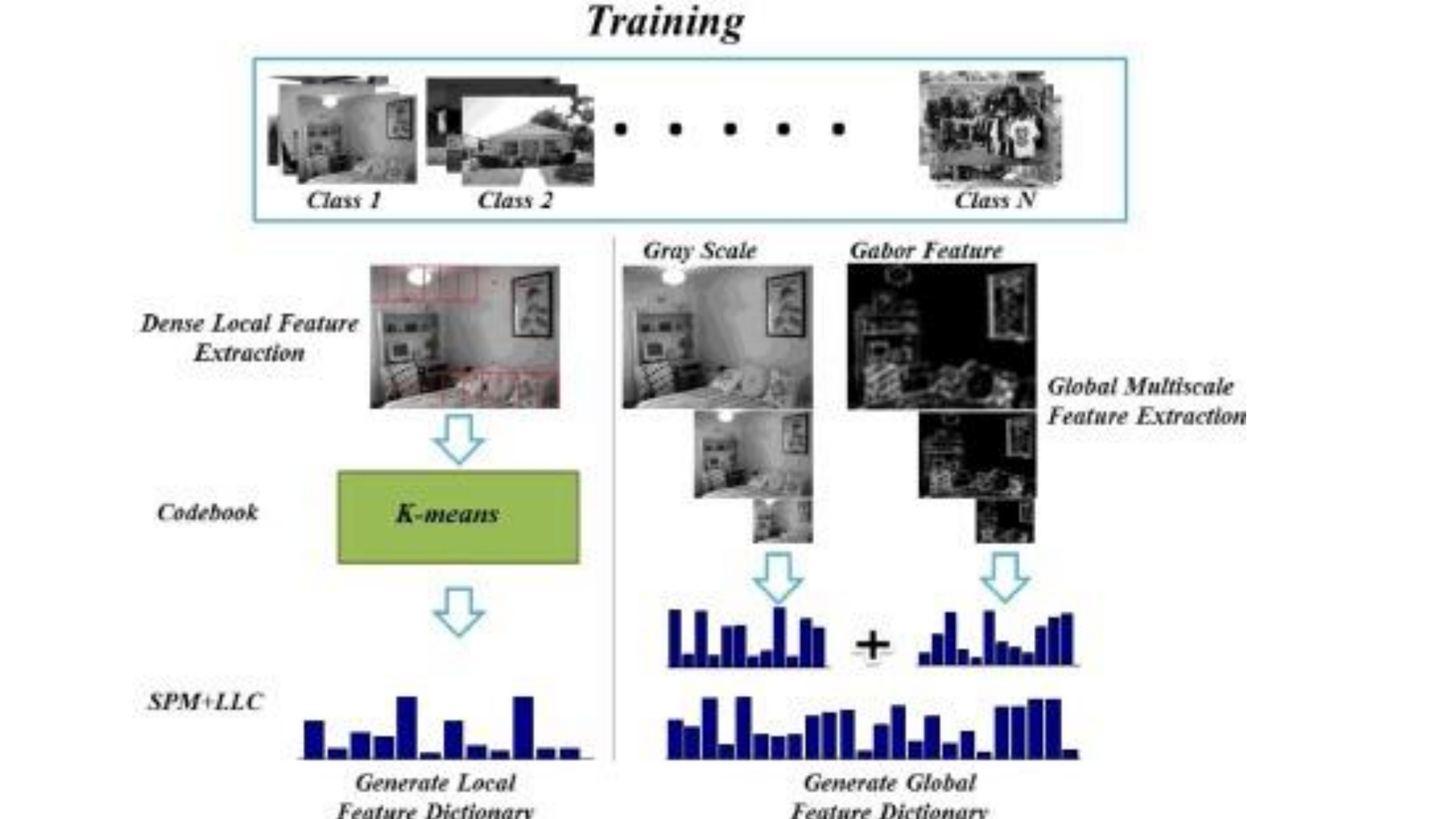



Original Target Image
20
40
60
80
100
120
20 40 60 80 100 120


edges
20
40
60
80
100
120
20 40 60 80 100 120


Target image with morphing lines
20
40
60
80
100
120
20 40 60 80 100 120


morphing lines
20
40
60
80
100
120
20 40 60 80 100 120


Training
Class 1
Class 2
Class N
Gray Scale
Gabor Feature
Dense Local Feature Extraction
Global Multiscale Feature Extraction
Codebook
K-means
SPM+LLC
Generate Local Feature Dictionary
Generate Global Feature Dictionary

- Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

- It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand, before getting them dirty with it.

The particular graphical techniques employed in EDA are often quite simple, consisting of various techniques of:

1. Plotting the raw data such as data traces, histograms, bihistograms, probability plots, lag plots, block plots, and Youden plots.
2. Plotting simple statistics such as mean plots, standard deviation plots, box plots, and main effects plots of the raw data.
3. Positioning such plots so as to maximize our natural pattern-recognition abilities, such as using multiple plots per page.

- The goal of a Data Scientist is to derive conclusions from the data. Through these conclusions, he is able to assist companies in making smarter business decisions.

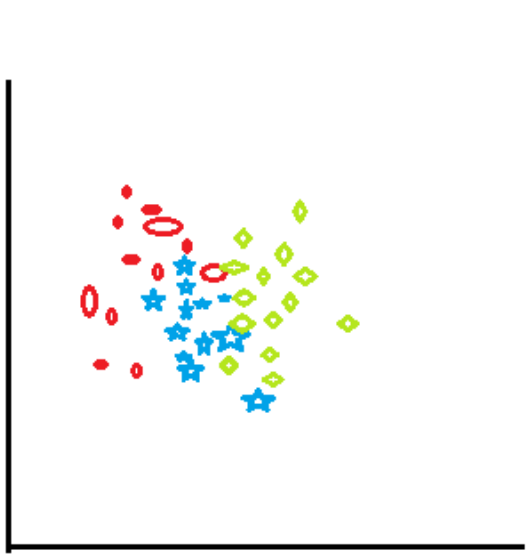


fig 1: before applying k-means clustering

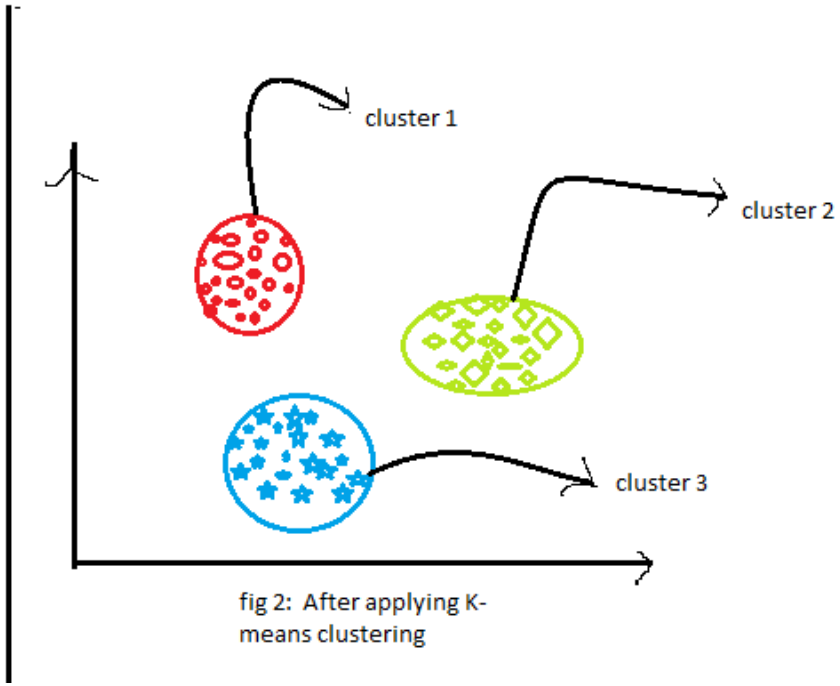
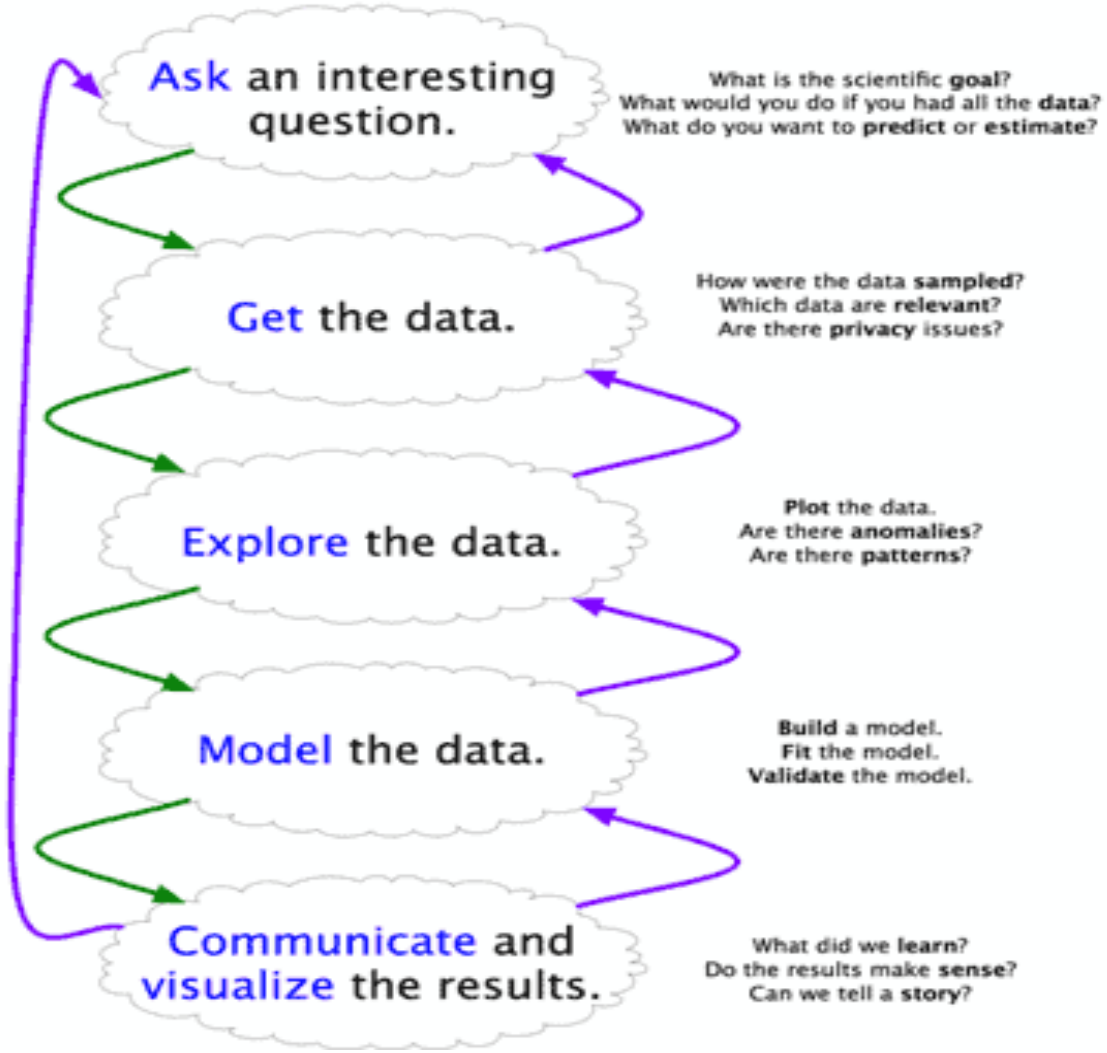


fig 2: After applying K-means clustering

Data Science Workflow Taught at Harvard



Blogs Describing a Data Science Workflow

- Aakash Tandel's Workflow
- For example, a workflow described by Aakash Tandel provides a high-level data science workflow, with a goal of serving as an example for new data scientists. It includes the following five logical steps:

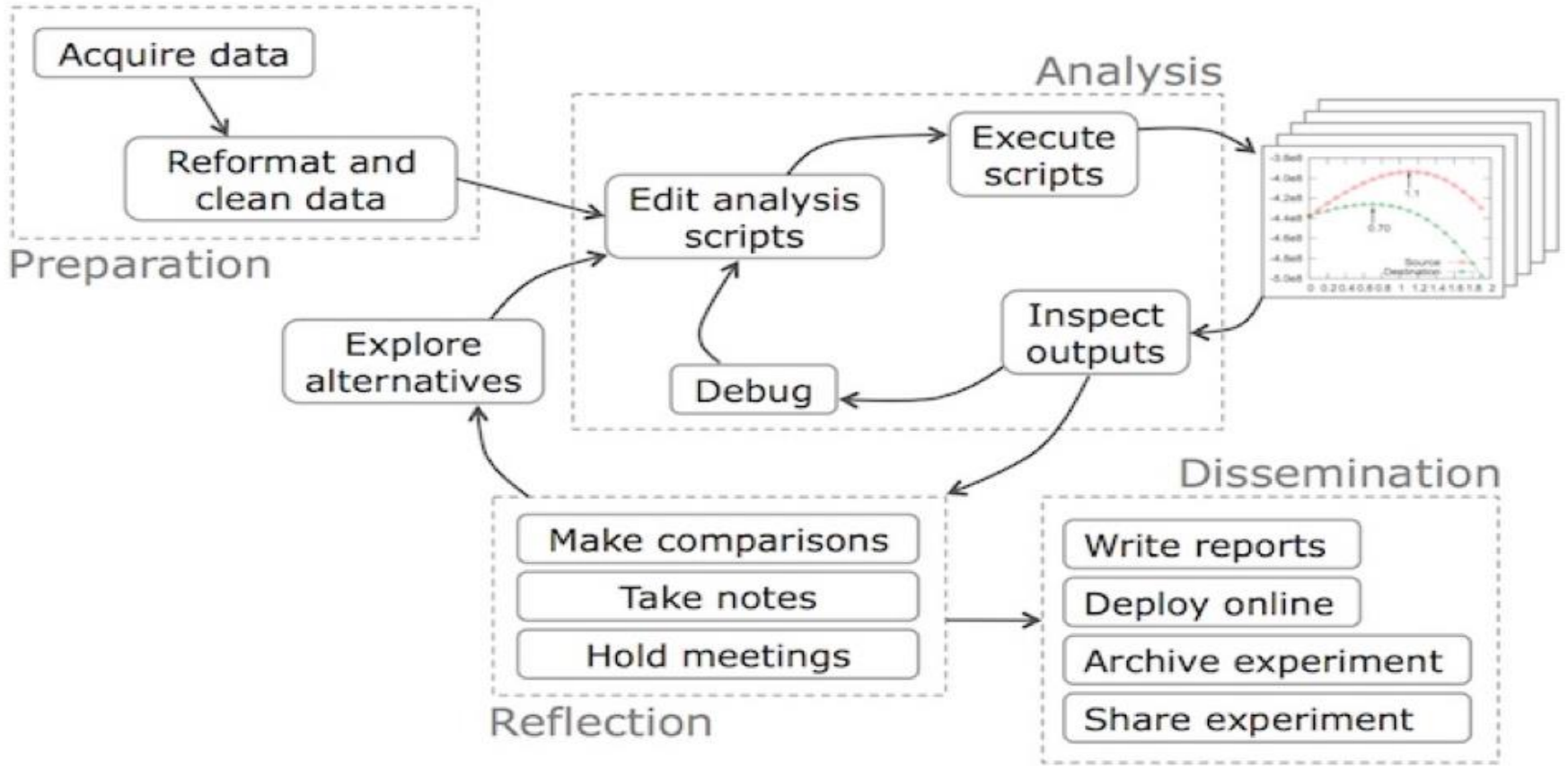
- Understand the objective
- Import the data
- Explore and clean the data
- Model the data
- Communicate the results.

- Aakanksha Joshi’s Workflow
- In a different blog, Aakanksha Joshi discussed using a data science workflow leveraging IBM’s Watson Studio Cloud, but the workflow could be useful independent of the technology stack used. In her blog, Joshi describes five linear phases:

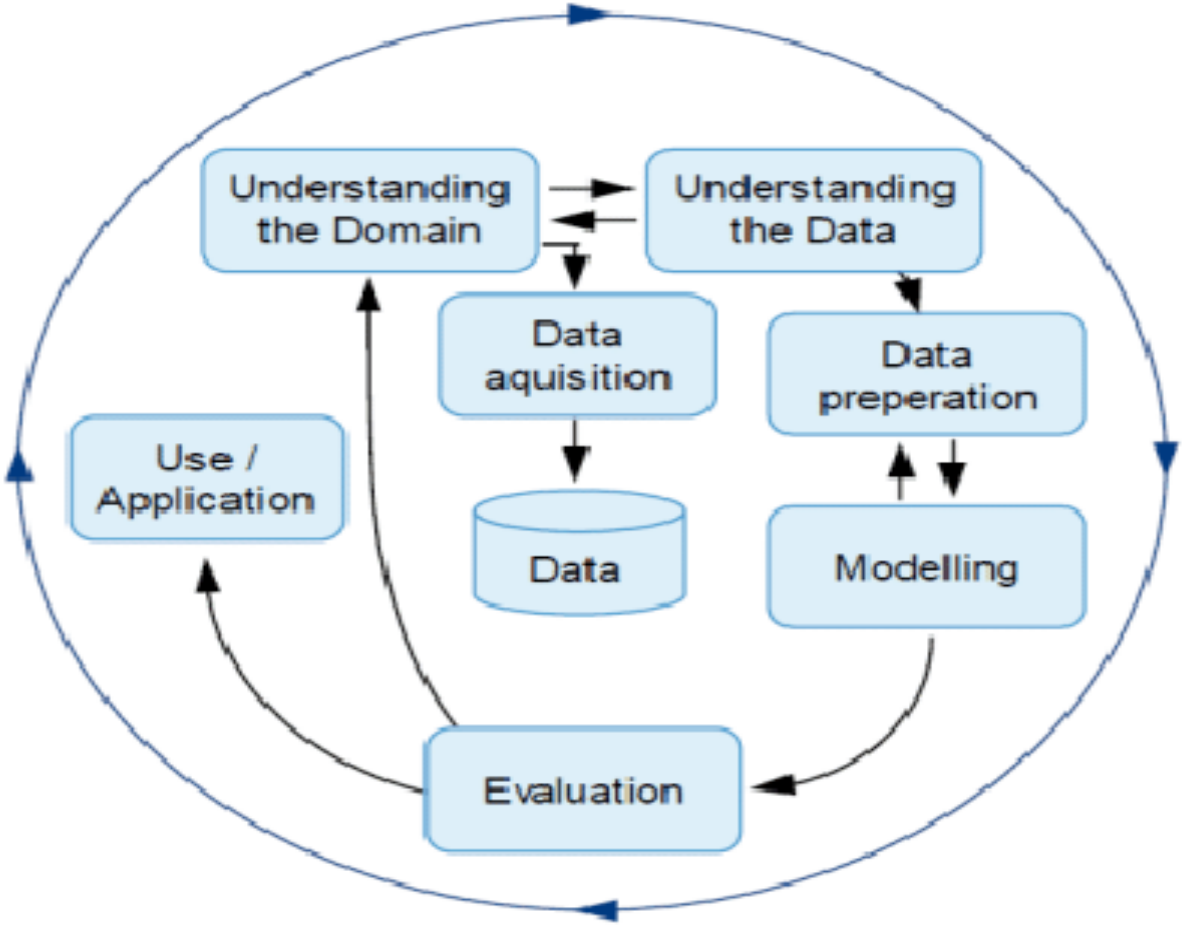
- Connect & access data
- Search and find relevant data
- Prepare for data analysis
- Build/train/deploy models
- Monitor/analyze/manage models

Philip Guo's Workflow

- A more advanced framework was described by Philip Guo. As shown below, it has four main phases.



- CRISP-DM: Defined to standardize a data mining process across industries, CRoss-Industry Standard Process for Data Mining (CRISP-DM) is the most well-known framework used to define a data science workflow



These frameworks all typically focus on the steps in a data science project (or skills needed by a data scientist).

	Harvard	CRISP-DM	OSEMN	Guo's	Tandel's	Joshi's
Understand	Ask an interesting question	Business understanding			Understand the objective	
Acquire	Get the data	Data understanding	Obtain	Prepare (acquire)	Import the data	Find, connect and access data
Clean		Data preparation	Scrub	Prepare (clean)	Clean the data	Prepare the data
Explore	Explore the data		Explore		Explore the data	
Model	Model the data	Modeling	Model	Analysis	Model the data	Build models
Evaluate		Evaluation	iNterpret	Reflect		
Communicate	Communicate / Visualize			Disseminate	Communicate results	
Deploy		Deployment				Deploy models
Monitor						Monitor models

This table might help you decide what phases are best for your team

DATA SCIENTIST

MUST-HAVE SKILLS

MATH & STATISTICS

- Machine Learning
- Statistical Modeling
- Experimental Analysis
- Clustering
- Regression Analysis

PROGRAMMING & DATABASE

- Computer Science Fundamentals
- Relational Database Management System
- Data Visualization
- Python
- Big Data

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passion for working with data
- Problem solver
- Strategic, proactive, and creative
- Collaborative by leading

COMMUNICATION & VISUALIZATION

- Storytelling skills
- Communicate data-based insights
- Collaborative skills for effective
- Knowledge of tools like Tableau
- Visual Design

One primary tip

- Coursera
- Udemy
- Data Camp
- Edx
- Udacity