01 Unit

INTRODUCTION TO

DATA SCIENCE & BIG DATA

Outline

- Basics and need of Data Science and Big Data
- Applications of Data Science | 5 V's of Big Data
- Relationship between Data Science and Information Science
- Business intelligence versus Data Science
- Data Science Life Cycle

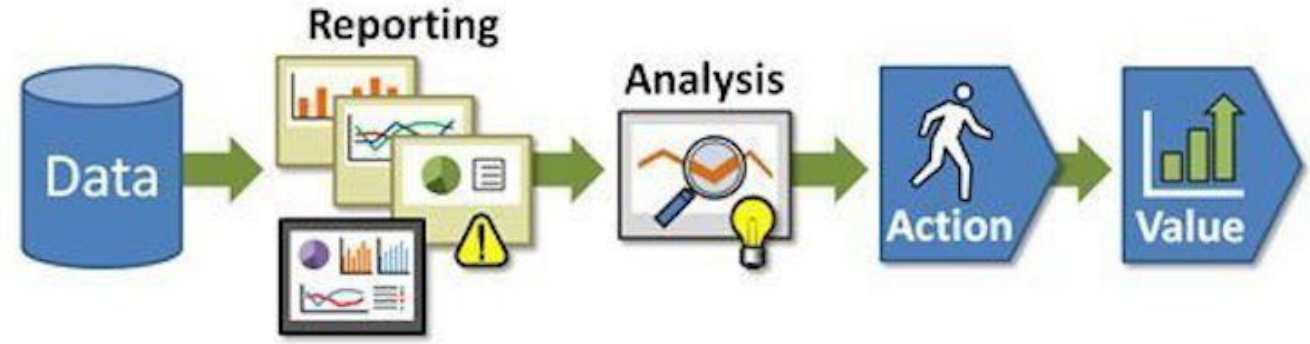
- Data
- Data Types
- Data Collection
- Need of Data wrangling

- Methods
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation
- Data Discretization

Analyze needs and challenges for Data Science and Big Data Analytics

Data Analysis

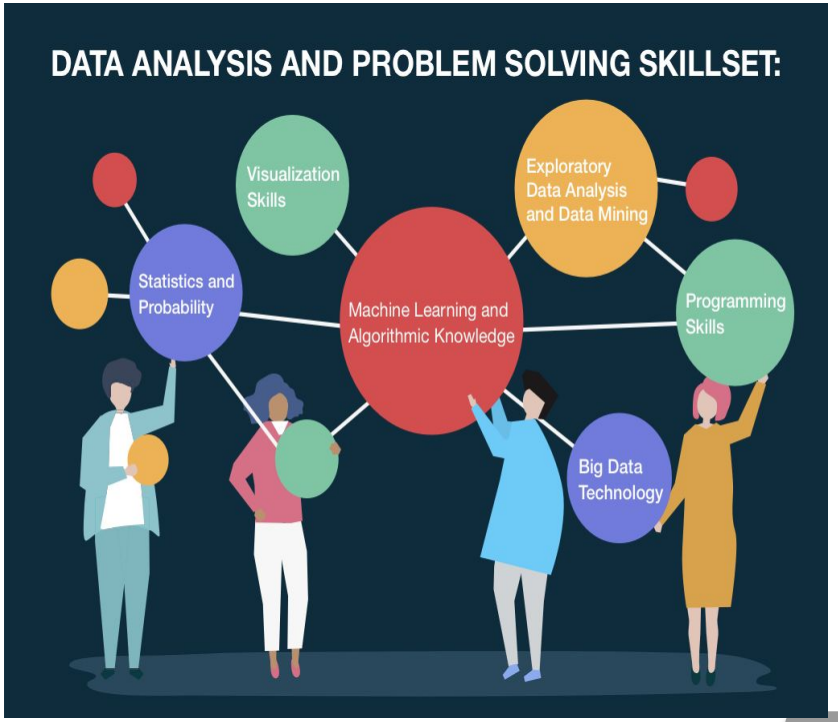
Data analysis is a procedure of investigating, cleaning, transforming, and training of the data with the aim of finding some useful information, recommend conclusions and helps in decision-making



Data Analytics

Analytics is utilizing data, machine learning, statistical analysis and computer-based models to get better insight and make better decisions from the data.

Analytics is defined as "a process of transforming data into actions through analysis and insight in the context of organisational decision making and problem-solving."



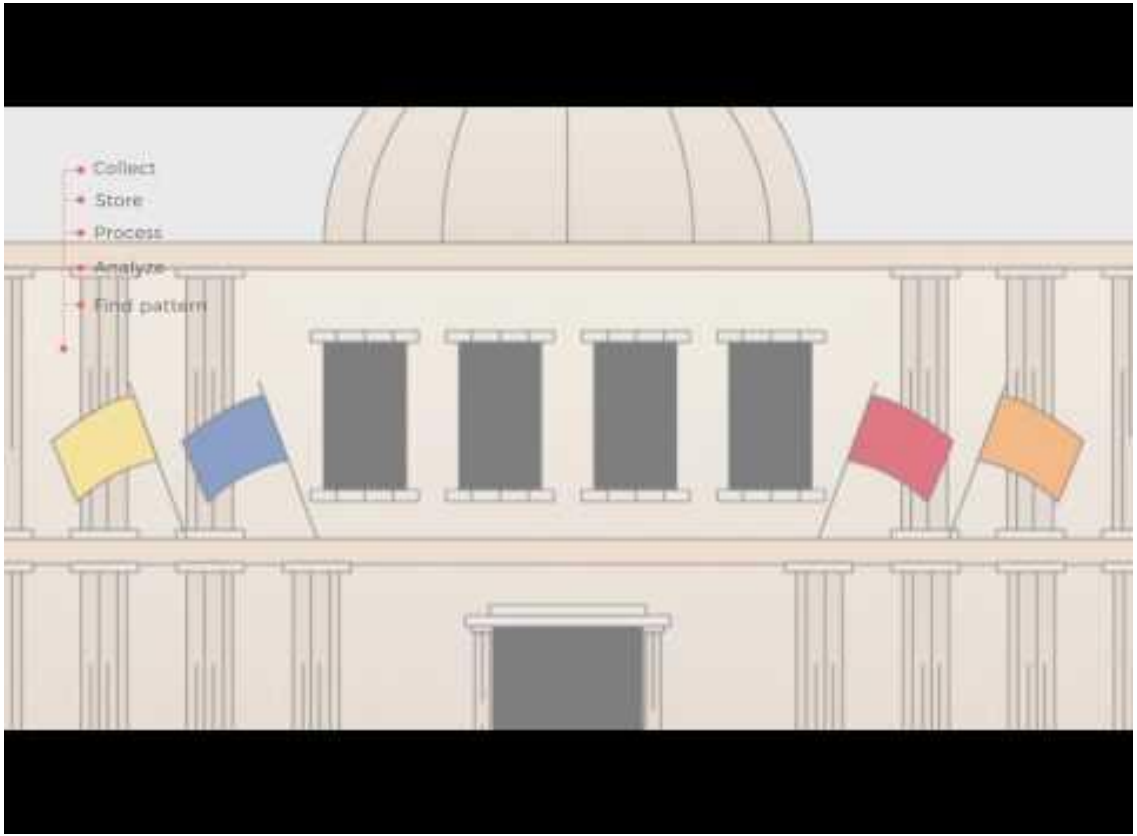
Data Analytics Vs Data Analysis

Data Analytics

There
are many analytics tools in a market but mainly R, Tableau Public, Python, SAS, Apache Spark, Excel are used.

Data Analysis

For
analyse the data OpenRefine, KNIME, RapidMiner, Google Fusion Tables, Tableau Public, NodeXL, WolframAlpha tools are used.


Collect
Store
Process
Analyze
Find pattern

Big data is a term 

related to

extracting meaningful data

by

huge amount of complex, variously formatted data generated at high speed

that cannot be

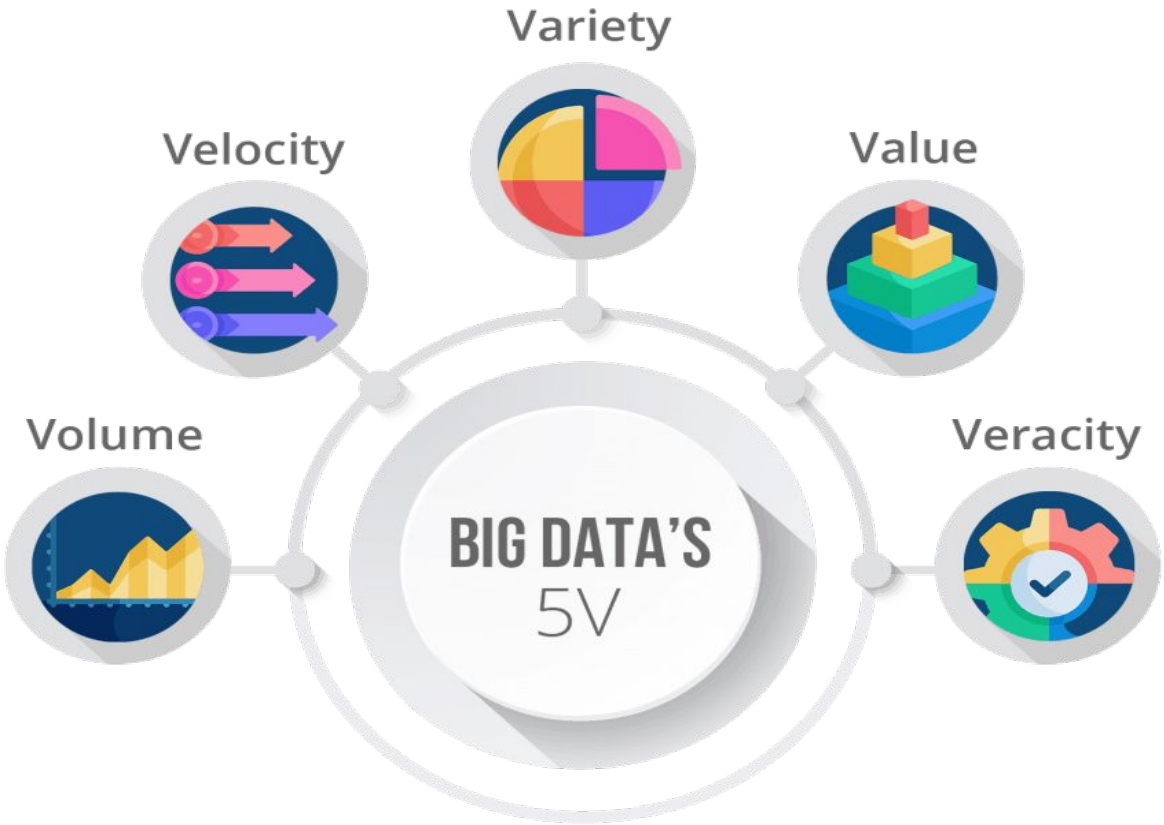
Handled | Processed

by

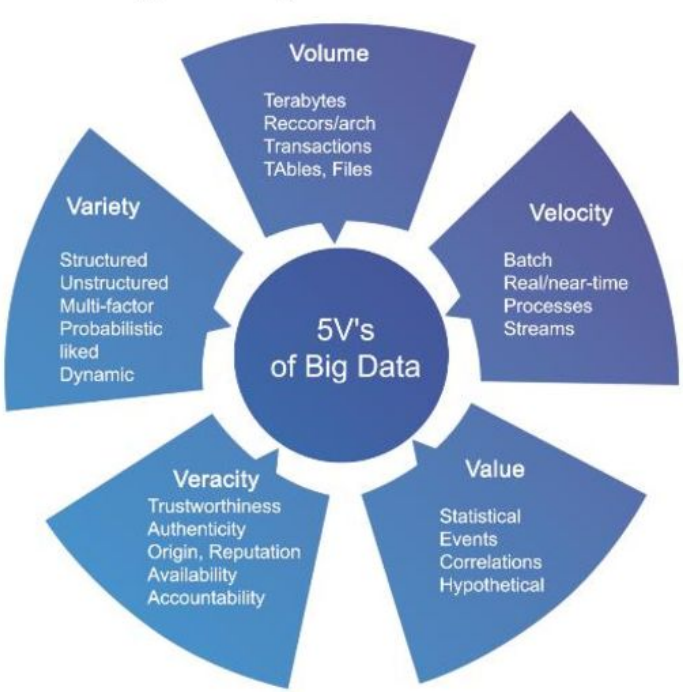
Traditional System


Big Data
Value
Veracity
Variety
Volume
Velocity

5 V's of Big Data



Source: https://medium.com/analytics-vidhya/big-data-3-vs-and-5-v-s-c1cae2a6d311

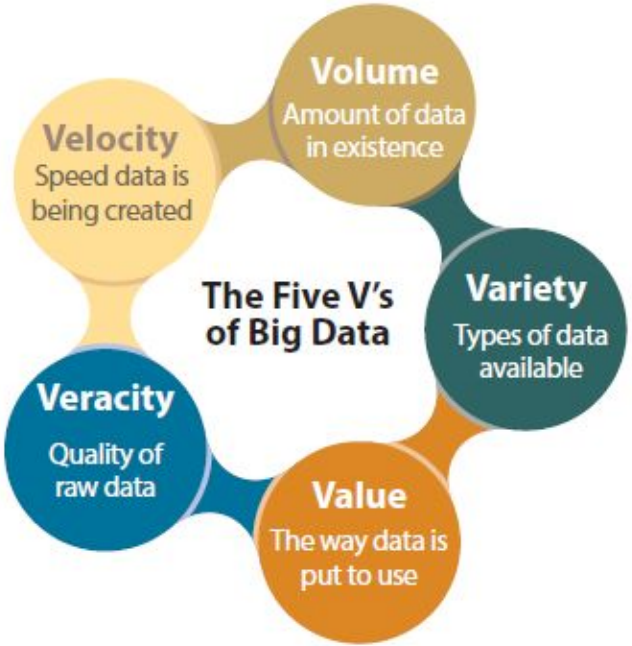


Volume

- Refers to the size of data (generally in Terabytes)
- This size aspect of data is referred to as Volume in the Big Data world.

Velocity

- Velocity refers to the speed at which the data is being generated.
- This speed aspect of data generation is referred to as Velocity in the Big Data world.



Variety

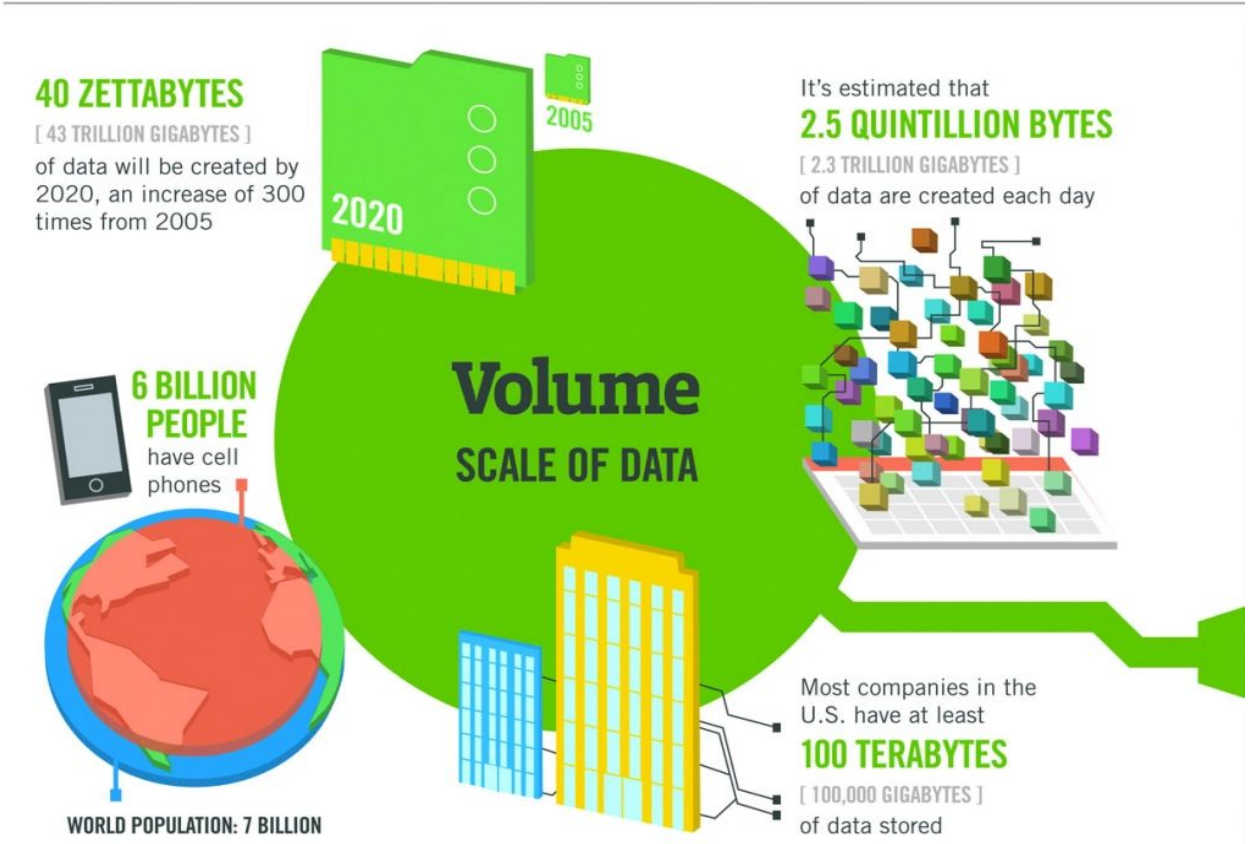
- Data in different formats
- Eg. Structured or Un-Structured Data
- Some more example: Data in text, image, animation, audio or video

Veracity

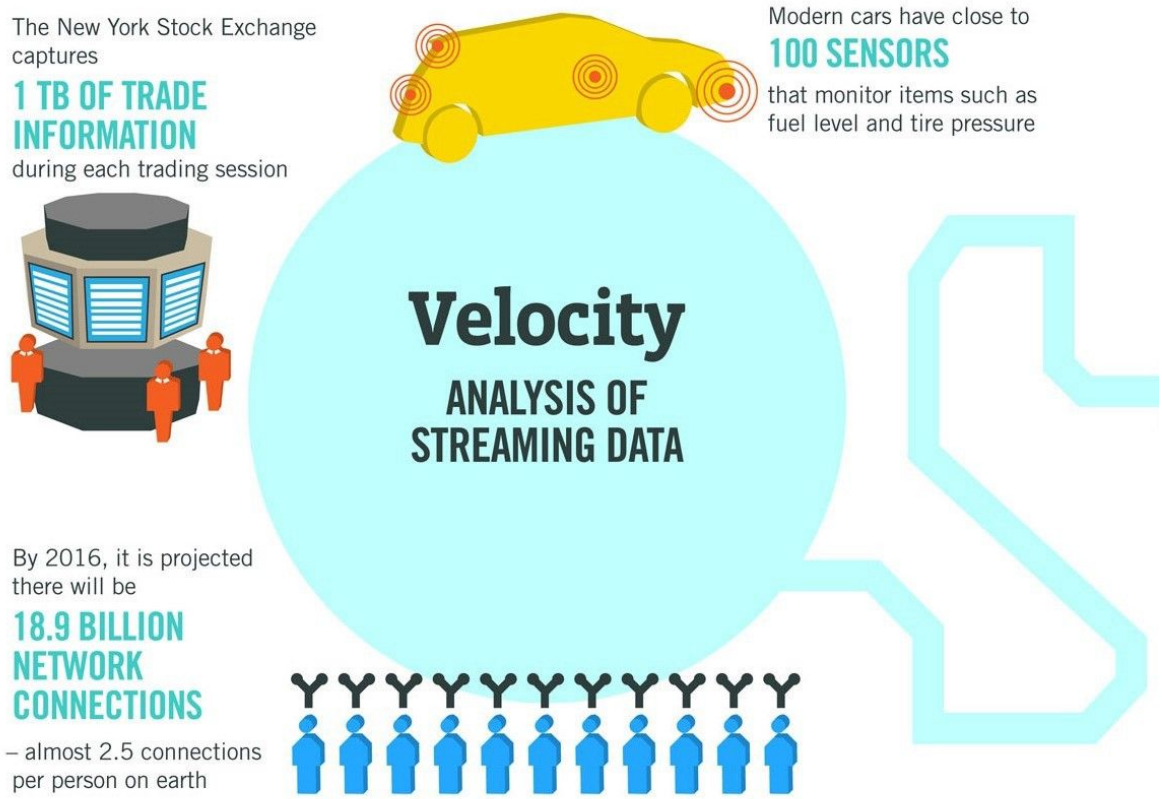
- Refers to the quality and accuracy of data.

Value

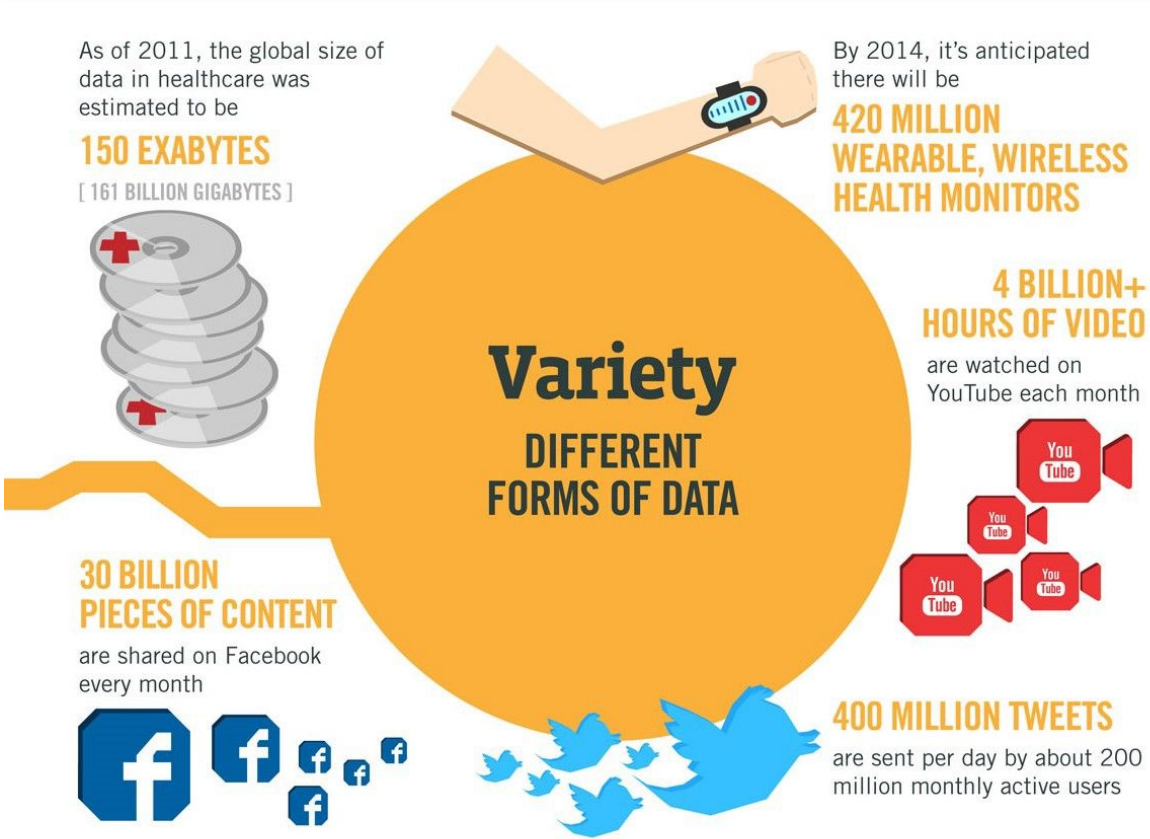
- what organizations can do with the collected data.



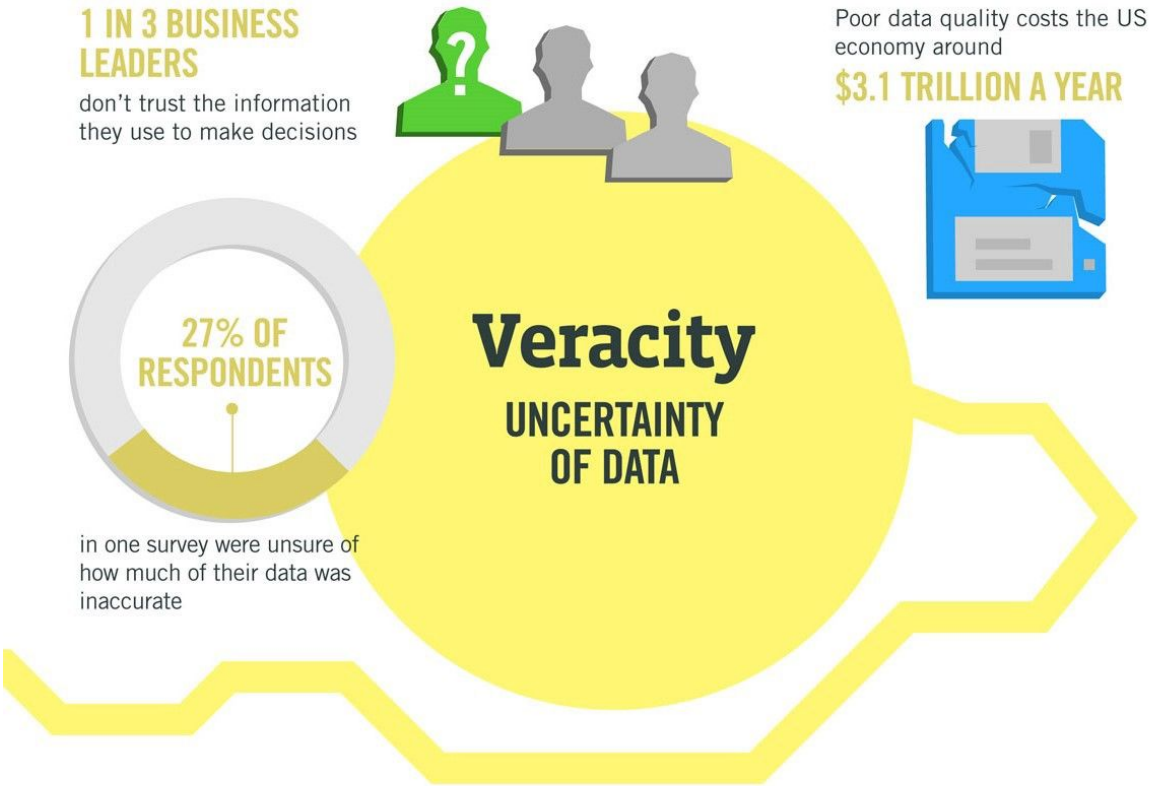
Source: https://medium.com/analytics-vidhya/big-data-3-vs-and-5-v-s-c1cae2a6d311

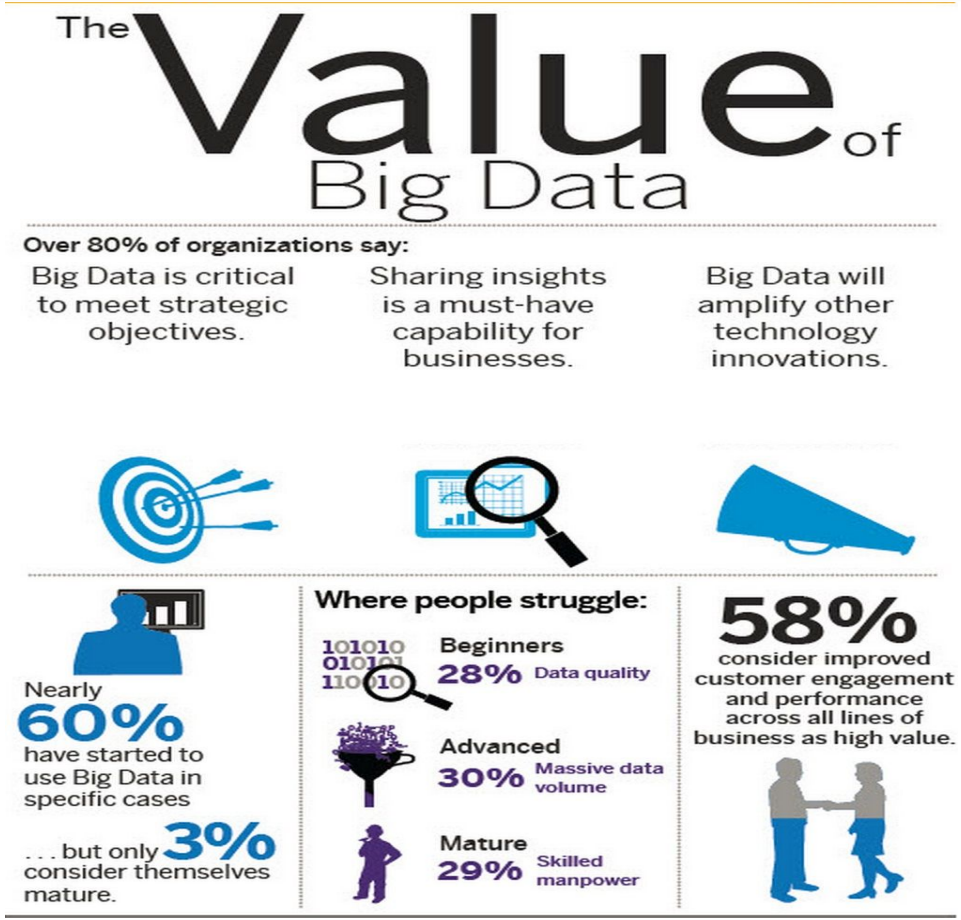


Source: https://medium.com/analytics-vidhya/big-data-3-vs-and-5-v-s-c1cae2a6d311

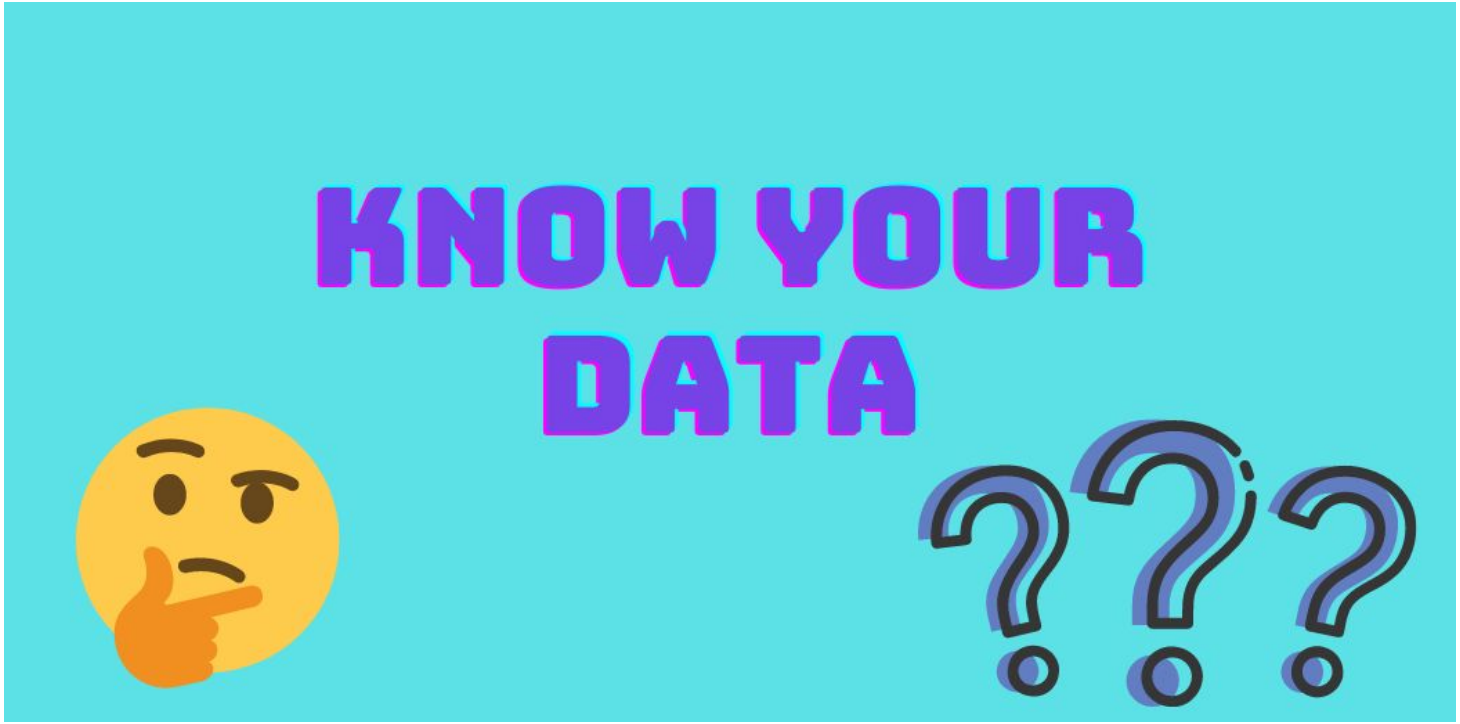


Source: https://medium.com/analytics-vidhya/big-data-3-vs-and-5-v-s-c1cae2a6d311


1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA


The Value of Big Data
Over 80% of organizations say:
Big Data is critical to meet strategic objectives.
Sharing insights is a must-have capability for businesses.
Big Data will amplify other technology innovations.
Nearly 60% have started to use Big Data in specific cases
. . . but only 3% consider themselves mature.
Where people struggle:
Beginners 28% Data quality
Advanced 30% Massive data volume
Mature 29% Skilled manpower
58% consider improved customer engagement and performance across all lines of business as high value.

5 V's of Big Data :Case study for Facebook



Sources of Big Data

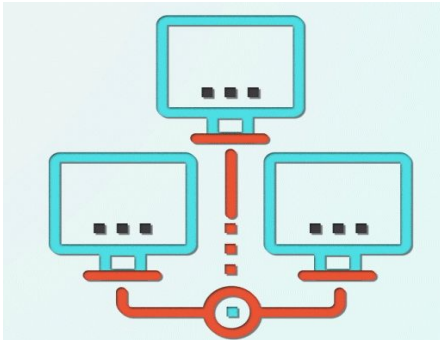
Sources of Big Data

Enterprise Data

Large Data With Different Formats

Sources of Big Data

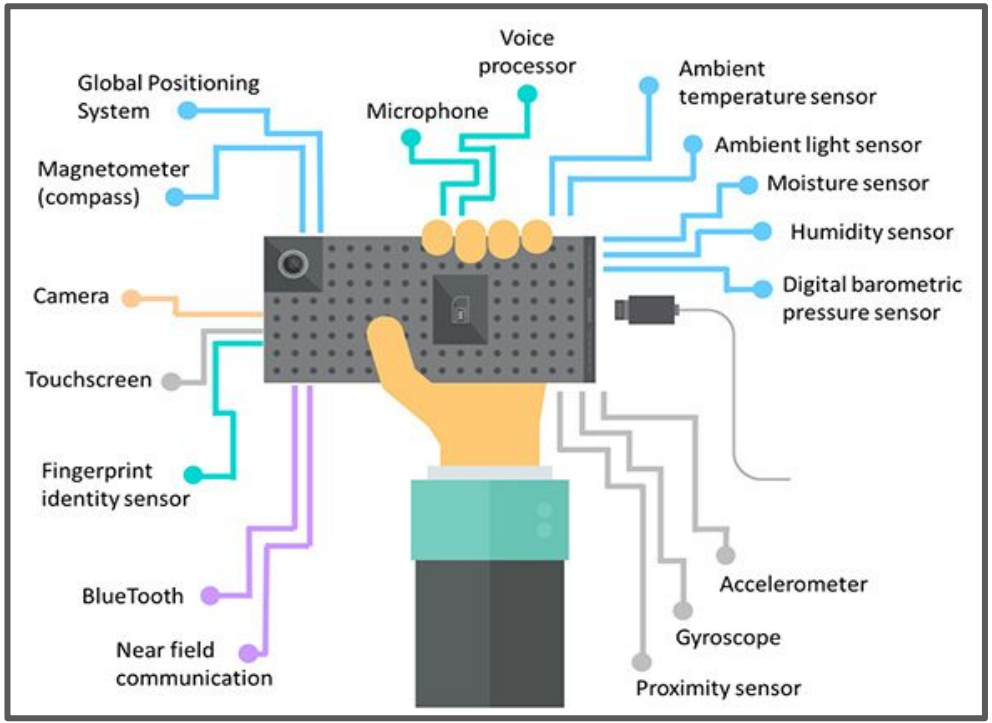
Transactional Data



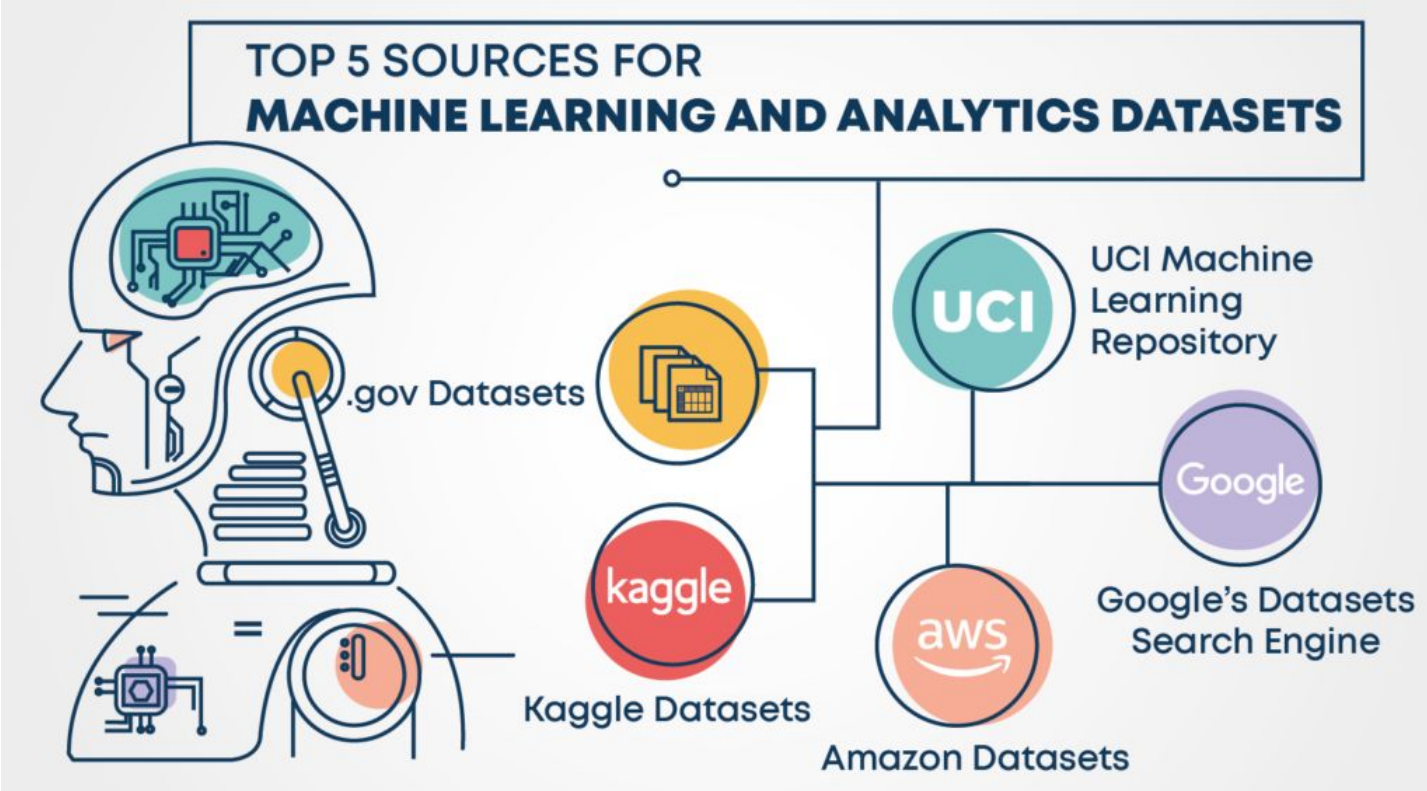
Sources of Structured Data

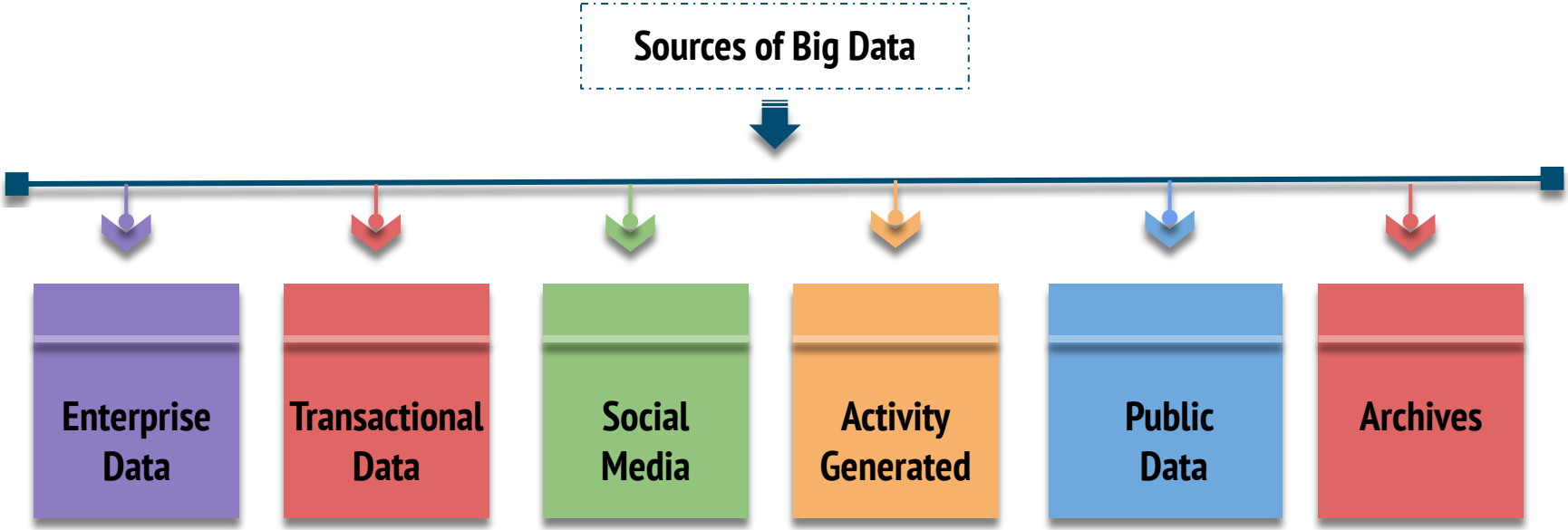
Social Media

Activity Generated Data



Public Data


TOP 5 SOURCES FOR
MACHINE LEARNING AND ANALYTICS DATASETS
.gov Datasets
UCI
UCI Machine Learning Repository
Google
Google's Datasets Search Engine
kaggle
Kaggle Datasets
aws
Amazon Datasets


Sources of Big Data
Enterprise Data
Transactional Data
Social Media
Activity Generated
Public Data
Archives

Archives

COMPLIANCE ARCHIVES

To meet regulatory & audit requirements

HISTORICAL ARCHIVES

To record data of your business

ANALYTICS ARCHIVES

For reporting and analytics

Big Data Statistics

Click Here

- 100 Terabytes of data is uploaded to Facebook every day
- Facebook Stores, Processes, and Analyzes more than 30 Petabytes of user generated data
- Twitter generates 12 Terabytes of data every day
- LinkedIn processes and mines Petabytes of user data to power the "People You May Know" feature
- YouTube users upload 48 hours of new video content every minute of the day
- Decoding of the human genome used to take 10 years. Now it can be done in 7 days
- 500+ new websites are created every minute of the day

Need of Big Data

Study of Data

It involves

developing methods

for

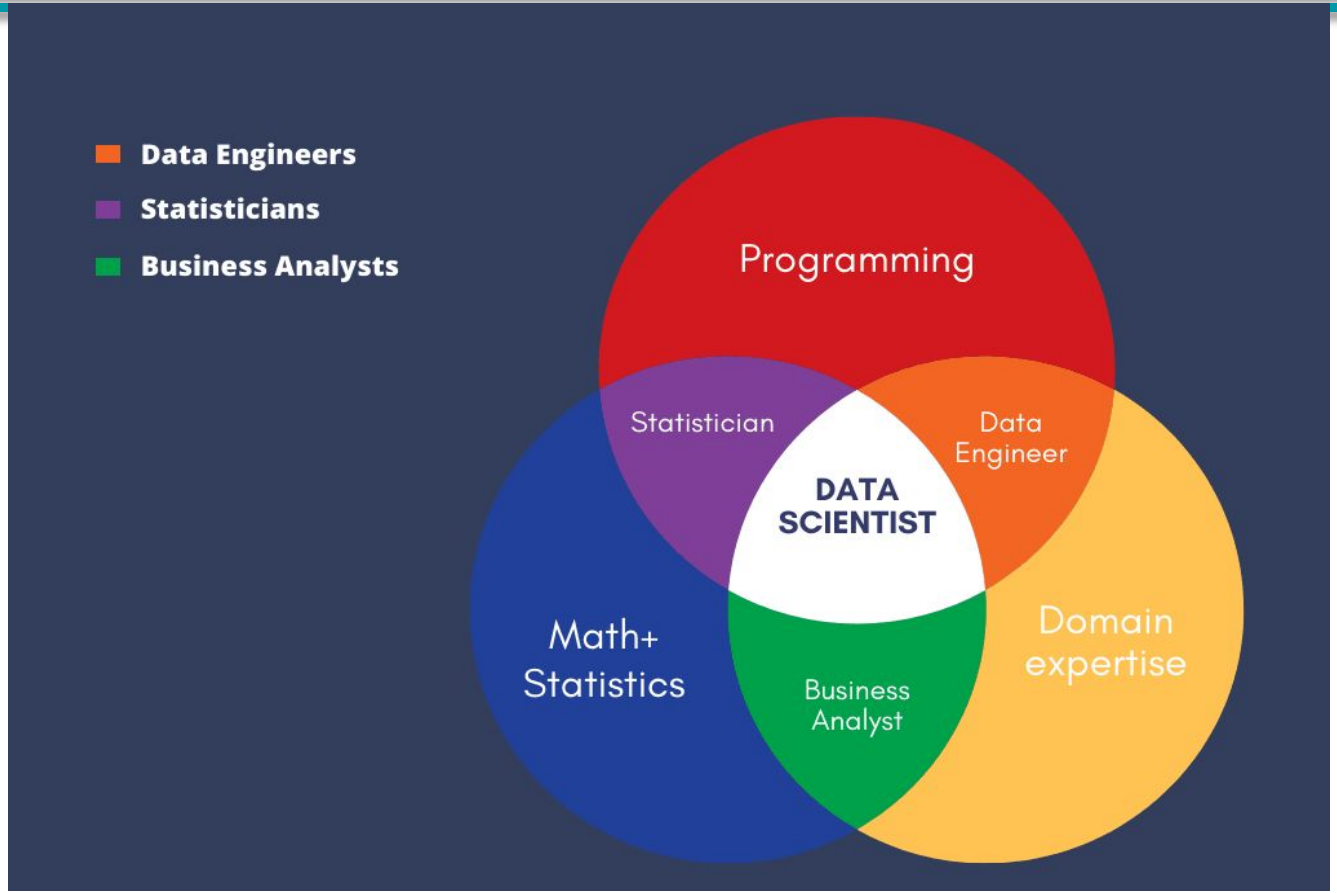
Recording, storing , analysing

To gain

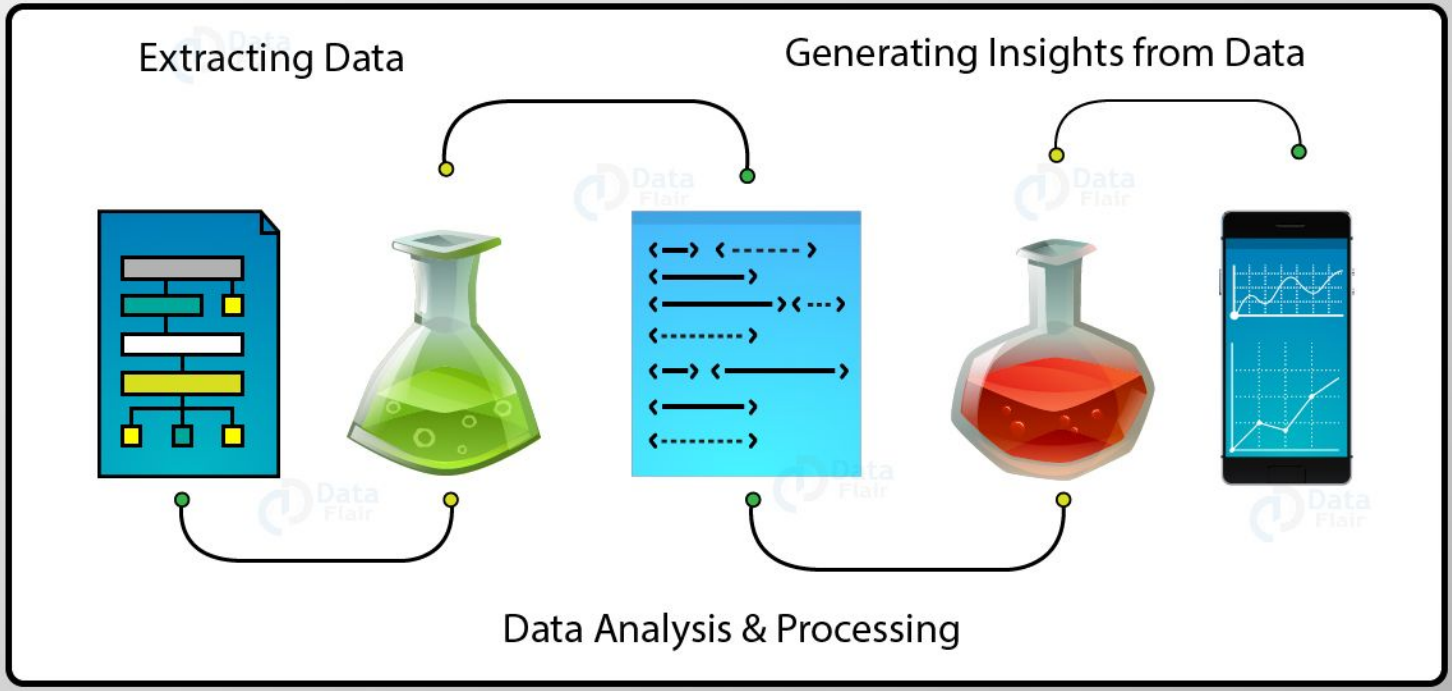
insights and knowledge

from

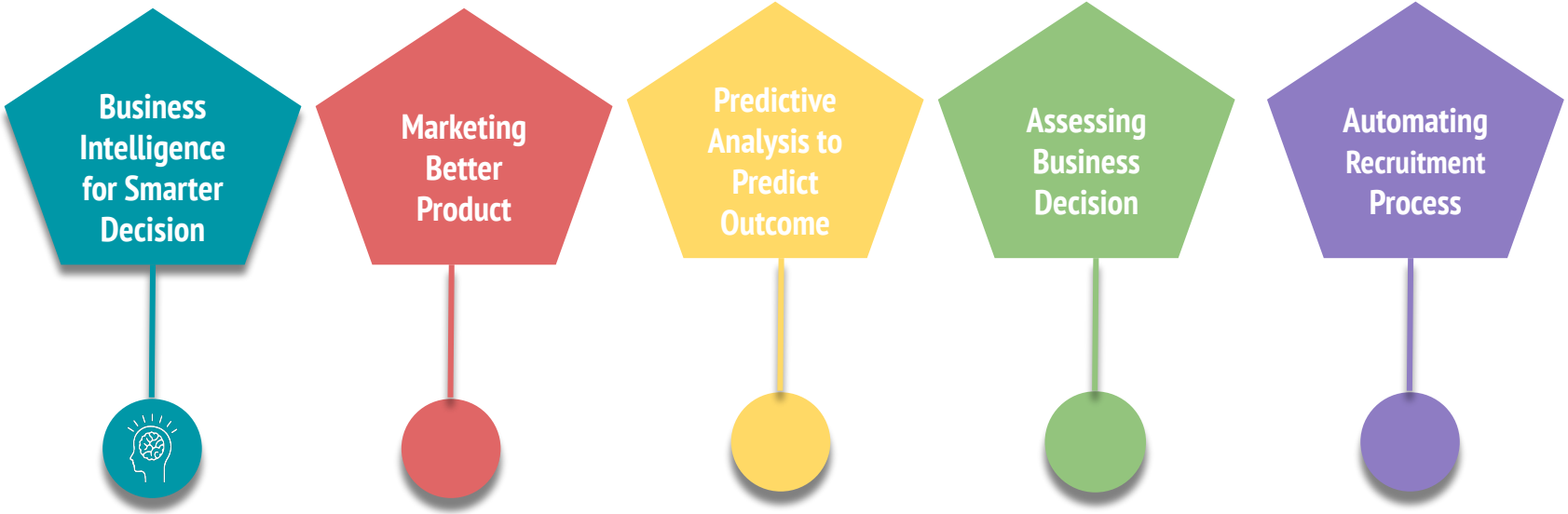
any type of data — both structured and unstructured.


Data Engineers
Statisticians
Business Analysts
Programming
Statistician
Data Engineer
DATA SCIENTIST
Math+ Statistics
Business Analyst
Domain expertise

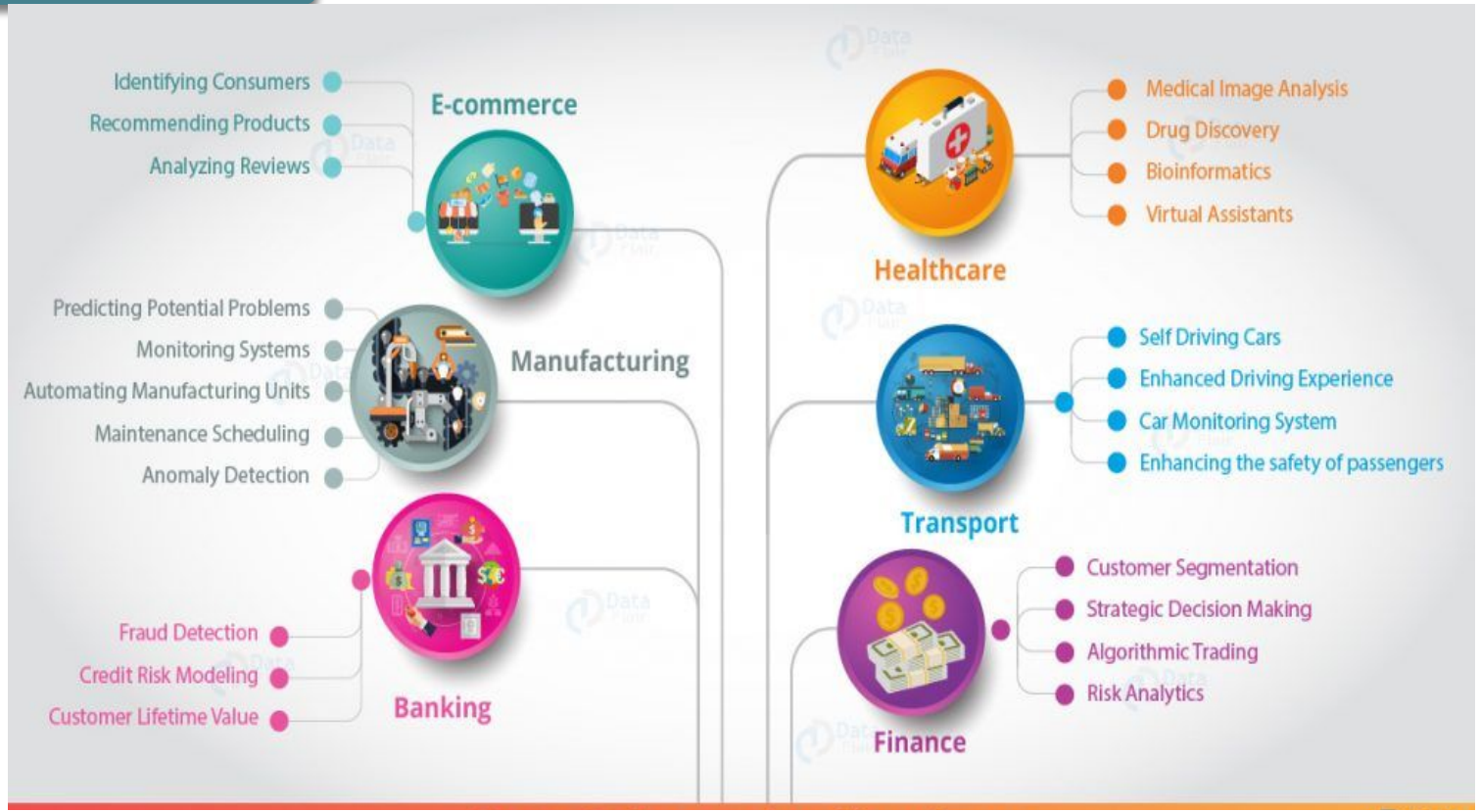
Purpose of Data Science

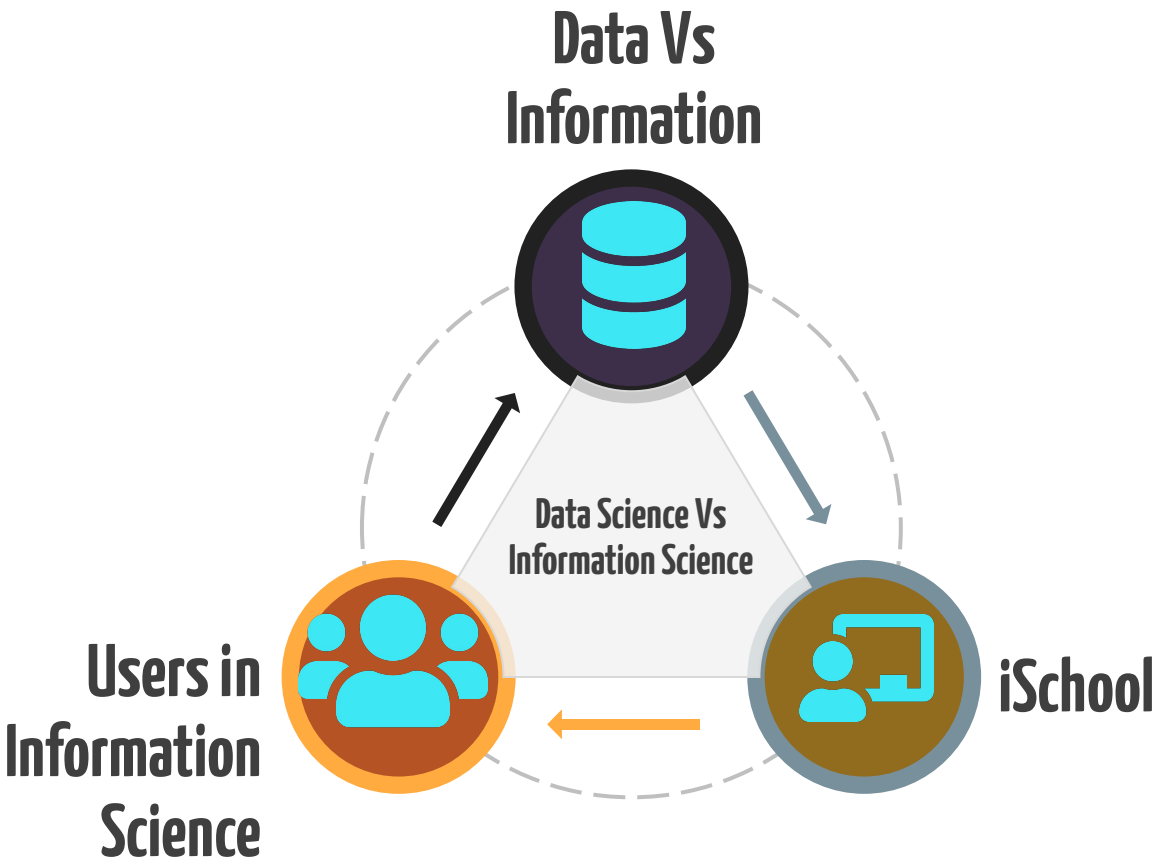


Importance of Data Science in Business



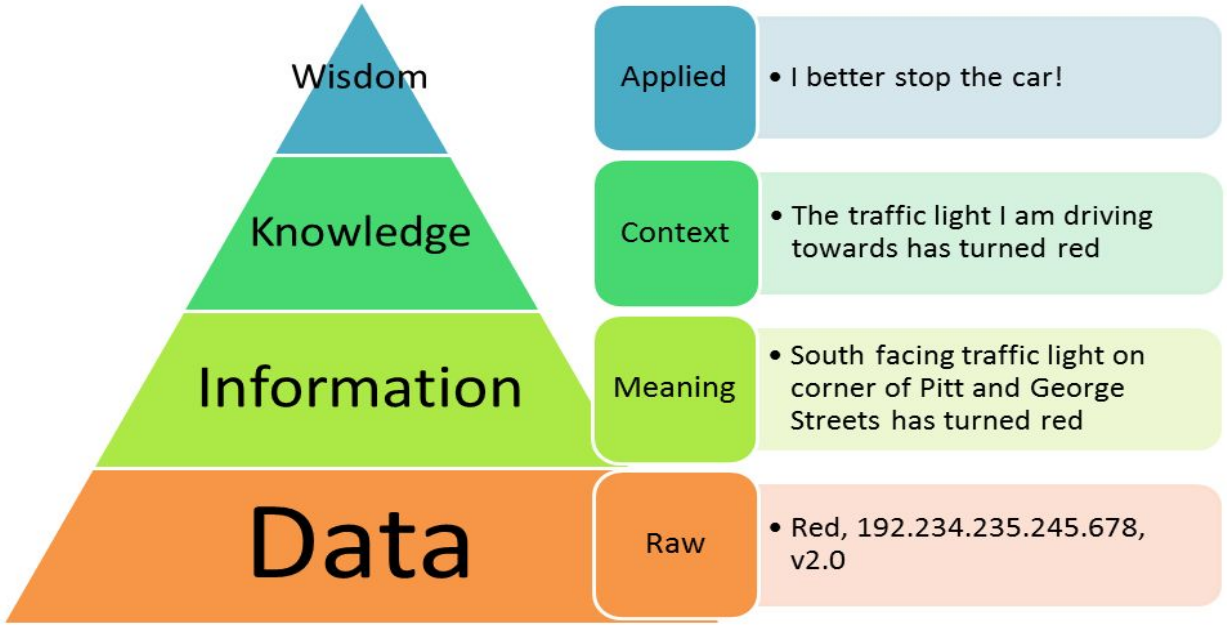
Applications of Data Science



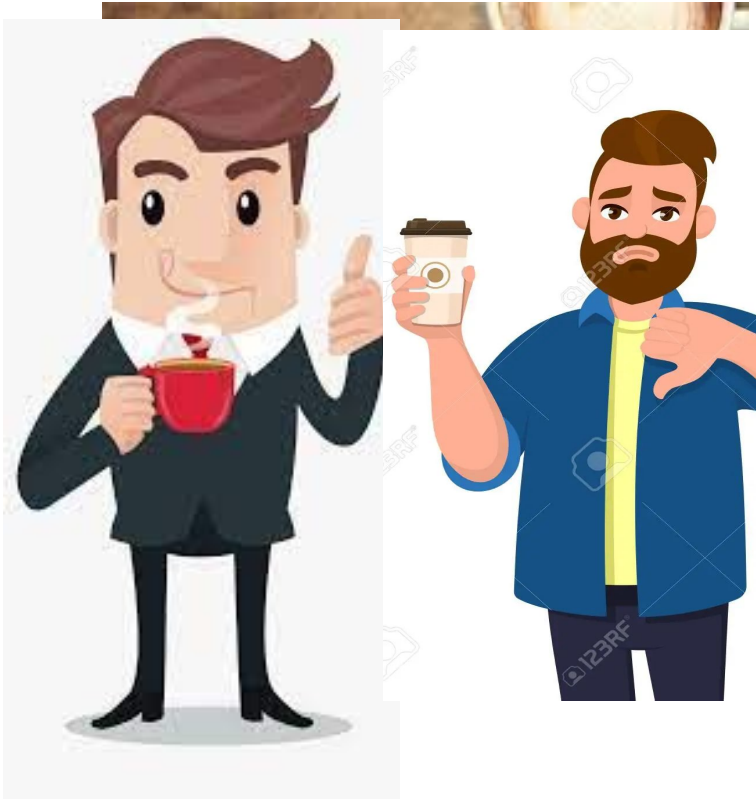

Data Vs
Information
Data Science Vs
Information Science
Users in
Information
Science
iSchool

reference

Data Vs Information



Users in Information Science



Users in Information Science

Usefulness is a criterion

That determines

the interaction between the user and the information

to

accomplishing the task or goal of the user.

studying

people's behaviors

to create

better methods for personalization and recommendation

Data Science in Information Schools

acquire diverse perspectives on data and information.

data science skills and knowledge

focus on analyzing data

extracting insightful information grounded in context.

"valuable resource in the creation of business and information technology strategies."

reference

Relationship between Data Science and Information Science

Data Science

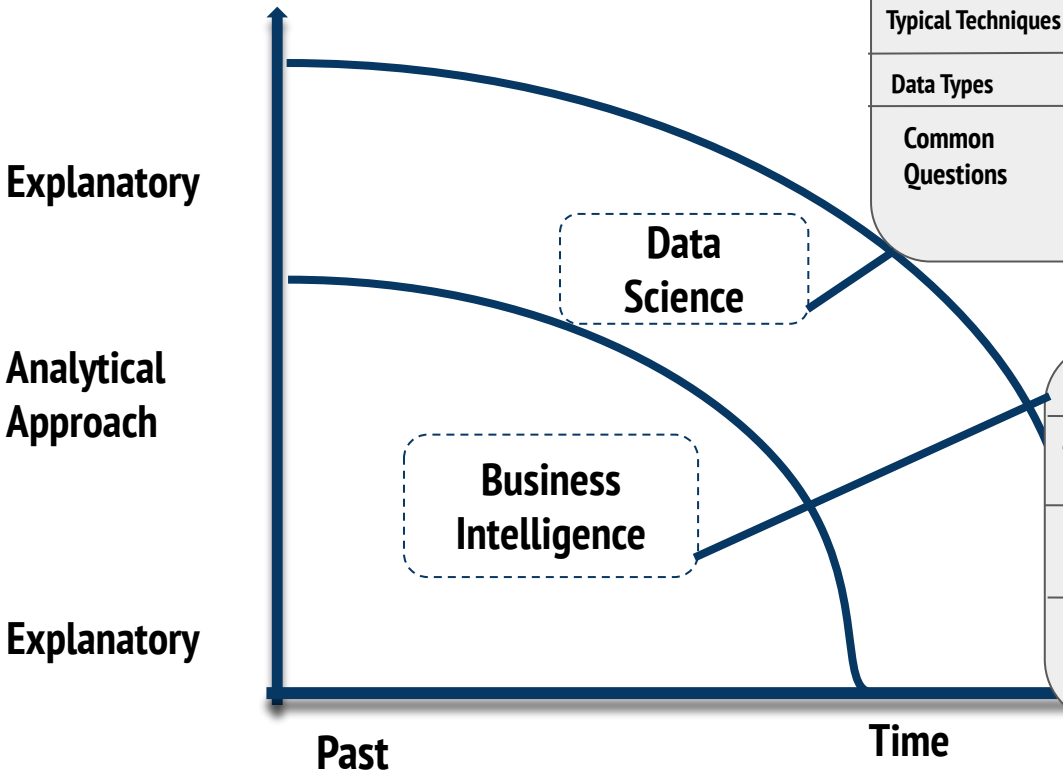
- Data Science is the discovery of knowledge or actionable information in data
- Data Science is heavy on computer science and mathematics
- Data Science is used in business functions such as strategy formation,decision making & operational process.

Information Science

- Information Science is the design of practices for sorting & retrieving information
- Information Science is more concerned with the areas such as library science, cognitive science & communication
- Information Science is used in areas such as knowledge management,data management & interaction design

reference

Business intelligence versus Data Science



Predictive Analytics & Data Mining (Data Science)	
Typical Techniques	Optimization,predictive modeling,forecasting,statistical Analysis
Data Types	Structured/Unstructured data,many types of sources,Very large datasets
Common Questions	What if...? What's the optimal scenario for our business ? What will happen next ? what if these trends continue ? why is this happening

Business Intelligence	
Typical Techniques	Standard & ad hoc reporting , dashboards,alerts,queries,detail on demand
Data Types	Structured data,traditional sources, manageable datasets
Common Questions	What happened lut quarter? How many units sold? Whore Is the problem? In which situations?

Business intelligence versus Data Science

Business intelligence

- It is basically a set of technologies, applications & processes that are used by the enterprises for business data analysis.
- It focuses the past and present.
- It mainly deals only with structured data.
- It makes the use of analytic method.
- It's tools are InsightSquared Sales Analytics, Klipfolio, ThoughtSpot, Cyfe, TIBCO Spotfire etc.

Data Science

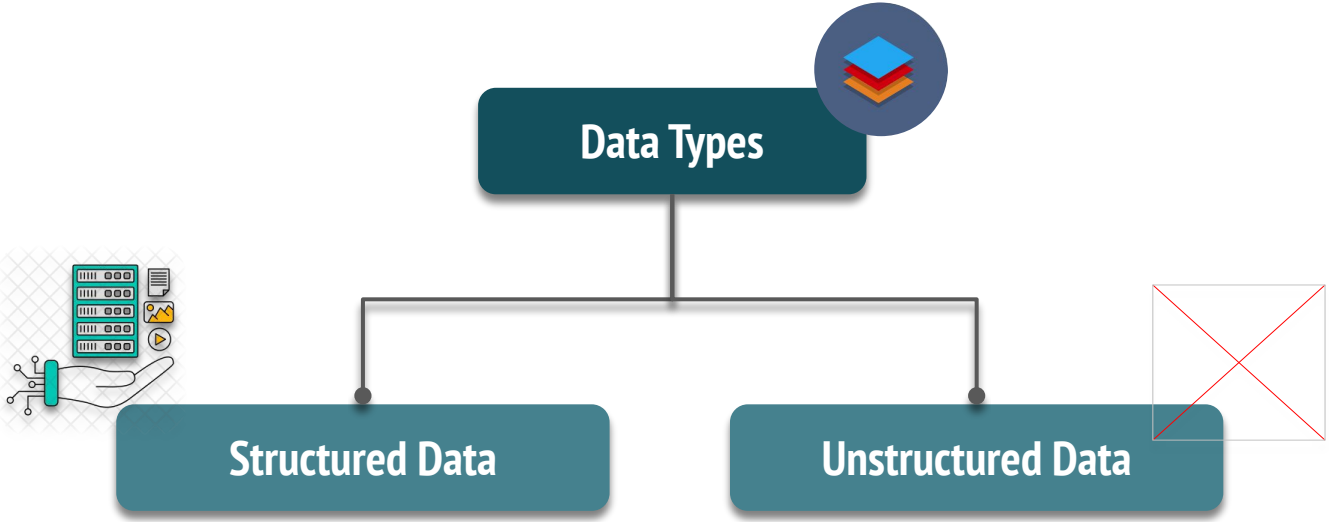
- It is a field that uses mathematics, statistics and various other tools to discover the hidden patterns in the data.
- It focuses on the future.
- It deals with both structured as well as unstructured data.
- It makes the use of scientific method.
- It's tools are SAS, BigML, MATLAB, Excel etc.

Reference

Data

- **What is Data?:** Raw or Fact
- **Sources of Data:** By Research, Human Creativity, Sensors, Transactions, Digital interactions, Calculations & Artificial Intelligence.
- **Data** is used to make decisions, solve problems, drive automation, execute transactions.

Data : Data Types



Structured Data

- Data having Predefined data model/schema/structure and is often either relational in nature or is closely resembling a relational model.
- Easily managed and consumed using the traditional tools / techniques.
- It includes data in the relational databases, data from ERP/CRM/SAP Systems, XML files etc.

Unstructured Data

- Data without labels.
- It includes flat files, spreadsheets, Word documents, emails, images, audio files, video files, feeds, PDF files, scanned documents, etc.

Data : Data Types

Characteristics	Structured Data	Unstructured Data
Data Model	Predefined data model and schema	No
Searchability	Easy	Difficult
Ease of Analysis	Easy	Difficult
Storage size	Less	More
Interpretation	Precise, ideal for databases and easy fit for machine interpretation	Multiple, depends on human perception
Expressivity	Precise	Variable
Degree of Information Organization	High	Sparse
Examples	RDF, XML, Key-Value	mages, Text, Video

Obstacles with Unstructured Data

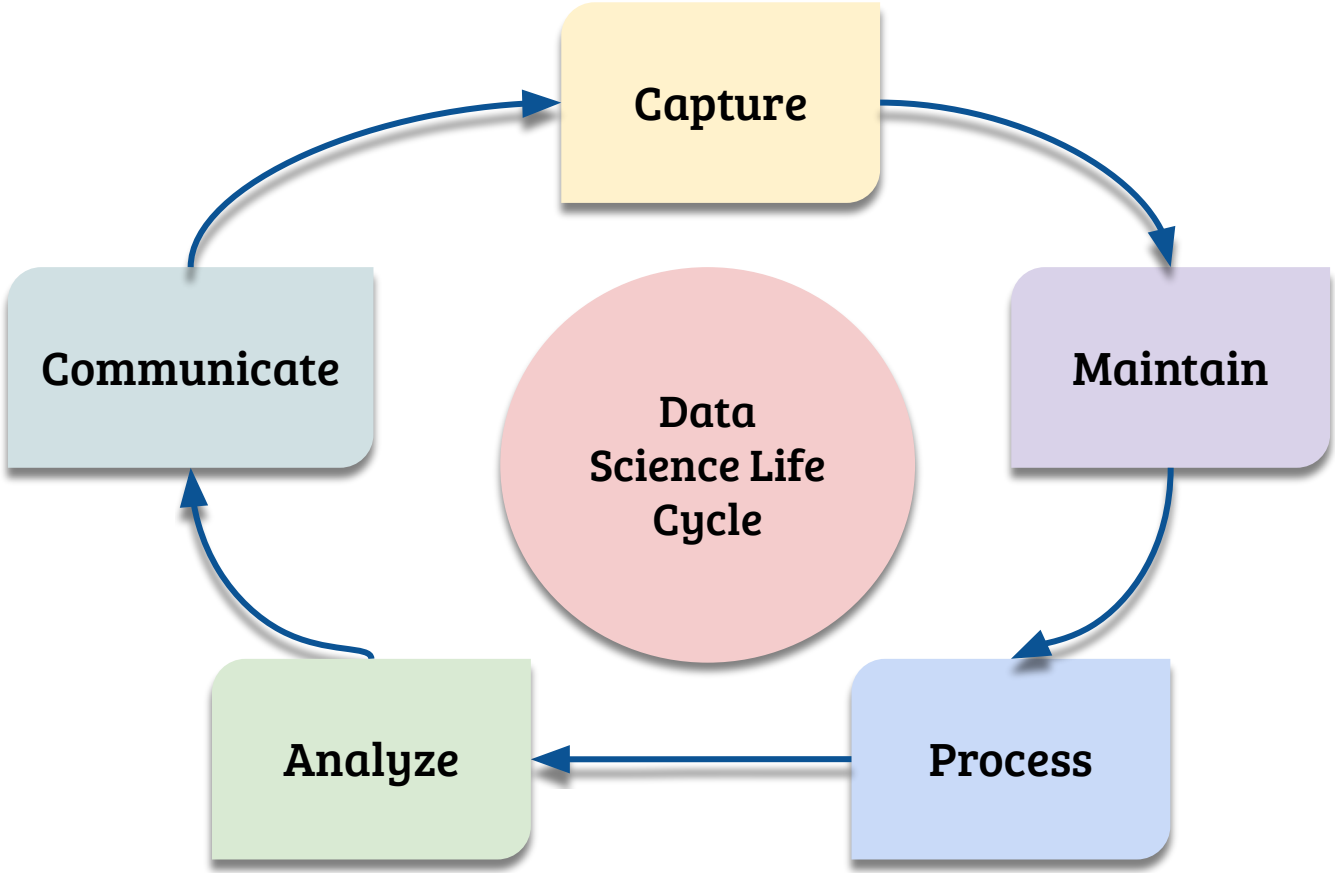
Growth Of Unstructured Data

Protection on identifiable information

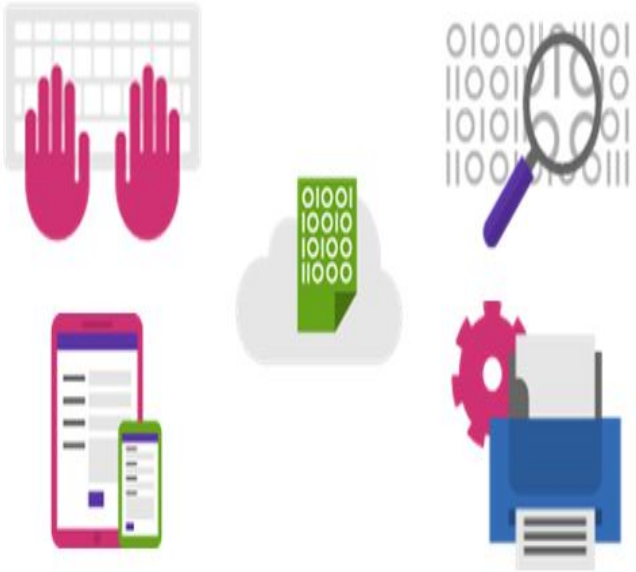
Revolving changes in type

Derive insights

Data Science Life Cycle



Data Capture

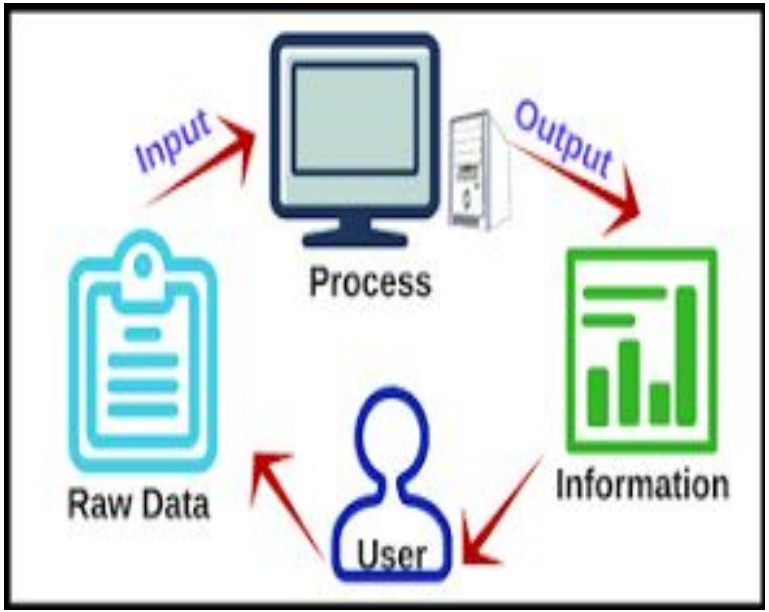


- Data Acquisition,
- Data Entry,
- Signal Reception,
- Data Extraction
- Collection of raw structured and unstructured data.

Data Maintenance

- Data Warehousing
- Data Cleansing
- Data Staging,
- Data Processing
- Data Architecture.
- taking the raw data and putting it in a form that can be used.

Data Process

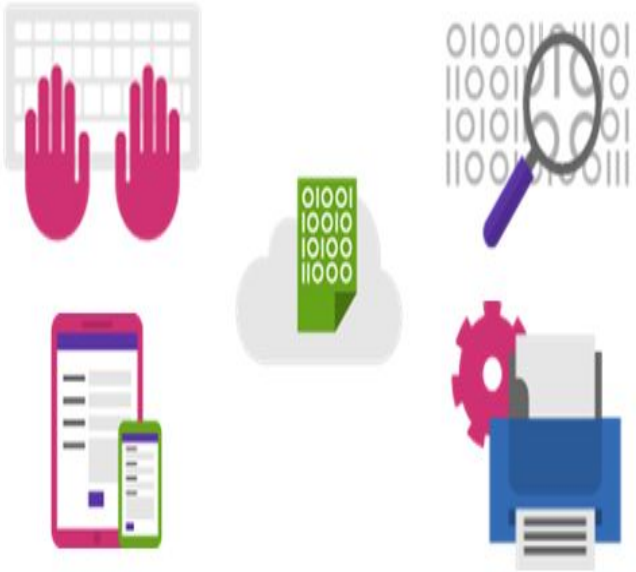


- Data Mining,
- Clustering/Classification,
- Data Modeling,
- Data Summarization.
- examine its patterns, ranges, and biases
- determine how useful it will be in predictive analysis.

Data Analyze

- Exploratory/Confirmatory
- Predictive Analysis
- Regression
- Text Mining
- Qualitative Analysis
- Chief real meat of the life cycle.
- performing the various analyses on the data

Data Capture

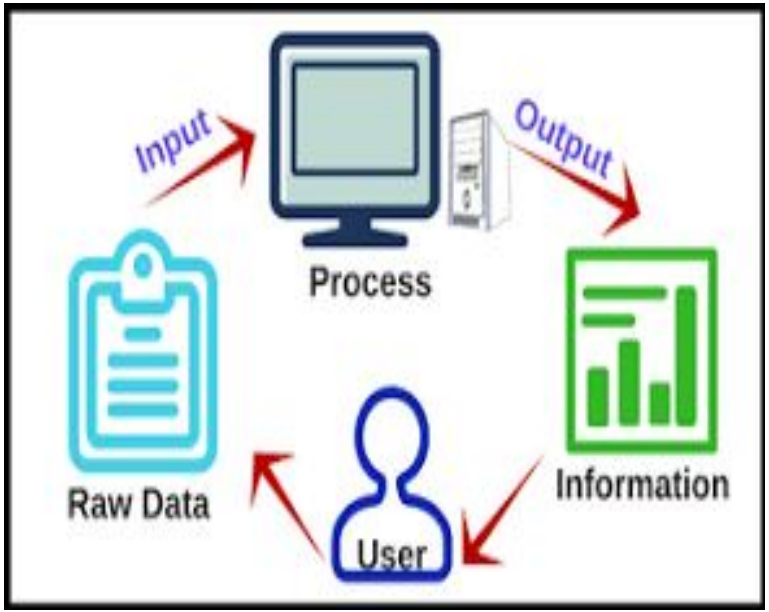


- Data Acquisition,
- Data Entry,
- Signal Reception,
- Data Extraction
- Collection of raw structured and unstructured data.

Data Maintenance

- Data Warehousing
- Data Cleansing
- Data Staging,
- Data Processing
- Data Architecture.
- taking the raw data and putting it in a form that can be used.

Data Process



- Data Mining,
- Clustering/Classification,
- Data Modeling,
- Data Summarization.
- examine its patterns, ranges, and biases
- determine how useful it will be in predictive analysis.

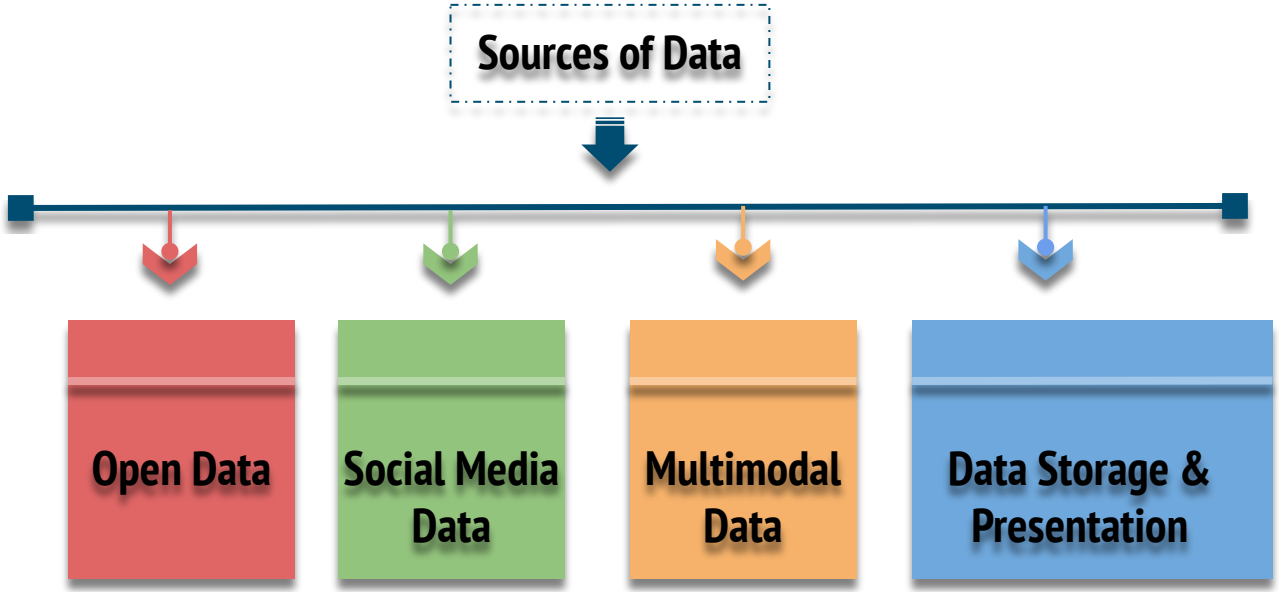
Data Analyze

- Exploratory/Confirmatory
- Predictive Analysis
- Regression
- Text Mining
- Qualitative Analysis
- Chief real meat of the life cycle.
- performing the various analyses on the data

Data Communication

- Data Reporting
- Data Visualization
- Business Intelligence
- Decision Making.
- prepare the analyses in easily readable forms such as charts, graphs, and reports.

- There are many places online to look for sets or collections of data



Open Data

In 2013, The White House developed Project Open Data

restrictions from copyright, patents

freely available in a public domain

Data : Data Collection

Sources of Data

In 2013

- WhiteHouse
- Developed
- Open Data

in

- Freely available
- Public Domain

From

- Restriction
- Copyrights , Patents

Open Data

Public:
- freely available in a public domain,
- without restrictions from copyright, patents

Reusable:
- made available under an open license that places no restrictions on their use.

Accessible:
- made available in convenient, modifiable
- open formats that can be retrieved, downloaded, indexed, and searched.

Complete:
- published in primary forms

Described:
- sufficient information to understand their strengths, weaknesses, analytical limitations, and security requirements

Timely:
- made available as quickly as necessary to preserve the value of the data

Social Data

gold mine for collecting data to analyze for research

facilitated by the Application Programming Interface (API)

API as a set of rule sand methods for asking and sending data

Multimodal Data

Every device is getting connected to the Internet

emerging trend of the Internet of Things(IoT)

generating and using much unstructured data

Sources of Data → **Data Storage & Presentation**

- Depending on its nature, data is stored in various formats.

CSV

XML

Json

TSV

RSS

Sources of Data → **Data Storage & Presentation**

CSV

```
1., Avatar, 18-12-2009, 7.8
2., Titanic, 18-11-1997,
3., Avengers Infinity War, 27-04-2018, 8.5
```

S.No	Movie	Release Date	Ratings (IMDb)
1.	Avatar	18-12-2009	7.8
2.	Titanic	18-11-1997	*Na*
3.	Avengers Infinity War	27-04-2018	8.5

- Most widely adopted
- But What If the data have "," in its content
- Normally "Banana, Orange" will be replaced by "Banana\,Orange

Open Data

Public:
- freely available in a public domain,
- without restrictions from copyright, patents

Reusable:
- made available under an open license that places no restrictions on their use.

Accessible:
- made available in convenient, modifiable
- open formats that can be retrieved, downloaded, indexed, and searched.

Complete:
- published in primary forms

Described:
- sufficient information to understand their strengths, weaknesses, analytical limitations, and security requirements

Timely:
- made available as quickly as necessary to preserve the value of the data

Social Data

gold mine for collecting data to analyze for research

facilitated by the Application Programming Interface (API)

API as a set of rule sand methods for asking and sending data

Multimodal Data

Every device is getting connected to the Internet

emerging trend of the Internet of Things(IoT)

generating and using much unstructured data

Sources of Data → **Data Storage & Presentation**

- Depending on its nature, data is stored in various formats.

CSV

XML

Json

TSV

RSS

Sources of Data → **Data Storage & Presentation**

CSV

```
1., Avatar, 18-12-2009, 7.8
2., Titanic, 18-11-1997,
3., Avengers Infinity War, 27-04-2018, 8.5
```

S.No	Movie	Release Date	Ratings (IMDb)
1.	Avatar	18-12-2009	7.8
2.	Titanic	18-11-1997	*Na*
3.	Avengers Infinity War	27-04-2018	8.5

- Most widely adopted
- But What If the data have "," in its content
- Normally "Banana, Orange" will be replaced by "Banana\,Orange

Sources of Data

Data Storage and Presentation

TSV

```
Name<TAB>Age<TAB>Address
Ryan<TAB>33<TAB>1115 W Franklin
Paul<TAB>25<TAB>Big Farm Way
Jim<TAB>45<TAB>W Main St
Samantha<TAB>32<TAB>28 George St
```

where <TAB> denotes a TAB character.[12]

- if the tab character is present, it may have to be removed.
- On the other hand, TSV is less common than other delimited formats such as CSV.

Sources of Data

Data Storage and Presentation

XML

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
    <book category="information science" cover="hardcover">
        <title lang="en">Social Information Seeking</title>
        <author>Chirag Shah</author>
        <year>2017</year>
        <price>62.58</price>
    </book>
    <book category="data science" cover="paperback">
        <title lang="en">Hands-On Introduction to Data
          Science</title>
        <author>Chirag Shah</author>
        <year>2019</year>
        <price>50.00</price>
    </book>
</bookstore>
```

- (eXtensible Markup Language)
- was designed to be both human- and machine-readable,
- one could write a program, a script, or an app that specifically parses this markup and uses it according to the context.

Sources of Data

Data Storage and Presentation

RSS

- (Really Simple Syndication)
- used to share data between **services, and which was defined in the 1.0 version of XML.**
- follows XML standard usage but in **addition defines the names of specific tags**
- Since RSS data is small and fast loading
- can easily be used with services such as **mobile phones, personal digital assistants (PDAs), and smart watches.**
- RSS is useful for websites that are updated frequently, such as:
  - **News sites** – Lists news with title, date and descriptions.
  - **Companies** – Lists news and new products.
  - **Calendars** – Lists upcoming events and important days.
  - **Site changes** – Lists changed pages or new pages.

Sources of Data

Data Storage and Presentation

RSS

```
<?xml version="1.0" encoding="UTF-8" ?>
<rss version="2.0">
    <channel>
        <title>Dr. Chirag Shah's Home Page</title>
        <link>http://chiragshah.org/</link>
        <description> Chirag Shah's webhome
        </description>
        <item>
            <title>Awards and Honors</title>
            <link>http://chiragshah.org/awards
            .php</link>
            <description>Awards and Honors
            Dr. Shah received</description>
        </item>
    </channel>
</rss>
```

Sources of Data

Data Storage and Presentation

Json

It is not only easy for humans to read and write, b**ut also easy for machines to parse and generate.**

It is based on a subset of the JavaScript Programming Language, Standard ECMA-262, 3rd Edition – December 1999.18

JSON is built on two structures:

- A **collection of name–value pairs**. In various languages, this is realized as an object, record, structure, dictionary, hash table, keyed list, or associative array.
- **An ordered list of values**. In most languages, this is realized as an array, vector, list, or sequence.

Sources of Data

Data Storage and Presentation

Json

```
<!DOCTYPE html>
<html>
<body>
<p id="demo"></p>
<script>
var obj = {"name":"John", "age":25, "state": "New Jersey"};
var obj_JSON = JSON.stringify(obj);
window.location = "json_Demo.php?x=" + obj_JSON;
</script>
</body>
</html>
```

color	director_name	duration	gross	movie_title	anguag	country	budget	title_year	imdb_score
Color	Martin Scorsese	240	116866727	The Wolf of Wall StreetÂ	English	USA	100000000	2013	8.2
Color	Shane Black	195	408992272	Iron Man 3Â	English	USA	200000000	2013	7.2
color	Quentin Tarantino	187	54116191	The Hateful EightÂ	English	USA	44000000	2015	7.9
Color	Kenneth Lonergan	186	46495	MargaretÂ	English	usa	14000000	2011	6.5
Color	Peter Jackson	186	258355354	The Hobbit: The Desolation of SmaugÂ	English	USA	225000000	2013	7.9
	N/A	183	330249062	Batman v Superman: Dawn of JusticeÂ	English	USA	250000000	202	6.9
Color	Peter Jackson	-50	303001229	The Hobbit: An Unexpected JourneyÂ	English	USA	180000000	2012	7.9
Color	Edward Hall	180		RestlessÂ	English	UK		2012	7.2
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
Color	Joss Whedon	173	623279547	The AvengersÂ	English	USA	220000000	2012	8.1
	Tom Tykwer	172	27098580	Cloud AtlasÂ	English	Germany	102000000	2012	-7.5
Color	Null	158	102515793	The Girl with the Dragon TattooÂ	English	USA	90000000	2011	7.8
Color	Christopher Spencer	170	59696176	Son of GodÂ	English	USA	22000000	2014	5.6
Color	Peter Jackson	164	255108370	The Hobbit: The Battle of the Five ArmiesÂ	English	New Zealand	250000000	2014	7.5
Color	Tom Hooper	158	148775460	Les MisÂ©rablesÂ	English	USA	61000000	2012	7.6
Color	Tom Hooper	158	148775460	Les MisÃ©rablesÂ	English	USA	61000000	2012	7.6

Duplicates

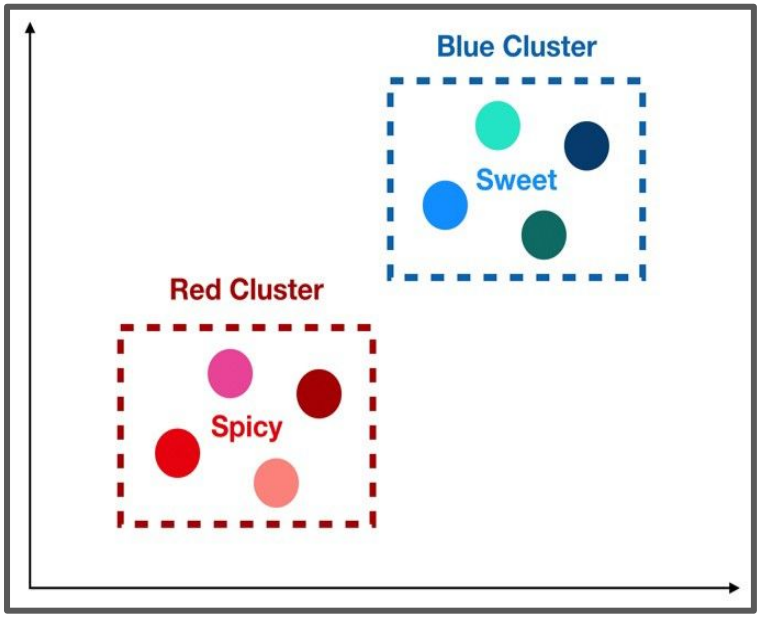
Outliers

Null or bad Characters

Missing Values

Bad Characters

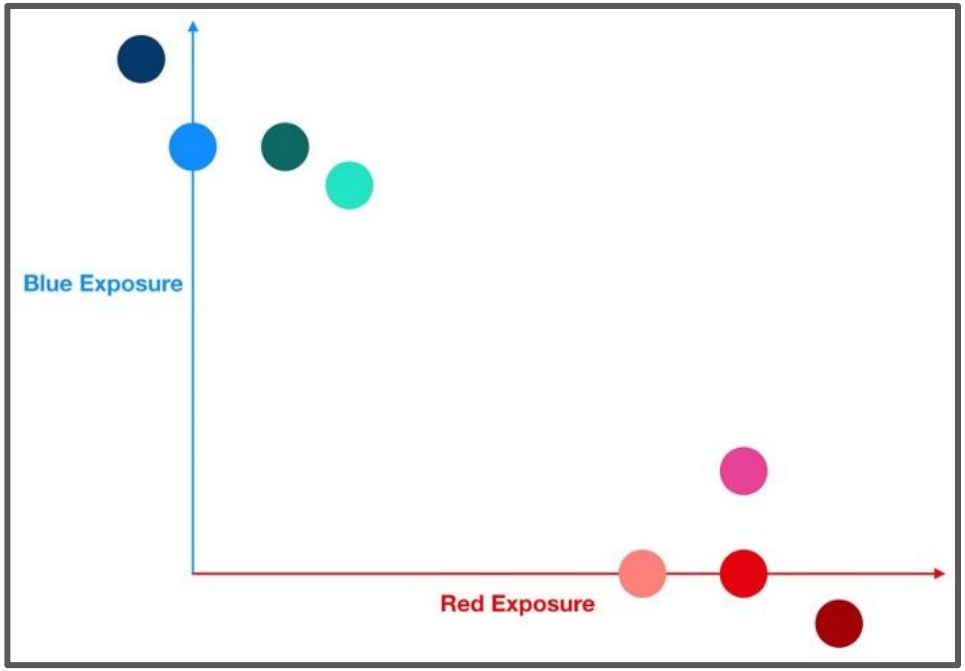
Need of Data Wrangling



	Red	Maroon	Pink	Flamingo	Blue	Turquoise	Seaweed	Ocean
	1	0	0	0	0	0	0	0
	0	1	0	0	0	0	0	0
	0	0	1	0	0	0	0	0
	0	0	0	1	0	0	0	0
	0	0	0	0	1	0	0	0
	0	0	0	0	0	1	0	0
	0	0	0	0	0	0	1	0
	0	0	0	0	0	0	0	1

Need of Data Wrangling

	Red	Blue
Red	1.00	0
Maroon	1.20	-0.10
Pink	1.00	0.20
Flamingo	0.80	0
Blue	0	1.00
Turquoise	0.25	0.90
Seaweed	0.15	1.00
Ocean	-0.10	1.20



Dimensionality Reduction

Need of Data Wrangling

- Duplicates
- Missing Values
- Bad Characters
- Outliers
- Null or bad Characters
- Dimensionality Reduction

Definition

the process of cleaning, organizing, and transforming raw data

In to

Desired format

for

Analyst

For

decision-making

Also known as

data cleaning or data munging, data wrangling

Methods of Data Wrangling

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation
- Data Discretization

Benefits of Data Wrangling

- Data wrangling helps **to improve data usability** as it converts **data into a compatible format** for the end system.
- It helps to quickly **build data flows within an intuitive user** interface and easily schedule and automate the data-flow process.
- **Integrates various types of information** and their sources (like databases, web services, files, etc.)
- Help users **to process very large volumes of data easily** and easily share data-flow techniques.

Data Cleaning

Data Pre-processing

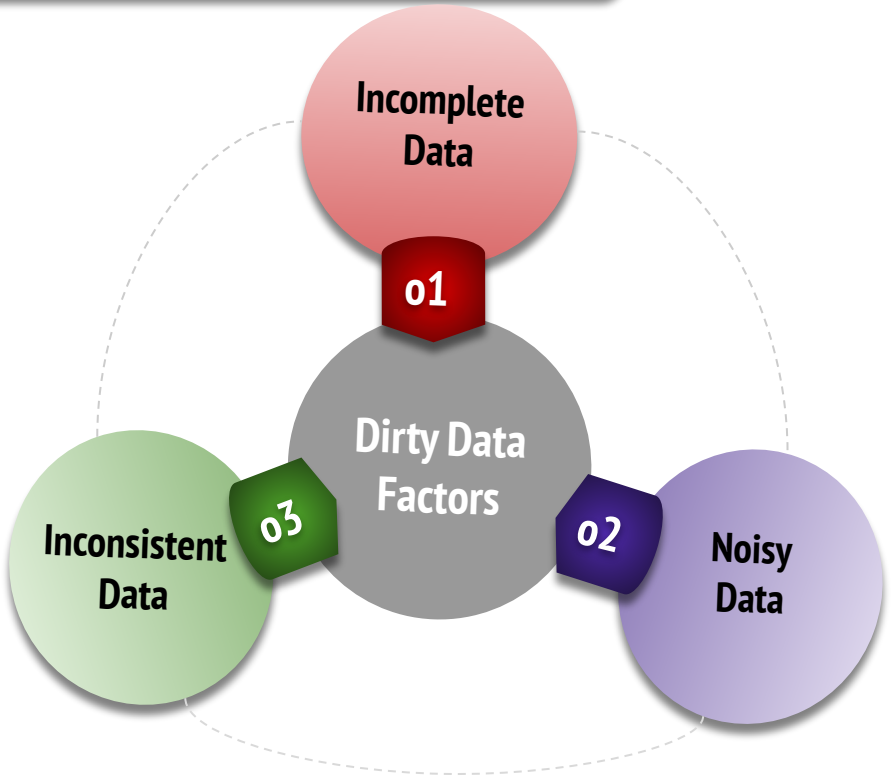
Data Cleaning

- Data in the real world is often dirty; that is, it is in need of being cleaned up before it can be used for a desired purpose. This is often called data pre-processing.

Data Cleaning

Data Pre-processing | Data Cleaning

What makes data “dirty”?



Data Cleaning

Data Pre-processing

Data Cleaning

What makes data "dirty"?

Student ID	Student Name	Age	GPA	Classification
100122014	Joseph	21	3.5	Junior
100232015	Patrick	200	3.2	Sophomore
100122012	Seller	24	3.0	Senior
100342013	Roger	23	234	Senior
100942012	Davis	2.8	3.7	Sophomore
	Travis	23	3.4	Sr
100982015	Alex	27		Sophomore
100982013	Trevor	-22	4.0	Senior
AUC2016XC	Aman	30	3.5	Jr

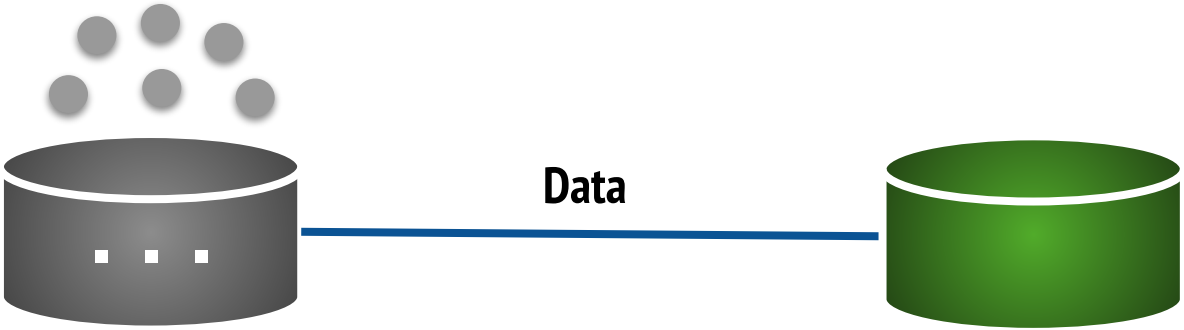
Missing Data

Inconsistent Data

Noisy Data

Data Cleaning

Data Cleaning



Methods of Data Wrangling

Data Cleaning

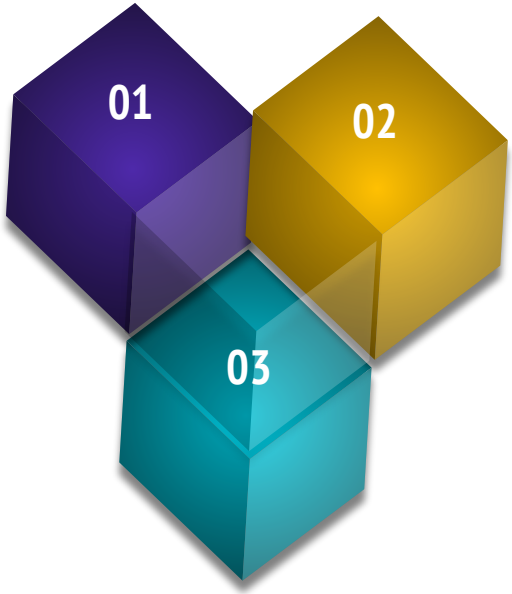
Data Cleaning

Key Methods

Data Munging

Handling Missing Data

Smooth Noisy Data



Data Cleaning

Key Methods

Data Munging

Data Cleaning

Data Cleaning

Key Methods

Data Munging

Data Cleaning

Consider the following text recipe.

"Add two diced tomatoes, three cloves of garlic, and a pinch of salt in the mix."

Table 2.2 Wrangled data for a recipe.

Ingredient	Quantity	Unit/size
Tomato	2	Diced
Garlic	3	Cloves
Salt	1	Pinch

Data Cleaning

Key Methods

Data Cleaning

Reasons for Missing Data

Forgot to answer

Refused to answer

Respondents failed to complete the survey.

Data Cleaning

Key Methods

Internet Connection lost

Network went down

Hard drive corrupt

Data Transfer was cut short

Data Cleaning

Reasons for Missing Data

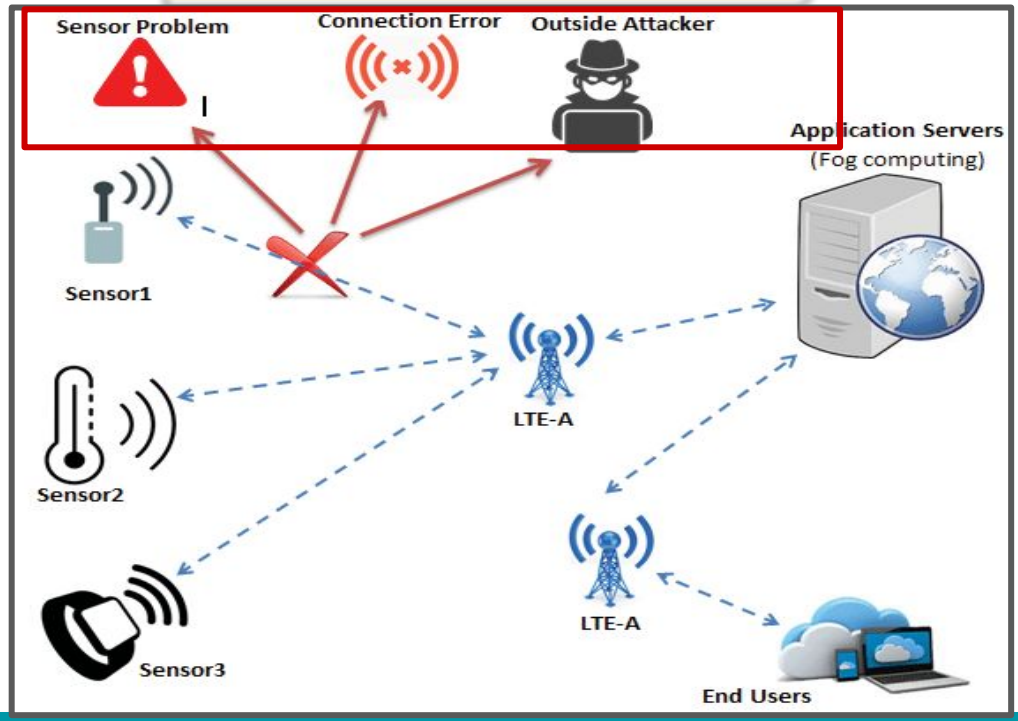
Data Cleaning

Key Methods

- Sensor Failed
- Purposely turned off equipment
- Power Failure
- Change Data Capture Method

Data Cleaning

Reasons for Missing Data



Methods of Data Wrangling

Data Cleaning

Key Methods

Data Cleaning

Types of Missing Data

Missing Completely at Random (MCAR)

Missing at Random (MAR)

Missing not at Random (MNAR)

Reference: Click here

Types of Missing Data

Missing Completely at Random (MCAR)

- There's no relationship between whether a data point is missing and any values in the data set
- The missing data are just a random subset of the data
- The missingness is nothing to do with any other variable .
- By the way , data are rarely MCAR.

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

Reference: Click here

Types of Missing Data

Missing at Random (MAR)

- Missing at Random means the data is missing relative to the observed data.
- It is not related to the specific missing values.
- The data is not missing across all observations but only within sub-samples of the data
- We could easily notice that IQ score is missing for youngsters (<40)

Reference: Click here

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	
29	
30	
30	
31	
44	118
46	93
48	141
51	104
51	116
54	97

Types of Missing Data

Missing not at Random (MNAR)

- It is nor Type I neither Type II , and the data will be missing based on the missing column itself
- The MNAR category applies when the missing data has a structure to it. In other words, there appear to be reasons the data is missing.
- The fact that data are missing on IQ score with only the people having a low score .

Reference: Click here

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	133
26	121
29	
30	
30	110
31	
44	118
46	
48	141
51	
51	116
54	

Methods of Data Wrangling

Data Cleaning

Key Methods

Data Cleaning

Types of Missing Data

Missing Completely at Random (MCAR)

Missing at Random (MAR)

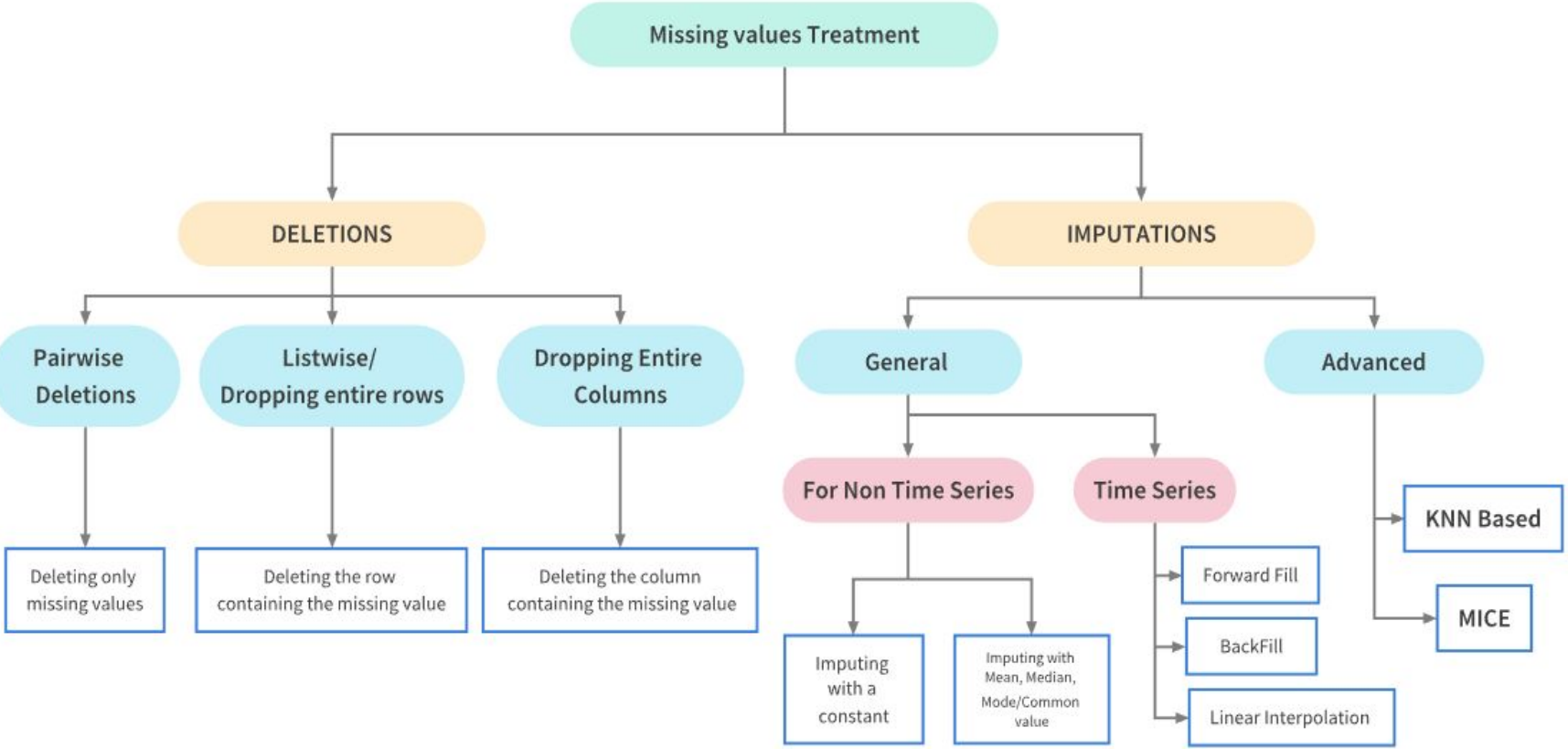
Missing not at Random (MNAR)

Reference: Click here

Methods of Data Wrangling

Data Cleaning

Handling Missing Data



Data Cleaning | Handling Missing Data | Discard Data

1) list-wise deletion

Mobile ID	Mobile Package	Download Speed	Data Limit Usage	
1	Fast+	157	80%	
2	Lite	99	70%	
3	Fast+	167	10%	
4	Fast+	N/A	80%	← Delete
5	Lite	76	70%	
6	Fast+	155	10%	
7	N/A	N/A	95%	← Delete
8	Lite	76	77%	
9	Fast+	180	N/A	← Delete

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
5	Lite	76	70%
6	Fast+	155	10%
8	Lite	76	77%

Data Cleaning | Handling Missing Data | Discard Data

2) Pairwise Deletion

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	N/A	N/A	95%
8	Lite	76	77%
9	Fast+	180	N/A

Delete
Delete
Delete

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+		80%
5	Lite	76	70%
6	Fast+	155	10%
7			95%
8	Lite	76	77%
9	Fast+	180	

Data Cleaning | Handling Missing Data | Discard Data

3) Dropping Variables

Delete

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	N/A	80%
2	Lite	N/A	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	N/A	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	77%

Mobile ID	Mobile Package	Data Limit Usage
1	Fast+	80%
2	Lite	70%
3	Fast+	10%
4	Fast+	80%
5	Lite	70%
6	Fast+	10%
7	Fast+	95%
8	Lite	77%
9	Fast+	77%

Data Cleaning | Handling Missing Data | Retain All Data

1) Mean, Median and Mode

Mean (Download Speed) = 130

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	130	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	130	95%
8	Lite	76	77%
9	Fast+	180	95%

Data Cleaning | Handling Missing Data | Retain All Data

1) Mean, Median and Mode

Median (Download Speed) = 155

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	155	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	155	95%
8	Lite	76	77%
9	Fast+	180	95%

Data Cleaning | Handling Missing Data | Retain All Data

1) Mean, Median and Mode

Mode (Download Speed) = 200

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	N/A	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	N/A	95%
8	Lite	200	77%
9	Fast+	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	200	80%
2	Lite	100	70%
3	Fast+	200	10%
4	Fast+	200	80%
5	Lite	50	70%
6	Fast+	200	10%
7	Fast+	200	95%
8	Lite	200	77%
9	Fast+	180	95%

Data Cleaning | Handling Missing Data | Retain All Data

2) Last Observation Carried Forward (LOCF)

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	N/A	86%
6	6-Jan	155	87%
7	7-Jan	N/A	89%
8	8-Jan	N/A	90%
9	9-Jan	180	92%

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	90	86%
6	6-Jan	155	87%
7	7-Jan	155	89%
8	8-Jan	155	90%
9	9-Jan	180	92%

Data Cleaning | Handling Missing Data | Retain All Data

3) Next Observation Carried Backward (NOCB)

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	N/A	86%
6	6-Jan	155	87%
7	7-Jan	N/A	89%
8	8-Jan	N/A	90%
9	9-Jan	180	92%

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	155	86%
6	6-Jan	155	87%
7	7-Jan	180	89%
8	8-Jan	180	90%
9	9-Jan	180	92%

Data Cleaning | Handling Missing Data | Retain All Data

4) Linear Interpolation

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	N/A	86%
6	6-Jan	150	87%
7	7-Jan	160	89%
8	8-Jan	N/A	90%
9	9-Jan	180	92%

Mobile ID	Date	Download Speed	Data Limit Usage
1	1-Jan	157	80%
2	2-Jan	99	81%
3	3-Jan	167	83%
4	4-Jan	90	84%
5	5-Jan	120	86%
6	6-Jan	150	87%
7	7-Jan	160	89%
8	8-Jan	170	90%
9	9-Jan	180	92%

(90+150)/2 = 120

(160+180)/2 = 170

Data Cleaning | Handling Missing Data | Retain All Data

5) Adding a category to capture NA

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	N/A	99	70%
3	Fast+	167	10%
4	Fast+	90	80%
5	Lite	76	70%
6	N/A	155	10%
7	Fast+	200	95%
8	Lite	76	77%
9	N/A	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Missing	99	70%
3	Fast+	167	10%
4	Fast+	90	80%
5	Lite	76	70%
6	Missing	155	10%
7	Fast+	200	95%
8	Lite	76	77%
9	Missing	180	95%

Data Cleaning | Handling Missing Data | Retain All Data

6) Frequent category imputation

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	N/A	99	70%
3	Fast+	167	10%
4	Fast+	90	80%
5	Lite	76	70%
6	N/A	155	10%
7	Fast+	200	95%
8	Lite	76	77%
9	N/A	180	95%

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Fast+	99	70%
3	Fast+	167	10%
4	Fast+	90	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	200	95%
8	Lite	76	77%
9	Fast+	180	95%

Data Cleaning | Handling Missing Data | Retain All Data

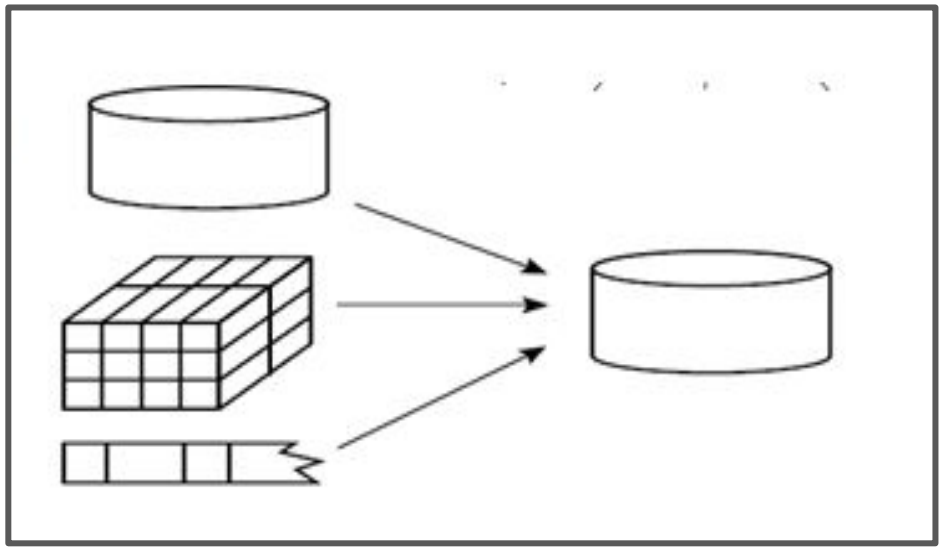
7) Arbitrary Value Imputation

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%

Arbitrary value 999

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	999	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	999	95%
8	Lite	76	77%
9	Fast+	180	95%

Data Integration



- It involves combining data from multiple heterogeneous data sources into a coherent data store
- provide a unified view of the data
- These sources may include multiple data cubes, databases, or flat files.

Data Integration

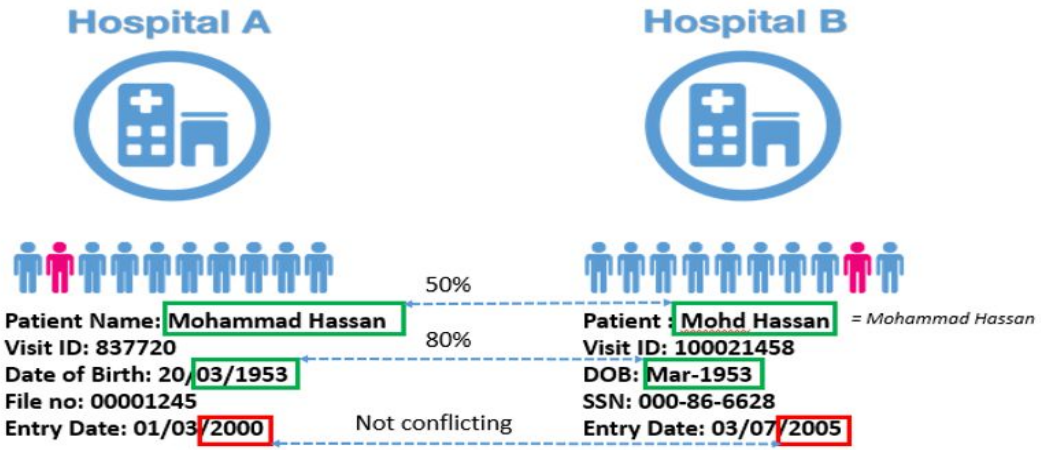
Issues in Data Integration:

Patient Name	Mohammad Hassan
Visit ID	837720
Date of Birth	20/03/1953
File No	00001245
Entry Date	01/03/2000

Table 3 - Record from Hospital A

Patient	Mohd Hassan
Visit ID	100021458
DOB	Mar-1953
SSN	000-86-6628
Entry Date	03/07/2005

Table 4 - Record from Hospital B



Data Integration

2 major approaches

Tight Coupling:

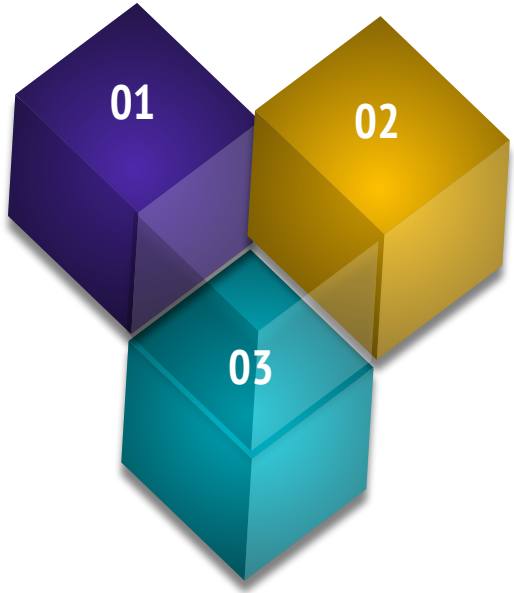
- Here, a **data warehouse is treated as an information retrieval component.**
- In this coupling, data is combined from different **sources into a single physical location through the process of ETL** – Extraction, Transformation, and Loading

Loose Coupling:

- Here, an interface is **provided that takes the query from the user, transforms it in a way the source database can understand, and then sends the query directly to the source databases to obtain the result.**
- And the d**ata only remains in the actual source databases.**

Data Integration

3 Steps of Data Integration:



- Schema Integration
- Data Value conflicts
- Handling Redundancy

Data Integration

3 Steps of Data Integration:

Schema Integration

Data Source 1

Cust_ ID	Cust_Name	DOB	Cust_Type	Discount
cust_1	Nisha	1991	Gold	20.00%
cust_2	Pooja	1992	Silver	10.00%
cust_3	Ankur	1991	silver	10.00%
cust_4	Shraddha	1996	Gold	20.00%
cust_5	Raj	1990	Silver	10.00%

Data Source 2

Cust_ Num	Cust_Name	Year	Cust_Type	Discount
cust_1	Ronak	31	Permanent	Free Lunch
cust_2	Rushi	26	Permanent	Free Lunch
cust_3	Rakhi	31	Temp	Free Breakfast
cust_4	Pooja	30	Temp	Free Breakfast
cust_5	Priya	32	Temp	Free Breakfast

Issues: Same Feature may **have different names in different databases**

Solution : Provide **metadata for each Feature**

MetaData:

- Name
- Meaning
- Data type
- Range of values permitted for the attribute

Data Integration

3 Steps of Data Integration:

Data value conflicts

Data Source 1

Cust_ ID	Cust_Name	DOB	Cust_Type	Discount
cust_1	Nisha	1991	Gold	20.00%
cust_2	Pooja	1992	Silver	10.00%
cust_3	Ankur	1991	silver	10.00%
cust_4	Shraddha	1996	Gold	20.00%
cust_5	Raj	1990	Silver	10.00%

Data Source 2

Cust_ Num	Cust_Name	Year	Cust_Type	Discount
cust_1	Ronak	31	Permanent	Free Lunch
cust_2	Rushi	26	Permanent	Free Lunch
cust_3	Rakhi	31	Temp	Free Breakfast
cust_4	Pooja	30	Temp	Free Breakfast
cust_5	Priya	32	Temp	Free Breakfast

Issues:

- **Feature Values** from **different sources are different**
- **Different Units , representation , scaling**

Solution : Methods of Data Cleaning

Data Integration

3 Steps of Data Integration:

Handling Redundancy

Data Source 1

Cust_ ID	Cust_Name	DOB	Cust_Type	Discount
cust_1	Nisha	1991	Gold	20.00%
cust_2	Pooja	1992	Silver	10.00%
cust_3	Ankur	1991	silver	10.00%
cust_4	Shraddha	1996	Gold	20.00%
cust_5	Raj	1990	Silver	10.00%

Data Source 2

Cust_ Num	Cust_Name	Year	Cust_Type	Discount
cust_1	Ronak	31	Permanent	Free Lunch
cust_2	Rushi	26	Permanent	Free Lunch
cust_3	Rakhi	31	Temp	Free Breakfast
cust_4	Pooja	30	Temp	Free Breakfast
cust_5	Priya	32	Temp	Free Breakfast

- Cust_Name Matching
- DOB Converted to Age (After Conversion 2022-1992= 30 Years) Matching
- Cust_Type silver in Data Source 1=Temp in Data Source 2
- Discount 10% in Data Source 1= Temp in DataSource 2
- It may be identical

Correlation Analysis

Data Integration

3 Steps of Data Integration:

1. **Schema Integration:**

- Integrate metadata from different sources.
- The real-world entities from multiple sources are matched referred to as the **entity identification problem**

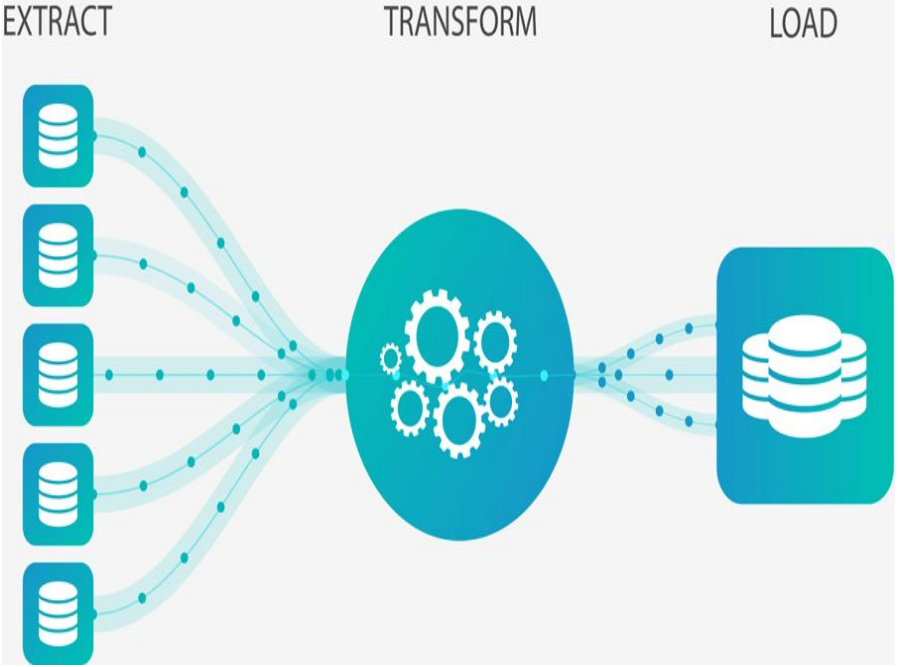
2. **Redundancy:**

- An attribute may be redundant **if it can be derived or obtaining from another attribute or set of attributes**.
- **Inconsistencies in attributes** can also cause **redundancies** in the resulting data set.

3. **Detection and resolution of data value conflicts:**

- Attribute values from different sources **may differ for the same real-world entity.**
- An attribute in **one system may be recorded at a lower level abstraction than the "same" attribute in another**

Data Transformation



- converting **data from one format to another**,
- typically from the format of **a source system into the required format of a destination system.**
- This entire process **is known as ETL (Extract, Load, Transform).**
- During the extraction phase, data is identified and pulled from many different locations or sources into a single repository.

5 Processes of Data Transformation

- Data Smoothing
- Data Aggregation
- Generalization
- Normalization
- Attribute or Feature Construction

Data Transformation

Data Smoothing

Year	Crimes
2000	14
2001	12
2002	15
2003	14
2004	14
2005	10
2006	44
2007	44
2008	12
2009	14
2010	12

- Crime Feature Values <= 15
- Only for 2016 nd 2017>44

- Case 1: Average **including** 2016 and 2017 is **19**
- Case 2: Average **excluding** 2016 and 2017 is **13**

- Difference is 32%

- By Observation Case 2 is realistic Average

Data Transformation

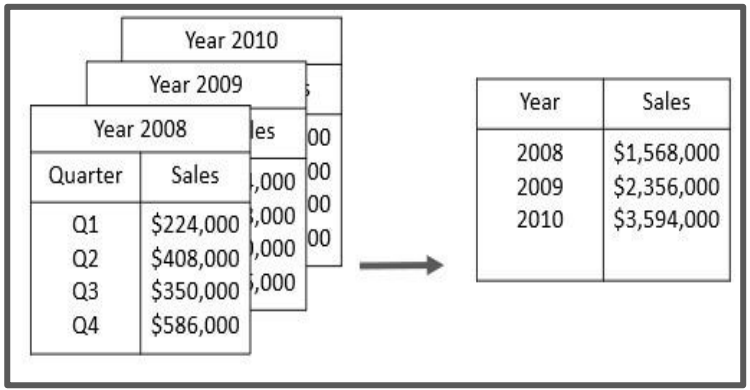
Data Smoothing

- Statistical approach of **removing noise from datasets**
- To make the **patterns more noticeable**
- It is achieved **using algorithms** to eliminate statistical noise from datasets.
- **The random method, simple moving average, random walk, simple exponential, and exponential moving average are some of the methods used for data smoothing.**
- The use of data smoothing can help **forecast patterns**, such as those seen in share prices, n identifying trends in businesses, financial securities, and the economy.

Reference

Data Transformation

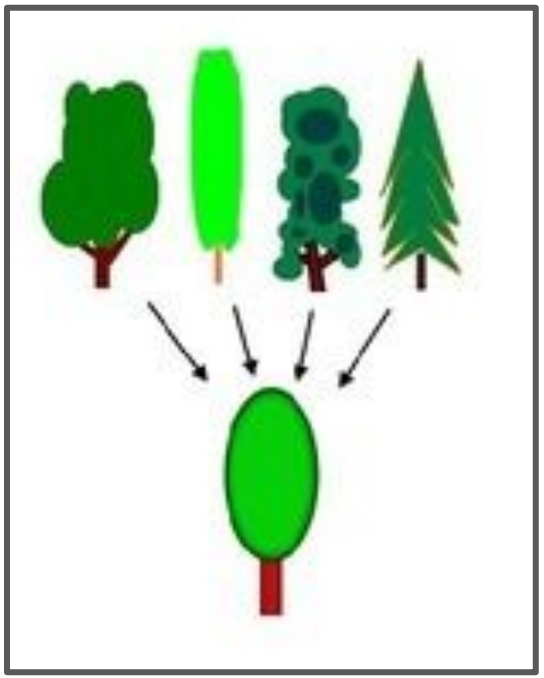
Data Aggregation



- The method of **storing and presenting data in a summary format.**
- The data may be obtained from **multiple data sources** to integrate these data sources into a data analysis description.
- The accuracy of **data analysis insights is highly dependent on the quantity and quality of the data** used.
- Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results

Data Transformation

Generalization



- It converts **low-level data attributes to high-level data attributes** using **concept hierarchy**.
- For Example Age initially in Numerical form **(22, 45) is converted into categorical value (young, old)**.
- For example, Categorical attributes, such as **house addresses, may be generalized to higher-level definitions, such as town or country.**

Data Transformation

Normalization

Data Transformation

Normalization

- Scaled to **fall within a small**, specified range and aggregation.
- Some of the techniques that are used for accomplishing normalization:
  - **Min-Max Normalization:**
  - **Z-Score Normalization:**
  - **Decimal Scaling**

Data Transformation | Normalization

Min-Max Normalization:

- It is an operation which **rescales a set of data.**
- **This can be useful when:**
  - Comparing data from two different scales
  - Converting data to a new scale
- In most situations, data is normalized to a fit a target range of [0, 1]
- The **smallest value** in the original set **would be mapped to 0**
- The **largest value** in the original set **would be mapped to 1**
- **Every other value would be mapped** to a value somewhere **between these two bounds**

Data Transformation | Normalization

Min-Max Normalization

Normalization Formula

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

https://www.codecademy.com/article/normalization
https://www.upgrad.com/blog/normalization-in-data-mining/

Data Transformation | Normalization

Example for Min-Max Normalization :

- X = [7, 21, 13,15]
- xmin=7
- xmax=21
- xnew= (7-7)/(21-7)
- Repeat for all the remaining values of x
  [0/14,1, 6/14,8/14]

Normalization

Z-Score Normalization

Refers to

Process

of

Normalizing every value

In a

Dataset

such that

- the mean of all of the values is 0
- the standard deviation is 1.

Formula ➡ **New value = (x – μ) / σ**

where:

- **x**: Original value
- **μ**: Mean of data
- **σ**: Standard deviation of data

Standard Deviation

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Normalization

Example

Consider the dataset

Data
3
5
5
8
9
12
12
13
15
16
17
19
22
24
25
134

Solution

1. the mean of the dataset is **21.2**
2. the standard deviation is **29.8.**
3. To perform a z-score normalization on the first value in the dataset

- New value = $(x - \mu) / \sigma$
- New value = (3 – 21.2) / 29.8

First value in dataset | Mean | Standard Deviation

- New value = -0.61

❑ We can use this formula to perform a z-score normalization on every value in the dataset

Normalization

z-score normalization on every value in the dataset

Data	Z-Score Normalized Value
3	-0.61
5	-0.54
5	-0.54
8	-0.44
9	-0.41
12	-0.31
12	-0.31
13	-0.28
15	-0.21
16	-0.17
17	-0.14
19	-0.07
22	0.03
24	0.09
25	0.13
134	3.79

- The mean of the normalized values is **0**
- The standard deviation of the normalized values is **1**.

- The normalized values represent the number of standard deviations that the original value is from the mean.

Example

- The first value in the dataset is **0.61** standard deviations below the mean.

Methods of Data Wrangling : Data Transformation

Normalization

z-score normalization on every value in the dataset

Data	Z-Score Normalized Value
3	-0.61
5	-0.54
5	-0.54
8	-0.44
9	-0.41
12	-0.31
12	-0.31
13	-0.28
15	-0.21
16	-0.17
17	-0.14
19	-0.07
22	0.03
24	0.09
25	0.13
134	3.79

The benefit of performing this type of normalization

Is that

clear outlier

Has been

Transformed

In such way that

No longer Massive Outlier

In the dataset

- If we then use this dataset to fit some type of machine learning model, the outlier will no longer have as big of an influence that it might have on the model fit.

Reference

Normalization

Decimal Scaling

- Decimal scaling is a data normalization technique like Z score, Min-Max, and normalization with standard deviation.
- In this technique, we move the decimal point of values of the attribute.
- This movement of decimal points totally depends on the maximum value among all values in the attribute.

Normalization

Decimal Scaling

Example : Open Spreadsheet

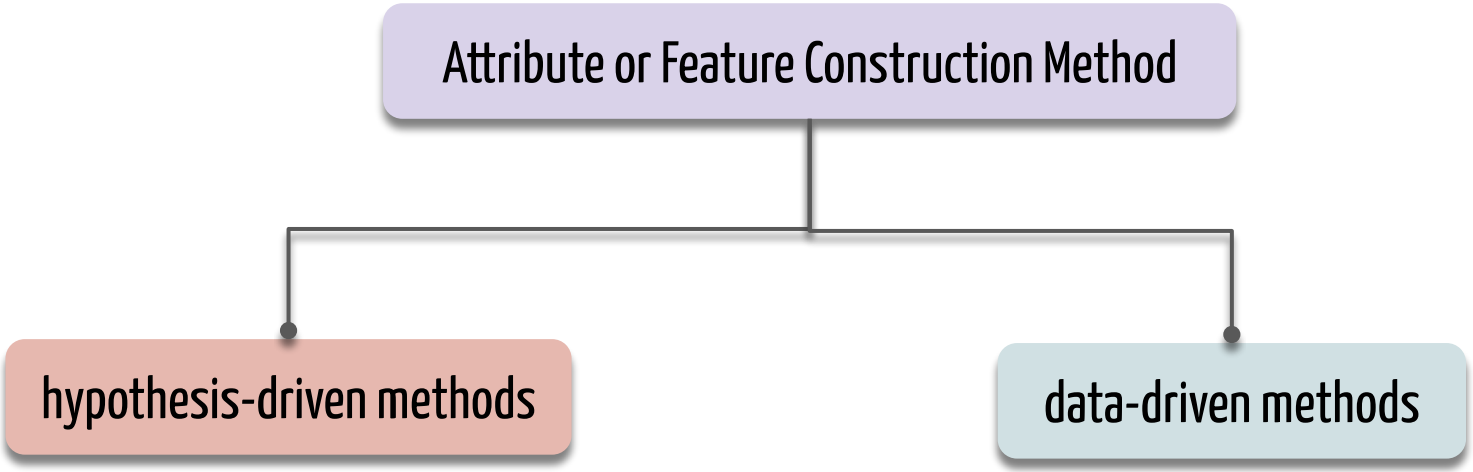
Data Transformation

Attribute or Feature Construction

Example : Open Spreadsheet

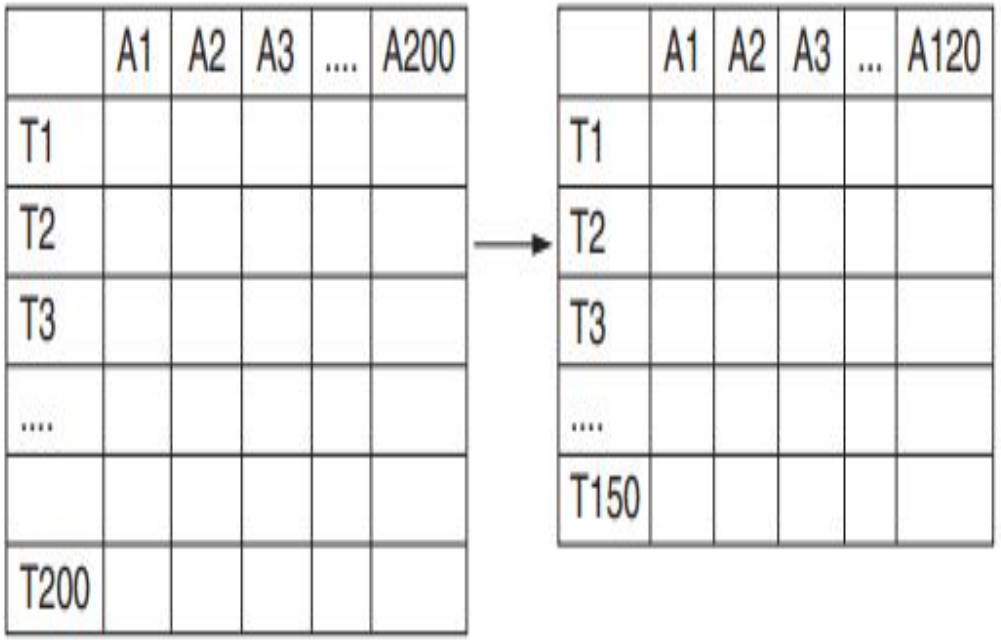
- **New attributes are constructed and added** from the given set of attributes to help the mining process.
- The goal of an attribute construction method is **to construct new attributes out of the original ones**
- Transforming the original data representation into a new one **where regularities in the data are more easily detected by the classification algorithm**
- which tends to improve the predictive accuracy.

Attribute or Feature Construction



Data Reduction

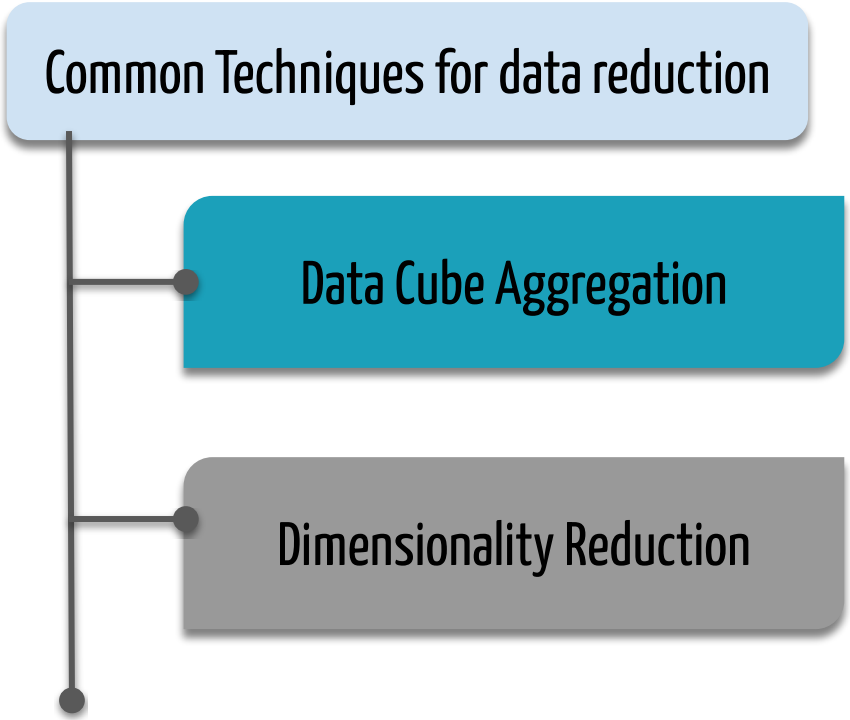
- The data set that is much smaller in volume
- maintains the integrity of the original data.



	A1	A2	A3	....	A200
T1					
T2					
T3					
....					
T200					

	A1	A2	A3	...	A120
T1					
T2					
T3					
....					
T150					

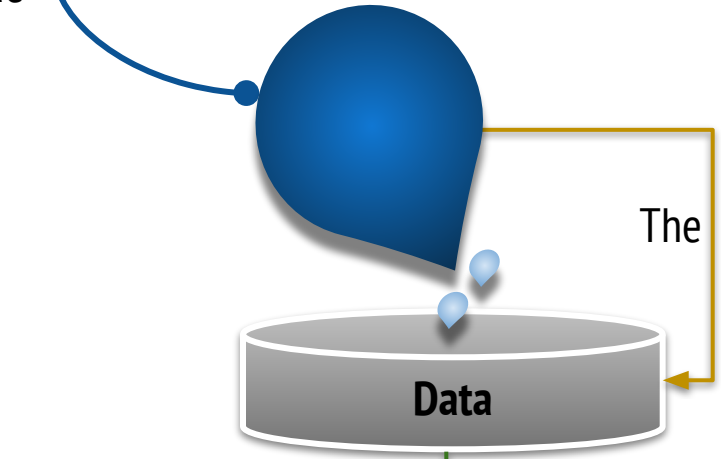
Data Reduction



Methods of Data Wrangling : Data Reduction

Data Cube Aggregation

Technique used to



Simpler Form

It is

Process

Gather information | summarized into smallest volume

Data Cube Aggregation

- This technique is used to aggregate data in a simpler form.
- Aggregation operations are applied to the data in the construction of a data cube.

Example

quarterly sales for the year 2010 to 2012 → Open Spreadsheet

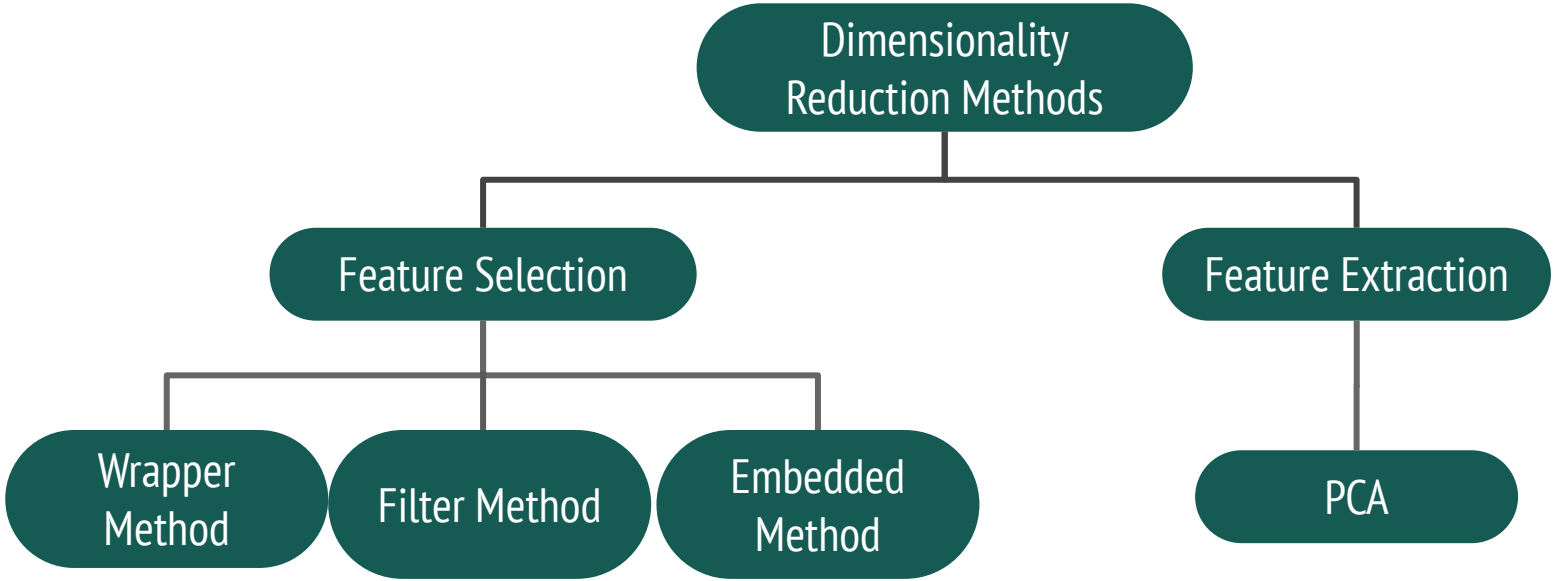
- summarizes the total sales per year instead of per quarter. It summarizes the data.

Data Reduction | Dimensionality Reduction

- In dimensionality reduction redundant attributes are **detected and removed which reduce the data set size.**
- we come across any data which is **weakly important, then we use the attribute required for our analysis.**
- It reduces **data size as it eliminates outdated or redundant features.**

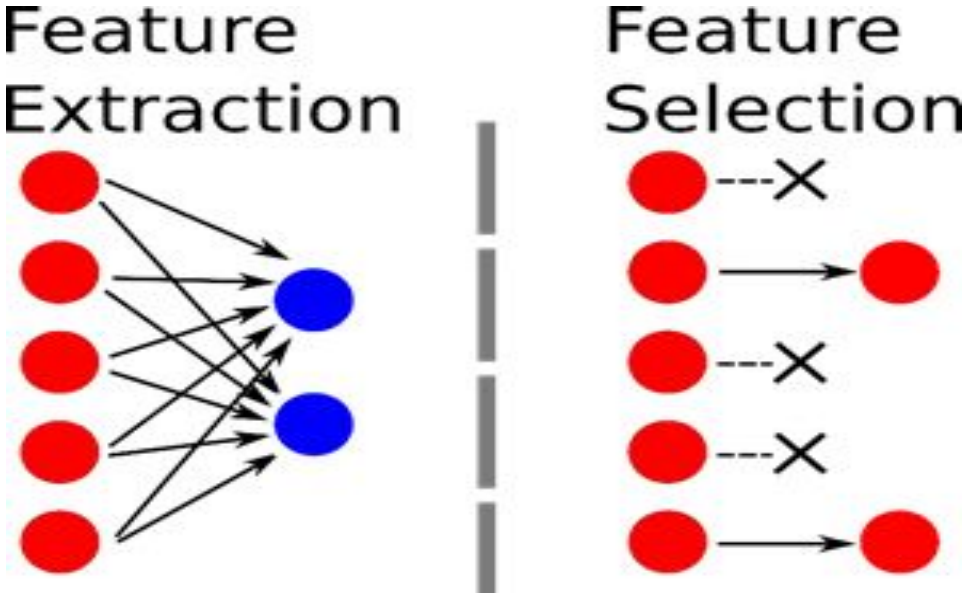
Data Reduction

Dimensionality Reduction



Data Reduction

Dimensionality Reduction



Data Reduction

Dimensionality Reduction

Feature Selection

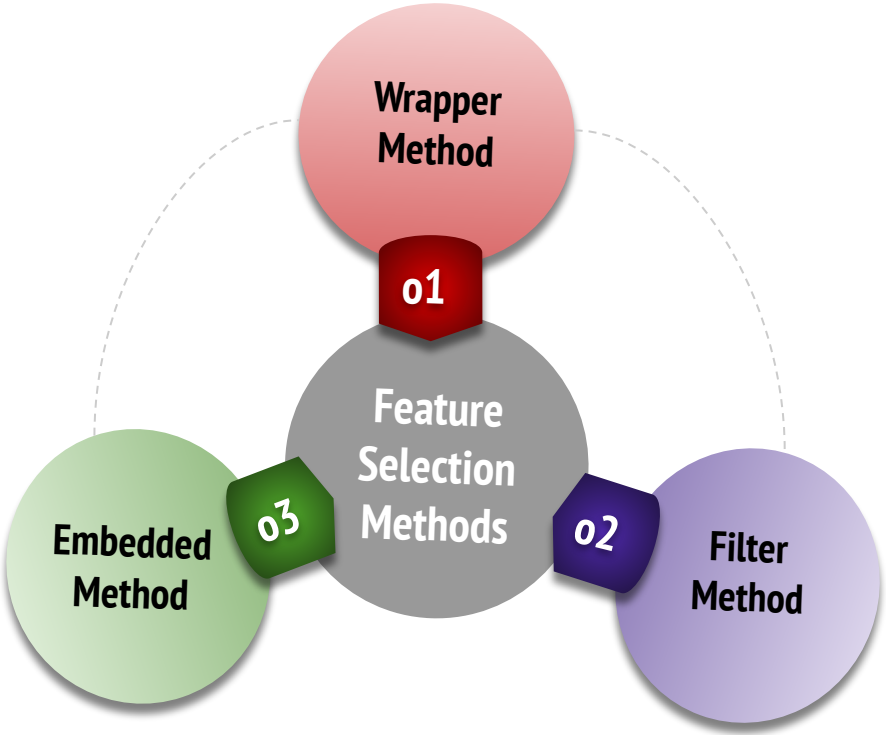
- Here, we select a subset of features from the original feature set.

Feature Extraction

- With this technique, we generate a new feature set by extracting and combining information from the original feature set.

Data Reduction

Dimensionality Reduction



Data Reduction | Dimensionality Reduction

01 Wrapper Method

- Wrapper methods iterate through different combinations of features and perform a model retrain on each.
- The feature combination which resulted in the best model performance metric (accuracy) is selected.

Data Reduction

Dimensionality Reduction

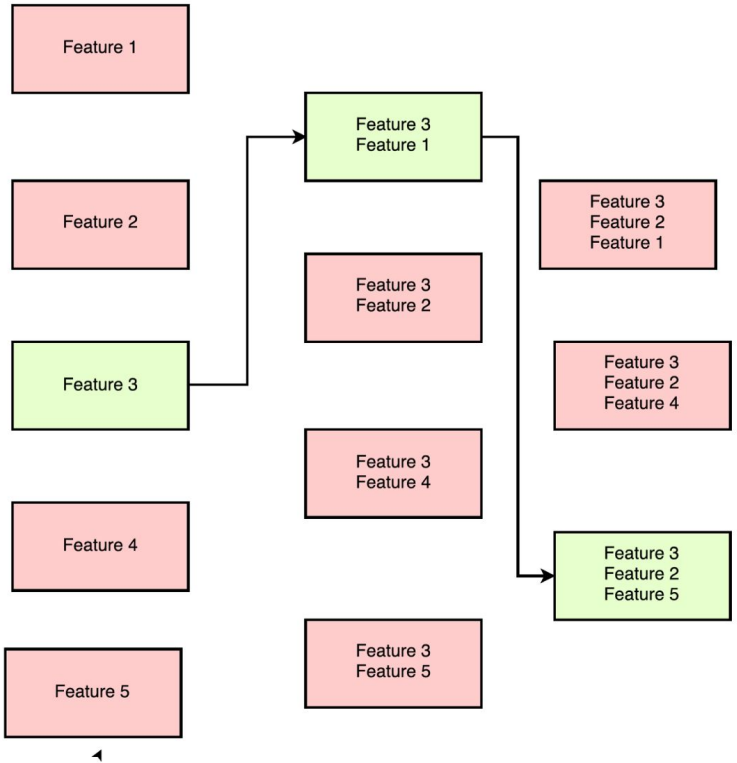
01 Wrapper Method

Data Reduction

Dimensionality Reduction

01 **Wrapper Method**

- Forward Selection



Data Reduction

Dimensionality Reduction

Example

fitness level prediction

01 **Wrapper Method**

- Forward Selection

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Data Reduction

Dimensionality Reduction

Example
fitness level prediction

01 **Wrapper Method**

- Forward Selection

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 87%

Data Reduction

Dimensionality Reduction

Example
fitness level prediction

01 **Wrapper Method**

- Forward Selection

ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 80%

Data Reduction

Dimensionality Reduction

Example
fitness level prediction

01 **Wrapper Method**

- Forward Selection

ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 85%

Methods of Data Wrangling

Data Reduction

Dimensionality Reduction

Example

fitness level prediction

01 **Wrapper Method**

- Forward Selection

ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 88%

Methods of Data Wrangling

Data Reduction

Dimensionality Reduction

Example
fitness level prediction

01 **Wrapper Method**

- Forward Selection

Variable used	Accuracy
Calories_burnt	87.00%
Gender	80.00%
Plays_Sport?	85.00%

Accuracy = 91%

Plays_Sport gives us a better accuracy when we combined it with the Calories_Burnt. Hence we will retain that and select it in our model.

Data Reduction

Dimensionality Reduction

Example

fitness level prediction

01

Wrapper Method

- Forward Selection

ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 91%

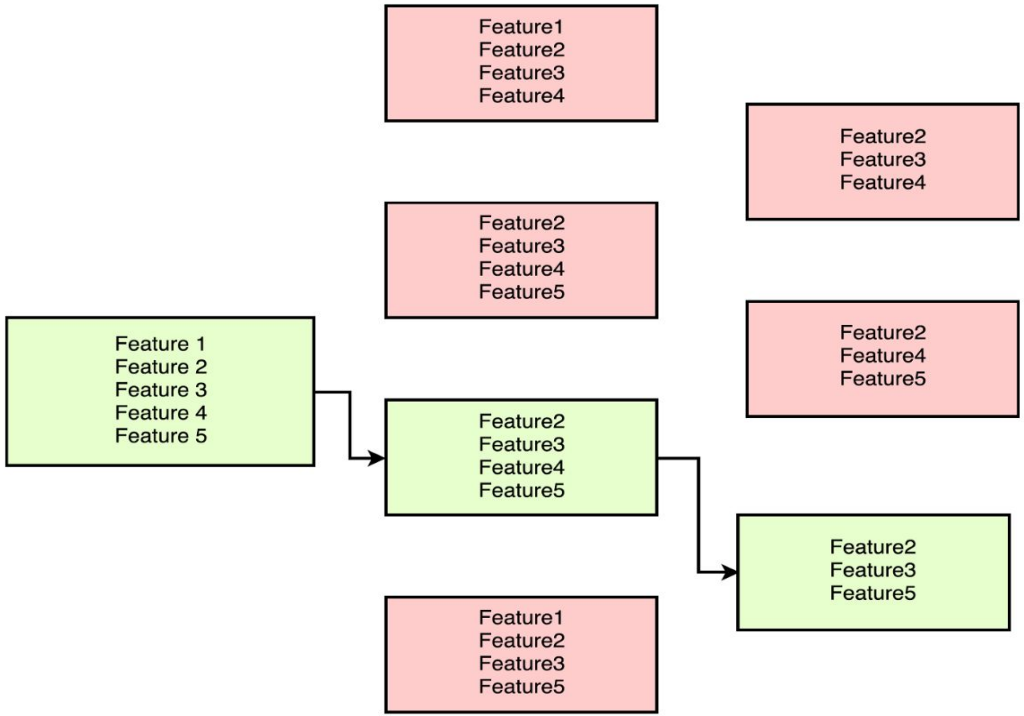
Methods of Data Wrangling

Data Reduction

Dimensionality Reduction

01 **Wrapper Method**

- Backward Elimination



Reference

Data Reduction

Dimensionality Reduction

Example
fitness level prediction

01 Wrapper Method

Backward Elimination

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy of 92%

Reference

Data Reduction

Dimensionality Reduction

Example
fitness level prediction

01 Wrapper Method

Backward Elimination

ID	Calories_burnt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Drop Gender Variable

Accuracy of 91.6%

Reference

Data Reduction

Dimensionality Reduction

Example
fitness level prediction

01 Wrapper Method

Backward Elimination

ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Drop Plays_Sports Variable
Accuracy of 88%

Reference

Data Reduction

Dimensionality Reduction

Example
fitness level prediction

01 Wrapper Method

- Backward Elimination

Accuracy using all the variables = 92%

Variable_dropped	Accuracy
Calories_burnt	90%
Gender	91.60%
Plays_Sport?	88%

Gender does not have a high impact on the Fitness_Level variable. And hence it can be dropped

Reference

Data Reduction

Dimensionality Reduction

01 Wrapper Method

- **Forward Selection**
  - ❑ Begin with zero features in your model and iteratively add the next most predictive feature until no additional performance is reached with the addition of another feature.
- **Backward Selection**
  - ❑ Begin with all features in your model and iteratively remove the least significant feature until performance starts to drop.

Note : ➡ These methods are typically less preferred as they take a long time to compute and tend to overfit.

Data Reduction

Dimensionality Reduction

02 **Filter Method**

Variance thresholds

- The dataset is filtered, and a subset that contains only the relevant features is taken.
- Some common techniques of filters method are:
  - Correlation
  - Chi-Square Test
  - ANOVA
  - Information Gain

Data Reduction

Dimensionality Reduction

02

Filter Method

ID	season	holiday	workingday	weather	f5	temp	atemp	humidity	windspeed	count
AB101	1	0	0	1	7	9.84	14.395	81	0.0000	16
AB102	1	0	0	1	7	9.02	13.635	80	0.0000	40
AB103	1	0	0	1	7	9.02	13.635	80	0.0000	32
AB104	1	0	0	1	7	9.84	14.395	75	0.0000	13
AB105	1	0	0	1	7	9.84	14.395	75	0.0000	1
AB106	1	0	0	2	7	9.84	12.880	75	6.0032	1
AB107	1	0	0	1	7	9.02	13.635	80	0.0000	2
AB108	1	0	0	1	7	8.20	12.880	86	0.0000	3
AB109	1	0	0	1	7	9.84	14.395	75	0.0000	8
AB110	1	0	0	1	7	13.12	17.425	76	0.0000	14

Reference

Data Reduction

Dimensionality Reduction

02 **Filter Method**

- Variance thresholds

Variance = 0

ID	season	holiday	workingday	weather	f5	temp	atemp	humidity	windspeed	count
AB101	1	0	0	1	7	9.84	14.395	81	0.0000	16
AB102	1	0	0	1	7	9.02	13.635	80	0.0000	40
AB103	1	0	0	1	7	9.02	13.635	80	0.0000	32
AB104	1	0	0	1	7	9.84	14.395	75	0.0000	13
AB105	1	0	0	1	7	9.84	14.395	75	0.0000	1
AB106	1	0	0	2	7	9.84	12.880	75	6.0032	1
AB107	1	0	0	1	7	9.02	13.635	80	0.0000	2
AB108	1	0	0	1	7	8.20	12.880	86	0.0000	3
AB109	1	0	0	1	7	9.84	14.395	75	0.0000	8
AB110	1	0	0	1	7	13.12	17.425	76	0.0000	14

Data Reduction

Dimensionality Reduction

02 Filter Method

- Variance thresholds

- ❑ The variables **with low variance** have **less impact on the target variable.**
- ❑ **a threshold value of variance** can be set
- ❑ if the variance of **a variable is less than that threshold,** drop that variable
- ❑ it's very important, **Variance is range-dependen**t,
- ❑ **do normalization before applying this technique.**

Data Reduction | Dimensionality Reduction

02 **Filter Method**

- **Do not depend on machine learning algorithms**.
- Instead, features are selected on the **basis of their scores in various statistical tests** for their **correlation with the outcome variable.**

Data Reduction | Dimensionality Reduction

02 **Filter Method**

❑ **After Unit 2**

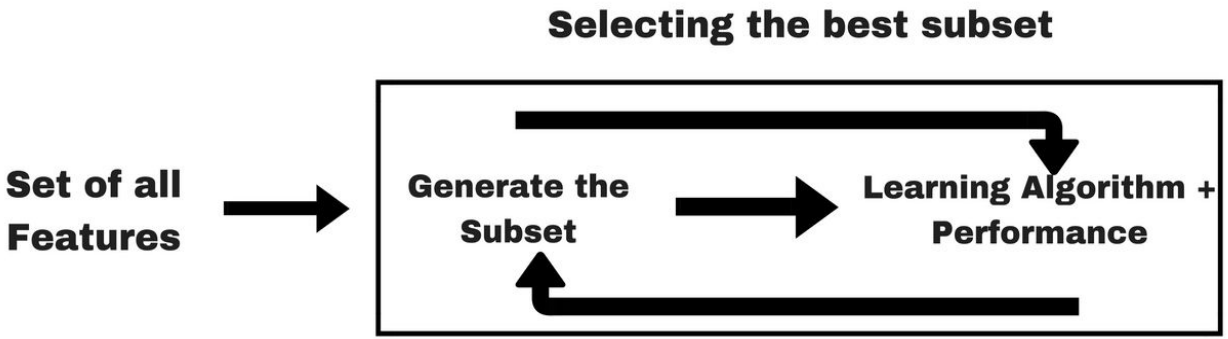
Feature\Response	Continuous	Categorical
Continuous	Pearson's Correlation	LDA
Categorical	Anova	Chi-Square

Data Reduction

Dimensionality Reduction

03 Embedded Method

- Combine the **qualities' of filter and wrapper methods.**
- It's implemented by **algorithms** that have their **own built-in feature selection methods**



Methods of Data Wrangling

Data Discretization

We are dealing with data

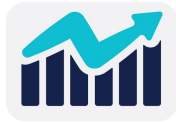
That are

collected from processes

That are

Continues

Such as



Stock Prices

Ambient Light

Data Discretization

- Data discretization converts a large number of data values into smaller once
- so that data evaluation and data management becomes very easy.

Transformation of a continuous attribute to categorical attribute requires two subtasks.

i.) To decide how many categories to have.
ii.) How to map values to these categories.

Data Discretization

The two types of discretization are:

i.) Unsupervised Discretization

ii.) Supervised Discretization

Data Discretization

i.) Unsupervised Discretization

- This type of discretization considers only the attribute being discretized and does not use class information.

ii.) Supervised Discretization

- This type of discretization considers the class value and will divide the continuous attributes in such a way so that it provides maximum information about the class.

Data Discretization

i.) UnSupervised Discretization

Example

- **Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28
- **Equal width**
  - Bin 1: 0, 4 [-,10)
  - Bin 2: 12, 16, 16, 18 [10,20)
  - Bin 3: 24, 26, 28 [20,+)
- **Equal frequency**
  - Bin 1: 0, 4, 12 [-, 14)
  - Bin 2: 16, 16, 18 [14, 21)
  - Bin 3: 24, 26, 28 [21,+)

Data Discretization

i.) UnSupervised Discretization

Example

- **Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28
- **Equal width**
  - Bin 1: 0, 4 [-,10)
  - Bin 2: 12, 16, 16, 18 [10,20)
  - Bin 3: 24, 26, 28 [20,+)
- **Equal frequency**
  - Bin 1: 0, 4, 12 [-, 14)
  - Bin 2: 16, 16, 18 [14, 21)
  - Bin 3: 24, 26, 28 [21,+)

Data Discretization

i.) UnSupervised Discretization

:Example : Equal Width Binning

Given data: $0, 4, 12, 16, 16, 18, 24, 26, 28$		
Number of Bins: 3		
Minimum and Maximum Attribute: $a_{min} = 0,\ a_{max} = 28$		
Interval of bin: $b_i = \frac{28-0}{3} = 9.33 \approx 10$		
Bin 1 $: [0 - 10]$	**Bin** 2 $: [10 - 20]$	**Bin** 3 $: [20 - 30]$
$0,4$	$12,16,16,18$	$24,26,28$

Data Discretization

i.) UnSupervised Discretization

:Example : Equal Frequency Binning

Given data:0, 4, 12, 16, 16, 18, 24, 26, 28		
Number of Bins: 3		
Bin 1	Bin 2	Bin 3
0,4,12	16,16,18	24,26,28

Data Discretization

ii.) Supervised Discretization

Entropy Based Binning

- Entropy based approach is the **most commonly used discretization measures**.
- This methods targets to find out the **best split for given attributes**.
- The spilt will be **recursively calculated for each bin and the split having maximum information gain** will be treated as best split.
- Here the bins are **more promising , means that the majority of the data attributes in a particular bins will be associated with the same class label**.

Data Discretization

ii.) Supervised Discretization

The formula to calculate entropy, Information and Gain are:

$$Entropy(S1) = -\sum_{i=1}^{m} p_i * \log\ p_i$$

$$where\ p_i\ is\ probablity\ \ of\ \ class\ i\ \ in\ S1$$

$$I(S1, S2) = \frac{|S1|}{S} Entropy(S1) + \frac{|S2|}{S} Entropy(S2)$$

$$where\ S1\ and\ S2\ are\ two\ samples\ in\ S$$

$$Gain(v, S) = \ Entropy(S) - Information$$

Data Discretization

ii.) Supervised Discretization

Example ➡ Entropy Based Binning

- In table ??, the attributes are associated with the class label:P and N. the entropy and information gain is calculated for two different possibilities.

Data Discretization

Example ➡ Entropy Based Binning

Given data: $(0, P), (4, P), (12, P), (16, N), (16, N), (18, P), (24, N), (26, N), (28, N)$	
S denotes the given instances $S = 9$	
Fraction of P Pairs	Fraction of N Pairs
p=$\frac{4}{9}$	n=$\frac{5}{9}$
$Entropy = -\frac{4}{9}\log\frac{4}{9} - \frac{5}{9}\log\frac{5}{9} = 0.991$	
Let V be a possible split. It is a mid point of given instances.	
S is divided into 2 parts: S1 and S2	
For S1 attribute value $\leq$V	For S2 attribute value >V

Data Discretization

Consider V=14 or 16	
For V=14	
S1=(0,P),(4,P),(12,P)	S2=(16,N),(16,N),(18,P),(24,N),(26,N),(28,N)
$I(S1, S2) = \frac{3}{9}Entropy(S1) + \frac{6}{9}Entropy(S2)$	
$= 0 + \frac{6}{9} * 0.65 = 0.433$	
$Entropy(S1) = 0,\ All\ Instances\ belongs\ to\ same\ class$	
$Gain(14, S) = Entropy(S) - 0.433 = 0.558$	

Data Discretization

For V=16	
S1=(0,P),(4,P),(12,P),(16,N),(16,N)	S2=,(18,P),(24,N),(26,N),(28,N)
$I(S1,S2) = \frac{5}{9}Entropy(S1) + \frac{4}{9}Entropy(S2)$ $= \frac{5}{9} * 0.97 + \frac{6}{9} * 0.81 = 0.89$ $Gain(14,S) = Entropy(S) - 0.89 = 0.101$	
The process will repeat for all possible split values. Here I(s1,S2) is maximum for V=14, So best split is at 14.	
Bin 1:	**Bin 2:**
(0,P),(4,P),(12,P)	(16,N),(16,N),(18,P),(24,N),(26,N),(28,N)