Data Science & Big Data Analytics

Subject Code: 310251

T. E. Computer (2019 Pattern)

Dr. S.R.Khonde

Course Objectives

Course Objectives:

- To understand the need of Data Science and Big Data
- To understand computational statistics in Data Science
- To study and understand the different technologies used for Big Data processing
- To understand and apply data modeling strategies
- To learn Data Analytics using Python programming
- To be conversant with advances in analytics

Course Outcomes

Course Outcomes:

After completion of the course, learners should be able to

CO1: Analyze needs and challenges for Data Science Big Data Analytics

CO2: Apply statistics for Big Data Analytics

CO3: Apply the lifecycle of Big Data analytics to real world problems

CO4: Implement Big Data Analytics using Python programming

CO5: Implement data visualization using visualization tools in Python programming

CO6: Design and implement Big Databases using the Hadoop ecosystem

UNIT I

Unit I	Introduction to Data Science and Big Data	07 Hours
Basics and need of Data Science and Big Data, Applications of Data Science, Data explosion, 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.		
#Exemplar/Case Studies	Create academic performance dataset of students and perform data pre-processing using techniques of data cleaning and data transformation.	
***Mapping of Course Outcomes for Unit I**	CO1	

Introduction to Data Science

Definition :

Data Science can be defined as the study of data in terms of,

- where it comes from,
- what it represents,
- the ways by which it can be transformed into valuable inputs and
- resources to create business and IT strategies.

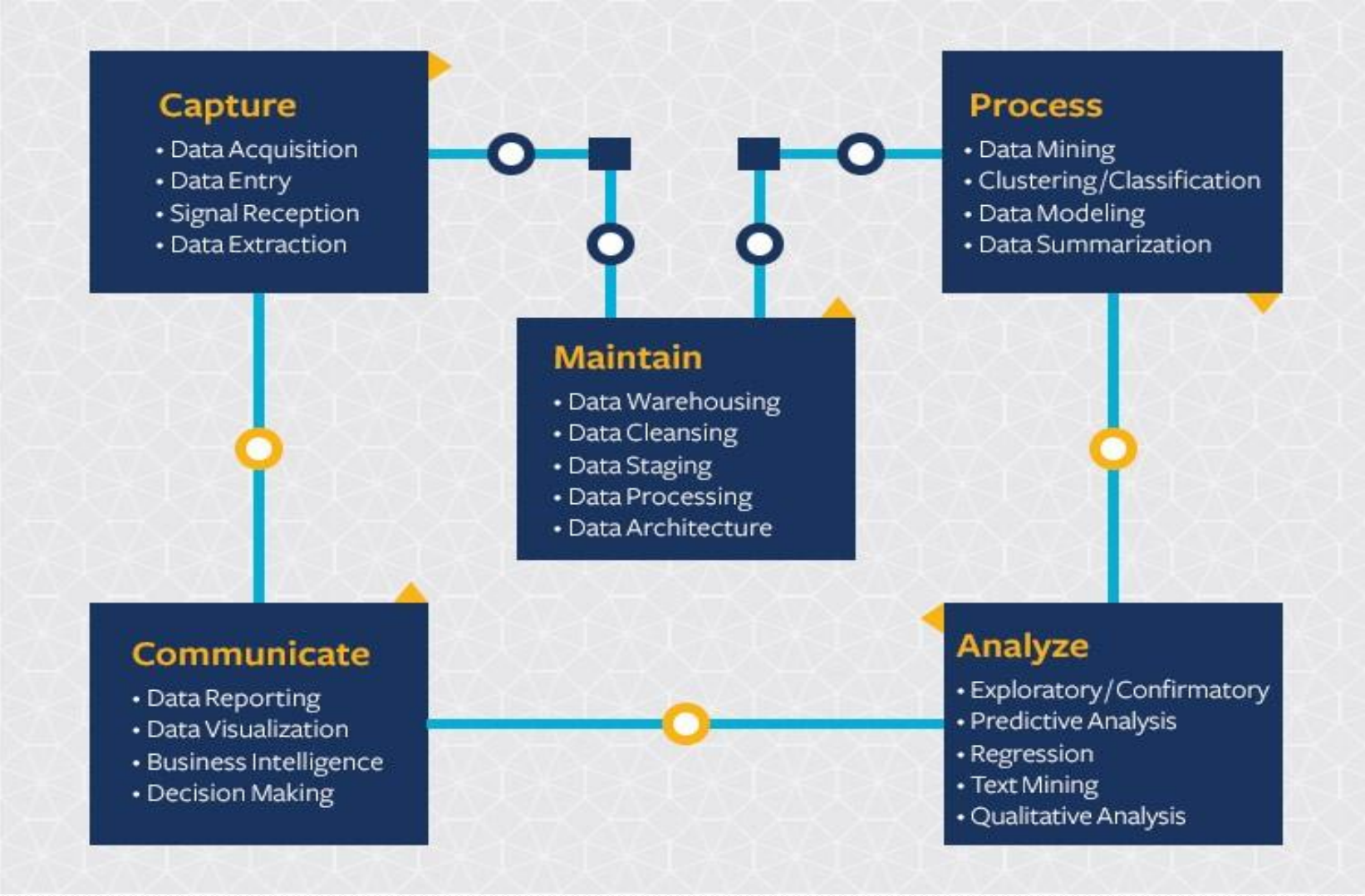
Introduction to Data Science

Definition :

Data science in simple words can be defined as an **interdisciplinary field of study** that **uses data** for various **research** and **reporting,** to **derive insights** and **meaning** out of that data.

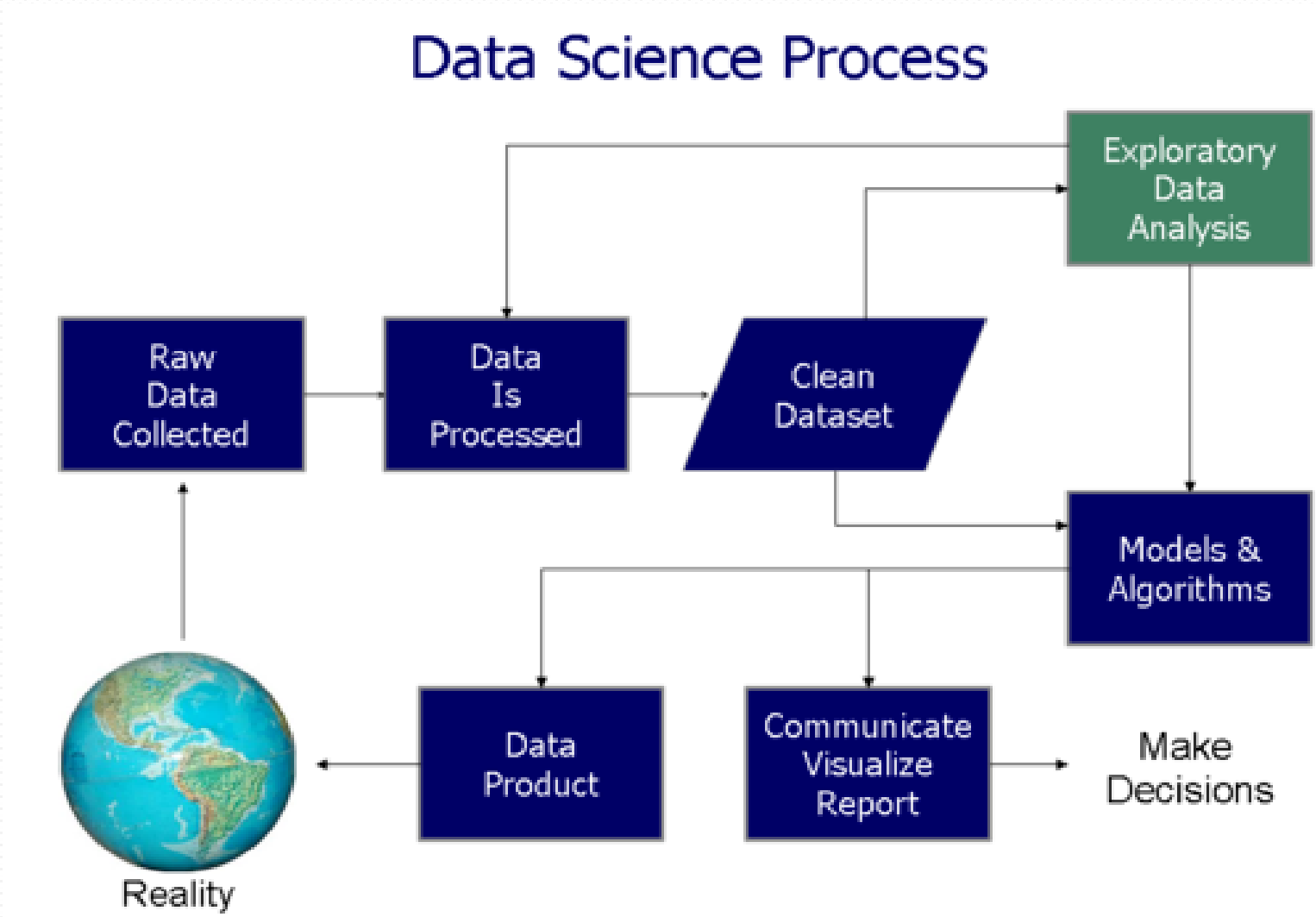
Data science requires a mix of different **skills** including **statistics**, **business acumen**, **computer science**, and more.

Introduction to Data Science

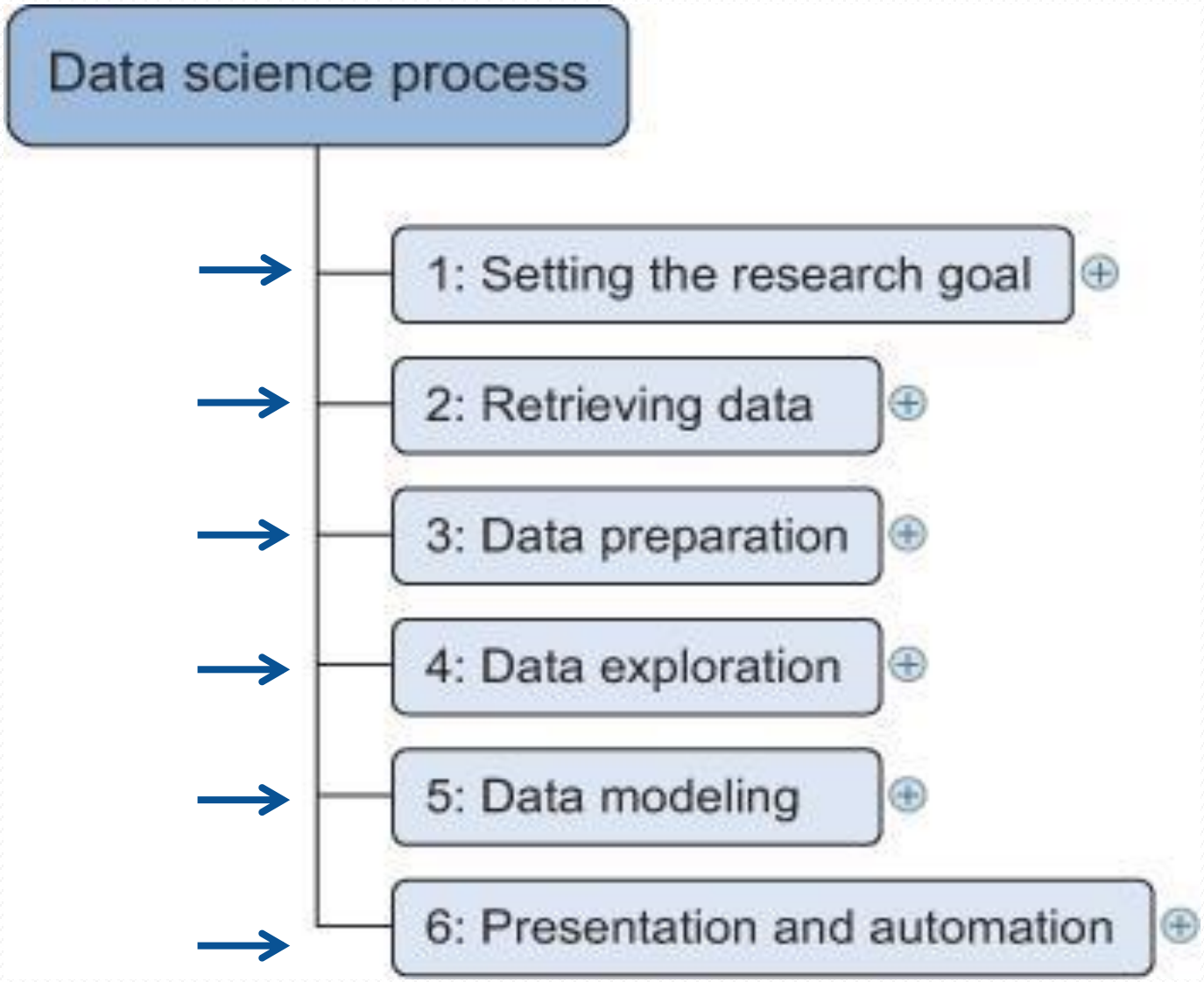


Application to Data Science

Data Science Process



Data Science Process



Data Science Process

Step 1: Frame the problem (Setting the research goal)

Before you solve a problem you have to define exactly what it is. You need to be able to translate data questions into something actionable Say you're solving a problem for the VP Sales of your company.
You should ask questions like the following:

1. Who are the customers?
2. Why are they buying our product?
3. How do we predict if a customer is going to buy our product?
4. What is different from segments who are performing well and those that are performing below expectations?
5. How much money will we lose if we don't actively sell the product to these groups?

Data Science Process

Step 2: Collect the raw data needed for your problem (Retrieving data)

Once you've defined the problem, you'll need data to give you the insights needed to turn the problem around with a solution.

This part of the process involves thinking through what data you'll need and finding ways to get that data, whether it's querying internal databases, or purchasing external datasets.

Data Science Process

Step 3: Process the data for analysis (Data Preparation)

Now that you have all of the raw data, you'll need to process it before you can do any analysis.
Oftentimes, data can be quite messy, especially if it hasn't been well-maintained. You'll see errors that will corrupt your analysis, common errors:

- Missing values, perhaps customers without an initial contact details, missing date of birth
- Corrupted values, such as invalid entries
- Timezone differences, perhaps your database doesn't take into account the different timezones of your users
- Date range errors, perhaps you'll have dates that makes no sense, such as data registered from before sales started

Data Science Process

Step 4: Explore the data

Steps in Data Exploration and Pre-processing:

1. Identification of variables and data types.
2. Analysing the basic metrics.
3. Univariate Analysis.
4. Bivariate Analysis.
5. Variable transformations.
6. Missing value treatment.
7. Outlier treatment.

Data Science Process

Step 5: Data Modelling (Perform in depth analysis)

This step of the process is where you're going to have to apply your statistical, mathematical and technological knowledge and leverage all of the data science tools at your disposal to crunch the data and find every insight you can.

In this case, you might have to create a predictive model that compares your underperforming group with your average customer. You might find out that the age and social media activity are significant factors in predicting who will buy the product.

Data Science Process

Step 6: Presentation and Automation

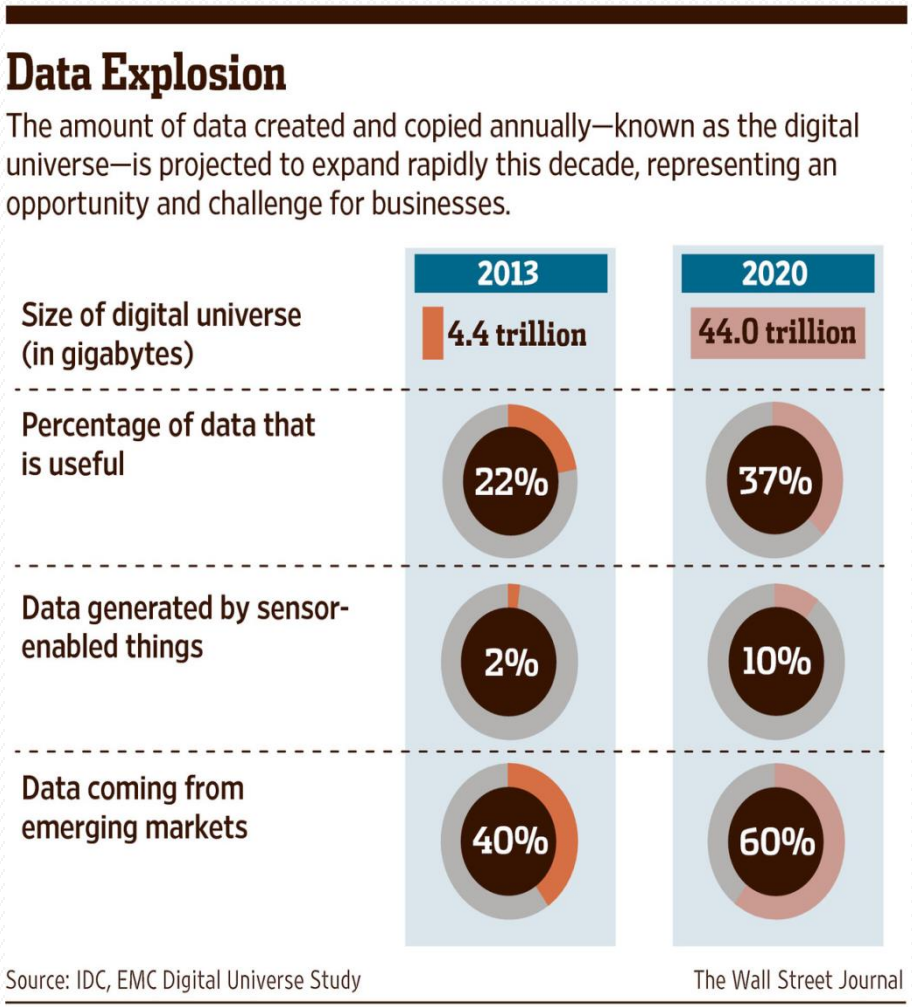
It's important that the VP Sales understand why the insights you've uncovered are important. Ultimately, you've been called upon to create a solution throughout the data science process. Proper communication will mean the difference between action and inaction on your proposals.

Data Explosion

The rapid or exponential increase in the amount of data that is generated and stored in the computing systems, that reaches level where data management becomes difficult, is called ***"Data Explosion".***

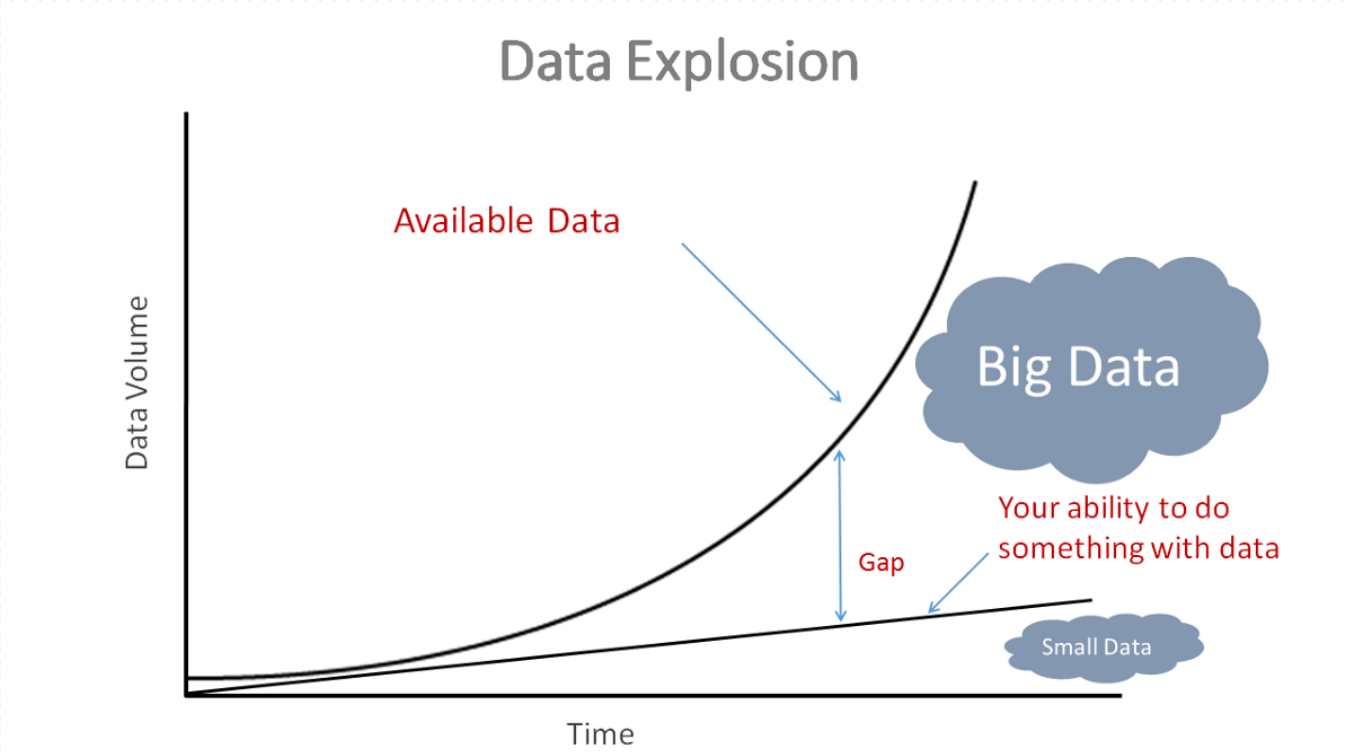
The key drivers of data growth are following :

- Increase in storage capacities.
- Cheaper storage.
- Increase in data processing capabilities by modern computing devices.
- Data generated and made available by different sectors.



Data Explosion

Data Explosion Solution ??



Big Data

No single definition; here is from Wikipedia:

- **Big data** is the term for a collection of data sets, which are large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.
- The challenges include capture, creation, storage, search, sharing, transfer, analysis, and visualization.
- The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data.

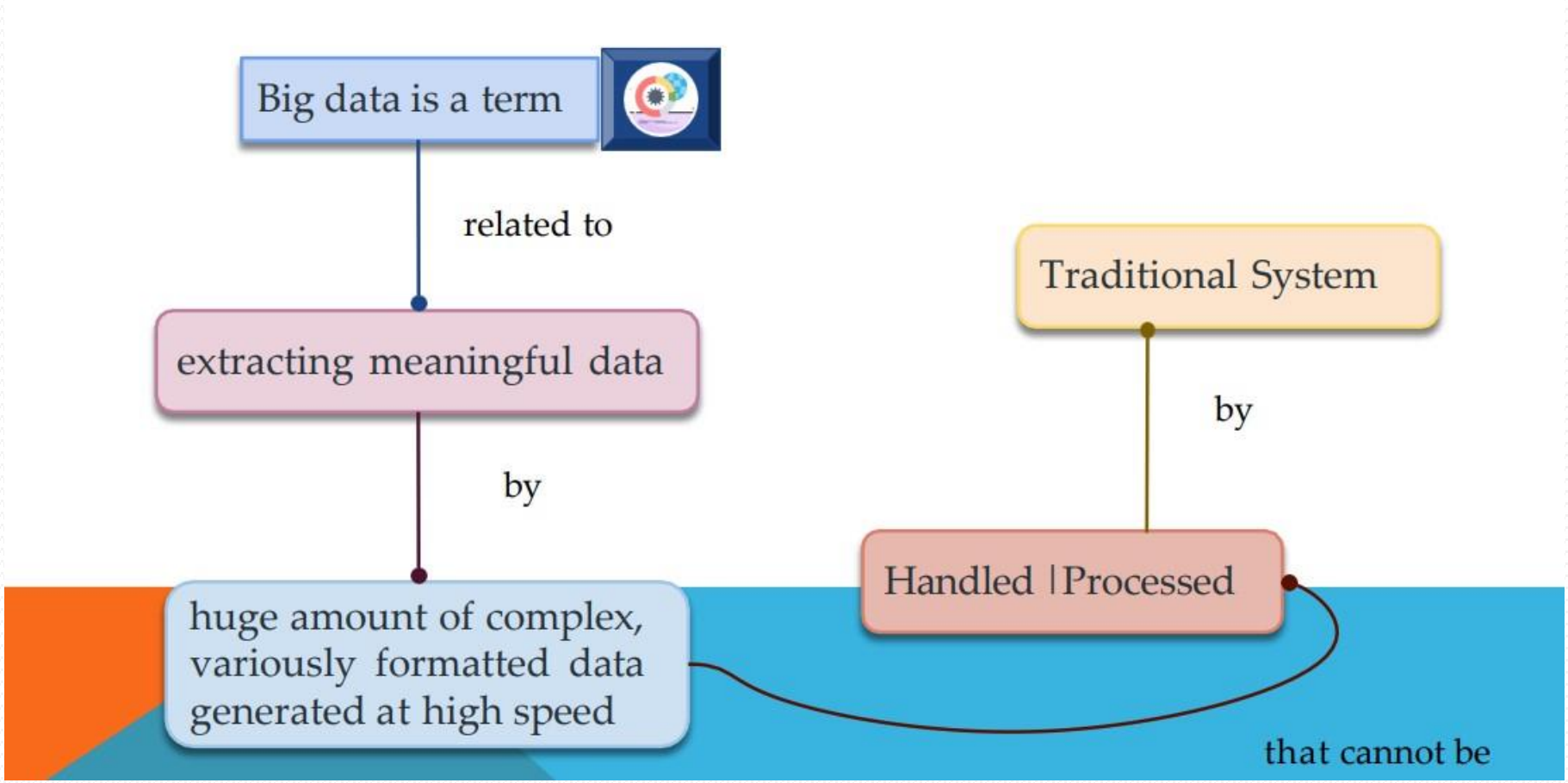
Big Data Example

• Credit card companies monitor every purchase their customers make and can identify fraudulent purchases with a high degree of accuracy using rules derived by processing billions of transactions.

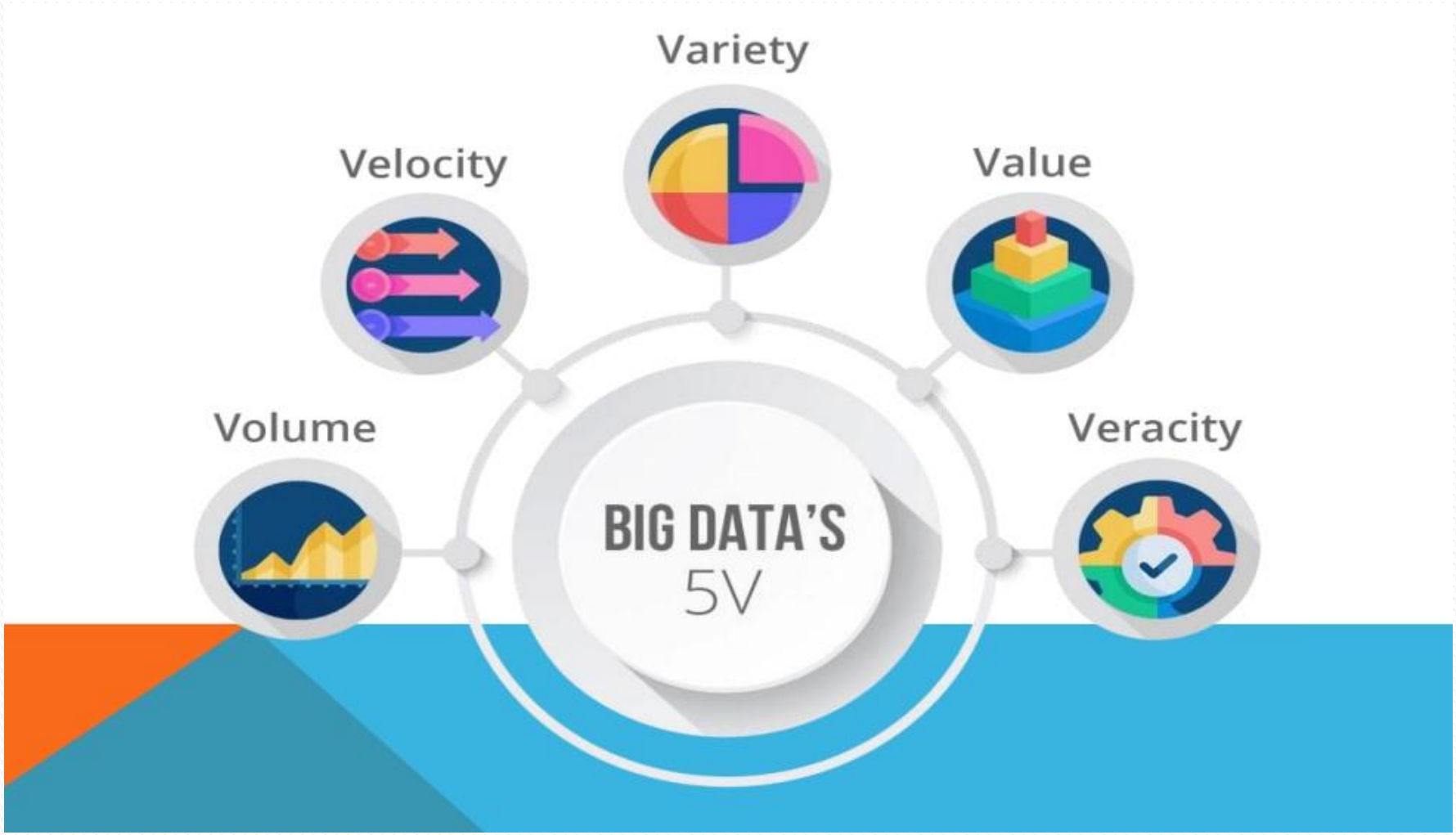
• Mobile phone companies analyze subscribers' calling patterns to determine, for example, whether a caller 's frequent contacts are on a rival network. If that rival network is offering an attractive promotion that might cause the subscriber to defect, the mobile phone company can proactively offer the subscriber an incentive to remain in her contract.

• For companies such as Linked In and Facebook, data itself is their primary product. The valuations of these companies are heavily derived from the data they gather and host, which contains more and more intrinsic value as the data grows.

Big Data


Big data is a term
related to
extracting meaningful data
by
huge amount of complex, variously formatted data generated at high speed
Traditional System
by
Handled |Processed
that cannot be

5 V's of Big Data



5 V's of Big Data

Characteristics and 5 V's of big data : ***Volume, Variety, Velocity, Value, Veracity***

- Huge volume of data (***Volume***) : Rather than thousands or millions of rows, Big Data can be billions of rows and millions of columns.

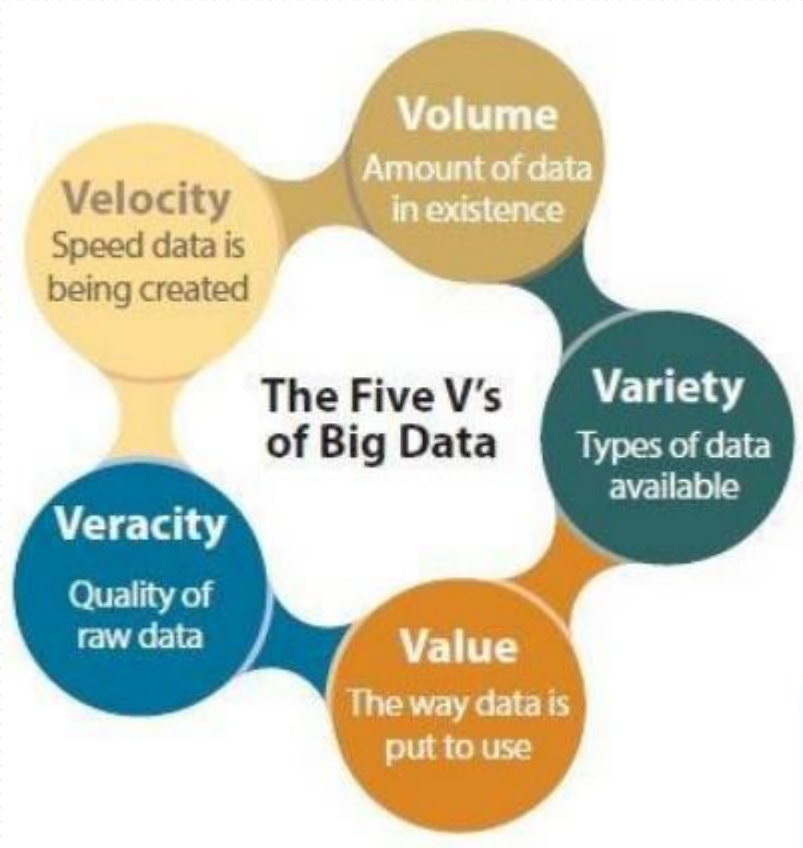
- Complexity of data types and structures (***Variety***): Big Data reflects the variety of new data sources, formats, and structures, including digital traces being left on the web and other digital repositories for subsequent analysis.

5 V's of Big Data

- Speed of new data creation and growth (***Velocity***): Big Data can describe high velocity data, with rapid data ingestion and near real time analysis.

- ***Veracity*** : Refers to the quality and accuracy of data.

- ***Value*** : what organizations can do with the collected data.



Sources of Big Data

The data now comes from multiple sources, such as these:

- Medical information, such as genomic sequencing and diagnostic imaging
- Photos and video footage uploaded to the World Wide Web
- Video surveillance, such as the thousands of video cameras spread across a city
- Mobile devices, which provide geospatial location data of the users, as well as metadata about text messages, phone calls, and application usage on smart phones
- Smart devices, which provide sensor-based collection of information from smart electric grids, smart buildings, and many other public and industry infrastructures
- Non-traditional IT devices, including the use of radio-frequency identification (RFID) readers, GPS navigation systems, and seismic processing

Sources of Big Data

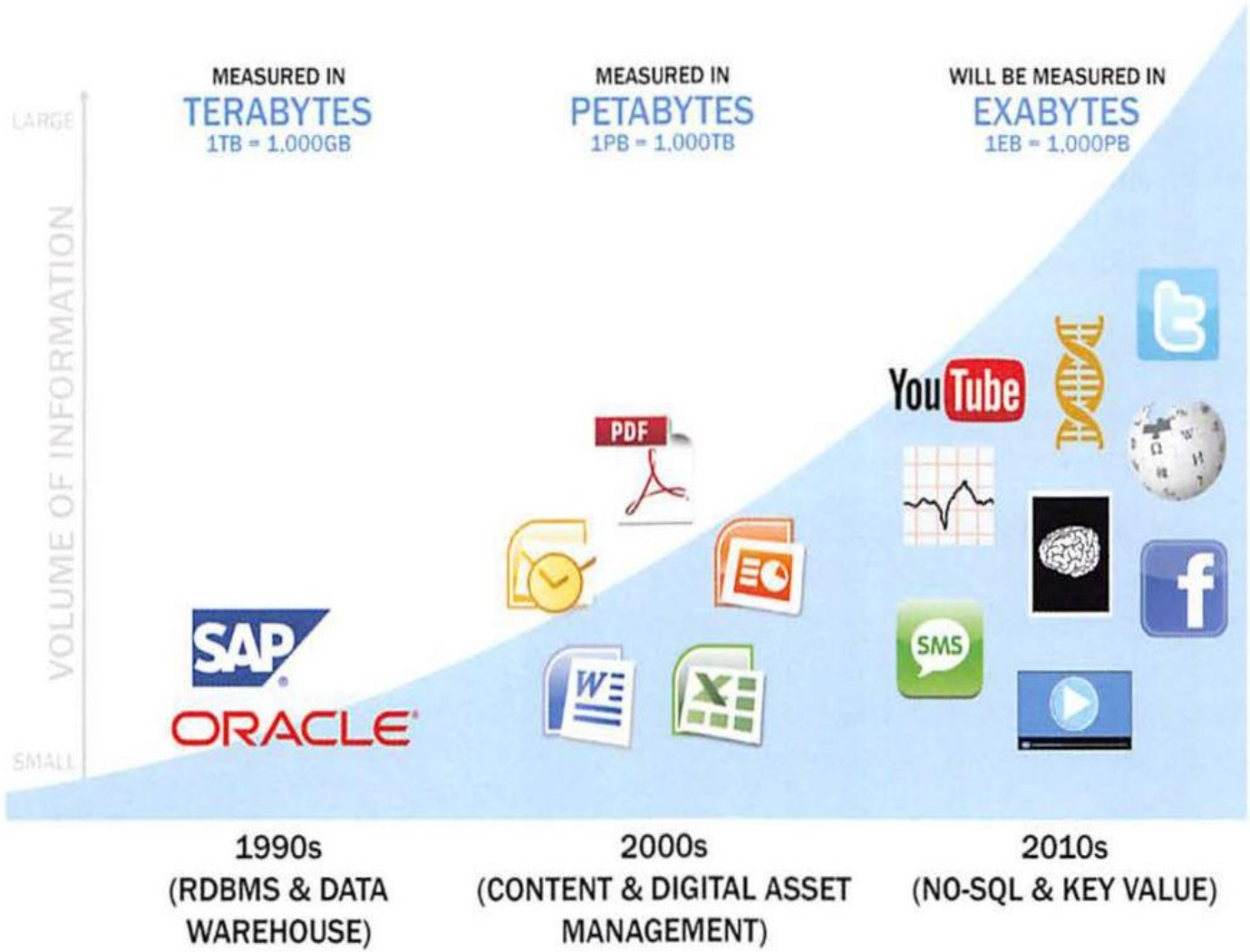
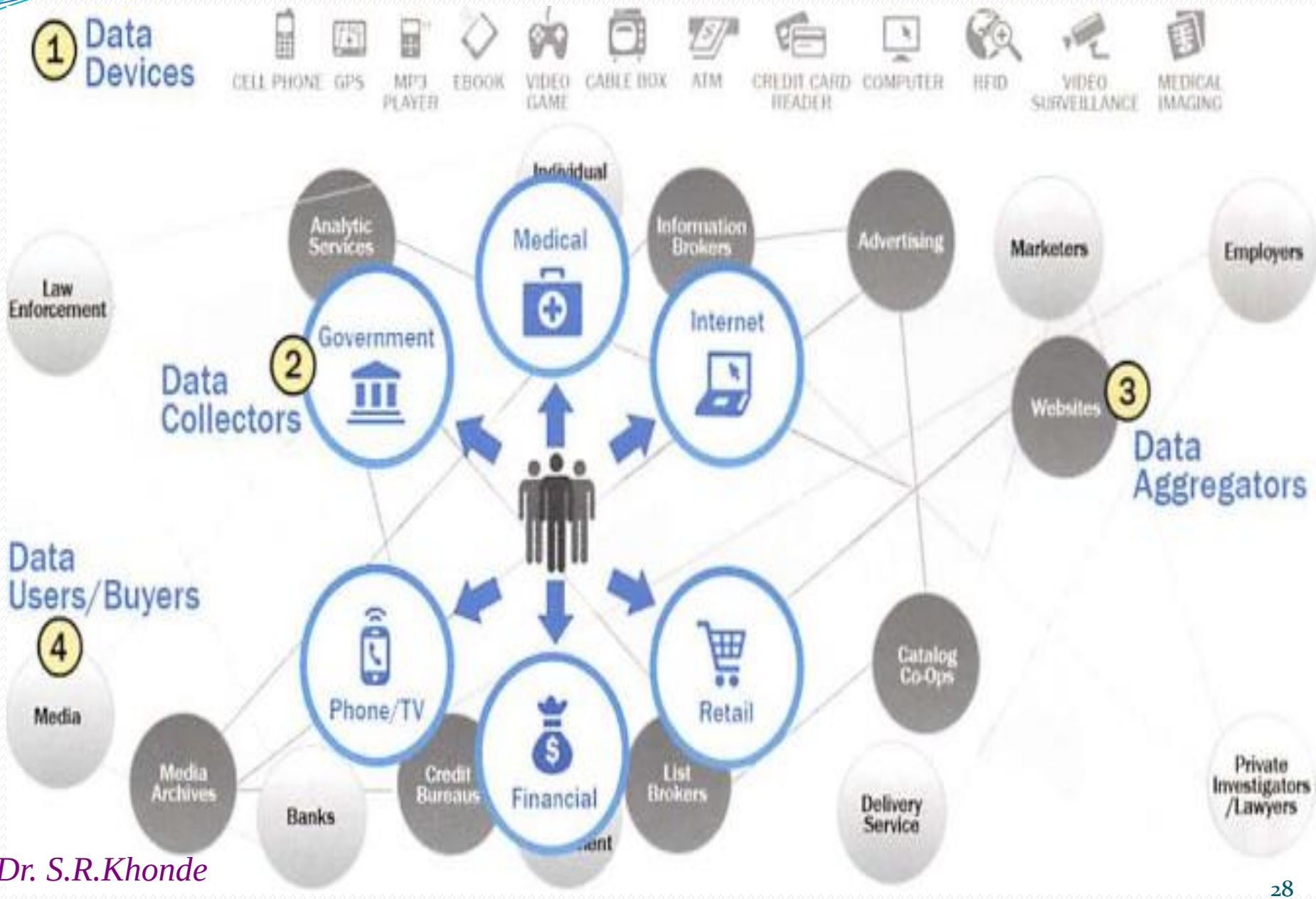


FIGURE 1-10 *Data evolution and the rise of Big Data sources*

Sources of Big Data

Big Data Generators



Business Intelligence Vs Data Science

1. **Perspective:** BI systems are designed to look backwards based on real data from real events. Data Science looks forward, interpreting the information to predict what might happen in the future.

2. **Focus:** BI delivers detailed reports, trends but it doesn't tell you what this data may look like in the future in the form of patterns and experimentation.

3. **Process:** Traditional BI systems tend to be static and comparative. They do not offer room for exploration and experimentation in terms of how the data is collected and managed.

Business Intelligence Vs Data Science

4. **Data sources:** Because of its static nature, BI data sources tend to be pre-planned and added slowly. Data science offers a much more flexible approach as it means data sources can be added on the go as needed.

5. **Transform:** How the data delivers a difference to the business is important. BI helps you answer the questions you know, whereas Data Science helps you to discover new questions because of the way it encourages companies to apply insights to new data.

6. **Storage:** Like any business asset, data needs to be flexible. BI systems tend to be warehoused and siloes, which means it is difficult to deploy across the business. Data Science can be distributed real time.

Business Intelligence Vs Data Science

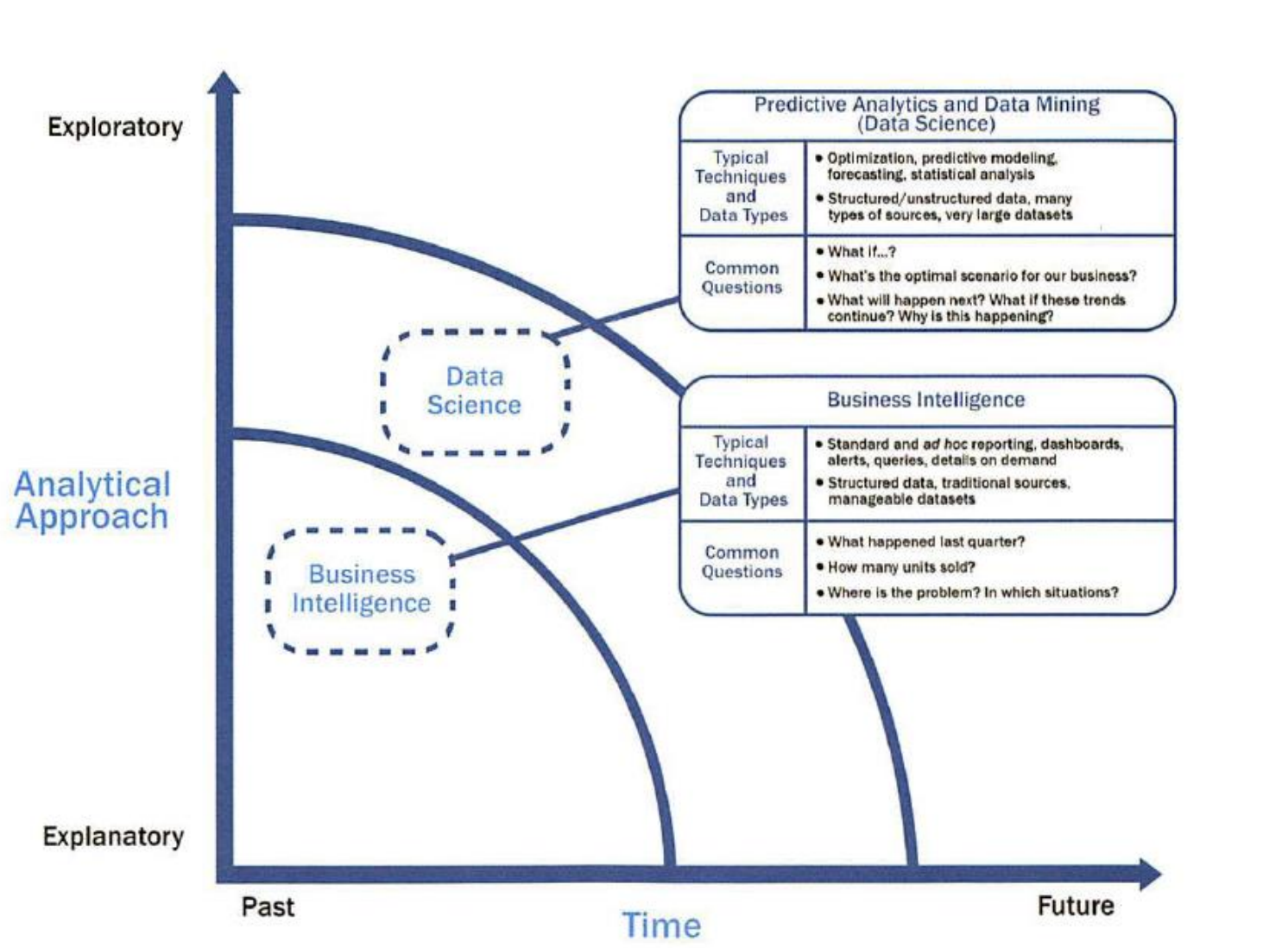
7. Data quality: Any data analysis is only as good as the quality of the data captured. BI provides a single version of truth while data science offers precision, confidence level and much wider probabilities with its findings.

8. IT owned vs. business owned

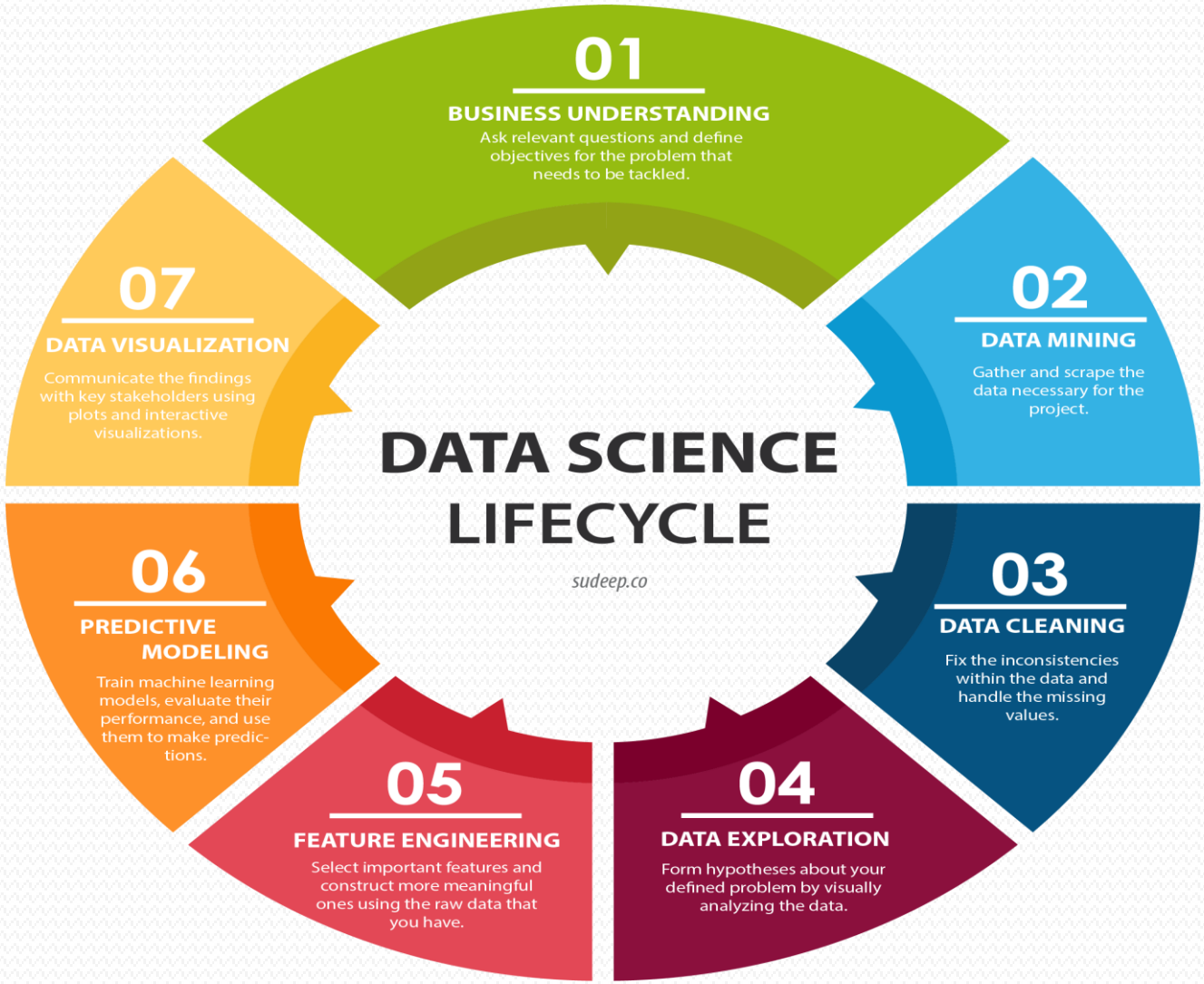
In the past, BI systems were often owned and operated by the IT department, sending along intelligence to analysts who interpreted it. With Data Science, the analysts are in charge. The new Big Data solutions are designed to be owned by analysts, who spend little of their time on 'IT housekeeping' and most of their time analyzing data and making predictions upon which to base business decisions.

9. Analysis: A retrospective and prescriptive BI system is much less likely to be placed to do this than a Predictive Data Science program.

Business Intelligence Vs Data Science



Data Science Lifecycle



Data Science Lifecycle

1. Business Understanding:

The complete cycle revolves around the enterprise goal. What will you resolve if you do no longer have a specific problem? It is extraordinarily essential to apprehend the commercial enterprise goal sincerely due to the fact that will be your ultimate aim of the analysis. After desirable perception only we can set the precise aim of evaluation that is in sync with the enterprise objective. You need to understand if the customer desires to minimize savings loss, or if they prefer to predict the rate of a commodity, etc.

Data Science Lifecycle

2. Data Understanding (Data Mining):

After enterprise understanding, the subsequent step is data understanding. This includes a series of all the reachable data. Here you need to intently work with the commercial enterprise group as they are certainly conscious of what information is present, what facts should be used for this commercial enterprise problem, and different information. This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data.

Data Science Lifecycle

3. Preparation of Data (Data Cleaning):

Next comes the data preparation stage. This consists of steps like choosing the applicable data, integrating the data by means of merging the data sets, cleaning it, treating the lacking values through either eliminating them or imputing them, treating inaccurate data through eliminating them, additionally test for outliers the use of box plots and cope with them. Constructing new data, derive new elements from present ones. Format the data into the preferred structure, eliminate undesirable columns and features. Data preparation is the most time-consuming but arguably the most essential step in the complete existence cycle. Your model will be as accurate as your data.

Data Science Lifecycle

4. Exploratory Data Analysis (Data Exploration):

This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model. Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps. Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

Data Science Lifecycle

5. Data Modelling (Feature Engineering):

Data modelling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output. This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem. After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them. We need to tune the hyper parameters of every model to obtain the preferred performance. We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

Data Science Lifecycle

6. Model Evaluation (Predictive Modelling):

Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved. Any data science solution, a machine learning model, simply like a human, must evolve, must be capable to enhance itself with new data, adapt to a new evaluation metric. We can construct more than one model for a certain phenomenon, however, a lot of them may additionally be imperfect. The model assessment helps us select and construct an ideal model.

Data Science Lifecycle

7. Model Deployment (Data Visualization):

The model after a rigorous assessment is at the end deployed in the preferred structure and channel. This is the last step in the data science life cycle. Each step in the data science life cycle defined above must be laboured upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste. For example, if data is no longer accumulated properly, you'll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work. If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.

Get to know your Data….

- Data Objects
- Attribute Types
- Types of Datasets
- Collection of Data
- Need of Data Wrangling
- Data Wrangling Methods

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.

- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses

- Also called *samples* , *examples*, *instances*, *data points*, *objects*, *tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns ->attributes.

Attributes

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
- *E.g., customer _ID, name, address*

- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled
  - Discrete and continuous attributes

Attributes Types

- **Nominal:** categories of states, or "names of things"
  - *Hair_color* = {*auburn*, *black*, *blond*, *brown*, *grey*, *red*, *white*}
  - marital status, occupation, ID numbers, zip codes

- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)

- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small*, *medium*, *large*}, grades, army rankings

Attributes Types

- Numeric : Quantitative (integer or real-valued)

- **Interval-Scaled**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point

- **Ratio-Scaled**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Attributes Types

- Discrete vs. Continuous Attributes

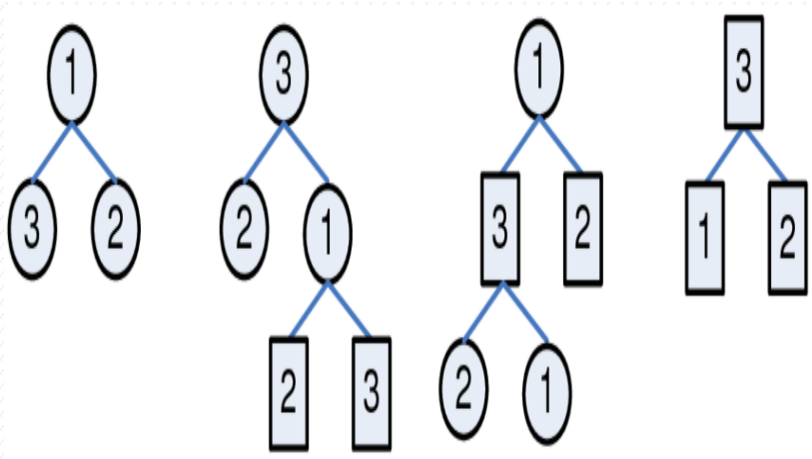
- **Discrete Attribute**
  - Has only a finite or countable infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes

- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Types of Data Sets

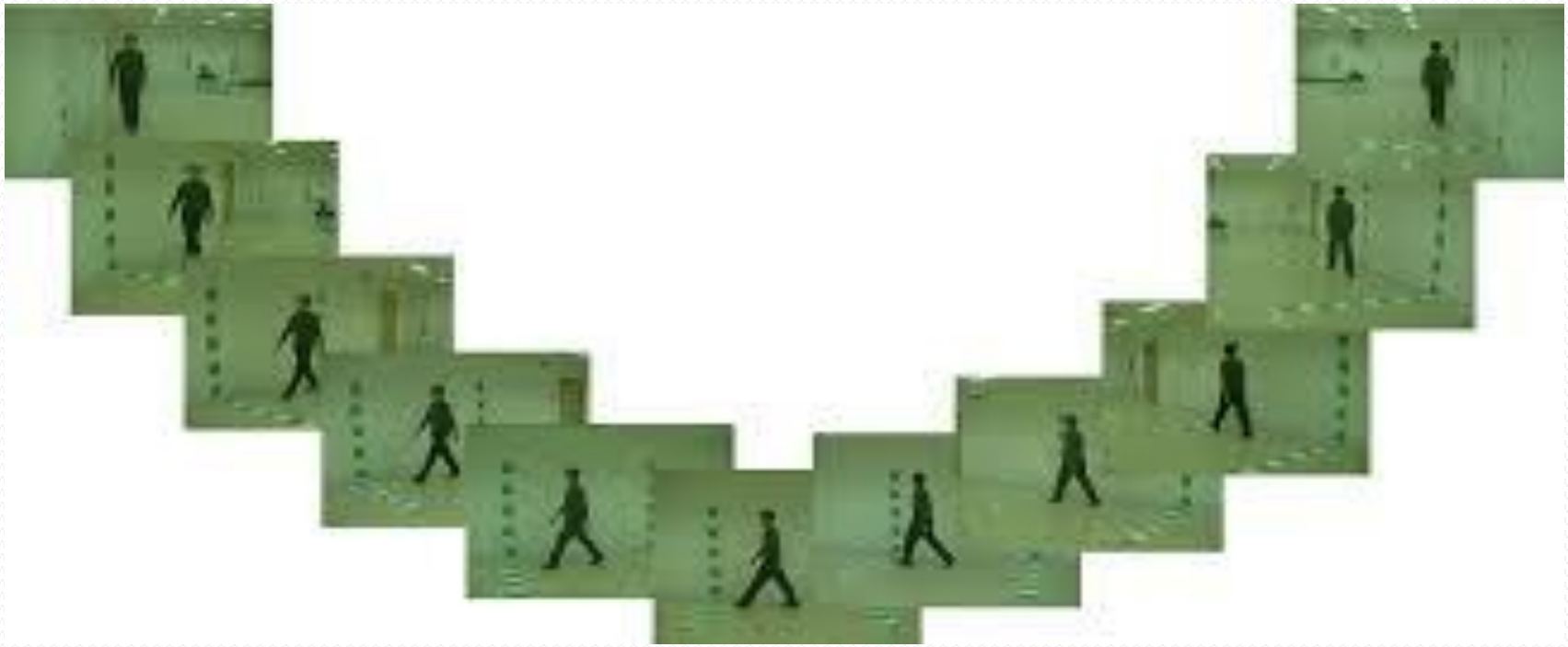
- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0



Types of Data Sets

- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:



Data Wrangling

- Data Preprocessing: An Overview
  - Data Quality
  - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

NEED OF DATA WRANGLING

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 MisÃ©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

77

NEED OF DATA WRANGLING

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
  - Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - Timeliness: timely update?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation*=“ ” (missing data)
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*=“−10” (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*=“142”, *Birthday*=“03/07/2050”
    - Was rating “1, 2, 3”, now rating “A, B, C”
    - discrepancy between duplicate records
  - Intentional (e.g., *disguised missing* data)
    - Jan. 1 as everyone’s birthday?

Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

METHODS OF DATA WRANGLING

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%

METHODS OF DATA WRANGLING

Unit - 1

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%	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%
4	Fast+		80%
5	Lite	76	70%
6	Fast+	155	10%
7			95%
8	Lite	76	77%
9	Fast+	180	

101

METHODS OF DATA WRANGLING

Unit - 1

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%

102

METHODS OF DATA WRANGLING

Unit - 1

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%

103

METHODS OF DATA WRANGLING

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%

106

METHODS OF DATA WRANGLING

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

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%

Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Binning method

Binning Methods for Data Smoothing:

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into equal-frequency (equi-depth) bins::
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Data Integration

- **Data integration**:
  - Combines data from multiple sources into a coherent store
- Schema integration: e.g., A.cust-id ≡ B.cust-#
  - Integrate metadata from different sources
- **Entity identification problem**:
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- **Detecting and resolving data value conflicts**
  - For the same real world entity, attribute values from different sources are different
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- Redundant data occur often when integration of multiple databases
  - *Object identification*: The same attribute or object may have different names in different databases
  - *Derivable data:* One attribute may be a "derived" attribute in another table, e.g., annual revenue
- Redundant attributes may be able to be detected by *correlation analysis* and *covariance analysis*
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

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

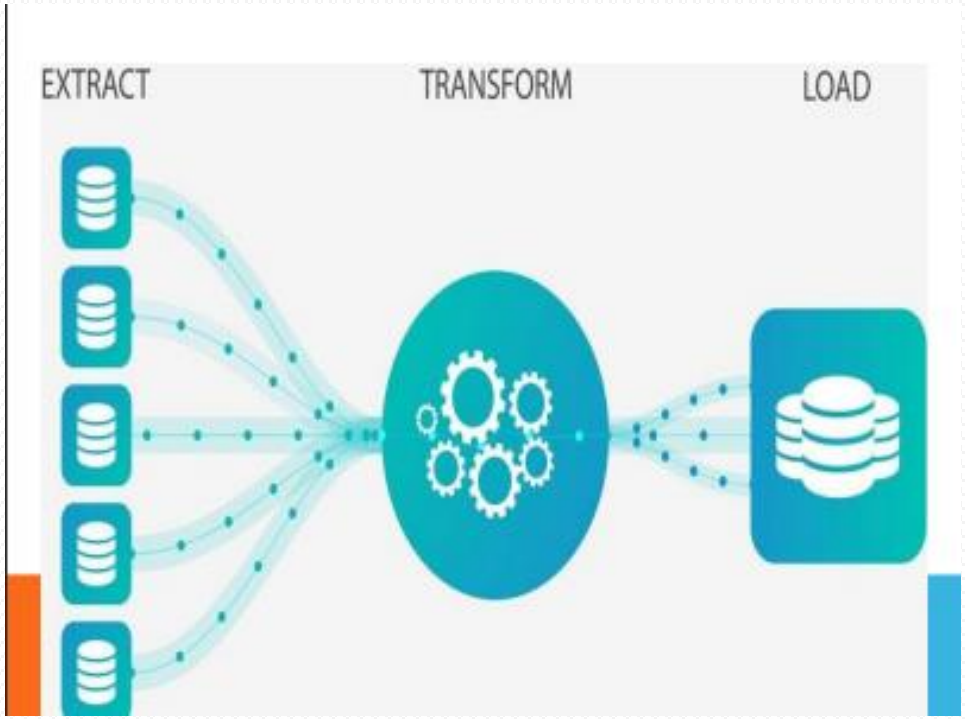
118

Data Reduction Strategies

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - Dimensionality reduction, e.g., remove unimportant attributes
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - Numerosity reduction (some simply call it: Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - Data compression

Data Transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values.
- Process of transformation is called as ETL (Extract, Transform and Load)



Data Transformation

- Methods
  - Smoothing: Remove noise from data
  - Attribute/feature construction
    - New attributes constructed from the given ones
  - Aggregation: Summarization, data cube construction
  - Normalization: Scaled to fall within a smaller, specified range
    - min-max normalization
    - z-score normalization
    - normalization by decimal scaling

Discretization

- Discretization: Divide the range of a continuous attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Reduce data size by discretization
  - Supervised : 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.
  - Unsupervised : This type of discretization considers only the attribute being discretized and does not use class information.
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
  - Prepare for further analysis, e.g., classification

Data Discretization Methods

- Typical methods: All the methods can be applied recursively
  - Binning
    - Top-down split, unsupervised
  - Histogram analysis
    - Top-down split, unsupervised
  - Clustering analysis (unsupervised, top-down split or bottom-up merge)
  - Decision-tree analysis (supervised, top-down split)
  - Correlation (e.g., χ^2) analysis (unsupervised, bottom-up merge)

Reference Books used for this unit:

1. Data Science and Big Data Analytics by EMC Education Services

2. Data Mining Concepts and Techniques by J Han, M Kamber and Jian Pei

END
of
UNIT I