

Collecting and Cleaning Data

Dr. Sandareka Wickramanayake
sandarekaw@cse.mrt.ac.lk

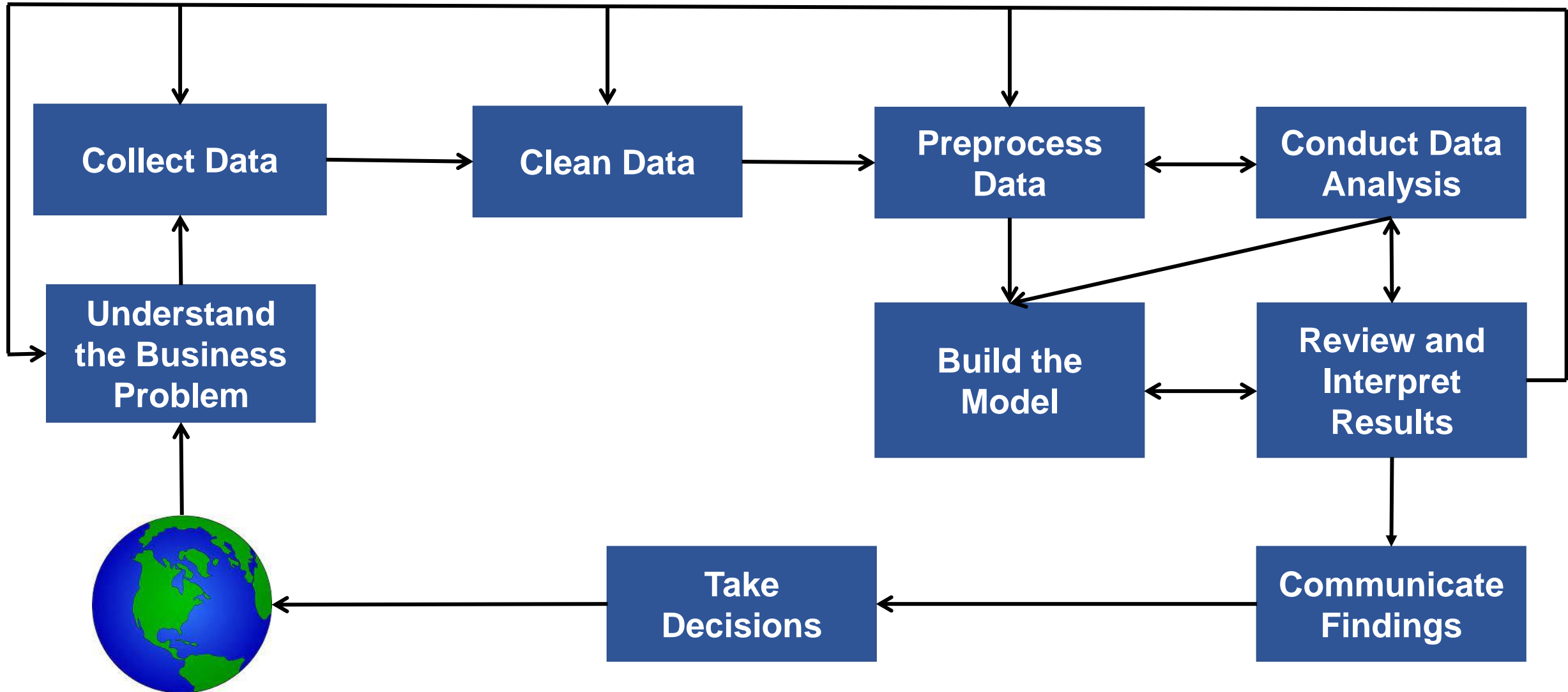
Acknowledgement

- These slides are based on

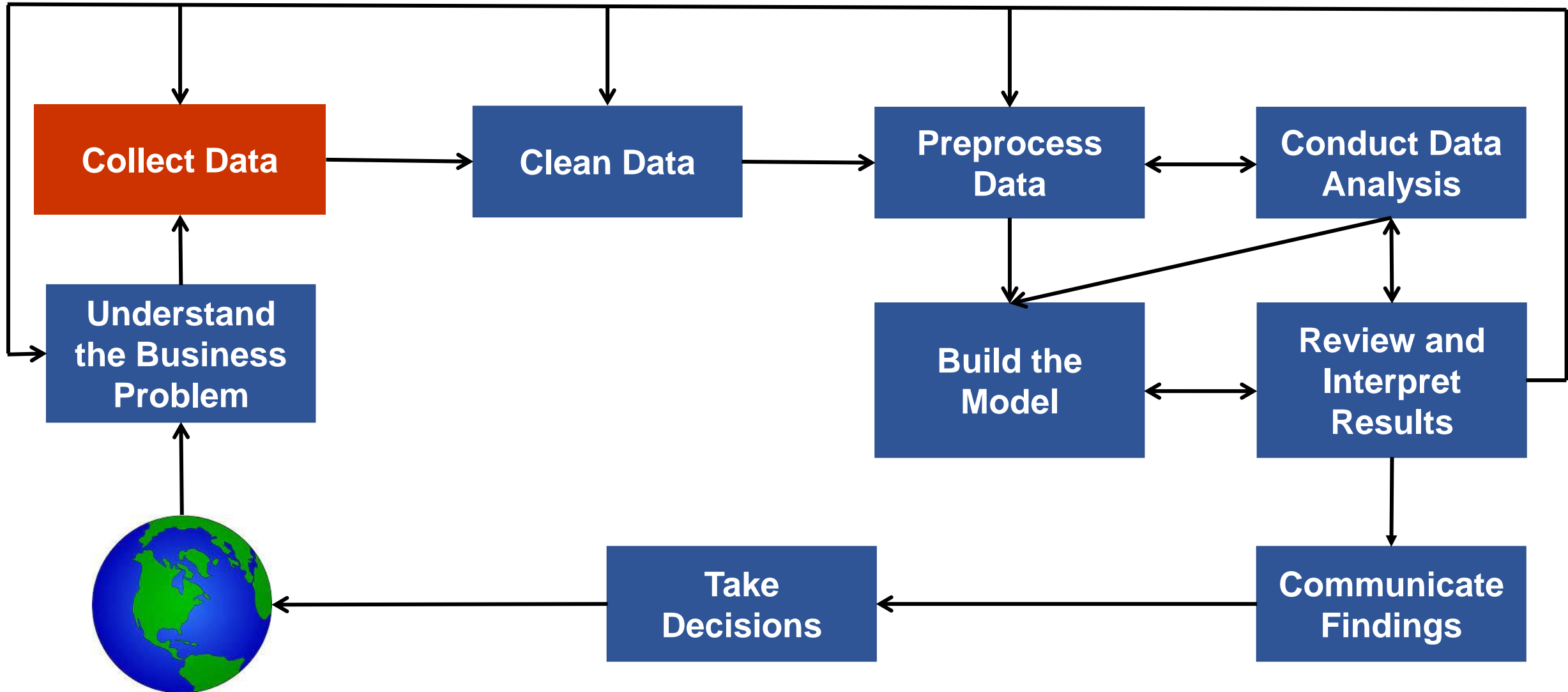
Chapter 3 - **Data Mining: Know It All**

by Soumen Chakrabarti, Earl Cox, Eibe Frank, Ralf Hartmut Güting, Jiawei Han, Xia Jiang, Micheline Kamber, Sam S. Lightstone, Thomas P. Nadeau, Richard E. Neapolitan, Dorian Pyle, Mamdouh Refaat, Markus Schneider, Toby J. Teorey, Ian H. Witten

Data Science Process



Data Science Process



Data Collection – 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, and tuples.
- Data objects are described by attributes.
- Database rows → data objects; columns → attributes.

Data Collection – Data Objects

Attribute: **Gender**

Values: {Male, Female}

FILE

HOME

INSERT

PAGE LAYOUT

FORMULAS

DATA

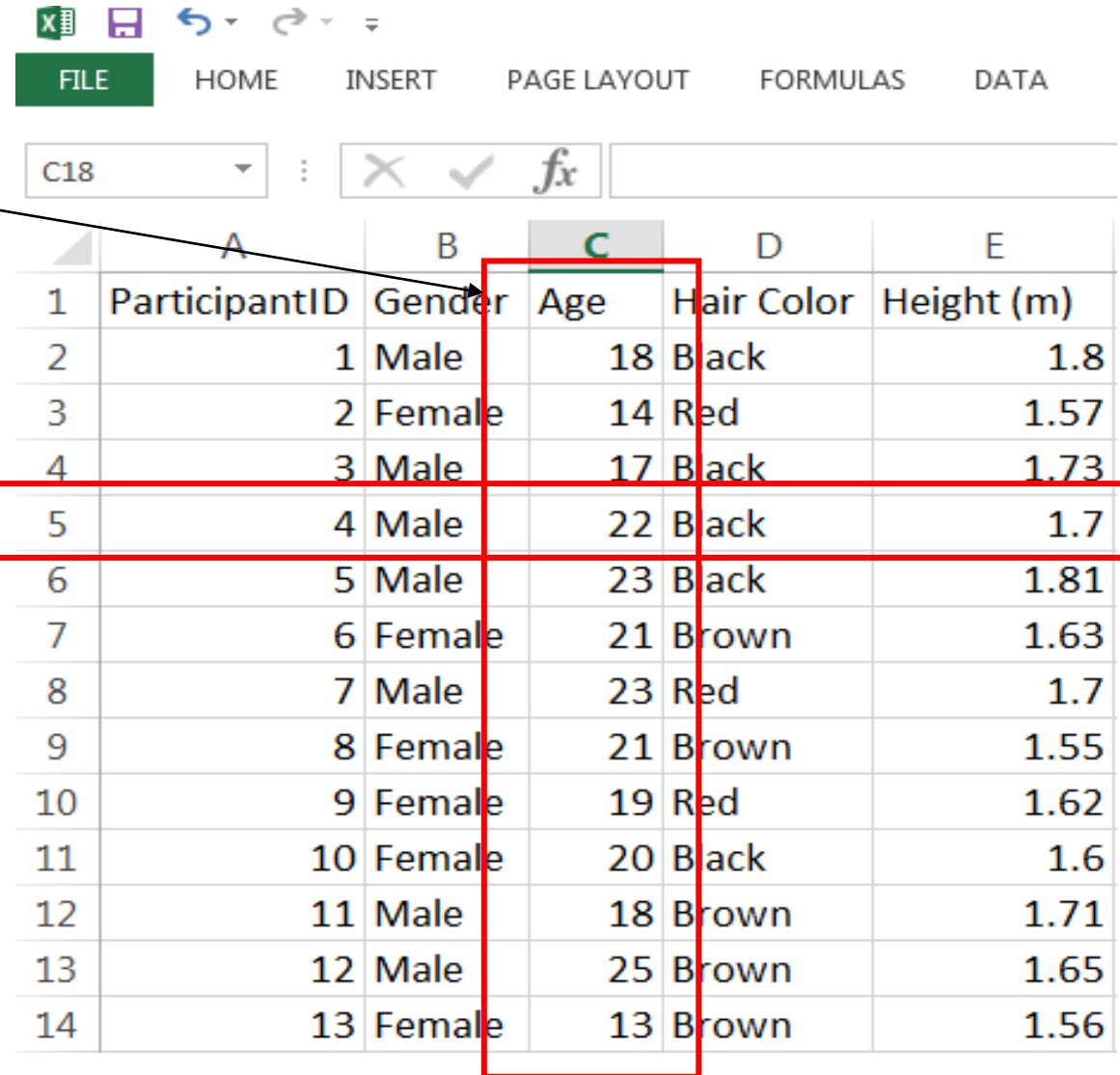
C18

A data object

Data Collection – Data Objects

Attribute: **Age**

Values: 1,2,3,4,5,6,7,.....,125



The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E
	ParticipantID	Gender	Age	Hair Color	Height (m)
1		1 Male	18	Black	1.8
2		2 Female	14	Red	1.57
3		3 Male	17	Black	1.73
4		4 Male	22	Black	1.7
5		5 Male	23	Black	1.81
6		6 Female	21	Brown	1.63
7		7 Male	23	Red	1.7
8		8 Female	21	Brown	1.55
9		9 Female	19	Red	1.62
10		10 Female	20	Black	1.6
11		11 Male	18	Brown	1.71
12		12 Male	25	Brown	1.65
13		13 Female	13	Brown	1.56

A data object

Data Collection – Data Objects

Attribute: **Height**
Values: 1.8, 1.57, ...

A data object →

	A	B	C	D	E
1	ParticipantID	Gender	Age	Hair Color	Height (m)
2	1	Male	18	Black	1.8
3	2	Female	14	Red	1.57
4	3	Male	17	Black	1.73
5	4	Male	22	Black	1.7
6	5	Male	23	Black	1.81
7	6	Female	21	Brown	1.63
8	7	Male	23	Red	1.7
9	8	Female	21	Brown	1.55
10	9	Female	19	Red	1.62
11	10	Female	20	Black	1.6
12	11	Male	18	Brown	1.71
13	12	Male	25	Brown	1.65
14	13	Female	13	Brown	1.56

Data Collection – Data Types

Data Types

```
graph TD; A[Data Types] --> B[Quantitative]; A --> C[Qualitative];
```

Quantitative

- Numbers, Tests, Counting, Measuring
- Quantitative data are data about numeric variables (E.g., how many, how much, or how often).

Qualitative

- Words, Images, Observations, Conversations, Photographs
- Qualitative data are data about categorical variables (E.g., what type).

Data Collection – Data Types

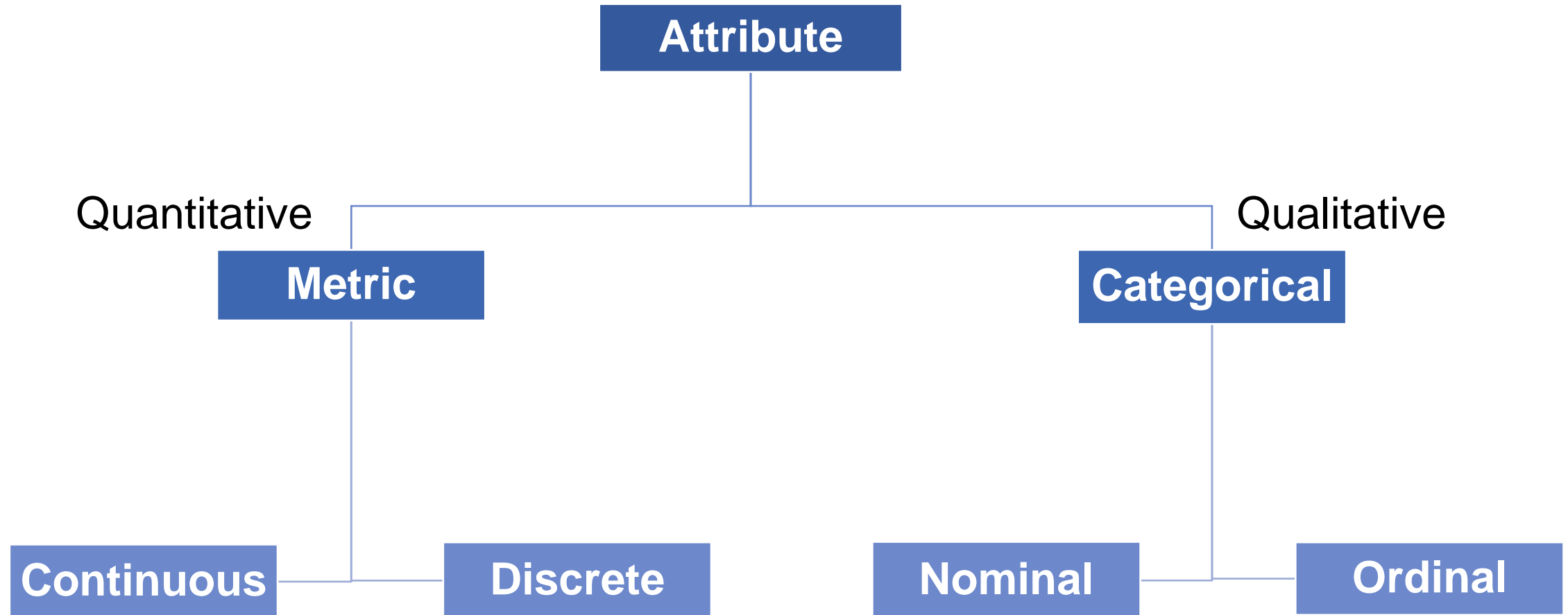
Data Types

Quantitative

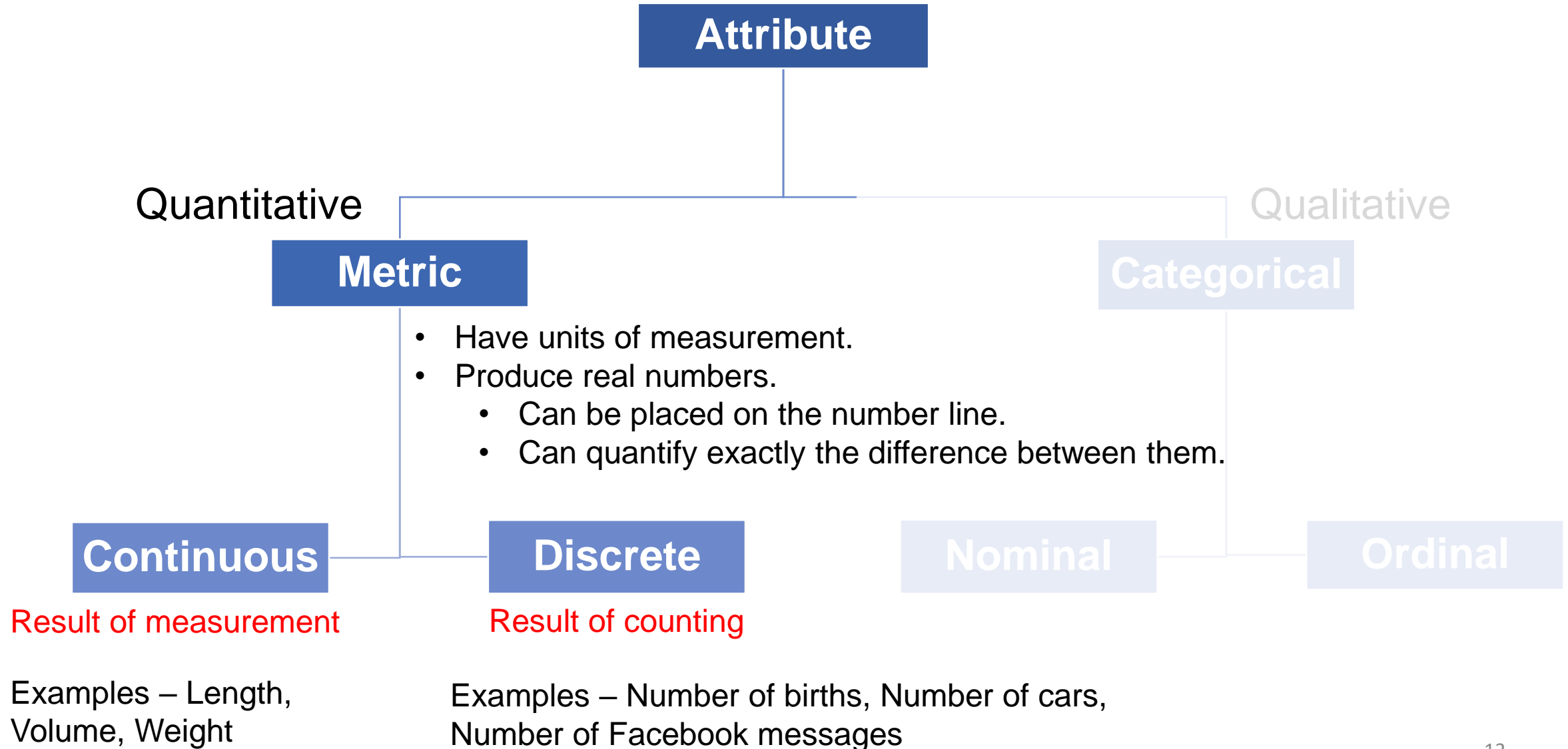
Qualitative

Data Unit	Numerical Variable	= Quantitative Data	Categorical Variable	= Qualitative Data
A person	"How many children do you have?"	4 Children	"In which country were your children born"	Sri Lanka
	"How much do you earn?"	\$60,000 p.a.	"What is your occupation"	Photographer
	"How many hours do you work?"	38 hours per week	"Do you work full-time or part-time?"	Full-time
A house	"How many square meters is the house?"	200 sq meters	"In which city or town is the house located?"	Colombo
A business	"How many workers are currently employed?"	264 employees	"What is the industry of the business?"	Retail
A farm	"How many milk cows are located on the farm?"	36 cows	"What is the main activity of the farm"	Diary

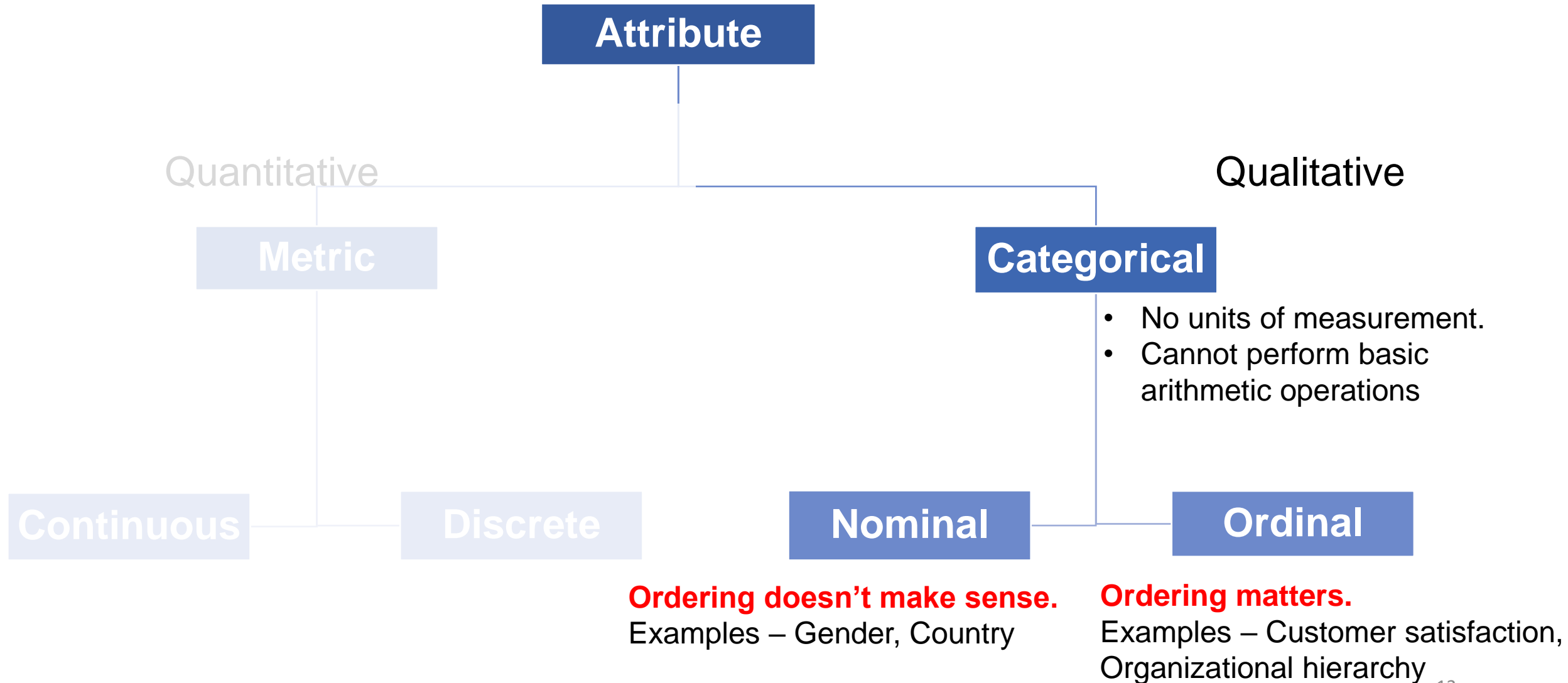
Data Collection – Data Types



Data Collection – Data Types



Data Collection – Data Types



Data Collection – Data Types

Data Types

Primary

- Collected fresh, for the first time thus, original.
- Has not been published.
- More reliable, authentic
- Has not been altered by a human.

Secondary

- Data collected by others
- Might be published or unpublished.
- Might have been collected for a different purpose than the problem at hand.

Pros	Cons
Targeted issues are addressed	Evaluation cost
Data interpretation is better	Time consuming
High accuracy of data	More recourses needed
Greater control	Might contain inaccuracies
	Required skill in labor

Pros	Cons
Quick and cheap	Might not fulfil our objective
Wider geo-socio area reach (/coverage)	Poor accuracy
	Might be outdated
	Poor accessibility

Data Collection

- Techniques: observations, tests, surveys, document analysis
- Sources of Data
 - Operational systems
 - E.g., - A banking transaction processing system that keeps records of opened and closed accounts, deposits, withdrawals, balances, and all other values related to the money moving among accounts, clients, and the outside world.
 - Data warehouses and data marts
 - Online analytical processing applications
 - Surveys

Data Collection – Factors to be Considered

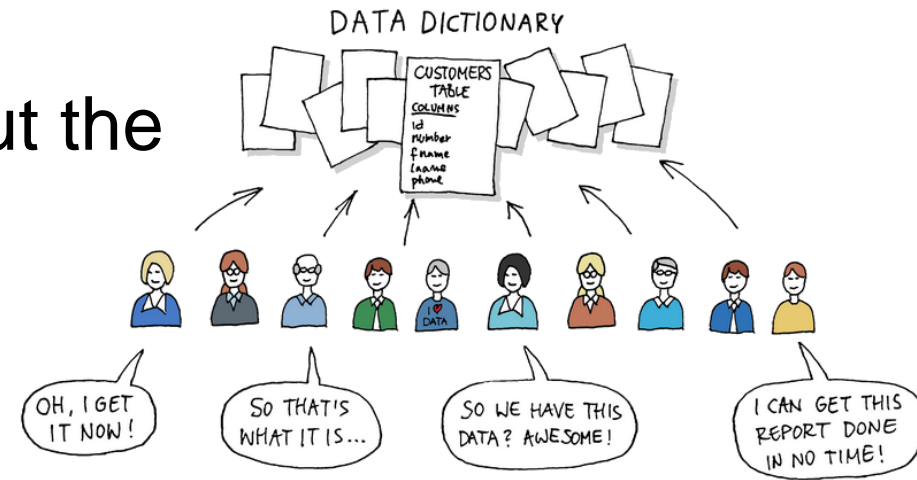
- Objective and scope
- Sources
- Technique of collection
- Representativeness
 - Does your sample represent the population you are studying? Must use random sample techniques.
- Eliminate biases
- Prevent data contamination by poor measurement or error in procedure

Data Documenting

- Sufficient descriptive information about your data so that it can be used properly by you, your colleagues, and other researchers in the future.
- Well-documented data is identifiable, understandable, and usable in the future

Data Dictionary – What is it?

- A summary of the dataset.
- Contains important meta-information about the data:
 - Where it comes from.
 - Who owns it.
 - Access rights/privacy, etc.
- Contains the generic, but also the most important, descriptive understanding of the dataset:
 - What are the attributes.
 - What are their data types.
 - What do their values look like.
 - Any additional information/cautions needed.



Data Dictionary – Why do we need it?

- To have a reference to the dataset when needed.
- To be used as a standard document among team members.
- Every team member has agreed upon the information provided.
- To communicate with external parties (e.g., to your customer, industry partners, etc.).

Field Name	Data Type	Data Format	Field Size	Description	Example
License ID	Integer	NNNNNN	6	Unique number ID for all drivers	12345
Surname	Text		20	Surname for Driver	Jones
First Name	Text		20	First Name for Driver	Arnold
Address	Text		50	First Name for Driver	11 Rocky st Como 2233
Phone No.	Text		10	License holders contact number	0400111222
D.O.B	Date / Time	DD/MM/YYYY	10	Drivers Date of Birth	08/05/1956

<https://thebadoc.com/ba-techniques/f/defining-a-data-dictionary>

Sheet_1				
Show rows with cells including: <input type="text"/>				
Variable	Variable name	Mesaurement unit	Allowed values	Description
Participant ID number	ID	Numeric	001-999	ID number assigned to participant in sequential order
Group number	GROUP	Numeric	1-30	Group assigned to participant based on ID number
Age in years	AGE	Numeric	18.0-65.0	Age of participant in years
Date of birth	DOB	mm/dd/yyyy	1-12/1-31/1951-1998	Participant's date of birth
Gender	SEX	Numeric	1 = male 2 = female	Participant's gender
Date of survey	SURVEY	mm/dd/yyyy	01/01/2015 – 01/01/2016	When the participant completed the survey
Self-reported consumer spending	SPEND	Numeric	0-100,000,000	Self-reported average yearly expenditure
Market sentiment	SENTIMENT	Numeric	1 = negative 2 = neutral 3 = positive	Sentiment towards US domestic economy
Actual GDP growth	GDP	Numeric	-5.0-5.0	Average US yearly GDP growth

<https://help.osf.io/article/217-how-to-make-a-data-dictionary>

Data Dictionary – Steps to build a data dictionary

	A	B	C	D	E	F	G
1	Purchase ID	Last name	First name	Birthday	Country	Date of purchase	Amount of purchase
2	1	Davidson	Michael	04/03/1986	United States	10/12/2016	37
3	2	Vito	Jim	09/01/1994	United Kingdom	02/02/2016	85
4	3	Johnson	Tom	23/08/1972	France	02/11/2016	83
5	4	Lewis	Peter	18/10/1979	Germany	22/11/2016	27
6	5	Koenig	Edward	13/05/1983	Argentina	26/03/2015	43
7	6	Preston	Jack	16/06/1991	United States	06/11/2016	77
8	7	Smith	David	11/03/1965	Canada	15/11/2016	23
9	8	Brown	Luis	03/09/1997	Australia	03/07/2015	74
10	9	Miller	Thomas	07/01/1980	Germany	07/11/2016	13
11	10	Williams	Bill	26/07/1960	United States	20/11/2015	80
12	11	Gemini	Alexia	12/09/1995	Canada	11/03/2017	35
13	12	Bond	James	25/02/1975	United Kingdom	12/08/2017	40
14	13	Burgle	Patricia	01/12/1990	United States	18/01/2015	55
15	14	Reding	Michelle	07/04/1985	Canada	23/02/2017	28
16	15	Harvey	Billy	14/07/1971	United Kingdom	12/01/2016	41
17							

Structured Dataset

Step 1

- Construct Dataset Description

Step 2

- Construct Attribute Description

Data Dictionary – Steps to build a data dictionary

- Step 1 – Construct dataset description
 - We need to provide information for:
 - Dataset Name
 - Dataset Size
 - Date of Release
 - No. of Attributes
 - No. of Data Records
 - Data Source Provider
 - Data Privacy
 - Notes
 - Prepared by
 - Point of Contact
 - Team Members

Data Dictionary – Steps to build a data dictionary

- Step 1 – Construct dataset description
 - Dataset Name
 - Specify the name of the dataset if known, if not, attempt to describe where the data from.
 - Dataset Size
 - Specify the size of the dataset, an approximate number is also okay. (e.g., ≈ 342.5 MB)
 - Date of Release
 - Specify the date the dataset is released or being updated. Choose the date that is most recent to the time the data was downloaded.
 - Number of Attributes
 - Specify the number of attributes if known exactly, otherwise give your best estimate. (e.g., for tabular data, it is usually the number of columns)

Data Dictionary – Steps to build a data dictionary

- Step 1 – Construct dataset description
 - Number of Data Records
 - Specify the number of data records, or an estimate of this number (e.g., it is usually the number of rows)
 - Data Source Provider
 - Specify the source of your data provider (e.g., the company name (if you are conducting analysis service for a company), URLs where the data comes from, etc.)
 - Data Privacy
 - Is the data publicly available or confidential? Any privacy attention needed (e.g., public, strictly confidential, internally used only, etc.)
 - Notes
 - Any additional information needed (e.g., the data has been pre-processing, modified in some ways)

Data Dictionary – Steps to build a data dictionary

- Step 1 – Construct dataset description
 - Prepared by
 - Specify the name of your team, organization, your name, etc.
 - Point of Contact
 - Specify contact detail (e.g., the team leader's name, email, telephone, etc.)
 - Team Members
 - If applicable, list all team members, otherwise this field can be removed.

Data Dictionary – Steps to build a data dictionary

- Step 1 – Construct dataset description

DATA SET NAME	Geelong WiFi Usage Subset
DATA SIZE	7.9 MB
DATE OF RELEASE	1/06/2015
NO OF ATTRIBUTES	20
NO OF DATA RECORDS	500
DATA SOURCE PROVIDER	http://data.gov.au/dataset/geelong-wi-fi-usage
DATA PRIVACY	Publicly Available
NOTES	For education purpose, only the first 500 data records are retrained for analytics tasks
Prepared by:	Teaching Team- SIT112 Data Science Concepts.
Point of Contact:	Dr Truyen Tran: truyen.tran@deakin.edu.au
Team Members:	Truyen Tran, Trang Pham

Data Dictionary – Steps to build a data dictionary

- Step 2 – Construct attribute description
 - For each attribute, we need to provide information for:
 - Attribute Name
 - Data Type
 - Data Subtype (if applicable)
 - Description
 - Examples
 - Additional Notes

Data Dictionary – Steps to build a data dictionary

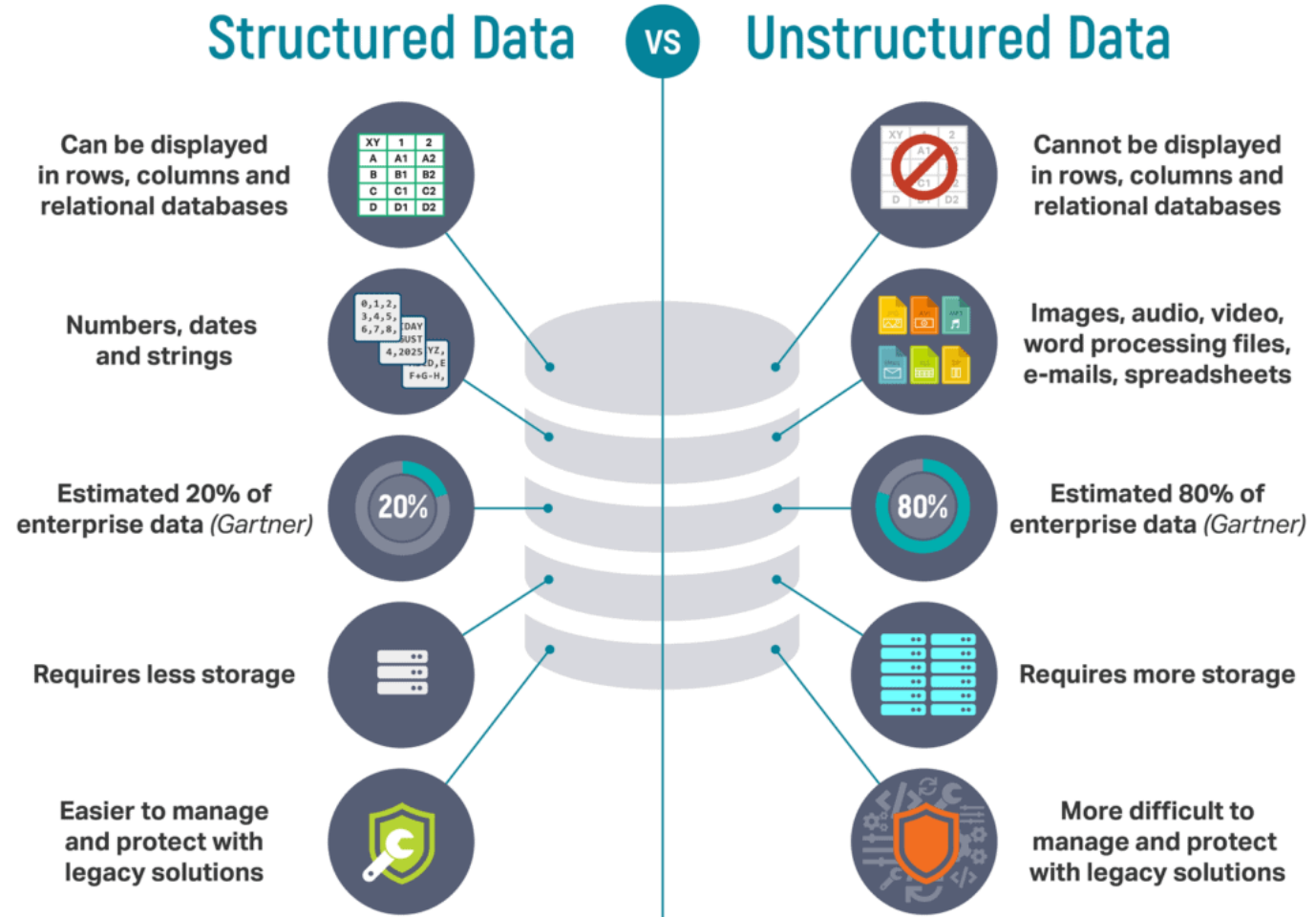
- Step 2 – Construct attribute description - Example

- Attribute Name
 - visitorid
- Data Type
 - Categorical Nominal
- Data Subtype
 - ID (identification)
- Description
 - Visitor Identification Number
- Examples
 - 1567009
- Additional notes
 - Data provider didn't give any specific description.

visitorid
1893824
1893825
1893826
1893833
1893836
1893838
1893839

Structured vs Unstructured Datasets

- Structured Data
 - Typically categorized as quantitative data.
 - Highly organized and easily understandable by humans and machines.
- Unstructured Data
 - Typically categorized as qualitative data
 - Cannot be processed and analyzed via conventional data tools and methods.
 - Does not have a predefined data model.

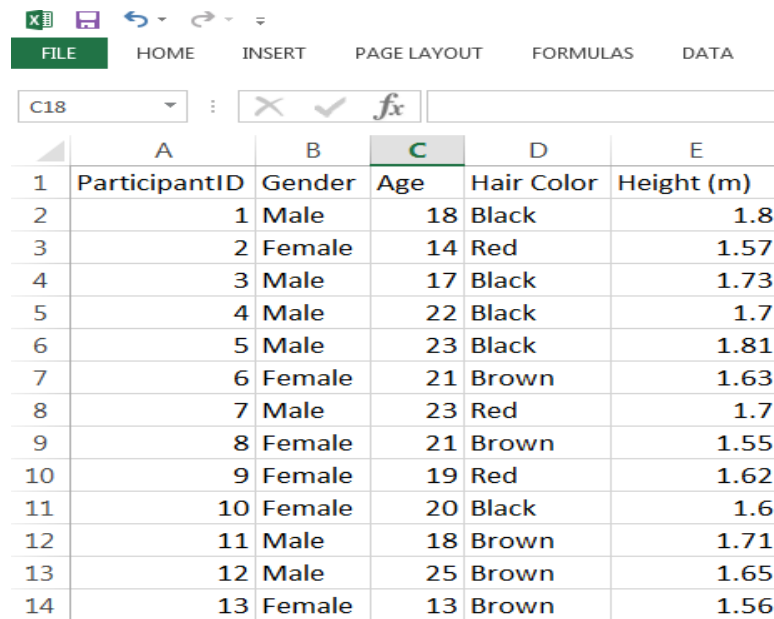


Characteristics of Structured Data

- Dimensionality / Cardinality
 - Curse of Dimensionality / Cardinality
- Sparsity
 - Only presence counts
- Resolution
 - Patterns depend on the scale
- Distribution
 - Centrality and dispersion

Types of Datasets

- Record
 - Relational records
 - Data matrix, e.g., numerical matrix, crosstabs
 - Document data: text documents: term-frequency vector
 - Transaction data



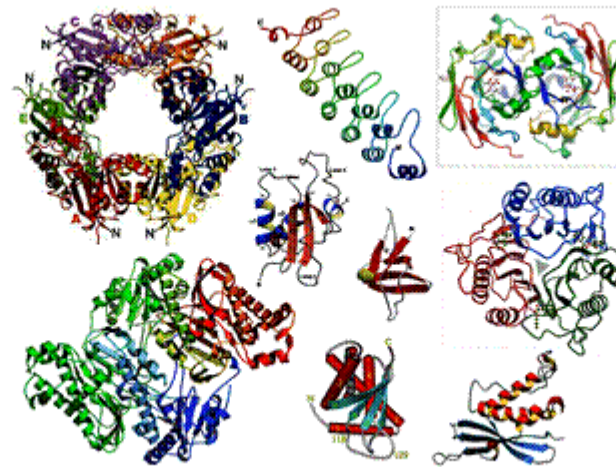
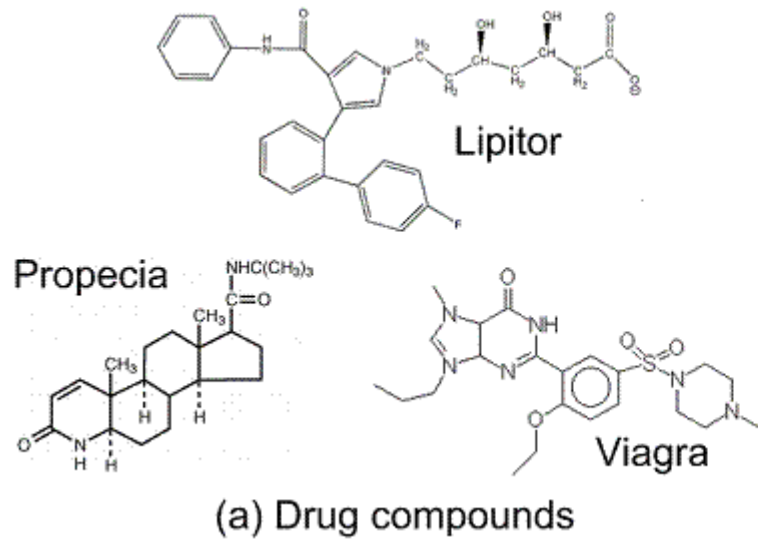
A screenshot of the Microsoft Excel interface. The ribbon at the top shows 'FILE', 'HOME', 'INSERT', 'PAGE LAYOUT', 'FORMULAS', and 'DATA'. The active cell is C18. The formula bar shows a multiplication formula. The spreadsheet contains a table with 5 columns: ParticipantID, Gender, Age, Hair Color, and Height (m). The data is as follows:

	A	B	C	D	E
	ParticipantID	Gender	Age	Hair Color	Height (m)
1	1	Male	18	Black	1.8
2	2	Female	14	Red	1.57
3	3	Male	17	Black	1.73
4	4	Male	22	Black	1.7
5	5	Male	23	Black	1.81
6	6	Female	21	Brown	1.63
7	7	Male	23	Red	1.7
8	8	Female	21	Brown	1.55
9	9	Female	19	Red	1.62
10	10	Female	20	Black	1.6
11	11	Male	18	Brown	1.71
12	12	Male	25	Brown	1.65
13	13	Female	13	Brown	1.56

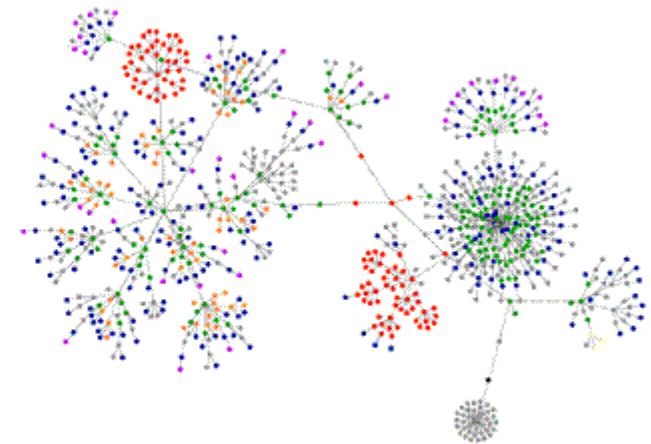
<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Types of Datasets

- Graph and network
 - World Wide Web
 - Social or information networks
 - Molecular Structures



(b) Protein structures

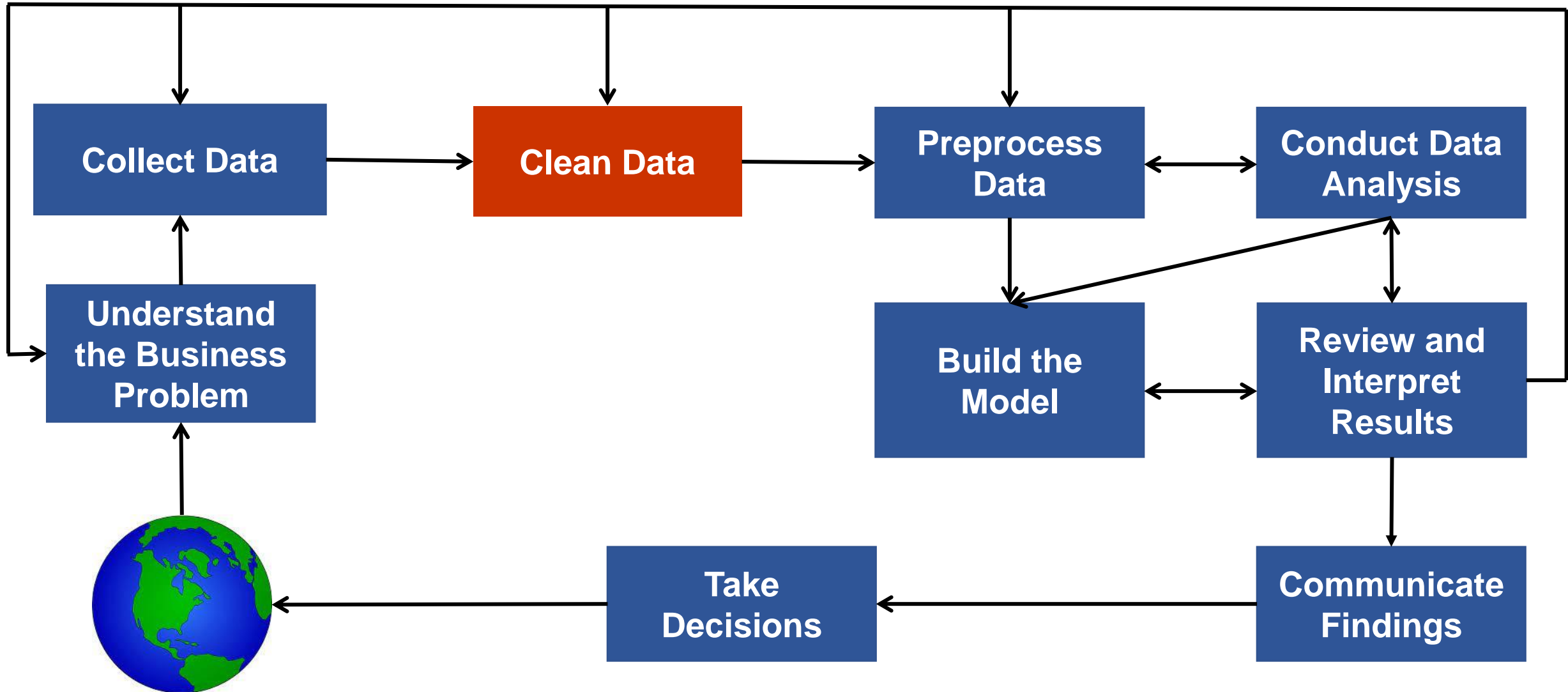


(c) Web site structure

Types of Datasets

- 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 Science Process



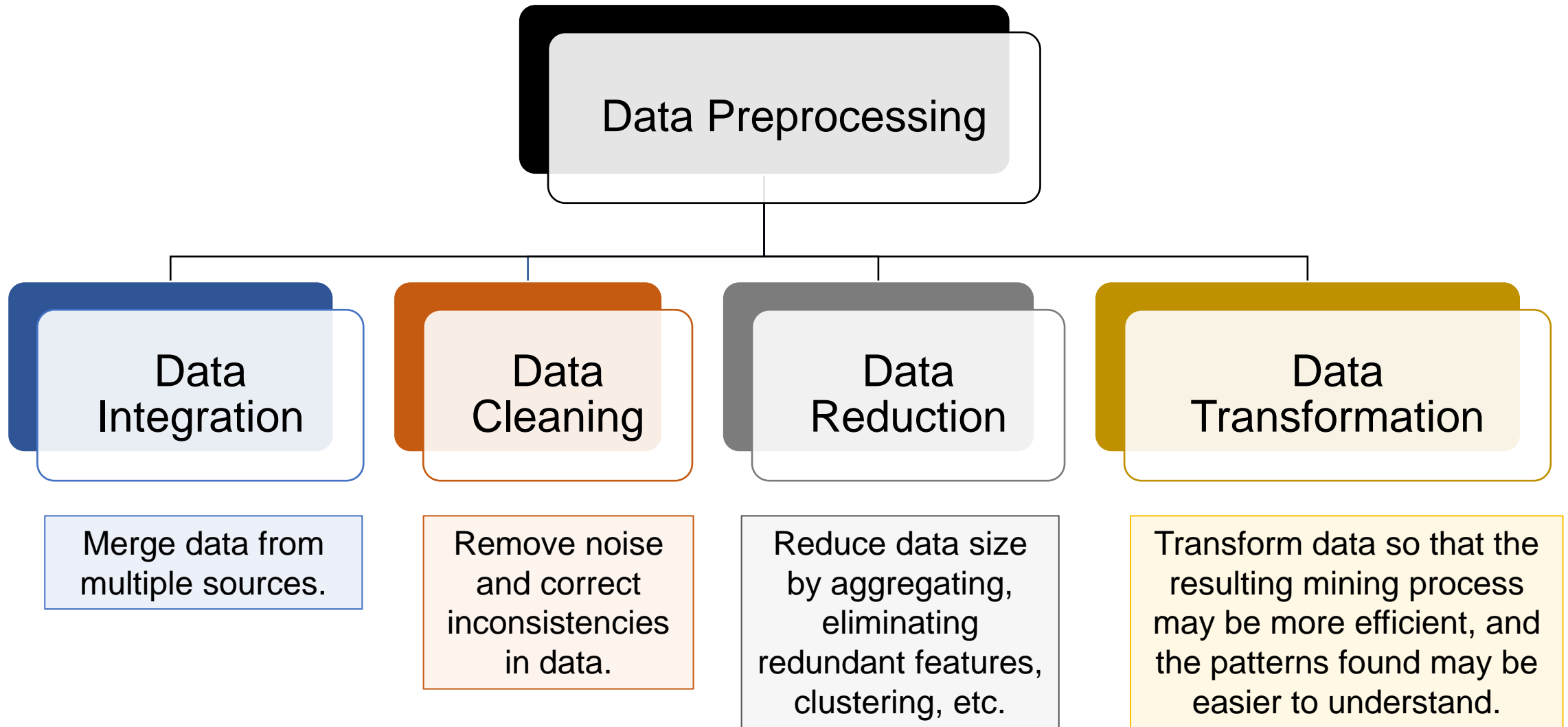
Why Preprocess Data?

Garbage In. Garbage Out.

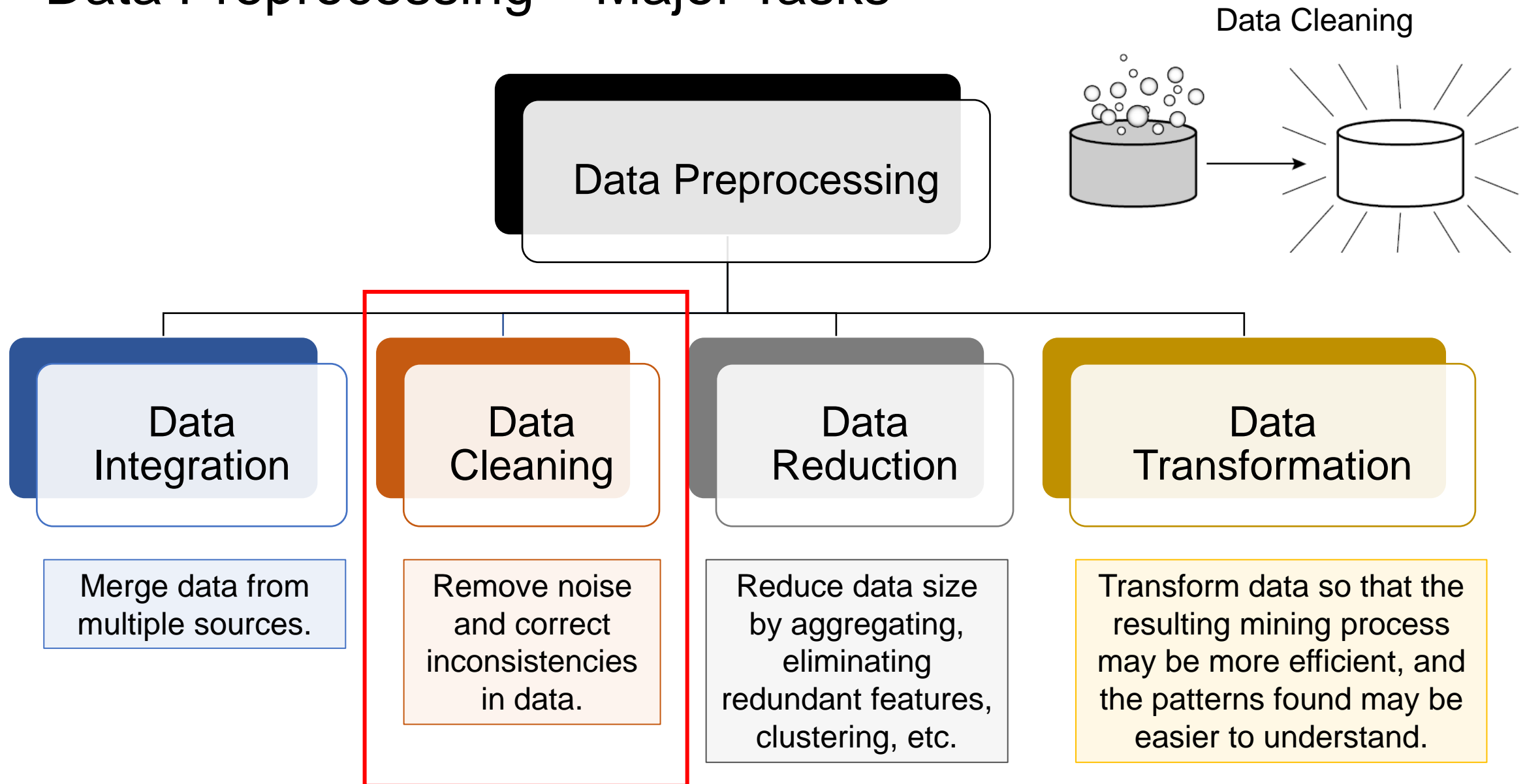


Low-quality data will lead to low-quality mining results.

Data Preprocessing – Major Tasks



Data Preprocessing – Major Tasks



Data Quality

- Data has quality if it satisfies the requirements of its intended use.
- A well-accepted multidimensional view:
- Accuracy :
 - The data should reflect actual, real-world scenarios; the measure of accuracy can be confirmed with a verifiable source.
- Completeness
 - A measure of the data's ability to effectively deliver all the required values that are available.
- Consistency / Reliability
 - Uniformity of data as it moves across networks and applications. The same data values stored in different locations should not conflict with one another.

Data Quality

- Timeliness :
 - Data that is available when it is required. Data may be updated in real-time to ensure that it is readily available and accessible.
- Believability
 - Trust by the users of the data
- Value added
 - Is it adding any value to already known things
- Interpretability
 - How easily can we understand the data
- Accessibility
 - Do we have access to the latest data if having latest data is relevant for the use

Example

Imagine that you are a manager at AllElectronics and have been charged with analyzing the company's data with respect to the sales at your branch. You immediately set out to perform this task. You carefully inspect the company's database and data warehouse, identifying and selecting the attributes or dimensions to be included in your analysis, such as item, price, and units sold. You notice that several of the attributes for various tuples have no recorded value. For your analysis, you would like to include information as to whether each item purchased was advertised as on sale, yet you discover that this information has not been recorded. Furthermore, users of your database system have reported errors, unusual values, and inconsistencies in the data recorded for some transactions.



Incomplete



Inaccurate



Inconsistent

Welcome to the real world!

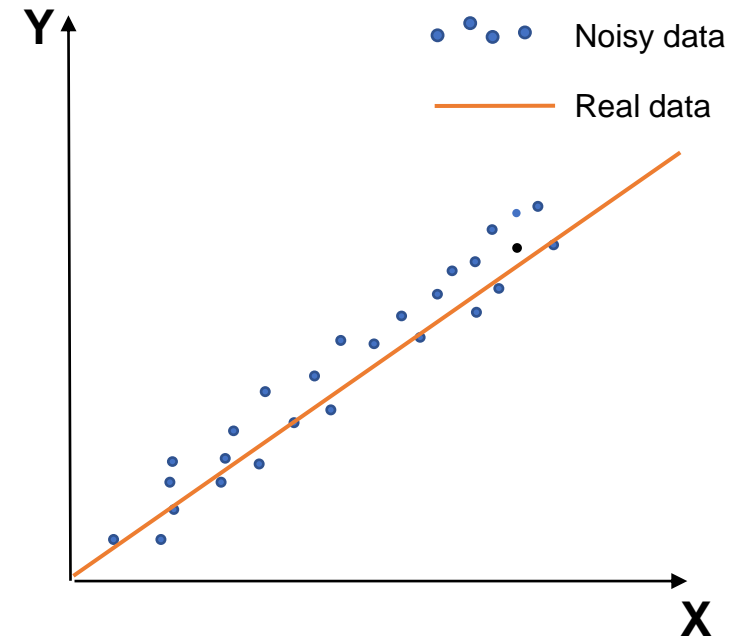
Source of Errors

- Human errors
 - Due to measurements (measure wrongly)
 - By accidents (spelling, typos, data entry errors, ...)
 - Due to human bias (e.g., coding disease code in hospital), etc.
- Errors due to machines
 - Software bugs
 - Data crawling errors
 - Data integration errors, etc....
- Others: errors may be deliberate, duplicates, stale data (an artifact of caching, not being refreshed).

Data Quality Issues

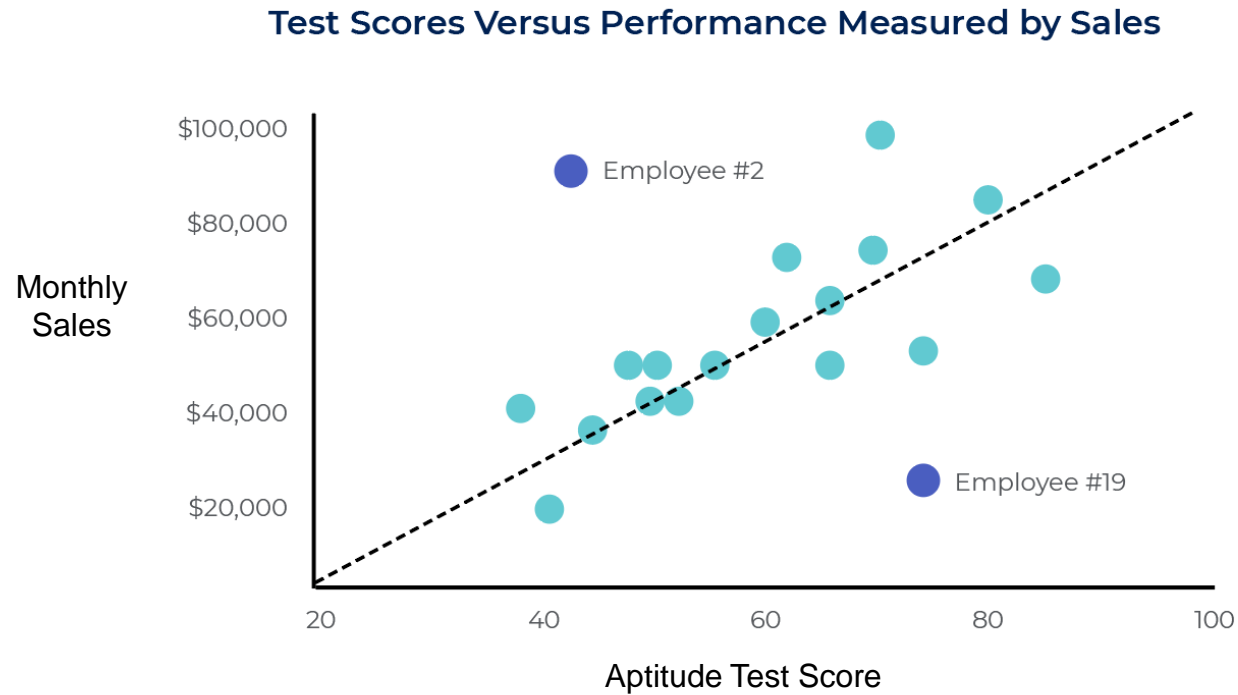
- Noise

- Noise refers to the **modification** of original values.
 - E.g., Distortion of a person's voice when talking on a poor phone
- What causes noise
 - Faults in instruments used for data collection, e.g., faulty sensors
 - Data entry problems
 - Data transmission problems
 - Technology limitations, e.g., cannot collect the real value
 - Inconsistencies in naming conventions



Data Quality Issues

- Outliers
 - Data objects with characteristics that are **considerably different** than most of the other data objects in the dataset.



<https://www.criteriacorp.com/resources/glossary/outlier>

Data Quality Issues

- Missing Values

- Data values are not always available - tuples may not have recorded values for one or more attributes.

- What causes missing values?

- Malfunctioning equipment.
 - E.g., faulty sensors.
- Information could not be collected.
 - E.g., people decline to provide their personal information.
- Attributes may not be applicable to all cases.
 - E.g., Annual income may not be applicable to children.
- Certain attributes maybe not be considered important at the time of entry.
 - *Remember data quality is accessed based on the intended use of the data!*

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			-
Carol		3	127	11	1	veggie	1	""
Mandy	44	2	68	8	1	chicken		null

<https://gallery.azure.ai/Experiment/Methods-for-handling-missing-values-1>

Data Quality Issues

- Duplicate Data
 - Datasets may consist of data objects that are duplicate or almost duplicates of one another
 - A major issue arises from data integration
 - E.g., Email addresses are not unique IDs.

Data Cleaning

- Data cleaning is the number 1 problem in data preprocessing.
- Data cleaning tasks
 - Handling missing values
 - Smoothing noisy data
 - Identifying and removing outliers
 - Correcting inconsistencies
 - Resolving redundancy caused by data integration

Data Cleaning – Handling Missing Values

- **Eliminate the tuple**

- Usually done when the class label is missing.
- Effective when a dataset has only a few tuples with missing values.

- **Fill in the missing values manually**

- Tedious and infeasible for large datasets with many missing values.

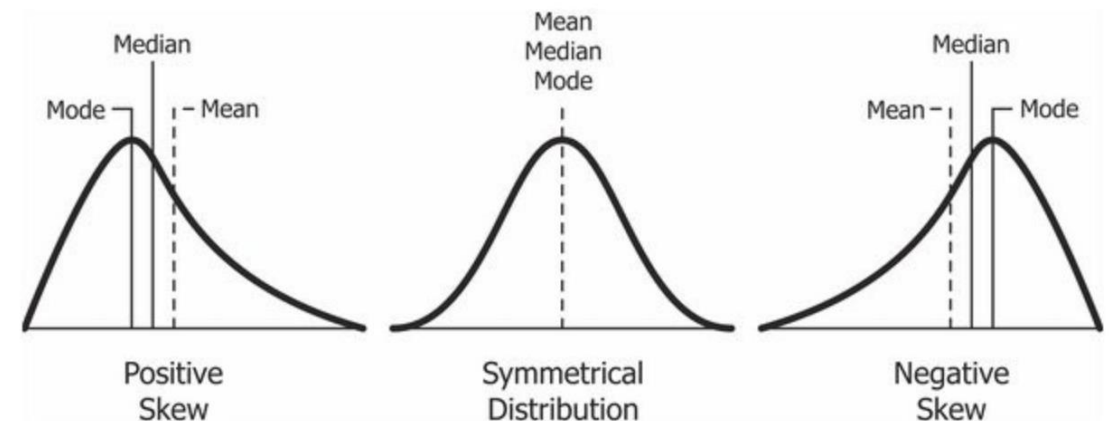
Data Cleaning – Handling Missing Values

- **Filling missing values automatically**

- A global constant (e.g., NA or “unknown”)
 - This may confuse the data mining algorithm to think that these tuples are common or similar.
- Use a measure of **central tendency** for the attribute
 - A measure of central tendency is a single value that describes an attribute indicating the central position of the values of that attribute.
 - **Mean**

$$Mean = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- When not to use the mean?
 - In the presence of outliers
 - When data distribution is skewed.



Data Cleaning – Handling Missing Values

- **Filling missing values automatically**
 - A global constant (e.g., NA or “unknown”)
 - This may confuse the data mining algorithm to think that these tuples are common or similar.
 - Use a measure of **central tendency** for the attribute
 - A measure of central tendency is a single value that describes an attribute indicating the central position of the values of that attribute.
 - **Mean**
 - **Median** – The middle value when values for an attribute are sorted.
 - Better representation than mean when there are outliers or the data distribution is skewed.
 - **Mode** - The mode is the value that appears most often in a set of data values. (Most popular option)

Data Cleaning – Handling Missing Values

- **Filling missing values automatically**
 - A global constant (e.g., NA or “unknown”)
 - This may confuse the data mining algorithm to think that these tuples are common or similar.
 - Use a measure of **central tendency** for the attribute

Type of Variable	Best Measure of Central Tendency
Nominal	Mode
Ordinal	Median
Metric (Fewer outliers and not skewed)	Mean
Metric (Fewer outliers and/or not skewed)	Median

Data Cleaning – Handling Missing Values

- **Filling missing values automatically**
 - Use a measure of central tendency for the attribute for samples belonging to the same class
 - Suitable for classification tasks
 - Most probable value
 - Inference based, such as Bayesian formula, decision tree, or other classification/regression models
 - Could be time-consuming

Data Cleaning – Handling Missing Values

- **Filling missing values automatically**
 - Bias the data.
 - The filled-in value may not be correct
 - Most probable value - Popular
- *A missing value is not always an error!*

Data Cleaning – Smoothing Noisy Data

- **Binning**

- Binning methods smooth a sorted data value by consulting its “neighborhood.”
- First, sort data and partition into (equal frequency) bins
- Then smooth by **bin means, bin median, or bin boundaries**
- E.g., Sorted data for Price (in \$) – 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
 - Partition into equal-frequency bins:
 - Bin 1: 4, 8, 9, 15 $(4+8+9+15)/4 = 9$
 - 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

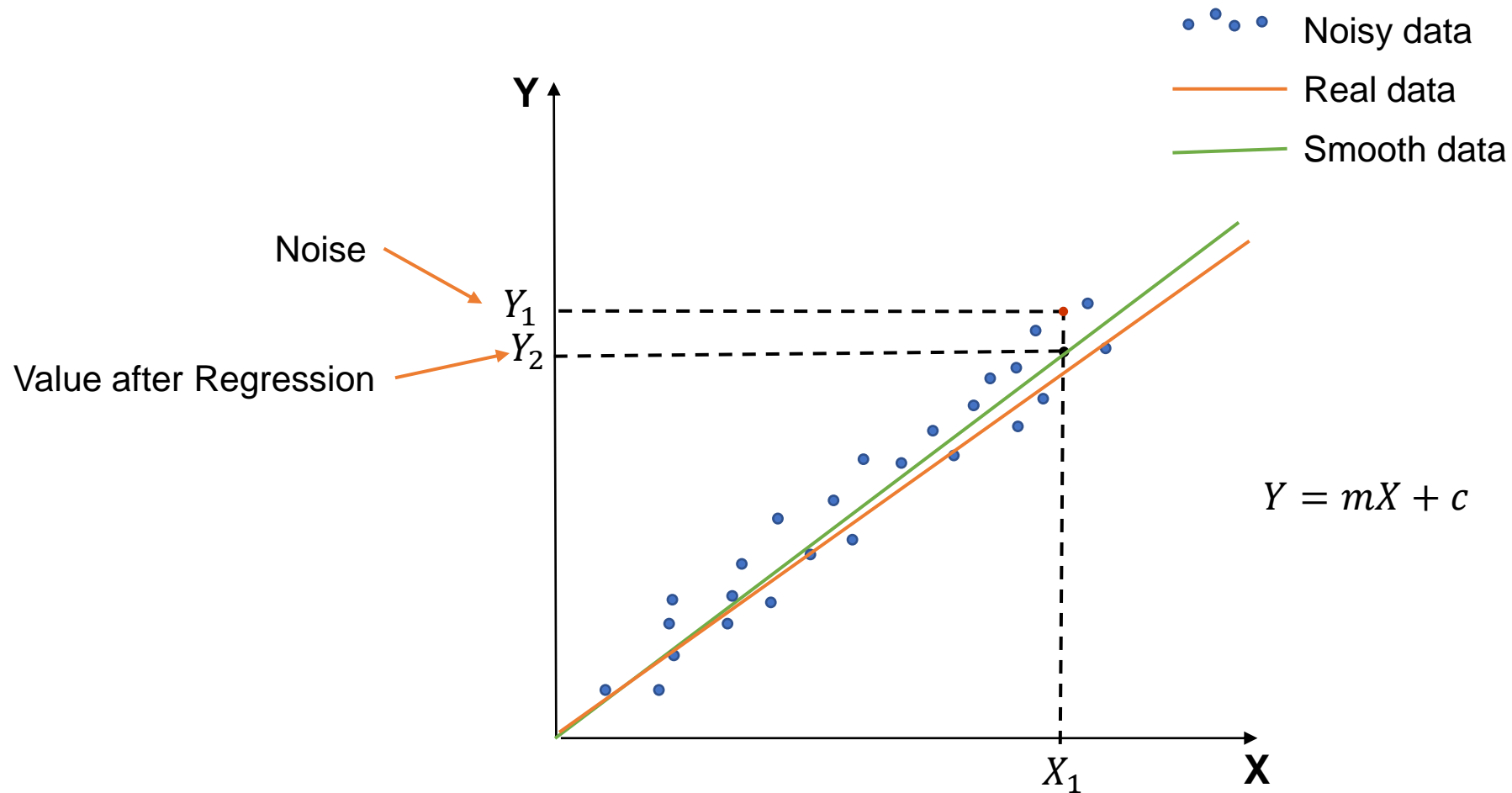
Data Cleaning – Smoothing Noisy Data

- **Binning**

- First sort data and partition into (equal frequency) bins
- Then smooth by bin means, bin median, or bin boundaries
- E.g., Sorted data for Price (in \$) – 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into equal-frequency bins:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
- **Smoothing by bin boundaries**
 - Each value needs to check which boundary it nears to.
 - Bin 1: 4, 4, 4, 15 (8 and 9 are more closed to 4 than 15)
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 26, 34
 - *Bin boundaries preserve more information than the bin means.*

Data Cleaning – Smoothing Noisy Data

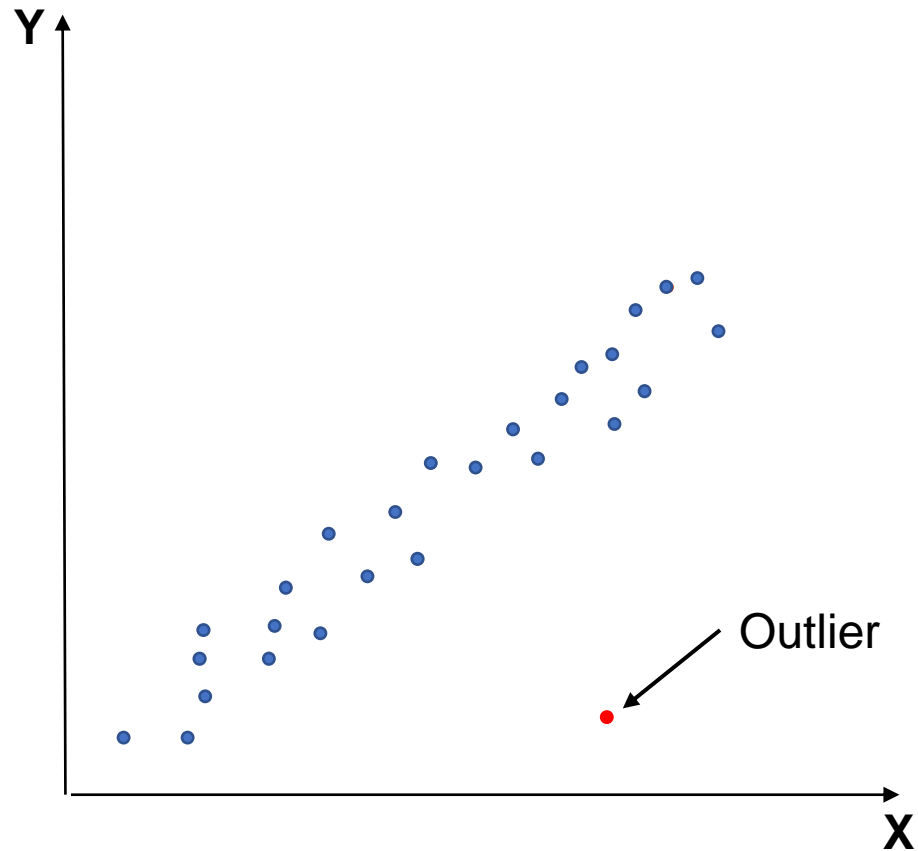
- **Regression**



Data Cleaning – Identifying Outliers (Noise)

- **Data visualization**

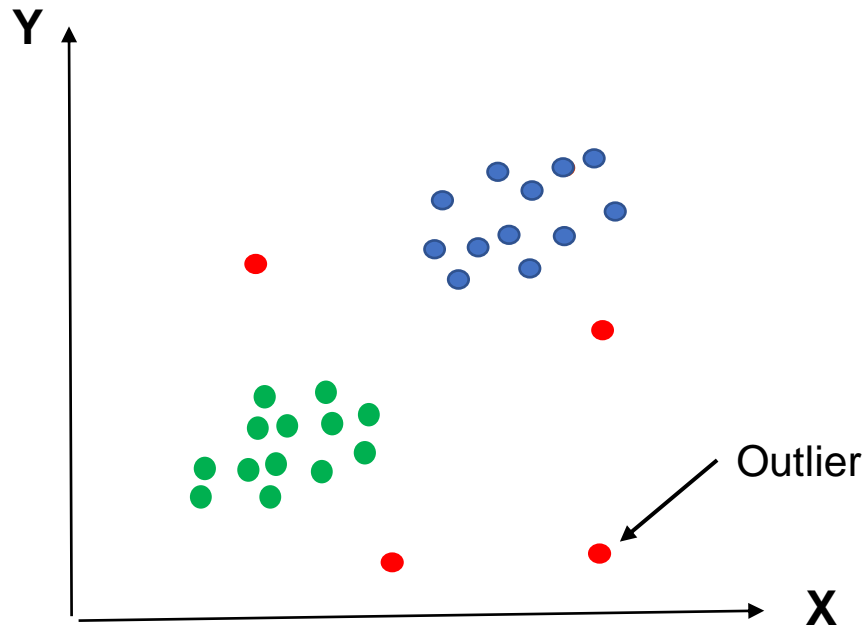
- More suitable for low-dimensional datasets



Data Cleaning

- **Clustering**

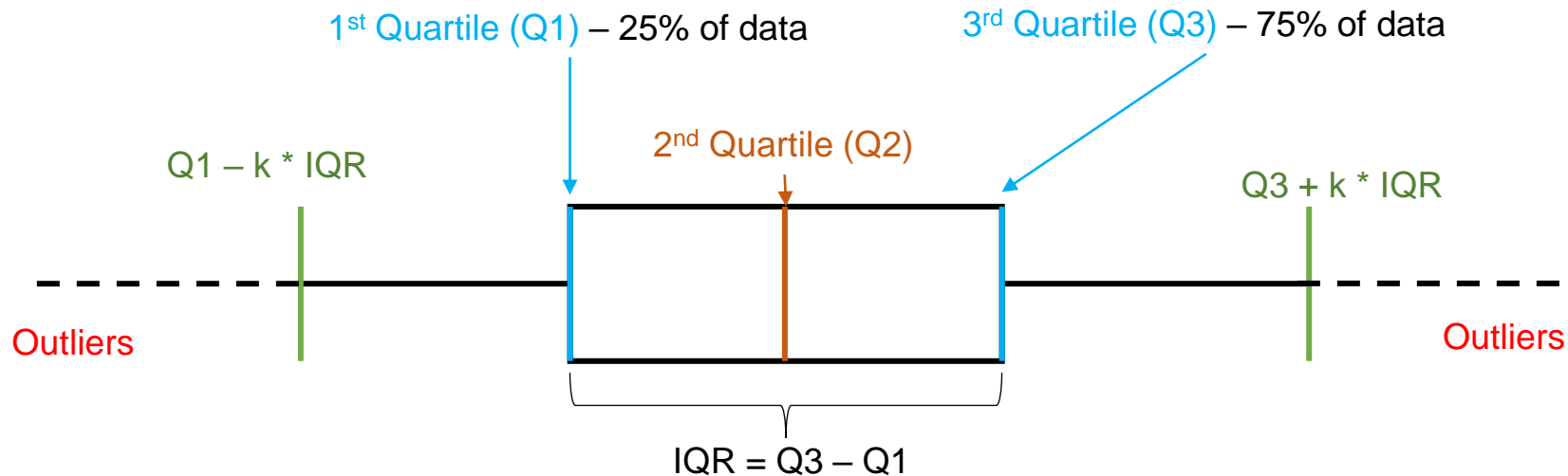
- Use a clustering technique to cluster the dataset.
- The points that are too far away from the cluster centers would be outliers.
- Be mindful of the clustering technique used. E.g., DBSCAN is widely used.



Data Cleaning

- **Quartile Analysis**

- We define the values falling outside the $k \cdot \text{IQR}$ range as outliers.
- A commonly used value for the multiplier k is 1.5.



Thank You!
Questions?