

Introduction to Data Science



Chapter 2

Data Collection and Acquisition

Papangkorn Inkeaw, PhD

Department of Computer Science, Faculty of Science
Chiang Mai University

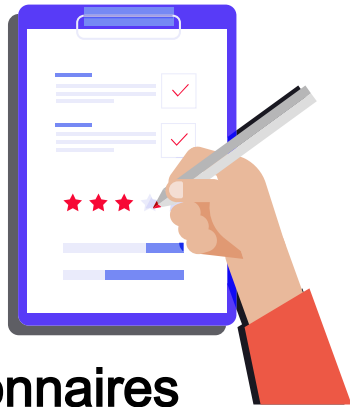


Outline

Data Collection and Acquisition

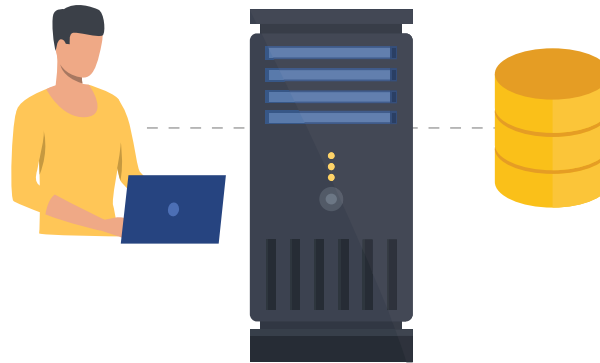
1. Data Sources
2. Data Representation
 - Data Matrix
 - Types of Data
 - Attributes
3. Preparing Data
 - Encoding of Categorical Data
 - Data Quality
 - Data Cleaning
 - Inconsistent Datatypes
 - Missing data
 - Duplicate data

Data Sources



Questionnaires

- Paper-based questionnaires
- Electronic-based questionnaires
- Online questionnaires



Web Servers

Server software, or hardware dedicated to running said software, that can satisfy World Wide Web client requests.



Web Services

A service offered by an electronic device to another electronic device, communicating with each other via the World Wide Web

Data Sources



Database

An organized collection of data, generally stored and accessed electronically from a computer system



Logs

- Records of events.
- In computer, for example, a file that records either events that occur in an operating system or other software runs, or messages between different users of a communication software.



Online Repositories

- A repository is a central place in which an aggregation of data is kept and maintained in an organized way, usually in computer storage.
- An online repository is a digital library or archive which is accessible via the internet.

Data Matrix

Data Representation

Example: Cosmic Dataset

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
x_1	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
x_2	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel

x_n	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel



Data Matrix

Data Representation

$$\text{Dataset } D = \begin{array}{cccc} & \underbrace{\hspace{10em}}_{\text{Attributes}} & & \\ & X_1 & X_2 & \dots & X_d & & \\ \left(\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{array} \right) & \mathbf{x}_1 & \mathbf{x}_2 & & \mathbf{x}_n \end{array} \left. \vphantom{\begin{array}{cccc} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nd} \end{array}} \right\} \begin{array}{l} \text{Entities, Instances,} \\ \text{Examples, Records,} \\ \text{Points, Feature vectors, ...} \end{array}$$

\mathbf{x}_i denotes the i th row which is a d -tuple given as

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})$$

X_j denotes the j th column which is a n -tuple given as

$$X_j = (x_{1j}, x_{2j}, \dots, x_{nj})$$

Data Matrix

Data Representation

Example: Cosmic Dataset

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
x_1	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
x_2	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel

x_n	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel



We can write an example x_2 as

$x_2 = (\text{Captain America (Steven Rogers)}, \text{Public}, \text{Good}, \text{Blue Eyes}, \text{White Hair}, \text{Male}, \text{Living Characters}, 3360, \text{Mar} - 41, \text{marvel})$

Types of Data

Data Representation



Quantitative Data

- This data can be described using **numbers**.
- **Basic mathematical procedures** are possible on the set.



Qualitative Data

- This data cannot be described using numbers and basic mathematics.
- This data is generally described using natural **categories and language**.

Attributes

Data Representation



Numeric Attributes - Quantitative

- One that has a real-valued or integer-valued domain.
- Such as age, height, grade, frequency, etc.

Categorical Attributes

- One that has a set-valued domain composed of a set of symbols.
- Such as Gender = {M,F},
Education = {High School, BS, MS, PhD},
etc.

Discrete

- Take on a finite or countably infinite set
- Such as integer, grade, number of object, etc.

Continuous

- Take on any real value
- Such as height, weight, size, etc.

Attributes

Data Representation



Categorical Attributes

Nominal

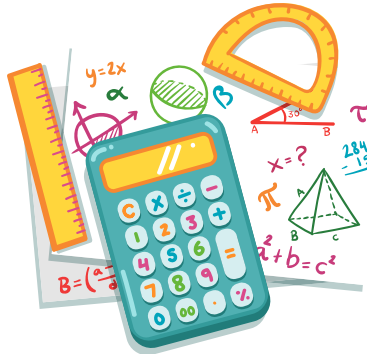
- Attribute values in the domain are unordered.
- Can only equality (=) compare.
- Such as gender, type of hair, etc.

Ordinal

- Attribute values are ordered.
- Can both equality (=) and inequality (<, >) compare.
- Such as education, feel (unhappy, OK, happy), etc.

Attributes

Data Representation



Numeric Attributes

Interval-scaled

- Can compute only differences (addition or subtraction)
- For example, temperature measured in °C or °F.
 - If it is 20 °C on one day and 10 °C on previous day
 - We **can** talk about a temperature drop of 10°C.
 - We **cannot** say that it is twice as cold as the previous day.

Ratio-scaled

- Can compute both differences and ratio between values,
- For example age.
 - If Jone is 20 years old and Jim is 10 years old.
 - We **can** say that Jone older than Jim with 10 years.
 - We **can** say that Jone is twice as old as Jim.

Attributes

Data Representation

Summary of data types and scale measures

Provides	Nominal	Ordinal	Interval-scaled	Ratio-scaled
The order of values is known		/	/	/
“Count,” aka “Frequency of Distribution”	/	/	/	/
Mode	/	/	/	/
Median		/	/	/
Mean			/	/
Can quantify the difference between each values			/	/
Can add or subtract values			/	/
Can multiple and divide values				/
Has “true zero”				/

<https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/>

Attributes

Data Representation

Cosmic Dataset

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
x_1	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
x_2	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel

x_n	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel



Practice: What is the type of each attribute?

Nominal, Ordinal, Interval-scaled or Ratio-scaled

Encoding of Categorical Data


Preparing Data

- Most of Machine learning algorithms can not handle categorical variables.
→ We convert them to numerical values.

Nominal variable

One Hot Encoding

- Map each category to a vector that contains 1 and 0
 - 1 - presence of the feature
 - 0 - absence of the feature

Gender		isMale	isFemale	isOther
Male		1	0	0
Female		0	1	0
Other		0	0	1

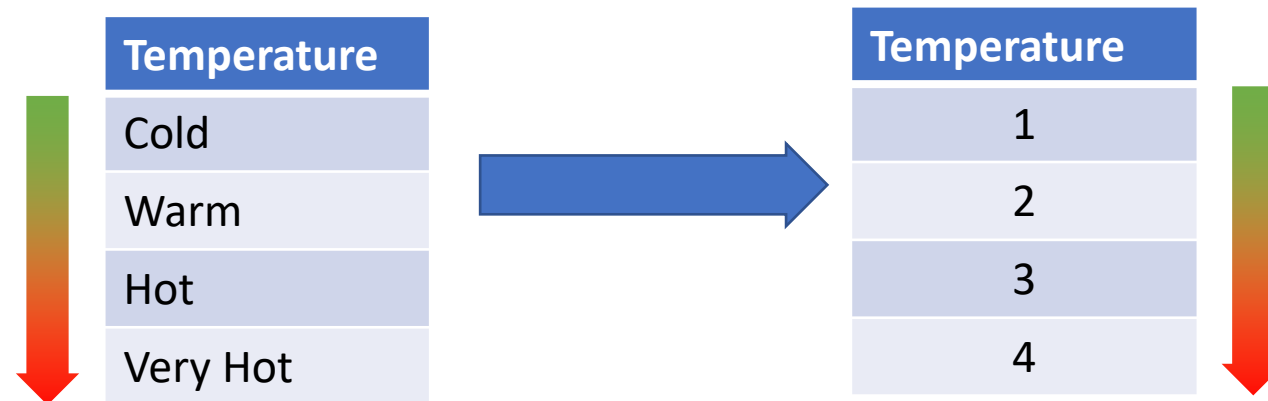
Encoding of Categorical Data

Preparing Data

Ordinal

Ordinal Encoding

- The encoding of variables retains the ordinal nature of the variable
- Each category is assigned a value from 1 through the number of possible values by considering the order of values.



Encoding of Categorical Data

Preparing Data

Practice

How can we encode the following categorical data?



Align
Bad
Neutral
Good

Hair
Black
Bronze
Brown
Gold
Gray

Data Quality

Preparing Data



Source:

<http://itsadeliverything.com/wordpress/images//accuracy-vs-precision.jpg>



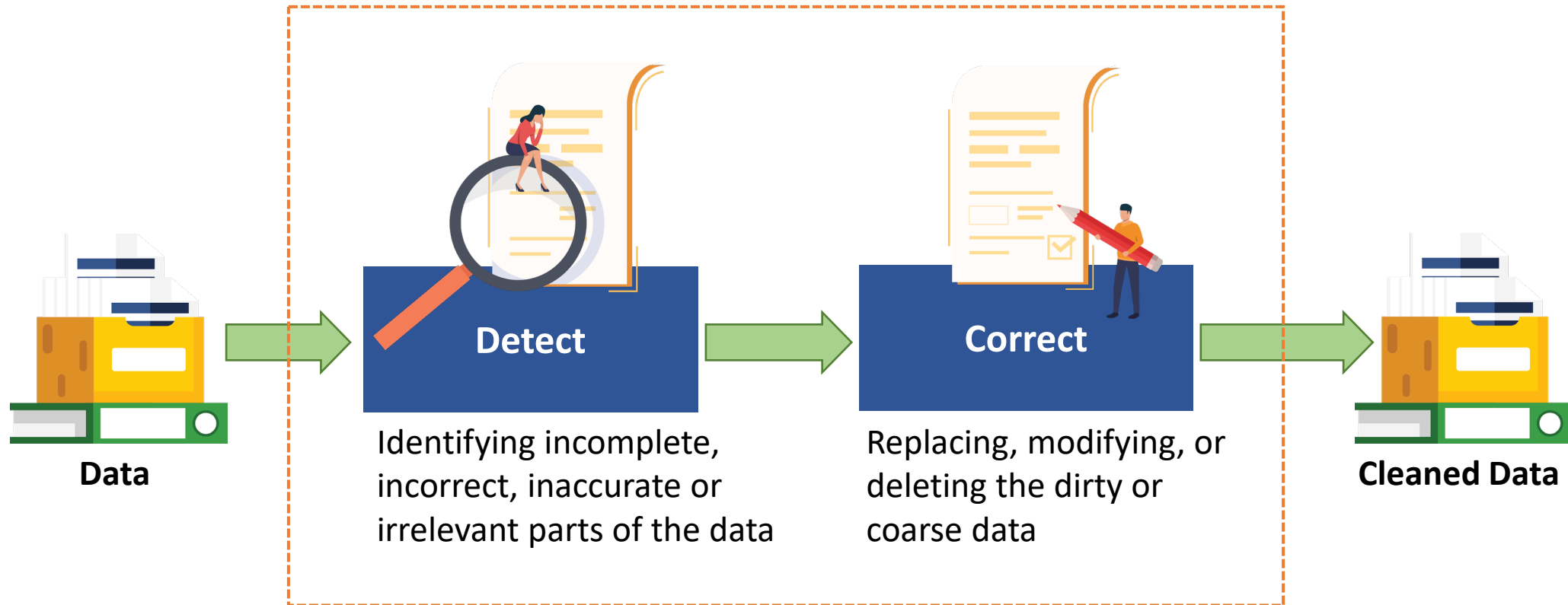
Data should be:

- **Accurate and Precise**
- **Complete** – Does not have "unknown" or "missing" values
- **Consistency** – Two data items in the data set contradict each other
- **Valid** – Conform to defined business rules or constraints
- **Uniform** – Using the same units of measure in all systems
- **Unique** – Does not contain duplicates

Data Cleaning

Preparing Data

Data Cleaning is the process of detecting and correcting/removing corrupt or inaccurate records from a record set



Inconsistent Datatypes

Preparing Data >> Data Cleaning

We expect that:

Values in a particular attribute must be of a particular datatype, e.g., Boolean, numeric (integer or real), date, etc.

Values in *align* and *alive* are inconsistent datatype

1 – Living Characters
0 – Deceased Characters

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
x_1	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	1	4043	Aug-62	marvel
x_2	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel

x_n	Natalia Romanova (Earth-616)	Public	1	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

1 – Good
0 – Bad

Inconsistent Datatypes

Preparing Data >> Data Cleaning

How to address the Inconsistent datatypes

- Choose an appropriate datatype
- Transform values in another datatype into the selected datatype

Values in *align* and *alive* are inconsistent datatype

1 – Living Characters
0 – Deceased Characters

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
x_1	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
x_2	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel

x_n	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

1 – Good
0 – Bad

Missing Value

Preparing Data >>> Data Cleaning

We expect that:

All required measures are known.

	IQ X_1	Job performance X_2
x_1	78	NA
x_2	84	NA
x_3	84	NA
x_4	85	NA
x_5	99	7
x_6	105	10
x_7	105	11
x_8	106	15
x_9	108	10
x_{10}	112	10
x_{11}	113	12
x_{12}	115	14
x_{13}	118	16
x_{14}	134	12

Job performances of x_1, x_2, x_3 and x_4 are unknown. They are missing value.

Missing Value

Preparing Data >> Data Cleaning

How to deal with the missing value

Single Imputation: Generate a single replacement value for each missing data point.

- **Arithmetic Mean Imputation**
 - replaces missing values with mean of available values
- **Regression Imputation**
 - replaces missing values with predicted scores from a regression equation
- **Hot-deck Imputation**
 - A collection of techniques that impute the missing values with scores from “similar” datapoints, such as nearest neighbor hot-deck and last observation carried forward.
- **and etc.**

Missing Value

Preparing Data >> Data Cleaning

	IQ X_1	Job performance X_2
x_1	78	11.70
x_2	84	11.70
x_3	84	11.70
x_4	85	11.70
x_5	99	7
x_6	105	10
x_7	105	11
x_8	106	15
x_9	108	10
x_{10}	112	10
x_{11}	113	12
x_{12}	115	14
x_{13}	118	16
x_{14}	134	12

Example of Arithmetic Mean Imputation

1. Compute the arithmetic mean of X_2 from available values
2. Replace the missing values of X_2 by the arithmetic mean

Mean = 11.70

Missing Value

Preparing Data >> Data Cleaning

	IQ X_1	Job performance X_2
x_1	78	7.529
x_2	84	8.267
x_3	84	8.267
x_4	85	8.390
x_5	99	7
x_6	105	10
x_7	105	11
x_8	106	15
x_9	108	10
x_{10}	112	10
x_{11}	113	12
x_{12}	115	14
x_{13}	118	16
x_{14}	134	12

$$JP = 0.123(78) + (-2.065) = 7.529$$

$$JP = 0.123(84) + (-2.065) = 8.267$$

$$JP = 0.123(84) + (-2.065) = 8.267$$

$$JP = 0.123(85) + (-2.065) = 8.390$$

$$JP = \beta_1(IQ) + \beta_0 = 0.123(IQ) + (-2.065)$$

incomplete variables

complete variables

Example of Regression Imputation

1. Estimate a set of regression equations
2. Generate predicted values for the incomplete variables
3. Fill in the missing values

Duplicate Data

Preparing Data >> Data Cleaning

We expect that:

A data should appear on the dataset one time

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
x_1	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
x_2	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
x_3	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Black Hair	Male	Living Characters	NA	Aug-62	marvel

x_{100}	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

We have two recodes of Spider-Man. So, the two recodes are duplicate data

Moreover, one contradicts each other

Duplicate Data

Preparing Data >> Data Cleaning

How to deal with the duplicate data

1. Select one recode that is up-to-date and accurate
2. Remove the others

	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
x_1	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
x_2	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel

x_{100}	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

We have two recodes of Spider-Man. So, the two recodes are duplicate data

Practice

Problem

- จงบอกชนิดข้อมูลของแอตทริบิวต์
 - Eye (Categorical / Numerical)
 - Hair (Categorical / Numerical)
 - Number of appearances (Categorical / Numerical)
- จงบอกวิธีการจัดการค่าข้อมูลสูญหาย (Missing Value) ที่เหมาะสม ของแอตทริบิวต์ต่อไปนี้
 - Eye
 - Hair
 - Number of appearances

Name	Eye	Hair	Number of appearances
Captain America	Blue	NA	3360
Thor	NA	NA	NA
Benjamin Grimm	Blue	NA	2255
Reed Richards	Brown	Brown	2072
Hulk	Brown	NA	2017
Scott Summers	Brown	Brown	1955
Jonathan Storm	Blue	NA	NA
Robert Drake	Brown	NA	1265
*NA – Missing Value			

Further Study

- **Book:**

- Zaki, M., & Meira, W. (2014). Data mining and analysis : Fundamental concepts and algorithms. New York: Cambridge University Press.
- Enders, C. (2010), Applied Missing Data Analysis. New York: Guilford Press.
- Sunil Kakade & Sinan Ozdemir (2018). Principles of Data Science. UK: Packt Publishing.

- **Website**

- <https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02>