Introduction to Data Science

Chapter 2
Data Collection and Acquisition

Papangkorn Inkeaw, PhD

Department of Computer Science, Faculty of Science

Chiang Mai University

Outline

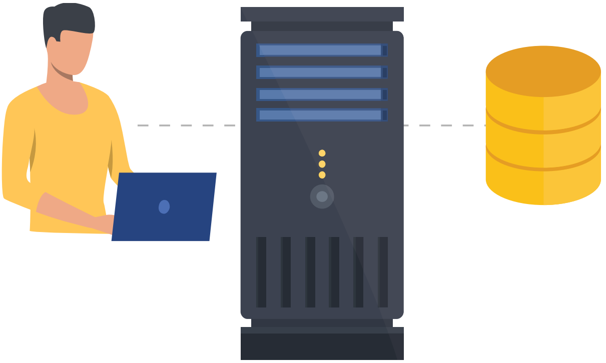
Data Collection and Acquisition

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

Data Sources

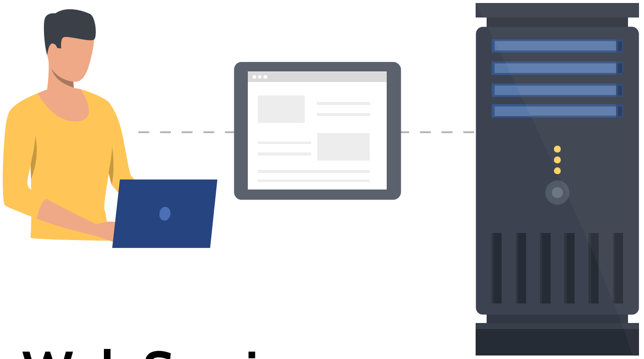
Questionnaires

- 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 Sources

Suggested Data Sources

- **UCI Machine Learning Repository**
  https://archive.ics.uci.edu/ml/index.php
- **Kaggle**
  https://www.kaggle.com/datasets
- **Open Government Data of Thailand**
  https://data.go.th/

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}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

Data Matrix

Data Representation

$$
\textbf{Dataset}\quad D = \begin{array}{c}
\overbrace{\begin{matrix} X_1 & \quad X_2 & \qquad & \qquad X_d \end{matrix}}^{\textbf{Attributes}} \\
\begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1d} \\
x_{21} & x_{22} & \cdots & x_{2d} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nd}
\end{pmatrix}
\end{array}
\left.\begin{matrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \\ \mathbf{x}_n \end{matrix}\right\}
\begin{array}{l}\textbf{Entities, Instances,} \\ \textbf{Examples, Records,} \\ \textbf{Points, Feature-vectors, ...}\end{array}
$$

$\mathbf{x}_i$ 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 jth column which is a n-tuple given as

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

Data Matrix

Data Representation

Example: Cosmic Dataset

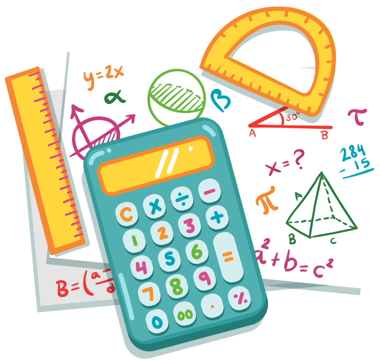
	name X_1	id X_2	align X_3	eye X_4	hair X_5	gender X_6	alive X_7	appearances X_8	first_appear X_9	publisher X_{10}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	Natalia Romanova (Earth-616)	Public	Good	Green Eyes	Red Hair	Female	Living Characters	1050	Apr-64	marvel

We can write an example $\mathbf{x}_2$ as

$$\mathbf{x}_2 = (Captain\ America\ (Steven\ Rogers), Public, Good, Blue\ Eyes, White\ Hair, \\ Male, Living\ Characters, 3360, Mar-41, marvel)$$

Types of Data

Data Representation



Quantitative Data

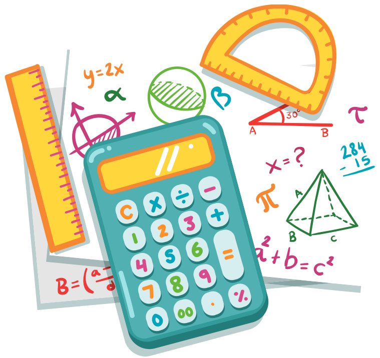
- This data can be described using **numbers.**
- **Basic mathematical procedures** are possible on the set.

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.

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.

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.

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}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	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
Male
Female
Other

isMale	isFemale	isOther
1	0	0
0	1	0
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.

Temperature
Cold
Warm
Hot
Very Hot

Temperature
1
2
3
4

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

Normalization and Standardization

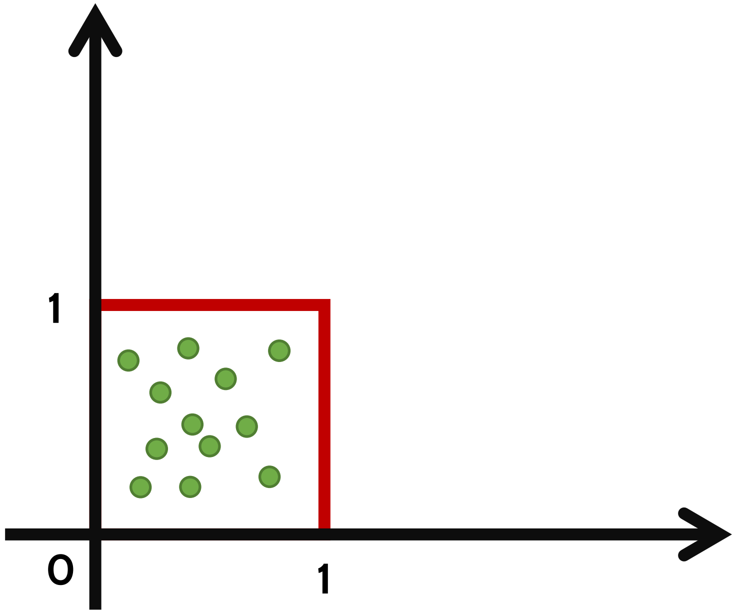
Preparing Data

Normalization

Scale a variable to have a values between 0 and 1

Min-Max Normalization:

$$x_{normalized} = \frac{x - x_{min}}{x_{max} - x_{min}}$$



Normalization and Standardization

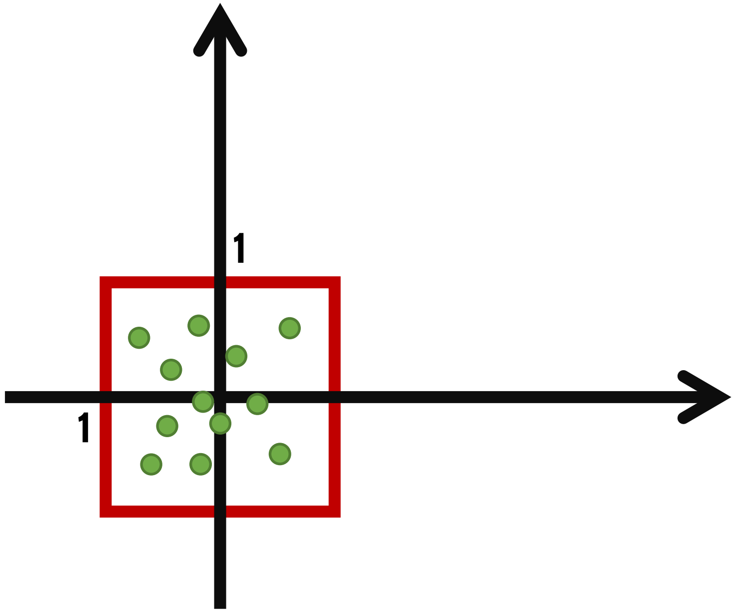
Preparing Data

Standardization

Transforms data to have a mean of zero and a standard deviation of 1.

Z-score Standardization

$$x_{standardized} = \frac{x - \bar{x}}{S.D.}$$



Data Quality

Preparing Data

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

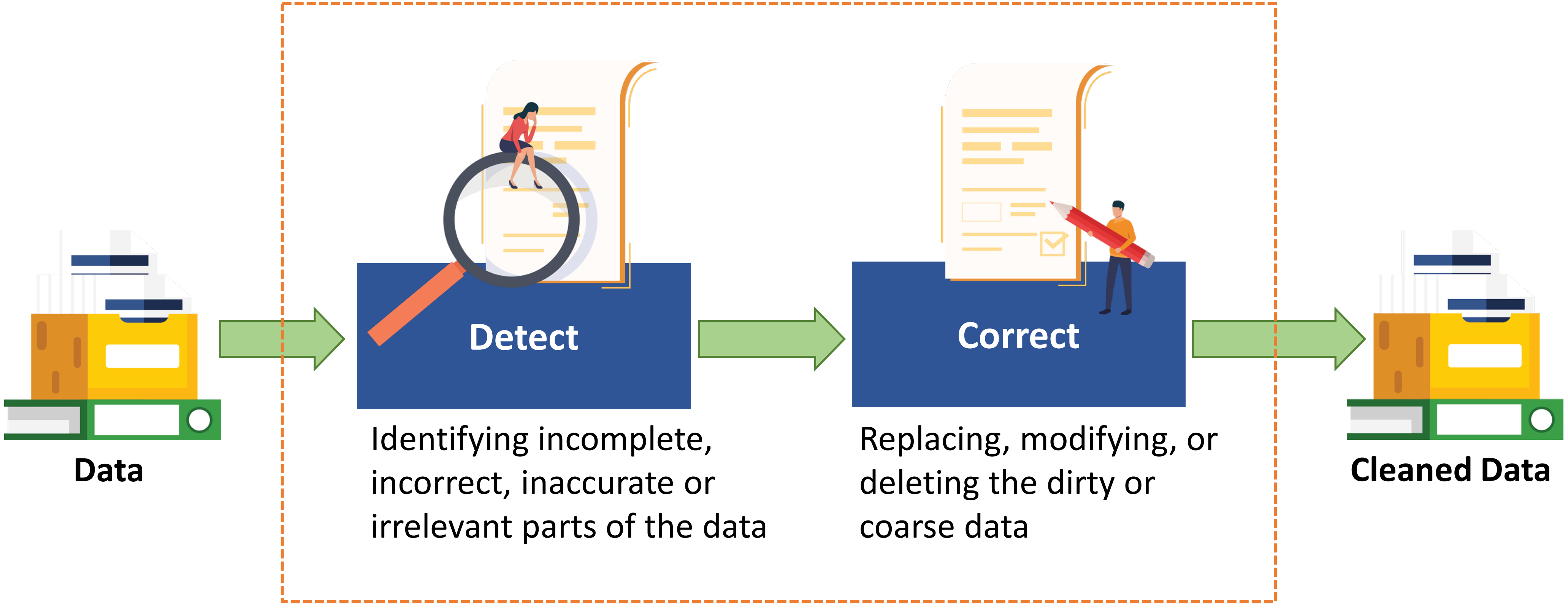
Data should be:

- **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}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	1	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	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}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_n$	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
$\mathbf{x}_1$	78	NA
$\mathbf{x}_2$	84	NA
$\mathbf{x}_3$	84	NA
$\mathbf{x}_4$	85	NA
$\mathbf{x}_5$	99	7
$\mathbf{x}_6$	105	10
$\mathbf{x}_7$	105	11
$\mathbf{x}_8$	106	15
$\mathbf{x}_9$	108	10
$\mathbf{x}_{10}$	112	10
$\mathbf{x}_{11}$	113	12
$\mathbf{x}_{12}$	115	14
$\mathbf{x}_{13}$	118	16
$\mathbf{x}_{14}$	134	12

Job performances of $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$ and $\mathbf{x}_4$ are unknow.
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
$\mathbf{x}_1$	78	11.70
$\mathbf{x}_2$	84	11.70
$\mathbf{x}_3$	84	11.70
$\mathbf{x}_4$	85	11.70
$\mathbf{x}_5$	99	7
$\mathbf{x}_6$	105	10
$\mathbf{x}_7$	105	11
$\mathbf{x}_8$	106	15
$\mathbf{x}_9$	108	10
$\mathbf{x}_{10}$	112	10
$\mathbf{x}_{11}$	113	12
$\mathbf{x}_{12}$	115	14
$\mathbf{x}_{13}$	118	16
$\mathbf{x}_{14}$	134	12

Mean = 11.70

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

Missing Value

Preparing Data >> Data Cleaning

	IQ X_1	Job performance X_2
$\mathbf{x}_1$	78	7.529
$\mathbf{x}_2$	84	8.267
$\mathbf{x}_3$	84	8.267
$\mathbf{x}_4$	85	8.390
$\mathbf{x}_5$	99	7
$\mathbf{x}_6$	105	10
$\mathbf{x}_7$	105	11
$\mathbf{x}_8$	106	15
$\mathbf{x}_9$	108	10
$\mathbf{x}_{10}$	112	10
$\mathbf{x}_{11}$	113	12
$\mathbf{x}_{12}$	115	14
$\mathbf{x}_{13}$	118	16
$\mathbf{x}_{14}$	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}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
$\mathbf{x}_3$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Black Hair	Male	Living Characters	NA	Aug-62	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_{100}$	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 <u>duplicate data</u>

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}
$\mathbf{x}_1$	Spider-Man (Peter Parker)	Secret	Good	Hazel Eyes	Brown Hair	Male	Living Characters	4043	Aug-62	marvel
$\mathbf{x}_2$	Captain America (Steven Rogers)	Public	Good	Blue Eyes	White Hair	Male	Living Characters	3360	Mar-41	marvel
	…	…	…	…	…	…	…	…	…	…
$\mathbf{x}_{100}$	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