Data Wrangling and Transformation

Riki Akbar
Ibrahim Saleh Siregar

About Us

Riki Akbar

Direktorat TIK - DJP

Ibrahim Saleh Siregar

Direktorat DIP - DJP

Standar Kompetensi

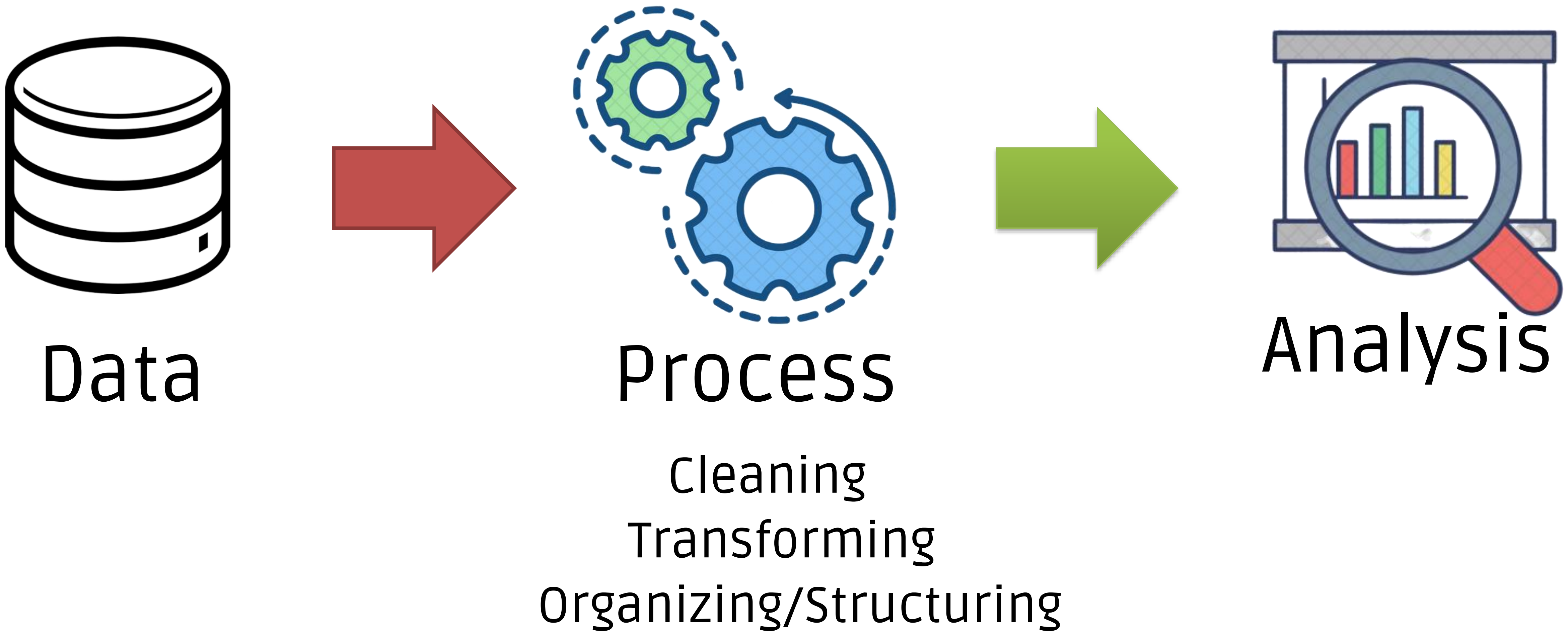
Konsep
Data Wrangling
Fundamentals

Advanced
Data Wrangling
Techniques

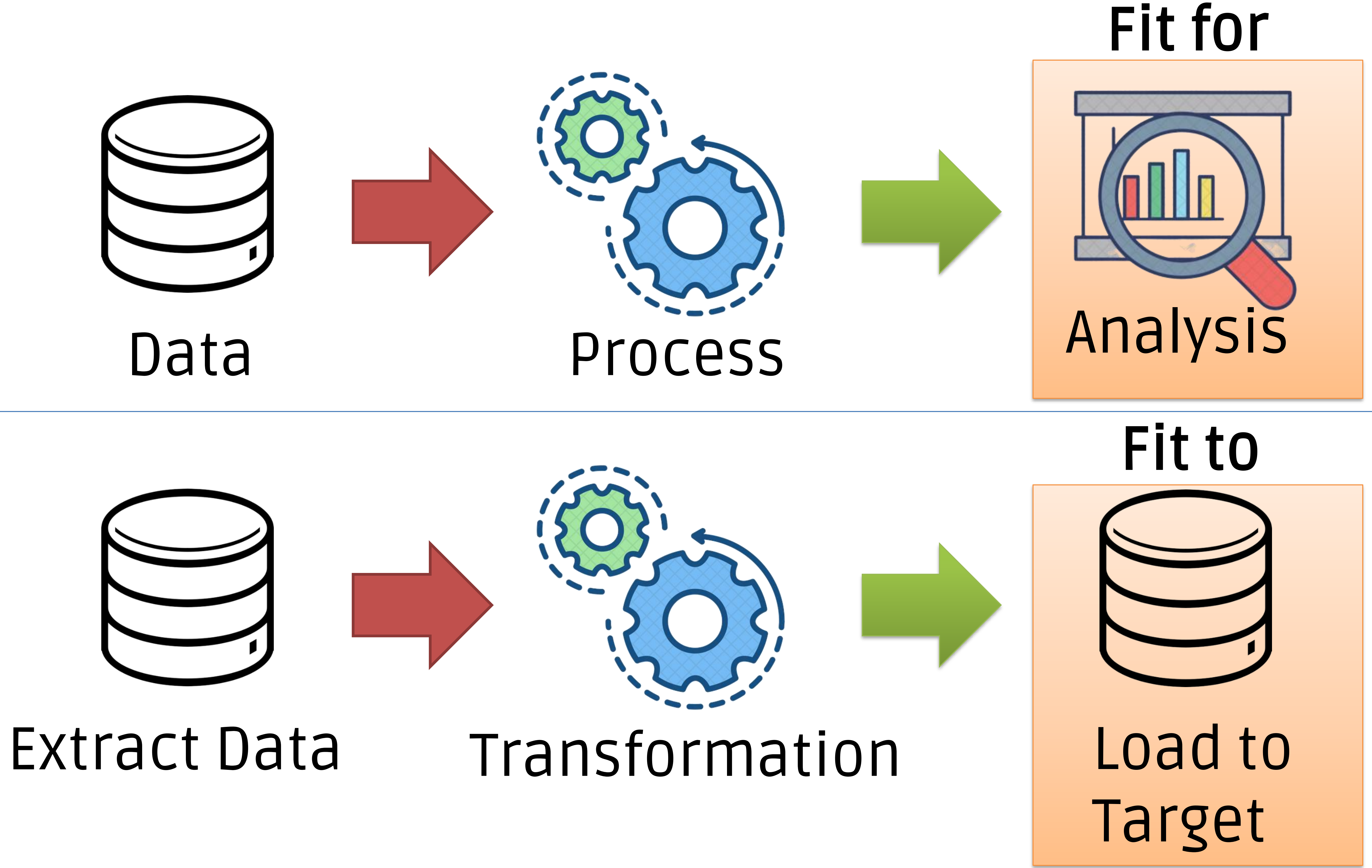
Data Cleaning
dan
Missing Values

What is Data Wrangling

Data Wrangling = Data Munging

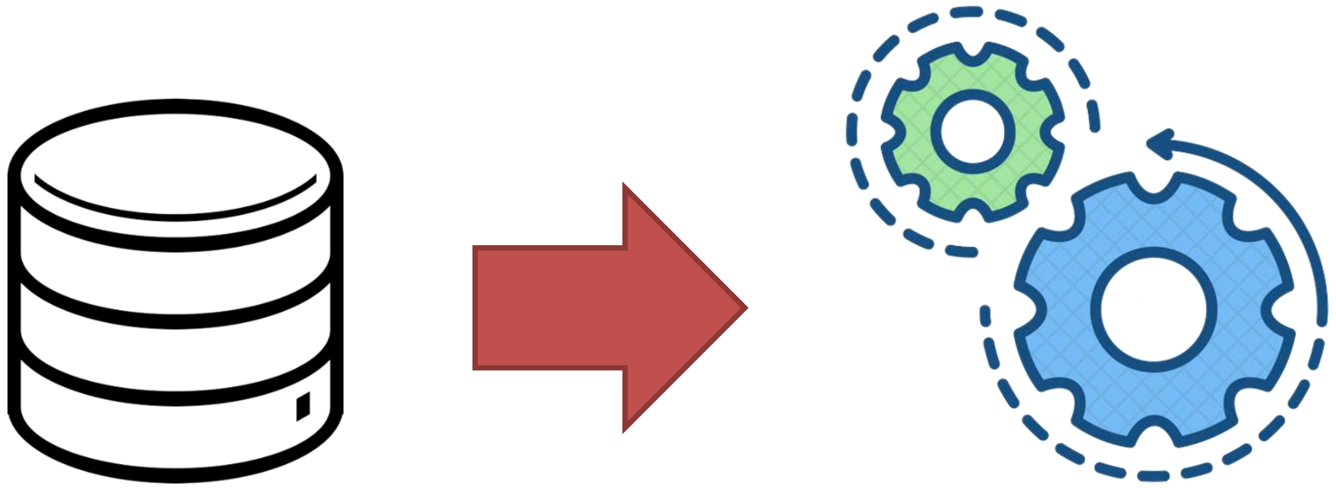


Data Wrangling vs ETL



Data Wrangling: How

SQL



Relational Data

SQL Command / Query

PANDAS

CSV, JSON, SQL, Parquet

pandas

Pandas: an Overview

Series (one-dimensional)

```
1 df['total_bedrooms']
```

```
0        1283.0
1        1901.0
2         174.0
3         337.0
4         326.0
          ...
16995     394.0
16996     528.0
16997     531.0
16998     552.0
16999     300.0
Name: total_bedrooms, Length: 17000, dtype: float64
```

Dataframe (two-dimensional)

```
1 df[['total_bedrooms','population','median_income']]
```

	total_bedrooms	population	median_income
0	1283.0	1015.0	1.4936
1	1901.0	1129.0	1.8200
2	174.0	333.0	1.6509
3	337.0	515.0	3.1917
4	326.0	624.0	1.9250
...	...	...	...
16995	394.0	907.0	2.3571
16996	528.0	1194.0	2.5179
16997	531.0	1244.0	3.0313
16998	552.0	1298.0	1.9797
16999	300.0	806.0	3.0147

17000 rows × 3 columns

Pandas: Access Data

By Index

Dataframe Name

Row Index

```
df.iloc[0:2, 0:2]
```

Column Index

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
0	-114.31	34.19	15.0	5612.0	1283.0	1015.0	472.0	1.4936	66900.0

1

```
1 df.iloc[0]
```

```
longitude              -114.3100
latitude                 34.1900
housing_median_age       15.0000
total_rooms            5612.0000
total_bedrooms         1283.0000
population             1015.0000
households              472.0000
median_income             1.4936
median_house_value    66900.0000
Name: 0, dtype: float64
```

2

```
1 df.iloc[0,0:]
```

```
longitude              -114.3100
latitude                 34.1900
housing_median_age       15.0000
total_rooms            5612.0000
total_bedrooms         1283.0000
population             1015.0000
households              472.0000
median_income             1.4936
median_house_value    66900.0000
Name: 0, dtype: float64
```

3

```
1 df.iloc[0:1,0:1]
```

	longitude
0	-114.31

4

```
1 df.iloc[0,0]
```

```
-114.31
```

Pandas: Access Data

By Label

Row Index/Label

Dataframe Name

df.loc['a':'c', 'x':'z']

Column Label

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value
0	-114.31	34.19	15.0	5612.0	1283.0	1015.0	472.0	1.4936	66900.0

1

```
1 df.loc[0]
```

```
longitude              -114.3100
latitude                 34.1900
housing_median_age       15.0000
total_rooms            5612.0000
total_bedrooms         1283.0000
population             1015.0000
households              472.0000
median_income             1.4936
median_house_value    66900.0000
Name: 0, dtype: float64
```

2

```
1 df.loc[0, 'longitude':]
```

```
longitude              -114.3100
latitude                 34.1900
housing_median_age       15.0000
total_rooms            5612.0000
total_bedrooms         1283.0000
population             1015.0000
households              472.0000
median_income             1.4936
median_house_value    66900.0000
Name: 0, dtype: float64
```

3

```
1 df.loc[0:1, 'longitude':'latitude']
```

	longitude	latitude
0	-114.31	34.19
1	-114.47	34.40

Pandas: Access Data

Filtering

```
df[df['latitude'] > 35]
```

	longitude	latitude
119	-115.93	35.55
157	-116.22	36.00
264	-116.57	35.43
568	-117.02	36.40
1863	-117.28	35.13

```
df[(df['latitude'] > 35) & df['housing_median_age'].isin([18,19])]
```

	longitude	latitude	housing_median_age	total_rooms	total_bed
119	-115.93	35.55	18.0	1321.0	
568	-117.02	36.40	19.0	619.0	
2638	-117.67	35.65	18.0	2737.0	
2745	-117.70	35.62	18.0	2657.0	
3054	-117.81	35.65	19.0	1124.0	

Data Wrangling Steps

1 Discovery	2 Structuring	3 Cleaning
4 Enriching	5 Verifying	6 Publishing

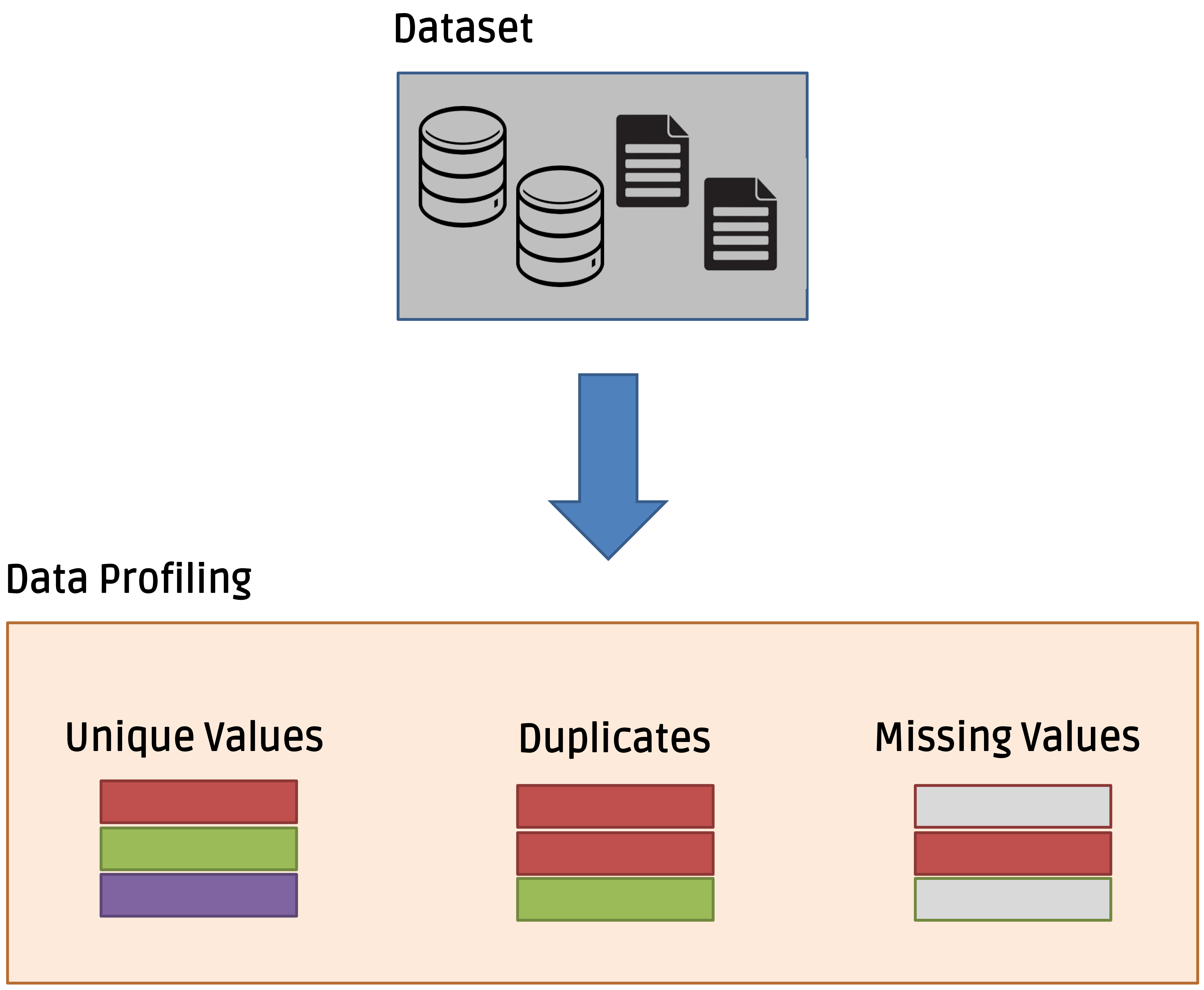
Reference: Harvard Business School Business Insights – [Data Wrangling: What it is and Why It's Important]

Data Wrangling Steps

1 Discovery	2 Structuring	3 Cleaning
4 Enriching	5 Verifying	6 Publishing

Reference: Harvard Business School Business Insights – [Data Wrangling: What it is and Why It's Important]

1. Discovery


Dataset
Data Profiling
Unique Values
Duplicates
Missing Values

Data Wrangling Steps

1 Discovery	2 Structuring	3 Cleaning
4 Enriching	5 Verifying	6 Publishing

2. Structuring

- **Transforming the raw data to be more readily leveraged**

- Operations Involved:
  - Handling Dates
  - Encode Categorical Attributes
    - One-Hot Encoding (OHE)
    - Label Encoding

2. Structuring

One-Hot Encoding

Original categorical column

Origin
USA
Japan
Europe
USA
Europe

One-Hot encoded columns

Origin_USA	Origin_Japan	Origin_Europe
1	0	0
0	1	0
0	0	1
1	0	0
0	0	1

Source: [4 Categorical Encoding Concepts to Know for Data Scientists by Cornellius Yudha Wijaya]

Pros

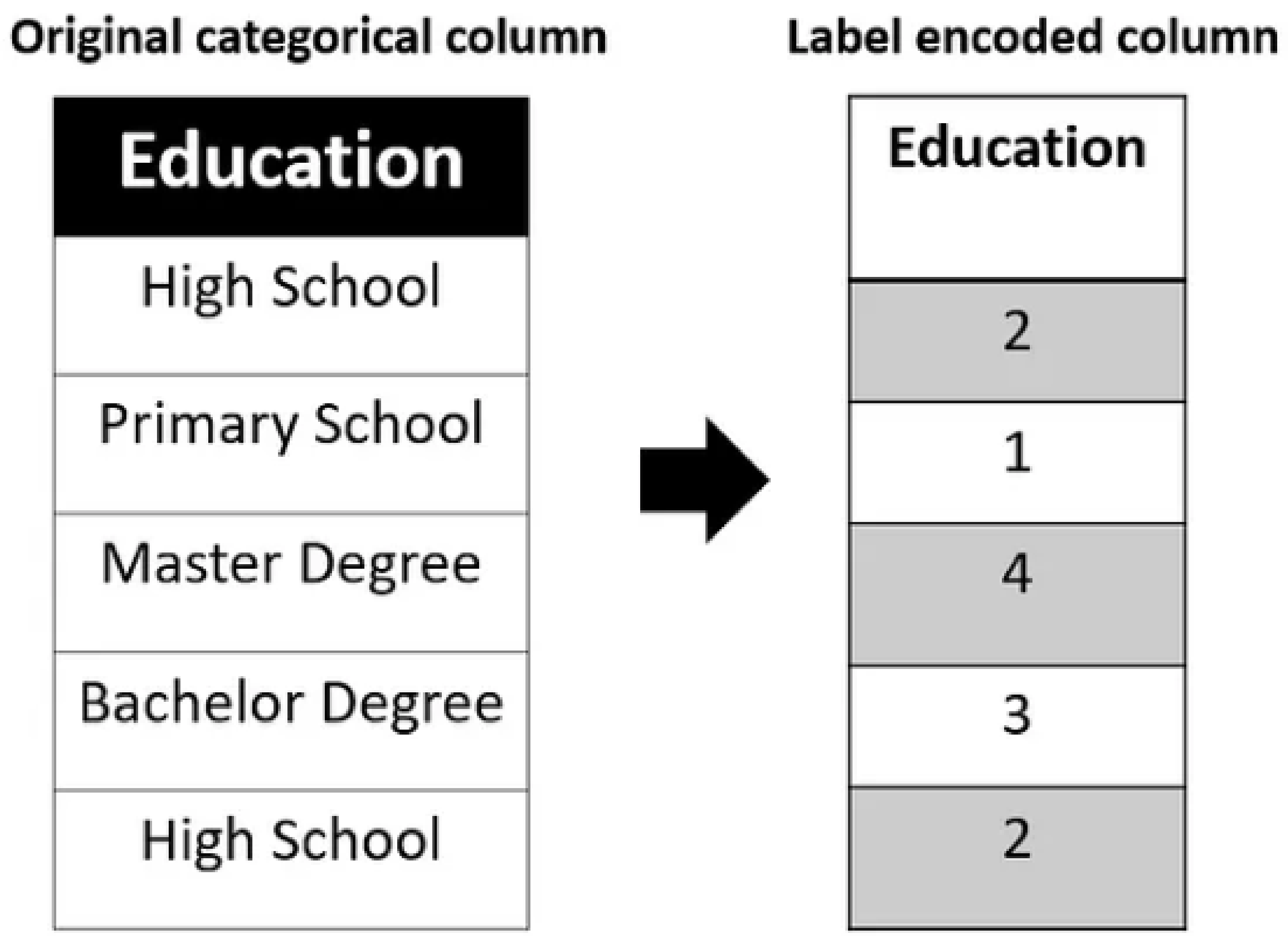
- Suitable for **non-ordinal categories, less cardinality**

Cons

- For high-cardinal categories, leading to the **Curse of Dimensionality**

2. Structuring

Label Encoding



Source: [4 Categorical Encoding Concepts to Know for Data Scientists by Cornellius Yudha Wijaya]

Pros

- Suitable for **ordinal categories**
- Produces only one encoded column

Cons

- Not suitable for **non-ordinal categories**

Data Wrangling Steps

1
Discovery

2
Structuring

3
Cleaning

4
Enriching

5
Verifying

6
Publishing

3. Cleaning

- Process of removing inherent errors in the data that distort the analysis
- Operations Involved:
  - Handling Duplicates
  - Handling Missing values
  - Aggregation/Grouping
  - Attributes Enrichment (New Measures)
  - Attribute Standardisation (with or without regex)
    - Text → Uppercase, Lowercase, Capitalisation, Other Patterns

3. Cleaning

Handling Missing Values

Types of Missing Data (Rubin, 1976)

- MCAR (Missing Completely at Random)
- MAR (Missing at Random)
- MNAR (Missing Not at Random)

Reference: Analytics Vidhya – Dealing with Missing Values for Data Science Beginners

3. Cleaning

Handling Missing Values - MCAR vs MAR vs MNAR

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

Source: Missing Completely at Random - Iris Eekhout | Missing data

3. Cleaning

Handling Missing Values - MCAR vs MAR vs MNAR

MCAR

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

- No relationship with any values, missing or observed. Hence, **completely random**
- Typically indicated by small number of missing values

MAR

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

- **Somewhat** related to another **observed** attribute

MNAR

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

- There is a relationship within the attribute, involving both missing and observed values
- Typically indicated by much higher number of missing values (*compared to* MAR)

Reference: [How to Identify Missingness Types With Missingno]

3. Cleaning

Handling Missing Values – Important Notes

- **We can never confirm if missing values are MNAR or MAR.**
  - The only available option via statistical testing is to test whether missing values are MCAR or not MCAR
- **Context and Common Sense are extremely important!**
  - We need to know the context of the missing data e.g. the value range, to be able to determine the type of missing data
- **Helpful guides:**
  - The best strategy to handle missing values is **getting new data**
  - If the context (e.g. range) for categorical and continuous attributes is known, **check if the attribute is somewhat related to the others**
  - If no context known for a categorical attribute and there is high number of missing values in that attribute, **delete the attribute**

Strategy vs Type	MCAR	MAR	MNAR
Deletion	Yes	No	No
Imputation	Yes	Yes (Advanced e.g. MICE)	No

Get New Data if Possible

Reference: [Assuming a Missing Data Mechanism]

3. Cleaning

Handling Missing Values – Strategies

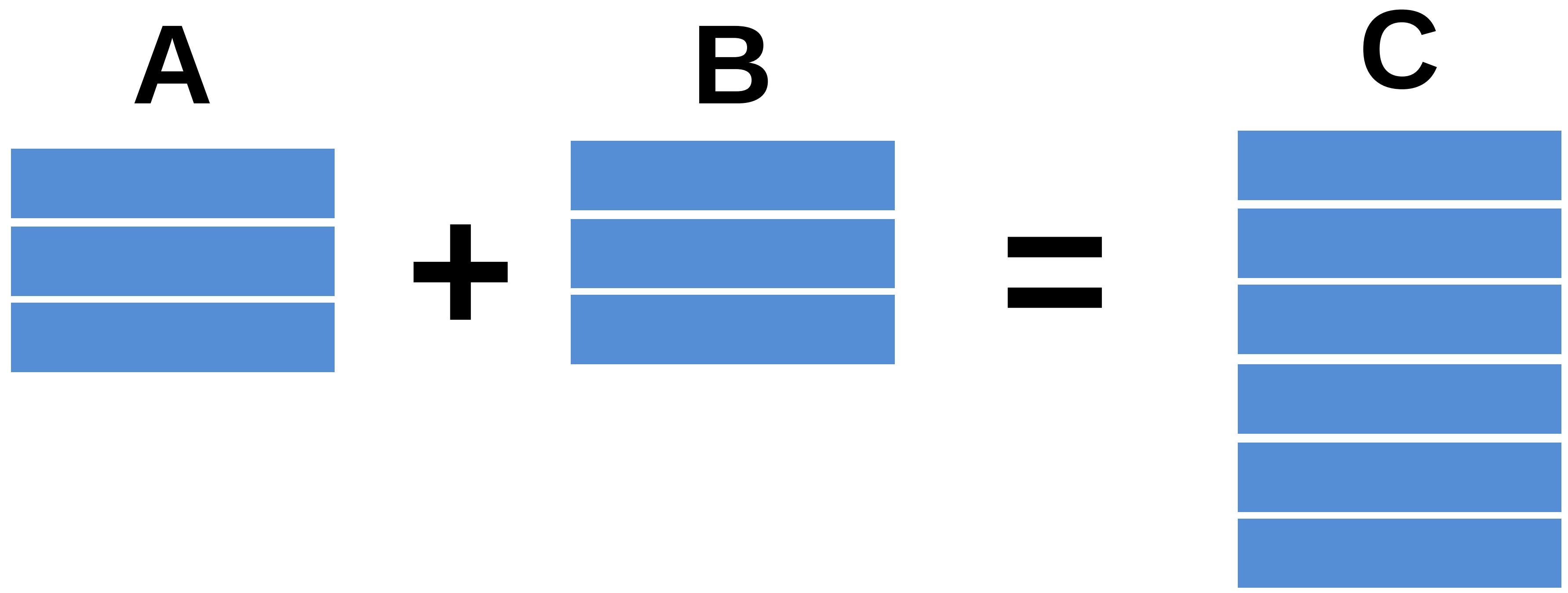
- **Deletion**
  - Rows Deletion (Listwise Deletion)
  - Column Deletion
  - If too many rows/columns containing missing values, deletion leads to information loss or even worse: **not fit for analysis!**
- **Imputation** **(Active Research Field)**
  - For Continuous Attribute(s)
    - Mean, Median, Mode
      - Introduces bias if too many rows are imputed
      - Mean imputation is sensitive to outliers
      - Median imputation assumes MCAR, which is not always the case
  - For Categorical Attribute(s)
    - Mode
    - “Missing” category for missing observations
  - MICE (Multiple Imputation by Chained Equations)
    - Fits predictive model

Data Wrangling Steps

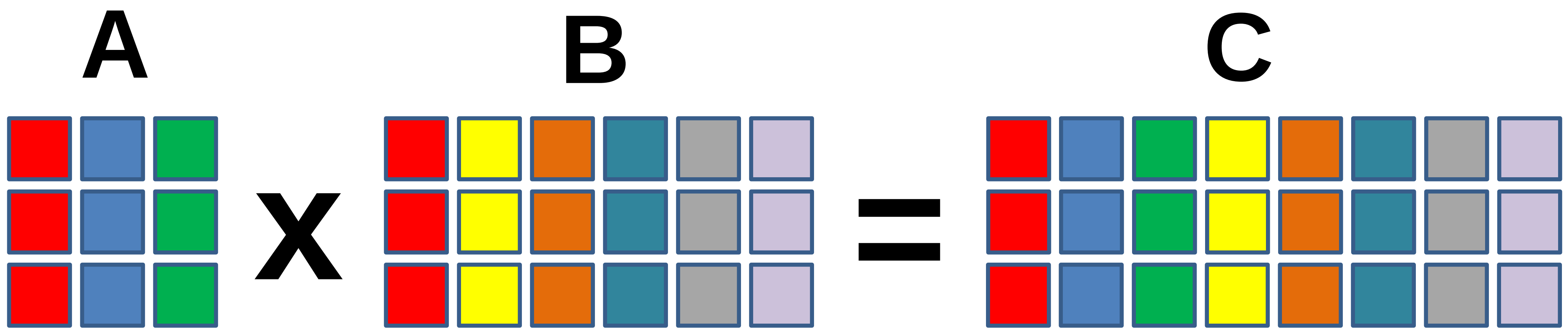
1 Discovery	2 Structuring	3 Cleaning
4 Enriching	5 Verifying	6 Publishing

4. Enriching

- Sometimes, it isn't just enough! Other data may be required
- Depends on the analysis objective(s)
- Operation involves:
  - Concatenation (Adding rows)



  - Merge (combine two different datasets on common keys)



Data Wrangling Steps

1
Discovery

2
Structuring

3
Cleaning

4
Enriching

5
Verifying

6
Publishing

5. Verifying

Ensure the final data:

- Satisfy business rules/common sense
  - The numeric representation of **month** should not exceed 12
  - No negative values for income attribute

- Consistent Formatting
  - The values of currency attributes need to be of numeric types
  - Date format of **yyyy-mm-dd**

In short: **fit for analysis**

Data Wrangling Steps

1
Discovery

2
Structuring

3
Cleaning

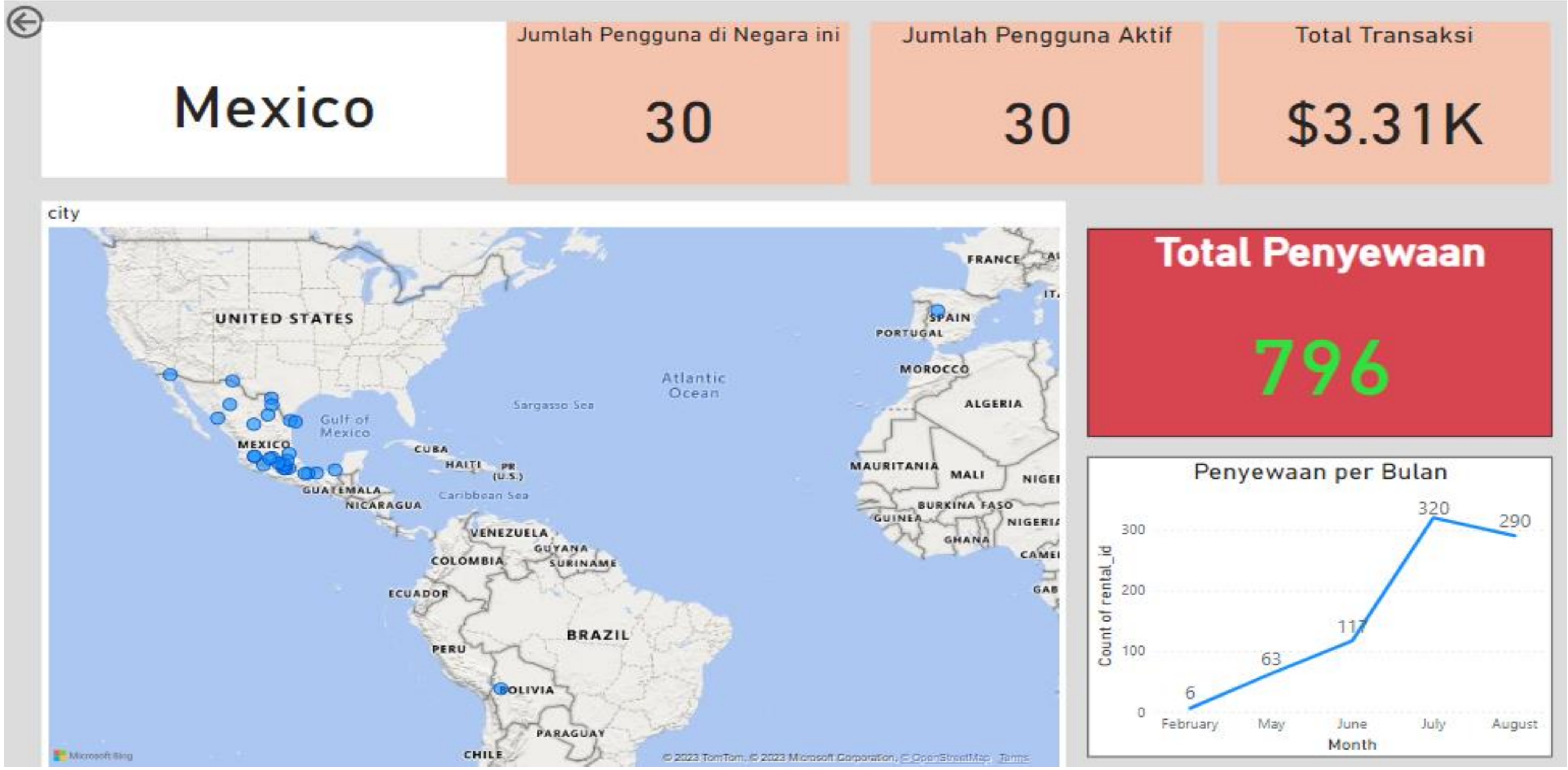
4
Enriching

5
Verifying

6
Publishing

6. Publishing

- Show your data!
- Make it available for others to:
  - Be informed → visualization



- Further analyse → clean data

Pop Quiz

Do those steps need to be in order?

That’s it

Now open your

Data Wrangling and Transformation

Cheers!

Riki Akbar

Ibrahim Saleh Siregar