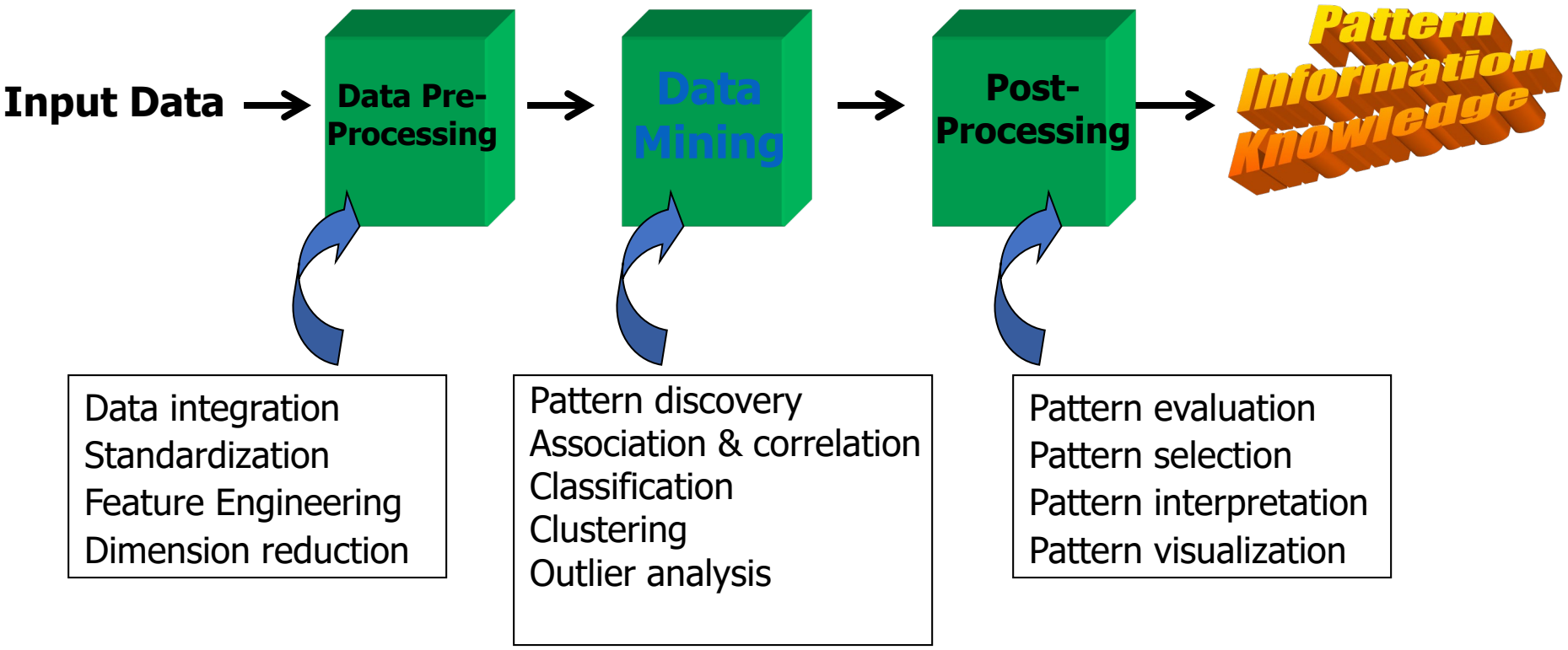
# Data Engineering

MG-GY 8441

# Processing Data

- Agenda
  - Data Cleaning
  - Data Transformation
  - Data Integration
  - Data Reduction
- References
  - Han, Kamber, Pei, *Data Mining: Concepts and Techniques*
    - Chapter 3.1 – 3.5

# Knowledge Discovery from Data



# Example

**Example from Human Resources**

Suppose you are a business analyst within the human resources division of your company. Your group has an annual survey employees. The goals of the survey include

- Better understand the background of new employees
- Gauge satisfaction among returning employees
- Solicit feedback to identify issues or areas for growth
  - Encourage communication between employees and management
- Chart professional development of employee
  - Encourage internal mobility within the company to improve retention

# Example

| | id | cs_python | cs_java | cs_c | ... | len_answer | season | experience_coded | experience |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 1 | 0 | 1 | ... | 74 | Summer | 1 | None, I just finished my undergrad! |
| **1** | 2 | 1 | 1 | 0 | ... | 597 | Spring | 1 | None, I just finished my undergrad! |
| **2** | 3 | 0 | 0 | 0 | ... | 548 | Fall | 4 | 5+ years, I'm a veteran! |
| **3** | 4 | 0 | 0 | 1 | ... | 954 | Spring | 4 | 5+ years, I'm a veteran! |
| **4** | 5 | 0 | 0 | 1 | ... | 612 | Fall | 3 | 2-5 years, I'm getting good at what I do! |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **153** | 154 | 1 | 1 | 0 | ... | 265 | Spring | 4 | 5+ years, I'm a veteran! |
| **154** | 155 | 1 | 1 | 0 | ... | 402 | Summer | 1 | None, I just finished my undergrad! |
| **155** | 156 | 1 | 1 | 0 | ... | 555 | Fall | 2 | < 2 years, I'm fresh! |
| **156** | 157 | 1 | 0 | 0 | ... | 0 | Summer | 1 | None, I just finished my undergrad! |
| **157** | 158 | 1 | 0 | 0 | ... | 222 | Spring | 2 | < 2 years, I'm fresh! |

# Example

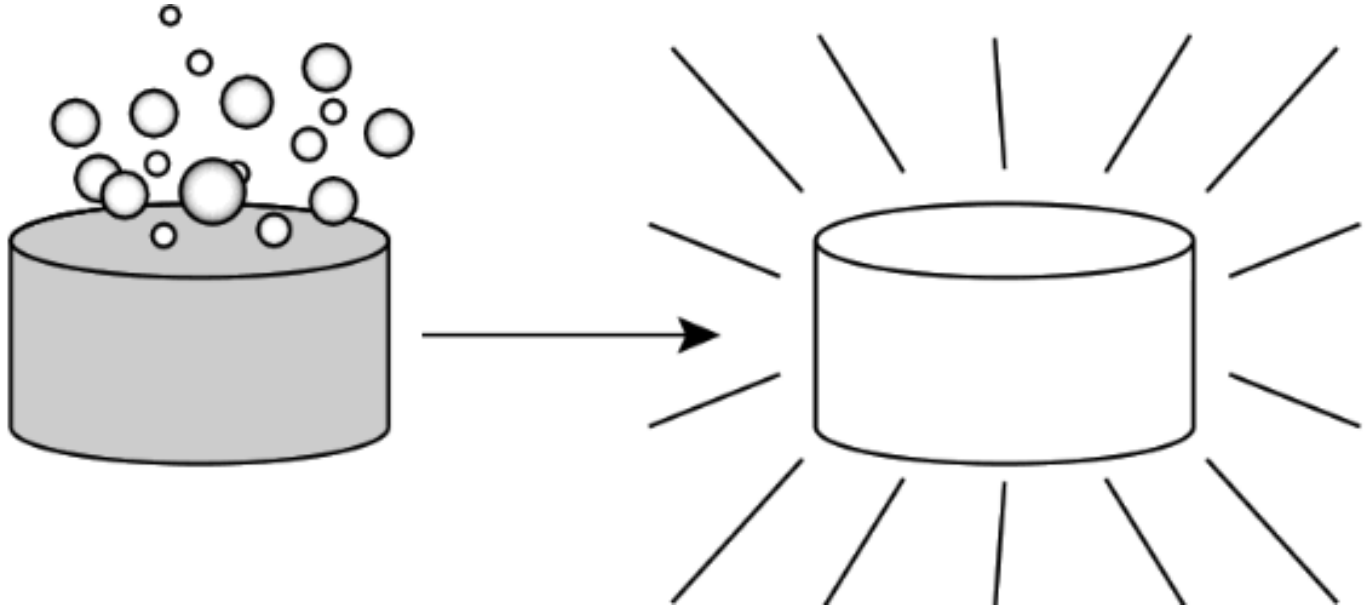
**Example from Human Resources**

You want to analyze the response to report to your group about trend amongst the workforce.

- Does the length of the short answer response correlate with years of professional experience?
- Do employees with experience in math have experience in statistics and vice versa?
- Does the company appear to have more expertise in some programming languages. For example, do we commonly find experience with both Python and R?

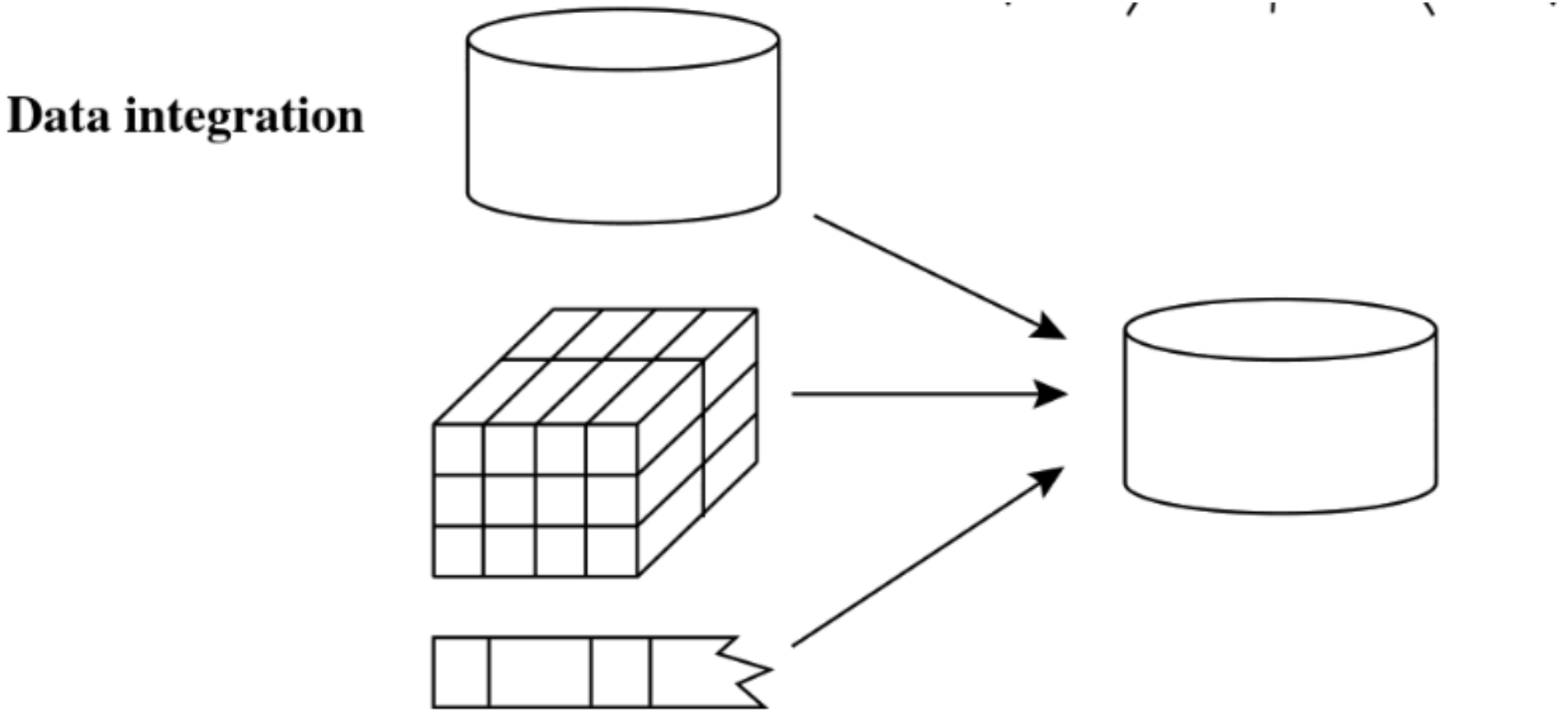
# Example

**Data cleaning**

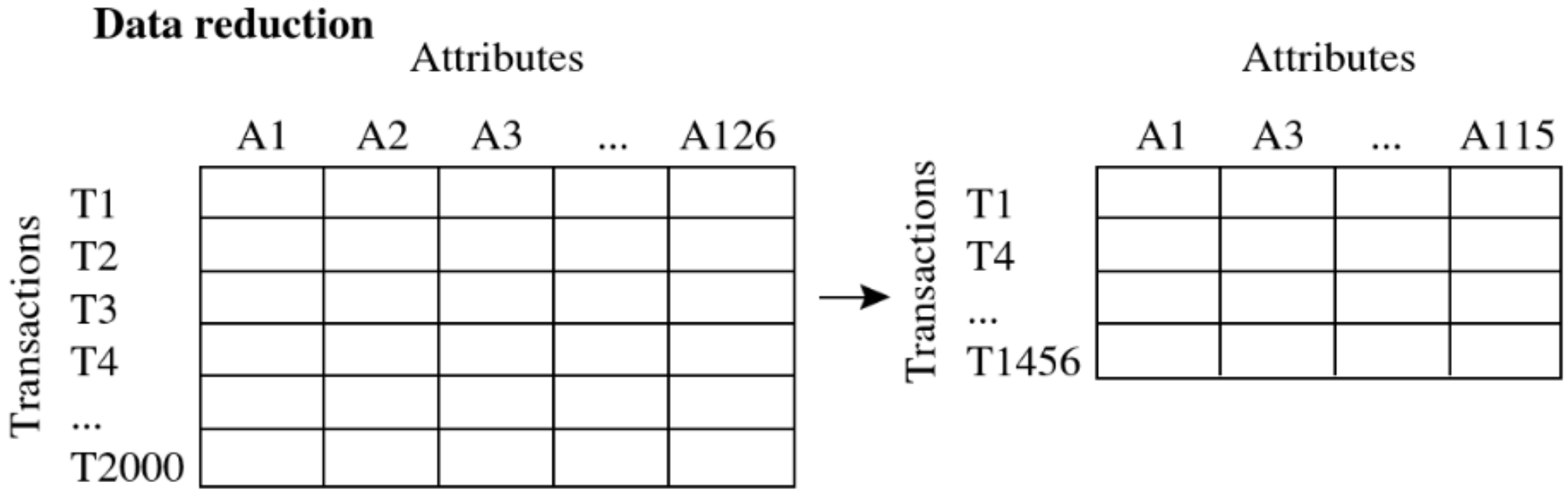


**Data transformation** $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

# Example

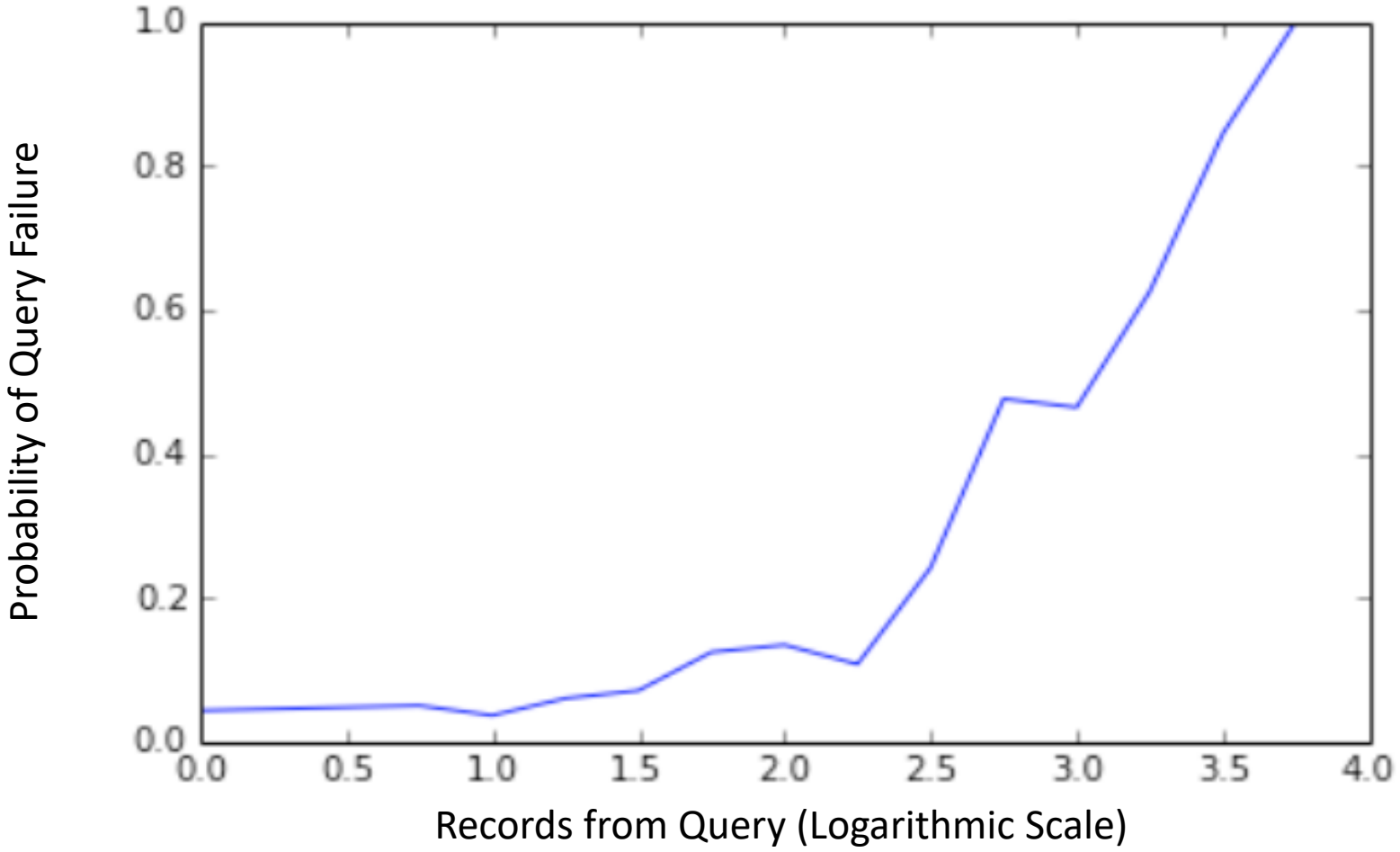


# Example


Data reduction
Attributes
A1
A2
A3
...
A126
Transactions
T1
T2
T3
T4
...
T2000
Attributes
A1
A3
...
A115
Transactions
T1
T4
...
T1456

# Example

Not missing at random



# Data Cleaning

# Accuracy

The accuracy depends on different factors such as the

- provenance of the data
  - source of the data, modifications to the data
- the collection of the numbers
  - system crash, transmission problem, equipment malfunction
- the entry of the records into the system
  - manual entry leads to mistakes
- possibility of falsification

Sometimes we refer to accuracy as the *faithfulness* of the dataset.

# Missing Values

Real world data is messy. Commonly records are missing values for some attributes. Which of the following approaches may be reasonable for dealing with this issue?

1. Drop the observations with missing values
2. Replace missing values with an average value
3. Replace missing values with comparable values from another dataset
4. Replace missing values with random values
5. Replace a missing value with the last present value
6. Ignore the missing values, they won't affect our study anyways

# Missing Values

Real world data is messy. Commonly records are missing values for some attributes. Which of the following approaches may be reasonable for dealing with this issue?

1. Drop the observations with missing values
2. Replace missing values with an average value (mean imputation)
3. Replace missing values with comparable values from another dataset (cold deck imputation)
4. Replace missing values with random values (hot deck imputation)
5. Replace a missing value with the last present value (forward-fill or back-fill)
6. ~~Ignore the missing values, they won't affect our study anyways~~

# Missing Values

Missing values are omissions from the dataset. However, we can represent the omissions in different ways.

## Blank

- Absence of a value have different implications.
- If the data was a census collected annually, then was the value for that year never collected?
- If the data was a survey, then did a respondent refuse to answer the question?

## Special Numbers

- Large numbers like 9999
  - If the dataset contains the age of mothers, then large value should be interpreted as missing value
- Small number like 0
  - 0 for latitude and longitude interpreted as 0°00'00.0"N+0°00'00.0"E island off the coast of Africa
  - 0 date interpreted as 1970-01-01T00:00:00Z

# Missing Values

Missing values are omissions from the dataset. However, we can represent the omissions in different ways.

## Special Characters

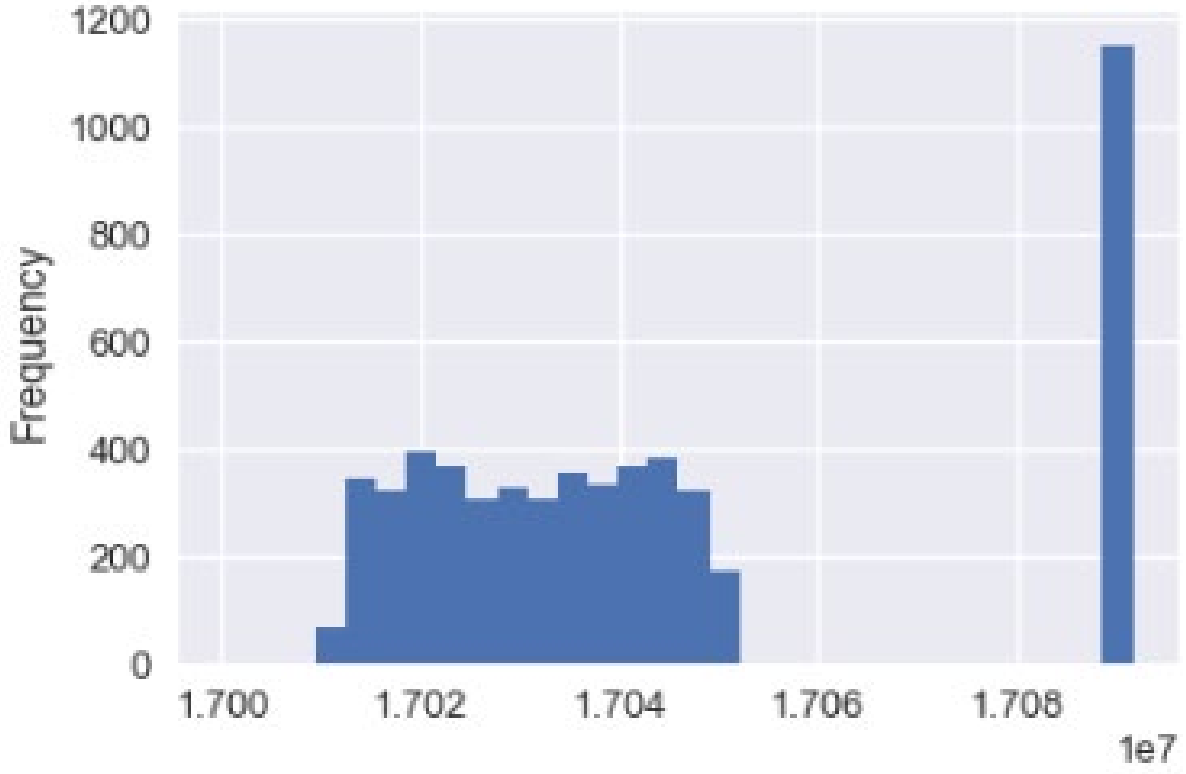
- NULL common for tables in databases
- #NA common for spreadsheets
- NaN common for programming languages
- In Python we use
- numpy.NaN
- to indicate missing values

| | | | | |
|---|---|---|---|---|
| 16 | Jeremy | Male | 9/21/2010 | 5:56 AM |
| 17 | Shawn | Male | 12/7/1986 | 7:45 PM |
| 18 | Diana | Female | 10/23/1981 | 10:27 AM |
| 19 | Donna | Female | 7/22/2010 | 3:48 AM |
| 20 | Lois | NaN | 4/22/1995 | 7:18 PM |
| 21 | Matthew | Male | 9/5/1995 | 2:12 AM |
| 22 | Joshua | NaN | 3/8/2012 | 1:58 AM |
| 23 | NaN | Male | 6/14/2012 | 4:19 PM |
| 24 | John | Male | 7/1/1992 | 10:08 PM |

# Data Transformation

# Consistency

Consistency refers to the coherence of the data. The records in the dataset cannot conflict with each other in terms of content or format.



# Standardization

Standardization refers to transformations of numbers to ranges

- min-max

$$v' = \frac{\mathrm{v} - \min}{\max - \min}$$

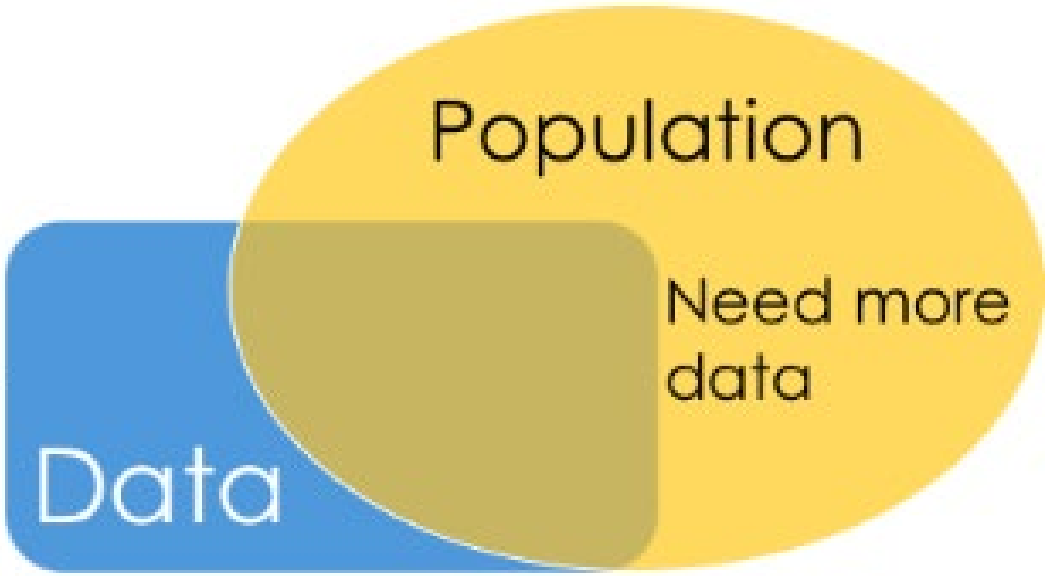
- z-score

$$\mathrm{v}' = \frac{\mathrm{v} - \mathrm{mean}}{\mathrm{standard\_deviation}}$$

- decimal scaling

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \mathrm{Max}(|v'|)<1$$

# Data Integration



# Scope and Temporality

Scope refers to the coverage of the dataset. Note that coverage depends on the problem.

How was the data collected in the survey?

- Was it voluntary? Could the survey have selection bias?

Who collected the data?

- Was the survey anonymous? If the surveyor was managements, then would the employee respond honestly?

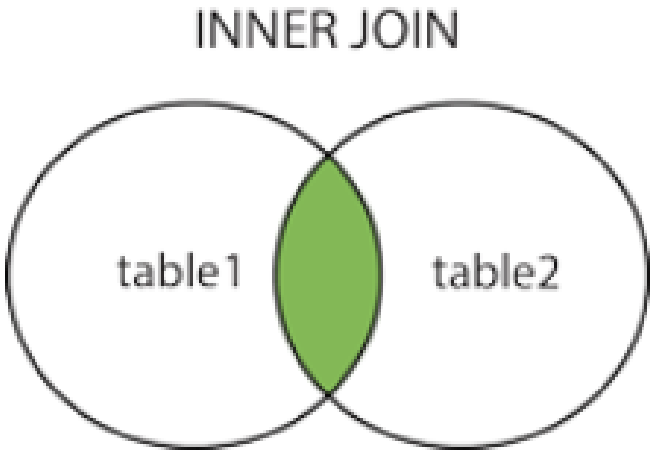
Temporality refers to date and time of the record along with data and time of the record keeping.

How timely is the data in the survey?

- Was the survey recent? Has the company changed significantly since the suvery?

# Joins

- Inner Join pairs each of the rows in the tables depending on the entries in specific columns.
- The entries of columns must match for the pair to appear in the Join.




s

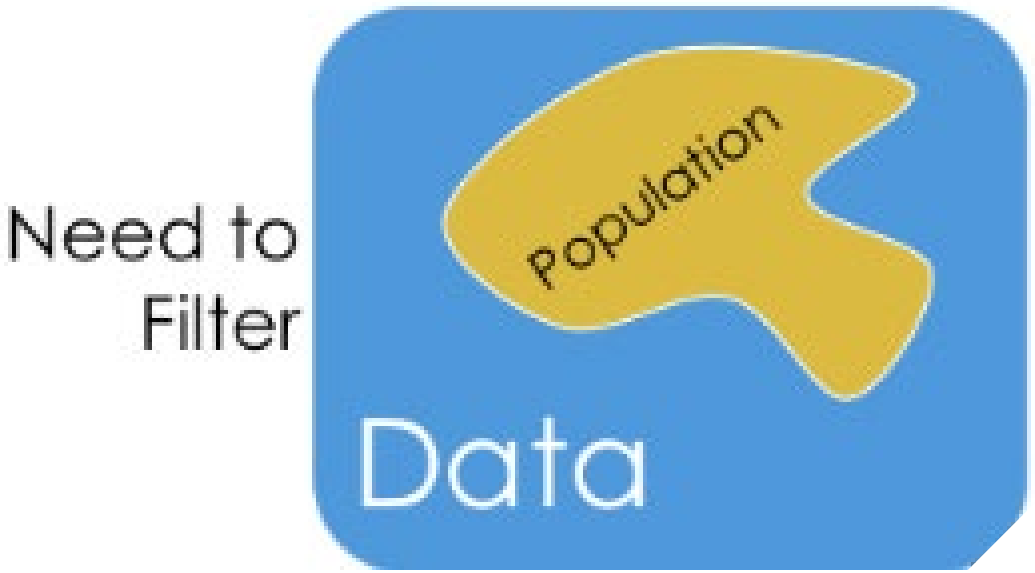
| M | U |
|---|---|
| 1 | W |
| 2 | X |
| 3 | X |
| 4 | Y |

t

| N | V |
|---|---|
| A | X |
| B | X |
| C | Y |
| D | Z |

| M | U | N | V |
|---|---|---|---|
| 2 | X | A | X |
| 3 | X | A | X |
| 2 | X | B | X |
| 3 | X | B | X |
| 4 | Y | C | Y |

# Data Reduction



# Accessibility

Granularity of the data means the information behind the data particularly the amount of detail. We can change the granularity to make the data more accessible.

## Could we group the surveys by employee?

- Comparing year over year would help us to chart growth within the firm

## What is the data was grouped by location?

- How were the number aggregated at each branch of the company?

## How can we find patterns through matching common entries?

- If we grouped the employees by seniority, then would be find patterns in skills? Could we use these patterns to identify employees for promotion?

# Grouping

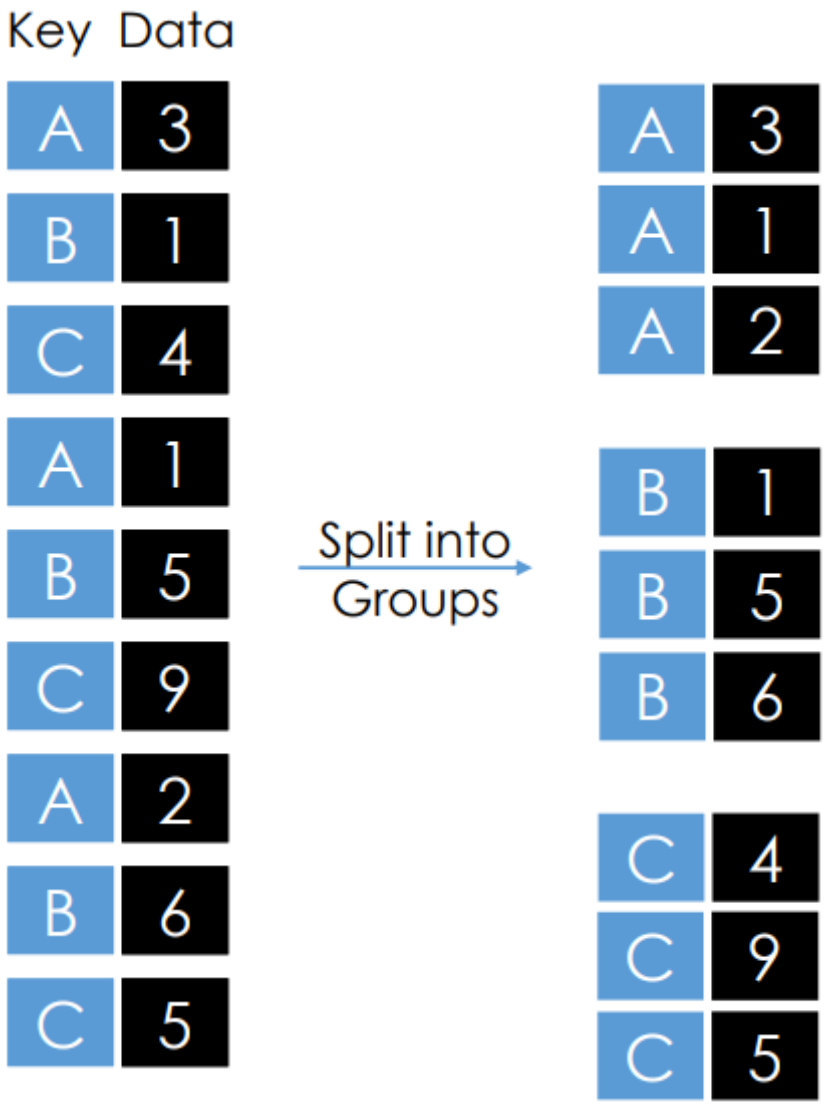
| Key | Data |
|---|---|
| A | 3 |
| B | 1 |
| C | 4 |
| A | 1 |
| B | 5 |
| C | 9 |
| A | 2 |
| B | 6 |
| C | 5 |

# Grouping

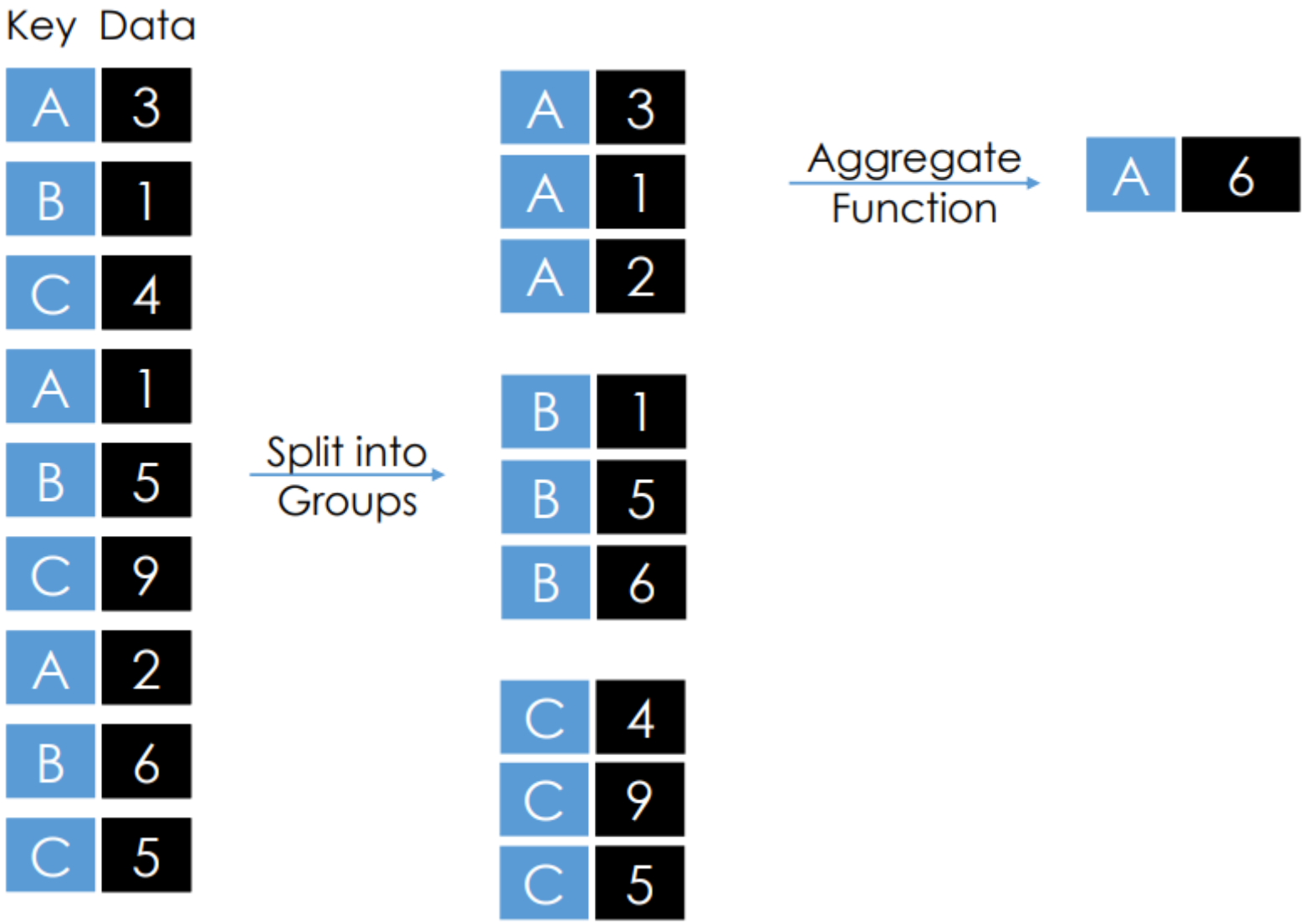
| Key | Data |
|---|---|
| A | 3 |
| B | 1 |
| C | 4 |
| A | 1 |
| B | 5 |
| C | 9 |
| A | 2 |
| B | 6 |
| C | 5 |

| | |
|---|---|
| A | 3 |
| A | 1 |
| A | 2 |

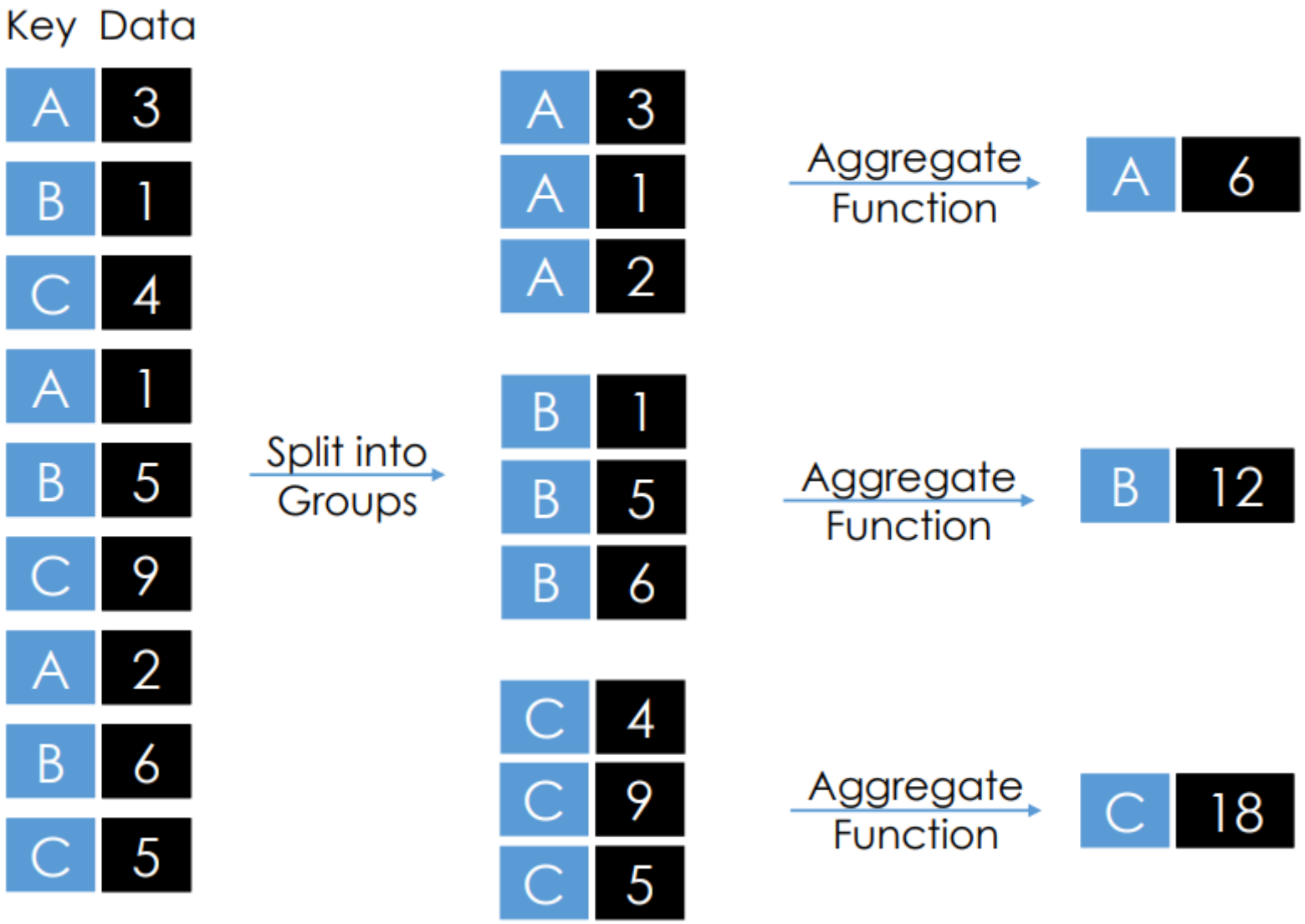
# Grouping


Key Data
A 3
B 1
C 4
A 1
B 5
C 9
A 2
B 6
C 5
Split into Groups
A 3
A 1
A 2
B 1
B 5
B 6
C 4
C 9
C 5

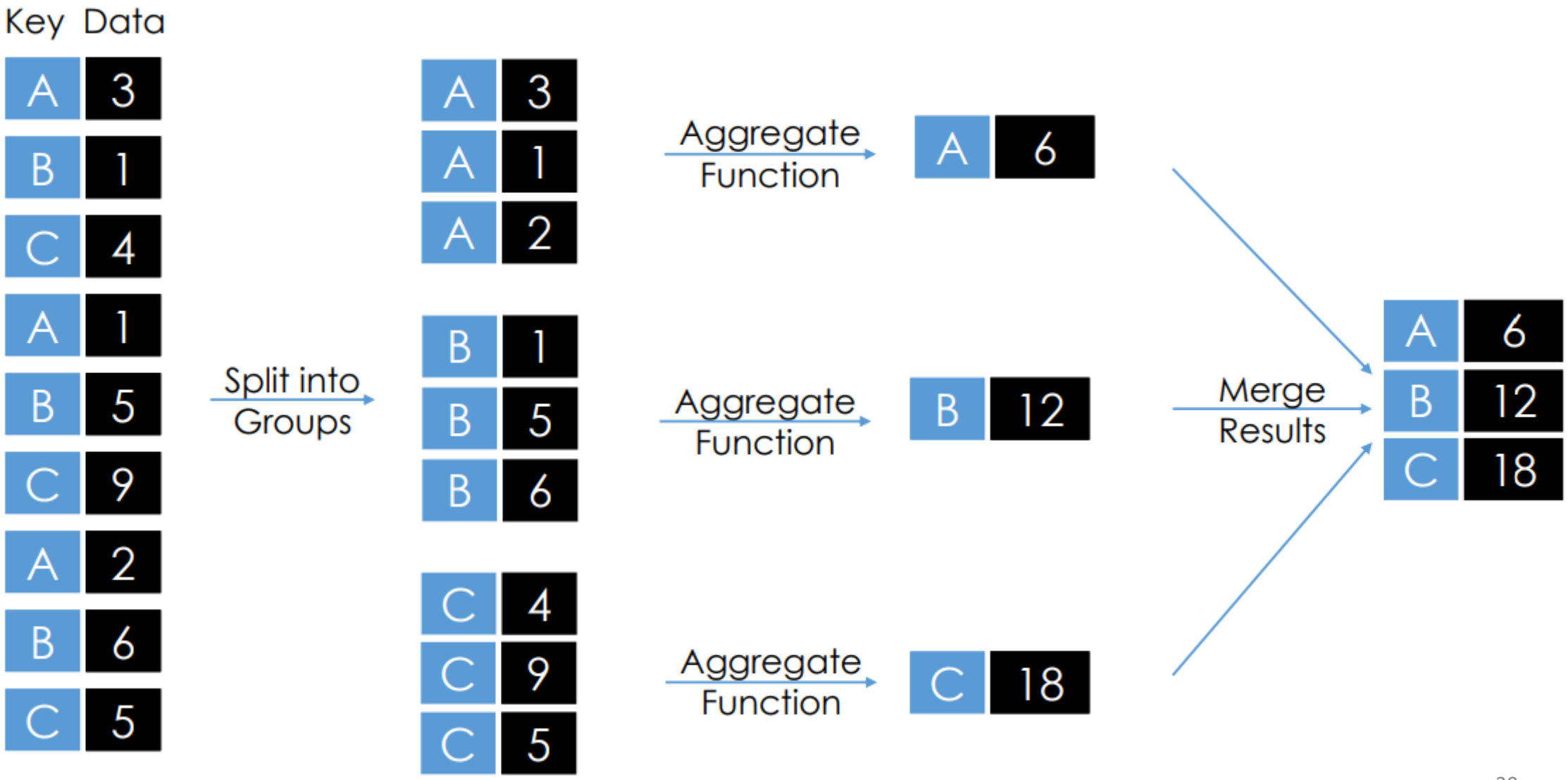
# Grouping


Key Data
A 3
B 1
C 4
A 1
B 5
C 9
A 2
B 6
C 5
Split into Groups
A 3
A 1
A 2
B 1
B 5
B 6
C 4
C 9
C 5
Aggregate Function
A 6

# Grouping


Key Data
A 3
B 1
C 4
A 1
B 5
C 9
A 2
B 6
C 5
Split into Groups
A 3
A 1
A 2
B 1
B 5
B 6
C 4
C 9
C 5
Aggregate Function
A 6
Aggregate Function
B 12
Aggregate Function
C 18

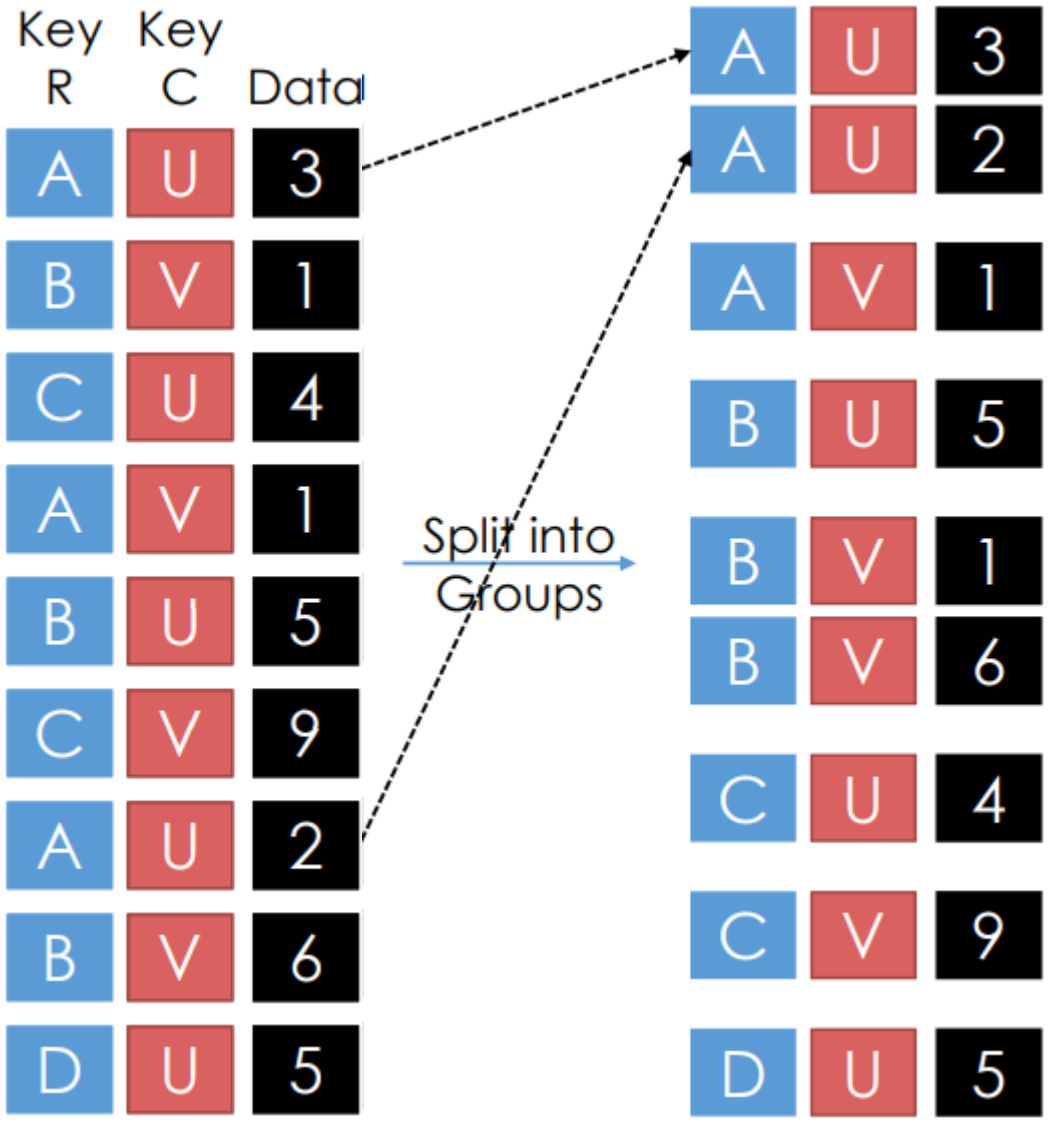
# Grouping


Key Data
A 3
B 1
C 4
A 1
B 5
C 9
A 2
B 6
C 5
Split into Groups
A 3
A 1
A 2
B 1
B 5
B 6
C 4
C 9
C 5
Aggregate Function
A 6
Aggregate Function
B 12
Aggregate Function
C 18
Merge Results
A 6
B 12
C 18

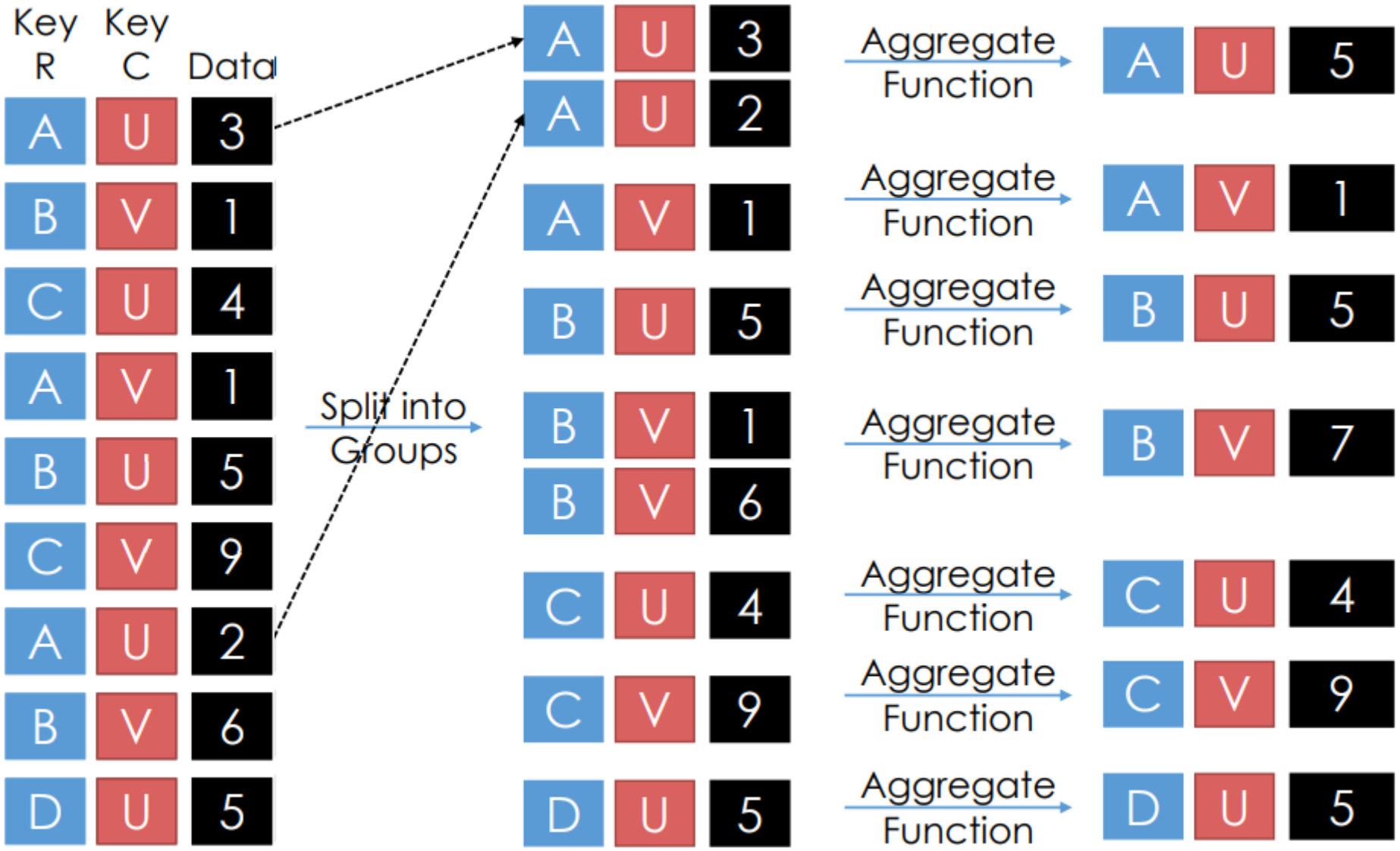
# Pivoting

| Key R | Key C | Data |
|---|---|---|
| A | U | 3 |
| B | V | 1 |
| C | U | 4 |
| A | V | 1 |
| B | U | 5 |
| C | V | 9 |
| A | U | 2 |
| B | V | 6 |
| D | U | 5 |

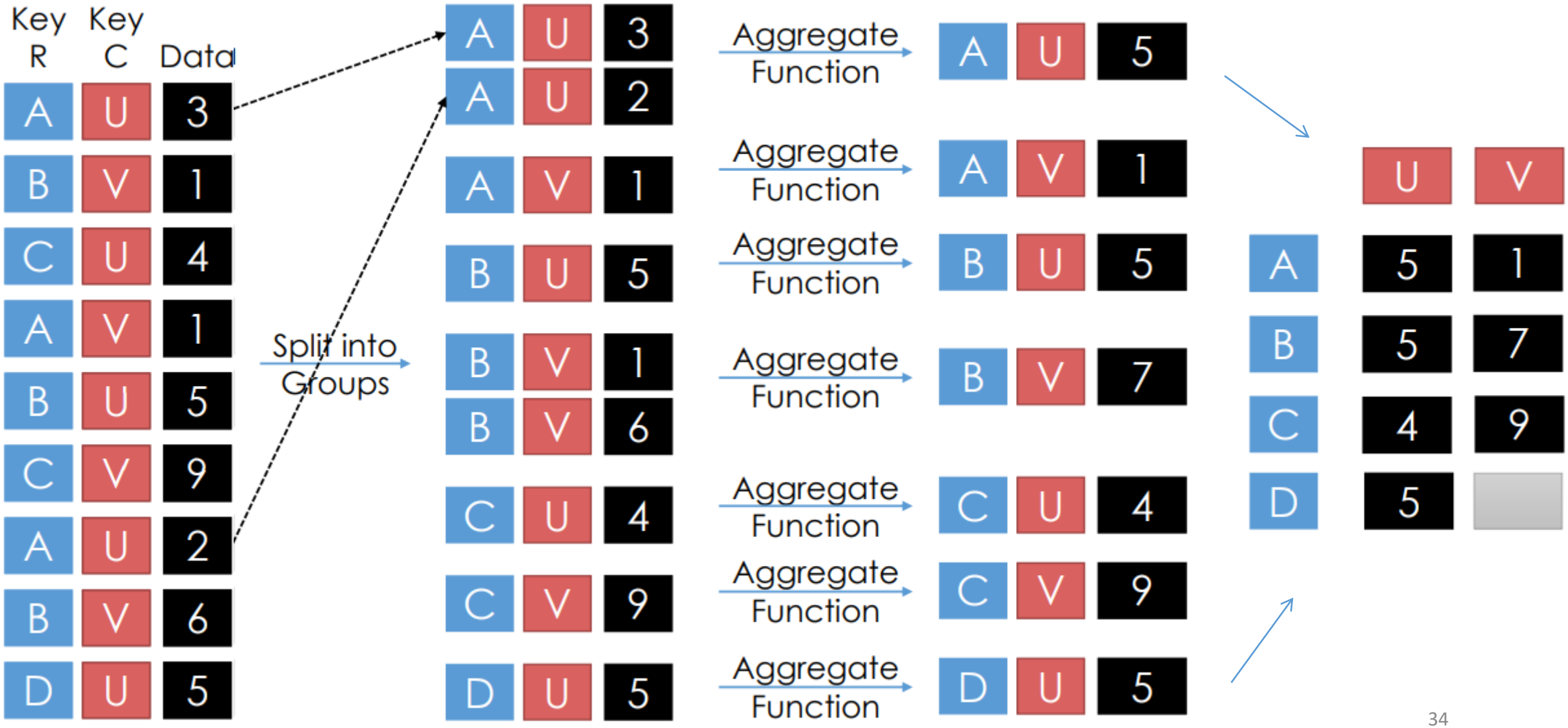
# Pivoting


Key R
Key C
Data
A U 3
B V 1
C U 4
A V 1
B U 5
C V 9
A U 2
B V 6
D U 5
Split into Groups
A U 3
A U 2
A V 1
B U 5
B V 1
B V 6
C U 4
C V 9
D U 5

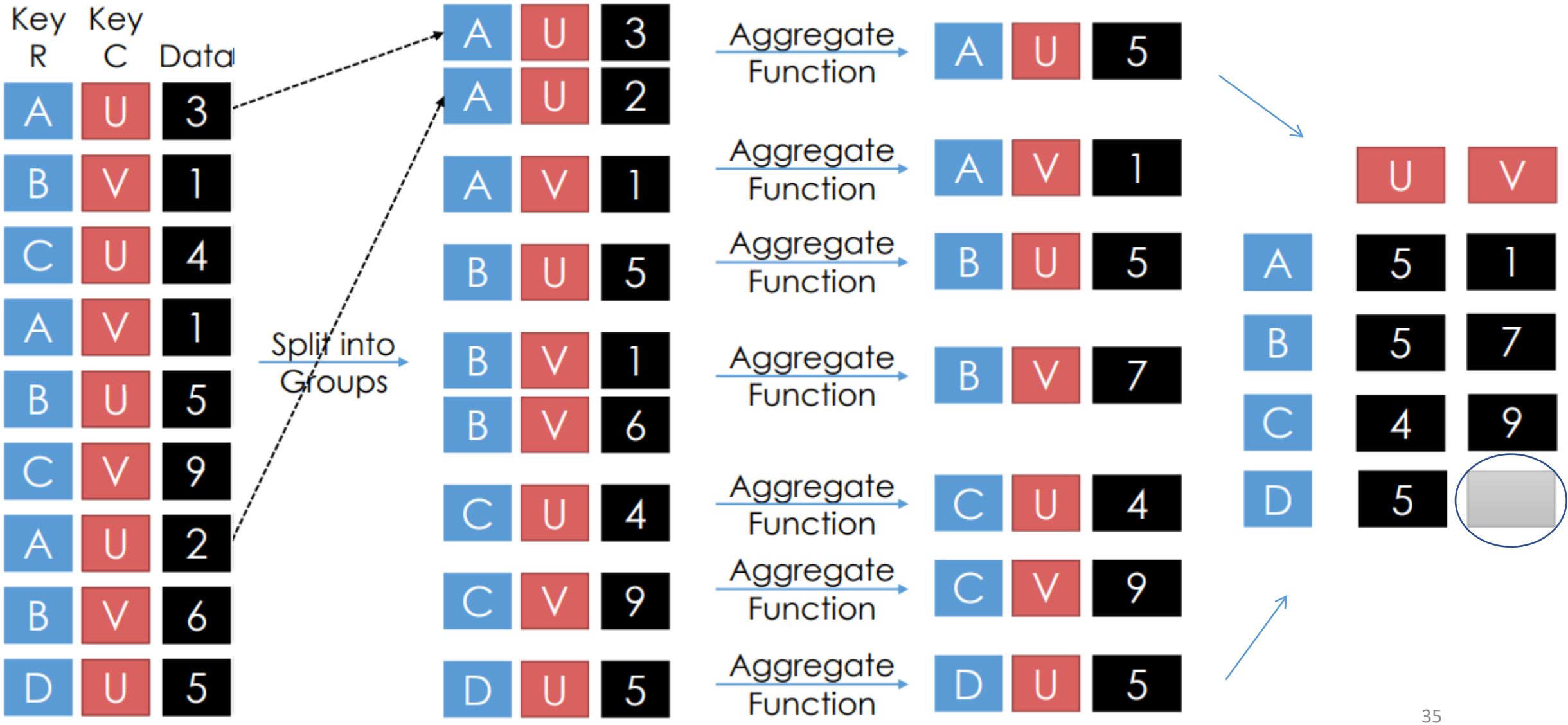
# Pivoting


Key R
Key C
Data
A U 3
B V 1
C U 4
A V 1
B U 5
C V 9
A U 2
B V 6
D U 5
Split into Groups
A U 3
A U 2
A V 1
B U 5
B V 1
B V 6
C U 4
C V 9
D U 5
Aggregate Function
A U 5
Aggregate Function
A V 1
Aggregate Function
B U 5
Aggregate Function
B V 7
Aggregate Function
C U 4
Aggregate Function
C V 9
Aggregate Function
D U 5

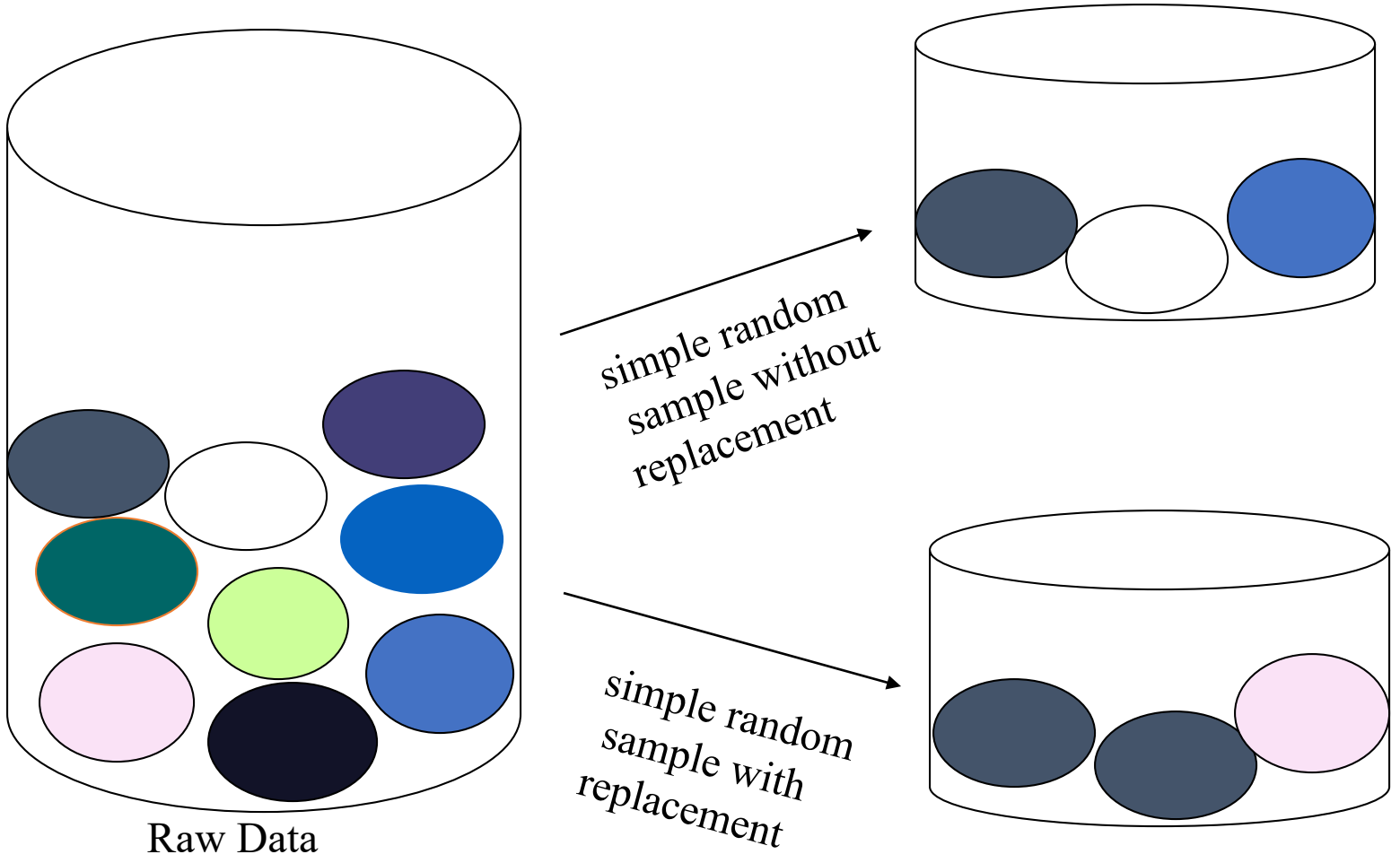
# Pivoting


Key R
Key C
Data
A U 3
B V 1
C U 4
A V 1
B U 5
C V 9
A U 2
B V 6
D U 5
Split into Groups
A U 3
A U 2
A V 1
B U 5
B V 1
B V 6
C U 4
C V 9
D U 5
Aggregate Function
A U 5
Aggregate Function
A V 1
Aggregate Function
B U 5
Aggregate Function
B V 7
Aggregate Function
C U 4
Aggregate Function
C V 9
Aggregate Function
D U 5
U V
A 5 1
B 5 7
C 4 9
D 5

# Pivoting


Key R
Key C
Data
A U 3
B V 1
C U 4
A V 1
B U 5
C V 9
A U 2
B V 6
D U 5
Split into Groups
A U 3
A U 2
A V 1
B U 5
B V 1
B V 6
C U 4
C V 9
D U 5
Aggregate Function
A U 5
Aggregate Function
A V 1
Aggregate Function
B U 5
Aggregate Function
B V 7
Aggregate Function
C U 4
Aggregate Function
C V 9
Aggregate Function
D U 5
U V
A 5 1
B 5 7
C 4 9
D 5

# Sampling


simple random
sample without
replacement
simple random
sample with
replacement
Raw Data

# Summary

- Data cleaning
  - Treat missing values
  - Handle spurious or corrupt values
- Data transformation
  - Resolve inconsistencies
  - Standardize
- Data integration
  - Assess scope and temporality
  - Link records
- Data reduction
  - Group
  - Aggregate