Data Preparation

Harvesting, Storing, and Retrieving Data

Data Prep

Data Preparation

- Deleting Unnecessary Data
- Consolidating/Separating Fields
- Changing Formats
- Transforming Data
- Making Corrections

Quick Review

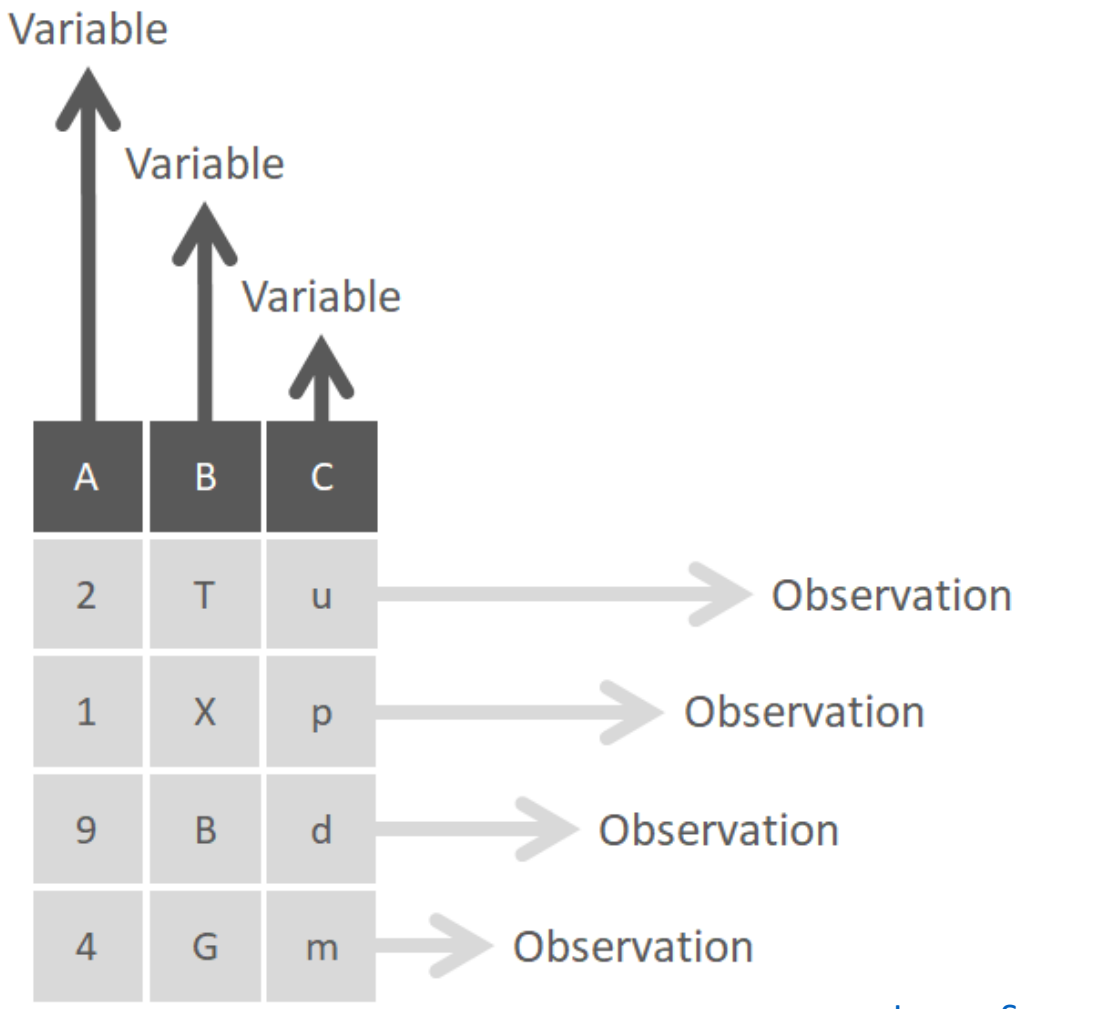


Image Source

Main Tasks for a Single Dataset

Sort/Arrange

Order rows by value or characters

Filter

When you only want certain rows/columns

Transform

Creation on new record fields from existing data

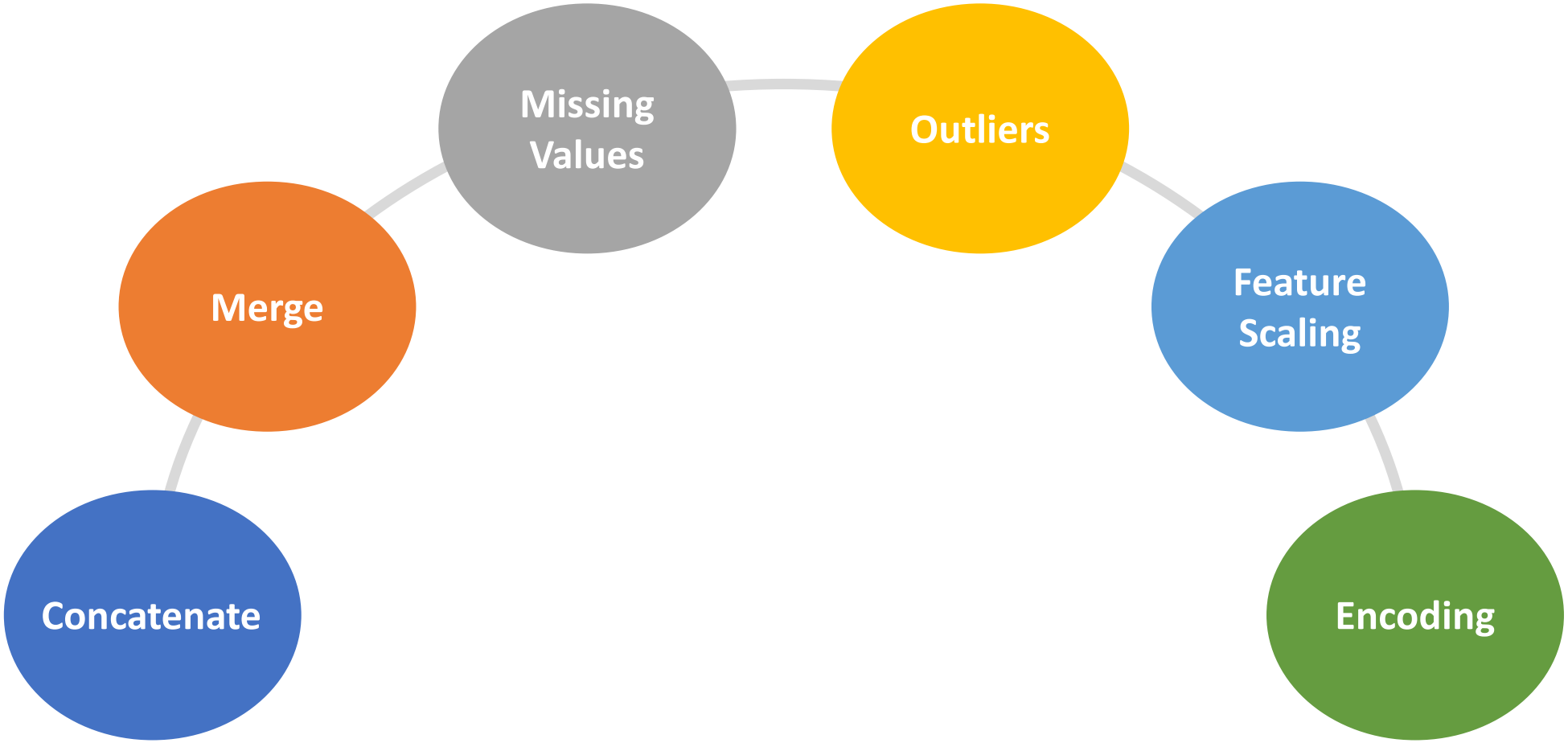
Select

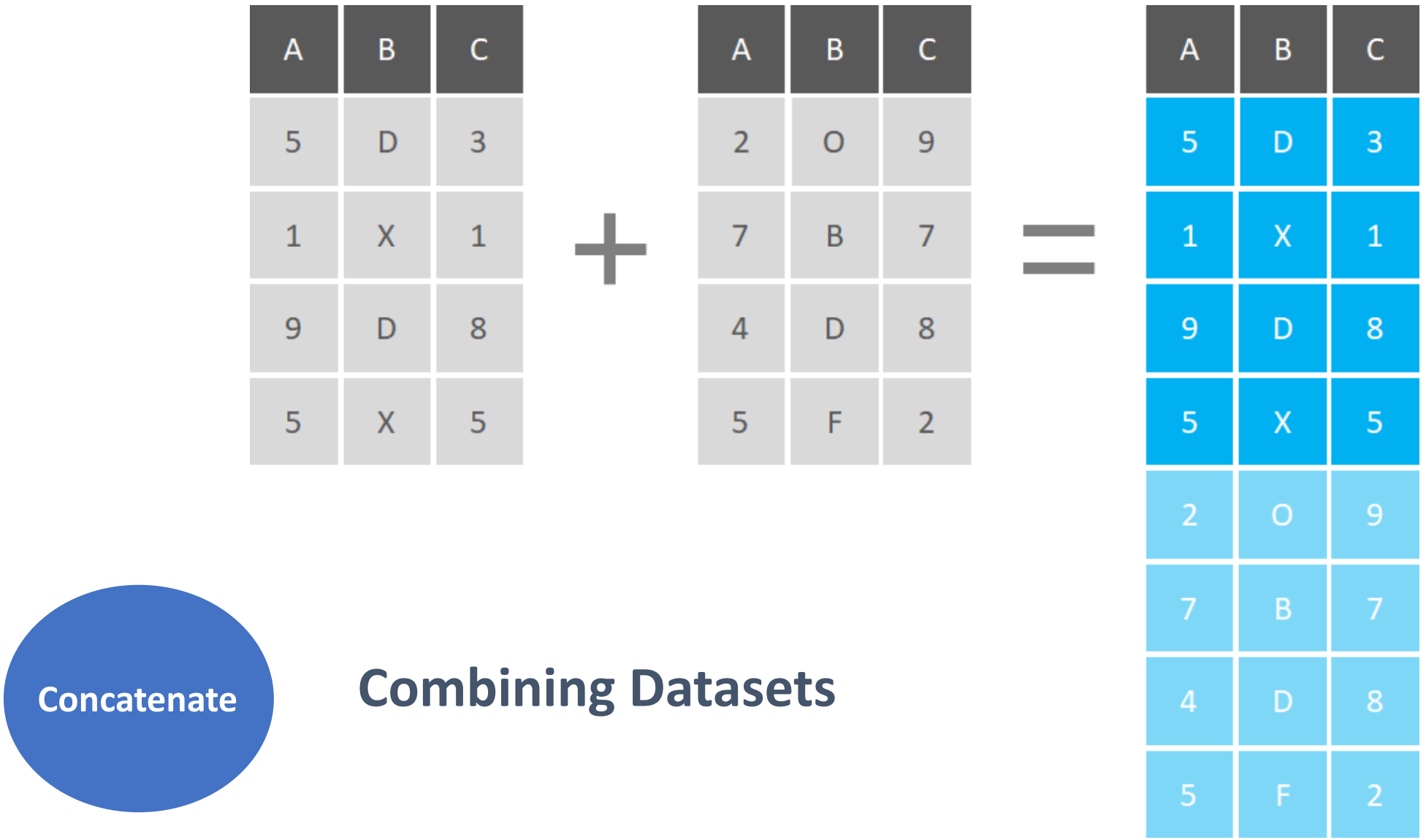
Choosing columns based on a defined criteria

Aggregate

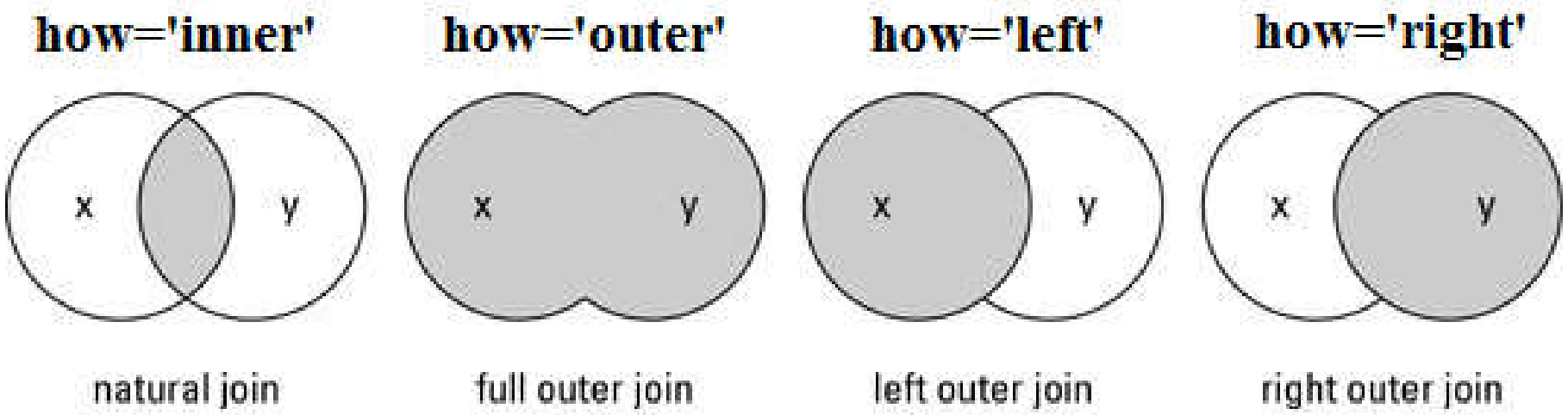
Gathering and/or summarizing information

Combining Datasets



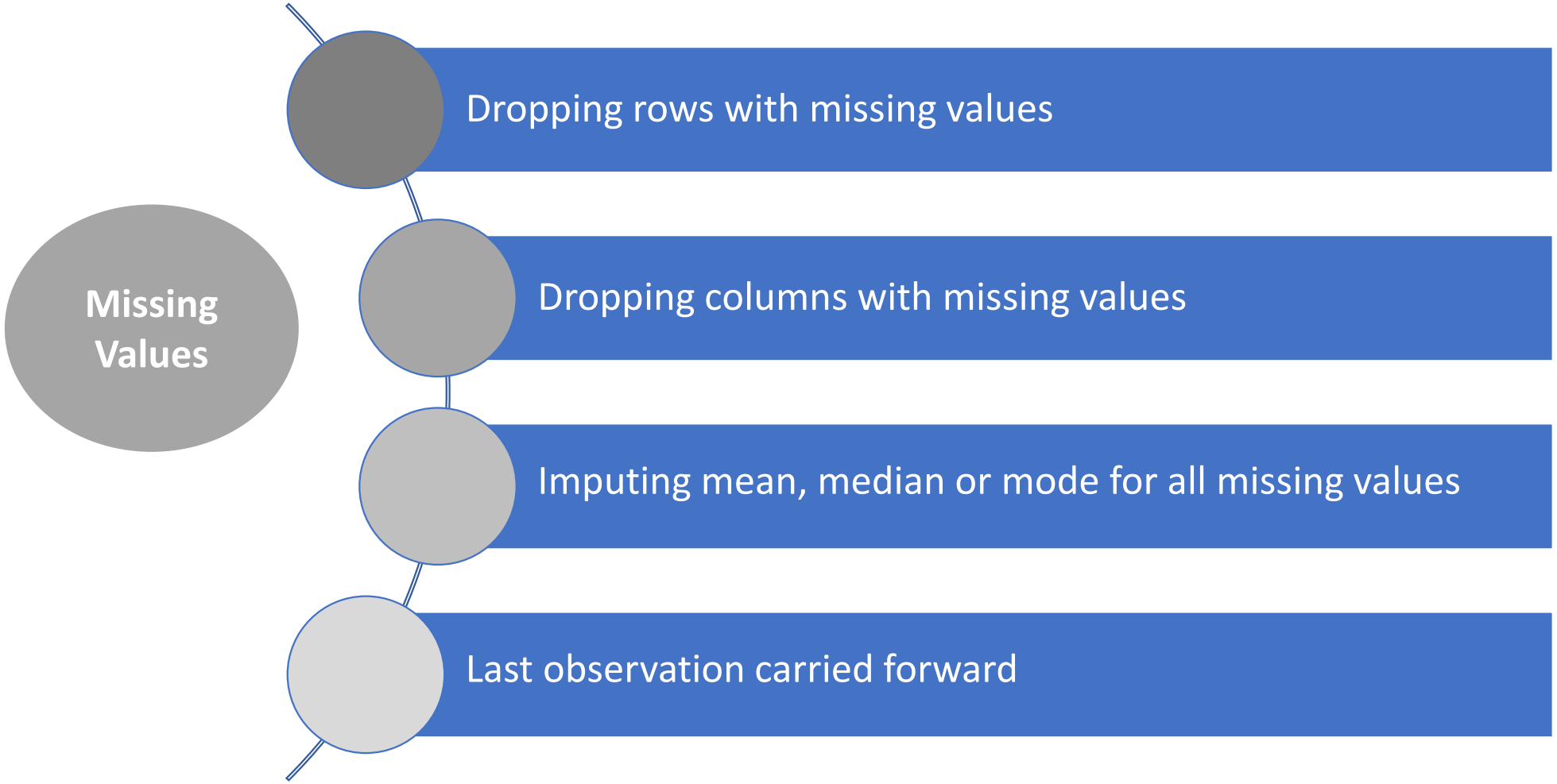

A B C
5 D 3
1 X 1
9 D 8
5 X 5
A B C
2 O 9
7 B 7
4 D 8
5 F 2
A B C
5 D 3
1 X 1
9 D 8
5 X 5
2 O 9
7 B 7
4 D 8
5 F 2
Concatenate
Combining Datasets

Combining Datasets

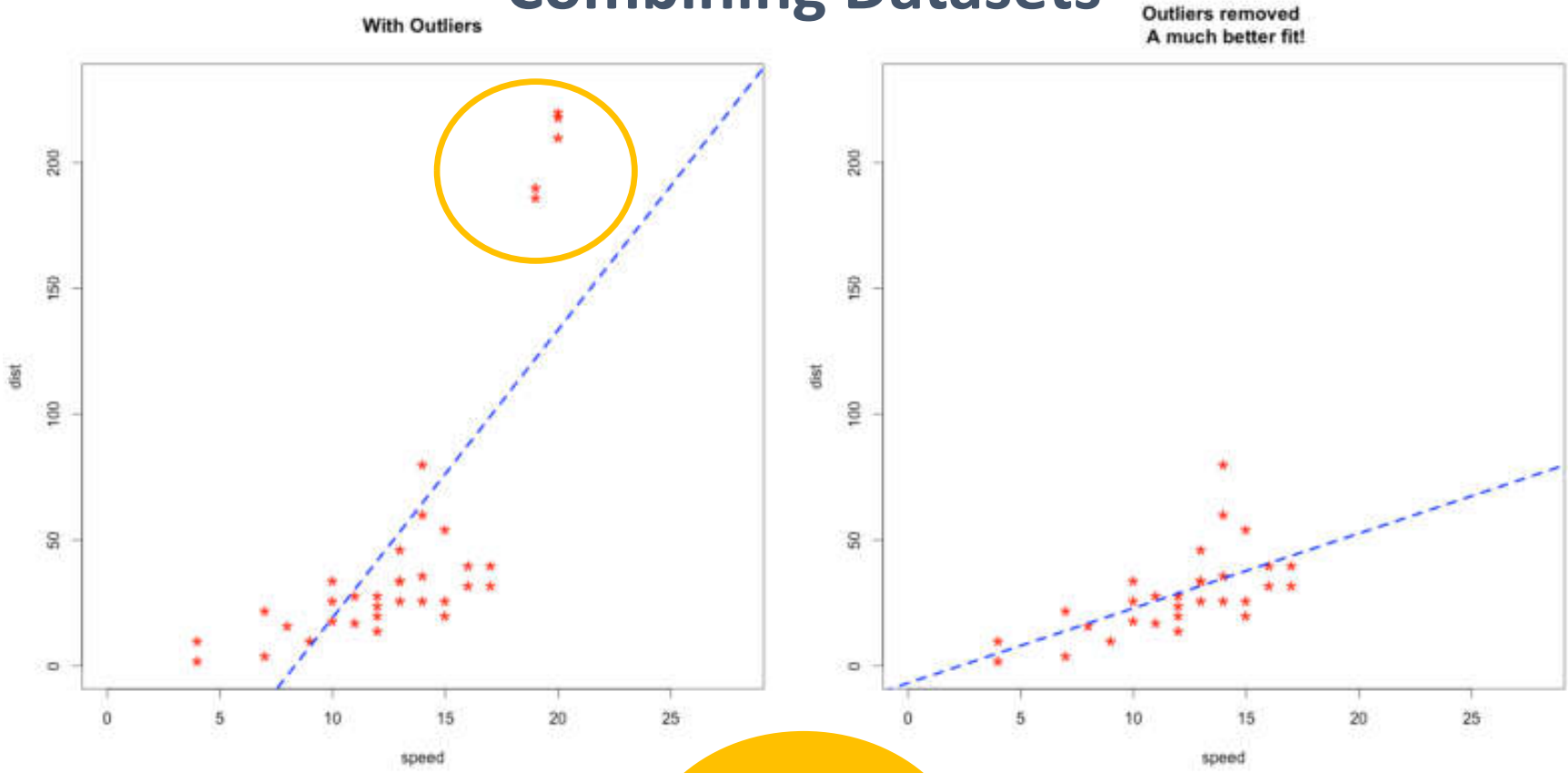


Merge

Combining Datasets



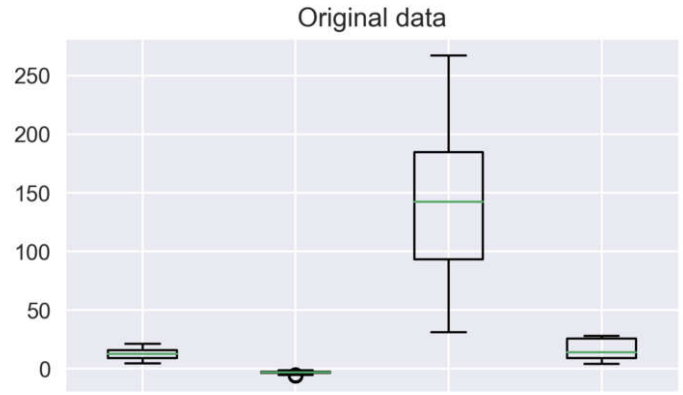
Combining Datasets

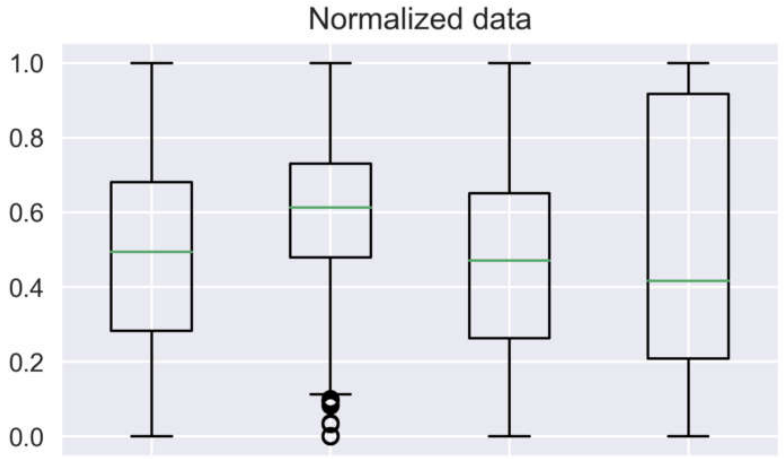


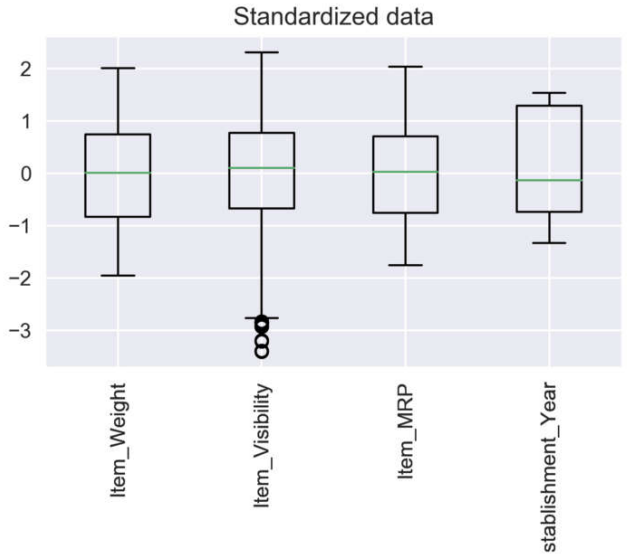
Outliers

Combining Datasets

Feature Scaling







SUMMARY

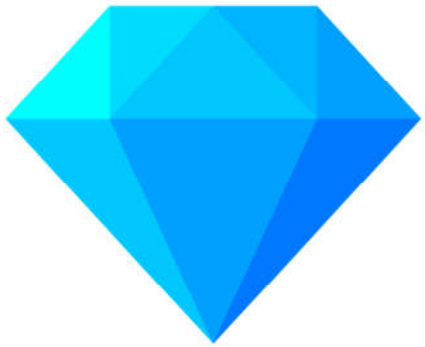
01 You may find data in multiple datasets and in many different formats.

02 You will need to make decisions in the process of data wrangling.

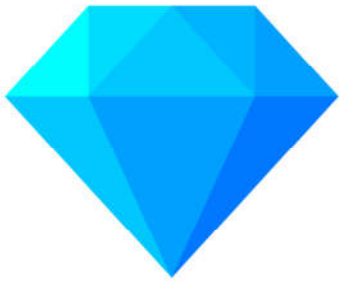
03 Keep in mind that your actions will have consequences.

Data Preparation: Homework

Harvesting, Storing, and Retrieving Data

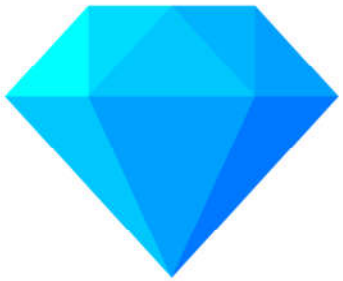


OpenRefine

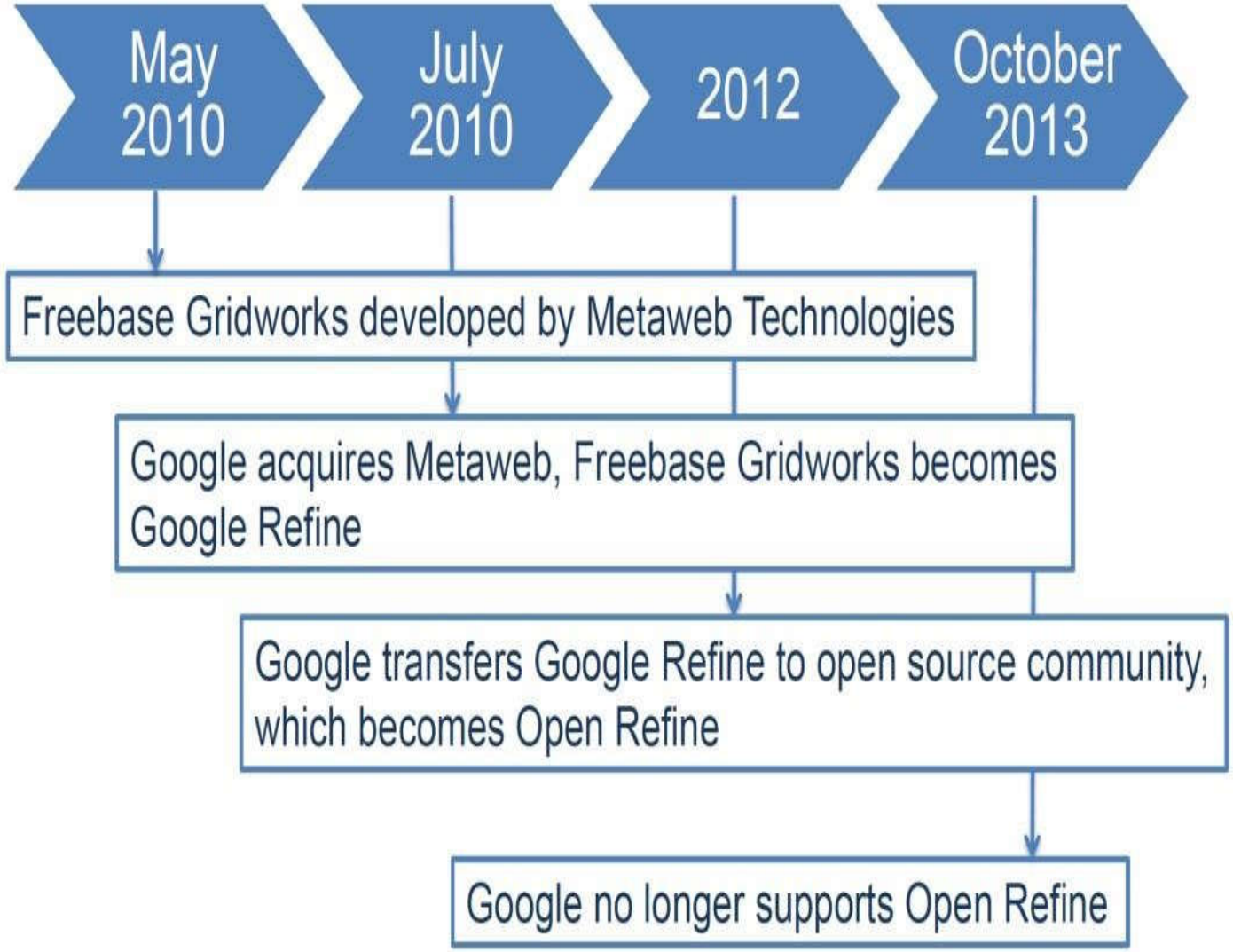


OpenRefine

- Open source
- Browser-based
- Designed for tabular data
- Supports CSV, Excel, XML, Google docs and RDF files



OpenRefine


May 2010
July 2010
2012
October 2013
Freebase Gridworks developed by Metaweb Technologies
Google acquires Metaweb, Freebase Gridworks becomes Google Refine
Google transfers Google Refine to open source community, which becomes Open Refine
Google no longer supports Open Refine