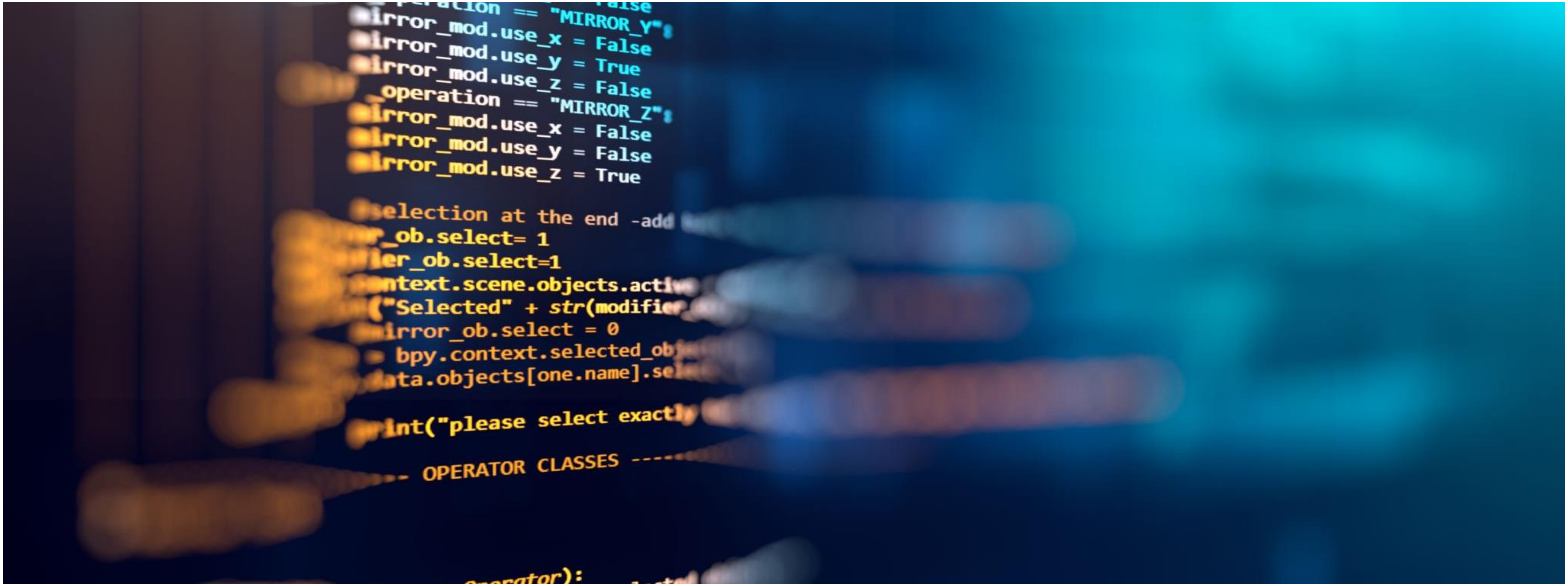
WORKSHOPS

Week 2 -Data cleaning
and data carpentry



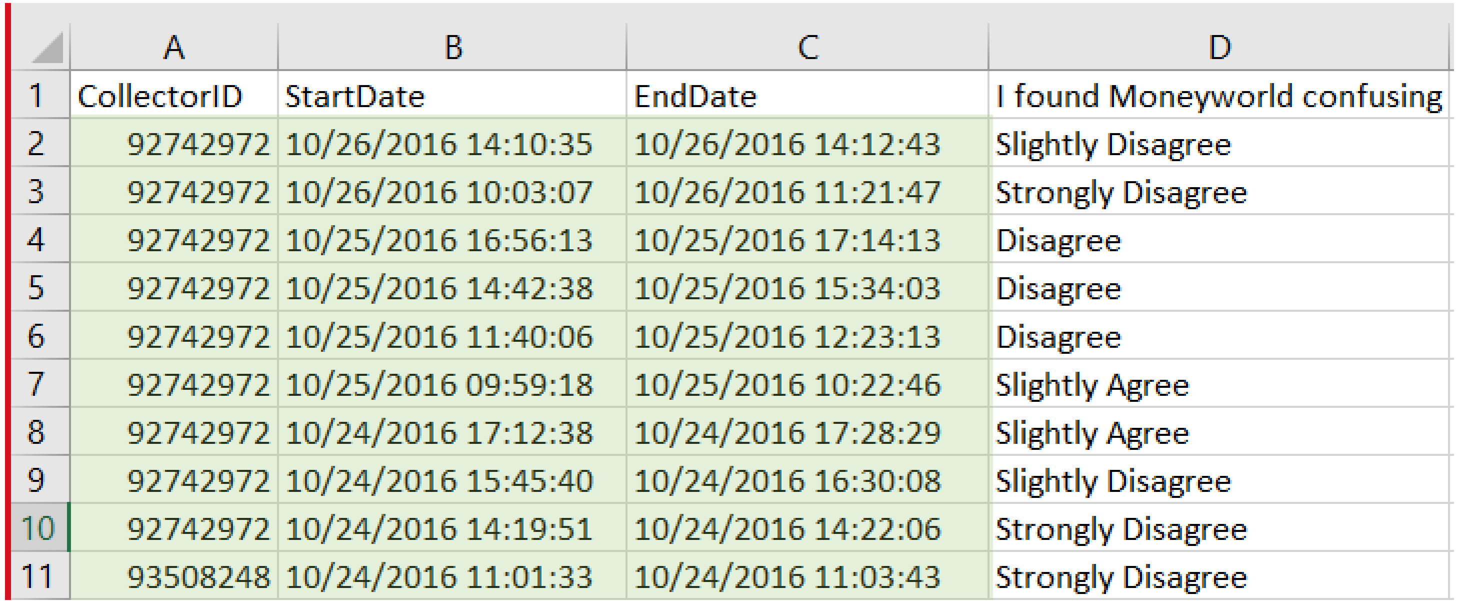
DATA: TYPES AND FORMATS

TYPES OF DATA

- Quantitative: numerical
- Qualitative: text

QUANTITATIVE

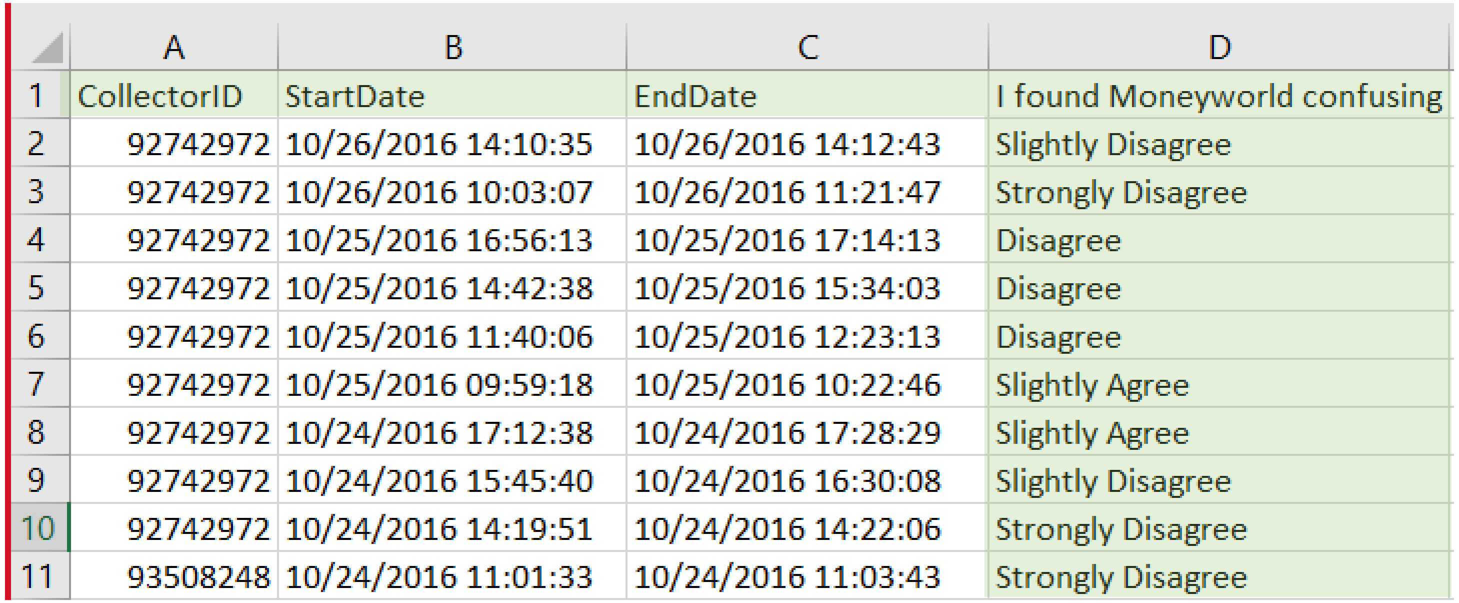
- Annual sales
- Profitability



	A	B	C	D
1	CollectorID	StartDate	EndDate	I found Moneyworld confusing
2	92742972	10/26/2016 14:10:35	10/26/2016 14:12:43	Slightly Disagree
3	92742972	10/26/2016 10:03:07	10/26/2016 11:21:47	Strongly Disagree
4	92742972	10/25/2016 16:56:13	10/25/2016 17:14:13	Disagree
5	92742972	10/25/2016 14:42:38	10/25/2016 15:34:03	Disagree
6	92742972	10/25/2016 11:40:06	10/25/2016 12:23:13	Disagree
7	92742972	10/25/2016 09:59:18	10/25/2016 10:22:46	Slightly Agree
8	92742972	10/24/2016 17:12:38	10/24/2016 17:28:29	Slightly Agree
9	92742972	10/24/2016 15:45:40	10/24/2016 16:30:08	Slightly Disagree
10	92742972	10/24/2016 14:19:51	10/24/2016 14:22:06	Strongly Disagree
11	93508248	10/24/2016 11:01:33	10/24/2016 11:03:43	Strongly Disagree

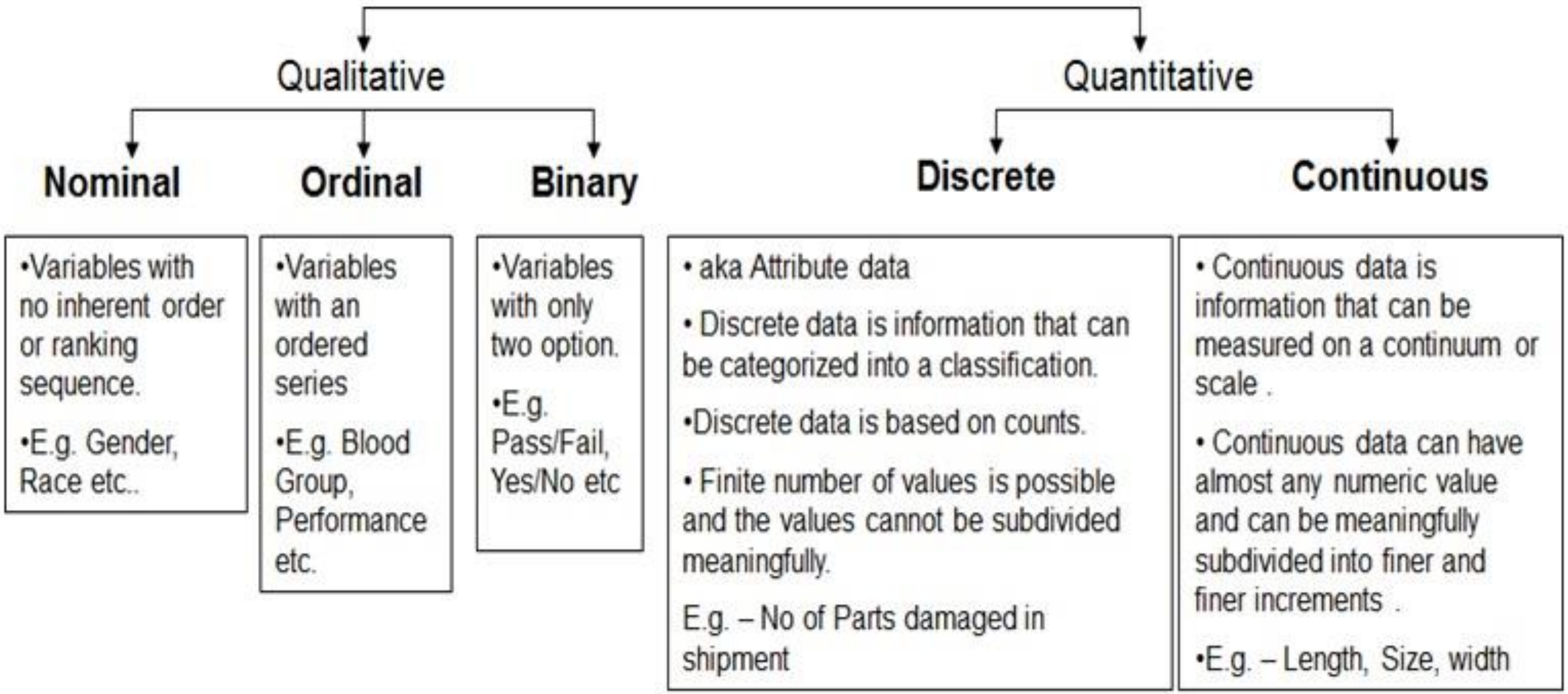
QUALITATIVE

- Customer reviews
- Explicit description of a malfunction in evaluation reports



	A	B	C	D
1	CollectorID	StartDate	EndDate	I found Moneyworld confusing
2	92742972	10/26/2016 14:10:35	10/26/2016 14:12:43	Slightly Disagree
3	92742972	10/26/2016 10:03:07	10/26/2016 11:21:47	Strongly Disagree
4	92742972	10/25/2016 16:56:13	10/25/2016 17:14:13	Disagree
5	92742972	10/25/2016 14:42:38	10/25/2016 15:34:03	Disagree
6	92742972	10/25/2016 11:40:06	10/25/2016 12:23:13	Disagree
7	92742972	10/25/2016 09:59:18	10/25/2016 10:22:46	Slightly Agree
8	92742972	10/24/2016 17:12:38	10/24/2016 17:28:29	Slightly Agree
9	92742972	10/24/2016 15:45:40	10/24/2016 16:30:08	Slightly Disagree
10	92742972	10/24/2016 14:19:51	10/24/2016 14:22:06	Strongly Disagree
11	93508248	10/24/2016 11:01:33	10/24/2016 11:03:43	Strongly Disagree

TYPES OF DATA



Qualitative

Nominal

- •Variables with no inherent order or ranking sequence.
- •E.g. Gender, Race etc..

Ordinal

- •Variables with an ordered series
- •E.g. Blood Group, Performance etc.

Binary

- •Variables with only two option.
- •E.g. Pass/Fail, Yes/No etc

Quantitative

Discrete

- • aka Attribute data
- • Discrete data is information that can be categorized into a classification.
- •Discrete data is based on counts.
- • Finite number of values is possible and the values cannot be subdivided meaningfully.

E.g. – No of Parts damaged in shipment

Continuous

- • Continuous data is information that can be measured on a continuum or scale .
- • Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
- •E.g. – Length, Size, width

Source: Six Sigma Institute

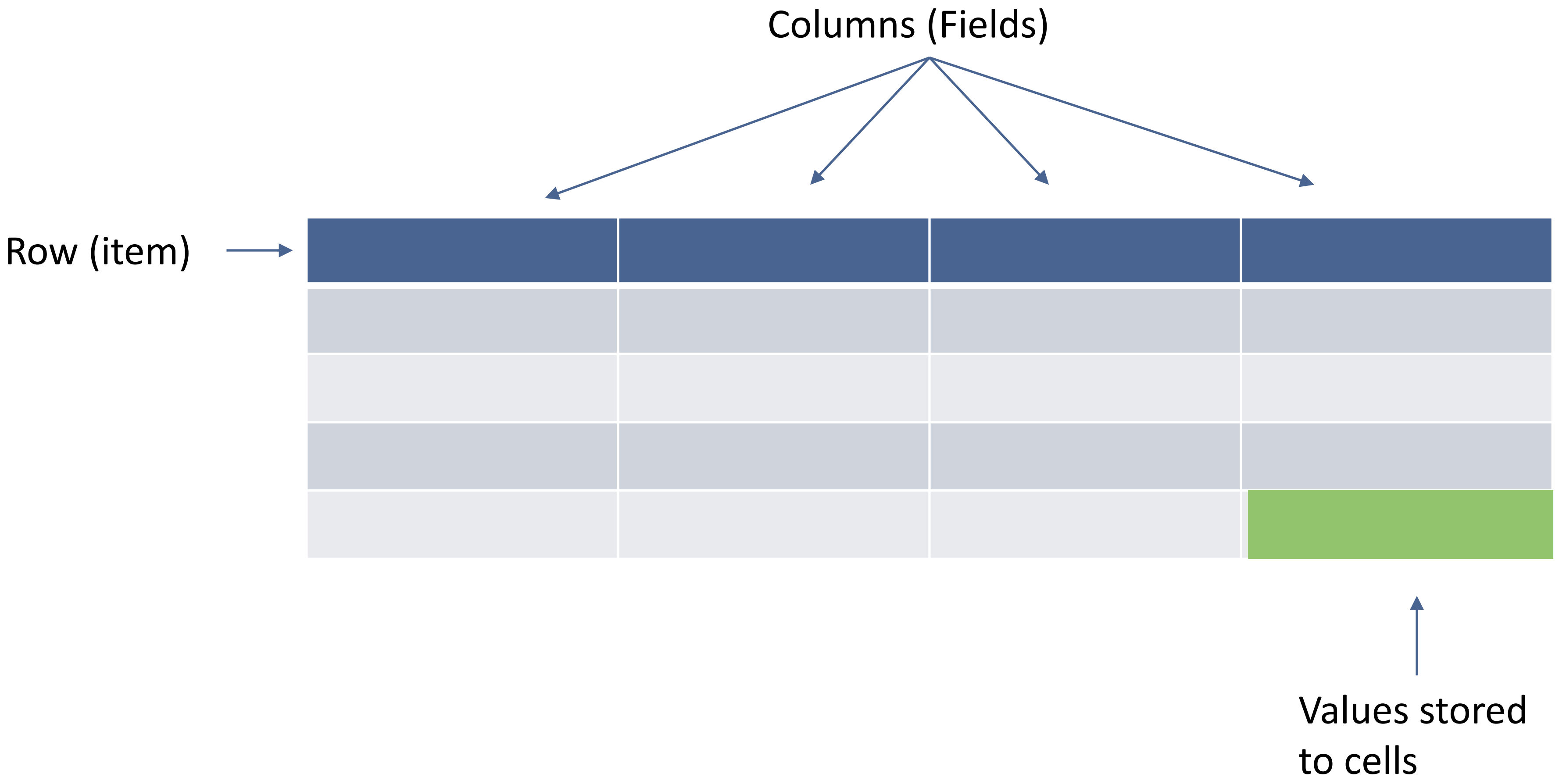
STRUCTURED DATA

- Tabular or spreadsheet-like data in which each column may be a different type (string, numeric, date, or otherwise). This includes most kinds of data commonly stored in relational databases or tab- or comma-delimited text files.
- Multidimensional arrays (matrices).
- Multiple tables of data interrelated by key columns (what would be primary or foreign keys for a SQL user).
- Evenly or unevenly spaced time series [1].

TABULAR AND CSV

- Tabular formats (ex. Excel)
- Comma-delimited text files, CSV format

TABULAR DATA

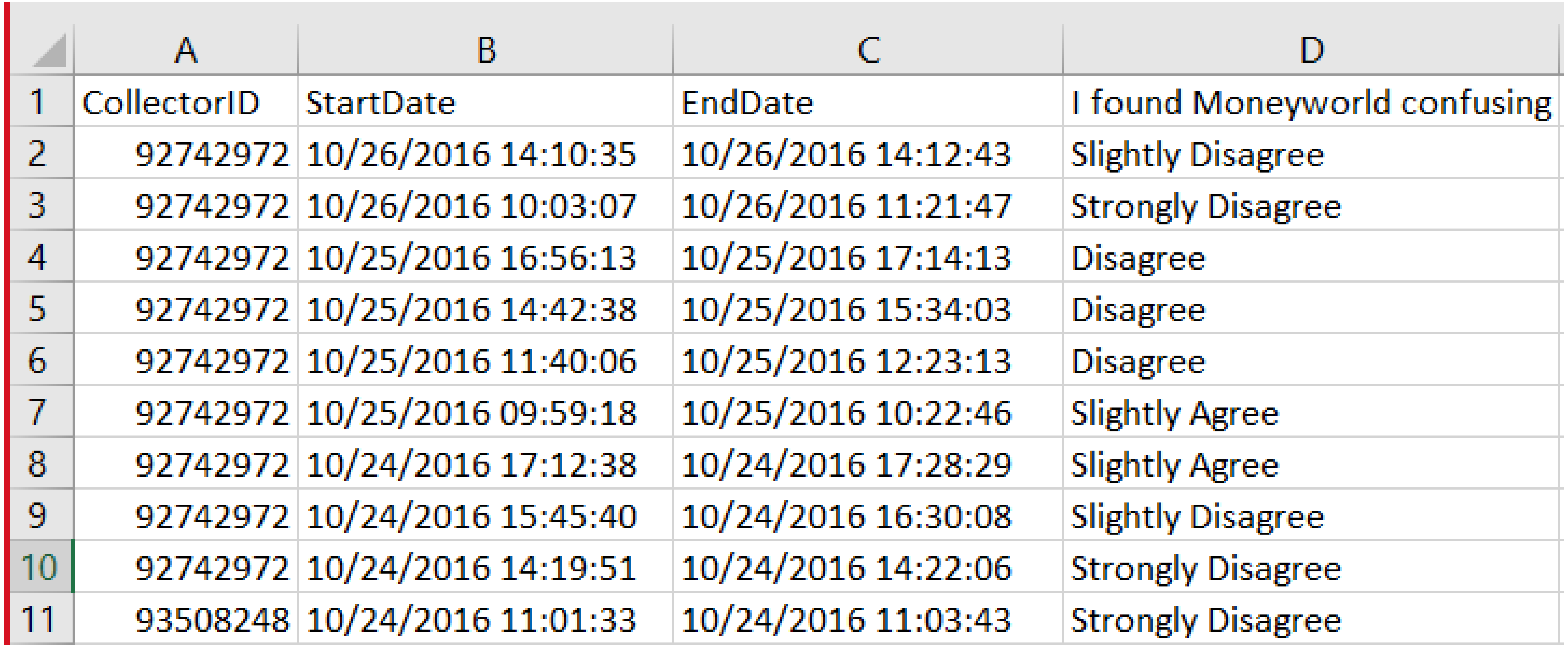


Columns (Fields)

Row (item)

Values stored to cells

TABULAR DATA

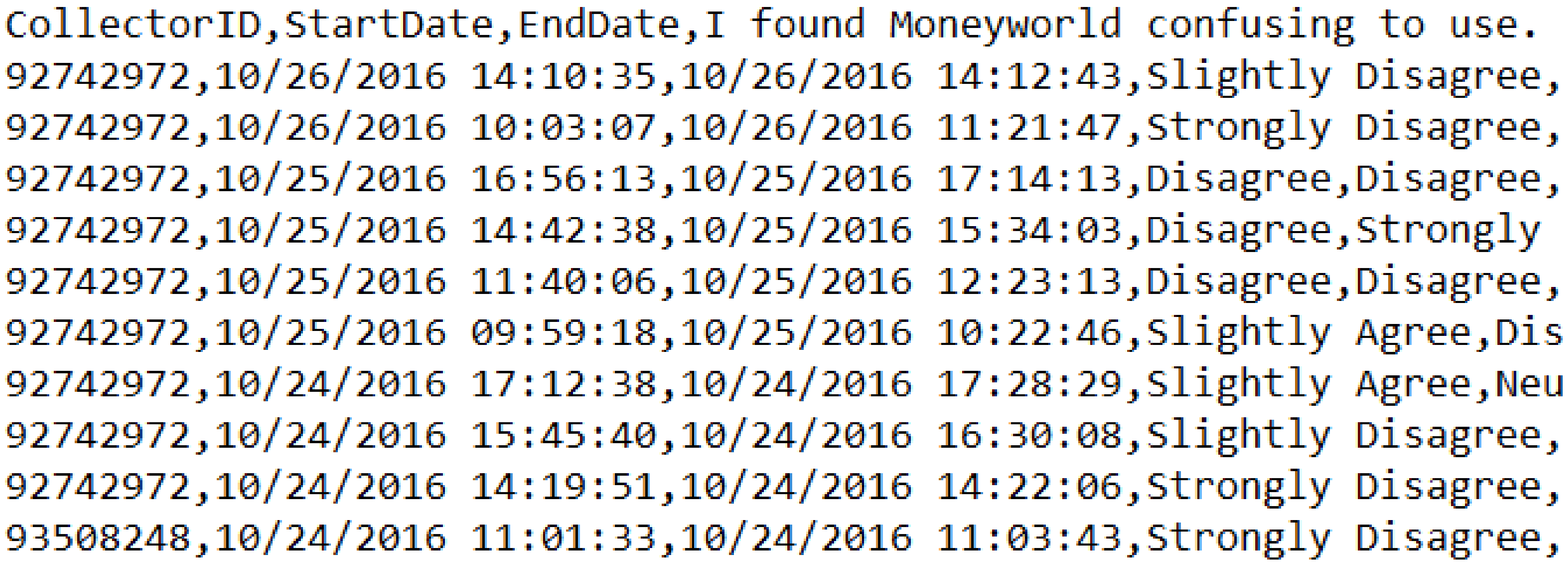


	A	B	C	D
1	CollectorID	StartDate	EndDate	I found Moneyworld confusing
2	92742972	10/26/2016 14:10:35	10/26/2016 14:12:43	Slightly Disagree
3	92742972	10/26/2016 10:03:07	10/26/2016 11:21:47	Strongly Disagree
4	92742972	10/25/2016 16:56:13	10/25/2016 17:14:13	Disagree
5	92742972	10/25/2016 14:42:38	10/25/2016 15:34:03	Disagree
6	92742972	10/25/2016 11:40:06	10/25/2016 12:23:13	Disagree
7	92742972	10/25/2016 09:59:18	10/25/2016 10:22:46	Slightly Agree
8	92742972	10/24/2016 17:12:38	10/24/2016 17:28:29	Slightly Agree
9	92742972	10/24/2016 15:45:40	10/24/2016 16:30:08	Slightly Disagree
10	92742972	10/24/2016 14:19:51	10/24/2016 14:22:06	Strongly Disagree
11	93508248	10/24/2016 11:01:33	10/24/2016 11:03:43	Strongly Disagree

CSV FORMAT

- CSV stands for Comma Separated Values
- Saves tabular information into a delimited text file with the series of values separated by commas
- It is lightweight and consumes less memory
- Each line of text is a single row
- It is human readable and can be opened using a text editor vs tabular data that are stored as binary files.

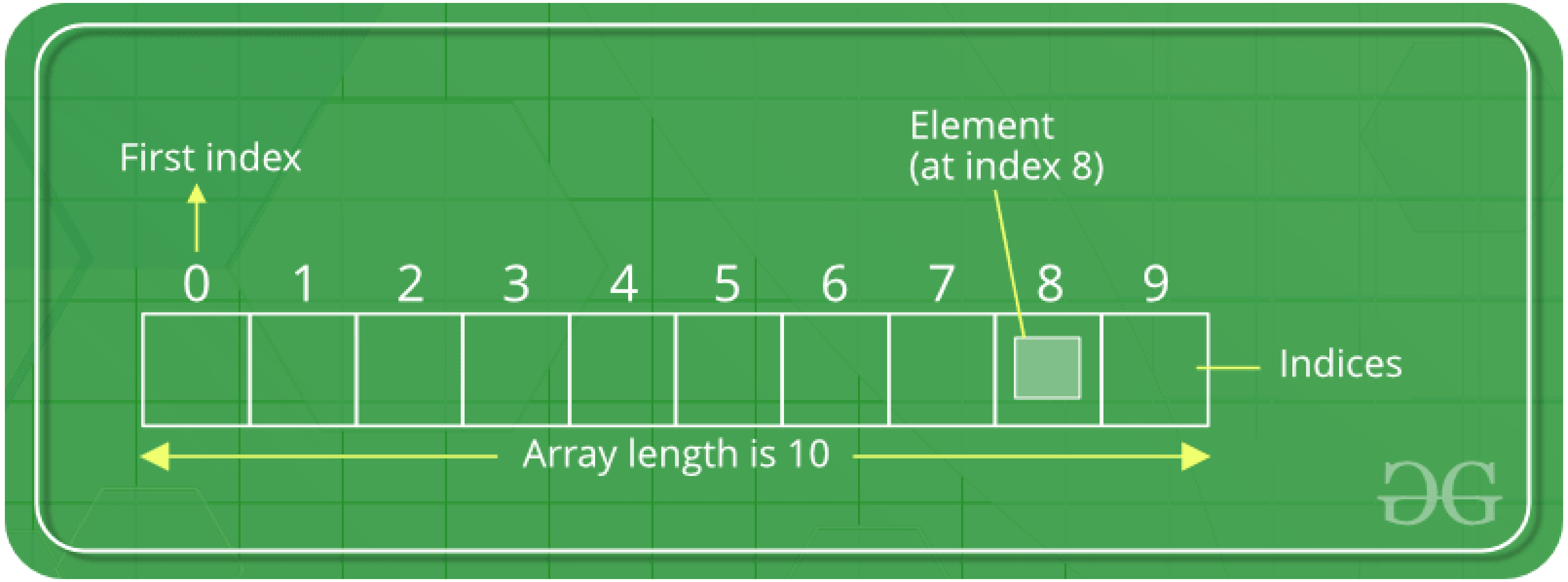
CSV FORMAT EXAMPLE



```
CollectorID,StartDate,EndDate,I found Moneyworld confusing to use.
92742972,10/26/2016 14:10:35,10/26/2016 14:12:43,Slightly Disagree,
92742972,10/26/2016 10:03:07,10/26/2016 11:21:47,Strongly Disagree,
92742972,10/25/2016 16:56:13,10/25/2016 17:14:13,Disagree,Disagree,
92742972,10/25/2016 14:42:38,10/25/2016 15:34:03,Disagree,Strongly
92742972,10/25/2016 11:40:06,10/25/2016 12:23:13,Disagree,Disagree,
92742972,10/25/2016 09:59:18,10/25/2016 10:22:46,Slightly Agree,Dis
92742972,10/24/2016 17:12:38,10/24/2016 17:28:29,Slightly Agree,Neu
92742972,10/24/2016 15:45:40,10/24/2016 16:30:08,Slightly Disagree,
92742972,10/24/2016 14:19:51,10/24/2016 14:22:06,Strongly Disagree,
93508248,10/24/2016 11:01:33,10/24/2016 11:03:43,Strongly Disagree,
```

ARRAYS

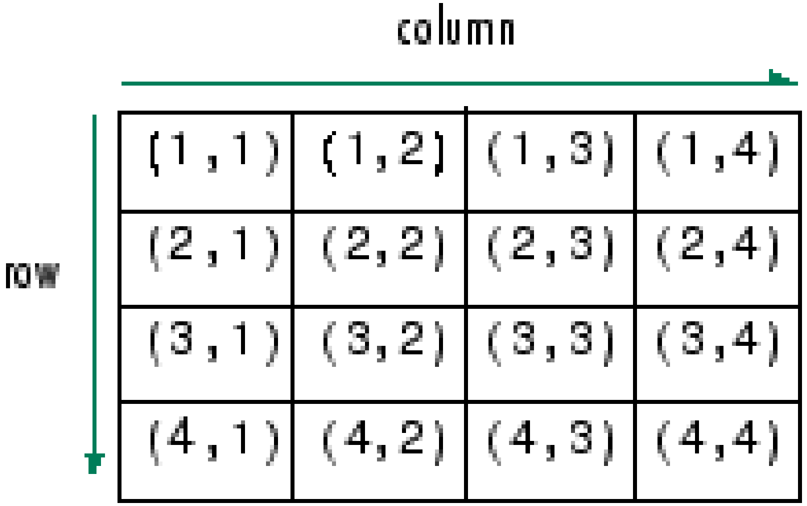
- An array is a special variable, which can hold more than one value at a time.



Source: geeksforgeeks

MULTIDIMENSIONAL ARRAYS

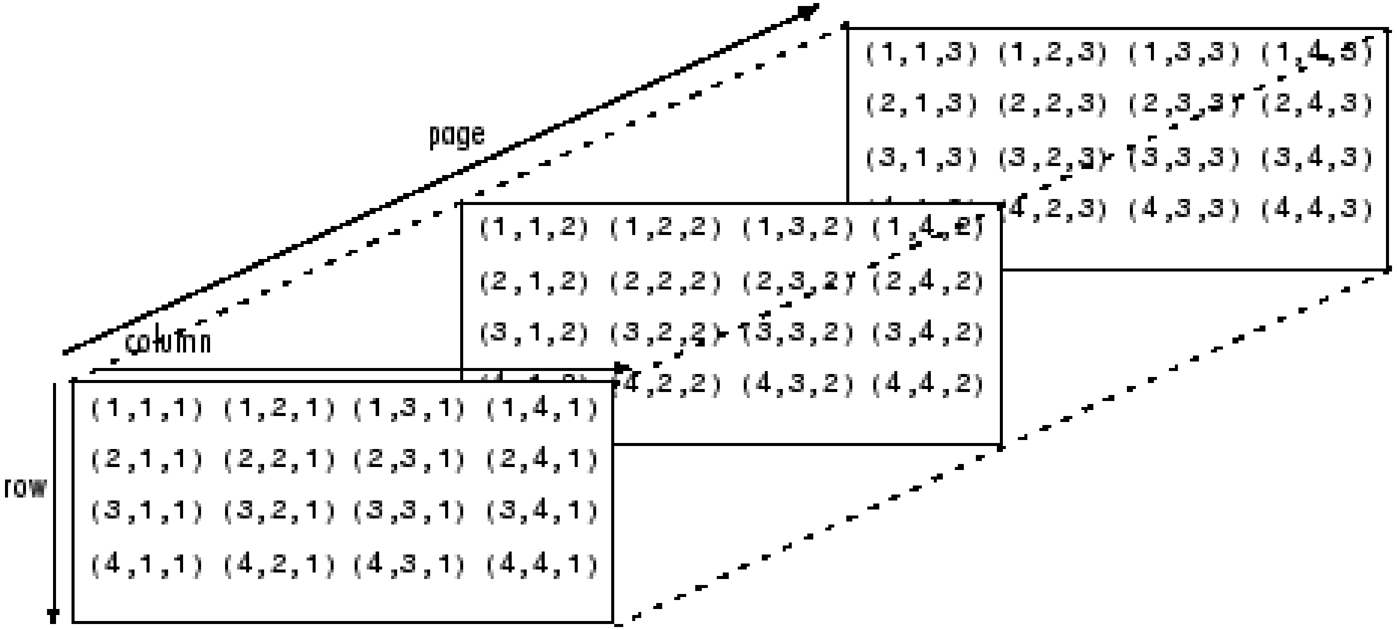
- Multidimensional arrays are data arrays with more than two dimensions. Each element is defined by two subscripts, the row index and the column index.



(1,1)	(1,2)	(1,3)	(1,4)
(2,1)	(2,2)	(2,3)	(2,4)
(3,1)	(3,2)	(3,3)	(3,4)
(4,1)	(4,2)	(4,3)	(4,4)

MULTIDIMENSIONAL ARRAYS

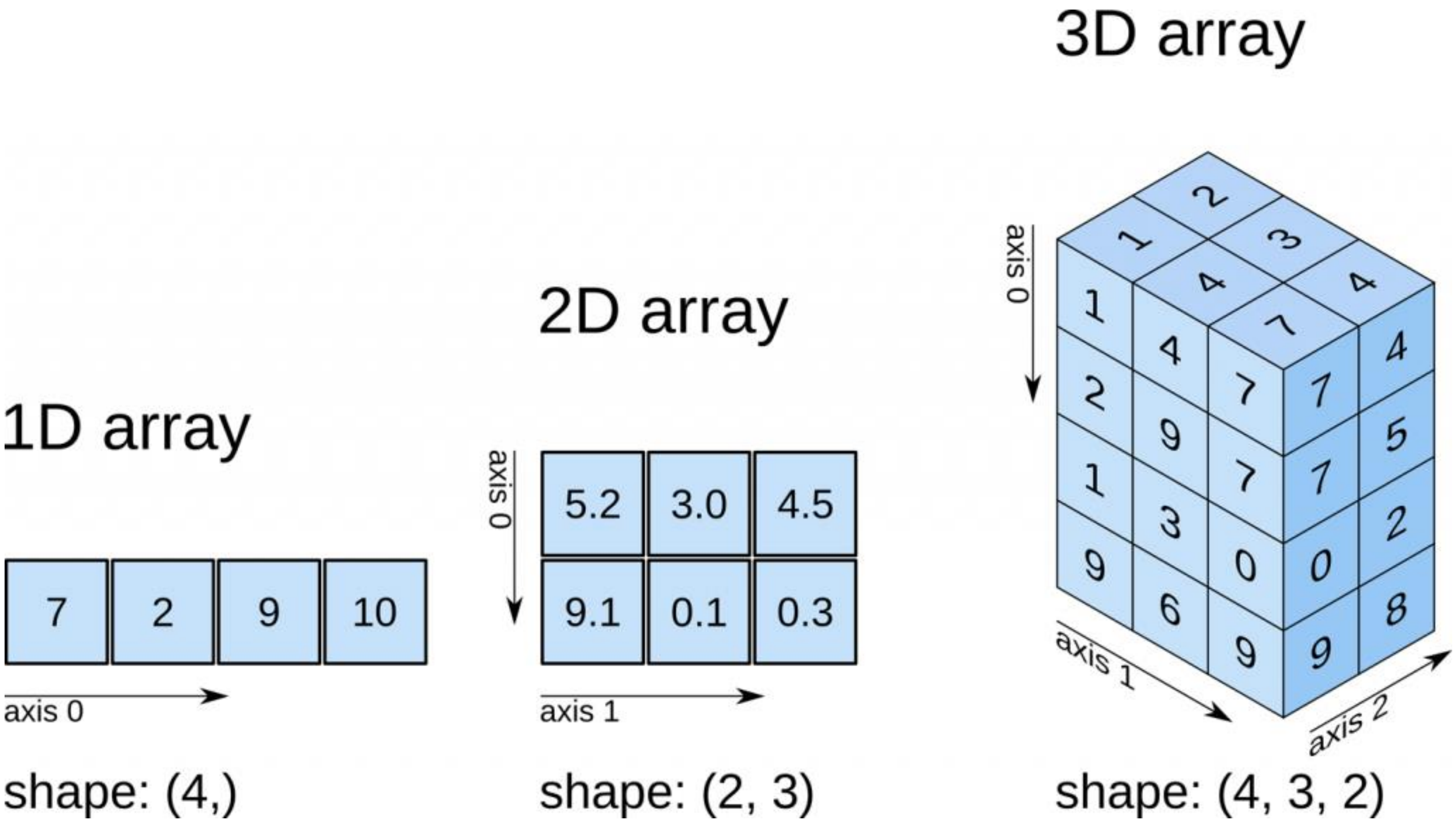
- Multidimensional arrays are an extension of 2-D matrices and use additional subscripts for indexing. A 3-D array, for example, uses three subscripts. The first two are just like a matrix, but the third dimension represents pages or sheets of elements [2].



(1,1,3) (1,2,3) (1,3,3) (1,4,3)
(2,1,3) (2,2,3) (2,3,3) (2,4,3)
(3,1,3) (3,2,3) (3,3,3) (3,4,3)
(4,1,3) (4,2,3) (4,3,3) (4,4,3)

(1,1,2) (1,2,2) (1,3,2) (1,4,2)
(2,1,2) (2,2,2) (2,3,2) (2,4,2)
(3,1,2) (3,2,2) (3,3,2) (3,4,2)
(4,1,2) (4,2,2) (4,3,2) (4,4,2)

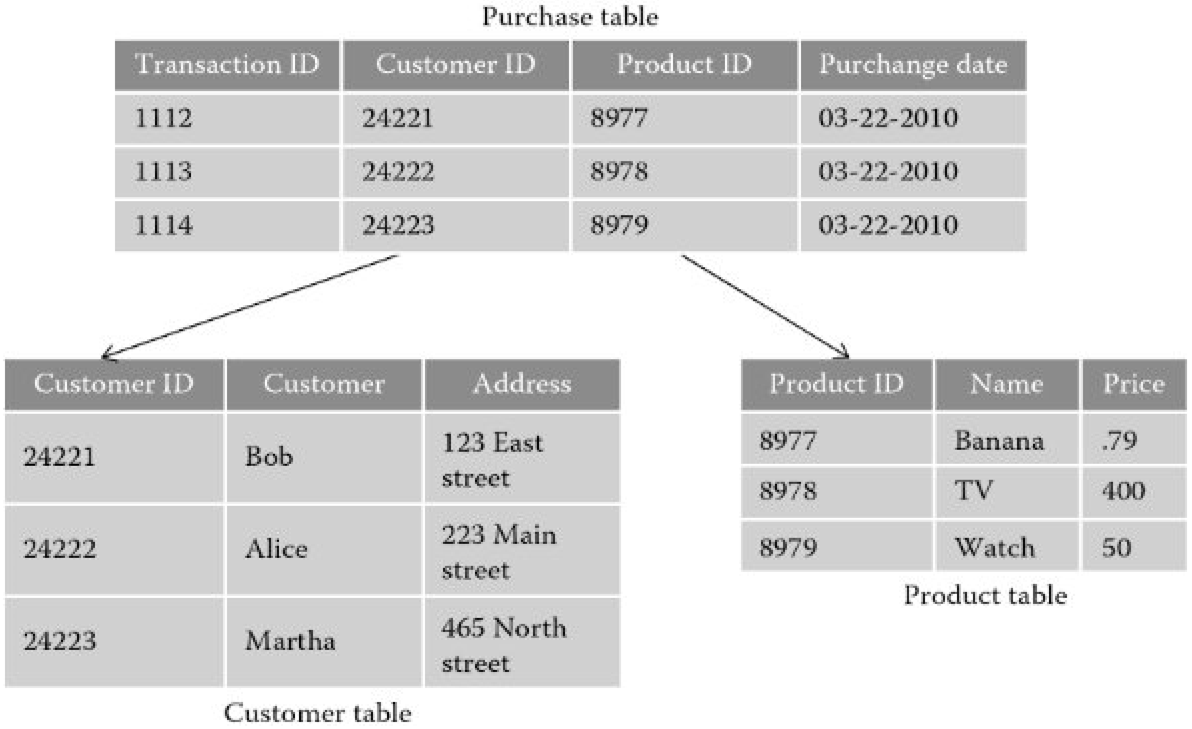
(1,1,1) (1,2,1) (1,3,1) (1,4,1)
(2,1,1) (2,2,1) (2,3,1) (2,4,1)
(3,1,1) (3,2,1) (3,3,1) (3,4,1)
(4,1,1) (4,2,1) (4,3,1) (4,4,1)



Source: predictivehacks

RELATIONAL TABLES

- Multiple tables of data interrelated by key columns (what would be primary or foreign keys for a SQL user).



Purchase table

Transaction ID	Customer ID	Product ID	Purchange date
1112	24221	8977	03-22-2010
1113	24222	8978	03-22-2010
1114	24223	8979	03-22-2010

Customer ID	Customer	Address
24221	Bob	123 East street
24222	Alice	223 Main street
24223	Martha	465 North street

Customer table

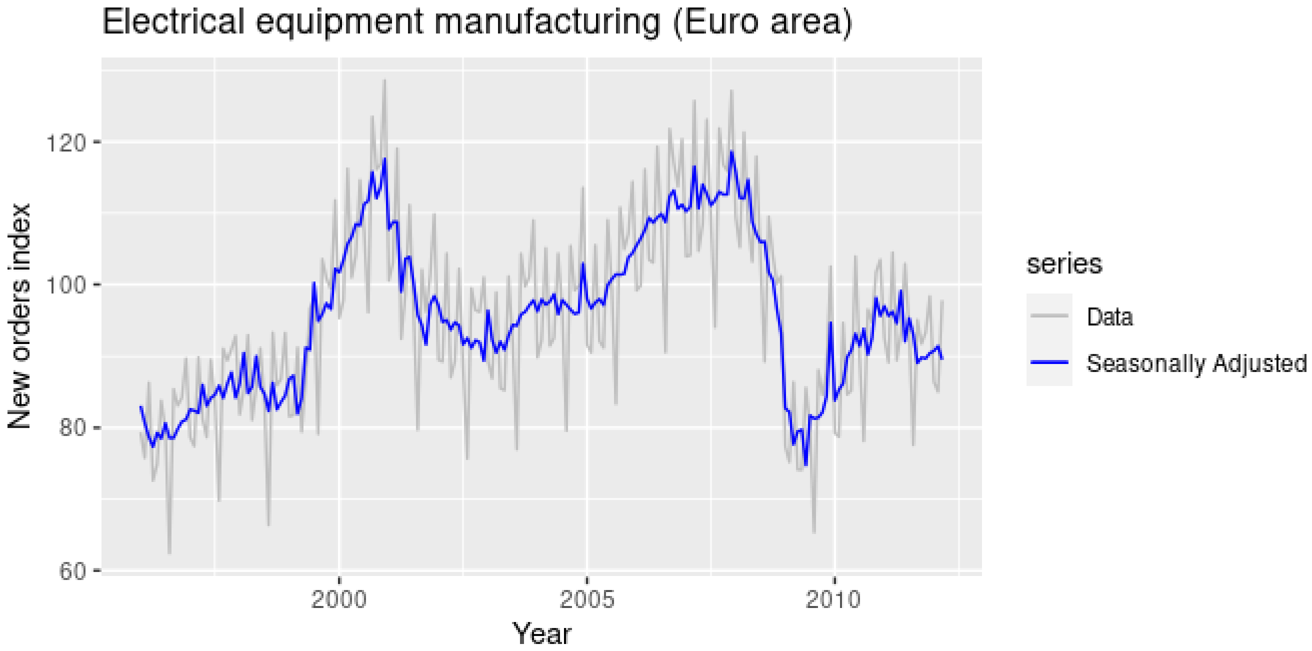
Product ID	Name	Price
8977	Banana	.79
8978	TV	400
8979	Watch	50

Product table

Image Source: Grace L. Samson Ph.D

TIME SERIES

- Time series data is a collection of quantities that are assembled over even intervals in time and ordered chronologically. The time interval at which data is collection is generally referred to as the time series frequency [3].



Source: otexts.com

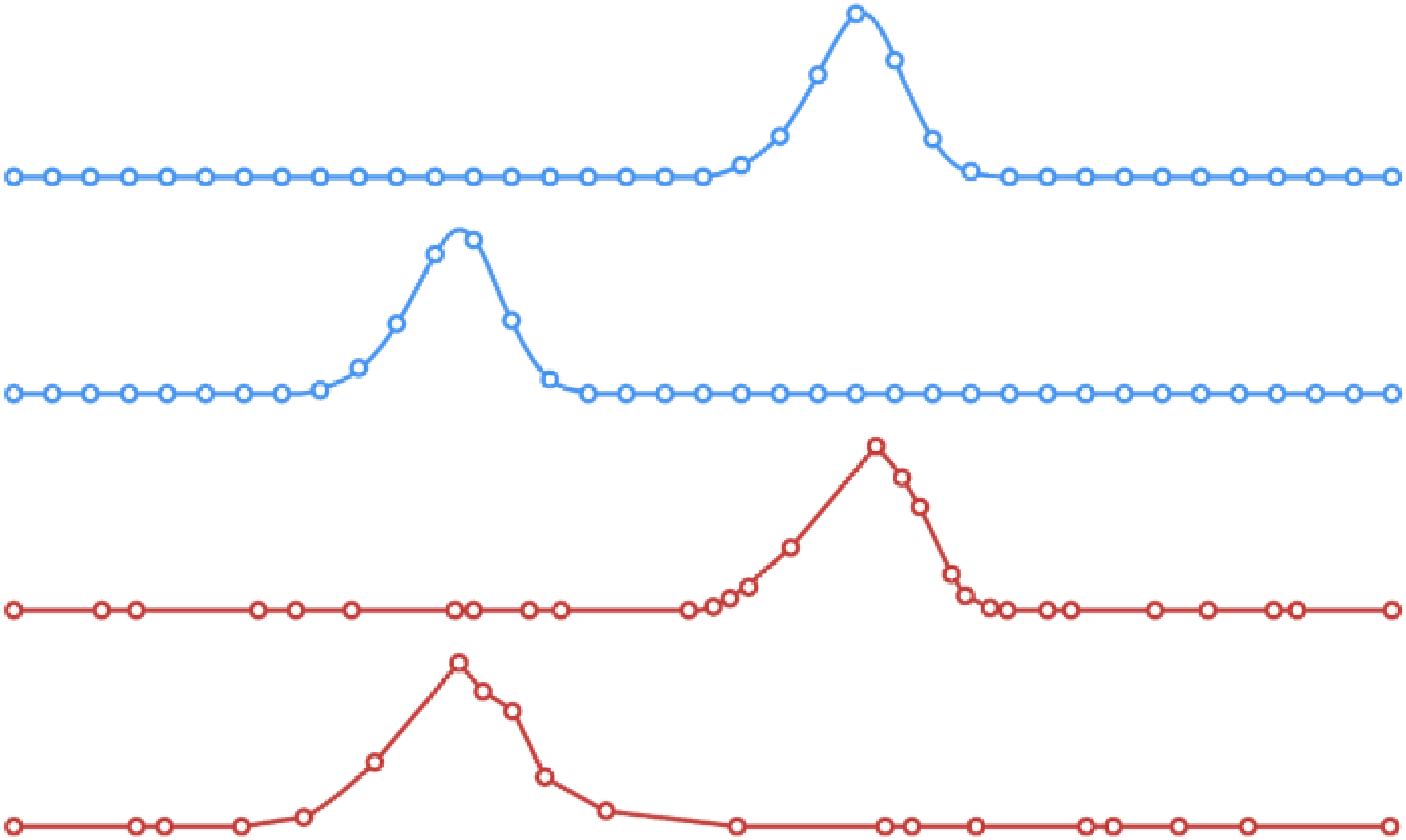
TIME SERIES

Metrics (Regular)

Measurements gathered at regular time intervals

Events (Irregular)

Measurements gathered at irregular time intervals



Source: influxdata

REFERENCES

1. Rosett C.M., Hagerty A. (2021) Data Wrangling. In: Introducing HR Analytics with Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-030-67626-1_13
2. Stefanski, R., Sinha, V. and Poddar, A., 2022. *Data Wrangling in 6 Steps: An Analyst's Guide For Creating Useful Data*. [online] Learn | Hevo. Available at: https://hevodata.com/learn/data-wrangling/#s2
3. Tripathi, S., Muhr, D., Brunner, M., Jodlbauer, H., Dehmer, M. and Emmert-Streib, F., 2021. Ensuring the Robustness and Reliability of Data-Driven Knowledge Discovery Models in Production and Manufacturing. *Frontiers in Artificial Intelligence*, 4.