Data Forms

Hemashree Kilari[1], Ram Kishore Malkari[2]

[1,2]Student, Department of Computer Science Engineering, GITAM University, Visakhapatnam, Andhra Pradesh, India

---***---

Abstract - *Data is the new electricity. The fourth industrial revolution is now underway. Artificial Intelligence and Big Data are the eras in which we live. There has been a tremendous data explosion, resulting in the emergence of new technology and smarter goods. On average, 2.5 exabytes of data are generated every day. In the last decade, the need for data has skyrocketed. Many businesses have made data the heart of their operations. In the IT industry, data has spawned new sectors. Industries require data in order for them to make informed judgments. Data Science is a technique that transforms raw data into actionable information. As a result, data is everywhere and changes depending on the sources from which it is generated. Examples include machines, devices like mobile phones, sensors, etc. With advancements in technology, processing power, and increased access to big data sets, businesses are increasingly adopting a data-driven approach. Rather than intuition, decisions are made based on data in a data-driven approach, which makes it simpler to be objective while making decisions. Organizations that are data-driven are more confident in their strategies. They will be able to make more precise projections in the future. They can swiftly assess results, test new techniques, and make necessary adjustments. They can adapt to market developments faster than the competitors as a result of this quick feedback. As a result, data-driven businesses become more agile and efficient. Therefore, data is the foundation for everything, and it is essential to understand all types of data. This paper will further discuss some of the most common data forms in more detail.*

***Key Words*: *Data, Data Forms, Data Mining, Data Science, Types of Data, Data-driven approach*.**

1.INTRODUCTION

Data is raw facts or a collection of information used to calculate, analyze, or plan to do something. Data volume triumphs over data quality. When it comes to how much to use, what to use is often more important. Even though it's an unbelievable truth that 90 percent of the world's data was generated over the last two years alone, less than 0.5% of all data generated is ever analyzed and used. At present, 2.5 quintillion bytes of data are created every day, and as the Internet of Things expands, this rate may rise much.

The data that is generated either falls under structured, semi-structured or unstructured. Structured data is clearly defined and easily searchable data. For example, structured data helps Google better understand your page content and is also used to enable rich results. Semi-structured, is the information that is not structured (like relational database) but still has some structure. Documents in the JavaScript Object Notation (JSON) format make up semi-structured data. It also comprises graph databases and key-value stores. On the other hand, unstructured data is disordered and is approximately 80% of the data that organizations process daily. Therefore, businesses must adapt to handle the increasing stores of unstructured data.

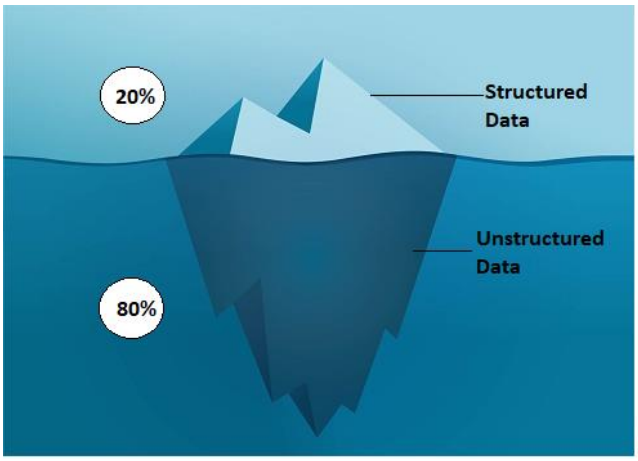

Figure1. Categorization of the Data

Every second in 2020, people generated 1.7 MB of data. And it is said that, by 2022, 70 percent of the world's GDP will be digitized. Imagine how much more data would be produced by 2022 if the entire digital universe was 44 zettabytes by the end of 2020. According to IDC, integrating IoT devices such as machines, sensors, and cameras would generate 79.4 zettabytes of data in 2025. Due to this rapid increase in the data rate, businesses are adopting a data-driven approach.

Data-driven refers to a strategy that is focused on data analysis to derive insights and interpretation of complex data rather than an observation. A data-driven approach is used to review and organize data in order to serve customers with a customer-centric approach better. It ensures that solutions and plans are supported by factual information and not just hunches, feelings, and anecdotal evidence. A data-driven approach helps to predict the future by using past and current information. With advances in technology, computing capabilities, and improved access to large data sets, there has been an increase in the data-driven approach. It is a lot easier to optimize with a data-driven approach. Therefore, not only in the data-driven approach, data is the foundation for everything, and it is essential to understand all types of data. This paper will further discuss some of the most common data forms in more detail.

2. TYPES OF DATA

Structured Data:

Under structured data, there will be a high degree of data organization and can easily be understood by machine language. Here the data is typically organized and stored in a tabular format like spreadsheets (Excel sheets, google sheets) or relational databases (traditional database management (DBMS)). Some examples of structured data include names, dates, addresses, credit card numbers, geolocation, etc. It is commonly said that about 20% of the world's data is structured. Due to this well-formatted structure, it is easy to input, search and manipulate the data relatively quickly, and it also requires less storage space. [1], [2]

Semi-structured Data:

Semi-structured data is a form of structured data that can't be organized in rational databases or has no strict structural framework. Yet, it has some organizational properties that make it easier to analyze, such as semantic tags. While in semi-structured entities that belong in the same class, they may have different attributes. Some of the examples include emails, HTML, XML, and other markup languages. This semi-structured data falls between structured and unstructured data, containing both the structured and unstructured aspects. Because of its higher level of organization than structured data, it is easy to analyze, and it is flexible, i.e., the schema can be easily changed. Also, this data is portable, and it can easily deal with the heterogenicity of sources. [3], [4]

Unstructured Data:

As the name says, the data is not structured, i.e., it is not structured via any predefined data models or schema. And it is comprised of data that is usually not as easily searchable. This Unstructured data may be textual or non-textual and human-generated. It includes formats like texts, blogs, documents, photos, videos, facts, sensor data, etc. According to IDC (International Data Corporation), 80% of unstructured data are never analyzed, also called dark data. To put it simply, we can compare this unstructured data to an iceberg because there is yet a massive amount of data to be analyzed for good decision-making. Therefore, this unstructured data has no proper schema and structure, so it isn't easy to store and manage and also indexing the data is complex and error-prone. [5], [6], [7]

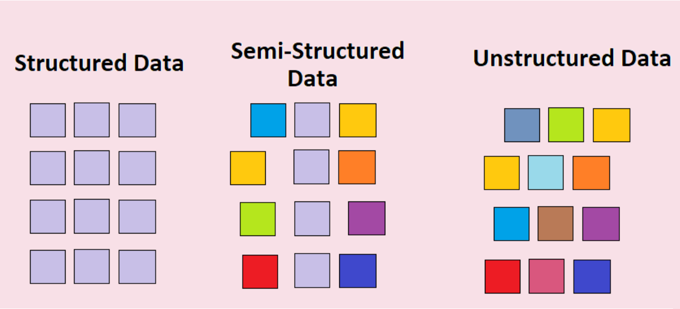

Figure2. Structured data Vs. Semi-structured data Vs. Unstructured data

Categorical Data:

Categorical data is statistical data which is a collection of information that consists of categorical variables divided into categories or groups. I.e., if some college is trying to get the biodata of its students, then the resulting data can be referred to as categorical. This grouped data could be derived from either quantitative data analysis grouped within the given intervals or from qualitative data analysis that is countable. This categorical data can take on numerical values. For example, we can take "1" for "yes" and "2" for "no," but it does not mean that those two numbers can have mathematical meaning. Some examples of these categorical variables are sex, race, age group, color, educational level, etc. In general, categorical data has observations and values that can be stored in groups or categories, and the best way to represent these data is by bar graphs, pie charts, Boolean plots, etc. This categorical data can be further divided into two groups, nominal and ordinal. On this categorical data, quantitative analysis cannot be applied. Therefore, arithmetic or numerical operations cannot be performed, and there is no limit to the kind of statistical analysis which can be performed on categorical data. [8]

Spatial Data:

Any data that directly or indirectly involves a specific geographical area or location can be categorized as spatial and can also be called geographical data. Generally, a location is described by coordinates along with a coordinate reference system (CRS). I.e., Spatial data = Coordinate + Topology. This data helps locate any feature or boundaries on earth, and the data can be represented in point, line, or polygon format. The temporal and attribute information is also required whenever an object/ event description is done along with the spatial information. For example, if for any city, if we want to prepare a disaster mitigation strategy, then along with the city boundary information, we also will be collecting extra information like the time of occurrence of the disaster, the number of people injured, amount of damage, etc. So, for collecting all this information, a Geographic Information System (GIS) is designed. Here all the geographic, temporal, and other information of an event/object will be captured, stored efficiently, analyzed, and presented in map format for the users. Spatial data can be stored either by using a raster or vector model. Therefore, the data can be stored in its original resolution and need not be generalized, and the data is always easy to program and quick to perform. However, the location of each vertex must be stored explicitly, and if large amounts of data exist, then processing attribute data may be inconvenient. [9], [10]

Quantitative Data:

Quantitative data is data that defines the range, count, or measurement. If any data collected from a population can be counted or measured, then that data is quantitative. This quantitative data has a unique numerical value associated with each data set, i.e., quantitative data are always numbers. For example, the amount of money you have, height, weight, the number of students in a class, etc. Quantitative data can be represented through graphs, charts, tables, and maps and can be displayed over time like a line chart. This kind of data can be used for mathematical computations and statistical analysis, which deals with real-life decisions like if a manufacturing company will need to answer a question, "How is the cost of production?". Now the company's production cost will be collected with the help of the question and can be informed to the company's selling cost. This quantitative data can be further divided into discrete and continuous data. This kind of data is precise and highly reliable, and also easy to communicate and understand. But due to the narrow collection of data set, there may be a problem of data loss and is highly complex. [11], [12]

Qualitative Data:

Qualitative data is data that describes the characteristics like the smell, taste, etc. This data type cannot be measured, but this can be categorized or described based on the attributes of the population. This data is represented by a bar graph, pie chart, and many more. This data can be categorized or described based on attributes and properties.

The Source of this data type comes from audio or video form interviews, electronic journals, social media posts, blog posts, responses to online surveys, etc. Qualitative data cannot be measured, i.e., any information that can be captured that is not numerical. This data can be categorized or described based on the attributes of a population. Qualitative data include audio or video form interviews, electronic journals, social media posts, blog posts, responses to online surveys, etc. This data is usually represented through pictures and visual representations, and they can be displayed graphically as a bar graph, pie chart, etc. Qualitative data can be further divided into binary, nominal, and ordinal. Using this qualitative data, we can have a better understanding and identification of behavior patterns. Also, it provides a good explanation. However, there can be a lesser reach, possibility bias, and can be time-consuming too. [13], [14]

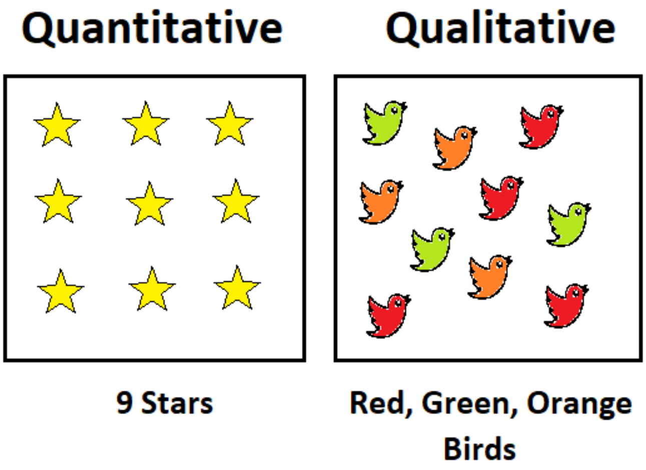

Figure3: Quantitative Vs. Qualitative

Nominal Data:

If the data can be named or labeled and do not contain any quantitative value, i.e., data with no numeric value, such data is nominal. It is are also called categorical data that are divided into various groups. This nominal data usually does not follow any order, so there won't be any change in the meaning even if the order is changed. There is no hierarchy. For example, the race is a nominal variable that can have a vast number of categories, but there is no hierarchy or any order like highest to lowest or vice versa. This nominal data will usually be collected via questions. For nominal data, data interpretation is complex and can never be quantified. [15],[16]

Ordinal Data:

Unlike nominal data, ordinal data involves some order, i.e., the values will be ordered. This data ordering is significant, but the differences between the values cannot be known. For example, in ordered levels of your education like elementary, high school, undergraduate, and graduate, we cannot decide that the difference between elementary and high school is the same as graduate and undergraduate. The measurements of these ordinal values are typically non-numeric concepts like happiness, unhappiness, satisfaction, discomfort, etc. Due to its "ordered" nature, ordinal data will be used to carry our questionaries or surveys. In ordinal data, as the gaps between the values are not equal, the mean cannot be used to assess the central tendency of data. [17], [15]

Discrete Data:

When values in a data set can only take specific and countable values, it is called discrete data. Unlike continuous data, discrete data can't be measured. Some examples of discrete data can be the number of players in a team, the number of planets in the solar system, etc. This data can also be categorical, i.e., it contains a finite number of data values, such as the gender of a person, etc. Discrete data can be easily visualized and demonstrated using simple statistical methods like bar graphs, frequency tables, line plots (number line), pie charts, etc. These discrete data are distributed discretely in terms of time and space, making data analysis more practical. [18], [15]

Continuous Data:
Continuous data is a type of numeric data or can distribute over date and time, and the data can be significantly divided into smaller increments, including fractional and decimal values. This type of data can't be counted but is measurable, and there can be an infinite number of possible values between any two ranges. For example, when you measure height, weight, temperature, etc., you have continuous data. As these numbers are not collected from precise measurements, they are always unclean and tidy. Some continuous data will change over time, like the weight of a baby in its first year, the speed of wind during a storm, etc. This continuous data is graphically represented using histograms. [19], [20]

Sequential Data:
Sequential data occurs when points in a dataset are reliant on other points in the dataset. A Timeseries, such as a stock price or sensor data, is an example of this, where each point represents an observation at a specific point in time. Sequential data also includes sequences, DNA sequences, and meteorological data. The RNN, or Recurrent Neural Network, has a technique that can handle sequential datasets. This is also the crux of the issue that the recurrent neural network is attempting to solve. [21]

Spatiotemporal data:
It is information that is related to both space (spatial) and time (temporal) [24]. When data is collected across both geography and time to characterize a phenomenon in a specific area and time period, it is employed in data analysis. The data can produce different results depending on how space is defined, a zip code, a state. Time can also provide conflicting answers depending on whether it is measured in seconds, minutes, hours, days, or years. Sources are geographical information systems (GIS), graphic processing units (GPU). The commonly used methods for data visualization of Spatio-temporal data are mainly the combination of graphs, statistical charts, time axes, maps(heat map), etc. The commonly used techniques such as highlighting, scaling, fish-eye technology, association updating, and dynamic change. Combined with human-computer interaction visualization technology, combined with white background maps, three-dimensional virtual and other scenes for visual display. Examples are Tracking of moving objects, which typically can occupy only a single position at a given time, geometry changing over time, Historical tracking of plate tectonic activity. [22], [23]

Boolean data:
This kind of data has two possibilities like, true or false, representing the truth values of Boolean algebra. In programming languages, we use Boolean data for some decision-making in programs as conditional statements, Boolean expressions, comparison of values, Boolean operations. It falls within the heading of structured qualitative data. Graphs and scatter plots can be used to depict Boolean data. [25], [26]

Machine data:
It's data that's been generated by a machine. It is also digital data produced by network devices such as mobile phones, laptops, websites, servers, and embedded systems. The vast majority of humans will not alter machine data. It is automatically generated and collected. A final significant category of machine data is Metadata., which is information associated with an event that describes the circumstances in which it occurred. Setups for monitoring oil and gas pipelines, natural disaster warning systems based on feeds from marine sensors, forecasting systems that use data from satellites and weather stations to help predict the weather in small geographic areas, and building energy are all examples of applications that use machine data. A management system that analyzes HVAC and elevator data to improve efficiency. Machine Data Used For Operations Analytics, Security Analytics, Business Analytics. Machine data can be represented using bar graphs, histograms, box plots, graphs, pie charts. [27]

Synthetic Data:
As the name suggests, synthetic data is an alternative to real-world data generated by computer simulations or algorithms. This data is artificial because it cannot be generated from actual events. I.e., Synthetic data is created in digital worlds and reflects real-world data, mathematically or statistically. This artificial data is generated mainly to preserve privacy, product testing, and training machine learning algorithms hence crucial for business. Synthetic data is divided into three types, fully synthetic data, partially synthetic data, and hybrid synthetic data. These can be in the form of text, media (image, video, audio, etc.), or tabular data. Also, synthetic data is cheap to produce and is used in a wide range of applications. [28]

Longitudinal Data:
Longitudinal data, also known as panel data, can be defined as at least three or more measurements on the same unit with multiple units involved. I.e., the data set contains observations on multiple entities (companies, countries, cities, individuals, etc.) wherein every entity is observed at two or more points in time. These units can often be individuals, but not always. For example, blood pressure in patients measured every week for six weeks or math test scores of students measured in grades 3 through 8, etc. Moreover, these measurements can take different forms based on the study's design, including the data that are continuous or dichotomous. This panel data is a combination of cross-sectional and time-series data. Longitudinal data allows researchers to follow their subjects in real-time and helps to find long-term patterns. However, it costs higher

and takes more research time with unpredictability factors always present. [29]

Bivariate Data:

Bi means two and the second half variate means variable, so bivariate data are datasets that store two variables measured from the same observation. That is, often, more than one variable is collected on each individual. For example, in extensive health studies of a population, it is very common to obtain variables such as sex, age, weight, height, total cholesterol, and blood pressure on each individual. On the other hand, economic studies may be interested in, among other things, personal income and years of education. Another example would be, most university admission committees ask for an applicant's high school grade point average and admission test scores like SAT, etc. This bivariate data can be analyzed using scatter plots, hex plots, or stacked plots. The bivariate data set shows no relationship or linear relationship if the scatter plot is random with no tilt. With the help of bivariate analysis, one can quickly determine to what extent it becomes easier to know and predict a particular value for a variable. Although it does not factor in how variables could influence each other. [30]

Real-Time Data:

We know that most of the things that one does happen in real-time. Real-time is just a phrase to describe the time in which a particular event or an action takes place. When it comes to data, real-time refers to the processing of data that co-occurs as it would in the real world. Whenever business operations run at a very high speed, generating substantial data volumes and operational complexity abounds, real-time data exploration and visualization become increasingly critical to manage regular operations. Examples include bubble charts with varying-sized bubbles, line charts that draw by themselves, changing lights based on polling or voting, computer games, etc. The usage of real-time data increases the response time and efficiency and minimizes the risk. However, the process is more expensive and time-consuming. [31]

3. CONCLUSION:

As seen, the data can be classified into numerous categories. For instance, quantitative (nominal and ordinal) and qualitative (discrete and continuous) are two categorizations. They might also be sequential, temporal, structured, unstructured, cross-sectional, bivariate, etc. Each of these data kinds has a unique way of expressing itself. Quantitative data, for example, can be represented in a variety of forms, such as graphs, charts, tables, and maps. Pictures and other visual representations, on the other hand, characterize qualitative data. These representations include statistical approaches and visualizations, which are just a way to examine data to understand it better. We live in an era where there is a massive urge for data science because of the increasing amounts of data. Data science is quickly gaining traction as a topic that is transforming both science and industry. With work becoming more data-driven across nearly all disciplines, it influences both the jobs offered and the skills required. As a result, aspects of the economy, society, and daily life will grow dependent on data as more data and means to analyze it become available. Hence, data plays a prominent role everywhere, including in the data-driven approach. So, knowing the various data forms and how they can be represented aids in a better grasp of the subject.

REFERENCES

[1] Devin Pickell, “Structured vs Unstructured Data – What’s the Difference?,” 2018. https://www.g2.com/articles/structured-vs-unstructured-data (accessed Sep. 28, 2021).

[2] “What is structured, semi structured and unstructured data? › Michael Gramlich.” https://www.michael-gramlich.com/what-is-structured-semi-structured-and-unstructured-data/ (accessed Sep. 28, 2021).

[3] “Semi Structured Data: A Comprehensive Guide In 7 Points.” https://www.jigsawacademy.com/blogs/big-data-analytics/semi-structured-data (accessed Sep. 28, 2021).

[4] “What Is Semi-Structured Data?” https://blog.hubspot.com/marketing/semi-structured-data (accessed Sep. 28, 2021).

[5] S. Sato, A. Kayahara, and S. I. Imai, “Unstructured data treatment for big data solutions,” IEEE International Symposium on Semiconductor Manufacturing Conference Proceedings, May 2017, doi: 10.1109/ISSM.2016.7934512.

[6] “Unstructured data treatment for big data solutions | IEEE Conference Publication | IEEE Xplore.” https://ieeexplore.ieee.org/document/7934512 (accessed Sep. 28, 2021).

[7] K. Adnan and R. Akbar, “An analytical study of information extraction from unstructured and multidimensional big data”, doi: 10.1186/s40537-019-0254-8.

[8] “Visualizing Multivariate Categorical Data - Articles - STHDA.” http://www.sthda.com/english/articles/32-r-graphics-essentials/129-visualizing-multivariate-categorical-data/ (accessed Sep. 28, 2021).

[9] “What is spatial data and how does it work?” https://searchsqlserver.techtarget.com/definition/spatial-data (accessed Sep. 28, 2021).

[10] I. Franch-Pardo, B. M. Napoletano, F. Rosete-Verges, and L. Billa, “Spatial analysis and GIS in the study of COVID-19. A review,” Science of The Total Environment, vol. 739, p. 140033, Oct. 2020, doi: 10.1016/J.SCITOTENV.2020.140033.

[11] “What is Qualitative Data?” https://searchcio.techtarget.com/definition/qualitative-data (accessed Sep. 28, 2021).

[12] “Analyzing Quantitative Data | SAGE Publications Ltd.” https://uk.sagepub.com/en-gb/eur/analyzing-quantitative-data/book210308 (accessed Sep. 28, 2021).

[13] “What is Qualitative Data?” https://searchcio.techtarget.com/definition/qualitative-data (accessed Sep. 28, 2021).

[14] “Qualitative Data Analysis | SAGE Publications Inc.” https://us.sagepub.com/en-us/nam/qualitative-data-analysis/book246128 (accessed Sep. 28, 2021).

[15] J. R. Dettori and D. C. Norvell, "The Anatomy of Data," Global Spine Journal, vol. 8, no. 3, p. 311, May 2018, doi: 10.1177/2192568217746998.
[16] "Nominal Data | What Is It and How Can You Use It?" https://www.scribbr.com/statistics/nominal-data/ (accessed Sep. 28, 2021).
[17] Johnson, Valen E., Albert, and James H., Ordinal Data Modeling.
[18] Santner, Thomas J., Duffy, and Diane E., The Statistical Analysis of Discrete Data.
[19] S. Schmitz, R. Adams, and C. Walsh, "The use of continuous data versus binary data in MTC models: A case study in rheumatoid arthritis," BMC Medical Research Methodology 2012 12:1, vol. 12, no. 1, pp. 1–17, Nov. 2012, doi: 10.1186/1471-2288-12-167.
[20] C. McCue, "Continuous Variable," Data Mining and Predictive Analysis, pp. 67–92, 2007, doi: 10.1016/B978-075067796-7/50027-1.
[21] P. Kröger, Y. Lu Sommer, A. Zimek, and Y. Lu, "Knowledge Discovery in Databases II Lecture 4-Sequential Data".
[22] A. v Manzhosov et al., "IOP Conference Series: Earth and Environmental Science Recent citations Research on the Visualization of Spatio-Temporal Data", doi: 10.1088/1755-1315/234/1/012013.
[23] K. R. Ferreira, A. G. de Oliveira, A. M. V. Monteiro, and D. B. F. C. de Almeida, "TEMPORAL GIS AND SPATIOTEMPORAL DATA SOURCES," Revista Brasileira de Cartografia, vol. 68, no. 6, pp. 1191–1202, 2016, Accessed: Sep. 28, 2021. [Online]. Available: http://www.seer.ufu.br/index.php/revistabrasileiracartografia/article/view/44492
[24] C. Yang, K. Clarke, S. Shekhar, and C. V. Tao, "Big Spatiotemporal Data Analytics: a research and innovation frontier," https://doi.org/10.1080/13658816.2019.1698743, vol. 34, no. 6, pp. 1075–1088, Jun. 2019, doi: 10.1080/13658816.2019.1698743.
[25] "Boolean data type - Wikipedia." https://en.wikipedia.org/wiki/Boolean_data_type (accessed Sep. 28, 2021).
[26] S. P. Pencheva, "Seminar Intermediate Report".
[27] C. Atik and B. Martens, "Competition Problems and Governance of Non-personal Agricultural Machine Data: Comparing Voluntary Initiatives in the US and EU," 2021, Accessed: Sep. 28, 2021. [Online]. Available: https://www.farmdatacode.
[28] "What Is Synthetic Data? - Unite.AI." https://www.unite.ai/what-is-synthetic-data/ (accessed Sep. 28, 2021).
[29] "Longitudinal Data - Definition, Latest News, and Why Longitudinal Data is Important?" https://cleartax.in/g/terms/longitudinal-data (accessed Sep. 28, 2021).
[30] A. Bertani, G. di Paola, E. Russo, and F. Tuzzolino, "How to describe bivariate data," Journal of Thoracic Disease, vol. 10, no. 2, p. 1133, Feb. 2018, doi: 10.21037/JTD.2018.01.134.
[31] D. Croushore, "FRONTIERS OF REAL-TIME DATA ANALYSIS," 2009.