https://doi.org/10.57017/jorit.v3.1(5).04

The What and How of Data Analysis

Sidharta CHATTERJEE
School of Economics, Andhra University[1], India

Abstract

In this paper, we discuss about what and how of data science and data analysis: i.e., the approach and the mechanism that analysts employ while working with data. A Philosophical approach to analysis of data and data science has been undertaken those peeks into the conceptual world of aspects of the epistemology of data science. The paper also highlights the role played by analysts, tools, and specialized techniques that analysts employ in data science to derive insights from data. The discussion demonstrates the complexities associated with data science, and by what mechanism and how organizations and businesses draw insights that constitute the real value of data, and that which lay hidden deep within datasets constituting as a form of resource and asset for the organizations.

Keywords: creativity, data analysis, big data, data types, heuristics, statistical tools, data science.

JEL Classification: O11.

Introduction

> *"Unanticipated novelty, the new discovery, can emerge only to the extent that his anticipations about nature and his instruments prove wrong."*
>
> *— Thomas S. Kuhn*[2]

In this paper, we will discuss the conceptual aspects of data science and data analysis (Carpineto & Romano, 2004). In data analysis, information is learned at each step (Peng & Matsui, 2015). Each step teaches the analyst about the mechanism of data analytic process. From the beginning until the end, the entire process involves several instructional steps on how to proceed further, what methods of data exploration to adopt, and how to carry out the entire process of data analysis. In the beginning, planning precedes. Following planning, models come into being. Models are checked based on the hypotheses constructed to design experiments. This is followed by organization of the collected data, its categorical classification, tabulation, dealing with missing values and numbers, etc., before the data is fed to the system. According to Peng & Matsui (2015), the entire process is non-linear and highly iterative, akin to something that can be considered the epicycle of data analysis. But before one attempt to analyse data, one must have in mind a framework to conduct data analysis. This is to ensure that a coherent study protocol is followed in an organized manner based upon which the core activities of data analysis should rest.

There are numerous methods of data analysis, each with its own advantages (and *drawbacks*) attached to them. When we encounter data, we should ask questions that help formulate hypotheses. These are as follows:

[1] Andhra University, Waltair Junction, Visakhapatnam-530003.

[2] Kuhn, T. S. (1997). The structure of scientific revolutions (Vol. 962). Chicago: University of Chicago press.

- What is the data about?
- How has it been obtained or collected? i.e., validate sources.
- How well is the data structured?
- Is the data *qualitative* or *quantitative* in nature?
- What are the dependent and independent variables, if they could be identified, and how do they relate to the data?
- What expectations can we set upon the data so obtained?
- How can we formulate hypotheses based on the nature of the data?
- What statistical methods would be best suited to examine and analyse the data?
- What visualization metrics should one employ to represent the results graphically?[3]
- Is the data appropriate for the question?

These are a few essential questions one should be asking when beginning to undertake data analysis. Why? Because it is necessary to ponder or think about what one should expect before conducting an analysis. To some extent, it may be considered important to accommodate the role of the analyst's intuition in the success of data analysis. Several experts working on the field of data science (Spicer, 2005; Sanger, 2006) have stressed this issue. The role of intuition in strategic and business decision-making (Khatri & Ng, 2000; Sinclair & Smith, 2009), on the other hand, is slowly gaining ground among the scholars studying this area. That intuition is an important factor to consider in both strategic decision-making and data analysis is now being stressed by workers active in this area of research. Nevertheless, it is often unconvincing to overrule the importance of "a priori" knowledge in data analysis as it is a given knowledge based on which our assumptions are often developed.

For most of the process, i.e., *forecasting* and *trend analysis*, using knowledge from past historical data is heavily relied upon. It depends, largely, on the analysts' expertise and ability to make the data analysis process a success, and intuition may play a significant role. In addition, several technical and methodological aspects rule the game as well. Equally important is the interpretation of the data following its analysis. Much like data analysis, *data interpretation* could be viewed as an art.

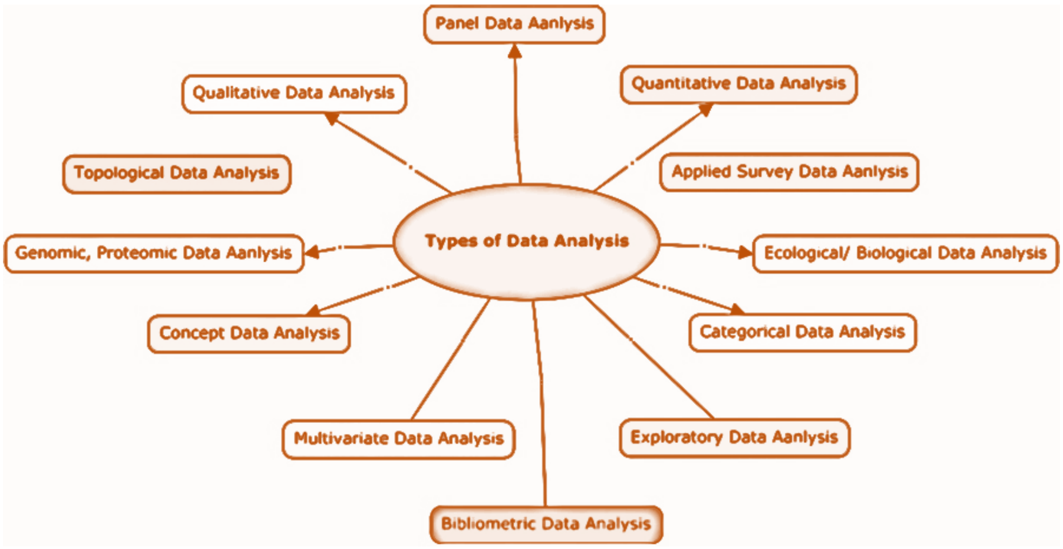
1. Types of Data Analysis

As data types are many and varied, so there exist different types of data analysis. These are as follows:

- Qualitative data analysis (Miles and Huberman, 1994);
- Quantitative data analysis (Sheard, 2018);
- Multivariate data analysis (Murtagh, and Heck, 2012);
- Exploratory data analysis (Tukey, 1977);
- Topological data analysis (Wasserman, 2018);
- Panel data analysis (Hsiao, 2022);
- Applied survey data analysis (Heeringa, West and Berglund, 2017);
- Concept data analysis (Carpineto and Romano, 2004);
- Categorical data analysis (Agresti, 2012);
- Bibliometric data analysis (Garfield, 1980; Battisti and Salini, 2013; Donthu, et al., 2021);
- Bioinformatics, Genomics, and Proteomics data analysis (Smith, 2000).

[3] See Sadiku et al. (2016).

Figure 1. Types of data analysis


Panel Data Aanlysis
Qualitative Data Analysis
Quantitative Data Analysis
Topological Data Analysis
Applied Survey Data Aanlysis
Genomic, Proteomic Data Aanlysis
Types of Data Analysis
Ecological/ Biological Data Analysis
Concept Data Analysis
Categorical Data Analysis
Multivariate Data Analysis
Exploratory Data Aanlysis
Bibliometric Data Analysis

In this paper, we discuss all these issues related to the effective analysis of data to derive coherent results. We will not go into the details of each type of analysis, but provide a simple outline of how different types of data are subjected to testing and validation using different methods and approaches of analysis. Data science has become an important domain on account of datafication that renders various aspects of the world that can be quantified for measurement and analysis. Besides, data science is a methodology that uses insights from analysis of data that has widespread applications in business and organizations. Data science relies heavily on exploratory statistics that uses different methodological approaches used extensively in business intelligence to take evidence-based decisions (Igual & Seguí, 2024).

2. Methodological Approaches to Data Analysis

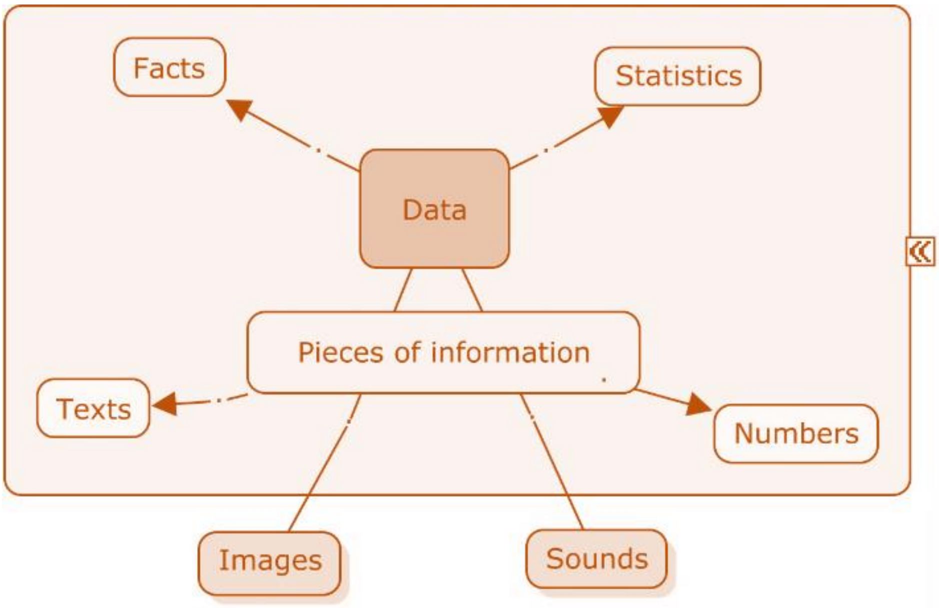
The methods of data exploration are generally varied and complex (Sinclair, 2009). They are dependent on formal statistical models employed in the analysis of data. However, with the penetration of artificial intelligence (AI) and its impact deeply felt in most organisational activities, data science is readily adopting AI for making smart decisions based on advanced technologies. Data science and AI are also finding their usefulness in crisis situations and in disaster management activities, in predicting weather patterns, cyclones and tornadoes, and Tsunami warning systems (Cao, 2023).

Data science involves the collaborative expertise of experts from various domains of learning, often involving the cross-disciplinary engagement of specialists who are capable of handling data and directing analysis, see, for instance, Nasution et al. (2023), Igual & Seguí (2024). They mostly adopt methodological approaches to the analysis of data. But the first thing that goes into the process is the careful inspection of the data. The analyst will be faced, if not with difficult, but specific problems involving big data handling[4]. What can one expect from data? What line of approach is best suited to conduct a specific test on data? How do you draw hypotheses, and which model will best fit the data? Here, some degree of intuition instilled into the process would help the analyst go a long way in serving the purpose of analysis. It is the core activities of the data analytic process that demand maximum attention from both the analytical and analyst's points of view. We shall delineate some commonly used methodologies and approaches used by researchers for data analysis that comprise functional models as statistical tests generally applied to data to perform analysis. Several models of data analysis could be outlined that have relevance to the different methodologies used for the purpose.

[4] Readers may wish to refer to Gartner IT Glossary on Big Data. http://www.gartner.com/it-glossary/big-data/

The renowned Austrian Philosopher Paul Feyerabands[5] always stressed the importance of adding depth and zest to an analysis. According to him, energetic and stimulatory exchange of ideas sparks thought that compels one to seek the grounds on which one is seeking something. In data science, too, this seems relevant, as to back any claim, one must be able to provide evidence (data/facts) based on strong grounds. Though he held less belief in methods and was a staunch opponent of fixed universals (methods) in science, he did not fail to claim his strong belief in reason and evidence. One of the topics that was being debated in his book "Three Dialogues on Knowledge" contends that "there cannot be any knowledge without statistics", in sociology, per se. Statistics imply "data."

Figure 2. Data types



Now, we come to discuss iterative cyclical data analysis, wherein data is analysed in a continuous loop. Business analytics utilises this kind of process where one can draw insights from each loop that help inform the next process within the cycle. Again, in the iterative nonlinear exploratory method of data analysis, data does not follow a linear path; i.e., it branches out, often forming loops to revisit previous steps. Different models appropriate for the analytical process could be adopted based on the nature of the analytic objectives. The entire process is reliant on a multi-step exploratory model chosen to fit the data. The analyst, however, must be able to evaluate and assess the relevance of the model used in analysing the data. Whenever necessary, parameters can be modified based on the results derived, and assumptions can be re-evaluated or refined as necessary. Following these, validation and interpretation are crucial steps in the data analysis process. Analysts must ensure that the model and the parameters are robust, reliable, and relevant to the context in question (Spector et al., 2022). Documentation and reporting would follow only if all these processes met the criteria and objectives of the entire process. Insights and recommendations can only be drawn if the rationale behind them remains sound. The results must, however, must be interpreted in light of the context of the original objectives (Rizk & Elragal, 2020).

3. Heuristics of Data Analysis

The science of data analysis is reliant on a system of logic based on the discovery of patterns, with the aim of linking causes to their effects, or simply, the discovery of causes. The procedures that define the heuristics of data analysis are not bound by a fixed set of rules that cannot vary. Analysis is generally performed on a set of data collected from observations of events. In other words, this is synonymous with the collection of evidence from

[5] See Feyerband's (1991). *Three Dialogues on Knowledge*, John Wiley and Sons.

observation. A heuristic in data analysis is a method that relies on an organised, coherent framework based on protocols or guidelines, where rules are defined but usually vary, to be used to bring maximum efficiency to the process (Dzemyda & Sakalauskas, 2011). Heuristics is used in problem-solving when estimation and modelling problems arise that become unstable when large data sets do not fit a model (Dzemyda & Sakalauskas, 2011). But they often turn out to be highly effective in solving problems when conventional and traditional methods simply do not yield satisfactory results from data analysis.

The meaning is couched in numbers that must be *crunched*, *munched*, and *chewed* by analysts using the logic of data analysis. This is performed to find the "link" in the connecting chain of evidence (data). Data is firmly correlated with events, where the correlation between data and effect (cause and effect) can only be positive or negative. It involves looking intently into the data to afford us an understanding of the subtle nature of patterns hidden inside the data. The methods employed in the discovery process may be varied and take the help of statistical models to support the claims of hypotheses thus constructed. It also involves a thorough search for causes and effects that underlie patterns hidden deep within the pockets of data. For one who is practically versed in data analysis, opportunities seem abundant for deeper exploration of the process. They often try different models or versions of the same model with slight modifications in order to obtain a variety of results. Expert analysts are able to discover new models that have not yet been tried and apply them to data. Expertise in handling data, therefore, also enables analysts to comprehend the abstruse, complex properties of the clusters of data that demand high technical competency in working with complex models.

Our best instruments are wisdom and reason. Reason promotes understanding. But it is a fact that data analysts employ a varied assortment of software tools to analyse data and then visualise the results in graphic format. Infographics make it easier to grasp the results of statistical data analysis. Software and statistical tools add to the wisdom of analysis, thus rendering it possible for the analyst to understand what the data means to them, what actually informs. The baffling tangle of relations and interrelationships within data among variables is often brought to light using core statistical tools. Descriptive statistics, exploratory data analysis, and regression models, along with many other statistical tests, aid in the diagnosis of causes and the discovery of correlations among variables. In effect, during the entire analytic process, data undergoes critical examination. It is checked for hidden patterns using infographics and data visualisation software (Sadiku et al., 2016). But when the primary purposes are prediction and forecasting, different models are employed to test the data for trends, and forecasting models are used to reveal future drifts.

> *Whenever data is obtained, it must be analysed. Raw data has little value to the naked eye. Measurements characterise the quantitative studies of phenomena under consideration (Brandt, 2014).*

Now, it is understood that the evidence for science is observation and experimentation. Researchers predict observations - *probability* calculations (Saha, 2003). Observations create either continuous or discrete data, from which information can be inferred[6]. The effects are primarily produced by means of instruments and tools, wisdom, and intellect. It might often be the case where data aids in the construction of propositions, i.e., axioms can be collected from facts by an inductive process. It is true that effects are discovered due to experiments and observations, but data can, and oftentimes it does, help, in deciphering the causes from the study of effects. For example, to conduct a bibliometric study, data is necessary, for without data, no conclusion can be reached that could give meaning to bibliometric research.

Researchers work with models that most likely fit the data obtained from experiments and observations. The nature of data determines those models that are likely to be applied to it to conduct analysis, i.e., how much randomness is there in the data, what's the nature of data distribution—normal or Gaussian, Bayesian or Poisson distribution, beta distribution or lognormal, etc., among other characteristics that get revealed during the process. The nature of the distribution depends on the variance of the individual elements within the data. It is often the case

[6] This is not to be confused with discrete and continuous random variables.

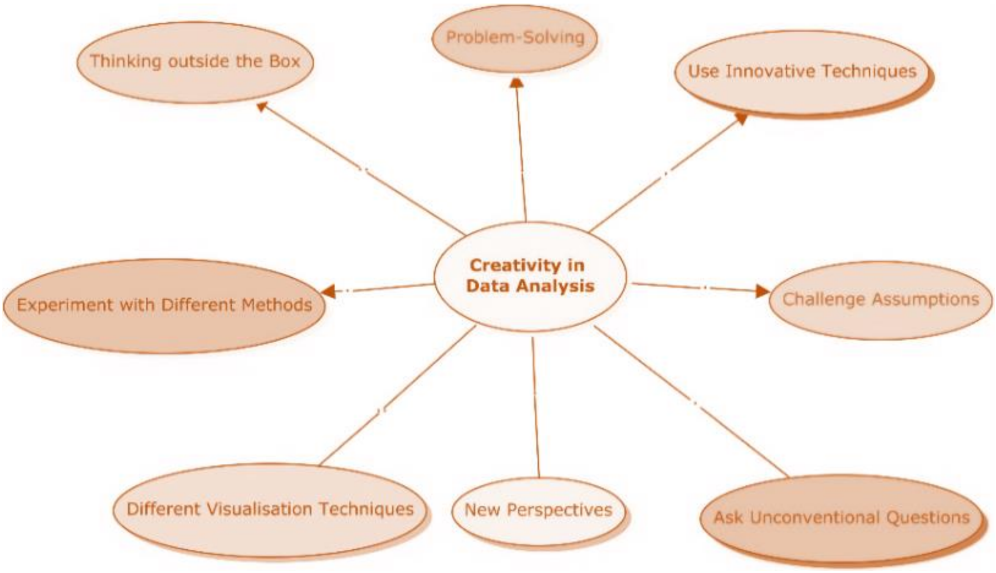
that the analysts do not collect the data themselves, but they gain access to data from various sources, e.g., web portals, online databases, periodicals, books, or from paid subscriptions through corporate databases. There are large multinational companies that deal with data as well. They collect, curate, organise, store, and sell data to their clients for a price.

How does one obtain data? One obtains data from measurements done on systems, products, operations, processes, and quality control assessments. Today, data ranges by many orders of magnitude, as most data is obtained from scientific sources, material sciences, biological and chemical experiments, image analysis, textual analyses, and data extraction using machine languages (Olivetti et al., 2020). Material physics, satellite-generated data, and Genome sequencing data are among the most abundant sources of scientific data that one could access. Data from experimental outcomes and results are regularly published, uploaded, and updated, and are shared by data portals being hosted by governmental agencies and organizations. Large centralised datasets exist in the NIH portals, Genome Databases, and the EMBL, PMC database, Medline, among others, from which they can be seamlessly downloaded to conduct analysis. The question is, in the abundance of data, what functions do they serve? They serve many functions. Using data, one can perform a diverse array of qualitative analyses, from semantic text mining and analytics to textual information extraction, rules-based heuristic data analysis, text summarization, biomedical text mining (Olivetti et al., 2020), parsing, content search and keyword analysis, word classification, etc. Apart from these, the primary function of data is to serve the purpose of quantitative analysis, statistical data analysis, exploratory data analysis, descriptive statistics, and financial and accounting analysis, among others. These are among the most important functions that one can perform using the collected data (Figure 1 above).

4. Creativity in Data Analysis

Art rekindles our will to create. And creativity may require the infusion of some degree of imagination into the process. In fact, according to Iacopini et al., (2018), creativity and innovation are the driving forces that underlie our economic growth and social development. Indeed, as Johnston (2014) states, vast amount of data are being collected, curated, and archived by researchers and organizations around the world - leading to data explosion. With increasing technological progress and product discovery powered by innovation, we have experienced explosive growth in both primary and secondary data and information. Supported by a buoyant economic growth in knowledge society, the utilisation of data has increased over time. The use of secondary data has increased as well, as it offers the advantage of knowing more about a topic through reviewing secondary sources. Data analysis - although being an intellectual activity, it does require some amount of *intuition* to foster innovation in data analysis (Sanger, 2006). According to Sanger (2006), there are seven types of creativity in data science from where insights can be derived, see Figure 3 below. These include among others thinking outside the box, problem solving, use of innovative tools (advance software) and visualisation techniques, and experimentation with new models.

Figure 3. Creativity in data analysis


Creativity in Data Analysis
Thinking outside the Box
Problem-Solving
Use Innovative Techniques
Experiment with Different Methods
Challenge Assumptions
Different Visualisation Techniques
New Perspectives
Ask Unconventional Questions

Some experts believe that data science is both art and science (Sanger, 2006; Peng and Matsui, 2015). With data, we can describe differences in reality and among things. We can say something about the behaviours and preferences of a group of individuals, things, or entities. At least, if we obtain data with care following scientific approaches, we can say something about things using numbers. In fact, data analyses serve different purposes depending on the needs of the inquirer. It is for this reason that data contains information and hidden patterns that require interpretation. Interpretation of data is, too, an art. The past is also captured in the data. The past records of numbers do provide clues to the future: With data supporting evidence, one can search for the specific cause(s) of an effect, for the causes can be identifiable from analysing the data. Cause gives rise to effect, and if there is an effect identifiable in the data analysed, there must be some causes behind it or something to have given rise to such an effect. This "something" is the cause. There also exists the lore of possibilities that seem probable when they present themselves as possibilities for a second time in future.

Data analysis increases the power of prediction and forecasting. At least, we have something (past data) at hand when we have data with which we can visualise the future. Now we call data a science. Why do we do so? It is because science is a particular approach to learning and an attitude towards understanding reality. Data science involves specific methods of analysis based on statistical models designed to peek into the depth of the data (Martinez, Viles & Olaizola, 2021). Particular methods of investigation determine the effectiveness of results obtained from an inquiry or analysis based on "facts" (Brandt, 1976). And here, there exist enough grounds and spaces for an analyst to be creative in her approaches or attitudes to data analysis—using big or small data (Dahlstedt, 2019). How? Let us examine.

Analysts are creative in their analysis when they develop affectivity towards the work to which they have been dedicated. One may ask, what is the nature of skill required for data analysis work? The practice of skill development allied to the field of scientific data analysis is an emerging field booming with prospects.

Affectivity and dedication transform an individual into a creative and passionate being. Here, the analyst assumes, quite correctly, that our past is what makes us what we are today. Hence, historical data is important for us to know the past in order to understand the present and, therefore, to be able to predict the future. With data at hand, analysts can not only predict, but also refute "Biased opinions", or oppose delusions. With a deep sense of dedication and affection towards their job, analysts can surprise us with their results. Historical data becomes critical when they use the past as a lens through which they view the present to understand and inform us about their predictions of the future.

To explore complex data, analysts use advanced visualisation techniques to identify historical trends and patterns that guide organisations to make informed decisions. There is power in predictive insights if they could be obtained as infographic visualisations that showcase interconnected patterns and identify relationships that are complex to depict human behaviours in the realms of social, economic, and cultural trends.

Data interpretation is an art (Peng and Matsui, 2015). How it is to be interpreted depends on the approach and methods the analyst has employed. Moreover, it depends on the analyst's ability to understand what the data says. Truth is obtainable from data, as facts refute illusions, delusions, and misconceptions. It depends a lot on the analyst's skills and how he or she is able to interpret what the data means. Special skills necessitate data analysis and interpretation. Analysts who are skilled in analysing data are equally adept at interpreting the results. The value of expertise is apparent when effort is materialized; i.e., made to find patterns in data. The importance of ratios is immense in the field of data analysis, for most relations between things are understood using ratios. Finding correlation among variables in data is also an important task that must be mastered for interrelating thoughts from data. Use of visual mapping tools (Ali et al., 2016; Healy, 2018) and advanced graphic visualisation software (Healy, 2018) are creative instruments that appeal to the eye of the investigator, where results are generated and interpreted in graphical formats for easier comprehension.

5. Quality Issues in Data Science and Data Analysis

Thanks to computers and the internet, we now deal with large amounts of data stored in *databases*, which are brimming with information at our fingertips awaiting to be tapped, exploited, modified, and shared. However, data quality is an important factor to consider (Wang, Reddy, & Kon, 1992). Some of these databases are not error-free, as they contain data that may be corrupt, incomplete, or faulty. It may result from errors in data collection and input that might have cropped up during the data entry process, or missing elements overlooked or ignored by data entry operators during the initial phase of collection, which, too, are counted as errors in data collection. In this respect, Ballou and Tayi (1999) offers a conceptual framework for enhancing the quality of data and its storage in data warehousing environments. Since many important decisions are nowadays driven by inferences drawn from data analysis, data quality is an important issue that must be addressed with care in order to derive potential benefits out of data.

Also, owing to more availability of data - and more importantly - big data, the issue of data quality and its usage has taken the centre stage of debate and discussion (Liu et al. 2015), since errors may be observed in the data collected before it is subjected to analysis. Here, too, data quality is a major issue with big data. It concerns reliability, consistency, and authenticity of the data collected (Liu et al. 2015). Different models - including 3V, 4V and 5V are emerging in an attempt to redefine big data. These V's refer to volume, variety and velocity, among others respectively, in relation to the explosion in big data (Laney, 2001). More than that, an exploration of 10 Vs within the context of psychological research provided a thorough understanding of the big data background in this area, shaping research methodologies and complete understandings of complex psychological phenomena (Nicola-Gavrilă & Dincă, 2023).

Nevertheless, despite such minor hitches and glitches, there appears to be a growing sense among data analysts regarding the quality of the data they have been working with. But it must also be acknowledged that it would be difficult to achieve *zero defect data*,[7] which is inconceivable given the large gamut of data produced and which is filling the ever-burgeoning databases, constraining resources and increasing costs to curate and manage such data. Nevertheless, given all such difficulties, there are different aspects and dimensions of data quality that the reader could refer to, to gain a formal understanding of the assessment of quality issues in data collection, storage, entry, and analysis (see Wang, Reddy & Kon,1992).

[7] A concept analogous to zero defect product discussed in the seminal paper by Wang, Reddy, and Kon, (1992).

6. The Science of Data Analysis: Methods and Principles

The scientific part that plays a pivotal role in data analysis is the use of statistical and formal models in analysing data and the methodology being used to add depth to an analysis. Data is counted as evidence, as it provides the grounds based on what one is seeking. When we have conceived a theory, we should be able to test and validate its claims based on evidence and data (Rizk & Elragal, 2020). If we have data backing up claims, such theories would prove to be powerful instruments to create and spread knowledge as it is understood. According to W.V. Quine, the renowned philosopher and thinker, a thesis built on scepticism has little value unless proved otherwise. However, in scientific research, hypotheses are often constructed upon a high degree of uncertainty as proposals intended to explain certain facts or phenomena, which constitute tentative insights.

> *W.V. Quine states that,*
>
> *“It is within science itself, and not in some prior philosophy, that reality is to be identified and described.”*
>
> *— W.V.O. Quine, Theories and Things*

When you have evidence but no numbers, it becomes difficult to conduct statistical tests because, to perform statistical analysis, you need numbers (data). The question remains, though, how successfully you can analyse data. This requires knowledge of the correct methods. Now, knowledge of methods is different from that of principles, but both have immense value to the analyst in order for them to derive coherent results from their analyses. Both methods and principles of data analysis fall within the ambit of scientific practices.

6.1. How Methods differ from the Principles of Data Analysis?

Methods imply techniques of data analysis. Principles govern the rules of analysis. In data science, there can hardly be any knowledge without the application of statistics. A skilled statistician is an able handler of data, and so is an efficient data analyst who must have some degree of proficiency in conducting data analysis employing statistical methods, tools, and techniques. Various tests are performed on big data after it is fed into an Excel spreadsheet or software that supports analysis and visualisation of results (Ali, et al., 2016; Healy, 2018), for example, using free data visualisation tools like Gephi, VOSviewer (for bibliometric network visualisations). Commonly used data visualisation techniques include different kinds of charts (pie, line, scatterplot, bar chart, etc.), histograms, and other advanced viewing techniques, among others (Myatt and Johnson, 2009; Sadiku, 2016).

Data entered into the system is subjected to various statistical tests performed by an analyst, where specific methods are employed to study and examine it. Models are designed using variables and parameters that define the quantities incorporated into the analysis. Regarding the availability of data, there are quite a large number of online databases that store data, and can be accessed freely. Data acquired from trustworthy databases or downloaded by the user to conduct analysis is first checked for accuracy, omissions, errors, and missing values.

A particular methodology is generally followed based on standard protocols and established procedures for the statistical and scientific evaluation of the data (Foster, Rzhetsky & Evans, 2015). Metadata used in the analysis is obtained from online databases, and results are analysed using parameters defined for the study. The strategy and procedure defined for the study constitute the ‘rules or principles’ of data analysis (Foster, Rzhetsky & Evans, 2015; Sinclair et al., 2009). Several simple tests, like, for example, ANOVA, probability distribution, Student’s t-test, regression analysis, and correlation coefficient analysis are among the most commonly performed tests on data, and the descriptive statistics that are obtained from the analysis also contribute greatly to the success of the research in question. Apart from this, the data evaluation process comprises an essential practical aspect of the entire process. Data evaluation increases the social value of data, as it validates the quality and examines the sources to see whether the data is obtained using standard protocols. If the data is secondary, reliability is checked concerning its primary source. All these involve the expenditure of energy and effort. Effort is represented by values. A combination of special skillsets and capabilities creates competent data analysts.

6.2. What Philosophy has to offer to Data Science?

Now, one may ask a paradoxical question: What can we learn about *data science* from ancient philosophers and their philosophies? There couldn't be a simple answer to this question, but if we consider ancient values as a basis for learning from them, then data science can learn valuable lessons from the. This are more likely related to principles of ethics, critical thinking (Miranda-Saavedra, 2022), and logical reasoning and morality, which could guide analysts in their data analysis to remain transparent and get the truth out for social and collective benefits. Falsifying of data is as immoral act, and it is here fairness and ethical standards should play a crucial role for the data analyst. Data transparency is also an important thing that is attached to public accountability and representation of truth. Data fabrication has become a cornerstone of unethical practice as in misrepresenting of facts and cooked data that belong to nowhere but to an analyst's dishonest effort.

> *"Order and falsehood cannot subsist together."*
>
> *—Thomas Carlyle, Lectures on Heroes*

Ancient philosophy has something more to teach modern data science, e.g., continuous learning, critical thinking, and undaunted effort that data analysts could incorporate in their practice of analysis, which could enhance decision-making and derive data-driven solutions. In simple terms, data scientists and data analysts should work with a thoughtful mind and add depth to their understanding of data. From an ethical perspective, data collection and analysis should adopt standard models that can enhance an analyst's ability to interpret and draw deeper insights from the data that has been analysed. Data analysts should handle data responsibly, and should ensure that the models employed insofar in analysing data are sound and appropriate for the data. This factor depends more on the ability and judgment of the analyst in helping others make informed choices based on the honest implications of the results of data analysis. Data scientists and analysts must place a strong emphasis on the ethical and moral sides of their work that may have important implications for the society. This is to ensure fairness and avoid biases in representing their results. When analysing data, analysts should be able to see the interconnectedness and interrelatedness of all things, and how they might affect the dependent variable. This is to derive a broader context and present a holistic perspective on the data being analysed.

6.3. Limitations of Big Data and Data Science

Today, we rely too much on data, for we have few other means of predicting or forecasting the future. But we must remember that there is knowledge beyond data—true, accurate knowledge, which can never be accurately predicted by analysing data alone. This, however, does not raise questions about the effectiveness of data, but points to the utility of formulas - mathematical or physics, that neatly explain so much of physics using one single formula, $e=mc^2$ discovered by Albert Einstein (Wigner, 1960; Halevy et al., 2009).

Data science and data analysis are lucrative fields that create jobs for prospective candidates. The implications and advantages of "big data" are ostensible and not clearly discernible, but there is far better knowledge beyond big data, the true scientific, epistemic, and transcendental knowledge of principles and processes that go far beyond the realms of modern data science. This notion, however, is contended by Haig (2020) and others who see big data as a support machinery to decision making, and as a data-intensive science in the aid of our understanding the nature of things using facts. According to Kar et al. (2023), big data is a handy tool for data-driven theory building guided by the principles and philosophy of data mining where data is used in the development of theory that tend to be more robust. A positive and pragmatic approach to big data along with recent advances in this domain has been highlighted by Shi (2022).

Data science is nothing but an accessory tool in the hands of an analyst who can get some meaning out of data. What has already happened could be known from facts and data, but what is likely to happen is difficult to predict with precision. Even the best models of probability using big data have failed to predict any major event so far with accuracy. Hence, the inference that big data is a big success is still a half-baked truth.

Nevertheless, data analysis is a highly scientific and rigorous field that involves statistical and mathematical models to analyse existing data (secondary data) and technical tools to visualise and interpret results. Models play important roles in data science as they are enablers for managing and using data (van Gils, 2023). It incorporates machine-learning techniques, deep learning, neural network models, generative models, and algorithms that help extract meaning, information, and knowledge from large datasets. Data science drives decision-making, growth, and helps optimise processes and operations across industries and businesses. Insights on customer behaviour are seamlessly derived from the interpretation of data, often showing certain trends in their behaviours and revealing patterns that help refine business decisions (Albright et al., 2011; Albright and Winston, 2020).

Marketing intelligence and market analysis are indispensable tools in the hands of business leaders to streamline decisions and make informed choices. Also, researchers across many fields use data to draw conclusions from their analyses of experimental results to assess outcomes. It makes possible to draw precise conclusions supported by data based on observed and experimental evidence that often helps support or contradict previously open hypotheses. Therefore, utilising existing data for research and analysis has become more prevalent (Johnston, 2014). It is also the basis for secondary data analysis to conduct inquiries using existing data collected from primary sources (Smith et al., 2011).

6.4. The Question of 'How' and 'What' of Data Analysis

Analysts often have to deal with massive data, which is objective, sometimes unstructured, and of a complex nature. For example, take the case of bibliometric analysis. Bibliometrics is fast extending to all disciplines, and is now used by businesses as well. Beyond science mapping and the emphasis on empirical contributions related to different scientific fields, it has spread to other domains, including business and management science, education, information science, and others. It also helps keep track of burgeoning academic publications, which are increasing at a rapid pace (Aria & Cuccurullo, 2017).

In bibliometric analysis, a researcher's primary aim is to interpret and inform the readers about certain trends, e.g., the number of cumulative citations, publications, H-index, and i-10 index, which place data at the centre stage of analysis (Donthu et al., 2021). Most often, rigorous methods are applied to test bibliometric data. Beyond these, bibliometric analyses are conducted to assess trends within the publishing industry, including the contribution of an author or groups of authors to a particular domain of study, their collaboration patterns, article performance analysis, etc. To measure the growth of a scientific field, for example, one needs data. This is exactly what Bormann and Mutz (2015) have done to measure the rate of growth of scientific domains over a long period of time based on bibliometric analysis. They have conducted the research with the help of two different sets of data gathered on the number of publications and cited references. Several other landmark studies concerning those conducted by Price (1965) and Tabah (1999) have previously used different methodologies and models to study the growth and progress of knowledge in specific domains - all using data based on the concept of "literature dynamics." All these reflect important information concerning assessment of the progress of diverse fields of study, understanding research patterns, and output analysis of top-performing authors (as well as individual academic departments) and experts in their respective fields, science and the arts. The question of how such studies is conducted depends on the techniques and methods employed by the researchers and analysts. Interested readers may find detailed step-by-step information concerning these in a recent thought-provoking article (Donthu et al., 2021).

Now, there is always a cost and effort attached to the entire process of data analysis, right from the initial steps concerning the collection of data to the reporting of results and interpretations of the data analysed. Then there is productive effort that goes into the analysis of the data. Effort is embodied in value: The value of a data analysis process depends on how effectively the data has been examined. It also depends on the analytic acumen and expertise of the analysts conducting data analysis. Of course, researchers and data analysts can augment their skills, strategies, and analytic acumen for better productivity in terms of success in data analysis (Foster, Rzhetsky & Evans, 2015). These constitute the necessary factors in the creation of value from data analysis.

Conclusions

In this paper on the what and how of data analysis, we discuss the methods and techniques that analysts most frequently employ while analysing data. Some discussions have taken place on the notion of creativity in data analysis, whereas, in the background, the paper reflects on the idea of efficient analysis through competence building and strategic use of intuition, heuristics and data visualisation techniques to augment the processes of analysis. Some philosophical aspects of data analysis have been discussed as well, concerning the role of knowledge (epistemology) of analytics in data science. It must be acknowledged that data science, including its sub discipline Big Data, have become indispensable tools for businesses and organizations on one hand, and for the managers and decision-makers on the other.

Credit Authorship Contribution Statement

The corresponding author, Sidharta Chatterjee has conceived the idea and contributed to writing the research paper, and drafted it thus performing the entire work.

Acknowledgments

Acknowledgements are due to the Central Library staff of Andhra University, Visakhapatnam. The author expresses his gratitude to some of his colleagues with whom fruitful discussions have led to the conception of ideas useful to this research.

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

[1] Agresti, A. (2012). Categorical data analysis, *Volume 792 din Wiley Series in Probability and Statistics*, John Wiley & Sons, 752 pp. ISBN: 978-0470463635

[2] Albright, S. C., & Winston, W. L. (2020). Business Analytics: Data Analysis and Decision Making. 7th Edition, Cengage Learning, Inc. ISBN: 978-0357109953

[3] Albright, S. C., Winston, W. L., Zappe, C. J., & Broadie, M. N. (2011). Data Analysis and Decision Making, (Volume 577). South-Western/Cengage Learning. https://www.wu.ac.at/fileadmin/wu/d/i/ifr/Data_Analysis_and_Decision_Making.pdf

[4] Ali, S. M, Noopur, G., Gopal, K. N., & Rakesh, K. L. (2016). Big data visualization: Tools and challenges, In: 2nd *IEEE International Conference on Contemporary Computing and Informatics (IC3I)*, pp. 656-660. https://doi.org/10.1109/IC3I.2016.7918044

[5] Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4), 959-975. https://doi.org/10.1016/j.joi.2017.08.007

[6] Ballou, D. P., & Tayi, G. K. (1999). Enhancing data quality in data warehouse environments. *Communications of the ACM*, *42*(1), 73-78. http://doi.org/10.1145/291469.291471

[7] Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11), 2215-2222. https://doi.org/10.1002/asi.23329

[8] Brandt, S. (1976). Statistical and computational methods in data analysis (No. 04). Amsterdam, The Netherlands: North-Holland Publishing Company. https://doi.org/10.1119/1.1986393

[9] Cao, L. (2023). AI and data science for smart emergency, crisis and disaster resilience. *International Journal of Data Science and Analytics,* 15(3), 231-246. https://doi.org/10.1007/s41060-023-00393-w

[10] Carlyle, T. (1910). Lectures on heroes: Hero-worship and the heroic in history. Clarendon Press. https://www.gutenberg.org/files/1091/1091-h/1091-h

[11] Carpineto, C., & Romano, G. (2004). *Concept Data Analysis: Theory and Applications*. John Wiley & Sons. https://DOI:10.1002/0470011297

[12] Dahlstedt, P. (2019). Big data and creativity. *European Review*, 27(3), 411-439. https://doi:10.1017/S1062798719000073

[13] Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285-296. https://doi.org/10.1016/j.jbusres.2021.04.070

[14] Dzemyda, G., & Sakalauskas, L. (2011). Large-scale data analysis using heuristic methods. *Informatica*, 22(1), 1-10. https://doi.org/10.15388/Informatica.2011.310

[15] Feyerabend, P. K. (1991). Three dialogues on knowledge. John Wiley & Sons. ISBN: 978-0-631-17918-4

[16] Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, 80(5), 87 5-908. https://doi.org/10.1177/0003122415601618

[17] Garfield, E. (1980). Citation indexing. *Journal of Information Science*, 2(1), 47-47. https://doi.org/10.1177/016555158000200109

[18] Haig, B. D. (2020). Big data science: A philosophy of science perspective. In: *Big Data in Psychological Research,* (pp. 15-33). American Psychological Association. https://psycnet.apa.org/doi/10.1037/0000193-002

[19] Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), 8-12. http://doi:10.1109/MIS.2009.36

[20] Healy, K. (2018). *Data Visualization: A Practical Introduction*. Princeton University Press.

[21] Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). Applied survey data analysis. Chapman and Hall CRC. https://doi.org/10.1201/9781315153278

[22] Hsiao, C. (2022). Analysis of panel data (No. 64). Cambridge University Press. https://doi.org/10.1017/9781009057745.016

[23] Iacopini, I., Milojević, S., & Latora, V. (2018). Network dynamics of innovation processes. *Physical Review Letters,* 120(4), 048301. https://doi.org/10.1103/PhysRevLett.120.048301

[24] Igual, L., & Seguí, S. (2024). Introduction to data science. In: *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications* (pp. 1-4). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50017-1

[25] Johnston, M. P. (2014). Secondary data analysis: A method of which the time has come. *Qualitative and Quantitative Methods in Libraries*, 3(3), 619-626. https://qqml-journal.net/index.php/qqml/article/view/169

[26] Kar, A. K., Angelopoulos, S., & Rao, H. R. (2023). Big data-driven theory building: Philosophies, guiding principles, and common traps. International *Journal of Information Management*, 102661. https://doi.org/10.1016/j.ijinfomgt.2023.102661

[27] Khatri, N., & Ng, H. A. (2000). The role of intuition in strategic decision making. *Human Relations*, 53(1), 57-86. https://psycnet.apa.org/doi/10.1177/0018726700531004

[28] Kuhn, T. S. (1997). The structure of scientific revolutions (Vol. 962). Chicago: University of Chicago Press.

[29] Laney, D. (2001). 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, 6. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[30] Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 134-142. https://ui.adsabs.harvard.edu/link_gateway/2016JPRS..115..134L/doi:10.1016/j.isprsjprs.2015.11.006

[31] Martinez, I., Viles, E., & Olaizola, I. G. (2021). Data science methodologies: Current challenges and future approaches. *Big Data Research*, 24, 100183. https://doi.org/10.1016/j.bdr.2020.100183

[32] Miles, M. B., & Huberman, A. M. (1994). Qualitative Data analysis: An Expanded Sourcebook. Sage. https://vivauniversity.wordpress.com/wp-content/uploads/2013/11/milesandhuberman1994.pdf

[33] Miranda-Saavedra, D. (2022). How to Think about Data Science. CRC Press. https://doi.org/10.1201/b23197

[34] Murtagh, F., & Heck, A. (2012). Multivariate data analysis (Vol. 131). Springer Science & Business Media. https://doi.org/10.1007/978-94-009-3789-5

[35] Myatt, G. J., & Johnson, W. P. (2009). Making sense of data II: A practical guide to data visualization, advanced data mining methods, and applications. 1st Edition, John Wiley & Sons. ISBN: 978-0470222805

[36] Nasution, M. K., Syah, R., & Elveny, M. (2023). What is data science. In Data Science with Semantic Technologies (pp. 1-25). CRC Press. http://dx.doi.org/10.1201/9781003310785-1

[37] Nicola-Gavrilă, L., & Dincă, S. (2023). Future Interdisciplinary Combination of AI Technologies and Psychology. *Journal of Contemporary Approaches in Psychology and Psychotherapy*, 1(1). https://doi.org/10.57017/jcapp.v1.1.02

[38] Peng, R. D., & Matsui, E. (2015). The Art of Data Science: A guide for anyone who works with Data. Skybrude Consulting, LLC. https://ci.nii.ac.jp/ncid/BC16512258?l=en

[39] Price, D. J. D. (1965). Networks of scientific papers. *Science*, 149, 3683, 510-515. https://doi.org/10.1126/science.149.3683.510

[40] Quine, W. V. O. (1981). *Theories and Things.* Harvard University Press. ISBN: 978-0674879263

[41] Rizk, A., & Elragal, A. (2020). Data science: developing theoretical contributions in information systems via text analytics. *Journal of Big Data,* 7, 1-26. https://doi.org/10.1186/s40537-019-0280-6

[42] Sadiku, M., Shadare, A. E., Musa, S. M., Akujuobi, C. M., & Perry, R. (2016). Data visualization. *International Journal of Engineering Research and Advanced Technology*, 2(12), 11-16. http://doi.org/10.31695/IJERAT

[43] Saha, P. (2003). Principles of Data Analysis. Cappella Archive. https://www.physik.uzh.ch/~psaha/pda/pda-a4.pdf

[44] Sanger, J. (1994). Seven types of creativity: looking for insights in data analysis. *British Educational Research Journal*, 20(2), 175-185. http://dx.doi.org/10.1080/0141192940200203

[45] Sheard, J. (2018). Quantitative data analysis. In *Research Methods: Information, Systems, and Contexts*, 2nd Edition, Williamson, K. & Johanson, G. (Eds.), pp. 429-452, Chandos Publishing. ISBN 978-0081022207. https://doi.org/10.1016/B978-0-08-102220-7.00018-2

[46] Shi, Y. (2022). *Advances in Big Data Analytics. Theory, Algorithms and Practices*. eBook. ISBN 978-981-16-3607-3. https://doi.org/10.1007/978-981-16-3607-3

[47] Sinclair, M., Sadler-Smith, E., & Hodgkinson, G. P. (2009). The role of intuition in strategic decision making. In: *Handbook of Research on Strategy and Foresight*. Edward Elgar Publishing. https://doi.org/10.4337/9781848447271.00032

[48] Smith, A. K., Ayanian, J. Z., Covinsky, K. E., Landon, B. E., McCarthy, E. P., Wee, C. C., & Steinman, M. A. (2011). Conducting high-value secondary dataset analysis: An introductory guide and resources. *Journal of General Internal Medicine*, 26, 920-929. https://doi.org/10.1007/s11606-010-1621-5

[49] SMITH, C. M. (2000). Bioinformatics, genomics, and proteomics. *The Scientist*. https://www.the-scientist.com/bioinformatics-genomics-and-proteomics-55317

[50] Spector, A. Z., Norvig, P., Wiggins, C., & Wing, J. M. (2022). Data science in context: Foundations, challenges, opportunities. http://www.cambridge.org/9781009272209

[51] Spicer, J. (2005). Making Sense of Multivariate Data Analysis. Sage. ISBN: 978-1412904018. http://dx.doi.org/10.4135/9781412984904

[52] Tabah, A. N. (1999). Literature dynamics: studies on growth, diffusion, and epidemics. *Annual Review of Information Science and Technology*, 34, 249-286. https://www.learntechlib.org/p/92548/

[53] Tukey, J. W. (1977). *Exploratory Data Analysis*, 1st Edition, Pearson. ISBN 978-0201076165

[54] van Gils, B. (2023). *Data in Context: Models as Enablers for Managing and Using Data*, 1st Edition, Springer. 240 pp. ISBN 978-3031355387

[55] Wang, R. Y., Reddy, M. P., & Kon, H. B. (1995). Toward quality data: An attribute-based approach. *Decision Support Systems,* 13(3-4), 349-372. https://doi.org/10.1016/0167-9236(93)E0050-N

[56] Wasserman, L. (2018). Topological data analysis. *Annual Review of Statistics and Its Application*, 5, 501-532. https://doi.org/10.1146/annurev-statistics-031017-100045

Cite this article

Chatterjee, S. (2024). The What and How of Data Analysis. Journal of Research, Innovation and Technologies, Volume III, 1(5), 51-65. https://doi.org/10.57017/jorit.v3.1(5).04

Article's history:

Received 5th of May, 2024; Revised 29th of May, 2024;
Accepted for publication 4th of June, 2024; Available online: 7th of June, 2024
Published 30th of June, 2024 as article in Volume III, Issue 1(5)