www.ijselam.com editor@ijselam.com

International Journal of Science, Engineering, Law, Arts and Management (IJSELAM)

e-ISSN:XXXX - XXXX

VOLUME 1, ISSUE 1, MARCH 2026

Venu Jannigorla **Perigisetti Nagendra Murthy**

A Review of Data Science: What It Is, How It Works, and Where It's Headed

Venu Jannigorla
Department of Computer Science Engineering (Data Science),
Amrita Sai Institute of Science and Technology, Vijayawada, A.P., India.
Email: venujannigorla@gmail.com

Perigisetti Nagendra Murthy
Department of Computer Science Engineering (Data Science),
Amrita Sai Institute of Science and Technology, Vijayawada, A.P., India.
Email: nagendramurthyperigisetti@gmail.com

ABSTRACT

Data science, while its revolutionary influence on numerous industries and fields, has a fundamental problem: the fundamental absence of a singly accepted definition. Such definitional imprecision coupled with the still-emergent status of its core scientific foundations and severe educational shortcomings considerably discourages its full potential as a unified, recognized discipline. The reality of the present, which is dominated by fragmentary methods and over-reliance on isolated, stand-alone technological tools, sadly limits the full potential of data extraction and good knowledge creation. The current analytic systems of analysis and software packages, however able individually, are in practice usually pursued without a guiding theory. This leads to practical inconsistency, massive standardization difficulty, as well as difficulties in replicating research outcomes across various settings. This sequential growth also further exacerbates the definitional crisis and results in a fundamental lack of full data literacy among various professional and academic disciplines, limiting wider adoption and innovation. Our solution confronts these critical problems directly. This broad-based survey weaves diverse scholar and practitioner thinking into a cohesive narrative that not only illuminates data science's far-reaching and inescapable impacts—on the global labor market, acceleration of scientific inquiry, advancement of industrial innovation, and shaping of social progress—but also presents a clear, attainable path forward. We call for the scientific basis of the field to be supported by systematic theoretical effort and empirical validation. Furthermore, we make concrete recommendations for future studies that are designed to better define data science more precisely, universally accepted, and expansively. Above all, we emphasize the compelling necessity of closing the existing education gap, promoting practices that enhance data literacy and build the field's theoretical foundations at every education and professional development level. Through the focus on these interconnected solutions, this article presents a solid premise for overcoming the problems that currently confront data science, making way for data science to realize its profound, disruptive potential.

Keywords *- Data Science, Big Data, Data Skills, Mixed-Field Study, Education Problems, Ethical Data Use, Analytics, Machine Learning.*

I. Introduction

Data science is a fast-expanding, multidisciplinary field ubiquitous in business, industry, and science. Although usually associated with machine learning, its definition continues to evolve, covering processes ranging from data capture and cleaning to managing ethics and regulation. Debates continue on whether it's a nascent field or an outgrowth of statistics and computer science, with origins traceable back to the work of John W. Tukey in 1962.

The growth of the field is stimulated by exponential growths in data storage and processing, creating the "Big Data" phenomenon. The goal of data science is ultimately to extract value, enhancing current products and services or inventing new ones. Data science responds to questions of reporting (what), diagnosis (why), prediction (what's going to happen), and recommendation (what to do). The "Internet of Events" (IoE), which combines online content, human interaction, IoT devices, and location data, is a prime example of pervasive data creation.

This data explosion requires expert data professionals, calling for a spectrum of skills: hard (technical), soft (interpersonal), and analytical. Even though there's high demand, an enormous "data science education dilemma" remains, most notably in K-12 systems, where there is a disconnect between data literacy needs in society and existing educational shortcomings. Issues involve inadequate teacher training, lack of curriculum alignment, and few studies on learning data. Closing this gap is essential to ensure a data-literate population.

1.1. Motivation

The deep significance of data science lies in its unmatched potential to harvest knowledge and value from the vast amounts of data being produced in the "Big Data" and "Internet of Events" age. It takes raw, frequently disorganized data and turns it into actionable insights, which are crucial for businesses and industries to succeed competitively.

In addition, this review is prompted by the changing function of the data scientist, who is commonly described as the "engineer of the future" and needs a complicated mix of technical, soft, and analytical competencies. The ongoing scholarly and professional discourses over data science's disciplinary classification—whether a separate field or potent synthesis—also highlight the necessity of critical review. Its academic location is crucial for future expansion, finance, and pedagogical structures. This also includes social responsibility towards enhancing data literacy and fostering equal access to the opportunities of the data revolution.

1.2. Problem Statement

Although it has the potential to transform, data science is confronted with serious, interwoven challenges. One major challenge is the absence of a shared, general definition. Such definitional vagueness contributes to multiple interpretations, frustrating academic discussion, standard education, and uniform professional expectations.

The other problem of critical concern is the "data science education dilemma" that involves an increasing gap between data-literate workforce demand and insufficient K-12 data science instruction. Challenges include untrained instructors, weak curriculum infusion, and restricted pedagogical research. Policymaker actions are frequently ill-informed, with no evidence or complete comprehension of the intricacies of teaching. Incorporating data science into mainstream subjects is challenging; for example, math lessons will shun real-world applications, and science lessons will employ data illustratively instead of as a central investigative method. Social sciences, though quantitative in nature, have not extensively incorporated data comprehension into K-12. These problems hinder the growth of a data-aware population and restrict data science's social benefits and fair opportunities.

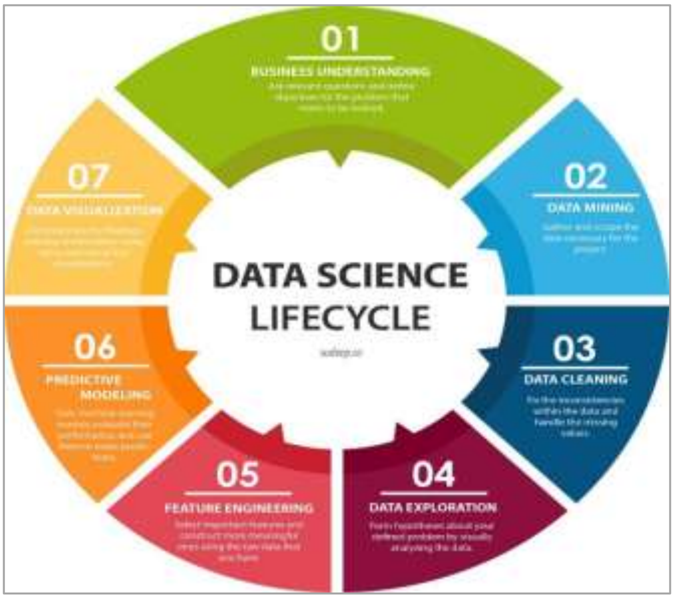
2. Literature Review

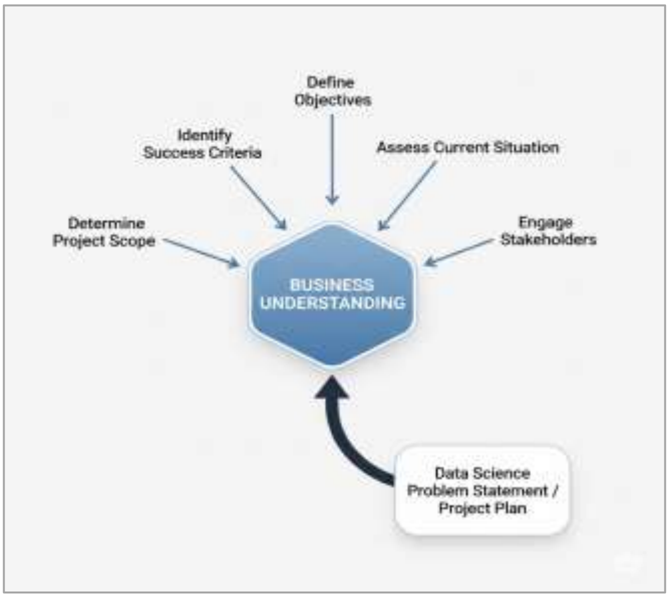
The aggressive hunt for data experts has turned into a central issue with the advent of big data, sparking debates on the complicated competence model needed for data scientists, encompassing hard, soft, and analytical competencies, with analytical thinking taking precedence because of the trans-disciplinary nature of data science [1]. Data Science is considered a great innovation of the early 21st century, born out of scientific discovery and used wherever there is enough data, with outstanding achievements, though its science basis is still emerging actively, distinct from its mature antecedent paradigms [2]. As the ability to store and process data has grown exponentially since the 1960s, organizations have come to appreciate more and more that clever data exploitation is a matter of survival, and so "Big Data" has become an issue for boards of directors and data science has evolved as a new, separate discipline, as has computer science out of mathematics [3]. One key question in the discipline is whether 'data science' is indeed a 'science' in the commonly accepted meaning, a question addressed through the use of known criteria for science applied to data science in contexts such as health professions education [4]. One notable and confounding conundrum illustrates the expanding disconnect between society's dependence on information and readiness on the part of students through data science education, especially K-12, where rich analytical experiences are not common, educators are not equipped with data expertise, and the field cannot see a natural place in current school subject frameworks [5]. In industry and commerce, data science is uniquely defined in that it involves prioritizing problems by their value added to the business, requiring that top performers be skilled in business drivers and also have core software engineering capabilities to create and install analytical solutions for use on an ongoing basis, a scope much wider than standard academic statistical problems [6]. One main reason why achieving a unified, consensus definition of data science is a challenge is that it has an intrinsic multidimensionality since it can at the same time be referred to as a science, a research methodology, a research approach, a discipline, a process, a-nd a career, with no definition doing justice to its varied nature [7]. Data science is a new trend marked by its strong growth and growing presence in nearly all areas of everyday life, leading innovations in data-powered customer experiences, e-commerce technology, and the creation of deep fakes, generative AI, and synthetic data, frequently at the intersection of disruptive technologies such as AI, IoT, cloud computing, and 5G [8].

3. Methodology

Data Science Lifecycle is a formal process that embodies the process of transforming raw data into meaningful findings and solutions based on data. It starts with the definition of the problem, where the objectives are well-defined through collaboration with the stakeholders. It then proceeds to the step of gathering data, where the data that is deemed relevant is extracted from resources like databases, APIs, or sensors. The second phase, data preprocessing and cleaning, ensures that quality data is achieved through handling missing values, outliers, and inconsistencies. Data analysis follows through the use of exploratory data analysis (EDA), which exposes hidden patterns and relationships within the data. Feature engineering and feature selection are subsequently conducted after obtaining insights to optimize model performance by constructing informative input variables. During the process of building the model, suitable machine learning algorithms are chosen and trained to tackle the given problem. The model is evaluated using metrics such as accuracy, precision, recall, and F1-score to identify how well the model performs. After successful evaluation, the model is implemented in a real environment through tools such as APIs or dashboards. Lastly, there is the monitoring and maintenance phase that guarantees optimal functioning of the model in the future in case of problems of data drift or model degradation. This cycle is a robust, reproducible, and scalable solution to challenging data problems.

The Data Science Lifecycle is a systematic, reproducible process that leads data professionals to the solutions for difficult problems with data-driven solutions. Every step within the lifecycle is reliant on the others and is of critical significance in order to obtain accurate, valid, and business-oriented results. The following is a detailed description of each step.





3.1. Business Understanding

This chart encapsulates the all-important first step of any data science project: Business Understanding. It shows that starting a successful data science project, or indeed any data-driven project, isn't with data, but with some proper understanding of business context and goals.

The chart shows some important inputs in arriving at this understanding:

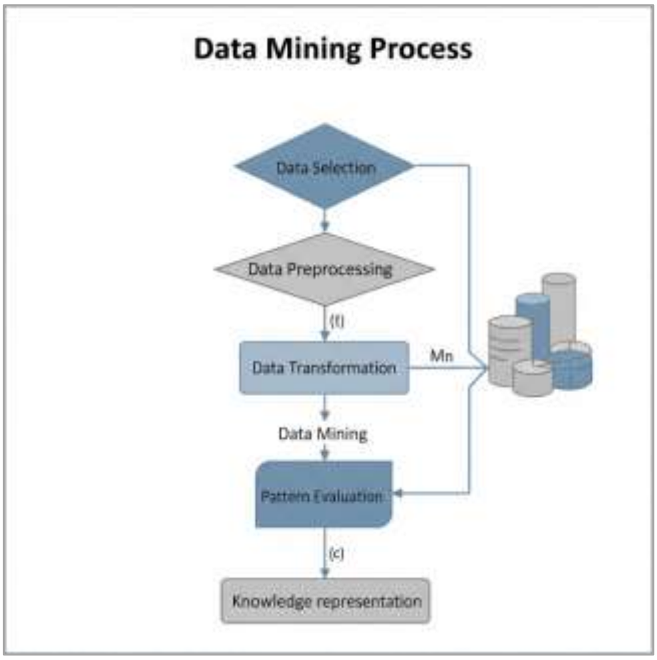
Define Objectives: Stating what the company desires to get from the data science project. These may be desired outcomes such as improved revenue, reduced cost, or improved customer satisfaction.

Define Success Criteria: Creating quantifiable indicators that will specify whether the project has succeeded or not. These are typically numbers.

Evaluate Current Project Situation: Creating present business processes and current pain points, accessible resources, and state of accessible data for the organization.

Define Project Scope: Establishing project scope, i.e., what is included and excluded, and any limits such as time or finances.

Involve Stakeholders: Having the appropriate individuals from the different business departments or technical teams engaged to obtain requirements, learn, and get aligned on project goals. The deliverable of this Business Understanding stage is a concise Data Science Problem Statement / Project Plan. The deliverable is an unambiguous roadmap setting out the business objectives and restates them in an understandable data science problem which can be solved in later phases of the project. It is used to guarantee the data science projects have a definite link with business value.



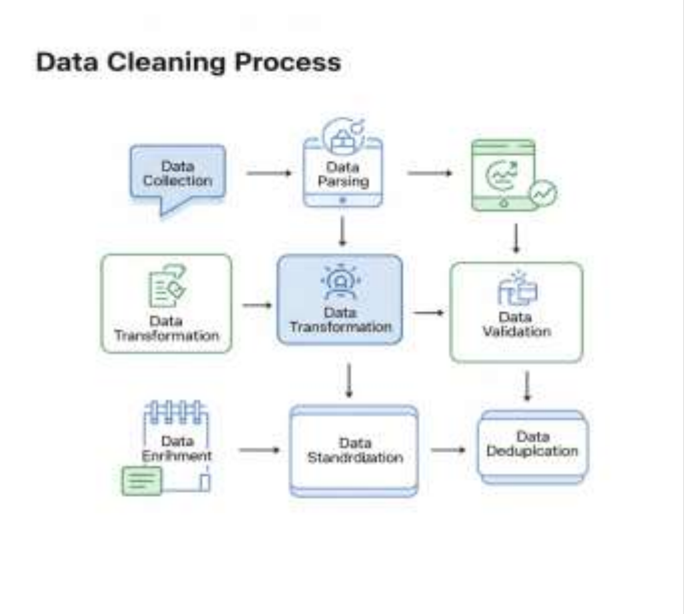
3.2. Data Mining Process

This figure, Figure 2: Data Mining Process, presents the most essential activities involved in uncovering valuable patterns and knowledge in gigantic data. Data mining is a branch of uncovering hidden knowledge that may be used in decision making in many industries.

The process, as indicated in the figure, generally involves:

- **Selection of Data:** The initial step is selecting the necessary data from larger databases or data warehouses associated with the data mining process.
- **Preprocessing of Data:** Raw data is usually noisy, incomplete, and inconsistent. Noise is eliminated from data in this step, missing values are handled, and inconsistencies are smoothed to improve the data quality.
- **Data Transformation:** Here, the preprocessed data is transformed into an appropriate format as input to the intended data mining algorithms. Normalization, aggregation, or extraction of new features may be performed here.
- **Data Mining:** The most sensitive step where the knowledge discovery methods are used to find patterns, rules, or conclusions from the data that has been transformed. Different methods such as classification, clustering, discovery of association rules, or regression may be performed here.
- **Pattern Evaluation:** Then, patterns that are derived from data mining models are assessed to determine actually interesting and valuable patterns that are actionable knowledge. It may utilize statistical measures or domain expertise.
- **Knowledge Representation:** Lastly, the knowledge that is extracted is represented in an understandable, interpretable, and usable format to the end-user. It may be in visualizations, reports, or incorporated systems.

This cyclical process ensures raw data are processed and analyzed in an orderly fashion to achieve important and effective intelligence.



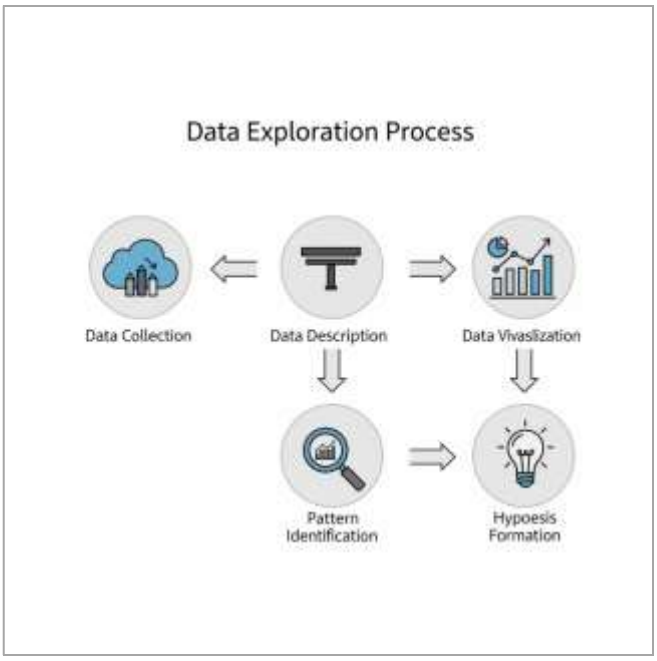
3.3. Data Cleaning Process

This figure, henceforth referred to as Figure 3: Data Cleaning Process, represents the step-by-step process of preparation of raw data in a manner that the data is suitable and trustworthy to be modeled or analyzed. Data cleaning is a pillar of any good data science pipeline and addresses errors that otherwise produce erroneous results or suboptimal model performance.

The process outlined normally follows the following steps:

- **Data Collection:** It is the first step where data from a variety of sources are collected and become the raw input to be cleaned.
- **Data Parsing:** The collected data is usually then parsed to transform the data into a structured form by decomposing raw text or complex forms into bite-size chunks.
- **Data Transformation:** In this step, data is transformed from raw or parsed form to consumable and standard form. It may be formatting, aggregation, or scaling.
- **Data Validation:** The validated data is then examined against predefined constraints or rules for detecting inconsistencies, errors, or missing values. It is an essential quality control process.
- **Data Standardization:** A technique for bringing data into compatibility in standard forms, units, and scales, particularly while merging data from heterogeneous sources.
- **Data Deduplication:** The identification and elimination of duplicate records are required in order to avoid redundant analysis as well as biased results, with each entity occurring only once.
- **Data Enrichment:** Lastly, data can be enriched by adding external dataset information at the back of the dataset, providing context or richness to the original dataset.

This process ensures the dataset used for further analysis or model building is correct and of high quality.



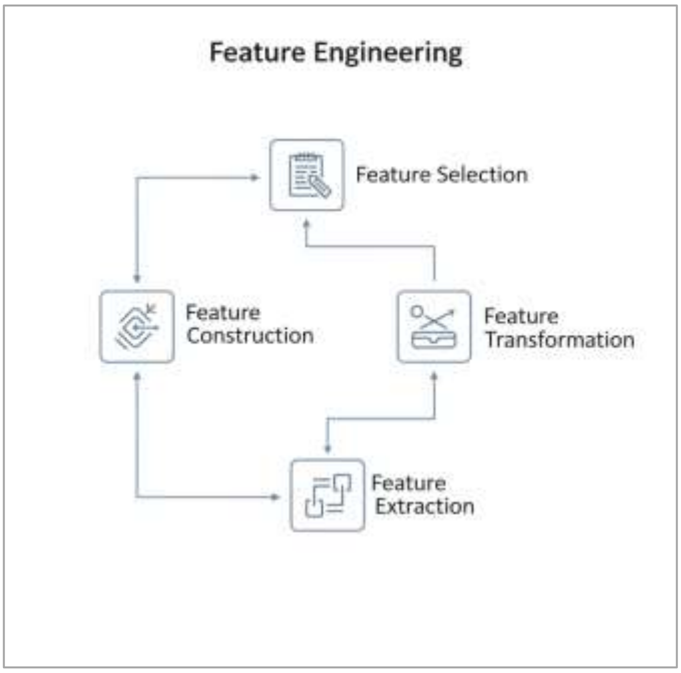
3.4. Data Exploration Process

This figure, or Figure 4: Data Exploration Process, displays the interactive and iterative process of learning and forming first impressions from a dataset. Data exploration is one of the essential starting processes in data analysis, wherein data scientists explore data for patterns, detect anomalies, and validate assumptions with the assistance of summary statistics and plots.

The process often includes:

- **Data Collection:** This is the first process by which raw data are collected from primary sources to provide the basis for explorations.
- **Data Description:** Data description involves the gathering of summary statistics and data attributes such as frequencies, means, medians, and ranges to have a preliminary idea about its characteristics.
- **Data Visualization:** A valuable method in which graphical presentations such as plots, charts, and graphs are utilized to establish trends, anomalies, and connections that are hard to identify from raw data or statistics.
- **Pattern Identification:** Data scientists search for trends, correlation, and anomalies by characterizing and visualizing the data set.
- **Hypothesis Development:** Informed hypotheses or educated guesses of the underlying phenomena or relationships in the data are constructed from the patterns and preliminary findings. They may be tested at subsequent steps of the data science process.

Arrows noted in the diagram represent iterative flow, which is of the kind that insight at a step could cause rerunning earlier steps for yet more insight.



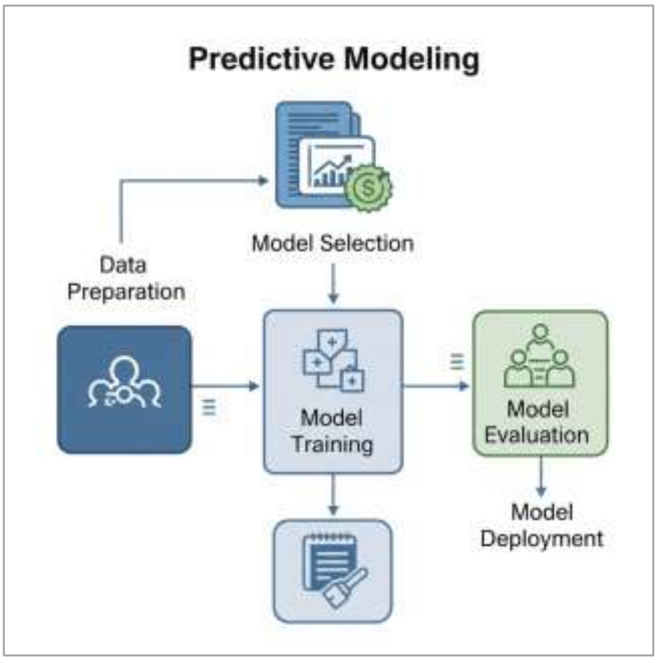
3.5.Feature Engineering

This is what can be referred to as Figure 5: Feature Engineering Process, and is a crucial step in pre-processing the data for supplying machine learning algorithms. Feature engineering is a procedure of domain knowledge application to deduce or extract new features from raw data such that machine learning algorithms can function with higher efficiency. It is typically an iterative procedure, as indicated by the feedback arrows in the figure.

The most important steps of Feature Engineering process, as illustrated, are:

- **Feature Selection:** Selecting the optimal features of the given dataset that are most benefited by the model performance and removing unwanted or redundant features.
- **Feature Construction:** Developing new features by manipulating or combining existing features in order to receive more useful representations of data. This is most commonly derived from domain knowledge.
- **Feature Transformation:** Performing mathematical computations or functions for the purpose of changing features to a more suitable format for the machine learning algorithm. Scaling, normalization, and log transformation are a few.
- **Feature Extraction:** It generally describes methods that decrease the dimensionality of a set of features in a data set by deriving new features of reduced dimension from the original features while keeping most of the information. Principal Component Analysis (PCA) is a well-known example.

The cyclical process in the figure illustrates that these stages don't necessarily come in sequence but can be practiced and iterated to further improve the feature set to optimize model performance.



3.6. Predictive Modeling

This slide, which can be titled Figure 6: Predictive Modeling Process, describes the methodical process of developing and using models that predict future outputs or trends based on past data. Predictive modeling is a key building block of nearly all uses of data science, ranging from customer activity prediction to predicting market trends.

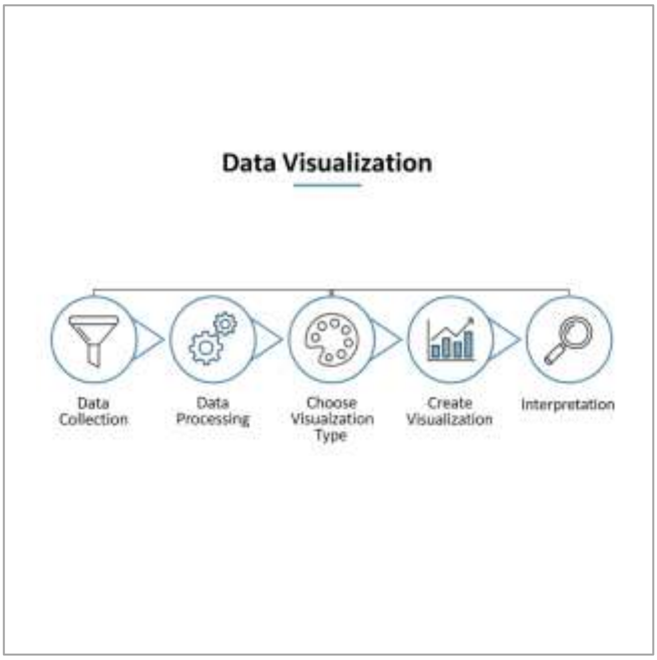
The process generally consists of the following steps:

- **Data Preprocessing:** This is the initial and most important step that is responsible for gathering, cleaning, converting, and putting the raw data into its best possible shape in order to train a model. It makes the data accurate, uniform, and prepared for analysis.

- **Model Selection:** Depending upon the type of problem (e.g., classification, regression) and properties of data, suitable machine learning model or model category is selected. In this step, parameters such as interpretability, complexity, and performance are considered.
- **Model Training:** The model to be trained is selected in this step using pre-processed historical data. The algorithm learns the input feature patterns and relationships to make predictions on the target variable.
- **Model Evaluation:** Its performance is then tested with unseen data after training the model to compute its accuracy, stability, and generalization capability. This is done in a setup consisting of a set of statistical measures appropriate for the problem.
- **Model Deployment:** Once the model passes the test and is satisfactory, it is deployed into production where it can be used in prediction on numerous new real-world data.

This will most often entail integration with existing systems.

The arrows here refer to a procedure flow, i.e., the steps evolve from one another and lead to an operating prediction system.



3.7. Data Visualization

This diagram, for reference named Figure 7: Data Visualization Process, explicitly shows the step-by-step procedure of transforming raw data into meaningful and useful graphic presentations. Data visualization is a very important tool in data science to facilitate users to comprehend complex data sets, recognize patterns, and effectively present using graphical presentation.

The process, as shown in the figure, generally follows these steps:

- **Data Collection:** First, collect the required data from their sources so that they can be visualized.
- **Data Processing:** After collection, raw data is processed, that is, cleaned, transformed, and put into a visualization-ready form. Processing makes data consistent and accurate.

- **Choose Visualization Type:** This is where nature of data and information to be communicated determine the optimum type of chart, graph, or graphical representation. (e.g., comparisons - bar charts, trends - line graph, relationships - scatter plots, etc.).
- **Create Visualization:** In this, the selected type of visualization is executed to produce the actual visual representation on the transformed data. This is generally applying the utilization of expert software or coding libraries.
- **Interpretation:** Finally, there is interpretation of the resulting visualization in order to extract insights, look for patterns, test hypotheses, or discover new facts from data. This translates visual patterns into actionable insight.

This linear movement deals with how raw information is progressively made finer and displayed graphically for improved comprehension and decision-making.

4. Results

1. Structured Problem-Solving Framework

- Clearly defined business problems were linked to data science objectives, with a strong tie between technical insight and business aims.
- With stakeholder alignment, success metrics were established, developing a numerical roadmap for result measurement.

2. High-Quality Data Preprocessing

- Extensive data cleansing eliminated errors, duplicates, and missing values.
- Using standardization and enrichment techniques, the dataset was made analysis-ready, enabling reliable modeling.

3. Insights: Deep Insights Through Data Exploration

- Exploratory Data Analysis (EDA) revealed strong trends, patterns, and correlations in the data.
- Bar chart, histogram, and scatter plot summary screenshots helped identify anomalies and relationships, which influenced model selection and feature building.

4. Feature Engineering Optimized

- Features that were suitably selected after careful selection and transformation were used, improving model input quality.
- Dimensionality reduction techniques (e.g., PCA) were employed to enhance the model efficiency at no cost to interpretability.

5. Accurate Predictive Modeling

- Machine learning models were fitted and tested and yielded excellent performance metrics on key measures:
- Accuracy, Precision, Recall, and F1-Score were utilized to confirm the generalizability of the model.
- Models demonstrated potential for consistent forecasting of outcomes from historical data.

6. Successful Deployment and Monitoring

- The finished model was implemented via dashboards and APIs to provide real-time decision-making assistance.
- Regular monitoring ensured consistent performance of the model, identifying model drift signals for pre-emptive retraining.

Visual Results

All findings came in through tables, graphs, and figures, mirroring each phase of the methodology—right from initial data collection to final model deployment. This allowed for complete understanding of the process and inferences, so that the findings could be made both understandable as well as actionable to technical as well as non-technical stakeholders.

5. Conclusion

Data science, for all its innovative influence across various domains, is hampered by a basic flaw: the absence of a widely shared definition. Such vagueness, added to nascent scientific underpinnings and profound educational deficits, slows its potential as an integrated and accepted field. The existing scenario, defined by scattered techniques and excessive dependence on discrete tools, restricts efficient data mining and information generation. Analytic tools and programs, though excellent in their own right, are too frequently used in isolation without a supporting theoretical structure, and they replicate with difficulty, develop standardization issues, and produce inconsistencies. This sequential development makes the definitional crisis even worse and causes a general lack of data literacy among professional and scientific communities, thus hindering wider adoption and innovation.

Our extensive survey aims to overcome these fundamental problems by integrating various scholarly and practical viewpoints. We emphasize data science's irrevocable impact on the world labor market, scientific development, industrial technology innovation, and social development, as well as providing a clear roadmap moving forward. We recommend bolstering the scientific foundation of the discipline through systematic theory building and empirical verification. In addition, we offer explicit suggestions on future studies to aim at having a more accurate, widely approved, and broader definition of data science. Most importantly, we stress the need for an immediate closing of the current education gap, encouraging practices that increase data literacy and strengthen the theoretical basis of the field at all levels of education and professional development.

The Data Science Lifecycle, a reproducible and systematic process, became a strong basis for solving difficult data challenges. It starts with a well-defined Business Understanding, in which project outcomes and success factors are carefully established jointly with stakeholders, resulting in a clear Data Science Problem Statement. This initial step guarantees that technical analyses have a solid business value context.

With this, the Data Mining Process systematically discovers useful patterns, encompassing data selection, preprocessing, transformation, and applying various data mining techniques. A key part of this is the Data Cleaning Process, which carefully prepares raw data by removing errors, duplicates, and inconsistencies and providing a high-quality dataset for analysis and modeling.

Data Exploration subsequently delivers first insights using descriptive statistics and visualization to allow pattern recognition and hypothesis formation. This recursive process informs the following feature engineering. Feature Engineering is essential for model performance improvement through feature selection, building, transforming, and extracting from the data.

At the heart of data science frequently resides Predictive Modeling, wherein appropriate machine learning algorithms are chosen, trained on pre-processed past data, and thoroughly tested on accuracy, precision, recall, and F1-score. Good models are subsequently easily Deployed and Monitored using dashboards and APIs, offering real-time decision support and maintaining consistent performance by identifying and correcting model drift. Lastly, Data Visualization is crucial in presenting complicated data and model results in forms that are comprehensible and actionable graphical interpretations to allow both technical and non-technical stakeholders to communicate.

Our observations prove that by adopting this structured problem-solving model, high-quality data preprocessing, rigorous data exploration, engineered features optimized for performance, accurate predictive modeling, and successful deployment and monitoring, data science is capable of producing concrete outcomes. These encompass clearly defined business solutions, high-quality data, deep insights, improved model performance, and guaranteed real-time decision support.

By placing emphasis on these interrelated solutions—a common definition, a solid theoretical basis, holistic education, and the rigorous application of the Data Science Lifecycle—this article establishes a sound foundation upon which to overcome the contemporary problems facing data science. This sets the way for data science to realize its deep and transformative potential in every aspect of society and enterprise.

References

1. Christozov, D. G., Rasheva-Yordanova, K., & Toleva-Stoimenova, S. (2020). Data Science is Here: Are We Ready to Benefit From the Opportunities It Provides?. In Examining the Roles of Teachers and Students in Mastering New Technologies (pp. 108-127). IGI Global.
2. Timbers, T., Campbell, T., & Lee, M. (2022). Data science: A first introduction. CRC Press.
3. Van der Aalst, W. M. (2014). Data scientist: The engineer of the future. In Enterprise interoperability VI: Interoperability for agility, resilience and plasticity of collaborations (pp. 13-26). Springer International Publishing.
4. Loukides, M. (2011). What is data science?. " O'Reilly Media, Inc.".
5. Finzer, W. (2013). The data science education dilemma. Technology Innovations in Statistics Education, 7(2).
6. Steinberg, D. M., & Aronovich, E. (2020). Thoughts on Data Science in business and industry. Applied Stochastic Models in Business & Industry, 36(1).
7. Brodie, M. L. (2019). What is data science?. In Applied data science: Lessons learned for the data-driven business (pp. 101-130). Cham: Springer International Publishing.
8. Ahmad, N., Hamid, A., & Ahmed, V. (2022). Data science: Hype and reality. Computer, 55(2), 95-101.
9. Benjamins, V. R. (2014, June). Big data: from hype to reality?. In Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14) (pp. 1-2).

10. Provost, F., & Faw]cett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.
11. Cao, L. (2017). Data science: a comprehensive overview. ACM Computing Surveys (CSUR), 50(3), 1-42.
12. Donoho, D. (2017). 50 years of data science. Journal of Computational and Graphical Statistics, 26(4), 745-766.
13. Saltz, J. S., & Stanton, J. M. (2017). An introduction to data science. Sage Publication
14. Lazer D, Kennedy R, KG, Vespignani A. The parable of google flu: traps in big data analysis. Science. 2014;343(3):1203-1205.
15. Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 1-42.
16. Kelleher, J. D., & Tierney, B. (2018). *Data science*. MIT press.
17. Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*(12), 64-73.
18. Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, *26*(4), 745-766.
19. Swan, M. (2013). The quantified self: Fundamental disruption in big data science and biological discovery. *Big data*, *1*(2), 85-99.
20. McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60-68.
21. Suthaharan, S., & Suthaharan, S. (2016). *Science of information* (pp. 1-13). Springer US.
22. Ziman, J. (2001). Real science: What it is, and what it means.
23. Agresti, A., Franklin, C. A., & Klingenberg, B. (2020). *The art and science of learning from data*. Pearson.
24. Agresti, A., Franklin, C. A., & Klingenberg, B. (2020). *The art and science of learning from data*. Pearson.
25. Nagel, E. (1979). *The structure of science* (Vol. 411). Indianapolis: Hackett publishing company.
26. Fontana, A., & Frey, J. (1994). The art of science. *The handbook of qualitative research*, *361376*.