ISSN 2616-7433 Vol. 6, Issue 6: 75-80, DOI: 10.25236/FSST.2024.060612

Scientific Nature and Scientific Problems of Data Science

Li Xinyao

Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, Guangdong, China, 519085

Abstract: *As an emerging discipline, data science is crucial for the scientific circle to explore the discussion of its scientific nature and core scientific issues. This research covers a comprehensive discussion from scientific paradigms to scientific attributes to core scientific issues, with particular emphasis on falsiability, reproducibility, scientific spirit, and the ability to iterate quickly.The theoretical system construction of data science also includes key issues such as the alignment of data and problem, the trust relationship between data and model, and the balance between performance and interpretability. Through this comprehensive discussion, the study aims to provide new perspectives on the scientific nature of data science and to find new ways to solve the scientific problems it faces, thus promoting the development of data science as a scientific discipline.*

Keywords: *data science; scientific; falsifiable; reproducibility; scientific research paradigm*

1. Introduction

Data science converts large amounts of complex data into insight to driving technology and society forward. However, as a discipline, what is the scientific nature of data science and what are the core science problems are still open problems. The discussion of these problems is of great significance for the academic construction and practical application of data science.This paper deeply analyzes the scientific research paradigm and the core scientific attributes of data science, and systematically discusses the main scientific problems it faces. The goal of this study is not only to clarify the scientific status of data science, but also to promote its methodology innovation and interdisciplinary application, and to provide the theoretical basis and practical guidance for the further development of data science.

2. The scientific research paradigm in data science

2.1. Comparison between traditional and modern scientific paradigms

When discussing the scientific research paradigm of data science, it is inevitable to compare the traditional scientific paradigm with the modern data-driven scientific paradigm. Traditional scientific paradigms, such as experimental and theoretical sciences, rely on rigorous experimental design, hypothesis verification, and theoretical derivation.These paradigms often emphasize controllable experimental conditions and repeated experiments to verify the reliability of the hypotheses. Relatively speaking, data science, as a representative of the modern science paradigm, breaks through the boundaries of the traditional paradigm, takes data and algorithms as the core, and focuses on discovering patterns and predicting future trends from a large number of data. Data science uses machine learning, statistical analysis, and computational models to process large-scale, unstructured, or semi-structured data collected from the real world, a methodology that emphasizes the ability for exploratory and predictive analysis of empirical data.Different from the dependence on theory in traditional scientific methods, data science uses more computational technology and mathematical models to extract knowledge directly from data, which marks the transformation of scientific methods from relying on theory to relying on data. This shift not only improves research efficiency, but also expands the boundaries of scientific research, allowing scientists to address previously unsolvable issues, such as using big data analysis to predict market trends or disease outbreaks.

ISSN 2616-7433 Vol. 6, Issue 6: 75-80, DOI: 10.25236/FSST.2024.060612

2.2. A unique methodology of data science

Research approaches to data science extend far beyond the bounds of traditional statistical inference, which extensively utilizes advanced machine learning techniques to process and analyze large-scale datasets. These technologies give data scientists the ability to reveal hidden patterns and deep relationships in complex data structures, and then derive new theories or optimize decision-making processes.In particular, deep learning technology, which has been applied in the field of image and speech recognition, demonstrates its powerful ability to automatically identify and learn patterns from large amounts of data. Furthermore, the methodology of data science also places special emphasis on the data processing and cleaning process, a key step in extracting useful information from raw data and translating it into an analytable format.These analyses are supported not only by complex algorithms but also by powerful computational resources that are able to automatically adjust their parameters to optimize model performance. The methodology of data science also includes the rigorous validation of the model, which ensures the accuracy and generalization ability of the model by implementing cross-validation, A / B testing and other methods. Through this comprehensive and multi-angle approach, data science not only improves the efficiency and accuracy of analysis, but also enhances the scientific nature and reliability of decision-making [1].

2.3. Model optimization and parameter adjustment

Data science is crucial in the process of model building, which directly affects the prediction accuracy and generalization ability of the model.When implementing model optimization, algorithm selection, parameter adjustment and balance of model complexity are the core tasks. Commonly used techniques such as grid search and stochastic search find the optimal model configuration by systematically exploring the parameter space.Regularization techniques such as L1 and L2 regularization are widely used to prevent model overfitting, maintain model simplicity by penalizing the complexity of models, and thus”Impr’ve their performance on unknown data.In deep learning models, batch normalization and discarding techniques are also often used to adjust the network structure, and these methods help to accelerate the training process and improve the stability of the model. The optimization of the parameters requires not only the support of algorithms and techniques, but also a comprehensive evaluation based on the model performance, such as using the performance of the validation set to guide the direction of the parameter tuning.Through these methods, data scientists are able to finely control the learning process of the model, ensuring that the model achieves optimal performance on a variety of tasks and datasets.

2.4. Construction and evaluation criteria of the datasets

The construction of the dataset includes the collection, cleaning, and preprocessing of the data, and each step requires a precise technical operation to ensure the quality and representativeness of the data. During data collection, ensuring the diversity of data sources and the comprehensiveness of data samples are key, which helps to improve the generalization ability of the model.During the data cleaning phase, handling missing values, detecting and handling outliers, and performing data transformation are common steps in preprocessing and can help improve data consistency and applicability.The missing value processing method can be deleted, interpolated or filled to select the most appropriate method according to the specific situation; outlier detection can be realized through statistical analysis or machine learning algorithm to ensure the authenticity and accuracy of data; data conversion, such as normalized or standardized, can help to eliminate the magnitude difference and improve the learning efficiency and effect of the model.When evaluating the validity of the dataset, adopting techniques such as cross-validation can ensure the consistency of the model performance on various data subsets, thus evaluating the stability and predictive power of the model. Cross-validation, by dividing the dataset into multiple subsets, which are recycled for training and testing, helps to reduce the bias and variance of the model and provides a more reliable evaluation of the performance.

3. The core scientific attributes of data science

3.1. Authenticability and reproducibility

Permiseptibility, as a standard of scientific research, requires that scientific hypotheses must be able to prove false theoretically or experimentally. Data science applies advanced statistical methods and

ISSN 2616-7433 Vol. 6, Issue 6: 75-80, DOI: 10.25236/FSST.2024.060612

machine learning algorithms to build models that generate testable predictions based on experimental or observational data, thus providing the opportunity to test hypotheses.During the modeling process, the data scientist sets the parameters of the specific algorithm, and uses the real-world data to train the model to predict future events. If the prediction results are significantly inconsistent with the actual observations, the null hypothesis may be rejected. Reproducibility refers to the ability that research results can be replicated by other researchers under the same conditions, which is particularly important in data science.Due to the large amount of data involved and complex data processing processes, ensuring reproducibility requires the disclosure of data sources and methods used, as well as ensuring the transparency of data processing and analysis. In the practice of data science, the transparency and tractability of research are enhanced by version-control data analysis scripts and the use of open-source tools, so that other researchers can copy the research process and verify the authenticity of the results.

3.2. Scientific spirit and rapid iteration

The rapid iterative culture of data science emphasizes the flexibility and efficiency in the process of scientific exploration, and this culture tends to adopt agile research methods and rapid feedback loops to optimize the analysis models and data processing algorithms. In data science projects, the implementation of rapid iteration means constant adjustments and improvements, from the acquisition and preprocessing of data, to model development and testing, both to better adapt to new discoveries, or to cope with new problems encountered in the analysis of experimental data.The data scientist may employ exploratory data analysis in the initial stages of the project to identify key features and potential outliers in the data set, subsequently adapting the data cleaning strategy and preprocessing steps based on these preliminary findings. In subsequent processes, they may iteratively adjust the parameters of the model or try different algorithmic frameworks to improve the accuracy of the model predictions. This rapid iterative process not only accelerates the speed of knowledge extraction from data, but also enables data science projects to adapt more flexibly to changing data environments and project requirements.The scientific spirit of data science is also reflected in the pursuit of the transparency and explanatory ability of models and algorithms, to ensure that the results of the analysis are not only accurate, but also understandable and credible. This emphasis on transparency and interpretability helps to improve the social acceptance of models and ensures that data science as a practice of a science can provide effective and reliable support in a variety of decision-making processes. Through these methods, data science is constantly pushing the boundaries of knowledge, while also constantly testing and optimizing its own scientific methodology to maintain its relevance and effectiveness in a rapidly developing technology environment [2].

3.3. Scientific research Program and theoretical system

As a scientific subject, the development of data science depends on the construction of scientific research program and theoretical system. This requires us not to simply apply the technology, but to systematically understand and framework the various research methods, theoretical hypotheses and their application in real scenarios.In constructing the theoretical system of data science, we must make clear the scientific research methods, and at the same time, deeply explore the nature of data, the transformation mechanism of information, and how to extract knowledge from the data. The core of this theoretical system lies in the construction and verification of the model, through which we can test the validity of the theoretical hypothesis and provide a profound explanation of the phenomena. Through such a methodological framework, data science can provide a new perspective to observe and parse the complex information world, making data not only regarded as a collection of numbers or symbols, but into actionable knowledge, further promoting decision support and innovative development. Building a sound theoretical system of data science not only helps to promote the systematization and standardization of data science as a discipline, but also provides a solid foundation for interdisciplinary research, so that it can play a greater role in solving real-world problems.For example, the theoretical system of data science can be illustrated by the following data table. The prediction performance of different models on the same data set is shown in Table 1 to verify the validity of different theoretical assumptions:

Table 1: shows the model performance comparisons and their corresponding theoretical hypothesis validation

Types of models	data set	definition	recall	F1 score	Theoretical hypothesis
model A	data set 1	0.92	0.89	0.9	hypothesis 1
model B	data set 1	0.85	0.90	0.87	hypothesis 2
model C	data set 1	0.88	0.93	0.9	hypothesis 3

ISSN 2616-7433 Vol. 6, Issue 6: 75-80, DOI: 10.25236/FSST.2024.060612

4. Core scientific issues of data science

4.1. Problem alignment of data or data alignment problem

Data science faces a fundamental scientific question in the exploration of problems and data alignment: whether to let the problem guide the direction of data collection and analysis (problem alignment data), or whether to find and define problems on the basis of existing data (data alignment problems).The former, problem alignment data, is usually seen in specific scientific research and application fields, research hypothesis first, researchers design experiments and collect data according to preset scientific problems to verify the theory or solve specific problems. The advantage of this approach is that the data collection and analysis process can be designed with highly relevant and reliable data support.There are also limitations to this approach, especially where data are less available or experiments are costly, and over-reliance on the problem lead may lead to neglect of other potential values in the available data. The corresponding method of data alignment problem is particularly important in the era of big data. In this mode, researchers take the existing large-scale data set as the starting point to explore the patterns and laws in the data, so as to define and refine the research question [3].

4.2. Trust data or a trust model

In the practice of data science, one of the problems of science, is to determine the level of trust in data or models during the analysis. This issue concerns the quality, completeness, representativeness of the data as well as the accuracy, robustness, and transparency of the model. The frequent challenge for data scientists is that data can be biased, noisy, or uncomplete, problems that can lead to inaccurate analytical results, and blind trust in the data can lead to erroneous decisions and conclusions.While models, especially complex machine learning models such as deep learning networks, perform well in processing large-scale datasets and identifying complex patterns, their black-box nature and dependence on training data may make the results difficult to interpret and verify.This leads to a key question: whether we should rely more on data intuition or model reasoning in making decisions and predictions. Over-reliance on the model may lead to the neglect of the real patterns in the data, while ignoring the analytical power of the model may waste the opportunity to dig deep into the potential of the data. The key to addressing this problem lies in the development and adoption of technologies that can improve data transparency and model interpretability, such as feature importance analysis and model interpretability frameworks, as well as enhanced quality control measures during data processing.

4.3. Balance of performance and interpretability

With the development of machine learning and artificial intelligence technologies, high-performance models such as deep learning networks have made remarkable achievements in image recognition and natural language processing. These models are widely used for their excellent predictive capabilities, but their high complexity often leads to their being "black boxes" and difficult to explain their internal working mechanisms and decision-making processes.This lack of transparency and interpretability may raise serious ethical and legal problems in high-risk areas such as healthcare, finance, because the opacity of decision-making can affect the impartiality, security and reliability of the model.Conversely, simple models such as decision trees and linear regression, while providing better interpretability, often underperform complex models when handling complex data or capturing high-dimensional patterns, and data scientists face challenges in finding the best balance between the performance and interpretability of the models.To solve this problem, researchers are developing new methods and techniques such as interpretable machine learning frameworks and algorithms designed to enhance the transparency of complex models and allow users to understand the predictive behavior and decision basis of the model making.Research is also exploring the balance of performance and interpretability through model fusion and multi-model systems, such as combining deep learning and decision trees, using the ability of deep learning to process data and the decision transparency of decision trees.In addition, research is also exploring the balance of performance and interpretability through model fusion and multi-model systems, such as combining deep learning and decision trees, using the ability of deep learning to process data and the decision transparency of decision trees. The performance indicators (e. g., precision, recall, F1 score) of the different models on the same data set are shown in Table 2:

ISSN 2616-7433 Vol. 6, Issue 6: 75-80, DOI: 10.25236/FSST.2024.060612

Table 2: Comparison of model performance and interpretability.

types of models	data set	definition	recall	F1 score	Interpretability rating
Deep learning network	data set X	0.94	0.91	0.92	low
decision tree	data set X	0.85	0.88	0.86	tall
linear regression	data set X	0.80	0.82	0.81	tall
Hybrid model (deep learning + decision tree)	data set X	0.90	0.89	0.89	centre

5. The methodology and practice of data science

5.1. Application of Statistics and machine learning

Traditional methods of statistics, such as regression analysis, analysis of variance, and hypothesis testing, have long been used for data analysis, providing a rigorous mathematical framework for model building and interpretation of results. The core advantage of these methods is their ability to provide reliable inferences about data relationships on a given dataset, and their theoretical basis makes the results highly interpretable.With increasing data volume and increasing complexity, traditional statistical methods have shown limitations in handling large-scale datasets or unstructured data. In this context, machine learning techniques have become a powerful supplement and alternative. Machine learning, especially deep learning, optimizes the ability to process big data and can identify complex patterns and structures in the data, which are often difficult to capture by traditional methods.Machine learning models such as random forests, support vector machines, and neural networks are powerful tools for predicting results from data, capable of handling high-dimensional data and automatically adapting to non-linear relationships in the data. Moreover, a key advancement of machine learning is its ability to improve its prediction performance without explicit programmatic instructions by learning from data[4].

5.2. Data segmentation and model validation

Data segmentation typically involves dividing large datasets into the training set, the validation set and the test set, for model learning and parameter adjustment, the validation set for model selection and hyperparameter tuning, and the test set for evaluating model performance on unseen data.The purpose of this segmentation method is to simulate the behavior of the model when facing new and unknown data and ensure the performance of the model in practical application. An effective data segmentation strategy can significantly reduce the risk of overfitting, which occurs when the model performs well on the training data but poorly on the new data.During the segmentation of data, researchers must pay attention to the representativeness and randomness of the data to ensure the consistency of the data distribution among different subsets, so as to avoid introducing bias. Model validation is another complex area in data science that involves assessing the predictive power and stability of models. Commonly used model validation techniques include cross-validation and the introduction of new independent datasets for external validation.Cross-validation provides a comprehensive assessment of model stability and reliability by randomly drawing different training and test data from the original data multiple times. While external validation, that is, to test the model using completely independent datasets, is the final step [5] to test the generalization ability of the model.

5.3. People in the ring road role

In the data science process, the concept of "Human-in-the-Loop" (HITL) refers to the practice of direct human involvement in data processing, model training, and result verification at key decision points. This approach underscores the importance of direct human intervention in ensuring the accuracy, transparency, and ethics of data science applications.Although automated technologies and machine learning algorithms are able to handle large-scale datasets and perform complex computational tasks, human intuition, experience, and critical thinking in data interpretation, outlier processing, and final decision making remain irreplaceable.For example, in the medical health field, machine learning models can help identify disease patterns and predict patient outcomes, but the final diagnostic decision often requires the professional judgment of the doctor to consider clinical factors that the model does not fully capture, and human practice in the loop is also extremely important for model training and

ISSN 2616-7433 Vol. 6, Issue 6: 75-80, DOI: 10.25236/FSST.2024.060612

tuning. Data scientists and field experts usually need to re-annotate or adjust the training data according to the results of the preliminary model output to optimize the performance and adaptability of the model. During the training process of machine learning, checking and correcting label errors or data inconsistency through manual intervention can significantly improve the learning efficiency and final accuracy of the model.More importantly, the practice of including people in the loop helps to build trust in data science models and algorithms, especially in application scenarios that require high accuracy and impartiality, such as law and financial services. By ensuring that every step of data science is supervised and verified by humans, bias can be reduced and the fairness and objectivity of the model can be enhanced, thus making data science work more in line with ethical standards and social responsibilities.

6. Conclusion

As a comprehensive and rapidly developing discipline, data science and its core scientific problems in practice are constantly put forward and discussed.Through a thorough analysis of the scientific research paradigm, the core scientific attributes and the key problems facing data science, this paper reveals the challenges and opportunities of data science in providing falsifiable, maintaining reproducibility, balancing performance and interpretability, and the important role of man in the loop.It is these explorations and challenges that drive the continuous progress of data science in theory and application, provide innovative solutions for multidisciplinary fields, and point out the direction for the future development of data science.

References

[1] Hao Shuhui. Wang Guodong uses good data science and digital technology to overcome the "black box" problem [N]. China Metallurgical News, 2024-04-25 (001).
[2] Zhao Yingjie, Hou Juan. Digital Humanities: New Challenges to the boundaries of Science [J]. Journal of Hunan University of Humanities, Science and Technology, 2024,41 (02): 8-15.
[3] Chao Lemen. Scientific nature of data science and analysis of scientific problems [J]. Computer Science, 2024,51 (01): 26-34.
[4] Liu Liang. Research on the Methodology innovation of Ideological and political education in the Digital age [D]. Jiangxi University of Finance and Economics, 2023.001862.
[5] Kang Chao, She Shuanghao. Discussion on the scientificity and controversy of big data methods in the study of ideological and political education [J]. China Audio-visual Education, 2021, (09): 59-63 + 87.