standards

MDPI

Article

The Expansion of Data Science: Dataset Standardization

Nuno Pessanha Santos [1,2]

[1] Portuguese Military Research Center (CINAMIL), Portuguese Military Academy (Academia Militar), R. Gomes Freire 203, 1169-203 Lisbon, Portugal; santos.naamp@academiamilitar.pt
[2] Portuguese Navy Research Center (CINAV), Portuguese Naval Academy (Escola Naval), Alfeite, 2800-001 Almada, Portugal

Abstract: With recent advances in science and technology, more processing capability and data have become available, allowing a more straightforward implementation of data analysis techniques. Fortunately, available online data storage capacity follows this trend, and vast amounts of data can be stored online freely or at accessible costs. As happens with every evolution (or revolution) in any science field, organizing and sharing these data is essential to contribute to new studies or validate obtained results quickly. To facilitate this, we must guarantee interoperability between existing datasets and developed software, whether commercial or open-source. This article explores this issue and analyzes the current initiatives to establish data standards and compares some of the existing online dataset storage platforms. Through a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis, it is possible to better understand the strategy that should be taken to improve the efficiency in this field, which directly depends on the data's characteristics. The development of dataset standards will directly increase the collaboration and data sharing between academia and industry, allowing faster research and development through direct interoperability.

Keywords: datasets; standards; standardization; guidelines; framework; interoperability

Citation: Pessanha Santos, N. The Expansion of Data Science: Dataset Standardization. *Standards* **2023**, *3*, 400–410. https://doi.org/10.3390/standards3040028

Academic Editors: Ramy A. Fathy and Georgios Dounias

Received: 4 May 2023
Revised: 24 November 2023
Accepted: 28 November 2023
Published: 30 November 2023

1. Introduction

With the verified and expected scientific evolution, a development existed from the standard statisticians or software engineers to data scientists today [1,2]. Even though a formal definition of data science does not yet exist, we can state that data science was created from a conjunction of several disciplines, such as data engineering, machine learning, and advanced analysis [3,4]. Some data science applications use data mining since they focus on discovering patterns and relationships in the dataset's variables [5] that correspond to the essential tasks that must be performed during data analysis. It is crucial to keep up with the constant increase in applications to stay updated with the latest developments in the field [6].

The existing advances in the available computer processing capability and the vast increase in the available data due to proper data logging make it possible to extract more information from data in real time, allowing the generation of knowledge [7,8]. Data analysis techniques do not have a specific field for the application being implemented for most of them, e.g., marketing [9], healthcare [10], or electronic commerce [11]. If we have data, we can retrieve essential knowledge from them by implementing data analysis algorithms [12,13]. If the knowledge is obtained in real-time, it can be a vital field advantage and one of the most significant contributions to the application's success. Real-time analysis has many advantages, including faster and more accurate interaction with a process or specific field of application and guaranteeing the convergence to a specific endpoint [14,15]. Data can be considered the new oil since they becomes vital to guarantee the quality of the provided products and services [16–18].

With the verified increase in online storage capacity over time, publicly available datasets and sharing platforms have emerged. Among existing worldwide platforms,

we have Kaggle [19] and the University of California Irvine Machine Learning Repository (UCIMLR) [20]. Most platforms can access and organize competitions using the stored datasets [21–23]. It is easy to state that several platforms provide the same service with different requirements and information about each stored dataset, which limits the interoperability [24,25] between the implemented analysis and the developed software. Uniformizing the requirements and dataset descriptions is essential to efficiently interpret and use the stored data.

For much easier pre-processing [26,27], software development, and data analysis, it is essential to decrease the needed requirements for adaptations necessary to ensure interoperability. Defining and having interoperability between different data sources and types is critical and can be considered a direct contribution to facilitating research and development. The *holy grail* of interoperability is a framework that allows datasets to be easily used by different software independently of its source. In the technological evolution history, we have some examples of success in standardization, e.g., the Portable Operating System Interface (POSIX) or computer graphics framework [28] and the Open System Interconnection (OSI) model that allowed rapid development in their areas. It is essential to learn from the good examples and ensure all efforts are made to guarantee interoperability.

A standard is a document that can define the characteristics of a product, process, or service [29]. Using a proper framework, it is possible to implement the defined standard structure to build something useful [30]. A common framework could help to achieve interoperability between different software and to decrease the time needed to repeat or implement additional data pre-processing [25,31]. Interoperability should guarantee basic principles, such as robustness to new implementations to leave room for evolution. It must be independent of the implementation, being as inclusive as possible, and independent of the technology used since the technology constantly evolves and methods can quickly become obsolete [32].

For ease of use, it is preferable that existing data be made public and are already pre-processed [33]. This interoperability will also contribute to validating the obtained scientific results since many more algorithms can be applied and compared against the same data. Accessing a vast quantity of data is useless or brings nothing if we cannot use them or understand them. As we have templates for scientific documents and other applications, we have to ensure that we have proper standards in this field of application.

It is essential to analyze and evaluate the current state of data standardization to have a clear perspective of context. As in any field of application, it is crucial to have a proper implementation strategy, and this strategy benefits from a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis [34–36]. The SWOT analysis categories are cross-referenced and correlated to obtain results that provide conclusions and recommendations. This analysis enhances the development of the organization, project, or business venture.

The main contributions of this article are (i) an analysis of the currently existing dataset standardization initiatives, (ii) a proposal of a SWOT analysis for the data standardization approach, and (iii) an initial analysis of the strategy that must be followed to ensure data standardization. Our research aims to address the following question: *What are the current challenges associated with dataset standardization, and what are the key recommendations that could contribute to a higher level of global standards adoption?*

Apart from this introduction, this article is divided into three sections. Section 2 describes existing dataset standardization initiatives. Section 3 presents a SWOT analysis and recommends a strategy for dataset standardization. Finally, Section 4 summarizes the conclusions and outlines future work.

2. Dataset Standardization

In recent years, the daily amount of generated and stored data has increased exponentially [37–39], creating the necessity to develop dataset standards to share, analyze, and manage these data. This has led to the development of various data management systems and analytical tools to handle large and complex datasets [40,41]. However, the absence of

globally adopted standardized data formats may hinder researchers from quickly sharing and comparing results. In a world where real-time analysis and processing are critical, we must ensure that we can easily compare results to respond quickly to the desired implementation or application.

Establishing dataset standards is crucial for collaboration and validation in academia and industry. Notable initiatives have already been implemented to establish data standards and best practices in this area, such as:

- **Findability, Accessibility, Interoperability, and Reuse (FAIR)** [42]—An initiative that started defining a set of principles applicable to research data to promote interoperability between data sources.
- **Document, Discover, and Interoperate (DDI)** [43]—An initiative that started defining social science research data metadata standards. It provides a framework for describing and documenting research data and promoting data reuse in this field of science.
- **Clinical Data Interchange Standards Consortium (CDISC)** [44]—An initiative that started defining principles applicable to clinical research data. It provides a framework for describing, acquiring, and documenting data used in this field of science.

There needs to be more than just developing a standard by itself since it must be adopted by worldwide dataset users to be considered an implementation success. Suppose this effort also incorporates the pre-processing [45] scripts and the developed software. In that case, it will undoubtedly increase the knowledge obtained from the data and decrease the time needed for pre-processing and analysis. Data heterogeneity is a limitation in the data-sharing process since each data format requires a different pre-processing approach, and no one-size-fits-all solution exists. Apart from that, and considering the existing limitations, the advantages of having data standardization easily surpass all the current disadvantages.

As also happened with the verified evolution in other fields of science, data are starting to be a significant revenue source in some companies' profits [46,47]. This will surely hinder data standardization and the increase in available public datasets. In the *eyes* of these companies and institutions, the existence of available public datasets decreases dataset exclusivity and allows others to provide similar data sharing or analysis services. Apart from the vast majority of the academic community, which mainly focuses on science dissemination, the possible lobby created by those companies is certainly a threat that must be considered when we start thinking about a dataset standardization strategy.

Some datasets may be considered confidential or can provide some personal information, and we must ensure dataset anonymization in these cases [48–50]. This process protects private and sensitive information (data) by applying pre-processing techniques to erase or encrypt unique identifiers that can connect the dataset to an individual while minimizing information loss [51,52]. At any time, and depending on the dataset's content, each individual who sees his/her rights as not respected must be able to trigger quick and easy actions to correct the situation. This must also be a global concern when we are dealing with data.

When there is a shortage of real data in research, synthetic data can be used to speed up a development process [53–56]. Typically, when conducting a study on data, we focus on real data for inference purposes. However, synthetic data should closely resemble real data and are often used to accelerate algorithm development. Since most algorithms are designed to be deployed in the real world, they are fine-tuned using real data [57,58]. When generating a dataset that includes synthetic data, it is crucial to label and distinguish all synthetically generated data clearly. Consistency in the standardization procedure should be maintained across all data types, whether synthetic or real.

Another issue that must be considered is data consistency and accuracy [59,60]. It is essential to consider the existing errors in the data acquisition process depending on the data type and sensor we are considering. A possible solution to guarantee dataset accuracy is to ensure that a third-party institution or organization without any economic relation

to the dataset owner certifies its level of precision to ensure transparency in the process. Sensor calibration is crucial for ensuring high accuracy and consistency in data acquisition systems [61,62], especially when providing data for datasets.

With the current emergence of dataset-sharing platforms, it is essential to analyze the most popular. Some of the most popular dataset-sharing platforms are:

- **Kaggle** [63]—A platform that allows access to a wide range of dataset topics intended for artificial intelligence and data science algorithms. Some of the datasets are intended for image segmentation [64,65], object detection [65,66], and image generation [65,67], among many other applications.
- **UCI** [20]—A platform that allows access to a wide range of dataset topics for machine learning [68], including standard classification [69] or clustering algorithms [70].
- **Data.gov** [71,72]—A platform that allows access to datasets collected from United States of America (USA) agencies, with a wide range of topics such as education, finance, health, and climate, among others [73,74].
- **Google Dataset Search (GDS)** [75]—A search engine that can be used to find datasets online, covering a wide range of topics, e.g., social sciences [76] or finance [77].
- **Amazon Web Services Open Data Registry (AWSODR)** [78]—A platform that allows access to a wide range of datasets hosted by Amazon, covering topics such as climate [79] or geospatial data [80].
- **Microsoft Research Open Data (MROD)** [81]—A platform that allows access to a wide range of dataset topics, including, e.g., computer vision [82,83] or natural language processing [84,85].
- **World Bank Open Data (WBOD)** [86]—A platform that allows access to datasets regarding global world development, including, e.g., poverty [87], education [88], or climate change [89,90].

A comparison between the characteristics of the described dataset-sharing platforms is made in Table 1. The table analysis indicates that most datasets come from open-access sources with varying user interfaces and requirements. Reducing the number of platforms may prove beneficial for having better global control over the accuracy and consistency of data provided. It might lead to an accelerated converged standardized view of the platform architectures and data exchange protocols. However, this premise has yet to be proven in future work. It should be highlighted that achieving this goal could be difficult, if not impossible, as it is impossible to control or prohibit the creation of new platforms or products globally. Most platforms also allow users to use an Application Programming Interface (API) to retrieve and manipulate data without explicitly having to download it. This can be particularly helpful when working with datasets with a significant amount of information (size). The organization and content of a dataset often depend on its creator or data-sharing platform due to a lack of defined standards beyond the standard data description.

Companies and organizations may want to create their own platforms for sharing datasets and providing access to their data to obtain a financial return [16,18,91]. This occurrence should be minimized or eliminated since it threatens global data sharing. A global effort is needed to optimize resources and ensure consistent and accurate global data dissemination in the future. The data must be consistent and accurate, follow a specific standard, and guarantee accuracy by passing a proper certification process, as stated before.

Next, we need to understand what strategy should be followed to mitigate the described limitations, improve the future of data science, and obtain knowledge [92] from data faster and simpler. In the next section, an initial study of a strategy is performed to generate a better future in this field. Defining long-term goals and objectives is crucial for creating a strategy that determines necessary actions and resources [93].

Table 1. Comparison of characteristics of popular platforms for sharing datasets.

Dataset	Open Data	Access	Update Frequency	User Interface	API Access
Kaggle [63]	No	Free Paid	Daily	User friendly (interactive)	Yes
UCIMLR [20]	Yes	Free	Irregular	User friendly (simple)	No
Data.gov [71]	Yes	Free	Irregular	User friendly (simple)	Yes
GDS [75]	Yes	Free	Irregular	Simple search interface	No
AWSODR [78]	Yes	Free	Irregular	User friendly (simple)	Yes
MROD [81]	Yes	Free	Irregular	Simple search interface	Yes
WBOD [86]	Yes	Free	Irregular	User friendly (simple)	Yes

3. Strategy Analysis

A strategy's success lies mainly in its existence and correct execution rather than the strategy content itself [94,95]. A strategy that can be implemented worldwide must have realistic objectives that justify each one and highlight the advantages to the end user, whether an individual or a governmental organization. As previously mentioned, it is vital to establish long-term goals and objectives when developing a strategy.

Most of the current efforts and initiatives in dataset standardization were discussed in the previous section. It is essential to consider the future and provide a feasible strategy. To achieve that objective, an initial SWOT analysis was suggested regarding the standardization of the datasets. The SWOT analysis is considered a strategic planning tool used to understand an organization, project, or business venture and makes it possible to enhance its development [34–36]. We can consider the internal characteristics of a company, organization, or institution since they should have similar objectives regarding dataset standardization. The strengths are internal characteristics or resources that can give some advantage (Positive vs. Internal factors), weaknesses are internal characteristics or resources that can bring some disadvantage (Negative vs. Internal factors), opportunities are external factors that can contribute to the success (Positive vs. External factors), and threats are external factors that can contribute to failure (Negative vs. External factors).

The identified strengths (Positive vs. Internal factors) of the standard implementation were:

- The increase in the consistency and accuracy of the data and datasets;
- The data become easier to interpret and analyze, also allowing faster technological innovation;
- It is possible to save time and resources in data pre-processing and analysis;
- Data sharing between systems becomes more accessible by ensuring interoperability;
- The ability to develop internal knowledge that directly increases productivity levels;
- The increased ease of collaboration between teams;
- The ability to provide data-related services easily, e.g., data analysis or sharing data that complies with a recognized standard.

The identified weaknesses (Negative vs. Internal factors) of the standard implementation were:

- The implementation and development of a data standard requires a considerable amount of time;
- The implementation and development of a data standard requires a significant amount of resources, e.g., workers and hardware;

- If the data standard is not flexible enough to accommodate the necessary data types or respective content, it can be impossible to implement;
- The developed standards may not be compatible with currently existing tools and data sources;
- Current teams may require training time to adopt the standards effectively;
- An initial investment is needed to implement a proper structure to perform data standardization more easily.

The identified opportunities (Positive vs. External factors) of the standard implementation were:

- The generalized adoption of a data standard makes it possible to increase external collaboration and interoperability;
- The developed applications and software provide direct interoperability with any dataset following the data standard;
- The well-known standards allow fast response since applications and services use consistent and accurate data;
- The boost in the research and development in the data science field;
- The growth in the economy since many companies can benefit from the advantages of data standardization;
- The recruitment process becomes easier for the company and the worker: since the worker already knows the dataset standard, he/she can become productive earlier without requiring the usual adaptation time.

The identified threats (Negative vs. External factors) to the standard implementation were:

- The existence of several redundant or competing standards with property formats not open to everyone;
- The data standards, if not correctly updated periodically, can become rapidly obsolete;
- Even with a standard, the data can be misinterpreted or manipulated;
- Some companies can develop standards to decrease interoperability and maintain service or application exclusivity;
- It is important to standardize data to ensure privacy and security;
- The cost-effectiveness of the data standardization investment since the data standards may not be accepted globally.

Using a SWOT analysis [35] can be beneficial when analyzing an environment. However, it should not be the only focus [96]. By systematically matching and exploring the relationships between opportunities, weaknesses, strengths, and threats, it is possible to gain a more nuanced understanding of how these elements interact [96,97]. This leads to a more comprehensive and tailored strategy recommendation. After thoroughly reviewing the relevant literature and analyzing the gathered data, we obtained crucial insights into the current context of the dataset standardization field. This information will help us develop a comprehensive strategy for addressing future challenges in the area [94,98].

Most identified strengths are based on increased data consistency, accuracy, internal knowledge development, and resource optimization. Weaknesses are mainly based on the vast resources needed to implement a dataset standard from scratch. Opportunities are linked to increasing external collaboration and interoperability between all the services, applications, and software, allowing a fast response. The threats are mainly focused on the standard since the cost of its implementation and the pursuit of service and application exclusivity by a specific company or group of companies can lead to competing or even redundant standards.

It is essential to balance the weaknesses and the existing opportunities since the resources needed for the internal implementation can be gathered from external collaboration and by ensuring interoperability. The strengths can also balance the threats since the increase in data consistency, accuracy, knowledge, and resource optimization can help to deal with the cost-effectiveness of the data standardization and overcome the necessity for a

company to have its own standard, in which case the company would lose its interoperability capacity and need to have very specialized workers that have to learn a very specific implementation with a limited field of application.

The SWOT analysis is merely the initial step in the strategic planning process. It is imperative to analyze and make necessary adjustments to ensure accuracy continually. With an environment or context change, new challenges and opportunities will emerge, and every possibility must be considered. Even after a suitable strategy formulation, the implementation is the main challenge that must be overcome [99]. Since we are talking about a worldwide strategy for dataset standardization, we must face significant challenges in the strategic alignment between different cultural, social, and economic characteristics [100,101]. Still, the expected results will compensate for all the existing adversities and difficulties.

A critical factor in every strategy is the evaluation and control of its implementation. Defining and using proper performance metrics to evaluate and control the strategy implementation is essential [102,103]. The performance metrics should be based on clear strategic objectives and provide a comprehensive picture of the obtained performance over the needed analysis dimensions [104]. As described before, in a world dominated by data, this strategic performance management can also be performed using data analysis techniques [105–107]. If the dataset standard is accepted and adopted as a worldwide strategy, each developed application and implementation must comply to ensure interoperability. The necessity to adhere to the standard will indirectly be a performance metric since the number of applications or implementations that comply can be quantified.

Standardizing datasets can transform raw data into valuable assets, ensuring consistency, accuracy, and efficiency in analysis. This accelerates innovation, collaboration, and informed decision-making while addressing complexity, change resistance, and compatibility challenges. By embracing this practice with adaptability and by considering privacy concerns, individuals and organizations can drive insights, cooperation, and growth.

4. Conclusions

As we enter an era of data-driven decision-making, the importance of consistent, accurate, and interoperable data cannot be overstated. Our study highlights the importance of standardized datasets for enhancing collaboration, driving innovation, and making informed decisions. Individuals and organizations that leverage standardization while navigating adaptability and foresight challenges will be able to harness the full potential of their data.

Data science is critical since it can help individuals and organizations make better decisions by retrieving knowledge from data. Fortunately for the field, more data can be used for multiple applications since higher data-logging and available online storage space exist nowadays. With the rise of new requirements for data, we must adapt ourselves and be able to deal with them. The easiest way is to create a clear standard that can guarantee data standardization and maintain its accuracy and consistency. However, a dataset standard must be able to cover a vast number of possibilities, and it needs to be updated periodically to ensure that it continues to make sense even with the expected evolutions in the data science field.

The proposed SWOT analysis should be continually updated to consider the natural development of the environment and access to new sources of information. Through its analysis, it is possible to verify that the main weaknesses and threats identified are mainly based on the vast resources needed to implement a dataset standard from scratch and adhering to the standard since a company looking for exclusivity can develop competing or even redundant standards. Taking the necessary actions to mitigate the identified weaknesses and threats is essential. Data standardization is not easy, and there is no one-size-fits-all solution, as there is the possibility that a dataset does not fit into the defined data standardization. The dataset standard should be continuously and quickly updated as soon as possible to deal with real-world implementation challenges. It is easy to state

after the review and analysis performed in this article that dataset standardization must be a worldwide concern and that in the end, even with some challenges and threats, the obtained implementation gain will compensate for these costs. As soon as possible, academia and industry team participants should consider the scenario and create the necessary documentation to follow and adopt worldwide considering the conclusions and recommendations obtained by our study.

After defining the dataset standardization process, the future depends on data analysis. Our objective in this field should be to maximize insights while streamlining processing. Machine learning and Artificial Intelligence (AI) can help unearth knowledge from datasets. Dedicated resources for automated data preprocessing and standardized formats can save time and effort. The future of data analysis will depend on seamless integration, efficiency, and actionable insights. Collaboration among experts, technologists, and stakeholders will continue to drive this transformation, making data a strategic advantage.

What is the present and future of dataset standardization? We believe that it is crucial to implement the actions resulting from our strategic analysis to ensure rapid development in the field and maximize its potential. Data are the present and future. Effective utilization requires well-defined standards.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the article.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Kim, M.; Zimmermann, T.; DeLine, R.; Begel, A. Data scientists in software teams: State of the art and challenges. *IEEE Trans. Softw. Eng.* **2017**, *44*, 1024–1038. [CrossRef]
2. Davenport, T.H.; Patil, D. Data scientist. *Harv. Bus. Rev.* **2012**, *90*, 70–76. [PubMed]
3. Gibert, K.; Horsburgh, J.S.; Athanasiadis, I.N.; Holmes, G. Environmental data science. *Environ. Model. Softw.* **2018**, *106*, 4–12. [CrossRef]
4. Nasution, M.K.; Sitompul, O.S.; Nababan, E.B. Data science. *J. Phys. Conf. Ser.* **2020**, *1566*, 012034. [CrossRef]
5. Coenen, F. Data mining: Past, present and future. *Knowl. Eng. Rev.* **2011**, *26*, 25–29. [CrossRef]
6. Sarker, I.H. Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective. *SN Comput. Sci.* **2021**, *2*, 377. [CrossRef]
7. Inmon, W.H. The data warehouse and data mining. *Commun. ACM* **1996**, *39*, 49–51. [CrossRef]
8. Mikut, R.; Reischl, M. Data mining tools. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2011**, *1*, 431–443. [CrossRef]
9. Sterne, J. *Artificial Intelligence for Marketing: Practical Applications*; John Wiley & Sons: Hoboken, NJ, USA, 2017.
10. Obenshain, M.K. Application of data mining techniques to healthcare data. *Infect. Control Hosp. Epidemiol.* **2004**, *25*, 690–695. [CrossRef]
11. Kohavi, R.; Provost, F. *Applications of Data Mining to Electronic Commerce*; Springer: Berlin/Heidelberg, Germany, 2001.
12. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. *AI Mag.* **1996**, *17*, 37.
13. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P. *Knowledge Discovery and Data Mining: Towards a Unifying Framework*; AAAI Press: Portland, OR, USA, 1996; Volume 96, pp. 82–88.
14. Sismanoglu, G.; Onde, M.A.; Kocer, F.; Sahingoz, O.K. Deep learning based forecasting in stock market with big data analytics. In Proceedings of the 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), Istanbul, Turkey, 24–26 April 2019; pp. 1–4.
15. Mohamed, N.; Al-Jaroodi, J. Real-time big data analytics: Applications and challenges. In Proceedings of the 2014 International Conference on High Performance Computing & Simulation (HPCS), Bologna, Italy, 21–25 July 2014; pp. 305–310.
16. Stach, C. Data Is the New Oil–Sort of: A View on Why This Comparison Is Misleading and Its Implications for Modern Data Administration. *Future Internet* **2023**, *15*, 71. [CrossRef]
17. Loi, M.; Dehaye, P.O. If data is the new oil, when is the extraction of value from data unjust? *Filos. Quest. Pubbliche* **2017**, *7*, 137–178.
18. Possler, D.; Bruns, S.; Niemann-Lenz, J. Data Is the New Oil—But How Do We Drill It? Pathways to Access and Acquire Large Data Sets in Communication Science. *Int. J. Commun. (19328036)* **2019**, *13*, 3894–3911.

19. Bojer, C.S.; Meldgaard, J.P. Kaggle forecasting competitions: An overlooked learning opportunity. *Int. J. Forecast.* **2021**, *37*, 587–603. [CrossRef]
20. Asuncion, A.; Newman, D. UCI Machine Learning Repository. 2007. Available online: https://archive.ics.uci.edu/ml/index.php (accessed on 21 April 2023).
21. Yang, X.; Zeng, Z.; Teo, S.G.; Wang, L.; Chandrasekhar, V.; Hoi, S. Deep learning for practical image recognition: Case study on kaggle competitions. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 923–931.
22. Iglovikov, V.; Mushinskiy, S.; Osin, V. Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. *arXiv* **2017**, arXiv:1706.06169.
23. Taieb, S.B.; Hyndman, R.J. A gradient boosting approach to the Kaggle load forecasting competition. *Int. J. Forecast.* **2014**, *30*, 382–394. [CrossRef]
24. Kasunic, M. Measuring systems interoperability: Challenges and opportunities. Defense Technical Information Center. 2001. Available online: https://apps.dtic.mil/sti/pdfs/ADA400176.pdf (accessed on 21 April 2023).
25. Tolk, A.; Muguira, J.A. The levels of conceptual interoperability model. In Proceedings of the 2003 Fall Simulation Interoperability Workshop, Orlando, FL, USA, 14–19 September 2003; Volume 7, pp. 1–11.
26. Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J.J.; Downey, G.; Blanchet, L.; Buydens, L.M. Breaking with trends in pre-processing? *TrAC Trends Anal. Chem.* **2013**, *50*, 96–106. [CrossRef]
27. Rinnan, Å. Pre-processing in vibrational spectroscopy–when, why and how. *Anal. Methods* **2014**, *6*, 7124–7129. [CrossRef]
28. Foley, J.D.; Van, F.D.; Van Dam, A.; Feiner, S.K.; Hughes, J.F. *Computer Graphics: Principles and Practice*; Addison-Wesley Professional: Boston, MA, USA, 1996; Volume 12110.
29. Geraci, A. *IEEE Standard Computer Dictionary: Compilation of IEEE Standard Computer Glossaries*; IEEE Press: Piscataway, NJ, USA, 1991.
30. Mora, A.; Riera, D.; Gonzalez, C.; Arnedo-Moreno, J. A literature review of gamification design frameworks. In Proceedings of the 2015 7th International Conference on Games and Virtual Worlds for Serious Applications (VS-Games), Skövde, Sweden, 16–18 September 2015; pp. 1–8.
31. Wegner, P. Interoperability. *ACM Comput. Surv. (CSUR)* **1996**, *28*, 285–287. [CrossRef]
32. Mellal, M.A. Obsolescence—A review of the literature. *Technol. Soc.* **2020**, *63*, 101347. [CrossRef]
33. Mihaescu, M.C.; Popescu, P.S. Review on publicly available datasets for educational data mining. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2021**, *11*, e1403. [CrossRef]
34. Sarsby, A. *SWOT Analysis—A Guide to SWOT for Business Studies Students*; Spectaris Ltd - Leadership Library: Suffolk, UK, 2016.
35. Benzaghta, M.A.; Elwalda, A.; Mousa, M.M.; Erkan, I.; Rahman, M. SWOT analysis applications: An integrative literature review. *J. Glob. Bus. Insights* **2021**, *6*, 55–73. [CrossRef]
36. Leigh, D. SWOT Analysis. In *Handbook of Improving Performance in the Workplace: Volumes 1–3*; Pfeiffer: San Francisco, CA, USA, 2009.
37. Larson, D.; Chang, V. A review and future direction of agile, business intelligence, analytics and data science. *Int. J. Inf. Manag.* **2016**, *36*, 700–710. [CrossRef]
38. Kumari, A.; Tanwar, S.; Tyagi, S.; Kumar, N. Verification and validation techniques for streaming big data analytics in internet of things environment. *IET Netw.* **2019**, *8*, 155–163. [CrossRef]
39. Acharjya, D.P.; Ahmed, K. A survey on big data analytics: Challenges, open research issues and tools. *Int. J. Adv. Comput. Sci. Appl.* **2016**, *7*, 511–518.
40. Anuradha, J. A brief introduction on Big Data 5Vs characteristics and Hadoop technology. *Procedia Comput. Sci.* **2015**, *48*, 319–324.
41. Sagiroglu, S.; Sinanc, D. Big data: A review. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, USA, 20–24 May 2013; pp. 42–47.
42. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 1–9. [CrossRef]
43. Vardigan, M.; Heus, P.; Thomas, W. Data documentation initiative: Toward a standard for the social sciences. *Int. J. Digit. Curation* **2008**, *3*. [CrossRef]
44. Sato, I.; Kawasaki, Y.; Ide, K.; Sakakibara, I.; Konomura, K.; Yamada, H.; Tanaka, Y. Clinical data interchange standards consortium standardization of biobank data: A feasibility study. *Biopreserv. Biobank.* **2016**, *14*, 45–50. [CrossRef]
45. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [CrossRef]
46. Akhigbe, A.; Stevenson, B.A. Profit efficiency in US BHCs: Effects of increasing non-traditional revenue sources. *Q. Rev. Econ. Financ.* **2010**, *50*, 132–140. [CrossRef]
47. Schüritz, R.; Seebacher, S.; Dorner, R. Capturing value from data: Revenue models for data-driven services. In Proceedings of the 50th Hawaii International Conference on System Sciences, San Diego, CA, USA, 4–7 January 2017. [CrossRef]
48. Byun, J.W.; Sohn, Y.; Bertino, E.; Li, N. Secure anonymization for incremental datasets. In Proceedings of the Secure Data Management: Third VLDB Workshop, SDM 2006, Seoul, Republic of Korea, 10–11 September 2006; Proceedings 3; Springer: Berlin/Heidelberg, Germany, 2006; pp. 48–63.

49. Bayardo, R.J.; Agrawal, R. Data privacy through optimal k-anonymization. In Proceedings of the 21st International Conference on Data Engineering (ICDE'05), Tokyo, Japan, 5–8 April 2005; pp. 217–228.
50. Murthy, S.; Bakar, A.A.; Rahim, F.A.; Ramli, R. A comparative study of data anonymization techniques. In Proceedings of the 2019 IEEE 5th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing,(HPSC) and IEEE International Conference on Intelligent Data and Security (IDS), Washington, DC, USA, 5–8 April 2019.
51. Ghinita, G.; Karras, P.; Kalnis, P.; Mamoulis, N. Fast data anonymization with low information loss. In Proceedings of the 33rd International Conference on Very Large Data Bases, Vienna, Austria, 23–27 September 2007; pp. 758–769.
52. Loukides, G.; Gkoulalas-Divanis, A. Utility-preserving transaction data anonymization with low information loss. *Expert Syst. Appl.* **2012**, *39*, 9764–9777. [CrossRef]
53. Raghunathan, T.E. Synthetic data. *Annu. Rev. Stat. Appl.* **2021**, *8*, 129–140. [CrossRef]
54. Bolón-Canedo, V.; Sánchez-Maroño, N.; Alonso-Betanzos, A. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.* **2013**, *34*, 483–519. [CrossRef]
55. Pessanha Santos, N.; Lobo, V.; Bernardino, A. Two-stage 3D model-based UAV pose estimation: A comparison of methods for optimization. *J. Field Robot.* **2020**, *37*, 580–605. [CrossRef]
56. Pessanha Santos, N.; Melício, F.; Lobo, V.; Bernardino, A. A ground-based vision system for UAV pose estimation. *Int. J. Robot. Mechatron.* **2014**, *1*, 138–144. [CrossRef]
57. Tremblay, J.; Prakash, A.; Acuna, D.; Brophy, M.; Jampani, V.; Anil, C.; To, T.; Cameracci, E.; Boochoon, S.; Birchfield, S. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 969–977.
58. Handa, A.; Patraucean, V.; Badrinarayanan, V.; Stent, S.; Cipolla, R. Understanding real world indoor scenes with synthetic data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4077–4085.
59. Cong, G.; Fan, W.; Geerts, F.; Jia, X.; Ma, S. *Improving Data Quality: Consistency and Accuracy*; VLDB: Marousi, Athens, Greece, 2007.
60. Han, J. *Data Mining: Concepts and Techniques*; Han, J., Kamber, M., Pei, J., Eds.; Elservier—Morgan Kaufman Publishers: Burlington, MA, USA, 2011.
61. Kirianaki, N.V.; Yurish, S.Y.; Shpak, N.O.; Deynega, V.P. *Data Acquisition and Signal Processing for Smart Sensors*; Wiley: New York, NY, USA, 2002.
62. Römer, K.; Blum, P.; Meier, L. Time synchronization and calibration in wireless sensor networks. *Handb. Sens. Netw. Algorithms Archit.* **2005**, 199–237. [CrossRef]
63. Kaggle. Available online: https://www.kaggle.com/ (accessed on 21 April 2023).
64. Kang, W.X.; Yang, Q.Q.; Liang, R.P. The comparative research on image segmentation algorithms. In Proceedings of the 2009 First International Workshop on Education Technology and Computer Science, Wuhan, China, 7–8 March 2009; Volume 2, pp. 703–707.
65. Zhang, X.; Dahu, W. Application of artificial intelligence algorithms in image processing. *J. Vis. Commun. Image Represent.* **2019**, *61*, 42–49. [CrossRef]
66. Yang, R.; Yu, Y. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Front. Oncol.* **2021**, *11*, 638182. [CrossRef]
67. Dosovitskiy, A.; Tobias Springenberg, J.; Brox, T. Learning to generate chairs with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1538–1546.
68. Mahmudur Rahman Khan, M.; Bente Arif, R.; Abu Bakr Siddique, M.; Rahman Oishe, M. Study and Observation of the Variation of Accuracies of KNN, SVM, LMNN, ENN Algorithms on Eleven Different Datasets from UCI Machine Learning Repository. *arXiv* **2018**, arXiv:1809.06186.
69. Kotsiantis, S.B.; Zaharakis, I.; Pintelas, P. Supervised machine learning: A review of classification techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* **2007**, *160*, 3–24.
70. Ezugwu, A.E.; Ikotun, A.M.; Oyelade, O.O.; Abualigah, L.; Agushaka, J.O.; Eke, C.I.; Akinyelu, A.A. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Eng. Appl. Artif. Intell.* **2022**, *110*, 104743. [CrossRef]
71. Data.gov. Available online: https://www.data.gov/ (accessed on 21 April 2023).
72. Ding, L.; DiFranzo, D.; Graves, A.; Michaelis, J.R.; Li, X.; McGuinness, D.L.; Hendler, J. Data-gov wiki: Towards linking government data. In Proceedings of the 2010 AAAI Spring Symposium Series, Stanford, CA, USA, 22–24 March 2010.
73. Krishnamurthy, R.; Awazu, Y. Liberating data for public value: The case of Data. gov. *Int. J. Inf. Manag.* **2016**, *36*, 668–672. [CrossRef]
74. Stevens, H. Open data, closed government: Unpacking data.gov.sg. *First Monday* **2019**, *24*. [CrossRef]
75. Google Dataset Search. Available online: https://datasetsearch.research.google.com/ (accessed on 21 April 2023).
76. Grimmer, J.; Roberts, M.E.; Stewart, B.M. Machine learning for social science: An agnostic approach. *Annu. Rev. Political Sci.* **2021**, *24*, 395–419. [CrossRef]
77. Dixon, M.F.; Halperin, I.; Bilokon, P. *Machine Learning in Finance*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 1170.
78. Amazon Web Services Open Data. Available online: https://registry.opendata.aws/ (accessed on 21 April 2023).

79. Kashinath, K.; Mustafa, M.; Albert, A.; Wu, J.; Jiang, C.; Esmaeilzadeh, S.; Azizzadenesheli, K.; Wang, R.; Chattopadhyay, A.; Singh, A.; et al. Physics-informed machine learning: Case studies for weather and climate modelling. *Philos. Trans. R. Soc. A* **2021**, *379*, 20200093. [CrossRef] [PubMed]
80. Kiwelekar, A.W.; Mahamunkar, G.S.; Netak, L.D.; Nikam, V.B. Deep learning techniques for geospatial data analysis. In *Machine Learning Paradigms: Advances in Deep Learning-Based Technological Applications*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 63–81.
81. Microsoft Research Open Data. Available online: https://msropendata.com/ (accessed on 21 April 2023).
82. Khan, A.I.; Al-Habsi, S. Machine learning in computer vision. *Procedia Comput. Sci.* **2020**, *167*, 1444–1451. [CrossRef]
83. Sebe, N.; Cohen, I.; Garg, A.; Huang, T.S. *Machine Learning in Computer Vision*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005; Volume 29.
84. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 604–624. [CrossRef]
85. Li, H. Deep learning for natural language processing: Advantages and challenges. *Natl. Sci. Rev.* **2018**, *5*, 24–26. [CrossRef]
86. World Bank Open Data. Available online: https://data.worldbank.org/ (accessed on 21 April 2023).
87. Adams, R.H. *Economic Growth, Inequality and Poverty: Findings from a New Data Set*; World Bank Publications: Washington, DC, USA 2003; Volume 2972 .
88. Altinok, N.; Angrist, N.; Patrinos, H.A. Global data set on education quality (1965–2015). *World Bank Policy Res. Work. Pap.* **2018**. Available online: http://hdl.handle.net/10986/29281 (accessed on 21 April 2023).
89. Rolnick, D.; Donti, P.L.; Kaack, L.H.; Kochanski, K.; Lacoste, A.; Sankaran, K.; Ross, A.S.; Milojevic-Dupont, N.; Jaques, N.; Waldman-Brown, A.; et al. Tackling climate change with machine learning. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–96. [CrossRef]
90. Ardabili, S.; Mosavi, A.; Dehghani, M.; Várkonyi-Kóczy, A.R. Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review. In *Engineering for Sustainable Future: Selected papers of the 18th International Conference on Global Research and Education Inter-Academia—2019, Budapest & Balatonfüred, Hungary, 4–7 September 2019* ; Springer: Cham, Switzerland 2019; pp. 52–62.
91. Javornik, M.; Nadoh, N.; Lange, D. Data is the new oil: How data will fuel the transportation industry—The airline industry as an example. In *Towards User-Centric Transport in Europe: Challenges, Solutions and Collaborations*; Springer: Cham, Switzerland, 2019.
92. Larose, D.T.; Larose, C.D. *Discovering Knowledge in Data: An Introduction to Data Mining*; John Wiley & Sons: Hoboken, NJ, USA, 2014; Volume 4.
93. Nickols, F. Strategy, strategic management, strategic planning and strategic thinking. *Manag. J.* **2016**, *1*, 4–7.
94. Olson, E.M.; Slater, S.F.; Hult, G.T.M. The importance of structure and process to strategy implementation. *Bus. Horizons* **2005**, *48*, 47–54. [CrossRef]
95. Okumus, F. Towards a strategy implementation framework. *Int. J. Contemp. Hosp. Manag.* **2001**, *13*, 327–338. [CrossRef]
96. Teece, D.J. SWOT Analysis. In *The Palgrave Encyclopedia of Strategic Management*; Augier, M., Teece, D.J., Eds.; Palgrave Macmillan: London, UK, 2018; pp. 1689–1690. [CrossRef]
97. Weihrich, H. The TOWS matrix—A tool for situational analysis. *Long Range Plan.* **1982**, *15*, 54–66. [CrossRef]
98. Mintzberg, H.; Ahlstrand, B.; Lampel, J.B. *Strategy Safari: A Guided Tour through the Wilds of Strategic Management*; Simon & Schuster Inc.: New York, NY, USA, 1998.
99. Hill, C.W.; Jones, G.R.; Schilling, M.A. *Strategic Management: Theory: An Integrated Approach*; Cengage Learning: Boston, MA, USA, 2014.
100. Doz, Y.L.; Prahalad, C.K. Managing DMNCs: A search for a new paradigm. *Strateg. Manag. J.* **1991**, *12*, 145–164. [CrossRef]
101. Ghemawat, P. Distance still matters—The hard reality of global expansion. *Harvard Bus. Rev.* **2001**, *79*, 137 .
102. Kaplan, R.S.; Norton, D.P. *The Balanced Scorecard: Translating Strategy into Action*; Harvard Business Press: Cambridge, MA, USA, 1996.
103. Lynch, R.L.; Cross, K.F. *Measure Up!: The Essential Guide to Measuring Business Performance*; Mandarin: London, UK, 1991.
104. Austin, R.D. *Business Performance Measurement: Theory and Practice*; Cambridge University Press: Cambridge, UK, 2002.
105. Mello, R.; Martins, R.A. Can big data analytics enhance performance measurement systems? *IEEE Eng. Manag. Rev.* **2019**, *47*, 52–57. [CrossRef]
106. Armstrong, M.; Baron, A. *Performance Management*; Kogan Page Limited: London, UK, 2000.
107. Ledolter, J. *Data Mining and Business Analytics with R*; John Wiley & Sons: Hoboken, NJ, USA, 2013.