Article

The Expansion of Data Science: Datasets Standardization

Nuno Pessanha Santos *

Posted Date: 8 May 2023

doi: 10.20944/preprints202305.0525.v1

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

Article

The Expansion of Data Science: Datasets 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 the recent advances in science and technology, more processing capability and data have become available, allowing a more straightforward implementation of data analysis techniques. Fortunately, the 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 this data is essential to contribute to new studies or validate the obtained results quickly. To facilitate this, we must guarantee interoperability between the existing datasets and the developed software, whether commercial or open source. This article explores this issue, analyzing the current initiatives to establish data standards and comparing some of the existing dataset online storage platforms. Through a Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis, it is possible to understand better the strategy that should be taken to improve the efficiency in this field, which directly depends on the data characteristics.

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

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, or 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.

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 the data in real time, allowing the generation of knowledge [6,7]. The data analysis techniques do not have a specific field of the application being traversal to most of them, e.g., marketing [8], healthcare [9], or electronic commerce [10]. If we have data, we can retrieve essential knowledge from it by implementing data analysis algorithms. 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. Data can be considered as the new *gold* since it becomes vital to guarantee the quality of the provided products and services.

With the verified increase in online storage capacity over time, publicly available datasets and sharing platforms have emerged. Among the existing worldwide platforms, we have Kaggle [11] or the University of California Irvine Machine Learning Repository (UCIMLR) [12,12]. Most platforms are free to access and even organize competitions using the stored datasets [13–15]. 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 [16,17] 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 a much easier pre-processing [18,19], software development, and data analysis, it is essential to decrease the needed requirements for adaptations being able to ensure interoperability. Defining and

 doi:10.20944/preprints202305.0525.v1

having interoperability between the different data sources and types is essential and can be considered a direct contribution to facilitating research and development. The *holy grail* of interoperability is a framework that allows the 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 [20] and the Open System Interconnection (OSI) model that allowed rapid development in their areas.

A standard is a document that can define the characteristics of a product, process, or service [21]. Using a proper framework, it is possible to implement the defined standard structure to build something useful [22]. 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 [17,23]. Interoperability should guarantee some basic principles, such as robustness to new implementations leaving room for evolution. It must be independent of the implementation, being the most inclusive as possible, and independent of the used technology since the technology constantly evolves and the methods can quickly become obsolete.

The existing data should be publicly available [24] and preferably pre-processed so everyone can easily use it. 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 it or if we cannot understand it. As we have templates for scientific documents and other applications, we have to ensure that we have proper standards in this field of application.

Analyzing and evaluating the current state of data standardization is essential. 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 [25–27]. This analysis allows an understanding of the organization, project, or business venture and makes it possible to enhance its development.

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.

This article is organized as follows. Section 2 presents some currently existing initiatives for dataset standardization. Section 3 presents a SWOT analysis and explores the best strategy for dataset standardization. Finally, we present the conclusions in Section 4.

2. Dataset Standardization

In recent years, the daily amount of generated and stored data has increased exponentially [28–30], creating the necessity of developing dataset standards to share, analyze, and manage this data. This has led to the development of various data management systems and analytical tools to handle large and complex datasets. However, the lack of standardization that can define the used data formats and accuracy requirements has made it difficult for researchers to share and compare the results quickly.

Developing the needed dataset standards is an important step to increase data sharing and collaboration between academia and industry, being essential to compare and even validate the obtained results. Nowadays, some initiatives have started to establish data standards and best practices, such as:

- **Findability, Accessibility, Interoperability and Reuse (FAIR)** [31] - An initiative that started defining a set of principles applicable to research data to promote interoperability between data sources;
- **Document, Discover and Interoperate (DDI)** [32] - An initiative that started defining social sciences research data metadata standards. It provides a framework for describing and documenting research data and promoting data reuse in this field of science;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0525.v1

- **Clinical Data Interchange Standards Consortium (CDISC)** [33] - 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.

Developing a standard by itself is not enough since it must be adopted by the worldwide dataset users to be considered an implementation success. Suppose this effort also incorporates the pre-processing [34] scripts and the developed software. In that case, it will undoubtedly increase the obtained knowledge from the data, decreasing the pre-processing and analysis needed time. 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 is starting to be a significant revenue in some companies' profit [35,36]. 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 the dataset exclusivity and allows others to provide similar data sharing or analysis services. Apart from the vast majority of the academic community mainly focused on science dissemination, the possible lobby created by those companies is certainly a thread that must be considered when we start thinking about strategy.

Some of the datasets may be considered confidential or can provide some personal information, and we must ensure dataset anonymization in these cases [37–39]. This process protects private and sensitive information by applying multiple pre-processing techniques to erase or encrypt unique identifiers that can connect the dataset to an individual. At any time, and depending on the dataset content, each individual that sees their 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.

Another issue that must be considered is data consistency and accuracy [40,41]. 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 the dataset accuracy is to ensure that a third-party institution or organization without any economic relations to the dataset owner certifies its level of precision to ensure transparency in the process.

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** [42] - A platform that allows access to a wide range of dataset topics intended to apply artificial intelligence and data science algorithms. Some of the datasets are intended to apply image segmentation [43,44], object detection [44,45], and image generator [44,46], among many others;
- **UCI** [47] - A platform that allows access to a wide range of dataset topics to apply machine learning [48], including the standard classification [49] or clustering algorithms [50];
- **Data.gov** [51,52] - 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 [53,54];
- **Google Dataset Search (GDS)** [55] - A search engine that can be used to find datasets online, covering a wide range of topics, e.g., social sciences [56] or finance [57];
- **Amazon Web Services Open Data Registry (AWSODR)** [58] - A platform that allows access to a wide range of datasets hosted by Amazon, covering topics such as climate [59] or geospatial data [60];
- **Microsoft Research Open Data (MROD)** [61] - A platform that allows access to a wide range of dataset topics, including, e.g., computer vision [62,63] or natural language processing [64,65];
- **World Bank Open Data (WBOD)** [66] - A platform that allows access to datasets regarding global world development, including, e.g., poverty [67], education [68], or climate change [69,70].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0525.v1

A comparison between the characteristics of the described dataset-sharing platforms is made in Table 1. The table analysis shows that most datasets are from the open-access type with different user interfaces and requirements. It will be much more helpful to have a single platform, or at least decrease the number of platforms, allowing better control of the provided data accuracy and consistency, guaranteeing standardization. Most platforms also allow users to use an Application Programming Interface (API) to retrieve and manipulate data without explicitly having to download it. This is particularly useful when dealing with datasets containing a large amount of data.

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

Dataset	Open Data	Access	Update Frequency	User Interface	API Access
Kaggle [42]	No	Free Paid	Daily	User friendly (interactive)	Yes
UCIMLR [47]	Yes	Free	Irregular	User friendly (simple)	No
Data.gov [51]	Yes	Free	Irregular	User friendly (simple)	Yes
GDS [55]	Yes	Free	Irregular	Simple search interface	No
AWSODR [58]	Yes	Free	Irregular	User friendly (simple)	Yes
MROD [61]	Yes	Free	Irregular	Simple search interface	Yes
WBOD [66]	Yes	Free	Irregular	User friendly (simple)	Yes

Some companies or organizations surely will want to have their data or platform that allows access to it since it will possibly bring, in the end, a financial return [71–73]. This occurrence should be minimized or eliminated since it threatens the global dataset implementation. A global effort must exist to optimize the existing resources, guaranteeing, in the future, that the data presents the needed consistency and accuracy. The data must be consistent and accurate, follow a specific standard, and guarantee accuracy by passing a proper certification process, as stated before.

Understanding what strategy should be followed to mitigate the described limitations, improve the future of data science, and obtain knowledge [74] 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.

3. Strategy Analysis

A strategy's success lies mainly in its existence and correct execution rather than the strategy content itself [75,76]. A worldwide implementable strategy must rely on realistic objectives justifying each one and the advantages to the end user, whether an individual or a governmental organization.

Most of this field's current efforts and initiatives were described in the previous section, being essential to look at the future and provide a feasible direction. An initial SWOT analysis regarding the datasets standardization was proposed for that goal. 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 [25–27]. 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 kind of advantage (Positive vs. Internal factors), weaknesses are internal characteristics or resources that can bring some kind of disadvantage (Negative vs. Internal factors), opportunities are external factors that can contribute to

 doi:10.20944/preprints202305.0525.v1

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 becomes easier to interpret and analyze, also allowing faster technological innovation;
- It is possible to save time and resources in the data pre-processing and analysis;
- The data sharing between systems becomes more accessible by ensuring interoperability;
- The ability to develop internal knowledge that directly increases productivity levels;
- The ability to provide data-related services easier, 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 number 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 initial investment needed to implement a proper structure to perform data standardization easier.

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 a fast response since the 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, it 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;
- The cost-effectiveness of the data standardization investment since the data standards may not be accepted globally.

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 with 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 services and application exclusivity by a specific or a group of companies can lead to competing or even redundant standards.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 8 May 2023 doi:10.20944/preprints202305.0525.v1

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 since the company loses 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 just the beginning of a strategic planning process, being essential to make a constant analysis and perform the needed adaptations to ensure the accuracy of the analysis. 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 [77]. 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 [78,79]. 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 [80,81]. The performance metrics should be based on clear strategic objectives and provide a comprehensive picture of the obtained performance over the needed analysis dimensions [82]. As described before, in a world dominated by data, this strategic performance management can also be performed using data analysis techniques [83–85]. 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.

4. Conclusions

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 exists nowadays. With the rise of new requirements for data, we must adapt ourselves and be able to deal with it. The easier way is to create a clear standard that can guarantee data standardization, maintaining 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, considering 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 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 not a one size fits all solution, having the possibility of having a dataset that does not fit into the defined data standardization. The dataset standard should be continuously and quickly updated to deal with real-world implementation challenges as soon as possible. It is easy to state, and after 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. The future surely relies on data analysis, and we must ensure that we maximize the retrieved knowledge from it and simultaneously decrease the needed efforts in the processing stage.

Author Contributions: Conceptualization, N.P.S.; methodology, N.P.S.; Not applicable; validation, N.P.S.; formal analysis, N.P.S.; investigation, N.P.S.; resources, N.P.S.; data curation, N.P.S.; writing—original draft preparation,N.P.S.; writing—review and editing, N.P.S.; visualization, N.P.S.; supervision, N.P.S.; project

administration, N.P.S.; funding acquisition, Not applicable. The author has read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The manuscript includes all the data and materials supporting the presented conclusions.

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 Transactions on Software Engineering* **2017**, *44*, 1024–1038.
2. Davenport, T.H.; Patil, D. Data scientist. *Harvard business review* **2012**, *90*, 70–76.
3. Gibert, K.; Horsburgh, J.S.; Athanasiadis, I.N.; Holmes, G. Environmental data science. *Environmental Modelling & Software* **2018**, *106*, 4–12.
4. Nasution, M.K.; Sitompul, O.S.; Nababan, E.B. Data science. Journal of Physics: Conference Series. IOP Publishing, 2020, Vol. 1566, p. 012034.
5. Coenen, F. Data mining: past, present and future. *The Knowledge Engineering Review* **2011**, *26*, 25–29.
6. Inmon, W.H. The data warehouse and data mining. *Communications of the ACM* **1996**, *39*, 49–51.
7. Mikut, R.; Reischl, M. Data mining tools. *Wiley interdisciplinary reviews: data mining and knowledge discovery* **2011**, *1*, 431–443.
8. Sterne, J. *Artificial intelligence for marketing: practical applications*; John Wiley & Sons, 2017.
9. Obenshain, M.K. Application of data mining techniques to healthcare data. *Infection Control & Hospital Epidemiology* **2004**, *25*, 690–695.
10. Kohavi, R.; Provost, F. *Applications of data mining to electronic commerce*; Springer, 2001.
11. Bojer, C.S.; Meldgaard, J.P. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting* **2021**, *37*, 587–603.
12. Asuncion, A.; Newman, D. UCI machine learning repository, 2007.
13. Yang, X.; Zeng, Z.; Teo, S.G.; Wang, L.; Chandrasekhar, V.; Hoi, S. Deep learning for practical image recognition: Case study on kaggle competitions. Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 923–931.
14. Iglovikov, V.; Mushinskiy, S.; Osin, V. Satellite imagery feature detection using deep convolutional neural network: A kaggle competition. *arXiv preprint arXiv:1706.06169* **2017**.
15. Taieb, S.B.; Hyndman, R.J. A gradient boosting approach to the Kaggle load forecasting competition. *International journal of forecasting* **2014**, *30*, 382–394.
16. Kasunic, M. Measuring systems interoperability: Challenges and opportunities **2001**.
17. Tolk, A.; Muguira, J.A. The levels of conceptual interoperability model. Proceedings of the 2003 fall simulation interoperability workshop. Citeseer, 2003, Vol. 7, pp. 1–11.
18. Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J.J.; Downey, G.; Blanchet, L.; Buydens, L.M. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry* **2013**, *50*, 96–106.
19. Rinnan, Å. Pre-processing in vibrational spectroscopy–when, why and how. *Analytical Methods* **2014**, *6*, 7124–7129.
20. Foley, J.D.; Van, F.D.; Van Dam, A.; Feiner, S.K.; Hughes, J.F. *Computer graphics: principles and practice*; Vol. 12110, Addison-Wesley Professional, 1996.
21. Geraci, A. *IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries*; IEEE Press, 1991.
22. Mora, A.; Riera, D.; Gonzalez, C.; Arnedo-Moreno, J. A literature review of gamification design frameworks. 2015 7th international conference on games and virtual worlds for serious applications (VS-Games). IEEE, 2015, pp. 1–8.
23. Wegner, P. Interoperability. *ACM Computing Surveys (CSUR)* **1996**, *28*, 285–287.
24. Mihaescu, M.C.; Popescu, P.S. Review on publicly available datasets for educational data mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2021**, *11*, e1403.
25. Sarsby, A. *SWOT Analysis - A guide to SWOT for business studies students*; Leadership Library, 2016.
26. Benzaghta, M.A.; Elwalda, A.; Mousa, M.M.; Erkan, I.; Rahman, M. SWOT analysis applications: An integrative literature review. *Journal of Global Business Insights* **2021**, *6*, 55–73.

27. Leigh, D. SWOT analysis. *Handbook of Improving Performance in the Workplace: Volumes 1-3* **2009**, pp. 115–140.
28. Larson, D.; Chang, V. A review and future direction of agile, business intelligence, analytics and data science. *International Journal of Information Management* **2016**, *36*, 700–710.
29. Kumari, A.; Tanwar, S.; Tyagi, S.; Kumar, N. Verification and validation techniques for streaming big data analytics in internet of things environment. *IET Networks* **2019**, *8*, 155–163.
30. Acharjya, D.P.; Ahmed, K. A survey on big data analytics: challenges, open research issues and tools. *International Journal of Advanced Computer Science and Applications* **2016**, *7*, 511–518.
31. 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.; others. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* **2016**, *3*, 1–9.
32. Vardigan, M.; Heus, P.; Thomas, W. Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation* **2008**, *3*.
33. 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. *Biopreservation and Biobanking* **2016**, *14*, 45–50.
34. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings* **2022**.
35. Akhigbe, A.; Stevenson, B.A. Profit efficiency in US BHCs: Effects of increasing non-traditional revenue sources. *The Quarterly Review of Economics and Finance* **2010**, *50*, 132–140.
36. Schüritz, R.; Seebacher, S.; Dorner, R. Capturing value from data: Revenue models for data-driven services **2017**.
37. Byun, J.W.; Sohn, Y.; Bertino, E.; Li, N. Secure anonymization for incremental datasets. Secure Data Management: Third VLDB Workshop, SDM 2006, Seoul, Korea, September 10-11, 2006. Proceedings 3. Springer, 2006, pp. 48–63.
38. Bayardo, R.J.; Agrawal, R. Data privacy through optimal k-anonymization. 21st International conference on data engineering (ICDE'05). IEEE, 2005, pp. 217–228.
39. Murthy, S.; Bakar, A.A.; Rahim, F.A.; Ramli, R. A comparative study of data anonymization techniques. 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS). IEEE, 2019, pp. 306–309.
40. Cong, G.; Fan, W.; Geerts, F.; Jia, X.; Ma, S. Improving Data Quality: Consistency and Accuracy. VLDB, 2007, Vol. 7, pp. 315–326.
41. Han, J. Data Mining: Concepts and Techniques/Han J., Kamber M., Pei J. *Elservier: Morgan Kaufman Publishers* **2011**.
42. Kaggle. https://www.kaggle.com/. Accessed: April 21, 2023.
43. Kang, W.X.; Yang, Q.Q.; Liang, R.P. The comparative research on image segmentation algorithms. 2009 First international workshop on education technology and computer science. IEEE, 2009, Vol. 2, pp. 703–707.
44. Zhang, X.; Dahu, W. Application of artificial intelligence algorithms in image processing. *Journal of Visual Communication and Image Representation* **2019**, *61*, 42–49.
45. Yang, R.; Yu, Y. Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in oncology* **2021**, *11*, 638182.
46. Dosovitskiy, A.; Tobias Springenberg, J.; Brox, T. Learning to generate chairs with convolutional neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1538–1546.
47. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/index.php. Accessed: April 21, 2023.
48. 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 e-prints* **2018**, pp. arXiv–1809.
49. Kotsiantis, S.B.; Zaharakis, I.; Pintelas, P.; others. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering* **2007**, *160*, 3–24.
50. 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. *Engineering Applications of Artificial Intelligence* **2022**, *110*, 104743.

51. Data.gov. https://www.data.gov/. Accessed: April 21, 2023.
52. Ding, L.; DiFranzo, D.; Graves, A.; Michaelis, J.R.; Li, X.; McGuinness, D.L.; Hendler, J. Data-gov wiki: Towards linking government data. 2010 AAAI spring symposium series, 2010.
53. Krishnamurthy, R.; Awazu, Y. Liberating data for public value: The case of Data. gov. *International Journal of Information Management* **2016**, *36*, 668–672.
54. Stevens, H. Open data, closed government: Unpacking data. gov. sg. *First Monday* **2019**, *24*.
55. Google Dataset Search. https://datasetsearch.research.google.com/. Accessed: April 21, 2023.
56. Grimmer, J.; Roberts, M.E.; Stewart, B.M. Machine learning for social science: An agnostic approach. *Annual Review of Political Science* **2021**, *24*, 395–419.
57. Dixon, M.F.; Halperin, I.; Bilokon, P. *Machine learning in Finance*; Vol. 1170, Springer, 2020.
58. Amazon Web Services Open Data. https://registry.opendata.aws/. Accessed: April 21, 2023.
59. Kashinath, K.; Mustafa, M.; Albert, A.; Wu, J.; Jiang, C.; Esmaeilzadeh, S.; Azizzadenesheli, K.; Wang, R.; Chattopadhyay, A.; Singh, A.; others. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A* **2021**, *379*, 20200093.
60. Kiwelekar, A.W.; Mahamunkar, G.S.; Netak, L.D.; Nikam, V.B. Deep learning techniques for geospatial data analysis. *Machine Learning Paradigms: Advances in Deep Learning-based Technological Applications* **2020**, pp. 63–81.
61. Microsoft Research Open Data. https://msropendata.com/. Accessed: April 21, 2023.
62. Khan, A.I.; Al-Habsi, S. Machine learning in computer vision. *Procedia Computer Science* **2020**, *167*, 1444–1451.
63. Sebe, N.; Cohen, I.; Garg, A.; Huang, T.S. *Machine learning in computer vision*; Vol. 29, Springer Science & Business Media, 2005.
64. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems* **2020**, *32*, 604–624.
65. Li, H. Deep learning for natural language processing: advantages and challenges. *National Science Review* **2018**, *5*, 24–26.
66. World Bank Open Data. https://data.worldbank.org/. Accessed: April 21, 2023.
67. Adams, R.H. *Economic growth, inequality and poverty: Findings from a new data set*; Vol. 2972, World Bank Publications, 2003.
68. Altinok, N.; Angrist, N.; Patrinos, H.A. Global data set on education quality (1965-2015). *World Bank Policy Research Working Paper* **2018**.
69. 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.; others. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)* **2022**, *55*, 1–96.
70. 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. Engineering for Sustainable Future: Selected papers of the 18th International Conference on Global Research and Education Inter-Academia–2019 18. Springer, 2020, pp. 52–62.
71. Javornik, M.; Nadoh, N.; Lange, D. Data is the new oil: How data will fuel the transportation industry—The airline industry as an example. *Towards User-Centric Transport in Europe: Challenges, Solutions and Collaborations* **2019**, pp. 295–308.
72. 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. *International Journal of Communication (19328036)* **2019**, *13*.
73. 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.
74. Larose, D.T.; Larose, C.D. *Discovering knowledge in data: an introduction to data mining*; Vol. 4, John Wiley & Sons, 2014.
75. Olson, E.M.; Slater, S.F.; Hult, G.T.M. The importance of structure and process to strategy implementation. *Business horizons* **2005**, *48*, 47–54.
76. Okumus, F. Towards a strategy implementation framework. *International journal of contemporary hospitality Management* **2001**.
77. Hill, C.W.; Jones, G.R.; Schilling, M.A. *Strategic management: theory: an integrated approach*; Cengage Learning, 2014.

 doi:10.20944/preprints202305.0525.v1

78. Doz, Y.L.; Prahalad, C.K. Managing DMNCs: a search for a new paradigm. *Strategic management journal* **1991**, *12*, 145–164.
79. Ghemawat, P. Distance still matters–the hard reality of global expansion", Hanvard Business Review, September, p. 137 **2001**.
80. Kaplan, R.S.; Norton, D.P.; others. *The balanced scorecard: translating strategy into action*; Harvard business press, 1996.
81. Lynch, R.L.; Cross, K.F. *Measure up!: The essential guide to measuring business performance*; Mandarin, 1991.
82. Austin, R.D. *Business performance measurement: theory and practice*; Cambridge University Press, 2002.
83. Mello, R.; Martins, R.A. Can big data analytics enhance performance measurement systems? *IEEE Engineering Management Review* **2019**, *47*, 52–57.
84. Armstrong, M.; Baron, A. *Performance management*; Kogan Page Limited, 2000.
85. Ledolter, J. *Data mining and business analytics with R*; John Wiley & Sons, 2013.