A Manifesto for Applied Data Science - Reasoning from a Business Perspective

Abstract
Due to the increasing complexity of the discipline Data Science, a partition into sub-disciplines seems appropriate. Therefore, we propose the division of Data Science in Pure Data Science, in which methods and tools are developed, and Applied Data Science, in which these methods and tools are adapted and applied to practical problems of a specific domain. This article focuses on Applied Data Science and how it should be positioned in relation to its adjoining disciplines. We also introduce the term Business Data Science as a specific form of Applied Data Science in the business domain and describe its relationship to existing terms like Business Analytics and Business Intelligence.
Keywords
Data Science, Applied Data Science, Pure Data Science, Business Data Science

1. Partitioning Data Science

Data Science (DS) is developing into an independent new subject area. Highly interdisciplinary by nature, it takes concepts and methods from mathematics/statistics and computer science, it combines, expands, and enhances them and applies them to new areas of application [1]. As an emerging new subject area, DS also develops its own research questions, processes and techniques, independent of its underlying disciplines [2].

Although DS now possesses generic methods and algorithms that can be applied in many domains, they often must be refined to the particular requirements of domain-specific data applications [3]. Systematic approaches are needed to address the complexities of DS problems inherent within these domains, which are not effectively accommodated within a single discipline [4]. To fill this gap, we propose the division of DS into new sub-disciplines.

DS can be divided into two sub-disciplines; a sub-discipline in which methods and tools are developed, and another sub-discipline in which these methods and tools are adapted and applied to practical problems of a specific domain. Following the conventions in Mathematics, where a similar distinction is made, we propose the terms Pure Data Science (PDS) to identify the sub-disciplines focused on the further development of the scientific field of the DS and Applied Data Science (ADS) which is mainly devoted to solve application domain challenges and provide an explanation of the obtained results. An overview of the sub-disciplines with their distinguished characteristics are shown in Table 1.

Table 1

Proceedings Name, Month XX–XX, YYYY, City, Country
EMAIL: email1@mail.com (A. 1); email2@mail.com (A. 2); email3@mail.com (A. 3)
ORCID: XXXX-XXXX-XXXX-XXXX (A. 1); XXXX-XXXX-XXXX-XXXX (A. 2); XXXX-XXXX-XXXX-XXXX (A. 3)

Distinguishing characteristics of Pure Data Science and Applied Data Science

	Pure Data Science	Applied Data Science
Main goal	Further development of the scientific field of data science	Solving application domain problems and explaining the results
Output	New analytical methods and the required theoretical foundations	Problem- and domain-specific analytical methods adaptation and insights from data
Beneficiary area	Data science	Application domain
Required competencies	In-depth knowledge	

This manifesto focuses on ADS and how it should be positioned in relation to its adjoining disciplines. Particularly since 2020, more and more degree programs are being launched worldwide under the name *Applied Data Science*. A search in early November 2021 identified 60 of such programs. The fact that so many universities are launching study programs with the same designation almost simultaneously can hardly be a coincidence. In this manifesto we suggest that this is an appropriate term for the demarcation of the sub-discipline from the other area of DS, which we call PDS.

A closer look at the degree programs named *Applied Data Science* shows that even these have distinctly different emphases, so that a further subdivision of the subdiscipline should be made, which will be presented in the following chapters. In this manifesto, we focus on the application of DS to business problems. There are two main reasons for this: First, a discussion of ADS is only partially possible without considering an explicit application domain. Second, the chosen application domain is particularly relevant due to its widespread use. However, it is also an area that requires differentiation from existing terms, such as business intelligence and business analytics, which are also described in this manifesto.

2. Applied Data Science

Having discussed the division of Data Science into the sub-disciplines of PDS and ADS, we now move ADS to the center of our discussion.

In particular, we consider ADS in the context of its adjoining relevant disciplines by showing their respective influences. **Figure 1** provides an overview of the relations of ADS with: *(1) PDS* and its related disciplines (2) *Focus dependent disciplines, (3) Domain and (4) Ethics and Regulations*.

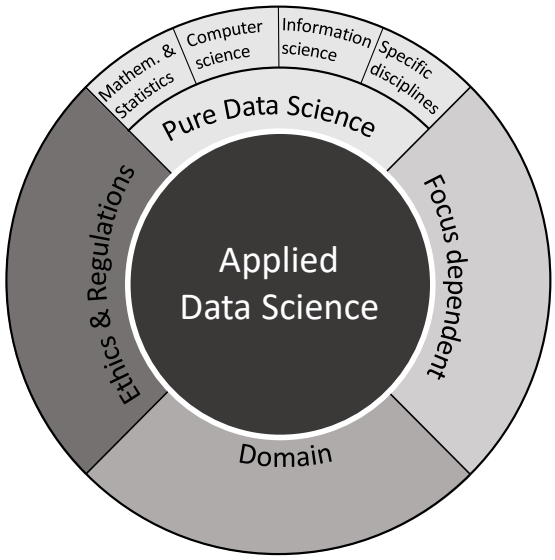

Figure 1: Applied Data Science in the context of its neighboring disciplines

The main purpose of ADS is to solve real-world problems in a domain using and possibly adapting methods and tools provided by PDS. Depending on the focus of an individual project, other disciplines such as linguistics or psychology may also be relevant.

The connections between Data Science and neighboring disciplines are often represented as a Venn diagram with Data Science as the intersection of other disciplines. This type of visualization incorrectly suggests that Data Science is merely a compilation of existing concepts from surrounding disciplines and does not adequately represent the increasing number of concepts and methods developed within Data Science – or with a more detailed view, within PDS and ADS. The type of diagram chosen in our figure is intended to highlight precisely this fact, which justifies the definition of a (sub)discipline in the first place. Self-developed methods of ADS include re-usable domain-specific strategies for applying data mining techniques, communication processes with decision makers, managing analytical projects, or quality attributes as privacy, explainability, debias, fairness, replicability, accuracy, etc. These aspects could be added to the center of the above figure. However, since a complete enumeration is not possible, we have refrained from doing so.

The domain-independent methods and tools developed by PDS rely heavily on the disciplines of mathematics/statistics, computer science and information science. Depending on the specific application domain, other specific disciplines will provide additional support. Therefore, all these disciplines play an important role in ADS as well.

We illustrate the relationships of the four neighboring areas to ADS in an example. Consider an insurance company that wants to identify fraudulent claims. For this purpose, knowledge from all four neighboring areas of ADS are relevant: From the domain (insurance) the contract-specific insurance terms and conditions, from the focus dependent disciplines (in this example linguistics) relevant methods for text analysis, from Ethics and Regulations the consideration, if social customer features should be included, and from Pure Data Science a suitable classification algorithm. Furthermore, ADS-specific concepts such as explainability and debiasing are relevant.

Domain-specific ADS forms

Having clarified that ADS is a sub-discipline of DS, a natural question is whether ADS also has sub-disciplines itself. The formation of an independent sub-discipline of ADS seems appropriate to us if the following criteria are met:

(a) The application domain is sufficiently large.
(b) There are many questions in the application domain that can be answered with the help of ADS.
(c) The questions to be answered make special adjustments and further developments of the ADS methods necessary.

Criteria (a) and (b) are intended to ensure that the application domain under consideration is significant enough to justify the creation of an independent ADS sub-discipline (with a specialized training program, special publications, etc.). Criterion (c) is to ensure that the domain questions cannot be answered using standard ADS methods alone. If that were the case, the creation of a dedicated sub-discipline would not be necessary.

There are a variety of application domains that, according to the above criteria, justify the creation of a separate sub-discipline. Besides the domain they differ in the following aspects:

- focus on specific types of data (structured databases, text data, sensor data, audio/video data)
- use different subsets of the "PDS toolbox"
- adapt general PDS methods to the needs of the domain
- may develop domain specific DS methods
- may use specialized tools/software that supports domain specific DS methods
- take into account domain specific requirements, e.g.
    - legal and ethical concerns
    - data privacy
    - organizational standards
    - scientific rigor
    - explainability of results
    - fairness/bias
    - performance

While a complete listing of all domain-specific ADS forms is difficult, a subdivision into the two forms *ADS in a scientific domain* (e.g., Biological Data Science, Climate Data Science, and in more general sense, Experimental Sciences) and *ADS in a non-scientific domain* (e.g., Business Data Science, Public Service Data Science, Engineering Data Science) seems useful due to the differences presented below. Table 1 shows some of the differences we consider between ADS in a scientific domain and ADS in a non-scientific domain.

Table 2
Applied Data Science in a scientific domain vs. Applied Data Science in a non-scientific domain

	Applied Data Science in a scientific domain	**Applied Data Science in a non-scientific domain**
Objectives	Scientific insights	Pragmatic insights and their usage
Communication barrier between domain and data science	Often low	Often high
Required level of problem understanding	Often deep	Mostly medium
Conduction of the Data Science project	Typically conducted by the domain experts (scientists) themselves	Typically conducted by Applied Data Scientists

3. Business Data Science

In contrast to the previous considerations, the classification of terms in the context of analytics with a business perspective has already been considered intensively in the literature, but a consensus has not yet been found. We first give an overview of the common usages of the terms Business Analytics (BA) and Business Intelligence (BI). Then we introduce the term Business Data Science and describe its relationship to BA/BI.

The literature reveals significant areas of ambiguity in relation to the demarcations that distinguish the disciplines and subfields that concern organizational decision making. For example, some authors still do not distinguish between the terms Business Analytics and Data Science [5, 6, 7, 8]. However, many other authors do distinguish the two [9, 10, 11]. Despite this ambiguity, reviews of the literature have identified emerging themes relating to the demarcation of relevant disciplines.

A characteristic that is often used in the literature to delimit the terms is the type of statistical methods used. It is therefore advantageous to first break down the statistics into its three sub-disciplines [12]:

- *Descriptive statistics* is used to describe a data set (representing a phenomena) using meaningful parameters such as location, dispersion, or correlation measures as well as the graphical visualization of the data. Such techniques are used to analyze operational business data to aggregate decision-supporting information by means of data-based dashboards, scorecards, or reports. In descriptive statistics, neither hypothesis tests based on mathematical models are used to generalize the results nor are probabilistic forecasts made. Descriptive statistics does not use probability theory.
- *Explorative statistics* deals with the extraction of structures from data. It uses the results of descriptive statistics and creates mathematical models or statistical hypotheses. Many analytical methods of explorative statistics (e.g. cluster analyses, factor analyses, principal component analyses) are used in data mining to reduce the complexity of large amounts of data and thus gain insight into underlying relationships [13].

- *Inferential statistics* tests the models and hypotheses generated from data by exploratory algorithms and makes intensive use of the probability theory. The aims are to produce reliable generalizations of results beyond the existing data set or forecast future developments. The application of such methods in a business context is usually summarized as predictive analytics.

Analytics has its origins within logic, mathematics, and science. Grammatically, the term Analytics includes the suffix "-ics", which refers to a body of knowledge or principles. Nelson defines Analytics as "the scientific process or discipline of fact-based problem-solving" [14]. Davenport and Harris define Analytics as the "extensive use of data, statistical and quantitative analysis, exploratory and predictive models, and fact-based management to drive decisions and actions" [15]. There are many other definitions within the literature, but this theme of applying advanced analytics and statistical techniques on data to drive decision making is often a common thread that links them together.

Analytics is a broad discipline that logically includes the sub-fields of BA and Data Science. Business Analytics is primarily concerned with business relevance and actionable insights from analyses. This concern is highlighted in a review performed by Phelps and Szabat [9]. They found that definitions of BA often contained a substantial focus on the statistical and quantitative analysis of data, and decision-making support within business domains. Their review also sought definitions of Data Science from the literature and they observed clear differences between the two. The definitions they found relating to Data Science include four key aspects: data (data modeling, taxonomy, data management, data optimization, onthology, ethical and legal usage of data, etc.), databases, computer systems (transformation of inputs into outputs) and advanced analytics including statistics. The degree of advanced analytics that each of these typically employs is, at least, one significant difference between these two sub-fields. The utilization of advanced analytics in Data Science additionally demands methods for quality assurance or improvement (privacy, debiasing, fairness, accuracy, etc.) and as a result more advanced tools than in BA. This is also true for statistics. While BA does apply explorative statistics frequently, the use of inferential statistics is more common in Data Science.

Previously, and to some extent currently, BI has been described in the literature as an umbrella term that includes the full range of available strategies and technologies that enable data driven decision making (e.g., [16]). Based on this very broad definition, BA could be considered as subfields of BI. However, other authors (e.g., [17]) define BI more narrowly, as technologies that apply descriptive statistics to improve strategic decision making. Certainly within industry, BI tools used to implement what are referred to as Business Intelligence solutions provide functionality that focuses primarily on summarizing and visualizing the data found in data warehouses, using technologies such as OLAP (online analytical processing). Typically, these tools analyze data using descriptive statistical techniques and provide decision makers with visual tools to monitor business performance against a variety of KPIs. Using this narrower definition of BI, it can be considered as a sub-field of BA due to its application of descriptive statistics to aid decision making.

Taking a business perspective, the term Big Data Analytics has been described as a discipline that has emerged from BI [18]. Big Data Analytics concerns the organizing of big data, analyzing and discovering knowledge, patterns and intelligence from big data, visualization and the reporting of discovered knowledge for assisting decision making [19]. The authors claim that the main components of big data analytics include descriptive statistics, predictive analytics and prescriptive analytics.

Introducing the term Business Data Science

Analogous to the term Business Informatics in Computer Science, we propose to introduce the term *Business Data Science* (*BDS*). In our opinion, BDS is a sub-discipline of ADS, in which the methods and instruments of ADS are applied to issues in the business domain. Methods and instruments are adapted to the requirements of the domain and, if necessary, expanded. BDS can therefore be defined without reference to BI and BA. It is sufficient to concretize a general data science definition (e.g. [20]) for application to the business domain:

Business Data Science is a field of interdisciplinary expertise in which scientific procedures are used to (semi)automatically generate business insights from conceivably complex data leveraging existing or newly developed analysis methods. The gained business insights are subsequently utilized mainly for decision support, taking into account the effects on society.

The business domain is one of the most important application areas of DS [21], and ADS addresses a wide variety of business related problems, many of which require modified or even specifically created ADS methods and tools. The establishment of a separate sub-discipline of ADS therefore seems justified. Though the term Business Data Science is not used frequently yet, it appears in a growing number of publications (e.g., [22, 23, 24].

The relationship between BI, BA and BDS can be represented as follows. As described above, the aim of BI and BA is to support data-driven decisions in the business domain. In contrast to BI, BA also uses advanced and complex methods of statistics, information science and computer science. Historically, BI is the older discipline that was later supplemented/expanded by BA. BA continues to grow strongly as more and more questions in the business domain are being addressed with increasingly complex methods.

The goals and methods of BA and BDS are very similar and differ more in their point of view than in their content. BA represents more the internal perspective of the business domain. It focuses on the business benefits and sees the DS primarily as an auxiliary science. BDS, on the other hand, represents more of a domain-external perspective, focuses on the correct application of methods and sees the DS as a guiding discipline. BA and BDS tend to develop towards each other, so that the terms can be used synonymously in perspective.

4. Summary

Data Science can be divided into two sub-disciplines: *Pure Date Science*, in which tools and methods are developed, and *Applied Data Science*, in which these tools and methods are applied and adapted to practical problems of a specific domain. The focus of this manifesto is on the latter.

We have presented ADS in the context of its four neighboring disciplines: (1) PDS, (2) Focus dependent disciplines, (3) Domain and (4) Ethics & Regulations. It should be emphasized that ADS is more than the intersection of the neighboring disciplines. This aspect was illustrated by an appropriate graphical visualization and a concrete example.

Analogous to the subdivision of DS into the subdisciplines PDS and ADS, ADS can be decomposed into numerous domain-specific subdisciplines such as Business Data Science. It is useful – regarding their separating characteristics - to classify these sub-disciplines into two categories: *ADS in a scientific domain* and *ADS in a non-scientific domain*.

An important sub-discipline of ADS focusing on a non-scientific domain is BDS. Traditionally, the terms BI and BA are often used for data-driven analytical activities in the business context. While BDS can be easily distinguished from BI, there are major methodological similarities to BA. The main difference is the perspective on the problem to be analyzed. Besides BDS, other sub-disciplines of ADS exist such as Public Service Data Science or Engineering Data Science. A closer look at these sub-disciplines was not the focus of this manifesto.

The relationships between the various sub-disciplines has been summarized in **Figure 2**: Structuring of the Data Science discipline.

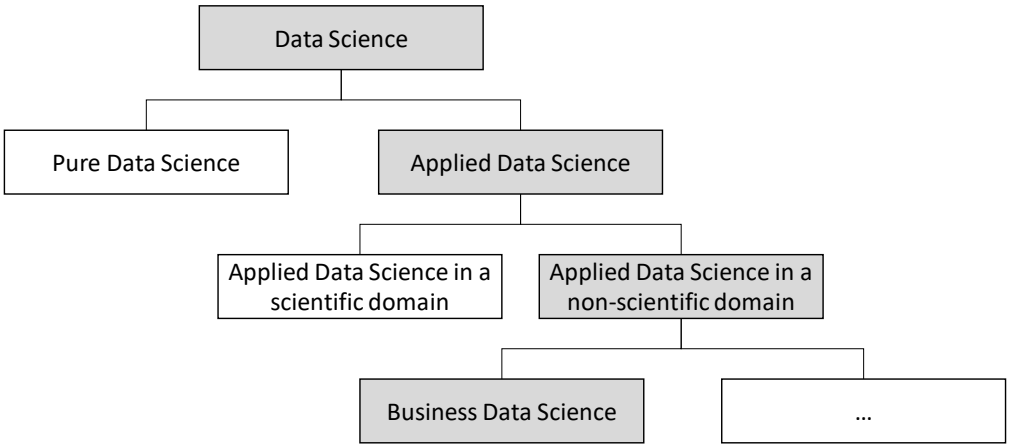

Figure 2: Structuring of the Data Science discipline

5. References

[1] Blei, D. M., and Smyth, P., "Science and data science", *Proceedings of the National Academy of Sciences*, 114.33 (2017): 8689-8692.

[2] Braschler, M., Stadelmann, T., and Stockinger, K., "Data science", in: Braschler M., Stadelmann T. and Stockinger K. (eds) Applied Data Science. Springer, Cham, 2019.

[3] Brodie, M., "What Is Data Science?", In: Braschler M., Stadelmann T. and Stockinger K. (eds) Applied Data Science. Springer, Cham, 2019.

[4] Longbing, C., "Data Science: A Comprehensive Overview", ACM Computing Surveys. 50 (2017): 1-42.

[5] Henry, R. and Venkatraman, S., "Big Data Analytics the Next Big Learning Opportunity", *Journal of Management Information and Decision Sciences,* 18.2 (2015): 17-29.

[6] Miller, S., "Collaborative Approaches Needed to Close the Big Data Skills Gap", *Journal of Organization Design*, 3.1 (2014): 26-30.

[7] Power, D.J., "Data science: supporting decision-making", *Journal of Decision Systems*, 25.4 (2016): 345-356.

[8] Kemper, S. and Mathews, T., "Earth Science Data Analytics: Definitions, Techniques and Skills", *Data Science Journal*, 16.6 (2017): 1-8.

[9] Phelps, A.L. and Szabat, K.A., "The Current Landscape of Teaching Analytics to Business Students at Institutions of Higher Education: Who is Teaching What?", *The American Statistician*, 71.2 (2017).

[10] Bichler, M., Heinzl, A., and Van der Aalst, W.M., "Business Analytics and Data Science: Once Again?", *Business & Information Systems Engineering*, 59.2 (2017): 77-79.

[11] Kambatla, K., Kollias, G., Kumar, V. and Grama, A., "Trends in big data analytics", *Journal of Parallel and Distributed Computing*, 74 (2014): 2561-2573.

[12] Laursen, G. H., and Thorlund, J., "Business analytics for managers: Taking business intelligence beyond reporting", John Wiley & Sons, 2016.

[13] Wixom, B., Ariyachandra, T., Douglas, D., Goul, M., Gupta, B., Iyer, L., and Turetken, O., "The current state of business intelligence in academia: The arrival of big data", *Communications of the Association for information Systems*, 34.1 (2014).

[14] Nelson, G., Difference between analytics and big data, data science and informatics. ThotWave Blog, 2017. Retrieved from https://www.thotwave.com/blog/2017/07/07/difference-between-analytics-and-bigdatadatascience-informatics/

[15] Davenport, T. H., and Harris, J. G., "Competing on analytics: The new science of winning", MA: Harvard Business School Press, 2007.

[16] Turban, E., Sharda, R., Delen, D. and King, D., "Business Intelligence: A Managerial Approach", Pearson, 2013.

[17] Kurniawan, Y., Gunawan, A., and Kurnia, S. G., "Application of Business Intelligence to Support Marketing Strategies: A case study approach", *Journal of Theoretical & Applied Information Technology*, *64*.1 (2014).
[18] Chen, H., Chiang, R. H., and Storey, V. C., "Business intelligence and analytics: From big data to big impact", *MIS quarterly* 36.4 (2012): 1165-1188.
[19] Sun, Z., Sun, L. and Strang, K., "Big Data Analytics Services for Enhancing Business Intelligence", *Journal of Computer Information Systems*, 58.2 (2018):162-169.
[20] Schulz, M., Neuhaus, U., Kaufmann, J., Kühnel, S., Alekozai, E.M., Rohde, H., Badura, D., Kerzel, U., Lanquillon, C., Daurer, S. Günther, M., Huber, L., Thiée, L.-W., zur Heiden, P., Passlick, J., Dieckmann, J., Schwade, F., Seyffarth, T., Badewitz, W., Rissler, R., Sackmann, S., Gölzer, P., Welter, F., Röth, J., Seidelmann, J., Haneke, U., „DASC-PM v1.1 - A Process Model for Data Science Projects, Elmshorn, 2022.
[21] Virkus, S. and Garoufallou, E., "Data Science from a Perspective of Computer Science", in *Research Conference on Metadata and Semantics Research* (pp. 209-219). Springer, Cham, 2019.
[22] Taddy, M., "Business Data Science: Combining Machine Learning and Economics to Optimize, Automate, and Accelerate Business Decisions", McGraw-Hill Education, New York ,2019.
[23] Davenport, T., "Beyond unicorns: Educating, classifying, and certifying business data scientists", Harvard Data Science Review, 2020.
[24] Miah, S. J., Solomonides, I., and Gammack, J. G., "A design-based research approach for developing data-focussed business curricula", *Education and Information Technologies* 25.1 (2020): 553–581.