The Science of Statistics versus Data Science: What is the Future?

Hossein Hassani, Christina Beneki, Emmanuel Sirimal Silva, Nicolas Vandeput and Dag Øivind Madsen

Abstract
The importance and relevance of the discipline of statistics with the merits of the evolving field of data science continues to be debated in academia and industry. Following a narrative literature review with over 100 scholarly and practitioner-oriented publications from statistics and data science, this article generates a pragmatic perspective on the relationships and differences between statistics and data science. Some data scientists argue that statistics is not necessary for data science as statistics delivers simple explanations and data science delivers results. Therefore, this article aims to stimulate debate and discourse among both academics and practitioners in these fields. The findings reveal the need for stakeholders to accept the inherent advantages and disadvantages within the science of statistics and data science. The science of statistics enables data science (aiding its reliability and validity), and data science expands the application of statistics to Big Data. Data scientists should accept the contribution and importance of statistics and statisticians must humbly acknowledge the novel capabilities made possible through data science and support this field of study with their theoretical and pragmatic expertise. Indeed, the emergence of data science does pose a threat to statisticians, but the opportunities for synergies are far greater.

Keywords: Perspective; science; statistics; data science; similarities; differences; pragmatism.

1. Introduction

In recent years, a growing debate in academia and industry has compared the importance and relevance of the discipline of statistics with the merits of the evolving field of data science (MacGillivray, 2021; Nachtsheim and Stufken, 2019; Ben-Zvi et al. 2018; Ribeiro et al. 2017; Davenport & Patil, 2012; Wickham, 2014; Wu, 1997). These debates have also extended to comparisons of software tools used for statistics and data science (Sardareh et al., 2021). While the discipline of statistics has a long history and is well established (Marquardt, 1987; Stigler, 1986), traditional statisticians have recently been overshadowed by the emergence of a new class of number crunchers – data scientists. Today, statistics is a profession that is both invaluable and invisible (Rodriquez, 2015) with data science being considered one of the hottest professions, and in the words of Davenport and Patil (2012) "the sexiest job of the 21st century."

Whilst the contribution of statistics to the progression of scientific knowledge across many disciplines continues to be acknowledged in this era of Big Data (McNutt, 2014), many organisations are actively seeking to employ data scientists. In fact, Baškarada and Koronios (2017) note that many organisations often seek "unicorn data scientists", a rare breed, almost mythical creatures that are experts in multiple specialties, from mathematics to computer science and artificial intelligence (AI). There are, however, commentators who remain critical and skeptical of these broad-based portrayals of data scientists as corporate saviours. For example, some researchers criticize data science and largely see it as a myth, suggesting instead a return to

https://doi.org/10.1016/j.techfore.2021.121111

conventional scientific approaches where data and methodology are just processual components (Learner and Phillips, 1993; Phillips, 2017).

Through a review of the relevant literature, this article aims to take stock of these differing views on the fields of statistics and data science. The goal is to generate a pragmatic perspective on the relationship and differences between statistics and data science.

Like Phillips (2017) who presents a perspective on Big Data, this article seeks to present a balanced and pragmatic perspective on the science of statistics and data science. Thus, this article can arguably be described as a "perspective article" since the overall aim is to discuss current debates and advances in these two fields and identify future directions both in academic research and in practice. Through critical analysis and discussion, the article holds the potential to stimulate the academic discourse about the pragmatic relationship that exists between these two fields, which goes far beyond semantic considerations. We subscribe to MacGillivray's (2021, p.S5) view that analysis and well-researched cautionary commentary by statisticians [*and data scientists*] can be extraordinarily valuable for both statistics and data science.

The article is conceptual and builds heavily on in-depth examinations of more than 100 scholarly and practitioner-oriented publications from statistics and data science. The literature is reviewed following a narrative review approach (Baumeister & Leary, 1997; Ferrari, 2015). As Ferrari (2015) notes, a narrative approach is particularly useful for appraising the current state of knowledge and for contributing to general debates in the research literature. That said, narrative reviews also suffer from some limitations, such as being not very explicit about the researchers' assumptions and biases related to selection and sampling of studies. While the authors of this article recognise the potential subjectivity of following such an approach, steps have taken to remedy these biases. For example, the authors hail from both the statistics and data science communities, which means that each field's underlying assumptions and viewpoints have been challenged during the research process.

The remainder of this article is structured as follows. Section 2 provides a background to statistics and data science and identifies some of the main distinguishing features of these fields. Section 3 discusses the challenges posed by the evolution of Big Data, while Section 4 discusses the limitations of data science. Section 5 follows a discussion of the distinctions between the two fields. Lastly, Section 6 concludes the article.

2. Background: Statistics and Data Science

The field of statistics has a rich and complex genealogy (Stigler, 1986). The term statistics is said to have originated around the year 1749 (Walker, 1929; Ribeiro et al., 2017). For example, Norton (1978) traces the development of the modern field of statistics to Karl Pearson and his work in mathematical biology and biometry. Today, the American Statistical Association defines statistics as "the science of learning from data, and of measuring, controlling and communicating uncertainty" (Ben-Zvi et al. 2018, p. 6). Statistical methods were developed for a world with scarce data where the lack of information called for models based on simplified assumptions to enable drawing conclusions from small datasets (Galeano and Pena, 2019). However, as Google's Chief Economist Hal Varian emphasises, the complexities inherent in modern world problems demands something more than statistics for understanding and extracting value from data (McKinsey Quarterly, 2009; Dayal, 2020). In Google searches worldwide (Figure 1, top graph), big data is the most popular search term

https://doi.org/10.1016/j.techfore.2021.121111

followed by data scientists, and statisticians respectively, thereby giving an indication of the interest or demand for the data scientist role in the modern world. However, it is noteworthy that Google Trends for the fields of 'statistics' and 'data science' shows the popularity of statistics over data science (Figure 1, bottom graph), perhaps fuelled by the increased number of statisticians than data scientists in the world. The trends indicate that whilst the role of a data scientist is increasingly more popular, as a field, the popularity of statistics continues to dominate over data science.

Over the years, the debate on the superiority of statistics and data science has resulted in varied views. Prof. Jeff Wu (1997) argued that "statistics" should be renamed "data science," but as Wickham (2014) explained, statistics is only part of data science, albeit a crucial part. The John Hopkins Data Science Specialisation[1] gives prominence to hypothesis testing, statistical model development and statistical inference as essential to the development of a data scientist (Ben-Zvi et al., 2018). Dunson (2018) noted that a significant portion of data science is not statistics. Even though, statistics not only supports but is also directly associated to data science, and statistical skills are very important for data scientists (Ribeiro et al., 2017), data science is closely connected to mathematics, statistics, and computer science (Saltz and Stanton, 2017).

The rise of 'Big Data' and 'data science' have given statistics a wake-up call (Breiman, 2001; Galeano and Pena, 2019) because the expansion of data science through the increasing availability of data and user-friendly software could result in the marginalisation of statistics (Ben-Zvi et al., 2018). Interestingly, as these two disciplines rely on a set of skills that often overlap (Diggle, 2015), data science and statistics frequently share distinguishing qualities. Nevertheless, some advocates of data science believe that you can be a good data scientist without a background in statistical theory (Granville, 2014; Davison, 2018). Others suggest that science is the only reality, and that data science is a myth as data and methodology are simply two of the four components that make up science (Phillips, 2017; Learner and Phillips, 1993). Some even questioned the longevity of the 'buzzword' that is data science (Press, 2013). However, data science is not merely a 'buzzword' and instead represents significant advances in capabilities for tackling highly complex challenges (MacGillivray, 2021). Challenges that statisticians alone would struggle to overcome given the growing volume, velocity, and variety of data.

[1] https://www.jhsph.edu/news/news-releases/2014/coursera-specialization.html

https://doi.org/10.1016/j.techfore.2021.121111

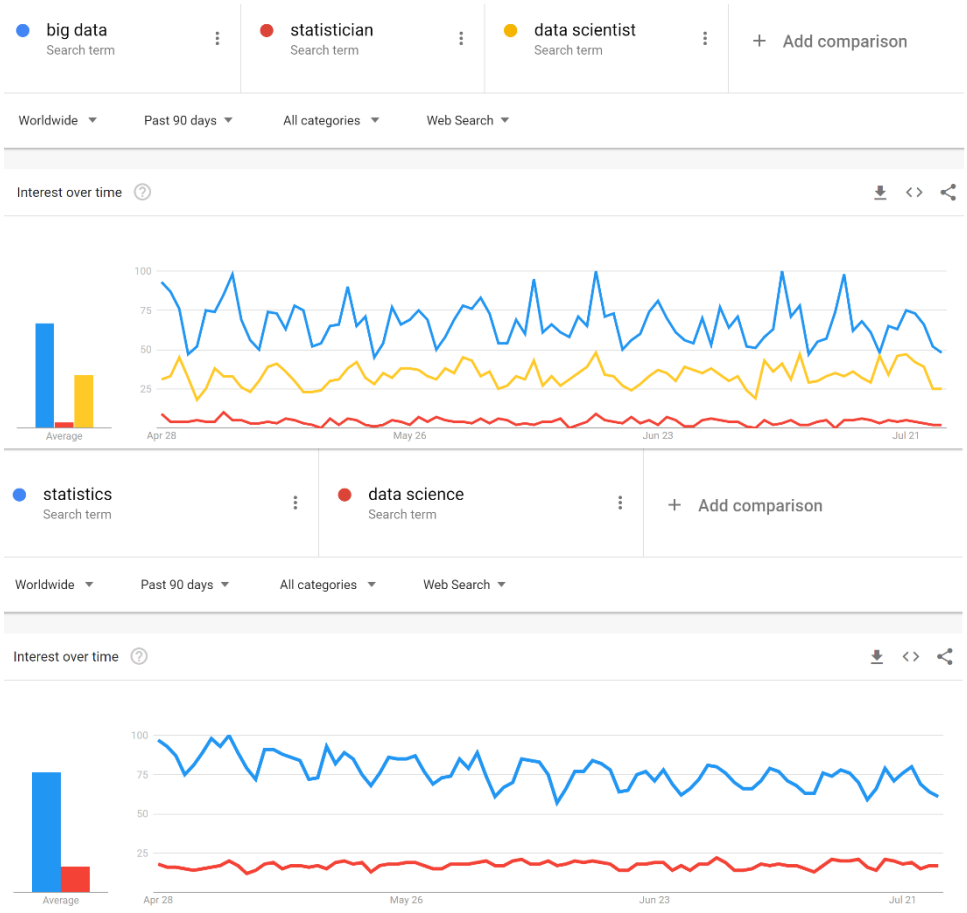

Figure 1. Worldwide search interest for 'big data', fields of 'statistics', and 'data science', and roles of 'statistician' and 'data scientist' over the last 90 days [accessed: 28.07.2021].

It is likely that if data science was to proceed without statistics, it would diminish both statistics and data science and worsen data-based decision-making in society (Ben-Zvi et al., 2018). Furthermore, in contrast to Granville and other advocates, Huang's (2019) view is that statistics is one of the three main data science skill sets (in addition to programming and business knowledge). For Huang (2019), the ability to use statistics to infer insights from smaller data sets onto larger populations is a fundamental law of data science. Patil (2018) supports this view as he notes the critical importance of understanding statistics, especially Bayesian probabilities for machine learning. Galeano and Pena (2019) note that machine learning has been successful through the integration of some methods developed for large data analysis with the methods in operations research, applied mathematics and statistics (for example, support vector machines, regularization methods, and network analysis). Rodriguez (2013) asserts that data scientists are in demand not only for their expertise in programming, machine learning, and strong communication skills, but also statistical modelling.

Some authors argue that most statisticians have not contributed much to recent progress in data science (van der Aalst, 2016). In fact, over the years, statisticians have been criticised for being too reliant on theory at the expense of computation (Carmichael and Marron, 2018). However, it is important to bear in mind that statisticians too have contributed to computation through software such as the R-Project (Members, 2017). Also, as Phillips (2017, p. 731) explains, "theory is a necessary

https://doi.org/10.1016/j.techfore.2021.121111

overlay for making sense of big data". As such, the incorporation of statistical theory within data science processes can bring some added reliability and validity to the findings and analysis. Furthermore, statisticians are also as infamous for creating complex models useful for solving well-defined problems based on assumptions that do not materialise in the real world. An argument that resonates somewhat with the famous quote by George Box that "all models are wrong; some models are useful" (Box et al. 2005, p. 440).

For example, in the field of inventory optimization, statisticians are infamous for creating extremely precise models that work under very restrictive assumptions (n.b., the existence of non-parametric methods that do not rely on assumptions) that fail when faced with 'real' supply chains. Nahmias (1979) summarises this aptly as he asserts that in the past, research has focused primarily on providing rigorous analysis of optimal policies for very simple problems, rather than focus on developing practical solutions to realistic problems.

Thus, as the emergence of data science has created a balance between theory and computation, the distinction between statisticians and non-statisticians has blurred (Cleveland, 2001). Historically, data analysis was associated with statisticians. However, the emergence of data science with its automation and machine learning has broken this barrier, enabling those who do not necessarily possess a background in statistics to also engage in meaningful data analysis. Even though automated software enables engagement with data analysis, the concept of garbage in, garbage out is very important to consider. Therefore, the rigour inherent within statisticians and statistical methods can provide a solid foundation for data scientists to ensure their output remains reliable and valid in practice.

In contrast to statistics, the story underlying data science began comparatively recently, over fifty years ago, when the American mathematician John Tukey referred to a novel science focused on learning from data (Tukey, 1962). Raban and Gordon (2020) carried out a search of Web of Science and found that there were more than 50 publications dealing with aspects of data science published during the 1960s, although it should be noted that the words "data" and "science" were not used conjointly in the titles of these articles. Moreover, Raban and Gordon (2020) found that the focus the research of this era was on "data collation in the social sciences, and not in a sense of extracting knowledge from data as referred to this area today" (p. 1565).

It was not until many years later that data science was formed into a field, when authors such as Cleveland (2001) and Wu (1997) started referring to the practices of Tukey and others as data science (Donoho, 2017; Raban and Gordon, 2020). The Data Science Association[2] defines data science as "the scientific study of the creation, validation and transformation of data to create meaning" and statistics as "the practice or science of collecting and analysing numerical data in large quantities." Herein lies the first hint of the relationship between statistics and data science, as the former definition appears to encompass the bread and butter of an applied statistician's daily routine (i.e., use methodology to make inferences from data) (Donoho, 2017). Nevertheless, data scientists tend to downplay the importance of the discipline of traditional statistics and intentionally obscure the evident overlap of the two fields (see for example, Granville 2014; Matteson 2020). For that reason, some statisticians feel that data science marginalises statistics (Donoho, 2017).

By contrast, De Veaux et al. (2017) noted that statistics forms part of the primary theoretical foundations of data science; Weihs and Ickstadt (2018) and Cao

[2] http://www.datascienceassn.org/code-of-conduct.html

https://doi.org/10.1016/j.techfore.2021.121111

(2017) noted that as a scientific discipline, data science is influenced by statistics. Professor Broman (2013) took a more authoritative stance in explicitly stating that data science is statistics, and anyone who analyses data is doing statistics. Carmichael and Marron (2018) added to the above ideology by pointing out that data science has its origins in statistics and is all about learning from data, which is traditionally the business of statistics.

However, one key difference is that statisticians are interested in developing models that are then confirmed by data. In contrast, data scientists are more interested in the application of machine learning and data mining without being restricted by models. Intellectuals such as Professor Andrew Gelman (2013), however, opines that statistics is the least important component of data science, and Hardin et al. (2015) asserted incorrectly (to the best of our knowledge) that the profession of statistics changed its name to data science! Overall, it is evident that there is disagreement between academics over the terminology, value, and contribution of both disciplines. It is our opinion that such extreme views from intellectuals in both disciplines are not aiding the collaborative advancement that is required for the benefit of statisticians and data scientists.

Whilst practitioners are more interested in user acceptance, results, and reliability, the M4-competition highlighted the need for statisticians and data scientists to work collaboratively. In particular, the M4-competition saw statisticians and data scientists compete at forecasting 100,000 time series. Interestingly, pure machine learning methods and pure statistical methods reported poor accuracy in relation to hybrid models that utilised both statistical and machine learning features (Makridakis et al., 2020). Whilst these findings were consistent with those in Makridakis et al. (2018), they differ from those in other machine learning studies such as Salaken et al. (2017). However, the results from machine learning studies claiming superior forecasting capabilities cannot be replicated or reproduced as the data and algorithms are not publicly available (Makridakis et al., 2020), thereby hindering its reliability and validity of the claims within machine learning studies.

Over a decade ago, Hal Varian predicted that the 'sexy' job in demand between 2009-2019 would be that of a statistician (Lohr, 2009; Davenport & Patil, 2012). However, a quick search of job opportunities on various platforms indicates that the number of roles for data scientists exceed the number for statisticians (in line with the Google Trends findings in Figure 1 where the data scientist role was more popular than that of a statistician). Thus, Hal Varian's prediction would appear flawed to those who do not see the complementary nature of statistics and data science. A closer look at the job adverts (Table 1) do uncover the continuing importance of statistics within the field of data science as evident in several roles that were advertised by high-profile companies. The authors of the current article do not suggest that all data scientist jobs in the market would follow a similar pattern, but below are few examples of excerpts from job adverts:

Table 1. Selected data scientist job adverts and excerpts from their descriptions.

Company	**Job Title**	**Excerpts from Job Descriptions**
Google	Data Scientist Engineering	Master's degree in a quantitative discipline (e.g**.,** Statistics, …), expertise with statistical data analysis such as linear models, skills in selecting the right

https://doi.org/10.1016/j.techfore.2021.121111

		statistical tools given a data analysis problem.
Dyson	Head of Data Science	Solid foundations on statistical and scientific methods.
Revolut	Head of Data Science	Deep understanding of fundamentals of probability and statistics.
Farfetch	Data Scientist	Master's degree, or higher, in a quantitative domain such as Mathematics, Operational Research, Statistics, or similar.
Facebook	Data Science Manager, Ads	Understanding of statistical analysis.
Playstation	Data Scientist	Solid theoretical and practical understanding of Statistics (e.g., hypothesis testing, experimentation, regressions).
Warner Bros. Entertainment	Data Scientist	Strong knowledge of statistical techniques.
Amazon	Data Scientist	Outstanding quantitative modelling and statistical analysis skills.
Deloitte UK	Data Scientist	Strong statistics skills including distributions, statistical testing, regression, etc.

Note: *These job adverts appeared on various online platforms during the year 2020.*

The debate about the relationship between statistics and data science is grounded in anecdote and is occasionally viewed as pointless or even non-sensical. The contentions have even given rise to the definition of a data scientist as someone who is better at statistics than any software engineer and better at software engineering than any statistician (Donoho, 2017). This attitude underlines the mounting issues regarding

appropriate definitions, assignments, and applications of these disciplines and problems around the incomplete understanding of what they involve (see for example, Carmichael and Marron (2018) and references therein). Importantly, this lack of clarity (Nantais, 2019) has arisen within the context of a world that, over the last decade, has placed data science and statistics at the heart of due process, research, and decision-making.

Some view the growth of data science as a threat to the long-term status of the discipline of statistics (Diggle, 2015) while others view data science as a challenging opportunity for statisticians (Ridgway, 2015). Barber (2018), for example, praises the relationship between statistics and data science. Van der Aalst (2016) included statistics as an ingredient contributing to data science (Figure 2), producing an outlook that supports the conclusion of Nantais (2019) that "data science is something more than statistics" and going so far as to claim that statistics can benefit from the emergence of data science. Van der Aalst (2016) also wrote that the discipline of data science is an amalgamation of classical disciplines such as statistics, data mining, databases, and distributed systems.

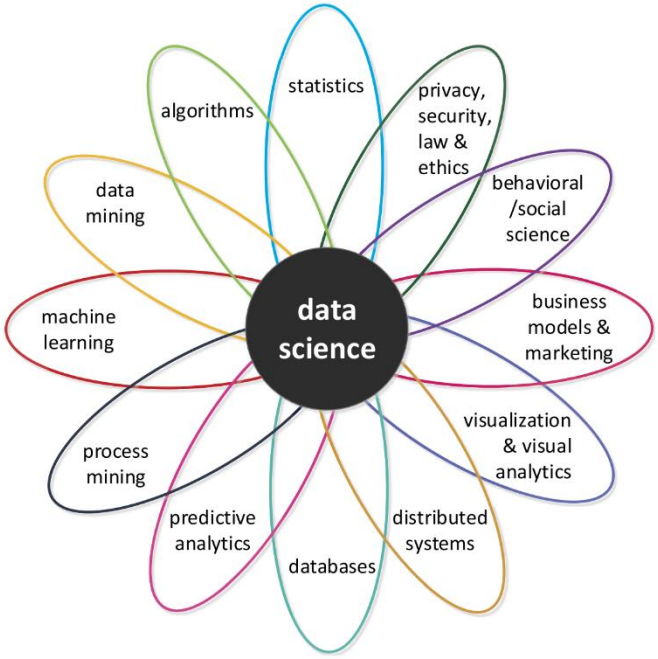

Figure 2. Data science ingredients (van der Aalst 2016, p. 12).

The modern world has undoubtedly come to embrace the importance of data (Bean, 2020), including Big Data (Mills, 2019), whatever the issue under investigation and this has led to an increased need for adequate data interpretation. In search of information, governments, corporations, academics, and other organisations have not only pushed to expand the pervasiveness of data-driven decision-making but have also allocated significant resources to data mining and the capture of still more information (see for example, Hassani et al., 2014; Geum et al. 2015; Li et al. 2019; Chen et al. 2018; Iqbal et al. 2020; McKinsey Analytics, 2018). It is unsurprising then, that technology companies are pushing technologies that can serve as a data source, notably, ambient intelligence, 'everyware', and the Internet of Things (see for example, Lo and Campos, 2018; Islam et al. 2020; Carayannis et al. 2018).

As technology evolves, so does the role and importance of the data scientist (Biswal, 2019). Data science specialists are in high demand throughout the public and

https://doi.org/10.1016/j.techfore.2021.121111

private sectors, across government departments and tech start-ups (Holak, 2019; Teichmann, 2019; Trivedi, 2018). Big Data has opened the door to creativity and innovation as well as scientific advances achieved through applied statistics and data science. As the variety, volume, and value of big data increases (Hassani et al. 2020), it becomes vital to identify the skills and qualities that data science and statistics should provide and re-think how we use and work with data. Moreover, this shift in thinking offers university systems worldwide an opportunity to update their current statistics curriculums and place a greater emphasis on computation skills, which are vital in the era of Big Data. The importance of ensuring that graduates are skilled at assessing data quality, finding meaning in data patterns, and understanding its business/social implications (Phillips, 2017) should not be underscored.

3. The Challenges Posed by the Evolution of Big Data

Evidence-based decision-making processes have always been heavily reliant on data collection. However, the development of more data collection procedures, mainly through ambient intelligence, is having different effects on the value, veracity, velocity, and volume of Big Data, which comprise Big Data's four distinguishing features. Typically, each of these features is supposed to be robust, and volume, value, veracity, and velocity are all expected to be high. The use of ambient intelligence in general or multi-disciplinary contexts creates a great deal of noise and little clarity. The data, information, knowledge, wisdom (DIKW) pyramid (Rowley, 2007) is one framework that is widely used to explain the inherent relationships within data. The application of this framework reveals that the rise in ambient intelligence does not result in greater knowledge or wisdom but simply more data as converting data into information becomes more and more difficult. It is up to the analyst or observer, that is, the person who utilises data to achieve a certain outcome, to address this challenge. The question of whether addressing this issue should fall under the purview of statistics, data science, or both may itself become a problem.

Academic institutions offering master's degrees often connect the importance of data science with the growth of Big Data (Donoho, 2017) instead of referring to the importance of statistics. However, it is noteworthy that statisticians (like data scientists are today) have been dealing with Big Data in the form of census data since the beginning and are comfortable with large datasets (Donoho, 2017; Carmichael and Marron, 2018). Nevertheless, the learning capabilities through data science is higher than with statistics as machine learning 'learns' more from data. Statisticians should appreciate the value generated by data science in scaling the application of statistics to Big Data via technology (Donoho, 2017; Greenhouse, 2013; Hardin et al., 2015). Such an appreciation would ensure that the next generation of Big Data analysts will be born out of the combination of sound statistical knowledge and data science skills that are mandatory for future employability and longevity of both disciplines.

In line with this argument, universities should give equal prominence to data science and statistics for two main reasons. First, Big Data can lead to inaccurate results by providing false positives at the hypothesis testing stage during statistical data analysis (McFarland and McFarland, 2015; Hassani and Silva, 2015). As such, a data scientist with mastery of machine learning, statistics, and analytics (Kozyrkov, 2018) can be extremely beneficial for businesses in such scenarios and can help ensure that the data analytics performed appreciates that common intuition does not always equate to mathematical correctness (Leetaru, 2019). Kozyrkov (2018) argues that a statistician is 'your best protection against fooling yourself in an uncertain world' as they

https://doi.org/10.1016/j.techfore.2021.121111

emphasize on determining whether the methods applied are apt for the problem and agonize over which inferences are valid from the information at hand.

The automation offered by machine learning generally comes at the expense of the rigorous process offered by statistics for data analysis through crucial steps such as sampling, exploratory and descriptive analysis, inference, prediction, measurement of uncertainty, and interpretation (Galeano and Pena, 2019). Here, it is important to note that data scientists would argue that they too provide careful consideration to inferences by relying on training, test and validation sets that do not require specific statistical or mathematical knowledge to identify issues.

However, Kozyrkov's (2018) emphasis is on the notion of combining statistics, machine learning and analytics to create a well-rounded data scientist. Such data scientists can help stakeholders take prudent risks by using the rigor of statistics to minimise the chance of unwise conclusions, be able to automate tricky tasks to pass the pure statistician's strict controls and have the necessary coding skills to visualise and mine Big Data with speed to uncover insights worthy of further investigation (Kozyrkov, 2018). In addition, business knowledge (as relevant to a field) is mandatory to avoid being fooled by data. Phillips (2017) asserts that the importance of context and motive for analytics will continue to remain mandatory even as data science matures. Galeano and Pena (2019) believe in the wider benefits from the convergence of machine learning and statistical approaches of data analysis under the data science umbrella.

Second, even though we live in an age of Big Data, not all problems are "Big Data problems" (for example, there are many small data problems in supply chain) where we continue to need statistics as a core foundation (Nantais, 2019). As an example of the importance of foundational statistics, Figure 3 shows a graph from an unpublished consultancy report on the ongoing COVID-19 pandemic. Many analyses of this crisis must rely on small sample sizes, and "machine learning often performs poorly on small datasets" (Faraway and Augustin 2018, p.144). In this case, a basic exponential model (with a series of additional assumptions, as statisticians do) was used to predict the emergence of COVID-19 cases in Sri Lanka over a 30-day period. The model was built on 28th March 2020 using only 18 data points, and it provided a considerably accurate forecast using simple foundational statistics.

This example showcases that in some situations, simple statistical models are still useful (Breiman, 2001; Koehrsen, 2019). This example does not intend to undermine the simultaneous contributions of data science during the pandemic (see for example, Marr, 2020), instead it offers evidence of the valuable role of statistics in an age of data science. In short, the authors believe that the science of statistics enables data science, and data science expands the application of statistics.

https://doi.org/10.1016/j.techfore.2021.121111

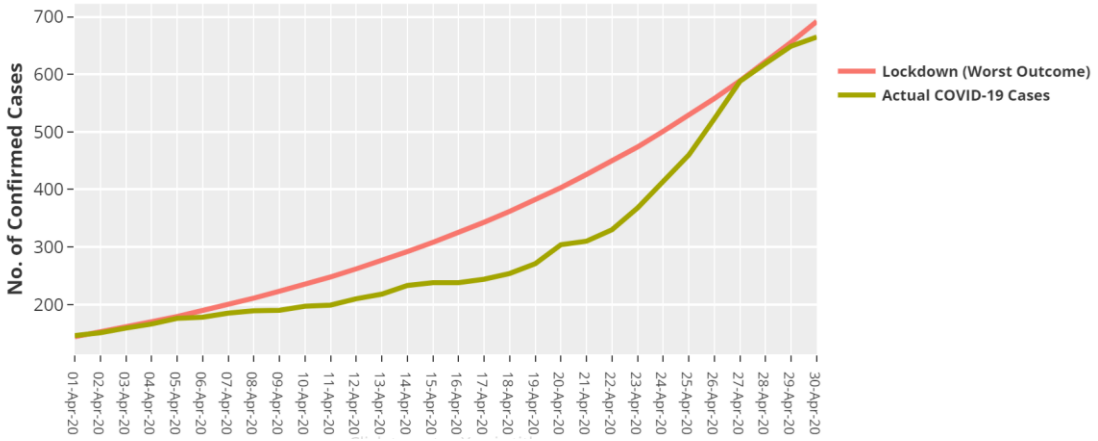

Figure 3. Daily forecast for COVID-19 cases in Sri Lanka[3].

This section ends with a strength-weaknesses-opportunities-threats (SWOT) analysis matrix (see e.g., Helms & Nixon, 2010). The SWOT matrix will be used as an organising framework to analyse the challenges posed by the evolution of Big Data. Here, the SWOT analysis is done from the perspective of a data scientist (Figure 4). A similar analysis from the perspective of a statistician is performed towards the end of Section 4.

As Figure 4 indicates, the field of data science has several strengths. First, data scientists are generally willing to adopt and apply new technologies. This is increasingly important in a world where the volume, velocity and veracity of data is expanding exponentially. Second, data science has expanded the application of statistics. Data scientists are also perceived as "Jacks of all trades" due to their skills across several areas, such as machine learning, statistics, and analytics. The flip side of being versatile and having knowledge in many areas, is of course that one may be "Master of none". Another weakness of data science is also that it is less rigorous than statistics. Since the field has not reached maturity, this might change over time. Finally, it may be too focused on data, and it is well-known that converting data into meaningful and actionable insights is often difficult.

There are, however, several opportunities and threats for the field of data science. One significant opportunity is related around the convergence of fields and technologies. For example, as discussed earlier, there is an on-going convergence of statistical analysis and machine learning methods. The general well-roundedness of data scientists may also position themselves to seize opportunities that may arise in the near future. Since technologies are developed at a fast pace, it is difficult to forecast where these fields will be some years from now. In terms of threats, there is a general realisation that not all problems are "Big Data problems" and that basic and foundational statistics may be sufficient to tackle the myriad of small(er) data problems. There is a risk that many will ask whether it is time to go "back to basics" (i.e., statistical analysis).

[3] Figure 3 was extracted from an unpublished consultancy report prepared by Emmanuel Sirimal Silva and RemediumOne (http://www.remediumone.com/) to provide insights into potential resource constraints during the COVID-19 pandemic.

https://doi.org/10.1016/j.techfore.2021.121111

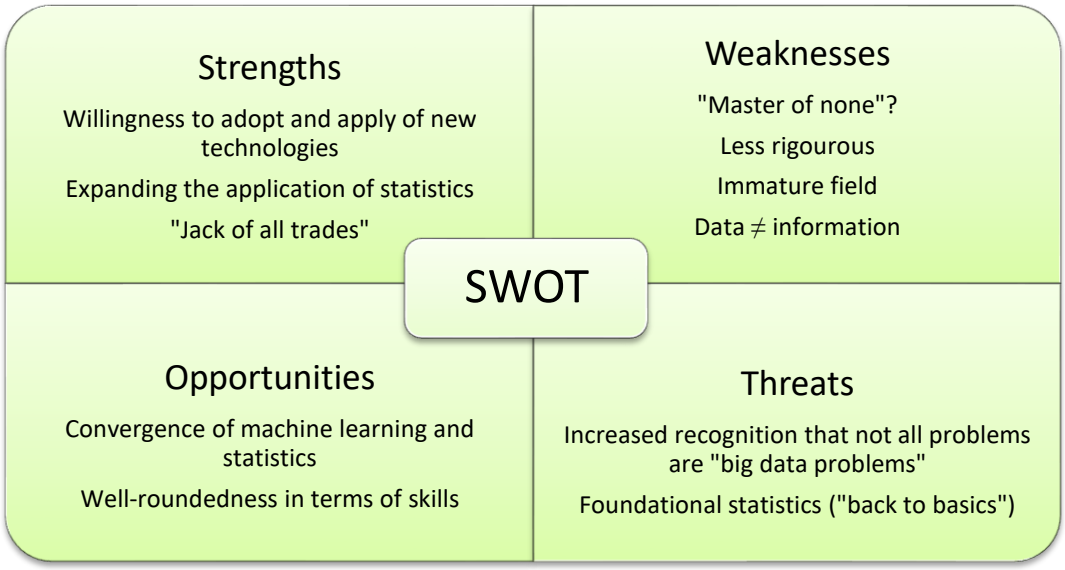

Figure 4. SWOT analysis from the perspective of data science.

4. The Limitations of Data Science

At first glance, data scientists appear to be the ideal specialists to come to grips with Big Data and its constantly evolving dynamics. However, they face barriers in terms of using datasets to model forecasts and achieve optimisation under certain conditions.

Machine learning experts have evaluated several options to overcome the barriers. For example, *k*-fold cross validation (Vandeput, 2020) is popular at present, but there is much potential through the development of 'AutoML' for overcoming the optimisation problem (Cronin, 2018). The idea underlying 'AutoML' is the generation of a machine learning algorithm that is then used to optimize the parameters of a second algorithm. Typically, data collection procedures are customised to respond to a set problem. Analytical and ethical reasoning skills, as well as a knowledge of data acquisition, archiving, and architecture are prerequisite tools that data science ultimately exploits to evaluate and present data. Data scientists often develop the necessary tools to interpret data or, alternately, if asked, may use pre-existing tools. Importantly, quite often it is not the processes used by data scientists that can cause errors but the data itself.

However, there are situations when the processes fail, resulting in data leakage (Pierre, 2018), which is one of the top ten machine learning mistakes (Nisbet et al., 2009) that can lead to flawed conclusions. In addition, overfitting (models with extremely low training errors but high testing errors) and underfitting (models with high training and testing errors) of models are also concerns for data scientists (Vandeput, 2020). Overfitting was identified as one of the shortcomings of pure machine learning methods during the M4 competition (Makridakis et al., 2020). Consequently, ensuring the quality of data is among a data scientist's challenges (Hazen et al., 2014; Ardagna et al., 2018; Alaoui and Gahi, 2019; Ghasemaghaei and Calic, 2019).

To determine the quality and thus suitability of data, it is vital to possess a full understanding of the origins of data and the data collection processes used. Statistics too is affected by data quality, but the rigor and accepted processes in place demands that statisticians pay careful attention to the design of data collection instruments, sampling bias, and the reliability and validity of the methods and data collected for analysis. Important information may be lost if it is inconsistently collected, or it may

https://doi.org/10.1016/j.techfore.2021.121111

be overlooked entirely. This incomplete set of data will lack predictive authority and will be unable to provide insights into the issue under investigation.

Understanding the trade-off between accuracy and interpretability is important for data analytics (Rane, 2018) and crucial when discussing the failures of data science. Interpretability is important for human curiosity and learning, finding meaning in the world, gaining knowledge, detecting bias, social acceptance, managing social interactions, and debugging and auditing (Molnar, 2020). Many of the complex data science led applications lack interpretability even though they produce highly accurate results (Rodriguez, 2018). See for example Figure 5.

Figure 5. Interpretability vs predictive power.

Statistical methods such as linear regression is highly interpretable but comparatively low in terms of predictive power whilst neural networks is an example of machine learning techniques that are highly accurate but lack interpretability. Depending on the problem at hand, one needs to determine whether they are interested in obtaining the best results or understanding how those results were produced (Rodriguez, 2018). The application of statistics is very useful where the objective calls for understanding relationships and deriving models that can interpret problems and generate forecasts (Galeano and Pena, 2019). Yet, relying on interpretability alone (for example using purely statistical models) can provide us with fairness, privacy, reliability, causality, and trust, at the expense of accuracy (Molnar, 2020).

However, it is important to remember that accuracy does not always give the full picture (Rawat, 2019) and interpretability is crucial unless the model has no significant impact or relates to a problem that has been well studied or is applied to situation where we are not concerned with potential manipulation of the system (Molnar, 2020). Thus, data scientists who also knowledgeable in statistics and statisticians who are equipped with data science skills are better placed to navigate the trade-off between accuracy and interpretability in practice. It also noteworthy that data scientists are now focusing on adding interpretability to machine learning to demystify the black box (Srinivasan, 2019). Murdoch et al. (2019) proposed the predictive, descriptive, relevant (PDR) framework for discussing interpretations whilst model-agnostic interpretation methods are another example of such efforts (Molnar, 2020).

Data scientists and statisticians are distinguished primarily by their different interests and approaches to problem solving, but aligned by their end goal, that is data analysis and prediction. It is expected that statistical analysis will continue to remain core to scientific modelling with well-structured data whilst machine learning and AI will succeed where relationships in data are not well understood (Galeano and Pena, 2019). However, it is no secret that AI is not yet advanced enough to identify anomalies in data that would require the expertise of ‘data preparers’ – a role that is increasingly important for ensuring ‘validity’ of big data (Phillips, 2017). As such, in a world where data analysis is becoming more valuable, existing statistical theory and methodology also becomes more valuable (Carmichael and Marron, 2018). Assuring the integrity of

https://doi.org/10.1016/j.techfore.2021.121111

data is becoming more and more important as ambient intelligence becomes increasingly pervasive as a new methodology for data collection is used.

The human brain is prone to finding patterns in random noise and this problem is even more prevalent in Big Data where the signal-to-noise ratio tends to be low (Silver, 2012; Hassani et al., 2015). Noise distracts us from the truth (Silver, 2012) and therefore data scientists can be misled if they fail to account for the signal and noise problem in Big Data. Here, statistical techniques such as decomposition and filtering can add value to a data scientist's tool kit when dealing with Big Data that requires differentiation between signal and noise (Galeano and Pena, 2019).

One might argue that the reiterative character of the statistical analysis process used to quantify data uncertainty puts statisticians in a stronger position to deal with the issues posed by Big Data. However, it is no secret that data scientists are better positioned to mine Big Data and thus, when their skills in machine learning are topped up with knowledge in statistics and business, then they will be better positioned to deal with the issues posed by Big Data. When building models, statisticians focus on the examination of the correlations, causality between variables, theory, and predictors, and emphasise the certainty of the applied parameters, as illustrated through margins of error or confidence intervals. In contrast, data scientists would be more interested in the prediction errors on a test set and the identification of which features to use.

Data science, like other scientific fields, should be precise in its identification and application of the correct tools to a problem. Sometimes the best tool for a use-case is statistics, exploratory data analysis or a simple yet understandable visualization of descriptive statistics. While understanding how AI and machine learning systems work is vital to a career in data science, these professionals often overlook the basics. For example, data science focuses on comparing many methods to create the best machine learning model while statistics instead seeks to improve a single, simple model to best suit the data. Overall, the main limitations of data science relate to small samples (Faraway and Augustin, 2018), and at predicting black swans (Taleb, 2007; Rodriguez, 2017) as the entire premise of machine learning is about learning from data. Furthermore, Dunson (2018) notes that automated methods (such as those increasingly used by data scientists) presents a lack of consideration of interpretability, quantification of uncertainty (or hypothesis testing), applications with limited training data, and selection bias.

Lastly, this section ends with a SWOT analysis from a statistician's perspective addressing the limitations of data science (Figure 6). The SWOT framework is used to examine the impact of the emergence of data science and Big Data on the future of the field of statistics, as well as the current status and future position of the statistics field.

As Figure 6 shows, the field of statistics is useful because it provides the foundation and oftentimes the basics are all that are needed, especially, when it comes to the analysis of structured data. Statistics also scores high in terms of interpretability. The flip side of this is of course that it is less useful for data mining and analysing unstructured data.

There are also opportunities as well as some threats looming in the horizon. For example, there are untapped potential to better apply statistics to Big Data via the use of technology. Data science also has the potential to expand the application of statistics. From a statistics point of view, the emergence and growth of the data science field is a threat, but one that has been around for a very long time (Diggle, 2015). For example, the new field of data science is generally perceived as being more relevant and having more real-world applications. In general, data scientists also seem to be more up-date

https://doi.org/10.1016/j.techfore.2021.121111

and ready to adopt new technologies such as AI and machine learning. For the field of statistics, it becomes important to be more adaptable and receptive to new technologies.

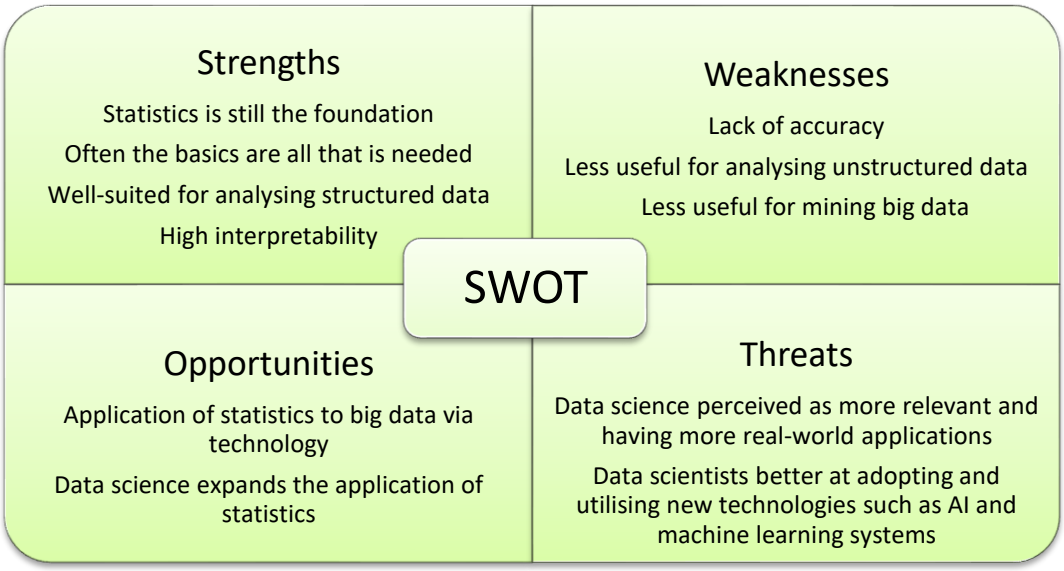

Figure 6. SWOT analysis from the perspective of statistics.

5. Discussion

This section discusses key distinctions between the fields of statistics and data science. These distinctions are summarised in Table 2. As reflected by the views of the various authors cited in the previous sections, the dividing line between the two fields is not clear. The emergence and evolution of Big Data has further blurred the distinction between these two fields.

In terms of theoretical origins, the field of statistics has deep roots that go back to the early work in mathematical biology and biometry (Norton, 1978), while data science is a relatively new field that builds directly on statistics and probability (Tayo, 2019). There are greater differences when it comes to the main foci of the two fields. Statistics strives for theoretical sophistication, while data science aims to provide practical solutions to real-world problems (van der Aalst, 2016). However, it is important to reflect whether practical solutions that cannot be theoretically supported add value to managerial decision making. This is because, as humans, we strive to obtain a deep understanding of phenomena and tend to prefer outcomes that can be explained.

This difference is also reflected in their main approaches. Statisticians focus on methodology/model development (Wild et al., 2018) and confirmation and prefer precise models with strict assumptions (Olhede and Wolfe, 2018). Data scientists, on the other, apply new techniques such as machine learning and try to avoid being restricted by models. Given that assumptions rarely hold in the real world, one would argue in favour of the data scientist's approach to analytics. However, statisticians too can rely on non-parametric approaches when assumptions are violated. The focus of model building is also different. Data scientists focus on prediction errors and identification of features whilst statisticians emphasize on the examination of correlations, causality between variables, theory, and predictors.

https://doi.org/10.1016/j.techfore.2021.121111

As noted earlier, another distinguishing feature is their relative emphasis on interpretability versus accuracy. The strength of statistics is interpretability, while for data sciences it is accuracy. The challenge and opportunity for statisticians and data scientists would be to collaborate on efforts at creating machine learning algorithms that are both interpretable and highly accurate. Another key difference is their preferred type of data for analysis. Statisticians prefer to work with well-structured data, while data science shines when the data are highly unstructured. Given that Big Data continues to generate large amounts of unstructured data that needs analysis, it is evident that data science skills are increasingly important on a day-to-day basis. The data wrangling skills which are a core focus of data science courses are of much value in the modern world. Statisticians can benefit by incorporating such skills and expertise into their own courses to ensure the skill set is developed. However, the application of machine learning is also becoming more accessible with automated algorithms enabling non-experts to benefit from its application and generation of output. Yet, interpretation of this output and the ability to make business sense of the data is key for analytics to be value adding.

The fact that statistics and data science can be complementary should be emphasised. Data science which relies on data mining and machine learning techniques are a mixture of statistics, AI, and searches in databases (Ribeiro et al., 2017; Gorunescu, 2011). The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology is a sound example which demonstrates the complementary nature of the two fields (Ribeiro et al., 2017). Przybyla (2020) notes several similarities between data scientists and statisticians, from the understanding of mathematics, investigating problems, exploratory data analysis, analysing trends, creating forecasts, visualisations, to reporting findings to non-technical users. Furthermore, there are two types of statistics called descriptive and inferential statistics (Singpurwalla, 2013; van der Aalst, 2016), and similarly, data mining is composed of two types of techniques called descriptive and predictive (Ribeiro et al., 2017). Saltz and Stanton (2017) presents the "four A's" of data science (i.e., data architecture, data acquisition, data analysis and data archiving) and notes that the analysis phase requires statistical aspects (Ribeiro et al., 2017). Therefore, despite the differences, the two fields share a common end goal. As such, it makes sense for experts in both fields to work collaboratively to develop data analysis and prediction capability further for the benefit of society. By working collaboratively with data scientists, statisticians can ensure statistics is placed firmly at the heart of data science and aid in protecting the identity of data science as an ongoing true science (MacGillivray, 2021). The complementary nature of data science and statistics was further emphasised by Weihs and Ickstadt:

> *Statistics is one of the most important disciplines to provide tools and methods to find structure in and to give deeper insight into data, and the most important discipline to analyse and quantify uncertainty…Finding structure in data and making predictions are the most important steps in Data Science. Here, in particular, statistical methods are essential since they are able to handle many different analytical tasks.*
>
> (Weihs and Ickstadt 2018, p.189-191)

Table 2. Key differences between statistics and data science.

	Statistics	**Data science**

https://doi.org/10.1016/j.techfore.2021.121111

Theoretical origins	Mathematical biology and biometry (Norton, 1978)	Statistics (De Veaux et al., 2017; Weihs and Ickstadt, 2018; Cao, 2017) and probability (Tayo, 2019)
Main focus	Theoretical sophistication (Carmichael and Marron, 2018; Olhede and Wolfe, 2018; van der Aalst, 2016)	Practical solutions to real problems (Cleveland, 2001; van der Aalst, 2016)
Main approach	Methodology/model development and confirmation (Wild et al., 2018) (precise models with strict assumptions, Olhede and Wolfe, (2018))	Application of machine learning and data mining (avoid being restricted by models) (Ribeiro et al., 2017; Gorunescu, 2011)
Focus of model building	Examination of correlations, causality between variables, theory, and predictors (Galeano and Pena, 2019)	Hyperparameter optimization and feature selection (Nantasenamat, 2020; Shaikh, 2018)
Interpretability vs. accuracy	High interpretability Low accuracy (Molnar, 2020; Olhede and Wolfe, 2018)	High accuracy Low interpretability (Olhede and Wolfe, 2018; Donoho, 2017; Hall, 2016; Rane, 2018; Rodriguez, 2018)
Preferred type of data	Well-structured data (Galeano and Pena, 2019; Olhede and Wolfe, 2018)	Unstructured data (Galeano and Pena, 2019; Olhede and Wolfe, 2018)
End goal	Data analysis and prediction	

6. Conclusions

Through a review of over 100 sources representing both fields of statistics and data science, this article developed a pragmatic perspective on the importance and relevance of the science of statistics in an age of data science. The research uncovered continuing debate and disagreement (in most cases) by statisticians and data scientists regarding the superiority of each other's disciplines. However, the evidence presented herewith clearly shows the growing need for and importance of positive collaborative efforts between data scientists and statisticians, as the science of statistics enables data science,

and data science expands the application of statistics. The SWOT analysis and discussions around the advantages and limitations around both disciplines further highlight the potential for synergies.

In summary, statisticians should embrace data science, approaching the collaboration with equal parts confidence in what statisticians can offer and humility to learn from the newer field (Diggle, 2015). Furthermore, the era of Big Data demands that statisticians broaden their understanding of statistical practice to be inclusive of all those who learn from data (Rodriguez, 2015). This is vital as data science has helped improve the reproducibility and communication of statistical outcomes, thereby adding to the reliability and validity of scientific studies (Carmichael and Marron, 2018). There is no doubt that in the absence of data science, statistics (as asserted by Diggle 2015) would make an essential but incomplete contribution in this age of Big Data. Likewise, data scientists should understand that statistics teaches the scientific method (Carmichael and Marron, 2018) that underlies data science. In addition, statisticians can develop new theories and methods to meet the upcoming challenges of data science (Olhede and Wolfe, 2018).

Furthermore, it is important to acknowledge that the frontier between the fields of statistics and data science is blurred and not easy to demarcate. In some cases, the two fields are indistinguishable from each other and therefore share a close association. For example, data scientists are expected to master statistics, machine learning and analytics (Kozyrkov, 2018), but statisticians themselves must align with data science or risk being left behind (Ben-Zvi et al. 2018). However, it is misleading to argue that data science is a rebranding of statistics (Carmichael and Marron, 2018), or vice versa. There is more to data science than statistics alone and vice versa (as there are problems that demand the application of statistics over data science). The need of the hour is for data scientists to genuinely appreciate statistics as an important element of data science, and for statisticians to celebrate the emergence of data science for making statistics more applicable and accessible across the globe (Carmichael and Marron, 2018).

At the end of the M4 forecasting competition, stakeholders concluded that the way forward was to exploit the advantages of both machine learning and statistical methods (Makridakis et al., 2020). He and Lin (2020) outlined 10 research challenge areas that have piqued the interest of statisticians from a data science perspective, and vice versa. Thus, universities have an active role to play in facilitating the exchange of ideas between statisticians and those aspiring to be data scientists to stimulate advances in all areas of knowledge (Galeano and Pena, 2019; Faraway and Augustin, 2018). The emergence and growth of Big Data ensures that data science will remain an extremely important and comparatively more popular field of study in relation to statistics. There are also interesting developments currently taking place in the field of data science. One notable example is the EDISON Data Science Framework (Manieri et al. 2015; Demchenko et al. 2016) which attempts to lay the foundation for the professionalisation of the data science field. This framework explicitly recognises the importance of statistics for data scientists.

Therefore, in the era of data science, statistics, “the most unselfish of science” (Rodriguez, 2013) is far from dead. Statistics lays the foundation for data science and adds to its reliability and validity whilst data science powers the application of statistics to Big Data through its incorporation of technology and AI. It is the authors’ view that a good data scientist could benefit from a solid (if not, at least foundational) theoretical and practical understanding of statistics, in addition to expertise in machine learning, analytics, and business knowledge (which differentiates a data scientist from a

https://doi.org/10.1016/j.techfore.2021.121111

statistician). As Brad Efron (2019) eloquently said whilst accepting the 2019 international prize in statistics:

> *"I tell my fretful friends that we have a strong positive regression coefficient with data science, as long as we remember not to let the inferential side of statistical thinking get lost in the excitement over new technology".*

So, what is the future? Based on the review, we foresee a future where the synergies made possible through the collaboration between statisticians and data scientists will drive reliable and valid data analytics and empower the continued relevance of both disciplines in a world where 'big' and 'small' data problems will continue to emerge.

Acknowledgements
The authors would like to acknowledge and thank the Editor, and the two anonymous referees for their constructive comments, guidance, and patience throughout the review process. The usual disclaimer applies. Dr. Hassani would also like to express his sincere gratitude to Professors Vahidi Asl, Faghihi, and Taheriyoun from Shahid Beheshti University (SBU), Iran for their valuable insights, comments, suggestions, and experiences shared with him on this topic.

References

Ardagna, D., Cappiello, C., Samá, W., & Vitali, M. (2018). Context-aware data quality assessment for big data. *Future Generation Computer Systems*, **89**, 548–562.

Alaoui, I. E., & Gahi, Y. (2019). The Impact of Big Data Quality on Sentiment Analysis Approaches. *Procedia Computer Science*, **160**, 803–810.

Barber, M. (2018). Data science concepts you need to know! Part 1. Towards Data Science. Available via: https://towardsdatascience.com/introduction-to-statistics-e9d72d818745 [Accessed: 09.05.2020].

Baškarada, S. and Koronios, A. (2017). Unicorn data scientist: the rarest of breeds. Program: electronic library and information systems, **51**(1), pp. 65-74

Baumeister, R. F., and Leary, M. R. (1997). Writing narrative literature reviews", *Review of General Psychology*, **1**(3), 311-320.

Bean, R. (2020). Now More Than Ever! – The Necessity Of Data, Analytics, And Expertise. Forbes. Available via: https://www.forbes.com/sites/ciocentral/2020/04/17/now-more-than-ever--the-necessity-of-data-analytics-and-expertise/#5dccc25a20f4 [Accessed: 10.05.2020].

Ben-Zvi, D., Makar, K., and Garfield, J. (2018). *International Handbook of Research in Statistics Education*, Springer International Publishing.

Biswal, M. (2019). How Technology Is Changing How We Treat The Elements Of Data Science. Medium. Available via: https://medium.com/@minatibiswal_85870/how-technology-is-changing-how-we-treat-the-elements-of-data-science-486f3f90c97 [Accessed: 10.05.2020].

Box, G., Hunter, J. S., and Hunter, W. G. (2005). *Statistics for Experimenters.* 2nd ed., John Wiley & Sons.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, **16**(3), 199–231.

Broman, K. (2013). Data science is statistics. Blog post. Available via: https://kbroman.wordpress.com/2013/04/05/data-science-is-statistics/ [Accessed: 09.05.2020].

Cao, L. (2017). Data Science: A Comprehensive Overview. *ACM Computing Surveys*, **50**(3), 43:1-43:42.

Carmichael, I., and Marron, J. S. (2018). Data science vs. statistics: two cultures? *Japanese Journal of Statistics and Data Science*. **1**, 117-138.

Carayannis, E. G., Del Guidice, M., and Soto-Acosta, P. (2018). Disruptive technological change within knowledge-driven economies: The future of the Internet of Things (IoT). *Technological Forecasting and Social Change*, **136**, 265-267.

Chen, C. P., Weng, J.-Y., Yang, C.-S., and Tseng, F.-M. (2018). Employing a data mining approach for identification of mobile opinion leaders and their content usage patterns in large telecommunications datasets. *Technological Forecasting and Social Change*, **130**, 88-98.

Cleveland, W. (2001). Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistics Review*, **69**, 21-26.

Cronin, S. K. (2018). What's auto ML? Available via: https://towardsdatascience.com/whats-auto-ml-b457d2710f9d [Accessed: 23.05.2020].

Davenport, T. H., and Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*, October. Available via: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century [Accessed: 22.05.2020].

Davison, J. (2018). No, Machine Learning is not just glorified Statistics. Available via: https://towardsdatascience.com/no-machine-learning-is-not-just-glorified-statistics-26d3952234e3 [Accessed: 22.05.2020].

Dayal, V. (2020). *Quantitative Economics with R: A Data Science Approach*. Springer Nature, Singapore.

Demchenko, Y., Belloum, A., Los, W., Wiktorski, T., Manieri, A., Brocks, H., ... & Brewer, S. (2016, December). EDISON data science framework: a foundation for building data science profession for research and industry. In *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)* (pp. 620-626). IEEE.

De Veaux et al. (2017). Curriculum Guidelines for Undergraduate Programs in Data Science. *Annual Review of Statistics and its Application*, **4**, 15-30.

https://doi.org/10.1016/j.techfore.2021.121111

Diggle, P. J, (2015). Statistics: a data science for the 21st century. *Journal of the Royal Statistical Society (Statistics in Society: Series A)*, **178**(Part 4), 793-813.

Donoho, D. (2017). 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, **26**(4), 745-766.

Dunson, D. B. (2018). Statistics in the big data era: Failures of the machine. *Statistics & Probability Letters*, **136**, 4-9.

Efron, B. (2019) Acceptance speech, 2019. Available via: https://statprize.org/pdfs/2019-Efront-AcceptanceSpeech.pdf (Accessed: 22.07.2021).

Faraway, J. J., and Augustin, N. H. (2018). When small data beats big data. *Statistics & Probability Letters*, **136,** 142-145.

Galeano, P., and Pena, D. (2019). Data science, big data and statistics. *TEST*, **28**, 289-329.

Gelman, A. (2013). Statistics is the least important part of data science. Blog post. Available via: https://statmodeling.stat.columbia.edu/2013/11/14/statistics-least-important-part-data-science/ [Accessed: 09.05.2020].

Geum, Y., Lee, H., Lee, Y., and Park, Y. (2015). Development of data-driven technology roadmap considering dependency: An ARM-based technology roadmapping. *Technological Forecasting and Social Change*, **91**, 264-279.

Ghasemaghaei, M., & Calic, G. (2019). Can big data improve firm decision quality? The role of data quality and data diagnosticity. *Decision Support Systems*, **120**, 38-49.

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques.* Springer-Verlag Berlin Heidelberg.

Granville, V. (2014). Data science without statistics is possible, even desirable. Data Science Central. Available via: https://www.datasciencecentral.com/profiles/blogs/data-science-without-statistics-is-possible-even-desirable [Accessed: 09.05.2020].

Greenhouse, J. B. (2013). Statistical Thinking: The Bedrock of Data Science. Huffpost. Available via: https://www.huffpost.com/entry/statistical-thinking-the-bedrock-of-data-science_b_3651121 [Accessed: 09.05.2020].

Hall, P. (2016). Predictive modeling: Striking a balance between accuracy and interpretability. Available via: https://www.oreilly.com/content/predictive-modeling-striking-a-balance-between-accuracy-and-interpretability/ [Accessed: 26.07.2021].

Hassani, H., Saporta, G., and Silva, E. S. (2014). Data mining and official statistics: the past, the present and the future, *Big Data*, **2**(1), 34-43.

Hassani, H., and Silva, E. S. (2015). Forecasting with Big Data: A Review. *Annals of Data Science*, **2**(1), 5-19.

https://doi.org/10.1016/j.techfore.2021.121111

Hassani, H., Silva, E. S., Unger, S., TajMazinani, M., and Mac Feely, S. (2020). Artificial Intelligence (AI) or Intelligence Augmentation (IA): What Is the Future? *AI,* **1,** 143-155.

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., Murrell, P., Peng, R., Roback, P., Temple Lang, D., and Ward, M. D. (2015). Data Science in Statistics Curricula: Preparing Students to "Think with Data". *The American Statistician,* **69**(4), 343-353.

Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, **154**, 72–80.

He, X., and Lin, X. (2020). Challenges and Opportunities in Statistics and Data Science: Ten Research Areas. *Harvard Data Science Review*, **2.3**, https://doi.org/10.1162/99608f92.95388fcb

Helms, M. M., & Nixon, J. (2010). Exploring SWOT analysis–where are we now?. *Journal of Strategy and Management*, **3**(3), 215-251.

Holak, B. (2019). Demand for data scientists is booming and will only increase. SearchBusinessAnalytics. Available via: https://searchbusinessanalytics.techtarget.com/feature/Demand-for-data-scientists-is-booming-and-will-increase [Accessed: 09.05.2020].

Huang, R. (2019). How to Learn Data Science Without a Degree. Available via: https://www.springboard.com/blog/learn-data-science-without-degree/ [Accessed: 22.05.2020].

Islam, N., Marinakis, Y., Majadillas, M. A., Fink, M., and Walsh, S. T. (2020). Here there be dragons, a pre-roadmap construct for IoT service infrastructure. *Technological Forecasting and Social Change,* **155**, 119073.

Iqbal, R., Doctor, F., More, B., Mahmud, S., and Yousuf, U. (2020). Big data analytics: Computational intelligence techniques and application areas. *Technological Forecasting and Social Change,* **153**, 119253.

Koehrsen, W. (2019). Thoughts on the Two Cultures of Statistical Modeling. Available via: https://towardsdatascience.com/thoughts-on-the-two-cultures-of-statistical-modeling-72d75a9e06c2 [Accessed: 22.05.2020].

Kozyrkov, C. What Great Data Analysts Do — and Why Every Organization Needs Them. *Harvard Business Review*, Available via: https://hbr.org/2018/12/what-great-data-analysts-do-and-why-every-organization-needs-them [Accessed: 09.05.2020].

Learner, D. B., and Phillips, F. Y. (1993). Method and Progress in Management Science. *Socio-Economic Planning Sciences*, **27**(1), 9–24.

Leetaru, K. (2019). How Data Scientists Turned Against Statistics. Forbes. Available via: https://www.forbes.com/sites/kalevleetaru/2019/03/07/how-data-scientists-turned-against-statistics/#15777d91257c [Accessed: 09.05.2020].

Li, X., Xie, Q., Jiang, J., Zhou, Y., and Huang, L. (2019). Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. *Technological Forecasting and Social Change*, **146**, 687-705.

Lo, F.-Y., and Campos, N. (2018). Blending Internet-of-Things (IoT) solutions into relationship marketing strategies. *Technological Forecasting and Social Change*, **137**, 10-18.

Lohr, S. (2009). For Today's Graduate, Just One Word: Statistics. The New York Times. Available via: https://www.nytimes.com/2009/08/06/technology/06stats.html?_r=1& [Accessed: 09.05.2020].

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, **36**(1), 54-74.

Makridakis S., Spiliotis E., and Assimakopoulos V. (2018). Statistical and machine learning forecasting methods: concerns and ways forward. *PLoS One*, **13**(3), 1-26.

Manieri, A., Brewer, S., Riestra, R., Demchenko, Y., Hemmje, M., Wiktorski, T., ... & Frey, J. (2015, November). Data Science Professional uncovered: How the EDISON Project will contribute to a widely accepted profile for Data Scientists. In *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)* (pp. 588-593). IEEE.

Marr, B. (2020). Coronavirus: How Artificial Intelligence, Data Science And Technology Is Used To Fight The Pandemic. Forbes. Available via: https://www.forbes.com/sites/bernardmarr/2020/03/13/coronavirus-how-artificial-intelligence-data-science-and-technology-is-used-to-fight-the-pandemic/#1aa797915f5f [Accessed: 10.05.2020].

Marquardt, D. W. (1987). The importance of statisticians. *Journal of the American Statistical Association*, **82**(397), 1-7.

Matteson, S. (2020). How to become a data scientist without getting a Ph.D., TechRepublic. Available via: https://www.techrepublic.com/article/how-to-become-a-data-scientist-without-getting-a-ph-d/ [Accessed: 09.05.2020].

McFarland, D. A., and MacFarland, H. R. (2015). Big Data and the danger of being precisely inaccurate. *Big Data & Society*, (July – December), 1-4.

MacGillivray, H. (2021). Statistics and data science must speak together. *Teaching Statistics*, **43**, S5–S10.

McKinsey Analytics. (2018). Analytics comes of age. McKinsey & Company. Available via:

https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Analytics/Our%20Insights/Analytics%20comes%20of%20age/Analytics-comes-of-age.ashx [Accessed: 10.05.2020].

McKinsey Quarterly. (2009). Hal Varian on how the Web challenges managers. Available via: https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/hal-varian-on-how-the-web-challenges-managers [Accessed: 22.05.2020].

McNutt, M. (2014). Raising the Bar, *Science*, **345**(6192), 9.

Members, R. P. (2017). The r project for statistical computing. Available via: https://www.r-project.org/ [Accessed: 22.05.2020].

Mills, T. (2019). Why Big Data And Machine Learning Are Important In Our Society. Forbes. Available via: https://www.forbes.com/sites/forbestechcouncil/2019/01/07/why-big-data-and-machine-learning-are-important-in-our-society/#6eeec6097aa2 [Accessed: 10.05.2020].

Molnar, C. (2020). *Interpretable machine learning: A guide for making black box models explainable.* Available via: https://christophm.github.io/interpretable-ml-book/index.html [Accessed: 22.05.2020].

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *PNAS*, **116**(44), 22071-22080.

Nachtsheim, A. C., and Stufken, J. (2019). Comments on: Data science, big data and statistics. *TEST*, **28,** 345-348.

Nahmias, S. (1979). Simple Approximations for a Variety of Dynamic Leadtime Lost-Sales Inventory Models. *Operations Research*, **27**(5), 857-1066.

Nantais, J. (2019). Data Science or Statistics? Towards Data Science. Available via: https://towardsdatascience.com/data-science-or-statistics-9e826ebf7fe2 [Accessed: 09.05.2020].

Nantasenamat, C. (2020). How to Build a Machine Learning Model: A visual guide to learning data science. Available via: https://towardsdatascience.com/how-to-build-a-machine-learning-model-439ab8fb3fb1 [Accessed: 28.07.2021].

Nisbet, R., Elder, J., and Miner, G. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press.

Norton, B. J. (1978). Karl Pearson and statistics: The social origins of scientific innovation. *Social Studies of Science*, **8**(1), 3-34.

Olhede, S. C., and Wolfe, P. J. (2018). The future of statistics and data science. *Statistics & Probability Letters*, **136**, 46-50.

https://doi.org/10.1016/j.techfore.2021.121111

Patil, A. (2018). How to self-learn statistics of data science. Available via: https://medium.com/ml-research-lab/how-to-self-learn-statistics-of-data-science-c05db1f7cfc3 [Accessed: 22.05.2020].

Phillips, F. (2017). A perspective on 'Big Data'. *Science and Public Policy*, **44**(5), 730-737.

Pierre, R. (2018). Data Leakage, Part I: Think You Have a Great Machine Learning Model? Think Again. Available via: https://towardsdatascience.com/data-leakage-part-i-think-you-have-a-great-machine-learning-model-think-again-ad44921fbf34 [Accessed: 22.05.2020].

Press, G. (2013). Data Science: What's The Half-Life Of A Buzzword? Available via: https://www.forbes.com/sites/gilpress/2013/08/19/data-science-whats-the-half-life-of-a-buzzword/ [Accessed: 22.07.2021].

Przybyla, M. (2020). The Difference Between Data Science and Statistics: Which role are you, should you change careers? Available via: https://towardsdatascience.com/the-difference-between-data-science-and-statistics-168c7062c201 [accessed: 26.02.2021].

Raban, D. R., & Gordon, A. (2020). The evolution of data science and big data research: A bibliometric analysis. *Scientometrics*, *122*(3), 1563-1581.

Rawat, S. (2019). Is accuracy EVERYTHING? Available via: https://towardsdatascience.com/is-accuracy-everything-96da9afd540d [Accessed: 22.05.2020].

Rane, S. (2018). The balance: Accuracy vs. Interpretability. Available via: https://towardsdatascience.com/the-balance-accuracy-vs-interpretability-1b3861408062 [Accessed: 22.05.2020].

Ribeiro, V., Rocha, A., Peixoto, R., Portela F. and Santos, M. F. (2017). Importance of Statistics for Data Mining and Data Science. *In:* 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), 2017, pp. 156-163, doi: 10.1109/FiCloudW.2017.86.

Ridgway, J. (2015). Implications of the data revolution for statistics education. *International Statistical Review*, **84**(3), 528–549.

Rowley, J. (2007). The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, **33**(2), 163-180.

Rodriguez, R. N. (2013). The 2012 ASA Presidential Address: Building the big tent for statistics. *Journal of the American Statistical Association*, **108**(501), 1–6.

Rodriguez, R. N. (2015). Who Will Celebrate Our 200th Anniversary? Growing the Next Generation of ASA Members, *The American Statistician*, **69**, 91–95.

Rodriguez, J. (2017). The Black Swan Problem in Artificial Intelligence: Part I. Available via: https://medium.com/@jrodthoughts/the-black-swan-problem-in-artificial-intelligence-part-i-74306aee0156 [Accessed: 23.05.2020].

https://doi.org/10.1016/j.techfore.2021.121111

Rodriguez. J. (2018). Interpretability vs. Accuracy: The Friction that Defines Deep Learning. Available via: https://towardsdatascience.com/interpretability-vs-accuracy-the-friction-that-defines-deep-learning-dae16c84db5c [Accessed: 22.05.2020].

Salaken S.M., Khosravi A., Nguyen T., and Nahavandi S. (2017). Extreme learning machine based transfer learning algorithms: A survey. *Neurocomputing,* **267**, 516-524.

Saltz, J. S., and Stanton, J. M. (2017). *An Introduction to Data Science.* SAGE Publications.

Sardareh, S.A., Brown, G.T.L., and Denny, P. (2021). Comparing four contemporary statistical software tools for introductory data science and statistics in the social sciences. *Teaching Statistics,* **43**, S157-S172.

Shaikh, R. (2018). Feature Selection Techniques in Machine Learning with Python. Available via: https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e [Accessed: 28.07.2021].

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail – but Some Don't.* The Penguin Press, New York.

Singpurwalla, D. (2013). *A Handbook of Statistics: An Overview of Statistical Methods*, Bookboon.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.

Srinivasan, P. (2019). Interpretable Machine Learning: An attempt to demystify the black-box. Available via: https://medium.com/walmartlabs/accuracy-vs-interpretability-paradox-382803f6a99d [Accessed: 22.05.2020].

Taleb, N. N. (2007). *The Black Swan: The Impact of the Highly Improbable.* Random House Trade Paperbacks, New York.

Tayo, B. O. (2019). Theoretical Foundations of Data Science— Should I Care or Simply Focus on Hands-on Skills? Available via: https://towardsdatascience.com/theoretical-foundations-of-data-science-should-i-care-or-simply-focus-on-hands-on-skills-c53fb0caba66 [Accessed: 26.02.2021].

Teichmann, J. (2019). The increasing demand for data scientists. An interview. Towards Data Science. Available via: https://towardsdatascience.com/the-increasing-demand-for-data-scientists-an-interview-6d74d98afba0 [Accessed: 09.05.2020].

Trivedi, A. (2018). Why data science jobs are in high demand? Medium. Available via: https://medium.com/cutshort/why-data-science-jobs-are-in-high-demand-c1b5614d3083 [Accessed: 09.05.2020].

Tukey, J. (1962). The future of data analysis. *The Annals of Mathematical Statistics,* **33**, 1-67.

Vandeput, N. (2020). *Data Science for Supply Chain Forecasting*. 2nd Edition. De Gruyter.

https://doi.org/10.1016/j.techfore.2021.121111

van der Aalst, W. (2016) Data Science in Action. *In: Process Mining.* Springer, Berlin, Heidelberg.

Walker, H. M. (1929). *Studies in the history of statistical method: With special reference to certain educational problems.* Williams & Wilkins Co.

Weihs, C., and Ickstadt, K. (2018). Data Science: the impact of statistic. *International Journal of Data Science and Analytics,* **6**, 189-194.

Wickham, H. (2014). Data science: how is it different to statistics? Institute of Mathematical Statistics. Available via: https://imstat.org/2014/09/04/data-science-how-is-it-different-to-statistics%E2%80%89/ [Accessed: 09.05.2020].

Wild C. J., Utts J. M., and Horton N. J. (2018). What Is Statistics?. *In:* Ben-Zvi D., Makar K., Garfield J. (eds) *International Handbook of Research in Statistics Education.* Springer International Handbooks of Education. Springer, Cham.

Wu, J. (1997). Statistics = Data Science? Inaugural lecture for the Carver Chair. Available via: https://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf [Accessed: 09.05.2020].

Bibliography

Dr. Hossein Hassani is the Statistical Systems Co-ordinator at the Organisation for Petroleum Exporting Countries in Vienna, Austria.

Dr. Christina Beneki is an Associate Professor at the Department of Tourism, Ionian University in Greece.

Dr. Emmanuel Sirimal Silva is Head of Research Coordination: Fashion Business School at London College of Fashion, University of the Arts London.

Nicolas Vandeput is a supply chain data scientist specialized in demand forecasting and inventory optimization.

Dr. Dag Øivind Madsen is a Professor at the USN School of Business, University of South-Eastern Norway.

https://doi.org/10.1016/j.techfore.2021.121111