Similarities and Differences Between Traditional Statistics and DSML: Implications for Data Scientists

Chong Ho Yu & Charlene Yang

Paper presented at 2023 IDEAS Global Conference, Online

November 18, 2023

Abstract

Data Science and Machine Learning (DSML) are taking over the world by storm, but the roots of DSML and their relationship to traditional statistics remain understudied, leading to misconceptions and potential misapplications. For example, although linear regression, which is based on the least square criterion discovered in 1805, is known to be unsuitable for big data analytics, it is still categorized as a DSML method by some data scientists. This presentation illustrates how the emergence of DSML as a reaction to the shortcomings of classical statistics by covering four major sources that contributed to the development of DSML: (1) Exploratory Data Analysis (EDA) advocated by John Turkey, (2) The notion of learning from the data proposed by John Chambers, (3) Data visualization and advanced computing methods developed by William Cleveland, and (4) the two-culture thesis suggested by Leo Brieman. In addition, differences between DSML and traditional statistics in seven aspects are discussed: (1) Dichotomous evidence and decision vs. pattern recognition and contextual decision, (2) Model-driven and assumption-based vs data-driven and assumption-free, (3) Single-modeling vs multiple-modeling, (4) Whole sample vs resampling and subsetting, (5) Small data vs. big data, (6) Inference and explanation vs prediction, and (7) Overall generalization vs. personalized recommendation. Nonetheless, although DSML methods are considered more versatile, powerful, and efficient than traditional statistics in many aspects, it does not necessarily imply that the former can completely supersede the latter. A number of examples are given to illustrate when DSML can outperform traditional statistics, and vice versa.

Data Science: Data Analytics, Computer Science, and Domain Knowledge

Although there are many definitions of data science, it is generally agreed that data science is an interdisciplinary field that integrates data analytics/statistics, computation, and domain knowledge (Blei & Smyth, 2017; Conway, 2010; Cleveland, 2001; Yu, 2014). The term “data science” was coined as early as the 1970s, and it started gaining popularity around the second decade of the 21st century as both data size and variety have been constantly growing. Given that data might be sourced from many channels, it is essential for the analyst to utilize

powerful computer technologies for complex data processing. In addition, as data analytics has become an integral component of many academic disciplines and industries, understanding the context of the data is imperative for meaningful analysis and interpretation.

The theory of interdisciplinary data science has been put into practice by several institutions. For example, the Michigan Institute for Data Science at the University of Michigan (2020) facilitates collaborative research across the departments of medical science, public health, engineering, chemistry, economics, psychology, and computer science. By the same token, the Data Sciences Initiative established by North Carolina State University (2023) also coordinates resources across ten departments in the university to enhance the infrastructure, expertise, and services needed to drive data-intensive research discoveries.

Data Science, Machine Learning, and Data Mining

On many occasions “data science” and “machine learning” are used interchangeably and the two phrases are even combined together as “data science and machine learning” (DSML; Kroese, Botev, Taimre, & Vaisman, 2021). Although “data science” and “machine learning” are two distinct concepts, they are strongly related to each other. Machine learning is a specific application of artificial intelligence (AI) that enables computers to learn and improve from examples without having to be programmed explicitly (Bonaccorso, 2018). This powerful technology has been widely implemented in various domains, such as facial recognition, image generation, video enhancement, natural language processing, robotics, self-driven vehicles, and data analytics. Specifically, machine-learning-based algorithms, such as artificial neural networks (ANN), ensemble methods (e.g. bagging and boosting), and text mining, have been incorporated into the toolkit of data scientists. For instance, although in some situations a nonlinear model is a better fit to the data, nonlinear modeling is laborious and technically challenging. To rectify the situation, neural networks can perform automated nonlinear transformation in the hidden layer by learning from the training data set (Yu, 2022).

The terms “data mining” and “data science” seem to be synonymous, but there is a subtle difference between them. Data mining is the process of extracting useful information and relationships from data. It is so named because while insight is buried under immense quantities of data, the analyst has to dig into the “mine.” Usually, data mining deals with structured data only, whereas text mining handles unstructured data. Data mining is a subset of data science; however, data science is an umbrella term which covers both data mining and text mining. Data science is an umbrella term covering both data mining and text mining whereas data mining is a subset of data science (Kelleher, Sorensen, Tierney, & Media, 2018; Yu, 2022).

Multiple Origins of Data Science

There is no definite founder of the school or the movement of data science, and different historical roots were suggested by different authors. For example, Peter Naur (1974) offered a prototypical definition of data science, which is the science of dealing with data that establishes the relationship between the data and other fields. A close examination of his notion indicates that he simply provided the literal string of "data science," but the idea of data science was not adequately developed (Carmichael & Marron, 2018). In a similar vein, Professor C. F. Jeff Wu called for renaming statistics to be data science and statisticians to be data scientists in his 1977 inaugural lecture for the H. C. Carver Chair in Statistics at the University of Michigan. However, no formal article of the presentation was written and no elaborated research on the suggestion was done before or after the talk (Donoho, 2017). Based on the criterion that pioneering the field must go beyond naming it, data science can be traced back to at least four major sources that contributed to its development: (1) Exploratory Data Analysis (EDA) advocated by John Turkey, (2) The notion of learning from the data proposed by John Chambers, (3) Data visualization and advanced computing methods developed by William Cleveland, and (4) The two-culture thesis suggested by Leo Brieman.

John Tukey: Exploratory Data Analysis

John Tukey was a professor and founding chairman of the statistics department at Princeton University. In his seminal article, *The future of data analysis*, John Tukey (1962) wrote,

> For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt.…All in all, I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data. (p. 2)

This passage epitomizes Tukey's philosophical orientation towards data analysis. In Tukey's view, traditional statistics aims to impose a theoretical distribution and a mathematical model on the data, but this approach neglects the patterns and trends of the data at hand which should be the focal interest of empirical science. Tukey's key point is that data analysis should

be a new science, rather than a branch of mathematics. His book *Exploratory Data Analysis* (1977) expanded the preceding notion by advocating data-driven exploration, rather than prematurely leaping into confirmatory data analysis. Data visualization, also known as revelation, is one of the cornerstones of EDA, and this technique is ubiquitous in modern data science. Another toolkit of EDA is data transformation, also known as data re-expression. Tukey argued that when the data structure cannot meet the model assumptions, such as linearity, data re-expression is necessary. While transformation is done manually in EDA, neural networks, which is one of the popular data science and machine learning methods, perform the transformation in the hidden layer. In this sense, EDA is a precursor of data science. Conversely, data science is an extension of EDA.

John Chambers: Learning From the Data

John Chambers was a researcher at AT&T Bell Labs and the co-inventor of the S programming language, which grew into the R language widely used in data science (Chambers, 2020). The message in his article titled, *Greater or Lesser Statistics, A Choice for Future Research* (1993), is provocative and controversial. According to Chambers (1993), lesser statistics is confined to academia; its focus is mathematical techniques and collaboration with other disciplines is rare. On the contrary, greater statistics is more inclusive, closely related to other disciplines, and practiced by analysts outside of academia. It is also more comprehensive in the sense that the process includes planning, data collection, data organization, data validation, data analysis, and presentation. More importantly, it aims to learn from the data. He warned that statistics might be marginalized unless it can go beyond the traditional realm. Further, statistical software must be integrated into the whole process of learning from data in order to be valuable. There is a common thread between Chambers's notion of learning from the data and Tukey's data-driven EDA. The workflow of a data analyst proposed by Chambers is a typical job description of a modern data scientist. Last but not least, his vision of incorporating advanced computer software into data analysis is also shared by William Cleveland, which will be discussed next.

William Cleveland: Data Visualization and Advanced Computing

William Cleveland II is a Professor of Statistics and Professor of Computer Science at Purdue University. He is well-known for his work on data visualization, which is in alignment with EDA. In his classic book *Visualizing Data*, Cleveland (1985, 1993) argued that data visualization can provide deep insight into the data structure. Many statistical graphing tools introduced in his

books, such as the coplot, are still widely used by data scientists today. Following this line of research methodology, Cleveland (2001) published *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics*, proposing a new discipline of data science that synthesizes modern computing methods and statistics. He suggested that statisticians should incorporate advanced computing into data analysis and learn how to engage with subject matter experts. Thus, his conceptualization of data science encompasses multiple disciplines.

Leo Breiman: Two-Culture Thesis

Leo Breiman was a distinguished statistician at the University of California, Berkeley, who invented several powerful DSML methods, including the classification and regression tree and the random forest (Breiman, 1984, 1996). His article *Statistical Modeling: The Two Cultures* (2001), as the name implies, outlined the two directions of data analysis. According to Brieman, there are two cultures in the application of statistical modeling. The traditional approach looks for inferring from sample statistics to the true population parameters based on the assumption that the data are generated by a stochastic model. Conversely, the prediction-centric approach uses algorithmic models without assuming the data mechanism. In Breiman's view, the first approach tends to produce irrelevant theories and questionable conclusions, and also it is unable to tackle a variety of interesting issues on the horizon. While traditional statistics is often used with smaller data sets, there has been rapid development of algorithmic modeling in fields other than statistics. These endeavors yielded predictive modeling techniques that can be used for both small data sets and large complex data sets. Hence, Breiman recommended moving away from over-reliance on statistical modeling and keeping an open mind to a more diverse set of tools. His vision is later validated by the fact that the growth of data availability has reduced the importance of the first culture and data analysts increasingly focus on how to utilize big data for prediction (Galeano & Pena, 2019).

Differences Between Traditional Statistics and Data Science

The preceding pioneers of data science shared a common vision: Data analytics must remediate the shortcomings of traditional statistics in order to cope with the trends in the modern world. The differences between traditional statistics and data science can be found in the following aspects. In this discussion, traditional or classical statistics is defined as hypothesis testing built upon the frequency school of probability. The Bayesian approach is beyond the scope of this discussion.

Dichotomous Evidence and Decision vs. Pattern Recognition and Contextual Decision

Hypothesis testing is based on the conviction that in the population there is only one fixed constant that represents the true parameter. Its conclusion is inevitably dichotomous: either to reject or not to reject a null hypothesis according to the *p*-value. It is noteworthy that when alpha = .05 or .01 is adopted, not only is the final decision dichotomous, but also the evidence is binary.

In contrast, data science seeks to identify the patterns and trends in data, which is consistent with Chambers' (2001) notion of learning from the data. On some occasions when data patterns are revealed from data visualization, numeric evidence is unnecessary (Yu, 2014). Further, instead of using any absolute cut-off, in data science the decision is contextual and comparative. For example, in data science Akaike's information criterion (AIC; Akaike, 1973) and Bayesian information criterion (BIC; Schwarz, 1978) are two common criteria for model selection, but neither has a cutoff level. Rather, the selection is made in the context of model comparison. After multiple models are explored, the one with the lowest AIC or BIC is chosen. Due to its capability of complex model selection, AIC is considered the first step toward AI and machine learning (Galeano & Pena, 2019).

Model-Driven and Assumption-Based vs. Data-Driven and Assumption-Free

In traditional statistics usually, the researcher starts with a pre-formulated hypothesis or a strong model and then collects data to confirm the hypothesis or the model. This model-based approach requires parametric assumptions derived from mathematical models. For example, according to R.A.Fisher (1935), "The normal distribution has only two characteristics, its mean and its variance. The mean determines the bias of our estimate, and the variance determines its precision" (p. 42). The estimation is more precise as the variance becomes smaller and smaller. However, in reality, a lot of data are not normally distributed. French physicist Gabriel Lippmann criticized the circular logic of proving normality by saying, "Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact" (cited in Thompson, 1959, p. 121). In response to the challenge of big data analytics, Carmichael and Marron (2018) argued that Gaussian models are no longer adequate for noisy data. Whenever data from multiple sources are combined, classical statistical approaches, such as the likelihood principle, are problematic. Besides normality, another strong assumption commonly imposed on the data structure is linearity. However, the assumption of linearity is seldom true in big data (Galeano and Pena, 2019).

To rectify the situation, most data science methods are data-driven and assumption-free. First, rather than pinning down one or two specific hypotheses, a data scientist might explore hundreds or even thousands of potential predictors of a phenomenon simultaneously. Second, most data science methods are nonparametric, meaning that no or very few assumptions of the data structure are required. In the real world, data are noisy and complex, rather than following the rules of mathematical theorems. The developers of data science algorithms take this cruel fact into account and let the data speak for themselves. This principle of data science is fully compatible with the philosophy of EDA, which also endorses assumption-free knowledge discovery.

Single-Modeling vs. Model Multiple-Modeling

The starting point in statistics is usually a simple model (e.g., linear regression), and the data are checked to find out whether the data structure meets the parametric assumptions. The model is improved by remediating the assumption violations. When there are many predictors, variable selection methods, such as stepwise regression, are employed to determine which predictors should be retained (Smith, 2018). Even though on some occasions multiple test procedures are employed, each one is dedicated to a specific purpose, not for examining the data through multiple perspectives. For example, if multiple dependent variables are involved in the study, the analyst uses MANOVA for initial screening. If a significant effect is found in the relationship between one or more dependent variables and the grouping factors, follow-up analysis is run using ANOVA. If the *F*-test result is significant, the next step is post hoc analysis (Toothaker, 1991). In short, traditional statisticians tend to make improvements on a single modeling technique to fit the data.

On the other hand, data scientists usually run many models using a variety of modeling techniques, such as support vector machines, decision trees, random forests, gradient boosting, neural networks, etc. Finally, the best model is chosen on the basis of predictive accuracy, variance explained, and error rate. This race-to-the-top analytical approach is in accordance with inference to the best explanation (Yu, 2006, 2022).

Whole Sample vs. Resampling and Subsetting

The entire sample is usually included in the modeling process in traditional statistics. As a result, this approach is prone to overfitting, which is caused by the model attempting to account for all observations in a given sample. If the researcher attempts to generalize the

finding to another sample, the model will not hold up. This limitation partly contributes to the replication crisis in psychology (Colling & Szucs, 2018; Open Science Collaboration, 2015).

To compensate for this shortcoming, in most data science methods the sample is randomly partitioned into two to three subsets for cross-validation (Kurtz, 1948). If there are two subsets, the first one is known as the training set whereas the second one is called the validation set. If there are three subsets, the names of the first two subsets remain the same and the last one is called the testing set. The training set, as the name implies, is used for training the algorithm. After an initial model is proposed by the preliminary analysis, it is fitted into the second or the third sub-sample. If the predictive accuracy, variance explained, and error rate are similar across three subsets, it is likely that the model is stable and generalizable. Consider this metaphor: the first run is a rehearsal and the final one is the actual performance on the stage. Thus, usually, the analyst should present the final model as the finding. Ensemble methods, such as bagging and boosting, go even further to generate hundreds or even thousands of sub-samples by bootstrapping (Efron, 1979). Cross-validation is resampling without replacement while bootstrapping is resampling with replacement. Group membership in the former is mutually exclusive, while cases in the latter can be reassigned to multiple subsets. By doing so ensemble methods can repeat the same analyses numerous times and then merge the results into the final model in the end.

Small Data vs. Big Data

When Karl Pearson, Fisher, and other statistical theorists developed their methodologies during the late 19th and early 20th centuries, these methods were suited to small-sample studies. When statisticians use a small sample to make an inference about the population, it is difficult to separate noise from signals, and thus, it is crucial to quantify the uncertainty of the estimation. Common ways for quantifying uncertainty are the confidence intervals and the margin of error. As mentioned previously, traditional statistics assumes there exists a single true parameter in the population. In a similar vein, classical test theory also postulates a true score. Following this line of reasoning, inferential uncertainty can be conceptualized as the probability of the strength of approximation to the truth (Lele, 2020).

On the contrary, in corporations like Google and Amazon, their data scientists are able to access population-level data, and thus, quantifying uncertainty becomes less important. Nonetheless, some researchers argued that its inability to quantify uncertainty is a major disadvantage of machine learning methods (Dunson, 2018; Faraway & Augustin, 2018). To

alleviate the problem, Andrew Gelman (2020) asserted that the Bayesian inference network should be used in machine learning in order to quantify uncertainty.

The meaning of big data is commonly misunderstood as nothing more than a bigger sample size. For example, David Donoho (2017) rejected big data as a criterion for separating traditional statistics from data science, since statistics dealt with census data before the emergence of data science. According to Faraway and Augustin (2018), “Big data deals with the large, observational and machine analysed . Small data results from the experimental or intentionally collected data of a human scale where the focus is on causation and understanding rather than prediction” (p. 142). First, some big data are collected through experimental methods (e.g. Kramer, Guillory, & Hancock, 2014). Second, big data entail multiple facets. In addition to high volume (big sample size and a large number of variables), big data are also characterized by high velocity, high variety, and high veracity.

Velocity refers to how fast data is generated and disseminated. For example, the AI-enabled algorithm of the bank can closely monitor incoming credit card information and recognizes any anomalies instantly. High variety means that the data type available to analysts is no longer structured data (numbers in a table) only. With the advance of digital technology, data could come in different forms other than numbers. Specifically, the incoming data might be semi-structured, such as web-based data in the form of Extensible Markup Language (XML), quasi-structured data, such as Webpage click-stream data, and unstructured data, such as text, image, audio, or video. The veracity of big data is concerned with the quality and trustworthiness of the data. Rather than collecting self-report data via survey, many tech companies directly acquire “behavioral data” of users (what they do instead of what they say). For example, instead of asking how many books the customers read and how much time they spend reading weekly, Amazon's Kindle records the number of pages and duration of the reader's readings.

When the sample size is small and the data format is simple, the researcher can focus on analysis only. As discussed at the beginning, data science is an intersection between data analytics, computing, and domain knowledge. Due to the complexity of big data, it is necessary for a data scientist to be proficient in data architecture, data management, and advanced computing (e.g. data warehousing, distributed file system, structured query language [SQL], NoSQL, high-performance computing [HPC], cloud computing, etc.). Cleveland foresaw this interdisciplinary nature of data science back in 2001.

Inference and Explanation vs. Prediction

The primary goal of traditional statistics is to infer from the sample statistics to the population parameters. Additionally, causal inference plays an important role in theoretical research, and therefore, traditional statistics and experimental design are generally tied together. Some experimenters insisted that without running a randomized experiment we cannot assert a causal explanation between variables. The logic behind this argument implies that causal inferences are weakened in quasi-experiments and that non-experimental data cannot be used to infer cause and effect relationships at all (Keppel & Zedeck, 1989). However, some researchers argued that many causal inferences are made without using the experimental framework (Kenny, 1979; Christensen, 1988). Since the introduction of Linear Structural Equation (LISREL) in the 1970s, Structural Equation Modeling (SEM) has been widely applied by statisticians for uncovering causal structures from non-experimental data (Kline, 2023; Yu, 2007).

Interestingly, the preceding debate did not occur in the realm of data science because prediction, recommendation, and search optimization for patterns and associations from big data, rather than causal inference, was more central to data science, especially in business analytics. For example, a company might be interested in assigning customers to different clusters for market segmentation. It may also be beneficial for the company to identify products or services that are frequently bought together. A concrete example is the a priori association rule used in a recommendation system (Soor, Dalal, & Vora, 2020). Businesses do not need to understand why customers like certain products or services or the psychological mechanism behind the reasoning as long as the system works. It is important to note that in the context of data science, prediction is not merely equated with forecasting events in the future. Rather, it could mean recovering past events. For example, based on the patterns of many junk emails (e.g. containing the phrase. “You received a payment of $1,000” or “Make money easily”), a spam-filtering algorithm could “predict” which message is likely to be a spam email. In this case, the algorithm does not predict what will happen in the future because the email is already in the mailbox. Rather, it is about predicting the missing value of an attribute, regardless of whether the event will take place in the future or has already occurred in the past (Kelleher et al., 2018). These diverging orientations of statistics and data science somewhat echo the two-culture thesis proposed by Breiman in 2001.

Overall Tendencies vs. Personalized Recommendation

The traditional approach to statistics emphasizes tendencies and representations derived from combining individual data points. An example is Adolphe Quetelet's (1835) *On Man: Essays on Social Physics*. Quetelet (1835) found that different scientists obtained different results even though they observed the same astronomical phenomenon. He applied this phenomenon of astrophysics to social studies by developing the concept of the "average man," which is a value aggregated from mass data representing the whole group. The subsequent development of statistics followed this line of reasoning (Tafreshi, 2022).

Although data scientists also analyze data patterns and trends, they do not stop at the overall picture. Rather, the implications are made to both the group and the individuals. Machine learning, for example, has been widely used in the development of recommendation systems that suggest appropriate content and items to users based on their preferences. You can find recommendation systems in a wide variety of domains, including e-commerce (e.g. Amazon), streaming services (e.g. Netflix and YouTube), social media platforms (e.g. Facebook), and many others (Banik, 2018). Furthermore, machine learning plays an important role in personalized treatment by using computation algorithms and models to analyze large volumes of data and provide personalized insights. By utilizing these algorithms, early detection of diseases can be identified and accurately diagnosed, allowing for timely intervention and personalized treatment strategies. By using genetic profiles, biomarkers, and clinical characteristics, machine learning algorithms can identify patient subgroups. The stratification of patients enables the design of trials that target specific patient populations, which in turn increases the likelihood of successful outcomes and reduces the risks associated with non-responsive patients (Sebastiani, Vacchi, Manfredi, & Cassone, 2022).

Relationships Between Traditional Statistics and Data Science

Nothing can emerge from a vacuum. As a matter of fact, certain DSML methods are based on concepts, procedures, and theorems from traditional statistics. For example, the LogWorth statistic used in the decision tree is derived from the *p-value* (Yu, 2022). Generalized Regression, which is a common DSML method for preventing the model from overfitting, is extended from the framework of Ordinary Least Squares regression (Tibshirani, 1996). In addition, several data/dimension reduction methods, such as cluster analysis and principal component analysis, are inherited from traditional multivariate statistical analysis (Yu, 2022).

It has been demonstrated by Fan, Han, and Liu (2014) that some statistical procedures are no longer applicable to big data. Although data science methods are considered more

versatile, powerful, and efficient than traditional statistics in many aspects, it does not necessarily imply that the former can completely supersede the latter. Even though we have entered the era of Big Data, not all research problems are big data problems and some simple problems can be resolved by traditional statistics (Hassani, Beneki, Silva, Vandeput & Madsen, 2021). For example, using neural networks on a data set with 50 cases and three variables is definitely overkill. Indeed, statistical inference works better on small data sets while machine learning algorithms often perform poorly on them (Faraway & Augustin, 2018).

In data science, predictive modeling takes precedence over causal inferences. Today, most causal modeling techniques for observational data still rely on traditional statistics. For example, direct acyclic graphs are derived from structural equation modeling (Pearl &Mackenzie, 2018) and propensity scoring (Connelly et al., 2013) is based on regression modeling. According to Pearl and Mackenzie (2018), big data analytics is not the answer to causal inferences. To develop a causal model, analysts need the right data, not necessarily big data. Taking all of the above into consideration, traditional statistics and data science can supplement each other, and the proper use of each depends on the research problem, the sample size, and the data type.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. pp. 267-281. In: B. N. Petrov and F. Csaki [Eds.]. *International Symposium on Information Theory*. Akademia Kiado, Budapest, Hungary.

Banik, R. (2023). *Hands-On recommendation systems with Python: Start building powerful and personalized, recommendation engines with Python*. Packt Publishing.

Blei, D. M., & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences*, *114*(33), 8689–8692.

Bonaccorso, G. (2018). *Machine learning algorithms : Popular algorithms for data science and machine learning* (2nd ed.). Packt Publishing.

Breiman, L. (1984). *Classification and regression trees*. Wadsworth International Group.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

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

Connelly, B. S., Sackett, P. R., & Waters, S. D. (2013). Balancing treatment and control groups in quasi-experiments: An introduction to propensity scoring. *Personnel Psychology*,*66*, 407–442. https://doi.org/10.1111/peps.12020

Donoho, D. (2017). 50 years of data science, *Journal of Computational and Graphical Statistics*, *26*(4), 745-766. DOI: 10.1080/10618600.2017.1384734

Carmichael, I., & Marron, J. S. (2018). Data science vs. statistics: Two cultures? *Japanese Journal of Statistics and Data Science*, *1*,117–138. https://doi.org/10.1007/s42081-018-0009-3

Chambers, J. M. (1993). Greater or lesser statistics: A choice for future research. *Statistics and Computing*, *3*, 182–184.

Chambers, J. M. (2020). S, R, and data science. *Proceedings of the ACM on Programming Languages*, *4*, 1-17. https://doi.org/10.1145/3386334

Christensen, L. B. (1988). *Experimental methodology*. Allyn and Bacon.

Cleveland, W. S. (1985). *The elements of graphing data*. Wadsworth Advanced Books and Software.

Cleveland, W. S. (1993). *Visualizing data*. Hobart Press.

Cleveland, W. S. (2001). Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*, 21–26.

Colling, L., & Szucs, D. (2018). Statistical inference and the replication crisis. *Review of Philosophy and Psychology*, *12*, 121-147. https://doi.org/10.1007/s13164-018-0421-4

Conway, D. (2010, September 30). *The data science Venn diagram*. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

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

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*, 1-26.

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review*, *1*(2), 293–314.

Faraway,J. J., & Augustin, N. H. (2018). When small data beats big data. *Statistics & Probability Letters*, 136, 142-145. https://doi.org/10.1016/j.spl.2018.02.031

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society*, *98*, 39-82.

Galeano, P., & Pena, D. (2019). Data science, big data and statistics. *TEST*, *28*, 289–329. https://doi.org/10.1007/s11749-019-00651-9.

Gelman, A. (2021). Reflections on Breiman's two cultures of statistical modeling. *Observational Studies*, *7*(1), 95-98. doi:10.1353/obs.2021.0025

Hassani, H., Beneki, C., Silva, E. S., Vandeput, N., & Madsen, D. (2021). The science of statistics versus data science: What is the future? *Technological Forecasting and Social Change*, *173*, 121111. https://doi.org/10.1016/j.techfore.2021.121111

Kelleher, J. D., Sorensen, C., Tierney, B., & Media, G. (2018). *Data science*. MIT Press.

Kenny, D. (1979). *Correlation and causality*. John Wiley.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs: Analysis of variance and multiple research/correlation approaches*. W. H. Freeman.

Khine, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788-8790. https://doi.org/10.1073/pnas.132004011

Kroese, D. P., Botev, I. B., Taimre, T., & Vaisman, R. (2021). *Data science and machine learning: Mathematical and statistical methods*. CRC Press.

Kurtz, A. K. (1948). A research test of Rorschach test. *Personnel Psychology*, *1*, 41-53. https://doi.org/10.1111/j.1744-6570.1948.tb01292.x

Lele, S. R. (2020). How should we quantify uncertainty in statistical inference? *Frontiers in Ecology and Evolution*, *8*. https://doi.org/10.3389/fevo.2020.00035

Naur, P. (1974). *Concise survey of computer methods*. Petrocelli Books.

North Carolina State University. (2023). *Interdisciplinary data science*. https://research.ncsu.edu/dsi/

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi: 10.1126/science.aac4716. http://science.sciencemag.org/content/349/6251/aac4716

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. doi:10.1214/aos/1176344136

Sebastiani, M., Vacchi, C., Manfredi, A., & Cassone, G. (2022). Personalized medicine and machine learning: A roadmap for the future. *Journal of Clinical Medicine*, *11*(14),4110. DOI:10.3390/jcm11144110

Smith, G. (2018). Step away from stepwise. *Journal of Big Data*, *5*, 1-12. https://doi.org/10.1186/s40537-018-0143-6

Soor, G. K., Dala, R. l., & Vora, D. (2020). Product recommendation system based on user trustworthiness and sentiment analysis. *ITM Web of Conferences*, *32*, 03030. https://doi.org/10.1051/itmconf/20203203030

Tafreshi, D. (2022). Adolphe Quetelet and the legacy of the 'average man' in psychology. *History of Psychology*, *25*(1), 34-55. DOI: 10.1037/hop0000202

Thompson, D. W. (1959) *On growth and form*. Cambridge University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, *B*,(58), 267–288.

Toothaker, L. (1991). *Multiple comparisons for researchers*. Sage.

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

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

University of Michigan. (2020). *Michigan Institute for Data Science*. https://midas.umich.edu/

Yu, B. (2014, October 1). *IMS presidential address: Let us own data science*. Institute of Mathematical Statistics. https://imstat.org/2014/10/01/ims-presidential-address-let-us-own-data-science/

Yu, C. H. (2006). *Philosophical foundations of quantitative research methodology*. University Press of America.

Yu, C. H. (2007). Causation in quantitative research methodologies from path modeling, SEM to TETRAD. *Theory and Science*, *9*. https://theoryandscience.icaap.org/content/vol9.3/chong.html

Yu, C. H. (2014). *Dancing with the data: The art and science of data visualization*. LAP LAMBERT Academic Publishing.

Yu, C. H. (2022). *Data mining and exploration: From traditional statistics to modern data science*. CRC Press.