International Journal of Data Science and Analytics (2018) 6:167–175
https://doi.org/10.1007/s41060-017-0090-x

REGULAR PAPER

What makes Data Science different? A discussion involving Statistics2.0 and Computational Sciences

Christophe Ley[1] · Stéphane P. A. Bordas[2,3]

Received: 21 March 2017 / Accepted: 12 December 2017 / Published online: 5 February 2018

Abstract
Data Science is today one of the main buzzwords, be it in business, industrial or academic settings. Machine learning, experimental design, data-driven modelling are all, undoubtedly, rising disciplines if one goes by the soaring number of research papers and patents appearing each year. The prospect of becoming a "Data Scientist" appeals to many. A discussion panel organised as part of the European Data Science Conference (European Association for Data Science (EuADS)) https://euads.org/edsc/ asked the question: "What makes Data Science different?" In this paper, we give our own, personal and multi-faceted view on this question, from a Statistics and an Engineering perspective. In particular, we compare Data Science to Statistics and discuss the connection between Data Science and Computational Sciences.

Keywords High-dimensional statistics · Data assimilation · Interdisciplinary research · Data Science · Data fusion · Computational Sciences · Machine Learning · Scientific Computing · Data-driven modelling · Modelling · Applied mathematics · Simulation · Digital twins · Training · Education · Research

1 Introduction

According to IBM, 90% of the data available today has been generated over the last 2 years [1]. We have been experiencing a data-flood, fuelled by a surge in (mobile) computing power which has enabled the creation of devices which can create, collect, store and transfer increasingly complex and large data sets. This accelerated data-gathering ability has been drastically changing the world of science and business. The "internet of things" and wearable technologies densely maculate our world with digital footprints. These massive amounts of data are continuously being gathered in geography, geophysics, medicine, genetics, social science (media), finance, climatology and engineering. Evidence suggests that the intensity of this surge will only increase with time. We are living in the "Big Data" era and this yet ill-defined concept is now ubiquitous, be it in science, business, healthcare, media, industry, business, politics or sports. The challenges posed by the Big Data phenomenon are numerous, and the discipline known as "Data Science" may well be a natural consequence of the data outpour we have been witnessing.

But what does Data Science actually stand for? What makes it different from other, well-established disciplines? Why has it become so popular over the past years? Is Data Science merely Statistics? Is it Computer Science, Machine Learning? Wikipedia provides the following answer to the first question:

> *Data Science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to Knowledge Discovery in Databases.* Wikipedia, accessed on 23 February 2017.

The authors would like to dedicate this work to the late Professor Sabine Krolak-Schwerdt, main organiser of the conference where this work was born.

✉ Stéphane P. A. Bordas
stephane.bordas@alum.northwestern.edu

Christophe Ley
christophe.ley@ugent.be

[1] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281, Campus Sterre, S9, 9000 Ghent, Belgium

[2] Department of Computational Engineering, University of Luxembourg, 6 Avenue de la Fonte, 4362 Esch-sur-Alzette, Luxembourg

[3] Institute of Mechanics and Advanced Materials, School of Engineering, Cardiff University, Wales, UK

This definition, like many others, remains vague and is surely insufficient to differentiate this discipline from its cousins. Many have attempted to define Data Science through articles [2–5] as well as numerous panel discussions at the highest level as well as conference/seminar presentations. There have also been significant discussions on teaching and education in Data Science [6,7], approaches to building Data Science teams [8] and the use of Data Science for various applications ranging from Social Sciences [9] to the material genome initiative [10].

But, in spite of all these developments, what has really changed since the 1962 article entitled *"The future of data analysis"* by John W. Tukey in *The Annals of Mathematical Statistics*?

We wish to contribute to this active discussion via the present paper, which is based on a panel discussion to which we contributed during the European Data Science Conference in November 2016 in Luxembourg. The originality of our approach is the combination of two apparently disjoint domains which Data Science may have already brought closer together, namely Statistics and Computational Sciences.

Should the reader wish to delve more into the details of particular Big Data disciplines, the review papers [1] and [11] are excellent sources of information.

2 A simple classification of Data Science approaches

Before looking at how Data Science approaches can be classified, let us first think of examples of typical Data Science problems. Data Science answers sharp and quantitative questions such as:

> Quantify: How many coffee bean futures should I order assuming the temperature in the tropics rises by 5 degrees? This is done using regression algorithms.
> Detect anomaly: Has this credit card been stolen?
> Classify and make predictions: Will this aircraft door fail within the next 2,000 flights? How likely is a returning customer to become a regular customer? Given images of a brain, what is the probability that the tumour is located within 10 mm of an eloquent region of this brain?
> Organise: How is the data organised? For example, clustering algorithms help organise data. This can be useful to predict behaviour and events.
> Choose the next step: What innovation directions should this country follow apart from maximising its GDP? These algorithms are known as "reinforcement learning" and can be used to control autonomous systems, for example self-driving cars or climate control systems. They learn by trial and error.

Clearly, attacking such problems in their full complexity requires a serious mathematical arsenal. The mathematical methods behind Data Science applications can seem mystical to the neophyte, see for instance [12] for open problems in the Mathematics of Data Science. We summarise here very briefly how we believe Data Science methods can be classified. We distinguish between bottom–up and top–down approaches.

In top–down approaches, a model is built which represents the information contained in the data. This is usually a statistical model, for example a regression approach, possibly Bayesian when information is scarce. In practice, what makes a top–down Data Science algorithm successful is the craft with which the above statistical models are used in concert. We discuss some of this orchestration and how Data Science relates to Statistics in the next section.

In bottom–up approaches, on the contrary, the starting point is the data and the model of this data is generated by a computer (and updated continuously as new data is acquired) to match observations. However powerful these methods have become, (skilled) human intervention is still necessary to filter outliers, optimise the learning paradigm to ensure the accuracy of classifications, tune the parameters involved in the model, etc.

In fact, successful Data Science algorithms are usually a combination of the top–down and bottom–up approaches. The top–down approach brings domain- (or application-) knowledge which leads to significant savings in the computing power required by the bottom–up approaches, for example by accelerating classification.

3 How does Data Science relate to Statistics?

3.1 Nomen est omen

The first author has recently asked the students of his Data Mining class what the word Data Science meant to them. After a long silence, the following answer came: "Data Science is the discipline that makes sense out of data". For a statistician such an answer is surprising, as this is precisely what Statistics aims to do. What causes this difference in perception between professional statisticians and non(or not yet)-statisticians? The reason is simple: Data Science seems, just by its name, to be a more data-oriented area than Statistics. And more attractive. If you say to a random person on the street that you are a statistician, the typical reaction of that person is to think you are dealing with spreadsheets, which can seem monotonous as job. However, if you happen to say you are a data scientist, then that same person will have no clue about your job, yet he/she will have the feeling your job must be exciting. The core task is in both cases data analysis, but the marketing effect of the name Data Science is incontestable.

A similar effect happens to prevail also among people with an advanced understanding of data analysis. While Statistics appears to be a rigid field, filled with rules to follow and warnings of how to correctly quantify the uncertainty inherent to any data set, Data Science seems to invite theoreticians and practitioners to play around with data in an unrestricted way. This is, again, just a subjective impression.

3.2 Statistics in the Big Data era

A core role of Statistics is the quantification of the uncertainty accompanying any data analysis. Sir Ronald Fisher has laid a solid mathematical background for this endeavour in the beginning of the 20th century. Estimation, testing and regression procedures were devised on the basis of this formalism. These methods, however, can no longer be blindly applied to 21st-century data which happen to be complex and occur in unprecedented quantities. We illustrate this statement through two classical statistical procedures:

- Linear regression: suppose we are interested in modelling the relationship between a one-dimensional outcome variable Y and p one-dimensional predictors $X_1, \ldots, X_p$, and we have good reasons to believe the relationship to be of the form

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon,$$

  where ϵ is an error term (typically assumed to follow a normal distribution) and $\beta_0, \beta_1, \ldots, \beta_p$ are the regression parameters we need to estimate. The standard solution to this estimation problem is least squares estimation. This approach works very well as long as the number n of observations $((Y_1, X_{11}, \ldots, X_{1p}), \ldots, (Y_n, X_{n1}, \ldots, X_{np}))$ is larger than the dimension p. However, in many data sets nowadays the situation is rather reversed, with p being larger than n. Think of Genetics, where every single gene should in principle be taken into account to measure the impact of a new treatment. Least squares estimation breaks down in such a context because the empirical covariance matrix is no longer invertible. As a response, variable selection methods have been proposed. This idea is based on the belief in sparsity: the majority of predictors, here genes, shall only have a very small, irrelevant impact on the outcome variable, hence should be discarded. Variable selection does precisely this: it focusses on a small number of predictors that really do matter in the linear regression. Linear regression combined with variable selection can deal with $p > n$ situations. The perhaps most famous example is the so-called Lasso regression of [13].
- Hypothesis testing: suppose we have n data points $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})', i = 1, \ldots, n$, of dimension p and we wish to perform a typical hypothesis testing problem of the form $\mathcal{H}_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ versus the alternative $\mathcal{H}_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ for $\boldsymbol{\mu}_0$ some particular value of the parameter $\boldsymbol{\mu}$ (which can be a parameter of location, scatter, skewness, etc.). Suppose that the classical (meaning $n \to \infty$ while p remains small) asymptotic distribution of the associated test statistic T_p^n follows a chi-square distribution with p degrees of freedom, which we denote χ_p^2, and that $\mathcal{H}_0$ is rejected whenever $T_p^n > \chi^2_{p;1-\alpha}$, the α-upper quantile of the χ_p^2 distribution. Now, when the dimension p itself becomes very large, potentially larger than n, this test becomes worthless as the chi-square distribution will diverge (recall that its expectation is p and its variance is $2p$). Consequently, the test statistic needs to be modified, for instance into

$$\tilde{T}_p^n := \frac{T_p^n - p}{\sqrt{2p}} \xrightarrow[n\to\infty]{\mathcal{D}} \frac{X_p - p}{\sqrt{2p}} \xrightarrow[p\to\infty]{\mathcal{D}} \mathcal{N}(0,1) \qquad (1)$$

  where X_p stands for a χ_p^2 random variable and $\mathcal{D}$ means convergence in distribution. From (1) we see that comparing the modified test statistic $\tilde{T}_p^n$ to quantiles of the standard normal distribution would allow us to have a new large-p test for our hypothesis of interest, provided the so-called (n, p)-asymptotic result from (1) holds true. Indeed, as both n and p grow large, there is no guarantee that the limit when both n and p go to infinity can be calculated by first letting n become large and then p. This must be formally proved. In certain cases it turns out to be a valid manipulation, but in other situations it does not and the initial test statistic must be changed more substantially. An example of such distinct situations is provided in the seminal paper by Ledoit and Wolf [14] who considered scatter matrices.

These two examples underline two novel challenges statisticians are facing when dealing with Big Data. The need to cope with such data has given rise to a popular new research direction, called high-dimensional statistics (see, e.g. [15]). Besides this new research line, the entire field of Statistics has undergone changes as a reaction to the new data paradigm[1]. Supervised and unsupervised learning, shrinkage techniques, graphical models, data mining, functional data analysis and methods to deal with intractable likelihood models are just a few of the new hot topics in statistical research. There is also an increasing trend towards Statistics occupying a

[1] Samworth in [16] provides a concise and very accessible overview on the new data-driven statistical research.

central role in Science in general, as discussed in the next section.

3.3 Data Science = Statistics2.0

The idea of a statistician (or mathematician) working in an Ivory Tower is obsolete. Several fields are in need of statisticians to help them analyse their data; conversely, significant advances in Statistics have been driven by such demands and the collaboration with experts having complementary knowledge. The Big Data era offers Statistics plenty of new possibilities and has brought this traditional field to the limelight of modern scientific research. The era of data may be that of the rebirth of Statistics. Hal Varian, chief economist of Google, said in 2009 "*I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?*".

Where precisely lies the boundary between modern Statistics and Machine Learning? How much Statistics is present in Computational Biology, in Bioinformatics? Health Sciences have benefitted enormously from tailor-made statistical research, see [17] for examples. The same holds true for Systems Biomedicine, Finance and Environmetrics, among many others. Diggle in [17] expresses his opinion that Statistics is actually the Data Science of our modern times. We concur with him and like to say that Data Science is actually Statistics2.0, hereby underlining the new orientation Statistics has taken.

4 How does Data Science relate to Computational Sciences?

The soaring amount of data has brought a new life to Statistics, and by doing so has also opened new doors to the discipline known as "Computational Sciences" or "Scientific Computing." We discuss briefly in this section how Data Science relates to Computational Sciences and how it may revolutionise the way we think about modelling, simulations and computations and enable a transformation of the engineering ecosystem.

First, let us agree that Science is defined as *the activity concerned with the systematic acquisition of knowledge and is an enterprise that builds and organises knowledge in the form of testable explanations and predictions about the universe*. Engineering we define as *the application of scientific and practical knowledge for the benefits of mankind*. For example, Theodore von Kármán, a leading mathematician, aerospace engineer and physicist, developed theories for aerodynamics, in particular supersonic and hypersonic airflow characterisation, which have been essential to the design and fabrication of modern jet engines and rockets. Computational Sciences have been an essential tool for such theories to bear upon modern design approaches.

To produce new knowledge and apply this knowledge to practical fields, scientists and engineers use the "scientific method" which tests statements that are logical consequences of scientific hypotheses (theories or computer models and simulations) through repeatable experiments and observations. This production of knowledge has been fuelled by a significant revolution which has taken place over the last 50 years, through which a new, inherently multi-disciplinary pillar of Science has emerged to complement these theories and observations: Computational Sciences. Computational Sciences is the tri-disciplinary endeavour concerned with the use of computational methods and devices to enable scientific discovery and engineering applications in Science.

In this new era, the wealth of Data has transformed the world of scientific discovery and engineering innovation. We believe that the fusion of Computational Sciences with Data Science will lie at the core of future scientific and engineering research. A new ability will play a central role, namely that of extracting knowledge from this wealth of information by storing, compressing, classifying, ordering and analysing data.

In particular, we will witness the emergence of smart systems, able to adapt to their environment through advanced data gathering and treatment approaches. These developments will be multi-disciplinary with Mathematics, in particular Statistics and Numerical Analysis, as well as Computer Science at its core.

In short, the fusion of Computational Sciences, a half-century old scientific field, with Data Science, which could be argued is a modern embodiment of Statistics, will fuel the development of exciting new research, technology and businesses. The interested reader can refer to [18,19].

5 Interdisciplinarity aspects

Data Science is, by definition, an interdisciplinary field. It incorporates knowledge from Statistics, Computer Science and Mathematics and hence can tackle challenging application domains which had remained out of reach because of a combined lack of data and computer power. In what follows we shall illustrate this interdisciplinary nature of Data Science by means of two case studies.

5.1 Case study 1: protein structure prediction

Predicting the correct three-dimensional structure of a protein given its one-dimensional protein sequence is a crucial issue in Life Sciences and Bioinformatics. Massive databases of DNA and protein sequences have become available, and

many research groups are actively pursuing their efforts to solve the protein folding problem.

A promising approach has been put forward by the research group of Prof. Thomas Hamelryck from the University of Copenhagen. It combines inputs from Biology, Statistics, Machine Learning, Physics and Computer Science, and hence is a nice example of Data Science in action. One of their main ingredients are graphical models from Machine Learning such as dynamic Bayesian networks, which they analyse from a statistical physics standpoint. An essential part of every protein sequence are the dihedral angles between certain atoms. Predicting their most likely values is a key component in understanding the protein structure at a local level. These pairs of angles, however, are no typical quantities since 0° and 360° represent the same value, hence pairs of angles need to be represented as data points on a torus. Devising statistical models and methods for such data is part of a research stream called Directional Statistics (see the book [20] for a recent account) and requires, besides Mathematics, also Computer Science skills. Finally, the Hamelryck group uses probability kinematics to combine their findings on local and non-local structures in a meaningful way.

We refer the interested reader to the monograph [21] for details about this approach.

5.2 Case study 2: Digital Twins in engineering and personalised medicine

Our second case study is concerned with the problem of data-driven model selection in engineering and medical simulations. We split the discussion in two parts, starting with engineering applications in which digital twins are the most advanced and where ethical considerations are more easily addressed.

All systems devised today in Engineering fall within the category of Complex Systems, i.e. *a system composed of many components which interact with each other.* Natural systems such as the human body or the environment are other examples of Complex Systems. It is not possible to study, design and optimise complex systems using analytical methods, i.e. hand calculations. Recourse is always made to some type of mathematical model, usually a set of partial differential equations (PDEs). The resulting problem is solved numerically using a wide variety of *discretisation methods* including finite element methods [22–26], finite differences, meshfree methods [27], isogeometric approaches [28,29], geometry independent field approximation [30,31], scaled-boundary finite elements [32–36], boundary element approaches [37], enriched boundary elements [38] or combinations thereof [39–41].

Discretisation methods have been subject to a large amount of research but a much more difficult task is the choice of a suitably descriptive mathematical model. In other words, computational engineers need to answer the question: "What is the best model for this system given computational constraints and the quantities I am interested in?" Once the model is chosen, selecting a suitable discretisation approach is usually straightforward.

Let us look at this problem of *model selection* via two connected examples. First, consider modern engineering materials, such as composites which have been developed to perform well in increasingly challenging environments[2]. The durability of gigantic composite structures such as the Airbus A380, over 79 m in wingspan, is influenced by physical phenomena occurring at the scale of carbon fibres which are around 5 microns in diameter. The brute-force approach consisting of including all carbon fibres in the simulation of 1 cubic millimetre of composite material would require solving a set of 8 billion equations in 8 billion unknowns, making the problem intractable over the size of the aircraft. The task of the computational engineer is therefore to select a model which can deal with engineering-scale simulations in a computationally affordable manner, but preserves the important effects taking place at the smaller scales.

Once a suitable model has been selected, the associated parameters must be identified in light of experimental observations, i.e. the model must be calibrated. In Materials Engineering, the traditional approach to this has been to perform experiments within laboratory conditions, which are most often far removed from those which the structure or system will undergo during its service life, in particular when harsh environmental effects are of interest. Statistical approaches can be used, but they only partially overcome the hurdle as they are reliant upon predefined statistical distributions, which do not account for "unknown unknowns" or in-service conditions which were not considered during the experimental campaigns, "rare events" in particular. Parameter and model identification and selection are, today still, open problems.

Increasingly miniatured and versatile sensing devices, embedded into engineering and natural systems offer an exciting alternative to traditional (and insufficient) "Experiment-in-the-lab-to-model-behaviour-in-the-field" approaches by leveraging (Big) Data gathered on the fly, during the service life of the system to drive model selection and parameter identification.

To achieve this, Statistics (namely Bayesian inference) and Machine Learning methods [42–47] have been leveraged for a few years. The Bayesian paradigm, in particular, enables

[2] In particular for space applications where not only mechanical but radiation and thermal effects become critical.

the enrichment of prior (expert) knowledge about the system with new data, as it is being acquired.[3]

Whilst important in Engineering, the need to update models on the fly as new data becomes available in order to better control Engineering Systems is strictly necessary in Personalised Medicine where all patients are different and in vivo experiments are not possible. In this field, it is necessary to infer the best possible model for a patient from *a priori* knowledge obtained from other patients. Successful approaches have been recently published [19,47] which enable predictive science in Medicine, for example for laser-treatment of tumours [42]. The reader is referred to [47] for a recent discussion of the emerging field known as "Computer-Guided Predictive Medicine", to [52] for applications to brain tumour model personalisation and to [53] for sparse Bayesian image registration.

This quest for on-the-fly data assimilation and fusion into computer models has been fuelling the development of "digital twins", a digital replica of the real system, which lives a "digital life" in parallel to the real system and can be interrogated to make decisions. These "twins" require predictive, high-fidelity models to learn from real-time data acquired during the life of the system, accounting for "real" conditions during predictions. These Twins could enable to predict the motion of target areas during surgery with predefined accuracy [54–56] or fuel virtual reality engines [57] enabling the surgeons to "see through" the patient, investigate the potential response of a patient to a given treatment [58]. Digital Twins could also enable to transition from "factors of safety" and associated over-engineering to adaptive structures and systems which adapt to their environment [59–64]. For this revolution to take place, Data Science approaches must be harnessed by computational scientists. This will require significant multi-disciplinary efforts in educating the next-generation computational and data scientists.

6 Conclusions and discussion

We discussed in this paper what we believe makes Data Science different. We offered various interpretations of Data Science and differentiated between bottom–up and top–down Data Science approaches. We also defined Science, Engineering and Computational Sciences/Scientific Computing and attempted to relate Data Science to these more established disciplines. Through personal examples and two case studies, we provided possible explanations for the singularity of Data Science.

In short, we conclude that Data Science enhances the traditional and more conservative world of Statistics with advanced algorithms to enable us to make sense out of soaring amounts of data. Here are our conclusions:

- Data Science fuses the fields of Statistics, Mathematics and Computer Science. Computers are of key importance in Data Science, in particular for bottom–up approaches, but the creation of suitable models, mandatory to make these approaches computationally tractable, requires expert knowledge which we believe will be brought forward by statisticians. In this sense, we perceive Data Science as a modernised version of Statistics, which we term Statistics2.0.
- Data Science has the potential to have strong impact in application domains, in particular on Engineering and Medicine. Some of the exciting applications of Data Science include the delivery of the next-generation smart and autonomous devices able to learn from and adapt to their environment.
- Through a crafty coupling with Computational Sciences, Data Science can help create "digital twins" of complex systems. Those are replicas of the actual system which live a parallel, virtual/digital life and can be interrogated in order to make decisions on the (cyber-)physical system itself.
- Data Science is an attractive name which makes Data Science sound young, exciting, innovative, and partially mysterious. This may endow those entering this field with a particularly creative and less conservative mindset than in other, more established disciplines.
- Data Science is the right discipline at the right time: the data deluge creates urgent needs and challenging problems, in academia, industry and business. Spurred by a rapid increase in computer power and the ability of mobile devices to generate large amounts of data everywhere we leave our digital footprints, Data Science appears to be the tailor-made discipline to help make sense out of (very) large amounts of data.

Having made the above reflections, there are a number of points which seem important to us going forward in the world of Data Science:

- Ensure that we do not fall for the "hype of Data Science" and ignore theories to the benefit of algorithms. There is need for a "scientist in the loop" even when bottom–up approaches are advocated.[4]
- Devise suitable training programmes at all levels, in particular through continuing education, in order to help

[3] A discussion of pros and cons of the Bayesian approach for model calibration is provided in [48–51].

[4] A referee pointed out to us that "part of the industry tends to confuse Data Science with the ability to use the (penalised)logistic regression in Python". Data Science is obviously much more than that, as our article and the numerous references clearly demonstrate.

create sound careers for data scientists, at the interface between Statistics and Computer Science, with robust mathematical foundations.
– Nurture an intellectually coherent core relying on Mathematics, Statistics and Computer Science to provide rigourous abstractions to application domains and receiving in return stimulating problems and challenges to address.
– Develop research and teaching programmes at the interface between Computational Sciences and Data Science.
– Foster communication between the disciplines at play by encouraging jargon-free discussions and joint conferences.

In our opinion, an exciting research direction lies at the interface between bottom–up and top–down approaches. In many systems, pure computing power and algorithms are insufficient to obtain results within a reasonable time frame. At the same time, full mathematical models involving the full complexity of the system at hand are also computationally intractable, for example in quantum physics [65–69]. Building such hybrid strategies, we expect, will continue to be exciting research directions, at the interface between Statistics, Computer Science and application domains, see, e.g. [70–72]. These hybrid approaches will provide users with a new way to design experiments, based on data acquired on the fly [73,74].

We presented what are but our personal opinions. The reader is free to disagree with us. We hope nonetheless to have contributed a fresh and multi-disciplinary view to the understanding of what makes Data Science different and hence so popular as discipline.

Acknowledgements We wish to thank the guest editors of this special issue as well as two anonymous referees for useful comments and suggestions on our paper, and Sabine Krolak-Schwerdt for inviting us to the discussion panel at the European Data Science Conference.

References

1. Sagiroglu, S., Sinanc, D.: Big data: a review. In: International Conference on Collaboration Technologies and Systems (CTS). IEEE 2013, pp. 42–47 (2013)
2. Hayashi, C.: What is data science? Fundamental concepts and a heuristic example. In: Hayashi, C., Yajima, K., Bock, H.H., Ohsumi, N., Tanaka, Y., Baba, Y. (eds.) Data Science, Classification, and Related Methods, pp. 40–51. Springer, New York (1998)
3. Loukides, M.: What is data science, Big Data Now. Current Prespectives from O'Reilly Radar. Media, Inc., Sebastopol (2011)
4. Akerkar, R., Sajja, P.S.: Intelligent Techniques for Data Science, 1st edn. Springer, Switzerland (2016). https://doi.org/10.1007/978-3-319-29206-9
5. Kitchin, R.: Big Data–hype or revolution? In: Sloan, L., Quan-Haase, A. (eds.) The SAGE Handbook of Social Media Research Methods, p. 27. SAGE, Beverley Hills (2017)
6. Hicks, S.C., Irizarry, R.A.: A guide to teaching data science. arXiv preprint arXiv:1612.07140 (2016)
7. Tang, R., Sae-Lim, W.: Data science programs in US higher education: an exploratory content analysis of program description, curriculum structure, and course focus. Educ. Inf. **32**(3), 269 (2016)
8. Patil, D.: Building Data Science Teams. O'Reilly Media, Inc., Sebastopol (2011)
9. Cioffi-Revilla, C.: Unpublished: link to paper on Researh Gate (2016)
10. McDowell, D.L., Kalidindi, S.R.: The materials innovation ecosystem: a key enabler for the materials genome initiative. MRS Bull. **41**(04), 326 (2016)
11. Cao, L.: Data science: a comprehensive overview. ACM Comput. Surv. (CSUR) **50**(3), 43 (2017)
12. Bandeira, A.S.: Ten lectures and forty-two open problems in the mathematics of data science. Lecture Notes (2015)
13. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) **58**(1), 267 (1996)
14. Ledoit, O., Wolf, M.: Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. Ann. Stat. **30**(4), 1081 (2002)
15. Bühlmann, P., Van De Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, New York (2011)
16. Samworth, R.J.: Big Data: A New Era for Statistics, pp. 43–46. The Eagle, Bryan (2014)
17. Diggle, P.J.: Statistics: a data science for the 21st century. J. R. Stat. Soc. Ser. A (Stat. Soc.) **178**(4), 793 (2015)
18. Willcox, K., Bounova, G.: Mathematics in engineering: identifying, enhancing and linking the implicit mathematics curriculum. In: Proceeding of the 2004 American Society for Engineering Education Annual Conference and Exposition, American Society for Engineering Education, pp. 1–13 (2004)
19. Rüde, U., Willcox, K., McInnes, L.C., De Sterck, H., Biros, G., Bungartz, H., Corones, J., Cramer, E., Crowley, J., Ghattas, O., Gunzburger, M., Hanke, M., Harrison, R., Heroux, M., Hesthaven, J., Jimack, P., Johnson, C., Jordan, K., Keyes, D., Krause, R., Kumar, V., Mayer, S., Meza, J., Morken, K., Oden, J.T., Petzold, L., Raghavan, P., Shontz, S.M., Trefethen, A., Turner, A., Voevodin, V., Wohlmuth, B., W.C.S.: Research and Education in Computational Science and Engineering. arXiv preprint arXiv:1610.02608 (2016)
20. Ley, C., Verdebout, T.: Modern Directional Statistics. Chapman and Hall, Boca Raton, FL (2017)
21. Hamelryck, T., Mardia, K., Ferkinghoff-Borg, J.: Bayesian Methods in Structural Bioinformatics. Springer, New York (2012)
22. Strang, G., Fix, G.J.: An Analysis of the Finite Element Method, vol. 212. Prentice-Hall, Englewood Cliffs, NJ (1973)
23. Zienkiewicz, O., Taylor, R.: The Finite Element Method, vol. 3. McGraw-Hill, London (1977)
24. Bathe, K.J.: Finite Element Method. Wiley Online Library, London (2008)
25. Dhatt, G., Lefrançois, E., Touzot, G.: Finite Element Method. Wiley, London (2012)
26. Hughes, T.J.: The Finite Element Method: Linear Static and Dynamic Finite Element Analysis. Courier Corporation, North Chelmsford (2012)

27. Nguyen, V.P., Rabczuk, T., Bordas, S., Duflot, M.: Meshless methods: a review and computer implementation aspects. Math. Comput. Simul. **79**(3), 763 (2008)
28. Hughes, T.J., Cottrell, J.A., Bazilevs, Y.: Isogeometric analysis: CAD, finite elements, NURBS, exact geometry and mesh refinement. Comput. Methods Appl. Mech. Eng. **194**(39), 4135 (2005)
29. Nguyen, V.P., Anitescu, C., Bordas, S.P., Rabczuk, T.: Isogeometric analysis: an overview and computer implementation aspects. Math. Comput. Simul. **117**, 89 (2015)
30. Marussig, B., Zechner, J., Beer, G., Fries, T.P.: Fast isogeometric boundary element method based on independent field approximation. Comput. Methods Appl. Mech. Eng. **284**, 458 (2015)
31. Atroshchenko, E., Xu, G., Tomar, S., Bordas, S.: Weakening the tight coupling between geometry and simulation in isogeometric analysis: from sub-and super-geometric analysis to Geometry Independent Field approximaTion (GIFT). arXiv preprint arXiv:1706.06371 (2017)
32. Song, C., Wolf, J.P.: The scaled boundary finite-element method? Alias consistent infinitesimal finite-element cell method? For elastodynamics. Comput. Methods Appl. Mech. Eng. **147**(3–4), 329 (1997)
33. Wolf, J.P., Song, C.: The scaled boundary finite-element method-a primer: derivations. Comput. Struct. **78**(1), 191 (2000)
34. Natarajan, S., Ooi, E.T., Saputra, A., Song, C.: A scaled boundary finite element formulation over arbitrary faceted star convex polyhedra. Eng. Anal. Bound. Elem. **80**, 218 (2017)
35. Ooi, E., Song, C., Natarajan, S.: A scaled boundary finite element formulation with bubble functions for elasto-static analyses of functionally graded materials. Comput. Mech. **60**, 1–25 (2017)
36. Saputra, A., Talebi, H., Tran, D., Birk, C., Song, C.: Automatic image-based stress analysis by the scaled boundary finite element method. Int. J. Numer. Methods Eng. **109**(5), 697 (2017)
37. Brebbia, C., Cruse, T.: The boundary element method. J. Appl. Mech. **46**, 718 (1979)
38. Simpson, R., Trevelyan, J.: A partition of unity enriched dual boundary element method for accurate computations in fracture mechanics. Comput. Methods Appl. Mech. Eng. **200**(1), 1 (2011)
39. Simpson, R.N., Bordas, S.P., Trevelyan, J., Rabczuk, T.: A two-dimensional isogeometric boundary element method for elasto-static analysis. Comput. Methods Appl. Mech. Eng. **209**, 87 (2012)
40. Scott, M.A., Simpson, R.N., Evans, J.A., Lipton, S., Bordas, S.P., Hughes, T.J., Sederberg, T.W.: Isogeometric boundary element analysis using unstructured T-splines. Comput. Methods Appl. Mech. Eng. **254**, 197 (2013)
41. Peng, X., Atroshchenko, E., Kerfriden, P., Bordas, S.: Isogeometric boundary element methods for three dimensional static fracture and fatigue crack growth. Comput. Methods Appl. Mech. Eng. **316**, 151 (2017)
42. Fuentes, D., Oden, J., Diller, K., Hazle, J., Elliott, A., Shetty, A., Stafford, R.: Computational modeling and real-time control of patient-specific laser treatment of cancer. Ann. Biomed. Eng. **37**(4), 763 (2009)
43. Oden, T., Moser, R., Ghattas, O.: Computer predictions with quantified uncertainty, part I. SIAM News **43**(9), 1 (2010)
44. Oden, J.T., Prudhomme, S.: Control of modeling error in calibration and validation processes for predictive stochastic models. Int. J. Numer. Methods Eng. **87**(1–5), 262 (2011)
45. Hawkins-Daarud, A., Prudhomme, S., van der Zee, K.G., Oden, J.T.: Bayesian calibration, validation, and uncertainty quantification of diffuse interface models of tumor growth. J. Math. Biol. **67**(6–7), 1457 (2013)
46. Prudencio, E., Bauman, P., Faghihi, D., Ravi-Chandar, K., Oden, J.: A computational framework for dynamic data-driven material damage control, based on Bayesian inference and model selection. Int. J. Numer. Methods Eng. **102**(3–4), 379 (2015)
47. Oden, J.T., Lima, E.A., Almeida, R.C., Feng, Y., Rylander, M.N., Fuentes, D., Faghihi, D., Rahman, M.M., DeWitt, M., Gadde, M., Cliff, Z.J.: Toward predictive multiscale modeling of vascular tumor growth. Arch. Comput. Methods Eng. **23**(4), 735 (2016)
48. Vanik, M.W., Beck, J.L., Au, S.: Bayesian probabilistic approach to structural health monitoring. J. Eng. Mech. **126**(7), 738 (2000)
49. Rappel, H., Beex, L.A., Hale, J.S., Bordas, S.: Bayesian inference for the stochastic identification of elastoplastic material parameters: introduction, misconceptions and insights. arXiv preprint arXiv:1606.02422 (2016)
50. Rappel, H., Beex, L.A., Bordas, S.P.: Bayesian inference to identify parameters in viscoelasticity. Mech. Time Depend. Mater. pp. 1–38 (2017)
51. Giffard-Roisin, S., Delingette, H., Jackson, T., Fovargue, L., Lee, J., Rinaldi, A., Ayache, N., Razavi, R., Sermesant, M.: Sparse Bayesian non-linear regression for multiple onsets estimation in non-invasive cardiac electrophysiology. In: International Conference on Functional Imaging and Modeling of the Heart. Springer, New York, pp. 230–238 (2017)
52. Lê, M., Delingette, H., Kalpathy-Cramer, J., Gerstner, E.R., Batchelor, T., Unkelbach, J., Ayache, N.: Bayesian personalization of brain tumor growth model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, New York, pp. 424–432 (2015)
53. Le Folgoc, L., Delingette, H., Criminisi, A., Ayache, N.: Sparse Bayesian registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, New York, pp. 235–242 (2014)
54. Cotin, S., Delingette, H., Ayache, N.: Real-time elastic deformations of soft tissues for surgery simulation. IEEE Trans. Vis. Comput. Graph. **5**(1), 62 (1999)
55. Bui, H.P., Tomar, S., Courtecuisse, H., Cotin, S., Bordas, S.: Real-time error control for surgical simulation. IEEE Trans. Biomed. Eng. (2017). http://ieeexplore.ieee.org/abstract/document/7932498/
56. Bui, H.P., Tomar, S., Courtecuisse, H., Audette, M., Cotin, S., Bordas, S.: Controlling the error on target motion through real-time mesh adaptation: applications to deep brain stimulation. arXiv preprint arXiv:1704.07636 (2017)
57. Haouchine, N., Dequidt, J., Peterlik, I., Kerrien, E., Berger, M.O., Cotin, S.: Image-guided simulation of heterogeneous tissue deformation for augmented reality during hepatic surgery. In: 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR). IEEE, pp. 199–208 (2013)
58. Beger, R.D., Dunn, W., Schmidt, M.A., Gross, S.S., Kirwan, J.A., Cascante, M., Brennan, L., Wishart, D.S., Oresic, M., Hankemeier, T., et al.: Metabolomics enables precision medicine:"a white paper, community perspective". Metabolomics **12**(10), 149 (2016)
59. Tuegel, E.J., Ingraffea, A.R., Eason, T.G., Spottswood, S.M.: Reengineering aircraft structural life prediction using a digital twin. Int. J. Aerosp. Eng. **2011**, 1 (2011)
60. Glaessgen, E., Stargel, D.: The digital twin paradigm for future NASA and US Air Force vehicles. In: 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA, p. 1818 (2012)
61. Cerrone, A., Hochhalter, J., Heber, G., Ingraffea, A.: On the effects of modeling as-manufactured geometry: toward digital twin. Int. J. Aerosp. Eng. **2014**, 10 (2014)
62. West, T.D., Pyster, A.: Untangling the digital thread: the challenge and promise of model-based engineering in defense acquisition. INSIGHT **18**(2), 45 (2015)
63. Gabor, T., Belzner, L., Kiermeier, M., Beck, M.T., Neitz, A.: A simulation-based architecture for smart cyber-physical systems. In: 2016 IEEE International Conference on Autonomic Computing (ICAC). IEEE, pp. 374–379 (2016)

64. Grieves, M., Vickers, J.: Digital twin: mitigating unpredictable, undesirable emergent behavior in complex systems. In: Kahlen, J., Flumerfelt, S., Alves, A. (eds.) Transdisciplinary Perspectives on Complex Systems, pp. 85–113. Springer, New York (2017)
65. Tkatchenko, A., Scheffler, M.: Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data. Phys. Rev. Lett. **102**(7), 073005 (2009)
66. Montavon, G., Hansen, K., Fazli, S., Rupp, M., Biegler, F., Ziehe, A., Tkatchenko, A., Lilienfeld, A.V., Müller, K.R.: Learning invariant representations of molecules for atomization energy prediction. In: Advances in Neural Information Processing Systems, pp. 440–448 (2012)
67. Rupp, M., Tkatchenko, A., Müller, K.R., Von Lilienfeld, O.A.: Fast and accurate modeling of molecular atomization energies with machine learning. Phys. Rev. Lett. **108**(5), 058301 (2012)
68. Montavon, G., Rupp, M., Gobre, V., Vazquez-Mayagoitia, A., Hansen, K., Tkatchenko, A., Müller, K.R., Von Lilienfeld, O.A.: Machine learning of molecular electronic properties in chemical compound space. New J. Phys. **15**(9), 095003 (2013)
69. Hansen, K., Biegler, F., Ramakrishnan, R., Pronobis, W., Von Lilienfeld, O.A., Müller, K.R., Tkatchenko, A.: Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. J. Phys. Chem. Lett. **6**(12), 2326 (2015)
70. Kerfriden, P., Schmidt, K.M., Rabczuk, T., Bordas, S.P.A.: Statistical extraction of process zones and representative subspaces in fracture of random composites. Int. J. Multiscale Comput. Eng. **11**(3), 253 (2013)
71. Mueller, T., Kusne, A.G., Ramprasad, R.: Machine learning in materials science: recent progress and emerging applications. Rev. Comput. Chem. **29**, 186 (2016)
72. Antony, P., Manujesh, P., Jnanesh, N.: Data mining and machine learning approaches on engineering materials—a review. In: IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, pp. 69–73 (2016)
73. Viti, F., Verbeke, W., Tampère, C.: Sensor locations for reliable travel time prediction and dynamic management of traffic networks. Transp. Res. Rec. J. Transp. Res. Board **2049**, 103 (2008)
74. Fonzone, A., Schmöcker, J.D., Viti, F.: New services, new travelers, old models? Directions to pioneer public transport models in the era of big data. J. Intell. Transp. Syst. **20**(4), 311–315 (2016). https://doi.org/10.1080/15472450.2016.1190553