Data Science in the Chemical Engineering Curriculum

Authors:

Thomas A. Duever

Date Submitted: 2019-12-16

Keywords: chemical engineering curriculum, statistics, Big Data, data science

Abstract:

With the increasing availability of large amounts of data, methods that fall under the term data science are becoming important assets for chemical engineers to use. Methods, broadly speaking, are needed to carry out three tasks, namely data management, statistical and machine learning and data visualization. While claims have been made that data science is essentially statistics, consideration of the three tasks previously mentioned make it clear that it is really broader than just statistics alone and furthermore, statistical methods from a data-poor era are likely insufficient. While there have been many successful applications of data science methodologies, there are still many challenges that must be addressed. For example, just because a dataset is large, does not necessarily mean it is meaningful or information rich. From an organizational point of view, a lack of domain knowledge and a lack of a trained workforce among other issues are cited as barriers for the successful implementation of data science within an organization. Many of the methodologies employed in data science are familiar to chemical engineers; however, it is generally the case that not all the methods required to carry out data science projects are covered in an undergraduate chemical engineering program. One option to address this is to adjust the curriculum by modifying existing courses and introducing electives. Other examples include the introduction of a data science minor or a postgraduate certificate or a Master's program in data science.

Record Type: Published Article

Submitted To: LAPSE (Living Archive for Process Systems Engineering)

Citation (overall record, always the latest version):	LAPSE:2019.1628
Citation (this specific file, latest version):	LAPSE:2019.1628-1
Citation (this specific file, this version):	LAPSE:2019.1628-1v1

DOI of Published Version: https://doi.org/10.3390/pr7110830

processes

MDPI

Commentary

Data Science in the Chemical Engineering Curriculum

Thomas A. Duever

Department of Chemical Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada; tom.duever@ryerson.ca; Tel.: +1-416-979-5140

Received: 30 August 2019; Accepted: 4 November 2019; Published: 8 November 2019

Abstract: With the increasing availability of large amounts of data, methods that fall under the term data science are becoming important assets for chemical engineers to use. Methods, broadly speaking, are needed to carry out three tasks, namely data management, statistical and machine learning and data visualization. While claims have been made that data science is essentially statistics, consideration of the three tasks previously mentioned make it clear that it is really broader than just statistics alone and furthermore, statistical methods from a data-poor era are likely insufficient. While there have been many successful applications of data science methodologies, there are still many challenges that must be addressed. For example, just because a dataset is large, does not necessarily mean it is meaningful or information rich. From an organizational point of view, a lack of domain knowledge and a lack of a trained workforce among other issues are cited as barriers for the successful implementation of data science within an organization. Many of the methodologies employed in data science are familiar to chemical engineers; however, it is generally the case that not all the methods required to carry out data science projects are covered in an undergraduate chemical engineering program. One option to address this is to adjust the curriculum by modifying existing courses and introducing electives. Other examples include the introduction of a data science minor or a postgraduate certificate or a Master's program in data science.

Keywords: data science; big data; statistics; chemical engineering curriculum

1. Introduction

The terms "Data Science", "Big Data" and "Data Analytics" are becoming pervasive, affecting many aspects of our lives, many professional disciplines and certainly the profession of chemical engineering. The reason for this is of course the increasing availability of data brought on by the proliferation of inexpensive sensors and instrumentation, new measurement capabilities related to the development of the Internet of Things and smart sensors, and improved data storage power like cloud computing. In order to exploit the data that is becoming available, chemical engineers need to use data science methods. The questions are then what are these methods and what role can the curriculum play in preparing graduates to be able to use them.

This paper reviews some of the trends and developments that are occurring related to Data Science, explores the relationship between Data Science and Statistics, addresses some of the limitations and failures of data science projects and lists some applications and methodologies in Data Science that are relevant to Chemical Engineering. Based on this discussion, the implications for an undergraduate chemical engineering curriculum are discussed. Several approaches for teaching Data Science at different institutions are reviewed and finally the desire to include Data Science in the curriculum is set in context with respect to many other pressures that exist for a modern engineering program.

2. Trends in Data Science

As mentioned above, the field of Data Science is affecting our lives in a multitude of ways and Institutions and Industry are recognizing its importance and making significant investments. For example, on 8 September 2015, the University of Michigan announced a $100 M "Data Science Initiative" involving the hiring of 35 new faculty members [1]. "Data Science has become the fourth approach to scientific discovery, in addition to experimentation, modeling and computation" said Provost Martha Pollack. Similarly, Purdue University announced that it plans to embed data science into every Major [2]. In addition, there is an urgent need to support Canadians in transitioning to new jobs in data science, machine learning and big data analytics so that they can confidently acquire the relevant skills to work in high-demand jobs that currently go unfilled. The Information and Communications Technology Council (ICTC) projects that by 2020, 43,300 data analytics specialists will be directly employed in Canada [3].

Chiang et al. [4] report that over 70 universities in the US offer Master's programs in related areas, a trend that is occurring at many Canadian institutions as well.

Broadly speaking, as discussed by Beck et al. [5], Data Science can be divided into three tasks. The first is data management, which is core to data science in that it deals with how the data is organized, stored, accessed and shared. The most basic tool for this purpose is the spreadsheet, however for large datasets they are inadequate and it is necessary to use relational database management systems (RDBMs) and some form of structured query language (SQL). The second task is statistical and machine learning which are methods that can be used for supervised or un-supervised learning. In supervised learning, the objective is to develop a predictive model that predicts the outputs from the inputs. In un-supervised learning, the objective is to determine the underlying structure of the data based on some features. The third task is data visualization which consists of methods that can be used to explore the data and help to make decisions based on the analysis of the data. The EDISON project [6] gives a broader, more detailed description of Data Science competencies, which include:

- Data Science Analytics (including Statistical Analysis, Machine Learning, Data Mining, Business Analytics, etc.)
- Data Science Engineering (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
- Data Management and Governance (including data stewardship, curation and preservation)
- Research Methods and Project Management for research related professions and Business Process Management for business related professions
- Domain Knowledge and Expertise (Subject/Scientific domain related)

The emergence of Data Science as a scientific and academic domain is related to the notion of Big Data, which is comprised of data sets that are too large for commonly used software tools to deal with and therefore require specialized approaches and tools. Some of the key concepts or characteristics related to Big Data are often referred to as the three Vs, namely Volume, Velocity and Variety [4]. The Volume refers to the amount of data that has to be managed, while the Velocity describes the rate of the incoming data. Variety is related to the type of data be it structured or unstructured. Two additional characteristics that are important are the Veracity, meaning the quality and accuracy of the data and Value which addresses the question: will the data lead to value? Is it useful and informative?

3. Data Science and Statistics

In trying to define what data science really is, the question arises is data science really different from statistics? In 1997, Jeff Wu in his inaugural lecture entitled "Statistics = Data Science" for his appointment to the H.C. Carver Professorship at the University of Michigan, suggested that statistics, which he defined as a trilogy of data collection, data modeling and analysis, and decision making, be renamed data science and statisticians be called data scientists [7]. As far back as 1962, Tukey promoted the idea that the field of statistics and its research scope needed to be broadened, enlarged

and redirected [8]. The emphasis on mathematical statistics with theorems and proofs was too narrow and he introduced the term "Data Analysis" with a focus on techniques for analyzing and interpreting data and the design of experiments for collecting data having high information content, which he felt should become the new areas of focus. In his 2001 paper entitled, "Data Science: An Action Plan for Expanding the Technical Areas in the Field of Statistics", William S. Cleveland suggested a plan for how academic statistics departments should reframe their work [9]. The abstract reads: "An action plan to expand the technical areas of statistics focuses on the data analyst. The plan sets out six technical areas of work for a university department and advocates a specific allocation of resources devoted to research in each area and to courses in each area. The value of technical work is judged by the extent to which it benefits the data analyst, either directly or indirectly. The plan is also applicable to government research labs and corporate research organizations". Cleveland then goes on to propose six areas of activity and even indicates what percentage a department should devote to each. They are: (1) Multidisciplinary investigations (25%), (2) Models and Methods for Data (20%), (3) Computing with Data (15%), (4) Pedagogy (15%), (5) Tool Evaluation (5%) and (6) Theory (20%). Donoho [1] points out that many departments overemphasize the last area, Theory, therefore resisting the recommendations by colleagues like Tukey and Cleveland towards a much broader definition of the field.

Despite the fact that statisticians, engineers and scientists working in the area of applied statistics have promoted the methods and ideas behind data science for many years, statistics and the role of statisticians seems to have become marginalized in the recent developments around Data Science. Leaders and leading initiatives in Data Science seem to be coming from various engineering disciplines, computer science, business schools, etc. This despite the fact that many descriptions of what Data Science is, will be very familiar to statisticians. Vincent Granville at the Data Science Central Blog writes "Data Science without statistics is possible, even desirable" [10]. He differentiates between "old" outdated statistical methods that are not particularly useful and "new" statistical methods that are, but that are not recognized by traditional statisticians as being "statistics".

In his 2013 paper, Vasant Dhar addresses what the terms data science and big data mean, what skills individuals working in data science need and what the implications might be for scientific inquiry [11]. He makes the distinction between data analysis, which he states has been used to explain phenomena and data science which "aims to discover, and extract actionable knowledge from the data, that is knowledge that can be used to make decisions and predictions not just to explain what is going on". He also points out that data takes many forms including text, videos and images. As far as the skills required Dhar lists (1) Machine Learning which must build on statistics including Bayesian statistics and multivariate analysis, (2) Computer Science including data structures, algorithms and systems including distributed computing, databases, parallel computing, and fault-tolerant computing (3) Knowledge about correlation and causation and finally (4) The ability to formulate problems in a way that results in effective solutions. On the last point, Wladawsky-Berger points out that being able to effectively formulate the problems requires domain expertise to identify what the important problems in an area are, how to formulate the questions properly and how to present the results so that they are useful to the domain practitioner [12]. In addition to the technical skills, Dhar also points out that a significant change in a manager's mindset from intuition and past practices to a data-driven decision-making approach is required. Here he reminds us of a quote from W.E. Demming "In God we trust—everyone else please bring data".

While there certainly have been many successes reported in the application of data science in industry, the financial sector, healthcare, transportation, education, professional services, etc., there have also been a number of reports about limitations and failures that have been experienced. An article in Wired Magazine [13] makes the controversial claim that the abundance of data that is becoming available will make the hypothesize–model–test approach to science obsolete. The latter was necessary in the age when only small samples of data were available, but now scientists have access to the entire population and therefore do not need statistics or theory. However, as Reis et al. point out [14], a sample no matter how big, may not accurately reflect the target population and give an excellent

example related to the 1936 US presidential election to illustrate the point. This supports the point that domain knowledge is needed even when massive data are available which is particularly true in the process industries. They go on to discuss a number of challenges that need to be considered. These include the issue of meaningful data referring to the difference between happenstance data, which may be suitable for process monitoring or fault detection, but not for prediction, where data from designed experiments for process optimization or system identification experiments for process control may be necessary. In addition, it may be possible that a very large industrial data set may be information poor, when interesting information happens only infrequently. They also discuss issues related to multiple data structures, heterogeneous data, multiple data management systems, the incorporation of a priori knowledge, uncertainty data, unstructured variability, data with high time resolution and adaptive fault detection and diagnosis. They conclude that while big data has the potential to be of great benefit in the process industries there are many issues that still need to be addressed and that data science and domain knowledge should be used synergistically.

A number of publications [15–18] also discuss organizational and managerial issues that need to be addressed to prevent failure of data science applications. These include:

- Lack of domain knowledge
- Lack of a trained workforce
- Lack of a clear vision and objective
- Data science being viewed as a technical and not a business initiative
- Data science for the sake of data science
- Failure to organize for big data
- Organizations that are resistant to change.

For data science to be successfully deployed in an organization, data scientists and data science projects must be managed properly, including deploying data scientists in the right spots, and giving them the tools and opportunities to make a convincing case.

4. Data Science in Chemical Engineering

There are of course many examples of the use of data science methodologies in the chemical engineering community that have been ongoing for many years, including multivariate analysis, on-line fault detection, inferential sensors, batch data analytics, experimental design approaches, parameter estimation and model discrimination. Chiang et al. [4] mention in their review paper that the process industries were for example early adopters of computer-based control. They point to the use of multivariate methods including principle component analysis, partial least squares and canonical variate analysis which have been used to analyze large volumes of data to develop predictive models and for fault detection. They provide examples from five different industries including chemicals, energy, semi-conductors, pharmaceuticals and food. A number of technical challenges, software platform challenges and culture challenges are also addressed. They conclude that the application of data science methods requires additional skills outside of the traditional chemical engineering curriculum and point to the large number of Master's programs in data science that have been introduced. Qin [19] points out that for well understood chemical mechanisms, first-principle mechanistic models can be derived for process operations, but that processes for which mechanisms are not well understood, data analytics are a valuable tool for gaining insight and developing predictive models. A four-part series in *Chemical Engineering Progress* 2016 special issue on big data analytics discusses what big data is [20], gives some success stories [21], describes how to get started [22] and what the challenges are [14]. In their 2016 paper, Beck et al. [23] address the question "What is Data Science and Why Should Chemical Engineers Care About It?" They also discuss a number of research areas in chemical engineering that have benefited from data science including computational molecular science and engineering, synthetic biology and energy systems and management. Finally,

Holdaway [24] uses specific case studies to show how data analytics can be used for optimization in exploration, development, production and rejuvenation of oil and gas assets.

5. Curriculum Implications

Chemical engineers have a very strong background in mathematics and problem solving and are therefore well poised to engage in data science. Historically, chemical engineering undergraduate programs, certainly in Canada, have also incorporated some form of statistical training. This has been accomplished by incorporating one or more courses in applied statistics into the undergraduate core curriculum, in many cases taught by chemical engineering faculty. However, it can be argued that chemical engineering education and training may not have kept pace with the proliferation of data described above and that therefore courses tend to teach approaches used in a data-poor era.

An informal survey of twelve chemical engineering programs in Canadian schools show that all contain a fundamental probability and statistics course, an introductory computing programming course, a course on engineering computation/numerical methods and in some a course on advanced statistics, usually experimental design taught as an elective. As mentioned above, the skills required to apply data science methods include data access and management, databases and data warehousing, statistical methods including classification and clustering, time series, various regression methods and multivariate statistics and data visualization. It is unlikely that existing courses are sufficient to cover all the required topics. The question then is what can be done to better prepare graduate chemical engineers for the realities of today when it comes to data analysis?

Beck et al. [23] propose that this can be accomplished by making small "tweaks" to the existing curriculum to include data science methods to existing courses, by adding elective course work or professional development workshops, or via the use of free on-line self-guided tutorials. The key is to ensure of course that this material is not added to the detriment of the core chemical engineering curriculum since as mentioned in Section 3 above, domain knowledge is an important component required for the successful outcome of data science projects. One also has to remember however that there are many pressures on engineering programs to add additional material on topics such as Life Cycle and Socio-Economic Analyses, Life Sciences, Nanotechnology, Renewable Energy, Advanced Materials and Additive Manufacturing, Virtual and Augmented Reality, etc.

The EDISON project [25] links the data science competencies listed in Section 2 above to learning outcomes and even proposes courses that could be used in a Master's program. There is also some discussion on how to accommodate students with diverse educational backgrounds by assessing their competencies and having students take pre-requisite courses and bootcamps. An approach offered by, for example the University of Calgary is to offer a Data Science Minor, which is taken co-currently with the student's particular core program. Another variant on this is the Certificate in Data Analytics offered by Ryerson University's Chang School of Continuing Education in which a six-course compressed program is offered to students who already have completed an undergraduate degree. The latter is a very high-touch program which provides one-on-one support, especially for students whose background does not include the competencies normally required for data science studies.

While adding data science components to an existing undergraduate program may be beneficial, examining the competencies required would indicate that acquiring a strong background may require additional studies. Targeted Master's programs in Data Science may be one option; however, many chemical engineering departments have faculty working in the area of process systems engineering and offer graduate programs which allow for the interdisciplinary training that is required.

Funding: This research received no external funding.

Acknowledgments: The author would like to acknowledge the background research for this paper carried out by Anne-Marie Brinsmead, Program Director, Engineering, Architecture and Science in the Chang School for Continuing Education at Ryerson University.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Donoho, D. 50 Years of Data Science. *J. Comput. Graph. Stat.* **2017**, *26*, 745–766. [CrossRef]
2. Schaffhauser, D. Purdue to Embed Data Science into Every Major, Campus Technology 2018. Available online: https:/campustechnology.com/articles/2018/04/19/purdue-to-embed-data-science-into-every-major.aspx (accessed on 24 August 2018).
3. Information and Communications Technology Council, Big Data and the Intelligence Economy. 2015. Available online: https://www.ictc-ctic.ca/wp-content/uploads/2015/12/BIG-DATA-2015.pdf (accessed on 6 November 2019).
4. Chiang, L.; Lu, B.; Castillo, I. Big Data Analytics in Chemical Engineering. *Annu. Rev. Chem. Biomol. Eng.* **2017**, *8*, 63–85. [CrossRef] [PubMed]
5. Beck, D.; Pfaendtner, J.; Carothers, J.; Subramanian, V. Data Science for Chemical Engineers. *Chem. Eng. Prog.* **2017**, *113*, 21–26.
6. Demchenko, Y. (Ed.) EDISON Data Science Framework: Part 1. Data Science Competence Framework (CF_DS) Release 3 v.10. 2018. Available online: https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome (accessed on 28 October 2019).
7. Wu, C.W.J. Statistics = Data Science. 1997. Available online: http://www2.isye.gatech.edu/~{}jeffwu/presentations/datascience.pdf (accessed on 21 October 2018).
8. Tukey, J.W. The future of data analysis. *Ann. Math. Stat.* **1962**, *33*, 1–67. [CrossRef]
9. Cleveland, W.S. Data Science: An Action Plan for Expanding the Technical Areas in the Field of Statistics. *Int. Stat. Rev.* **2001**, *69*, 21–26. [CrossRef]
10. Granville, V. Available online: http:/www.datasciencecentral.com/profiles/blogs/data-science-without-statistics-is-possible-even-desirable (accessed on 24 August 2018).
11. Dhar, V. Data Science and Prediction. *Commun. ACM* **2013**, *56*, 64–73. [CrossRef]
12. Wladowsky-Berger, I. Why Do We Need Data Science When We've Had Statistics for Centuries? *Wall Street J.* Available online: https://blogs.wsj.com/cio/2014/05/02/why-do-we-need-data-science-when-weve-had-statistics-for-centuries (accessed on 24 August 2018).
13. Anderson, C. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired Mag.* **2008**. Available online: www.wired.com/2008/06/pb-theory (accessed on 23 June 2008).
14. Reis, M.S.; Braatz, R.D.; Chiang, L.H. Big Data Challenges and Future Research Directions. *Chem. Eng. Prog.* **2016**, *112*, 46–50.
15. Redman, T.C. Are You Setting Your Data Scientists Up to Fail. *Harvard Bus. Rev.* **2018**. Available online: https://hbr.org/2018/01/are-you-setting-your-data-scientists-up-to-fail (accessed on 25 January 2018).
16. Demirkan, H.; Dal, B. The Data Economy: Why Do So Many Analytics Projects Fail? *Anal. Mag.* **2014**. Available online: http://analytics-magazine.org/the-data-economy-why-do-so-many-analytics-projects-fail/ (accessed on 26 August 2018).
17. Elprin, N. Data Science: 4 Reasons Why Most Are Failing to Deliver. Available online: https://kdnuggets.com/2018/05/data-science-4-reasons-failing-deliver.html (accessed on 24 August 2018).
18. 70% of Big Data Projects in the UK Fail to Realize Full Potential. Available online: https://www.consultancy.uk/news/16839/70-if-big-data-projects-in-uk-fail-to-realise-full-potential (accessed on 24 August 2018).
19. Qin, S.J. Process data analytics in the era of big data. *AIChE J.* **2014**, *60*, 3092–3100. [CrossRef]
20. White, D. Big Data: What is it? *CEP Mag.* **2016**, *112*, 33–35.
21. Garcia-Munoz, S.; MacGregor, J.F. Big Data: Success Stories in the Process Industries. *CEP Mag.* **2016**, *112*, 36–40.
22. Colegrove, L.F.; Seasholtz, M.B.; Khare, C. Big Data: Getting Started on the Journey. *CEP Mag.* **2016**, *112*, 41–45.
23. Beck, D.A.C.; Carothers, J.M.; Subramanian, V.R.; Pfaendtner, J. Data Science: Accelerating Innovation and Discovery in Chemical Engineering. *AIChE J.* **2016**, *62*, 1402–1416. [CrossRef]

24. Holdaway, K.R. *Harness Oil and Gas Big Data with Analytics: Optimization, Exploration and Production with Data Driven Models*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
25. Demchenko, Y. (Ed.) EDISON Data Science Framework: Part 3. Data Science Model Curriculum (MC-DS) Release 3 v.5. 2018. Available online: https://github.com/EDISONcommunity/EDSF/wiki/EDSFhome (accessed on 28 October 2019).