MRS Advances © 2020 Materials Research Society
DOI: 10.1557/adv.2020.135

Learnings from developing an applied data science curricula for undergraduate and graduate students

Roger H. French [1,2,3,4] and Laura S. Bruckman [1,2]

[1] *SDLE Research Center, Case Western Reserve University, Cleveland OH, 44106*

[2] *Dept. of Materials Science & Engineering, Case Western Reserve University, Cleveland OH, 44106*

[3] *Dept. of Macromolecular Science & Engineering Case Western Reserve University, Cleveland OH, 44106*

[4] *Dept. of Computer & Data Sciences, Case Western Reserve University, Cleveland OH 44106*

ABSTRACT

Data science has advanced significantly in recent years and allows scientists to harness large-scale data analysis techniques using open source coding frameworks. Data science is a tool that should be taught to science and engineering students in addition to their chosen domain knowledge. An applied data science minor allows students to understand data and data handling as well as statistics and model development. This move will improve reproducibility and openness of research as well as allow for greater interdisciplinarity and more analyses focusing on critical scientific challenges.

TRANSFORMATIVE CHANGES DRIVING DATA SCIENCE, BIG DATA ANALYTICS AND DEEP LEARNING

Digital Transformation and Big Data Analytics

Data science has arisen from combined advances in computing, communication, and data that are driving the digital transformation [1-3]. Digital transformation has benefited from the computer science concepts developed by organizations such as Google, which have enabled big data analytics and led to the

development of the open source projects such as Hadoop, Hbase, and most recently Spark. These projects allow for data-driven modeling of massive petabyte scale datasets [4-7]. These "distributed computing" approaches and the ease of acquiring petabyte scale datasets have given rise to Facebook and other data-centric technologies and enable the digital transformation across industry, science and technology, and society itself. Distributed computing complements the petaflop computing characteristic of high performance computing allowing for the rise of a new computing paradigm of distributed and high performance computing (D&HPC).

Openness

Another major driver of change has been the move towards "openness" [8] in which restrictive copyrights and licenses limit innovation and creativity [9]. Several key events have shown that the open source approach drives collaboration and community development, and accelerates innovation [10,11]. These events include:

- the establishment of the Free Software Foundation in the 1980's;
- the initial release of the Linux kernel by Linus Torvalds in 1991 as an open source version of Unix;
- the release of R an open source version of the S language for data analysis in the 1992; and
- the initial Python release by Guido Rossum in 1991 as an open source language.

Today most software companies acknowledge the benefits of open source communities and software [12], with Python and R the dominant programming languages used in data science. Another important contributor to the advance of data science is open data, whereby governments (e.g., the US [13] and the G8 nations [14]) make data openly available. In science, the recent goal is open access to the results of government-funded research [15]. The trend in scientific publishing is to have datasets along with the data analysis codes used in journal articles openly available under an open source license to aide scientific reproducibility and contribute to the furtherance of science [16,17]. Today openness is the pathway to better science [18].

Deep learning

The introduction of graphics processing units (GPUs) [19] and even tensor processing units (TPUs) [20] for fitting large, nonlinear neural network models of large datasets, has led to the introduction of deep learning [21], which goes beyond typical neural network modeling. Deep learning was initially guided by the ImageNet dataset assembled by Fei-Fei Li [22] and the associated annual ImageNet Large Scale Visual Recognition Challenge [23]. This challenge included an initial collection of 1 million labeled images (now increased to 14 million) and competing teams of computer scientists to develop the best performing models able to predict the contents of these images [24]. ImageNet ran from 2006-2017 and the research on the ImageNet dataset continues to inform the development of deep learning [25]. Another step in the development of deep learning has been the development of deep learning frameworks such as Theano [26] (which is no longer in active development) and TensorFlow [27], and interface packages such as Keras that enable building deep learning models using Python or R languages [28]. The field of deep learning is very active and its full capabilities are the focus of much discussion [29,30]. The most recent advance is AlphaGo Zero [31,32], a TensorFlow deep learning machine designed to play Go, has beaten the best Go players, an unexpected outcome. All of these elements and aspects highlight the fast pace of

change in the data science field. The ultimate capabilities and applications of data science are not clearly evident, but have already been transformational.

CHALLENGES IN SCIENCE THAT DEMAND NEW APPROACHES

Reproducibility

Reproducibility in science, or the lack thereof, has driven the call for open sharing of codes and datasets, to facilitate replication of published results [33]. This has been a concern for many years, but more recently the concern around reproducibility has led to a focus on the methods of data analysis and interpretation [34-36].

Challenges at the Frontiers of Materials Science

Just as the advances in computer science have provided new tools and approaches, materials science is constantly striving to develop new approaches to address the arising challenges in materials to enable many critical societal advancements. For example materials science has a critical role to play in sustainability and climate change [37], in the areas of solar energy, [38,39] and energy storage [40].

MATERIALS DATA SCIENCE, THE NEXT STEP IN INNOVATION

In materials science, the ubiquity of high performance petaflop computing gave rise to integrated computational materials engineering (ICME) initiatives in 2008 [41] and the materials genome initiative (MGI) in 2011 [42]. One challenge for the current development of materials data science is that typically materials science datasets have been small and sparse in comparison to the datasets developed in epidemiological studies in the life sciences. This has meant that often materials science research is inherently observational, since large enough datasets to satisfy the central limit theorem are relatively uncommon [43].

Data science and the digital transformation are playing a major role in the transformation of industry referred to as Industry 4.0 [44,45]. This is where the ease of acquiring data in the factory, combined with the above mentioned arrival of distributed and high performance computing, transform foundational aspects of manufacturing, materials processing, and product assembly and qualification [46,47]. We are just now at the beginning of this Industry 4.0 transition and the implementation of the digital transformation is still currently unclear.

Data Science combines advances in statistics, computer science, and domain science (e.g., materials science) to enable new understandings through the application of statistical and machine learning and most recently deep learning. There is a need to develop broader data science skills across the workforce to produce T-shaped graduates who have deep skills in a domain science area (e.g., materials science, physics, chemistry), while also having broad skills in data science [48].

Applied and Materials Data Science Programs

In 2013, we launched a 1 year study to design an applied data science (ADS) undergraduate minor, available to students across Case Western Reserve University. These ADS students learn programming, inferential statistics, exploratory data analysis, modeling and prediction, and complete a semester long data science project [49]. The

ADS minor started in academic year 2015 and has grown to include 100 undergraduate and graduate students this academic year. The ADS curricula is taught using an open data science tool chain focused on open and reproducible science, based on R [50] / Rstudio [51,52], Python [53-55], Git [56-58], Markdown [59], and LaTeX [60] to produce, compilable and reproducible data analyses. In R, advances such as the TidyVerse package of pipes and pipelined code [61] and ggplot2 for data visualization [62] are major steps towards realizing Donald Knuth's vision of literate programming and are well matched to today's multi-disciplinary team research [63]. Students end the ADS minor with a semester long data science project. The student must take a dataset through cleaning, exploring, and analyzing to draw conclusions about the data and be able to communicate their data analysis to a varied audience.

For materials data science, we now offer a data science concentration, focusing the ADS courses on materials problems while addressing the core challenges of integrating data science with the physical and chemical sciences foundations of materials science. Essential to adoption of data-driven modeling is demonstrating that these models do not replace physical and chemical theories, models, and experimental experience. Instead, data-driven modelling is a tool that adds statistical power and significance with improved inference and prediction. These material data analyses must be subject to robust validation, using training and testing splits of the data.

Materials data science is not only an educational challenge, but also calls for advancing how we perform our research experiments and acquire data for analysis. A study protocol, encompassing the samples, their exposures and the evaluations performed on them, constitutes the basis of the metadata, the predictors and the responses of the experiment. In many experiments, it is possible to augment the experiment with additional predictors measured in sufficient numbers to provide statistically sound results. Additionally, capturing the data from experiments that were successful as well as those that failed is important to reduce the success bias in datasets. Having materials scientists knowledgeable about these data issues is important to advancing materials research methods.

CONCLUSION

The advances in distributed computing and open source tools has moved materials data science forward. These advances are needed in university curricula and classrooms to develop future scientists capable of handling large data analysis problems and communicate these data science results to a broad audience. This requires students to learn open source coding languages and to have hands on practice with various types of real world datasets. Data science exposure for domain specific science students will benefit students as science becomes more multi-disciplinary.

ACKNOWLEDGMENTS

The authors acknowledge the helpful input provided by our corporate and industrial collaborators and the Business Higher Education Forum. The support and contributions of CWRU's Applied Data Science Faculty and the Materials Science Faculty and undergraduate and graduate students have aided the development of the Applied and Materials Data Science curricula.

REFERENCES

[1] T. Wackler: Strategy for American Leadership in Advanced Manufacturing, National Science and Technology Policy, White House, 40 (2018). https://www.whitehouse.gov/wp-content/uploads/2018/10/Advanced-Manufacturing-Strategic-Plan-2018.pdf. (accessed 4 January 2020).

[2] B. Weinelt: Digital Transformation Initiative, World Economic Forum, (2015). http://wef.ch/2hU0x7I (accessed 4 January 2020).

[3] R. Grossman, The Industries That Are Being Disrupted the Most by Digital, *Harvard Business Review*, (2016). https://hbr.org/2016/03/the-industries-that-are-being-disrupted-the-most-by-digital (accessed January 4, 2020).

[4] M. I. Jordan, editor, Frontiers in Massive Data Analysis, National Research Council, National Academies Press, (2013). http://www.nap.edu/catalog.php?record_id=18374. (accessed 4 January 2020).

[5] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M. Burrows, T. Chandra, A. Fikes, R.E. Gruber, Bigtable: A distributed storage system for structured data, *ACM Transactions on Computer* Systems, 26, 4 (2008). http://dl.acm.org/citation.cfm?id=1365816. (accessed January 26, 2016).

[6] R.C. Taylor, An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics, *BMC Bioinformatics*. 11, S1 (2010). http://www.biomedcentral.com/1471-2105/11/S12/S1. (accessed October 28, 2014).

[7] M. Zaharia, R.S. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M.J. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, I. Stoica, Apache Spark: A Unified Engine for Big Data Processing, Commun. *ACM*. 59, 56-65 (2016). https://doi.org/10.1145/2934664. (accessed 4 January 2020).

[8] E. Maxwell: Harnessing Openness to Improve Research, Teaching and Learning in Higher Education. *Innovations: Technology*, *Governance*, *Globalization*, 5(2), 155 (2010). http://dx.doi.org/10.1162/inov_a_00019. (accessed 4 January 2020).

[9] E. Maxwell, Open Standards, Open Source, and Open Innovation: Harnessing the Benefits of Openness, *Innovations: Technology, Governance, Globalization,* 1, 119–176 (2006). https://doi.org/10.1162/itgg.2006.1.3.119. (accessed 4 January 2020).

[10] D. C. Ince, L. Hatton, and J. Graham-Cumming: The case for open computer programs. *Nature*, 482, 7386, 485 (2012). http://www.nature.com/nature/journal/v482/n7386/full/nature10836.html. (accessed 4 January 2020).

[11] J. Andraka: Open Access: The Pathway to Innovation, OSTP, (2013). https://obamawhitehouse.archives.gov/blog/2013/06/20/open-access-pathway-innovation. (accessed 4 January 2020).

[12] J. S. S. Lowndes, B. D. Best, C. Scarborough, J. C. Afflerbach, M. R. Frazier, C. C. O'Hara, N. Jiang, and B. S. Halpern: Our path to better science in less time using open data science tools. *Nat. Ecol. Evol.*, 1(6), 160 (2017). https://dx.doi.org/10.1038/s41559-017-0160. (accessed 4 January 2020).

[13] B. Obama: Executive Order -- Making Open and Machine Readable the New Default for Government Information, The White House (2013). https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-. (accessed 4 January 2020).

[14] Group of 8 (G8): G8 Open Data Charter and Technical Annex (Gov.UK), (2013). https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex. (accessed 4 January 2020).

[15] J. P. Holdren: Increasing Access to the Results of Federally Funded Scientific Research, Executive Office of the President: Office of Science and Technology Policy, (2013). https://obamawhitehouse.archives.gov/blog/2016/02/22/increasing-access-results-federally-funded-science. (accessed 4 January 2020).

[16] C. Wadia, M. Stebbins: It's Time to Open Materials Science Data, Executive Office of the President: Office of Science and Technology Policy, (2015). https://obamawhitehouse.archives.gov/blog/2015/02/06/its-time-open-materials-science-data. (accessed 4 January 2020).

[17] F. S. Collins and L. A. Tabak, "Policy: NIH plans to enhance reproducibility," *Nature*, 505, 7485, 612–613, (Jan. 2014). http://www.nature.com/news/policy-nih-plans-to-enhance-reproducibility-1.14586. (accessed 4 January 2020).

[18] H. V. Fineberg, "Reproducibility and Replicability in Science," National Academies Press, (May 2019) https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science. (accessed 4 January 2020).

[19] Y.E. Wang, G.-Y. Wei, D. Brooks, Benchmarking TPU, GPU, and CPU Platforms for Deep Learning, *ArXiv*:1907.10701 [Cs, Stat]. (2019). http://arxiv.org/abs/1907.10701 (accessed January 8, 2020).

[20] N.P. Jouppi, et al., In-Datacenter Performance Analysis of a Tensor Processing Unit, *ArXiv*:1704.04760 [Cs]. (2017). http://arxiv.org/abs/1704.04760 (accessed January 8, 2020).

[21] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature*. 521, 436-444 (2015). https://doi.org/10.1038/nature14539. (accessed 4 January 2020).

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, *Proc. of IEEE Computer Vision and Pattern Recognition*, 8, (2009). https://wordnet.cs.princeton.edu/papers/imagenet_cvpr09.pdf. (accessed 4 January 2020).

[23] ImageNet, (n.d.). http://image-net.org/ (accessed January 8, 2020).

[24] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, 1097-1105, (2012). https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf. (accessed 4 January 2020).

[25] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, *ArXiv*:1409.1556 [Cs]. (2014). http://arxiv.org/abs/1409.1556. (accessed 4 January 2020).

[26] R. Al-Rfou, et al., Theano: A Python framework for fast computation of mathematical expressions, *ArXiv*:1605.02688 [Cs]. (2016). http://arxiv.org/abs/1605.02688 (accessed January 8, 2020).

[27] M. Abadi, et al., TensorFlow: A System for Large-Scale Machine Learning, *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, 265-283, (2016). https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi (accessed January 8, 2020).

[28] F. Chollet, J. J. Allaire, *Deep Learning with R*, Manning Publications, (2018). https://www.manning.com/books/deep-learning-with-r (accessed May 29, 2019).

[29] G. Marcus, Deep Learning: A Critical Appraisal, *ArXiv*:1801.00631 [Cs, Stat]. (2018). http://arxiv.org/abs/1801.00631 (accessed January 8, 2020).

[30] J. Dean, The Deep Learning Revolution and Its Implications for Computer Architecture and Chip Design, *ArXiv*:1911.05289 [Cs, Stat]. (2019). http://arxiv.org/abs/1911.05289 (accessed January 8, 2020).

[31] D. Silver et al., "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, (Oct. 2017). https://www.nature.com/articles/nature24270. (accessed 4 January 2020).

[32] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature*, 529, 7587, 484–489, (Jan. 2016). https://www.nature.com/articles/nature16961. (accessed 4 January 2020).

[33] E. E. David, Jr.: Responsible Science, Volume I: Ensuring the Integrity of the Research Process, National Academies Press, (1992). http://www.nap.edu/catalog/1864/responsible-science-volume-i-ensuring-the-integrity-of-the-research. (accessed 4 January 2020).

[34] R. D. Peng: Reproducible Research in Computational Science. *Science,* 334, 6060, 1226 (2011). https://dx.doi.org/10.1126/science.1213847. (accessed 4 January 2020).

[35] Announcement: Reducing our irreproducibility. *Nature,* 496(7446), 398 (2013). http://www.nature.com/news/announcement-reducing-our-irreproducibility-1.12852. (accessed 4 January 2020).

[36] J. T. Leek and R. D. Peng: Statistics: P values are just the tip of the iceberg. *Nature*, 520, 7549, 612 (2015). http://www.nature.com/doifinder/10.1038/520612a. (accessed 4 January 2020).

[37] A. Guterres, "The Sustainable Development Goals Report 2018," United Nations, Department of Economic and Social Affairs, (2018) https://www.un.org/development/desa/publications/the-sustainable-development-goals-report-2018.html. (accessed 4 January 2020).

[38] R. H. French et al., "Degradation science: Mesoscopic evolution and temporal analytics of photovoltaic energy materials," *Current Opinion in Solid State and Materials Science*, 19, 4, 212–226, (Aug. 2015). http://www.sciencedirect.com/science/article/pii/S1359028614000989. (accessed 4 January 2020).

[39] H. E. Yang, R. H. French, L. S. Bruckman, Eds., *Durability and Reliability of Polymers and Other Materials in Photovoltaic Modules*, 1st Edition. Amsterdam: Elsevier, William Andrew Applied Science Publishers, (2019). https://www.sciencedirect.com/book/9780128115459/durability-and-reliability-of-polymers-and-other-materials-in-photovoltaic-modules. (accessed 4 January 2020).

[40] International Energy Agency, *World Energy Outlook 2019*, (2019). https://www.iea.org/weo/weo2019/secure/data/. (accessed 4 January 2020).

[41] T. M. Pollock: Integrated Computational Materials Engineering, National Academies Press, (2008). https://nae.edu/25043/Integrated-Computational-Materials-Engineering. (accessed 4 January 2020).

[42] J. P. Holdren: Goals of the Materials Genome Initiative (2011). https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf. (accessed 4 January 2020).
[43] R.M. Dudley, R.M. Dudley, *Uniform Central Limit Theorems*, Cambridge University Press, (1999). https://doi.org/10.1017/CBO9780511665622. (accessed 4 January 2020).
[44] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann: Industry 4.0. *Business & Information Systems Engineering*, 6, 4, 239 (2014). DOI: 10.1007/s12599-014-0334-4. (accessed 4 January 2020).
[45] L. D. Xu, E. L. Xu, and L. Li: Industry 4.0: State of the Art and Future Trends. *International Journal of Production Research*, 56, 8, 2941 (2018). DOI: 10.1080/00207543.2018.1444806. (accessed 4 January 2020).
[46] J. Lee, B. Bagheri, and H.-A. Kao: A Cyber-Physical Systems Architecture for Industry 4.0-based Manufacturing Systems. *Manufacturing Letters*, 3, 18 (2015). http://dx.doi.org10.1016/j.mfglet.2014.12.001. (accessed 4 January 2020).
[47] Y. Lu: Industry 4.0: A Survey on Technologies, Applications and Open Research Issues. *Journal of Industrial Information Integration*, 6, 1 (2017). DOI: 10.1016/j.jii.2017.04.005
[48] D. Hughes and R. H. French, "Crafting a Minor to Produce T-Shaped Graduates," National Academies, Washington DC, 21 March 2016. http://tsummit.org/files/T-Summit_Speaker_Abstracts-2016.pdf. (accessed 4 January 2020).
[49] Business Higher Education Forum, "Creating a Minor in Applied Data Science | BHEF," The Business Higher Education Forum, Case Study, Aug. 2016. Available: http://www.bhef.com/publications/creating-minor-applied-data-science. (accessed 4 January 2020).
[50] R Core Team, "R: The R Project for Statistical Computing", (2019). https://www.r-project.org/. (accessed 4 January 2020).
[51] *RStudio: Integrated Development Environment for R*, RStudio, Inc., Boston, MA (2015). http://www.rstudio.com/. (accessed 4 January 2020)
[52] H. Wickham, G. Grolemund, "R for Data Science: Import, Tidy, Transform, Visualize, and Model Data", 1 edition, O'Reilly Media, (2017). http://r4ds.had.co.nz/. (accessed 4 January 2020).
[53] G. van Rossum, *Python tutorial, technical report CS-R9526*, National Research Institute for Mathematics and Computer Science, Amsterdam, The Netherlands (1995), p.71. https://ir.cwi.nl/pub/5007/05007D.pdf. (accessed 4 January 2020).
[54] G. Van Rossum and Fred L. Drake, *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA (2009).
[55] Python Software Foundation: Python 3.8.1 documentation", (n.d.). https://docs.python.org/3.8/contents.html. (accessed 4 January 2020).
[56] H. Van Styn, Git – Revision Control Perfected, *Linux Journal*, 208 (2011). https://www.linuxjournal.com/content/git-revision-control-perfected. (accessed 4 January 2020).
[57] Z. Brown, A Git Origin Story, *Linux Journal,* 288 (2018). https://www.linuxjournal.com/content/git-origin-story. (accessed 4 January 2020).
[58] K. Ram, "Git can facilitate greater reproducibility and increased transparency in science," *Source Code for Biology and Medicine*, 8, 1, 7, (Feb. 2013). https://doi.org/10.1186/1751-0473-8-7. (accessed 4 January 2020).
[59] A. Swartz, "Aaron Swartz's A Programmable Web: An Unfinished Work," Synthesis Lectures on the Semantic Web: Theory and Technology, 3, 2, 1–64, (Feb. 2013). https://www.morganclaypool.com/doi/abs/10.2200/S00481ED1V01Y201302WBE005. (accessed 4 January 2020).
[60] M. Kline, Modern LaTeX, 2nd Ed. (2018). https://assets.bitbashing.io/modern-latex.pdf. (accessed 4 January 2020).
[61] H. Wickham et al., "Welcome to the Tidyverse," Journal of Open Source Software, vol. 4, no. 43, p. 1686, (Nov. 2019). https://joss.theoj.org/papers/10.21105/joss.01686. (accessed 4 January 2020).
[62] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, 2nd ed. Springer International Publishing, (2016). https://www.springer.com/gp/book/9783319242750. (accessed 4 January 2020).
[63] D. E. Knuth, "Literate Programming," Comput J, 27, 2, 97–111, (Jan. 1984). https://academic.oup.com/comjnl/article/27/2/97/343244/Literate-Programming. (accessed 4 January 2020).