Top 10 Essential Data Science Topics to Real-World Application from the Industry Perspectives

Victor S.Y. Lo[1]

[1]AI & Data Science Center of Excellence, Workplace Investing, Fidelity Investments

1. Introduction

Statistics and data science are more popular than ever in this era of data explosion and technological advances. Decades ago, John Tukey (Brillinger, 2014) said, "The best thing about being a statistician is that you get to *play in everyone's backyard.*" More recently, Xiao-Li Meng (2009) said, "We no longer simply enjoy the privilege of playing in or cleaning up everyone's backyard. We are now being invited into *everyone's study or living room,* and trusted with the task of being their offspring's first quantitative nanny."

Professors Wing, He, and Lin have written two excellent summaries of *research* opportunities for computer scientists and statisticians. *Research* and *education* offer key knowledge and talents (*supply*) to the society and industry. We provide comments on Wing (2020) and He & Lin (2020) from the industry perspectives and introduce additional topics that are *practical* and *important* to meet the *industry demand,* and hence are also key to career development for students and professionals. Our discussion may also be helpful for researchers interested in industry applications. An analytics framework is first introduced below to support our comments.

2. Three Types of Analytics and Their Link to Causality

Figure 1 provides an industry framework for three types of analytics. Traditional statistical inference is more about taking a random sample to learn about the population, including data relationships. Prediction is typically not the primary objective of statistical inference, but the industry today tends to welcome predictive analytics. The top level, prescriptive analytics, is about guiding *decision making* if you know something may happen. For example, in weather forecasting, reporting what the past weather and identifying patterns by region are descriptive. Weather forecast is a form of predictive analytics. For prescriptive analytics, if the forecast says there will be a big snowstorm tomorrow, you assess the impact of going to work/school versus that of working from home. Given experience, you may have a causal assessment of each decision on safety and work effectiveness. Then you choose the right decision by balancing between safety and benefit. In other words, prescriptive analytics needs understanding of what may happen if we take action A as opposed to B, requiring causal inference to estimate the effect of each action. See Bojinov et al. (2020) for further causality discussion.
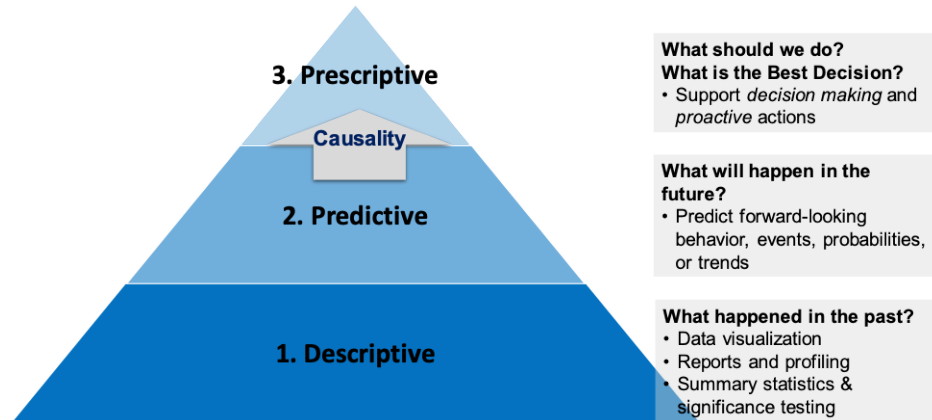
Figure 1. Three Types of Analytics

3. Data Science Project Process and Typical Skill Requirement

Figure 2 describes a typical data science project, similar to Wing's (2019) "Data Life Cycle"—starting with analytic consulting to understand the problem and define scope, then gathering and processing data. Next, models (analytics) are developed, with insights extracted and reports presented. Finally, models are deployed by the business and implemented in a system. These stages require four skill sets: Computer Science, Statistics/Mathematics, Subject Matter Expertise (important for the application area), and Soft Skills (*see* Lo, 2019). Our 10 topics will be centered around these.
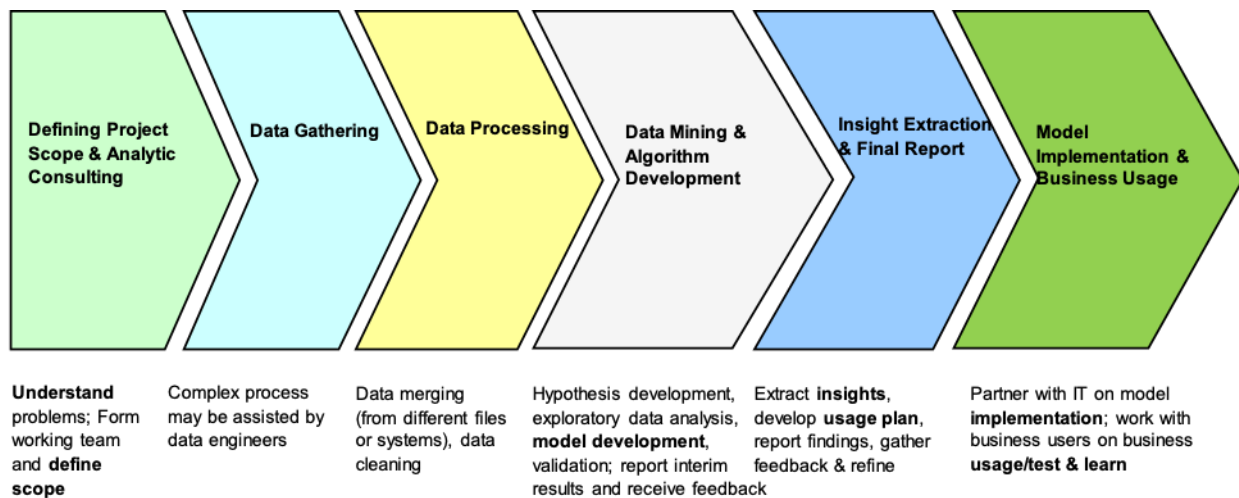

Figure 2. Data Science Project Process

Table 1. Causality and Optimization Questions in Marketing

| 4P's	Causal Inference	Prescriptive Analytics
Price	Would a price discount generate a higher demand?	What is the optimal price?
Promotion	What are the effects of direct marketing and advertising campaigns?	How to optimally allocate marketing investment to direct marketing and advertising?
Place	What is the impact of new store location on business outcomes?	Where to open new stores?
Product	Would an improvement in product feature be valuable to customers?	What are the best product features?
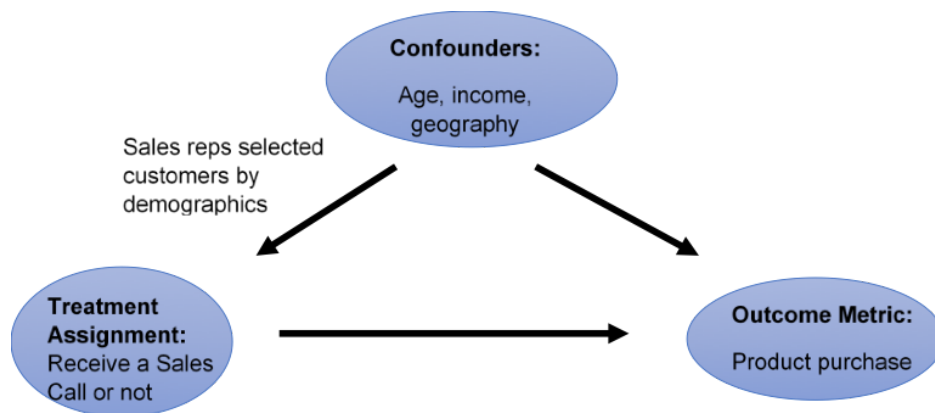
Figure 3. Sales Measurement using Observational Data

4. Comments on Wing and He & Lin Papers and Additional Topics

The first five topics below are in line with Wing and He & Lin, augmented with industrial perspectives and business examples. The next five topics are practical topics highly relevant to the industry that are mostly additional to their lists.

4.1. Causal Inference

He & Lin's seventh area, Causal Inference which falls between predictive and prescriptive analytics in Figure 1, is widely applicable not only to medicine but also marketing, political election, and policy. To illustrate its significance in business, see the middle column of Table 1 with examples based on the 4P's of marketing. While randomized controlled trial (RCT) is the *gold standard* for causal measurement

and is popular for business experiments such as online advertising (e.g., comparing advertisements A and B), it is not always feasible in practice. Absence of RCT, causal inference techniques are widely available. As a business example, you are interested in assessing the effect of a sales call on purchase rate using observational data, see Figure 3. Historically, sales associates tended to contact customers with characteristics associated with a higher purchase rate, e.g., older and wealthier. This is a classic confounding situation as treatment's link to outcome has a 'backdoor' path through the confounders. A common solution to handle this is through propensity score matching (see Rubin, 2006, and Imbens & Rubin, 2015). Methodologies developed by statisticians, epidemiologists, and economists tend to focus on estimating effects of causes. Another powerful branch of techniques from AI employs Directed Acyclic Graph and Bayesian Network to discover causal relationships, which is more taught in computer science (see Pearl, 2000, and Pearl and MacKenzie, 2019). Guidelines on usage of different approaches would be highly beneficial to researchers and practitioners.

4.2 Heterogeneous Treatment Effect / Uplift Modeling

Related to Causal inference is He & Lin's first area, Quantitative Precision-X. Causal inference is about measuring the overall treatment effect (Average Treatment Effect or ATE) while the latter aims at individual/subgroup treatment effect (Heterogeneous Treatment Effect or HTE) for *personalization*. Estimating HTE in business is known as Uplift Modeling pioneered by practitioners such as Radcliffe & Surry (1999) and Lo (2002; 2008), sharing similarity with subgroup analysis. Due to its business impact, it has grown to a subfield on its own with terminology, techniques (including validation), and packages dedicated to it, and has drawn wide attention not only from the industry but also academics who followed similar terminology and approaches. See Rzepakowski & Jaroszewicz (2012), Yong (2015), Zhao (2017), and Zhang et al. (2020).
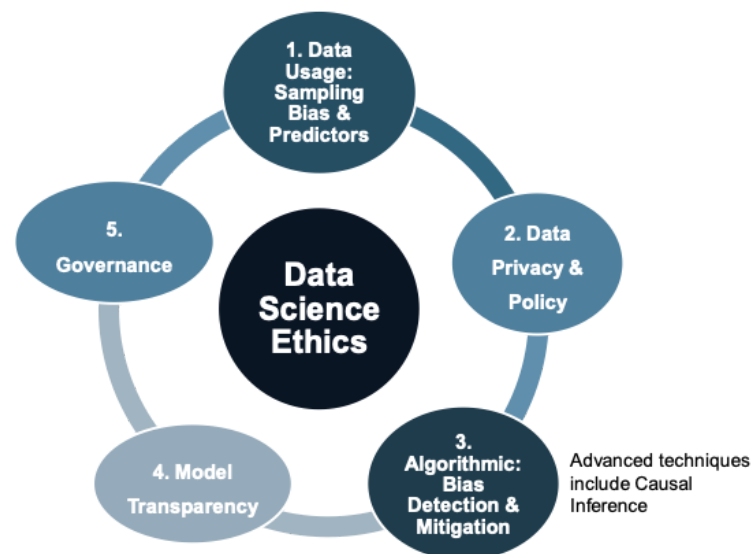

Figure 4. Multi-disciplinary Data Science Ethics

4.3. Data Science Ethics

The second area in He & Lin, Fair and Interpretable Learning and Decision Making, along with sixth (Trustworthy AI) and tenth(Ethics) areas in Wing are increasingly critical, grouped under AI/Data Science Ethics. This involves disciplines across quantitative and non-quantitative fields as outlined in Figure 4.

- When developing models, we should ask: is the training data unbiased? Are we using appropriate predictors that are ethically acceptable? Although we echo the Precious Data view in the third area of Wing that solid methods are needed to tackle the challenges of precious data, data scientists need to be mindful of the potential bias in *found* data.

- Data privacy and policy are necessary requirements. As mentioned in the ninth area of Wing, data scientists have been developing methods to protect individual identities. It may be helpful to mention that the collective approaches in this space are known as Privacy Preserving Data Mining, where data are transformed or perturbed through various ways prior to model development, while retaining value for knowledge discovery, as summarized in Aldeen et al. (2015) and Mendes & Vilela (2017).

- After a model is developed, can we detect whether it is biased against protected classes through fairness metrics? Can we algorithmically remove any bias?

- There is a general need for model transparency—would we rather develop an interpretable model OR a 'black box' that requires efforts to explain it? See Rudin (2019).

- Should we set up a governance process to oversee the data science process and address gray areas? See Sandler & Basl (2019).

None of the above above are simple, requiring expertise from multiple fields. See O'Neil (2017), Boddington (2017), ASA (2018), Russell (2019), IFoA and RSS (2019).

4.4. Deep Learning and Unstructured Data

Wing's first area starts with a profound understanding of deep learning, a pervasive technique with wide applications. While her point that we have a limited understanding of why it is so effective is true, early research by Hornik (1991) showed that multi-layer perceptron is a universal approximator that can mimic pretty much any functions, and modern techniques have taken it to the next level through several layers, weight sharing, and regularization. Although deep learning is mostly a subclass of predictive analytics (level 2 in Figure 1), it has broad impact especially for unstructured data

(images, text, and speeches) which represent majority of Big Data. These data are briefly mentioned in the tenth area of He & Lin and fourth of Wing, but the success of neural networks for unstructured data deserves a highlight—e.g., convolutional neural network for images, recurrent neural network for sequential data, and word embedding for converting text into numbers. Among the types of unstructured data, the largest opportunity today may belong to text, because lots of "data" in words are waiting to be analyzed by Natural Language Processing (NLP) and deep learning. For instance, legal and contractual documents are key data sources for legal and compliance analyses and hence are a natural opportunity for NLP to uncover rules and agreements. Similarly, doctors' notes, electronic health records, survey verbatim are also opportunities. A website owner may establish a user search capability for keywords, which requires NLP to interpret questions and return answers.

4.5. Computational Tools and Technology

We cannot execute projects without these. Reinsel et al. (2020) reported that the global data size would increase by ~3X in the next few years, or 175 ZB. With dramatic data growth and rise of deep learning, more computational power is needed. So, what are the computational technologies we should acquire? Wing's seventh point highlights the significance of computing systems for data-intensive applications and recommends new system designs with efficient data access and processing, and He & Lin's fifth point mentions cloud-based distributed statistical inference. While these are important to research, we would add the following important skills for industry applications:

- **Data Knowledge:** To analyze data properly, we must have a good understanding of the data. This is particularly important when data are large and reside in multiple sources, including structured and unstructured data. Knowing the data is often related to understanding the business.

- **Extract, Transform, and Load (ETL):** This is an essential skill especially when handling Big Data, and can often be assisted by professional data engineers.

- **Model Production:** While classical statistical analyses lead to findings that may not be deployed in a production environment, data science tends to complete the production cycle with ongoing prediction and decision support (see Figure 2). Modern model deployment involves containerization tools which make it robust and efficient to deploy scoring codes and adapt necessary changes, see Rao (2019), and can be incorporated in the Agile development process, see Kelleher and Kelleher (2019).

4.6. Analytic Consulting, Communication and Soft Skills

These are essential skills for industry applications as described in Kruhse-Lehtonen & Hofman (2020): "The highest level of AI maturity is when the whole company moves in unison, silos are dissolved, and data and AI are used by everyone as part of their daily business." We classify them as follows:

- **Business Consulting:** Academic programs may incorporate data science consulting, ranging from identifying opportunities, initiating projects, defining scope, drafting proposals, identifying resources, and ultimately leading project development and model deployment.

- **General Business Communication:** This includes understanding the business audience, speaking in their language, presenting in ways appealing to the business with visualization and storytelling. Some of these are part of business analytics programs.

- **Communication with IT Professionals:** As in Figure 2, data scientists are involved in multiple phases, partnering with data/tech professionals who have their own language. Being able to effectively communicate with them can lead to efficient development and deployment.

4.7. Descriptive Analytics

The first level of the analytics in Figure 1 is Descriptive Analytics, including data visualization. Although this was highly proposed by pioneers such as Florence Nightingale and John Tukey, the latest data science and statistics programs may have less focus on it.

- **Data Visualization and Statistical Graphics:** Visually attractive graphics are powerful for learning insights and sharing with business stakeholders, and they often require description, as mentioned by Unwin (2020) that "A picture is not a substitute for a thousand words; it needs a thousand words…" See also Dykes (2020) and Knaflic (2020) for data storytelling with graphics.

- **Reports, Summary Statistics, Profiling:** These are the essential tools for data scientists. For example, to learn about a customer base, one can run frequency distributions, crosstabs, and perform behavior analysis (on product ownership and usage, say). Significance testing can be used to assess differences in the data by customer group.

- **Feature Selection:** The above steps could serve as feature selection, an input to predictive analytics.

4.8. Prescriptive Analytics

The top level of analytics in Figure 1, prescriptive analytics for decision making, tends to be under-focused in statistics and data science programs. Recall the marketing 4P's in Table 1, if we can answer these questions *causally*, we can use the results to *optimize*, as listed in the last column of the table, linking predictive and prescriptive analytics through causal inference. Optimization is covered in Operations Research / Industrial Engineering with a collection of techniques including linear programming, dynamic programming, and stochastic programming. Reinforcement learning, in particular, has caught much attention lately given its success in autonomous vehicles and chess games

by integrating predictive and prescriptive analytics, and also has emerging business and healthcare applications (related to the brief mention of dynamic treatment regimes in He & Lin's first point). The key to prescriptive analytics is the *optimization mindset*—start with an objective function and constraints and solve the problem systematically. A common business application is customer relationship management which optimizes customer contacts—Figure 5 illustrates a marketing campaign where there are multiple channels, several messages/offers, and millions of customers. How to best assign the channel and message to customer? While the channel/message impact on customer can be estimated by uplift modeling, the presence of lots of combinations leads to a complex problem that requires efficient optimization algorithms, see Lo and Pachamanova (2015).

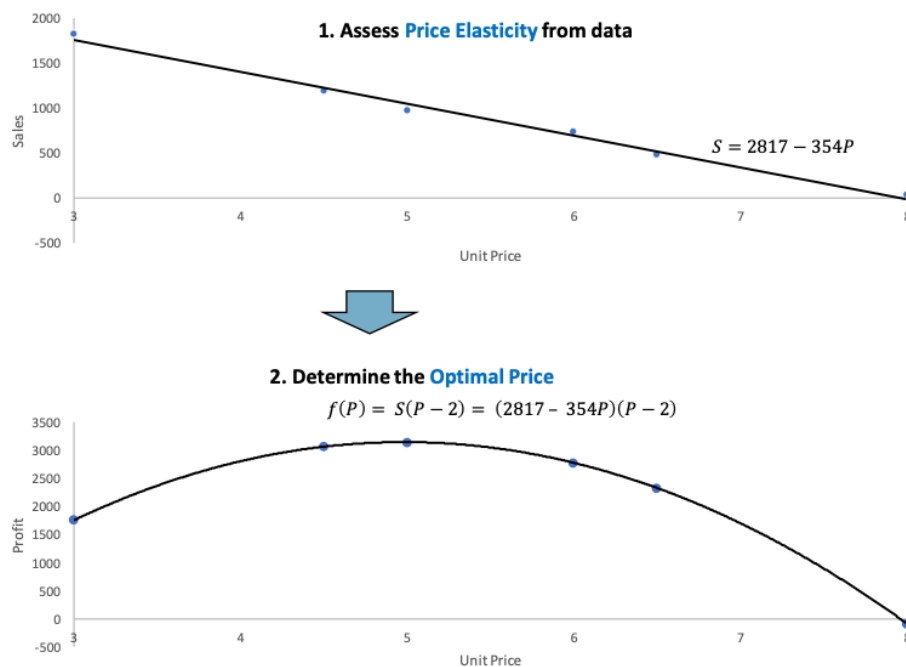**Figure 5. Optimization for Customer Relationship Management**

1. Assess Price Elasticity from data

$$S = 2817 - 354P$$

2. Determine the Optimal Price

$$f(P) = S(P - 2) = (2817 - 354P)(P - 2)$$

Figure 6. Price Determination

4.9. Social Sciences

Social Sciences is a collection of fields which data scientists can integrate to enhance analytics and insights.

- **Microeconomics:** This field studies individual and business decisions. As an illustration, suppose you sell ice cream and want to determine the best price to maximize profit. You gather historical data with various price points and corresponding sales, and fit a line to the data (Figure 6), you see a downward relationship (demand curve) between sales and price. The slope of this curve is the price elasticity[1]. Multiplying the estimated sales by unit profit is the total profit, where the unit profit is price minus unit cost (say, $2). We can see that the best price is $5. This illustration connects microeconomics, marketing, statistics, and optimization.

- **Macroeconomics:** If you are tasked with forecasting outcomes such as sales, revenue, or risk, a causal diagram of how macroeconomic variables (e.g., GDP, unemployment rate, market indices) may drive the outcomes based on economic/business knowledge would be helpful. Although future macroeconomic conditions are unknown, one can integrate forecasting and simulation to simulate future conditions in order to support outcome forecasting, see Oxelheim & Wihlborg (2008) and Leamer (2010).

- **Behavioral Economics:** By introducing Psychology to Economics, we have Behavioral Economics, leading to a couple of Nobel prizes, including the Nudge Theory summarized in Thaler & Sunstein (2009). For example, if you want your students to attend a webinar and you just provide the link, they usually will not opt-out. If they need to register first (opt-in), only a small portion may follow. Therefore, opt-out is more successful than opt-in. Another example is choice architecture—if you sell something with a few choices versus lots of choices, typically too many choices turns people off. Some of these phenomena are well studied and others are still open questions. We can borrow what is known and unknown and put them through a randomized experiment and analyze them to positively impact human behavior.

Figure 7. Common Data Science Applications

4.10. Application Domain Knowledge

These are some common application areas, with extensive domain knowledge in each (see Figure 7 for examples). Certain analytics programs have a whole course dedicated to applications, while others offer them as elective courses. Statistics and data science degrees may not include these subjects and thus students/practitioners may have to acquire themselves. For instance, if you are applying data science in marketing & sales, you would be working with professionals in that area and it would be more efficient to learn what they studied such as customer relationship management and marketing mix. Similarly, in risk management, data scientists should have some knowledge of market, credit, and/or operational risk, where each has its own set of knowledge. Once you have enough familiarity with a domain, you can effectively identify opportunities and apply appropriate analytics.

5. Summary

Data Science touches a wide range of academic and applied disciplines. Statisticians and data scientists can diverse their knowledge by acquiring modern day techniques such as NLP, Deep Learning, and other computational approaches as well as application-oriented areas such as Business and Social Sciences. This article suggested 10 topics from industry perspectives that can be considered for expanding the scope of educational programs and knowledge of professionals:

1. Causal Inference
2. Heterogeneous Treatment Effect / Uplift Modeling
3. Data Science Ethics

4. Deep Learning and Unstructured Data
5. Computational Tools and Technology
6. Analytic Consulting, Communication and Soft Skills
7. Descriptive Analytics
8. Prescriptive Analytics
9. Social Sciences
10. Application Domain Knowledge

Some of these are also applied research opportunities. To bridge the gap between *supply* of research & education and industry *demand*, extensive academic and industry *collaboration* is encouraged. Early exposure to real-world applications by students and interns can stimulate interests, expand skills, and enhance career opportunities.

Disclosure Statement

The author has nothing to disclose.

References

American Statistical Association (2018). "Ethical Guidelines for Statistical Practice." https://www.amstat.org/asa/files/pdfs/EthicalGuidelines.pdf

Boddington, P. (2017) *Towards a Code of Ethics for Artificial Intelligence*. Springer.

Bojinov, I., Chen, I., & Liu, M. (2020). "The Importance of Being Causal," *Harvard Data Science Review*, 2(3). https://doi.org/10.1162/99608f92.3b87b6b0

Brillinger, D.R. (2014). ". . . how wonderful the field of statistics is. . ." *Past, Present, and Future of Statistical Science*, edited by Xihong Lin, Christian Genest, David L. Banks, Geert Moleberghs, David W. Scott, & Jane-Ling Wang. CRC Press.

Dykes, B. (2020). *Effective Data Storytelling*. Wiley: Hoboken, NJ.

Leslie, D. (2019). "Understanding Artificial Intelligence Ethics and Safety." The Alan Turing Institute. https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf

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

Hornik, K. (1991). "Approximating Capabilities of Multilayer Feedforward Network," *Neural Networks*, 4(2), 251-257. https://web.njit.edu/~usman/courses/cs675_spring20/hornik-nn-1991.pdf

Imbens, G. W. & Rubin, D.B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press: New York, NY.

Independent High-Level Expert group on Artificial Intelligence, set up by the European Commission (8 April 2019) "Ethics Guidelines for Trustworthy AI." https://ai.bsa.org/wp-content/uploads/2019/09/AIHLEG_EthicsGuidelinesforTrustworthyAI-ENpdf.pdf

Kelleher, A. & Kelleher, A. (2019). *Machine Learning in Production: Developing and Optimizing Data Science Workflows and Applications.* Addison-Wesley.

Knaflic, C. N. (2020). *Storytelling with Data: Let's Practice!* Wiley: Hoboken, NJ.

Kowalski, R. (2011). *Computational Logic and Human Thinking: How to be Artificially Intelligent.* Cambridge University Press: Cambridge, UK.

Kruhse-Lehtonen, U. & Hofman, D. (2020). "How to Define and Execute Your Data and AI Strategy." *Harvard Data Science Review*, 2(3). https://doi.org/10.1162/99608f92.a010feeb

Leamer, E. E. (2010). *Macroeconomic Patterns and Stories: A Guide for MBAs,* Springer.

Lo, V.S.Y. (2002). "The True Lift Model – A Novel Data Mining Approach to Response Modeling in Database Marketing." *SIGKDD Explorations* 4(2),78-86. https://www.researchgate.net/profile/Victor_Lo3/publication/220520042_The_True_Lift_Model_-_A_Novel_Data_Mining_Approach_to_Response_Modeling_in_Database_Marketing/links/00463530681bd8f8de000000/The-True-Lift-Model-A-Novel-Data-Mining-Approach-to-Response-Modeling-in-Database-Marketing.pdf

Lo, V.S.Y. (2008). "New Opportunities in Marketing Data Mining." *Encyclopedia of Data Warehousing and Mining*, 2nd ed., edited by John Wang. Idea Group Publishing. https://www.igi-global.com/chapter/new-opportunities-marketing-data-mining/11006

Lo, V.S.Y. (2019). "Searching for the Perfect Unicorn." *Analytics Magazine.* https://pubsonline.informs.org/do/10.1287/LYTX.2019.04.02/full/

Lo, V.S.Y. & Pachamanova, D. (2015). "A Practical Approach to Treatment Optimization While Accounting for Estimation Risk." *Journal of Marketing Analytics*, 3(2), 79-95. https://link.springer.com/article/10.1057/jma.2015.5

Mendes, R & and Vilela, J.P. (2017). "Privacy-Preserving Data Mining: Methods, Metrics, and Applications." *Digital Object Identifier*, 5, 10562 – 10582. https://www.researchgate.net/publication/317630234_Privacy-Preserving_Data_Mining_Methods_Metrics_and_Applications

Meng, X-L. (2009). "Desired and Feared – What do we do now and over the next 50 years?" *The American Statistician*, 63(3), 202-210. https://www.tandfonline.com/doi/abs/10.1198/tast.2009.09045

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.

Oxelheim, L. & and Wihlborg, C. (2008). *Corporate Decision-Making with Macroeconomic Uncertainty*. Oxford University Press.

Pearl, J. (2000). *Causality*. Cambridge University Press.

Pearl, J. & MacKenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect.* Basic Books.

Radcliffe, N.J. & Surry, P. (1999). "Differential response analysis: modeling true response by isolating the effect of a single action." *Proceedings of Credit Scoring and Credit Control VI*. Credit Research Centre, University of Edinburgh Management School.

Rao, D. (2019). *Keras to Kubernetes: The Journey of a Machine Learning Model to Production*. Wiley: Indianapolis, IN.

Reinsel, D., Gantz, J., & Rydning, J. (2020). "The Digitization of the World: From Edge to Core." *IDC White Paper*. https://blog.seagate.com/business/enormous-growth-in-data-is-coming-how-to-prepare-for-it-and-prosper-from-it/

Rzepakowski, P. & Jaroszewicz, S. (2012). "Decision Trees for Uplift Modeling with Single and Multiple Treatments." *Knowledge Information Systems,* 32, 303-327. https://link.springer.com/content/pdf/10.1007/s10115-011-0434-0.pdf

Rubin, D.B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press.

Rubin, D.B. & Waterman, R.P. (2006). "Estimating the Causal Effects of Marketing Interventions Using Propensity Score Methodology." *Statistical Science*, 206-222. https://projecteuclid.org/download/pdfview_1/euclid.ss/1154979822

Rudin, C. (2019). "Stop Explaining Black Box Machine Learning for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence*, 1, 206-215. https://www.nature.com/articles/s42256-019-0048-x

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Sandler, R. & Basl, J. (2019). "Building Data and AI Ethics Committees." Northeastern University Ethics Institute and Accenture. https://cssh.northeastern.edu/informationethics/wp-content/uploads/sites/51/2019/08/811330-AI-Data-Ethics-Committee-Report_V10.0.pdf

Thaler, R. H. & Sunstein, C.R. (2009). *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Penguin Group.

The Institute and Faculty of Actuaries (IFoA) & the Royal Statistical Society (RSS) (2019) "A Guide for Ethical Data Science." https://www.rss.org.uk/Images/PDF/influencing-change/2019/A-Guide-for-Ethical-Data-Science-Final-Oct-2019.pdf

Unwin, A. (2020). "Why is Data Visualization Important? What is Important in Data Visualization?" *Harvard Data Science Review*, 2(1). https://doi.org/10.1162/99608f92.8ae4d525

Wing, J. M. (2019). "The Data Life Cycle," *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.e26845b4

Wing, J. M. (2020). "Ten Research Challenge Areas in Data Science," *Harvard Data Science Review*, 2(3). https://doi.org/10.1162/99608f92.c6577b1f

Yong, F. H. (2015), "Quantitative Methods for Stratified Medicine." *PhD Dissertation*, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University. https://dash.harvard.edu/handle/1/17463130

Yousra, A., Salleh, M., & Razzaque, M.A. (2015). "A Comprehensive Review on Privacy Preserving Data Mining." *SpringerPlus* 4:694. https://link.springer.com/article/10.1186/s40064-015-1481-x

Zhao, Y. (2017). "Uplift Modeling with Multiple Treatments." *PhD Dissertation*, Department of Electrical Engineering and Computer Science, MIT. https://dspace.mit.edu/handle/1721.1/113979

Zhang, W., Li, J., & Liu, L. (2020). "A Unified Survey on Treatment Effect Heterogeneity and Uplift Modeling." https://arxiv.org/pdf/2007.12769.pdf

indicated with respect to particular material included in the article. The article should be attributed to the author identified above.

Footnotes

1. For illustration, we assume this is a reasonable causal assessment. Proper causal inference is recommended in practice. ↩