TOWARD HOLISTIC DATA SCIENCE EDUCATION

RICHARD DE VEAUX
Williams College
rdeveaux@williams.edu

ROGER HOERL
Union College
roger.hoerl@gmail.com

RONALD SNEE
Snee Associates, LLC
Ron@sneeassociates.com

PAUL VELLEMAN
Cornell University
pfv2@cornell.edu

ABSTRACT

Holistic data science education places data science in the context of real-world applications, emphasizing the purpose for which data were collected, the pedigree of the data, the meaning inherent in the data, the deploying of sustainable solutions, and the communication of key findings for addressing the original problem. As such it spends less emphasis on coding, computing, and high-end black-box algorithms. We argue that data science education must move toward a holistic curriculum, and we provide examples and reasons for this emphasis.

Keywords: *Statistics education research; Data science; Data provenance; Human-machine interaction; Data analysis ethics; Problem solving*

1. INTRODUCTION

Data Science has been a STEM phenomenon. The data science community site at www.datascience.community/colleges lists 632 college and university programs offering degrees or certificates in Data Science as of the middle of 2022—and the list continues to grow. The U.S. Bureau of Labor Statistics predicts that Data Science/Analytics will create 11.5 million jobs by 2026.[i]

Despite all those graduates and jobs, however, the field itself seems unsettled. A *Kaggle* survey showed that most people working in the field spend 1 to 2 hours a week looking for a new job.[ii] And, although it has had successes, data science has also had a number of well-publicized failures. For example, the *Google* flu trends model, which at first seemed to be able to predict flu outbreaks faster than the Centers for Disease Control and Prevention, suffered what Wired called an "epic failure" in subsequent years.[iii] Another example is the model that won the widely publicized *Netflix* recommender systems competition. It was never actually implemented, reportedly due to "engineering costs."[iv] A longer list of disappointments resulting from naïve approaches to data science can be found in De Veaux et al. (2016).

Hutson (2018) pointed out the angst shared by many in this field:

> Ali Rahimi, a researcher in artificial intelligence (AI) at Google in San Francisco, California, took a swipe at his field last December—and received a 40-second ovation for it. Speaking at an AI conference, Rahimi charged that machine learning algorithms, in which computers learn through trial and error, have become a form of "alchemy". Researchers, he said, do not know why some algorithms work and others don't, nor do they have rigorous criteria for choosing one AI architecture over another. … As Rahimi puts it, "I'm trying to draw a distinction between a machine learning system that's a black box and an entire field that's become a black box."

Statistics Education Research Journal, 21(2), Article 2. https://doi.org/10.52041/serj.v21i2.40

This anxiety is not coming from outside the field, but from within it. Moreover, the financial gains achieved to date from data science appear to have been exaggerated. Poletti from *Market Watch* reported in April 2018 that "IBM earnings show AI is not paying off yet."[v] The MITSloan Management Review reported, "The reality is that many companies still struggle to figure out how to use analytics to take advantage of their data. The frustration of managers grappling with ever-increasing amounts of data and sophisticated analytics is often more the rule than the exception."[vi]

There are also growing concerns over ethical issues. *Amazon*'s use of a model to sort through job applications turned out to be biased against women.[vii] *Facebook*'s troubles were initially exposed by the Cambridge Analytica scandal,[viii] and they continue to deal with many ethical issues.[ix] O'Neil, in her book, *Weapons of Math Destruction* (2017), points out the dangers and injustices that can arise by using complex models to determine credit worthiness, the price of insurance policies, whether someone should receive parole, or even which crimes police should investigate. Public concern over the ethical use of data is growing. According to Pew Research, "roughly half of Americans do not trust the federal government or social media sites to protect their data."

This discussion is not to suggest that data science is unimportant or has not achieved dramatic success; on the contrary, we view it as one of the most critical technology developments in recent times. But, beneath the surface of success, the discipline seems to be floundering in search of direction. Our hope is that more holistic data science education can help the field find that direction and realize its potential. In this article, we suggest ways that data science education should evolve to serve that goal.

2. TOWARDS A HOLISTIC APPROACH

Many of the current limitations of data science arise from the simple fact that it is not truly a science (Donoho, 2017). Or, perhaps we should say, not yet a science. *Wikipedia* defines science as: "a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe" ()[x]. Although there is certainly knowledge (intellectual content) in data science, much of it is borrowed from statistics and computer science. There is, however, no consensus about what constitutes the essential elements of data science outside of the union of statistics and computer science. A cursory survey of the syllabi of some of the many data science majors and degree programs shows a diversity of focus, a point made previously by Donoho (2017). There is, of course, typically a core of topics that include data acquisition, data management, data processing, and model-fitting by automated algorithms such as machine learning.

Two major professional societies, the Association for Computing Machinery (ACM) and the American Statistical Association (ASA) have both endorsed curriculum guidelines that contain this core of topics (Danyluk et al., 2019; De Veaux et al., 2017, respectively). However, many data science curricula fail to give sufficient attention to the original motivation that gave rise to data science—namely attention to:

- the purpose for which the data were collected,
- the pedigree and quality of the data,
- the meaning inherent in the data,
- model validation – beyond train/test splits on one data set,
- the deployment of sustainable solutions producing tangible results, and
- the communication of key findings from addressing the original problem.

We propose a holistic approach to data science education that is strongly informed by the scientific method. Our view of holistic data science shares much with Donoho's (2017) discussion of greater data science (GDS). It can also be viewed as the application of the emerging discipline of Statistical Engineering (see www.ISEA-change.org) to Data Science.

Specifically, we define holistic data science as an interdisciplinary approach to data-based problem solving and knowledge discovery, which applies the scientific method to each phase of the problem-solving life cycle, from problem definition through verification of solution sustainability. To teach this, an holistic data science curriculum should address the following areas:

1. ***Motivation, problem definition and context.*** A data analysis must be motivated by a goal—even if that goal is just finding a good prediction versus understanding cause and effect

relationships (Shmueli 2010). And it must be located within a clear context where it is expected to apply and inform. A good data science application solves a clearly elucidated problem or answers a specific question. This is the hard work that must be done before applying the automated tools. And it is some of the most difficult material for students to learn and internalize.

2. ***Data provenance and pedigree.*** The most sophisticated analysis is worthless if it is based on weak data (Kenett & Redman, 2019). The context of the problem to be addressed informs judgements about the required relevance and quality of the data. Too often we have seen data analyses proceed blindly; applied to data that are inappropriate or riddled with errors and gaps. By contrast, students are often given tidy data ready to plug into sophisticated algorithms, thus missing a key step in working with real data. Students must learn to document their data sources and pedigree. And—even more important—to be skeptical about the soundness of their data. We argue that data should be considered "guilty until proven innocent".
3. ***Scientific inference.*** It is the rare (and weak) analysis that applies only to the data at hand. Students must learn to cautiously respect and intelligently use statistical inference and other confirmation methods such as cross-validation. They must be prepared to position their analyses in the larger context of scientific or business understanding. A result at variance with other related results requires a skeptical eye, greater care, and (likely) better data.
4. ***Human machine interaction and decisions***. Analytics must be a collaboration between human analysts and computer algorithms, with the algorithms serving as tools to be wielded by humans. It is the human analyst who can adapt to changing circumstances, perceive the limits of the model, understand the limitations of the data set, evaluate extraordinary and outlying values and correct, exclude, or account for them, and understand the possible unintended consequences of any model that optimizes one criterion.
5. ***Ethics***. Increasingly, the ethical consequences of data science analytics have been exposed—usually to the detriment of the analysis and analysts. Here again, we cannot rely on algorithms and must train our students to think and behave ethically and to apply those principles to their work. Students should be taught to ask *why* an analysis is being pursued and to consider the ethical consequences of the answer.
6. ***Problem Solving.*** Yes, we need to teach coding, machine learning algorithms, and big data issues. But they should not be the main focus of a Data Science syllabus any more than calculus should be the main focus of a physics curriculum. They are tools, and students should become facile users of them—but first they must learn why and how to use them. Solving the problem at hand, by delivering sustainable solutions that produce tangible impact, is the ultimate measure of success. The *Netflix* model, discussed above, won a million-dollar competition, and although some elements of the solution were utilized by *Netflix*, it didn't solve *Netflix*' original problem, and the full model was never deployed. In fact, the entire business model for *Netflix* changed soon thereafter. Should it be considered a success or a failure?

3. ELABORATION OF MAIN POINTS

Many practicing data scientists are already aware of the issues presented above and use an holistic approach in their professional practice. Data science education, however, needs to focus more purposefully on the six areas noted in the previous section, to avoid curricula becoming just collections of techniques and coding practice.

3.1. MOTIVATION, PROBLEM DEFINITION, AND CONTEXT

At a recent data science conference one of the authors saw an analysis of New York City taxi data. The data set contains information for each ride, including pickup and drop off locations, time of day, fare, and duration of the trip. The graphics were impressive and the dataset was legitimately big: about 2GB for each month. But the purpose of the study (other than to show that the data could be displayed)

was never made clear, leaving one with the feeling that an opportunity to advise the troubled taxi industry had been missed. The analysis was interesting, but not particularly useful.

Most data science competitions focus on narrow performance measures, such as minimizing root mean squared error (RMSE) for continuous response variables or overall error rates for categorical responses. But optimality will ignore slightly "suboptimal" solutions that may more appropriately address the problem or avoid ethical pitfalls of an "optimal" model.

An algorithm that plays chess or *Go* need only optimize its performance playing in those well-fenced playgrounds. But if data science is to fulfill its promise for science and business, its playground must be the open and dynamic field of the world at large. Algorithms typically cannot incorporate the meaning of the data, so they cannot benefit from any understanding brought to the work by knowledgeable human participants, and cannot refer to other related data that has not been offered. The failure of *Google*'s flu prediction algorithm may have been due in part to blind overfitting, including such seasonal search terms as "high school basketball."

Algorithms inherently treat the data they are given as static. When underlying conditions change, a human can often adapt, but computer-based methods are likely to be confounded. In the case of the Flu Trends Model, *Google* introduced a suggested search feature and several health-based features to help people find relevant information. But those changes encouraged new search strategies, changing some of the behaviors that the original flu prediction algorithm had implicitly depended on. Dileep George, a computer scientist at Vicarious (as quoted by Clive Thomson (https://www.magzter.com/stories/Science/WIRED/The-Miseducation-Of-Artificial-Intelligence) noted:

> Humans engage in reasoning, making logical inferences about the world around us; we have a store of common-sense knowledge that helps us figure out new situations. … The neural net, on the other hand, … (could only) follow the pattern. When the pattern changed, it was helpless.[xi]

Data scientists have tried to overcome this problem with automated updating of models, or so-called "continuous learning". However, even continuous learning was not able to save the Google Flu Trend Model, nor other practical problems involving dynamics.

Despite these well-publicized experiences, respondents to a KDnuggets poll in 2015[xii] thought that most expert-level predictive analytics/Data Science tasks will be automated by 2025. Evidently, they didn't think human participants were essential. So, evidently their Data Science education missed some of these points. We argue below that problem solving must be a fundamental part of any Data Science Curriculum, as part of every course rather than a stand-alone offering. And part of this must be attention to the ethical consequences of models. See De Veaux et al. (2016) for details of appropriate projects.

3.2. DATA PROVENANCE AND PEDIGREE

No algorithm can overcome bad data and bad data are everywhere (see e.g., De Veaux & Hand, 2005). Needed is critical thinking about data that begins with questioning the pedigree of the data (Hoerl & Snee, 2019). Knowing the data pedigree provides a holistic view of the data enabling one to assess and identify:

- Data Quality – Are the data fit for use in the application being studied?
- Data Integrity – Can the data be trusted?
- Sources of variation in the data help one to decide how to analyze the data.
- Chain of custody – What persons and organizations have had access to the data from first measurement to current values?
- Metadata – Information about the data.

To produce useful models, it is essential that we know how the values were measured and recorded, and the chain of custody that led to the version of the data we have. Students must be able to recognize the qualitative difference between observational data and data from randomized experiments. They must also understand why sample size does not by itself justify trust in a data set. Understanding whether data are correctly measured and collected or are representative of the population under study requires human judgment. The chain of custody documents whether others have made changes or

adjustments to the data. Do we have a copy of the entire original dataset? If the data are about people, were any groups not included?

Of course, when analyzing observational data, we have less control over the data pedigree. But it is essential that the analyst understands the pedigree and its consequences for any analysis. Concern with the data's pedigree should start with the original data collection and calibration of the measuring instruments. This includes attending to the design of studies and the selection of appropriate measures to ensure (or at least encourage confidence) that the data at hand are representative of corresponding situations in the world at large.

Recently, Braynard, a former Trump campaign strategist enlisted the help of Prof. Steven Miller of Williams College to demonstrate fraud in the 2020 Presidential Election in Pennsylvania[xiii,xiv]. Taking Braynard's data at face value and assuming that it was a representative sample from the population of votes, Miller extrapolated from a simple binomial confidence interval formula to an entire state and concluded that more than 50,000 fraudulent Republican absentee ballots had probably been requested. This fact was re-tweeted by Trump himself to over 70,000,000 followers in an attempt to bolster public support for his failed attempt to have the election results overturned. After more carefully considering the pedigree of the data and the motivations of those handling the data, Miller retracted his analysis and submitted an apology. Automated methods have no ability to take such a second look.

Errors or deletions can be introduced in the handling of the data, so the "chain of custody" can help the human analyst account for (and possibly, correct) anomalous values. Reasons why the data may not be homogeneous may arise from understanding its context and the nature of the problems to be solved. This can be a serious issue even with massive data sets, because there may be a substantial portion of inaccurate, faulty, missing, or misunderstood data. For example, we have seen data sets in which one of the variables was 40% zeros. It turned out that these were actually missing cases, not true zeros. Some software codes missing cases with extreme values, such as 999, expecting that they will be noticed and dealt with. In very large datasets, they may go unnoticed but still influence any model or conclusion. We should add that we do not recommend automated "data cleaning" algorithms. These are not algorithms selected for the purpose of filtering out important data, but algorithims which, as an unintended consequence, can filter out the most important data, which stands out because of signal rather than noise or error.

3.3. THE SCIENTIFIC METHOD AND INFERENCE

There is a large literature describing the scientific method stretching back to Sir Francis Bacon in his *Novum Organum*. There he lays out a plan for understanding the world that calls for a sequence of steps. Guided by our best understanding, we seek data. Then guided by the data, we adjust our understanding. This process continues with no real end. It has been understood for several decades that statistical and analytical methods work best when integrated with the sequential approach of the scientific method. For example, see Figure 1, which is based on a similar figure in Box et al. (1978).

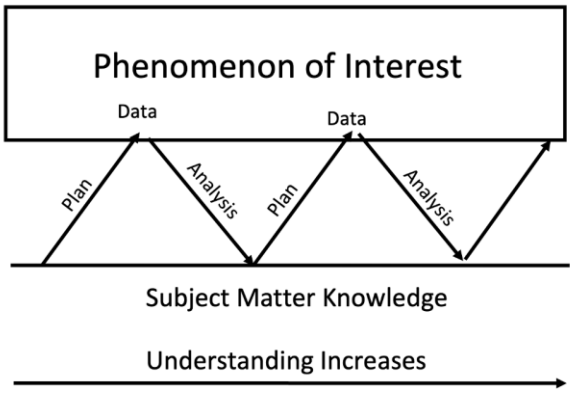

Figure 1. The sequential nature of statistical studies

At no point do scientists think that their understanding is perfect, but intermediate models are helpful. The challenge of scientific thinking is that it calls for skepticism—something computer algorithms are not good at. You must think about things that are not in front of you and imagine ways in which things might have gone wrong. In effect, you should go looking for trouble.

Data alone does not create science. In "*The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*," Anderson (2008) claims that data can "speak for themselves", without the need for science or even models.

> There is now a better way. Petabytes allow us to say: Correlation is enough. We can stop looking for models. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot …. There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?[xv]

This naïve reasoning was the basis of the *Google* flu model failure. It is unfortunately prevalent in too much current data analysis and accepted in too much data science education. More important to our thesis here, students must be taught not only that this approach is likely to fail, but why it rarely works.

Instead, what we learn from our initial analysis should inform our subsequent data collection and analysis—and we should expect there to be a subsequent analysis. This is true even for "classical" designed experiments, but it is especially true of the exploratory analyses that make up much of Data Science.

The principle that analysis often involves a sequence of models is rooted in the scientific method. It guides the CRISP-DM methodology developed by the European Union in early days of data mining (Shearer, 2000). This methodology has much wisdom about integration of the scientific method into analysis of large data sets. The PPDAC model (Problem, Plan, Data, Analysis, Conclusions), also a sequential approach, goes back to MacKay and Oldford (1994).

Of course, there is also a need for statistical inference, but with appropriate adjustments. Very large datasets can render an estimated standard error silly or irrelevant due to the sample size. Yet the assumptions underlying those methods are neither silly nor irrelevant. There is still a need to understand the assumptions of any model before making predictions or inferences outside the current data set. Cross validation using hold-out data is certainly useful, but still focuses on the current data set collected under specific conditions at a specific point in time. What gives us confidence that the model might predict new data, collected under different circumstances (Sambasivan et al., 2021)?

3.4. HUMAN-MACHINE INTERACTION

We have noted that even the most powerful data science algorithms should be seen as tools guided by human data analysts. O'Neil (2016) stated, "Big data processes codify the past. They do not invent the future. Doing that requires moral imagination, and that's something only humans can provide" p. 204). Perhaps they could invent the future if properly guided by human imagination and ethics.

Some Data Science curricula and texts that we have seen pay insufficient attention to exploratory data analysis (EDA) as a first step in a data analysis. John Tukey, who coined the EDA term, suggested some methods in his book *Exploratory Data Analysis* (Tukey 1977), and in many related articles. But many of his methods are pencil-and-paper tools and not suited to big data. The insight often missing in data science education is that Tukey also elucidated a philosophy of how to address data—and this philosophy is not merely relevant, but essential to sound analyses with common data science methods.

The philosophy of EDA is based on five fundamental principles: display, re-expression, residuals, resistance, and iteration (Velleman & Hoaglin, 2022). In particular: "…the principal contribution of EDA is philosophical. EDA advocates exploring data for patterns and relationships without requiring prior hypotheses"

Of course, this is often what a data scientist must do, as much data comes without clear hypotheses and many data science problems have loosely stated goals. Unsupervised analyses are, almost by definition, free of prior hypotheses. A key point is that each of these five principles involves an iterative approach, requiring interaction among the data, the computer, and the analyst.

One of the principles of exploratory data analysis is the importance of re-expressing variables to achieve simple structure. In a review of a competition to find a model to predict baseball player's

salaries from performance statistics, Velleman and Hoaglin (1995) found that most participants had built complex models using unsupervised algorithms. Although re-expressing salaries by logarithms led to a simple linear model with a single predictor (runs scored) that fit better and could be understood, only two participants in the competition had found that model.

Re-expression can clarify the presence of outliers. A skewed distribution may appear to have outlying values at the tip of its longer tail, and these may be automatically deleted by cleaning algorithms. Yet a re-expression that makes the distribution more symmetric may pull in those values, so they are seen as appropriately part of the main body of the data.

3.5. ETHICS

The fifth principle reminds us that ethical considerations are an essential part of every step of Data Science. It is especially important that students internalize this principle because decisions that may have ethical consequences are often hidden in the details of an analysis and will not be visible to most consumers of the results or understood by those not well-versed in the methods. Were variables used that would be illegal for a human to use, such as in sentencing? Were the variables standardized? Should they have been? If not, did a larger subpopulation dominate the conclusions? Was some identifiable group excluded from the data or under-represented? If there was a focus on a subgroup, was it part of an ethically questionable hunt for statistical significance ("p-hacking") when the full sample didn't show the desired relationship?

In *Weapons of Math Destruction*, O'Neil (points out the many ways that models can and are being used in unethical ways. A recidivism algorithm by the Northpointe company, used widely in the US for helping make parole decisions, is based on 137 questions from both defendant responses and criminal records (Larson et al., 2016). The algorithm produces a risk score (called COMPAS) from 1 to 10 (10 is high risk) to help guide judges when assessing whether an inmate should be paroled. The algorithm is complex and proprietary. Angwin et al. (2016) from the ProPublica group publicly criticized the algorithm for being racially biased.[xvi] One striking example was the contrast of Brisha Borden, a young black woman serving a sentence for some minor juvenile crimes who was given an 8 out of 10, while Vernon Prater, a middle-aged white man, serving time for a series of armed robberies, was given only a 3.

While the question of whether an algorithm can be "biased" can be debated, it is clear that the conclusions of a model can be biased if the data on which it is based are biased. Google's photo-categorization software labeling some Black people as gorillas is a prime example. In the Northpointe case, it turned out that data quality was a significant factor. Rudin and co-authors (2020) reproduced much of the Northpointe scores on a subset of the data but using much simpler models. As they concluded:

> In the past, there have been documented cases where individuals have received incorrect COMPAS scores based on incorrect criminal history data and no mechanism to correct it after a decision was made based on that incorrect score. We do not know whether this happens often enough to influence the scatter plot in a visible way. However, this type of miscalculation is one of the biggest dangers in the use of proprietary models.

Another example involves policing algorithms that tend to send police patrols to poor neighborhoods, because there are high crime rates there. The natural result from more patrols is more arrests, not only for violent crimes, but for things like underage drinking and marijuana. The additional arrests cause the algorithm to send even more patrols to those same areas, in a "vicious cycle" of more patrols, more arrest, more incarceration and criminal records. The criminal records make it harder for people to obtain legitimate jobs, leading to more crime, and the cycle continues. Of course, underage drinking and marijuana use, not to mention sexual assaults, occur at expensive private universities as well, but the algorithms are much less likely to send the police to exclusive private universities.

Many data scientists may think models are objective and unbiased, but O'Neil (2017, p. 21) defines models as "opinions embedded in mathematics." Mathematics gives the model the appearance of objectivity, but someone built the model and decided which data to use, which variables to include, which model form to use, and so on. A model is in fact an opinion, reflecting both the biases of the

modeler and of the data. Students of data science need to be sensitive to these ethical issues and taught to avoid bias and discrimination in their models.

3.6. PROBLEM SOLVING

Students should learn early in their education that, fundamentally, it's NOT about the tools. Data science tools, no matter how powerful, are a "how," not the "what." The ultimate objective is to determine and deploy sustainable solutions to challenging problems, not just to know and apply the tools. Unfortunately, education that is tools-oriented produces students who may understand and know how to use tools, but do not know how to solve problems with them. Students must learn about the problem-solving process itself.

It is essential that students understand how data science tools can fit together into an overall approach to solving real problems. In this, we can learn from the statistics community. The ASA guidelines for undergraduate programs in statistical science state:[xvii]

> Undergraduates need practice using all steps of the scientific method to tackle real research questions. All too often, undergraduate statistics majors are handed a "canned" dataset and told to analyze it using the methods currently being studied. This approach may leave them unable to solve more complex problems out of context, especially those involving large, unstructured data …. Students need practice developing a unified approach to statistical analysis and integrating multiple methods in an iterative manner.

The ASA guidelines are not alone in calling for a problem-solving approach. Franklin and Bargagliotti (2020), in discussing the updated guidelines for precollege statistics and data science education (GAISE II), note:

> It is critical that statisticians, or anyone who uses data, be more than just data crunchers. They should be data problem solvers who interrogate the data and utilize questioning throughout the statistical problem-solving process to make decisions with confidence... (p. 8)

An engineering mindset, focused on solving the problem at hand, can augment students' ability to understand and apply tools (Snee & Hoerl, 2020). Of course, we need to teach coding, machine learning algorithms, and big data processing. But these should not be the main focus of a data science syllabus any more than calculus should be the main focus of a physics curriculum. They are tools, and students should become facile users of them—but first they must learn *why* and *how* to use them.

Similarly, students should understand that prediction accuracy from train/test splits, cross validation results, and so on, are intermediate model evaluation results, and not ultimate measures. It may be that a "suboptimal" solution is the one that best solves the real problem. Solving real problems sustainably, regardless of the complexity of the tools used, is the ultimate objective. Otherwise, the old saying "the operation was a success, but the patient died" becomes all too real.

How can problem solving be taught in an academic setting? Obviously, only by solving problems! The best way to accomplish this in academia, in our opinion, is through student projects. In particular, these should be a sequence of projects in which students learn to use multiple tools in a logical sequence to reach actionable conclusions. Such projects are very different from "data fests," in which students focus on in-depth analysis of a static data set, with a narrow criterion, such as minimizing out-of-sample root mean square error.

More realistic projects help immerse students in the problem-solving process, and help them gain understanding beyond optimizing a numeric objective function. We have frequently heard from students' comments such as: "I thought I understood this stuff but I found out that there were gaps in my knowledge and understanding when I worked to solve a real problem complete with real data."

In our experience, students can work in teams of 2–4 people, which teaches teamwork and reduces the number of problems that have to be identified for the class to use as projects. Team projects also prepare students for careers in which the vast majority of problems will be addressed through teams. In most cases students can identify worthwhile problems that are of interest to them.

Other options for developing problem-solving skills include:

- Attending presentations by others who have worked on real problems, including external speakers.

- Assigning case studies for homework and requiring students to critique the problem-solving approach.
- Pairing of students studying data science with students in other disciplines, to work on data-oriented problems in those disciplines.
- Having informal lunch sessions to discuss the challenge of deploying sustainable solutions in the real world.

3.7. PRACTICAL IMPLEMENTATION

While there is no "cookie cutter" approach to integration of these principles into specific curricula, we would like to offer sample approaches that we have taken in our own courses. Some of the specific modifications that we have utilized include (in haphazzard order):

- Augmenting a technical textbook with one more focused on data science ethics, such as O'Neil (2017).
- Incorporating ethical issues into classroom discussion.
- Incorporating discussion of big data failures, such as the *Google* flu trends model, in addition to the successes.
- Requiring course projects that utilize sequential analysis and are graded not solely on the basis of fit, but on what was learned, explaining how such a model might be used in the future, documentation of its limitations, and ideas for obtaining better data to enhance it in the future. Students should report on their analysis progress early enough in the term that they have time to respond to feedback. If class presentations are a part of the plan, feedback can come from other students—an excellent learning experience for them as well.
- Formally covering the scientific method in the classroom and discussing how machine learning and other methods fit into it. This includes beginning with a clear problem statement prior to data analysis or model building.
- Formally covering data pedigree in the classroom and requiring students to document the pedigree of the data used for their project, as well as other data used in the class or on homework. This usually requires more work than the students expect, especially for data sets from public data repositories or Kaggle.com.
- Formally covering sequential approaches to problem solving in the classroom and expecting students to utilize them (or others of their choosing) on their projects.

While we do not claim to have an "optimal" course structure, we have found these modifications to be very effecgtive at broadening students' understanding of data science, and what it can/can't do. Of course, to add to a syllabus, one must subtract. We understand that incorporation of these modifications will require reduction in time spent on other topics, such as coding or machine learning tools.

4. CONCLUSION

The tools of data science include powerful algorithms. These tools are best used with guidance informed by human understanding of the questions to be addressed. Figure 2 illustrates our view on how humans, data, and algorithms should interact in any data analysis, including statistics and data science. This also illustrates how the principles we have discussed can work together in a data science curriculum.

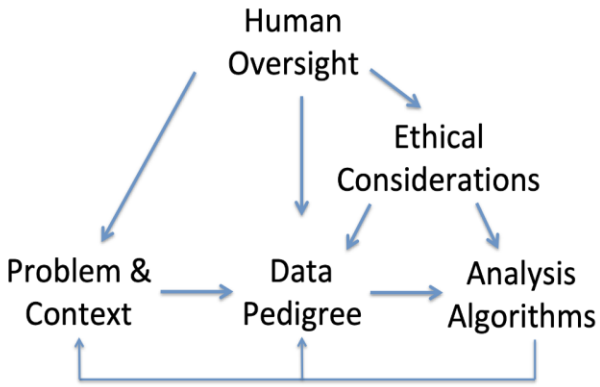

Figure 2. Relationship among people, data, ethics, and algorithms

It can be tempting for analysts to simply "let the algorithm run," turning over control to the computer. Students may prefer this because it relieves them of the requirement to think and make decisions and appears to relieve them of responsibility for those decisions. "The computer did it" is never an acceptable excuse for a model gone wrong. We must design our curricula to push back against this tendency and instead encourage a holistic view. Removing human oversight impairs adequate consideration of ethical issues.

We believe that a curriculum that places the techniques of data science within the context of scientific approaches to data, understanding of problem context, and thoughtful human-guided analyses, will enable data science to help shape a brighter future for everyone. A more holistic approach is needed that incorporates elements of the scientific method, takes a broad view of problem solving, incorporates subject-matter knowledge, considers the data pedigree, and evinces a concern for the ethical consequences of analyses.

Rosenberg (2017) wrote an online article for Wired.com entitled, "*Why AI is Still Waiting for its Ethics Transplant*".[xviii] He quotes Kate Crawford on the importance of understanding data pedigree:

> Data will always bear the marks of its history. That is human history, held in those data sets. So, if we're going to try to use that to train a system, to make recommendations or to make autonomous decisions, we need to be deeply aware of how that history has worked.

We could not agree more. We are confident that the changes needed can be made. To do so, requires a conscious move away from the glamor of media "hype," to the open and sometimes painful scrutiny of scientific criticism. Hutson's (2018) article,[xix] quoted previously, is one example of such scrutiny. The view will certainly be worth the journey.

REFERENCES

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978), *Statistics for experimenters*. John Wiley & Sons.

Danyluk, A., Leidig, P., Cassel, L., & Servin, C. (2019). ACM task force on data science education: Draft report and opportunity for feedback. *SIGCSE 19: The 50th ACM Technical Symposium on Computer Science Education*, Minneapolis, February 27–March 2 (pp. 496–497). https://doi.org/10.1145/3287324.3287522

De Veaux, R. D., & Hand, J. L. (2005). How to lie with bad data, *Statistical Science*, *20*(3), 231–238.

De Veaux, R. D., Hoerl, R. W., & Snee, R. D. (2016). Big data and the missing links. *Statistical Analysis and Data Mining*, *9*(6), 411–416.

De Veaux, R. D. et al. (2017). Curriculum guidelines for undergraduate programs in data science. *Annual Review of Statistics and Its Application, 4*, 15–30.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*(4), 745–766.

Franklin, C., & Bargagliotti, A. (2020). Introducing GAISE II: A guideline for precollege statistics and data science education. *Harvard Data Science Review, Issue 2.4.* https://doi.org/10.1162/99608f92.246107bb

Hoerl, R. W., & Snee, R. D. (2019, January). Show me the pedigree: Evaluating data quality includes analyzing its origin and history. *Quality Progress,* pp. 16–23.

Hutson, M. (2018, May 3). AI researchers allege that machine learning is alchemy. *Science.org*. https://www.science.org/content/article/ai-researchers-allege-machine-learning-alchemy

Kennet, R. S., & Redman, T. C. (2019). *The real work of data science.* Wiley and Sons.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016, May 23). *How we analyzed the COMPAS recidivism algorithm.* ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

MacKay, R. J., & Oldford, W. (1994). *Stat 231 course notes full 1994*. University of Waterloo.

O'Neil, C. (2017). *Weapons of math destruction.* Broadway Books.

Ransbotham S., Kiron D. & Prentice, P. K. (2016). Beyond the hype: The hard work behind analytics success: Why competitive advantage from analytics is declining and what to do about it. *MITSloan Management Review*. https://sloanreview.mit.edu/projects/the-hard-work-behind-data-analytics-strategy/

Rosenberg, S. (2017, November 1). *Why AI is still waiting for its ethics transplant.* Wired. https://www.wired.com/story/why-ai-is-still-waiting-for-its-ethics-transplant/

Rudin, C., Wang, C. & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, *Issue 2.1*. https://doi.org/10.1162/99608f92.6ed64b30

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. K., & Aroyo, L. M. (2021). Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI. *Proceedings of CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan, May 8–13. https://research.google/pubs/pub49953/

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Shearer C., (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, *5*, 13–22.

Snee, R. D., & Hoerl, R. W. (2020, July). It's not about the tools. *Quality Progress,* pp. 44–46

Tukey, J. W. (1962). The future of data analysis. In L. V. Jones (Ed.), *The collected works of John W. Tukey*. Vol. III (1986). Wadsworth.

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

Velleman, P. F., & Hoaglin, D. C. (1995). A critical look at some analyses of major league baseball salaries. *The American Statistician*, *49*(3), 277–285.

Velleman, P. F., & Hoaglin, D. C. (2022). Exploratory data analysis. In H. Cooper (Ed), *APA handbook of research methods in psychology: Vol 3. Research designs: Quantitative, qualitative, neuropsychological, and biological.* American Psychological Association.

[i] https://www.kdnuggets.com/2018/09/how-many-data-scientists-are-there.html

[ii] https://www.ft.com/content/49e81ebe-cbc3-11e7-8536-d321d0d897a3

[iii] https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

[iv] https://www.wired.com/2012/04/netflix-prize-costs/

[v] https://www.marketwatch.com/story/ibm-earnings-show-ai-is-not-paying-off-yet-2018-04-17

[vi] https://sloanreview.mit.edu/projects/the-hard-work-behind-data-analytics-strategy/

[vii] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[viii] https://www.businessinsider.com/facebook-87-million-cambridge-analytica-data-2018-4

[ix] http://fortune.com/2018/04/06/facebook-scandals-mark-zuckerberg/

[x] https://en.wikipedia.org/wiki/Science

[xi] "How to Teach Artificial Intelligence Some Common Sense," *Wired* 11.3.18, https://www.wired.com/story/how-to-teach-artificial-intelligence-common-sense

[xii] https://www.kdnuggets.com/2015/05/data-scientists-automated-2025.html

[xiii] https://magazine.amstat.org/blog/2021/02/01/miller-controversy-perspective/

[xiv] https://www.berkshireeagle.com/news/local/williams-prof-disavows-own-finding-of-mishandled-gop-ballots/article_9cfd4228-2e03-11eb-b2ac-bb9c8b2bfa7f.html

[xv] https://www.wired.com/2008/06/pb-theory/

[xvi] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[xvii] http://www.amstat.org/education/curriculumguidelines.cfm

[xviii] https://www.wired.com/story/why-ai-is-still-waiting-for-its-ethics-transplant/

[xix] https://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy