The Data Science Pipeline: Automation and Ethics

Scott Rubey

Computer Science, Portland State University, Portland, Oregon, USA, scrubey@pdx.edu

ABSTRACT

The data science pipeline has historically been labor-intensive and time consuming. It consists of myriad steps, including (but not limited to) data ingestion, cleaning, and modeling. Recent advancements in artificial intelligence and machine learning have eased this burden in many ways, paving a path to faster output and saved dollars. Advantages of automating the data science pipeline are numerous but can be fraught with ethical dilemmas due to algorithmic bias and the potential for lost jobs. This paper explores the processes and technologies being studied for the purpose of data science automation, in addition to the underlying ethical considerations that accompany them.

KEYWORDS

Automation, artificial intelligence, data science, predictive analytics, data mining, data wrangling, ethics

1 Introduction

To the outsider, the data science pipeline may appear relatively straightforward: in order to arrive at a given output, one must simply ingest and process a dataset. The naiveté of this misconception becomes apparent upon examination of the sub-steps required in completing each of these overarching tasks. Once data ingestion is complete, one must clean the data, transform the data (i.e. account for outliers and populate any missing fields), analyze and interpret the data, and finally prepare a model for the client/end user. Two problems make this process all the more arduous for the data scientist. First, the sheer size of the datasets being processed in today's world of big data can be enormous. Tens of millions of records with hundreds of millions of fields can represent a modest domain in some contexts. Secondly, few (if any) datasets are without error in some form. Missing fields, inconsistent formatting and erroneous values are a natural occurrence in most datasets, and such imperfections only grow with the size of the data.

Historically, data scientists were required to perform these tasks manually, a tedious and costly procedure that was prone to human error. Some studies have indicated that 50-80% of a data scientist's time – even today – is tied up in data wrangling, the process of cleaning and processing raw data for the purpose of analysis or modeling later on. Recent technological advancements in artificial intelligence and machine learning have lightened the manual workload substantially in some areas (though the considerable number of tools available to modern users can present obstacles in and of themselves). Automation of this nature provides benefits in several forms. First, less time is required to produce the final model/product, as less time is spent on manual cleaning and other processing of data. Secondly, automation of otherwise complicated tasks makes data science more accessible to those without special training. And third, the resulting model can be produced at a financial discount in the form of fewer human-hours devoted to the aforementioned tasks, benefitting the producer, the client, or both.

Studies have shown that AI automation has the potential to benefit the US economy into the trillions of dollars. [2] While this is a considerable figure by any metric, it does not reveal the hidden cost to individuals in the form of lost jobs, underserved communities, and even racial discrimination. Algorithmic bias has become a topic of concern in recent years and is an area of active research. Only through acknowledgement and understanding of these ethical concerns can the computer- and data-science communities begin forging a path toward technologies that benefit their practitioners and clients without rejecting other sectors of our society.

This paper describes the steps involved in the data science pipeline, the technologies being developed for the purpose of automating these steps, as well as the ethics involved in attempting to solve human problems with software algorithms. Section 2 will provide a description of the steps in the pipeline described above; Section 3 summarizes the automation technologies developed in recent years; and Section 4 digs deeper into the ethics and principles surrounding the use of artificial intelligence and machine learning and their impact on individuals and communities.

2 The Data Science Pipeline

Microsoft Azure's product documentation describes the data science lifecycle in the following terms: 1) Business understanding; 2) Data acquisition and understanding; 3) Modeling; 4) Deployment; and 5) Customer acceptance.[1] This paper focuses on the subprocesses involved in Step 2, Data acquisition and understanding, with a short discussion on Modeling in Section 3. These subprocesses are summarized by many in the data science community as "data wrangling." Koehler *et al* [3] describe data wrangling as "the process of preparing potentially large and complex datasets for further analysis or manual examination…" A routine Google search divides this process into the following steps, which are echoed by numerous sources: data extraction, structuring, cleaning, enriching, validating, and publishing. For purposes of this paper, data structuring, cleaning, and enriching will be approached under the umbrella of data "transformation." While data science companies might encourage/deploy proprietary workflows, these steps provide a foundation for understanding pre-modeling phases from a high level. A short description of each step in the data wrangling process follows.

2.1 Data Extraction

Data extraction, simply put, is the process of capturing data from one or more sources. Sources may include (but are not limited to) documents, webpages, images, PDF's, and spreadsheets/CSV files. At this stage, the data is often incomplete, inconsistently formatted, and structured in a way that must be modified prior to analysis. Extracted data may be stored in an intermediate repository for further processing.

2.2 Data Transformation

Data transformation, as mentioned above, is a term that broadly encompasses the structuring, cleaning and enriching of raw data into a state that can be validated and packaged. Those who are familiar with relational databases are well conditioned toward the notion of structured data. Each record in a relational database exhibits uniformity in terms of the associated data fields and their types. Records can be subdivided into assorted tables that can reference each other through keys. The underlying structure of the database is known as its *schema*.

Raw data, by nature, has no explicit schema. In order to assemble data from varied sources in a fashion that allows for analysis and modeling, the data scientist (historically speaking) has been required to possess extensive knowledge of the problem domain such that they can manually map raw data features to a set of target data features. For example, suppose the problem requires the determination of projected property tax revenue from a given jurisdiction; from a collection of public records obtained through local title/escrow companies along with assessment and taxation records from county and state web documents, the data scientist might create a target schema for structuring a database of property addresses, assessed values, and millage rates.

Once the raw data has been compiled and appropriately structured, the data must be "cleaned." Data cleaning refers to the process of accounting for missing data fields, identifying and rectifying incorrect records, and ensuring uniform formatting in the target dataset. Extending our example from the previous paragraph, let's

[1] https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

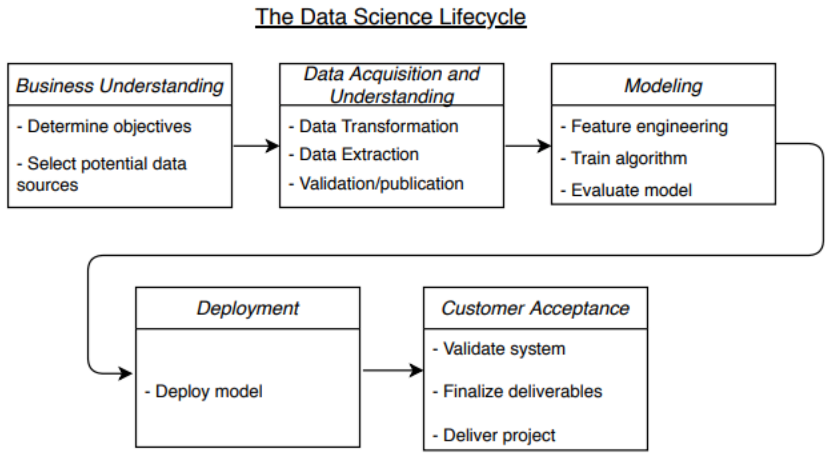

Fig. 1 The Data Science Lifecycle. Note how many subprocesses exist within nearly all phases of the pipeline. Many of these subprocesses contain further subprocesses, making the overall data science lifecycle substantially more complex than the layman can at first appreciate. Studies have shown that 50-80% of a data scientist's efforts are spent during *Data Acquisition and Understanding.*

suppose the end user requires that all zip codes be represented in their abbreviated five-digit form, but the raw data includes nine-digit zip codes. The data scientist must develop a method of locating the offending fields and transforming them to meet the user's specification. Furthermore, missing fields may be inferred from other data of similar nature or proximity; it the data scientist's challenge to determine when and how to apply such techniques, as misapplication can be costly.

Data enrichment, an optional third step in our data transformation procedure, involves augmenting the previously assembled dataset with further data; for some applications, for instance consumer behavior analytics, enriching a target dataset in this fashion can provide a broader, more complete picture of a given subject prior to packaging for analysis and modeling.

2.3 Validation / publication

Data validation and publication require little explanation, which is why they appear bundled in the same section. Validation refers to the process of ensuring the efforts made in our prior steps were fruitful, i.e. that the target dataset is complete, accurate, and consistently formatted – in short, that it is ready for further analysis. A simple validation might include type checking of fields (making sure street numbers are integers, for instance), along with checking for inclusion of certain character groupings in text fields (such as Blvd., St., Rd.). Upon validation, the data is "published," making it available for use in subsequent steps in the data science pipeline.

3 Automating Workflow

Section 2 provided a description of the steps involved in the data science pipeline. By no means is this an all-inclusive list, but it provides a foundation for understanding the complex workload one must undertake in order to arrive at a useable model from raw data. Reflecting upon the fact that each of these tasks were undertaken manually for decades, it should come as no surprise that a great deal of research has been performed with regard to task automation. Such studies have targeted very specific steps in the pipeline, such that there now exists a plethora of tools providing assistance in each of the above referenced tasks. This can be a double-edged sword, as data scientists must now possess knowledge of the many tools at their disposal, how to use them, and how one tool may provide better results than a similar one given the dataset in question.

Automation algorithms have only grown more powerful in the modern age of artificial intelligence and machine learning. Koehler *et al* [3] provide a description of one such methodology as it applies, specifically, to data transformation and validation (i.e. "wrangling"). Their technique initiates via a process known as "schema matching," which seeks to address the problem presented by schema-less raw data, as discussed in Section 2. In the schema matching phase of their automation workflow, the algorithm is provided with metadata compiled and managed by domain experts. This metadata may include keywords, data types, or other elements that may be useful in training the algorithm to recognize congruences between raw and target data items. Such recognition takes place through analysis of parsed tokens, which are then scored based on a confidence factor and output accordingly.

Schema matching, in this workflow, is followed by schema *mapping*, which uses the information accumulated in the previous step to rearrange the raw data such that it attains a structured, cohesive state. Prior to executing this step, the AI is provided with a training dataset; the algorithm evaluates this training data, which is effectively a smaller subset of the target dataset. With the information extracted from this training data, the algorithm can then initiate the process of distributing the raw data over a target schema. This is handled iteratively. Candidate mappings are evaluated and scored according to a confidence metric; upon completion, the algorithm returns the most viable candidate.

The next step in Koehler's data transformation process is referred to as *value format transformation*. This is the phase in which nine-digit zip codes might be reformatted as five-digit zip codes, "St." is expanded to "Street," and so on. The algorithm that provides this functionality does so by automatically detecting data that requires transformation, thereby triggering the appropriate modifications. As in prior steps, this process is informed through the evaluation of training data, as prepared and supplied by the data scientist.

By providing training data at each step in the aforementioned process, Koehler's team managed to achieve 95-98% accuracy when comparing their AI output with a reference ground-truth dataset.[2] Given these numbers, it is no mystery as to why the data science industry is seeing an explosion of AI-based data cleaning, transformation, analysis and modeling tools. With such an abundant variety of mechanisms to choose from, how can the data scientist begin to determine the appropriate tool for a given project?

Fear not: there are even tools that help us select tools. Biem *et al* [1] detail one such mechanism in their paper "Towards Cognitive Automation of Data Science." The solution depicted therein is an AI for extreme front-end preprocessing of raw data, along with a dynamic repository of algorithms used for automation of subsequent steps such as data transformation. The process is initiated by prompting the user for certain data and preferences relative to their project. This information is provided as input to an analytics repository, which is effectively a search engine for algorithms. (Indeed, their mechanism is capable of combing web-based resources, such as research papers and articles, in order to gain insights regarding the best procedures and algorithms for the given project or dataset.) Selected algorithms are then analyzed by a learning controller, which outputs the top candidates to the user. At each step in the process, the user is allowed interaction with the mechanism for the purpose of providing real-time feedback. For instance, if the user likes certain aspects of a recommended tool but not others, she may provide this as input to the learning controller, which can subsequently attempt to find a superior alternative.

These are simply a few of the AI tools available to the modern data scientist. Specialized tools are becoming widely available for all aspects of the data science pipeline. A comprehensive analysis is very likely impossible and goes beyond the scope of this paper.

---

[2] There are additional processes included in this method that go beyond the scope of this paper. As of this writing, one may view their paper through IEEE's archive for a complete description.

4 Ethical considerations

Given the tremendous amount of time and energy the data scientist spends on labor-intensive tasks in any given step of their workflow, it requires little cognitive effort to understand why automating these steps yields a tremendous advantage. Not only can the data scientist produce a model in far less time, but s/he can now do so using data elements that number into the millions (even tens or hundreds of millions) and at fraction of the cost. Boire [2] notes that some studies predict a multi-trillion-dollar boon to the US economy on account of computerized automation.

While such figures paint a rosy picture for the future of data science, they fail to reveal several lurking hazards. Section 4 will discuss the human cost of AI and automation as it relates to the potential for lost jobs, as well as algorithmic discrimination. It will conclude with a discussion of measures currently being explored that seek to mitigate such bias.

4.1 Jobs

In his paper "Artificial Intelligence (AI), Automation, and its Impact on Data Science", Boire [2] asserts that the fundamental motivator behind AI automation can be reduced to financial gain. Furthermore, automation can be seen as the next logical extension in a history of businesses sacrificing workers for profit. Says Boire, "This enhanced focus on cost effectiveness has also resulted in organizations exploring options to outsource tasks to countries with a lower standard of living. Over the years, we have seen how outsourcing of tasks has evolved from the more routine mundane tasks to the more highly advanced knowledge-based tasks." The upshot of such cost-cutting measures, naturally, is the loss of jobs such that presently only the most highly-trained individuals with the most advanced technical skills have the possibility of earning a living wage; this further perpetuates disparity between the income classes.

While outsourcing of this nature remains prevalent, it is not an all-purpose business solution; furthermore, as automated services become more mainstream and less expensive, outsourcing to cheaper laborers may become less cost effecting than leveraging of machines. Indeed, machines have the potential to provide a far greater breadth of services in the long term, and at greater cost savings. Take the familiar example of the self-checkout kiosk at your local supermarket, for instance. According to a 2018 article published by Vox.com [9], a single station requires an investment of approximately $30,000 to $60,000. The mid- to high-end automated "checker" will recoup the cost of a full-time worker's wages in one to two years; add health insurance, retirement contributions, paid vacation/sick leave, and raises to the mix, and one can see this investment paying for itself within the first year. Supplement all this with the fact that machines (arguably) do not need breaks and can work around the clock, and it's no mystery as to why one might see a single human worker overseeing a pod of fifteen or more machines at a large grocery or home-supply store. And, while the trade-off of human workers for their automated counterparts in this industry may not be one-to-one, when viewed through a grander lens – automotive, manufacturing, customer service and agricultural workers all being replaced by machines – one can truly begin to see the impact on low- to middle-income wage earners.

Data science, of course, falls into a different category. The data scientist, by and large, comes to the table with an advanced skillset obtained through schooling and experience; their skills stretch from programming and statistics to business solutions. While advancements in end-to-end AI solutions may put some data scientists on edge, most tools in use today target individual segments of the pipeline and still require extensive human intervention to achieve optimal results.

Fast forward a decade or two and the outlook gets murkier. Algorithms will improve, cloud-computing will make processing large datasets less time consuming, and end-to-end AI solutions may hit their stride. Several possible scenarios emerge as the new reality for data scientists. In the first scenario, they are simply rendered obsolete, replaced wholesale via a combination of AI's and computer scientists. This may seem alarmist from a ten-thousand-foot perspective; however, the Twentieth Century saw the end of telephone operators, typesetters, and travel agents due to technological advancements; are grocery store clerks and perhaps even data scientists soon to follow?

Examination of a second scenario reveals a much more likely outcome. Boire predicts that data scientists in the coming decades will be split into two camps: those who specialize in business problem solving, and those who specialize in creating the algorithms and AI technology necessary to automate the more technical tasks; his prediction goes so far as to suggest that colleges and universities will begin offering separate degrees for each sub-specialty. The emphasis for the data scientists of the future will be on thinking and problem solving rather than performance of laborious tasks such as data wrangling. In a related but slightly divergent scenario, automation may well make the data science industry more accessible to those without highly specialized training, effectively *creating* jobs in the longer term. This leads to the question of whether the loss of jobs in data science will appear linear, as in the first scenario related about, or take on a parabolic character as new segments of the population gain access to the field.

It is not unlikely that our future holds some hybrid of the above scenarios. Jobs will be lost due to automation in some areas of the industry, while jobs will be gained in others. In other words, the data science industry will see changes, but the possibility of its wholesale elimination seems remote at best. In this modern age of information, the only certainty is that big data will remain as a prime mover in business decision making.

4.2 Bias

There are algorithms that deal with weightier topics than business decisions, of course, which leads to an ethical discussion of a different nature. These algorithms result in predictive models which can be used to make decisions about topics ranging from sports teams to policing, depending upon the data being studied; they are also highly subject to discriminative bias.

Bias in algorithms can be intentional or unintentional, explicit or implicit. One jarring example of seemingly unintentional yet explicit bias stems from a 2016 paper in which the authors, Wu and Zhang [10], claimed to have created an AI that could determine the criminality of an individual by their facial features. As subsequently pointed out by Bergstrom and West [11], however, the training data possessed one fatal flaw: all images of convicted criminals provided to the AI reflected frowning faces, while all images of non-criminals reflected neutral or smiling faces. The thoughtful reader can draw their own conclusions regarding the veracity of the resulting predictive model.

Faulty algorithms are not exclusively the result of dubious science, however. To be sure, bias can originate from any number of sources, many of which remain undetected by programmers for quite some time. For instance, remember from Section 2 that raw source data is often incomplete at best, incorrect at worst. This may be of little consequence when attempting to use such data to predict consumer behavior (presuming we consider “little consequence” to encompass financial loss); however, what if the model is used to predict rates of recidivism for use in a convicted criminal’s rehabilitation plan? In this case, missing/incorrect data (or fields that are manually repaired with incorrect information) can result in, at worst, racial disparities in predictive models that may impact an individual’s probability of qualifying for parole.

More troubling is the notion that even complete, correct datasets can yield biased models. Lux *et al* [4] point to a scenario in which historical data is used to train an AI for the purpose of assisting an employer with hiring decisions. Unbeknownst to this employer, this historical dataset may contain protected information such as race/ethnicity, age, religion, sex, gender identity, national origin or disability. By law, employers may not use protected information of this nature in their hiring decisions, yet the recommendations made by the AI may be partially or wholly determined by such data. In a scenario such as this one, the employer may be participating in discriminatory hiring practices without even knowing as much.

The cases grow darker from here. In a piece published by Smithsonian Magazine, Rieland [12] details the use of an experimental software tool known as PredPol by the Los Angeles Police Department. The UCLA scientists who developed the technology, according to Rieland, purported their software to be "twice as accurate as human analysts when it comes to predicting where crimes will happen."[3] The software makes its predictions through analysis of crime data. Specifically, three metrics are used: the type of crime committed, the location of the crime, and the date and time that the crime occurred. Civil rights organizations, including (but not limited to) the American Civil Liberties Union and the Brennan Center for Justice, were quick to sound the alarm, alleging that such metrics had the potential to generate feedback loops: would police officers be more zealous in making arrests in neighborhoods that generated higher crime scores, and would higher crime scores result from these arrests? While accusations of systemic racism have kindled spirited debates in the policing community of late, there is a caustic irony to the notion that the future of policework may hinge on algorithms that result in disproportionate incarceration of targeted individuals.

Even in today's era of renewed awareness, there exists a pervasive naiveté when it comes to the benefits of technology in the eradication of inequity. Indeed, the layman would ask how a machine could produce biased results when it cannot detect color or sex, whereas the experienced scientist recognizes that machines simply carry out human intolerance with greater efficiency. After all, humans – with all of their imperfections – created the machine.

4.3 Solutions

Racial and gender-based discrimination in predictive models has been well documented and is an area of active research. While the present age may spring forth renewed urgency due to mainstream recognition of the Me Too and Black Lives Matter movements, studies on algorithmic discrimination began to emerge as early as the 1970s. More recent studies have attempted to scale this multifaceted dilemma into a comfortably limited problem set, thus allowing scientists to concentrate their efforts. One such study, produced by Calmon *et al* [7], reduced algorithmic bias into two prevailing categories: disparate treatment, in which discrimination occurs due to explicit consideration of federally protected class values, and disparate impact, in which an algorithm considers variables that hold some correlation to protected class values but does not consider such values directly.[4] The study pointed out that disparate impact could be addressed through two common but sometimes conflicting goals, the first according to the principles of group fairness, and the second according to principles of individual fairness. Group fairness principles dictate that an algorithm should yield similar results for all groups, whereas individual fairness principles seek comparable treatment for like individuals regardless of group identity. Because error rates are the easiest benchmark upon which to measure success, the team sought to produce a method that resulted in equal error rates amongst all groups while preserving the accuracy of the output when tested against a ground-truth comparison data set.

This team and others have noted that the data science pipeline contains three stages at which algorithmic discrimination can be addressed: 1) the administration of training data (aka "pre-processing"); 2) within the learning algorithm itself (aka "in-processing"); and 3) during human analysis and modeling of the results (aka "post-processing"). Lux's study [4], for instance, focused on a pre-processing solution in which human interaction was key to bias prevention; the resulting paper proposed that, through intervention on a smaller subset of data, potential for bias could be determined through algorithmic analysis, as well as by the users themselves. By detecting problem spots in the raw dataset, discriminatory values could be pinpointed and cleaned prior to evaluation.

Calmon's team, however, points out that discarding protected class values from the original dataset is akin to skirting the issue of disparate impact rather than addressing it head-on; more problematic is the notion that

[3] The author pointed out that this assertion had not seen independent verification; the claim was conspicuously absent from their website, predpol.com, as of this writing.
[4] Note that disparate impact is the result of indirect discrimination but can be either intentional or unintentional depending upon context.

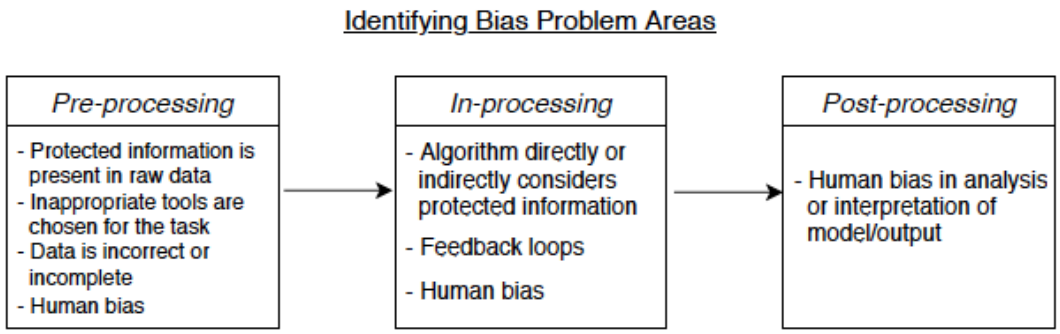

Fig. 2 Discriminatory practices can occur in any of three phases: pre-processing, in-processing, or post-processing. Bias may be explicit or implicit; implicit bias can be particularly difficult to detect. Note that human bias may be present in any/all of these phases; it is left as an exercise to the reader as to how this might manifest in each phase.

cleaning such values from raw data can inadvertently serve to exacerbate algorithmic bias. (The irony that indirect discrimination can be aggravated through efforts to rectify it has not been overlooked and is amongst the factors that make this an area of active research.) Once Calmon's team ruled out cleaning protected data as a viable solution, they were left with an optimization problem that sought a finely tuned balance between group and individual fairness parameters while minimizing loss in accuracy. More specifically, given a source dataset D * X * Y, in which D represents discriminatory variables (such as race), X represents non-discriminatory variables (such as credit score), and Y represents an outcome variable (such as a loan-approval decision), their goal was to "determine a randomized mapping…that (i) transforms the given dataset into a new dataset…which may be used to train a model, and (ii) similarly transforms data to which the model is applied." Discrimination, therefore, could be reasonably expected if, given a random sample of data, some protected class value D could be consistently deduced relative to some outcome variable Y. This principle helped guide the team in formulating and calibrating a solution that not only achieved balance between group- and individual-fairness but maintained integrity in the subsequent output.

The resulting algorithm was tested on recidivism data for which prior algorithmic models had been applied, and for which ground-truth comparison data was available. They found that their mapping succeeded in greatly reducing discrimination, particularly amongst African American males, at a small cost to accuracy when used to train applicable AI algorithms. (Complete results may be examined in their paper Data Pre-Processing for Discrimination Prevention: Information-Theoretic Optimization and Analysis.") The team noted that their work could be used as the basis for future research on algorithmic bias and offered myriad prospects for continued development.

Calmon's and Lux's studies are but several of many investigations into algorithmic bias, its causes, and its potential solutions. Awareness surrounding implicit bias in computer/data science has been gaining traction in recent years, though we remain on the front end of attaining meaningful solutions. Expect novel contributions to reach fever pitch in the coming decade as this heretofore technical issue gains salience in the public eye.

Conclusion

Data science was born a hybrid industry, melding algorithms and statistical studies with business profitability and community development. Predictive modeling, its logical extension, was developed for the sole purpose of anticipating needs, and artificial intelligence has made the field accessible to a broader body of users. Carried out successfully, such models can be used to effect social change; executed poorly, and they hold the power to proliferate racist policies and leave communities wanting for needed supplies.

This paper discussed the complexity of the data science pipeline, but more importantly, it pointed out the potential for human and algorithmic bias in each phase thereof. We have already seen the adverse impacts of ill-conceived models in studies on recidivism and, arguably, policing; other such cases will undoubtedly come to light in the not-so-distant future. Awareness alone is incapable of preventing bias and discrimination within the data science pipeline and the models subsequently produced. If those in this community seek to develop the technologies and processes that would carry us into the future, then they must actively and tirelessly act to remedy the blunders of our past. Moreover, if the community maintains sight of the notion that this industry was created by people for the betterment of people – and can balance technological advancements with a wealth of human ingenuity – its future will remain bright.

ACKNOWLEDGMENTS

I would like to thank Professor Kristin Tufte, the instructor for this course, for providing advice, ideas, and resources for this project.

REFERENCES

[1] Alain Biem, Maria A. Butrico, Mark D. Feblowitz, Tim Klinger, Yuri Malitsky, Kenney Ng, Adam Perer, Chandra Reddy, Anton V. Riabov, Horst Samulowitz, Daby Sow, Gerald Tesauro, Deepak Turaga. 2015. Towards Cognitive Automation of Data Science. AAAI Conference on Artificial Intelligence. [Online]

[2] Richard Boire. 2017. "Artificial Intelligence(AI), Automation, and its Impact on Data Science." *2017 IEEE International Conference on Big Data (Big Data),* Boston, MA, pp. 3571-3574 DOI: 10.1109/BigData.2017.8258349

[3] Martin Koehler, Alex Bogatu, Cristina Civili, Nikolaos Konstantinou, Edward Abel, Alvaro A.A. Fernandes, John Keane, Leonid Libkin, Norman W. Paton. 2017. "Data Context Informed Data Wrangling," *2017 IEEE International Conference on Big Data (Big Data),* Boston, MA, pp. 956-963, DOI: 10.1109/BigData.2017.8258015

[4] Thomas C.H. Lux, Stefan Nagy, Mohammed Almanaa, Sirui Yao, Reid Bixler. 2019. "A Case Study on a Sustainable Framework for Ethically Aware Predictive Modeling," *2019 IEEE International Symposium on Technology and Society (ISTAS),* Medford, MA, pp. 1-7, DOI: 10.1109/ISTAS48415.2019.8937885

[5] S. G. Shaila, M S M Prasanna, L Niveditha, A Vadivel. 2017. "Support Automation Connectivity System for Big Data Applications," *2017 International Conference on Inventive Computing and Informatics (ICICI)*, Coimbatore, 2017, pp. 805-809, DOI: 10.1109/ICICI.2017.8365248

[6] D. Vianna, A. Yong, C. Xia, A. Marian and T. Nguyen, "A tool for personal data extraction," *2014 IEEE 30th International Conference on Data Engineering Workshops,* Chicago, IL, 2014, pp. 80-83, DOI: 10:1109/ICDEW.2014.6818308

[7] F. d. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy and K.R. Varshney, "Data Pre-Processing for Discrimination Prevention: Information-Theoretic Optimization and Analysis," in *IEEE Journal of Selected Topics in Signal Processing,* vol. 12, no. 5, pp. 1106-1119, Oct. 2018, DOI: 10.1109/JSTSP.2018.2865887

[8] Charu Aggarwal, Djallel Bouneffouf, Horst Samulowitz, Beat Buesser, Thang Hoang, Udayan Khurana, Sijia Liu, Tejaswini Pedapati, Parikshit Ram, Ambrish Rawat, Martin Wistuba, Alexander Gray, "How can AI Automate End-to-End Data Science?" Oct. 2019, arXiv:1910.14436

[9] Kaitlyn Tiffany, "Wouldn't it be better if self-checkout just died?", Oct. 2018, vox.com/the-goods/2018/10/2/17923050/self-checkout-amazon-walmart-automation-jobs-surveillance

[10] Wu, Xiaolin and Xi Zhang. "Automated Inference on Criminality using Face Images." 2016, arXiv: 1611.04135

[11] Carl Bergstrom and Jevin West, "Criminal Machine Learning", April 2017, callingbullshit.org/case_studies/case_study_criminal_machine_learning.html

[12] Randy Rieland, "Artificial Intelligence Is Now Used to Predict Crime. But Is It Biased?", March 2018, smithsonianmag.com/innovation/artificial-intelligence-is-now-used-predict-crime-is-it-biased-180968337/