Data, what is it good for? We don't know!

Matthew Littlejohn
SAIC
Austin, TX
matthew.littlejohn@saic.com

ABSTRACT

As organizations collect an increasing volume of data, their employees and vendors are tasked with making sense of it. This is done for a multitude of reasons – to create future strategy, develop products, understand customers, and a myriad of others. But data for data's sake can lead to confusion. Without a clear vision or idea, being given troves of data without instruction can lead to bad analytics, flawed insight, and ineffective strategy. In this paper, we will provide examples, based on our experience with the US Air Force (USAF) Air Education Training Command (AETC) Pilot Training Next (PTN) initiative, of what happens when too much data is presented, and common pitfalls that arise when analysts are not given a clear mission or guidance. We will discuss PTN's data driven system architecture and analytic strategy as it evolved, while providing a state of the art learning platform to new student pilots. Lastly, we will provide recommendations for organizations to prevent these mistakes, as well as provide best practices and behaviors to nurture better data literacy, strategy, and mission creation which will help accelerate training system adaptation and innovation to an unexpected future. This paper is intended for program managers and leaders, and no data analytics background is assumed.

ABOUT THE AUTHOR

Matthew Littlejohn is a data science senior manager at SAIC where he leads data and analytics teams to provide customers with compelling insights through machine learning, visualization, and data strategy. With over a decade of analytics experience, Mr. Littlejohn has worked across multiple industries, including automotive, CPG, government, H&PS, and others, receiving praise and recognition for his work in developing novel solutions, strategy, and training. He holds a master's degrees in bioinformatics from Texas A&M University and biotechnology from Southeast Missouri State University, and lives in Austin, TX.

Data, what is it good for? We don't know!

Matthew Littlejohn
SAIC
Austin, TX
matthew.littlejohn@saic.com

INTRODUCTION

Throughout human history, and as long as we as a species have needed to infer the world around us, we have made observations, noted changes as they came, and recorded these findings to support various claims and arguments. These findings and notes came to be known as data, and they are the backbone for all forms of research, evidence, and identifying what is "true" and what is not. As technology advanced, so too did how we collect and use data. Not only did our collection methods grow, but the volume of data we were able to collect and analyze grew along with it.

As statistics and computer science both evolved and become more powerful, data analysis became more advanced and precise. When the Internet came to be, this led to an exponential growth in data being generated – and collected. As organizations began learning more about people and the choices they made, they began noticing patterns that would change or be modified along with the situation and environment. Whether it was at a certain location or using a cell phone instead of a computer, people acted differently and in ways that ran counter to what others like them would do. This led to more data sources being created and more data being captured, cumulating to what is aptly defined as "big data."

But while the number of tools, data collection points, and volume of data itself all grew, so too did the number of answers that could be inferred. Schools were able to begin identifying how to help students pick more interesting courses, marketing firms were better presenting products to targeted customers with ease and gusto, and social media sites can predict human behavior and decisions up to weeks before an individual realizes he or she has made a decision. All of these phenomenal insights came to be because of the advances made in data analytics, machine learning, and artificial intelligence (AI).

That does not mean these insights came easily. Throughout the development of these remarkable tools, a myriad of mistakes were made and several inaccuracies surfaced. And while certainly all of these major companies had budgets to afford these misfires, for those new to or just entering the world of advanced analytics, there is no guide on where to start. And with the hype and growing number of papers praising the benefits of data, you must forgive people for jumping in head first, thinking "the data will tell the story for us." More often than not, these teams and managers will surface with countless hours spent and little to show, forcing them to question why data gets so much praise with seemingly little actionable outcome.

The purpose of this paper is to address this frustration head-on. In working with data and analytics for years, we have observed and corrected countless instances where the data may tell an extremely compelling story, but not the one that needed to be told. Throughout this essay, we shall provide instances where organizations thought their data could do more than it was able, undervalued the insights that lie waiting in their databases, focused too much on the wrong piece of information, and misinterpreted otherwise invaluable results. By the end, our hope is that you will have a better understanding of how to use data to form hypotheses, develop powerful data-driven insights, and set up data systems and teams for success.

BACKGROUND

Before there was data science and advanced analytics, there was statistics. Even without the computational power we are able to harness today, researchers and curious minds used probability and statistics to solve problems, with one of the most famous examples being Gregor Mendel's use of probability to calculate heredity in peas (De Castro, 2016).

Scientists were not the only ones utilizing these resources, though – entrepreneurs and businesses eventually began to understand the power of statistical prowess to solve problems and optimize operations, marketing, and budgets; Guinness Brewing, for example, hired teams of statisticians and mathematics researchers for such purpose as early as 1900 (Stingler, 1990).

Throughout the 20th century, multiple advances in statistics and analytics began to flourish and be widely accepted through multiple disciplines. Improvements in computational power, performance, and capability exponentially led to the development of more accurate and comprehensive statistical methods (Weihs and Ickstadt, 2018). These advancements coincide with the emergence of data analysts becoming more frequently utilized in multiple industries, initially in marketing and finance. The former was where customer behaviors were able to be tracked and analyzed, leading to multiple businesses using data analytics to drive strategy and better anticipate customer needs and market changes (Duhigg, 2012). Eventually, researchers and educators discovered that behavior could be modified and the most effective methods of persuasion could be predicted and used to best influence decisions (Thaler and Sunstein, 2008).

Buildup to a Knowledge Gap

As these benefits became more apparent, the need for analytics professionals grew through the 2000s (BLS, 2021). However, many organizations have found it difficult to fill vacant roles (Ramachandran and Watson, 2021). Figure 1 shows that while the interest and vacancies in data analytics has grown, the number of available candidates has not grown at a similar pace (Flowers, 2019). With such a shortage, many organizations and leaders have been forced to look inward towards their current staff to solve data and analytics problems. Unfortunately, many of these employees do not have the necessary knowledge of statistics, programming, analytics, or problem solving to fill the role of an analytics professional (Schirf and Serapiglia, 2017; Mikalef and Krogstie, 2019). These skill gaps could potentially be resolved by continuing education, training, and other forms up upskilling, but attempts to help current employees improve their skills and knowledge in technical fields has been met with resistance, either from organizations not having sufficient funding or desire, employees' resistance to learning new skills or lack of available time, or some combination therein (WEF, 2021).

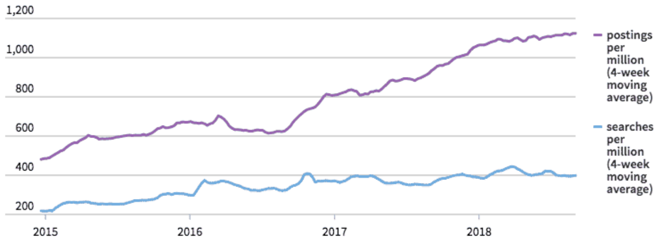

Figure 1. Data Analytics Job Postings (purple) vs. Job Searches (blue)

Arguably, in lieu of hiring additional staff, the largest hurdle of filling this gap is convincing more experienced and senior employees to pursue continuing education or training. As these senior employees have built methods and habits to improve their performance and deliver exemplary results, it is unnatural for them, as with anyone else, to break these habits and apply new methods, especially if they have no prior knowledge or experience (Duhigg, 2012). Additionally, management or leadership might find it easier to expand the search for new talent to perform these new tasks, regardless of candidate shortages (Ramachandran and Watson, 2021). This may be due to a belief that new talent or younger employees have a better understanding of these new methods, or that the more experienced employees would not be as effective in performing these tasks even after receiving additional training (Ng and Law, 2014 and Vallor, 2015). Regardless of the reason, leaving this gap open will inevitably result in misinformation and a decline in quality (Mikalf and Krogstie, 2019).

While the above example of skills gap is a common occurrence, there exists a related and equally difficult situation that can cause problems. Take for example a senior team member who has worked for a long time but has not stayed up to date with advances or more recent developments in their field. This lack of exposure gives these talented team members the impression that their contributions are still state of the art or novel, when in actuality it does not follow best practices (Ferdows, 2018 and Ng and Law, 2014). One could argue that convincing these individuals to upskill is more difficult, as they feel their knowledge and experience is still relevant and that any training would be redundant (Johansson and Abrahamsson, 2021). Just as was the case with not having personnel with the knowledge to perform data-oriented tasks, not staying relevant with the field or refusing to learn to attend trainings can leave employees and entities at a competitive disadvantage.

Missing Knowledge, Missed Opportunities

While companies are looking to either upskill or hire new talent, they are still expected to meet the needs of their customers in the form of products, strategy, or other solutions. In the case of data solutions, the knowledge gap does not make these companies exempt from delivering such products. This is where missteps and objectionable products will be delivered by the organization. Without the knowledge to properly design, develop, and deliver data solutions, institutions must rely on their current resources and employees, who will rely on their limited knowledge and expertise in the field. This can and frequently does lead to misinformation, outdated or legacy solutions, and potential loss of future business (Bloch, 2011 and Latin, 1993).

Ultimately, this can lead to both businesses, employees, and customers losing trust in the work being delivered, which in turn leads to mistrust in the data and analytics process. As frustration with problem solving grows along with inadequate results, one could not bemoan an organization from abandoning data services or products and instead focus on other objectives or products (De Santis, Scannapieco, and Catarci, 2003, Bahramirad, Svachula, and Juna, 2014). This distrust in analytics and data science can lead to improper analysis and planning, a reduction in the number and quality of offerings or products, and an inevitable failure to be competitive in the landscape (Bahramirad, Svachula, and Juna, 2014).

PROBLEMS THAT HAVE ARISEN

Despite the apparent risk and demonstrable need for robust and capable data teams, an alarming number of organizations do not realize their deficiencies until they have already engaged in projects. In our experiences in working with customers to solve their analytics and data problems, the most common point of realization occurs when a stakeholder or product manager asks to see the minimum viable product (MVP). As the data team begins gathering their results and findings, they realize that they do not have enough high quality data to build an MVP. This leads to poor deliveries and a disappointed PM left asking “what’s the point of collecting all this data when you can’t do anything with it?” While no professional wants to be in the unenviable position of having to explain how their team is doing the best with what they have, it is a difficult conversation that has been played out too many times and puts leadership in the difficult position of deciding whether to abandon the project, hire additional resources, or seek outside help from consultants or similar firms.

Conversely, a team can have seemingly unlimited resources, be staffed by stars within the organization, and have full backing from leadership, but lack guidance or direction. The often uttered phrase “we’ll let the data tell the story” may sound appealing and give the impression that exploration and creativity will be unobstructed, but in practice this has caused more problems than answers. In our experiences, this has been a result of customers and their leadership or stakeholders wanting to be more data driven but not knowing where to begin or how to build around that mindset. On multiple occasions, leadership has told us that they thought data scientists and engineers could read code, perform calculations, and produce high quality results within days, not realizing the setup and analysis could take weeks, if not months. Such examples are not uncommon, but anecdotally have been less frequent as stakeholders, product owners, and others have let analytics professionals estimate these timelines. Despite this, there are select situations that appear more frequently than others and should be addressed

Letting Data Tell the Story

As previously mentioned, customers and organizations that want to be data conversant and more data driven often make the mistake of not formulating a goal, strategy, or infrastructure, insisting that the data will provide all the necessary information. In theory this appears sensible – data is simply information presented in a way that can be analyzed and studied, so in the hands of capable professionals this should be a simple task. The problems arise, however, when there is a lack of understanding of the data itself and what the end goal is. While data professionals can indeed use data to tell a story, they do not always know what story to tell. As an example, a former automotive client of mine wanted us to analyze sales data for a particular region and use those findings tell them “what [they] should do.” When pressed for clarification, the customer’s product owner (PO) insisted that “the data will tell the story” and that we would be able to interpret the necessary actions and strategy. With these cryptic instructions, we were indeed able to identify sales trends for those regions, identify which customer segments responded best to certain promotions and ads, and developed an optimization engine to help this client better manage their marketing budgets. After we delivered our work to the PO, he decided to use these analyses and results in a completely different region and different vehicle class. While the data told the story for this customer set and how they reacted to smaller cars, it was not the right story for the target region to sell larger vehicles, and the PO's assumptions proved disastrous for his company.

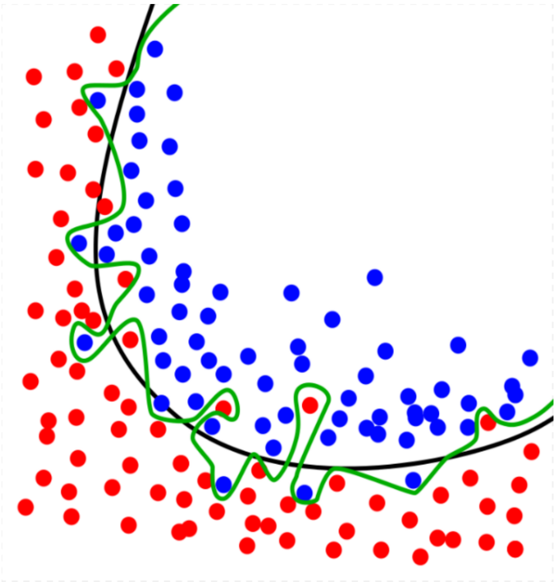

Figure 2. Example of Over-fitting. Here, the Black line shows an appropriately fitted regression line, while the Green line represents an over-fitted regression line.

Another example of letting the data tell the story was a misguided strategy while working with a vendor in the consumer packaged goods (CPG) industry. While developing how to best deploy coupons to shoppers, this vendor spent a large amount of money on their in-house data collection as well as on customer data provided by multiple third parties. When we were brought in to make sense of the big data ecosystem, we were tasked with finding new trends and correlations between the multiple data sources. To complicate matters, because this customer had spent so much money on third party data, they were insistent that we utilize it so they would not be seen as wasting their budget. While we were able to identify trends and patterns using these data, it came at the cost of the models being too specific for individual data interactions which, in practice, did not represent real world scenarios. This process is referred to as "over-fitting" a model, and occurs when a statistical model is built to fit too closely with sample data to the point that it cannot be applied to new, real data. Figure 2 is a visualized an example. Ultimately, the data did indeed tell a story, but it was not one that could be used to drive strategy or give confident results. Once we were allowed to trim back the number of required data sources, we were able to deliver the products the customer needed.

Even in instances where letting the data tell the story is beneficial, this can be an extremely time consuming process. Within our Pilot Training Next (PTN) program, the Air Force initially did not know the exact questions to ask. Over the course of two years, there were realizations that data was missing or incomplete and leadership changes on both the customer and our side shifted the scope and direction, which led to frequent requirement changes. While PTN is now on track and has a clear goal, the changes in this process meant that most of our historical data cannot be used going forward, as the sources, intent, and systems deviated too drastically. While the old data can be used as a snapshot, the drastic changes meant it could not be used as a comparison or benchmark to more recent data.

Poor Data Quality

Proper data collection, classification, and access are necessary for all analytics tasks, as is ensuring robust systems are in place to maintain these processes. Many organizations and practitioners new to data analytics frequently make the mistake of assuming data collection is a plug-and-play solution and that the data they are collecting will contain all necessary information in the onset. Regrettably, they only realize their data is lacking or not formatted properly as it is streaming in, making troubleshooting more urgent and potentially losing important entries.

The first type of poor data quality is improper formatting. This can occur either from the data source, within the data collection or extract/transform/load (ETL) processes, or within the database or data lake. The last example is more problematic when the data is stored in an unsuitable format for future analytics, visualization, or other tasks. This is the common and fortunately easy to fix. When data formatting problems occur in the collection or ETL stages, this usually is a result of not setting up the data collection protocols properly, or a data engineer making a mistake somewhere in the process. Fortunately, a competent data engineer should have little problem modifying the ETL procedures to ensure the data is presented in an appropriate format. But there do exist instances where data being pushed from the source requires the software or other development teams to make changes to the source.

The second type of poor data quality has to do with the data itself, regardless of formatting. This can occur when data being presented uses excessive special characters, contains characters that double as delimiters (for example, excessive comma use in a comma delimited file), or uses foreign language characters. An example I experienced was working with a Japanese *juku*, or after-school tutoring service, who sent all their data in Japanese. While the numbers and some symbols were not problematic, our systems could not read most data cells because the Japanese characters were not compatible. This led to a costly delay in our deliverables, as we required additional resources to translate everything into English.

The third type of poor data quality is deprecated and legacy data. This is becoming less of an issue as organization implement more industrial and global standards, but it is still prevalent in many industries, including education, healthcare, and the public sector. Data and software are considered legacy if they relate to an old or previously outdated computer system that is still in use (but not commonly accepted) and require translation to a modern system in order to be more usable. Legacy programs include early versions of Fortran and MS-DOS, while legacy data examples include older versions of SQL, data stored on floppy disks, and SAS files containing deprecated procedures. While most legacy data is archival, historical data is often used to understand the evolution of processes, especially in longitudinal studies, or needs to be used as training data for analyses. Within PTN, this became an issue when we changed how the data was collected and used for reporting. Between PTN version 2 and version 3, the Air Force realized they needed student curricula and academic data to be better classified and structured. In making this change, we not only changed the grading and reporting systems, but how the courses flowed together and were organized changed so drastically that we could not confidently measure year over year improvement between the cohorts. Additionally, many organizations still rely on legacy software in their day to day activities. This can cause problems with delivering quality analytics, as there is a risk of improper formatting and loss of data once these files are converted into a workable format.

The final type of poor data quality has nothing to do with what is contained in the databases or data sources. Rather, it relates to inaccurate, incomplete, or cryptic requirements and expectations. Unlike letting the data tell the story, this is considered poor data quality because the project requirements and needs themselves are considered data – specifically, the points of data that results will be measured against. As an example, while working with a public health department, we were tasked with improving their county-level health index reports. When asked what needed improvement and what they wanted, the product owners very cryptically and bluntly told us to "show how care has changed over time across all demographics." When asked for clarification, we were only instructed to look at what was already available and "make it better." This was an extremely subjective ask, since what we thought was better did not align with the customer's expectations. During our bi-weekly updates, we would present status updates only to be told that what we had made didn't align with what the customer wanted. It was during these meetings that we were able to gradually get an understanding of what the customer needed. As demonstrated, without a clear understanding of the project expectations or the questions needing to be answered, the chance of performing the wrong analytics, solving the wrong problem, addressing the wrong audience, and other missteps occurring will become increasingly probable, which can prove detrimental to all parties and stakeholders. To these stakeholders, the expense for data services could be called into question. And to the data team, this can prove frustrating and demoralizing, which in turn can have a negative effect on the final product and discourage working with similar customers.

Biases

It is important to note that bias can have two very distinct meanings when working with data. In statistics and machine learning, "bias" refers to the difference between an estimator's expected value and its true value (Millsap and Everson, 1993). Here, "bias" will refer to its more common definition of "a tendency, inclination, or prejudice toward or against something or someone."

Data professionals have developed a reputation for impartiality and letting their work and findings speak for itself. In many instances, these findings run counter to what a customer or organization expects, or worse, needs. While everyone has moments where their biases interfere with decision making, data is meant to be objective and free of outside influences as much as possible. Regrettably, it is not uncommon for organizations to make bold assumptions or promises with the notion that upcoming data will back their claims. Friction will inevitably form when reality does not match these expectations, and subsequent decisions can call an analysis's validity and impartiality into question.

The most frequent way a bias can prove to be problematic is when a confirmation bias has been established. The most common sort of cognitive bias, confirmation bias occurs when a researcher, program manager, or organization has a preconceived notion about a subject and only includes data that corroborate this notion. It is not limited to just a belief about a subject and can also exist when a stakeholder believes a certain resource, group, or method is either more or less important than it actually is. Cognitive bias leads to inaccurate assumptions being treated as fact, disregards new strategies or ideas, and can prove hostile amongst teams. As an example, when helping a state health department build an electric health record (EHR) system, we were told to only focus on data coming out of hospitals and doctors' offices while ignoring local clinics. When asked, the research lead felt that clinics would keep horrible records, not see any major foot traffic or activity, and would not contribute anything helpful. After a week of data mining and analyses, we were not getting significant results and the lead could not understand why. Once we added clinics to the EHR network, the large gaps began to disappear and we were able to accurately build records around patients. The lead's bias against clinics led to him ignoring the state's demographics for our target area, which had a high population of uninsured workers who could not afford visits to doctors' offices or hospitals and instead had to rely on clinics for healthcare.

When working with data, it makes sense that you have to rely on all the data that is being collected and what is available to you. Specifically, if data cannot be quantified somehow, then it is best left ignored in favor of data that can be quantified. Yes, this can be considered a confirmation bias, but when working with large amounts of data this is also seen as a more efficient way of working. Yet by doing this, and as has been mentioned in previous examples, many data professionals fail to consider factors that are not being collected. This fallacy is often a result of not truly knowing the data or the context in which it is presented, or thinking context is irrelevant. This frequently leads to important considerations being ignored and explanations being dismissed, as was the case in the above EHR exercise.

But what about instances where a specific result is deemed necessary? As alluded to in the first example of this section, what if there was an instance where a customer or researcher needed a certain result to sell a product or convince a committee? This would be considered a confirmation bias, but more specifically, it would be referred to as the observer-expectancy effect. As the name indicates, this is a cognitive bias where a decision has already been made by a researcher or leader, and any results that do not support, confirm, or otherwise run counter to that decision are disregarded. This usually manifests in high pressure and time-sensitive situations and can often lead to poor analytics practices, inaccurate results, and misrepresentation or spin of information to appease an audience. This is not necessarily a bad thing – there may be instances where it's best to leave bad news for a closed-door conversation with a senior leader while the rest of the audience is only informed of key features. But in instances where unexpected outcomes arise and the hypothesized ones are proven wrong, reverting to the observer-expectancy effect can cause a myriad of problems, and skewing the data to spin it to appear different is becoming increasingly easier to spot and can call the data and analyses of all other projects into question.

AVOIDING PROBLEMS AND CHAMPIONING DATA

Despite the severity and number of factors that could draw a data project's validity and reliance into question, these can easily be negated by proper strategy and planning. While some of the examples provided did spell trouble, they were all resolved and able to avoid future conflicts. Yet the ideal time to fix issues should not be in the middle of a project or near its delivery – these are things that should be taken into account during the proposal phase. With a handful of planning strategies and checks, most problems should be easy to avoid or, if they arise, will not require significant time or effort to resolve.

Know the Objective

As mentioned, "letting the data tell the story" is infrequently the ideal strategy. In academic, exploratory, or similar applications where there are no set expectations or timelines, this can be performed with little consequence. Unfortunately, most are not awarded this situation, so preparedness is crucial to ensure punctual and actionable results. Rather than letting the data tell the story, it is imperative that questions, goals, and objectives be clearly defined and communicated prior to beginning any data work. This may seem like a fairly straightforward course of action and one that is likely performed before every task, and indeed it is likely done within the data team. But there is frequently a break in communication between the data team, leadership, and other stakeholders, which can lead to confusion and misunderstanding of the goals and value the data can provide. Ensuring these stakeholders and the data team are aligned ensures the tasks are well defined, expectations and limitations are known, and challenges that may arise can be anticipated and efficiently dealt with.

In both software development and data projects, the Agile methodology has been used as a template for how to properly plan, manage, and deliver products. In Agile, the goal is not to produce a product over a long timeline using a waterfall approach, but rather to break the development down into a series of 2-4 week periods of work known as "sprints." The goal is to gradually build to a final product, with the first step being the minimum viable product, or MVP. Prior to the project and each sprint, the team lead will outline tasks to be performed in the sprint using what's aptly known as a planning phase. The goal of the project planning phase is to work with the product owner and other stakeholders to outline what the expectations are, the desired final product delivery, and list resources, tools, staff, and other required components to complete the project. The goal of each sprint planning session is to determine a reasonable amount of work that can be performed in the upcoming sprint to work towards the MVP, updates, or final product. With both cases, the objectives are outlined, expectations measured, and risks and limitations are identified and communicated throughout the team and to the stakeholders. Ensuring open communication and clarified objectives keeps expectations in check, while better avoiding and resolving problems.

As previously mentioned, PTN was unsure about their initial objectives, which led to drastic changes in scope that led to historical data becoming unusable. But once the main objective was identified, we began adopting Agile practices and incremental updates along with open and clear communication. Throughout the current cohort, we have been able to deliver novel visualizations and reports, better understand change requests and new instructor needs, and drastically reduced our turnaround time for all asks. Using Agile has also ensured the Air Force better anticipates release cycles and what will be included with each one. This has enabled them to be better informed and ask more poignant questions about the work being done and how the data will be used.

Know the Data

Once you have received approval to work on a project, it is in both the data team and customer's best interest to be open with the data and resources available. Knowing what systems are available, the security requirements, ability to modify or change systems, and ability to add new sources is crucial to ensure necessary tasks are able to be performed and meet the customers' expectations. Not only should all parties know the tools and databases they are using, but they must also have an intimate knowledge of the data that is being provided and stored in those systems. If it is discovered that there are not enough data points being collected or shared, steps should be taken to ensure that these pieces of data are indeed needed, a risk/benefit analysis should be performed to determine the impact of missing data on final deliverables, and the cost and turnaround of adding these required sources needs to be determined. Not having all the necessary data components will greatly ebb the value of the data and limit the ability to perform meaningful analysis. It could make the data seem insignificant or worthless, questing the reasoning behind its collection.

Equally important to having access to these data is understanding the intricacies and context. If you do not know why certain outliers, combinations, or other pieces of the data stand out, they should be investigated. Similarly, if the stakeholders or other subject matter experts know why some data points appear out of sync with the rest, these should be well documented and made aware to anyone working with the data. For example, if you were investigating ways to improve the time it takes a student to master a subject, you may want to start by looking at how mu ch time students spend on learning modules. Let's say a group of students spends an unusually long time on introductory materials, and they frequently submit their tasks at the last minute. At first glance, these students may seem like a good group to work with, as they seem to have a harder time with the base material than the rest of their cohorts. However, when investigating the history of these students, you notice that they also tend to receive better-than-average grades and had no problems with in-person assignments. This was one such situation we had encountered with the previously mentioned Japanese after-school program. The issue had nothing to do with the students' poor performance, but rather was a result of using older computers with poor internet connection. Had we not known the context of these students' hardware issues, we would have erroneously chosen to have them be the subject of our study rather than a more suitable group.

Identify and Remove Biases

While most biases are unconscious and may appear without realization, it is still paramount that they be identified and mitigated. Bias towards or against data sources, points, methods, and other factors that should be taken into consideration with experimental design and execution can cause viable options to be overlooked and skew results. These biases can come from within the data team, from a lead, stakeholders, or other external factors. It is best to identify and remove biases as early as possible, ideally during the planning phase. When discussing strategy and planning the project, pay attention to when someone mentions the team should or should not take a certain course of action. While many may be perfectly reasonable (for example, not wanting using a certain tool because of cost restrictions), you should discuss and ask for elaboration on other tasks that you feel are a result of bias. In addition to the previously mentioned EHR example where the lead wanted to ignore clinics, other examples could include refusing to consider certain methods or tools because a team member doesn't like them, or ignoring a particular source. By discussing the reasons behind these, you could potentially reveal and help dispel biases or gain an understanding you can use for future tasks.

Fortunately, when done in the planning stage, biases can sort themselves out without deliberate intervention. Many biases came to be due to poor experiences with other projects, so by discussing the goals of a new project those with a bias could realize that those bad experiences were isolated to a previous engagement. Additionally, by exploring and familiarizing yourselves with the data and tasks, biases can be remediated by the context of the data, scope of the tasks, and ultimate goal or deliverable of the project.

You must also make sure you do not confuse a bias with legitimate learnings from someone's previous experience. If, for example, an employee cautions against using a certain tool with a particular dataset or platform, this could be due to the employee's prior attempts and learnings. While bias should be avoided, treating every argument or caution as bias can be just as damaging as letting biases slip in. Ultimately, clear communication and understanding will be your best tool against preventing bias while also ensuring legitimate risks are mitigated.

CONCLUSION

For centuries, we have relied on data to solve problems and develop new and exciting advancements. Over the past decade, data has been championed as one of the most important assets any enterprise can use, and being able to harness its capabilities is one of the most sought after skills in today's day and age. But despite this, without proper training and understanding, it is far too easy to be overwhelmed by these tools, and even easier to dismiss data as much ado about nothing. Throughout this paper, common problems that arise when working with data have been addressed with the hope that readers can understand what leads people into the belief that working with data is overrated and not worth the effort. Additionally, this paper provided resolutions and preventative measures to aid in helping novices avoid common problems and ensure that customers and stakeholders maintain and grow their confidence in data projects. We have demonstrated examples of where the Pilot Training Next program encountered these issues, as well as how they were able to resolve them, albeit at some cost. As more people become data conversant and use it to drive strategy and garner new and powerful insights, our hope is that these issues will be less common and easier to circumvent. But until then, we must perform due diligence. We hope this paper inspired you to not dismiss data when things do not go as they are expected and instead encourage you to ask more questions and dig deeper into your data stores.

REFERENCES

Bahramirad, S., Svachula, J., and Juna, J. (2014). Trusting the Data: ComEd's Journey to Embrace Analytics. *IEEE Power and Energy Magazine*, 12(2), 107-111.

Bloch, P.H. (2011). Product Design and Marketing: Reflections after Fifteen Years. *Journal of Product Innovation and Management*, 28(3), 378-380.

De Castro, M. (2016). Johann Gregor Mendel: paragon of experimental science. *Molecular genetics & genomics medicine*, *4*(1), 3-8. https://doi.org/10.1002/mgg3.199.

De Santis L., Scannapieco M., Catarci T. (2003). Trusting Data Quality in Cooperative Information Systems. In: Meersman R., Tari Z., Schmidt D.C. (eds) *On The Move to Meaningful Internet Systems 2003*: CoopIS, DOA, and ODBASE. OTM 2003. *Lecture Notes in Computer Science*, vol 2888. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-39964-3_23

Duhigg, C. (2012). *The Power of Habit: Why we do what we do in life and business*, New York, NY: Random House.

Ferdows, K. (2018). Keeping up with growing complexity of managing global operations. *International Journal of Operations & Production Management*, 38(2), 390-402.

Flowers, A. (2019). Data Scientist: A Hot Job that Pays Well. *Indeed Hiring Lab Occupation Spotlight*. Retrieved May 28, 2021, from https://www.hiringlab.org/2019/01/17/data-scientist-job-outlook/.

Johansson, J.J. and Abrahamsson, L. (2021). Digitalisation [sic] and Sustainable Work: obstacles and pathways. *European Journal of Workplace Innovation*, 6(2), 187-197.

Latin, H. (1993). *"Good" Warnings, Bad Products, and Cognitive Limitations*. *UCLA L. Rev. 41*, 1193.

Mikalf, P. and Krogstie. J. (2019). Investigating the Data Science Skill Gap: An Imprical Analysis. 2019 IEEE *Global Engineering Education Conference (EDUCON)*. 2019, 1275-1284. https://doi.org/10.1109/EDUCON.2019.8725066.

Millsap, R.E. and Everson, H.T. (1993). Methodology Review: Statistical Approaches for Assessing Measurement Bias. *Applied Psychological Measurement*, 17(4), 297-334.

Ng, E.S.W. and Law, A. (2014). Keeping up! Older Workers' Adaptation in the Workplace after Age 55. *Canadian Journal on Aging/La Revue canadienne du vieillissement*, 33(1), 1-14.

Ramachandran, K. and Watson, J. (2021). Tech looks to analytics skills to bolster its workforce: Addressing the analysis talent shortage. *Deloitte Insights*. Retrieved May 14, 2021, from https://www2.deloitte.com/us/en/insights/industry/technology/data-analytics-skills-shortage.html

Schirf, E. and Serapiglia, A. (2017). Identifying the Real Technology Skills Gap: A Qualitative Look Across Discplines. *Information Systems Education Journal (ISEDJ)*, 15(6), 72-82.

Stingler, S. M. (1990). *The History of Statistics: The Measurement of Uncertainty before 1900 (Reprint Edition)*, Cambridge, MA: The Belknap Press of Harvard University Press.

Thaler, R. H. and Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*, New York, NY: Penguin Group.

U.S. Bureau of Labor Statistics (BLS), (2021). *Occupational Outlook Handbook (Updated April 9, 2021)*. Retrieved May 14, 2021, from https://www.bls.gov/ooh/math/operations-research-analysts.htm.

Vallor, S. (2015). Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. *Philosophy & Technology*, 28, 107-124.

Weihs, C., Ickstadt, K. (2018). Data Science: the impact of statistics. *International Journal of Data Science and Analytics*, 6, 189–194. https://doi.org/10.1007/s41060-018-0102-5.

World Economic Forum (WEF) (2021). Upskilling for Shared Prosperity. *World Economic Forum Insight Reports*, January 2021.