Data is The Mission: A Guide to Data Science for Nonprofit Organizations

Sean Reidy

University Of Maryland Baltimore County

sreidy1@umbc.edu

DATA 606: Capstone in Data Science

Dr. Tony Diana

August 18th, 2022

Table of Contents

Abstract

Nonprofit organizations are eager to modernize into data-driven organizations but are disadvantaged by economic and cultural factors. The industry of nonprofit consultants supplying data science services disempowers nonprofits through a loss of autonomy. Through policy changes, Nonprofits should incorporate data as a core of the mission statement that can capture long-term value from data-driven decision-making. Sentiment analysis concludes that the opinion of data-driven nonprofits is generally positive, and there is clear evidence of an underserved demand for data science projects that appropriately match the scale of issues facing nonprofits. Nonprofits become data-driven by quantifying the goals of their mission statements, refocusing on small-scale data science that is feasible under limited resources, and investing in recruiting new qualified talent. In doing so, nonprofit organizations mitigate many large-scale data science projects' disempowering effects. A case study of the Easter Seals data lake project illustrates the benefits and limitations of ambitious data modernization. Evidence recommends a policy change for nonprofits to adopt a 'data first' culture that directly measures and tracks the impact of a mission-driven organization.

Overview of Nonprofit Organizations (NPOs)

Nonprofit organizations are decades late to the data science and big data hype. Eager to catch up, NPOs jump headfirst into data science initiatives, hoping that data-driven approaches will work like a magic treasure map, directing leadership to continuous growth and donations. These ambitions are cut short; NPOs are intimated and invalidated by their data. Recovering from the COVID-19 pandemic, the social, economic, and moral responsibilities NPOs possess in American society are impossible to overlook. Understanding that NPOs have a vital role in the function of many individuals and entire communities demonstrates how essential NPOs use limited recourses most optimally. Every year it is becoming challenging to argue that a decision is in the best interest of an NPOs core mission without supporting data. Despite this, a majority NPOs are ill-equipped to maximize the utility of the critical data they generate and often lack the capital and human resources to develop and maintain their tools and data team. “A generous estimate suggests that any form of advanced analysis is used at only 3% of the nonprofit organizations in the United States” (MacLaughlin, 2016). A lot has changed since that figure, and turn-key cloud services offer tools that NPOs could only dream of having access to a few short years ago. Still, this opening of a proverbial data science flood gate has its own more debilitating set of challenges that leaders must address.

What Makes NPOs Unique

According to the United Nations, NPOs must meet the following criteria: self-governing, not for profit & nonprofit distributing, institutionally separate from the government, and non-compulsory as in “that membership and contributions of time and money are not required or enforced by law or otherwise made a condition of citizenship”(United Nations, 2003).   This definition suggests that the nonprofit sector is rigidly defined, but the distinction between the nonprofit, public, and private sectors often blur. Organizations migrate from private to nonprofit; for example, a hospital network can change from public to nonprofit as government adjacency previously running them spins them off to be independent. Fueled by a continuous increase in demand for human services since the turn of the millennia and accelerated by the recent Covid-19 pandemic, the nonprofit sector continues to be a significant economic and social force globally. In America, an estimated 1.3 million charitable NPOs spend a combined 2 trillion dollars annually; expressly, NPO jobs represent 12.9% of private employment in Maryland. (National Council of Nonprofits, 2022) Without a doubt, NPOs have a substantial role in modern American society; every person in the U.S. has, at one point or another, directly benefited or contributed from an NPO. Reflecting on a personal experience volunteering or participating in a civil rights protest is easy to understand the good NPOs bring qualitatively. Quantifying that impact proves to be more challenging; understanding NPOs' role in society is limited by the patchy and often nonexistent data coverage. Measuring and quantifying NPOs' economic and social impacts is critical as the nonprofit sectors have become a primary focus of policy initiatives (Anheier, p.12 2014). On August 7th, 2022, The Build Back Better Act passed the Senate, which included 1.7 trillion in spending, including provisions for but not limited to: expanded child care, universal preschool, home care services, and climate change initiatives; all sectors were NPOs play a majority role (National Council of Nonprofits, 2022). Decisions of federal policymakers will have substantial implications for NPOs, and it is urgently vital that these organizations have the tools, talent, and tech to defend their existence and show evidence of their social good with data-focused decision-making.

Economic Challenges Facing NPOs

Our nation slowly emerged from the Covid-19 pandemic into economic peril. We are seeing rising inflation and warning signs of a pending recession. As of July 2022, the consumer price index rose by 8.5% (U.S. Bureau of Labor Statistics, 2022). Consequently, individuals have changed their spending priorities and allocating less disposable income for charitable donations. While rising inflation is an alarming worry for NPOs, individual contributions have been slowly declining for decades. Generous donations have remained stagnant at 2% of GDP since the 1970s. Perhaps more concerning is that the percentage of GDP and contributions from disposable income have remained constant for 40 years, as the number of nonprofits and the diversity of their missions has rapidly increased. In 1996, approximately 1 million NPOs registered with the IRS; this grew to 1.57 million by 2016. (MacLaughlin, p.10, 2016). Fundraising is broken; an increasing number of NPOs are all competing for the same limited and shrinking population of donors; this is not sustainable. Data-powered new donor identification and retention programs provide a solution for an individual NPO but also have unique challenges. While alarming, individual donations only represent a fraction of total funding; the remaining comes from service fees, government grants, and contract work. The previously mentioned revenue sources are considered more stable and dependable, but evidence indicates that even these are in decline too. Research by the Urban Institute in 2021 concluded that all revenue sources were under stress. In 2020, 40% of all nonprofit originations reported a loss of revenue, averaging a loss of 31% (Faulk et al., 2021). NPOs were already limited, and that pressure will only increase as we move forward into economic uncertainty. Any new ventures NPOs must embark on be within their means and vital to the mission.

The Problems with Nonprofit Data Science

A cliché statistic echoed in the data science field is a reminder that 90% of all the data in the world was created in the last two years (Dragland, 2013). Despite this statistic being almost a decade old, it continues to be a compelling reminder of how much data has changed how we all live our lives. However, the seemingly endless march for more and more data, NPOs feel left behind, unable to jump into the fast-moving currents of data, unwilling to risk, and unsure of the value data could bring to their organizations. It is not for lack of interest, as a staggering 97% of 2018 surveyed NPOs are interested in data science. Furthermore, 90% of respondents collected data, yet only 5% claim to have data-driven decisions making at the core of their operations (Every Action, 2018). NPOs are painfully aware of the growing importance of data science but lack the tools, knowledge, and personnel. NPOs cannot utilize data because they do not have the people or tools to use it, but they also cannot recruit the right people because they do not utilize data. There is a demand that the existing data science market is failing to meet. NPOs have faced unique challenges attempting to embark on tech innovation, an inherent risky endeavor. Among these challenges are the ethical issues when working with underserved populations, the nebulous criteria for what success means in charitable work, and the fear of media backlash and loss of funding if innovation efforts fail (Jaskyte, 2011). Against these odds, NPOs possess one strategic advantage over for-profit companies: they are flexible and adaptable because they are not profit-driven and do not have to be beholden to shareholders. By tailoring services to the unique needs of NPOs, there is a potentially enormous social and economic benefits to society. It is possible to break this ‘Catch-22’ that many NPOs find themselves in by adopting a policy that makes data part of the mission by focusing on the tools, talent, and tech.

Sentiment Analysis of New York Times Articles

Bad press from a failed innovation endeavor is one reason nonprofits are hesitant to lean too heavily on data science. News media can provide information on the attitudes and big picture sentiments around various topics. While reading a single article on nonprofits can only provide a single perspective, looking at an aggregate summary of articles will highlight connections and associations between NPOs and data. What are the average perceptions of NPOs in connection with data, and what nonprofit missions are most associated with data?

Data & Methodology. News articles were sources from the New York Times; all articles were published between January 2020 and August 2022. Using the NYT API, articles were scraped that met the following criteria: the words “Non Profit” and “Data Science” were present within the main body of the text, and the article metadata geotagged to the USA. A total of 1200 articles were downloaded. From the raw data, article headline, the publication date, abstract, and leading paragraph of each was saved into a CSV File. Two additional data sets of equal size were scraped, the first containing articles with just the ‘data science’ keyword, and the other with just the “Non Profit” keyword. This data was then cleaned and processed for natural language processing. The visualizations of the results can be found in the appendix.

Results. Assessing the summary statistics, the mean polarity of sentiment was 0.081, indicating a slightly positive attitude towards NPOs when associated with data science. The sentiment of articles filtered for the NPO topic exclusively exhibited a mean polarity slightly lower at 0.074. There is a slight edge that NPOs associated with data science are more positively viewed when compared to the average sentiment of nonprofit organizations. NPO fears of media ridicule from data failures may not be substantiated as, on average, associations with data science provide a slight bump in general sentiment.

“Data for Good” Initiatives are Flawed

A patchwork of volunteers, corporations, and educators donate their time and expertise under the umbrella of the “data for good” community. Projects that qualify as “data for good” follow a similar goal: skilled developers volunteer their time to deliver a data product for nonprofits at a free or subsidized cost (Hooker, 2018). Relying on volunteers causes problems as flashier cutting-edge projects are more likely to attract interest. Most data questions NPOs want to solve are not exciting; they are mundane, time-consuming, and would never make for exciting media coverage. Data for good programs exist on an expectation that NPOs are more empowered and supercharged by their newfound data skills. But if these volunteers complete all the work, the NPO loses the prospect of developing data skills. Still, this naïve but well-meaning optimism can provide nonprofits with a potent tool, but it raises questions about the longevity of these tools.

Large tech companies such as Google, Microsoft, and Amazon all have nonprofit services divisions dedicated to giving tools and cloud resources to NPOs at a free or discounted price. Amazon Web Services awards AWS Imagine Grants to nonprofit organizations seeking to transform and modernize their technology by migrating infrastructure to AWS cloud services. For the 2021-2022 fiscal year, Imagine Grant recipients included March of Dimes, Black Girls Code, and Easter Seals DC MD VA. Amazon’s goal for the grant is to improve the “…integrations with advanced cloud services, such as artificial intelligence (A.I.), machine learning (ML), high-performance computing, Internet of Things (IoT), and more” (Amazon Web Services, 2022). Amazon’s ambition for the grant program is well intended but wildly out of touch with the current needs of the vast majority of NPOs. Founder of the nonprofit Delta Analytics: Sara Hooker, emphasizes, “we forget that the vast majority of organizations still use

Excel and consider moving their data to Salesforce to be a big technical step. Providing cloud credits, hardware or expensive visualization licenses for free is useful to an extremely small group of organizations" (Hooker, 2018). These programs do not address the lack of technical training, as once a pro-bono big tech advisor completes the requested tools, the NPO now possesses advanced technology and instructions written in a foreign language and no internal staff to maintain it. At best corporate "data for good" helps a small number of top nonprofits. At worst, these programs are hollow P.R. tools to absolve big tech from a civic duty for more meaningful participation in charitable work.

Data Disempowerment

A whole industry exists around providing data tools to nonprofits. These services comprise a mix of the large tech giants, "data for good" volunteers, and nonprofit consultancy firms. NPOs hire these outside consultants to aid small internal teams or sometimes even entirely outsource data initiatives—these third parties are not always invested in the organization's corporate mission or long-term social impact goals. In a 2017 University of Colorado Boulder study, a survey of 13 mission-driven organizations and NPOs with data experience found that three unforeseen and negative consequences that are mutual reinforcement led to a cycle of increasing disempowerment. These consequences are The Erosion of Autonomy, Data Drift, and Data Fragmentation. (Bopp et al. 2017). The autonomy of nonprofit organizations is undermined by a variety of external stakeholders, whose biased opinions on what features are valuable shape data practices by prescribing the metrics that organizations should track. The information systems (owned by whom) that are used to collect data, and the formats in which such data should be reported. "The shifting of metrics and data collection foci in response to externally re-framed missions and priorities, moving the organization towards a mission that is both undefined and unknowable" (Bopp et al. 2017) . NPOs rely on outside expertise; these contractors come and go, this frequent stakeholder churn fragments the data, and the NPOs rely on the outsider knowledge to solve issues. These three create a sinister feedback loop where the loss of autonomy leads to data drift as outside funders introduce new goals that further the data fragmentation as data moves between systems. The lack of direction leads to a high internal turnover rate, leading to greater reliance on outside consultants for data work, thus perpetuating the cycle. It is not fair to blame the third-party consultants for disempowerment; if they did not believe in the mission of NPOs, then they would not have volunteered their expertise. Nonprofits who have neglected data for years have put in motion the ingredients that will eventually lead to disempowerment. As they join a project to assist, third-party stakeholders lack equity in the mission of NPO's and make decisions that unintentionally sabotage the ability of these organizations to take charge of their data independently. Truthfully, the well-meaning contractors have no incentive to do so; why help the NPO improve their numeracy when they could turn NPOs into recurring customers for a consultancy?

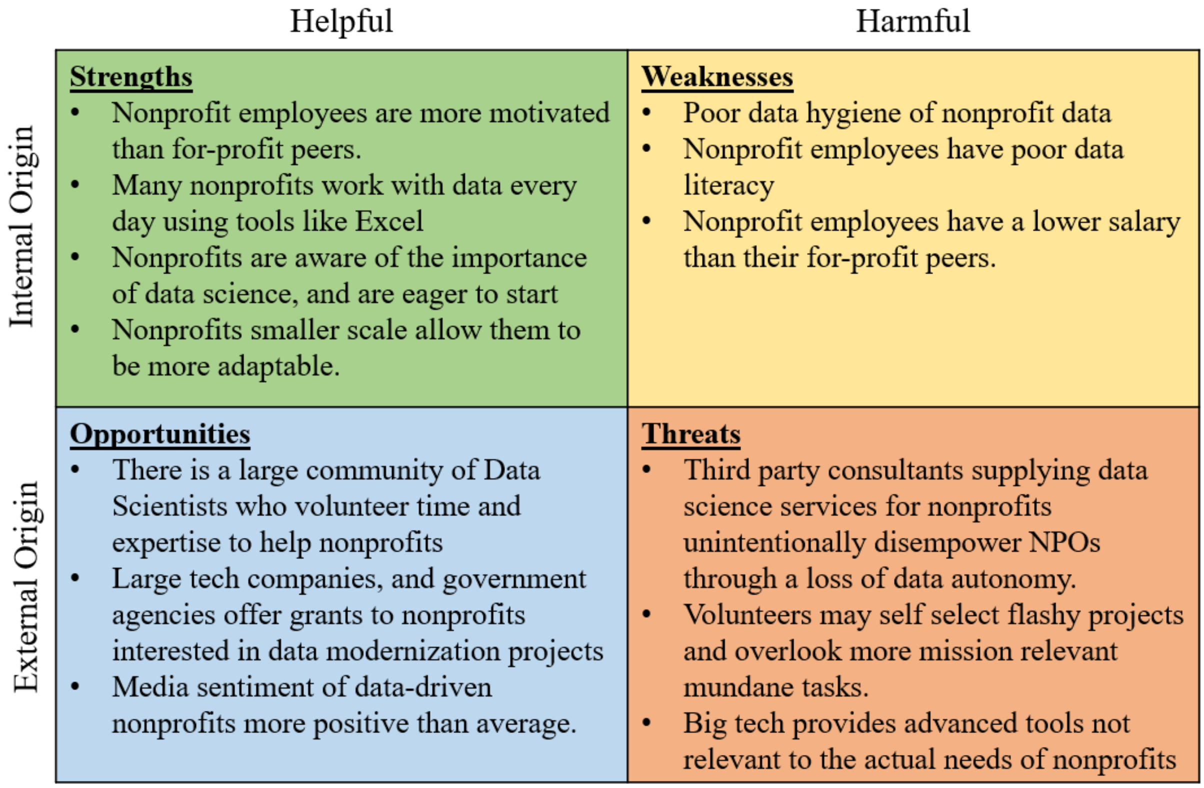
Recruiting Challenges

Access to modern technology does not guarantee that it is used correctly, if at all. Data Science is a tool, and having it does not instantly create value, much like owning a power saw, will not guarantee that individual becomes an expert craftsman. NPOs lack staff trained in data science and leaders who can help direct and advocate for data initiatives. In a competitive job market, NPOs find hiring employees skilled in data science and data engineering challenging enough for for-profit companies and doubly so for nonprofits. Staffing and high turnover disrupt data science programs. A 2021 survey by the National Council of Nonprofits found that "eight out of ten nonprofits responding to the survey identified salary competition as a factor preventing them from filling job openings" (National Council of Nonprofits, 2021). Nonprofits do not have the same capital resources and will not be able to compete dollar to dollar with for-profit companies. Employees at NPOs, on average, earned 4% to 8% less than their peers at for-profit companies, and for high-skilled jobs such as management the discrepancy was more significant at 17.8% (Kearney, 2018). Pay is not the only factor that influences hiring trends, as a 2008 survey of nonprofits found that while NPOs found it more challenging to acquire qualified candidates, NPO have a more significant advantage recruiting employees with higher than average worth ethic (Colins, 2008). Employees who actively look to work for NPOs choose to do so because they want to make a difference and are motivated by the mission-driven culture. These ace employees empower nonprofits to do more with fewer resources, which will be relevant when constructing an internal data team.

Data Hygiene

Due to the staff innumeracy, NPOs lack robust data compliance. Databases filled with errors and missing fields and with little to no schema will never provide long-term value, no matter how skilled a data science team is. Data health is at the heart of good data science, exemplified by the timeworn expression "garbage in, garbage out." Surprisingly, this turn of phrase is not assumed knowledge; NPOs are among the lowest ranking organizations in terms of data health. A Target Analytics study analyzed the quality of over a thousand NPO donor address records. Results found that "the average nonprofit was missing email address for 74% of their constituents. The worst are missing 96% of their email address. For the best nonprofits, 43% of their email addresses are missing." (Maclaughlin, 2018 p.28). Unfortunately, this purported 'data rot,' much like natural rust on the steel frame of a car, is exponentially easier to prevent than to treat once it has occurred. Once the data rot embeds itself, the sad truth is that it is more straightforward and cost-effective to toss out the existing data and start over. By doing so, NPOs potentially lose decades of untapped insights and an incalculable loss of potential.

Summary of Challenges



	Helpful	Harmful
Internal Origin	**Strengths** • Nonprofit employees are more motivated than for-profit peers. • Many nonprofits work with data every day using tools like Excel • Nonprofits are aware of the importance of data science, and are eager to start • Nonprofits smaller scale allow them to be more adaptable.	**Weaknesses** • Poor data hygiene of nonprofit data • Nonprofit employees have poor data literacy • Nonprofit employees have a lower salary than their for-profit peers.
External Origin	**Opportunities** • There is a large community of Data Scientists who volunteer time and expertise to help nonprofits • Large tech companies, and government agencies offer grants to nonprofits interested in data modernization projects • Media sentiment of data-driven nonprofits more positive than average.	**Threats** • Third party consultants supplying data science services for nonprofits unintentionally disempower NPOs through a loss of data autonomy. • Volunteers may self select flashy projects and overlook more mission relevant mundane tasks. • Big tech provides advanced tools not relevant to the actual needs of nonprofits

There are many stakeholders nonprofit organizations need to work around before engaging in discussions of data science policy. Our analysis finds that the greatest threat to any data initiative is external to the NPO. In order for success, the influence of third-party consultants who do not operate with the mission at heart should be kept to a minimum. NPOs can leverage their strengths and flexibility to tackle the weakness of poor data hygiene. While the state of data within nonprofits is overall quite bleak, there is substantial evidence to remain optimistic that with the right policy. There is hope that the average NPO can learn to harness the untapped potential of their data.

Theory of Change: Practical Data Science for NPOs

The existing market solution for nonprofit data science creates NPOs that are neither empowered nor equipped to think and plan for the long term, despite our communities' needs for such strategic planning. The crux is an absence of policy that focuses on data as core to the mission statement. It is not simply a lack of caring but instead not having a voice in the Mission of NPOs. Across thousands of NPOs is the same narrative: resource-limited and data illiterate staff use their worth ethic advantage to put forth a valiant effort to navigate a complicated web of actors that perpetuate a cacophony of data demands. Conflicting interests draw the staff's focus away from the organization. If data staff picture data as a secondary product, it will always be viewed as something that detracts from the mission. For most small NPO, cold switching to a data-driven operation is impossible without significant policy changes. The mission itself must be adapted to focus on quantitative data-driven goals.

Advanced data science and big data are useless average NPO. These projects carry too much risk, and the project scope is simply impractical and out of touch with the humble data needs of nonprofits. NPOs must keep the project tight, focused, and outsource as little of the work to third parties to maintain autonomy. So, start small, maybe a database or two, and grow naturally and methodically. Nonprofits who want to invest in their longevity and truly become data-driven must use policy to recontextualize the mission stament to be measurable with data. Conventionally a good mission statement answers the who, what, and why in a single statement. Now we must add 'by how much?" and quantify the good nonprofits accomplish.

Value, not Volume. Most NPOs should focus on the value of the data above all else. A commonly used idiom in data science to describe the pillars of exemplary data projects are the five V's of big data: volume, velocity, variety, variability, and value. (Jain, 2016). Big data in

2022 is a well-established discipline, and the novelty of the vastness of massive databases has worn off. While large databases are compelling, NPOs are not ready for big data's scale and complexity. What good are petabytes of data if there is no question to be answered? That is why nonprofits need a solid research hypothesis and a knowledgeable leader who can oversee data operations and, more importantly, hire talented data scientists. Before organizations can begin the process of data analysis, they must first frame a research question, learn what data is needed to answer that question, promote a data first culture ensuring organizational buy in, and invest in talent (Azam, 2022).

A data-focused mission statement will not produce results if nonprofits do not invest in hiring qualified staff. Nonprofits can overcome their hiring disadvantages by looking for talented employees whom well-established for-profit companies often overlook. Hiring managers can look for data scientists early in their professional careers who may lack years of experience, who may be willing to accept a lower-paid position in return for greater flexibility and the potential to work their talents for a cause they are enthusiastic about. In addition to hiring new staff, existing employees exhibit strong work ethics and providing training in data literacy would be a valuable investment.

Case Study: Easter Seals DC MD VA

In 2021, the NPO Easter Seals DC MD VA worked to modernize the organization’s data infrastructure. The Easter Seals is a nationally recognized charity that provides various disability services. Including but not limited to “early intervention programs, inclusive childcare, medical rehabilitation, and autism services for young children and their families; job training and coaching, employment placement, and transportation services for adults with disabilities, including veterans; adult day programs and employment opportunities for older adults” (Easter

Seals, 2022). The key to this accomplishment is a two-sided approach that considers both the technical and cultural ideals surrounding data and its role at Easter Seals. Eater Seals' previous data infrastructures were fragmented, where each team within the organization used and maintained their isolated systems. There was little interoperability between these tools, requiring a lot of manual labor and data cleaning when sharing information with others in the organization. The Amazon Imagine Grant provided funding for this effort. To remedy the fragmentation, the Easter Seals worked with Amazon Web Services to develop a full-stack data lake that would ingest various data. On top of this data lake, AWS Quicksight dashboards visualized data in real-time, such as student achievement scores.

Summary of Challenges

- Nonprofit organizations lack stable financial resources to develop and implement a modern data pipeline.

- Nonprofit organizations' work scope and topics do not align with the needs of most big data customers, hard to compare to example use-cases.

- Organizations need to have an intrinsic perspective on their data before gaining real operational insights from data. More colloquially: you must know what they are looking for before you can find it.

- Leaders struggle to get empirical evidence of how the organization is performing. For example, at Easter Seals, it took a week of work to get an answer to a question as seemingly straightforward as: How many students did we have enrolled in a given timeframe?

- Data cannot be an obstacle that gets in the way of the mission goals of nonprofits. "Nonprofits need to be able to answer urgent questions like, "Who else is working on homelessness in my town?" or "Has anyone else ever tried this approach for reducing teen pregnancy?" or "What do the people in our job training program think of our work?" The lack of data makes day-to-day tactical decisions hard and long-term planning practically impossible. And who pays the price? The people, communities, and ecosystems that nonprofits serve." (Harold, 2013).

- Front-end employees who interact and log data have years of experience with existing tools, and organizations cannot afford to retrain fully.

- Existing legacy tools run on antiquated systems and lack API or SQL access, making them challenging to integrate historical Data into modern big data storage.

- Historical data is full of errors and typos, primarily due to a lack of robust data compliance.

The Easter Seals dealt with the two most challenging issues when starting a data initiative, innumeracy and poor data hygiene. Easter Seals lacked trained staff in data science and had only one intern on the project with a technical background in the field. The lack of internal data science knowledge led the Easter Seals to rely heavily on outside contractors like AWS and Deloitte to perform the programming and heavy lifting. Looking at the data itself, it was messy, with little to no standardization. For example, many conflicting and overlapping categories for ethnicity were used as the data was hand-entered by a teacher in a classroom.

What is a Data Lake?

Big data technologies enable much greater scalability and cost-efficiency than can be achieved with traditional data management infrastructure. When evaluating any Big Data system, we judge it based on the three V's of Big Data: Volume, Variety, and Velocity. When an organization, such as the Easter Seals, has a broad scope of services, all operating independently from one another, the data created has many tools, formats, and needs. The Easter Seals is not a fast-moving company, nor does it generate multiple terabytes of data. The Easter Seals need greater flexibility with their sheer variety of data generated by the enormous scope of services it provides to the MD/DC/VA region. A Data Lake fits the flexibly requirements while remaining at a somewhat manageable scale. The goal of a data lake is to preserve data in a raw format as it was initially received, with as minimal processing as possible. This raw data rested alongside any transformed data (visualizations, dashboards, regression models, etc.). A similar tool used for long-term data archiving is the Data Warehouse, but the Data Lake differs because it utilizes schema on read. A Data Lake is more flexible and easily scalable to Big Data-sized pools than a warehouse's predefined schema. Data lakes allow interoperability between data scores lowering the barrier of entry for data science. Self-service has replaced the labor-intensive and carefully crafted approaches that are currently used. "The data lake is a daring new approach that harnesses the power of big data technology and marries it with the agility of self-service. Most large enterprises today either have deployed or are in the process of deploying data lakes" (Gorelik, 2019). Data Lakes are flexible to the applications of many different parts of the organization, utilizing a schema on read approach where data is recontextualized into a format and file that is most relevant to a given user. Consequently, data lakes collect information from operational sources "as is", often without any upfront analysis. (Lemahieu et al, 2018). Data Lakes are ideal for nonprofits looking for a modern tool to catalog and store data in an agile system, allowing for diverse data from across the organization to be integrated, stored, and used immediately to assist in charitable goals.

Technical Overview of Easter Seals Prototype Data Lake

The Data Lake deployed at The Easter Seals was always intended to be a scalable proof of concept; whereas the data needs of the organization grew, new data sources could easily be added to the document store. The team found it essential to initially implement the data lake as a supplement to existing software to not disrupt the operations of the various child development centers. The goal was to ingest various primary data sources without interrupting or changing the functionality of the existing software.

Data Sources: Procare: A Microsoft SQL-based local server-side application that runs the day-to-day operations of Child Development Centers, including but not limited to class rosters, child disability surveys, and billing T.S. Goal: A online tool for lesson planning and tracking student progress. External Surveys: various surveys run by Easter Seals staff can vary to any topic. External Relationships Tracking: Various Excel documents and CVS files log the involvements and relationships of donors, regional partners, and universities.

Data Pipeline: The Data Lake was built using Amazon Web Services Cloud service, Utilizing the S3 data and document storage system. Within S3 data, copies of data will be kept in one of three "buckets" Raw, Cleansed, and Curated. When new data from local sources is synced to the Data Lake, that data is uploaded to the Raw bucket, where preliminary ETL processes begin utilizing AWS Lambda functions. These functions work similar to map reduce functions found in Hadoop and similar distributed computing systems. After preliminary ETL cleaned data is stored in the Cleansed S3 bucket. Directing the whole data lake is Amazon Athena, who infers schema during read, and performs analytics on data, saving that data into the Curated bucket. Easter Seals has limited staff and cannot have a dedicated team working on enforcing strict schemas, so Amazon Athena works well, freeing up resources "Amazon Athena make it easy to run interactive queries against data directly in Amazon S3 without worrying about formatting data or managing infrastructure." (Amazon Web Services, 2021) . Tracking all the data stored in the lake is the Data catalog, which serves as the central metadata repo, logging changes and location of all data in each S3 bucket. The final but arguably most important for proving the value of data to Easter Seals is Amazon Quicksight, a real-time dashboard tool that uses infrared schemas from curated data and can quickly show real-time visualization of data.

All these systems working together attempted to solve the most significant problem at the Easter Seals: people who need information don't know who has that data, and where to get that data. With the S3 and Amazon Athena based Data Lake, all information, including class rosters, student health records, fundraising goals and surveys, will be in a secure and easily accessible cloud-based tool. This dramatically reduced the time needed when measuring the progress of each child development center, as that data will be queried in real time, as compared to taking a week or more to generate a handwritten report.

Lessons Learned

The nonprofit sector is trapped in slow motion, unable to take risks on large, expensive data science projects with abstract benefits. How can an NPO genuinely understand the benefits if the organization does not view its data as a product to aid in its mission? Reflecting on the case of the Easter Seals Data Lake project, data literacy among employees and executives is poor. The literacy issues are no fault of the individuals employed by the Easter Seals; data has never played a pivotal role in their day-to-day responsibilities but rather as a tool or a means to an end. Easter Seals experiences many issues, including missing data or unreported expenses, originating from communication breakdowns between different departments. In other words, data was never considered in the mission, and the project has little to no organizational buy-in as crucial to the future of the Easter Seals.

From the onset of the data lake project, no straightforward research question guided the team's efforts. The project went forward to put all existing data into the new data lake but lacked focus on what value this could bring the Easter Seals beyond just migrating the data to the cloud. It was never clear what the obvious value a data lake brought to teachers and bookkeepers. Raising the question if Amazon considered whether or not supplying a data lake and code support was really in the best interest of the long-term ability of the Easter Seals to build an interval data science team. A data lake was one of many possible solutions and would have the capacity to scale as the organization grows, but it is unlikely that the Easter Seals will grow at the rate that is best suited to the rapid scalability of a data lake. Unfortunately, the Easter Seals data lake project was exhibiting the common symptom of over-ambitious projects that eventually led to the following: Disempowerment cycle: erosion of autonomy, data drift, and data fragmentation. (Bopp et al. 2017). Perhaps the Easter Seals would have been better served by migrating data out of the closed-source ProCare and into an open-source Postgres SQL database running on top of AWS. This has the extra advantage of not being a proprietary Amazon technology and frees the Easter Seals from some third-party influence while still utilizing the cloud resources provided by the Imagine grant.

One of the project's successes was using AWS Quicksight to reveal metrics and stats that the organization had never seen in real-time. These tools have a lower barrier of entry, easily overcoming issues of employee innumeracy, so that staff can use their nonprofit expertise to extract value in the data without waiting on a data report from a different department. The most effective way is to use real-time dashboards, which can give immediate information about operations without a week of lead time for an overworked staff member to make an excel spreadsheet to find the same information. The dashboard allows immediate value to be extracted from underutilized data—looking into the future. The Data Lake allows for all data to be stored in a controlled and organized way. When future technology allows, previously cost-prohibitive tools such as modeling and data forecasting could be used.

Policy Recommendations for Data Success

Recent public policy, most notably Build Back Better, will supercharge the nonprofit sector. This is a once in a generation opportunity for NPOs to break out of the past and enter the modern world of data. To prepare for this, Nonprofits must begin to implement their own policies and put in place the mechanisms for practical and obtainable data science.

1. Recontextualize the mission statement to include quantitative measurement of the desired impact of the organization's work. Promoting data as a core pillar of the organization's culture will ensure greater organization by in, improving morale.
2. Invest in hiring talented and motivated staff skilled at data science. Hiring managers can circumvent direct competition with large private companies by looking for less experienced but equally qualified candidates.
3. Develop a data science road map and evaluate at what stage the organization is currently at. Even the most modest of goals can be achievable and provide long-term value from the wisdom within the data.
4. Keep data science projects internal to gain operational learning from years of trial and error. Data is a long-term investment, and every year the wisdom compounds like interest on a loan.
5. Utilize free and reduced-price SAS (Software as a Service) platforms to lower the capital requirements for data hardware but avoid using any cutting-edge tools

Conclusions & Future Work

Data is intimate to the organization it originates from, and it should be treated as a valuable asset to be protected and invested in. When NPOs outsource data science to consultants, they lose out on the valuable educational experience. In order to know what questions to ask, what research to pursue, and what data to grow, an organization must initially possess knowledge of that data. Being data-driven does not remove the need for the ever-so-important human element core to the compassion many vital to the identity of NPOs. People will always be at the core of the work NPOs do, and data will never fully replace this. This is why it is important to view data as equal to the human element and not a superior law that can magically fix all the issues and challenges NPOs face.

Further research should be done on identifying NPOs with small internal data teams. Given more time, I would develop a survey to administer to a pool of Maryland-based nonprofits asking questions about the scope of their data ambitions. Inquire if the organizations have planned or have partnered with a third-party consultant to assist or fully outsource data analysis and their experiences with the data disempowerment cycle.

An additional case study about a successful data science project at an average-sized NPO would also be helpful evidence for the policy of medium data. Reviewing the steps this organization took would provide a valuable guide and best practice for how to run a data science project. A potential future project would be to scrape the actual mission statements and various press release writings of a sample of nonprofits. Then using NLP topic recognition to classify NPOs as data-focused or not data-focused. This could be used to develop a tool to help individuals looking to donate to a charity to identify what organizations are more likely to be data-driven.

I have gained valuable experience by writing this report and breaking out of my comfort zone. I am confident saying that my time working on this capstone has ignited a personal passion for public policy that I was previously unaware of. I want to thank both the instructors and peers of the UMBC Data Science program for their assistance in helping me become a more well-rounded data professional.

References

Amazon Web Services. (2022). *AWS IMAGINE Grant*. Amazon Web Services, Inc. https://aws.amazon.com/government-education/nonprofits/aws-imagine-grant-program/

Amazon Web Services. (2021). *Amazon Athena FAQ's*

Anheier, H. K. (2014). *Nonprofit Organizations: Theory, Management, Policy* (2nd ed.). Routledge.

Azam, H. (2022, January 5). 14 Tips for Nonprofits Working with Data. Medium.

Bopp, C., Harmon, E., & Voida, A. (2017). Disempowered by Data: Nonprofits, Social Enterprises, and the Consequences of Data-Driven Work. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3608–3619. https://doi.org/10.1145/3025453.3025694

Collins, B. K. (2008). What's the Problem in Public Sector Workforce Recruitment? A Multi-Sector Comparative Analysis of Managerial Perceptions. *International Journal of Public Administration*, *31*(14), 1592–1608. https://doi.org/10.1080/01900690802434214

Dragland, Å. (2013, May 22). *Big Data – for better or worse*. SINTEF. https://www.sintef.no/en/latest-news/2013/big-data-for-better-or-worse/

Easter Seals. (2022). *Frequently Asked Questions*. https://www.easterseals.com/who-we-are/faqs/

Gorelik, A. (2019). The Enterprise Big Data Lake: Delivering the Promise of Big Data and Data Science (1st ed.). O'Reilly Media.

Harold, J. (2013). Nonprofits: Master "medium data" before tackling big data. Harvard Business Review Blog Network.

Hooker, S. (2018, July 25). *Why "data for good" lacks precision.* Medium. https://towardsdatascience.com/why-data-for-good-lacks-precision-87fb48e341f1

Jain, A., MD. (2022, June 30). *The 5 V's of big data*. IMB Watson Health Perspectives. https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/

Jaskyte, K. (2011). Predictors of Administrative and Technological Innovations in Nonprofit Organizations. *Public Administration Review*, *71*(1), 77–86. https://doi.org/10.1111/j.1540-6210.2010.02308.x

Kearney, C. (2022, June 23). *The Price of Doing Good: Measuring the Nonprofit Pay Cut - Payscale*. Payscale. https://www.payscale.com/research-and-insights/nonprofit-pay-cut/

Lemahieu, W., Broucke, V. S., & Baesens, B. (2018). Principles of Database Management: The Practical Guide to Storing, Managing and Analyzing Big and Small Data (1st ed.). Cambridge University Press.

MacLaughlin, S. (2016). *Data Driven Nonprofits*. Saltire Press.

National Council of Nonprofits. (2021, December). *The Scope and Impact of Nonprfoit Workforce Shortages*.

National Council of Nonprofits. (2022). *Economic Impact*. https://www.councilofnonprofits.org/economic-impact

New York Times. (2022). *New York Times Arrticle Search API* [Look up NYTs articles by keyword]. https://developer.nytimes.com/

United Nations. Statistical Division, Nations Unies. Division de statistique, United Nations. Statistical Division, & United Nations. (2003). *Handbook on Non-profit Institutions in the System of National Accounts*. United Nations.

Appendices

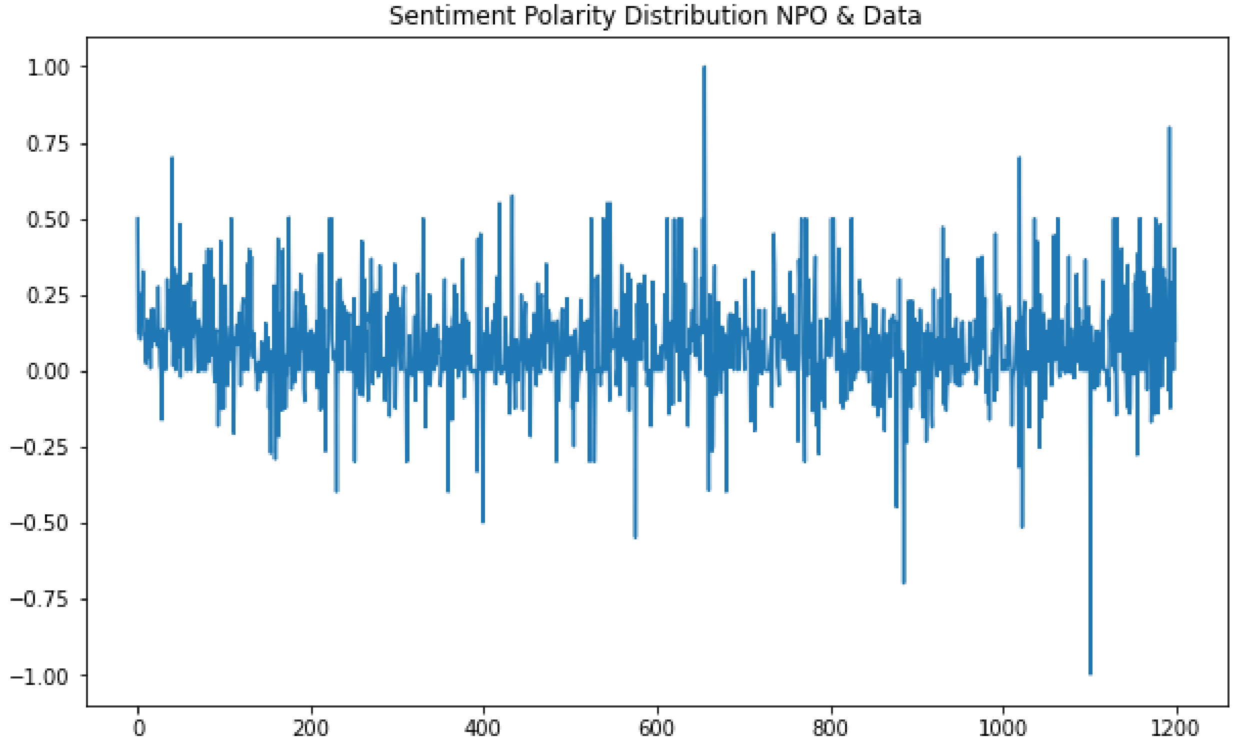


Figure 1: Sentiment polarity plot from January 2020 to August 2022 of New York Times articles containing both the keywords "Non Profit Organization" and "Data Science"

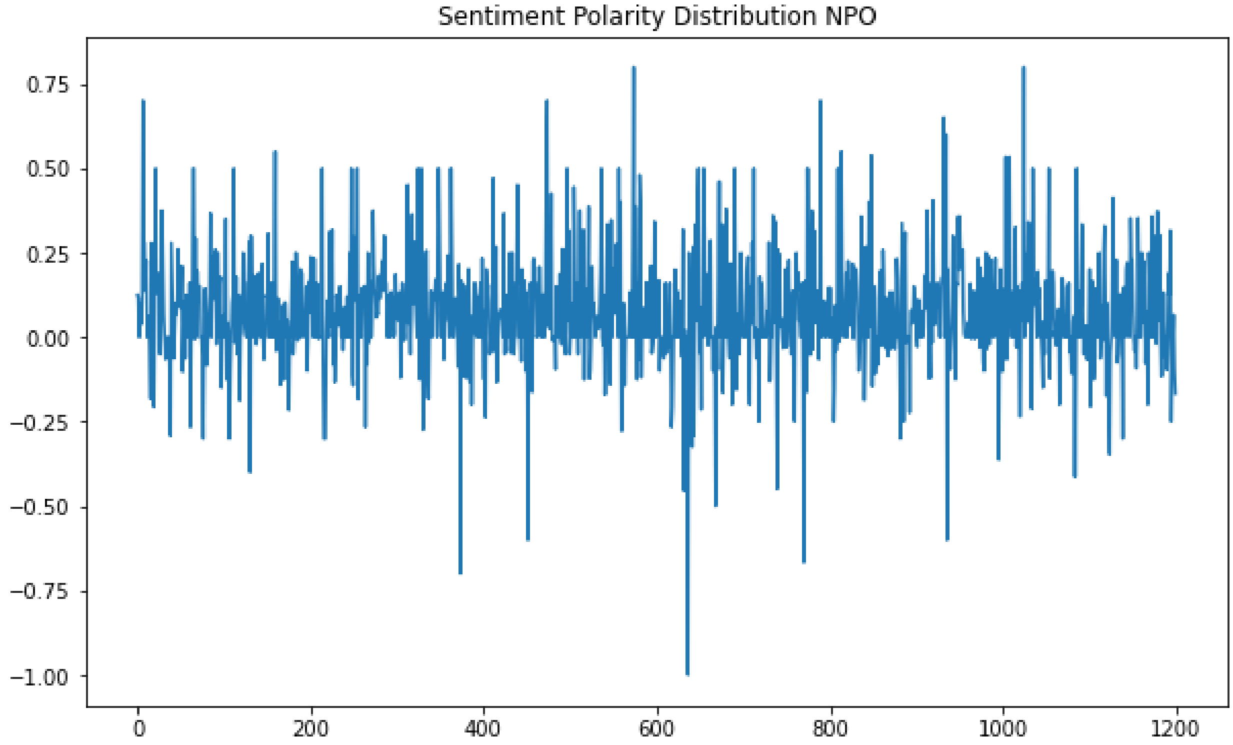


Figure 2: Sentiment polarity plot from January 2020 to August 2022 of New York Times articles containing only the keywords "Non Profit Organization"