1. **Introduction** — Data Mining

The National Institute of Standards and Technology (NIST) defines data mining as, "An analytical process that attempts to find correlations or patterns in large data sets for the purpose of data or knowledge discovery." (Gallagher. 2013) Essentially, data mining enables an organization to transform raw data into valuable information.

The process of data mining can be parsed into five steps. First–An organization must compile data, and then load it into their data warehouse. Second–The organization stores and manages the data they have complied. By loading the data into the data warehouse, the organization's data analyst(s), management team(s) and IT professional(s) can access the data. Third–The organization runs the data through their application software and the data is sorted based on the organization's criteria. Lastly–The organization's data analyst(s), management team(s) and/or IT professional(s) format the information obtained in an easy-to-share format, such as a table or graph. (Twin. 2020)

Data mining facilitates an organizations ability to obtain valuable information regarding customer's behavior. This information can be used to better-develop an effective business model. Information derived from data mining is critical to an organization's success. This is because of the improved strategic decisions that can be inferred from this information. Organizations that possess the personnel too mine data have a considerable competitive advantage over organizations that cannot.

The value of information generated from data mining heavily relies on the data's quality.

1.1. Data Quality

According to the textbook, *Data Mining Concepts and Techniques*, states, "Today's real-word databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogenous sources." (Han, Kamber, & Pei. 2012) This illustrates why quality data is so important, especially when mining data. Low-quality data will generate low-quality mining results; and high-quality data will generate high-quality mining results. Therefore, emphasizing high-quality data practices will enable an organization to obtain more valuable information from data mining.

Analytics India Magazine identified six major data quality issues that pose obscure obstacles for many organizations: poor data quality, duplicate data entries, incomplete data, inconsistent data formats, data's accessibility, system upgrades and data purging & storage. (Hebbar. 2019)

1.1.1. *Duplicate Entries*

Storing the same information in multiple places can generate a negative impact on computation power and storage; and can lead to negligent decisions based off misinformation. The more susceptible an organization's data entry method is to human error, the more likely a database is to contain duplicate entries. A way to avoid this from occurring is data deduplication. This is a process that combines human insight, data processing software, and algorithms to identify and eliminate probable duplicate records.

1.1.2. *Incomplete Data*

Data can be incomplete for several reasons, such as simply missing data, censored data and truncated data. An example of incomplete data is when a data set containing addresses is missing the zip code. This significantly decreases this data set's value because determining a geographical location without a zip code is very difficult.

1.1.3. *Inconsistent Formats*

Formatting discrepancies are a common problem that decreases the quality of a data set. If an organization's data formatting is not uniform throughout their data sets, it can require a ton of time for a computer to process. When formatting issues are severe, it can be impossible to mine and/or process queries from the data. To avoid issues regarding formatting, an organization can enforce a strict formatting method for storing generic data (first name, last name, date of birth, etc.).

1.1.4. *Accessibility*

Protecting data is a fundamental practice of every organization. The key problem is finding the fine line of not too much protection, but also not limited protection. Granting each employee with an appropriate level of access can aid an organization's data usage. As a result, data will be available when needed. Identifying *who* needs access to *what* data is vital for an organization.

1.1.5. *System Upgrades*

Every system requires updates. Some more than other. Regardless, backing up data before updating a system is crucial. Anytime an update occurs, there is a chance the data can become corrupted or destroyed. Practicing regular backups on all data can help mitigate this issue.

1.1.6. Data Purging & Storage

Within every organization, there is a possibility of their data being deleted—on accident or a malicious attack. Therefore, ensuring data is stored in a secure yet accessible manner is significant. Analytics India Magazine contributed an example— "As business users grow frustrated that they cannot get answers when they need them, they may give up waiting and revert to flying blind without data." (Hebbar. 2019) Therefore, implementing an effective storage method that is secure and accessible is essential.

2. The Impact of Low-Quality Data

Organizations are becoming more reliable on data. "Corporate leaders consider data and analytics capability a top investment priority." (Forbes Staff. 2017) Therefore, to maximize the value of information derived from data, organizations must enforce high-quality data practices. This seems fundamental. However, according to the Forbes Insights and KPMG 2016 Global CEO Outlook, "84% of CEOs are concerned about the quality of the data they are basing their decisions on." (Redman. 2017) Data's role enables an organization to identify and capitalize on opportunities, however, this role will become considerably diminished if organizations do not prioritize high-quality data practices.

Formulating strategic decisions via low-quality data can result in costly consequences.

2.1. Low-Quality Data Costs

Opportunity cost, in simple terms, is the loss of a potential gain from choosing other alternatives. A great example of how *costly* an opportunity cost can be, comes from the early stages of developing the renown tech company, Apple Inc. The short version of this example: Apple's third co-founder sold his 10% share of Apple for $800 in 1976. Apple's 2019 market value was appraised at roughly $1 trillion. Therefore, Apple's third co-found suffered an *opportunity cost* of roughly $10 billion. (Martin. 2018) This example stresses the significance of an organization's ability to identify and capitalize opportunities through high-quality data. If Apple's third co-founder had access to data that illustrated how successful Apple would become, he could have avoided this massive opportunity cost.

Forbes identified three major costs generated from low-quality data: reputational damage, lost revenue and missed opportunities. (Forbes Staff. 2017)

2.1.1. Reputational Damage

Low-quality data's impact can affect more than simply monetary components. For example, contacting the same user(s) multiple times with the repetitive/irrelevant information can damage an organization's reputation. Often, an organization's reputation can be tarnished with far less effort and resources than required to restore a reputation.

On a more impactful level, low-quality data in the banking industry can result in unintentionally engaging in trade with sanctioned governments and/or suspected terrorist capitalist. This can happen when a banking institution lacks adequate data regarding their business partners. As a result, the organization will endure a major reputational hit; and will require a lot of time and many resources to repair.

2.1.2. Lost Revenue

IBM estimates that mitigating low-quality data effects costs roughly $3.1 trillion a year, in the U.S. alone. (Redman. 2017) This is because organizational decision-makers deduce strategic decisions from data alone. Strategic decisions are made with the intention of achieving a substantial gain. When these decisions are based off low-quality data, they can do the opposite—produce a substantial cost.

Thomas C. Redman, also known as, "the Data Doc," suggests another reason for low-quality data being so costly. He believes the low-quality data costs so much is due to the number of individuals who need it for their everyday work tasks. "The data they need has plenty of errors, and in the face of a critical deadline, many individuals simply make correlations themselves to complete the tasks at hand." (Redman. 2017) Correlations inferred from incomplete data can lead to expensive consequences.

2.1.3. Missed Opportunities

As briefly covered, missed opportunities can result in colossal opportunity costs. The example involving Apple Inc. is a bit extreme. Nonetheless, low-quality data can result in a decreased ability to identify opportunities. An organization can fail to identify a significant opportunity regarding product development, customer service, strategic decision-making, etc. Failing to realize an advantageous opportunity can result in a competitor with a more thorough understanding of data exploiting that same opportunity. As a result, the competitor can increase upon their competitive advantage.

2.2. Data Mining's Future

The effectiveness of data mining heavily relies on the quality of data. With the increasing rate at which technology advancements are forthcoming, the possibilities regarding data usage seem unlimited. However, with the amount of low-quality data circulating throughout the world—More $3 trillion spent yearly in the U.S. to mitigate low-quality data—it is difficult to envision a sensible future for data.

Individual organizations must ensure they are enforcing high-quality data practices. Organizations that fail to accomplish this will endure increased fees overtime mitigating their low-quality data.

As data will only become more integrated within organizations, the cost to mitigate the adverse effect of low-quality data will continue to rise, but at what cost? Will it be feasible for U.S. organizations to spend $10 trillion in ten-fifteen years? Maybe more importantly, will it be responsible? At what point will the cost of recouping low-quality data cost more than its worth? Organizations must enforce high-quality practices and decrease the amount of low-quality data that is circulating.

3. Controls to Mitigate the Impact of Low-Quality Data

Low-quality data is the source of so many complex issues within data mining. We have laid out an action plan that can be used to help mitigate the effects of low-quality data. If an action plan is not initiated, organizations ability to benefit from data will decrease. While these recommendations are not full proof, they will contribute to eliminating low-quality data. As a result, organizations can save time and resources.

3.1. Data Duplication

The first control that can limit low-quality data is detecting duplicate data entries. Duplicated data entries are something an organization will deal with if they use multiple information systems that store the same type of information. (IFP) This can increase the difficulty of running queries because there is twice the amount of information. This can also lead to an organization drawing conclusions on inaccurate data. The use of data duplication tools like Druva inSync, and Barracuda Backup provide methods to prevent data duplications from happening and protect an organizations data. (TrustRadius)

3.2. Formatting Issues

Formatting controls will also prevent low-quality data. For example, say an organization is merging two databases that contain people's home address. If the format of the address column in one data source contains the full name for "Road," "Street," "Avenue," "Drive," etc. And then in the other data source, the formatting of the address column contains abbreviations for "Rd.," "St.," "Ave.," "Dr.," etc. This can cause the database to treat the same home address as two different pieces of information. As a result, organizations can be led on to make a decision that may not be best suited for them. Organizations should implement a universal data format within their organization.

This will encourage all organizational data to be structured similarly. As a result, quality of data will be consistent, precise and readable.

A quick real-life example—With the COVID-19 situation going on, there is a ton of data circulating regarding the virus. I have seen many data charts, from many different sources, and they all have a disclaimer essentially saying, "All of these numbers could be inaccurate due to the number of tests administered being difficult to quantify." The Center for Disease & Control (CDC) should have advised a standard data format that is required to abide by for a data set regarding COVID-19 to be considered credible. This would have aided the ability of combining many sets of data and drawing more accurate conclusions. This data format could be fundamental, so anyone could contribute. This would have enabled the CDC to process much more data.

3.3. Artificial Intelligence

Another control that will be important withing the future of data mining is the use of artificial intelligence (A.I.). With technology expanding in our society, the rise of A.I. has been a major contributor to developing strategic decisions. Human errors are a huge issue within data that can be difficult to identify. Such as a "0" being in the place of an "O." Using artificial intelligence can mitigate the effect of human errors. Artificial intelligence can identify the source of an issue and eliminate it from reoccurring. While it can be an expensive control; it will undoubtedly assist in eliminating and avoiding low-quality data.

3.4. Increase Quality Assurance

According to the Harvard Business Review Article, "If Your Data is Bad, Your Machine Tools are Useless." (Redman. 2018) Thomas C. Redman, the Data Doctor, states that quality assurance is the process of ensuring that an organization's quality program consistently provides desired results. Meaning, that quality assurance is needed from a different department within an organization to assure that high-quality data practices are being enforced. This allows organizations to obtain feedback on their data practices and continuously improve their high-quality data habits.

3.5. Charge Responsibility

Another control that can prevent low-quality data is putting someone in charge of maintaining quality of data, such as a Database Administrator (DBA). If no one is responsible for maintaining an organizations data; then data accountability issues will inevitably arise. (Redman 2018) To prevent this happening, an organization must identify or hire an individual with a high level of database expertise. By establishing responsibility within data practices, accountability can be held for data mishaps. As a result, organizations can identify where their low-quality data practices occur.

3.6. Conclusion

In conclusion, data mining is a process that can enable an organization to improve strategic decision-making, and better develop a successful business model. Organizations are integrating data within their business functions at an increasing rate. This is because high-quality data is a great source to base decisions on. High-quality data facilitates the ability to identify patterns. From these patters strategic decisions can be made that result in desired results.

Strategic decisions that are based on low-quality data can be extremely costly too an organization.

Data mining enables an organization too visualize data. Meaning, it allows them to see *what* the data is saying. As a result, organizations can implement strategic decisions that increase several productivity measures. The controls recommended were developed around the ideas of storing only necessary data and enforcing accountability for low-quality data practices. By eliminating unnecessary and duplicate records, organizations will not waste time sorting through futile data records. By establishing accountability, organizations can reprimand individuals that are sourcing low-quality data practices. As a result, organizations will improve their data management practices; and can eliminate the source of their low-quality data.

With improved data management practices and less low-quality data, organizations information derived from data mining will be much more valuable.

4. References

4.1. Gallagher, Patrick D. “Security and Privacy Controls for Federal Information Systems and Organizations.” *NIST Special Publication 800-53*, Apr. 2013, pp. B-6-B-6., https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r4.pdf

4.2. Han, Jiawei, et al. *Data Mining: Concepts and Techniques*. Third ed., Morgan Kaufmann, 2012, http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.

4.3. Hebbar, Prajakta. “6 Major Data Quality Issues That Haunt Almost All Major Organizations.” *Analytics India Magazine*, 27 Sept. 2019, http://analyticsindiamag.com/6-major-data-quality-issues-haunt-almost-major-organisations/.

4.4. Forbes Staff. “Poor-Quality Data Imposes Costs and Risks on Businesses, Says New Forbes Insights Report.” *Forbes*, Forbes Magazine, 31 May 2017, www.forbes.com/sites/forbespr/2017/05/31/poor-quality-data-imposes-costs-and-risks-on-businesses-says-new-forbes-insights-report/#1479a2b4452b.

4.5. Redman, Thomas C. “Bad Data Costs the U.S. $3 Trillion Per Year.” *Harvard Business Review*, Harvard Business School, 4 Oct. 2017, www.hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year.

4.6. Martin, Emmie. “Apple Just Hit a $1 Trillion Market Cap-Here's Why Its Little-Known Third Co-Founder Sold His 10% Stake for $800.” *CNBC.com*, CNBC, 2 Aug. 2018, www.cnbc.com/2018/08/02/why-ronald-wayne-sold-his-10-percent-stake-in-apple-for-800-dollars.html.

4.7. “Insights for Professionals.” *Insights For Professionals*, IFP, www.insightsforprofessionals.com/.

4.8. Redman, Thomas C. “If Your Data Is Bad, Your Machine Learning Tools Are Useless.” *Harvard Business Review*, Harvard Business School, 3 Apr. 2018, www.hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless.

4.9. TrustRadius. “List of Top Data Deduplication Tools 2020.” *TrustRadius.com*, www.trustradius.com/data-deduplication.

4.10 Twin, Alexandra. “Data Mining: How Companies Use Data to Find Useful Patterns and Trends.” Investopedia, Investopedia, 29 Jan. 2020, www.investopedia.com/terms/d/datamining.asp.

4.11 Koh, Hian Chye, and Gerald Tan. “Data Mining Applications in Healthcare.” Journal of Healthcare Information Management : JHIM, *U.S. National Library of Medicine*, 2005, www.ncbi.nlm.nih.gov/pubmed/15869215.

4.12 Clifton, Christopher. “Data Mining.” Encyclopædia Britannica, Encyclopædia Britannica, Inc., 20 Dec. 2019, www.britannica.com/technology/data-mining.

4.13 Tibco. “5 Data Quality Problems and Their Solutions.” Https://Www.insightsforprofessionals.com, 13 Nov. 2018, www.insightsforprofessionals.com/en-us/it/storage/data-quality-problems-solutions.

4.14 Saunders, Asena Atilla. “The History of Data Mining.” Exastax, Exastax, 31 Jan. 2018, www.exastax.com/big-data/the-history-of-data-mining/.