[Research Causes of Bad Data](https://learning.rasmussen.edu/webapps/assignment/uploadAssignment?content_id=_5882931_1&course_id=_65722_1&group_id=&mode=view)

Shaun Pritchard

Rasmussen College

QMB3000

Kevin Watts

April 26, 2020

[**Research Causes of Bad Data**](https://learning.rasmussen.edu/webapps/assignment/uploadAssignment?content_id=_5882931_1&course_id=_65722_1&group_id=&mode=view)

**Let us explore the definition of the word integrity for a minute. Integrity by definition means; "The quality of being honest and having strong moral principles". When I was in the military USAF, they told me that this word means doing the right thing and doing what to expect it when no one else is around. There is however one more definition to this word which refers to the Integrity of data. Which means the accuracy, consistency, and validity of data over its life cycle. Hypothetically let us assume this word was interpreted from some random user submission on a web-form. How would a computer processing this as raw data know which definition to apply to this word?**

**There are over 150 words in the English language which are spelled the same and have multiple meanings. When these words are processed and interpreted as raw data to be transformed into information businesses rely on to make vital decisions. It is important to have quality data provisions in place to provide this integrity. Ultimately interpreting the data into the correct transformations for this business Analytics. These provisions determine the quality of data any misinterpretations cause poor data quality.**

**This is but one example, in this research we are going to discuss 4 of the main causes of poor data quality, some solutions, provisions, and indicators to prevent and find bad data. To understand this more clearly, we need to ask a few questions? Then we need to associate four categories in which these questions are relevant.**

**Data questions we should ask:**

* **What are the main issues is that businesses are unaware that they have data quality issues.**
* **How good is an organization’s data?**
* **How does an organization analyze and improve the quality of its data?**
* **Does the organization have the correct infrastructure?**
* **Is the correct technology being incorporated for the extraction and transformation of data into viable data to make correct business decisions?**
* **Is the data complete?**
* **Is the data collected on time?**
* **Is the data correct?**

**These are just a few very important questions outlined in this research. Questions that need to be asked when analyzing data. Poor data quality can take many abstract shapes & forms; not just incorrect metrics or entries, but a lack of completeness, lack of organization, too much, too little, or the data being too old (for meaningful use). There are unlimited potential reasons for poor quality data as follows.***(Guess, 2011)*

* **Excessive amounts of data being collected; too much data to be collected leads to less time to do it, and “shortcuts” to finish reporting**
* **Many manual steps; moving figures, summing up, etc. between different paper forms**
* **Unclear definitions: wrong interpretation of the fields to be filled out**
* **Lack of use of information: no incentive to improve quality**
* **Fragmentation of information systems; can lead to duplication of reporting**

**The main 4 factors of bad data**

**The main 4 factors we will discuss below that are the main cause bad data are as follows. keep in mind that many of the stated issues above can be encompassed within these factors attributing to the problem and ultimately the integrity of the data** *(Spotless Data, 2020)****.***

***Outline of the 4 main factors:***

1. **Poor Organization**

**If you are not able to easily search through your data, you’ll find that it becomes significantly more difficult to make use of. Through different organizational methods and procedures, there are dozens of ways that data can be represented.**

1. **Too Much Data**

**40% of people reported that there is often too much data to properly work off of inside a database. While it might seem like “too much data” can never be a bad thing, more often than not, a good portion of the data simply isn’t usable, which is going to mean that you’re spending more time digging through the bad so they can get to the good.**

1. **Inconsistent Data**

**When dealing with multiple data sources, inconsistency is a big indicator that there is more than likely going to be a data quality problem. In many circumstances, the same records might exist multiple times in a database. Duplicate data is one of the biggest problems that exist for data-driven businesses and can bring down revenue faster than any other data issue.**

1. **Poorly Defined Data**

**Oftentimes data is poorly defined, which causes great confusion around the proper methodology for management. For example, data that’s sectioned into the wrong category, like a company account being filed as a single person’s contact, is going to really mess things up in your database and make the whole thing more difficult to understand and sort through.**

**Elaboration of the 4 main factors of bad data**

**Poor Organization**

**Improving data quality is a long-term task, and many of the measures are organizational in nature. However, data quality should be an issue from the start of any implementation process. Part of the issues with poor organization contributing to bad data is ineffective controls. Where not only to employees, managers, and such have issues with adapting to new technology they begin to establish bad practices that might be built into technology systems. Also, they avoid them to get things done more quickly** *(Greene, 2015)***.**

**They also allow different parts of the business to become isolated from one another. On top of this, we can add human error such as when employees manually enter data simple typos and spelling mistakes account for 58% of errors in data. on top of technological errors and bad organizational practices. Companies who just chalk up the loss are implementing negative occurrences. There are few things that can be done to facilitate the avoidance of bad data though***(Spotless Data, 2020)****.***

**Poor organization Solutions:**

* **Perform consistent data process audit and utilize gatekeeper roles**
* **Changes in data collection forms, harmonization of forms**
* **Promote information use at the local level, where data is collected**
* **Develop routines on checking data quality (especially for siloed departments)**
* **Include data quality in training**
* **Ensure proper training and systems on new Technologies Incorporated into the organization.**
* **Challenge professionalism**
* **Avoid manual entry (let technology do its job)**

**Too Much Data**

**Many problems we see with poor data quality partially consistent with the overloads of historical data. The collection of data which facilitates no purpose in providing strategic metrics. Also, migrating and mixing this type of data to new systems carries an inherent risk to data quality. As a report shows 2013 an analysis was published on how Motorola undertook to answer the question of how much money does bad data cost our company. They answer this question by applying it to activity-based (ABC) to a data cleanup project. The specific focus on cleansing large amounts of data. They initially predetermined a set of specific metrics as a guideline. Then apply these metrics honest Pacific data element they referred to as a planning yield. Which they were able to quickly identify which information was considered bad and which sets of information would be valuable and allowed for the project team to process. To implement a simple process system, they were able to go through what historical data to essentially invaluable data elements** (Grime, 2013)**.**

**Cleaning data especially Legacy or historical data proves to be very beneficial in this case. They are able to develop a plan of action to turn otherwise dead-weight data into viable metrics data. This is about one instance the causality between having too much data it is inconsistent with the business goals in mind. Here are a few recommendations so not only avoid but to determine and clean this specific type of data issue***(Syniti, 2014)****.***

**Too much data Solutions.**

* **Implement data cleansing sessions**
* **Implement a CMMS**
* **Establish business rules and standard operating procedures**
* **Use the angle and create metrics four guidelines of weeding out bad data**
* **demonstrate failure data analytics showing how this data will be used to help the company stay in business**
* **Grade data quality**
* **Implement Predictive Analytics.**
* **Use Machine Learning and sort algorithms on data**

**Inconsistent Data**

**As mentioned before this is one of the best biggest issues of data. Combining multiple data inconsistency datasets redundant or invalid information not only causes for bad data but can cost time and resources and processing and transforming this data. Also combining different data of different data types. Can contribute to the consistency problem afford air quality and create Integrity constraints** *(Brisaboaa, 2015)***.**

**For example, this could be telephone numbers with no dialing code or too many spaces, or money in different currencies, or a name written as an initial rather than fully. These are serious issues, here are a few solutions to reconcile inconsistent data. Also, I might mention this is one of the hardest issues to tackle in data science. It really all comes down to the types of data being processed as well as the systems in place.**

**Inconsistent data solutions:**

* **Automated processes**
* **Establishing value/importance to the working level**
* **Consistent audits in data cleansing**
* **Developing a grading system for data quality**
* **Use OLAP and normalize databases**
* **Research new technologies being implemented**
* **Develop Data source quality criteria**
* **Implement machine learning and AI models on data**

**Poorly Defined Data**

**The top three issues above all relate to the consistency of poorly defined data. Whether it is human error, lack of understanding in common technologies, outdated database models, or organizational isolation. Business departments become isolated creating the Department of silos. Once the silos have been created and taking not only hard to break down, they also encourage the use of data definitions that are different from those used by other parts of the organization. These actions facilitate unstructured communication which in turn creates the initial cause to poorly defined data. Below are a few solutions to avoid poorly defined data** *(Cragg, 1994)****.***

**Poorly Defined Data Solutions:**

* **Structured models of categories**
* **Predefined management categorization**
* **Implement business rules & standard operating procedures for data**
* **Consistent management to reconcile data between teams and applications**
* **Audit checks for inconsistent input data**
* **Implement business rules & standard operating procedures**

**Conclusion**

**The negative impact of poor data quality on companies has been documented in the industry. Data quality problems result in lower customer satisfaction and increased costs. It has been estimated that data quality problems can cost companies as much as 8–12 percent of revenue in the industry or higher Often the magnitude of these problems may be unclear, unquantified, or unrecognized by managers who have become accustomed to accepting the cost associated with data quality problems as the normal cost of doing business** *(Journey to Data Quality, 2006)***.**

**Implementing new procedures and systems as recommend what are the four basic categories above. It can ultimately facilitate diverting organizations from the negative implications associated with what's poor data quality.**

**Due to the prevalence of poor data, extraction transform load ETL and data quality management Technologies have to be implemented and trained internally to avoid the cost of bad data. Through these best practices noted above. organizations no longer have to settle to accepting the high percentage cost of lost revenues. Simply by obtaining and implementing data systems, structure, and standards companies can convert poor data into ROI converting data.**

**We can establish a solid basis by observing the three categories of costs: (1) costs of non-quality information, (2) costs of assessment or inspection, and (3) costs associated with process improvement and defect prevention. Included in the costs of non-quality of information would be processed failure costs, li-ability and exposure costs, information scrap and rework costs, and lost and missed opportunity costs** *(Journey to Data Quality, 2006)***.**

# References

Brisaboaa, N. R. (2015). Rank-based strategies for cleaning inconsistent spatial databases. *International Journal of Geographical Information Science*, 26.

Cragg, J. g. (1994). Making good inferences from bad data. *Canadian Journal of Economics*, 26.

Greene, K. B. (2015). Bad Data States are embracing the promise. *Governing*, 9.

Grime, L. J. (2013). How Much Does Bad Data Cost Your Company? *Journal of Corporate Accounting & Finance* , 4.

Guess, A. (2011). Root Causes of Data Quality Problems. *Journal of Information Science and Engineering* , 1.

Spotless Data. (2020). *14 CAUSES OF DATA QUALITY PROBLEMS*. Retrieved from Spotlessdata.com: https://spotlessdata.com/blog/14-causes-data-quality-problems-part-1

Syniti. (2014, 3 24). *The Three Root Causes of Poor Big Data Quality*. Retrieved from syniti.com: https://resources.syniti.com/syniti-blog/the-three-root-causes-of-poor-big-data-quality

W., L. Y. (2006). Journey to Data Quality. *American Psychological Assoc*, 1.