



Validating “Data Validation”

Norita Ahmad, American University of Sharjah

Kevin Dias, MRM

We have more access to data than ever before, but it is difficult to make sense of it when the data is incomplete or inaccurate. As such, data validation is necessary to increase data quality for effective decision making.

Data validation is an essential part of data handling, notably in the fields of data science, artificial intelligence, and the Internet of Things. As the name suggests, it is the checking of quality and accuracy of data prior to processing it. Data validation is not a new topic; however, most data handling processes do not follow a systematized data validation approach and hence lead to small- or large-scale error-prone conclusions.¹⁻³ The use of a proper data validation process could prevent life-threatening scenarios such as incorrectly classifying a cancerous tissue as benign,¹ prevent the loss of millions of dollars by ensuring the accuracy of a price prediction model,² or prevent a fatal accident of a self-driving car that failed to recognize a jaywalking pedestrian.³ As such, there is an urgent need

for efficient and effective data validation procedures.

Today, with the advances in big data and analytics, businesses, governments, and individuals can apply diverse data mining and machine-learning algorithms to bring new business opportunities and improve quality of life. The International Data Corporation predicted that, by 2025, data will

grow to 175 trillion gigabytes worldwide, and businesses will become even more reliant on big data for decision making.⁴ However, given the nature of big data (that is, huge volume of generated data, fast velocity of arriving data, and large variety of heterogeneous data), the quality of data can easily be compromised.⁵ There are anecdotes where collected data was so voluminous that people eventually discarded it since it derived no benefit.⁶

WHAT IS DATA VALIDATION?

There are different definitions of data validation but, in general, data validation is a process of delivering clean and accurate data to specific programs, applications, and services.⁷ For example, in machine learning, data validation means checking the quality and accuracy of source data before training a new model.⁸ Different types of validation can be performed depending on the objectives and constraints of a given data set.

Data validation usually occurs during the transform stage of the extract, transform, and load (ETL) data process, where data are first extracted from a data source (Stage 1); validated, cleaned, merged, formatted, and/or appended with other data set extracts (Stage 2); and finally ready to load (Stage 3) and process as per the given use case.⁹

Different types of validation can be performed depending on the objectives and constraints of a given data set.

WHY DATA VALIDATION?

Depending on a specific industry, there are numerous reasons why data validation is critical to data-driven projects. The two most important reasons that are often taken too lightly are: 1) early detection of errors and 2) the cost of time saved.^{5,8,10} These two go hand in hand, but it is important to highlight them independently as various decision makers and data experts tend to exclude data validation due to a lack of knowledge on the many consequences from each of the two.

Early detection of errors

Validating details of data are necessary to mitigate any project defects. Given that businesses rely on high-quality data to make critical decisions, they run the risk of basing decisions on data that are not accurately representative of the situation at hand if data validation is not properly performed. According to Gartner, on average, the financial impact of poor data quality on organizations is US\$9.7 million per year.²

Cost of time saved

It is without question that data scientists, analysts, and engineers are in one of the highest paid lines of work. The reported median annual wage for these positions in 2020 was higher than the median annual wage for all

other occupations.¹¹ However, it was reported that one out of three data analysts spend over 40% of their time “vetting and validating their analytics data before it can be used for strategic decision-making.”¹² In the past, data prep tasks have occupied around 80% of a data scientist’s time—challenging the wisdom of asking highly paid data scientists to spend most of their time

preparing data instead of using them for the actual analysis and decision making. Imagine the value of time saved if there is a solution to this efficiency gap.

HOW IS DATA VALIDATION PERFORMED?

The most straightforward rules used in data validation are rules that ensure data integrity, for example, the correct format to enter a phone number or the minimum password length. These basic data validation rules help to uphold standards that will effectively make working with data more efficient. During the data validation process, it is important that the standards and structure of the data model is also well understood. Structured data are data that has been formatted into a well-defined data model while unstructured data are data in a raw form. There is also semistructured data that fall in between structured and unstructured data.¹³ Understanding the difference between these types of data will help in maintaining compatibility with applications and other data sets with which data are integrated and stored.

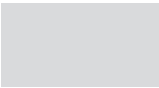
There are many methods available for data validation. The most common methods are using script and software programs. Writing a script may be an option for those who are fluent in coding languages such as Python and

R. Script allows for a better interpretation of the data, such as comparing data values and structure against predefined rules and verifying all the necessary information within the required quality parameters. This method can however be very time consuming depending on the complexity and size of the data set. Today, there are many commercial software programs that can be used to perform data validation. For most people, this is the preferred method since the program has been developed to understand the common rules and file structures used in data validation in the industry. No in-depth understanding of the underlying format is required from the user. However, this method can be a bit costly. As an alternative, a more technical user would opt for an open source software such as SourceForge and OpenRefine because they are more cost-effective.

Another important aspect of data validation is the evaluation of the validity of range measurements. The top use case is in satellites or deep-space exploration systems such as those used by NASA. Given that the system state estimates can be altered significantly by erroneous sensor data, an algorithm which decides whether to assimilate or reject data are required. Algorithms such as least squares, linear combinations, Bayesian, or Kalman filtering can be used to select sensors that are consistent with the computed estimates and reject those that are far from the predicted values.^{14,15}

WHAT ARE THE CHALLENGES?

Let’s face it, data validation is by far the least exciting data-related buzzword that exists. This is a fundamental failure in both industry as well as academia. The only people you see actively talking about it are veterans in the field who have firsthand experience in realizing its importance too late. Furthermore, as of the writing of this column, there is a surplus of data practitioners dominated by millennials, with a demand from employers who



have a flawed view of recruiting based on quantity and not quality. It is only being part of a “trend” for 20-year-olds to hone in on their ability to build predictive models rather than verify the credibility of their data sources.⁵ Additionally, the majority of data analytics teams today is so time constrained in delivering “quick insights” that data validation is unfortunately classified as a waste of time. To further stress on the broad lack of love for data validation in the industry, there are over a hundred different data science/analysis certifications available but how many data validation certifications are there?

The ETL processes are inherently so focused on “getting the job done” with speed and efficiency, with accuracy being a low priority. In the top-down team structures where there is a cascading effect of pressure, the data analyst doing the heavy lifting is left with a difficult choice between spending time sense checking their data or keeping their job. Data governance and data ethics-related roles such as the chief data officer are only recently being birthed into the industry and will hopefully bring more emphasis on the importance of allocating engineering hours specifically to data validation.

Additionally, the management teams often fail to make a clear distinction between a data science data validation and a human-oriented data validation. There is an industry-coined phrase called “human-in-the-loop” validation that is used for situations when it isn’t enough to run if/else error tests but also necessary for a person to sense-check the data themselves. Examples of this are phone numbers (that require an actual phone call), email addresses (email bounce-back test), and websites (that need a Google search). When a human successfully makes these validation tests, machine learning can be applied to both learn from the validations and also to assess newly recruited “human data validators.”

It is, however, important to note that personally identifiable information

(PII) is sensitive, personal data that is protected by data regulatory laws such as the General Data Protection Regulation¹⁶ and can make data validation tedious. This is particularly a challenge with the types of data mentioned earlier (that is, email address and phone

validation is a part of a project that exists to help a human with a data-driven decision. For example, before digital marketing existed, traditional marketers with little to no data, would make decisions based on gut instinct and “intuition” from their experiences—

Decision makers need to have fundamental standards for quality and reliable data because they are critical for corporate survival.

number) that can be tied back to a person. Since these regulations make attaining, storing, and using PII data a costly liability for a company, it can be risky when PII data are used as part of the data validation approach or as the data being validated itself.

Perhaps the most challenging of all is scalability. Scalability of data validation is how decision makers find the easiest excuse to turn a blind eye on. It is a costly aspect of the data pipeline that needs more research and attention. To go from 95% accuracy to 99% accuracy, in confidence intervals, can cost a company an exponential amount more versus their standard resourcing cost in data validation due to the human aspect of it. When a business decision maker is more comfortable to go with 95% over 99% accuracy, it is often too late to realize that it could cost thousands of dollars to a startup or billions of dollars to a Fortune 500 company in a worst-case scenario. This needs to change. Decision makers need to have fundamental standards for quality and reliable data because they are critical for corporate survival.²

DATA VALIDATION DASHBOARD

We have discussed some of the essential key concepts and challenges regarding data validation. It is evident that when it comes to data validation, human oversight cannot be completely removed from the practice. Data

invaluable human traits that cannot be learned. Even today, no matter how advanced data validation software is, or automated methodologies are, humans are unlikely to proceed with decisions that are not vetted by another human. The final decision would probably be made by someone at a senior level, with experience and contextual knowledge, who can tell whether the data validation process was successful.

A *data validation dashboard* might serve as a means for an experienced data practitioner to monitor data analysis processes from start to finish. The dashboard could also be used as a tool that, over time, improves recommendations for handling data validation by suggesting the appropriate time to be spent at each step, the number and types of people involved, the format of the data validation technique, and the expected results.


Two of the core functions of such a dashboard would be

- a. a checklist to help decide and sequence the most appropriate tasks to apply for a given problem
- b. a reference list of learned data quality issues filtered specific to the data set/problem at hand.

The dashboard could enable teams or project managers to allocate resources, tasks, and also more effectively supervise the status and the best outcome possible for projects.

Since many people are comfortable with knowing that they understand data validation at the fundamental level, it is most important for you to take away the more subtle nuances about it that this column article highlights. Data validation generally occurs but isn't exclusively at the data source level. Data sources or data sets should be acknowledged as being comparatively "more" or "less" accurate, as perfect accuracy is impossible to conclude. Records and data points, although confirmed accurate, can run the risk of inaccuracy over time if not routinely verified.

Records and data points, although confirmed accurate, can run the risk of inaccuracy over time if not routinely verified.

As marketing automation and data-driven retargeting becomes more widespread, there will be an increasing need to guarantee the accuracy of data pertaining to real people and business entities. A data analytics operation can produce ground-breaking data-driven solutions or recommendations, but when presented to an already risk-averse business decision maker, the last thing you want is to allow the risk of having employed inaccurate data from the very beginning. 

REFERENCES

1. C. Farr, "This patient's medical record said she'd given birth twice—In fact, she'd never been pregnant," CNBC, Dec. 9, 2018. <https://www.cnbc.com/2018/12/09/medical-record-errors-common-hard-to-fix.html> (accessed June 21, 2021).
2. "Poor - Quality data imposes costs and risks on business, says new Forbes insights report," Forbes, May 31, 2017. <https://www.forbes.com/sites/forbespr/2017/05/31/poor-quality-data-imposes-costs-and-risks-on-businesses-says-new-forbes-insights-report/?sh=4eb96655452b> (accessed June 21, 2021).
3. A. Marshall and A. Davies, "Uber's self-driving car saw the woman it killed, report says," WIRED, May 24, 2018. <https://www.wired.com/story/uber-self-driving-crash-arizona-ntsb-report/> (accessed June 21, 2021).
4. D. Reinsel, J. Gantz, and J. Rydning, "The digitization of the world from edge to core," International Data Corporation, Needham, MA, White Paper, 2018. Accessed: June 19, 2021. [Online]. Available: <https://resources.moredirect.com/white-papers/idc-report-the-digitization-of-the-world-from-edge-to-core>
5. J. Gao, C. Xie, and C. Tao, "Big data validation and quality assurance - Issues, challenges, and needs," in *Proc. IEEE Symp. Service-Oriented Syst. Eng. (SOSE)*, Mar. 2016, pp. 433–441. doi: 10.1109/SOSE.2016.63.
6. J. Anderson and L. Rainie, "The future of big data - Main findings: Influence of big data in 2020," Pew Research Center, 2012. <https://www.pewresearch.org/internet/2012/07/20/main-findings-influence-of-big-data-in-2020/> (accessed July 8, 2021).
7. "Techopedia explains data validation," Techopedia, 2017. <https://www.techopedia.com/definition/10283/data-validation> (accessed June 15, 2021).
8. N. Polyzotis, M. Zinkevich, S. Roy, E. Breck, and S. Whang, "Data validation for machine learning," in *Proc. Mach. Learn. Syst.*, 2019, vol. 1, pp. 334–347.
9. T. Jun, C. Kai, F. Yu, and T. Gang, "The research & application of ETL tool in business intelligence project," in *Proc. Int. Forum Inf. Technol. Appl.*, May 2009, vol. 2, pp. 620–623. doi: 10.1109/IFITA.2009.48.
10. N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, "Data lifecycle challenges in production machine learning: A survey," *ACM SIGMOD Rec.*, vol. 47, no. 2, pp. 17–28, June 2018. doi: 10.1145/3299887.3299891.
11. "Salary guide 2020," Robert Half Technology, Menlo Park, CA, 2020. [Online]. Available: https://www.roberthalf.com/sites/default/files/documents_not_indexed/2020_Salary_Guide_Technology_NA.pdf (accessed June 20, 2021).
12. C. Taylor, "Structured vs. unstructured data," Datamation, May 21, 2021. <https://www.datamation.com/big-data/structured-vs-unstructured-data/> (accessed July 8, 2021).
13. M. Goetz, G. Leganza, E. Miller, and J. Vale, "Data performance management is essential to prove data's ROI," FORRESTER, 2018. <https://www.forrester.com/report/Build+Trusted+Data+With+Data+Quality/-/E-RES83344> (accessed June 20, 2021).
14. M. Fernandez and H. F. Durrant-Whyte, "An information-theoretic approach to data-validation," in *Proc. IEEE Amer. Contr. Conf.*, June 1993, pp. 2351–2355. doi: 10.23919/ACC.1993.4793308.
15. A. El-Mowafy, "Diagnostic tools using a multi-constellation single-receiver single-satellite data validation method," *J. Navig.*, vol. 68, no. 1, pp. 196–214, 2015. doi: 10.1017/S0373463314000526.
16. C. Bernstein, "Personally identifiable information (PII)," TechTarget, Feb. 2020. <https://searchsecurity.techtarget.com/definition/personally-identifiable-information-PII> (accessed June 27, 2021).

NORITA AHMAD is an associate professor of management information systems at American University of Sharjah, Sharjah, United Arab Emirates. Contact her at nahmad@aus.edu.

KEVIN DIAS is a performance and analytics manager at MRM, Toronto, M5V 0N6, Canada. Contact him at kevinrosedias@gmail.com.