Acta Scientific COMPUTER SCIENCES

Volume 3 Issue 12 December 2021

Mini Review

On the Preeminence of Data Quality

Christian Mancas*

DATASIS ProSoft Srl, Bucharest, Romania

***Corresponding Author:** Christian Mancas, DATASIS ProSoft Srl, Bucharest, Romania.

Received: October 18, 2021

Published: November 10, 2021

Abstract

This editorial paper pinpoints the paramount importance of data quality in both computer science and information technology, especially nowadays, when data is the key world asset.

Keywords: Data Quality; Data Plausibility; Correctness Proofs; Object-oriented Programming; Structured Programming; Automated Software Testing; Automatic Code Generation; Databases; Constraints; Coherence and Minimality of Constraint Sets; Social Media Platforms; Fake News; Artificial Intelligence; Machine Learning

Introduction

Data was always important to mankind: from the dawn of civilization (Where do you find good shelter, water, and wood for fire? Where do you find best places to capture animals, fish, fruits, etc.?) and until nowadays (be it scientific, commercial, governmental, private, etc.), its importance ever grew exponentially.

Historically, first efforts towards acquiring data were made in astronomy, starting some 4 millennia ago, for correctly dividing time in years, seasons, months, days, hours, minutes, for computing distances from Earth to other celestial bodies, predicting eclipses, navigating, etc., by both Babylonians, Egyptians, Greeks, Indians, Chinese, etc. Then, Euclidian geometry emerged mainly from the need to correctly delimit agricultural land between adjacent owners.

Today, data is a key asset of us all (e.g. [23]). According to The Economist [14], data is already the most valuable resource in the world, ahead of oil. This is no surprise, as 97% of companies use data to power their business opportunities and for 76% of the businesses data serves as an integral part of forming their business strategies (e.g. [28]).

Very probably, one of the most memorable sayings of the previous century that is increasingly meaningful in the current one as well is due to the famous Dr. W. Edwards Deming (known, among other, as the father of continual improvement of quality): "In God we trust; all others must bring data" [12].

But is any data worth? Especially in this more and more polluted by fake news century (e.g. [33]), obviously not. And it is not at all only about fake data: it is also about unplausible one (e.g. [17]), be it generated by human or machine errors.

Driven by Mathematics, their Queen, Sciences (starting with Logic, then Geography, Anatomy, Physics, Chemistry, History, etc.) did, are doing, and always will struggle to measure and infer only quality data - the only data guaranteeing Truth and worthy scientific results.

Of course that, in the process, lot of poor, unplausible data was measured or inferred, but such data was eventually eliminated in time.

Moreover, lot of fake data was generated as well, for deceiving adversaries, gaining battles or advantages by all possible means - which still and will always exist, but that is out of the scope of this editorial that focuses only on computer science and information technologies.

Hence, we are among those who consider that data quality should be of preeminence in any field of endeavor. Defined as a measure of how well suited a data set is to serve a specific purpose, data quality is, in fact, based on several measures of its main characteristics: accuracy, completeness, consistency, relevance, validity, uniqueness, timeliness, etc. (e.g. [27]).

State-of-the-art

In computer science (but also in logic, analysis, statistics, etc.), "Garbage in, garbage out" (GIGO) [25] was from the beginning, is, and will always be a supreme axiom. Consequently, checking parameter data is always a first task for any professional in this field.

But quality input data may not guarantee alone quality output one: algorithms' quality is at least as crucial. Therefore, algorithms' correctness proofs (e.g. [1]), structured ([5,7]) and then object-oriented programming (e.g. [9,24]), first manual and then, more and more, automated software testing (e.g. [3]), and, ultimately, automatic code generation (e.g. [20-22,30]) were developed and are consolidated.

Especially in commercial databases (dbs) data quality is crucial; for example, nobody, be it supplier or customer, would accept wrong orders, products, quantities, prices, delivery dates, invoices, payments, etc. Therefore, data models include constraints that guarantee at least data plausibility (e.g. [16-18,22]). Moreover, whenever data models provide lot of constraint types, the coherence and minimality of constraint sets are crucial as well (e.g. [19,22]). As such, lot of db architects and cohorts of IT employees work daily to prevent and correct any missing or incorrect data in these dbs.

To begin with, social media platforms (generally powered by nosql db management systems, which do not enforce data plausibility constraints) were almost not at all concerned with data quality; even worse, for example, they are still compressing photo and video files, which is significantly reducing their quality. Recently, however, confronted with both bullying, fake news, and terrorism propagation, they had to act against them, by monitoring posts, enforcing stricter policies, and even banning trolls, but also the former POTUS (e.g. [32]). Unfortunately, success in this fight is yet not at all complete (e.g. [2]).

In Artificial Intelligence, things are much more complicated (e.g. [31]). On one hand, there are lot of data cleansing tools [8] that help obtaining high quality data. On the other, the data quality issues are seriously hindering the speed of AI system implementations (e.g. [11]). Biases underlying many historical datasets are a huge problem as well (e.g. [15]). Even technological giants like Facebook and Tesla, which pored lot of money into AI, are merely disappointed by the outcome: Mark Zuckerberg is dealing with algorithms that are failing to stop the spread of harmful content, while Elon Musk with software that has yet to drive a car in the ways he has frequently promised (e.g. [26], which recalls as well that AI has also been falling short in healthcare). This proves that AI still needs more training and better data.

In particular, in Machine Learning too, until recently, accent was merely only on models, with much less attention paid to the quality of the data sets on which models were trained. Fortunately, in this field too, data sets quality has recently become a priority as well (e.g. [29]).

Last, but not least, data quality is the cornerstone of bioinformatics too, especially in the current pandemic framework: governments, healthcare facilities, MDs, and we all as patients may hopefully benefit from quality data in this field, but also suffer potential problems and even catastrophic consequences from poor data (see, e.g. [4,10,13]).

Unfortunately, research papers on COVID-19 (that are published in an unprecedented rate) sometimes need to be retracted, due to issues with their data quality [6].

Conclusion

Data was always important to mankind, but recently it became its most valuable asset. To be successful in any field of endeavor,

from science and technology to business, from governing to healthcare, we all need more and more data, but only quality one.

Dually, poor data is compromising anything that is based on it, resulting in wrong decisions, money, effort, and time waste, and even in increased illnesses and death tolls.

Therefore, since 2012, there is even an international professional association for those interested in improving effectiveness through quality data and information: the IQ International (abbreviated as IQint, website home page IQ International - Information Quality International).

We strongly believe that much more intelligence, effort, time, and money are and will always be needed for guaranteeing data quality.

To conclude with, we would rather amend Dr. Deming's quote as: "In God we trust; all others must bring quality data".

Bibliography

1. Aspnes J. "Correctness Proofs". Univ. Nacional de Colombia, Bogota, verification.pdf (unal.edu.co) (2003).
2. Augustin F. "Troll farms peddling misinformation on Facebook reached 140 million Americans monthly ahead of the 2020 presidential election, report says". Insider (2021).
3. Blokdyk G. "Automated Software Testing A Complete Guide - 2020 Edition". 5STARCooks, BookShout, Plano, TX (2021).
4. Costa-Santos C., *et al*. "COVID-19 surveillance - a descriptive study on data quality issues". *medRxiv* (2020).
5. Dahl OJ., *et al*. "Structured Programming". Academic Press, London and NY (1972).
6. Data.Europa.EU. "COVID-19 open data quality in research papers". data.europa.eu (2020).
7. Dijkstra E W. "Notes on Structured Programming". Techn. Univ. Eindhoven, Math. Dept., NL (1965).
8. Dilmegani C. "Data Quality Tools and Criteria for Right Tools". AI Multiple, Data Quality Tools and Criteria for Right Tools (2021).
9. Freeman E., *et al*. "Design Patterns (A Brain Friendly Guide)". O'Reilly, Sebastopol, CA (2004).
10. GAO. "COVID-19 Data Quality and Considerations for Modeling and Analysis". GAO-20-63SSP, GAO-20-635SP, Accessible Version (2020).
11. Ghosh P. "Challenges of Data Quality in the AI Ecosystem". datadiversity.net, Challenges of Data Quality in the AI Ecosystem - DATAVERSITY (2019).
12. Hansen H L. "In God we trust. All others must bring data". IBM, Big Data, "In God we trust. All others must bring data". IBM Nordic Blog (2019).
13. Larsen T. "Healthcare Data Quality: Five Lessons Learned from COVID-19". HealthCatalyst, Healthcare Data Quality: 5 Lessons from COVID-19 (healthcatalyst.com) (2021).
14. Leaders. "The world's most valuable resource is no more oil, but data". *The Economist* (2017).
15. Liu L T. "When bias begets bias: A source of negative feedback loops in AI systems". *Microsoft Research Blog* (2020).
16. Mabine V J and Balderstone S J. "The World of the Theory of Constraints". A Review of the International Literature. CRC Press, Boca Raton, FL (1999).
17. Mancas C. "Conceptual Data Modeling and Database Design: A Completely Algorithmic Approach". Volume I: The Shortest Advisable Path. Apple Academic Press/CRC Press (Taylor and Francis Group), Palm Bay, FL (2015).
18. Mancas C. "On the Paramount Importance of Database Constraints". *Journal of Information Technology and Software Engineering* 5.3 (2015): 1-4.
19. Mancas C. "MatBase Constraint Sets Coherence and Minimality Enforcement Algorithms". In: Benczur, A., Thalheim, B., Horvath, T. (eds.), Proc. 22nd ADBIS Conf. on Advances in DB and Inf. Syst., LNCS 11019 (2018): 263-277.
20. Mancas C. "MatBase - a Tool for Transparent Programming while Modeling Data at Conceptual Levels". In: Proc. 5th Int. Conf. on Comp. Sci. and Inf. Techn. (CSITEC 2019) (2020): 15-27.

21. Mancas C. "On Modelware as the 5th Generation of Programming Languages". *Acta Scientific Computer Science* 2.9 (2020): 24-26.

22. Mancas C. "Conceptual Data Modeling and Database Design: A Completely Algorithmic Approach". Volume II: Refinements for an Expert Path. Apple Academic Press/CRC Press (Taylor and Francis Group), Palm Bay, FL (in press) (2022).

23. McKeen R. "Data as a key asset - maximizing value and minimizing risk in a changing legal landscape". Financier Worldwide, Data as a key asset - maximising value and minimising risk in a changing legal landscape — Financier Worldwide (2013).

24. McLaughlin B., *et al*. "Object-Oriented Analysis and Design: A Brain Friendly Guide to OOA&D: The Best Introduction to Object Orientated Programming". O'Reilly, Sebastopol, CA (2006).

25. Mellin W D. In Work with New Electronic 'Brains' Opens Field for Army Math Experts. The Times, Clipping from The Times - Newspapers.com (1957).

26. Olson P. "For Tesla, Facebook and Others, AI's Flaws Are Getting Harder to Ignore". Bloomberg Opinion, Artificial Intelligence Ain't That Smart. Look at Tesla, Facebook, Healthcare - Bloomberg (2021).

27. Rahanti R. "Data Quality: Dimensions, Measurement, Strategy, Management, and Governance". Quality Press, Milwaukee, WI (2019).

28. Redman T C. "Data driven: Profiting from Your Most Important Business Asset". Harvard Business Press, Boston, MA (2008).

29. Redman T C. "If Your Data Is Bad, Your Machine Learning Tools Are Useless". Harvard Business Review, If Your Data Is Bad, Your Machine Learning Tools Are Useless (hbr.org) (2018).

30. Thalheim B and Jaakkola H. "Models as Programs: The Envisioned and Principal Key to True Fifth Generation Programming". In: Proc. 29th European-Japanese Conf. (EJC 2019) (2019): 170-189.

31. Upadrashta P. "AI-Enabled Data Quality: Improve Data Quality Across Your Enterprise". Mastech InfoTrellis, AI Enabled Data Quality for data quality across enterprise (mastechinfotrellis.com) (2021).

32. Wells JR and Winkler C A. "Facebook Fake News in the Post-Truth World". Harvard Business School Case (2017): 717-473.

33. Zhou X and Zafarani R. "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities". Argxiv.org (2020).

Volume 3 Issue 12 December 2021