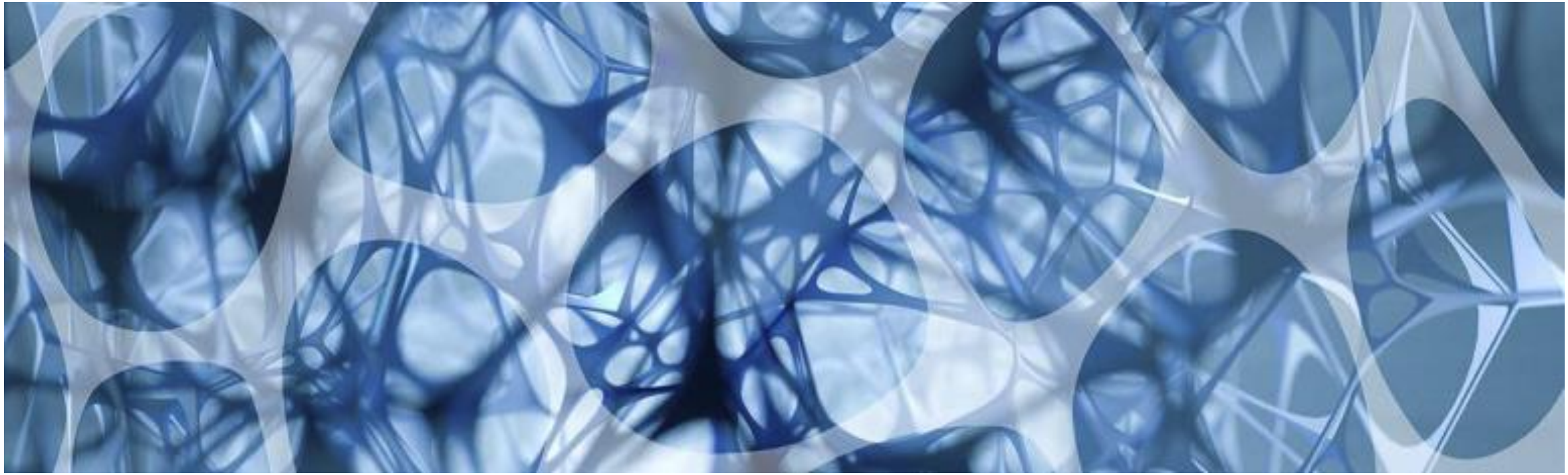
DATA QUALITY IN BIG DATA ENVIRONMENTS

Marcel Altendeitering



Fraunhofer ISST

„Garbage in, Garbage out“? – Why data quality matters

What a lack of data quality looks like…

- Research on two data sets from the domains of stock markets and airline flights, usually considered as highly reliable, showed that 70% of data items were inconsistent and 50% of them had ambiguous values [1].
- According to Gartner [3] poor data quality is responsible for an average of $15 million in costs per year and company. HBR [4] sums that up to $3 Trillion per year for the US economy.

Improved data quality leads to…

Improved data-driven decision making [5]

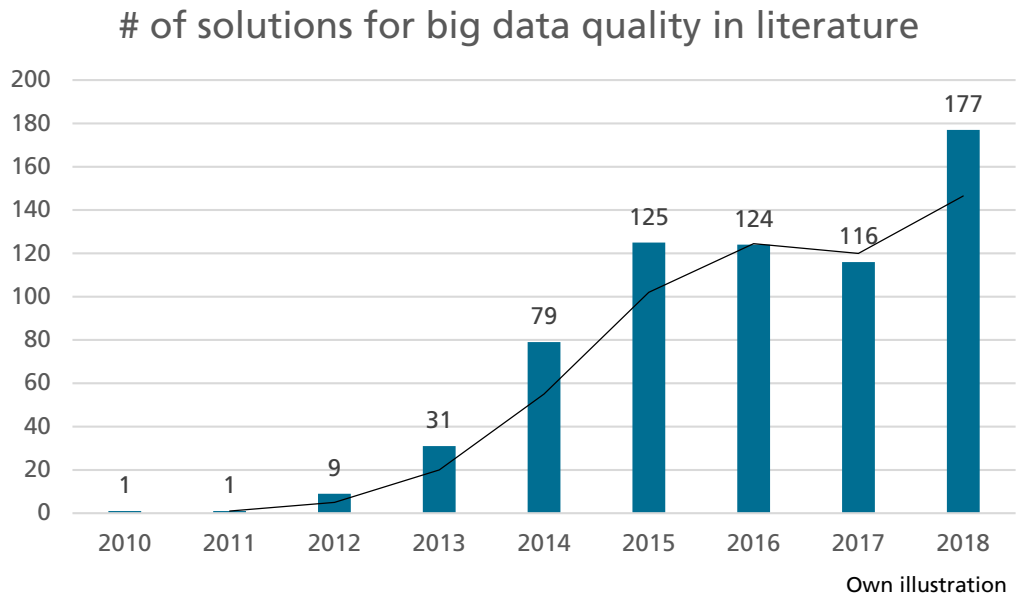
better performance of a company [2]

What makes big data quality so difficult?

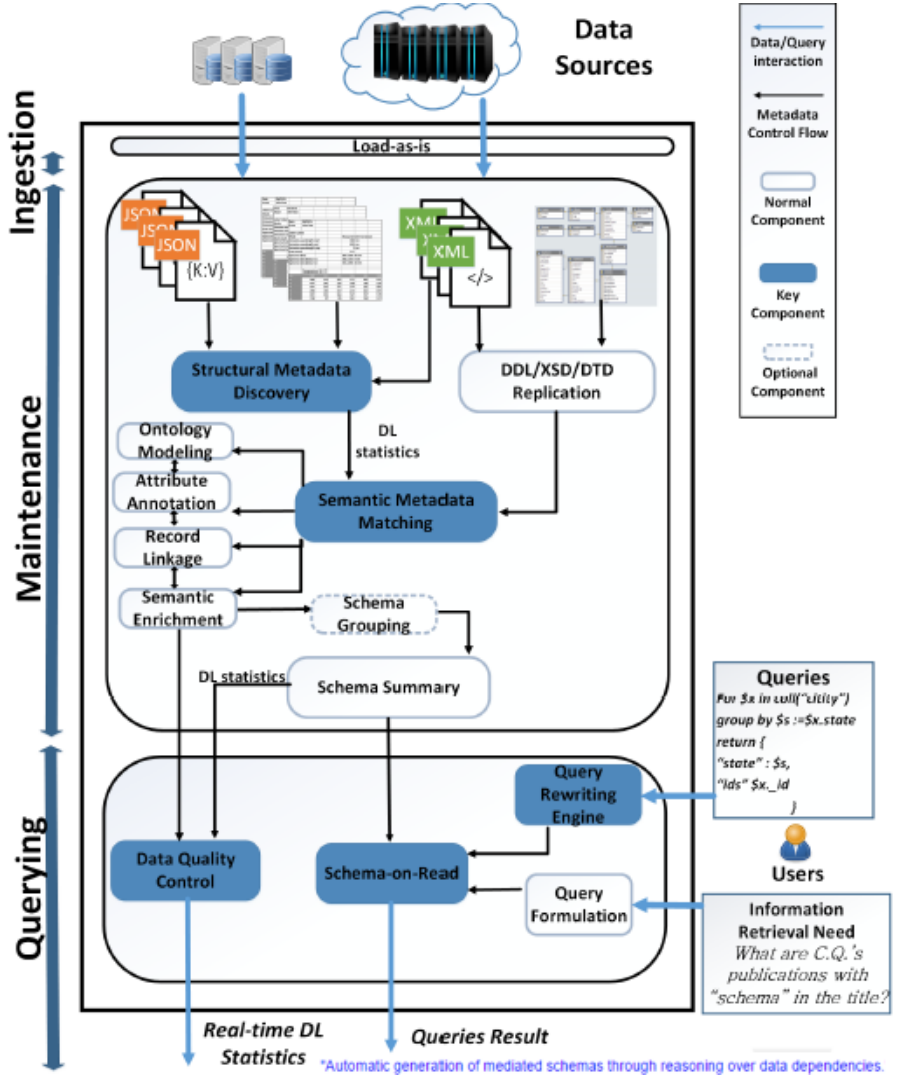
- Typical data quality issues (e.g. inconsistent, incomplete or duplicate values) are more difficult in big data environments due to the four V's [6].
- Generally data quality issues increase proportionally to volume and variety, however some errors can increase exponentially [7].
- Traditional solutions like Data Warehouses are not sufficient to deal with data quality in big data [8].

Tackling big data quality by combining organizational and technical actions

- A combination of organizational (e.g. data governance rules) and technical (e.g. automated data quality control) actions is required [4, 9, 10]
- Research on solutions for big data quality is accelerating based on literature



Own illustration



Data lake architecture proposed by [9]

Fraunhofer ISST

Marcel Altendeitering

Wissenschaftlicher Mitarbeiter

Emil-Figge-Straße 91, 44227 Dortmund
Telefon: +49 (0) 231 / 9 76 77-461

marcel.altendeitering@isst.fraunhofer.de

Fraunhofer ISST

Questions / Feedback?

Fraunhofer ISST

Sources

[1]: Li, X., Dong, X. L., Lyons, K., Meng, W., & Srivastava, D. (2012). Truth finding on the deep web: Is the problem solved?. *Proceedings of the VLDB Endowment*, *6*(2), 97-108.

[2]: Tallon, Paul P., Ronald V. Ramirez, and James E. Short. (2013). "The information artifact in IT governance: toward a theory of information governance." *Journal of Management Information Systems,* 30(3), 141-178.

[3]: https://www.gartner.com/smarterwithgartner/how-to-stop-data-quality-undermining-your-business/

[4]: https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year

[5]: Kwon, Ohbyung, Namyeon Lee, and Bongsik Shin. (2014). "Data quality management, data usage experience and acquisition intention of big data analytics." *International journal of information management* 34.3, 387-394.

[6]: Taleb, Ikbal, Mohamed Adel Serhani, and Rachida Dssouli. (2018). "Big data quality: A survey." *2018 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 2018.

[7]: Becker, David, Trish Dunn King, and Bill McMullen. (2015). "Big data, big data quality problem." *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 2015.

[8]: Berndtsson, Mikael & Forsberg, Daniel & Stein, Daniel & Svahn, Thomas. (2018). "Becoming a data-driven organisation". In *Proceedings of the 2018 European Conference on Information Systems (ECIS 2018)*.

[9]: Hai, R., Geisler, S., & Quix, C. (2016). „Constance: An Intelligent Data Lake System". *Proceedings of the 2016 International Conference on Management of Data (SIGMOD '16)*. ACM, New York, NY, USA, 2097-2100.

[10]: Farid, M., Roatis, A., Ilyas, I. F., Hoffmann, H. F., & Chu, X. (2016). "CLAMS: bringing quality to Data Lakes". In *Proceedings of the 2016 International Conference on Management of Data* (pp. 2089-2092). ACM.