Harvard Data Science Review • Issue 6.1, Winter 2024

When Data Science Goes Wrong: How Misconceptions About Data Capture and Processing Causes Wrong Conclusions

Peter Christen[1] **Rainer Schnell**[2]

[1]School of Computing, The Australian National University, Canberra, Australia,
[2]Methodology Research Group, University Duisburg-Essen, Duisburg, Germany

Published on: Jan 31, 2024

DOI: https://doi.org/10.1162/99608f92.34f8e75b

Column Editor's Note: *In an era of large, complex data, it is important for data scientists to understand how data being analyzed have been acquired, processed and linked. In this Diving Into Data column piece, Professors Christen and Schnell explore and categorize key aspects of data provenance, highlighting issues that can arise and providing recommendations to help readers identify problems and avoid resulting errors.*

Keywords: data quality, record linkage, data linkage, personal data, administrative data, data errors

Introduction

In September 2020, over 15,000 of the daily COVID-19 cases (nearly 20%) in the United Kingdom were lost for a week due to a mistake in handling spreadsheet data, resulting in delayed contract tracing and a significant increase in deaths in the affected population (Fetzer & Graeber, 2021). Also, during the pandemic, a health study (Badker et al., 2021) found that data aggregation across jurisdictions was challenged because a positive COVID-19 case was reported as either with the date of symptoms onset, the date of sample collection, or the date of a positive test diagnosis. In the summer of 2023, a German newspaper uncovered an inflated success rate of solved crime cases being reported by crime agencies due to omitted data transfers between police databases. In 2019, a bioinformatics study (Abeysooriya et al., 2021) reported that gene name errors due to the autocorrect function of spreadsheet software were found in as many as 30% of supplementary files in PubMed Central, leading to the Human Gene Name Consortium to change the names of many of the affected genes to be less susceptible to such autocorrections.

These are just some of many examples that illustrate aspects of how data are captured (collected), processed, and linked (integrated) can lead to bad outcomes and wrong decisions being made. While data scientists are generally aware of the quality of the data they are working with, they often do not know exactly how these data have been captured, what kind of processing has been conducted on the data and how these data have been linked and integrated before they are analyzed or used for decision-making. This lack of knowledge about the provenance of data can lead to a variety of assumptions that can potentially result in costly mistakes made in data science projects. Reasons for this to happen are manifold, and this article aims to categorize them and provide recommendations that help the reader identify misconceptions and avoid the resulting traps.

Characteristics of Data

Much of the data used in data science projects—especially in governments and the commercial sector—were not primarily collected for research and analytics purposes. In the public sector, governments collect data about diverse aspects of their citizens' lives and about the economy to regulate some aspects of society. Commercial organizations collect data about their customers, including their spending habits and social interactions from supposedly free online apps. Businesses furthermore collect data about suppliers, competitors, and supply chains to better understand the challenges and opportunities of the environment they are working in.

In the context of research, while in the past primary data collection has been explicit—such as via experiments and surveys—there is now a shift toward the use of administrative data in domains ranging from health to the social sciences. Administrative data are collected by an organization for some operational purpose, such as billing customers or tracking the medications given to patients (Hand, 2018). The use of such data for data science is therefore secondary. In some application areas and research domains, for example, polling (Bailey, 2023), novel methods of collecting data may seem to challenge proven research methodologies, but their possible advantages still have to be systematically tested.

Because much of the data used in data science are not explicitly collected by data scientists themselves, they have much less control over how these data have been captured, processed, and possibly linked, and only limited ways to learn about their data's provenance. This can make it unclear if such data are fit for the purpose of a specific data science project. Quoting Brown et al. (2018), "in science, three things matter: the data, the methods used to collect the data (which give them their probative value), and the logic connecting the data and methods to conclusions." *When both the collection and processing of data are outside the control of a data scientist, conducting proper research can become challenging.*

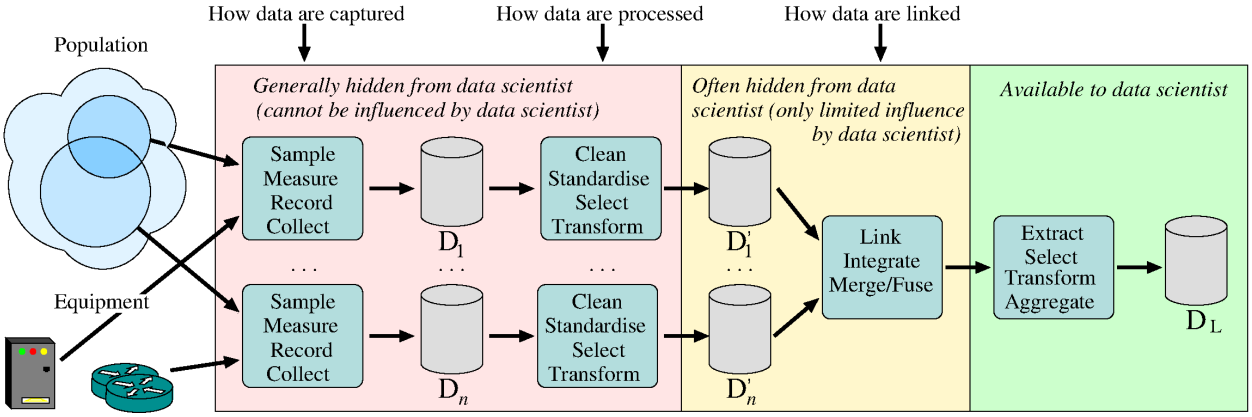

Figure 1. The general pathway of data from their source to a data scientist's computer, with D_1 to D_n being the source databases, D'_1 to D'_n the processed databases, and D_L the final linked database.

In Figure 1 we illustrate the general pathway of data from how they are collected or generated until they end up on the computer of a data scientist. While we mostly consider data about people, generally representing a population with some characteristics of interest, data can also be collected (semi-) automatically through some equipment (such as sensors or machinery) or from log files. Overall, we, however, believe that data about people—all aspects of the lives of individuals at the scale of populations (McGrail et al., 2018)—are the most valuable type of data that can be available to governments, businesses, and research organizations. For example, the value of the United Kingdom's National Health Service data, if curated, was estimated to be as much as GBP 5 billion (USD 6 billion) per year (Wayman & Hunerlach, 2019). Personal data are also the type of data most vulnerable to the various misconceptions we discuss. A main aspect of data about people is that in almost any case, some form of human intervention is happening in the pipeline from when, how, and where data are captured until they appear in the computer of a data scientist.

Even fully automated data collection methods generally involve some human efforts and interactions, such as setting up, tuning, calibrating, and maintaining equipment, and adapting them to changing environments. Automated data collection also includes communication systems that can be misconfigured or even be manipulated. Despite extensive testing, developers are often not able to anticipate any possible future abnormal situations or conditions that might occur, and therefore a system might go into a failure mode (where it stops normal operation and instead only provides some diagnostic output).

Misconceptions of the Data-Generating Process

Many of the issues we discuss below are due to humans being involved in the processes that generate data, including the mistakes and choices people make, changing requirements, novel computing and data entry systems, limited resources and time, as well as decision-making influenced by political or economic reasons. Research managers, and policy and decision makers, often also assume any kind of question can be answered with highly accurate and unbiased results when analyzing large databases (Biemer et al., 2017; Meng, 2018).

While the literature on general data quality is broad (Batini & Scannapieco, 2016), given the widespread use of personal data at the level of populations, it is surprising that only little published work seems to discuss data-quality aspects specific to personal data (Christen et al., 2020; Christen & Schnell, 2023; Smith et al., 2018; Tufiş & Boratto, 2021). One reason is due to the perceived sensitivity of this kind of data. Personal data are generally covered by privacy regulations, such as the European General Data Protection Regulation or the U.S. HIPAA, and the processes and methods employed with such data are often covered by confidentiality agreements. Furthermore, detailed data aspects are generally not included in scientific publications where the focus is on presenting the results obtained in a research study rather than the steps taken to obtain these results.[1]

Following Figure 1, we categorize misconceptions due to how data are captured, how data are processed, and how they are linked (integrated). We do not discuss any misconceptions related to the analysis of data—how to prevent pitfalls in statistical data analysis and machine learning has been discussed extensively elsewhere (Hastie et al., 2009; Riley, 2019). For detailed descriptions of misconceptions relevant to population data, see Christen and Schnell (2023).

Misconceptions Due to Data Capturing

We refer to data capture as any processes and methods that convert information about people, businesses, events, and so on, from a source into electronic format. This involves decisions about selecting or sampling individuals from an actual population to be included into a database and how to measure the characteristics of these individuals, or how sensors and equipment are configured to collect data, as well as the methods employed to collect, record, and communicate these data.

Many data-capturing methods and processes involve humans who can make mistakes or behave in unexpected ways, or equipment that can malfunction or be misconfigured. Data-capturing methods include manual data entry, optical character recognition, automatic speech recognition, and sensor readings (like biometrics from fingerprint readers and smart watches, location traces from smartphones, weather sensors, traffic cameras, and many others). While each of these methods can introduce specific data-quality issues (such as keyboard typing or scanning errors), there are common misconceptions about data capturing, as we list in Box 1.

Box 1: Misconceptions Due to Data Capturing

- A population database contains all individuals in a population.
- The population covered in a database is well defined.
- A database contains complete information for all its records.
- All records in a database are within the scope of interest.
- Each individual in a population is represented by a single record in a database.
- There are no duplicate measurements in a database.
- Records in a population database always refer to real people.
- Errors in personal data are not intentional.
- Certain personal details do not change over time.
- Coding systems do not change over time.
- Data definitions are unambiguous.
- Temporal data aspects do not matter.
- The meaning of data is always known.
- Missing data have no meaning.
- All records in a database were captured using the same process.
- All attribute/field values are correct and valid.
- Data values are in their correct attributes/fields.
- Data validation rules produce correct data.
- All relevant data have been captured.
- Automatically collected data are always correct, complete, and valid.
- Population data provide the same answers as survey data.
- Data are always of value.
- Hardware and software used to capture data are error free.

Misconceptions Due to Data Processing

It is rare for a database to be used for data science without any processing. The organization(s) that collect data, and those that further aggregate, link, or otherwise integrate such data, as well as the data scientists who will analyze the data, all will likely apply some form of data processing.

Processing can include data cleaning and standardization, parsing of free text values, transformation of values, numerical normalization, recoding into categories, imputation of missing values, removal of outliers, data aggregation, and even generating synthetic data. The use of different database management systems and data analysis software can furthermore result in data being reformatted internally before being stored and later extracted for further processing and analysis. Each component of a data pipeline can result in both explicit (user applied) as well as implicit (internally to software) data processing being conducted, leading to various misconceptions, as we show in Box 2.

Box 2: Misconceptions Due to Data Processing

- Data processing can be fully automated.
- Data processing is always correct.
- Aggregated data are sufficient for research.
- Metadata are correct, complete, and up to date.
- Synthetically generated data fully reflect reality.
- Software used to process data is error free.

Misconceptions Due to Data Linkage

Most data science projects require data from more than one source, involving some form of data integration to combine multiple data sets into a form suitable for analysis (Bleiholder & Naumann, 2009; Doan et al., 2012). The lack of unique identifiers in all source databases often prevents the simple joining of records that refer to the same entity. Techniques known as data or record linkage (Christen, 2019) are required to identify records that refer to the same person (or entity). These techniques are generally based upon comparing the quasi-identifying values of individuals, such as people’s names, addresses, and other personal details (Christen et al., 2020). Such values, however, can contain errors, be missing, and they can change over time, potentially leading to incorrect results even when modern linkage methods are employed (Binette & Steorts, 2022). As Box 3 shows, linking databases can be the source of various misconceptions about a linked data set.

Box 3: Misconceptions Due to Data Linkage

- A linked data set corresponds to an actual population.
- Linked databases represent the conditions of individuals/entities at the same time.
- A linked data set contains no duplicates.
- A linked data set is unbiased.
- Attribute/field values in linked records are correct.
- Linkage error rates are independent of database size.
- Modern record linkage techniques can handle databases of any size.
- Linkage techniques and their settings are easily transferable across domains.
- Software used to link data is error free.

Conclusions and Recommendations

Due to misconceptions such as the ones we have discussed, the promises of data science require some careful considerations. Many data scientists have potentially little control over the quality of the data they are using for their work, and any processing done on these data. They likely will also have only limited metadata that are needed to fully understand the characteristics and quality of their data. Because data are commonly sourced from organizations other than where they are being analyzed, these limitations are inherent to the kind of data used in many data science projects.

There are no (simple) technical solutions to detect and correct many of the misconceptions we have listed above. What is required is heightened awareness by anybody working with data, especially data about people. While our lists are unlikely to be exhaustive, our aim was to show that there is a broad range of issues that can lead to misconceptions. The following recommendations might help to recognize and overcome such potential misconceptions.

- If possible, data scientists should aim to get involved in the capturing, processing, and linking of any data they plan to use for their work, while those developing data-capturing, processing, and linkage systems should collaborate closely with data scientists. It is crucial for successful data science projects to form multidisciplinary teams with members skilled in data science, statistics, domain expertise, as well as ‘business’ aspects of research (Jorm, 2015).
- Cross-disciplinary training should be aimed at improving complementary skills, while training for data scientists should cover how modern data processing, linkage, and analytics methods can introduce bias and errors into the data they are working with. Training in data exploration and data-cleaning methods, as well as data-quality issues, should be part of any degree that deals with data.
- While methodologies about how to deal with uncertainties, bias, and data quality in surveys have been developed (Blair et al., 2014), there is a lack of corresponding rigorous methods that can be employed on large (administrative) databases. The big data paradox (Meng, 2018), the illusion that large databases automatically mean valid results, requires new statistical techniques and novel data-exploration methods. As already recognized by Tukey (1986, pp. 74–75): “The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.”
- It is crucial to have detailed metadata about a database, including how it was captured, and any processing and linkage applied to it. Relevant information about all sources and types of uncertainties should be collected.
- Existing guidelines, frameworks, and checklists, such as RECORD (Benchimol et al., 2015), GUILD (Gilbert et al., 2017), and the Big Data Total Error method (Biemer et al., 2017) should be adopted. Data management principles such as FAIR (Wilkinson et al., 2016) should be adhered to, although in some situations the sensitive nature of personal data might limit or prevent their application.

- The lack of publications that describe practical challenges when dealing with data can result in the misconceptions we have discussed here. We therefore encourage increased publication of data issues and the sharing of experiences with the scientific community about lessons learned and best practice approaches being implemented when dealing with data.

We have discussed some aspects in modern scientific processes that are rarely considered when data are used for research studies or decision-making. *Since good data management is a key aspect of good science, it is vital for anybody who uses data to be aware of underlying assumptions concerning the kind of data they are using* (Meng, 2022; Wu, 2022). We hope the misconceptions and recommendations given here will help to identify and prevent misleading conclusions and poor real-world decisions.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful comments and suggestions.

Disclosure Statement

Peter Christen and Rainer Schnell have no financial or non-financial disclosures to share for this article.

References

Abeysooriya, M., Soria, M., Kasu, M. S., & Ziemann, M. (2021). Gene name errors: Lessons not learned. *PLoS Computational Biology*, *17*(7), Article e1008984. https://doi.org/10.1371/journal.pcbi.1008984

Badker, R., Miller, K., Pardee, C., Oppenheim, B., Stephenson, N., Ash, B., Philippsen, T., Ngoon, C., Savage, P., Lam, C., & Madhav, N. (2021). Challenges in reported COVID-19 data: best practices and recommendations for future epidemics. *BMJ Global Health*, *6*(5), Article e005542. https://doi.org/10.1136/bmjgh-2021-005542

Bailey, M. A. (2023). A new paradigm for polling. *Harvard Data Science Review*, *5*(3). https://doi.org/10.1162/99608f92.9898eede

Batini, C., & Scannapieco, M. (2016). *Data and information quality*. Springer, Heidelberg. https://doi.org/10.1007/978-3-319-24106-7

Benchimol, E. I., Smeeth, L., Guttmann, A., Harron, K., Moher, D., Petersen, I., Sørensen, H. T., von Elm, E., & Langan, S. M. (2015). The REporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLOS Med*, *12*(10), Article e1001885. https://doi.org/10.1371/journal.pmed.1001885

Biemer, P. (2017). Errors and inference. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science* (pp. 265–297). CRC Press. https://doi.org/10.1201/9781315368238

Binette, O., & Steorts, R. (2022). (Almost) all of entity resolution. *Science Advances*, *8*(12), Article eabi8021. https://doi.org/10.1126/sciadv.abi8021

Blair, J., Czaja, R. F, & Blair, E. A. (2014). *Designing surveys: A guide to decisions and procedures* (3rd ed.). Sage. https://us.sagepub.com/en-us/nam/designing-surveys/book235701

Bleiholder, J., & Naumann, F. (2009). Data fusion. *ACM Computing Surveys* , *41*(1), 1–41. https://doi.org/10.1145/1456650.1456651

Boyd, R. J., August, T. A., Cooke, R., Logie, M., Mancini, F., Powney, G. D., Roy, D. B., Turvey, K., & Isaac, N. J. (2023). An operational workflow for producing periodic estimates of species occupancy at national scales. *Biological Reviews*, *98*(5), 1492–1508. https://doi.org/10.1111/brv.12961

Brown, A. W., Kaiser, K. A., & Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences*, *115*(11), 2563–2570. https://doi.org/10.1073/pnas.1708279115

Christen, P. (2019). Data linkage: The big picture. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.84deb5c4

Christen, P., Ranbaduge, T., & Schnell, R. (2020). *Linking sensitive data*. Springer, Heidelberg. https://doi.org/10.1007/978-3-030-59706-1

Christen, P., & Schnell, R. (2023). Thirty-three myths and misconceptions about population data: From data capture and processing to linkage. *International Journal of Population Data Science*, *8*(1). https://doi.org/10.23889/ijpds.v8i1.2115

Doan, A., Halevy, A., & Ives, Z. (2012). *Principles of data integration*. Elsevier. https://doi.org/10.1016/C2011-0-06130-6

Fetzer, T., & Graeber, T. (2021). Measuring the scientific effectiveness of contact tracing: Evidence from a natural experiment. *Proceedings of the National Academy of Sciences*, *118*(33), Article e2100814118. https://doi.org/10.1073/pnas.2100814118

Gilbert, R., Lafferty, R., Hagger-Johnson, G., Harron, K., Zhang, L.-C., Smith, P., Dibben, C., & Goldstein, H. (2017). GUILD: Guidance for information about linking data sets. *Journal of Public Health*, *40*(1), 191–198. https://doi.org/10.1093/pubmed/fdx037

Hand, D. J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society Series A*: *Statistics in Society*, *181*(3), 555–605. https://doi.org/10.1111/rssa.12315

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Jorm, L. (2015). Routinely collected data as a strategic resource for research: Priorities for methods and workforce. *Public Health Research Practice*, *25*(4), Article e2541540. https://doi.org/10.17061/phrp2541540

McGrail, K. M., Jones, K., Akbari, A., Bennett, T. D., Boyd, A., Carinci, F., Cui, X., Denaxas, S., Dougall, N., Ford, D., Kirby, R., Kum, H. C., Moorin, R., Moran, R., O'Keefe, C., Preen, D., Quan, H., Sanmartin, C. Schull, M., Smith, M., ... Kotelchuck, M. (2018). A position statement on population data science: The science of data about people. *International Journal of Population Data Science*, *3*(1). https://doi.org/10.23889/ijpds.v3i1.415

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, *12*(2), 685–726. https://doi.org/10.1214/18-AOAS1161SF

Meng, X.-L. (2022). Comments on “Statistical inference with non-probability survey samples” – Miniaturizing data defect correlation: A versatile strategy for handling non-probability samples. *Survey Methodology*, *48*(2), 339–360. http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00002-eng.htm

Riley, P. (2019). Three pitfalls to avoid in machine learning. *Nature*, *572*(7767), 27–29. https://doi.org/10.1038/d41586-019-02307-y

Smith, M., Lix, L. M., Azimaee, M., Enns, J. E., Orr, J., Hong, S., & Roos, L. L. (2018) Assessing the quality of administrative data for research: A framework from the Manitoba Centre for Health Policy. *Journal of the American Medical Informatics Association*, *25*(3), 224–229. https://doi.org/10.1093/jamia/ocx078

Tufiş, M., & Boratto, L. (2021). Toward a complete data valuation process: Challenges of personal data. *ACM Journal of Data and Information Quality*, *13*(4), 1–7. https://doi.org/10.1145/3447269

Tukey, J. W. (1986). Sunset salvo. *The American Statistician*, *40*(1), 72–76. https://doi.org/10.1080/00031305.1986.10475361

Wayman, C., & Hunerlach, N. (2019). *Realising the value of health care data: A framework for the future*. Ernst & Young. https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/life-sciences/life-sciences-pdfs/ey-value-of-health-care-data-v20-final.pdf

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, A., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), Article 160018. https://doi.org/10.1038/sdata.2016.18

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology*, *48*(2), 283–311. http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00002-eng.htm

Footnotes

1. There are initiatives toward sharing code as well as data sets that commonly include some information on how these data came to be. There are also data-centric articles that describe data workflows or data processing aspects in a certain research domain (see, for example, Boyd et al., 2023), and some journals provide data-centric manuscript types (see for example https://ijpds.org/calls/population-data-notes). ↩