Patterns

CellPress
OPEN ACCESS

Opinion

# Rethinking data and metadata in the age of machine intelligence

Martin-Immanuel Bittner[1,2,*]
[1]Arctoris Ltd., Oxford, UK
[2]Young Academy of the German National Academy of Sciences, Berlin, Germany
*Correspondence: martin-immanuel.bittner@arctoris.com
https://doi.org/10.1016/j.patter.2021.100208

**A continuous cycle of hypotheses, data generation, and revision of theories drives biomedical research forward. Yet, the widely reported lack of reproducibility requires us to revise the very notion of what constitutes relevant scientific data and how it is being captured. This will also pave the way for the unique collaborative strength of combining the human mind and machine intelligence.**

> It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.—Richard Feynman

Our scientific understanding of the world evolves through iterative feedback loops between theory and experiment. Data informs theory building, theory guides empirical explorations, new data challenges old theories, and so on. Resolving discrepancies between theory and experiment is what propels science forward.

To see how common explanations can be completely subverted by new empirical evidence, consider the transformation in our understanding of duodenal and peptic ulcers. Up until the 1980s, these disorders were ascribed to stress, poor diet, smoking, alcohol, and susceptible genes, with severe forms of the disease sometimes leading to gastrectomy. Then, Barry Marshall and Robin Warren put forth a radically different view of the disease. If they were right, effective treatments for peptic ulcers were readily available in the form of antibiotics. In 1982, Warren discovered a new type of bacteria—*Helicobacter pylori*—capable of surviving and proliferating in the stomach's hostile environment. Biopsying ulcer patients and culturing the bacteria in the laboratory, Marshall built a carefully curated dataset, revealing a strong correlation between ulcers and the presence of *H. pylori*. However, the gastroenterology community was dismissive of his explanation as it challenged conventional wisdom. Marshall then decided to try his hypothesis in the only human patient he could ethically recruit—himself. And indeed, he did develop a peptic ulcer, which he then successfully treated with antibiotics.[1] Marshall's and Warren's hypothesis was vindicated, and in 2005 they were awarded the Nobel Prize for Medicine.

As the scientific enterprise continued to expand over the past decades, the effort has become increasingly data intensive. Driven by Moore's law, data acquisition and processing capabilities have become more powerful and affordable, and this trend is expected to continue. These new technologies enable an exponential accumulation of scientific data in all disciplines. Consider the European Molecular Biology Laboratory data bank, where sequence data doubles every 51 months. Or the Large Hadron Collider, where experiments produce about 90 petabytes of data every year. The published literature is also witnessing a rapid expansion; on average, every two minutes a new scientific paper is indexed in PubMed. This wealth of data holds an enormous potential to help scientists refine their understanding of nature. However, there are critical challenges to address before this promise can truly materialize.

Beyond the quantity of the available data, it is paramount to consider its quality. Arguably, the tools used for data storage and distribution have evolved faster than our tools for and appreciation of full (meta)data capture and data stewardship. This results in the widespread distribution of unduly curated data, which has severe consequences, being at the core of the "reproducibility crisis" affecting several disciplines, especially the biomedical sciences. The magnitude of this issue was brought to the limelight in 2011 when German pharma company Bayer's research teams looked at 4years' worth of target validation projects to find that less than 25% could be reproduced.[2] Often, the reason for this troubling lack of reproducibility was the absence of detailed protocols and (meta)data crucial for experimental execution. This is not an isolated incident. It is estimated that every year, approximately USD 28 billion are spent on irreproducible laboratory-based biomedical research in the United States alone.[3]

Solving this crisis will require more than merely improving data management practices for the results of scientific inquiries. The very notion of what constitutes relevant data must be carefully considered. Frequently, researchers report only the data and procedures they consider essential to support their conclusions. Yet, what might appear to them as inconsequential information—changes in the laboratory's temperature during the experiment, reagent batch numbers, etc.—omitted from the records can and will result in experimental discrepancies. This information (collectively denoted as metadata) must be captured and accounted for to ensure experimental reproducibility.

A *Cell* Commentary[4] portraying world-class laboratories struggling to reproduce each other's results lucidly illustrates the perils of metadata omissions. Two research groups, from Harvard Medical School and UC Berkeley respectively, were collaborating on a project studying the heterogeneous nature of breast cancer. Despite running seemingly identical experiments, they continued to see different results for almost two years. The discrepancy was buried in unreported metadata; at Berkeley, cells were prepared with a shaking platform, whereas at Harvard, more vigorous rotational stirrers were used. This story highlights that encompassing stewardship of data and metadata is an absolute necessity, even more so in this era of data-intensive multi-center collaborations.

Over the past few years, awareness of this challenge has led to efforts to establish standards and procedures for better data and metadata collection and sharing practices. These include the development of the FAIR principles[5] for scientific data management and stewardship. These principles stipulate that experimental data have to be findable, accessible, interoperable, and reusable and that they highlight the critical importance of metadata. Since their publication in 2016, the FAIR guidelines have become an internationally accepted guidebook for increasing transparency and improving reproducibility in research.

Another crucial aspect closely related to FAIR data practices and metadata capture is machine readability. With the exponential increase in scientific data and reports, we are reaching the point where researchers cannot navigate the vastness of available data anymore without the support of machine intelligence. This means that, first of all, databases must be suitable for a new mode of exploration. Moreover, the partial transfer of decision-making from humans to machines makes the question of data quality even more pressing. Artificial intelligence (AI) tools are highly capable of finding patterns in datasets. However, if the data used to train these models is faulty, skewed, or inaccurate, it can easily lead the models astray. Therefore, as AI's role in scientific discovery grows, the question of data quality becomes even more important, and the entire scientific community must make a concerted effort to ensure that only well-curated, reliable, and reproducible datasets are deployed for AI model training.

Undoubtedly, scientific endeavours augmented with AI tools will change how we approach the scientific discovery process in the years to come. Challenges which stood unsolved for decades are now starting to yield to promising new approaches. A noteworthy and very recent example includes the tremendous advances made in predicting protein folding, with DeepMind's AI system AlphaFold 2 showing an impressive performance.[6] Looking at biomedical research more broadly, we can expect the way we discover and develop new drugs to change profoundly with the increasing adoption of AI tools in various parts of the process, from target identification to molecule design to clinical trial recruitment and many others. As one prominent example, Insilico Medicine reported in 2019 that by relying on deep reinforcement learning techniques, they managed to speed up the generation of lead candidates from the pharma average of 1.8 years to just 46 days.[7]

We can expect that the unique collaborative strength emerging from the combined power of the human mind and machine intelligence will lead to impressive advances over the coming years. For this promise to come through, machines will have to play a more active role not just in finding patterns in existing data, but in generating and capturing data in the first place. Using automation to create machine-readable data at scale, with the depth, reliability, and annotation necessary for successful AI deployment, will be game-changing.[8] Furthermore, the increasing adoption of automation will allow scientists to turn their attention from manually conducting experiments to more genuine scientific tasks, such as experiment planning, data analysis and interpretation, communication and discussion of findings, etc. We are facing an exciting new paradigm built on both humans and machines which will accelerate the iterative feedback loops of hypothesis, experiment, and theory that are the engine of science.

## DECLARATION OF INTERESTS

M.-I.B. is a co-founder, shareholder, and director of Arctoris Ltd.

## REFERENCES

1. Marshall, B.J., and Warren, J.R. (1984). Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet *1*, 1311–1315.
2. Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. *10*, 712.
3. Freedman, L.P., Cockburn, I.M., and Simcoe, T.S. (2015). The economics of reproducibility in preclinical research. PLoS Biol. *13*, e1002165, https://doi.org/10.1371/journal.pbio.1002165.
4. Hines, W.C., Su, Y., Kuhn, I., Polyak, K., and Bissell, M.J. (2014). Sorting out the FACS: a devil in the details. Cell Rep. *6*, 779–781.
5. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data *3*, 160018.
6. Jumper, J., et al. (2020). High Accuracy Protein Structure Prediction Using Deep Learning. In Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book), pp. 22–24.
7. Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A., et al. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat. Biotechnol. *37*, 1038–1040.
8. Schneider, P., Walters, W.P., Plowright, A.T., Sieroka, N., Listgarten, J., Goodnow, R.A., Jr., Fisher, J., Jansen, J.M., Duca, J.S., Rush, T.S., et al. (2020). Rethinking drug design in the artificial intelligence era. Nat. Rev. Drug Discov. *19*, 353–364.

## About the author

**Martin-Immanuel Bittner** is the Chief Executive Officer of Arctoris, the automated drug discovery platform that he co-founded in 2016. He graduated as a medical doctor from the University of Freiburg in Germany, followed by his DPhil in Oncology as a Rhodes scholar at the University of Oxford. Martin-Immanuel Bittner has extensive research experience covering both clinical trials and preclinical drug discovery, and in recognition of his research achievements, he was elected a member of the Young Academy of the German National Academy of Sciences in 2018.