Harvard Data Science Review • Issue 8.4, Fall 2026

An Invitation to Dive Into Data

Radu V. Craiu[1]

[1]**Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada**

The MIT Press

Published on: Oct 29, 2026

DOI: https://doi.org/10.1162/99608f92.40aced1c

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)

***Column Editor's Note**: I am thrilled to step in as the editor of Diving Into Data (DivInD),* Harvard Data Science Review*'s column for stories about interesting data, methods, and concepts in data science. My vision for the column is colored by my experience as a method developer and as a collaborator on scientific projects. In these roles, I have managed data features that challenge an analysis, worked with theory strained by the failure of canonical assumptions, redesigned algorithms around computational obstacles, and learned from negative results. In my university teaching, I draw on these experiences, and those of others, to show students the surprising and creative sides of data science. I invite researchers and practitioners to share similar experiences in short, readable articles that stay grounded in the scholarly literature and from which other data scientists can learn. The column aims to make the often-hidden lessons of data science practice visible through rigorous, accessible narratives.*

Keywords: complex models, data, synthetic data, privacy, fairness

Data and Storytelling

The DivInD column presents short articles describing ideas, concepts, methods, and tools used in data science. The overriding aim is that the articles should be enlightening, useful, and accessible, avoiding obscure technicalities. DivInD exists also because more traditional outlets often present data analysis as a clean path from question to collection, model, computation, and conclusion. Anyone who has finished an analysis with practical motivations and consequences knows that this is rarely how things go. We choose what to measure, whom to observe, which imperfections to live with, and how to interpret what passes our filters. Those choices belong to the analysis itself. DivInD articles can make that usually hidden part of data science visible.

The relationship between evidence and narrative is ancient. At Leang Karampuang on the Indonesian island of Sulawesi, a cave-art scene of humanlike figures hunting with a wild pig has been dated to at least 51,200 years ago. It is currently the oldest known surviving example of representational visual storytelling (Oktaviana et al., 2024). It is a reminder that people have long selected, modeled, and preserved observations in forms that others could learn from.

Modern science often treats images, texts, records, and traces of behavior as data. Data science principles are used to turn experience into data and data into insights about the world. We define variables, mark out populations and time windows, compress measurements into summaries, and let models bring some patterns forward while others recede. A scientific story is convincing when these choices can be examined and tied to evidence.

This is particularly relevant now, when we rely more and more on automated summaries, predictions, and explanations. Artificial intelligence (AI) systems can reveal structure that would otherwise be hard to see, but a fluent answer does not settle questions of measurement, selection, or interpretation. These tools are powerful, but they are not magic. They require care, skepticism, refinement, and, thankfully,

human's ability to make the total insight larger than the sum of logical parts. At a time when many scientists and researchers in data science question their place in a quickly evolving AI environment, learning how a system was trained, what it was asked to optimize, where it failed, and how its output and human intervention changed the chain of evidence is useful and belongs in DivInD.

The Changing Challenges of Data Science

Useful lessons in data science appear when quantity and quality pull in opposite directions. A sample does not have to mirror every feature of a population in order to support a valid analysis. We do need to understand the selection mechanism and any differential representation well enough for the question we are asking, and we need to say what that understanding does not cover. Large samples do not make those questions go away. Yet big data are not our only challenge. Anyone who has worked with collaborators in the social sciences, astronomy, public health, ecology, or many other fields knows that large data need not be informative data. The 'Big Data Paradox' shows how noise, missingness, selection, and dependence can turn abundance into effective sparsity (Meng, 2018). For instance, Bradley et al. (2021) showed that during the COVID-19 pandemic, surveys with roughly 250,000 weekly responses substantially overestimated vaccine uptake, while a carefully designed panel of about 1,000 respondents produced more reliable estimates. A DivInD article might describe how an analyst spotted such a selection problem, which diagnostic made it impossible to ignore, or why a data set of modest size proved more informative than a much larger one.

Differential representation can also interact with the target chosen for an algorithm. A widely used health management algorithm showed racial bias because it treated health care cost as a proxy for health need; unequal access made that proxy systematically misleading (Obermeyer et al., 2019). Human genetics research offers another example: genomic studies remain concentrated in populations of European ancestry, which limits both discovery and the transferability of risk prediction (Sirugo et al., 2019). Not every data set must be representative in the same way, but DivInD would be interested in publishing the path—the more sinuous and undulating the better—from population to sample and from scientific goal to computational target.

Synthetic data raise a related set of opportunities and cautions. Simulation, latent-variable models, and generative models have long been part of statistics. What is changing is their scale, their flexibility, and how close they now influence real decisions. Language models can be prompted to generate 'pseudosamples' meant to reproduce response patterns in human populations (Argyle et al., 2023). In biomedicine, generative systems are used to create synthetic health records, predict protein structures (Jumper et al., 2021), and model single-cell expression (Lopez et al., 2018). The outputs can expand what we are able to try, or make restricted data easier to share. Looking like the original data is not enough, however, to establish validity, fairness, or privacy. A recent review found substantial variation in how synthetic medical data are evaluated and noted that privacy risks are often not assessed or are underestimated (Kaabachi et al., 2025).

It may seem easy to separate real and pseudo data, but in high-volume settings that distinction is becoming harder to achieve. Even when we can tell the difference, difficult questions remain. Which features of the original data are preserved? Which are smoothed away or invented? Should synthetic observations be discarded, down-weighted, separately modeled, or integrated with real ones? What uncertainty should travel from the generating model? How exposed are the individuals whose records contributed to it? Differential privacy offers one rigorous way to quantify a privacy–information trade-off (Dwork & Roth, 2014), but putting such ideas to work in a scientific analysis still requires choices with significant consequences. DivInD should be a place where those choices, including the unsuccessful ones, can be examined honestly.

Scale does not make classical ideas obsolete. Design, uncertainty, identifiability, and validation still matter when the data become enormous. What often has to change is the algorithm: exact calculations may give way to approximations, centralized procedures to distributed ones, and statistically ideal solutions to explicit trade-offs among data, time, and risk (Jordan, 2013).

We also continue to face problems in which data are sparse, costly, missing, or highly dependent. In such settings, extracting the information remains both difficult and essential. The challenges of huge data sets and tiny samples often mirror one another. How a team recognized either kind of constraint, and what they did about it, is the kind of contribution DivInD will feature.

Why Diving Into Data?

Research articles and reviews serve distinct purposes: research articles establish theory, introduce methods, or document empirical findings in sufficient detail for scrutiny and replication, while reviews synthesize a literature. Neither format always leaves room for the episode inside a project that actually changed the work: the sampling decision that redirected the analysis, the apparently minor data error that explained a large discrepancy, the algorithm that worked only after an elegant first version failed, or the collaboration that changed the question. DivInD is meant for that episode. Original research is welcome, but it is not required. A contribution may return to published work to draw out a lesson others can use, explain a useful technique through one revealing application, discuss a negative result, or give a cautionary account of an approach that did not work. In other words, tell us what surprised you, what failed, what worked, and why it matters. The goal is to have the reader come away with a clear and useful insight.

The articles will be relatively short, so mathematical detail, code, and technical terminology should appear when they are needed to understand or evaluate the lesson; they should support the narrative rather than bury it.

In a world plagued by short attention spans, a DivInD article is the printed equivalent of a TED talk—concentrated wisdom, compelling narrative, and actionable insights—but one that remains academically rigorous and grounded in the scholarly literature. I very much look forward to reading your contributions.

Disclosure Statement

Radu V. Craiu is on the editorial board of *Harvard Data Science Review* and is the current editor of the Diving into Data column. He has no financial disclosures to share for this article.

References

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(7890), 695–700. https://doi.org/10.1038/s41586-021-04198-4

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4), 211–407. https://doi.org/10.1561/0400000042

Jordan, M. I. (2013). On statistics, computation and scalability. *Bernoulli*, *19*(4), 1378–1390. https://doi.org/10.3150/12-BEJSP17

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Kulynych, B., Prasser, F., & Raisaro, J. L. (2025). A scoping review of privacy and utility metrics in medical synthetic data. *npj Digital Medicine*, *8*(1). https://doi.org/10.1038/s41746-024-01359-3

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, *15*(12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2

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

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Oktaviana, A. A., Joannes-Boyau, R., Hakim, B., Burhan, B., Sardi, R., Adhityatama, S., Hamrullah, Sumantri, I., Tang, M., Lebe, R., Ilyas, I., Abbas, A., Jusdi, A., Mahardian, D. E., Noerwidi, S., Ririmasse, M.

N. R., Mahmud, R., Duli, A., Aksa, L. M., … Aubert, M. (2024). Narrative cave art in Indonesia by 51,200 years ago. *Nature*, *631*(8022), 814–818. https://doi.org/10.1038/s41586-024-07541-7

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell*, *177*(1), 26–31. https://doi.org/10.1016/j.cell.2019.02.048

©2026 Radu V. Craiu. This article is licensed under a Creative Commons Attribution (CC BY 4.0) International license, except where otherwise indicated with respect to particular material included in the article.

References

- Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337–351. https://doi.org/10.1017/pan.2023.2

  ↩

- Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature*, *600*(7890), 695–700. https://doi.org/10.1038/s41586-021-04198-4

  ↩

- Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4), 211–407. https://doi.org/10.1561/0400000042

  ↩

- Jordan, M. I. (2013). On statistics, computation and scalability. *Bernoulli*, *19*(4), 1378–1390. https://doi.org/10.3150/12-BEJSP17

  ↩

- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

  ↩

- Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Kulynych, B., Prasser, F., & Raisaro, J. L. (2025). A scoping review of privacy and utility metrics in medical synthetic data. *npj Digital Medicine*, *8*(1). https://doi.org/10.1038/s41746-024-01359-3

  ↩

- Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods*, *15*(12), 1053–1058. https://doi.org/10.1038/s41592-018-0229-2 ↩
- 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 ↩
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342 ↩
- Oktaviana, A. A., Joannes-Boyau, R., Hakim, B., Burhan, B., Sardi, R., Adhityatama, S., Hamrullah, Sumantri, I., Tang, M., Lebe, R., Ilyas, I., Abbas, A., Jusdi, A., Mahardian, D. E., Noerwidi, S., Ririmasse, M. N. R., Mahmud, R., Duli, A., Aksa, L. M., … Aubert, M. (2024). Narrative cave art in Indonesia by 51,200 years ago. *Nature*, *631*(8022), 814–818. https://doi.org/10.1038/s41586-024-07541-7 ↩
- Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell*, *177*(1), 26–31. https://doi.org/10.1016/j.cell.2019.02.048 ↩