Journal of Data Science 19 (2), 173–177
April 2021
DOI: 10.6339/21-JDS192EDI
Editorial

Welcome to the Philosophy of Data Science Section

Glen Wright Colopy[1,*]
[1]*Independent Researcher, Cary, North Carolina, USA*

1 Introduction

This issue of the Journal of Data Science is the first to feature our new section titled "***The Philosophy of Data Science***". The aim of this section is to highlight the role of critical scientific reasoning in data science by providing dedicated commentary by exemplars in the field. Instead of taking the typical approach of discussing how data science advances science, this section will discuss how the foundations of science underpin data science.

2 What Niche Will This Fulfill?

This section will broaden the discussion of scientific reasoning in data science. Many valuable insights of successful data science projects are under-discussed in the literature. This is due to many factors including, but not limited to, page limits, technical focus, and idiosyncrasies in the review process. A dedicated section prioritizing the coverage of scientific reasoning can give the complexity of this topic its due space while retaining the academic format of journal publications

Invited contributions are encouraged to explicitly state the critical reasoning component(s) in data scientists' scientific process. Critical reasoning is a part of the scientific process that transcends particular scientific domains. Critical reasoning can be studied and replicated.

The section will discuss subjective considerations in data science. Subjective considerations pervade data science but are no less rigorous in the critical thought required. Instead of being swept under the rug, they should be made explicit and justified.

Invited contributions are encouraged to emphasize the process of prioritization in data science projects. Identifying priorities is an essential part of practical execution of data science projects. Many components compete for a data scientist's time and attention. Strong scientific thinkers are clear-minded when prioritizing their efforts. Priorities are frequently set when domain experts and technical experts collaborate.

The section will expand the literature and language of data science. Writing helps the writer organize, evaluate, and articulate his or her thoughts. Reading helps the reader "find the words" for new ideas so that they can critically evaluate those ideas as well. Readers will find the discussion of critical reasoning less abstract and more practical if the discussion comes from experts within the field, whose success confirms the value of the ideas.

I will briefly expand on several of these ideas below.

3 Replicating Critical Reasoning

The reproducibility and replication crisis in science does not just stem from widespread inability to reproduce or replicate scientific results. It also stems from our inability to replicate scientific

* Email: gwcolo@gmail.com.

processes. Critical scientific reasoning is a part of the scientific process that can be appreciated and efficiently replicated by a wide range of data scientists.

Understandably, many researchers are reluctant to take time from their original (more publishable) research in order to replicate another researcher's work. In contrast, many modes of critical reasoning can be replicated (albeit inexactly) to accelerate original research. Inspiration from another researcher's line of reasoning is rarely perceived to reduce the "novelty" of original research. Since the validity, soundness, and/or cogency of critical arguments can be appreciated by a wide range of data scientists, it seems that critical reasoning is a promising area where replicability is possible with few, if any, opportunity costs.

4 Data Scientists Are United by the Same Modes of Reasoning

Data science is a big tent covering diverse technical professions that contribute to a range of application domains. To our detriment, the demand for specific technical solutions causes us to narrow our focus. As we specialize, we feel we that have less in common. In reality, we share many common threads, one being critical reasoning.

There are 3 major forms of critical reasoning used in the day-to-day practice of data science. Deductive, inductive, and abductive reasoning are all daily mental exercises of a data scientist. (The boundaries between these modes are not cut-and-dry. Nor is their number. If you would like a fourth, I would add causal reasoning.) These modes of reasoning are essential to scientific thought with rich histories of intellectual investigation. Unfortunately, terms such as deduction, induction, and abduction rarely surface in the data science literature despite their constant use as a mental process. Many early career data scientists are unaware of these terms at all.

To get everyone on the same page, some informal definitions are:

Deductive Reasoning Reasoning from general premises/principles to a conclusion. Deductive reasoning has the useful property that if an argument is valid and its premises are true then the conclusion is guaranteed to be true. Examples of deductive reasoning include mathematical proofs and hypothesis testing.

Inductive Reasoning Reasoning from specific observations to general statements. Examples of inductive reasoning include inferring population characteristics from a sample, prediction, and forecasting. Cross-validation is an attempt to measure the uncertainty inherent in inductive reasoning by repeatedly measuring how well a rule or algorithm that was learned from one partition of the data set performs on a held-out or complimentary partition of the data.

Abductive Reasoning Inference to the best explanation. Examples of abductive reasoning include model inference (such as maximum likelihood or maximum a posteriori estimation) and the likelihood principle. More broadly, integration over a parameter space could be considered a better explanation than selecting a single set of parameter values. So Markov chain Monte Carlo methods could be considered a form of abductive reasoning as well.

These modes of reasoning may seem abstract at first glance, but they are critical tools in practice. Just as models (like linear and logistic regression) are tools and data scientists are expected to know when each tool is appropriate, data scientists should know which "tool" they

are using to reason critically. Just as models are compared for performance at different tasks, our modes of reasoning should also be compared for their ability to perform certain tasks. The former comparisons are well-covered in data science literature. This section will be a platform for the latter comparisons.

5 Priorities in Data Science

The best scientists are clear-minded when setting priorities in the face of challenging trade-offs. "Bias versus variance" is classic trade-off. "Performance versus interpretability" is another.

There are other trade-offs that are harder (or impossible) to measure. For example,

- Whether an algorithm's convergence guarantees or asymptotic properties matter in the face of empirical performance metrics.
- Whether a model with solid foundational assumptions is preferable to a model with stronger performance but clearly-violated assumptions when deployed "into the wild".
- Whether the requirements of hypothesis-driven science conflict with data-driven science
- Whether to combine multiple layers of simple analyses versus a single complex model to handle the data in its totality.

I have strong opinions on some of these questions and am fairly neutral on others. It's likely that my own opinions would be very successful in some scenarios and disastrous in others. Some phenomenal scientists hold nearly-opposite views to my own and to great success. This section will aim to show how critical scientific reasoning was used to resolve these questions in certain cases.

6 Editorial Format

The section will appear in issues of the Journal of Data Science. Each section will have an introductory editorial followed by one or more invited contributions from world-class scientists whose tool is data science, data analysis, or another technical field under the big tent of data science.

The introductory editorial for each section will provide commentary and discuss key points of the invited contributions. The invited contributions will follow an academic publication format.

To attract the best authors, invited contributions are citable, just like any other journal publication. The editorial and invited contributions are open access to maximize readership. Furthermore, since the invited contributions do not require technical results, authors will be free to opine without forgoing a technical publication in a journal of their choice.

7 Our Inaugural Editorial Piece

I would like to thank Myungjin Kim, Zhiling Gu, Shan Yu, Guannan Wang, and Li Wang for bravely stepping forward to contribute our inaugural invited contribution (Kim et al., 2021)! Their contribution was both different from and better than what I was anticipating, which shows the value of allowing researchers to set their own priorities on what to discuss in their publications. Their article comes at a time when the predictive scientific models are highly scrutinized, both by professional scientists and the general public. Their article contains valuable discussion both for early-career and experienced data scientists, as well as non-scientists who would like a deeper understanding of the intrinsic differences in various COVID-19 forecasting methods.

Kim et al. covers several topics that are essential to a critical understanding of COVID-19 forecasting. I will highlight two topics in particular that, in my opinion, have been under-discussed both within the scientific community and among the general public.

7.1 The Distinction Between Mechanistic and Phenomenological Models

While mechanistic and phenomenological models both attempt to model the course of the pandemic, they differ fundamentally in approach. Simply put, mechanistic models explicitly model the interactions of underlying mechanisms of the pandemic, whereas phenomenological models describe how observed outcomes vary according to observed characteristics. The difference between these approaches carries many similarities to the difference between deductive and inductive approaches to reasoning. Practically, the difference between these two approaches has important implications for whether the model is reliable in the face of violated assumptions, the availability of data, the accuracy of data, and the appropriateness of the model as the true underlying mechanisms of the pandemic change.

7.2 Scientific Inference Using Limited Data

The data used in COVID-19 forecasting is limited in many ways: COVID-19 has no historical data, which hampers inference on transmission characteristics (such as infectivity and seasonality). Furthermore, the most salient pandemic data is definitionally observational and suffers from inconsistent clinical definitions, sampling techniques, granularity, accuracy, and a host of other characteristics that are required for straight-forward analysis. How these data challenges affect specific predictive models is under-discussed and I appreciate the insights that Kim et al. provided when describing these interactions.

The article contains several more interesting critical evaluations on the interplay between data, models, and the application domain. I appreciate that the authors seamlessly integrated their analysis with a review of popular COVID-19 forecasting models to help readers get up to speed. Thank you again to Drs. Kim, Gu, Yu, Wang, and Wang for an excellent inaugural contribution.

8 Welcome to the Philosophy of Data Science Section

I will leave you with a parting thought on why this section will be interesting:

For most data science articles, we already know from the start how things will end. The novel (and well-tuned) model will outperform the older model. The added complexity will result in improved performance... or the performance can be maintained while reducing complexity. The treatment effect will be non-zero. There's nothing wrong with this, I enjoy a success story too.

In this section, I don't know where the discussions will lead or what path they will take. When scientists describe the path they took to success, their reasons and strategies are frequently counter-intuitive. There's a lot going into "the secret sauce" that just isn't covered in the technical publication.

I don't know where this is leading but I look forward to finding out.

I also look forward to you joining me and think that it will be worth your time!

References

Kim M, Gu Z, Yu S, Wang G, Wang L (2021). Methods, challenges, and practical issues of COVID-19 projection: A data science perspective. *Journal of Data Science*, 19(2): 219–242. https://doi.org/10.6339/21-JDS1013