Harvard Data Science Review •

Exploratory Analysis and Its Malcontents

Jeffrey Heer[1]

[1]**University of Washington**

Published on: Jul 30, 2021

DOI: 10.1162/99608f92.3b3cf5be

Over a half-century ago, Tukey and Wilk (1966) launched a rhetorical broadside against purely formal approaches to data analysis. In the intervening decades, software support for statistical modeling and exploratory visualization has made great strides. Despite these advances—or perhaps because of them—we also find ourselves in the midst of a reckoning on issues of analytic transparency, reliability, and replication. Might the time be right for rapprochement? In their thought-provoking article, Hullman and Gelman question the divide between exploratory (EDA) and confirmatory (CDA) data analysis, and offer theoretical directions to inform the future study and development of software tools for analysis.

What Exactly is EDA?

Hullman and Gelman (2021, this issue) share an analogy in which exploratory analysis is akin to a detective "developing hunches," while confirmatory analysis is "likened to a jury deciding whether a defendant is guilty." I find this analogy telling of our current predicament, given how much is left out: going from hunches to evidence collection, weighing and relating the evidence to build up evidentiary arguments, preparing for trial, and so on. If we map this analogy back to the domain of data analysis, it appears that most of the actual 'work' of analysis (of data preparation, quality assessment, operationalization, model building, etc.) lies in-between the given examples.

Confusion further echoes in my work with Leilani Battle (Battle & Heer, 2019), where our literature review finds that data visualization researchers espouse conflicting notions of what does and does not qualify as 'exploratory' analysis. For example, if one has a well-formed goal, is it still 'exploration'? Perhaps a more helpful framing is to consider the diversity of activities that may fall under the umbrella of 'exploratory' work. Such tasks might include:

- Identify specific data points or relations of interest to form an evidentiary chain. For example, tracking a flow of transactions when investigating money laundering.
- Assess data quality (gaps, outliers, etc.) and assumptions prior to modeling. For example, ensuring sufficient coverage of time-series data and checking that variables satisfy the distributional assumptions of intended models.
- Flexibly explore sales data in a visualization tool, producing charts that will subsequently be used as the basis for making business decisions.

While just a small slice of potential tasks, these examples already imply a diversity of needs and failure cases. They range over logical (but not necessarily statistical)

inference, to ‘exposure’ in service of more reliable modeling, to ‘free-wheeling’ exploration with implicit forms of statistical inference from visualized data.

Following the anti-dichotomous lean of the current zeitgeist (e.g., the movement away from null hypothesis significance testing towards estimation-based approaches [Cumming, 2013]), it may be well past time to retire the EDA / CDA distinction, re-focus on the broader and iterative nature of end-to-end analysis workflows (Liu et al., 2020; Gelman et al., 2020), and, as Hullman and Gelman suggest, “strengthen, rather than separate, the links between purely exploratory and model-driven analysis.”

Formalizing Exploration?

Hullman and Gelman propose the use of Bayesian model checks as an organizing theory for inference within ‘exploratory’ analysis tools. They argue that this framework subsumes existing work on graphical inference (Buja et al., 2009), and relate their perspective to applications of Bayesian theories of cognition to visualization evaluation. In order to be useful, the theory need not be an accurate behavioral model of analysis. Rather it can serve as a *normative* model: we might evaluate behavior in terms of how it deviates from theory, or in a prescriptive fashion we might try to instantiate the theory in new software tools.

The latter goal implies that data exploration tools be extended to represent data-generating processes. A tool might fit a reference model (an expectation) based on observed data, sample from this model to get replicated datasets, and then support comparison to the actual data. The authors envision a “grammar” of statistical constructs (not unlike what is provided by model formulae in existing tools such as the lme4 [Bates et al., 2014] and brms [Bürkner, 2017] packages in R) that can be used to specify “pseudo-statistical models that help [analysts] make inferences about real-world phenomena.”

I find this goal worthwhile and exciting, but also share the authors’ concerns about the interface complexity and statistical expertise required. To add to the litany of such quotes, Tukey and Wilk (1996) also wrote that “approaches and techniques need to be structured so as to facilitate human involvement and intervention… Some implications for effective analysis are: (1) it is essential to have convenience of interaction of people and intermediate results and (2) at all stages of data analysis, the outputs need to be matched to the capabilities of the people who use it and want it.” In a quest to bridge ‘exploratory’ and ‘confirmatory’ might the pendulum swing too far, beyond the capabilities of people who might otherwise benefit?

These considerations raise the question of how to design appropriate intermediate representations (such as ‘pseudo-statistical models’) to structure the interaction between an analyst and their software. Indeed, for ‘exploratory’ tasks, many results traditionally remain implicit or informal, for example as realizations within an analyst’s mind. Hullman and Gelman consider exposing statistical modeling mechanisms to support model checks, but perhaps a different starting point, one which tries to make concrete the observations and assumptions accrued both prior to and during exploration, might prove more accessible?

For a reductive, simplistic account, we might consider three forms of expertise on the part of analysts: software expertise (fluency with a graphical tool or programming language), statistical expertise (knowledge of probability theory and statistical modeling methods), and domain expertise. As Hullman and Gelman seek to re-examine ‘exploratory’ analysis relative to both software design and statistical theory, the also-nebulous concept of ‘domain knowledge’ seems due for its own reckoning.

In “Statistical Rethinking,” McElreath (2018) delineates among conceptual hypotheses, causal models underlying hypotheses (e.g., represented as directed acyclic graphs [Pearl, 2009]), and statistical models. The conceptual and causal phases involve ‘domain knowledge’ in terms of the overarching goals of an analysis, the variables considered, and hypothesized causal influences among them. McElreath further notes that, in general, there is not a one-to-one mapping between phases, so a revision to a statistical model may carry unexpected consequences in terms of the causal or conceptual model one is actually assessing. Carrying these notions further, Jun et al. (2021) characterize the process of *hypothesis formalization*, in which an analyst develops conceptual hypotheses and then operationalizes them in statistical models, finding that this process is under-researched and largely unsupported by existing tools.

To advance Hullman and Gelman’s vision, these additional concerns could play a central role. To wit, how might future tools reify conceptual models and aspects of domain knowledge within a software specification, serving to inform and guide subsequent analysis phases? As analysts progress, they may revise or annotate such a specification (e.g., adding previously overlooked causal influences, or noting expected distributions), thus helping to make concrete the results of exploration. This alone could carry benefits for documentation and transparency of flexible (or ‘free-wheeling’) phases of analysis. Concrete conceptual and/or casual specifications could then provide the desired link to statistical modeling (Jun et al., 2021), serving as a starting

point for the 'pseudo-statistical' models envisioned by Hullman and Gelman, and even enabling the partial automation of analysis they ponder at their article's end.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*. https://arxiv.org/pdf/1406.5823.pdf

Battle, L., & Heer, J. (2019). Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in Tableau. *Computer Graphics Forum (Proc. EuroVis)*. https://doi.org/10.1111/cgf.13678

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., & Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906), 4361–4383. https://doi.org/10.1098/rsta.2009.0120

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1), 1-28. https://doi.org/10.18637/jss.v080.i01

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P. C., & Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*. https://arxiv.org/pdf/2011.01808.pdf

Jun, E., Birchfield, M. de Moura, N. Just, R., and Heer., J. (2021). Hypothesis Formalization: Empirical Findings, Limitations of Software, and Design Implications. To appear in *ACM Transactions on Computer-Human Interaction*.

Liu, Y., Althoff, T., & Heer, J. (2020). Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis. *ACM Human Factors in Computing Systems (CHI)*. https://doi.org/10.1145/3313831.3376533

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Tukey, J. W., & Wilk, M. B. (1966). Data analysis and statistics: An expository overview. *Proceedings of the November 7-10, 1966, Fall Joint Computer Conference*, 695–709. https://doi.org/10.1145/1464291.1464366
