Harvard Data Science Review • Issue 3.3, Summer 2021

The Foundation Is Available for Thinking About Data Visualization Inferentially

Dianne Cook¹ Nancy Reid² Emi Tanaka¹

¹Department of Economics and Business Statistics, Monash Data Futures Institute, Monash Business School, Monash University, Melbourne, Australia,

²Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

Published on: Jul 30, 2021

DOI: <https://doi.org/10.1162/99608f92.8453435d>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

[Hullman and Gelman's paper](#) is a refreshing read. Their thinking about how data graphics research, particularly interactive graphics systems, need to evolve in the near future is revealing. Their commentary on the current state of interactive graphics systems is quite discursive, and with a focus primarily on systems that have emerged from research in computer science. The authors are vexed by the lack of a solid modeling framework built into visualization systems, which leads to a lack of rigor in statements made from data plots.

The authors make several recommendations for the design of interactive graphical systems. We applaud the effort, and agree. It is time to tighten up graphical thinking, broadly across discipline boundaries, and to generalize the endeavors into comprehensive theoretical frameworks. Disappointingly, the paper provides little practical guidance on how the recommended paradigms might be realized. For example, one of their primary recommendations is to design systems that conduct explicit model checks such as comparison plots of the data relative to a null or reference distribution. However, they do not address a clear pathway for practical implementation of these model checks, with the result that their recommendations are speculative. Theory needs practice, and usable tools, just as a building cannot be built from an architect's plans alone, but needs construction workers and engineers. We'd really like to see the authors code up their advice and let others give it a spin.

In this response, we discuss Hullman and Gelman's discussion of exploratory and confirmatory analysis, outline existing foundations for graphical inference, and wade into the ideas around a grammar of statistical models.

We feel compelled to note that the authors have omitted some important historical efforts. The software Dataviewer (Buja et al., 1988; Buja et al., 1986; Hurley, 1987) incorporated null, or reference, distributions for comparison with the data, primarily using permutation tests. Developers following Hullmann and Gelman's advice to provide mechanisms for generating reference or null distributions would be advised to investigate this pioneering work. It is the source that motivated the lineup protocol (Buja et al., 2009).

The Interplay of Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA)

The authors (2021, this issue) argue that "scholars have continued to stress a division between EDA and CDA." The reference to Tukey's writings stressing that "exploratory analysis and model fitting go hand in hand" misses the point. Modeling can be both exploratory and confirmatory. In the exploratory sense, modeling can be thought of as applying a camera lens to a data plot, to sharpen the focus, reducing the blur.

Figure 1 illustrates how we might contrast exploratory and confirmatory analysis. The authors are correct that there are close connections between EDA and CDA, and any particular analysis might have some elements of each. Even pure EDA can contain some inferential elements today, given the availability of modern computing

techniques to generate null sets, reference sets, simulations of different scenarios, and probabilities under different dependence relationships. Graphics are only part of these analytic pipelines, and there are several related areas that are not explicitly referenced by Hullman and Gelman: initial data analysis (IDA), model diagnostics (MD) and information visualization (info vis).

Figure 1. An illustration of the distinction between exploratory data analysis (EDA) and confirmatory data analysis (CDA), and the related areas of initial data analysis (IDA), model diagnostics (MD) and information visualization (info vis). The dashed boxes indicate areas where data graphics are common practice. Inference enters these analysis pipelines in different places. Importantly, the starting points for EDA and CDA are different.

The difference between EDA and CDA is most prominent at the outset of an analysis. *EDA starts from the data.* Reading Tukey's classic text one finds it brimming with the sheer pleasure of simply 'scratching down numbers' (today data is more than numbers). This is the essence of EDA, an essence which has become lost in most of what is called EDA today. What most people call EDA today (summarized in Staniak & Biecek, 2019) should be considered to be initial data analysis (IDA), which emerges from the British school of data analysis, with the term coined by Chatfield (1985). Chatfield stresses "the need to see IDA and classical inference as complementary and not as rivals" similar to the Hullman and Gelman viewpoint. For Chatfield, IDA is step four in the analysis process, after clarifying the objectives, collecting data appropriately, investigating the structure and quality of data, and it is followed by carrying out a formal statistical analysis, comparing results with previous findings, and lastly communicating results. As such, IDA is preliminary to the formal statistical analysis, even though, Chatfield does argue that sometimes IDA is all that is necessary. This is very different

from Tukey’s EDA, which is more that of the discoverer, who is handed some exotic beast and tasked with informing us all about its magic, or lack thereof.

When you start from the data, the first step is abstraction (which may seem to be a step backwards) to define what types of variables are present, how the data was collected. The types of variables will inform the types of plots that are appropriate, types of aggregations, or models that might be fitted, to understand the patterns within a variable and between variables. (This is what Tukey might call ‘scratching down numbers.’) It is also a good spot to sit and ponder, and explicitly delineate what one might expect to see when we do apply the techniques to the data. Because this helps to make explicit what would be interesting (alternative hypothesis) and not interesting (null hypothesis, or reference sets). Now in reality, the taxonomy provided by Figure 1 doesn’t reflect that in practice the lines between EDA and CDA can be much more blurred. That often one might start with CDA, and take a detour orthogonal to the objective, and find unexpected patterns among the measurements. Conversely, one starting with EDA might employ the wealth of today’s computational tools to compute how likely it might be to see a particular pattern. The introduction of Cook and Swayne (2007) provides a clear and simple example of EDA and CDA side-by-side, and a discussion of the interplay between the two, and of issues of data snooping, false discovery vs tragedy of non-discovery. It should be noted that model building is not synonymous with CDA, and is often an EDA endeavor. All the data snooping criticisms leveled at data graphics are appropriately leveled also at model building.

An example of the close connection between modeling and EDA can be found in Hand et al. (2000), describing an analysis of credit card transactions for UK petrol station purchases. An interesting pattern of modes at £10, £15, £20, ... is seen from a histogram, and this is followed by building a mixture model to capture this pattern, one expects might be used in the future with new data.

Tight-coupling of modeling and interactive graphics has been the objective of statistical graphics research from the outset. It is reflected in the earliest work of Tukey, and can be seen repeatedly in historical videos available from the ASA video library (ASA Statistical Graphics Section, 2021). The gold standards are XLipsStat (Tierney, 1991), and DataDesk (Velleman, 2012). The emergence of R (R Core Team, 2018) as the next level from S (Becker et al., 1988), actually reflects the tight-coupling of modeling and graphics. What is missing from R is the high-level interactive graphics—a very active current area of effort now by several researchers.

In addition, recent formalisms, including tidy data (Wickham, 2014) and the grammar of graphics (Wickham, 2016; L. Wilkinson, 2005), strengthen inference from EDA using graphics.

Inference and Visualization

It is natural to build theories of graphical inference from classical statistical theory. The most “model-light” form of statistical inference is arguably provided by what Cox and Hinkley (1974) call *pure significance tests*, which require only a test statistic T , a null hypothesis H_0 , and knowledge of the distribution of T under H_0 .

The line-up protocol is an implementation of this for graphical displays; the test statistic is the graphical display of selected from the line-up by an observer, and the null hypothesis is roughly, ‘each graphical display in the lineup is equally likely to be selected,’ leading to a simple binomial distribution with sample size equal to the number of displays k in the line-up and probability of success $1/k$. As with pure significance tests, this simplicity has drawbacks: the conclusions are limited to a statement that the data are or are not consistent with H_0 under the proposed model. For line-up inference this is compounded because there may be several plausible null hypotheses for the same display, as noted by Hullman and Gelman. As well, the conclusion is dependent on the observer, or the expertise of the observer; this complication can be addressed by incorporating subject-specific effects and repeated measurements, as used in Majumder et al. (2013).

The authors propose to move to a more nuanced form of inference by treating each visual comparison as a form of either implicit or explicit model-checking. This fits naturally with the widespread use of, for example, residual plots in regression (MD in Figure 1). It is not clear to us however that all such graphical displays rely on the observer comparing the plot mentally to what might be obtained from a statistical, or even pseudo-statistical model; it seems that lineups are designed exactly to make this more precise, and not leave the comparison in the observer’s head. It also seems a tall order to relate this to a posterior predictive distribution, which requires both a likelihood function and a prior probability distribution, both rather fuzzy in the case of typical statistical graphics, and a relatively complex averaging operation over these functions.

We have been exploring a step intermediate between line-ups and a full, possibly implicit, Bayesian analysis, by considering whether or not we can think of graphical displays as estimators, or perhaps identifiers, of well-defined features. This may enable comparison of different types of graphical displays according to how well, or poorly, the corresponding estimator behaves in repeated samples.

Plausibility of the “Grammar” of Statistical Models

Hullman and Gelman make several mentions of the need for a ‘grammar’ of statistical models, using various phrases—“grammar for model recommendations,” “grammar of flexible yet robust model specifications,” and “grammar of model components,”—acknowledging, however, that the development of such a grammar is a formidable challenge. We interpret these to mean explicitly formulating reference models to use with graphics. This grammar is a critical building block of the authors’ suggestions, yet there is no explicit pathway to achieving such a task. This lack of specificity leaves a large gap to practical implementation. In linguistics, Chomsky (1956) describe the grammar as a finite set of rules that can describe an infinite number of sentences. L. Wilkinson (2005) expanded this notion to graphics to describe a formal object-oriented system that is flexible to describe a large range of graphics. The so-called grammar of statistical models, at the minimum, will require systems of mapping variables in the data to particular models (model specification) and describe ways to encode the method for estimation of the model parameters (model fitting) in a computational system.

A formula system to specify linear models in the R language (R Core Team, 2018) is described in Chambers and Hastie (1992), stemming from G. N. Wilkinson and Rogers (1973), with similar systems employed in other languages, e.g., in *patsy* library (Smith et al., 2018) for Python (Van Rossum & Drake Jr, 1995). This system, however, has limitations in describing more complex models such as mixed models (or multi-level or panel data models) that require additional specification around the variance-covariance structure. Tanaka and Hui (2019) show that the one-to-one correspondence of the symbolic terms to the terms in particular forms of the model equation, aid the user when inputting the desired model into a computational system. They also describe extensions of the symbolic formula for specifying classes of linear mixed models.

The other aspect of model description is specifying the fitting procedure, e.g. least squares, least absolute deviation, M-estimation and so on. The *parsnip* package (Kuhn & Vaughan, 2021), which is part of the *tidymodels* ecosystem (Kuhn & Wickham, 2020) in the R language, is an effort in offering a unified interface to fit a range of statistical and machine learning models but noticeably it does not refer to this system as the ‘grammar.’ In the end, using a finite set of rules to describe the large space of models, and fitting procedures, may be infeasible and may require restricting attention to a small subset.

Beyond the model grammar, the next step is to specify the mapping of model parameters and associated data values to graphical elements. The authors describe this step as a model check primarily with the null or reference distribution. In this sense it is similar to line-up protocol in Buja et al. (2009). A contrasting viewpoint not mentioned by the authors is in Wickham et al. (2015), who suggest that fitted models should more often be visualized in the data space. This is similar to a Bayesian mindset of conditioning on observed data, except surprisingly one hardly ever sees the observed data plotted along with posterior distribution visualizations.

Human Visual System or Interpretation System

Little is said by Hullman and Gelman about subjectivity vs objectivity when reading plots. For example, care is taken when employing the lineup protocol to remove any context of the data: axis labels are translated to generic names like X, Y, titles removed, category levels changed to generic names. The reason for this is that the purpose is to ‘see’ the patterns in the plot, unhindered by prior beliefs or personal prejudices. This was not practiced in example 1 provided by the authors, taken from Nguyen et al. (2020). Subjects in the study are asked to provide an interpretation of structure in a selection of plots (called generalizations), but technically, the results are confounded. The generalizations could be based on prior notions and not on any pattern in the data plot.

Hullman and Gelman discuss the possibility of automating an interactive visual analysis. This needs to be built on a solid foundation. This might involve, as stated by the authors, “setting up a full probability model—a joint probability distribution”—so that inference is based on the entirety of the data, taking into account the joint distribution of the multiple variables. Interestingly, this is also one of the differences seen in the design of

interactive graphics systems relative to static graphics systems. An interactive system does need to expose the full iceberg, not just the surface. Thus, it is a natural fit to think about inference in context of a joint data distribution, here too.

Final Thoughts

Inspired by Norman (1988), the focus of developing interactive graphics systems with an intuitive interface design for humans is important because it allows the analyst to focus on the data, not which button to press or menu to click on. However, as Hullman and Gelman argue, this is not enough. Interactive (and static) data visualization systems need a tight-coupling with statistical thinking, be it Bayesian, frequentist or non-parametric. We look forward to research in this direction, and thank the authors for surveying the literature and prescribing actions to take.

Disclosure Statement

Dianne Cook, Nancy Reid, and Emi Tanaka have no financial or non-financial disclosures to share for this article.

References

- ASA Statistical Graphics Section. (2021). Video Library.
- Becker, R., Chambers, J., & Wilks, A. (1988). *The new s language: A programming environment for data analysis and graphics*. Wadsworth & Brooks/Cole.
- Buja, A., Asimov, D., Hurley, C., & McDonald, J. A. (1988). Elements of a viewing pipeline for data analysis. In W. S. Cleveland & M. E. McGill (Eds.), *Dynamic graphics for statistics* (pp. 277–308). Wadsworth.
- Buja, A., Hurley, C., & McDonald, J. A. (1986). A data viewer for multivariate data. *Computing Science and Statistics*, 17(1), 171–174.
- 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 Statistical Society A*, 367(1906), 4361–4383. <https://doi.org/10.1098/rsta.2009.0120>
- Chambers, J., & Hastie, T. (1992). *Statistical models in s*. Wadsworth & Brooks/Cole Advanced Books & Software.
- Chatfield, C. (1985). The initial examination of data. *Journal of the Royal Statistical Society: Series A*, 148(3), 214–253. <https://doi.org/10.2307/2981969>

- Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory*, 2(3), 113–124. <https://doi.org/10.1109/TIT.1956.1056813>
- Cook, D., & Swayne, D. (2007). *Interactive and Dynamic Graphics for Data Analysis with examples using R and GGobi* [With contributions from Buja, A., Temple Lang, D., Hofmann, H., Wickham, H. and Lawrence, M. and additional data, R code and demo movies at <http://www.ggobi.org>]. Springer. <https://doi.org/10.1007/978-0-387-71762-3>
- Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. Chapman; Hall.
- Hand, D., Blunt, G., Kelly, M., & Adams, N. (2000). Data mining for fun and profit. *Statistical Science*, 15(2), 111–131. <https://doi.org/10.1214/ss/1009212753>
- Hurley, C. (1987). The data viewer: An interactive program for data analysis. [PhD thesis, University of Washington, Seattle].
- Kuhn, M., & Vaughan, D. (2021). *Parsnip: A common api to modeling and analysis functions* [R package version 0.1.5]. <https://CRAN.R-project.org/package=parsnip>
- Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles*. <https://www.tidymodels.org>
- Majumder, M., Hofmann, H., & Cook, D. (2013). Validation of visual statistical inference, applied to linear models. *Journal of American Statistical Association*, 108(503), 942–956. <https://doi.org/10.1080/01621459.2013.808157>
- Nguyen, F., Qiao, X., Heer, J., & Hullman, J. (2020). Exploring the effects of aggregation choices on untrained visualization users' generalizations from data. *Computer Graphics Forum*, 39(6), 33–48. <https://doi.org/10.1111/cgf.13902>
- Norman, D. (1988). *The psychology of everyday things*. Basic Books.
- R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>
- Smith, N. J., Hudon, C., broessli, Seabold, S., Quackenbush, P., Hudson-Doyle, M., Humber, M., Leinweber, K., Kibirige, H., Davidson-Pilon, C., & Portnoy, A. (2018). *Pydata/patsy: V0.5.1* (Version v0.5.1). Zenodo. <https://doi.org/10.5281/zenodo.1472929>
- Staniak, M., & Biecek, P. (2019). The landscape of R packages for automated exploratory data analysis. *The R Journal*, 11(2), 347–369. <https://doi.org/10.32614/rj-2019-033>

- Tanaka, E., & Hui, F. K. C. (2019). Symbolic formulae for linear mixed models. In *Communications in Computer and Information Science: Vol. 1150. Statistics and Data Science* (pp. 3–21). https://doi.org/10.1007/978-981-15-1960-4_1
- Tierney, L. (1991). *LispStat: An Object-Orientated Environment for Statistical Computing and Dynamic Graphics*. Wiley.
- Van Rossum, G., & Drake Jr, F. L. (1995). *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands. <http://www.python.org>
- Velleman, P. F. (2012). DataDesk: an interactive package for data exploration, display, model building, and data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(4), 407–414. <https://doi.org/10.1002/wics.1208>
- Wickham, H., Cook, D., & Hofmann, H. (2015). Visualizing statistical models: Removing the blindfold. *Statistical analysis and data mining*, 8(4), 203–225. <https://doi.org/10.1002/sam.11271>
- Wickham, H. (2014). Tidy data. *The Journal of Statistical Software*, 59(10), 1–23. <https://doi.org/10.18637/jss.v059.i10>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. <https://ggplot2.tidyverse.org>
- Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society: Series C*, 22(3), 392–399. <https://doi.org/10.2307/2346786>
- Wilkinson, L. (2005). *The grammar of graphics*. Springer. <https://doi.org/10.1007/0-387-28695-0>
-

©2021 Dianne Cook, Nancy Reid, and Emi Tanaka. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.