Harvard Data Science Review •

Challenges in Incorporating Exploratory Data Analysis into Statistical Workflow

Jessica Hullman, Andrew Gelman

Published on: Jul 30, 2021

DOI: 10.1162/99608f92.9d108ee6

It is a pleasure to take part in such a lively discussion about the relationship between exploratory and confirmatory data analysis (EDA and CDA), and what is and is not feasible and likely to be valuable in supporting visual analysis. Graphics and data exploration have long been the ugly duckling of statistics even while they have become important aspects of data science, so we are thrilled to see the *Harvard Data Science Review* give this topic a prominent place of discussion.

Several discussants point out that we offer a framework, more of a placeholder for a theory than a theory itself, and no novel graphical or analytical methods. And, indeed, our immediate goal in writing this paper is to not to develop or present new methods so much as to point to the potential value for integration of existing practical and theoretical ideas. To put it another way, when we write, "Designing for interactive exploratory data analysis requires theories of graphical inference," we do not claim to offer any comprehensive theories ourselves, beyond the meta-theory that such a theory would be useful, and a discussion of alternatives and their implications.

Remember, however, that outside the confines of this discussion are the vast majority of practitioners and theorists of statistics and data science, for whom statistics is all about the use of pre-chosen models and analytics tools, with exploratory data analysis playing a minor role at best, producing some graphs before pulling a model out of the toolkit. And separate from them are researchers on graphical perception who run experiments comparing different visualizations but rarely with a clear sense of how this fits into statistical practice outside of exploration, as well as researchers of interactive visual interfaces who design interface idioms, affordances, and optimizations but are not always practicing statisticians. We believe that theories of a fuller integration of exploratory analysis with theories of inference can serve two purposes: it can facilitate the greater use of EDA in data science, and it can point to ways of making research and development in interactive statistical graphics more useful.

As a computer scientist (Hullman) and an author of statistics textbooks (Gelman), we are focused on software paradigms, implementations and defaults as much as on solutions to particular data-analysis problems. This may give our article a slightly different perspective that would be seen in most statistics articles.

Most of the discussants of our paper agree with our goal of tighter integration between exploratory and confirmatory modes of analysis testing in graphical user interface systems for visualization. Several others seem to disagree with our goals or proposal

that interactive visual analysis tools should support model checks as a means of bridging between ‘model-free’ EDA and confirmatory analysis, due to the difficulty of supporting specification of arbitrary reference distributions and the potential cognitive load of interacting with model predictions. We are grateful to the commenters for supplying what we think should be requisite reading alongside our article, as they delve deeper into many of the ideas in our work, raise some challenges, and help us clarify our intentions and delineate where future research focus might be directed.

Characterizing Data Analysis

Neither EDA nor CDA is precisely defined. Exploratory data analysis can range from simple graphics or even seminumerical displays, Tukey's "scratching down numbers," as Cook et al. (2021, this issue) put it, to dynamic multicolored displays, as discussed by Unwin and illustrated by Pfister et al. In addition to the range of ways in which data can be displayed, there are different purposes of plots. Adding to our discussion of conflicting notions of EDA in the visualization literature, Heer suggests, as we do, that there is a need to retire superficial notions of the EDA / CDA divide in favor of a focus on end-to-end workflows.

Cook et al. extend our discussion of phases of exploratory analysis, summarizing differences between Tukey’s notion of EDA and the concept of initial data analysis (IDA). We emphasized Tukey's goal of learning the unexpected (from which we can deduce an important, if implicit, role of *the expected* in interpreting exploratory graphs), but Cook et al. make a good point that EDA graphics are not just about learning the unexpected; they are also about *creating the conditions* by which one can learn new things from data, and to create such conditions does not necessarily require any model or expectations beyond the general sense that a certain set of data is likely to be rich enough to reveal insights in some unknown directions.

Confirmatory data analysis, too, means different things to different people. In the days of Tukey, CDA most likely referred to formal hypothesis testing, and when we say exploratory and confirmatory data analysis are two aspects of the same thing, we are referring to EDA as open-ended implicit model checking and CDA as focused explicit testing. But CDA can also refer to classical or Bayesian estimation and, in a modern data science sense, prediction as confirmed by out-of-sample accuracy. If we think of data science workflow as involving the fitting and checking of a series of statistical models and scientific hypotheses, then the role of EDA in these steps is indeed more complex than the sort of generalized posterior predictive checking we discuss in our paper. As Heer and Fekete discuss, these graphs play another important role in

hypothesis generation. However, we still think that it is useful to formulate EDA relative to 'the expected,' as a model check may, in many subroutines within a branching exploratory analysis, represent the goal state, whether or not the analyst conceives of themselves as conducting implicit model checks.

Multiple discussants share our views on the importance of prior domain knowledge to exploratory workflows. Heer and Pfister et al. second our argument that more explicit representations of relevant prior knowledge are likely to play an important role in the future of interactive data analysis. Fekete and Pfister et al. emphasize the complementarity of visual perception and operations on symbolic representations in analysis workflows, and the fact that some visual patterns are unsuitable for analysis or even explicit articulation. Indeed, the data scenarios that drive a good deal of research aimed at developing novel interactive visualization systems and encodings involve identifying visual signatures facilitated by the parallel processing capabilities of the human visual system, from structural features of complex networks, to multi-scale relationships in genomics data, to conjunctions of features thought to capture cause-effect dynamics in time series. The question of what makes a pattern or deviation from expectation amenable to graphical model checks is important and likely to require research in graphical perception.

Tests of Theory Against Behavior Versus Design Hypotheses

VanderPlas (2021, this issue) writes, "Unfortunately, the theoretical framework for model checking during exploratory and confirmatory data analysis proposed in this paper is just another conceptual and theoretical framework that is difficult to test or falsify as presented," and "[w]ithout [an] empirical analysis, it is very difficult to see what this proposal adds to the two empirical methods discussed within as sub-cases of the model-check system, Bayesian Cognition, and Visual Inference."

Our article poses a question: Is it better to leave it to the analyst using a GUI visualization system to determine if and how they will validate conclusions they draw from visually inspecting graphs, or to build in tools that encourage thinking about possible models to describe the data? From a theoretical perspective, our work proposes a Popperian or Lakatosian interpretation, based on Gelman (2003, 2004), that what an analyst is often already doing is subjecting visualizations to a form of internal model check.

VanderPlas overlooks the ways in which Bayesian cognitive models such as those underlying our proposed model checks actually do yield testable hypotheses. A

Bayesian model checking formalism can drive the generation of new knowledge about graphical inference. In an experimental context, we can propose an instantiation of a Bayesian model (endowing or eliciting prior knowledge) and evaluate its predictions against human behavior (in the form of elicited posterior beliefs). Comparisons between human behavior and model predictions can drive new insights into how people seem to draw inferences from graphs and where they struggle. As Pfister et al. note, it is the differences between analysts' informal mental models and statistical models that might be most fruitful to explore.

This form of testing already occurs in the papers we cite in Bayesian cognition applied to visualization, though that research does not necessarily consider the design implications of such models with regard to interactive graphical model checking tools. The graphical visual inference literature provides examples of testing human judgment against statistical models, though prior beliefs have not been integrated to their formalisms. A Bayesian model check formalism draws from both these bodies of empirical work, emphasizing their relation, and can similarly be used to drive testable hypotheses.

From a design perspective, our work takes the model checking formulation as prescriptive: if we believe that thinking about data generating processes and comparing their predictions to observed data is what good analysts do, then we should design interactive interfaces that encourage analysts to interact with data generating processes. It would be fair to say that we are arguing for system designers and researchers to make less of a sharp distinction between the data and the possible models that might be used to explain it. We consider what it might look like to complement optimizations for visual pattern finding with graphical tools for estimating how plausible patterns perceived in the data are under different assumptions.

Rather than proposing that all visualization be accompanied by strict confirmatory testing as in a NHST paradigm, we argue that being able to interact with predictions from data-generating processes more informally could improve the robustness of inferences drawn using widely available visual analysis tools. As Cook et al. note, model building is often an EDA endeavor. As Heer notes, the goal of system builders is often to identify the appropriate intermediate representations to structure interactions between analysts and their software. We propose providing more explicit support for exploring a space of possible models during visual analysis as a natural bridge between pattern-finding and confirmatory testing.

VanderPlas and Cook et al. request confirmation, through the development and testing of a prototype, of some subset of these hypotheses, and indeed the first author has current projects exactly in that vein. There is much work to be done in designing and testing specific approaches with users—to evaluate the utility of the model specifications that can be supported, the visual mappings between models and visualizations, and possible interaction designs—in order to learn what works and what doesn't. But for the sake of understanding the implications of the interfaces we build, we think it's also important to do the hard work of synthesizing the often conflicting arguments in the literature about EDA versus CDA to identify points of confusion, and to identify broader theoretical frameworks with which to make sense of what works and doesn't and what visual analysis might mean.

In fields like information visualization and statistical graphics it can be easy to produce technical work focused on particular interaction techniques, visual encodings, or systems without jumping up a level to consider what the configurations of these contributions in tools imply about analysis and human goals when working with data. We see the graphical statistical inference work coming out of statistics (including the work of Buja, Cook, VanderPlas, and their collaborators), as an exception, in that these literatures do engage with how visualizations play a role in statistical inference. Part of our intention in writing the article was to provide an entry point into thinking about what it might look like to base empirical research on information visualization, and the design of general purpose visual analysis software, on a formal model of graphical inference. As Cook et al. note, the statistical graphics literature has been engaging with these ideas since the 1980s. Our concern is that the potential applicability of these works to broadly available GUI visual analysis tools is not widely recognized.

Can Imperfect Interactive DGPs Have Value?

VanderPlas makes the important point that it can be difficult to formulate a null generating mechanism for an arbitrary visualization and visual judgment, because our visual systems are so efficient at examining multiple properties of data at once. In other words, it's easy in a lineup context to produce null visualizations that draw the analyst's attention for reasons other than the presence of some target relationship. VanderPlas (2021, this issue) concludes that "[s]oftware which is capable of discerning the specific features the analyst is using to declare a graphic 'surprising' or 'significant' would need to either explicitly ask […] or be psychically linked to the analyst in order to design a null generation model which is suitable for investigating the likelihood of the pattern being real."

The reason that we are not as concerned as VanderPlas about this challenge is that we don't think that a reference model needs to be a perfect expression of a null model for a target visual judgment or set of judgments to have value. We think software designers should consider giving people interactive representations of DGPs to scaffold thinking about DGPs, and that techniques like line-ups and posterior predictive checks overlaid with data in a visualization can have some practical value, even if imperfect.

We also appreciate Cook et al.'s summary of available building blocks (and challenges) in specifying a grammar, which we used as a catch-all to refer to the symbolic representations of models (specifications) and estimation processes, as well as the graphical mappings between visualizations and model specifications. As Cook et al. describe, we should expect the space of supported models to be limited. We are not convinced that this fact should discourage future inquiry into what such a grammar would look like and offer analysts.

The imperfect nature of graphical model checking makes the way that predictions of the model are communicated to the user an essential consideration. We should not be presenting predictions from automatically inferred or elicited models to people as if they are representations of some true data generating process. Rather we should offer them to the user as a collection of crude "golems" (McElreath, 2018) which, while imperfect, represent the building blocks by which we can gain insight into our data in light of questions and expectations we bring to it. Our hope is that system developers and researchers in computer science and statistics won't let the perfect be the enemy of the good when it comes to exploring some of these options.

As we acknowledge in the article, how useful these interactive tools for model checking will be will depend on the analyst's experience. A novice who is relying primarily on superficial visual subroutines to determine what patterns 'matter' has more to gain from being exposed to predictions from possible DGPs (assuming the tool helps them grasp the idea of a process that produces data) than a seasoned analyst who feels comfortable specifying and fitting models. There are important questions regarding visualization literacy, as Fekete points out, and there may be feedback loops between what the analyst is comfortable with and what the GUI system offers. Nonetheless we expect this latter group to be more likely to be familiar with statistical software that makes model fitting easy to begin with, so they make less sense to us as targets for our proposal than users with less statistics experience.

It seems worth noting that our use of the term line-up in the article, outside of specific references to prior work on line-ups, is broader than the typical grid presentation of some number of *null* distribution plots alongside the observed data plot used to approximate null hypothesis significance testing. In writing the article, we use the term line-up to refer to the visualization of a set of views representing samples drawn from *any* reference model, null or otherwise, in line with some prior work in graphical statistical inference. But perhaps our use of this term has led some readers to interpret our goals as centered around perfect null distribution specification, which is not our intention. We are inspired by how line-ups attempt to make implicit reference distribution comparisons to data more precise, but well aware of the various challenges associated with the 'pure significance test' view of a lineup that Cook et al. summarize. Our view is that to be realistic, we shouldn't expect these tools to be perfectly precise, but that lack of perfection may in fact be better for encouraging a modeling-oriented mindset, where the goal is to learn about what assumptions describe the data better or worse, rather than to pose more dichotomous questions.

Moving Beyond Instinctual Arguments about Cognitive Load

All of our commenters mention the question of cognitive load, as do we in our article. We have two responses.

First, we think it's worth thinking creatively about how much we can reduce the mental overhead of interacting with model predictions through expressive and interactive visual interfaces, as this is where information visualization and human-computer interaction research excels. For example, VanderPlas presents an example workflow to illustrate her concerns about the complexity of adding model check subroutines to a visualization system. While her proposed workflow is one valid way to realize model checking, we can imagine simpler workflows that require less explicit knowledge articulation to offer value. Current systems make it relatively easy to switch between encodings and filters; why not also support easy exploration of different candidate model specifications that might capture aspects of the data generating process? For example, the analyst might load their data, generate some views (which is as simple as dragging and dropping variables to shelves in a tool like Tableau, and even simper in other visualization recommenders), perhaps refine one or more views as they attempt to answer a sub-question, then, prior to concluding their exploration on the analysis 'branch,' click on a button to turn on reference model predictions and tabs through a set of automatically-calculated models, which are overlaid on the chart where possible. The available model specifications might include parametric or non-

parametric resampling and model specifications inferred from the visualization specification. Default priors might be inferred from variable domains or previously interacted with datasets that share the same schema, if available. An analyst browsing predictions of different models might observe that none of the default models seem to predict the data well, and think about what it is about the pattern they perceive that is difficult to capture formally. Or they might attend to the fact that a pattern they thought they saw disappears upon replication under different reasonable DGPs, and place less weight on any hypotheses or causal explanations they may have generated to explain the pattern to themselves going forward. Or maybe the analyst notices that one or two models come closer than the others to capturing a difference they perceive in several trends, so they click a 'Model Summary' tab and get the details about these. They might refine the prior, invoking a Bayesian model fitting process. Even if the analyst doesn't choose to follow up their graphical analysis with further modeling, they have explored potential DGPs, to think about what assumptions seem supported by the data. Again, how these predictions are expressed will be important: a good analyst generally considers more than one model specification, and doesn't commit to a model without compelling evidence of fit.

Pfister et al. seem to share our vision of the power that expressive interactive visual interface patterns can have in realizing future visualization tools that enable representation of predictions and beliefs. We agree with their view that interactive data journalism is a natural source of motivation through its demonstration of how, through sketching and simulation, statistical concepts can be made palatable to broad audiences. How users understand and use graphical elicitation and animated simulations are topics being increasingly explored in the interactive visualization literature, providing empirical knowledge around the interpretation and potential effects of these techniques on non-expert analysts.

Our second reaction to concerns about cognitive load is to ask, How might we extend theoretical frameworks for graphical inference to also consider the value of representations of uncertainty like model checks in an analysis workflow? As Fekete (2021, this issue) writes, "Analysts should be able to choose their trade-offs between robustness and time/resources, in an accurate and confident way." We agree: presumably, whether the added cognitive load of interacting with graphical simulations and uncertainty representations is tolerable depends on how much error can be tolerated in the user's answer. How can interactive visualization research and tools more directly address this intuition?

Fekete's discussion of how statistical problems might be decomposed into lower-level operations that can be solved using logical deduction, computation, and graphical inference suggests what level and units of analysis are likely to be important. We can imagine, for example, defining visual queries of different formats (many of which might be specified using the model check framework), and asking, under what conditions would we expect human perception of summary visualizations that employ statistical or visual aggregation be sufficient to answer a particular query? Beyond edge cases (e.g., more data points than available pixels), there is always a choice of how much satisficing can be tolerated, but the visualization literature provides little by way of theory to guide reasoning on this topic (for an exception, see van Wijk [2005]). How do existing GUI visual analysis tools support analysts in confidently deciding which routines are enough? Economic and decision theoretic accounts also seem to have a role to play in the future design of analysis tools.

Formalization in this direction is also likely to be useful in enabling more effective pairings of human efforts and automated efforts, not unlike the growing literature in human-in-the-loop AI on complementarity. Pfister et al. allude to how paying more attention to which statistical tasks the human visual system excels at is a natural step toward machine augmentations that better capture models that may underlie data. Heer describes how given causal representations of analysts' expectations, automating certain analyses may be possible. These are exciting areas of future work.

On Figure Design

Beyond noting the value of domain knowledge and visualization in early stages of analysis, both of which we agree with, Unwin focuses on the figures in our paper, implying that better graphical defaults may be a more important problem than integrating EDA and CDA. We included figures to nod to the different kinds of graphical tools we were discussing, using a range of datasets beyond the research domains we usually work in, to evoke the kinds of data business users of systems like Tableau Software, Power BI, or other programs might use. We intentionally constructed the figures in Tableau, to evoke the design feel of a graphical user interface visualization tool, and did post-hoc editing in image editing software. While in many cases we chose to manually override defaults, in some cases (Figures 4, 5) we failed to get the specifications to our liking on all details. This is as much the result of our (perhaps unwise) choice to work outside of tools in our comfort zone as it is poor defaults as Unwin suggests. We see how the figures might be confusing to readers who interpret them as exact recommendations or examples of what a domain expert would

create. We regret that we did not adopt a practice like sketching-by-hand or wireframing, which designers use to illustrate abstractions rather than low-level details as was our aim in these figures.

Final Comments

In conclusion, we think that both computer scientists and statisticians will be important to the future of visual analysis tools. We are delighted to see, through this discussion, representatives from both camps weighing in on the broader idea of integrating EDA and CDA. We are grateful to all the commenters for their willingness to engage with our ideas.

Acknowledgements

We thank Alex Kale for comments on this response.

References

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

Van Wijk, J. J. (2005, October). The value of visualization. In *VIS 05. IEEE Visualization, 2005.* (pp. 79-86). IEEE.
