Discussion of "Visualizing statistical models: Removing the blindfold"

Prasad Patil and Jeffrey T. Leek

Department of Biostatistics, Johns Hopkins University

April 30, 2015

Wickham and colleagues have provided a nice summary of a data-driven approach to exploratory analysis of statistical models. We would summarize the central themes as:

1. *Plot as much of the raw data as possible and overlay the model fits and parameters.*
2. *Display multiple models from a collection to understand structure in the data.*

Some of the concepts discussed are fairly common in applied data analysis - for example it is common to visualize the boundaries of a decision rule. Other ideas are more infrequently applied, like taking "grand tours" through the high-dimensional data space via a sequence of two-dimensional slices. In general the intuition of displaying as much data as possible is natural among practicing data analysts, but the intuition is frequently based on *ad hoc* experience. This paper is a nice synthesis of these intuitive ideas and provides some examples of new potential visualizations that expose specific features of complex models.

Our intuition agrees with Wickham and colleagues: that displaying more data is better than displaying less. In considering the paper we were motivated by the question: *What is the objective of visualizing models?*. In their culminating example in Section 6, Wickham and colleagues use grand tours and decision boundary plots to understand the operating characteristics of the neural network procedure. The path suggested by this example moves from data visualization, to model fitting, to model visualization, and culminates in model understanding (**Figure 1**, teal path).

We envision this path as a useful approach for teaching about model fitting methods on idealized data sets. For example visualizing the hidden nodes within the neural network (Wickham and colleagues Figure 22) highlights the way that a neural network combines multiple logistic boundaries to arrive at a non-linear classification boundary. This type of plot is an excellent teaching tool for ensembling methods - similar plots are helpful in explaining the AdaBoost boosting algorithm (e.g. Slides 17-19 of http://bit.ly/1EveDiP, re-hosted from: http://webee.technion.ac.il/people/rmeir/BoostingTutorial.pdf). This plot can certainly be used to improve students' understanding of the mechanics of a

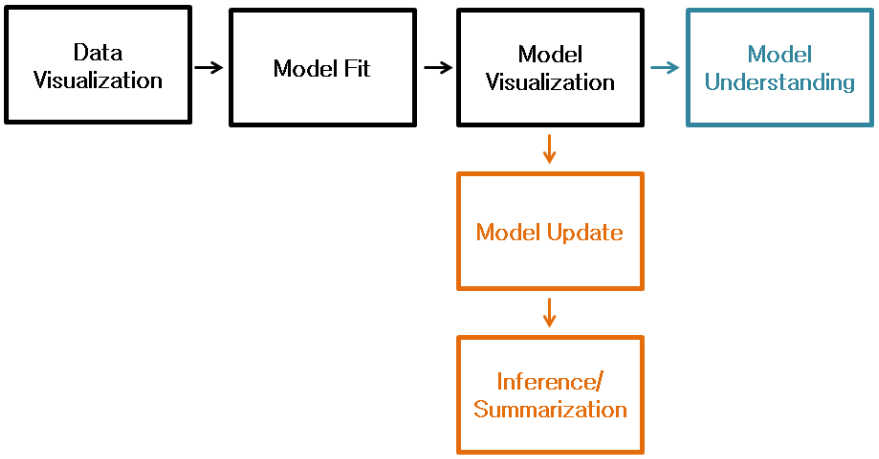

Figure 1: **Flow chart of data analysis incorporating model visualization** The black path represents general steps that are taken in any data analysis with a model visualization step appended.The teal path shows the conclusion presented in the paper, namely that we gain a better understanding of the model fit in our data. The orange path is what we would like to accomplish: use what we learn from visualization to update the model fit and make final inferences.

particular class of models and could unearth ways to improve the algorithm itself.

In our experience, a more typical objective in day-to-day data analysis is to use model visualization for the purpose of checking and updating model fits before summarization (**Figure 1**, orange path). It is less clear to us that the proposals of Wickham and colleagues are suitable for the objective of model correction. Human beings have a difficult time inferring correlations or statistical significance from even simple scatterplots or boxplots [2, 3, 6]. While we agree that showing as much of the data as possible makes intuitive sense, we wonder how people will actually interact with the visualizations suggested in the paper. Plots such as those in Figures 1,6, or 19 of the paper, which present a particular two-dimensional slice of the n-dimensional data space, will perhaps be even more difficult to comprehend.

For similar reasons, visualizing collections of models may not directly address the objective of improving model fit or model inference. There may be many cases where simple, well-understood plots can provide the same insights as visualizing collections. Wickham and colleagues introduce an example from linear models, where six predictor variables from the `NewHavenResidential` dataset in the `barcode` R package [5] are permuted to create a collection of 63 models (note that the `NewHavenResidential` dataset has been moved from the `YaleToolkit` [4] package to the `barcode` package). Figures 12 and 13 in the paper show standardized coefficient estimates from all model fits and employ linked brushing to connect coefficient estimates from the same model and R^2 values for each model. These figures are used to conclude that there may be collinearity between `bedrms`, the number of bedrooms in a home, and `livingArea`, the square footage of a home.

While comprehensive, this collection-based analysis may complicate inference and model selection. No statistical analysis is necessary to reach the conclusion that a larger home is likely to contain more bedrooms. Even in cases where the relationship is not as obvious, we could easily ascertain the same information from a standard scatterplot of the log living area and the number of bedrooms (**Figure 2**). A correlation coefficient of 0.765 would also alert the analyst to check for collinearity.

The compelling work by Wickham and colleagues demands a discussion about "what's next"? We believe the next steps after model visualization are a crucial, and often neglected, component of the data analysis process. The paper makes it clear how a model may be visualized: in the data space, as part of a collection, or stripped to its operating characteristics. But how to proceed in the analysis of the data with these visualizations in hand remains unclear.

We believe, as the authors do, that displaying as much information about the model and the data is a sensible idea. If the goal is to update a model fit and eventually make inferences or summaries from a final model fit, then it is crucial to understand how an analyst will react to these visualizations. We have called this approach to understanding data analyst behavior "evidence based data analysis" (EBDA) [7] and this type of experiment has a rich tradition in the data visualization community [2, 3]. We would be eager to hear how the authors propose to combine the visualizations described in this work with EBDA given the leadership role the authors have played in EBDA for data visualization [8, 1].

Now that Wickham and colleagues have removed the blindfold from statis-

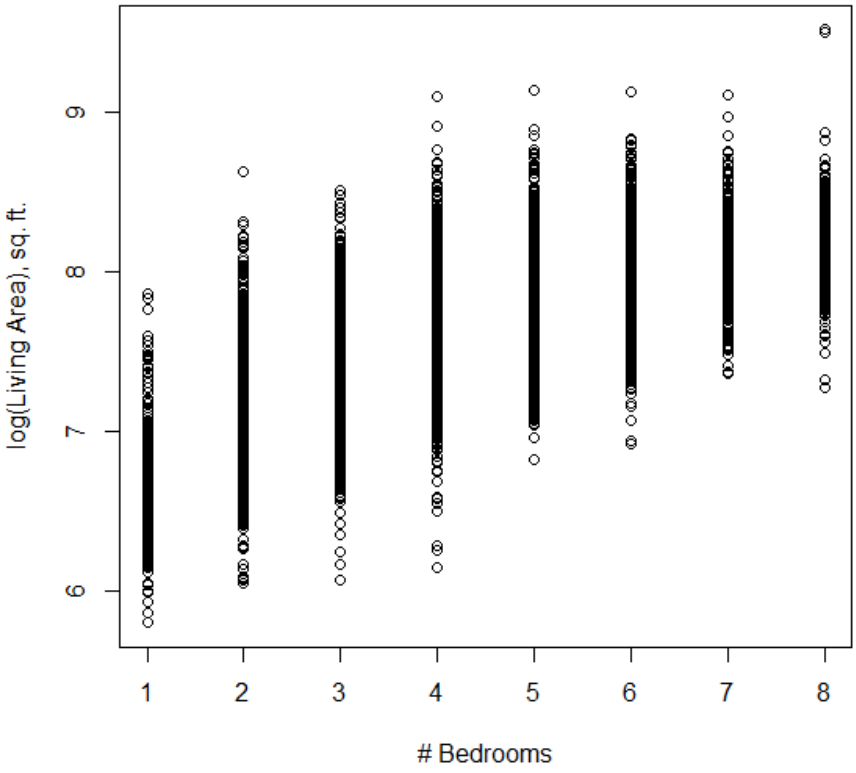

Figure 2: **Scatterplot of log living area against number of bedrooms** From the `NewHavenResidential` dataset in the `barcode` package, which provides information about housing prices and attributes. A potential linear relationship between these two predictors is apparent.

tical model visualization, we need to figure out which path to walk.

References

[1] Niladri Roy Chowdhury, Dianne Cook, Heike Hofmann, Mahbubul Majumder, Eun-Kyung Lee, and Amy L Toth. Using visual statistical inference to better understand random class separations in high dimension, low sample size data. *Computational Statistics*, pages 1–24, 2014.

[2] William S Cleveland, Persi Diaconis, and Robert McGill. Variables on scatterplots look more highly correlated when the scales are increased. *Science*, 216(4550):1138–1141, 1982.

[3] William S Cleveland and Robert McGill. Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833, 1985.

[4] John W. Emerson and Walton A. Green. *YaleToolkit: Data exploration tools from Yale University.*, 2014. R package version 4.2.2.

[5] John W. Emerson, Walton A. Green, and John A. Hartigan. *barcode: Barcode distribution plots*, 2012. R package version 1.1.

[6] Aaron Fisher, G Brooke Anderson, Roger Peng, and Jeff Leek. A randomized trial in a massive online open course shows people don't know what a statistically significant relationship looks like, but they can learn. *PeerJ*, 2:e589, 2014.

[7] Jeffrey T Leek and Roger D Peng. Opinion: Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, 112(6):1645–1646, 2015.

[8] Mahbubul Majumder, Heike Hofmann, and Dianne Cook. Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*, 108(503):942–956, 2013.