Rejoinder

Garrett Grolemund and Hadley Wickham

July 2, 2013

We thank Peter Huber and Leland Wilkinson for their thought provoking and insightful comments. They provide a valuable critique of our model and useful suggestions for future work. We agree with almost everything that Wilkinson says and welcome Huber's disagreements.

Huber takes us to task for not clearly defining sensemaking and its relationship to data analysis, a situation that we are eager to rectify. Sensemaking is a specific algorithm that is associated with Schema theory. Schema theory posits that the human mind stores information as mental models, known as schemas, as described in Section 3 of our paper. The sensemaking hypothesis extends this theory to describe how the mind encodes information into schemas. The hypothesis contends that the mind compares information to schemas with a simple algorithm, Figure 1 of our paper. We refer to this algorithm as sensemaking, which should not be confused with "making sense", a common phrase that can mean many things.

If the sensemaking hypothesis is true, then the sensemaking algorithm is hardwired into the human mind. It can be done consciously or unconsciously, but it cannot be circumvented. In this scenario, sensemaking would serve as the basis of more familiar methods of learning. For example, experimental science is a method that helps us learn because it collects information and identifies schemas that we can use as input in a sensemaking algorithm. Names, animistic worldviews, and theories of physics should not be misunderstood as algorithms for learning. They are only the results of learning, schemas that help us interpret new information.

How should we understand data analysis in relation to sensemaking? The similarities between various descriptions of data analysis suggest that a common algorithm underlies most attempts at analyzing data. We argue that this common algorithm is the sensemaking algorithm, which would make data analysis a subset of sensemaking. However, the human mind cannot apply the sensemaking algorithm to data without modifying the algorithm in predictable ways. As a result, data analysis shares the broad outlines of sensemaking, but differs in the details.

Huber's description of sensemaking as a *counterpart* to data analysis captures this relationship nicely.

The sensemaking algorithm is a counterpart of the data analysis process that sheds light on the process. This light, along with the considerations of Section 4, allows us to deduce a specific model of the data analysis process, which is the model we use in Figure 3.

The difference between data analysis and other instances of sensemaking stems solely from the use of data. Data imposes functional constraints on the sensemaking algorithm, which give it a modified form.

We do not mean to establish a dichotomy of objective data analysis done on quantitative data vs. subjective sensemaking done on qualitative information. First, neither data analysis nor other types of sensemaking can be completely objective because they all rely on the learner's prior experiences through their schemas. Second, there is nothing about data analysis that limits it to quantitative data. In the paper, we emphasize quantitative data because it is easy to see that collections of numbers do not resemble our mental models. However, this is a red herring and Huber is right to chide us for it. Both qualitative and quantitative data can be collected and precisely measured, and qualitative data imposes the same functional constraints on sensemaking that quantitative data does.

This then is the relationship between data analysis and sensemaking: Data analysis overlaps with other types of sensemaking; it differs only in its details, and then only because it uses data. What is it about data that makes sensemaking with data different from sensemaking without data? To put it simply, data surpasses our ability to attend to it.

Apparently, statisticians know that they cannot include more than "seven plus or minus two" different symbols in a graph, but the reason for this has much broader implications than graph construction. G.A. Miller speculated in 1956 that humans could only attend to about seven pieces of novel information at once (Miller, 1956). Studies that followed confirmed this, but placed the "magic number" even lower. For example, Cowan (2000) suggests the number is closer to four than seven.

We do not normally notice how small our attention span is, but it creates a bottleneck when we try to comprehend large, complex chunks of information, such as data sets. Specifically, we cannot easily comprehend new information that contains more than 4 - 7 separate elements that must be considered in relation to each other. This insight has become the foundation of two major theories of learning: Cognitive Load Theory (Sweller et al., 2011) and Multimedia Learning Theory (Mayer, 2001). It can also shed light on data analysis. Statisticians may not be used to thinking of data sets in relation to attention spans, but the two are closely linked. Data sets almost always contain more pieces of related information than we can attend to.

To see how our attention span mediates our understanding of data sets, consider the relationship shown

in Figure A. Two data sets are shown, each containing a pattern. You cannot discern the pattern described in the table on the left until you consider each data point in relation to the others, something that is difficult to do. In contrast, you can immediately see the pattern of a second data set on the right because the data appears in a graph. Why are the results so different? Among other things, visualization sneaks information past the bottleneck in our attention span; the human brain has evolved a separate, high bandwidth pathway for processing visual information (Baddeley and Hitch, 1974).

(4.28, -0.96)	(2.8, 0.18)	(2.33, 0.59)	(2.62, 0.54)
(3.44, -0.12)	(1.59, 0.96)	(4.55, -0.86)	(2.5, 0.85)
(4.04, -0.84)	(3.34, -0.29)	(0.52, 0.39)	(5.67, -0.29)
(4.69, -1.19)	(3.81, -0.65)	(5.5, -0.65)	(1.79, 0.87)
(4.2, -0.94)	(2.77, 0.32)	(5.25, -1.03)	(2.36, 0.82)
(0.47, 0.5)	(5.53, -0.66)	(4.57, -1.03)	(5.52, -0.92)
(4.53, -1.21)	(4.29, -1.02)	(3.08, 0.25)	(2.06, 0.68)
(1.24, 0.78)	(0.76, 0.74)	(4.65, -1.05)	(4.47, -0.8)
(2.89, 0.45)	(0.94, 0.75)	(1.34, 0.97)	(5.95, -0.58)
(3.4, -0.28)	(0.99, 0.81)	(5.91, -0.09)	(0.28, 0.41)
(5.44, -0.75)	(2.49, 0.59)	(4.28, -1)	(1.24, 1.09)
(4.23, -0.9)	(3.21, 0.02)	(5.51, -0.55)	(3.86, -0.73)
(4.07, -0.85)	(1.91, 0.72)	(1.84, 1.13)	(1.1, 1.08)
(4.85, -1.03)	(1.31, 0.94)	(5.25, -0.72)	(0.11, 0.05)
(1.54, 1.15)	(4.71, -0.88)	(2.29, 0.46)	(4.21, -0.83)
(5.76, -0.56)	(4.77, -0.89)	(2.16, 1.11)	(5.04, -1.06)
(0.09, 0.17)	(3.9, -0.61)	(4.12, -0.63)	(0.49, 0.55)
(2.45, 0.75)	(2.46, 0.63)	(3.82, -0.88)	(3.55, -0.45)
(3.66, -0.59)	(1.83, 0.93)	(6.24, 0.05)	(6.25, 0.01)
(4.71, -0.84)	(1.91, 0.93)	(2.05, 1.01)	(5.55, -0.61)
(2.12, 0.69)	(3.67, -0.32)	(3.19, -0.31)	(3.41, -0.26)
(1.45, 0.93)	(3.3, 0.01)	(5.39, -0.75)	(0.67, 0.68)
(3.01, 0.1)	(0.77, 0.58)	(6.11, -0.29)	(2.28, 0.72)
(0.63, 0.7)	(5.47, -0.72)	(3.77, -0.72)	(4.68, -1.06)
(4.46, -1.12)	(0.17, 0.07)	(5.42, -0.56)	(6, -0.23)

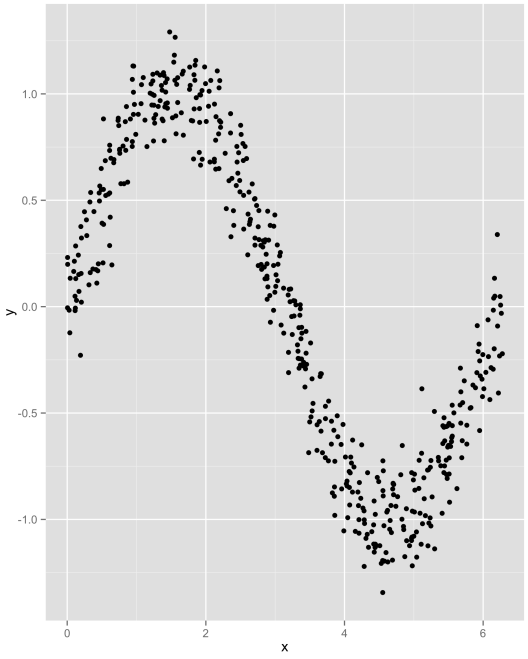

Figure A: It is impossible to attend to enough data points at once in tabular form to comprehend a pattern (*left*). In contrast, we instantly comprehend the pattern of a similar, visualized data set (*right*).

The relationship between data analysis and sensemaking becomes clear if we define data in cognitive terms: *Data is any collection of information that overwhelms our ability to attend to it*. Sensemaking is a mental process that requires us to attend to the information we evaluate. We normally cannot perform sensemaking with a data set because we cannot attend to the complete data set. To perform sensemaking with a data set, we must do one of two things. We can modify the data set in a way that allows us to attend to it. Or, we can use external cognitive aids to help us perform sensemaking without attending to the complete data set. Analysts take the first approach when they reduce a data set to a visualization, a model, or a set of descriptive statistics. Analysts take the second approach when they quantify their schemas into a model or a hypothesis that they can test against the data.

This definition of data emphasizes why it is difficult to compare schemas to data. Our schemas evolve

from information that we attend to: general observations, previously held schemas, and meaningful facts acquired a few at a time. Since we cannot easily attend to data sets, few of our schemas will have evolved to resemble data sets.

This arrangement also hints at why sensemaking should be so ill-suited to handle variation, a feature of the real world. Since sensemaking usually operates on information filtered by our attention, it only encounters isolated pieces of information. Unusual observations are easily dismissed without harmful consequences, and schemas are preserved that adequately explain typical observations.

We like the cognitive definition of data because it parallels the colloquial definition of big data. Big data is a collection of information so large that it cannot be easily stored and processed with a modern computer. Data is a collection of information so large that it cannot be easily stored and processed with the human mind. The definition also distinguishes data analysis from mathematics and much of statistics, fields that focus on logical tasks. It also separates data analysis from computer science, which focuses on computational tasks.

Of course, our ultimate goal is not to know more precisely what data and data analysis are. Our ultimate goal is to successfully analyze data. In sensemaking, learning succeeds when the sensemaker adopts an accurate, useful schema. This is true of data analysis as well. Wilkinson draws attention to an important question of data analysis, how do we find a correct schema when our current schemas will not do? Finding a correct schema is the cognitive equivalent of learning correct knowledge or discovering new truths about the world. In this way, finding a correct schema is a universal goal of all academic inquiry.

We cannot offer a foolproof method for generating correct schemas, but we agree with the many scholars who describe this as a creative task. Beveridge is not the only researcher to view hypothesis generation as an art. Huber also describes data analysis as an art in his critique of our paper, and Karl Popper, the proponent of the hypothetico-deductive model of science, described the generation of theories as a matter of inspiration and creativity (Popper, 1959). We think that this is a savvy description of schema generation. An element of creativity and inspiration underlies all new ideas, including new schemas. However, cognitive science also has much to tell us about schema acquisition.

Sweller (2003) has proposed an evolutionary model for schema generation, where new schemas evolve from old schemas by re-arranging small amounts of information in an iterative fashion. Qu and Furnas (2008) have demonstrated a second way to generate new schemas: subjects in their study instinctively sought out structural relationships in the environment to borrow and use to organize information within a schema. This suggests that we can build potentially useful schemas by analogy. The usefulness of analogies

may underlie Beveridge's comments. One reason visualization may be so useful is that it invites analogies between abstract thoughts and the physical world.

Furthermore, we do not always have to generate our own schemas to assess; we can build a schema from the ideas of others. By including interdisciplinary collaborators in our research, we expand the pool of schemas that we have access to. Our chance of accessing a correct, or at least useful, schema increases with the size of this pool. In a way, organized science is a method of sharing the largest pool of useful schemas possible. (Nonetheless, there is no guarantee that a correct schema is available in the realm of science at any given time, as the occasional scientific revolution demonstrates.)

To admit that there is a creative element to data analysis does not mean that there is no place for theory. Box's aphorism is clever wordplay, but a false analogy. The theory of buoyancy has led to safe flotation devices, which help many people learn to swim. A model of data analysis can help students attempt data analysis in a similar way. Students can rely on the model, and they can use it to assess where they have failed and succeeded. A model can also help professional analysts build new tools and avoid making dangerous mistakes. To extend Box's analogy, you do not learn how to swim from books and lectures, but you might learn how to build a boat, sail it, and keep it from sinking.

Moreover, while generating new schemas is an important topic that deserves future research, there is still work to do in the meantime. In many areas of science and policy, alternative competing schemas already exist. For example, does gun control prevent crime or increase crime in the United States of America? Do minimum wage laws raise unemployment or leave it steady? Research in these areas is often controversial, consequential, and inconclusive. While a new and better schema would be welcomed, we could move forward by judging better between the mental models that already exist. To make progress, we must better understand how to compare schemas against data and how to spot and avoid the pitfalls that lead to faulty conclusions.

References

A. Baddeley and G. Hitch. Working memory. *The psychology of learning and motivation*, 8:47–89, 1974.

N. Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(01):87–114, 2000.

R. Mayer. *Multimedia learning*. Cambridge University Press, New York, NY, 1st edition, 2001.

G. A. Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

K. Popper. The logic of scientific discovery. *London: Hutchinson*, 1959.

Y. Qu and G.W. Furnas. Model-driven formative evaluation of exploratory search: A study under a sense-making framework. *Information Processing & Management*, 44(2):534–555, 2008.

J. Sweller. Evolution of human cognitive architecture. *Psychology of Learning and Motivation*, 43:215–266, 2003.

J. Sweller, P. Ayers, and S. Kalyuga. *Cognitive load theory*. Springer, New York, NY, 2011.