A Cognitive Interpretation of Data Analysis

Garrett Grolemund and Hadley Wickham

August 7, 2012

Abstract

This paper proposes a scientific model to explain the data analysis process. We argue that data analysis is primarily a procedure to build understanding and as such, it dovetails with the cognitive processes of the human mind. Data analysis tasks closely resemble the cognitive process known as sensemaking. We demonstrate how data analysis is a sensemaking task adapted to use quantitative data. This identification highlights a universal structure within data analysis activities and provides a foundation for a theory of data analysis. The competing tensions of cognitive compatibility and scientific rigor create a series of problems that characterize the data analysis process. These problems form a useful organizing model for the data analysis task while allowing methods to remain flexible and situation dependent. The insights of this model are especially helpful for consultants, applied statisticians, and teachers of data analysis.

1 Introduction

This paper proposes a scientific model to explain the data analysis process, which attempts to create understanding from data. Data analysis tasks closely resemble the cognitive process known as sensemaking. We demonstrate how data analysis is a sensemaking task adapted to use quantitative data. This identification highlights a universal structure within data analysis activities and provides a foundation for a theory of data analysis. The proposed view extends existing models of data analysis, particularly those that describe data analysis as a sequential process (Tukey, 1962; Tukey and Wilk, 1966; Box, 1976; Wild, 1994; Chatfield, 1995; Wild and Pfannkuch, 1999; Cook and Swayne, 2007). The paper follows the suggestion of Mallows and Walley (1980) to build on insights from psychology and the examples of Lakoff and Núñez (1997) and Lakoff and Núñez (2000), who documented the influence of cognitive mechanisms on mathematics. The paper was motivated by the authors' need to find criteria on which to compare and optimize the usefulness of data analysis tools; however, the paper's discussion is relevant to all users of data analysis techniques, such as consultants, applied statisticians, and teachers of data analysis.

The paper is organized as follows. Section 2 defines data analysis and explains the shortcomings of the current treatment of data analysis in statistics. Section 3 examines the relationship between cognitive science and data analysis. It outlines areas of cognitive science research that are relevant to the data analysis process, such as mental representations of knowledge, the sensemaking process, and the use of external cognitive tools to complete sensemaking tasks. Section 4 identifies how the use of precise, measured data disrupts the sensemaking process. It then describes the adaptations to general sensemaking that measured data require. Section 5 proposes that data analysis is a sensemaking task adapted to the use of measured data. This provides a theoretic model of data analysis that explains existing descriptions of the data analysis process. In Section 6 we examine a prediction of this model: data analysis inherits the known shortcomings of sensemaking. We examine two of these shortcomings with case studies of well known data analyses. These shortcomings include the tendency to retain false schemas and the inability of sensemaking to prove its conclusions. We conclude by discussing the usefulness of the cognitive model of data analysis as a guiding theory for data analysis.

2 A theory of data analysis

Data analysis is the investigative process used to extract knowledge, information, and insights about reality by examining data. Common data analysis activities include specifying a hypothesis, collecting data relevant to a problem, modelling data with quantitative methods, and interpreting quantitative findings. This process relies on statistics, a field with useful methods for specific data analysis tasks, but has an applied focus; data analysts focus less on the properties of a method and more on the connections between the data, the method, its results, and reality. Data analysis is sometimes referred to as "applied statistics" (Mallows, 1998) or the "wider view" of statistics (Wild, 1994), but we prefer the term data analysis because it does not suggest that statistics is the only tool to be applied when analyzing data.

Data analysis is a widely used technique that is relevant to many fields. This relevance has increased sharply in the past decades as data has become more ubiquitous, more complex, and more voluminous. Large data sets, such as online customer review ratings, social network connections, and mappings of the human genome, promise rewarding insights but overwhelm past methods of analysis. The result is a "data deluge" (Hey and Trefethen, 2003) where current data sets can far exceed scientists' capacity to understand them. Despite this difficulty, the rewards of understanding data are so promising that data analysis has been labelled the sexiest field of the next decade (Varian, 2009).

Future advancements in data analysis will be welcomed by the scientific community, but progress may be limited by the currently sparse theoretical foundations. Little theory exists to explain the mechanisms of data analysis. By theory, we mean a conceptual model that synthesizes relevant information, makes predictions, and provides a framework for understanding data analysis. This definition is more pragmatic than formal: a useful theory of data analysis would help analysts understand what data analysis is, what its goals are, how it achieves these goals, and why it fails when it falls short. It should go beyond description to explain how the different parts of a data analysis, such as experimental design, visualization, hypothesis testing, and computing relate to each other. Finally, a theory of data analysis should allow analysts to predict the success or failure of possible data analysis methods.

It is hard to prove such a theory does not exist, but Unwin (2001) points out that there are few texts and little theory to guide a data analysis. Similar concerns have been expressed by Mallows and Walley (1980), Breiman (1985), Wild (1994), Huber (1997), Velleman (1997), Mallows (1998), Wild and Pfannkuch (1999), Viertl (2002), Mallows (2006), Cobb (2007), Huber (2011) and in the discussion of Breiman (2001). Huber (1997) identifies one reason for the lack of data analysis theory: techniques are developed by researchers who work with data in many different fields. Often knowledge of the technique remains localized to that field. As a result, data analysis ideas have been balkanized across the fields of statistics, computer science, economics, psychology, chemistry, and other fields that proceed by collecting and interpreting data. The subject matter of data analysis is also hard to generalize. The methods of each analysis must be flexible enough to address the situation in which it is applied. This malleability resists a top-down description and led Unwin (2001) to suggest a bottom-up pattern language to stand in for data analysis theory.

A well defined theory of data analysis would provide many benefits. First, it would facilitate the development of better techniques. In many fields, advancements accrue through the extension and development of theories (Unwin, 2001). Advancements in data analysis techniques may lead to many potential rewards. The areas of applications for data analysis have developed more in recent decades than they have during any previous period in the history of statistics (Breiman, 2001). Despite this, many statistics courses still teach methods typical of the first half of the 20th century, an era characterized by smaller data sets and no computers (Cobb, 2007). The development of theory could hasten the speed with which data fields adapt to emerging challenges. A theory of data analysis may also curtail the development of bad techniques. Technology and large data sets do not guarantee useful results. Freedman (2009) argues that "many new techniques constitute not progress but regress" because they rely on technical sophistication instead of realistic assumptions. A better understanding of data analysis will help ground future innovations to sound practice.

A theory of data analysis will also improve the education of future analysts. Statistics curricula have been criticized for teaching data analysis techniques without teaching how or why statisticians should use them (Velleman, 1997). This undermines students' attempts to learn. As Wild and Pfannkuch (1999) explain "the cornerstone of teaching in any area is the development of a theoretical structure with which to make sense of experience, to learn from it and to transfer insights to others." The lack of data analysis theory means that little structure exists with which to teach statistical thinking. As a result, some graduates from statistics programs have been poorly trained for their profession; they know the technical details of statistical methods but must undertake an on-the-job apprenticeship to learn how to apply them (Breiman, 1985; Mallows, 1998). The focus on technique also fails non-statisticians, who are the primary consumers of introductory statistics courses. Without a grasp of statistical thinking, non-statisticians are less likely to recognize the need for a trained statistician and therefore less likely to hire one (Wild, 1994).

A theory of data analysis may also benefit the field of statistics by providing unity and direction. At the end of his 1997 assessment of statistics, Huber predicted that statistics would dissolve as a field unless statisticians replaced their focus on techniques with a focus on "meta-methods" and "meta-statistics" (Huber, 1997). Three years later in 2000, a panel on data analysis called for statistics to evolve into a data science organized by a general theory of data analysis (Viertl, 2002). These conclusions echo Tukey's argument that statistics should be "defined in terms of a set of problems (as are most fields) rather than a set of tools, namely those problems that pertain to data" (Tukey, 1962). A theory of data analysis would offer a unifying theme for statistics and its applications. It would also offer a common language that would promote collaboration by analysts in various fields.

Finally, a theory of data analysis would improve data analysis practice. A theory would aid practitioners because theoretical concerns guide practice (Gelman and Shalizi, 2010). Theory also improves practice; people problem solve more successfully when they "have suitably structured frameworks" to draw upon (Pea, 1987; Resnick, 1988).

Where should we look for such a theory? Many published papers involve a data analysis. But as Mallows and Walley (1980), Cox (2001), and Mallows (2006) point out, most studies do not provide a detailed description of the analysis involved. Instead, they focus on results and implications. We could narrow our focus to statistics and computer science; both fields develop tools to analyze data. However, statistics articles usually focus on the mathematical properties of individual techniques, while computer science articles focus on algorithmic efficiency. As a result, little research deals explicitly with the data analysis process. We propose an alternative source for data analysis insights: cognitive science.

3 The role of cognition in data analysis

Cognitive science offers a way to understand data analysis at a theoretic level. Concerns of cognitive science may seem far from the field of statistics, but they have precedent in the early literature of exploratory data analysis. Tukey and Wilk (1966) highlight the role of cognitive processes in their initial descriptions of EDA (emphasis added): “The basic general intent of data analysis is simply stated: to seek through a body of data for interesting relationships and information and to exhibit the results *in such a way as to make them recognizable to the data analyzer*” (emphasis added). And again, “...at all stages of data analysis the nature and detail of output, both actual and potential, *need to be matched to the capabilities of the people who use and want it*” (emphasis added.) Cognitive concerns also appear in recommendations for improving data analysis. Tukey (1962) suggested that “those who try may even find new [data analysis] techniques evolving ... from studies of the nature of ‘intuitive generalization.’” Mallows and Walley (1980) list psychology as one of four areas likely to support a theory of data analysis.

Cognitive science also addresses a commonality of all data analyses. Data analyses rely on the mind’s ability to learn, analyze, and understand. Each analysis attempts to educate an observer about some aspect of reality. Usually, this requires data to be manipulated and preprocessed, but the end result of these efforts must be a knowledge product that can be interpreted by the human mind. An analysis cannot be useful if it fails to provide this. Even “black box” analyses, which may rely on methods that are incomprehensible to the analyst, must produce a result that the analyst can assign meaning to. If they do not, they will not be useful. This last step of assigning meaning is not a statistical or computational step, but a cognitive one. In this way, each data analysis is part of a larger, cognitive task. The success of each data analysis depends on its ability to interact with this cognitive process.

This alone is good reason for data analysts to learn about cognition. However, cognitive processes also shed insights on the preprocessing stages of a data analysis; mental processes closely parallel the preprocessing stages of data analyses. Moreover, untrained analysts can and do “analyze” data with only their natural mental abilities. The mind performs its own data analysis-like process to create detailed understandings of reality from bits of sensory input. In this section, we examine these mental processes. In Sections 4 and 5 we argue that data analysis is a specific extension of a mental process known as sensemaking.

3.1 Schemas and sensemaking

Studies suggest that the average person can only hold two to six pieces of information in their attention at once (Cowan, 2000). Yet people are able to use this finite power to develop detailed understandings of reality, which is infinitely complex. The mind builds this understanding in a process that is similar to many descriptions of data analysis. The mind creates and manages internal cognitive structures that represent aspects of external reality. These structures consists of mental models and their relationships (Rumelhart and Ortony, 1976; Carley and Palmquist, 1992; Jonassen and Henning, 1996). Mental models have been studied under a number of different names. Examples include frames (Goffman, 1974; Minsky, 1975; Rudolph, 2003; Smith et al., 1986; Klein et al., 2003), scripts (Schank and Abelson, 1977), prototypes (Rosch and

Mervis, 1975; Rosch, 1977; Smith, 1978) and schemas (Bartlett, 1932; Neisser, 1976; Piaget and Cook, 1952). A schema is a mental model that contains a breadth of information about a specific type of object or concept. Schemas are organized into semantic networks based on their relationships to other schemas (Wertheimer, 1938; Rumelhart and Ortony, 1976). This arrangement helps the brain process its experiences: instead of storing every sensory observation, the brain only needs to maintain its schemas, which are sufficient summaries of all previous observations. Some "memories" may even be complete recreations built with a schema (Bartlett, 1932; Klein et al., 2003). Once the brain associates an event with a schema, it can use the schema to access unobserved information related to the event. The mind uses this information to assign meaning to sensory inputs and predict the relationships between data points (Klein et al., 2003). In this way, the mind uses schemas and semantic networks to construct our perception of reality from limited sensory input (Neisser, 1967).

People maintain their schemas in a process known as *sensemaking*. Russell et al. (1993); Klein et al. (2003); Pirolli and Card (2005) and Zhang (2010) have each proposed a description of the sensemaking process. These models vary in their details, but they all contain the same basic components, shown in Figure 1. Variations of this basic model have been utilized by scientists in the fields of cognitive science (Lundberg, 2000; Klein et al., 2003; Helsdingen and Van den Bosch, 2009); organizational studies (Weick, 1995; Weick et al., 2005); computer science (Attfield and Blandford, 2009; Russell et al., 1993); knowledge management (Dervin, 1998); intelligence analysis (Pirolli and Card, 2005); InfoVis (Yi et al., 2008); and Visual Analytics (Wu et al., 2010).

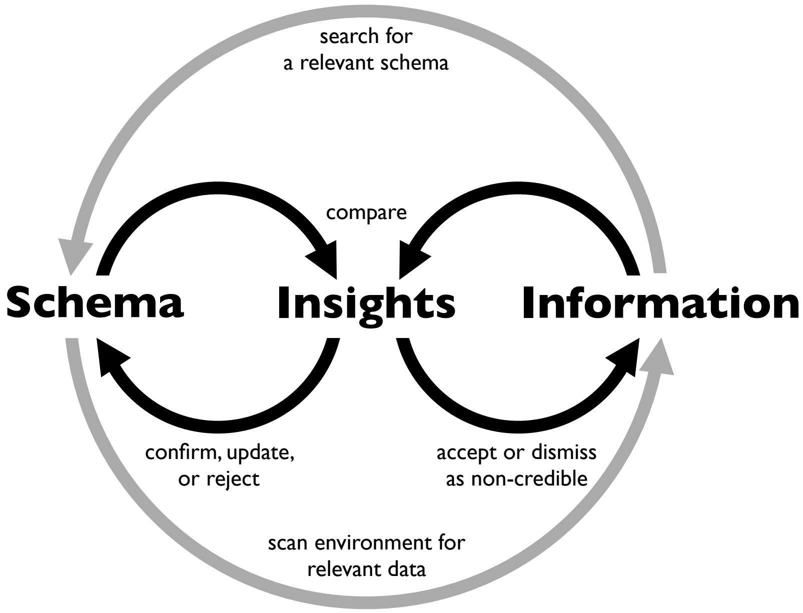

Figure 1: A simplified summary of the sensemaking process. Schemas are compared to observed information. If any discrepancies (i.e, insights) are noticed, the schema is updated or the information is disregarded as untrustworthy.

The sensemaking process revolves around noticing discrepancies between schemas and reality. To understand an event, the brain selects a relevant schema. This selection may be guided by context clues or a few initial observations that serve as anchor points (Klein and Crandall, 1995). The brain then uses this schema to scan the environment for relevant sensory inputs. The schema helps the brain build information from the inputs by assigning meaning to them. This is similar to Moore's definition of data as numbers *that have been given a context* (Moore, 1990). As new information is constructed, the brain tries to fit it into the schema. If a piece of relevant information does not fit, the schema may be flawed. The sensemaking literature calls these unique, non-fitting pieces of information *insights* (Pirolli and Card, 2005). If new information contains no insights, the brain retains the schema as it is. If insights are present, the brain either updates the schema to account for them, dismisses the information as non-credible, or abandons the schema entirely. In the last outcome, the insights and information would then guide the selection of a new schema. This process repeats itself whenever further information becomes available.

Data analysis is a sensemaking task. It has the same goals as sensemaking: to create reliable ideas of reality from observed data. It is performed by the same agents: human beings equipped with the cognitive mechanisms of the human mind. It uses the same methods. Experts in data analysis such as John Tukey and George Box have offered descriptions of the data analysis process. These descriptions show that data analysis proceeds like sensemaking by comparing theory to fact, searching for discrepancies, and modifying theory accordingly. According to Box (1976), "matters of fact can lead to a tentative theory. Deductions from this tentative theory may be found to be discrepant with certain known or specially acquired facts. These discrepancies can then induce a modified, or in some cases, a different, theory. Deductions made from the modified theory now may or may not be in conflict with fact and so on." Tukey's view of data analysis also stresses comparison, discrepancy, and iteration: "Data analysis is a process of first summarizing [the data] according to the hypothesized model [theory] and then exposing what remains [discrepancies], in a cogent way, as a basis for judging the model or the precision of this summary, or both" (Tukey and Wilk, 1966). Both Tukey and Box also emphasize the iterative nature of data analysis and the importance of successive approximations of the truth.

Sensemaking also explains both exploratory data analysis and confirmatory data analysis. Many researchers separate data analysis tasks into exploratory and confirmatory parts (for example, Mulaik (1985), Chatfield (1995)). As Mulaik (1985) explains, "exploratory statistics are usually applied to observational data collected without well-defined hypotheses for the purpose of generating hypotheses. Confirmatory statistics, on the other hand, are concerned with testing hypotheses." In other words, confirmatory analyses focus on a hypothesis (the schema) and seek to validate the schema against data. Exploratory analyses focus on the data and seek to find schemas that explain the data. Many sensemaking descriptions begin with a schema and then proceed to collecting data as in a confirmatory analysis. However, sensemaking can also begin with data and then seek a plausible schema as in exploratory analysis. Qu and Furnas (2008) demonstrates the data to schema direction to sensemaking. In pilot studies of information search tools, Qu and Furnas found that people use sensemaking to develop schemas that explain available data. Early definitions of sensemaking also reflect its bi-directional nature. For example, Russell et al. (1993) define sensemaking as "a process of searching for a representation and encoding data in that representation to answer task-specific questions." To summarize, sensemaking is an integrated, iterative process with multiple points of entry. Exploratory data analysis follows a sensemaking loop that begins with data. Confirmatory data analysis follows a sensemaking loop that begins with a schema (in the form of a hypothesis), Figure 2.

While the general structure of data analysis aligns with sensemaking, its results differ. The results of unguided sensemaking are too unreliable to meet the goals of science. Science requires objective results that can be recreated under consistent conditions. Sensemaking creates

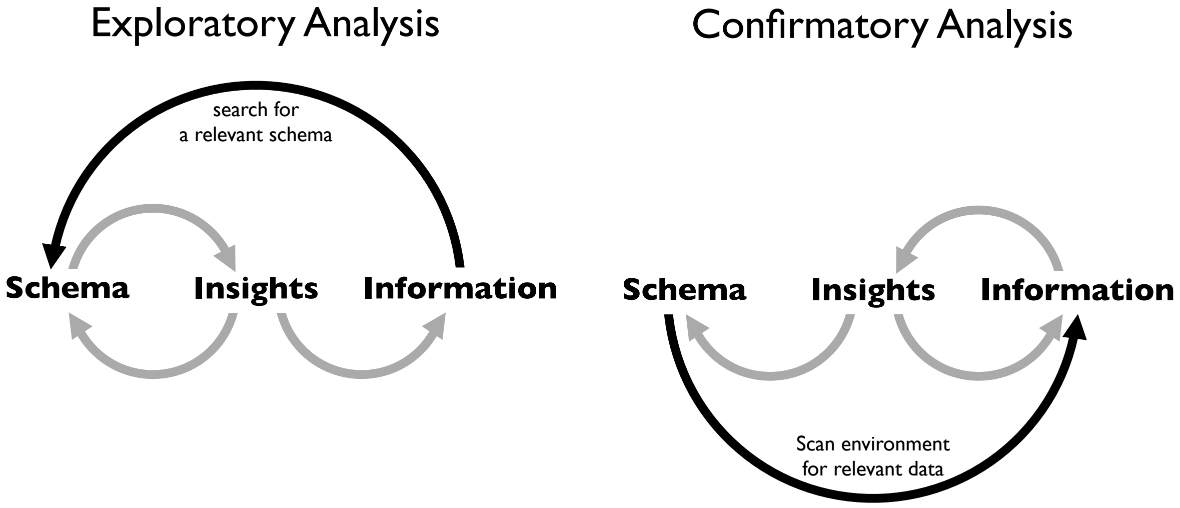

Figure 2: Exploratory and confirmatory data analysis both follow the sensemaking process. Exploratory analysis begins with a received data set. A confirmatory analysis begins with a received schema (often a hypothesis or model).

subjective results that can vary from person to person and from time to time. It is common experience that different people come to different conclusions when presented with the same information. This subjectivity occurs because people have and use different sets of schemas when analyzing information. Unguided sensemaking also has other flaws that increase subjectivity. Tversky and Kahneman (1974) showed that people express predictable biases when they try to make sense of uncertain information. Tversky and Kahneman (1981) showed that meaningless changes in the way information is presented can result in complete changes in the conclusions people draw. These are only two of the most well known biases in human thinking, many more exist. Fortunately, sensemaking can be augmented in ways that reduce these biases and foster objective results. Data analysis is shaped by these augmentations.

3.2 External tools of cognition

We can augment sensemaking with external methods of cognition. The human mind has evolved to rely on external, artificial tools to aid thought (Donald, 1991). These external tools allow us to perform cognitive feats that would not be possible otherwise. For example, a child may use a paper and pencil to perform math that they could not do in their head, or an adult may rely on a written list when grocery shopping (Zhang, 2000). External cognition tools can also be used to reduce the subjectivity of sensemaking. Data analysis relies on two external tools: data, which is an external representation of knowledge; and logic, particularly mathematics, which is an external system for processing information.

External representations of knowledge are information that is stored in the environment. This information can be stored as physical symbols (e.g, written numbers), as relations embedded in physical configurations (e.g, the beads of an abacus or lines on a map), as systems of rules and constraints (e.g, the laws of algebra), or in other ways (Zhang, 1997). External representations play an important role in many cognitive tasks. They can extend a person's working memory, permanently archive large amounts of data, make abstract information directly accessible, suggest solutions by reducing the number of options, and change the nature of a cognitive task to something easier (Zhang, 2000). Well chosen external representations can even provide access to knowledge and skills unavailable from internal representations. For example, the invention of arabic numerals enabled the development of algebra, something that was not possible with roman numerals or purely internal representations of counts. External representations of knowledge guide a sensemaker's attention and give schemas and observations a form that can be shared among sensemakers.

Data analysis relies heavily upon an external representation of knowledge: measured and recorded data. Data provides many benefits that reduce the subjectivity of sensemaking. Recorded data allows large quantities of information to be stored outside of the memory. Here it can be quickly and easily accessed to support cognition. Recorded data can also be manipulated outside of the working memory (e.g. with computers) and shared with other sensemakers. Data is usually collected in a prescribed manner, which reduces the role that schemas play in attending to and interpreting observations. Measurement also allows data to be defined with more consistency and precision than the human senses can supply. Finally, precise measurement facilitates the use of other external cognitive tools such as math and logic.

Systems of rules and constraints can also be external cognitive tools. These systems automate the extraction and transformation of knowledge. As a result, information can be processed outside of the working memory. This allows more data to be processed at once, more complex operations to be performed, and fewer errors to occur during processing. Data analysis relies heavily on math and logic, which are external systems of information processing. Logic and math reduce the subjectivity of data analysis by mandating which conclusions can be drawn from which facts. As Norman (1993) summarizes, "logic is reliable: provide the same information and it will always reach the same conclusion." This is not true of unguided sensemaking. Logic also allows sensemakers to work with entire data sets instead of just the collection of data points they can mentally attend to. As mentioned at the start of this section, the working memory seems to only be able to hold two to six pieces of information at once Cowan (2000). Although, the mind uses various strategies to augment this ability (see for example, Sweller et al. (1998)), the average modern data set exceeds the capacity limits of the working memory. Finally, math and logic allow us to perform our reasoning externally, where we can examine it for errors and biases.

Data analysis can be distinguished from general sensemaking by its reliance on measured data and math and logic. Data and logic reduce the subjectivity of sensemaking. The use of these external cognitive tools makes sensemaking more fit for science, which prefers objective results. Unmodified, our internal knowledge building processes are too subjective to provide these results. Data, in particular, resists the internal forces that create subjectivity. Data reduces the tendency of schemas to screen out observations. Data expands our storage and processing powers. Data can be manipulated and examined externally, which allows us to police our reasoning during sensemaking. But using data introduces new problems: how do we compare abstract schemas to specific, often quantitative, data? How do we identify discrepancies between schema and data when data contains its own type of variation?

4 Making sense of measured data

The sensemaking process must be adapted to accommodate measured data. First, schemas must be made precise to allow comparison against precisely measured data. Schemas must be made quantitative to be easily compared against quantitative data. Second, a test must be developed to identify discrepancies between schema and data in the presence of variation. If data analysis is a sensemaking process, as we propose, each instance of data analysis will exhibit these accommodations. We discuss these accommodations below.

4.1 Abstract schema, quantified data

Sensemaking proceeds by identifying discrepancies between schemas and reality. These two objects must have similar forms to allow accurate comparison. However, schemas do not usually resemble measured data. A typical schema may be as simple as an idea that can be expressed in a sentence or as well developed as what Kuhn (1962) calls a paradigm, a world view that not only contains a theory, but also defines what questions are acceptable, what assumptions are permissible, what phenomena deserve attention, and more. How should an analyst compare schemas against data? The common solution is to deduce a prediction from a schema that can be tested against the data. The predictions can be simple or complex, but they must take the same precise or quantitative form as the data. A linear regression model of the form $Y = \alpha + \beta X + \epsilon$ is one example of a quantified prediction deduced from a schema. A set of data simulated from $Y = \alpha + \beta X + \epsilon$ would be a further prediction from the schema. The underlying schema includes additional non-quantitative information, such as model assumptions, contextual information, and any other beliefs about the subject matter, data sources, and their relationships. The direction of causal relationships and the assumption that there are no lurking variables are two examples of information contained in a schema but not the quantitative hypothesis deduced from the schema.

Data analysis proceeds by testing these quantitative predictions against data in the usual sensemaking fashion. We should not confuse these predictions with the actual underlying schema. They are only deductions that must be true if the schema is true. The validation of a prediction does not validate the underlying schema because the same prediction may also be associated with other competing schemas. This ambiguity is most clear in exploratory data analyses. Exploratory analyses begin with data and then attempt to fit a model to the data. Often more than one model can be fit, which presents one layer of ambiguity. Then the analyst must grapple with a second layer of ambiguity: which explanation of reality (i.e., schema) does the fitted model support? If smoking is correlated with lung cancer, does this suggest that smoking causes lung cancer (schema 1), that lung cancer causes smoking (schema 2), or that a third variable causes both (schema 3)? Analysts can reduce ambiguity by using multiple lines of argument, collecting more data, iterating between confirmatory and exploratory analyses, and deducing and testing as many predictions as can be had from each schema.

Transforming schemas is not the only way to facilitate comparison. Often it is also useful to transform the data to resemble a schema or model. Schemas parallel the way humans think, which rarely involves sets of measured numbers. More often a schema will only describe a characteristic of the data, such as the mean, maximum, or variance. In other occasions, a schema may focus on a variable that must be derived from the data, such as a rate ($count/time$) or density ($mass/volume$). Mathematical calculations can transform the data into the appropriate quantity prior to comparison. Exploratory analysis can be made simpler by transforming data to resemble familiar situations. For example, "curved" scatterplots can be unbent with a log transformation to resemble linear scatterplots. This aids schema search: humans have more schemas to explain familiar situations than they do to explain unfamiliar ones. It also facilitates comparison: humans are better at perceiving differences on a linear scale than a curved one. Visualization is another way to transform data that allows analysts to use their strongest perceptual abilities.

In summary, human cognitive processes are unaccustomed to sets of measured data. To use such data, a sensemaker must transform his or her schemas to resemble data. This can be done by deducing precise predictions from the schema (such as the models commonly used by statisticians). Often it can be helpful to transform the data as well. The need to navigate between schema and prediction/model characterizes all data analyses and distinguishes data analyses from general sensemaking.

4.2 Omnipresent variation

Variation creates a second distinction between general sensemaking and data analysis. Variation in quantitative data is an omnipresent and demonstrable reality (Wild and Pfannkuch, 1999). In usual sensemaking tasks, this variation goes unnoticed. Observers assign observations to general categories (Rosch and Mervis, 1975). Variation is only noticed when it is large enough to place an observation in an unexpected category. Measurement, however, reveals even small variations. These variations disrupt the sensemaking process. A model will appear discrepant with data if it does not account for all of the sources of variation that affect the data. This is not a failure of sensemaking. Afterall, a schema can not be a very accurate model of reality if it does not account for variation that exists in the real world. However, it is unlikely that any model used in data analysis will describe all of the relevant sources of variation. Cognitive, computational, and financial constraints will intervene before every associated variable can be identified and measured. Moreover, many sources of variation will have little to do with the purpose of the analysis. To summarize, the omnipresence of variation in quantitative data derails the sensemaking process. Discrepancy ceases to be an informative signal; unobserved sources of variation will create minute discrepancies between predictions and observations even if a schema correctly describes the relationships between observed variables.

Data analysis proceeds by examining schemas and models that predict a *pattern* of outcomes. This pattern can then be compared against the pattern of the data. Models that predict a pattern do not need to be very complex. Probability theory provides a concise, expressive, and mathematical toolbox for describing patterns. A deterministic model can be transformed into a model that predicts a pattern by adding a probability term. This term acts as a “catch all” that describes the combined effects of all sources of variation that are not already explicitly accounted for in the model.

Comparing patterns changes the task of identifying discrepancies in an important way. To accurately diagnose a discrepancy between two patterns, an analyst must observe the entirety of both patterns, which is rarely an option. The entire patterns may contain a large or even infinite number of points. Research budgets will intervene before the observation can be completed. However, comparing subsets of two patterns can be misleading; a subset of a pattern may look very different than the overall pattern. The data analyst must decide whether or not an observed discrepancy between sub-patterns implies a genuine difference between the entire patterns. This introduces a new step into confirmatory analyses: the analyst must decide whether observed differences between the hypothesis and data imply actual differences between the hypothesis and reality. In exploratory analyses, an analyst must decide how closely to fit a model to the data. At what point does the model begin to fit the sub-pattern of the observed data more closely than the implied pattern of the unobserved data? These variance related judgements provide a second characterization of data analysis.

These judgements become harder when data is contaminated with measurement bias and sampling bias. Both types of bias obscure the true pattern of unobserved reality, which invalidates the results of sensemaking. Bias can be minimized by ensuring that the observed data accurately represent reality and that measurements are made accurately. This may require identifying (but not measuring) all of the data points contained in the pattern, which is sometimes referred to as the population of interest, as well as identifying the relationships between unobserved points and the observed points. These considerations make data collection a more salient part of data analysis than information collection is in sensemaking. Obviously, data analysts can not always control how their data is collected. However, data analysts should always seek out and consider evidence of bias when making variance related judgements. Avoiding and considering bias may be considered a third characteristic of data analysis that distinguishes it from general sensemaking.

5 A conceptual model of data analysis

Data analysis combines sensemaking with two data related considerations: how can we compare abstract schemas to precise data? And, how can discrepancy between schema and data be distinguished from variance? These considerations combine with the general sensemaking structure to create a conceptual model of the data analysis process, see Figure 3. Data analyses proceed as a series of iterations through sub-loops of this process. Individual analyses will vary by the paths they take and the methods they use to achieve each step.

A generalized exploratory task proceeds as follows:

1. Fit a tentative model to available data
2. Identify differences between the model and data
3. Judge whether the differences suggest that the model is misfit, overfit, or underfit (discrepancies)
4. Retain or refine the model as necessary
5. Select a plausible schema that interprets the model in the context of the research

A generalized confirmatory task proceeds in the opposite direction:

1. Select an appropriate schema to guide data collection.
2. Deduce a precise hypothesis from the schema. Multiple hypotheses may be developed to test multiple aspects of the schema.
3. Identify the set of data that would be relevant for testing the hypothesis
4. Collect a representative subset of the data.
5. Identify differences between data and hypothesis
6. Judge whether the discrepancies imply a meaningful difference between the hypothesis and reality or result from random variation or faulty data
7. Confirm, update, or reject the hypothesized model (and its associated schema)

This model parallels the descriptions of data analysis offered by Chatfield (1995), Wild and Pfannkuch (1999), MacKay and Oldford (2000), Cox (2007), and Huber (2011) as well as the

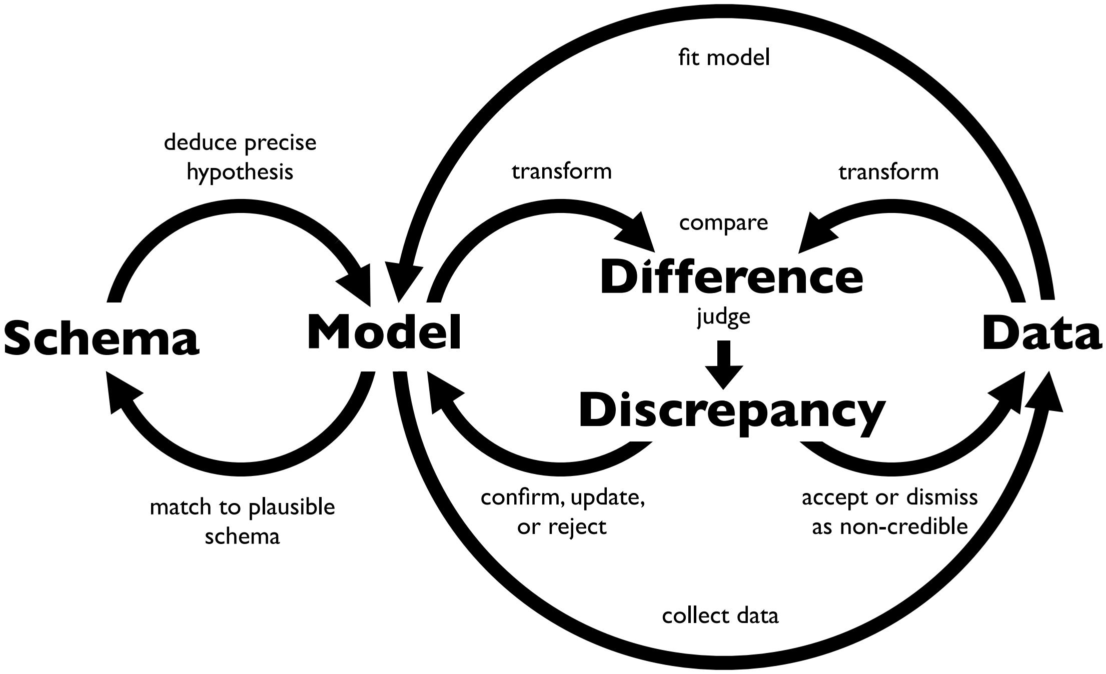
fit model
deduce precise hypothesis
transform
transform
compare
Difference
Schema
Model
judge
Data
Discrepancy
match to plausible schema
confirm, update, or reject
accept or dismiss as non-credible
collect data

Figure 3: Data analysis parallels sensemaking. Analysts deduce a precise hypothesis (model) from the schema, which they compare to the data or a transformation of the data. Analysts must attempt to distinguish discrepancies between schema and data from differences that result from variance and bias. Analysts must also match each accepted model back to a schema to provide interpretation in real world concepts.

description of data analysis offered by Tukey and Wilk (1966), and Box (1976) which we discussed before. The model also lends these descriptions explanatory power: data analysis follows consistent stages because it is a sensemaking process that has been adapted to accommodate data. We briefly discuss the alignment of these descriptions with the cognitive model of data analysis below.

5.1 Chatfield (1995)

Chatfield (1995) divides an idealized statistical investigation into seven stages. As with the proposed model, the methods used in each stage will vary from situation to situation. The seven stages loosely follow our proposed model:

1. Understand the problem and clarify objectives (*begin with a schema*)
2. Collect data in an appropriate way (*collect data*)
3. Assess the structure and quality of the data, i.e, clean the data
4. Examine and describe the data (*transform data into words, visuals, etc.*)
5. Select and carryout appropriate statistical analyses
   (a) Look at data (*transform into visuals*)
   (b) Formulate a sensible model (*make schema precise*)
   (c) Fit the model to the data (*fit model*)
   (d) Check the fit of the model (*identify discrepancies*)
   (e) Utilize the model and present conclusions
6. Compare findings with further information, such as new data or previous findings (*iterate*)
7. Interpret and communicate the results

Many of Chatfield's stages directly map to steps in the cognitive model (shown in italics above). Other stages align with sub-loops of the cognitive model, such as step 3, which requires comparing the data to a schema of "clean" data and then updating the data set. Chatfield's final stage does not match the cognitive model. We agree that communication is an important part of the data analyst's job; however, it occurs after sensemaking has finished. As such, it deals with a different set of cognitive concerns and we refrain from examining it in this article.

5.2 Wild and Pfannkuch (1999)

Wild and Pfannkuch (1999) develop a model of the "thought processes involved in statistical problem solving." This model has four dimensions, but the first dimension is a description of the phases of a data analysis: problem, plan, data, analysis, conclusions (PPDAC). These phases were developed by Mackay and Oldford and later published in MacKay and Oldford (2000). The problem stage involves defining the problem and understanding the context. In these respects, it resembles selecting an initial schema. The plan and data stages involve collecting data relevant to the problem. The analysis stage includes data exploration and both planned and unplanned analyses. These activities search for relevant models and identify discrepancies between the model and the data, when they exist. The final stage, conclusions, encapsulates communicating and using the understanding developed by the analysis. Wild and Pfannkuch (1999) develop connections between data analysis and cognition in other ways as well. They conceptualize applied statistics as "part of the information gathering and learning process."

Wild and Pfannkuch also argue that scientists utilize statistical modeling because we are incapable of handling the enormous complexity of real world systems, which include variation in innumerable components. Modeling provides data reduction, which allows understanding. Schemas play the same role in sensemaking by distilling data and assigning meaning. Wild and Pfannkuch further argue that statistical models become the basis of our mental models, where understanding accumulates, an observation supported by the cognitive model of data analysis.

5.3 Cox (2007)

Cox (2007) discusses the main phases of applied statistics with a focus on technical considerations. Like Chatfield (1995) and Wild and Pfannkuch (1999), Cox divides data analysis into general phases that parallel the sensemaking model: formulation of objectives, design, measurement, analysis of data, and interpretation. The formulation phase parallels selecting a schema. The design and measurement phases focus on acquiring relevant data. The data is analyzed by searching for discrepancies with the model. Cox's interpretation phase focuses on parsing the results of analysis into new understanding. Our model describes this in cognitive terms as matching the accepted model to a schema.

5.4 Huber (2011)

Huber (2011) divides data analysis into the following activities.

1. Planning and conducting the data collection (*collect data*)
2. Inspection (*transform data*)
3. Error checking
4. Modification (*transform data*)
5. Comparison (*identify discrepancies*)
6. Modelling and model fitting (*model fitting*)
7. Simulation (*make schema precise*)
8. What if analyses
9. Interpretation (*match model to a schema*)
10. Presentation of conclusions

Most of these activities directly appear in the cognitive model of data analysis. Other activities, such as error checking, play a general support role to the distinct phases of data analysis that appear in the cognitive model. Like Chatfield (1995), Huber also highlights the important role of communication, which is not covered by the cognitive model. Huber parts with the cognitive model by asserting that "ordering the [above] pieces is impossible." However, Huber's explanation of this agrees with the cognitive model: "one naturally and repeatedly cycles between different actions."

The model of data analysis proposed in this section synthesizes insights provided by prominent descriptions of data analysis. The model explains why these descriptions take the form that they do, and the model provides a framework for understanding data analysis: data analysis is a sensemaking process adapted to measured data. The cognitive model of data analysis also offers an immediate implication for the practice of data analysis, which we discuss in the next section.

6 Implications for data analysis practice

The cognitive model of data analysis predicts a set of problems that may undermine data analysis practice. The mind uses sensemaking to build knowledge of the world, but the process has known flaws. If data analysis is built upon sensemaking as we propose, it will inherit these flaws. Each flaw poses challenges for a data analyst. In this section, we discuss two of these flaws and illustrate each with a case study of a notable data analysis failure.

6.1 Data analysis is biased towards accepted schemas

The nature of cognition tends to undermine the sensemaking mechanism for detecting faulty schema. People only attend to a small portion of the information in their environment, and schemas direct where this attention is placed (Klein et al., 2003). To understand an event, the brain selects a relevant schema. This selection may be guided by context clues or a few initial observations that serve as anchor points (Klein and Crandall, 1995). The brain then uses this schema to scan the environment for additional relevant sensory inputs. The schema then helps the brain build information from the inputs by assigning meaning to them. In other words, schemas determine where attention will be placed and how observations will be interpreted (Klein et al., 2003). Information that contradicts a schema is less likely to be noticed (Klein et al., 2003), correctly interpreted (DeGroot, 1965), or recalled later (Woodworth, 1938; Miller, 1962). As a result, the mind is prone to retain incorrect schemas. This tendency has been well documented in educational research. Students are more likely to misinterpret new information than update their misconceptions. For example, when children are told that the world is round, they are more likely to picture a pancake than a sphere (Vosniadou et al., 1989). High school students are likely to retain an Aristotelian worldview even after completing a year long curriculum in Newtonian physics (Macabebe et al., 2010). Statisticians are not immune to this schema inertia either. The "hot hand" effect in basketball (Gilovich et al., 1985) and the Monty Hall problem (Tierney, 1991) are two well known examples where students (and sometimes professors) have been unable to update their schemas despite statistical training.

The mind tends to discredit observations before beliefs whenever it is easy to do so. A direct experience that requires minimal interpretation is often necessary to impugn an accepted schema. In the classroom, schema change can be initiated by having the student examine their beliefs and then creating an experience that directly contradicts the faulty schema (Bransford et al., 2000). In naturalistic settings, schema change usually does not occur until experience violates expectation, creating a shock or surprise (Klein et al., 2003).

The discovery of the hole in the ozone layer illustrates the inertia that incorrect schemas can have in an analysis. In 1974, Molina and Rowland (1974) predicted that industrial use of chlorofluorocarbons (CFCs) could deplete levels of atmospheric ozone, which could have dangerous environmental effects. According to Jones (2008), NASA's Nimbus-7 satelite began to record seasonal drops in ozone concentrations over Antarctica just two years later. These drops went unnoticed for eight years until the British Antarctic Survey spotted the decrease in ozone through its own measurements in 1984 (Farman et al., 1985). Why did analysis of the Nimbus-7 data fail to reveal such a dramatic phenomenon for eight years?

The Nimbus-7 delay demonstrates the need to address low level schemas during a data analysis. Analysts normally focus on quantifiable models, which are deductions from low level schemas. But it is the cognitive schema that will dictate where analysts direct their attention and how they will explain their findings. These cognitive schemas are particularly dangerous because they often persist in the presence of contradictory information.

NASA programmed the Nimbus-7 to flag observations of low ozone as unreliable, which accords with an initial belief that ozone values should fall in a known range. When NASA scientists encountered these values, the flag made it easy to explain away the data and preserve their schema. Moreover, the unreliability hypothesis was easy to believe because the Nimbus-7 observations depended upon numerous mechanical, electrical, and communication systems. In other words, the observations were collected through a process too complex for the analysts to cognitively comprehend or mentally check. This could explain why the data did not raise any alarm bells; evidence suggests that observations that seem less certain than direct experience will be ineffective for removing faulty schemas.

The BAS team had two advantages on the NASA team. First, the BAS team did not receive a pre-emptive flag of unreliability with their low ozone measurements. Second, the BAS team measured ozone in person in Antarctica and used much simpler equipment than the NASA team. This imbued their observations with the jolt of direct experience, which facilitates schema change. The lack of complexity in the measurement process allowed the BAS team to assign the same confidence to the measurements that they assign to their everyday sensory experiences.

Analysts can not always collect their data in person with simple tools. However, analysts can guard against faulty schemas by addressing the mechanisms that allow them to persist: mis-attention and premature data rejection. Analysts should consider whether or not they have sought out the type of data that would be likely to disprove their basic beliefs should they be wrong. Analysts can further avoid mis-attention by focusing on all plausible schemas. Tukey (1960) advocates for this approach. According to Tukey, science examines “a bundle of alternative working hypotheses.” Conclusion procedures reduce the bundle to only those hypotheses “regarded as still consistent with the observations.” Considering all plausible schemas helps prevent the adoption of a faulty schema, which may then mis-direct an analyst’s attention.

Once data has been collected, analysts should be circumspect about rejecting data. Individuals are prone to reject data as erroneous when it violates their basic ideas about what data should say. However, this prevents the analyst from discovering that their basic ideas are wrong. Data cleaning is a useful and often necessary part of an analysis, but analysts should be wary of using part of a schema under consideration to filter their data. Instead, data points should only be rejected when a source of error can be found in the data collection or generation mechanism.

Finally, we suspect that analysts can approximate the jolt of direct experience by visualizing their data. NASA’s flagged ozone observations were highly structured. They occurred in a temporal pattern (ozone dips low each Antarctic spring and then recovers). They also occurred in the same geographical location in the Southern Hemisphere. We speculate that had the NASA team noticed this by visualizing their data, the pattern would have been as striking as direct experience and prompted a schema change.

6.2 Data analysis does not prove its conclusions

Data analysis inherits a second flaw from sensemaking; it relies on an unsound logical connection between premise and conclusion. As a result, data analysis does not prove its conclusions with logical certainty, and hence, does not completely remove the subjectivity of sensemaking. The reasoning an analyst uses to adopt a schema on the basis of data is as follows:

If schema P is true, data should look like Q
The data looks like Q

Therefore schema P is true

This type of reasoning is not rare, nor is it useless. It is so common in science that it has been given a name: abduction. Abduction was introduced and broadly explored by Peirce (1932). It has been discussed more recently in a statistical context by Rozeboom (1997). Abduction does not prove its conclusions unless there is a one to one mapping between P and Q. More often, alternative schemas R, S, etc. exist that also predict that the data should look like Q. If the data looks like Q, this increases the likelihood that P is true, but it does not rule out the possibility that P is false and R or S is instead true. Yet this is how the human mind functions when sensemaking, and it is how data analysts must function as well.

Data analysts can improve the success of abduction with statistical techniques. Many statistical techniques perform an optimized abduction within a constrained set of models. For example, maximum likelihood estimation chooses the model of the form $P(X = x) \sim f(x|\theta)$ that is most likely to explain the data. However, maximum likelihood does not guarantee that the true explanation is in the set of models of the form $P(X = x) \sim f(x|\theta)$ to begin with. Many statistical methods, such as the method of moments, statistical learning methods, and bayesian estimation methods are all also ways to guide abductive selection. Statistical methods provide a significant advantage over unguided sensemaking. Humans are extremely prone to be biased by emotionally salient information when reasoning about likelihoods (Tversky and Kahneman, 1974). Statisticians frequently use models as tools without assuming the models are true. This mitigates reliance on abduction. Nevertheless, the abductive nature of data analysis requires caution and corroboration before data is used to make weighty decisions.

The space shuttle *Challenger* accident demonstrates the need to strengthen abductive arguments with further analysis. NASA decided to launch the space shuttle in 31°F weather despite worries that the shuttle's O-rings would leak at that temperature. The O-rings failed, killing all aboard. Prior to launch, engineers from NASA and Morton Thoikol, the manufacturer of the space shuttle, examined data on the relationship between O-ring failure and temperature. They concluded that no relationship existed. This analysis has been widely scrutinized and criticized (see for example, Dalal et al. (1989), Tufte (1997), Presidential Commission on the Space Shuttle *Challenger* Accident (1986), etc.). However, the data that NASA examined *could* be construed to support the belief that temperature does not affect O-ring performance. The seven data points that NASA examined could be seen as random cloud that does not vary over temperature, Figure 4 (top). Alternatively, they could be seen as a parabola that suggests increasing danger at extreme temperatures. This is the nature of abduction, it does not rule out competing plausible explanations.

To strengthen its conclusions, NASA should have sought to corroborate its view with a second line of evidence. NASA had access to 17 additional data points from shuttle launches that it could have examined. These points would have cast doubt on NASA's conclusion; a trend between O-ring failure and temperature appears when the additional data points are considered, Figure 4 (bottom).

Even more analysis, however, may have been needed to avert disaster. Lavine (1991) points out that any attempt to predict performance at 31°F from the data would be an extrapolation since the observed data all occurred between 53°F and 81°F. In other words, multiple models could be fit to the existing data and each would predict the performance at 31°F differently. In fact, a careful statistical analysis that considered the leverage of each available data point could potentially be seen as evidence that 31°F would increase the risk of O-ring erosion, but not by enough to pose a severe danger (Lavine, 1991).

In summary, abduction even assisted by statistics could not differentiate between an eventless launch and catastrophe based on the available data. Data analysis could be used to support either argument, although the argument for danger would have appeared stronger. To validate

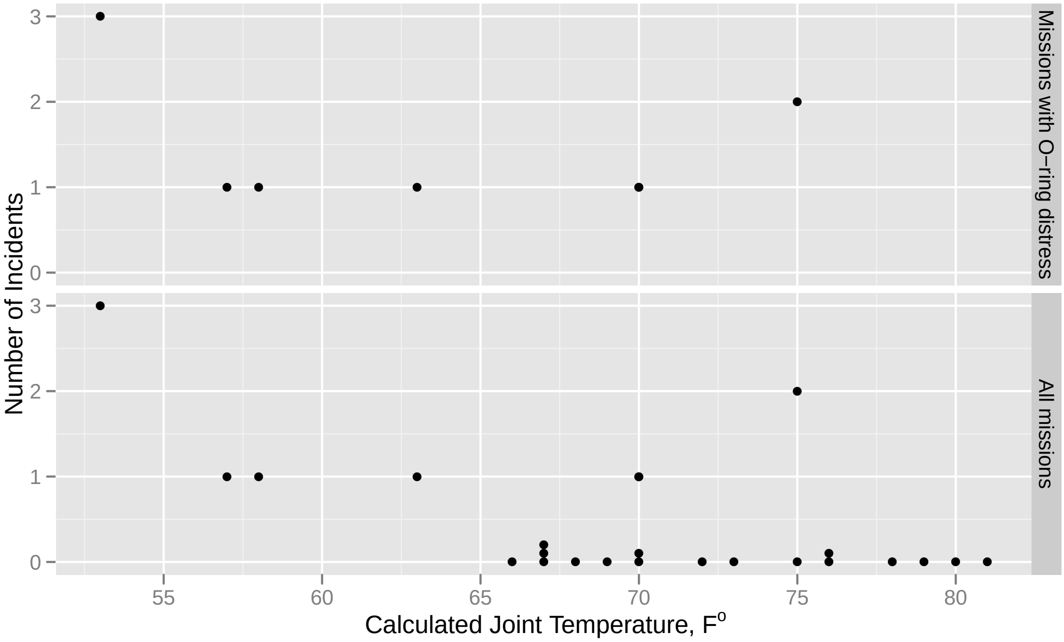

Figure 4: The seven flights examined by NASA and Morton Thoikol managers (above). The 24 flights for which information was available (below). Recreated from Presidential Commission (1986) (p. 146).

one of the arguments, NASA and Morton Thoikol would have had to collect new data near 31°F that could distinguish between competing models. Such data was collected during the investigation of the *Challenger* disaster. A controlled experiment of O-ring performance at different temperatures reported by Presidential Commission on the Space Shuttle *Challenger* Accident (1986) (p. 61–62) demonstrated conclusively that O-rings would not perform safely at 31°F.

In general, analysts can avoid trouble by acknowledging the abductive nature of data analysis. Researchers should view an analysis as an argument for, but not proof of its conclusions. An analyst can strengthen this argument by judging the strengths and weaknesses of the argument *during the analysis* and adjusting for them. An analyst can also continue an analysis — often by collecting new data — until one schema appears much more plausible than all others.

Controlled experiments, expertise, and corroboration can also be used to strengthen the abductive step of data analysis. An experiment can be designed to limit the amount of plausible schemas that can be abduced from, which increases the likelihood of success. Subject matter expertise provides the analyst with more relevant schemas to select from, which allows a better mental approximation of reality. Expertise also helps ensure that the analyst will know of a correct schema, which is a prerequisite for selecting one during abduction. Expertise also broadens the amount of previous information and data that the analyst can utilize when matching a schema. Finally, an independent line of argument can corroborate the results of an abduction if it comes to the same conclusion.

7 Conclusion

This paper identifies data analysis as an extension of the internal cognitive processes that build knowledge. In particular, we propose that data analysis is a sensemaking process that has been modified to use precisely measured data. This improves the performance of sensemaking, but creates a new set of problems that exist in every data analysis. Every data analysis must choose a way to express abstract concepts precisely (often quantitatively). Every data analysis must also find a way to identify discrepancies between a schema and reality in the presence of variation. These problems characterize data analyses and give them a recognizable pattern. Moreover, data analysis inherits weaknesses from the sensemaking processes upon which it is built. In this paper, we identify two such weaknesses: the unusual persistence of false schemas and the unavoidable subjectivity of model and schema selection.

We began this paper by pointing to the need for a formal theory of data analysis. Does the cognitive model of data analysis qualify as a formal theory of data analysis? Perhaps not. Philosophers of science have offered multiple definitions of a scientific theory. These range from the axiomatic to the semantic and usually require a degree of mathematical precision that our conceptual model does not offer. However, the cognitive model of data analysis meets our pragmatic view of a theory. It offers an explanatory framework for data analysis that synthesizes available information and makes predictions about data analysis tasks.

The cognitive model of data analysis may not change the way data analysis is practiced by experienced statisticians. We believe that the prescription offered by the model is very similar to current expert practices. The value of the model lies instead in its portability. Current methods of statistical training have been criticized because novices must acquire years of experience before they settle into expert data analysis practices. In contrast, the cognitive model can be taught to novice statisticians to guide data analysis practices from the get go. It is easy to understand that a data analysis seeks to minimize discrepancies between theory and reality. It is easy to accept that the mind goes about this in an innate way. It is also easy to see that this task can be hindered by cognitive, logistical, and epistemological obstacles. The details of data analysis emerge as these problems arise and are overcome.

The cognitive model also provides a way to unify the field of statistics, as advocated by Huber (1997); Viertl (2002) and others. The model focuses on cognition, but it does not ignore the contributions of statistics to data analysis. Instead it organizes them. Statistical pursuits can be associated with the steps of data analysis that they perform or support. Individual techniques of data analysis, such as design of experiments, data visualization, etc., can be categorized and criticized by identifying which problems they solve. This arrangement highlights how different areas of statistics interact with each other. It also provides a global framework for students trying to master the field of statistics.

The cognitive model also offers guidance for adapting data analysis to new contexts. Small sample statistical methods may become less applicable as the size and nature of data sets change, but the general structure and challenges of data analysis will remain. The cognitive model identifies these challenges: analysts will need methods that facilitate comparisons between data and schema and allow judgements of dissimilarity in the presence of variation. Analysts will need ways to develop abstract schemas into precise models that describe patterns of observation, and they will need guidance for transforming the best fitting models into real world explanations.

Finally, a cognitive interpretation of data analysis also offers a way to improve current data analyses. A cognitive view suggests that cognitive phenomena may adversely affect data analysis – often in unnoticed ways. We examined two such effects in Section 6. Other cognitive phenomena with other effects should also be looked for. Each would provide new opportunities to improve data analysis. This focus on the human analyst distinguishes the cognitive model of data analysis from other models of science, which it may appear similar to. A focus on the human analyst is necessary. When errors in analysis occur, they will do harm because they violate Aristotelian logic or Sir Karl Popper's principles of falsification. But the cause of these errors will be ingrained human tendencies. To prevent such errors, data analysts must understand and watch for these tendencies.

References

S. Attfield and A. Blandford. Improving the cost structure of sensemaking tasks: Analysing user concepts to inform information system design. *Human-Computer Interaction–INTERACT 2009*, pages 532–545, 2009.

F. C. Bartlett. Remembering: A study in experimental and social psychology. 1932.

G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356): 791–799, 1976.

J. D. Bransford, A. L. Brown, and R. R. Cocking. *How people learn: Brain, mind, experience, and school*. National Academies Press, Washington, DC, 2000.

L. Breiman. Nail finders, edifices, and oz. In Lucien M. Le Cam and Richard A. Olshen, editors, *Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 1, pages 201–214, Hayward, CA, 1985. Institute of Mathematical Sciences.

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

K. Carley and M. Palmquist. Extracting, representing, and analyzing mental models. *Social Forces*, 70(3):601–636, 1992.

C. Chatfield. *Problem solving: a statistician's guide*. Chapman & Hall/CRC, 1995.

G. W. Cobb. The introductory statistics course: A ptolemaic curriculum? *Technology Innovations in Statistics Education*, 1(1), 2007.

D. Cook and D. F. Swayne. *Interactive and dynamic graphics for data analysis with R and GGobi*. Springer Publishing Company, Incorporated, 2007.

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.

D. R. Cox. Comment on statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

D.R. Cox. Applied statistics: A review. *The Annals of Applied Statistics*, 1(1):1–16, 2007.

S. R. Dalal, E. B. Fowlkes, and B. Hoadley. Risk analysis of the space shuttle: Pre-challenger prediction of failure. *Journal of the American Statistical Association*, pages 945–957, 1989.

A. D. DeGroot. Thought and mind in chess. *The Hague: Mouton*, 1965.

B. Dervin. Sense-making theory and practice: an overview of user interests in knowledge seeking and use. *Journal of Knowledge Management*, 2(2):36–46, 1998.

M. Donald. *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Harvard Univ Pr, 1991.

J. C. Farman, B. G. Gardiner, and J. D. Shanklin. Large losses of total ozone in antarctica reveal clox/nox interaction. *Nature*, 315:207–201, 1985.

D. Freedman. *The development of theory could hasten the speed with which data fields adapt to emerging challenges. Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. Cambridge University Press, 2009.

A. Gelman and C. R. Shalizi. Philosophy and the practice of bayesian statistics. *Arxiv preprint arXiv:1006.3868*, 2010.

T. Gilovich, R. Vallone, and A. Tversky. The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17(3):295–314, 1985.

E. Goffman. *Frame analysis: An essay on the organization of experience.* Harvard University Press, 1974.

A. S. Helsdingen and K. Van den Bosch. Learning to make sense. In *Cognitive Systems with Interactive Sensors conference*, November 2009.

T. Hey and A. Trefethen. The data deluge: An e-science perspective. pages 809–824, 2003.

P. Huber. *Speculations on the Path of Statistics*. Princeton University Press, December 1997.

P. Huber. *Data Analysis What Can Be Learned from the Past 50 Years*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2011.

D. H. Jonassen and P. Henning. Mental models: Knowledge in the head and knowledge in the world. In *Proceedings of the 1996 international conference on Learning sciences*, pages 433–438. International Society of the Learning Sciences, 1996.

A. E. Jones. The antarctic ozone hole. *Physics Education*, 43:358, 2008.

G. Klein and B. W. Crandall. The role of mental simulation in problem solving and decision making. *Local applications of the ecological approach to human-machine systems*, 2:324–358, 1995.

G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso. A data/frame theory of sense making". In *Expertise out of context: proceedings of the sixth International Conference on Naturalistic Decision Making*, pages 113–155, 2003.

T. S. Kuhn. The structure of scientific revolutions. 1962.

G. Lakoff and R. Núñez. The metaphorical structure of mathematics: Sketching out cognitive foundations for a mind-based mathematics. *Mathematical reasoning: Analogies, metaphors, and images*, pages 21–89, 1997.

G. Lakoff and R. Núñez. *Where mathematics comes from: How the embodied mind brings mathematics into being*. Basic Books, 2000.

M. Lavine. Problems in extrapolation illustrated with space shuttle o-ring data. *Journal of the American Statistical Association*, 1991.

C. G. Lundberg. Made sense and remembered sense: Sensemaking through abduction. *Journal of Economic Psychology*, 21(6):691–709, 2000.

E. Q. B. Macabebe, I. B. Culaba, and J. T. Maquiling. Pre-conceptions of Newton's Laws of Motion of Students in Introductory Physics. In *AIP Conference Proceedings*, volume 1263, page 106, 2010.

R.J. MacKay and R.W. Oldford. Scientific method, statistical method and the speed of light. *Statistical Science*, 15(3):254–278, 2000. ISSN 0883-4237.

C. Mallows. The zeroth problem. *The American Statistician*, 52(1):1–9, 1998.

C. Mallows. Tukey's paper after 40 years (with discussion). *Technometrics*, 48(3):319–325, August 2006.

C. Mallows and P. Walley. A theory of data analysis? *Proceedings of the business and economics section, American Statistical Association*, 1980.

G. A. Miller. Some psychological studies of grammar. *American Psychologist*, 17(11):748, 1962.

M. Minsky. A framework for the representation of knowledge. *The psychology of computer vision*, pages 211–277, 1975.

M. J. Molina and F. S. Rowland. Stratospheric sink for chlorofluoromethanes. *Nature*, 249: 810–812, 1974.

D. S. Moore. Uncertainty. *On the shoulders of giants: New approaches to numeracy*, pages 95–137, 1990.

S. A. Mulaik. Exploratory statistics and empiricism. *Philosophy of Science*, 52(3):410–430, 1985.

U. Neisser. *Cognitive psychology*. Appleton-Century-Crofts New York, 1967.

U. Neisser. *Cognition and reality: Principles and implications of cognitive psychology.* WH Freeman/Times Books/Henry Holt & Co, 1976.

D.A. Norman. *Things that make us smart: Defending human attributes in the age of the machine*. Number 842. Basic Books, 1993.

R. D. Pea. Cognitive technologies for mathematics education. *Cognitive science and mathematics education*, pages 89–122, 1987.

C. S. Peirce. *Collected papers of Charles Sanders Peirce*, volume 1. Belknap Press, 1932.

J. Piaget and M. T. Cook. The origins of intelligence in children. 1952.

P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. May 2005.

Presidential Commission on the Space Shuttle *Challenger* Accident. *Report of the Presidential Commission on the Space Shuttle Challenger Accident (pbk).*, volume 1. Author, Washington, DC, 1986.

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

L. B. Resnick. Treating mathematics as an ill-structured discipline. 1988.

E. Rosch. Classification of real-world objects: Origins and representations in cognition. In P. N. Johnson-Laird and P. C. Watson, editors, *Thinking: Reading in cognitive science*, pages 212–222. Cambridge University Press, 1977.

E. Rosch and C.B. Mervis. Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, 7(4):573–605, 1975.

W. W. Rozeboom. Good science is abductive, not hypothetico-deductive. *What if there were no significance tests*, pages 335–392, 1997.

J. Rudolph. Into the big muddy and out again: Error persistence and crisis management in the operating room. *Unpublished doctoral dissertation, Boston College, Chestnut Hill, Mass., at http://escholarship. bc. edu/dissertations/AAI3103269*, 2003.

D. E. Rumelhart and A. Ortony. *The representation of knowledge in memory*. Center for Human Information Processing, Dept. of Psychology, University of California, San Diego, 1976.

D. M. Russell, M. J. Stefik, P. Pirolli, and S. K. Card. The cost structure of sensemaking. In *CHI '93: Proceedings of the INTERACT '93 and CHI '93 conference on Human factors in computing systems*, pages 269–276, New York, NY, USA, 1993. ACM.

R. C. Schank and R. P. Abelson. Scripts, plans, goals, and understanding:: An inquiry into human knowledge structures. 1977.

E.E. Smith. Theories of semantic memory. *Handbook of learning and cognitive processes*, 6:1–56, 1978.

P. J. Smith, W. C. Giffin, T. H. Rockwell, and M. Thomas. Modeling fault diagnosis as the activation and use of a frame system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 28(6):703–716, 1986.

J. Sweller, J. J. G. van Merrienboer, and F. G. W. C. Paas. Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3):251–296, 1998.

J. Tierney. Behind monty halls doors: Puzzle, debate and answer. *New York Times*, 140(1):20, 1991.

E. R. Tufte. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, first edition edition, February 1997. ISBN 0961392126.

J. W. Tukey. Conclusions vs decisions. *Technometrics*, 2(4):423–433, 1960.

J. W. Tukey. The future of data analysis. *The Annals of Mathematical Statistics*, 33(1):1–67, 1962.

J. W. Tukey and M. B. Wilk. Data analysis and statistics: an expository overview. In *Proceedings of the November 7-10, 1966, fall joint computer conference*, pages 695–709. ACM, 1966.

A. Tversky and D. Kahneman. Judgement under uncertainty. *Science*, 185:1124–1131, 1974.

A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, January 1981. ISSN 0036-8075.

A. Unwin. Patterns of data analysis? *Journal of the Korean Statistical Society*, 30(2):219–230, 2001.

H. Varian. Hal varian on how the web challenges managers. *McKinsey Quarterly*, 1, 2009.

P. F. Velleman. *The Philosophical Past and the Digital Future of Data Analysis*. Princeton University Press, December 1997.

R. Viertl. On the future of data analysis. *Austrian Journal of Statistics*, 31(2&3):241–244, 2002.

S. Vosniadou, W. F. Brewer, and University of Illinois at Urbana-Champaign. Center for the Study of Reading. *The concept of the earth's shape: A study of conceptual change in childhood*. University of Illinois at Urbana-Champaign, 1989.

K. E. Weick. *Sensemaking in Organizations (Foundations for Organizational Science)*. Sage Publications, Inc, May 1995.

K. E. Weick, K. M. Sutcliffe, and D. Obstfeld. Organizing and the process of sensemaking. *Organization Science*, 16(4):409–421, 2005.

M. Wertheimer. Laws of organization in perceptual forms. *A source book of Gestalt psychology*, pages 71–88, 1938.

C. J. Wild. Embracing the "wider view" of statistics. *The American Statistician*, 48(2):163–171, 1994.

C. J. Wild and M. Pfannkuch. Statistical thinking in empirical enquiry. *International Statistical Review/Revue Internationale de Statistique*, 67(3):223–248, 1999.

R. S. Woodworth. Experimental psychology. 1938.

A. Wu, X. L. Zhang, and G. Cai. An interactive sensemaking framework for mobile visual analytics. In *Proceedings of the 3rd International Symposium on Visual Information Communication*, page 22. ACM, 2010.

J. S. Yi, Y. A. Kang, J. T. Stasko, and J. A. Jacko. Understanding and characterizing insights: how do people gain insights using information visualization? In *BELIV '08: Proceedings of the 2008 conference on BEyond time and errors*, pages 1–6, New York, NY, USA, 2008. ACM.

J. Zhang. The nature of external representations in problem solving. *Cognitive science*, 21(2): 179–217, 1997.

J. Zhang. External representations in complex information processing tasks. *Encyclopedia of library and information science*, 68:164–180, 2000.

P. Zhang. Sensemaking: Conceptual changes, cognitive mechanisms, and structural representations. a qualitative user study. 2010.