

# Elements and Principles of Data Analysis

Stephanie C. Hicks<sup>1</sup> and Roger D. Peng<sup>1</sup>

<sup>1</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

March 20, 2019

## Abstract

The data revolution has led to an increased interest in the practice of data analysis. As a result, there has been a proliferation of “data science” training programs. Because data science has been previously defined as an intersection of already-established fields or union of emerging technologies, the following problems arise: (1) There is little agreement about what is data science; (2) Data science becomes secondary to established fields in a university setting; and (3) It is difficult to have discussions on what it means to learn about data science, to teach data science courses and to be a data scientist. To address these problems, we propose to define the field from first principles based on the activities of people who analyze data with a language and taxonomy for describing a data analysis in a manner spanning disciplines. Here, we describe the elements and principles of data analysis. This leads to two insights: it suggests a formal mechanism to evaluate data analyses based on objective characteristics, and it provides a framework to teach students how to build data analyses. We argue that the elements and principles of data analysis lay the foundational framework for a more general theory of data science.

## 1 Introduction

The data revolution has led to an increased interest in the practice of data analysis and increased demand for training and education in this area [Cleveland, 2001, Nolan and Lang, 2010, Workgroup, 2014, Baumer, 2015, PricewaterhouseCoopers, 2019, Hardin et al., 2015, Kaplan, 2018, Hicks and Irizarry, 2018]. This revolution has also led to the creation of the term *data science*, which is often defined only in relation to existing fields of study [Conway, 2010, Tierney, 2012, Matter, 2013, Harris, 2013], such as the intersection of computer science, statistics, and substantive expertise. For example, Drew Conway’s data science Venn diagram [Conway, 2010] defines data science as the intersection of hacking skills, substantive expertise, and statistics. An alternate approach is to define data science as the union of emerging methods, technologies or languages, including such tools as R, Python, SQL, Hadoop, or any of a long list of applications [Mason and Wiggins, 2010]. However, defining data science only in relation to other fields or technologies creates an existential problem for anything that we might consider as the *field* of data science. Two major problems with this approach are (1) there is no agreement on what are the foundations of data science and what are the core activities of data scientists, and (2) because data science is defined only in relation to existing fields of study or the tools that others use, it makes data science secondary in stature to those other fields in the university setting. The result is a lack of an identity for what data science

is and what data scientists should be doing, and an inability to communicate and teach consistent topics about data science.

We propose to use an alternative definition of *data science* based on what a data scientist does, as others have taken a similar approach [Mason and Wiggins, 2010, Hochster, 2014], but here we formalize these ideas:

*Data science is the science and design of (1) actively creating a question to investigate a hypothesis with data, (2) connecting that question with the collection of appropriate data and the application of appropriate methods, algorithms, computational tools or languages in a data analysis, and (3) communicating and making decisions based on new or already established knowledge derived from the data and data analysis.*

These three states do not necessarily occur in a linear order, and data scientists constantly move or iterate between states.

Within the core definition of data science, a data scientist builds a *data analysis* [Tukey, 1962, Tukey and Wilk, 1966, Box, 1976, Wild, 1994, Chatfield, 1995, Wild and Pfannkuch, 1999, Cook and Swayne, 2007], which has been defined as “the investigative process used to extract knowledge, information, and insights about reality by examining data” [Grolemund and Wickham, 2014]. In a typical data analysis, the data scientist starts with a question in mind and a set of collected data to investigate a hypothesis, all of which may evolve over time. However, there is a great amount of variation in defining what is a data analysis and how a data scientist can construct or create a data analysis. Furthermore, there is little in the way of language available that data scientists can use to describe what makes analyses different from each other.

Other fields, such as art or music, have overcome similar challenges by defining a language that can be used to construct or create works of art and to characterize the variation between different pieces of art. More formally, an artist can create a piece of art using the *elements* and *principles* specific to that area. The elements of art are color, line, shape, form, and texture [of Art, 2019]; and the principles of art are the means by which the artist uses to compose or to organize the elements within a work of art [Marder, 2018]. For example, an artist can use the principle of contrast (or emphasis) to combine elements in a way that stresses the differences between those elements, such as combining two contrasting colors, black and white. The principles of art, by themselves, are not used to evaluate a piece of art, but they are meant to be objective characteristics that can describe the variation between pieces of art.

Here, we define a language for describing data analyses that can be used to construct and to create a data analysis and to characterize the variation between data analyses. We denote this language as the *elements and principles of data analysis* and we use them to describe the fundamental concepts for the practice and teaching of creating a data analysis. Furthermore, this

language provides the vocabulary and framework to have an informed debate and discussion on what a data scientist does in the larger context of defining data science. We argue that this is a more productive way to define data science, where different individuals or disciplines can emphasize different elements and principles of a data analysis. In addition, having a descriptive language for data analyses gives us an efficient way to communicate about data analyses and to describe lessons learned from members of this emerging field.

## 1.1 Defining a data analysis based on elements and principles

In scientific fields, a data analysis serves to quantify and characterize evidence in data. Scientists develop hypotheses about the world and collect data in a manner guided by those hypotheses. A data analysis is then conducted by a data analyst to formally determine the strength of evidence in favor of one hypothesis vis-a-vis an alternative hypothesis. There are a variety of methods and tools available to the data analyst for accomplishing this goal, some of which may provide different summaries of the data and the evidence. Part of the data analyst's job is to choose the appropriate set of methods, algorithms, computational tools or languages to assess the strength of evidence for a particular hypothesis.

Decisions about how the analysis is conducted are made while simultaneously considering three aspects, which are all fundamental to the larger field of data science, but are treated as key inputs to the data analysis here: (i) the hypotheses, or more broadly the scientific questions being addressed, (ii) the data, and (iii) the audience that will ultimately view or digest the data analysis. It is important to note that by the word *hypothesis*, we are not referring to statistical hypothesis testing, but instead we refer to it as *a proposed explanation for a phenomenon*, which can be broadly investigated by descriptive, exploratory, inferential, predictive, casual, mechanistic analyses [Leek and Peng, 2015].

These decisions by the analyst are critical because, while analysts will choose to analyze the data in a manner that they believe is valid for the hypothesis, data, and audience, analyses of the same data set can vary widely depending on the specific choices made by the analyst [Silberzahn et al., 2018], including variation in the methods, tooling, and workflow. However, we currently lack a language to describe how to construct and to create a data analysis and to characterize the variation between data analyses. To address this, we can leverage the ideas from other fields, such as art and music, to define a language for describing data analyses based on the elements and principles chosen by the data analyst. Briefly, the elements of an analysis are the individual basic components of the analysis that, when assembled together by the analyst, make up the entire analysis. The principles of the analysis are prioritized qualities or characteristics that are relevant to the analysis, as a whole or individual components, and that can be objectively observed or

measured. In the sections below, we go into further detail on these two important topics, namely the *elements* (Section 2) and *principles* (Section 3) of data analysis.

Finally, an analysis will usually result in potentially three basic outputs. The first is the analysis itself, which we imagine as living in an *analytic container* which might be a set of files including a Jupyter notebook or R Markdown document, a dataset, and a set of ancillary code files. The analytic container is essentially the “source code” of the analysis and it is the basis for making modifications to the analysis and for reproducing its findings. In addition to the container, there will usually be an *analytic product*, which is the executed version of the analysis in the analytic container, containing the executed code producing the results and output that the analyst chooses to include, which might be a PDF document or HTML file. Finally, the analyst will often produce an *analytic presentation*, which might be a slide deck, PDF document, or other presentation format, which is the primary means by which the data analysis is communicated to the audience. Elements included in the analytic presentation may be derived from the analytic container, analytic product, or elsewhere.

## 1.2 A sample data analysis

To provide a concrete example, we present here a brief vignette that describes a hypothetical data analysis scenario and presents examples of the ideas we develop below.

Roger has a long-term collaboration with a pediatrician who has just finished executing a clinical trial to see if a new asthma treatment can decrease exhaled nitric oxide (eNO, a measure of pulmonary inflammation) in children with asthma, as compared to the standard of care. She sends Roger the data upon completion of the data collection and asks for a summary of the findings as they relate to the primary outcome.

Roger receives the data and immediately notices that the "treatment" variable is coded as "0" and "1". He calls his collaborator to inquire about the coding scheme and she clarifies that "0" indicates standard of care and "1" indicates the new drug. Upon hearing that, Roger re-codes the data to have more informative labels. Roger then makes a histogram of the outcome (which is continuous) to see if there is anything unusual. He then makes side-by-side boxplots of eNO by treatment group to see if there is any difference between the groups. As a final step he runs a two-sample *t*-test to test for a difference in eNO between groups. Upon seeing those results he puts together a brief slide deck presentation of the results for his collaborator and schedules a call to discuss next steps. A representation of the code and output Roger used for this initial analysis is presented in Figure 1.

The unexecuted code presented above is the analytic container which describes what was done and provides a mechanism for reproducing or modifying the analysis. The executed code is the

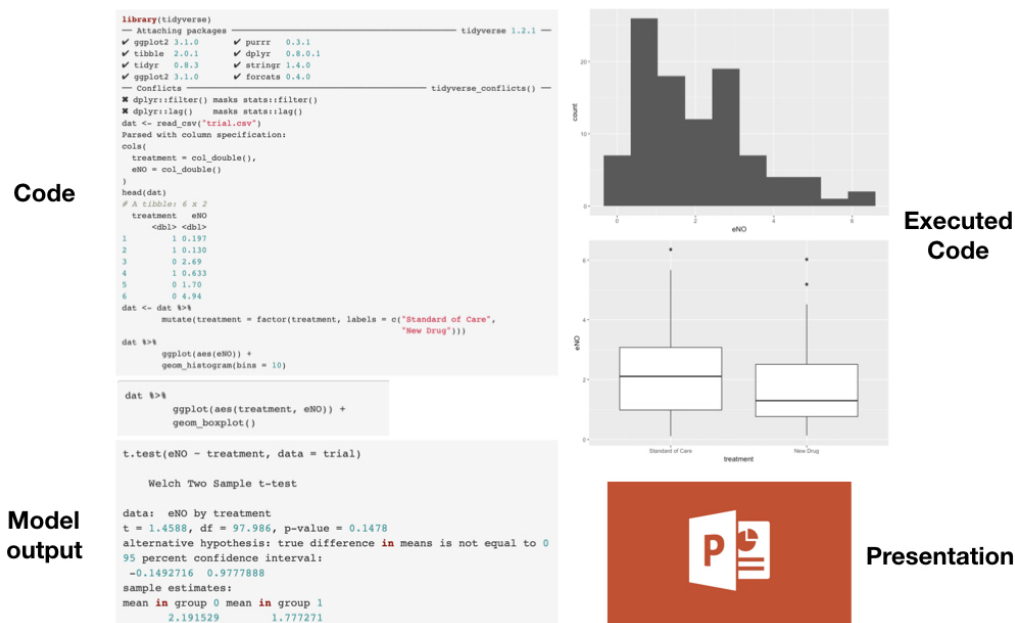


Figure 1: Sample analytic container, analytic product, and analytic presentation.

analytic product and includes model output (the  $t$ -test) and two plots. Finally, the slide deck is the analytic presentation and is a key vehicle by which the collaborator will interact with the analysis. All of the items presented above are elements of a data analysis and will be described in greater detail in the next section.

## 2 Elements of data analysis

The *elements* of a data analysis are the fundamental components of a data analysis used by the data analyst: code, code comments, data visualization, non-data visualization, narrative text, summary statistics, tables, and statistical models or computational algorithms [Breiman, 2001] (Table 1).

Code and code comments are two of the most commonly used elements by the data analyst to describe the executable programmatic instructions to execute a set of operations or computations and the non-executable instructions that describe the action or result of the surrounding code. These can be an entire line, multiple lines or a short snippet. Examples of code include defining variables or writing functions. Code comments and narrative text are related because they both can include expository phrases or sentences that describe what is happening in the data analysis in a human readable format. However, the difference between the two is the code comment has a symbol in front of the narrative text, which instructs the document container to not execute this element.

There are two types of visualizations elements used in data analysis, data visualization and non-data visualization, where the former can be a plot, figure or graph illustrating a visual repre-

sensation of the data and the latter is figure relevant to the data analysis but does not necessarily contain data, such as a diagram or flowchart.

There are two main types of summary elements of a data analysis: summary statistics and tables, where the former are one (or more than one) dimensional numerical quantities derived from the data, such as mean or standard deviation, while the latter is an ordered arrangement of either data or summaries of the data into a row and column format. The last element of a data analysis is the statistical model or computational algorithm, which an analyst uses to investigate the data-generation process or predictive ability of different mathematical models or computational algorithms.

Element	Description
Narrative text	Expository phrases or sentences that describe what is happening in the data analysis in a human readable format
Code	A series of programmatic instructions to execute a particular programming or scripting language
Code comment	Non-executable code or text near or inline with code that describes the expected action/result of the surrounding code or provides context
Data visualization	A plot, figure or graph illustrating a visualize representation of the data.
Narrative diagram	A diagram or flowchart without data
Summary statistics	Numerical quantities derived from the data, such as the mean, standard deviation, etc.
Table	An ordered arrangement of data or summaries of data in rows and columns
Statistical model or computational algorithm	Mathematical model or algorithm concerning the underlying data phenomena or data-generation process, predictive ability, or computational algorithm

Table 1: **Elements of a data analysis.** This table describes eight elements that are used by the data analyst to build the data analysis. These elements include code, code comments, data visualization, non-data visualization, narrative text, summary statistics, tables, and statistical models or computational algorithms.

In addition to these elements, there are also *contextual inputs* to the data analysis, such as the main scientific question or hypothesis, the data, the choice of programming language to use, the audience, and the document or container for the analysis, such as a Jupyter or R Notebooks. We do not include these as elements of data analysis, because these inputs are not necessarily decided or fundamentally modified by the analyst. Often an upstream entity such as a manager at a company, a collaborator at a university or scientific institute, an educator in the classroom provides the framework for these contextual inputs. However, we note that often the data analyst will be expected to decide or contribute to these contextual inputs. In addition, it may be the analyst’s job to provide feedback on some of these inputs in order to further refine or modify them. For example, an analyst may be aware that a specific programming language is more appropriate for a planned analysis than the currently selected one.

### 3 Principles of data analysis

The *principles* illustrated by a data analysis are prioritized qualities or characteristics that are relevant to the analysis, as a whole or individual components, and that can be objectively observed or measured. Their presence (or absence) in the analysis is not dependent on the characteristics of the audience viewing the analysis, but rather the relative weight assigned to each principle by the analyst can be highly dependent on the audience’s needs. In addition, the weighting of the principles by the analyst can be influenced outside constraints or resources, such as time, budget, or access to individuals to ask context-specific questions, that can impose restrictions on the analysis. The weighting of the principles *per se* is not meant to convey a value judgment with respect to the overall quality of the data analysis. Rather, the requirement is that multiple people viewing an analysis could reasonably agree on the fact that an analysis gives high or low weight to certain principles. In Section 5 we describe some hypothetical data analyses that demonstrate how the various principles can be weighted differently. Next, we describe six principles that we believe are informative for creating and characterizing data analyses.

**Data Matching.** Data analyses with high *data matching* have data readily measured or available to the analyst that directly matches the data needed to investigate a hypothesis or problem with data analytic elements (Figure 2). In contrast, a scientific question may concern quantities that cannot be directly measured or are not available to the analyst. In this case, data matched to the hypothesis may be surrogates or covariates to the underlying data phenomena. While we consider the main scientific question or hypothesis and the data to be contextual inputs to the data analysis, we consider this a principle of data analysis because the analyst selects data analytic elements that are used to investigate the hypothesis, which depends on how well the data are matched. If the data are poorly matched, the analyst will not only need to investigate the main hypothesis with one set of data analytic elements, but also will need to use additional elements that describe how well the surrogate data is related to the underlying data phenomena to investigate the main hypothesis.

It is important to note, problems and hypotheses can be more or less specific, which will impose strong or weak constraints on the range of data matching to the problem. Highly specific hypotheses or problems tend to induce strong constraints to investigate with data analytic elements. Less specific problems emit a large range of potential data to investigate the hypothesis. Data that can be readily measured or are available to the analyst to directly address a specific hypothesis results in high data matching, but depending on the problem specificity, can result in a narrow or broad set of data to consider.

**Exhaustive.** An analysis is *exhaustive* if specific questions or hypotheses are addressed using multiple, complementary elements (Figure 3). For example, using a  $2 \times 2$  table, a scatter plot,

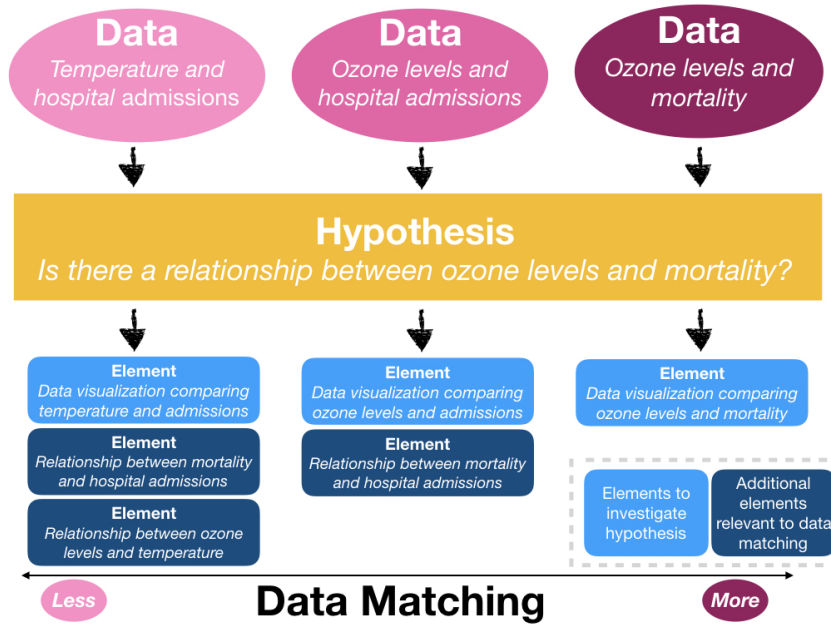


Figure 2: **The data matching principle of data analysis.** Data analyses with high data matching have data readily measured or available to the analyst that directly matches the data needed to investigate a hypothesis or problem with data analytic elements. In contrast, a hypothesis may concern quantities that cannot be directly measured or are not available to the analyst. In this case, data matched to the hypothesis may be surrogates or covariates to the underlying data phenomena that may need additional elements to describe how well the surrogate data is related to the underlying data phenomena to investigate the main hypothesis.

and a correlation coefficient are three different elements that could be employed to address the hypothesis that two predictors are correlated. Analysts that are exhaustive in their approach use complementary tools or methods to address the same hypothesis, knowing that each given tool reveals some aspects of the data but obscures other aspects. As a result, the combination of elements used may provide a more complete picture of the evidence in the data than any single element.

**Skeptical.** An analysis is *skeptical* if multiple, related hypotheses are considered using the same data (Figure 4). Analyses, to varying extents, consider alternative explanations of observed phenomena and evaluate the consistency of the data with these alternative explanations. Analyses that do not consider alternate explanations have no skepticism. For example, to examine the relationship between a predictor X and an outcome Y, an analyst may choose to use different models containing different sets of predictors that may potentially confound that relationship. Each of these different models represents a different but related hypothesis about the X-Y relationship. A separate question that arises is whether the configuration of hypotheses or alternative explanations are relevant to the problem at hand. However, often that question can only be resolved using contextual information that is outside the data.

**Second-Order.** An analysis is *second-order* if it includes elements that do not directly address



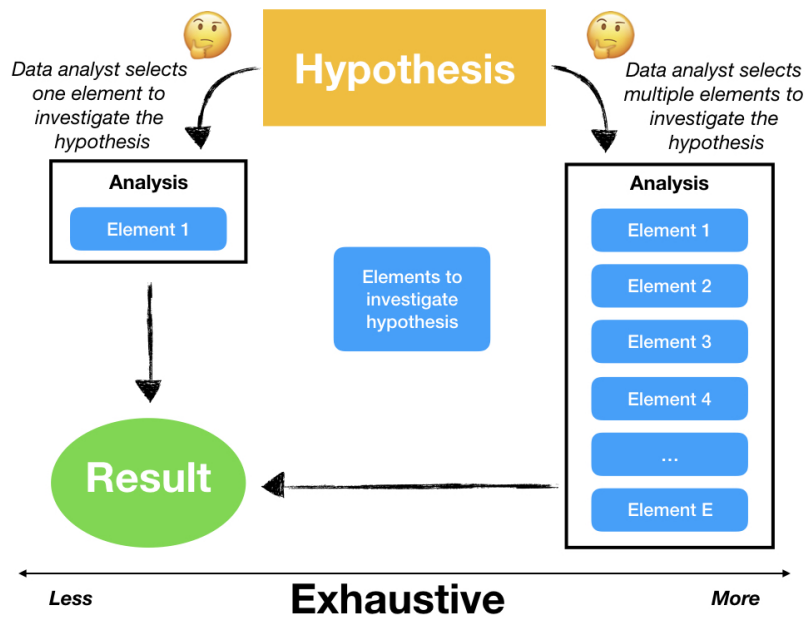


Figure 3: **The exhaustive principle of data analysis.** An analysis is exhaustive if specific questions or hypotheses are addressed using multiple, complementary elements. Given a hypothesis or scientific question, the analyst can select an element or set of complementary elements to investigate the hypothesis. The more complementary elements that are used to investigate the hypothesis, the more exhaustive the analysis is, which provides a more complete picture of the evidence in the data than any single element.

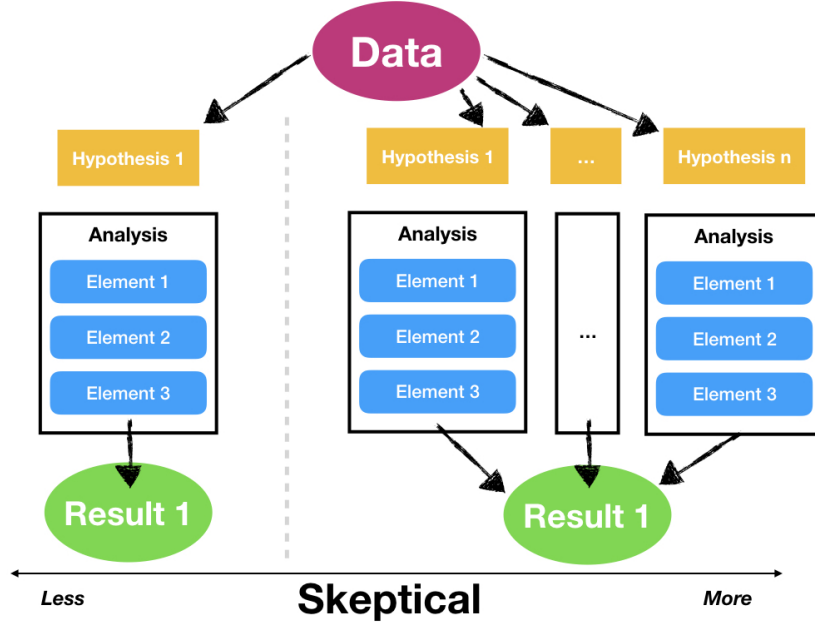


Figure 4: **The skeptical principle of data analysis.** An analysis is skeptical if multiple, related hypotheses or alternative explanations of observed phenomena are considered using the same data and offer consistency of the data with these alternative explanations. In contrast, analyses that do not consider alternate explanations have no skepticism.

the primary hypothesis or question, but give important context or supporting information to the analysis (Figure 5). Any given analysis will contain elements that directly contribute to the results

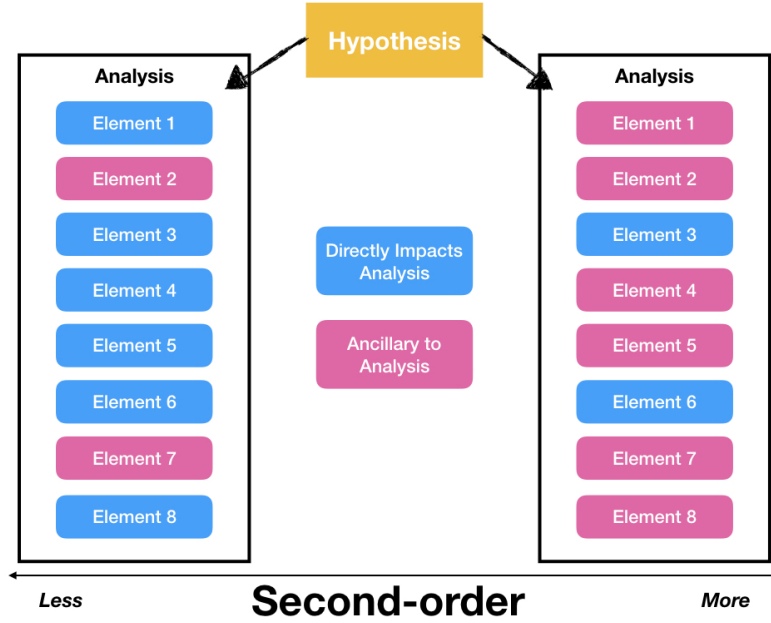


Figure 5: **The second-order principle of data analysis.** An analysis is second-order if it includes ancillary elements that do not directly address the primary hypothesis/question but give important context to the analysis. Examples of ancillary elements could be background information of how the data were collected, and expository explanations or analyses comparing different statistical methods or software packages. While these details may be of interest and provide useful background, they likely do not directly influence the analysis itself.

or conclusions, as well as some elements that provide background or context or are needed for other reasons, such as if the data are less well matched to the investigating the hypothesis (Figure 2). Second-order analyses contain more of these background/contextual elements in the analysis, for better or for worse. For example, in presenting an analysis of data collected from a new type of machine, one may include details of who manufactured the machine, or why it was built, or how it operates. Often, in studies where data are collected in the field, such as in people’s homes, field workers can relay important about the circumstances under which the data were collected. In both examples, these details may be of interest and provide useful background, but they may not directly influence the analysis itself. Rather, they may play a role in interpreting the results and evaluating the strength of the evidence.

**Transparent.** *Transparent* analyses present an element or subset of elements summarizing or visualizing data that are influential in explaining how the underlying data phenomena or data-generation process connects to any key output, results, or conclusions (Figure 6). While the totality of an analysis may be complex and involve a long sequence of steps, transparent analyses extract one or a few elements from the analysis that summarize or visualize key pieces of evidence in the data that explain the most “variation” or are most influential to understanding the key results or conclusion. One aspect of being transparent is showing the approximate mechanism by which the data inform the results or conclusion.

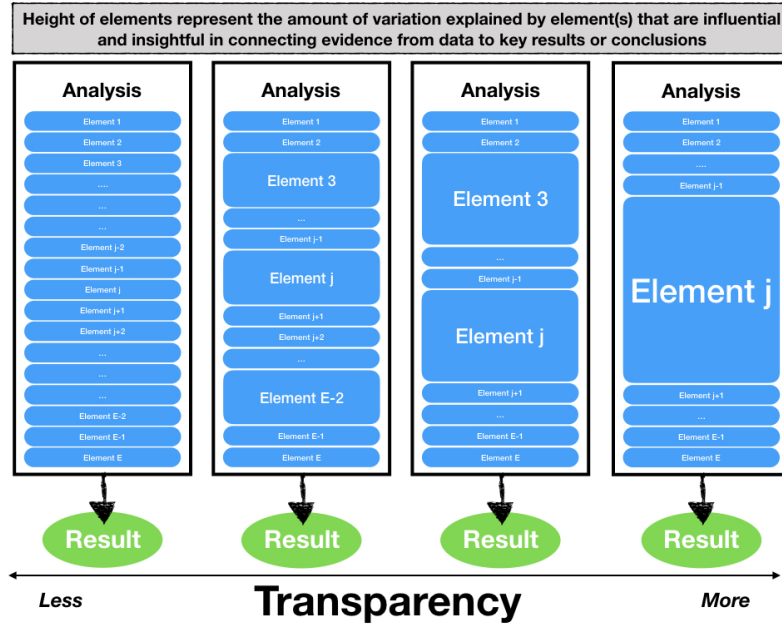


Figure 6: **The transparency principle of data analysis.** Transparent analyses present an element or set of elements summarizing or visualizing data that are influential in explaining how the underlying data phenomena or data-generation process connects to any key output, results, or conclusions. While the totality of an analysis may be complex and involve a long sequence of steps, transparent analyses extract one or a few elements from the analysis that summarize or visualize key pieces of evidence in the data that explain the most “variation” or are most influential to understanding the key results or conclusion.

**Reproducible.** An analysis is *reproducible* if someone who is not the original analyst can take the published code and data and compute the same results as the original analyst (Figure 7). In the terminology of our framework, given the elements of the data analysis, we can produce the exact same results of the analysis. Critical to reproducibility is the availability of the analytic container to others who may wish to re-examine the results. Much has been written about reproducibility and its inherent importance in science, so we do not repeat that here [Peng, 2011]. We simply add that reproducibility (or lack thereof) is usually easily verified and is not dependent on the characteristics of the audience viewing the analysis. Reproducibility also speaks to the coherence of the workflow in the analysis in that the workflow should show how the data are transformed to eventually become results.

## 4 The atomic unit of data analysis

The framework of elements and principles described thus far allows us to define a data analysis as a collection of elements whose choice is guided by the relative weight given to a set of principles. Given the contextual inputs, the data analysis is built upon *atomic units* which occur both on the computer and in the analyst’s head (Figure 8). In each atomic unit, the data analyst first chooses a principle to investigate a hypothesis or scientific question. Then, the analyst alternates between

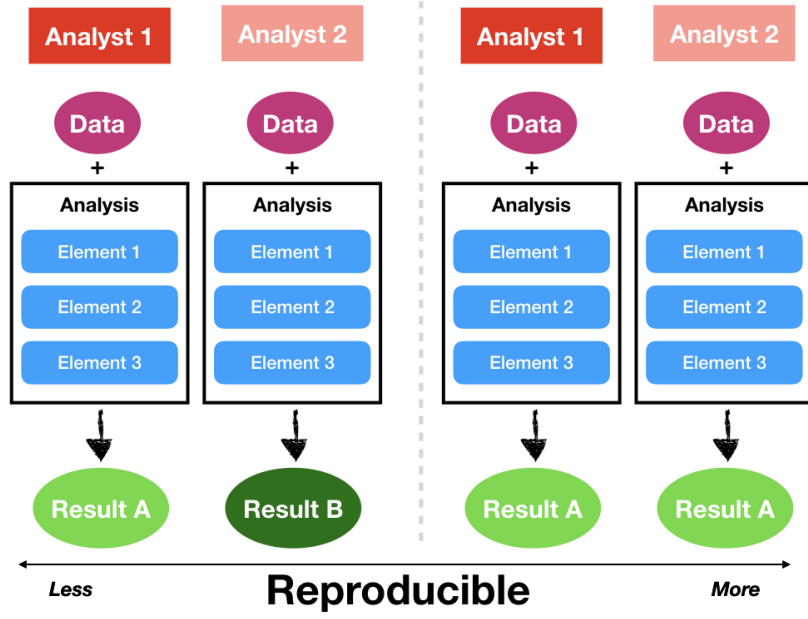


Figure 7: **The reproducible principle of data analysis.** An analysis is reproducible if someone who is not the original analyst (Analyst 2) can take the same data and the same elements of the data analysis and produce the exact same results as the original analyst (Analyst 1). In contrast, analyses that conclude in different results are less reproducible.

two stages until the analyst exits the atomic unit. The actions in the atomic unit are denoted in Figure 8 by (1) the analyst selects an element to investigate a hypothesis or scientific question, (2) the analyst interprets the output, (3) the analyst decides whether or not to place the element from step (1) into the analytic container or analytic product or not, and (4) the analyst decides whether to continue in the atomic unit by selecting another element to investigate the hypothesis or question or to exit the atomic unit entirely and end this line of investigation.

While action (2) happens on the computer or console, action (3) happens in the analyst’s head. In action (3), if an element is not recorded into the analytic container or analytic product, then the audience does not see it as part of the analysis. At the conclusion of the atomic unit, the analyst can use the information gained or not gained to guide the choice of what will be the next atomic unit.

As an example, let us assume our hypothesis is that two features X and Y are associated. The analyst begins the atomic unit by choosing a principle to investigate this hypothesis (1), for example the *principle of exhaustiveness*, where the analyst begins by picking an element (2), such as a 2-dimensional scatter plot, to investigate the hypothesis. The element is executed in the console, which might reveal evidence for or against an association. The analyst then interprets the result in his or her head. At this point, the analyst must make two decisions: if he or she wants to place the data visualization element in the analytic container (or analytic product) or not (3) and if he or she would like to continue the investigation with another element or exit the atomic

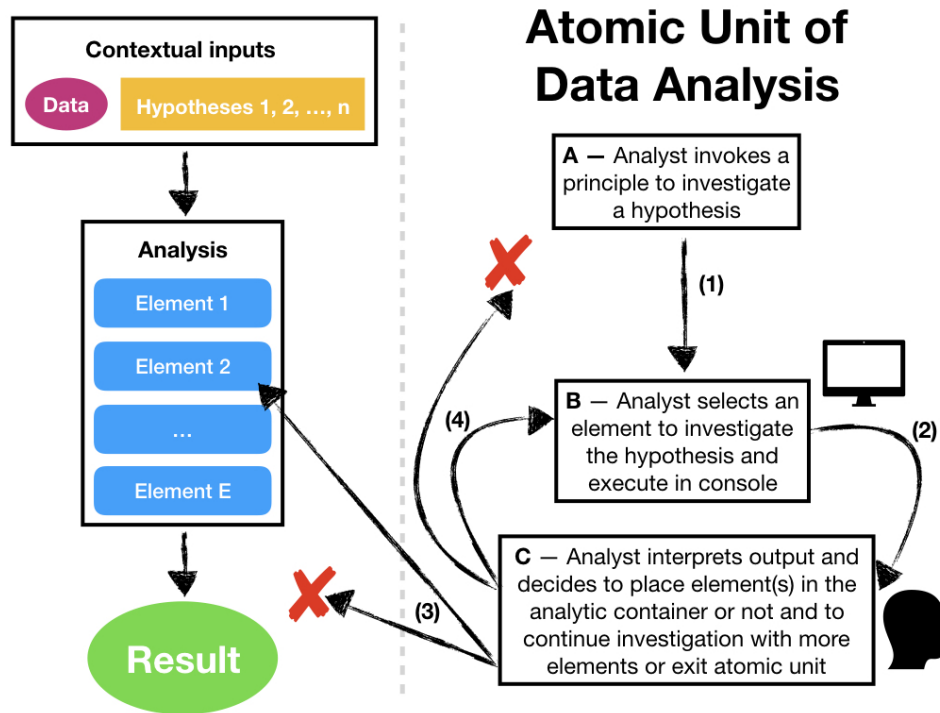


Figure 8: **The atomic unit of data analysis.** Given the contextual inputs, the data analysis is built upon atomic units of data analyses, which occurs both on the computer and in the analyst's head. In each atomic unit, the data analyst first chooses a principle to investigate a hypothesis or scientific question (Stage A). Then, the analyst alternates between Stages B and C until the analyst exits the atomic unit, either by choosing to end the line of investigation or choosing to invoke a new principle. The actions in the atomic unit are denoted in by (1) the analyst selects an element to investigate a hypothesis or scientific question, which is executed in the computer or console, (2) the analyst interprets the output in his or her head, (3) the analyst decides whether or not to place the element from (1) in the data container or data product or not, which means the element is never recorded in the data container or product and the audience does not see it as part of the analysis, and (4) the analyst decides to whether to continue in the atomic unit by selecting another element to investigate the hypothesis or question or to exit the atomic unit entirely and end this line of investigation.

unit (4). Let's assume the analyst decides to incorporate the element into the data container and also decides to continue the investigation of the hypothesis because the principle of exhaustiveness has been invoked by the analyst. Then, let's assume the analyst decides to quantitatively assess the strength of the association between X and Y using another element. If the analyst selects an element (4), such as the mean of X, to quantify the strength of association between X and Y, then the analyst would not learn anything (2) because a univariate summary statistic (or element) will not provide information on a 2-dimensional relationship between X and Y. Next, the analyst chooses not to record this element into the data container or data product (3) and chooses to select a different element (4), such as Pearson correlation summary statistic, which would be a more informative element to use when assessing the strength of the association between X and Y. The analyst interprets the results (2) and then makes the decision to place the element into the data container or product (3) and to end this investigation of the original hypothesis (4). In this

way, the analyst has completed one atomic unit of data analysis to form a portion of the final data analysis by invoking the principle of exhaustiveness by incorporating multiple, complementary elements to investigate if two features X and Y are associated.

## 5 Vignettes

To make these ideas more concrete, we provide four vignettes in this section where we describe how a data analysis could invoke or not invoke certain principles of data analysis.

### 5.1 Vignette 1

*Background.* Roger is interested in understanding the relationship between outdoor air pollution concentrations and variables predictive of pollution levels (e.g. temperature, wind speed, distance to road, traffic counts, etc.). However, monitoring of air pollution is expensive and time-consuming, and so he is interested in developing a prediction model for predicting air pollution levels where there is no monitoring data.

*Analysis.* Using available data on air pollution concentrations as well as 20 other variables that he thinks would be predictive of pollution levels, he fits a linear model using measured monitor-level pollution concentrations as the outcome. Once the analysis is complete, he includes all of the code and the writeup in a Jupyter Notebook. The document and all the corresponding data are uploaded to GitHub and are made publicly available.

*Results.* In his analysis report, he indicates that temperature had a large coefficient in the model and was statistically significant. He further reports that a 1 degree increase in temperature was associated with a 2.5 unit increase in pollution. Other coefficients that were statistically significant were the coefficients for distance to road and wind speed.

#### 5.1.1 Mapping to Principles

The stated goal of this analysis is to build a prediction model for predicting unobserved levels of air pollution. A linear model is fit and then the coefficients of the model are interpreted<sup>1</sup>.

- **Matching to the Data:** The data appear highly appropriate for addressing the problem of building a prediction model for pollution. Observed monitoring data are available for the outcome and 20 covariates that potentially related to pollution are used as predictors.

---

<sup>1</sup>While the goal was to build a predictive analysis, the analyst ultimately built an inferential analysis. The principles of data analysis do not characterize the validity or success of the analysis, nor the strength or quality of evidence for the hypothesis of interest. However, we propose a framework for how the elements and principles of data analysis might be used for these ideas in Section 6.3.

- **Exhaustive:** There is little evidence of exhaustiveness in this report. There is no attempt to try alternative approaches to see if improvements can be made. Essentially, one model was fit and the results reported.
- **Skeptical:** There is little skepticism in the analysis as only one approach was taken and only one question/hypothesis was explored. There is also no checking of assumptions or the linear modeling choices made.
- **Second-order:** No second-order details are provided in the summary.
- **Transparent:** The level of transparency is low as there is no data visualization or data summary included in the analysis that highlights the evidence in the data or data-generation process that reveal how the reported results are influenced by features in the data.
- **Reproducible:** The code and data are made available on GitHub and the code for implementing the model is organized in an R Markdown document. The analysis therefore appears reproducible.

## 5.2 Vignette 2

*Background.* Stephanie works as a data scientist at a small startup company that sells widgets over the Internet through an online store on the company’s web site. One day, the CEO comes by Stephanie’s desk and asks her how many customers have typically shown up at the store’s website each day for the past month. The CEO waits by Stephanie’s desk for the answer.

*Analysis.* Stephanie launches her statistical analysis software and, typing directly into the software’s console, immediately pulls records from the company’s database for the past month. She then groups the records by day and tabulates the number of customers. From this daily tabulation she then calculates the mean and the median count.

*Results.* Stephanie verbally reports the daily mean and median count to the CEO standing over her shoulder. She also notes that in the past month the web site experienced some unexpected down time when it was inaccessible to the world for a few hours.

### 5.2.1 Mapping to Principles

This scenario is a typical “quick analysis” that is often done under severe time constraints and where perhaps only an approximate answer is required. In such circumstances, there is often a limited ability to weight certain principles very heavily.

- **Matching to Data:** The data are essentially perfectly matched to the problem. The database tracks all visitors to the web site and the analysis used data directly from the database.

- **Exhaustive:** There is some evidence of exhaustiveness here as the analysis presented both the mean and the median as a summary of the typical number of customers per day.
- **Skeptical:** The analysis did not address any other hypotheses or questions.
- **Reproducible:** Given that the results were verbally relayed and that the analysis was conducted on the fly in the statistical software’s console, the analysis is not reproducible.
- **Second-order:** Noting that the web site experienced some downtime is a second order detail that suggests that the result presented may not be typical for a monthly mean. However, the information does not imply that the summary statistic is incorrect and therefore does not necessarily directly impact the analysis.
- **Transparent:** The analysis lacks transparency, but is nevertheless fairly simple. One way to make this analysis more transparent would be to include a histogram or a time series plot of daily values.

### 5.3 Vignette 3

*Background.* Stephanie is conducting a data analysis to see if unconventional oil and gas extraction (e.g. hydraulic “fracking”) methods are affecting the health of populations living near oil and gas deposits in Colorado. Because precise data on the operation of oil and gas wells is not publicly available, she can only obtain the location of each well and a five-year window in which the well was operating. This isn’t ideal, but she considers doing the analysis with this coarsened data more valuable than not doing it. She is able to obtain insurance records from Medicare from everyone in the state aged 65 years and older and so can examine health claims within this population and connect them to oil and gas activity.

*Analysis.* For her primary analysis, Stephanie decides to compare hospitalizations for respiratory disease during times when a well was in operation to otherwise similar times before the well was active. She first makes a time series plot of hospitalizations and annotates the plot with the windows of time when oil and gas wells were in operation. As part of her report, she provides an extensive description of how the oil wells operate and the schedule of activity by which they tend to adhere. She then runs a generalized linear regression model (GLM) with numbers of hospitalizations for respiratory diseases as the outcome and an indicator of well activity as the key predictor. Potential confounders are adjusted for directly in the regression model. To test the sensitivity of her findings to her chosen model, she fits a series of additional models using different functional forms on the various confounders. Finally, as an alternative modeling approach, Stephanie implements a propensity score model for the indicator of well operation and then uses a doubly robust (DR) estimator to estimate the effect of well operation on hospitalizations.



*Results.* The time series plot, generalized linear regression model, and doubly robust estimator all appear to provide similar and consistent results. Adjusting for different sets of potential confounders in the GLM does not appear to modify the primary results much. Stephanie decides that the time series plot effectively tells the story of the relationship between well activity and hospitalizations and decides to include that in her final analysis. The results from the GLM, doubly robust estimator, and various sensitivity analyses are all presented in the analysis. Because the topic is of great interest to many stakeholders, she writes the entire analysis in an R Markdown document where the code for all of the analyses and models can be easily accessed. The R Markdown document and well activity data are placed on a GitHub repository. Because the hospitalization data are considered protected health information, she cannot post that dataset publicly.

### 5.3.1 Mapping to Principles

- **Matching to Data:** The data are not particularly well-matched to the problem because coarseness of the data on well activity is not ideal.
- **Exhaustive:** The analysis demonstrates some exhaustiveness as multiple modeling approaches (GLM and DR estimator) were taken to address the primary question and a time series plot was made.
- **Skeptical:** Multiple hypotheses were considered in the GLM in the consideration of different sets of confounders, demonstrating some amount of skepticism in the analysis. There did not appear to be strong evidence that any of the potential confounders played a key role in explaining the relationship between well activity and hospitalization.
- **Second-order:** The information about how wells operate is second-order and does not pertain directly to the analysis, but is relevant background.
- **Transparent:** The presentation of a single element, the time series plot, demonstrates how the data are connected to the results/conclusions.
- **Reproducible:** The analysis is partially reproducible because the modeling code and well data are available to others. However, because the hospitalization data are not available, it is not possible for others to recreate the analysis results in their entirety.

## 5.4 Vignette 4

*Background.* Roger is considering is trying to determine whether ambient nickel being generated by an oil-burning power plant is correlated with respiratory illness in a nearby community. He is given data on both variables and must decide what methods to apply to address this question

*Analysis.* Roger initially decides to calculate the Pearson correlation coefficient between the two variables. Upon doing so he sees that the coefficient is positive and concludes that the two variables are positively correlated. He then decides to consider one more element, a scatter plot of the two variables, to further investigate this question of correlation. Upon seeing the scatter plot it is clear that there is an outlier in the nickel variable that is driving the positive correlation. If the outlier were removed, the correlation would likely be zero (although he does not actually compute this value). Seeing the scatter plot, Roger decides not to include it in the final analysis.

*Results.* The final analysis includes the Pearson correlation coefficient along with a conclusion that ambient nickel and respiratory illness are “likely positively correlated”.

#### 5.4.1 Mapping to Principles

At this point, the analysis consists of the correlation coefficient and its interpretation.

- **Matching to Data:** The data appear reasonably matched to the problem.
- **Exhaustive:** Although the analyst used two elements (the correlation coefficient and the scatter plot) to explore the problem, only one is presented in the final analysis. As a result, the analysis is not exhaustive.
- **Skeptical:** There is no skepticism in the final analysis. One way to the skepticism could have been increased is if the correlation coefficient had been calculated with the outlier removed.
- **Second-order:** There is no second order information in the analysis.
- **Transparent:** This analysis is simple, yet lacks transparency. Including the scatter plot would have increased transparency.
- **Reproducible:** There is no evidence that the analysis is reproducible.

## 6 Discussion

In developing the *elements* and *principles* of data analysis, our goal is to define a language for describing *data analyses* that can be used to characterize the variation between analyses. Ultimately, we hope that this language can serve as the foundational framework for developing a *theory* of data analysis or data science. While the elements are the building blocks for a data analysis, the principles can be wielded by the analyst to create data analyses with diversity in content and design. It is important to reiterate that the inclusion or exclusion of certain elements or the weighting of different principles in a data analysis do not determine the overall quality or success of a data analysis.

The framework we have developed here provides a mechanism for describing differences between analyses in a manner that is not specifically tied to the science or the application underlying the analysis. Being able to describe differences in this manner allows data analysts, who may be working in different fields or on disparate applications, to have set of concepts that they can use to have meaningful discussions. The elements and principles therefore broaden the landscape of data analysts and allows people from different areas to converse about their work and have a distinct shared identity.

When considering the foundations of data analysis, one cannot ignore the work of Tukey [Tukey, 1962], where he suggested that data analysis be thought of as a scientific field, as opposed to more like mathematics. In thinking of data analysis that way, the field’s identity would be centered around a set of problems rather than a set of tools. Our work builds on this work and that of many others in that we similarly believe that data analysis should be thought of in a different way. However, perhaps different from Tukey’s time, there is a substantial heterogeneity in the population of people doing data analysis today. As a result, there is a need to develop a language for discussing data analysis that is targeted at a somewhat higher level of abstraction than Tukey’s.

The work of Grolemund and Wickham [Grolemund and Wickham, 2014] is closely related to what we propose here, and in particular, the atomic unit of data analysis discussed in Section 4. They draw on ideas from the field of cognitive science and characterize data analysis as a sense-making task whereby theories or expectations are set and then compared to reality (data). Any discrepancies between the two are further examined and then theories are possibly modified.

Our work and the work of Grolemund and Wickham (as well as Tukey and others) share a desire for a more formal theory of data analysis. However, a key distinction of our work is our focus on observable outputs from a data analysis. Grolemund and Wickham’s cognitive model is useful for describing the data analysis process and for shedding light on how it might be improved, but much of that process is typically not observed by outsiders. We attempt to characterize observed outputs—the analytic container and analytic product—of a data analysis so that individual analyses can be described and compared to other analyses in a concrete manner. Having a language that can be used to describe specific analyses has benefits in particular for teaching data analysis because it allows students and teachers to discuss the various aspects of an analysis and debate which principles should be weighted more or less heavily.

The elements and principles we have laid out here do not make for a complete framework for thinking about data analysis and leave a number of issues unresolved. We touch on some of those issues here and point to how they could be addressed in future work.

## 6.1 Honesty and Intention

The principles described above are designed to be descriptive and to allow one to characterize the wide range of data analyses that might be conducted in different fields and areas of study. In describing the principles, we do make some assumptions about the intentions of both the analyst conducting the analysis and any potential audiences that may view the analysis. However, it is clear from the historical record that some analyses are not done with the best of intentions. The possibilities range from benign neglect, to misunderstandings about methodology, to outright fraud or intentional deceit. These analyses, unsavory as their origins may be, are nevertheless data analyses. Therefore, they should be describable according to the principles outlined here.

Vignette 4 describes a scenario where an element was employed (a scatter plot) which showed evidence that appeared to contradict a previous element (the correlation coefficient). As part of the atomic unit of data analysis, the data analyst is constantly making decisions about whether to include an element in the final analysis or not. In this hypothetical example, the analyst chose not to include the scatter plot. The final analysis did provide evidence concerning the correlation between variables  $X$  and  $Y$ . However, it could be argued that the evidence is *misleading* because, unknown to us at the present time, a scatter plot would reveal that the bulk of the data exhibit a negative correlation.

The fact that an analyst has produced misleading evidence is troubling, but such an outcome does not necessarily have a one-to-one relationship with dishonest intention. On the contrary, misleading evidence can arise from even the most honest of intentions. In particular, when sample sizes are small, models do not capture hidden relationships, there is significant measurement error, or for any number of other analytical reasons, evidence can lead us to believe something for which the opposite is true. However, such situations are not generally a result of fraud or intentional deceit. They are often a result of the natural iteration and incremental advancement of science.

Can the principles be used to “detect” dishonest data analyses? Perhaps. For example, the hypothetical posed above is neither exhaustive nor skeptical. If the analyst had been forced in some way to consider other data analytic elements or alternative hypotheses, it is possible that the other aspects of the data would have been revealed. However, a truly wily analyst will be able to make an analysis appear exhaustive and skeptical, while still being misleading. There is no guarantee that dishonest analyses will always give certain principles specific weights.

Concern about dishonest analyses is of course highly relevant, but we do not envision the set of principles outlined here as being able to definitively discriminate them from honest analyses. Our goal here is to provide a method for characterizing data analyses on a *prima facie* basis, using nothing but the analysis presented. Furthermore, it is likely not possible to design an instrument, which takes solely the data analysis as its input, that can detect such phenomena. Rather, it

would make more sense to implement other methods to either deter or prevent such analyses from happening in the first place.

## 6.2 Analysis and Analyst

The characterization of a data analysis is bound to conflate two important, but distinct, entities. First is the *analysis* as presented and second is the *analyst* who conducted that data analysis. For example, it is perhaps reasonable to think that an analysis presented by an inexperienced analyst might require more scrutiny than an analysis presented by a seasoned veteran. A data analyst that has established a solid track record of analyzing data in a specific area might be more trustworthy than a data analyst with no track record. When reviewing data analysis, it is natural to consider the entire picture.

While both the analysis presented and the analyst behind it are important in the evaluation of the conclusions of an analysis, it is important to consider them separately. A key reason is because an analysis is presented to an audience, and therefore the audience by definition has all of the information about that analysis before them. Specific information about the data analyst is seldom available unless the audience has a personal relationship with the analyst or if the audience is very familiar with their work and has seen past examples. Therefore, requiring any characterization of an analysis to include information about the analyst would be entirely unworkable.

The audience may have partial information about the person doing the analysis, such as who has provided money to fund the analysis. Disclosures about funding sources are common in academic publications because their is shared understanding that conducting research or data analysis in an area where significant financial incentives are at stake can cause one to be biased in certain directions. Readers of a data analysis on the effects of smoking on lung cancer may be more skeptical of the conclusions drawn if they know the analysis was funded by a tobacco company which has a significant financial stake in selling cigarettes. However, in general, it is important that the audience does not over-correct and evaluate the entire analysis based solely on information about the analyst. The audience may of course question other things outside the analysis, such as the study design, the data collection, or various other aspects that are often more influential than what was chosen for the data analysis.

## 6.3 Future Directions

The practice of data analysis is a rich, complicated, challenging topic because involves not only the data analysis (analytic container, analytic product, or analytic presentation) built by the data analyst, but also requires consideration of contextual inputs such as the hypotheses, data, audience, and any additional constraints on the analysis. Furthermore, the practice of data analysis

also requires the ability to characterize the validity or the success of the analysis, and the strength or quality of evidence for the hypothesis of interest. Creating a framework for the elements and principles of a data analysis as defined here points us in a few clear directions towards addressing these topics.

First, this language and taxonomy for describing data analyses can be used to have a healthy debate and discussion on the larger **definition of data science** (1). We argue that describing data analyses from first principles and defining data science based on what a data scientist does is a more productive way for describing variation in data analysis and variation in the meaning of data science in a manner that spans disciplines, where different data analysts or disciplines may emphasize different elements and principles.

Second, this framework can be used to describe how a data analyst can select **informative elements** to investigate a given hypothesis, where informative is defined as, if for a given hypothesis, there is a collection of elements presented that provides evidence for or against that hypothesis (Figure 9). In contrast, if a data analyst chooses elements that do not provide evidence for or against an hypothesis, then he or she does not learn anything from the data, and the analysis is not informative. For example, in Vignette 1 the goal was to build a predictive analysis, the analyst built an inferential analysis, which is not informative for prediction. Another example is, if the analyst wants to learn if two features X and Y are correlated, they could make a histogram of X and a histogram of Y. However, these two histograms do not provide any information about the correlation of the two features (although they may provide other information about the data). Therefore, the use of histograms to address the question of correlation is not informative. Instead, a 2-dimensional scatter plot would reveal evidence for or against a hypothesis of non-zero correlation. The definition of informative is inherently making a statement about the quality of the analysis (where more informative analyses are of a higher quality), and therefore, *informativeness* is not a principle of data analyses as we have defined them here. Rather, informativeness would play a role in assessing the overall quality of the analysis.

In addition to informative analyses, one could also imagine this framework being used to understand the robustness of conclusions to the choices made by an analyst in terms of evaluating the quality of a data analysis. This might be a combination of principles, such as the combination of exhaustiveness and skepticism with the goal of critically thinking about the nature and quality of evidence that a data analyst is able to extract from a particular data source. For example, this might entail evaluating the strength and quality of evidence for the particular hypothesis of interest, or identify possible sources of bias in data.

Third, describing the variation between data analyses as variation in the weighting of different principles suggests a formal mechanism for **evaluating the success of a data analysis**. In

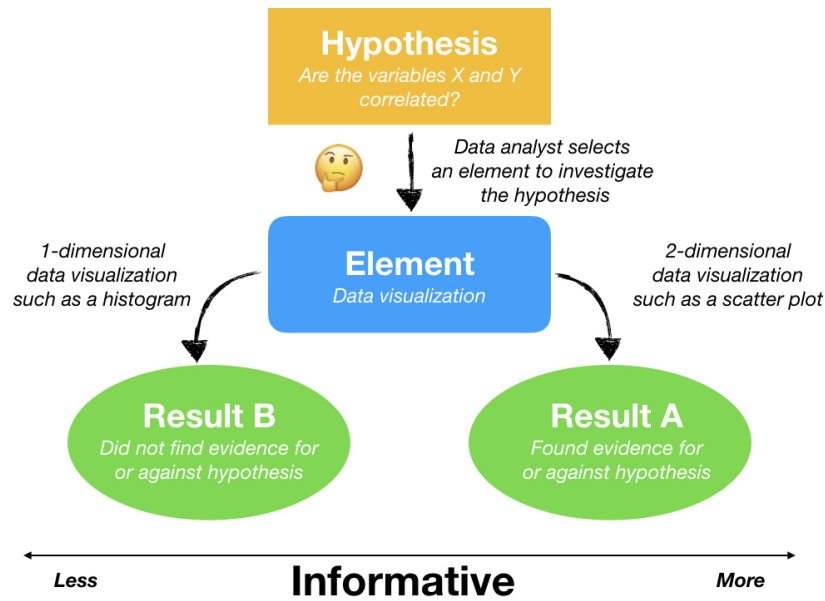


Figure 9: **Selecting informative elements in a data analysis.** An analysis is informative if for a given hypothesis, there is a collection of elements that provides evidence for or against that hypothesis. Given a hypothesis or scientific question, the analyst selects an element to investigate the hypothesis, such as a data visualization. If the analyst selects an uninformative element, such as a 1-dimensional data visualization, this would lead to not finding any evidence for or against the hypothesis. In contrast, if the analyst selects an informative element, such as a 2-dimensional data visualization, which would lead to finding evidence for or against the hypothesis.

particular, every data analysis has an audience that views the analysis and the *audience* may have a different idea of how these various principles should be weighted. One audience may value reproducibility and exhaustiveness while another audience may value interactivity and brevity. Neither set of weightings is correct or incorrect, but the success of an analysis may depend on how well-matched the analyst's weightings are to the audience's. Similarly, data analysts may be put in constrained situations where certain principles must be down-weighted or up-weighted. In Vignette 2 above, speed was of the essence and so there was little time for extensive skepticism or reproducibility. In Vignette 3 above, the substantial scrutiny placed on an analysis of that nature might demand an up-weighting of exhaustiveness, skepticism, and transparency. Regardless of the situation, an analyst who goes against the principle weightings that are demanded by the constraints may have some explaining to do. That said, audiences may be open to such explanation if the analyst can make a convincing argument in favor of a different set of weightings.

Fourth, another important area of consideration is the teaching of data science. We do not discuss this topic at length here, but the elements and principles may provide an efficient framework to **teach students data science at scale**, which is a significant problem given the demand for data science skills in the workforce. Because much data science education involves experiential learning with a mentor in a kind of apprenticeship model, there is a limit on how quickly students

can learn the relevant skills while they gain experience. Having a formal language for describing different aspects of data science that does not require mimicking the actions of a teacher or through a time-consuming mentorship may serve to compress the education of data scientists and to increase the bandwidth for training.

Finally, the development of elements and principles for data analysis provides a **foundation for a more general theory of data science**. For example, one could imagine defining mathematical or set operators on the elements of data analysis and consider the ideas of independence and exchangeability. One could define the formal projection mapping between a given data analysis and a principle of data analysis. Alternatively, one could combine one or more elements into coherent activities to define *units* or sections of a data analysis, such as the “introduction”, “setup”, “data import”, “data cleaning”, “exploratory data analysis”, “modeling”, “evaluation”, “communication”, and “export” units. There might not be a formal ordering of the units and the units can appear in a data analysis once, more than once or not at all. Then, a set of units can be assembled together into *canonical forms* of data analyses, which are likely to vary across disciplines.

## 7 Summary

The demand for data science skills has grown significantly in recent years, leading to a re-examination of the practice and teaching of data analysis. Having a formal set of elements and principles for characterizing data analyses allows data analysts to describe their work in a manner that is not confounded by the specific application or area of study. Having concrete elements and principles also opens many doors for further exploration and formalization of the data analysis process. The benefits of developing elements and principles include setting the basis for a distinct identity for the field of data science and providing a potential mechanism for teaching data science at scale.

## 8 Back Matter

### 8.1 Author Contributions

SCH and RDP equally conceptualized, wrote and approved the manuscript.

### 8.2 Acknowledgements

We would like to thank Elizabeth Ogburn, Kayla Frisoli, Jeff Leek, Brian Caffo, Kasper Hansen, Rafael Irizarry and Genevera Allen for the discussions and their insightful comments and suggestions on how to improve the presented ideas.



## References

- Ben Baumer. A data science course for undergraduates: Thinking with data. *The American Statistician*, 69:334–342, 2015.
- G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356): 791–799, 1976.
- Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215, 2001. ISSN 08834237. URL <http://www.jstor.org/stable/2676681>.
- C. Chatfield. *Problem solving: a statistician’s guide*. Chapman and Hall/CRC, 1995.
- William S. Cleveland. Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review / Revue Internationale de Statistique*, 69(1):21–26, 2001. ISSN 03067734, 17515823. URL <http://www.jstor.org/stable/1403527>.
- Drew Conway. *THE DATA SCIENCE VENN DIAGRAM*, 2010. URL <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>.
- D. Cook and D. F. Swayne. *Interactive and dynamic graphics for data analysis with R and GGobi*. Springer Publishing Company, Incorporated, 2007.
- Garrett Grolemund and Hadley Wickham. A cognitive interpretation of data analysis. *International Statistical Review*, 82(2):184–204, 2014.
- Johanna Hardin, Roger Hoerl, Nicolas J. Horton, and Deobrah Nolan. Data science in statistics curricula: Preparing students to “think with data”. *The American Statistician*, 69:343–353, 2015.
- Harlan Harris. *The Data Products Venn Diagram*, 2013. URL <http://www.datacommunitydc.org/blog/2013/09/the-data-products-venn-diagram>.
- Stephanie C. Hicks and Rafael A. Irizarry. A guide to teaching data science. *The American Statistician*, 72(4):382–391, 2018. doi: 10.1080/00031305.2017.1356747. URL <https://doi.org/10.1080/00031305.2017.1356747>.
- Michael Hochster. *What is data science?* Quora, 2014. URL <https://www.quora.com/What-is-data-science>.
- Daniel Kaplan. Teaching stats for data science. *The American Statistician*, 72(1):89–96, 2018. doi: 10.1080/00031305.2017.1398107. URL <https://doi.org/10.1080/00031305.2017.1398107>.
- Jeffery T Leek and Roger D Peng. Statistics. what is the question? *Science*, 347(6228):1314–5, 03 2015. doi: 10.1126/science.aaa6146.

- Lisa Marder. *The 7 Principles of Art and Design*, 2018. URL <https://www.thoughtco.com/principles-of-art-and-design-2578740>.
- Hilary Mason and Chris Wiggins. *A Taxonomy of Data Science*, 2010. URL <http://www.dataists.com/2010/09/a-taxonomy-of-data-science/>.
- Ulrich Matter. *Data Science in Business/Computational Social Science in Academia?*, 2013. URL <http://giventhedata.blogspot.com/2013/03/data-science-in-businesscomputational.html>.
- Deborah Nolan and Duncan Temple Lang. Computing in the statistics curricula. *The American Statistician*, 64(2):97–107, 2010. doi: 10.1198/tast.2010.09132. URL <https://doi.org/10.1198/tast.2010.09132>.
- National Gallery of Art. *The Elements of Art*. National Gallery of Art, 2019. URL <https://www.nga.gov/education/teachers/lessons-activities/elements-of-art.html>.
- R. D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 12 2011.
- PricewaterhouseCoopers. *What’s next for the data science and analytics job market?*, 2019. URL <https://www.pwc.com/us/en/library/data-science-and-analytics.html>.
- R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, A. BahnÄk, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. HÄgden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. SchlÄeter, F. D. SchÄnbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. SpÄrlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 2018. doi: 10.1177/2515245917747646.
- Brendan Tierney. *Data Science Is Multidisciplinary*, 2012. URL <https://www.oralytics.com/2012/06/data-science-is-multidisciplinary.html>.
- John W Tukey. The future of data analysis. *The annals of mathematical statistics*, 33(1):1–67, 1962.

- W. Tukey and M. B. Wilk. Data analysis and statistics: an expository overview. In *In Proceedings of the November 7-10, 1966, fall joint computer conference*, pages 695–709, 1966.
- 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*, 1999.
- American Statistical Association Undergraduate Guidelines Workgroup. *2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science*. American Statistical Association, 2014.  
URL <http://www.amstat.org/education/curriculumguidelines.cfm>.