Essays on Data Analysis

Roger D. Peng

Essays on Data Analysis

Roger D. Peng

This book is for sale at
http://leanpub.com/dataanalysisessays

This version was published on 2021-11-17

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

Also By Roger D. Peng

R Programming for Data Science

The Art of Data Science

Exploratory Data Analysis with R

Executive Data Science

Report Writing for Data Science in R

Advanced Statistical Computing

The Data Science Salon

Conversations On Data Science

Mastering Software Development in R

Tidyverse Skills for Data Science in R

Contents

I Defining Data Analysis

1. The Question

In 2011 I was teaching the course "Methods in Biostatistics" to graduate students in public health at Johns Hopkins University. The students in this course were getting Master's and PhD degrees in the variety of disciplines that make up public health and were all very bright and very motivated to learn. They would need the skills taught in my course to complete their thesis research and to do research in the future. I taught the final two terms of a year-long sequence. The topics covered in these two terms could broadly be classified as "regression modeling strategies", including linear regression, generalized linear models, survival analysis, and machine learning.

I distinctly remember that at the end of my second-to-last lecture for the entire school year, one student came up to me at the end to ask a question. She was Michelle[1], and at this point in the term I already knew she was going to ace the course. She was one of the best in the class this year. She came up to me and said, "This entire year, I feel like I've learned so many tools and techniques. But I still don't know, when I open a new dataset, **what should I do?**"

Data analysis is often taught in the negative. Don't do this, don't do that, that's a bad idea. It's rarely taught in the affirmative. You should always do this, you should definitely do that. The reason is because it's not possible to do so. Affirmative statements like that do not hold for all possible data analysis scenarios. To make use of a common phrase uttered by data analysts the world over, "It depends."

After she asked me this question I let out a nervous laugh. The truth is, I *hadn't* told her what to do. What I and the professors that came before me had done was given her a list of tools

[1] Not her real name.

and descriptions of how they worked. We taught them about when some tools were appropriate for certain kinds of data and when they were not. But we had not outlined a sequence of steps that one could take for any data analysis. The truth is, no such sequence exists. Data analysis is not a rote process with a one-size-fits-all process to follow.

But then, how is it that all these people out there are doing data analysis? How did *they* learn what to do?

When Michelle asked that question I still had one more lecture to go in the term. So I threw out whatever it was I was going to talk about and replaced it with a lecture naively titled "What to Do." When I revealed the title slide for my last lecture, I could tell that people were genuinely excited and eager to hear what I had to say. It seems Michelle was not the only one with this question.

I've tried to dig up those lecture slides but for whatever reason I cannot find them. Most likely I deleted them out of disgust! The truth is, I was unable to articulate what exactly it was that I did when I analyzed data. I had honestly never thought about it before.

At Johns Hopkins University in the Department of Biostatistics, we don't have a course specifically titled "Data Analysis". Data analysis isn't taught in courses. We teach it using a kind of apprenticeship model. As an advisor, I watch my students analyze data one at a time, tell them what they could improve, steer them when they go wrong, and congratulate them when I think they've done something well. It's not an algorithmic process; I just do what I think is right.

Some of what I think is right comes from my training in statistics. There we learned about many of the important tools for data analysis; the very same tools I was teaching in my Methods in Biostatistics course when Michelle asked her question. The guiding principle behind most of graduate education in statistics goes roughly as follows:

> If I teach you everything there is to know about a tool, using mathematics, or simulation, or real data

> examples, then you will know when and where it is appropriate to use that tool.

We repeat this process for each tool in the toolbox. The problem is that the end of the statement doesn't follow from the beginning of the statement. Just because you know all of the characteristics of a hammer doesn't mean that you know how to build a house, or even a chair. You may be able to infer it, but that's perhaps the best we can hope for.

In graduate school, this kind of training is arguably okay because we know the students will go on to work with an advisor who will in fact teach them data analysis. But it still raises the question: Why can't we teach data analysis in the classroom? Why must it be taught one-on-one with an apprenticeship model? The urgency of these questions has grown substantially in recent times with the rise of big data, data science, and analytics. Everyone is analyzing data now, but we have very little guidance to give them. We still can't tell them "what to do".

One phenomenon that I've observed over time is that students, once they have taken our courses, are often very well-versed in data analytic tools and their characteristics. Theorems tell them that certain tools can be used in some situations but not in others. But when they are given a real scientific problem with real data, they often make what we might consider elementary mistakes. It's not for a lack of understanding of the tools. It's clearly *something else* that is missing in their training.

The goal of this book is to give at least a partial answer to Michelle's question of "What should I do?" Perhaps ironically, for a book about data analysis, much of the discussion will center around things that are *outside* the data. But ultimately, that is the part that is missing from traditional statistical training, the things outside the data that play a critical role in how we conduct and interpret data analyses. Three concepts will emerge from the discussion to follow. They are **context**, **resources**, and **audience**, each of which will get their own chapter. In addition, I will discuss an expanded picture of

what data analysis is and how previous models have omitted key information, giving us only a partial view of the process.

Data analysis should be thought of as a separate field of study, but for too long it has been subsumed by either statistics or some other field. As a result, almost no time has been spent studying the process of data analysis. John Tukey, in an important (but somewhat rambling) article published in 1962 titled "The Future of Data Analysis", argued similarly that data analysis should be thought of as separate from statistics. However, his argument was that data analysis is closer to a *scientific* field rather than a *mathematical* field, as it was being treated at the time.

In my opinion, Tukey was right to say that data analysis wasn't like mathematics, but he was wrong to say that it was like science. Neither mathematics nor science provides a good model for how to think about data analysis, because data analysis is a process that is neither deductive or inductive. Rather, it is a process that solves a problem given the data available. It is the cycling back and forth of proposing solutions and asking further questions that is characteristic of the data analytic process and often results in the problem space expanding and contracting over time. There is tremendous ambiguity in most data analyses and accepting that ambiguity while also providing a useful result is one of the great challenges of data analysis.

Ultimately, what the data analyst is left with at the end is not truth (as with science) or logical certainty (as with mathematics), but a *solution*; a solution to a problem that incorporates the data we have collected. Whether we discover some fundamental law is not important to the data analyst (although it may be very important to the scientist!). This is not to say that data analysts should be disinterested in the topics on which they are working. It is just to say that their process leads to a different result than the scientists' process. The data analyst's process leads to something that other people can *use*.

One book will not answer the complex question that Michelle raised back in 2011, but if I had to answer her now, I might say, "Think of data analysis as a design process. You are

designing a solution to a problem that has been handed to you with a dataset attached. This problem has many components, constraints, and possible solutions and you will need to put it all together into something that is coherent and useful."

Here we go!

2. What is Data Analysis?

What is data analysis? It seems that sometimes even the best statisticians have a bit of difficulty answering this basic question. Consider the following from Daryl Pregibon, a long time statistics researcher at AT&T's Bell Laboratories and then later at Google, Inc. In a U.S. National Research Council report titled *The Future of Statistical Software*, he wrote,

> Throughout American or even global industry, there is much advocacy of statistical process control and of understanding processes. Statisticians have a process they espouse but do not know anything about. It is the process of putting together many tiny pieces, the process called data analysis, and is not really understood.

This report was written in 1990 but the essence of this quotation still rings true today. Pregibon was writing about his effort to develop "expert systems" that would guide scientists doing data analysis. His report was largely a chronicle of failures that could be attributed to a genuine lack of understanding of how people analyze data.

Top to Bottom

One can think of data analysis from the "top down" or from the "bottom up". Neither characterization is completely correct but both are useful as mental models. In this book we will discuss both and how they help guide us in doing data analysis.

The top down approach sees data analysis a product to be *designed*. As has been noted by many professional designers

and researchers in this area, the design process is full of vagaries, contradictions, constraints, and possibilities. This is also true of data analysis. Data analyses must be designed under a set of constraints that are ever-shifting and often negotiable. In addition, the problem that the data analyst is attempting to solve often begins life as a hazy statement that gains clarity through the data analyst proposing different solutions to the problem. The "conversation" that the data analyst has with the problem is a critical aspect of the problem and the solution.

Data analysts are tasked with designing solutions that are useful to an audience. The solution may involve summarizing evidence, visualizing data, or writing a report. Regardless of the concrete form of the solution, data analysis is a rhetorical activity that requires careful design in order to produce a convincing argument. We will discuss this top down view in the next chapter.

The bottom up approach to data analysis is perhaps more traditional and sees data analysis as a sequence of steps to achieve a result. These steps can often be modeled as a directed graph with boxes and arrows pointing boxes. This mental model is useful because it is prescriptive and gives one a sense of what needs to be done and in which order. It is the "engineer's" view of the problem, where each step can be optimized and made more efficient. However, previous depictions of this mental model have been critically incomplete and have therefore given us only a partial understanding of the data analysis process. In the section below, I will attempt to fill out this mental model a bit to better reflect what the data analyst actually does. Hopefully, with an improved model for the data analysis process, we can learn more about how data analysis is done and translate that information more rapidly to others.

A Bottom Up Mental Model

One reason for our lack of understanding about data analysis is that data analysis has been viewed from a very narrow perspective, thus constraining our ability to completely characterize it and develop generalizable lessons on how to do it. The typical picture of data analysis is that you have a rough sequence of steps:

1. The raw data come in;
2. You have to extensively wrangle the data to get it into a format suitable for analysis (perhaps a tidy format[1]);
3. Then you conduct exploratory data analysis (EDA) using various visualization and dimension reduction techniques;
4. Once EDA is complete you can goto modeling the data (if necessary);
5. Once modeling and EDA is done we can produce results;
6. The results are then communicated (maybe just to yourself).

The sequence is not wrong in the sense that data analysis encompasses all of these things. Pretty much all data analysts engage in these activities, roughly in that order. You may need to cycle back and forth a bit through the steps, especially steps 2–4.

The problem with this mental model for data analysis is that it is incomplete. If you told me that you did some data analysis, and you provided me all of the output from steps 1–6, I would have a lot of information, but *I would not have sufficient information to determine whether the analysis was good or not*. There are some key missing elements that need to be incorporated in order to have a full picture; to determine if the analysis has been *designed* properly.

Three key aspects that are missing from this mental model are context, resources, and audience. We will go through

[1] https://en.wikipedia.org/wiki/Tidy_data

each of these briefly here and in greater detail in subsequent chapters.

- **Context**. The context of a problem covers many factors, including the question that gave rise to the dataset and the motivations of the people asking the questions. Understanding what kind of solution needs to be created and what kind of answer might be useful is a key part of the context. Datasets and statistical methods can be deemed useful or not based solely on the context of a problem. Data analysts must make an effort to understand the context of a problem in order to develop a complete understanding of the problem. Getting a better understanding of the context may require communicating with subject matter experts or other people involved in posing the question.
- **Resources**. Every data analysis is conducted with a limited set of resources. The resources available determine the completeness and complexity of the solution developed. The most critical resource for a data analyst is time, but computing power, technology, and money also play a critical role. Understanding what resources were available when a data analysis was done allows us to evaluate the quality of an analysis relative to what else could have been done.
- **Audience**. Who is this data analysis for? Every data analysis has an audience, even if it is the data analyst alone. Knowing who the audience is and what they need is important for determining whether an analysis is successful or not. Further, if one considers data analysis as a rhetorical activity, then it can be important to understand how an audience receives information and evidence. Many factors can play a role here, including cultural differences, differences in background knowledge, and technical proficiency. While it may not be possible for a data analyst to know all of these things as an analysis is being conducted, the analyst must make a reasonable attempt to obtain this information.

Each of these three aspects of data analysis are “outside” the

data and statistical methodology. What this means is that a given data analysis can be viewed and evaluated differently depending on the nature of the context, resources, and audience. For example, the imputation of missing values can be useful in one context but totally inappropriate in a different context. The methodology is the same, but the context changes. Similarly, certain kinds of model results may be useful for one audience but meaningless to another audience. A data analyst can argue that they did the "right" or "optimal" thing, but ultimately an audience can choose to accept an analysis or not independent of what an analyst has done.

A complete picture of the activity of a data analyst would show a constant back and forth between the technical aspects of a data analysis and the context, resources, and audience surrounding the analysis. Designing a good data analysis requires a complete understanding of all these things, both the data and the things outside the data. *This is what makes data analysis so hard*. It requires a constant juggling of many different factors, some technical in nature and some all-too-human in nature.

The analyst mindset regularly has to deal with multiple conflicting and contradictory requirements while still producing something useful at the end. Overcoming these challenges requires solid technical knowledge as well as creativity and an ability to manage human relationships. All of this also is why data analysis is so fun. It is a very stimulating activity and can be highly rewarding when everything comes together.

In the chapters to follow I will be discussing the ideas of context, resources, and audience in greater detail. I will also present some other important aspects of doing data analysis well. My hope is that this book will serve as a useful complement to the many existing (and much more technical) books out there on statistics and data analysis.

3. Data Analysis as a Product

Data analyses are not naturally occurring phenomena. You will not run into a data analysis while walking out in the woods. Data analyses must be created and constructed by people.

One way to think about a data analysis is to think of it as a product to be designed. Data analysis is not a theoretical exercise. The goal is not to reveal something new about the world or to discover truth (although knowledge and truth may be important by-products). The goal of data analysis is to produce something useful. Useful to the scientist, useful to the product manager, useful to the business executive, or useful to the policy-maker. In that sense, data analysis is a fairly down-to-Earth activity.

Producing a useful product requires careful consideration of who will be using it. Good data analysis can be useful to just about anyone. The fact that many different kinds of people make use of data analysis is not exactly news, but what is new is the tremendous availability of data in general.

If we consider a data analysis as something to be designed, this provides for us a rough road map for how to proceed.

Questioning the Question

A favorite quotation from John Tukey, a legendary statistician and data analyst at Princeton, is

> Far better an approximate answer to the right question, which is often vague, than the exact answer

> to the wrong question, which can always be made precise.

What does these words mean in the context of data analysis? In data analysis, we often start with a dataset or a question. But good data analysts do not solve the problem that is handed to them. The reason is not necessarily arrogance. often the problem as initially stated is only a first attempt. And that's okay.

A good data analyst recognizes that the problem itself requires examination. For example, someone might ask "Is air pollution bad for your health?" That's a great question, and a critical one for public policy, but it's difficult to map to a specific data analysis. There are many different types of air pollution and there are many health outcomes we might be worried about. Prioritizing and refining the original problem is a key initial step for any data analysis. In my experience, this process usually leads to a question that is more meaningful and whose answer could lead to clear action.

The first job of the data analyst is to *discover the real underlying problem*. The fact that the problem we end up with may not be the problem we started with is not anyone's fault in particular. It's just the nature of things. Often, scientific collaborators will come to my office essentially just to talk. They come in with a clear question, but as I probe and ask questions–"What kinds of data are available?"; "What action might result in answering this question?"; "What resources are available to do this analysis?"; "Is it possible to collect new data?"–the question can shift and evolve. Good collaborators are not offended by this process, but rather appreciate it as it hones their thinking.

The bad collaborator comes in with the goal of handing off a question and waiting for the solution to appear. I have seen my fair share of these and it almost never works except perhaps with the most trivial problems. The process of designing a good data analysis cannot be modularized where we cleanly move from question, to data, to analysis, and to results, with each person involved doing their job and not

talking to anyone else. One may wish it were so, because it would be much simpler this way, but wishing doesn't make it so. This initial discussion of figuring out the right problem is an important part of designing a data analysis. But if one has a thorough discussion, we're not done questioning the question yet.

Sometimes a problem does not become clear until we've attempted to solve it. A data analyst's job is to propose solutions to a problem in order to explore the problem space. For example, in the above air pollution example, we might first express interest in looking at particulate matter air pollution[1]. But when we look at the available data we see that there are too many missing values to do a useful analysis. So we might switch to looking at ozone pollution instead, which is equally important from a health perspective.

At this point it is important that you not get too far into the problem where a lot of time or resources have to be invested. Making that initial attempt to look at particulate matter probably didn't involve more than downloading a file from a web site, but it allowed us to explore the boundary of what might be possible. Sometimes the proposed solution "just works", but more often it raises new questions and forces the analyst to rethink the underlying problem. Initial solution attempts should be "sketches", or rough models, to see if things will work. The preliminary data obtained by these sketches can be valuable for prioritizing the possible solutions and converging on the ultimate approach.

This initial process of questioning the question can feel frustrating to some, particularly for those who have come to the data analyst to get something done. Often to the collaborator, it feels like the analyst is questioning their own knowledge of the topic and is re-litigating old issues that have previously been resolved. The data analyst should be sensitive to these concerns and explain why such a discussion needs to be had. The analyst should also understand that the collaborator *is* an expert in their field and probably does know what they're

[1] https://www.epa.gov/pm-pollution

talking about. A better approach for the analyst might be to frame this process as a way for the analyst to learn more about the subject matter at hand, rather than simply challenging long-standing assumptions or beliefs. This has the co-benefit of actually being true, since the analyst is likely not an expert in the data before them. By asking simple questions in an attempt to learn the subject matter, often collaborators will be forced to re-visit some of their own ideas and to refine their thinking about the topic.

Engineering the Solution

Once we have refined the question we are asking and have carefully defined the scope of the problem we are solving, then we can proceed to engineer the solution. This will similarly be an iterative process, but we will be iterating over different things. At this point we will need a reasonably precise specification of the problem so that tools and methodologies can be mapped to various aspects of the problem. In addition, a workflow will need to be developed that will allow all of the interested parties to participate in the analysis and to play their proper role.

The analyst will need to *setup the workflow for the analysis* and adapt it to the various needs and capabilities of the collaborators. Every project will likely have a different workflow, especially if every project has a different set of collaborators involved. This is not just a comment about the tools involved for managing the workflow, but more generally about how information is exchanged and relayed to different people. Sometimes the analyst is a "central hub" through which all information is passed and sometimes it's more of a "free for all" where everyone talks to everyone else. There's no one right approach but it's important that everyone understands what approach is being used.

The analyst is also responsible for selecting the methodologies for obtaining, wrangling, and summarizing the data. One might need to setup databases to store the data or retrieve

it. Statistical methodologies might be a *t*-test for a two group comparison or a regression model to look at multivariate relationships. The selection of these tools and methodologies will be guided in part by the specification of the problem as well as the resources available and the audience who will receive the analysis. Identifying the optimal methodological approach, given the constraints put on the problem, is the unique job of the analyst. The analyst may need to delegate certain tasks based on the expertise available on the team.

Wrangling the Team

Any interesting or complex data analysis will likely involve people from different disciplines. In academia, you might be working with a biologist, an engineer, a computer scientist, and a physician at the same time. In business, you might need to interact with finance, marketing, manufacturing, and data engineering in a given analysis. A difficult part of an analyst's job is managing all of these people's interests simultaneously while integrating them into the final analysis.

The challenge facing the analyst is that each discipline is likely to think that their interests take priority over everyone else's. Furthermore, collaborators are likely to think that the problems relevant to their discipline are the most important problems to address. However, the analyst cannot accept that every discipline is "tied for first place"; priorities must be set and a reasonable assessment must made regarding which problems should addressed first, second, and so on. This is a delicate operation on the part of the analyst and doing it successfully requires open communication and good relationships with the collaborators.

The fight for priority is also why it can be so difficult for a data analysis to be modularized. If an analysis is passed from person to person, then each person is inclined to take their specific disciplinary perspective on the problem and ignore the others. This is a naturally occurring phenomenon and it

is the analyst's job to prevent it from happening, lest the analysis be watered down to be incoherent or too tightly focused on one aspect. Ultimately, the analyst must take responsibility to see the "big picture" of the analysis, weigh each collaborator's views, and choose a path that is acceptable to everyone. It's not an easy job.

4. Context Compatibility

All data arise within a particular context and often as a result of a specific question being asked. That is all well and good until we attempt to use that same data to answer a different question within a different *context*. When you match an existing dataset with a new question, you have to ask if the original context in which the data were collected is compatible with the new question and the new context. If there is *context compatibility*, then it is often reasonable to move forward. If not, then you either have to stop or come up with some statistical principle or assumption that makes the two contexts compatible. Good data analysts will often do this in their heads quickly and may not even realize they are doing it. However, I think explicitly recognizing this task is important for making sure that data analyses can provide clear answers and useful conclusions.

Understanding context compatibility is increasingly important as data science and the analysis of existing datasets continues to take off. Existing datasets all come from somewhere and it's important to the analyst to know where that is and whether it's compatible with where they're going. If there is an imcompatibility between the two contexts, which in my experience is almost always the case, then any assumption or statistical principle invoked will likely introduce *uncertainty* into the final results. That uncertainty should at least be communicated to the audience, if not formally considered in the analysis. In some cases, the original context of the data and the context of the new analysis will be so incompatible that it's not worth using the data to answer the new question. Explicit recognition of this problem can save a lot of wasted time analyzing a dataset that is ultimately a poor fit for answering a given question.

I wanted to provide two examples from my own work where context compatibility played an important role.

Example: Data Linkage and Spatial Misalignment

Air pollution data tends to be collected at monitors, which can reasonably be thought of as point locations. Air pollution data collected by the US EPA is primarily collected to monitor *regulatory compliance*. The idea here (roughly) is that we do not want any part of a county or state to be above a certain threshold of air pollution, and so we strategically place the monitors in certain locations and monitor their values in relation to air quality standards. The monitors *may* provide representative measurements of air pollution levels in the general area surrounding the monitor, but how representative they are depends on the specific pollutant being measured and the nature of pollution sources in the area. Ultimately, for compliance monitoring, it doesn't really matter how representative the monitors are because an exceedance at even one location is still a problem (the regulations have ways of smoothing out transient large values).

Health data tends to be measured at an aggregate level, particularly when it is coming from administrative sources. We might know daily counts of deaths or hospitalizations in a county or province or post code. Linking health data with air pollution data is not possible because of a context mismatch: Health data are measured areally (counts of people living within some political boundary) and pollution data are measured at point locations, so there is an incompatibility in the spatial measurement scale. We can only link these to data sources together if we do one of the following:

1. *Assume* that the monitor values are representative of population exposure in the entire county
2. Develop a *model* that can make predictions of pollution levels at all points in the county and then take the average of those values as a representative of the average county levels

This problem is well-known in spatial statistics and is referred

to as *spatial misalignment* or as *change of support*. The misalignment of the pollution and health data is the context mismatch here and arises because of the different measurement schemes that we use for each type of data. As a result, we must invoke either an assumption or a statistical model to link the two together.

The assumption of representativeness is easier to make because it requires no additional work, but it can introduce unknown uncertainties into the problem if the pollution values are *not* representative of population exposure. If a pollutant is regional in nature and is spatially homogeneous, then the assumption may be reasonable. But if there a lot of hyperlocal sources of pollution that introduce spatial heterogeneity, the assumption will not hold. The statistical modeling approach is more work, but is straightforward (in principle) and may offer the ability to explicitly characterize the uncertainty introduced by the modeling. In both cases, there is a statistical price to pay for linking the datasets together.

Data linkage is a common place to encounter context mismatches because rarely are different datasets collected with the other datasets in mind. Therefore, careful attention must be paid to the contexts within which each dataset was collected and what assumptions or modeling most be done in order to achieve context compatibility.

Example: The GAM Crisis

A common way to investigate the acute or short-term associations between air pollution levels and health outcomes is via time series analysis. The general idea is that you take a time series of air pollution levels, typically from an EPA monitor, and then correlate it with a time series of some health outcome (often death) in a population of interest. The tricky part, of course, is adjusting for the variety of factors that may confound the relationship between air pollution and health outcomes. While some factors can be measured and adjusted for directly (e.g. temperature, humidity), other

factors are unmeasured and we must find some reasonable proxy to adjust for them.

In the late 1990s investigators started using generalized additive models[1] to account for unmeasured temporal confounders in air pollution time series models. With GAMs, you could include smooth functions of time itself in order to adjust for any (smoothly) time-varying factors that may confound the relationship between air pollution and health. It wasn't a perfect solution, but it was a reasonable and highly flexible one. It didn't hurt that there was already a nice S-PLUS software implementation that could be easily run on existing data. By 2000 most investigators had standardized on using the GAM approach in air pollution time series studies.

In 2002, investigators at Johns Hopkins discovered a problem with the GAM software with respect to the default convergence criterion. The problem was that the default convergence criterion used to determine whether the backfitting algorithm used to fit the model had converged was set to 0.0001, which for most applications of GAM was more than sufficient. (In all iterative algorithms, there is some convergence parameter that determines when the algorithm stops. Usually, we monitor successive differences between iterations of the algorithm and then stop when they are small.)

The typical application of GAM was for scatterplot smoothing to look at potential nonlinearities in the relationship between the outcome and a predictor. However, in models where the nonparametric terms were highly correlated (a situation referred to as "concurvity"), then the default criterion was not stringent enough. It turned out that concurvity was quite common in the time series models and the end result was most models published in the air pollution literature had not actually converged in the fitting process.

Around this time, the US EPA was reviewing the evidence from time series studies and asked the investigators who published most of the studies to redo their analyses using alternate approaches (including using a more stringent con-

[1] https://youtu.be/f9Rj6SHPHUU

vergence criterion). To make a long story short, many of the published risk estimates dropped by around 40%, but the overall story remained the same. There was still strong evidence of an association between air pollution and health, it just wasn't as large of an effect as we once thought. Francesca Dominici wrote a comprehensive post-mortem[2] of the saga that contains many more details.

The underlying problem here was an undetected shift in context with respect to the GAM software. In previous usage of GAMs, the default convergence criterion was likely fine because there were not strong dependencies between the various smoother components in the model and the relationships being modeled did not have time series characteristics. Furthermore, the goal was not the estimation of a parameter in the model; it was more to visualize certain bivariate relationships.

However, when the same GAM software was used in a totally different context, one which the original authors likely did not foresee, suddenly the same convergence criterion was inadequate. The low-concurvity environment of previous GAM analyses was incompatible with the high-concurvity environment of air pollution time series analysis. The lesson here is that software used in a different context from which it was developed is essentially *new software*. And like any new software, it requires testing and validation.

Summary

Context shifts are critically important to recognize because they can often determine if the analyses you are conducting are valid or not. They are particularly important to discuss in data science applications here often the data are pre-existing but are being applied to a new problem or question. Methodologies and analytic approaches that are totally reasonable

[2] https://www.ncbi.nlm.nih.gov/pubmed/12142253

under one context can be inappropriate or even wrong under a different context. Finally, any assumptions made or models applied to achieve context compatibility can have an effect on the final results, typically in the form of increased uncertainty. These additional uncertainties should not be forgotten in the end, but rather should be communicated to the audience or formally incorporated in the analysis.

5. The Role of Resources

When learning about data analysis in school, you don't hear much about the role that resources such as time, money, and technology play in the development of analysis. This is a conversation that is often had "in the hallway" when talking to senior faculty or mentors. But the available resources do play a significant role in determining what can be done with a given question and dataset. It's tempting to think that the situation is binary–either you have sufficient resources to do the "right" analysis, or you simply don't do the analysis. But in the real world there a quite a few shades of gray in between those two endpoints.

All analyses must deal with constraints on time and technology and that often shapes the plan for what can be done. For example, the complexity of the statistical model being used may be constrained by computing power available to the analyst, the ability to purchase more computing power, and the time available to run, say, complex Markov chain Monte Carlo simulations.

Time

Time is usually the biggest constraint and is obviously related to money. However, even if money is in abundance, it cannot buy more time if none is available from the analyst. Complex analyses often involve many separate pieces, and complex data must be validated, checked, and interrogated before one can be confident about the results. All of this takes time and having less time leads to doing less of all those things. Deadlines are always looming and the analyst must develop a reasonable plan in the time allotted. Occasionally, it is necessary to stop an analysis if time is insufficient, but this cannot be done every time.

Similarly some analyses may require multiple people's time, if one person cannot fit it all into their schedule. If multiple people are not currently available, that will change the nature of the analysis done. Communication between collaborators must be fluid and efficient or else analyses will take longer to do.

Technology

I use the word "technology" broadly to refer to both computing resources and statistical "resources". Some models may be more optimal than others, but characteristics of the dataset (such as its size) may prevent them from being applied. Better analyses may be done with more computing power, but a constraint on available computing power will determine what models get fit and how much extra work is done.

Technological constraints may also be related to the audience that will receive the analysis. Depending on how sophisticated the audience is, one may tune the technology applied to do the analysis. It may not be worth applying complex methodology if the audience for the analysis will not be able to easily interpret the results.

The analyst may need to get creative to overcome constraints on technology. Often approximations may be used that are faster to compute but sacrifice a little accuracy. If reasonable justification can be made for the loss in accuracy, then such an approximation may be a good way to circumvent a time constraint.

Trustworthiness

Resource constraints can affect how *trustworthy* the analysis is. In a trustworthy analysis, what is presented as the analysis is often backed up by many facts and details that are not presented. These other analyses have been done, but the

analyst has decided (likely based on a certain *narrative* of the data) that they do not meet the threshold for presentation. That said, should anyone ask for those details, they are readily available. With greater resources, the sum total of all of the things that can be done is greater, thus giving us hope that the things left *undone* are orthogonal to what was done.

However, with fewer resources, there are at least two consequences. First, it is likely that fewer things can be done with the data. Fewer checks on the data, checks on model assumptions, checks of convergence, model validations, etc. This increases the number of undone things and makes it more likely that they will have an impact on the final (presented) results. Secondly, certain kinds of analysis may require greater time or computing power than is available. In order to present any analysis at all, we may need to resort to approximations or "cheaper" methodology. These approaches are not necessarily incorrect, but they may produce results that are noisier or otherwise sub-optimal.

That said, the various other parties involved in the analysis, such as the audience or the patron[1], may prefer having any analysis done, regardless of optimality, over having no analysis. Sometimes the question itself is still vague or a bit rough, and so it's okay if the analysis that goes with it is equally "quick and dirty". Nevertheless, analysts have to draw the line between what what is a reasonable analysis and what is not, given the available resources. If the only approach that is feasible is likely to produce very weak evidence with great uncertainty, the analyst may need to decide that the analysis is not worth doing at all

The Analyst's Job

The data analyst's job is to manage the resources available for analysis. The availability of resources may not be solely up to the analyst, but the job is nevertheless to recognize what

[1] https://simplystatistics.org/2018/04/30/relationships-in-data-analysis/

is available, determine whether the resources are sufficient for completing a reasonable analysis, and if not, then request more from those who can provide them. I've seen many data analyses go astray as a result of a mismatch in the understanding of the resources available versus the resources required.

A good data analyst can minimize the chance of a gross mismatch and will continuously evaluate the resource needs of an analysis going forward. If there appears to be a large deviation between what was expected and the reality of the analysis, then the analyst must communicate with others involved (the patron or perhaps subject matter expert) to either obtain more resources or modify the data analytic plan.

6. The Role of the Audience

Statisticians, in my experience, do not often discuss the topic of *who* will be the recipient of their data analysis. But the audience for a data analysis plays an important role in determining the success of the analysis. One important measure of success in data analysis is that the *audience* two whom it is presented *accepts* the results. I would pose this as a necessary condition, but not sufficient. There are a number of ideas to unpack here, so I will walk through them. Two key notions that I think are important are the notions of *acceptance* and the *audience*.

Acceptance

The first idea is the notion of *acceptance*. It's tempting to confuse this with *belief*, but they are two different concepts that need to be kept separate. Acceptance of an analysis involves the analysis itself–the data and the methods applied to it, along with the narrative told to explain the results. Belief in the results depends on the analysis itself as well as many other things outside the analysis, including previous analyses, existing literature, and the state of the science (in purely Bayesian terms, your *prior*). A responsible audience can accept an analysis without necessarily believing its principal claims, but these two concepts are likely to be correlated.

For example, suppose a team at your company designs an experiment to collect data to determine if lowering the price of a widget will have an effect on profits for your widget-making company. During the data collection process, there was a problem which resulted in some of the data being missing in a potentially informative way. The data are then handed to you. You do your best to account for the missingness and the

resulting uncertainty, perhaps through multiple imputation or other adjustment methods. At the end of the day, you show me the analysis and conclude that lowering the price of a widget will increase profits 3-fold.

I may accept that you did the analysis correctly and trust that you did your best to account for the problems encountered during collection using state-of-the-art methods. But I may disagree with the conclusion, in part because of the problems introduced with the missing data (not your fault), but also in part because we had previously lowered prices on another product that we sell and there was no corresponding increase in profits. Given the immense cost of doing the experiment, I might ultimately decide that we should abandon trying to modify the price of widgets and leave things where they are (at least for now). Your analysis was a success.

This simple example illustrates two things. First, acceptance of the analysis depends primarily on the details of the analysis and my willingness to *trust* what the analyst has done. Was the missing data accounted for? Was the uncertainty properly presented? Can I reason about the data and understand how the data influence the results? Second, my belief in the results depends in part on things *outside* the analysis, things that are primarily outside the analyst's control. In this case, these are the presence of missing data during collection and a totally separate experience lowering prices for a different product. How I weigh these external things, in the presence of your analysis, is a personal preference.

Acceptance vs. Validity

In scientific contexts it is tempting to think about *validity*. Here, a data analysis is successful if the claims made are true. If I analyze data on smoking habits and mortality rates and conclude that smoking causes lung cancer, then my analysis is successful if that claim is true. This definition has the advantage that it removes the subjective element of acceptance, which depends on the audience to which an analysis is presented. But validity is an awfully high bar to meet for

any given analysis. In this smoking example, initial analyses of smoking and mortality data could not be deemed successful or not until decades after they were done. Most scientific conclusions require multiple replications occurring over many years by independent investigators and analysts before the community believes or concludes that they are true. Leaving data analysts in limbo for such a long time seems impractical and, frankly, unfair. Furthermore, I don't think we want to penalize data analysts for making conclusions that turn out to be false, as long as we believe they are doing good work. Whether those claims turn out to be true or not may depend on things outside their control.

A related standard for analyses is essentially a notion of intrinsic validity. Rather than wait until we can validate a claim made by an analysis (perhaps decades down the road), we can evaluate an analysis by whether the *correct* or *best* approach was done and the correct methods were applied. But there are at least two problems with this approach. In many scenarios it is not possible to know what is the best method, or what is the best combination of methods to apply, which would suggest that in many analyses, we are uncertain of success. This seems rather unsatisfying and impractical. Imagine hiring a data analyst and saying to them "In the vast majority of analyses that you do, we will not know if you are successful or not."

Second, even in the ideal scenarios, where we know what is correct or best, intrinsic validity is necessary but far from sufficient. This is because the *context* in which an analysis performed is critical in understanding what is appropriate. If the analyst is unaware of that context, they may make critical mistakes, both from an analytical and interpretative perspective. However, those same mistakes might be innocuous in a different context. It all depends, but the analyst needs to know the difference.

One story that comes to mind comes from the election victory of George W. Bush over Al Gore in the 2000 United States presidential election. That election hinged on votes counted in the state of Florida, where Bush and Gore were

very close. Ultimately, lawsuits were filed and a trial was set to determine exactly how the vote counting should proceed. Statisticians were called to testify for both Bush and Gore. The statistician called to testify for the Gore team was Nicolas Hengartner, formerly of Yale University (he was my undergraduate advisor when I was there). Hengartner presented a thorough analysis of the data that was given to him by the Gore team and concluded there were differences in how the votes were being counted across Florida and that some ballots were undercounted.

However, on cross-examination, the lawyer for Bush was able to catch Hengartner in a "gotcha" moment which ultimately had to do with the manner in which the data were collected, about which Hengartner had been unaware. Was the analysis a success? It's difficult to say without having been directly involved. Nobody challenged the methodology that Hengartner used in the analysis, which was by all accounts a very simple analysis. Therefore, one could argue that it had intrinsic validity. However, one could also argue that he should have known about the issue with how the data were collected (and perhaps the broader context) and incorporated that into his analysis and presentation to the court. Ultimately, Hengartner's analysis was only one piece in a collection of evidence presented and so it's difficult to say what role it played in the ultimate outcome.

Audience

All data analyses have an audience, even if that audience is you. The audience may accept the results of an analysis or they may fail to accept it, in which case more analyses may need to be done. The fact that an analyst's success may depend on a person different from the analyst may strike some as an uncomfortable feature. However, I think this is the reality of all data analyses. Success depends on human beings, unfortunately, and this is something analysts must be prepared to deal with. Recognizing that human nature plays

a key role in determining the success of data analysis explains a number of key aspects of what we might consider to be good or bad analyses.

Many times I've had the experience of giving the same presentation to two different audiences. One audience loves it while the other hates it. How can that be if the analyses and presentation were exactly the same in both cases? The truth is that an analysis can be accepted or rejected by different audiences depending on who they are and what their expectations are. A common scenario involves giving a presentation to "insiders" who are keenly familiar with the context and the standard practices in the field. Taking that presentation verbatim to an "outside" audience that is less familiar will often result in failure because they will not understand what is going on. If that outside audience expects a certain set of procedures be applied to the data, then they may demand that you do the same, and refuse to accept the analysis until you do so.

I vividly remember one experience that I had presenting the analysis of some air pollution and health data that I had done. In practice talks with my own group everything had gone well and I thought things were reasonably complete. When giving the same talk to an outside group, they refused to accept what I'd done (or even interpret the results) until I had also run a separate analysis using a different kind of spline model. It wasn't an unreasonable idea, so I did the separate analysis and in a future event with the same group I presented both analyses side by side. They were not wild about the conclusions, but the debate no longer centered on the analyses themselves and instead focused on other scientific aspects. In retrospect, I give them credit for accepting the analyses even if they did not necessarily believe the conclusion.

Summary

I think the requirement that a successful data analysis be accepted by its audience is challenging (and perhaps unsettling) because it suggests that data analysts are responsible for

things outside the data. In particular, they need to understand the context around which the data are collected and the audience to which results will be presented. I also think that's why it took so long for me to come around to it. But I think this definition explains much more clearly *why it is so difficult to be a good data analyst*. When we consider data analysis using traditional criteria developed by statisticians, we struggle to explain why some people are better data analysts than others and why some analyses are better than others. However, when we consider that data analysts have to juggle a variety of factors both internal and external to the data in order to achieve success, we see more clearly why this is such a difficult job and why good people are hard to come by.

Another implication of this definition of data analysis success is that it suggests that human nature plays a big role and that much of successful data analysis is essentially a successful negotiation of human relations. Good communication with an audience can often play a much bigger role in success than whether you used a linear model or quadratic model. Trust between an analyst and audience is critical when an analyst must make choices about what to present and what to omit. Admitting that human nature plays a role in data analysis success is difficult because humans are highly subjective, inconsistent, and difficult to quantify. However, I think doing so gives us a better understanding about how to judge the quality of data analyses and how to improve them in the future.

7. The Data Analysis Narrative

Data analysis is supposed to be about the data, right? Just the facts? And for the most part it is, up until the point you need to communicate your findings to an audience. The problem is that in any data analysis that would be meaningful to others, there are simply too many results to present, and so *choices* must be made. Depending on who the audience is, or who the audience is composed of, you will need to tune your presentation in order to get the audience to accept the analysis. How is this done?

In data analysis we often talk about dimension reduction. The idea here is that when we have a lot variables or predictors in a dataset, we want to reduce the number of variables that we are looking at without eliminating too much information in the dataset. The reason we can often do this is because many variables will be correlated with each other and therefore provide redundant information. Methods like principal components analysis, multi-dimensional scaling, or canonical correlation analysis are often used for dimension reduction and exploratory analysis.

Narrative serves as dimension reduction for the results in a data analysis. There will be many different kinds of results in a dataset–tables, plots, diagrams–coming from different models, different approaches, and different assumptions. We must figure out some way to choose between all of these different results. Often, the story that we tell about the data and the evidence guide us in choosing what to present. However, things can go wrong if the narrative is not chosen carefully.

When Narratives Go Wrong

In the worst case scenario, the narrative is plainly false and the presenter resorts to trickery and obfuscation. Graphs with messed up axes, or tables that obscure key data; we all know the horror stories. A sophisticated audience might detect this kind of trickery and reject the analysis, but maybe not. That said, let's assume we are pure of heart. How does one organize a presentation to be successful?

We all know another other horror story, which is the *data dump*. Here, the analyst presents everything they have done and essentially shifts the burden of interpretation on to the audience. In this case, the analyst has failed to produce any narrative and therefore has not summarized the results. Rarely is this desired. In some cases the audience will just want the data to do their own analyses, but then there's no need for the analyst to waste their time doing *any* analysis. They can just hand over the data.

Decisions, Decisions, Decisions

Ultimately, the analyst must *choose what to present*, and this can cause problems. The choices must be made to fit the analyst's narrative of "what is going on with the data". They will choose to include some plots and not others and some tables and not others. These choices are directed by a narrative and an interpretation of the data.

When an audience is upset by a data analysis, and they are being honest, they are usually upset with the chosen narrative, not with the facts per se. They will be upset with the combination of data that the analyst chose to include and the data that the analyst chose to *exclude*. Why didn't you include *that* data? Why is this narrative so focused on this or that aspect?

The choices that an analyst make are a critical form of leader-

ship. Decisions about what to present will often move other aspects of a project forward and so settling on the narrative for an analysis is very important. Deciding what the data and the results *mean*, and hence what story they tell, is also dependent on many things outside the data (the context in particular), and so the analyst must be in contact with other collaborators to make sure that the narrative matches these other aspects.

As someone who has some control over how the narrative of a data analysis is crafted, the data analyst exercises an important power. Care must be taken to make sure that input from collaborators is taken into account and used appropriately to shape the narrative. Often team members will want to force the narrative in one direction or another, sometimes in a direction not supported by the data. This situation is not unlike the one early on in the data analysis where different disciplines will argue over who's discipline takes priority. In either case, the analyst must be careful to hear everyone, but construct the narrative only in a manner that is supported by the data.

8. The Four Jobs of the Data Scientist

In 2019 I wrote a post about The Tentpoles of Data Science[1] that tried to distill the key skills of the data scientist. In the post I wrote:

> When I ask myself the question "What is data science?" I tend to think of the following five components. Data science is (1) the application of design thinking to data problems; (2) the creation and management of workflows for transforming and processing data; (3) the negotiation of human relationships to identify context, allocate resources, and characterize audiences for data analysis products; (4) the application of statistical methods to quantify evidence; and (5) the transformation of data analytic information into coherent narratives and stories.
>
> My contention is that if you are a good data scientist, then you are good at all five of the tentpoles of data science. Conversely, if you are good at all five tentpoles, then you'll likely be a good data scientist.

I still feel the same way about these skills but my feeling now is that actually that post made the job of the data scientist seem easier than it is. This is because it wrapped all of these skills into a single job when in reality data science requires being good at **four** jobs. In order to explain what I mean by this, we have to step back and ask a much more fundamental question.

[1] https://simplystatistics.org/2019/01/18/the-tentpoles-of-data-science/

What is the Core of Data Science?

This is a question that everyone is asking and I think struggling to answer. With any field there's always a distinction between the questions of

- What is the core of the field?
- What do people in that field do on a regular basis?

In case it's not clear, these are not the same question. For example, in Statistics, based on the curriculum from most PhD program, the core of the field involves statistical methods, statistical theory, probability, and maybe some computing. Data analysis is generally not formally taught (i.e. in the classroom), but rather picked up as part of a thesis or research project. Many classes labeled "Data Science" or "Data Analysis" simply teach more methods like machine learning, clustering, or dimension reduction. Formal software engineering techniques are also not generally taught, but in practice are often used.

One could argue that data analysis and software engineering is something that statisticians *do* but it's not the core of the field. Whether that is correct or incorrect is not my point. I'm only saying that a distinction has to be made somewhere. Statisticians will always *do* more than what would be considered the core of the field.

With data science, I think we are collectively taking inventory of what data scientists tend to do. The problem is that at the moment it seems to be all over the map. Traditional statistics does tend to be central to the activity, but so does computer science, software engineering, cognitive science, ethics, communication, etc. This is hardly a definition of the core of a field but rather an enumeration of activities.

The question then is can we define something that *all* data scientists do? If we had to teach something to all data science students without knowing where they might end up afterwards, what would it be? My opinion is that at some point,

all data scientists have to engage in the *basic data analytic iteration*.

Data Analytic Iteration

The basic data analytic iteration comes in four parts. Once a question has been established and a plan for obtaining or collecting data is available, we can do the following:

1. Construct a **set of expected outcomes**
2. Design/Build/Apply a **data analytic system** to the data
3. Diagnose any **anomalies** in the analytic system output
4. Make a **decision** about what to do next

While this iteration might be familiar or obvious to many, its familiarity masks the complexity involved. In particular, each step of the iteration requires that the data scientist play a different role involving very different skills. It's like a one-person play where the data scientist has to change costumes when going from one step to the next. This is what I refer to as the *the four jobs of the data scientist*.

The Four Jobs

Each of the steps in the basic data analytic iteration requires the data scientist to be four different people: (1) Scientist; (2) Statistician; (3) Systems Engineer; and (4) Politician. Let's take a look at each job in greater detail.

Scientist

The scientist develops and asks the question and is responsible for knowing the state of the science and what the key gaps are. The scientist also designs any experiments for collecting

new data and executes the data collection. The scientist must work with the statistician to design a system for analyzing the data and ultimately construct a *set of expected outcomes* from any analysis of the data being collected.

The scientist plays a key role in developing the system that results in our set of expected outcomes. Components of this system might include a literature review, meta-analysis, preliminary data, or anecdotal data from colleagues. I use the term "Scientist" broadly here. In other settings this could be a policy-maker or product manager.

Statistician

The statistician, in concert with the scientist, designs the analytic system that will analyze the data generated by any data collection efforts. They specify how the system will operate, what outputs it will generate, and obtain any resources needed to implement the system. The statistician draws on statistical theory and personal experience to choose the different components of the analytic system that will be applied.

The statistician's role here is to apply the data analytic system to the data and to produce the data analytic output. This output could be a regression coefficient, a mean, a plot, or a prediction. But there must be something produced that we can compare to our set of expected outcomes. If the output deviates from our set of expected outcomes, then the next task is to identify the reasons for that deviation.

Systems Engineer

Once the analytic system is applied to the data there are only two possible outcomes:

1. The outputs meet our expectations, or
2. The output does not meet our expectations (an anomaly).

In the case of an anomaly, the systems engineer's responsibility is to diagnose the potential root causes of the anomaly, based on knowledge of the data collection process, the analytic system, and the state of scientific knowledge.

Recently, Emma Vitz wrote on Twitter[2]:

> How do you teach debugging to people who are more junior? I feel like it's such an important skill and yet we seem to have no structured framework for teaching it

For software and for data analysis alike, the challenge is that bugs or unexpected behavior can originate from anywhere. Any complex system is composed of multiple components, some of which may be your responsibility and many of which are someone else's. But bugs and anomalies do not respect those boundaries! There may be an issue that occurs in one component that only becomes known when you see the data analytic output.

So if you are responsible for diagnosing a problem, it is your responsibility to investigate the behavior of each component of the system. If it is something you are not that familiar with, then you need to *become* familiar with it, either by learning on your own or (more likely) talking to the person who is in fact responsible.

A common source of unexpected behavior in data analytic output is the data collection process, but the statistician who analyzes the data may not be responsible for that aspect of the project. Nevertheless, the systems engineer who identifies an anomaly has to go back through and talk to the statistician and the scientist to figure out exactly how each component works.

Ultimately, the systems engineer is tasked with taking a broad view of all the activities that affect the output from a data analysis in order to identify any deviations from what we

[2] https://twitter.com/EmmaVitz/status/1330697959156027392?s=20

would expect. Once those root causes have been explained, we can then move on to decide how we should act on this new information.

Politician

The politician's job is to make decisions while balancing the needs of the various constituents to achieve a reasonable outcome. Most statisticians and scientist that I know would recoil at the idea of being considered a politician or that politics in any form would play a role in doing any sort of science. However, my thinking here is a bit more basic: In any data analysis iteration, we are constantly making decisions about what to do, keeping in mind a variety of conflicting factors. In order to resolve these conflicts and come to a reasonable agreement, one has to engage a key skill, which is negotiation.

At various stages of the data analytic iteration the politician must negotiate about (a) the definition of success in the analysis; (b) resources for executing the analysis; and (c) the decision for what to do after we have seen the output from the analytic system and have diagnosed the root causes of any anomalies. Decisions about what to do next fundamentally involve factors outside the data and the science.

Politicians have to identify who the stakeholders of the problem are and what is it that they ultimately *want* (as opposed to what their *position* is). For example, an investigator might say "We need a p-value less than 0.05". That's their position. But what they *want* is more likely "a publication in a high profile journal". Another example might be an investigator who needs to meet a tight publication deadline while another investigator who wants to run a time-consuming (but more robust) analysis. Clearly, the positions conflict but arguably both investigators share the same goal, which is a rigorous high-impact publication.

Identifying positions versus underlying needs is a key task in negotiating a reasonable outcome for everyone involved.

Rarely, in my experience, does this process have to do with the data (although data may be used to make certain arguments). The dominating elements of this process tend to be the nature of relationships between each constituent and the constraints on resources (such as time).

Applying the Iteration

If you're reading this and find yourself saying "I'm not an X" where X is either scientist, statistician, systems engineer, or politician, then chances are that is where you are weak at data science. I think a good data scientist has to have some skill in each of these domains in order to be able to complete the basic data analytic iteration.

In any given analysis, the iteration may be applied anywhere from once to dozens if not hundreds of times. If you've ever made two plots of the same dataset, you've likely done two iterations. While the exact details and frequency of the iterations may vary widely across applications, the core nature and the skills involved do not change much.

II Data Analytic Process

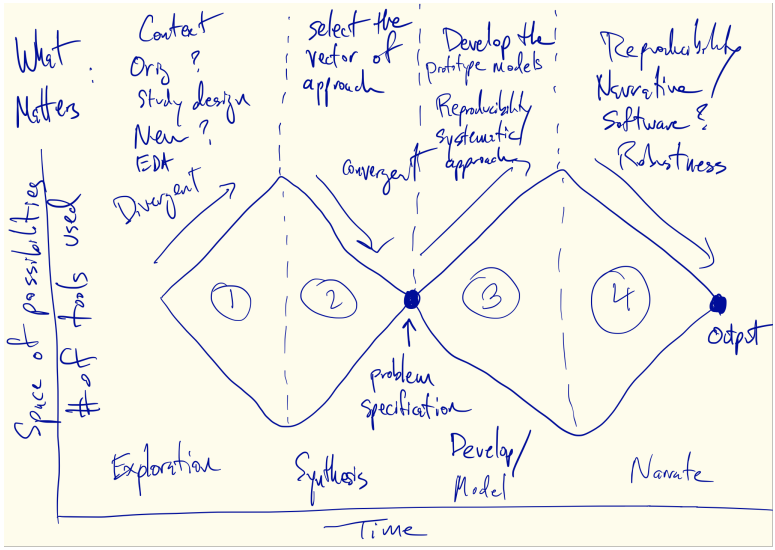
9. Divergent and Convergent Phases of Data Analysis

There are often discussions within the data science community about which tools are best for doing data science. But one thing that is often missing from many discussions about tooling in data analysis is an acknowledgment that data analysis tends to advance through different phases and that different tools can be more or less useful in each of those phases. In fact, a tool that is very useful in one phase can be less useful or even detrimental in other phases. My goal for this chapter is to expand a bit on this idea.

A Double Diamond for Data Analysis

One image that is commonly found in the literature on design thinking is the "double diamond"[1], which is a model for the design process. I think this model can be usefully adapted to help us think about data analysis and my best representation is here (apologies for the scribble).

[1] https://www.designcouncil.org.uk/news-opinion/design-process-what-double-diamond



Double Diamond for Data Analysis

In this figure I've identified four phases of data analysis that alternate between divergent and convergent forms of thinking. The x-axis is roughly "time" or the timeline of the analysis. The y-axis can be thought of as indicating the range of possibilities. Okay, it's not exactly a numerical graph, but you get the idea. The wide parts of each diamond are meant to indicate that there are many possibilities under consideration while the narrow "choke points" indicate that there should be one or a few things under consideration.

How does all this relate to data analysis? Let's first take a closer look at the four phases.

Phase 1: Exploration

The goal of Phase 1 is to explore the possibilities inherent in the data. This part is familiar to all data analysts.

The dataset lands in your lap and there are a lot of things to do and a lot of questions to answer. Did you get the right dataset? Are all the data there? Are they in the right format for

analysis? This is often where a lot of data wrangling occurs. We must consider what question is being asked and whether the data are appropriate for that question[2] (at least without making unreasonable assumptions). We might also consider what question motivated the creation of the data[3].

At this point we may think we have a question to ask, but usually that question is only vaguely formed or is in need of further information. Here is where we need the data to help us out. For example, an important general question is "Can the data even be used to answer my question?" We need to look at the data to figure that out. How we look at the data will vary across people, context, and many other factors. Regardless of the specific situation, we likely will need to make lots and lots of plots, summaries, and tables. We'll need to look at the data, perhaps even in a program like Excel, and just get a sense of the data.

This phase of analysis is highly divergent, with many possibilities being considered for how to ask the question and what approach to take. In my experience, I'm making tons of plots and looking at various transformations of individual variables and bivariate relationships. I've never bothered to count, but it wouldn't surprise me if there were thousands of plots made in this phase. This is the "sketching" phase of analysis, figuratively speaking, but sometimes literally speaking. Sketches of plots or tables are often a useful planning device.

This phase of analysis is almost always fun, because we are opening up the possibilities. But all good things must eventually come to an end.

Phase 2: Refining the Problem

Phase 2 is challenging because it involves making decisions and choices. Nobody likes to do that. Inevitably, most of the work you did in Phase 1 will be left on the cutting room floor.

[2] https://simplystatistics.org/2018/05/24/context-compatibility-in-data-analysis/

[3] https://simplystatistics.org/2018/07/06/data-creators/

You might love all your children equally but you still need to pick a favorite. The reason is nobody has the resources[4] to pursue every avenue of investigation. Furthermore, pursuing every avenue would likely not be that productive. You are better off sharpening and refining your question. This simplifies the analysis in the future and makes it much more likely that people (including you!) will be able to act on the results that you present.

This phase of analysis is convergent and requires synthesizing many different ideas into a coherent plan or strategy. Taking the thousands of plots, tables, and summaries that you've made and deciding on a problem specification is not easy, and to my surprise, I have not seen a lot of tools dedicated to assisting in this task. Nevertheless, the goal here is to end up with a reasonably detailed specification of what we are trying to achieve and how the data will be used to achieve it. It might be something like "We're going to fit a linear model with this outcome and these predictors in order to answer this question" or "We're building a prediction model using this collection of features to optimize this metric".

In some settings (such as in consulting) you might need to formally write this specification down and present it to others. At any rate, you will need to justify it based on your exploration of the data in Phase 1 and whatever outside factors may be relevant. Having a keen understanding of your audience[5] becomes relevant at the conclusion of this phase.

Phase 3: Model Development

This phase is the bread and butter of most statisticians and statistics education programs. Here, we have a reasonably well-specified problem, a clear question, an appropriate dataset, and we're going to engineer the solution. But that doesn't mean we just push a button and wait for the result to come out.

[4] https://simplystatistics.org/2018/06/18/the-role-of-resources-in-data-analysis/

[5] https://simplystatistics.org/2018/04/17/what-is-a-successful-data-analysis/

For starters, what will the results look like? What summaries do we want to produce and how will they be presented? Having a detailed specification is good, but it's not final. When I was a software engineer, we often got highly detailed specifications of the software we were supposed to build. But even then, there were many choices to make.

Thus begins another divergent phase of analysis, where we typically build models and gauge their performance and robustness. This is the data analyst's version of prototyping. We may look at model fit and see how things work out relative to our expectations set out in the problem specification. We might consider sensitivity analyses or other checks on our assumptions about the world and the data. Again, there may be many tables and plots, but this time not of the data, but of the results. The important thing here is that we are dealing with concrete models, not the rough "sketches" done in Phase 1.

Because the work in this phase will likely end up in some form in the final product, we need to develop a more formal workflow and process to track what we are doing. Things like version control play a role, as well as scriptable data analysis packages that can describe our work in code. Even though many aspects of this phase still may not be used, it is important to have reproducibility in mind as work is developed so that it doesn't have to be "tacked on" after the fact (an often painful process).

Phase 4: Narration

In the final convergent phase of data analysis, we must again make choices amongst the many pieces of work that we did in Phase 3. We must choose amongst the many models and results and decide what will make it into the final product, whether it is a paper, a report, a web site, or a slide deck.

In order to make these choices, it is useful to develop a *narrative* of the analysis. Building the narrative is dimension reduction for results and it allows you to choose from the

various results the ones that follow your narrative. Simply "presenting the data" is first, not really possible, and second, not desirable. It's information overload and rarely allows the audience to make an intelligent conclusion. Ultimately, the analyst must decide on a narrative and select the various results that tell that story.

Selecting a narrative doesn't mean that everything else is thrown out. There are often many parts of an analysis that cannot make it into the main product but can be put in some form of "supplementary materials". For example, FiveThirtyEight.com[6] doesn't put its election statistical model on its web site, it puts it in a PDF document that you can download separately. In many scientific papers, there can be dozens of tables in the supplementary materials that didn't make it into the paper. Many presenters often have backup slides hidden at the end of the deck in case there are questions. Some people may disagree with the choice of narrative, but that doesn't eliminate the need for choices to be made.

Implications

At this point it's worth recalling that all models are wrong, but some are useful. So why is this model for data analysis useful? I think there are a few areas where the model is useful as an explanatory tool and for highlighting where there might be avenues for future work.

Decision Points

One thing I like about the visual shape of the model is that it highlights two key areas–the ends of both convergent phases–where critical decisions have to be made. At the end of Phase 2, we must decide on the *problem specification*, after sorting through all of the exploratory work that we've done. At the

[6] https://fivethirtyeight.com

end of Phase 4, we must decide on the *narrative* that we will use to explain the results of an analysis.

At both decision points, we must balance a set of competing priorities and select from a large number of outputs to create a coherent product. When developing the problem specification, we must avoid the desire to say "We're going to do all the things" and when producing the final product we must avoid the "data dump" where we shift the burden of interpretation and distillation on to the audience.

In my experience, data analysts really enjoy the divergent phases of analysis because no choices have to be made. In fact, in those phases we're doing the *opposite* of making choices– we're trying out everything. But the convergent phases cannot be avoided if we want to produce good data analysis.

Tooling

When it comes to discussions about tooling in data science it's useful to preface such discussion with an indication of which phase of data analysis we are talking about. I cannot think of a single tool that is simultaneously optimal for every phase of analysis. Even R is sub-optimal for Phases 2 and 4 in my opinion. For example, when developing data visualizations for presentation, many people resort to tools like Adobe Illustrator.

I find most debates on tooling to be severely confounded by a lack of discussion about what phase of analysis we are focusing on. It can be difficult for people to accept that a given tool may be optimal for one phase but detrimental in another phase. I think the four-phase model for analysis can explain some of the recent "debates" (a.k.a flame wars) on tooling in the data science world.

Let's start with notebooks. Many people argue that notebooks (like Juypter notebooks) are bad for data analysis because of problems with reproducibility. I think the issue here is essentially that notebooks can be useful for Phase 1 but potentially

dangerous for Phase 3. Reproducibility is a critical consideration for developing the final results, but is less important when exploring the data in Phase 1. The need for consistent state and in-order execution of commands is more important in Phases 3 than in Phase 1 when lots of ideas are being tried and will likely be thrown out. There's no either-or decision to be made here with regard to notebooks. It just depends on what phase of analysis you're in. I generally prefer R Markdown but can see the attractiveness of notebooks too.

Consider the age-old base graphics vs. ggplot2[7] debate. This is another debate confounded by the phases of data analysis. In the divergent phases (like Phase 1), there is a need to do things very quickly so that good ideas can be discovered and bad ideas can be discarded. Having a system like ggplot2 that allows you to rapidly "sketch" out these ideas is important. While the base graphics system gives you fine control over every aspect of a plot, such a "feature" is really a detriment in Phase 1 when you're exploring the data. Nathan Yau seems to edge towards base graphics[8] but I think this is largely because he is considering developing "final" data presentations in Phase 4. If you look at the visualizations on his web page, they are akin to works of art. They are not sketches. When developing those kinds of visualizations, having fine control over every detail is exactly what you want.

Education

In my experience as both a student and a teacher, statistical education tends to focus on Phases 3 and 4 of the data analysis process. These phases are obviously important and can be technically demanding, making them good things to emphasize in a statistics program. In the classroom, we typically start with a reasonably well-specified problem and proceed with engineering the solution. We discuss the many options for developing the solution and the various strengths and

[7] https://simplystatistics.org/2016/02/11/why-i-dont-use-ggplot2/

[8] https://flowingdata.com/2016/03/22/comparing-ggplot2-and-r-base-graphics/

weaknesses of those options. Because the work in Phase 3 ends up in the final results, it's important to get this phase right.

When getting a graduate degree, Phases 1 and 2 are usually taught as part of developing the thesis using an apprenticeship model with an advisor ("just watch what I do for a few years"). Some good programs have courses in exploratory data analysis that fall squarely into Phase 1, but I don't think I've ever seen a course that covers Phase 2. Phase 4 is sometimes covered in visualization courses which discuss the design process for good data presentation. There are also a number of good books that cover this phase. But we rarely discuss the role of developing a coherent narrative of an analysis in the classroom. Usually, that is covered using an apprenticeship approach.

Data Analytic Errors

One general problem that occurs when I see others do data analysis is confusing the different phases. It's common for people to think that they are in one phase when they're really in another. This can lead to data analytic problems and even mistakes.

Confusing Phase 3 for Phase 1 is arguably the most common and the most dangerous problem that one can encounter in science. In Phase 1 we are developing the specification of the problem, but in Phase 3 that specification should be more or less locked down. If we discover something else interesting in Phase 3, we cannot pretend like that's what we *meant* to conclude in Phase 1. That's a recipe for p-hacking. Phase 1 should be broad and exploratory with the data. But once you've decided on a problem in Phase 2, Phases 3 and 4 should be solely dedicated to solving that problem. Changing up the problem in mid-stream in Phases 3 and 4 is often tempting but almost always a mistake.

A related problem is confusing Phase 2 with Phase 4. I have observed people who explore the data extensively, making

quick plots, sketches, and summaries, and then immediately draw a conclusion like "X causes Y" or "A is associated with B". Essentially, they jump from Phase 1 straight to Phase 4. This is problematic because often in the exploratory process we only look at rough sketches, often just bivariate plots or 2x2 tables which reveal strong relationships. Making inferences from this kind of exploratory work can be misleading without carefully considering second order factors like confounding or the effects of missing data. Often, those issues are more easily dealt with when using formal models or more detailed visualization.

Thinking through and formally specifying a problem in Phase 2 can often help to solve problems without having to fit a single model. For example, if exploratory analysis suggests strongly that missing data is missing completely at random, then there may be no need to model the missing data process in Phase 3. In the extreme case, you may conclude that the question of interest cannot be answered with the data available, in which case other data needs to be obtained or the analysis is over.

I think awareness of the different phases of an analysis can help to address this problem because it provides the analyst with a way to ask the question of "Where am I?" When using a map, in order to figure out where you're going, you first need to figure out where you are. The four-phase model serves as a roadmap to which analysts can point to identify their location.

Future Work

I will end this with some thoughts on what the four-phase model implies for potential future work in the field of data analysis. Seeing the model written down, some immediate gaps jump out at me that I think the data science community should consider addressing.

Tooling

Much of the software that I use for data analysis, when I reflect on it, is primarily designed for the divergent phases of analysis. Software is inherently designed to help you do things faster so that you can "analyze at the speed of thought". Good plotting and data wrangling software lets you do those activities faster. Good modeling software lets you execute and fit models faster. In both situations, fast iteration is important so that many options can be created for consideration.

Software for convergent phases of analysis are lacking by comparison. While there are quite a few good tools for visualization and report writing in Phase 4, I can't think of a single data analytic software tool designed for Phase 2 when specifying the problem. In particular, for both Phases 2 and 4, I don't see many tools out there for helping you to *choose* between all the options that you create in the divergent phases. I think data scientists may need to look outside their regularly scheduled programming to find better tools for those phases. If no good tools exist, then that might make for a good candidate for development.

Education

Data analytic education tends to focus on the model development and presentation phases of data analysis. Exploratory data analysis is an exception that largely falls into Phase 1, but EDA is not the only element of Phase 1. We often do not teach aspects of challenging a question or expanding beyond what is given to us. Typical data analytic education takes the original problem as the final specification and goes from there. There is no re-litigating the problem itself. In other words, we start at the end of Phase 2.

I think more could be done to teach students aspects of both convergent phases of analysis (Phases 2 and 4). For Phase 2, teaching students to hone and refine a question is critical for doing research or working elsewhere in data science. I

think providing them with greater exposure to this process and the chance to practice repeatedly would be a shift in the right direction. I also think we could do more to teach more narrative-development skills for Phase 4 of analysis, whereby results are synthesized into a coherent presentation.

Regardless of what we might do more or less of, I think the four phase model allows you to consider what aspect of data analysis you might want to teach at any given moment. It can be difficult to simulate a full-fledged data analysis from start to finish in a single classroom setting. However, splitting up the process and focusing on individual phases could be useful way to modularize the process into more manageable chunks.

Summary

Stealing the "double diamond" model from design thinking and adapting it for describing data analysis has some important benefits as a mental model for the process of analyzing data. It lays out four phases of analysis that all of unique activities (and potentially a unique toolbox) associated with them. I find it a useful model for explaining the various debates going on in the data science community, for exposing gaps in the tooling and in data science education, and for describing certain classes of mistakes that can be made during data analysis.

10. Abductive Reasoning

There are various forms of reasoning that play an important role in science and other data-related areas. Two of them, inductive and deductive reasoning, are well-discussed and commonly referred to. A third kind, abductive reasoning, is not nearly so commonly discussed, but I believe it plays a critical role in thinking about *data analysis*. I don't know why it so rarely comes up in the data analysis literature, but it serves as a useful way of thinking about the process.

Deductive and inductive reasoning are relatively straightforward to consider in the context of science. Deductive reasoning says there's a rule, say air pollution causes death, and that if we observe air pollution in the presence of people, then we can expect or predict that people will die at higher rates than if there weren't any air pollution. With inductive reasoning, we see lots of examples of air pollution going up and down and mortality rates going up and down, and conclude that there must be a rule along the lines of "air pollution causes death".

I think many people see examples of inductive reasoning and say "That sounds like data analysis", but it isn't. In particular, seeing a single instance of air pollution and mortality (i.e. a single data analysis), would not lead the responsible scientist to conclude that there must be some rule. Only after seeing many instances, i.e. many data analyses, and after considering other factors, might one conclude that such a rule exists. This is the value of replicating experiments and is a critical aspect of scientific synthesis. I would argue that inductive reasoning is what we engage in *after* conducting or observing many data analyses. But how do we characterize the process of doing a *single* data analysis?

With abductive reasoning, we know there's a rule (air pollution causes death) and we observe the outcome, say mortality.

We then reason that people might have been exposed to air pollution. Now, that conclusion isn't necessarily true, because there are many possible causes of death. However, concluding that people were exposed to air pollution is not necessarily unreasonable. Exactly how reasonable it is depends on the *context* of the problem. If you're living in Antarctica, then it's probably not the best explanation. But if you're living in Beijing, air pollution starts to look a bit more reasonable.

Let's take another example: smoking and lung cancer. Decades of research have shown that cigarette smoking causes lung cancer. If we observe someone with lung cancer, we might reason that they were at some point a smoker. In this case we likely would be correct! That's because although there is more than one cause of lung cancer, cigarette smoking has one of the strongest effects and is a fairly common activity. An example of abductive reasoning that is often given is *diagnosis*, and it's easy to see why. Given a set of outcomes, we are attempting to reason about the cause.

Designing a Data Analysis

Abductive reasoning is sometimes discussed in the context of design. There, we have requirements for a product or object, and we need to reason about what the "best fitting" solution will be. Here, we are less concerned with "cause and effect" and more concerned with "problem and solution" or perhaps "product and solution". The nature of that solution will depend critically on the context in which it is being designed, and furthermore, designers are less concerned with discovering or knowing what any underlying rule may be. As long as a solution can be found that satisfies the requirements, success can be claimed. That said, the process of discovering the actual requirements for a product is an essential part of the design process.

Data analysis follows a similar process. The most critical aspect of any data analysis is figuring out what the underlying problem is and making sure that it can be mapped to the data.

However, once the problem can be reasonable well-specified, it is up to the data analyst to design the solution.

For example, suppose the problem is to compare the means of two groups and generate evidence regarding whether the underlying means of the two groups are different. There are multiple ways to solve this problem, each with their strengths and weaknesses.

- A simple approach might be a *t*-test comparing the two groups. This approach makes some assumptions about the underlying probability distribution of the data, but also has some robustness properties. With this approach you can get a *p*-value which allows you to make probability statements about the underlying process (of course, after making assumptions about the sample space). A downside of this approach is that it is purely numerical and does not provide any visual information about the data.
- Another approach might be a side-by-side boxplot. This is a graphical approach that provides five summary statistics about the data in each group. While using the plot doesn't require any critical assumptions about the data, it's usefulness tends to depend on the data having a symmetric distribution. This plot allows one to easily check for things like outliers and one can make a crude eyeball calculation of the difference between the medians of the groups. There are no numerical outputs that we get from the plot and we cannot make any probability statements.
- Side-by-side dot plots are another approach that we can take, which simply presents all the data on a plot (rather than the five summary statistics given by the boxplot). If the data are not very numerous, this can be a very reasonable plot to make (perhaps with some jittering), and gives a lot more information about the distribution of the data across their range. With two groups, I've found the plot easy to look at. But if there were 50 groups, I might prefer to look at boxplots.

- One last approach might be to use a linear model to model the data using the group indicator as a predictor. This gives you a numerical result for the average difference between the two groups and a *p*-value for the statistical significance. It also gives you the ability to generate predictions if desired. Furthermore, if additional variables need to be accounted for, they can be included within the same framework, unlike with the *t*-test approach.

There are many other approaches or variations on these approaches that could be taken and that, of course, is one big reason why the data analyst's job is so hard! Which approach *should* be taken?

At this point, I've discussed possible solutions but I haven't discussed what the requirements of the problem might be. Some example scenarios might be

- The output for this analysis will be fed into a different algorithm that make predictions about some other phenomenon. The method you choose must be fast, automatic, and easily repeatable, potentially for large datasets.
- You will be making a presentation to an executive who will ultimately make a decision based on the evidence. This executive is interested in seeing how the data inform the output and wants to be able to see what influence individual data points might have on the conclusion.
- The output from this analysis will be presented on a web-based dashboard displayed to a colleague and there will be no opportunity to discuss or explain anything with that colleague about the results.

I think depending on which scenario you are in, you might choose to use one or another method or perhaps some combination of methods. I'm not sure there's one approach that would be appropriate for all scenarios, nor do I think any one approach is dramatically better or worse than the others.

Ultimately, I think a good data analyst assesses the situation they are in and assembles an appropriate solution. While some time might be spent considering what might be the "optimal" situation, such effort is mostly wasted in this case.

Summary

Data analysis is largely about fashioning reasonable solutions to problems, given their specific requirements. To that end, the process can perhaps be best described as abductive reasoning. While in some cases we can identify "optimal" approaches that should be used for specific cases (i.e. there is a *rule* that can be applied), most real-world problems and scenarios are not specific enough to neatly identify such optimal procedures. Rather, it is up to the data analyst to find a "best fitting" solution that meets the requirements of the problem and provides the necessary information.

11. Tukey, Design Thinking, and Better Questions

Roughly once a year, I read John Tukey's paper "The Future of Data Analysis"[1], originally published in 1962 in the *Annals of Mathematical Statistics*. I've been doing this for the past 17 years, each time hoping to really understand what it was he was talking about. Thankfully, each time I read it I seem to get *something* new out of it. For example, in 2017 I wrote a whole talk[2] around some of the basic ideas.

Probably the most famous line from this paper is

> Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise.

The underlying idea in this sentence arises in at least two ways in Tukey's paper. First is his warning that statisticians should not be called upon to produce the "right" answers. He argues that the idea that statistics is a "monolithic, authoritarian structure designed to produce the 'official' results" presents a "real danger to data analysis". Second, Tukey criticizes the idea that much of statistical practice centers around optimizing statistical methods around precise (and inadequate) criteria. One can feel free to identify a method that minimizes mean squared error, but that should not be viewed as the *goal* of data analysis.

But that got me thinking—what *is* the ultimate goal of data analysis? In 64 pages of writing, I've found it difficult to

[1] https://projecteuclid.org/euclid.aoms/1177704711
[2] https://youtu.be/qFtJaq4TlqE

identify a sentence or two where Tukey describes the ultimate goal, why it is we're bothering to analyze all this data. It occurred to me in this year's reading of the paper, that maybe the reason Tukey's writing about data analysis is often so confusing to me is because his goal is actually quite different from that of the rest of us.

More Questions, Better Questions

Most of the time in data analysis, we are trying to answer a question with data. I don't think it's controversial to say that, but maybe that's the wrong approach? Or maybe, that's what we're *not* trying to do at first. Maybe what we spend most of our time doing is figuring out a better question.

Hilary Parker and I have discussed at length the idea of design thinking on our podcast, Not So Standard Deviations[3]. One of the fundamental ideas from design thinking involves identifying the problem. It's the first "diamond" in the "double diamond" approach to design.

Tukey describes the first four steps in a data analysis as:

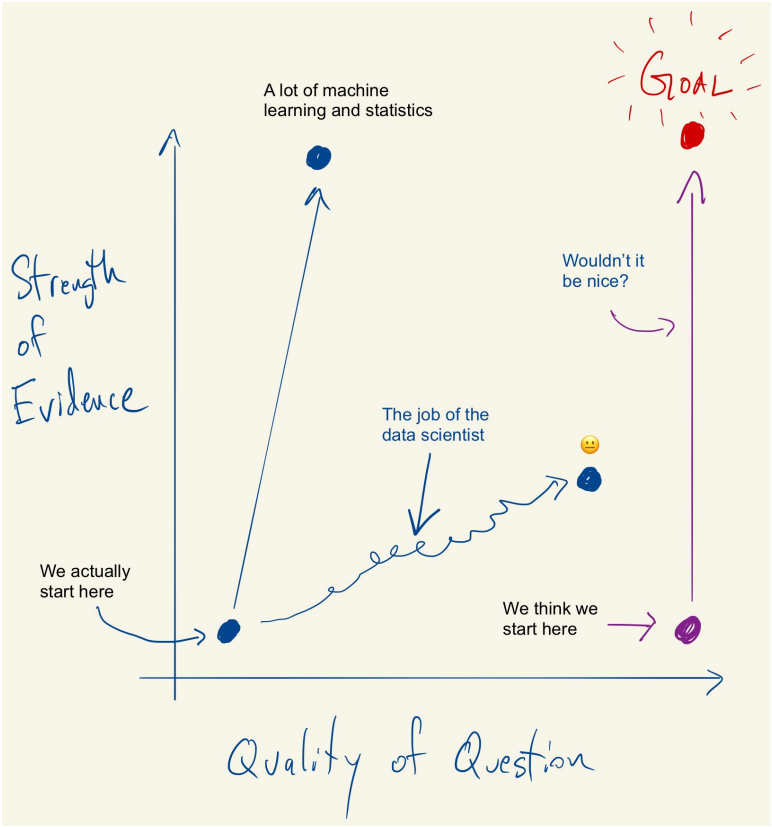
1. Recognition of problem
2. One technique used
3. Competing techniques used
4. Rough comparisons of efficacy

In other words, try one approach, then try a bunch of other approaches! You might be thinking, why not just try *the best* approach (or perhaps the *right* approach) and save yourself all that work of trying different approaches? Well, that's the kind of path you go down when you're trying to answer the question. Stop doing that! There are two reasons why you should stop thinking about answering the question:

[3] http://nssdeviations.com

1. You're probably asking the wrong question anyway, so don't take yourself too seriously;
2. The "best" approach is only defined as "best" according to some arbitrary criterion that probably isn't suitable for your problem/question.

After thinking about all this I was inspired to draw the following diagram.



Strength of Evidence vs. Quality of Question

The goal in this picture is to get to the upper right corner, where you have a high quality question and very strong evidence. In my experience, most people assume that they are

starting in the bottom right corner, where the quality of the question is at its highest. In that case, the only thing left to do is to choose the optimal procedure so that you can squeeze as much information out of your data. The reality is that we almost always start in the bottom left corner, with a vague and poorly defined question and a similarly vague sense of what procedure to use. In that case, what's a data scientist to do?

In my view, the most useful thing a data scientist can do is to devote serious effort towards improving the quality and sharpness of the question being asked. On the diagram, the goal is to move us as much as possible to the right hand side. Along the way, we will look at data, we will consider things outside the data like context, resources and subject matter expertise, and we will try a bunch of different procedures (some optimal, some less so).

Ultimately, we will develop some of idea of what the data tell us, but more importantly we will have a better sense of what kinds of questions we can ask of the data and what kinds of questions we actually want to have answered. In other words, we can learn more about ourselves by looking at the data.

Exploring the Data

It would seem that the message here is that the goal of data analysis is to explore the data. In other words, data analysis *is* exploratory data analysis. Maybe this shouldn't be so surprising given that Tukey wrote the book[4] on exploratory data analysis. In this paper, at least, he essentially dismisses other goals as overly optimistic or not really meaningful.

For the most part I agree with that sentiment, in the sense that looking for "the answer" in a single set of data is going to result in disappointment. At best, you will accumulate evidence that will point you in a new and promising direction. Then you can iterate, perhaps by collecting new data, or by

[4] https://en.wikipedia.org/wiki/Exploratory_data_analysis

asking different questions. At worst, you will conclude that you've "figured it out" and then be shocked when someone else, looking at another dataset, concludes something completely different. In light of this, discussions about p-values and statistical significance are very much beside the point.

The following is from the very opening of Tukey's book *Exploratory Data Analysis*:

> It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it

(Note that the all caps are originally his!) Given this, it's not too surprising that Tukey seems to equate exploratory data analysis with essentially all of data analysis.

Better Questions

There's one story that, for me, totally captures the spirit of exploratory data analysis. Legend has it that Tukey once asked a student what were the benefits of the median polish technique[5], a technique he invented to analyze two-way tabular data. The student dutifully answered that the benefit of the technique is that it provided summaries of the rows and columns via the row- and column-medians. In other words, like any good statistical technique, it *summarized the data* by reducing it in some way. Tukey fired back, saying that this was incorrect—the benefit was that the technique created *more data*. That "more data" was the residuals that are leftover in the table itself after running the median polish. It is the residuals that really let you learn about the data, discover whether there is anything unusual, whether your question is well-formulated, and how you might move on to the next step. So in the end, you got row medians, column medians, *and* residuals, i.e. more data.

[5] https://en.wikipedia.org/wiki/Median_polish

If a good exploratory technique gives you more data, then maybe good exploratory data analysis gives you more questions, or *better* questions. More refined, more focused, and with a sharper point. The benefit of developing a sharper question is that it has a greater potential to provide discriminating information. With a vague question, the best you can hope for is a vague answer that may not lead to any useful decisions. Exploratory data analysis (or maybe just *data analysis*) gives you the tools that let the data guide you towards a better question.

12. The Role of Creativity

I've often heard that there is a need for data analysts to be *creative* in their work. But why? Where and how exactly is that creativity exercised?

On one extreme, it could be thought that a data analyst should be easily replaced by a machine. For various types of data and for various types of questions, there should be a deterministic approach to analysis that does not change. Presumably, this could be coded up into a computer program and the data could be fed into the program every time, with a result presented at the end. For starters, this would eliminate the notorious researcher degrees of freedom[1] problem. If there were substantial institutional knowledge[2] of data analysis, this might indeed be possible. How is it that every data analysis is so different that a human being is needed to craft a solution?

Well, it's *not* true that every analysis is different. Many power calculations, for example, are identical or very similar, and can be automated to some degree. However, exactly how those power calculations are used or interpreted can vary quite a bit from project to project. Even the very same calculation for the same study design can be interpreted differently in different projects, depending on the context. The same is true for other kinds of analyses like regression modeling or machine learning.

Creativity is needed because of the *constraints* placed on

[1] http://journals.sagepub.com/doi/abs/10.1177/0956797611417632

[2] https://simplystatistics.org/2018/06/15/people-vs-institutions-in-data-analysis/

the analysis by context[3], resources[4], and the audience[5], all things that we might think of as being "outside" the data. The context around which the data are created, the resources (time, money, technology) available to do the analysis, and the audience to which the results will be presented, all play a key role in determining the strategy that an analyst develops to analyze the data. The analyst will often have to employ some amount of creativity in order to execute a strategy that produces useful output.

The Role of Context

The context of a problem has great influence over how we frame a question, how we translate it into a data problem, and how we go about collecting data. The context also allows us to answer questions regarding *why* the data appear to be the way they do. The same number for the same type of measurement can have different interpretations based on the context.

Missing Data

Missing data are present in almost every dataset and the most important question a data analyst can ask when confronted with missing data is "Why are the data missing?" It's important to develop some understanding of the mechanism behind what makes the data missing in order to develop an appropriate strategy for dealing with missing data (i.e. doing nothing, imputation, etc.) But the data themselves often provide little information about this mechanism; often the mechanism is coded outside the data, possibly not even written down but

[3] https://simplystatistics.org/2018/05/24/context-compatibility-in-data-analysis/

[4] https://simplystatistics.org/2018/06/18/the-role-of-resources-in-data-analysis/

[5] https://simplystatistics.org/2018/04/17/what-is-a-successful-data-analysis/

stored in the minds of people who originally collected the data.

Take a two-arm clinical trial with an experimental treatment and a placebo. Sometimes with experimental treatments, there are side effects and people will drop out of the trial (or even die) because they cannot handle the side effects. The result is more missing data in the experimental arm of the trial than in the placebo arm. Now the data themselves will reveal a differential in the rate of missing data between the arms as it will be clear that the treatment arm has a higher rate. But the data will not reveal the exact reason why they dropped out. Depending on the nature of the trial and the question being asked, there might be a few different ways to handle this problem. Imputation may be feasible or perhaps some sort of matching scheme. The exact choice of how to proceed will depend on what external data are available, how much data are missing, and how the results will be used, among many other factors.

Another example might be in the analysis of outdoor particulate matter air pollution data. Monitors run by the US EPA typically take measurements once every six days. The reason is that it is expensive to process the filters for PM data and so the 1-in-6 day schedule is a designed compromise to balance cost with quantity of data. Of course, this means that in the data records 5 out of every 6 days is "missing", even though the missingness was introduced deliberately. Again, the data don't say why they are missing. One could easily imagine a scenario where the monitor doesn't record data when the values of PM are very high or very low, a kind of informative missingness. But in this case, the missing data can be ignored and typically doesn't have a large impact on subsequent modeling. In fact, imputation can be detrimental here because it does not provide much benefit but can greatly increase uncertainty.

In both cases, the data analyst's job is to assess the situation, look at the data, obtain information about the context and why the data are missing (from a subject matter expert), and then decide on an appropriate path forward. Even with these

two scenarios, there is no generic path forward.

The Role of the Audience

The audience is another key factor that primarily influences *how* we analyze the data and present the results. One useful approach is to think about what final products need to be produced and then work backwards from there to produce the result. For example, if the "audience" is another algorithm or procedure, then the exact nature of the output may not be important as along as it can be appropriately fed into the next part of the pipeline. A priority will be placed on making sure the output is machine readable. In addition, interpretability may not weigh that heavily because no human being will be looking at the output of this part. However, if a person *will* be looking at the results, then you may want to focus on a modeling approach that lets that person reason about the data[6] and understand how the data inform the results.

In one extreme case, if the audience is another data analyst, you may want to do a relatively "light" analysis (maybe just some preprocessing) but then prepare the data in such a way that it can be easily distributed to others to do their own analysis. This could be in the form of an R package or a CSV file or something else. Other analysts may not care about your fancy visualizations or models; they'd rather have the data for themselves and make their own results.

A data analyst must make a reasonable assessment of the audience's needs, background, and preferences for receiving data analytic results. This may require some creative guessing. If the audience is available to the analyst, the analyst should ask questions about how best to present results. Otherwise, reasonable assumptions must be made or contingencies (e.g. backup slides, appendices) can be prepared for the presentation itself.

[6] https://simplystatistics.org/2017/11/16/reasoning-about-data/

Resources and Tools

The data analyst will likely have to work under a set of resource constraints[7], placing boundaries on what can be done with the data. The first and foremost constraint is likely to be time. One can only try so many things in the time allotted, or some analyses may take too long to complete. Therefore, compromises may need to be made unless more time and resources can be negotiated. Tooling will be limited in that certain combinations of models and software may not exist and there may not be time to develop new tools from scratch.

A good data analyst must make an estimate of the time available and determine whether it is sufficient to complete the analysis. If resources are insufficient, then the analyst must either negotiate for more resources or adapt the analysis to fit the available resources. Creativity will almost certainly be required when there are severe resource constraints, in order to squeeze as much productivity out of what is available.

Summary

Context, audience, and resources can all place different kinds of constraints on a data analysis, forcing the analyst to employ different kinds of creativity to get the job done. Although I've presented each context, audience, and resources separately here, in most analyses all of these factors will play a role simultaneously. The complexity of the constraint environment (and their various interactions) can grow quickly, placing increasing pressure on the analyst to think creatively to produce useful results.

[7] https://simplystatistics.org/2018/06/18/the-role-of-resources-in-data-analysis/

13. Should the Data Come First?

One conversation I've had quite a few times revolves around the question, "What's the difference between science and data science?" If I were to come up with a simple distinction, I might say that

> Science starts with a question; data science starts with the data.

What makes data science so difficult is that it starts in the *wrong place*. As a result, a certain amount of extra work must be done to understand the context surrounding a dataset before we can do anything useful.

Procedure for procedurals

One of my favorite shows growing up was *Law & Order*, one of the longest running "procedural" shows that also produced a number of spinoffs. I remember coming home from school in the afternoon, flipping on the TV and watching whatever *Law & Order* marathon was currently being aired (there was always some marathon happening).

One distinct feature of the show was that pretty much every episode followed the exact same format:

1. The episode starts with the discovery that a crime has occurred, usually a murder.
2. The first half of the episode involves the police retracing history trying to figure out who committed the crime (the "Order" part).

3. The police catch a suspect at exactly 22 minutes into the episode.
4. In the second half of the episode, the district attorneys take over and prosecute the suspect in court (the "Law" part). Usually, they win.

The format of the show is in fact highly unusual. While it starts off with a mystery (who committed this murder?) as many shows do, the mystery is solved half way through the episode! I can imagine some TV executive 30 years ago wondering, "What the heck are you going to do for the rest of the episode if you solve the mystery half way through?"

What made the show interesting to me was that in the second half of every episode, the lawyers for the government had to convince a jury that they had the right suspect. They had to present evidence and make an argument that this person was guilty. Of course, because this is TV, they usually won the argument, but I think many TV shows would have been satisfied with just catching the criminal. In most shows, presenting the data and the evidence to an audience is not interesting, but on *Law & Order*, it was.

The Law & Order Metaphor for Data Science

Many data science projects essentially follow this format, because we start with the data. Data are available. We often don't get to participate in its generation. Perhaps the data were collected as part of some administrative records system, or as part of some logging system, or as part of some other project addressing a different question. The initial challenge of any data science project is figuring out the *context* around the data and *question* that motivated its origination. A key milestone is then figuring out how exactly the data came to be (i.e. who committed this "crime"?).

Once we've figured out the context around the data, essentially retracing the history of the data, we can then ask "Are these data appropriate for the question that **I want to ask**?" Answering this question involves comparing the context surrounding the original data and then ensuring that it is compatible with the context for your question. If there is compatibility, or we can *create* compatibility via statistical modeling or assumptions, then we can intelligently move forward to analyze the data and generate evidence concerning a different question. We will then have to make a separate argument to some audience regarding the evidence in the data and any conclusions we may make. Even though the data may have been convincing for one question, it doesn't mean that the data will be convincing for a different question.

If we were to develop a procedure for a data science "procedural" TV show, it might like the following.

1. Datasets are cobbled together from various sources
2. The first task is to retrace the history of each dataset and determine the context around which it was created and what question, if any, motivated its generation.
3. We then determine whether the data are appropriate for answering our question or can be transformed to be appropriate for our question. (If not, we have to stop here.)
4. We analyze the data.

Data science often starts with the data, but in an ideal world it wouldn't. In an ideal world, we would ask questions and carefully design experiments to collect data specific to those questions. But this is simply not practical and data need to be shared across contexts. The difficulty of the data scientist's job is understanding each dataset's context, making sure it is compatible with the *current* question, developing the evidence from the data, and then getting an audience to accept the results.

14. Partitioning the Variation in Data

One of the fundamental questions that we can ask in any data analysis is, "Why do things vary?" Although I think this is fundamental, I've found that it's not explicitly asked as often as I might think. The problem with *not* asking this question, I've found, is that it can often lead to a lot of pointless and time-consuming work. Taking a moment to ask yourself, "What do *I* know that can explain why this feature or variable varies?" can often make you realize that you actually know more than you think you do.

When embarking on a data analysis, ideally *before* you look at the data, it's useful to *partition the variation* in the data. This can be roughly broken down into to categories of variation: fixed and random. Within each of those categories, there can be a number of sub-categories of things to investigate.

Fixed variation

Fixed variation in the data is attributable to fixed characteristics in the world. If we were to sample the data again, the variation in the data attributable to fixed characteristics would be *exactly the same*. A classic example of a fixed characteristic is seasonality in time series data. If you were to look at a multi-year time series of mortality in the United States, you would see that mortality tends to be higher in the winter season and lower in the summer season. In a time series of daily ozone air pollution values, you would see that ozone is highest in the summer and lowest in the winter. For each of these examples, the seasonality is consistent pretty much every year. For ozone, the explanation has to do with the nature of

atmospheric chemistry; for mortality the explanation is less clear and more complicated (and likely due to a combination of factors).

Data having fixed variation doesn't imply that it always has the same values every time you sample the data, but rather the general patterns in the data remain the same. If the data are different in each sample, that is likely due to *random variation*, which we discuss in the next section.

Understanding which aspects of the variation in your data are fixed is important because often you can collect data on those fixed characteristics and use them directly in any statistical modeling you might do. For example, season is an easy covariate to include because we already know when the seasons begin and end. Using a covariate representing month or quarter will usually do the trick.

Explaining the variation in your data by introducing fixed characteristics in a model can reduce uncertainty and improve efficiency or precision. This may require more work though, in the form of going out and collecting more data or retrieving more variables. But doing this work will ultimately be worth it. Attempting to model variation in the data that is inherently fixed is a waste of time and will likely cost you degrees of freedom in the model.

In my experience looking at biomedical data, I have found that a lot of variation in the data can be explained by a few fixed characteristics: age, sex, location, season, temperature, etc. In fact, often so much can be explained that there is little need for explicit models of the random variation. One difficult aspect of this approach though is that it requires a keen understanding of the context surrounding the data as well as having a good relationship with a subject matter expert who can help inform you about the sources of variation. Investing in learning more about the data, before digging into the data itself, can save you a lot of time in the modeling phase of data analysis.

Random variation

Once we've partitioned out all of the variation in the data that can be attributed to fixed characteristics, what's left is random variation. It is sometimes tempting to look at data and claim that all of the variation is random because then we can model it without collecting data on any more covariates! But as I mentioned above, it's useful to at least hypothesize what might be driving that variation and collect the extra data that's needed.

Random variation causes data to look different every time we sample it. While we might be quite sure that ozone is going to be high in the summer (versus the winter), that doesn't mean that it will always be 90 parts per billion on June 30. It might be 85 ppb one year and 96 ppb another year. Those differences are not easily explainable by fixed phenomena and so it might be reasonable to characterize them as random differences. The key thing to remember about random variation in the data is

> Random variation must be independent of the variation attributable to fixed characteristics

It's sometimes said that random variation is just "whatever is leftover" that we could not capture with fixed features. However, this is an uncritical way of looking at the data because if there are fixed characteristics that get thrown in the random variation bin, then you could be subjecting your data analysis to hidden bias or confounding. There are some ways to check for this in the modeling stage of data analysis, but it's better do what you can to figure things out beforehand in the discovery and exploration phases.

One application where random variation is commonly modeled is with financial market data, and for good reason. The efficient-market hypothesis[1] states that, essentially, if there

[1] https://en.wikipedia.org/wiki/Efficient-market_hypothesis

were any fixed (predictable) variation in the prices of financial assets, then market participants would immediately seize upon that information to profit through arbitrage opportunities. If you knew for example that Apple's stock price was always low in the winter and high in the summer, you could just buy in the winter and sell in the summer and make money without risk. But if *everyone* did that, then eventually that arbitrage opportunity would go away (as well as the fixed seasonal effect). Any variation in the stock price that is leftover is simply random variation, which is why it can be so difficult to "beat the market".

Is it really random?

When I see students present data analyses, and they use a standard linear model that has an outcome (usually labeled Y), a predictor (X), and random variation or error (e), my first question is always about the error component. Usually, there is a little confusion about why I would ask about that since that part is just "random" and uninteresting. But when I try to lead them into a discussion of why there is random variation in the data, often we discover some additional variables that we'd like to have but don't have data on.

Usually, there is a very good explanation of why we don't have those data. My point is not to criticize the student for not having data that they couldn't get, but to emphasize that those features are potential confounders and are not random. If data cannot be collected on those features, it might be worth investigating whether a reasonable surrogate can be found.

Summary

Partitioning your data into fixed and random components of variation can be a useful exercise even before you look at the data. It may lead you to discover that there are important

features for which you do not have data but that you can go out and collect. Making the effort to collect additional data when it is warranted can save a lot of time and effort trying to model variation as if it were random. More importantly, omitting important fixed effects in a statistical model can lead to hidden bias or confounding. When data on omitted variables cannot be collected, trying to find a surrogate for those variables can be a reasonable alternative.

15. How Data Analysts Think - A Case Study

In episode 71 of Not So Standard Deviations[1], Hilary Parker and I inaugurated our first "Data Science Design Challenge" segment where we discussed how we would solve a given problem using data science. The idea with calling it a "design challenge" was to contrast it with common "hackathon" type models where you are presented with an already-collected dataset and then challenged to find something interesting in the data. Here, we wanted to start with a problem and then talk about how data might be collected and analyzed to address the problem. While both approaches might result in the same end-product, they address the various problems you encounter in a data analysis in a different order.

In this post, I want to break down our discussion of the challenge and highlight some of the issues that were discussed in framing the problem and in designing the data collection and analysis. I'll end with some thoughts about generalizing this approach to other problems.

You can download an MP3 of this segment of the episode[2] (it is about 45 minutes long) or you can read the transcript of the segment[3]. If you'd prefer to stream the segment you can start listening here[4].

[1] http://nssdeviations.com/

[2] https://www.dropbox.com/s/yajgbr25dbh20i0/NSSD%20Episode%2071%20Design%20Challenge.mp3?dl=0

[3] https://drive.google.com/open?id=11dEhj-eoh8w13dS-mWvDMv7NKWXZcXMr

[4] https://overcast.fm/+FMBuKdMEI/00:30

The Brief

The general goal was to learn more about the time it takes for each of us to commute to work. Hilary lives in San Francisco and I live in Baltimore, so the characteristics of our commutes are very different. She walks and takes public transit; I drive most days. We also wanted to discuss how we might collect data on our commute times in a systematic, but not intrusive, manner. When we originally discussed having this segment, this vague description was about the level of specification that we started with, so an initial major task was to

1. Develop a better understanding of what question each of us was trying to answer;
2. Frame the problem in a manner that could be translated into a data collection task; and
3. Sketch out a feasible statistical analysis.

Framing the Problem

Hilary and I go through a few rounds of discussion on the topic of how to think about this problem and the questions that we're trying to answer. Early in the discussion Hilary mentions that this problem was "pressing on my mind" and that she took a particular interest in seeing the data and acting on it. Her intense interest in the problem potentially drove part of her creativity in developing solutions.

Hilary initially mentions that the goal is to understand the variation in commute times (i.e. estimate the variance), but then quickly shifts to the problem of estimating average commute times for the two different commute methods that she uses.

> HILARY: you…maybe only have one commute method and you want to understand the variance of that. So…what range of times does it usually take for me

> to get to work, or…I have two alternatives for my commute methods. So it might be like how long does it take me in this way versus that way? And for me the motivation is that I want to make sure I know when to leave so that I make it to meetings on time….

In her mind, the question being answered by this data collection is, "When should I leave home to get to meetings on time?" At this point she mentions two possible ways to think about addressing this question.

1. Estimate the variability of commute times and leave the house accordingly; or
2. Compare two different commute methods and then *choose* a method on any given day.

Right off the bat, Hilary notes that she doesn't actually do this commute as often as she'd thought. Between working from home, taking care of chores in the morning, making stops on the way, and walking/talking with friends, a lot of variation can be introduced in to the data.

I mention that "going to work" and "going home", while both can be thought of as commutes, are not the same thing and that we might be interested in one more than the other. Hilary agrees that they are different problems but they are both potentially of interest.

Question/Intervention Duality

At this point I mention that my commuting is also affected by various other factors and that on different days of the week, I have a different commute pattern. On days where I drop my son off at school, I have less control over when I leave compared to days when I drive straight to work. Here, we realize a fundamental issue, which is that different days of the week indicate somewhat different interventions to take:

- On days where I drive straight to work, the question is "When should I leave to arrive on time for the first meeting?"
- On days where I drop my son off at school, the question is "When is the earliest time that I can schedule the first meeting of the day?"

In the former situation I have control over, and could potentially intervene on, when I leave the house, whereas in the latter situation I have control over when I schedule the first meeting. While these are distinct questions with different implications, at this point they may both require collecting travel time data in the same manner.

Earlier in this section I mention that on days when I drop my son off at school it can take 45 minutes to get to work. Hilary challenges this observation and mentions that "Baltimore is not that big". She makes use of her knowledge of Baltimore geography to suggest that this is unexpected. However, I mention that the need to use local roads exclusively for this particular commute route makes it indeed take longer than one might expect.

Designing the Data Collection System

In discussing the design of her data collection system, Hilary first mentions that a podcast listener had emailed in and mentioned his use of Google Maps to predict travel times based on phone location data. While this seemed like a reasonable idea, it ultimately was not the direction she took.

> HILARY: At first I was thinking about that because I have location history on and I look at it a lot, but there's also a fair degree of uncertainty there. Like sometimes it just puts me in these really weird spots or…I lose GPS signal when I go underground and also I do not know how to get data in an API sense from that. So I knew it would be manual data.

> In order to analyze the data, I would have to go back and be like let me go and collect the data from this measurement device. So I was trying to figure out what I could use instead.

Then she describes how she can use Wi-Fi connections (and dis-connections) to serve as surrogates for leaving and arriving.

> And at some point I realized that two things that reliably happen every time I do a commute is that my phone disconnects from my home Wi-Fi. And then it reconnects to my work Wi-Fi. And so I spent some time trying to figure out if I could log that information, like if there's an app that logged that, and there is not. But, there is a program called If This, Then That, or an app. And so with that you can say "when my phone disconnects from Wi-Fi do something", and you can set it to a specific Wi-Fi. So that was exciting.

Other problems that needed solving were:

- *Where to store the data*. Hilary mentions that a colleague was using Airtable[5] (a kind of cloud-based spreadsheet/-database) and decided to give it a try.
- *Indicating commute method*. Hilary created a system where she could send a text message containing a keyword about her commute method to a service that would then log the information to the table collecting the travel time data.
- *Multiple Wi-Fi connects*. Because her phone was constantly connecting and disconnecting from Wi-Fi at work, she had to define the "first connection to Wi-Fi at work" as meaning that she had arrived at work.

[5] https://airtable.com

- *Sensing a Wi-Fi disconnect*. Hilary's phone had to be "awake" in order to sense a Wi-Fi disconnect, which was generally the case, but not always. There was no way to force her phone to always be awake, but she knew that the system would send her a push notification when it had been triggered. Therefore, she would at least know that if she didn't receive a push notification, then something had gone wrong.

Hilary mentions that much of the up front effort is important in order to avoid messy data manipulations later on.

> HILARY: I think I'll end up—I did not do the analysis yet but I'll end up having to scrub the data. So I was trying to avoid manual data scrubbing, but I think I'm going to have to do it anyway.

Ultimately, it is impossible to avoid all data manipulation problems.

Specifying the Data

What exactly are the data that we will be collecting? What are the covariates that we need to help us understand and model the commute times? Obvious candidates are

- The start time for the commute (date/time format, but see below)
- The end time for the commute (date/time)
- Indicator of whether we are going to work or going home (categorical)
- Commute method (categorical)

Hilary notes that from the start/end times we can get things like day of the week and time of day (e.g. via the lubridate[6] package). She also notes that her system doesn't exactly

[6] https://cran.r-project.org/web/packages/lubridate/index.html

produce date/time data, but rather a text sentence that includes the date/time embedded within. Thankfully, that can be systematically dealt with using simple string processing functions.

A question arises about whether a separate variable should be created to capture "special circumstances" while commuting. In the data analysis, we may want to exclude days where we know something special happened to make the commute much longer than we might have expected (e.g. we happened to see a friend along the way or we decided to stop at Walgreens). The question here is

> Are these special circumstances part of the natural variation in the commute time that we want to capture, or are they "one-time events" that are in some sense predictable?

A more statistical way of asking the question might be, do these special circumstances represent *fixed or random variation*[7]? If they are random and essentially uncontrollable events, then we would want to include that in the random portion of any model. However, if they are predictable (and perhaps controllable) events, then we might want to think of them as another covariate.

While Hilary believes that she ultimately *does* have control over whether these time-consuming detours occur or not, she decides to model them as essentially random variation and that these events should be lumped in with the natural variation in the data.

Specifying the Treatment

At this point in the discussion there is a question regarding what *effect* we are trying to learn about. The issue is that

[7] https://simplystatistics.org/2018/07/23/partitioning-the-variation-in-data/

sometimes changes to a commute have to be made on the fly to respond to unexpected events. For example, if the public transportation system breaks down, you might have to go on foot.

> ROGER: Well it becomes like a compliance situation, right? Like you can say, do you want to know how long does it take when you take MUNI or how long does it take when you intend to take MUNI?

In this section I mention that it's like a "compliance problem". In clinical trials, for example when testing a new drug versus a placebo, it is possible to have a situation where people in the treatment group of the study are given the new drug but do not actually take it. Maybe the drug has side effects or is inconvenient to take. Whatever the reason, they are not complying with the protocol of the study, which states that everyone in the treatment group takes the new drug. The question then is whether you want to use the data from the clinical trial to understand the actual effect of the drug or if you want to understand the effect of telling someone to take the drug. The latter effect is often referred to as the *intention to treat* effect while the former is sometimes called the *complier average effect*. Both are valid effects to estimate and have different implications in terms of next steps.

In the context of Hilary's problem, we want to estimate the average commute time for each commute method. However, what happens if Muni experiences some failure that requires altering the commute method? The potential "compliance issue" here is whether Muni works properly or not. If it does not, then Hilary may take some alternate route to work, even though she *intended* to take Muni. Whether Muni works properly or not is a kind of "post-treatment variable" because it's not under the direct control of Hilary and its outcome is only known *after* she decides on which commute method she is going to take (i.e. the "treatment"). Now a choice must be made: Do we estimate the average commute time when taking Muni or the average commute time when she *intends* to take Muni, even if she has to divert to an alternate route?

Hilary and I both seem to agree that the intention to treat effect is the one we want to estimate in the commute time problem. The reason is that the estimation of this effect has direct implications for the thing that we have control over: choosing which commute method to use. While it might be interesting from a scientific perspective to know the average commute time when taking Muni, regardless of whether we intended to take it or now, we have no control over the operation of Muni on any given day.

Starting from the End

I ask Hilary, suppose we have the data, what might we do with it? Specifically, suppose that we estimate for a given commute method that the average time is 20 minutes and the standard deviation is 5 minutes. What "intervention" would that lead us to take? What might we do differently from before when we had no systematically collected data?

Hilary answers by saying that we can designate a time to leave work based on the mean and standard deviation. For example, if we have to be at work at 9am we might leave at 8:35am (mean + 1 standard deviation) to ensure we'll be arrive at 9am most of the time. In her answer, Hilary raises an important, but perhaps uncomfortable, consideration:

> HILARY: I think in a completely crass world for example, I would choose different cutoffs based on the importance of a meeting. And I think people do this. So if you have a super important meeting, this is like a career-making meeting, you leave like an hour early.... And so, there you're like "I am going to do three standard deviations above the mean" so...it's very unlikely that I'll show up outside of the time I predicted. But then if it's like a touch-base with someone where you have a really strong relationship and they know that you value their

> time, then maybe you only do like one standard deviation.

Later I mention one implication for statistical modeling:

> Roger: Well and I feel like...the discussion of the distribution is interesting because it might come down to like, what do you think the tail of the distribution looks like? So what's the worst case? Because if you want to minimize the worst case scenario, then you really, really need to know like what that tail looks like.

Thinking about what the data will ultimately be used for raises two important statistical considerations:

- We should think about the extremes/tails of the distribution and develop cutoffs that determine what time we should leave for work.
- The cutoffs at the tail of the distribution might be dependent on the "importance" of the first meeting of the day, suggesting the existence of a cost function that quantifies the importance of arriving on time.

Hilary raises a hard truth, which is that not everyone gets the same consideration when it comes to showing up on time. For an important meeting, we might allow for "three standard deviations" more than the mean travel time to ensure some margin of safety for arriving on time. However, for a more routine meeting, we might just provide for one standard deviation of travel time and let natural variation take its course for better or for worse.

Statistical Modeling Considerations

I mention that thinking about our imaginary data in terms of "mean" and "standard deviation" implies that the data have a

distribution that is akin to a Normal distribution. However, given that the data will consist of travel times, which are always positive, a Normal distribution (which allows positive and negative numbers) may not be the most appropriate. Alternatives are the Gamma or the log-Normal distribution which are strictly positive. I mention that the log-Normal distribution allows for some fairly extreme events, to which Hilary responds that such behavior may in fact be appropriate for these data due to the near-catastrophic nature of Muni failures (San Francisco residents can feel free to chime in here).

From the previous discussion on what we might do with this data, it's clear that the right tail of the distribution will be important in this analysis. We want to know what the "worst case scenario" might be in terms of total commute time. However, by its very nature, extreme data are rare, and so there will be very few data points that can be used to inform the shape of the distribution in this area (as opposed to the middle of the distribution where we will have many observations). Therefore, it's likely that our choice of model (Gamma, log-Normal, etc.) will have a big influence on the predictions that we make about commute times in the future.

Study Design Considerations

Towards the end I ask Hilary how much data is needed for this project? However, before asking I should have discussed the nature of the study itself:

- Is it a fixed study designed to answer a specific question (i.e. what is the mean commute time?) within some bound of uncertainty? Or
- Is it an ongoing study where data will be continuously collected and actions will be continuously adapted as new data are collected

Hilary suggests that it is the latter and that she will simply collect data and make decisions as she goes. However, it's

clear that the time frame is not "forever" because the method of data collection is not zero cost. Therefore, at some point the costs of collecting data will likely be too great in light of any perceived benefit.

Discussion

What have we learned from all of this? Most likely, the problem of estimating commute times is *not* relevant to everybody. But I think there are aspects of the process described above that illustrate how the data analytic process works before data collection begins (yes, data analysis includes parts where there are no data). These aspects can be lifted from this particular example and generalized to other data analyses. In this section I will discuss some of these aspects and describe why they may be relevant to other analyses.

Personal Interest and Knowledge

Hilary makes clear that she is very personally interested in this problem and in developing a solution. She wants to apply any knowledge learned from the data to her everyday life. In addition, she used her knowledge of Baltimore geography (from having lived there previously) to challenge my "mental data analysis".

Taking a strong personal interest in a problem is not always an option, but it can be very useful. Part of the reason is that it can allow you to see the "whole problem" and all of its facets without much additional effort. An uninterested person can certainly learn all the facets of a problem, but it will seem more laborious. If you are genuinely interested in the subject of a problem, then you will be highly motivated to learn everything about that problem, which will likely benefit you in the data analysis. To the extent that data analysis is a *systems problem* with many interacting parts, it helps to learn as much as possible about the system. Being *interested*

in knowing how the system works is a key advantage you can bring to any problem.

In my own teaching, I have found that students who are keenly interested in the problems they're working on often do well in the data analysis. Partly, this is because they are more willing to dig into the nitty gritty of the data and modeling and to uncover small details that others may not find. Also, students with a strong interest often have strong expectations about what the data should show. If the data turn out to be different from what they are expecting, that surprise is often an important experience, sometimes even delightful. Students with a more distant relationship with the topic or the data can never be surprised because they have little in the way of expectations.

Problem Framing

From the discussion it seems clear that we are interested in both the characteristics of different commute methods and the variability associated with individual commute methods. Statistically, these are two separate problems that can be addressed through data collection and analysis. As part of trying to frame the problem, we iterate through a few different scenarios and questions.

One concept that we return to periodically in the discussion is the idea that every question has associated with it a potential intervention. So when I ask "What is the variability in my commute time", a potential intervention is changing the time when I leave home. Another potential intervention is rescheduling my first meeting of the day. Thinking about questions in terms of their potential interventions can be very useful in prioritizing which questions are most interesting to ask. If the potential intervention associated with a question is something you do not have any control over, then maybe that question is not so interesting for you to ask. For example, if you do not control your own schedule at work, then "rescheduling the first meeting of the day" is not an option for

you. However, you may still be able to control when you leave home.

With the question "How long does it take to commute by Muni?" one might characterize the potential intervention as "Taking Muni to work or not". However, if Muni breaks down, then that is out of your control and you simply cannot take that choice. A more useful question then is "How long does it take to commute when I *choose* to take Muni?" This difference may seem subtle, but it does imply a different analysis and is associated with a potential intervention that is completely controllable. I may not be able to take Muni everyday, but I can definitely *choose* to take it everyday.

Fixed and Random Variation

Deciding what is fixed variation and what is random is important at the design stage because it can have implications for data collection, data analysis, and the usefulness of the results. Sometimes this discussion can get very abstract, resulting in questions like "What is the meaning of 'random'?". It's important not to get too bogged down in philosophical discussions (although the occasional one is fine). But it's nevertheless useful to have such a discussion so that you can properly model the data later.

Classifying everything as "random" is a common crutch that people use because it gives you an excuse to not really collect much data. This is a cheap way to do things, but it also leads to data with a lot of variability, possibility to the point of not even being useful. For example, if we only collect data on commute times, and ignored the fact that we have multiple commute methods, then we might see a bimodal distribution in the commute time data. But that mysterious bi-modality could be explained by the different commute methods, a *fixed* effect that is easily controlled. Taking the extra effort to track the commute method (for example, via Hilary's text message approach) along with the commute time could dramatically reduce the residual variance in the data, making for more precise predictions.

That said, capturing every variable is often not feasible and so choices have to made. In this example, Hilary decided not to track whether she wandered into Walgreens or not because that event did have a random flavor to it. Practically speaking, it would be better to account for the fact that there may be an occasional random excursion into Walgreens rather than to attempt to control it every single time. This choice also simplifies the data collection system.

Sketch Models

When considering what to do with the data once we had it, it turned out that mitigating the worst case scenario was a key consideration. This translated directly into a statistical model that potentially had heavy tails. At this point, it wasn't clear what that distribution would be, and it isn't clear whether the data would be able to accurately inform the shape of the tail's distribution. That said, with this statistical model in mind we can keep an eye on the data as they come in and see how they shape up. Further, although we didn't got through the exercise, it could be useful to estimate how many observations you might need in order to get a decent estimate of any model parameters. Such an exercise cannot really be done if you don't have a specific model in mind.

In general, having a specific statistical model in mind is useful because it gives you a sense of *what to expect*. If the data come in and look substantially different from the distribution that you originally considered, then that should lead you to ask *why do the data look different*? Asking such a question may lead to interesting new details or uncover aspects of the data that hadn't been considered before. For example, I originally thought the data could be modeled with a Gamma distribution. However, if the data came in and there were many long delays in Hilary's commute, then her log-Normal distribution might seem more sensible. Her choice of that distribution from the beginning was informed by her knowledge of public transport in San Francisco, about which I know nothing.

Summary

I have spoken with people who argue that are little in the way of generalizable concepts in data analysis because every data analysis is uniquely different from every other. However, I think this experience of observing myself talk with Hilary about this small example suggests to me that there are some general concepts. Things like gauging your personal interest in the problem could be useful in managing potential resources dedicated to an analysis, and I think considering fixed and random variation is important aspect of any data analytic design or analysis. Finally, developing a sketch (statistical) model before the data are in hand can be useful for setting expectations and for setting a benchmark for when to be surprised or skeptical.

One problem with learning data analysis is that we rarely, as students, get to observe the thought process that occurs at the early stages. In part, that is why I think many call for more experiential learning in data analysis, because the only way to see the process is to do the process. But I think we could invest more time and effort into recording some of these processes, even in somewhat artificial situations like this one, in order to abstract out any generalizable concepts and advice. Such summaries and abstractions could serve as useful data analysis texts, allowing people to grasp the general concepts of analysis while using the time dedicated to experiential learning for studying the unique details of their problem.

III Human Factors

16. Trustworthy Data Analysis

The success of a data analysis depends critically on the audience. But why? A lot has to do with whether the audience *trusts* the analysis as well as the person presenting the analysis. Almost all presentations are incomplete because for any analysis of reasonable size, some details must be omitted for the sake of clarity. A good presentation will have a structured narrative that will guide the presenter in choosing what should be included and what should be omitted. However, audiences will vary in their acceptance of that narrative and will often want to know if other details exist to support it.

The Presentation

Consider the following scenario:

> A person is analyzing some data and is trying to determine if two features, call them X and Y, are related to each other. After looking at the data for some time, they come to you and declare that the Pearson correlation between X and Y is 0.85 and therefore conclude that X and Y are related.

The question then is, do you trust this analysis?

Given the painfully brief presentation of the scenario, I would imagine that most people experienced in data analysis would say something along the lines of “No”, or at least “Not yet”. So, why would we not trust this analysis?

There are many questions that one might ask before one were to place any trust in the results of this analysis. Here are just a few:

- What are X and Y? Is there a reason why X and Y might be causally related to each other? If mere correlation is of interest, that's fine but some more detail could be illuminating.
- What is the sampling frame here? Where did the data come from? We need to be able to understand the nature of uncertainty.
- What is the uncertainty around the correlation estimate? Are we are looking at noise here or a real signal. If there is no uncertainty (because there is no sampling) then that should be made clear.
- How were the data processed to get to the point where we can compute Pearson's correlation? Did missing data have to be removed? Were there transformations done to the data?
- The Pearson correlation is really only interpretable if the data X and Y are Normal-ish distributed. How do the data look? Is the interpretation of correlation here reasonable?
- Pearson correlation can be driven by outliers in X or Y. Even if the data are mostly Normal-ish distributed, individual outliers could make the appearance of a strong relationship even if the bulk of the data are not correlated. Were there any outliers in the data (either X or Y)?

The above questions about the presentation and statistical methodology are all reasonable and would likely come up in this scenario. In fact, there would likely be even more questions asked before a one could be assured that the analysis was trustworthy, but this is just a smattering.

Done but Not Presented

I think it's reasonable to think that a *good* analyst would have concrete answers to all of these questions even though they were omitted from the presentation.

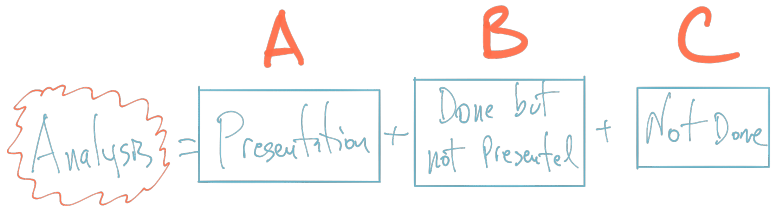
- They would know what X and Y were and whether it made sense to determine if they were correlated, based on the context of the data[1] and questions being asked.
- They would know how the data came to them, whether they represented a sample, and what the population was.
- They would have calculated the uncertainty around the correlation estimate (e.g. a 95% confidence interval) and would have it on hand to present to you.
- Ideally, they would have processed the data themselves and would be able to explain what had to be done in order to get the data to the point where correlations could be computed. If they didn't process the data, they would at least be able to describe what was done and whether those actions have a major impact on the results.
- To justify the use of Pearson correlation, they would have made a histogram (or a Q-Q plot) to look at the marginal distributions of X and Y. If the data weren't Normal looking they would have considered some possible transformations if possible.
- To check for outliers, a scatterplot of X and Y would have been made to examine if any small number of points was driving the observed correlation between X and Y. Even though they didn't show you the scatterplot, they might have it on hand for you to examine.

One might think of other things to do, but the items listed above are in direct response to the questions asked before.

[1] https://simplystatistics.org/2018/05/24/context-compatibility-in-data-analysis/

Done and Undone

My decomposed representation of a data analysis is roughly



Data Analysis Decomposition

Here we have

- **A**: The presentation, which in the above example, is the simple correlation coefficient between X and Y
- **B**: The answers to all of the questions that likely would come up after seeing the presentation
- **C**: Anything that was not done by the analyst

We can only observe A and B and need to speculate about C. The times when I most trust an analysis is when I believe that the C component is relatively small, and is essentially orthogonal to the other components of the equation (A and B). In other words, were one to actually do the things in the "Not Done" bucket, they would have no influence on the overall results of the analysis. There should be nothing surprising or unexpected in the C component.

No matter what data is being analyzed, and no matter who is doing the analysis, the **presentation** of an analysis must be limited, usually because of time. Choices must be made to present a selection of what was actually done, therefore leaving a large number of items in the "Done but not Presented" component. An analogy might be when one writes slides for a presentation, often there are a few slides that are left in the back of the slide deck that are not presented but are easily retrieved should a question come up. The material in those

slides was important enough to warrant making a slide, but not important enough to make it into the presentation. In any substantial data analysis, the number of "slides" presented as the results is relatively small while the number of "slides" held in reserve is potentially huge.

Another large part of a data analysis concerns *who* is presenting. This person may or may not have a track record of producing good analyses and the background of the presenter may or may not be known to the audience. My response to the presentation of an analysis tends to differ based on who is presenting and my confidence in their ability to execute a good analysis. Ultimately, I think my approach to reviewing an analysis comes down to this:

1. If it's a first presentation, then regardless of the presenter, I'll likely want to see A and B, and discuss C. For a first presentation, there will likely be a number of things "Not Done" and so the presenter will need to go back and do more things.
2. If we're on the second or third iteration and the presenter is someone I trust and have confidence in, then seeing A and part of B may be sufficient and we will likely focus just on the contents in A. In part, this requires my trust in their judgment in deciding what are the relevant aspects to present.
3. For presenters that I trust, my assumption is that there are many things in B that are potentially relevant, but I assume that they have done them and have incorporated their effects into A. For example, if there are outliers in the data, but they do not seem to introduce any sensitivity into the results, then my assumption is that they looked at it, saw that it didn't really make a difference, and moved on. Given this, my confidence that the elements of C are orthogonal to the results presented is high.
4. For presenters that I'm not familiar with or with whom I have not yet built up any trust, my assumptions about what lies in B and C are not clear. I'll want to see more of

> what is in B and my skepticism about C being orthogonal to the results will be high.

One of the implications of this process is that two different presenters could make the *exact same presentation* and my response to them will be different. This is perhaps an unfortunate reality and opens the door to introducing all kinds of inappropriate biases. However, my understanding of the presenters' abilities will affect how much I ask about B and C.

At the end of the day, I think an analysis is *trustworthy* when my understanding of A and B is such that I have reasonable confidence that C is orthogonal. In other words, there's little else that can be done with the data that will have a meaningful impact on the results.

Implications for Analysts

As an analyst it might be useful to think of what are the things that will fall into components A, B, and C. In particular, how one thinks about the three components will likely depend on the audience to which the presentation is being made. In fact, the "presentation" may range from sending a simple email, to delivering a class lecture, or a keynote talk. The manner in which you present the results of an analysis is *part of the analysis* and will play a large role in determining the *success* of the analysis. If you are unfamiliar with the audience, or believe they are unfamiliar with you, you may need to place more elements in components A (the presentation), and perhaps talk a little faster. But if you already have a long-term relationship with the audience, a quick summary (with lots of things placed into component B) may be enough.

One of the ways in which you can divide up the things that go into A, B, and C is to develop a good understanding of the audience. If the audience enjoys looking at scatterplots and making inquiries about individual data points, then you're going to make sure you have that kind of detailed understanding in the data, and you may want to just put that kind

of information up front in part A. If the audience likes to have a higher level perspective of things, you can reserve the information for part B.

Considering the audience is useful because it can often drive you to do analyses that perhaps you hadn't thought to do at first. For example, if your boss always wants to see a sensitivity analysis, then it might be wise to do that and put the results in part B, even if you don't think it's critically necessary or if it's tedious to present. On occasion, you might find that the sensitivity analyses in fact sheds light on an unforeseen aspect of the data. It would be nice if there were a "global list of things to do in every analysis", but there isn't and even if there were it would likely be too long to complete for any specific analysis. So one way to optimize your approach is to consider the audience and what they might want to see, and to merge that with what you think is needed for the analysis.

If you *are* the audience, then considering the audience's needs is a relatively simple task. But often the audience will be separate (thesis committee, journal reviewers/editors, conference attendees) and you will have to make your best effort at guessing. If you have direct access to the audience, then a simpler approach would be to just ask them. But this is a potentially time-consuming task (depending on how long it takes for them to respond) and may not be feasible in the time frame allowed for the analysis.

Trusting vs. Believing

It's entirely possible to trust an analysis but not believe the final conclusions. In particular, if this is the first analysis of it's kind that you are seeing, there's almost no reason to believe that the conclusions are true until you've seen other independent analysis. An initial analysis may only have limited preliminary data and you may need to make a decision to invest in collecting more data. Until then, there may be no way to know if the analysis is *true* or not. But the analysis may still

be trustworthy in the sense that everything that should have been done was done.

Looking back at the original “presentation” given at the top, one might ask “So, is X correlated with Y?”. Maybe, and there seems to be evidence that it is. However, whether I ultimately believe the result will depend on factors outside the analysis.

17. Relationships in Data Analysis

In Steve Coll's book *Directorate S*[1], which is a chronicle of the U.S. war in Afghanistan after 9/11, one line stuck out for me as I thought it had relevance for data analysis. In one chapter, Coll writes about Lieutenant Colonel John Loftis, who helped run a training program for U.S. military officials who were preparing to go serve in Afghanistan. In reference to Afghan society, he says, "Everything over there is about relationships." At the time, Afghanistan had few independent institutions and accomplishing certain tasks depended on knowing certain people and having a good relationship with them.

I find data analysis to be immature as an independent field. It uses many tools–mathematics, statistics, computer science–that are mature and well-studied. But the act of analyzing data is not particularly well-studied. And like any immature organization (or nation), much of data analysis still has to do with human relationships. I think this is an often ignored aspect of data analysis because people hold out hope that we can build the tools and technology to the point where we do not need to rely on relationships. Eventually, we will find the approaches that are universally correct and so there will be little need for discussion.

Human relationships are unstable, unpredictable, and inconsistent. Algorithms and statistical tools are predictable and in some cases, optimal. But for whatever reason, we have not yet been able to completely characterize all of the elements that make a successful data analysis in a "machine readable" format. We haven't developed the "institutions" of data analysis

[1] https://www.penguinrandomhouse.com/books/529288/directorate-s-by-steve-coll/9781594204586/

that can operate without needing the involvement of specific individuals. Therefore, because we have not yet figured out a perfect model for human behavior, data analysis will have to be done by humans for just a bit longer.

In my experience, there are a few key relationships that need to be managed in any data analysis and I discuss them below.

Data Analyst and Patron

At the end of the day, someone has to pay for data analysis, and this person is the patron. This person might have gotten a grant, or signed a customer, or simply identified a need and the resources for doing the analysis. The key thing here is that the patron provides the *resources* and determines the tools available for analysis. Typically, the resources we are concerned with are time available to the analyst. The Patron, through the allocation of resources, controls the scope of the analysis. If the patron needs the analysis tomorrow, the analysis is going to be different than if they need it in a month.

A bad relationship here can lead to mismatched expectations between the patron and the analyst. Often the patron thinks the analysis should take less time than it really does. Conversely, the analyst may be led to believe that the patron is deliberately allocating fewer resources to the data analysis because of other priorities. None of this is good, and the relationship between the two must be robust enough in order to straighten out any disagreements or confusion.

Data Analyst and Subject Matter Expert

This relationship is critical because the data analyst must learn the context around the data they are analyzing. The subject matter expert can provide that context and can ask the questions that are relevant to the area that the data inform. The expert is also needed to help interpret the results and

to potentially place them in a broader context, allowing the analyst to assess the practical significance (as opposed to statistical significance) of the results. Finally, the expert will have a sense of the potential impact of any results from the analysis.

A bad relationship between the analyst and the expert can often lead to

- **Irrelevant analysis**. Lack of communication between the expert and the analyst may lead the analyst to go down a road that is not of interest to the audience, no matter how correct the analysis is. In my experience, this outcome is most common when the analyst does not have *any* relationship with a subject matter expert.
- **Mistakes**. An analyst's misunderstanding of some of the data or the data's context may lead to analysis that is relevant but incorrect. Analysts must be comfortable clarifying details of the data with the expert in order to avoid mistakes.
- **Biased interpretation**. The point here is not that a bad relationship leads to bias, but rather a bad relationship can lead the expert to not *trust* the analyst and their analysis, leading the expert to rely more strongly on their preconceived notions. A strong relationship between the expert and the analyst could lead to the expert being more open to evidence that contradicts their hypotheses, which can be critical to reducing hidden biases.

Data Analyst and Audience

The data analyst needs to find some way to assess the needs and capabilities of the audience, because there is always an audience[2]. There will likely be many different ways to present the results of an analysis and it is the analyst's job to figure

[2] https://simplystatistics.org/2018/04/17/what-is-a-successful-data-analysis/

what might be the best way to make the results acceptable to the audience. Important factors may include how much time the audience has to see the presentation, how mathematically inclined/trained they are, whether they have any background in the subject matter, what "language" they speak, or whether the audience will need to make a clear decision based on your analysis. Similar to the subject matter expert, if the analyst has a bad relationship with the audience, the audience is less likely to trust the analysis and to accept its results. In the worst case, the audience will reject the analysis without seriously considering its merits.

Analysts often have to present the same analysis to multiple audiences, and they should be prepared to shift and rearrange the analysis to suit those multiple audiences. Perhaps a trivial, but real, example of this is when I go to give talks at places where I know a person there has developed a method that is related to my talk, I'll make sure to apply their method to my data and compare it to other approaches. Sometimes their method is genuinely better than other approaches, but most of the time it performs comparably to other approaches (alas, that is the nature of most statistical research!). Nevetheless, it's a simple thing to do and usually doesn't require a lot of extra time, but it can go a long way to establishing trust with the audience. This is just one example of how consideration of the audience can change the underlying analysis itself. It's not just a matter of how the results are presented.

Implications

It's tempting to think that the quality of a data analysis only depends on the data and the modeling applied to it. We're trained to think that if the data are consistent with a set of assumptions, then there is an optimal approach that can be taken to estimate a given parameter (or posterior distribution). But in my experience, that is far from the reality. Often, the quality of an analysis can be driven by the relationships between the analyst and the various people that have a stake

in the results. In the worst case scenario, a breakdown in relationships can lead to serious failure[3].

I think most people who analyze data would say that data analysis is easiest when the patron, subject matter expert, the analyst, and the audience are *all the same person*. The reason is because the relationships are all completely understood and easy to manage. Communication is simple when it only has to jump from neuron to anther. Doing data analysis "for yourself" is often quick, highly iterative, and easy to make progress. Collaborating with others on data analysis can be slow, rife with miscommunication, and frustrating. One common scenario is where the patron, expert, and the audience are the same person. If there is a good relationship between this person and the analyst, then the data analysis process here can work very smoothly and productively.

Combining different roles into the same person can sometimes lead to conflicts:

- **Patron is combined with the audience**. If the audience is *paying* for the work, then they may demand results that confirm their pre-existing biases, regardless of what evidence the data may bring.
- **Subject matter expert and the analyst are the same person**. If this person has strong hypotheses about the science, they may be tempted to drive the data analysis in a particular direction that does not contradict those hypotheses. The ultimate audience may object to these analyses if they see contradictory evidence being ignored.
- **Analyst and audience are the same**. This could lead to a situation where the audience is "too knowledgeable" about the analysis to see the forest for the trees. Important aspects of the data may go unnoticed because the analyst is too deep in the weeds. Furthermore, there is potential value and forcing an analyst to translate their findings for a fresh audience in order to ensure that the

[3] https://simplystatistics.org/2018/04/23/what-can-we-learn-from-data-analysis-failures/

narrative is clear and that the evidence as strong as they believe.

Separating out roles in to different people can also lead to problems. In particular, if the patron, expert, analyst, and audience are all separate, then the relationships between all four of those people must in some way be managed. In the worst case, there are 6 pairs of relationships that must be on good terms. It may however be possible for the analyst to manage the "information flow" between the various roles, so that the relationships between the various other roles are kept separate for most of the time. However, this is not always possible or good for the analysis, and managing the various people in these roles is often the most difficult aspect of being a data analyst.

18. Being at the Center

One of the aspects of the job as a data analyst is taking a "systems approach" to solving problems. Coupled with that approach is the role of balancing the various priorities of members of the research or product team. It involves skillfully navigating the roles of dictator and diplomat.

When doing data analysis there will be many people contributing who have specific expertise. It is their job to focus on what they think is the highest priority, but it is the analyst's job to put the whole analysis together and, on the way, raise some priorities and lower other priorities. The analyst is at the center of activity and must have good relationships[1] with every member of the team in order for everything to come together on time and on budget.

Data Analyst at the Center

A mentor once told me that in any large-ish coordinated scientific collaboration there will usually be regular meetings to discuss the data collection, data analysis, or both. Basically, a meeting to discuss data. And that these meetings, over time, tend to become the most important and influential meetings of the entire collaboration. It makes sense: Science is ultimately about the data and any productivity that results from the collaboration will be a function of the data collected. My mentor's implication was that as a statistician directing the analyses of the group, these data meetings were an important place to be.

I have been in a few collaborations of this nature (both small and large) and can echo the advice that I got. The data-related

[1] https://simplystatistics.org/2018/04/30/relationships-in-data-analysis/

meetings tend to be the most interesting and often are where people get most animated and excited. For scientific collaborations, that is in fact where the "action" occurs. As a result, it's important that the data analyst running the analyses know what their job is.

If these meetings are about data analysis, then it's important to realize that the product that the group is developing is the data analysis[2]. As such, the data analyst should play the role of designer. Too often, I see analysts playing a minor role in these kinds of meetings because it's their job to "just run the models". Usually, this is not their fault. Meetings like this tend to be populated with large egos, high-level professors, principal investigators, and the like. The data analyst is often a staff member for the team or a junior faculty, so comparatively "low ranked". It can be difficult to even speak up in these meetings, much less direct them.

However, I think it's essential that the data analyst be at the center of a meeting about data analysis. The reason is simply that they are in the best position to balance the priorities of the collaboration. Because they are closest to the data, they have the best sense of what information and evidence exists in the data and, perhaps more importantly, what is *not* available in the data. Investigators will often have assumptions about what might be possible and perhaps what they would *like* to achieve, but these things may or may not be supported by the data.

It's common that different investigators have very different priorities. One investigator wants to publish a paper as quickly as possible (perhaps they are a junior faculty that needs to publish papers or they know there is a competitor doing the same research). Another wants to run lots of models and explore the data more. Yet another thinks that there's nothing worth publishing here and yet another wants to wait and collect more data. And there's always one investigator who wants to "rethink the entire scientific question". There's no one thing to be done here, but the analyst is often the only one who can mediate all these conflicts.

[2] https://simplystatistics.org/2018/08/24/constructing-a-data-analysis/

What happens in these situations is a kind of "statistical horse trading". You want a paper published quickly? Then we'll have to use this really fast method that requires stronger assumptions and therefore weakens the conclusions. If you want to collect more data, maybe we design the analytic pipeline in such manner that we can analyze what we have now and then easily incorporate the new data when it arrives. If there's no time or money for getting more data, we can use this other model that attempts to use a proxy for that data (again, more assumptions, but maybe reasonable ones).

Managing these types of negotiations can be difficult because people naturally want to have "all the things". The data analyst has to figure out the relative ordering of priorities from the various parties involved. There's no magical one-liner that you can say to convince people of what to do. It's an iterative process with lots of discussion and trust-building. Frankly, it doesn't always work.

The analyst, as the designer of the ultimate product, the data analysis, must think of solutions that can balance all the priorities in different ways. There isn't always a solution that threads the needle and makes everyone happy or satisfied. But a well-functioning team can recognize that and move forward with an analysis and produce something useful.

19. Economic Models for Reproducible Analysis

Every once in a while, I see a tweet or post that asks whether one should use tool X or software Y in order to "make their data analysis reproducible". I think this is a reasonable question because, in part, there are so many good tools out there! This is undeniably a good thing and quite a contrast to just 10 years ago when there were comparatively few choices.

The question of toolset though is not a question worth focusing on too much because it's the wrong question to ask. Of course, you should choose a tool/software package that is reasonably usable by a large percentage of your audience. But the toolset you use will not determine whether your analysis is reproducible in the long-run.

I think of the choice of toolset as kind of like asking "Should I use wood or concrete to build my house?" Regardless of what you choose, once the house is built, it will degrade over time without any deliberate maintenance. Just ask any homeowner! Sure, some materials will degrade slower than others, but the slope is definitely down.

Discussions about tooling around reproducibility often sound a lot like "What material should I use to build my house so that it *never* degrades?" Such materials do not exist and similarly, toolsets do not exist to make your analysis permanently reproducible.

I've been reading some of the old web sites from Jon Claerbout's group at Stanford (thanks to the Internet Archive), the home of some of the original writings about reproducible research. At the time (early 90s), the work was distributed on CD-ROMs[1], which totally makes sense given that CDs could

[1] https://web.archive.org/web/20070911061035/http://sepwww.stanford.edu/research/redoc/cdvswww.html

store lots of data, were relatively compact and durable, and could be mailed or given to other people without much concern about compatibility. The internet was not quite a thing yet, but it was clearly on the horizon.

But ask yourself this: If you held one of those CD-ROMs in your hand right now, would you consider that work reproducible? Technically, yes, but I don't even have a CD-ROM reader in my house, so I couldn't actually read the data. And a larger problem is that a CD from the 90s probably degraded to the point where it is likely unreadable anyway.

Claerbout's group obviously knew about the web and were transitioning in that direction, but such a transition costs money. As does keeping a keen eye on emerging trends and technology usage.

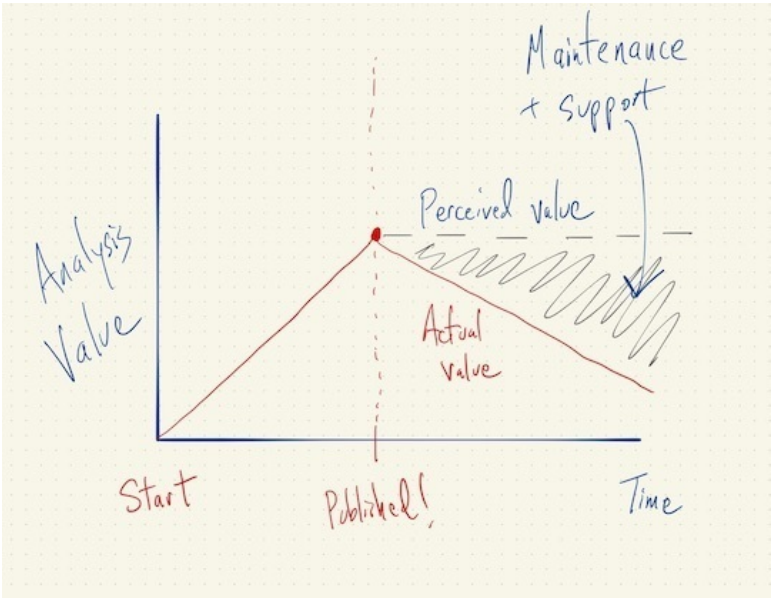
Hilary Parker and I recently discussed[2] the how the economics of academic research are not well-suited to support the reproducibility of scientific results. The traditional model is that a research grant pays for the conduct of research over a 3-5 year period, after which the grant is finished and there is no more funding. During (or after) that time, scientific results are published. While the funding can be used to prepare materials (data, software, and code) to make the published findings reproducible at the instant of publication, there is no funding afterwards for dealing with two key tasks:

1. Ensuring that the work *continues* to be reproducible given changes to the software and computing environment (maintenance)
2. Fielding questions or inquiries from others interested in reproducing the results or in building upon the published work (support)

These two activities (maintenance and support) can continue to be necessary in perpetuity for *every study* that an investigator publishes. The mismatch between how the grant funding

[2] http://nssdeviations.com/106-podcasts-are-a-process

system works and the requirements of reproducible research is depicted in the diagram below.



Analysis Depreciation

When I say “value” in the drawing above, what I really mean is the “reproducibility value”. In the old model of publishing science, there was no reproducibility value because the work was generally not reproducible in the sense that data and code were available. Hence, this whole discussion would be moot.

Traditional paper publications held their value because the text on the page did not generally degrade much over time and copies could easily be made. Scientists did have to field the occasional question about the results but it was not the same as maintaining access to software and datasets and answering technical questions therein. As a result, the traditional economic model for funding academic research really did match the manner in which research was conducted and then published. Once the results were published, the

maintenance and support costs were nominal and did not really need to be paid for explicitly.

Fast forward to today and the economic model has not changed but the "business" of academic research has. Now, every publication has data and code/software attached to it which come with maintenance and support costs that can extend for a substantial period into the future. While any given publication may not require significant maintenance and support, the costs for an investigator's publications *in aggregate* can add up very quickly. Even a single paper that turns out to be popular can take up a lot of time and energy.

If you play this movie to the end, it becomes soberingly clear that reproducible research, from an *economic* stand point, is not really sustainable. To see this, it might help to use an analogy from the business world. Most businesses have capital costs, where they buy large expensive things – machinery, buildings, etc. These things have a long life, but are thought to degrade over time (accountants call it depreciation). As a result, most businesses have "maintenance capital expenditure" costs that they report to show how much money they are investing every quarter to keep their equipment/buildings/etc. up to shape. In this context, the capital expenditure is worth it because every new building or machine that is purchased is designed to ultimately produce more *revenue*. As long as the revenue generated exceeds the cost of maintenance, the capital costs are worth it (not to oversimplify or anything!).

In academia, each new publications incurs some maintenance and support costs to ensure reproducibility (the "capital expenditure" here) but it's unclear how much each new publication brings in more "revenue" to offset those costs. Sure, more publications allow one to expand the lab or get more grant funding or hire more students/postdocs, but I wouldn't say that's universally true. Some fields are just constrained by how much total funding there is and so the available funding cannot really be increased by "reaching more customers". Given that the budgets for funding agencies (at least in the U.S.) have barely kept up with inflation and the number of publications increases every year, it seems the goal of

making all research reproducible is simply not economically supportable.

I think we have to concede that at any given moment in time, there will always be some fraction of published research for which there is no maintenance or support for reproducibility. Note that this doesn't mean that people don't publish their data and code (they should still do that!), it just means they don't support or maintain it. The only question is which fraction should *no*t be supported or maintained? Most likely, it will be older results where the investigators simply cannot keep up with maintenance and support. However, it might be worth coming up with a more systematic approach to determining which publications need to maintain their reproducibility and which don't.

For example, it might be more important to maintain the reproducibility of results from huge studies that cannot be easily replicated independently. However, for a small study conducted a decade ago that has subsequently been replicated many times, we can probably let that one go. But this isn't the only approach. We might want to preserve the reproducibility of studies that collect unique datasets that are difficult to re-collect. Or we might want to consider term-limits on reproducibility, so an investigator commits to maintaining and supporting the reproducibility of a finding for say, 5 years, after which either the maintenance and support is dropped or longer-term funding is obtained. This doesn't necessarily mean that the data and code suddenly disappear from the world; it just means the investigator is no longer committed to supporting the effort.

Reproducibility of scientific research is of critical importance, perhaps now more than ever. However, we need to think harder about how we can support it in both the short- and long-term. Just assuming that the maintenance and support costs of reproducibility for every study are merely nominal is not realistic and simply leads to investigators not supporting reproducibility as a default.

IV Towards a Theory

20. The Role of Theory in Data Analysis

In data analysis, we make use of a lot of theory, whether we like to admit it or not. In a traditional statistical training, things like the central limit theorem and the law of large numbers (and their many variations) are deeply baked into our heads. I probably use the central limit theorem everyday in my work, sometimes for the better, and sometimes for the worse. Even if I'm not directly applying a Normal approximation, knowledge of the central limit theorem will often guide my thinking and help me to decide what to do in a given data analytic situation.

Theorems like the central limit theorem or the law of large numbers ultimately tell us something about the world around us by serving as models. When we say "Let X1, X2, ... be random variables", we're usually thinking of X1 and X2 as abstract representations of real world phenomena. They might be people, or time points in a time series, or cars on the road. What's nice about the theory here is that with a single statement (and a critical set of assumptions), we can make general statements about the world, regardless of whether we're talking about people or cars. In this sense, statistical theory is analogous to scientific theory, which also tries to make general statements about the world. Statistical theory also contains statements about the tools we use, to help us understand their behavior and their properties. Most of this theory is a combination of well-understood mathematical transformations (derivatives, integrals, approximations) and models of the physical world.

Other Theories

There are other kinds of theory and often their role is not to make general statements about the natural world. Rather, their goal is to provide quasi-general summaries of what is commonly done, or what might be typical. So instead of making statements along the lines of "X is true", the aim is to make statements like "X is most common". Often, those statements can be made because there is a written record of what was done in the past and the practitioners in the area have a collective memory of what works and what doesn't.

On War

In his book *On War*, Carl von Clauswitz writes at length about the role of theory in warfare. What's the point of discussing war in the abstract when the reality of war is complicated and highly dependent on the facts on the ground? He sees theory in warfare as a form of training, "a compression of the past transmitting experience, while making minimal claims about the future."

Clauswitz did not envision theory as commanding a general to do something, but rather as an efficient summary of what had worked (or not worked) before. The only alternative to having a theory in this case would be for each general to re-learn the material from scratch. In the practice of warfare, such an approach could easily lead to death. The language of Clauswitz is reminiscent of John Tukey. In his 1962 paper *The Future of Data Analysis*, Tukey writes that theory should "guide, not command" the practice of data analysis.

On Music

Another area that contains a significant amount of theory is music. As I'm a bit more familiar with this area than with warfare, I will draw out a few more examples here.

Music theory is largely a descriptive practice that attempts to draw out underlying patterns that appear in various forms and instances of music. Out of music theory we get things like "Sonata Form", which says that a piece of music has an exposition, a development, and a recapitulation. This form is very common in western classical music and has strong ties to written work.

We also get tonal harmonic analysis, which provides a language for describing the harmonic transitions in a piece of music. For example, most western music has a "key signature", which can be thought of as the primary or "tonic" chord (C-major, for example). All other chords in the scale revolve around that primary chord. These chords are usually referred to using Roman numerals, so the primary or tonic chord is denoted with Roman numeral I. In most pieces of music, the commonly used chords are the tonic (the I chord), the dominant (the V chord), and the sub-dominant (the IV chord). The harmonic pattern of I-IV-V chords is instantly recognizable in many forms of western music written across centuries. We can find chorales written by Johann Sebastian Bach that follow similar harmonic patterns as songs written by The Beatles. In this way, the tonal theory of harmony allows us to draw connections between very disparate pieces of music.

One thing that the tonal theory of harmony does *not* give us is a recipe for what to do when creating new music. Simply knowing the theory does not make one a great composer. It's tempting to think that the theory of harmony is formulated in a manner that tells us what sounds good and what doesn't (and sometimes it is taught this way), but this is misleading. A rote application of the theory will lead to something that is passable, but likely not very inspiring, as you will essentially end up with a reformulation of what has been done before. Part of what makes music great is that it is new and different from what came before.

Arnold Schoenberg, in his textbook *Harmonielehre*, argued strongly against the idea that there were certain forms of music that inherently "sounded good" versus those that "sounded

bad". His thinking was not that the theory of music tells us what sounds good versus bad but rather tells us what is commonly done versus not commonly done. One could infer that things that are commonly done are therefore good, but that would be an individual judgment and not an inherent aspect of the theory.

Knowledge of music theory is useful if only because it provides an efficient summary of what is expected. You can't break the rules if you don't know what the rules are. Creating things that are surprising or unexpected or interesting relies critically on knowing what your audience is expecting to hear. The reason why Schoenberg's "atonal" style of music sounds so different is because his audiences were expecting music written in the more common tonal style. Sometimes, we can rely on music theory to help us avoid a specific chord progression (e.g. parallel fifths) because that "sounds bad", but what we really mean is that such a pattern is not commonly used and is perhaps unexpected. So if you're going to do it, feel free, but it should be for a good reason.

Humans in the Loop

In both examples of theory presented here–warfare and music–the theory only takes you so far, perhaps frustratingly so. Both theories leave room for a substantial human presence and for critical human decision-making. In warfare, Clauswitz acknowledges that there are significant uncertainties that every general encounters and that the role of any theory should be to "educate the mind of the future commander…not to accompany him to the battlefield."

Similarly in music, theory provides an efficient way to summarize the useful aspects of what has come before in a manner that can be fit into something like a semester-long course. It also provides a language for describing characteristics of music and for communicating similarities and differences across musical pieces. But when it comes to the creation of new music, theory can only provide the foundation; the composer must ultimately build the house.

What Does a Theory of Data Analysis Look Like?

If asked to make generally true statements about data analysis, I think most practitioners would struggle. Data analysis relies critically on the details. How could one make a statement that was true for all data analyses when the details differ so much between analyses? And yet, often one person is capable of analyzing data from vastly different domains. Two people who do data analysis in different areas can still have things to talk about related to their work. What are the things that transfer across domains?

One obvious candidate is the methods themselves. A linear regression in biology is the same linear regression in economics. If I am an economist, I may not know much about biology but I could still explain the concepts of linear regression. Similarly, a scatterplot is the same no matter what field it is applied to, even if the things being plotted are different. So the bread and butter of statistics, the study of methods of data analysis, is important. Yet, in my experience, people with solid training in a wide array of statistical methods can still be poor data analysts. In that case, what are they missing?

At this point, many would argue that what is missing is the "experience of doing data analysis". In other words, data analysis is learned through doing it. Okay, fine, but what exactly is it that we are learning? It's worth spending a little time asking this question and considering possible answers because any theory of data analysis would include the answers to this question.

General Statements

The closest thing to a general statement about data analysis that I can come up with is **a successful data analysis is reproducible**. (Note that I do not believe the converse is true.) The concept of reproducibility, whereby code and data accompany a data analysis so that others can re-create the results,

has developed over more than 20 years and has only grown in importance. With the increase in computational power and data collection technology, it is essentially impossible to rely on written representations of data analysis. The only way to truly know what has happened to produce a result is to look at the code and perhaps run it yourself.

When I wrote my first paper on reproducibility in 2006[1] the reaction was hardly one of universal agreement. But today, I think many would see the statement above as true. What has changed? Mostly, data analysts in the field have gained substantially more experience with complex data analyses and have increasingly been bitten by the non-reproducibility of certain analyses. With experience, both good and bad, we can come to an understanding of what works and what doesn't. Reproducibility works as a mechanism to communicate what was done, it isn't too burdensome if it's considered from the beginning of an analysis, and as a by-product it can make data available to others for different uses.

There is no need for a new data analyst to learn about reproducibility "from experience". We don't need to lead a junior data analyst down a months-long winding path of non-reproducible analyses until they are finally bitten by non-reproducibility (and therefore "learn their lesson"). We can just *tell* them

> In the past, we've found it useful to make our data analyses reproducible. Here's a workflow to guide you in your own analyses.

With that one statement, we can "compress" over 20 years of experience.

Another statement I can think of that is applicable to most data analyses is to **discover the data generating mechanism**. When I talk to other data analysts, one of the favorite category of "war stories" to tell is the one where you were bitten by some detail in data collection that went unnoticed. Many data

[1] https://academic.oup.com/aje/article/163/9/783/108733

analysts are not involved in the data collection process or the experimental design and so it is important to inquire about he process by which the data came to them.

For example, one person told me a story of an analysis she did on a laboratory experiment that was ostensibly simple (basically, a t-test). But when she visited the lab one day to see how the experiments were done, she discovered that the experimental units were all processed in one batch and the control units were all processed in a different batch at a different time, thereby confounding any treatment effect with the batch. There's not much a data analysis can do to rescue that situation and it's good for the analyst to know that before spending a lot of time considering the right methodology.

I've written previously about the Law & Order principle of data science[2] where a data analyst must retrace the footsteps of the data to see how they were generated. Such an activity is time-consuming, but I've never come across a situation where it was actively detrimental to the analysis. At worst, it's interesting knowledge to have but it plays no critical role in designing the ultimate analysis.

Most analysts I know have indeed learned "through experience" the dangers of not being informed of the data generating mechanism. But it seems like a waste of time to force new analysts to go through the same experience, as if it were some sort of fraternity hazing ritual. Why not just *tell* them?

Theory Principles

At this point, I think a theory of data analysis would look more like music than it would like physics or mathematics. Rather than produce general truths about the natural world, a theory of data analysis would provide useful summaries of what has worked and what hasn't. A "compression of the past", so to speak. Along those lines, I think a theory of data analysis should reflect the following principles:

[2] https://simplystatistics.org/2018/08/15/the-law-and-order-of-data-science/

- The theory of data analysis is not a theory that instructs us what to do or tells us universal truths. Rather, it is a descriptive or constructive theory that guides analysts without tying their hands.
- A theory should speed up the training of an analyst by summarizing what is commonly done and what has been successful. It should reduce the amount of learning that must be done experientially.
- The theory should serve the practice of data analysis and make data analysis better in order to justify its existence. This is in contrast to traditional statistical theory, whose existence could be justified by virtue of the fact that it allows us to discover truth about the natural world, not unlike with mathematics or physics. In other words, a theory of data analysis would have relatively little intrinsic value.
- The theory should not be dependent on specific technologies or types of data. It should be reasonably applicable to a wide range of data analyses in many different subject areas. A biologist talking to an economist should be able to understand each other when discussing the theory of data analysis.
- The theory should go far beyond the instruction of different methods of analysis, including aspects of data analysis that don't strictly involve the data.

The Scope of Data Analysis

Tukey writes that data analysis should be thought of more as a scientific field, not unlike biochemistry. The key aspect of that comparison is that scientists in any field are comfortable acknowledging that there are things they *don't know*. However, data analysts often feel that they have to have an answer to every question. I've felt this myself–when someone presents a problem to me for which there isn't an obvious solution, I feel a bit embarrassed, as if there *must* be an answer and I just don't know it.

The view of data analysis as a scientific field though is a bit too simplistic in that we shouldn't view solutions to problems as either known or unknown. Often, we can design a solution to a problem even if we are unaware of the "optimal" one. Such a sub-optimal solution will likely be based on intuition, judgment, and our memory of past experience. We may not be happy about the solution, but it might nevertheless be useful. If we are unhappy about the solution developed, we may be inspired to search for the "best" or optimal procedure…or not. Whether such a best solution can be found will depend on whether a realistic optimality criterion can even be specified.

Much of what is written about data analysis (in particular older material) tends to be about activities involving the data. But a distinction needs to be made between what data analysis *is* and what a data analyst *does*. The theory of data analysis will largely focus on what a data analyst *does*, as I think this aspect is potentially more generalizable across disciplines and includes many critically important activities that we don't often discuss.

I think most would agree that what data analysis *is* includes things that involve the data (like fitting a statistical model). But many things that data analysts *do* include things completely outside the data. For example, refining the question to be answered and consulting with experts is a key aspect of what data analysts do, but typically does not involve the analysis of any data (the data may not even be collected yet). Experimental design is another important job where data analysts need to be involved but often does not involve data (although there is substantial theory surrounding this topic already). Allocating resources[3] for a data analysis and building trust[4] with collaborators are also critical things that data analysts *do*.

Because data analysis has become such a critically valuable skill in so many areas of the world, statisticians will have to think harder about what makes for a good data analyst.

[3] https://simplystatistics.org/2018/06/18/the-role-of-resources-in-data-analysis/

[4] https://simplystatistics.org/2018/06/04/trustworthy-data-analysis/

Further, we need to develop better ways to train analysts to do the right thing. Learning by doing will always be a critical aspect of data analytic training, if only because practice is essential (not unlike with music). But we should make sure we are not wasting time in areas where we have a rich collective experience. In other words, we need a *theory of data analysis* that binds together this collective experience, summarizes what has worked and what hasn't, compresses the past, and provides useful guidance for the uninitiated.

21. The Tentpoles of Data Science

What makes for a good data scientist? This is a question I asked a long time ago[1] and am still trying to figure out the answer. Back in 2012, I wrote:

> I was thinking about the people who I think are really good at data analysis and it occurred to me that they were all people I knew. So I started thinking about people that I don't know (and there are many) but are equally good at data analysis. This turned out to be much harder than I thought. And I'm sure it's not because they don't exist, it's just because I think good data analysis chops are hard to evaluate from afar using the standard methods by which we evaluate people.

Now that time has passed and I've had an opportunity to see what's going on in the world of data science, what I think about good data scientists, and what seems to make for good data analysis, I have a few more ideas on what makes for a good data scientist. In particular, I think there are broadly five "tentpoles" for a good data scientist. Each tentpole represents a major area of activity that will to some extent be applied in any given data analysis.

When I ask myself the question "What is data science?" I tend to think of the following five components. Data science is

- the application of **design thinking** to data problems;

[1] https://simplystatistics.org/2012/05/07/how-do-you-know-if-someone-is-great-at-data-analysis/

- the creation and management of **workflows** for transforming and processing data;
- the negotiation of **human relationships** to identify context, allocate resources, and characterize audiences for data analysis products;
- the application of **statistical methods** to quantify evidence; and
- the transformation of data analytic information into coherent **narratives and stories**

My contention is that if you are a good data scientist, then you are good at all five of the tentpoles of data science. Conversely, if you are good at all five tentpoles, then you'll likely be a good data scientist.

Design Thinking

Listeners of my podcast[2] know that Hilary Parker and I are fans of design thinking. Having recently spent eight episodes discussing Nigel Cross's book *Design Thinking*[3], it's clear I think this is a major component of good data analysis.

The main focus here is developing a proper framing of a problem and homing in on the most appropriate question to ask. Many good data scientists are distinguished by their ability to think of a problem in a new way. Figuring out the best way to ask a question requires knowledge and consideration of the audience and what it is they need. I think it's also important to frame the problem in a way that is personally interesting (if possible) so that you, as the analyst, are encouraged to look at the data analysis as a systems problem. This requires digging into all the details and looking into areas that others who are less interested might overlook. Finally, alternating between

[2] http://www.nssdeviations.com

[3] https://www.amazon.com/Design-Thinking-Understanding-Designers-Think/dp/1847886361

divergent and convergent thinking[4] is useful for exploring the problem space via potential solutions (rough sketches), but also synthesizing many ideas and bringing oneself to focus on a specific question.

Another important area that design thinking touches is the solicitation of *domain knowledge*. Many would argue[5] that having domain knowledge is a key part of developing a good data science solution. But I don't think being a good data scientist is *about* having specific knowledge of biology, web site traffic, environmental health, or clothing styles. Rather, if you want to have an impact in any of those areas, it's important to be able to *solicit* the relevant information—including domain knowledge—for solving the problem at hand. I don't have a PhD in environmental health sciences, and my knowledge of that area is not at the level of someone who does. But I believe that over my career, I have solicited the relevant information from experts and have learned the key facts that are needed to conduct data science research in this area.

Workflows

Over the past 15 years or so, there has been a growing discussion of the importance of good workflows in the data analysis community. At this point, I'd say a critical job of a data scientist is to develop and manage the workflows for a given data problem. Most likely, it is the data scientist who will be in a position to observe how the data flows through a team or across different pieces of software, and so the data scientist will know how best to manage these transitions. If a data science problem is a *systems problem*, then the workflow indicates how different pieces of the system talk to each other. While the tools of data analytic workflow management are constantly changing, the importance of the idea persists and

[4] https://simplystatistics.org/2018/09/14/divergent-and-convergent-phases-of-data-analysis/

[5] https://simplystatistics.org/2018/11/01/the-role-of-academia-in-data-science-education/

staying up-to-date with the best tools is a key part of the job.

In the scientific arena the end goal of good workflow management is often reproducibility of the scientific analysis. But good workflow can also be critical for collaboration, team management, and producing *good* science (as opposed to merely reproducible science). Having a good workflow can also facilitate sharing of data or results, whether it's with another team at the company or with the public more generally, as in the case of scientific results. Finally, being able to understand and communicate how a given result has been generated through the workflow can be of great importance when problems occur and need to be debugged.

Human Relationships

In previous posts I've discussed the importance of context[6], resources[7], and audience[8] for producing a successful data analysis. Being able to grasp all of these things typically involves having good relationships[9] with other people, either within a data science team or outside it. In my experience, poor relationships can often lead to poor work.

It's a rare situation where a data scientist works completely alone, accountable to no one, only presenting to themselves. Usually, resources must be obtained to do the analysis in the first place and the audience (i.e. users, customers, viewers, scientists) must be characterized to understand how a problem should be framed or a question should be asked. All of this will require having relationships with people who can provide the resources or the information that a data scientist needs.

[6] https://simplystatistics.org/2018/05/24/context-compatibility-in-data-analysis/

[7] https://simplystatistics.org/2018/06/18/the-role-of-resources-in-data-analysis/

[8] https://simplystatistics.org/2018/04/17/what-is-a-successful-data-analysis/

[9] https://simplystatistics.org/2018/04/30/relationships-in-data-analysis/

Failures in data analysis can often be traced back to a breakdown in human relationships and in communication between team members. As the Duke Saga[10] showed us, dramatic failures do not occur because someone didn't know what a *p*-value was or how to fit a linear regression. In that particular case, knowledgeable people reviewed the analysis, identified exactly all the serious the problems, raised the issues with the right people, and...were ignored. There is no statistical method that I know of that can prevent disaster from occurring under this circumstance. Unfortunately, for outside observers, it's usually impossible to see this process happening, and so we tend to attribute failures to the parts that we *can* see.

Statistical Methods

Applying statistical methods is obviously essential to the job of a data scientist. In particular, knowing what methods are most appropriate for different situations and different kinds of data, and which methods are best-suited to answer different kinds of questions. Proper application of statistical methods is clearly important to doing *good* data analysis, but it's also important for data scientists to know what methods can be reasonably applied given the constraint on resources. If an analysis must be done by tomorrow, one cannot apply a method that requires two days to complete. However, if the method that requires two days is the *only* appropriate method, then additional time or resources must be negotiated (thus necessitating good relationships with others).

I don't think much more needs to be said here as I think most assume that knowledge of statistical methods is critical to being a good data scientist. That said, one important aspect that falls into this category is the *implementation* of statistical methods, which can be more or less complex depending on the size of the data. Sophisticated computational algorithms

[10] https://simplystatistics.org/2018/04/23/what-can-we-learn-from-data-analysis-failures/

and methods may need to be applied or developed from scratch if a problem is too big to work on off-the-shelf software. In such cases, a good data scientist will need to know how to implement these methods so that the problem can be solved. While it is sometimes necessary to collaborate with an expert in this area who can implement a complex algorithm, this creates a new layer of communication and another relationship that must be properly managed.

Narratives and Stories

Even the simplest of analyses can produce an overwhelming amount of results and being able to distill that information into a coherent narrative or story is critical to the success of an analysis. If a great analysis is done, but no one can understand it, did it really happen? Narratives and stories serve as *dimension reduction for results* and allow an audience to navigate a specified path through the sea of information.

Data scientists have to prioritize what is important and what is not and present things that are relevant to the audience. Part of building a good narrative is choosing the right presentation materials to tell the story, whether they be plots, tables, charts, or text. There is rarely an optimal choice that serves all situations because what works best will be highly audience- and context-dependent. Data scientists need to be able to "read the room", so to speak, and make the appropriate choices. Many times, when I've seen critiques of data analyses, it's not the analysis that is being criticized but rather the choice of narrative. If the data scientist chooses to emphasize one aspect but the audience thinks another aspect is more important, the analysis will seem "wrong" even though the application of the methods to the data is correct.

A hallmark of good communication about a data analysis is providing a way for the audience to reason about[11] the data[12]

[11] https://simplystatistics.org/2017/11/16/reasoning-about-data/

[12] https://simplystatistics.org/2017/11/20/follow-up-on-reasoning-about-data/

and to understand how the data are tied to the result. This is a *data analysis* after all, and we should be able to see for ourselves how the data inform the conclusion. As an audience member in this situation, I'm not as interested in just trusting the presenter and their conclusions.

Describing a Good Data Scientist

When thinking of some of the best data scientists I've known over the years, I think they are all good at the five tentpoles I've described above. However, what about the converse? If you met someone who demonstrated that they were good at these five tentpoles, would you think they were a good data scientist? I think the answer is yes, and to get a sense of this, one need look no further than a typical job advertisement for a data science position.

I recently saw this job ad[13] from my Johns Hopkins colleague Elana Fertig. She works in the area of computational biology and her work involves analyzing large quantities of data to draw connections between people's genes and cancer (if I may make a gross oversimplification). She is looking for a postdoctoral fellow to join her lab and the requirements listed for the position are typical of many ads of this type:

- PhD in computational biology, biostatistics, biomedical engineering, applied mathematics, or a related field.
- Proficiency in programming with R/Bioconductor and/or python for genomics analysis.
- Experience with high-performance computing clusters and LINUX scripting.
- Techniques for reproducible research and version control, including but not limited to experience generating knitr reports, GitHub repositories, and R package development.
- Problem-solving skills and independence.

[13] https://fertiglab.com/opportunities

- The ability to work as part of a multidisciplinary team.
- Excellent written and verbal communication skills.

This is a job where complex **statistical methods** will be applied to large biological datasets. As a result, knowledge of the methods or the biology will be useful, and knowing how to implement these methods on a large scale (i.e. via cluster computing) will be important. Knowing techniques for reproducible research requires knowledge of the proper **workflows** and how to manage them throughout an analysis. Problem-solving skills is practically synonymous with **design thinking**; working as part of a multidisciplinary team requires negotiating **human relationships**; and developing **narratives and stories** requires excellent written and verbal communication skills.

Summary

A good data scientist can be hard to find, and part of the reason is because being a good data scientist requires mastering skills in a wide range of areas. However, these five tentpoles are not haphazardly chosen; rather they reflect the interwoven set of skills that are needed to solve complex data problems. Focusing on being good at these five tentpoles means sacrificing time spent studying other things. To the extent that we can coalesce around the idea of convincing people to do exactly that, data science will become a distinct field with its own identity and vision.

22. Generative and Analytical Models

Describing how a data analysis is created is a topic of keen interest to me and there are a few different ways to think about it. Two different ways of thinking about data analysis are what I call the "generative" approach and the "analytical" approach. Another, more informal, way that I like to think about these approaches is as the "biological" model and the "physician" model. Reading through the literature on the process of data analysis, I've noticed that many seem to focus on the former rather than the latter and I think that presents an opportunity for new and interesting work.

Generative Model

The generative approach to thinking about data analysis focuses on the process by which an analysis is created. Developing an understanding of the decisions that are made to move from step one to step two to step three, etc. can help us recreate or reconstruct a data analysis. While reconstruction may not exactly be the goal of studying data analysis in this manner, having a better understanding of the process can open doors with respect to improving the process.

A key feature of the data analytic process is that it typically takes place inside the data analyst's head, making it impossible to directly observe. Measurements can be taken by asking analysts what they were thinking at a given time, but that can be subject to a variety of measurement errors, as with any data that depend on a subject's recall. In some situations, partial information is available, for example if the analyst writes down the thinking process through a series of reports or if

a team is involved and there is a record of communication about the process. From this type of information, it is possible to gather a reasonable picture of "how things happen" and to describe the process for generating a data analysis.

This model is useful for understanding the "biological process", i.e. the underlying mechanisms for how data analyses are created, sometimes referred to as "statistical thinking"[1]. There is no doubt that this process has inherent interest for both teaching purposes and for understanding applied work. But there is a key ingredient that is lacking and I will talk about that more below.

Analytical Model

A second approach to thinking about data analysis ignores the underlying processes that serve to generate the data analysis and instead looks at the observable outputs of the analysis. Such outputs might be an R markdown document, a PDF report, or even a slide deck (Stephanie Hicks and I refer to this as the analytic container[2]). The advantage of this approach is that the analytic outputs are real and can be directly observed. Of course, what an analyst puts into a report or a slide deck typically only represents a fraction of what might have been produced in the course of a full data analysis. However, it's worth noting that the elements placed in the report are the *cumulative result* of all the decisions made through the course of a data analysis.

I've used music theory as an analogy for data analysis many times before[3], mostly because it's all I know, but also because it really works! When we listen to or examine a piece of music, we have essentially no knowledge of how that music came to be. We can no longer interview Mozart or Beethoven about how they wrote their music. And yet we are still able to do a few important things:

[1] https://projecteuclid.org/euclid.ss/1009212754
[2] https://arxiv.org/abs/1903.07639
[3] https://youtu.be/qFtJaq4TlqE

- *Analyze and Theorize*. We can analyze the music that we hear (and their written representation, if available) and talk about how different pieces of music differ from each other or share similarities. We might develop a sense of what is commonly done by a given composer, or across many composers, and evaluate what outputs are more successful or less successful. It's even possible to draw connections between different kinds of music separated by centuries. None of this requires knowledge of the underlying processes.
- *Give Feedback*. When students are learning to compose music, an essential part of that training is the play the music in front of others. The audience can then give feedback about what worked and what didn't. Occasionally, someone might ask "What were you thinking?" but for the most part, that isn't necessary. If something is truly broken, it's sometimes possible to prescribe some corrective action (e.g. "make this a C chord instead of a D chord").

There are even two whole podcasts dedicated to analyzing music—Sticky Notes[4] for classical music and Switched on Pop[5] for pop—and they generally do not interview the artists involved (this would be particularly hard for Sticky Notes). By contrast, the Song Exploder[6] podcast takes a more "generative approach" by having the artist talk about the creative process.

I referred to this analytical model for data analysis as the "physician" approach because it mirrors, in a basic sense, the problem that a physician confronts. When a patient arrives, there is a set of symptoms and the patient's own report/history. Based on that information, the physician has to prescribe a course of action (usually, to collect more data). There is often little detailed understanding of the biological processes underlying a disease, but they physician may have a wealth of personal experience, as well as a literature of clinical trials comparing various treatments from which to

[4] https://stickynotespodcast.libsyn.com
[5] https://www.switchedonpop.com
[6] http://songexploder.net

draw. In human medicine, knowledge of biological processes is critical for designing new interventions, but may not play as large a role in prescribing specific treatments.

When I see a data analysis, as a teacher, a peer reviewer, or just a colleague down the hall, it is usually my job to give feedback in a timely manner. In such situations there usually isn't time for extensive interviews about the development process of the analysis, even though that might in fact be useful. Rather, I need to make a judgment based on the observed outputs and perhaps some brief follow-up questions. To the extent that I can provide feedback that I think will improve the quality of the analysis, it is because I have a sense of what makes for a *successful* analysis.

The Missing Ingredient

Stephanie Hicks and I have written about what are the elements of a data analysis as well as what might be the principles[7] that guide the development of an analysis. In a separate paper[8], we describe and characterize the *success* of a data analysis, based on a matching of principles between the analyst and the audience. This is something I have touched on previously, both in this book and on my podcast with Hilary Parker[9], but in a generally more hand-wavey fashion. Developing a more formal model, as Stephanie and I have done here, has been useful and has provided some additional insights.

For both the generative model and the analytical model of data analysis, the missing ingredient was a clear definition of what made a data analysis *successful*. The other side of that coin, of course, is knowing when a data analysis has failed. The analytical approach is useful because it allows us to separate the analysis from the analyst and to categorize analyses

[7] https://arxiv.org/abs/1903.07639
[8] https://arxiv.org/abs/1904.11907
[9] http://nssdeviations.com/

according to their observed features. But the categorization is "unordered" unless we have some notion of success. Without a definition of success, we are unable to formally criticize analyses and explain our reasoning in a logical manner.

The generative approach is useful because it reveals potential targets of intervention, especially from a teaching perspective, in order to improve data analysis (just like understanding a biological process). However, without a concrete definition of success, we don't have a target to strive for and we do not know how to intervene in order to make genuine improvement. In other words, there is no outcome on which we can "train our model" for data analysis.

I mentioned above that there is a lot of focus on developing the generative model for data analysis, but comparatively little work developing the analytical model. Yet, both models are fundamental to improving the quality of data analyses and learning from previous work. I think this presents an important opportunity for statisticians, data scientists, and others to study how we can characterize data analyses based on observed outputs and how we can draw connections between analyses.

23. Thinking About Failure in Data Analysis

What does it mean for a data analysis to fail? I've come to feel that this is an important question because considering the ways that data analyses can fail is a good way to prevent an analysis from actually failing. However, thinking about the ways an analysis can fail is easier said than done. It requires an active imagination, an ability to think about what *might* happen or what *might have* happened. It requires thinking about what was *not* observed rather than what was observed.

One thing to clarify here is that when I think of failure, I think of something that is more or less immediately observable. So once you've done the analysis, you know whether you've failed or not. This rules out a number of possibilities that I think people are used to thinking about. For example, it rules out the question of whether the claims of an analysis are true or not. Often, it is not possible to determine the truth of an analysis until much later. It also rules out considerations of replication or reproducibility because both of those require future analyses to be done. Therefore, I'm not going to consider these ideas for now and focus on what we as analysts can do in the moment.

Types of Failure

An analysis failure, broadly speaking, is any outcome that does not meet our expectations. This might sound like a bit of an anti-climactic definition of "failure" but ultimately such an outcome represents a failure of our understanding about something: the data (and data generation process), the methods, or the science. As a result, a success is any outcome

that is as-expected. That also seems like a bit of low bar for success, but we can make other distinctions later regarding whether a success is good or not.

My taxonomy of failure falls into two large bins:

- **Verification failure**: This is a narrow class a failure when an analysis produces an outcome that is counter to our expectation in any way. If we are expecting a correlation coefficient to be between 0.4 and 0.6 and we compute a correlation coefficient that is actually 0.1, then that is a verification failure. This failure indicates that there is something we don't quite understand about the underlying correlation or the data.
- **Validation failure**: This is a broader class of failure that includes considerations that are outside the data. Validation failure is typically a result of a poor design process for the analysis that results in flawed requirements. For example, if we are interested in the association between predictor X and outcome Y adjusted for Z, but we build a prediction model for Y that omits Z, then this is a validation failure. The prediction model may operate as-expected and do a good job of predicting Y, but it doesn't answer the question of whether X and Y are associated adjusted for Z.

I find a simple non-data related example can be helpful. A person can ask an architect to design a house with no roof. If the architect builds the house, then the finished house meets expectations from a verification standpoint. It is a verification success. However, when it rains and the inside of the house gets wet, that is a validation failure.

This failure can be traced back to the design process and the setting of the requirements for the project. The actual building of the house went without a hitch. So how could the failure be prevented?

One possibility is the architect could have pushed back and argued that a house with no roof is a bad idea and is therefore

a bad requirement (and probably against local code). The architect might have even refused to build the house. Or the architect could simply ask why the person doesn't want a roof. If it turned out that the person simply wanted a lot of light in the house, then a different requirement could be specified, like a glass roof, that might satisfy everyone's needs.

In general, I think the things we teach about statistics and data analysis in the classroom (textbook knowledge, so to speak) are focused around preventing verification failure. Given a set of requirements, we'd like people to know how to build the widget with the right tools and to be knowledgeable about the assumptions they are making. In my experience, the things that are taught "outside the classroom" are focused on preventing validation failure. Here, we want people to know whether they are building something useful and not simply building something that is "correct".

Potential Outcomes for Analysis

Statisticians like to talk about data analysis failure because it's a little like Monday morning quarterbacking. It's often easy to recognize a failure after the analysis is done. But what about *before* the analysis is done?

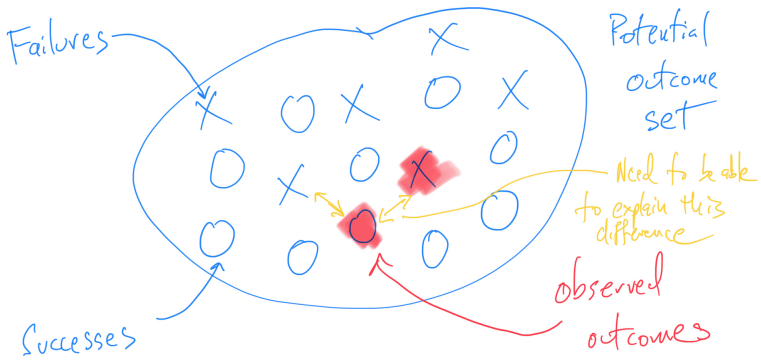
Catching a failure before it happens requires an understanding of the *potential outcomes* of an analysis. Most data analysis problems will admit a range of possible analysis plans and it is up to the analyst to choose one. Given an analysis plan, there is then a set of potential outcomes that this plan can produce before it is applied to the data. Once we apply the plan to the data, we will observe one of these potential outcomes.

For example, if our analysis is to take the arithmetic mean of a set of numbers, the set of potential outcomes is some subset of the real line. If we are making a scatter plot of two continuous variables, then our set of potential outcomes is the set of all possible scatter plots of two variables. Clearly, the

second example with the scatter plot is harder to characterize than the first. But most analysts will intuitively understand what this set looks like.

Once we can characterize the set of potential outcomes, we can divide that set into two broad regions: expected outcomes and unexpected outcomes. When we apply our analysis plan to the data, the outcome will fall in one of these two regions. Outcomes that fall into the "unexpected" region are what I am characterizing as failures here.

Continuing our two examples from above, for the mean we might expect that the outcome will fall into the interval [3, 7]. If the observed mean ended up being 10, that would be unexpected. If it were 4 then that would be as-expected. For the scatter plot, if we believed the two variables were positively correlated, then we might expect the scatter plot to look like a nice cluster of dots shaped like a "football" (American, that is) leaning at roughly a 45 degree angle. If the actual scatter plot looked like a circle (i.e. no correlation) or maybe a circle with a large outlier to the right, that would be unexpected.



Data analysis potential outcome set.

Going From Success to Failure

The traditional notions of success and failure would seem to suggest that we should favor success over failure. But in the data analysis context, what we need to consider is how an analysis can go from success to failure and vice versa. If an analysis outcome is a success and is as-expected, it is important to ask "How could this have failed?" If an analysis outcome is a failure and is unexpected, it is important to ask "How could this have succeeded?"

There is therefore a common task when it comes to the observed output of a data analysis, regardless of whether it could be considered a success or a failure. That task is to consider the entire potential outcome set (given the chosen plan) and determine what could cause one to observe a different outcome than what was actually observed.

In the case of failure, this scenario is a bit easier to understand. When one observes an unexpected outcome, usually one is highly incentivized to get to the bottom of what caused that to occur. It might have been an error in the dataset, or a problem in the data wrangling, or a misunderstanding of how the methods or software work. Finally, there might have been a misunderstanding in our expectation (i.e. in the underlying science) and that perhaps this outcome should have been expected.

In the case of success, it is critical that we apply essentially the same thinking, especially in the early stages of analysis. We should be getting to the bottom of what caused this (success) to occur. In this case, it is still possible to ask whether their might have been an error in the dataset, or a problem in the wrangling, or a misunderstanding of the methods, software or underlying science. In the case of success, it is sometimes valuable to ask what would happen if we *induced* some sort of problem, like an error in the dataset, or an outlier, or a misapplied statistical model (sometimes this is called sensitivity analysis).

In both success and failure it is valuable to consider the

unobserved outcomes in the potential outcome set and ask whether we actually should be observing one of those other outcomes, perhaps because there exists a better model or because the data should be processed in a different way. It is this consideration of the potential outcomes of an analysis, as well as the alternating between success and failure, that drives the iterative nature of data analysis. Ultimately, we want to come to a place where we feel we understand the data, the data generation process, and the analytic process that leads to our result.

Generalizing From Failure

When I hear people (including myself) say that data analysis is learned through experience, I realize now that what we mean is that experience is what allows one to build up that active imagination of what *could* happen. Producing the observed data analytic outcome requires skills-—fitting statistical models, data wrangling, visualization—-that can largely be taught in the classroom. But building the set of potential outcomes becomes easier and faster with experience as we observe more outcomes in other data analyses.

The way that we generalize our experience across different data analyses is to enrich and expand our set of potential outcomes for use in future analyses. What was once unexpected now becomes somewhat as-expected. And because such outcomes are expected, we know to watch out for them and we learn the techniques for checking on them.

But there is another way that we can "learn from experience" that has the potential to take a lot less time——collaborating with other people. Other people may have more experience or they may have different experience. Combining people with different experiences on the same analysis can produce a similar effect to a single person having more experience. Each person has seen different outcomes in their past and together they can produce a potential outcome set that is much larger than they could produce on their own. In this way, people

can gain “experience” by working together. And the more diverse the experiences of the individual collaborators, the richer and larger the potential outcome set will be that they can construct and imagine.