Chapter 2: Learning the “Hello World” of Data

It is natural to assume that the first step in data analysis is getting data. Of course getting the data is necessary, but *starting* with data collection is like running a race in whatever direction one is facing. In order to perform data analysis well, there must be a purpose to the analysis and context it is performed in and jumping

and context surrounding the work both a purpose and context is required. While it’s helpful to have a very specific purpose such as “which of our developers write more secure code?” the purpose may also be quite vague such as, “what traffic patterns exist in our network?” Overall, a good data analysis life cycle has these we want to touch on these points. Everyone doing data analysis does each of these to some degree or another.

Four stages of data analysis:

Frame (set context, questions, purpose)

Acquire (gathering and munging data)

Analyze (data analysis, generate conclusion)

Learn (learning from feedback)

Framing the Analysis

The framing stage is where the “what” and the “why” of the analyses are defined and most analyses are doomed in this stage long before any data are collected. Without a sense of purpose for the analysis, data collection will favor the convenient, analysis will uncover the obvious and conclusions, while perhaps interesting on the surface, will run the risk of being irrelevant. For example, figure 2.1 shows the amount and categories of spam blocked at an organization during a given month. Thanks to the logs generated by an email filtering system, it is entirely possible to collect and show this information. However, the questions and actions that this data addresses are of little interest to the typical organization. Outcomes like figure 2.1 are the result of a poor frame selection and/or skipping the frame selection altogether.

Putting some thought towards the what and why will set up how we proceed through the rest of the stages. Continuing on with the spam example, even though organizations don’t care how much spam is successfully blocked, they may be keenly interested in how many valid emails are blocked as spam or how much time is lost dealing with spam that bypasses our filtering. After thinking about the purpose and context of the analysis we would have concluded that the logs from the email filter may not help our analyses and we should seek other sources the data to address that frame.

A good frame for data analysis will establish both a purpose and a context in which the analysis is performed. The purpose is the specific goal of the analysis, while the context accounts for domain-specific information otherwise lacking. For example, let’s say we want to learn more about the challenges with phishing emails. Just hearing the “what” and skipping any context, we may craft some phishing emails to see how many people are duped by our cleverness. We could then conclude phishing is indeed a problem and we should do something about it. But if we add context of a limited budget for user awareness training and we can only target a subset of employees. Now the data collection and analysis takes on a dramatically different direction. Now we have to understand attributes of who is falling victim instead of how many people clicked. It is essential that we establish and understand both the purpose and context of the analysis before we seek any data.

It may also be worthwhile to understand that any frame that is set will not cover similar situations. It is usually necessary to consider secondary questions or “further” areas to research. Often times data collection is costly and/or time consuming and it’s much easier to add one more question to a survey or pull out one more field in context from the data source then it is to go back later for the data.

We do not cover this step further in this book, however that should not diminish the importance of this step.

Acquiring Data

Once we know what the “what” and “why” of the analysis, it’s time to seek the data that would best serve the frame of the analysis. The data collection is sometimes relatively straightforward. If we want to know (for some strange reason), the proportion of spam emails we block that offer discount prescriptions we can look at all of all of the blocked spam (known as the *population*) and count how many were in the prescription category. This method allows us to count and *describe* the proportion (*descriptive statistics is discussed later in this chapter)*. But if we want to know something a bit more useful like how much spam makes it past our filter, we may have a bigger challenge. We will not be able to gather the population of all emails in order to count the proportion of spam. Instead we could take a *sample* of emails and infer an estimate based on how many were spam (inferential statistics will be discussed in chapter 4). Perhaps we could take a sample of users and have them self-report how many spam messages they receive.

Once we move away from being able to count everything, we run into the need for a sample and how we gather the data will greatly influence the level of confidence we should have in the resulting analysis. For example, if we want to ask users about their level of spam and we ask those sitting near us in the I.T. department, we would be introducing bias into the sample because, as everyone knows, I.T. people are not representative of typical users. We will be discussing sample selection in chapter 4.

Let’s assume we’ve identified where we can gather data from and we’ve figured how the best method of collecting that data. Chances are really good that the data is imperfect and this is where the real world diverges from textbooks. Usually in statistics classes the sample data has either been completely prepared and is free of defects, or there is a single step to prepare it (ya know, as an exercise). In real life we have incomplete data, we have redundant data, and we also must cope with the problem of normalizing data since a single entity could be represented in multiple ways. The topic of data preparation will be covered in chapter 3.

1. Identify sources of data
2. Identify method of data collection
3. clean/munge the data

select a handful of people

want to answer and why we need it answered we can seek sources of data that support that direction. Again, it is critical that we don’t start with data and seek answers without a clear purpose. But

Taking a “encircling” approach

clean data is essential

protecting privacy and “sum of the whole is greater than the parts” effect of data aggregation

Sometimes data doesn’t come from systems but from stochastic models and/or just plain thinking.

What am I trying to do? What am I trying to decide or learn?

What can I meaure/monitor/observe to help that decision?

If I gather this data, will it represent what I’m measuring?

what I’m asking?

looking for data is like running a race by but even before the first datum

Exploratory Data Analysis

This is a slight variation on the four stages where the context or frame of the data analysis is not specifically defined before we begin. In this case we do begin with the data, but not to arrive at a conclusion but to arrive back at the framing of the analysis. In other words, exploratory data analysis helps us understand what questioned are answered are in the data, and we’d want to then use that to see if any of those questions are useful. If that exploration

of the data analyses is not known and before we are able to establish the frame of analysis we have to understand what’s in the data to begin with. So *prior to* establishing the frame of our analyses we may explore the data in an attempt to discover answers to questions we didn’t know to ask.

Acquire

Analyze

Frame

Acquire (if needed)

Analyze (refining)

Learn

Where is time spent? (acquire)