Chapter 2: Beginning with a Question

“My job was to find questions about baseball that have objective answers, that’s all that I do, that’s all that I’ve done.”

-- Bill James, Sabermetrician

It is natural to assume that the first step in data analysis is getting the data to analyze but data analysis is never performed for its own sake. It is always performed within a larger context and understanding that context is the key to a successful data analysis. But before we get into what makes data analysis good, we want to discuss what skills contribute to a good data analysis. We know there is a natural allure to data science and everyone wants to achieve that sexy mystique (oh we know) surrounding data analysts, so we will begin this chapter by talking about how we achieve that mystique. There are multiple disciplines and skills that come together in the science of data analysis and we will walk through each one.

The two most important personality traits of a data analyst are curiosity and communication. Working with data can at times be a bit like how we imagine archeology: spending hour after hour with small tools in the hope of uncovering even the tiniest of insights in the dirt. So it is with data analysis: pearls of wisdom are nestled deep within a mound of data just waiting to be discovered and presented to an eagerly awaiting audience. It is only with that sense of wonder and curiosity that the hours spent cleaning and preparing data are not just tolerable, but somehow worth every moment.

Once those pearls are discovered and cleaned up, they must be communicated to others. Make no mistake, the complexity of both our environment and analysis is difficult to convey to others. Often times it takes a combination of words, numbers and pictures to communicate the insights in data and that’s the personality trait that’s helpful, being able to condense complexity into a paragraph, table and/or graphic. Curiosity is especially important because doing data analysis well requires gaining and maintaining a broad mixture of skills. It takes some dedication and motivation to develop the other skills needed and with a healthy dose of curiosity the skills seem to develop effortlessly in an individual.

While it may be difficult to create an exhaustive list of skills to be a good data analyst[[1]](#footnote-1), we are going to cover the following skills/domains that data analysts benefit from knowing within information security: **domain expertise** (setting and maintaining purpose to the analysis), **data management** (being able to prepare, store and access data), **programming** (the glue that connects data to the analysis), **statistics** (to avoid being lied to by data) and **visualization** (communicating the analysis graphically).

Not all of these skills are required to be present in a single individual. While smaller shops may seek a single individual to cover these areas, as the work load increases it is entirely possible (and a lot more feasible) to spread these tasks across several individuals (or even individual teams). Wherever these skills come from, when they come together we create an environment where data analysis has its best chance at success. It’d be easy for us to label each one of these skills as the most important, but in reality, the whole is greater than the sum of its parts. Each of these contribute a significant piece of the puzzle.

Domain Expertise

Saying that a data scientist needs domain expertise should go without saying. It seems so obvious when we lay it out like this, but this cannot be talked about enough and there are some important points to discuss about the benefits and pitfalls within domain expertise. On one hand, domain expertise sets the entire context for data analysis and can help prevent or identify spurious results in the analysis. On the other hand, domain experts may push away data science, either as a threat to the status quo or because it challenges long-held beliefs which are treated as conventional wisdom.

We will cover how to begin with a question in the next

: context, overfitting, ignorance, pride/vanity

Creating a good question, as we’ll see in the next section when we talk about stages, is probably the most critical step in any data analysis.

Overfitting: data analysts putting meaning where none exists (overfitting is a bad term, mention it ,but it’s not the title).

Ignorance: there is an ignorance of what data science can bring and there is a defensiveness about it for some.

Pride/Vanity: check the assumptions at the door.

On one hand, we have

On one hand, data analysis is never done in isolation, there is

Data analysts need to bend to domain expertise and domain experts need to practice humility

egolessness

There is a natural struggle between those who have domain expertise and those who think they can do it better with data.

Data science is the result of multiple areas of expertise coming together. While it is tempting to say that the most important skill of a security data scientist is expertise in information security, that is just not true.

Domain expertise is absolutely critical, just as are the other areas of expertise. If the domains are not equally and appropriately represented we fall off from the ideal mixture and may lose confidence in the results.

There is also some resistance that we should address in our industry. There is a perception that we (proponents of data science and statistics) want to somehow replace information security

**A warning to information security practitioners:**

Many of the risk-based decisions within information have been made without any science of data and it’s been that way for years. Many of the opinions have gone unchallenged and

For years information security has struggled to make their voice heard and now that there are budgets and a seat at the big table There is a natural resistance to change, but the (and in some organizations, incredibly strong) resistance to the introduction of data as a science because it is seen as threatening to the freedom we’ve had for years.

some may see the introduction of “science” as threatening. That

To describe domain expertise and the context it brings, we want to paint two extremes. On one hand, we have the dedicated and passionate information security practitioner who has dedicated years to intuitively analyzing the security of applications and systems. Perhaps they have helped develop policies of best practices or written hardening guides for various platforms. They have no exposure to statistical methods (or worse, they had that one class back in college) and they scoff at the mere suggestion that a data geek could capture the complexity of their experience in a spreadsheet.

On the other hand, we have a data geek who just wrapped up their masters thesis on data mining techniques with large data sets. They have spent countless hours studying in a classroom and researching in journal articles. They have a combination of models at the ready to toss at whatever scenario they encounter. They scoff at the blind-faith and over-confidence that the information security practitioner has built up over the years on what they see as incredibly faulty data collection.

The truth, as with most complicated things, lies between the two extremes. First and foremost, a data geek will never, ever, ever be able to replace the experience and intuition of an experienced information security practitioner. Data analysis is a supporting function, and serves to support, illuminate and perhaps at times, offset the judgment of the practitioner. Data analysis, as we mentioned at the beginning of this chapter, is always done in a larger context. Ignoring that context or assuming domain expertise where none exists is done at the peril of the conclusion. Even though data may just be a collection of numbers or counts, the relationship and connotations of those data points cannot be guessed.

Overfitting

Programming

Hacking in this sense is the traditional meaning: someone who is masterful at the keyboard, commanding the software and hardware to efficiently complete a challenging task. Hacking skills are the gears of the data analysis watch, rarely seen or understood from the outside, but essential to the function. Every step of the data analysis lifecycle can be (and often is exclusively) performed on computers. Therefore, knowledge of how to command the software and hardware to complete tasks is rather important. If this is not a core skill of a data analyst, then it should be an area of active learning. Actually, even if it is a core skill, this should be an area of active learning -- new techniques, packages and methods are constantly appearing.

In the early state of data analysis, we start with the data. As opposed to the data handed out in statistic classes, data in the real world data comes in a nice neat bundle. It may be embedded in semi-structured log files or it maybe it needs to be scraped from a website, or in really bad cases, data comes in an overly complex and thoroughly confusing format known as an XML document. Somehow this data must be collected, coaxed and massaged into a format that supports further analysis and this is easily done. While this could be done with a lot of patience and a text editor, learning a programming language is way more efficient in the long run. Most every modern language will support basic data manipulation tasks, but the scripting languages (python, R, perl) are used more often in data analysis then their compiled counter parts (Java, C).

There is a “gateway drug” between a text editor and programming known as the spreadsheet (MS Excel, OpenOffice Calc). It allows non-programmers to do some amazing things and enables programmers to do some easy tasks quickly. While spreadsheets have their own sets of challenges and draw-backs, they also have their benefit. If the amount of data is not too large and the task is not deciding the future of the world economy (see case study), then excel may be the best tool for the job. We would strongly suggest seeing excel as a temporary solution. It does well at quick one-shot tasks. But if there is a repeating analytic task or model that is used consistently, moving to a programming language is a very good idea.

type="caseStudy"

The Limits of Spreadsheets

On January 16th, 2013, J.P. Morgan issued a report to shareholders titled “Report of JPMorgan Chase & Co. Management Task Force Regarding 2012 CIO Losses” (<http://files.shareholder.com/downloads/ONE/2532388207x0x628656/4cb574a0-0bf5-4728-9582-625e4519b5ab/Task_Force_Report.pdf>) in which they investigate the loss of $6 billion in trades. In an appendix they have this:

“During the review process, additional operational issues became apparent. For example, the model operated through a series of Excel spreadsheets, which had to be completed manually, by a process of copying and pasting data from one spreadsheet to another. In addition, many of the tranches were less liquid, and therefore, the same price was given for those tranches on multiple consecutive days, leading the model to convey a lack of volatility. While there was some effort to map less liquid instruments to more liquid ones (i.e., calculate price changes in the less liquid instruments derived from price changes in more liquid ones), this effort was not organized or consistent.”

“…the model was approved despite observed operational problems. The Model Review Group noted that the VaR computation was being done on spreadsheets using a manual process and it was therefore “error prone” and “not easily scalable.” …” (p. 105)

“CIO’s implementation of the model was flawed. CIO relied on the model creator, who reported to the front office, to operate the model. Data were uploaded manually without sufficient quality control. Spreadsheet-based calculations were conducted with insufficient controls and frequent formula and code changes were made. Inadequate information technology resources were devoted to the process. Contrary to the action plan contained in the model approval, the process was never automated.”

Aside from the challenges of spreadsheets, they can be a valuable tool in the data analyst’s tool belt. Throughout this book, we will have examples in Python, R and Excel though don’t think data analysis is limited to these tools. They serve as an example of what’s possible and how we may go about solving specific problems within data analysis. Excel makes the list because it is fairly ubiquitous these days and many people attempt to leverage it already for data analysis. Python and R make the list because they are open source efforts and free to download and they are incredibly robust languages with many convenient features that make data analysis pleasurable.

The hacking skill is not just for data preparation. Many of the languages we mentioned have robust data analysis features built into (or onto) the language. So the same programming language we used for data preparation can be used for the analysis itself. For example, the R language was developed by statisticians for performing data analysis. Python, with the addition of packages like NumPy, SciPy and Pandas offers a rich and comparable data analysis environment as well.

But we’re not done there, just preparing and analyzing the data is not enough, we also need to communicate our results and one of the most effective methods for that is data visualization (of which we devote several chapters to here). Again, Excel has the ability to produce graphics and with judicial modification of the default settings, good visualization can be done with Excel. However, in our opinion, flexibility and detail in data visualization is best achieved through programming. Both Python and R have some feature-rich packages for generating and exporting data visualization and we will cover that in later chapters.

Statistics

If hacking skills are the gears of a watch, then math and statistics dictate how all those gears come together to give them purpose and meaning (okay, the analogy was stretched a bit there). We need a way to learn from the data we collect, but more importantly, we also need a way to know when the data may be misleading and when we cannot learn from the data. As with R.A. Fisher’s story at Rothamsted in chapter 1, we may have a lot of data, and perhaps some metrics were derived from it, but without understanding the lessons from statistics, we may have undue confidence in the data.

Generations of people have collected data and tried to learn from it. Nathan Yau, author and maintainer of flowingdata.com, summarized the common misperception of statistics, “To most people statistics means plugging numbers into an advanced calculator that spits out values, without much thought involved. Those people don’t work with data.” (“a life in statistics: Nathan Yau”, significance magazine). This makes a lot of sense if we stop and think about it. The world is messy and it produces messy data. There are complex interdependencies and variations not just in the world but also in our measurements of the world. It makes sense that the majority of statistical methods generate probabilistic statements. Intuitively, a simple “advanced calculator” would simplify the world to the point of being useless, which sadly is how some

hacking (media definition)

All of this is done within some domain, in our case, we’re talking about a rather environment of

The majority of statistics has advanced by new generations of statisticsians showing that the previous methods

developing slightly better ways not to be deceived by the data and hopefully more accurate methods

Statistics is not a precise science, despite its use of mathematics and that’s because the world i

combined with If we stop and think about it, there is no way

Most people are familiar with the basic descriptive statistics (mean, median, etc). But when it comes to

The uninitiated, especially in engineering-based industries like computer science, see statistics as a collection of math formulas and applying statistics is just plugging in the right numbers to get the right answer. This type of thinking paints statistics as a cold unthinking science and nothing could be further from the truth. Most people naturally understand that the real world is filled with complex interactions and natural variations. On top of that, we have variation in our measurements and measuring everything is often impractical. Therefore, an essential concept within statistics is probability.

One of the common questions within statistics is asking how likely our observations are given some assumption. Notice the word “likely” in that statement. The outcome of basic statistics is either to describe the data we’ve seen, which is not statements in probability but are limited in their applicability (and we will see in chapter 4), or we want to make inferential statements about a larger population from limited observations. Descriptive statistics is limited in applicability to scope of the observations we made, while inferential statistics is only capable of making probabilistic statements.

But statistics is more than applied mathematics and models. Statistics also helps us tackle challenges with data collection and how we design our experiments. Understanding how the data was generated and collected help us determine how much confidence we can place on our results. Poor data collection techniques all but negate any inferential outcomes. put into the data representing the world

limit the inferences we can make or in some cases the data is coming from and how it was generated how our analysis are constructed and the data collected can influence how much confidence we should have in the inference we can pull from the data.

Even though we separated out statistics into its own domain, this function is for all practically purposes, exclusively done with computers (wish someone would tell the educational system that little secret). So the same language that we used to prepare the data, may be the same platform we use to perform the analysis.

The choice of programming language is not all that important though scripting languages (python, perl, R) appear to be more prevalent over compiled languages (Java, C). Though most every modern language has its own strengths when it comes to data manipulation, our personal favorites are python and R.

Python is a

. . leads to the ability to programming is an important skill to develop.

Knowledge of programming, however rudimentary

Being able to scrape a web site, transfer

Having at least cursory knowledge of regular expressions (for pattern matching in text)

, and often has a combination of challenges that must be addressed before analysis can begin. Sometimes the (mention python, perl, regex).

The tasks to complete within a data analysis will range from situation to situation. On one hand, every data analyst will have to be comfortable handling and manipulation data (chapter 3) and knowledge (however basic) of programming can serve an analyst well. Also, some data outgrow operations in memory and require a a more robust data management platform (databases are covered in chapter 6).

However, being able to manipulate and transform data is the key among them. We are leaving that definition vague because the tasks are decidedly vague. Obviously we have simple data manipulation functions

Perhaps it is to understand how something is functioning now, or to project how it could be done better, or to compare disparate entities out what is our understanding of

* + aim to find out what makes up some phenomenon: substance, function and rationale (why).
  + comparative - relationships (relation, equiivalance, difference)
  + explanatory - correlation, conditionality, causality
  + normative - how something should be done (best case world)

Perhaps it is to support a critical decision or to define to the uncertainty within a critical decision or define supporting a decision Analysts that lose sight of the larger context do so with disastrous results. It is essential that data analysis occurs with the so

analysis should always be driven by

Need good transition here

Now let’s move on to the stages within data analysis:

Four stages of data analysis:

Goals and Questions (set context, questions, purpose)

Acquire Data (gathering and prepare data)

Analyze (data analysis, generate conclusion)

Learn (learning from feedback)

Picking a Question

This first stage is hands down the most critical part of any data analysis effort. The entire point of this stage is summarized in the opening quote from Bill James. We don’t opening quote to this chapter is from Bill James, who’s work influenced Michael Lewis’, *Moneyball*.

the most critical and … framing stage is where the “why” and the “what” of the analyses are defined and most analyses are doomed in this stage because it is rushed through or skipped altogether.. 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.

frames simplify and organize, frames put a limit on the complexity. a challenge of over-framing - over-simplification, framing out possible off-shoots or discovery paths.do not define the problem with out thought about the assumptions, may overlook options.

Frames are limiting, which are both good and bad things. How a frame is set, will ultimately determine the range of possible conclusions that may be reached.

Careful margining frames, look at separate problems as separate analysis (see below on supporting future though). Not choosing a frame just means that a poorly thought out frame has been chosen (a frame has been chosen with little to no thought).

What proportion of viruses are caught by AV? (notice the proportion, meaning we want to know “out of how many” within the frame.)

Just choosing to count signature matches may be limiting, could we also correlate the payload of the virus?

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

Framing does not mean draw a conclusion, expect to be proven wrong, expect to be consistently wrong. (may have a quote here)

(from Wikipedia: The Research Question serves two purposes:

1. it determines where and what kind of research the writer will be looking for and
2. it identifies the specific objectives the study or paper will address.

From me:

* The frame should be set at the outset, but not set in stone, it must be fluid and able to updated and modified as challenges in data collection or analysis arise.

From “Constructin Research Questions”:

* “Some researchers aim for prediction and explanation while others search for understanding”
* “A fundamental step in all theory development is the formulation of carefully grounded research questions.”
* “If we do not pose innovative research questions, it is less likely that our research oefforts will generate interesting and influential theories”
* “the primary discussion revolves arouns how to formulate feasible research question in a particular sequential order. We are advised to first define the topic (for example. leadership, adult vocationsl learning, diversity among male engineers, middle-class status ansxiety in UK higher ed ), then to clarify the domain of the research, that is, what objects should be studied (individuals, social interaction and so on), state a purpose and finally to decide the type of research questions such as descriptive, explanatory and prescriptive questions.
* Others considering contextual issues, such as how various stakeholders, may influence the formaultion of research questions.
* “from gap-spotting to problemization”
* “with the risk of stating the obvious, therefore, we think it is critical to spell out an answer, one which is simple but fundamental: **questions are the core ingredient in all knowledge development**.” (p. 10)
* Quote: Gadamer “the path of all knowledge leads through the question” (1994, 363) - follow up on this.
  + Gadamer, H.G. (1994) Truth and Method. New York: continuum. (first published in 1960)
* “It is important to note that research questions can be understood in terms of different levels and with more or less precision and focus. … we see research questions not necessarily as very detailed questions or as specific objectives (close to testing an hypothesis). Instead, we regard research questions as setting the somewhat broader intellectual motive of a study, whether it is empirical and/or theoretical, that is, the rationale and direction of a study.”
* Questions provide the necessary starting point and path for all forms of knowledge development. It is by asking questions that we are able to generate knowledge about things. Similarily, asking questions forms the basis of every kind of research investigation.
* not all questions are research questions. Questions must be “researchable” or “investigable” - “they can be investigated scientifically and answered empirically” (p. 11, good stuff here)
* “It is important to note that some questions, which appear to be genuinely open, are in fact ‘closed’ question in the sense that they do not open up the subject matter and instead preserve it. The prime example of closed questions is found in educational settings where the teacher pretends to ask students a genuinely open question, but already has a ready-made answer.” … “however a research questions can never be completely open…”
* Types of research questions: descriptive, comparative, explanatory, and normative.
  + “a central observation made by Dillon is that these questions are to a large extent hierarchically related to each other in the sense that descriptive questions are the most basic, followed by comparative questions, and then explanatory and normative questions”
  + Descriptive - aim to find out what makes up some phenomenon: substance, function and rationale (why).
  + comparative - relationships (relation, equiivalance, difference)
  + explanatory - correlation, conditionality, causality
  + normative - how something should be done (best case world)

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 re

al 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)

Exploratory data analysis … employes a variety of techniques to:

* Uncover underlying structure
* Extract important variables
* Detect outliers and anomalies
* Test underlying assumptions
* Develop parsimonious models
* Determine optional factor setetings
* <http://www.unitedbiosource.com/pdfs/webinars/20121031-exploratory-wasiakr.pdf>

1. For example, we may argue that playing a musical instrument helps teach the creative and critical thinking necessary for good analysis. But alas, that did not make it on the broad list of skills we cover here. [↑](#footnote-ref-1)