Insert Chapter 2 Title Here

A good story can engage the listener and build on the power of our belief. Our goal here, as we talked about in the first chapter, is to build our stories and consequently our beliefs on the feedback we collect from our environment. Fortunately, we don’t have to reinvent the wheel here, there have been generations of clever people trying to figure out we can better learn from our environment and they’ve place a label on it: Statistics.

Before we dive into this topic (and we won’t dive deep), we understand that the word “statistics” can cause even the bravest of tech warriors to reach for the nearest pitchfork. Data and statistics are often abused and misused (and in some cases flat out made up) for the sake of serving an ulterior motive. However, we are in a different situation. We are sitting on a mound of data and we need to understand the stories lying undiscovered. We cannot afford to ignore the work of honest and hardworking stats-geeks that have already found dozens of methods that don’t work. The quickest way to be fooled by a complicated environment is through the avoidance of statistics. But let’s not rush in headfirst because the second quickest way to be fooled by a complicated environment is to use statistics. This chapter will help the reader take a big first step towards making sense of data, but before we get to that we need to take a step back and build a foundation.

As we were discussing our plans for this book, we talked about taking the reader on a statistical journey. We thought we could begin with some descriptive statistics, cross the river into inferential statistics (maybe pontificate about the strengths and weaknesses of regression analysis) and end the journey pointing to the horizon of machine learning and data mining. We quickly realized that wasn’t the path we should take. Not only would this section be at risk of becoming a sleeping aide, we truly believe that the first obstacle isn’t our avoidance of statistics or math -- it’s our avoidance of data and analytics in general that causes us problems and that’s the topic we’re going to tackle first.

Playing with human intuition

In a 2010 presentation to statistics students at the University of Kansas, Bill James said these words: “In my view, anyone who imagines that he can anticipate what will happen next in any area of life is delusional, and people who think that “experts” should be able to do this are children and fools.”

[ “Battling Expertise with the power or ignorance” 4/13/2010 from <http://vimeo.com/16538462> retrieved 3/26/2013 ]

This quote comes across as confusing for anyone who knows anything about the life and work of Bill James (or has seen the movie “Moneyball” where his work in baseball statistics is portrayed). He has spent years trying to anticipate what wins baseball games by collecting and analyzing data. So what or who is delusional here? What he is describing is a flaw we have as people who think we can figure out to anticipate a very complicated system simply by showing up and paying attention, which is a long-winded to say, “relying on our intuition.”

On one hand, everyone knows that relying on our intuition can be a very powerful thing. Most of us choose to get out of bed in the morning because we anticipate our day will be better out of bed then in it. Some of us shower because we anticipate it will feel better (and that we’ll smell better) then if we didn’t. We don’t need to record data and apply a statistical model to tell us these things: they were intuitive and by all accounts, good decisions. Just by living our lives, we have learned to make many valid decisions based on nothing but gut-feel and intuition. But (and this a really big but), there is a limit to these judgments. We wouldn’t want to take a good long look at the sky and attempt to predict the weather matter how many days we’ve looked at the sky in the past. According to the work of Bill James, we also cannot simply watch baseball and intuitively deduce the true value of a fast player or when bunting is a good strategy. Somewhere between the simplicity of getting out of bed and showering and the complexity of weather forecasting and building a winning baseball team there is a tipping point. At some point, the world quietly becomes too complex to understand by simple observation. What’s more is we may not realize it until either the pain of being wrong or the benefit of being right overcomes our faith in the existing set of beliefs.

We may not realize that we’re being fleeced by complexity because our helpful brains fill in any confusing gaps with familiar stories[[1]](#footnote-1) (remember the power of stories?). We make sense of the world by drawing upon our own experiences (or the experiences of others) to conjure a seamless logical flow. When the world is relatively simple this mechanism serves us well. We can make valid comparisons and connections between the qualities of our day in and out of bed. But when the world becomes more complex, familiar stories may be too simple or worse, completely inapplicable and inaccurate. The downside is that when the human minds connects the dots logically we begin to have faith and belief in the validity of our conclusion and changing our beliefs become more difficult with time.

Battling an intelligent adversary

As an example, back in the late 1990’s, I [Jay] was working as an I.T. consultant and felt lucky to be surrounded by some very talented people. One of those people was Alan, a seasoned 30-something system administrator. During one engagement, Alan and I were hired by a medium-sized organization to test their computer security by hacking into it (now called “pentesting” or penetration testing). We spent several weeks mapping their network and slowly gaining access through various exploits and hacks. Alan had developed respect for their system administrator (yes, their one system administrator) as we played an advanced game of cat and mouse with him. It seemed for every attack we’d launch at their systems, it looked like the administrator would fire back with something to prevent our attempts. As we wrapped up the engagement and prepared for our final presentation, Alan was a little nervous about meeting his nemesis face to face. Would he be upset that we had bested him on the field of battle? Would he have the same respect for us as we had for him? I will never forget the look on Alan’s face when the system administrator walked into the meeting. Alan’s nemesis turned out to be a high-school student working a summer job. He had no idea what the meeting was about and responded with indifference when we described pitched battles we had just fought. Turns out that all of the obstacles we found were just the random effect of a complex environment. We envisioned this story of an intelligent adversary because they were the most convenient for us and it was the most logical for us to believe.[[2]](#footnote-2)

Our intuitive explanations for what we observed were completely wrong. We had ventured into a relatively complex environment where our intuitive judgments were simply overwhelmed and we locked onto the first plausible explanation. In this particular case being wrong didn’t cause any problems, just some wasted time and effort avoiding what we thought was a skilled administrator. There are other situations where the complexity of environment and the consequences of being wrong are just too great to blindly trust our intuitive beliefs.

The birth of epidemiology

The field of medicine is a perfect example of both a complex environment and motivational (health) consequences. In our modern times, we expect medicine and science to go hand-in-hand. We wouldn’t consider beginning a medical treatment simply because someone felt it should work out okay. Statistics, clinical trials and years of research and experiments have advanced (and continue to advance) the science of medicine into areas not even fathomable to previous generations. As an example, let’s go back over 150 years ago to a cholera outbreak in the Soho district of London that took hundreds of lives. The root cause of cholera was not known at the time, though the common belief is that it was [food related?] and most of the treatments were based on second-hand stories (anecdotal information) with little or no application of a scientific method.

In the book *Ghost Map*, Dr. John Snow is quoted, as he talks about the prevailing intuitive approaches to an ongoing cholera epidemic in 1854:

…the scientific method rarely intersected with the development and testing of new treatments and medicines. When you rest through the endless stream of quack cholera cures published in the daily papers, what strikes you most is not that they are all, almost without exception, based on anecdotal evidence. What's striking is that they never apologize for this shortcoming. They never pause to say, "Of course, this is all based on anecdotal evidence, but hear me out." There's no shame in these letters. No awareness of the imperfection of the method, precisely because it seemed eminently reasonable that the local observation of a handful of cases might serve up the cure for cholera, if you looked hard enough.[[3]](#footnote-3)

Snow is appalled at not just the lack rigor being applied to this problem, but the apparent comfort level with the overall approach. This is much different then our pentest where the consequence of being was not wasted time -- it is death. Dr. John Snow used data and one very powerful visualization, which we will talk about in a bit, to turn the tables on cholera and it’s said that one event was the birth of a brand new field: epidemiology. It was the power of data combined with the cleverness to simply look at it that started the world down the path of learning like it has never done before.

While there is certainly the perception that statistics can be an enemy, it is not an enemy here. The real adversary here is relatively silent and much more pervasive within technology: the anecdotal story. This type of story creates a bond between an incorrect explanation of cause and effect and our belief system. It is this comfort level with the anecdotal story that we are battling intuition that we must battle and the best weapon to take on this adversary is data.

*[this transition to too abrupt]*

Tapping into the power of data

Saying that we’re surrounded with data should not come as a shock to anyone reading this book. We have the benefit (some may say curse) of living at a time in history when we have more data at our fingertips then ever before in human history and the data just keeps growing. We are also fortunate to have at our disposal the analytic power to make sense of this data and visualize it in ways that were unheard of even 20 years ago. These things are combining to collectively give us power to battle the anecdotal story and learn about our environment like no other generation before us.

This power of data analysis and visualization comes at a price. Even though we have accumulated century’s worth of lessons from statisticians and research, we would be naïve to think we will only experience an upside from this power. With great power comes great responsibility and even greater failure. We undoubtedly have many undiscovered successes and failures ahead of us and it is with that in mind that we can begin to talk about what we can and cannot expect from data analysis and visualization techniques.

Variability [Always|Never] means something

Before we can talk about what data analysis is, we wanted to point out that getting the world to sit still while we measure it is like getting a toddler to eat vegetables. We often can’t account for everything (like the random walk peas seam to take) and it rarely as clean and easy as we had envisioned it. In the physical world, we can often repeat an action over and over and get a slightly different result each time. If we repeatedly flip a coin ten times, we would expect some variation in the number of heads we see. We can hit a golf ball over and over and never see it land in the same place. Most things in the natural world have variation and this generally drives computer professionals crazy.

Computers are built to be deterministic, meaning they are expected to do the same thing over and over and produce the same result time and time again. Even random number generators are referred to as pseudo-random number generators, because given the same initial state; it will produce the same string of “random” numbers. If a computer has variation and does not perform deterministically, it either has broken, it is buggy or it’s Windows 98. Computer professionals are trained to expect little to no variation and frown upon its presence in our systems. Unfortunately, this presents an obstacle in data analysis because variability is the norm, not the exception within data. When people come from a binary world where any deviation from expectation is significant, it’s easy to project that onto data and see any variation in the data as significant.

Let’s say we find ourselves in charge of a small development team. As a way to measure the effectiveness of developers, we decide to track the number of bugs found from each developer during the code review process. After one round of code reviews we find that Andy had 35 while Brody had 44. The common reaction would be to ask Brody to step up his game because it’s obvious he's got more bugs than Andy. After all, we had two developers in the same space where the only perceivable difference was the developers, right? Unfortunately no. In reality there may be natural variations that could inflate or deflate the developers rating. We should expect some difference in the developers’ performance day after day. The complexity of the tasks or inconsistency in the code review process can contribute to a range of probable outcomes for each developer.

What we may find in reality is that the difference between the two developers is not significant giving this natural variation in the overall process. Even though the numbers seem clear, 44 is definitely more than 35, we may not be able to attribute of the difference entirely to the skill of the developer. We must be careful not to place significance where variation is present, and it’s present nearly everywhere. But at the same time, we don’t want to miss out when the data is telling us that a developer has opportunities to improve, right?

Therein lies the challenge: find ways to separate the natural variations in a system from the patterns in the data. Pause for a moment and try to think of some methods for doing that in our scenario. Perhaps just more data, right? If we had more data points, surely the variations would just even themselves out, right? Welcome to a big challenge of data analysis: getting the right data, but before we do that, we have to focus on getting the data right.

Getting the data right

Most people have heard of the famous quote that there are three types of lies in the world, lies, damned lies and statistics [citation/discussion]. This leads to the conclusion that calculation of statistics is somehow a shady practice or not worthy of trust. While statistics are often abused, it is *people* lying and using statistics, not statistics tricking people. However, there are times when honest people reach erroneous conclusions after applying statistical methods. But more often then not, problems are introduced (or not addressed) long before a measurement is taken and we end up not lying with statistics, but lying with data. To better understand lying with data, we turn to a lesson from history.

The magazine Literary Digest ran a large public opinion poll in an attempt to predict the 1936 presidential race. They gathered names from a variety of sources including the telephone directory, club memberships and magazine subscriptions. They ended up with over 2 million responses and predicted a clear winner: Alfred Landon (for those not up on their American history, the democratic candidate, Theodore Roosevelt, won that election carrying 46 states). The problem with the Literary Digest poll began long before a single response was collected or counted. Their trouble began with where they went looking for the data. Remember the year was 1936 and the great depression hadn’t let up yet and they ended up polling people with phones, club memberships and magazine subscriptions. They systematically polled the middle and upper class, which generally leaned towards Landon, and arrived at an answer that was mathematically correct and yet completely wrong.

Let’s go back to our developers. Our first reaction to gather more data is not a bad idea (since we only had one measurement). However, we must also seek to identify and remove sources of bias. How is biased removed? Luckily we can stand on the shoulders of giants here and rely on a nifty concept called the central limit theorem. It’s a very powerful concept (and we’d even say interesting, plus you can geek on it with our tutorial here: <tbd>) but it’s brilliance lies in its dependence on randomness.

Randomness is a funny beast. There was once a statistics professor (statistics is filled with stories about statistics professors) having her students flip a coin 500 times by hand and writing our the results on a piece of paper. When the students brought in their homework the next day, the professor amazed her class by quickly flipping through the results and announcing who had faked their coin flips and who hadn’t. How’d she do that? She relied on the fact that humans are very poor random number generators and they underestimate the probability of patterns within randomness. Turns out that when a fair coin is flipped 500 times, it’s highly unlikely (less then 2% chance) to not have at least 7 consecutive results (heads or tails) in a row. So the professor would perform magic by scanning through the records looking for these missing consecutive results as a sign of cheating.

As a general rule, randomness removes bias. We don’t have to measure everything to have confidence in the outcome, if we can only measure just a few things, but *do so randomly*, we can have confidence that what we measured represents the larger population. The problem with introducing randomness into a real-world problem is that we think we can stir things up a bit and create randomness. People are notoriously poor at generating randomness: quick - pick a random number between 1 and 20! If you are like most people, you chose either seventeen or seven, or at the very least selecting an odd over an even number. But on top of that, we are incapable of recognizing the lack of randomness as well. People thinking randomly should not generate randomness and it shouldn’t be casually applied. Both the generation and application must be intentional and deliberate.

What this means for our developers is we’d have to ask if every step in the development (and code review) process has the same opportunities as others. For example, if the code review is done by hand and Andy’s code review always occurs in the morning, or on Mondays or with the same subset of reviewers, there may be a bias in our measurements (number of bugs found). We’d want to see if we could inject randomness into the mixture by randomly assigning reviewers on random days at random times. Once we introduce randomness, we should have more confidence in drawing conclusions from the data.

One last point on this: if bias is present in the data adding more data does not remove the bias. The Literary Digest was one of the largest (and expensive) polls of it’s day with over 2 million responses. Incidentally a young man named George Gallup did predict the outcome of the election by collecting just 50,000 responses, but did so with randomness.

<I think my sample chapter is going to end here at the moment>

1. There needs to be a citation here, probably to Kahneman, Klein or Gigerenzer, could even expand on this concept more [↑](#footnote-ref-1)
2. There is a reference here to NDM and klein if we have long end-notes or footnote kind of thing. Klein proposes that we make decisions by finding the first plausible scenario and finding ways it won’t work. In our case, the first plausible story was an intelligent adversary and we had no reason to every doubt this, not until we were face to face with the administrator that is. [↑](#footnote-ref-2)
3. To see this quote adapted for information security, see http://rud.is/b/2013/03/27/a-wish-for-snow-in-spring/ [↑](#footnote-ref-3)