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>

Overcoming Analutophobia

Analutophobia (from the Greek words “analutos” meaning “solvable”, and “phobia” meaning “fear”) is a psychological condition in which a person experiences excessive anxiety about the application of data and analytic techniques to solve real world problems currently being solved without data and analytic techniques. There are many reasons for this condition (some of which are valid), but the root cause of this particular phobia is a general lack of understanding of the role of data analysis, what it is and what it is not.

Analytics and numbers are abused and misused everywhere we turn. Every politician keeps a suite of “supporting” statistics at the ready, marketing efforts use numbers that end up being “mostly accurate” and the media sensationalize analysis to stretch headlines well beyond the conclusions of scientific research. These statistics are intended to influence but are misused so often that any presentation of statistics may illicit analutophobic responses in otherwise normal and healthy individuals. But….

There is also a fear that these cold, heartless statistics will replace free thought as if the two cannot possibly exist in the same analysis. As one statistician put it, “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). In fact the opposite is true, the data rarely speaks clearly and rarely goes to the extremes. Most of the time the data is ambiguous

. <calculator reference here?>

(ludicrous) or that the analysis will create a false sense of confidence and we’ll be blinded to the reality of the situation (not so ludicrous) or that

overpromise, under-deliver, sales cycles, marketing, -- within our industry, thrown statistics at us by vendors, juniper 2.5 months ago, claiming value, but was

data analysis replaces human intelligence, (I offer no value), (the data is something and it’s something than - challenging conventional intelligence.

treatment: persistence, small battles, showing small victories.

not all forms of analutophobia are treatable, sometimes it just cannot be overcome.

“waste of time” - sometimes we just want to make a decision,

data analysis can draw spurious conclusions

problems typically solved without data, we’re getting by without statistics now

Arithmetic, Buckets and Comparisons.

Being less wrong versus being right

Information technology is built on deterministic systems. We can produce highly complex systems and given the same input, we can very precisely predict the output. For a variety of reasons, I.T. people have a belief that if something is possible it much be accounted for and uncertainty and There is a certain expectation that if something

still expect absolutely perfect execution. Engineering principles are built on striving for perfection. The world of statistics on the other hand, is built entirely on the world of probability. There is never a right answer in statistics, there is only an answer that is probably more wrong then another.

Simple, most people can connect the keyboard, monitor and mouse to a computer and turn it on.

But there is an obvious limit to that predictive ability and that’s what Bill James was talking about, in a complex environment (human performance in a competitive sport like baseball) our intuitive judgment can fill in the gaps with some wildly crazy stories and through compounding factors becomes conventional wisdom, one of the most dangerous things to get to a complex environment.

Our ability to predict depends on many things, but a couple of things influence greatly: the underlying complexity of patterns in the environment and two, our ability to make sense of those patterns. Our desire to be right (or avoid consequences of being wrong) and how predictable the environment is, that is how well we can observer and learn from the environment. The relatively simple physics of hitting a ball with a bat is observable and understandable by children, but predicting how a complex system of circuitry and software will fail next is not as easily observable.

Our ability to make intuitive prediction is directly and inversely related to the complexity of the environment.

The debate against “algorithms versus intuition” has a simple answer: intuition excels at both speed and accuracy in simple environments and algorithms consistently outperform intuition in complex environments. There is no blanket statement of one is better then the other and to further stop the bickering, they are not exclusive. Algorithms support intuition.

and the next failure will occur in a computer system composed of some of the most complex hardware and software we are capable of making may not be as predictable.

Nathan Yau: “To most people statistics means plugging numbers into an advanced calculate that spits out values, without much thought involved. Those people don’t work with data.” (“a life in statistics: Nathan Yau”, significance magazine)

this is not to say that we are unable to learn with-feel is worthless.

build expertise simply by showing up with our eyes open. learn anything if we just hang around it long enough. A good baseball scout is someone who has been around the game for a while.

We can figure out how

“I don’t really know much of anything about the workings or applications of statistical methods. I could not describe myself as a statistician because I could not meet the standards that people in this room would expect of a professional statistician”

“Experts are people who claim to know things or claim to understand how things works.”

Expertssay that baseball players are in their prime 28-32 - it turned out that this is “totally, wildly and completely untrue, it doesn’t match the data in anyway shape or form”

“large shared body of expertise”

“bunting doesn’t increase the number of runs the team can expect to score or their chance scoring a single run.

speed is not the key element of success of baseball teams

runner on first, no one out, bunt

“There was a great deal of misunderstanding, they thought that I as trying to supplant the experts, or anti-expert. or anti-scout.

speed was important.

experts good starting pitcher would draw a few thousand fans.

“axioms of expertise” to

“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 for the last 35 years.”

Our theory is that he was addressing his critics. Those who swore it was better to bunt in some situation, or steal bases.

But this isn’t at all what he was describing with that quote. He was describing a flaw we have as humans to not know when our own gut-feel is fooling us.

Intuition (as defined by some fancy researchers) is nothing more than the recognition of patterns in our memory. What Mr. James is saying is that our ability to make a prediction based on just the patterns in our memory is impossible.

ability to the recognize patterns in our memory. of patterns in memorycollection of memories.

project the simplicity of our intuition onto a highly complex world. The “experts” he is referring to are those who collected enough stories and memories that they seem like they are the most worthy of trust. But as James showed in his own work, human gut-feel can be remarkably poor at predicting the future.

But with respect to Mr. James, I knew that prediction was achievable without math and can be much simpler then that. I chose to get out of bed this morning because I predicted my day would be better out of bed then in. I showered because I predicted I’d feel better (and smell better) then if I didn’t. I ate breakfast because I predicted it’d give me energy for the morning. I didn’t need to record data and apply a statistical model to tell me these things: they were intuitive and by all accounts, good decisions. Just by living my life, I’ve learned to make many valid decisions based on nothing but a gut-feel. Day after day, our lives are filled with predictions that are (thankfully) fairly accurate. But (and this a really big but), there is a limit to gut-feel judgments. I wouldn’t want to take a good long look at the sky and attempt to predict the weather for the day no matter how many days I’ve looked at the sky in the past. Somewhere between the simplicity of getting out of bed and the complexity of weather forecasting there is a tipping point when our intuitive judgment ceases to be enough. There is a tipping point in system complexity where our ability to predict based on gut-feel should stop receiving our trust.

Strengths and limitations of statistics

Robin W decided to conduct a longitudinal study on the health among I.T. system administrators. The study followed the lives of 30 system administrators all of which died eventually (which greatly affected their performance at work). Robin ran the numbers and concluded the average proportion of system administrators who eventually die is between 88.6% and 100% (with 95% confidence).

There are two takeaways from that story. First, even though it’s absurd, there is a tiny bit of truth in that story. People see statistics being a mindless activity and can envision a math nerd presenting that conclusion while pushing their glasses up their nose and doing a laugh-snort at the end. Statistics is perceived as blindly shoving data into a magic box of math and getting an answer without any real thought. While we are technically able to calculate a confidence interval on 30 deaths out of 30 people, it doesn’t make much sense in the daylight of reality. Everyone understands the probability of death is 100%.

However, what if during the study, Robin found that 30 out of 30 system administrators developed bad eyesight and required glasses? It’s the same exact proportion (30 out of 30), yet we cannot come to the same conclusion as before. It’d be absurd to jump to the conclusion and pronounce, “All system administrators need glasses!” There is a perception that in order to do statistics you have to turn off all forms of logic, reasoning and common sense. Those people have obviously never actually worked with data.

The second takeaway from Robin’s story is that it conveys a certain amount of uncertainty in the statistical conclusion. This is rather important because another misperception of statistics is that it will provide one clear answer. Again we have the users and abusers of statistics to thanks for this. When statistics are misused, the assumptions, qualification and uncertainty in the conclusions are removed or relegated to a minor footnote. Because we’re not focusing on infographics here, we are much more interested in what we can learn from the data. We want to find the stories within the data and unfortunately most of the data is just a reflection of the uncertainty and variability we intuitively see in the world around us.

Given the uncertainty, it’s easy to form the initial (and incorrect) conclusion that statisticians have got the market cornered on CYA (cover your a-posterior) politics. Think about this statement: “Given that we observed 30 out of 30 system administrators in need of eyeglasses, we conclude, with 95% confidence, that the average (mean) rate of eyeglass usage among system administrators is between 88% and 100%.” Which means if we were to study more system administrators, we could expect the average to be less than 88% once in every 20 groups we look at. But what if we do one more sample and see that an average of 70% of system administrators wear eyeglasses? It could just be 1 of the 20 of groups that we’d expect and we’ve got CYA!

[insert Anscomb’s quartet in this section!]

However, once we can get past the initial wishy-washiness in that particular statement, it turns out to be the best statement (and the most honest) that we can communicate given the data in front of us.

But let’s take a more realistic example. Let’s take the problem of phishing emails. Everyone knows what these are, emails that claim to be from some authority or acquaintance and enticing the user to click a link or open an attachment in the hopes of infecting their system. We want to measure just how bad this is so we construct the following email:

Hello Alicia,

I noticed an irregularity in the salary information for your department, could you take a look at the attached list of salaries and see if they match with what you have on record?

Then we send the fake email that looks like it is coming from Human Resources to 100 people in Alicia’s group (or perhaps not in Alicia’s group). The attachment within the email is set to simply record whoever clicks. This enables us to measure who clicks and hopefully get a feel for the success of a real phishing campaign.

Maybe make two/three groups, executive leadership,

Let’s say out of 100 people 18 people click on the attachment. What could we conclude? While the easy answer is to head to a local pub and wishfully think up ways to patch people. The real answer is , to a conclusion that we can’t patch people and hope for the best, in reality we’re going to have some variation but in reality

The problem of course is that within I.T. and engineering we generally seek perfection. And when we don’t see perfection, it’s tossed out as flawed or broken or worthless. We want to know that if we build redundancy into an application that it will work. We confuse measurement with precision instead of accuracy .

get to know statisticians e most wishy-washy folks you could ever imagine.

that the answer. Just because we saw 30 out of 30 system administrators in need of eyeglasses (we’ll stick with that example for now), there will be some natural variation that will occur

Simplification occurs

But before we get into that we should go back to what was brought up at the beginning of this chapter. We are not talking about the type of statistics (or should we say “hype of statistics”) portrayed in the average Infographic. o <http://www.phdcomics.com/comics.php?f=1174> ?)

But thanks in large parts to the science news cycle

Lying with data

It’s easy to focus on the mechanics of mathematics and blame the cold-hard world of numbers for the inaccuracies in our stories. But unfortunately the math is actually the easy part (especially given that most of it done in applications now) and it’s rarely a source of misunderstanding. Most inaccuracies are caused either by the data collection methodology or introduced via error in the preparation and handling prior to any analysis of data has a chance to mess things up.

of data is analyzed.

are caused by the data (typically the collection) itself and not how we handle the data.

Some problems are established before any data are collected. of the problems are either set forth before any bit of data is collected

1. Applying probability to an uncertain world
2. Lying with data
3. Instilling a culture of analytics
4. Planning for analytics
5. Diagnosing & treating analutophobia
6. The role of visualization in data analysis

This breeds the “experts” (and we can toss in “pundits” into that category) that James was referring to. Experts are not those who

When the environment is complex and

unaware s environments we we are trick ourselves into thinking we know how things work when in reality, the system is just too complex.

At some point the pain of being wrong is too great or the benefit of being right outweighs our apathy.

When things become too complex our wonderful brains help us out by projecting stories: droughts are the gods being angry just as much as organizations who suffer a breach must have been idiots. Wearing the same socks during a winning streak in baseball and requiring a new complex passwords every 90 days all become plausible reasons correlations to the patterns we see.

when our intuitive gut-feel ceases to be enough. There is a tipping point in system complexity where our ability to predict based on gut-feel should stop receiving our trust.

The trick to this then is to know when we should become distrustful of a gut-feel approach and start looking for ways to enhance our ability to make decisions. Can we look at enough computer systems to know when there is enough risk present to warrant action just by looking at one more? Better yet, can we identify when not to make something more secure by identifying even though the system isn’t perfect, it is secure enough?

(every decision involving the future is a prediction) machines and mathematics can produce better results, but the challenge is identifying when that is. When do we learn to distrust our gut and learn to trust the data?

Ironically, the answer to that question can only be found accurately with data.

Interesting Quote: “We are shifting from a world in which *we think we know* into a world in which *we know and we can prove it*.” http://blogs.sas.com/content/corneroffice/2013/04/03/q2-2013-intelligence-quarterly-bring-big-data-to-life-with-visual-analytics/

Side note: self-selected survey is questionable in value because of what’s called a “non-response bias”. there is a large part of the selected sample that didn’t answer, perhaps they are hard workers, or apathetic, perhaps what we’re trying to measure is influenced by the reason we did not see a response. Getting medical advice from the internet is similar, those motivated to talk about a procedure are those who may have been horribly wronged by it. This sample bias is important to understand and minimize. As we talk about breach records for instance, we are already starting with a convenience sample, within that convenience sample, we see a 15% identification rate… why is that? can we trust that the 15% we were able to record represents not just the 100% in the full sample, but other breaches? How capable is a potentially biased 15% of the records of a potentially biased sample able to describe the larger population?

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)