Positioning Analytics in Information Technology Security

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 beliefs on the feedback we collect from our environment. Fortunately, we don’t have to reinvent the wheel here, there have been generations of people attempting to the exact same thing and they’ve place a label on it: Statistics.

Before we dive into this topic though, 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. But we are in a different situation, we are sitting on a mound of data and we want to understand the stories lying therein. We cannot afford to ignore the work of honest hardworking stats-geeks who have already found dozens of methods that don’t work. The quickest way to be fooled by data is through the avoidance of statistics. But let’s not rush in headfirst because the second quickest way to be fooled by data is to use statistics. Overall, we want to be able to collect data and draw the right meaning from it. We want to be sure the data doesn’t lie to us, so we don’t end up lying to others, or worse, ourselves.

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>

This quote comes across as confusing for anyone who knows anything about the life and work of Bill James (or has seen him portrayed by Brad Pitt in “Moneyball”). 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.”

Everyone knows that relying on our intuition can be a very powerful thing but there are limits. On one hand, we 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 showered because we anticipated it’d feel better (and we’d 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 a gut-feel and intuition. But (and this a really big but), there is a limit to gut-feel judgments. We 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 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 recruiting baseball players there is a tipping point 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.

On being right versus being less wrong

Information technology is built on deterministic systems. Which means we can produce highly complex systems and given the same input, we can very precisely predict the output.

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.

Patterns and Randomness

Randomness is a funny beast. There was once a statistics professor (statistics is filled with stories about statisticians by the way) 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 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 (2% chance) to not have at least 7 consecutive results in a row. So the professor would scan through the records looking for runs of heads tails in a row as an indication of a valid coin flip.

Humans are not just poor random number generators, but we are also naturally poor at intuitively understanding the patterns that can occur randomly.

The quickest way to be fooled by data is through ignorance of statistics. The second quickest way to be fooled by data is by using statistics.

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 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!

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.

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