siness Series

that help senior-level managers with

clude:

tutions: Driving Bottom-Line Results by

h Social Media and Mobility by Bernie

lert Laursen
s Intelligence beyond Reporting by Gert

m Approach to Maximizing Competitive
gam, and Stefanie Gerlach
ning Your Business in the Global Economy

de from the Experts by Tony C. Adkins
Information Technology, Second Edition

for Borrowers, Lenders, and Investors by

ting Intelligent Credit Scoring by Naeem

ion of the Truth, by Jill Dyche and Evan

ch to Forecasting by Charles Chase
chieving Strategic Objectives by Gregory

am, Jason Wahl, and Stuart Rose
cations for Credit Risk Management by

to Yen to Yuan: A Guide to Fundamental
wan
olution Model to Grow Your Business by

rity and Product Quality by Bobby Hull
ective Direct Marketing by Jeff LeSueur
to Make Knowledge Sharing Work by

Pieces (to Close the Intelligence Gap) by

Execution, Methodologies, Risk, and

Cox
Carlos Andre Reis Pinheiro
ices and Providing Practical Solutions by

heir Data for Business Success by Tony

a Strategy: How Social Networks Are
homas and Mike Barlow
by Thornton May
th to Profitability by Evan Stubbs
an Cox, Marie A Gaudard, Philip J.

les, please visit **www.wiley.com**.

Taming the Big Data Tidal Wave

Finding Opportunities in Huge Data Streams with Advanced Analytics

Bill Franks

John Wiley & Sons, Inc.

CHAPTER 7

What Makes a Great Analysis?

Computing statistics, writing a report, and applying a modeling algorithm are each only one step of many required for generating a great analysis. There is no "easy" button that lets you take one simple step and get a solid result. Not understanding and focusing on what is required to do an analysis right can cause a lot of pain, lead to wrong decisions, and generate enormous levels of extra work.

This chapter will explore several themes. We'll start by clarifying a few definitions, and then we'll discuss a variety of themes that relate to creating a great analysis. Each theme will contain a lesson in the nuances that separate reporting or statistic generation from analysis, as well as meaningful analysis from useless analysis.

The principles discussed apply broadly and are not specific to big data. However, with big data adding even more complexity to the mix than organizations are used to dealing with, it's more crucial than ever to keep the principles in mind. Your organization won't be able to tame the big data tidal wave by reports alone. Nor will you be able to tame it through substandard analytics.

ANALYSIS VERSUS REPORTING

Too many organizations mistakenly equate reporting with analysis. That may seem harsh at first glance, so let's clarify what is meant by

the statement. Reports are important and can be valuable. Reports used correctly will add value. But reports have their limits, and it is important to understand what they are.

In the end, an organization will need both reporting and analysis to succeed in taming big data, just as both reporting and analysis have been utilized to tame every other data source that's come along in the past. The key is to understand the difference between a report and an analysis. It is also critical to understand how they both fit together. Without that understanding, your organization won't get it right.

THOUGHT IS WHAT CREATES AN ANALYSIS

An analysis can lead to reports, and reports can lead to an analysis. It is even possible to have an analysis based entirely off of reports. For example, you might run 10 reports, line them up on the desk, identify the key things you see in each, and write a summary of what you found and what it means. That's an analysis. It is the thought that a person puts into the business implications of data or statistics that makes an analysis. Data and statistics without any interpretation are useless.

Reporting

Let's start by defining "reporting." A reporting environment, as we will define it here, is also often called a business intelligence (BI) environment. Such an environment is where users go to select the reports they want to run, get the reports executed, and view the results. The reports may contain tables, graphs, and charts in any combination. The key factors that define a report include:

- A report will provide back to the user the data that was asked for.
- That data will be provided in a standardized, predefined format.
- There is no person involved in generating a report outside of the user who requested the report through his or her reporting interface. (This assumes the report template itself has already been created and deployed.)
- As a result, reports are fairly inflexible.

Let's clarify that last point. Complicated report templates can be created with a variety of prompts and filters. There may be many

options in such reports, but within the constraints of those predefined options they are fairly inflexible. The average user is not typically able to generate a completely new report or overhaul how the predefined prompts and filters work. The user can simply fill in the prompts and filters that are already in place.

One way that reports are often misused is when having a bunch of reports available is mistaken for having a lot of analysis available. There is a common phenomenon that can be found in many organizations. The person in IT who is in charge of the business intelligence environment will say, "We have a world-class BI environment. We've got 500 reports available that cover every possible aspect of the business that anybody could want. Our businesspeople have everything they need."

At the same time, a business user will say, "I am so frustrated! We spent a year or two building this reporting system and I still don't have what I need." If the businesspeople get in a room with the IT people, the conversation often starts with the businesspeople complaining they don't have what they need. The IT people will tell them they are crazy since there are 500 reports available. It can devolve into an argument of finger pointing and accusation.

The disconnect stems from the fact that buried somewhere within those 500 reports probably is more or less what the business users want. But when they are overwhelmed with 500 reports it's very difficult for them to find what they need. In addition, any two people might want to look at things just a little bit differently. Each business user might want to have one extra metric on a report or to have it organized in a different manner. There may be 500 reports out there, but none of them are exactly what any given businessperson wants.

IN REPORTING, SIZE DOESN'T MATTER!

Many IT organizations focus on building as many reports that cover as many topics as possible. This can be driven by business users who submit requirements that cover anything they may ever possibly want, rather than what they actually need and will use. As a result, huge suites of reports often overwhelm users and they don't get what they need. Focus on providing a more limited set of relevant reports. Don't fall into the trap of assuming whoever has the most reports available wins!

It's far better to produce a handful of reports that are exactly what end users want than it is to create an all-encompassing suite of 500 reports. It isn't the number of reports, but the relevance of the reports, that matters. Too often, the number of reports gets the focus (with more being assumed to be better, of course!) rather than the relevance. As we'll discuss next, even having the perfect mix of reports for every business user still doesn't provide analysis. It simply makes a lot of data available to feed into the analysis process.

There are times where further analysis of a report really isn't required. For example, assume you have a report of sales by product by week and you want to know if your products hit their sales target last week. By running the report, the answer is right in front of you, and no further work or analysis is required to get your answer. This is one way that reports can add a lot of value. They can be configured to answer common questions quickly and simply. If everything looks good, there isn't a need for any further work. If something is seen that doesn't match expectations, then further analysis to determine why will be necessary.

Analysis

Now that we've defined reporting, let's define analysis. From there, it will be possible to compare and contrast the two. The key points that define an analysis are:

- An analysis provides answers to the questions being asked.
- An analysis process takes any steps needed to get the answers to those questions.
- An analysis is therefore customized to the specific questions being addressed.
- An analysis involves a person who guides the process.
- By its very nature, the analysis process is flexible.

An analysis is really about saying: "I've heard the problem. I'm going to put together what is needed to address the problem." It's an interactive process of a person tackling a problem, finding the data required to get an answer, analyzing that data, and interpreting the

results in order to provide a recommendation for action. The differences between analysis and reporting are summarized in Table 7.1.

Interplay between reporting and analysis is common and necessary. In fact, each makes the other more effective. For example, consider a sales manager with a basic sales summary report showing monthly sales by region. It is a very simple report that he looks at every day so he can get a feel for whether or not the business is on track. One day he sees something incredibly unusual that he doesn't understand. As a result, he walks down the hall and alerts the analytics team that there's something weird on the sales summary report. He asks them to dig in and find out what's going on. His request based on that report just spawned an analysis, which is exactly what it should be doing.

On the flip side, consider the analytic professional assigned to investigate that problem. She goes through and identifies what some of the underlying causes are. She comes back and shows the sales manager what she found. The manager may comment that the data she just put together is very, very useful. While she generated it to identify what caused this specific issue today, he'd like to see the same information on an ongoing basis, even if things do look to be on track.

What just happened? Her analysis of a problem today has led to a new standard report. She automates what she did, and it becomes a standard report moving forward.

One thing to keep in mind as your organization attempts to tame big data is that a great analysis can be created by simply piecing

Table 7.1 Summary of Analysis versus Reporting

Reporting	Analysis
Provides data	Provides answers
Provides what is asked for	Provides what is needed
Is typically standardized	Is typically customized
Does not involve a person	Involves a person
Is fairly inflexible	Is extremely flexible

together information you already have in different ways for a new purpose. It's looking at the business in a way that hasn't been done before. As much as analytics professionals love to talk about all the fancy stuff they do, a huge portion of what any of them do isn't that exciting. It is getting the data ready for their analysis and often doing a lot of simplistic computations as a starting point.

THE VALUE OF ANALYSIS IS IN LOOKING AT DATA DIFFERENTLY

The point of analysis is to not make a problem harder than it needs to be. Sometimes a simple analysis will do the trick and provide all the answers needed. Just looking at data differently can often yield powerful insights. If there isn't a need to get fancy, then don't. Instead, be happy a simple solution was found and move on to the next problem.

It often isn't necessary to get too fancy before an answer becomes clear. The value is in doing things in a different way more than doing something fancy. For example, perhaps some anomalies in sales are noted at a retail chain. One solution would be to build a complex predictive model that attempts to determine what drivers went into creating those anomalies. However, a first step might be to look at whether there were any supply-chain issues. Perhaps a shipment was delayed or a major weather event kept customers home. If it is possible to identify such a cause, there is no need to build a fancy model. You've found your explanation via a simple analysis and can stop there.

ANALYSIS: MAKE IT G.R.E.A.T.!

For analysis to have an impact, it has to be done well. There are a number of factors that need to come together for an analysis to truly add value. What separates a great analysis from a poor analysis? A great analysis will meet the G.R.E.A.T. criteria! Let's briefly cover what those are.

Guided

A great analysis will be guided by a business need. It won't be an analysis done just because it is interesting or fun. With big data in particular, it is easy to get drawn into a lot of interesting but irrelevant work. A great analysis is one that starts through the identification of a specific business problem. Once underway, the analysis is guided by what is required to solve that problem. Every step of the analysis should be guided by the needs of the problem being addressed.

Relevant

Clearly, any great analysis has to be relevant to the business. This means more than just choosing an arbitrary business problem. The problem needs to be one that the business feels needs a solution, and it has to be a problem that the business has an ability to address. There is no point to figuring out how sensitive different segments of customers are to the price point of a product if the product is being discontinued. It just isn't relevant.

Explainable

A great analysis will need to be explained effectively to those tasked with acting on it. It is possible to get carried away with formulas, algorithms, and statistics. While technical details may be the proof required behind the scenes that an analysis is valid, the results need to be explained in terms that decision makers can understand and digest. A great analysis will be explainable and easy for them to make use of.

Actionable

A great analysis will be actionable. It will point to specific steps that can be taken to leverage the results to improve a business. There is no point to an analysis showing that moving a few stores a mile down the road will increase sales if there is no way the company would ever

actually move the stores. Without providing the ability to be acted upon, an analysis is just noise.

Timely

A great analysis will be delivered in a timely fashion so that it is available when decisions need to be made. Having the answer to a question next month doesn't do any good if the decision needs to be made next week. It is possible for an analysis to be great in every aspect, but it just can't be completed in time for the decision it supports. If so, look for another problem to focus effort on. A late analysis isn't great.

CORE ANALYTICS VERSUS ADVANCED ANALYTICS

This book talks a lot about advanced analytics. This raises the question of how "advanced" analytics are different from other analytics. Let's refer to nonadvanced analytics as core analytics to keep it simple. Core analytics tend to ask simple questions and provide simple answers. A core analytics process is going to investigate what happened, when it happened, and what the impact was.

Let's illustrate with an example. A product manager needs to know how a sales promotion performed last month. Did the company get a lot of new subscribers to sign up as planned? How would a core analysis look into this question? A core analysis might look at how many subscribers signed up. That's what happened. How did the sign-ups occur by day? That's when it happened. How much money did the new subscribers bring in, and how did that compare to the baseline? That's the impact.

Note that all the data for the core analysis in this case can be provided by standardized reports. The analysis itself is the process of examining those reports, making inferences, and suggesting action. In this case, the analysis will consist of looking at the numbers and determining if the goals were met or not. The product manager can then determine whether or not the promotion can be considered to be a success.

The problem is that a core analysis such as this leaves a couple of questions unanswered. Specifically, why did the promotion produce these results, and what can be done about it in the future?

ADVANCED ANALYTICS GOES DEEPER

Advanced analytics goes beyond what happened, when it happened, and what the impact was. It also tries to identify what caused it to happen and what can be done about it in the future. Advanced analytics encompasses a range of activities including complex ad hoc SQL, predictive modeling, data mining, forecasting, optimization, and other similar activities.

Advanced analytics goes further than core analytics. Advanced analytics includes everything from complex ad hoc SQL, to forecasting, to data mining, to predictive modeling. One question that often arises is how advanced analytics is different from data mining, forecasting, or predictive modeling. The answer is that everything you would think about when you think of those activities is encompassed within advanced analytics. However, advanced analytics also includes other processes that aren't necessarily algorithm-intensive, such as ad hoc SQL—not basic everyday SQL queries, but highly complex SQL queries that involve combining data sources in complex ways.

The reason activities like advanced SQL are included in the definition is that the main goal of advanced analytics is to quantify the cause of events, predict when they might happen again, and identify how to influence those events in the future. Sometimes it doesn't require a fancy model to get the insights you require to answer those questions.

As an example, imagine a company is doing an initial exploration of customer web activity. An analysis is commissioned to identify if viewing a product on the web increases the likelihood of purchase or not. It will take quite a bit of work to parse the web data and combine it with other customer data since the web data is new. A starting point can be as simple as a correlation analysis. For the first effort, there isn't a need to build a fancier model and process. If a strong correlation is found between browsing and sales, then the company can be comfortable marketing to people who browsed but did not buy. Perhaps later they'll want to quantify the relationship more precisely. But in the short term, they're confident they've found a pattern that they can profit from, so they use it.

Advanced analytics is an important part of an organization's overall analytical strategy. It can help take an organization to the next level. Advanced analytics involves very complex SQL or data manipulation along with modeling, forecasting, data mining, and similar disciplines. While an organization won't have as many people with the skills to do advanced analytics, those people can provide very powerful insights that would not otherwise be possible.

LISTEN TO YOUR ANALYSIS

No analysis is great unless it is taken seriously. One common trap to be aware of is the trap of "cherry picking" analysis findings. Often there is an executive who has been around for a very long time. He or she is from back in the days when there was no choice but to make decisions based largely on gut feel. That executive's guess is probably usually very good. It's hard to find a high-level person in an organization who isn't pretty good at getting it right with his or her gut. Such executives get where they are because their gut has been good at making the right choices. The goal of analysis isn't to completely replace executives' gut instincts, but to enhance it with facts.

There are many cases where executives like those described in the preceding paragraph are resistant to letting numbers and data tell them what to do. Some executives will request an analysis to see whether the data supports an action being considered. When the analysis comes back and supports the action, the analysis is proudly used to further justify the decision and show that the decision is backed by the data. The analytic professionals who generated it are thanked profusely. That sounds like a terrific outcome, doesn't it?

The problem comes into play when the analysis comes back and suggests that perhaps the executive's plan isn't looking so good. If a company and an executive are committed to analytics and fact-based decisions, it is necessary to heavily reconsider that plan. What sometimes happens, however, is that the results are swept under the table. The recommendation to proceed is made anyway. There is no mention of the analysis when explaining the decision, but there is mention of all the other reasons why the company should proceed.

The preceding example is cherry picking. It's only using analytics when the results serve your purpose. If you're going to use analysis, you've got to use it across the board and consistently. Cherry picking and using the results only when they bolster your case isn't using analytics to make your business any better whatsoever. It's simply doing what you always would have done and using an analysis as extra justification in those cases where it supports you. Since no decisions will actually be changed by the analytics performed, there really is no point to doing the analysis and no benefit from it.

DON'T SUPPORT CHERRY PICKING!

One of the worst abuses of analytics is to cherry pick results. Cherry pickers tout analysis findings when the results serve the purpose at hand. But, they ignore the findings when the results conflict with the original plan. An organization claiming it uses analytics to make decisions when cherry picking is standard practice is dishonest. Nothing will change or improve in such an environment. There will just be a lot of extra time and money spent on analysis efforts that change nothing.

FRAMING THE PROBLEM CORRECTLY

In order to have a great analysis, it is necessary to ask the right question, gather the right data to address it, and design the right analysis to answer it. Perhaps the most important part of all of the distinctions between a great analysis and something less is framing the problem correctly up front. We need to start at the beginning, before the analysis process begins.

Framing the problem means ensuring that important questions have been asked and critical assumptions have been laid out up front. For example, is the goal of a new initiative to drive more revenue or more profit? The choice made leads to a huge difference in the analysis and actions that follow. Is all the data required available, or is it necessary to collect some more data? Have alternatives been considered in terms of how to design an analysis to address the problem? Without framing the problem, all the rest of the work is junk. It will result in a classic garbage in, garbage out scenario.

Consider the example of a team of consultants building a customer segmentation model for a client. The client had a business-to-business component and a business-to-consumer component. While the consultants were aware that the client had a business-to-business component, it was relatively small and had never been mentioned in any of the meetings leading up to the project.

The client sent the consultants data for the project. The consultants started struggling with the models because there were customers that were very extreme in their behavior. The consultants informed the client that something unusual was going on, as they were seeing very odd patterns that they couldn't explain. The client immediately said, "That's our business-to-business customers." The client acknowledged that only consumers had been discussed, but as it ends up, they provided all the data for both businesses and consumers in the data feed.

FRAMING IS EVERYTHING

The way you frame your problem and design your analysis are more important than anything done after that. An analysis can't be accurate and useful if the problem is poorly framed and a poor analysis is designed. Place the proper emphasis on the framing and design process to make sure it is done right. Otherwise the result won't be a great analysis.

The inclusion of businesses is an important thing to know to say the least! The consultants had not framed and designed the analysis appropriately to account for the business customers, and those customers were messing up the models. The consultants ended up building the client two models: one for businesses and one for consumers. It was necessary to have the business-to-business customers separated because they had such totally different patterns of behavior. In order to frame the problem correctly, it was necessary to either focus on only one type of customer, or to build a model for each type of customer.

Great analysis starts with framing the problem correctly. This includes assessing the data correctly, developing a solid analysis plan,

and taking into account the various technical and practical considerations that are in play. Arguably, framing the problem is the most critical step of an analysis, because if it isn't done right, neither will be anything that follows.

STATISTICAL SIGNIFICANCE VERSUS BUSINESS IMPORTANCE

Analytic professionals put a lot of focus on statistical significance, and that's not a bad thing. The key is that statistical significance is only part of the story in delivering a great analysis. Statistical significance testing takes a set of assumptions and determines the probability that the results seen would happen if the assumptions are correct.

For example, if it is assumed that a coin is fair, then it will land heads and tails each 50% of the time. With a fair coin, the odds of getting 10 tails in a row are very small. If 10 tails in a row are seen, there are only two possibilities. The first is that a streak of luck that occurs in only one in 1,024 attempts was just witnessed. The second is that the coin isn't really fair after all. A significance test related to a run of 10 tails would say that you can be 99.9 percent or so confident that the coin isn't fair. This is because a fair coin will only yield such a result 0.1 percent of the time. Such a computation is what statistical significance is all about.

It is necessary to differentiate between statistical significance and business importance. They are not the same. Let's examine why.

Statistical Significance

Statistical significance is used frequently for averages and percentages. It's also used to evaluate the parameter estimates that come out of statistical models. Statistical significance testing can be very, very valuable for helping make sure that data doesn't fool you. It will say from a mathematical perspective whether a difference is large enough to be of merit or not. There are times when differences that appear to be significant will not be and times that differences that appear small will be found significant. A statistical test will make sure the right conclusions are reached.

There is an entire discipline built around testing. A common term in the business world for this discipline is test and learn. Test and learn is really just the basic experimental design concepts taught in statistics classes in college. In a test and learn environment, an experiment is designed so that it is possible to specifically measure the effects of one or more options and identify which of the options is going to work best.

Businesses have to be diligent in making sure they follow the correct approach and don't simply run with the "obvious" answer. One of my favorite examples of something completely counterintuitive is from a problem in graduate school. Take a look at Table 7.2. There are two baseball players who played together for five seasons. What can be seen is that Joe had a higher batting average than Tom in every single one of those five seasons. So a very simple question to ask is: "Who has the higher batting average across those five seasons?" Take a moment, think about it, and lock in your answer.

And the answer is . . . This may surprise you: We don't know who has the best overall batting average! There is not enough information in Table 7.2 to know who had the best batting average across all five seasons. How can that be? If we knew that Joe and Tom had the same number of at bats in each season, then the answer is as simple as it seems. Joe would be the winner. But what if they had different numbers of at bats? What if, in the season when both Joe and Tom had their best average Joe was hurt for a few months and had only a fraction of the at bats as Tom? Similarly, what if Tom was injured in

Table 7.2 Baseball Batting Averages by Season

Season	Tom	Joe	Winner
1	.252	.255	Joe
2	.259	.266	Joe
3	.237	.241	Joe
4	.253	.255	Joe
5	.256	.257	Joe

the season with the lowest averages so Joe had a lot more at bats? It ends up that Tom can have a higher aggregate batting average than Joe even though Tom was beaten in every single season! It's not going to be the most common scenario, but it is absolutely possible.

NEVER TAKE SHORTCUTS

When you are given only part of the story, it is possible to be completely wrong in the conclusions you reach. Never take the easy way out and decide that results are so compelling that it isn't necessary to go through the formality of proving their statistical significance. Always ensure you have all the data you need and do all the tests required against that data before reaching your conclusions.

Without knowing the number of at bats, it isn't possible to say who did better overall. Take a look at Table 7.3 to see an example of how Tom can be the winner across the five-year period in total. The difference between Tom's and Joe's averages are not statistically significant in this case based on a t-test. So instead of what appears to be the obvious answer that Joe beat Tom, we find that Tom actually beat Joe. But it isn't that simple, either. Even though Tom won, the

Table 7.3 Full Comparison of Batting Averages

Year	Tom: Avg.	Tom: At Bats	Tom: Hits	Joe: Avg.	Joe: At Bats	Joe: Hits	Winner
1	.252	123	31	.255	341	87	Joe
2	.259	355	92	.266	109	29	Joe
3	.237	139	33	.241	377	91	Joe
4	.253	304	77	.255	294	75	Joe
5	.256	363	93	.257	206	53	Joe
Total	**.254**	**1,284**	**326**	**.252**	**1,327**	**335**	**Tom!!! ***

*Tom did win, but by an amount that is not statistically significant. From a statistical perspective, Tom and Joe are tied.

margin of victory is not statistically significant. They are tied from a statistical perspective. The answer is more nuanced than it appears.

Most people would see Table 7.2 and not bother giving the question any more thought. They would go with what appeared to be the obvious answer that Joe had a better overall average. Never do that. Always make sure you test and validate.

There is one last point to cover related to statistical significance. Most people start to feel pretty comfortable when they're 95 percent or 99 percent certain that their experiment crossed a given hurdle. The thing to keep in mind is that when you're 95 percent certain you are right, there is still a 5 percent chance you are wrong. That means that one out of every 20 times you do a similar experiment, you can expect to be wrong when you accept the results.

Make sure that the level of certainty tested for matches the level of risk that can be taken comfortably and affordably. For example, if a company would go completely bankrupt with an incorrect decision, then 95 percent certainty doesn't seem so great. Perhaps 99.9 percent or higher would be a better level of certainty to aim for.

Over a large a series of actions, the chances of being wrong at least once begin to get quite large. You have to be ready to absorb those mistakes. Or, you need to set your significance bar very, very high to keep the risk very, very low. Clinical trials for new drugs utilize a very high bar because the impacts of a bad drug are so large, including even death. The bar for deciding if a company should put Image A or Image B at the top of a web page for the rest of the day can be much lower.

Business Importance

We've covered how statistical significance is meaningful, how it is necessary to be careful to get complete data and do the right tests, and that nobody can ever be 100 percent sure the right decision is being made. The story doesn't end there. The final step is to assess the importance of a statistically significant finding to the business.

Let's assume statistical significance is found in an analysis. There is another layer of equally important, or even more important, questions to ask. There is a statistically significant result, which is terrific. But is it important to the business? How is the business going to use

and take action on the statistically significant results? A real effect has been found, but is it large enough to have a meaningful impact?

Always put the results in a business context as part of the final validation process. Maybe it is possible to be 99 percent confident that the lift in response from a given change to an offer is at least 10 percent. That's good. But what if the baseline is a basic offer and the change tested is a bonus offer that costs twice as much? In that case, getting an extra 10 percent of response may not cover the extra costs. The fact that the response rate is significantly higher doesn't actually matter. It still isn't important from a business perspective.

Look beyond significance tests and take into account the bigger picture. What are the costs associated with making the recommended changes? How much additional revenue might be generated over time? Is the new approach consistent with the overall corporate strategy? Are people and man-hours available to make the process changes that will be required? Statistical significance is critically important, but it only matters to the extent that it can be validated to be something that is important from a business perspective.

A GREAT ANALYSIS PROVIDES VALUE, NOT NOISE

It is crucial to understand the difference between statistical significance and business importance. This is especially true as big data comes of age. Analytic professionals are going to find some really interesting things in big data. As numbers geeks, they may say, "Wow! This is cool!" But it is important to ask whether or not the business will care about it. Part of an analysis is to ask if what has been found just happens to exist, or if it's also relevant and actionable. Otherwise, it is just noise.

SAMPLES VERSUS POPULATIONS

It used to be that sampling was a necessary and common practice. A huge concern was often whether or not there would be enough sample size for the problem at hand. With big data, having enough data for a sufficient sample is certainly not an issue. Using today's scalable systems, it's often possible to work with an entire population. It is no longer necessary to take a 10 percent sample of customers because

that's all that can be handled. There are some areas, such as clinical trials, where small sample sizes can still cause trouble. Those areas are the exceptions rather than the rule today. However, it is still important to consider when sampling should be a part of an analysis plan. When a sampling process is needed, it needs to be done correctly.

The next time you're reading a newspaper, look at the surveys that are invariably contained within it. What you'll see at the bottom of any survey result is a stated margin of error. This is typically in the plus or minus 3 to 5 percent range. You'll also see the size of the random sample used, which is typically in the 800 to 1,200 person range. These margins of error and sample sizes are going to be very consistent regardless of the question, regardless of the topic, and regardless of the size of the population being sampled. All that is needed to be within a few percentage points is around 1,000 responses.

The bigger a sample is made, the tighter the margin of error will be and the higher the certainty that the "real" answer is pretty darn close to what was seen in the sample. Big data can yield sample sizes so large that common summary statistics end up showing high, high levels of statistical significance. These differences can be extremely tiny, and irrelevant from a business perspective.

Perhaps a sample of hundreds of millions of web sessions is explored to study how many people clicked on link A or link B. It very well might be found that 2.5235 percent click A and 2.5237 percent click B. That difference of .0002 percent can be statistically significant if the sample is big enough. However, it is such a small difference that even if it's statistically significant, it doesn't matter. It doesn't meet the business importance or relevance criteria we've discussed. As an old statistical guideline simply states, "A difference is a difference only if it makes a difference."

It used to be that analysts would stress over having enough sample. The concern was that the margin of error of the analysis would be too big with a small sample. When a sample is too small, differences need to be relatively large to be found statistically significant. Many analyses would be effectively pointless under those conditions. These days it is almost necessary to make sure that too big of a sample isn't used. Having too much sample seems an odd concept. But it is a concept to consider.

If there is a specific problem that only requires exploring 200,000 random customers to get the precision needed, processing 20 million

just because it is possible is a waste of time and resources. Consider choosing a sample size so that it will start to find a statistically significant difference at the same time the difference becomes important and relevant. If a 1 percent difference must be found to take action, choose a sample size where 1 percent is going to be where statistical significance will start to be found. Using a huge sample may lead to finding differences of small fractions of a percent to be significant. This leads to a lot of extra processing for no practical gain. Ensure there is a large enough sample, but not one that is vastly larger than required. Taming big data will require trimming data down to the essential pieces.

There are times when 100 percent of the data is absolutely needed. One of the most common examples is where it is necessary to find a "Top N" list based on some criterion. For example, it may be necessary to identify the 100 highest-spending customers. By definition, any random customer sample won't have all the top 100 customers, but only a random subset of them. It is necessary to search all the customers to get the full top 100. As before, the problem will dictate if a sample is needed, and what size the sample should be. Make good use of samples when possible.

Another common misconception is that a single sample will work for many different problems. A marketing department may only need a 10 percent sample of customers for their work. So, marketing takes a sample of 10 percent of customers and then gets all of the activity for that 10 percent of customers. This sample won't work for other departments. Why? Let's explore.

YOU NEED ALL YOUR DATA!

As many different samples are pulled for many different problems, you will eventually touch 100 percent of the underlying data. Don't make the shortsighted mistake of throwing away data not needed for a specific problem. Sampling doesn't negate the need to collect and store all of the relevant data that you can. An enterprise data environment is not built from a sample. Samples are pulled from an enterprise data environment!

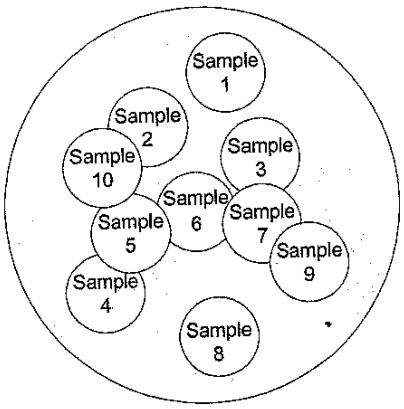
Consider a telecommunications company. A 10 percent customer sample works great for the customer relationship management (CRM)

team. Soon, however, the retail team needs to analyze performance of retail locations. What that group needs is a 10 percent sample of locations and all of the transactions tied to those locations. It's a sample striped a whole different way. They might not have all the information on any given customer, but they'll have every piece of information for a given store. Similarly, a product manager may need a 10 percent sample of all of the transactions that included their product. This sample won't necessarily have all the transactions for any given customer or any given store. All three departments need a different type of sample.

The point is that any given problem may require just a 10 percent sample. But every problem might require a *different* 10 percent sample from the last, as Figure 7.1 shows. Over time, as different samples are pulled for all sorts of different problems, 100 percent of the data will be required at some point. Therefore, all data must be kept and made available, even if never more than 10 percent of it is used at once!

MAKING INFERENCES VERSUS COMPUTING STATISTICS

This topic is the heart of the difference between analysis and reporting, and between great analysis and poor analysis. Imagine that an analysis



Any given problem may require only a small sample of the data. However, over many samples chosen for many problems all of the data will be needed.

Figure 7.1 Different Samples Require Different Data

has uncovered statistics that are significant. The analytic professional behind it has also validated that the findings are important and relevant to the business. Now he has to infer what might be done as a result. To produce a great analysis, it is necessary to infer potential actions that can be taken and to provide some guideposts about what can be done as a result of the findings. In addition, if there are actions that the analysis doesn't support, then those too need to be documented.

A great analysis makes the decision process for a decision maker as easy as possible. The decision maker is going to have to make the final call. The important thing is that an analysis summary provides suggestions as a starting point. A great analysis needs to make initial inferences and not just compute statistics. Just as a report isn't an analysis, simply providing statistics or other technical information isn't an analysis, either.

YOU NEED ANALYTIC PROFESSIONALS, NOT REPORTERS!

The job of an analytic professional is to provide analysis and recommendations, not reports, data, and statistics. Just as there is value in reports, there is value in someone who can analyze data and provide the output necessary to address a problem. The big value, however, is added when those results are interpreted and an action plan is generated. That is what turns reports into analysis and reporters into analytic professionals.

It isn't enough to point out that option #1 outperformed option #2 by 10 percent. Given all of the results, what decision should be made? A great analysis will include recommended steps. If option #2 outperformed option #1 by 10 percent, then include the statement that option #2 should be implemented. In a simple case like this, that is pretty obvious. Many analyses will be more complex, however. In those cases, guidance on what action the results imply will be immensely helpful. The decision maker shouldn't need to figure out the options on his or her own. He or she should be given options to accept or reject.

WRAP-UP

The most important lessons to take away from this chapter are:

- Reporting is not analysis. Generating a report is only the starting point for an analysis. Analysis and reporting both help make the other more effective when used appropriately.
- Analysis is about doing whatever it takes to enable a fact-based decision to address a business issue. Anything from reports to predictive models can play a role in the analysis process.
- A G.R.E.A.T. analysis is Guided by a business problem, Relevant, Explainable, Actionable, and Timely.
- Advanced analytics goes beyond the simple questions of what happened, when it happened, and what the impact was. It also looks into why it happened and what can be done about it.
- One of the worst ways to pursue analytics in an organization is to cherry pick positive results and ignore negative results. Such behaviors negate the purpose and value of analysis.
- The most important part of any analysis happens before it begins. The way the problem is framed up-front can determine the success or failure of the analysis.
- Statistical significance is not the same as business importance! Do not rely exclusively on statistical measures to determine what the important findings of an analysis are.
- Statistical significance tests provide only a probability of being correct. Tie the significance level tested for to the ramifications of the rare cases where the wrong call will be made.
- Even if it is possible to work with an entire population, it may add expense and effort without practical benefit. Sampling is a good strategy in many cases, including with big data.
- A great analysis involves offering inferences and potential actions, not simply reporting statistics and facts.