CHAPTER 13

Data Science and Business Strategy

Fundamental concepts: *Our principles as the basis of success for a data-driven business; Acquiring and sustaining competitive advantage via data science; The importance of careful curation of data science capability.*

In this chapter we discuss the interaction between data science and business strategy, including a high-level perspective on choosing problems to be solved with data science. We see that the fundamental concepts of data science allow us to think clearly about strategic issues. We also show how, taken as a whole, the array of concepts is useful for thinking about tactical business decisions such as evaluating proposals for data science projects from consultants or internal data science teams. We also discuss in detail the curation of data science capability.

Increasingly we see stories in the press about how yet another aspect of business has been addressed with a data science-based solution. As we discussed in Chapter 1, a confluence of factors has led contemporary businesses to be strikingly data rich, as compared to their predecessors. But the availability of data alone does not ensure successful data-driven decision-making. How does a business ensure that it gets the most from the wealth of data? The answer of course is manifold, but two important factors are: (i) the firm's management must think data-analytically, and (ii) the management must create a culture where data science, and data scientists, will thrive.

Thinking Data-Analytically, Redux

Criterion (i) does not mean that the managers have to be data scientists. However, managers have to understand the fundamental principles well enough to envision and/or appreciate data science opportunities, to supply the appropriate resources to the data science teams, and to be willing to invest in data and experimentation. Furthermore, unless the firm has on its management team a seasoned, practical data scientist, often the management must steer the data science team carefully to make sure that the team stays on track toward an eventually useful business solution. This is very difficult if the

managers don't really understand the principles. Managers need to be able to ask probing questions of a data scientist, who often can get lost in technical details. We need to accept that each of us has strengths and weaknesses, and as data science projects span so much of a business, a diverse team is essential. Just as we can't expect a manager necessarily to have deep expertise in data science, we can't expect a data scientist *necessarily* to have deep expertise in business solutions. However, an effective data science team involves collaboration between the two, and each needs to have some understanding of the fundamentals of the other's area of responsibility. Just as it would be a Sisyphean task to manage a data science team where the team had no understanding of the fundamental concepts of business, it likewise is extremely frustrating at best, and often a tremendous waste, for data scientists to struggle under a management that does not understand basic principles of data science.

For example, it is not uncommon for data scientists to struggle under a management that (sometimes vaguely) sees the potential benefit of predictive modeling, but does not have enough appreciation for the process to invest in proper training data or in proper evaluation procedures. Such a company may "succeed" in engineering a model that is predictive enough to produce a viable product or service, but will be at a severe disadvantage to a competitor who invests in doing the data science well.

A solid grounding in the fundamentals of data science has much more far-reaching strategic implications. We know of no systematic scientific study, but broad experience has shown that as executives, managers, and investors increase their exposure to data science projects, they see more and more opportunities in turn. We see extreme cases in companies like Google and Amazon (there is a vast amount of data science underlying web search, as well as Amazon's product recommendations and other offerings). Both of these companies eventually built subsequent products offering "big data" and data-science related services to other firms. Many, possibly most, data-science oriented start-ups use Amazon's cloud storage and processing services for some tasks. Google's "Prediction API" is increasing in sophistication and utility (we don't know how broadly used it is).

Those are extreme cases, but the basic pattern is seen in almost every data-rich firm. Once the data science capability has been developed for one application, other applications throughout the business become obvious. Louis Pasteur famously wrote, "Fortune favors the prepared mind." Modern thinking on creativity focuses on the juxtaposition of a new way of thinking with a mind "saturated" with a particular problem. Working through case studies (either in theory or in practice) of data science applications helps prime the mind to see opportunities and connections to new problems that could benefit from data science.

For example, in the late 1980s and early 1990s, one of the largest phone companies had applied predictive modeling—using the techniques we've described in this book—to the problem of reducing the cost of repairing problems in the telephone network and

to the design of speech recognition systems. With the increased understanding of the use of data science for helping to solve business problems, the firm subsequently applied similar ideas to decisions about how to allocate a massive capital investment to best improve its network, and how to reduce fraud in its burgeoning wireless business. The progression continued. Data science projects for reducing fraud discovered that incorporating features based on social-network connections (via who-calls-whom data) into fraud prediction models improved the ability to discover fraud substantially. In the early 2000s, telecommunications firms produced the first solutions using such social connections to improve marketing—and improve marketing it did, showing huge performance lifts over traditional targeted marketing based on socio-demographic, geographic, and prior purchase data. Next, in telecommunications, such social features were added to models for churn prediction, with equally beneficial results. The ideas diffused to the online advertising industry, and there was a subsequent flurry of development of online advertising based on the incorporation of data on online social connections (at Facebook and at other firms in the online advertising ecosystem).

This progression was driven both by experienced data scientists moving among business problems as well as by data science savvy managers and entrepreneurs, who saw new opportunities for data science advances in the academic and business literature.

Achieving Competitive Advantage with Data Science

Increasingly, firms are considering whether and how they can obtain competitive advantage from their data and/or from their data science capability. This is important strategic thinking that should not be superficial, so let's spend some time digging into it.

Data and data science capability are (complementary) strategic assets. Under what conditions can a firm achieve competitive advantage from such an asset? First of all, the asset has to be valuable to the firm. This seems obvious, but note that the value of an asset to a firm depends on the other strategic decisions that the firm has made. Outside of the context of data science, in the personal computer industry in the 1990s, Dell famously got substantial competitive advantage early over industry leader Compaq from using web-based systems to allow customers to configure computers to their personal needs and liking. Compaq could not get the same value from web-based systems. One main reason was that Dell and Compaq had implemented different strategies: Dell already was a direct-to-customer computer retailer, selling via catalogs; web-based systems held tremendous value given this strategy. Compaq sold computers mainly via retail outlets; web-based systems were not nearly as valuable given this alternative strategy. When Compaq tried to replicate Dell's web-based strategy, it faced a severe backlash from its retailers. The upshot is that the value of the new asset (web-based systems) was dependent on each company's other strategic decisions.

The lesson is that we need to think carefully in the business understanding phase as to how data and data science can provide value in the context of our business strategy, and also whether it would do the same in the context of our competitors' strategies. This can identify both possible opportunities and possible threats. A direct data science analogy of the Dell-Compaq example is Amazon versus Borders. Even very early, Amazon's data on customers' book purchases allowed personalized recommendations to be delivered to customers while they were shopping online. Even if Borders were able to exploit its data on who bought what books, its brick-and-mortar retail strategy did not allow the same seamless delivery of data science-based recommendations.

So, a prerequisite for competetive advantage is that the asset be valuable in the context of our strategy. We've already begun to talk about the second set of criteria: in order to gain competitive advantage, competitors either must not possess the asset, or must not be able to obtain the same value from it. We should think both about the data asset(s) and the data science capability. Do we have a unique data asset? If not, do we have an asset the utilization of which is better aligned with our strategy than with the stragegy of our competitors? Or are we better able to take advantage of the data asset due to our better data science capability?

The flip side of asking about achieving competitive advantage with data and data science is asking whether we are at a competitive disadvantage. It may be that the answers to the previous questions are affirmative for our competitors and not for us. In what follows we will assume that we are looking to achieve competitive advantage, but the arguments apply symmetrically if we are trying to achieve parity with a data-savvy competitor.

Sustaining Competitive Advantage with Data Science

The next question is: even if we can achieve competitive advantage, can we *sustain* it? If our competitors can easily duplicate our assets and capabilities, our advantage may be short-lived. This is an especially critical question if our competitors have greater resources than we do: by adopting our strategy, they may surpass us if they have greater resources.

One strategy for competing based on data science is to plan to always keep one step ahead of the competition: always be investing in new data assets, and always be developing new techniques and capabilities. Such a strategy can provide for an exciting and possibly fast-growing business, but generally few companies are able to execute it. For example, you must have confidence that you have one of the best data science teams, since the effectiveness of data scientists has a huge variance, with the best being much more talented than the average. If you have a great team, you may be willing to bet that you can keep ahead of the competition. We will discuss data science teams more below.

The alternative to always keeping one step ahead of the competition is to achieve sustainable competitive advantage due to a competitor's inability to replicate, or their ele-

vated expense of replicating, the data asset or the data science capability. There are several avenues to such sustainability.

Formidable Historical Advantage

Historical circumstances may have placed our firm in an advantageous position, and it may be too costly for competitors to reach the same position. Amazon again provides an outstanding example. In the "Dotcom Boom" of the 1990s, Amazon was able to sell books below cost, and investors continued to reward the company. This allowed Amazon to amass tremendous data assets (such as massive data on online consumers' buying preferences and online product reviews), which then allowed them to create valuable data-based products (such as recommendations and product ratings). These historical circumstances are gone: it is unlikely today that investors would provide the same level of support to a competitor that was trying to replicate Amazon's data asset by selling books below cost for years on end (not to mention that Amazon has moved far beyond books).

This example also illustrates that the data products themselves can increase the cost to competitors of replicating the data asset. Consumers value the data-driven recommendations and product reviews/ratings that Amazon provides. This creates switching costs: competitors would have to provide extra value to Amazon's customers to entice them to shop elsewhere—either with lower prices or with some other valuable product or service that Amazon does not provide. Thus, when the data acquisition is tied directly to the value provided by the data, the resulting virtuous cycle creates a catch-22 for competitors: competitors need customers in order to acquire the necessary data, but they need the data in order to provide equivalent service to attract the customers.

Entrepreneurs and investors might turn this strategic consideration around: what historical circumstances now exist that may not continue indefinitely, and which may allow me to gain access to or to build a data asset more cheaply than will be possible in the future? Or which will allow me to build a data science team that would be more costly (or impossible) to build in the future?

Unique Intellectual Property

Our firm may have unique intellectual property. Data science intellectual property can include novel techniques for mining the data or for using the results. These might be patented, or they might just be trade secrets. In the former case, a competitor either will be unable to (legally) duplicate the solution, or will have an increased expense of doing so, either by licensing our technology or by developing new technology to avoid infringing on the patent. In the case of a trade secret, it may be that the competitor simply does not know how we have implemented our solution. With data science solutions, the actual mechanism is often hidden; with only the result being visible.

Unique Intangible Collateral Assets

Our competitors may not be able to figure out how to put our solution into practice. With successful data science solutions, the actual source of good performance (for example with effective predictive modeling) may be unclear. The effectiveness of a predictive modeling solution may depend critically on the problem engineering, the attributes created, the combining of different models, and so on. It often is not clear to a competitor how performance is achieved in practice. Even if our algorithms are published in detail, many implementation details may be critical to get a solution that works in the lab to work in production.

Furthermore, success may be based on intangible assets such as a company culture that is particularly suitable to the deployment of data science solutions. For example, a culture that embraces business experimentation and the (rigorous) supporting of claims with data will naturally be an easier place for data science solutions to succeed. Alternatively, if developers are encouraged to understand data science, they are less likely to screw up an otherwise top-quality solution. Recall our maxim: *Your model is not what your data scientists design, it's what your engineers implement.*

Superior Data Scientists

Maybe our data scientists simply are much better than our competitors'. There is a huge variance in the quality and ability of data scientists. Even among well-trained data scientists, it is well accepted within the data science community that certain individuals have the combination of innate creativity, analytical acumen, business sense, and perseverence that enables them to create remarkably better solutions than their peers.

This extreme difference in ability is illustrated by the year-after-year results in the KDD Cup data mining competition. Every year, the top professional society for data scientists, the ACM SIGKDD, holds its annual conference (the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining). Each year the conference holds a data mining competition. Some data scientists love to compete, and there are many competitions. The Netflix competition, discussed in Chapter 12, is one of the most famous, and such competitions have even been turned into a crowd-sourcing business (see Kaggle). The KDD Cup is the granddaddy of data mining competitions and has been held every year since 1997. Why is this relevant? Some of the best data scientists in the world participate in these competitions. Depending on the year and the task, hundreds or thousands of competitors try their hand at solving the problem. If data science talent were evenly distributed, then one would think it unlikely to see the same individuals repeatedly winning the competitions. But that's exactly what we see. There are individuals who have been on winning teams repeatedly, sometimes multiple years in a row and for multiple tasks each year (sometimes the competition has more than

one task).[1] The point is that there is substantial variation in the ability even of the best data scientists, and this is illustrated by the "objective" results of the KDD Cup competitions. The upshot is that because of the large variation in ability, the best data scientists can pick and choose the employment opportunities that suit their desires with respect to salary, culture, advancement opportunities, and so on.

The variation in the quality of data scientists is amplified by the simple fact that top-notch data scientists are in high demand. Anyone can call himself a data scientist, and few companies can really evaluate data scientists well as potential hires. This leads to another catch: you need at least one top-notch data scientist to truly evaluate the quality of prospective hires. Thus, if our company has managed to build a strong data science capability, we have a substantial and sustained advantage over competitors who are having trouble hiring data scientists. Further, top-notch data scientists like to work with other top-notch data scientists, which compounds our advantage.

We also must embrace the fact that data science is in part a craft. Analytical expertise takes time to acquire, and all the great books and video lectures alone will not turn someone into a master. The craft is learned by experience. The most effective learning path resembles that in the classic trades: aspiring data scientists work as apprentices to masters. This could be in a graduate program with a top applications-oriented professor, in a postdoctoral program, or in industry working with one of the best industrial data scientists. At some point the apprentice is skilled enough to become a "journeyman," and will then work more independently on a team or even lead projects of her own. Many high-quality data scientists happily work in this capacity for their careers. Some small subset become masters themselves, because of a combination of their talent at recognizing the potential of new data science opportunities (more on that in a moment) and their mastery of theory and technique. Some of these then take on apprentices. Understanding this learning path can help to focus on hiring efforts, looking for data scientists who have apprenticed with top-notch masters. It also can be used tactically in a less obvious way: if you can hire one master data scientist, top-notch aspiring data scientists may come to apprentice with her.

In addition to all this, a top-notch data scientist needs to have a strong professional network. We don't mean a network in the sense of what one might find in an online professional networking system; an effective data scientist needs to have deep connections to other data scientists throughout the data science community. The reason is simply that the field of data science is immense and there are far too many diverse topics for any individual to master. A top-notch data scientist is a master of some area of technical expertise, and is familiar with many others. (Beware of the "jack-of-all-trades, master of none.") However, we do not want the data scientist's mastery of some area of

1. This is not to say that one should look at the KDD Cup winners as necessarily the best data miners in the world. Many top-notch data scientists have never competed in such a competition; some compete once and then focus their efforts on other things.

technical expertise to turn into the proverbial hammer for which all problems are nails. A top-notch data scientist will pull in the necessary expertise for the problem at hand. This is facilitated tremendously by strong and deep professional contacts. Data scientists call on each other to help in steering them to the right solutions. The better a professional network is, the better will be the solution. And, the best data scientists have the best connections.

Superior Data Science Management

Possibly even more critical to success for data science in business is having good *management* of the data science team. Good data science managers are especially hard to find. They need to understand the fundamentals of data science well, possibly even being competent data scientists themselves. Good data science managers also must possess a set of other abilities that are rare in a single individual:

- They need to truly understand and appreciate the needs of the business. What's more, they should be able to anticipate the needs of the business, so that they can interact with their counterparts in other functional areas to develop ideas for new data science products and services.
- They need to be able to communicate well with and be respected by both "techies" and "suits"; often this means translating data science jargon (which we have tried to minimize in this book) into business jargon, and vice versa.
- They need to coordinate technically complex activities, such as the integration of multiple models or procedures with business constraints and costs. They often need to understand the technical architectures of the business, such as the data systems or production software systems, in order to ensure that the solutions the team produces are actually useful in practice.
- They need to be able to anticipate outcomes of data science projects. As we have discussed, data science is more similar to R&D than to any other business activity. Whether a particular data science project will produce positive results is highly uncertain at the outset, and possibly even well into the project. Elsewhere we discuss how it is important to produce proof-of-concept studies quickly, but neither positive nor negative outcomes of such studies are highly predictive of success or failure of the larger project. They just give guidance to investments in the next cycle of the data mining process (recall Chapter 2). If we look to R&D management for clues about data science management, we find that there is only one reliable predictor of the success of a research project, and it is *highly* predictive: the prior success of the investigator. We see a similar situation with data science projects. There are individuals who seem to have an intuitive sense of which projects will pay off. We do not know of a careful analysis of why this is the case, but experience shows that it is. As with data science competitions, where we see remarkable repeat performances by the same individuals, we also see individuals repeatedly envisioning new data

science opportunities and managing them to great success—and this is particularly impressive as many data science managers never see even one project through to great success.

- They need to do all this within the culture of a particular firm.

Finally, our data science capability may be difficult or expensive for a competitor to duplicate because *we can hire data scientists and data science managers better*. This may be due to our reputation and brand appeal with data scientists—a data scientist may prefer to work for a company known as being friendly to data science and data scientists. Or our firm may have a more subtle appeal. So let's examine in a little more detail what it takes to attract top-notch data scientists.

Attracting and Nurturing Data Scientists and Their Teams

At the beginning of the chapter, we noted that the two most important factors in ensuring that our firm gets the most from its data assets are: (i) the firm's management must think data-analytically, and (ii) the firm's management must create a culture where data science, and data scientists, will thrive. As we mentioned above, there can be a huge difference between the effectiveness of a great data scientist and an average data scientist, and between a great data science team and an individually great data scientist. But how can one confidently engage top-notch data scientists? How can we create great teams?

This is a very difficult question to answer in practice. At the time of this writing, the supply of top-notch data scientists is quite thin, resulting in a very competitive market for them. The best companies at hiring data scientists are the IBMs, Microsofts, and Googles of the world, who clearly demonstrate the value they place in data science via compensation, perks, and/or intangibles, such as one particular factor not to be taken lightly: data scientists like to be around other top-notch data scientists. One might argue that they *need* to be around other top-notch data scientists, not only to enjoy their day-to-day work, but also because the field is vast and the collective mind of a group of data scientists can bring to bear a much broader array of particular solution techniques.

However, just because the market is difficult does not mean all is lost. Many data scientists want to have more individual influence than they would have at a corporate behemoth. Many want more responsibility (and the concomitant experience) with the broader process of producing a data science solution. Some have visions of becoming Chief Scientist for a firm, and understand that the path to Chief Scientist may be better paved with projects in smaller and more varied firms. Some have visions of becoming entrepreneurs, and understand that being an early data scientist for a startup can give them invaluable experience. And some simply will enjoy the thrill of taking part in a fast-growing venture: working in a company growing at 20% or 50% a year is much different from working in a company growing at 5% or 10% a year (or not growing at all).

In all these cases, the firms that have an advantage in hiring are those that create an environment for nurturing data science and data scientists. If you do not have a critical mass of data scientists, be creative. Encourage your data scientists to become part of local data science technical communities and global data science academic communities.

A note on publishing

Science is a social endeavor, and the best data scientists often want to stay engaged in the community by publishing their advances. Firms sometimes have trouble with this idea, feeling that they are "giving away the store" or tipping their hand to competitors by revealing what they are doing. On the other hand, if they do not, they may not be able to hire or retain the very best. Publishing also has some advantages for the firm, such as increased publicity, exposure, external validation of ideas, and so on. There is no clear-cut answer, but the issue needs to be considered carefully. Some firms file patents aggressively on their data science ideas, after which academic publication is natural if the idea is truly novel and important.

A firm's data science presence can be bolstered by engaging academic data scientists. There are several ways of doing this. For those academics interested in practical applications of their work, it may be possible to fund their research programs. Both of your authors, when working in industry, funded academic programs and essentially extended the data science team that was focusing on their problems and interacting. The best arrangement (by our experience) is a combination of data, money, and an interesting business problem; if the project ends up being a portion of the Ph.D. thesis of a student in a top-notch program, the benefit to the firm can far outweigh the cost. Funding a Ph.D. student might cost a firm in the ballpark of $50K/year, which is a fraction of the fully loaded cost of a top data scientist. A key is to have enough understanding of data science to select the right professor—one with the appropriate expertise for the problem at hand.

Another tactic that can be very cost-effective is to take on one or more top-notch data scientists as scientific advisors. If the relationship is structured such that the advisors truly interact on the solutions to problems, firms that do not have the resources or the clout to hire the very best data scientists can substantially increase the quality of the eventual solutions. Such advisors can be data scientists at partner firms, data scientists from firms who share investors or board members, or academics who have some consulting time.

A different tack altogether is to hire a third party to conduct the data science. There are various third-party data science providers, ranging from massive firms specializing in business analytics (such as IBM), to data-science-specific consulting firms (such as Elder

Research), to boutique data science firms who take on a very small number of clients to help them develop their data science capabilities (such as Data Scientists, LLC).[2] You can find a large list of data-science service companies, as well as a wide variety of other data science resources, at KDnuggets. A caveat about engaging data science consulting firms is that their interests are not always well aligned with their customers' interests; this is obvious to seasoned users of consultants, but not to everyone.

Savvy managers employ all of these resources tactically. A chief scientist or empowered manager often can assemble for a project a substantially more powerful and diverse team than most companies can hire.

Examine Data Science Case Studies

Beyond building a solid data science team, how can a manager ensure that her firm is best positioned to take advantage of opportunities for applying data science? Make sure that there is an understanding of and appreciation for the fundamental principles of data science. Empowered employees across the firm often see novel applications.

After gaining command of the fundamental principles of data science, the best way to position oneself for success is to work through many examples of the application of data science to business problems. Read case studies that actually walk through the data mining process. Formulate your own case studies. Actually mining data is helpful, but even more important is working through the connection between the business problem and the possible data science solutions. The more, different problems you work through, the better you will be at naturally seeing and capitalizing on opportunities for bringing to bear the information and knowledge "stored" in the data—often the same problem formulation from one problem can be applied by analogy to another, with only minor changes.

It is important to keep in mind that the examples we have presented in this book were chosen or designed for illustration. In reality, the business and data science team should be prepared for all manner of mess and contraints, and must be flexible in dealing with them. Sometimes there is a wealth of data and data science techniques available to be brought to bear. Other times the situation seems more like the critical scene from the movie *Apollo 13*. In the movie, a malfunction and explosion in the command module leave the astronauts stranded a quarter of a million miles from Earth, with the CO_2 levels rising too rapidly for them to survive the return trip. In a nutshell, because of the constraints placed by what the astronauts have on hand, the engineers have to figure out how to use a large cubic filter in place of a narrower cylindrical filter (to literally put a square peg in a round hole). In the key scene, the head engineer dumps out onto a table all the "stuff" that's there in the command module, and tells his team: "OK, people …

2. Disclaimer: The authors have a relationship with Data Scientists, LLC.

we got to find a way to make *this* fit into the hole for *this*, using nothing but *that*." Real data science problems often seem more like the Apollo 13 situation than a textbook situation.

For example, Perlich et al. (2013) describe a study of just such a case. For targeting consumers with online display advertisements, obtaining an adequate supply of the ideal training data would have been prohibitively expensive. However, data were available at much lower cost from various other distributions and for other target variables. Their very effective solution cobbled together models built from these surrogate data, and "transferred" these models for use on the desired task. The use of these surrogate data allowed them to operate with a substantially reduced investment in data from the ideal (and expensive) training distribution.

Be Ready to Accept Creative Ideas from Any Source

Once different role players understand fundamental principles of data science, creative ideas for new solutions can come from any direction—such as from executives examining potential new lines of business, from directors dealing with profit and loss responsibility, from managers looking critically at a business process, and from line employees with detailed knowledge of exactly how a particular business process functions. Data scientists should be encouraged to interact with employees throughout the business, and part of their performance evaluation should be based on how well they produce ideas for improving the business with data science. Incidentally, doing so can pay off in unintended ways: the data processing skills possessed by data scientists often can be applied in ways that are not so sophisticated but nevertheless can help other employees without those skills. Often a manager may have no idea that particular data can even be obtained—data that might help the manager directly, without sophisticated data science.

Be Ready to Evaluate Proposals for Data Science Projects

Ideas for improving business decisions through data science can come from any direction. Managers, investors, and employees should be able to formulate such ideas clearly, and decision makers should be prepared to evaluate them. Essentially, we need to be able to formulate solid proposals and to evaluate proposals.

The data mining process, described in Chapter 2, provides a framework to direct this. Each stage in the process reveals questions that should be asked both in formulating proposals for projects and in evaluating them:

- Is the business problem well specified? Does the data science solution solve the problem?
- Is it clear how we would evaluate a solution?

- Would we be able see evidence of success before making a huge investment in deployment?
- Does the firm have the data assets it needs? For example, for supervised modeling, are there actually labeled training data? Is the firm ready to invest in the assets it does not have yet?

Appendix A provides a starting list of questions for evaluating data science proposals, organized by the data mining process. Let's walk through an illustrative example. (In Appendix B you will find another example proposal to evaluate, focusing on our running churn problem.)

Example Data Mining Proposal

Your company has an installed user base of 900,000 current users of your Whiz-bang® widget. You now have developed Whiz-bang® 2.0, which has substantially lower operating costs than the original. Ideally, you would like to convert ("migrate") your entire user base over to version 2.0; however, using 2.0 requires that users master the new interface, and there is a serious risk that in attempting to do so, the customers will become frustrated and not convert, become less satisfied with the company, or in the worst case, switch to your competitor's popular Boppo® widget. Marketing has designed a brand-new migration incentive plan, which will cost $250 per selected customer. There is no guarantee that a customer will choose to migrate even if she takes this incentive.

An external firm, Big Red Consulting, is proposing a plan to target customers carefully for Whiz-bang® 2.0, and given your demonstrated fluency with the fundamentals of data science, you are called in to help assess Big Red's proposal. Do Big Red's choices seem correct?

> **Targeted Whiz-bang Customer Migration—prepared by Big Red Consulting, Inc.**
>
> We will develop a predictive model using modern data-mining technology. As discussed in our last meeting, we assume a budget of $5,000,000 for this phase of customer migration; adjusting the plan for other budgets is straightforward. Thus we can target 20,000 customers under this budget. Here is how we will select those customers:
>
> We will use data to build a model of whether or not a customer will migrate given the incentive. The dataset will comprise a set of attributes of customers, such as the number and type of prior customer service interactions, level of usage of the widget, location of the customer, estimated technical sophistication, tenure with the firm, and other loyalty indicators, such as number of other firm products and services in use. The target will be whether or not the customer will migrate to the new widget if he/she is given the incentive. Using these data, we will build a linear regression to estimate the target variable. The model will be evaluated based on its accuracy on these data; in particular, we want to ensure that the accuracy is substantially greater than if we targeted randomly.
>
> To use the model: for each customer we will apply the regression model to estimate the target variable. If the estimate is greater than 0.5, we will predict that the customer will migrate; otherwise, we will say the customer will not migrate. We then will select at ran-

dom 20,000 customers from those predicted to migrate, and these 20,000 will be the recommended targets.

Flaws in the Big Red Proposal

We can use our understanding of the fundamental principles and other basic concepts of data science to identify flaws in the proposal. Appendix A provides a starting guide for reviewing such proposals, with some of the main questions to ask. However, this book as a whole really can be seen as a proposal review guide. Here are some of the most egregious flaws in Big Data's proposal:

Business Understanding

- The target variable definition is imprecise. For example, over what time period must the migration occur? (Chapter 3)
- The formulation of the data mining problem could be better-aligned with the business problem. For example, what if certain customers (or everyone) were likely to migrate anyway (without the incentive)? Then we would be wasting the cost of the incentive in targeting them. (Chapter 2, Chapter 11)

Data Understanding/Data Preparation

- There aren't any labeled training data! This is a brand-new incentive. We should invest some of our budget in obtaining labels for some examples. This can be done by targeting a (randomly) selected subset of customers with the incentive. One also might propose a more sophisticated approach (Chapter 2, Chapter 3, Chapter 11).
- If we are worried about wasting the incentive on customers who are likely to migrate without it, we also should observe a "control group" over the period where we are obtaining training data. This should be easy, since everyone we don't target to gather labels would be a "control" subject. We can build a separate model for migrate or not given no incentive, and combine the models in an expected value framework. (Chapter 11)

Modeling

- Linear regression is not a good choice for modeling a categorical target variable. Rather one should use a classification method, such as tree induction, logistic regression, k-NN, and so on. Even better, why not try a bunch of methods and evaluate them experimentally to see which performs best? (Chapter 2, Chapter 3, Chapter 4, Chapter 5, Chapter 6, Chapter 7, Chapter 8)

Evaluation

- The evaluation shouldn't be on the training data. Some sort of holdout approach should be used (e.g., cross-validation and/or a staged approach as discussed above). (Chapter 5)
- Is there going to be any domain-knowledge validation of the model? What if it is capturing some weirdness of the data collection process? (Chapter 7, Chapter 11, Chapter 14)

Deployment

- The idea of randomly selecting customers with regression scores greater than 0.5 is not well considered. First, it is not clear that a regression score of 0.5 really corresponds to a probability of migration of 0.5. Second, 0.5 is rather arbitrary in any case. Third, since our model is providing a ranking (e.g., by likelihood of migration, or by expected value if we use the more complex formulation), we should use the ranking to guide our targeting: choose the top-ranked candidates, as the budget will allow. (Chapter 2, Chapter 3, Chapter 7, Chapter 8, Chapter 11)

Of course, this is just one example with a particular set of flaws. A different set of concepts may need to be brought to bear for a different proposal that is flawed in other ways.

A Firm's Data Science Maturity

For a firm to realistically plan data science endeavors it should assess, frankly and rationally, its own *maturity* in terms of data science capability. It is beyond the scope of this book to provide a self-assessment guide, but a few words on the topic are important.

Firms vary widely in their data science capabilities along many dimensions. One dimension that is very important for strategic planning is the firm's "maturity," specifically, how systematic and well founded are the processes used to guide the firm's data science projects.[3]

At one end of the maturity spectrum, a firm's data science processes are completely ad hoc. In many firms, the employees engaged in data science and business analytics endeavors have no formal training in these areas, and the managers involved have little understanding of the fundamental principles of data science and data analytic thinking.

3. The reader interested in this notion of the maturity of a firm's capabilities is encouraged to read about the Capability Maturity Model for software engineering, which is the inspiration for this discussion.

A note on "immature" firms

Being "immature" does *not* mean that a firm is destined to failure. It means that success is highly variable and is much more dependent on luck than in a mature firm. Project success will depend upon the heroic efforts by individuals who happen to have a natural acuity for data-analytic thinking. An immature firm may implement not-so-sophisticated data science solutions at a large scale, or may implement sophisticated solutions at a small scale. Rarely, though, will an immature firm implement sophisticated data science solutions at a large scale.

A firm with a medium level of maturity employs well-trained data scientists, as well as business managers and other stakeholders who understand the fundamental principles of data science. Both sides can think clearly about how to solve business problems with data science, and both sides participate in the design and implementation of solutions that directly address the problems of the business.

At the high end of maturity are firms who continually work to improve their data science *processes* (and not just the solutions). Executives at such firms continually challenge the data science team to instill processes that will align their solutions better with the business problems. At the same time they realize that pragmatic trade-offs may favor the choice of a suboptimal solution that can be realized today over a theoretically much better solution that won't be ready until next year. Data scientists at such a firm should have the confidence that when they propose an investments to improve data science processes, their suggestions will be met with open and informed minds. That's not to say that every such request will be approved, but that the proposal will be evaluated on its own merits in the context of the business.

Note: Data science is neither operations nor engineering.

There is some danger in making an analogy to the Capability Maturity Model from software engineering—danger that the analogy will be taken too literally. Trying to apply the same sort of processes that work for software engineering, or worse for manufacturing or operations, will fail for data science. Moreover, misguided attempts to do so will send a firm's best data scientists out the door before the management even knows what happened. The key is to understand *data science processes* and how to data science well, and work to establish consistency and support. Remember that data science is more like R&D than like engineering or manufacturing. As a concrete example, management should consistently make available the resources needed for solid evaluation of data science projects early and often. Sometimes this involves investing in data that would not otherwise have been obtained. Often this involves assigning engineering resources to sup-

> port the data science team. The data science team should in return work to provide management with evaluations that are as well aligned with the actual business problem(s) as possible.

As a concrete example, consider yet again our telecom churn problem and how firms of varying maturity might address it:

- An immature firm will have (hopefully) analytically adept employees implementing ad hoc solutions based on their intuitions about how to manage churn. These may work well or they may not. In an immature firm, it will be difficult for management to evaluate these choices against alternatives, or to determine when they've implemented a nearly optimal solution.
- A firm of medium maturity will have implemented a well-defined framework for testing different alternative solutions. They will test under conditions that mimic as closely as possible the actual business setting—for example, running the latest production data through a testbed platform that compares how different methods "would have done," and considering carefully the costs and benefits involved.
- A very mature organization may have deployed the exact same methods as the medium-maturity firm for identifying the customers with the highest probability of leaving, or even the highest expected loss if they were to churn. They would also be working to implement the processes, and gather the data, necessary to judge also the effect of the incentives and thereby work towards finding those individuals for which the incentives will produce the largest expected increase in value (over not giving the incentive). Such a firm may also be working to integrate such a procedure into an experimentation and/or optimization framework for assessing different offers or different parameters (like the level of discount) to a given offer.

A frank self-assessment of data science maturity is difficult, but it is essential to getting the best out of one's current capabilities, and to improving one's capabilities.