CHAPTER FOURTEEN:
DATA MINING ETHICS

WHY DATA MINING ETHICS?

It has been said that when you are teaching someone something, you should leave the thing that you want them to remember most to the very end. It will be the last thing they remember hearing from you, the thing they take with them as they depart from your instruction. It is in harmony with this philosophy that the chapter on data mining ethics has been left to the end of this book. Please don't misconstrue this chapter's placement as an afterthought. It is here at the end so you'll take it with you and remember it. It is believed that especially if you make a big deal out of it, the last thing you share with your audience will end up being what they remember from your teaching, so here is our effort at making a big deal about data mining ethics:

FIGURE 14-1. This Just In:

<u>BEING AN ETHICAL DATA MINER IS IMPORTANT</u>

In all seriousness, when we are dealing with data, those data represent peoples' lives. In this book alone, we have touched on peoples' buying behaviors, ownership of creative works, and even serious health issues. Imagine the ethical ramifications of using a decision tree to predict the risk levels of juvenile delinquents as just one example. You'd be profiling, potentially branding these youth, so any attempt to do so must be done ethically. But what does this mean? **Ethics** is the set of moral codes, above and beyond the legally required minimums, that an individual uses to make right and respectful decisions. When mining data, questions of an ethical nature will invariably arise. Simply because it is legal to gather and mine certain data does not make it ethical.

Because of these serious matters, there are some in the world who fear, shun and even fight against data mining. These types of reactions have led some data mining advocates and leaders to respond with attempts to defend and explain data mining technologies. One such response came in the year 2003. The Association for Computing Machinery (ACM) is the world's foremost professional organization for computing professionals in all disciplines. This includes the ACM Special Interest Group for Knowledge Discovery and Data Mining (SIGKDD). At that time, a number of criticisms and calls against data mining were occurring, mostly driven by concerns over citizens' privacy as the United States government increased its use of data mining in anti-terrorism activities in the years following the September 11th terrorist attacks. Certainly any time a government increases its scrutiny of its own citizens and those of other countries, it can be unsettling; however the leaders of ACM SIGKDD were likewise unsettled by the blame being placed on data mining itself. These leaders felt that the tool should be separated from the way it was being used. In response, the executive committee of ACM SIGKDD, a group that included such pioneers as Gregory Piatetsky-Shapiro, Usama Fayyad, Jiawei Han, and others, penned an open letter titled "*Data Mining" is NOT Against Civil Liberties*. (The two-page text of their letter is easily available on the Internet and you are encouraged to read and consider it). Their objective in writing this letter was not to defend government, or any data mining programs, but rather, to help individuals see that there is a large difference between a technology and the choices people make in the ways they use that technology.

In truth, every technology will have its detractors. It may seem a silly example, but consider a chair as a technology. It is a tool, invented by mankind to serve a purpose: sitting. If it is ergonomically designed and made of the right materials, it can facilitate very comfortable sitting. If it is fancy enough, it may exclude certain socio-economic classes from being able to own it. If pointed into a corner and associated with misbehavior, it becomes an object for punishment. If equipped with restraining straps and voltage high enough to take someone's life, it becomes a politicized object of controversy. If picked up and used to strike another person it becomes a weapon; and yet, it is still a chair. So it is with essentially all technologies—all tools invented by mankind to do work. It is not the tool, but the choices we make in how to use it, that create and answer the questions of ethics.

This is not a simple proposition. Every one of us have a different moral compass. Each is guided by a different set of values and influenced by a unique set of backgrounds, experiences and forces. No one set of ethical guidelines is completely right or completely wrong. However, there are ways for each of us to reflectively evaluate, at least for our own, and hopefully for our organizations' purposes, what our ethical parameters will be for each data mining activity we undertake. In order to aid in this process, we offer here a series of…

ETHICAL FRAMEWORKS AND SUGGESTIONS

- The brilliant legal scholar Lawrence Lessig has offered four mechanisms whereby we can frame and contain computing activities within reasonable bounds. These are:
  - **Laws**: These are statutes, enacted by a government and enforced by the same. If these are violated, they carry with them a prescribed punishment, adjudicated by a judge or a jury. Adherence to laws as a mechanism for right behavior represents the basest form of ethical decision making, because at this level, a person is merely doing what they have to do to stay out of trouble. Lessig suggests that while we often look to laws first as a method to enforce good behavior, there are other more reasonable and perhaps more effective methods.
  - **Markets**: Here Lessig suggests an economic solution to guiding behavior. If bad behavior is not profitable or would not enable an organization to stay in business, then bad behavior will not be prevalent. There are many ways that market forces, such as a good reputation for high quality products, excellent customer service, reliability, etc., can help guide good actions.
  - **Code**: In computing disciplines, code is a powerful guide of behavior, because it can be written to allow some actions while stopping others. If we feel that although it would not be illegal for members of a web site to access one another's accounts,

but that it would be unethical, we can write code to require usernames and passwords, making it more difficult for users to get into each other's personal information. Further, we can write a code of conduct, usually referred to as an Acceptable Use Policy, which dictates what users can and cannot do. The policy is not a law, that is, it is not enacted or enforced by a government, but it is an agreement to abide by certain rules or risk losing the privilege of using the site's services.

- **Social Norms**: This form of determining what is ethical is based on what is acceptable in our society. As we look around us, interact with our friends, family, neighbors, and associates, ethical bounds can be established by what is acceptable to these people. Often, if we would be embarrassed, humiliated or otherwise shamed by our behavior, if we find ourselves wanting to hide what we're doing from others, we have a strong indication that our activity is not ethical. We can also contribute to the establishment of social norms as ethical guides by making our own expectations of what is acceptable clear to others.

- **Organizational Standard Operating Procedures**: Ethical standards can often be established by creating a set of acceptable practices for your organization. Such an effort should be undertaken by company leadership, with input from a broad cross-section of employees. These should be well-documented and communicated to employees, and reviewed regularly. Checks and balances can be built into work processes to help ensure that workers are adhering to established procedures.

- **Professional Code of Conduct**: Similar to organizational operating standards, professional codes of conduct can help to establish boundaries of ethical conduct. The aforementioned Association for Computing Machinery maintains a Code of Ethics and Professional Conduct that is an excellent resource for computing professionals seeking guidance (http://www.acm.org/about/code-of-ethics). Other organizations also have codes of conduct that could be consulted in order to frame ethical decision making in data mining.

- **Immanuel Kant's Categorical Imperative**: Immanuel Kant was a German philosopher and anthropologist who lived in the 1700's. Among his extensive writings on ethical morality, Kant's Categorical Imperative is perhaps his most famous. This maxim states that if a given action cannot ethically be taken by anyone in a certain situation, then it should not be taken at all. In data mining, we could use this philosophy to determine: Would it be ethical for any business to collect and mine these data? What would be the outcome if every business mined data in this way? If the answers to such questions are negative and appear to be unethical, then we should not undertake the data mining project either.

- **Rene Descartes' Rule of Change**: Rene Descartes was a French philosopher and mathematician who like Kant, wrote extensively about moral decision making. His rule of change reflects his mathematical background. It states that if an act cannot be taken repeatedly, it is not ethical to do that act even once. Again to apply this to data mining, we can ask: Can I collect and mine these data on an ongoing basis without causing problems for myself, my organization, our customers or others? If you cannot do it repeatedly, according to Decartes, then you shouldn't do it at all.

There are a few other ways that are not quite as specifically defined that you can use to seek out ethical boundaries. There is the old adage known as the Golden Rule, which dictates that we should treat others the way we hope they would treat us. There are also philosophies that help us to consider how our actions might be perceived by others and how they might make them feel. Some ethical frameworks are built around actions that will bring the greatest good to the largest number of people.

CONCLUSION

We can protect privacy by aggregating data, anonymizing observations through removal of names and personally identifiable information, and by storing it in secure and protected environments. When you are busy working with numbers, attributes and observations, it can be easy to forget about the people behind the data. We should be cautious when data mining models might brand a person as a certain risk. Be sensitive to peoples' feelings and rights. When appropriate, ask for the their permission to gather and use data about them. Don't rationalize a justification for your data mining project—ensure that you're doing fair and just work that will help and benefit others.

Regardless of the mechanism you use to determine your ethical boundaries, our hope is that you will always keep ethical behavior in mind when mining data. Remember the personal side of what you are doing. As we began this book, we talked about the desire to introduce the subject of data mining to a new, non-traditional audience. We hope you are gaining confidence in your data mining skills, and that your creativity is helping you to envision your own data mining solutions to real-world problems you might be facing. Go exploring both within RapidMiner and through other tools for ways to find unexpected and interesting patterns in your data. The purpose of this book from the outset was to be a beginner's guide, a way to get started in data mining—even if you don't have a background in computer science or data analysis. Hopefully through the chapter examples and exercises, you've learned a lot and are well on your way to becoming an accomplished, and ethical, data miner. You've learned enough to be dangerous…don't be. Apply what you've learned to use data as a powerful and beneficial advantage. And so as we close this book, let us do so as we began it: Let's start digging!