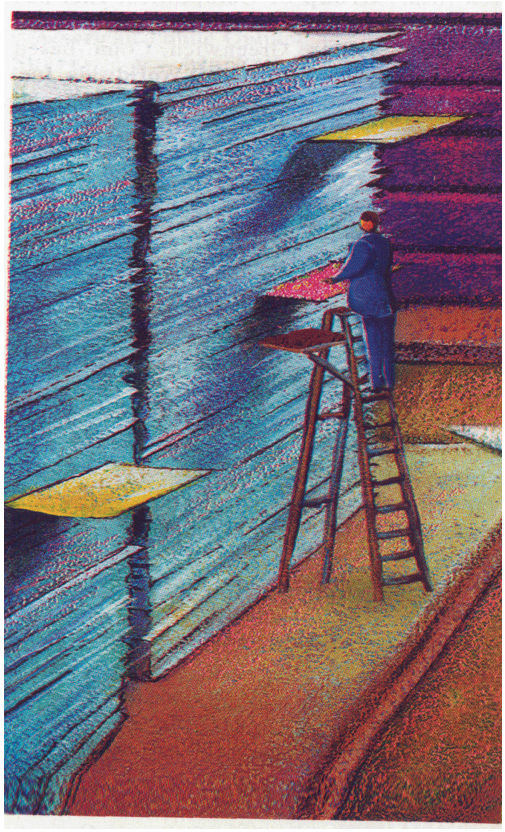


Data Mining Fact and Fiction

A discipline comes of age, more or less

By Steven Struhl, (July 2007)

Overview

In its early days, what was called "data mining" had little form or substance. Respectable researchers looked down upon it from great heights. At advanced research forums, it had as much appeal as rare roast beef at a vegetarian's convention. "Old" data mining often amounted to no more than going on fishing expeditions in data, looking for whatever might be found, with inappropriate, or no, statistical rigor—and fleeting reference to broader corporate and informational goals.

New data mining, however, is an organized, rigorous set of data analysis procedures, and more importantly, a systematic process for uncovering useful insights. Methods and approaches have been honed to distinguish between real, robust findings and the mere pseudo-findings that the "old" data mining all too often produced.

We will start with a brief discussion of the bad old days, including some reasons that problems started. Following this, we will discuss a few once-prevalent fictions, and set them straight. These are the common errors that we hope to address:

- Data mining always is awfully complex
- Dig enough and you will find something
- You have to look at mountains of data
- Data mining should be automatic
- Data mining is all about predictive accuracy
- Data mining is all about advanced algorithms

We will continue with a very brief section on what's new in methods, since no self-respecting paper on data mining could ever go without this. The concluding summary

will review what we have learned that makes data mining work better

Data mining: How the trouble began

Going back to the mid-1990s, searching for a good definition meant hours of frustrating and fruitless endeavor. Carefully reviewing the extensive literature on the subject led to one inescapable conclusion: We could say little definitive about what data mining meant.

This initial lack of clarity likely arose because "data mining" attracted wide attention, going well beyond the community of unfortunate souls who actually get their hands dirty with data.

Masses of material originated in the various "information technology" communities. People connected with IT then, as now, were a diverse and considerable force, including hardware mavens, networking experts, Internet wizards, content masters, and various groups boasting that they wore black hats.

Trailing close behind were legions of software vendors, nearly fainting at the prospect of selling plenty of "enterprise class" applications for data mining. As a reminder, "enterprise class" software is a special industry code-phrase for something costing between 10 times and 100,000,000 times whatever you pay for lowly "desktop" software.

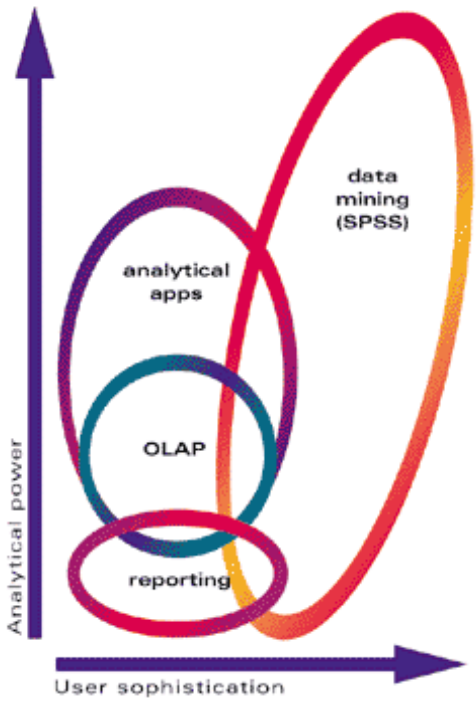
All parties generated a great deal of material (if that's the right term), in print and of course especially on the Internet. As we know, we can find postings of highly variable quality online, as the ability to post is the only qualification.

Actual published definitions

Just a few years back, non-useful definitions of *data mining* ran rampant. Perhaps this was the champion for brevity among them: data mining "uses statistical algorithms to discover patterns in data." (Unfortunately, this appeared in many places, so attribution to any source is impossible). Other expert sources considered the situation carefully, and then made sure to add that data mining discovers "useful patterns" in data. This certainly will relieve all of you who suspected that data mining intended to capture useless patterns.

What did the experts actually say when they spoke in more depth? As just one example, we could find this definition prominently displayed on the SPSS Web site:

> Data mining is a 'knowledge discovery process of extracting previously unknown, actionable information from very large databases.'
>
> —The META Group



Now, we have no intention of singling out SPSS, the META Group, or any of its employees for criticism, since nobody else was saying anything much more prescient. Their definition, though, is a good example of a statement that raises more questions than it answers.

- For instance, they talk about extracting "previously unknown information." Is this in opposition to data that are *(already) known*?
- Also, they specifically identify these data as "actionable." Does this

imply that other methods look for data that are *pointless*?

- And why "very large" databases? What would you call the same activities if the database were merely *large*—or medium, or worse, small?

Here is just one more quote, showing the variety of expression, if not content, in discussions:

> The process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.
>
> —The Gartner Group

We see the same lapses in logic, with a new unnecessary condition (that this applies specially to data "stored in repositories"). Beyond this we must ask, "What does this add to the earlier definition?" Admittedly, it does redundantly use, employ, or utilize, largely synonymous terms that also mean mostly the same thing—and so are largely equal. But in terms of the new, we unfortunately see not much, if anything, or a minimal if not a vanishing amount—and little, or nothing, in the bargain.

Early days: Overly simple concepts give rise to damaging fiction

As the quotes illustrate, prevailing views of data mining were very much like the diagram to the right. That is, it more or less resembled a large blob:

- Not exactly data analysis,
- Not exactly reporting,
- Not even the magic of OLAP cubes.

Again, this is not meant to pick on SPSS (or even OLAP cubes, although perhaps you have not heard much about these recently, as their perceived utility seems to have waned somewhat). As we will see, SPSS was bold enough to take a highly visible role then, and it now remains visible while speaking much more clearly and sensibly.

Much activity in all quarters was unfocused or aimless. As in the illustration, claims ran rampant that data mining could not exist without:

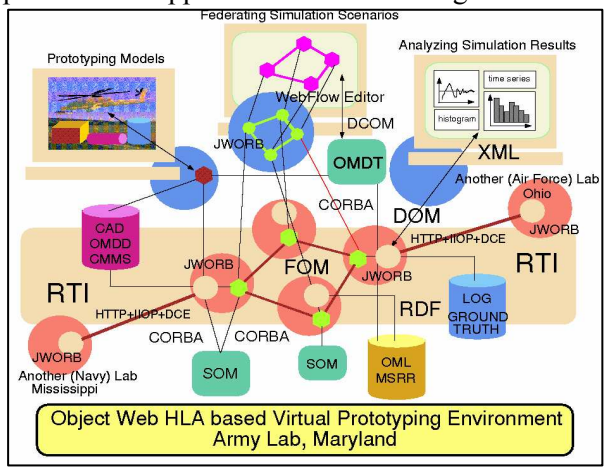
- Heavy analytical power, and
- Extreme software user sophistication.

Self-identified experts sometimes called it an "art" or worse, "a black art." And so, because data mining lacked a sound foundation, it accumulated a kind of bad "mythology."

These fictions still hurt many efforts today. Let's look at some of these fictions and their effects.

Fiction: Data mining is awfully complex

This has frustrated efforts going all the way back. It needs to vanish quickly, but is going slowly. For instance, the following recent diagram intends to *explain* data mining processes. It appears to have something to



do with defective soccer balls.

What could anybody expect to emerge from this process? How could anybody apply this to a real problem?

The mantle of excessive complexity often settles on fields where useful results are scarce. This can help the suppliers. When nothing emerges, the disappointed client sometimes can be led to believe that they just weren't smart enough to see whatever they needed to.

Emerging fact: Complexity should be weighed against usefulness

The facts are slowly chipping away at this long-held misconception. We can now find growing recognition that complexity does not ensure the best answer.

Some newer methods even balance the complexity of explanations vs. gains in accuracy. The cost of making each extra description is balanced against the added value of whatever extra information it imparts. For the statisticians in the audience, this of course is not the same as significance. Especially with large data sets, large models can emerge that have little value in their details. Commonly used analytical methods even will produce exquisitely detailed models when given enough completely random data.[1]

Perhaps more surprisingly, we find that small models often work very nicely. The seminal article that pointed this out was written by Holte in 1993 ("Very simple classification rules perform well on most commonly used datasets.")[2]

This article was written some time ago, and appears to be well known by some serious people in the data mining community. However, data mining fiction, like other fiction, enjoy wide circulation among the vast group who remain less involved and less informed.

As importantly, we see a rising awareness about the danger of overly complex algorithms. These tend not to get applied in many organizational settings. Indeed many organizations are restricted in what they use by the state of their software or hardware. This sometimes dismays technology experts, who may find a lot of complexity really interesting.[3]

However, data mining works well only when driven by domain expertise, not just numerical savvy. It almost inevitably fails in strategic applications if it cannot use input from non-technical professionals—and involve these people in putting the results to use.

Fiction: Dig enough and you will find something

Here we find perhaps the most harmful fallacy: Meaning must reside in the data, waiting to be unearthed!

You can't just dig and find rewards.

The first problem is that substantial effort must go into preparing data for analysis.

- *Typically between 60% and 95% of project time is required just to get the data into shape*.

"Free-form data," in particular, is full of inconsistencies, mysterious values, data outside acceptable ranges, missing values, and misunderstood coding. (You can take "free-form" to mean anything other than a survey you personally designed and oversaw.)

Cleaning and assembling data requires much more than meets the eye. The range and scope of problems sometimes surprise even the most experienced practitioners.

For instance, if you have different regions, they may have used different coding; areas have may been exceedingly lax, or worse, just not collected data. Different offices within a region can prove to be inconsistent, and sometimes even different individuals in the same department have their own "special" systems. Also, it is hard to overestimate the havoc that just one grossly incompetent data entry person can wreak.

Sometimes simple efforts may not have been taken to clean data. For instance, that "request for address correction" you sometimes sees on mail actually serves an important function, but companies fail to use it because it costs more than regular postage.

This function is to let the mailing company know whether the intended recipient is still at the address, and perhaps as importantly, if the address was entered correctly. So smarter companies use the returns for "cleaning" errors from their important databases.

Yet not too long ago, we actually encountered a *billing* database with enough address errors to fill the phone book of Eaglebutte, ND. For instance, we learned there were actually 67 states in the United States, including one called "7." (We believe the company might have had trouble collecting payments from residents there.) Perhaps the coup de grace was that the company did business only in eight states.

Worse, some corporate systems even encourage hiding data. For instance, we have encountered more than one instance where business account managers kept the sales records of good customers to themselves, and turned worthless ones over to the central corporate database. (That makes the good customers much more portable, and so more valuable to the manager when changing jobs.)

One major financial institution commissioned a survey among their corporate customers, only to discover that 60% of the contacts listed were no good, and 25% were, in fact, dead. (Although this may strain the reader's credulity, the database management team was headed by—we are not making it up—a man named John Schmuck.)

Reality: Applying knowledge of the domain and working hard are critical

Early data mining routinely failed because it did not have the required expertise behind it, or lacked the needed constant focus on organizational and analytical objectives. Clearly defined business and analytical goals, *from the outset,* are critical, as are clearly stated plans for using the findings.

Nothing is simpler than getting lost indefinitely in ad hoc explorations of the "interesting to know." This is the stuff data miner's nightmares are made of, and they have given it appropriately dismissive names:

- Data fishing,
- Data snooping, and (most accurately)
- Data dredging.

You cannot even assemble and clean up data without substantial knowledge of the domain. Many errors obvious to those with requisite knowledge just look like more data to others. For instance, to those unfamiliar with diabetes, what's wrong with an HbA1c reading of 105? (Hint: this is average blood sugar over a period of about six weeks; diabetics have elevated blood sugar and so higher readings—and risk organ and nerve damage from too much sugar. Normal values are below 7, while 10 is considered poor sugar control; over 15 shows serious illness, or worse.)

Experts in the methods cannot, for instance, tell you:

- "Only these responses make sense here"
- "Customers can have only one record of data"
- "These codes belong here," etc., etc.

Only a person with domain expertise can tell if the values in the data set make sense, and if key patterns of relationships shown make sense.

Fiction: You have to look at mountains of data

A myth arose early that you always needed huge amounts of data, if not all possible data. Indeed, you may remember that experts then typically singled out data mining as applying to "very large" data sets. Yet earlier efforts often failed when analyzing huge data sets, because these could run far beyond the capabilities of computers at hand.

Analysis time expanded often drastically. For many methods, time required was found to expand in proportion with the number of pieces of information *squared*, or to the third power, or worse. Unsurprisingly, programs broke down, memories overflowed, and sometimes frazzled computers even ruined the data along the way. (Your writer can attest to seeing an old computer spitting out a whole, tall stack of data "cards," in numerous directions, when it encountered a problem it found uncongenial. Experienced

analysts knew to write a sequence number in the corner of each computer card with a ball-point pen.)

Very large data sets also make specious relationships look very significant. Even the weakest effects can look very solid with large enough samples, for instance:

- People with the astrological sign Leo watch more TV than others;[4]
- Hard enough searching for obscure sequences has found all sorts of hidden "messages" in the Bible;
- The same methods yield equally as many "missives" in *War and Peace* or the Microsoft license agreement.[5]

Reality: Too much can be too much

Hugeness *may* be required—but not as a rule. It may even not be beneficial in many cases. Going through all of a huge data set, or perhaps just a few boxcars of data, seems to involve a strange unstated assumption. That is, if you somehow handle every piece of data you can, this will improve your analysis.

Here a rather bald evaluation may be best. Anybody who believes that manipulating tons of data will make you smarter has fallen for rank nonsense.

The chief proponents of this approach seem to be people with no understanding of sampling or statistical methods. Activity of this wasteful nature would never be tolerated where actual, physical work was needed, since squandering resources there can have significant costs.

For instance, think about actual mining, as in digging for minerals. Even though the equipment there also has gotten tremendously powerful over the years, nobody does real mining by tearing an entire mountain (or county, or country) to bits, and then sifting through the debris.

Rather, actual miners do careful testing of selected regions, find promising areas, and then dig further. They use specialized tools and methods to determine the possible worth of an area before setting up the heavy equipment, and constantly monitor to see if they are still following a worthwhile lead.

Anybody doing anything else would be considered, at best, foolish. One unfortunate aspect of "data mining" is that it manipulates only bits (or bytes). Because of this, the unwise have a far easier time running the really heavy analytical equipment, and slowly reducing huge masses of data to rubble.

Reality: too much could cost too much

This is not to say that massive efforts never are warranted. Huge projects done for the right reasons can pay off tremendously. For instance, Google has done pretty well mining the whole Web. Wal-Mart has attained almost mythic status from vigorously mining its 20,000,000+ daily transactions.

There is even an urban legend, often attributed to Wal-Mart, about data mining, in which beer and diaper sales were found to rise together before weekends. The story goes that some genius realized recent dads shop before the weekend; that is, while they are stocking up on Pampers, they just stop by and pick up a few sixes of beer. In a further stroke of genius, the stores then put beer promotions near the diaper aisle, or maybe vice versa (it makes more sense to think that really heavy beer drinkers might find the diapers handy). Sometimes the story is embellished with a set of related factoids, such as saying sales went up some specific amount, like 27%.

This story is absolutely untrue, which we can tell not only because it is also attributed to 7-11, and "a major supermarket chain." The person who started the story has been located. He said it all was only a joke.

One critic points out that no strategy could be based on assembling such factoids, and that all this story proves is that data mining is still in the diaper stage. [6] It is sort of depressing to see how many data mining books solemnly report this legend as if it were a solid fact

Monumental mining efforts even may be self-defeating for those who are not Google or Wal-Mart. Manipulating huge datasets requires massive hardware investments. It may require much more staffing than imagined to gather and assemble the data. Finally, it could also force you into using "enterprise-class software." (Recall that this is defined as software with 5 to 99 zeroes in its price.)

At best, moving heaps of data chews up huge amounts of time. Even though computers are much faster, and computer memory can expand to levels of vastness almost unimaginable just several years ago, there are limits. In line with an old adage, a few hundred terabytes here and a few hundred terabytes there, and soon you're talking about real numbers.

The goal, then, is not to move the most data, or even move data most quickly. Data mining tools that organizations choose should strive to optimize the user's time, not to optimize processing.

Reality: Focusing can work for you

Data mining has worked really well with data sets far smaller than the "very large." In fact, the approaches and methods work just as well on large, medium and even small data sets. (Here "small" is taken to mean hundreds of records.) Focusing on what you really need from the data, instead of wallowing in all the data, can reduce terabytes of data to files of reasonable sizes.

Here's an example: The client has a million customers and a 20% annual attrition ("churn") rate. Do we need to plot graphs and build models using all million examples, or even 500,000? Consider the following questions, with answers from domain experts:

- Q: How many different "churn profiles" do we expect to find?
- A: No more than ten.
- Q: What is the largest number of examples of each profile we need?
- A: Maybe a thousand.

Therefore, a sample of 10,000-20,000 churners and as many non-churners likely will suffice for this analysis. The terabytes have disappeared![7]

Fiction: Data mining should be automatic

A really harmful early fiction was that machines should do mining themselves. Some thought that smart enough algorithms would let the computer do evaluative work. Others wondered about the wisdom machines making unsupervised judgments in novel situations.

Machines have been more challenging to teach than many imagined, and difficult in ways hardly conceived. As a broad rule, machines have a much easier time generating output than processing the uncertainties of the real world. (Remember those old science fiction movies where the computer completely understood casual conversation, but then spoke haltingly in a metallic voice? Now you can get a $20 answering machine with a voice like a radio announcer—admittedly one over a small radio—and exactly where are all those voice activated robots?)

Many attempts at automation foundered precisely because machines need rules and can falter on the irregular and unexpected. Programs stumbled as they seized on nearly invisible anomalies in the data. Without the right guidelines in place, they seized on obviously wrong variables. Many highly "futuristic" programs, like early neural networks (hyped heavily as learning like human brains), could not communicate what they were doing. In the worst cases, "predictions" were put into action and problems became apparent after results emerged as completely and obviously wrong.

Reality: Automation works in specific circumstances, but is wrong in many others

Machines do well automatically detecting unusual behaviors in specific areas. For instance, this is why the new AT&T can disable your cell phone number so quickly when numerous calls start going to Slovenia or Nigeria.

However, these machine making these decisions are carefully trained by experts who sweated mightily over what distinguishes regular from irregular. The "skill" of the machine reflects how well programmers have captured the rules by which experts would detect unusual usage or fraud.

For data mining to work otherwise, it needs to be a *process* with many elements, namely:

- Formulating corporate or business goals,
- Mapping these goals to data mining goals,
- Acquiring, understanding and preprocessing the data,
- Evaluating and presenting the results of analyses, and
- Using these results to achieve desired goals.

Fiction: Data mining is all about predictive accuracy

One rampant error dating from the early days: "Predictive" accuracy is enough, even if causes are not understood.

Let's take a look at a few well-documented problems with this approach.

- Football results "predicted" stock market performance well from 1955 to 1977: 17 NFC victories predicted a market rise and 5 AFC losses a market decline.
- Elizabeth Taylor did well over an even longer time : Her marriages in 1951, 1953, 1958, 1960, 1965, 1976 and 1977 all coincided with strong stock market gains;
- Best of all: The GDP-adjusted "S & P" Index nearly coincides with the number of 45-50 year olds in the country. Over the period 1946-1997, their correlation is 0.927.[8]

Prediction without understanding typically falls into disuse—and may be just as correct as these examples.

Fiction: Data mining is all about advanced algorithms

Many proponents of data mining still indulge in this fallacy. At any typical data mining conference, you would bet all involved believe that data mining is all about advanced analytical algorithms. Further, the better the algorithms, the better the data mining. Of course, as this implies: increasing the effectiveness of data mining means advancing our knowledge of algorithms.

A subtext seems to underlie this. All we need to do is keep looking and that magic method will appear. If anything, this fiction may well be gaining ascendancy. We are in the midst of a golden age of seeking instant answers. Those promising instant answers, and especially those promising that you need not think about these answers, are gathering very large followings.

Reality: Methods are much better now, but are not enough by themselves

This does not minimize the importance of new or improved data mining algorithms. In fact, new and interesting methods are arriving at a remarkable rate. The pace of innovation itself may be part of the problem—it is hard just to keep up with advances in the field. We will get to just a few of the shiny new approaches and methods next, since no self-respecting paper on this topic could avoid them entirely.

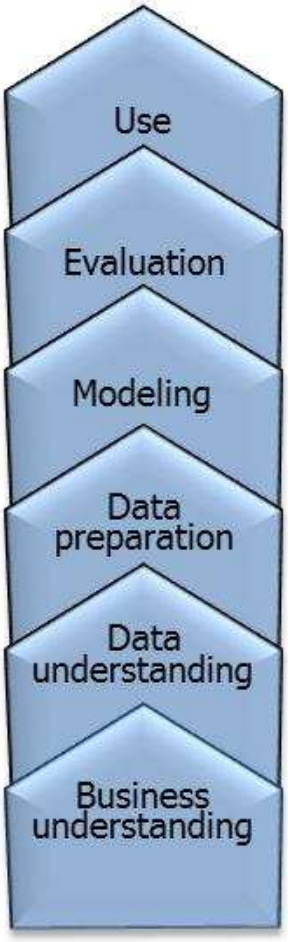
Still, having so many shining new methods for the analyst really does not pose the most difficulty. As one writer puts it: "The problem occurs when data miners focus…on the algorithms and ignore the other 90—95% of the data mining process." [9]

For the last several years, we have had standards for this other 95% of activities, developed by a consortium called CRISP-DM. (This stands for *Cr*oss *I*ndustry

*S*tandard for *D*ata *M*ining. Few poetry majors go into data mining.)

The CRISP-DM approach is large and multifaceted. It came about as many established professionals, scattered across many fields (hence the "cross-industry" part of the acronym) grappled with the fact that so much data mining went forward without any plan or conceptual framework. The standards document is readily available and reasonable to read. As instructive as the descriptions of the processes described, at least to this reader, is the almost palpable sense of frustration the document conveys about the confusion and formlessness of early data mining.

Perhaps to make sure no confusion could ever arise again, the document deals with each step in great detail. Here is a broad overview of the steps:



What's new in methods

Over 100 new and often useful methods have developed or had major overhauls since 2000. Data mining is now intermingled with the field of "machine learning." This discipline has its own vocabulary and encompasses approaches that can become mind-bendingly difficult even for those thoroughly versed in traditional statistics. We will just touch on a few highlights (or more accurately, a glimpse at the highlights, as just skimming the key methods fills a 500+ page book[10]).

It is remarkable, at least to your author, that we have had such a torrent of new approaches, after so many years of promises that delivered little good and new. Not all of these involve true "learning," in the sense that the computer does a sequence of actions, with the later ones changing based on the results of the earlier ones. However, many of them do involve taking many "passes" through the data. The idea—which works—is that looking at all these different "cuts" of the data greatly reduces the effects of any anomalous values on the results.

Perhaps one of the most amazing findings from machine learning is that a number of weak models can be put together (either by averaging or voting) and the results typically are better than any of them.

One approach takes a lot of samples drawn at random that overlap each other (we will spare you the precise details). This goes by the unfortunate term "bagging" (which comes from "bootstrap aggregating," and shows again that computer experts find it hard to resist the lure of a truly ugly name). Still, "bagging" often improves results sharply over taking just one pass through the data.

Related methods build many "classification tree" (CHAID or CART) models, taking many random samples of either observations (e.g., respondents) or variables. These are called random forests and random trees, and the power these have in prediction can be amazing, at times nearing 100%. These approaches, for instance, have successfully

taught computers how to recognize handwritten numbers, even when the handwriting is terrible. In the examples shown, the computers did as well as your author in guessing what people were trying to write.

Other approaches run models several or many times, and use the results from earlier models to shape the later ones. This is called "boosting" and one particularly useful approach is called AdaBoost (short for "adaptive boosting"). AdaBoost can use many basic methods, often strongly improving performance from what we would get running a model just once.

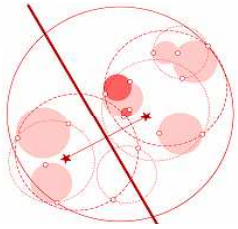
Unfortunately, many of the models that emerge from machine learning methods are highly complex. Some are so large that only a computer can understand them fully. Complex models pose their own problems. One of them is that all the parts they contain can "over-fit" the data you are analyzing. That is, with a very detailed model, you can end up fitting irregularities that are specific to the set of data you have at hand, but that do not appear elsewhere in the outside world.

Machine learning now routinely includes a procedure called "cross-fold" validation to reduce the problem of over-fitting. The entire data set is broken into "folds," usually 10, each of which is a sample containing most of the data in the entire set. The basic model is built using one of the folds, and then tested on the other nine. The answer that emerges averages results from all the folds not used to make the model.

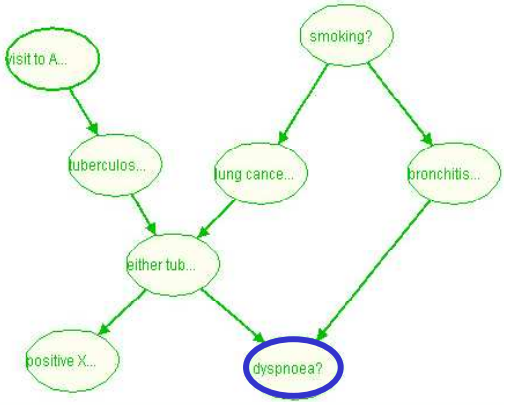
Traditional statisticians will find this approach familiar, since it is similar to splitting the sample and "holding out" part of it, so that the model developed on the "main" part can be tested with data that was not used to develop the model. However, cross-fold validation uses the data much more efficiently, and so requires much smaller samples than the long-established method.

Machine learning also includes new method for looking at similarities and patterns. In some of them, we can see influences from information theory, for instance, balancing the effort needed to describe more details vs. the gains in accuracy from the added description. This is an important concept with very large samples. As you get to hundreds of thousands, or millions, of cases, even minuscule differences can appear terrifically significant—and so we need methods like this, that go beyond traditional statistical tests to determine what truly matters.

Some approaches show strong influences from graphically oriented forms of analysis—as interpreted by computers. For instance, the diagram to the left, called a "ball tree" (another strikingly inelegant name), actually summarizes a highly efficient set of computations for getting people (or items) into groups. It is truly strange looking, and indeed reflects a type of thinking most of us would find alien—but it works. We hope readers can contain their disappointment if we skip the details.



Other methods showing very strong promise require excursions in other directions away from traditional methods. For instance, "Bayesian networks" show how a set of predictor (or independent) variables relate to



Bayesian network
(dependent variable highlighted)

each other and a dependent variable. The diagrams they produce explain which variables can best be seen as having a direct effect on the dependent variables, which have indirect effects, and which relate mostly to each other.

These look something like the structural equation models that may be familiar to some readers. Structural equations require painstaking analysis from highly advanced users. But Bayesian networks build and organize themselves. That is, they allow the data to determine the relationships and require only common-sense judgments from the data analyst.

We could go over many other innovations—but that's another paper.

The experts speak: wiser now

You likely recall that early definitions of data mining lacked precision and conceptual rigor. While this has not disappeared, many new definitions show signs of hard-won insight. For instance, look at this example:

> [Data mining is] a hot buzzword for a class of database applications that look for hidden patterns in a group of data...commonly misused to describe software that presents data in new ways.
>
> True data mining software doesn't just change the presentation, but actually discovers previously unknown relationships among the data
>
> —Data jargon site, Chun Li

We find one strong sign of progress in the simple existence of a Web site devoting itself to data jargon. The author not only calls "data mining" a *buzz word* (or if there is such a thing, a "buzz phrase")—but also a *hot* one. This neatly reflects the torrents of over-promising, inflated expectations, and limited performance that have surrounded this field.

The quote also points out one prevalent error that has dogged this field since its inception: much that is called data mining falls far short of the real, or perhaps is not "for real." You need to have plans and make judgments, not just poke around and make pictures.

Here is another definition from two formidable experts in creating, not just using, data mining software:

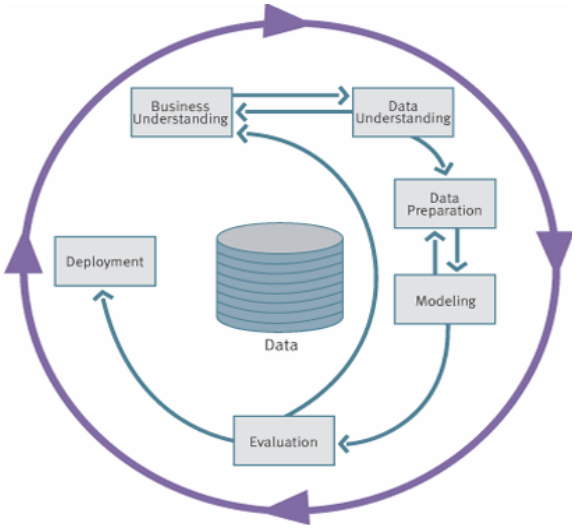
> Extracting implicit, previously unknown, potentially useful information from data. Needed: programs that detect patterns and regularities in the data
>
> —Witten and Frank, *Data Mining*

This quote encapsulates a solid truth: Software programs are just starting to grow up enough that the ideas underlying data mining can be fulfilled. Although the authors have an understandable bias toward the state of analytical algorithms, we have also seen that application of these new methods is just starting to attain its needed form and discipline.

Of course, we have not yet reached the analytical equivalent of nirvana. Note the both quotes still mention finding the "previously unknown." Perhaps this language has become a defensive reflex, given how many efforts have ended with overly obvious "findings" or no findings.

The difference 10 years makes

The same organization that provided our earlier example now strongly endorses the system in the illustration. (It shows the data mining process put forward by the CRISP-DM group.) The figure underlines the need for effective data mining—even when



automated, with machines doing much of the heavy lifting—to work as an iterative process. Understanding of corporate goals, understanding of the domain and the data, data preparation, and modeling all encompass the data. Use of the results takes place only after evaluation; and this evaluation needs to refer back to a solid understanding of the domain and the data. Nearly hidden in this grand cyclic diagram we find one learning about the difficulties involved in data mining. Namely, we need to return to the data repeatedly for more preparation while trying to refine analytical models. Your writer cannot recall one data mining project where attempts to draw forecasts or find underlying relations did not reveal ways in which the data needed more work to become fully useful.

Summary: What we have learned that makes data mining work

Data mining does not need to be awfully complex. Complexity needs to be weighed against usefulness. Simple rules often work surprisingly well. Overly complex algorithms tend not to get applied in many settings.

We must do much more than digging and poking around. Just slogging in data, hoping for something, does not produce useful insights. Data mining needs to be a multifaceted process:

Preparing the data to do the needed analyses typically takes most of the project time. This must be done always referring to business and informational objectives.

Data mining is an iterative process, requiring a lot of thought and labor. Machines can do some forms of data mining, but hardly ever with the research data we typically use.

You can do really well with large, medium and even small data sets. Relatively few applications require sifting through mountains of data. Domain expertise really helps define how much data you will need.

Too much data can have salient disadvantages. Massive data sets can slow or stall analyses, and can make it harder to tell trivial differences from real ones with traditional statistical tests (with enough data, everything becomes hugely significant). More practically, huge data sets may require massive hardware and software investments.

Data mining always is about far more than predictive accuracy. In-depth understanding is critical, not just getting a good score on a model. Domain expertise is critical for knowing if a model makes sense.

We have many new and powerful algorithms, but these alone are not sufficient. If applied as part of the entire data mining process, though, these can produce really strong results.

References

[1]Oates, T. and D. Jensen. (1998) Large datasets lead to overly complex models: An explanation and a solution. *Proceedings of The Fourth International Conference on Knowledge Discovery and Data Mining*, pp. 294-298.

[2]Holte, in *Machine Learning, 11, 1* (April 1993), pp. 63 – 90.

[3]Jensen, D. (2000) Data snooping, dredging and fishing : the dark side of data mining a SIGKDD99 panel report, *CMSIGKDD, Vol. 1, 2*, pp. 52- 56 (January 2000)

[4]Based on 5 years of survey data gathered by NORC, about 15,000 records in total.

[5]Jensen, D. (2000) Data snooping, dredging and fishing: the dark side of data mining (a SIGKDD99 panel report), *CMSIGKDD, Vol. 1, 2,* pp. 52-56 (January 2000)

[6]From "The land mines of data mining" on http://www.praxagora.com

[7]Example from Khabaza, T., (2005) Hard hats for data miners, *DM Direct Special Report* (4/3/2005 edition)

[8]Examples courtesy of Leamer, E. (1978) *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* Wiley. 1978, and updated in a talk by Leamer in 2000.

[9] Khabaza, T., (2005) Hard hats for data miners, *DM Direct Special Report* (4/3/2005 edition)

[10]Witten, I. H, and E. Frank (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufman Publishers.