Theory & Science (2007)

ISSN: 1527-5558

Is automated data mining an emerging paradigm that can supersede hypothesis testing?

Chong Ho Yu, Ph.D.

PO Box 612
Tempe AZ 85280
USA
Email: chonghoyu@gmail.com

What is data mining?

The objective of this article is to evaluate the bold claim that automated search in the form of data mining can supersede conventional hypothesis testing as a new paradigm. Data mining is a cluster of techniques, including classification trees, neural networks, and K-mean clustering, which has been employed in the field Business Intelligence (BI) for years. According to Larose (2005), data mining is the process of automatically extracting useful information and relationships from immense quantities of data. Data mining does not start from a strong pre-conception, a specific question, or a narrow hypothesis, rather it aims to detect patterns that are already present in the data and these patterns should be considered relevant to the data miner. Of a similar vein, Luan (2002) views data mining as an extension of Exploratory Data Analysis (EDA). In short, data mining possesses the following characteristics: (a) the utilization of automated algorithms, (b) the utilization of large quantities of data, and (c) an emphasis of exploration and pattern recognition instead of confirming or disconfirming a pre-defined hypothesis or model.

However, it is impossible for a researcher to completely do away with any pre-conception. At most we can attempt to keep an open mind to other possibilities that may remotely resemble to our existing conception. Take Search for Extraterrestrial Intelligence (SETI) as an example. On one hand scientists should not impose a strict definition of what an intelligent life form is on other planets based upon our own image. But on the other hand they must at least have some loose conceptions of what the signal from an intelligent civilization might sound like. How much pre-conception is considered acceptable remains an issue in data mining.

Data mining vs. testing single hypothesis

Inspired by data mining, Glymour (2004) asserted that we are in the midst of a revolution premised on the automation of scientific discovery made possible by modern computers and new methods of acquiring data. He gave many examples and almost all of them share a common theme: automated data mining either *confirms* or *disconfirms* previously inconclusive findings. For example, in the 1990s a team of Dutch scientists re-analyzed a data set collected in 1970s and “confirmed” that low-level lead exposure is more dangerous to children’s cognitive development than had previously been thought. Using similar techniques, climate researchers were able to trace the global increase in vegetation and its consequences over the last twenty years. Another success story is regarding the use of TETRAD, a causal discovery algorithm invented by Glymour and his colleagues, that leads to correct classification of mineral composition of rocks in order to aid NASA in planetary missions (Moody, Silva, Vanderwaart, Ramsey, & Glymour, 2002). In addition, in a study regarding the causal factors of college student retention, Druzdzel and Glymour (1994) claimed that “one apparently *robust* finding of our study is that student retention is directly related to the average standardized test scores of the incoming freshmen” (p.1, Italics added by me). Druzdzel and Glymour stated that other environmental variables, such as student faculty ratio, faculty salary, and university’s educational expenses per student, are independent of graduation rates.

Glymour argued that the existing scientific paradigm allows just one or very few hypotheses to be entertained and tested by few experiments. Alternatively, the emerging paradigm enables researchers to use computational methods to examine vast numbers of hypotheses and to identify those few that have a reasonable chance of being true. In this fashion, oversights of human judgment can be corrected by computer. This automation of scientific inquiry is chiefly driven by novel abilities to acquire, store, and access previously inconceivable amounts of data, far too much for humans to survey by hand and eye. Put bluntly, Glymour (2004) asserts that there exists a war of paradigms:

> Kuhn said that scientific revolutions generally meet fierce resistance- and the automation of discovery in science is no exception. In some cases the animosity stems from nothing more than conservatism, an effort to preserve academic turf, or simple snobbery. Above all, automated science competes with a grand craft tradition that assumes that science progresses only by scientists advancing a single hypothesis, or a small set of alternative hypotheses, and then devising a variety of experiments to test it. This tradition, most famously articulated by Sir Karl Popper, is championed by many historians and philosophers of science, and resonates with the accounts of science that many senior scientists learned in graduate school. The “Popperian” method of trial and error dominated science from the sixteenth through the twentieth century not because the method was ideal, but because of human limitations, including limitations in our ability to compute (p.75-76).

In addition, Glymour directs his focus to the aspect of automation in data mining instead of exploration. To him using the conventional approach for causal discovery is extremely inefficient. Glymour (2004) mocked hypothesis testing by using the following analogy: “So how does one find a needle in a haystack? Pick something out of the haystack. Subject it to a severe test, e.g., see if it has a hole in one end. If so, conjecture it’s a needle; otherwise, pick something else out of the haystack and try again. Continue until you find the needle or until civilization comes to an end” (p.76-77). With the advance of ultra-fast computers, in Glymour’s view, scientists should move away from the old paradigm to embrace automation, which would be equivalent to setting the haystack on fire and blowing away the ashes to expose the needle.

The above statements are exaggerated. First, it is not historically accurate. It is doubtful that theory testing was predominantly conducted in the Popperian falsification fashion in the last few centuries (Saunders, 2000). It seems that Glymour’s remarks are about what philosophers of science have thought of scientific investigations, but in practice what he described may not be the actual case. However, this point is not related to the main theme here and discussion of Popper is far beyond the scope of this paper. Second, the old paradigm that Glymour severely criticized is the hypothetico-deductive approach, which is manifested in use of significance testing. However, whether this paradigm should be abandoned remains inconclusive. As defenders of the hypothetical-deductive approach in the form of significance testing have argued, this approach provides us with the criteria by which provisionally to distinguish results due to chance variations from results that represent systematic effects in data at hand (Harlow, Mulaik, & Steiger, 1997). Third, data mining has been widely employed in the field of Business Intelligence (BI) for years and has been gradually adopted by educational researchers (Serban & Luan, 2002). Fourth, in actuality, Exploratory Data Analysis and Confirmatory Data Analysis are accepted by many researchers as complementary rather than competing methodologies (Behrens, 1997; Behrens & Yu, 2003). Thus, the techniques of exploration and pattern recognition are not strangers to data analysts and scientists. If data mining is treated as an extension of EDA, widespread acceptance of automated inquiry and data mining is foreseeable. Fifth, by citing many examples of how new data mining techniques corrected previously flawed research or confirmed existing correct beliefs, it seems that Glymour is not trying to suggest the approximate location of the needle inside the haystack, but to pinpoint the exact parameter of the needle. This addresses two key issues: how can we know the target in the research must be a needle, and how can we guarantee that the needle is really inside the haystack? Scientists frequently deal with proxy measures or latent constructs, in which definitions are open to debate.

If we are not sure what proxy measures are good indictors, what the constructs in a model mean or what variables should be loaded into a construct, how can we be so positive that the true causal structure can be unveiled by "blowing away" all other incorrect rival models? In actuality, automation is not the focal point of data mining. In the beginning data miners Berry and Linoff announced data mining as a tool using automatic or semi-automatic means for discovering patterns out of large quantities of data. Later they revised their view by saying, "If there is anything we regret, it is the phrase 'by automatic or semi-automatic means' … because we feel there has come to be too much focus on the automatic techniques and not enough on the exploration and analysis" (cited in Larose, 2005, p.4).

Data mining and the problem of induction

Nevertheless, the preceding arguments are not the main point of this article. Instead of indulging in the cat and mouse game of he-said she-said, I will approach this issue from an epistemological perspective. In brief, using large amount of data in automated data mining does not guarantee theory confirmation, because no matter how large the data set and how sophisticated the searching algorithm, the conclusion yielded by this process is still subject to the new problem of induction, which will be discussed later. To philosophers there is nothing new about the problems of induction; nonetheless, it is the objective of this article to raise the awareness of this issue among data miners. To facilitate the discussion, a first brief definition of induction will be provided. Next, the relationship between data mining and induction will be illustrated. Afterwards, the old and new problems of induction will be thoroughly examined.

What is induction?

In the 1600s the English philosopher Francis Bacon (1620) defined the use of inductive reasoning as drawing conclusions from an exhaustive body of facts. According to Bacon, one should proceed regularly and gradually from one thesis to another based on the generalization of empirical input and particular instances by collecting all relevant data without any presuppositions, so that the generalization is not reached till the last available instance is examined. In other words, each thesis is thoroughly tested by observation and experimentation before the next step is taken. In effect, each confirmed thesis becomes a building block for a higher level concept, with the most generalized thesis representing the last stage of the inquiry. In brief, induction is an inference from observed facts to generalizations.

Goldman (2006) asserts that no scientist has ever been a strict Baconian. First, the scientist would go nowhere if Baconian induction is literally followed for no one could inductively exhaust all facts. Second, the so-called presupposition-less approach inevitably presupposes that reasoning about nature begins with uninterpreted “input” data that are simply given to the mind in experience. Third, it also presupposes the availability of objective relevance criteria. But if the mind is truly passive in reasoning, how do hypotheses arise? Long before Goldman, Carnap (1952) had argued that induction might lead to the generalization of empirical laws but not theoretical laws. For instance, even if we observe thousands of stones, trees and flowers, we never reach a point at which we observe a molecule if we do not engage in theorizing. After we heat many iron bars, we can infer the empirical generalizations that metals will bend when they are heated. But we will never discover the physics of expansion coefficients in this way.

Indeed, superficial empirical–based induction could lead to wrong conclusions. For example, by repeated observations, it seems that heavy bodies (e.g. metal, stone) fall faster than lighter bodies (paper, feather). Without taking air resistance into account, ancient Greeks conjectured that heavy objects are more attracted to the ground than light objects. This Aristotelian belief had misled European scientists for over a thousand years for people did not go beyond superficial empirical–based induction to theoretical principles, but Galileo argued that indeed gravity produces equal acceleration to both heavy and light objects. There is a popular myth that Galileo conducted an experiment in the Tower of Pisa to prove his point. Probably he never performed this experiment. Actually this experiment was performed by one of Galileo’s critics and the result supported Aristotle’s notion. Galileo did not get the correct physical law from observation, but by a chain of logical arguments (Kuhn, 1985).

Nonetheless, to counter the preceding weaknesses in classical induction, researchers today who employ inductive reasoning would not claim that they approach data without any pre-conception; rather, theorizing must be involved in scientific inquiry.

Relationships between data mining and induction

In the context of data mining, induction involves the logic of both generalization and statistical syllogism, but there is a subtle difference between inductive reasoning in data mining and that in a conventional sense. A typical inductive generalization proceeds from a premise about a sample to a conclusion concerning the population. For example,

After examining sample S, it was found that S has attribute A.

Henceforth, it is probable that population P has attribute A.

On the other hand, a statistical syllogism reaches a conclusion in a reverse order: proceeding from the premise about a population to the conclusion of an individual or a sample. For example,

Population P has attribute A.

An individual I or a sample S is a member of P.

Henceforth, there is a probability that I or S has A.

Although data mining can be conceptualized as an integration of both preceding approaches, the reference class and the inference target in data mining, unlike those in conventional induction, are not called population and sample. Take a popular data mining method, induction-based classification tree, as an example. In a typical classification problem, a huge data set is partitioned into a training set and a testing set. The training set, TR, is used to learn about the classifying attributes. After the initial process is completed, the known attributes and the model would be used as a guideline to examine the remaining observations, namely, the testing set, TE (Srivastava, Han, Kumar, Vipin & Singh, 1999). However, unlike a classical inductive inference in which either a generalization is made from a sample to the population or vice versa, data mining attends to the data at hand. Thus, the form of inductive reasoning in data mining is as follows:

After examining TR, it was found that group 1 has attribute A and group 2 has B.

Henceforth, it is probable that in both TR and TE, group X has A and group Y has B.

An individual I in TE has A.

I is a member of X.

After discussing the general relationship between induction and data mining, now we examine how TETRAD and induction are related. Inspired by the formal theory of inductive causation introduced by Pearl and Verma (1990), in which causal graphs are constructed by computing probability distributions in a data-driven iterative learning (updating) process, Glymour, Madigan, Pregibon, and Smyth (1997) tie automated data mining to induction, in the sense of making generalizations of recurring patterns to a broader context. To be explicit, rather than building a coherent global model which includes all variables of interest, data mining algorithms in TETRAD set the rules to inductively produce sets of statements about local dependencies among variables based upon the Causal Markov Condition. Without the aid of algorithms, a human inquirer has to generate categories using empirical data according to his/her judgment, and then further classify other data based on the data-driven categories. Given that the data set is huge, the number of variables is enormous, and thus there are many different strategies, it is very likely that that the human inductor will be locked into inefficient learning strategies. TETRAD and other inductive-based data mining approach utilize machine learning, in which inductive algorithms are provided with training data from a previous stage of the knowledge discovery process. The initially produced model tends to suffer from the problem of over-fitting. Nonetheless, the structure learned from the previous step can be used by the algorithms to inspect another data set, and then the over-fitted model is revised in the light of new information. This iterative process is said to be self-correcting by the algorithms. Thus, the learned computer program can inductively build a model as more and more data sets are supplied to the program (Cooper & Herskovits, 1992).

Goodman: New riddle of induction

While the limitations of simple empirical generalizations are surmountable by categorizing the data and formulating theories, the most serious challenges to justification of induction comes Goodman. Before introducing Goodman’s challenge, it is necessary to mention Hume because Goodman introduced the “new riddle of induction” by analogy with Hume's classical problem of induction. Hume (1777) argued that induction could be justified if and only if we know that instances of which we have no experience resemble those of which we have experience. But we have no grounds that it is not question-begging for believing the statement that “the future will resemble the past”. Although this seems to be a serious challenge to our ability to give good reasons for using induction, in practice it does not stop us from using it, nor does it present any specific problems concerning how we can use induction.

The Humean problem is sometimes known as “the old riddle of induction .” Goodman (1954/1983) introduced the “new riddle of induction ,” in which our conceptualization of kinds plays an important role. Goodman argued that whenever we reach a conclusion based upon inductive reasoning, we could use the same rules of inference, but different criteria of classification, to draw an opposite conclusion. Goodman's example is: Suppose all emeralds examined before time t are green. At t, our observations support the statement that all emeralds are green. Inductively speaking, our evidence statements assert that emerald A is green, that emerald B is green, and so on; and each confirms the hypothesis that all emeralds are green. However, what would happen if another predicate, "grue" is introduced? "Grue" applies to all things examined before t if they are green but to other things if they are blue and not examined before t. Then at t for each evidence statement asserting that a given emerald is green, there is a parallel statement asserting that the emerald is grue. And the statement that emerald A is grue, emerald B is grue, and so on, will each confirm the hypothesis that all emeralds are grue. In this case emeralds A, B, C, examined after time t should be grue, and therefore blue. In other words, the prediction that all emeralds are green and the prediction that all emeralds are grue are both confirmed by evidence statements describing the same observations. Goodman and others further argue that it is difficult to find a principle that supports our preference for using "green" rather than "grue." Thus, the new riddle is also known as "the grue problem."

The new riddle focuses on the problem of projectibility . Whether an "observed pattern" is projectible depends on how we conceptualize the pattern. Skyrms (1975) used a mathematical example to illustrate this problem: If this series of digits (1, 2, 3, 4, 5) is shown, what is the next projected number? Without any doubt, for most people the intuitive answer is simply "6." Skyrms argued that this seemingly straight–forward numeric sequence could be populated by this generating function: (A–1)(A–2)(A–3)(A–4)(A–5)+A and let A be the input digit. However, using the preceding function, the sixth number is 126 and it substantively deviates from the intuitive projection. Skyrms pointed out that whatever number we want to predict for the sixth number of the series, there is always a generating function that can fit the given members of the sequence and that will yield the projection we want. This indeterminacy of projection is a mathematical fact.

The new riddle, an instantiation of the problem of theory under–determination , is germane to quantitative researchers in the context of "model equivalency " and "factor indeterminacy " (DeVito, 1997; Forster, 1999; Forster & Sober, 1994; Kieseppa, 2001; Raykov, & Marcoulides, 2001; Turney, 1999). The new riddle arises because scientific theories are under–determined by our limited evidence in the sense that the same phenomenon can be explained by rival models that are logically incompatible. In factor analysis, for example, the choice of adopting a one–factor or a two–factor model may have tremendous impact on subsequent inferences.

One may argue that the example used by Goodman is too unrealistic for scientists to obtain meaningful implications. How could one be unsure about what color a piece of rock should be? In Goodman's example, there is an association between being an emerald and possessing certain color. But it is possible that this association arises for the wrong reasons. All objects labeled as emeralds result from classification based upon conceptualization, but classification and conceptualization affect the application of the predicate "grue." For example, some emeralds grow on natural colorless beryl seeds, which become coated on both sides, and their growth rate is as slow as 1 mm per month. If these "baby" emeralds are classified as emeralds, what conclusions would we come up regarding the color of emeralds? In addition, natural emeralds appear in a wide variety of green and bluish green because there is a wide spectrum of clarity, along with various numbers of inclusions (an inclusion is any material that is trapped inside a mineral during its formation). Almost all natural emeralds are highly included and it is quite rare to find an emerald with only minor inclusions. There is an old Chinese story about how a King mistakenly tortured an expert on germ stones who donated his most treasured jade to the Royal court, but the King failed to identify the rare jade for his eyesight could not “pierce through” the inclusions of the jade.

There is an equivalent story in the West. When William Atherstone announced that he found a 21-carat diamond in South Africa in 1867, no one believed him because since the fourth century India had been the only source of diamonds for a thousand years. In addition, geologists at that time had strong pre-conceptions about the geological compositions of South Africa and the formation process of diamonds. Diamonds in the raw form are buried at great depths inside the earth. When a volcano erupts, diamonds are thrown out of the top of the volcano along with molten rock, and therefore the best place to find diamonds is in the center of an extinct volcano. However, there are no volcanoes on the mainland of South Africa, and only two are found in the south Indian Ocean, namely, Marion Island and Prince Edward Island. In 1868 England sent one of the best mineralogists, James Gregory, to South Africa for further investigation. After examining many rock samples, Professor Gregory “inductively” concluded that there were no diamonds in the whole of South Africa due to his pre-conceptions of what one might expect from South Africa. He asserted that any genuine diamonds found in South Africa had most likely been swallowed and excreted by wandering ostriches from a far off land. You may think that this mistake is laughable because today indeed there are many diamond mines in South Africa, but you must realize that Professor Gregory had used the best scientific apparatus accessible to him at his time. Today our best equipment is high-power computer. Had computers been available to Professor Gregory, would he have been discovered diamonds in South Africa? Probably the answer is still “no” if Professor Gregory had programmed the computer based on faulty conceptualization of geology, such as attending to traces of volcanoes. What was unknown to Professor Gregory at that time is the Kimberlite pipe, which is resulted from explosive volcanism deep down in the earth. These explosions produce vertical columns of rock, commonly known as dikes, in which raw diamonds are embedded. The diameter of a kimberlite pipe at the surface is typically a few hundred meters to a kilometer only. The first place where kimberlite pipes were recognized is Kimberley, South Africa, and thus this kind of mineral was named after the location of the discovery (Nigel, 1980; Morton, 1877). When Professor Gregory did not even know what a kimberlite pipe is, how could any automated program help him to recognize the potential presence of diamonds?

The morals of this story are: first, as Kuhn pointed out, scientists are not independent of habit, custom, and tradition. Rather they tend to stay within their comfort zone no matter how a complex explanation added to the existing paradigm violates the principle of Occam Razor (birds ate the diamonds in a faraway land and then traveled to South Africa). Second, our conceptualization of mineralogy, in the first place, determines how we classify stones, and subsequently this affects what attributes we can see in particular categories of stones. Just like inductive projection, in causal inferences different conceptualizations can lead to different conclusions on the causal structure.

Take classifying rocks as an example again. Based on the TETRAD approach, Ramsey, Gazis, Roush, Spirites, and Glymour (2002) developed automated methods for mineral identification from reflectance spectra, which is detectable by remote infra-red sensing of terrestrial and extraterrestrial surfaces. It is noteworthy that this kind of mineral classification is not merely descriptive; rather it involves tacit causal inferences because the composition of rocks are tied to specific geological formation processes. In other words, there are theoretical causal links between mineral formation processes and their reflectance spectra detected by infra-red sensors. However, when they supplied the algorithms with training examples for carbonate identification, they found that the trained algorithms did not perform well if the test data contained significant fractions of minerals not in the training set. A further problem is that in reality the reflectance spectra of rocks, soils and other materials are not in general linear or even additive functions of the spectra of their component minerals, and such training procedures therefore lack realistic training sets. No doubt machine learning is more accurate and efficient than humans in processing a large number of observations. However, we must keep in mind that in the initial stage we humans collect the observations and supply the computer with the predicates.

Can Akaike Information Criterion solve the new riddle?

Some quantitative researchers may argue that various criteria, such as the Akaike Information Criterion (AIC) (Akaike, 1973), have been developed to guide us in model selection, and thus the issue of Goodman's riddle and model equivalency is exaggerated. This optimism is unwarranted. Ockham's razor has been taken for granted by many researchers since its introduction by the 14th-century English logician William of Ockham. Not surprisingly, AIC is in alignment to Ockham's razor: Given all things being equal, the simplest model tends to be the best one; and simplicity is a function of the number of adjustable parameters. [1] Actually, AIC does not provide a fool-proof method for choosing between models. The problem with AIC, as well as other model selection criteria, is that the number of parameters associated with a model is a matter of conceptualization. DeVito (1997) gave AIC a litmus test by applying AIC to the grue problem. Let's revisit the two predictive models formulated by Goodman with respect to the observations of emeralds:

- Grue model: If E is an emerald and is observed before time t, then E is green; otherwise, if E is an emerald and is observed after t, then E is blue.
- Green model: If E is an emerald, then it is green.

According to AIC, when both of these two models fit the data, we should favor the Green model because it is the most parsimonious. To be specific, the Grue model has one adjustable parameter, namely, t, while the Green model has no adjustable parameters at all. Apparently, the Green model is simpler and thus is considered better than the Grue model. Could this approach solve the Goodman's riddle once and for all?

No. Because both the Grue model and the Green model can be re-conceptualized in the way that their numbers of adjustable parameters are swapped. If we change the way of how we conceptualize the world, emeralds can no longer be just green or blue; instead, they can be thought to be grue or bleen. As a result, the grue model would have no adjustable parameters while the Green model would have one adjustable parameters. Consider the following two models.

- Grue model: If E is an emerald, then E is grue.
- Green model: If E is an emerald and is observed before time t, then E is grue; otherwise, if E is an emerald and is observed after t, then E is bleen.

In this case, the Green model, according to AIC, seems to be more complex and hence should be rejected. The results of applying AIC are relative to how we conceptualize the world, which is the very essence of the Goodman's riddle. At the present time, there are no commonly agreed solutions to either the new riddle or the model selection criteria. To make the inductive system of TETRAD more defensible, Glymour and his colleagues need to take the Goodman's challenge into account by addressing the issue of how conceptualization of constructs affects the subsequent modeling.

Skinner's challenge

Granted that the latent constructs being put into the model are well-understand and clearly defined, and by automation a pattern among these constructs eventually emerges out of the "haystack." But Hume may be right that our psychological disposition to see the pattern as a causal link is illusory. Even if social scientists dare to ignore Hume and other philosophers, the same warning is established in the realm of psychology. One of Skinner's experiments (1947) demonstrated how an accidental reinforcement schedule could lead to superstitious behaviors. In the experiment, a pigeon is put into a box and occasionally a food hopper is swung into place so that the pigeon can eat from it. If a clock is arranged to present the food hopper at regular intervals with no reference whatsoever to the bird's behavior, operant conditioning usually takes place; the bird developed certain senseless behaviors to beg for food based on the perception that the clock has something to do with food delivery.

Collaborated with some psychologists, Glymour announced the discovery that humans have conducted inquiry in the form of Bayesian network by the age of five (Gopnik, Schulz, & Glymour, 2001; Glymour, 2003; Gopnik & Schulz, 2004). At first glance, this cognitive disposition seems to support making data-driven causal conclusions, but this begs for further questioning. Hume had said that it is our natural instinct to see faulty causal links. Even if psychologists had confirmed that it is natural for humans to engage in causal reasoning in certain ways, it does not provide any justification of the causal conclusion. On the contrary, it makes the matter worse by "naturalizing" a causal model as psychological. When a causal structure is proclaimed, is it considered a model depending on our cognition? Further explanation of what a true causal structure means is on the shoulders of Glymour. Further, other psychologists found that the frequency approach appears to be more natural to learners in the context of quantitative reasoning (Gigerenzer & Edwards, 2003; Hoffrage, Gigerenzer, & Martignon, 2002). Proclaiming a particular reasoning mode as the universal human mind structure, needless to say, would lead to immediate protest. The issue of circularity in justification for induction remains unsettled because psychologists could not reach a consent pertaining to the human reasoning process.

Conclusion

It is important to emphasize that the objective of raising the issues surrounding induction is not to negate the validity of conclusions yielded from data mining or TETRAD. After all, up to the present time no one could sufficiently solve the new riddle of induction, which can be conceptualized in a broader context: under-determination of theory by data. In this sense, a recurring pattern in the data could be explained by a genuine causal link or a psychological illusion. Paradoxically speaking, when a problem is so pervasive that all your rival schools of thought suffer from the same problem, this so-called problem ceases to be a problem (Laudan, 1977). Consider this hypothetical example: a NASA engineer complains that the Hubble telescope cannot transmit real-time images of another galaxy, and thus he coins a new term "the problem of under-determination of image by temporal gap." This alleged weakness of imaging technology is hardly devastating because so far no telescope can work against physics by transmitting a real-time image of something that happened light years away. "The problem of under-determination of image by temporal gap" becomes serious if and only if someday a brilliant scientist is able to punch a wormhole in the space to enable the telescope instantaneously captures the image of a remote galaxy. By the same token, if Methodology A is subject to the problem of under-determination of theory by data, but Methodology B is not, we can assert that A is an inferior methodology. If all known methodologies cannot escape from the problem, the only solution is to try our best in scientific theory. Nonetheless, researchers must take both the old and new riddles into consideration while interpreting a recurring pattern resulting from automated data mining.

Nonetheless, when a problem is pervasive and insurmountable, it does not necessarily imply that the methodologist can do nothing about it. One advisable action is to admit the fallible nature of the methodology by changing the *confirmatory* tone to *exploratory*. At the beginning of this paper, a "robust" finding regarding university student retention by data mining was cited (Druzdzel & Glymour , 1994). However, according to the Student Integration Model (Tinto, 1975, 1982, 1997), many sociological and demographic variables must also be taken into consideration while studying retention. On the other hand, the Student Attrition Model (Bean, 1980, 1983, Bean & Metzner, 1985) approaches the problem of retention with an interest in psychological factors, thus numerous latent constructs are included. Both Tinto's and Bean's models began with far more relevant causal factors than Druzdzel and Glymour's model. Interestingly, in a recent retention study using numerous variables extracted from the Arizona State University data Warehouse (Yu, DiGangi, Jannasch-Pennell, Lo, & Kaprolet, 2007), it was found that retention is strongly tied to "spatial" factors, including residence (in state/out of state) and living location (on campus/off campus), while average standardized test scores of incoming freshmen do not seem to affect the retention rate. This discrepancy is a typical example of how the initial input (selection of variables used for modeling) could affect the output. Weighing this evidence, Druzdzel and Glymour should be more cautious about claiming "robust" results.

It is worth repeating that data mining, as an extension of EDA, aims to detect a pattern and suggest a plausible explanation rather than confirming a conclusion. One of common criticisms against data mining is that this automated methodology draws scientists away from a rigorous and thorough evaluation of each hypothesis in the presence of rival explanations. In the batch processing mode, researchers tend not to devote specific attention to any particular hypotheses. As a remedy, findings based upon data mining should never be treated as "robust conclusions." Reconsider the example of university student retention. When different studies lead to different conclusions, a careful comparison between these studies by examining each theoretical model and each set of input variables is strongly recommended.

Interestingly, Glymour is not the only scholar who uses "the needle in a haystack" metaphor for data mining. Elser (2006) also stated that making a fundamental biological discovery today is similar to finding a needle in a haystack because of the bewildering array of species and species interaction. Throughout previous decades the "haystack" in biology, referring to the contents of genetic information, has been growing at an exponential rate due to the advent of gene sequencing technologies. Faced with this challenge, some biologists viewed algorithms as a viable solution. On one hand, Elser did not deny the possibility that the development and application of mathematical tools can narrow the scope for searching the haystack. On the other hand, he put his hope on developing clear conceptual frameworks by integrating different branches of biology. In my view, aside from biological researchers, scholars in other disciplines should also go beyond making a dichotomous decision, in which one must avoid search algorithms altogether or one must fully embrace data mining without reservations. Both theoretical advances and quantitative formulations must happen simultaneously to cope with the expanding data sources.

Again, this article is by no means intended to dismiss the value of data mining or automated model search algorithms. On the contrary, I agree with Glymour that automated data mining can compensate for several weaknesses of conventional methodologies. However, I express my skepticism toward the claim that automated data mining will lead to a paradigm shift in causal discovery. It is important to address that the essence of data mining is exploration instead of confirmation. Data mining and conventional hypothesis testing should work hand in hand rather than promoting the former as the emerging dominant paradigm as a replacement of conventional methodologies.

Endnotes

1. AIC is a fitness index for trading off the complexity of a model against how well the model fits the data. The general form of AIC is: AIC = 2k – 2lnL where k is the number of parameters and L is the likelihood function of the estimated parameters. Increasing the number of free parameters to be estimated improves the model fitness, however, the model might be unnecessarily complex. To reach a balance between fitness and parsimony, AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. This penalty discourages over-fitting and complexity. Hence, the "best" model is the one with the lowest AIC value. Since AIC attempts to find the model that best explains the data with a minimum of free parameters, it is considered an approach favoring simplicity.

Acknowledgement

Special thanks to Dr. Brad Armendt, Dr. Bernard Kobes, and Dr. Samuel Green for their valuable input to this article.

REFERENCES

Akaike, Hirotsugu. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *International Symposium on Information Theory* (pp. 267–81). Budapest: Akademia Kiado.

Bacon, Francis. (1620 / 1960). *The new organon, and related writings.* New York: Liberal Arts Press.

Bean, John. & Metzner, Barbara. (1985). A conceptual model of nontraditional student attrition. *Review of Educational Research, 55*, 485-540.

Bean, John. (1980). Dropouts and turnover: The synthesis and test of a casual model of student attrition. *Research in Higher Education, 12,* 155-187.

Bean, John. (1983). The application of a model of turnover in work organizations to the student attrition process. *The Review of Higher Education, 6,* 129-148.

Behrens, John. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2,* 131–160.

Behrens, John. and Yu, Chong-ho. (2003). Exploratory data analysis. In J. A. Schinka & W. F. Velicer, (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 33–64). New Jersey: John Wiley & Sons, Inc.

Carnap, Rudolf (1952). *The cognition of inductive methods.* Chicago, IL: University of Chicago Press.

Cooper, Gregory. and Herskovits, Edward. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309-347.

DeVito, Scott (1997). A gruesome problem for the curve–fitting solution. *British Journal for the Philosophy of Science, 48,* 391–396.

Druzdzel, Marek. and Glymour, Clark. (1994). Application of the TETRAD II program to the study of student retention in U.S. colleges. *Proceedings of the AAAI--94 Workshop on Knowledge Discovery in Databases (KDD--94)*, pp. 419-430, Seattle, WA.

Elser, James. (2006). Biological stoichiometry: A chemical bridge between ecosystem ecology and evolutionary biology. *The American Naturalist, 168,* S25–S35.

Forster, Malcolm (1999). Model selection in science: The problem of language variance. *British Journal for the Philosophy of Science, 50,* 83–102.

Forster, Malcolm. and Sober, Elliot (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45,* 1–35.

Gigerenzer, Gerd. and Edwards, Adrian (2003). Simple tools for understanding risks: from innumeracy to insight. *BioMedical Journal, 327,* 741–744.

Glymour, Clark (2003). Learning, prediction and causal Bayes nets. *Trends in cognitive sciences, 7*, 43-48.

Glymour, Clark. Madigan, David. Pregibon, Daryl. and Smyth, Padhraic. (1996). Statistical inference and data mining. *Communications of ACM, 39*, 35-41.

Glymour, Clark. (2004). The automation of discovery. *Dædalus, 133*, 69-77.

Goldman, Steven (2006). *Science wars: What scientists know and how they know it* [CD-ROM]. Teaching Company.

Goodman, Nelson (1954/1983). *Facts, fictions, and forecast.* Indianapolis, IN: Hackett.

Gopnik, Alison. and Schulz, Laura (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences, 8,* 371–377.

Gopnik, Alison. Sobel, David. Schulz, Laura. and Glymour, Clark (2001). Causal learning mechanisms in very young children: two-, three-, and four-year olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*, 620-629.

Harlow, Lisa. Mulaik, Stanley. and Steiger, James (Eds). (1997). *What if there were no significance tests?* NJ: Lawrence Erlbaum Associates.

Hoffrage, Ulrich. Gigerenzer, Gerd. Krauss, Stefan. and Martignon, Laura. (2002). Representation facilities reasoning: What natural frequencies are and what they are not. *Cognition, 2002,* 343–352.

Hume , David (1777/1912). *An enquiry concerning human understanding, and selections from a treatise of human nature.* Chicago: Open Court.

Kieseppa , I. A. (2001). Statistical model selection criteria and the philosophical problem of underdetermination. *British Journal for the Philosophy of Science, 52,* 761–794.

Kuhn , Thomas (1985). *The Copernican revolution.* Massachusetts, MA: Harvard University Press.

Larose, Daniel (2005). *Discovering knowledge in data: An introduction to data mining*. NJ: Wiley-Interscience.

Laudan, Larry (1977). *Progress and its problems: Toward a theory of scientific growth.* Berkeley, CA: University of California Press.

Luan, Jing (2002). Data mining and its applications in higher education. In A. Serban & J. Luan (Eds.), *Knowledge management: Building a competitive advantage in higher education* (pp. 17-36). PA: Josey-Bass.

Moody, Jonathan, Silva, Ricardo, Vanderwaart, Joseph, Ramsey, Joseph, and Glymour, Clark (2002). Classification and filtering of spectra: A case study in mineralogy. *Intelligent Data Analysis, 6,* 517-530.

Morton, William (1877). South African diamond fields, and a journey to the mines. *Journal of the American Geographical Society of New York, 9,* 66-83.

Nigel, Blundell (1980). *The world's greatest mistakes*. London: Octopus Books.

Pearl , Judea. and Verma, Thomas (1990). Equivalence and synthesis of causal models. *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 255 – 270 .

Raykov, Tenko. And Marcoulides, Georg . (2001). Can there be infinitely many models equivalent to a given covariance structure model? *Structural Equation Modeling, 8,* 142–149.

Saunders, Alan (2000). A portrait of Sir Karl Popper. Retrieved September 16, 2006, from http://www.abc.net.au/rn/science/ss/stories/s75303.htm

Serban, Andreea. and Luan, Jing (Eds.). (2002). *Knowledge management: Building a competitive advantage in higher education.* PA: Josey-Bass.

Skinner, Burrhus (1947). Superstition in the pigeon. *Journal of Experimental Psychology, 38*, 168-172.

Skyrms, Brian (1975). *Choice and chance: An introduction to inductive logic (2nd ed.).* Chicago, IL: University of Illinois Press.

Srivastava, Anurag. Han, Eui-Hong. Kumar, Vipin. and Singh, Vineet (1999). Parallel formulations of decision-tree classification algorithms. *Data Mining and Knowledge Discovery, 3*, 237-261.

Tinto, Vincent (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research, 45*, 89-125.

Tinto, Vincent (1982). Limits of theory and practice in student attrition. *Journal of Higher Education, 53,* 687-700.

Tinto, Vincent (1997). Classrooms as communities: Exploring the educational character of student persistence. *Journal of Higher Education, 68,* 599-623.

Turney, Peter (1999). The curving fitting problem: A solution. *British Journal for the Philosophy of Science, 41,* 509–530.

Yu, Chong-ho. DiGangi, Samuel. Jannasch-Pennell, Angel. Lo, Wen-juo. Kaprolet, Charles. and Kim, Chan. (2007, February). *A data mining approach to differentiate predictors of retention between online and traditional students.* Paper presented at the 2007 EDUCAUSE Southwest Regional Conference, Austin, TX.