The Data Avalanche is Here

Harvey J. Miller
Department of Geography
University of Utah
260 S. Central Campus Dr. Room 270
Salt Lake City, Utah 94112-9155
harvey.miller@geog.utah.edu

1. Introduction

For most of its history, science has operated in a data-poor environment. Measurements of reality were difficult, expensive and cumbersome to obtain, store and manipulate. Consequently, much of the apparatus of science is designed to tease information from scarce observations. This has changed dramatically in recent decades. Science has moved from a data-poor to a data-rich environment. The costs of capturing, storing and manipulating digital data have collapsed to a stunning degree, and communications and information technologies are widely deployed in professional and personal settings.

In a recent paper in *Science,* Lazer et al. (2009) note that a computational social science is emerging that is based on the capacity to collect and analyze massive amounts of data on individual and group behavior. However, it is emerging in the private sector such as Yahoo and Google and in government agencies such as the U.S. National Security Agency. Little evidence of this approach appears in the major journals in the social and economic sciences. The authors fear that computational social science will become the exclusive domain of private companies, government agencies and a privileged set of academics working with these entities on research that cannot be critiqued, published and replicated. This will not facilitate the advancement of science or serve the broader public interest in the accumulation and dissemination of knowledge.

In regional science, we have access to an unprecedented amount of fine-grained data on cities, transportation, economies and societies, much of these data referenced in geo-space and time. There is a tremendous opportunity to discover new insights and knowledge about spatial economies that can inform theory and modeling in regional

science, as well as policy and infrastructure decisions. Yet, mirroring the larger trend identified by Lazer et al. (2009), there is little presence in the mainstream literature in regional science. While some activity is emerging, the level of apparent activity is far short of the potential. Computational approaches to discovering patterns in spatio-temporal data mostly reside in the technical literatures in computer science, spatial analysis and geographic information science. While this work is valuable, it does not reflect the specific needs of researchers in regional science. It also does not reflect the rich body of theory and models in regional science; valuable sources of background knowledge that can help guide the exploration of massive spatio-temporal databases.

This paper addresses the potential for discovering new knowledge in regional science through exploration of spatio-temporal databases. There are well-established methods for knowledge discovery from databases, and a growing body of techniques tailored for spatio-temporal data. A reason for the slow adoption in regional science may be due to a lack of awareness about these techniques. But an equally formidable obstacle is misunderstanding of the role of data mining and knowledge discovery in regional science. Rather than being atheoretical or anti-theoretic, the knowledge discovery process harmonizes well with traditional avenues to knowledge construction in science. In fact, the knowledge discovery process benefits from domain expertise and theory to focus searching through vast information spaces and distinguish between real and spurious patterns discovered in these spaces.

The next section of this paper provides a brief history of knowledge discovery from databases, with special reference to the parent disciplines of regional science, namely, economics and geography. Section 3 discusses the new sources of data that are relevant to regional science: fine-grained data on the individual people, objects and money flowing through a spatial economic system. Section 4 reviews the general process of discovering new knowledge from databases, while Section 5 make the case for a specialized methods for geo-spatial data. Section 6 discusses the relationships between knowledge discovery and regional science, with special emphasis on the role of theory. This section also raises several challenges to closer integration between knowledge

discovery and theory-building in regional science. Section 7 concludes the paper by identifying some additional challenges.

Throughout the remainder of this paper, I will use the term "knowledge discovery" to describe the general process of exploring massive digital databases for novel information. "Data mining" refers to the specific techniques applied in one stage of this process. This is a critical distinction that will be more apparent later in the paper (Section 4).

2. A Brief History of Knowledge Discovery from Databases

To date, developments in knowledge discovery and data mining have mostly been driven by computer scientists and others in related fields. The main contributors are the subfields of machine learning and database research. The former involves the study of how machines and humans can learn from data. Machine learning has its origins in artificial intelligence, so early work in the 1950s and 1960s attempted to simulate human learning in computers. This focus subsequently became more pragmatic and continued research focused on developing algorithms and methods that could learn and perform well on specific tasks. Database researchers, in contrast, became interested in knowledge discovery as data warehousing began to grow in the 1990s. A *data warehouse* archives the records in the transactional databases of an organizational, often for liability purposes. As data accumulated, interest grew in exploring these repositories for information to support strategic and tactical decision making. Consequently, many of the exploratory techniques developed from database research focus explicitly on the relational data model (Smyth 2000).

A legacy of knowledge discovery's origins in machine learning and database research is a focus on algorithms and rules and a relative lack of traditional statistical concepts such as parameter estimation and testing. In addition, early applications of digital computers for data analysis in the 1960s lead to the conclusion that one can always search long enough and find a complex but often spurious model that will fit a dataset arbitrarily well. Consequently, "data mining" still has a negative connotation in many fields such as econometrics (Smyth 2000).

One of the more visible debates about data mining in the economics literature concerns issues surrounding specification-search in statistical model-building. An early study by Lovell (1983) analyzes specification-search methods with a simulated dataset and suggests rules for deflating exaggerated claims of significance. Hoover and Perez (1999) extended this simulation design to evaluate *general-to-specific* modeling where one starts with a complex model and reduces to a more elegant one. The theory is, given enough data, only the true specification will survive a sufficiently stringent battery of statistical tests designed to pare variables from the model. This "data mining" approach contrasts with the traditional *specific-to-general* strategy where one starts with a spare model based on theory and conservatively builds a more complex model. Hoover and Perez (1999) report generally favorable results about the ability of the general-to-specific approach to uncover the true model underlying the data. Campos and Ericsson (1999) and Hendry and Krolzig (1999) also support this positive conclusion. However, Hand (1999) dismisses this as irrelevant, claiming that given the size of the information space involved, one would be extraordinarily lucky to configure an initial model that contains the true one. Since one can only hope to discover an approximation of the true model, the only relevant criterion is predictive performance not explanatory validity. This is a controversial position, particularly in a theory-driven science such as economics (Feelders 2002).

Knowledge discovery methods in geography are less controversial than in economics due to the rise of Geographic Information Systems (GIS) as a vehicle for spatial data management and analysis. However, there is a growing concern about potential privacy violations that may occurs from exploring geospatial data, especially when integrating heterogeneous spatial databases. The concept of *locational privacy* results from these concerns, and protocols are emerging to diminish or eliminate these risks without destroying the utility of these data in basic and applied research (see Dobson and Fisher 2003; Duckham, Kulik and Birtley 2006).

As I will suggest below, controversy surrounding the scientific value of knowledge discovery methods results from a misunderstanding of the relationship knowledge discovery and statistics, as well as the relationship between knowledge discovery and theory. Knowledge discovery complements rather than replaces traditional statistics by providing an enhanced process for hypothesis generation; these hypotheses can (and should!) be evaluated against theory as well as through the techniques of confirmatory hypothesis testing. Rather than being antagonistic towards theory, knowledge discovery benefits from a strong theoretical base: this can help guide the search process as well as help evaluate the patterns that emerge from this process.

3. New Data Sources

The key opportunity for economic geographers and regional scientists at this juncture is not just the increasing power of computers, the improving sophistication of knowledge discovery methods, or even the avalanche of available data. Rather, it is the increasing availability of fine-grained data on the location of individual people and objects densely with respect to time, as well as data about online searches, transactions, social connections and commentaries. These data can allow an unprecedented view of societies and economies from the "bottom-up", as well as the aggregate structures, processes and dynamics that emerge from the interactions of individuals in markets, institutions, and regions. Although massive databases have been available for decades, the fine-grained nature of these data is relatively new, and techniques are emerging for exploring these data. This section discusses some of these new data sources.

3.1. Data collection technologies

Point-of-sale data. **A** motivation behind the wider diffusion of knowledge discovery techniques beyond the database and machine learning communities is the availability of point of sale (POS) data through product barcoding. POS data facilitates *market basket analysis*: the analysis and prediction of customer buying habits by finding associations among items that customers place in their shopping basket (Han and Kamber 2006). Indeed, one of the earliest data mining techniques used outside the database and machine learning communities was association rule mining (Smyth 2000). These data remain

important for knowledge discovery in economics, marketing and related fields, particularly when combined with data associated with customer loyalty programs. However, new sources of data, especially data on individual behavior in space and time, is making knowledge discovery more relevant for geography and regional science.

Location-aware technologies. *Location-aware technologies* (LATs) are devices that can report their geographic location densely with respect to time. Major strategies for locational referencing include *radiolocation methods* based on the time, time difference, or angle of the signals' arrivals at base stations from mobile clients, the *global positioning system* (GPS) that exploits time differences of signals arriving from satellites in Earth orbit and *interpolation methods* that use distances and directions from a known location along a route to determine the current location (Grejner-Brzezinska 2004). LATs enable *location-based services* (LBS). LBS provide information to individuals based on their geographic location though client devices such as mobile phones (Benson 2001). Typical LBS include concierge services, navigation, social networking, emergency response, fleet management, local news and tourist information (Spiekermann 2004). Worldwide deployment levels may reach 1 billion devices by 2010 (Bennahum 2001; Smyth 2001). LBS could be a very rich source of data on human activities in space and time if issues regarding privacy and propriety can be resolved (Miller 2007).

An increasingly important LAT is *radiofrequency identification* (RFID) tags. RFID tags attached to objects can transmit data to fixed readers using passive or active methods. Passive tags are cheaper, smaller, and lighter, but have a very limited range. Also, readers cannot track multiple passive tags simultaneously. Active tags are heavier and more expensive since they contain a power source, but have a longer range and readers than can track multiple tags simultaneously. In both systems, the RFID tags must self-identify because the reader conducts the location calculations. This means that RFID systems have greater potential for surveillance and privacy violations than systems such as the GPS where the client conducts the location referencing (Morville 2005). RFID tags are becoming a central feature in the retailing industry due to their capabilities for real-time supply chain management (Roberti 2003). Other RFID applications include

automated toll collection, passports, airline baggage tracking, and VIP services at hotels, clubs and resorts. These applications can allow individual tracking through the products and services that individuals use, but have significant potential for privacy violations (Eckfeldt 2005; McGinty 2004; Shih et al. 2005).

Geosensor networks. Another technology that can capture data on activities in space and time are *geosensor networks*. These are interconnected, communicating, and georeferenced computing devices that monitor a geographic environment. The geographic scales monitored can range from a single room to an entire city or ecosystem. The devices are typically heterogeneous, ranging from temperature and humidity sensors to video cameras and other imagery capture devices. Geosensor networks can also capture the evolution of the phenomenon or environment over time. Geosensor networks can provide fixed stations for tracking individual objects, identify traffic patterns and determine possible stops for a vehicle, as it travels across a given domain in the absence of mobile technologies such as GPS or RFID (Stefanidis 2006; Stefanidis and Nittel 2004).

The Internet. LATs and geosesnor networks only capture human activities in real space. An increasing number of activities and interactions are occurring in *cyberspace*: the domain implied by the world's collective information and communication technologies. Phone calls, emails, texts and other forms of interpersonal communication leave records that can be used to understand social networks, group interactions and social dynamics. The World Wide Web provides a vast repository of what people are saying, buying, searching and connecting with each other. Advances in natural language processing as well as methods for analyzing media data are improving the ability to evaluate social and economic behavior as mediated by these technologies (Lazer et al. 2009).

Brockmann, Hufnagel and Geisel (2006) provide a clever illustration of using web-based data in knowledge discovery about human behavior. The website "Where's George?"

(www.wheresgeorge.com) allows users to input the serial number of a U.S. or Canadian bill and the current location of the user. If the bill has been registered previously, a list of all the locations and times where the bill has been appears; if the bill has not been registered, a new list is created. This is clearly an incomplete, and likely biased, dataset. Nevertheless, Brockmann, Hufnagel and Geisel (2006) analyzed the trajectories of approximately 460,000 bills using this website and were able to conclude that human travel as evidenced by the money flow patterns is consistent with a continuous-time random walk process. Although not a knowledge discovery exercise, this research illustrates that the utility of the noisy, non-scientifically sampled surrogate data that is readily available on the web.

3.2. Simulation

The increasing availability of fine-grained empirical data is only part of the story. Another part is the increasing ability to *generate* vast amounts of synthetic data by simulating large and complex systems at the individual level. The avalanche of CPU cycles bequeathed by improvements in computing engineering (and as described by the now famous Moore's Law) is only one motivation. Equally important is a growing recognition that complex systems such as cities and societies cannot be understood through reductionist approaches. Rather, complex patterns and dynamics emerge from the interactions of individual components of the systems: the whole is more than the sum of the parts (Flake 1998).

Two major traditions for disaggregate modeling in regional science are microsimulation and agent-based modeling. *Microsimulation* is the older tradition: this refers to the modeling and analysis of phenomena at a disaggregate level to order to better understand its aggregate behavior. Microsimulation has a substantial history in social science, dating back to attempts to modeling the US economy in the 1950s (Clarke and Holm 1987). There are well-established standards and techniques for model estimation and validation (Boman and Holm 2004).

Agent-based modeling (ABM) simulates the dynamics of complex systems through the behaviors and interactions of its individual units or *agents*. An agent is an independent, goal-driven that are typically *autonomous* (it makes decisions based on its inputs and goals without an external controlling mechanism) and *adaptive* (its behavior can improve over time through a learning process). Agents interact by exchanging physical or virtual (informational) resources (Maes 1995). Agents can represent people, households, animals, firms, organizations, regions, countries, and so on, depending on the scale of the analysis and the elemental units hypothesized for that scale. The increasing availability of high-resolution data and GIS tools for handling these data facilitate ABM in simulating human spatial systems such as cities and economies (Benenson and Torrens 2004). Applications of ABM include economics (Epstein 1999; Tesfatsion 2009), land-use/land-cover change (Parker et al. 2003), social dynamics (Epstein and Axell 1996), transportation (Balmer et al. 2004), and human movement at micoscales (Batty et al. 2003). ABM offers a rigorous but rich approach to simulating human phenomena from the bottom-up, as well as the concepts of adaptation, self-organization, and emergence to capture linkages between individual behavior and aggregate dynamics (Boman and Holm 2004).

Both microsimulation and ABM have potential for facilitating deeper understanding of complex physical and human systems. Both techniques generate voluminous and intricate results; essentially, massive spatio-temporal databases. Making sense of the results of a large-scale simulation is the same challenge as understanding similarly detailed and massive empirical data about the system in the real-world. This is particularly challenging when conducting multiple simulation runs within an experimental design, as is good practice.

4. What is Knowledge Discovery from Databases?

Knowledge discovery from databases (KDD) is based on a belief that information in the form of *interesting* patterns is hidden in massive databases. "Interesting" means that the information is *easily understood* by humans, *valid* for generalization, potentially *useful*

and *novel* (Han and Kamber 2001). KDD is also predicated on the belief that traditional analysis methods are not capable of discovering the hidden and interesting information in massive, heterogeneous and nonscientific databases.

KDD accommodates data not normally amenable to statistical analysis. Statistical techniques typically requires a clean (relatively noise free) numeric database scientifically sampled from a large population with specific questions in mind. Many statistical models require strict assumptions (such as independence, stationarity of underlying processes, and normality). In contrast, the empirical and synthetic data being generated and stored in many databases are noisy, non-numeric and possibly incomplete. These data are also collected in an open-ended manner without specific questions in mind or were generated as a byproduct of another activity (Hand 1998). KDD in its contemporary form encompasses techniques from statistics, machine learning, pattern recognition, numeric search and scientific visualization to accommodate the new data types and data volumes being generated through information technologies.

4.1. The Knowledge Discovery Process

The KDD process usually consists of several stages, corresponding to major tasks involved in preparing and exploring the data as well as interpreting results (Adriaans and Zantinge 1996; Brachman and Anand 1996; Fayyad, Piatetsky-Shapiro and Smyth 1996; Han and Kamber 2006; Matheus, Chan and Piatetsky-Shapiro 1993). Although the order of the stages below represent a standard progression though the KDD process, in practice they may not be executed in any particular order. Some steps may be skipped and others repeated, depending on the judgment of the analyst guiding the process and the intermediate results. This is a key point. KDD is not an automated, push-button process: it demands intelligence decision-making, domain expertise and thoughtful reflection. These are skills that human minds still do much better than computers.

- *Data selection* involves determining a subset of the records or variables in a database for knowledge discovery.

- *Data pre-processing* involves "cleaning" the selected data to remove noise, eliminating duplicate records, and handling missing data fields and domain violations. The pre-processing step may also include *data enrichment* through combining the selected data with other, external data (e.g., census data, market data).
- *Data reduction and projection* diminishes the dimensionality of the data through transformations to equivalent but more efficient representations of the information space.
- *Data mining* involves the application of low-level algorithms to uncover hidden patterns in the data.
- *Interpreting and reporting* stage involves evaluating, understanding and communicating the information extracted from the data.

4.2. Data Mining Techniques

Data mining involves the application of low-level functions or algorithms for revealing hidden information in a database (Klösgen and Żytkow 1996). There are several major classes of data mining functions: the type of information being sought determines the particular type of data mining function to be applied (Han and Kamber 2006). Table 1 summarizes major data mining tasks. All of these techniques share a common characteristic: *scalability*, or the ability to handle massive databases without unreasonable increases in computational time.

Data mining task	Description	Techniques
Segmentation or clustering	Determine a finite set of implicit groups that describes the data.	• Cluster analysis
Classification	Predict the class label that a set of data belongs to based on some training datasets	• Bayesian classification • Decision tree induction • Artificial neural networks • Support vector machine
Association	Find relationships among data objects; predict the value of some attribute based on the value of other attributes	• Association rules • Bayesian networks
Deviations	Find data items that exhibit unusual deviations from expectations	• Cluster analysis • Outlier detection • Evolution analysis
Trends	Lines and curves summarizing the database, often over time	• Regression • Sequential pattern extraction
Generalizations	Compact descriptions of the data	• Summary rules • Attribute-oriented induction

Table 1: Data mining tasks and techniques (Miller and Han 2009)

Segmentation or *clustering* involves partitioning a selected set of data into meaningful groupings or classes. The commonly used data mining technique of *cluster analysis* determines a set of classes and assignments to these classes based on the relative proximity of data items in the information space. Cluster analysis methods for data mining must accommodate the large data volumes and high dimensionalities of interest in data mining; this usually requires statistical approximation or heuristics. *Classification* refers to finding rules or methods to assign data items into pre-existing classes. There are many classification methods developed in many years of research in statistics, pattern recognition, machine learning and data mining, including decision tree induction, naïve Bayesian classification, neural networks and support vector machines. *Associations* are rules that predict the relationships between data objects or the value of some attribute based on the value of other attributes. *Deviations* are data items that exhibit unexpected deviations or differences from some norm. These cases are either errors that should be corrected/ignored or represent unusual cases that are worthy of additional investigation.

Trends are lines and curves fitted to the data; techniques include linear and logistic regression analysis that are very fast and easy to estimate. These methods are often combined with filtering techniques such as stepwise regression. *Generalization and characterization* are compact descriptions of the database. Techniques include *summary rules* that generate a relatively small set of logical statements that condense the information in the database.

4.3. Scientific Visualization and Knowledge Discovery

Visualization is a powerful strategy for integrating high-level human intelligence and knowledge into the KDD process. The human visual system is extremely effective at recognizing patterns, trends and anomalies. The visual acuity and pattern spotting capabilities of humans can be exploited in many stages of the KDD process, including OLAP, query formulation, technique selection and interpretation of results. These capabilities have yet to be surpassed by machine-based approaches (Fayyad, Grinstein and Wierse 2001).

5. Knowledge Discovery in Regional Science

5.1. Geospatial Knowledge Discovery

Geospatial knowledge discovery (GKD) is the process of extracting knowledge massive georeferenced databases. GKD has emerged as a subdomain of KDD due to the unique requirements of geospatial data, information and knowledge. There are several reasons why GKD is unique.

Geospatial information is not only highly dimensioned, but also have the property that up to four dimensions (representing space and time) form a framework for the remaining dimensions. Fidelity with the real-world requires embedding georeferenced observations within a formal space that reflects empirical geographic and temporal relationships as faithfully as possible (or appropriate). Euclidean space is the most common framework, and often adopted implicitly by default, but it is only one of an infinite number of possibility. Techniques exist for estimating the geo-space implied by a matrix of

distance, cost or flows between geographic locations and analyzing differences among these spaces (see Ahmed and Miller 2007; Tobler 1994). Whatever space is adopted, it carries with it a rich spectrum of implied relations among the embedded spatial objects, including proximity, connectivity and direction (Miller and Wentz 2003). The information implicit in the geographic measurement framework is ignored in many knowledge discovery tools (Gahegan 2000a).

Geographic data usually exhibit the properties of *spatial dependency* and *spatial heterogeneity*. Spatial dependency is the tendency of attributes to be more related with proximity in geographic space due to the impeding effects of distance (Tobler 1970). Spatial heterogeneity refers to the non-stationarity of most geographic processes. The "friction" of distance combined with the relative uniqueness of each location means that geographic processes are local. Spatial dependency and spatial heterogeneity have historically been regarded as nuisances confounding standard statistical techniques that typically require independence and stationarity assumptions. However, these can also be valuable sources of information about the geographic phenomena under investigation.

Spatial objects tend to be more complex than the non-spatial objects, particularly with respect to their geometric footprint. Spatial objects often cannot be reduced to points in some information without doing great harm to the representation of the real-world phenomenon. Size, shape and boundary properties of geographic entities often affect geospatial processes, sometimes due to measurement artifacts (e.g., recording flow only when it crosses some fiat boundary), but often due to physical and human properties in the real world (e.g., a mountain range forming a natural boundary between two nations, or a divided highway creating an equally pervasive boundary between two urban neighborhoods). Relationships such as distance, direction and connectivity are more complex with dimensional objects (see Egenhofer and Herring 1994; Okabe and Miller 1996; Peuquet and Ci-Xiang 1987). The number and complexity of implied spatial relationships increase dramatically with dimensional spatial objects such as lines, polygons, surfaces and solids.

In addition to changes in their non-spatial attributes, spatial objects can also undergo change over time with respect to their geometry and spatial identity. We often refer to the former type of change as *motion*: this can include changes in morphology, location, or both. Motion is deceptively complex concept: it can also occur in the whole object or its composite parts, it can be continuous or discrete with respect to space and time and can be conceptualized at the individual or collective scales (Galton 1995, 1997). Spatial identity changes include events such as being created, destroyed, aggregated, disaggregated, fusion with other spatial objects and fission into new spatial objects. New spatial objects may also be spawned or cloned from existing objects (Frank 2001; Hornsby and Egenhofer 2000; Medak 2001).

5.2. Geospatial Data Mining Techniques

Many of the traditional data mining tasks discussed above have analogous tasks in the geographic data mining (Ester, Kriegel and Sander 1997; Han and Kamber 2006). *Spatial classification* builds up classification models based on a relevant set of attributes and attribute values that determine an effective mapping of spatial objects into predefined target classes. *Spatial clustering* groups spatial objects such that objects in the same group are similar and objects in different groups are unlike each other. Clustering can be based on combinations of non-spatial attributes, spatial attributes (e.g., shape) and proximity of the objects or events in space, time and space-time. *Spatial trend detection* involves finding patterns of change with respect to the neighborhood of some spatial object. *Spatial characterization and generalization* is therefore an important geographic data mining task. Generalization-based data mining can follow one of two strategies in the geographic case. *Spatial dominant generalization* first spatially aggregates the data and then applies standard attribute-oriented induction method at each geographic aggregation level. *Non-spatial dominant generalization* generates aggregated spatial units that share the same high-level semantic description. *Spatial association rules* are association rules that include spatial predicates in the precedent or antecedent.

5.3. Geovisualization

Geographic visualization (GVis) is the integration of cartography, GIS, and scientific visualization to explore geographic data and communicate geographic information to private or public audiences (see MacEachren and Kraak 1997). GVis and GKD are highly complementary, perhaps even more than KDD and scientific visualization due to importance of geometry and position in spatial data. Integration between GVis and GKD can occur at several levels, including determining high-level goals for the GKD process, specification of appropriate geographic data mining tasks for achieving the high-level goals and choices include specific tools and algorithms to achieve the data mining task specified (MacEachren et al. 1999).

6. Regional Science and Knowledge Discovery: The Role of Theory

KDD is sometimes dismissed by those outside its community as a black-box technique, or as a "fishing expedition." It should be clear by this point in the paper that this is not the case: KDD is complex and require thoughtful guidance by an interested and knowledgeable expert. Nevertheless, too often the "expert" is a computer scientist who enjoys building and testing tools for the sake of the tools themselves. KDD in general has not spread widely into the domain sciences beyond the fields of marketing and some of the physical sciences such as chemistry and medicine. GKD is also not well known or widely used outside of the GIS community: it is almost invisible in economic geography and regional science. This is unfortunate since regional scientists are missing opportunities to advance their science by exploiting newly available data on economic systems and increasingly sophisticated tools for exploring these data. They are also missing an opportunity to influence the development of these tools for the greater benefit of regional scientists.

GKD has a supporting role in modeling and theory-building: it does not attempt to replace or diminish these processes. Also, there is a crucial role for theory in supporting the GKD process.

6.1. Knowledge Discovery to Support Theory

KDD is essentially a hypothesis generation process; much more so than traditional inferential statistics. Although statistical analysis is appropriately viewed as an inductive process, it is often embedded within a broader deductive process. Statistical models are confirmatory, requiring the analyst to specify a model based on some theory, test the hypotheses generated by that model, and revise the theory depending on the results. In contrast, the deeply hidden patterns sought through KDD are difficult or impossible to specify *a priori*, at least with any reasonable degree of completeness. The number of potential hypothesis implied by a massive database is often too large to test exhaustively, even if many can be dismissed *a priori* as trivial or absurd. KDD is more concerned about prompting investigators to formulate *new* predictions and hypotheses from data: KDD seek novel information which (by definition) is surprising and would not have otherwise come to mind (Elder and Pregibon 1996; Hand 1998). In this sense, KDD is similar to a telescope or microscope: it is a way for researchers to look at the data in different ways, discover something new, and formulate theories, models or hypothesis based on this novel information within the context of existing theory and knowledge.

In addition to supporting the hypothesis formulation stage of scientific knowledge construction, Roddick and Lees (2009) argue that the knowledge discovery process can work directly in concert with traditional knowledge construction in science. Echoing the concerns expressed by Hand (1999), they note that exponentially large information space implied by a massive database may lead to an induction fallacy where hypothesis developed from the data are consistent but do not reflect the true model. Roddick and Less (2009) suggest a strategy that not only focuses the search for patterns through data selection and reduction, but also can support conceptual model building. The investigator supplies a conceptual model as a starting point for generating experimental hypotheses and a set of associated null hypotheses. Data mining techniques focus on the subregions of the information space indicated by these hypotheses. In accordance with the usual notion of scientific induction, confidence in the conceptual model increases if discovered patterns support its hypotheses. Unsupported hypotheses suggest a modification to the conceptual model or the investigator's reasoning about the phenomenon. This process can also accept competing conceptual models and provide a ranking based on support for

the data. However, this is a modeling building and data focusing strategy, not a modeling testing procedure: final acceptance of rejection of a model should be based on standard confirmatory procedures and tests.

6.2. Theory to Support Knowledge Discovery

Despite data reduction techniques and the computational efficiency of data mining tools, the information space implied by a massive, heterogeneous database may be so large or so complex that it cannot be exhaustively explored. A good data mining system should be able to generate all of the interesting patterns in a database, but only the interesting ones. Unfortunately, while some dimensions of interestingness can be such as validity can be assessed automatically, the qualities of comprehensible, useful and novel are more difficult to assess using automated methods. Consequently, it is typical for a data mining system to generate an unacceptably large number of patterns, most of which could be dismissed as difficult to interpret, useless or trivial. This is particularly true in GKD: the number of spatial predicates and transformations over time is so large that the number of candidate patterns to evaluate can be overwhelming (Roddick and Lees 2009).

Background knowledge is strategy to maximize the completeness of a data mining system. It is a computational representation of domain knowledge to focus data mining techniques on subregions of the information space that are likely to contain interesting patterns. Background knowledge reflects known facts about the domain being investigated in a way that can be exploited by data mining algorithms. For example, a common type of background knowledge is a *concept hierarchy*: this is a sequence of mappings from low-level to high-level semantics. It is typically organized as a tree whose leaves correspond to measured attributes in the database and parents represent higher level semantics that synoptically summarize the lower-level semantics of its children. Concept hierarchies provide a logical map for aggregating or drilling-down attributes automatically during the data mining process (Han and Kamber 2006). Background knowledge can be derived from assumptions about the system, observable facts, expert knowledge, or theory.

Interestingness measures quantitatively separate interesting patterns from uninteresting ones by assessing the simplicity, certainty, utility and novelty of the generated patterns (Silberschatz and Tuzhilin 1996; Tan, Kumar, and Srivastava 2002). These can be used to guide data mining algorithms using feedback, or evaluate the discovered patterns after discovery. There are different types of interestingness measures defined across the dimensions mentioned above, and some are specific to the type of data mining technique. For example, the resulting *rule length* when expressing the pattern in conjunctive normal form is a measure of simplicity (and therefore legibility). A simple *confidence* measure for association rules is the number of times two facts appear together relative to the appearance of one fact alone. *Support* refers to the number of times in the database for which a pattern is true; this measures the general utility of the pattern (Han and Kamber 2006).

Computer scientists have become very clever at developing ways to focus the knowledge discovery process. However, they are starving for content: the domain-specific facts that can guide discovery. Regional science, on the other hand, have a very rich body of theory and models that can be used as background knowledge to guide the discovery process and as interestingness measure to filter spurious patterns from potentially interesting ones. However, there are three challenges to translating domain knowledge from regional science to the discovery process

First, similar to many human sciences, regional science concepts can be abstract, vague, fluid and multi-level. For example, the core concept of a "region" can very depending on the context. Similarly, definitions of entities such as "household," "firm" and "city" are not always clear or consistent across different research inestigations. This is understandable given the complex nature of the phenomena of interest in regional science, but it does point to a need for more consistent definitions, even if these are flexible. In brief, regional science needs a clear *ontology* that provides formal definitions of its basic core concepts and the relationships among these concepts. Jackson (1994) advocates the use of object-oriented modeling languages in regional science, arguing that this can improve not only the modeling process itself through facilitating incremental

model building, but also the communication and collaboration among regional scientists. The latter point is the one that is particularly relevant for the discussion here: the development of standards and definitions through object-oriented modeling languages can allow the creation and maintenance of common and extensible libraries of core regional science concepts that can be shared across researchers and applications.

A second challenge concerns differences in knowledge representation between regional science versus computer science. In regional science, most knowledge is implicit: it is embedded in formal theories, models and equations. However, knowledge discovery techniques require *explicit* representations in the form of rules, hierarchies and concept networks. Translating knowledge embedded in regional science theories and models to a form that can be exploited by automated data mining techniques is not trivial: these explicit representations must not only be *effective* (can capture all of the implicit knowledge, but not more) but *efficient* (can be applied without undue computational burden).

A third challenge is a need for effective methods for spatial pattern evaluation when comparing reality to theory. If an observed spatial pattern matches a theoretical prediction, this is interesting but not novel (unexpected). At the other extreme, if an observed spatial pattern matches a spatial null (the trivial pattern that would be expected by chance) is neither interesting nor novel. The spatial patterns that fall between theory and null (in other words, those that are not trivial and at variance with theory), may be both interesting and novel, and therefore worthy of further investigation. However, there is no clear metric for comparing observed spatial patterns to null and theoretical spatial patterns. One difficulty is the lack of a good spatial null. Michael Goodchild has argued persuasively that complete spatial randomness (CSR) is a "straw target:" the likelihood of any spatial pattern following CSR is so remote that this null model is likely to be rejected in all cases. But, what is the alternative? It is unclear that there is a general spatial null model across all domains and applications, and the appropriate spatial null will depend on the behavioral and other processes assumed or postulated in the specific theory or model. Also, it is unclear how to measure deviations from null or expected spatial patterns.

Differences between observed and null or expected patterns can be expressed geometrically, but it is not clear how a geometric difference can be translated into one-dimensional, unambiguous quantitative measure of departure from a norm.

7. Other Challenges

In additional to the challenges mention above, there are challenges facing the wider use of knowledge discovery techniques in regional science.

Semantically poor data. LATs and ICTs can generate a vast amount of fine-grained data on peoples' activities in geo-space and cyberspace. However, much of these data are semantically poor. For example, while it is easy to use vehicle-based GPS receivers or location-enabled mobile phones to capture individual trajectories in space, it is more difficult to get the attributes associated with these trajectories such as the characteristics of the person, the activities conducted, or the planned activities that could not be conducted. Some of the information can be recovered by integrating heterogeneous data (for example, map-matching trajectories to detailed land-use data). However, this can be cumbersome, raise privacy concerns, and introduce error. Although existing efforts to exploit surrogate and auxiliary data to recover or infer data semantics should continue, working with voluminous but semantically poor data may require asking some research questions differently, or asking the same questions in a different way.

Algorithms and infrastructures. Spatial models and techniques can be computationally complex. In theory, a spatial model requires evaluation of pairwise distances among all locations in space. This implies quadratic computational times in the worse-case: this is unacceptably slow for knowledge discovery and data mining. The worse-case can be avoided in most cases since we know that most spatial processes attenuate with distance (indeed, this is why Waldo Tobler originally coined the often-cited "Tobler's First of Law of Geography" in Tobler (1970): it was a justification for skipping calculations between entities that are spatially remote). However, in general this is a heuristic, and the implications of these shortcuts for the knowledge discovery process must be understood.

There is also a need for more effective use of parallel and high-performance computing environments in spatial knowledge discovery. The use of parallel, grid and cloud-computing environments is seriously under-researched in regional science, geography and even Geographic Information Science despite the fact that many spatial models are easily adapted to these environments due to the inherent parallelism of spatio-temporal data, as well as parallelism in the computational tasks required (see Armstrong, Wang and Cowles 2005).

Education. Finally, educating the next generation of economists, regional scientists, and quantitative geographers is a major challenge. It is already difficult to provide students with the necessary background in economic and geographic theory to conduct cutting-edge research in regional science: adding computer science to this mix will be challenging, especially since the language and ways of thinking in these disparate fields are in many ways orthogonal to each other. Also, given the dynamic nature of digital technologies, it will be necessary to continue education throughout one's career in order to keep abreast of developments in the field. I am not saying that regional scientists should become computer scientists: in fact, I think this would be a very bad thing. But, regional scientists need the appropriate expertise to choose, implement and interpret knowledge discovery and data mining techniques effectively.

Literature Cited

Adriaans P. and Zantinge, D. (1996) *Data Mining*; Harlow, U.K.: Addison-Wesley

Ahmed, N. and Miller, H. J. (2007), "Time-space transformations of geographic space for exploring, analyzing, and visualizing transportation systems," *Journal of Transport Geography*, 15, 2-17.

Armstrong, M.P., Wang, S., and Cowles, M. (2005) "Using a computational grid for geographic information analysis," *Professional Geographer*, 57, 365-375.

Balmer, M., Nagel, K., and Raney, B. (2004) "Large-scale multi-agent simulations for transportation applications," *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations,* 8, 205–221.

Batty, M., Desyllas, J., and Duxbury, E. (2003) "The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades," *International Journal of Geographical Information Science* 17, 673–697.

Benenson, I., and Torrens, P. M. (2004). *Geosimulation: Automata-based Modeling of Urban Phenomena*. Chichester, UK: John Wiley & Sons.

Bennahum, D. S. (2001) "Be here now," *Wired,* 9 (11), 159–163

Benson, J. (2001) "LBS technology delivers information where and when its needed," *Business Geographics*, 9 (2), pp. 20–22.

Boman, M., and Holm, E. (2004) "Multi-agent systems, time geography and microsimulations," in Olsson, M.-O. and Sjöstedt, G. (eds) *Systems approaches and their applications*. Dordrecht, The Netherlands: Kluwer Academic, 95–118.

Brachman, R. J. and Anand, T. (1996) "The process of knowledge-discovery in databases: A human-centered approach," in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press 37-57.

Brockmann, D., Hufnagel, L. and Geisel, T. (2006) "The scaling laws of human travel," *Nature*, 439, 462-465.

Campos, J. and Ericsson, N. R. (1999) "Constructive data mining: Modeling consumers' expenditures in Venezuela," *Econometrics Journal*, 2, 226-240.

Clarke, M., and Holm, E. (1987) "Microsimulation methods in spatial analysis and planning," *Geografiska Annaler* 69B, 145–164.

Dobson, J. E. and Fisher, P. F. (2003) "Geoslavery," *IEEE Technology and Society Magazine* 22(1): 47-52

Duckham, M., Kulik, L. and Birtley, A. (2006) “A spatiotemporal model of strategies and counter-strategies for location privacy protection,” in M. Raubal, H. J. Miller, A. U. Frank and M. F. Goodchild (eds.) *Geographic Information Science: 4th International Conference, GIScience 2006, Münster, Germany, September 2006 Proceedings*, Springer, 47-64.

Eckfeldt, B. (2005) “What does RFID do for the consumer?” *Communications of the ACM* 48 (9), 77–79.

Egenhofer, M. J. and Herring, J. R. (1994) "Categorizing binary topological relations between regions, lines and points in geographic databases," in M. Egenhofer, D. M. Mark and J. R. Herring (eds.) *The 9-intersection: Formalism and its Use for Natural-language Spatial Predicates*, National Center for Geographic Information and Analysis Technical Report 94-1, 1-28.

Elder, J. and Pregibon, D. (1996) “A statistical perspective on knowledge discovery,” pp. 83-113 in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press, 83-113

Epstein, J. M. (1999). “Agent-based computational models and generative social science,” *Complexity* 4 (5), 41–60.

Epstein, J. M., and Axtell, R. (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. Cambridge, MA: MIT Press

Ester, M., Kriegel, H.-P. and Sander, J. (1997) "Spatial data mining: A database approach," M. Scholl and A, Voisard (eds.) *Advances in Spatial Databases*, Lecture Notes in Computer Science 1262, Berlin: Springer, 47-66.

Fayyad, U., Grinstein, G. and Wierse, A. (2001) *Information Visualization in Data Mining and Knowledge Discovery*, San Matel, CA: Morgan Kaufmann.

Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P. (1996) “From data mining to knowledge discovery: An overview” in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.) *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press, 1-34.

Feelders, A. (2002) “Data mining in economic science,” in J. Meij (ed.) *Dealing with the Data Flood: Mining Data, Text and Multimedia*, The Haag, The Netherlands: STT/Beweton, 165-175.

Flake, G. W. (1998) *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation*, Cambridge, MA: MIT Press.

Frank, A. (2001) “”Socio-economic units: Their life and motion,” in A. Frank, J. Raper, J.-P. Cheylan (eds.) *Life and Motion of Socio-economics Units*, London: Taylor and Francis, 21-34.

Galton, A. (1995) “Towards a qualitative theory of movement,” in A. Frank and W. Kuhn (eds.) *Spatial Information Theory: Proceedings of the European Conference on Spatial Information Theory (COSIT),* Springer-Verlag, 377-396.

Galton, A. (1997) “Space, time and movement,” in O. Stock (ed.) *Spatial and temporal Reasoning,* Kluwer, 321-353

Gahegan, M. (2000a) "On the application of inductive machine learning tools to geographical analysis," *Geographical Analysis*, 32, 113-139

Grejner-Brzezinska, D. (2004). “Positioning and tracking approaches and technologies,” in Karimi, H. A. and Hammad, A. (eds) *Telegeoinformatics: Location-based Computing and Services*, Boca Raton, FL: CRC Press, pp. 69–110.

Han, J. and Kamber, M. (2006), *Data Mining: Concepts and Techniques*, 2nd ed., San Matel, CA: Morgan Kaufmann.

Hand, D. J. (1998) "Data mining: Statistics and more?" *American Statistician,* 52, 112-118

Hand, D. J. (1999) “Discussion contribution on ‘Data mining reconsidered: Encompassing and the general-to-specific approach to specification search’ by Hoover and Perez,” *Econometrics Journal*, 2, 241-243.

Hendry, D. F. and Krolzig, H.-M. (1999) “Improving on ‘Data mining reconsidered,’ by K. D. Hoover and S. J. Perez,” *Econometrics Journal*, 2, 202-219.

Hoover, K. D. and Perez, S. J. (1999) “Data mining reconsidered: Encompassing and the general-to-specific approach to specification search,” *Econometrics Journal*, 2, 167-191.

Hornsby, K. and Egenhofer, M. J. (2000) “Identity-based change: A foundation for spatio-temporal knowledge representation,” *International Journal of Geographical Information Science*, 14, 207-224

Jackson, R. W. (1994) “Object-oriented modeling in regional science: An advocacy view,” *Papers in Regional Science*, 73, 347-367.

Klösgen, W. and Żytkow, J. M. (1996) "Knowledge discovery in databases terminology," in U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Ulthurusamy (eds.)

Advances in Knowledge Discovery and Data Mining, Cambridge, MA: MIT Press 573-592.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D. Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D. and Van Alstyne, M. (2009) "Computational Social Science," *Science*, 323, 721 – 723.

Maes, P. (1995) "Modeling adaptive autonomous agents," in Langton, C. (ed.) *Artificial Life: An Overview*. Cambridge, MA: MIT Press, 135–162.

MacEachren, A. M. and Kraak, M.-J. (1997) "Exploratory cartographic visualization: Advancing the agenda," *Computers and Geosciences*, 23, 335-343.

MacEachren, A. M., Wachowicz, M., Edsall, R., Haug, D. and Masters, R. (1999) "Constructing knowledge from multivariate spatiotemporal data: Integrating geographic visualization with knowledge discovery in database methods," *International Journal of Geographical Information Science*, 13, 311-334

Matheus, C. J., Chan, P. K. and Piatetsky-Shapiro, G. (1993) "Systems for knowledge discovery in databases," *IEEE Transactions on Knowledge and Data Engineering*, 5, 903-913.

McGinty, M. (2004) "RFID: is this game of tag fair play?" *Communications of the ACM* 47 (1), pp. 15–18.

Medak, D. (2001) "Lifestyles," in A. Frank, J. Raper, J.-P. Cheylan (eds.) *Life and Motion of Socio-economics Units*, London: Taylor and Francis, 139-153.

Miller , H. J. (2007) "Place-based versus people-based geographic information science," *Geography Compass*, 1, 503-535.

Miller, H. J. and Han, J. (2009) "Geographic data mining and knowledge discovery: An overview," in Miller, H. J. and Han, J. *Geographic Knowledge Discovery and Data Mining*, second edition. CRC Press.

Miller, H. J. and Wentz, E. A. (2003) "Representation and spatial analysis in geographic information systems," *Annals of the Association of American Geographers*, 93, 574-594.

Morville, P. (2005) *Ambient findability: What we Find Changes Who we Become*. Sebastopol, CA: O'Reilly Media.

Okabe, A. and Miller, H. J. (1996) "Exact computational methods for calculating distances between objects in a cartographic database," *Cartography and Geographic Information Systems*, 23, 180-195.

Parker, D. C., Manson, S. M., Janssen, M. A., Hoffmann, M. J. and Deadman, P. (2003). "Multi-agent systems for the simulation of land-use and land-cover change: a review," *Annals of the Association of American Geographers* 93, 314–337.

Peuquet, D. J. and Ci-Xiang, Z. (1987) "An algorithm to determine the directional relationship between arbitrarily-shaped polygons in the plane," *Pattern Recognition*, 20, 65-74.

Roberti, M. (2003) "Wal-Mart spells out RFID vision," *RFID Journal* [online]. Retrieved on 16 June 2003 from http://www.rfidjournal.com/

Roddick, J. F. and Lees, B. G. (2009) "Spatio-temporal data mining: Paradigms and methodologies," in H. J. Miller and J. Han (eds.) *Geographic Data Mining and Knowledge Discovery,* 2ed., CRC Press, in press.

Shih, D.-H., Lin, C.-Y., and Lin, B. (2005) "RFID tags: Privacy and security aspects," *International Journal of Mobile Communications,* 3, 214–230.

Silberschatz, A. and Tuzhilin, A. (1996), "What makes patterns interesting in knowledge discovery systems", *IEEE Transactions on Knowledge and Data Engineering*, 8, 970-974.

Smyth, C. S. (2001). "Mining mobile trajectories," in Miller, H. J. and Han, J. (eds) *Geographic Data Mining and Knowledge Discovery*. London: Taylor and Francis, 337–361.

Spiekerman, S. (2004) "General aspects of location-based services," in Schiller, J. and Voisard, A. (eds) *Location-based services*. San Francisco, CA: Morgan-Kaufmann, pp. 9–26.

Stefanidis, A. (2006). The emergence of geosensor networks. *Location Intelligence* [online]. Retrieved on 27 February 2006 from http://locationintelligence.net

Stefanidis, A., and Nittel, S. (eds) (2004). *Geosensor networks*. Boca Raton, FL: CRC Press.

Tan, P.-N., Kumar, V. and Srivastava, J. (2002) "Selecting the right interestingness measure for association patterns", *Proceedings, 2002 ACM SIGKDD International Conference on Knowledge Discovery in Databases (KDD'02)*, Edmonton, Canada, 32-41

Tesfatsion, L. (2009) *Agent-Based Computational Economics: Growing Economies from the Bottom Up*. Website; http://www.econ.iastate.edu/tesfatsi/ace.htm. Last accessed 2 April 2009.

Tobler, W. R. (1970) "A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–40.

Tobler, W. R. (1994) "Bidimensional regression," *Geographical Analysis*, 26, 187–212.