ICA Commission on Visualization: Working Group on Database-Visualization Links.

The Integration of Geographic Visualization with Knowledge Discovery in Databases and Geocomputation

Mark Gahegan [1], **Monica Wachowicz** [2], **Mark Harrower** [1] and **Theresa-Marie Rhyne** [3].

1 Department of Geography, 302 Walker Building, The Pennsylvania State University, University Park, PA 16802, USA. Email: mark@geog.psu.edu, mah282@psu.edu
2 Geo-Information Center, Wageningen UR - CGI, The Netherlands. Email: M.Wachowicz@sc.dlo.nl
3 Lockheed Martin Technical Services, US EPA Scientific Visualization Center, 86 Alexander Drive
Research Triangle Park, North Carolina 27711. Email:

Abstract

This paper details the research agenda of the International Cartographic Association commission on Visualization: working group on Database-Visualization links. The paper stresses the need for the closer integration of three largely disparate technologies: geographic visualization, knowledge discovery in databases and geocomputation. The introduction explains the meaning behind these terms, the ethos behind their practice and their connections within the broad realm of knowledge construction activities. The state of the art is then described for different approaches to knowledge construction, concentrating where possible on visual and geographically oriented methods. From these sections, a research agenda is synthesized in the form of three sets of research questions addressing: (1) visual approaches to knowledge discovery, (2) visual support for knowledge construction and geocomputation and (3) databases and data models that must be satisfied to make visually-led knowledge construction a reality in the geographic sphere of interest. Conclusions relate this agenda to issues of (1) data, (2) geographic knowledge and (3) the visualization environment and pose significant challenges to the way we currently represent geographic information and knowledge.

1 Introduction

The International Cartographic Association (ICA) has established a Commission on Visualization (http://www.geog.psu.edu/ica/icavis/ICAvis_overview.html), and associated research agenda. To support the agenda, this paper addresses the use of visualization within exploratory analysis, data mining and geocomputation, with the overall focus directed to the task of knowledge construction. Background information about this and other ongoing ICA visualization research initiatives may be found at http://www.geog.psu.edu/ica/icavis/poland1.html.

This paper is organized as follows: the first section introduces task of knowledge discovery using visualization and geocomputation tools and motivates interest in this process, the second section defines the state of the art in the joint areas of interest and the third section distills the research challenges arising from them. These challenges are summarized and framed in a broader context in the concluding section. The areas of interest covered here are: (1) data mining and exploratory data analysis, (2) knowledge construction and (3) database and data model issues. These themes are approached from the perspective of scientific inference, specifically the types of inferential mechanisms that knowledge discovery requires, the degree to which they are supported in the various tools and methods described and the roles played by the system and the user.

1.1 The Problem

Geographic datasets continue to become more complex—many areas of geographic analysis now have access to vast digital datasets, and conversely many conventional datasets now contain either explicit or implicit spatial references, providing huge and previously untapped data resources.

Furthermore, and for the first time, integration of data from different scientific and business communities is becoming a practical reality. To address social and environmental concerns, geographers are linking disparate datasets together across place, scale, time, theme and discipline and are now beginning to ask new types of questions that were hitherto not possible. The resulting databases are *rich* in terms of attribute depth and *large* in the sense of having many records or objects represented. Consequently, uncovering and understanding real-world patterns, or real-world processes that can be found in the data stored in these databases presents a difficult challenge; they are vast, dense and may span many different areas of expertise.

1.2 Contributing Technologies

1.2.1 Knowledge discovery

Knowledge discovery stems from the computer science and very large databases (VLDB) communities, where it has experienced an intense research focus of late (as evidenced by many new conference series and journals). The aims of knowledge discovery are very similar to those of visual approaches to exploratory data analysis: i.e. to find useful and valid structure in large volumes of data, and to provide some means of explaining it (Fayyad, 1996a). Fayyad *et al.* go (1996b) describe the knowledge construction process as comprising five stages: data selection, pre-processing, transformation, data mining and interpretation/evaluation. These stages progressively refine a large dataset to the point where it makes sense to propose object structures and relationships, they also decompose the general phases shown in Figure 1 into specific tasks.

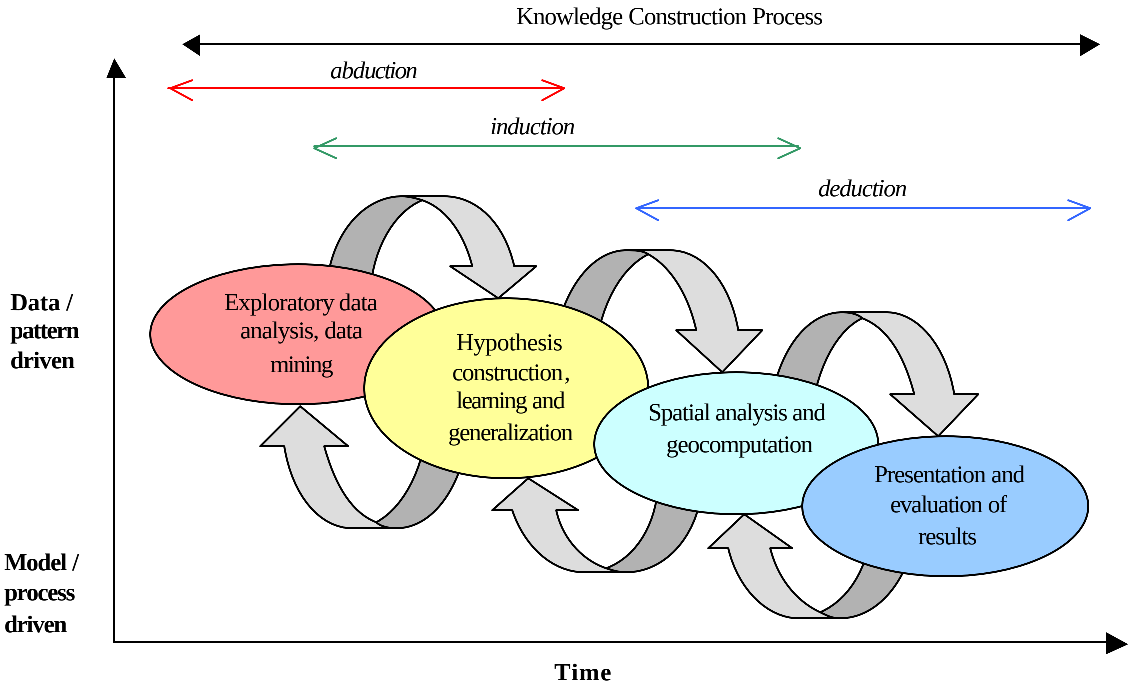


Figure 1. Different phases in the scientific method and the modes of scientific inference that apply to each phase.

Knowledge discovery methods and their implemented tools can help us to extract or uncover knowledge from very large and (possibly poorly-understood) databases. Before such databases can be utilized in the traditional sense, we must first extract from them useful patterns, objects, events, categories and structures with which to construct a model of the relationships between patterns and the processes or events they represent. Knowledge discovery subsumes the field of data mining, which can be viewed as comprising the early tasks of pattern recognition. Knowledge discovery methods (induction, compression, querying, and approximation) now facilitate the next step in understanding what makes one pattern structurally different from another, which can lead to the 'discovery' of useful trends or characteristics that might be associated with higher level entities, relationships or processes within some domain of interest. We can view KDD as a whole knowledge construction process with a series of interconnected phases linking together precursory activities with analysis and evaluation. Figure 1 shows the phases of the scientific method (e.g. Leedy, 1993); beginning with a focus on the data and the development of working hypotheses; these lead in turn to the construction of knowledge (objects, events, rules, categories, etc.) around which analyses can be built and results produced and evaluated.

Chen *et al.* (1996), and Rainsford and Roddick (1999) give useful reviews of approaches to data mining. The massive digital datasets now being gathered for a wide variety of applications such as marketing, sales, telecommunications, epidemiology, medical diagnosis, investment trading and geography pose some difficult problems (Koperski *et al*., 1999; Roddick & Spiliopoulou, 1999). Apart from their sheer size, little is known regarding the patterns and relationships embedded within them (e.g. Elder and Pregibon, 1996). The premise is that if these can be uncovered, they might offer significant insight into complex domains that might result in commercial advantage, improved decision making or deeper scientific insight. Candidate datasets are considered too large for textual browsing and thought to contain structure that is unexpected or hitherto unrecognized. Of late, data mining activities have focused on geospatial data as a rich source of structure and pattern (e.g. Koperski & Han, 1995; Ester *et al*., 1996, 1998; Knorr & Ng, 1996; Koperski *et al,* 1999). Most of the data mining activities are related to tasks that are dependent on the intended outcome of the overall knowledge construction process. Some examples are: prediction, classification, clustering, and association. Different methods are used to perform these tasks, such as: statistical association, case-based reasoning, neural networks, decision trees, rule induction, Bayesian belief networks, genetic algorithms, fuzzy and rough sets theory (see Mitchell, 1997 for details of the workings of these methods).

1.2.2 Geocomputation

The remit of geocomputation is perhaps wider than that of KDD, since it also addresses spatial analysis, but nevertheless also includes many examples of knowledge construction and data mining methods applied in geography. Geocomputation aims to "enrich geography with a toolbox of methods to model and analyze a range of highly complex, often non-deterministic problems" (Gahegan, 1999a). In practice this often involves a learning or mining stage in the analysis of a complex dataset (e.g. Openshaw *et al*., 1990; McGill & Openshaw, 1998; ***others***). There is mounting evidence that traditional (deductive and statistically based) approaches to analysis do not scale well to very large and rich datasets, partly because of the complexity of the attribute or feature space and partly as a result of the computational inefficiency of existing methods (e.g. Landgrebe, 1999; Gahegan, 2000b). Perhaps the major defining difference between geocomputational and knowledge discovery methods thus far is that the major focus of geocomputation is on providing solutions to given problems, and not the discovery of knowledge per-se. Many of the methods employed in geocomputation (such as neural and Bayesian networks, cellular automata, decision trees and genetic algorithms) are effectively 'learning tools'. However, since geocomputation is not usually concerned with distilling the implicit knowledge that may be contained in these tools after use (e.g. the positions of hyperplanes in a neural network), there is usually no way in which it can be saved and re-used in related circumstances. What geocomputation can offer is a wealth of experience in applying and configuring such tools within a geographic setting.

1.2.3 Exploratory visual analysis

Geographic visualization (geovisualisation) is already well-established for presentation and evaluation activities in many disciplines, and cartography is certainly no exception (e.g. MacEachren & Taylor, 1994; Hearnshaw & Unwin, 1994; MacEachren, 1995). However, geovisualization can also play an important role as an intermediary between the human and the machine, by facilitating data exploration tasks and improving human-computer interaction. Visually-led approaches to data exploration often employ sophisticated graphical techniques to uncover structure in data. A typical methodology consists of four main stages, namely: exploration, confirmation, synthesis, and presentation. In this case the goal of portraying the data in visual form is to stimulate pattern recognition and hypothesis generation rather than simply presenting a result or outcome. Visual representations might allow us to construct knowledge from visual displays, whether in propositional form (suggesting something for consideration), analogical form (explaining something by comparison), or procedural forms (defining an action or a set of actions necessary for doing something).

Visual approaches to exploratory data analysis have simultaneously arisen in many different scientific communities and are now well-established. The database community has coined the phrase 'visual data mining' to describe their efforts (e.g. Keim & Kriegel, 1994; Card *et al*., 1998; Ribarsky *et al*., 1999), while the terms *exploratory visual analysis* or *exploratory data analysis* are more common within the statistics community (Tukey, 1977; Chernoff, 1978; Asimov, 1985; Tufte, 1990; Haslett *et al*. 1991; Mihalisin *et al*., 1991). In geography, the term, *Exploratory Spatial Data Analysis* (ESDA) is perhaps in most common use, and describes the application of exploratory techniques (including visualization) in a specifically geographic setting. However arrived at, these strands are largely convergent, aiming to capitalize on the pattern recognition and hypothesis generation abilities of human experts. For sighted humans, the visual senses have a unique status, offering a very broadband channel for information flow, excellent pattern recognition capabilities and access to a rich cognitive structure from which hypotheses, theories and explanation can be imparted. Hence, visual approaches to analysis and mining attempt to harness of our abilities to perceive pattern and structure in visual form and to make sense of, or interpret, what we see.

Visualization has been used as the vehicle for exploration in the work of Monmonier (1990), MacDougall (1992), Tang (1992), Fotheringham and Charlton (1994), Gahegan (1996), Dykes (1997), MacEachren and Kraak (1997), Wise *et al.* (1998), Kraak and MacEachren (1999), MacEachren *et al*. (1999) and many others. Recently, visualization tools to aid specifically in the early data mining activities have also been proposed and developed (e.g. Lee & Ong, 1996; Keim & Kriegel, 1996; Keim & Herrmann, 1998). Useful overviews of visual data mining are provided by Wong (1999) and Hinneburg *et al.* (1999).

So far, limited research has been carried out on the use of visualization as enabling technology, and possibly as the interface, for a complete knowledge construction process, although some visually-supported methods have been proposed (e.g. MacEachren *et al*., 1999; Hao *et al*., 1999; Inselberg & Avidan, 1999). One possibility, Query By Pictorial Example (QBPE), is described by Rhyne (2000). Visually-based knowledge construction can offer two distinct advantages. Firstly, it creates an opportunity for humans and machines to work together in constructing and evaluating the objects and relationships required by the analysis phase, making the best use of the abilities of each. Secondly, it can provide an environment within which expert scientists can collaborate (sometimes remotely) on complex modeling and analysis activities (e.g. Brodlie *et al*., 1998). In both ways, visualization can play an important role in “process-pattern tracking” (visual representations that display key aspects of a process as it unfolds) and “processing steering” (interactive environments that provide controlling parameters of a knowledge construction process to shape and modify its behavior).

The major objective of this paper is to investigate the possibilities of applying geographic visualization into each of the problem-solving phases of knowledge discovery, specifically as they relate to geography, and from this to propose a suitable research agenda. Connections between generic knowledge discovery and information visualization are described by Card *et al.* (1999), this paper extends some of those thoughts and ideas into the geographical realm.

1.3 Modes of Reasoning: Inference in Knowledge Construction

The focus in data mining (which sets it apart from traditional query-processing in very large databases) is on detecting the unknown or unspecified. Where a specific data question is known *a-priori* then the ‘mining’ metaphor does not apply and the problem becomes a question of computing the state of some given object, for example if a condition is true or not. The broader process of knowledge construction may use the outcomes of mining or may begin with a more formed data question.

One goal of knowledge construction is to build tools that can function in the absence of pre-determined hypotheses, training examples or rules, i.e. that utilize an abductive form of reasoning.

Abduction is flexible because it is not restricted to using existing knowledge structures (e.g. categories) but is instead free to create new structures that help to explain the data presented. However, not all visualization, data mining and knowledge construction methods operate in this way, most either attempt to bcate pre-defined patterns (*deduction*) or else learn from examples that are presented or selected (*induction*) (Gahegan, 2000a; Baker, 1999).

Abduction can be viewed as the most flexible inference mode because it requires neither the target nor the hypothesis to be pre-defined. It is therefore highly suited to the initial exploratory phase, especially if little is known concerning likely structures in the data (e.g. when searching for a hypothesis to test analytically). Induction becomes more useful when examples have been identified and is arguably the most reliable means of knowledge construction due to the availability of many robust, automated methods developed by the machine learning community (see Mitchell, 1997 for many examples of inductive tools). Deduction can only be applied when objects and categories are already defined, and usually forms the basis of most inferential analysis and modeling because it can be verified straightforwardly[1]. For example, expert systems tend to be deductive whereas decision trees use inductive learning (Simoudis *et al.*, 1996). The colored arrows at the top of Figure 1 show the appropriateness of these three types of reasoning to the phases in scientific process.

Ramakrishman (REF!) presents a taxonomy in which a knowledge construction process is described according to four perspectives: induction, compression, querying, and approximation. The initial stage involves learning or discovery, the latter stages are concerned with generalizing and validating the findings. Since computationally-based approaches to knowledge construction do not have access to the richness of understanding and background context that human experts do, there is little surprise that abduction is missing from this list. Perhaps the single most important factor that sets visualization apart from other knowledge discovery methods is the connection to the expert as a rich source of interpretation for the uncovered structure, thus opening the way for abductive reasoning to take place.

2 State of the Art

This section reviews the current state of the art in three distinct areas: (1) direct visual approaches to data exploration and data mining, and the variety of different visual techniques that have been proposed (2) the support of knowledge construction and geocomputation through interactive and collaborative means and (3) the effects that choices over databases and data models ***and integrated knowledge construction environments*** can have on the outcomes of knowledge construction activity.

2.1 Visual Approaches to Data Exploration and Knowledge Construction

2.1.1 Techniques

Several distinct types of visual data mining and exploratory analysis techniques can be identified. From the perspective of data mining, Hinneburg *et al*. (1999) describe four: geometric projection, iconographic, pixel-based and hierarchical. When considering also those techniques used by geographers and statisticians involved in exploratory analysis chart-based and map-based techniques can also be added. A brief description follows of some of the more common families of techniques; greater detail is given in Gahegan (2000b). Notice that there is no consistency to the way these various groups are defined; some are named after their data representation methods (e.g. map-based and hierarchical techniques) and others by methods applied to this representation (e.g. projection techniques).

Map-based techniques allow the mapped data and its visual appearance to be changed interactively (Dykes, 1997). Map legends are often used as the basis for interaction, as shown by Peterson (1999) and Andrienko and Andrienko (1999), permitting the user to change the appearance of the objects

---

[1] It is interesting to note that models and analysis performed with deduction are often defended vigorously in statistical terms. But in many geographic applications, the objects and relationships used in the models are often the products of abductive and inductive inference, which cannot be defended in the same way.

mapped and thereby define and possibly explain clusters. *Chart-based techniques* plot the data on a chart or graph, common examples being scatterplots and parallel coordinate plots. Scatterplots use a simple 2D or 3D graph with dots or spheres to mark the position of individual data items (Cleveland & McGill, 1988). Parallel coordinate plots employ a (usually larger) number of parallel axes through which a trace of each data item can be made (Inselberg, 1986; 1997). These techniques are often accompanied by linking and brushing methods, allowing selected data points to be viewed in different ways or within different axes (e.g. Buja *et al.*, 1991, 1995; Edsall, 1999). *Projection techniques* use statistical transformations such as principle component analysis and, multi-dimensional scaling to project structure or trends from the data (e.g. Asimov, 1985; Haslett *et al*., 1996; Cook *et al*., 1995). They are also often based around graphs, particularly scatterplots, so most examples could be seen as building on the chart techniques defined above. *Pixel techniques* map data values to individual pixels that are ordered on the screen so that data-streams of similar values produce visible clusters in 2D (Keim, 1996; Jerding & Stasko, 1998). The screen can be divided into separate windows if several attributes are to be visualized (Keim & Kriegel, 1994; 1996) Such techniques may present a useful overview of a very large dataset. *Iconographic techniques* use complex symbols, such as stick figures (Pickett & Grinstein 1988) or faces (Chernoff, 1973; Dorling, 1994) to encode many data dimensions simultaneously. The aim of iconographic displays is to promote perception of the 'whole' while still allowing some differentiation of individual variables. *Hierarchical and network techniques* strictly organize data according to a specific data structure, such as a tree (Robertson, 1991) or network (Huffaker *et al*., 1999; GeoBoy®, http://www.ndg.com.au/), with progressive levels refining the display into subspaces.

Figure 2 shows four different approaches to exploratory visualization, conducted on the same dataset. Note that three of these rely heavily on the map paradigm, using real and artificial landform surfaces as the dominant visual metaphor.

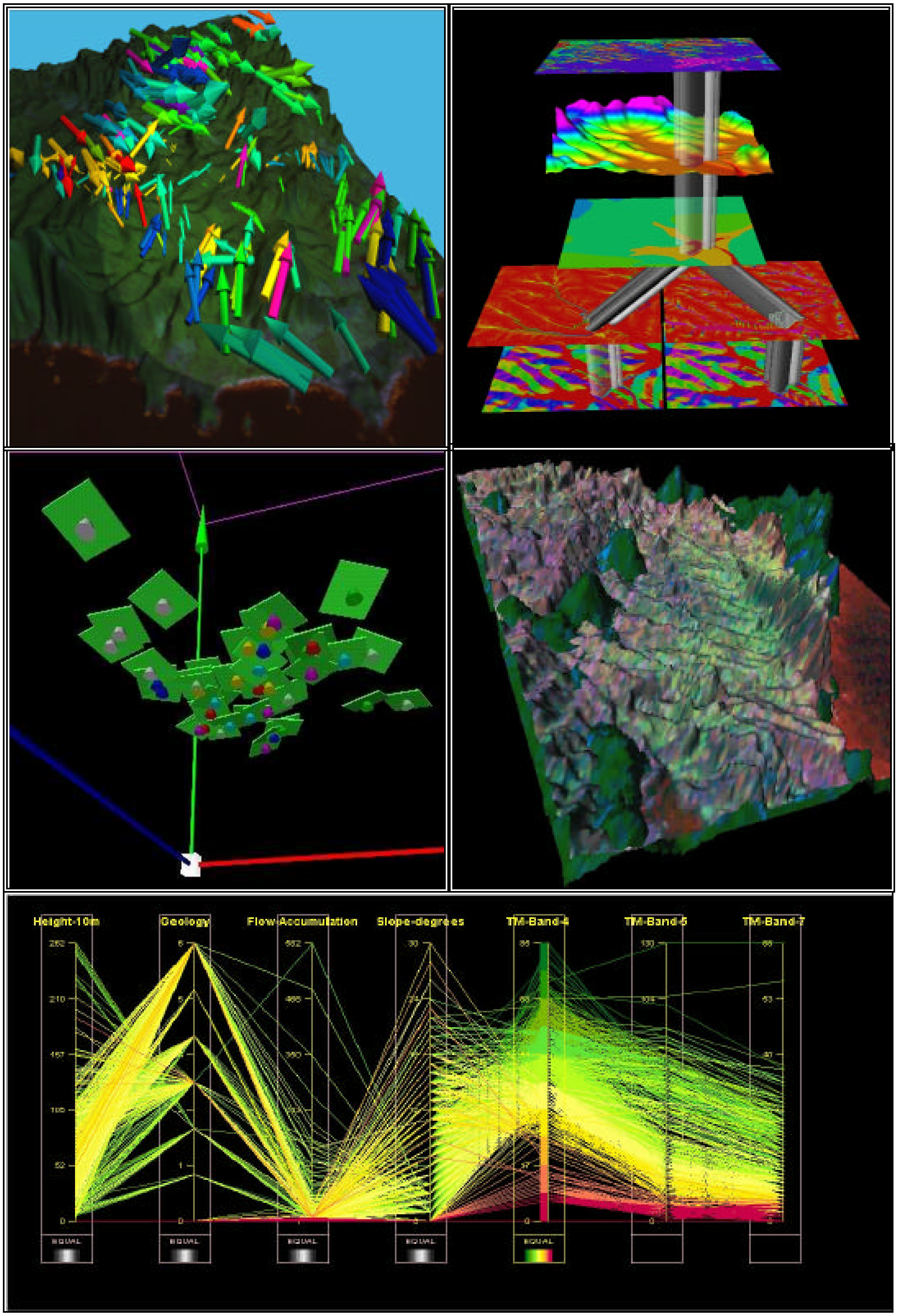


Figure 2. Five scenes depicting a range of visualization techniques applied to the same data (an environmental dataset of a coastal region of New South Wales, Australia). Top left, mark composition using arrows draped on an elevation model. Top right, 'interactors' describing relationships between different data layers. Mid left, a scatterplot, enhanced with planar icons to

encode additional information. Mid right, two different environmental surfaces are dynamically 'intersected'. Bottom, a parallel coordinate plot, showing the distribution of attribute values in seven dimensions.

2.1.2 Reasoning and Inference

One important consideration when proposing visualization to support data mining, exploratory data analysis and knowledge construction relates to the type of inference that is to be employed (described above in Section 1.3). Let us assume for the moment that the mode is to be abduction. The aim of abduction is to discover patterns within the data while simultaneously proposing a hypothesis by which the patterns might have come to be. In computational data mining, the hypothesis is usually constructed as some kind of classification scheme or set of association rules (e.g. Han, 1999; Agrawal *et al*., 1993) by which useful structure can be imposed on the data. For example, AutoClass performs repeated unsupervised Maximum Likelihood classification to (a) discover categories within the data and (b) discover feasible examples by which these categories might be defined. In some forms of visual exploration an abductive task is performed collaboratively between the observer and the visualization; patterns are observed as a consequence of how scene is constructed and rendered and how the observer perceives and comprehends it. The simultaneous task of hypothesis generation is also similarly split, the mappings used to create the scene themselves convey a hypothesis and an observer may generate one or more internal theories to explain the observed structure.

If induction is to be used instead, then some form of target must already have been identified, and the task becomes one of uncovering causes or related hypotheses to *explain* the examples. Consequently, induction is most commonly applied during the later stages in knowledge construction (see Section 2.2), following after category construction and requiring training examples. In the case of deduction, the knowledge construction activity is entirely procedural and is not influenced by individual characteristics of the data under consideration. The role of the user is somewhat reduced since both targets and explaining hypotheses are externally provided.

Visual approaches to data mining and exploratory data analysis differ in the emphasis they place on the observer and the system and the roles that each are expected to play. For example, pixel-based and projection methods tend to use pre-defined transformations applied to the data, and these also explicitly define a hypothesis by which any observed pattern might be explained. If the transformations are entirely pre-defined, in other words they cannot learn from or be changed by the specifics of the data under consideration, then the system is operating by deduction. The observer may not be, however. What a user notices in the scene and how they chose to interpret it is not defined (although it may be severely biased). Alternative visualization methods, such as the linked views in scatterplots and other dynamic exploratory techniques, offer a less deterministic structure, since the data is controlled less by the system and more by the user (and possibly the data itself). This provides greater flexibility, and less external bias to the perception and reasoning process. Abductive methods may seem more appealing but they can also be less reliable because the visual patterns produced may be irrelevant, anomalous or misleading, and there is no way of guaranteeing their statistical significance, since that would presuppose the existence of a model (e.g. a deductive approach). Kenwright *et al*. (1999) discuss some of these different perspectives from the standpoint of visualization.

2.2 Visual Support for Knowledge Construction and Geocomputation

Once interesting patterns have been discovered within the data they must be represented and stored for future use (in the analysis phase). As described earlier, data mining is often subsumed within the larger task of knowledge discovery (e.g. Piatetsky-Shapiro *et al*., 1996), for example to discover new 'facts' or to make new inferences about the data. The goal of knowledge construction is usually to 'learn' and generalize findings to a form that makes them useful for defining some entity such as a category or a process. Construction can be targeted at many different types of knowledge, including: discriminant rules, characteristic rules, association rules, functional relationships and dependencies and various other forms of clusters.

A visualization approach might bring two advantages to knowledge construction. The first is the high degree of interaction, allowing the user to explore data from several perspectives and steer the process (e.g. Ankerst *et al*., 1999). The second is the rich visual environment, allowing many dimensions of data to be viewed concurrently. Within such a setting, one might envision *geo-agents*—autonomous pattern analysis methods that attempt to learn structure within the data (e.g. MacGill & Openshaw, 1998). Geo-agents could be abductive or inductive; that is, they could attempt to *uncover* patterns and bring them to the user's attention or could be used to *explain* patterns that the user has observed. A highly interactive visual environment might prove to be a very intuitive setting for defining training examples from which to construct classes, for example, or otherwise explaining category definitions (Inselberg & Avidan, 1999).

2.2.1 Reasoning and Inference

Knowledge construction requires that example patterns have been previously identified. The primary mode of inference used is therefore induction, and specifically generalization from a specific set of observed patterns to a more widely useful, general form. In machine learning, this is often referred to as meta-learning or 'learning about learning' (Haralich, 1994). If patterns were extracted via inductive or deductive means, then a generalized form may already be available in the extraction methods themselves. Failing that, hypotheses, categories, objects or processes will need to be constructed to represent them. In either case, knowledge construction can also include a 'refinement' stage, where evaluation or significance testing takes place. In computational knowledge discovery, knowledge representation must take a symbolic form, defined in terms of classes, objects, attributes and relations (e.g. Quinlan, 1993; Gains, 1996). However, there is no reason why the goal should not be to instill knowledge in the observer as well as, or even instead of, the computer. The relationship between knowledge construction and data modeling is taken up in the following sub-section.

Geocomputational approaches to analysis also often begin with a learning stage, in the form of a classification or function approximation task, and again it is usual that some *a-priori* training examples are available. In this case the learning task can also be viewed as one of knowledge construction, since category descriptions or functional relationships are being learned.

2.3 Databases and Data Models

A knowledge construction environment consists of tools that can perform a range of tasks. There are over 200 such tools currently available in the public domain (see www.kdnuggets.com/software), ranging from data preparation, classification, visualization and web mining, to clustering and other forms of discovery tasks. However, they have so far not had much impact on spatio-temporal databases. The methods implemented in these tools are often not 'spatially aware' and where they are, they use very simple data models of spatio-temporal objects and relationships; for example, snapshots of point objects and Euclidean distances. Other complex spatio-temporal objects (e.g. moving points, lines, polygons) and their respective relationships (e.g. direction, connectivity, non-Euclidean distances) need also to be integrated into a knowledge construction environment. This requires a full range of conceptual, logical, and physical database models of spatio-temporal objects.

Understanding the underlying data model or schema is crucial in a knowledge construction environment and users may gain new insights when inspecting, interpreting, and validating them. Supporting tools to examine models and schema will require advanced algorithms and user interfaces to be built outside of the databases (Wachowicz 2000c). A visually-led approach to data mining and knowledge construction implies the integration of large geo-spatial databases (from well-structured vector and raster models to unstructured models such as georeferenced multimedia data) with artificial intelligence tools and visualization functionality. This integration can take many forms (Rhyne, 1997; Slocum, 1994), the extremes being a monolithic, single system and a component-based, loose federation. Hao *et al*. (1999) describe a visual data mining infrastructure, implemented using a Java component architecture. Gahegan *et al.* (2000) show how such an architecture can provide useful exploration of the data, leading to improvements in category construction.

Other system issues that may affect knowledge construction relate to the richness of models and concepts in the underlying database(s), specifically, their organization and the facilities they can provide for representing knowledge once it has been discovered (e.g. Sheth, 1997; Drew and Ying, 1998, Sowa, 1999). The most important missing functionality in current databases is 'data modeling management', allowing users to update the model as the process of knowledge construction evolves and to monitor these changes over time as objects, categories and relationships are uncovered. For example, when a decision tree is used to inductively learn a landcover category or predict a stream outflow, it is also generating new knowledge in the form of rules. This knowledge also belongs within the database, and as a result, the schema must be updated to accommodate it. Technically, it is an enormous challenge to determine how to incorporate this functionality into a database.

3 Research Challenges

This section considers the research challenges arising from the application of knowledge construction methods applied in geography and cartography. The means to achieve effective knowledge discovery is still largely an open question. Many of the challenges described are shared across a number of scientific disciplines, but others are unique to a geographic or cartographic setting, and relate either to the unique properties of spatio-temporal information or to the complexity of defining categories, objects and relationships in the natural sciences.

3.1 Visual approaches to knowledge discovery: Research Questions

Although data mining methods have mainly be developed for non-spatial data sets, their use with spatial and spatio-temporal data have been proven to be complementary (Wachowicz 2000a, 2000b). Some approaches have specifically concentrated on spatial, temporal or spatio-temporal datasets (e.g. Koperski *et al*., 1996) and a useful online bibliography is provided by Roddick and Spiliopoulou (1999). The following research questions result from the integration of visualization and spatial data mining, i.e. the 'early tasks' of locating interesting structure within the data as described in Section 2.1.

1. *From the perspective of the scientific method, what are the different tasks involved in a knowledge construction process applied to a geographical database and which visualization techniques are the most suitable for them?*

   This question explores the range of tasks that might be useful for process-pattern tracking early in the knowledge construction sequence (e.g. Fayyad, 1996b), and stems from the observation that distinct tasks might benefit from different visualization techniques, as the above discussion shows. For example, a scatterplot is very useful for determining outliers or data anomalies, whereas a projection via principle components or multi-dimensional scaling might be more useful to find global patterns or define a category. Other 'early' tasks may require data cleaning, data reduction, and data transformation, all operating on huge data volumes. New visualization methods are needed to support these tasks also.

2. *To what extent do the different visualization techniques affect logical inference?*

   The description in Section 2.1 also shows that visualization techniques offer a variety of styles of interaction, degrees of participation and support different modes of inference. If the strengths and weaknesses of individual techniques were well understood, it would be more straightforward to match them to the tasks uncovered in Question 1. The eventual goal here is to develop an integrated suite of interactive, visual techniques, all suited to specific tasks and logically flowing one from another, as described by MacEachren *et al*. (1999).

3. *How can developments in human computer interface design aid in knowledge construction?*

Many different types of interaction devices are now available that might improve utility or productivity of the interface, including: gesture detection for selecting, panning, zooming, rotating, and other manipulations of mining objects, and eye tracking. Some examples of new possibilities include:

- physical tools (e.g. markers and erasers) and virtual tools (lenses and sticky notes) for allowing users to concentrate on the task at hand;
- eye-contact (gaze) for tracking of head and eye movements on specific areas or objects of interest within a visualized scene or immersive environment;
- voice recognition and natural language processing (NLP) for improving the communication between the knowledge construction processes and their users.

4. *What kinds of visual encoding and visual metaphors work best for particular tasks, applications or data?*

This question searches after expertise and knowledge (either supplied by an expert or learned from the data) that can improve the effectiveness of scene construction (e.g. Robertson, 1991; 1997) and choice of visualization technique, e.g. from those described in Section 2.1.1 (Senay & Ignatius (1998). For example, in what situations is the *landscape metaphor* the most useful means of portraying a geographical dataset?

5. *To what degree should we rely on automated reasoning in producing the visual scenes used for data exploration (automated scene construction) and what implications does reliance bring?*

A related issue concerns the degree to which the presentation of the data is pre-determined or biased. Predefined methods are repeatable, more easily interpretable and convey an explicit hypothesis, but they are also inflexible and may overlook patterns that were not anticipated and possibly those that are not global. See the discussion above on inference for examples.

6. *If hypotheses are used (explicitly or implicitly) by the machine in scene construction, how should they be communicated to the observer, or utilized in the knowledge construction process?*

Another related issue is the role of geographic visualization in presenting the *a priori* knowledge of a domain expert in such a way that it will help the understanding of the outcomes of analysis (i.e. the knowledge used to set up and drive a visual data mining technique). In order to be transformed into useful knowledge, observed structure in the data must be related to a supporting hypothesis of some sort. Visual projection techniques, for example, might supply a transformation equation, whereas more ad-hoc methods might only *imply* a hypothesis by way of the data used and the visual encoding by which it is rendered. An observer or a machine will need to understand this hypothesis to 'learn' it. If the transformations and visual encodings used are complex, then the observer may well not understand the significance of a particular observation. More research is needed to determine how this kind of knowledge should be visualized effectively.

7. *How does the presence or absence of different data sets affect the success of visual data mining and what is the optimum arrangement of data to locate a given object or relationship?*

One of the preliminary steps in data mining is to apply filters to remove anomalies or outliers in the data, or pre-processing methods such as principle components analysis (PCA) to reduce data volume. This begs the question as to what should be retained and what ignored. If the aim is to discover major trends in the data then global techniques such as PCA will perform well, but if the goal is to uncover localized anomalies, such projections might actually remove them; locally adaptive techniques might meet with greater success. Likewise with noise; in the absence of a model it is not possible to know for certain what artifacts in the data are interesting and what are misleading. Choosing the most useful data layers, and pre-processing them in a suitable manner are therefore critical to success.

8. *To what extent, and for what tasks, must a data mining process consider the spatial and temporal properties of the data along with other attributes?*

This key area of concern relates to the nature of geographic data, and the intrinsic notion that location matters, because space forms the framework that both defines patterns and determines their significance. For many tasks, effectiveness will depend on the simultaneous presentation of attribute values and their spatio-temporal location, forming some minimal context by which data is given meaning. This fact should affect the design of, or selection of, effective visualization methods; they must be able to display at least three variables for spatial datasets (*x*, *y* and *attribute*) concurrently and four variables (*x*, *y*, *t* and *attribute*) for spatio-temporal datasets, preferably in an integrated manner. A deeper and related question involves the general problem of using only attributes to identify objects of interest. Humans use relationships and other expertise that extends beyond simple property values when identifying objects or defining categories (e.g. Rosch, 1973). Most knowledge construction methods do not have very sophisticated mechanisms to replicate this expertise, nor indeed is it straightforward to gather them from a human expert nor represent in a machine.

9. *How should exploratory visualization techniques take advantage of newer developments in information visualization?*

Improvements to the visualization systems available continually change what is possible, and how much the required hardware will cost! Consequently, this question will remain open for some time. Newer low-level examples of improvements offering benefits include the ‘level of detail node’ in VRML and the dynamic properties (behavior object) of scenegraph variables in Java3D. These provide opportunities to enrich user interaction with the scene and to render more data dimensions without sacrificing effectiveness. Future possibilities include QBPE (section 1.3), virtual environments, 3D-web applications, visual agents, and visually-led forms of collaboration with humans and/or machines.

3.2 Visual Support for Knowledge Construction and Geocomputation: Research Questions

The use of visualization as a tool for knowledge construction has few practical examples at present. Consequently, these research questions are more speculative and represent a picture of what might be possible, rather than building on a large body of established work and reflecting on how it might be improved (as in the previous section).

1) *Can a taxonomy of knowledge construction operations be constructed to relate tasks users are trying to accomplish with appropriate tools and functionality?*

Knowledge construction is a process that reflects not the end result a user would want, or what the system should produce, but rather the operations a user will perform using the tools of a knowledge construction environment. This is an important distinction because it places the user first in the knowledge construction process, but it also highlights the semantic 'gap' between desired outcomes and the available tools. Existing knowledge environments focus on identifying what kind of knowledge the system needs to discover successfully. However, it is very important to take the design initiative of enabling the users' expertise within a knowledge construction process, rather than attempting to supplant it.

2) *What kinds of visual interaction and metaphor are useful to support the knowledge creation process?*

   This question is designed to investigate how knowledge discovery and data mining can be portrayed visually. The flow of activities needs to reflect a structure for knowledge creation, such as that given in Figure 1 and requires a visual representation that a user can understand and work with.

3) *How should knowledge construction methods (e.g. geo-agents) be configured by visual interaction with the user?*

   Continuing on from the previous question, the visualization environment must facilitate a two-way exchange of knowledge. The user might be able to highlight examples in the data, select interesting or relevant parts of the dataset, refine the outcomes of machine learning and perhaps select appropriate data mining methods or goals. Conversely, the results of mining and knowledge construction must be conveyed to the user in a form that encourages assimilation and promotes understanding. For example, to define the functionality required for the user to point to a region in the visual space and say: “Learn that description, then go and find more.”

4) *How might human and computational methods for knowledge discovery be integrated in a collaborative (human-computer) strategy for exploring data?*

   Sometimes human experts may know best how to recognize or extract knowledge, but at other times a 'brute force' computing approach might be more fruitful. Collaborative approaches will need to work out how to arbitrate between expert opinion and machine learning and which to favor in a given situation.

5) *In what way do various expert configuration and collaboration strategies affect knowledge discovery?*

   The challenge here is to empirically demonstrate that a combination of human and machine 'intelligence' really does improve knowledge construction capabilities and to build environments that take advantage of the best that each has to offer. For example, the *a-priori* knowledge of a domain expert is difficult to pass on to a data mining algorithm. Conversely, the results of data mining are biased by the search criteria used, but these too might be difficult to communicate to the expert.

6) *What effect does the visual environment have on the knowledge discovery process?*

   As environments become more immersive and better integrated with databases and analytic functionality, their effectiveness for certain tasks may increase. One particular issue of interest is the degree to which visual realism helps or hinders understanding in various exploratory settings. As workstation and Internet technologies become more advanced, new possibilities for knowledge construction may arise. Distance or remote collaboration of experts within virtual environments is one such possibility.

7) *Do users working collaboratively have more or less success in knowledge construction, and how does their pooled expertise affect knowledge discovery?*

   When experts collaborate, group dynamics play an important role in determining outcomes. Tools to visually represent knowledge are necessary so that it may be shared, or examined from another's perspective. Also, pooled knowledge may have conflicts, tensions and biases within it, and these too will need to be visually represented.

3.3 Databases and Data Models: Research Questions

This set of questions addresses the more practical, engineering problems of bringing together disparate technologies, each with their own established tools, systems, data structures and interfaces, as described in Section 2.3.

1. *What kind of computational architectures are most efficient and effective for integrating standard databases with visualization environments?*

   Systems integration is becoming more straightforward, and close physical integration is no longer a pre-requisite for logical integration (Abel *et al.*, 1994; Bishr, 1997). Environments such as Java can provide platform independence, Internet access and simplify the exchange of data, functionality and control between collaborating applications (e.g. Gahegan *et al.*, 2000). However, there are some performance bottlenecks that still require attention (Bryson *et al*., 1999; Gahegan, 1999b). Data mining is inherently data intensive, so any visual approaches are likely to be so too, making considerable demands on graphics hardware and information retrieval structures. If visual data mining is to be interactive, then these demands are likely to increase.

2. *What effect does the underlying structure or representation of the data (the data model) have on the process of knowledge discovery?*

   The data that we visualize is seldom (one could argue 'never') free of some form of interpretation or conceptual model. These underlying models in part determine what we are able to observe within the data and thus also affect the conclusions we may and may not be able to draw. A related question is: "How should this 'model bias' be presented to the user?"

3. *How can the objects we discover be worked back into a consistent data model?*

   This question relates to closing the loop between the data model, the visualization environment and the cognitive systems of the user. Objects we discover are likely to be at a more abstract level (semantically) than those already represented in the database, and may require new class descriptions or other conceptual structures (e.g. Sowa, 1999; Luger and Stubblefield, 1998) to be created before they can be stored within the database. In turn, this may require changes to the underlying data model (model evolution). Data models, such as those in current GIS may also need to increase in semantic richness so that they can represent the structures uncovered.

4 Conclusions: Priority Areas for Research

The 1999 ICA Conference (Ottawa, August 14th to 21st) (http://www.ccrs.nrcan.gc.ca/ica1999/) hosted working group sessions for the ICA Commission on Visualization The *Database-Visualization Links* group summarized its findings in this way:

> "*Data models are the heart of a geographic visualization system. Knowledge discovery techniques and related fields empower the exploration engine and facilitate geographic visualization. Thus, research is needed to establish linkages between these fields. We believe that knowledge discovery through data mining, query and design of databases and geocomputation are necessary capabilities of geographic visualization. Thee notions need to be further explored in a research agenda.*"

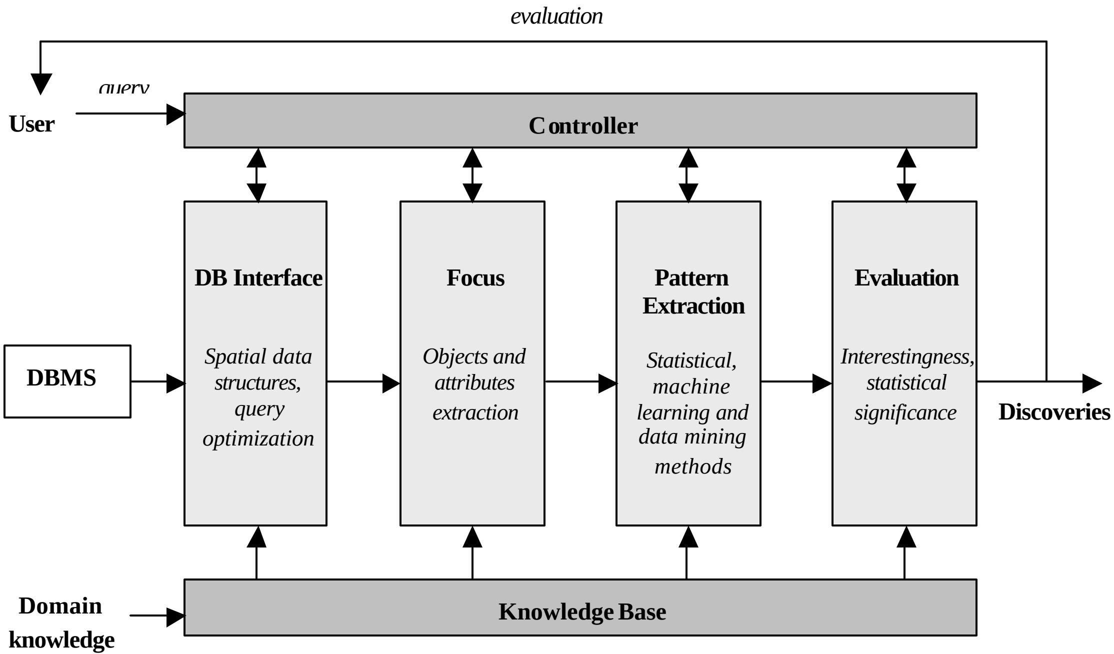
As the previous sections show, this challenge embraces many different research questions from the perspective of the data, the system, visual techniques, modes of inference and collaboration. Most of these questions could apply across a broad range of application domains, but central to the cartographic focus of the ICA are the geographic properties of data, of the constructed knowledge and of the visualization. These are summarized in turn below.

Data. As argued in Section 1.1, detecting pattern in geographic data often requires that location be explicitly presented, and this affects the utility of visualization techniques and metaphors (it will often be necessary to visualize the data as some form of map). The need for locational information may also have consequences for data indexing and retrieval. The huge volumes of geospatial data now available and the highly-multivariate nature of some of the inherent trends and patterns pose additional challenges when designing useful visualization techniques. It is not clear how location should be represented, for example, by name of a city or its geographical coordinates. Similar options exist for time also, e.g. as a linear scale (days of a year) or a cyclic scale (hours of the day, seasons of a year). Mappings between these different representations will be required in many cases.

Geographic knowledge. The representation of geographic knowledge is problematic. At present, there are no universal languages for geographic representation, data is represented and manipulated by the ad-hoc models developed by GIS and database vendors. Specifically, rich conceptual structures (e.g. Rosch, 1975) are all but absent from current, computationally-based geographical models. The only structure that is regularly present is that of the category, although some more recent systems based on the object-oriented data model may also contain generalization hierarchies and component relations (e.g. "is a part of'). Perhaps the first task is to specify a list of geographically-oriented concepts we wish to be able to identify or 'mine', and a means of representing them in current GIS or database schema. Following from this, we must define computational and visualization methods to detect, observe and communicate them.

Visualization environment. As well as supporting map metaphors, the visualization environment will need to communicate geographical meaning and allow the user to interact with tools that construct meaning. The difficulty in designing this kind of functionality should not be underestimated. Careful attention will need to be given to the symbols, legends, displays, interactive behaviors and so forth (e.g. MacEachren, 1995). The ultimate challenge is to construct an environment that can seamlessly address all mining and knowledge construction activities.

Figure 3 shows a diagram of an architecture for knowledge discovery in databases (from Koperski *et al*., 1999). The hypothesis posed in this paper is that visualization has great utility for each stage of this process, by being able to (1) provide deeper insight to the user, (2) facilitate interaction with a user or group of users, (3) enable some form of steering to control the process and (4) provide tools to inspect, verify and edit the extracted knowledge. In this case, we could replace or augment the 'controller' shown at the top of Figure 3 with an integrated visualization environment, and as a result, should improve on the efficiency and effectiveness of knowledge construction and at the same time provide better communication of the processes by which higher level entities are formed. In turn, this should improve understanding, interpretation and utilization of these entities by the user.



Fi[illegible] architecture for knowledge discovery in databases (from Koperski *et al*., 1999). Could visualization assume the role of the controller, overseeing and integrating all analytical tasks and providing interactive tools for pattern extraction and knowledge evaluation?

5 Authors' Note

The authors would like to thank all members and attendees of the ICA Visualization working group on *Database-Visualization Links*, and also Diansheng Guo for pointing out some useful data mining references.

6 References

Abel, D. J., Kilby, P. J. and Davis, J. R. (1994). The systems integration problem. *International Journal of Geographical Information Systems*, Vol. 8, No. 1, pp. 1-12.

Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD*, pp. 207-216.

Andrienko, G. L. and Andrienko, N. V. (1999). Interactive maps for visual data exploration. *International Journal of Geographic Information Science*, Vol. 13, No. 4, pp. 355-374.

Ankerst, M., Elsen, C., Ester, M., and Kriegel, H.-P. (1999), Visual classification: An interactive approach to decision tree construction, in *KDD'99 Proceedings of Fifth ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 392-396

Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data. *SIAM Journal of Science and Statistical Computing*, Vol. 6, pp. 128-143.

Baker, V. R. (1999). Geosemiosis. *GSA Bulletin*, May 1999, Vol. 111, No. 5, pp. 633-645.

Bishr Y (1997). *Semantic aspects of interoperable GIS*. Ph.D. Thesis, ITC Publication Series Number 56, PO BOX 6, 7500 AA Enschede, Netherlands.

Brodlie, K. W., D. A. Duce, J. R. Gallop, and J. D. Wood (1998). Distributed cooperative visualization. *State of the Art Reports at Eurographics98*, (Eds. d. Sousa, A. A. and. Hopgood, F. R. A.) Eurographics Association, pp. 27-50.

Bryson, S. Kenwright, D., Cox, M., Ellsworth, D. and Haimes, R. (1999). Visually exploring gigabyte data sets in real time. *Communications of the ACM*, August 1999, Vol. 42, No. 8, pp.83-90.

Buja, A., McDonald, J. A., Michalak, J. and Stuetzle, W. (1991). Interactive data visualization focusing and linking. Proc. *IEEE Conference on Visualization (Visualization '91)*, San Diego, CA:IEEE Computer Society, pp. 156-163.

Buja, A., Cook, D. and Swayne, D. (1996). Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*, Vol. 5, No. 1, pp. 78-99.

Card, S., Mackinlay, J. and Shneiderman, B. (1998). Information visualization. Readings in Information Visualization (Eds. Card, S., Mackinlay, J. and Shneiderman, B.), San Francisco: Morgan-Kaufmann, pp. 1-34.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, Vol. 68, pp. 361-36

Chernoff, H. (1978). Graphical representations as a discipline. In: *Graphical Representations of Multivariate Data* (Ed. Wang, P. C. C.), Academic Press, New York, USA.

Cleveland, W. S. and McGill, M. E. (1988). *Dynamic Graphics for Statistics*, Wadsworth & Brookes/Cole, Belmont, California, USA.

Cook, D., Buja, A., Cabrera, J. and Hurley, C. (1995). Grand tour and projection pursuit. *Computational and Graphical Statistics*, Vol. 4, No. 3, pp. 155-172.

Dorling, D (1994). Cartograms for human geography. In: *Visualization in Geographical Information Systems* (Eds. Hearnshaw, H. M. and Unwin, D. J.), Wiley, Chichester, England, pp. 85-102.

Drew, P. and Ying, J. (1998). Metadata management for geographic information discovery and exchange. In: Sheth, A. and Klas, W. (Eds.). Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media. McGraw-Hill, pp. 89-121.

Dykes, J. A. (1997). Exploring spatial data representation with dynamic graphics. *Computers & Geosciences*, Vol. 23, No. 4, pp.345-370.

Edsall, R. M (1999). The dynamic parallel coordinate plot: visualizing multivariate geographic data. Proc. 19th *International Cartographic Association Conference*, Ottawa, May 1999. URL: http://www.geog.psu.edu/~edsall/JSM99/paper.htm.

Elder, J. F. and Pregibon, D. (1996). A statistical perspective on knowledge discovery in databases. In: *Advances in Knowledge Discovery and Data Mining* (Eds. Fayyad, U., Piatetsky-Shapiro, G, Smyth, P. and Uthurusamy, R.), Cambridge, MA: AAAI/MIT Press, pp. 83-113

Ester M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. *2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 226-231.

Ester, M., Kriegel, H.-P. and Sander, J. (1998). Algorithms for characterization and trend detection in spatial databases. Proc. *4th International Conference on Knowledge Discovery and Data Mining (KDD'98)*, New York, USA, pp. 44-50.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996a). From data mining to knowledge discovery: An overview. In: *Advances in Knowledge Discovery and Data Mining* (Eds. Fayyad, U., Piatetsky-Shapiro, G, Smyth, P. and Uthurusamy, R.), Cambridge, MA: AAAI/MIT Press, pp. 1-34.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996b). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, Vol. 39, No. 11, pp. 27-34.

Fotheringham, A. S. and Charlton, M. (1994). GIS and exploratory data analysis: An overview of some basic research issues. *Geographical Systems*, Vol. 1, No. 4, pp. 315-327.

Gahegan, M. N. (1996), Visualization strategies for exploratory spatial analysis. Proc: *Third International Conference on GIS and Environmental Modeling*, Santa Fe, NM: NCGIA. URL: http://www.geog.psu.edu/~mark/santafe.html

Gahegan, M. (1999a). What is Geocomputation? *Transactions in GIS*, Vol. 3, No. 3, pp. 203-206.

Gahegan, M. (1999b). Four barriers to the development of effective exploratory visualization tools for the geosciences. *International Journal of Geographic Information Science*, Vol. 13, No. 4, pp. 289-310.

Gahegan, M. (2000a). Visual exploration in geography: Analysis with light. To appear in: *Geographic knowledge discovery and spatial data mining* (Eds. Miller, H. J. and Han, J.), London: Taylor & Francis.

Gahegan, M. (2000b). On the application of inductive machine learning tools to geographical analysis. *Geographical Analysis*, Vol. 32, No. 2, pp. 113-139.

Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F. (2000.) GeoVISTA *Studio*: A Geocomputational Workbench. Proc. 4th Annual Conference on GeoComputation, UK, August 2000. URL: http://www.ashville.demon.co.uk/gc2000/

Gains, B. R., (1996). Transforming Rules and Trees into Comprehensible Knowledge Structures, In: *Advances in Knowledge Discovery and Data Mining* (Eds. Fayyad, U., Piatetsky-Shapiro, G, Smyth, P. and Uthurusamy, R.), Cambridge, MA: AAAI/MIT Press.

Han, J. (1999). Characteristic rules. In: *Handbook of Data Mining and Knowledge Discovery* (Ed. Kloegen, W. and Zytkow, J.), Oxford University Press, Oxford, UK (in press). URL: http://db.cs.sfu.ca/sections/publication/kdd/kdd.html.

Hao, M. Dayal, U., Hsu, M. Baker, J. and D'Eletto, R. (1999). A Java-based visual mining infrastructure and applications. Proc. *InfoVis'99*, Oct 24-29, San Francisco, CA, pp. 124-127.

Haslett, J., Bradley, R., Craig, P., Unwin, A. and Wills, G. (1991). Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *The American Statistician*, Vol. 45, No. 3, pp. 234-242.

Hearnshaw, H. M. and Unwin, D. J. (Eds.) (1994). *Visualization in Geographical Information Systems*. Wiley, Chichester, England,

Hinneburg, A., Keim, D. and Wawryniuk, M. (1999). HD-Eye: Visual mining of high dimensional data. IEEE Computer Graphics and Applications, September/October 1999, pp. 22-31.

Holt, A. and Benwell, G (1999) Applying case-based reasoning techniques to GIS. *International Journal of Geographic Information Science*, Vol. 13, No. 1, pp.9-25.

Huffaker, B., E. Nemeth, and K. Claffy (1999). Tools to visualize the Internet multicast backbone. *Inet'99* Proceedings, http://www.isoc.org/inet99/proceedings/4e/4e_3.htm. San Jose, CA, June 22-25, 1999.

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer*, Vol. 1, pp. 69-97.

Inselberg, A. (1997). Multidimensional detective. Proc. *IEEE conference on Visualization (Visualization '97)*, Los Alamitos, CA: IEEE Computer Society, pp. 100-107

Inselberg, A. and Avidan, T. (1999). The automated multidimensional detective. IEEE *InfoVis'99*, Oct 24-29, San Francisco, CA, pp. 112-119.

Jerding, D. F. and Stasko, J. T (1998). The information mural: a technique for displaying and navigating large information spaces. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 4, No. 3, pp. 257-271.

Keim, D. A. (1996). Pixel-oriented database visualizations, *Sigmod Record*, Special Issue on Information Visualization, Dec. 1996.

Keim, D. and Kriegel, H.-P.(1994). VisDB: Database exploration using multidimensional visualization. *Computer Graphics and Applications*, September 1994, pp. 44-49.

Keim, D. and Kriegel, H.-P.(1996) Visualization techniques for mining large databases: a comparison. *IEEE Transactions on Knowledge and Data Engineering* (Special Issue on Data Mining),

Keim D. (1996). Pixel-oriented visualization techniques for exploring very large databases. *Journal of Computational and Graphical Statistics*, Vol. 5, No. 1, pp. 58-77.

Keim, D. A. and Herrmann, A. (1998). The gridfit algorithms: An efficient and effective approach to visualizing large amounts of spatial data. *IEEE Visualization'98*.

Kenwright, D., Banks, D. Bryson, S. Haimes, R., van Liere, R. and Uselton, S. (1999). Automation or interaction: What's best for big data? Proc. *IEEE Visualization '99*, San Francisco, CA.

Knorr, E. M. and Ng, R. T., (1996). Finding aggregate proximity relationships and commonalities in spatial data mining, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 884-897.

Koperski, K. and Han, J. (1995), Discovery of Spatial Association Rules in Geographic Information Databases, Proc. *4th International Symposium on Large Spatial Databases* (SSD95), Maine, pp. 47-66.

Koperski, K., Adhikary, J. and Han, J., (1996), Knowledge Discovery in Spatial Databases: Progress and Challenges, Proc. *ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, Montreal, Canada, pp. 55-70.

Koperski, K. Han, J. and Adhikary, J. (1999). Mining knowledge in geographic data. *Comm. ACM* (to appear). URL: http://db.cs.sfu.ca/sections/publication/kdd/kdd.html

Kraak, M.-J. and MacEachren, A. M. (Eds.) (1999). *International Journal of Geographic Information Science*: special issue on exploratory cartographic visualization, Vol. 13, No. 4.

Landgrebe, D. (1999). Information extraction principles and methods for multispectral and hyperspectral image data. In: *Information Processing for Remote Sensing* (Ed. Chen, C. H.). River Edge, NJ, USA: World Scientific.

Lee, H. Y. and Ong, H. L. (1996). Visualization support for data mining. *IEEE Expert Intelligent Systems and their Applications*, Vol. 11, No. 5, pp. 69-75.

Leedy, P. D. (1993). Practical Research: Planning and Design (5th edition). MacMillan, New York, USA.

Luger, G. F. and Stubblefield, W. A (1998). *Artificial Intelligence: structures and strategies for complex problem solving*. Reading, MA: Addison-Wesley.

MacDougall, E. B. (1992). Exploratory analysis, dynamic statistical visualization and geographic information systems. *Cartography and Geographical Information Systems*, Vol. 19, No. 4, pp. 237-246.

MacEachren, A. M. (1995) *How Maps Work*. Guilford Press, NY, USA.

MacEachren, A. M. and Kraak, M.-J. (Eds.) (1997). Exploratory Cartographic Visualization. *Computers & Geosciences* (special issue on exploratory cartographic visualization), Vol. 23, No. 4.

MacEachren, A. M. and Taylor, D. R. F. (Eds.) (1994). *Visualization in Modern Cartography*, Pergamon, Oxford, UK, pp. 91-122.

MacEachren, A. M., Wachowicz, M., Edsall, R., Haug, D. and Masters, R. (1999). Constructing knowledge from multivariate spatio-temporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographic Information Science*, Vol. 13, No. 4, pp. 311-334.

MacGill, J. and Openshaw, S. (1998). The use of flocks to drive a geographical analysis machine. Proc. *GeoComputation'98*. URL: http://www.ccg.leeds.ac.uk/james/geocomp/gc24_01.htm.

Mihalisin, T. Timlin, J. and Schwegler, J. (1991). Visualizing multivariate functions, data and distributions. *IEEE Computer Graphics and Applications*, Vol. 19, No. 13, pp. 28-35.

Mitchell, T. M. (1997). *Machine Learning*, New York, USA, McGraw Hill.

Monmonier, M. S. (1990). Strategies for the interactive exploration of geographic correlation. Proc. *4th International Symposium on Spatial Data Handling* (Ed. Brassel, K. & Kishimoto, H.), Dept. of Geography, University of Zurich, Switzerland, Vol. 1, pp. 381-389.

Openshaw, S., Cross, A. and Charlton, M. (1990), Building a Prototype Geographical Correlates Machine. *International Journal of Geographical Information Systems*, Vol. 4, No. 4, pp. 297-312.

Peterson, M. P. (1999). Active legends for interactive cartographic animation. *International Journal of Geographic Information Science*, Vol. 13, No. 4, pp. 375-384.

Piatetsky-Shapiro G., U. Fayyad & P. Smith (1996) From Data Mining to Knowledge Discovery: An Overview. In: *Advances in Knowledge Discovery and Data Mining*, (U. Fayyad, G. Piatetsky-Shapiro, P. Smith & R. Uthurusamy, eds.), AAAI/MIT Press, pp. 1-35.

Pickett, R. M. and Grinstein, G. G. (1988). Iconographic displays for visualizing multidimensional data. *Proc. IEEE conference on Systems, Man and Cybernetics*, IEEE Press, Piscataway, NJ, USA, pp. 514-519.

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, California, USA.

Rainsford, C. P., and Roddick, J. F. (1999). Database issues in knowledge discovery and data mining, *Australian Journal of Information Systems*, Vol. 6, No. 2, pp. 101-128.

Ramakrishman (REF!)

Rhyne, T.-M. (1997). Going virtual with geographic information and scientific visualization. *Computers & Geosciences*, special issue on Exploratory Cartographic Visualization, Vol. 23, No. 4, pp. 489-492.

Rhyne, T.-M. (2000), "Scientific Visualization in the Next Millennium", *IEEE Computer Graphics and Applications*, January/February 2000, Vol. 20, No. 1, pp. 20 - 21.

Ribarsky, W., Katz, J. and Holland, A. (1999). Discovery visualization using fast clustering. *IEEE Computer Graphics and Applications*, September/October 1999, pp. 32-39.

Robertson, G. G. (1991). Cone trees: Animated 3D visualization of hierarchical information. Proc. *ACM CHI'91*, New Orleans, LA: ACM Press, pp. 189-194.

Robertson, P. K. (1991). A methodology for choosing data representations. *IEEE Computer Graphics and Applications*, Vol. 11, No. 13, pp. 56-62.

Robertson P.K. (1997). Visualizing spatial data: the problem of paradigms. *International Journal of Pattern recognition and Artificial Intelligence*, Vol. 11, No. 2, pp. 263-273.

Roddick, J. F. and Spiliopoulou, M. (1999). A bibliography of temporal, spatial and spatio-temporal data mining research. SIGKDD Explorations. Vol. 1, No. 1, (in press) URL: http://www.cis.unisa.edu.au/~cisjfr/STDMPapers/.

Rosch, E. (1975). Cognitive representations of semantic concepts. *Journal of Experimental Psychology*, Vol. 104, No. 3, pp.192-233.

Senay, H. and Ignatius, E. (1998). Rules and principles of scientific data visualization. URL: http://homer.cs.gsu.edu/classes/percept/visrules.htm .

Sheth, A. (1997). Data semantics: What, where and how? In: Meersman, R. and Mark, L. (Eds.) *Database Application Semantics*. Chapman and Hall, pp. 601-610.

Simoudis, E., Livezey, B. and Kerber, R. (1996). Integrating inductive and deductive reasoning for data mining, In: *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, Cambridge, Mass.

Slocum, T. A. (in collaboration with Egbert, S., Weber, C., Bishop, I., Dungan, J., Armstrong, M., Ruggles, A., Demetrius-Kleanthis, D., Rhyne, T., Knapp, L., Carron, J., and Okazaki, D.) (1994). Visualization software tools. In: *Visualization in Modern Cartography*, Eds. MacEachren, A. M. and Taylor, D. R. F.), London: Pergamon, pp. 91-122.

Sowa, J. F. (1999). *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Pacific Grove, CA: Brooks/Cole.

Tang, Q. (1992). A Personal Visualization System for Visual Analysis of Area-Based Spatial Data: Proc. *GIS/LIS' 92, Vol. 2*, American Society for Photogrammetry and Remote Sensing, Bethesda, Maryland, USA, pp. 767-776.

Treinish, L. A. (1999). Task-specific visualization design. *IEEE Computer Graphics and Applications*, September/October 1999, pp. 72-77.

Treisman, A. (1986). Properties, parts and objects. In: *Handbook of Perception*, Volume II, pp. 35.1-35.70.

Tufte, E. R. (1990). *Envisioning Information*, Graphics Press, Cheshire, Connecticut, USA.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA, USA: Addison-Wesley.

Wise, S., Haining, R. and Signoretta, P. (1998). The role of visualization for exploratory spatial data analysis of area-based data. Proc. *Fourth International Conference on Geocomputation (GeoComputation'98)*, Bristol, UK. URL http://www.geog.port.ac.uk/geocomp/geo98/

Wong, P. C. (1999) Visual data mining. *IEEE Computer Graphics and Applications*, Vol. 19, No. 5, pp. 20-21.