"With me everything turns into mathematics."

—Rene Descartes (1596-1650)

4

Data Representations and Transformations

Visual analytics is fueled by data. These data must be represented, combined, and transformed to enable users to detect the expected and discover the unexpected. The volume and complexity of data, combined with their dynamic nature, provide a strong motivation to create innovative and scalable approaches to transformation and representation.

What Are Data Representations and Transformations?

To permit visualization, analysis, and reporting, data must be transformed from their original raw state into a form, or representation, that is suitable for manipulation. These data representations and transformations are the foundation on which visual analytics is built.

Data representations are structured forms suitable for computer-based transformations. These structures must exist in the original data or be derivable from the data themselves. They must retain the information and knowledge content and the related context within the original data to the greatest degree possible.

The structures of underlying data representations are generally neither accessible nor intuitive to the user of the visual analytics tool. They are frequently more complex in nature than the original data and are not necessarily smaller in size than the original data. The structures of the data representations may contain hundreds or thousands of dimensions and be unintelligible to a person, but they must be transformable into lower-dimensional representations for visualization and analysis.

Data representations may illuminate key features in the data, rather than showing every detail, so they are important to the process of data abstraction. The degree to which a visual analytics software tool can address the challenges of scale is also influenced by the data representation selected by the tool developer.

Data representations underlie the interactive visualizations described in Chapter 3. The creation of appropriate data representations is essential to producing meaningful visual representations. The data representation method must facilitate the analytical reasoning methods and capture the intermediate and final results of the reasoning

processes described in Chapter 2. These analytical results must be communicated via the production, presentation, and dissemination processes described in Chapter 5.

A *data transformation* is a mathematical procedure that converts data into different representations that may provide more insight for an analyst. Data transformations are required to convert data into structured forms that permit them to be visualized and analyzed. Data transformations are used to augment data by deriving additional data. For example, clustering is used to organize data into groups. Data transformations convert data into new and meaningful forms. For example, linguistic analysis can be used to assign meaning to the words in a text document. Data transformations make it possible to create more useful visual representations that support more sophisticated analyses. Data transformations can be applied iteratively, with each transformation producing a new representation and potentially leading to new insights. Data transformations may be used to find convenient layouts for displays, such as by creating a low-dimensional display space from a high-dimensional data space.

A major challenge of visual analytics is to find the most useful ways to couple data transformations with interactive visual representations and analytical reasoning techniques. Data representation and transformation techniques should not introduce biases that would affect the analytic conclusions based on the data. At the same time, they should preserve the inherent biases, uncertainties, and other quality attributes of the original data.

Currently, within visual analytics, the sources for data representation and transformation are primarily within the areas of mathematics and statistics, modeling and simulation, and natural language processing (NLP). Generally, much of the knowledge representation work going on in the area of information sciences and technology has some accompanying structure that can be leveraged for automatically generating visual representations and supporting analysis. Without this structure, analytical options are limited because computer processing is constrained.

Transformations to Support Visual Representation

To facilitate the analysis of large and intrinsically complex data repositories, data transformations can be used not only to generate raw analysis results but also to generate representations that can be mapped into spatial representations.

The task of creating representations and transformations to support visualization builds on the foundational work of Euclid (325 to 265 BC, relative to his treatise, *The Elements*) and of Rene Descartes (AD 1596 to 1650, inventor of analytic geometry, i.e., the Cartesian coordinate system). For every point we represent as a pixel on the screen, we leverage Euclid's and Descartes' visions and creativity [Bell, 1965].

In combining geometry and analytics to generate a point on a screen, we must have a value on the horizontal axis (x) and a value on the vertical axis (y). For example, to represent an individual as a point on the screen, we must have some associated spatial structure, say weight, x = 165 pounds, and height, y = 71 inches, thus yielding a location on the screen, or a point in the appropriate visualization space.

About This Chapter

The field of data representations and transformations is so large that it cannot be addressed completely here. Instead, we describe some representative examples and address the data representation and transformation topics that are most central to the advancement of visual analytics. We focus this chapter primarily on representations and transformations to support the creation of visual representations for analysis. The methods described in this chapter also address some of the needs for capturing and presenting the artifacts of the analytical reasoning process. These topics are described in more depth in Chapters 2 and 5.

We identify the need for research in data representation and transformation to better facilitate visual analytics. We highlight areas that must be pursued to address the challenges of understanding complex, diverse, dynamic, and uncertain data.

We also describe the research needed to deal with the linguistic and culturally related structure associated with language data. These data must undergo transformation before they can be represented in a way that supports visualization and analysis. The levels of linguistic structure inform the representation of language data, and some text transformations enrich the semantics of the resulting visualization. Culture affects language data, which in turn affect the visualization of language data, but the community is only in the early stages of research into data transformations to account for these cultural effects.

Because the analytic process often involves the comprehensive consideration of data of multiple types and sources, we present a discussion of the need for synthesizing this diverse information into a single environment in which it can be analyzed. The goal is to allow the analyst to focus on understanding the meaning of the information, rather than being burdened by artificial constraints associated with the form in which the information was originally packaged.

Data Representations

Data come in a variety of types, and each type must be represented in ways that facilitate computational transformations to yield analytical insight. Visualizations that combine multiple types of data are also needed to support comprehensive analytic reasoning in certain situations.

Analytic insights can hinge on the proper data representation underlying the visual representation. The data representation must faithfully preserve the inherent structure of the data as well as the data's real-world context and meaning. In most cases, that inherent structure will be known for a given data source. For example, a given sensor will produce data in a consistent format. If it is unknown, then technical analysis must be done to choose the proper representation for the data. It is important for the data representation to portray the quality of the data as collected. If information is missing, purposefully hidden, or misleading, the analyst must be able to detect that.

Data may be characterized from multiple perspectives, each of which has a bearing on the data representation:

- **Data type**. Data may be numeric, non-numeric, or both. Numeric data often originate from sensors or computerized instruments, and the scientific community has developed a variety of techniques for representing these data. Non-numeric data can include anything from language data, such as textual news stories, to categorical, image, or video data. Although techniques and formats exist for representing individual elements of the raw data, techniques for representing the key features or content of the data are far less mature.
- **Level of structure**. Data may range from completely structured, such as categorical data, to semi-structured, such as an e-mail message containing information about sender and receiver along with free-form text, to completely unstructured, such as a narrative description on a web page. The term *unstructured* does not mean that the data are without pattern, but rather that they are expressed in such a way that only humans can meaningfully interpret their construct. Structure provides information that can be interpreted to determine data organization and meaning. It provides a consistent context for the information. The inherent structure in data can form a basis for data representation. Unstructured data lack the same clues for automatic interpretation for data. Any structure to be applied to the data must be derived in some way.
- **Geospatial characteristics**. When data are associated with a particular location or region, this information must be represented. Any type of data, whether numeric data from a specific sensor, textual data, or image data from a specific satellite, may have a geospatial association.
- **Temporal characteristics**. Some data, such as reference data, are static and not presumed to change over time. However, data of all types may have a temporal association, and this association may be either discrete or continuous.

This section provides a high-level description of some of the considerations associated with representing data of varying types, levels of structure, and temporal and geospatial characteristics. Note that none of these data characteristics can truly be considered independently of the others. All facets must be considered collectively, in conjunction with knowledge about the structure of the source and limitations of the data-gathering technology, to create an appropriate representation. Here, we address some of the elements of data representation that are most significant with respect to visual analytics, but this only scratches the surface of the work that has been done by the computer science community in data representation.

Numeric Data

Numeric data are those data that are quantitative and result from sensors or other instruments, including other computers. These data are unique because they are produced by instruments that automatically format their data and may also be accompanied by software that collects and stores the output as data are being produced. Depending on the analytic tools available, these data may or may not require additional manipulation and re-representation before visually based analysis can begin.

Numeric data have long been the focus of data representation methods, even for manual analysis. There are classical computer-based methods for numeric data representation, many of which reduce the amount or complexity of the data. The current pervasiveness of massive collections of numeric data, such as high-energy physics data [Jacobsen, 2004] and data from the Earth Observation Satellites (EOS) [Braverman, 2004], has spurred development of data representation techniques. These classical techniques provide a basis on which visual analytics can build.

Under normal operational conditions, numeric data would be scientifically analyzed using computational tools designed for the formatted input being received. Research efforts may include investigation of analytic techniques to determine the data structure, the quality of the data source, or predictive indicators. However, the research may also focus on the methods used to represent the data or to detect and mitigate formatting errors in the data.

In emergency conditions and other situations where speed is critical, data representation may play a significant role in making massive data cognitively available to the analyst. Any methods used at this stage in the processing must make special effort to represent the original data content as faithfully as possible so as not to mislead the analyst.

Representing or modeling numeric data appropriately is the key to solving problems. Appropriate numeric data representations and transformations allow the visual representations to speed the analytic process.

Language Data

Linguistically organized data encompass all data that represent human language. While language data are typically processed in textual form, they may also be derived from sound waves or images. Regardless of the original source, representation of the language data content presents many common challenges.

It is difficult to automatically interpret even well-edited English text as well as a native English-speaking reader would understand the text. However, there have been advances in NLP of printed, spoken, and scanned forms in multiple languages that can make a difference in the visual analysis of large amounts of data. In this section, we address the representation of language data. In a later section, the transformation of these representations will be addressed to semantically enrich the resulting visual representation.

Language data can be processed without any acknowledgment of their linguistic structure because meaning is inherent in the communication of the originator. The originator intended to communicate a message to an audience, so the language can be presumed to be meaningful to the reader without automated linguistic analysis. So-called "bag of words" methods, in which a document is treated as a collection of words occurring with some frequency, work because they do not obscure this inherent meaning when presented to the analyst. For many analytic purposes, these methods are ideal. The first mechanized methods were developed by Salton [1968] for information retrieval, and his work continues to be foundational to all language processing as well as other inherently meaning-bearing sources of data [Salton et al., 1975]. His work on identifying salient terms in a corpus, indexing, and constructing

high-dimensional signature vectors that represent a corpus' topics or articles remains key to most of the current effective tools for analyzing large volumes of text. High-dimensional vectors can be projected into 2D to 3D representations to support visualizations that analysts can navigate.

In addition to Salton's work, centuries of general linguistic study of language provide a foundation for the computer-based analysis of language. The general structure of language provides a framework for the eventual reduction of text to its meaningful logical form for computer-based analysis. While computer-based linguistic analysis is not a solved problem, current capabilities provide some reliable results that add semantic richness to the "bag of words" approach.

Linguistics defines the levels of structure based on analysis across and within languages, and computational linguistics provides the methods for assigning structure to textual data. As shown in Figure 4.1, the major levels of structure applicable here are phonological, morphological, syntactic, semantic, and the pragmatic (or discourse) level.

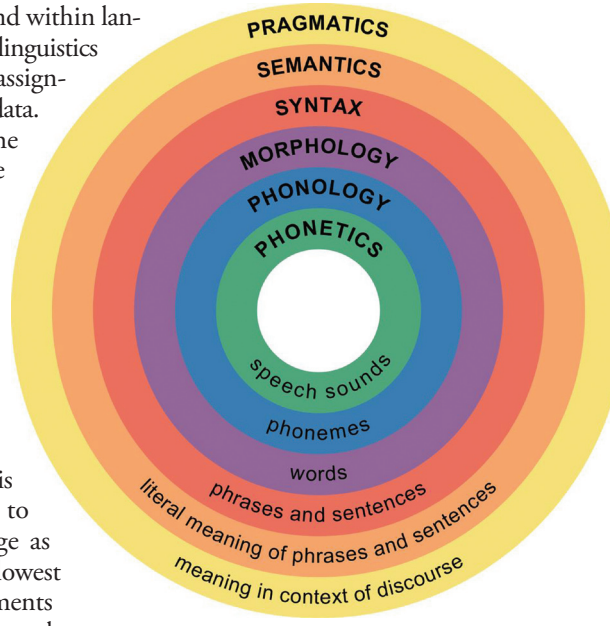

***Figure 4.1**. Major levels of linguistic structure*

The *phonological level* deals with the structure of the sounds that convey linguistic content in a language. However, this level of structure applies to writing and sign language as well. It is basically the lowest level containing the elements that distinguish meaning and can be defined physically as a means of linguistic production. Each language has its own set of sounds that are used in words to convey meaning at any time in its history. These elements are not usually equivalent to the graphemic elements (the smallest elements of meaning) in the writing system. Instead, phonological elements are related to graphemic elements by rules, to a greater or lesser degree. The graphemic system can influence the phonological system. For example, in Mongolia, the Russians replaced the Mongolian script with the Cyrillic alphabet, which caused a change in the vowel harmony rules of the spoken language because the full set of vowels in their verbal context could not be represented in Cyrillic. Usually, however, the phonological system is thought to dominate the written system.

The *morphological level* of a language is the level at which meaning can be assigned to parts of words and the level that describes how morphemes (the smallest meaning elements of words) are combined to make a word. Some written systems, such as

English and Chinese, are morphological in nature. For example, the morpheme "sign" is not always pronounced the same way in English words in which it appears (*sign, signal, signature, resign, resignation*). In highly agglutinative languages, words are built by affixing morphemes to one another, making word boundaries sparse within sentences. In such languages, such as Turkish and many Native American languages, an entire sentence can appear in one word. Obviously, this fact plays havoc with the "bag of words" approach because word boundaries are not easily identifiable using information within the corpus. Even in non-agglutinative languages, segmentation may be required because of the written system's lack of word delimitation. Chinese is such a language.

The *syntactic level* of structure concerns the structure of the sentence, i.e., the categories of words and the order in which they are assembled to form a grammatical sentence. The categories used in syntax are known as parts of speech. The main parts of speech are nouns and verbs. Verbs govern the roles that the nouns in the sentence can play, and the ordering and/or case marking of nouns determine their roles. Roles can be characterized at various levels. Most commonly, the syntactic roles are those like subject and object. The roles can also be viewed with degrees of semantic content, such as agent, instrument, or location.

The predicate-argument structure of the sentence is used to represent the logical form of the sentence, which is a semantic representation. The *semantic level* of structure of the sentence is computationally defined to be the level of representation supporting inferencing and other logical operations. Within linguistic theory, Montague semantics was one of the bases for this approach [Montague, 1974].

Other representations important to the semantic level include, but are not limited to, the meanings of the words. Lexicology is as old as writing, perhaps older, but modern lexicology includes psycholinguistic knowledge concerning how the brain stores the words in memory. WordNet is the preeminent lexicon structured along psycholinguistic principles [Miller, 1998]. The utility of WordNet for computational linguistics has been immeasurable. It contains an ontology, or hierarchical structuring, of the words of English and allows the user to find synonyms, antonyms, hypernyms (more general terms), and hyponyms (more specific terms). It also distinguishes the sense of the words. Other languages have WordNets developed for them and the senses of the words have been linked cross-lingually for use in sense disambiguation within and across languages (see EuroWordNet at http://www.illc.uva.nl/EuroWordNet/).

The *discourse structure* of language is the level of structure of the exchange or presentation of information in a conversation or a written piece. It describes the principles and order of information in the exchange.

All of these levels of structure are opportunities for linguistic representation that could support visualization. Advances like the language-independent UNICODE encoding standard are important at a very basic level, because visual analytics data representations must support analysis of multilingual data. However, the problems of assigning a structure to text are not solved by any means. The real issue is how useful the available linguistic representations are in creating data representations to support visualization and analysis. Techniques, such as tokenization and segmentation, morphology/stemming, and part-of-speech tagging, are sufficiently advanced to provide value in the representation of textual data. None of these techniques are

perfect, but all can be leveraged for many languages of interest. Still other techniques, such as automatic speech transcription, optical character recognition, and handwriting recognition, are currently adequate only under ideal conditions (high-fidelity input equipment, rich contextual clues, and lack of noise in the original expression or signal).

Image and Video Data

The largest volume of data is generally agreed to come in the form of imagery. Images come from a large number of sources, including satellites, surveillance cameras, professional and amateur photojournalism, microscopes, telescopes, and other visual instruments. In addition to the large volume, there is also the possibility of deception because of the tools widely available for editing digital imagery. There is a vast array of research underway in aspects of image and video analysis. Rather than cataloguing this work here, we focus on the particular areas of most significant concern for visual analytics.

State-of-the-art image processing allows edge detection, identification of regions of interest [Glassner, 1995], and the reconstruction of 3D objects from a set of still pictures [Debevec, 1996]. The state of the art in automatic imagery analysis has largely been achieved for computer vision, especially in robotic applications.

The key challenge for visual analytics is to derive semantic content or meaning from images in real time. We must make the leap from the representation of the image itself to the representation of the information contained in the image. The exploding volume of available imagery will stretch data storage and processing limitations. To realize value from these data, the potentially important content must be derived from the data rapidly and accurately so that unneeded data may be discarded and the remaining data can be compressed and offloaded to less accessible storage hardware.

Thus far, good results have only been achieved when the general domain of the imagery is understood, such as face matching and identification of military objects. Inferring that a set of pixels is a particular object in a scene from an unknown source has not been adequately addressed. This is an area of active research. Novel techniques used for massive textual data have shown promise in handling imagery. Ilgen et al. [2000] have applied their vector approach to imagery producing a pixel vector. The method may be useful for identification and verification of objects as well as corrupted images. The detection of hidden messages in imagery is in its infancy [Johnson et al., 2000]. Only under special internet transmission conditions has this been a problem up until now.

Video data are imagery sequences with an associated temporal dimension. Although not routinely exploited today, this temporal dimension can be useful in automated analysis of scenes and identification of objects and events. Research progress has been made in analyzing video content. Capabilities exist to search through and organize large video libraries (see the Virage website at http://www.virage.com/products/). Work has been done to partition video into key sequences [Kasik, 2004], and several companies are working on techniques to index videos for quick retrieval and review.

The proper representation of image and video for analysis is critical to homeland security. We must gain leverage from the extensive research and development efforts already underway in this field to advance the capability to represent not just the image or video itself but also the meaning it contains.

Structural Characteristics of Data

The level of structure within data directly affects their representation. In general, numeric data are well-structured. Imagery have their own unique structure, but the information content within the image is implicit and thus without inherent structure. Language data may exist in forms that range from highly structured to completely unstructured.

Often, metadata exist to describe a particular data element, such as information about the source of an image. These metadata are generally of a known structure. Categorical data, such as survey data, may contain a mix of language data and numeric data, but are also highly structured.

One important example of structured data is transaction data. Transaction data are highly structured records that document an individual event, such as a telephone call or a border crossing. Transaction data contain very small amounts of information in each record and generally do not have a clear context. However, transaction data are generally voluminous. Businesses use transaction data for many purposes, including tracking buying patterns and identifying potential credit card fraud. Security and privacy protection are especially important concerns for working with transaction data.

Many types of data lack the structure that is apparent in transaction data. When structure is not apparent within data, it must be identified through the use of innovative data transformations. Data transformations are discussed in more detail in the remainder of this chapter. When structure exists within data or metadata, that structure must be preserved and represented. Structured data are generally formatted as a field name followed by one or more field values of a specific type. The classic representation of structured data is a set of relations stored in a relational database management system (RDBMS). RDBMSs form the backbone of the commercial database industry. Significant investments have been made in relational databases, and large amounts of data are stored in such databases.

The selected data representation has significant influence over its range of possible analytic uses. In the case of a database, the schema describing field names and types is one such data representation. When possible, schema design should be done knowing the analytic uses to which the data will be put.

For example, the schemas for most databases are designed with transactional efficiency in mind. Names are normally replaced with unique numeric identifiers to avoid the issue of duplicate names. Databases are also normalized so that updates can be done efficiently. However, information shown to an analyst must be represented in the most meaningful way, using familiar names rather than obscure identifiers. There are tools that support the transformation or mapping of one schema into another. However, until recently, schema-mapping tools have not scaled well to large schemas or mappings [Robertson et al., 2005]. The size and complexity of databases for homeland security applications of visual analytics will grow so large

that new techniques will need to be developed to support schema mapping in the cases where analytic uses are not fully known at the time the database schemas are designed. Object-oriented database structures add flexibility but also are not easily changed after an analytic suite of tools is developed. Document metadata also suffer from the same representational limitations.

Another limitation of traditional RDBMSs is that their performance is optimized for transactions per second rather than analytical queries per second. The types of queries used for analysis are quite different. Rather than searching for and updating a single record, analysis usually requires scanning the entire database to find complex relationships of interest. During the scan, filters are applied, aggregations computed, and other calculations performed. In the commercial sector, this has led to a new class of database systems called online analytical processing (OLAP) systems that pre-compute aggregates and support more complex calculations and data modeling.

Interactive analytic workloads are different from traditional queries for other reasons. For example, most analysis is incremental. A subsequent query is a refinement of a previous query. Data management and caching, as well as query optimizers, could be improved to support analysis. Major breakthroughs are needed in these areas to support analytics.

Geospatial Characteristics of Data

Geospatial phenomena have a number of distinguishing characteristics.

First, natural boundaries tend to be very convoluted and irregular, and as a result do not lend themselves to compact definition or mathematical prediction. Geospatial databases tend to quickly become large as a result because of the detailed coordinate data that must be stored.

Second, geospatial phenomena tend to be scale-specific, and phenomena at different scales are interrelated. For example, global weather patterns affect the occurrence of excessive rain in California, which affects the risk of local landslides. Often, problems must be considered at multiple resolutions simultaneously. For example, in detection of a disease epidemic, information may need to be considered at the level of individual hospitals, at the city, state, and national levels in order to identify pockets of illness and identify both localized and widespread outbreaks. Accommodating the multi-resolution nature of geospatial data is a research challenge.

Third, locational definitions of geospatial entities are often inexact and can be scale or context dependent. For example, the boundaries between specific vegetation types in any given area and the location of shorelines when examining at a very local scale are conceptually transition zones and not sharp boundaries. If viewing the same information from a state-wide or national scale, these boundaries would most often be viewed as discrete. City boundaries are also fuzzy transition zones if seen from the point of view of an economist, but cities do have sharply defined boundaries for the purpose of political jurisdiction.

Fourth, locations are commonly recorded using specialized Cartesian, spherical, or other types of coordinate systems including latitude and longitude, Universal Transverse Mercator (UTM) grids, township and range, or street addresses. Location

expressed in some of these coordinate systems cannot be converted algorithmically into other systems with a predictable degree of accuracy, such as the conversion of street addresses into latitude and longitude. This often forces the storage of more than one type of coordinate for entities within the same database. Not only does this make the required storage volumes even larger but it also presents an additional level of complexity in maintaining the integrity of the database as data are added and updated.

The combination of these properties makes representation of geospatial data particularly difficult. Boundaries are represented as sharp demarcations in currently used data models in part because of the discrete nature of computing hardware. An additional problem arises in the transformation of a space that is inherently multi-dimensional into computer memory, which is normally one-dimensional in nature. Representation of geospatial phenomena in a way that retains their essential nature has proven to be a particularly challenging problem [Burrough & Frank, 1995; Mark et al., 1999; Peuquet, 2002; Yuan et al., 2004].

Although relational databases provide significant flexibility for representing many types of data, this does not extend into the geospatial realm, and we still lack representational techniques that are up to the task of modern requirements for visual or quantitative analysis of such data. It is a well-known principle that how the data are represented determines what can and cannot be done effectively with those data. Data representations can also suggest approaches for visual display. A classic example is the use of sequenced snapshots as a data model for storing space-time data; this model coincides with the visual representation of "digital movies"—a series of still images shown in quick succession to visually display movement and change through space-time.

The two basic representation schemes used currently for geospatial data (raster grids and vector) operate as independent and distinct representations within current geospatial data-handling systems. Most current commercial geographic information system (GIS) tools use a multi-representational database design and incorporate both raster- and vector-type representations for coordinate data, as well as links to an RDBMS for storing non-coordinate attribute data.

A fundamental theoretical framework has developed over the past 15 years or so that can serve as a robust basis for moving forward—the notion of the discrete versus the continuous view. These can be briefly defined as follows.

In the *discrete view*, or entity-based view, distinct entities, such as a lake, a road, or a parcel of land are the basis of the representation. Their spatial and temporal extents are denoted as attributes attached to these entities. Vector models fall within this category. In the *continuous view*, or field-based view, the basis of the representation is space and/or time. Individual objects are denoted as attributes attached to a given location in space and time. Using land ownership information as an example, the particular parcel number would be an attribute of the entire space it occupies, with locations denoted in some continuous coordinate field. Raster grids fall within this type of view.

For both discrete and continuous views, there may by attributes that are either absolute in nature (e.g., a lake may have associated with it measured values of specific pollutants, etc.), or relative in nature (e.g., entities adjacent to the lake), or both.

Object-oriented data modeling techniques seem particularly well-suited for specific implementations of this representational framework.

Although visual analytics can capitalize on the wealth of existing research, additional work is needed to address aspects of geospatial representation that are central to the homeland security challenge. Work is needed to address the representation of uncertainty in geospatial information, to address the challenges of information analysis at multiple resolutions, and to develop methods that permit integrated analysis of both geospatial and temporal aspects of data.

Temporal Characteristics of Data

Some phenomena can only be sensed through time. Seismic activity and sound are examples of numeric data that must have a temporal component to be of analytic value. Other non-numeric data, such as video, transportation data, and textual news reports, also have a temporal aspect. When an event generates the data as opposed to an object in stasis, then time must also be measured and associated with the data points or the sampling rates must be set.

The presence of a temporal component changes the types of analysis that may be done and consequently affects the data representation as well. Data must be stored in a structure that preserves metadata about the temporal characteristic of the data. It must also facilitate transformations that permit examination of data in temporal sequence, aggregation of data along temporal lines, and temporal alignment of data.

Just as discrete and continuous views can be applied to geospatial data, they can be applied to temporal data as well. A discrete view can be applied to entities in space-time (dynamic entities) or to events. Examples of purely temporal events would be a bankruptcy or an election. Events that occur in space and time would include an earthquake or a storm. Whether the temporal (or spatial) extent of any object is a point or some interval is dependent upon the temporal (or spatial) scale being used to record the data.

Data may have multiple temporal attributes. For example, a news story has multiple times associated with it: the time of the event it describes, the time at which it was written, and the time at which it was distributed. Any one of these times may be important, depending on the analytical need. When temporal attributes of data are represented explicitly, they can be harnessed to support analysis. However, further research is needed to be able to reliably extract and exploit temporal features that are embedded in unstructured data such as narrative text.

Data Representation Research Needs

Research in data representations is needed to improve our capabilities to fully characterize massive data volumes efficiently and enable effective visual representations.

Recommendation 4.1

Advance the science of data representation to create representations that faithfully preserve the information and knowledge content of the original data.

Among the major breakthroughs needed are:

- Automatic or semi-automatic approaches for identifying content of imagery and video data
- Improved approaches for extracting semantic content from unstructured language data
- Approaches for consistent representation of mixed data-type collections
- Representation of complex space-time relationships within data at multiple levels of resolution
- Representation of dynamic data collections in ways that facilitate real-time analysis and discovery processes.

Textual Data Transformations

The key to making a difference in transforming incoming textual data for visualization is determining the semantic units for the data and visualization method that will improve the analysis in speed, coverage, and/or accuracy. This key is essential even in the use of structured and semi-structured linguistic data, such as databases and tables, where the semantic units may seem to be preset.

This section describes a few representative examples of approaches to textual data transformation, including both statistical and linguistically based methods. An exhaustive survey is beyond the scope of this chapter.

Vector-Based Approaches

Among the widely used statistically based approaches to text transformation are vector-based approaches. In this class of approaches, the content of each document is represented in the form of a vector representing its content. There are many good examples of vector-based approaches. Here, we discuss three of them for illustration.

Latent Semantic Indexing

One example is Latent Semantic Indexing (LSI) [Deerwester, 1990]. LSI looks at patterns of word distribution, specifically, word co-occurrence, across a set of documents.

Natural language is full of redundancies, and not every word that appears in a document carries content. Articles such as "the" and "a" are obvious examples of words that do not carry content. These words are called *stop-words* and are ignored in LSI and other vector-based approaches. LSI condenses documents into sets of content-bearing words that are used to index the collection. A matrix of terms and documents is created using these content-bearing words. The value placed in a cell corresponding to document d and term t is some measure of the importance of term t in document d. There are several alternative approaches for calculating this measure of importance, ranging from simple binary approaches indicating the presence or absence of a term, to counts of word frequency, to derived measures. The resulting

matrix is transformed using singular value decomposition (SVD) [Forsythe et al., 1997] to create a more compact representation of document content. This compact representation can support document grouping and retrieval based on content rather than on keywords.

System for Information Discovery

System for Information Discovery (SID) is an example of a statistically based system for computing high-dimensional knowledge vector representations. Developed at the Pacific Northwest National Laboratory, SID characterizes natural language documents as vector-based knowledge representations so that they may be organized, related, navigated, and retrieved based on content similarity. SID autonomously identifies the working vocabulary of terms that best differentiate and describe a collection of text documents, defines a tangible anchoring vocabulary that is represented in a measurement matrix, determines weighted probability distributions for the working vocabulary in terms of these tangible anchors, and uses these results to construct an interpretable, high-dimensional vector representation of each document. The use of compact probability distributions and a tangible anchoring vocabulary allows interactive steering of representation based on user need for multiple points of view and specialized knowledge understanding frameworks. SID offers the advantages of scalability and speed of computation. It requires no training by the user, so it offers great flexibility. It supports processing of dynamic data sets and permits efficient incremental addition of documents to existing data sets.

Visualization systems can render interactive representations of document collections with an underlying vector knowledge representation by applying clustering, self-organizing maps, and dimensionality reduction techniques to form low-dimensional visualizations of the high-dimensional knowledge space. In IN-SPIRE™, the document vectors produced by SID are used to generate Galaxy and ThemeView™ visualizations. These visualizations allow users to rapidly understand the relationship between documents and themes throughout the document space [Hetzler & Turner, 2004].

MatchPlus

Still another example of a vector-based approach is that used by the MatchPlus system [Caid & Onig, 1997; Caid & Carleton, 2003]. This approach uses an adaptive neural network-based approach to creation of document vectors. Relationships among terms are calculated with reference to the given data set. Unlike LSI and SID, MatchPlus uses a training set. However, while LSI and SID treat each document as a "bag of words," MatchPlus considers the proximity of terms in a document, providing an increased sensitivity to uncovering the multiple meanings that a word can have within a document set. MatchPlus produces vector representations for words, documents, and clusters. This vector representation provides a structure that can be leveraged in the generation of visual representations of the type discussed in Chapter 3 [Caid & Onig, 1997; Caid & Carleton, 2003].

Probabilistic extensions to the vector space model approach

Several techniques are being developed that harness the combined power of vector space models and probability distribution approaches. A new class of research in generative models brings machine-learning techniques to the characterization of text data.

Probabilistic Latent Semantic Indexing (PLSI) [Hofmann, 1999] refines the LSI approach by grounding it in theoretical statistical foundations. It models topics as probabilistic combinations of terms; individual words are associated with a given topic. Blei et al. [2003] developed a generative probabilistic model known as latent Dirichlet allocation (LDA) to model documents as mixtures of topics. This topic-modeling approach has been further extended by Rosen-Zvi et al. [2004] to model both topic and author.

The Association Grounded Semantics (AGS) approach is based on the premise that an entity, such as a word or other object in an information space, is the totality of all that is associated with it. Using vector representations in combination with probability distributions, one can develop techniques for representing semantics and other aspects of meaning and knowing, such as unsupervised identification of entities (people, places, and other unsupervised activities) [Wang et al., 2005].

Natural Language Techniques

The techniques described above all derive their representation of a text collection without considering the semantic structure of the language. Natural language techniques offer a different approach that considers the text from a meaning-centered approach. These techniques offer a good complement to the "bag of words" approaches described above.

Named entity recognition (NER) and multi-word expression detection algorithms are improving and offer potential value in visual analytics. NER is the automated identification and categorization of proper names, while multi-word expression detection involves the automatic identification of multi-word phrases used to describe a single thing. Despite the results of public evaluations (see http://www.itl.nist.gov/iad/programs.html), NER is still not much more than 70% reliable on realistic data, even with extensive tuning by a computational linguist. This level could be inadequate for visual analytics if high recall—that is, the automatic identification of a high percentage of named entities present in the data—is critical to the analytic results. Techniques as simple as n-grams [Jurafsky & Martin, 2000], which involve examining sliding windows of n consecutive words, and as complex as full-fledged linguistic processing with co-reference resolution have been tried successfully for certain data sets under ideal conditions. Unsupervised learning has produced initial high-recall results above 80% with a handful of analyst examples. However, NER is not currently a solved problem for realistic, streaming data, let alone the volume of streaming data that must be analyzed rapidly and accurately in emergency situations.

Sense disambiguation, or the identification of the correct sense of meaning of a word in a particular context, currently relies on extensive lexical resources such as WordNet and EuroWordNet, but statistical methods show promise. Machine translation is only viable for data triage or the narrowing of a collection to the most

promising sets of documents. A linguist must be called in to translate crucial materials. Topic detection, summarization, and question answering have been possible in English but work poorly across languages.

At this time, combinations of NLP techniques may quickly degrade the data because of the multiplicative nature of the error rates. More significantly, no normalization of names or co-references across documents has been successful enough to support visual analytics; analysts must still assist in completing the pre-processing or finding workarounds. As reasoning capabilities are visualized, the current understanding of negation [Polanyi & Zaenan, 2004], affect (that is, the emotional content of an expression) and attitude, and modals (auxiliary verbs that change the logic of a sentence and have abnormal time references) will need to be integrated into data representation for computational visualization.

Semi-structured data in tables with labeled rows and columns or other formats in written language have proven to be more difficult to interpret semantically than first thought. The data are often put in these special formats because of their semantic salience, so it is appropriate to find a data representation for semi-structured data that will support visual analytics. Semi-structured data have been worked on in speech recognition in the areas of air traffic control and air traffic information services. Hidden Markov modeling lends itself well to semi-structured speech data because of the relative simplicity of the language models required. However, it is unclear that the same techniques would work in cases where the structure of the semi-structured data is unknown ahead of the transmission. Work needs to be done to detect semi-structured data and to exploit their inherent semantics.

Unstructured linguistic data are the most common linguistic data. The nuances of communication through language can only be generated and interpreted by humans. Automatic multilingual NLP has been attempted for decades. However, to determine the semantic units useful for visualization and analysis, much research is still needed. Such research would support applications well beyond the needs of homeland security.

Intercultural Analysis

Culture has a significant effect on the appropriate interpretation of textual data. The incorporation of cultural considerations into data transformations has not been systematic. Successful prevention of terrorist activities and response in the event of a homeland security emergency could hinge on knowledge of the subcultures.

Theories in fields such as ethnography are emerging to provide experience-informed theories of how to understand and work across cultures. Hooker [2003] describes the dynamics of different cultures and theorizes that the best way to adapt to and be effective in another culture is to use that culture's mechanisms for stress management. These mechanisms are the dynamics of the culture for a reason and one must adapt them to become aware of and be assisted by the culture in managing the stress of acclimatization and everyday life.

The concepts used in Hooker's classifications of culture include relationship-based and rule-based cultures, shame-based versus guilt-based cultures, and polychronic (multi-tasking) versus monochronic (serial) cultures. These concepts are important

to both the appropriate analytical interpretation of textual information and the communication of instructions to the community in the event of an emergency.

These culture-based concepts cannot be quickly identified or separated from incoming language data to make the language data culture-neutral. Instead, new methods of data transformation are needed that build upon ethnographic theories to identify and understand the dynamics of the cultures that are evident in the data and to appropriately reflect effects of cultural differences in analyses and models.

Text Transformation Research Needs

Much work has been done in the transformation of textual data for analysis, but the problem remains a difficult one. Technology advancements are necessary to advance the state of the art in statistically based representations. In addition, there are technology needs in all areas of NLP in dealing with unstructured, semi-structured, and structured linguistic data whether they come in as text, sounds, or images. Structured data perhaps require the least research. Normalization of names, locations, dates, and times within and across languages must be fully addressed if we are to equip the analyst to cope with multi-source information in circumstances ranging from emergency operations to long-range reporting to planners and policymakers.

Recommendation 4.2

Advance the state of the art for statistical transformation of textual data collections so that the resulting representations restructure the data into forms suitable for computer manipulation without compromising the information and knowledge content of that data.

Research is needed to develop new statistically grounded transformations that can support:

- Real-time characterization of documents as they are added to a dynamic collection. New approaches are needed that can handle massive data volumes in a computationally efficient manner.
- Multi-resolution characterization of document content. Techniques are necessary for characterizing documents at a finer level of detail than a "bag of words." Additional techniques are needed for characterizing documents at the sentence and paragraph level, and the section level, in addition to the overall document level.

Recommendation 4.3

Invest in the synergistic combinations of NLP and "bag of words" data transformation techniques to create higher-fidelity, more meaningful data representations.

Current successes in NLP and limits of "bag of words" approaches allow for synergistic combinations that have not yet been explored extensively. The combination of these techniques has the potential to dramatically transform the volume

problem by maximizing the use of human cognitive channels through the presentation of only semantically salient data in normalized form. Integration of NER techniques, sense disambiguation algorithms, and multi-word phrase identification are all feasible augmentations of vector-based approaches.

Recommendation 4.4

Extend NLP to infuse visual analytics data representations with semantic richness.

The current state of the art of NLP is not adequate for the needs of visual analytics. Important research areas include:

- Pre-processing multilingual text, speech, and written or printed input
- Normalizing names, locations, dates, and times
- Using and developing intermediate representations from computational linguistics processing to support cognitive absorption
- Developing co-reference techniques to tie information from different languages, files, and databases to the correct topics, events, and entities
- Developing logical models of modals, negation, attitude, and affect to support reasoning.

Recommendation 4.5

Leverage work done in ethnography and computational linguistics to develop data transformations that can capture the cultural context inherent in textual data.

The need to incorporate culture within visual analytics systems is largely unaddressed today. This represents a significant challenge and a major opportunity for research.

Additional Approaches to Data Transformation

The challenge of data transformation is central to the success of visual analytics tools. Earlier, we outlined approaches for transformation of language data. While some of those approaches may be used for non-language data, they are most commonly used with textual data. Here, we consider more general data transformation approaches that are broadly applicable to a variety of data. These represent only a subset of the data transformation techniques that hold promise for visual analytics.

We describe three classes of data transformations: dimensionality reduction approaches that simplify data collections; discrete mathematics techniques that represent discrete objects through a combination of data and models; and modeling- and simulation-based approaches.

Dimensionality Reduction

Dimensionality reduction techniques provide generalized methods for data simplification. The ability to transform large, high-dimensional structured data sets into lower-dimensional representations is important for the generation of the visual representations.

Dimensionality reduction may be accomplished in two ways: by reducing the number of observations that must be managed or by reducing the number of variables that must be considered. Dimensionality reduction methods can be linear or nonlinear. The more straightforward linear methods identify global homogeneities to collapse while the more complex nonlinear methods search for local homogeneities.

For reduction of the number of variables, we consider two example techniques. Principal components analysis (PCA) is an example of a linear variable reduction technique in which new variables are produced by creating linear combinations of the original variables. A second approach to reduction of variables is automatic feature selection. The set of variables to be considered is identified through automated means such as statistical or machine-learning algorithms, or in simple cases they can be identified directly by users exercising their expert judgment.

Multi-Dimensional Scaling (MDS) techniques are an example of a well-established approach to dimensionality reduction. MDS techniques may be either linear or nonlinear. MDS techniques create smaller pseudovectors that approximate the high-dimensional structure of data in a lower-dimensionality representation that attempts to preserve the proximity characteristics of the high-dimensionality structure. Because there are many different ways to analyze proximity, and because of the nonlinear nature of the algorithm, it is difficult to interpret the results of this algorithm. One of the major challenges is to preserve the information and knowledge content of the original data set that was used to generate the original high-dimensional set. It is at least important to preserve the information and knowledge of interest to the visualization's user [Steyvers, 2002].

Newer, nonlinear techniques are also being pursued in the area of manifold learning for reducing the number of variables. An example is Tennebaum et al.'s [2000] Isometric Mapping (ISOMAP) procedure, which combines graph theoretic approaches and MDS methods to approximate the structure of the interpoint distance matrix in a lower-dimensional space. Roweis and Saul [2000] suggested Local Linear Embeddings (LLE). This method solves for the manifold using local linear patches. An alternate approach to the manifold learning problem is based on the eigenvalues/eigenvectors solutions that characterize the behavior of an operator on the manifold. Belkin and Niyogi [2003] and Lafon [2004] studied this problem from the context of a machine-learning problem.

Clustering of homogeneous data is a common method for reducing the number of observations to be managed. With large data sets, statistical sampling is often proposed as a means of obtaining computable data sets. The merit of the approach depends, to some extent, upon whether the task is to find subtle hidden evidence, in which case the information exists only in trace amounts that are unlikely to be discovered through sampling, or more widespread trends, in which case sampling is likely to suffice. There are challenges in producing random and stratified samples from databases and from streaming data. For example, there are numerous ways to

pick half a million items from a billion items and it is non-trivial to guarantee that all possible combinations of one-half million have an equal chance of being selected. In general, it is much better to represent all data rather than to statistically sample the collection to reduce data volume.

Recursive partitioning methods are examples of nonlinear methods that can also be used to simplify data. These methods provide models that span categorical, ordered, and different kinds of numerical data. While not always obvious, many of the modern modeling methods developed in recent years are based on a combination of data partitioning followed by local parametric modeling.

Both established and new data structures have the potential to enable calculation of previously expensive statistics on large amounts of data [Moore, 2004]. The current trend is to use fast, cheap statistics to emulate more computationally expensive statistics. Applications include the fast calculation of likelihoods. Another current research topic is the scaling of algorithms to accommodate massive data volumes. New approaches take advantage of fast approximation methods to accomplish less important computations quickly, so that more of the computation time can be focused on tasks for which accuracy is more critical.

Another paradigm for analysis of large information spaces is to analyze the statistics of scattered data in very-high-dimensional (VHD) spaces, consisting of hundreds or thousands of variables. Sparse data also affect much lower-dimensional data sets, if the ratio of the number of observations to the number of dimensions being measured is small.

Analysis of sparse data is difficult for two reasons. First, the emptiness of these sparse spaces ("curse of dimensionality") [Bellman, 1961] makes it difficult to reliably establish neighborhood relationships. As noted in Chapter 3, interesting structures in these spaces may be non-planar (nonlinear), meaning that they cannot be represented easily in very low dimensions. Thus, analysis of sparse data requires more sophisticated tools than those used for linear analysis.

Fundamental developments in several fields suggest opportunities for new strategies and tools for sparse data. For instance, statisticians have found that the curse of dimensionality is not as dire as the theory predicts [Scott, 1992] and that many naturally occurring data sets fall on low-dimensional manifolds within VHD spaces. Newer, better tools in nonparametric density estimation, based on information theory, promise to be a good foundation for exploring such data [Haykin, 2000]. Another area of development is in computational topology, where researchers have proposed new methods for robustly parameterizing such manifolds and characterizing their structure, dimensionality, and topology.

Discrete Mathematics

Whereas continuous objects can be characterized by a formula, discrete objects are characterized by a method and require a mathematical model or abstraction [Maurer & Ralston, 2004]. These models or abstractions transform the data in ways that aid in analytical reasoning. Discrete mathematics provides mathematical models for a large number of discrete objects: induction, graphs, trees, counting methods, difference equations, discrete probabilities, algorithms, and n-order logics. Once

discrete mathematical methods have provided the model, the data must be transformed into a form that fits the model.

One example of such a model is a semantic graph. A graph consists of entities as nodes and relationships as links. The entities and the relationships often have attributes associated with them in the graph. These entities may be extracted from the source data through a combination of semantic, statistical, and mathematical techniques. A relational data model is normally used as the representation for a graph.

A semantic graph is a type of graph that encodes semantic meaning about entities and relationships. Examples include relationships between people, places, and events. The transformations that produce a semantic graph are generally natural-language-based and, as discussed previously, are not without error. Consequently, measures of data integrity must be represented if analytics is to be served. Other government programs dealing with knowledge representation and data filtering have created sophisticated approaches to such noisy data and will continue to provide technology to transform the graphs and provide probabilistic query functions. However, research is needed to address the challenge of creating meaningful visual representations for the voluminous, complex semantic graph structure.

Modeling and Simulation

Modeling and simulation are useful in gaining understanding of the interaction of large numbers of variables and a dynamic situation in which there are many possible outcomes. Modeling and simulation transform data into sophisticated representations that depict the evolution of a situation over time. These outputs can be challenging to analyze, but they offer rich insights into complex systems. Visual analytics provides distinct advantages for analyzing these outputs because it can help the analyst clearly understand the phenomena these outputs depict through a combination of visualization and analytical reasoning tools.

Many modeling and simulation techniques are relevant to visual analytics. We consider agent-based modeling, neural networks, and genetic algorithms as examples here.

An agent-based model is a specific, individual-based computational model for computer simulation extensively related to themes in complex systems, emergence, computational sociology, multi-agent systems, and evolutionary programming. The idea is to construct the computational devices, known as *agents*, with some properties, and then simulate them in parallel to model the real phenomena. Because of the interactions that take place over time, new patterns and properties emerge [Axelrod, 1997].

A neural network is a processing device, either an algorithm or actual hardware, whose design was inspired by computer simulation of the design and functioning of human brains and components thereof. An artificial neural network (ANN) is a network of usually simple processors, units, or neurons. The units may have local memory and are tied together by unidirectional communication connections, which carry numeric data. The units operate only on their local data and on the inputs they receive via the connections. Most neural networks have some sort of training rule whereby the weights of connections are adjusted based on presented patterns. In other words, neural networks learn from examples, just as the brain learns to recognize things from examples, and exhibit some structural capability for generalization.

Neurons are often elementary nonlinear processors. Another feature of ANNs that distinguishes them from other computing devices is a high degree of interconnection that allows a high degree of parallelism. Further, there is no idle memory containing data and programs, but rather each neuron is pre-programmed and continuously active [Fausett, 1994].

Genetic algorithms use simple representations (bit strings) to encode complex structures and simple transformations to improve those structures [Davis, 1987; Holland, 1975]. The transformations are inspired by the computer simulation of natural genetics to evolve a population of bit strings in a way analogous to the way populations of animals evolve. Genetic algorithms have many of the characteristics of neural networks, in that they are parallel and can learn from examples to detect extremely complex patterns. Genetic algorithms are also the basis of evolutionary programming mentioned earlier.

These and other modeling and simulation techniques transform data into new representations that offer the opportunity for insight into complex situations. We need to conduct research to identify the additional transformations and representations necessary to effectively present this information to analysts for understanding and action in urgent situations.

Data Transformation Research Needs

Data transformation is central to the success of analysis of massive and dynamic data sets. The visual analytics community must stay abreast of the advancements being made by the thriving research community that is already addressing many of these topics. We must take advantage of new capabilities as they are discovered.

There are a few areas of special interest to the visual analytics community that are of lesser focus in the data transformation community as a whole. The visual analytics community must help drive the development of new transformation methods in these areas.

Recommendation 4.6

Pursue research in data transformations that improve our understanding and reaction to new and unexpected situations.

Research is needed to develop data transformations that facilitate characterization of current situations through the real-time identification of relationships, categories, and classifications. Specifically, new transformations must be created to facilitate the dynamic identification of new and emerging events and situations that were not previously identified or anticipated. Techniques that rely on a priori knowledge or training sets for characterization must be augmented with approaches that recognize novelty or detect surprise. Multi-resolution techniques are needed to allow the detection of both broad, emerging characteristics and very subtle, trace-level characteristics.

Recommendation 4.7

Develop a theoretical basis to represent and transform discrete data sets of varying scale and complexity.

Continuous mathematical theory has been successfully applied to natural science and engineering. However, discrete mathematical techniques require an additional theoretical base, especially when applied to massive data. Existing techniques are ad hoc and often break down as the amount of input data increases. Huge gains appear to be possible in the scale of data for which routine analyses can be pragmatically accomplished.

Discrete mathematical models show real promise in addressing the challenges of analysis of massive and dynamic data sets. Visual analytics must support the transformation of these models and associated data into a form that can be visually represented for analysis.

Information Synthesis

The techniques described thus far in this chapter address the challenges of representing and transforming data that are all relatively homogeneous in form and format. Similarly, most current commercial tools and research techniques focus on the representation of the unique characteristics of static collections containing a single type of data. Consequently, many existing visual analytics systems are data-type-centric. That is, they focus on a particular type of data or, in some cases, provide separate but linked environments for analysis of different types of information, such as textual data, structured data, and geospatial data.

Information synthesis will extend beyond the current data-type-centric modes of analysis to permit the analyst to consider dynamic information of all types in a seamless environment. The user should not have to be concerned with, or restricted by, the original form or data type, but should instead be able to consider the relevant elements of the content of that data in concert with data of other types. We must eliminate the artificial analytical constraints imposed by data type so that we can aid the analyst in reaching deeper analytical insight. As with all data transformations, the resulting data representations must preserve, to the best degree possible, the information and knowledge content of the original data, but these representations must also integrate the information content across multiple data types and sources. By giving the analyst the ability to assemble facts and examine alternatives without imposing artificial barriers on data, information synthesis will help the analyst gain rich insights.

To achieve the desired information synthesis, data transformations must permit the combination of data of multiple types into integrated collections via unifying mathematical representations. Because of the dynamic nature of the data, we must develop techniques to identify and represent significant changes in data. Methods for coping with missing, sparse, and inconsistent data are important in all visual analytics data representations and transformations but take on special significance in

synthesized information spaces where the heterogeneous nature of the data adds complexity to the analytical challenge. Furthermore, methods for preserving and representing data quality, pedigree, and uncertainty must also be considered in order to produce a more powerful, information-rich structure to support visual analytics. Each of these subjects is considered in more detail below.

Combining Multiple Sources

Synthesizing data across sources allows an analyst to form a semantic model. This, in turn, leads to discovery of previously unknown or unsuspected behavior. Because the data streams are so large, contain multiple data representations and transformations, and represent multiple domains, data synthesis provides techniques to facilitate the cognitive merge that may not take place otherwise. A person's visual channel alone cannot overcome the limitations of formulating a model or set of viable models. Data synthesis addresses both the quantitative and qualitative aspects of the task and helps the analyst identify what is interesting and what is not. Otherwise, an analyst would have to sort through a huge set of mappings and views in disparate data forms to be able to gain a similar level of understanding [Hetzler & Turner, 2004].

One example of rapidly evolving scientific endeavors that parallels the homeland security need for information synthesis is the analysis of genomics and proteomic data in bioinformatics. The field increasingly uses the exponentially growing body of metadata to provide both broader knowledge and more focused analysis of new quantitative data. The metadata include gene ontologies (GO), the mapping of genes to GO, measures of the mapping quality, and the corpora of abstracts, papers, and data related to the genes and gene products of interest. These metadata provide a structure for a global information space that lends context to support multi-type analysis [Gentleman, 2004].

Combining data of multiple sources and formats, also called multiple media, into a single data representation constitutes an important research challenge for visual analytics. We envision creation of cross-media representational techniques such as a modified use of the context vectors discussed above. In a cross-media analytical environment, the vectors themselves may represent data content and context in a universal information space. This universal information space is the product of multiple data types and multiple data sets. The vectors contain sub-vectors that contain common cross-media content and context, while other sub-vectors contain specific, within-media content and context. This combination of a unified information space and a data-type-specific representation allows for maximum flexibility so that data of all types may be analyzed together or within homogeneous collections as needed for a particular task.

Identifying Change

Common data transformation techniques are oriented toward transformation of a single data snapshot. However, data are dynamic, making the detection of change in data fundamental to analysis. Changes in monitored data are often good early

warning signs about emerging events of interest, even if the change is only partly or poorly understood. The quantity, variety, and complexity of data relevant to homeland security require novel approaches to change detection.

Change detection arose from research in industrial quality control during the 1950s. Today, change detection is used in a variety of applications, including machinery diagnostics, computer network failure detection, authorship change detection in text documents, and scene change detection in video. A common viewpoint in change detection is to consider a sequence of random measurements that are to be monitored for a possible change. The goal is to discover a specific time such that a sequence of measurements before that time differs statistically from a sequence of measurements after that time. The objective is to find the first such time while minimizing the rate of false positives.

A basic algorithm for change detection is the cumulative sum, or CUSUM, algorithm. CUSUM monitors a recursively defined quantity defined from the set of measurements and represents the log-odds ratio that any specific measurement is a post-change measurement [Poor, 2004]. Over the past couple of decades, a number of approaches for change detection have been developed that extend the CUSUM algorithm in different ways.

However, detecting change within homeland security data sources is more complex than the original industrial quality-control applications. First, the much larger scale of the data and the multi-source, multi-type nature of the data demand new approaches. Furthermore, data in the homeland security context are driven by discrete events. As a result, new methods for change detection are needed.

Accommodating Incomplete, Uncertain, and Contradictory Data

It is important that data representations preserve all of the quality attributes of the original data that they represent. In the case of visual analytics, data are often incomplete, uncertain, and contradictory. The data representation and transformation techniques used in visual analytics must both accommodate these data characteristics and facilitate management of them in an analytical context.

In dealing with the uncertainty associated with data, one must consider: 1) identification of uncertainty, which is frequently treated as a given but usually is actually well hidden and fuzzy, 2) representation of uncertainty, 3) aggregation of uncertainty, and 4) communication of uncertainties. Gaps, uncertainty, contradiction, and deception are characteristics of homeland security data requiring special consideration. For example, the internet, as is the case with many data sources, is rife with misinformation [Mintz, 2002]. Providing some automatic assistance for identifying, or even hypothesizing, that information is missing or incorrect is of substantial benefit to analysts.

Information can be missing for a variety of reasons, ranging from failure to have an observer or instrument in place to obtain the information to the inability to retrieve the information in a sufficiently timely manner. Examples of contradictory information and misinformation are readily available in financial, political, and

social settings. Missing, incorrect, and contradictory data are conditions that frequently occur in scientific and business data processing, and practitioners have several years' experience managing these conditions. Dasu and Johnson [2003], among others, comprehensively review data preparation, quality, and exploration issues. Pattern analysis for contradictory data is also explored in the literature [Leser & Freytag, 2004]. Intentional deception is not typically considered in these domains, although it does occur in situations such as competitive intelligence [Mintz, 2002]. There is general, established theory for addressing the missing data in specific domains (e.g., financial, political, and social settings) [Little & Rubin, 1987; Allison, 2002].

For the purposes of visual analytics, a different slant on these data conditions must be taken [Berkowitz & Goodman, 1989]. Analysts deal with a combination of known facts that can be verified with a high degree of confidence and data with known gaps and ambiguities. Problems arise because analysts are required to make a best estimate using available data. They bring forward assumptions to help drive their evaluations. This can result in disagreements among different analysts reviewing the same data.

Situations in which different hypotheses are strongly supported by facts and in which gap-filling assumptions drive different interpretations must be made known to those outside the analysis community. Information consumers and decision makers would like definitive answers, but often the best product contains areas of uncertainty, unclear meaning, and suspect origins. When estimates and evaluations are made, descriptive yet subjective terms, such as "highly likely" or "unlikely" appear. Confidence levels are affected by specific factors that Donald Rumsfeld recently (and Sherman Kent, earlier; see http://www.cia.gov/cia/publications/Kent_Papers/vol2no3.htm) referred to as "known unknowns" [Rumsfeld, 2002]—items known to be important, yet unable to be estimated with a sufficient level of confidence.

Deceptive data or disinformation is provided by adversaries to attempt to deceive or mislead analysts. Deception and disinformation can cause intelligence assessments to go awry, distort confidence levels in intelligence channels, and cause broad questioning of related assessments even to the level of creating discomfort about the overall quality of intelligence processes and products. The typical approach to detecting deception in information, other than by directly identifying it, is through examination of patterns of anomalies. This is a difficult process because of the enormous amounts of information that need to be processed.

Automated identification of cues used in deception in text-based communications is preliminary but promising [Zhou et al., 2003]. Current theory for dealing with contradictory information applies in a focused technical area, in which the pattern of missing information does not carry information about the underlying model or phenomena. This theory and methodology are typically applied in settings, such as surveys (product warranty information, opinion polls, etc.), in which a template for all the possible information is available. This technical approach does not apply here; however, the theory might provide some cues or approaches for this technical area. Additional, less mathematical, reference areas include information on internet hoaxes and library science perspectives on evaluating internet references and using peer reviews (e.g., Wikipedia, http://www.wikipedia.org/).

Two aspects of uncertainty representation must be considered. One is the representation of uncertainty attributes that are known or can be identified. The other aspect is the identification, representation, or propagation of the uncertainty attributes that are not necessarily known and quite likely not intuitively obvious to the user. In the transformation of data to support visual analytics, it is important to transform the original uncertainty attributes in a way that they can be presented within the visualization for exploitation by the user consciously or subconsciously. The uncertainty can be made available to the user's mental processing capabilities independent of the uncertainty attribute ever being specifically identified for processing by the relevant algorithm.

Confidence Assessment

To ensure the appropriate use and interpretation of data, confidence levels must be represented. This confidence has its origins in the value and uncertainties associated with the data or lack thereof, with the source of the data, in the analytical methods used in an assessment, and in the perceptual aspects of the end user of the assessment. The identification and communication of confidence values are not easy tasks. Therefore, it is important that transformation of the original data preserves all uncertainty attributes that influence the confidence assessment.

We need to facilitate both the assessment of confidence and its subsequent communication so that that the user can understand both the information being conveyed and the level of confidence that should be placed in that information.

Information Synthesis Research Needs

Information synthesis is central to the major goal of visual analytics. To achieve this vision, several important research objectives must be achieved.

Recommendation 4.8

Pursue mathematical and statistical research in the creation of data representations and transformations to permit unified representation of dynamic data of multiple types and sources.

These techniques are central to achieving the goal of information synthesis. These techniques must produce high-fidelity representations of the original data. The representations must be versatile enough to not only permit cross-media analysis but also allow for more detailed analysis of data-type-specific attributes in homogeneous collections.

We must identify transformations that combine different data representations into more meaningful supersets to improve an analyst's ability to comprehend complexity. Current transformations offer solutions for a single data type and rely on a user's ability to look at and integrate the separate data streams. Extracting common threads in a more automated fashion will allow an analyst to derive clear mental models of the situation.

We need to explore other areas of scientific endeavor in which multi-type data analysis is emerging as a challenge, such as the biological sciences, and consider opportunities to adapt methodologies.

Recommendation 4.9

Develop new approaches to identify changes in multi-source, multi-type, and massive data collections.

Change detection is essential to identifying emerging trends and events. In emergency situations, rapid change detection is central to effective response. Change-detection methods are required for novel structures, such as those arising from discrete events, graphs, or spatial-temporal representations.

Recommendation 4.10

Develop new methods and technologies for capturing and representing information quality and uncertainty.

Quality and uncertainty measures must be preserved throughout the data transformation process and must be represented in a form that will permit their incorporation into visual representations. Accurate understanding of uncertainties is essential to the analytical process.

Recommendation 4.11

Determine the applicability of confidence assessment in the identification, representation, aggregation, and communication of uncertainties in both the information and the analytical methods used in their assessment.

The focus should be on leveraging the visual and spatial ability of the human brain in dealing with uncertain dynamic information. Any assistance in assessing the confidence of an analysis is of direct benefit to an analyst.

Summary

Data representation and transformation provide the mathematical foundations for visual analytics. They are essential to the success of visual analytics approaches.

Advancing the state of the art in data representation and transformation will facilitate computer processing and communicating the information and knowledge content of large, complex, dynamic, and diverse data repositories. Crosscutting research in information and knowledge representation approaches and into methods for transformation of these representational sets is essential to provide the underlying structure to support visualization.

Analysts need a complete set of tools to help them understand massive amounts of data assembled from numerous sources. We strongly believe that the techniques and recommendations in this chapter will expand even further. It is much too early

in the evolution of visual analytics to know what data representation and transformation techniques will work best in a given situation. We will explore individual techniques and document the results to build long-term selection guidelines that will be based on the value of particular transformation techniques.

Summary Recommendations

The following high-level recommendations summarize the detailed recommendations from this chapter. These represent the path forward for continued research and development to provide the data representations and transformations needed for use in generation of the visual forms necessary for visual analytics.

Recommendation

Develop both theory and practice for transforming data into new scalable representations that faithfully represent the content of the underlying data.

From the standpoint of the analyst, border guard, or first responder, information provides guidance, insight, and support for assessments and decisions. Our goal is to illuminate the potentially interesting content within the data so that users may discover important and unexpected information buried within massive volumes of data. Each type of data presents its own challenges for data representation and transformation. In most cases, data representations are not meant to replace the original data but to augment them by highlighting relevant nuggets of information to facilitate analysis.

We must develop mathematical transformations and representations that can scale to deal with vast amounts of data in a timely manner. These approaches must provide a high-fidelity representation of the true information content of the underlying data. They must support the need to analyze a problem at varying levels of abstraction and consider the same data from multiple viewpoints.

Data are dynamic and may be found in ever-growing collections or in streams that may never be stored. New representation methods are needed to accommodate the dynamic and sometimes transient nature of data. Transformation methods must include techniques to detect changes, anomalies, and emerging trends.

Methods exist at varying levels of maturity for transforming data. For example, there are a variety of methods for transforming the content of textual documents using either statistical or semantic approaches. Combining the strengths of these two approaches may greatly improve the results of the transformation.

Recommendation

Create methods to synthesize information of different types and from different sources into a unified data representation so that analysts, first responders, and border personnel may focus on the meaning of the data.

Complex analytical tasks require the user to bring together evidence from a variety of data types and sources, including text sources in multiple languages, audio, video, and sensor data. Today's analytical tools generally require that the user consider data of different types separately. However, users need to be able to understand the meaning of their information and to consider all the evidence together, without being restricted by the type of data that the evidence originally came in. Furthermore, they need to be able to consider their information at different levels of abstraction.

Synthesis is essential to the analysis process. While it is related to the concept of data fusion, it entails much more than placing information of different types on a map display. The analytical insight required to meet homeland security missions requires the integration of relationships, transactions, images, and video at the true meaning level. While spatial elements may be displayed on a map, the non-spatial information must be synthesized at the meaning level with that spatial information and presented to the user in a unified representation.

Recommendation

Develop methods and principles for representing data quality, reliability, and certainty measures throughout the data transformation and analysis process.

By nature, data are of varying quality, and most data have levels of uncertainty associated with them. Furthermore, the reliability of data may differ based on a number of factors, including the data source. As data are combined and transformed, the uncertainties may become magnified. These uncertainties may have profound effects on the analytical process and must be portrayed to users to inform their thinking. They will also make their own judgments of data quality, uncertainty, and reliability, based upon their expertise. These judgments must be captured and incorporated as well. Furthermore, in this environment of constant change, assessments of data quality or uncertainty may be called into question at any time based on the existence of new and conflicting information.

The complexity of this problem will require algorithmic advances to address the establishment and maintenance of uncertainty measures at varying levels of data abstraction.

References

Allison PD. 2002. *Missing Data*. SAGE Publications, Thousand Oaks, California.

Axelrod R. 1997. *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*. Princeton University Press, Princeton, New Jersey.

Belkin M and P Niyogi. 2003. "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation." *Neural Computation* 15(6):1373-1396.

Bell ET. 1965. *Men of Mathematics*. Simon and Schuster, New York.

Bellman R. 1961. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, New Jersey.

Berkowitz B and A Goodman. 1989. *Strategic Intelligence*. Princeton University Press, Princeton, New Jersey.

Blei DM, AY Ng, and MI Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993-1022.

Braverman A. 2004. "Statistical Challenges in the Production and Analysis of Remote Sensing Earth Science Data at the Jet Propulsion Laboratory." In *Proceedings of the Statistical Analysis of Massive Data Streams Workshop*. National Academies Press, Washington, D.C.

Burrough PA and AU Frank. 1995. "Concepts and Paradigms in Spatial Information: Are Current Geographical Information Systems Truly Generic?" *International Journal of Geographical Information Systems* 9(2):101-116.

Caid R and JL Carleton. 2003. "Context Vector-Based Text Retrieval." A Fair Isaac White Paper available at http://www.fairisaac.com/NR/rdonlyres/635C0BCA-2226-4C17-AD07-FD25913B331B/0/contextvectorwhitepaper.pdf.

Caid W and P Onig. 1997. "System and Method of Context Vector Generation and Retrieval." U.S. Patent 5,619,709.

Dasu T and T Johnson. 2003. *Exploratory Data Mining and Data Cleaning*. Wiley-Interscience, New York.

Davis L. 1987. *Genetic Algorithms and Simulated Annealing*. Morgan Kaufmann, San Francisco.

Debevec PE. 1996. *Modeling and Rendering Architecture from Photographs*, Ph.D. Dissertation, University of California at Berkeley, Berkeley, California.

Deerwester S, S Dumais, T Landauer, G Furnas, and R Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the Society for Information Science* 41(6):391-407.

Fausett L. 1994. *Fundamentals of Neural Networks*. Prentice-Hall, Englewood Cliffs, New Jersey.

Forsythe GE, MA Malcolm, and CB Moler. 1977. "Least Squares and the Singular Value Decomposition." Chapter 9 in *Computer Methods for Mathematical Computations*. Prentice Hall, Englewood Cliffs, New Jersey.

Gentleman R. 2004. "Using GO For Statistical Analyses." In *Proceedings of the 16th COMPSTAT Conference*, pp. 171-180. J Antoch, ed., Physica Verlag, Heidelberg, Germany.

Glassner AS. 1995. *Principles of Digital Image Synthesis*. Morgan Kaufmann, San Francisco.

Haykin S. ed. 2000. Vols. 1 & 2. *Unsupervised Adaptive Filtering*. Wiley-Interscience, New York.

Hetzler E and A Turner. 2004. "Analysis Experience Using Information Visualization." *IEEE Computer Graphics and Applications* 24(5):22-26.

Hofmann T. 1999. "Probabilistic Latent Semantic Indexing." In *Proceedings of the Twenty-Second Annual SIGIR Conference on Research and Development Information Retrieval*, pp. 50-57. ACM Retrieval Press, New York.

Holland J. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, Michigan.

Hooker J. 2003. *Working Across Cultures*. Stanford University Press, Stanford, California.

Ilgen M, J Sirosh, and W Chonghua. 2000. *Novel Self-Organizing Neural Network Methods for Semantically Accessing Unstructured, Multilingual, Multimedia Databases, Final Report*. DARPA Collaboration, Visualization, and Information Management Project: Multilingual and Multimedia Information Retrieval.

Jacobsen R. 2004. "Statistical Analysis of High Energy Physics Data." In *Proceedings of the Statistical Analysis of Massive Data Streams Workshop*. National Academies Press, Washington, D.C.

Jurafsky D and JH Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Processing*. Prentice-Hall, Englewood Cliffs, New Jersey.

Johnson NF, Z Duric, and S Jajodia. 2001. *Information Hiding: Steganography and Watermarking – Attacks and Countermeasures*. Kluwer Academic, Boston.

Kasik DJ. 2004. "Strategies for Consistent Image Partitioning." *IEEE Multimedia* 11(1):32-41.

Lafon S. 2004. *Diffusion Maps and Geometric Harmonics*, Ph.D. dissertation, Yale University, New Haven, Connecticut.

Leser U and J Freytag. 2004. "Mining for Patterns in Contradictory Data." In *Proceedings of the 2004 International Workshop on Information Quality in Information Systems*. June 18, 2004, Paris, France. Available at http://www.informatik.uni-trier.de/%7Eley/db/conf/iqis/iqis2004.html.

Little RJ and DB Rubin. 1987. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York.

Mark DM, C Freksa, SC Hirtle, R Lloyd, and B Tversky. 1999. "Cognitive Models of Geographical Space." *International Journal of Geographical Information Science* 13(8):747-774.

Maurer SB and A Ralston. 2004. *Discrete Algorithmic Mathematics*. AK Peters, LTD, Wellesley, Massachusetts.

Miller G. 1998. "Five Papers on WordNet." In *WordNet: An Electronic Lexical Database*, ed. C Fellbaum. MIT Press, Cambridge, Massachusetts.

Mintz AP. 2002. *Web of Deception: Misinformation on the Internet*. Cyberage Books, Medford, New Jersey.

Montague R. 1974. "The Proper Treatment of Quantification in Ordinary English." In *Formal Philosophy: Selected Papers of Richard Montague*, R Thomason, ed. Yale University Press, New Haven, Connecticut.

Moore A. 2004. "Kd-, R-, Ball-, and Ad- Trees: Scalable Massive Science Data Analysis." In *Proceedings Statistical Analysis of Massive Data Streams Workshop*. National Academies Press, Washington, D.C.

Peuquet DJ. 2002. *Representations of Space and Time*. The Guilford Press, New York.

Polanyi L and A Zaenan. 2004. "Contextual Lexical Valence Shifters." In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*. Technical Report SS-04-07. AAAI Press, Menlo Park, California.

Robertson G, M Czerwinski, and J Churchill. 2005. "Visualization of Mappings Between Schemas." *Proceedings of CHI 2005*, pp. 431-439. ACM Press, New York.

Rosen-Zvi M, T Griffiths, M Steyvers, and P Smyth. 2004. "The Author-Topic Model for Authors and Documents." In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 487-494. Banff, Canada.

Roweis, S and L Saul. 2000. "Nonlinear Dimensionality Reduction by Locally Linear Embedding." *Science* 290(5500):2323-2326.

Rumsfeld D. 2002. "Secretary Rumsfeld Press Conference at NATO Headquarters, Brussels, Belgium." Accessed April 28, 2005 at http://www.defenselink.mil/transcripts/2002/t06062002_t0606sd.html.

Salton G. 1968. *Automatic Information Organization and Retrieval*. McGraw-Hill, New York.

Salton G, A Wong, and C Yang. 1975. "A Vector Space Model for Automatic Indexing." *Communications of the ACM* 18(11):613-620. ACM Press, New York.

Scott DW. 1992. *Multivariate Density Estimation*. Wiley-Interscience, New York.

Steyvers M. 2002. "Multi-Dimensional Scaling." In *Encyclopedia of Cognitive Science*. Nature Publishing Group, London.

Tenenbaum, JB, V de Silva, and JC Langford. 2000. "A Global Geometric Framework for Nonlinear Dimensionality Reduction." *Science* 290(5500):2319-2323.

Wang Z, E Chow, and R Rohwer. 2005. "Experiments with Grounding Spaces, Pseudo-counts, and Divergence Formulas in Association-Grounded Semantics." In *Proceedings of the 2005 Conference on Intelligence Analysis*. Available at https://analysis.mitre.org/proceedings/Final_Papers_Files/13_Camera_Ready_Paper.pdf.

Yuan M, D Mark, M Egenhofer, and D Peuquet. 2004. "Extensions to Geographic Representation: A Research Agenda for Geographic Information Science." In *Research Challenges in Geographic Information Science*, pp. 129-156, eds. R McMaster and L Usery. CRC Press, Boca Raton, Florida.

Zhou L, D Twitchell, T Qin, J Burgoon, and J Nunamaker. 2003. "An Exploratory Study into Deception Detection in Text-based Computer-Mediated Communication." In *Proceedings of 36th Hawaii International Conference on System Sciences*, January 6-9, 2003.