Contents

1 Statement of Work

1.1 Summary of Proposed Work

Goal: We propose to develop a new visual exploration paradigm emphasizing integrated data/metadata analysis and to use it to enhance the capabilities of the DEVise [LRM96] visualization and exploration tool. DEVise is a public-domain tool that brings together advanced database technology, a flexible visualization framework, and intuitive interfaces to support *retrieval, integration and visual analysis of large datasets from one or more sources*, including tabular datasets, time-series and imagery. By building upon this tool, we will conduct our research in an open software environment and will leverage current funding from NASA and ORD. In addition to space science applications, DEVise has been applied to analyse logs from parallel computations and the UW Clinics, and to support financial applications that use data from a variety of sources including Web repositories, Compustat, CRSP and ISSM, and sales information from a catalog merchant.

Research Topics: As larger and larger datasets are collected and generated by scientists, and the number and diversity of these datasets increases, effective tools for data retrieval, integration and analysis become essential. Given that the Web is quickly becoming the most widely used means of connecting users to datasets, tools that can efficiently and conveniently access local data together with remote data over the Web are especially timely and useful. From a research standpoint, several issues require attention in developing such tools:

1. *Dataset Size:* Tools should be able to deal with datasets larger than main memory by using buffering to cache data from secondary and tertiary storage devices. While this is standard for database query engines, it is not so for visualization tools. Further, as the amount of graphical data increases, indexing techniques are required for ensuring interactive response times.
2. *Integrated Presentation of Data and Metadata:* As datasets increase in number, diversity and size, *metadata* becomes an essential tool for managing this complexity. We use the term 'metadata' in a general sense to mean *a summary of a subset of data*. A powerful paradigm is to allow users to visualy explore data and (various kinds of) metadata in a uniform manner, and to *drill down* to lower levels interactively. How to achieve this, especially in the context of a visual querying environment, is a key element of this proposal.
3. *Diversity, Distribution and Autonomy of Data Sources:* The Web can be thought of as a gigantic distributed database, and queries over data from multiple sites can be thought of as distributed database queries, but there are important differences: typically, there is little or no data processing capability at the site where the data is stored, the number and diversity of data sources is enormous, and there is complete autonomy across sites.
4. *Integration of Database-style Querying and Visualization:* Allowing users to define visual presentations in a flexible manner, and to pose complex queries in terms of these visual presentations, greatly facilitates data retrieval and analysis for users who are not database experts.

Expected Results:

1. An enhanced version of DEVise that is capable of supporting data integration and analysis tasks, with emphasis on visual exploration and queries over the Web. In contrast to other visual tools, DEVsie will allow database-style queries to be combined easily with visual data exploration, will support uniform querying of data and metadata, and will implement queries over the Web efficiently.

2. An environment for analyzing data at multiple levels of detail, within the context of DEVise. We believe that understanding how to use metadata and summary data effectively in exploring data is central to interactive analysis of large datasets. Various tools can be used to produce a high-level summary of data, and the visualization tool must support integrated viewing of the summary data, or *metadata*, as well as the original data. This requires the ability to inform the visualization tool about the mapping from a summary object to the data records that correspond to it.

   Specifically, we will develop an enhanced version of the BIRCH [ZRL96] clustering package (which produces summaries that are 'clusters' in the data) integrated fully with DEVise, to support interactive classification tasks. More generally, we will provide interfaces to enable other kinds of tools to generate summaries and pass to DEVise the mapping from summary objects to data records.

3. A Carbon Budget Visual Workbench for the Boreal Ecosystem Atmoshere Study (BOREAS). Using the enhanced version of DEVISe and BIRCH, we will build a system that visually integrates the datasets needed to study and to characterize the carbon budget of the boreal forest. This involves handling images from aircraft and satellites, long time series of weather data and carbon fluxes, classified images of vegetation type and density, images to characterize forest canopy architecture collected with a multiband vegetation imager and integration of all these data with models.

Project Team: This project involves faculty members at the University of Wisconsin-Madison, graduate students, and a staff programmer. Prof. Livny has led the development of the DeNet simulation package and the Condor management system for workstation clusters. Prof. Ramakrishnan has led the development of the Coral deductive system. These are research software packages that are considered to be state-of-the-art, and have been used widely in universities and research labs in the US and abroad. Profs. Livny and Ramakrishnan have collaborated in the past, with NASA support, to develop the current versions of the DEVise and BIRCH tools and a system (called SEQ) for querying sequence data, and have also combined Coral and Condor (using ORD support) to provide query evaluation capabilities utilizing idle resources. Prof. Norman has worked with measurements and integrative modeling in the area of environmental biophysics. He has developed a soil- plant-atmosphere model, named Cupid, that has been applied to problems in remote sensing, irrigation scheduling, integrated pest management and water and carbon budgets of ecosystems. Prof. Norman has been the domain scientist driving the development and use of DEVise and BIRCH.

1.2 Introduction

Many large and diverse datasets are being generated and made available through the World-Wide Web by scientists, in large projects such as NASA's EOS and BOREAS, as well as numerous smaller projects. Ease of access and effective analysis tools are crucial to the successful exploitation of these datasets for scientific research and economic growth. The following trends have been motivating our work, and we aim to continue to develop paradigms and build tools that address the attendant problems:

1. Increasing size of datasets, which implies that visualization tools must be able to deal with data spread over secondary and tertiary storage devices. Further, these tools must allow users to examine data at several levels of detail, ranging from individual observations to *metadata* that summarizes a large volume of related data.

2. Increasing variety and number of data sources, which means that there must be a convenient way to integrate and analyse data from multiple remote sources, and to combine this information with local data.

3. The need for visual exploration of data to complement analytical tools. This is especially important when data is noisy or when the user is looking for loosely defined patterns or trends in the data.

4. The Web as a medium for connecting researchers to data sources. Ideally, users should be able to efficiently query data from several Web sites as if all the data were stored on the browsing machine;

This proposal is a logical continuation of the work carried out under our current AISRP NASA grant, in which we have developed the DEVise [LRM96] visual exploration tool and the SEQ system for analytical queries over sequences (as well as the BIRTCH data clustering tool). We intend to expand the scope of our work to include more general data sources, datasets, and queries, and to emphasize two new issues:

1. Integrated querying and visualization of data and *metadata*.

2. Analytical data transformation and querying capabilities, with an emphasis on efficient access to data sources through the Web.

These two changes to our research direction are based largely on experience with the current tools and feedback form our users. Originally, DEVise was intended to address just visual exploration. Computation and transformation of the original dataset were carried out by the CORAL system for deductive queries and the SEQ system for queries over sequences. While this paradigm of loosely coupling the data transformation tools with the visualization tool does provide a powerful paradigm, many of our users clearly wanted a tool that seamlessly combines data transformation and visualization. This experience has led us to enhance the data transformation and querying capabilities within DEVise. A direction we intend to continue and follow as part of this proposal.

The issue of integrated access to data and metadata also arose out of user feedback on using DEVise in conjunction with the BIRCH [ZRL96] clustering tool. The clusters produced by BIRCH can be seen as a summary of the original data, or as a form of *metadata*, providing a compact, high-level description. Each cluster provides summery information and represents a collection of records in the original dataset. From our interaction with users of BIRCH and DEVise, we learned that users would like to cluster a dataset with a particular choice of clustering parameters (e.g., weights for the different dimensions, choice of distance metric), visualize the clusters, then select all records in the *original dataset* corresponding to one or more clusters and process these records further. Clusters are by no means the only useful kind of summary, but it is a very good canonical example. Other examples of summaries include statistical measures, lossy compressed versions of images, moving averages over a time-series. Our goal is to provide a convenient environment for generating summaries using an appropriate tool, and exploring both the summaries and the original data together in a natural manner. The key here is to enable the visualization tool to connect between a summary to the original data.

As concrete applications of our tool, we envisage the following:

- **Quality Assurance:** Is the data noisy? Is there a lot of missing data? Is the loss/noisiness concentrated in certain 'regions' of the data?

- **Exploratory Data Analysis:** Can we spot interesting patterns or trends in the data, especially when the datasets are very large and come from different sources?

Such exploratory analysis is carried out routinely in many different application domains, but software environments supporting this activity are in their infancy. The problems that must be addressed and the techniques that can be applied while studied in Statistics, Artificial Intelligence, and in Databases, have not been yet integrated with visualization tools.

- **Hypothesis Verification:** If we believe that there is a causal relationship between ozone levels in the atmosphere and the incidence of skin cancer, it would be helpful to be able to quickly browse relevant datasets to see if they might support such a hypothesis. Then, additional tools can be brought to bear upon such promising datasets.

- **Queries Over the Data:** Scientists often need to identify parts of the data that satisfy certain criteria. Sometimes, determining which part of a dataset satisfies a criteria involves complex analysis; often, however, it can be expressed as a database-style query, e.g., "Find days when the temperature was higher than some value,".

- **Concept Identification and Classification:** Can we identify instances of domain-specific concepts in the data? For example, given several images collected by a digital camera, can we identify *sunlit leaves*, *shaded leaves*, and *branches*? (This is an example of an application that we can already handle, although with more manual intervention than we would like to have ultimately, with the current versions of the DEVise and BIRCH tools.)

1.3 Benefits of Proposed Work

A major objective of BOREAS is to evaluate the carbon budget of a 1000 km X 1000 km area in central Canada and determine whether this region is a sink or source of atmospheric carbon. A comprehensive field experiment began in August 1993 (will end in October 1996) to combine measurement expertise of ecologists, atmospheric and soil scientists with the capability of remote-sensing scientists through the use of integrative models. These models, Cupid being an example, require an extensive array of input data related to the atmosphere, soil, vegetation, topography, ground water levels, etc., and an intensive period of validation where model outputs are compared with point measurements of carbon fluxes from the field experiment. This model validation is exceedingly labor intensive because of the need to acquire and evaluate a vast array of model-input data and validation data from numerous sources including the BOREAS Information System (BORIS), individual investigators on the World Wide-Web and image data generated locally. Key model input data that are required includes vegetation type maps (generated by NASA), weather inputs, and forest canopy architectural characteristics associated with various vegetation types.

DEVise has proven to be an excellent tool for processing our input data and the data generated by our models. It has dramatically improved our ability to assess suitable periods when appropriate weather events occur and data is available coincident with measurements of output validation data. By enhancing DEVise to support intergarted visualization of data and metdata from different sources, it will provide us with visual capabilities needed to study the carbon budget of the BOREAS forest. It will allow us to visually evaluate and analyze forest canopy images that will provide canopy architectural characteristics for various vegetation types, including leaf area index, radiation extinction coefficients and crown clumping factors. These vegetation characteristics are essential to all carbon-budget models of the boreal region. Considering that dozens of variables are measured every 30 minutes at numerous sites over several years, the capability of DEVise to provide interactive visualization of multiple, long, time series would save time and effort and simultaneously provide greater insight into the quality and characteristics of these complex data sets.

1.4 Software that We Propose to Enhance and Build Upon

Our overall goal is to develop a tool that provides integrated visualization of data and metadata of large datasets from local and Web sources. There are many aspects to such a tool, and we intend to enhance existing tools and software, rather than to develop a new tool from scratch. By doing so, we significantly leverage existing and prior support from NSF, NASA, ORD and the Packard Foundation. In this section, we describe the tools that we intend to enhance, and related software that we intend to build upon.

1.4.1 DEVise: A Tool for Visual Data Analysis and Exploration

Large and complex data sets present a challenging task for visualization systems. The number of records in a data stream can easily be in the millions, its size may reach tens or hundreds of megabytes, and each record in the stream may contain hundreds of data attributes. It is important for visualization systems to avoid overwhelming the main memory or virtual memory of a workstation, and to allow the user to navigate through the data and metadata without being buried in it.

We have developed a data visualization and exploration system called DEVise that attempts to solve both the data size and metadata complexity issues. The system uses off-the-shelf hardware and has minimal main memory requirements. It can handle data that is large in size, complex in structure (multidimensional and/or hierarchical), and that may be imported from different sources such as database servers, external programs, and World Wide Web resources. Commercial and scientific databases can also be linked to DEVise to allow the user to visualize and analyze related information from heterogeneous sources.

As an example of the use of DEVise, consider the following scenario, which is based on our collaboration with a Soil Sciences user-group. Extensive manual effort is currently required to extract useful data from the FIS or BORIS data bases (maintained by NASA). For example, to compare a prediction of infrared temperature from a soil-plant-atmosphere model (named Cupid) with aircraft and satellite observations, input data (soil, vegetation and weather data) required by the model must be obtained from the data base subject to the constraint that the site was cloud-free at the time of satellite and aircraft observation. Furthermore, when a cloud-free period is obtained all the model input data must be available (weather data is collected on half-hour intervals, surface soil moisture measured daily, leaf area and deep soil moisture measured weekly, with the time of observation varying from station to station) from the previous day of rain to the time of coincidence of clear sky and satellite overpass either by direct measurement or interpolation in time or space. Extracting the appropriate data from the database is itself time consuming, but it does not require as much time as just finding an appropriate time sequence that meets the criteria, which can take months. We note that SQL is wholly inadequate for tasks such as this.

Missing or erroneous data are among the major complicating factors that make this selection process difficult. After the inputs required by the model have been extracted, the model is run; then model predictions of infrared temperatures are compared with measurements from aircraft or satellites, which also need to be extracted from the data system. Typically such comparisons involve time sequences of predictions and measurements. For example, predicted and measured infrared temperature could be plotted as a function of time; they may agree at some times and disagree at other times. On other occasions the two curves may just be shifted in time by some fixed phase. Alternatively they may agree at high wind speeds and disagree at low wind speeds. Clearly, the ability to visually browse and correlate such datasets is invaluable.

Defining Visual Queries in DEVise

In essence, we can think of *visual queries* in DEVise as definitions of views that integrate the source data, which can come from a variety of sources including text files, sequence databases, and the World-Wide Web. Using a point-and-click interface, a user can define a 'visual query' that determines how the source data is to be interpreted and displayed on the screen.

In DEVise, a *data stream* consists of binary or ASCII records where each record contains one or more *attributes*. A data stream is combined with a *schema* to produce *TData*. The schema describes the data attributes in the data stream. Each TData record is mapped to a *GData* record where each GData record consists of graphical attributes. This *mapping* can be defined by the user, and provides a very high degree of flexibility in defining visual queries, as the following examples illustrate.

Example Applications of DEVise

We now present some example applications of DEVise, to illustrate the power of the current tool and its scope. The first example, shown in Figure DEVise-1, illustrates the combined use of DEVise and BIRCH and is discussed in Section 1.5.2. The second example, shown in Figure DEVise-2, illustrates the use of DEVise to visualize the logs of the Distributed Batch Controller (DBC) [CSL96]. Here we present information on the execution of 5000 inversion jobs of TOVS brightness temperatures to produce temperature and humidity profiles of the Earth's polar regions submitted by the Polar Science Center at the University of Washington [PSC96] to a DBC system that consists of Condor pools at the University of Washington, University of Wisconsin, and NASA's Goddard Space Flight Center in Maryland.

The third example, shown in Figure DEVise-3, illustrates the flexibility of DEVise. The data that is visualized is a collection of traces that reflect the execution of a parallel material sciences application on a cluster of workstations. The application is written in PVM and uses the Resource Management services of Condor via the Condor Application Resource Management Interface.

The fourth example, shown in Figure DEVise-4, illustrates the use of DEVise to interpret output from the Cupid model. Photosynthesis is predominantly influenced by light, but the influence of larger values of vapor pressure deficit can be seen clearly.

1.4.2 BIRCH: A Tool for Clustering Multidimensional Datasets

Finding natural 'clusters' in a multidimensional dataset is a well-studied problem in Statistics, and has been applied in a variety of domains for concept identification and classification. It can be regarded as a variant of the problem of computing a *density function*; this has applicability for constructing histograms of data, for example. Clearly, clustering is a powerful technique for exploratory analysis.

We have developed a clustering algorithm called BIRCH [ZRL96] that is capable of clustering very large datasets while using very little memory. It incrementally and dynamically clusters incoming multi-dimensional metric data points, and can typically find a good clustering with a single scan of the data, and improve the quality further with a few additional scans. A novel feature is that BIRCH exploits its knowledge of the portion of the input dataset that it has already clustered to fine-tune the clustering of the rest of the dataset. Thus, it adjusts dynamically to the input dataset to try and produce the best quality clustering with the available resources (i.e., available memory and time constraints). The low memory requirements of BIRCH enable us to cluster on different subsets of the data attributes (i.e., dimensions) in parallel, and to collect useful statistical information on attributes that do not take part in the clustering.

1.4.3 Database Query Tools

Over the years we have accumulated a large and rich software-base for query processing from database systems like CORAL, SEQ and more recently Minibase that we have developed.

1.4.4 Multidimensional Constraint Queries

Constraints offer a very powerful way of formulating queries over multidimensional point datasets, and can be used in a variety of applications, including scientific and business. The idea is that a desired region can be specified as a conjunction of polynomial constraints. As an example, a convex polygon can be represented as a *conjunction of two-dimensional linear constraints.* Each constraint identifies a half-plane; a conjunction of constraints can therefore be used to identify any convex region.

The R^*-Tree index is a popular multidimensional index structure and can be used to store multi-dimensional point datasets (as well as non-point data). Boxes that are orthogonal to the axes are the standard approximation for data objects and queries in such indexes. However, a box is often a poor approximation, especially for queries and data that are irregularly shaped and not suitably oriented with respect to the axes.

An alternative is to use a conjunction of constraints to describe the bounding envelope of data objects and queries. Storing conjunctions of linear constraints can be expensive in terms of space; we do not consider this approach. We do propose, however, that queries be approximated by such conjunctions; the only overhead is a more complex overlap test, which we demonstrate can be carried out very efficiently [GRSY96].

To illustrate the kinds of queries that can be supported with such index structures, let us consider a portion of the Compustat stock and business data set with attributes *year number, earnings per share, stock low, stock high,* and *stock value at year close.* One example query using constraints is: *Retrieve all tuples that have a price-earnings (P/E) ratio less than some constant.* Investors frequently use the P/E ratio as a measure of how undervalued or overvalued a company is. In terms of constraints, the above query can be expressed as: $price_{year-end} = r * earnings \leq 0$ (where r is the ratio value). As another example, consider: *Retrieve all tuples that correspond to companies whose value didn't vary by more than some percent of the price at year's end.* In terms of constraints: $price_{high} - price_{low} - p * price_{year-end} \leq 0$ (where p is the percentage above). This measure is indicative of a stock's *volatility.*

We have developed an extension to the R^*-Tree index that can support multidimensional constraint queries very efficiently, and implemented a package based upon this extension. We intend to use this package both within the visual query processing engine of DEVise (to index graphical data) and to provide constraint query capabilities in the data transformation engine.

1.5 Research Goals

In this section, we present the motivation for and description of the appraoch we plan to develop and how we propose to apply it to the tools and software mentioned above. In essence, we propose to extend the capabilities of DEVise in three main directions: improved *visual query* capabilities, a framework and interfaces for *integrated data and metadata exploration*, and increased support for *data transformations.* For each of these directions, we outline the functionality needed and then indicate the extensions required to the current version of DEVise to provide these functionalities. Of the currently available tools for visual data exploration (e.g., Collage [COL96], LinkWinds [JBO94], and VISTAS [VIS95]), we believe that DEVise [LRM96] is best suited to serve as the software environment for our research and as a vehicle to place the results of our work in the hands of the scientific community. DEVise brings together advanced database technology, a flexible visualization franmework, and intuitive user interfaces to support the visual exploration needs of ordinary scientists with ordinary hardware resources. It was developed by us with prior NASA support and is currently in use by several groups (e.g., the business school, scientists in soil science, the NMR data bank, the Paradyn parallel processing research team, the UW clinics, a catalog merchant; it is also under consideration to serve as a frontend to a well known Data Mining product, and by LBL Laboratories).

1.5.1 Visual Queries

Visual browsing or **visual querying** of the data can be an extremely powerful way to explore large datasets. Several properties of a visual browser are important for the applications that we target:

1. In addition to providing some convenient way of using common visual presentation 'templates' such as scatter plots or bar graphs, the system should provide a flexible way to construct a broad class of new templates and to inter-link them so that each dataset can be viewed in the most natural manner for any given application. To view data, the user must describe how it is to be visually presented; this includes a choice of existing templates or construction of new templates, establishing links between them, and a specification of how the data records are to be mapped into a visual presentation based on these templates. The graphical data in linked views changes automatically in natural ways when one view is manipulated through a visual operation (e.g., a cursor operation, zooming, color filtering, size selection, etc.). We refer to the combination of templates, links, and mappings used to describe a visual presentation as a *visual query*.

2. The definition of a 'visual query' should be cleanly separated from the actual dataset, so that it can be applied to other similar datasets, and the query definition facilities should be flexible, in order to allow users to 'see' their data in as many different ways as possible.

3. The browser should be capable of dealing with very large datasets, even if the datasets are much larger than available memory.

Extending DEVise Support for Visual Queries

DEVise already meets many of our goals in the area of visual queries, but there are some important directions in which it needs to be enhanced. We intend to work on the following tasks:

1. Separate query definitions from details of the underlying data sets. This will make DEVise a powerful tool for *collaborative* computing: a user can navigate the data, and other connected users will see the same data display simultaneously, including all the changes performed by the original user. Since *each* user has the ability to do such navigation, in effect each user sees an *active* data presentation.

2. Improve various implementation-level aspects of DEVise, e.g., 3D symbols, color management, network connections, etc.

3. Build 'drivers' for data sources currently not supported by DEVise. (We have drivers for text files, SEQ data streams, Web URLs, Compustat data, etc., but not for HDF, for example).

Extensions related to other aspects such as integrated data and metadata browsing and data transformations are described in Section 1.5.2 and 1.5.3.

1.5.2 Integrated Access to Data and Metadata

Visualization of very large datasets presents several challenges. A basic problem is that the dataset size may exceed the amount of main memory available. This problem has been addressed in DEVise by incorporating a *buffer manager*, which brings in data records as needed, if necessary making space in the buffer pool by discarding records that are not currently needed. Even with intelligent buffer management policies, however, interactive response times cannot be achieved for very large datasets, and too much information is lost by compressing a very large volume of data onto a single screen.

A powerful paradigm for addressing this fundamental problem is to let users create summaries of data (which are typically much smaller than the original dataset) and to browse the summaries, or metadata, to get an overview of the entire dataset. Subsequently, users can look at interesting portions of the data in more detail; in our experience, users in fact want to be able to interleave the browsing of data and metadata. The familiar operation of *zooming* is just one instance of this paradigm; more generally, there are several levels of summary data, or metadata, and users want to interleave browsing of the metadata and browsing of relevant portions of the actual data as needed.

As a concrete example, we have demonstrated the feasibility of integrating the BIRCH clustering tool with DEVise. To illustrate the combined use of BIRCH and Devise, we present an actual application involving the interpretation of forestry-canopy images. We have a pair of images of a forest-canopy with a partly cloudy sky as the background, taken at two different wavelengths. One image is in the near-infrared band (NIR), and the other image is in visible wavelength band (VIS). Each image contains 512x1024 pixels, and each pixel has a pair of brightness values corresponding to NIR and VIS. As part of the BOREAS field experiments, soil scientists will receive hundreds of such image pairs and will have to separate background, sunlit leaves, shaded leaves, and branches.

We applied BIRCH to the (NIR,VIS) value pairs for pixels in an image (512x1024 2-d tuples) by using 400 kbytes of memory (about 5% of the dataset size) and 80 kbytes of disk space (about 20% of the memory size), and weighting NIR and VIS values equally. We obtained 5 clusters that correspond to (1) very bright part of sky, (2) blue sky, (3) clouds, (4) sunlit leaves (5) tree branches and shaded leaves on the trees. This step took 284 seconds.

However the branches and shaded leaves were too similar to be distinguished from each other, although we could separate them from the other cluster categories. So we extracted the part of the data corresponding to (5) (146707 2-d tuples) and used BIRCH again. But this time, (1) NIR was weighted 10 times heavier than VIS because we observed that branches and shaded leaves were easier to tell apart from the NIR image than from the VIS image; (2) BIRCH was able to draw finer distinctions because it processed a smaller dataset with the same amount of memory. The two clusters corresponding to branches and shadows were obtained in 71 seconds.

The result of clustering using BIRCH is shown in Figure DEVise-1. Visually, we can see that it is a satisfactory filtering of the original image according to the user's intention. While this was a very successful effort in terms of the clustering results that we obtained, the interaction between BIRCH (to do clustering) and DEVise (to display the results) was tedious.

The important point to be noted in this example of using BIRCH and DEVise together is that the clusters produced by BIRCH can be seen as a summary of the original data: these clusters can thus be regarded as a form of *metadata*, providing a compact, high-level description. This kind of summary is by no means the only useful one, but it is a very good canonical example. In our interaction with users of BIRCH and DEVise, it has become very clear that users want to be able to cluster a dataset with a particular choice of clustering parameters (e.g., weights for the different dimensions, choice of distance metric), see the clusters in a visual presentation, then select all records in the *original data* corresponding to one or more clusters and process these records further. In other words, they use the clusters produced by BIRCH to obtain a high-level overview of the data, and narrow the scope of subsequent detailed analysis to interesting portions of the data.

Extending DEVise to Handle Integrated Data/Metadata Browsing

The clusters produced by BIRCH are only one example of a summary description of data. Other examples of summaries include:

1. Statistical measures over subsets of the data. Indeed, such summaries are so useful that some support is built directly into the current version of the visualization engine of DEVise.

2. Compressed versions of images [LR96].

3. Moving averages over a time-series, with the window size controlling the degree of summarization. Such a summary could be generated using SEQ, for instance.

4. The result of an SQL Groupby query, which summarizes all records with the same value in the grouped field into a single record (using statistical functions like *sum* or *average*).

5. Patterns extracted by data mining techniques for finding association rules [AS94] or sequential patterns.

As an example of integrated data/metadata browsing using some of these other kinds of summaries for metadata, we can obtain a monthly average of a series of daily temperature readings, using SEQ. A user may then want to see a scatterplot of daily temperature versus rainfall for all months in which the average temperature is above a certain value.

As another example, a user may want to browse a large collection of satellite images (perhaps with additional selection criteria such as the time and location the images were recorded, which can be expressed using the data transformation capabilities described in Section 1.5.3). By compressing the images, it becomes feasible to look directly at several images interactively; but having browsed this 'metadata' representation, the user may want to see a subset of the images in greater detail. Our work on quality-controlled compression [LR96] can be used to generate compressed images, and a natural interleaving of compression and browsing can be supported by incorporating this technology into DEVise;

More generally, image summaries can be generated in several other ways as well, for example, by using mathematical tranforms (e.g., wavelets, DCT, Fourier, Karhunen-Loewe) that concentrate most of the 'information' in an image into a small vector of numbers. Each such vector summarizes an image, and comparing these summaries can be a useful filter in further processing of the image set: similar images have summaries that are 'close', using an appropriate distance metric. While we do *not* propose to directly address image processing issues, we note that visualizing such summaries, and then interactively looking at specific images, is done very naturally in the DEVise data/metadata browsing paradigm.

Our goal is to turn DEVise into a convenient environment for generating summaries using an appropriate tool, and analysing both the summaries and the original data together in a natural manner. The key here is to enable the visualization tool to understand the mapping between a summary object and the original data. We intend to develop BIRCH further (e.g., to cluster non-metric data), and to integrate it fully with DEVise. In doing so, we expect to abstract from the experience and put in place a framework and a set of interfaces that will enable users to browse other kinds of summaries.

1.5.3 Data Transformation Capabilities

The DEVise tool was originally intended to address just visual exploration, with database style querying and, in general, computation and transformation of the original dataset carried out using other tools. Specifically, we provided our users with the CORAL for deductive queries and SEQ for queries over sequences. While this paradigm of loosely coupling the data transformation tools with the visualization tool did provide a powerful framework, users clearly wanted to be able to combine data transformation and visualization seamlessly within a single tool. We therefore added data transformation capabilities to DEVise, and this has proven to be very popular.

One of the objectives of this proposal is to add a data transformation, or *analytical query*, component to DEVise. This would allow source data to be combined and transformed, before it is presented visually. More generally, it would make DEVise a very powerful front-end for SQL-style query languages, in addition to its other applications, given the tight interaction between visual operations and data retrieval in DEVise: a broad class of SQL queries over a relation, including queries utilizing grouping and aggregate operations, can be expressed in terms of mouse operations on a DEVise presentation.

Developing a powerful database style query facility is a significant task, but fortunately, we have a significant code-base—and feedback from a wide user-base!—to draw upon in this effort, and indeed, we have a preliminary prototype in place. The system architecture for DEVise with all the extensions that we propose is shown in Figure 1; it closely follows the architecure of the current system.

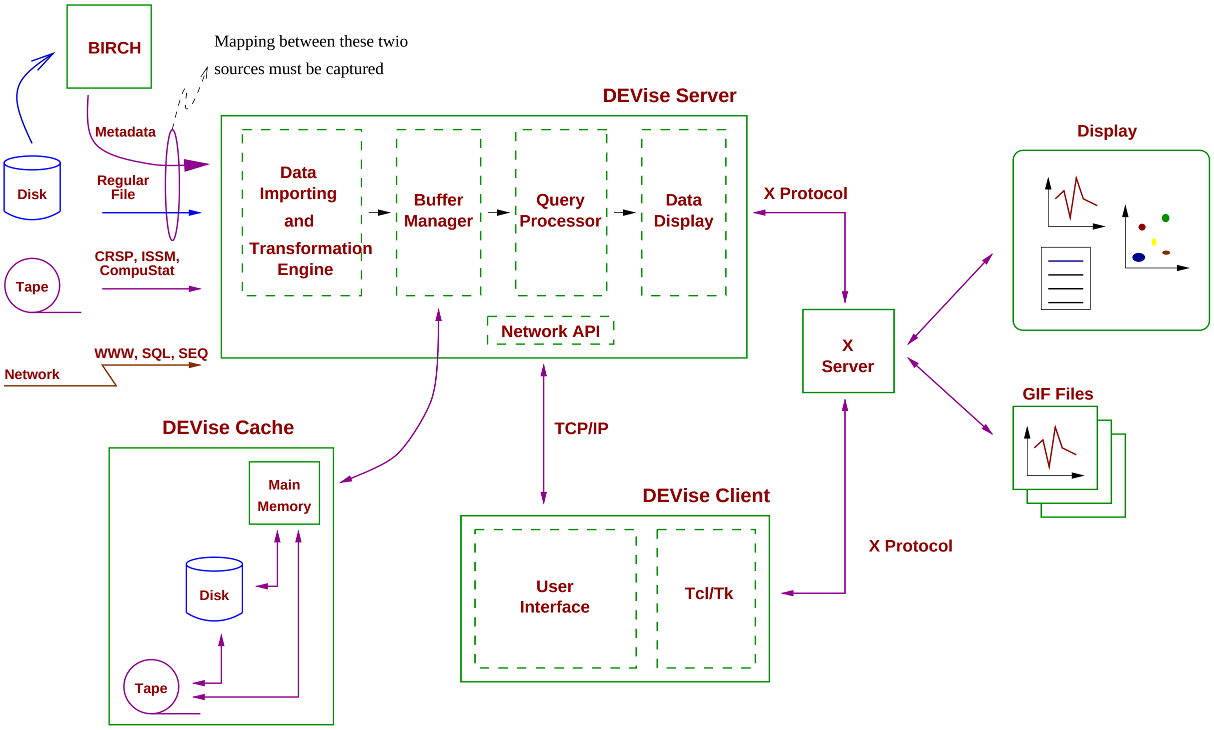

Figure 1: DEVise System Architecture

Extending the Data Transformation Capabilities of DEVise

In extending and refining the prototype data transformation facility, we intend to draw upon code from the following systems that we have developed (and distributed widely): the CORAL deductive query system, the Minibase relational DBMS, and the SEQ sequence query system.

Queries Over the Web: The data transformation component allows new tables and time-series to be defined in terms of existing tables and time-series, using the *transformation language* (or *query language*, depending on one's point of view). Efficient database-style query optimization is important, and a query optimizer taken from Minibase has already been incorporated into our prototype data transformation facility.

DEVise views can be defined over datasets stored at several Web sites, and distributed query optimization is essential to good performance. We therefore propose to extend the optimizer in the prototype to do database-style distributed query optimization for queries that access multiple sources over the Web. This would greatly enhance the use of our tool for accessing large datasets over the Web, and we have already initiated work on this extension.

A novel aspect of our paradigm is that we are developing light-weight *data wrappers* that can be used to provide efficient access in conjunction with the query optimization in the browser. For example, if a site contains a lot of tabular data (e.g., weather tables, financial time-series), the owner of the site can download the data wrapper package and use it to provide an alternative way to access the data. The wrapper supports index creation, and includes an optimizer and evaluation engine for simple queries, plus a well-defined interface that is understood by the optimizer in the DEVise browsing tool. Thus, without modifying the original files or moving them into a DBMS, local query processing capability is added, and the browser can exploit this capability to minimize data traffic and compute queries efficiently. This benefits both the owner of the data site, since more queries can be supported, given that each query requires less information to be shipped (although some local processing cost is involved), and the user of the browsing tool, who receives a response faster.

We emphasize that our design goal is to let users maintain their data in whatever format they happen to be in currently, without requiring that the data be imported into a specialized system such as a conventional DBMS or converted to a different file format. We will use drivers to read data in various file formats, and include any necessary processing capabilities in our data transformation engine and the source wrappers. Further, the use of source wrapper code is *entirely optional*. If a user does not wish to install and use this software or provides selection and projection services via other means (like the NCDC Web repository), his or her data can still be retrieved, transformed and visualized using the DEVise browser, only perhaps less efficiently since filtering at the source site is not guaranteed in this scenario.

Class of Queries to be Supported: Designing the data transformation language is an important task. Based on feedback from our users, it seems clear that basic SQL-style queries supporting selection, projection of attributes, relational joins, grouping and aggregate operations is required. Further, a number of queries over time-series such as computing moving averages arise frequently in the scientific applications that we have encountered. A third class of queries that seems useful to support in this context is the class of *conjunctive constraint queries on k-dimensional datasets* described in Section 1.4.3. These are queries specified as a conjunction of linear or nonlinear polynomial constraints over one or more of the k attributes, and intuitively, define regions in k-dimensional space.

We are currently in the process of designing a language with support for the kinds of queries described above.

1.6 Related Work

In comparison to other visualization tools DEVise is unique in its combination of database style query capabilities and flexible visualization support. Unlike most visualization tools, it is designed to deal with large datasets, and buffering and caching (from disk, from tape, and from remote sites) are integral aspects of the DEVise architecture. By mapping each data record to a visual element, and associating the record id of the data record with the corresponding visual element, DEVise has the ability to go from the visual presentation all the way back to the original data record. In conjunction with the ability to associate summary data with the original data records, this gives DEVise (with the extensions that we propose) unique capabilities with respect to the levels of detail at which a user can explore a dataset: a user can switch interactively from a metadata representation (e.g., a cluster) to the visual representation of the records in the cluster, to the actual records themselves, and back again to the cluster.

To our knowledge there is no other tool or research that aims to support the paradigm of integrated querying and browsing of data together with a wide range of metadata, independently of how the metadata is generated. It is worth noting however that tools such as IDL have recently integrated a number of packages that allow data to be processed. This indicates that the importance of using metadata in the visual exploration process.

Specifically, the following visualization tools are similar in certain respects to DEVise: Vis5D [HPS94], Collage [COL96], LimkViews [JBO94], and VISTAS [VIS95]. We regard some of these tools, e.g. Vis5D, as being complementary to DEVise in that they both provide some valuable and distinct capabilities.

2 Implementation Plan and Milestones

We plan to perform the proposed effort in three years and to progress in three parallel activities that refelect the three research goals stated in section 1.5.

2.1 Year 1

Visual Queries:

1. Add image symbols to DEVise. Here we plan to build on the Quality Controlled Lossy Image Compression system that we have developed [LR96].
2. Enhance the capabilities of DEVise to derive statistical information on graphical time-series and visualize it. Here we plan on taking advantage of code from SEQ [SLR96].
3. Separate visual query definitions from details of the underlying data sets in DEVise.
4. Enhance the Visual Query Processor to support inter-query optimization.

Data/Metadata Browsing:

1. Extend the visualization framework of DEVise to support links between metadata and data.

Data Transformation:

1. Build 'drivers' for HDF sources and improve the way network connections are managed by DEVise.
2. Extend and strengthen the relational query optimizer and run-time query evaluation code.
3. Develop a buffer manager interface that allows the visual engine and the data transformation engine to share the buffer pool.

2.2 Year 2

Visual Queries:

1. Improve color management in DEVise to ensure that a consistent palette of colors is used when a session is stopped and re-started later on the same or another display.
2. Incorporate efficient indexing of graphical data in the visual engine, using our extended R^*-Tree package [GRSY96].

Data/Metadata Browsing:

1. Understand the interaction of BIRCH and DEVise, in particular, the nature of the intermediate datasets produced as a result of BIRCH and DEVise operations in the course of a long session in which both kinds of operations are interleaved.
2. Develop an interface between BIRCH and DEVise that allows a user to guide the interaction between them by using a point-and-click interface.

Data Transformation:

1. Design the data transformation/query language, to support a combination of efficiently implementable relational, sequence and constraint queries.
2. Extend the buffering and caching modules to deal cleanly with remote data and data from sequential media such as tape drives and files from the Web. Here we plan to utilize code we developed to support relational joins for data on tapes.
3. Develop the run-time query evaluation code in the prototype to support the extended transformation language.

2.3 Year 3

Visual Queries:

1. Add 3-D symbols to DEVise.
2. Optimize evaluation of visual operations by improving the caching techniques for graphical data.

Data/Metadata Browsing:

1. Design and develop interfaces between DEVise and other tools (e.g. SEQ, QCLICS) that can generate metadata.

Data Transformation:

1. Develop the prototype source wrapper package further to make it more robust and integrate efficient indexing based upon the extended R^*-Tree package that we have developed.
2. Extend the optimizer to do distributed query optimization.
3. Develop a GUI front-end for the source wrapper package to allow users to conveniently make their data accessible to DEVise.
4. Develop interfaces for data source wrappers, following industry standards such as OLE/DB as closely as possible.

3 Supporting Facilities

The Computer Sciences department currently has 110 Sparcs and 50 i86pc running Solaris, 20 DecStations running Ultrix4.3, 42 HP workstations (Snakes) running HPUX8, and 14 Alphas running OSF1 In addition to these workstations the Computing Facility of the Computer Sciences department includes a 64-node CM5. Presently, every faculty, staff member, and graduate student, has a workstation on his/her desk.

All the computing resources, research and instructional, are managed and maintained by the Computer System Lab (CSL) of the department.

These facilities are more than adequate to support the compute cycles needed for the proposed work; we are therefore only asking for three workstations for software developement and testing for the research assistants to be hired as part of this project.

4 Biographical Sketches

Miron Livny

University of Wisconsin Deptt. of Computer Sciences

Madison, WI 53706, (608)262-0856, miron@cs.wisc.edu

EDUCATION Ph.D., Computer Science, Weizmann Institute of Science, February, 1984.

RESEARCH AND PROFESSIONAL EXPERIENCE University of Wisconsin-Madison,CS Dept., Professor, September, 1995 - present.

University of Wisconsin-Madison, CS Dept., Associate Professor, September, 1989 - 1995.

University of Wisconsin-Madison, CS Dept., Assistant Professor, April, 1984 - August 1989. *Program Committee Member:* PODS, SIGMOD, VLDB, PDIS, SIGMETRICS, IDCS, HCW, ARTDB.

SOFTWARE DEVELOPED The DeNet simulation package, the CONDOR cluster management system, the DBC batch system, the QCLICS image compression and browsing system, the DEVise visualization environment, the BIRCH clustering tool, and the ZOO experiment management system. All are distributed freely, and some have been used extensively worldwide.

SELECTED RECENT PUBLICATIONS L. Haas, M. Carey, M. Livny, and A. Shukla, "SEEKing the Truth About Ad Hoc Join Costs", to appear in *VLDB Journal.*

J. Pruyne, and M. Livny, "Interfacing Condor and PVM to harness the cycles of workstation clusters," *Journal on Future Generations of Computer Systems, Vol. 12, No. 1,* May 1996

E. Haber,Y. Ioannidis, and M. Livny, "Foundations of Visual Metaphors for Schema Display," *Journal of Intelligent Information Systems (special issue on management of visual information), Vol. 3, No. 3/4*, July 1994.

C. Chen, K. Salem, and M. Livny, "The DBC: Processing Scientific Data Over the Internet," *16th International Conference on Distributed Computing Systems*

K. Brown, M. Carey, and M. Livny, "Goal-Oriented Buffer Management Revisited." *Proc. 1996 ACM SIGMOD International Conference on Management of Data.*

J. Myllymaki and M. Livny, "Disk-Tape Joins: Synchronizing Disk and Tape Access" *The 1995 ACM SIGMETRICS - PERFORMANCE Conference*

V. Ratnakar, M.Livny, J. M. Norman,and K. Kucharik, "Classification on compressed images with bounded loss," *The 1995 International Geoscience and Remote Sensing Symposium*

M. Livny, B. Melamed, and A. K. Tsiolis, "The Impact of Autocorrelation on Queueing Systems", *Management Science, Vol, No. 3*, March 1993

V. Ratnakar, M.Livny, "Extending RD-OPT with Global Thresholding for JPEG Optimization" *The 1996 IEEE Data Compression Conferenceq*

V. Anjur, Y. Ioannidis,and M.Livny, "Frog and Turtle: Visual Bridges Between Files and Object-Oriented Data", *8th International Conference on Scientific and Statistical Database Management.*

John M. Norman
University of Wisconsin - Madison
Department of Soil Science
Madison, WI 53706
(608)262-4576
norman@calshp.cals.wisc.edu

EDUCATION Ph.D., Soil Science, University of Wisconsin-Madison, 1971.

Dissertation Topic: "Sunfleck-Size and Light-Intensity Distributions in Plant Canopies"
Dissertation Advisor: "Champ B. Tanner"

M.S., Soil Science, University of Minnesota, St. Paul, MN, 1967.

B.S., Physics, University of Minnesota, Mpls., MN, 1964.

RESEARCH AND PROFESSIONAL EXPERIENCE University of Wisconsin-Madison, Department of Soil Science, Professor, September, 1988 - present.

University of Nebraska-Lincoln, Department of Agronomy, Professor, August, 1978 - August 1988. Pennsylvania State University, Department of Meteorology, Associate Professor September 1975-1978, Assistant Professor September 1972-1975. University of Aberdeen, Aberdeen, Scotland, Botany Department, Research Fellow, March 1971 - May 1972.

SELECTED RECENT PUBLICATIONS "Instrumentation For Studying Vegetation Canopies For Remote Sensing In Optical And Thermal Infrared Regions", Harwood Academic Publishers, London, 360 pp. November 1990. (Edited with N.S. Goel).

"Biospheric Models, Measurements And Remote Sensing Of Vegetation", *ISPRS J. Photogrammetry and Remote Sensing, Vol 47, pp. 163-188*, 1992, (with N.S. Goel).

"Scaling Processes Between Leaf And Canopy Levels", *Scaling Physiological Processes: Leaf To Globe, eds J.R.Ehleringer and C.B.Field*, Academic Press, Inc. New York. pp. 41-76. 1993,

"Soil Surface CO2 Fluxes And The Carbon Budget Of A Grassland", *J. Geophys. Res. Vol 97, pp. 18,845-18,853*, (with R. Garcia and S.B. Verma).

"Algorithms for Extracting Information From Remote Thermal-IR Observations of the Earth's Surface", *Remote Sens. of Environment,* Vol. 51, pp. 157-168, 1995, (with M. Divakarla, and N.S. Goel).

"Terminology in Thermal Infrared Remote Sensing of Natural Surfaces" *Remote Sensing Reviews, Vol. 12, pp. 159-173*, 1995 (with F. Becker).

"Leaf Bidirectional Reflectance And Transmittance In Corn And Soybean", *Remote Sensing of Environment, Vol 29, pp. 161-174.*, 1989, (with E.A. Walter-Shea and B.L.Blad).

"Contrasts Among Bidirectional Reflectance Of Leaves, Canopies And Soils", *IEEE Trans. On Geoscience And Remote Sensing, Vol GE-23, pp.659-667*, September 1985, (with J.M. Welles and E.A. Walter).

"From Artificial Life To Real Life: Computer Simulation of Plant Growth", *International J. General Systems, Vol 18, pp. 291-319*, 1991, (with N.S. Goel and L.B. Knox).

Raghu Ramakrishnan
University of Wisconsin, Dept. of Computer Sciences
Madison, WI 53706, (608)262-9759, raghu@cs.wisc.edu

EDUCATION Ph.D., Computer Science, University of Texas at Austin, May 1987.

RESEARCH AND PROFESSIONAL EXPERIENCE University of Wisconsin-Madison, CS Dept., Associate Professor, September, 1992 - present.

University of Wisconsin-Madison, CS Dept., Assistant Professor, July, 1987 - August 1992.

Editorial Board: Journal of Logic Programming (JLP), Journal of Constraints, Journal of Intelligent Information Systems (JIIS).

Program Committee Member: Databases (PODS, SIGMOD, VLDB, ICDE, PDIS, CIKM, DOOD, EDBT) and logic programming (ICLP, ILPS, NACLP).

AWARDS NSF Presidential Young Investigator, 1990 - 1995, Packard Fellowship in Science and Engineering, 1989 - 1994, IBM Faculty Development Award, 1988 - 1990.

SOFTWARE DEVELOPED The CORAL deductive system. Source code in C++ for CORAL is available freely, and is now installed at several universities and research labs.

The Minibase DBMS. Currently in beta-tests with Ramakrishnan's database text at several schools. Includes a relational DBMS plus several visualization tools.

The DEVise visualization tool and the BIRCH clustering tool. These are available in beta versions, and were developed jointly with Prof. Livny.

SELECTED RECENT PUBLICATIONS M. Cheng, M. Livny, and R. Ramakrishnan, "Visual Analysis of Stream Data", *Proc. SPIE Conf. on Visual Data Exploration and Analysis*, 1995.

R. Ramakrishnan and J.D. Ullman, "A Survey of Deductive Database Systems", *Journal of Logic programming, Vol. 23, No. 2*, 1995.

P. Seshadri, M. Livny and R. Ramakrishnan, "Sequence Data Processing", *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, 1994.

R. Miller, Y. Ioannidis and R. Ramakrishnan, "Translation and Integration of Heterogeneous Schemas: Bridging the Gap between Theory and Practice", *Information Systems, Vol. 19, No. 1*, 1994. *(Also in Proc. EDBT 93, and received the best paper award.)*

J.F. Naughton and R. Ramakrishnan, "How to Forget the Past Without Repeating It", *Journal of the ACM, Vol. 41, No. 6*, 1994.

R. Ramakrishnan, D. Srivastava and S. Sudarshan, "CORAL: A Database Programming Language", *Proc. VLDB*, 1992.

R. Ramakrishnan. "Magic Templates: A Spellbinding Approach to Logic Programs", *Journal of Logic Programming, Vol. 11, No. 3*, Oct. 1991.

I.S. Mumick, S. Finkelstein, H. Pirahesh and R. Ramakrishnan, "Magic is Relevant", *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, 1990.

F. Bancilhon and R. Ramakrishnan, "An Amateur's Introduction to Recursive Query Processing Strategies", *Proc. ACM SIGMOD Intl. Conf. on Management of Data*, 1986. *Reprinted in "Readings in AI and Databases", Eds. M. Brodie and J. Mylopoulos, 1988, pp. 376-430.*

5 Bibliography

References

[AS94] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1994.

[COL96] ftp://ftp.ncsa.uiuc.edu/Visualization/Collage/

[CSL96] C. Chen, K. Salem, M. Livny. The DBC: Processing Scientific Data Over the Internet, In *16th International Conference on Distributed Computing Systems*, May 1996. (http://www.cs.umd.edu/users/jimduff/dbc.html)

[GRSY96] J. Goldstein, R. Ramakrishnan, U. Shaft and J. Yu. Using Constraints to Query R^*-Trees *Technical Report No. 1301 , Comp. Sci. Dept., Univ. of Wisconsin-Madison, 1996.*

[HPS94] W.L. Hibbard, B.E. Paul, D.A. Santek, C.R. Dyer, A.L. Battaiola and M-F. Voidrot-Martinez. Interactive Visualization of Earth and Space Science Computations. *IEEE Computer*, 27(7):65–72, July 1994.

[JBO94] A.S. Jacobson, A.L. Berkin, and M.N. Orton, LinkWinds: Interactive Scientific Data Analysis and Visualization In *Communications of the ACM,Vol.37, No.4* April 1994

[LR96] M. Livny, and V. Ratnakar, Quality-Controlled Compression of Sets of Images In *Proceedings of International Workshop on Multimedia DBMS*, August,1996.

[LLM88] M. Litzkow, M. Livny, and M. Mutka. Condor—A Hunter of Idle Workstations. In *Proc. of the 8th Int. Conf. on Distributed Computing Systems*, San Jose, CA, 1988.

[LRM96] M. Livny, R. Ramakrishnan, and J. Myllymaki. Visual Exploration of Large Data Sets. In *Proceedings of the IS&&T/SPIE Conference on Visual Data Exploration and Analysis,* January, 1996. (http://www.cs.wisc.edu/ jussi/devise/devise.html)

[Min96] http://www.cs.wisc.edu/coral/minibase/minibase.html

[PSC96] The Polar Science Center (PSC) (http://psc.apl.washington.edu/)

[RSSS93] R. Ramakrishnan, D. Srivastava, S. Sudarshan, and P. Seshadri. Implementation of the CORAL deductive database system. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1993.

[SLR96] P. Seshadri, M. Livny and R. Ramakrishnan. Design and Implementation of a Sequence Database System In *Proceedings of the 21th International Conference on Very Large Databases*, September 1996. (http://www.cs.wisc.edu/ praveen/projects/seq.html)

[VIS95] ftp://daac.gsfc.nasa.gov/pub/contrib/VISTAS_1.1/

[ZRL96] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: A Clustering Algorithm for Large Multidimensional Datasets. In *Proceedings of the ACM SIGMOD Conference on Management of Data*, 1996.