Project 301

Cyber-Infrastructure for Climate-Change Research

Sudarshan S. Chawathe

Project 301 is an effort to develop cyberinfrastructure for climate-change research, with the goal of accelerating scientific discoveries through more effective and efficient use of large and diverse datasets. For printing and reference, please use the PDF version of this document: `p301.pdf`.

People

This project is a collaborative effort that includes the Computer Science and Earth Sciences departments, and the Climate Change Institute, at the University of Maine. The current members of the project are:

Joseph Bane
: Computer Science

Sudarshan S. Chawathe
: Computer Science & Climate Change

Andrei Kurbatov
: Climate Change

Paul Mayewski
: Climate Change & Earth Sciences

Mark Royer
: Computer Science

Motivation

The amount of data available to climate-change researchers has grown rapidly in the past few years, and the trend is likely to continue. This growth of data has two components: First, advances in data collection and storage technologies permit us to generate data faster (e.g., from sources such as ice cores). Second, advances in the Web and related technologies permit us to easily access data produced by others. By itself, the first component holds the potential to significantly boost scientific discovery. This potential is amplified many fold by the second component. The amplification is not only quantitative, due to the availability of a larger quantity of data generated by others, but also qualitative, due to the benefits of integrating data from multiple disciplines (e.g., integrating ice-core data with data on the activities and health of prehistoric communities). There are already several examples of the kinds of discoveries enabled by such an integrated view of available data [1].

However, the potential benefits of the above integrated approach are currently difficult to realize. Although researchers are able to integrate disparate data by carefully and painstakingly studying them, they are limited in their ability to do so with the large volumes of data now available. In other words, the conventional method of integrating data, in which all integration decisions are made by a human, simply does not scale up.

Certainly, there are many important aspects of the data integration and comprehension process that we cannot realistically expect to automate in the near future. These are aspects that require deep background knowledge, scientific training, and creativity. There are also, however, very many aspects of this process that may be partly or fully automated with a focused effort on extending current body of work on data integration and data mining. It is these latter aspects that are the target of this work. By automating as much of the process as we can, we hope to enable researchers to better focus their time and energy on aspects that are truly deserving of such attention.

Research Topics

The research plan is based on developing *domain-specific solutions* to general data management problems, with the hypothesis that by carefully limiting the scope of operations to the needs of climate-change research, we can devise and implement effective and efficient tools. We may classify the research topics into four groups.

Data integration The goal of this component is to provide an integrated logical view of datasets relevant to climate-change research, shielding the researcher from low-level details such as the data formats, storage location, database schema, document formats, varying nomenclature, etc.

Data mining Once we have an integrated view of data, we will develop a suite of tools that re-

searchers can use to identify interesting patterns and features in the integrated data.

Provenance When consulting an integrated view of data, it is important to know of the source from which a displayed value or fact is derived. While this task poses few challenges when the integration is simple (e.g., the displayed value being the average of a set of values from different sources), it is much more complex with the interaction uses more sophisticated operations (which are necessary for effective integration), such as extraction of numerical data from text, semantic mapping of terms, and schema transformations.

Workflows If we visualize the lifecycle of data, from the point of origin to the scientific discovery or other product that they enable, they typically go through several steps that include both human and automated processing. While our earlier components are designed to ease and enhance the individual steps, the goal of this component is to develop methods for effectively managing the entire collection of steps (workflow).

Our initial work has focused on the integration and interactive analysis of data, as described next.

Exploratory Data Analysis

Our early work on P301 has resulted in a *data explorer* application, P301DX, that is designed to help scientists study large and often disparate datasets. Currently, the application allows scientists to explore various collections of climate and ice-core datasets. The tool is highly responsive and allows users to resample and display large datasets very quickly. The graphs that are created can then be exported as high-quality vector images in SVG, EPS, and PDF formats, as well as bitmapped images in PNG format.

A screenshot of a typical P301DX session appears as Figure 1. The chart in this session has been built by processing and plotting data from the Siple Dome core, with additional markers corresponding to notable dates extracted from the PDF version of a research paper [2]. When one of these markers is selected, the relevant text from the paper is displayed.

Interface As suggested by Figure 1, the graphical interface is composed of five panels. The largest panel displays the *charts* and other graphical products resulting from data manipulation operations. The two small panels above the chart panel are used by the system to provide *feedback* and to solicit *input*. The two panels on the left are the most interesting ones. The upper one is a *dynamic palette of functional tools*: Each tool is a function that takes one or more objects as arguments and produces one or more objects as output. Input data objects are drawn from the *working objects* panel in the bottom right corner, and the resulting data objects are deposited there as well. Many tools, such as those that expand objects or smooth time-series data, take one or more data objects as their functional arguments. However, tools are also permitted to use other tools as arguments. As a simple example, a composition tool takes two tools, T_1 and T_2 as arguments and produces a new tool, T_3, the composition of T_1 and T_2, that is deposited in the tool palette.

Charts An important feature of P301DX is its ability to produce the kind of detailed, carefully formatted and annotated charts customarily used in climate-change research. To this end, the system provides the expected interactive tools for manipulating chart components such as markers, labels, scales, and colors. Further, all charts can be exported in both vector and raster formats, to enable easy processing by other tools. Figure 2 depicts a simple example of such a chart, the unedited output of P301DX.

Additional examples and screenshots appear in the next section.

Examples and Screenshots

References

[1] P.A. Mayewski, E. Rohling, C. Stager, K. Karlén, K. Maasch, L.D. Meeker, E. Meyerson, F. Gasse, S. van Kreveld, K. Holmgren, J. Lee-Thorp, G. Rosqvist, F. Rack, M. Staubwasser, and R. Schneider. Holocene climate variability. *Quaternary Research*, 62:243–255, 2004.

[2] Daniel H. Sandweiss, Heather McInnis, Richard L. Burger, Asunción Cano, Bernardino Ojeda, Rolando Paredes, María del Carmen Sandweiss, and Michael D. Glascock. Quebrada Jaguay: Early South American maritime adaptations. *Science*, 281:1830–1832, September 1998.

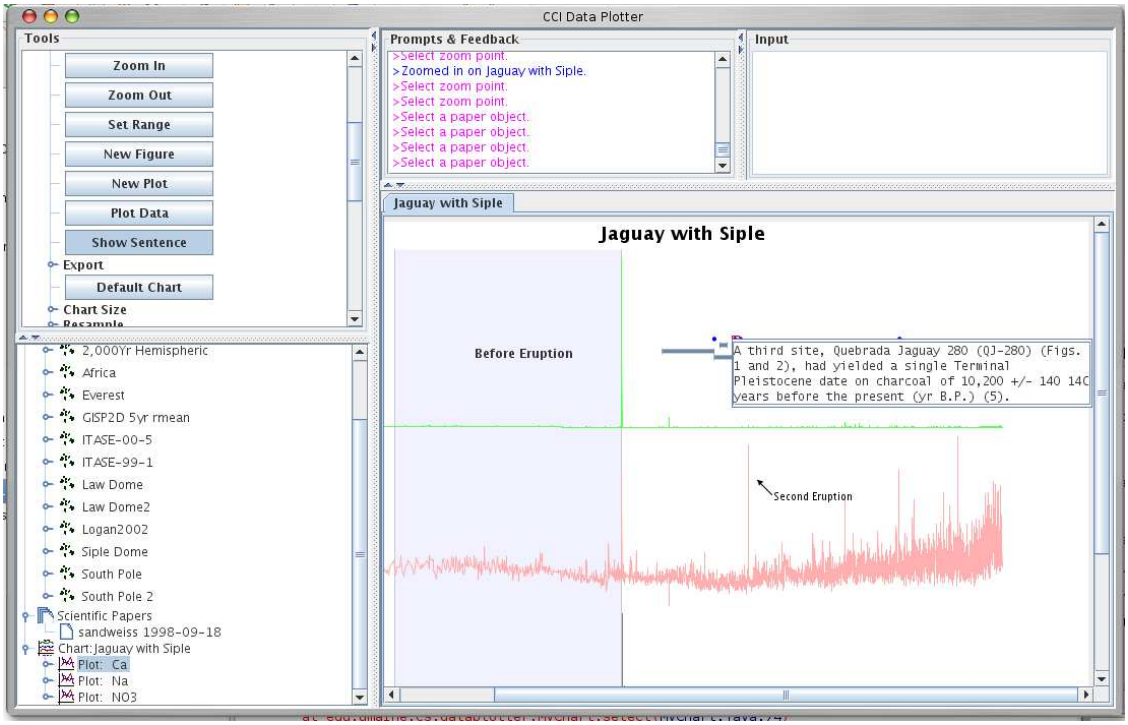

Figure 1: A screenshot of a P301DX session.

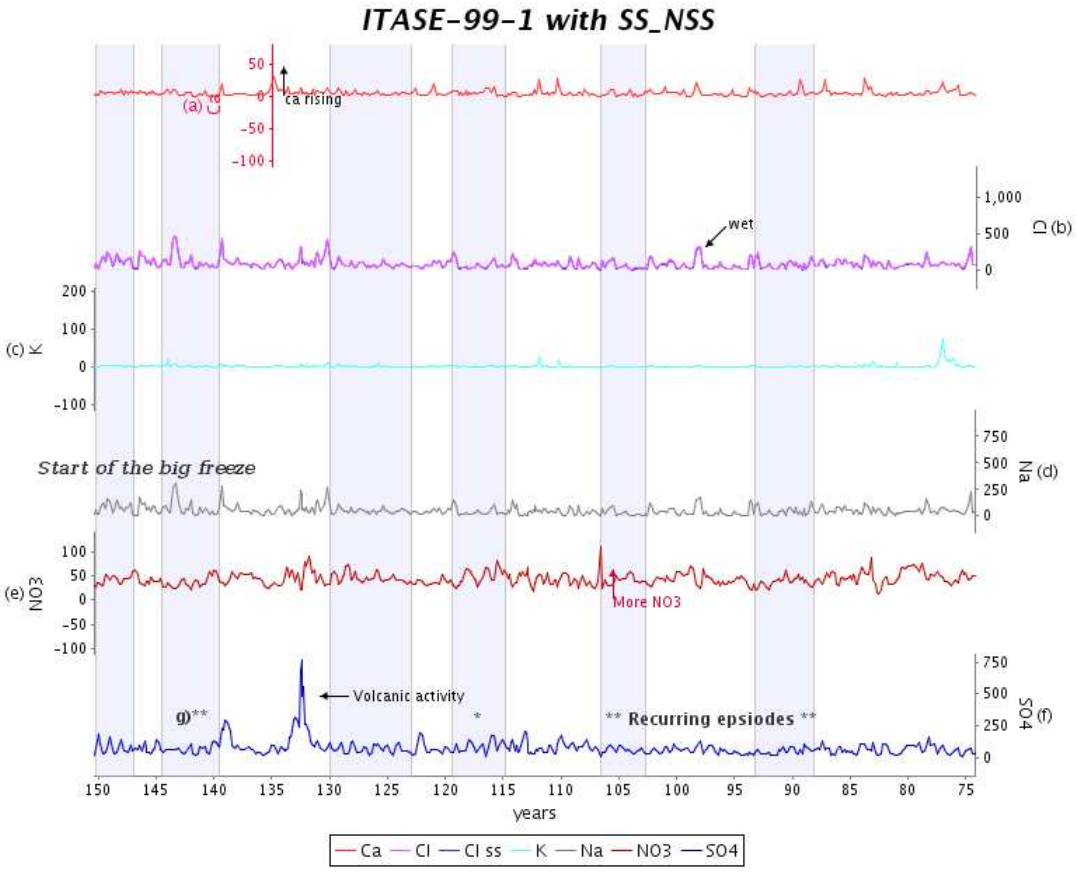

Figure 2: Sample output.

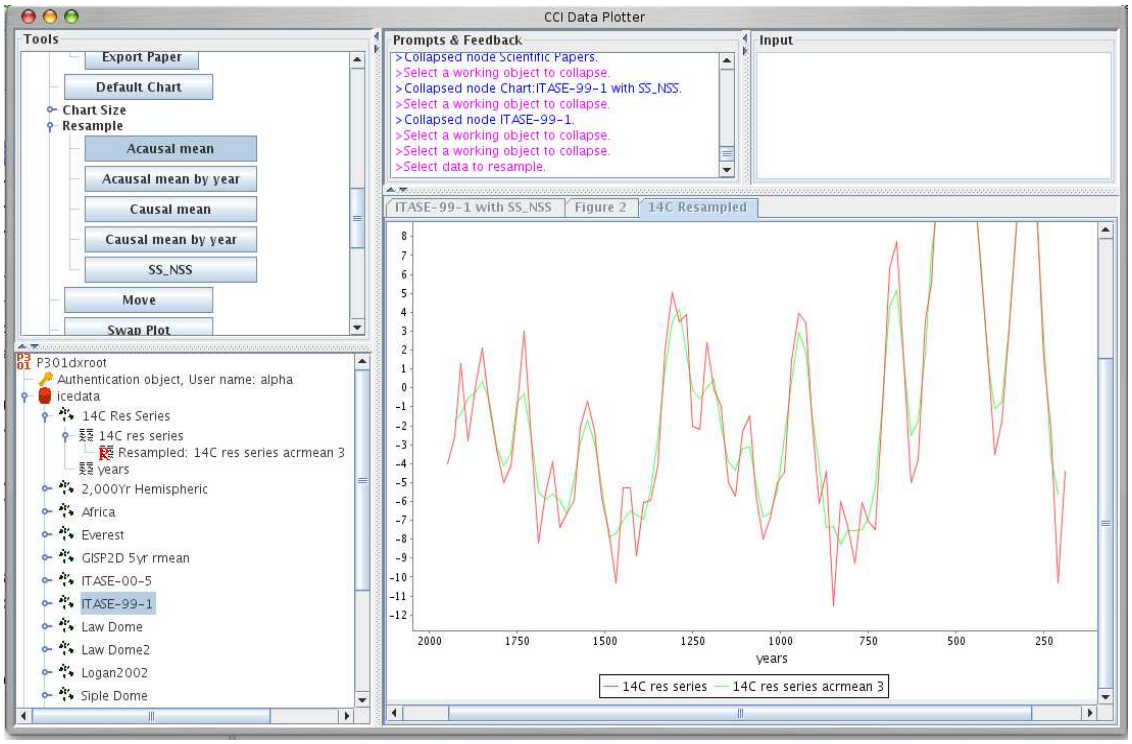

Figure 3: A chart with interactively resampled data.

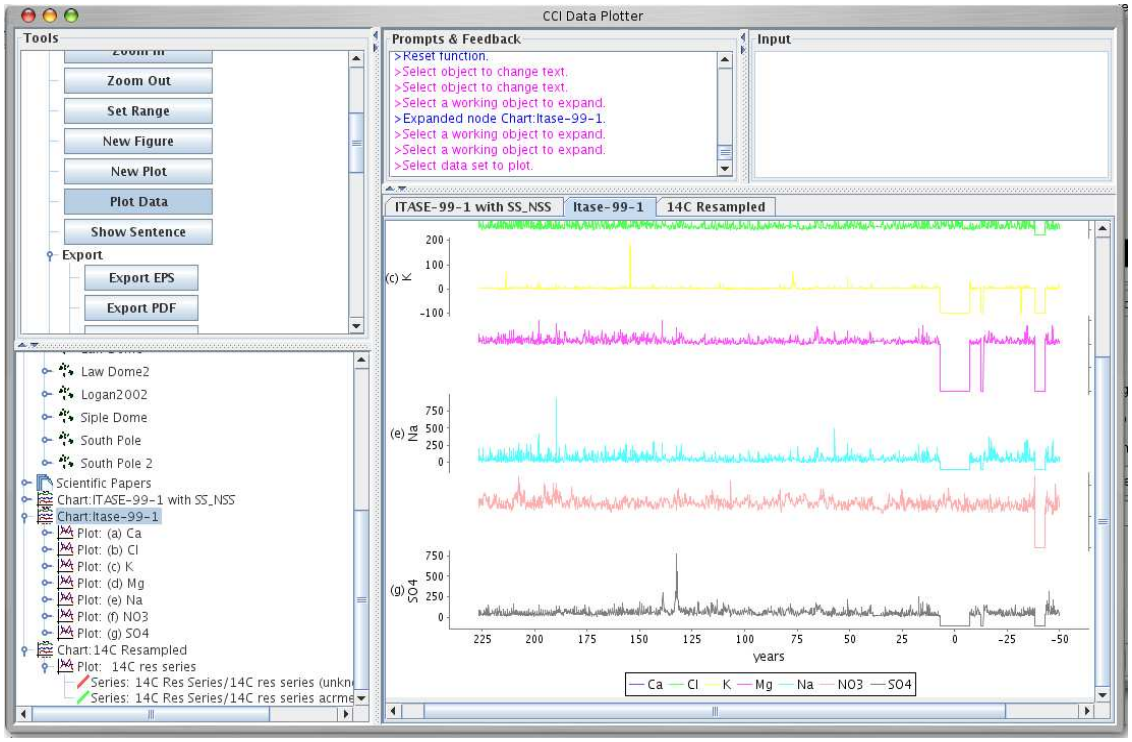

Figure 4: A chart with multiple ice-core data plots.

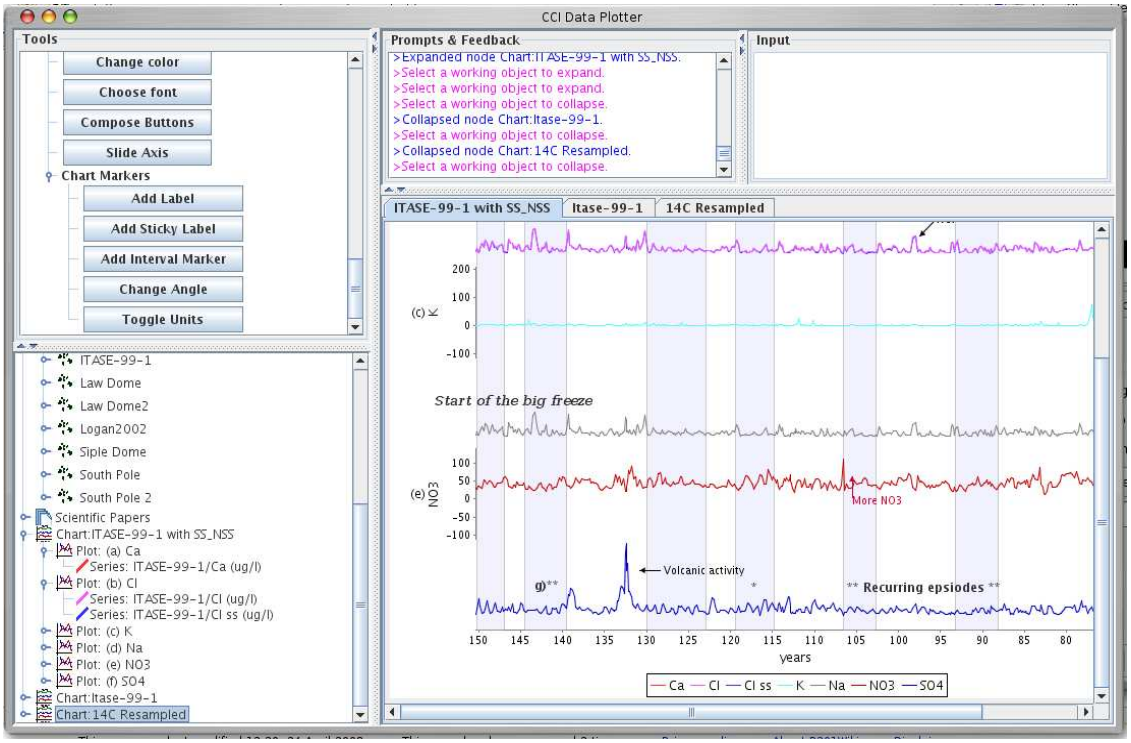

Figure 5: A chart illustrating the use of markers.

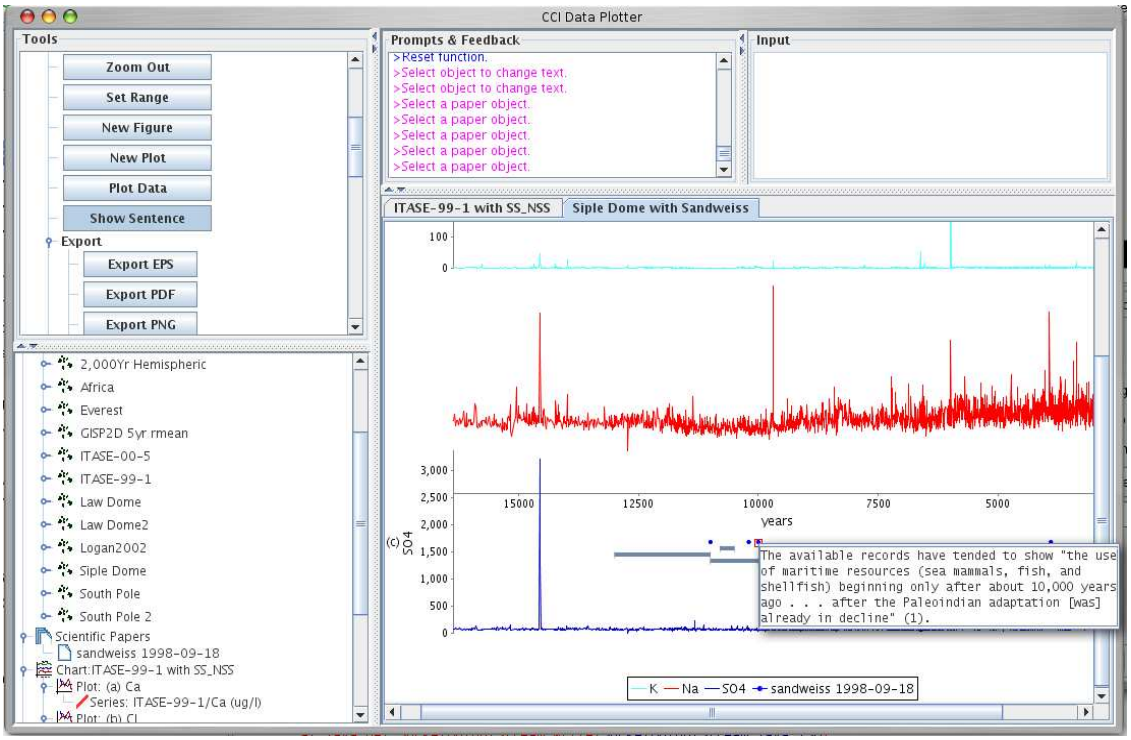

Figure 6: A chart that integrates data extracted from a paper [2] with data from ice cores.