A Characterization of the Scientific Data Analysis Process

Rebecca R. Springmeyer, Meera M. Blattner, and Nelson L. Max

Lawrence Livermore National Laboratory and University of California, Davis
L-301, Livermore, California 94550, {springme, blattner, max2}@llnl.gov

Abstract

Extensible scientific visualization tools are often offered as data analysis tools. While images may be the goal of visualization, Insight is the goal of analysis. Visualization tools often fail to reflect this fact both in functionality and in their user interfaces, which typically focus on graphics and programming concepts rather than on concepts more meaningful to end-user scientists. This paper presents a characterization which shows how data visualization fits into the broader process of scientific data analysis. We conducted an empirical study, observing scientists from several disciplines while they analyzed their own data. Examination of the observations exposed process elements outside conventional image viewing. For example, analysts queried for quantitative information, made a variety of comparisons, applied math, managed data, and kept records. The characterization of scientific data analysis reveals activity beyond that traditionally supported by computer. It offers an understanding which has the potential to be applied to many future designs, and suggests specific recommendations for improving the support of this important aspect of scientific computing.

Introduction

Scientists who study physical phenomena often face complex puzzles. For example, for a physicist studying material interactions with large-scale calculations, the puzzle pieces can include dozens of files, each with several hundred thousand zones involving a dozen or more variables. The disparity between the rapid generation of scientific data and its time-consuming analysis has caused widespread interest in more effective computer support for scientific data analysis [1,11,12,15]. The effectiveness of analysis tools depends not only on adequate functionality, but on a human-computer interface which reflects the goals of the scientific data analysis process. Recent interfaces for data analysis have emphasized scientific visualization, focusing on the rendering and playback of images [1,2,13]. Although this is a natural focus for visualization, it may be less appropriate for analysis.

A survey of tools for the visual analysis of scientific data indicates a wealth of public domain and commercial products which are similar in functionality as well as in their view of analysis as the filter-map-render cycle introduced by Upson [13]. This model provides an excellent representation of the visualization aspects of analysis. The leading tools based on this model have delivered impressive graphics capabilities, portability to many platforms, and advanced visual programming techniques. But do they address all the basic needs of scientific data analysts? This paper will show that the answer to this question is no.

Tool designers have several sources of information on which to base their designs. Some talk extensively with potential users in order to draft functional requirements [1]. Others begin with the functionality of existing tools, and some designers make decisions based on their own intuition and experience. It is usually assumed that these sources of information furnish a sufficiently accurate understanding of data analysis and, therefore, an adequate basis for design. However, this is often not the case; designers may end up with a list of requirements without having a complete understanding of the process to be supported. An incomplete understanding of the process can be compounded by the propagation of inadequacies and misconceptions from previous tools. Clearly, a detailed understanding of the scientific data analysis process should lie at the foundation of analysis tool design.

Scientific data analysis is the process of distilling potentially large amounts of measured or calculated data into a few simple rules or parameters which characterize the phenomenon under study. This may include quantifying known analytical relations between data parameters and inferring previously unknown relations. Images are a by-product of the analysis process rather than its desired result, which is an understanding of the phenomenon under investigation. A more appropriate focus for scientific data analysis interfaces would reflect the broader process which includes what scientists do before, during, and after generating images and other representations of data.

This paper reports on an empirical study designed to characterize the scientific data analysis process. There were two main objectives:

- to expose fundamental elements of scientific data analysis through observation of analysis activity, and
- to consider how technology can be used to support these elements more effectively.
In practice, this process is complex and depends on the physical system under study, the type of data, the analyst, and the available tools. Nonetheless, we assume this process can be characterized by a few fundamental elements common to most types of analysis. In order to expose those elements, we adapted methods used by developers in studies of other processes. Interaction analysis studies have resulted in detailed descriptions and design recommendations for tools to support engineering design [10, 5]. The contextual inquiry technique has been used to increase the effectiveness and usability of tools in other areas [3, 14]. These methods are used to inform designers about what end-users do and how they do it. This information, along with iterative design techniques and participation of users in design, helps developers to create effective tools.

We adapted the techniques of contextual inquiry and interaction analysis to observe scientists analyzing their own data and to characterize the data analysis process. This characterization exposed elements outside the conventional visualization model which defines analysis in terms of image generation. In addition to viewing images, scientists queried for quantitative information, made several kinds of comparisons, applied mathematics, managed data, and kept records. Many of these activities were only indirectly supported by computer. A description of the scientific data analysis process was developed to provide a broad-based foundation of understanding which is rooted in empirical fact, reasonably comprehensive, and applicable to a range of scientific environments. This characterization led to design recommendations for more effective support. It is intended that analysis tool designers supplement their own understanding with these results, adapting the information to their particular situations.

An empirical study of the scientific data analysis process

In this section the method used to study the scientific data analysis process is described. The collection and characterization of the observational information focused on the process, rather than on the scientists, their style of work, or the particular problem and tools. In addition to the visualization tools which are the main concern of the current paper, scientists also used data manipulation tools, laboratory experimental equipment, and statistics packages. Some of these tools were quite well suited to their analysis subtask, and some were lacking in functionality as well as usability. However, it is important to note that the point of the observations was to start at the beginning — with the process of data analysis — and not with tools.

Scientist participation

Interviews and observation sessions were conducted with scientists who represented several different scientific subdisciplines, including physics (laser, semiconductor, nuclear), biochemistry, and numerical analysis. The purpose of interviewing was to become familiar with a range of analysis methods and environments, while observation sessions were designed to record analysis activity as it occurred in practice. Interviews and observation sessions were intermingled over several months, with observation sessions being scheduled as candidates were identified through recommendations and pre-observation interviews.

In searching for scientists to participate in observation sessions, we considered several criteria. The study required subjects who routinely analyzed data in their work as laboratory scientists studying physical phenomena, who agreed to being observed and audiotaped while working in their lab or office, and who together represented several different fields and institutions. Ten scientists participated in the study, two of them in both pre-observation interviews and observation sessions. All were Ph.D. scientists working at either a national laboratory, an industrial research laboratory, or a research university.

Interviews and observation sessions

Two types of interviews were used: initial interviews with computer scientists and experienced data analysts, and pre-observation interviews with scientists on their specific analysis problems. Participants were selected from three institutions with a range of scientific data analysis environments and problem complexity. Initial interviews provided an introduction to analysis methods and environments and were also used to identify potential candidates for further interviews and observation sessions. Pre-observation interviews were conducted with subjects from seven different research groups. Questions involved the general nature of the work, the type of data, and analysis techniques. Some of the interviews included demonstrations of current analysis tools.

While interviews provided ample background information, the bulk of our "data" came from observation sessions. Eight observation sessions were conducted during the normal routine of five scientists who were analyzing their own data. Three of the scientists were observed twice for comparison purposes. Each audiotaped session lasted between 45 and 90 minutes.

No tasks were assigned; rather, whatever analysis problem was under study at the time was the focus of that session. The scientist was asked to proceed as usual, but to explain what was happening as it occurred, and to allow the observer to watch, listen, and ask questions. This was an open-ended dialogue, structured somewhat like an apprentice sitting in on an expert’s work session in order to listen, question, and learn.

The primary emphasis was on observing and recording the actions, but the observer asked for clarification of some actions in order to more fully understand their purpose. Although interaction analysis does not usually include observer participation [10], some communication was required to capture the intent of actions in this
individual activity. The method of asking the scientists to "talk aloud" [6] can provide the context of actions, which would otherwise not be fully understood. Therefore the open-ended interview style of contextual inquiry was combined with interaction analysis to collect the observational data.

Characterization of observations

The technique for characterizing the observations followed the example of interaction analysis, with its goal to "capture as much of the phenomena, and presuppose as little, as possible" [9]. A transcribed record of each audiotape was condensed into a summary which noted the main focus of the analysis and the observable actions, such as viewing a plot, noting the timestep, summing a list of values, or making a comparison between two numbers or images. In a method similar to the "collectibles" approach of interaction analysis, like actions were collected in order to study and describe their common aspects. Both summary and transcript were consulted in identifying actions and listing them in the order of their occurrence. Actions were noted whether supported by computer or not; examples of actions accomplished by hand included calculating the slope of a line and summing lists of values.

<table>
<thead>
<tr>
<th>Observation session #: 1</th>
<th>Action in Chronological Order</th>
<th>Compare</th>
<th>Examine</th>
<th>Number</th>
<th>Navigate</th>
<th>Decide</th>
</tr>
</thead>
<tbody>
<tr>
<td>resize two plots</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>look for curve peaks</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>compare shapes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exact point line-up of peaks:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zoom</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shift</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>show grid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>get # from axis</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shift curve</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>view relative curve behavior</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>decide 2nd calculation best</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>view plot</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zoom</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shift curve, as above</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>exact num. query</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>calculate slope</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>calculate elec. volt. difference</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>compare timing diff (number)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>execute assumption succeed; greater drop than expected</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Navigate through filesystem</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Chronology of actions.

Categories began to emerge as similar actions were contrasted both within and across sessions. For example, comparison figured prominently in the sessions, ranging from comparing two numbers to comparing curve shapes on overlaid plots. Chronological lists, like that of Figure 1, were developed to classify actions in terms of both chronological occurrence (rows) and type of element (columns). Figure 1 illustrates an early list with five headings; the number and description of these elements gradually evolved as more information was collected and our understanding of the process grew.

The method for grouping elements into categories can be illustrated with the example of comparison. As noted above, comparisons occurred between both pictures and numbers; they also occurred between precise numerical values and abstract shapes. Another element, queries, applied to exact values of variables as well as notes and drawings. Both comparisons and queries can be classified in terms of their specificity (precise or general) and how the data is represented (numerical or graphical). However, these classifications were too simplistic to adequately describe the common aspects of comparisons and queries.

A difference between quantitative and qualitative analysis was suggested by an earlier informal study of physicists analyzing simulation data [41]. These terms were adapted to describe the common aspects of the observed types of comparisons and queries. Qualitative was defined to mean an interaction which identifies characteristics of data; quantitative was defined to mean an interaction which determines amounts and proportions of previously identified characteristics. Considering specific instances of queries and comparisons led to drawing a distinction not between the representations, but between the scientists' interactions with representations. For example, a scientist can extract qualitative information from viewing a curve, and that curve might be plotted on an axis from which quantitative information can be extracted. Thus one representation can provide information useful in both types of interactions.

Other elements were reviewed to consider whether this distinction applied to them as well. In all, six elements were categorized as "interactions with representations:"

- Generate: produce and display representations of data;
- Examine: visually inspect a representation;
- Orient: alter a representation or view it differently;
- Query: request information, particularly numbers;
- Compare: examine two or more items for similarity;
- Classify: categorize or identify items;

Each one of these elements can be employed with the goal of obtaining a qualitative or quantitative understanding, or both. For example, objects might be classified by shape or numerical value. Orientation can apply to moving around a 3-d surface, or to identifying the point in time during a simulation which an image represents.

After all of the actions were compared and contrasted with each other, there was a list of fourteen elements.
These were grouped into four categories, including Interactions with Representations. A second category concerned mathematical operations, ranging from arithmetic to elaborate numerical calculations. A third category captured the idea of movement within and among programs. This includes elements which focus attention on data preparation or interaction with a user interface rather than the data to be analyzed. A final category developed more slowly. One of these elements addressed actions which were part of the reasoning and decision process of the investigation. Another addressed the act of recording information during analysis. These elements shared the distinction of being spoken or written ideas and facts involved in the assimilation of information.

Further consideration of these four categories of elements led to drawing a broader distinction between two classes of analysis activity: Investigation, and Integration of Insight. These classes form the two main branches of a decomposition of the scientific data analysis process, which is presented in the following section. The elements and categories of the decomposition were not presupposed, and are firmly based on the empirical evidence of this study rather than intuition or guesswork. Some, like comparison, were obvious from the outset. Others, like the elements to be presented under the heading "Integration of Insight," are not immediately apparent and extend our understanding of the process beyond the obvious. The aim of the decomposition is not a rigid classification of all possible actions, but rather a comprehensive overview of important elements. No matter how complex the data, tools, or methods, it should be possible to describe most instances of scientific data analysis as a collection of these few basic elements.

Results of empirical study

The study resulted in a decomposition of the scientific data analysis process into classes of activity common to all of the observation sessions. The decomposition provides a general frame of reference for:

- comparing different processes,
- categorizing and evaluating existing tools for scientific data analysis, and
- suggesting functional requirements for the design of new tools.

Here we concentrate on the third activity: suggesting functional requirements for design. In this section, we describe the fundamental elements observed, and discuss major implications for the design of tools to support a broader complement of process elements.

The decomposition contains two primary branches of activity: exploring the data to extract information or confirm results (Investigation), and assimilating the resulting knowledge (Integration of Insight). These main branches share four categories of elements as shown in Figure 2. The Investigation branch includes several activities traditionally described as data analysis, including statistical data analysis (part of Applying Math) and visual analysis (part of Interacting with Representations). The second branch of the decomposition highlights an aspect of analysis which is often ignored in the design of tools. If the purpose of analysis is insight rather than images, then designers must ask how tools can support scientists in integrating that insight into their base of knowledge and experience. This important branch of analysis includes decomposition elements which relate to how ideas are expressed, and how information and data are organized.

This portrayal of the process highlights activities observed to be of importance in data analysis. Namely, it:

- appropriately reflects insight as the data analysis goal;
- includes traditional visualization among many other actions and highlights the quantitative as well as qualitative nature of investigation;
- includes a range of mathematical activities, from simple to sophisticated;
- points out that computer-supported analysts must often attend to maneuvering tasks;
- introduces the concept of the expression of ideas, which concerns human qualities of organizing work and decision making.

The elements of each category, listed in Figure 3, are described and illustrated below with examples from the observation sessions. Each should be considered when designing tools for complete scientific data analysis environments.

Interacting with Representations

Generation of representations included 2-d parameter plots, color contours and 3-d surfaces, numeric tables, hand-drawn figures, and digitized microscope images.
Examination modes included viewing a series of plots or numbers, overlaying plots, and searching for peaks or shapes. No single type of representation was viewed as superior to the rest. Three-dimensional views were used to gain a qualitative understanding and to present that understanding to other people. Two-dimensional plots were used to establish precise relationships between parameters and to quantify parameter behavior as a function of time.

Orientation refers to actions which alter a representation, such as flipping axes, shifting one curve above another, and viewing a 3-d surface to gain perspective before viewing related parameter plots over time. Cycling through a series of color tables was a form of orientation used to highlight different features of data.

Queries were used in each session and in a variety of forms, from computer-assisted questions to manual retrieval of information in print-outs, notes, books and other material unavailable on-line. Quantitative queries included using a pick-and-query command to retrieve values by pointing at an image, using a ruler to estimate a value on a screen, and using a command language (get x where y=0). A bi-color split of a 3-d surface was used to query for points greater than some threshold.

Comparison was the primary purpose of several sessions, with calculational results being compared to experimental results. Specific examples of comparisons included those between pairs or groups of timesteps, between pick-and-query values, and relative differences in curves, numbers, patterns, shapes, points, ratios, and estimates.

Classification was the primary purpose of one session. A biologist classified shapes of nuclei which were treated with fluorocarbons and analyzed with an imaging system which measured length, perimeter, and sphericity, for example. Although the imaging was supported by computer, classification was accomplished manually, based on the biologist's experience and understanding of "shape."

Conclusions: The decomposition distinguishes between examination and orientation to stress the importance of interaction. Visualization tools offer increasingly interactive control and more powerful visual effects, such as cutting away layers of 3-d objects. Designers should ask how they can provide analysts with even further control over their data explorations, particularly for quantitative information. Effective designs do not relegate the analyst to a passive audience presented with inflexible representations. As designers speak of steering simulations with visual feedback, this question of control and orientation should be considered carefully. The examples demonstrate the need for multiple representation types, multiple windows, and multiple

<table>
<thead>
<tr>
<th>Interacting with Representations</th>
<th>Applying Mathematics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generate</td>
<td>Calculate</td>
</tr>
<tr>
<td>Examine</td>
<td>Estimate</td>
</tr>
<tr>
<td>Orient</td>
<td>Transform</td>
</tr>
<tr>
<td>Query</td>
<td>Derive new conditions</td>
</tr>
<tr>
<td>Compare</td>
<td>Generate statistics</td>
</tr>
<tr>
<td>Classify</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Maneuvering</th>
<th>Expressing Ideas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigate</td>
<td>Record</td>
</tr>
<tr>
<td>Data Management</td>
<td>Describe</td>
</tr>
<tr>
<td>Data Culling</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Scientific data analysis elements. Reference mechanisms in the active analysis of scientific data. The classification example demonstrates another important point in the use of the decomposition. Process elements are not identified in order to enable the automation of human work, but to understand what it is that scientific data analysts do. The classification of these nuclei demonstrates that a crucial element in analysis is human experience and intuition. Empowering analysts to exercise that experience, intuition, and knowledge in computer-supported exploration of scientific data is the challenge of designing for investigative data analysis.

Applying Math

Applying Math involves the derivation of mathematical quantities or the generation of new data, such as new boundary conditions. Ranging from rough estimates to complex calculations with mathematical tools, this type of activity was grouped into three elements.

Calculations can take the form of both estimations (e.g. approximate slopes, temperature decreases) and transformations (e.g. integration, differentiation, Fourier transform). An example of an advanced estimation concerns the change in volume of a complex shape over time. The analyst approximated the shape by a sphere and estimated its changing volume by calculating an effective radius. This allowed a quick understanding of the relative change in volume without the difficulty of a complex volume calculation of unnecessarily high accuracy.

Deriving new conditions was demonstrated in a session where an offset from a distorted curve was required as the initial condition of a new calculation. A plot was generated and an interactive pick-and-query capability was used to click on several points offset from the distorted curve. The resulting x-y coordinates were used as input to the simulation code. Without the pick-and-query feature, it
would have been necessary to use a pencil, ruler, and calculator to estimate the offset from the curve.

The generation of statistics was a major part of one session with a biologist who calculated statistics such as mean and standard deviation and performed an analysis of variance. Another example involved counting electrons in certain states during the electron recombination process which occurs as a plasma cools. With these statistics, one can compare spectrometer data with modeling calculations to judge whether the process is correctly modeled.

Conclusions: The observations of this study support the conclusion that estimates and transformations are frequently employed elements of analysis. The derivation of new conditions and the use of statistics are further evidence of the importance of supporting the application of math. Visualization often begins the analysis process, but scientific data analysis does not end with the examination of images. This category illustrates that what follows often involves some form of mathematics. Math tools often appeared to be separated from advanced display tools. This study suggests that analysis tools should provide for either integrating calculations directly, or at least improving the interface between math tools and other analysis tools.

Maneuvering

Maneuvering refers to tasks involved in organizing the data, and choosing and setting up representations. It existed as part of analysis long before computers, in the form of scanning tables, flipping through notes and books, and setting up plots and drawings. With the use of computers, maneuvering has become substantially more complex. This is an important element of large-scale scientific data analysis, since one would be stranded without the ability to maneuver through data and representations. Yet it can also become cumbersome if too many details must be specified, or if interaction techniques are inconsistent or awkward. Maneuvering is the only category shared by both branches of the decomposition. The following elements are involved not only in extracting information, but also in synthesizing and organizing the information acquired in investigation.

Navigation includes sizing image windows, controlling the viewing of animations, and scrolling through textual representations. Computer tools can provide capabilities unavailable with manual methods, like quickly resizing plots and replaying animations. This category focuses on interactions with user interfaces of systems, in contrast to the previously described orientation, which focuses on interactions with the data as part of exploration. Orientation increases a scientist’s understanding of the data, whereas navigation is concerned with the mechanics of directing a system and can sometimes be accomplished before data is even read into the program.

Data management has been defined as concerning generic data formats and transportation of data among systems and groups of people [11]. In this study, the time cost associated with data management was observed to be an important consideration. The coordination and execution of programs may require learning several user interfaces, formats, and command languages, and it may require considerable data transfer and translation time. When a question could be answered by either a crude method on a supercomputer or a sophisticated method on a workstation, the scientist tended to choose the cruder but faster version.

For the effective analysis of large data sets, scientists must first focus on an area of interest and extract a manageable portion of data. Three of the five observed scientists have dealt with the problems of data culling, although during the sessions the approximate area of interest was already known. For example, one scientist selected three particular files (one timestep per file) to be displayed to get a feel for the evolution over time; he knew from experience which three were appropriate. A key problem for first-time analysis of a large-scale problem is that one does not know where to begin to look. If no clues exist, one must jump from one data file to the next, searching until an area of interest is discovered. Unfortunately, many tools assume that such a section of data has already been pinpointed, and offer no help in this time consuming and often frustrating part of large-scale data analysis.

Conclusions: Maneuvering can be distracting. For example, some tools pop up windows in random spots or on top of each other and require users to move and resize each window before beginning their analysis. Designers should consider that maximum flexibility for data analysis may be different than maximum flexibility in terms of low level control over windows and widgets. Furthermore, several of the sessions included pauses and long sequences of commands which indicated a lack of integration, and mismatches between tools and tasks. This interrupted the analysis process and focused energy on time consuming manual tasks. For example, a tool might provide a visual image, but no access to the underlying raw data. When quantitative information was needed, it would be necessary to view the raw data manually, with an editor, microfiche, or hardcopy printouts.

Animations have been offered as the answer to data selection, and clearly they are the answer when scanning for visual features in image data. However, it seemed clear from our observations that a quantitative approach might be an effective and faster solution in cases where the scientist has a quantitative understanding of the culling operation. For example, "I want to know where pressure exceeds some amount and density equals another amount," is a different class of question than "I need to see where the materials become intertwined." Thus the conditions of some culling operations on large data sets are such that a quantitative computer search could possibly surpass, in
speed and precision, a qualitative, visual search by a person who manually scans images in an animation.

Expressing Ideas

The assimilation of facts and ideas acquired during investigation involves judgements, conjectures, and deductions. Such inferential activity does not easily lend itself to observation. However, some insight may be gained from considering the subsequent expression of ideas, once formulated, through writing (record) and speaking (describe).

One form of recording pertains to contextual information preserved throughout an investigation, such as lists of calculated numbers, statements, drawings of models and conditions, and file-related records. Notes were written in various media, including notebooks, scratch paper, and post-it notes. In one session, notes were written directly on a connected series of hard-copy plots, as a means of recording the course of analysis and points of interest. This tangible record served as a history of the session to be recalled in later analyses. Notes were taken in some form in every session and previous notes were inspected. A second type of recording might be recalled "working notes," temporary notes which are used during analysis to recall information but are often not kept for later retrieval (occasionally to the dismay of the analyst who wants to remember where the calculation or conclusion came from.)

The describe category captures points during integration of insight which were indicated by a conclusion or shift during analysis activity. Examples of descriptions include decisions, such as concluding that a calculation was worth pursuing in more detail based on the initial findings. Another example was the comparison of results from three calculations; the scientist stopped short at the second, having concluded that the first two cases bracketed the behavior under investigation and the third case was not needed.

Conclusion: Expressing Ideas appears to be an important but often overlooked aspect of scientific data analysis. The notes and descriptions generated during analysis were used for organization as well as records for later reference. In no case was the integration and organization of this information directly supported by computer, except in the form of file-naming conventions and print-outs of isolated figures and lists of numbers. Records and descriptions can become forms of communication as well as histories of a session. The observations suggest that supporting this class of activity could make a valuable contribution to analysis tools. An example in this direction is MediaView, a multi-media document tool designed to support scientific visualization tasks [7].

Functional requirements

Characterizing the scientific data analysis process led to considering how the elements could be supported more effectively. Five main implications for the design of tools are stated as design recommendations. In order to support scientific data analysis effectively, designers should:

Facilitate active exploration
1) provide not only visual, qualitative representations of data, but allow interactive, quantitative exploration through, for example, comparing, querying, applying math, and incorporating intermediate results;

Capture the context of analysis
2) assist analysts in maintaining a record of analysis sessions, including not only notes, but graphics, sketches, and perhaps even audio and video records;

3) link materials from different stages of a study, such as experimental records, analysis materials, and documents for publication;

Decrease Cumbersome Maneuvering
4) minimize unnecessary or distracting navigation requirements;

5) provide computer support for culling large data sets.

Not every analysis tool can necessarily encompass all of the elements. However, an effective data analysis environment should provide an integrated set of tools which supports not only visualization, but some of the additional functionality noted here. Prototype systems should be used in an iterative development cycle in order to adapt these ideas to specific application areas. Such a project was undertaken for one of the elements, data culling. A prototype tool was designed and developed to explore a quantitative rather than image-based approach to selecting conditional areas of interest. The design and results of evaluation exercises with end-users are presented in [8].

Some of the problems with tools which were exposed during the observation sessions could be addressed by extending existing tools. However, added functionality alone will not solve the problems of inadequate integration between tools, and mismatches between tools and tasks. Domain specific knowledge plays an important role in improving the functionality of software. A designer can apply knowledge of how domain activities are actually practiced to improve the effectiveness and usability of software tools to support data analysis. The descriptions of process elements presented in this paper are intended to inspire designs which focus on data analysis activity rather than depend on human adaptability to adapt a visualization or programming model to analysis tasks.
Summary and conclusions

An empirical investigation of the scientific data analysis process was conducted by observing scientists who were analyzing their own data. Fundamental elements of the process were identified, described, and categorized. The characterization shows that the scientific visualization focus supports only part of the scientific data analysis process. Functional requirements were specified for designing tools to support a broader complement of elements.

Several important elements of the process were identified which reach beyond visualization of data. While some of these elements may appear to be obvious, they are firmly based on the empirical evidence of this study rather than intuition or guesswork. Moreover, many important elements which were not obvious also resulted from the study. For example, the observations revealed the need for a history of analysis sessions, and the many ways in which quantitative information is obtained.

These results provide a starting point from which future research can proceed. In particular, problems peculiar to the interactive analysis of very large scientific simulations and collaborative analysis are two topics that have yet to be fully explored. Decompositions of other processes can be compared to scientific data analysis, and different types of data analysis can be compared to each other. For example, the analysis of physical phenomena could be contrasted with the analysis of survey data by examining how elements differ in frequency and use. Another direction would be to use the decomposition to describe and evaluate existing tools in terms of the process elements.

Acknowledgements

We would like to acknowledge Sara Bly and Carol Hunter for helpful discussions and critiques of the descriptions of scientific data analysis.

This work was performed under the auspices of the U.S. Department of Energy by LLNL under contract No. W-7405-Eng-48.

References


