A Case Study in Astrophysical Data Visualization

Gitta Domik

CU-CS-622-92 October.1992

University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

A Case Study in Astrophysical Data Visualization

Gitta Domik

CU-CS-622-92 October 1992

University of Colorado at Boulder

Technical Report CU-CS-622-92
Department of Computer Science
Campus Box 430
University of Colorado
Boulder, Colorado 80309

Abstract

This study reports on a collaboration between astrophysicists and computer scientists, in which the visualization needs of the physicists have been researched. A focus on user-centered design while building a visualization system uncovered shortcomings present in most current visualization systems. These shortcomings relate to the integration of visualization tools into the complexity of the existing data analysis environment. The relationship between data, scientific interpretation intent, and visual representation was closely observed in the case of a multi-spectral data set. Observations on preferences and future needs for new visualization tools are reported.

A Case Study in Astrophysical Data Visualization

Gitta O. Domik

October 1992

1. Introduction

This study reports on astrophysical data visualization as performed at the University of Colorado. While it is not an attempt to capture general visualization needs of astrophysicists, it concentrated on the needs of one specific group of scientists. Between Fall of 1989 and summer of 1992, graduate and undergraduate students of the Department of Computer Science worked together with me and the scientists at the Center for Astrophysics and Space Astronomy (CASA) to better understand the physicists' needs for data visualization.

2. The Scientific Environment

CASA hosts about fifteen scientists and about the same number of graduate students. Their scientific interests are on stars, interstellar matter, galaxies, star and planetary system formation and cosmology. The data they base their assumptions on consist mainly of images, spectra and point source catalogs. Much of their data analysis deals with two dimensional, even gridded images (e.g. the IRAS[1] Skyflux images). These data sets are described as a function $z = f(x,y)$, where z denotes the emission at the spatial location (x,y) at a specific wavelength or wavelength range. The observation of changes in the emission depending on the wavelength is a clue to many scientific discoveries.

A typical research scenario at CASA consists of subsequent

• retrieval of data promising for the specific research,

1. InfraRed Astronomy Satellite

• preprocessing to remove noise and correct instrumental effects,
• numerical calculations, often in the form of statistical analysis,
• visual and interactive data processing.

Available software at CASA concentrated on the first three steps (retrieval, preprocessing and quantitative analysis). Interaction and qualitative analysis through visual browsing was to be added with our help. The new tools were collected under the name STAR (Scientific Toolkit for Astrophysical Research).

The computing environment consists of a conglomerate of workstations (SUN Sparcstations, DECstations and VAXstations) of relatively low computational power. Most of the software used for retrieval and preprocessing of data items has been developed by CASA's scientists, students and staff. Public domain software covers most of the numerical and visual analysis modules. Due to the various characteristics of space and ground sensor data from different wavelength ranges (such as radio, infrared, visible or x-ray), different software packages are being used to work with different sensor data. Input and output data streams are not fully standardized, reflecting the lack of standard data formats, and resulting in "islands" of software systems (Nadeau et al., 1991). Each software package has its own strengths and weaknesses in its ability to preprocess, analyze and visualize astronomical data of specific characteristics.

The scientists have varying degrees of computer expertise. Some spend more time on developing computer programs as opposed to analyzing the results, whereas some scientists spend hardly any time at the computer at all. Our interaction was stronger with the first type of scientists, because they were willing to experiment and saw direct gain for their own research in working with us.

3. The integration of visualization tools into the scientists' environment

During the first two years the goals of the study were mainly driven by the scientists themselves. To ensure the understanding of their needs, a user-centered approach to any design/development of user interface and tools was taken. A study performed by Mickus-Miceli (Mickus, 1990) describes the efforts of using cognitive design techniques to solicit feedback from the scientists at every step throughout the development of STAR.

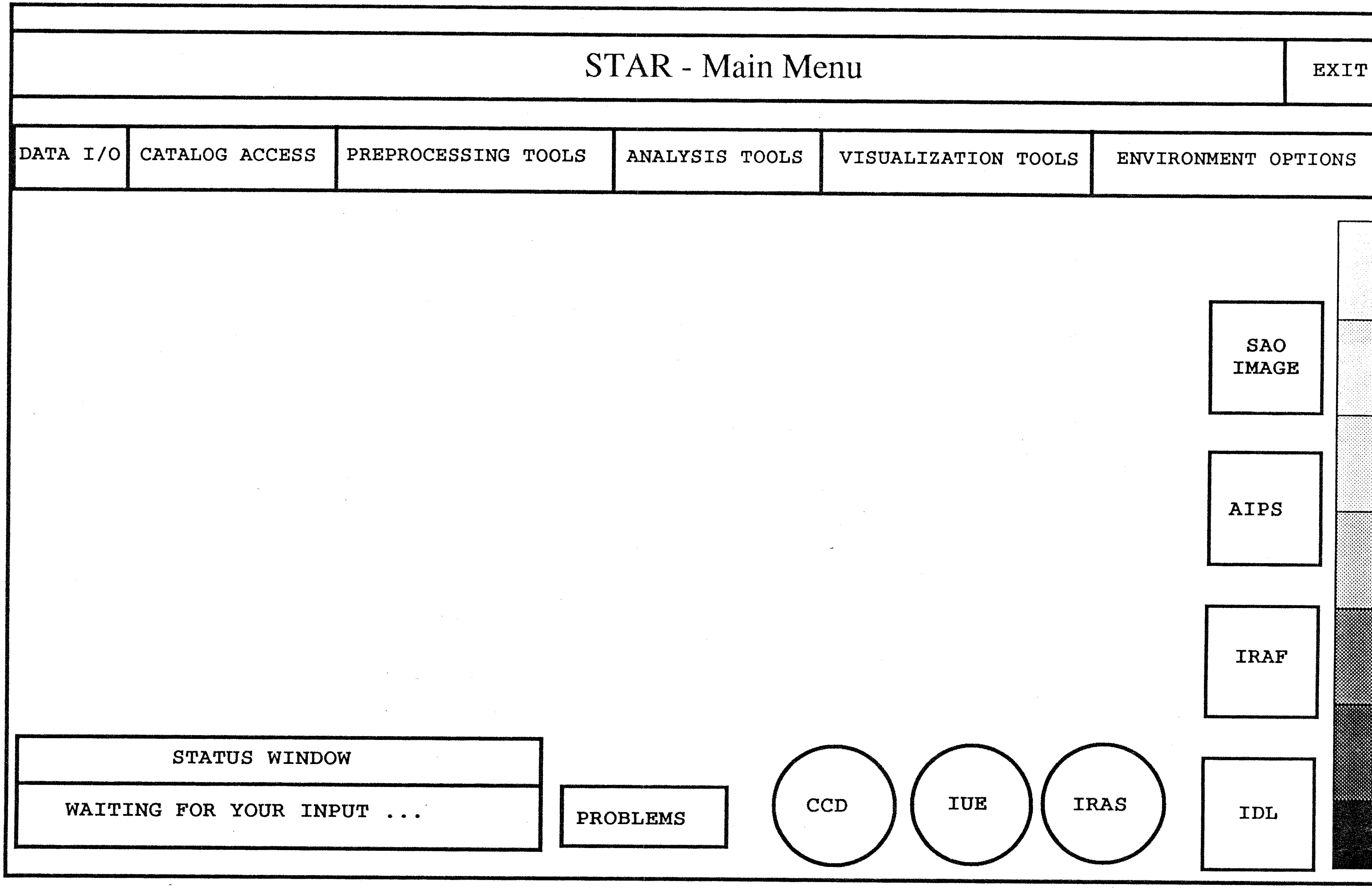


Figure 1: STAR's toplevel user interface

Figure 2: Switch board for the available color tables.

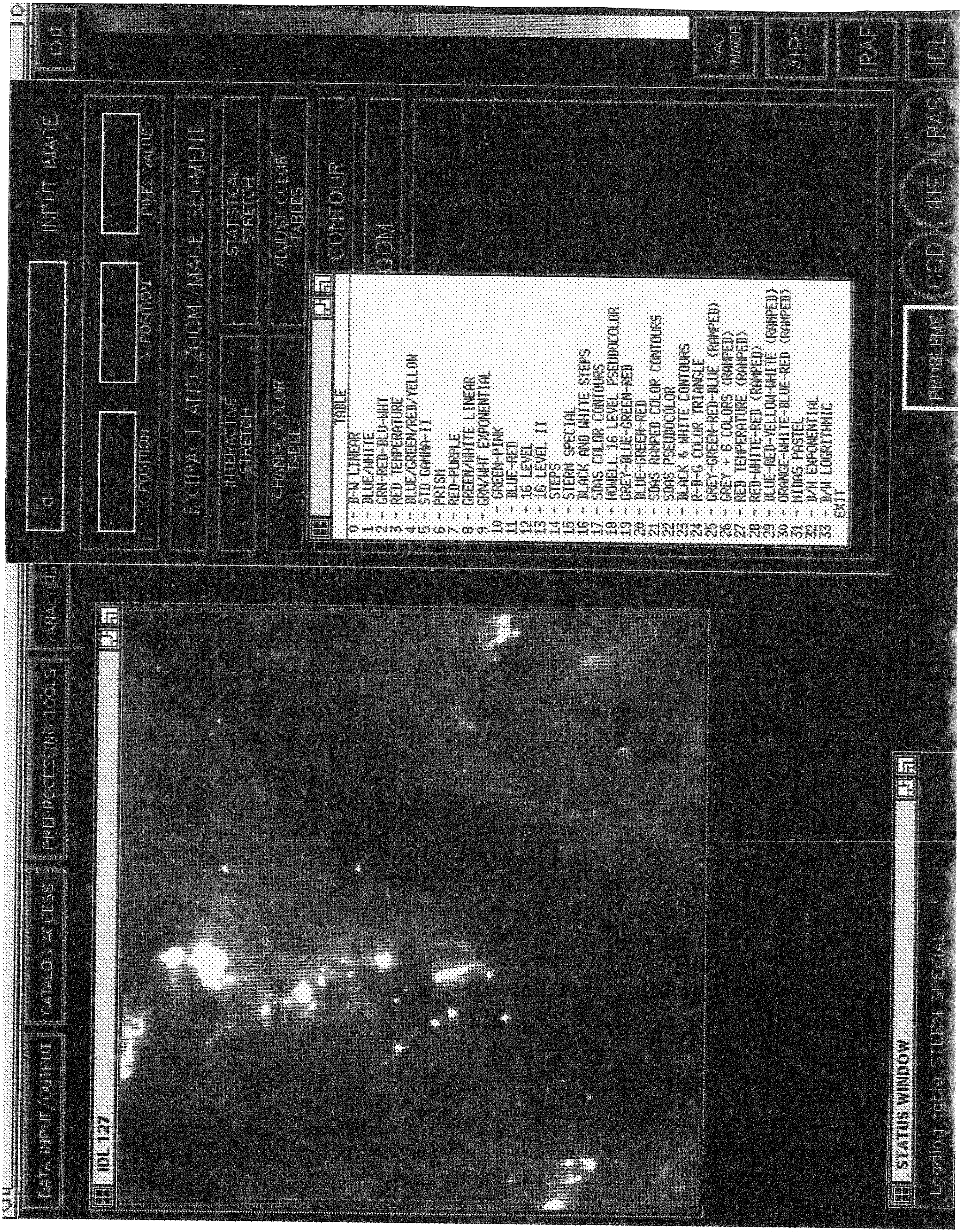
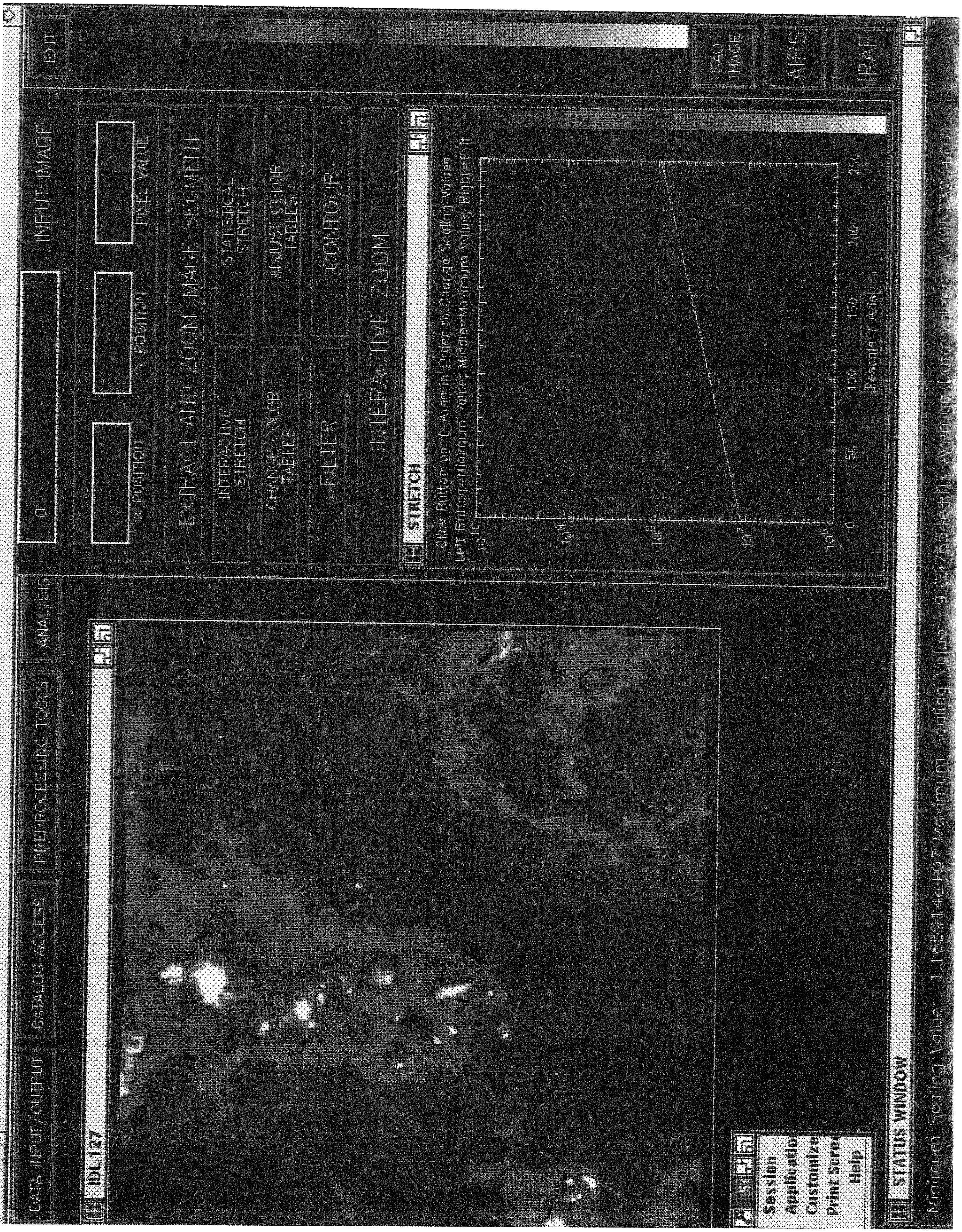


Figure 3: Manual/interactive adjustment of the mapping from data value to display value.



The scientists' first desire was to integrate any newly developed software with the existing software. Additionally any new software should allow direct interaction with the tools (point-and-click). As the existing software was not integrated in any way, we developed a new user interface that would allow to integrate existing software and additionally the new tools. The new user interface is pictured in Figure1. Most of the new software tools appear under the menu "Visualization". Software packages accessible as routines from STAR's development platform (IDL[2]) are also called by clicking on the menus or buttons. "Foreign" (non-IDL based) software packages are activated by clicking at the corresponding square buttons to the right of the user interface. Data format conversions between different formats the various packages favor, is performed by functions under the menu "Data I/O". This integrated approach was to have an influence on data visualization: by being able to combine data sets from various sensors, merging of data sets through visual means becomes feasible.

A more elaborate description of the integrated software design is given by Mickus-Miceli and Domik (1991) and Domik and Mickus-Miceli (1992).

In this first stage, the visualization tools demanded by the scientists were simple, interactive tools, ranging from visual user interfaces to color transformations (see Figure 2 and 3). The attempt to create and visually explore data cubes from multi-spectral data ($z = f(x,y,\lambda)$, with an additional dimension expressing wavelength) failed, because there was no availability of sufficient multi-spectral data and therefore the needs and interest of the scientists was not given at that time.

4. Complex Visualization Demands

Recently Prof. John Bally joined CASA. His research involves scientific interpretation of data cubes as described above; he is familiar with the representation of spatial-spectral data (data containing spatial as well as spectral dimensions) and was interested in the development of tools to interact with his data. In order to present the scientist with expressive and effective visual representations, we studied the nature of his data as well as his scientific interpretation intents.

The role of data visualization is to stimulate mental processes different from quantitative data analysis. Visual data analysis offers an overview of data characteristics through browsing, often leading to an intuitive understanding of data characteristics and their relationships by sacrificing accuracy in interpreting the data values. Because the human visual system emphasizes spatial relationships, up to three data characteristics can be represented in a natural, intuitive way in form of spatial dimensions. Data visualization is an indirect way of interpreting data: instead of

2. Interactive Data Language by Research Systems, Inc.

being interpreted from its natural, usually quantitative characteristics, it is first encoded into a pictorial representation. The encoding process bears the danger of creating artifacts and therefore missing the correct interpretation: e.g. abrupt color changes may mislead by pointing to discontinuities in a data set or subjective assessments of patterns may lead to misinterpretations.

A visual representation of data values should take into account the data characteristics as well as the interpretation intent, as suggested by (Mackinlay, 1986; Wehrend and Lewis, 1990; Robertson, 1990). De Ferrari (1991) adds the influence of other visualization specifications, such as user imposed restrictions, besides the interpretation intent.

In the case of visualizing the astrophysical data cubes, we used a simple mapping model as shown in Figure 4 to map numbers into pictures and map the pictures into a valid scientific interpretation (Domik, 1991).

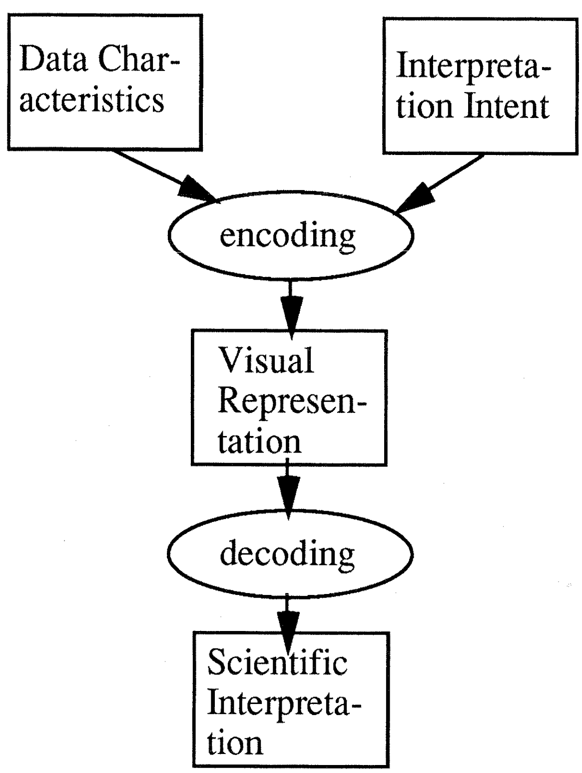


Figure 4: Going from numbers to pictures to a valid scientific interpretation of the numbers.

4.1 Data Characteristics

The data is collected by a 7 m telescope dish owned by ATT Bell Labs. It operates at a frequency between 20 to 40 Mhz, corresponding to a wavelength of 1.3 cm to .7 cm. The collected data is in form of 2-d image tiles for each measured frequency. Processing of the collected raw data values from the heterodyne receiver results in even gridded data values defined in three dimensions (spatial, spatial, frequency). The data values correspond to a count of carbon monoxide molecules at that specific spatial location and frequency. Data values range between -32000 and +32000.

4.2 Interpretation Intent

Carbon monoxide is used to trace molecular clouds. Molecular clouds are the material from which new stars and planets are formed. It is important to understand the changes of the molecular cloud in space as well as in frequency: Is the cloud expanding? Collapsing? In what direction is it moving?

4.3 Encoding numbers into pictures

It is important to express essential data characteristics in the resulting visual representations. In the case of the astrophysical data cubes, such essential characteristics are spatial location as well as frequency and the data value itself. Leaving both spatial dimensions in their natural form and mapping frequency into a third spatial dimension created an even gridded cube with the data values expressed as voxels. This geometric representation will be referred to as “data cube” throughout this paper.

However, the various slices of spatial data values could also collapse into one single slice, where spatial dimensions are represented in their natural form, but various data values along one frequency dimension appear clustered together. This geometric representation will be called “collapsed slices”.

It is also important to represent the data is an effective way, so that the decoding process from pictures to scientific interpretation is quick and accurate. The following visual representations were chosen and discussed with John Bally:

a) Iso surfaces:

Data values of a certain threshold were connected to create iso surfaces. This is a well known rendering technique of the data cube representation. In this representation, the overall shape of the data can be observed as well as isolated volumes (see

Figure 5). Understanding the overall distribution of the carbon monoxide in the given spatial-spectral dimensions is important in order to understand the detailed quantitative information. The iso surface representation can be enhanced by using several transparent iso surfaces, and by adding individual slices through the data cube.

b) Translucent representation:

Rays penetrate the data cube from a chosen point-of-view and accumulate values of opacity assigned to the data values. This representation inflicts a translucent characteristics on the molecular clouds, very much like the visual form of real clouds. It allows to look into the cloud as opposed to observe the surface only. Because the scientist felt a natural understanding of this representation, it was favored as compared to any other representation.

Figure 6 shows a translucent rendering of the cube by looking at the data from the side: one spatial dimension increases to the right, the frequency increases from bottom up. The rapid changes of the data values in the mid-frequencies show special characteristics of carbon monoxide at these frequencies. Figure 7 shows the same data cube using the same representation looking from top down onto the cube.

c) Data slicer:

To monitor the change of one data value in relation to its neighbor values, a data slicer was used. Even though a data slicer can only monitor the neighbors surrounding a certain data value inside a plane, flexibility in placing the slices inside the cube can monitor various changes. Figure 8 shows four slices cutting through the cube parallel to the x/y plane, enhancing the understanding of the movement of the cloud through frequency.

d) Collapsed data slices:

To collapse all (or a subset of) data slices along the frequency dimension into one single two dimensional image, one must be careful to maintain all information of all three dimensions. By using principles of Gestalt theory (Gordon, 1989), all data values along one frequency dimension are mapped into one complex "glyph" or "icon". In order to be seen as belonging to the same Gestalt, the set of data values along one frequency dimension is mapped into one figure that can be distinguished from its neighbors. The difference of one figure from its neighbor relates to the spectral characteristics of carbon monoxide and can be interpreted accordingly.

The resulting image can also be seen as one entity, therefore allowing interpretation of the overall distribution and change of carbon monoxide in the data cube.

Visual representations of collapsed data slices leave it up to the human visual system to decide if the focus is on large-scale or small-scale structures.

Figure 9 shows a representation using color to indicate the various carbon monoxide counts of nine consecutive slices of a subset of the original data values; the spatial location inside each red square is used to indicate the various spectral responses. Figure 10 encodes five slices into five characteristics of a cube: width, height, depth, color and view point.

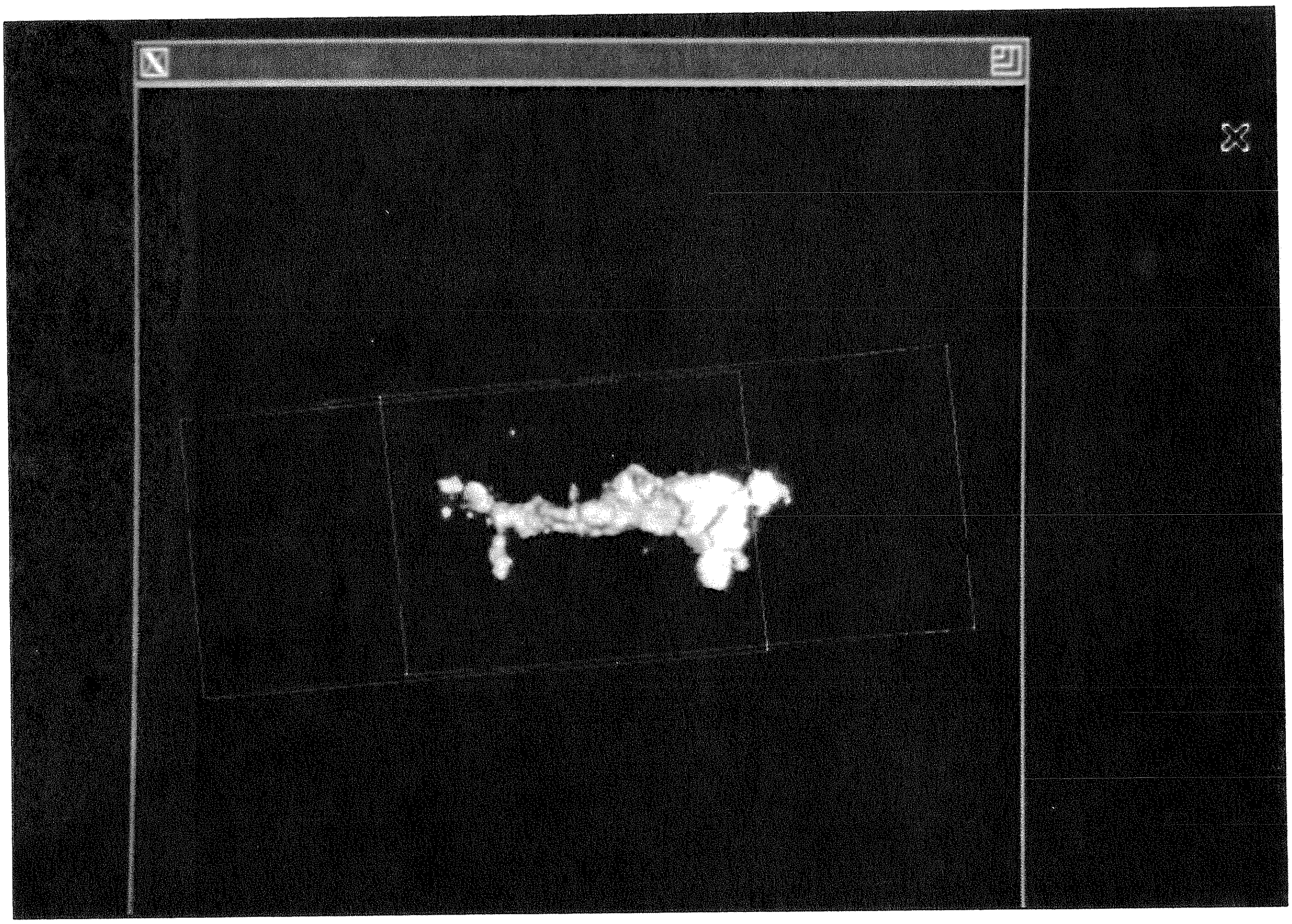
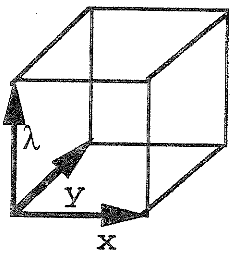


Figure 5: Isosurfaces

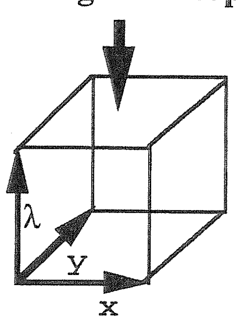
Figure 6: Translucency



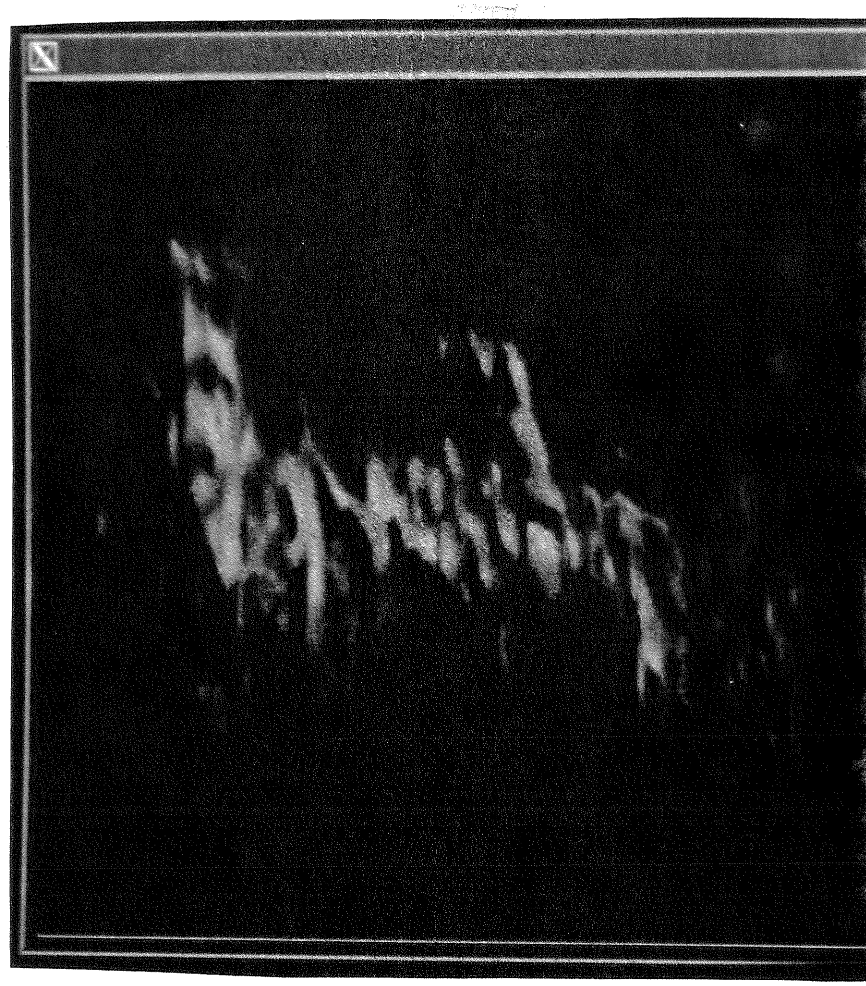
viewing from side

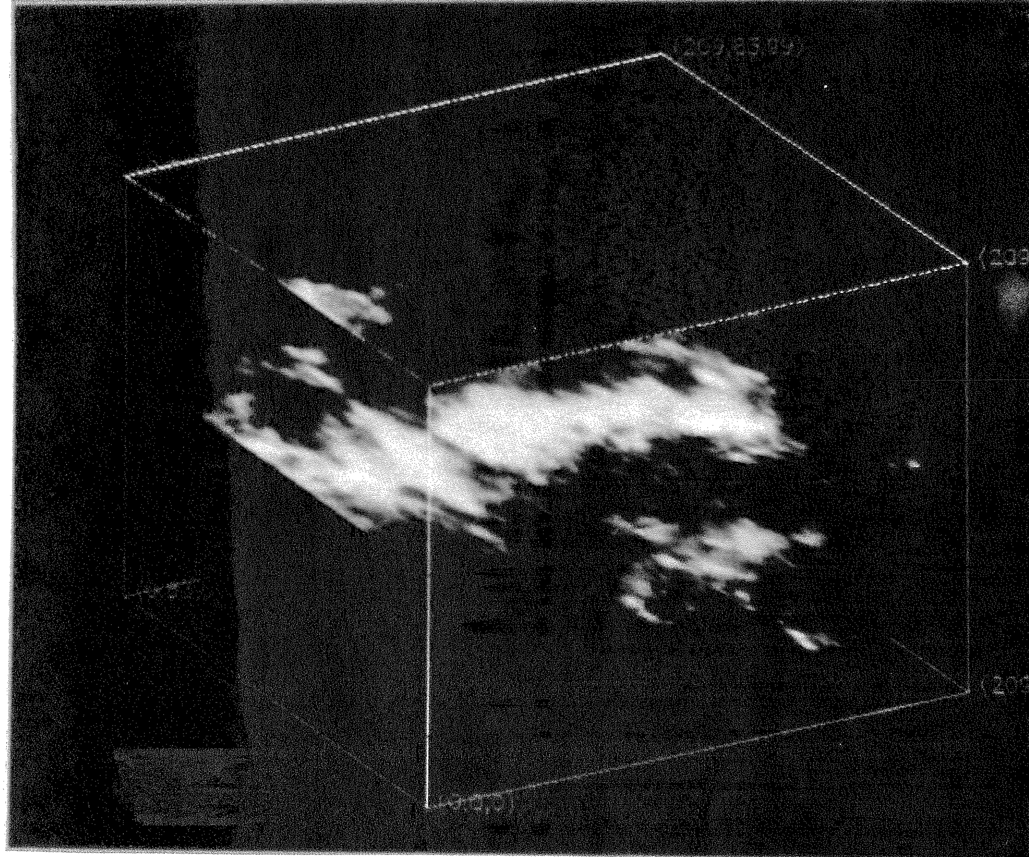
dark blue: low molecule count
cyan: medium molecule count
green: high molecule count
yellow: highest molecule count

Figure 7: Translucency
viewing from top



dark blue: low molecule count
cyan: medium molecule count
green: high molecule count
yellow: highest molecule count





igure 8: <u>Data slicer</u> (viewing
our slices parallel to x/y plane)

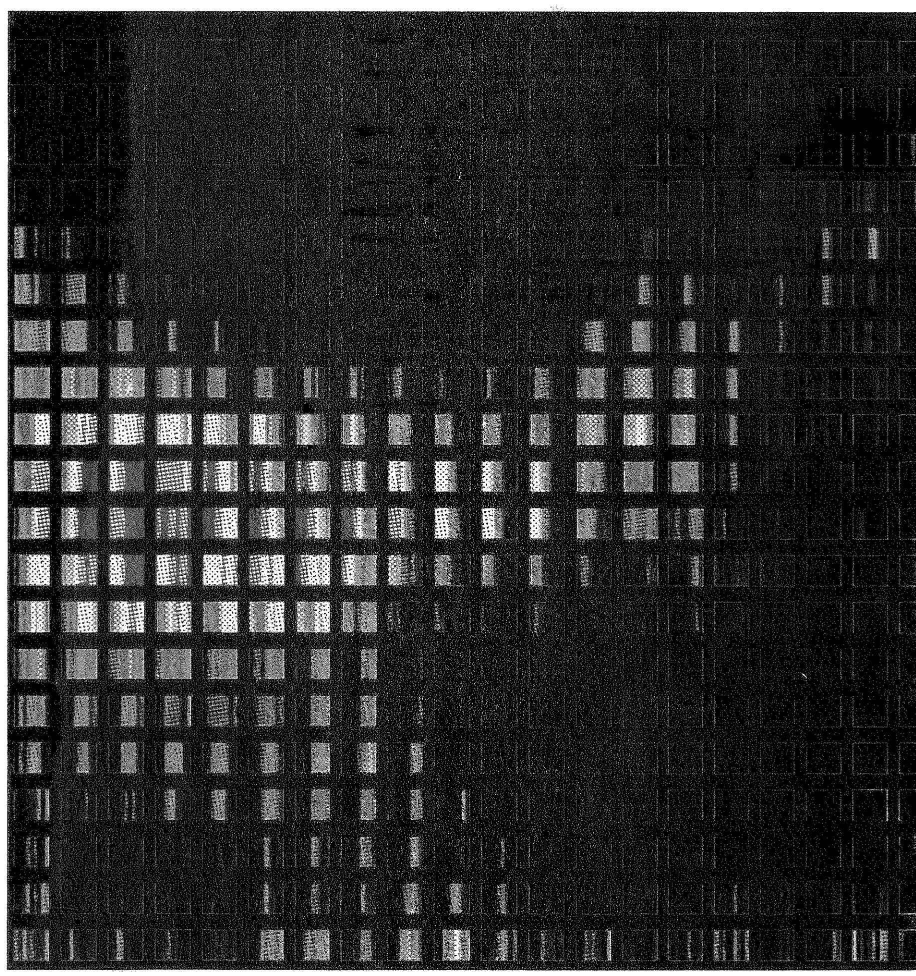


Figure 9: Nine slices of a subset of the data cube (20 x 20 x 9 voxels) are visually correlated. Each red square contains 9 colored slices relating (from right to left) to the spectral responses at that spatial location. Low values are blue and cyan, medium values are green and yellow, and higher values are red and magenta.

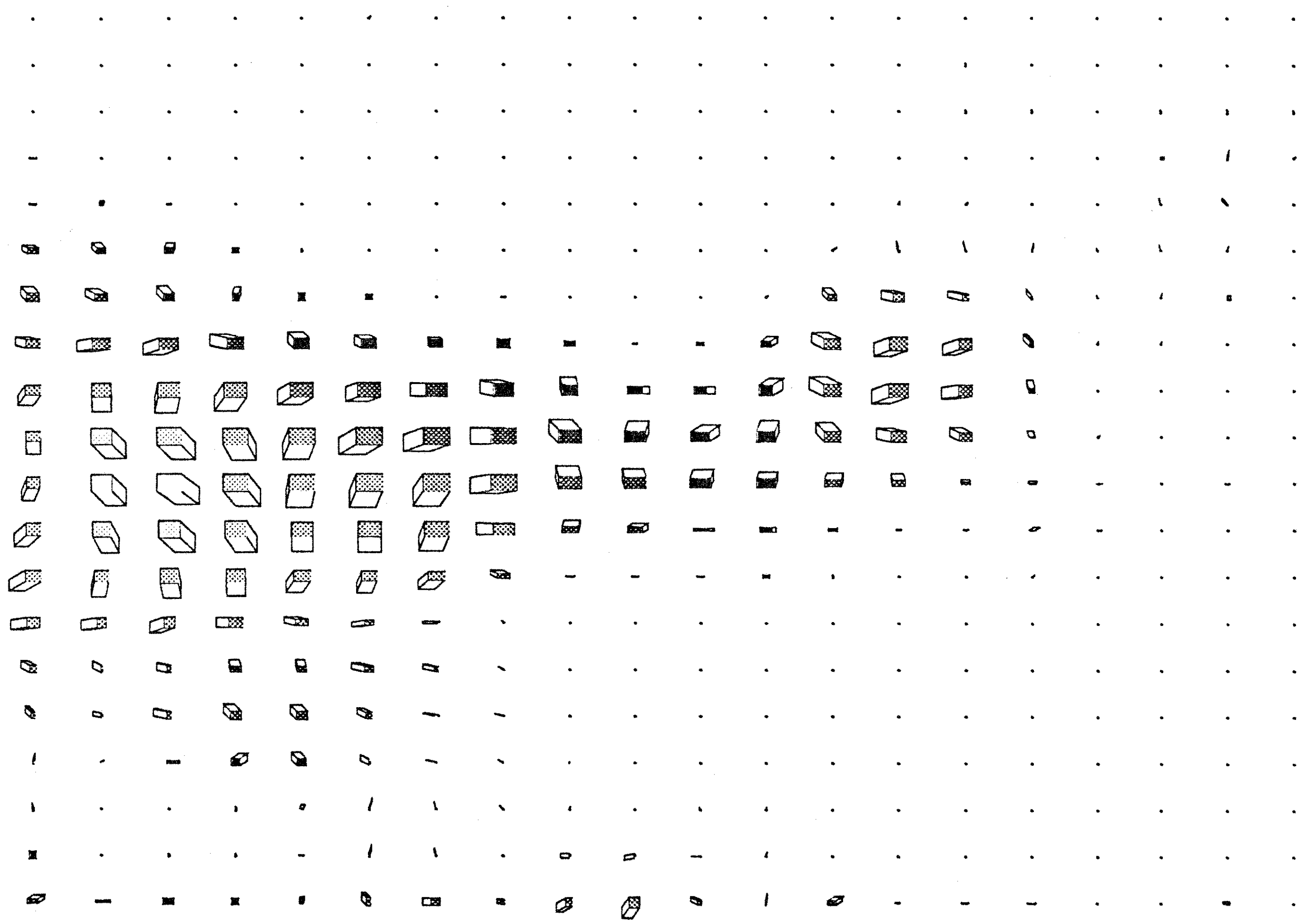


Figure 10: Five slices of a subset of the data cube (20 x 20 x 5voxels) are encoded into five characteristics of a cube: width, height, depth, color and point-of-view. Data of the first three slices was scaled down to a range beween 0 and 20 to indicate width, height and depth. The fourth slice is mapped into a range between 0 and 255 to show color; the last slice contains a range of values between 0 and 360 and indicates the point-of-view. It is interesting to observe an animated version of this representation by visually correlating each five consecutive slices in the data set of 100 slices: growth and rotation relate to changes in the spectral response from slice to slice.

4.4 Reaction of the scientist

The reaction of John Bally to the representations was very positive. However, the representations fell short of his hopes in various ways:

a) The speed of the workstations do not allow real time animations of iso surfaces or translucent representations. This hampers the understanding of the third dimension and greatly limits interpretation. However, we can make up by storing an appropriate set of pictures to create a movie that allows the scientist a 3-d experience. But this will hamper the direct interaction with the data.

b) On some tools, the user interfaces were not adequate to allow the scientist a direct manipulation of the data. Without "the tool in his hands" to explore the cube, John Bally feels that he cannot give us an adequate feedback of its use.

c) As soon as a tool was useful enough to enable John Bally to identify interesting features on the pictures, he wanted to perform computations on the corresponding data values (e.g. integrate data values in the neighborhood; compare the median value of one subset with the median value of another data cube subset). It is not hard for a programmer to apply computations to single data values or a collection of such data values; however, the tools in their current status did not allow an easy handling of such manipulations.

The understanding derived from the work with John Bally points towards a strong need for interaction with the data, direct manipulation of the representations, and much more iteration between the encoding and decoding process than presented in Figure 4.

5. Conclusions

We have found that there is need to sacrifice generality and concentrate on specific needs of individual scientific groups in order to better understand the potential of scientific visualization. This means spending sufficient time in the environment of scientists to follow various steps in the data analysis process that lead to the need for data visualization. In our case, cognitive techniques were helpful in soliciting feedback from scientists. Obstacles that hamper the use of visualization often lie before the rendering process that is part of a visualization system, e.g. in accessing special data formats and in the use of a complex conglomerate of software packages to fulfill various aspects of data analysis. Other visualization researchers collaborating closely with scientists have pointed out similar problems (e.g. Treinish, 1989 and 1990).

New methodologies are only of interest to a scientist if the need to use them exists. New methods that may have a strong potential for future applications are not attractive enough to spend time with. In the case where the need existed to explore multi-dimensional data sets, the scientist was interested in experimenting with new means of visual representations, however, was still mainly attracted to visual representations that seemed "natural". Scientists need to go beyond looking at pictures representing their data: they need to interact with the picture, relate screen positions back to original data values, and perform computations on these numbers.

Acknowledgment

The author is grateful to all the scientists and students at the Center for Astrophysics and Space Astronomy, specifically to Prof. John Bally, for the collaboration. Kristina Mickus-Miceli performed the user interface study for STAR as a graduate student during 1990 and 1991 and did all of the initial software development. Scott Merkle, Anthony Veale and Ron Chaiken greatly supported the development of STAR. The project is supported by NASA under grant NAGW-1902. Gitta Domik's efforts were partly also sponsored under NSF grant ASC-9001115.

Many thanks to John Wilson, who produced Figures 5 through 8 as part of an independent study on volumetric rendering (using AVS and IDL). Thanks also to Walter Fortson (Figure 9) and Chris Joslyn (Figure 10) who produced visualizations of collapsed data slices as part of my undergraduate course on scientific visualization. These were two of a number of good representations on collapsed data slices developed by students after discussing Gestalt theory and work done at the University of Lowell and by Jeff Beddow.

References

De Ferrari, L., 1991, *New Approaches in Scientific Visualization,* CSIRO Division of Information Technology, GPO Box 664, Canberra ACT 2601, Australia, Technical Report CSIRO-DIT TR-HJ-91-06.

Domik, G. O. and Mickus-Miceli, K.D. (in print for Vol 15, 1992), *Design and Development of a Data Visualization System in a Workstation Environment,* J. Microcomputer Applications.

Domik, G.O., 1991, *The Role of Visualization in Understanding Data,* Lecture notes on Computer Science 555, Springer Verlag, "New Trends and Results in Computer Science", pp 91-107.

Gordon, I.E., 1989, *Theories of Visual Perception,* John Wiley & Sons, ISBN 0 471 92196 3.

Mackinlay, J., 1986, *Automating the Design of Graphical Presentations of Relational Information*, ACM Trans. Graphics, Vol. 5, No. 2, April 1986, pp. 110-141.

Mickus, K.D., 1991, *Participatory User Interface Design for Scientific Visualization Systems,* Master's Thesis at the Department of Computer Science, May 1991.

Mickus-Miceli, K.D. and Domik, G.O., 1991, *Participatory Design in the Development of a Scientific Software System*, Technical Report, Department of Computer Science, University of Colorado, Boulder, CO. 80309-0430.

Nadeau, D.R., Elvins, T.T. and Bailey, M.J., 1991, *Image Handling in a Multi-vendor Environment*, Proceedings Visualization '91, pp. 276-283, Ed.: G.M. Nielson and L. Rosenblum, IEEE Computer Society Press.

Robertson, P.K., 1990, *A Methodology for Scientific Data Visualization: Choosing Representations Based on a Natural Scene Paradigm*, Proceedings of the Conference on Visualization '90, October 23-26, San Francisco, pp. 114-123.

Treinish, L.A., 1989, *An interactive, discipline-independent data visualization system*, Computers in Physics, July/August 1989, 55-64.

Treinish, L.A., 1990, *Interactive Visualization Requirements for Scientists*, presented at the Panel on Interaction Issues in Visualization at the Conference on Visualization /90, Oct. 23-26, San Francisco.

Wehrend, S., and Lewis, C., 1990, *A Problem-Oriented Classification of Visualization Techniques*, Proceedings of the Conference on Visualization '90, October 23-26, San Francisco, pp. 139-143.