Exploring a High-Dimensional Galaxy Dataset

Nicole Atherly* Mahir Kothary† Grace Telford‡

ABSTRACT

We present a tool that facilitates interactive exploration of correlations between parameters in a high-dimensional astronomy dataset. This dataset contains measurements of 25 properties of 10,000 galaxies derived from spectra from the Sloan Digital Sky Survey. Our web-based tool, implemented using D3 and JavaScript, allows the user to rapidly generate one and two-dimensional orthogonal projections of the dataset and dynamically change these projections. It also enables brushing and linking between plots so that the user can search for variations in the distribution of galaxies in different regions of parameter space. We obtain feedback on this tool from astronomers and other scientists, many of whom express interest in applying our system to enable exploration of other multidimensional datasets.

1 INTRODUCTION

In astronomy, studying correlations between various galaxy properties (e.g., the mass in stars, the rate of star formation, or the chemical composition of the gas between stars) for large sample sizes can be used to make inferences about the physical processes that govern galaxy evolution. However, measurements of such properties are derived from modeling the distribution of light emitted by a galaxy, and therefore are model-dependent and likely biased. Before drawing conclusions about physical mechanisms based on the correlations between such galaxy properties, it is important to understand the systematic errors that may be affecting the observed correlations.

The process of looking for biases in a high-dimensional dataset can be time consuming, hindering progress in research. It is difficult to choose useful and interesting low-dimensional projections of the data to investigate, given the very large number of possibilities. Further, just looking at the distributions of the data in one or two-dimensional projections does not help in identifying systematics; it is necessary to compare the distribution of the full dataset to subsamples of the data in different regions of parameter space. This can be accomplished by color coding points meeting some filtering criterion, but manually generating many such figures that efficiently sample parameter space using traditional analysis methods (e.g. Python or MATLAB) is prohibitively slow.

A tool that facilitates rapid, interactive exploration of the distributions of a dataset in many one and two-dimensional orthogonal projections would help astronomers to better understand the biases in multidimensional datasets. To be useful, such a tool must allow the user to quickly change the parameter(s) shown in a given projection. Further, the user must be able to to compare data points in a small region of parameter space to the full dataset and quickly change the region of interest in parameter space.

This paper describes our design of a tool that enables detailed exploration of a high-dimensional astronomy dataset containing measurements of 25 different properties of 10,000 star-forming galaxies. Parameters obtained from the MPA/JHU Catalog [4] were measured from spectra taken by the Sloan Digital Sky Survey [10]. The Python module `pyqz` [1] was used to calculate some parameters.

2 RELATED WORK

The problem of effectively visualizing high-dimensional data is an active area of research, especially as large and complex datasets are becoming more common. Visual representations are limited to two or three dimensions, and as the number of dimensions increases, the time required to manually search for an informative low-dimensional projection of the data increases rapidly.

Various techniques have been devised to speed up the process of finding a useful projection of the data. The technique of parallel coordinates is used to display many dimensions of a dataset in two dimensions and allow the user to see structure in the data [3]; however, the ordering of the parameter axes strongly affects the utility of the resulting visualization. `TOPCAT` [8] is an astronomy-specific tool that allows rapid manipulation of large data tables; however, this tool does not allow for easy and rapid exploration of changing distributions of data in different regions of parameter space. `Ggobi` [6] is an interactive visualization tool that "tours" the dataset by rotating through many low-dimensional projections of the data so that the user can rapidly see many different views and quickly identify interesting projections. It also allows the user to use brushing/linking between multiple graphs to explore structure in the data.

Several authors have devised methods of automating the choice of useful projections, using orthogonal [7] and non-orthogonal [9] projections. Studies that analyze the relative effectiveness of different methods of choosing low-dimensional projections have found that the optimal dimensionality reduction technique is highly dependent on the nature and size of the dataset [5, 2]. At present, there exist many methods of finding potentially useful low-dimensional projections of multidimensional datasets, but the success of each technique in identifying interesting structure in the data quickly varies by dataset and by task.

Since this dataset is known to be noisy (both due to Poisson distributed photon counting errors and systematic effects) and the errors are usually non-Gaussian, we choose to avoid machine-learning approaches to finding useful non-orthogonal projections, as such projections would likely be dominated by

*athern@uw.edu
†mahirk@uw.edu
‡otelford@uw.edu

noise. Given that the best method for automatically selecting useful low-dimensional projections is so task-dependent, we design our own system for exploring this multidimensional dataset of galaxy parameters. This enables us to include the features tailored to the specific problem of searching for biases in the data.

3 METHODS

A thorough description of our implementation in D3.

4 RESULTS

A short summary of what we have produced and what it allows the user to do to accompany the two screenshots.

5 DISCUSSION

5.1 Design Tradeoffs

Discussion of tradeoffs in the design.

5.2 Informal Feedback

We obtain feedback from two sources: (1) people who viewed our poster describing this project, and (2) astronomy graduate students who were asked to play with our galaxy dataset exploration tool and give their impressions and suggestions.

Overall, both the astronomers and poster session attendees had positive comments about this tool, and many of the people with whom we spoke expressed interest in adapting our code to explore other multi-dimensional datasets in fields ranging from astronomy to biology to materials testing. Most people who used the tool said that our tool would be useful for searching for interesting correlations and biases in a high-dimensional dataset.

Of course, even though people generally liked the tool, we received some useful suggestions for improvement. The following is a compilation of some suggestions from both astronomers and poster session attendees:

- Change appearance of drop down menus to be more modern, perhaps including a mini-view of histogram next to each variable name

- Allow for non-rectangular brushing area (e.g., triangular or freeform)

- Switch to a density map to avoid obscuration issues and better show distribution of data points in scatter plot

5.3 Usefulness for Intended Task

Co-author Grace Telford uses this dataset in her research. In the course of playing with this tool during development, she has already seen some interesting trends in plots that she had never thought to make before (e.g., bifurcation in one of the scatter plots generated). Before this project began, a few biases in the dataset had already been found, and those biases are indeed noticeable in the plots generated using our tool. Therefore, from the point of view of the intended user, our design is indeed useful for searching for bias in this dataset.

6 FUTURE WORK

6.1 Design Alterations

We would like to incorporate many of the suggestions listed in Section 5.2, particularly switching from a scatter plot to a heat map or adding a toggle between the two views. This would provide a more informative view of the distribution of galaxies in the various two-dimensional projections. Of course, we would have to alter the brushing area to snap to the edges of the bins used to generate the heat map.

Generally, this tool can be made even more flexible by allowing the user more control over what types of plots are generated; e.g., each histogram could have a toggle to switch between a histogram and a scatter or density plot.

We could also move the drop down menus for selecting the parameters used in the scatter plot to the locations of the axis labels. This would free up space next to the scatter plot, which could then be devoted to providing more details-on-demand. If a user clicks on a point on the scatter plot (or bin in the heat map), then a box could appear next to the plot showing more information about that galaxy (or statistics of the bin).

6.2 Extending to Other Datasets

Given that so many people have already expressed interest in applying this tool to their data, we would like to make it as easy as possible for users to adapt our code to analyze other multi-dimensional datasets. We plan to add a license to our Github repository and advertise this project on personal webpages so that others can fork our repository and use our code. Of course, this will require thorough documentation and a description of how to change the axis labels, which are currently hard-coded.

Because D3 is used to render our figures (and the tool contains a scatter plot with one circle rendered per data point), we are limited in the amount of data that our tool can handle. Currently, it is working well on a dataset of 10,000 galaxies with 25 measured parameters. However, this is a subsample of the full dataset of about 140,000 galaxies. The clear next step is to add a database backend to this system to speed up binning and selecting data. Our goal is to scale up this tool to handle hundreds of thousands of data points.

REFERENCES

[1] Dopita, M. A. et al., 2013, *New Strong-line Abundance Diagnostics for H II Regions: Effects of κ-distributed Electron Energies and New Atomic Data*, The Astrophysical Journal Supplement, 208, 1

[2] Etemandpour, R., et al., 2014, *Perception-Based Evaluation of Projection Methods for Multidimensional Data Visualization*, IEEE Transactions on Visualization and Computer Graphics, 21, 81 - 94

[3] Inselberg, A., 1997, *Multidimensional Detective*, IEEE Symposium on Information Visualization, 100 - 107

[4] Kauffmann, G. et al., 2003, *Stellar masses and star formation histories for 10^5 galaxies from the Sloan Digital Sky Survey*, Monthly Notices of the Royal Astronomical Society, 341, 33

[5] Keim, D. & Krigel, H., 1996, *Visualization Techniques for Mining Large Databases: A Comparison*, IEEE Transactions on Knowledge and Data Engineering, 8, 923 - 938

[6] Swayne, D., Lang, D. T., Buja, A., and Cook, D., 2003, *GGobi: evolving from XGobi into an extensible framework for interactive data visualization*, Computational Statistics & Data Analysis, 43, 423 - 444

[7] Tatu, A., et al., 2009, *Combining Automated Analysis and Visualization Techniques for Effective Exploration of High-Dimensional Data*, IEEE Symposium on Visual Analytics Science and Technology, 59 - 66
[8] Taylor, M., 2005, *TOPCAT & STIL: Starlink Table/VOTable Processing Software*, Astronomical Data Analysis Software and Systems XIV ASP Conference Series, 347, 25
[9] Yang, J., Peng, W., Ward, M. O., and Rundensteiner, E. A., 2003, *Interactive Hierarchical Dimension Ordering, Spacing, and Filtering for Exploration of High-Dimensional Datasets*, IEEE Symposium on Information Visualization, 105 - 112
[10] York, D. G. et al., 2000, *The Sloan Digital Sky Survey: Technical Summary*, The Astronomical Journal, 120, 1579

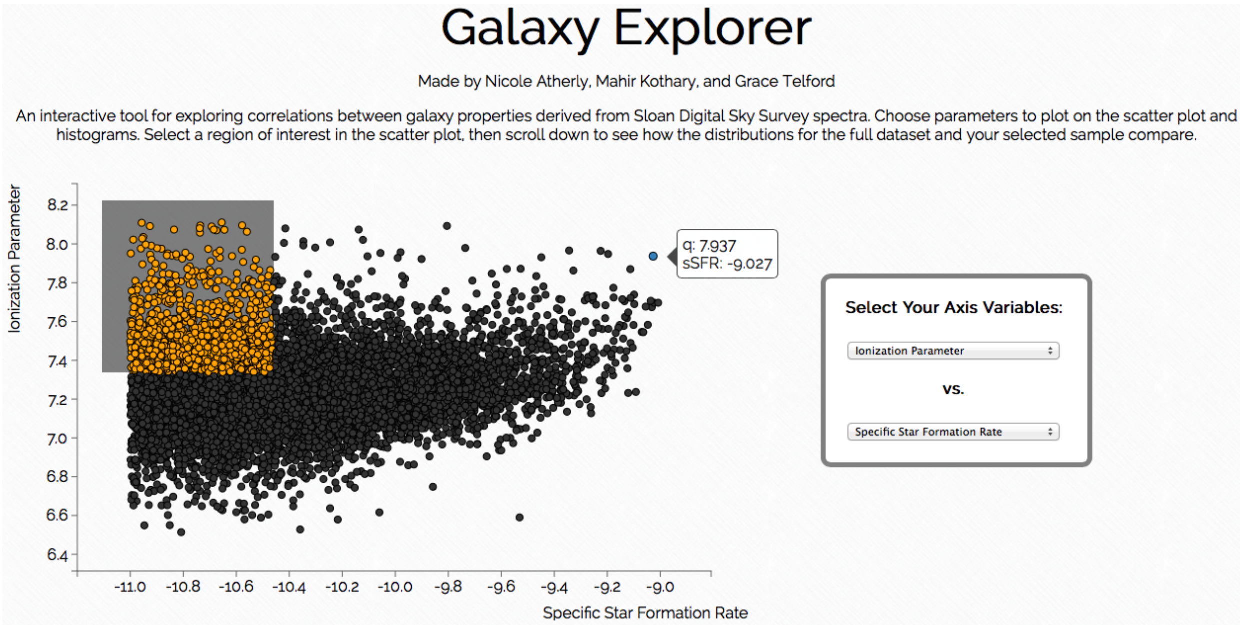

Figure 1: The main view of our design. The user is able to select any two parameters out of 25 available parameters to plot against each other on a scatter plot, where each point shows the values of those parameters for one galaxy. A tooltip shows the exact values of those quantities on mouseover. The user may then draw a rectangular brushing area to highlight points in a region of interest; these selected points are linked to histograms below (see Figure 2.)

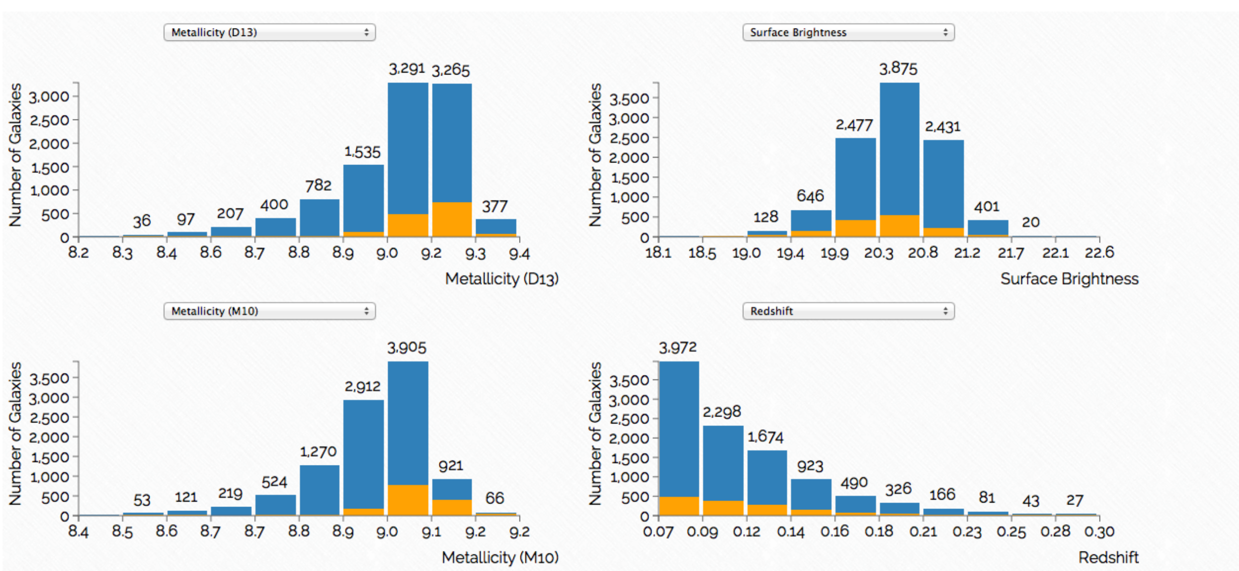

Figure 2: The lower half of our design. Four histograms show the distribution of the full dataset in four parameters, which the user can choose. The selected points inside the brushing area in the scatter plot above (see Figure 1) are over plotted as orange bars, showing how the distribution of that subsample of the data compares to the full sample. A tooltip shows the number of galaxies in the orange bars and the median value of the plotted parameter in each bin on scroll over.