# SpRay - an R-based visual-analytics platform for large and high-dimensional datasets

J. Heinrich[1,2], J. Dietzsch[1], D. Bartz[2] and K. Nieselt[1]
[1]Center for Bioinformatics, University of Tübingen Sand 14, 72076 Tübingen, Germany
email: {juheinri, dietzsch, nieselt}@informatik.uni-tuebingen.de
[2]ICCAS/VCM, University of Leipzig, Germany
email: dirk.bartz@medizin.uni-leipzig.de

March 31, 2008

Recently developed high-throughput methods produce increasingly large and complex datasets. For instance, microarray-based gene expression studies generate data for several thousands of genes under numerous different conditions, yielding large, heterogeneous, potentially incomplete or conflicting datasets. From both technical and analytical points of view, extracting useful and relevant information - known as the knowledge discovery process - from these large data sets is a challenge. While the technical capacity to collect and store such data grows rapidly, the ability to analyze it does not advance at the same pace. The extraction of relevant information from large and high-dimensional data is very difficult and requires the support of automated extraction algorithms based on statistical computing. Unfortunately, the unsupervised application of these statistical measures does not guarantee the successful extraction of relevant information, but requires critical consideration itself. Hence, the use of interactive visualization methods for the simultaneous evaluation of the applied statistical models is of central relevance and plays therefor a key role in the emerging field of visual analytics.

The aim of the work is to combine statistical methods with modern visualization techniques in an extendable, hardware-accelerated visual-analytics framework. We are currently developing SpRay (viSual exPloRation and AnalYsis of high-dimensional data), which provides for the explorative analysis of large, high-dimensional datasets in accordance with the visual-analytics paradigma. Similar to GGobi [SLBC03], the statistical backend is provided through R, as a plugin. The performance-oriented design of SpRay, which uses hardware-accelerated graphics (OpenGL), C++ and Qt, also allows very large datasets to be explored with greatly reduced response times. The use of modern GPUs (OpenGL) further accelerates the application of different transparency-modulations and color maps to the currently implemented plugins, such as refined parallel coordinates and scatterplots. All plugins (currently: parallel coordinates, scatterplots, TableLens, TableView, Histogram, R-Console, Brushing) are linked by means of a common data model which is particularly useful to tightly integrate R along with all its extensions via packages. Hence, adequate statistical values may be defined and interactively visualized together with the raw data, providing an iterative, interactive and integrated approach to the analytical reasoning process as proposed by the visual-analytics-paradigm. The benefit of the currently implemented features has succesfully been demonstrated with different gene-expression datasets [DHNB06, DHNB08].

## References

[DHNB06] J. Dietzsch, J. Heinrich, K. Nieselt, and D. Bartz. Poster: Extended parallel coordinates for bioinformatical applications. In *German Conference on Bioinformatics (GCB)*, Tübingen, 2006.

[DHNB08] J. Dietzsch, J. Heinrich, K. Nieselt, and D. Bartz. Visual Analysis of Microarray Data from Bioinformatics Applications. Technical Report WSI-2008-1, ISSN 0946-3852, Dept. of Computer Science (WSI), University of Tübingen, 2008.

[SLBC03] D. Swayne, D. Lang, A. Buja, and D. Cook. GGobi: Evolving from XGobi into an Extensible Framework for Interactive Data Visualization. *Computational Statistics & Data Analysis*, 43:423–444, 2003.