SpRay

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

J. Heinrich[1] J. Dietzsch[1] D. Bartz[2] K. Nieselt[1]

[1]Center for Bioinformatics, University of Tübingen

[2]ICCAS/VCM, University of Leipzig

August 12, 2008

Outline

1. Introduction
2. SpRay
3. Discussion
4. Future Work

Data Sets Become Increasingly Large

High-Throughput techniques yield a huge amount of data

- Microarrays
- CT scanner
- Simulation data

Many data sets are high-dimensional

- Time series: 100 experiments, 5 replicates, 10000 oligos
- 10000 rows $\times$ 500 columns = $5 \cdot 10^6$ data points

. . . and complex

- Heterogeneous data (categorical, metric)
- Invalid data (NA, NaN)

Knowledge Discovery Becomes Increasingly Difficult

Effects of Large and High-Dimensional Datasets for the Analysis

- Storage: obvious
- Speed: time to read, locate, compute, render, display the data
- Quality: errors, administration
- Complexity: more variables, more detail, special cases...
- Visualization: Dimensionality, Occlusion, Identification

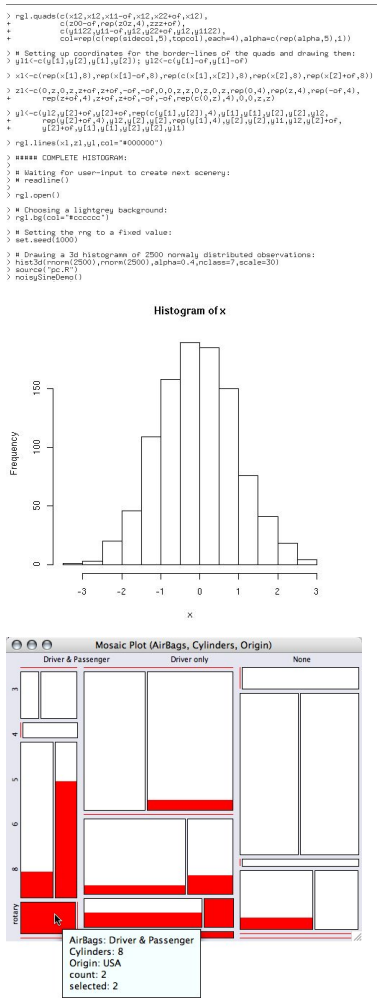
Visual Analytics with R

Analytical Reasoning

- Gain insight into data
- Reveal underlying structure and model
- Extract information contained

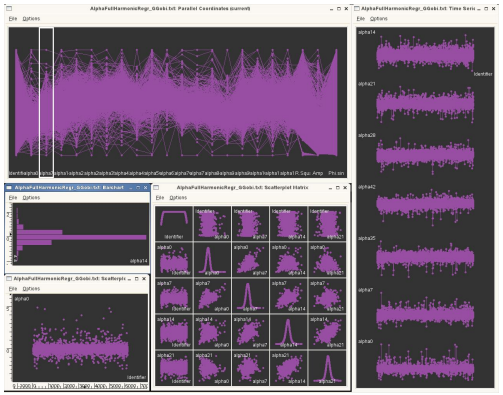
Techniques

- Data Analysis
- Visualization
- Interaction



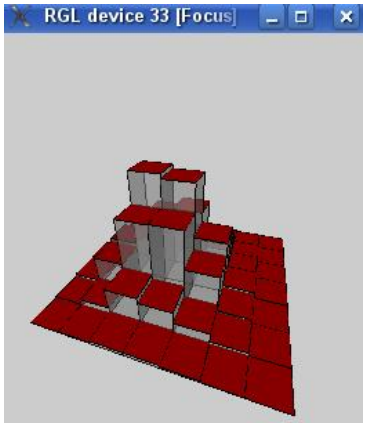
Visual Analytics with R
Related Work

GGobi[1]



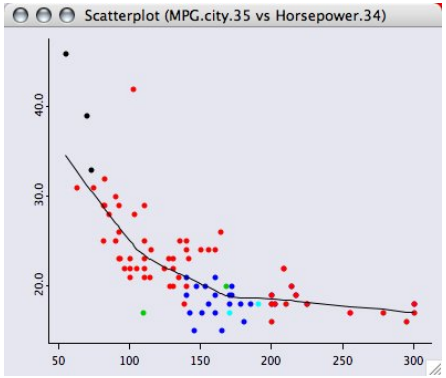
- linked views
- CPU only
- R optional

RGL[2]



- no linked views
- CPU/GPU
- depends on R

iPlots[3]

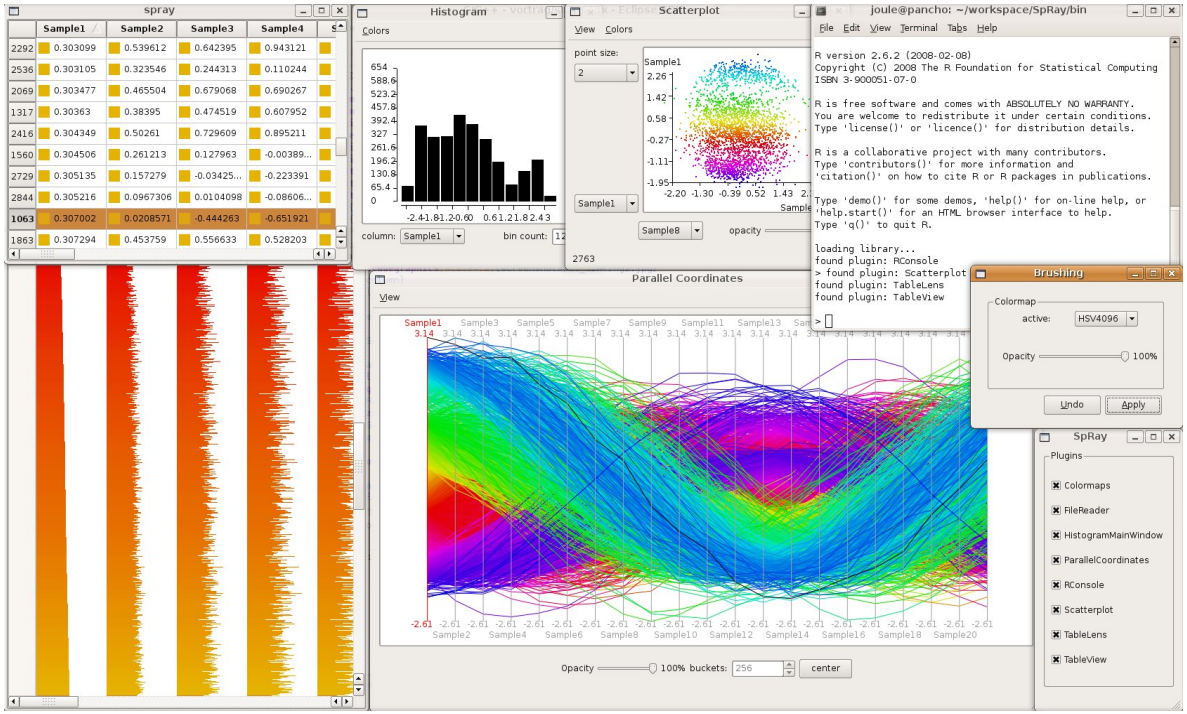


- linked views
- CPU/GPU
- depends on R

[1] [Swayne et al., 2003]
[2] [Adler and Nenadic, 2003]
[3] [Urbanek and Theus, 2003]

SpRay

viSual exPloRation and anAlYsis of high-dimensional data



- linked views
- CPU/GPU
- R optional

SpRay
Objectives

Objectives

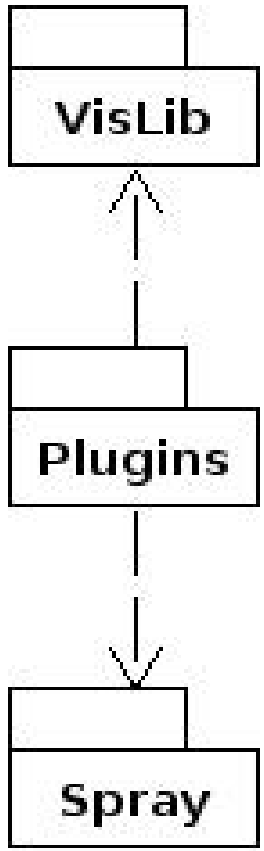
- Extendable
- Interactive
- Portable
- Statistical Backend
- High-Performance

SpRay
Architecture

VisLib
- Independent Visualization Library

Plugins
- Implement the plugin-interface
- Make use of VisLib (optional)

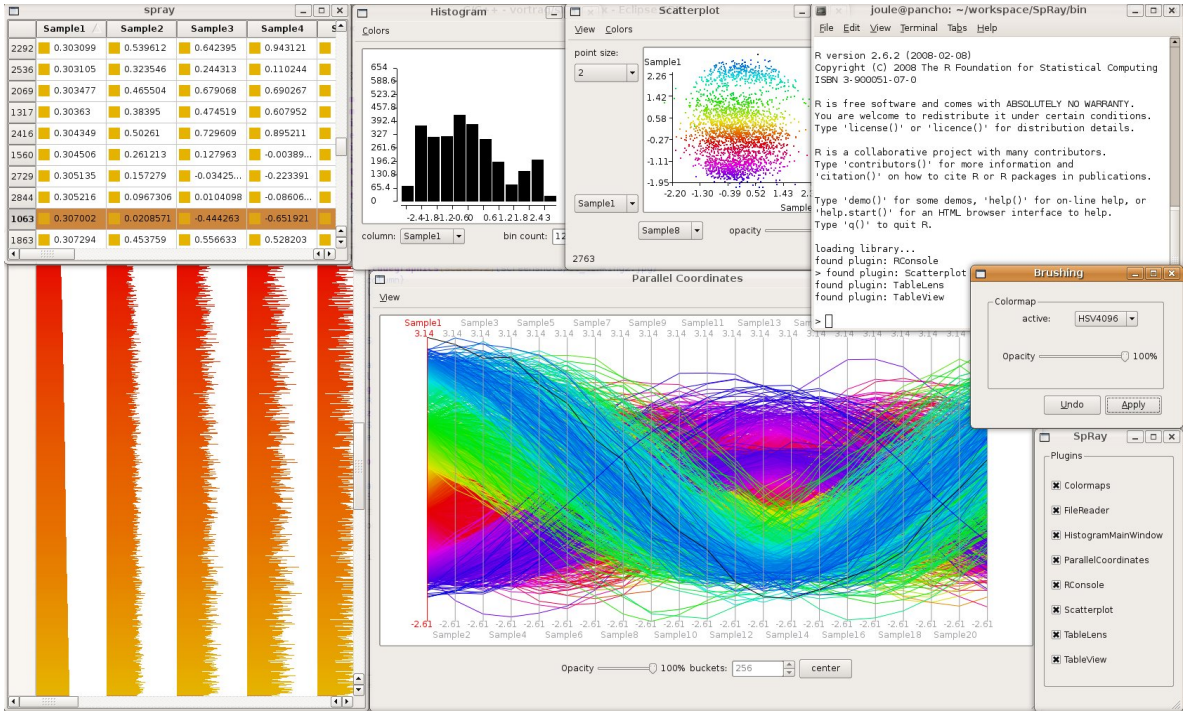
Host Application
- Defines the plugin-interface
- Organizes communication



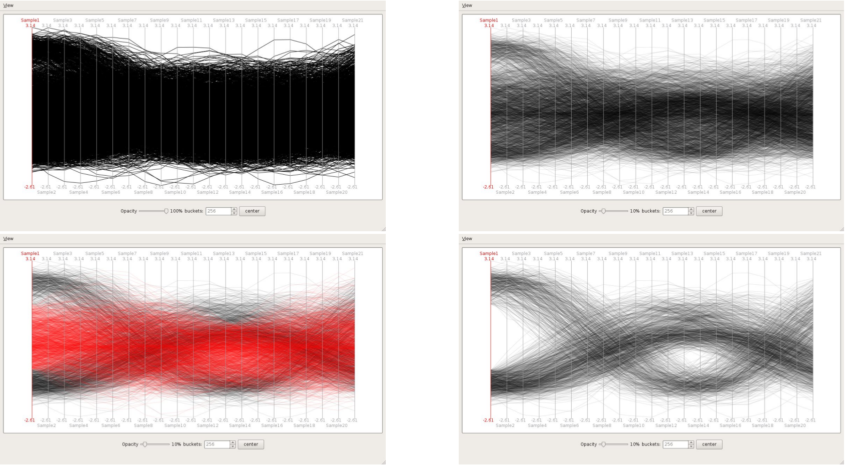
Plugins

Currently available

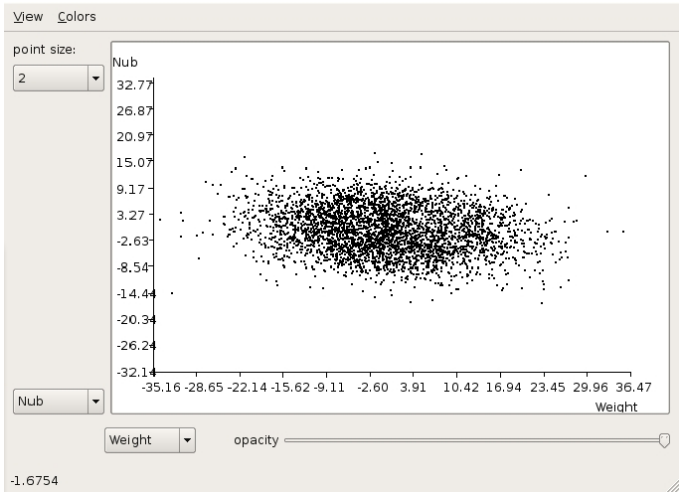
- Parallel Coordinates
- Scatterplot
- Histogram
- Data Table
- TableLens
- R-Console
- Brushing

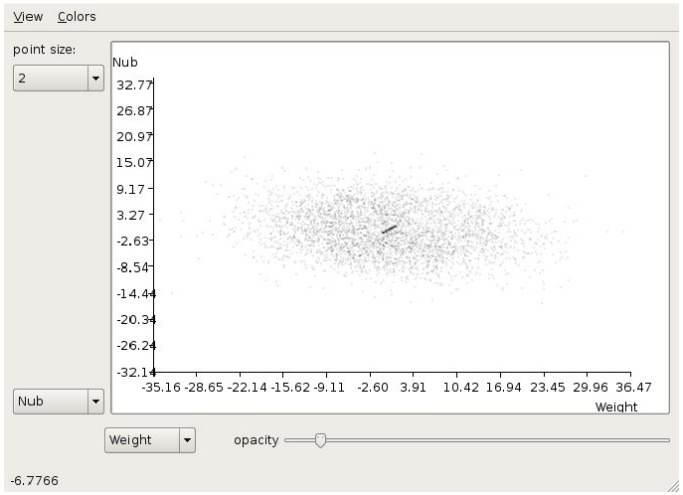


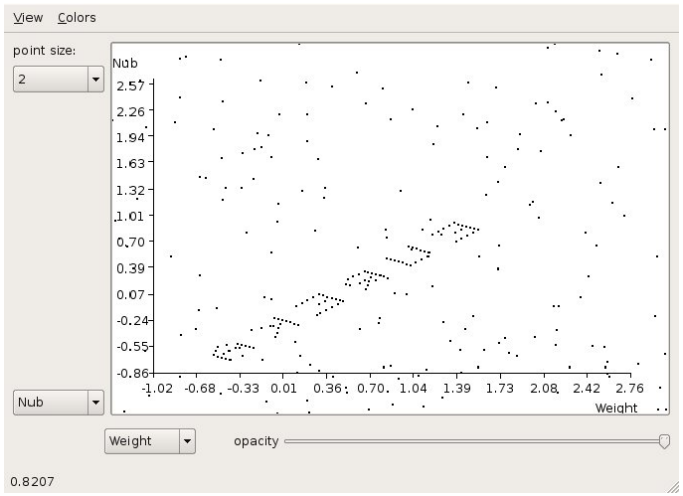
Parallel Coordinates

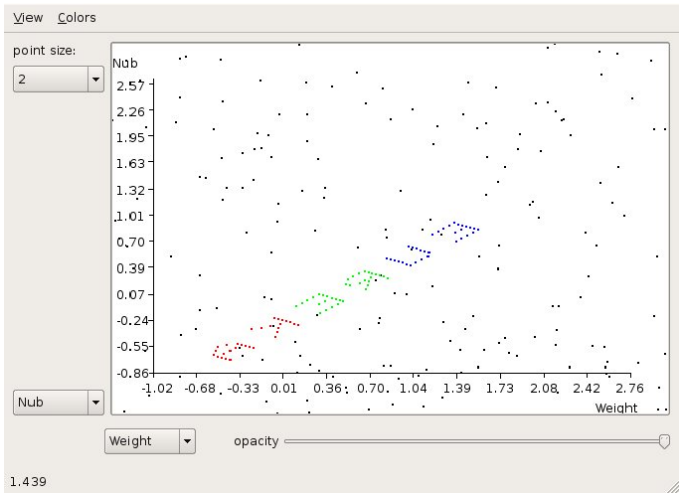


Scatterplot







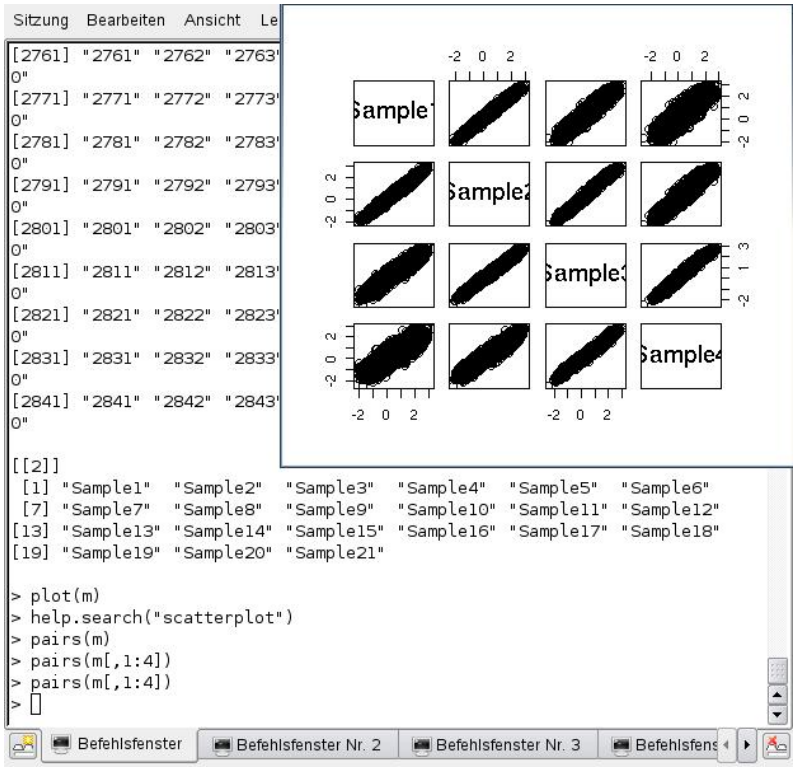


Data Table and R-Console

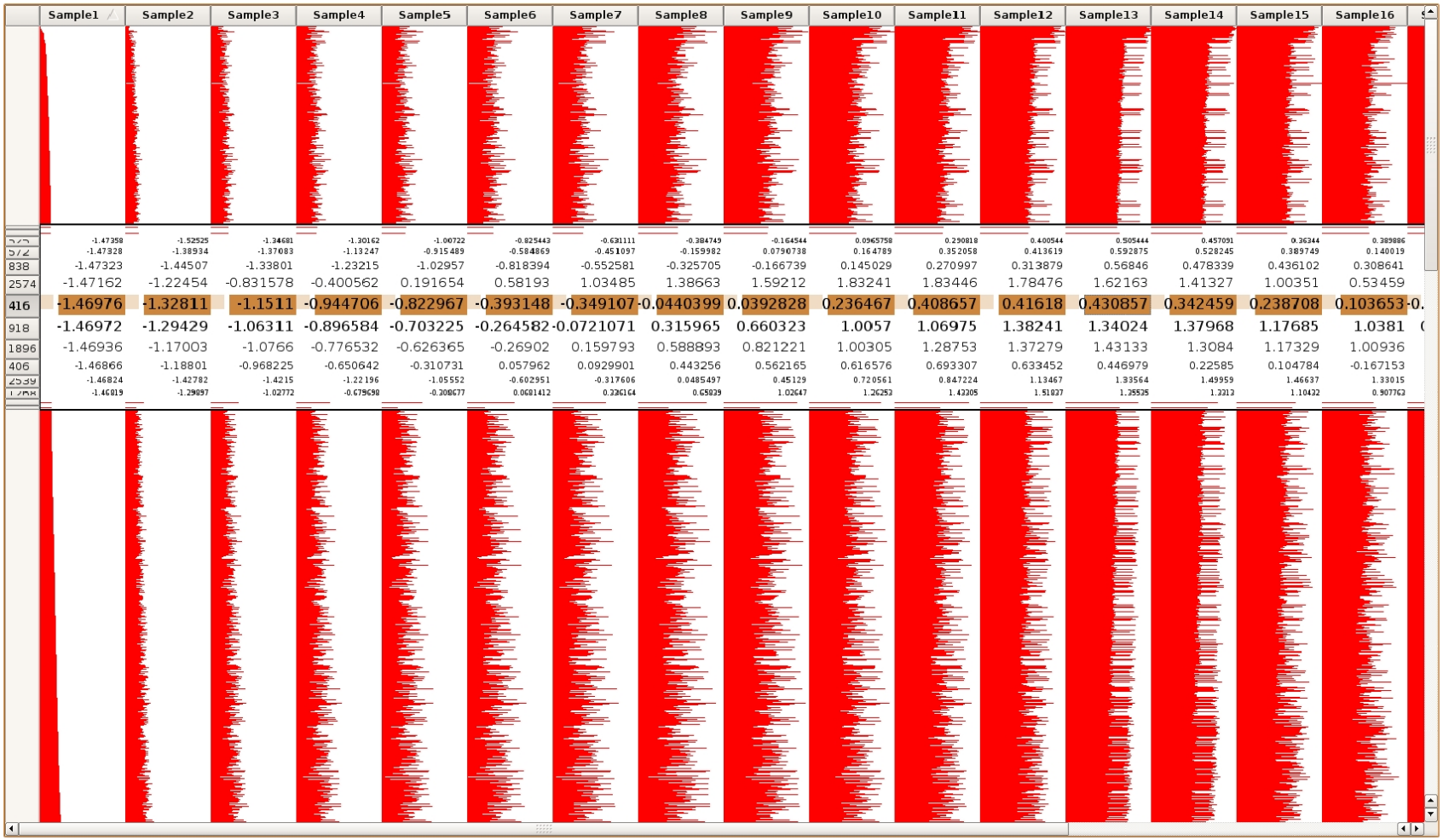
Data Table

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
105	2.3415859163...	2.2376690546...	1.8982071874...	1.5555907577...	1.3721192240...	1.0038125142...
106	2.4401413565...	2.5040031651...	2.2856312029...	2.3950072936...	1.9836212244...	1.4697343178...
107	1.3811222629...	1.0465605603...	0.5781103422...	-0.233239960...	-0.621642829...	-0.807906840...
108	2.5384502012...	2.6588043399...	2.3529196256...	2.0231697166...	1.4417099692...	1.1541864958...
109	2.6865113599...	2.5672115712...	2.2875285175...	1.7987117521...	1.6185984467...	0.9827138999...
110	1.5442642534...	1.0196378953...	0.7877514254...	0.1382789851...	-0.403685698...	-0.692011846...
111	2.0488761740...	2.4978078321...	2.4989904982...	2.3638465612...	2.4162556514...	2.3486733609...
112	2.4515904970...	2.4102145324...	2.2785772702...	2.3462009364...	1.6576799128...	1.1253601772...
113	2.8350728203...	2.6358395454...	2.7894649536...	2.4301788917...	2.3818830086...	1.5301226056...
114	2.6457625081...	2.6191339249...	2.2891254100...	2.0749760720...	1.7682662302...	1.2335163649...
115	2.7380323543...	2.8372698299...	2.8199595139...	2.4995996580...	2.0364326709...	1.6025039883...
116	2.2889389588...	2.3867752810...	2.2150238929...	1.8784653441...	1.7698521355...	1.3650019018...
117	2.0939331355...	2.1002774908...	1.6969150501...	1.2723382349...	0.7364192866...	0.3901431583...
118	1.9801059987...	1.6898704150...	1.5512342968...	0.9184525149...	0.6300543819...	0.1504721183...
119	2.0147595769...	2.1199225332...	2.2438563815...	2.3026739011...	2.2815192030...	2.0488097932...
120	1.7817203160...	2.0466779923...	2.1316428308...	2.2598791103...	2.2426233845...	2.1510957519...
121	2.6658253103...	2.5174836628...	2.3540442541...	1.8444133751...	1.2336569553...	0.8281764888...
122	2.3363089591...	2.0727448585...	1.6733367610...	1.5319254814...	1.1105856698...	0.6047233089...
123	1.6921383910...	2.0874461781...	2.4729319179...	2.6812284408...	2.3453695262...	2.5820502901...

R-Console

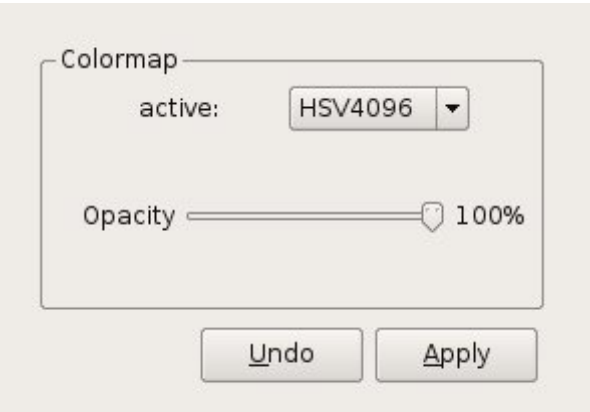


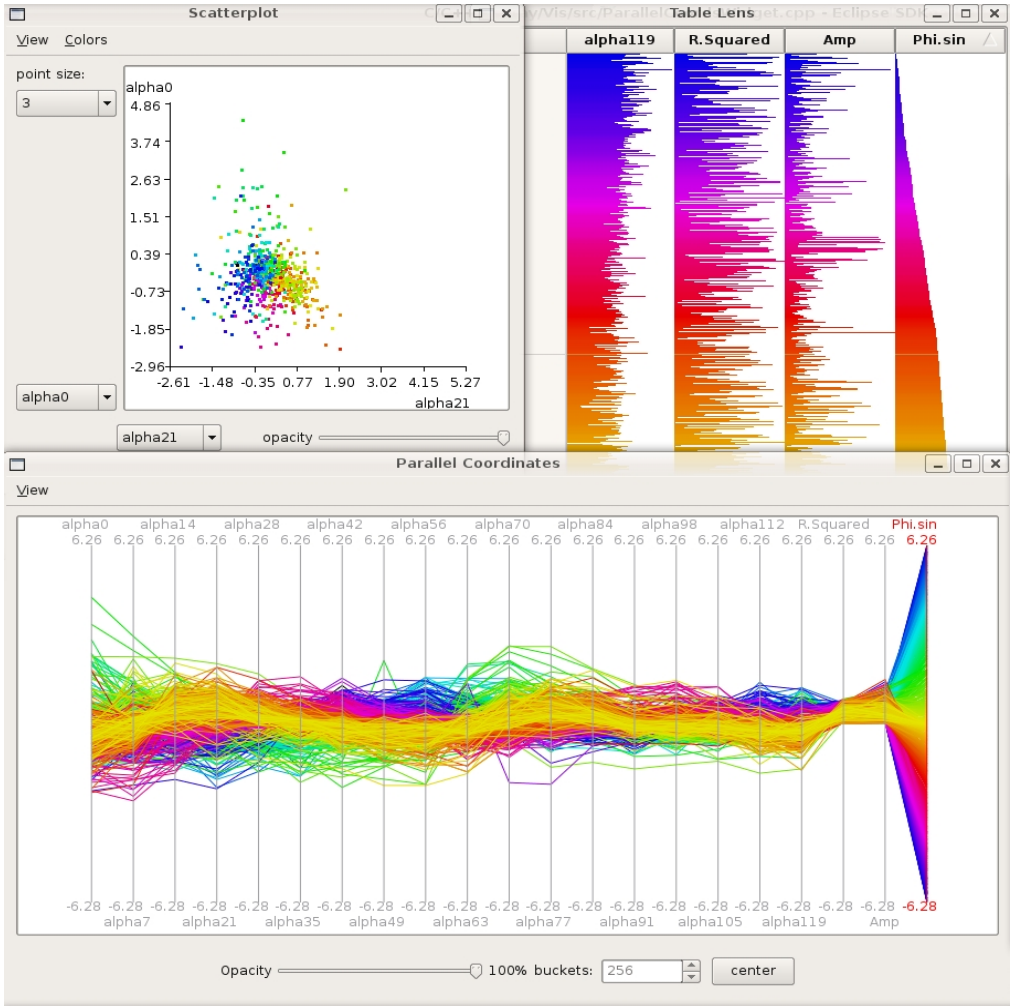
TableLens



[Rao and Card, 1994]

Linking and Brushing





Performance

Depends on

- Size of the data set
- Number of plugins loaded
- Operation in progress
- Available hardware (GPU?)

Results

- Lower response times than GGobi/iPlots/RGL/Mondrian
- Good performance for middle-sized datasets

Discussion

Objectives achieved

- Extendable Visual-Analytics-Framework
- Independent Visualization Library
- Hardware-accelerated Graphics
- Statistical Backend using R
- Interactivity
- Good performance / Low response times

Problems

- Redundancy in frequently used calculations
- Very basic interface to R
- categorical data only supported via the R-plugin

Future Work

Future Work

- Incorporate meta-information into datamodel to avoid redundancy (e.g. maxima)
- Add/Improve plugins (Heatmap, 3D Plots, . . .)
- Extend interface to R (hot-linking, selections)
- Improve GPU-usage (textures, framebufferobjects . . .)

Thank You!

References I

Adler, D. and Nenadic, O. (2003).

A Framework for an R to OpenGL Interface for Interactive 3D graphics.
In *Proc. of the 3rd International Workshop on Distributed Statistical Computing*.

Rao, R. and Card, S. K. (1994).

The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular i nformation.
In *Proc. of SIGCHI conference on Human factors in computing systems*, pages 318–322, New York, NY, USA. ACM.

Swayne, D. F., Lang, D. T., Buja, A., and Cook, D. (2003).

GGobi: evolving from XGobi into an extensible framework for interactive data visualization.
Computational Statistics and Data Analysis, 43(4):423–444.

Urbanek, S. and Theus, M. (2003).

iPlots - High Interaction Graphics for R.
In *Proc. of the 3rd International Workshop on Distributed Statistical Computing*.