SpRay

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

J. Heinrich\textsuperscript{1} \hspace{1em} J. Dietzsch\textsuperscript{1} \hspace{1em} D. Bartz\textsuperscript{2} \hspace{1em} K. Nieselt\textsuperscript{1}

\textsuperscript{1}Center for Bioinformatics, University of Tübingen
\textsuperscript{2}ICCAS/VCM, University of Leipzig

August 12, 2008
Outline

1. Introduction
2. SpRay
3. Discussion
4. Future Work
Data Sets Become Increasingly Large

<table>
<thead>
<tr>
<th>High-Throughput techniques yield a huge amount of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Microarrays</td>
</tr>
<tr>
<td>- CT scanner</td>
</tr>
<tr>
<td>- Simulation data</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Many data sets are high-dimensional</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Time series: 100 experiments, 5 replicates, 10000 oligos</td>
</tr>
<tr>
<td>- 10000 rows $\times$ 500 columns $= 5 \cdot 10^6$ data points</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>...and complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Heterogeneous data (categorical, metric)</td>
</tr>
<tr>
<td>- Invalid data (NA, NaN)</td>
</tr>
</tbody>
</table>
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

- linked views
- CPU only
- R optional

RGL

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

iPlots

- linked views
- CPU/GPU
- depends on R

---

1[Swain 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

- Extendable
- Interactive
- Portable
- Statistical Backend
- High-Performance
SpRay - an R-based visual-analytics platform

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
Data Table

<table>
<thead>
<tr>
<th>Sample1</th>
<th>Sample2</th>
<th>Sample3</th>
<th>Sample4</th>
<th>Sample5</th>
<th>Sample6</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3419059163</td>
<td>2.2976690546</td>
<td>1.8982071874</td>
<td>1.555907577</td>
<td>1.3721192240</td>
<td>1.0358125142</td>
</tr>
<tr>
<td>2.401413565</td>
<td>2.504003165</td>
<td>2.2856312029</td>
<td>2.365072935</td>
<td>1.965212244</td>
<td>1.497343178</td>
</tr>
<tr>
<td>1.381122629</td>
<td>1.0456056603</td>
<td>0.5781103422</td>
<td>0.233230960</td>
<td>0.621642829</td>
<td>0.807306840</td>
</tr>
<tr>
<td>2.5384002012</td>
<td>2.6580493999</td>
<td>2.372916257</td>
<td>2.023169156</td>
<td>1.441769962</td>
<td>1.154164902</td>
</tr>
<tr>
<td>2.6866113539</td>
<td>2.5672117572</td>
<td>2.287585175</td>
<td>1.798717521</td>
<td>1.618598446</td>
<td>0.982713699</td>
</tr>
<tr>
<td>1.5442649534</td>
<td>1.019369953</td>
<td>0.7677514254</td>
<td>0.138788815</td>
<td>0.403865898</td>
<td>0.652011846</td>
</tr>
<tr>
<td>2.0487671740</td>
<td>2.407878321</td>
<td>2.498904989</td>
<td>2.368345612</td>
<td>2.416255615</td>
<td>2.348763609</td>
</tr>
<tr>
<td>2.451504379</td>
<td>2.4102145324</td>
<td>2.278577202</td>
<td>2.346509384</td>
<td>1.657679318</td>
<td>1.1253601772</td>
</tr>
<tr>
<td>2.8350723203</td>
<td>2.6383058545</td>
<td>2.780464536</td>
<td>2.4301786917</td>
<td>2.368130086</td>
<td>1.5301226068</td>
</tr>
<tr>
<td>2.6457625081</td>
<td>2.6191330429</td>
<td>2.289125410</td>
<td>2.0749760720</td>
<td>1.760262302</td>
<td>1.235163649</td>
</tr>
<tr>
<td>2.7389034364</td>
<td>2.8372682299</td>
<td>2.8190595139</td>
<td>2.496596589</td>
<td>0.963452709</td>
<td>1.602503868</td>
</tr>
<tr>
<td>2.2897870588</td>
<td>2.3667521810</td>
<td>2.215023629</td>
<td>1.8748653441</td>
<td>1.769852185</td>
<td>1.365910168</td>
</tr>
<tr>
<td>2.0392913295</td>
<td>2.102274909</td>
<td>1.696715905</td>
<td>1.273382349</td>
<td>0.736152865</td>
<td>0.390143188</td>
</tr>
<tr>
<td>1.980159987</td>
<td>1.809870415</td>
<td>1.515342366</td>
<td>0.918145218</td>
<td>0.630554318</td>
<td>0.150472189</td>
</tr>
<tr>
<td>2.0147597679</td>
<td>2.1199253322</td>
<td>2.2435653817</td>
<td>2.302673991</td>
<td>2.2821521020</td>
<td>2.408597932</td>
</tr>
<tr>
<td>1.781703161</td>
<td>2.066779092</td>
<td>2.131642808</td>
<td>2.259871104</td>
<td>2.204262884</td>
<td>2.1518057519</td>
</tr>
<tr>
<td>2.0638253103</td>
<td>2.5174836528</td>
<td>2.3524042451</td>
<td>1.844413751</td>
<td>1.233505933</td>
<td>0.828176488</td>
</tr>
<tr>
<td>2.3386369591</td>
<td>2.0727449365</td>
<td>1.6738867610</td>
<td>1.539254068</td>
<td>1.110585668</td>
<td>0.604720894</td>
</tr>
</tbody>
</table>

R-Console

```r
> plot(m)
> help.search("scatterplot")
> pairs(m)
> pairs(m[,1:4])
> pairs(m[,2:4])
```

useR! 2008

SpRay - an R-based visual-analytics platform
TableLens

[Rao and Card, 1994]
Linking and Brushing

- Colormap
  - active: HSV4096
- Opacity: 100%

- Undo
- Apply
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

- 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 ...)

SpRay - an R-based visual-analytics platform
Thank You!
A Framework for an R to OpenGL Interface for Interactive 3D graphics.

The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information.

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

iPlots - High Interaction Graphics for R.