Final Report

Michael Gottlieb *mikemgottlieb@gmail.com*
Emily Hindalong *ehindalong@gmail.com*
Dmitry Tebaykin *dmitry.tebaykin@gmail.com*

December 16, 2015

Abstract

concise summary of your project. do not include citations.

Introduction

**give the big picture. establish the scope of what you did, some background material may be appropriat

Neuroelectro is a Django text mining application [http://neuroelectro.org/] and was published as an article in Frontiers in Neuroinformatics [1]. Through the combined use of text mining and manual curation, it currently hosts experimental data from over 500 articles and is expected to grow to host the experimental data of about 5,000 articles. The data of each article can be accessed by type of neuron, electrophysiological property, or via a table of articles.

Neuroelectro is unique amongst text mining projects in that it allows end users to interact with curated data directly. Most text-mining tools in the biomedical domain assume that the end user will want an association matrix for terms in a controlled vocabulary, such as MEDLINE or MeSH terms [26,27]. These tools automatically generate and output an association matrix without providing the user with a way to interface with the primary data. This limits the analyses the user can perform.

Related Work

**include both work aimed at similar problems and work that employs similar solutions to yours
structure into subsections based on your own synthesis of themes in the related work
although there is no requirement to establish research novelty since it's a course project, you should s
definitely cover academic work; it's often good to cover non-academic work as well (commercial software
you may choose to reorder sections to put this one after Abstractions, if it will help you write this s

Solutions to Similar Problems

Neuroelectro provides some crude data visualizations in the form of static scatterplots and a single PCA analysis plot. Most text-mining tools in biomedicine do not use visualization at all, and those that do are restricted to analyses on the derived association matrix. For example, VOSviewer [2] uses colour and spatial position to visualize the semantic clustering and strength of association across text mined terms. The Trading Consequence project [3] focuses on mined trading documents supported by controlled vocabularies to generate maps of commodity trading over time.

Exploring relationships Exploration of relationships between the different properties in Neuroelectroâs dataset is supported by the current version of Neuroelectro. However, it is a static set of strip plots that do not account for all combinations of variables and do not allow any interaction.

Exploration of relationships is a common task in many analytics platforms such as Tableau [17], SAP BOBJ AOLAP [32] and Microsoft Excel [33]. Generally, these platforms provide a tabular view of the data in

addition to customizable visualizations to enhance usersâ exploration of the data. Our goal differs from these platforms as we are not including a tabular view, we are limiting the usersâ choices to provide a simpler experience, and we are using plots that are not easily achieved with these platforms (e.g. interactive plots, hive-plots).

Providing an overview of data Essentially, we are facing a problem of visualizing a network when we are trying to give an overview of our data. Over the years many solutions have been proposed for this type of task: hairball [35], matrix [34], arc diagram [31], call network [37], hive plot [29], etc. Simply visualizing the network as a collection of nodes connected with edges (the hairball approach) seems impractical due to a large number of nodes and connections (currently: 150 nodes and ~10k edges), scaling is also a problem since the hairball only gets bigger with time. The matrix approach deserves some credit in terms of data visibility and it is a familiar visualization style to biologists, but there are 2 issues with utilizing matrices for this task: 1) Our data is 3 dimensional (neuron type, ephys. property, metadata) and 3D matrices are usually very hard to interpret, we could provide a faceted view of 1 matrix per metadata as a possible solutions, but the amount of screen space that would require is enormous; 2) Matrices do not scale well, the labels get too small to be legible at some point. A call network visualization would end up looking very similar to a hairball in our case, as a result we had to discard this possibility due to scalability issues. Arc diagrams came in as a close second as our visualization of choice - they are easy to interpret, pleasant to look at and they can scale reasonably well with the amount of evidence in the database. The problem with arc diagrams is that all nodes would end up being on one line and that does not represent the 3 distinct groups of nodes (neuron types, ephys. properties, metadata) in our data. As a result, we decided to use hive plots for providing an overview of our data. Krzywinski, Birol, Jones and Marra [30] describe the advantages of hive plots in terms of gaining quantitative understanding when visualizing networks. They also support: multiple axes, information encoding in the nodes and edges, scaling. The one issue with hive plots is that they are a fairly new visualization style and researchers may have trouble understanding what they are looking at. However, we plan to provide a guide to interpreting the hive plot as well as provide links to the supporting literature. This feature would be on-demand and can be disabled in the application preferences. Applications of Similar Solutions

Filter panels The filter panel paradigm, where one panel is used to control what data appears in the main panel, is well established in visualization domain [24,25,28]. An alternative solution is the filter bar, which uses less screen real estate [22]. However, we have opted to stick with filter panels because they will never interfere with the main view and will make it easier for the user to track which filters are applied at any given time. Furthermore, the number of filtering options that we offer will require a larger section of the screen.

There are two basic attribute-based filtering paradigms: drill-down and parallel selection [21]. As the referenced blog post describes, Amazon uses drill-down filtering and Kayak uses parallel filtering. Our solution uses a hybrid of these, allowing the user to drill-down categories and apply parallel selection within. We intend to refer to this blog post when designing the specific details of our filter panel.

Connected scatterplots Since Neuroelectro data is rather diverse (dozens of electrophysiology properties for each of over one hundred neuron types), we plan to utilize scatterplots [16] and connected scatterplots [7] for answering research-oriented questions. Haroz et al. [7] showed the effectiveness of the latter in representing time-series data: even though connected scatterplots are novel to many users, they are excellent at being intuitive to understand and capturing and holding the viewer's attention.

Linked highlighting There are a number of interaction approaches to linked highlighting in scatter plots [23]. Through our consultation with the stakeholder, linked highlighting on hover was emphasized as a critical element. However, this not the only means of linked highlighting available. For example, linked brushing, where the user selects a subset of points to be highlighted, is a popular choice for multi-selection [23, 36]. Further conversations with our stakeholder as we gain more exposure will allow us to finalize our design decisions.

Hive plot Overviews of the data will make use of the work done by Krzywinski et al. [5] and Hanson [9] for our proposed hive plot. Various good examples of hive plot visualizations of networks are shown on the hive plot website [29]. We will be developing our visualization based on the features that are most meaningful for our data and the questions we are trying to answer with this view (sparsity of data, node degree, node centrality and reachability, outliers and general trends in the data).

Colour We have decided to use colour as a channel for linked-highlighting. Previous research suggests that colour is one of the most powerful visualization channels [20, 8] and, when used correctly, it can provide insight into the data so intuitively that the user wouldnât even need a legend to understand what kind of data the channel encodes for.

Data and Task Abstractions

```
**you should analyze your domain problem according to the framework of the book, translating from domai
typically data will need to come first, since you will need to refer to that data in your task descripti
it is very likely that you will need to first have domain-specific descriptions, followed by the abstra
it is often helpful to split these sections into two pieces: first have subsection where data is descril
similarly, it can be easier to write the tasks section by first providing a domain-specific list of tasl
you should decide whether to split by domain vs abstract (first have both data and task domain-specific
```

Solution

```
**describe your solution idiom, analyze it according to the framework of the book, and justify your desi
if you have done any significant algorithmic work, discuss the algorithm and data structures
you might choose to split out Interface into its own section**
```

Implementation

```
**medium-level implementation description. you must include specifics of what you did yourself versus wh
```

We will likely take advantage of the work done in interactive visualizations such as D3.js [4] and AngularJS [14] or RShiny [15]. Although D3.js and Angular.js are arguably more powerful in terms of possible customizations and optimizations than RShiny, both of them would require a steep learning curve and time input we do not have. In the interest of time and given the teamâs expertise in R, we are leaning towards using RShiny and will most likely take advantage of previous work available via github and the Shiny website. For example, we have found solutions to set up multiple plots [10], use linked views [11], filter data [12], and rearrange plot layout [13]. The hive plot library HiveR.R developed by Hanson [9] provides us with the tools we need to create an interactive representation of the whole dataset.

Results

```
**should include scenarios of use and multiple screenshots of your software in action. walk the reader t
if you did any evaluation (deployment to target users, computational benchmarks), do report on that here
```

Discussion and Future Work

```
**Strengths, weaknesses, limitations (reflect on your approach)
Lessons learned (what do you know now that you didn't when you started?)
Future work (what would you do if you had more time?)**
```

Conclusions

```
**summarize what you've done in a way that's different from the abstract because you can count on the r
```

Bibliography

```
**make sure to use real references for any work that's been published academically, not just URLs
do pay particular attention to my instructions for checking reference consistency**
```


1. Tripathy, Shreejoy J., et al. "NeuroElectro: a window to the world's neuron electrophysiology data." Frontiers in neuroinformatics 8 (2014).
2. Van Eck, Nees Jan, and Ludo Waltman. "Text mining and visualization using VOSviewer." arXiv preprint arXiv:1109.2058 (2011).
3. Hinrichs, Uta, et al. "Trading Consequences: A Case Study of Combining Text Mining and Visualization to Facilitate Document Exploration." Digital Scholarship in the Humanities (2015).
4. Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer. "DÂ³ data-driven documents." Visualization and Computer Graphics, IEEE Transactions on 17.12 (2011): 2301-2309.
5. Krzywinski, Martin, et al. "Hive plotsâ"rational approach to visualizing networks." Briefings in bioinformatics 13.5 (2012): 627-644.
6. Rieder, Christian, et al. "Interactive visualization of multimodal volume data for neurosurgical tumor treatment." Computer Graphics Forum. Vol. 27. No. 3. Blackwell Publishing Ltd, (2008).
7. Steve Haroz, Robert Kosara, Steven L. Franconeri. âThe Connected Scatterplot for Presenting Paired Time Series.â Transactions on Visualization and Computer Graphics, (2016): https://research.tableau.com/sites/default/files/Haroz-TVCG-2016.pdf
8. Vidya Setlur, Maureen C. Stone. âA Linguistic Approach to Categorical Color Assignment for Data Visualizationâ The IEEE Information Visualization Conference (Chicago, October 25-30, 2015): https://research.tableau.com/sites/default/files/setlurstoneinfovis2015_2.pdf
9. Bryan A. Hanson. âHiveR: 2D and 3D Hive Plots for R.â (2015): academic.depauw.edu/~hanson/HiveR/HiveR.html
10. https://gist.github.com/wch/5436415
11. http://shiny.rstudio.com/gallery/plot-interaction-zoom.html
12. http://shiny.rstudio.com/gallery/basic-datatable.html
13. http://shiny.rstudio.com/articles/layout-guide.html
14. https://material.angularjs.org/latest/
15. http://shiny.rstudio.com/
16. Michael Friendly, Daniel Denis. âThe early origins and development of the scatterplotâ Journal of the History of the Behavioral Sciences (2005): Vol. 41(2), 103â"130
17. http://www.tableau.com/
18. http://www.tableau.com/gartner-magic-quadrant-2015
19. http://datablick.com/2015/04/13/circular-and-hive-plot-network-graphing-in-tableau-by-chris-demartini/
20. Tamara Munzner. âVisualization Analysis and Designâ (2014) A K Peters/CRC Press, Print ISBN: 978-1-4665-0891-0, eBook ISBN: 978-1-4665-0893-4
21. http://www.uxmatters.com/mt/archives/2009/09/best-practices-for-designing-faceted-search-filters.php/
22. http://www.uxforthemasses.com/filter-bars/
23. https://www.perceptualedge.com/articles/visual_business_intelligence/coordinated_highlighting_in_context.pdf
24. Bearavolu, Ratna, et al. "A visualization tool for situational awareness of tactical and strategic security events on large and complex computer networks." Military Communications Conference, 2003. MILCOM'03. 2003 IEEE. Vol. 2. IEEE, 2003.,

25. Dumas, Maxime, Jean-Marc Robert, and Michael J. McGuffin. “Alertwheel: radial bipartite graph visualization applied to intrusion detection system alerts.”Network, IEEE 26.6 (2012): 12-18.
26. French, Leon, and Paul Pavlidis. “Informatics in neuroscience.” Briefings in bioinformatics 8.6 (2007): 446-456.
27. Rebholz-Schuhmann, Dietrich, Anika Oellrich, and Robert Hoehndorf. “Text-mining solutions for biomedical research: enabling integrative biology.” Nature Reviews Genetics 13.12 (2012): 829-839.
28. Sopan, Awalin, et al. “Community Health Map: A geospatial and multivariate data visualization tool for public health datasets.” Government Information Quarterly 29.2 (2012): 223-234.
29. hive plot website: www.hiveplot.net
30. Krzywinski M, Birol I, Jones S, Marra M. âHive Plots â Rational Approach to Visualizing Networks.â Briefings in Bioinformatics, doi: 10.1093/bib/bbr069
31. Arc diagrams in R (network visualization): http://gastonsanchez.com/blog/got-plot/how-to/2013/02/02/Arc-Diagrams-in-R-Les-Miserables.html
32. http://help.sap.com/boaolap40
33. https://products.office.com/en-ca/excel
34. Henry N, Fekete JD. Matrixexplorer: a dual-representation system to explore social networks. IEEE Trans Visu Comput Graph (2006). 12(5):677-84
35. Aitaluk M, Sedova M, Ray A, et al. âBiological Networks: visualization and analysis tool for systems biology.â Nucleic Acids Res (2006). 34:W466-71
36. https://github.com/rstudio/ggvis/blob/master/demo/rmarkdown/linked_brush.Rmd
37. Yan KK, Fang G, Bhardwaj N, Alexander RP, Gerstein M. 2010. âComparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks.â Proc Natl Acad Sci USA. 107(20): 9186-9191