Research Statement

Alexander Lex
http://alexander-lex.com

In my research **I develop interactive data analysis methods for experts and scientists**. My goal is to help them gain insights and derive knowledge from big, complex, and heterogeneous data, and to realize the opportunities that come with the increasing amount of data our society generates. While certain analytical questions can or will be solvable through automatic means, I concern myself with the challenges that require human reasoning. Data analysis in support of addressing these challenges requires an **interactive and visual approach that tightly integrates algorithms, statistics, and machine learning**. Such a continuous and iterative interplay between humans and computers is preferable to purely visual or analytical methods as it leverages the domain knowledge of experts, the highly effective human visual system, and the power of computation.

My past research broadly addresses the visualization of heterogeneous and high dimensional data, the visualization of large and complex multivariate networks, the visualization of relationships between multiple views, and the visualization of hidden content. I follow the principle that big datasets are not amenable to "show everything" visualization approaches, but have to be filtered and partitioned into smaller and meaningful subsets in order to be meaningful for a human user.

My research is both concerned with fundamental visual analysis challenges and with domain problems of collaborators. Examples of the former are the techniques for the visualization of multi-attribute rankings, large numbers of sets and their attributes, and my work on rendering perceptually efficient visual links. My domain specific projects often are in molecular biology, a field that is increasingly data-driven and that can greatly benefit from visual exploration methods. I choose problems that are specific and relevant but also generalizable across domains so that the underlying concepts have a broad impact.

Using both approaches I have developed a number of visualization systems. Most of them are part of the **Caleydo biomolecular visualization framework (`http://caleydo.org`), the development of which I co-lead**. Caleydo is used by researchers in academia and industry from all over the world, including scientists from the Harvard Medical School and pharmaceutical companies such as Novartis and Boehringer Ingelheim.

Cancer Subytpe Analysis

In a close collaboration with domain experts at the Harvard Medical school we developed **StratomeX**, a visual analysis technique that enables scientists to efficiently generate and test hypothesis about disease subtypes [6, 10] (`http://stratomex.caleydo.org`). Subtypes are detailed classifications of diseases that where historically treated as one condition. In cancer, for example, molecular biology methods have lead to the discovery of many new subtypes. Knowing about subtypes can lead to improved diagnosis, prognosis, and targeted treatment. The goal of subtype analysis is to find clinically meaningful groups of patients and reason about the underlying causes of their differences. This requires the integration of many datasets and observations, multiple often diverging automatic classifications, and established knowledge from various databases. Figure 1 shows StratomeX during an analysis of kidney cancer. Each column represents a different dataset. Columns are split into subsets of patients and the bands show the correlations between the

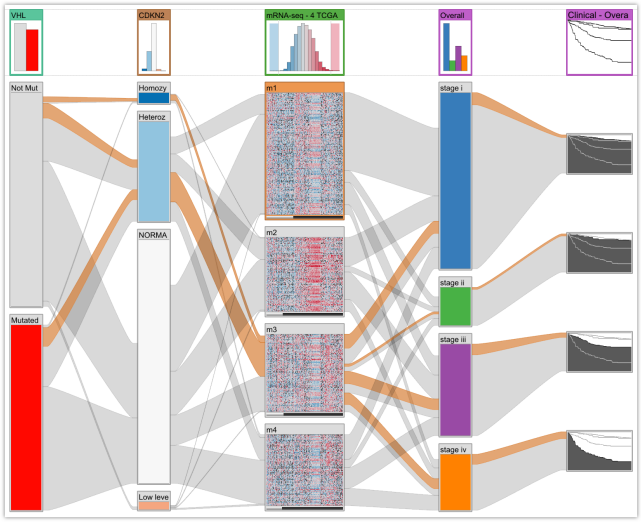

Figure 1: StratomeX showing relationships between subsets of multiple datasets for cancer subtype analysis.

columns. The columns are selected from a vast space of datasets and partitions using a semi-automatic approach that guides analysts to other datasets that support a given observation. Analysts can, for example, query the system for datasets that contain patient clusters similar to another dataset, or that show significant differences in patient survivals.

Our work on cancer subtype visualization has received the **3rd Best Paper Award at EuroVis 2012** and the integrated analysis tool was **published in Nature Methods**. StratomeX was used, for example, in the comprehensive characterization of clear cell renal carcinoma by the Cancer Genome Atlas (TCGA) consortium. Our follow-up work on StratomeX, Domino [2] (`http://domino.caleydo.org`), generalizes the approach, allowing free arrangement and combinations of subsets of datasets. The paper describing Domino won a **Best Paper Honorable Mention at IEEE InfoVis 2014**.

Multi-Attribute Rankings

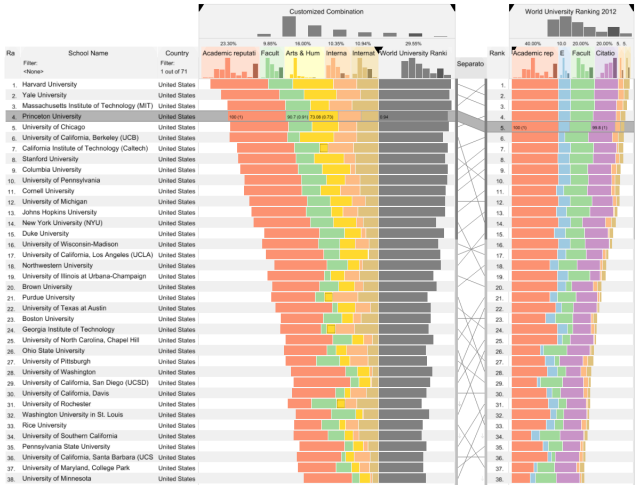

Figure 2: Two alternative rankings of universities based on multiple weighted attributes in LineUp.

Rankings are easy to use and understand and can convey the relative importance of items in a list. They are very popular as they allow us to identify the "best" items, and to easily compare two otherwise potentially complex items. The problem of rankings is that the rank is often determined by subjectively weighted combinations of selected attributes. This is evident from the multiple diverging rankings of universities, for example, which use different methodologies and weights. As it is clear that there is no objective method to compare complex entities such as universities, we need a solution that communicates the composition of the scores used in the rankings and that can be adjusted to reflect the values and preferences of the individual who explores the data. With LineUp [3] (Figure 2, see also `http://lineup.caleydo.org`), our visualization technique, users can clearly see the relative importance of each attribute, can interactively combine multiple attributes, and set custom weights for each of them. Every adjustment is immediately reflected, while animated transitions make changes comprehensible. LineUp can also be used to compare multiple rankings, e.g., at different time points, or to compare different weights and combinations of the same dataset.

While LineUp is a general purpose tool, we also use it as part of larger analysis workflows (e.g., with StratomeX) to rank and select datasets based on the output of various algorithms. The LineUp paper received the **Best Paper Award at IEEE InfoVis 2013** and was widely featured in the press.

Set Visualization

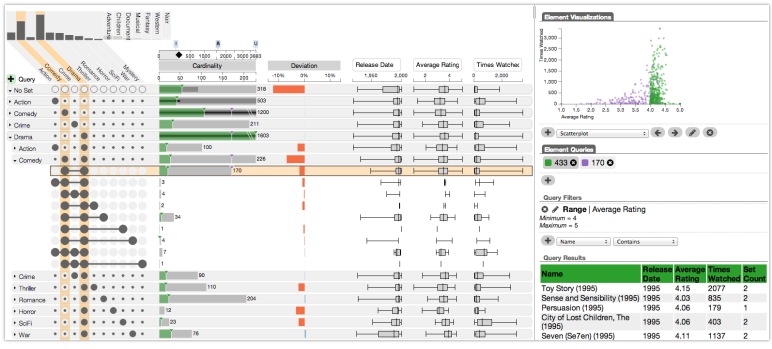

Figure 3: UpSet, a set visualization technique showing how genres co-occur for movies and how these movies differer.

Sets are a common form of data that are used in many analysis scenarios. Examples are the identification of genes shared by multiple species, or to find out which social networking tools are commonly used in combination. Sets can also be used to explore slices of high-dimensional heterogeneous data, i.e., to partition a dataset based on inclusion or exclusion in a set, or based on more complex logical combinations. The challenge of exploring set data is the exponential growth of intersections, which makes traditional approaches such

as Venn diagrams impractical for non-trivial cases. To address these problems we developed UpSet [4] (see Figure 3) an interactive, web-based visualization technique for set data with heterogeneous attributes (live demo at `http://vcg.github.io/upset/`). UpSet is designed so that it only uses the most perceptually efficient visual encoding (position) to represent the underlying data. Analysts can answer specific questions by using a variety of aggregation methods as well as queries based on boolean logic. The data associated with sets and their intersections is visualized in multiple linked views. UpSet can show intersections of dozens of sets and can aggregate these intersections to show all logical set combinations. UpSet has been used, e.g., by economists at the Harvard Kennedy School of Government to analyze trade data, by geneticists at the Harvard Medical School to evaluate variant calling algorithms, and by chemists at Novartis to evaluate the specificity of drugs at inhibiting genes.

Large and Partitioned Networks

While networks are a crucial tool in biology research they are also difficult to understand and use. The challenge stems from both, the size of the networks (i.e., topology related challenges) and the large datasets associated with the nodes and edges (i.e., attribute related challenges). As a remedy to the topology related challenges biologists divide the large overall network into meaningful subnetworks called pathways. These pathways contain only nodes that have a clear association with a specific topic, such as a disease or a metabolic process. One pathway contains, for example, all nodes that are known to be relevant for diabetes. The problem of this approach is that these pathways are (by design) incomplete, yet a deeper analysis often requires connections that were left out. For example, interventions in one part of the network, e.g., through a drug inhibiting a gene, can have effects in many other processes. Thus analysts have to simultaneously explore multiple pathways. Also, comparisons between pathways are a common task to identify, for example, how the molecular processes of two cancers diverge from each other.

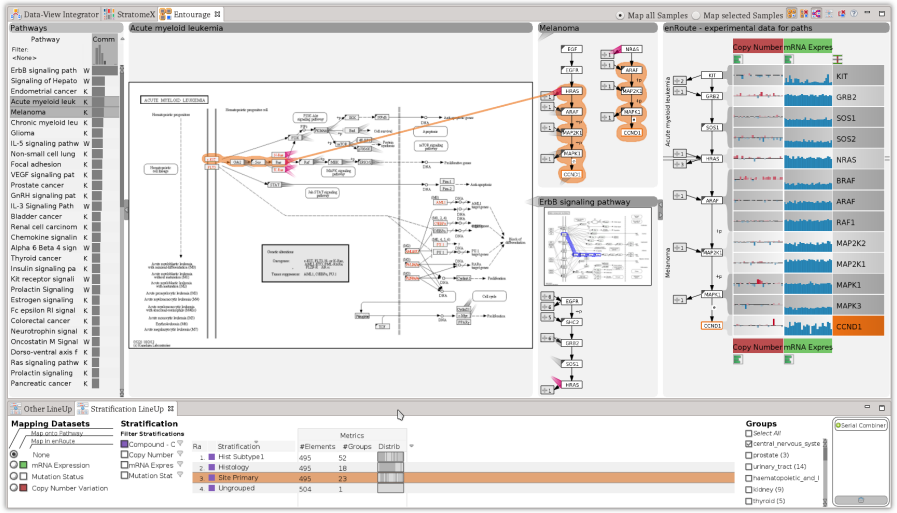

Figure 4: Visualization of connections between multiple biological pathways (Entourage) in the context of heterogeneous experimental datasets (enRoute).

To address these challenges we have worked closely with scientists at the Novartis Institutes for BioMedical Research to develop Entourage, a method that lets users focus on the pathway most relevant to their analysis question yet still conveys the relationships to the larger network [5]. The technique, shown in Figure 4 (`http://entourage.caleydo.org`), displays one focus pathway at full scale, while contextually relevant parts of other pathways are dynamically supplemented. Which contextual information is shown depends on a combination of user selections based on automatic recommendations.

Multivariate Networks

Pathways only represent the state of the art and do not account for variations between patients. Many diseases are caused by a deregulation of a specific path in a pathway, but this information is not contained in the pathways. Associating attributes with the nodes and edges can rectify this situation, as they allow inferring whether a path, for example, is shut off in a group of patients. It is common that each node is associated with hundreds or thousands of measurements from many patients and datasets. Showing both the topology of a network and a large number of node or edge

attributes at the same time is a hard problem. A common approach is to color the nodes with average values or to show multiple instances of the pathway, each with a different set of experimental data mapped to it. However, both approaches do not scale well. An interactive solution to this problem, enRoute (`http://enroute.caleydo.org`), is shown on the right of Figure 4, integrated with Entourage. Instead of attempting to show all attributes for all nodes, analysts can chose a path in a network (highlighted in orange) that is then extracted and displayed in a linear form [7, 8]. The linearization of this path makes it possible to visualize large quantities of attributes from different datasets and from multiple groups for the selected path. Figure 4 shows gene expression and copy number data for cancer cell lines.

Our collaborators at Novartis employ enRoute to understand why certain cell lines show deviating behavior from others when exposed to a compound. This will allow them to identify which drugs are effective and safe for which patients. The paper has won the **Best Paper Award at IEEE BioVis 2012** and we were invited to publish an **extended version in BMC Bioinformatics** [8].

Showing Relationships & Guiding Attention

A central aspect of data analysis is to understand relationships between items that are displayed in multiple views. Most commonly, these relationships are highlighted interactively using color so that analysts can identify cross-links and reason about the connections. An alternative to color highlighting that has gained widespread attention in recent years are visual links, i.e., explicit edges that connect related pieces of information. Visual links have the benefit of being more salient—they stand out—and we have shown that search tasks are completed faster using visual links compared to highlighting with color [9]. A drawback of visual links is that they introduce visual clutter. To remedy this we developed an intelligent routing technique that measures the salient regions of a scene and routes the links around them [9] (Figure 5).

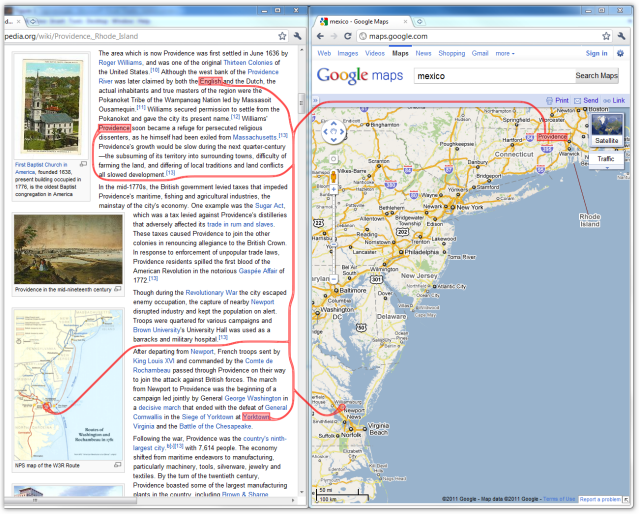

Figure 5: Context-preserving visual links route around important regions in a scene.

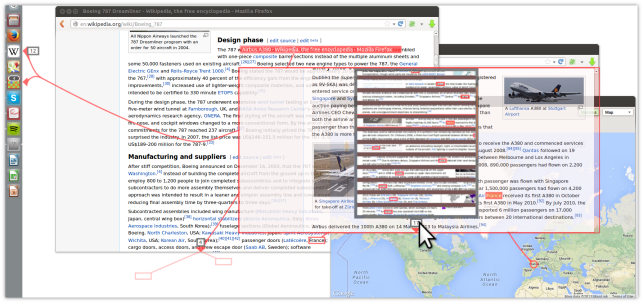

Figure 6: Visualization of hidden content using visual links, see-through interfaces and smart previews.

A related topic is the integration of multiple otherwise independent applications for visualizing relationships. Together with collaborators, we have developed a system [11] that integrates content from multiple independent applications and connects related items with visual links, rendering them on top of the desktop. A recent extension of this work [1], shown in Figure 6, not only connects visible content, but also conveys hidden elements, such as information that is scrolled away, occluded, or in unopened files, by using smart previews, see-through techniques and other visual elements, thus guiding analysts to content that could be easily overlooked. These techniques were validated in formal user studies.

Our work on connecting independent applications [11] has received the **Best Student Paper Award at Graphics Interface 2010**, the routing technique for visual links [9] has won the **Best Paper Award at IEEE InfoVis 2011**, and the work on guidance to hidden content [1] won an **Honorable Mention Award at ACM CHI 2014**.

Research Agenda & Future Work

While there have certainly been great achievements in analysis and visualization of large and heterogeneous data, significant challenges remain. The general strategy of using meaningful data subsets and tightly interweaving interactive visualization with algorithms is essential for addressing big data problems in the future. There are many open research questions in this area that are well suited as individual PhD thesis topics.

Representing Data on Multiple Levels of Detail. As datasets grow bigger, there is a need for multi-scale representations. As we combine multiple datasets and slices thereof we need to represent the underlying data at various levels of detail. Here, semantic aggregation methods that automatically convey the relevant aspects of the data at various scales are important. Also it is essential that these aggregations are not static but react to the context of the analysis. We need to develop such techniques for all major classes of data, including multidimensional datasets and graphs.

Slicing Datasets. An essential component of the subsets approach is the division step and it is crucial that these divisions are meaningful and interpretable. I intend to investigate two approaches to this problem: for black box methods such as clustering, I will investigate how to best visually evaluate subsets (clusters) and communicate the differences between multiple subsets efficiently. An alternative is to use simple, interpretable rules to create the splits in the first place. Here, the challenge is to identify the rules that are interpretable yet yield good results. Employing these methods will lead to better and more trusted analysis results.

Ensuring Coverage. How can we make sure that we see all meaningful features of a large and heterogeneous dataset if we cannot look at everything? This is an important issue for avoiding selection bias, which we would encounter if we only investigate and report data that fits a story. We need to develop visual analysis systems that not only convey data that fits a hypothesis, but also search for and communicate alternative scenarios, a burden which currently is solely the analyst's responsibility.

Visualizing Relationships. The necessary fragmentation and the heterogeneity of multiple large datasets require improved methods for visualizing relationships between disjoint representations of the parts. We need to develop methods that scale, reduce clutter, and are able to convey complex relationships, such as varying relationship types and magnitudes.

High-Impact Domain Problems. While many data analysis methods can be addressed with general purpose methods, there is a certain class of important problems that are highly domain specific and require tailored solutions. An example for such a problem is the representation of data on a genome scale: traditional genome browser use the DNA sequence as its coordinate system, which for many tasks is not the ideal representation as it is oblivious to functional relationships between distant parts. This calls for a new representation that focuses on functional relationships. While such custom solutions are sometimes not easily transferable to other domains, the potential impact is highly rewarding.

Domain Relevance and Reproducibility

When working on visualization and analysis methods for real world problems it is critical to ensure that the resulting systems also solve the domain analysis problem, which translates into a strong requirement for high-quality software development. I strive to not only produce visualization research but also to provide high-impact analysis tools that are useful to a large community of domain scientists. My commitment to maintaining, documenting and improving software systems such as Caleydo clearly demonstrates this.

I take reproducibility of my research very seriously. To achieve this I always demonstrate my techniques with publicly available data, even if the original datasets of collaborators are proprietary. To ensure reproducibility I also make all my source code available to the public under a permissive open source license.

References

[1] T. Geymayer, M. Steinberger, A. Lex, M. Streit, and D. Schmalstieg. Show me the invisible: Visualizing hidden content. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, pp. 3705–3714. ACM Press, 2014. ISBN 978-1-4503-2473-1. doi:10.1145/2556288.2557032.

[2] S. Gratzl, N. Gehlenborg, A. Lex, H. Pfister, and M. Streit. Domino: Extracting, comparing, and manipulating subsets across multiple tabular datasets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '14)*, 2014. To appear.

[3] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. LineUp: Visual analysis of multi-attribute rankings. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '13)*, vol. 19, no. 12, pp. 2277–2286, 2013. doi:10.1109/TVCG.2013.173.

[4] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '14)*, 2014. To appear.

[5] A. Lex, C. Partl, D. Kalkofen, M. Streit, S. Gratzl, A. M. Wasserman, D. Schmalstieg, and H. Pfister. Entourage: Visualizing relationships between biological pathways using contextual subsets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '13)*, vol. 19, no. 12, pp. 2536–2545, 2013. doi:10.1109/TVCG.2013.154.

[6] A. Lex, M. Streit, H.-J. Schulz, C. Partl, D. Schmalstieg, P. J. Park, and N. Gehlenborg. StratomeX: Visual analysis of large-scale heterogeneous genomics data for cancer subtype characterization. *Computer Graphics Forum (EuroVis '12)*, vol. 31, no. 3, pp. 1175–1184, 2012. doi:10.1111/j.1467-8659.2012.03110.x.

[7] C. Partl, A. Lex, M. Streit, D. Kalkofen, K. Kashofer, and D. Schmalstieg. enRoute: Dynamic path extraction from biological pathway maps for in-depth experimental data analysis. In *Proceedings of the IEEE Symposium on Biological Data Visualization (BioVis '12)*, pp. 107–114, 2012. doi:10.1109/BioVis.2012.6378600.

[8] C. Partl, A. Lex, M. Streit, D. Kalkofen, K. Kashofer, and D. Schmalstieg. enRoute: Dynamic path extraction from biological pathway maps for exploring heterogeneous experimental datasets. *BMC Bioinformatics*, vol. 14, no. Suppl 19, p. S3, 2013. doi:10.1186/1471-2105-14-S19-S3.

[9] M. Steinberger, M. Waldner, M. Streit, A. Lex, and D. Schmalstieg. Context-preserving visual links. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '11)*, vol. 17, no. 12, pp. 2249–2258, 2011. doi:10.1109/TVCG.2011.183.

[10] M. Streit*, A. Lex*, S. Gratzl, C. Partl, D. Schmalstieg, H. Pfister, P. J. Park, and N. Gehlenborg. Guided visual exploration of genomic stratifications in cancer. *Nature Methods*, vol. 11, no. 9, pp. 884–885, 2014. doi:10.1038/nmeth.3088. *equal contribution.

[11] M. Waldner, W. Puff, A. Lex, M. Streit, and D. Schmalstieg. Visual links across applications. In *Proceedings of the Conference on Graphics Interface (GI '10)*, pp. 129–136. Canadian Human-Computer Communications Society, 2010. ISBN 1568817125.