Half-Day Tutorial: Visual Analytics for High-Dimensional Data

Instructors: Klaus Mueller and Shenghui Cheng (Computer Science Department, Stony Brook University)

Purpose:
Analyzing high-dimensional data and finding hidden patterns in them is a difficult problem and has attracted numerous research efforts in the visualization community and beyond. Gaining insight into high dimensional data is at the core of big data analysis and data science. Automated methods can be useful to some extent but bringing the data analyst into the loop via interactive visual tools can help the discovery process tremendously. All of these visual tools use some kind of projection strategy to convey the high dimensional space within the confines of the two screen dimensions. Since this projection is an inherently ill-posed problem in all but the most trivial cases, all methods will bear certain trade-offs. Knowing the strengths and weaknesses of the various paradigms existing in the field can inform the design of the most appropriate visualization strategy for the task at hand. It can help practitioners in selecting the best among the many tools available, and it can help researchers in devising new tools to advance the state of the art. This tutorial aims to serve both of these factions of the visualization community.

Level: Beginner

Syllabus:
We will begin the course with theoretical aspects of high dimensional data spaces and then delve into the various existing visualization paradigms and present examples for each, complemented by live demos. A related field is data mining and course attendees will also gain some knowledge in this area when taking this tutorial.

Section 1 – the basics (1 hour):
This section will endow attendees with the fundamental knowledge needed to follow through later sections. It will begin with some motivating examples of high-dimensional data, analysis tasks, and real-world applications. Later sections will refer to these examples to put the new material in context. The course will then enumerate common data types occurring in high-dimensional data spaces, such as numerical, categorical, ordinal, nominal, etc., which might require special treatment in certain visualization schemes. Since it is often a mystery to people how all this relates to non-numerical data, such as images, video, text, and the like we will present standard mechanisms and feature transforms that can convert these data into vectors of numerical data suitable for visual high-dimensional data exploration. A very fundamental concept is the curse of dimensionality. Presenting it in detail will give attendees a feel what types of schemes developed for low-dimensional data are feasible and which ones are hopelessly infeasible due to the vastness of the space and their explosive polynomial complexity. The curse of dimensionality is intimately connected to the distance between data objects, and henceforth we will end this section with a discussion of distance and similarity metrics, such as Euclidian, cosine, correlation, structural, geodesic, Mahanalobis, the Gabriel graph, and others. Distance metrics are an integral component of many visualization schemes and form essential basic knowledge.

Section 2 – data processing and wrangling (1 hour):
In practice, more often than not, the data are too large to be visualized directly and need to be decimated, either preprocessing or adaptively during the visualization, within a detail-on-demand framework. We will discuss available techniques for data reduction by sampling and discuss their tradeoffs, such as random, stratified, well-scattered point selection, and others, with a focus on detail

and outlier preservation. All these measure reduce the number of data points that need to be carried through subsequent tasks, such as clustering, and prevent possible cluttering in the final display. Decimating the number of data points can also be decisive in making an application interactive. Next, we will present methods for dimension reduction and attribute selection, describing their strengths and tradeoffs, such as PCA, LDA, MCA, factor analysis, entropy, and others, emphasizing that reducing the number of dimensions make distance calculations more efficient and also helps reduce the complexity of several visualization methods discussed in section 3. Important data wrangling steps are also noise removal, outlier analysis, and coping with missing values for which we will present established techniques from data mining. We will end this section with a discussion on data clustering and their tradeoffs. We will begin with the use of data clustering as a means for data redundancy removal, where clustering is used with tight proximity bounds or with entropy measures, mostly in the form of k-means or k-medoids. Then we will generalize k-means to EM using probabilistic boundaries, DBSCAN, CURE, TaxMap, spectral, and others. We will stress their abilities to preserve shape and outliers.

Section 3 – visualization and interaction (1.5 hours):
Having established the fundamental knowledge on high dimensional data spaces, the course attendees will now be able to recognize the strengths and weaknesses of the various visualization techniques available, appreciate the motivation that led to their development, and their application domains. We will differentiate among the major classes of visualization schemes. We will begin with bivariate scatterplots and extend them to small multiples and scatterplot matrices (SPLOMs). We will point out that these give access to (piecewise) bivariate relationships and describe situations in which one might want to see relations of more than two variables in a joint display, as opposed to a disjoint SPLOM. This will motivate a discussion of generalized scatterplots and biplots, which address this need. We will discuss projection pursuit and the Grand Tour, and we will describe effective view quality metrics (also applicable to SPLOMs), that can help the system to select those projections among the vast number possible that can discern interesting data phenomena, such as clusters, spread, and others.

Following we will show how scatterplots can be transformed into parallel coordinates and star plots. We will discuss the important topics of interaction, abstraction, as well as dimension ordering. For the latter we will discuss mechanisms available to assist users in determining good dimension orderings, such as the traveling salesman paths across a correlation map. An important class of visualization schemes is that of interior displays, such as Radvis, star plots and generalized barycentric coordinates. We will discuss both their relations and differences to parallel coordinates as well as scatterplots and the associated tradeoffs.

The methods presented so far are all based on linear projections. Next, we will expose the attendees to methods that are all based on constraint optimization, leading to non-linear projections or space-warping. We will motivate why this can be advantageous, such as overcoming the problem of projection ambiguities that plague the linear displays, and we will highlights what the tradeoffs are, such as the loss of the context to the data attributes, at least partially. We will discuss the most popular methods of low-dimensional embeddings such as MDS (simple and pivot-based), t-SNE, isomap, LLE, and SOM, and we will discuss their strengths and weaknesses.

An important development in recent years is also the visualization of high-dimensional data by decomposing them into salient subspaces. We will discuss the available techniques, such as our own TripAdvisorND framework. We will also discuss how these subspaces can be managed in exploration tasks, for example by ways of maps In addition, methods based on a factorization of the data matrix (SVD, NNMF, and others) are also important to mention in this context. We will close this section by discussing feature-based visualizations, such as correlation networks and rank by feature methods.

The course will build and expand on material the first instructor has developed for his annual graduate and undergraduate visualization courses, taught every year over the past decade. The present graduate course can be accessed here: http://www3.cs.stonybrook.edu/~mueller/teaching/cse564. In addition, the first instructor has also recently developed and taught a course on data science fundamentals which can be accessed here http://www3.cs.stonybrook.edu/~mueller/teaching/cse590_dataScience/.

Instructors:

Klaus Mueller received a PhD in computer science from the Ohio State University. He is currently a professor in the Computer Science Department at Stony Brook University and is also an adjunct scientist in the Computational Science Initiative at Brookhaven National Labs. He is the director of the Visual Analytics and Imaging (VAI) lab at Stony Brook University, a lab with more than 10 PhD students, and has graduated 14 PhD students over the past 16 years. His current research interests are visualization, visual analytics, data science, medical imaging, and high-performance computing, He won the US National Science Foundation CAREER award in 2001 and the SUNY Chancellor Award for Excellence in Scholarship and Creative Activity in 2011. Mueller has authored more than 170 peer-reviewed journal and conference papers (more than 30 specifically on visual high-dimensional data exploration), which have been cited more than 6,500 times. The papers on visual high-dimensional data exploration have addressed all topics discussed in this tutorial, such as distance metrics, generalized projections, optimized layouts, subspace analysis, feature-extraction from text and images, correlation networks, radial displays, evaluation, and model-building, in the context of real-world applications. He is a frequent speaker at international conferences and he has participated in more than 20 tutorials on various topics, many of them as lead organizer. One of these was on "Visual Medicine" which he taught 8 times with various co-instructors at IEEE VIS, Eurographics, and MICCAI. He has also been teaching a half-day course on High-Performance Computing for Medical Imaging for the past 8 years at the annual SPIE Medical Imaging conference. Mueller has chaired a number of conferences, among these IEEE VIS 2009 in Atlantic City, and he was until recently the chair of the IEEE Technical Committee on Visualization and Computer Graphics. He is also back on the editorial board of IEEE Transactions on Visualization and Computer Graphics and he is a senior member of the IEEE. For more information, please see http://www.cs.sunysb.edu/~mueller He can be reached at mueller@cs.stonybrook.edu

Shenghui Cheng is a PhD candidate at Visual Analytics and Imaging (VAI) Lab, Computer Science Department, Stony Brook University. Over the past 5 years he has developed a number of interactive high-dimensional visualization system, some of them in frequent use at collaborating institutes in the US, Asia, and Europe, where he was a visiting research scientist at several occasions. His primary research interest includes visual analytics, information visualization and scientific visualization with a special focus on high-dimensional and multivariate data. Specifically, he worked on high dimensional data layouts (embeddings), time-series data visual analytics, geospatial data visual analytics, social and supercomputing network/graph visual analytics, bioinformatics etc. Notable is his recent paper at VAST 2015 on data context maps that was also published in IEEE TVCG, special issue on VIS. The data context builds on a deep understanding of the intricacies of high-dimensional data spaces. It allows users to simultaneously appreciate (1) the similarity of data objects, (2) the similarity of attributes in the specific scope of the collection of data objects, and (3) the relationships of data objects with attributes and vice versa. The contextual layout also allows data regions to be segmented and labeled based on the locations of the attributes. This enables, for example, the map's application in selection tasks where users seek to identify one or more data objects that best fit a certain configuration of factors, using the map to visually balance the tradeoffs. For more information on Shenghui's diverse research projects see http://www3.cs.stonybrook.edu/~shecheng/ Shenghui is a student member of the IEEE and can be reached at shenghui.cheng@stonybrook.edu