An important goal of my research is to design novel techniques and systems to support data exploration. Since real-world data is often complex, uncertain, and heterogeneous, to analyze and derive insights from these data requires extensive expertise in many disciplines. Thus, an important theme in my work has been on the integration of visualization-based approaches with techniques and algorithms from related areas such as machine learning, topology, and text processing, to improve and guide the data analysis process. The ever-increasing scale and complexity of data creates additional challenges. Data sets can not only be large, but analyses often require data from different sources and in different formats.

To overcome these challenges, I have (i) devised novel visual encodings and visual analytics approaches to help domain experts generate insights out of complex data, and (ii) designed efficient algorithms that cut across multiple areas, such as, machine learning, topology and optimization, and aim to attain balance between the available resources and the ability to interactive analyze the data.

During my doctoral studies, I was involved in two main projects: Multifaceted Climate Data Visualization, where we collaborated with climate scientists to develop data analytics techniques and visual representation to describe climate models; and Spatio-Temporal Urban Data Visualization, where we (1) proposed a new visual model to query origin-destination data and an efficient index structure to evaluate these queries at interactive rates, and (2) used techniques from vector field visualization to identify mobility patterns using New York City taxi data. During my masters, I worked on problems where I used multidimensional projection to explore High-Dimensional Data and proposed novel algorithms to accelerate the projections and allow users’ feedback to re-project the data interactively. These projects led not only to 15 publications but also to real systems that are currently being used by domain experts.

1 Multifaceted Climate Data Visualization
Gauging consensus among predictions and outputs of multiple simulation models is a critical problem for understanding global climate change patterns. This requires similarity analysis of climate models which typically involve multiple data facets, including space, time, input parameters, and output variables. Model inter-comparison enables scientists to explore and develop different hypotheses about ecosystem processes and climate change indicators. The main contributions of this two year long collaboration with climate scientists\(^1\) are summarized below.

**Design Space Analysis.** Our first step in this project was to understand the state-of-the-art (static) visualizations techniques that climate scientists use. In [1] we describe the results of an exploratory study of climate data visualizations conducted in collaboration with climate scientists. The outcome of the study was a scheme that categorizes the design problems in the form of a descriptive taxonomy.

During our interactions with the scientists we found that the Rainbow colormap, despite ample research advocating against its use, is considered the de-facto standard for climate model comparison. To study this phenomenon we set up a user study aimed at assessing and comparing their objective performance to their subjective assessments while using the Rainbow colormap and other perceptually balanced colormaps. We provided evidence of the effectiveness of perceptually balanced color maps in climate science. In addition, through our study, we also captured instances of mismatch between subjective and objective perception in visualization [2].

---

\(^1\)Multi-scale Synthesis and Terrestrial Model Intercomparison Project: [http://nacp.ornl.gov/MsTMIP.shtml](http://nacp.ornl.gov/MsTMIP.shtml)
Visual Exploration. Reflecting on the inadequacies of the static visualization approaches, we introduced the scientists to the Ultrascale Visualization Climate Data Analysis Tools (UV-CDAT) [3, 4]. UV-CDAT is a workflow-based, provenance-enabled system that integrates climate data analysis libraries and visualization tools in an end-to-end application. We contributed to this project by allowing scientists to assemble and execute complex workflows that consist of dataset selection, a specification of a series of operations that must be applied to the data, and the creation of appropriate visual representations, before finally viewing and analyzing the results. All these additional functionalities provided to UV-CDAT new capabilities for model inter-comparison. Scientists found this tool very useful, especially for its provenance support, which aid them in the reproducibility of their results (See Figure 1(a)).

Multifaceted data analysis is inherently challenging on two counts: i) preserving the mental model about the different facets necessitates an encoding strategy that preserves visual symmetry, and ii) exploring these facets at multiple levels of granularity and understanding their relationships necessitates a systematic interaction strategy. Existing visualization tools are only capable of integrating one or two facets. To fill this gap, we developed the SimilarityExplorer tool [5] which enables multi-faceted visual analysis of climate models. Using our tool, climate scientists were able to get an overview of model similarity across space and time, and then drill down to further explore where, when, and by how much models were similar or different (See Figure 1(b)).

Visual Analytics Approaches. We then pursued two threads of research.

Exploring the inter-relationship among groups defined by alternative similarity criteria is a challenging problem. We derived this problem from our work on model comparison in climate science where climate modelers are faced with the challenge of making sense of alternative ways to describe their models: one through the output they generate, another through the large set of properties that describe them. Ideally, they want to understand whether groups of models with similar spatio-temporal behaviors share similar sets of criteria or, conversely, whether similar criteria lead to similar behaviors. To tackle this problem, we proposed a visual analytics paradigm: visual reconciliation [6], which is an iterative, human-in-the-loop refinement strategy for reconciling alternative similarity spaces. The reconciliation technique involves synergy among computational methods, adaptive visual representations, and a flexible interaction model, for communicating the relationships among the similarity spaces. While iterative refinement strategies are not new in visual analytics, sense-making of diverse characterization of data spaces is still an emerging area of research. The strength of the reconciliation model stems from transparency in presentation and communication of the similarity relationships among diverse data descriptors, with minimal abstraction, and effective visual guidance through visual cues and direct manipulation of the data.

Second, Species Distribution Models (SDM) which are high dimensional functions that combine observations of species occurrence or abundance with environmental estimates. They are used to gain ecological and evolutionary insights and to predict distributions across landscapes. The predictions from the different

---

2UV-CDAT: http://uvcdat.llnl.gov/
SDMs can vary substantially even when created with the same observation data. Such discrepancies not only highlight the uncertainties associated with the predictions, but also hinders the analyses of these predictions by ecologists and environmental decision makers. Studying the topology of high dimensional data sets helps better understand and visualize the data. In this work, we proposed a topology-based framework to help study the differences in various SDMs directly in the high dimensional domain. The topology of the differences between a given pair of SDMs was first used to identify the disparity between the two functions. Then, a visualization interface was designed to explore these differences at multiple scales. This is an ongoing project, we intend to present our results on IEEE VIS 2015.

2 Spatio-Temporal Urban Data Visualization

Data captured in urban environments (e.g., taxi, subway, energy consumption, weather) provide valuable information about the behavior of many components of a city. The analysis of such data has the potential to derive knowledge that can be used to make cities more efficient, inform policies and planning decisions, and improve the lives of citizens. Over the past two years, we have worked on new methods and tools to support the interactive exploration of spatio-temporal data.

Data Exploration. Analyzing spatio-temporal data presents many challenges. The data are complex, containing geographical and temporal components in addition to multiple variables associated with these components. Consider for example the New York City taxi data collected by the Taxi & Limousine Commission (TLC). Each taxi trip consists of two spatial attributes (pickup and dropoff locations), two temporal attributes (pickup and dropoff times), and other attributes including taxi id, distance traveled, fare and tip amount. Formulating queries against these data and comparing multiple spatio-temporal slices is challenging. This problem is compounded due to the size of the data: every day, there are over 500,000 taxi trips, totalling over 170 million trips in a year.

We proposed (i) a new visual query model that supports complex spatio-temporal queries over origin-destination (OD) data [7]. We showed that this model is able to express a wide range of spatio-temporal queries, and it is also flexible in that not only can queries be composed but also different aggregations and visual representations can be applied, allowing users to explore and compare results. (ii) We built a scalable system TaxiVis⁴ (See Figure 2(a)) that implements this model and supports interactive response times through a new spatio-temporal index structure; makes use of an adaptive level-of-detail rendering strategy to generate clutter-free visualization for large results; and shows hidden details to the users in a summary through the use of overlay heat maps.

TaxiVis is currently being used at the NYC Department of Transportation (DoT) and the TLC. We should note that while the original motivation to build TaxiVis was to analyze taxi data, we have used the system to explore other spatio-temporal data sets, including: NYC CitiBike, property ownership patterns, 311 complaints, geo-tagged tweets, and energy consumption.

TrafficFlow. Given the high penetration rate of taxis in large cities, it is reasonable to assume that taxis can be used as probe vehicles, and taxi movement and travel times are representative of the overall traffic and provides a broad coverage of the city in space and time. Our goal in this project was to support interactive analysis and visualization of traffic mobility dynamics. The first hurdle we had to address was missing data: while the TLC data set contains the start and end points for trips, it does not have information about the route. We proposed a scalable model that takes as input data about taxi trips and derives the required traffic information [8]. The model relies on an efficient data-driven closest path algorithm to compute plausible routes for trips, which are subsequently used to infer traffic speed. We validated our model against speed values obtained from EZ-pass tag readers, and the results showed that our model derives accurate predictions, outperforming existing models.

Exploring traffic mobility dynamics through visualization is also a challenging task. Previous works had predominantly focused on the global movement of objects, and proposed techniques for visualizing trajectories either directly on a map, or through the use of specialized metaphors like the space-time cube and flow map. While these techniques are effective for analyzing object movement, they do not capture

⁴TaxiVis: http://taxivis.org/
the local mobility dynamics resultant from their collective motion. Furthermore, in the presence of a large number of objects, such visualizations quickly get cluttered. Vector field visualization techniques, on the other hand, are capable of visualizing localized flow and have been effectively used in various applications in scientific visualization and computer graphics. In order to support interactive exploration of the traffic data, the predicted traffic speed for all roads was then used to derive a traffic function, which corresponds to a time-varying probabilistic vector-valued function defined on the graph representing the road network. We adapted two well-known vector field visualization techniques, particle advection techniques and global techniques, to visualize local traffic flow and to provide global traffic overviews, respectively (See Figure 2(b)).

3 High-Dimensional Data Visualization

Document and image collections, time series, or multiple scalar fields related to a single phenomenon are just a few examples of high-dimensional data. Creating visual representations that provide insight into the overall behavior of such data is challenging. Multidimensional projection techniques offer a unifying framework in this scenario, by mapping data to a low-dimensional visual space (i.e., 2D or 3D) suitable for user interaction.

**PLP.** However, existing projection methods are either computationally expensive or not flexible enough to enable interactive data manipulation. We proposed the Piecewise Laplacian-based Projection (PLP) [9], a multidimensional projection technique. PLP has a local character that renders it more versatile than other projection schemes. Additionally, we presented a mechanism for locally changing the projection, in accordance with user interaction, in such a way that the mapping itself adapts to user manipulations of the layout during visual exploration. This mechanism can be combined with the local nature of PLP so as to allow for drastic changes in the projection map, enabling the exploration and organization of the data in a flexible and dynamic way. Figure 3(a) shows an application of PLP creating music playlists.

**3D LSP.** While 2D maps afford easy interaction, 3D projections decrease information loss allowing for better group discrimination. However, interacting in 3D in everyday applications is more difficult. We proposed a framework for projecting high-dimensional data to 3D visual spaces [10] based on a generalization of the Least-Square Projection (LSP). We investigated its effectiveness using quantitative measures and a user study (See Figure 3(b)). For the quantitative evaluation, we apply the similarity metrics of neighborhood hits and neighborhood preservations, which confirm the intuition that 3D projections outperform 2D in terms of precision. The user study compares the suitability of 2D and 3D projections to perform analysis tasks. We found that better separation in 3D projections allowed for a more precise analysis result.
Fiber tracks exploration. Diffusion Tensor Imaging (DTI) quantifies, the diffusion of the movement of molecules in biological tissues caused by intern thermal energy. An example is the brain white matter. In the white matter the water molecules move faster in the direction of the length of the neuron cells, composing the brain fibers. A common approach to visualize DTI data is to convert tensors into vector and show pathways. However, exploration in visual space remains a challenging task due to the large number of derived pathways and the complexity of fiber space geometry. We presented a strategy for the exploration of DTI datasets [11], employing spatial and curvature features to define feature spaces, and Local Affine Multidimensional Projection (LAMP) technique to represent the feature space in a projected space. Additionally, we allowed the user to organize groups of fibers based on projected feature space and relate that to his or her findings, providing, therefore, a novel, fast, and intuitive way for the user to control possible limitations of the projection process and of the feature space (See Figure 3(c)).

4 Future Work

To leverage the growing volumes of data being collected, e.g., urban, social, scientific, or commercial, we need to develop methods and tools that draw from multiple disciplines to allow domain experts to analyze their data. The research projects I pursued during my PhD are good examples, and they also showed that multi-disciplinary efforts are not only a great source of open problems for computer science, but they can also lead to substantial practical impact. After I graduate, I would like to continue to work on the (i) design of algorithms, methods and infrastructure to support the analysis of massive data sets, (ii) design of new visual encodings and interaction modes that are effective in the exploration process.

An interesting example for (i) would be the domain of web-analytics. Some of the techniques used during my research such as multidimensional projections or topological algorithms can be directly applied to this domain. Web data can be considerate as points in a high-dimensional space which contains some connections among them. We can embed those points in a visual space to visualize and explore, but it would also be interesting to recover a topological network. Additionally we can exploit the temporal concept on this data. As we can see there are multiple ways to tackle this problem, but it requires multiple areas of expertise.

Additionally, I look forward to continuing interdisciplinary collaborations. I believe as computer scientists it is important for us to interact with different domain experts for gaining knowledge about different problems and for understanding how to better apply the techniques we develop in diverse analysis scenarios. Such interactions can not only help drive new research directions and opportunities, but also create real-world impact.
References


