Dagstuhl Seminar 17461: Connecting Visualization and Data Management Research

Participant list

http://www.dagstuhl.de/program/calendar/partlist/?semnr=17461&SUOG

Daily Schedule

Monday	
9:00AM-10:00AM	Intro: Remco Chang (5 minutes) General introduction: 30s - 1 minute each (30-45 mins.) Speakers: - **Danyel Fisher** Note-taker: Juliana Freire
10:15AM-12:15PM	Speakers: - **Eugene Wu** - **Jason Dykes** Note-taker: Carlos Scheidegger
2:00-3:30PM	Speakers: - **Alexandra Meliou** - **Sihem Amer-Yahia** - **Leilani Battle** Note-taker: Jean-Daniel Fekete
4:00-5:30PM	- **Surajit Chaudhari** - **Tim Kraska** Discussion, emerging topics for rest of week Note-taker: Remco Chang

Tuesday	
9:00-10:00AM	Intro: Carlos Scheidegger (5 minutes) Speakers: - **Arnab Nandi** - **Richard Wesley** Note-taker: Juliana Freire
10:15AM-12:15PM	Speakers: - **Harish Doraswaimy** - **Carsten Binnig** - **Michael Gleicher** Note-taker: Carlos Scheidegger
2:00-3:30PM	Volker Markl, 15 mins, report and next steps Lightning Talks Hannes Mühleisen - Frictionless data movement for analytics Themis Palpanas - Visualizations for Analytics on Very Large Data Series Collections Zhicheng Liu - Event Sequence Data Queries and Visualization Note-taker: Jean-Daniel Fekete
4:00-5:30PM	Wesley Willett - Dominik Moritz - Grammars for Interactive Visualizations Eugene Wu - perceptual functions?? Jean-Daniel Fekete - ProgressiVis Michael Sedlmair - A bit more on human factors, and why they might matter Lightning Talks Discussion, emerging topics for rest of week Note-taker: Remco
Wednesday	
9:00AM-12:15PM (15-min break @10AM)	Intro: Juliana Freire (5 minutes) Breakout sessions and discussion
	Note-takers: distributed per group
1:30PM-6:00PM	**Excursion: Trier**

Thursday	
9:00AM-12:15PM (15-min break @10AM)	Intro: Jean-Daniel Fekete (5 minutes) Breakout sessions (see 5.1) Breakout sessions, writing reports and preparing presentation Note-takers: distributed per group
2:00PM-5:30PM (30-min break @3:30PM)	Writing
5:30PM-6:00PM	Writing Plenary
Friday	
9:00AM-12:15PM	Wrap-up/next steps (new venue? Next dagstuhl? report?) Report -- assign tasks - decide venue
10:00AM-12:15PM	Next steps - community, events, venues

Introduction Talks

Danyel Fisher

Title: **Data Exploration requires collaboration between visualization and data infrastructures**

Abstract: As datasets grow to tera- and petabyte sizes, exploratory data visualization becomes very difficult: a screen is limited to a few million pixels, and main memory to a few tens of millions of data points. Yet these very large scale analyses are of tremendous interest to industry and academia. This paper discusses some of the major challenges involved in data analytics at scale, including issues of computation, communication, and rendering. It identifies techniques for handling large scale data, grouped into “look at less of it,” and “look at it faster.” Using these techniques involves a number of difficult design tradeoffs for both the ways that data can be represented, and the ways that users can interact with the visualizations. In this talk, I argue that solving these problems requires collaboration between visualization and data management skills. I outline one class of important problems, around the data flow pipeline, that is fundamentally in need of connection between these communities.

Eugene Wu

Title: **Closing the Loop on Data Analysis**

Abstract: The rapid democratization of data has placed its access and analysis in the hands of the entire population. While the tools for rapid and large-scale data processing have continued to reduce the time to compute analysis results, the techniques to help users better and more easily visualize their data, clean and prepare their data, and understand what their results mean are still lacking. In this talk, I will provide an overview of our lab's recent work on addressing each stage of data analysis—data cleaning, data visualization, and explanation.

Jason Dykes

Title: **vis: a geographer’s perpective**

Abstract: In a non-technical talk I introduced some of the kinds of questions that we attempt to answer in the giCentre at City, University of London. We use space as a framework for considering limited data about complex phenomena in context through our cartography and are exploring the
Examples relating to the dynamics of the London cycle hire scheme, the social geography of the UK, survey non-response bias, ballot ordering bias, landscape morphology and historical geography
Geographic patterns that we see and detect are scale dependent, with data recorded analysed and presented at a range of (often arbitrary) scales. The visual channel is rich and can both capture variations in scale and support concurrent multi-scale analysis of spatially related data, which may be informative in characterising phenomena and places. Patterns also vary if data are dynamic data or progressively loading and again, visualization can help us consider the stability or dynamism in signals and use such signatures to characterise phenomena. Considered design is import if we are to make the most of visualization in these ways. Thinking about layout and design within the page as a hierarchical specification of design options is beneficial.
Open questions include: developing perspectives that acknowledge visualization as an adequate answer from which actionable knowledge can be established; learning more about concurrent multi-scale thinking and its use in the context of task; developing approaches for participatory, exploratory, applied, problem driven research; integrating techniques from databases and machine learning to support interactive exploratory visualization and enhance the dialogue between data, design and discovery.

Alexandra Meliou

Title: **Understanding data: diagnosis, diversity, and fairness**

Abstract: The validity of data insights relies on proper interpretations of data and the results of data analyses. Thus, it is critical to develop tools and methodologies that facilitate understanding of data and data-driven processes. I describe three dimensions of this problem: diagnosis, diversity, and fairness.
(1) Diagnosis deals with finding causes of errors that are systemic to the data derivation processes. This allows repairing the processes, rather than just their effects, and improving data quality.
(2) Diversity focuses on retrieving small (top-k) representative subsets of data that a human can effectively process and interpret. But, retrieving diverse sets efficiently is challenging; I discuss our contributions towards a novel index structure that allows for simultaneous range search and diversification.
(3) Biases in data-driven software can have devastating societal consequences. Detecting when systems exhibit bias is crucial to understanding and correcting system behavior. I discuss our contributions in fairness testing, and open a discussion on the fairness-related challenges in creating tools that facilitate data insights.

Sihem Amer-Yahia

Title: **User Data Exploration**

Abstract: User data can be acquired from various domains and is characterized by a combination of demographics such as age and occupation and user actions such as rating a movie or recording one's blood pressure. User data exploration has been formulated as identifying group-level behavior such as {\em Asian women who publish regularly in databases}. Group-level exploration enables new findings and addresses issues raised by the peculiarities of user data such as noise and sparsity. The talk reviews some work on user data exploration via an architecture composed of 4 layers: raw user data preparation, group mining, user data exploration, and visualization. The talk highlights the relationship between exploration and mining via the optimization of new dimensions such as coverage, diversity and uniformity of mined groups. The talk ends with an example of health trajectory exploration that requires to think about how exploration and visualization can be combined via new dimensions, readability and informativeness.

Leilani Battle

Title: Behavior-Driven Optimizations for Big Data Exploration

Abstract: One of the key issues for people who work with large datasets is efficient visualization of their data to extract patterns, observe anomalies, and debug their workflows. Though a variety of visualization tools exist to help people make sense of their data, these tools often rely on database management systems (or DBMSs) for data processing and storage; and unfortunately, DBMSs fail to process the data fast enough to support a fluid, interactive visualization experience. My work blends optimization techniques from databases and methodology from HCI and visualization in order to support interactive and iterative exploration of large datasets. In this talk, I discuss ForeCache, a visual exploration system that learns user exploration patterns automatically, and exploits these patterns to pre-fetch data ahead of users as they explore. I show that ForeCache's pre-fetching techniques provide significant performance benefits compared to existing systems. I also discuss four current research directions for ForeCache, and my broader research interests.

Surajit Chaudhari

Title: Approximation in query processing: leveraging for visualization

Abstract: Approximation, using sketches and sampling, offer the opportunities to support

low-latency/progressive visualization. However, to realize this vision, having a comprehensible error model and query processing processing techniques to support such error models need to be developed. Another challenge is to identify what visualization-related operations need particular support from database technology.

Tim Kraska

Title: **Interactive Data Science**

Abstract: Unleashing the full potential of Big Data requires a paradigm shift in the algorithms and tools used to analyze data towards more interactive systems with highly collaborative and visual interfaces. Ideally, a data scientist and a domain expert should be able to make discoveries together by directly manipulating, analyzing and visualizing data on the spot, instead of having week-long forth-and-back interactions between them. Current systems, such as traditional databases or more recent analytical frameworks like Hadoop or Spark, are ill-suited for this purpose. They were not designed to be interactive nor to support the special requirements of visual data exploration. Similarly, most machine learning algorithms are not able to provide initial answers at "human speed" (i.e., sub-seconds), nor are existing methods sufficient to convey the impact of the various risk factors, such as multi hypothesis problem. In this talk, I will present our vision of a new approach for conducting interactive exploratory analytics and explain why integrating the aforementioned features requires a complete rethinking of the full analytics stack, from the interface to the "guts". I will present recent results towards this vision including our novel interface, analytical engine and automatic error detection, and outline what challenges are still ahead of us.

Arnab Nandi

Title: **Query, Feedback, Result**

Abstract: New computing interfaces that use "natural" modes of interaction — such as multitouch and gestures — are rapidly becoming more popular than traditional keyboard-based interaction. Applications for such devices are highly interactive, and pose a fundamentally different set of expectations on the underlying data infrastructure. To address these expectations, we propose a new paradigm that includes “feedback” into the query-result paradigm. We rethink various aspects of the data infrastructure stack to include feedback, from the query specification process to distributed query execution, to address interactive workloads, and walk through a few examples of including feedback.

Richard Wesley

Title: **Like Pulling Teeth: Data Extraction in Tableau**

Abstract: A large part of Tableau's history with data processing involves copying the data being visualised into extracts. This talk is a brief history of our experience with six different extraction systems, followed by a summary of the problems we encountered with traditional query execution engines.

Harish Doraswaimy

Title: **Interactive Visual Exploration of Large Urban Data**

Abstract: Visualization systems are widely used to explore large spatial data. These systems utilize visual spatial queries allowing users to interactively visualize multiple data sets. Supporting the interactive response times they require is challenging. Many applications use polygons of arbitrary shapes and sizes, and thus involve computationally-intensive operations (eg. point-in-polygon tests to associate points with the regions containing them) to accomplish their tasks. This problem is compounded for large data sets, since more such operations are required. In this talk, I will briefly motivate how the use of GPUs can immensely help obtain real time performance for spatial queries.

Carsten Binnig

Title: **IDEBench - Towards a Benchmark for Interactive Data Exploration**

Abstract: Existing benchmarks for analytical database systems such as TPC-DS and TPC-H are designed for static reporting scenarios. The main metric of these benchmarks is the performance of running different SQL queries over a predefined database. In this talk, I argue that such benchmarks are not suitable for evaluating modern interactive data exploration (IDE) systems, which allow data scientists of varying skill levels to explore large data set at interactive speeds. While query performance is still important for data exploration, I believe that a much better metric would reflect the number and complexity of insights users gain in a given amount of time. I will therefore discuss the challenges of creating such a benchmark and present a first concrete implementation of such a new benchmark called IDEBench (http://cs.brown.edu/~peichmann/idebench/) that simulates typical user behavior and allows IDE systems to be compared in a reproducible way.

Mike Gleicher

Title: **Some Things I Got Right, Some Things I Got Wrong**

Abstract: This talk will provide a survey of some recent work, with an aim of helping seed the conversation between visualization and data management. I will provide a definition of what visualization is, to focus our discussion on user tasks. Tasks will motivate a framework for thinking about user-centric challenges in working with data more broadly. I will provide an example of how scalability must be addressed as a design challenge, not just as a computational one. I will talk about how human perceptions provides challenges and opportunities for design.

Wesley Willett

Title: **Capturing and Communicating Context for Open Data**

Abstract: Visualization tools designed for public use have the potential to drive substantially greater citizen engagement with open government and institutional datasets. However, metadata documenting the complex processes of creating and managing these datasets are still often unavailable. As part of a collaborative multi-year engagement with Canada's National Energy Board, we are exploring how institutions might capture this context, and how public-facing visualizations can surface it — supporting conversations more deeply grounded in data.

Breakout Groups

Group 1 - Algebras/Languages for Vis

Group 2 - Understandability in Data Discoveries: a DB and Vis perspective

Group 3 - Evaluation in DB+Vis / User studies vs benchmarks

Group 4 - Help VIS to get to the DB happy place