# DSIA: Data Systems for Interactive Analysis

Remco Chang
Tufts University

Danyel Fisher
Microsoft Research

Jeffrey Heer
University of Washington

Carlos Scheidegger
University of Arizona

## 1 Background: DSIA 2015

The primary goal of the DSIA workshop is to bring together researchers from the Database community to participate in VIS. The premise of the workshop is that the development of back-end data systems is of increasing importance for visualization tools due to the growing size and complexity of data and the increased demand of interactivity. However, tackling the issue of data systems cannot be addressed by VIS researchers alone. Collaboration with other communities is essential to ensure integration between the front-end visualization and back-end storage and computation.

The first DSIA workshop was held in VIS 2015 with somewhere around 150 to 200 attendees. We were able to bring in young, but prominent researchers in Databases, including Eugene Wu (Columbia), Arnab Nandi (Ohio State), Aditya Parameswaran (Illinois), Aaron Elmore (Univ of Chicago), Jenny Duggan (Northwestern), Leilani Battle (MIT); as well as senior researchers like Marti Hearst (Berkeley) and Joe Hellerstein (Berkeley, who was also the invited keynote speaker). Overall, the event was successful – we observed the researchers from both sides interacting with each other, and anecdotally we are aware of a few collaborative projects borne from these discussions. For example, the collaboration between Jean-Daniel Fekete and Tim Kraska (Brown University) resulted in a recent TVCG paper [14].

## 2 Goals

The goal of the 2017 workshop is to continue the community-building activities that started in 2015, which is to foster innovative work in the backend that will support the next generation of interactive data analysis tools. We envision such backends to require novel insights in database technology, algorithm design, and information visualization techniques.

In the current design of database and analytics systems, interactive data visualization and analysis concerns are usually an afterthought. Although recent architectural changes like columnar databases have enabled faster visualization systems, these databases are often not specifically optimized for data visualization purposes but for general acceleration of queries. In this workshop, we will explore the consequences of pushing the concerns of interactive data visualization and analysis deep into the "computational fabric" of our systems: the backend infrastructure of data storage, organization, and retrieval. If we design with the knowledge that every analysis is eventually read by a human, bound by perceptual limitations and latency constraints, what are the consequences for database design, or data analysis and visualization algorithms? Are there fundamental tradeoffs? What can we give up?

Beyond DSIA in 2015, here have been previous workshops in VIS that have investigated issues of large data visualization, such as LDAV. Our focus is markedly different from that of LDAV. While LDAV focuses on simulation and imaging data, we believe current backend techniques in information visualization and visual analytics are close to hitting a ceiling in terms of performance. We propose this workshop as a venue to discuss these limits, and fundamental new ways to attack them.

The workshop most similar to DSIA is the HILDA workshop, which was held at SIGMOD in 2016 (and will be held again in 2017 in SIGMOD). The key difference between the two is that the database researchers at HILDA approach the problem from the opposite end as DSIA – in DSIA we seek to leverage the user's abilities (and limitations) in designing backend systems, in HILDA the researchers start with the backend system and analyze how the system might be useful to the user in data analysis tasks.

## 3 Technical Scope

The scope of this workshop will span, among possibly other fields, information visualization, visual analytics, algorithm design, database architecture, and statistics. Although this is clearly a broad scope, the workshop will focus on backends for interactive tools. As a community, visualization researchers have made significant progress in our understanding of perceptual and human-interaction issues behind interactive visualization design. However, much of this understanding has not, in general, been translated into new designs for our backend infrastructure. We think now is the right time to have a discussion of backend design for interactive data analysis. Following the principle of "No One Size Fits All" in designing databases [13], recent publications by the organizers have shown that by taking advantage of the needs and limitations of interactive visualization systems, databases can be optimized by making tradeoffs in novel and interesting ways. For example, imMens [7], profiler [5], and nanocubes [6] have shown that careful algorithm design can enable extremely fast interactive visualizations of massive data. Scalr [2] has shown that traditional query planners can fruitfully take advantage of perceptual and resolution limits. ForeCache [1] utilizes a predictive prefetching strategy by learning a user's interaction patterns [3] and prefetching data from the server to support real-time visual data exploration. Lastly, SampleAction [4] takes advantage of streaming databases and incorporates the HCI principle that users need immediate feedback, but are willing to wait for high-fidelity results.

In addition to the work by the organizers, researchers in the database community have also made significant strides in developing databases specifically for interactive visualization, analysis, and exploration of data. For example, systems like Tableau [12] utilize columnar databases which provide speedups for typical visual analytics tasks that require operations across columns (instead of the traditional rows). BlinkDB [9] uses a sampling strategy to build a smaller, but representative database from a larger one. MapD [11] takes advantage of modern graphics card capabilities and resulted in a specialized database that runs entirely within (an array of) graphics cards. Finally, SciDB [10] breaks away from the traditional relational data structure and is designed to efficiently store and query multidimensional scientific data that are stored in an array-based structure.

Given the increasing number of publications and backend systems that can be tailored or applied to interactive visual exploration of large data, we believe that it is very important for the InfoVis and the VAST communities to begin having discussions and to develop a research agenda that focuses on backend technologies. In this proposed workshop, we seek to invite members from the database community (in-

- *Remco Chang is with Tufts University. email: remco@cs.tufts.edu*
- *Danyel Fisher is with Microsoft Research. email: danyelf@microsoft.com*
- *Jeffrey Heer is with the University of Washington. email: jheer@uw.edu*
- *Carlos Scheidegger is with the University of Arizona. email: cscheid@cs.arizona.edu*

*Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014; date of publication xx xxx 2014; date of current version xx xxx 2014.*
*For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.*

cluding authors of the aforementioned papers) to meet with the VIS community and begin such dialogs. As many of these database researchers are current collaborators of the organizers (see Section 5 below for a list of intended invitees), we believe that the organizing team is well suited to influence research not only in information visualization, but in these other related fields as well.

## 4 Format, Length, and Planned Activities

We propose a half-day workshop that focuses on generating discussions between VIS and database researchers. The workshop will solicit submissions of position papers as well as short research papers from both the VIS and database communities. These papers will be reviewed by the organizing team. The authors of the selected top papers will be invited to each give a short presentation of their work (or position).

The goal of the presentation is to allow the attendees to gain an overview of the different perspectives and approaches of data storage, organization, and retrieval techniques that can be tailored specifically to support interactive data visualization and exploration. We anticipate that these presentations will take up approximately two thirds of the workshop.

Following a short break, the last one third of the workshop will be oriented around open discussions and possibly a question-and-answer session. As we anticipate that members of the two different communities will not have a common language or viewpoint, we believe that the open-ended format will facilitate discussions and general knowledge sharing. Based on our experience in 2015, a round-table discussion involving 150 to 200 participants is not possible. As a result, we envision a panel style discussion where the database researchers sit on stage and the participants from the VIS community can ask them questions and initiate a dialog.

It is relevant to note that the organizing team is also currently seeking a follow-up "closed-door" (invitation only) meeting that will take place in the afternoon outside of the main VIS venue. The purpose of this closed-door meeting is to allow for a more intimate discussion between enthusiastic researchers who are actively seeking collaborations across the aisle. Although it is not the goal of the organizing team to be exclusive, it is our belief that a more focused discussion will have a higher likelihood towards developing a cohesive research agenda. Note that this is not an extraordinary proposal: such an off-site format has been successfully employed by other workshops in previous years at VIS, most notably by the BELIV workshop [8].

## 5 List of Participants and Invitees

As noted earlier, it is our hope that this workshop will include both attendees from the VIS and database communities. However, since the workshop will be located at VIS, we will specifically reach out to members of the database community and invite them to attend the workshop. We will seek a mix of researchers in different stages of their careers, together with researchers from different companies in the industry. If possible, we will also seek to identify a more senior person from this invited list as the keynote speaker for the workshop.

### Senior researchers

- Sam Madden (MIT)
- Joe Hellerstein (Berkeley, Trifacta)
- Marti Hearst (Berkeley)
- Mike Stonebraker (MIT)
- Ugur Cetintemel (Brown)
- Stan Zdonik (Brown)
- Juliana Freire (NYU)

### Early or mid-career researchers

- Eugene Wu (Columbia)
- Aaron Elmore (Chicago)
- Tim Kraska (Brown)
- Carsten Binning (Brown)
- Aditya Parameswaran (UIUC)
- Arnab Nandi (Ohio State)
- Joey Gonzalez (Berkeley AMPLab, CMU, GraphLab/Dato)
- Bill Howe (UW)
- Arnab Nandi (OSU)
- Cecilia Aragon (UW)
- Jenny Duggan (Northwestern)
- Mike Cafarella (Michigan)

### Junior researchers (students, etc.)

- Manasi Vartak (MIT)
- Todd Mostak (MapD)
- Leilani Battle (MIT)
- Dominik Moritz (UW)

### Relevant companies we will reach out to

- Tableau
- Trifacta
- Paradigm4
- Graphistry
- Autodesk

## 6 Requested Facilities

In order to facilitate both paper presentations and open discussions, we seek to have a room that has space for both seats and for posters. A podium with audio/video support will be necessary for the paper presentations. In addition, some supplies for sketching and drawing would be beneficial in facilitating group discussions (e.g. big Easel Pads with markers).

## 7 Intended Result and Impact of Workshop Results

There are three goals for this workshop. First, we hope to engage both researchers in the VIS and the database communities. Should we be successful in attracting and inviting database researchers, the workshop can serve as a forum for true interdisciplinary discussion and collaboration.

Second, as we will accept both position papers and (short) research papers, we aim to compile the accepted papers into a short proceeding that we will disseminate to the workshop attendees. While we do not anticipate this proceeding to contain cutting-edge research contributions, we do envision the content to serve as a starting point for an open dialog between the communities.

Lastly, depending on the outcomes of these discussions, we aim to sketch out an initial outline for a research agenda for this burgeoning field. This research agenda will summarize the current state-of-the-art in both the VIS and database communities concerning backend supports for interactive data visualization and analysis. In addition, it will highlight the top challenges that need to be addressed that will hopefully spur future research in this field. The four organizers will enlist attendees of the workshop to help author this research agenda with the hope that the resulting paper will be published in a venue that is visible to both the VIS and database communities.

## 8 Important Dates

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Need to specify 3 dates: (1) date of participation (when the workshop will be held); (2) submission deadline (for the workshop), (3) date by which we'll notice by the authors

## 9 Organizer Bios

Remco Chang is an Associate Professor in the Computer Science Department at Tufts University. He received his BS from Johns Hopkins University in 1997 in Computer Science and Economics, MSc from Brown University in 2000, and PhD in computer science from UNC Charlotte in 2009. Prior to his PhD, he worked for Boeing developing real-time flight tracking and visualization software, followed by a position at UNC Charlotte as a research scientist. His current research interests include visual analytics, information visualization, HCI, and databases. His research has been funded by NSF, DHS, MIT Lincoln Lab, and Draper. He received the NSF CAREER Award in 2015.

Danyel Fisher is a Senior Researcher in information visualization and human-computer interaction at Microsoft Research's VIBE group. His research focuses on ways to help users interact with data more easily. His recent work has looked at ways to make big data analytics faster and more interactive. Danyel received his MS from UC Berkeley, and his PhD from UC Irvine. He tweets at @FisherDanyel.

Jeffrey Heer is an Associate Professor of Computer Science & Engineering at the University of Washington, where he directs the Interactive Data Lab and conducts research on data visualization, human-computer interaction and social computing. The visualization tools developed by his lab (D3.js, Vega, Protovis, Prefuse) are used by researchers, companies and thousands of data enthusiasts around the world. His group's research papers have received awards at the premier venues in Human-Computer Interaction (ACM CHI, UIST, CSCW) and Information Visualization (IEEE InfoVis, VAST, EuroVis). Other awards include MIT Technology Review's TR35 (2009), a Sloan Foundation Research Fellowship (2012), and a Moore Foundation Data-Driven Discovery Investigator award (2014). Jeff holds BS, MS and PhD degrees in Computer Science from UC Berkeley, whom he then betrayed by joining the Stanford faculty (2009-2013). Jeff is also a co-founder of Trifacta, a provider of interactive tools for scalable data transformation.

Carlos Scheidegger is an Assistant Professor in the Department of Computer Science at the University of Arizona since 2014. Prior to joining the Arizona faculty, Carlos worked at AT&T Research from 2009 to 2014, where he helped develop award-winning, open-source techniques for massive data sources (http://nanocubes.net). His current research interests are in the intersection of large-scale data analysis, information visualization, data management, and software infrastructure for scientific collaboration. His research has received multiple awards at top venues in data visualization (IEEE InfoVis, IEEE SciVis, EuroVis).

## Publications by Organizers


[1] Leilani Battle, Remco Chang, and Michael Stonebraker. Dynamic prefetching of data tiles for interactive visualization. In *Proceedings of the 2016 International Conference on Management of Data*, pages 1363–1375. ACM, 2016.

[2] Leilani Battle, Michael Stonebraker, and Remco Chang. Dynamic reduction of query result sets for interactive visualizaton. In *Big Data, 2013 IEEE International Conference on*, pages 1–8. IEEE, 2013.

[3] Eli T Brown, Alvitta Ottley, Helen Zhao, Quan Lin, Richard Souvenir, Alex Endert, and Remco Chang. Finding waldo: Learning about users from their interactions. *IEEE Transactions on visualization and computer graphics*, 20(12):1663–1672, 2014.

[4] Danyel Fisher, Igor Popov, Steven Drucker, and m.c. schraefel. Trust me, i'm partially right: Incremental visualization lets analysts explore large datasets faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1673–1682, New York, NY, USA, 2012. ACM.

[5] Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 547–554. ACM, 2012.

[6] Lauro Lins, James T Klosowski, and Carlos Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2456–2465, 2013.

[7] Zhicheng Liu, Biye Jiang, and Jeffrey Heer. immens: Real-time visual querying of big data. *Computer Graphics Forum*, 32(3pt4):421–430, 2013.


## Other References


[8] BELIV: BEyond time and errors: novel evaLuation methods for Information Visualization. http://beliv.cs.univie.ac.at/about.php.

[9] Sameer Agarwal, Barzan Mozafari, Aurojit Panda, Henry Milner, Samuel Madden, and Ion Stoica. Blinkdb: queries with bounded errors and bounded response times on very large data. In *Proceedings of the 8th ACM European Conference on Computer Systems*, pages 29–42. ACM, 2013.

[10] Philippe Cudré-Mauroux, Hideaki Kimura, K-T Lim, Jennie Rogers, Roman Simakov, Emad Soroush, Pavel Velikhov, Daniel L Wang, Magdalena Balazinska, Jacek Becla, et al. A demonstration of scidb: a science-oriented dbms. *Proceedings of the VLDB Endowment*, 2(2):1534–1537, 2009.

[11] Todd Mostak. An overview of mapd (massively parallel database). http://map-d.com/docs/mapd-whitepaper.pdf, 2014.

[12] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):52–65, 2002.

[13] Michael Stonebraker and Ugur Cetintemel. " one size fits all": an idea whose time has come and gone. In *Data Engineering, 2005. ICDE 2005. Proceedings. 21st International Conference on*, pages 2–11. IEEE, 2005.

[14] E. Zgraggen, A. Galakatos, A. Crotty, J. D. Fekete, and T. Kraska. How progressive visualizations affect exploratory analysis. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2016.