Symposium on Visualization in Data Science (VDS at IEEE VIS 2015)

(/2015/)

VDS 2015

Monday, October 26th, 2015

Menu

Sponsors

 (http://www.disneyresearch.com)

 (http://iacs.seas.harvard.edu)

Program

Summary

Sunday, October 25th, 2015

Location: Red

14:00-17:55	**Affiliated Tutorial: A Practical Introduction to Data Science in Python** Speaker: Stéfan van der Walt	

Monday, October 26th, 2015

Location: Red

08:30-10:10	**Data Science and Visualization for Scientific Discovery** Invited Speakers: Torsten Möller, Jeff Heer, Rumi Chunara
10:30-12:00	**Health and Spatio-Temporal Data** Panel Discussion, Paper Presentations
02:00-03:30	**Journalism, Sports, and Data Exploration** Invited Speakers: Rachel Schutt, Luke Bornn, Feifei Li
04:00:-05:50	**Databases and Algorithms** Panel Discussion, Paper Presentations
06:00:-07:30	**Poster Reception**

Affilliated Tutorial

Sunday, October 25th, 2015

A Practical Introduction to Data Science in Python
Stefan J. Van Der Walt, Berkeley Institute for Data Science
http://mentat.za.net/ (http://mentat.za.net/)

Abstract: From a wider perspective, Data Science be seen as the management and interpretation of data through computation and statistics. This tutorial highlights several of these core elements through an interactive computational workshop. To work with data, we need to access a data source, whereafter the data can be visualized to explore its structure. Based on intuitions gained about this structure, exploratory statistical analyses can then be made. Finally, more sophisticated machine learning models can be fit to the data to draw inferences and make predictions about data yet unseen. This tutorial systematically leads attendees through these steps by way of practical, real-world examples, augmented by hands-on computations in the Python language.

VDS

Monday, October 26th, 2015

Session 1: Data Science and Visualization for Scientific Discovery

8:30-10:10, Session Chair: Marc Streit

Opening Address

Visual data science - Advancing science through visual reasoning

Download Slides (/2015/files/slides_moeller.pdf)

Torsten Möller, University of Vienna

http://vda.cs.univie.ac.at (http://vda.cs.univie.ac.at)

Abstract: Modern science is driven by computers (computational science) and data (data-driven science). While visual analysis has always been an integral part of science, in the context of computational science and data-driven science it has gained new importance. In this talk I will demonstrate novel approaches in visualization to support the process of modeling and simulations. Especially, I will report on some of the latest approaches and challenges in modeling and reasoning with uncertainty. Visual tools for ensemble analysis, sensitivity analysis, and the cognitive challenges during decision making build the basis of an emerging field of visual data science which is becoming an essential ingredient of computational thinking.

Bio: Torsten Möller is a professor at the University of Vienna, Austria, since 2013. Between 1999 and 2012 he served as a Computing Science faculty member at Simon Fraser University, Canada. He received his PhD in Computer and Information Science from Ohio State University in 1999 and a Vordiplom (BSc) in mathematical computer science from Humboldt University of Berlin, Germany. He is a senior member of IEEE and ACM, and a member of Eurographics. His research interests include algorithms and tools for analyzing and displaying data with principles rooted in computer graphics, image processing, visualization and human-computer interaction. He heads the research group of Visualization and Data Analysis. He served as the appointed Vice Chair for Publications of the IEEE Visualization and Graphics Technical Committee (VGTC) between 2003 and 2012. He has served on a number of program committees and has been papers co-chair for IEEE Visualization, EuroVis, Graphics Interface, and the Workshop on Volume Graphics as well as the Visualization track of the 2007 International Symposium on Visual Computing. He has also co-organized the 2004 Workshop on Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration as well as the 2010 Workshop on Sampling and Reconstruction: Applications and Advances at the Banff International Research Station, Canada. He is a co-founding chair of the Symposium on Biological Data Visualization (BioVis). In 2010, he was the recipient of the NSERC DAS award. He received best paper awards from IEEE Conference on Visualization (1997), Symposium on Geometry Processing (2008), and EuroVis (2010), as well as two second best paper awards from EuroVis (2009, 2012).

© Universität Wien / Barbara Mair

Visualization for Discovery

Download Slides (/2015/files/slides_heer.pdf)

Jeff Heer, University of Washington

http://homes.cs.washington.edu/~jheer/ (http://homes.cs.washington.edu/~jheer/)

Abstract: While visualizations are often used to convey research results or tell data-driven stories, they can also be powerful tools for discovery. General visualization tools typically require manual specification of views: analysts must select data variables and then choose which transformations and visual encodings to apply. While well-suited to answering targeted questions, this interaction model may impose a tedious specification process that impedes broad consideration of the data, particularly in the early stages of analysis. In this talk, I will present our work designing tools for more comprehensive and efficient data exploration. Our primary approach is to develop mixed-initiative systems for steerable, interactive browsing of recommended visualizations, chosen according to statistical and perceptual measures.

Bio: 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 and Information Visualization (ACM CHI, ACM UIST, IEEE InfoVis, IEEE 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 to go teach at Stanford from 2009 to 2013). Jeff is also a co-founder of Trifacta, a provider of interactive tools for scalable data transformation.

---

Visualization in Public Health

Download Slides (/2015/files/slides_chunara.pdf)

Rumi Chunara, NYU

http://rumichunara.com/ (http://rumichunara.com/)

Abstract: Population-level public health entails many spatial and temporal challenges. Questions often arise about which spatio-temporal factors, from a slew of environmental, medical and other variables, affect health outcomes. Additionally there are issues of privacy, noise and resolution which must all be considered. Further, when thinking about how to action on and improve health outcomes, this information must be communicated to diverse parties. In this talk I will describe the goals and audiences involved in public health, survey efforts of how visualization is used in public health, and articulate areas where there is room for improved work in visualization.

Bio: Rumi Chunara is Assistant Professor of Computer Science and Engineering and Global Public Health at New York University. Her research focuses on building novel information sources and computational techniques to describe and predict population-level public health issues. Dr. Chunara received her Ph.D. in Electrical and Medical Engineering at the Harvard-MIT Division of Health Sciences and Technology, her S.M. in Electrical Engineering and Computer Science at MIT and her B.Sc. in Electrical Engineering at Caltech. She is a recipient of a Caltech Merit Scholarship, the MIT Presidential Fellowship, and was named an MIT Technology Review Innovator Under 35 in 2014. Her research has been funded by multiple sources including the National Science Foundation and National Institutes of Health.
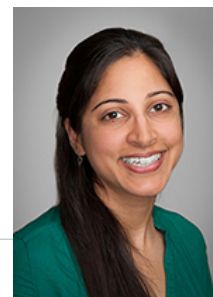
Session 2: Health and Spatio-Temporal Data

10:30-12:00, Session Chair: Hanspeter Pfister

Panel I: Challenges in Visualization for Data Science
Torsten Möller, Jeff Heer, Rumi Chunara
Moderator: Hanspeter Pfister

Paper: Service Oriented Development of Information Visualization of the Electronic Health Records for Population Data Set
Jaehoon Lee, Thomas Oniki, Nathan Hulse, Stanley Huff

Abstract: In this paper we propose a service-oriented architecture that provides visualization of clinical data of patient population as a service. We derived a service model including essential information of visualization transformation and developed a service manager based on the model to interface between clinical applications and visualization resources. The proposed framework was implemented in development environment at Intermountain Healthcare to be connected with our internal clinical applications. It demonstrated that the applications benefit by standardization and reuse of the services, elimination of duplicated development, use case of shared visualization between researchers, and enhanced security control.

Paper: RioBusData: Visual Data Analysis of Outlier Buses in Rio de Janeiro
Aline Bessa, Fernando de Mesentier Silva, Rodrigo Frassetto Nogueira, Enrico Bertini, Juliana Freire

Abstract: Buses are the main source of public transportation of the city of Rio de Janeiro, having around 100 million passengers every month. Recently, the city hall of Rio de Janeiro released the real-time GPS coordinates of all their operating public buses. The system generates almost 1 million GPS entries every day, but a reasonable amount of them are outliers - i.e. trajectories that do not follow their expected behavior. In this paper we present RioBusData, a tool that help users identify and understand, through different visualizations, the behavior of outlier buses that were automatically detected by a Convolutional Neural Network (CNN). RioBusData enables a better comprehension of the flow and service of outlier buses in Rio, and collaterally improves the understanding of Rio's main source of transportation as a whole.

Paper: Quality of Movement Data: from Data Properties to Problem Detection
Gennady Andrienko, Natalia Andrienko, Georg Fuchs

Abstract: Understanding of data quality is essential for choosing suitable analysis methods and interpreting their results. Investigation of quality of movement data, due to their spatio-temporal nature, requires consideration from multiple perspectives at different scales. We review the key properties of movement data and, on their basis, create a typology of possible data quality problems and suggest approaches to identifying these types of problems.

Session 3: Journalism, Sports, and Data Exploration

2:00-3:30, Session Chair: Alexander Lex

Doing Data Science at News Corp
Rachel Schutt, News Corp & Columbia University
http://datascience.columbia.edu/rachel-schutt-0 (http://datascience.columbia.edu/rachel-schutt-0)

Abstract: Increasingly, large corporations are building data science and engineering teams. These teams can be differentiated from traditional analytics teams in several ways including that they tend to run more like software engineering teams, that the methods and the practices often use open source technology, and that they are focused not solely on providing insights to the organization but also in building data products. News Corp is an interesting place to work because its history is in journalism and publishing, and the ways that data science and journalism intersect plays out in different ways including data-driven decision making in the news room, building sustainable business models for journalism and story-telling with data. Also most relevant to this conference are ways in which data visualization is used to communicate to internal business stakeholders for decision-making purposes, as well as to our readership.

Bio: Dr. Rachel Schutt is the Chief Data Scientist at News Corp, which includes The Wall Street Journal, Dow Jones, New York Post, Times of London, The Sun, The Australian; Harper Collins and Amplify. In this role, she is responsible for setting the global data strategy. Rachel was named a World Economic Forum Young Global Leader in 2015, and is on the 2014 Crain's New York Business 40 under 40 list. She is the co-author of the book "Doing Data Science" based on a class she created and taught at Columbia University, where she is an adjunct professor in the Department of Statistics. Rachel is a member of the Education Board for the Institute for Data Science at Columbia. Previously, Rachel was a statistician at Google Research and holds patents based on her work in the areas of social networks, large data sets, experimental design and machine learning. She earned her PhD in Statistics from Columbia University, a Masters degree in mathematics from NYU, and a Masters degree in Engineering-Economic Systems and Operations Research from Stanford University. Her undergraduate degree is in Honors Mathematics from the University of Michigan.

Space, Time, and Skill: Understanding High Performance Sport   Download Slides (/2015/files/slides_bornn.pdf)
Luke Bornn, Simon Fraser University
http://www.people.fas.harvard.edu/~bornn/ (http://www.people.fas.harvard.edu/~bornn/)

Abstract: In this talk I will explore how players perform, both individually and as a team, on a basketball court. By blending advanced spatio-temporal models with geography-inspired mapping tools, we are able to understand player skill far better than either individual tool allows. Using optical tracking data consisting of hundreds of millions of observations, I will demonstrate these ideas by characterizing defensive skill and decision making in NBA players.

Bio: Luke Bornn is an assistant professor in the Department of Statistics and Actuarial Science, Simon Fraser University. His work focuses on space-time modeling and statistical computation, with applications to structural engineering, climate, and sports.

Interactive Online Data Exploration and Analytics
Feifei Li, University of Utah
http://www.cs.utah.edu/~lifeifei/ (http://www.cs.utah.edu/~lifeifei/)

Abstract: How to store, process, and analyze large heterogeneous data is a fundamental challenge. In particular, it is important to support interactive analysis over such data, while reducing query latency and improving system throughput. We extend the concept of online aggregation to online analytics, and enable the support for interactive query conditions as well. Our goal is to provide query and analytical results continuously from the start of the query and analytical task execution, while providing quality guarantees and refining the query accuracy throughout the query execution. We implemented our system based on both traditional relational database engine and NoSQL system. When exact results are required for interactive query execution over big data, we introduce an in-memory interactive query and analytical engine based on Spark SQL that supports rich query semantics and analytics through both SQL and DataFrame APIs.
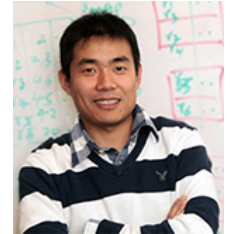
Bio: Feifei Li is an associate professor at the School of Computing, University of Utah. He obtained his B.S. in computer engineering from Nanyang Technological University, Singapore in 2002 (transferred from Tsinghua University, China) and PhD in computer science from Boston University in 2007. His research focuses on the scalability, efficiency, and effectiveness issues, as well as security problems in database systems and big data management. He was a recipient for an NSF career award in 2011, two HP IRP awards in 2011 and 2012 respectively, a Google App Engine award in 2013, the IEEE ICDE best paper award in 2004, the IEEE ICDE 10+ Years Most Influential Paper Award in 2014, a Google Faculty award in 2015, and the SIGMOD Best Demonstration Award in SIGMOD 2015. He is/was the demo chair for VLDB 2014, general chair for SIGMOD 2014, PC area chair for ICDE 2014 and SIGMOD 2015, and associate editor for IEEE TKDE.

Session 4: Databases and Algorithms

4:00-5:50, Session Chair: Claudio Silva

Panel II: Challenges in Visualization for Data Science
Rachel Schutt, Luke Bornn, Feifei Li
Moderator: Claudio Silva

Paper: Off-Screen Visualization Perspectives: Tasks and Challenges
Dominik Jaekle, Bum Chul Kwon, Daniel A. Keim

Abstract: The visual exploration of large data spaces often requires zooming and panning operations to obtain details. However, drilling down to see details results in the loss of contextual overview. Existing overview-plus-detail approaches provide context while the user examines details, but typically suffer from distortion or overplotting. This is why, there is great potential for off-screen visualization. Off-screen visualization is a family of techniques which provide data-driven context with the aid of visual proxies. Visual proxies can be visually encoded and adapted to the necessary data context with respect to scalability and visualization of high dimensional data. In this paper, we uncover the potential of off-screen visualization in visual data exploration by introducing its application examples to different domains through three derived scenarios. Furthermore, we categorize supporting tasks of off-screen visualizations and show areas of improvement. Then, we derive research challenges of off-screen visualizations and draw our perspectives on the issues for future research. This paper will provide guidance for future researchers on off-screen visualization techniques in visual data analysis.

Paper: Comparing Dimensionality Reduction Methods Using Data Descriptor Landscapes
Bastian Rieck, Heike Leitte

Abstract: Dimensionality reduction (DR) methods are commonly used in data science to turn high-dimensional data into 2D representations. Since data sets contain different structural features that need to be preserved by this process, there is a multitude of DR methods, each geared towards preserving a different aspect. This makes choosing a suitable algorithm for a given data set a challenging task. In this paper, we propose a comparative analysis of DR methods based on how well their embedding preserves structural features in the high-dimensional point cloud. To this end, we develop a set of data descriptors that assess local and global structural features of point clouds. These features are computed for the high-D point cloud and the 2D embedding. We then use persistent homology to robustly compare the feature functions. An interactive landscape of the data descriptors, based on their topological differences, permits visually exploring the embeddings and their quality. We demonstrate the utility of our workflow by analysing multiple embeddings on high-dimensional data sets from real-world applications.

Paper: Comprehension of data/model differences through diagrammatic reasoning
Kim Frederic Albrecht, Burcu Yucesoy

Abstract: How is celebrity status or popularity related to performance in sports? This initial question lead to the acquisition and analysis of large amounts of data regarding the performance of tennis players and the amount of public attention paid to them. In this paper we will explore the relationship between discoveries from this data and the visualization tools used to reveal them throughout the project. There were three steps of visual inquiry that followed one another organically: First, we created diagrams to help answering 'who' questions, in the sense that whose popularity does (or not) follow the trajectory of their performance. Then we needed diagrams to answer the 'why' questions as in why different athletes' careers would proceed differently. Finally we developed a tool to combine and connect the various visual forms to gain a deeper understanding of the entirety of the data set and we plan to use this tool to explore similar data sets in the future.

Paper: Feature-Based Visual Exploration of Text Classification
Florian Stoffel, Lucie Flekova, Daniela Oelke, Iryna Gurevych, Daniel A. Keim

Abstract: There are many applications of text classification such as gender attribution in market research or the identification of forged product reviews on e-commerce sites. Although several automatic methods provide satisfying performance in most application cases, we see a gap in supporting the analyst to understand the results and derive knowledge for future application scenarios. In this paper, we present a visualization driven application that allows analysts to gain insight in text classification tasks such as sentiment detection or authorship attribution on feature level, built with a practitioner's way of reasoning in mind, the Text Classification Analysis Process.

Closing

Poster Reception

6:00-7:30

Pathfinder: Visual Analysis of Paths in Heterogeneous Graphs
Christian Partl, Samuel Gratzl, Marc Streit, Hanspeter Pfister, Dieter Schmalstieg, Alexander Lex

NetSet: Interactive visualization for analyzing set in large network
Heungseok Park, Hongjun Lim, Kyungwon Lee

Seeing The Web of Microbes
Annette Greiner, Trent Northen, Suzanne Kosina, Richard Baran, Benjamin Bowen, Stefan Jenkins, Tami Swenson

Internet Review Opinion Mining utilizing Opinion Mining and Data Visualization
Seongmin Mun, Ginam Kim, Raja Mubashar Karim, Kyungwon Lee

Interactive exploration and verification of latent factor in large scale biophysical networks
Yosuke Onoue, Naohisa Sakamoto, Koji Koyamada

Visual analytics system for finding a causal relationship between physical quantities from multivariate volume datasets
Naohisa Sakamoto, Zhao Kun, Koji Koyamada

Minerva Taxi: Interacting with the Social Media and Transportation Landscape of Cities at Scale on the Web
David Manthey, Roni Choudhury, Jeffrey Baumes, Jonathan Beezley, Aashish Chaudhary

Show me the spot : Mapping & Parallel visualization of traffic accident pattern analysis in highway
Sang bin Han, Ba rom Kang, Seong yeop Han, Jin ki Kim, Kyung won Lee

Interactive Visualization for Interdisciplinary Research
Mohamed aly etman, Naomi Keena, Anna Dyson