Data Science Work Practices from a Ubiquitous Analytics Perspective

Søren Knudsen
University of Calgary
Calgary, AB, Canada
sknudsen@ucalgary.ca

Sheelagh Carpendale
Simon Fraser University
Vancouver, BC, Canada
sheelagh@sfu.ca

ABSTRACT

In this position paper, we discuss our prior and current work on data science work practices and how these might be supported by ubiquitous analytics tools. We argue collaborative tools might reduce risks of false conclusions based on ill-conceived understandings of data sets. Aiming to study this area, we outline research questions within these topics.

CCS CONCEPTS

• **Human-centered computing** → **Information visualization**; Human computer interaction (HCI);

KEYWORDS

Information visualization; data science, ubiquitous analytics.

ACM Reference Format:
Søren Knudsen and Sheelagh Carpendale. 2019. Data Science Work Practices from a Ubiquitous Analytics Perspective. In *Proceedings of ACM CHI Workshop on Human-Centered Study of Data Science Work Practices (CHI'2019)*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.475/123_4

CHI'2019, May 2019, Glasgow, Scotland, UK

Context of analysts' work

The analysts' work is characterized by constant adaptation to changing healthcare policies. To keep up with changing healthcare policies, the analysts revise the scripts used to create the rate foundation table. Thus, a large part of their work lay in 'data wrangling' [4, 5], including revising data flows and understanding where errors have occurred in the process.

Due to policy changes, information codes might be added, changed, or removed. For example, new administrative patient pathway codes, changes of codes describing in- and outpatients, and introductions of new medical procedures, might require new description codes.

Thus, while the analysts have scripts from previous years available, they will rarely work due to syntactic, semantic, and structural changes in the data between the previous years and the current. To develop scripts for the current year, they use scripts from the previous years, rely on the knowledge in the team, and debugging based on the output from the scripts.

INTRODUCTION

It is commonly suggested that the massive and complex data sets at our disposal holds value and that we can use these data sets to solve many aspects of personal, organisational, and societal challenges [1]. That we can use data to gain insights, to communicate knowledge, and to explore information. On the other hand, we are also starting to realise the very real consequences of inherent biases in the data. Data sets that we expected could be used to answer a broad set of questions, might only allow for correctly answering specific questions. Thus, incomplete knowledge of a data set might lead to false conclusions [2]. As data scientists, it is therefore crucial to have a good understanding of a data set, including how it was created.

Currently, most tools used by data scientists are created without consideration for collaboration. However, since it is important to have knowledge about a data set, collaboration on or communication about the background of a data set in some form is often necessary. From this perspective, we think collaboration should be considered as an integral part of data science work practises.

Although many other approaches are used in data analysis, it is rare to consider data analysis without some form of representation. Of potential representations, visual ones serve a wide variety of tasks and activities. However, we have also noticed that visual representations tend to have an air of trust associated with them — they tend to represent data in a "clean" manner.

In our work, we consider collaborative information visualization (InfoVis) and visual analytics (VAST) tools for working with data. While the former concerns mainly how to represent data and interact with it, the latter concerns the broader work surrounding visual representations in data analysis. We think it makes sense particularly to consider VAST tools in the context of collaboration. While we have mainly focused on co-located interdisciplinary data analysis [6–8], we have started to gain interest in understanding collaborative analysis processes more broadly.

We find that all the interesting and complex aspects of data analysis that we outlined above are present in the health domain. We see a diverse mix of people working together to collaboratively analyses massive and complex data sets. They are interested in gaining insights on diseases and treatments, and they are interested in educating and communicating their knowledge.

Through studies of data analysts' work practises, we have seen a high level of collaboration at a concrete site of analysis practise. Through prototyping and interventions of work practices, we have glimpsed at new possibilities for supporting collaborative analysis work, based on tools for ubiquitous analytics [3]. In the following, we describe these efforts. Afterwards, we describe our current plans for advancing the state. Finally, we discuss questions relating to our prior and current work.

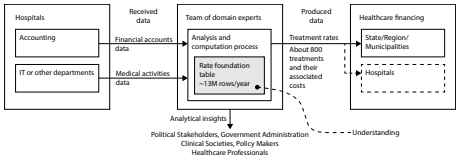

Figure 1: The analysts' collaborations and computation process.

Data characterization

The analysts work with data that comprise medical activities and financial accounts data. Medical activities data describe what has happened at a hospital (e.g., patient admittance and discharge dates from the wards and blood test metadata from clinical biochemistry labs). Financial accounts data describe the expenses incurred at a hospital (e.g., doctor and nurse salary expenses, implant costs, and overhead costs for each hospital department).

To compute the rates, the analysts map medical activities and financial accounts data in what they refer to as the rate foundation table. They construct the table based on a set of scripts—a subset of which relate to individual hospitals. The table contains a row for each patient (about 13 million per year). Each row describes a patient contact (an admission and discharge for inpatients and comparable information for outpatients) and comprises columns of patient information (e.g., age, gender, diagnoses), treatment information (e.g., procedures, duration, ward, hospital), and cost information (e.g., diagnosis-related group, salaries, overhead).

For example, codes describe operation procedures in a hierarchy of about 9,000 codes and hospital and ward definitions in another hierarchy of about 20,000 wards that describe physical locations that change both name and id over the years.

STUDIES AND PROTOTYPING

We have conducted observations, interviews, and workshops with a team of healthcare data analysts, who analyse data from and for the Danish healthcare system. The analysts' work reminds of tasks and contexts characterized by [5], and their level of expertise falls somewhere between hackers and scripters. See the sidebar "Context of analysts' work" for details. The analysts compute annual rates for hospital treatments. The analysts obtain data from about 50 hospitals, which they primarily use to map hospital activities to expenses (see also Figure 1) and sidebar "Data characterization". To discuss ongoing work, the analysts held weekly team meetings with their manager. In these meetings, each analyst provided a brief status update, including data, analysis, and scripting problems. Then, the issues were discussed between the analysts. After the meeting, the analysts returned to their desks. For example, in one meeting, an analyst presented a scripting problem relating to implant costs from a specific hospital. Other analysts asked if they had 'look[ed] into whether [the data] contained all implants,' since they had experienced implant types that caused problems in their scripts. This question required the analyst to look at the data again in order to be able to answer the question. In addition to the weekly analysis meetings, pairs of analysts often met for smaller scheduled and unscheduled one-half to two-hour meetings in front of a computer to work on a shared task. The meetings took place in the context of an informal work environment dominated by three- to four-person offices in which the analysts frequently interrupted each other with quick questions such as 'do you remember the code for the new cancer treatment?'—reminiscent of the blast-emails described by Kandel et al. [5].

We created a range of low- and high-fidelity prototypes for co-located collaboration based on large displays and evaluated these with the analysts described above through lab- and deployment-based studies to gain insight on how the tools we created could be used concretely, and how they might change the data science work. A main insight from our work is that any tool that is introduced in such an environment need to integrate tightly to existing work practices, in addition to usual expectations (e.g., be useable and useful on its own). In our situation, this means considering how analysts might use existing tools together with tools designed for multiple devices and form factors, which is discussed by Elmqvist & Irani [3]. In our interventions, we realised that this could be solved by simple techniques. For example, analysts considered the idea of receiving screenshots from an analysis tool running on a large display, which could be sent via email or stored in a shared folder. They also considered receiving SQL queries that corresponded to the queries they had performed using touch interaction techniques on a large display.

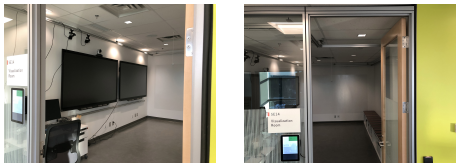

Figure 2: The visualization room at the Center for Health Informatics (C-H-I).

OUR PLANS

Based on the above insights, we are currently establishing a major long-term collaboration with an interdisciplinary group of health care data analysts. As the first task in this collaboration, we were working with them to create a shared interdisciplinary space: "Center for Health Informatics (C-H-I)". As part of this space, we have set up a visualization room equipped with large displays and a motion capture system (see Figure 2). The room is situated in the center of the C-H-I and in a location that people naturally pass on their way to and from meetings, lunch, and coffee breaks. To foster efficient collaborations, we have a drop-in office, which we hope will enable us to engage with the analysts on a day-to-day basis.

As first questions for this collaboration, we are planning to conduct an interview study aiming to explore the possibilities for using the visualization room. We also consider running surveys on current tool use.

OPEN QUESTIONS

We are interested in discussing how we might support highly collaborative analysis work, both co-located and remote, synchronous and asynchronous (all four quadrants of the CSCW matrix). In particular, we are interested in studying how collaboration impacts a teams' understanding of complex phenomena in data sets. Can collaborative tools reduce risks of faulty conclusions in data science?

On a more concrete level, we expect to create tools for a heterogeneous environment, where different people use different tools and have different background knowledge. We wonder how we might design such tools and see this as an interesting and concrete discussion.

CONCLUSION

We have briefly described some of our prior and current work on considering co-located collaborative data analysis and outlined questions we find important. We see many interesting discussions on this topic, and much work still ahead.

REFERENCES

[1] Ritu Agarwal and Vasant Dhar. 2014. Editorial—Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. 25, 3 (2014), 443–448. https://doi.org/10.1287/isre.2014.0546

[2] Danah Boyd and Kate Crawford. 2012. Critical Questions for Big Data. 15, 5 (2012), 662–679. https://doi.org/10.1080/1369118X.2012.678878

[3] Niklas Elmqvist and Pourang Irani. 2013. Ubiquitous analytics: Interacting with big data anywhere, anytime. *Computer* 46, 4 (2013), 86–89.

[4] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. 2011. Research directions in data wrangling: Visualizations and transformations for usable and credible data. 10, 4 (2011), 271–288. https://doi.org/10.1177/1473871611415994

Author bio (first author)
Søren Knudsen is a Marie Curie post-doctoral research fellow at the Department of Computer Science at University of Calgary. He received his PhD in 2015 from University of Copenhagen in Denmark. His research interests centers on Human–Computer Interaction and Information Visualization, with particular focus on multiple views in visualizations, and how this might facilitate individual and shared data understanding. Working with the National Energy Board of Canada, the notions of truth, trust, and provenance, became crucial in discussing visualizations for the general public. This conceptual understanding now underpins much of his research on information visualization.

[5] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Enterprise data analysis and visualization: An interview study. 18, 12 (2012), 2917–2926. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6327298

[6] Søren Knudsen. 2015. *How Does Abundant Display Space Support Data Analysis?: Interaction Techniques for Information Visualizations on Large, High-Resolution Displays.* Ph.D. Dissertation. Department of Computer Science, Faculty of Science, University of Copenhagen.

[7] Søren Knudsen, Mikkel Rønne Jakobsen, and Kasper Hornbæk. 2012. An exploratory study of how abundant display space may support data analysis. In *Proc NordiCHI* (2012). ACM, 558–567. http://dl.acm.org/citation.cfm?id=2399102

[8] M. Tobiasz, P. Isenberg, and S. Carpendale. 2009. Lark: Coordinating Co-located Collaboration with Information Visualization. 15, 6 (2009), 1065–1072. https://doi.org/10.1109/TVCG.2009.162