Understanding alternatives in data analysis activities

Jiali Liu, Nadia Boukhelifa, James Eagan

Understanding Alternatives in Data Analysis Activities

Jiali Liu
LTCI, Telecom ParisTech,
Université Paris-Saclay
46, rue Barrault
Paris, France
jiali.liu@telecom-paristech.fr

Nadia Boukhelifa
INRA, Université Paris-Saclay
1 av. Brétignières, 78850,
Thiverval-Grignon, France
nadia.boukhelifa@inra.fr

James R. Eagan
LTCI, Telecom ParisTech,
Université Paris-Saclay
46, rue Barrault
Paris, France
james.eagan@telecom-paristech.fr

Abstract
Data workers are non-professional data scientists who engage in data analysis activities as part of their daily work. In this position paper, we share our past work and our multidisciplinary approaches on understanding data workers' sense-making practices and the human-tool partnerships. We introduce our current research ideas on the role of alternatives in data analysis activities. Finally, we conclude with open questions and research directions.

Author Keywords
Data science, sense-making, visual analytics

Introduction
Data analysis shares similar characteristics with experimental [15] and design processes [12]: both are open-ended, highly interactive, and iterative—where a broad space of possible solutions are explored until a satisfactory result emerges [4, 7]. As such, data workers generate and compare multiple schemas [13], formulate solutions by combining parts from different exploration paths [7], and deal with uncertainty that could arise at different layers of the analysis [2]. Moreover, multiple types of data workers need to collaborate together to perform more complex analytic tasks [1, 6].

While real-world analyses tend to be messy and complex,

today's tools still largely rely on single-state models involving one user at a time. This disconnect contributes to making analytic practices cumbersome and increases cognitive load [9, 7]. Worse still, a lack of support for exploration of alternatives, explicit management of uncertainty in analysis, and support for collaboration in sense-making can lead to bad problem solving and decision-making [11].

Our research more generally focuses on understanding how data workers make sense of data and on providing them with rich, flexible tools that are adapted to the highly situated nature of such interactions. Our current work focuses more specifically on understanding the role of alternatives in data workers' analysis practices.

In this position paper, we describe our past and on-going work on understanding data workers' practices. We describe our background and our general approach , and close with open research questions drawn from our experience.

Background & Approach

We are researchers in information visualisation and more generally human-computer interaction, with a background in computer science. One co-author is a researcher in an applied laboratory for agronomics research, while the other two are in an engineering school.

Our work focuses on data workers: professionals who engage in data analysis activities as part of their daily work but who would generally not characterize themselves as a data analyst or data scientist. Our approach tends to combine different qualitative methods to understand these users' needs and practices, tool-building to address these needs, and other qualitative and quantitative evaluation methods.

In prior work, for example, we conducted interviews and walk-through scenarios with data workers to understand how they think about and manage the various kinds of uncertainty that arise in their work [2]. In a recent study, we examined how multiple types of data workers collaborate to explore complex data sets (*e. g.*, model simulations) using a visualization tool [1].

Beyond analyzing the data workers' general sense-making strategies, we also looked at how they explore large search spaces and how they reconcile conflicting optimisation criteria. The results of this study revealed an iterative analysis approach adopted by our data workers, that interleaves different types of analysis scenarios.

We found exploration scenarios where data workers examine together new and refined research questions and hypotheses, and other scenarios where they learn to appropriate the exploration tool and setup, and attempt to recap and establish common ground (storytelling). This type of study can help improve our understanding of the role of human expertise and its inter-play with visual analytics in reaching new insights and building common ground during collaborative data work.

Other work focuses more on building better tools for analysts—usually within a specific context. For example, the EvoGraphDice prototype [3] combines the automatic detection of visual features with *human interpretation* to aid in the exploration of multidimensional datasets. Our recent work on Codestrates [14] builds literate computing capabilities—similar in concept to Jupyter notebooks—on top of Webstrates [8] to provide a malleable, collaborative environment in which users can not only collaboratively write code but can also dynamically extend the environment from within the environment. Such systems aim to reduce the barriers

for users to adapt the tool to their contextual needs and further a co-adaptive relationship between tool and user [10].

Alternatives in Data Analysis Activities

In our current study, we take our first step at trying to gain a deeper understanding of the nuance and complexity of how data workers explore alternatives. For example, in the early stage of data analysis, alternative ideas are explored to better define the problem [6]; in the implementation stage, alternative data features, models, and evaluation methods are explored for different reasons [1, 5]. However, it remains unclear what kinds of alternatives exist within the data sense-making loop—what characteristics do they share and how do these considerations of alternatives influence sense-making.

We focus on understanding: (1) When do data workers consider and try alternatives (if at all)? (2) What kinds of alternatives do they consider? (3) What strategies do data workers adopt to explore and manage alternatives? (4) What are the triggers and barriers to "alternative exploration"? (5) How do existing tools support the exploration of alternatives?

We conducted semi-structured interviews with 12 participants (4 from an enterprise setting and 8 from the research domain). We collected 827 minutes of recordings yielding 585 unique observations. We are analyzing these data with a combination of affinity diagramming and workflow diagram methods with the goal of characterizing the role of alternatives in these kinds of sense-making activities.

Our aim is to provide a richer understanding of these practices and to identify specific gaps that call out the need for new tools or new ways of designing such tools.(nb: I wonder if we have a results teaser to add here, even if very high level, couple of sentences or so ...)

Open questions & research directions

Data work involves a rich body of disciplines and methods. While significant work has identified and formalized various sense-making processes, there is still much to learn about these practices in impromptu or less-formal settings.

Moreover, much analysis work takes place in the head of the user. How can we build better tools and processes to help get such tacit learning out of the head of the individual and into a form that can be more easily shared within collaborative contexts.

(jl: Data workers often need to explore a large scale of alternatives. How do we design tools to enable better focus on exploring and comparing alternatives, as opposed to just managing them?)

Alternatives can be generated, updated, revisited, and recombined in different stages along the analysis. How do we design tools to actively support and to record this process in a fluid manner.?

Data workers need to revisit past and on-going data analysis projects, both to share key insights with others, but also to reflect upon and to adapt their sense-making strategies. How do we facilitate reflective and collaborative sensemaking in a team of data workers having differing skills and expertise?

REFERENCES

1. Nadia Boukhelifa, Anastasia Bezerianos, Ioan Cristian Trelea, Nathalie Mejean Perrot, and Evelyne Lutton. 2019. An Exploratory Study on Visual Exploration of Model Simulations by Multiple Types of Experts. In *CHI '19: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 14. DOI:

http://dx.doi.org/10.1145/3290605.3300874

2. Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. 2017. How Data Workers Cope with Uncertainty: A Task Characterisation Study. In *CHI '17: Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3645–3656. DOI: http://dx.doi.org/10.1145/3025453.3025738

3. Waldo Cancino, Nadia Boukhelifa, and Evelyne Lutton. 2012. Evographdice: Interactive evolution for visual analytics. In *Evolutionary Computation (CEC), 2012 IEEE Congress on*. IEEE, 1–8.

4. Björn Hartmann, Loren Yu, Abel Allison, Yeonsoo Yang, and Scott R. Klemmer. 2008. Design as exploration: Creating interface alternatives through parallel authoring and runtime tuning. *UIST 2008 - Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology* (01 2008), 91–100. DOI: http://dx.doi.org/10.1145/1449715.1449732

5. C. Hill, R. Bellamy, T. Erickson, and M. Burnett. 2016. Trials and tribulations of developers of intelligent systems: A field study. In *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. 162–170. DOI: http://dx.doi.org/10.1109/VLHCC.2016.7739680

6. Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Enterprise Data Analysis and Visualization: An Interview Study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (Dec. 2012), 2917–2926. DOI: http://dx.doi.org/10.1109/TVCG.2012.219

7. Mary Beth Kery, Amber Horvath, and Brad Myers. 2017. Variolite: Supporting Exploratory Programming by Data Scientists. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1265–1276. DOI: http://dx.doi.org/10.1145/3025453.3025626

8. Clemens N. Klokmose, James R. Eagan, Siemen Baader, Wendy Mackay, and Michel Beaudouin-Lafon. 2015. Webstrates: Shareable Dynamic Media. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*. ACM, New York, NY, USA, 280–290. DOI: http://dx.doi.org/10.1145/2807442.2807446

9. Aran Lunzer and Kasper Hornbæk. 2008. Subjunctive Interfaces: Extending Applications to Support Parallel Setup, Viewing and Control of Alternative Scenarios. *ACM Trans. Comput.-Hum. Interact.* 14, 4, Article 17 (Jan. 2008), 44 pages. DOI: http://dx.doi.org/10.1145/1314683.1314685

10. Wendy E. Mackay. 1990. *Users and Customizable Software: A Co-Adaptive Phenomenon*. Ph.D. Dissertation. Massechusetts Institute of Technology.

11. David S. McLellan. 1980. Presidential Decisionmaking in Foreign Policy: The Effective Use of Information and Advice. By Alexander L. George. (Boulder, Colo.: Westview Press, 1980. Pp. xviii 267. $24.00, cloth$;10.00, paper.). *American Political Science Review* 74, 4 (1980), 1082–1083. DOI: http://dx.doi.org/10.2307/1954355

12. Roger Peng. 2018. Divergent and Convergent Phases of Data Analysis. (2018). https://simplystatistics.org/2018/09/14/divergent-and-convergent-phases-of-data-analysis/

13. Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. (2005), 2–4.

14. Roman Rädle, Midas Nouwens, Kristian Antonsen, James R. Eagan, and Clemens Klokmose. 2017. Codestrates: Literate Computing with Webstrates. 715–725. DOI: http://dx.doi.org/10.1145/3126594.3126642

15. J. W. Tukey and M. B. Wilk. 1966. Data Analysis and Statistics: An Expository Overview. In *Proceedings of the November 7-10, 1966, Fall Joint Computer Conference (AFIPS '66 (Fall))*. ACM, New York, NY, USA, 695–709. DOI: http://dx.doi.org/10.1145/1464291.1464366