Towards Data Science for the Masses: A Study of Data Scientists and their Interactions with Clients

Abigail Mosca* Tufts University
Shannon Robinson† Tufts University
Meredith Clarke‡ Tufts University
Rebecca Redelmeier§ Tufts University
Sebastian Coates ¶ Tufts University
Dylan Cashman ‖ Tufts University
Remco Chang** Tufts University

ABSTRACT

There is a growing gap between the general public and advanced analytics. In an effort to better understand this gap, and potential ways to overcome it, we performed 14 semi-structured interviews with data scientists who work with clients with little to no expertise in data science. Interview recordings were coded and analyzed, resulting in three major data scientist - client tasks: (1) initialization, (2) communication, and (3) iteration. We find data scientists spend the majority of their time on task 1–defining analysis queries based on broad, high-level client questions and running the appropriate analysis. In this poster, we explore the tactics data scientists employ for this task and potential avenues for automation.

1 INTRODUCTION

Data democratization is the idea that data should be available for anyone to access, analyze, and benefit from. In the spirit of data democratization, collection and availability of data has expanded dramatically [4]. Similarly, data science has been rapidly advancing analytic techniques, producing increasingly sophisticated and accurate tools. These tools, coupled with greater availability of data, allow for analyses and insights previously unattainable even in recent years.

While there has been an increase in tools and systems available for analyzing data most of these tools still require the user to have certain levels of proficiency in data science or programming to perform advanced analyses. For users lacking such proficiency, there is a gap between available data science techniques and the ability to utilize them without outside help.

We seek to learn ways in which visualization can bridge this gap between non-technical people and advanced analytics. To do this, we performed semi-structured interviews with 14 data scientists from a variety of data intensive fields. In each field, we sought out professionals who directly interface with clients and either perform data analytics themselves or have a high-level understanding of advanced analytics. Our goal is to learn the techniques and tactics employed by data scientists during their interactions with non-technical people so we may incorporate them into visualization systems that make advanced analytics more accessible.

We analyzed interviews quantitatively and qualitatively and found that despite domain and background data scientists regularly perform three tasks: (1) ***initialization***: translating clients' broad, high-level questions into analytic queries and running the appropriate analysis, (2) ***communication***: explaining the results of analytic queries to clients, and (3) ***iteration***: following up on analyses based on feedback from clients. However, time is not evenly split between the three; the majority of time is spent on initialization. In this poster we discuss the tactics data scientists employ for ***initialization***, and how we might be able to automate them.

*e-mail: abigail.mosca@tufts.edu
†e-mail: shannon.robinson@tufts.edu
‡e-mail: meredith.clarke@tufts.edu
§e-mail: rebecca.redelmeier@tufts.edu
¶e-mail: sebastian.coates@tufts.edu
‖e-mail: dylan.cashman@tufts.edu
**e-mail: remco@cs.tufts.edu

2 STUDY DESIGN

Our work differs but is complementary to the recent study by Kandel et al. [3] where the authors interviewed data scientists to understand the tools they use and the challenges they face when analyzing data. Our interviews with data scientists instead focus on the interactions between the data scientists and the clients, particularly clients who have limited or no knowledge of data science techniques, but who nonetheless have data science needs.

We performed semi-structured interviews with fourteen data scientists working in market research, biomedical research, policy research, and epidemiological and health research. Some of these data scientists were mathematicians or statisticians by training, while others had backgrounds in marketing or other areas. Each interview lasted 30 to 60 minutes and occurred in person (11 interviews) or via a conference call (3 interviews). All interviewees signed a consent form to participate in the study, which included agreeing to an audio recording of their interview. Recorded interviews were coded via an iterative coding process, and analyzed. Below, we discuss the three data scientist - client tasks resulting from analysis as well as qualitative findings related to ***initialization***.

3 DATA SCIENTIST - CLIENT TASKS

The major outcome of the coding process is three distinct tasks which data scientists undertake when working with clients: ***initialization***, ***communication***, and ***iteration*** . During ***initialization***, data scientists aim to determine the precise needs of the client. This task covers the process of first meeting with a client and defining the client's problem, up through the initial trials of running an analysis. During ***communication***data scientists are focused on explaining the results of their analyses and confirming clients' understanding. Finally, during ***iteration*** data scientists aim to engage the client in clarifying analysis results and adding depth to analyses.

Though these three tasks are essential to successful projects between data scientists and clients, analysis revealed that data scientists spend the vast majority of their time on ***initialization***. In order to run an analysis, data scientists need to understand their clients' goals and questions, as well as the nuances of the client's domain area. However, clients may come to data scientists not fully aware of their own analysis needs or of the capabilities of advanced analytics, increasing the difficulty of this task. We explore stopping points and tactics pertaining to ***initialization*** in the next section.

4 QUALITATIVE FINDINGS ON INITIALIZATION

During ***initialization*** clients' analysis needs can present with different levels of clarity. A need may be expressed as broad and

high-level as *"can you tell me about [a topic]?"*, or as specific as a testable hypothesis such as *"are 5th graders in New Hampshire better at math than 5th graders in Massachusetts?"*. Regardless of this range in question clarity, it is up to the data scientist to formulate and formalize an analysis that reflects the clients' analysis goals. Interviews reveal that data scientists have an arsenal of tactics to use for this, discussed in the following subsections.

4.0.1 High Clarity Needs: Working Backwards

In the case of high clarity analysis needs, clients have a solid sense of what they want from the analysis but may not know what to ask in order to achieve their end goal. For example, a market researcher may have a clear goal of seeking evidence to support targeting millennials in ads. However, because the client lacks analysis expertise, they may ask the data scientist to *"tell me about millennials"*.

In order to construct a formal analysis query for a high clarity request, data scientists explained that they will start by asking clients to define the end goal of their project and try to find the *why* behind a request. We call this technique "working backwards." To continue the example above, the data scientist would want the client to express their end goal, evidence to support targeting millennials in ads. One data scientist explained that this tactic is effective because it provides a well-defined analysis outcome from which the data scientist can extrapolate the appropriate approach. Essentially, with a well-defined outcome, the data scientist can identify and perform the necessary analytics, independent of the client.

4.0.2 Medium Clarity Needs: Probing

Sometimes a client will struggle in answering the data scientist's questions about a specific end goal, but still have a gist of what they would like to know from the analysis. This is characteristic of clients with medium clarity needs. For example, a client might know that they would like to write a report that says *"drinking milk is healthy"*. However, the population for which they want to make this claim is not clear, nor is the definition of 'healthy'. In this case, data scientists will employ probing questions to understand the clients' problem space and determine an analysis strategy. We call this technique "probing".

One data scientist explained that the purpose of probing is to get clients talking so that they provide enough background information for the data scientist to *"fill in the blanks"*: *"Once they get [talking] they can talk for days, so that gives me some information and some background... we need open-endedness so that we can fill the blanks of what they're talking to us about"*.

Essentially, probing is a way for the data scientist to learn about clients' backgrounds, needs, and restrictions. Once the data scientist has a strong handle on these, they can fill in the blanks to define an end goal themselves and proceed as they would for a high clarity need.

4.0.3 Low Clarity Needs: Recommending

When probing and working backwards fail, the client probably has a low clarity need. For example, a market researcher may say to the data scientist *"tell me about millennials"*. In contrast to a client with a high clarity need, they may not be able to communicate what they specifically want to say about millennials, or provide additional specificity on their interest in millennials when asked probing questions. In cases like this, most data scientists will run several plausible analyses based on their domain expertise and present them to the client to see which they prefer. This process of running and presenting analyses will continue until the client identifies one as useful. We call this tactic "recommending".

When probing, data scientists either utilize their expertise in the clients' domain, or spend time talking with the client to gain a deeper understanding. From there, they guess what the client's need may be, provide an analysis result based on that understanding, and then adapt based on client feedback.

5 Discussion

The regularity in our findings for ***initialization*** techniques of data scientists is an indication that it is possible to automate this task for data scientists. Based on the above analysis, we can formally define the "inputs" and "outputs" to ***initialization***.

Initialization takes as input a high-level question from a user and outputs an analogous analytic query. Taking cues from the tactics of data scientists, we can build systems that achieve this goal by emulating working backwards, probing, and recommending. For example, working backwards could be achieved through systems that accept an analysis end goal and work backwards to select and perform an analysis, similar to Query by Example techniques [12, 7], or Visual Query Systems [6]. Probing could be achieved through systems that leverage incremental query construction techniques [11]. Finally, recommending could be achieved through systems that recommend analyses to users, similar to visualization recommendation systems such as Voyager [10], VizDek [5], Small-Multiples-Large-Singles [9], and Foresight [1].

Identifying inputs and outputs for ***initialization*** is a first step toward automating this task. Our next steps include analyzing interview data for ***communication***, and ***iteration*** to identify inputs and outputs for those tasks so that we can explore methods of automation. With these three pieces in place, we can begin to conceive of a framework that could serve as a blueprint to building systems that perform data scientist - client tasks on the surface, and perform advanced analytics under the hood. Such systems would make advanced analytics more accessible to people who do not have the technical expertise to perform such analyses on their own, and who cannot afford to contract with data scientists. In other words, such a system would be a step towards making advanced analytics more accessible to the masses.

Acknowledgments

This work was supported by NSF 1452977 and DARPA FA8750-17-2-0107.

References

[1] Ç. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati. Foresight: Rapid data exploration through guideposts. *CoRR*, abs/1709.10513, 2017.

[2] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[3] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. In *IEEE Visual Analytics Science & Technology (VAST)*, 2012.

[4] A. Katal, M. Wazid, and R. H. Goudar. Big data: Issues, challenges, tools and good practices. In *2013 Sixth International Conference on Contemporary Computing (IC3)*, pages 404–409, Aug 2013.

[5] A. Key, B. Howe, D. Perry, and C. Aragon. Vizdeck: self-organizing dashboards for visual analytics. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 681–684. ACM, 2012.

[6] J. Lloret-Gazo. A survey on visual query systems in the web era (extended version). *CoRR abs/1708.00192*, 2017.

[7] Query by Example. Query by example – wikipedia, the free encyclopedia. Online; accessed 4-November-2017, 2005.

[8] C. Stolte, D. Tang, and P. Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, Jan 2002.

[9] S. van den Elzen and J. J. van Wijk. Small multiples, large singles: A new approach for visual data exploration. In *Computer Graphics Forum*, volume 32, pages 191–200. Wiley Online Library, 2013.

[10] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(11):649–658, Jan. 2016.

[11] G. Zenz, X. Zhou, E. Minack, W. Siberski, and W. Nejdl. From keywords to semantic queriesincremental query construction on the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):166 – 176, 2009. The Web of Data.

[12] M. M. Zloof. Query by example. In *Proceedings of the May 19-22, 1957, nationl computer conference and exposition*, pages 431–438. ACM, 1957.