Statement of Purpose

How can we propagate breakthroughs in the scientific community to the real world? With the explosion of big data, how can we help fields outside of computer science (CS) extract and leverage its value? Inspired by these questions, my current research focuses on facilitating user interaction with databases. Specifically, I apply visualization and machine learning techniques to alleviate the barriers between users and databases to help users access and make sense of data. By helping users better explore and understand the data they have collected, I hope to enable data-driven decision-making in a wide range of fields. It is with these broad goals in mind that I am applying to pursue a PhD.

Finding My Research Interests. With a focus on data provenance summarization, my research journey began under Dr. Rachel Pottinger at the University of British Columbia (UBC) Data Management and Mining Lab. The provenance of a query over a database is a subset of the data in the database that contributed to the query answer. While comprehensive, query provenance consists of large volumes of data and hence is overwhelming for users to explore. We presented an approach to provenance exploration that builds on data summarization techniques and provides an interface to visualize the summary. This work led to the first two papers I co-authored, *Summarizing Provenance of Aggregation Query Results in Relational Databases* (ICDE'21) and *Pastwatch: On the Usability of Provenance Data in Relational Databases* (ICDE'20). My main contributions include identifying the limitations in the existing methods, implementing the existing and our summarization methods, and running the experiments. Our work experienced a few submissions. Although I felt discouraged at first, I learned to reflect and was encouraged by how much our work had improved after each round. I also enjoyed my experience in research more than the industry for the autonomy and ownership over my work. However, I had some burning questions regarding my research interests going forward. Although I was engaged by the technical aspects of solving open-ended problems, I wanted to find something that would really excite me – what is the thing that would get me out of bed every morning? And how could I find it?

My next project, *Developing a Data-Driven Electric Vehicle (EV) Strategy in Surrey, BC, Canada*, helped me answer those questions. Working with another undergraduate student under the supervision of Dr. Raymond Ng, we set out to address the challenge of how the city of Surrey should place EV charging stations. Prior to our work, the approach to determine where to install an EV charging site was solely based on expert opinions, despite a large volume of data collected by the city of Surrey. To help city planners make strategic decisions informed by evidence, I developed a web application to give them a user-friendly way to explore and make sense of the data. I used interactive maps and graphs to visualize the spatial distribution and time trends of Surrey's vehicle stock, traffic flows, and land use. In September 2019, the city used my tool to choose 20 charger locations for a Canadian federal funding proposal, and I was proud to co-present this work at the SIGKDD'20 Social Impact Session this summer. Through zooming in and out on a pressing, real-world issue, I realized what I should be looking for in the research I pursue: the possibility of helping others and the insight into real-world issues that would spark that possibility. I started to envision making an impact on the real world through my research. The value of our work in the scientific community can only be actualized when our tools are adopted by downstream users such as domain experts and decision-makers. Hence alleviating user-database barriers is a vital step in advancing data-driven decision-making in a wide range of fields.

With that overarching goal in mind, I initiated a project to facilitate user interaction with databases by identifying the major stakeholders and their challenges when interacting with databases, and then mapped that to their needs. Database users often interact with databases via SQL query sessions. From our analysis, users pose a variety of SQL queries in sequence with changes in SQL keywords and query fragments such as tables and attributes. However, the existing approaches only consider queries individually and make recommendations based on query similarity and popularity. We presented a new approach to recommend query information by learning from the sequential knowledge exploration patterns of historical users. We modelled our query recommendation problem as a query prediction task and used sequence-to-sequence models to predict the next query. Supervised by Dr. Pottinger, this work led to *Sequence-Aware Query Recommendation Using Deep Learning*, submitted to VLDB'21. As the lead researcher, I identified knowledge gaps in the existing work, defined and scoped the research problem, analyzed the workload data, implemented the deep learning models, ran the experiments, discussed the results, and wrote the paper. Seeing a connection between my work and the quantifiable impact gives me a rush of excitement that I am contributing to help those real-world users in need. I found myself enjoying both scoping and solving open-ended problems and hope to further improve with additional formal training in graduate studies.

Future Work. All my experiences collectively shaped my research interests and motivated me to pursue graduate studies. Today, database systems provide a vital infrastructure to access high volumes of data in a variety of applications. Seeing the user-database barriers and the potential of data-driven decision-making in areas outside of CS (e.g., city planning and sustainability) incites my urge to build my work around the theme of facilitating user interaction with databases. With a deep understanding of the problem space and skills gained through solving problems in this space, I hope to continue this line of work by applying visualization and ML techniques to help database users access and make sense of data.

Specifically, I would be excited to work with Dr. Aditya Parameswaran and Dr. Joseph M. Hellerstein. Dr. Parameswaran has made outstanding contributions to building end-to-end systems with a focus to simplify data analytics for users. My work in provenance summarization extended his smart drill-down system to numerical attributes, and we built an interface to give users visualization of the summary. Following SIGMOD'20, I had an opportunity to chat with Dr. Parameswaran and was thrilled to learn that there is a good fit in our research interests. I would be excited to work with Dr. Parameswaran by combining our interests in making data more accessible to users through improved interfaces. My research interests also greatly overlap with Dr. Hellerstein's recent work (e.g., *Learning to Optimize Join Queries With Deep Reinforcement Learning*), on building data-driven systems to facilitate user-database interaction using ML-based techniques. I would have strong support to extend my work on query recommendation using ML-based techniques under his supervision.

Where I See Myself. Through these valuable experiences, I not only learned about the many real-world challenges that people face on the job, but also discovered research interests that would allow me to address some of those challenges. After graduate studies, I aim to pursue a career in academia, so that I can develop the research and tools to address these challenges and more. Furthering my education at Berkeley would bring me one step closer to my goal of advancing data-driven decision-making in a wide range of fields.