Statement of Purpose

Eugenie Y. Lai, UC Berkeley EECS PhD applicant for Fall 2021

I am a 6th-year Business and Computer Science Combined Major undergraduate at the University of British Columbia (UBC). The combined major means I am graduating with the same CS degree requirement as UBC's core CS majors with additional business courses. In the past six years, I did grad-level courses and two years of co-op with experiences in the industry, non-profit, and academia. Since May 2019, I have been working with Dr. Rachel Pottinger. My current research focuses on facilitating user interaction with databases. Specifically, I apply concepts of visualization and machine learning to alleviate the barriers between users and databases to help users access and make sense of data. I worked on two papers as the 2nd author, presented at SIGKDD'20, and am working on a paper as the lead researcher and 1st author.

My two 2nd-authored short papers, *Summarizing Provenance of Aggregation Query Results in Relational Databases* (ICDE'21) and *Pastwatch: On the Usability of Provenance Data in Relational Databases* (ICDE'20), focus on data provenance summarization. Our work contributes a new approach to provenance exploration that builds on data summarization techniques. We extended the smart drill-down system to aggregate queries and numerical attributes and built an interface to provide users with visualizations of the summary. I had been involved with this problem for a year and a half, including the ideation, the implementation of the backend system, and the experiments. We experienced a few submissions. Although I felt discouraged at first, I was fascinated by how much our work had improved after each round, which made me appreciate the value of reviewers' comments even more, especially with such an inherently long feedback loop in academia.

I was excited to co-present my Data Science for Social Good project, *Developing a Data-Driven Electric Vehicle Strategy in Surrey, BC, Canada*, at the SIGKDD'20 Social Impact Session. This work was done by a team of two undergraduates supervised by Dr. Raymond Ng. Without adequate tech support, the existing process to determine where to install an EV charging site was solely based on expert opinions, despite a large volume of data owned by the city of Surrey. In addition to an electric vehicle charging site utility model, I developed the visualizer to give the city planners a user-friendly way to interact with the data, including the spatial distribution and time trends of Surrey's vehicle stock and traffic flows. The city used our tools to choose 20 charger locations for a federal funding proposal in 2019. Through this project, I started to envision making the real-world impact as a chain effect. 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 my work on alleviating user-database barriers is a vital step in advancing data-driven decision-making in a wide range of fields.

I plan to submit my first 1st-authored paper, *QueryTeller: Sequence-Aware Query Recommendation Using Deep Learning*, to VLDB'21 by December 2020. We present a new approach to recommend query information by learning from the sequential knowledge exploration patterns of historical users. We leverage the sequential feature of query sessions using deep learning techniques. We model our query recommendation problem as a query prediction task and use sequence-to-sequence models to predict the content and structure information of the next query. Under the supervision of Dr. Rachel Pottinger, I identified knowledge gaps in the existing work, defined and scoped the research problem, analyzed the datasets, and implemented the deep learning models. I am currently running experiments, analyzing results, and writing the paper. Along the way, I acquired the ability to unstuck myself and learned that progress in research is not always linear. Although it is still easy for me to downplay the heavy lifting I did in thinking and planning before actualizing my ideas, I found that keeping consistent documentation and reflecting on what I accomplished help crystallize my learning and keep myself motivated.

These experiences have given me valuable research skills and have helped shape my future research direction. Today, database systems provide a vital infrastructure for users to access high volumes of data in a variety of applications, which makes both field-specific and database-related expertise required for users. Seeing the user-database barriers incites my urge to build my work around the theme of facilitating user interaction with databases, which greatly overlaps with Dr. Aditya Parameswaran's research interests in end-to-end systems with a focus to simplify data analytics for individuals. My work in provenance summarization also extends his smart drill-down system. Following SIGMOD'20, I had an opportunity to chat with Dr. Parameswaran. I was thrilled to learn that I would be a good fit for his lab as my current interests are also in the theme of human-in-the-loop data analytics.

Besides the technical skills, e.g., deep learning, software engineering workflow, data cleaning and exploratory analysis, I gained in the past two years in research, my biggest takeaway is that exploring my research interests is an on-going, iterative process, and I am most motivated when I see a connection between my work and a real-world problem. My goal for the next five years is to become an independent researcher in databases. Specifically, I am ready to further my training in finding and defining research problems, while when given a problem, I want to be able to ask intelligent questions to poke holes, identify if it is ill-defined, and make it right. By the end of my PhD, I aim to have a broad understanding of computer science, develop a long-term vision of my field, master a specific area in my research, and make real-world impact using my knowledge.