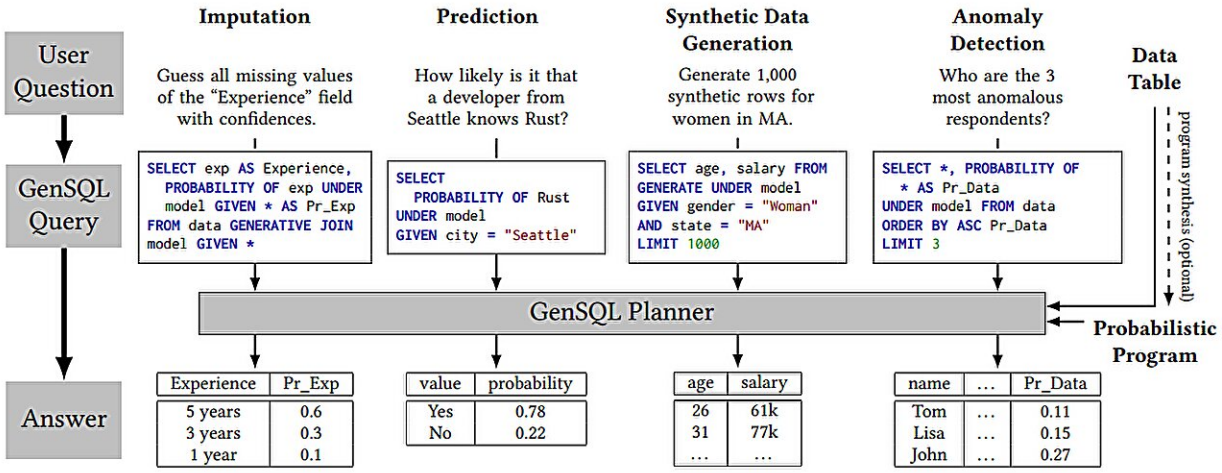
Researchers introduce generative AI to analyze complex tabular data

July 8 2024, by Adam Zewe



Overview of GenSQL. Credit: *Proceedings of the ACM on Programming Languages* (2024). DOI: 10.1145/3656409

A new tool makes it easier for database users to perform complicated statistical analyses of tabular data without the need to know what is going on behind the scenes.

GenSQL, a generative AI system for databases, could help users make predictions, detect anomalies, guess missing values, fix errors, or generate synthetic data with just a few keystrokes.

For instance, if the system were used to analyze medical data from a

patient who has always had high blood pressure, it could catch a blood pressure reading that is low for that particular patient but would otherwise be in the normal range.

GenSQL automatically integrates a tabular dataset and a generative probabilistic AI model, which can account for uncertainty and adjust their decision-making based on new data.

Moreover, GenSQL can be used to produce and analyze synthetic data that mimic the real data in a database. This could be especially useful in situations where sensitive data cannot be shared, such as patient health records, or when real data are sparse.

This new tool is built on top of SQL, a programming language for database creation and manipulation that was introduced in the late 1970s and is used by millions of developers worldwide.

"Historically, SQL taught the business world what a computer could do. They didn't have to write custom programs, they just had to ask questions of a database in high-level language.

"We think that, when we move from just querying data to asking questions of models and data, we are going to need an analogous language that teaches people the coherent questions you can ask a computer that has a probabilistic model of the data," says Vikash Mansinghka, senior author of a paper introducing GenSQL and a principal research scientist and leader of the Probabilistic Computing Project in the MIT Department of Brain and Cognitive Sciences.

The research is published in the journal *Proceedings of the ACM on Programming Languages*.

When the researchers compared GenSQL to popular, AI-based

approaches for data analysis, they found that it was not only faster but also produced more accurate results. Importantly, the probabilistic models used by GenSQL are explainable, so users can read and edit them.

"Looking at the data and trying to find some meaningful patterns by just using some simple statistical rules might miss important interactions. You really want to capture the correlations and the dependencies of the variables, which can be quite complicated, in a model.

"With GenSQL, we want to enable a large set of users to query their data and their model without having to know all the details," adds lead author Mathieu Huot, a research scientist in the Department of Brain and Cognitive Sciences and member of the Probabilistic Computing Project.

They are joined on the paper by Matin Ghavami and Alexander Lew, MIT graduate students; Cameron Freer, a research scientist; Ulrich Schaechtel and Zane Shelby of Digital Garage; Martin Rinard, an MIT professor in the Department of Electrical Engineering and Computer Science and member of the Computer Science and Artificial Intelligence Laboratory (CSAIL); and Feras Saad, an assistant professor at Carnegie Mellon University.

The research was recently presented at the ACM Conference on Programming Language Design and Implementation (PLDI 2024).

Combining models and databases

SQL, which stands for structured query language, is a programming language for storing and manipulating information in a database. In SQL, people can ask questions about data using keywords, such as by summing, filtering, or grouping database records.

However, querying a model can provide deeper insights, since models can capture what data imply for an individual. For instance, a female developer who wonders if she is underpaid is likely more interested in what salary data mean for her individually than in trends from database records.

The researchers noticed that SQL didn't provide an effective way to incorporate probabilistic AI models, but at the same time, approaches that use probabilistic models to make inferences didn't support complex database queries.

They built GenSQL to fill this gap, enabling someone to query both a dataset and a probabilistic model using a straightforward yet powerful formal programming language.

A GenSQL user uploads their data and probabilistic model, which the system automatically integrates. Then, she can run queries on data that also get input from the probabilistic model running behind the scenes. This not only enables more complex queries but can also provide more accurate answers.

For instance, a query in GenSQL might be something like, "How likely is it that a developer from Seattle knows the programming language Rust?" Just looking at a correlation between columns in a database might miss subtle dependencies. Incorporating a probabilistic model can capture more complex interactions.

Plus, the probabilistic models GenSQL utilizes are auditable, so people can see which data the model uses for decision-making. In addition, these models provide measures of calibrated uncertainty along with each answer.

For instance, with this calibrated uncertainty, if one queries the model

for predicted outcomes of different cancer treatments for a patient from a minority group that is underrepresented in the dataset, GenSQL would tell the user that it is uncertain, and how uncertain it is, rather than overconfidently advocating for the wrong treatment.

Faster and more accurate results

To evaluate GenSQL, the researchers compared their system to popular baseline methods that use neural networks. GenSQL was between 1.7 and 6.8 times faster than these approaches, executing most queries in a few milliseconds while providing more accurate results.

They also applied GenSQL in two case studies: one in which the system identified mislabeled clinical trial data and the other in which it generated accurate synthetic data that captured complex relationships in genomics.

Next, the researchers want to apply GenSQL more broadly to conduct largescale modeling of human populations. With GenSQL, they can generate synthetic data to draw inferences about things like health and salary while controlling what information is used in the analysis.

They also want to make GenSQL easier to use and more powerful by adding new optimizations and automation to the system. In the long run, the researchers want to enable users to make natural language queries in GenSQL. Their goal is to eventually develop a ChatGPT-like AI expert one could talk to about any database, which grounds its answers using GenSQL queries.

More information: Mathieu Huot et al, GenSQL: A Probabilistic Programming System for Querying Generative Models of Database Tables, *Proceedings of the ACM on Programming Languages* (2024). DOI: 10.1145/3656409