Probabilistic databases extend statistical inference from limited, hand-crafted statistical models to an entire database. Data analysts can discover trends, test hypothesis, and run what-if scenarios by simply running SQL queries. The technical challenge in a probabilistic database is the query processor, which needs to perform a probabilistic inference for every row output by a SQL query: the general-purpose probabilistic inference algorithms used in this step do not scale beyond small or medium-sized databases. Overcoming this limitation will require major advances in the optimization of probabilistic inference in databases. In this talk, I will describe one line of research in this direction, which relies on a combination of probabilistic views and safe queries. Like a traditional view, a probabilistic view is defined by a SQL query, and like a probabilistic database, its rows are random variables; their probabilities are computed offline, presumably at high expense. "Safe queries" are a restricted class of SQL queries for which the probabilistic inference can be done quite efficiently. The idea in this approach is to rewrite the user query as a safe query over the probabilistic views, thus benefiting from the probabilities that have been computed offline. This talk will give the necessary background on probabilistic databases, and describe some of the technical challenges associated to probabilistic views.

*Joint work with Christopher Re.*