DATABASE THEORY:
Where Has It Been?
Where Is It Going?

Philip A. Bernstein*
Harvard University

Database theory is the application of mathematical techniques to the solution of problems related to the design, implementation, and use of database mangement systems. It is not a tightly integrated mathematical theory. Rather, it is a collection of results on multifarious topics that are connected loosely or not at all. Some of these topics are data models, views, dependencies, power of query languages, dynamic storage structures, query optimization, concurrency control, recovery, security, and semantic integrity.

The field began with Codd's insight that relations and predicate calculus provide a powerful and easy to use interface to databases. Codd's application of a mathematical language and techniques to database management are the most important contributions to database theory. They gave the mathematical framework for much of the field.

Functional dependencies, also defined by Codd, were the first area of intense mathematical analysis. Dependency structures are still one of the most popular topics. Useful application has been achieved for database design and universal relation interfaces and, to a lesser extent, for view updating and query simplification.

View updating has benefited from a mathematical treatment that demonstrates that most useful views are not updatable. The theory also gives methods for proving that a translation of view updates into databases is correct.

There are a number of interesting results comparing the power of relational style query languages. For example, it is known that nestings of aggregate queries has the same expressive power as alternations of quantifiers, and that transitive closure is not expressible in relational calculus.

Some of the most popular dynamic storage structures, such as B-trees and dynamic (extendible) hashing, were developed with the help of analyses of the computational complexity of retrievals and updates.

There are several important theoretical results on query optimization: Wong and Youseffi's decomposition algorithm, query simplification using tableaux, and semijoin theory. However, many open problems remain, such as the optimization of queries containing aggregates, quantifiers, or disjunctions.

Concurrency control is among the most popular topics in database theory. It is difficult to reason intuitively about the correctness of concurrency control algorithms, so mathematical proofs have accompanied most proposed algorithms. Interesting algorithms have been analyzed for deadlock handling and synchronizing access to distributed replicated data.

More recently, there has been a burst of research in problems related to transaction recovery in the face of media and site failure. There are now correctness proofs for the major centralized recovery algorithms. In distributed recovery, analyses have focused on transaction commitment (e.g. two-phase commit) and the related problem of reaching agreement among processes in an unreliable computing environment (e.g. the Byzantine Generals problem). Work in this area is likely to continue for some time.

Other active areas of theoretical interest are security, semantic integrity, and applications of logic to databases.

As long as there are mathematically inclined researchers working on computer science problems, there will be research on database theory. It's a growing field; over 100 papers were submitted to each of the first two ACM SIGACT-SIGMOD Symposia on Principles of Database Systems (PODS). Not all of the reported results will be of earth-shaking importance. And more than one area of inquiry will lead to a dead end. However, as in the past, the field will have its share of major successes, some of which we can already classify as breakthroughs.

*Author's address: Aiken Computation Lab
Harvard University, Cambridge, MA 02138