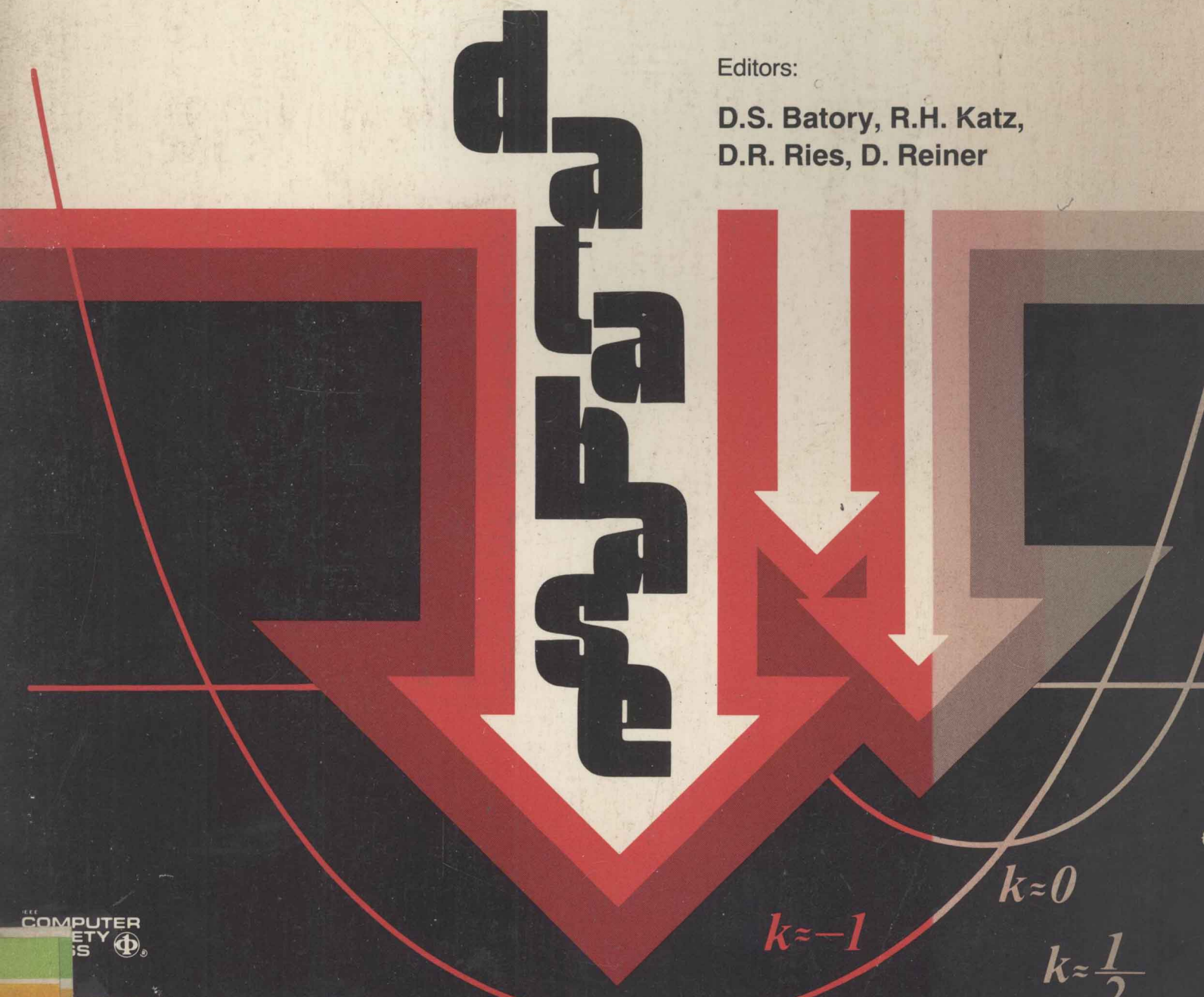
Database Engineering

Editors:

D.S. Batory, R.H. Katz,
D.R. Ries, D. Reiner



IEEE COMPUTER SOCIETY

ISBN 0-8186-0672-X
Library of Congress Number 84-102452
IEEE Catalog Number EHO 233-7
IEEE Computer Society Order Number 672

1984 THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

DATABASE Engineering

ISBN 0-8186-0672-X
Library of Congress Number 84-102452
IEEE Catalog Number EHO 233-7
IEEE Computer Society Order Number 672

COMPUTER SOCIETY PRESS

IEEE COMPUTER SOCIETY

1884 1984 A CENTURY OF ELECTRICAL PROGRESS

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

Published by IEEE Computer Society Press
1109 Spring Street
Suite 300
Silver Spring, MD 20910

ISBN 0-8186-0672-X
Library of Congress Number 84-102452
IEEE Catalog Number EHO 233-7
IEEE Computer Society Order Number 672

Order from: IEEE Computer Society
Post Office Box 80452
Worldway Postal Center
Los Angeles, CA 90080

IEEE Service Center
445 Hoes Lane
Piscataway, NJ 08854

1884 1984 A CENTURY OF ELECTRICAL PROGRESS

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

Database Engineering Bulletin is a quarterly publication of the IEEE Computer Society Technical Committee on Database Engineering. Its scope of interest includes: data structures and models, access strategies, access control techniques, database architecture, database machines, intelligent front ends, mass storage for very large databases, distributed database systems and techniques, database software design and implementation, database utilities, database security and related areas.

Contribution to the Bulletin is hereby solicited. News items, letters, technical papers, book reviews, meeting previews, summaries, case studies, etc., should be sent to the Editor. All letters to the Editor will be considered for publication unless accompanied by a request to the contrary. Technical papers are unrefereed.

Opinior.s expressed in contributions are those of the individual author rather than the official position of the TC on Database Engineering, the IEEE Computer Society, or organizations with which the author may be affiliated.

Membership in the Database Engineering Technical Committee is open to individuals who demonstrate willingness to actively participate in the various activities of the TC. A member of the IEEE Computer Society may join the TC as a full member. A non-member of the Computer Society may join as a participating member, with approval from at least one officer of the TC. Both full members and participating members of the TC are entitled to receive the quarterly bulletin of the TC free of charge, until further notice.

Preface

1984 Issues of Database Engineering

This book binds together the four 1984 issues of DBE. Don Batory led off the year with a special issue summarizing working group discussions at the Second International Workshop on Statistical Databases, held in September 1983 at Los Altos, California. Subjects treated included user interfaces, workstations and special purpose hardware, time series and econometric database management, special data types, models, and operations, metadata, and storage and implementation issues.

Randy Katz followed with an issue on Engineering Data Management, covering recent CAD/CAM work. This included a literature survey, descriptions of several large CAD systems, extending relational databases to handle CAD requirements, and performance questions.

In September came Dan Ries's issue on Multimedia Data Management. Topics were optical disks, spatial data management, and modeling and querying multimedia systems.

I put out the last issue of the year, on Database Design Aids, Methods, and Environments. This covered comprehensive design environments now under development, conceptual, logical, distributed, physical, update, and form-based design, early prototyping, and modeling transactions.

Editorial Staff of Database Engineering

Since its revival in 1981, Database Engineering has gained a reputation as a timely and carefully written publication, covering current research and development work in the database area. Won Kim, Editor-in-Chief through mid-1984, was an effective, hard-working, and knowledgeable editor and organizer. I took over fron Won with the September 1984 issue of DBE, hoping to continue its traditions.

One of the major strengths of DBE is its staff of Associate Editors, active researchers all, who are responsible for editing individual issues. Don Batory, Randy Katz, and Dan Ries, whose excellent issues are included in this compendium, have now moved on. As a fellow editor, I have appreciated their enthusiasm and breadth of expertise.

The current Associate Editors, and their issues for 1985, are:

Dr. Haran Boral, MCC, "DBMS Performance," March 1985.

Dr. C. Mohan, IBM Research, San Jose, "Concurrency Control and Recovery in DBMSs," June 1985.

Professor Yannis Vassiliou, Graduate School of Business Administration, NYU, "Natural Languages and Databases," December 1985.

Professor Fred Lochovsky, Department of CS, University of Toronto, "Object Oriented Systems and DBMSs," December 1985.

Gio Wiederhold is the new Chairperson of the Technical Committee on Database Engineering, replacing Bruce Berra as coordinator of TC activities. In addition to publishing DBE, the TC sponsors or cosponsors one or more conferences each year.

The orientation of DBE continues to be toward engineering aspects of databases, rather than abstract theory. Although submissions to DBE are not subject to a formal review process, the editors generally read articles carefully, and work with the authors to achieve both clarity and brevity.

My thanks again to the associate editors represented in this volume, and most certainly to the authors of the excellent papers in it. Chip Stockton of the IEEE Computer Society Press has been very helpful. Finally, I appreciate the support of Bea Yormark, SIGMOD Chairperson, who arranged to have copies of the December 1984 issue distributed to all SIGMOD members.

David S. Reiner

David S. Reiner
Cambridge, Massachusetts
February 1985

Table of Contents

September 1984, Vol. 7, No. 3 (D.R. Ries, Editor)

December 1984, Vol. 7, No. 4 (D. Reiner, Editor)

Letter from the Editor

This issue is about statistical databases. It contains tutorial articles that present concensus opinions on the current knowledge, problems, and anticipated research directions regarding statistical databases. This issue differs from previous issues of Database Engineering as its articles are not status reports of the current research of specific groups or individuals. Instead, it contains results of working group discussions which were held at the Second International Workshop on Statistical Databases (September 27-29, 1983). Among the workshop participants were experienced practitioners, leading researchers, and recognized pioneers in the statistical database field. The authors of the papers in this issue were usually the working group leaders; the opinions expressed in each article reflect the concensus of the working group and are not necessarily just those of the authors. To acknowledge the contributions of the group members, their names are listed at the start of each article.

The issue begins with a short note from John McCarthy and Roy Hammond, the general chairman and program chairman of the workshop, respectively. They describe the workshop context and give some insights about common themes that emerged from the workshop as a whole. Next is an overview paper that appeared in the workshop Proceedings. In it, Dorothy Denning, Wesley Nicholson, Gordon Sande, and Arie Shoshani present a concise introduction to the problems and research topics of statistical database management. As noted there and in other papers, statistical databases present problems and requirements that current data management and statistical software do not fully address. The subsequent articles represent summaries from individual working groups on the following topics: user interface issues, workstations and special purpose hardware, connecting heterogeneous systems, time series and econometric database management, special data types and operations, logical data models, metadata management, and physical storage and implementation issues. A keyword index is provided at the end of this issue to facilitate the cross-referencing of major topics.

Readers of this issue will be struck by the enormity of the problems that confront statistical database practitioners and researchers alike. Statistical database research, as a whole, is still in its infancy. Almost all of the major problems can be traced to an inadequate understanding of the fundamental needs and basic tools for statistical database management. It is hoped that this issue will contribute to the improvement of this understanding, and will stimulate further research and better solutions to the problems at hand.

Finally, I thank John McCarthy for his help, enthusiasm, and support. I also thank the other contributors of this issue for all the hard work they put in to make this publication possible.

D.S. Batory

December 1983
Austin, Texas

The Second International Workshop on Statistical Database Management: Common Themes and Issues

John L. McCarthy, General Chairman
Lawrence Berkeley Laboratory
Building 50B, room 3238
Berkeley CA 94720

Roy Hammond, Program Chairman
Statistics Canada, EPSD
2405 Main Bldg, Tunney's Pasture
Ottawa, Canada K1A0T6

1. Introduction

The Second International Workshop on Statistical Database Management was held in Los Altos, California, on September 27-29, 1983. One hundred computer scientists and statisticians from North America, Europe, and Japan attended. The workshop was sponsored by the Lawrence Berkeley Laboratory and the United States Department of Energy, in cooperation with the Association for Computing Machinery's Special Interest Group on Management of Data, the American Statistical Association's Statistical Computing Section, the IEEE Computer Society's Technical Committee on Database Engineering, and Statistics Canada.

2. Workshop Goals and Working Groups

Like the First LBL Workshop on Statistical Database Management, which was held in December, 1981, the Second Workshop brought together researchers and system designers from both computer science and statistics to discuss current work on problems of statistical and scientific database management. It was intended not only to facilitate interchange of ideas, but also to stimulate some systematic and collective thought about research directions. Although the purpose of the Second Workshop was the same as its predecessor, the format and content differed in a number of ways.

Participants came prepared to work in small groups and to produce the reports summarized in this publication. About half of the time was spent in parallel working group sessions, with each group composed of five to ten experienced practitioners from a variety of backgrounds. Members of each working group were asked individually and collectively to discuss and produce written summaries of questions that need to be addressed and promising research ideas in selected subtopics of statistical database management. Each working group then presented its conclusions at a plenary session to get comments from other participants prior to preparation of the summaries presented here.

We hope that these reports will help focus attention of the larger database community on some of the special problems of statistical database management. We are grateful to Don Batory, Won Kim, and IEEE's *Database Engineering* for providing a forum in which to report the results of our working groups.

3. Common Themes and Issues

At the First Workshop in 1981, definition of "statistical database management" and a working vocabulary that computer scientists and statisticians could both use were major issues. At the Second Workshop there seemed to be more agreement among participants on "statistical database management" and a common vocabulary. In addition, several new themes emerged from written contributions to the *Proceedings*, presentations at the workshop, and working group reports summarized in this issue. Four major ideas which recurred frequently were the importance of metadata, the need for richer semantics, the limitations of current relational systems, and the growing role of microprocessors.

First, there was a widespread recognition of the key role that meta-data, or data about data, can play in different aspects of statistical database management. Meta-data is necessary to specify information about statistical data for both human beings and computer programs. It can provide definition of logical models as well as more mundane documentation details for both database administrators and users. Well-defined and differentiated meta-data is necessary to permit software linkages between different logical and physical representations; between statistical databases, application programs, and user interfaces; as well as between multiple distributed and heterogeneous systems.

A second general theme was the need for richer semantics and operators for statistical data. A number of papers and group reports discussed the need to define and manipulate complex, multi-dimensional data structures. For example, many scientific databases require capabilities for defining and operating directly on vectors, time-series, and multi-dimensional matrices. There also seemed to be widespread agreement on the desirability of using self-describing files for both input and output, with functions automatically using and producing meta-data as well as data.

Although the relational model has become the standard for academic research, a number of groups and individuals noted serious limitations of current relational systems for statistical database applications. Of particular concern are the limited number of data types and operators for both data and meta-data. Some felt such limitations might be overcome by extending the relational model to include complex or abstract data types.

Finally, there was a growing recognition of the wide range of opportunities and challenges for statistical database management inherent in the microprocessor revolution. There is an accelerating trend towards transferring statistical data and meta-data from central databases to microprocessor workstations *and vice-versa,* with many attendant problems of distributed data management. High resolution terminals, large local memory and disk storage, fast local processing, and higher data transmission rates are bringing quantum changes in user interfaces and the way in which statistical analysts work. There promises to be an increasing emphasis on interactive graphical display of *pictures* as well as numbers and words for data, data models, meta-data, control options, and so on.

4. Proceedings for First and Second Workshops

Copies of papers, research reports, and issues outlines are available in *Proceedings* for the First and Second Workshops. For either, contact the Computer Science and Mathematics Department, Lawrence Berkeley Laboratory, Berkeley, CA 94720, or the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161.

5. Future Workshops

Preliminary planning has begun for a Third International Workshop on Scientific and Statistical Database Management in Seattle during the fall of 1985, Wes Nicholson and David Hall of Pacific Northwest Laboratories will organize the program and local arrangements. One topic that will probably get more emphasis at the Third Workshop is *scientific* database management, particularly for physical science data. Direct inquires to D. Hall, Math 1137/ 3000 Area, PNL, Box 999, Richland WA 99352; telephone (509) 375-2369.

Research Topics in Statistical Database Management

Dorothy Denning
SRI International

Wesley Nicholson
Battelle-Pacific Northwest Labs

Gordon Sande
Statistics Canada

Arie Shoshani
Lawrence Berkeley Labs

Abstract

This report identifies research topics in statistical database management. These topics are grouped into four major areas: characteristics of statistical databases, functionality/usage, metadata, and logical models.

1. Statistical Databases Characteristics

Computer scientists, especially designers of database systems, commonly ask statisticians and data analysts to identify the characteristics or features of a database that identify it as a statistical database. Searching for a profound answer to this question has perplexed data analysts. Many conclude that there are no characteristics which uniquely identify a statistical database. In principle, any collection of quantitative information residing in a computer is a candidate statistical database. As soon as the body of information is interrogated and statistically analyzed, either in total or by sampling or subsetting, it becomes a statistical database.

There are, however, important characteristics that should be built into a database if it is going to be useful for statistical analysis. These characteristics involve adequate description of the quantitative information in the database (i.e., the inclusion of appropriate metadata as defined in Section 3 below.). Such description is essential to understanding inferences evolving from data analysis. Certain kinds of description or definition are almost always included in the database because it is well known that the particular description is critical to understanding the data. On the other hand, certain other information is almost never included even though a detailed analysis will uncover subtleties that are correlated with such description and often cannot be modeled without it. A simple example will serve to illustrate the point. In a database of hospital records, the subject is always described as male or female. This description is important for prognosis and treatment. Periodic readings of blood pressure are also included in the database. On the other hand, the conditions under which the blood pressure was taken – patient lying down, standing up, sitting; recording made on the left or right arm – are almost never included. If the protocol dictates taking the blood pressure on the left arm with the patient lying down, then that information should be included in the database. If there is a variety of conditions, then each blood-pressure reading should be accompanied with a descriptor. When does such detailed information become important? When blood pressure is correlated with treatment protocol, we wish to minimize the random error in the measurements. Clearly if systematic changes in readings can be associated with the position of the patient or the arm on which the reading was made, then that random variability is reduced and a more precise statement can be made about the effect of a specified treatment.

There are distinct types of quantitative data that may be recorded in the database. For each type, there are general conditions which should be met if the information is to be described adequately for detailed statistical analysis.

1.1. Missing Data

Almost every statistical database has incomplete records. Proper statistical treatment of missing data usually depends on the reason for the missing data. For example, in a seismology file listing individual station seismometer magnitudes associated with particular earthquakes, values missing because a station was not operational should be ignored in an estimate of earthquake magnitude. On the other hand, values missing because the signal was either below the seismometer threshold or beyond the seismometer range and off scale, bound the magnitude of the earthquake and should be utilized in an estimate of earthquake magnitude.

As in the seismometer example, there are several possible reasons for a missing value. A set of tags to identify the particular type of missing value should be included in the file. In the seismology example, the tags would at least include "non-operational," "below threshold," and "offscale."

In some situations, such as with questionnaires, the logical structure may influence the interpretation of a missing value; e.g., whereas for males it is not important

whether a question on the number of pregnancies is answered, for females, it is critical to distinguish between a nonresponse and zero.

Most database management systems identify missing values but lack proper tagging capability. Research is needed to improve missing value treatment, and, in particular, to include sufficient information in retrievals so that missing values (either included or excluded) can be properly handled during data analysis.

1.2. Data Quality

Knowing the quality of data is important for statistical analysis. For example, if data are keyed into a file from a remote terminal, how frequently are typographical errors made? Are the data cross checked before being accepted? If data come from a measurement instrument, what is the resolution of that instrument? What is the reproducibility of independent measurements on that instrument? Has that instrument undergone modification during the time that the total set of data was collected? Or further, is that instrument recalibrated every day prior to data collection? These are all important questions; their answers may well influence the way the data are handled in any statistical evaluation. The file should include such data quality information. If the quality is uniform over the entire file, this information can be included in the file descriptor; if it varies in a haphazard fashion, it may be necessary to attach it to each datum.

Further considerations with respect to data quality involve the frequency of spurious measurements through either a breakdown in the data-generating system or the introduction of a rare physical phenomenon which grossly changes the measurement process. For example, in a chemical analysis for trace constituents a contaminant in the apparatus could cause major variation in the measurement. Here explanatory flags should accompany the data corroborating the presence of a contaminant or suggesting the possibility of a contaminant.

Finally, when data are collected over a period of time, there may be changes in the data-collection process; e.g., in the method of reporting, measuring, validating, or summarizing. To sort out such effects, a time stamp should be associated with each datum giving the time when the data were generated, and the time of the particular file update when the data were included.

In many situations it is useful to have a "degree of believability" associated with data. For example, economic data on developing countries may be obtained by estimates. Using such data for economic forecasts or evaluation should take into account the believability of the data. Another source of imprecise data is introduced by imputation. Imputed data values should be marked as such and not interpreted as reliable data.

Current database management systems do not have facilities for keeping track of data quality. Research is needed to find economical ways of storing information about data quality, and to find ways of passing this information to the data analyst.

1.3. Data Sparseness

In many data sets, there are structured patterns of missing data. This is particularly the case for designed experiments where the "design" is an optimum sparse coverage of the independent variable levels. Here the structure allows encoding which could materially reduce database storage requirements.

To reduce storage requirements, designers of databases often change the logical structure of the data. For example, a file may be partitioned into multiple segments, or data values (e.g., year) included with a data element name. This practice can obscure the meaning of the data and complicate retrieval.

Research is needed on the handling of sparse data to find ways to economize storage, to describe metadata, and to optimize retrieval while keeping the logical description independent of storage considerations.

1.4. File Freezing

Many databases are dynamic in the sense that they are continually being updated. If a statistical analysis is to be performed, there will be a natural time cutoff. All data resident in the file as of the cutoff point must be identifiable. Thus there must be a capability to segment on time so that information that comes in after the cutoff will not erroneously get into the statistical analysis and possibly bias the results. As a consequence of file freezing, there may be several versions of the same file in existence.

Research is needed to find techniques that impose proper time constraints on retrievals. Research is also needed to find techniques for efficiently storing multiple versions of large files.

1.5. Imprecise Keys

In statistical analysis, information may be needed from various parts of a single file or from several files. Often, this must be done by making a cross reference linkage using imprecise keys. For example, in a hospital database system, all the information on a patient might be retrieved using the patient's name as an imprecise key to search portions of the same file or several files (name is usually an imprecise key because there may be several people in a database with the same name). A file structure that allows cross referencing with such imprecise keys is very useful for statistical analysis. In statistical databases, subsetting and retrieval using imprecise keys is a difficult question that needs research.

1.6. Security

When a statistical evaluation is to be done on a file that contains sensitive information, the question of privacy protection arises. The confidentiality dilemma is to provide useful summary information while protecting the privacy of the individuals. Suitable mechanisms for protecting information may depend on the logical data model. Research is needed to determine what is obtainable within the constraint of summary information

criteria, and how to provide security mechanisms in a multiuser environment.

2. Functionality/Usage

Several issues were raised regarding the desired functionality or usage of statistical databases.

2.1. Subsetting

The key to successful data analysis lies in finding interesting subsets of the data. This requires the capability for multiple key retrievals or, more generally, for retrieval of any identifiable subset of data (e.g., all PhD's in the age bracket 25-40 living in California and earning more than $50,000 annually). Once a subset of data has been formed and analyzed, it is often desirable to retain the subset for further analysis, for aggregation, or for decomposition into smaller subsets. For example, the salaries for the preceding subset of PhD's may be aggregated by profession or by sex, or the subset of PhD's in the computer industry may be extracted for a more detailed analysis. Because subsets are obtained or retained for the purpose of aggregating or summarizing over certain attributes, they are often called summary sets.

Many commercial database systems have facilities for specifying and retrieving arbitrary subsets. The storage and retrieval mechanisms of these systems are not always efficient, however, for statistical database structures, e.g., sparse data. Research is needed to find efficient techniques for statistical databases; transposed files are a good beginning.

Some commercial database systems support view definitions, which permit subset definitions to be saved and managed by the database system. The data in a view is derived from the current state of the database when the view is retrieved, rather than being stored as a separate data set. With large statistical databases, views may not allow efficient enough access to certain subsets; hence, it may be preferable to store these subsets separately. Additional metadata is then needed for describing the subsets and their relationship to the main database. Research is needed to develop techniques for managing these retained subsets.

2.2. Sampling

In addition to forming identifiable subsets of data, it is often desirable to extract samples of the data. This is particularly true for large databases, where it may be infeasible or impractical to analyze the entire database. Sampling can also provide a means of protecting the confidentiality of sensitive data.

Most existing database systems do not support data sampling. Research is needed to develop efficient techniques for defining, retrieving, and retaining samples, and for combining sampling with other subsetting operators.

2.3. Data Analysis

Many existing database systems have operators for computing counts, sums, maxima, minima, and means. Although full data analysis capability should not be the goal of statistical database management systems (see Section 2.6), research is needed to determine which data analysis operators can and should be included in such systems. For example, it is quite efficient to perform the sampling operations in the data management system. In addition, new methods are needed for accessing complex data structures, e.g., hierarchies, by data analysis programs.

The results of data analysis should be self-documenting; that is, they should contain metadata describing the resulting structure. Existing systems do not provide this capability, and research is needed to develop analysis tools that produce self-documenting structures.

2.4. Adaptive Data Analysis

Data analysis is an adaptive process, where intermediate results determine subsequent steps in the analysis. It is often desirable to go back to an earlier step and try a different path. With appropriate computer graphics, much of the analysis could be done on-line without recourse to hard copy.

Existing database systems do not support this form of adaptive analysis. Research is needed to develop techniques for recording analysis paths, and to develop graphical aids for moving along these paths.

2.5. Historical Data

Traditionally, historical data has been difficult to assemble for analysis. If it is saved at all, it is usually archived on tapes. With on-line database systems, historical data can be retained and retrieved by the database system. Research is needed to determine how historical data is best managed.

2.6. Data Management and Statistical Analysis Interface

The data management software and statistical analysis software should not form a single monolithic system that attempts to provide all capabilities for all users. Even if we could predict what capabilities would be required, it would be difficult to develop and maintain such a monolith. On the other hand, the user interface should provide the image of a single system. The data management and statistical analysis capabilities should be constructed from building blocks that allow their easy interface. Research is needed to determine what building blocks are needed, and to develop a methodology for constructing and interfacing them. Several interfacing styles are possible; for example, the database system may drive the statistical analysis system or vice-versa, or both systems may operate as coroutines.

2.7. Distributed Systems

Local and nonlocal computer networks can provide access to distributed databases and to computing resources not available at the user's personal work station. Several scenarios are possible; for example, data from one or more sites may be assembled at a user's personal work station for analysis; data collected at different sites may be analyzed at the sites (e.g., to reduce the volume), and then transmitted to a central database system for further

analysis; data managed at a personal work station may be sent to a more powerful machine for analysis, and the results returned to the work station, possibly for additional analysis. Before any of these scenarios can be fully realized, research is needed to develop mechanisms for managing distributed statistical data and distributed analysis.

3. Metadata

Metadata is information about data. The panel has repeatedly emphasized the importance of metadata for statistical data. Often data becomes obsolete because the information about its content and meaning is nonexistent or lost. The following is a collection of metadata issues that could benefit from further research.

3.1. Meaning of Data

Most data management systems, as well as statistical packages, have a data definition capability for the specification of a data field descriptors such as type, size and acronym. This type of information is necessary for computer manipulation of the data. However, this information is not sufficient to characterize the meaning of the data to people. A description of the origin of the data, how it was collected, when it was generated and modified, and who is the responsible person for its collection is also needed. The description should include the full names of data entities and an explanation of what they represent. Data types of statistical databases are often complex, such as time series, vectors, or categorical variables. In addition, special types of data values may be required, such as codes for missing, unavailable, or suppressed values.

The lack of metadata is even more acute when data is collected through automatic data systems. Here it is necessary to be able to collect some of the metadata automatically as well.

3.2. Metadata of Subsets

As was mentioned in section 2, a large number of subsets can be generated in the data analysis process. In addition, new data values can be generated by computations over previous data values. The metadata for these newly created data sets include the origin from which the data sets were obtained, the operations (selection, sampling, computations) involved, descriptions of the data elements, who created the data sets, and time of generation.

Most of this information can (and should) be automatically obtained by the system at the time of subset creation. Some additional semantic information must be obtained from the user if he wants to keep these data sets for future use. The open research issues are how to capture and store this information efficiently. In particular, if data sets are generated from each other, they would have much descriptive information in common that should not be stored repeatedly.

3.3. Metadata Management

It is necessary to organize and manage metadata, just as it is the case with data. However, metadata typically contains much text, and its structure can be more complex than just text strings. It is therefore necessary to manage metadata with tools that can handle text. Most data management systems and statistical packages have very limited capabilities in this area.

One should be able to retrieve and search metadata, just as one does with data. For example, it should be possible to ask the system for the data sets generated by John Smith after February of this year, or to search for all data sets that have information about a certain topic in a hierarchical fashion. Research is needed to determine how to organize the (mostly) textual information so that it can be searched, retrieved, updated, and automatically maintained.

3.4. Consistency

Unfortunately, the meaning of terms change over time, and they may be inconsistent across data sets. This occurs often when similar data is collected over long periods of time. For example, the boundaries of a county may be redefined in a certain election year, but the change is not reflected in the name of the county. Clearly, it is invalid to compare data collected for that county over several years which include the change, yet it is commonly done because the corresponding metadata does not reflect the change.

Another reason for confusion is the use of the same terms for different data elements. This occurs often when new data sets are generated from existing ones. For example, one data set may contain information about income generated by an average over the entire set, while another may be generated by averaging over a sample. If both data elements are labeled the same (e.g. income), it is easy to make mistakes in comparing them. These changes should be captured in the metadata, and be readily available when the data sets are used. At the same time there should be a way to indicate that the data elements are related.

The reverse problem is one of using different terms for the same data element. It is particularly important if the same data element, such as "state", is used by more than a single file, since this information is necessary to determine if the files are comparable (joinable) over this data element. Using different terms in the same file requires the support of a synonym capability.

Another related need is the use of metadata for comparing or merging data from data sets whose parameters are similar but not identical. For example, suppose that the partitioning of ages into age groups in two data sets is not the same. In order to compare or merge these data sets on the basis of age groups, one needs the metadata describing the age groups.

3.5. Reformatting

It is not realistic to assume that at some point there will be a standard for data formats over all systems. Therefore, the need for reformatting data is inevitable. Metadata should be used to facilitate the automatic reformatting of databases. Research is needed to determine how to organize the metadata and how to use it for the purpose of reformatting. Perhaps a standard for metadata

specifications can be developed.

3.6. Distributed Data

There is additional metadata that is necessary when databases are distributed over several nodes of a computer network. For example, suppose that data is collected and analyzed at several hospital nodes on patients response to a certain drug. If one was to combine such information, it is necessary to synchronize the state of these databases as well as the correspondence between the items involved. Research is necessary to determine what status information should be kept, and how to coordinate such information for queries that involve several nodes.

There is very little development of distributed systems that can handle statistical data, mainly because the difficulties in implementing such systems seem too great. But, as was discussed by many members of the panel, the trend is indeed towards distributed systems of work stations. As powerful personal work stations come down in price, so it is more likely that future data analysis will be performed on a work station that is connected to other work stations and central machines through a computer network. The central machines are likely to contain data that are of interest and are shared by many users, while the work stations will contain temporary or private data sets that analysts currently work on. Thus, we believe that it is not too early to conduct research in the area of metadata in distributed systems.

4. Logical Models

Logical modeling is that part of database management concerned with the meaning of data collected about the real world. The typical logical model encountered in a statistical textbook is the rectangular array or observation on a case by attribute basis. The current status is that the real world is more complex than the logical models of database systems, but that logical database models are more complex and diverse than the logical models handled by standard statistical algorithms.

4.1. Complexity of Data

The data organizations encountered in statistical textbooks are data matrices or contingency tables. The mathematical machinery used is the matrix and vector algebras or calculus. The traditional interface with computer science has been the numerical analysis of the computational processes needed to implement the arithmetical processes.

When the data becomes more complex, of which the hierarchical relationship of individuals to a family is an example, differing information is relevant in different subsets of the data, and the classical notations quickly loose their elegance and power. In complex situations, the identification of an appropriate unit of analysis, and the collection of data for that unit, may become substantive problems. All of this may have the additional complication of missing and erroneous values. The notation needed to deal with other types of relationships, such as networks, is often weak and has weak associated theory. With complex data structures, the interface with computer science grows to include algorithms and data structures, computational complexity, and database management.

4.2. Missing Data

A common characterization of complex situations is the need to use and identify insightful subsets. In the presence of missing and erroneous data, this may be difficult. The missing data may arise for many reasons - not observed and not defined or relevant are the standard cases. The ability of database systems to approximately deal with the various types of missing data is weak in current practice. The initial machinery typified by the not-a-number symbols (NaNs) of the IEEE floating point standard have not been expanded or integrated into control mechanisms (query languages) of database systems.

4.3. Data Aggregation

The various attributes of data may be more complex than is realized. Hierarchical relationships may be multifaceted in practice. For example, in geographic aggregations, the notion of county and metropolitan area are intermediate between municipality and state and of equal standing; either may be embedded in a strict hierarchy. The form of the aggregation may change over time so that both analysis and representation are further complicated. Simple responses may be either multiple or repeated in practice. The representation of complex data which has been fully and correctly observed is now possible, but the methods to deal with partially or incorrectly observed data have not been developed.

4.4. Documentation

The logical data model is part of the description of the data and should be included in the documentation of the data. The metadata has the role of communicating both the internal technical facts about the data, including the data models used in its representation, and the external information available about the data. The meaning of the data may be derived both from the data models and the external knowledge about the data.

Logical data models should be associated with good analysis methods. The models that are available await analysis techniques, some of which may arise in the interaction of statistics and algorithm design. Some of the known problems with existing models are the identification of appropriate analysis units, and the bringing of data to those units. The current algorithms often are weak in the presence of the various forms of missingness and errors present in data.

Acknowledgements

Mervin Muller joined some of our discussions, and we are grateful to him for sharing with us his experience and insight.

References

There is an extensive literature covering the different aspects of statistical databases and statistical software. Instead of giving a long list of references, we mention a few surveys and collections of papers, all of which con-

tain many references.

Reference 1 below is an introductory paper to the area of statistical databases. It discusses several problem areas and surveys some existing solutions and work in progress. Reference 2 discusses extensively metadata structures and needs. Reference 3 discusses the security aspects of statistical databases, and surveys existing and proposed controls. Reference 4 contains numerous papers and abstracts presented at a specialized workshop on statistical database management. Reference 5 is a large volume that describes and compares statistical packages and other noncommercial statistical software. Reference 6 is the proceedings of an annual conference that has been held over the last 15 years, and that contains (especially in the more recent issues) several papers on statistical databases.

1. Shoshani, A., Statistical Databases: Characteristics, Problems, and Some Solutions, *Proc. Eighth International Conference on Very Large Data Bases*, Sept. 1982, pp. 208-222. (Copies available from: VLDB Endowment, P.O.Box 2245, Saratoga, CA 95070.)

2. McCarthy, J.L., Metadata Management for Large Statistical Databases, *Proc. Eighth International Conference on Very Large Data Bases*, Sept. 1982, pp. 234-243. (Copies available from: VLDB Endowment, P.O.Box 2245, Saratoga, Ca. 95070.)

3. Denning, D.E. and Schlorer, J., "Inference Controls for Statistical Databases," *IEEE Computer*, (to appear July 1983).

4. *Proceedings of the First LBL Workshop on Statistical Database Management*, Dec. 1981. (Copies available from: Computer Science and Mathematics Dept., Lawrence Berkeley Laboratory, Berkeley, Cal. 94720.)

5. Francis, Ivor (Editor), *A Comparative Review of Statistical Software*, 1977. (Copies available from: The International Association for Statistical Computing, 428 Princes Beatrixlaan, 2270 AZ Voorburg, Netherlands.)

6 *Proceedings of the Computer Science and Statistics: Annual Symposium on the Interface.* (Copies available from different places, depending on the year of the symposium.)