

PROJECT DESCRIPTION

1 Rationale for the Project

At the same time as Federal statistical databases continue to grow in complexity and size, more and more users are accessing the information they contain, for a multitude of purposes. Since much of the data is gathered under pledges or legal requirements of confidentiality, there is a clear trade-off between privacy protection and confidentiality on the one hand and user access to high-quality statistical data on the other [31]. To enable safe, effective dissemination of information from Federal databases, three interacting, data-driven challenges must be dealt with *now* — data confidentiality, data quality and data integration.

Responding to these challenges, we propose a large-scale, cross-disciplinary research and development effort to create abstractions, theory, implementable methodology and software prototypes that will be applied to actual statistical databases. The research confronts *simultaneously*:

Data confidentiality,¹ together with accompanying issues of privacy of the individuals and establishments that are data subjects. Inability to assure privacy increases the chance that systems with important societal benefits will be delayed, implemented excessively conservatively or even abandoned.

Data quality,² which is increasingly acute as huge amounts data are collected and new dissemination systems allow and even encourage data to be put to uses for which they were not originally intended.

Data integration,³ which raises deep new questions about its effects on confidentiality and quality.

The research involves fundamental questions in statistical and computer science, as well as issues of implementation and scalability of methods. The ultimate result will be Federal dissemination systems that serve both citizens and the government well.

To achieve the goals and impact set forth in §2, we have assembled an integrated, cross-disciplinary team of statistical and computer scientists from the National Institute of Statistical Sciences (NISS), Carnegie Mellon University (CMU), the University of Maryland College Park (UMd), the Institute for Social Research (ISR) at the University of Michigan (UMi), Purdue University (Purdue), Southern Methodist University (SMU) and the Los Alamos National Laboratory (LANL). Five leading Federal statistical agencies — the Bureau of Labor Statistics (BLS), Bureau of Transportation Statistics (BTS), Census Bureau (Census), National Agricultural Statistics Service (NASS) and National Center for Education Statistics (NCES) — are partners in the project. Through access to data and participation in development and evaluation of methods and software systems, they will help ensure that the research is relevant, timely and applicable.

Research advances in multiple disciplines necessary to confront difficult issues of data confidentiality (DC), data quality (DQ) and data integration (DI) are the thrust of the proposal. **Computer science** will formulate abstractions and algorithms for dissemination systems that accommodate interactions among DC, DQ and DI. The **statistical sciences** will provide decision-theoretic formulations that account for both disclosure risk and the utility of disseminating information, models of the processes that affect DQ, and characterizations of the consequences for inference of DC, DQ and DI. Domain knowledge, particularly from the **social sciences**, will link uses of information to requirements for DC and DQ. **Visualization** tools will enable operational understanding of the abstractions, as well as support user interaction and evaluation of the

¹Protection of data subjects from both identity disclosure [15] and disclosure of sensitive attributes, such as health status.

²The capability of data to serve multiple uses and users, with dimensions that include accuracy, accessibility, completeness, timeliness and interpretability.

³The combining of related data from multiple databases, often assembled by different organizations for different purposes.

usage and performance of the systems. **Software and systems engineering** will design and build prototype systems that operate at realistic scales, and support evaluation and refinement of theory and methodology.

The disciplines intersect throughout, but especially sharply in the cross-cutting issues of *complexity* and *scalability*. Many techniques for ensuring DC, improving DQ and performing DI are untried for the size and dimension of the databases, the diversity of user needs and the range of analyses that we will address. One central research challenge, indeed, is to *build systems that work*, implementing correct solutions to technical problems for the large, complex databases maintained by Federal statistical agencies.

2 Setting, Goals and Impact of the Research

Setting. The project will address research issues lying within the three-way intersection of DC, DQ and DI, as shown in Figure 1, as well as the pairwise intersections. It builds strongly on the current NISS digital government (DG) project *A Web-Based Query System for Disclosure-Limited Statistical Analysis of Confidential Data* (see §6), but goes far beyond that project in terms of scope — adding DQ and DI to DC, scale — with more, and more varied, researchers drawn from more organizations, and depth.

The underlying issues drive the research in different ways.

Data Confidentiality is mature scientifically, but, as evidenced by the National Research Council (NRC) report [71], electronic dissemination has created both challenges (means to break confidentiality are available widely and inexpensively) and opportunities (computationally intensive risk evaluation and reduction techniques [18, 19] have become scalable). The most pressing need is problem formulations and tools that scalably accommodate both *disclosure risk* and the *utility* of disseminating information derived from data.

Data Quality. Concern about DQ is widespread; however, as a field, DQ lacks fundamental quantifications, usable models and scalable methods [63]. These we will create, focused by partner agency needs and the perspective that *data are a product*, with users, uses, costs, value and quality [37, 63, 70, 97].

Data Integration. Tools to integrate (synonymously, “fuse”) data from multiple databases exist in profusion, standards are being created [68, 102] and rapid development is proceeding. The fundamental gap to be addressed is lack of understanding of the consequences of DI for DC, DQ and statistical inference.

The research we propose lies in the interactions among the three issues. To illustrate, poor DQ actually improves DC: for example, adding noise to data (either bias or increased variability) reduces the probability of re-identification [50]. To some, the error rate of approximately 10% in the 1990 US Decennial Census gave better protection to Census Bureau releases than data swapping alone would have provided [2, 9, 39]. Conversely, improved DQ threatens DC: decreased error rates of approximately 6% in the 2000 Decennial Census add concerns to the Bureau’s re-examination of its disclosure limitation strategy [1, 12].

Goals. Currently, however, the abstractions, quantifications, methods and software tools necessary to assess tradeoffs between DC and DQ do not exist. Nor can we yet describe or study, either theoretically or for real databases, how different methods for DI, such as record linkage, affect DC and DQ. The goal of the research is to fill these and other, equally significant, gaps.

To meet partner agency needs, the research will emphasize equally (1) Development of new abstractions, theory and methodology; (2) Construction of prototype systems implementing them; and (3) Use of the prototypes as testbeds to evaluate and refine theory and methodology, and to achieve scalability.

Federal government-unique problems define and inform the research. Beyond DC, DQ and DI, we will address multiple, diverse user communities, including hard-to-serve users, and dissemination constraints, such as lack of persistent cookies, that inhibit user service and understanding of system usage.

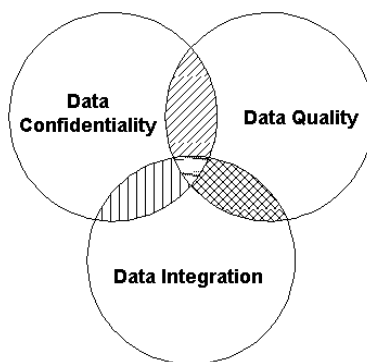


Figure 1: *Scientific Organization of the Project*. The research addresses primarily issues in the pairwise DC–DQ intersection and the three–way DC–DQ–DI intersection.

Impact. Political attention to privacy and increasing research on the subject [91] underscore the significance of the project. Statements by some, including Congressman Edward Markey, that privacy is the “civil rights issue of the decade” [56] may be rhetorical, but without credible ways to ensure DC in the face of strong concern about DQ and the growing need and capability for DI, the risk is real that Federal data warehouses will become data cemeteries instead. That similar challenges must also be faced elsewhere, especially in E-commerce and electronic medical records [31, 80, 91], extends the impact.

The research points to a paradigm shift, in which statistical analyses and other information derived from Federal databases are disseminated as a complement to, or even in place of, (sanitized) microdata records themselves.⁴ However, use of the results of statistical analyses as the medium of exchange, especially for research purposes, is controversial. For example, researchers will always want not merely the estimated coefficients in a linear model, but also to use various residual analyses to check assumptions and model specifications, even though residual analysis involving individual cases can compromise confidentiality.

The project will create the concepts and tools that enable the partner agencies to address these kinds of tensions, and allow the paradigm shift to occur sensibly and efficiently.

3 Research Program

Suppose that information is to be disseminated whose derivation requires integration of databases \mathcal{D}_1 and \mathcal{D}_2 . For example (see §3.1), \mathcal{D}_1 may contain confidential household-level economic and energy consumption data, while \mathcal{D}_2 contains data on travel behavior, and the information sought is the relationship among economics, energy use and travel.

To release information derived by analysis of an integrated database $\mathcal{D}_1 \bowtie \mathcal{D}_2$ requires understanding both the confidentiality properties and the quality of $\mathcal{D}_1 \bowtie \mathcal{D}_2$.⁵ The project addresses the abstractions, quantifications, techniques and prototype software systems needed to carry out this process, enveloped by the pervasive issue of scalability. We begin with the DC–DQ intersection in Figure 1, because DI is in many senses more a compounding factor for DC–DQ issues than an independent dimension.

⁴This concept is part of the access plan for Luxembourg Income Study; see <http://lisweb.ceps.lu/dataaccess.htm>.

⁵We use the join symbol \bowtie to denote all forms of DI.

3.1 Data Confidentiality – Data Quality Linkages

There is a strong confidentiality–quality duality. Many techniques, such as record linkage — a method for DI [42, 43, 54, 100], used to break confidentiality can also be employed to improve DQ. Conversely, strategies to protect DC often have a deleterious effect on DQ. For example, the 1997 Residential Energy Consumption Survey (RECS) [36] conducted by the Energy Information Administration (EIA) includes demographic data collected on households as well as information on their energy consumption. In addition to removing direct identifiers, EIA “masked” the data in the public use file by adding noise independently to several key consumption variables, potentially distorting (to some users, central) relationships between energy consumption and household characteristics.

Moreover, promised confidentiality is essential to DQ, especially for surveys and self-reported data, although this effect has never been quantified. The issues can be subtle: in drug abuse surveys respondents must be protected not just from “external” intruders but even from members of their own families.

3.1.1 Data Quality Foundations

Several foundation issues for DQ will be confronted.

Abstractions and Quantification. Data quality can be abstracted into such diverse attributes as accuracy, credibility, accessibility, consistency, relevance, timeliness, interpretability and confidentiality [37, 59, 63, 93, 98]. The first step will be to operationalize these abstractions as quantified *metrics for data quality* at multiple resolutions — individual records, tables within databases, entire relational databases and integrated databases (§3.2.2).

Detection and Characterization of Data Anomalies. Anomalous, and therefore potentially low-quality, data are widely prevalent. For example, NISS work on the EPA’s Toxic Release Inventory (TRI) [38] has identified anomalies that include abrupt changes, especially order–of–magnitude decreases, in otherwise “flat” data series, gaps in series, systematic non-changes over time ($y_t = y_{t-1} = \dots$) and suspiciously systematic changes ($y_t = \{1, .5, 2/3, \dots\} \times y_{t-1}$).⁶ Anomalous survey data may represent “socially correct” responses to questions or be deliberately incorrect, especially if confidentiality is not perceived as assured.

Methods to detect and characterize anomalies will be developed that are *automatic* and *scalable*. These link forward to techniques for statistical inference that accommodate the characterized anomalies. They also link backward to models of data generation processes and ultimately to techniques to design data collection processes that assure DQ. Statistical methods for outlier detection [13, 53, 85], while relevant, are not well-suited to anomaly detection because they typically focus on extreme data values rather than aberrant patterns. Therefore, the techniques to be created will draw as well on pattern recognition concepts from computer science [28, 83]. The methods will also capture domain knowledge of diverse types.

Models for Data Quality. One strategy to increase DQ is to “build it into the processes” by which the data are generated. As for industrial products, this is more cost-effective than re-work of data after they have been collected. To underlie such strategies, and to relate DQ to DC and DI, we will develop *models for data generation processes*.

Special attention will be devoted to the central role of people in such processes, and the models will involve significant descriptions of human behavior. To illustrate, cognitive and social processes influence the answers to survey questions: the wording and ordering of questions affect respondents’ answers [92]. For example, in the 1990 and 2000 US Decennial Censuses, a question on race was followed by a question

⁶The TRI epitomizes use of data for purposes other than the original one: it was initiated *solely* to force public disclosure of toxic releases, but is now used in contexts ranging estimation of trend to environmental justice.

on Hispanic origin. Tests in which the order of these questions was reversed showed significantly decreased use of the racial category “other” by those who identified themselves as Hispanics [3, 73]. Changing the race question in 2000 to allow respondents to choose more than one category produced surprising numbers of multi-racial respondents [3]. The models will not only represent phenomena at this level of intricacy but also be sufficiently detailed to reflect multiple uses of data (for example, administrative use as opposed to inference), as well as accommodate DC and DI concerns.

3.1.2 DC – DQ Foundations

Because DQ depends strongly on the use of the data or information, measures of the *utility* of released data or information will play a central role in relating DQ to DC. At the same time, several foundation issues for DC will be addressed.

To date, most work on disclosure limitation has focused on reducing disclosure risk [41, 99], ignoring the utility — to the public, the agency or even the data subjects — of disclosing information. Research has begun on decision-theoretic formulations that accommodate both risk and utility; see §6 and [87]. Trotini’s approach [94, 95] extends that of Duncan and Keller–McNulty [32, 64] by recognizing the differing perspectives of the statistical agency, the users, and intruders in assessing the extent of disclosure and the quality of the inferences associated with different forms of data release.

To formulate the problem, given a database \mathcal{D} , the abstraction for *released information* (for instance, in response to a query to a Web-based system; see §3.3) is a function $\text{Rel}(\mathcal{D})$. Examples are the results of statistical analyses of \mathcal{D} , “sanitized” transformations of \mathcal{D} by removal of identifying attributes, top-coding and other procedures that mask extreme attribute values and swapping of attribute values between data elements [21], statistical models of \mathcal{D} , and virtual data (§3.2) synthesized from a model of \mathcal{D} .

Inference as a Measure of Data Quality. Statistical inference is a principal use of databases, especially for policy and research purposes. Consequently, the accuracy (equivalently, uncertainty) of inference is a central measure of DQ, and use of it as a metric for DQ provides a path to investigate statistical implications of disclosure limitation strategies, as well as DI methods such as record linkage.

One approach, as in [69], will be strongly Bayesian, using Markov chain Monte Carlo (MCMC) [52] to characterize what other underlying data could have yielded the same released information. This permits characterization of the uncertainty (as noted above, a metric for DQ) in statistical analyses that protect DC. This research will also build strongly on existing Bayesian approaches to record linkage [7, 51, 67].

A similar formulation yields techniques to evaluate and optimize disclosure limitation strategies. Let $f_{\text{low-dim}}(\mathcal{D})$ be a set of low-dimensional characteristics of \mathcal{D} that can be released, and let $f_{\text{high-dim}}(\mathcal{D})$ be a set of high-dimensional characteristics to be protected. For example, $f_{\text{low-dim}}(\mathcal{D})$ may be a low-order cross-tabulation from a large contingency table, and $f_{\text{high-dim}}(\mathcal{D})$ information about small cell entries. Then the released information will be chosen to solve the optimization problem

$$\begin{aligned} \min & d(f_{\text{low-dim}}(\mathcal{D}), f_{\text{low-dim}}(\text{Rel}(\mathcal{D}))) \\ \text{s.t.} & d(f_{\text{high-dim}}(\mathcal{D}), f_{\text{high-dim}}(\text{Rel}(\mathcal{D}))) \geq \alpha, \end{aligned} \tag{1}$$

where d is an appropriate distance measure and α is a risk threshold. We must also protect against inadvertently disclosing other high-dimensional characteristics $f'_{\text{high-dim}}(\mathcal{D})$, which might be discovered, e.g., by data mining. In addition, dynamic versions of (1) will be developed that account for previous releases.

To illustrate how these kinds of techniques would be applied, an intruder could attempt to break confidentiality by developing a distribution over the possible high-dimensional characteristics that are consistent with the released low-dimensional characteristics. In the context of tables of counts, a natural entry point to

such induced Bayesian distributions is NISS work on exact distributions given marginals [23, 24, 49]. There has yet to be a full Bayesian formulation of the problem, however, which is needed, in particular, in order to update risk assessments given the release of additional low-dimensional marginals.

More General Risk–Utility Formulations. An initial formulation is similar to (1). The released information is a statistical model \mathcal{M} (for example, estimated coefficients and standard errors in a linear regression) of the database \mathcal{D} . Let $\mathcal{R}(\mathcal{M})$ and $\mathcal{U}(\mathcal{M})$ represent the risk and utility of releasing \mathcal{M} in lieu of the data. For example, \mathcal{M} may be a parametric model, with $\mathcal{R}(\mathcal{M})$ and $\mathcal{U}(\mathcal{M})$ the accuracies with which two of the parameters can be estimated from the model. The first parameter is to be protected, while the second conveys releasable, useful information. Then, \mathcal{M} may be determined by solving the optimization problem

$$\begin{aligned} \max \mathcal{U}(\mathcal{M}) \\ \text{s.t. } \mathcal{R}(\mathcal{M}) \leq \beta \end{aligned} \tag{2}$$

of maximizing the utility of the released model, subject to a prescribed upper bound on the risk. (This problem is more natural than the dual problem of minimizing risk subject to a lower bound on utility, but the latter will be considered as well.) The work will build on [32, 64], where *RU frontiers* are calculated, for some parametric models, that quantify the tradeoffs between risk and utility.

Risk–Quality Formulations. A principal challenge is to extend formulations such as (2) to databases themselves, with the utility $\mathcal{U}(\mathcal{M})$ replaced by a quality metric $\mathcal{Q}(\mathcal{D})$. Indeed, risk–quality tradeoffs are the epitome of risk–utility tradeoffs, since improvements to DQ often endanger DC.

By analogy to (2), we will address the following formulation: given a database \mathcal{D} , released information $\text{Rel}(\mathcal{D})$ and a context-dependent DQ metric \mathcal{Q} , an optimal quality improvement q^* of \mathcal{D} would satisfy

$$\begin{aligned} q^* = \arg \max_q \mathcal{Q}(q(\mathcal{D})) \\ \text{s.t. } \mathcal{R}(\text{Rel}(\mathcal{D})) \leq \beta. \end{aligned} \tag{3}$$

By incorporating models of the impact of DQ [82] as a function of quality metrics, measures of risk and uses of the data, we will develop strategies that map the use of data and information onto requirements for DC and DQ. These indicate the levels of DC and DQ compatible with the desired impact, leading to standards for Federal databases that accurately reflect user needs.

3.2 Data Confidentiality – Data Quality – Data Integration Linkages

We first address the effects of DI on DC and DQ, then the three-way interaction among DC, DQ and DI.

3.2.1 Confidentiality Consequences of Integration

Methodology will be developed and evaluated that relates the disclosure risk for an integrated database to that of the components *and* to the method \mathcal{I} used to effect the integration. Such methodology will be embodied, initially, in relationships of the form

$$\mathcal{R}_{\{1,2\}}(\mathcal{D}_1 \bowtie \mathcal{D}_2, S) = f(\mathcal{R}_1(\mathcal{D}_1), \mathcal{R}_2(\mathcal{D}_2), S, \mathcal{I}), \tag{4}$$

where S is the statistical analysis whose results are to be released.

In (4), \mathcal{I} is the method by which \mathcal{D}_1 and \mathcal{D}_2 are integrated. Specific examples on which we will focus are exact and statistical record linkage [22, 54, 88, 101]. In most cases, the former will have more severe DQ consequences; the challenge is to quantify these. Approaches to record linkage from the computer science

literature raise further concerns about confidentiality [91], which will be incorporated into the statistical framework we are proposing.

Apportioning $\mathcal{R}_{\{1,2\}}(\mathcal{D}_1 \bowtie \mathcal{D}_2, S)$ in (4) between \mathcal{D}_1 and \mathcal{D}_2 will be approached computationally through sensitivity analyses that in effect estimate “partial derivatives” of f with respect to its arguments.

3.2.2 Quality Consequences of Integration

We will construct techniques to characterize DQ for integrated databases [5]. The RECS data described in §3.1 are illustrative: some users wish to link them to household data from surveys carried out by the BTS on energy consumption related to travel.

The entry point will be to characterize the quality of the join of two tables \mathcal{T}_1 and \mathcal{T}_2 in the same relational database, through relationships of the form

$$\mathcal{Q}(\mathcal{T}_1 \bowtie \mathcal{T}_2) = f(\mathcal{Q}(\mathcal{T}_1), \mathcal{Q}(\mathcal{T}_2)), \quad (5)$$

where \mathcal{Q} is a DQ metric (§3.1.1).

More challenging issues will be confronted when distinct databases are integrated, as in the BTS’ Intermodal Transportation Database (ITDB) [10]. Multiple DQ metrics may be involved, and (5) must be generalized to take account of the method \mathcal{l} by which the data are integrated:

$$\mathcal{Q}_{\{1,2\}}(\mathcal{D}_1 \bowtie \mathcal{D}_2) = f(\mathcal{Q}_1(\mathcal{D}_1), \mathcal{Q}_2(\mathcal{D}_2), \mathcal{l}). \quad (6)$$

For the setting of §3.1, the implications of DI for statistical questions of inference and characterization of uncertainty will be explored by means of the Bayesian approach described there, providing partner agencies fundamental insight into DI processes.

3.2.3 Confidentiality–Quality–Integration Synthesis

Fundamentals. The first step in the DC–DQ–DI synthesis is to merge (4) and (6) to address the consequences of DI on DC and DQ *jointly*:

$$\left(\mathcal{R}_{\{1,2\}}(\mathcal{D}_1 \bowtie \mathcal{D}_2, S), \mathcal{Q}_{\{1,2\}}(\mathcal{D}_1 \bowtie \mathcal{D}_2) \right) = f \left(\mathcal{R}_1(\mathcal{D}_1), \mathcal{R}_2(\mathcal{D}_2), S, \mathcal{Q}_1(\mathcal{D}_1), \mathcal{Q}_2(\mathcal{D}_2), \mathcal{l} \right). \quad (7)$$

As with (4) and (6), sensitivity of (7) to the measures of risk, DQ metrics, released information and the DI method will be explored in detail. Much of this exploration will be empirical, relying on the prototype software systems described in §3.3 as testbeds. The sensitivities are both quantitative, as numerical parameters of DI methods are varied, and qualitative, as methods themselves are varied.

Metadata are an essential additional ingredient in (7), and especially in the sensitivity analyses outlined above. While the term carries multiple meanings [27, 66], we take it to mean the information necessary to create relationships such as (7). Examples include database schema, models of data generation processes (§3.1.1), and descriptions of DQ metrics, measures of risk and characteristics of DI methods. Building on a variety of resources, including other DG projects [57], we will address two principal research issues: development of the ontologies necessary to represent metadata and the software tools to capture them.

Dynamic Utility. Another need is to extend risk–utility formulations to accommodate dynamic utility. Table and regression servers currently being developed by NISS (§6) incorporate dynamic risk, taking previously released information into account when release decisions are made. The same will be done for utility, in a manner that also reflects DQ and DI considerations.

We illustrate in a setting analogous to (2). Let \mathcal{D} be the database of interest, and let $\mathcal{P} = \text{Rel}(\mathcal{D})$ represent previously released information. An additional release $\text{Rel}^*(\mathcal{D})$ would then be chosen to solve

$$\begin{aligned} \max & \mathcal{U}(\text{Rel}^*(\mathcal{D}) \oplus \mathcal{P}) \\ \text{s.t.} & \mathcal{R}(\text{Rel}^*(\mathcal{D}) \oplus \mathcal{P}) \leq \beta. \end{aligned} \tag{8}$$

The issues to be confronted are the domains of the risk and utility measures, the nature of the combination operation \oplus in (8) and how the optimization is to be performed.

One point of departure is NISS work on table servers (§6). In this case, there is a well-defined query space, and risk and utility can be defined for subsets of it. Moreover, $\text{Rel}^*(\mathcal{D}) \oplus \mathcal{P}$ is a computable, tractably small subset of the query space (the maximal frontier of $\{\text{Rel}^*(\mathcal{D})\} \cup \mathcal{P}$), and calculation of $\mathcal{U}(\text{Rel}^*(\mathcal{D}) \oplus \mathcal{P})$ and $\mathcal{R}(\text{Rel}^*(\mathcal{D}) \oplus \mathcal{P})$ is feasible. Risk is measured using verifiable bounds on confidentiality-threatening small cell values, while utility is measured by the accuracy with which the entire table can be reconstructed using iterative proportional fitting [8]. Generalization to more complex settings poses challenges ranging from extended abstractions of risk and utility to computational algorithms.

Complex Analyses. To this point, the statistical analysis in relationships such as (7) has been left generic. Systems built or being built by NISS offer common, simple analyses — aggregations, cross-tabulations (implicitly, log-linear models) and linear regressions. More complex analyses must be available, especially when access to microdata is limited severely (or forbidden). We will investigate several classes of analyses, some entailing DI, that fall within the setting of (7).

Generalized linear models [74] represent a response Y as a function of predictors X_1, \dots, X_n as

$$Y = a + h\left(\sum_{i=1}^n b_i X_i\right) + \text{Error},$$

where a and the b_i are unknown constants, h is a known function and Error is a random error. *Generalized additive models* [55] have the form

$$Y = a + \sum_{i=1}^n h_i(X_i) + \text{Error},$$

where a is an unknown constant and the h_i are unknown functions. Both classes of models generalize ordinary linear regressions, and inherit the query space structure (response and predictors), as well as abstractions of risk and utility, from regression servers (§6), so that the paths to approach them are discernible.

Economic Models. Expanding the emphasis to date on demographic data and models, we will also address two classes of economic models using data synthesized from the Health and Retirement Study (HRS) conducted by the ISR at UMI: *probit models* and *two-stage regressions* [6].

Data Mining. While a precise definition is elusive, we take data mining to mean techniques for classification and clustering of large data sets that, typically, do not require detailed distributional assumptions [40]. The goal of data mining is generally to identify rules and patterns that hold across many entities, which seems unlikely to compromise individual privacy. Nevertheless, there exists significant, albeit untested, fear that confidentiality of microdata released by Federal statistical agencies may be vulnerable to data mining.

In conjunction with DI it may be possible to find factors linking releasable individual data with data that are not supposed to be linked to individuals. Worse yet, this link may not always hold, but actions may be taken as if it did. An example of this was the (now illegal) practice of red-lining in home mortgages: if the default rate was high in a given neighborhood, then all applicants in that neighborhood were treated as likely to default, resulting in *de facto* racial discrimination. Data mining shows promise of identifying even more

subtle factors linking public data such as addresses with protected attributes such as ethnicity. A second issue arises when the relationships discovered by data mining are themselves the confidential information, even though any individual item in the database is releasable. These problems have received some study [4, 16, 17, 58], which we will continue within the framework of the DC, DQ, and DI models of the project.

Virtual Data. As an alternative to releasing a model of a database, as in (2), virtual data synthesized from the model could be released instead. Such an approach has intriguing advantages — the released records do not correspond to real entities, so that there is no risk of identity disclosure, nor is there any limitation on the analyses that can be performed by users. However, there are also clear shortcomings: users cannot know whether the results of analyses performed on the virtual data are the same as for the original data. We will investigate the risk, quality and integration implications of virtual data.

New Forms of Data, such as images, sound, video and physical objects, such as DNA samples, call for new kinds of analyses [14]. Gaps include abstractions for disclosure as well as for risk, means to compute and reduce risk, metrics for quality and fundamentals of DI. Despite discussion of the issues that multiple-media databases raise for privacy and confidentiality [44, 47], and related concerns with blood and genetic samples [72], as yet there exist few concrete ideas on how to address disclosure limitation.

Design. The ultimate impact of the research will be realized through design of data collection systems and databases in ways that accommodate DC, DQ and DI. Realizing it fully is beyond the scope of this project, but we will identify issues to be confronted and paths to be pursued.

3.3 Software Prototypes

Full impact of the research requires construction of prototype systems that implement the methodologies and models described in §3.1 and 3.2. Their performance must be assessed and compared on *real* Federal databases, whose scale can be formidable, but whose structure often allows for efficient algorithms. Such assessments and comparisons yield both scientific understanding and the path to refine methodologies. Prototypes also provide essential insight into performance and scalability, and inform decisions about which methods to implement in practice.

The software prototypes will primarily be *dynamic query systems*, similar to those described in §6 and [60, 61, 62, 65], in which risk, utility and quality change over time as more queries are answered, including those that require integration of multiple databases.

The high-level architecture of such a system is shown in Figure 2. User access is via a Web browser, while screen display and XML [102] are the principal forms of output. Although all details cannot be foreseen, the prototypes are likely to be deployed as multi-tier distributed systems, with components such as the database layer, Web server–browser applications for the user interface, and computational modules written in C/C++. We currently favor the standardized platform of Java 2, Enterprise Edition (J2EE) [89] as the framework for middleware that provides services to the components.

The fundamental abstractions underlying such a system are its *query space* and time-increasing subsets representing information released to date and the information that has become unreleasable as a consequence. To illustrate, for table servers a query specifies which attributes are to be included, and because the query space is partially ordered by set inclusion, the released information is summarized by its frontier of maximal elements and the unreleasable information by its frontier of minimal elements. Figure 3 illustrates these for a Java prototype.

Scalable Data Structures and Algorithms. Implementing prototype systems on a realistic scale is beyond the capabilities of commercial off-the-shelf software or traditional algorithms. For instance, most standard techniques for performing computations on multi-dimensional contingency tables are “in-memory”

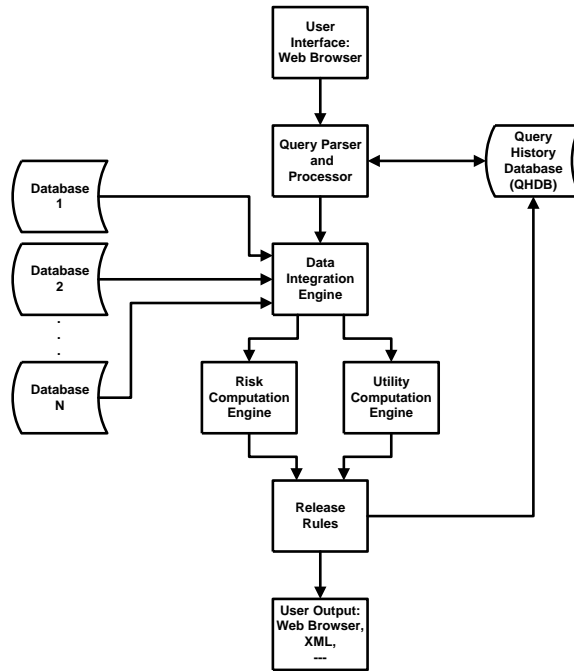


Figure 2: *High-level architecture of a dynamic query system.* User interaction takes place through a Web browser interface. Computational engines perform DI, compute the risk and utility of requested releases, and apply release rules to determine whether queries are to be answered. The query history database maintains the history and system state, and allows risk and utility to depend dynamically on previous queries.

algorithms, which are viable only for tables having around ten (or fewer) dimensions, while real tables with as few as fifteen dimensions can have hundreds of millions of cells.

Whereas the current NISS DG project emphasizes statistical solutions that compute the quantities of interest *correctly*, we will now focus on algorithms whose computational complexity [75], which will be characterized either theoretically or empirically, allows them to scale to the hundreds of attributes and thousands-to-millions of cases in Federal statistical databases. In particular, we will develop algorithms that scale effectively with respect to both the number of cases and the number of attributes, by using cleverly crafted data structures [76, 77] and exploiting problem-specific characteristics such as sparsity or special structure. (Many demographic tables contain structural zeros representing attribute combinations, such as 2-year old parents, that cannot, as opposed to “do not,” occur.) An alternative approach to computation, also to be explored, is to recognize the parallel nature of algorithms such as those for calculating bounds on table entries and to exploit the availability of machines that can carry out such computations.

Release Rules determine, on the basis of risk and utility computations, which queries will be answered, and whether risk reduction strategies [18, 19, 20, 21, 99] will be invoked. They implement the optimizations and risk-utility tradeoffs described in §3.1 and §3.2, and must also account for equity and other agency-user community issues. How particular release rules behave for real databases will be explored empirically and in detail, because there is no other way to assess which ones will work when.

Database Issues. Both the history and the system state are maintained in a single query history database (QHDB). The schema for the QHDB includes tables for transactions (queries), users, the frontier of released

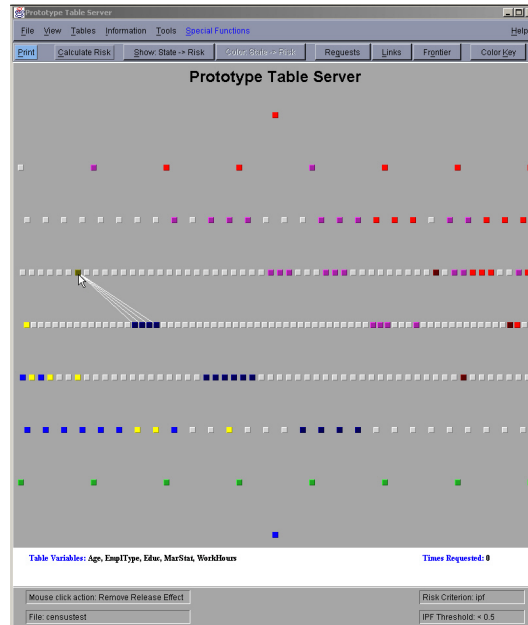


Figure 3: *Visualization of the query space for a table server, showing released information and its frontier, queries that are already risky to answer, and additional queries, some of dimension as low as three, that would become unreleasable if the 5-dimensional query highlighted by the arrow cursor were answered.*

information and the frontier of unreleasable information. Depending on the release rules, the QHDB may also need to maintain additional information about as-yet-unresponded-to queries. The principal issues to be faced are incorporation of problem-specific details and representation of utility, DQ and metadata.

Data mining (§3.2) merges scalability and database issues. Building on research on moving data mining capabilities into database query processing [96], we will explore the use of database capabilities to help scale the statistical algorithms developed during the project.

System Instrumentation and Performance. On the basis of prototypes it may not be possible to determine which query systems will operate effectively in real-world settings, where they are burdened by large numbers of users, concurrency problems and performance degradation resulting from security considerations. Nevertheless, the prototypes will incorporate the instrumentation necessary to monitor system and algorithm performance without affecting either adversely. The design issues of where to place the instrumentation probes, and the analysis issues of how to interpret the data they collect, as well as predict system performance under more stressful circumstance, will be addressed.

User and Operator Interfaces. Java Servlets [90] will form the basis of the system interfaces for both users and operators; they have powerful capabilities for interactive graphics and database access, while minimizing overhead for interprocess communication. The main issue is scalable, comprehensible presentation of the query space and query results. Figure 3 conveys visually the query space and the system state for a modest-sized table server in a form that is very comprehensible but not clearly scalable.

Visualization fulfills a spectrum of roles, from system interfaces (Figure 3 visualizes the query space of a table server.) to risk reduction. In the latter, visualizations allow discovery and exploration of high-level structure while preventing access to confidentiality-threatening details. In addition, visualizations are

robust against certain DQ problems, and can also assist in detection of anomalies (§3.1.1). Not only will we investigate such roles empirically, building on *visual scalability* concepts in [35], but also we will provide techniques for formal treatment of visualizations as information releases, in the setting, for example, of (1).

4 Dissemination and Training

4.1 Dissemination

Project Web Site. A project Web site will be established, which will contain, as does the Web site for the current NISS DG project [78], technical reports and publications produced by participants, presentations made at scientific conferences and project briefings, monthly progress reports and links to related material. Prototype software systems will be available on the Web site, providing opportunity for partner agencies and multiple user communities to explore and evaluate them, as well as propose refinements and alternatives.

Briefings. At least twice annually we will brief the partner agencies on progress and priorities for the research, and the problems encountered. When appropriate, these briefings will be structured as, or coupled with, public workshops that present the products of the project to wider audiences.

Course Development. Courses will be developed in the Department of Statistics at CMU on statistical disclosure control and data quality, which currently lack a prominent place in graduate statistics programs, and associated issues in computer science and engineering. These courses will also be made available to students in Electrical and Computer Engineering, the School of Computer Science and the Information Networking Institute, as well as at UMi and Purdue.

4.2 Training and Human Resources

There will be a strong training emphasis, at multiple levels, that immerses young participants in the science of the project *and* exposes them to the personnel and environments of the partners. From this, they will gain appreciation of the challenges and rewards of research in the Federal statistical agencies.

Postdoctorals, two of whom will be based at NISS and one at CMU, will receive the most intensive training. We expect that they will make definitive contributions to the formulation and conduct of the project. Through experience in the cross-disciplinary collaborations needed to perform the research, their careers will be strengthened significantly. Postdoctorals will be mentored by senior project personnel.

Graduate Students will participate at CMU (one each in computer science and statistics), UMi (one associated with the ISR) and Purdue (one each in computer science and statistics). Emphasis will be on dissertation research that advances the project.

Graduate Summer Interns will be hired by NISS. Typically, these will be students in the early stages of their programs, and often from NISS-affiliated universities [79] not otherwise involved in the project. They will be mentored by senior personnel and postdoctorals.

Undergraduate Students, to be employed at CMU and NISS, will receive exposure to challenging cross-disciplinary research, which can influence their career decisions.

5 Project Organization and Management

Personnel. The project team consists of experienced researchers with knowledge of the problems and strategies to address them, and a history of collaboration with one another and the partner agencies. **Alan F. Karr**,

Director of NISS and PI on the current NISS DG project, as well as director of other NISS projects on DQ, will be project director. In addition to ensuring efficient resource allocation and strong communication, Karr will lead research on DQ foundations and visualization and oversee development of software prototypes. **Stephen E. Fienberg**, Maurice Falk University Professor, Statistics and Social Science, and Acting Director, Center for Automated Learning and Discovery (CALD), CMU, will be co-principal investigator, with overall responsibility for activities at CMU. He and Karr will communicate at least weekly.

Senior investigators will be **Mary Ellen Bock**, Professor and Head, Statistics, Purdue; **Christopher W. Clifton**, Associate Professor, Computer Science, Purdue; **George T. Duncan**, Professor of Public Policy and Statistics, CMU; **Pradeep Khosla**, Philip and Marsha Dowd Professor of Engineering and Head, Electrical and Computer Engineering, CMU; **Sallie Keller-McNulty**, Group Leader, Statistical Sciences, LANL; **Thomas P. Minka**, Visiting Assistant Professor, Computer Science, CMU; **Andrew W. Moore**, A. Nico Haberman Associate Professor, Computer Science, CMU; **Adam A. Porter**, Associate Professor, Computer Science, UMD; **Trivellore E. Raghunathan**, Associate Professor, Biostatistics, and Senior Researcher, ISR, UMi; **Stephen F. Roehrig**, Associate Professor, Information Systems and Public Policy, CMU; **Ashish Sanil**, Research Statistician, NISS; and **S. Lynne Stokes**, Professor, Statistical Sciences, SMU.

Organizations. NISS will be responsible for coordination of activities among project sites at CMU, UMi and Purdue, which will be subawardees, and LANL. **Stokes** will participate primarily by means of collaboration with researchers at LANL, whose own participation will be funded by means other than this award, including directly by NICES. **Porter**, an expert on software instrumentation, will work directly with NISS and CMU.

Schedule and Workplan. We propose a four-year project, whose schedule and principal activities are shown in Table 1, which also shows key senior personnel associated with each activity. Postdoctorals and students will be involved deeply in the entire spectrum of activities.

Absent from the table are feedbacks of later components to earlier ones. Nor can the entire course of the research be foreseen with as much clarity as the table suggests. Each year, in conjunction with senior investigators and the Project Advisory Council (PAC), Karr and Fienberg will prepare an *Annual Workplan* setting forth in detail the priorities of the project and the roles of all personnel. An initial Workplan, for Year 1, will be prepared between the times when an award is made and the project commences.

A major review of the project will take place at the end of Year 2.

Communication. Karr and Fienberg will be responsible for maintaining strong, effective communication within the project and between the project and partner agencies, using an internal area on the project Web site (§4.1) at which participants will exchange ideas and problems, monthly teleconferences that include postdoctorals and students as well as senior personnel, and visits to other sites. Twice per year the project participants will meet at NISS or CMU, at which time Karr, Fienberg and the senior investigators will set the scientific agenda and priorities for the next six months. Briefings to agency personnel also facilitate communication within the project.

Advisory Council. The PAC will provide rapid feedback and assessment, as well as ensure that partner agency and user interests are represented throughout the project. It will be chaired by **Melvyn Ciment**, a moving force behind the creation of NSF's DG program. Other members will be senior researchers from each of the partner agencies, and academic statistical, computer or social scientists to be designated. The PAC will meet at least twice annually, sometimes by teleconference.

Leveraging. The project will be leveraged by ongoing NISS research on DQ, funded by BTS and EPA. A number of non-Federal NISS affiliates [79] also have strong interest in one or more of DC, DQ and DI. Should the pending proposal of Duke University, North Carolina State University, the University of

Time	Activity or Milestone	Personnel
[1, 12]	DQ Foundations: §3.1.1 Abstractions and Quantification Anomaly Detection and Characterization Models for Data Quality	Bock, <i>Karr</i> , Sanil <i>Karr</i> , Raghunathan, Sanil <i>Karr</i> , Raghunathan, Sanil
[6, 15]	Inference as a Measure of DQ: §3.1.2	Bock, <i>Fienberg</i> , <i>Karr</i> , Sanil, Raghunathan
[1, 24]	Risk–Utility Formulations: §3.1.2	<i>Duncan</i> , Keller–McNulty, Stokes
12	Algorithms and Prototypes for Risk–Utility Formulations	<i>Karr</i> , Minka, Moore, Porter, Roehrig, Sanil
[18, 30]	Risk–Quality Formulations: §3.1	Duncan, <i>Fienberg</i> , <i>Karr</i> , Keller–McNulty
24	Algorithms and Prototypes for Risk–Quality Formulations	<i>Karr</i> , Minka, Moore, Porter, Roehrig, Sanil
[15, 36]	Confidentiality Consequences of DI: §3.2.1 (initial focus on record linkage as the means of DI)	Bock, Clifton, <i>Fienberg</i>
[18, 36]	Quality Consequences of DI: §3.2.2	Bock, Clifton, <i>Fienberg</i> , <i>Karr</i>
24	Mid-Project Review with Partner Agencies	All Participants
[24,48]	More Complex Analyses: §3.2 GLM and GAM Economic Models Data Mining Virtual Data Multimedia Data	<i>Karr</i> , Keller–McNulty, <i>Sanil</i> Duncan, <i>Raghunathan</i> Bock, Clifton, <i>Karr</i> , <i>Sanil</i> <i>Duncan</i> , Keller–McNulty, Stokes <i>Fienberg</i> , Moore
30	Algorithms and Prototypes to Assess DC and DQ Consequences of DI	<i>Karr</i> , Moore, Porter, Roehrig, Sanil
[30, 48]	Confidentiality–Quality–Integration Synthesis: §3.2	Bock, Duncan, <i>Fienberg</i> , <i>Karr</i> , Keller–McNulty
42	Algorithms and Prototypes for DC–DQ–DI Synthesis (§3.2) (including More Complex Analyses)	<i>Karr</i> , Minka, Moore, Porter, Roehrig, Sanil
48	Project Completion	
Continual	Evaluation and Refinement of Methods	All Participants

Table 1: Project Timeline and High-Level Workplan. Shown are [start, end] times in months of major activities, **milestones** and key personnel. The leader of each activity is shown in italics.

North Carolina and NISS to establish a Statistical and Applied Mathematical Sciences Institute (SAMSI) be funded, planned SAMSI programs in the social sciences would interact strongly with this project.

The project will have full access to the CALD [11] at CMU, as well as the multidisciplinary Center for Secure Information Systems, directed by **Khosla**. The latter, currently under development, will perform research with technology and policy components addressing such issues as distributed and secure information systems and privacy and confidentiality of distributed information, which will be integrated into software prototypes as appropriate.

The European Community (EC), through *Eurostat*, has supported statistical research programs on confidentiality, with which several members of the current project team have had close contact. We plan to develop formal relationships, involving workshops and direct research collaborations, with one or more statistics/computer science groups funded under the EC’s *Fifth Framework Research Programme*. This initiative includes among its foci “Statistical data mining, statistical modelling, representation and analysis of non-numerical data, the use of administrative data (particularly business registers) for statistical purposes, statistical disclosure control and improvements in quality and in timely and low-cost data production.”

6 Results from Prior NSF Support

Digital Government.⁷ Proof of concept has been demonstrated for Web-based systems that disseminate statistical analyses rather than microdata. A NISS-developed system [60, 61, 62] uses geographical aggregation to allow NASS to disseminate chemical use survey data in far greater detail than in the past. Presaging the research proposed in §3.1.2, detailed study was conducted of the consequences for inference [69].

Table servers have been developed that disseminate marginal subtables from large contingency using dynamic measures of disclosure risk — for example, the accuracy with which small, and therefore risky, entries in the table can be bounded using released information — and utility — for example, the fidelity to the full table of reconstructions performed using iterative proportional fitting [8] on the released information.

Scalable methods have been developed, in special circumstances corresponding to decomposable and reducible graphical models, to calculate bounds for entries in contingency tables in terms of released marginal subtables [25], and are being incorporated into table servers. A general algorithm based on branch and bound methods that computes bounds for entries in contingency tables in terms of released marginal subtables has been developed and applied (initially, to relatively small problems) [26]. Current work is focused on scaling the general algorithm to deal with large tables.

Related work has addressed Monte Carlo computation of the exact distribution of contingency tables under log-linear models conditional on marginal subtables corresponding to the model's minimal sufficient statistics. This work examines and implements the theory of Gröbner bases [23] for multi-way contingency tables [49, 84], but it is computationally intensive. Implementable results on Gröbner bases have been derived for the special cases of decomposable and reducible graphical models [24].

Risk measures have traditionally taken a variety of forms, and significant effort has gone into extending those from sample to population uniques [48]. Initial risk–utility formulations have been developed, treating such widely employed disclosure limitation strategies as data swapping [32, 64, 94, 95].

Regression servers that disseminate the results of regressing one attribute in a database on a subset of the others are being built at NISS, which contain entirely new abstractions for disclosure risk and utility.

Finally, several members of the project team have written expository papers on topics related to statistical disclosure limitation methodology and philosophy [2, 29, 30, 45, 46, 47].

Software Engineering.⁸ This collaborative effort with Lucent Technologies quantified, measured and predicted *code decay* — the increasing difficulty to change large software systems over time without negative consequences. Contributions relevant to the proposal include tools for visualization of large-scale data [34, 35] and *Live Documents* that define new modes of interaction between readers of scientific documents and the underlying data [33], and can form part of the infrastructure of Web-based dissemination systems.

Large Data Sets.⁹ This collaboration with AT&T Research, involving computer scientists, network engineers and statisticians, developed methods for detection of network-based intrusion into computer systems [86] and analysis of customer behavior for Internet service providers [81]. Relevant results include techniques for inference from relational databases, scalable algorithms for dealing with data whose massive scale precludes analysis even if all data could be retained, and techniques for detection of data anomalies.

⁷EIA-9876619, *A Web-Based Query System for Disclosure-Limited Statistical Analysis of Confidential Data*, Alan F. Karr, Principal Investigator. BLS, Census, NASS, NCES and National Center for Health Statistics (NCHS) are agency partners.

⁸SBR-9529926, *Code Decay in Legacy Software Systems*, Alan F. Karr, Principal Investigator.

⁹DMS 97-11365, *Pilot Projects to Explore Large Data Sets*, Jerome Sacks and Alan F. Karr, Principal Investigators

References

- [1] M. Anderson and S. E. Fienberg. Counting and estimation: Methodology for improving the quality of Censuses. The US 2000 Census adjustment decision. In *Proc. International Conference on Quality in Official Statistics*. Statistics Sweden, 2001.
- [2] M. Anderson and S. E. Fienberg. US Census confidentiality: Perception and reality. *Bull. Internat. Statist. Inst.*, 53rd Session, 2001. To appear.
- [3] M. Anderson and S. E. Fienberg. *Who Counts? The Politics of Census-Taking in Contemporary America*. Russell Sage Foundation, New York, 2001.
- [4] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios. Disclosure limitation of sensitive rules. In *Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, pages 25–32, 1999.
- [5] D. P. Ballou and G. Tayi. Methodology for allocating resources for data quality enhancement. *Comm. ACM*, 32(3):320–329, 1989.
- [6] B. H. Baltagi. *Econometrics*. Springer–Verlag, New York, 1998.
- [7] T. R. Belin and D. B. Rubin. A method for calibrating false match rates in record linkage. *J. Amer. Statist. Assoc.*, 90:694–707, 1995.
- [8] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA, 1975.
- [9] B. E. Bryant. Census-taking for a litigious, data-driven society. *Chance*, 6(3):44–49, 1993.
- [10] Bureau of Transportation Statistics. Intermodal Transportation Database. Available on-line at www.itdb.bts.gov.
- [11] Carnegie Mellon University. Center for Automated Learning and Discovery. Information available on-line at www.cs.cmu.edu/afs/cs.cmu.edu/project/cald/www/.
- [12] US Census Bureau Executive Steering Committee for Accuracy and Coverage Evaluation Policy. Report of tabulations of population to states and localities. Federal Register, March 8, 2001. Available on-line at www.census.gov/dmd/www/EscapRep.html.
- [13] K. Chaloner and R. Brant. A Bayesian approach to outlier detection and residual analysis. *Biometrika*, 75:651–659, 1988.
- [14] W. Chang, D. Murthy, A. Zhang, and T. Syeda Mahmood. Metadatabase and search agent for multimedia database access over Internet. In *Proc. Fourth IEEE International Conference on Multimedia Computing and Systems (ICMCS'97)*, 1997.
- [15] G. Chen and S. Keller–McNulty. Estimation of identification disclosure risk in microdata. *J. Official Statist.*, 14(1):79–96, 1998.
- [16] C. Clifton. Using sample size to limit exposure to data mining. *J. Computer Security*, 8(4):281–307, 2000.

- [17] C. Clifton and D. Marks. Security and privacy implications of data mining. In *Proc. ACM SIGMOD Workshop on Data Mining and Knowledge Discovery*, 1996.
- [18] L. H. Cox. Suppression methodology and statistical disclosure control. *J. Amer. Statist. Assoc.*, 75:337–385, 1980.
- [19] L. H. Cox. Linear sensitivity measures in statistical disclosure control. *J. Statist. Planning Inf.*, 5:152–164, 1981.
- [20] T. Dalenius. Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, 5:429–444, 1977.
- [21] T. Dalenius and S. Reiss. Data-swapping: A technique for disclosure control. *J. Statist. Planning Inf.*, 6:73–85, 1982.
- [22] M. H. DeGroot. Record linkage and matching systems. In *Encyclopedia of Statistical Sciences*, volume 7, pages 649–654, New York, 1986. Wiley.
- [23] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.*, 26:363–397, 1998.
- [24] A. Dobra. Gröbner bases for decomposable and reducible graphical models. Technical report, Department of Statistics, Carnegie Mellon University, 2001.
- [25] A. Dobra and S. E. Fienberg. Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. *Proc. Nat. Acad. Sci.*, 97(22):11885–11892, 2000.
- [26] A. Dobra and S. E. Fienberg. Bounds for cell entries in contingency tables induced by fixed marginals. In *2nd Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, 2001. Eurostat.
- [27] Dublin Core Metadata Initiative. Information available on-line at www.purl.org/dc/.
- [28] R. Duda and P. Hart. *Pattern Recognition and Scene Analysis*. Wiley, New York, 1973.
- [29] G. T. Duncan. Confidentiality and statistical disclosure limitation. In N. J. Smelser and P. B. Baltes, editors, *International Encyclopedia of the Social and Behavioral Sciences*. Elsevier Science, Amsterdam, 2001. To appear.
- [30] G. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman, and S. F. Roehrig. Disclosure limitation methods and information loss for tabular data. In J. Lane, editor, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier, New York, 2001. To appear.
- [31] G. T. Duncan, T. B. Jabine, and V. A. de Wolf, editors. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. National Academy Press, Washington, 1993. Report of a Panel on Confidentiality and Data Access, Committee on National Statistics.
- [32] G. T. Duncan and S. Keller–McNulty. Mask or impute? *Res. Official Statist.*, 2001. To appear.

- [33] S. G. Eick, T. L. Graves, A. F. Karr, and A. Mockus. A Web laboratory for software data analysis. *World Wide Web*, 1:55–60, 1997.
- [34] S. G. Eick, T. L. Graves, A. F. Karr, A. Mockus, and P. Schuster. Visualizing software changes. *IEEE Trans. Software Engrg.*, 2001. To appear.
- [35] S. G. Eick and A. F. Karr. Visual scalability. *J. Computational and Graphical Statist.*, 2001. To appear.
- [36] Energy Information Administration. Residential Energy Consumption Survey. Information available on-line at www.eia.doe.gov/emeu/recs/contents.html.
- [37] L. P. English. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. Wiley, New York, 1999.
- [38] US Environmental Protection Agency. Toxic Release Inventory Database. Available on-line at www.epa.gov/triexplorer.
- [39] E. P. Ericksen and T. K. DeFonso. Beyond the net undercount: How to measure Census error. *Chance*, 6(4):38–43, 1993.
- [40] U. M. Fayyad, G. Piatetsky–Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky–Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 1–34. AAAI Press, Menlo Park, CA, 1996.
- [41] Federal Committee on Statistical Methodology. *Report on Statistical Disclosure Limitation Methodology*. US Office of Management and Budget, Washington, 1994.
- [42] Federal Committee on Statistical Methodology. *Record Linkage Techniques – 1997: Proceedings of an International Workshop and Exposition*. US Office of Management and Budget, Washington, 1997.
- [43] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *J. Amer. Statist. Assoc.*, 64:1183–1210, 1969.
- [44] S. E. Fienberg. Towards multiple-media survey and census data: Rethinking fundamental issues of design and analysis. In *Symposium 97: New Directions in Surveys and Censuses, Proceedings*, pages 7–18, Ottawa, Canada, 1998. Statistics Canada.
- [45] S. E. Fienberg. Statistics and privacy in the new millennium: Continuing the dialogue for increased access to research data. In *ASA Proceedings of the Section on Survey Research Methods*, pages 86–88, Alexandria, VA, 1999. American Statistical Association.
- [46] S. E. Fienberg. Confidentiality and data protection through disclosure limitation: Evolving principles and technical advances. In *IAOS Conference on Statistics, Development and Human Rights*, 2000.
- [47] S. E. Fienberg. Statistical perspectives on confidentiality and data access in public health. *Statist. in Medicine*, 20:1347–1356, 2001.
- [48] S. E. Fienberg and U. E. Makov. Uniqueness and disclosure risk: Urn models and simulation. *Res. Official Statist.*, 2001. To appear.

- [49] S. E. Fienberg, U. E. Makov, M. M. Meyer, and R. J. Steele. Computing the exact distribution for a multi-way contingency table conditional on its marginal totals. In A. K. E. Saleh, editor, *Data Analysis from Statistical Foundations: Papers in Honor of D. A. S. Fraser*. Nova Science Publishing, 2001. To appear.
- [50] S. E. Fienberg, U. E. Makov, and A. P. Sanil. A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *J. Official Statist.*, 13:75–89, 1997.
- [51] M. Fortini, B. Liseo, A. Nuccitelli, and M. Scanu. On Bayesian record linkage. *Res. Official Statist.*, 4, 2001. To appear.
- [52] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.
- [53] R. Gnanadesikan and J. Kettenring. Robust estimates, residuals, and outlier detection. *Biometrics*, 28:81–124, 1980.
- [54] P. K. Goel and T. Ramalingam. *The Matching Methodology: Some Statistical Properties*. Springer-Verlag, New York, 1989.
- [55] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall, London, 1990.
- [56] D. Hawkins. Privacy is under siege at work, at home, and online. *US News and World Report*, October 10, 2000.
- [57] E. Hovy, A. Philpot, J. L. Ambite, Y. Arens, J. Klavans, W. Bourne, and D. Saros. Data acquisition and integration in the DGRC’s energy data collection project. In *dg.o 2001: Proc. First National Conference on Digital Government Research*, pages 60–67, Marina del Rey, CA, 2001. Digital Government Research Center.
- [58] T. Johnsten and V. Raghavan. Impact of decision-region based classification algorithms on database security. In *Proc. Thirteenth Annual IFIP WG 11.3 Working Conference on Database Security*, 1999.
- [59] D. Kaplan and R. Krishnan. Assessing data quality in accounting information systems. *Comm. ACM*, 41(2):72–78, 1998.
- [60] A. F. Karr, J. Lee, A. P. Sanil, J. Hernandez, S. Karimi, and K. Litwin. Disseminating information but protecting confidentiality. *IEEE Computer*, 34(2):36–37, 2001.
- [61] A. F. Karr, J. Lee, A. P. Sanil, J. Hernandez, S. Karimi, and K. Litwin. Web-based systems that disseminate information but protect confidentiality. In A. K. Elmagarmid and W. M. McIver, editors, *Digital Government*. Kluwer, Amsterdam, 2001. To appear.
- [62] A. F. Karr and A. P. Sanil. Web-based systems that disseminate information but protect confidentiality. In *dg.o 2001: Proc. First National Conference on Digital Government Research*, pages 159–166, Marina del Rey, CA, 2001. Digital Government Research Center.
- [63] A. F. Karr, A. P. Sanil, J. Sacks, and E. Elmagarmid. Workshop report: Affiliates workshop on data quality. Technical Report, National Institute of Statistical Sciences. Available on-line at www.niss.org/affiliates/dqworkshop/report/dq-report.pdf, 2001.

- [64] S. Keller–McNulty and G. T. Duncan. A progress report to the National Center for Education Statistics: Disclosure-limited statistical analysis of confidential data to support NSF-sponsored Digital Government grant. Technical Report LA-UR-01-1673, Los Alamos National Laboratory, 2001.
- [65] S. Keller–McNulty and E. A. Unger. A remote access database system prototype for the release of confidential data. *J. Official Statist.*, 14(4):347–360, 1998.
- [66] C. Lagoze, C. A. Lynch, and R. Daniel, Jr. The Warwick framework: A container architecture for aggregating sets of metadata, 1996. Available on-line at www.ifla.org/documents/libraries/cataloging/metadata/tr961593.pdf.
- [67] M. D. Larsen and D. B. Rubin. Iterative automated record linkage using mixture distributions. *J. Amer. Statist. Assoc.*, 96:32–41, 2001.
- [68] O. Lassila and R. Swick. Resource Description Framework (RDF), Model and Syntax Specification, W3C Recommendation, 1999. Available on-line at www.w3.org.
- [69] J. Lee, C. Holloman, A. F. Karr, and A. P. Sanil. Analysis of aggregated data in survey sampling with application to fertilizer/pesticide usage surveys. *Res. Official Statist.*, 2001. To appear.
- [70] D. Loshin. *Enterprise Knowledge Management*. Morgan Kaufmann, San Francisco, 2001.
- [71] C. Mackie and N. Bradburn, editors. *Improving Access to and Confidentiality of Research Data: Report of a Workshop*. National Academy Press, Washington, 2000.
- [72] B. Malin and L. Sweeney. Determining the identifiability of DNA database entries. In *Proc. AMIA Symp. 2000*, pages 537–541, 2000.
- [73] E. Martin, T. J. Demaio, and P. C. Campanelli. Context effects for Census measures of race and Hispanic origin. *Public Opinion Quarterly*, 54:551–566, 1990.
- [74] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman & Hall, London, 2nd edition, 1989.
- [75] T. M. Mitchell. *Machine Learning*. McGraw–Hill, New York, 1997.
- [76] A. Moore. Very fast EM–based mixture model clustering using multiresolution *kd*-trees. In *Neural Information Processing Systems*, 1998.
- [77] A. Moore and M. Lee. Cached sufficient statistics for efficient machine learning with large datasets. *J. Artificial Intell. Res.*, 8:67–91, 1998.
- [78] National Institute of Statistical Sciences. Digital Government Project Web Site. Available on-line at www.niss.org/dg.
- [79] National Institute of Statistical Sciences. NISS Affiliates Program Web Site. Available on-line at www.niss.org/affiliates/affiliates-main.html.
- [80] National Research Council, Computer Science and Telecommunications Board. *For the Record: Protecting Electronic Health Information*. National Academy Press, Washington, 1997.

- [81] N. Raghavan, R. Bell, A. F. Karr, M. Schonlau, and D. Pregibon. Defection detection: Using online activity profiles to predict ISP customer vulnerability. In *Proc. Sixth Internat. Conf. on Knowledge Discovery and Data Mining*, pages 506–515, 2000.
- [82] T. C. Redman. The impact of poor data quality on the typical enterprise. *Comm. ACM*, 41(2):79–82, 1998.
- [83] B. Ripley. *Pattern Recognition and Neural Networks*. Oxford University Press, Oxford, UK, 1996.
- [84] S. Roehrig. Auditing disclosure in multi-way tables with cell suppression: Simplex and shuttle solutions. In *Proc. Joint Statistical Meetings*, Alexandria, VA, 1999. American Statistical Association.
- [85] P. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.
- [86] M. Schonlau, W. DuMouchel, W.–H. Ju, A. F. Karr, M. Theus, and Y. Vardi. Computer intrusion: Detecting masqueraders. *Statist. Sci.*, 16, 2001. To appear.
- [87] A. C. Singh, M. Feder, G. Duntelman, and F. Yu. Use of optimal subsampling, substitution, and calibration for protecting quality and confidentiality of public use microdata. In *Proc. ENAR/IBS Spring Meeting*, 2001.
- [88] A. C. Singh, H. J. Mantel, M. D. Kinack, and G. Rowe. Statistical matching: Use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19:59–79, 1993.
- [89] Sun Microsystems, Inc. Java 2 Platform, Enterprise Edition Specification. Information available on-line at java.sun.com/j2ee/.
- [90] Sun Microsystems, Inc. Java Servlet Technology. Information available on-line at java.sun.com/products/servlet/.
- [91] L. Sweeney. *Computational Disclosure Control*. PhD thesis, MIT, 2001.
- [92] J. M. Tanur, editor. *Questions About Questions*. Russell Sage Foundation, 1992.
- [93] G. Tayi and D. Ballou. Examining data quality: Guest editors’ introduction. *Comm. ACM*, 41(2):54–57, 1998.
- [94] M. Trottni. A decision-theoretic approach to data disclosure problems. In *2nd Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, 2001. Eurostat.
- [95] M. Trottni. A decision-theoretic approach to data disclosure problems. *Res. Official Statist.*, 2001. To appear.
- [96] S. Tsur, J. D. Ullman, S. Abiteboul, C. Clifton, R. Motwani, S. Nestorov, and A. Rosenthal. Generalization of association-rule mining. In *Proc. SIGMOD Conference*, pages 1–12, 1998.
- [97] R. Wang. A product perspective on total data quality management. *Comm. ACM*, 41(2), 1998.
- [98] R. Y. Wang, M. Ziad, and Y. W. Lee. *Data Quality*. Kluwer, Amsterdam, 2000.

- [99] L. C. R. J. Willenborg and T. de Waal. *Statistical Disclosure Control in Practice*. Springer-Verlag, New York, 1996.
- [100] W. E. Winkler. Matching and record linkage. In *Record Linkage Techniques – 1997: Proceedings of an International Workshop and Exposition*, pages 374–403, Washington, 1997. US Office of Management and Budget.
- [101] W. E. Winkler. Machine learning, information retrieval and record linkage. Available on-line at www.niss.org/affiliates/dqworkshop/papers.html, 2000.
- [102] World Wide Web Consortium. Extensible Markup Language (XML). Information available on-line at www.w3.org/TR/REC-xml.