Chapter 15
Data Privacy

Ágoston Reguly and Zsigmond Pálvölgyi

Abstract This chapter examines how data privacy-preserving methods have affected empirical research over the last decade, emphasizing their implications for identification, estimation, inference, and reproducibility. It distinguishes between revealable confidentiality, where accurate microdata exist but cannot be openly shared, and unrevealable confidentiality, where privacy concerns prevent direct observation of the sensitive variables. First, we review the historical evolution of data protection and discuss lessons from data leaks, which primarily governed shifts in the approaches used. Then, we overview the history of econometric modeling with imprecisely observed sensitive variables and highlight key takeaways for data privacy considerations. We argue that modern empirical analysis with sensitive data must intentionally choose both the privacy-preserving method and the parameter of interest specified by the econometric model. These choices directly affect identification and inference. Looking to the future, we emphasize that privacy concerns should receive greater focus and urge the development of new methods that jointly account for the risk-utility trade-off in data protection, thereby satisfying the basic requirements of econometric analysis.

15.1 Introduction

In empirical research and policy analysis, data privacy has moved from a secondary administrative concern to a central methodological and institutional issue. The growing availability of detailed microdata on individuals and firms has substantially

Ágoston Reguly ✉
Corvinus University of Budapest, Budapest, Hungary and Georgia Institute of Technology, Atlanta, Georgia, USA, e-mail: agoston.reguly@uni-corvinus.hu

Zsigmond Pálvölgyi
Corvinus University of Budapest, Budapest, Hungary and CERGE-EI Foundation, New York, USA, e-mail: zsigmond.palvolgyi@uni-corvinus.hu

expanded the scope of empirical inquiry, particularly in official statistics, applied economics, health research, and administrative data analysis. At the same time, this development has brought stronger legal mandates to protect confidentiality, heightened concern for maintaining respondent trust, and increased pressure to ensure that empirical findings remain transparent and reproducible even when the underlying data cannot be openly shared. Statistical agencies and researchers, therefore, nowadays operate in a more demanding environment than in earlier decades. They must produce informative, policy-relevant evidence from increasingly rich data, while also ensuring that participation does not expose respondents to harm and that confidential information is not disclosed, either directly or accidentally. In this sense, data privacy is no longer merely a constraint on access, but it has become an important feature of the modern empirical workflow.

In this chapter, we provide an overview of the history of data privacy-preserving techniques, which are mainly embedded in information theory and the statistical tradition. This implies that key elements of econometric analysis have not played a central role in the development of such methods. This became evident in the data privacy field over the last decade, where consistent, unbiased, and efficient estimators across models are not the main focus of research, but rather an occasional byproduct. On the contrary, the econometric tradition often treated sensitive, hence modified data as given and developed methods that could handle data in such form. However, the confidentiality problem in empirical economic analysis, as we argue, should be treated jointly: preserving privacy, while allowing the use of a wide variety of econometric models. Here, we discuss both traditions and provide historical lessons and potential synergies between the two fields to address the growing pressure to use micro level sensitive variables in empirical analysis.

The main challenge – learned from history – is not simply that some data are confidential, but linkage across datasets enables re-identification. This was a painful lesson that the data privacy profession had to learn the hard way during the 2000s. Data protection is now understood as a fundamental risk–utility trade-off. The more detailed and analytically useful a release is, the greater the likelihood that information about particular individuals or firms can be reconstructed, whereas stronger privacy protection reduces statistical applicability by altering the feasible estimands of interest. These tensions are especially important for econometric practice because confidentiality protection affects not only what data may be shared, but also what can be learned from them, under what assumptions, and with what degree of inferential validity. For this reason, data privacy should be understood not simply as a matter of legal compliance, but as a substantive issue in statistical design with consequences to empirical research.

A useful point of departure for this chapter is that confidentiality is not a single problem, but two conceptually distinct ones. In one setting, the researcher or statistical agency accurately observes the relevant microdata, yet legal, ethical, or institutional obligations prohibit disclosure in identifiable form. In the other, the difficulty arises already at the stage of data collection: individuals or firms may be unwilling to reveal sensitive information truthfully, or at all, so that the variable of interest is observed only indirectly through some (discretization) mechanism. These two situations differ

not merely in degree but in kind because privacy enters the empirical process at different stages, thereby generating distinct econometric challenges.

The first of these can be described as the problem of *revealable* confidentiality. Here, the underlying information is measured with precision, often in administrative records, censuses, data centers, or confidential surveys. However, access to the microdata is restricted because disclosure could reveal the identity or sensitive attributes of particular individuals or firms. The empirical challenge is therefore not how to observe the variable, but how to use the data statistically without violating confidentiality constraints. This is the familiar case of sensitive but observed data: income records, medical histories, financial transactions, and linked demographic databases may all be highly informative for econometric modeling, yet their direct disclosure would create legal and ethical risks. Historically, this problem has given rise to legal safeguards, statistical disclosure control, secure access environments, synthetic data generation, and, more recently, differential privacy. Across these approaches, the central concern remains the same: how to preserve the analytical value of well-measured data while ensuring that no identifiable unit can be reconstructed or too precisely inferred from the released information.

The second problem may be called *unrevealable* confidentiality. In this case, privacy concerns affect the measurement process itself. Respondents may refuse to state exact income, wealth, health status, or other sensitive attributes. Instead, individuals or firms are more willing to provide bracketed responses, incomplete answers, or no response at all. In many surveys, the relevant quantity is therefore not observed in its exact form, but through intervals, Likert-type scales, censored outcomes, or noisy self-reports. Here, sensitivity is not merely a restriction on dissemination, but part of the 'data-generating process'. As a consequence, the econometric problem is deeper: one must draw inference about an underlying quantity that is not directly observed, hence confidential information cannot be revealed. Unlike the first setting, where the 'true' value often exists in a secure environment, here the analyst may only observe a transformed, coarsened, or partial signal of the latent variable of interest.

This distinction has important methodological consequences. Under revealable confidentiality, the analyst typically begins from a correctly measured confidential dataset and asks how disclosure-control mechanisms alter the estimands of interest, statistical inference, or the reproducibility of the results. The associated literature therefore focuses on how to transform, restrict, or mediate access to accurate microdata while, where possible, retaining the statistical properties needed for credible analysis. The key question for the analysis's credibility is whether the mechanism used to protect anonymity preserves not only one or some target quantities (e.g., mean, median, etc.), but also allows for consistent parameter estimation across different econometric models. Some procedures are designed to leave selected aggregates or low-dimensional margins approximately unchanged; others, however, provide stronger privacy guarantees at the cost of additional noise, reduced precision, or limits on the admissible range of analyses. From an econometric perspective, the issue is not merely whether privacy is protected, but whether the protection mechanism leaves the estimand identifiable and the resulting inference valid. The central issue, therefore, is

how to protect confidentiality without undermining econometric usefulness. As we show in Section 15.2, some methods completely ignore this econometric perspective, while others offer insights into the models used; however, limitations and remaining challenges must be emphasized.

Under unrevealable confidentiality, by contrast, identification is already part of the problem. Since the variable of interest is not directly observed, the analyst must model the reporting or elicitation mechanism jointly with the behavioral relationship of interest. This naturally leads to latent-variable models, models with censoring, truncation, interval data, missing data, ordered-response models, and measurement error. When point identification via model assumptions is too strong, partial identification and set-valued inference become natural alternatives. The main methodological question here is therefore not how to protect well-measured data, but how to recover model parameters from variables that are observed only indirectly or imperfectly. In such cases, we must explicitly account for that loss when formulating the model and interpreting the results.[1] As we discuss in Section 15.3, some of the methods used are widely accepted in empirical practice due to their simplicity (although relying on strong assumptions), there are, however, newer more comprehensive approaches, which have been replacing other more 'traditional' ones.

Taken together, revealable and unrevealable confidentiality have generated two broad methodological responses: protecting accurately observed but sensitive data and modeling variables that are not directly observed. Although these traditions are often treated separately, they are linked by a common underlying feature: confidentiality modifies the relationship between the substantive variable of interest and the information available to the analyst. The perspective of this chapter is that confidentiality should not be treated as an external administrative inconvenience but rather as a fundamental feature of modern data environments that directly affects identification, estimation, and inference. Recognizing the distinction between *revealable* and *unrevealable* confidentiality is useful not only for organizing the historical development of privacy-preserving methods and discussing which methods aged well, but also for clarifying why different approaches disappeared. As we argue, some empirical problems are best understood through the lens of disclosure limitation and controlled access, whereas others demand models with specific attributes. In both cases, however, privacy protection inevitably alters the information available for analysis, and empirical methodology must explicitly account for that alteration rather than treating it as a technical detail. These lenses also allow us to view the future of data privacy in a new light. Namely, to learn from both revealable and unrevealable confidentiality, and to promote the development of approaches that protect the anonymity of individuals, but also allow the use of a wide range of econometric models without pre-specification before the data becomes available.

[1] Let us mention here that there are many cases where variables are measured similarly, but it is not related to data confidentiality. Examples include happiness indices, utility measures, and other behavioral-psychological measures that, by their nature, can only be observed using these scales. Many of the aforementioned methods have been developed to model these types of problems, not to address confidentiality issues.

The chapter is structured as follows. First, we overview the history of revealable confidentiality before 1970, and then for each decade after 1970 until recent years. Then, we turn to unrevealable confidentiality, where we overview the main approaches and then how they evolved. Finally, we compare these two traditions, discuss some recent developments, and elaborate on the remaining difficulties and the likely directions for future research.

15.2 Revealable Confidentiality: Data Available but Sensitive

Let us first consider settings in which the data are accurately measured but cannot openly be released. The central question is how to protect confidentiality without destroying the statistical usefulness of the data. The discussion begins with legal and institutional safeguards. Then it follows the historical shift toward statistical disclosure control, synthetic data, and differential privacy, emphasizing throughout the problems these newer methods solve and how they affect econometric inference.

15.2.1 Before 1970s: Institutional Norms and Legal Safeguards

Before the 1970s, there were no explicit 'privacy-preserving techniques,' especially as a unified mathematical field, but rather a collection of institutional norms and legal safeguards implemented in a pragmatic fashion. In the United States, early censuses (1790–1840) operated with essentially no confidentiality protections. Enumerators were legally required to post locally identifiable values (registries) in public places so that communities could correct errors, effectively trading privacy for accuracy and accountability (U.S. Census Bureau, 2003, 2019). As the census expanded from headcounts toward richer demographic and economic information, resistance to 'prying' first surfaced around business information. By 1840, officials explicitly instructed enumerators to assure manufacturers that names would not appear in published tables and to treat communications about 'the business' as confidential, foreshadowing the modern idea that statistical reporting can be protected by separating micro-level collection from macro-level publication (U.S. Census Bureau, 2003; Davis, 1973).

From the mid-to-late 19th century through the early 20th century, confidentiality gradually evolved from informal directives into enforceable commitments. Administrative guidance between 1850 and 1870 emphasized nondisclosure by field staff. Statutory provisions during the 1880 and 1890 Census Acts required oaths. They penalized disclosure, and by 1910, presidential proclamations and stronger penalties protected census responses from harm and misuse, while later legislation culminated in felony-level sanctions for unauthorized disclosure of census information. In parallel, agencies confronting the risks of indirect disclosure in published aggregates developed early operational disclosure controls. Manual eyeballing of tables, suppressing or

compressing categories, and, when necessary, withholding small-area tabulations. This anticipates the later Statistical Disclosure Control (SDC) doctrine even before it was named. The shift was propelled by automation (tabulating machines, then computers) that enabled much more granular cross-tabulation, thereby increasing the risk of reidentification.

By the mid-20th century, confidentiality principles were consolidated into durable legal regimes, such as Title 13 of the U.S. Code (enacted 1954, U.S. Government Publishing Office), which formalized strong restrictions on disclosure and established that statistical agencies must protect respondent information throughout the lifecycle of the data. At the same time, two approaches that would later become central to modern privacy technology were already taking shape. The first strand is information-theoretic secrecy in communication, crystallized by Shannon (1949), who formalized 'perfect secrecy' and the trade-offs between key information, redundancy, and the adversary's posterior beliefs. These concepts inspired privacy definitions based on bounded information gain. The second strand is randomized masking mechanisms for data collection. Most notably, Warner (1965) introduced the randomized response technique, one of the first methods to add noise to responses to protect individuals' anonymity.

Finally, just before the 1970s, the emergence of large-scale administrative files and the prospect of linking records across sources motivated the paper Fellegi and Sunter (1969), which laid the foundation for the probabilistic record linkage framework. This paper sharpened awareness that privacy risks are amplified not only by what is released, but also by what can be inferred through linkage. Despite its importance, this paper was quite ahead of its time, and its significance became only evident during the linkage threat after the 1990s.

15.2.2 The 1970s: Emergence of Statistical Disclosure Control

The 1970s are widely viewed as the decade in which confidentiality concerns in official statistics hardened into recognizable protocols, which later became known as Statistical Disclosure Control (SDC). During this time, statistical offices were simultaneously expanding the scope of what they published (more cross-tabulations, finer geography, emerging 'summary tapes' and microdata infrastructures) and facing new risks from computerized storage and retrieval (see e.g., Hansen, 1972; Fellegi & Phillips, 1974). A rapid emergence of research in the field can be observed over this period, including studies on protecting individual records in statistical data systems (United States Federal Committee on Statistical Methodology, 1978), addressing formal confidentiality issues (Dalenius, 1978), and implementing concrete table-protection mechanisms such as cell suppression and rounding (Cox, 1980).

This shift reflects that the issue had become operationally urgent across various agencies rather than merely a legal obligation. For example, in the U.S., the Census Bureau's primary approach around 1970 relied on whole-table suppression keyed to small-area counts. This strategy reduced usability and still left vulnerabilities when

complementary tables were not suppressed. This prompted an active exploration of alternatives such as random rounding, ordinary rounding, table redesign, and later complementary suppression rules (see more in McKenna, 2018).

Methodologically, these advancements coalesced around a two-stage workflow: (i) *risk identification* (e.g., flagging 'sensitive' cells, small counts, dominance situations, or uniquely identifying combinations) and (ii) *risk mitigation* via controlled information loss. This information loss is achieved through suppression (primary and complementary), coarsening/recoding, rounding/perturbation, and record-level transformations in microdata-like environments. Among these, randomized perturbation in tabular outputs was especially influential. Fellegi (1975) showed that aggregates constructed by random rounding produce the same expectation as the original (unreleased) quantity. This was especially attractive for econometric use, as it suggested that certain estimators of population means, totals, and linear contrasts remain unbiased conditional on the perturbation mechanism. Even though valid inference required inflating variance (or otherwise accounting for added disclosure noise) when forming standard errors and tests (see e.g., Fellegi, 1975; Cox, 1987).

One of the core concepts emerging during the 1970s was the notion of 'inferential disclosure', proposed by Dalenius (1977). With inferential disclosure, access to a release should not increase what can be learned about any particular individual beyond what could be learned without the release. This idea captures the 'no additional learning' criterion, which reframes confidentiality as a statement about inference, not merely about removing identifiers (see more in Dwork & Naor, 2008). This perspective clarified why ad hoc rules must be evaluated against an adversary with auxiliary information and implicitly laid the groundwork for the modern risk–utility framing. In practice, statistical offices translated these ideas into SDC toolkits whose econometric implications depend on which statistics are preserved. For example, record swapping (Dalenius & Reiss, 1982) was motivated by preserving selected low-order quantities (e.g., mean) so that inferences relying only on those sufficient statistics are unaffected. At the same time, other joint relationships (e.g., regression slopes) may be attenuated or distorted unless analysts explicitly model the perturbation. This distinction matters for econometrics, as early SDC methods often kept key first-order aggregates approximately unbiased but did not generally guarantee unbiased parameter estimates for arbitrary structural or reduced-form models. This highlights the enduring tension between econometric modeling and data privacy, which ensured the safety of official releases through newly designed confidentiality protocols. Still, the econometric validity of downstream estimation hinges on whether the chosen SDC preserves the estimands and whether the sampling distributions relevant to the model analysts actually fit. This forced econometricians to develop different (often complicated) methods to account for these issues, as discussed in the next section. The main contribution of the 1970s was to turn confidentiality from an informal obligation into an explicit statistical design problem.

15.2.3 The 1980s: From Suppression to Perturbation-Based SDC

In the 1980s, data privacy in official statistics moved from largely procedural safeguards toward more explicit statistical disclosure limitations, driven by two converging pressures: (i) user demand for increasingly granular tabulations and microdata-like products, and (ii) growing awareness that suppressing 'too much' data undermines the analytic mission of statistical offices. A telling example is the U.S. Census Bureau's shift from the 1970-era reliance on whole-table suppression to the 1980 Census, where table-level suppression remained central but was augmented by more detailed 'complementary' suppression rules. This resulted in more reported information (e.g., race-by-tenure structures), but prevented researchers from directly using it in regression analysis due to suppression.

The increase in the information provided (despite suppression), posed new challenges. One of the most notable examples is the absence of *complementary* suppression across geographic hierarchies, thereby enabling subtraction attacks (e.g., recovering a suppressed county table from state totals minus other counties; see more in McKenna, 2018). The methodological evolution in the 1980s followed this problem: rather than removing outputs entirely, agencies increasingly adopted structured perturbation designed to preserve key statistical properties. Causey, Cox and Ernst (1985) argued for the use of controlled rounding,[2] and showed how it could be made operationally feasible for large tabulations and connected it directly to disclosure limitation. Importantly, the decade also saw a sharper conceptualization of disclosure itself. Duncan and Lambert (1986) framed disclosure limitation via predictive distributions and uncertainty functions, and formalized the idea that releases are disclosive precisely when they affect posterior beliefs about a target too much.

The 1980s developments clarify a recurring lesson for econometrics. Some SDC mechanisms preserve unbiasedness for particular (linear) estimands, such as totals under properly designed random/controlled rounding. Still, they generally do not guarantee unbiasedness or valid classical inference for arbitrary regression parameters or nonlinear functionals unless the disclosure mechanism is explicitly incorporated into estimation (e.g., as a known perturbation/measurement process). The 1980s, therefore, made clear that privacy protection could no longer be evaluated only by whether disclosure was prevented; it also had to be judged by how much useful inference it preserved.

15.2.4 The 1990s: New Linkage Threat – Combining Multiple Sources

The 1990s marked a decisive operational and conceptual turn in data privacy for official statistics. Statistical agencies increasingly abandoned blunt publication controls (e.g., whole-table suppression) in favor of record- and mechanism-based disclosure avoidance. Simultaneously, due to the increasing number of publications,

[2] Interestingly, they used methods from the transportation literature.

data privacy was threatened by linkage risks. This early problem addressed by Fellegi and Sunter (1969) became a reality by cheap computation and proliferating external lists.

Record- and mechanism-based disclosure avoidance became evident in the U.S. decennial census program. For example, in the 1990 'Confidentiality Edit' the Bureau implemented a rule-based data swapping operation on a set of households, matched on a small set of key characteristics such as race or unit counts by vacancy status. These methods became popular, and the Bureau generated all published tabulations using similar techniques. This allowed statistical offices to avoid pervasive table suppression and made disclosure protection largely invisible (Griffin, Navarro & Flores-Baez, 1989; McKenna, 2018). These changes mattered as the earlier suppression regime ended, sometimes with missing tables.

However, record- and mechanism-based methods remained vulnerable when complementary suppression did not span, e.g., geographic additivity. To maintain usability at finer levels, offices started to inject uncertainty about units. For example, in terms of geographical location, noise means perturbing households across small neighboring areas. Mid-decade methodological reviews within the Census Bureau then focused on improving this procedure by explicitly weighing, e.g., record swapping and adding noise to data. As both methods add certain distortion to the data – foreshadowing the privacy–utility frontier – the weighing of the two methods becomes an empirical design problem rather than a purely administrative one during this time (R. A. Moore, 1996).

During this decade, the linkage threat became apparent. The broader research community demonstrated that removing direct identifiers does not guarantee anonymity. Sweeney (2000) showed that quasi-identifiers can be linked. Using 1990 Census summary data, she showed that combinations like {5-digit ZIP, sex, date of birth} can uniquely (or near-uniquely) identify a very large share of individuals (famously, 87% of the U.S. population), and she illustrated practical re-identification by linking 'de-identified' health-type records to purchased voter rolls. This evidence directly challenged the prevailing 'release-and-forget' intuition of the era. Linkage demonstrations catalyzed a late-1990s shift toward formal criteria of microdata anonymization. A well-known method of that time is the k-anonymity framework pioneered by Samarati and Sweeney (1998). They proposed that each released record should be indistinguishable from at least $k - 1$ others with respect to its quasi-identifiers. For instance, exact birth dates may be converted to years of birth, or detailed ZIP codes to broader regions, until each equivalence class contains at least k individuals. The intuition is that an intruder who knows the quasi-identifiers cannot isolate one person in the released data, but can only narrow the match to a group of at least k candidates.

In parallel, another approach emerged, namely the synthetic data approach as a response to a widening gap between (i) growing demand for public use microdata and reproducible empirical research, and (ii) rising disclosure risks under traditional masking and suppression rules. The pioneering work of Rubin (1993) proposed releasing no actual microdata at all, but instead disseminating synthetic microdata

generated via the multiple-imputation framework.[3] He argued that users could analyze synthetic files using standard methods, while agencies reduced the risk of direct identification because released records are simulated rather than actual values. Little (1993) articulated a complementary and more conservative variant—what later became known as partially synthetic data, in which only the most disclosure-prone fields (e.g., sensitive outcomes or quasi-identifiers) are replaced, leaving much of the dataset unchanged to preserve analytic utility better, but at the cost of potentially higher residual disclosure risk. Throughout the late 1990s, these ideas remained largely conceptual for two reasons. The first is that there were no universally accepted model(s) to generate synthetic data from the agencies. Simple models (e.g., linear Gaussian models) might be too restrictive, whereas overly complex models lack theoretical and practical acceptance from the research community. The second reason was related to the previous point, but it was due to the dataset's size. Typically, sensitive variables are interesting alongside many other variables. In a high-dimensional environment, synthetic records were computationally infeasible during this time.

15.2.5 The 2000s: Failure of De-identification and the Rise of Differential Privacy

In the 2000s, the privacy landscape for public microdata releases shifted from debating whether de-identification works to understanding how repeated releases fail under adversarial attacks with auxiliary information and linkage opportunities. This turn was motivated by two emblematic leakage episodes. The first data leak is AOL's[4] 2006 release of 'anonymized' search logs (usernames replaced by numeric IDs). These data were re-identified from query content and patterns, with journalistic reconstruction of individual users from ostensibly anonymous records (see more in Barbaro & Zeller, 2006). The second case is the Netflix Prize dataset (released as anonymous movie ratings for roughly 500,000 subscribers), which was de-anonymized by Narayanan and Shmatikov (2008) via robust linkage to external ratings (using information from e.g., IMDb[5]), demonstrating that even partial background knowledge about a person's ratings can isolate their record in a sparse, high-dimensional space and thereby reveal sensitive preferences. These incidents reinforced a key practical lesson for the limitation of k-anonymity developed by Samarati and Sweeney (1998) in the previous decade: in high-dimensional microdata, sparsity makes many records essentially

[3] One of the simplest examples of such synthetic microdata is the following. Take random variables of (X_i, Y_i), where Y_i is the sensitive data. Fit a model to the confidential data, $Y_i|X_i \sim f(Y_i|X_i;\theta)$, where $f(\cdot)$ is the conditional density function with parameters θ. For simplicity let it be a Gaussian linear regression $Y_i = X_i\beta + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Next, generate synthetic $\tilde{Y}_i$ values using the fitted model; in this simple case, $\tilde{Y}_i = X_i\hat{\beta}$ + random draw from $\mathcal{N}(0, \hat{\sigma}^2)$ and parameters are estimated via e.g., OLS. In this case, $(X_i, \tilde{Y}_i)$ is the released synthetic microdata.

[4] AOL Media LLC (formerly a predecessor company known as AOL Inc. and originally known as America Online). It is an American web portal and a former online service provider.

[5] Internet Movie Database (IMDb), is an online database of information related to films, television series, podcasts, video games, and streaming content online.

unique, so releasing rich behavioral attributes creates attack surfaces that are not well managed by traditional de-identification checklists (Aggarwal, 2005).

Papers from this era (see e.g., Ganta, Kasiviswanathan & Smith, 2008; Aggarwal, 2008) showed that k-anonymity primarily addresses identity disclosure under a narrow quasi-identifier model, and it can fail for attribute disclosure even when identity is protected, because equivalence classes may exhibit homogeneity (little variation in sensitive attributes) or be vulnerable to background knowledge. This recognition drove a sequence of refinements explicitly designed to patch k-anonymity's weaknesses. Machanavajjhala, Kifer, Gehrke and Venkitasubramaniam (2007) introduced ℓ-diversity (an improved version of k-anonymity), which formalized the requirement that each equivalence class contains sufficiently diverse sensitive values to resist homogeneity and background-knowledge attacks. Nevertheless, ℓ-diversity itself was shown to be neither necessary nor sufficient in some regimes (e.g., skewed global distributions where 'diversity' can be misleading). This recognition motivated the emergence of t-closeness (Li, Li & Venkatasubramanian, 2007), which tied protection to the closeness of the the sensitive-attribute distribution within each equivalence class to the overall distribution. These contributions primarily focused on formal privacy guarantees and protection against re-identification or attribute disclosure. Although they recognized a privacy–utility trade-off, utility was typically assessed using information-loss criteria, feasibility, or illustrative empirical performance rather than the statistical properties of downstream econometric estimators. In particular, this literature did not develop any (asymptotic) theory for estimation and inference.

Re-identification problems in the early 2000s also marked a turning point for the synthetic data approach in practice. Not only does the synthetic data approach provide a more robust answer to privacy concerns than methods before the 2000s, but it also enables reliable parameter estimation for wider range of models, an advantage over k-anonymity and subsequent methods. Via design-based simulations using common survey designs, Reiter (2002) demonstrated that properly generated synthetic datasets can support valid estimation for broad classes of estimands. He also made suggestions in practical design choices, such as the number and size of released synthetic datasets, and highlighted the importance of including design variables during the data generation to preserve inferential targets. Shortly thereafter, Raghunathan, Reiter and Rubin (2003) supplied the missing *combining rules* for fully synthetic data, explaining why standard missing-data with multiple-imputation pooling formulas are not generally valid for synthetic releases. [6] They also provided a corrected variance decomposition that allows analysts to fit familiar models (including standard econometric regressions) on each synthetic file and then obtain

[6] The 'combining rules' specify how analysts should pool estimates and standard errors obtained from multiple synthetic datasets. If $\hat{Q}^{(l)}$ is the parameter of interest from the synthetic dataset l with $l = 1, \ldots, L$, then the pooled estimate is just the simple average $\bar{Q} = L^{-1} \sum_{l=1}^{L} \hat{Q}^{(l)}$. However, for synthetic data, the total variance must be corrected and the proper variance estimator is $T = (1 + L^{-1})\hat{B} - \bar{U}$, where $\hat{B} = 1/(L-1) \sum_{l=1}^{L} (\hat{Q}^{(l)} - \bar{Q})^2$ is the between-synthetic-dataset variance of the estimates and $\bar{U} = 1/L \sum_{l=1}^{L} Var(\hat{Q}^{l})$ is the average variance estimate within the synthetic datasets. This adjustment is needed because, when the whole dataset is simulated, the usual formula would double-count part of the uncertainty.

valid inference by pooling results. Reiter (2005) proposed a solution to nonlinear problems via a classification and regression tree-based method to create partially synthetic microdata, which better captures complex interactions without the need for subjective decision-making from the econometrician.

By the late 2000s, research attention shifted toward operational governance of synthetic releases. The main task was to quantify the risk–utility trade-off by choosing the number of released imputations and by evaluating whether synthetic data preserves key analytic relationships (Reiter, 2010). Drechsler and Reiter (2009) illustrated this problem by providing multiple examples for empirical studies with partially synthetic business microdata. This period also saw refinements in implementation, such as two-stage strategies that reduce computational burden and allow different numbers of imputations for different variables, reflecting the maturation of synthetic data from a theoretical concept into a production workflow (Reiter & Drechsler, 2010; Drechsler, 2011). Across this trajectory, the central lesson has remained the same: synthetic data can preserve familiar analysis workflows, but its inferential fidelity is inherently model-mediated. This means that the utility of this approach depends on whether the synthesis models capture the joint structure needed for downstream econometric estimands, and that confidentiality evaluation depends on risk measures that necessarily embed assumptions about potential attacks and linkage opportunities (Drechsler, 2011). This is easy to accept if the synthetic data is created for specific tasks, but less so for various non-pre-specified analyses.

During the mid 2000s, the 'multiple releases' problem became visible, which was not addressed properly by the synthetic data approach. When different organizations independently publish on overlapping populations, attackers can combine releases to narrow equivalence classes. Differential privacy (DP) emerged from this realization. A key precursor was the reconstruction attack of Dinur and Nissim (2003), which showed that sufficient number of subset-sums, each answered with small error, can allow efficient recovery of much of the underlying database. Motivated by this insight, the seminal papers of Dwork (2006) and Dwork, Kenthapadi, McSherry, Mironov and Naor (2006) proposed differential privacy, formalizing privacy as a stability requirement with respect to other databases that differ by a single individual record. Formally, let a randomized mechanism $\mathcal{M}$ is (ε, δ)-differentially private if for all neighboring databases D, D' and all measurable output events S, $\Pr\big(\mathcal{M}(D) \in S\big) \leq e^{\varepsilon} \Pr\big(\mathcal{M}(D') \in S\big) + \delta$. $\varepsilon \geq 0$ is the privacy loss parameter (often called the 'privacy budget'). The smaller ε forces the output distributions under D and D' to be closer in a multiplicative sense, meaning that including (or excluding) any single individual's data cannot change the likelihood of any outcome by more than a factor of e^{ε}. The second parameter, $\delta \in [0, 1)$, is a small failure probability allowing the privacy guarantee to be violated on a set of outputs of probability at most δ (e.g., leakage). Setting $\delta = 0$ yields pure ε-DP, while $\delta > 0$ gives approximate DP. This definition captures the idea that participation should not substantially increase an individual's risk, regardless of what auxiliary information an adversary may hold. The next milestone was the mechanism-design toolkit in Dwork, McSherry, Nissim and Smith (2006), which showed how to achieve (ε, δ)-DP by adding random noise

to a sensitive variable,[7] thereby turning privacy into an engineering discipline in which one composes private primitives into more complex analyses while explicitly accounting for the resulting (ε, δ) privacy loss.

After these advancements, DP had expanded beyond 'add noise to numeric variables' and become a new paradigm in the field. An important problem during this time was the use of auxiliary information and multi-release environments. Ganta et al. (2008) showed that combining independently 'anonymized' releases about overlapping populations can defeat many traditional protections (such as k-anonymity), but they proved that DP based methods provide appropriate protection against such attacks. A final important insight emerged in Dwork and Naor (2010), where they showed that absolute disclosure prevention is impossible. This result refers back to the definition of Dalenius (1977) from the 1970s, with 'learn nothing new' requirement from newly published data. The implication is important not only from a historical point of view, but it also shows that complete privacy cannot be achieved for econometrically useful releases in the presence of side information. From this point onward, the data protection literature explicitly balances privacy risks and statistical utility rather than trying to minimize privacy risks.

15.2.6 2010s: From Theory to Deployment of Operational Differential Privacy

During the 2010s, differential privacy (DP) matured from a primarily theoretical definition into a family of models that differ in who is trusted and where randomization occurs. In the *central* (often called 'global') model, a trusted curator holds the raw microdata and releases privatized statistics or outputs satisfying DP criteria. These settings emphasized in the core DP toolkit and its composition theorems (e.g., Laplace/Gaussian noise for numeric variables). In contrast, *local differential privacy* (LDP) treats the curator as untrusted. Each respondent randomizes their own data before transmission, so privacy holds even against the data collector (Dwork & Roth, 2014). The literature formalized LDP as a strong 'trust-minimizing' model and developed sharp privacy–utility characterizations. Duchi, Jordan and Wainwright (2013) and Duchi, Jordan and Wainwright (2014) provided minimax and information-theoretic analyses that show that LDP generally reduces statistical efficiency relative to central DP.[8] From a practitioner's perspective, involving economic problems, the 2010s saw further refinements in the application of DP. Chen, Mohammed, Fung, Desai and Xiong (2011) proposed a method to privatize set-valued variables. Holohan, Leith and Mason (2016) extended the methodology to categorical values.

A central message of DP became clearer: DP imposes an explicit trade-off between privacy and statistical accuracy. Improving privacy generally requires greater injected

[7] To be more precise, Dwork, McSherry et al. (2006) talks about calibrating random noise to the query's global sensitivity, which relates the maximum change in the query induced by altering one individual's record. This can be related to the (ε, δ)-DP definition.

[8] Often interpretable as a reduction in effective sample size for fixed privacy.

randomness, which reduces statistical accuracy. Statistically, this shows up as slower convergence rates and wider confidence intervals when inference is done correctly on privatized outputs. However, if the exact downstream econometric task is not incorporated during the privatization mechanism,[9] the results can be biased and inconsistent. For official statistics and policy, this trade-off can be framed as a social choice problem: statistical agencies jointly allocate accuracy and privacy, and increasing one decreases the other. Economic treatments emphasize choosing operating points where the marginal social benefits of accuracy balance the marginal privacy costs under a DP production possibility frontier (see, e.g., Abowd & Schmutte, 2019).

15.2.7 Recent Progress and The Challenge of Remaining Relevant

As is apparent, differential privacy has shifted from a definition and noise-mechanism toolkit to a deployment-ready engineering discipline with a much richer methodological interface to statistics. Even though DP has many strands of research, let us mention here only two of them — the most important from an econometric point of view.

First, a large literature studies inference and hypothesis testing with DP. During the last decade, the literature has moved beyond 'DP breaks classical inference' toward constructing valid uncertainty quantification, including confidence intervals and tests that explicitly model privatization noise. Sheffet (2017) investigated the linear regression framework and shows under which assumptions confidence intervals and t-like statistics can be approximated. In parallel, hypothesis testing under DP has matured into both optimal finite-sample constructions (e.g., uniformly most powerful tests in certain discrete settings) and general frameworks that can systematically 'privatize' classical tests with analyzable power losses, addressing the practical need for p-values under privacy constraints (see e.g., Awan & Slavković, 2018; Kazan, Shi, Groce & Bray, 2023). Aligned with this approach, Farzam and Sapiro (2024) discussed the use of DP for causal inference with the potential outcome framework. They showed that applying DP mechanisms to outcomes preserves the unbiasedness of average treatment effect (ATE) estimates, but it can increase their variance and severely distort conditional average treatment effect (CATE) estimates. Similar results were discussed by Niu et al. (2022), who showed that heterogeneous effect estimation is typically fragile as it depends on finer-grained conditional structure that is disproportionately affected by privacy noise and composition.

The second research agenda is formal DP variants and accounting. Modern large-scale DP systems increasingly rely on privacy-loss accounting frameworks such as Concentrated DP (CDP, Dwork & Rothblum, 2016), zero-Concentrated DP (zCDP, Bun & Steinke, 2016), and related Rényi-divergence (Mironov, 2017; Wang, Balle & Kasiviswanathan, 2021). These are formalisms that account for repeated queries

[9] A privatization mechanism is a procedure that maps confidential data into a released output designed to protect privacy while preserving some statistical utility.

and iterative procedures. This is extremely relevant, as applied economic analysis involves multiple requests for privatized data due to changes in the objectives of the downstream analysis. Examples of such changes include using different model specifications, using model selection, reporting robustness checks, or interactive exploration in the initial phase. These multiple requests consume privacy budget, hence increasing the combined probability of leakage. Another implementation of such a formalism is when a vast amount of data is used in the privatization mechanism, and many tables are created. Abowd et al. (2022) provided a detailed example on how the U.S. Census Bureau's 2020 Disclosure Avoidance System (the TopDown Algorithm) could be combined with zCDP accounting, and then translated to (ϵ, δ) for communication, to manage cumulative privacy loss while producing national-to-block-level tabulations.

Although it sounds profound in theory, in practice, there are shortcomings in the DPs implemented at a very large scale. The U.S. Census Bureau used the DP framework when releasing the 2020 Census tabulations and encountered a trade-off between privacy and efficiency in the process. Del Vasto-Terrientes, Sánchez and Domingo-Ferrer (2025) describes the process in detail; here, let us mention the main conclusion. The initial privacy budget parameter of the Bureau was $\varepsilon = 4.5$, which is considered as a moderate privacy parameter.[10] However, this value still seemed too narrow, and after 4 years of work, the largest individual ε value was 39.9, whereas the combined value of the whole report is 52.83. These values are vastly larger than the original, posing a serious threat to privacy[11]. Meanwhile, Eurostat's most recent guideline on SDC for census and demographic data releases does not contain a DP recommendation (Eurostat, 2024). Eurostat suggested that national statistical offices use a combination of Targeted Record Swapping (TRS) and the Cell Key Method (CKM), approaches that originated in the 1990s.[12] However, as de Vries, Golmajer, Tent, Giessing and de Wolf (2021) reported, only around half of national statistical offices implemented TRS, and 60% of them used CKM in 2021. Even though the main statistical offices are struggling with the implementation of DP, private sector companies such as LinkedIn, Facebook or Google successfully used DP with ε values between 1-2 (see more in Del Vasto-Terrientes et al., 2025).

Despite the rapid theoretical progress, integrating DP into mainstream econometric workflows remains nontrivial. Although there is an emerging literature on statistical

[10] Recall that 0 is no information leakage that is uninformative due to the large noise added. 1 is the proposed value by Dwork, 2008, which shows a useful balance between accuracy and privacy. Above that, there are only rules of thumb. E.g., around 5, there is a noticeable risk of leakage, and around 40, very weak privacy guarantees in practice. Unfortunately, there is no direct interpretation nor agreement in the literature on the recommended value of ε, see Dwork, Kohli & Mulligan, 2019.

[11] Note that there is an exponential component; hence one needs to compare $e^{4.5}$ with $e^{52.83}$, when multiplying with the probability value.

[12] TRS protects privacy by swapping unique and sensitive information (like high income or rare occupation) between similar households in small geographic areas, which prevents identification of households based on the swapped information. CKM assigns random noise to each observation (each cell) and stores it in a separate "key" table. During aggregation, the random noises stored in the key are also aggregated and added to the population counts. Thanks to the key, the added noise for each observation is the same across data releases and geographic aggregations.

inference with DP,[13] in general, standard asymptotic logic, along with model specification, is disrupted by privacy randomization. Furthermore, econometric workflows routinely involve different models and specification comparisons, such as variable selection, functional form checks, and many robustness specifications. Under DP, each specification consumes privacy budget through composition, so either (a) the privacy cost becomes large, forcing heavy noise that undermines power and may destroy a useful signal, or (b) analysts must pre-commit to a small set of specifications, which clashes with common empirical practice. Finally, even for a canonical task like linear regression, achieving usual inferential outputs (p-values, confidence intervals, etc.) under DP typically requires nonstandard algorithmic constraints (bounded covariates and outcomes, regularization, careful random projections, or specialized privatized estimators) and additional conditions under which classical-looking inference can be approximately recovered (Sheffet, 2017). This makes the econometric analysis nontrivial, whose results may depend on many parameters, quite the opposite of what is actually preferred in most empirical econometric applications.

15.3 Unrevealable Confidentiality: Sensitive Data that are Not Directly Observable

Next let us turn to cases in which privacy concerns affect measurement itself. In these settings, respondents may be unwilling to report exact values for sensitive variables such as income, wealth, or health status. Instead, one observes brackets, ordered categories, censored values, missing responses, or other indirect signals. The econometric problem is therefore not how to protect confidentiality, but how to draw inference when the variable of interest is not directly observed.

15.3.1 The Fundamental Challenge of Unobserved Values

In many mid-19th-century surveys and administrative settings, economically important variables (income, expenditures, health status, etc.) could not be collected precisely due to confidentiality commitments. Consequently, agencies and researchers relied on designs that increased response rates, such as coarsened categories, censoring/truncation, and indirect questioning. This practice resulted in variables whose values are essentially distorted and observed only when they exceed (or fall below) a limit or fall into a bracket. In contrast, the underlying unobserved variable remained continuous, and the analyst was interested in doing inference on the parameter(s) of that process. During this time, it was already clear that the way a variable is distorted plays an important role in identifying and estimating the parameter(s) of interest.

[13] For discussion of DP applications to survey data, see Drechsler and Bailie (2024); Seeman, Si and Reiter (2026).

First, let us discuss censoring and truncation. Censoring arises when observations at the limit are recorded (e.g., $y = \bar{y}$ for top-coded income). In contrast, truncation arises when observations beyond the limit are not observed at all (e.g., due to eligibility thresholds or survey designs that exclude part of the population), which is conceptually important because it changes both the likelihood and the sampling frame. The Tobin (1958) model formalized censoring by defining a latent linear index $y_i^* = x_i'\beta + u_i$ and an observed outcome $y_i = \max\{0, y_i^*\}$ (or, more generally, censoring at known bounds), enabling consistent estimation when privacy or survey design generates mass points at boundaries. Complementarily, Amemiya (1973) provided a rigorous treatment of truncated-normal regression, clarifying identification and asymptotic properties when the dependent variable is only observed conditional on being above a truncation point, which is directly relevant for privacy-constrained or administratively filtered samples. Heckman (1976) unified modeling approaches on truncation, sample selection, and limited dependent variables, as all of them share a common structure that can be handled in a unified framework. This unification helped apply work when confidentiality policies or field protocols determine who is observed and which values are released. These models and their historical lessons are detailed more in Chapter 10.

Another case in which the variables are observed through ordered categories. These types ordered the continuous latent variable into ordered values. Models that handle such variables are embedded in the maximum likelihood framework, which treats ordinal values as thresholded observations of a continuous variable. The privacy connection is that ordered categories (e.g., income bands, self-rated health, risk scales) reduce identifiability by limiting granularity, while still permitting estimation of systematic relationships through a latent-variable model. Formally, these models use a similar latent variable $y_i^* = x_i'\beta + \epsilon_i$, but now, there are $j = 1, \ldots, J$, cutpoints with values of κ_j. Then the observed response satisfies $y_i = j$ if and only if $\kappa_{j-1} < y_i^* \leq \kappa_j$, so confidentiality-induced coarsening is treated as an explicit measurement layer. Aitchison and Silvey (1957) provided an early probit framework for multi-category responses in biometrics, establishing the latent-threshold logic that later became standard for ordered outcomes. McKelvey and Zavoina (1975) then generalized and popularized the approach for social-science data with multiple covariates and maximum likelihood estimation, making ordered probit/logit a popular tool in survey-based econometrics. These models have been used with sensitive variables that are observed in ranked form. This format aligns with confidentiality discussions of that decade (recall that in the 1970s, the practice of coarsening was widespread in statistical offices). A natural follow-up to these models was the random-utility/discrete-choice models, in which the modeled variable is unobserved (not only due to data privacy) and only choices are recorded. The privacy connection to these models is that surveys and administrative records can often store (or release) a discrete choice (e.g., a selected option, participation decision, etc.) with lower disclosure risk than a full cardinal valuation or a detailed attribute report, while still allowing researchers to statistically infer underlying utilities. In random utility models, the analyst assumes latent utilities $U_{ij} = V_{ij}(x) + \epsilon_{ij}$ and observes only the maximizing alternative, so the structure is not directly observable. McFadden

(1974) conditional logit framework showed how to estimate choice probabilities as functions of alternative attributes, turning qualitative choice data into estimable preference parameters without requiring disclosure of an individual's full latent utility. Retrospective assessments emphasize that this contribution was foundational for the econometric analysis of discrete choice and catalyzed wide application in settings where collecting or releasing richer microdata would create confidentiality concerns (see Manski, 2001).

Another type of variable is when the unobserved variable is mixed with additional noise. This type is particularly relevant when privacy constraints require replacing direct measurement with noisy indicators, such as with controlled/random rounding (Fellegi, 1975), or record swapping (Dalenius & Reiss, 1982). The econometric handling of such data was models with measurement errors. In the classical errors-in-variables framework, naïve estimators (e.g., OLS) are generally biased and inconsistent when regressors are measured with error, and point identification requires extra information such as validation data, repeated measurements, instrumental variables, or parametric restrictions on the error process (see e.g., Goldberger, 1972; Chamberlain & Griliches, 1975; Pollard, Bobbitt & Bergner, 1978). These are the kinds of auxiliary inputs that agencies can provide (under secure access) even when the microdata themselves are perturbed for confidentiality. In survey-based economics specifically, validation-study evidence and synthesis work emphasize that measurement error is often nonclassical[14] and usually biased downward, leading to too narrow confidence intervals (see e.g., Heeringa, Berglund & Khan, 2011).

Let us mention an interesting research agenda at that time, in which measurement error and a latent-variable system were jointly used. Jöreskog (1970) or Fuller (1987) are such examples: multiple perturbed releases or multiple noisy proxies are indicators of a latent sensitive variable. In such settings, one can exploit cross-equation restrictions for identification and estimation, aligning closely with the idea that privacy protection often replaces a single exact report with several less-revealing signals. These approaches provided a complicated but working solution under the assumption that the model is well specified. Their limitation, however, is that they fail to identify and estimate the true parameter(s) when the original data-generating process is unknown or the correction mechanism is only approximate.

The last case is when one observes no response in surveys or values are removed as they would reveal identity (e.g., extreme values). No response in surveys falls into two categories: complete nonresponse (the respondent refuses to complete the survey) and item nonresponse (the respondent refuses to answer a specific question). Both can be correlated with the sensitive value being protected, which leads to biased estimates in naïve modeling frameworks (e.g., linear regressions with OLS). Rubin (1976) introduced the 'missing at random' and 'ignorability' conditions as conceptual milestones, stating that correct inference requires modeling or justifying the decision to ignore the missingness mechanism. These mechanisms were primarily modeled using likelihood-based methods, which were also developed during this era.

[14] For example, nonclassical measurement errors occur when the error is correlated with true values. See the early discussion in Griliches (1974).

15.3.2 Advancements Between 1980 and Early 2000

During the early 1980s, the common understanding across econometrics and official statistics was that disclosure avoidance often behaves like measurement error or partial observability, thereby enabling the use of the aforementioned techniques to yield identification and valid standard error calculations under the *properly* specified model. In the following decades, these assumptions were challenged step by step, and newer methods offered more flexible solutions within the same paradigm.

First, the missing-data paradigm matured into a coherent likelihood-based toolkit, including weighting, likelihood factorization under ignorability, and principled sensitivity analyses. Little and Rubin (1987) made explicit that incomplete data undermine naïve point identification and require modeling assumptions about the response mechanism, which might be true or not.

Another methodological consolidation from the 1980s through the early 2000s affected ordered models by relaxing parametric assumptions. Turnbull (1976) proposed a nonparametric estimator for arbitrarily grouped, censored, and truncated data, and defined the nonparametric maximum likelihood estimator (NPMLE) of the distribution function. He also provided a monotone 'self-consistency' algorithm (closely related in spirit to expectation maximization or EM) which converges to the likelihood maximizer under interval censoring. This extension not only provided nonparametric distributional treatment but also widened the scope of discretized variables that can be used within this framework. Building on this, the 1980s literature developed likelihood-based regression under interval censoring, most prominently extending semiparametric survival models. Finkelstein (1986) derived methods for fitting Cox-type proportional hazards models with left-, right-, and overlapping interval-censored observations and provided practical testing procedures that generalize rank-based comparisons when exact failure times are unavailable. The theoretical and computational maturation was synthesized in Groeneboom and Wellner (1992), who develop information-bound arguments and clarified when root-n rates and efficient influence-function representations are attainable. They also analyzed the distribution theory of the NPMLE in canonical interval-censoring models and present practical algorithms, including EM-type procedures and the iterative convex minorant (ICM) approach, to compute the estimators. This likelihood-centric perspective also fed back into econometric practice for discrete and coarsened outcomes. Interval-censored dependent variables (e.g., bracketed income) can be treated as observations of a latent continuous variable falling between reported bounds, yielding interval-regression and ordered-response likelihoods as special cases, and motivating systematic sensitivity checks that relax strict parametric assumptions on the latent distribution. From a data-privacy standpoint, these developments are relevant because official statistics and survey practice increasingly use bracket/unfolding-bracket designs for sensitive items to reduce item nonresponse and respondent discomfort while limiting identifiability, thereby making interval censoring an intentional measurement channel rather than an accidental nuisance (see e.g., J. C. Moore & Loomis, 2000). The broader lesson of the 1980s–2000s MLE program are thus twofold: (i) likelihood methods can recover valid inference under incomplete observation by modeling

the observation mechanism explicitly, and (ii) nonparametric and semiparametric MLEs (as in Turnbull-type NPMLEs and Cox-type interval-censored models) offer principled ways to relax distributional assumptions that would otherwise be imposed for theoretical or computational convenience.

Alongside maximum likelihood methods, the measurement error literature also experienced dynamic development. A new line of research established why different noise-inducing procedures lead to different parameter estimation results. In the early 2000s, the literature increasingly emphasized how measurement error leads to generally biased and inconsistent results (see, e.g., Wansbeek & Meijer, 2000; Hausman, 2001; Bound, Brown & Mathiowetz, 2001). Alternatives to classical error – such as Berkson-type error and 'optimal prediction' reporting models – became more prominent during this era, and newer results showed how these nonclassical errors change the direction and location of bias relative to the classical case. For example, Hyslop and Imbens (2001) argued that different information sets can yield different estimands even under the same observed data, so the privacy mechanism must be modeled (or bounded) jointly with the behavioral relationship of interest.

Measurement error methods also expanded beyond linear regression. Carroll, Ruppert, Stefanski and Crainiceanu (2006) provided an overview of different techniques, such as regression calibration, corrected score methods, likelihood and quasi-likelihood approaches, and Bayesian treatments, showing how to recover inference in nonlinear models when the analyst observes only noisy proxies or interval/coarsened measures. In the survey context, Bound et al. (2001) argued that bounding arguments, external records, or dedicated validation subsamples can improve identification and estimation as well, especially where access to confidential 'truth' files is restricted but can be used to calibrate public use perturbations.

To summarize the measurement error literature from a data privacy perspective, this literature treated disclosure-avoidance restrictions (top-coding, coarsening, etc.) as data limitations that must be combined with properly defined models and assumptions. To make the measurement error approach more feasible in this era, many researchers relaxed linearity conditions and/or used models with more complex distributional assumptions. However, the resulting estimates inherently depend on these assumptions, and when such assumptions are violated, the resulting parameters might be (vastly) different.

Although the aforementioned advancements provided more flexible insights into the parameters of interest, they all rely on point identification via distributional assumptions. By the late 1980s and early 1990s, identification-centered questions became more prominent. As official statistical practices shifted to inject uncertainty about small-area or sensitive cells (see, e.g., Griffin et al., 1989), the privacy protection mechanism became so nuanced that it cannot be supported by modeling the mechanism by simple (linear) equations. This led to questioning whether the downstream econometric estimands are point-identified from the released data without complicated assumptions about the protection mechanism and its interaction with the analyst's model. Along with this fact, econometric identification theory increasingly embraced the possibility that point identification is unattainable in settings with nonrandom selection and privacy-induced data limitations. Manski's (1990) work

on nonparametric bounds on treatment effects made this explicit by showing how to derive set identification when only one potential outcome is observed and strong identifying assumptions are not credible. These insights foreshadow the emergence of the 'partial identification' boom during the 2000s, which provided a different treatment.

15.3.3 The Identification Problem

A defining development of the 2000s was the formalization of 'partial identification' as an inferential paradigm. This approach allowed econometricians to think about confidentiality procedures not as a measurement error but, more fundamentally, as an identification problem. Manski and Tamer (2002) studied regression problems where one regressor or the outcome is observed only through bounds (interval data) and show how independence restrictions yield identification regions and set estimators (e.g., modified minimum-distance or modified maximum-score procedures). While the proposed approach provided identified sets rather than point identification under less restrictive assumptions, the resulting regions are often economically too wide, which limits routine adoption in applied work when practitioners desire sharp point estimates rather than ranges. Manski (2003) consolidated this logic across missing-data, contaminated-measurement, and treatment-response problems, emphasizing that credible inference can proceed via set-valued conclusions that transparently track the strength of the maintained assumptions. A second breakthrough was the systematic development of estimation and inference methods for moment (in)equalities, which generalized the idea of partially identified parameters developed in Manski and Tamer (2002). Chernozhukov, Hong and Tamer (2007) provided a general framework for estimation and confidence regions for identified parameter sets, defined by criterion functions and moment inequalities/equalities, with applications explicitly including regressions with missing or mismeasured data.

Subsequent work improved power and robustness in these models. D. Andrews and Soares (2010) introduced generalized moment selection (GMS) confidence sets with correct uniform size and substantially better power relative to subsampling or plug-in asymptotics. D. Andrews and Shi (2013) extended the framework to conditional moment inequalities via instrument transformations and GMS critical values. These advances are relevant to data privacy because many confidentiality-preserving collection and dissemination strategies intentionally replace exact values with intervals or inequalities, thereby creating the incomplete-information structures that partial identification is designed to handle (see more in J. C. Moore & Loomis, 2000). A practical lesson of the 2000s is therefore the importance of emphasizing parameter identification. Rather than treating interval measurement as a latent variable or a measurement error problem and imposing assumptions on the data-generating process that enable point identification, attention shifted towards the realization of partial identification and the limitations of what can be learned from such data without restrictive assumptions. Although this approach is theoretically appealing, in practice

it lacks several important features that have prevented its spread. First, the resulting parameters remained too wide in practice, and with a larger dataset[15] computational burden of set inference becomes an issue. Finally, and probably most importantly, the difficulty of communicating set-valued results compared to point estimators to policy audiences remained a persistent barrier to wider routine use.

Alongside the breakthroughs in partial identification in the 2000s, simulation-based and Bayesian/quasi-Bayesian tools began to emerge. These methods are used to extract information from incomplete models without committing to full structural specification. A key computational innovation was the 'Laplace-type' (LTE) approach of Chernozhukov and Hong (2003), which allowed for estimating partially identified models (as well) efficiently. LTE treats a generic extremum criterion[16] as a quasi-log-likelihood, forms a quasi-posterior, and then uses Markov Chain Monte Carlo (MCMC) to compute posterior means (or quantiles). This approach avoided fragile global optimization in ill-shaped problems, which reduced the computational burden for many partially identified estimators.

Another typical empirical problem in partially identified models is overly wide confidence intervals. Liao and Jiang (2010) derived posteriors for partially identified parameters by moment inequalities using a limited-information likelihood. They showed that the posterior mass concentrates on the identified region, and provided simulation evidence that informative priors can yield sharper inference within the identified set, effectively turning wide identified regions into more concentrated posterior beliefs. Moon and Schorfheide (2012) provided further argument for the Bayesian approach in partially identified models, by showing that the highest-posterior-density regions may lie strictly within estimated identified sets (unlike frequentist confidence sets, which must account for boundary uncertainty), and they recommended reporting the identified set together with the conditional prior information that drives the tightening. A related development mapped posterior uncertainty about point-identified reduced-form objects into posterior statements about membership and functionals of the identified set, providing Bayesian summaries without exhaustive search over the partially identified parameter space (see, e.g., Kline & Tamer, 2016). These approaches gained popularity precisely because they offer a principled way to incorporate external information, thereby enabling sharper inference than the typically wide nonparametric identified sets delivered under weak assumptions. This is particularly important in settings that involve privacy-sensitive variables. When collecting sensitive information using methods such as brackets or unfolding brackets to minimize nonresponse and disclosure risk, the resulting interval data can often yield broad identification regions. To narrow these regions, it is beneficial to incorporate reliable auxiliary information using Bayesian or simulation-based methods to obtain narrower parameter regions.

[15] 1k+ observations with dozens of variables.

[16] E.g., GMM, empirical likelihood, censored/IV quantile objectives.

15.3.4 Clarification of Differences in Different Methods and the Remaining Challenges

After 2010, methodological discussions increasingly emphasized that many familiar 'latent-variable' models yield interpretable estimates only after normalizations and distributional choices that are not innocuous with respect to the estimand (Ho & Rosen, 2017; Molinari, 2020). For ordered and binary latent-index models, identification typically requires fixing the scale of the disturbance (e.g., $\mathsf{var}(\epsilon) = 1$ in ordered probit), so coefficients are inherently identified only up to scale and can shift in meaning across specifications and samples when the normalization is changed or implicitly differs across groups (Choe, Jung & Oaxaca, 2019). As marginal effects and predicted probabilities depend on both the coefficient vector and the assumed link function, changing the link (e.g., probit vs. logit, or a more flexible nonparametric model) can alter the implied economic object even when the qualitative fit looks similar. Although in practice these estimands are usually similar, the theoretical gap persists, preventing practitioners from using theoretical proofs when comparing different model results.

In the partial identification literature, Ho and Rosen (2017) surveyed various models and emphasized that identifying assumptions via moment conditions is crucial. They draw parallels to IV estimation, where the chosen IV alters the conditional distribution of the endogenous variable, thereby altering the result. Similarly, the imposed moment conditions are specific to each empirical problem, and different sets of conditions yield different results. This result assumes a high level of knowledge of the field, enabling the researcher to provide justification for the best available set of moment constraints. Apart from the moment constraint, there are other technical barriers in partial identification, such as test inversion, tuning choices, and scaling with nuisance-parameter dimension. All these help to explain why moment-(in)equality and set-inference tools remain specialized despite their conceptual appeal (see, e.g., Canay, Illanes & Vélez, 2023; I. Andrews, Roth & Pakes, 2023).

Finally, the selection/treatment literature highlighted an important point: models often identify different (causal) parameters (e.g., complier-specific effects versus population averages) under different (identifying) assumptions. Lewbel's (2019) paper, entitled 'The identification zoo', provides a great comparison of these different estimands. The lesson is important. When one compares models with different identifying assumptions, the resulting parameters are not merely a comparison of the same object under more or less restrictive assumptions, but often identify a different object that refers to a different question. Therefore, apparent disagreements across methods frequently reflect differences in the target estimands rather than purely statistical efficiency.

Consequently, the landscape is best described as 'many competing solutions' that are not directly comparable because they refer to different quantities with different sets of assumptions. Unfortunately, none of these econometric methods was developed primarily for data privacy, which explains why there is no single superior method for such a context. Therefore, a specific approach that handles such sensitive variables and combines with the econometric toolbox remains for future work.

15.4 Future Trends and Comparative Discussion

A central future trend is clear: researchers and statistical agencies will keep collecting more data, often richer, higher-frequency, and more linkable, because digital systems make curation and reuse inexpensive, while demand for fine-grained evidence keeps rising. At the same time, this expansion tightens the privacy–accuracy trade-off, since stronger privacy protection typically requires injecting more randomness or releasing coarser summaries. As a result, some form of imperfect observability is likely to remain the default for most publicly released data. Even without formal privacy measures, surveys and administrative practices show significant distortions, while confidentiality-protecting data releases introduce additional inaccuracies due to algorithmic noise.

In practice, the key empirical question is in which situations this matters. Firstly, there are cases where the research question allows the analysis to rely on aggregated statistics that are preserved by design (e.g., certain totals under unbiased rounding). In this case, point estimates of those specific linear estimands can remain approximately unbiased. The efficiency loss shows up in standard errors, reflecting the fact that the analysis does not use all the possibly available information. By contrast, research questions that require the use of individual- or firm-specific values can exacerbate the effect of privacy-preserving methods. However, there are many cases in which such aggregates do not meet the requirements of the analysis, and micro-level data is needed. In such cases, the parameters of interest (e.g., regression slopes, nonlinear functionals, tail risk, heterogeneous effects, etc.) can be materially distorted or even weakly identified unless the privatization mechanism (or credible bounds on it) is explicitly incorporated. This problem can be solved in three ways. The first solution is to hold the data in a secure environment, where the analyst must comply with strong physical and legal protections. Examples are servers located in secure rooms with cameras, confidentiality agreements, etc. The second is when the goal of the econometric analysis is known before the privatization mechanism, so the privacy-preserving method already incorporates the downstream task and modifies the microdata so that the target parameters are (the least) unbiased. The third option is to know the privatization mechanism and incorporate this fact structurally into the econometric model. Let us briefly discuss the details of these approaches.

In restricted-access settings (e.g., research data centers), in addition to high data access costs, the main challenge is the reproducibility of results. Disclosure review processes constrain what can be exported (tables, model output, diagnostic plots), which limits the extent of open replication packages and makes ‘exact reproduction’ contingent on secure access and output vetting (U.S. Census Bureau, 2024). Empirical results, therefore, are hard to verify unless one has access to the same data and can re-run the shared code.[17] Typically, in practice, a common-sense validity check is done instead of such a meticulous check. Another side effect of such access-based privacy-preserving methods is that researchers can build monopolies by restricting

[17] Agencies, therefore, often substitute reproducibility with detailed documentation and internal review for publicly available data (U.S. Census Bureau, 2002).

access to data for privacy reasons. This constrains the use of such rich databases and explains why sharing such data is beneficial, leading us to the second option.

The current leading approach to data privacy is based on differential privacy. This approach offers explicit protections against re-identification and composition attacks, and it aligns well with agency mandates to publish statistics while bounding incremental disclosure risk under arbitrary auxiliary information. Its core weakness, though, is specifying the core model(s) before applying the privatization mechanism; otherwise, the estimands are altered, invalidating statistical inference. A recent paper by Bi and Shen (2023) shows that a modified DP algorithm can be used to estimate the parameters of interest precisely, only if the method is adjusted for a pre-defined (linear) model, which is rarely the only interest in economic research. This incorporates one of the main challenges in empirical analysis: how to investigate patterns, causal estimands, etc., in such data, where classical back-and-forth investigation is not possible. Multiple employment of privatization weakens privacy, whereas incorporating multiple iterations would require such large noise that the resulting parameters would be useless. Recent proposals for such a problem, like Vilhuber (2025), emphasize solutions in which researchers develop code on synthetic or proxy data and then submit encapsulated runs for execution and verification on confidential data, enabling the generation of results without disclosing microdata. This is a mixture of secure servers and the provision of a 'toy dataset' where researchers can learn from the data. Again, one potential threat to such a solution is that the 'toy dataset' has been constructed so that the existing pattern in the real data is washed out. The challenge of replication remains relevant to this approach, as in many cases the privatized dataset is shared under additional nondisclosure agreements; therefore, privatization itself is not a sufficient guarantee.

By contrast, if one takes the data as given, different econometric models can (and will) yield different results, depending on the (identifying) assumption if the exact privatization mechanism is not known. Obviously, if the privatization mechanism is fully disclosed, then there is no data protection. Partial solutions may exist, but they must mitigate the privacy-utility trade-off, and we are not aware of any real-life setups that do so. However, if such a context is possible, it is feasible to build a structural model that incorporates the privacy mechanism directly. Measurement error based models are excellent examples of how to build such models. In practice, there is usually limited or no knowledge of the privatization mechanism. The most intellectually compelling approach nowadays is partial identification, as it demands the least restrictive assumptions. The applicability of partial identification, however, remains limited due to i) an economically too wide range of estimated parameter regions, with hard(er) interpretation and ii) computational and specification burden, especially with many observations and covariates. Due to these limitations, ordered logit/probit models remain one of the most widely used approaches, with average marginal effects reported. Still, the identifying assumptions are strict and apply to the model and distribution specification as well, while different model variants can be compared only up to scale. This makes this family of models less appealing from a theoretical point of view. Lastly, the measurement error literature had a compelling advantage over the ordered model before 2000, as it can, under a well-specified

model, recover the true parameter. However, over the last two decades, these models have rarely been used in practice, unless there is further information about the data-generating process.

A useful decision rule for when to use privacy-preserving methods or simply take the latent-variable approach is to consider using DP with its constraints and to define the target parameter. When an agency must release information broadly (public use outputs) and must remain robust to unknown auxiliary information and repeated releases, privacy-preserving mechanisms with explicit composition properties are the natural choice, even if they require accepting wider uncertainty. However, this wider uncertainty can still be lower than that from an econometric method, where the same uncertainty arises from built-in assumptions.

At a broad level, the choice among methods depends on the institutional setting and the inferential goal. If the main constraint is public release under unknown external information, differential privacy and related privacy-preserving mechanisms are attractive. If the main constraint is incomplete observation of a sensitive variable, latent-variable and partial-identification approaches may be more natural. In practice, many modern applications combine both problems, which is why hybrid designs are likely to become increasingly important.

A promising bridge between 'privacy-preserving' and 'latent-variable' paradigm is proposed by Chan, Matyas and Reguly (2025). They use *split sampling*, in which the sensitive variable is never collected or released in its exact form. Instead, different subsamples receive different discretization schemes, and the analyst combines the resulting information from *both* the observed data and the discretization mechanism to reconstruct the conditional moments needed for point identification while preserving confidentiality by design. In this approach, discretization serves as a privacy mechanism, reducing individual exposure. Whereas the unobserved variable distributional characteristic can be characterized by the combination of multiple discretizations, helping to estimate the parameter of interest in a large variety of models. In this sense, the paper operationalizes a 'measurement system' that is privacy-aware rather than privacy-agnostic. Methodologically, split sampling also clarifies a practical compromise: one can obtain consistent estimators for linear models with OLS, but only because the sampling design supplies additional identifying variation that would be unavailable under a single discretization. Although there is potential in the work of Chan et al. (2025), the exact privacy-preserving properties of their method require further research.

Overall, future work is likely to converge on integrated pipelines that treat confidentiality as part of the data-generating process and deliver data suitable for multi-purpose modeling. For example, DP-aware regression and hypothesis testing procedures that return standard errors and confidence intervals adjusted for privacy noise, and postprocessing or probabilistic inference layers that lower uncertainty without spending additional privacy budget. A second direction is the expansion of privacy-preserving access architectures (verification servers and containerized validation) designed to scale reproducible research on sensitive data while keeping disclosure review tractable and reducing researcher friction. Finally, an important econometric frontier should concern the conditions under which privacy materially

changes conclusions. Developing diagnostics for estimand drift (i.e., which parameter is actually being learned) and robustness checks that account for privacy mechanisms would enhance the credibility of empirical research that uses privatized data.

15.5 Conclusion

This chapter has shown how data privacy evolved from institutional practice and legal protection into a statistical and algorithmic discipline with direct consequences for empirical research. A central contribution of the chapter is the distinction between *revealable* and *unrevealable* confidentiality, which provides the main lens for analyzing how privacy affects econometric work. In the first case, the data are accurately observed but cannot be openly disclosed, in the second, privacy concerns affect measurement itself, so the variable of interest is observed only indirectly or incompletely.

This distinction helps clarify why privacy is not a single methodological problem and why different settings require different tools. Revealable confidentiality naturally leads to disclosure control, secure access, synthetic data, and differential privacy, whereas unrevealable confidentiality calls for latent-variable models, measurement error corrections, or partial-identification approaches. Across both settings, privacy protection can alter estimands, weaken identification, and invalidate conventional inference unless the protection mechanism is explicitly taken into account.

Looking ahead, as data become richer and more linkable, imperfect observability will remain a standard feature of empirical work. Progress will depend on methods and research workflows that treat confidentiality as part of the data-generating process and that make clear when privacy protection materially changes substantive conclusions. We expect the revealable/unrevealable distinction to remain a useful framework for organizing both future methodological developments and applied empirical practice.

References

Abowd, J. M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., . . . Zhuravlev, P. (2022). The 2020 census disclosure avoidance system topdown algorithm. *Harvard Data Science Review*. doi: 10.1162/99608f92.529e3cb9

Abowd, J. M. & Schmutte, I. M. (2019). An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, *109*(1), 171–202. doi: 10.1257/aer.20170627

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. In *Proceedings of the 31st international conference on very large data bases* (p. 901–909). VLDB Endowment.

Aggarwal, C. C. (2008). Privacy and the dimensionality curse. In C. C. Aggarwal & P. S. Yu (Eds.), *Privacy-preserving data mining: Models and algorithms* (pp.

433–460). Boston, MA: Springer US. doi: 10.1007/978-0-387-70992-5_18

Aitchison, J. & Silvey, S. D. (1957). The generalization of probit analysis to the case of multiple responses. *Biometrika*, *44*(1-2), 131–140. doi: 10.1093/biomet/44.1-2.131

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, *41*(6), 997–1016.

Andrews, D. & Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, *81*(2), 609–666. doi: 10.3982/ECTA9370

Andrews, D. & Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, *78*(1), 119–157. doi: 10.3982/ECTA7502

Andrews, I., Roth, J. & Pakes, A. (2023). Inference for linear conditional moment inequalities. *The Review of Economic Studies*. doi: 10.1093/restud/rdad004

Awan, J. & Slavković, A. (2018). Differentially private uniformly most powerful tests for binomial data. In *Advances in neural information processing systems.*

Barbaro, M. & Zeller, T. J. (2006, August). *A face is exposed for aol searcher no. 4417749.* Retrieved from https://w2.eff.org/Privacy/AOL/exhibit_d.pdf (Reprint of New York Times article (Aug. 9, 2006) hosted by EFF)

Bi, X. & Shen, X. (2023). Distribution-invariant differential privacy. *Journal of Econometrics*, *235*(2), 444-453. doi: https://doi.org/10.1016/j.jeconom.2022.05.004

Bound, J., Brown, C. & Mathiowetz, N. (2001). Measurement error in survey data. In J. J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (Vol. 5, pp. 3705–3843). Amsterdam: Elsevier. doi: 10.1016/S1573-4412(01)05012-7

Bun, M. & Steinke, T. (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. *arXiv preprint*. Retrieved from https://arxiv.org/abs/1605.02065 doi: 10.48550/arXiv.1605.02065

Canay, I., Illanes, G. & Vélez, A. (2023). *A user's guide to inference in models defined by moment inequalities* (Tech. Rep. No. Working Paper 31040). National Bureau of Economic Research. Retrieved from https://www.nber.org/system/files/working_papers/w31040/w31040.pdf

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). New York: Chapman & Hall/CRC. doi: 10.1201/9781420010138

Causey, B. D., Cox, L. H. & Ernst, L. R. (1985). Applications of transportation theory to statistical problems. *Journal of the American Statistical Association*, *80*(392), 903–909.

Chamberlain, G. & Griliches, Z. (1975). Unobservables with a variance-components structure: Ability, schooling, and the economic success of brothers. *International Economic Review*, *16*(2), 422–449.

Chan, F., Matyas, L. & Reguly, A. (2025). *Modelling with sensitive variables.* Retrieved from https://arxiv.org/abs/2403.15220

Chen, R., Mohammed, N., Fung, B. C. M., Desai, B. C. & Xiong, L. (2011, August). Publishing set-valued data via differential privacy. *Proc. VLDB Endow.*, *4*(11), 1087–1098. doi: 10.14778/3402707.3402744

Chernozhukov, V. & Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, *115*(2), 293–346. doi: 10.1016/S0304-4076(03)00100-3

Chernozhukov, V., Hong, H. & Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, *75*(5), 1243–1284. doi: 10.1111/j.1468-0262.2007.00794.x

Choe, C., Jung, S. & Oaxaca, R. L. (2019). Identification and decompositions in probit and logit models. *Empirical Economics*. doi: 10.1007/s00181-019-01716-2

Cox, L. H. (1980). Suppression methodology and statistical disclosure control. *Journal of the American Statistical Association*, *75*(370), 377–385. doi: 10.1080/01621459.1980.10477481

Cox, L. H. (1987). A constructive procedure for unbiased controlled rounding. *Journal of the American Statistical Association*, *82*(398), 520–524. doi: 10.2307/2289455

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistik Tidskrift*, *15*, 429–444. Retrieved from https://ecommons.cornell.edu/entities/publication/dd721733-1958-4298-b3c6-8b9c4af06a6b

Dalenius, T. (1978, July). *Information privacy and statistics: A topical bibliography* (Tech. Rep. No. Working Paper No. 41). U.S. Bureau of the Census. Retrieved from https://ia800606.us.archive.org/12/items/informationpriva00dale/informationpriva00dale.pdf

Dalenius, T. & Reiss, S. P. (1982). Data-swapping: A technique for disclosure control. *Journal of statistical planning and inference*, *6*(1), 73–85.

Davis, R. C. (1973). *Appendix c: Confidentiality and the census, 1790–1929.* Retrieved from https://archive.epic.org/privacy/hew1973report/appenc.htm

Del Vasto-Terrientes, L., Sánchez, D. & Domingo-Ferrer, J. (2025). *Critical analysis of real-world differential privacy applications in data releases.* Retrieved from https://unece.org/sites/default/files/2025-10/SDC2025_Sd_URV_DelVasto_D.pdf (UNECE Expert Meeting on Statistical Data Confidentiality)

de Vries, M., Golmajer, M., Tent, R., Giessing, S. & de Wolf, P.-P. (2021). *An overview of used methods to protect the european census 2021 tables* (Tech. Rep.). Retrieved from https://unece.org/sites/default/files/2023-08/SDC2023_S3_2_Netherlands_deVries_D.pdf

Dinur, I. & Nissim, K. (2003). Revealing information while preserving privacy. In *Proceedings of the twenty-second acm sigmod-sigact-sigart symposium on principles of database systems* (p. 202–210). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/773153.773173

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: Theory and implementation* (Vol. 201). Springer. doi: 10.1007/978-1-4614-0326-5

Drechsler, J. & Bailie, J. (2024). The Complexities of Differential Privacy for Survey Data. *National Bureau of Economic Research*, Working Paper 32905.

Drechsler, J. & Reiter, J. P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the german iab establishment survey. *Journal of Official Statistics*, *25*(4), 589–603.

Duchi, J. C., Jordan, M. I. & Wainwright, M. J. (2013). Local privacy and statistical

minimax rates. In *2013 ieee 54th annual symposium on foundations of computer science* (p. 429-438). IEEE Computer Society. doi: 10.1109/FOCS.2013.53

Duchi, J. C., Jordan, M. I. & Wainwright, M. J. (2014). *Local privacy, data processing inequalities, and statistical minimax rates.* Retrieved from https://arxiv.org/abs/1302.3203

Duncan, G. T. & Lambert, D. (1986). Disclosure-limited data dissemination. *Journal of the American Statistical Association*, *81*(393), 10–18.

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming, part ii* (Vol. 4052, pp. 1–12). doi: 10.1007/11787006_1

Dwork, C. (2008). Differential Privacy: A Survey of Results. In M. Agrawal, D. Du, Z. Duan & A. Li (Eds.), *Theory and Applications of Models of Computation* (pp. 1–19). Berlin, Heidelberg: Springer. doi: 10.1007/978-3-540-79228-4_1

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I. & Naor, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In S. Vaudenay (Ed.), *Advances in cryptology - eurocrypt 2006* (pp. 486–503). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dwork, C., Kohli, N. & Mulligan, D. (2019, Oct.). Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality*, *9*(2). doi: 10.29012/jpc.689

Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference (tcc 2006)* (Vol. 3876, pp. 265–284). doi: 10.1007/11681878_14

Dwork, C. & Naor, M. (2008, August). *On the difficulties of disclosure prevention in statistical databases or the case for differential privacy* (Tech. Rep.). Microsoft Research / Weizmann Institute of Science. Retrieved from https://www.wisdom.weizmann.ac.il/~naor/PAPERS/imp_disclosure.pdf

Dwork, C. & Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, *2*(1). doi: 10.29012/jpc.v2i1.585

Dwork, C. & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, *9*(3–4), 211–407. doi: 10.1561/0400000042

Dwork, C. & Rothblum, G. N. (2016). *Concentrated differential privacy.* Retrieved from https://arxiv.org/abs/1603.01887

Eurostat. (2024). *Guidelines for statistical disclosure control methods for census and demographics data.* Luxembourg. Retrieved from https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/w/ks-01-24-014 doi: 10.2785/4927485

Farzam, A. & Sapiro, G. (2024). Causal inference under differential privacy: Challenges and mitigation strategies. In *Neurips 2024 causal representation learning workshop.*

Fellegi, I. P. (1975). Controlled random rounding. *Survey Methodology*, *1*(2), 123–133.

Fellegi, I. P. & Phillips, J. L. (1974). Statistical confidentiality: Some theory and applications to data dissemination. *Annals of Economic and Social*

Measurement, *3*(2), 399–409.

Fellegi, I. P. & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183–1210. doi: 10.1080/01621459.1969.10501049

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, *42*(4), 845–854. doi: 10.2307/2530698

Fuller, W. A. (1987). *Measurement error models*. New York: John Wiley & Sons.

Ganta, S. R., Kasiviswanathan, S. P. & Smith, A. (2008). Composition attacks and auxiliary information in data privacy. In *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining (kdd)* (pp. 265–273). doi: 10.1145/1401890.1401926

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica*, *40*(6), 979–1001.

Griffin, R. A., Navarro, A. & Flores-Baez, L. (1989). Disclosure avoidance for the 1990 census. In *Proceedings of the survey research methods section, american statistical association.*

Griliches, Z. (1974). Errors in variables and other unobservables. *Econometrica*, *42*(6), 971–998.

Groeneboom, P. & Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*. Basel: Birkhäuser. doi: 10.1007/978-3-0348-8621-5

Hansen, M. H. (1972). Insuring confidentiality of individual records in data storage and retrieval for statistical purposes. In *Proceedings of the november 16-18, 1971, fall joint computer conference* (p. 579–585). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/1479064.1479167

Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives*, *15*(4), 57–67. doi: 10.1257/jep.15.4.57

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, *5*(4), 475–492.

Heeringa, S. G., Berglund, P. A. & Khan, A. (2011). Sampling error estimation in design-based analysis of the PSID data. *PSID Technical Report 11-05*. Retrieved from https://psidonline.isr.umich.edu/Publications/Papers/tsp/2011-05_Heeringa_Berglung_Khan.pdf#page=1

Ho, K. & Rosen, A. M. (2017). Partial identification in applied research: Benefits and challenges. In B. Honoré, A. Pakes, M. Piazzesi & L. Samuelson (Eds.), *Advances in economics and econometrics: Eleventh world congress.* Cambridge: Cambridge University Press.

Holohan, N., Leith, D. J. & Mason, O. (2016). Differentially private response mechanisms on categorical data. *Discrete Applied Mathematics*, *211*, 86-98. doi: https://doi.org/10.1016/j.dam.2016.04.010

Hyslop, D. R. & Imbens, G. W. (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics*, *19*(4), 475–481.

Jöreskog, K. G. (1970). *A general method for estimating a linear structural equation system* (Tech. Rep. No. RB-70-54). Educational Testing Service. doi: 10.1002/j.2333-8504.1970.tb00783.x

Kazan, Z., Shi, K., Groce, A. & Bray, A. (2023). The test of tests: A framework for differentially private hypothesis testing. In *Proceedings of the 40th international conference on machine learning*.

Kline, B. & Tamer, E. (2016). Bayesian inference in a class of partially identified models. *Quantitative Economics*, *7*(2), 329-366. doi: https://doi.org/10.3982/QE399

Lewbel, A. (2019, December). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, *57*(4), 835–903. doi: 10.1257/jel.20181361

Li, N., Li, T. & Venkatasubramanian, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In R. Chirkova, A. Dogac, M. T. Özsu & T. K. Sellis (Eds.), *Proceedings of the 23rd international conference on data engineering, ICDE 2007, the marmara hotel, istanbul, turkey, april 15-20, 2007* (pp. 106–115). IEEE Computer Society. doi: 10.1109/ICDE.2007.367856

Liao, Y. & Jiang, W. (2010). Bayesian analysis in moment inequality models. *The Annals of Statistics*, *38*(1), 275–316. doi: 10.1214/09-AOS714

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, *9*, 407–426.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Machanavajjhala, A., Kifer, D., Gehrke, J. & Venkitasubramaniam, M. (2007, March). L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, *1*(1), 3–es. doi: 10.1145/1217299.1217302

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review*, *80*(2), 319–323.

Manski, C. F. (2001). Daniel mcfadden and the econometric analysis of discrete choice. *Scandinavian Journal of Economics*, *103*(2), 217–229. doi: 10.1111/1467-9442.00241

Manski, C. F. (2003). *Partial identification of probability distributions*. New York: Springer.

Manski, C. F. & Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica*, *70*(2), 519–546.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). New York: Academic Press.

McKelvey, R. D. & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *The Journal of Mathematical Sociology*, *4*(1), 103–120. doi: 10.1080/0022250X.1975.9989847

McKenna, L. (2018). *Disclosure avoidance techniques used for the 1970 through 2010 decennial censuses of population and housing* (Tech. Rep.). U.S. Census Bureau, Center for Economic Studies. Retrieved from https://www.census.gov/content/dam/Census/library/working-papers/2018/

adrm/Disclosure%20Avoidance%20Techniques%20for%20the%201970-2010%20Censuses.pdf

Mironov, I. (2017). Rényi differential privacy. In *2017 ieee 30th computer security foundations symposium (csf)* (p. 263-275). doi: 10.1109/CSF.2017.11

Molinari, F. (2020). Microeconometrics with partial identification. In S. N. Durlauf, L. P. Hansen, J. J. Heckman & R. L. Matzkin (Eds.), *Handbook of econometrics, volume 7a* (Vol. 7, p. 355-486). Elsevier. doi: https://doi.org/10.1016/bs.hoe.2020.05.002

Moon, H. R. & Schorfheide, F. (2012). Bayesian and frequentist inference in partially identified models. *Econometrica*, *80*(2), 755–782. doi: 10.3982/ECTA8360

Moore, J. C. & Loomis, L. S. (2000). Using alternative question strategies to reduce income nonresponse. In *Proceedings of the section on survey research methods.*

Moore, R. A. (1996). *Preliminary recommendations for disclosure limitation for the 2000 census: Improving the 1990 confidentiality edit procedure* (Tech. Rep.). U.S. Census Bureau, Statistical Research Division. Retrieved from https://www.census.gov/content/dam/Census/library/working-papers/1996/adrm/rr96-06.pdf

Narayanan, A. & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the 2008 ieee symposium on security and privacy* (pp. 111–125). doi: 10.1109/SP.2008.33

Niu, F., Nori, H., Quistorff, B., Caruana, R., Ngwe, D. & Kannan, A. (2022). Differentially private estimation of heterogeneous causal effects. In *Proceedings of the conference on causal learning and reasoning*.

Pollard, W. E., Bobbitt, R. A. & Bergner, M. (1978). Examination of variable errors of measurement in a survey-based social indicator. *Social Indicators Research*, *5*, 279–301. doi: 10.1007/BF00352935

Raghunathan, T. E., Reiter, J. P. & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, *19*, 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics*, *18*(4), 531–544.

Reiter, J. P. (2005). Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics*, *21*(3), 441–462.

Reiter, J. P. (2010, Apr.). Multiple imputation for disclosure limitation: Future research challenges. *Journal of Privacy and Confidentiality*, *1*(2). doi: 10.29012/jpc.v1i2.575

Reiter, J. P. & Drechsler, J. (2010). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*, *20*(1), 405–421.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. doi: 10.1093/biomet/63.3.581

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics*, *9*, 461–468.

Samarati, P. & Sweeney, L. (1998). *Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression* (Tech.

Rep. No. SRI-CSL-98-04). Computer Science Laboratory, SRI International. Retrieved from https://www.csl.sri.com/papers/sritr-98-04/

Seeman, J., Si, Y. & Reiter, J. (2026). Toward Differentially Private Finite Population Estimation: An Approach Based on Survey Weight Regularization. *Harvard Data Science Review*, *8*(2). Retrieved from https://hdsr.mitpress.mit.edu/pub/2f0veaek/release/2 doi: 10.1162/99608f92.fdb2338d

Shannon, C. E. (1949). Communication theory of secrecy systems. *Bell System Technical Journal*, *28*(4), 656–715. doi: 10.1002/j.1538-7305.1949.tb00928.x

Sheffet, O. (2017, 06–11 Aug). Differentially private ordinary least squares. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 3105–3114). PMLR.

Sweeney, L. (2000). *Simple demographics often identify people uniquely* (Tech. Rep.). Data Privacy Lab / Carnegie Mellon University. Retrieved from https://dataprivacylab.org/projects/identifiability/paper1.pdf

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, *26*(1), 24–36.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *38*(3), 290–295. doi: 10.1111/j.2517-6161.1976.tb01597.x

United States Federal Committee on Statistical Methodology. (1978, May). *Statistical policy working paper 2: Report on statistical disclosure and disclosure-avoidance techniques* (Tech. Rep.). Office of Federal Statistical Policy and Standards, U.S. Department of Commerce. Retrieved from https://statspolicy.gov/assets/fcsm/files/docs/spwp2.pdf doi: 10.21949/1529887

U.S. Census Bureau. (2002). *Transparency and reproducibility.* Retrieved from https://www.census.gov/about/policies/quality/guidelines/transparency.html (Information Quality Guidelines page)

U.S. Census Bureau. (2003). *Census confidentiality and privacy: 1790–2002* (Tech. Rep.). U.S. Census Bureau. Retrieved from https://www2.census.gov/library/publications/2003/comm/monograph-confidentiality-privacy.pdf

U.S. Census Bureau. (2019). *A history of census privacy protections.* Retrieved from https://www2.census.gov/library/visualizations/2019/communications/history-privacy-protection.pdf

U.S. Census Bureau. (2024). *Federal statistical research data center disclosure avoidance review procedures: A handbook for researchers.* Retrieved from https://www2.census.gov/adrm/FSRDC/Resources/FSRDC-Disclosure-Avoidance-Procedures-Handbook.pdf (FSRDC guidance document)

U.S. Government Publishing Office. (2020). *U.s.c. title 13 — census (enacted aug. 31, 1954).* Retrieved from https://www.govinfo.gov/content/pkg/USCODE-2020-title13/html/USCODE-2020-title13.htm

Vilhuber, L. (2025). Using containers to validate research on confidential data at scale. *Harvard Data Science Review*. doi: 10.1162/99608f92.4d1853ce

Wang, Y.-X., Balle, B. & Kasiviswanathan, S. (2021). Subsampled rényi differential privacy and analytical moments accountant. *Journal of Privacy and Confidentiality*, *10*(2). doi: 10.29012/jpc.723

Wansbeek, T. J. & Meijer, E. (2000). *Measurement error and latent variables in econometrics*. Amsterdam: Elsevier / North-Holland.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63–66. doi: 10.1080/01621459.1965.10480775