A Qualitative Analysis of Practical De-Identification Guides

Wentao Guo
University of Maryland
College Park, MD, USA
wguo5@umd.edu

Aditya Kishore
University of Maryland
College Park, MD, USA
adityak1@umd.edu

Adam J. Aviv
The George Washington University
Washington, DC, USA
aaviv@gwu.edu

Michelle L. Mazurek
University of Maryland
College Park, MD, USA
mmazurek@umd.edu

Abstract

De-identifying microdata is necessary yet difficult. Myriad techniques exist, which reduce risk and preserve utility to varying, often unclear extents. We conducted a thematic analysis of 38 online de-identification guides for practitioners, to understand what content they contain and how they are designed to support decision-making and execution. We highlight trends and differences between guides, and we find some concerning patterns, including inconsistent definitions of key terms, gaps in coverage of threats to de-identification, and areas for improvement in usability. We identify directions for future research and suggest changes to de-identification guidance in order to better support practitioners in conducting effective de-identification.

CCS Concepts

• **Security and privacy → Usability in security and privacy**; **Privacy protections**; **Data anonymization and sanitization**.

Keywords

De-identification; anonymization; guidance; practitioners

ACM Reference Format:
Wentao Guo, Aditya Kishore, Adam J. Aviv, and Michelle L. Mazurek. 2024. A Qualitative Analysis of Practical De-Identification Guides. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3690270

1 Introduction

Sensitive personal data is collected and shared for a variety of purposes that do not require individuals to be identified: tech companies collect usage data to improve their services [5, 7], statistics agencies publish administrative data to create transparency for public funding and policy [1, 18], and researchers share study data to enable replication, meta-analysis, and follow-up work [64, 117]. Tying sensitive information back to individuals can result in various harms: data about sexual behavior may be socially stigmatizing, data about reproductive care may result in legal consequences, and data about political opinion in conflict zones may lead to physical violence. Practitioners aim to reduce this risk by *de-identifying* data—modifying data, or the interface for viewing it, to make it more difficult to re-identify or learn information about individuals.

De-identification techniques are many, and striking a balance between risk and utility can be difficult in the presence of various attacks [21, 74, 96]. Many traditional approaches are time-consuming [6] and offer questionable protection against re-identification [21, 74], while newer ones such as differential privacy [26] may be unfamiliar or unacceptable to practitioners [14, 72] or lack accessible, fully featured tools to assist in implementation [37, 38]. In many cases, as a secondary concern [3] that can interfere with the primary goal of publishing data, de-identification may be treated as an afterthought with limited time and resources.

Online guidance has an impact on how other professionals approach security and privacy tasks [4, 9, 114]. Similarly, well designed guidance could help practitioners de-identify data efficiently and effectively by teaching approaches that are proven, context-appropriate, and accessible. Indeed, the Internet abounds with de-identification advice, ranging from short corporate blog posts to guides containing hundreds of pages. To assess the quality and consistency of these existing resources, we conducted a thematic analysis of 38 recent online guides that explain how to de-identify microdata.[1] We investigate two research questions.

RQ1: What content do de-identification guides contain, particularly with regard to techniques and attacks? We provide a detailed breakdown of techniques and attacks covered by guides, finding that some basic techniques are covered near-universally, but technically complex ones such as differential privacy are infrequently mentioned—especially in guides for researchers. We find that terms such as *anonymization*, *aggregation*, and *differential privacy* are defined inconsistently across guides, sowing potential for confusion. And we observe notable gaps in threat coverage, including overselling the outcomes of de-identification, claims that variables such as salary and medical diagnosis are non-identifying, and patchy coverage of reverse-engineering attacks. We recommend terminology and content updates to address this patchwork of contradictory and misleading information.

This work is licensed under a Creative Commons Attribution International 4.0 License.

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3690270

[1] Microdata refers to data about individuals (e.g., each record in the dataset represents one person or a small group of people), as opposed to aggregate statistics, where released data is about groups or the entire population (e.g., cells in a table contain counts of how many people fall into certain categories).

RQ2: Are guides designed to help readers decide on a de-identification strategy and carry it out? We find that guides do discuss trade-offs to help readers choose between de-identification techniques; however, much of this is vague, merely stating that choosing one approach over another will change the balance between privacy and utility. Several guides contain encyclopedic tables of techniques or instructions, which have the potential to be useful resources but also overwhelming or overly prescriptive. And we find that few guides provide thorough examples of de-identification or real-world case studies of disclosure. We recommend that guides include more examples and case studies, and we identify directions for future work that would help inform how best to inform and teach practitioners about de-identification.

2 Background and terminology

We first cover background on disclosure and de-identification, including both traditional and state-of-the-art perspectives, as both are important context for our analysis of guides. We then turn to related work on challenges and guidance for practitioners, on both de-identification and other security and privacy tasks.

De-identification does not have one definition and is often used semi-synonymously with other terms including *anonymization*, *pseudonymization*, *statistical disclosure control*, and *masking*.[2] In this paper, we define de-identification as a process of modifying data, or the interface for interacting with it, to reduce the risk that someone with access to the data could learn information about individuals. Thus, we are interested in how guides cover topics that range from modifying raw data to prohibiting misuse through data use agreements, but we consider securing data against unintended access (e.g., data breaches) in the first place to be out of scope.

2.1 Disclosure

Any release of de-identified data carries risk of *disclosure*: that attackers may learn information about individuals from the data. To make sense of how de-identification guides talk about disclosure, we break attacks into two component parts: disclosure *types* (what information is revealed) and *mechanisms* (how attackers uncover this information).

In this paper, we refer to three non-overlapping disclosure types, which mirror the three types defined in traditional literature [55]:

- **Identity disclosure**: deducing which record in a dataset refers to an individual. E.g., finding direct identifiers such as email addresses in the dataset, or *singling out* records with unique combinations of indirect identifiers such as occupation and race.[3]
- **Attribute disclosure**: deducing traits of an individual without identifying one specific record. E.g., narrowing an individual down to a subset of records that have certain values in common.
- **Probabilistic inference**: estimating information about an individual with higher likelihood of accuracy, but not certainty. This (as well as attribute disclosure, occasionally) can apply to individuals not in the dataset.

We also refer to three disclosure mechanisms that help to tie de-identified data back to specific individuals:

- **Linking**: connecting de-identified records to external data that is identified. E.g., matching a record to voter registration data using zip code, birth date, and gender.
- **Personal knowledge**: an attacker leveraging their own knowledge of individuals to aid in disclosure. E.g., a work supervisor using their knowledge of co-workers' backgrounds to re-identify them in a company survey.
- **Reverse engineering**: undoing de-identification techniques using knowledge or assumptions about how they were applied. E.g., uncovering a deterministic method of replacing names with pseudonyms. More examples are given in Appendix C in the supplementary materials.[4]

Traditional perspectives on disclosure often divide variables into different classes: direct identifiers, which can identify an individual on their own (e.g., email address, ID number), indirect identifiers, which can identify an individual in combination (e.g., race, location, occupation), and non-identifiers, which are left out of risk considerations (e.g., outcome of a medical treatment, response to an opinion question). A landmark study demonstrated the risk of traditional indirect identifiers, estimating that 87% of the U.S. population is uniquely identifiable by their zip code, birth date, and gender [96].

In reality, though, all variables exist on a spectrum of risk, as demonstrated by a study that identified individuals in the Netflix Prize dataset, by linking de-identified movie and show ratings to named accounts on a public ratings website [74]. Thus, experts on disclosure discourage simplistic distinctions between identifying and non-identifying variables [75, 80].

2.2 De-identification techniques and standards

Basic de-identification techniques include deleting data, pseudonymizing named entities, and generalizing values into less granular categories. These produce data that is still accurate but less specific to individuals. Other techniques produce data that is actually inaccurate, by adding noise, swapping values between records, or generating synthetic data based on properties of the original data. Table 1 contains an overview of the de-identification techniques covered in this work. Carvalho et al. go into much greater depth on *what researchers know* about a wide range of de-identification techniques [17]; our work instead aims to make sense of *how practitioners are informed* about de-identification.

Some standards lay out instructions for how to de-identify data. The Safe Harbor provision of the U.S. Health Insurance Portability and Accountability Act (HIPAA) lays out 18 types of identifiers for healthcare providers to remove from their data [108]. Independent groups may also develop their own standards: the U.S. Agency for International Development's Feed the Future project defines rules for how funded researchers should generalize and add noise to geolocations, dates, and ages before publishing data [45].

[2] Among these variations is the spelling of de-identification (vs. deidentification); here, we elect de-identification, which is significantly more popular according to the Google Books Ngram Viewer [40].

[3] This work focuses on identity disclosure through singling out, as it is more complex to defend against than identity disclosure through direct identifiers.

[4] https://osf.io/mz4p5/

Table 1: An overview of the de-identification techniques and standards we focus on in this work. Carvalho et al. discuss much of the same material in greater detail [17]. Note we include techniques because they were discussed in de-identification guides, not necessarily because we recommend them. Indeed, some are actively discouraged by parts of the research community.

Name	Description	Examples
Delete	Remove data entirely, leaving values blank or replacing them with values such as *N/A*; sometimes called *suppression*.	Remove all values for gender. Remove ages *85* and above.
Pseudonymize	Replace named entities with persistent identifiers. Enables multiple records involving the same entity to be associated without naming them.	Replace all instances of *Martin* with *P1*. Replace *Madison* with *City 1*.
Hash or encrypt	Replace values with meaningless hashes and ciphertexts. Ideally, values should be hashed with a salt that is random, long, and secret, in order to prevent brute forcing. Encryption may or may not be done in such a way that the same value produces consistent ciphertexts, depending on the aims.	Replace *Martin* with its salted SHA-256 hash.
Partially obfuscate	Truncate, scramble, or otherwise modify values in a way that still leaks information about their original form. Can be considered a form of generalization (see below), but we distinguish partial obfuscation in that the modified values are not generally intended to be meaningful.	Replace *Martin* with *M.* (truncated to first letter), *trnima* (scrambled), or ****** (six characters replaced by asterisks).
Subsample	Release only a portion of the entire dataset.	Randomly remove 50% of records.
Generalize	Replace granular values with categories or ranges that encompass them; sometimes called *recoding* or *aggregation*. Special forms include *top-coding* and *bottom-coding*, which group all values above or below a threshold, and *rounding*, which effectively generalizes numbers into ranges.	Replace *Madison* with *Wisconsin*. Replace *Chinese* with *Other*. Replace age *93* with *85+* (top-coding). Round *38* to *40*.
Change data type	Replace values with a different form of data, typically less sensitive or identifying.	Replace hospital stay start and end dates with length of stay.
Add noise	Distort values, e.g. by applying calibrated numerical offsets, or by randomly changing a portion of categorical values. Noise is usually semi-random, though we include manually chosen offsets for the purposes of this work.	Change the date *04-05* to *04-07*. Flip 10% of responses between *Yes* and *No*.
Swap values	Exchange some or all values between records, usually in a systematic way (randomly, or following an algorithm).	Select two records with ages *20* and *24*, and then swap their ages.
Micro-aggregate	Replace groups of individuals with records averaging their values.	Replace three records, ages *20*, *21*, and *22*, with one record, age *21*.
Synthetic data	Generate fake data based on the real data. This can be released as a fully synthetic dataset, or synthetic data can be mixed in with the real data.	Use Bayesian inference to generate data with the same distributions.
k-anonymity	Ensure each individual in the data is identical to at least k-1 others in a specified set of indirect identifiers [97].	Delete and/or generalize data until $k = 5$ for race, gender, and age.
Differential privacy	Algorithmically modify the data (usually add noise) to achieve a formal guarantee limiting the impact of each individual's inclusion in the dataset [26].	Use the Laplace mechanism, $\epsilon = 1.5$, to add noise to all values in a dataset.

Other standards, such as *k*-anonymity, lay out properties of de-identified datasets without hard rules for how to achieve those properties. A dataset is *k*-anonymous if each individual in the data is identical to at least *k-1* others in a set of indirect identifiers that are deemed risky for linking with public data [97]; intuitively, this provides safety in numbers by preventing attackers from linking a known individual to one specific record in the data. *k*-anonymous data can be vulnerable to re-identification using non-traditional identifiers; variants have been developed to address this and other limitations [61, 62], but these may still be vulnerable to reverse engineering attacks, depending on implementation [21].

Many consider differential privacy to be the state of the art for de-identification. Differential privacy algorithms provide a probabilistic upper bound on the impact of any one individual's data on the de-identified dataset, usually by adding noise [26, 28]; this necessarily limits an attacker's ability to re-identify and/or learn information about individuals. The upper bound, which can be tuned using the ϵ parameter, is formally defined and cannot be exceeded (assuming proper randomness), even by processing de-identified data or linking new data [29]. Researchers are still developing differential privacy implementations and evaluating outcomes for new contexts [15, 34, 42, 49]. Differential privacy has attracted controversy from the social science community, primarily due to perceived utility trade-offs [14, 72, 113]. Some privacy researchers [10, 12, 35, 70] have also expressed caution towards differential privacy for similar reasons, though others have demonstrated in specific contexts that differential privacy can actually result in good [69] or even better [20, 60] utility than other de-identification approaches.

2.3 Challenges and guidance for practitioners

A wide variety of practitioners are involved in de-identification—not only privacy experts and programmers at companies and national statistics agencies, but also many who may have fewer resources and expertise, such as social and behavioral researchers who are expected or required by their funding sources to publish sensitive research data [43, 76]. This is important, because de-identification can be quite difficult. Balancing privacy and utility is difficult in practice, whether practitioners employ deletion [50], k-anonymity [6], or synthetic data [94]. Differential privacy also poses challenges: for examples, the lack of satisfactory generalized implementations means that practitioners have had to create their own [37, 38, 42], and choosing an appropriate value for ϵ is not straightforward [27].

Many researchers have aimed to help practitioners carry out de-identification. Some have created and evaluated frameworks for de-identification in different contexts [6, 16, 33, 63, 65]. Others have created visualization tools, for example to help choose a value for ϵ [71]. A plethora of tools exist to assist directly with de-identification; one study evaluated two open-source tools, focusing especially on effectiveness and usability, and found ARX to be better than Amnesia [100].

However, there has not been a systematic review of de-identification guidance online. In a related domain, previous work found that software developers learn to write secure code partly from online sources [9, 114] and that the type of source affects their performance [2]; an analysis of online guidance identified gaps in topics and in learning aids such as examples and tutorials [4]. Our work similarly aims to understand the content and quality of online guidance on de-identifying data.

3 Method

We collected a dataset of 65 de-identification guides between May and December 2023 and conducted qualitative coding on a sample of 38. We describe how we assembled our dataset in Section 3.1 and how we analyzed it in Section 3.2. Though data collection and analysis are presented in distinct sections, we note that these were actually interleaved in iterative phases.

3.1 Collecting guides

Web searches. We began collecting guides through exploratory web searches to better understand the variety of de-identification guides available online. We analyzed an initial sample of 24 using the methods described in Section 3.2. Next, we determined a set of inclusion criteria and systematically searched for each of the following strings using both Google and Bing:[5] "How to de-identify data," "How to anonymize data," "De-identification guide," and "Anonymization guide.

We collected any de-identification guides that were among the top 20 search results[6] and fit the following criteria:

- **Recency:** was published, updated, or reviewed in 2018 or later.[7] We focus on recent guides, as norms and methods for de-identification are evolving rapidly; in particular, 2018 is the year in which the General Data Protection Regulation (GDPR) took effect and the use of differential privacy was announced for the U.S. Census [1].
- **Document type:** is a document published online by humans. We excluded presentations, books, crowdsourced guidance such as forum posts, and AI-written articles.
- **Data type:** covers de-identification of microdata, not exclusively aggregate statistics. We focus on microdata because its granular nature increases the complexity of de-identification.
- **Technical focus:** discusses specific techniques and how they work. We excluded documents that mainly cover a high-level philosophy or describe regulations.
- **Intent:** aims to teach practitioners how to de-identify data. We excluded research papers, as well as webpages informing data subjects about how their data is de-identified.
- **Availability:** is freely accessible online or through our university's library.

The first author reviewed candidate guides and consulted with the second author in cases of ambiguity. We collected at most one guide per organization.

Recommendations. While search engines are often gatekeepers of information online, our systematic searches do not necessarily provide a good representation of high-quality or impactful de-identification guides: search results are prone to manipulation by savvy writers seeking to sell products or ad space, and results vary widely based on specific combinations of search terms. Therefore, we supplemented our data collection by consulting guides recommended by eight organizations (listed in Appendix A) and asking for recommendations from 28 researchers participating in a separate study who have de-identification experience.

Dataset composition. In total, we collected 65 guides: 41 appeared only in online searches, 16 were recommended, and 8 were found both ways. All are publicly available, except one recommended internal guide (R13) that was provided confidentially by a social science research company. We separate guides into two groups: 39 published by government agencies, universities, or other researchers and practitioners, and 26 *blog posts* typically published by a business or tech-related website (which tend to be shorter, are often one of many similar pieces of content on the same website, and also sometimes advertise specific products to help with de-identification).

3.2 Analyzing content

In order to both characterize the content of guides and explore interpretive themes, we followed a template analysis approach to qualitative coding [56]. The first author developed a partial initial codebook covering de-identification techniques and other parts of the process, based on an informal initial review of several guides.

[5] We cleared browsing data between searches to reduce personalization.

[6] We only considered standard web page results as part of this count; i.e., we ignored ads, videos, and AI-generated content. We did, however, include Google's featured snippets, as those are often lifted from among the top standard results.

[7] If a date was not readily available, we looked for references to recent years in the text, checked webpage metadata, and/or identified the earliest snapshot of the current version on the Internet Archive's WayBack Machine (https://web.archive.org/).

The first two authors iteratively refined the codebook while coding guides from our dataset, adding new categories, including outcomes of de-identification, attacks, and rationales for how to choose a de-identification strategy.

The first two authors initially coded one guide collaboratively to flesh out the codebook structure; they then double-coded all remaining guides separately, meeting after every 1–5 to resolve coding differences, update the codebook, and discuss observations.[8] When changes to the codebook potentially affected previously coded guides, the first author recoded previous guides, consulting with the second author in cases of ambiguity. Our codebook is in the supplementary materials.

We used purposive sampling to select guides from our dataset to code, prioritizing guides that appeared different from those we had already analyzed in the following dimensions:

- **Context:** e.g., regulations mentioned, intended audience.
- **Content:** e.g., de- and re-identification techniques.
- **Format:** e.g., presence of examples.
- **Philosophy:** e.g., framing of outcomes.

We also prioritized guides that appeared higher or more frequently in search results, as well as the guides that were recommended.

We analyzed 38 guides in total, continuing until we reached saturation [41]. These included 27 of 39 guides published by government agencies, universities, or other researchers and practitioners, as well as 11 of 26 blog posts. Table 2 contains information about each guide we analyzed, and PDFs are in the supplementary materials. The remaining 27 unanalyzed guides are listed in Appendix D in the supplementary materials.

The two regulations referenced most often by the 38 guides we analyzed are GDPR (N = 15) and HIPAA (N = 13). In total, 29 mention at least one law or regulatory framework pertaining to de-identification.

3.3 Limitations

As this is a qualitative study with 38 guides, we do not claim that results—such as the proportion of guides covering various de-identification techniques—will generalize beyond our dataset. For example, we do not necessarily expect results to generalize across cultural contexts, as our guides are predominantly from organizations based in North America and Europe.

Qualitative data analysis is inherently subjective, especially when interpreting documents that are sometimes unclear or self-contradictory. We aim to reduce the impact of individual bias and error by having two researchers code each guide separately. We also prioritize transparency by providing definitions and examples when describing our methods, including a detailed codebook and documentation of the guides we did and did not analyze.

4 Findings

Now, we present findings from our analysis of 38 de-identification guides. We first synthesize the content covered by these guides, including techniques, other parts of the process, and attacks. (For definitions and context, refer back to Sections 2.1 and 2.2.) Next, we highlight some problematic themes involving inconsistent definitions and gaps in coverage of threats. We conclude with elements broadly related to usability, analyzing stated rationales for how to decide between possible de-identification approaches, as well as examples and other learning aids.

4.1 Content coverage

Table 3 contains an overview of the techniques, attacks, and learning aids included in the guides we analyzed. Notes on how we defined codes and used them to produce these results are in our qualitative codebook, and Appendix B in the supplementary materials elaborates specifically on how we decided whether a guide provided enough detail for us to code a technique.

Differentiation in content. Some techniques are covered by almost all guides, such as generalizing values into broader categories (N = 36) and replacing named entities with pseudonyms (N = 28). Others are less common, such as subsampling only a portion of the collected data (N = 9) and applying k-anonymity (N = 17).

Among disclosure types, singling out is more commonly discussed (N = 30) than attribute disclosure (N = 12) and probabilistic inference (N = 6). This is not surprising, given the focus on singling out in many foundational works related to de-identification [74, 96]. Even when other disclosure types are brought up, they are not typically a focus, with some guides deeming them irrelevant. G3 reads, "Inferential disclosure is generally not addressed . . . since microdata is distributed precisely so that researchers can make statistical inference and understand relationships between variables. In that sense, inference cannot be likened to disclosure." O2 goes further and also excludes attribute disclosure, explaining that "most traditional anonymisation techniques aim to protect against re-identification and not necessarily other types of disclosure risks."

Coverage of disclosure mechanisms is moderately high for linking with external data (N = 27) and reverse engineering de-identification (N = 20), but lower for attackers leveraging their personal knowledge to achieve disclosure (N = 6).

Some guides are restrained in the diversity of content covered—R1 only recommends 3 out of the 15 techniques and standards we focus on—while G1 and G4, at the other end of the scale, both discuss 13. We note that more does not necessarily mean better or worse.

Differences across different types of guides. We note some tendencies that distinguish types of guides from one another. Guides intended for researchers generally leave out technically complex methods, which see far greater coverage in other guides: these methods include differential privacy (N = 1 out of 15 researcher guides vs. 10 out of 23 other guides), k-anonymity (N = 2/15 vs. 15/23), synthetic data (N = 1/15 vs. 12/23), and hashing/encryption (N = 2/15 vs. 15/23). Potential reasons include differing norms or active skepticism [14, 72] in certain communities of researchers, limited awareness of newer techniques such as differential privacy, and perceived limits to researchers' capacity. In contrast, guides

[8] We used coding software to freely apply codes to excerpts from the guides ranging from sentences to paragraphs to graphical elements. When resolving differences, we often compared individual coded excerpts for more nuanced codes (e.g., discussing each excerpt for which either coder had *Examples* or *Impossible to re-identify individuals*) but checked at the document level for more straightforward codes (e.g., confirming that both coders had *Differential privacy* for the guide in question).

Table 2: For each analyzed guide, this table describes who authored it, whom it was written for, when it was published, how many words it contains, and how we found it (see Section 3.1 for details). To help distinguish different types of guides, IDs for all blog posts start with B, while non-blog guides begin with G if written for government agencies; R for researchers; and O for other audiences.

ID	Author	Audience	Year	Words	Source[1]
G1	U.S. National Institute of Standards and Technology [36]	Government agencies	2023	45k	🔍 👍
G2	Canada Secretariat Treasury Board [101]	Government agencies	2023	4k	🔍
G3	International Household Survey Network [11]	Government agencies	2021	78k	👍
G4	eHealth Queensland [30]	Government agencies	2021	13k	🔍
G5	New South Wales Information and Privacy Commission [77]	Government agencies	2020	2k	🔍
R1	Millennium Challenge Corporation [67]	Data submitters	2020	12k	💬
R2	U.S. Agency for International Development [107]	Data submitters	2020	9k	💬
R3	Inter-university Consortium for Political and Social Research [47]	Data submitters	2019*	22k	👍 💬
R4	Johns Hopkins University Data Services [51]	Researchers	2023	6k	🔍! 💬
R5	La Trobe University [59]	Researchers	2023	1k	🔍!
R6	University of Groningen [104]	Researchers	2023	4k	🔍
R7	UK Data Service [103]	Researchers	2021*	4k	👍
R8	Abdul Latif Jameel Poverty Action Lab [57]	Researchers	2020	5k	🔍 👍 💬
R9	The New School Information Security & Privacy Office [99]	Researchers	2020	2k	🔍!
R10	Vrije Universiteit Brussel [109]	Researchers	2020	13k	👍
R11	Portage Network [84]	Researchers	2020	6k	👍
R12	San José State University [88]	Researchers	2019*	1k	🔍 👍
R13	[anonymous company]	Researchers	2018	11k	💬
R14	University of Washington Privacy Office [106]	Researchers/employees	2019	2k	🔍
R15	University of South Carolina Institute for Families in Society [105]	Health researchers	2018	3k	🔍
O1	U.S. Department of Health and Human Services [108]	Healthcare providers	2022	12k	🔍! 👍
O2	Singapore Personal Data Protection Commission [82]	Non-specific	2022	14k	🔍! 👍
O3	UK Anonymisation Network [31]	Non-specific	2020	41k	👍
O4	Irish Data Protection Commission [48]	Non-specific	2019	7k	🔍
O5	Spanish Data Protection Agency [93]	Non-specific	2019*	4k	🔍
O6	Office of the Australian Information Commissioner [78]	Non-specific	2018	8k	🔍 👍
O7	Office of the Victorian Information Commissioner [79]	Non-specific	2018	2k	🔍
B1	International Association of Privacy Professionals [86, 87]	Lawyers	2020	2k	🔍!
B2	Immuta [95]	Businesses	2023	3k	🔍!
B3	Corporate Finance Institute [22]	Businesses	2023	1k	🔍!
B4	k2view [52]	Businesses	2023	4k	🔍
B5	MOSTLY AI [68]	Businesses	2023	2k	🔍
B6	Privitar [85]	Businesses	2022	2k	🔍
B7	Imperva [46]	Businesses	2020*	1k	🔍!
B8	Okera [81]	Businesses	2020	4k	🔍
B9	Pangeanic [89]	Non-specific	2023	2k	🔍!
B10	Satori [90]	Non-specific	2021	16k	🔍!
B11	Towards Data Science [66]	Non-specific	2020	3k	🔍!

[1]🔍 = web search 🔍! = among the top 5 results in a systematic web search
👍 = recommendation by an organization 💬 = recommendation by a researcher
*Indicates a date that represents our best estimate.

for government agencies tend to cover a wide variety of content—with 14 coded techniques and attacks per guide on average, as compared to 9 for other guides—though only contained examples of de-identification. Meanwhile, blog posts discuss partial obfuscation of data much more frequently (N = 9 out of 11) than other guides (N = 4 out of 27), but they discuss attacks less frequently, with 2 out of 11 not covering attacks in any detail at all (compared to 1 out of 27 other guides).

While there are some trends in the kinds of content contained by different types of guides, there is a great deal of variation among

Table 3: Techniques and standards, attacks, and learning aids covered in each guide that we analyzed. Definitions for techniques and standards are in Table 1, while more detail on how we coded for each concept is in the codebook.

	Techniques and standards															Attacks						Aids		
ID	Delete direct IDs	Delete indirect IDs	Pseudonymize	Hash or encrypt	Delete values	Partially obfuscate	Subsample	Generalize	Change data type	Add noise	Swap values	Micro-aggregate	Synthetic data	k-anonymity	Differential privacy	Singling out	Attribute disclosure	Probabilistic inference	Linking	Personal knowledge	Reverse engineering	Examples[1]	Case studies	Tools
G1	●	○	●	●	●	○	●	●	●	●	●	●	●	●	●	●	●	●	●	○	●	○	●	●
G2	●	○	●	●	○	●	○	●	○	●	○	○	●	○	○	●	○	○	●	○	●	○	○	○
G3	●	○	○	○	●	○	●	●	○	●	●	●	○	●	○	●	●	●	●	●	●	●	●	●
G4	●	●	●	●	●	○	●	●	●	●	●	○	●	●	●	●	●	●	●	●	●	○	○	○
G5	●	○	●	●	●	○	○	●	○	○	●	●	○	○	●	●	○	○	●	○	●	○	●	○
R1	●	●	○	○	○	○	○	●	○	○	○	○	○	○	○	●	○	○	●	○	●	○	○	○
R2	●	●	●	○	○	○	○	●	●	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○
R3	●	●	●	○	●	○	●	●	●	●	●	○	○	○	○	●	○	○	●	○	○	○	○	●
R4	●	●	●	○	●	●	○	●	○	●	●	●	○	○	○	●	●	○	●	○	○	◐	○	●
R5	●	○	●	○	○	○	○	●	●	●	○	○	○	○	○	●	○	○	●	○	○	●	○	○
R6	●	○	●	○	○	●	○	●	○	●	●	○	○	○	○	●	○	○	○	○	○	●	○	○
R7	●	○	●	○	○	○	○	●	●	○	○	○	○	○	○	●	○	○	●	○	○	●	○	●
R8	●	●	●	○	○	○	○	●	●	●	○	○	○	○	○	●	○	○	○	○	●	●	●	●
R9	○	●	●	●	○	○	○	●	○	●	○	○	○	○	○	●	○	●	●	○	●	○	●	○
R10	●	●	●	○	●	○	●	●	○	●	●	●	●	●	○	●	●	○	●	○	●	●	○	●
R11	●	●	●	○	●	○	○	●	○	●	○	○	○	●	○	●	●	○	●	○	●	○	○	●
R12	○	○	●	●	●	○	●	●	○	●	●	○	○	○	●	○	○	○	○	○	●	○	○	○
R13	●	●	●	○	○	○	○	●	○	●	●	○	○	○	○	●	●	●	●	●	●	◐	●	○
R14	●	●	●	○	○	○	○	○	○	○	○	○	○	○	○	●	○	○	●	○	●	●	○	○
R15	●	●	○	○	●	○	○	●	○	●	○	●	○	○	○	○	○	○	○	○	○	◐	○	○
O1	●	●	○	●	●	○	○	●	○	●	○	○	○	●	○	●	○	○	●	●	●	●	○	○
O2	●	●	●	●	●	●	●	●	○	●	●	○	●	●	○	●	●	●	●	●	●	●	○	●
O3	○	○	○	○	○	○	●	●	○	●	○	○	●	●	●	●	○	○	●	○	●	○	○	●
O4	●	○	●	○	○	○	○	●	○	●	●	○	○	●	○	●	○	○	●	●	○	○	○	○
O5	●	○	○	○	●	○	○	●	○	○	○	○	○	●	○	●	●	○	●	○	○	●	○	●
O6	●	●	○	●	○	○	●	●	○	●	●	○	●	○	○	●	●	○	●	○	○	○	○	○
O7	●	●	○	○	●	○	○	●	○	○	○	○	○	●	●	●	○	○	●	○	○	◐	○	○
B1	○	○	○	●	●	●	○	●	○	○	○	○	○	●	●	●	○	○	○	○	●	○	●	○
B2	●	●	●	●	○	○	○	○	○	○	○	○	○	●	●	●	●	○	○	○	○	○	●	●
B3	○	○	●	○	○	○	○	●	○	●	●	○	●	○	○	○	○	○	●	○	○	○	○	○
B4	●	○	●	○	○	●	○	●	○	●	●	○	○	●	●	●	○	○	●	○	○	○	○	○
B5	○	○	●	○	○	●	○	●	○	●	○	○	●	○	○	○	○	○	○	○	○	●	○	●
B6	●	○	●	●	○	●	○	●	○	●	○	○	○	●	○	●	○	○	●	○	○	○	○	●
B7	○	○	●	●	○	●	○	●	○	●	●	○	●	○	○	○	○	○	●	○	○	◐	○	●
B8	●	○	○	●	○	●	○	●	○	○	○	○	○	●	●	○	○	○	○	○	●	◐	○	○
B9	○	○	●	●	○	●	○	●	○	●	●	○	●	○	○	○	○	○	○	○	○	○	○	○
B10	●	●	●	●	●	●	○	●	○	●	●	●	●	●	●	●	○	○	●	○	●	●	○	●
B11	●	●	●	●	○	●	○	●	○	○	●	○	●	○	○	○	○	○	○	○	●	●	○	●

[1]For the Examples column,
● = detailed examples showing multi-variable data before and after de-identification
◐ = basic examples illustrating how de-id works

guides within each type. For this reason, along with the limited size of our dataset, we ultimately decided against comparing guide types as a pillar of our analysis.

Differential privacy. Interestingly, while 27 guides discuss adding noise, only 11 discuss differential privacy. Some guides indicate that this relates in part to differential privacy's relative newness—though differential privacy was proposed in 2006 [26], mainstream adoption is more recent, with the U.S. Census notably announcing their decision to use differential privacy in 2018 [1]. Referencing challenges encountered by the Census, R10 recommends against differential privacy (except for "statistically proficient" users) until usable tools are developed for researchers. In additional to technical difficulties, guides note a lack of research on differential privacy's impact on utility: G1 reads, "The theory and practice of differential privacy is still in their infancy, and at present, they are not sufficiently well-developed enough to produce privacy-protecting synthetic microdata that preserve interactions between more than a few independent variables." R13 explains that its organization may adopt differential privacy in the future, but only after assessing various considerations including usability, research funders' expectations, and privacy-utility trade-offs.

Other parts of the process. In addition to the information in Table 3, guides also comment on other parts of the de-identification process. 25 guides describe threat modeling, sometimes at a high level and sometimes detailing specific procedures. O1 defines three risk factors to assist in assessing risk for individual or groups of variables; O3 explains how to search for individuals that are unique with respect to indirect identifiers, recommending the Special Uniques Detection Algorithm [32]. G3 tells readers that "the first step for any agency is to undertake an exercise in which an inventory is compiled of all datasets available in the country," and G4 instructs readers to test de-identified data by attacking it.

In addition, 18 guides describe documenting the process of de-identification, often emphasizing the importance of documentation for oversight and to help end users interpret de-identified data. At the same time, 4 of these also warn against documenting too much, for fear of enabling reverse engineering: these guides suggest, for example, leaving out random seeds and avoiding specifying which values have been modified in a targeted fashion.

13 guides discuss soliciting review from other experts, and 12 describe continuing review even after de-identified data is shared or published, in order to account for developments in research on disclosure risk as well as new sources of external data for linking. We note that continuing review may have limited utility, especially for data made public online, which can be difficult to completely retract; R10 is the only guide to bring up this potential concern. Nonetheless, continuing review may still be useful if new risks are detected before data is downloaded by bad actors.

Finally, 26 guides describe adding access controls to de-identified data or asking end users to sign data use agreements before gaining access. Several guides presented these as a necessary trade-off, given the challenge of balancing risk and utility in de-identification. R10 takes a particularly strong stance, advocating for releasing minimally de-identified data with restricted access instead of making rigorously de-identified data public: "It is often better to abandon open access archiving of fully anonymized data and, instead, archive the pseudonymized (= non-anonymous!) person-related research data in restricted access. . . . This is still a form of open science, but disclosure risks are in this case managed through a restricted access policy and not (only) by means of anonymization methods." However, other guides recommend against restricting access if possible: R11 describes it as "not preferred."

4.2 Inconsistent definitions

Across de-identification guides, key terms are defined and used inconsistently. Here we highlight *de-identification*, *anonymization*, *inference*, *aggregation*, *perturbation*, and *differential privacy* as especially inconsistent.

De-identification and anonymization. Some guides distinguish de-identification from anonymization, generally attributing stronger outcomes to anonymization. Several say that de-identification lowers re-identification risk, while anonymization eliminates it: G2 describes the result of anonymization as "personal information that has been de-identified to the point that there is no serious possibility of re-identification, by any person or body using any additional data or technology at this point in time," while de-identification "carries a residual risk of re-identification." Several also define de-identification as only impacting direct identifiers, while anonymization also applies to other variables: R10 reads, "In our view, de-identification refers to the removal of direct identifiers such as name, address etc., while anonymization is a broader concept, which also deals with the disclosure potential coming from (combinations of) indirect identifiers such as sex, age, geographical region etc."

These distinctions conflict with other guides. Many emphasize that anonymization does not eliminate risk. G3 reads, "The aim of anonymizing microdata is to transform the datasets to achieve an 'acceptable level' of disclosure risk. The level of acceptability of disclosure risk and the need for anonymization are usually at the discretion of the data producer and guided by legislation." And O4 argues that due to the future potential for new data processing techniques and new linking datasets, "It is not possible to say with certainty that an individual will never be identified from a dataset which has been subjected to an anonymisation process." Likewise, many guides do not limit de-identification to direct identifiers: O6 addresses this explicitly, saying, "Sometimes de-identification is used to refer to the removal of 'direct identifiers,' such as name and address. In this guide, de-identification is used in a broader sense consistently with the meaning defined in the Privacy Act."

This splintering of definitions can be partially traced to our evolving understanding of disclosure risk. In light of attacks on de-identified data, some guides have loosened former definitions of anonymization that involve elimination of risk: acknowledging "the difficulty in achieving perfect anonymisation at the present time," O5 says, "We will use the term 'anonymisation' whether or not the identification of the data subject is reversible to a greater or lesser degree." Other guides have instead added qualifiers to key terms: O3 refers to *functional anonymization*, which "does not assume that anonymisation can be zero-risk or irreversible; it is meant instead to bring anonymisation practice in line with the art of the possible,"

and R9 encourages *partial anonymization* to reduce risk, as "full anonymization is often difficult to attain and for research, often not desirable." Yet other guides have dropped the term anonymization altogether: G1 avoids using the term due to "inconsistencies in the use and definitions."

Inference. The term *inference* is also used inconsistently. Some guides provide definitions based on statistical properties of de-identified data, which fall within the bounds of our definition of probabilistic inference (Section 2.1): R9 defines inference as "the possibility of deducing, with significant probability of correctness, the value of an attribute from the values of a set of other attributes." Some of these, such as R13, focus specifically on such disclosures about individuals not in the dataset.

However, other guides use inference to describe entirely different disclosure types and mechanisms. B4 appears to simply describe singling out: "One example of an inference attack is to cross-reference location, and browsing histories, to infer their identity." O4 provides an unclear definition that seems to result in identity disclosure: "In some cases, it may be possible to infer a link between two pieces of information in a set of data, even though the information is not expressly linked. This may occur, for example, if a dataset contains statistics regarding the seniority and pay of the employees of a company. Although such data would not point directly to the salaries of individuals in the dataset, an inference might be drawn between the two pieces of information, allowing some individuals to be identified."

Aggregation, perturbation, and differential privacy. Finally, there is substantial disagreement in describing certain techniques. Aggregation is an especially common offender, with some guides referring to the process of replacing microdata with aggregate data (i.e., summary statistics such as counts, averages, and ranges), but many—including G2, R7, R8, and B11—instead using it synonymously with generalization.

We also observed many different definitions of perturbation. Many are in line with definitions in academic work [19, 54], including O1, which describes adding noise in a limited way intended to preserve statistical properties of the data as a whole. However, some, such as R12, also include fully random alteration of data within the scope of perturbation, while others restrict perturbation to an extremely limited scope with negligible and even predictable distortion: G2 defines it as "replacing specific values with other values that are consistent for each individual. For example, adding or subtracting two years from each individual's actual age." Other guides are even broader, essentially equating perturbation to de-identification: R1 defines it as "methods used to alter data in order to mitigate risks to data provider[s] (i.e. removal of PII/sensitive data; top/bottom coding of outliers)."

Definitions of differential privacy sometimes fail to distinguish it from adding noise more broadly. R6 only mentions differential privacy as an unexplained parenthetical, suggesting synonymity: "perturbation or adding noise (differential privacy)." Like several other guides, B8 leaves out the privacy guarantee in its description: "Differential privacy introduces statistical noise (or jitter) to a numeric value, such that the actual value is not known, but the new value does not adversely distort the aggregate analysis."

Guides also sometimes contain misconceptions about differential privacy. For example—lumping it together with k-anonymity—B1 incorrectly states that differential privacy's core guarantee can be broken by new linking data: "While PETs like k-anonymization and differential privacy can offer mathematical guarantees for individual datasets, it's important to note these guarantees are based on assumptions about the availability of other data that can change over time. The availability of new data, for example, can create new indirect identifiers." This is in contradiction with differential privacy's property of immunity to post-processing [29].

4.3 Gaps in threat coverage

Overselling outcomes. As we touched on in Section 4.2, some guides describe de-identification as making re-identification impossible. This can be misleading, as eliminating all risk entails commensurate loss of utility that is rarely desirable; even for differential privacy, which does provide a guarantee against definite re-identification, higher ϵ values permit re-identification of individuals with high probability. While 22 guides describe de-identification as lowering, rather than eliminating, this risk, 10 contain language describing the total elimination of re-identification.

Interestingly, 9 of those 10 contain both kinds of statements, which at first may appear to be self-contradictory. However, in some cases, this is due to guides attributing different outcomes to *anonymization* and *de-identification*. In other cases, it is due to guides starting with simplistic definitions and adding nuance later. For example, R9 starts by defining anonymization as the "complete and irreversible removal of any information from a dataset that could lead to an individual being identified," before clarifying two pages later that full anonymization is rarely desirable and that partial anonymization can lower risk while preserving utility.

Non-identifying variables. All variables in a dataset pose some amount of disclosure risk [75, 80]; however, some guides draw distinctions between identifying and non-identifying variables. In some cases, guides make extremely broad claims about the degree of non-identifiability. For example, O2 says that target variables—giving examples of transactions, salary, credit rating, insurance policy, medical diagnosis, and vaccination status—"cannot be used for re-identification as they are typically proprietary." In reality, these types of information are often for sale, available online, or known to specific adversaries such as employees of a healthcare organization.

Patchy coverage of reverse engineering. While many guides present other disclosure types and mechanisms—particularly singling out and linking—as key concepts for readers to understand, reverse engineering is rarely presented as a class of disclosure mechanisms, instead popping up in ad hoc examples. Some guides mention attacks that use knowledge or assumptions about how specific de-identification techniques were applied to undo them: e.g., unmasking pseudonyms that were assigned non-randomly in alphabetical order, downcoding k-anonymous data [21], and brute-forcing improperly hashed values. Other guides mention reconstructing missing data by reasoning about available data: e.g.,

Figure 1: Table 9 from G3, an example of local suppression that is vulnerable to reverse engineering.

Table 9 Local suppression illustration - sample data before and after suppression

Variable	Before local suppression			After local suppression		
ID	Gender	Region	Education	Gender	Region	Education
1	female	rural	higher	female	rural	NA/missing [5]
2	male	rural	higher	male	rural	higher
3	male	rural	higher	male	rural	higher
4	male	rural	higher	male	rural	higher
5	female	rural	lower	female	rural	lower
6	female	rural	lower	female	rural	lower
7	female	rural	lower	female	rural	lower

deducing an individual's region of residence based on whether they live in an urban or rural area, or guessing the identity of a pseudonymized city from the demographics of individuals in the dataset who live in that city.

Despite this great variety of reverse engineering attacks, many guides fail to address them, even when presenting techniques or examples that are vulnerable. For example, even though a long salt is needed to keep hashed values from being brute forced in many cases, only 7 of 14 guides that mention hashing also mention a salt or key to prevent brute forcing, and only 3 of those mention a length requirement. Guides sometimes recommend adding noise in a way that is minimally random, increasing the likelihood that it could be reversed: both G2 and R13 suggest adjusting values such as ages and dates by an offset that is consistent across all records. In some egregious cases, guides recommend techniques that are blatantly reversible: for example, B10 suggests scrambling the letters in a name.

Even guides that are generally conscious of reverse engineering attacks may still let vulnerable examples fall through the cracks. G3 discusses reverse engineering extensively, covering examples of both reversing de-identification techniques and reconstructing missing data; however, it also includes an example of local suppression, shown in Figure 1, that has a clear, undiscussed vulnerability. In the example, the deleted value can either be *higher* or *lower* education. Based on G3's own explanation, if it were *lower*, then it would not have needed to be removed because the record would not have been unique in the datasest; therefore, an attacker could correctly deduce that the original value must have been *higher*.

Similarly, though Page 6 of R8 stresses the importance of assigning pseudonyms in an unpredictable fashion ("at random and not linked to a sort order (e.g., by alphabet) or any pre-existing ID variable from another database") to prevent reverse engineering, it misses another vulnerability in the very same example. The proposed scheme—in which pseudonyms for villages and other administrative divisions are encoded hierarchically—could enable attackers to, for example, identify pseudonymized provinces by counting the number of pseudonymized districts within them.

4.4 Limited explanation of *how* to de-identify

We analyzed two criteria broadly related to usability: how guides support readers in choosing a de-identification strategy, and how they teach readers to actually carry it out. We leave a fuller exploration of usability, including testing with users, to future work.

How to choose between de-identification approaches. With a wide array of de-identification techniques, each with its own trade-offs, it is potentially daunting for practitioners to choose an approach that suits their own capabilities and needs, as well as the best interests of their data subjects. One way in which guides attempt to help is by discussing trade-offs of specific techniques.

Most commonly, 31 guides discuss specific techniques' impact on risk, and 31 discuss utility. Often, these go hand in hand: O7 notes, "in relation to differential privacy, the more that the data is altered (that is, the more noise added), the more that privacy is preserved, yet this comes at a high cost to data utility," and O5 similarly says, "Higher values of K correspond to more stringent privacy requirements . . . In the obtainment of higher values of K, we may lose fidelity in the source data." While true, these kinds of general statements are questionably helpful: arguably, every single de-identification choice has a risk-utility trade-off.

Some guides provide more actionable recommendations, though these often take the place of de-identification standards such as k-anonymity and differential privacy rather than complementing them. R2 explains that their repository will not publish datasets containing both student-level and school-level information in public access, requiring datasets either to be restricted-access or to remove one of the information types. To avoid losing data utility, they recommend the former. R13 also provides several actionable recommendations, including generalizing values that appear fewer than five times (with flexibility depending on context), as well as echoing HIPAA Safe Harbor rules: deleting or top-coding ages 90 and up, and pseudonymizing geographic regions containing fewer than 20,000 people.

17 guides comment on the usability or accessibility of specific de-identification techniques. This includes techniques that require expertise to carry out: multiple guides state that differential privacy is difficult to implement. It also includes techniques that are computationally expensive: G3 notes that using microSDC to conduct local suppression or shuffling can require prohibitive amounts of computational power, suggesting that these methods be combined with other de-identification techniques that can reduce complexity in the dataset first. Occasionally, guides present other rationales for choosing specific techniques. Several consider the potential for the appearance of de-identified data to be misleading in various ways, for example contrasting adding noise with other techniques such as deletion and generalization: R8 warns that adding noise to location data could create "the illusion of precision," and G3 warns that adding noise generally could cause participants to believe their data was not de-identified, reducing their willingness to participate in future surveys.

G1 lays out two kinds of standards—risk-based and prescriptive—created by government agencies and regulations, designed to guide practitioners through the process of de-identification. We see these

distinctions among our guides as well. Risk-based guidance emphasizes the importance of thinking critically, taking context into account, and avoiding one-size-fits-all solutions; it also sometimes emphasizes that de-identification is a subjective process. Risk-based guidance often lists de-identification techniques and their trade-offs to help readers understand and judge for themselves: G4 and R12 both contain tables listing techniques and descriptions alongside details about their impact on risk and utility (G4) or their pros and cons (R12). Prescriptive guidance, on the other hand, often provides suggested thresholds or rules, and it may refer back to prescriptive standards such as HIPAA Safe Harbor. Prescriptive guidance sometimes organizes content around data types to provide unambiguous direction on how to de-identify data in different situations: O2 provides type-by-type instructions for de-identifying gender, date of birth, geographical location, and more (as do others, including R8 and R4, though in much less detail), while R13 provides different instructions for different situations, including how to treat values with low counts, continuous variables, outliers, and identifiers within text fields.

How to implement de-identification. Just half (N = 19) of the guides we analyzed provide examples of data that help illustrate how de-identification works; these largely consist of tables of example data before and/or after de-identification. Of these, 13 guides provide detailed examples, which (1) illustrate data both before and after de-identification and (2) demonstrate de-identification across multiple variables in a meaningful way (e.g., different techniques are applied to different variables, or several variables are considered in the process of achieving k-anonymity). R5 provides a table with three individuals' data, showing how names are pseudonymized, ages and addresses are generalized, dates are distorted with noise, and IP addresses are deleted; interestingly, this example is followed by reflection questions encouraging readers to think about the choices made in the example and consider alternatives.

Not all examples are good. In addition to examples that fail to protect against reverse engineering (Section 4.3), we also observed some examples that likely undermine utility past the point of usefulness. Examples in B5 replace all salaries and addresses with entirely random synthetic values; examples in O2 and B7 randomly swap all values for all individuals. Unless specific reasons are given for these niche approaches, most readers would probably benefit from either simpler examples such as deleting data, or more nuanced ones that reduce risk while doing more to preserve utility.

Beyond examples, 3 guides describe de-identification case studies, with G3 including several detailed case studies from the authors' own experience that contain descriptions of both techniques and metrics of risk and utility. Moving into attacks, 8 guides provide case studies of disclosure. Most case studies arefrom research and journalism focused on the topic of de-identification, with 5 mentioning Sweeney's study estimating that 87% of the U.S. population is identifiable by their zip code, birth date, and gender [96]. Just one case study discussed data being exploited for other purposes: G1 mentioned the case of a news website *The Pillar* outing a Catholic priest using poorly de-identified data from the app Grindr [13].

18 guides suggest tools to help with de-identification. Common examples include ARX [8], sdcMicro [98], and μ-ARGUS [44]. Some guides strongly recommend using specialized tools for de-identification, including G1, which recommends that government agencies do so because de-identifying data in general-purpose software such as spreadsheets "typically lack the key functions required for sophisticated de-identification" and "may encourage the use of simplistic de-identification methods, such as deleting columns that contain sensitive data categories and manually searching and removing individual data cells that appear sensitive." On the other hand, some guides express reservations that tools are not designed for and tested with a diversity of end users: R13 notes that several tools including ARX have primarily been used on large datasets in specific domains such as health, rather than on the smaller-scale kinds of data its organization regularly deals with.

5 Discussion

In light of our findings, we now discuss suggestions for improving de-identification guides, as well as directions for future work.

Practitioners face a maze of contradictory and misleading definitions. The guides we analyzed contained conflicting definitions for key terms such as *anonymization*, *inference*, *aggregation*, and *differential privacy*. This increases risk of misunderstandings and misaligned expectations, making productive discussion—between researchers and data repository staff, privacy engineers and stakeholders, social scientists and cryptographers—more difficult.

As others have before us [36], we find existing attempts to distinguish *anonymization* from *de-identification* problematic, overselling the degree of protection against unwanted disclosure or overly limiting the scope of de-identification techniques. Ideally, a single term would be used by all guides to indicate a spectrum of approaches and outcomes. Unfortunately, this is unrealistic, as different terms are already deeply encoded in regulation and practice: for example, *anonymous* data and *pseudonymisation* are GDPR terms, and *deidentified* data is a CCPA term. Others have already tried and failed to coalesce various communities around a single term [80]. To avoid confusion, our recommendation to guide writers, policymakers, practitioners, and others is to always be explicit about (1) whether given methods are limited to direct identifiers only, and (2) what kind of risk reduction is the goal (being cautious of promising unreasonable privacy outcomes, such as the total elimination of risk). This should resolve the most notable conflicts in terminology we observed in our dataset. Explicitly acknowledging the existence of conflicting terminology may also reduce confusion.

It is similarly tricky to provide an agreeable definition for *inference*. This is partly because it has also been defined variously by previous work. Some works have scoped inferential disclosure incredibly broadly, encompassing any increase in an attacker's ability to learn or guess information about any individual, whether or not they are in the dataset [24, 25]. Finding this too broad to be useful, Kifer et al. created a new definition that distinguishes between valid scientific inferences and privacy-breaching inferences (including identity and attribute disclosure), by considering whether a disclosure is caused by an individual's inclusion in a dataset [55]. Given this smorgasbord of definitions, guides should use precise language when describing inference. In many cases, guides may consider

avoiding the term altogether and instead explaining in other words how disclosure can occur even in the absence of certainty.

Our recommendations for definitions of techniques are more prescriptive. *Aggregation*, confusingly, has two entirely distinct definitions, and we urge guide writers to define it as converting microdata into aggregate data, rather than grouping values into coarser categories (i.e., *generalization*). *Perturbation* is not always presented in a way that helps readers navigate the balance between privacy and utility, with guides describing heavy-handed approaches such as totally random alteration of data, as well as quite limited approaches such as shifting values the same amount for all individuals. We argue perturbation should be scoped similarly to its usage in published work—that is, adding noise to data in a way that aims to preserve overall statistical properties. Last but not least, guides should avoid common misconceptions about *differential privacy*, most of which involve failing to distinguish it from adding noise broadly. Explaining differential privacy is not easy and is an active area of research in the context of the general population [23, 53, 58, 73, 92, 112, 115]; while these can be a starting point, future work investigating explanations specifically for practitioners would provide better insight. While prior works have studied *how* to help practitioners implement differential privacy [27, 71], more work is needed in this crucial earlier step of communicating *when* and *why* to use differential privacy in the first place. Nanayakkara and Hullman take initial steps by analyzing discourse around differential privacy for the U.S. Census, to understand the causes of controversy [72].

Guides could provide more recommendations *against* certain techniques. Recommendations for how *not* to de-identify data are nearly nonexistent in the guides we analyzed. As a result, what coverage exists of several de-identification techniques that we believe should be retired is largely positive or neutral. We suggest that guides consider actively discouraging use of the following techniques except in niche situations:

- Partial obfuscation risks reverse engineering and generally lacks a meaningful utility benefit. If it is recommended, a specific reason should be given for why the retained data is more useful than simply deleting it altogether.
- Hashing and encrypting also risk reverse engineering and lack a meaningful utility benefit. Creating properly random pseudonyms (and retaining a key until it is no longer needed) offers the same level of protection.

Despite problems with other traditional de-identification techniques, such as generalization and deleting individual values, they still have a necessary place in de-identification guides until usable differential privacy tools exist for a range of practitioners whose workflows include Excel, R, and Stata; and until differential privacy techniques are rigorously demonstrated to meet the needs of stakeholders such as data users and research funding agencies.

Guides could discuss attacks more systematically. Previous work has suggested that understanding how attacks occur is important for software developers and security professionals to defend against them [91, 110, 111]. In the context of de-identification, at a minimum, we believe singling out and linking should be covered by all guides, as they are components of disclosure that are well researched and carry significant potential for harm. While most guides did mention singling out and linking, not all did (especially blog posts), and some mentioned them offhand rather than highlighting them as key concepts.

We believe reverse engineering should be treated as its own category of disclosure mechanisms—though guides should also continue to highlight vulnerabilities for specific techniques, such as some of the examples in Appendix C in the supplementary materials. Guides should also discuss attribute disclosure and/or probabilistic inference, to avoid giving the impression that preventing singling out will eliminate all risk. This is a notable gap within our corpus—26 guides do not mention either type of disclosure.

Future work could inform the proper balance between de-identification techniques and access control. Many guides advocate for using access control and data-use agreements to guard the release of de-identified data, in order to manage the difficulty of balancing utility and risk with de-identification techniques alone. This is a sensible approach that speaks to the success of researchers who have studied and spoken out about issues with traditional de-identification techniques.

However, access control comes with its own challenges. In interviews with privacy practitioners at companies, Garrido et al. found that "analysts suffer from lengthy bureaucratic processes for requesting access to sensitive data, yet once granted, only scarcely-enforced privacy policies stand between rogue practitioners and misuse of private information" [39]. They conclude that differential privacy might help reduce barriers to accessing data while also decreasing risk. Similarly, Yoon and Copeland interviewed staff at community-based organizations who struggle to access government data on important topics such as domestic violence and sexual assault, because the practice of granting access through individual consultations means staff must contend with long delays, inconsistency between different government agencies, and lawyers' "nervousness" about confidentiality [116]. Further, Tyler analyzed documents and interviewed staff at data repositories, identifying inconsistencies in how different repositories evaluate researchers' requests for data access [102].

Balancing the trade-offs between de-identification techniques and access control is a thorny issue. Future user-centered research and dialogue could provide paths forward by further characterizing (1) the robustness of protecting data through access control, (2) the practical challenges faced by diverse practitioners tasked with applying de-identification techniques, and (3) the barriers introduced by both approaches for data users. Ultimately, we hope these investigations would make it back into de-identification guides in the form of a framework for deciding what combination of de-identification techniques and access control should be employed to protect a given dataset, and whether it is worth the effort to create multiple data releases with different combinations of these measures. This echoes a call by researchers at the Inter-university Consortium for Political and Social Research to seek a "better understanding of the workflows of researchers accessing confidential data, how they work, and with whom they work or collaborate," in order to inform future design of data access mechanisms [83].

Improving usability is key. De-identification is a complicated process, and guides for practitioners must be designed to be clear, readable, and actionable. More examples would help—especially detailed examples showing the before and after of de-identification in the context of multiple variables, which are present in just 13 of the guides we analyzed. Case studies of re-identification would both reinforce key concepts and highlight the importance of de-identification; real-world case studies outside of the academic and journalistic settings, such as the outing of a Catholic priest using poorly de-identified data from Grindr [13], would provide even greater motivation.

Guides should also ideally be evaluated with real users. This includes testing specific guides before release to determine whether they are clear and facilitate effective learning. This also includes future work aimed at answering certain high-level questions about how best to structure guides: e.g., are risk-based or prescriptive standards, discussed in Section 4.4, more effective at teaching various aspects of de-identification? Do practitioners find big tables of techniques useful, and do they prefer for them to be organized by technique or by data type? Answers to these questions likely depend on the target audience, meaning that there is fruitful research to be done on how to tailor advice to a variety of practitioners.

Acknowledgments

This project was supported in part by NSF grant OAC-2232863. We are grateful to the researchers who participated in a separate study and recommended de-identification guides for us to analyze, particularly the participant who helped us confidentially access their organization's internal guide.

References

[1] John M. Abowd. 2018. Protecting the Confidentiality of America's Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau. https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html.

[2] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. 2016. You Get Where You're Looking for: The Impact of Information Sources on Code Security. In *IEEE S&P '16*. https://doi.org/10.1109/SP.2016.25.

[3] Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. 2016. You Are Not Your Developer, Either: A Research Agenda for Usable Security and Privacy Research beyond End Users. In *IEEE Cybersecurity Development Conference*. https://doi.org/10.1109/SecDev.2016.013.

[4] Yasemin Acar, Christian Stransky, Dominik Wermke, Charles Weir, Michelle L. Mazurek, and Sascha Fahl. 2017. Developers Need Support, Too: A Survey of Security Advice for Software Developers. In *IEEE Cybersecurity Development*. https://doi.org/10.1109/SecDev.2017.17.

[5] Hal Ali. 2023. Privacy-Preserving Usage Data: Under the Hood. https://blog.1password.com/privacy-telemetry-deep-dive/.

[6] Olivia Angiuli, Joe Blitzstein, and Jim Waldo. 2015. How to De-Identify Your Data: Balancing Statistical Accuracy and Subject Privacy in Large Social-Science Data Sets. *Queue* 13, 8 (2015). https://doi.org/10.1145/2838344.2838930.

[7] Apple. [n. d.]. Apple Differential Privacy Technical Overview. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf.

[8] ARX. 2024. ARX: Data Anonymization Tool. https://arx.deidentifier.org/.

[9] Rebecca Balebako, Abigail Marsh, Jialiu Lin, Jason Hong, and Lorrie Faith Cranor. 2014. The Privacy and Security Behaviors of Smartphone App Developers. In *USEC '14*. Internet Society. https://doi.org/10.14722/usec.2014.23006.

[10] Jane Bambauer, Krishnamurty Muralidhar, and Rathindra Sarathy. 2014. Fool's Gold: An Illustrated Critique of Differential Privacy. *Vanderbilt Journal of Entertainment and Technology Law* 16, 4 (2014). https://scholarship.law.vanderbilt.edu/jetlaw/vol16/iss4/1/.

[11] Thijs Benschop and Matthew Welch. 2021. Statistical Disclosure Control for Microdata: A Practice Guide for sdcMicro. https://sdcpractice.readthedocs.io/en/latest/.

[12] Alberto Blanco-Justicia, David Sánchez, Josep Domingo-Ferrer, and Krishnamurty Muralidhar. 2022. A Critical Review on the Use (and Misuse) of Differential Privacy in Machine Learning. *Comput. Surveys* 55, 8 (Dec. 2022). https://doi.org/10.1145/3547139.

[13] Michelle Boorstein, Marisa Iati, and Annys Shin. 2021. Top U.S. Catholic Church Official Resigns after Cellphone Data Used to Track Him on Grindr and to Gay Bars. *The Washington Post* (July 2021). https://www.washingtonpost.com/religion/2021/07/20/bishop-misconduct-resign-burrill/.

[14] Danah Boyd and Jayshree Sarathy. 2022. Differential Perspectives: Epistemic Disconnects Surrounding the U.S. Census Bureau's Use of Differential Privacy. *Harvard Data Science Review* Special Issue 2 (June 2022). https://hdsr.mitpress.mit.edu/pub/3vj5j6i0/release/3.

[15] Quentin Brummet, Patrick Coyle, and Brandon Sepulvado. 2021. Effects of Differential Privacy Techniques: Considerations for End Users. *Research in Social and Administrative Pharmacy* 17, 5 (May 2021). https://doi.org/10.1016/j.sapharm.2020.07.029.

[16] Loredana Caruccio, Domenico Desiato, Giuseppe Polese, Genoveffa Tortora, and Nicola Zannone. 2022. A Decision-Support Framework for Data Anonymization with Application to Machine Learning Processes. *Information Sciences* 613 (Oct. 2022). https://doi.org/10.1016/j.ins.2022.09.004.

[17] Tânia Carvalho, Nuno Moniz, Pedro Faria, and Luís Antunes. 2023. Survey on Privacy-Preserving Techniques for Microdata Publication. *Comput. Surveys* 55, 14s (July 2023). https://doi.org/10.1145/3588765.

[18] CDC. 2022. 2022 Behavioral Risk Factor Surveillance System. https://www.cdc.gov/brfss/annual_data/annual_2022.html.

[19] M. A. P. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil. 2018. Efficient Data Perturbation for Privacy Preserving and Accurate Data Stream Mining. *Pervasive and Mobile Computing* 48 (2018). https://doi.org/10.1016/j.pmcj.2018.05.003.

[20] Miranda Christ, Sarah Radway, and Steven M. Bellovin. 2022. Differential Privacy and Swapping: Examining de-Identification's Impact on Minority Representation and Privacy Preservation in the U.S. Census. In *IEEE S&P '22*. https://doi.org/10.1109/SP46214.2022.9833668.

[21] Aloni Cohen. 2022. Attacks on Deidentification's Defenses. In *USENIX Security '22*. https://www.usenix.org/conference/usenixsecurity22/presentation/cohen.

[22] Corporate Finance Institute. [n. d.]. Data Anonymization. https://corporatefinanceinstitute.com/resources/business-intelligence/data-anonymization/.

[23] Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. 2021. "I Need a Better Description": An Investigation into User Expectations for Differential Privacy. In *ACM CCS '21*. https://doi.org/10.1145/3460120.3485252.

[24] Tore Dalenius. 1977. Towards a Methodology for Statistical Disclosure Control. *Statistik Tidskrift* 15 (1977). https://hdl.handle.net/1813/111303.

[25] George T. Duncan, Thomas B. Jabine, Virginia A. de Wolf, Panel on Confidentality and Data Access, Committee on National Statistics, Commission on Behavioral and Social Sciences and Education, National Research Council, and Social Science Research Council. 1993. *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. National Academies Press.

[26] Cynthia Dwork. 2006. Differential Privacy. In *International Colloquium on Automata, Languages, and Programming (Lecture Notes in Computer Science)*. https://doi.org/10.1007/11787006_1.

[27] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. 2019. Differential Privacy in Practice: Expose Your Epsilons! *Journal of Privacy and Confidentiality* 9, 2 (Oct. 2019). https://doi.org/10.29012/jpc.689.

[28] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography Conference (Lecture Notes in Computer Science)*. https://doi.org/10.1007/11681878_14.

[29] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014). https://doi.org/10.1561/0400000042.

[30] eHealth Queensland. 2021. De-Identification and Anonymisation of Data Guideline.

[31] Mark Elliot, Elaine Mackey, and Kieron O'Hara. 2020. *The Anonymisation Decision-Making Framework: European Practitioners' Guide* (2nd ed.). UKAN. https://eprints.soton.ac.uk/445373/.

[32] M. J. Elliot, A. M. Manning, and R. W. Ford. 2002. A Computational Algorithm for Handling the Special Uniques Problem. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002). https://doi.org/10.1142/S0218488502001600.

[33] Khaled El Emam. 2013. *Guide to the De-Identification of Personal Health Information*. Auerbach Publications. https://doi.org/10.1201/b14764.

[34] Joseph Ficek, Wei Wang, Henian Chen, Getachew Dagne, and Ellen Daley. 2021. Differential Privacy in Health Research: A Scoping Review. *Journal of the American Medical Informatics Association* 28, 10 (Oct. 2021). https://doi.org/10.1093/jamia/ocab135.

[35] Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. 2014. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In *USENIX Security*

'14. https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/fredrikson_matthew.
[36] Simson Garfinkel, Barbara Guttman, Joseph Near, Aref Dajani, and Phyllis Singer. 2023. *De-Identifying Government Datasets: Techniques and Governance*. NIST Special Publication 800-188. https://csrc.nist.gov/pubs/sp/800/188/final.
[37] Simson L. Garfinkel, John M. Abowd, and Sarah Powazek. 2018. Issues Encountered Deploying Differential Privacy. In *WPES '18*. https://doi.org/10.1145/3267323.3268949.
[38] Simson L. Garfinkel and Philip Leclerc. 2020. Randomness Concerns When Deploying Differential Privacy. In *WPES '20*. https://doi.org/10.1145/3411497.3420211.
[39] Gonzalo Munilla Garrido, Xiaoyuan Liu, Floria Matthes, and Dawn Song. 2023. Lessons Learned: Surveying the Practicality of Differential Privacy in the Industry. *PoPETs* 2023, 2 (2023). https://doi.org/10.56553/popets-2023-0045.
[40] Google Books Ngram Viewer. 2024. De-Identify, Deidentify, de-Identification, Deidentification. https://books.google.com/ngrams/graph?content=de-identify%2Cdeidentify%2Cde-identification%2Cdeidentification&year_start=1800&case_insensitive=on.
[41] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods* 18, 1 (Feb. 2006). https://doi.org/10.1177/1525822X05279903.
[42] Samuel Haney, Ashwin Machanavajjhala, John M. Abowd, Matthew Graham, Mark Kutzbach, and Lars Vilhuber. 2017. Utility Cost of Formal Privacy for Releasing National Employer-Employee Statistics. In *ACM International Conference on Management of Data*. https://doi.org/10.1145/3035918.3035940.
[43] John P. Holdren. 2013. Increasing Access to the Results of Federally Funded Scientific Research. https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.
[44] Anco Hundepool, Ramya Ramaswamy, Peter-Paul de Wolf, Luisa Franconi, Ruth Brand, and Josep Domingo. 2021. μ-ARGUS. https://research.cbs.nl/casc/mu.htm.
[45] ICF. 2022. Protocol for Preparing Non-Public, Restricted, and Public Access Datasets: Zone of Influence Surveys. https://github.com/FTF-Survey-Methods-Toolkit/Feed-The-Future-Survey-Methods-Toolkit-Baseline/blob/main/4.02%20Protocol%20for%20Preparing%20Datasets/Protocol%20for%20Preparing%20Non-Public%2C%20Restricted%2C%20and%20Public%20Access%20Datasets_20220405.docx.
[46] Imperva. [n. d.]. Anonymization. https://www.imperva.com/learn/data-security/anonymization/.
[47] Inter-university Consortium for Political and Social Research. [n. d.]. Guide to Social Science Data Preparation and Archiving. https://www.icpsr.umich.edu/files/deposit/dataprep.pdf.
[48] Irish Data Protection Commission. 2019. Guidance on Anonymisation and Pseudonymisation. https://www.dataprotection.ie/sites/default/files/uploads/2022-04/Anonymisation%20and%20Pseudonymisation%20-%20latest%20April%202022.pdf.
[49] Honglu Jiang, Jian Pei, Dongxiao Yu, Jiguo Yu, Bei Gong, and Xiuzhen Cheng. 2023. Applications of Differential Privacy in Social Network Analysis: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 35, 1 (Jan. 2023). https://doi.org/10.1109/TKDE.2021.3073062.
[50] Samantha Joel, Paul W. Eastwick, and Eli J. Finkel. 2018. Open Sharing of Data on Close Relationships and Other Sensitive Social Psychological Topics: Challenges, Tools, and Future Directions. *Advances in Methods and Practices in Psychological Science* 1, 1 (2018). https://doi.org/10.1177/2515245917744281.
[51] Johns Hopkins University Data Services. 2023. Protecting Human Subject Identifiers. https://guides.library.jhu.edu/protecting_identifiers/overview.
[52] k2view. 2023. What Is Data Anonymization? Techniques, Pros, Cons, and Use Cases. https://www.k2view.com/hubfs/K2view_Data%20Anonymization.pdf.
[53] Farzaneh Karegar, Ala Sarah Alaqra, and Simone Fischer-Hübner. 2022. Exploring User-Suitable Metaphors for Differentially Private Data Analyses. In *SOUPS '22*. https://www.usenix.org/conference/soups2022/presentation/karegar.
[54] Hillol Kargupta, Souptik Datta, Qi Wang, and Krishnamoorthy Sivakumar. 2003. On the Privacy Preserving Properties of Random Data Perturbation Techniques. In *IEEE International Conference on Data Mining*. https://doi.org/10.1109/ICDM.2003.1250908.
[55] Daniel Kifer, John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Philip Leclerc, Ashwin Machanavajjhala, William Sexton, and Pavel Zhuravlev. 2022. Bayesian and Frequentist Semantics for Common Variations of Differential Privacy: Applications to the 2020 Census. http://arxiv.org/abs/2209.03310. arXiv:2209.03310
[56] Nigel King and Joanna M. Brooks. 2017. *Template Analysis for Business and Management Students*. SAGE.
[57] Sarah Kopper, Anja Sautmann, and James Turitto. 2020. J-PAL Guide to de-Identifying Data. https://www.povertyactionlab.org/sites/default/files/research-resources/J-PAL-guide-to-deidentifying-data.pdf.
[58] Patrick Kühtreiber, Viktoriya Pak, and Delphine Reinhardt. 2022. Replication: The Effect of Differential Privacy Communication on German Users' Comprehension and Data Sharing Attitudes. In *SOUPS '22*. https://www.usenix.org/conference/soups2022/presentation/kuhtreiber.
[59] La Trobe University. [n. d.]. Sensitive Data. https://latrobe.libguides.com/sensitivedata/introduction.
[60] Hyukki Lee and Yon Dohn Chung. 2020. Differentially Private Release of Medical Microdata: An Efficient and Practical Approach for Preserving Informative Attribute Values. *BMC Medical Informatics and Decision Making* 20 (2020). https://doi.org/10.1186/s12911-020-01171-5.
[61] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t-Closeness: Privacy beyond k-Anonymity and ℓ-Diversity. In *IEEE International Conference on Data Engineering*. https://doi.org/10.1109/ICDE.2007.367856.
[62] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. ℓ-Diversity: Privacy beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data* 1, 1 (March 2007). https://doi.org/10.1145/1217299.1217302.
[63] Abdul Majeed and Seong Oun Hwang. 2022. A Practical Anonymization Approach for Imbalanced Datasets. *IT Professional* 24, 1 (2022). https://doi.org/10.1109/MITP.2021.3132330.
[64] Ana P. Martínez-Donate and Gudelia Rangel Gómez. 2023. Project Migrante: Health Status and Access to Health Care among Migrants on Mexico's Northern Border, 2020-2021. https://doi.org/10.3886/ICPSR38601.V2.
[65] Alishah Mawji, Holly Longstaff, Jessica Trawin, Dustin Dunsmuir, Clare Komugisha, Stefanie K. Novakowski, Matthew O. Wiens, Samuel Akech, Abner Tagoola, Niranjan Kissoon, and J. Mark Ansermino. 2022. A Proposed De-Identification Framework for a Cohort of Children Presenting at a Health Facility in Uganda. *PLOS Digital Health* 1, 8 (Aug. 2022). https://doi.org/10.1371/journal.pdig.0000027.
[66] MC. 2020. Anonymizing Data Sets. https://towardsdatascience.com/anonymizing-data-sets-c4602e581a35.
[67] Millennium Challenge Corporation. 2020. MCC Guidelines for Transparent, Reproducible, and Ethical Data and Documentation (TREDD). https://www.mcc.gov/resources/doc/guidance-mcc-guidelines-tredd/.
[68] MOSTLY AI. 2023. Data Anonymization in Python. https://mostly.ai/blog/data-anonymization-in-python.
[69] Soumya Mukherjee, Aratrika Mustafi, Aleksandra Slavković, and Lars Vilhuber. 2023. Assessing Utility of Differential Privacy for RCTs. http://arxiv.org/abs/2309.14581. arXiv:2309.14581 [cs, econ, stat]
[70] Krishnamurty Muralidhar and Josep Domingo-Ferrer. 2023. A Rejoinder to Garfinkel (2023) – Legacy Statistical Disclosure Limitation Techniques for Protecting 2020 Decennial US Census: Still a Viable Option. *Journal of Official Statistics* 39, 3 (Sept. 2023). https://doi.org/10.2478/jos-2023-0019.
[71] Priyanka Nanayakkara, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers. 2022. Visualizing Privacy-Utility Trade-Offs in Differentially Private Data Releases. *PoPETs* 2022, 2 (April 2022). https://doi.org/10.2478/popets-2022-0058.
[72] Priyanka Nanayakkara and Jessica Hullman. 2023. What's Driving Conflicts around Differential Privacy for the U.S. Census. *IEEE Security & Privacy* 21, 5 (2023). https://doi.org/10.1109/MSEC.2022.3202793.
[73] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. 2023. What Are the Chances? Explaining the Epsilon Parameter in Differential Privacy. In *USENIX Security '23*. https://www.usenix.org/conference/usenixsecurity23/presentation/nanayakkara.
[74] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-Anonymization of Large Sparse Datasets. In *IEEE S&P '08*. https://doi.org/10.1109/SP.2008.33.
[75] Arvind Narayanan and Vitaly Shmatikov. 2010. Myths and Fallacies of "Personally Identifiable Information". *Commun. ACM* 53, 6 (June 2010). https://doi.org/10.1145/1743546.1743558.
[76] Alondra Nelson. 2022. Ensuring Free, Immediate, and Equitable Access to Federally Funded Research. https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf.
[77] New South Wales Information and Privacy Commission. 2020. De-Identification of Personal Information. https://www.ipc.nsw.gov.au/sites/default/files/2021-03/Fact_Sheet_De-identification_of_personal_information_May_2020.pdf.
[78] Office of the Australian Information Commissioner. 2018. De-Identification and the Privacy Act. https://www.oaic.gov.au/privacy/privacy-guidance-for-organisations-and-government-agencies/handling-personal-information/de-identification-and-the-privacy-act.
[79] Office of the Victorian Information Commissioner. 2018. An Introduction to De-Identification. https://ovic.vic.gov.au/privacy/resources-for-organisations/an-introduction-to-de-identification/.
[80] Paul Ohm. 2010. Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review* 57 (2010). https://papers.ssrn.com/abstract=1450006.
[81] Okera. 2022. A Practical Guide to Data Anonymization and Soft Deletion. https://www.okera.com/wp-content/uploads/2022/06/Okera_Whitepaper_Practical_Guide_Data_Anonymization_R4-1.pdf.
[82] Personal Data Protection Commission of Singapore. 2022. Guide to Basic Data Anonymization. https://www.pdpc.gov.sg/help-and-resources/2018/01/basic-anonymisation.

[83] Amy Pienta, Joy Jang, and Margaret Levenstein. 2023. Beyond Legal Frameworks and Security Controls for Accessing Confidential Survey Data: Engaging Data Users in Data Protection. *Journal of Privacy and Confidentiality* 13, 2 (2023). https://doi.org/10.29012/jpc.845.
[84] Portage COVID-19 Working Group, Kristi Thompson, Erin Clary, Lucia Costanzo, Beth Knazook, Nick Rochlin, Felicity Tayler, Jane Fry, Chantal Ripp, Kathy Szigeti, Qian Zhang, Roger Reka, Minglu Wang, Rebecca Dickson, Mark Leggott, and Melanie Parlette-Stewart. 2020. De-Identification Guidance. https://doi.org/10.5281/zenodo.4270551.
[85] Privitar. 2022. The Top 7 Techniques for De-Identifying Data. https://www.privitar.com/wp-content/uploads/2022/09/P1006-EB-Top-7-De-identification-Techniques-DP.pdf.
[86] Alfred Rossi, Andrew Burt, and Sophie Stalla-Bourdillon. 2020. Deidentification 101: A Lawyer's Guide to Masking, Encryption and Everything in Between. https://iapp.org/news/a/de-identification-101-a-lawyers-guide-to-masking-encryption-and-everything-in-between/.
[87] Alfred Rossi, Andrew Burt, and Sophie Stalla-Bourdillon. 2020. Deidentification 201: A Lawyer's Guide to Pseudonymization and Anonymization. https://iapp.org/news/a/de-identification-201-a-lawyers-guide-to-pseudonymization-and-anonymization/.
[88] San José State University. [n. d.]. Table of De-Identification Techniques. https://www.sjsu.edu/research/docs/irb-deidentification-techniques-table.pdf.
[89] Carles Durá Santonja. 2022. 6 Personal Data Anonymization Techniques You Should Know About. https://blog.pangeanic.com/6-personal-data-anonymization-techniques.
[90] Satori. 2022. Guide: Data Masking. https://satoricyber.com/data-masking/data-masking-8-techniques-and-how-to-implement-them-successfully/.
[91] Koen Schoenmakers, Daniel Greene, Sarah Stutterheim, Herbert Lin, and Megan J. Palmer. 2023. The Security Mindset: Characteristics, Development, and Consequences. *Journal of Cybersecurity* 9, 1 (2023). https://doi.org/10.1093/cybsec/tyad010.
[92] Mary Anne Smart, Dhruv Sood, and Kristen Vaccaro. 2022. Understanding Risks of Privacy Theater with Differential Privacy. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022). https://doi.org/10.1145/3555762.
[93] Spanish Data Protection Authority. 2019. *k*-Anonymity as a Privacy Measure. https://www.aepd.es/es/documento/nota-tecnica-kanonimidad-en.pdf.
[94] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic Data – Anonymisation Groundhog Day. In *USENIX Security '22*. https://www.usenix.org/conference/usenixsecurity22/presentation/stadler.
[95] Sophie Stalla-Bourdillon. 2023. What Is Data De-Identification and Why Is It Important? https://www.immuta.com/blog/what-is-data-de-identification/.
[96] Latanya Sweeney. 2000. Simple Demographics Often Identify People Uniquely. https://dataprivacylab.org/projects/identifiability/.
[97] Latanya Sweeney. 2002. *k*-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (Oct. 2002). https://doi.org/10.1142/S0218488502001648.
[98] Matthias Templ, Bernhard Meindl, Alexander Kowarik, Johannes Gussenbauer, Organisation For Economic Co-Operation And Development, Statistics Netherlands, and Pascal Heus. 2024. sdcMicro. https://cloud.r-project.org/web/packages/sdcMicro/index.html.
[99] The New School Information Security & Privacy Office. 2020. Guidelines for Anonymization & Pseudonymization. https://ispo.newschool.edu/guidelines/anonymization-pseudonymization/.
[100] Joana Tomás, Deolinda Rasteiro, and Jorge Bernardino. 2022. Data Anonymization: An Experimental Evaluation Using Open-Source Tools. *Future Internet* 14, 6 (2022). https://doi.org/10.3390/fi14060167.
[101] Treasury Board of Canada Secretariat. 2023. Privacy Implementation Notice 2023 01: De-Identification. https://www.canada.ca/en/treasury-board-secretariat/services/access-information-privacy/access-information-privacy-notices/2023-01-de-identification.html.
[102] Allison R. B. Tyler. 2020. Facilitating Access to Restricted Data: Operationalizing Trust in Data Users. *International Journal of Digital Curation* 15, 1 (2020). https://doi.org/10.2218/ijdc.v15i1.602.
[103] UK Data Service. [n. d.]. Research Data Management. https://ukdataservice.ac.uk/learning-hub/research-data-management/.
[104] University of Groningen. 2023. Data Minimization & De-Identification. https://www.rug.nl/digital-competence-centre/privacy-and-data-protection/data-protection/de-identification.
[105] University of South Carolina Institute for Families in Society. 2018. Guidelines and Methods for De-Identifying Protected Health Information. https://www.schealthviz.sc.edu/Data/Sites/1/media/images/USC_IFS_PHIDataDeIdentification_18.pdf.
[106] University of Washington Privacy Office. 2019. Data Anonymization and De-Identification: Challenges and Options. https://privacy.uw.edu/wp-content/uploads/sites/7/2021/03/DataAnonymization_Aug2019.pdf.
[107] U.S. Agency for International Development. 2020. DDL Roadmap for Education Programs. https://www.edu-links.org/sites/default/files/media/file/DDL%20Roadmap_FINAL_07Aug2020_508.pdf.
[108] U.S. Department of Health and Human Services Office for Civil Rights. 2012. Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.
[109] Niek Van Wettere. 2020. Anonymization Tools and Techniques. https://zenodo.org/records/3843319.
[110] Daniel Votipka, Desiree Abrokwa, and Michelle L. Mazurek. 2020. Building and Validating a Scale for Secure Software Development Self-Efficacy. In *CHI '20*. https://doi.org/10.1145/3313831.3376754.
[111] Daniel Votipka, Rock Stevens, Elissa Redmiles, Jeremy Hu, and Michelle Mazurek. 2018. Hackers vs. Testers: A Comparison of Software Vulnerability Discovery Processes. In *IEEE S&P '18*.
[112] Zikai Alex Wen, Jingyu Jia, Hongyang Yan, Yaxing Yao, Zheli Liu, and Changyu Dong. 2023. The Influence of Explanation Designs on User Understanding Differential Privacy and Making Data-Sharing Decision. *Information Sciences* 642 (2023). https://doi.org/10.1016/j.ins.2023.03.024.
[113] Michael Wines. 2022. The 2020 Census Suggests That People Live Underwater. There's a Reason. *The New York Times* (April 2022). https://www.nytimes.com/2022/04/21/us/census-data-privacy-concerns.html.
[114] Xin Xia, Lingfeng Bao, David Lo, Pavneet Singh Kochhar, Ahmed E. Hassan, and Zhenchang Xing. 2017. What Do Developers Search for on the Web? *Empirical Software Engineering* 22, 6 (2017). https://doi.org/10.1007/s10664-017-9514-4.
[115] Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. 2020. Towards Effective Differential Privacy Communication for Users' Data Sharing Decision and Comprehension. In *IEEE S&P '20*. https://doi.org/10.1109/SP40000.2020.00088.
[116] Ayoung Yoon and Andrea Copeland. 2020. Toward Community-Inclusive Data Ecosystems: Challenges and Opportunities of Open Data for Community-Based Organizations. *Journal of the Association for Information Science and Technology* 71, 12 (Dec. 2020). https://doi.org/10.1002/asi.24346.
[117] William Zimmerman, Sharon Werning Rivera, and Kirill Kalinin. 2023. Survey of Russian Elites, Moscow, Russia, 1993-2020. https://doi.org/10.3886/ICPSR03724.V8.

A Sources of guide recommendations

We collected qualifying guides from the websites of six universities, research groups, and data repositories:

- Australian Research Data Commons: https://ardc.edu.au/resource/identifiable-data/
- University College Dublin: https://libguides.ucd.ie/data/ethics
- University of California San Francisco: https://data.ucsf.edu/cdrp/de-identification
- University of Winnipeg: https://libguides.uwinnipeg.ca/rdm/anonymization
- The World Bank Development Impact department: https://dimewiki.worldbank.org/De-identification
- Dryad: https://datadryad.org/docs/HumanSubjectsData.pdf

We also collected qualifying guides listed in the appendices of G1 (U.S. National Institute of Standards and Technology) [36] and as further/recommended reading in R6 (University of Groningen) [104].

Supplementary materials containing Appendices B–D are located at https://osf.io/mz4p5/.