



1

NIST Special Publication NIST SP 800-188 3pd

2

3

De-Identifying Government Data Sets

4

5

Third Public Draft

6

Simson Garfinkel

7

Phyllis Singer

8

Joseph Near

9

Aref N. Dajani

10

Barbara Guttman

11

This publication is available free of charge from:

12

<https://doi.org/10.6028/NIST.SP.800-188.3pd>

13

14

NIST Special Publication NIST SP 800-188 3pd

15

16

De-Identifying Government Data Sets

17

18

Third Public Draft

19

Simson Garfinkel

20

Barbara Guttman

21

Software Quality Group

22

Software and Systems Division

23

Joseph Near

24

Department of Computer Science

25

University of Vermont

26

Aref N. Dajani

27

Phyllis Singer

28

Center for Enterprise Dissemination

29

US Census Bureau

30

This publication is available free of charge from:

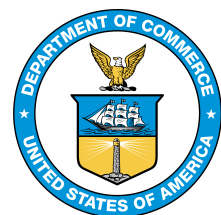
31

<https://doi.org/10.6028/NIST.SP.800-188.3pd>

32

November 2022

33



34

US Department of Commerce

35

Gina M. Raimondo, Secretary

36

National Institute of Standards and Technology

37

Laurie E. Locascio, NIST Director and Under Secretary of Commerce for Standards and Technology

38 Certain commercial entities, equipment, or materials may be identified in this document in order to describe
39 an experimental procedure or concept adequately. Such identification is not intended to imply
40 recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to
41 imply that the entities, materials, or equipment are necessarily the best available for the purpose.

42 There may be references in this publication to other publications currently under development by NIST in
43 accordance with its assigned statutory responsibilities. The information in this publication, including
44 concepts and methodologies, may be used by federal agencies even before the completion of such
45 companion publications. Thus, until each publication is completed, current requirements, guidelines, and
46 procedures, where they exist, remain operative. For planning and transition purposes, federal agencies may
47 wish to closely follow the development of these new publications by NIST.

48 Organizations are encouraged to review all draft publications during public comment periods and provide
49 feedback to NIST. Many NIST cybersecurity publications, other than the ones noted above, are available at
50 <https://csrc.nist.gov/publications>.

51 This document is presented with the hope that its content may be of interest to the general privacy
52 community. The views in this document are those of the authors, and do not represent those of the US
53 Census Bureau.

54 **Authority**

55 This publication has been developed by NIST in accordance with its statutory responsibilities under the
56 Federal Information Security Modernization Act (FISMA) of 2014, 44 U.S.C. § 3551 et seq., Public Law
57 (P.L.) 113-283. NIST is responsible for developing information security standards and guidelines, including
58 minimum requirements for federal information systems, but such standards and guidelines shall not apply to
59 national security systems without the express approval of appropriate federal officials exercising policy
60 authority over such systems. This guideline is consistent with the requirements of the Office of Management
61 and Budget (OMB) Circular A-130.

62 Nothing in this publication should be taken to contradict the standards and guidelines made mandatory and
63 binding on federal agencies by the Secretary of Commerce under statutory authority. Nor should these
64 guidelines be interpreted as altering or superseding the existing authorities of the Secretary of Commerce,
65 Director of the OMB, or any other federal official. This publication may be used by nongovernmental
66 organizations on a voluntary basis and is not subject to copyright in the United States. Attribution would,
67 however, be appreciated by NIST.

68 **NIST Technical Series Policies**

69 [Copyright, Fair Use, and Licensing Statements](#)
70 [NIST Technical Series Publication Identifier Syntax](#)

71 **Publication History**

72 Approved by the NIST Editorial Review Board on YYYY-MM-DD [will be added upon final publication]

73 **How to cite this NIST Technical Series Publication:**

74 Garfinkel S, Guttman B, Near J, Dajani AN, Singer P (2022) De-Identifying Government Data
75 Sets. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Special Publication (SP)
76 NIST SP 800-188 3pd. <https://doi.org/10.6028/NIST.SP.800-188.3pd>

77 **Author ORCID iDs**

78 Simson Garfinkel: 0000-0003-1294-2831
79 Joseph Near: 0000-0002-3203-3742
80 Aref N. Dajani: 0000-0003-0361-5409
81 Phyllis Singer: 0000-0002-8885-7273

82 **Public Comment Period**

83 November 15, 2022 – January 15, 2023

84 **Submit Comments**

85 sp800-188-draft@nist.gov

86 National Institute of Standards and Technology

87 Attn: Software and Systems Division, Information Technology Laboratory

88 100 Bureau Drive (Mail Stop 8970) Gaithersburg, MD 20899-8970

89 **All comments are subject to release under the Freedom of Information Act (FOIA).**

90 **Abstract**

91 De-identification is a process that is applied to a dataset with the goal of preventing or
92 limiting informational risks to individuals, protected groups, and establishments while still
93 allowing for meaningful statistical analysis. Government agencies can use de-identification
94 to reduce the privacy risk associated with collecting, processing, archiving, distributing,
95 or publishing government data. Previously, NISTIR 8053, *De-Identification of Personal*
96 *Information* [51], provided a survey of de-identification and re-identification techniques.
97 This document provides specific guidance to government agencies that wish to use de-
98 identification. Before using de-identification, agencies should evaluate their goals for us-
99 ing de-identification and the potential risks that de-identification might create. Agencies
100 should decide upon a de-identification release model, such as publishing de-identified data,
101 publishing synthetic data based on identified data, or providing a query interface that incor-
102 porates de-identification. Agencies can create a Disclosure Review Board to oversee the
103 process of de-identification. They can also adopt a de-identification standard with measur-
104 able performance levels and perform re-identification studies to gauge the risk associated
105 with de-identification. Several specific techniques for de-identification are available, in-
106 cluding de-identification by removing identifiers and transforming quasi-identifiers and the
107 use of formal privacy models. People performing de-identification generally use special-
108 purpose software tools to perform the data manipulation and calculate the likely risk of
109 re-identification. However, not all tools that merely mask personal information provide
110 sufficient functionality for performing de-identification. This document also includes an
111 extensive list of references, a glossary, and a list of specific de-identification tools, which is
112 only included to convey the range of tools currently available and is not intended to imply
113 a recommendation or endorsement by NIST.

114 **Keywords**

115 data life cycle; de-identification; differential privacy; direct identifiers; Disclosure Re-
116 view Board; the five safes; *k*-anonymity; privacy; pseudonymization; quasi-identifiers;
117 re-identification; synthetic data.

118 **Reports on Computer Systems Technology**

119 The Information Technology Laboratory (ITL) at the National Institute of Standards and
120 Technology (NIST) promotes the U.S. economy and public welfare by providing technical
121 leadership for the Nation's measurement and standards infrastructure. ITL develops tests,
122 test methods, reference data, proof of concept implementations, and technical analyses to
123 advance the development and productive use of information technology. ITL's responsi-
124 bilities include the development of management, administrative, technical, and physical
125 standards and guidelines for the cost-effective security and privacy of other than national
126 security-related information in federal information systems. The Special Publication 800-

127 series reports on ITL’s research, guidelines, and outreach efforts in information system
128 security, and its collaborative activities with industry, government, and academic organiza-
129 tions.

130 **Call for Patent Claims**

131 This public review includes a call for information on essential patent claims (claims whose
132 use would be required for compliance with the guidance or requirements in this Information
133 Technology Laboratory (ITL) draft publication). Such guidance and/or requirements may
134 be directly stated in this ITL Publication or by reference to another publication. This call
135 also includes disclosure, where known, of the existence of pending U.S. or foreign patent
136 applications relating to this ITL draft publication and of any relevant unexpired U.S. or
137 foreign patents.

138 ITL may require from the patent holder, or a party authorized to make assurances on its
139 behalf, in written or electronic form, either:

- 140 1. assurance in the form of a general disclaimer to the effect that such party does not
141 hold and does not currently intend holding any essential patent claim(s); or
- 142 2. assurance that a license to such essential patent claim(s) will be made available to ap-
143 plicants desiring to utilize the license for the purpose of complying with the guidance
144 or requirements in this ITL draft publication either:
 - 145 (a) under reasonable terms and conditions that are demonstrably free of any unfair
146 discrimination; or
 - 147 (b) without compensation and under reasonable terms and conditions that are demon-
148 strably free of any unfair discrimination.

149 Such assurance shall indicate that the patent holder (or third party authorized to make assur-
150 ances on its behalf) will include in any documents transferring ownership of patents subject
151 to the assurance, provisions sufficient to ensure that the commitments in the assurance are
152 binding on the transferee, and that the transferee will similarly include appropriate provi-
153 sions in the event of future transfers with the goal of binding each successor-in-interest.

154 The assurance shall also indicate that it is intended to be binding on successors-in-interest
155 regardless of whether such provisions are included in the relevant transfer documents.

156 Such statements should be addressed to: sp800-188-draft@nist.gov

159 **Table of Contents**

160	Executive Summary	1
161	1. Introduction	3
162	1.1. Document Purpose and Scope	7
163	1.2. Intended Audience	7
164	1.3. Organization	7
165	2. Introducing De-Identification	8
166	2.1. Historical Context	8
167	2.2. Terminology	10
168	3. Governance and Management of Data De-Identification	17
169	3.1. Identifying Goals and Intended Uses of De-Identification	17
170	3.2. Evaluating Risks that Arise from De-Identified Data Releases	18
171	3.2.1. Probability of Re-Identification	19
172	3.2.2. Adverse Impacts of Re-Identification	22
173	3.2.3. Impacts Other Than Re-Identification	23
174	3.2.4. Remediation	24
175	3.3. Data Life Cycle	24
176	3.4. Data-Sharing Models	29
177	3.5. The Five Safes	30
178	3.6. Disclosure Review Boards	31
179	3.7. De-Identification Standards	36
180	3.7.1. Benefits of Standards	36
181	3.7.2. Prescriptive De-Identification Standards	36
182	3.7.3. Performance-Based De-Identification Standards	37
183	3.8. Education, Training, and Research	38
184	3.9. Defense in Depth	38
185	3.9.1. Encryption and Access Control	38
186	3.9.2. Secure Computation	38
187	3.9.3. Trusted Execution Environments	39
188	3.9.4. Physical Enclaves	39
189	4. Technical Steps for Data De-Identification	40

190	4.1. Determine the Privacy, Data Usability, and Access Objectives	40
191	4.2. Conducting a Data Survey	41
192	4.3. De-Identification by Removing Identifiers and Transforming Quasi-Identifiers	43
193	4.3.1. Removing or Transforming of Direct Identifiers	44
194	4.3.2. Special Security Note Regarding the Encryption or Hashing of Di-	
195	rect Identifiers	46
196	4.3.3. De-Identifying Numeric Quasi-Identifiers	46
197	4.3.4. De-Identifying Dates	48
198	4.3.5. De-Identifying Geographical Locations and Geolocation Data . . .	49
199	4.3.6. De-Identifying Genomic Information	49
200	4.3.7. De-Identifying Text Narratives and Qualitative Information	51
201	4.3.8. Challenges Posed by Aggregation Techniques	51
202	4.3.9. Challenges Posed by High-Dimensional Data	52
203	4.3.10. Challenges Posed by Linked Data	52
204	4.3.11. Challenges Posed by Composition	52
205	4.3.12. Potential Failures of De-Identification	53
206	4.3.13. Post-Release Monitoring	54
207	4.4. Synthetic Data	54
208	4.4.1. Partially Synthetic Data	55
209	4.4.2. Test Data	56
210	4.4.3. Fully Synthetic Data	56
211	4.4.4. Synthetic Data with Validation	58
212	4.4.5. Synthetic Data and Open Data Policy	58
213	4.4.6. Creating a Synthetic Dataset with Differential Privacy	58
214	4.5. De-Identifying with an Interactive Query Interface	59
215	4.6. Validating a De-Identified Dataset	60
216	4.6.1. Validating Data Usefulness	60
217	4.6.2. Validating Privacy Protection	60
218	4.6.3. Re-Identification Studies	61
219	5. Software Requirements, Evaluation, and Validation	63
220	5.1. Evaluating Privacy-Preserving Techniques	63
221	5.2. De-Identification Tools	64

222	5.2.1. De-Identification Tool Features	64
223	5.2.2. Data Provenance and File Formats	64
224	5.2.3. Data Masking Tools	64
225	5.3. Evaluating De-Identification Software	65
226	5.4. Evaluating Data Accuracy	65
227	6. Conclusion	66
228	References	80
229	Appendix A. Standards	81
230	A.1. NIST Publications	81
231	A.2. Other U.S. Government Publications	82
232	Selected Publications by Other Governments	83
233	Reports and Books	83
234	How-To Articles	84
235	Appendix B. List of Symbols, Abbreviations, and Acronyms	86
236	Appendix C. Glossary	89

237 **List of Tables**

238	Table 1. Reading levels at a hypothetical school, as measured by entrance examinations, reported at the start of the school year on October 1.	51
239		
240	Table 2. Reading levels at a hypothetical school, as measured by entrance examinations, reported one month into the school year on November 1 after a new student has transferred to the school.	51
241		
242		
243	Table 3. Adjectives used for describing data in data releases.	55

244 **List of Figures**

245	Fig. 1. The data life cycle as described by Michener et al. [80]	25
246	Fig. 2. Chisholm’s view of the data life cycle is a linear process with a branching point after data usage [25]	25
247		
248	Fig. 3. Altman’s “modern approach to privacy-aware government data releases” [75] [26]	26
249	Fig. 4. Altman’s conceptual diagram of the relationship between post-transformation identifiability, level of expected harm, and suitability of selected privacy controls for a data release [75]	27
250		
251		
252	Fig. 5. Advice for Practitioners: A Summary	68

253 **Acknowledgments**

254 The authors wish to thank the U.S. Census Bureau for its help in researching and preparing this
255 publication, with specific thanks to John Abowd, Ron Jarmin, Christa Jones, and Laura McKenna.
256 The authors would also like to thank Luk Arbuckle, Andrew Baker, Daniel Barth-Jones, Christi
257 Dant, Khaled El Emam, Robert Gellman, Tom Krenzke, Bradley Malin, Kevin Mangold, John
258 Moehrke, Linda Sanchez, Denise Sturdy, and Chris Traver for providing comments on previous
259 drafts and their valuable insights, all of which were helpful in creating this publication.

260 The authors also wish to thank several organizations that provided useful comments on previous
261 drafts of this publication: the Defense Contract Management Agency (DCMA) Information As-
262 surance Directorate; the Office of Chief Privacy Officer within the U.S. Department of Education;
263 the Office of Planning, Research, and Evaluation (OPRE) within the Administration for Children
264 and Families at the U.S. Department of Health and Human Services; the Millennium Challenge
265 Corporation (MCC) Department of Policy and Evaluation; Integrating the Healthcare Enterprise
266 (IHE), an ANSI-accredited standards organization focused on healthcare standards; and the Privacy
267 Tools project at Harvard University (including Micah Altman, Stephen Chong, Kobbi Nissim, David
268 O'Brien, Salil Vadhan, and Alexandra Wood).

269 **Author Contributions**

270 **Simson Garfinkel:** Conceptualization, Supervision, Writing (original draft preparation); **Joseph**
271 **Near:** Writing (original draft preparation); **Aref N. Dajani:** Writing (original draft preparation of
272 the section on reidentification studies); **Phyllis Singer:** Writing (original draft preparation of the
273 section on reidentification studies).

274 **Executive Summary**

275 Every federal agency creates and maintains internal datasets that are vital for fulfilling its
276 mission. The Foundation for Evidence-based Policymaking Act of 2018 [2] mandates that
277 agencies also collect and publish their government data in open, machine-readable formats,
278 when it is appropriate to do so. Agencies can use de-identification to make government
279 datasets available while protecting the privacy of the individuals whose data are contained
280 within those datasets.

281 Many Government documents use the phrase *personally identifiable information* (PII) to
282 describe private information that can be linked to an individual [62, 79], although there are
283 a variety of definitions for PII. As a result, it is possible to have information that *singles*
284 *out* individuals but that does not meet a specific definition of PII. This document therefore
285 presents ways of removing or altering information that can identify individuals that go
286 beyond merely removing PII.

287 For decades, de-identification based on simply removing of identifying information was
288 thought to be sufficient to prevent the re-identification of individuals in large datasets. Since
289 the mid 1990s, a growing body of research has demonstrated the reverse, resulting in new
290 privacy attacks capable of re-identifying individuals in “de-identified” data releases. For
291 several years the goals of such attacks appeared to be the embarrassment of the publishing
292 agency and achieving academic distinction for the privacy researcher [50]. More recently,
293 as high-resolution de-identified geolocation data has become commercially available, re-
294 identification techniques have been used by journalists and activists [100, 140, 70] with the
295 goal of learning confidential information.

296 These attacks have become more sophisticated in recent years with the availability of ge-
297 olocation data, highlighting the deficiencies in traditional

298 Formal models of privacy, like k -anonymity [122] and differential privacy, [39] use math-
299 ematically rigorous approaches that are designed to allow for the controlled use of confi-
300 dential data while minimizing the privacy loss suffered by the data subjects. Because there
301 is an inherent trade-off between the accuracy of published data and the amount of privacy
302 protection afforded to data subjects, most formal methods have some kind of parameter
303 that can be adjusted to control the “privacy cost” of a particular data release. Informally, a
304 data release with a low privacy cost causes little additional privacy risk to the participants,
305 while a higher privacy cost results in more privacy risk. When they are available, formal
306 privacy methods should be preferred over informal, *ad hoc* methods.

307 Decisions and practices regarding the de-identification and release of government data can
308 be integral to the mission and proper functioning of a government agency. As such, an
309 agency’s leadership should manage these activities in a way that assures performance and
310 results in a manner that is consistent with the agency’s mission and legal authority. One way
311 that agencies can manage this risk is by creating a formal Disclosure Review Board (DRB)
312 that consists of legal and technical privacy experts, stakeholders within the organization,

313 and representatives of the organization’s leadership. The DRB evaluated applications for
314 data release that describe the confidential data, the techniques that will be used to mini-
315 mize the risk of disclosure, the resulting protected data, and how the effectiveness of those
316 techniques will be evaluated.

317 Establishing a DRB may seem like an expensive and complicated administrative under-
318 taking for some agencies. However, a properly constituted DRB and the development of
319 consistent procedures regarding data release should enable agencies to lower the risks as-
320 sociated with each data release, which is likely to save agency resources in the long term.

321 Agencies can create or adopt standards to guide those performing de-identification, and
322 regarding regarding the accuracy of de-identified data. If accuracy goals exist, then tech-
323 niques such as differential privacy can be used to make the data sufficiently accurate for the
324 intended purpose but not unnecessarily more accurate, which can limit the amount of pri-
325 vacy loss. However, agencies must carefully choose and implement accuracy requirements.
326 If data accuracy and privacy goals cannot be well-maintained, then releases of data that are
327 not sufficiently accurate can result in incorrect scientific conclusions and policy decisions.

328 Agencies should consider performing de-identification with trained individuals using soft-
329 ware specifically designed for the purpose. While it is possible to perform de-identification
330 with off-the-shelf software like a commercial spreadsheet or financial planning program,
331 such programs typically lack the key functions required for proper de-identification. As a
332 result, they may encourage the use of simplistic de-identification methods, such as deleting
333 sensitive columns and manually searching and removing data that appears sensitive. This
334 may result in a dataset that appears de-identified but that still contain significant disclosure
335 risks.

336 Finally, different countries have different standards and policies regarding the definition and
337 use of de-identified data. Information that is regarded as de-identified in one jurisdiction
338 may be regarded as being identifiable in another.

339 1. Introduction

340 The U.S. Government collects, maintains, and uses many kinds of datasets. Every federal
341 agency creates and maintains internal datasets that are vital for fulfilling its mission, such
342 as delivering services to taxpayers or ensuring regulatory compliance. There are also 13
343 principal federal statistical agencies, three recognized statistical units, and over 100 other
344 federal statistical programs that collect, compile, process, analyze, and distribute informa-
345 tion for statistical purposes [126, 92].

346 Government programs collect information from individuals and organizations for taxation,
347 public benefits, public health, licensing, employment, censuses, and the production of of-
348 ficial statistics. While privacy is integral, many individuals and organizations that provide
349 information to the Government do not typically have the right to opt-out of such requests.
350 For example, people and establishments in the United States are required by law to respond
351 to mandatory U.S. Census Bureau surveys.

352 Agencies make many of their datasets available to the public. The U.S. Government
353 publishes data to promote commerce, scientific research, and public transparency. Many
354 datasets contain some data elements that should not be made public, and it is necessary to
355 remove such information before making the rest of the dataset available. Some datasets
356 are so sensitive that they cannot be made publicly available at all but can be available on a
357 limited basis to qualified, vetted researchers in protected enclaves. In some cases, agencies
358 may also elect to release summary statistics of sensitive data or create synthetic datasets
359 that resemble the original data but that have a lower disclosure risk [8].

360 There is frequent tension between the goals of privacy protection and the release of useful
361 data to the public. One way that the Government attempts to resolve this tension is with
362 an official promise of confidentiality to individuals and organizations regarding the infor-
363 mation that they provide [102]. A bedrock principle of official statistical programs is that
364 data provided to the Government should generally remain confidential and not be used in a
365 way that could harm the individual or the organization providing the data. One justification
366 for this principle is that it helps to ensure high data accuracy. If data providers did not feel
367 that the information they provide would remain confidential, they might not be willing to
368 provide information that is accurate.

369 Other information is created by the Government as a consequence of providing government
370 services. This information – sometimes called *administrative data* – is also increasingly
371 being used and made available for statistical purposes and must be protected.

372 In 2018, the U.S. Congress passed three laws that significantly increased the need for ex-
373 pertise regarding privacy-preserving data analysis and data publishing techniques, such as
374 de-identification:

- 375 1. **The Foundations for Evidence-Based Policymaking Act of 2018** [2], commonly
376 called the Evidence Act, requires federal agencies to track all of their data in data

377 inventories, report public datasets to <https://data.gov>, perform systematic evidence-
378 making and evaluation activities, and engage in capacity-building so that the federal
379 workforce can meet the requirements of data-centric, evidence-based operations. The
380 Evidence Act is based on the findings of the U.S. Commission on Evidence-Based
381 Policymaking [27] and is implemented in part by OMB Memorandum M-19-23
382 [139].

383 The Evidence Act contains specific guidance requiring that agencies publishing data
384 take into account “(A) risks and restrictions related to the disclosure of personally
385 identifiable information, including the risk that an individual data asset in isolation
386 does not pose a privacy or confidentiality risk but when combined with other available
387 information may pose such a risk;” and “(B) security considerations, including the
388 risk that information in an individual data asset in isolation does not pose a security
389 risk but when combined with other available information may pose such a risk” [2].

390 2. **The Open Government Data Act**, which was passed as part of the Evidence Act,
391 requires that the U.S. Government publish data in machine-readable, open, non-
392 proprietary formats when possible. This act largely codified presidential Executive
393 Order 13642 of May 9, 2013, “Making Open and Machine Readable the New Default
394 for Government Information” [88] and its implementation in OMB Memorandum M-
395 13-13 [18].

396 3. **The Geospatial Data Act of 2018**, which requires that government agencies make
397 inventories of their geospatial data and that public geospatial data be registered on
398 the U.S. Government’s public geospatial platform, <https://www.geoplatform.gov/>.

399 Other laws, regulations, and policies that govern the release of statistics and data to the
400 public enshrine this principle of confidentiality. For example:

401 • **The Confidential Information Protection and Statistical Efficiency Act of 2002**
402 states, “data or information acquired by an agency under a pledge of confidentiality
403 for exclusively statistical purposes shall not be disclosed by an agency in identifi-
404 able form for any use other than an exclusively statistical purpose, except with the
405 informed consent of the respondent.” [126, §512 (b)(1)] Commonly called CIPSEA,
406 the act further requires that federal statistical agencies “establish appropriate admin-
407 istrative, technical, and physical safeguards to ensure the security and confidentiality
408 of records and to protect against any anticipated threats or hazards to their security
409 or integrity which could result in substantial harm, embarrassment, inconvenience,
410 or unfairness to any individual on whom information is maintained.”

411 • **US Code Title 13, Section 9** governs the confidentiality of information provided to
412 the Census Bureau and prohibits “any publication whereby the data furnished by any
413 particular establishment or individual under this title can be identified” [130].

414 • **US Code Title 26, Section 6103** governs the confidentiality of information provided
415 to the U.S. Government on tax returns and other return information. These rules are

416 now spelled out in IRS Publication 1075, “Tax Information Security Guidelines for
417 Federal, State and Local Agencies,” published by the IRS Office of Safeguards [93].

418 • **The Privacy Act of 1974** covers the release of personal information of U.S. citizens
419 and Lawful Permanent Residents by the Government. The Act recognizes that the
420 disclosure of records for statistical purposes is acceptable if the data are not “indi-
421 vidually identifiable” [103, at a(b)(5)].

422 Minimizing privacy risk is not an absolute goal of federal laws and regulations. Guidance
423 from the U.S. Department of Health and Human Services (HHS) on the Health Insurance
424 Portability and Accountability Act (HIPAA) de-identification standards notes that “[b]oth
425 methods [the safe harbor and expert determination methods for de-identification], even
426 when properly applied, yield de-identified data that retains some risk of identification. Al-
427 though the risk is very small, it is not zero, and there is a possibility that de-identified data
428 could be linked back to the identity of the patient to which it corresponds” [136].

429 U.S. law also balances privacy risk with other factors, such as transparency, accountabil-
430 ity, and the opportunity for public good. An example of this balance is the handling of
431 personally identifiable information collected by the Census Bureau as part of the decennial
432 census: this information remains confidential for 72 years and is then transferred to the
433 National Archives and Records Administration where it is released to the public [131, 5].

434 *De-identification* is a process that is applied to a dataset with the goal of preventing or lim-
435 iting privacy risks to individuals, protected groups, and establishments while still allowing
436 for the production of aggregate statistics.¹ De-identification is not a single technique, but
437 a collection of approaches, algorithms, and tools that can be applied to different kinds of
438 data with differing levels of effectiveness. In general, the potential risk to privacy posed by
439 a dataset’s release decreases as more aggressive de-identification techniques are employed,
440 but data accuracy and – in some cases – the ultimate utility of the de-identified dataset
441 decreases as well.

442 *Accuracy* is traditionally defined as the “closeness of computations or estimates to the exact
443 or true values that the statistics were intended to measure” [9]. The *data accuracy* of
444 de-identified data, therefore, refers to the degree to which inferences drawn on the de-
445 identified data will be consistent with inferences drawn on the original data. Data accuracy
446 can be measured by the ratio of a value computed with de-identified data to the same value
447 computed using the underlying true confidential value.

448 In economics, *Utility* is traditionally defined as “the satisfaction derived from consumption
449 of a good or service”[138]. *Data utility* therefore refers to the value that data users can de-
450 rive from data in general. When speaking of de-identified data, utility comes from two pub-

¹In Europe, the term *data anonymization* is frequently used as a synonym for de-identification, but the terms may have subtly different definitions in some contexts. For a more complete discussion of de-identification and data anonymization, see NISTIR 8053, *De-Identification of Personal Data* [51].

451 lic goods: the uses of the data and the privacy protection afforded by the de-identification
452 process.

453 This document uses the phrase *data accuracy* to refer to the abstract characteristic of the
454 data as determined by a specific, measurable statistic, whereas *data utility* refers to the ben-
455 efit derived from the application of the data to a specific use. Although there has previously
456 been a tendency within official statistical organizations to conflate these two terms, it is im-
457 portant to keep them distinct because they are not necessarily correlated. Data may have low
458 accuracy because they contain errors or substantial noise, yet users may nevertheless derive
459 high value from the data, giving the data high utility. Likewise, data that are very close to
460 the reality of the thing being measured may have high accuracy but may be fundamentally
461 worthless and, thus, have low utility.

462 In general, data accuracy decreases as more aggressive de-identification techniques are
463 employed. Therefore, any effort that involves the release of data that contain personal
464 information typically involves making a trade-off between identifiability and data accuracy.
465 However, increased privacy protections do not necessarily result in decreased data utility.

466 Some users of de-identified data may be able to use the data to make inferences about
467 private facts regarding the data subjects. They may even be able to re-identify the data
468 subjects. Both of these uses undo the privacy goals of de-identification. Agencies that
469 release data should understand what data they are releasing, what other data may already
470 be publicly or privately available, and the risk of re-identification. Agencies should aim to
471 make an informed decision about the fidelity of the data that they release by systematically
472 evaluating the risks and benefits and choosing de-identification techniques and data sharing
473 models that are tailored to their requirements. In addition, when telling individuals that
474 their de-identified information will be released, agencies should disclose that privacy risks
475 may remain despite de-identification.

476 Planning is essential for successful de-identification and data release. In a research envi-
477 ronment, this planning should include the research design, data collection, protection of
478 identifiers, disclosure analysis, and data-sharing strategy. In an operational environment,
479 this planning includes a comprehensive analysis of the purpose of the data release and the
480 expected use of the released data, the privacy-related risks, and the privacy protecting con-
481 trols. Both cases should review the appropriateness of various privacy controls given the
482 risks, intended uses, and the ways that those controls could fail.

483 De-identification can have significant costs, including time, labor, and data processing
484 costs. However, when properly executed, this effort can result in data that have high value
485 for a research community and the general public while still adequately protecting individual
486 privacy.

487 **1.1. Document Purpose and Scope**

488 This document provides guidance on the selection, use, and evaluation of de-identification
489 techniques for U.S. Government datasets. It also provides a framework that can be adapted
490 by federal agencies to shape the governance of de-identification processes. The ultimate
491 goal of this document is to reduce disclosure risks that might result from an intentional data
492 release.

493 **1.2. Intended Audience**

494 This document is intended for use by government engineers, data scientists, privacy offi-
495 cers, disclosure review boards, and other officials. It is also designed to be generally infor-
496 mative to researchers and academics involved in the technical aspects of the de-identification
497 of government data. While this document assumes a high-level understanding of informa-
498 tion system security technologies, it is intended to be accessible to a wide audience.

499 **1.3. Organization**

500 The remainder of this publication is organized as follows:

- 501 • **Section 2, “Introducing De-Identification,”** presents a background on the science
502 and terminology of de-identification.
- 503 • **Section 3, “Governance and Management of Data De-Identification,”** provides
504 guidance to agencies on the establishment of or improvement to a program that makes
505 privacy-sensitive data available to researchers and the public.
- 506 • **Section 4, “Technical Steps for Data De-Identification,”** provides specific tech-
507 nical guidance for performing de-identification using a variety of mathematical ap-
508 proaches.
- 509 • **Section 5, “Software Requirements, Evaluation, and Validation,”** provides a rec-
510 ommended set of features that should be in de-identification tools, which may be use-
511 ful for potential purchasers or developers of such software. This section also provides
512 information for evaluating both de-identification tools and de-identified datasets.
- 513 • **Section 6, “Conclusion,”** Section 6 is the conclusion.

514 Following the conclusion, this document provides a list of all publications referenced
515 in this document, as well as an Appendix that includes standards, related NIST pub-
516 lications, other selected publications by the US and other governments, reports and
517 books, and a few articles of interest. A second appendix provides a list of symbols,
518 abbreviations and acronyms. The third appendix contains a glossary.

519 2. Introducing De-Identification

520 This document presents recommendations for de-identifying government datasets.

521 If the information derived from personal data remains in a de-identified dataset, the dataset
522 might inadvertently reveal attributes related to specific individuals, specific de-identified
523 records could be linked back to specific individuals. When this happens, the privacy pro-
524 tection provided by de-identification is compromised. Even if a specific individual cannot
525 be matched to a specific data record, de-identified data can be used to improve the accu-
526 racy of inferences regarding individuals whose de-identified data are in the dataset. This
527 so-called *inference risk* cannot be eliminated if there is any information in the de-identified
528 data, but it can be minimized. Thus, the decision of how or whether to de-identify data
529 should be made in conjunction with decisions over how the de-identified data will be used,
530 shared, or released.

531 De-identification is especially important for government agencies, businesses, and other or-
532 ganizations that seek to make data available to outsiders. For example, significant medical
533 research resulting in societal benefit is made possible by the sharing of de-identified patient
534 information under the framework established by the HIPAA Privacy Rule, the primary U.S.
535 regulation that provides for the privacy of medical records; billing records; enrollment, pay-
536 ment, and claims records; and “other records that are used, in whole or in part, by or for
537 the covered entity to make decisions about individuals” [90]. The HIPAA Privacy Rule de-
538 identification framework applies to both government organizations charged with protecting
539 government datasets as well as to private sector organizations, such as health plans and
540 health care providers.

541 Agencies may also be required to de-identify records when responding to a Freedom of
542 Information Act (FOIA) [134, 133] request in a manner that is consistent with Exemption
543 6, which protects information about individuals in “personnel and medical files and similar
544 files” when the disclosure of such information “would constitute a clearly unwarranted in-
545 vasion of personal privacy,” and Exemption 7(C), which is limited to information compiled
546 for law enforcement purposes and protects personal information when disclosure “could
547 reasonably be expected to constitute an unwarranted invasion of personal privacy.” The
548 meaning of these exemptions has been clarified by multiple cases before the US Supreme
549 Court [105, 117, 118].

550 2.1. Historical Context

551 The modern practice of de-identification comes from three overlapping intellectual tradi-
552 tions.

- 553 1. For four decades, official statistical agencies have researched and investigated meth-
554 ods broadly termed *Statistical Disclosure Limitation* (SDL) or *Statistical Disclosure*

555 *Control* [29, 36].² Statistical agencies created these methods so that they could re-
556 lease statistical tables and *public use files* (PUF) to allow users to learn information
557 and perform original research while protecting the privacy of the individuals in the
558 dataset. SDL is widely used in contemporary statistical reporting.

559 2. In the 1990s, there was a significant increase in the release of *microdata* files for
560 public use in the form of both individual responses from surveys and administrative
561 records. Initially, these releases merely stripped obviously identifying information,
562 such as names and social security numbers (what are now called *direct identifiers*).
563 Following some releases, researchers discovered that it was possible to re-identify
564 individuals' data by triangulating with some of the remaining data (now called *quasi-*
565 *identifiers* or *indirect identifiers* [28]). The research resulted in the invention of the
566 *k*-anonymity model for protecting privacy [124, 108, 109, 123], which is reflected
567 in the Office of Civil Rights guidance on how to apply de-identification in a manner
568 consistent with the HIPAA Privacy Rule [89]. Today, variants of *k*-anonymity are
569 commonly used to allow for the sharing of medical microdata. This intellectual tra-
570 dition is typically called *de-identification*, although this document uses that term to
571 describe all three intellectual traditions.

572 3. In the 2000s, research in theoretical computer science and cryptography developed
573 the theory of *differential privacy* [40], which is based on a mathematical definition
574 of the privacy loss to an individual that results from queries on a database containing
575 that individual's personal information. Differential privacy is termed a *formal model*
576 *for privacy protection* because its definitions for privacy and privacy loss are based on
577 mathematical proofs.³ This does not mean that algorithms that implement differen-
578 tial privacy cannot result in increased privacy risk. Rather, it means that the amount
579 of privacy risk that results from the use of these algorithms can be mathematically
580 bounded. These mathematical limits on privacy risk have created considerable inter-
581 est in differential privacy in academia, commerce, and business. To date, however,
582 only a few systems that utilize differential privacy have been operationally deployed.

583 During the first decade of the 21st century, there was a growing awareness within the U.S.
584 Government about the risks that could result from the improper handling and inadvertent
585 release of personal identifying and financial information. This realization, combined with
586 a growing number of inadvertent data disclosures within the U.S. Government, resulted
587 in President George Bush signing Executive Order 13402, which established an Identity
588 Theft Task Force on May 10, 2006 [19]. One year later, the Office of Management and
589 Budget issued Memorandum M-07-16 [62], which required federal agencies to develop
590 and implement breach notification policies. As part of this effort, NIST issued Special

²A summary of the history of Statistical Disclosure Limitation can be found in *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics* [102].

³Other formal methods for privacy include cryptographic algorithms and techniques with provably secure properties, privacy-preserving data mining, Shamir's secret sharing, and advanced database techniques. A summary of such techniques appears in [128].

591 Publication (SP) 800-122, *Guide to Protecting the Confidentiality of Personally Identifiable*
592 *Information (PII)* [79]. These policies and documents had the specific goal of limiting the
593 accessibility of information that could be directly used for identity theft but did not create
594 a framework for processing government datasets so that they could be released without
595 impacting the privacy of the data subjects.

596 In 2015, NIST published NISTIR 8053, *De-Identification of Personal Information* [51],
597 which provided an overview of de-identification issues and terminology. It also sum-
598 marized significant publications involving de-identification and re-identification. How-
599 ever, NISTIR 8053 did not make recommendations regarding the appropriateness of de-
600 identification or specific de-identification algorithms. The following year, NIST convened
601 a Government Data De-Identification Stakeholder’s Meeting [52].

602 De-identification is one of several models for allowing the controlled sharing of personal
603 data and other kinds of sensitive data.⁴ Other models include the use of data processing en-
604 claves, where computations are performed with confidential data using computers that are
605 physically isolated from the outside world. That isolation might be performed with locked
606 doors and guards, or it might be performed using silicon and encryption, as is the case
607 with enclaves implemented on some modern microprocessors. Another approach is to use
608 mathematical techniques – such as secure multiparty computation – so that computations
609 can be carried out on confidential data held by multiple parties without ever bringing all of
610 the confidential data together in a single location.

611 Techniques for privacy-preserving data-sharing and analysis can be layered to provide
612 stronger protection than any single technique would provide in isolation. Such comple-
613 mentary models are discussed in Section 3.4. For a more complete description of data-
614 sharing models, privacy-preserving data publishing, and privacy-preserving data mining,
615 see NISTIR 8053.

616 Many of the techniques discussed in this publication (e.g., fully synthetic data and differen-
617 tial privacy) have limited use within the Federal Government due to cost, time constraints,
618 and the sophistication required of practitioners. However, these techniques are likely to
619 see increased use as agencies seek to make datasets that include identifying information
620 available.

621 **2.2. Terminology**

622 While each of the de-identification traditions has developed its own terminology and math-
623 ematical models, they share many underlying goals and concepts. Where terminology
624 differs, this document relies on the terminology developed in previous documents by the
625 U.S. Government and standards organizations.

⁴For information on characterizing the sensitivity of information, see NIST SP 800 Volume I, Revision 1 [119].

626 *De-identification* is a process that is applied to a dataset with the goal of preventing or
627 limiting informational risks to individuals, protected groups, and establishments while still
628 allowing for the production of aggregate statistics.⁵ De-identification takes an *original*
629 *dataset* and produces *de-identified data*.

630 *Re-identification* is the general term for any process that restores the association between a
631 set of de-identified data and the data subject. Re-identification is not the only way that de-
632 identification techniques can fail to protect privacy. Improperly de-identified information
633 can also be used to infer private facts about individuals that were thought to have been
634 protected.

635 *Re-identification risk* is the likelihood that a third party can re-identify data subjects in a
636 de-identified dataset. Re-identification risk is typically a function of the adverse impacts
637 that would arise if the re-identification were to occur and the likelihood of occurrence.
638 Re-identification risk is a specific form of privacy risk.

639 *Redaction* is the removal of information from a document or dataset for legal or security
640 purposes. Also known as *suppression*, redaction is a kind of de-identifying technique that
641 relies on the removal of information. In general, redaction alone is not sufficient to provide
642 formal privacy guarantees, such as differential privacy. Redaction may also reduce the data
643 accuracy of the dataset since the use of selective redaction may result in the introduction of
644 non-ignorable bias.

645 *Anonymization* is a “process that removes the association between the identifying dataset
646 and the data subject” [66]. This term is reserved for de-identification processes that cannot
647 be reversed.

648 Some authors use the terms *de-identification* and *anonymization* interchangeably. In some
649 contexts, the term *anonymization* is used to describe the destruction of a table that maps
650 pseudonyms to real identifiers.⁶ Both of these uses are potentially misleading, as many
651 de-identification procedures can be readily reversed if a dataset is discovered that maps a
652 unique attribute or combination of attributes to identities. For example, a medical dataset
653 may contain a list of names, medical identifiers, the rooms where a patient was seen, the
654 time that the patient was seen, and the results of a medical test. Such a dataset could
655 be de-identified by removing the name and medical identification numbers. However, the
656 dataset of medical test results should not be considered anonymized because the tests can
657 be re-identified if the dataset is joined with a second dataset of room numbers, times, and

⁵ISO/TS 25237:2008 defines de-identification as the “general term for any process of removing the association between a set of identifying data and the data subject.” [66]. This document intentionally adopts a broader definition for de-identification that allows for noise-introducing techniques, such as differential privacy and the creation of synthetic datasets that are based on privacy-preserving models.

⁶For example, “Anonymization is a step subsequent to de-identification that involves destroying all links between the de-identified datasets and the original datasets. The key code that was used to generate the new identification code number from the original is irreversibly destroyed (i.e., destroying the link between the two code numbers)” [127].

658 names. Since it is not possible to know whether such an auxiliary dataset exists, this publi-
659 cation recommends avoiding the word *anonymization* and using the word *de-identification*
660 instead.

661 Because of the inconsistencies in the use and definitions of the word “anonymization,” this
662 document avoids the term except in this section and in the titles of some references. Instead,
663 it uses the term “de-identification” with the understanding that sometimes de-identified
664 information can be re-identified, and sometimes it cannot.⁷

665 *Pseudonymization* is a “particular type of [de-identification]⁸ that both removes the asso-
666 ciation with a data subject and adds an association between a particular set of character-
667 istics relating to the data subject and one or more pseudonyms” [66]. The term *coded*
668 is frequently used in healthcare settings to describe data that has been pseudonymized.
669 Pseudonymization is commonly used so that multiple observations of an individual over
670 time can be matched and so that an individual can be re-identified if there is a policy reason
671 to do so. Although pseudonymous data are typically re-identified by consulting a key that
672 may be highly protected, the existence of the pseudonym identifiers frequently increases
673 the risk of re-identification through other means.

674 Many U.S. Government documents use the phrase *personally identifiable information* (PII)
675 to describe private information that can be linked to an individual [62, 79], although there
676 are a variety of definitions for PII in various laws, regulations, and agency guidance docu-
677 ments. Because of these differing definitions, it is possible to have information that *singles*
678 *out* individuals but that does not meet a specific definition of PII. An added complication
679 is that some documents use the term PII to denote any information that is attributable to
680 individuals or information that is uniquely attributable to a specific individual, while others
681 use the term strictly for data that are directly identifying.

682 This document avoids the term *personally identifiable information*. Instead, it uses the
683 phrases *personal data* or *personal information* to denote information related to individu-
684 als and *identifying information* for “information that can be used to distinguish or trace an
685 individual’s identity, such as their name, social security number, biometric records, etc.,
686 alone, or when combined with other personal or identifying information which is linked or
687 linkable to a specific individual, such as date and place of birth, mother’s maiden name,
688 etc.” [62]. Under this definition, identifying information is personal information, but per-
689 sonal information is not necessarily identifying information.

690 *Non-public personal information* is used to describe personal information that is in a dataset
691 that is not publicly available. Non-public personal information is not necessarily identify-
692 ing.

⁷Thus, where other references (e.g. [104]) might use the term *anonymized file* or *anonymized dataset* to describe a dataset that has been de-identified, this publication will use the terms *de-identified file* and *de-identified dataset* since the term *de-identified* is descriptive while the term *anonymized* is aspirational.

⁸Here, the word *anonymization* in the ISO 25237 definition is replaced with the more accurate and descriptive term *de-identification*.

693 The definition of identifying information above suggests that it is easy – or at least possible
694 – to distinguish personal information from identifying information. Indeed, many tech-
695 niques for de-identification require an expert to make this distinction and protect only the
696 identifying information. However, as understanding of privacy risk develops, it is increas-
697 ingly apparent that *all* information is potentially identifying information.

698 This document envisions a *de-identification process* in which an *original dataset* that con-
699 tains personal information is algorithmically processed to produce *de-identified data*. The
700 result may be a *de-identified dataset*, *aggregate statistics* such as summary tables, or a
701 *synthetic dataset*, in which the data are created by a model. This kind of de-identification
702 is envisioned as a batch process. Alternatively, the de-identification process may be a
703 system that accepts queries and returns responses that do not leak more identifying infor-
704 mation than is allowable by policy. De-identified results may be corrected or updated and
705 re-released on a periodic basis. The accumulated leakage of information from multiple
706 releases may be significant, even if the leakage from a single release is small. Issues that
707 arise from multiple releases are discussed in Section 3.4, “Data-Sharing Models.”

708 *Disclosure* is generally the exposure of data beyond the original collection use case. How-
709 ever, when the goal of de-identification is to protect privacy, disclosure

710 ...relates to inappropriate attribution of information to a data subject, whether
711 an individual or an organization. Disclosure occurs when a specific individual
712 can be associated with a corresponding record(s) in the released dataset with
713 high probability (*identity disclosure*), when an attribute described in a dataset is
714 held by a specific individual, even if the record(s) associated with that individ-
715 ual is (are) not identified (*attribute disclosure*), or when it is possible to make
716 an inference about an individual, even if the individual was not in the dataset
717 prior to de-identification (*inferential disclosure*). [47, emphasis in original]

718 More information about disclosure can be found in Section 3.2.1, “Probability of Re-
719 Identification.”

720 *Disclosure limitation* is a general term for the practice of allowing summary information
721 or queries on data within a dataset to be released without revealing information about spe-
722 cific individuals whose personal information is contained within the dataset. Thus, de-
723 identification is a kind of disclosure limitation technique. Every disclosure limitation pro-
724 cess introduces inaccuracy into the results [14, 11].

725 A primary goal of disclosure limitation is to protect the privacy of individuals while avoid-
726 ing the introduction of *non-ignorable biases* [7] (e.g., bias that might lead a social scientist
727 to come to the wrong conclusion) into the de-identified dataset. One way to measure the
728 amount of bias that has been introduced by the de-identification process is to compare
729 statistics or models generated by analyzing the original dataset with those that are gener-
730 ated by analyzing the de-identified datasets. Such biases introduced by the de-identification

731 process are typically unrelated to any statistical biases that may also exist in the original
732 data.

733 Formal models of privacy can quantify the amount of privacy protection offered by a de-
734 identification process. With methods based on differential privacy, this measurement takes
735 the form of a number called *privacy loss*, which quantifies the additional risk that an ad-
736 versary might learn something new about an individual as a result of a de-identified data
737 release. When a de-identification process is associated with low privacy loss, releasing the
738 data it produces results in little additional risk for individuals in the input dataset. Some
739 formal models, such as differential privacy, allow *composing* the privacy losses of multiple
740 data releases to quantify the *total risk* to individuals of the combined releases, while others
741 – such as *k*-anonymity – do not have this capability.

742 An upper bound on the total acceptable privacy loss of many data releases is often called
743 a *privacy loss budget* or simply a *privacy budget*. This number quantifies the *total* privacy
744 risk to an individual who participates in all of the releases.

745 *Differential privacy* [40] is a model based on a mathematical definition of privacy that con-
746 siders the risk to an individual from the release of a query on a dataset containing their
747 personal information. Statisticians, mathematicians, and other kinds of privacy engineers
748 then develop mathematical algorithms, called mechanisms, that process data in a way that
749 is consistent with the definition. Differential privacy limits both identity and attribute dis-
750 closure by adding non-deterministic noise (random values) to the results of mathematical
751 operations before the results are reported. Unlike *k*-anonymity and other de-identification
752 frameworks, differential privacy is based on information theory and makes no distinction
753 between what is private data and what is not. Differential privacy does not require that val-
754 ues be classified as direct identifiers, quasi-identifiers, and non-identifying values. Instead,
755 differential privacy assumes that *all values* in a record might be identifying and therefore
756 all must be de-identified.

757 Differential privacy’s mathematical definition requires that the result of an analysis of a
758 dataset should be roughly the same with or without the data of any single individual. The
759 definition is usually satisfied by adding random noise to the result of a query, ensuring
760 that the added noise masks the contribution of any individual. The degree of sameness
761 is defined by the parameter ϵ (epsilon). The smaller the parameter ϵ , the more noise is
762 added, and the more difficult it is to distinguish the contribution of a single individual. The
763 result is increased privacy for all individuals – both those in the sample and those in the
764 population from which the sample is drawn who are not present in the dataset. The research
765 literature describes differential privacy being used to solve a variety of tasks, including
766 statistical analysis, machine learning, and data sanitization [38]. Differential privacy can
767 be implemented in an online query system or in a batch mode in which an entire dataset
768 is de-identified at one time. In common usage, the phrase “differential privacy” is used
769 to describe both the formal mathematical framework for evaluating privacy loss and for
770 algorithms that provably provide those privacy guarantees.

771 The use of differential privacy algorithms does not guarantee that privacy will be preserved.
772 Instead, the algorithms guarantee that the amount of privacy risk introduced by data pro-
773 cessing or data release will reside within specific mathematical bounds. It is also important
774 to remember that the impact on privacy risk is limited to reducing the risk of identity and
775 attribute disclosures (see §3.2.1, “Probability of Re-Identification”) and not inferential dis-
776 closure.

777 *K-anonymity* [108, 123] is a framework for quantifying the amount of manipulation re-
778 quired of the quasi-identifiers to achieve a desired level of privacy. The technique is based
779 on the concept of an *equivalence class* – the set of records that have the same values on the
780 quasi-identifiers⁹. A dataset is said to be *k-anonymous* if there are no fewer than *k* match-
781 ing records for every specific combination of quasi-identifiers. For example, if a dataset
782 that has the quasi-identifiers (birth year) and (state) has *k=4* anonymity, then there must
783 be at least four records for every combination of (birth year, state). Subsequent work has
784 refined *k-anonymity* by adding requirements for diversity of the sensitive attributes within
785 each equivalence class (known as *l-diversity* [76]) and requiring that the resulting data be
786 statistically close to the original data (known as *t-closeness* [73]).

787 *K-anonymity* and its subsequent refinements define formal privacy models but come with
788 two important drawbacks. First, they require an expert to determine the set of quasi-
789 identifiers by distinguishing between identifying and non-identifying information. As de-
790 scribed earlier, this task can be difficult or impossible in some contexts. If identifying
791 information is not marked as a quasi-identifier, then the resulting *k-anonymous* dataset will
792 not prevent the re-identification of data subjects. Second, *k-anonymity* and related tech-
793 niques are not compositional – they do not quantify the cumulative privacy loss of multiple
794 data releases, and multiple releases can result in a catastrophic loss of privacy.

795 When data releases containing information about the same individual accumulate, then
796 privacy loss accumulates. This accumulation of privacy loss is not reflected in *k-anonymity*,
797 nor is it reflected in HIPAA privacy rule guidance [136]. Nevertheless, the accumulation
798 is real. In 2003, Dinur and Nissim discovered that it was possible to reconstruct private
799 microdata from a query interface even if the results of each query were systematically
800 infused with small amount of noise [33]. The researchers showed that the amount of
801 noise added to prevent an accurate reconstruction increases as the amount of queries on the
802 dataset increase. If a query interface allows for an unlimited number of queries, no amount
803 of noise is sufficient. Organizations should keep this in mind and try to assess the overall
804 accumulated risk. The discovery in this paper led directly to the invention of differential
805 privacy.

806 Some agencies (notably those that publish data for accountability and enforcement pur-
807 poses) view perturbative Statistical Disclosure Limitation methods (e.g., those that add
808 noise, such as differential privacy) as being inherently unacceptable, since the noise intro-

⁹A quasi-identifier is a variable that can be used to identify an individual through association with other information.

809 duced by the methods can void their ability to be used for accountability. For example, if
810 a school would lose funding if the promotion rate for any class fell below a certain thresh-
811 old, then a method that protects the privacy of students within each class by introducing
812 noise could mask whether the school did or did not make that target. Thus, despite their
813 weaknesses and flaws, program agencies often prefer to use suppression as the preferred
814 protection method for these purposes because the data are either reported as is or sup-
815 pressed, eliminating the uncertainty. Agencies should realize that suppression alone is not
816 sufficient to protect privacy, and if a large enough number of queries is released based on
817 the same confidential dataset, it is frequently possible to reconstruct even data that have
818 been suppressed.

819 Traditional disclosure limitation and k -anonymity start with specific disclosure limitation
820 mechanisms that were designed to hide information while allowing for useful data analysis
821 and attempting to reach the goal of privacy protection. In contrast, differential privacy starts
822 with an information-theoretic definition of privacy and has attempted to evolve mechanisms
823 that produce useful (but privacy-preserving) results. These techniques are currently the
824 subject of academic research, so it is reasonable to expect new techniques to be developed
825 in the coming years that simultaneously increase privacy protection while providing for the
826 high accuracy of resulting de-identified data. Indeed, some authors have shown that the
827 models can be viewed synergistically [114] under some circumstances.

828 Finally, privacy harms are not the only kinds of harms that can result from the release of
829 de-identified data. Analysts working with de-identified data often have no way of knowing
830 how inaccurate their statistical results are due to statistical distortions introduced by the
831 de-identification process. Thus, de-identification operations intended to shield individuals
832 from harm could result in inaccurate research findings. Such research might also cause
833 harm if it is used to support harmful policies.

834 3. Governance and Management of Data De-Identification

835 The decisions and practices regarding the de-identification and release of government data
836 can be integral to the mission and proper functioning of a government agency. As such,
837 these activities should be managed by an agency’s leadership in a way that assures that
838 performance and results that are consistent with the agency’s mission and legal authority.
839 As discussed above, the need for attention arises because of the conflicting goals of data
840 transparency and privacy protection. Although many agencies once assumed that it was
841 relatively straightforward to remove privacy-sensitive data from a dataset so that the re-
842 mainder could be released without restriction, history shows that this is not the case [51,
843 §2.4, §3.6].

844 Given this history, there may be a tendency for government agencies to either over-protect
845 data or to simply avoid its release. Limiting the release of data clearly limits the privacy risk
846 that might result from a data release. However, limiting the release of data also creates costs
847 and risks for other government agencies (which will then not have access to the identified
848 data), external organizations, and society. For example, absent the data release, external
849 organizations will suffer the cost of recollecting the data (if it is possible to do so) or the
850 risk of incorrect decisions that might result from having insufficient information.

851 This section begins with a discussion of why agencies might wish to de-identify data and
852 how agencies should balance the benefits of data release with risks to the data subjects. It
853 then discusses where de-identification fits within the data life cycle. Finally, it discusses
854 options that agencies have for adopting de-identification standards.

855 3.1. Identifying Goals and Intended Uses of De-Identification

856 Before engaging in de-identification, agencies should clearly articulate their goals regard-
857 ing transparency and disclosure limitation in making a data release. They should then
858 develop a written plan that explains how de-identification will be used to accomplish those
859 goals.

860 For example:

- 861 • **Federal Statistical Agencies** collect, process, and publish data for use by researchers,
862 business planners, and other well-established purposes. These agencies are likely to
863 have established standards and methodologies for de-identification. As these agen-
864 cies evaluate new approaches for de-identification, they should document their rati-
865 onale for adopting legacy versus new approaches, evaluate how successful their
866 approaches have been over time, and address inconsistencies between data releases.
- 867 • **Federal Awarding Agencies** are allowed under OMB Circular A-110 to require that
868 institutions of higher education, hospitals, and other non-profit organizations that
869 receive federal grants provide the U.S. Government with “the right to (1) obtain,
870 reproduce, publish or otherwise use the data first produced under an award; and

871 (2) authorize others to receive, reproduce, publish, or otherwise use such data for
872 Federal Purposes” [91, see §36 (c) (1) and (2)]. To realize this policy, awarding
873 agencies can require that awardees establish data management plans for making re-
874 search data publicly available. Such data are used for a variety of purposes, including
875 transparency and reproducibility. In general, research data that contain personal in-
876 formation should be de-identified by the awardee prior to public release. Awarding
877 agencies may establish de-identification standards to ensure the protection of per-
878 sonal information and may consider audits to assure that awardees have performed
879 de-identification in an appropriate manner.

- 880 • **Federal Research Agencies** may wish to make de-identified data available to the
881 public to further the objectives of research transparency and allow others to reproduce
882 and build upon their results. These agencies are generally prohibited from publishing
883 research data that contain personal information, requiring the use of de-identification.
- 884 • **All Federal Agencies** that wish to make administrative or operational data available
885 for transparency, accountability, or program oversight or to enable academic research
886 may wish to employ de-identification to avoid sharing sensitive personally identifi-
887 able information of employees, customers, or others. These agencies may wish to
888 evaluate the effectiveness of simple field suppression, de-identification that involves
889 aggregation, and the creation and release of synthetic data as alternatives for realizing
890 their commitment to open data.

891 3.2. Evaluating Risks that Arise from De-Identified Data Releases

892 Once the purpose of the data release is understood, agencies should identify the risks that
893 might result from the data release. As part of this risk analysis, agencies should specifically
894 evaluate the anticipated negative actions that might result from re-identification, as well as
895 strategies for remediation. NIST provides detailed information on how to conduct risk
896 assessments in NIST SP 800-30 [23].

897 Risk assessments should be based on objective scientific factors and consider the best inter-
898 ests of the individuals in the dataset, the responsibilities of the agency holding the data, and
899 the anticipated benefits to society. The goal of a risk evaluation is not to eliminate risk but
900 to identify which risks can be reduced while still meeting the objectives of the data release
901 and then deciding whether the residual risk is justified by the goals of the data release. An
902 agency decision-making process may choose to accept or reject the risk that might result
903 from a release of de-identified data, but participants in the risk assessment should not be
904 empowered to prevent risk from being documented and discussed. Centralized processes
905 also allow for standardization of the risk assessment and the amount of “acceptable risk”
906 across different programs’ releases.

907 It is difficult to measure re-identification risk in ways that are both general and meaningful.
908 For example, it is possible to measure the similarity between individuals in the dataset

909 under a variety of different parameters and to model how that similarity is impacted when
910 the larger population is considered. However, such calculations may result in different
911 levels of risk for different groups. There may be some individuals in a dataset who would
912 be significantly adversely impacted by re-identification and for whom the likelihood of
913 re-identification might be quite high, but these individuals might represent a tiny fraction
914 of the entire dataset. This represents an important area for research in the field of risk
915 communication.

916 3.2.1. Probability of Re-Identification

917 As discussed in Section 2.2, “Terminology,” the potential impacts on individuals from the
918 release and use of de-identified data include [143]:

919 **Identity disclosures:** Associating a specific individual with the corresponding record(s)
920 in the dataset with high probability. Identity disclosure can result from insufficient
921 de-identification, re-identification by linking, or pseudonym reversal.

922 **Attribute disclosure:** Determining that an attribute described in the dataset is held by a
923 specific individual with high probability, even if the records associated with that indi-
924 vidual are not identified. Attribute disclosure can occur without identity disclosure if
925 the de-identified dataset contains data from a significant number of relatively homo-
926 geneous individuals [51, p.13]. In these cases, traditional de-identification does not
927 protect against attribute disclosure, although differential privacy can. Membership
928 inference is an example of attribute disclosure.

929 **Inferential disclosure:** Being able to make an inference about an individual (typically a
930 member of a group) with high probability, even if the individual was not in the dataset
931 prior to de-identification. “Inferential disclosure is of less concern in most cases
932 as inferences are designed to predict aggregate behavior, not individual attributes,
933 and thus are often poor predictors of individual data values” [60]. Traditional de-
934 identification does not protect against inferential disclosure. Such disclosures can
935 never be eliminated; they can only be controlled.

936 *Re-identification probability*¹⁰ is the estimated probability that an outside party will be able
937 to use information contained in a de-identified dataset to make identity-related inferences
938 about individuals. This outside party was originally termed a *data intruder*, although the
939 terms *adversary* and *attacker* are also used, borrowing from the colorful language of infor-
940 mation security. Different kinds of re-identification probabilities for this data intruder can
941 be calculated.

¹⁰Previous publications described identification probability as “re-identification risk” and used scenarios such as a journalist seeking to discredit a national statistics agency or a prosecutor seeking to find information about a suspect as the bases for probability calculations. That terminology is not presented in this document because of the possible unwanted connotations of those terms and in the interest of bringing the terminology of de-identification into agreement with the terminology used in contemporary risk analysis processes [42].

942 Here are several kinds of probabilities, as well as proposals for new, declarative, self-
943 describing names:

944 **Known inclusion re-identification probability (KIRP)** is the probability of finding the
945 record that matches a specific individual known to be in the sample corresponding to
946 a specific record. KIRP can be expressed as the probability for a specific individual
947 or the probability averaged over the entire dataset (AKIRP).¹¹

948 **Unknown inclusion re-identification probability (UIRP)** is the probability of finding the
949 record that matches a specific individual without first knowing whether the individual
950 is in the dataset. UIRP can be expressed as a probability for an individual record in
951 the dataset averaged over the entire population (AUIRP).¹²

952 **Record matching probability (RMP)** is the probability of finding the record that matches
953 a specific individual chosen from the population. RMP can be expressed as the prob-
954 ability for a specific record (RMP), the probability averaged over the entire dataset
955 (ARMP), or the maximum probability over the entire dataset.

956 **Inclusion probability (IP)** is the probability that a specific individual's presence in the
957 dataset can be inferred.

958 Whether it is necessary to quantitatively estimate these probabilities depends on the specifics
959 of each intended data release. For example, many cities publicly disclose whether taxes
960 have been paid on a property. Given that this information is already a matter of public
961 record, it may not be necessary to consider inclusion probability when a dataset of property
962 taxpayers for a specific dataset is released. Likewise, there may be some attributes in a
963 dataset that are already public and may not need to be protected with disclosure limitation
964 techniques. However, the existence of such attributes may pose a re-identification risk for
965 other information in the dataset or in other de-identified datasets. The fact that information
966 is public may not negate the responsibility of an agency to provide protection for that in-
967 formation, as the aggregation and distribution of information may cause privacy risk that
968 was not otherwise present. Agencies may also be legally prohibited from releasing copies
969 of information that is similar to information that is already in the public domain.

970 Although disclosures are commonly thought to be discrete events involving the release of
971 specific data, such as an individual's name matched to a record, disclosures can result from
972 the release of data that merely changes a data intruder's probabilistic belief. For example,
973 a disclosure might change an intruder's estimate that a specific individual is present in a
974 dataset from a 50% probability to 90%. The intruder still does not *know* if the individual
975 is in the dataset or not (and the individual might not, in fact, be in the dataset), but a

¹¹Some texts refer to KIRP as "prosecutor risk." The scenario is that a prosecutor is looking for records that belong to a specific, named individual.

¹²Some texts refer to UIRP as "journalist risk." The scenario is that a journalist has obtained a de-identified file and is trying to identify one of the data subjects, but the journalist fundamentally does not care *who* is identified.

976 probabilistic disclosure has still occurred because the intruder’s estimate of the individual
977 has been changed by the data release.

978 It may be difficult to estimate specific re-identification probabilities, as the ability to re-
979 identify depends on the original dataset, the de-identification technique, the technical skill
980 of the data intruder, the intruder’s available resources, and the availability of additional
981 data (publicly available or privately held) that can be linked with the de-identified data.
982 It is likely that the true probability of re-identification increases over time as techniques
983 improve and more contextual information becomes available to potential data intruders.
984 Indeed, some researchers have claimed that computing these probabilities “is a fundamen-
985 tally meaningless exercise” because the calculations are based on assumptions that cannot
986 be validated (e.g., the lack of a database that could link specific quasi-identifiers or sensi-
987 tive, non-identifying values to identities) [83].

988 De-identification practitioners have traditionally quantified re-identification probability, in
989 part, based on the skills and abilities of a potential data intruder. Datasets that were thought
990 to have little possibility for exploitation were deemed to have a lower re-identification
991 probability than datasets containing sensitive or otherwise valuable information. Such ap-
992 proaches are not appropriate when attempting to evaluate the re-identification probability
993 of government datasets that will be publicly released.

- 994 • Although a specific de-identified dataset may not be recognized as sensitive, re-
995 identifying that dataset may be an important step in re-identifying another dataset
996 that is sensitive. Alternatively, the data intruder may merely wish to embarrass the
997 government agency. Thus, adversaries may have a strong incentive to re-identify
998 datasets that are seemingly innocuous.
- 999 • Although the public may not generally be skilled in re-identification, many resources
1000 on the internet make it easy to acquire specialized datasets, tools, and experts for
1001 specific re-identification challenges. Family members, friends, colleagues, and oth-
1002 ers may also possess substantial personal knowledge about individuals in the data
1003 that can be used for re-identification.

1004 Instead, de-identification practitioners should assume that de-identified government datasets
1005 could be subjected to sustained, worldwide re-identification attempts, and they should
1006 gauge their de-identification requirements accordingly. Of course, it is unrealistic to as-
1007 sume that all of the world’s resources will be used to attempt to re-identify every publicly
1008 released file. Therefore, de-identification requirements should be gauged using a risk as-
1009 sessment [75]. More information on conducting risk assessments can be found in NIST SP
1010 800-30, *Guide for Conducting Risk Assessments* [23].

1011 Members of vulnerable populations (e.g., prisoners, children, people with disabilities) may
1012 be more susceptible to having their identities disclosed by de-identified data than non-
1013 vulnerable populations because the thing that makes these individuals vulnerable may also
1014 make them stand out in the dataset. Likewise, residents of areas with small populations

1015 may be more susceptible to having their identities disclosed than residents of urban areas.
1016 Individuals with multiple traits will generally be more identifiable if the individual's loca-
1017 tion is geographically restricted. For example, data belonging to a person who is labeled as
1018 a pregnant, unemployed female veteran will be more identifiable if restricted to Baltimore
1019 County, Maryland, than to all of North America.

1020 If agencies determine that the potential for harm is large in a contemplated data release, one
1021 way to manage the risk is by increasing the level of de-identification and accepting a lower
1022 data accuracy level. Other options include data controls, such as restricting the availability
1023 of data to qualified researchers in a data enclave.

1024 **3.2.2. Adverse Impacts of Re-Identification**

1025 As part of a risk analysis, agencies should attempt to enumerate specific kinds of adverse
1026 impacts that can result from the re-identification of de-identified information. These can
1027 include potential impacts on individuals, the agency, and society.

1028 Potential adverse impacts on individuals include:

- 1029 • Increased availability of personal information that leads to an increased risk of fraud,
1030 identity theft, discrimination, or abuse
- 1031 • Increased availability of an individual's location that puts that person at risk for bur-
1032 glary, property crime, assault, or other kinds of violence
- 1033 • Increased availability of an individual's non-public personal information that causes
1034 psychological harm by exposing potentially embarrassing information or information
1035 that the individual may not otherwise choose to reveal to the public or to family
1036 members and that potential affects opportunities in the economic marketplace (e.g.,
1037 employment, housing, college admission)

1038 Potential adverse impacts on agencies include:

- 1039 • Mandatory reporting under breach reporting laws, regulations, or policies
- 1040 • Embarrassment or reputational damage
- 1041 • Harm to agency operations if some aspect of those operations required that the de-
1042 identified data remain confidential (e.g., an agency that is forced to discontinue a
1043 scientific experiment because the data release may have biased the study participants)
- 1044 • Financial impacts that result from the harm to the individuals (e.g., lawsuits)
- 1045 • Civil or criminal sanctions against employees or contractors that result from a data
1046 release contrary to U.S. law

1047 Potential adverse impacts on society include:

- 1048 • Undermining the reputation of researchers in general and the willingness of the pub-
1049 lic to support/tolerate research and provide accurate information to government agen-
1050 cies and researchers
- 1051 • Engendering a lack of trust in government – individuals may stop consenting to the
1052 use of their data, may stop providing data, or may provide false data
- 1053 • Damaging the practice of using de-identified information – de-identification is an
1054 important tool for promoting research and accountability, and poorly executed de-
1055 identification efforts may negatively impact the public’s view of this technique and
1056 limit its use

1057 One way to calculate an upper bound on impact to an individual or the agency is to es-
1058 timate the impact that would result from the inadvertent release of the original dataset.
1059 This approach will not calculate the upper bound on the societal impact, however, since
1060 that impact includes reputational damage to the practice of de-identification itself. That is,
1061 every time data are compromised because of a poorly executed de-identification effort, it
1062 becomes harder to justify the use of de-identification in future data releases.

1063 As part of a risk analysis process, organizations should enumerate specific measures that
1064 they will take to minimize the risk of successful re-identification. Organizations may wish
1065 to consider both the actual risk and the perceived risk to those in the dataset and in the
1066 broader community.

1067 As part of the risk assessment, an organization may determine that there is no way to
1068 achieve the de-identification goal in terms of data accuracy and identifiability. In these
1069 cases, the organization will need to decide whether it should adopt additional measures to
1070 protect privacy (e.g., administrative controls or data use agreements), accept a higher level
1071 of risk, or choose not to proceed with the project.

1072 **3.2.3. Impacts Other Than Re-Identification**

1073 The use of de-identified data can lead to adverse impacts other than those that might re-
1074 sult from re-identification. Risk assessments that evaluate the risks of re-identification can
1075 address these other risks as well. Such risks might include:

- 1076 • The risk of excessive inferential disclosures
- 1077 • The risk that the de-identification process might introduce bias or inaccuracies into
1078 the dataset that result in incorrect decisions¹³

¹³For example, a personalized warfarin dosing model created with data that had been modified in a manner consistent with the differential privacy de-identification model produced higher mortality rates in simulation than a model created from unaltered data [49]. Educational data de-identified with the *k*-anonymity model can also result in the introduction of bias that leads to spurious results [14, 125].

- 1079 • The risk that releasing a de-identified dataset might reveal non-public information
1080 about an agency’s policies or practices

1081 It is preferable to use de-identification processes that include assessments of accuracy (e.g.,
1082 confidence intervals) with respect to the bias and precision of statistical properties of the
1083 data. Where it does not provide information that may aid data intruders, it is also useful to
1084 reveal the de-identification process itself so that analysts can understand any potential in-
1085 accuracies that might be introduced by the de-identification. This is consistent with *Kerck-*
1086 *hoffs’ principle* [67], a widely accepted system design principle that holds that the security
1087 of a system should not rely on the secrecy of the methods that it employs.

1088 3.2.4. Remediation

1089 As part of a risk analysis process, agencies should attempt to enumerate techniques that
1090 could be used to mitigate or remediate harm that would result from a successful re-identification
1091 of de-identified information. Remediation could include victim education, the procurement
1092 of monitoring or security services, the issuance of new identifiers, or other measures.

1093 3.3. Data Life Cycle

1094 The *NIST Big Data Interoperability Framework* defines the data life cycle as “the set of
1095 processes in an application that transform raw data into actionable knowledge” [85]. The
1096 data life cycle can be used in the de-identification process to help analyze the expected
1097 benefits, intended uses, privacy threats, and vulnerabilities of de-identified data. As such,
1098 the data life cycle concept can be used to select appropriate privacy controls based on a
1099 reasoned analysis of the threats. For example, privacy-by-design concepts [22] can be
1100 employed to decrease the number of identifiers collected, minimizing requirements for de-
1101 identification prior to data release. The data life cycle can also be used to design a tiered
1102 access mechanism based on this analysis [12].

1103 Several data life cycles have been proposed, but none are widely accepted as a standard.

1104 Michener et al. [80] (Figure 1) describe the data life cycle as a true cycle:

1105 → Assure → Describe → Deposit → Preserve → Discover → Integrate → Analyze →
1106 Collect

1107 Stobierski [120] also describes the data life cycle as a cycle with different steps:

1108 Generation → Collection → Processing → Storage → Management → Analysis →
1109 Visualization → Interpretation → Generation

1110 De-identification does not fit into a circular data life cycle model, as the data owner typ-
1111 ically retains access to the identified data. However, if the organization employs de-
1112 identification, it could be performed during Collect or between Collect and Assure if iden-
1113 tified data were collected but the identifying information was not actually needed. Alter-

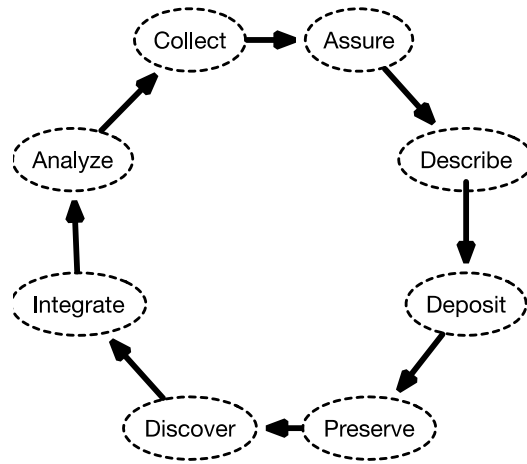


Fig. 1. The data life cycle as described by Michener et al. [80]

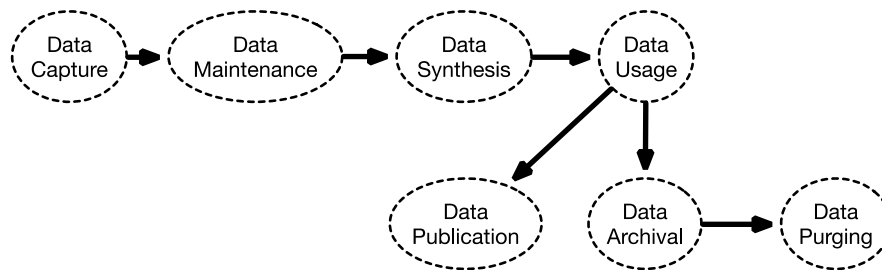


Fig. 2. Chisholm's view of the data life cycle is a linear process with a branching point after data usage [25]

1114 natively, de-identification could be applied after Describe and prior to Deposit to avoid
1115 archiving identifying information.

1116 Chisholm and others [25] (Figure 2) describe the data life cycle as a linear process with a
1117 fork for data publication:

1118 Data Capture → Data Maintenance → Data Synthesis → Data Usage →
1119 {Data Publication & Data Archival → Data Purging}

1120 Using this formulation, de-identification can take place either during Data Capture or fol-
1121 lowing Data Usage. However, agencies should consider data release requirements from
1122 the very beginning of the planning process for each new data collection. By knowing in
1123 advance how they intend to publish and for what purposes and by having a plan for how
1124 disclosure limitation will be applied, agencies can tailor information collection accordingly.

1125 For example, if specific identifiers are not needed for maintenance, synthesis, and usage,
1126 then those identifiers should not be collected. If fully identified data are needed within the

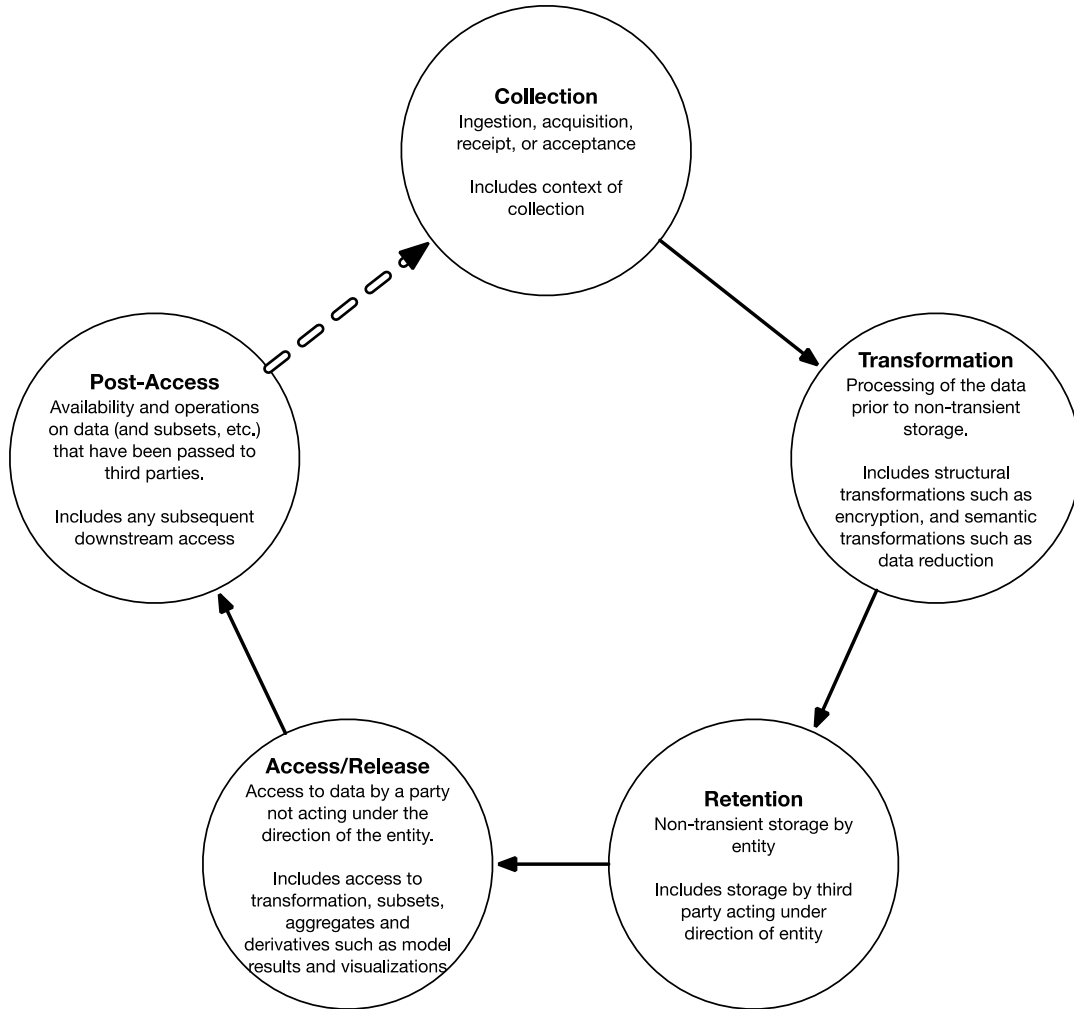


Fig. 3. Altman’s “modern approach to privacy-aware government data releases” [75]

1127 organization, the identifying information can be removed prior to the data being published,
1128 shared, or archived. Applying de-identification throughout the data life cycle minimizes
1129 privacy risk and significantly eases the process of public release. However, agencies should
1130 be cognizant of the potential loss of future utility if identifiers are permanently removed.
1131 For this reason, agencies may wish to retain an identified dataset or data linking informa-
1132 tion, as it may be difficult to predict future needs.

1133 Altman et al. [75] (Figures 3 and 4) propose a “modern approach to privacy-aware gov-
1134 ernment data releases” that incorporates progressive levels of de-identification as well as
1135 different kinds of access and administrative controls in line with the sensitivity of the data.

1136 Agencies that perform de-identification should document that:

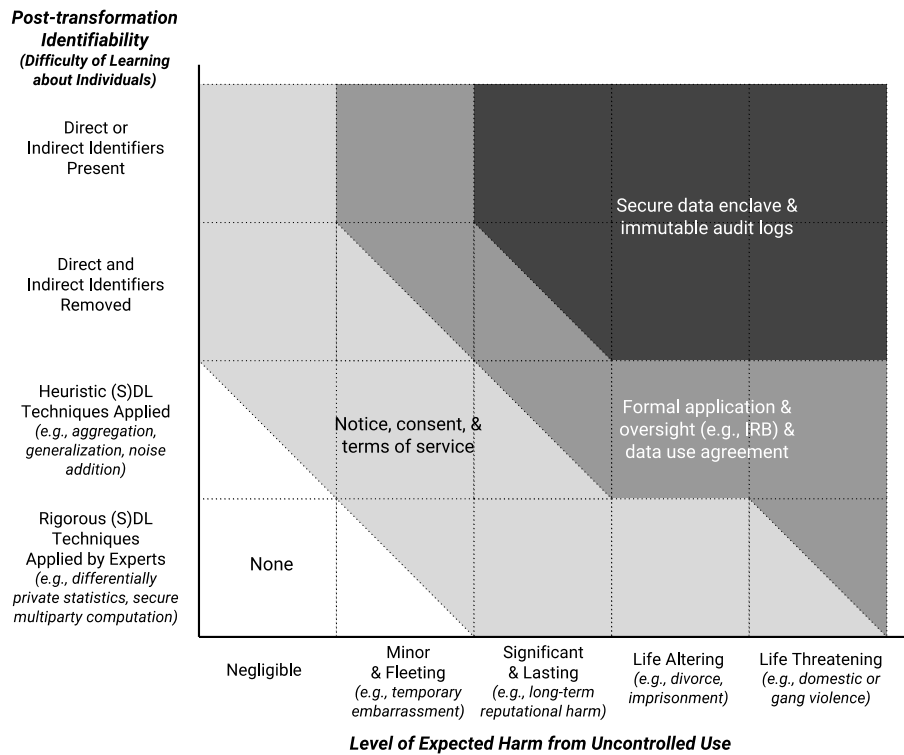


Fig. 4. Altman’s conceptual diagram of the relationship between post-transformation identifiability, level of expected harm, and suitability of selected privacy controls for a data release [75]

- 1137 • The techniques used to perform the de-identification are theoretically sound and gen-
1138 erally accepted.¹⁴
 - 1139 • The software used to perform the de-identification is reliable for the intended task.
 - 1140 • The individuals who performed the de-identification were suitably qualified.
 - 1141 • The tests that were used to evaluate the effectiveness of the de-identification were
1142 validated for that purpose.
 - 1143 • Ongoing monitoring is in place to ensure the continued effectiveness of the de-
1144 identification strategy.
- 1145 No matter where de-identification is applied in the data life cycle, agencies should docu-
1146 ment the answers to the following questions for each de-identified dataset:
- 1147 • Are direct identifiers collected with the dataset?
 - 1148 • Even if direct identifiers are not collected, is it still possible to identify the data
1149 subjects through the presence of quasi-identifiers?
 - 1150 • Where in the data life cycle is de-identification performed? Is it performed in only
1151 one place or in multiple places?
 - 1152 • Is the original dataset retained after de-identification?
 - 1153 • Is there a key or map retained so that specific data elements can be re-identified later?
 - 1154 • How are decisions made regarding de-identification and re-identification?
 - 1155 • Are there specific datasets that can be used to re-identify the de-identified data? If so,
1156 what controls are in place to prevent intentional or unintentional re-identification?
 - 1157 • Is it a problem if some records in a dataset are re-identified?

¹⁴To determine that a technique is theoretically sound and generally accepted, agencies that wish to adopt guidance that mirrors the language that the HHS November 26, 2012 *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule* [136]. uses in its discussion of the Privacy Rule’s “expert determination method,” which states on page 7:

“A covered entity may determine that health information is not individually identifiable health information only if:

- (1) A person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable:
 - (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and
 - (ii) Documents the methods and results of the analysis that justify such determination;”

- 1158 • Is there a mechanism that will inform the de-identifying agency if there is an attempt
1159 to re-identify the de-identified dataset? Is there a mechanism that will inform the
1160 agency if the attempt is successful?

1161 3.4. Data-Sharing Models

1162 Agencies should decide on the data-sharing model that will be used to make the data avail-
1163 able outside of the agency after the data have been de-identified [51, p.14]. Specific models
1164 combine *security* and *privacy* techniques to reduce privacy risks to individuals. *Security*
1165 refers to techniques that limit *who* can view the data. Encryption is an example of a secu-
1166 rity technique – it allows only the party holding the encryption key to view the data. *Pri-*
1167 *vac*y refers to techniques that limit *what information* the data contains. The two concepts
1168 can be considered orthogonally. In practice, however, *who* has access to the data makes
1169 a significant difference in the expected risk of disclosure and therefore influences the ex-
1170 tent to which privacy techniques must be used to limit the presence of sensitive personally
1171 identifiable information in the data.

1172 A number of possible models exist at different points in the spectrum of security and pri-
1173 vacy protections. Figure 4 summarizes this spectrum: its x-axis describes various privacy
1174 techniques that can limit the informational content of the data; its y-axis describes how
1175 much harm would occur if the underlying information were disclosed; and the regions of
1176 the graph are labeled with suggested security techniques. Some common combinations of
1177 security and privacy techniques include:

1178 **The Release and Forget Model [94].** The de-identified data may be released to the pub-
1179 lic, typically by being published on the internet. It can be difficult or impossible for
1180 an organization to recall the data once released in this fashion and may limit infor-
1181 mation for future releases.

1182 **The Data Use Agreement (DUA) Model.** The de-identified data may be made available
1183 under a legally binding data use agreement that details what can and cannot be done
1184 with the data. Typically, data use agreements may prohibit attempted re-identification,
1185 linking to other data, and redistribution of the data without a similarly binding DUA.
1186 A DUA will typically be negotiated between the data holder and qualified researchers
1187 (the “qualified investigator model” [44]) or members of the general public (e.g., cit-
1188 izen scientists or the media), although they may be simply posted on the internet
1189 with a click-through license agreement that must be agreed to before the data can be
1190 downloaded (the “click-through model” [44]).

1191 **The Synthetic Data with Verification Model.** Statistical disclosure limitation techniques
1192 are applied to the original dataset and used to create a synthetic dataset that contains
1193 many of the aspects of the original dataset but does not contain disclosing infor-
1194 mation. The synthetic dataset is released, either publicly or to vetted researchers.
1195 The synthetic dataset can then be used as a proxy for the original dataset, and if

1196 constructed well, the results of statistical analyses should be similar. If used in con-
1197 junction with an enclave model as below, researchers may use the synthetic dataset
1198 to develop queries and/or analytic software. These queries and/or software can then
1199 be taken to the enclave or provided to the agency and be applied on the original data.

1200 **The Enclave Model [44, 87, 113].** The de-identified data may be kept in a segregated en-
1201clave that restricts the export of the original data and instead accepts queries from
1202qualified researchers, runs the queries on the de-identified data, and responds with
1203results. Enclaves can be physical or virtual and can operate under a variety of dif-
1204ferent models. For example, vetted researchers may travel to the enclave to perform
1205their research, as is done with the Federal Statistical Research Data Centers operated
1206by the U.S. Census Bureau. Enclaves may be used to implement the verification step
1207of the Synthetic Data with Verification Model. Queries made in the enclave model
1208may be vetted automatically or manually (e.g., by the DRB). Vetting can try to screen
1209for queries that might violate privacy or are inconsistent with the stated purpose of
1210the research.

1211 Sharing models should consider the possibility of multiple or periodic releases. Just as
1212repeated queries to the same dataset may leak personal data from the dataset, repeated de-
1213identified releases (whether from the same dataset or from different datasets containing
1214some of the same individuals) by an agency may result in compromising the privacy of in-
1215dividuals unless each subsequent release is viewed in light of the previous release. Even if
1216a contemplated release of a de-identified dataset does not directly reveal identifying infor-
1217mation, federal agencies should ensure that the release – combined with previous releases
1218– will also not reveal identifying information [137].

1219 Instead of sharing an entire dataset, the data owner may choose to release a sample. If only
1220a sample is released, the probability of re-identification decreases because a data intruder
1221will not know if a specific individual from the data universe is present in the de-identified
1222dataset [43]. However, releasing only a sample may decrease the statistical power of tests
1223on the data, may cause users to draw incorrect inferences if proper statistical sampling
1224methods are not used, and may not align with agency goals regarding transparency and
1225accountability.

1226 **3.5. The Five Safes**

1227 Agencies that make data available to outsiders should use a repeatable methodology for
1228evaluating the terms under which that data will be made available. The Five Safes [31] is
1229such a framework.

1230 The Five Safes was created in the United Kingdom to assist a national statistical agency in
1231evaluating proposed collaborative projects with the larger research community. The frame-
1232work is designed to assist in “designing, describing and evaluating” data access systems.
1233 Here, the term “data access system” is viewed broadly as any mechanism that allows out-

1234 siders to gain access to the agency’s confidential data. That is, a data access system might
1235 include setting up an enclave for academic researchers who undergo extensive background
1236 checks, but it also includes publishing data on the internet.

1237 The Five Safes framework gets its name from the use of five categories (called “risk” or
1238 “access” dimensions) that are used in the evaluation. They are:

- 1239 1. **Safe projects** Is this use of the data appropriate?
- 1240 2. **Safe people** Can the researchers be trusted to use it in an appropriate manner?
- 1241 3. **Safe data** Is there a disclosure risk in the data itself?
- 1242 4. **Safe settings** Does the access facility limit unauthorized use?
- 1243 5. **Safe outputs** Are the statistical results non-disclosive?

1244 Each of these dimensions is independent. That is, the legal, moral, and ethical review of
1245 each dimension is independent of the others. In practice, this might mean that the project
1246 is safe (the proposed use of the data is appropriate), the people are safe (the researchers are
1247 noted academics with respected histories of collaborative work), the data are safe (there is
1248 no disclosure risk in the data), and the output is safe (it will not disclose personal infor-
1249 mation). However, because the setting is not safe (perhaps the facility has poor internal
1250 security), the project should not go forward. In this example, the Five Safes framework
1251 would provide a decision-maker with the tools to separate each of these dimensions and
1252 resolve the problems so that the project could proceed.

1253 One of the positive aspects of the Five Safes framework is that it forces data controllers
1254 to consider many different aspects of data release when evaluating data access proposals.
1255 Frequently, the authors write, it is common for data owners to “focus on one, and only
1256 one, particular issue (such as the legal framework surrounding access to their data or IT
1257 solutions).” With the Five Safes, people who may be specialists in one area are forced to
1258 consider (or to explicitly not consider) aspects of privacy protection with which they may
1259 not be familiar and might otherwise overlook.

1260 The Five Safes framework can be used as a tool for designing access systems, for evaluating
1261 existing systems, for communication, and for training. Agencies should consider using a
1262 framework such as The Five Safes for organizing risk analyses of data release efforts.

1263 3.6. Disclosure Review Boards

1264 Disclosure Review Boards (DRBs), also known as Data Release Boards, are administrative
1265 bodies created within an organization that are charged with ensuring that intended dis-
1266 closures meet the policy and procedural requirements of that organization. DRBs should
1267 be governed by a written *mission statement* and *charter* (or equivalent document) that
1268 are ideally approved by the same mechanisms that the organization uses to approve other
1269 organization-wide policies.

1270 The DRB should have a mission statement that guides its activities. For example, the U.S.
1271 Department of Education’s DRB has the mission statement:

1272 The Mission of the Department of Education Disclosure Review Board (ED-
1273 DRB) is to review proposed data releases by the Department’s principal offices
1274 (POs) through a collaborate technical assistance, aiding the Department to re-
1275 lease as much useful data as possible, while protecting the privacy of individ-
1276 uals and the confidentiality of their data, as required by law. [41]

1277 The DRB charter specifies the mechanics of how the mission is implemented. A formal,
1278 written charter promotes transparency in the decision-making process and ensures consis-
1279 tency in the applications of its policies.

1280 Most DRBs will be established to weigh the interests of data release against those of in-
1281 dividual privacy protection. However, a DRB may also be chartered to consider *group*
1282 *harms* [51, p.13] that can result from the release of a dataset. Such harms go beyond the
1283 harm to the privacy interests of a specific individual.

1284 The DRB charter should frame the DRB’s responsibilities in reference to existing orga-
1285 nizational policies, regulations, and laws. Some agencies may balance these concerns by
1286 employing data use models other than de-identification (e.g., by establishing data enclaves
1287 where a limited number of vetted researchers can access sensitive datasets in a way that
1288 provides data value while minimizing the possibility for harm or by authorizing the use
1289 of secure multi-party computation, homomorphic encryption, or other Privacy Preserving
1290 Data Analytics to compute various statistics). In those agencies, a DRB would be empow-
1291 ered to approve the use of such mechanisms.

1292 Certain agencies may engage in data disclosure on a routine basis (such as research and
1293 evaluation agencies), in which case it may be beneficial for the DRB to establish policies
1294 and procedures for de-identification rather than being responsible for every review. In
1295 these cases, the DRB charter should clearly specify how the group will provide oversight
1296 and ensure organizational accountability to the agreed-upon policies.

1297 The DRB charter should specify the DRB’s composition. To be effective, the DRB should
1298 include representatives from multiple groups and experts in both technology and privacy
1299 policy. Specifically, DRBs may wish to have as members:

- 1300 • Individuals who represent the interests of potential users (such individuals need not
1301 come from outside of the organization)
- 1302 • Representation from among the public, specifically from groups represented in the
1303 datasets if they have a limited scope
- 1304 • Representation from the organization’s leadership team, such as a representation of
1305 the Senior Agency Official for Privacy [4, Appendix II, section 4] (such representa-
1306 tion helps to establish the DRB’s credibility with the rest of the organization)

- 1307 • A representative of the organization’s senior privacy official
- 1308 • Subject matter experts
- 1309 • Outside experts

1310 The charter should establish rules for ensuring a quorum and specify whether members can
1311 designate alternates on a standing or meeting-by-meeting basis. The DRB should specify
1312 the mechanism by which members are nominated and approved, their tenure, conditions
1313 for removal, and removal procedures.¹⁵

1314 The charter should set policy expectations for record keeping and reporting, including
1315 whether records and reports are considered public or restricted. For example, the char-
1316 ter could specify that a DRB issue an annual report with a list of every dataset that was
1317 approved for release. The charter should indicate whether it is possible to exclude sensitive
1318 decisions from these reporting requirements and the mechanism for doing so. Ideally, the
1319 charter should be a public document to promote transparency.

1320 To meet its requirement of evaluating data releases, the DRB should require that writ-
1321 ten applications be submitted to the DRB that specify the nature of the dataset, the de-
1322 identification methodology, and the result. An application may require that the proposer
1323 present the re-identification risk, the risk to individuals if the dataset is re-identified, and
1324 a proposed plan for detecting and mitigating successful re-identification. In addition, the
1325 DRB should require that when individuals are informed that their information will be de-
1326 identified, they also be informed that privacy risks may remain despite de-identification.

1327 The DRB should keep accurate records of its request memos, their associated documen-
1328 tation, the DRB decision, and the actual files released. These records should be appropri-
1329 ately archived and curated so that they can be recovered. In the case of large data releases,
1330 the definitive version of the released data should be curated using an externally validated
1331 procedure, such as a recorded cryptographic hash value or signature, and a digital object
1332 identifier (DOI) [64].

1333 DRBs may wish to institute a two-step process in which the applicant first proposes and
1334 receives approval for a specific de-identification process that will be applied to a specific
1335 dataset and then submits and receives approval for the release of the dataset that has been
1336 de-identified according to the proposal. However, because it is theoretically impossible
1337 to predict the results of applying an arbitrary process to an arbitrary dataset [26, 129],
1338 the DRB should be empowered to reject a proposed release of a dataset even if it has
1339 been de-identified in accordance with an approved procedure because performing the de-
1340 identification may demonstrate that the procedure was insufficient to protect privacy. The

¹⁵For example, in 2022, the Census Bureau’s DRB had 12 voting members: two technical co-chairs, a repre-
sentative from the Policy Coordination Office, a representative from the Associate Director for Communica-
tions, two representatives from the Center for Enterprise Dissemination-Disclosure Avoidance (CED-DA),
two representatives from the Economic Programs Directorate, two representatives from the Demographic
Programs Directorate, and two representatives from the Decennial Programs Directorate [24].

1341 DRB should be able to delegate the responsibility of reviewing the de-identified dataset,
1342 but such responsibility should not be delegated to the individual or group that performed
1343 the de-identification.

1344 The DRB charter should specify whether the DRB needs to approve each data release by
1345 the organization or if it may grant blanket approval for all data of a specific type that is de-
1346 identified according to a specific methodology. The charter should specify the duration of
1347 the approval. Given advances in the science and technology of de-identification, it is inad-
1348 visable that a Board be empowered to grant release authority for an indefinite or unlimited
1349 amount of time.

1350 In most cases, a single privacy protection methodology will be insufficient to protect the
1351 varied datasets that an agency may wish to release. That is, different techniques might best
1352 optimize the trade-off between re-identification risk and data usability, depending on the
1353 specifics of each kind of dataset. Nevertheless, the DRB may wish to develop guidance, rec-
1354 ommendations, and training materials regarding specific de-identification techniques that
1355 are to be used. Agencies that standardize on a small number of de-identification techniques
1356 will gain familiarity with these techniques and are likely to have results with a higher level
1357 of consistency and success than those that have no such guidance or standardization.

1358 Although it is envisioned that DRBs will work in a cooperative, collaborative, and conge-
1359 nial manner with those inside an agency seeking to release de-identified data, there will
1360 at times be a disagreement of opinion. For this reason, the DRB's charter should state
1361 whether the DRB has the final say over disclosure matters or if the DRB's decisions can be
1362 overruled, by whom, and by what procedure. For example, an agency might give the DRB
1363 final say over disclosure matters but allow the agency's leadership to replace members of
1364 the DRB as necessary. Alternatively, the DRB's rulings might merely be advisory, with all
1365 data releases being individually approved by agency leadership or its delegates.¹⁶

1366 Finally, agencies should decide whether the DRB charter will include any kind of perfor-
1367 mance timetables or be bound by a service-level agreement (SLA) that defines a level of
1368 service to which the DRB commits.

1369 The key elements of a Disclosure Review Board include:

- 1370 • A written mission statement and charter
- 1371 • Members represent different groups within the organization, including leadership
- 1372 • The Board receives written applications to release de-identified data
- 1373 • The Board reviews *both* the proposed methodology *and* the results of applying the
1374 methodology

¹⁶At the Census Bureau, “staff members [who] are not satisfied with the DRB’s decision . . . may appeal to a steering committee consisting of several Census Bureau Associate Directors. Thus far, there have been few appeals, and the Steering Committee has never reversed a decision made by the Board” [130, p.35].

- 1375 • Applications should identify the risks associated with data release, including re-
1376 identification probability, potentially adverse events that would result if individuals
1377 are re-identified, and a mitigation strategy if re-identification takes place
- 1378 • Approvals may be valid for multiple releases but should not be valid indefinitely
- 1379 • Reliable records management for applications, approvals, and released data
- 1380 • Mechanisms for dispute resolution
- 1381 • Timetable or service-level agreement (SLA)
- 1382 • Legal and technical understanding of privacy

1383 Example outputs of a DRB include specifying access methods for different kinds of data
1384 releases, establishing acceptable levels of re-identification risk, and maintaining detailed
1385 records of previous data releases that ideally include the dataset that was released and the
1386 privacy-preserving methodology that was employed.

1387 There is some similarity between DRBs as envisioned here and the Institutional Review
1388 Board (IRBs) system created by the Common Rule¹⁷ for regulating human subject research
1389 in the United States. However, there are also important differences:

- 1390 • While the purpose of IRBs is to protect human subjects involved in human subject
1391 research, DRBs are charged with protecting data subjects, institutions, and – poten-
1392 tially – society as a whole.
- 1393 • Whereas IRBs are required to have “at least one member whose primary concerns
1394 are in nonscientific areas” and “at least one member who is not otherwise affiliated
1395 with the institution and who is not part of the immediate family of a person who is
1396 affiliated with the institution,” there does not appear to be a requirement for such
1397 members on a DRB.
- 1398 • Whereas IRBs give approval for research and then typically receive reports only dur-
1399 ing an annual review or when a research project terminates, DRBs may be involved
1400 at multiple points during the process.
- 1401 • Whereas approval of an IRB is required before research with human subjects can
1402 commence, DRBs are typically involved after research has taken place and prior to
1403 data or other research findings being released.
- 1404 • Whereas service on an IRB requires knowledge of the Common Rule and an under-
1405 standing of ethics, service on a DRB requires knowledge of statistics, computation,
1406 public policy, and some familiarity with the data being considered for release.

¹⁷The Federal Policy for the Protection of Human Subjects or the “Common Rule” was published in 1991 and codified in separate regulations by 15 federal departments and agencies. The Revised Common Rule was published in the Federal Register (FR) on January 19, 2017, and was amended to delay the effective and compliance dates on January 22, 2018, and June 19, 2018 [135].

1407 **3.7. De-Identification Standards**

1408 Agencies can rely on de-identification standards to provide standardized terminology, pro-
1409 cedures, and performance criteria for de-identification efforts. Agencies can adopt existing
1410 de-identification standards or create their own. De-identification standards can be prescrip-
1411 tive or performance-based.

1412 **3.7.1. Benefits of Standards**

1413 De-identification standards assist agencies with the process of de-identifying data prior to
1414 public release. Without standards, data owners may be unwilling to share data, as they may
1415 be unable to assess whether a procedure for de-identifying data is sufficient to minimize
1416 privacy risk.

1417 Standards can increase the availability of individuals with appropriate training by iden-
1418 tifying a specific body of knowledge and practice that training should address. Absent
1419 standards, agencies may forego opportunities to share data. De-identification standards can
1420 help practitioners develop a community, as well as certification and accreditation processes.

1421 Standards decrease uncertainty and provide data owners and custodians with best practices
1422 to follow. Courts can consider standards as acceptable practices that should generally be
1423 followed. In the event of litigation, an agency can point to the standard and say that it
1424 followed good data practice.

1425 **3.7.2. Prescriptive De-Identification Standards**

1426 A prescriptive de-identification standard specifies an algorithmic procedure that – if fol-
1427 lowed – results in data that are de-identified to an established benchmark.

1428 The “Safe Harbor” method of the HIPAA Privacy Rule [3] is an example of a prescriptive
1429 de-identification standard. The intent of the Safe Harbor method is to “provide covered en-
1430 tities with a simple method to determine if the information is adequately de-identified” [89].
1431 It does this by specifying that health information is considered to be de-identified through
1432 the removal of 18 kinds of identifiers and the assurance that the entity does not have *actual*
1433 *knowledge* that the remaining information can be used to identify an individual who is the
1434 subject of the information. Once de-identified, the dataset is no longer subject to HIPAA
1435 privacy, security, and breach notification regulations. Nevertheless, “a covered entity may
1436 require the recipient of de-identified information to enter into a data use agreement to ac-
1437 cess files with known disclosure risk” [89].

1438 The Privacy Rule states that a covered entity that employs the Safe Harbor method must
1439 have no “actual knowledge” that the information – once de-identified – could still be used
1440 to re-identify individuals. However, covered entities are not obligated to employ experts
1441 or mount re-identification attacks against datasets to verify that the use of the Safe Harbor
1442 method has in fact resulted in data that cannot be re-identified.

1443 Prescriptive standards have the advantage of being relatively easy for users to follow, but
1444 developing, testing, and validating such standards can be burdensome. Because prescrip-
1445 tive de-identification standards do not depend on the particulars of a specific case, there
1446 is a tendency for them to be more conservative than is necessary, resulting in an unneces-
1447 sary decrease in data for corresponding levels of risk. Even so, there is no assurance that
1448 following a prescriptive standard actually produces the intended outcome.

1449 Agencies that create prescriptive de-identification standards should ensure that data de-
1450 identified according to the standards have a sufficiently small risk of being re-identified
1451 consistent with the intended level of privacy protection. Such assurances frequently can-
1452 not be made unless formal privacy techniques, such as *differential privacy*, are employed.
1453 However, agencies may determine that public policy goals furthered by having an easy-
1454 to-use prescriptive standard outweighs the risk of a standard that does not have provable
1455 privacy guarantees.

1456 Prescriptive de-identification standards carry the risk that the standard may not sufficiently
1457 de-identify to avoid the risk of re-identification, especially as methodology advances and
1458 more data sources become available.

1459 A second risk when adopting prescriptive standards is that different agencies (or govern-
1460 ments) may adopt inconsistent rules. In such a case, information that is legally de-identified
1461 for one purpose or in one jurisdiction may not be legally de-identified in another.

1462 **3.7.3. Performance-Based De-Identification Standards**

1463 Performance-based de-identification standards specify the properties that de-identified data
1464 must have. For example, under the “Expert Determination” method of the HIPAA Privacy
1465 Rule, a technique for de-identifying data is sufficient if an appropriate expert applying
1466 generally accepted statistical and scientific principles and methods “determines that the
1467 risk is very small that the information could be used, alone or in combination with other
1468 reasonably available information, by an anticipated recipient to identify an individual who
1469 is a subject of the information” [89]. The rule requires that experts document their methods
1470 and the results of their analyses.

1471 Performance-based standards have the advantage of allowing users many different ways to
1472 solve a problem by leaving room for innovation. Another advantage is that they can require
1473 the desired outcome rather than specifying an aspirational mechanism.

1474 Performance-based standards should be sufficiently detailed to perform in a manner that is
1475 reliable and repeatable. For example, standards that call for the use of experts can specify
1476 how an expert’s expertise should be determined. Standards that call for the reduction of
1477 risk to an acceptable level should provide a procedure for determining that level.

1478 **3.8. Education, Training, and Research**

1479 De-identifying data in a manner that preserves privacy can be a complex mathematical,
1480 statistical, administrative, and data-driven process. Frequently, the opportunities for iden-
1481 tity disclosure will vary from dataset to dataset. Privacy-protecting mechanisms developed
1482 for one dataset may not be appropriate for others. For these reasons, agencies that engage
1483 in de-identification should ensure that their workers have adequate education and training
1484 in the subject domain. Agencies may wish to establish education or certification require-
1485 ments for those who work directly with the datasets or to adopt industry standards such
1486 as the HITrust De-Identification Framework [77]. Because de-identification techniques are
1487 modality-dependent, agencies using de-identification may need to institute research efforts
1488 to develop and test appropriate data release methodologies.

1489 **3.9. Defense in Depth**

1490 In addition to de-identification, there are other technologies and methodologies that can
1491 secure sensitive data. Many of these approaches can complement de-identification and
1492 further reduce privacy risk to data subjects. Combining techniques is an example of *defense*
1493 *in depth* and should be considered whenever possible.

1494 **3.9.1. Encryption and Access Control**

1495 Encrypting sensitive data *at rest* can prevent attackers from obtaining the data directly
1496 (e.g., by compromising the server that stores it). Encryption can also serve as a form of
1497 *access control* (i.e., it can control *who* can access the data) because examining the data
1498 requires access to the encryption keys. If the original data (with identities) are retained,
1499 they should be stored encrypted, and access should be limited. Even after de-identification,
1500 more sensitive data not intended for public release can be provided to select individuals by
1501 limiting access via encryption.

1502 **3.9.2. Secure Computation**

1503 Two technologies enable computing on encrypted data *without decrypting it*:

- 1504 1. **Fully-homomorphic encryption (FHE)** [55] allows a server to compute a function
1505 $f(x)$ on an encrypted value x without decrypting it. The result is a new encrypted
1506 value that can only be decrypted by someone who holds the original encryption key.
- 1507 2. **Secure multi-party computation (MPC)** [74] allows multiple servers to *jointly*
1508 compute a function $f(x_1, \dots, x_k)$, where each server provides one of the inputs x_i ,
1509 and no server learns any of the others' inputs.

1510 Both of these approaches are general-purpose in that they can be used to compute any
1511 function, and both are considerably slower than performing the equivalent computation

1512 with unencrypted data on a single computer. Nevertheless, both approaches are now suf-
1513 ficiently performant that they can be used for many practical kinds of privacy-preserving
1514 data analysis.¹⁸

1515 **3.9.3. Trusted Execution Environments**

1516 *Trusted Execution Environments* (TEEs) (also called *trusted hardware enclaves* or *secure*
1517 *hardware enclaves*) are another approach for computing on encrypted data. TEEs are im-
1518 plemented in computer hardware, typically within the silicon of a modern CPU, and pro-
1519 tect programs that run on that CPU from the surrounding environment. For example, a
1520 TEE can cause data from a computer’s CPU to be automatically encrypted when written
1521 to main memory and decrypted when read back to the CPU. In this way, data in memory
1522 are protected from other devices that can access memory, such as a network interface card.
1523 In addition to encryption, TEEs typically support attestation so that a program running on
1524 a TEE can attest to a remote system that the program is a true, legitimate, and faithful
1525 execution of the program.

1526 Traditional cloud services require trusting the cloud provider, who may have a compro-
1527 mised environment (e.g., an operating system that records encryption keys). A TEE de-
1528 creases the need for trust because it allows a user to validate that they are communicating
1529 with the remote program and offers assurance that no other program running in the cloud
1530 provider can access the program’s data. Secure enclaves can thus be used to allow untrusted
1531 infrastructure to operate on sensitive data in much the same way as technologies like FHE
1532 and MPC.

1533 Intel’s Software Guard Extensions (SGX) [112], ARM’s TrustZone [101], and AMD’s Se-
1534 cure Encrypted Virtualization (SEV) [13] are all examples of secure hardware enclaves.
1535 All of these products are designed to provide similar security to cryptographic techniques
1536 while also providing performance similar to a single CPU operating on unencrypted data.
1537 These secure hardware products are necessarily complex, and various implementation er-
1538 rors have been discovered that can allow attackers to defeat their security protections. Se-
1539 cure hardware enclaves certainly offer increased security for data compared to plaintext
1540 computation, but agencies should carefully consider the trade-off between performance
1541 and security when choosing between secure hardware and cryptographic techniques.

1542 **3.9.4. Physical Enclaves**

1543 For extremely sensitive data, a *physical enclave* (see Section 3.4) may provide additional
1544 security. In this model, data are stored on a computer not connected to any network and
1545 are accessible only via physical access to a particular room. Access to the data is then
1546 controlled by limiting access to the room. This approach can be quite cumbersome.

¹⁸More information about these and other kinds of secure computation can be found on the NIST Privacy-
Enhancing Cryptography (PEC) project website at <https://csrc.nist.gov/projects/pec>.

1547 **4. Technical Steps for Data De-Identification**

1548 The goal of de-identification is to transform data in a way that protects privacy while pre-
1549 serving the validity of inferences drawn on that data within the context of a target use-case.
1550 This section discusses technical options for performing de-identification and verifying the
1551 result of a de-identification procedure.

1552 Agencies should adopt a detailed, written process for de-identifying data prior to com-
1553 mencing work on a de-identification project. The details of the process will depend on the
1554 particular de-identification approach that is pursued. In developing technical steps for data
1555 de-identification, agencies may wish to consider existing de-identification standards, such
1556 as the HIPAA Privacy Rule, the IHE De-Identification Handbook [61], or the HITRUST
1557 De-Identification Framework [77].

1558 **4.1. Determine the Privacy, Data Usability, and Access Objectives**

1559 Agencies intent on de-identifying data for release should understand the nature of the
1560 data that they intend to de-identify and determine the policies and standards that will be
1561 used to determine acceptable levels of data accuracy, de-identification, and the risk of re-
1562 identification. For example:

- 1563 • Where did the data come from?
- 1564 • What promises were made when the data were collected?
- 1565 • What are the legal and regulatory requirements regarding data privacy and release?
- 1566 • What is the purpose of the data release?
- 1567 • What is the intended use of the data?
- 1568 • What data-sharing model (Section 3.4) will be used?
- 1569 • Which standards for privacy protection or de-identification will be used?
- 1570 • What is the level of risk that the project is willing to accept?
- 1571 • What are the goals for limiting re-identification? For example:
 - 1572 – No one can be re-identified.
 - 1573 – Only a few people can be re-identified.
 - 1574 – Only a few people can be re-identified in theory, but no one will actually be
1575 re-identified in practice.
 - 1576 – Only outliers can be re-identified.
 - 1577 – Only people who are not outliers can be re-identified.

- 1578 – There is a small percentage chance of re-identification that is shared by every-
1579 one in the dataset.
- 1580 – There is a small percentage chance of re-identification, but some people in the
1581 dataset are significantly more likely to be re-identified, and the re-identification
1582 probability is somehow bounded.
- 1583 • What harm might result from re-identification, and what techniques will be used to
1584 mitigate those harms?
- 1585 • How should compliance with that level of risk be determined?

1586 Some goals and objectives are synergistic, while others are in opposition.

1587 4.2. Conducting a Data Survey

1588 Different kinds of data require different kinds of de-identification techniques. As a result,
1589 an important early step in the de-identification of government data is to identify the data
1590 modalities that are present in the dataset and formulate a plan for de-identification that takes
1591 into account goals for data release, data accuracy, privacy protection, and the best available
1592 science.

1593 For example:

- 1594 • **Tabular numeric and categorical data** is the subject of the majority of de-identification
1595 research and practice. These datasets are most frequently de-identified by using tech-
1596 niques based on the designation and removal of direct identifiers and the manipula-
1597 tion of quasi-identifiers. The chief criticism of de-identification based on direct and
1598 quasi-identifiers is that administrative determinations of quasi-identifiers may miss
1599 variables that can be uniquely identifying when combined and linked with external
1600 data, including data that are not available at the time the de-identification is per-
1601 formed but become available in the future.

1602 *K*-anonymity [122] is a common framework for performing and evaluating the de-
1603 identification of tabular numeric and categorical data. However, *risk determinations*
1604 *based on this kind of de-identification will be incorrect if direct and quasi-identifiers*
1605 *are not properly classified.* For example, if there exist quasi-identifiers that are not
1606 identified as such and not subjected to *k-anonymity*, then it may be easy to re-identify
1607 records in the de-identified dataset.

1608 Tabular data may also be used to create a synthetic dataset that preserves some infer-
1609 ence validity but does not have a one-to-one correspondence to the original dataset.

- 1610 • **Dates and times** require special attention when de-identifying because temporal in-
1611 formation is inherently linked to an external dataset: the natural progression of time.
1612 Some dates and times (e.g., February 22, 1732) are highly identifying, while others
1613 are not. Dates that refer to matters of public record (e.g., date of birth, death, or home

1614 purchase) should be routinely taken as having high re-identification potential. Dates
1615 may also form the basis of linkages between dataset records or even within a record.
1616 For example, a record may contain the date of admission, the date of discharge, and
1617 the number of days in residence. Thus, care should be taken when de-identifying
1618 dates to locate and properly handle potential linkages and relationships. Applying
1619 different techniques to different fields may result in information being left in a dataset
1620 that can be used for re-identification. Specific issues regarding date de-identification
1621 are discussed in Section 4.3.4, “De-Identifying Dates.”

1622 • **Geographic and map data** also require special attention when de-identifying, as
1623 some locations can be highly identifying, other locations are not identifying at all,
1624 and some locations are only identifying at specific times. As with dates and times,
1625 de-identifying geographic locations is challenging because locations inherently link
1626 to an external reality, and some locations during specific time periods are highly
1627 correlated with specific individuals (e.g., 38.8977° N, 77.0365° W). Identifying lo-
1628 cations can be de-identified through the use of perturbation or generalization. The
1629 effectiveness of such de-identification techniques for protecting privacy in the pres-
1630 ence of external information has not been well-characterized [51, p.37][115]. Spe-
1631 cific issues regarding geographical de-identification are discussed in Section 4.3.5,
1632 “De-Identifying Geographical Locations.”

1633 • **Unstructured text** may contain direct identifiers, such as a person’s name, or may
1634 contain additional information that can serve as a quasi-identifier. Finding such iden-
1635 tifiers invariably requires domain-specific knowledge [51, p. 30]. Note that unstruc-
1636 tured text may be present in tabular datasets and require special attention.¹⁹

1637 • **Photos and video** may contain identifying information, such as printed names (e.g.,
1638 name tags), as well as metadata in the file format. A range of biometric techniques
1639 also exists for matching photos of individuals against a dataset of photos and identi-
1640 fiers [51, p. 32].

1641 • **Medical imagery** poses additional problems over photographs and video due to the
1642 presence of technical, medically specific information. For example, identifying in-
1643 formation may be present in the image itself (e.g., a photo may show an identifying
1644 scar or tattoo), an identifier may be “burned in” to the image area (e.g., an identifi-
1645 cation plate containing a patient name that is included in an X-Ray), or an identifier
1646 may be present in the file metadata. The body part in the image itself may also be
1647 recognized using a biometric algorithm and dataset [51, p.35].

1648 • **Genetic sequences** and other kinds of sequence information can be identified by
1649 using existing databanks that match sequences and identities. There is also evi-

¹⁹For an example of how unstructured text fields can damage the policy objectives and privacy assurances of a larger structured dataset, see Andrew Peterson’s article, “Why the names of six people who complained of sexual assault were published online by Dallas police” [98].

1650 dence that genetic sequences from individuals who are not in datasets can be matched
1651 through genealogical triangulation – a process that uses genetic information and other
1652 information as quasi-identifiers to single out a specific identity [51, p.36]. At present,
1653 there is no known method to reliably de-identify genetic sequences. Specific issues
1654 regarding the de-identification of genetic information is discussed in Section 4.3.6,
1655 “De-Identifying Genomic Information.”

1656 In many cases, data are complex and contain multiple modalities. Such mixtures may
1657 complicate risk determinations.

1658 **4.3. De-Identification by Removing Identifiers and Transforming Quasi-Identifiers**

1659 De-identification based on the removal of identifiers and the transformation of quasi-identifiers
1660 is one of the most common approaches currently in use. It has the advantage of being con-
1661 ceptually straightforward, and there is a long institutional history of using this approach
1662 within both federal statistical agencies and the healthcare industry. This approach has the
1663 disadvantage of not being based on formal methods for assuring privacy protection. The
1664 lack of formal methods does not mean that this approach cannot protect privacy, but it does
1665 mean that privacy protection is not assured.

1666 Below is a sample process for de-identifying data by removing identifiers and transforming
1667 quasi-identifiers:²⁰

- 1668 1. Determine the re-identification risk threshold. The organization determines accept-
1669 able risk for working with the dataset and possibly mitigating controls based on
1670 strong precedents and standards.²¹
- 1671 2. Determine the information in the dataset that could be used to identify the data sub-
1672 jects. Identifying information can include:
 - 1673 **Direct identifiers** including names, phone numbers, and other information that un-
1674 ambiguously identifies an individual.
 - 1675 **Quasi-identifiers** that could be used in a linkage attack. Typically, quasi-identifiers
1676 identify multiple individuals and can be used to triangulate a specific individual.
 - 1677 **High-dimensional data** [10] that can be used to single out data records and thus
1678 constitute a unique pattern that could be identifying if the values exist in a
1679 secondary source to link against.²²

²⁰This protocol is based on a protocol developed by Professors Khaled El Emam and Bradley Malin [44].

²¹See the Federal Committee on Statistical Methodology’s Data Protection Toolkit at <https://nces.ed.gov/fcsm/dpt>.

²²For example, Narayanan and Shmatikov demonstrated that the set of movies that a person had watched could be used as an identifier given the existence of a second dataset of movies that had been publicly rated [84].

- 1680 3. Determine the direct identifiers in the dataset. An expert determines the elements in
1681 the dataset that only serve to identify the data subjects.
- 1682 4. Mask (transform) direct identifiers. The direct identifiers are either removed or re-
1683 placed with pseudonyms. Options for performing this operation are discussed in
1684 Section 4.3.1.
- 1685 5. Perform threat modeling. The organization determines the additional information
1686 they might be able to use for re-identification, including both quasi-identifiers and
1687 non-identifying values that a data intruder might use for re-identification.
- 1688 6. Determine minimal acceptable data accuracy. The organization determines what uses
1689 can or will be made with the de-identified data.
- 1690 7. Determine the transformation process that will be used to manipulate the quasi-
1691 identifiers. Pay special attention to the data fields that contain dates and geographical
1692 information, removing or recoding as necessary.
- 1693 8. Import (sample) data from the source dataset. Because the effort to acquire data from
1694 the source (identified) dataset may be substantial, some researchers recommend a test
1695 data import run to assist in planning [44].
- 1696 9. Review the results of the trial de-identification. Correct any coding or algorithmic
1697 errors that are detected.
- 1698 10. Transform the quasi-identifiers for the entire dataset.
- 1699 11. Evaluate the actual re-identification risk, which is calculated. As part of this evalua-
1700 tion, every aspect of the released dataset should be considered in light of the question,
1701 “Can *this* information be used to identify someone?”
- 1702 12. Compare the actual re-identification risk with the threshold specified by the policy-
1703 makers.
- 1704 13. If the data do not pass the actual risk threshold, adjust the procedure and repeat Steps
1705 11 and 12. For example, additional transformations may be required. Alternatively,
1706 it may be necessary to remove outliers. Removing data will of course impact data
1707 quality, but it will also protect the privacy of the individuals whose data has been
1708 removed.

1709 **4.3.1. Removing or Transforming of Direct Identifiers**

1710 There are many possible processes for removing direct identifiers from a dataset, including:

- 1711 • **Removal and replacement.** Replace identifiers with the value used by the database
1712 to indicate a missing value, such as NULL or NA.
- 1713 • **Masking.** Replace identifiers with a repeating character, such as XXXXXX or 999999.

- 1714 • **Encryption.** Encrypt the identifiers with a strong encryption algorithm. After en-
1715 cryptation, the key can be discarded the cryptographic key to prevent decryption. How-
1716 ever, if there is a desire to employ the same transformation at a later point in time,
1717 the key should not be discarded but rather stored in a secure location separate from
1718 the de-identified dataset. Encryption used for this purpose carries special risks that
1719 need to be addressed with specific controls (see Section 4.3.2 below for further in-
1720 formation).
- 1721 • **Hashing with a keyed hash.** A keyed hash is a special kind of hash function that
1722 produces different hash values for different keys. The hash key should have sufficient
1723 randomness to defeat a brute force attack aimed at recovering the hash key (e.g.,
1724 SHA-256 HMAC [20] with a 256-bit randomly generated key). As with encryption,
1725 the key should be discarded unless there is a desire for repeatability. Hashing used
1726 for this purpose carries special risks that need to be addressed with specific controls
1727 (see Section 4.3.2 below for further information).
- 1728 • **Replacement with keywords.** This approach transforms identifiers such as George
1729 Washington to PATIENT. Note that some keywords may be equally identifying, such
1730 as transforming George Washington to PRESIDENT.
- 1731 • **Replacement with realistic surrogate values.** This approach transforms identifiers
1732 such as George Washington to surrogates that blend in, such as Abraham Polk.²³

1733 Encryption, hashing with a keyed hash, and replacement with realistic surrogate values are
1734 pseudonymization techniques. The technique used to remove direct identifiers should be
1735 clearly documented for users of the dataset – especially if the technique of replacement by
1736 realistic surrogate names is used – so that future data users have documentation that the
1737 dataset has been de-identified.

1738 If the agency plans to make data available for longitudinal research and contemplates mul-
1739 tiple data releases, then the transformation process should be repeatable, and the resulting
1740 transformed identities should be *pseudonyms*. The mapping between the direct identifier
1741 and the pseudonym is performed using a lookup table or a repeatable transformation. In
1742 either case, the release of the lookup table or the information used for the repeatable trans-
1743 formation will result in compromised identities. Thus, the lookup table or the information
1744 for the transformation must be highly protected. When using a lookup table, the pseudonym
1745 must be randomly assigned.

1746 A significant risk of using a repeatable transformation is that a data intruder may be able
1747 to determine the transformation and – in so doing – gain the capability to re-identify all of
1748 the records in the dataset.

²³A study by Carrell et. al found that using realistic surrogate names in de-identified text like John Walker and 3900 Pennsylvania Ave instead of generic labels like PATIENT and ADDRESS could decrease or mitigate the risk of re-identifying the few names that remained in the text because “the reviewers were unable to distinguish the residual (leaked) identifiers from the...surrogates” [21].

1749 When multiple organizations use the same pseudonymization scheme, they can trade data
1750 and perform matching on the pseudonyms. However, this practice also allows the orga-
1751 nizations to re-identify each other’s shared datasets. As an alternative, organizations can
1752 participate in a *private set intersection protocol*, of which there are many in the crypto-
1753 graphic literature [78, 34, 69].

1754 **4.3.2. Special Security Note Regarding the Encryption or Hashing of Direct** 1755 **Identifiers**

1756 The transformation of direct identifiers through encryption or hashing carries special risks,
1757 as errors in procedure or the release of the encryption key can compromise identities for
1758 the entire dataset.

1759 When information is protected with encryption, the security of the encrypted data depends
1760 entirely on the security of the encryption key. If a key is improperly chosen, it may be
1761 possible for a data intruder to discover the key using a brute force search. Because there
1762 is no visual difference between data that are encrypted with a strong encryption key and
1763 data that are encrypted with a weak key, organizations must utilize administrative controls
1764 to ensure that keys are both unpredictable and suitably protected. The use of encryption or
1765 hashing to protect direct identifiers is, therefore, *not recommended*.

1766 **4.3.3. De-Identifying Numeric Quasi-Identifiers**

1767 Once a determination is made regarding quasi-identifiers, they should be transformed. A
1768 variety of techniques are available to transform quasi-identifiers:

- 1769 • **Top and bottom coding.** Outlier values that are above or below certain values are
1770 coded appropriately. For example, the HIPAA Privacy Rules calls for ages over 89
1771 to be “aggregated into a single category of age 90 or older” [132, § 164.514 (b)].
- 1772 • **Micro aggregation.** Individual microdata are combined into small groups that pre-
1773 serve some data analysis capability while providing for some disclosure protec-
1774 tion [110].
- 1775 • **Generalize categories with small counts.** When preparing contingency tables, sev-
1776 eral categories with small values may be combined. For example, rather than re-
1777 porting that there is one person with blue eyes, two people with green eyes, and one
1778 person with hazel eyes, it may be reported that there are four people with blue, green,
1779 or hazel eyes.
- 1780 • **Data suppression.** Cells in contingency tables with counts lower than a predefined
1781 threshold can be suppressed to prevent the identification of attribute combinations
1782 with small numbers [141].
- 1783 • **Blanking and imputing.** Specific values that are highly identifying can be removed
1784 and replaced with imputed values.

- 1785 • **Attribute or record swapping.** Attributes or data values are swapped within a set
1786 of similar records. For example, data that represent families in two similar towns
1787 within a county might be swapped with each other. “Swapping has the additional
1788 quality of removing any 100-percent assurance that a given record belongs to a given
1789 household” [130, p.31] while preserving the accuracy of regional statistics, such as
1790 sums and averages. In this case, the average number of children per family in the
1791 county would be unaffected by data swapping. However, swapping may damage
1792 or destroy important relationships within the data and introduce systematic biases,
1793 depending on how the swapping candidates are selected.
- 1794 • **Noise infusion.** Also called “partially synthetic data,” this approach adds small ran-
1795 dom values to attributes. For example, instead of reporting that a person is 84 years
1796 old, the person may be reported as being 79 years old. Noise infusion increases
1797 variance in reported statistics and leads to attenuation bias in estimated regression
1798 coefficients and correlations among attributes [36, 7]. When combined with a re-
1799 quirement for non-negative reporting of attributes, such as age or population, noise
1800 infusion also introduces systematic bias since more values are increased in value than
1801 decreased.

1802 These techniques (and others) are described in detail in several publications, including:

- 1803 • **Statistical Policy Working Paper #22.** (Second version, 2005) by the Federal Com-
1804 mittee on Statistical Methodology [47]. This 137-page paper includes worked exam-
1805 ples of disclosure limitation, specific recommended practices for federal agencies,
1806 profiles of federal statistical agencies conducting disclosure limitation, and an ex-
1807 tensive bibliography. This document has been superseded by the Data Protection
1808 Toolkit.
- 1809 • **The Data Protection Toolkit (BETA).** A website maintained by the Federal Com-
1810 mittee on Statistical Methodology for the purpose of promoting data access while
1811 protecting confidentiality throughout the federal statistical system [48]. [https://nces.](https://nces.ed.gov/fcsm/dpt)
1812 [ed.gov/fcsm/dpt](https://nces.ed.gov/fcsm/dpt)
- 1813 • **The Anonymisation Decision-Making Framework.** By Mark Elliot, Elaine MacKey,
1814 Kieron O’Hara and Caroline Tudor, UKAN, University of Manchester, Manchester,
1815 UK, 2016. This 156-page book provides tutorials and worked examples for de-
1816 identifying data and calculating risk.
- 1817 • **IHE IT Infrastructure Handbook: De-Identification.** (Integrating the Health-
1818 care Enterprise, June 6, 2014) IHE offers a variety of guides, including one on de-
1819 identification at http://www.ihe.net/User_Handbooks/.

1820 Swapping and noise infusion both introduce noise into the dataset, such that records lit-
1821 erally contain incorrect data. Certain kinds of noise infusion have been mathematically
1822 proven to provide formal privacy guarantees. Swapping has no such guarantees.

1823 All of these techniques impact data accuracy, but whether they impact data *utility* depends
1824 on the downstream uses of the data. For example, top-coding household incomes will not
1825 impact a measurement of the 90-10 quantile ratio, but it will impact a measurement of the
1826 top 1% of household incomes [99].

1827 Prior to the adoption of differential privacy by the U.S. Census Bureau, federal statistical
1828 agencies largely did not document the specific statistical disclosure techniques they used
1829 when performing statistical disclosure limitation. Likewise, statistical agencies did not
1830 document the parameters used in the transformations nor the amount of data that have been
1831 transformed, as documenting these techniques can allow a data intruder to reverse-engineer
1832 the specific values, eliminating privacy protection [7]. This lack of transparency sometimes
1833 resulted in erroneous conclusions on the part of data users. This is another example of why
1834 it is important for documentation of the de-identification process to accompany the release
1835 of de-identified data and is one of the motivations for the U.S. Census Bureau to to adopt
1836 data privacy techniques that do not rely on secrecy for their effectiveness [56, 121, 58, 6].

1837 4.3.4. De-Identifying Dates

1838 Dates can exist in many ways in a dataset. Dates may be in particular kinds of typed
1839 columns, such as a date of birth or the date of an encounter. Dates may be present as
1840 a number, such as the number of days since an epoch like January 1, 1900. Dates may
1841 be present in the free text narratives or in photographs (e.g., a photograph that shows a
1842 calendar or a picture of a computer screen with date information).

1843 Several strategies have been developed for de-identifying dates:

- 1844 • Under the HIPAA Privacy Rule, dates must be generalized to no greater specificity
1845 than the year (e.g., July 4, 1776, becomes 1776).
- 1846 • Dates within a single person’s record can be systematically adjusted by a random
1847 amount. For example, dates of a hospital admission and discharge might be system-
1848 atically moved the same number of days – a date of admission and discharge of July
1849 4, 1776, and July 9, 1776, become Sept. 10, 1777, and Sept. 15, 1777 [89]. However,
1850 this does not eliminate the risk that a data intruder will make inferences based on the
1851 interval between dates.
- 1852 • In addition to a systematic shift, the intervals between dates can be perturbed to
1853 protect against re-identification attacks that involve identifiable intervals while still
1854 maintaining the order of events.
- 1855 • Some dates cannot be arbitrarily changed without compromising data accuracy. For
1856 example, it may be necessary to preserve the day of the week, whether a day is a
1857 workday or a holiday, or a relationship to a holiday or event.
- 1858 • Some ages can be randomly adjusted without impacting data accuracy while others
1859 cannot. For example, in many cases, the age of an individual can be randomly ad-

1860 justed ± 2 years if the person is over the age of 25 but not if their age is between
1861 one and three. However, individuals become eligible for specific benefits at specific
1862 ages, such as Social Security retirement at age 62, so changes to ages around these
1863 milestones may also result in data accuracy problems.

1864 **4.3.5. De-Identifying Geographical Locations and Geolocation Data**

1865 Geographical data can exist in many ways in a dataset. Geographical locations may be
1866 indicated by map coordinates (e.g., 39.1351966, -77.2164013), a street address (e.g., 100
1867 Bureau Drive), or a postal code (e.g., 20899). Geographical locations can also be embedded
1868 in textual narratives.

1869 Some geographical locations are not identifying (e.g., a crowded train station), while others
1870 may be highly identifying (e.g., a house in which a single person lives). Other locations
1871 may be identifying at some times of day and not others or during some months or some
1872 years. The amount of noise required to de-identify geographical locations significantly
1873 depends on the availability of external data, including geographical surveys. Identity may
1874 be shielded in an urban environment by adding ± 100 m, whereas a rural environment may
1875 only require ± 5 km or more to introduce sufficient ambiguity.

1876 A prescriptive de-identification rule – even one that accounts for varying population densi-
1877 ties – may still be insufficient for de-identification if the rule fails to consider the interaction
1878 between geographic locations and other quasi-identifiers in the dataset. Noise should be
1879 added with caution to avoid the creation of inconsistencies in underlying data (e.g., mov-
1880 ing the location of a residence along a coast into a body of water or across geopolitical
1881 boundaries).

1882 Single locations may become identifying if they represent locations linked to a single in-
1883 dividual that are recorded over time (e.g., a work/home commuting pair). Such behavioral
1884 time-location patterns can be quite distinct and allow for re-identification even with a small
1885 number of recorded locations per individual [82, 81]. Research in 2021 concluded that
1886 “[t]he risk of re-identification remains high even in country-scale location datasets” [46].

1887 Data that are of higher resolution are typically more identifying. For example, in July
1888 2021, the Catholic publication *The Pillar* published a report in which it had purchased
1889 the de-identified geolocation information for users of a homosexual dating platform. With
1890 this data, the journalists identified a prominent Catholic official as a user of the platform
1891 by simply matching the geolocation data to the official’s official residence. The official
1892 promptly resigned [100].

1893 **4.3.6. De-Identifying Genomic Information**

1894 Deoxyribonucleic acid (DNA) is the molecule inside human cells that carries genetic in-
1895 structions used for the proper functioning of living organisms. DNA present in the cell

1896 nucleus is inherited from both parents, while DNA present in the mitochondria is only
1897 inherited from an organism's mother.

1898 DNA is a repeating polymer that is made from four chemical bases: adenine (A), guanine
1899 (G), cytosine (C), and thymine (T). Human DNA consists of roughly 3 billion bases, of
1900 which 99% are the same in all people [54]. Modern technology allows for the complete
1901 specific sequence of an individual's DNA to be chemically determined, although this is
1902 rarely done in practice. With current technology, it is far more common to use a DNA
1903 microarray to probe for the presence or absence of specific DNA sequences at predeter-
1904 mined points in the genome. This approach is typically used to determine the presence
1905 or absence of specific single nucleotide polymorphisms (SNPs) [53]. DNA sequences and
1906 SNPs are the same for monozygotic (identical) twins, individuals resulting from divided
1907 embryos, and clones. With these exceptions, it is believed that no two humans have the
1908 same complete DNA sequence.

1909 Individual SNPs may be shared by many individuals, but a sufficiently large number of
1910 SNPs that show sufficient variability is generally believed to produce a combination that
1911 is unique to an individual. Thus, there are some sections of the DNA sequence and some
1912 combinations of SNPs that have high variability within the human population and oth-
1913 ers that have significant conservation between individuals within a specific population or
1914 group. When there is high variability, DNA sequences and SNPs can be used to match an
1915 individual with a historical sample that has been analyzed and entered into a dataset. The
1916 inheritability of genetic information has also allowed researchers to determine the surnames
1917 and even the complete identities of some individuals [57].

1918 As the number of individuals who have their DNA and SNPs measured increases, scientists
1919 are realizing that the characteristics of DNA and SNPs in individuals may be more com-
1920 plicated than the preceding paragraphs imply. DNA changes as individuals age because of
1921 senescence, transcription errors, and mutation. DNA methylation, which can impact the
1922 functioning of DNA, also changes over time [17]. Individuals who are made up of DNA
1923 from multiple individuals – typically the result of the fusion of twins in early pregnancy
1924 – are known as *chimera* or *mosaic*. In 2015, a man in the United States failed a paternity
1925 test because the genes in his saliva were different from those in his sperm [68]. A hu-
1926 man chimera was identified in 1953 because the person's blood contained a mixture of two
1927 blood types: A and O [37]. The incidence of human chimeras is unknown.

1928 Because of the high variability inherent in DNA, complete DNA sequences may be iden-
1929 tifiable by linking with an external dataset. Likewise, biological samples for which DNA
1930 can be extracted may be identifiable. Subsections of an individual's DNA sequence and
1931 collections of highly variable SNPs may be identifiable unless it is known that there are
1932 many individuals who share the region of DNA or those SNPs. Furthermore, genetic infor-
1933 mation may not only identify an individual but could also identify an individual's ancestors,
1934 siblings, and descendants.

Reading Level at Start of School Year	# of Students
Below grade level	30-39
At grade level	50-59
Above grade level	20-29

Table 1. Reading levels at a hypothetical school, as measured by entrance examinations, reported at the start of the school year on October 1.

Reading Level at Start of School Year	# of Students
Below grade level	30-39
At grade level	50-59
Above grade level	30-39

Table 2. Reading levels at a hypothetical school, as measured by entrance examinations, reported one month into the school year on November 1 after a new student has transferred to the school.

1935 **4.3.7. De-Identifying Text Narratives and Qualitative Information**

1936 Researchers must devote specific attention when they de-identify text narratives and other
1937 kinds of qualitative information. Many approaches developed in the 1980s and 1990s that
1938 provided reasonable privacy assurances at the time may no longer provide adequate protec-
1939 tion in an era with high-quality internet search and social media [96, 97]. This is an area of
1940 active research.

1941 **4.3.8. Challenges Posed by Aggregation Techniques**

1942 Aggregation does not necessarily provide privacy protection, especially when data are pre-
1943 sented in multiple data releases. Consider a hypothetical example of a school that reports
1944 on its website the number of students performing below, at, and above grade level at the
1945 start of the school year (table 1). Then consider that a new student enrolls at the school on
1946 October 15, and the school updates the table on its website (table 2).

1947 By comparing the two tables, it is possible to infer that the student who joined the school
1948 is likely performing above grade level. This reveals protected information. Moreover, if
1949 a person who views both tables knows the specific student who enrolled in October, they
1950 have learned a private fact about that student.

1951 Aggregation does not inherently protect privacy, and thus aggregation alone is not sufficient
1952 to provide formal privacy guarantees. However, the differential privacy literature does pro-
1953 vide methods for performing aggregation that are both formally private and highly accurate
1954 when applied to large datasets. These methods work through the addition of carefully cali-
1955 brated noise.

1956 **4.3.9. Challenges Posed by High-Dimensional Data**

1957 Even after removing all of the unique identifiers and manipulating the quasi-identifiers,
1958 data can still be identifying if it is of sufficiently high dimensionality and if there exists a
1959 way to link the supposedly non-identifying values to an identity.²⁴

1960 **4.3.10. Challenges Posed by Linked Data**

1961 Data can be linked in many ways. Pseudonyms allow data records from the same individual
1962 to be linked together over time. Family identifiers allow data from parents to be linked with
1963 their children. Device identifiers allow data to be linked to physical devices and potentially
1964 link together all data coming from the same device. Data can also be linked to geographical
1965 locations.

1966 Data linkage increases the risk of re-identification by providing more attributes that can be
1967 used to distinguish the true identity of a data record from others in the population. For ex-
1968 ample, survey responses that are linked together by household are more readily re-identified
1969 than survey responses that are not linked. Heart rate measurements may not be considered
1970 identifying, but given a long sequence of tests, each individual in a dataset would have a
1971 unique constellation of heart rate measurements, and the dataset could be susceptible to
1972 being linked with another dataset that contains the same values.²⁵ Geographical location
1973 data can – when linked over time – create individual behavioral time-location patterns that
1974 can be used to classify and identify unlabeled data, even with a small number of recorded
1975 locations per individual [82, 81].

1976 Dependencies between records may result in record linkages even when there is no explicit
1977 linkage identifier. For example, it may be that an organization has new employees take a
1978 proficiency test within seven days of being hired. This information would allow links to be
1979 drawn between an employee dataset that accurately reported an employee’s start date and a
1980 training dataset that accurately reported the date that the test was administered, even if the
1981 sponsoring organization did not intend for the two datasets to be linkable.

1982 **4.3.11. Challenges Posed by Composition**

1983 In computer science, the term *composition* refers to combining multiple functions to create
1984 more complicated ones. One of the defining characteristics of complex systems is that they
1985 have unpredictable behavior, even when they are composed of very simple components. A
1986 challenge of composition is to develop approaches for limiting or eliminating such unpre-
1987 dictable behavior. Typically, this is done by proactively limiting the primitives that can be

²⁴For example, consider a dataset of an anonymous survey that links together responses from parents and their children. In such a dataset, a child might be able to find their parents’ confidential responses by searching for their own responses and then following the link [84].

²⁵This is a different approach than characterizing an individual’s heartbeat pattern so that it can be used as a biometric. In this case, it is a specific sequence of heartbeats that is recognized.

1988 composed. De-identification is such a primitive that statisticians and data scientists must
1989 carefully control to ensure that the results of de-identification efforts can be composed.
1990 Without such controls, the results of composition can become unpredictable.

1991 Specifically, it is important to understand whether the techniques used for de-identifying
1992 will retain their privacy guarantees when they are subject to composition. For example, if
1993 the same dataset is made available through two different de-identification regimes, what
1994 will happen to the privacy guarantees if the two downstream datasets are recombined? One
1995 of the primary advantages of differential privacy is that its operators are composable. This
1996 is not true of most other de-identification techniques.

1997 Composition concerns can arise when:

- 1998 • The same dataset is provided to multiple downstream users.
- 1999 • Snapshots of a dataset are published on a periodic basis.
- 2000 • Changes in computer technology result in new aspects of a dataset being made avail-
2001 able.
- 2002 • Legal proceedings require that aspects of the dataset (attributes or a subset of records)
2003 are made available without de-identification.

2004 Privacy risk can result from unanticipated composition, which is one of the reasons that
2005 released datasets should be subjected to periodic review and reconsideration.

2006 4.3.12. Potential Failures of De-Identification

2007 The de-identification process outlined in this section can fail to prevent a disclosure for a
2008 number of different reasons. In addition, failures of data *utility* can also occur, in which
2009 the de-identification process removes *too much* information, and the de-identified dataset
2010 is not useful for its intended purpose.

- 2011 • If an **inappropriate risk threshold** is selected, then the risk of re-identification may
2012 be higher than intended. Agencies should select risk thresholds conservatively to
2013 address this issue.
- 2014 • If **direct or quasi-identifiers are missed**, then identifying information may remain in
2015 the de-identified dataset, leading to increased re-identification risk. Agencies should
2016 be mindful of the ways in which personal information can be used to identify indi-
2017 viduals and – in ambiguous situations – assume that such information is identifying.
- 2018 • If **threats are missed** during threat modeling, then the re-identification risk could
2019 be higher than intended. In particular, if other datasets that could be linked with
2020 the de-identified dataset are not considered, then the risk could be much higher than
2021 anticipated. Agencies should carefully consider existing and future data releases
2022 during threat modeling.

- 2023 • If the **selected transformations fail to remove identifying information**, then the
2024 risk of de-identification could be higher than intended. Agencies should select trans-
2025 formations with well-understood properties and a history of successful use.
- 2026 • If the de-identified dataset **does not produce accurate results for its intended use**,
2027 then it may not satisfy the goals of the data release. Future data custodians may be
2028 forced to oversee additional data releases, and those future releases might be com-
2029 bined with the already released datasets in ways that are unforeseen. Agencies should
2030 understand how the de-identified data will be used and make sure to carefully evalu-
2031 ate its utility for those purposes before releasing it.

2032 4.3.13. Post-Release Monitoring

2033 Following the release of a de-identified dataset, the releasing agency should monitor it to
2034 ensure that the assumptions made during the de-identification remain valid. This is because
2035 the identifiability of a dataset can only increase over time. For example, the de-identified
2036 dataset may contain information that can be linked to an internal dataset that is later the
2037 subject of a data breach. In such a situation, the data breach could also result in the re-
2038 identification of the de-identified dataset. The de-identified dataset might also be linked
2039 to an external dataset released by a completely separate organization. Agencies have no
2040 control over the release of such datasets, and even monitoring may be challenging in this
2041 situation. In some cases, the de-identified dataset might be linked with privately held data,
2042 making monitoring impossible.

2043 Agencies may wish to make releasing units responsible for post-release monitoring or to
2044 centralize the post-release monitoring in a single location. However, proper post-release
2045 monitoring requires knowledge of the datasets that have been released and the kinds of data
2046 that would allow for a re-identification attack. These requirements are likely to increase
2047 costs to organizations that wish to delegate post-release monitoring to other organizations
2048 or third parties. One way to decrease the requirement for post-release monitoring is to
2049 perform the de-identification using a formal privacy model (e.g., differential privacy) that
2050 provides for privacy without making assumptions about background information available
2051 to the data intruder.

2052 4.4. Synthetic Data

2053 An alternative to de-identifying using the technique presented in the previous section is to
2054 use the original dataset to create a synthetic dataset [35, p.8].

2055 Synthetic data can be created by two approaches:

- 2056 1. Sampling an existing dataset and either adding noise to specific cells likely to have a
2057 high risk of disclosure or replacing those cells with imputed values. This is known
2058 as a “partially synthetic” dataset (see Table 3).

Data adjective	Description
<i>Datasets without formal guarantees:</i>	
Partially synthetic	Data for which there may be one-to-one mappings between records in the original dataset and the synthetic dataset but for which some attributes may have been altered or swapped between records. This approach is sometimes called <i>blank-and-impute</i> .
<i>Datasets with formal guarantees if the original dataset is not used to create the data:</i>	
Test	Data that resemble the original dataset in terms of structure and the range of values but for which there is no attempt to ensure that inferences drawn on the test data will be like those drawn on the original data. Test data may also include extreme values that are not in the original data but are present for testing software.
Realistic	Data that have a characteristic that is like the original data but that is not developed by modifying original data and which contains no information that is privacy-sensitive.
<i>Datasets with formal guarantees when formal techniques are used:</i>	
Fully synthetic	Data for which there is no one-to-one mapping between any record in the original dataset and the synthetic dataset.

Table 3. Adjectives used for describing data in data releases.

2059 2. Using the existing dataset to create a model and then using that model to create a
2060 synthetic dataset. This is known as a “fully synthetic” dataset (see Table 3).

2061 In both cases, formal privacy techniques can be used to quantify the privacy protection
2062 offered by the synthetic dataset.

2063 **4.4.1. Partially Synthetic Data**

2064 A partially synthetic dataset is one in which some of the data have been altered from the
2065 original dataset using probabilistic models. For example, data that belong to two families
2066 in adjoining towns may be swapped to protect the identity of the families. Alternatively, the
2067 data for an outlier variable may be removed and replaced with a range value that is incorrect
2068 (e.g., replacing the value “60” with the range “30-35”). It is considered best practice for
2069 the data publisher to indicate that some values have been modified or otherwise imputed
2070 but not to reveal the specific values that have been modified.

2071 **4.4.2. Test Data**

2072 It is also possible to create *test data* that is syntactically valid but does not convey accu-
2073 rate information when analyzed. Such data can be used for software development. When
2074 creating test data, it is useful for the names, addresses, and other information in the data to
2075 be conspicuously non-natural so that the test data are not inadvertently confused with true
2076 confidential data. For example, use the name “FIRSTNAME1 LASTNAME2” rather than
2077 “JOHN SMITH.”

2078 **4.4.3. Fully Synthetic Data**

2079 A fully synthetic dataset is a dataset for which there is no one-to-one mapping between
2080 data in the original dataset and data in the de-identified dataset. One approach to creating a
2081 fully synthetic dataset is to use the original dataset to create a high-fidelity model and then
2082 to use a simulation to produce individual data elements that are consistent with the model.
2083 Special efforts must be taken to maintain marginal and joint probabilities when creating
2084 partially or fully synthetic data.

2085 Fully synthetic datasets cannot provide more information to the downstream user than was
2086 contained in the original model. Nevertheless, some users may prefer to work with the fully
2087 synthetic dataset instead of the model for a variety of reasons:

- 2088 • Synthetic data provides users with the ability to develop queries and other techniques
2089 that can be applied to the real data without exposing real data to users during the
2090 development process. The queries and techniques can then be provided to the data
2091 owner, who can run the queries or techniques on the real data and provide the results
2092 to the users.
- 2093 • Many hypotheses not represented exactly in the original model may be informed by
2094 the synthetic data because they are correlated with hypotheses (effects) that are in the
2095 model.
- 2096 • Some users may place more trust in a synthetic dataset than in a model.
- 2097 • When researchers form their hypotheses from synthetic data and then verify their
2098 findings on actual data, they can be protected from pretest estimation and false-
2099 discovery bias [7, p.257].

2100 Because of the possibility of false discovery, analysts should be able to validate their dis-
2101 coveries against the original data to ensure that the things they discover are in the original
2102 data and not artifacts of the data generation process.

2103 Both high-fidelity models and synthetic data generated from models may leak personal
2104 information that is potentially re-identifiable. The amount of leakage can be controlled us-
2105 ing formal privacy models (e.g., differential privacy) that typically involve the introduction

2106 of noise. Section 4.4.6 describes the construction of fully synthetic data with differential
2107 privacy.

2108 There are several advantages for agencies that choose to release de-identified data as a fully
2109 synthetic dataset:

- 2110 • It can be very difficult or even impossible to map records to actual people.
- 2111 • The privacy guarantees can potentially be mathematically established and proven (cf.
2112 the section below on “Creating a synthetic dataset with differential privacy”).
- 2113 • The privacy guarantees can remain in force even if there are future data releases.

2114 Fully synthetic data also have these disadvantages and limitations:

- 2115 • It is not possible to create pseudonyms that map back to actual people because the
2116 records are fully fabricated.
- 2117 • The data release may be less useful for accountability or transparency. For example,
2118 investigators equipped with a synthetic data release would be unable to find the actual
2119 “people” who make up the release because they would not actually exist.
- 2120 • It is difficult to find meaningful correlations or abnormalities in synthetic data that
2121 are not represented in the model. For example, if a model contains only main effects
2122 and first-order interactions, then all second-order interactions can only be estimated
2123 from the synthetic data to the extent that their design is correlated with the main or
2124 first-order interactions.
- 2125 • Users of the data may not realize that the data are synthetic. Simply providing doc-
2126 umentation that the data are fully synthetic may not be sufficient public notification
2127 since the dataset may be separated from the documentation. Instead, it is best to
2128 indicate in the data itself that the values are synthetic. For example, names like
2129 “SYNTHETIC PERSON” or “FIRSTNAME1 LASTNAME1” may be placed in the
2130 data.
- 2131 • Releasing a synthetic dataset may not be regarded by the public as a legitimate act of
2132 transparency, or the public may question the validity of the data based on its perceived
2133 lack of relationship to the original dataset. These concerns can be addressed with
2134 public education and by documenting the accuracy of the synthetic dataset.

2135 In addition, it can be extremely challenging to construct the high-fidelity models that enable
2136 good synthetic datasets. The best known techniques for constructing these models are
2137 designed around ensuring that specific properties of the data (e.g., correlations between
2138 certain data attributes) are preserved when the model is constructed. Models constructed
2139 this way may not necessarily reflect *other* properties that were present in the original data.

2140 It is often possible to construct very high-fidelity models when the desirable properties
2141 of the synthetic data are known in advance (e.g., when it is known what questions future

2142 analysts will want to answer using the synthetic data). Constructing synthetic data that
2143 faithfully represents *all* properties of the original data is much more challenging.

2144 **4.4.4. Synthetic Data with Validation**

2145 Agencies that share or publish synthetic data can optionally provide a validation service
2146 that takes queries or algorithms developed with synthetic data and applies them to actual
2147 data. The results of these queries or algorithms can then be compared with the results
2148 of running the same queries on the synthetic data, and the researchers can be warned if
2149 the results are different. Alternatively, results can be provided to the researchers after the
2150 application of additional statistical disclosure limitation.

2151 **4.4.5. Synthetic Data and Open Data Policy**

2152 Releases of synthetic data can be confusing to the lay public.

- 2153 • It may not be clear to data users that the synthetic data release is actually synthetic.
2154 Members of the public may assume instead that the synthetic data are simply an
2155 operational dataset that has had identifying columns suppressed.
- 2156 • Synthetic data may contain synthetic individuals who appear similar to actual indi-
2157 viduals in the population.
- 2158 • Fully synthetic datasets do not have a zero-disclosure risk because they still contain
2159 information derived from non-public information about individuals. The disclosure
2160 risk may be greater when synthetic data are created with traditional statistical mod-
2161 eling or data imputing techniques rather than those based on formal privacy models,
2162 such as differential privacy, as the formal models have provisions for tracking the
2163 accumulated privacy loss that results from multiple data operations, as discussed in
2164 Section 4.4.6.

2165 **4.4.6. Creating a Synthetic Dataset with Differential Privacy**

2166 A growing number of mathematical algorithms have been developed for creating syn-
2167 thetic datasets that meet the mathematical definition of privacy provided by differential
2168 privacy [40]. Most of these algorithms will transform a dataset containing private data into
2169 a new dataset that contains synthetic data that nevertheless provides reasonably accurate
2170 results in response to a variety of queries. However, there is no algorithm or implemen-
2171 tation currently in existence that can be used by a person who is unskilled in the area of
2172 differential privacy.

2173 The idea of differential privacy is that the result of a data analysis function κ applied to a
2174 dataset should not change very much if an arbitrary person p 's data is added to or removed
2175 from a dataset D . That is, $\kappa(D) \approx \kappa(D - p)$. The degree to which the two values are
2176 approximately equal is determined by the privacy loss parameter ϵ .

2177 In the mathematical formulation of differential privacy, ϵ can range from 0 to ∞ . When
2178 $\epsilon = 0$, the output of κ does not depend on the input dataset. When $\epsilon = \infty$, the output of κ
2179 is entirely dependent upon the input dataset, such that changing a single record results in
2180 an unambiguous measurable change in κ 's output. Thus, larger values of ϵ provide for more
2181 accuracy but result in increased privacy loss.

2182 When ϵ is set appropriately, differential privacy limits the privacy loss that a data subject
2183 experiences from the use of their private data to the maximum privacy loss necessary for
2184 a given statistical purpose. Note that this particular notion of privacy does *not* protect all
2185 secrets about a person. It only protects the secrets that an observer would not have been
2186 able to learn if the person's data was not present in the dataset. Stated another way, differ-
2187 ential privacy protects individuals from *additional harm resulting from their participation*
2188 *in the data analysis* but does not protect them from harm that would have occurred *even*
2189 *if their data were not present*. For example, if a study concludes that residents of Ver-
2190 mont overwhelmingly drive 4-wheel-drive vehicles, one might conclude that a *particular*
2191 *Vermont*er drives a 4-wheel-drive vehicle even if that individual did not participate in the
2192 study. Differential privacy does not attempt to prevent inferences of this type.

2193 Many academic papers on differential privacy assume a value of 1.0 for ϵ but do not explain
2194 the rationale of the choice. Some researchers working in the field of differential privacy
2195 have tried mapping existing privacy regulations to the choice of ϵ , but these efforts invari-
2196 ably result in values of $\epsilon \neq 1$. Principled approaches for setting ϵ is a subject of current
2197 academic research [72].

2198 There are relatively few scholarly publications regarding the deployment of differential pri-
2199 vacy in real-world situations, and there are few papers that provide guidance in choosing
2200 appropriate values of ϵ . Thus, agencies that are interested in using differential privacy al-
2201 gorithms to allow for querying of sensitive datasets or the creation of synthetic data should
2202 ensure that the techniques are appropriately implemented and that the privacy protections
2203 are appropriate to the desired application.

2204 **4.5. De-Identifying with an Interactive Query Interface**

2205 Another model for granting public access to de-identified agency information is to construct
2206 an interactive query interface that allows members of the public or qualified investigators to
2207 run queries over the agency's dataset. This option has been developed by several agencies,
2208 and there are many ways that it can be implemented. For example:

- 2209 • If the queries are run on actual data, the results can be altered through the injection
2210 of noise to protect privacy, potentially satisfying a formal privacy model such as
2211 differential privacy. Alternatively, individual queries can be reviewed by agency staff
2212 to verify that privacy thresholds are maintained.
- 2213 • Queries can be run on synthetic data. In this case, the agency can also run queries
2214 on the actual data and warn the external researchers if the queries run on synthetic

2215 data deviate significantly from the queries run on the actual data (ensuring that the
2216 warning itself does not compromise the privacy of some individual).

- 2217 • Query interfaces can be made freely available on the public internet, or they can be
2218 made available in a restricted manner to qualified researchers operating in secure
2219 locations.

2220 A significant privacy risk with interactive queries is that each query results in additional
2221 privacy loss [33].²⁶ For this reason, query interfaces should also log both queries and
2222 query results in order to deter and detect malicious use.

2223 One of the advantages of synthetic data is that the privacy loss budget can be spent on
2224 creating the synthetic dataset rather than on responding to interactive queries.

2225 **4.6. Validating a De-Identified Dataset**

2226 Agencies should validate datasets after they are de-identified to ensure that the resulting
2227 dataset meets the agency's goals in terms of both data usefulness and privacy protection.

2228 **4.6.1. Validating Data Usefulness**

2229 De-identification decreases data accuracy and the usefulness of the resulting dataset. It is
2230 therefore important to ensure that the de-identified dataset is still useful for the intended
2231 application. Otherwise, there is no reason to go through the expense and added risk of
2232 de-identification.

2233 Several approaches exist for validating data usefulness. For example, insiders can perform
2234 statistical calculations on both the original dataset and the de-identified dataset and compare
2235 the results to see if the de-identification resulted in unacceptable changes. Agencies can
2236 engage trusted outsiders to examine the de-identified dataset and determine whether the
2237 data could be used for the intended purpose.

2238 Recognizing that there is an inherent trade-off between data accuracy and privacy protec-
2239 tion, agencies can adopt accuracy goals for the data that they make available to a broad
2240 audience. An accuracy goal specifies how accurate data must be in order to be fit for an
2241 intended use. Limiting data accuracy to this goal is an important technique for protecting
2242 the privacy of data subjects.

2243 **4.6.2. Validating Privacy Protection**

2244 Several approaches exist for validating the privacy protection provided by de-identification,
2245 including:

²⁶If a finite privacy loss budget is allocated, the data controller needs to respond by increasing the amount of noise added to each response, accepting a higher level of privacy risk, or ceasing to answer questions as the budget nears exhaustion. This can result in equity issues if the first users to query the dataset obtain better answers than later users.

- 2246 • Examining the resulting data files to make sure that no identifying information is
2247 unintentionally included in file data or metadata.
- 2248 • Examining the resulting data files to make sure that the data meet stated goals for
2249 ambiguity under a k -anonymity model, if such a standard is desired.
- 2250 • Critically evaluating all default assumptions used by software that performs data
2251 modification or modeling.
- 2252 • Conducting a *motivated intruder test* to see if reasonably competent outside indi-
2253 viduals can perform re-identification using publicly available datasets, commercially
2254 available datasets, or even private datasets that might be available to certain data
2255 intruders. Motivations for an intruder can include prurient interest, causing embar-
2256 rassment or harm, revealing private facts about public figures, or engaging in a rep-
2257 utation attack. Details for how to conduct a motivated intruder test can be found in
2258 *Anonymisation: Managing data protection risk code of practice*, published by the
2259 United Kingdom’s Information Commissioner’s Office [63].
- 2260 • Providing the team conducting the motivated intruder test with some confidential
2261 agency data to understand how a data intruder might be able to take advantage of
2262 data leaked as a result of a breach or a hostile insider.

2263 These approaches do not provide provable guarantees on the protection offered by de-
2264 identification, but they may be useful as part of an overall agency risk assessment.²⁷ Ap-
2265 plications that require provable privacy guarantees should rely on formal privacy methods,
2266 such as differential privacy, when planning their data releases.

2267 Validating the privacy protection of de-identified data is greatly simplified by using vali-
2268 dated de-identification software, as discussed in Section 5, “Evaluation.”

2269 4.6.3. Re-Identification Studies

2270 Re-identification studies are motivated intruder tests. These studies can identify issues that
2271 would allow external actors to successfully re-identify de-identified data. Re-identification
2272 studies look for vulnerabilities in a dataset that could be used for re-identifying data sub-
2273 jects. They do not determine whether someone with intimate knowledge of a specific re-
2274 spondent can find that respondent in the database. The only way to protect a single specific

²⁷Although other documents that discuss de-identification use the term *risk assessment* to refer to a specific calculation of ambiguity using the k -anonymity de-identification model, this document uses the term *risk assessment* to refer to a much broader process. Specifically, risk assessment is defined as, “The process of identifying, estimating, and prioritizing risks to organizational operations (including mission, functions, image, reputation), organizational assets, individuals, other organizations, and the Nation, resulting from the operation of an information system. Part of risk management incorporates threat and vulnerability analyses and considers mitigations provided by security controls planned or in place. Synonymous with risk analysis” [23].

2275 individual perceived to be at high risk of re-identification is through data perturbation (e.g.,
2276 noise injection) or information reduction (e.g., removing the observation altogether).

2277 The key statistic calculated in re-identification studies is the *conditional re-identification*
2278 *rate*. This statistic is a proxy for disclosure risk. The rate is the number of confirmed links
2279 between the dataset and another dataset divided by the number of putative (suspected)
2280 links, unduplicated by “defender” ID, expressed as a percentage. If the conditional re-
2281 identification rate falls above an agreed upon threshold for any publication strata, it suggests
2282 that the data should not be released outside of a controlled environment.

2283 Re-identification studies are often an iterative process. If a re-identification study uncovers
2284 problems with the de-identified data, the data curator can engage with subject matter ex-
2285 perts, make changes to the dataset, and perform another re-identification study. Changes to
2286 the dataset might involve coarsening linking variables, eliminating highly disclosive link-
2287 ing variables from the microdata to be released, or coarsening strata. This continues until
2288 the study concludes that the de-identified data can be disseminated.

2289 There are two very different types of re-identification studies:

- 2290 1. **Micro (or targeted) re-identification studies**, where one is looking for a specific
2291 person. A well-known example is that of former Governor William Weld of Mas-
2292 sachusetts, whose medical records in a hospital discharge summary were record
2293 linked to voter records [16]. As noted earlier, individual targets are supremely hard
2294 to protect as there is often extensive publicly available information about them.
- 2295 2. **Macro (or wholesale) re-identification studies**, where one seeks to embarrass or
2296 discredit the organization releasing the data. This is accomplished by linking easily
2297 procurable external intruder data to the protected microdata that are being released.
2298 Several metrics can be calculated to uncover putative links, and several methods can
2299 be used to confirm putative links. Python has record linkage objects that probabilis-
2300 tically link files using a wide variety of metrics.

2301 Formal privacy parameters often appear opaque and elusive to non-theoreticians. Subject
2302 matter experts and decision-makers more clearly understand disclosure risk after reviewing
2303 the results of re-identification studies.

2304 External intruders may calculate low or high suspected re-identification rates, given the
2305 information they have available to them. They may even purport to have successfully linked
2306 their external data to a de-identified dataset. By conducting a re-identification study *a*
2307 *priori*, those seeking to disseminate the de-identified data know how successful the external
2308 intruder’s re-identification attempt was if both parties have access to the same external
2309 internal data.

2310 The conditional re-identification rate is identical to the metric of *precision* in the record
2311 linkage and health science literature. It represents the ratio of true positives to the sum
2312 of true positives and false positives. Data owners should not be alarmed if an external

2313 organization reports a relatively high suspected re-identification rate as long as they know
2314 that the conditional re-identification rate is low [45, 59, 111].

2315 Confirmed re-identification rates are defined in Section 3.2.1 as *re-identification probabil-*
2316 *ities*. On its own, a low confirmed re-identification probability does not indicate that an or-
2317 ganization should disseminate a de-identified dataset. Even when a confirmed rate is low,
2318 a high *conditional* rate should direct an organization to not disseminate the de-identified
2319 microdata.

2320 Re-identification studies may identify problems that can direct improvements to any or-
2321 ganization’s disclosure avoidance methods. Re-identification studies are not designed to
2322 replace legacy or modern provable privacy methods but to act as a quality control to vali-
2323 date that the methods – old and new – protect as they were designed.

2324 **5. Software Requirements, Evaluation, and Validation**

2325 Agencies should clearly define the requirements for de-identification algorithms and the
2326 software that implements those algorithms. They should be sure that the algorithms that
2327 they intend to use are validated, that the software implements the algorithms as expected,
2328 and that the data that result from the operation of the software are correct.

2329 Today, there a growing number of algorithms and tools for performing de-identification,
2330 data masking, and performing other privacy-preserving operations. NIST maintains a list of
2331 some of these tools at [https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/
2332 collaboration-space/focus-areas/de-id/tools](https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/de-id/tools). Such tools are also increasingly being eval-
2333 uated in academic literature [116] and by NIST [107, 1], although there are no widely
2334 accepted performance standards or certification procedures at present.

2335 **5.1. Evaluating Privacy-Preserving Techniques**

2336 There have been decades of research in the field of statistical disclosure limitation and
2337 de-identification, and understanding in the field has evolved over time. Agencies should
2338 not base their technical evaluation of a technique solely on the fact that the technique has
2339 been published in peer-reviewed literature or that the agency has a long history of using
2340 the technique and has not experienced any problems. Instead, it is necessary to evaluate
2341 proposed techniques through the totality of scientific experience and with regard to current
2342 threats.

2343 Traditional statistical disclosure limitation and de-identification techniques base their risk
2344 assessments – in part – on an expectation of what kinds of data are available to a data
2345 intruder to conduct a linkage attack. Where possible, these assumptions should be docu-
2346 mented and published along with a description of the privacy-preserving techniques that
2347 were used to transform the datasets prior to release so that they can be reviewed by external
2348 experts and the scientific community.

2349 Because our understanding of privacy technology and the capabilities of privacy attacks
2350 are rapidly evolving, techniques that have been previously established should be periodi-
2351 cally reviewed. New vulnerabilities may be discovered in techniques that have been pre-
2352 viously accepted. Alternatively, new techniques may be developed that allow agencies to
2353 re-evaluate the trade-offs they have made with respect to privacy risk and data usability.

2354 **5.2. De-Identification Tools**

2355 A de-identification tool is a program that is involved in the creation of de-identified datasets.

2356 **5.2.1. De-Identification Tool Features**

2357 De-identification tools may perform many functions, including:

- 2358 • Detecting identifying information
- 2359 • Calculating re-identification risk
- 2360 • Performing de-identification
- 2361 • Mapping identifiers to pseudonyms
- 2362 • Providing for the selective revelation of pseudonyms

2363 De-identification tools may handle a variety of data modalities. For example, tools may
2364 be designed for tabular data or for multimedia. Tools may attempt to de-identify all data
2365 types or be developed for specific modalities. A potential risk of using de-identification
2366 tools is that a tool could be equipped to handle some but not all of the different modalities
2367 in a dataset. For example, a tool could de-identify the categorical information in a table
2368 according to a de-identification standard but might not detect or attempt to address the
2369 presence of identifying information in a text field. For this reason, de-identification tools
2370 should be validated for the specific kinds of data that the agency intends to use.

2371 **5.2.2. Data Provenance and File Formats**

2372 Output files created by de-identification tools and data masking tools can record provenance
2373 information, such as metadata regarding input datasets, the de-identification methods used,
2374 and the resulting decrease in data accuracy. Output files can also be explicitly marked to in-
2375 dicate that they have been de-identified. For example, de-identification profiles that are part
2376 of the Digital Imaging and Communications in Medicine (DICOM) specification indicate
2377 which elements are direct versus quasi-identifiers and which de-identification algorithms
2378 have been employed [32, Appendix E, “Attribute Confidentiality Profiles”].

2379 **5.2.3. Data Masking Tools**

2380 Data masking tools are programs that can remove or replace designated fields in a dataset
2381 while maintaining relationships between tables. These tools can be used to remove direct

2382 identifiers but generally cannot identify or modify quasi-identifiers in a manner consistent
2383 with a privacy policy or risk analysis.

2384 Data masking tools were developed to allow software developers and testers access to
2385 datasets that contain realistic data while providing minimal privacy protection. Absent
2386 additional controls or data manipulations, data masking tools should not be used for the
2387 de-identification of datasets that are intended for public release nor as the sole mechanism
2388 to ensure confidentiality in non-public data sharing.

2389 **5.3. Evaluating De-Identification Software**

2390 Once techniques are evaluated and approved, agencies should ensure that the techniques are
2391 faithfully executed by their chosen software. Privacy software evaluation should consider
2392 the trade-off between data usability and privacy protection. Privacy software evaluation
2393 should also seek to detect and minimize the chances of tool error and user error.

2394 For example, agencies should verify:

- 2395 • **Correctness.** The software properly implements the chosen algorithms.
- 2396 • **Containment.** The software does not leak identifying information in expected or
2397 unexpected ways, such as through the inaccuracies of floating-point arithmetic or the
2398 differences in execution time (if observable to a data intruder).
- 2399 • **Usability.** The software can be operated efficiently and with minimal error, and users
2400 can detect and correct errors when they happen.

2401 Agencies should also evaluate the performance of the de-identification software, such as:

- 2402 • **Efficiency.** How long does it take to run on a dataset of a typical size?
- 2403 • **Scalability.** How much does it slow down when moving from a dataset of N to $100N$?
- 2404 • **Repeatability.** If the tool is run twice on the same dataset, are the results similar? If
2405 two different people run the tool, do they get similar results?

2406 Ideally, software should be able to track the accumulated privacy leakage from multiple
2407 data releases.

2408 **5.4. Evaluating Data Accuracy**

2409 Finally, agencies should evaluate the accuracy of the de-identified data to verify that it is
2410 sufficient for the intended use. For example, researchers at MIT and Harvard applied k -
2411 anonymity de-identification to educational data collected by a massive open online course
2412 operated by MITx and HarvardX on the edX platform and found that de-identification
2413 resulted in meaningful biases that changed the meaning of some statistics. For example, in
2414 one case, de-identification decreased the number of enrolled female students from 29% to
2415 26% because of the need to suppress attributes for specific microdata [30].

2416 The field of statistical disclosure control has developed approaches for gauging the impact
2417 of SDC techniques on microdata [142]. The literature examines the mathematical impact
2418 of SDC procedures (e.g., sampling, recoding, suppression, rounding, and noise infusion)
2419 and computes the possible impact on various statistical measurements.

2420 Approaches for evaluating data accuracy include [71]:

- 2421 • Demonstrating that machine learning algorithms trained on the de-identified data can
2422 accurately predict the original data and vice versa
- 2423 • Verifying that statistical distributions do not incur undue bias because of the de-
2424 identification procedure
- 2425 • Publishing sufficient information about the statistical properties of the disclosure lim-
2426 itation methods to permit the correction of inferences using these properties

2427 Agencies can create or adopt standards regarding the accuracy of de-identified data. If
2428 data accuracy cannot be well-maintained along with data privacy goals, then the release of
2429 data that is inaccurate for statistical analyses could potentially result in incorrect scientific
2430 conclusions and incorrect policy decisions.

2431 **6. Conclusion**

2432 Government agencies can use de-identification technology to make datasets available to
2433 researchers and the public without compromising the privacy of the people contained within
2434 the data.

2435 There are currently three primary models available for de-identification:

- 2436 1. agencies can make data available with traditional de-identification techniques that
2437 rely on the suppression of identifying information (direct identifiers) and the manip-
2438 ulation of information that partially identifies (quasi-identifiers);
- 2439 2. agencies can create synthetic datasets; and
- 2440 3. agencies can make data available through a query interface.

2441 These models can be mixed within a single dataset to provide different kinds of access for
2442 different users or intended uses.

2443 Privacy protection can be strengthened when agencies employ formal models for privacy
2444 protection, such as differential privacy, because the mathematical models that these sys-
2445 tems use are designed to ensure privacy protection irrespective of future data releases or
2446 developments in re-identification technology. However, the mathematics underlying these
2447 systems is very new, and there is little experience within the Government in using these
2448 systems. Thus, agencies should understand the implications of these systems before de-
2449 ploying them in place of traditional de-identification approaches that do not offer formal
2450 privacy guarantees.

2451 Agencies that use de-identification should establish appropriate governance structures to
2452 support de-identification, data release, and post-release monitoring. Such structures will
2453 typically include a Disclosure Review Board as well as appropriate education, training,
2454 and research efforts.

2455 A summary of this document's advice for practitioners appears in Figure 5.

2456 In closing, it is important to remember that different jurisdictions may have different stan-
2457 dards and policies regarding the definition and use of de-identified data. Information that
2458 is considered de-identified in one jurisdiction may be regarded as being identifiable in an-
2459 other.

Governance and Management (Section 3) The management of de-identification includes identifying the goals of the de-identification process and considering risks to participants in the data release. To guide this process, this document describes several tools:

- Consider all phases of the **Data Life Cycle** (Section 3.3).
- Consider different **Data Sharing Models** (Section 3.4), including complementary protections like Data Use Agreements, Synthetic Data, and Enclaves.
- Leverage the **Five Safes** (Section 3.5), a methodology for evaluating risk.
- Form a **Disclosure Review Board** (Section 3.6) to oversee the implementation of de-identification policies.
- Follow existing **de-identification standards** when possible (Section 3.7).

Technical Steps (Section 4) The technical process of de-identification should leverage the best practices developed over the past several decades. In particular, NIST recommends that agencies:

- Conduct a **Data Survey** (Section 4.2) to identify de-identification requirements specific to the data.
- Determine **identifiers** and **quasi-identifiers** in the data, and select a method for de-identifying each one (Section 4.3).
- Consider the existing **auxiliary data** (Section 4.3) that could be used to enable a re-identification attack.
- Practice **defense in depth** by combining security measures with de-identification when possible, and consider using **Synthetic Data** (Section 4.4) or an **Interactive Query Interface** (Section 4.5).
- When possible, use **formal privacy techniques** to quantify privacy loss associated with the release of de-identified data (Section 4.4.6).
- Validate the **utility** and **privacy** of the de-identified data (Section 4.6). In particular, establish accuracy goals for de-identification data so that the data is not more accurate than required for the intended purpose.

Software (Section 5) In general, agencies should:

- Utilize automated, repeatable, software-based approaches for performing de-identification.
- Carefully consider the software used to implement de-identification to ensure that the algorithms used have been validated and that the software correctly implements those algorithms.
- Consider the efficiency, scalability, and repeatability properties of software tools, and evaluate the accuracy of the tool's output.

Fig. 5. Advice for Practitioners: A Summary

2460 **References**

- 2461 [1] July 2020. URL: <https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-privacy-preserving-data-analysis-introduction-our>.
- 2462
- 2463 [2] 115th Congress (2017–2018). *Public Law 115-435: The Foundations for Evidence-Based Policymaking Act of 2018*. 2018. URL: <https://www.congress.gov/bill/115th-congress/house-bill/4174>.
- 2464
- 2465
- 2466 [3] *45 CFR 164 Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule Safe Harbor method Standard: De-identification of protected health information*.
- 2467
- 2468
- 2469 [4] *81 FR 49689: Revision of OMB Circular No. A-130, “Managing Information as a Strategic Resource*. July 2016. URL: <https://www.cio.gov/policies-and-priorities/circular-a-130/>.
- 2470
- 2471
- 2472 [5] 95th Congress. *Public Law 95-416*. Oct. 1978. URL: https://www.census.gov/history/pdf/NARA_Legislation.pdf.
- 2473
- 2474 [6] John Abowd et al. “The 2020 Census Disclosure Avoidance System TopDown Algorithm”. In: *Harvard Data Science Review* Special Issue 2 (June 2022). URL: <https://hdsr.mitpress.mit.edu/pub/7evz361i>.
- 2475
- 2476
- 2477 [7] John M. Abowd and Ian M. Schmutte. *Economic Analysis and Statistical Disclosure Limitation*. Mar. 2015. URL: <https://www.brookings.edu/bpea-articles/economic-analysis-and-statistical-disclosure-limitation/>.
- 2478
- 2479
- 2480 [8] John M. Abowd and Lars Vilhuber. “How Protective are Synthetic Data?” In: *Lecture Notes in Computer Science: Privacy in Statistical Databases 5262* (2008), pp. 239–246.
- 2481
- 2482
- 2483 [9] “Accuracy”. In: *Glossary of Statistical Terms* (Sept. 2001). Last accessed June 23, 2022. URL: <https://stats.oecd.org/glossary/detail.asp?ID=21>.
- 2484
- 2485 [10] Charu C. Aggarwal. “On K-Anonymity and the Curse of Dimensionality”. In: *Proceedings of the 31st International Conference on Very Large Data Bases. VLDB ’05*. Trondheim, Norway: VLDB Endowment, 2005, pp. 901–909. ISBN: 1595931546.
- 2486
- 2487
- 2488 [11] J. Trent Alexander, Michael Davern, and Betsey Stevenson. “Inaccurate age and sex data in the census PUMS files: Evidence and implications”. In: *Public Opinion Quarterly* 74 (3 2010), pp. 551–569. URL: <https://doi.org/10.1093/poq/nfq033>.
- 2489
- 2490
- 2491 [12] Micah Altman et al. “Towards a Modern Approach to Privacy-Aware Government Data Releases”. In: *Berkeley Technology Law Journal* 30 (3), pp. 1967–2072. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2779266.
- 2492
- 2493
- 2494 [13] AMD. *AMD Secure Encrypted Virtualization (SEV)*. Last accessed July 13, 2022. 2022. URL: <https://developer.amd.com/sev/>.
- 2495

- 2496 [14] Olivia Angiuli, Joe Blitzstein, and Jim Waldo. “How to De-Identify Your Data”.
2497 In: *Commun. ACM* 58.12 (Nov. 2015), pp. 48–55. ISSN: 0001-0782. DOI: [10.1145/](https://doi.org/10.1145/2814340)
2498 [2814340](https://doi.org/10.1145/2814340). URL: <https://doi.org/10.1145/2814340>.
- 2499 [15] ASTM International. *ASTM E1869-04 (Reapproved 2014) Standard Guide for Con-*
2500 *fidentiality, Privacy, Access, and Data Security Principles for Health Information*
2501 *Including Electronic Health Records*. 2014.
- 2502 [16] Daniel Barth-Jones. *The ‘Re-Identification’ of Governor William Weld’s Medical*
2503 *Information: A Critical Re-Examination of Health Data Identification Risks and*
2504 *Privacy Protections, Then and Now*. July 2012. URL: [https://ssrn.com/abstract=](https://ssrn.com/abstract=2076397)
2505 [2076397](https://ssrn.com/abstract=2076397)[%20or%20http://dx.doi.org/10.2139/ssrn.2076397](http://dx.doi.org/10.2139/ssrn.2076397).
- 2506 [17] Hans T Bjornsson et al. “Intra-individual Change Over Time in DNA Methylation
2507 with Familial Clustering”. In: *JAMA* 299 (24 June 2008), pp. 2877–2833. URL:
2508 <https://pubmed.ncbi.nlm.nih.gov/18577732/>.
- 2509 [18] Sylvia M. Burwell. *Open Data Policy-Managing Information as an Asset*. May
2510 2013. URL: [https://obamawhitehouse . archives . gov / sites / default / files / omb /](https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf)
2511 [memoranda/2013/m-13-13.pdf](https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf).
- 2512 [19] George Bush. *Executive Order 13402:Strengthening Federal Efforts to Protect Against*
2513 *Identity Theft*. May 2006. URL: [https://www.gpo.gov/fdsys/pkg/FR-2006-05-](https://www.gpo.gov/fdsys/pkg/FR-2006-05-15/pdf/06-4552.pdf)
2514 [15/pdf/06-4552.pdf](https://www.gpo.gov/fdsys/pkg/FR-2006-05-15/pdf/06-4552.pdf).
- 2515 [20] G. Camarillo, C. Holmberg, and Y. Gao. *Re-INVITE and Target-Refresh Request*
2516 *Handling in the Session Initiation Protocol (SIP)*. RFC 6141 (Proposed Standard).
2517 Internet Engineering Task Force, Mar. 2011. URL: www.ietf.org/rfc/rfc6141.txt.
- 2518 [21] David Carrell et al. “Hiding in plain sight: Use of realistic surrogates to reduce
2519 exposure of protected health information in clinical text”. In: *Journal of the Ameri-*
2520 *can Medical Informatics Association : JAMIA* 20 (2 July 2012), pp. 342–348. DOI:
2521 [10.1136/amiajnl-2012-001034](https://doi.org/10.1136/amiajnl-2012-001034).
- 2522 [22] Ann Cavoukian. *Privacy by Design: The 7 Foundational Principles*. Ontario, CA,
2523 Jan. 2011. URL: [https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.](https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf)
2524 [pdf](https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf).
- 2525 [23] Jennifer Cawthra et al. *Securing Telehealth Remote Patient Monitoring Ecosystem*.
2526 2022. DOI: [10.6028/NIST.SP.1800-30](https://doi.org/10.6028/NIST.SP.1800-30). URL: [https://nvlpubs.nist.gov/nistpubs/](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1800-30.pdf)
2527 [SpecialPublications/NIST.SP.1800-30.pdf](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1800-30.pdf).
- 2528 [24] Census Bureau Data Stewardship Program. *DS025: Organization of the Disclosure*
2529 *Review Board*. Dec. 2019. URL: [https://www2.census.gov/foia/ds_policies/ds025.](https://www2.census.gov/foia/ds_policies/ds025.pdf)
2530 [pdf](https://www2.census.gov/foia/ds_policies/ds025.pdf).
- 2531 [25] Malcolm Chisholm. “7 Phases of a Data Life Cycle”. In: *Information Manage-*
2532 *ment* (July 2015). URL: [http://www.information-management.com/news/data-](http://www.information-management.com/news/data-management/Data-Life-Cycle-Defined-10027232-1.html)
2533 [management/Data-Life-Cycle-Defined-10027232-1.html](http://www.information-management.com/news/data-management/Data-Life-Cycle-Defined-10027232-1.html).

- 2534 [26] A. Church. “A Note on the ‘Entscheidungsproblem’”. In: *Journal of Symbolic*
2535 *Logic* 1 (1936), pp. 40–41.
- 2536 [27] Commission on Evidence-Based Policymaking. *The Promise of Evidence-Based*
2537 *Policymaking*. Sept. 2017. URL: [https://www.acf.hhs.gov/opre/project/commission-](https://www.acf.hhs.gov/opre/project/commission-evidence-based-policymaking-cep)
2538 [evidence-based-policymaking-cep](https://www.acf.hhs.gov/opre/project/commission-evidence-based-policymaking-cep).
- 2539 [28] Tore Dalenius. “Finding a Needle in a Haystack, or Identifying Anonymous Census
2540 Records”. In: *Journal of Official Statistics* 2 (3 1986), pp. 329–336.
- 2541 [29] Tore Dalenius. “Towards a methodology for statistical disclosure control”. In: *Statis-*
2542 *tik Tidskrift* 15 (1977), pp. 429–444.
- 2543 [30] Jon P. Daries et al. “Privacy, Anonymity, and Big Data in the Social Sciences”. In:
2544 *Communications of the ACM* 57 (6 Sept. 2014), pp. 56–63.
- 2545 [31] Tanvi Desai, Felix Ritchie, and Richard Welpton. Economics Working Paper Series
2546 1601. 2016. URL: <http://dx.doi.org/10.13140/RG.2.1.3661.1604>.
- 2547 [32] DICOM Standards Committee. *DICOM PS3.15 2016e — Security and System Man-*
2548 *agement Profiles*. 2016. URL: [http://dicom.nema.org/medical/dicom/current/](http://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E)
2549 [output/html/part15.html#chapter_E](http://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E).
- 2550 [33] Irit Dinur and Kobbi Nissim. “Revealing Information While Preserving Privacy”.
2551 In: *Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Sympo-*
2552 *sium on Principles of Database Systems*. PODS ’03. San Diego, California: ACM,
2553 2003, pp. 202–210. ISBN: 1-58113-670-6. DOI: [10.1145/773153.773173](https://doi.org/10.1145/773153.773173). URL:
2554 [doi.acm.org/10.1145/773153.773173](https://doi.org/10.1145/773153.773173).
- 2555 [34] Changyu Dong, Liqun Chen, and Zikai Wen. “When Private Set Intersection Meets
2556 Big Data: An Efficient and Scalable Protocol”. In: *Proceedings of the 2013 ACM*
2557 *SIGSAC Conference on Computer and Communications Security*. CCS ’13. Berlin,
2558 Germany: Association for Computing Machinery, 2013, pp. 789–800. ISBN: 9781450324779.
2559 DOI: [10.1145/2508859.2516701](https://doi.org/10.1145/2508859.2516701). URL: <https://doi.org/10.1145/2508859.2516701>.
- 2560 [35] Jörg Drechsler, Stefan Bender, and Susanne Rässler. *Comparing fully and partially*
2561 *synthetic datasets for statistical disclosure control in the German IAB Establish-*
2562 *ment Panel (Working paper 11)*. New York, 2007. URL: [http://fdz.iab.de/342/](http://fdz.iab.de/342/section.aspx/Publikation/k080530j05)
2563 [section.aspx/Publikation/k080530j05](http://fdz.iab.de/342/section.aspx/Publikation/k080530j05).
- 2564 [36] George T. Duncan, Mark Elliot, and Juan-José Salazar-Gonzalez. *Statistical Confi-*
2565 *dentiality: Principles and Practice*. Springer, 2011, p. 113.
- 2566 [37] I. Dunsford et al. “A human blood-group chimera”. In: *British Medical Journal* 81
2567 (July 1953). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2028470/>.
- 2568 [38] Cynthia Dwork and Aaron Roth. “The Algorithmic Foundations of Differential Pri-
2569 vacy”. In: *Foundations and Trends in Theoretical Computer Science*. Vol. 9. 3–4.
2570 NOW, 2014, pp. 211–407.

- 2571 [39] Cynthia Dwork et al. “Calibrating Noise to Sensitivity in Private Data Analysis”.
2572 In: *Theory of Cryptography*. Ed. by Shai Halevi and Tal Rabin. Berlin, Heidelberg:
2573 Springer Berlin Heidelberg, 2006, pp. 265–284. ISBN: 978-3-540-32732-5.
- 2574 [40] Cynthia Dwork et al. “Calibrating Noise to Sensitivity in Private Data Analysis”.
2575 In: *Proceedings of the Third Conference on Theory of Cryptography*. TCC’06. New
2576 York, NY: Springer-Verlag, 2006, pp. 265–284. ISBN: 3540327312. DOI: [10.1007/
2577 11681878_14](https://doi.org/10.1007/11681878_14). URL: doi.org/10.1007/11681878_14.
- 2578 [41] Department of Education (ED) Disclosure Review Board (DRB). *The Data Disclo-*
2579 *sure Decision*. Version 1.0. 2015. URL: [https://s3.amazonaws.com/sitesusa/wp-
2580 content/uploads/sites/1151/2016/10/The-Data-Disclosure-Decision-Department-
2581 of-Education-Case-Study_Mar-2015.pdf](https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/1151/2016/10/The-Data-Disclosure-Decision-Department-of-Education-Case-Study_Mar-2015.pdf).
- 2582 [42] Mark Elliot and Angela Dale. *Scenarios of attack: the data intruder’s perspective*
2583 *on statistical disclosure risk*. Spring 1999.
- 2584 [43] El Emam. *Methods for the de-identification of electronic health records for genomic*
2585 *research*. 2011. URL: [https://genomemedicine.biomedcentral.com/articles/10.
2586 1186/gm239](https://genomemedicine.biomedcentral.com/articles/10.1186/gm239).
- 2587 [44] K. El Emam and B. Malin. “Appendix B: Concepts and Methods for De-Identifying
2588 Clinical Trial Data”. In: *Sharing Clinical Trial Data: Maximizing Benefits, Mini-*
2589 *mitizing Risk*. Washington, DC: Institute of Medicine of the National Academies,
2590 The National Academies Press, 2015.
- 2591 [45] Khaled El Emam and Luk Arbuckle. *Anonymizing Health Data: Case Studies and*
2592 *Methods to Get you Started*. Sebastopol, CA: O’Reilly Media, 2013.
- 2593 [46] Ali Farzanehfar, Florimond Houssiau, and Yves-Alexandre de Montjoye. “The risk
2594 of re-identification remains high even in country-scale location datasets”. In: *Pat-*
2595 *terns* 2.3 (2021), p. 100204. ISSN: 2666-3899. DOI: [https://doi.org/10.1016/j.
2596 patter.2021.100204](https://doi.org/10.1016/j.patter.2021.100204). URL: [https://www.sciencedirect.com/science/article/pii/
2597 S2666389921000143](https://www.sciencedirect.com/science/article/pii/S2666389921000143).
- 2598 [47] Confidentiality and Data Access Committee. *Statistical Policy Working Paper 22:*
2599 *Report on Statistical Disclosure Limitation Methodology*. Tech. rep. Federal Com-
2600 mittee on Statistical Methodology, 2005. URL: [https://www.hhs.gov/sites/default/
2601 files/spwp22.pdf](https://www.hhs.gov/sites/default/files/spwp22.pdf).
- 2602 [48] Federal Committee on Statistical Methodology. *Data Protection Toolkit*. Sept. 2020.
2603 URL: <https://nces.ed.gov/fcsm/dpt>.
- 2604 [49] Matthew Fredrikson et al. “Privacy in Pharmacogenetics: An End-to-End Case
2605 Study of Personalized Warfarin Dosing”. In: *23rd USENIX Security Symposium*.
2606 San Diego, CA. URL: [https://www.usenix.org/system/files/conference/usenixsecurity14/
2607 sec14-paper-fredrikson-privacy.pdf](https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-fredrikson-privacy.pdf).

- 2608 [50] Simson Garfinkel. *De-Identification of Personally Identifiable Information*. Tech.
2609 rep. NIST IR 8053. National Institute of Science and Technology, Nov. 2015. URL:
2610 <http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf>.
- 2611 [51] Simson L. Garfinkel. *De-identification of personal information*. 2015. DOI: 10.
2612 6028/NIST.IR.8053. URL: [https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.](https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf)
2613 [pdf](https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf).
- 2614 [52] Simson L. Garfinkel. *Government Data De-Identification Stakeholder's Meeting*
2615 *June 29, 2016 Meeting Report*. 2016. DOI: 10.6028/NIST.IR.8150. URL: [https:](https://nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8150.pdf)
2616 [//nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8150.pdf](https://nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8150.pdf).
- 2617 [53] Genetics Home Reference. *What are single nucleotide polymorphisms (SNPs)?*
2618 Last access June 16, 2022. 2022. URL: [https://ghr.nlm.nih.gov/primer/genomicresearch/](https://ghr.nlm.nih.gov/primer/genomicresearch/snp)
2619 [snp](https://ghr.nlm.nih.gov/primer/genomicresearch/snp).
- 2620 [54] Genetics Home Reference. *What is DNA*. Last access June 16, 2022. 2022. URL:
2621 <https://ghr.nlm.nih.gov/primer/basics/dna>.
- 2622 [55] Craig Gentry. "A Fully Homomorphic Encryption Scheme". AAI3382729. PhD
2623 thesis. Stanford, CA, USA, 2009. ISBN: 9781109444506.
- 2624 [56] Ruobin Gong, Erica L. Groshen, and Salil Vadhan. "Harnessing the Known Un-
2625 knowns: Differential Privacy and the 2020 Census". In: *Harvard Data Science Re-*
2626 *view* Special Issue 2 (June 2022). URL: <https://hdsr.mitpress.mit.edu/pub/fgyf5cne>.
- 2627 [57] Melissa Gymrek et al. "Identifying Personal Genomes by Surname Inference". In:
2628 *Science* 339 (6117 Jan. 2013), pp. 321–329.
- 2629 [58] Michael B. Hawes. "Implementing Differential Privacy: Seven Lessons From the
2630 2020 United States Census". In: *Harvard Data Science Review* 2.2 (Apr. 2020).
2631 URL: <https://hdsr.mitpress.mit.edu/pub/dgg03vo6>.
- 2632 [59] TN Herzog, FJ Scheuren, and WE Winkler. *Data Quality and Record Linkage Tech-*
2633 *niques*. New York/London: Springer, 2007.
- 2634 [60] Vagelis Hristidis, ed. *Information Discovery on Electronic Health Records*. 1st.
2635 Chapman and Hall/CRC, 2009. ISBN: 1420090380. URL: [https://doi.org/10.1201/](https://doi.org/10.1201/9781420090413)
2636 [9781420090413](https://doi.org/10.1201/9781420090413).
- 2637 [61] IHE IT Infrastructure Technical Committee. *IHE IT Infrastructure Handbook: De-*
2638 *Identification*. Integrating the Healthcare Enterprise, Mar. 2014. URL: [https://ihe.](https://ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_Handbook_De-Identification_Rev1.0_2014-03-14.pdf)
2639 [net/uploadedFiles/Documents/ITI/IHE_ITI_Handbook_De-](https://ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_Handbook_De-Identification_Rev1.0_2014-03-14.pdf)
2640 [Identification_Rev1.](https://ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_Handbook_De-Identification_Rev1.0_2014-03-14.pdf)
[0_2014-03-14.pdf](https://ihe.net/uploadedFiles/Documents/ITI/IHE_ITI_Handbook_De-Identification_Rev1.0_2014-03-14.pdf).
- 2641 [62] Clay Johnson III. *OMB Memorandum M-07-16: Safeguarding Against and Re-*
2642 *sponding to the Breach of Personally Identifiable Information*. May 2007. URL:
2643 [https://georgewbush-whitehouse.archives.gov/omb/memoranda/fy2007/m07-](https://georgewbush-whitehouse.archives.gov/omb/memoranda/fy2007/m07-16.pdf)
2644 [16.pdf](https://georgewbush-whitehouse.archives.gov/omb/memoranda/fy2007/m07-16.pdf).

- 2645 [63] Information Commissioner’s Office. *Anonymisation: code of practice, managing*
2646 *data protection risk*. 2012. URL: [https://ico.org.uk/media/1061/anonymisation-](https://ico.org.uk/media/1061/anonymisation-code.pdf)
2647 [code.pdf](https://ico.org.uk/media/1061/anonymisation-code.pdf).
- 2648 [64] *ISO 26324:2012, Information and documentation – Digital object identifier system*.
2649 Geneva, Switzerland, 2012. URL: <https://www.iso.org/standard/43506.html>.
- 2650 [65] *ISO/IEC 24760-1:2011, Information technology – Security techniques – A frame-*
2651 *work for identity management – Part 1: Terminology and concepts*. 2011.
- 2652 [66] *ISO/TS 25237:2008(E) Health Informatics — Pseudonymization*. Geneva, Switzer-
2653 land, 2008.
- 2654 [67] Auguste Kerckhoffs. “II. Desiderata De La Cryptographie Militaire”. In: *Journal*
2655 *des sciences militaires* IX (Jan. 1883), pp. 5–38.
- 2656 [68] Shehab Khan. ““Human chimera”: Man fails paternity test because genes in his
2657 saliva are different to those in sperm”. In: *The Independent* (Oct. 2015).
- 2658 [69] Vladimir Kolesnikov et al. “Efficient Batched Oblivious PRF with Applications to
2659 Private Set Intersection”. In: *Proceedings of the 2016 ACM SIGSAC Conference on*
2660 *Computer and Communications Security*. CCS ’16. Vienna, Austria: Association
2661 for Computing Machinery, 2016, pp. 818–829. ISBN: 9781450341394. DOI: [10.](https://doi.org/10.1145/2976749.2978381)
2662 [1145/2976749.2978381](https://doi.org/10.1145/2976749.2978381). URL: <https://doi.org/10.1145/2976749.2978381>.
- 2663 [70] Leah Krehling. *De-Identification Guideline*. Tech. rep. WL-2020-01. Department
2664 of Electrical and Computer Engineering, Western University, 2020, p. 45.
- 2665 [71] Sandra Lechner and Winfried Pohlmeier. “To Blank or Not to Blank? A Compari-
2666 son of the Effects of Disclosure Limitation Methods on Nonlinear Regression Es-
2667 timates”. In: *Privacy in Statistical Databases, Lecture Notes in Computer Science*
2668 *3050* (2004), pp. 187–200.
- 2669 [72] Jaewoo Lee and Chris Clifton. “How Much Is Enough? Choosing ϵ for Differential
2670 Privacy”. In: *Information Security*. Ed. by Xuejia Lai, Jianying Zhou, and Hui Li.
2671 Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 325–340. ISBN: 978-3-
2672 642-24861-0.
- 2673 [73] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. “t-Closeness: Privacy
2674 Beyond k-Anonymity and l-Diversity”. In: *2007 IEEE 23rd International Confer-*
2675 *ence on Data Engineering*. 2007, pp. 106–115. DOI: [10.1109/ICDE.2007.367856](https://doi.org/10.1109/ICDE.2007.367856).
- 2676 [74] Yehuda Lindell. “Secure Multiparty Computation”. In: *Commun. ACM* 64.1 (Dec.
2677 2020), pp. 86–96. ISSN: 0001-0782. DOI: [10.1145/3387108](https://doi.org/10.1145/3387108). URL: [https://doi.org/](https://doi.org/10.1145/3387108)
2678 [10.1145/3387108](https://doi.org/10.1145/3387108).
- 2679 [75] M. Altman M et al. “Towards a Modern Approach to Privacy-Aware Government
2680 Data Release”. In: *Berkeley Journal of Technology Law Internet* (2016). URL: [https:](https://lawcat.berkeley.edu/record/1127405?ln=en)
2681 [//lawcat.berkeley.edu/record/1127405?ln=en](https://lawcat.berkeley.edu/record/1127405?ln=en).

- 2682 [76] Ashwin Machanavajjhala et al. “l-diversity: Privacy beyond k-anonymity”. In: *Proc.*
2683 *22nd Intl. Conf. Data Engg. (ICDE)*. 2006.
- 2684 [77] Sean Martin. *When De-identifying Patient Information, Follow the HITRUST Frame-*
2685 *work*. Sept. 2016. URL: [https://hitrustalliance.net/de-identifying-patient-information-](https://hitrustalliance.net/de-identifying-patient-information-follow-hitrust-framework/)
2686 [follow-hitrust-framework/](https://hitrustalliance.net/de-identifying-patient-information-follow-hitrust-framework/).
- 2687 [78] Daniel A. Mayer et al. “Implementation and Performance Evaluation of Privacy-
- 2688 Preserving Fair Reconciliation Protocols on Ordered Sets”. In: *Proceedings of the*
2689 *First ACM Conference on Data and Application Security and Privacy*. CODASPY
2690 ’11. San Antonio, TX, USA: Association for Computing Machinery, 2011, pp. 109–
2691 120. ISBN: 9781450304665. DOI: [10.1145/1943513.1943529](https://doi.org/10.1145/1943513.1943529). URL: [https://doi.org/](https://doi.org/10.1145/1943513.1943529)
2692 [10.1145/1943513.1943529](https://doi.org/10.1145/1943513.1943529).
- 2693 [79] E McCallister, T Grance, and K A Scarfone. *Guide to protecting the confiden-*
2694 *tiality of Personally Identifiable Information (PII)*. Gaithersburg, MD, 2010. DOI:
2695 [10.6028/NIST.SP.800-122](https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf). URL: [https://nvlpubs.nist.gov/nistpubs/Legacy/SP/](https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf)
2696 [nistspecialpublication800-122.pdf](https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-122.pdf).
- 2697 [80] William K. Michener et al. “Participatory design of DataONE—Enabling cyber-
- 2698 infrastructure for the biological and environmental sciences”. In: *Ecological In-*
2699 *formatics* 11 (2012). Data platforms in integrative biodiversity research, pp. 5–15.
2700 ISSN: 1574-9541. DOI: <https://doi.org/10.1016/j.ecoinf.2011.08.007>. URL:
2701 <https://www.sciencedirect.com/science/article/pii/S1574954111000768>.
- 2702 [81] Yves-Alexandre de Montjoye et al. “Unique in the Crowd: The Privacy Bounds of
- 2703 Human Mobility”. In: *Nature Scientific Reports* 3 (1376 2013).
- 2704 [82] Yves-Alexandre de Montjoye et al. “Unique in the Shopping Mall: On the Reiden-
- 2705 tifiability of Credit Card Metadata”. In: *Science* 347 (536 2015).
- 2706 [83] Arvind Narayanan and Ed Felten. *No silver bullet: De-identification still doesn’t*
2707 *work*. Working Paper. July 2014. URL: [http://randomwalker.info/publications/no-](http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf)
2708 [silver-bullet-de-identification.pdf](http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf).
- 2709 [84] Arvind Narayanan and Vitaly Shmatikov. “Robust De-anonymization of Large Sparse
- 2710 Datasets”. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 2008,
2711 pp. 111–125. DOI: [10.1109/SP.2008.33](https://doi.org/10.1109/SP.2008.33).
- 2712 [85] *NIST Big Data Interoperability Framework: volume 1, definitions, version 2*. Gaithers-
- 2713 burg, MD, 2018. DOI: [10.6028/NIST.SP.1500-1r1](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1r1.pdf). URL: [https://nvlpubs.nist.gov/](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1r1.pdf)
2714 [nistpubs/SpecialPublications/NIST.SP.1500-1r1.pdf](https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-1r1.pdf).
- 2715 [86] *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk*
2716 *Management, Version 1.0*. Gaithersburg, MD, 2022. DOI: [10.6028/NIST.CSWP.10](https://doi.org/10.6028/NIST.CSWP.10).
2717 URL: <https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf>.

- 2718 [87] Christine M. O’Keefe and James O. Chipperfield. “A Summary of Attack Meth-
2719 ods and Confidentiality Protection Measures for Fully Automated Remote Analysis
2720 Systems”. In: *International Statistical Review / Revue Internationale de Statistique*
2721 81.3 (2013), pp. 426–455. ISSN: 03067734, 17515823. URL: [http://www.jstor.org/
2722 stable/43299645](http://www.jstor.org/stable/43299645) (visited on 06/28/2022).
- 2723 [88] Barack Obama. *Executive Order 13642—Making Open and Machine Readable the*
2724 *New Default for Government Information*. May 2013. URL: [https://obamawhitehouse.
2725 archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-
2726 machine-readable-new-default-government-](https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-).
- 2727 [89] Office of Civil Rights, US Department of Health and Human Services. *Guidance*
2728 *Regarding Methods for De-identification of Protected Health Information in Ac-*
2729 *cordance with the Health Insurance Portability and Accountability Act (HIPAA)*
2730 *Privacy Rule*. Nov. 2012. URL: [http://www.hhs.gov/hipaa/for-professionals/
2731 privacy/special-topics/de-identification/](http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/).
- 2732 [90] Office of Civil Rights, US Department of Health and Human Services. *Individuals’*
2733 *Right under HIPAA to Access their Health Information 45 CFR § 164.524*. Last
2734 accessed June 17, 2022. 2022. URL: [https://www.hhs.gov/hipaa/for-professionals/
2735 privacy/guidance/access/index.html](https://www.hhs.gov/hipaa/for-professionals/privacy/guidance/access/index.html).
- 2736 [91] Office of Management and Budget. *Circular A110 Revised 11/19/93, as further*
2737 *amended 9/30/99*. URL: [https://obamawhitehouse.archives.gov/omb/circulars-
2738 a110/](https://obamawhitehouse.archives.gov/omb/circulars-a110/).
- 2739 [92] Office of Management and Budget. *Statistical Programs and Standards*. Last ac-
2740 cessed July 15, 2022. 2022. URL: [https://www.whitehouse.gov/omb/information-
2741 regulatory-affairs/statistical-programs-standards/](https://www.whitehouse.gov/omb/information-regulatory-affairs/statistical-programs-standards/).
- 2742 [93] Office of Safeguards, US Internal Revenue Service. *Publication 1075: Tax Infor-*
2743 *mation Security Guidelines For Federal, State and Local Agencies*. 2021. URL:
2744 <https://www.irs.gov/pub/irs-pdf/p1075.pdf>.
- 2745 [94] Paul Ohm. “Broken Promises of Privacy: Responding to the Surprising Failure of
2746 Anonymization”. In: *UCLA Law Review* 57 (July 2012), pp. 1701–1778.
- 2747 [95] *OHRP-Guidance on Research Involving Private Information or Biological Spec-*
2748 *imens*. Aug. 2008. URL: <http://www.hhs.gov/ohrp/policy/cdebiol.html>.
- 2749 [96] Joanne Pascale et al. *Issue Paper on Disclosure Review for Information Products*
2750 *with Qualitative Research Findings*. Mar. 2020. URL: [https://www.census.gov/
2751 library/working-papers/2020/adrm/rsm2020-01.html](https://www.census.gov/library/working-papers/2020/adrm/rsm2020-01.html).
- 2752 [97] Joanne Pascale et al. “Protecting the Identity of Participants in Qualitative Re-
2753 search”. In: *Journal of Survey Statistics and Methodology* 10.3 (Jan. 2022), pp. 549–
2754 567. ISSN: 2325-0984. DOI: 10.1093/jssam/smab048. eprint: [https://academic.oup.
2755 com/jssam/article-pdf/10/3/549/44275508/smab048.pdf](https://academic.oup.com/jssam/article-pdf/10/3/549/44275508/smab048.pdf). URL: [https://doi.org/10.
2756 1093/jssam/smab048](https://doi.org/10.1093/jssam/smab048).

- 2757 [98] Andrew Peterson. “Why the names of six people who complained of sexual assault
2758 were published online by Dallas police”. In: *The Washington Post* (Apr. 2016).
2759 URL: [https://www.washingtonpost.com/news/the-switch/wp/2016/04/29/why-the-
2760 names-of-six-people-who-complained-of-sexual-assault-were-published-online-
2761 by-dallas-police/](https://www.washingtonpost.com/news/the-switch/wp/2016/04/29/why-the-names-of-six-people-who-complained-of-sexual-assault-were-published-online-by-dallas-police/).
- 2762 [99] Thomas Piketty and Emmanuel Saez. “Income Inequality in the United States 1913-
2763 1998”. In: *Quarterly Journal of Economics* 118 (1 2003), pp. 1–41.
- 2764 [100] “Pillar Investigates: USCCB gen sec Burrill resigns after sexual misconduct alle-
2765 gations”. In: *The Pillar* (July 2021). URL: [https://www.pillarcatholic.com/p/pillar-
2766 investigates-usccb-gen-sec](https://www.pillarcatholic.com/p/pillar-investigates-usccb-gen-sec).
- 2767 [101] Sandro Pinto and Nuno Santos. “Demystifying Arm TrustZone: A Comprehensive
2768 Survey”. In: *ACM Comput. Surv.* 51.6 (Jan. 2019). ISSN: 0360-0300. DOI: [10.1145/
2769 3291047](https://doi.org/10.1145/3291047). URL: <https://doi.org/10.1145/3291047>.
- 2770 [102] *Private Lives and Public Policies: Confidentiality and Accessibility of Government
2771 Statistics*. Panel on Confidentiality and Data Access, National Research Council,
2772 p. 288. ISBN: 0-309-57611-3. URL: <http://www.nap.edu/catalog/2122/>.
- 2773 [103] *Public Law 93-579: The Privacy Act*. 88 Stat. 1896, 5 U.S.C. § 552a.
- 2774 [104] Balaji Raghunathan. *The Complete Book of Data Anonymization: From Planning
2775 to Implementation*. USA: Auerbach Publications, 2013. ISBN: 1439877300.
- 2776 [105] William H. Rehnquist. *Department of State v. Washington Post Co.*, 456 U.S. 595
2777 (1982). 1982. URL: <https://www.loc.gov/item/usrep456595/>.
- 2778 [106] *Report 08-536, Privacy: Alternatives Exist for Enhancing Protection of Personally
2779 Identifiable Information*. May 2008. URL: [http://www.gao.gov/new.items/d08536.
2780 pdf](http://www.gao.gov/new.items/d08536.pdf).
- 2781 [107] Diane Ridgeway et al. *Challenge Design and Lessons Learned from the 2018 Dif-
2782 ferential Privacy Challenges*. 2021. DOI: [10.6028/NIST.TN.2151](https://doi.org/10.6028/NIST.TN.2151). URL: [https:
2783 //nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.2151.pdf](https://nvlpubs.nist.gov/nistpubs/TechnicalNotes/NIST.TN.2151.pdf).
- 2784 [108] Pierangela Samarati and Latanya Sweeney. “Protecting privacy when disclosing
2785 information: k-anonymity and its enforcement through generalization and suppres-
2786 sion”. In: *Proceedings of the IEEE Symposium on Research in Security and Privacy*
2787 (May 1998).
- 2788 [109] Pierangela Samarti. “Protecting Respondents’ Identities in Microdata Release”.
2789 In: *IEEE Transactions on Knowledge and Data Engineering* 13 (6 Nov. 2001),
2790 pp. 1010–1027.
- 2791 [110] Josep Sanz and Josep Domingo-Ferrer. “A Comparative Study of Microaggrega-
2792 tion Methods”. In: *Questiio: Quaderns d’Estadística, Sistemes, Informàtica i In-
2793 vestigació Operativa* 22 (3 Aug. 2000), pp. 511–526.

- 2794 [111] M Scaiano et al. “unified framework for evaluating the risk of re-identification of
2795 text de-identification tools”. In: *Journal of Biomedical Informatics* 63 (Oct. 2016),
2796 pp. 174–183.
- 2797 [112] Matthias Schunter. “Intel Software Guard Extensions: Introduction and Open Re-
2798 search Challenges”. In: *Proceedings of the 2016 ACM Workshop on Software PRO-*
2799 *tection*. SPRO ’16. Vienna, Austria: Association for Computing Machinery, 2016,
2800 p. 1. ISBN: 9781450345767. DOI: [10.1145/2995306.2995307](https://doi.org/10.1145/2995306.2995307). URL: [https://doi.org/](https://doi.org/10.1145/2995306.2995307)
2801 [10.1145/2995306.2995307](https://doi.org/10.1145/2995306.2995307).
- 2802 [113] M Seastrom. “Licensing”. In: *Confidentiality, Disclosure and Data Access: Theory*
2803 *and Practical Application for Statistical Agencies*. Ed. by P. Doyle et al. Elsevier
2804 Science, 2001.
- 2805 [114] Jordi Soria-Comas and Josep Domingo-Ferrer. “Connecting privacy models: syn-
2806 ergies between k-anonymity, t-closeness, and differential privacy”. In: Working
2807 Paper (English Only). Ottawa, Canada, Oct. 2013. URL: [https://www.unece.](https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_2_soria-comas_domingo-ferrer.pdf)
2808 [org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_2_soria-](https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_2_soria-comas_domingo-ferrer.pdf)
2809 [comas_domingo-ferrer.pdf](https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_2_soria-comas_domingo-ferrer.pdf).
- 2810 [115] Philip Steel and Jon Sperling. *The Impact of Multiple Geographies and Geographic*
2811 *Detail on Disclosure Risk: Interactions between Census Tract and ZIP Code Tab-*
2812 *ulation Geography*. 2001. URL: [https://www.census.gov/content/dam/Census/](https://www.census.gov/content/dam/Census/library/working-papers/2001/adrm/steel-sperling-2001.pdf)
2813 [library/working-papers/2001/adrm/steel-sperling-2001.pdf](https://www.census.gov/content/dam/Census/library/working-papers/2001/adrm/steel-sperling-2001.pdf).
- 2814 [116] Jackson M. Steinkamp et al. “Evaluation of Automated Public De-Identification
2815 Tools on a Corpus of Radiology Reports”. In: *Radiol Artif Intell* 2 (6 Oct. 2020).
2816 URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8082401/>.
- 2817 [117] John Paul Stevens. *U.S. Dept. of Justice v. Reporters Committee For Freedom of*
2818 *Press*, 489 U.S. 749 (1989). 1988. URL: <https://www.loc.gov/item/usrep489749/>.
- 2819 [118] John Paul Stevens. *United States Department of State v. Ray et al.*, 502 U.S. 164
2820 (1991). 1991. URL: <https://www.loc.gov/item/usrep502164/>.
- 2821 [119] Kevin Stine et al. *Volume I: guide for mapping types of information and infor-*
2822 *mation systems to security categories*. Gaithersburg, MD, 2008. DOI: [10.6028/](https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-60v1r1.pdf)
2823 [NIST.SP.800-60v1r1](https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-60v1r1.pdf). URL: [https://nvlpubs.nist.gov/nistpubs/Legacy/SP/](https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-60v1r1.pdf)
2824 [nistspecialpublication800-60v1r1.pdf](https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-60v1r1.pdf).
- 2825 [120] Tim Stobierski. In: *Business Insights* (Feb. 2021). URL: [https://online.hbs.edu/](https://online.hbs.edu/blog/post/data-life-cycle)
2826 [blog/post/data-life-cycle](https://online.hbs.edu/blog/post/data-life-cycle).
- 2827 [121] Teresa A. Sullivan. “Coming to Our Census: How Social Statistics Underpin Our
2828 Democracy (and Republic)”. In: *Harvard Data Science Review* 2.1 (Jan. 2020).
2829 URL: <https://hdsr.mitpress.mit.edu/pub/1g1cbvkx>.
- 2830 [122] Latanya Sweeney. “k-anonymity: a model for protecting privacy”. In: *International*
2831 *Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10 (5 2002),
2832 pp. 557–570.

- 2833 [123] Latanya Sweeney. “*k*-anonymity: a model for protecting privacy”. In: *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10 (5 Oct. 2002), pp. 557–570. URL: <http://dx.doi.org/10.1142/S0218488502001648>.
- 2834
- 2835
- 2836 [124] Latanya Sweeney. “Weaving Technology and Policy Together to Maintain Confidentiality”. In: *Journal of Law, Medicine and Ethics* 25 (1997), pp. 98–110.
- 2837
- 2838 [125] “The Debate Over ‘Re-Identification’ Of Health Information: What Do We Risk?”. In: *Health Affairs Blog* (Aug. 2012). DOI: [10.1377/hblog20120810.021952](https://doi.org/10.1377/hblog20120810.021952). URL: <https://www.healthaffairs.org/doi/10.1377/forefront.20120810.021952>.
- 2839
- 2840
- 2841 [126] *Title V of the E-Government Act of 2002: Confidential Information Protection and Statistical Efficiency Act (CIPSEA) PL 107–347, 116 Stat. 2899, 44 USC § 101 Section 502(8)*.
- 2842
- 2843
- 2844 [127] TransCelerate Biopharma, Inc. *Data De-identification and Anonymization of Individual Patient Data in Clinical Studies—A Model Approach*. 2013.
- 2845
- 2846 [128] Michael Carl Tschantz and Jeannette M. Wing. *Formal Methods for Privacy*. Tech. rep. CMU-CS-09-154. Pittsburg, PA: Carnegie Mellon University, Aug. 2009. URL: <http://reports-archive.adm.cs.cmu.edu/anon/2009/CMU-CS-09-154.pdf>.
- 2847
- 2848
- 2849 [129] A. M. Turing. “On Computable Numbers, with an Application to the Entscheidungsproblem”. In: *Proceedings of the London Mathematical Society, Series 2* (42 1936–37), pp. 230–265.
- 2850
- 2851
- 2852 [130] US Census Bureau. *Census Confidentiality and Privacy: 1790-2002*. 2003. URL: <https://www.census.gov/prod/2003pubs/conmono2.pdf>.
- 2853
- 2854 [131] US Census Bureau. *The “72-Year Rule”*. Jan. 2022. URL: https://www.census.gov/history/www/genealogy/decennial_census_records/the_72_year_rule_1.html.
- 2855
- 2856 [132] US Congress. *Public Law 104-191: Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. Aug. 1996. URL: <https://www.congress.gov/bill/104th-congress/house-bill/3103>.
- 2857
- 2858
- 2859 [133] US Congress. *Public Law 114-185: FOIA Improvement Act of 2016*. 2016. URL: <https://www.congress.gov/114/plaws/publ185/PLAW-114publ185.pdf>.
- 2860
- 2861 [134] US Congress. *The Freedom Of Information Act, 5 U.S.C. § 552*. 2022. URL: <https://www.justice.gov/oip/freedom-information-act-5-usc-552>.
- 2862
- 2863 [135] US Department of Health and Human Services. *45 CFR Part 46: Federal Policy for the Protection of Human Subjects*. Jan. 2017. URL: <https://www.govinfo.gov/content/pkg/FR-2017-01-19/pdf/2017-01058.pdf>.
- 2864
- 2865
- 2866 [136] US Department of Health and Human Services. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. 2012. URL: <https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html>.
- 2867
- 2868
- 2869
- 2870

- 2871 [137] *Joel Havermann, plaintiff—Appellant v. Carolyn W. Colvin, Acting Commissioner*
2872 *of the Social Security Administration, Defendant—Appellee, No. 12-2453, US Court*
2873 *of Appeals for the Fourth Circuit, 537 Fed. Appx. 142; 2013 US App. Aug 1,*
2874 *2013. Joel Havemann v. Carolyn W. Colvin, Civil No. JFM-12-1325. 2015 US Dist.*
2875 *LEXIS 27560. Mar. 2015.*
- 2876 [138] “Utility”. In: *Glossary of Statistical Terms* (Aug. 2002). Last accessed June 23,
2877 2022. URL: <https://stats.oecd.org/glossary/detail.asp?ID=4884>.
- 2878 [139] Russell T. Vought. *Phase 1 Implementation of the Foundations for Evidence-Based*
2879 *Policymaking Act of 2018: Learning Agendas, Personnel, and Planning Guidance.*
2880 URL: <https://www.whitehouse.gov/wp-content/uploads/2019/07/M-19-23.pdf>.
- 2881 [140] Charlie Warzel and Stuart A. Thompson. “How Your Phone Betrays Democracy”.
2882 In: *The New York Times* (Dec. 2019). URL: [https://www.nytimes.com/interactive/](https://www.nytimes.com/interactive/2019/12/21/opinion/location-data-democracy-protests.html)
2883 [2019/12/21/opinion/location-data-democracy-protests.html](https://www.nytimes.com/interactive/2019/12/21/opinion/location-data-democracy-protests.html).
- 2884 [141] Cathy Wasserman and Eric Ossiander. *Department of Health Agency Standards for*
2885 *Reporting Data with Small Numbers.* May 2018. URL: [https://doh.wa.gov/sites/](https://doh.wa.gov/sites/default/files/legacy/Documents/1500/SmallNumbers.pdf)
2886 [default/files/legacy/Documents/1500/SmallNumbers.pdf](https://doh.wa.gov/sites/default/files/legacy/Documents/1500/SmallNumbers.pdf).
- 2887 [142] Leon Willenborg and Ton de Waal. “Chapter 3 Data Analytic Impact of SDC
2888 Techniques on Microdata”. In: *Elements of Statistical Disclosure Control* (2012),
2889 pp. 72–92.
- 2890 [143] Li Xiong et al. “Privacy-Preserving Information Discovery on EHRs”. In: *Informa-*
2891 *tion Discovery on Electronic Health Records.* Ed. by Vagelis Hristidis. 1st. Chap-
2892 man and Hall/CRC, 2009. ISBN: 1420090380. URL: [https://doi.org/10.1201/](https://doi.org/10.1201/9781420090413)
2893 [9781420090413](https://doi.org/10.1201/9781420090413).

2894 **Appendix A. Standards**

- 2895 • ASTM E1869-04(2014) Standard Guide for Confidentiality, Privacy, Access, and
2896 Data Security Principles for Health Information Including Electronic Health Records.
- 2897 • DICOM PS3.15 2016d – Security and System Management Profiles Chapter E At-
2898 tribute Confidentiality Profiles, DICOM Standards Committee, NEMA 2016. [http:](http://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E)
2899 [//dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E](http://dicom.nema.org/medical/dicom/current/output/html/part15.html#chapter_E)
- 2900 • HITRUST De-Identification Working Group (2015, March). De-Identification Frame-
2901 work: A Consistent, Managed Methodology for the De-Identification of Personal
2902 Data and the Sharing of Compliance and Risk Information. Frisco, TX: HITRUST.
2903 Retrieved from <https://hitrustalliance.net/de-identification-license-agreement/>
- 2904 • ISO 8000-2:2012(E) Data quality – Part 2: Vocabulary, 2012. ISO, Geneva, Switzer-
2905 land. 2012.
- 2906 • ISO/IEC 27000:2014 Information technology -- Security techniques -- Information
2907 security management systems -- Overview and vocabulary. ISO, Geneva, Switzer-
2908 land. 2012.
- 2909 • ISO/IEC 24760-1:2011 Information technology -- Security techniques -- A frame-
2910 work for identity management -- Part 1: Terminology and concepts. ISO, Geneva,
2911 Switzerland. 2011.
- 2912 • ISO/TS 25237:2008(E) Health Informatics – Pseudonymization. ISO, Geneva, Switzer-
2913 land. 2008.
- 2914 • ISO/IEC 20889 WORKING DRAFT 2016-05-30, Information technology – Secu-
2915 rity techniques – Privacy enhancing data de-identification techniques. ISO, Geneva,
2916 Switzerland. 2016.
- 2917 • IHE IT Infrastructure Handbook, De-Identification, Integrating the Healthcare Enter-
2918 prise, June 6, 2014. http://www.ihe.net/User_Handbooks/

2919 **Appendix A.1. NIST Publications**

- 2920 • *NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk*
2921 *Management, Version 1.0*. Gaithersburg, MD, 2022. DOI: [10.6028/NIST.CSWP.10](https://doi.org/10.6028/NIST.CSWP.10).
2922 URL: <https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf>
- 2923 • Kevin Stine et al. *Volume I: guide for mapping types of information and information*
2924 *systems to security categories*. Gaithersburg, MD, 2008. DOI: [10.6028/NIST.SP.800-](https://doi.org/10.6028/NIST.SP.800-60v1r1)
2925 [60v1r1](https://doi.org/10.6028/NIST.SP.800-60v1r1). URL: [https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-](https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-60v1r1.pdf)
2926 [60v1r1.pdf](https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-60v1r1.pdf)
- 2927 • Simson L. Garfinkel. *De-identification of personal information*. 2015. DOI: [10.6028/](https://doi.org/10.6028/NIST.IR.8053)
2928 [NIST.IR.8053](https://doi.org/10.6028/NIST.IR.8053). URL: <https://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf>

- 2929 • Simson L. Garfinkel. *Government Data De-Identification Stakeholder's Meeting*
2930 *June 29, 2016 Meeting Report*. 2016. DOI: [10.6028/NIST.IR.8150](https://doi.org/10.6028/NIST.IR.8150). URL: [https:](https://nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8150.pdf)
2931 [//nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8150.pdf](https://nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8150.pdf)

2932 **Appendix A.2. Other U.S. Government Publications**

- 2933 • *Census Confidentiality and Privacy: 1790-2002*, US Census Bureau, 2003. [https:](https://www.census.gov/prod/2003pubs/conmono2.pdf)
2934 [//www.census.gov/prod/2003pubs/conmono2.pdf](https://www.census.gov/prod/2003pubs/conmono2.pdf)
- 2935 • *Data De-identification: An Overview of Basic Terms*, Privacy Technical Assistance
2936 Center, US Department of Education. May 2013. [http://ptac.ed.gov/sites/default/](http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf)
2937 [files/data_deidentification_terms.pdf](http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf)
- 2938 • *Data Disclosure Decision*, Department of Education (ED) Disclosure Review Board
2939 (DRB), A Product of the Federal CIO Council Innovation Committee. Version 1.0,
2940 2015.
- 2941 • *Disclosure Avoidance Techniques at the US Census Bureau: Current Practices and*
2942 *Research*, Research Report Series (Disclosure Avoidance #2014-02), Amy Lauger,
2943 Billy Wisniewski, and Laura McKenna, Center for Disclosure Avoidance Research,
2944 US Census. Bureau, September 26, 2014. [https://www.census.gov/srd/CDAR/cdar2014-02_](https://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf)
2945 [Discl_Avoid_Techniques.pdf](https://www.census.gov/srd/CDAR/cdar2014-02_Discl_Avoid_Techniques.pdf)
- 2946 • *Frequently Asked Questions – Disclosure Avoidance, Privacy Technical Assistance*
2947 *Center*, U.S. Department of Education. October 2012 (revised July 2015). [http:](http://ptac.ed.gov/sites/default/files/FAQ_Disclosure_Avoidance.pdf)
2948 [//ptac.ed.gov/sites/default/files/FAQ_Disclosure_Avoidance.pdf](http://ptac.ed.gov/sites/default/files/FAQ_Disclosure_Avoidance.pdf)
- 2949 • *Guidance Regarding Methods for De-identification of Protected Health Information*
2950 *in Accordance with the Health Insurance Portability and Accountability Act (HIPAA)*
2951 *Privacy Rule*, U.S. Department of Health & Human Services, Office for Civil Rights,
2952 November 26, 2012. [http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/](http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/De-identification/hhs_deid_guidance.pdf)
2953 [De-identification/hhs_deid_guidance.pdf](http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveridentities/De-identification/hhs_deid_guidance.pdf)
- 2954 • <http://www.hhs.gov/ohrp/policy/cdebiol.html>
- 2955 • http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf
- 2956 • The Data Disclosure Decision, Department of Education (ED) Disclosure Review
2957 Board (DRB), A Product of the Federal CIO Council Innovation Committee.
- 2958 • https://www.cdc.gov/nchs/data/nchs_microdata_release_policy_4-02a.pdf
- 2959 • http://www.cdc.gov/nchs/nvss/dvs_data_release.htm
- 2960 • *Linking Data for Health Services Research: A Framework and Instructional Guide.*,
2961 *Dusetzina SB, Tyree S, Meyer AM, Meyer A, Green L, Carpenter WR. (Prepared*
2962 *by the University of North Carolina at Chapel Hill under Contract No. 290-2010-*

- 2963 000141.) *AHRQ Publication No. 14-EHC033-EF. Rockville, MD: Agency for Health-*
2964 *care Research and Quality; September 2014.*
- 2965 • *National Center for Health Statistics Data Release and Access Policy for Micro-data*
2966 *and Compressed Vital Statistics File, Centers for Disease Control, April 26, 2011.*
2967 http://www.cdc.gov/nchs/nvss/dvs_data_release.htm
 - 2968 • *National Center for Health Statistics Policy on Micro-Data Dissemination, Centers*
2969 *for Disease Control, July 2002.* [https://www.cdc.gov/nchs/data/nchs_microdata_release_policy_4-](https://www.cdc.gov/nchs/data/nchs_microdata_release_policy_4-02a.pdf)
2970 [02a.pdf](https://www.cdc.gov/nchs/data/nchs_microdata_release_policy_4-02a.pdf)
 - 2971 • *OHRP-Guidance on Research Involving Private Information or Biological Specimens*
2972 *(2008), Department of Health & Human Services, Office of Human Research*
2973 *Protections (OHRP), August 16, 2008.* <http://www.hhs.gov/ohrp/policy/cdebiol.html>
 - 2974 • *OMB Circular A-130, Managing Information as a Strategic Resource, July 2016.*
 - 2975 • *Privacy and Confidentiality Research and the U.S. Census Bureau, Recommendations*
2976 *Based on a Review of the Literature, Thomas S. Mayer, Statistical Research Di-*
2977 *vision, US Bureau of the Census. February 7, 2002.* [https://www.census.gov/srd/papers/pdf/rsm2002-](https://www.census.gov/srd/papers/pdf/rsm2002-01.pdf)
2978 [01.pdf](https://www.census.gov/srd/papers/pdf/rsm2002-01.pdf)
 - 2979 • *Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical*
2980 *Disclosure Limitation Methodology, Federal Committee on Statistical Methodology,*
2981 *December 2005.*

2982 **Selected Publications by Other Governments**

- 2983 • *Privacy business resource 4: De-identification of data and information, Office of the*
2984 *Australian Information Commissioner, Australian Government, April 2014.* [http://www.oaic.gov.au/ima-](http://www.oaic.gov.au/ima-resources/privacy-business-resources/privacy_business_resource_4.pdf)
2985 [resources/privacy-business-resources/privacy_business_resource_4.pdf](http://www.oaic.gov.au/ima-resources/privacy-business-resources/privacy_business_resource_4.pdf)
- 2986 • *Opinion 05/2014 on Anonymisation Techniques, Article 29 Data Protection Working*
2987 *Party, 0829/14/EN WP216, Adopted on 10 April 2014.*
- 2988 • *Anonymisation: Managing data protection risk, Code of Practice 2012, Information*
2989 *Commissioner's Office.* [https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-](https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf)
2990 [code.pdf](https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf). 108 pages.
- 2991 • *The Anonymisation Decision-Making Framework, Mark Elliot, Elaine Mackey, Kieron*
2992 *O'Hara and Caroline Tudor, UKAN, University of Manchester, July 2016.* [http:](http://ukanon.net/ukan-resources/ukan-decision-making-framework/)
2993 [//ukanon.net/ukan-resources/ukan-decision-making-framework/](http://ukanon.net/ukan-resources/ukan-decision-making-framework/)

2994 **Reports and Books**

- 2995 • *Private Lives and Public Policies: Confidentiality and Accessibility of Government*
2996 *Statistics (1993), George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf,*

- 2997 Editors; Panel on Confidentiality and Data Access; *Commission on Behavioral and*
2998 *Social Sciences and Education; Division of Behavioral and Social Sciences and Ed-*
2999 *ucation*; National Research Council, 1993. <http://dx.doi.org/10.17226/2122>
- 3000 • *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, Committee on
3001 Strategies for Responsible Sharing of Clinical Trial Data, Board on Health Sciences
3002 Policy, Institute of Medicine of the National Academies, The National Academies
3003 Press, Washington, DC. 2015.
 - 3004 • P. Doyle and J. Lane, *Confidentiality, Disclosure and Data Access: Theory and Prac-*
3005 *tical Applications for Statistical Agencies*, North-Holland Publishing, Dec 31, 2001.
 - 3006 • George T. Duncan, Mark Elliot, Juan-José Salazar-Gonzalez, *Statistical Confiden-*
3007 *tiality: Principles and Practice*, Springer, 2011.
 - 3008 • Cynthia Dwork and Aaron Roth, *The Algorithmic Foundations of Differential Pri-*
3009 *vac* (Foundations and Trends in Theoretical Computer Science). Now Publishers,
3010 August 11, 2014. <http://www.cis.upenn.edu/~aaroht/privacybook.html>
 - 3011 • Khaled El Emam, *Guide to the De-Identification of Personal Health Information*,
3012 CRC Press, 2013.
 - 3013 • Khaled El Emam and Luk Arbuckle, *Anonymizing Health Data*, O'Reilly, Cam-
3014 bridge, MA. 2013.
 - 3015 • K El Emam and B Malin, "Appendix B: Concepts and Methods for De-Identifying
3016 Clinical Trial Data," in *Sharing Clinical Trial Data: Maximizing Benefits, Minimiz-*
3017 *ing Risk*, Institute of Medicine of the National Academies, The National Academies
3018 Press, Washington, DC. 2015.
 - 3019 • Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte
3020 Nordholt, Keith Spicer, Peter-Paul de Wolf, *Statistical Disclosure Control*, Wiley,
3021 September 2012.

3022 **How-To Articles**

- 3023 • Leah Krehling, De-Identification Guideline, WHISPERLAB, Technical Report WL-
3024 2020-01, Department of Electrical and Computer Engineering, Western University,
3025 2020.
- 3026 • Olivia Angiuli, Joe Blitstein, and Jim Waldo, How to De-Identify Your Data, Com-
3027 munications of the ACM, December 2015.
- 3028 • Jörg Drechsler, Stefan Bender, Susanne Rässler, Comparing fully and partially syn-
3029 thetic datasets for statistical disclosure control in the German IAB Establishment
3030 Panel. 2007, United Nations, Economic Commission for Europe. Working paper,
3031 11, New York, 8 p. <http://fdz.iab.de/342/section.aspx/Publikation/k080530j05>

- 3032 • Ebaa Fayyumi and B. John Oommen, A survey on statistical disclosure control and
3033 micro-aggregation techniques for secure statistical databases. 2010, *Software Prac-*
3034 *tice and Experience*. 40, 12 (November 2010), 1161-1188. DOI=10.1002/spe.v40:12
3035 <http://dx.doi.org/10.1002/spe.v40:12><http://dx.doi.org/10.1002/spe.v40:12>
- 3036 • Jingchen Hu, Jerome P. Reiter, and Quanli Wang, Disclosure Risk Evaluation for
3037 Fully Synthetic Categorical Data, *Privacy in Statistical Databases*, pp. 185-199,
3038 2014. https://link.springer.com/chapter/10.1007/978-3-319-11257-2_15
- 3039 • Matthias Templ, Bernhard Meindl, Alexander Kowarik and Shuang Chen, Introduc-
3040 tion to Statistical Disclosure Control (SDC), IHSN Working Paper No. 007, Inter-
3041 national Household Survey Network, August 2014. [http://www.ihsn.org/home/sites/](http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf)
3042 [default/files/resources/ihsn-working-paper-007-Oct27.pdf](http://www.ihsn.org/home/sites/default/files/resources/ihsn-working-paper-007-Oct27.pdf)
- 3043 • Natalie Shlomo, Statistical Disclosure Control Methods for Census Frequency Ta-
3044 bles, *International Statistical Review* (2007), 75, 2, 199-217. [https://www.jstor.org/](https://www.jstor.org/stable/41508461)
3045 [stable/41508461](https://www.jstor.org/stable/41508461)

3046 **Appendix B. List of Symbols, Abbreviations, and Acronyms**

3047 Selected acronyms and abbreviations used in this paper are defined below.

3048 **ACM** Association for Computing Machinery

3049 **AHRQ** Agency for Healthcare Research and Quality

3050 **AMD** Advanced Micro Devices

3051 **ARM** Advanced RISC Machines (formerly Acron RISC Machine)

3052 **ARMP** average record matching probability

3053 **ASTM** ASTM (formerly the American Society for Testing and Materials)

3054 **CED-DA** Center for Enterprise Dissemination-Disclosure Avoidance

3055 **CFR** Code of Federal Regulations

3056 **CIO** chief information officer

3057 **CIPSEA** The Confidential Information Protection and Statistical Efficiency Act of 2002

3058 **CNSS** Committee on National Security Systems

3059 **CNSSI** Committee on National Security Systems instruction

3060 **CPU** central processing unit

3061 **CRC** (formerly the Chemical Rubber Company)

3062 **DC** District of Columbia

3063 **DCMA** Defense Contract Management Agency

3064 **DICOM** Digital Imaging and Communications in Medicine

3065 **DNA** deoxyribonucleic acid

3066 **DOI** digital object identifier

3067 **DRB** disclosure review board

3068 **DUA** data use agreement

3069 **EDDRB** Department of Education disclosure review board

3070 **FCSM** Federal Committee on Statistical Methodology

3071 **FHE** Fully-homomorphic encryption

3072 **FISMA** Federal Information Security Modernization Act

3073 **FOIA** Freedom of Information Act

3074 **HHS** Health and Human Services
3075 **HIPAA** Health Insurance Portability and Accountability Act
3076 **HITRUST** (formerly the Health Industry Trust Alliance)
3077 **IAB** Institut für Arbeitsmarkt-und Berufsforschung (Germany's Institute for Employment
3078 and Research)
3079 **ICSP** Interagency Council on Statistical Policy
3080 **ID** Identification number
3081 **IEC** International Electrotechnical Commission
3082 **IHE** Integrating the Healthcare Enterprise
3083 **IHSN** International Household Survey Network
3084 **IP** internet protocol
3085 **IR** inter-agency report
3086 **IRB** institutional review board
3087 **IRS** Internal Revenue Service
3088 **ISO** (formerly International Organization for Standardization)
3089 **ISO/TS** ISO Technical Standard
3090 **IT** information technology
3091 **ITL** Information Technology Laboratory
3092 **KIRP** Known inclusion re-identification probability
3093 **MA** Massachusetts
3094 **MCC** Millennium Challenge Corporation
3095 **MD** Maryland
3096 **MIT** Massachusetts Institute of Technology
3097 **MPC** multi-party computation
3098 **NEMA** National Electrical Manufacturers Association
3099 **NIST** National Institute of Standards and Technology
3100 **NISTIR** National Institute of Standards and Technology interagency report
3101 **OECD** Organisation for Economic Co-operation and Development
3102 **OHRP** Office for Human Research Protections

- 3103 **OMB** Office of Management and Budget
- 3104 **OPRE** Office of Planning, Research and Evaluation
- 3105 **PDF** portable document file
- 3106 **PEC** privacy enhancing cryptography
- 3107 **PHI** protected health information
- 3108 **PII** personally identifiable information
- 3109 **PL** public law
- 3110 **PUF** public use file
- 3111 **RMP** record matching probability
- 3112 **SDC** statistical disclosure control
- 3113 **SDL** statistical disclosure limitation
- 3114 **SHA** secure hash algorithm
- 3115 **SLA** service-level agreement
- 3116 **SP** special publication
- 3117 **TEE** trusted execution environments
- 3118 **TX** Texas
- 3119 **UIRP** Unknown inclusion re-identification probability
- 3120 **UK** United kingdom
- 3121 **UKAN** United Kingdom Advocacy Network
- 3122 **US** United States
- 3123 **USC** United States Code
- 3124 **WHISPERLAB** Western Information Security and Privacy Research Laboratory
- 3125 **WP** working paper

3126 **Appendix C. Glossary**

3127 Selected terms used in the publication are defined below. Where noted, the definition is
3128 sourced from another publication.

3129 **anonymization** A process that removes the association between the identifying dataset
3130 and the data subject. (ISO 25237-2008)

3131 **attribute** An inherent characteristic. (ISO 9241-302:2008)

3132 **attribute disclosure** Re-identification event in which an entity learns confidential infor-
3133 mation about a data principal, without necessarily identifying the data principal.
3134 (ISO/IEC 20889 WORKING DRAFT 2 2016-05-27)

3135 **anonymity** Condition in identification whereby an entity can be recognized as distinct,
3136 without sufficient identity information to establish a link to a known identity. (ISO/IEC
3137 24760-1:2011)

3138 **anticipated re-identification rate** When an organization contemplates performing re-identification,
3139 the re-identification rate that the resulting de-identified data are likely to have.

3140 **attacker** A person who seeks to exploit potential vulnerabilities of a system.

3141 **attribute** Characteristic or property of an entity that can be used to describe its state, ap-
3142 pearance, or other aspect. (ISO/IEC 24760-1:2011)[65]

3143 **brute force attack** In cryptography, an attack that involves trying all possible combina-
3144 tions to find a match.

3145 **characteristic** Distinguishing feature. (ISO 8000-2:2012(E))

3146 **coded** 1. Identifying information (such as name or social security number) that would
3147 enable the investigator to readily ascertain the identity of the individual to whom
3148 the private information or specimens pertain has been replaced with a number, let-
3149 ter, symbol, or combination thereof (i.e., the code); 2. A key to decipher the code
3150 exists, enabling linkage of the identifying information to the private information or
3151 specimens. [95]

3152 **control** Measure that is modifying risk. Note: controls include any process, policy, device,
3153 practice, or other actions which modify risk. (ISO/IEC 27000:2014)

3154 **covered entity** Under HIPAA, a health plan, a health care clearinghouse, or a health care
3155 provider that conducts certain health care transactions electronically (e.g., billing).
3156 (HIPAA Privacy Rule)

3157 **data** Re-interpretable representation of information in a formalized manner suitable for
3158 communication, interpretation, or processing. (ISO 8000-2:2012(E))

- 3159 **data accuracy** Closeness of agreement between a property value and the true value. (ISO
3160 8000-2:2012(E))
- 3161 **data dictionary** collection of data dictionary entries that allows lookup by entity identifier.
3162 (ISO 8000-2:2012(E))
- 3163 **data dictionary entry** Description of an entity type containing, at a minimum, an unam-
3164 biguous identifier, a term, and a definition. (ISO 8000-2:2012(E))
- 3165 **data intruder** A data user who attempts to disclose information about a population through
3166 identification or attribution. (OECD Glossary of Statistical Terms)
- 3167 **data life cycle** The set of processes in an application that transform raw data into action-
3168 able knowledge. (NIST SP 1500-1)
- 3169 **data subjects** Persons to whom data refer. (ISO/TS 25237:2008)
- 3170 **data use agreement** Executed agreement between a data provider and a data recipient that
3171 specifies the terms under which the data can be used.
- 3172 **data universe** All possible data within a specified domain.
- 3173 **dataset** A collection of data.
- 3174 **dataset with identifiers** A dataset that contains information that directly identifies indi-
3175 viduals.
- 3176 **dataset without identifiers** A dataset that does not contain direct identifiers.
- 3177 **de-identification** A process that is applied to a dataset with the goal of preventing or lim-
3178 iting informational risks to individuals, protected groups, and establishments, while
3179 still allowing for the production of aggregate statistics.²⁸
- 3180 **de-identification model** An approach to the application of data de-identification tech-
3181 niques that enables the calculation of re-identification risk. (ISO/IEC 20889 WORK-
3182 ING DRAFT 2 2016-05-27)
- 3183 **de-identification process** A general term for any process of removing the association be-
3184 tween a set of identifying data and the data principal. (ISO/TS 25237:2008)
- 3185 **de-identified information** Records that have had enough PII removed or obscured such
3186 that the remaining information does not identify an individual, and there is no rea-
3187 sonable basis to believe that the information can be used to identify an individual.
3188 (SP800-122)
- 3189 **direct identifying data** Data that directly identify a single individual. (ISO/TS 25237:2008)

²⁸ISO/TS 25237:2008 defines de-identification as the “general term for any process of removing the association between a set of identifying data and the data subject” [66, p.3]. This document intentionally adopts a broader definition for de-identification that allows for noise-introducing techniques, such as differential privacy and the creation of synthetic datasets that are based on privacy-preserving models.

- 3190 **disclosure** Divulging of, or provision of access to, data. (ISO/TS 25237:2008)
- 3191 **disclosure limitation** Statistical methods used to hinder anyone from identifying an indi-
3192 vidual respondent or establishment by analyzing published data, especially by ma-
3193 nipulating mathematical and arithmetical relationships among the data. [p.21][130]
- 3194 **effectiveness** The extent to which planned activities are realized and planned results achieved.
3195 (ISO/IEC 27000:2014)
- 3196 **entity** An item inside or outside an information and communication technology system,
3197 such as a person, an organization, a device, a subsystem, or a group of such items
3198 that has recognizably distinct existence. (ISO/IEC 24760-1:2011)
- 3199 **expert determination** Within the context of de-identification, refers to the Expert Deter-
3200 mination method for de-identifying protected health information in accordance with
3201 the HIPAA Privacy Rule de-identification standard.
- 3202 **Federal Committee on Statistical Methodology (FCSM)** An interagency committee ded-
3203 icated to improving the quality of Federal statistics. The FCSM was created by the
3204 Office of Management and Budget (OMB) to inform and advise OMB and the Inter-
3205 agency Council on Statistical Policy (ICSP) on methodological and statistical issues
3206 that affect the quality of Federal data. (fscm.sites.usa.gov)
- 3207 **genomic information** Information based on an individual's genome, such as a sequence
3208 of DNA or the results of genetic testing.
- 3209 **harm** Any adverse effects that would be experienced by an individual (i.e., that may be
3210 socially, physically, or financially damaging) or an organization if the confidentiality
3211 of PII were breached. (SP 800-122)
- 3212 **Health Insurance Portability and Accountability Act of 1996 (HIPAA)** A federal statute
3213 that called on the federal Department of Health and Human Services to establish reg-
3214 ulatory standards to protect the privacy and security of individually identifiable health
3215 information. See <https://www.hhs.gov/hipaa/for-professionals/index.html>.
- 3216 **HIPAA** See *Health Insurance Portability and Accountability Act of 1996*.
- 3217 **HIPAA Privacy Rule** Establishes national standards to protect individuals' medical records
3218 and other personal health information and applies to health plans, health care clear-
3219 inghouses, and those health care providers that conduct certain health care transac-
3220 tions electronically. (HIPAA Privacy Rule, 45 CFR 160, 162, 164). See [https://www.hhs.gov/hipaa/for-](https://www.hhs.gov/hipaa/for-professionals/privacy/index.html)
3221 [professionals/privacy/index.html](https://www.hhs.gov/hipaa/for-professionals/privacy/index.html).
- 3222 **identification** The process of using claimed or observed attributes of an entity to single
3223 out the entity among other entities in a set of identities. (ISO/TS 25237:2008)
- 3224 **identifying information** Information that can be used to distinguish or trace an individ-
3225 ual's identity (e.g., their name, social security number, biometric records, etc.) alone

3226 or when combined with other personal or identifying information that is linked or
3227 linkable to a specific individual (e.g., date and place of birth, mother’s maiden name,
3228 etc.). (OMB M-07-16)

3229 **identifier** Information used to claim an identity, before a potential corroboration by a cor-
3230 responding authenticator. (ISO/TS 25237:2008)

3231 **imputation** A procedure for entering a value for a specific data item where the response is
3232 missing or unusable. (OECD Glossary of Statistical Terms)

3233 **inference** Refers to the ability to deduce the identity of a person associated with a set of
3234 data through “clues” contained in that information. This analysis permits determi-
3235 nation of the individual’s identity based on a combination of facts associated with
3236 that person even though specific identifiers have been removed, like name and social
3237 security number. (ASTM E1869-04)[15]

3238 **information** Knowledge concerning objects, such as facts, events, things, processes, or
3239 ideas, including concepts, that within a certain context has a particular meaning.
3240 (ISO 8000-2:2012(E))

3241 **k-anonymity** A technique “to release person-specific data such that the ability to link to
3242 other information using the quasi-identifier is limited” [122]. *k*-anonymity achieves
3243 this through suppression of identifiers and output perturbation.

3244 **l-diversity** A refinement to the *k*-anonymity approach that assures that groups of records
3245 specified by the same identifiers have sufficient diversity to prevent inferential dis-
3246 closure. [76]

3247 **masking** The process of systematically removing a field or replacing it with a value in a
3248 way that does not preserve the analytic utility of the value, such as replacing a phone
3249 number with asterisks or a randomly generated pseudonym. [45]

3250 **motivated intruder test** The ‘motivated intruder’ is taken to be a person who starts with-
3251 out any prior knowledge but who wishes to identify the individual from whose per-
3252 sonal data the anonymised data has been derived. This test is meant to assess whether
3253 the motivated intruder would be successful. [63]

3254 **noise** A convenient term for a series of random disturbances borrowed through communi-
3255 cation engineering, from the theory of sound. In communication theory, noise results
3256 in the possibility of a signal sent, *x*, being different from the signal received, *y*, and
3257 the latter has a probability distribution conditional upon *x*. If the disturbances con-
3258 sist of impulses at random intervals, it is sometimes known as “shot noise.” (OECD
3259 Glossary of Statistical Terms)

3260 **non-deterministic noise** A random value that cannot be predicted.

3261 **non-ignorable bias** A bias introduced into data or an analytics procedure that results in a
3262 change that cannot be ignored.

3263 **non-public personal information** Information about a person that is not publicly known;
3264 called “private information” in some other publications.

3265 **personal identifier** Information with the purpose of uniquely identifying a person within
3266 a given context. (ISO/TS 25237:2008)

3267 **personal data** Any information relating to an identified or identifiable natural person (*data*
3268 *subject*). (ISO/TS 25237:2008)

3269 **personal information** See *personal data*.

3270 **personally identifiable information (PII)** Any information about an individual maintained
3271 by an agency, including (1) any information that can be used to distinguish or trace
3272 an individual’s identity, such as name, social security number, date and place of birth,
3273 mother’s maiden name, or biometric records; and (2) any other information that is
3274 linked or linkable to an individual, such as medical, educational, financial, and em-
3275 ployment information. [106](SP 800-122)

3276 **perturbation-based methods** Perturbation-based methods falsify the data before publica-
3277 tion by introducing an element of error purposely for confidentiality reasons. This
3278 error can be inserted in the cell values after the table is created, which means the
3279 error is introduced to the output of the data and will therefore be referred to as output
3280 perturbation, or the error can be inserted in the original data on the microdata level,
3281 which is the input of the tables one wants to create; the method with then be referred
3282 to as data perturbation—input perturbation being the better but uncommonly used
3283 expression. Possible methods are: rounding; random perturbation; [and] disclosure
3284 control methods for microstatistics applied to macrostatistics. (OECD Glossary of
3285 Statistical Terms)

3286 **privacy** Freedom from intrusion into the private life or affairs of an individual when that
3287 intrusion results from undue or illegal gathering and use of data about that individual.
3288 (ISO/IEC 2382-8:1998, definition 08-01-23)

3289 **privacy risk**

3290 **privacy loss** A measure of the extent to which a data release may reveal information that
3291 is specific to an individual.

3292 **privacy loss budget** An upper bound on the cumulative total privacy loss for individuals.

3293 **property value** Instance of a specific value together with an identifier for a data dictionary
3294 entry that defines a property. (ISO 8000-2:2012(E))

3295 **protected health information (PHI)** Individually identifiable health information: (1) Ex-
3296 cept as provided in paragraph (2) of this definition, that is: (i) Transmitted by elec-
3297 tronic media; (ii) Maintained in electronic media; or (iii) Transmitted or maintained
3298 in any other form or medium. (2) *Protected health information* excludes individu-
3299 ally identifiable health information in: (i) Education records covered by the Fam-

3300 ily Educational Rights and Privacy Act, as amended, [20 USC. 1232g](#); (ii) Records
3301 described at [20 USC. 1232g\(a\)\(4\)\(B\)\(iv\)](#); and (iii) Employment records held by a
3302 covered entity in its role as employer. (HIPAA Privacy Rule, 45 CFR 160.103). See
3303 <https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html>.

3304 **pseudonymization** A particular type of de-identification that both removes the association
3305 with a data subject and adds an association between a particular set of characteristics
3306 related to the data subject and one or more pseudonyms.²⁹ Typically, pseudonymiza-
3307 tion is implemented by replacing direct identifiers with a pseudonym, such as a ran-
3308 domly generated value.

3309 **pseudonym** Personal identifier that is different from the normally used personal identifier.
3310 (ISO/TS 25237:2008)

3311 **quality** Degree to which a set of inherent characteristics fulfils requirements. (ISO 8000-
3312 2:2012(E))

3313 **quasi-identifier** A variable that can be used to identify an individual through association
3314 with another variable.

3315 **recipient** Natural or legal person, public authority, agency, or any other body to whom
3316 data are disclosed. (ISO/TS25237:2008)

3317 **redaction** The removal of information from a document or dataset for legal or security
3318 purposes.

3319 **re-identification** A general term for any process that restores the association between a
3320 set of de-identified data and a data subject.

3321 **re-identification risk** The likelihood that a third party can re-identify data subjects in a
3322 de-identified dataset.

3323 **re-identification rate** The percentage of records in a dataset that can be re-identified.

3324 **re-identificaion probability** TBD

3325 **requirement** A need or expectation that is stated, generally implied or obligatory. (ISO
3326 8000-2:2012(E))

3327 **risk** A measure of the extent to which an entity is threatened by a potential circumstance
3328 or event, and typically a function of: (i) the adverse impacts that would arise if the
3329 circumstance or event occurs; and (ii) the likelihood of occurrence. (CNSSI No.
3330 4009)

3331 **risk assessment** The process of identifying, estimating, and prioritizing risks to organi-
3332 zational operations (including mission, functions, image, reputation), organizational

²⁹This definition is the same as the definition in ISO/TS 25237:2008, except that the word “anonymization” is replaced with the word “de-identification.”

3333 assets, individuals, other organizations, and the Nation, resulting from the operation
3334 of an information system. Part of risk management, incorporates threat and vulner-
3335 ability analyses, and considers mitigations provided by security controls planned or
3336 in place. Synonymous with risk analysis. (NIST SP 800-39)

3337 **safe harbor** Within the context of de-identification, refers to the Safe Harbor method for
3338 de-identifying protected health information in accordance with the Health Insurance
3339 Portability and Accountability Act (HIPAA) Privacy Rule. See [https://www.hhs.gov/hipaa/for-](https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html)
3340 [professionals/privacy/special-topics/de-identification/index.html](https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html).

3341 **statistical disclosure control** The set of methods to reduce the risk of disclosing informa-
3342 tion on individuals, businesses or other organizations. Such methods are only related
3343 to the dissemination step and are usually based on restricting the amount of or modi-
3344 fying the data released. (OECD Glossary of Statistical Terms)

3345 **suppression** One of the most commonly used ways of protecting sensitive cells in a table is
3346 via suppression. It is obvious that in a row or column with a suppressed sensitive cell,
3347 at least one additional cell must be suppressed, or the value in the sensitive cell could
3348 be calculated exactly by subtraction from the marginal total. For this reason, certain
3349 other cells must also be suppressed. These are referred to as secondary suppressions.
3350 (OECD Glossary of Statistical Terms)

3351 **synthetic data generation** A process in which seed data are used to create artificial data
3352 that have some of the statistical characteristics as the seed data.