NISTIR 8053

De-Identification of Personal Information

Simson L. Garfinkel

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8053

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8053

De-Identification of Personal Information

Simson L. Garfinkel
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8053

October 2015

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Under Secretary of Commerce for Standards and Technology and Director

National Institute of Standards and Technology Internal Report 8053
vi + 46 pages (October 2015)

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8053

National Institute of Standards and Technology
Attn: Computer Security Division, Information Technology Laboratory
100 Bureau Drive (Mail Stop 8930) Gaithersburg, MD 20899-8930

Reports on Computer Systems Technology

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

Abstract

De-identification removes identifying information from a dataset so that individual data cannot be linked with specific individuals. De-identification can reduce the privacy risk associated with collecting, processing, archiving, distributing or publishing information. De-identification thus attempts to balance the contradictory goals of using and sharing personal information while protecting privacy. Several U.S laws, regulations and policies specify that data should be de-identified prior to sharing. In recent years researchers have shown that some de-identified data can sometimes be re-identified. Many different kinds of information can be de-identified, including structured information, free format text, multimedia, and medical imagery. This document summarizes roughly two decades of de-identification research, discusses current practices, and presents opportunities for future research.

Keywords

De-identification; HIPAA Privacy Rule; k-anonymity; differential privacy; re-identification; privacy

Acknowledgements

John Garofolo and Barbara Guttman provided significant guidance and support in completing this project. The author would also like to thank Daniel Barth-Jones, David Clunie, Pam Dixon, Khaled El Emam, Orit Levin, Bradley Malin, Latanya Sweeney, and Christine M. Task for their assistance in answering questions and reviewing earlier versions of this document. More than 30 sets of written comments were received on Draft 1 of this document from many organizations including Anonos, Center for Democracy and Technology, Future of Privacy Forum, GlaxoSmithKline, IMS Health, Microsoft, Optum Privacy, Patient Privacy Rights, Privacy Analytics, and World Privacy Forum. We also received comments from the members of COST Action IC1206 "De-identification for Privacy Protection in Multimedia Content," Dr. Marija Krlic, Prof. Samir Omarovic, Prof. Slobodan Ribarić and Prof. Alexin Zoltan.

Audience

This document is intended for use by officials, advocacy groups, researchers and other members of communities that are concerned with policy issues involving the creation, use and sharing of data containing personal information. It is also designed to provide technologists and researchers with an overview of the technical issues in the de-identification of data. Data protection officers

in government, industry and academia will also benefit from the assemblage of information in this document.

While this document assumes a high-level understanding of information system security technologies, it is intended to be accessible to a wide audience. For this reason, this document minimizes the use of mathematical notation.

Table of Contents

1 Introduction 1
1.1 Document Purpose and Scope 1
1.2 Intended Audience 1
1.3 Organization 1
1.4 Notes on Terminology 2
1.4.1 “de-identification,” “redaction,” “pseudonymization,” and “anonymization” 2
1.4.2 “Personally Identifiable Information (PII)” and “Personal Information” 3
2 De-identification, Re-Identification, and Data Sharing Models 3
2.1 Motivation 3
2.2 Models for Privacy-Preserving use of Private Information 6
2.2.1 Privacy Preserving Data Mining (PPDM) 7
2.2.2 Privacy Preserving Data Publishing (PPDP) 8
2.3 De-Identification Data Flow Model 9
2.4 Re-identification Attacks and Data Intruders 9
2.5 Release models and data controls 14
3 Approaches for De-Identifying and Re-Identifying Structured Data 15
3.1 Removal of Direct Identifiers 15
3.2 Pseudonymization 16
3.3 Re-identification through Linkage Attacks 17
3.4 De-identification of Quasi-Identifiers 19
3.5 De-identification of Protected Health Information (PHI) under HIPAA 22
3.5.1 The HIPAA Expert Determination Method 22
3.5.2 The HIPAA Safe Harbor Method 23
3.5.3 Evaluating the effectiveness of the HIPAA Safe Harbor Method 25
3.5.4 HIPAA Limited Datasets 26
3.6 Evaluation of Field-Based De-identification 26
3.7 Estimation of Re-Identification Risk 29
4 Challenges in De-Identifying Unstructured Data 30
4.1 De-identifying medical text 30
4.2 De-identifying Photographs and Video 32
4.3 De-Identifying Medical Imagery 35

4.4 De-identifying Genetic information and biological materials 36
4.5 De-identification of geographic and map data 37
5 Conclusion 38

List of Appendices

Appendix A Glossary 39
Appendix B Resources 44
B.1 Official publications 44
B.2 Law Review Articles and White Papers: 45
B.3 Reports and Books: 46
B.4 Survey Articles 46

1 Introduction

De-identification is a tool that organizations can use to remove personal information from data that they collect, use, archive, and share with other organizations.

De-identification is not a single technique, but a collection of approaches, algorithms, and tools that can be applied to different kinds of data with differing levels of effectiveness. In general, privacy protection improves as more aggressive de-identification techniques are employed, but less utility remains in the resulting dataset.

De-identification is especially important for government agencies, businesses, and other organizations that seek to make data available to outsiders. For example, significant medical research resulting in societal benefit is made possible by the sharing of de-identified patient information under the framework established by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, the primary US regulation providing for privacy of medical records.

As long as any utility remains in the data derived from personal information, there also exists the possibility, however remote, that some information might be linked back to the original individuals on whom the data are based. When de-identified data can be *re-identified* the privacy protection provided by de-identification is lost. The decision of how or if to de-identify data should thus be made in conjunction with decisions of how the de-identified data will be used, shared or released, since the risk of re-identification can be difficult to estimate.

Risks to individuals can remain in de-identified data. These risks include allowing inferences about individuals in the data without re-identification, and impacts on groups represented in the data.

1.1 Document Purpose and Scope

This document provides an overview of de-identification issues and terminology. It summarizes significant publications to date involving de-identification and re-identification. It does not make recommendations regarding the appropriateness of de-identification or specific de-identification algorithms.

1.2 Intended Audience

This document is intended for use by officials, advocacy groups, researchers, academics, and other members of communities that are concerned with the practical and policy issues involving the creation, use, archiving, and sharing of data containing personal information. It is also designed to provide researchers and academics with a limited overview of the technical issues involving the de-identification of data. Although this document presents research from around the world and is written for a world-wide audience, it is primarily concerned with policy and legal issues in the United States. While this document assumes a high-level understanding of information system security technologies, it is intended to be accessible to a wide audience. For this reason, this document minimizes the use of mathematical notation.

1.3 Organization

The remainder of this report is organized as follows: Section 2 introduces the concepts of de-

identification, re-identification and data sharing models. Section 3 discusses approaches for de-identifying structured data, typically by removing, masking or altering specific categories such as names and phone numbers. Section 4 discusses challenges of de-identification for non-tabular data, such as free-format text, images, and genomic information. Section 5 provides this report's conclusion that de-identification, while not perfect, is a significant technical control that may protect the privacy of data subjects. Appendix A is a glossary, and Appendix B provides a list of additional resources.

1.4 Notes on Terminology

1.4.1 "de-identification," "redaction," "pseudonymization," and "anonymization"

Some authors and publications use the terms "de-identification" and "anonymization" interchangeably. Others use "de-identification" to describe a process and "anonymization" to denote a specific kind of de-identification that cannot be reversed. In some healthcare contexts the terms "de-identification" and "pseudonymization" are treated equivalently, with the term "anonymization" being used to indicate that the mapping pseudonyms to subject identities has been erased. The term "redaction" is sometimes used in a government context to describe the straightforward removal of information that is identifying or otherwise sensitive.

This document bases its terminology choices on ISO/TS 25237:2008(E), "Health Informatics—Pseudonymization." Unfortunately, the standard does not provide unambiguous guidance. In particular, the standard contains the following definitions::

> *de-identification:* "general term for any process of removing the association between a set of identifying data and the data subject." [p. 3]
>
> *anonymization:* "process that removes the association between the identifying dataset and the data subject." [p. 2]
>
> *pseudonymization:* "particular type of anonymization that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms."[1] [p. 5]

Later, ISO/TS 25237:2008(E) provides explanatory text stating:

> "NOTE—Anonymization is another subcategory of de-identification. Unlike pseudonymization, it does not provide a means by which the information may be linked to the same person across multiple data records or information systems. Hence re-identification of anonymized data is not possible." [p. 6]

The problem with these definitions is that some anonymization attempts have resulted in data have been re-identified, implying that the date thought to be anonymized actually weren't.

[1] The term *coded* is sometimes used to describe private information or biological specimens where the identifiers have been replaced with pseudonyms. See *OHRP-Guidance on Research Involving Private Information or Biological Specimens*, Department of Health & Human Services, Office of Human Research Protections (OHRP), August 16, 2008. http://www.hhs.gov/ohrp/policy/cdebiol.html

In medical imaging, the term "de-identification" is used to denote "the process of removing real patient identifiers or the removal of all subject demographics from imaging data for anonymization," while the term "de-personalization" is taken to mean "the process of completely removing any subject-related information from an image, including clinical trial identifiers."[2]

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

1.4.2 "Personally Identifiable Information (PII)" and "Personal Information"

The phrase Personally Identifiable Information (PII) is typically used to indicate information that contains identifiers specific to individuals, although there are a variety of definitions for PII in various law, regulation, and agency guidance documents. Because of these multiple different definitions, it is possible to have information that singles out individuals but which does not meet a particular definition of PII. An added complication is that some documents use the phrase PII to denote any information that is attributable to individuals, or information that is uniquely attributable to a specific individual, while others use the term strictly for data that are in fact identifying

Because of these inconsistences, this document avoids the term "personally identifiable information." Instead, the phrase "personal information" is used to denote information from individuals, and "identifying information" is used to denote information that identifies individuals. Therefore, identifying information is personal information, but personal information is not necessarily identifying information.

2 De-identification, Re-Identification, and Data Sharing Models

This section explains the motivation for de-identification, discusses the use of re-identification attacks to gauge the effectiveness of de-identification, and describes models for sharing de-identified data. It also introduces the terminology used in this report.

2.1 Motivation

Increasingly, organizations that are collecting and maintaining data are being challenged to protect the data while using and sharing them as widely as possible. For government data, sharing can increase transparency, provide new resources to private industry, and lead to more efficient government as a whole. Private firms can realize benefits from dataset sharing in the form of increased publicity, civic engagement, and even increased revenue from the sale of data or analytic results.

When data contain identifying information such as names, email addresses, geolocation information, or photographs, there can be a conflict between the goals of data use and privacy protection. De-identification attempts to resolve this conflict, allowing for some privacy sensitive

[2] Colin Miller, Joe Krasnow, Lawrence H. Schwartz, *Medical Imaging in Clinical Trials,* Springer Science & Business Media, Jan 30, 2014.

data that identifies individuals to be removed, while allowing other useful information to remain.

De-identification is thus an important tool that organizations can use to minimize the privacy risk associated with creating, using, archiving, sharing and even publishing data containing personal information. De-identifying data at the time of collection or after minimal processing can reduce the costs associated with using and archiving data, by reducing the privacy risk associated with inadvertent release (i.e., a data breach). De-identifying data that are shared can reduce the need for technical and policy controls. De-identification can therefore allow organizations to make greater use of data than might otherwise be possible.

Several U.S. laws and regulations specifically recognize the importance and utility of data de-identification. Examples include, but are not limited to:

- The Department of Education has held that the restrictions imposed by the Family and Educational Records Privacy Act (FERPA) do not apply to de-identified student records. "Educational agencies and institutions are permitted to release, without consent, educational records, or information from educational records that have been de-identified through the removal of all personally identifiable information."[3]
- The requirements for the security and privacy of health information established by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule[4] and the Health Information Technology for Economic and Clinical Health Act (HITECH Act) [5] explicitly do not apply to the protected health information that has been de-identified, provided that there is "no reasonable basis to believe that the information can be used to identify an individual."[6]
- The Foodborne Illness Surveillance System operated by the Centers for Disease Control and Prevention[7] is required to allow "timely public access to aggregated, de-identified surveillance data."[8]
- Entities contracted by Health and Human Services to provide drug safety data must have the ability to provide that data in de-identified form.[9]
- Voluntary safety reports submitted to the Federal Aviation Submission are not protected from public disclosure if the data that they contain are de-identified.[10]

---

[3] Dear Colleague Letter about Family Educational Rights and Privacy Act (FERPA) Final Regulations, U.S. Department of Education, December 17, 2008. http://www2.ed.gov/policy/gen/guid/fpco/hottopics/ht12-17-08.html

[4] 45 CFR 160, 45 CFR 162, and 45 CFR 164. See also "Combined Regulation Text of All Rules," U.S. Department of Health and Human Services, Office for Civil Rights, Health Information Privacy. http://www.hhs.gov/ocr/privacy/hipaa/administrative/combined/index.html

[5] 42 USC 17935

[6] 42 CFR 164.514

[7] http://www.cdc.gov/foodsafety/fdoss/surveillance/index.html

[8] 21 USC 2224

[9] 21 USC 355

These laws and regulations treat de-identification as a technical control that can be applied to data, removing personal information and allowing the data that remains to be used in a way that will not disclose the identities of the data subjects.

However, de-identification approaches based on suppressing or generalizing specific fields in a database cannot provide absolute privacy guarantees, because there is always a chance that the remaining data can be re-identified using an auxiliary dataset. Section 3.6 discuss some of the well-publicized cases in which data that were thought to be properly de-identified were published and then later re-identified by researchers or journalists. Some of these re-identification demonstrations disclosed the identity of the data subjects. Additional privacy issues can result from the disclosure of specific attributes that the dataset linked to the identities. Nevertheless, a test of the HIPAA "Safe Harbor" method, one of two approaches that can be used to meet the HIPAA de-identification standard (which is based on field suppression) found that less than 1% of the de-identified records could be re-identified when the only identifying information that remained was year of birth, sex, and 3-digit ZIP code (see Section 3.5).

Because of the re-identification risk, some organizations sharing de-identified data may wish to execute a data use agreement (DUA) with downstream users. For example, a DUA could prohibit a recipient of de-identified data from attempting to re-identify the data subjects, from linking to external data, or from sharing the data without permission.[11]

As shown in Figure 1, all data exist on an identifiability spectrum. At one end (the left) are data that are not related to individuals (for example, historical weather records) and therefore pose no privacy risk. At the other end (the right) are data that are linked directly to specific individuals. Between these two endpoints are data that can be linked with effort, that can only be linked to groups of people, and that are based on individuals but cannot be linked back. In general, de-identification approaches are designed to push data to the left while retaining some desired utility, lowering the risk of distributing de-identified data to a broader population or the general public.

Some privacy advocates maintain that the Fair Information Practice Principles (FIPPs)[12] require data subjects be notified that their personal data will be shared, even if the data being shared will be de-identified, and even if there is no legal obligation to notify the subjects. However, current US policy and law gives organizations considerable lattitude in the uses that can be made of de-identified data. These policies were typically developed based on an attempt to balance the societal benefit resulting from the use of de-identified data with the perceived risks to subjects that might result from having the data re-identified. Because these risks may change as technology evolves, it is important to periodically review policies regarding the use of de-identified data.

---

[10] 49 USC 44735

[11] Under the HIPAA Privacy Rule, data use agreements are required when sharing *limited datasets*, which are de-identified to a lesser standard, and not necessarily when sharing datasets that have been de-identified according to the Privacy Rule (*e.g.*, under the Expert Determination method, the expert may require a DUA).

[12] National Strategy for Trusted Identities in Cyberspace, Appendix A—Fair Information Practice Principles. April 15, 2011. http://www.nist.gov/nstic/NSTIC-FIPPs.pdf

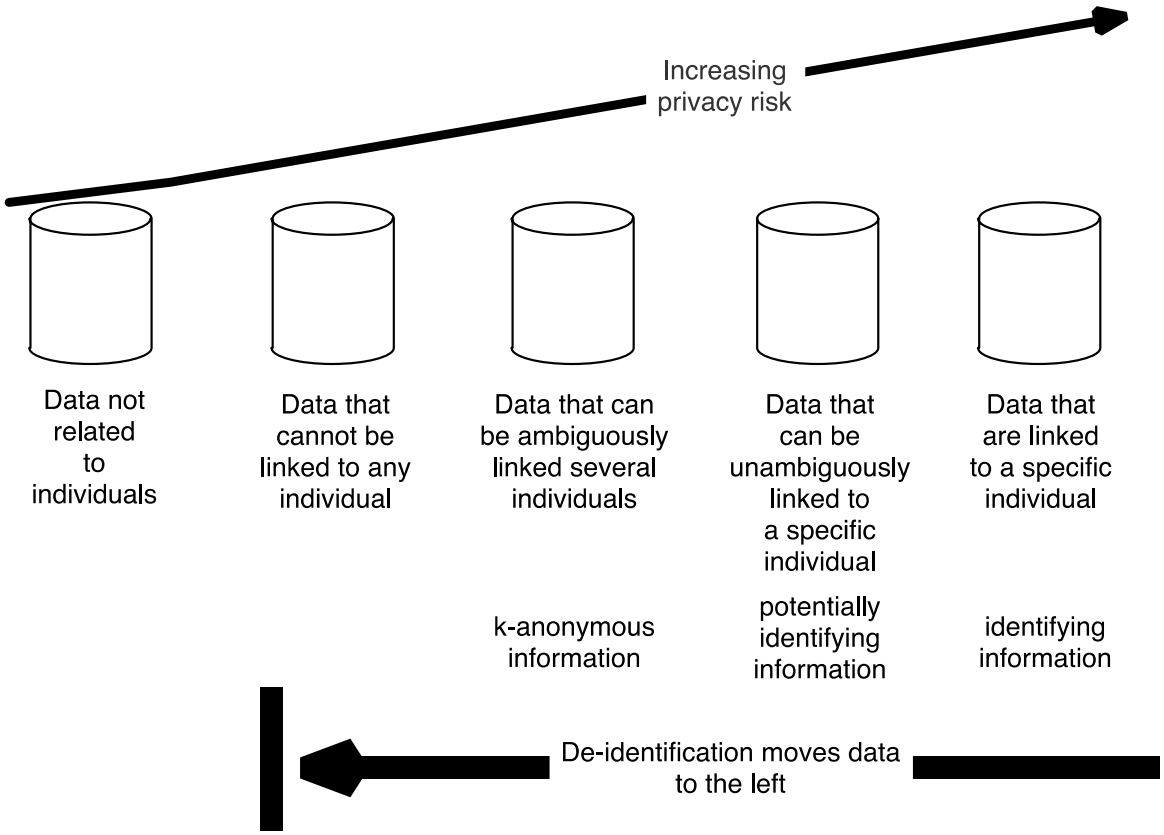

Figure 1: The Data Identifiability Spectrum.

2.2 Models for Privacy-Preserving use of Private Information

Academics have identified two distinct models for using personal information in a database while protecting the privacy of the data subjects:

- ***Privacy Preserving Data Mining (PPDM).*** In this model, data are not released, but are used instead for statistical processing or machine learning.[13] The results of the calculations may be released in the form of statistical tables based on summarization and aggregation, classifiers that implement machine learning algorithms, and other kinds of results.
- ***Privacy Preserving Data Publishing (PPDP).*** In this model, data are processed to produce a new, de-identified or synthetic data product that is distributed to users.

Both of these models are said to be "privacy preserving" in that they are intended to allow the

[13] *Machine Learning* is a class of computer algorithms and techniques that can allow computers to classify information and recognize patterns without being explicitly programmed to do so.

release of some information (*e.g.*, aggregated information, statistical results, classifiers, or synthetic data) without revealing information that can be attributed to a specific individual within the original dataset.

2.2.1 Privacy Preserving Data Mining (PPDM)

PPDM is a broad term for any use of sensitive information to publish public statistics. Statistical reports that summarize confidential survey data are an example of PPDM.

Statistical Disclosure Limitation *"is the discipline concerned with the modification of statistical data in order to prevent third parties working with these data to recognize individuals in the data."*[14] Techniques developed for disclosure limitation include generalization of reported information to broader categories, swapping data between similar entities, and the addition of noise in reports.[15]

Differential Privacy is a set of techniques based on a mathematical definition of identity disclosure and information leakage from operations on a dataset. Differential privacy prevents disclosure by adding non-deterministic noise (usually small random values) to the results of mathematical operations before the results are reported.[16] Differential privacy's mathematical definition holds that the result of an analysis of a dataset should be roughly the same before and after the addition or removal of a single data record (which is usually taken to be the data from a single individual). This works because the amount of noise added masks the contribution of any individual. The degree of sameness is defined by the parameter **ε** (epsilon). The smaller the parameter **ε**, the more noise is added, and the more difficult it is to distinguish the contribution of a single record. The result is increased privacy for all of the data subjects. In its most basic form, differential privacy applies only to online query systems, but differential privacy can also be used to produce machine-learning statistical classifiers and synthetic data.[17]

Differential privacy is an active research area, but to date it has only been applied to a few operational systems, including:

- The Census Bureau's "OnTheMap" website, which uses differential privacy to create reasonably accurate block-level synthetic census data.[18]

- Google's "Chrome" web browser, which uses randomized responses to collect aggregate statistics about the Windows process names running on the user's computer and the user's home page. Although the statistics are accurate in aggregate, the use of

---

[14] Leon Willenborg and Ton de De Wall, *Elements of Statistical Disclosure Control, 2001. Springer*

[15] Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, December 2005.

[16] Cynthia Dwork, Differential Privacy, in ICALP, Springer, 2006

[17] Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron, Zhiwei Steven Wu, Dual Query: Practical Private Query Release for High Dimensional Data, Proceedings of the 31st International Conference on Machine Learning, Beijing, China. 2014. JMLR: W&CP volume 32.

[18] Abowd *et al.*, "Formal Privacy Guarantees and Analytical Validity of OnTheMap Public-use Data," Joint NSF-Census-IRS Workshop on Synthetic Data and Confidentiality Protection, Suitland, MD, July 31, 2009.

> randomization makes it impossible to reliably determine a users' processes or home page.[19]

Differential privacy comes at the cost of decreased result accuracy. For example, a study conducted by Fredrikson *et al.* determined the impact of using differential privacy to create a statistical model for correlating genomic information and warfarin dosage based on clinical trial data.[20] The study found that the models constructed using differential privacy would result in worse clinical outcomes for a significant number of patients compared to those models created without differential privacy, although this finding was only tested in simulation and not in an actual clinical trial.

2.2.2 Privacy Preserving Data Publishing (PPDP)

PPDP allows for information based on private data to be published, allowing other researchers to perform novel analyses. The goal of PPDP is to provide data that have high utility without compromising the identity of the data subjects.[21]

De-identification is the "general term for any process of removing the association between a set of identifying data and the data subject."[22] De-identification is designed to protect individual identity, making it hard or impossible to learn if the data in a dataset is related to a specific individual, while preserving some of the dataset's utility for other purposes. De-identification is one of the primary tools for achieving PPDP.

Synthetic data generation uses some PPDM techniques to create a dataset that is similar to the original data, but where some or all of the resulting data elements are generated and do not map to actual individuals. As such, synthetic data generation can be seen as a fusion of PPDM and PPDP.

---

[19] Úlfar Erlingsson, Vasyl Pihur, Aleksandra Korolova, RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response, In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (CCS '14). ACM, November 3-7, 2014, Scottsdale, AZ. http://dl.acm.org/citation.cfm?doid=2660267.2660348

[20] Fredrikson *et al.*, Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing, 23rd Usenix Security Symposium, August 20-22, 2014, San Diego, CA.

[21] Benjamin C. M. Fung, Ke Wang, Rui Chen and Philip S. Yu, Privacy-Preserving Data Publishing: A Survey on Recent Developments, Computing Surveys, June 2010.

[22] ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. ISO, Geneva, Switzerland. 2008.

2.3 De-Identification Data Flow Model

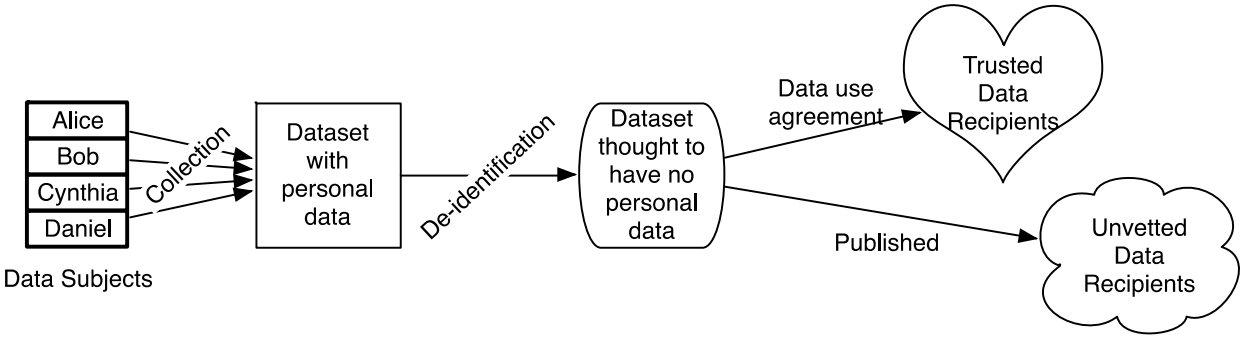

Figure 2: Data Collection, De-Identification and Use

Figure 2 provides an overview of the de-identification process. Data are collected from *Data Subjects***,** the "persons to whom data refer." (ISO/TS 25237-2008) These personal data are combined into a *dataset* containing *personal information*. De-identification creates a new dataset thought to have no identifying data. This dataset may be internally used by an organization instead of the original dataset to decrease privacy risk. The dataset may also be provided to trusted data recipients who are bound by additional administrative controls such as data use agreements.[23] Alternatively, the data might be made broadly available to a larger (potentially limitless) number of unknown and unvetted data recipients—for example, by publishing the de-identified data on the Internet.

As Figure 2 shows, de-identified data need not be publicly released. Furthermore, de-identification may be merely one of several controls applied to protect the identity of the data subjects.

De-identification can be performed *manually* by a human, by an *automated* process, or by a combination of the two. Once de-identified, data can be manually reviewed or otherwise audited to determine if any identifying information remains, or to see if it is possible to learn the identities that have been removed by correlating with other data sources.

2.4 Re-identification Attacks and Data Intruders

Re-identification is the process of attempting to discern the identities that have been removed from de-identified data. Because an important goal of de-identification is to prevent unauthorized re-identification, such attempts are sometimes called re-identification *attacks.*

The term "attack" is borrowed from the literature of computer security, in which the security of a computer system or encryption algorithm is analyzed through the use of a hypothetical "attacker" in possession of specific skills, knowledge, and access. A risk assessment involves cataloging the range of potential attackers and, for each, the probability of success.

[23] As noted previously, the HIPAA Privacy Rule only requires data use agreements with *limited data sets*, and not with data sets that have been de-identified according to either the Expert Determination method or the Safe Harbor method. Nevertheless, there is always a risk that de-identified data may be re-identified—even if the data are de-identified according to the HIPAA standard. For that reason, organizations that wish to exchange de-identified data while minimizing the risk of re-identification may wish to employ data use agreements as an additional administrative control.

There are many reasons that an individual or organization might attempt a re-identification attack:

- **To test the quality of the de-identification.** For example, a researcher might conduct the re-identification attack at the request of the data controller. If the data originally contained some kind of legally protected information, such as protected health information or education records, researchers should have appropriate confidentiality agreements in place in advance and exercise appropriate security procedures, otherwise a successful re-identification might be considered a "breach" and trigger mandatory reporting requirements under applicable statutes or policies.
- **To gain publicity or professional standing for performing the re-identification.** Several successful re-identification efforts have been newsworthy and professionally rewarding for the researchers conducting them. Note that these attacks were conducted legally, without a DUA prohibiting re-identification.
- **To embarrass or harm the organization that performed the de-identification.** Organizations that perform de-identification generally have an obligation to protect the personal information contained in the original data. As such, demonstrating that inadequate privacy protecting measures were employed can embarrass or harm these organizations, especially if the de-identified data were publicly released.
- **To gain direct benefit from the re-identified data.** For example, a marketing company might purchase de-identified health information and attempt to match up the information with identities, so that the re-identified individuals could be sent targeted coupons for prescription medicines.
- **To cause problems such as embarrassment or harm to an individual whose sensitive information can be learned by re-identification.** The problem might result from the direct publicizing of the information, or the information might be used to threaten or blackmail the individual, to force a resignation, or to force some other negative outcome.

In the literature, re-identification attacks are sometimes described as being performed by a hypothetical *data intruder* who is in possession of the de-identified dataset and some additional *background information*. (Note that in the case of a publicly released data set, the data intruder does not need to have gained unauthorized access to a computer system or data set.)

Re-identification risk[24] is the measure of the risk that the identifiers and other information about individuals in the dataset can be learned from the de-identified data. It is complex to quantify this risk, as the ability to re-identify depends on the original dataset, the de-identification technique, the technical skill of the attacker, the attacker's available resources, and the availability of additional data that can be linked with the de-identified data. In many cases, the risk of re-

[24] Elliot M, Dale A. Scenarios of attack: the data intruders perspective on statistical disclosure risk Netherlands Official Statistics 1999;14(Spring):6-10.

identification will increase over time as techniques improve and more contextual information become available (*e.g.*, publicly or through a purchase). For this reason, it is not possible to algorithmically determine what kinds of contextual information can be used to assist in future re-identification efforts.

Researchers have taken various approaches for computing and reporting the re-identification risk. Such efforts typically include a scenario that describes the measure of success and some role used to gauge the attacker's resources and abilities. Re-identification scenarios include:

- The risk that a specific person in the dataset can be re-identified when the attacker knows they are in the dataset. (The "prosecutor scenario.")
- The risk that there exists at least one person in the dataset who can be re-identified. The point is to prove that someone can be re-identified. In this case, the goal of the re-identification is frequently to embarrass or discredit the organization that performed the de-identification. (The "journalist scenario.[25]")
- The percentage of identities in the dataset that can be correctly re-identified. (The "marketer scenario.")
- The distinguishability between an analysis performed on a dataset containing an individual and the same analysis performed on a dataset that does not contain the individual. (The "differential identifiability" scenario.[26])

Different standards used to describe the abilities of the data intruder include:

- A member of general public who has access to public information ("general public")
- A computer scientist skilled in re-identification ("expert")
- A member of the organization that produced the dataset ("insider")
- A member of the organization that is receiving the de-identified data but may have access to more background information than the general public ("insider recipient")
- An information broker that systematically acquires both identified and de-identified information, with the hope of combining the data to produce an enriched information product that can then be used internally or resold ("information broker")
- A friend or family member of the data subject with specific context ("nosy neighbor")

The purpose of de-identifying data is to allow some uses of the de-identified data while providing for some privacy protection by shielding the identity of the data subjects. These two

[25] Paass, G. (1988) Disclosure risk and disclosure avoidance for microdata. J. Bus. Econ. Statist., 6, 487-500.

[26] Jaewoo Lee and Chris Clifton. 2012. Differential identifiability. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12). ACM, New York, NY, USA, 1041-1049. DOI=10.1145/2339530.2339695 http://doi.acm.org/10.1145/2339530.2339695

goals are antagonistic, in that there is a trade-off between the amount of de-identification and the utility of the resulting data. However, de-identification opens up new uses for the data that were previously prohibited due to privacy concerns. It is thus the role of the data controller, standards bodies, regulators, lawmakers and courts to determine the appropriate level of security, and thereby the acceptable trade-off between de-identification and utility.

In some cases, the use of de-identified data can result in a harm or adverse privacy action for individuals in the data set. A common taxonomy divides these risks into three categories: identity disclosure, attribute disclosure, and inferential disclosure.[27]

Identity disclosure happens when an attacker can link a specific data item to a specific individual. Several scenarios can result in identity disclosure:

- ***Insufficient de-identification.*** Identifying information may inadvertently remain in the de-identified dataset, with insufficient controls in place. This was the case in search query data released by AOL in 2006: AOL removed some identifying information but preserved the search terms that the users had typed. The data identified the user with a single numeric code. Because the code was a randomly generated pseudonym it could not itself be tied back to the users' identity. Identifying information did appear in the search queries themselves (for example, people who searched for information about their property), and the existence of the pseudonym allowed matching multiple queries from the same user with one another. Journalists were able to identify several users from those terms and contacted the users for comments.[28] In another case, Sweeney showed that some patients in de-identified Washington State hospital discharge records could be identified by manually correlating information in the discharge records with newspaper articles describing the accident that caused the hospitalization.[29]
- ***Re-identification by linking.*** It may be possible to re-identify specific records by linking some of the remaining data with similar attributes in another, identifying dataset. For example, de-identification of search records might remove the searcher's name but leave an IP address, allowing the data to be linked against a database that maps IP addresses to names.
- ***Pseudonym reversal***. If the data were pseudonymized, and if the pseudonyms are derived from identity information, it may be possible to reverse the pseudonymization process. This was the case in a dataset of taxi rides released by the New York City Taxi and Limousine Commission, discussed in Section 3.2.

Attribute disclosure happens if a piece of confidential information can be attributed to a subject.

---

[27] Li Xiong, James Gardner, Pawel Jurczyk, and James J. Lu, "Privacy-Preserving Information Discovery on EHRs," in Information Discovery on Electronic Health Records, edited by Vagelis Hristidis, CRC Press, 2009.

[28] Barbaro M, Zeller Jr. T. A Face Is Exposed for AOL Searcher No. 4417749 New York Times. 9 August, 2006.

[29] Sweeney L. Matching Known Patients to Health Records in Washington State Data. Harvard University. Data Privacy Lab. 1089-1. June 2013.

Identity disclosure invariably results in attribute disclosure if the dataset includes any confidential information. However, attribute disclosure can happen without identity disclosure if a dataset reveals that all individuals who share a characteristic have a particular attribute, and if the adversary knows of an individual in the sample who has that characteristic. For example, if a hospital releases information showing that all 20-year-old female patients treated had a specific diagnosis, and if Alice Smith is a 20-year-old female that is known to have been treated at the hospital, then Alice Smith's diagnosis can be inferred, even though her individual de-identified medical records cannot be distinguished from the others.[30] The L-diversity technique can help protect against inferential disclosure by assuring that each group of records matching a specific criteria has a certain amount of diversity. [31] This approach should be applied with care, however, as it can lower data utility even when it is not necessary to do so—for example, when there is no risk of stigmatization or harm to data subjects.

Inferential disclosure *"*occurs when information can be inferred with high confidence from statistical properties of the released data. For example, the data may show a high correlation between income and purchase price of a home. As the purchase price of a home is typically public information, a third party might use this information to infer the income of a data subject."[32]

Inferential disclosure may result in ***group harms*** to an entire class of individuals, including individuals whose data do not appear in the dataset. For example, if a specific demographic group is well represented in a data set, and if that group has a high rate of a stigmatizing diagnosis in the data set, then all individuals in that demographic may be stigmatized, even though it may not be statistically appropriate to do so.

US privacy policy reflected in laws and regulations is generally concerned with identity disclosure but not with other harms or problems that can result from the use or distribution of de-identified data. Organizations that wish to address these risks can attempt to address them through an explicit ethics review. Such a review should balance:

- The effort that the organization can spend performing and testing the de-identification process.
- The utility desired for the de-identified data (*i.e.*, benefits to the individuals and/or groups represented in the data).
- The problems for individuals or groups that might arise from the use of the de-identified data.

---

[30] El Emam, Methods for the de-identification of electronic health records for genomic research. Genome Medicine 2011, 3:25 http://genomemedicine.com/content/3/4/25

[31] The ***l-diversity*** technique can help protect against inferential disclosure by assuring that groups of records within a de-identified dataset represent a range of values. The disadvantage of this technique is that it lessens the fidelity of the de-identified data to the original dataset. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

[32] Glossary of statistical terms, OECD, November 30, 2005. https://stats.oecd.org/glossary/detail.asp?ID=6932

- The ability to use other controls that can minimize the risk.
- The probability that an attacker will attempt to re-identify the data and the amount of effort that the attacker might be willing to spend.

2.5 Release models and data controls

One way to limit the chance of re-identification is to place controls on the way that the data may be obtained and used. These controls can be classified according to different release models. Several named models have been proposed in the literature, ranging from no restrictions to tightly restricted:

- **The Release and Forget model:**[33] The de-identified data may be released to the public, typically by being published on the Internet. It can be difficult or impossible for an organization to recall the data once released in this fashion.
- **The Data Use Agreement (DUA) model:** The de-identified data may be made available to under a legally binding data use agreement that details what can and cannot be done with the data. Typically, data use agreements prohibit attempted re-identification, linking to other data, or redistribution of the data. A DUA will typically be negotiated between the data holder and qualified researchers (the "qualified investigator model" [34]), although they may be simply posted on the Internet with a click-through license agreement that must be agreed to before the data can be downloaded (the "click-through model" [35]).
- **The Enclave model:**[36,37] The de-identified data may be kept in some kind of segregated enclave that restricts the export of the original data, and instead accepts queries from qualified researchers, runs the queries on the de-identified data, and responds with results.

Robert Gellman, a privacy and information policy consultant, has proposed model legislation that would strengthen data use agreements.[38] Gellman's proposal would recognize a new category of information that he calls *potentially identifiable personal information (PI^2)*. Consenting parties could add to their data-use agreement a promise from the data provider that the data had been stripped of personal identifiers but still might be re-identifiable. Recipients would then face civil and criminal penalties if they attempted to re-identify. Thus, the proposed

[33] Ohm, Paul, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. UCLA Law Review, Vol. 57, p. 1701, 2010

[34] K El Emam and B Malin, "Appendix B: Concepts and Methods for De-identifying Clinical Trial Data," in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015

[35] Ibid.

[36] Ibid.

[37] O'Keefe, C. M. and Chipperfield, J. O. (2013), A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems. International Statistical Review, 81: 426–455. doi: 10.1111/insr.12021

[38] Gellman, Robert; *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 Fordham Intellectual Property, Media & Entertainment Law Journal 33 (2010), http://iplj.net/blog/wp-content/uploads/2013/09/Deidentificatin-Dilemma.pdf.

legislation would add to the confidence that de-identified data would remain so. "Because it cannot be known at any time whether information is re-identifiable, virtually all personal information that is not overtly identifiable is PI^2," Gellman notes.

3 Approaches for De-Identifying and Re-Identifying Structured Data

Most approaches for de-identifying structured data attempt to designate and the remove the specific identifying data elements from a dataset. This section introduces the terminology used by such schemes, discusses the two methods of the de-identification standard of the Health Insurance Portability and Privacy Act (HIPAA) Privacy Rule, and discusses critiques of these techniques and efforts that have appeared in the academic literature.

3.1 Removal of Direct Identifiers

Many de-identification approaches are easiest to understand when applied to a dataset containing a single table of data in which each row contains data for a different individual. A hypothetical dataset is shown in Table 1.

Direct identifiers, also called *directly identifying variables* and *direct identifying data*, are "data that directly identifies a single individual." [39] Examples of direct identifiers include names, social security numbers, and email addresses.

ISO/TS 25237:2008(E) defines direct identifiers as "data that can be used to identify a person without additional information or with cross-linking through other information that is in the public domain."[40] However, it is advisable to treat other individualized information such as medical record numbers and phone numbers as direct identifiers, even though additional information is required to link them to an identity, because these forms of identification are extensively used and thus available for linking to identities.

The HIPAA Privacy Rule further expands the definition of direct identifiers to include 18 specific data types, including names, telephone numbers, email addresses, and other unique identifying numbers, characteristics or codes; the complete list appears in Section 3.5.

Direct identifiers must be removed or otherwise transformed during de-identification. Approaches include:

- The direct identifiers can be removed.
- The direct identifiers can be replaced with either category names or data that are obviously generic. For example, names can be replaced with the phrase "PERSON NAME", addresses with the phrase "123 ANY ROAD, ANY TOWN, USA", and so on.
- The direct identifiers can be replaced with symbols such as "*****" or "XXXXX".

[39] ISO/TS 25237:2008(E), p.3

[40] ISO/TS 25237:2008(E), p.3

- The direct identifiers can be replaced with random values. If the same identity appears twice, it receives two different values. This preserves the form of the original data, allowing for some kinds of testing, but makes it harder to re-associate the data with individuals.
- The direct identifiers can be systematically replaced with pseudonyms, allowing records referencing the same individual to be matched. Pseudonymization is discussed in the next section.

Direct Identifiers								
Name	Address	Birthday	ZIP	Sex	Weight	Diagnosis	…	…

Table 1: A hypothetical data table showing direct identifiers

Early efforts to de-identify data stopped with the removal of direct identifiers. The resulting data could be re-identified through linkage attacks, discussed below in Section 3.3.

3.2 Pseudonymization

Pseudonymization is a specific kind of transformation in which the names and other information that directly identifies an individual are replaced with pseudonyms.[41] Pseudonymization allows linking information belonging to an individual across multiple data records or information systems, provided that all direct identifiers are systematically pseudonymized.

Pseudonymization can be readily reversed if the entity that performed the pseudonymization retains a table linking the original identities to the pseudonyms, or if the substitution is performed using an algorithm for which the parameters are known or can be discovered.

Pseudonymization frequently allows for the pseudonyms to be *reversed* at some time in the future, re-identifying the data subjects. A pseudonymized dataset can be reversed if the mapping between the direct identifiers and the pseudonyms is preserved or can be re-generated.[42] For example, an identifier can be encrypted with a secret key to create a pseudonym; decrypting the key reversed the pseudonymization process, producing the original identifier. Under HIPAA, reversing direct identifiers back to subject identities may only be performed by an organization covered by HIPAA's rules (mostly healthcare providers), known as a *covered entity*.[43]

Guidance from the HHS Office for Human Research Protection (OHRP) states that if

[41] ISO/TS 25237:2008

[42] HIPAA specifically allows properly salted one-way cryptographic hashes to be used for pseudonymization, but only as part of the expert determination method, and only if the cryptographic keys are not disclosed. See Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, Office of Civil Rights, Health and Human Services, November 26, 2012, pages 21-22.

[43] 45 CFR 164.514 (c)

pseudonymization can be readily reversed, the data are considered to be "coded" and not anonymous under the Common Rule.[44] This guidance applies to federally funded research that may be performed using such data sets outside a HIPAA covered entity. However, if a data use agreement is in place that forbids the code key to be shared, then the data are not considered to be identifiable private information.

If a pseudonymized dataset is released without a data use agreement that prohibits re-identification, a recipient may attempt to reverse the pseudonyms or to re-identify based on identifying information. For example, in 2014 the New York City Taxi and Limousine Commission released a dataset of the 173 million taxi trips taken in New York City during the previous year. In an attempt to de-identify the dataset, the Commission replaced the taxi medallion numbers and driver license numbers with a one-way cryptographic hash. Users of the dataset discovered the hash algorithm and were able to reverse the pseudonymization by iterating through all possible medallion numbers and license numbers, determining the cryptographic hash of each, and replacing the hash with the original number.[45] Thus, the attempted de-identification did not actually protect the identity of the drivers. This is known as a *brute force attack*.

The ability to reverse pseudonyms depends on many factors, including whether the pseudonyms are generated randomly or by an algorithm (if the algorithm employed a random key), the availability of the key, and whether the pseudonyms are unique or reused. Randomly generated pseudonyms can only be reversed if the mapping is retained. The HIPAA Privacy Rule allows pseudonyms that are coded identifiers to be made public, because they are not derived from information connected to the individual.

Even if the mapping is not retained, the use of a consistent pseudonym across multiple datasets may make it easier to re-identify data using a linkage attack. For this reason, the Article 29 Data Protection Working Party (a working group of the European Commission) states that "pseudonymized data cannot be equated to anonymized information as they continue to allow an individual data subject to be singled out and linked across different data sets."[46] However, because pseudonymization reduces the linkability of data with the original identity, the opinion also describes pseudonymization as "a useful security measure."

Unique pseudonyms that are used for an extended period may pose an increased privacy risk as they are linked to an increasing amount of information. Likewise, long-lived device pseudonyms may pose a similar privacy risk, especially when compared to pseudonyms that are dynamically reassigned to multiple individuals.

3.3 Re-identification through Linkage Attacks

Another way to re-identify a dataset that has been de-identified is through a *linkage attack*. In

[44] U.S. Department of Health and Human Services, "OHRP—Guidance on Research Involving Coded Private Information or Biological Specimens," October 16, 2008. http://www.hhs.gov/ohrp/policy/cdebiol.html

[45] "Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset," Anthony Tockar, September 15, 2014, http://research.neustar.biz/author/atockar/

[46] Opinion 05/2014 on Anonymisation Techniques, Article 29 Data Protection Working Party, 0829/14/EN WP216, Adopted on 10 April 2014

this attack, each record in the de-identified dataset is linked with similar records in a second dataset that contains both the linking information and the identity of the data subject.

One of the most widely publicized linkage attacks was performed by Latanya Sweeney, who re-identified the medical records of Massachusetts governor William Weld as part of her graduate work at MIT in the 1990s. At the time, the state of Massachusetts was distributing a research dataset containing de-identified insurance reimbursement records of Massachusetts state employees that had been hospitalized. To protect the employees' privacy, their names were stripped from the dataset, but the employees' date of birth, zip code, and sex was preserved to allow for statistical analysis.

Knowing that Weld had recently been treated at a Massachusetts hospital, Sweeney was able to re-identify the governor's records by searching for the "de-identified" records that matched the Governor's date of birth, zip code, and sex. She learned this information from the Cambridge voter registration list, which she purchased for $20. Sweeney then generalized her findings, arguing that up to 87% of the U.S. population could be uniquely identified by their 5-digit ZIP code, date of birth, and sex based on the 1990 census.[47] Follow-up work by privacy researcher Phillip Golle computed a re-identification rate of 62% using the year 2000 census.[48] In a critique of the study, Columbia University professor Daniel C. Barth-Jones notes that only 55% of the Cambridge population was registered to vote in the year 1996-97, and thus no more than 55% of the Cambridge city population could have been re-identified by linking with voter roles.[49]

Sweeney's linkage attack can be demonstrated graphically:

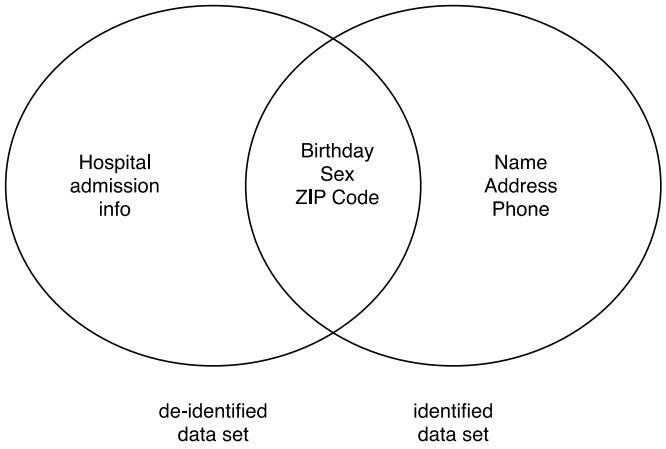

Figure 3: Linkage attacks combine information from two or more datasets to re-identify records

[47] Sweeney L., Simple Demographics Often Identify People Uniquely, Carnegie Mellon University, Data Privacy Working Paper 3, Pittsburgh, 2000. http://dataprivacylab.org/projects/identifiability/paper1.pdf

[48] Golle, Phillip, Revisiting the Uniqueness of Simple Demographics in the U.S. Population, in proceedings of WEPS 2006.

[49] Daniel C. Barth-Jones, "The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now," SSRN-id2116396, June 4, 2012. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397

Many factors complicate such linkage attacks:

- In order to be linkable, a person needs to be in both datasets. Sweeney knew that Weld was in both datasets because she knew that Weld has been hospitalized and that, as Governor, he was a registered voter.
- Only records that are uniquely distinguished by the linking variables in both sets can be linked uniquely, otherwise one needs to take a probabilistic view of the linking (*e.g.*, if there are two possible matches, then the probability a correct match is 0.5). In the case of a unique match, a person's records can only be linked if no one else shares their same birthday, sex, and ZIP in either dataset. As it turned out, no other person in Cambridge voter registration list shared Weld's date of birth. However, if Weld's records had been de-identified to today's HIPAA standard, the records would have only included his birth year and the first three digits of his ZIP code. Under these conditions, the unique match would probably not have been possible.
- If the variables are not the same in both datasets, then the data must be normalized or otherwise made consistent for linking to take place. This normalization can introduce errors. This was not an issue in the Weld case, but it could be an issue if one dataset reported "age" and another reported "birthday."
- A person could appear unique in the dataset but not be unique in the population. That is, there might have been another person in Cambridge with Weld's birthday, but the analysis would have missed that person if the individual had not been registered to vote.
- Verifying whether a link is correct requires using information that was not used as part of the linkage operation. In this case, Weld's medical records could be verified using newspaper accounts of Weld's hospitalization.

Re-identification can be administratively prohibited through the data use agreements and click-through license agreements. However, these agreements may be difficult to enforce for de-identified data sets that are made freely available (for example, by publishing them on the Internet).

3.4 De-identification of Quasi-Identifiers

Quasi-identifiers, also called *indirect identifiers* or *indirectly identifying variables*, are identifiers that by themselves do not identify a specific individual but can be aggregated and "linked" with other information to identify data subjects.[50] In Sweeney's re-identification of William Weld's medical records because birthday, ZIP, and sex are quasi-identifiers.

[50] Dalenius, Finding a Needle in a Haystack, or Identifying Anonymous Census Records, Journal of Official Statistics 2:3, 329-336, 1986.

Direct Identifiers		Quasi-Identifiers						
Name	Address	Birthday	ZIP	Sex	Weight	Diagnosis	…	…

Table 2: A hypothetical data table showing direct identifiers and quasi-identifiers

Quasi-identifiers pose a significant challenge for de-identification. Whereas direct identifiers can be removed from the dataset, quasi-identifiers generally convey some sort of information that might be important for a later analysis and removing them may damage the utility of the dataset. As such, they require careful consideration to balance the risk of re-identification with the utility gained by their inclusion.

Several methods are used for de-identifying quasi-identifiers:

1) **Suppression:** The quasi-identifier can be suppressed or removed. Removing the data maximizes privacy protection, but may decrease the utility of the dataset.

2) **Generalization:** Specific quasi-identifier values can be reported as being within a given range or as a member of a set. For example, the ZIP code 12345 could be generalized to a ZIP code between 12000 and 12999. Generalization can be applied to the entire dataset or to specific records—for example, identifying outliers.

3) **Perturbation:** Specific values can be replaced with other values in a manner that is consistent for each individual, within a defined level of generalization. For example, all ages may be randomly adjusted (-2 ... 2) years of the original age, or dates or hospital admissions and discharges may be systematically moved the same number of (-1000 ... 1000) days.[51]

4) **Swapping:** Quasi-identifier values can be exchanged between records, within a defined levels of generalization. Swapping must be handled with care if it is necessary to preserve statistical properties.

5) **Sub-sampling:** Instead of releasing an entire dataset, the de-identifying organization can release a sample. If only a subsample is released, the probability of re-identification decreases.[52]

K-anonymity[53] is a framework developed by Sweeney for quantifying the amount of

[51] Office of Civil Rights, "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule", US Department of Health and Human Services, 2010. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

[52] El Emam, Methods for the de-identification of electronic health records for genomic research, Genome Medicine 2011, 3:25 http://genomemedicine.com/content/3/4/25

[53] Latanya Sweeney. 2002. *k*-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.* 10, 5

manipulation required of the quasi-identifiers to achieve a given desired level of privacy. The technique is based on the concept of an *equivalence class***,** the set of records that match on all quasi-identifier values. A dataset is said to be *k-anonymous* if, for every combination of quasi-identifiers, there are at least *k* matching records. For example, if a dataset that has the quasi-identifiers birth year and state has k=4 anonymity, then there are at least four records for every combination of (birth year, state). Subsequent work has refined *k-anonymity* by adding requirements for diversity of the sensitive attributes within each equivalence class (*l-diversity*),[54] and requiring that the resulting data are statistically close to the original data (*t-closeness*).[55]

Professors Khaled El Emam and Bradley Malin[56] have developed an 11-step process for de-identifying data based on the classification of identifiers and quasi-identifiers:

- **Step 1: Determine direct identifiers in the dataset.** An expert determines the elements in the dataset that serve only to identify the data subjects.
- **Step 2: Mask (transform) direct identifiers.** The direct identifiers are either removed or replaced with pseudonyms.
- **Step 3: Perform threat modeling.** The organization determines "plausible adversaries," the additional information they might be able to use for re-identification, and the quasi-identifiers that an adversary might use for re-identification.
- **Step 4: Determine minimal acceptable data utility.** In this step, the organization determines what uses can or will be made with the de-identified data, to determine the maximal amount of de-identification for each field that could take place.
- **Step 5: Determine the re-identification risk threshold.** The organization determines acceptable risk for working with the dataset and possibly mitigating controls, based on strong precedents and standards (*e.g.*, Working Paper 22: Report on Statistical Disclosure Control) .
- **Step 6: Import (sample) data from the source database.** Because the effort to acquire data from the source (identified) database may be substantial, the authors recommend a test data import run to assist in planning.
- **Step 7: Evaluate the actual re-identification risk.** The actual identification risk is calculated.

(October 2002), 557-570. DOI=10.1142/S0218488502001648 http://dx.doi.org/10.1142/S0218488502001648

[54] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

[55] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian (2007). "t-Closeness: Privacy beyond k-anonymity and l-diversity". ICDE (Purdue University).

[56] K. El Emam and B. Malin, "Appendix B: Concepts and Methods for De-identifying Clinical Trial Data," in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015

- **Step 8: Compare the actual risk with the threshold.** The result of step 5 and step 7 are compared.
- **Step 9: Set parameters and apply data transformations.** If the actual risk is acceptable, the de-identification parameters are applied and the data is transformed. If the risk is too high, then new parameters or transformations need to be considered.
- **Step 10: Perform diagnostics on the solution.** Perform analyses on the de-identified data to make sure that it has sufficient utility and that re-identification is not possible within the allowable parameters.
- **Step 11: Export transformed data to external dataset.** Finally, the de-identified data are exported and the de-identification techniques are documented in a written report.

The chief criticism of de-identification based on direct and quasi-identifiers is that administrative determinations of quasi-identifiers may miss variables that can be uniquely identifying when combined and linked with external data—including data that are not available at the time the de-identification is performed, but become available in the future.

It is important to be realistic and consider plausible attacks, especially when there are data use agreements that prohibit re-identification, linking to other data, and sharing without permission. Besides the standards which give direction on the selection of identifiers and precedents for acceptable levels of risk, an evaluation or re-identification risk can be limited to the amount of information that an adversary can realistically know (the "attacker's power"[57]).

3.5 De-identification of Protected Health Information (PHI) under HIPAA

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule describes two approaches for de-identifying Protected Health Information (PHI): The Expert Determination Method (§164.514(b)(1)) and the Safe Harbor method (§164.514(b)(2)). Neither method promises a foolproof method of de-identification with zero risk of re-identification. Instead, the methods are intended to be practical approaches to allow de-identified healthcare information to be created and shared with a low risk of re-identification.

3.5.1 The HIPAA Expert Determination Method

The "Expert Determination" method provides for an expert who examines the data and determines an appropriate means for de-identification that minimizes the risk of re-identification. The specific language of the Privacy Rule states:

> "(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

[57] Yabo Xu, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. 2008. Anonymizing transaction databases for publication. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '08). ACM, New York, NY, USA, 767-775. DOI=10.1145/1401890.1401982 http://doi.acm.org/10.1145/1401890.1401982

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;"

The El Emam and Malin methodology in Section 3.4 is an example of an expert determination method.

Apart from this section, neither the Privacy Rule nor the implementation guidance provided by the Department of Health and Human Services Office of Civil Rights[58] specify standards or qualifications for the expert, nor does they specify requirements for organizations using experts to release the expert's determination or even to acknowledge that an expert determination has been made. They likewise do not specify how the risk of re-identification should be computed or quantified, nor do they specify the minimum allowable risk of re-identification, other than saying that it must be "very small."

The Expert Determination method specifies that "generally accepted statistical and scientific principles and methods" must be known and employed by the expert, which would imply an understanding of the relevant literature on statistical disclosure control and de-identification methods. Strong precedents for de-identification and risk levels also exist[59] and should be employed in an effective de-identification methodology.

3.5.2 The HIPAA Safe Harbor Method

The "Safe Harbor" method allows a covered entity to treat data as de-identified by removing 18 specific types of data for "the individual or relatives, employers, or household members of the individual." The 18 types are:

"(A) Names
(B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:
(1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
(2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000
(C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
(D) Telephone numbers

[58] Office of Civil Rights, "Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule", US Department of Health and Human Services, 2013. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

[59] For example, see Working Paper 22: Report on Statistical Disclosure Control.

(E) Fax numbers
(F) Email addresses
(G) Social security numbers
(H) Medical record numbers
(I) Health plan beneficiary numbers
(J) Account numbers
(K) Certificate/license numbers
(L) Vehicle identifiers and serial numbers, including license plate numbers
(M) Device identifiers and serial numbers
(N) Web Universal Resource Locators (URLs)
(O) Internet Protocol (IP) addresses
(P) Biometric identifiers, including finger and voiceprints
(Q) Full-face photographs and any comparable images
(R) Any other unique identifying number, characteristic, or code"

Compared to the Expert Determination method, the Safe Harbor method offers the promise of a straightforward application of rules, a repeatable process, and a known result: a dataset that is legally de-identified.

Under either method, the covered entity performing the de-identification must not "have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information."[60] However, HHS has issued specific guidance that simply knowing about the existence of re-identification techniques does not meet the "actual knowledge" standard:

> **Q:** "3.7: If a covered entity knows of specific studies about methods to re-identify health information or use de-identified health information alone or in combination with other information to identify an individual, does this necessarily mean a covered entity has actual knowledge under the Safe Harbor method?
>
> **A:** "No. Much has been written about the capabilities of researchers with certain analytic and quantitative capacities to combine information in particular ways to identify health information.[61,62,63,64] A covered entity may be aware of studies about methods to identify remaining information or using de-identified information alone or in combination with other information to identify an individual. However, a covered entity's mere

[60] §164.514 (c)

[61] K. El Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk. Evaluating the risk of re-identification of patients from hospital prescription records. Canadian Journal of Hospital Pharmacy. 2009; 62(4): 307-319.

[62] G. Loukides, J. Denny, and B. Malin. The disclosure of diagnosis codes can breach research participants privacy. *Journal of the American Medical Informatics Association Annual Symposium*. 2010; 17(3): 322-327

[63] B. Malin and L. Sweeney. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. *Journal of Biomedical Informatics*. 2004; 37(3): 179-192.

[64] L. Sweeney. Data sharing under HIPAA: 12 years later. A presentation at the Workshop on the HIPAA Privacy Rule's De-Identification Standard. Washington, DC. March 8-9, 2010.

knowledge of these studies and methods, by itself, does not mean it has "actual knowledge" that these methods would be used with the data it is disclosing. OCR does not expect a covered entity to presume such capacities of all potential recipients of de-identified data. This would not be consistent with the intent of the Safe Harbor method, which was to provide covered entities with a simple method to determine if the information is adequately de-identified."[65]

3.5.3 Evaluating the effectiveness of the HIPAA Safe Harbor Method

The HIPAA Safe Harbor method is heavily influenced by Sweeney's research: it cites her work and pays specific attention to the quasi-identifiers that she identified for generalization (ZIP code and birthdate). The method appears designed to strike a balance between the risk of re-identification and the need to retain some utility in the dataset—for example, by allowing the reporting of the first three digits of the ZIP code and the year of birth.

There is disagreement regarding the effectiveness of the HIPAA Safe Harbor method at de-identifying medical records and in the re-identification risk of the resulting data. Sweeney estimated a nation-wide unique re-identification risk of 0.04%, meaning that 4 in 10,000 records could be re-identified uniquely.[66] Kathleen Benitez and Bradley Malin at the Vanderbilt Health Information Privacy Lab performed a state-by-state study and concluded that the risk of unique re-identification was between 0.01% and 0.25%.[67] But this is only limited to data in which the only identifying information is year of birth, sex, and 3-digit ZIP code. In her attack of the health data sold by Washington State, Sweeney showed that there is also a risk from other quasi-identifiers besides those mentioned in Safe Harbor.

In 2010, the Office of the National Coordinator for Health Information Technology (ONC HIT) at the U.S. Department of Health and Human Services conducted a test of the HIPAA Safe Harbor method. As part of the study, researchers were provided with 15,000 hospital admission records belonging to Hispanic individuals from a hospital system for the years 2004–2009. Researchers then attempted to match the de-identified records to a commercially available dataset of 30,000 records from InfoUSA, a company that claims to have data on 235 million US consumers.[68] Based on the U.S. Census data the researchers estimated that the 30,000 commercial records covered approximately 5,000 of the hospital patients. When the experimenters matched using sex, ZIP3 (the first 3 digits of the ZIP code, as allowed by HIPAA), and age, they found 216 unique records in the hospital data, 84 unique records in the InfoUSA data, and 20 records that matched on both sides. The researchers then examined each of

---

[65] *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*, US Department of Health and Human Services, Office for Civil Rights, 2010. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#_edn32

[66] Latanya Sweeney, Testimony before that National Center for Vital and Health Statistics Workgroup for Secondary Uses of Health information. August 23, 2007. As quoted in Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, Office of Civil Rights, Health and Human Services, November 26, 2012. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

[67] Kathleen Benitez and Bradley Malin, Evaluating re-identification risks with respect to the HIPAA privacy rule," J. Am Med Inform Assoc. 2010; 17:169-177.

[68] http://www.infousa.com, accessed June 4, 2015

these 20 matches and determined the only two out of the 20 had the same last name, street address, and phone number.[69] This represents a re-identification rate of 0.013% uniques; the researchers also calculated a re-identification risk of 0.22% uniques using a more conservative methodology. These rates are not a nationwide average, since they are based on a single ethnic population in a single healthcare system.[70]

3.5.4 HIPAA Limited Datasets

In addition to data that contain protected health information and data that have been de-identified, the HIPAA Privacy Rule recognizes a third category of data: *limited datasets*. Under the Privacy Rule, a limited dataset is a dataset that has been partially de-identified but that still includes dates, city, state, zip code, and age. Such data are considered protected health information but may be shared for research, public health, or health care operations if the organizations sharing the data execute a *data use agreement*. At a minimum, the data use agreement must require security safeguards, require that all users of the data be similarly limited, and prohibit re-identification, linking to other data without permission, and contacting of the data subjects.[71]

Benitez and Malin determined that the risk of re-identification in a limited dataset ranges from 10% to 60%,[72,73] further demonstrating that limited datasets should be not be considered properly de-identified.

3.6 Evaluation of Field-Based De-identification

The basic assumption of de-identification is that some of the data fields in a dataset might contain useful information without being potentially identifying.

In recent years, a body of academic research has shown that many data fields may be identifying, and that it is frequently possible to single out individuals in high-dimensional data where there is access to suitable data with the identities of data subjects and no prohibitions on re-identification or linking:

- ***The Netflix Prize:*** Narayanan and Shmatikov showed in 2008 that in many cases the set of movies that a person had watched could be used as an identifier.[74] Netflix had released a dataset of movies that some of its customers had watched and ranked as part of its "Netflix Prize" competition. Although there was no direct identifiers in the dataset, the

[69] Because the researcher's "actual matches" only considered last name, the matched records might represent the same person or family members living together.

[70] Peter K. Kwok and Deborah Lafky, "Harder Than You Think: A Case Study of Re-identification Risk of HIPAA-Compliant Records," Joint Statistical Meeting, August 2, 2011 (abstract only).

[71] http://privacyruleandresearch.nih.gov/pr_08.asp

[72] Deborah Lafky, The Safe Harbor Method of De-Identification: An Empirical Test, Department of Health and Human Services, Office of the National Coordinator for Health Information Technology, October 8, 2009. http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf

[73] Benitez and Malin, ibid.

[74] Narayanan, Arvind and Shmatikov Vitaly: Robust De-anonymization of Large Sparse Datasets. IEEE Symposium on Security and Privacy 2008: 111-125

researchers showed that a set of movies watched (especially less popular films, such as cult classics and foreign films) could frequently be used to match a user profile from the Netflix dataset to a single user profile in the Internet Movie Data Base (IMDB), which had not been de-identified and included user names, many of which were real names. The threat scenario is that by rating a few movies on IMDB, a person might inadvertently reveal *all* of the movies that they had watched, since the person's IMDB profile could be linked with the Netflix Prize data.

- ***Medical Tests:*** Atreya *et al.* showed in 2013 that 5-7 laboratory results from a patient could be used "as a search key to discover the corresponding record in a de-identified biomedical research database."[75] Using a Vanderbilt University dataset of 8.5 million de-identified laboratory results from 61,280 patients, the researchers found that four consecutive laboratory test results uniquely distinguished between 34% and 100% of those in the dataset, depending on the test. The two most common test results, CHEM7 and CBC, respectively distinguished 98.9% and 98.8% of the test subjects. The threat scenario is that a person who intercepted a single lab identified lab report containing a CHEM7 or CBC result (perhaps by finding it in the trash) could use that report to search the de-identified biomedical research database for other records belonging to the individual.

- ***Credit Card Transactions:*** Working with a collection of de-identified credit card transactions from a sample of 1.1 million people from an unnamed country, Montjoye *et al.* showed that four distinct points in space and time were sufficient to specify uniquely 90% of the individuals in their sample.[76] Lowering the geographical resolution and binning transaction values (*e.g.*, reporting a purchase of $14.86 as between $10.00 and $19.99) increased the number of points required.

- ***Mobility Traces:*** Montjoye *et al.* showed that people and vehicles could be identified by their "mobility traces" (a record of locations and times that the person or vehicle visited). In their study, trace data from a sample of 1.5 million individuals was processed, with time values being generalized to the hour and spatial data generalized to the resolution provided by a cell phone system (typically 10-20 city blocks). The researchers found that four randomly chosen observations of an individual putting them at a specific place and time was sufficient to uniquely identify 95% of the data subjects.[77] Space/time points for individuals can be collected from a variety of sources, including purchases with a credit

[75] Atreya, Ravi V, Joshua C Smith,Allison B McCoy, Bradley Malin and Randolph A Miller, "Reducing patient re-identification risk for laboratory results within research datasets," J Am Med Inform Assoc 2013;20:95–101. doi:10.1136/amiajnl-2012-001026.

[76] Yves-Alexandre de Montjoye *et al.*, Unique in the shopping mall: On the reidentifiability of credit card metadata, Science, 30 January 2015, Vol 347, Issue 6221

[77] Yves-Alexandre de Montjoye *et al.*, Unique in the Crowd: The privacy bounds of human mobility, Scientific Reports 3 (2013), Article 1376.

card, a photograph, or Internet usage. A similar study performed by Ma *et al.* found that 30%-50% of individuals could be identified with 10 pieces of side information.[78] The threat scenario is that a person who revealed five place/time pairs (perhaps by sending email from work and home at four times over the course of a month) would make it possible for an attacker to identify his or her entire mobility trace in a publicly released dataset. As above, the attacker would need to know that the target was in the data.

- ***Taxi Ride Data:*** In 2014 The New York City Taxi and Limousine Commission released a dataset containing a record of every New York City taxi trip in 2013 (173 million in total). The data did not include the names of the taxi drivers or riders, but it did include a 32-digit alphanumeric code that could be readily converted to each taxi's medallion number.[79] A data scientist intern at the company Neustar discovered that he could find time-stamped photographs on the web of celebrities entering or leaving taxis in which the medallion number was clearly visible.[80] With this information, the intern was able to discover the other end-point of the ride, the amount paid, and the amount tipped for two of the 173 million taxi rides. A reporter at the Gawker website was able to identify another nine.[81]

The experience with the taxi data demonstrates there are many unanticipated sources of data that might correlate with other information in the data record and shows extreme care must be taken when transforming direct identifiers. The taxi and mobility trace studies demonstrate the strong identification power of geospatial information: just a few observations of a person's location and time can be highly identifying, even in a dataset that is generalized and noisy. Furthermore, some locations are highly identifying—either because they are isolated or well photographed.

A critique of these types of studies is that they frequently fail to distinguish between uniqueness in the sample and uniqueness in the population. For example, Montjoye *et al.*'s study finding that an individual can be distinguished from four geospatial points only applies if the attacker has additional information that the targeted individual is in the sample. If the sample size is increased to the entire population, the number of points required for uniqueness might increase as well.[82,83]

---

[78] Ma, C.Y.T.; Yau, D.K.Y.; Yip, N.K.; Rao, N.S.V., "Privacy Vulnerability of Published Anonymous Mobility Traces," Networking, IEEE/ACM Transactions on, vol.21, no.3, pp.720,733, June 2013

[79] The 32-digit code was the MD5 cryptographic hash of the taxi medallion number. The MD5 algorithm cannot be inverted, but there are such a small number of possible taxi medallion numbers that it was straightforward to use a brute force attack and hash them all. The problem could have been avoided if the Taxi and Limousine Commission had used a keyed hash and not released the key, if the medallion numbers had been encrypted instead of hashed (and the encryption key had not been released), or if the TLC had created a randomly generated code for each medallion number.

[80] "Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset," Anthony Tockar, September 15, 2014, http://research.neustar.biz/author/atockar/

[81] "Public NYC Taxicab Database Lets You See How Celebrities Tip," J. K. Trotter, GAWKER, October 23, 2014. http://gawker.com/the-public-nyc-taxicab-database-that-accidentally-track-1646724546

[82] Daniel Barth-Jones and Khaled El Emam, Assessing data intrusion threats, Science, 10 April 2015, VOl 348 Issue 6231, p. 194.

[83] Jane Yakowitz Bambauer, Is De-Identification Dead Again, INFO/LAW Blog, April 28, 2015.

Furthermore, both the medical tests and taxi studies also show that relatively small perturbations to the data may make re-identification difficult or impossible. In Atreya *et al.*'s examination of medical test results, the authors present a simple de-identification algorithm that adds adaptive noise to the clinical values in a way that eliminates the distinguishing characteristics without significantly changing clinical importance. In the case of the taxi study, the celebrities were identified because the taxi medallion pseudonyms could be reversed. If the medallion number had been properly protected and if the GPS location data had been aggregated to a 100-meter square grid, the risk of re-identification would have been considerably reduced, and the privacy impact would have been negated.

El Emam *et al.*[84] reviewed 14 re-identification attempts published between 2001 and 2010. For each, the authors determined whether or not health data had been included, the profession of the adversary, the country where the re-identification took place, the percentage of the records that had been re-identified, the standards that were followed for de-identification, and whether or not the re-identification had been verified. The researchers found that the successful re-identification events typically involved small datasets that had not been de-identified according to existing standards. In many cases the re-identification researchers had re-identified just a few records, and those re-identifications may not have been validated. As such, drawing scientific conclusions from these cases is difficult.

3.7 Estimation of Re-Identification Risk

As the examples above demonstrate, individuals and organizations working with personal data are in need of easy-to-use and reliable procedures for calculating the risk of re-identification given a specific de-identification protocol. Calculating this risk is complicated, as it depends on many factors, including the distinctiveness of different individuals within the sampled dataset, the de-identification algorithm, the availability of linkage data, and the range of individuals that might mount a re-identification attack.[85]

There are also different kinds of re-identification risk and many ways to report the risk. A model might report the average risk of each subject being identified, the risk that *any* subject will be identified, or the risk that individual subjects might be identified as being 1 of *k* different individuals, etc.

Dankar *et al.* propose a statistical model and decision rule for estimating the distinctiveness of different kinds of data sources.[86] El Emam *et al.* developed a technique for modeling the risk of re-identifying adverse drug event reports based on two attacker models: a "mildly motivated adversary" whose goal is to identify a single record, and a "highly motivated adversary" that

---

https://blogs.law.harvard.edu/infolaw/2015/04/28/is-de-identification-dead-again/

[84] K El Emam, E Jonker, L Arbuckle, B Malin (2011) A Systematic Review of Re-Identification Attacks on Health Data. PLoS ONE 6(12): e28071. doi:10.1371/journal.pone.0028071

[85] Catherine Marsh, Chris Skinner, Sara Arber, Bruce Penhale, Stan Openshaw, John Hobcraft, Denise Lievesley and Nigel Walford, The Case for Samples of Anonymized Records from the 1991 Census *Journal of the Royal Statistical Society. Series A (Statistics in Society)* Vol. 154, No. 2 (1991), pp. 305-340

[86] Dankar *et al.* Estimating the re-identification risk of clinical data sets, BMC Medical Informatics and Decision Making 2012, 12:66.

wishes to identify and verify all matches "and is only limited by practical or financial considerations."[87] In general, organizations should balance the decision to treat all data elements (continuous and categorical) as quasi-identifiers with the plausibility of re-identification and the usefulness of releasing data that has not been substantially modified. De-identification may result in significant damage to statistical relationships within the data, limiting the usefulness of data that have been de-identified.

4 Challenges in De-Identifying Unstructured Data

Whereas the last section was concerned mostly with the de-identification of structured data, this section concerns itself with the open challenges of de-identifying unstructured text and multimedia.

4.1 De-identifying medical text

Medical records contain significant amounts of unstructured text. In recent years, there has been an effort to develop and evaluate tools designed to remove the 18 HIPAA data elements from free-format text (*e.g.*, narrative intake reports) using natural language processing techniques. The two primary techniques explored have been rule-based systems and statistical systems. Rule-based systems tend to work well for specific kinds of text but may not work well when applied to new domains. Statistical tools generally perform less accurately than rule-based systems and require labeled training data, but are easier to repurpose to new domains.

Multiple factors combine to make de-identifying text narratives hard:

1) Direct identifiers such as names and addresses may not be clearly marked.
2) Important medical information may be mistaken for personal information and removed. This is especially a problem for eponyms that are commonly used in medicine to describe diseases (*e.g.*, Addison's Disease, Bell's Palsy, Reiter's Syndrome, etc.)
3) Even after the removal of the 18 HIPAA Safe Harbor elements, information may remain that allows identification of the medical subject.

Several researchers have performed formal evaluations of text de-identification tools:

- In 2012 Deleger *et al.* at Cincinnati Children's Hospital Medical Center tested The MITRE Identification Scrubber Toolkit (MIST)[88] against MCRF, an in-house system developed by CCHMC based on the MALLET machine-learning package. The reference

[87] El Emam *et al.*, Evaluating the risk of patient re-identification from adverse drug event reports, BMC Medical Informatics and Decision Making 2013, 13:114 http://www.biomedcentral.com/1472-6947/13/114

[88] Aberdeen J, Bayer S, Yeniterzi R, *et al.* The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int. J Med Inform 2010;79:849e59.

corpora were 3503 clinical notes selected from 5 million notes created at CCHMC in 2010, the 2006 i2b2 de-identification challenge corpus[89] and the PhysioNet corpus.[90,91]

- In 2013 Ferrández *et al.* at the University of Utah Department of Biomedical Informatics performed an evaluation of five automated de-identification systems against two reference corpora.[92] The test was conducted with the 2006 i2b2 de-identification challenge corpus, consisting of 889 documents that had been de-identified and then given synthetic data,[93] and a corpus of 800 documents provided by the Veterans Health Administration that was randomly drawn from documents with more than 500 words dated between 4/01/2008 and 3/31/2009.
- In 2013 The National Library of Medicine issued a report to its Board of Scientific Counselors entitled "Clinical Text De-Identification Research" in which the NLM compared the performance of its internally developed tool, the NLM Scrubber (NLM-S), with the MIT de-identification system (MITdeid) and MIST.[94] The test was conducted with an internal corpus of 1073 Physician Observation Reports and 2020 Patient Study Reports from the NIH Clinical Center.

Both the CCHMC and the University of Utah studies tested the systems both "out-of-the-box" and after they were tuned by using part of the corpus as training data. The Utah study found that none of the de-identification tools worked well enough to de-identify the VHA records for public release, and that the rule-based systems excelled at finding certain kinds of information (*e.g.*, SSNs and phone numbers), while the trainable systems worked better for other kinds of data. Although there are minor variations between the systems, they all had similar performance. The NLM study found that NLM-S significantly outperformed MIST and MITdeid on the NLM dataset, removing 99.2% of the identifiers matching the HIPAA Safe Harbor standard. The authors concluded that the remaining identifiers would not pose a significant threat to patient privacy.

A study by Meystre, Shen et. al showed the automatically de-identified discharge summaries from the Salt Lake City VHA Medical Center were not recognized by the patient's doctor or treating professional.[95] A study by Carrell *et. al* found that using realistic surrogate names in the

---

[89] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de- identification. J Am Med Inform Assoc. 2007;14:550e63.

[90] Neamatullah I, Douglass MM, Lehman LW, *et al.* Automated de-identification of free-text medical records. BMC Med Inform Decis Mak 2008;8:32.

[91] Goldberger AL, Amaral LA, Glass L, *et al.* PhysioBank, PhysioToolkit, and Physionet: components of a new research resource for complex physiologic signals. Circulation 2000;101:E215e20.

[92] Ferrández, O., South, B. R., Shen, S., Friedlin, F. J., Samore, M. H., & Meystre, S. M. (2013). BoB, a best-of-breed automated text de-identification system for VHA clinical documents. *Journal of the American Medical Informatics Association : JAMIA*, *20*(1), 77–83. doi:10.1136/amiajnl-2012-001020

[93] Uzuner O, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc 2007;14:550e63.

[94] Kayaalp M *et al.*, A report to the Board of Scientific Counselors, 2013, The Lister Hill National Center for Biomedical Communications, National Library of Medicine.

[95] Meystre S *et al.*, Can Physicians Recognize Their Own Patients in De-Identified Notes? In Health – For Continuity of Care C. Lovis *et al.* (Eds.) © 2014 European Federation for Medical Informatics and IOS Press.

de-identified text like "George Washington" and "1600 Pennsylvania Ave" instead of generic labels like "PATIENT" and "ADDRESS" could decrease or mitigate the risk of re-identification of the few names that remained in the text, because "the reviewers were unable to distinguish the residual (leaked) identifiers from the … surrogates."[96]

None of these systems attempt to de-identify data beyond removal of the 18 HIPAA Safe Harbor data elements, leaving the possibility that individuals could be re-identified using other information. For example, regulations in both the U.S. and Canada require reporting of adverse drug interactions. These reports have been re-identified by journalists and researchers by correlating reports of fatalities with other data sources, such as news reports and death registers.[97] Additional research is required to develop systems that can automatically redact such reports against motivated attackers that have access to all of the world's publicly available data.

4.2 De-identifying Photographs and Video

Still photographs, consumer videos, and surveillance video potentially contain a wealth of information that can be used to identify an individual. De-identification seeks to remove this identifying information so that individuals cannot be identified, while still allowing specific uses of the resulting videos.

Identifying data that was generated by the camera when a photo or video was taken, or was added by post-processing tools, may exist along with the image data in the form of embedded metadata in the image file. For example, a GPS address of the person's house, the serial number associated with a camera, or a person's name may be embedded in a header. Issues involved in de-identifying this information are similar to those discussed in Section 3 for de-identifying other kinds of structured information. An added complication, however, is that the data may be stored in binary structures that are not widely understood or documented. Thus, identifying data may be present but either unintelligible or hidden.

More complex are the range of issues that arise in attempting to de-identify the multimedia content itself. Multimedia content is rich in information that may identify a specific individual or make it possible to impose constraints on a person's identity—for example, allowing an observer to infer that an individual is a tall elderly female.

ICT COST Action IC1206, "De-identification for privacy protection in multimedia content," is exploring approaches for removing identifying information "such … as face, voice, silhouette and gait," from multimedia content.[98] As part of that work the committee has developed a taxonomy of identifiers in multimedia content:

---

[96] Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B., & Hirschman, L. (2013). Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. Journal of the American Medical Informatics Association, 20(2), 342-348.

[97] K El Emam, E Jonker, L Arbuckle, B MalinB (2011) A Systematic Review of Re-Identification Attacks on Health Data. PLoS ONE 6(12): e28071. doi:10.1371/journal.pone.0028071

[98] ICT COST Action IC1206, "De-identification for privacy protection in multimedia content," 27 November 2012, http://www.cost.eu/COST_Actions/ict/Actions/IC1206.

- ***Biometric identifiers*** that are distinctive, measurable, generally unique and permanent personal characteristics used to identify individuals. This includes physiological biometrics (face, iris, ear, fingerprint) and behavioral biometrics (voice, gait, gesture, lip-motion, and typing style);
- ***Soft biometrics*** of some vague physical, behavioral or adhered human characteristic that is not necessarily permanent or distinctive (height, weight, eye color, silhouette, age, gender, race, moles, tattoos, birthmarks, and scars);
- ***Non-biometric identifiers*** including text context, speech context, specific social-political and environmental context, dressing style, and hairstyle. [99]

De-identifying multimedia content involves modifying the imagery and audio information (if present) to make identification difficult or impossible. Biometric, soft biometric, and non-biometric identifiers are frequently present in concert and must all be de-identified in order to protect the privacy of individuals. This can be termed *multimodal de-identification*.

Early research had the goal of transforming imagery so that individuals could not be reliably identified using automated face recognition systems. For example, face blurring is used by Google Street View, one of the largest deployments of photo de-identification technology.[100] Some researchers have developed systems that can identify and blur bodies,[101] as research has shown that bodies can be identifiable without faces.[102] An experimental system can locate and remove identifying tattoos from some kinds of still images,[103] although it is unclear how the system would perform with stills with poor quality data, complex scenes, and partially occluded or highly angled views of the tattoos, or in video where many perspectives are captured.

In addition to de-identifying images of people, it is also necessary to address identifying cues in audio and text. Cunningham and Truta explored using "controlled audio distortion" to help produce a corpus that protected the identity of the speakers with minimal information loss.[104] Of course, simple audio distortion can only prevent identification based on the speaker's voice, but cannot remove other identifying cues that might exist in the audio track, such as speaking

---

[99] This taxonomy was provided by Slobodan Ribaric as part of the COST Action IC1206 comments on a previous draft of this report.

[100] Frome, Andrea, *et al.*, "Large-scale Privacy Protection in Google Street View," IEEE International Conference on Computer Vision (2009).

[101] Prachi Agrawal and P. J. Narayanan. 2009. Person de-identification in videos. In Proceedings of the 9th Asian conference on Computer Vision - Volume Part III (ACCV'09), Hongbin Zha, Rin-ichiro Taniguchi, and Stephen Maybank (Eds.), Vol. Part III. Springer-Verlag, Berlin, Heidelberg, 266-276. DOI=10.1007/978-3-642-12297-2_26 http://dx.doi.org/10.1007/978-3-642-12297-2_26

[102] Rice, Phillips, *et al.*, Unaware Person Recognition From the Body when Face Identification Fails, Psychological Science, November 2013, vol. 24, no. 11, 2235-2243 http://pss.sagepub.com/content/24/11/2235

[103] Darijan Marčetić *et al.*, An Experimental Tattoo De-identification System for Privacy Protection in Still Images, MIPRO 2014, 26-30 May 2014, Opatija, Croatia

[104] Scot Cunningham and Traian Marius Truta. 2008. Protecting privacy in recorded conversations. In Proceedings of the 2008 international workshop on Privacy and anonymity in information society (PAIS '08), Farshad Fotouhi, Li Xiong, and Traian Marius Truta (Eds.). ACM, New York, NY, USA, 26-35. DOI=10.1145/1379287.1379295 http://doi.acm.org/10.1145/1379287.1379295

patterns or the mention of names.

Evaluating the effectiveness of multimedia de-identification is a multidimensional problem, including:

- ***The precision and accuracy of identifying objects requiring de-identification.*** Google reports that its completely automatic system is able to blur 89% of faces and 94-96% of license plates.[105] Nevertheless, journalists have criticized Google for leaving many faces unblurred.[106] Journalists have also criticized Google for blurring the faces of religious effigies.[107,108] In some contexts, it may be unacceptable to blur or otherwise adulterate certain objects, symbols, or individuals.
- ***The reversibility of the transformation.*** Care must also be taken if pixelation or blurring are used for obscuring video, as technology exists for de-pixelating and de-blurring video by combining multiple images. As an alternative to pixilation or blurring, some researchers have developed systems that can replace faces with a composite face,[109,110] or with a face that is entirely synthetic.[111,112]
- ***The visual quality of the resulting imagery.*** Blurring and pixelation have the disadvantage of creating a picture that is visually jarring and could potentially affect one's interpretation of the scene.
- ***The effectiveness of the chosen identity obscuring techniques in actually obscuring identity.*** While some researchers may score the algorithms against face recognition software, other factors such as clothing, body pose, or geo-temporal setting might make the person identifiable by associates.

Whereas several attempts to quantify the effectiveness of de-identifying structured data and text were discussed in previous sections of this report, we have found no significant efforts to quantify the effectiveness of multimedia de-identification. A proper test of de-identification is

---

[105] Frome, Andrea, *et al.*, "Large-scale Privacy Protection in Google Street View," IEEE International Conference on Computer Vision (2009).

[106] Stephen Chapman, "Google Maps, Street View, and privacy: Try harder, Google," ZDNet, January 31, 2013. http://www.zdnet.com/article/google-maps-street-view-and-privacy-try-harder-google/

[107] Gonzalez, Robbie. "The Faceless Gods of Google Street View," io9, October 4, 2014. http://io9.com/the-faceless-gods-of-google-street-view-1642462649

[108] Brownlee, John, "The Anonymous Gods of Google Street View," Fast Company, October 7, 2014. http://www.fastcodesign.com/3036319/the-anonymous-gods-of-google-street-view#3

[109] Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando de la Torre, and Simon Baker. In: Protecting Privacy in Video Surveillance, A. Senior, editor. Springer, 2009 Preserving Privacy by De-identifying Facial Images. http://dataprivacylab.org/projects/facedeid/paper.pdf

[110] E. Newton, L. Sweeney, and B. Malin. Preserving Privacy by De-identifying Facial Images, Carnegie Mellon University, School of Computer Science, Technical Report, CMU-CS-03-119. Pittsburgh: March 2003.

[111] Saleh Mosaddegh, Löıc Simon, Frederic Jurie. Photorealistic Face de-Identification by Aggregating Donors' Face Components. Asian Conference on Computer Vision, Nov 2014, Singapore. pp.1-16.

[112] Umar Mohammed, Simon J. D. Prince, and Jan Kautz. 2009. Visio-lization: generating novel facial images. In ACM SIGGRAPH 2009 papers (SIGGRAPH '09), Hugues Hoppe (Ed.). ACM, New York, NY, USA, Article 57, 8 pages. DOI=10.1145/1576246.1531363 http://doi.acm.org/10.1145/1576246.1531363

complex as it could include a variety of re-identification scenarios:

- Attempted re-identification by an automated identification system.
- Re-identification by a trained human, but one who does not have personal knowledge of the data subject.
- Re-identification by associates or friends of the data subject.

4.3 De-Identifying Medical Imagery

There is a strong desire to share medical imagery to further a wide range of both scientific and operational goals, including broadening the pool of research data, allowing for replication and validation of scientific findings, and quality assessment. Privacy concerns and the HIPAA Privacy Rule require that imagery be de-identified prior to being shared.

Medical imagery consists of header information (metadata), typically in DICOM[113] (Digital Imaging and Communications in Medicine) format, and "pixels" (image data) generated by the imaging device. Information that may identify an individual can be present in either the header or in the pixels. Medical imagery thus poses a broad range of challenges for de-identification, including:

- The DICOM header may contain the patient's name or other direct and quasi-identifiers.
- The pixels may contain photographic information or other another biometric that is individually distinguishing. For example, a hospital might photograph an injury to a person's face, from which the person could be identified.
- The pixels may contain information that can be mathematically processed to produce a recognizable image or biometric.[114,115]
- Direct identifiers may be embedded as human-readable text in the pixels. For example, a patient name may be hand-drawn on records that have been digitized or photographically captured at the time that an image was created. Alternatively, the text may be "burned in" by software, often at a predictable or specified location within the imagery.

Unlike other file formats, the DICOM standard has provisions[116] for designating whether or not identifying information is present in either the header or pixel area of an image and for indicating

---

[113] The DICOM Standard 2015a, Medical Imaging and Technology Alliance, National Electrical Manufacturers Association, 2015. http://medical.nema.org/

[114] J. Chen, K. Siddiqui, L. Fort, R. Moffitt, K. Juluru, W. Kim, N. Safdar, and E. Siegel, "Observer success rates for identification of 3D surface reconstructed facial images and implications for patient privacy and security," in Proc. SPIE Int. Soc. Opt. Eng., 2007, pp. 65161B-1–65161B-8

[115] F. W. Prior, B. Brunsden, C. Hildebolt, T.S. Nolan, M. Pringle, S.N. Vaishnavi, L.J. Larson-Prior. Facial Recognition From Volume-Rendered Magnetic Resonance Imaging Data. IEEE Transactions On Information Technology In Biomedicine, 13:1, Jan 2009 p.5-9.

[116] Digital Imaging and Communications in Medicine (DICOM) Supplement 142: Clinical Trial De-Identification Profiles, DICOM Standards Committee, Working Group 18 Clinical Trials, Jan 25, 2011. ftp://medical.nema.org/medical/dicom/final/sup142_ft.pdf

whether a file has been de-identified. For example, DICOM's "Clean Pixel Data Option," if specified, can indicate whether identifying information has been burned in to the pixel data. The "Clean Recognizable Visual Features Option," if specified, can indicate if the information in the pixel data could be used to recognize the individual. DICOM PS3.15 Annex E, "Attribute Confidentiality Profiles," specifies a structured process for producing de-identified versions of DICOM images: identifying information can either be redacted or it can be encrypted and stored in an "Encrypted Attributes Dataset," so that de-identified files can be re-identified at a later point in time.

The DICOM standard does not specify the actual algorithms or techniques for de-identification, however. Such techniques would be specific to each imaging modality and are active research areas. For example:

- Freymann *et al.* developed an open-source software suite for de-identifying DICOM header information prior to images being stored in a public archive.[117] The authors note that support for the DICOM burned-in indicators is lacking among manufacturers, so it is necessary to for tools to scan for burned-in protected health information.
- Several researchers have demonstrated techniques for mathematically reconstructing a person's face using three-dimensional models generated from computed tomography (CT) and magnetic resonance (MR) imaging.[118,119] To counteract this threat, researchers have developed techniques for automatically removing the portion of an MR image that could be used to produce a face but which was not useful for diagnostic purposes.[120,121]

4.4 De-identifying Genetic information and biological materials

Genetic sequences and other kinds of sequence information are not considered to be identifying of individuals under the HIPAA Privacy Rule. Some commentators have faulted the Privacy Rule for this omission, since some genetic sequences are highly individualistic and genetic information is now routinely collected and placed in databanks for the express purpose of identifying individuals. Furthermore, because genetic information is inherited, genetic sequences have been used identify individuals who were not themselves sequenced but instead had relatives who were sequenced and placed in an identification database.

For example:

---

[117] John B. Freymann, Justin S. Kirby, *et al.*, Image Data Sharing for Biomedical Research—Meeting HIPAA Requirements for De-Identification, J. Digit Imaging (2012) 25:14-24.

[118] Kermi, A.; Marniche-Kermi, S.; Laskri, M.T., "3D-Computerized facial reconstructions from 3D-MRI of human heads using deformable model approach," *Machine and Web Intelligence (ICMWI), 2010 International Conference on*, vol., no., pp.276,282, 3-5 Oct. 2010 doi: 10.1109/ICMWI.2010.5648144

[119] Fred W. Prior, Barry Brunsden, Charles Hildebolt, Tracy S. Nolan, Michael Pringle, S. Neil Vaishnavi, and Linda J. Larson-Prior. 2009. Facial recognition from volume-rendered magnetic resonance imaging data. *Trans. Info. Tech.* Biomed. 13, 1 (January 2009), 5-9. DOI=10.1109/TITB.2008.2003335 http://dx.doi.org/10.1109/TITB.2008.2003335

[120] Bischoff-Grethe A, Ozyurt IB, Busa E, *et al.* A Technique for the Deidentification of Structural Brain MR Images. *Human brain mapping*. 2007;28(9):892-903. doi:10.1002/hbm.20312.

[121] Milchenko M, Marcus D, Obscuring surface anatomy in volumetric imaging data, Neuroinformatics 2013 Jan; 11(1):65-75.

- In 2005, a 15-year-old teenager used the DNA-testing service FamilyTreeDNA.com to find his sperm donor father. The service, which cost $289, did not identify the boy's father, but it did identify two men who had matching Y-chromosomes. The two men had the same surname but with different spellings. As the Y-Chromosome is passed directly from father to son with no modification, it tends to be inherited the same way as European surnames. With this information and with the sperm donor's date and place of birth (which had been provided to the boy's mother), the boy was able to identify his father using an online search service,[122] a technique now known as "genealogical triangulation."
- In 2013, a group of researchers at MIT extended the experiment, identifying surnames and complete identities of more than 50 individuals who had DNA tests released on the Internet as part of the Study of Human Polymorphisms (CEPH) project and the 1000 Genomes Project.[123]

Erlich and Narayanan have published a detail review of techniques for re-identifying genetic sequences and technical measures that could provide for protection against such attacks.[124]

Currently, there is no scientific consensus on the minimum size of a genetic sequence necessary for re-identification. There is also no consensus on an appropriate mechanism to make de-identified genetic information available to researchers without the need to execute a data use agreement that would prohibit re-identification.

4.5 De-identification of geographic and map data

De-identification of geographic data is an area of active research. Current methods rely on perturbation and generalization. Perturbation is problematical in some cases, because perturbed locations can become nonsensical (*e.g.*, moving a restaurant into a body of water).[125] Generalization may not be sufficient to hide identity, however, especially if the population is sparse or if multiple observations can be correlated—especially if those observations use different generalization areas.[126]

Without some kind of generalization or perturbation, there is so much diversity in geographic data that it may be extremely difficult to de-identify locations. For example, measurement of cell phone accelerometers taken over a timeframe can be used to infer position by fitting movements to a street grid.[127] Thus, acceleration motions sampled over time may be identifying when

---

[122] Sample, Ian. Teenager Finds Sperm Donor Dad on Internet. The Guardian, November 2, 2005. http://www.theguardian.com/science/2005/nov/03/genetics.news

[123] Gymrek *et al.*, Identifying Personal Genomes by Surname Inference, Science 18 Jan 2013, 339:6117.

[124] Yaniv Erlich and Arvind Narayanan, "Routes for breaching and protecting genetic privacy," Nature Reviews Genetics, June 2014, volume 15, p. 409

[125] Christopher A. Cassa, Shannon C. Wieland, and Kenneth D. Mandi, Re-identification of home addresses from spatial locations anonymized by Gaussian skew, International Journal of Health Geographics, 12 August 2008. http://www.ij-healthgeographics.com/content/7/1/45

[126] Philip Steel and Jon Sperling, "The Impact of Multiple Geographies and Geographic Detail on Disclosure Risk: Interactions between Census Tract and ZIP Code TRabulation Geography," U.S. Census Bureau staff report, 2001. https://www.census.gov/srd/sdc/steel.sperling.2001.pdf

[127] Jun Han; Owusu, E.; Nguyen, L.T.; Perrig, A.; Zhang, J., "ACComplice: Location inference using accelerometers on

combined with publicly available information.

5 Conclusion

De-identification techniques are intended to remove identifying information from a dataset while retaining some utility in the remaining data. De-identification can be used within an organization to minimize the privacy risk associated with data use or storage. De-identification can also be used prior to sharing or release of a dataset. In some cases, de-identified data can be released without further controls, while in others it is necessary to additional protection measures such as data use agreements that administratively restrict what recipients of the data can do with it.

A variety of problems can result from the distribution or use de-identified data. Some of these problems result if the de-identified data can be re-identified—that is, if the data can be matched back up with the original data subjects. Two important problems that can result are *disclosure of private facts* and *reputational damage*. Disclosure of private facts affects the individuals whose data were re-identified. Damage to reputation affects the organization that performed the de-identification.

Researchers or journalists have performed most of the publicized re-identifications, and many of those re-identified were public figures. Studies that use public figures to gauge the risk of re-identification may yield risks that are misleadingly high, as public figures are inherently easier to re-identify than private figures for the simple reason that more information about public figures is available in the public domain.

There remains among researchers considerable disagreement regarding the issue of re-identification risk. Among some researchers, the term is taken to mean the percentage of de-identified records that can be re-identified. In these cases, re-identification risk can be directly measured by performing a re-identification attack. Others take re-identification risk as the probability of record re-identification in the future. In this case, the risk must be estimated but is ultimately impossible to quantify, as the risk depends upon the availability of data in the future that may not be available now. The re-identification risk may also be estimated as the probability that *any* record can be re-identified. As a result, it is always necessary when discussing re-identification risk that the specific risk under the specific re-identification scenario be described.

Organizations endeavoring to share such data might consider employing a combination of several approaches to mitigate re-identification risk. These include technical controls, such as removing quasi-identifiers and other kinds of information that might be used to re-identify the data subjects; continuously surveying for data that could be linked to the de-identified information that they are sharing; controls on the de-identified data, such as data use agreements and click-through agreements that prohibit re-identification, linking to other data, or sharing with others; and technical controls that limit the activities of data recipients.

De-identification techniques are increasingly important as organization seek to use, share, and

---

smartphones," *Communication Systems and Networks (COMSNETS), 2012 Fourth International Conference on*, pp.1,9, 3-7 Jan. 2012

even monetize data that contain personal data. Yet after more than a decade of research, there is comparatively little known about the underlying science of de-identification. Many of the current techniques and procedures in use, such as the HIPAA Privacy Rule's Safe Harbor de-identification standard, are not firmly rooted in theory. There are no widely accepted standards for testing the effectiveness of a de-identification process or gauging the utility lost as a result of de-identification. Given the growing interest in de-identification, there is a clear need for standards and assessment techniques that can measurably address the breadth of data and risks described in this paper.

Appendix A Glossary

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

anonymity: "condition in identification whereby an entity can be recognized as distinct, without sufficient identity information to establish a link to a known identity" (ISO/IEC 24760-1:2011)

anonymization: "process that removes the association between the identifying dataset and the data subject" (ISO/TS 25237:2008)

anonymized data: "data from which the patient cannot be identified by the recipient of the information" (ISO/TS 25237:2008)

anonymous identifier: "identifier of a person which does not allow the unambiguous identification of the natural person" (ISO/TS 25237:2008)

attacker: person seeking to exploit potential vulnerabilities of a system

attribute: "characteristic or property of an entity that can be used to describe its state, appearance, or other aspect" (ISO/IEC 24760-1:2011)[128]

brute force attack: in cryptography, an attack that involves trying all possible combinations to find a match

coded: "1. identifying information (such as name or social security number) that would enable the investigator to readily ascertain the identity of the individual to whom the private information or specimens pertain has been replaced with a number, letter, symbol, or combination thereof (i.e., the code); and 2. a key to decipher the code exists, enabling linkage of the identifying information to the private information or specimens."[129]

control: "measure that is modifying risk. Note: controls include any process, policy, device,

[128] ISO/IEC 24760-1:2011, Information technology -- Security techniques -- A framework for identity management -- Part 1: Terminology and concepts

[129] *OHRP-Guidance on Research Involving Private Information or Biological Specimens*, Department of Health & Human Services, Office of Human Research Protections (OHRP), August 16, 2008. http://www.hhs.gov/ohrp/policy/cdebiol.html

practice, or other actions which modify risk." (ISO/IEC 27000:2014)

covered entity: under HIPAA, a health plan, a health care clearinghouse, or a health care provider that electronically transmits protected health information (HIPAA Privacy Rule)

data linking: "matching and combining data from multiple databases" (ISO/TS 25237:2008)

data subjects: "persons to whom data refer" (ISO/TS 25237:2008)

data use agreement: executed agreement between a data provider and a data recipient that specifies the terms under which the data can be used.

de-identification: "general term for any process of removing the association between a set of identifying data and the data subject" (ISO/TS 25237-2008)

de-identified information: "records that have had enough PII removed or obscured such that the remaining information does not identify an individual and there is no reasonable basis to believe that the information can be used to identify an individual" (SP800-122)

direct identifiers: see *direct identifying data*

direct identifying data: "data that directly identifies a single individual" (ISO/TS 25237:2008)

directly identifying variables: a category of data that contains direct identifiers

disclosure: "divulging of, or provision of access to, data" (ISO/TS 25237:2008)

distinguishable information: "information that can be used to identify an individual" (SP800-122)

effectiveness: "extent to which planned activities are realized and planned results achieved" (ISO/IEC 27000:2014)

entity: "item inside or outside an information and communication technology system, such as a person, an organization, a device, a subsystem, or a group of such items that has recognizably distinct existence" (ISO/IEC 24760-1:2011)

Family and Educational Records Privacy Act (FERPA): the primary law in the United States that governs the privacy of student educational records

Fair Information Practice Principles (FIPP): a set of principles that have been developed world-wide since the 1970s that provide guidance to organizations in the handling of personal data

genomic information: information based on an individual's genome, such as a sequence of DNA or the results of genetic testing

harm: "any adverse effects that would be experienced by an individual (i.e., that may be socially, physically, or financially damaging) or an organization if the confidentiality of PII were

breached" (SP800-122)

Health Insurance Portability and Accountability Act of 1996 (HIPAA): the primary law in the United States that governs the privacy of healthcare information

healthcare identifier: "identifier of a person for exclusive use by a healthcare system" (ISO/TS 25237:2008)

HIPAA: see *Health Insurance Portability and Accountability Act of 1996*

HIPAA Privacy Rule: "establishes national standards to protect individuals' medical records and other personal health information and applies to health plans, health care clearinghouses, and those health care providers that conduct certain health care transactions electronically" (HIPAA Privacy Rule, 45 CFR 160, 162, 164)

Health Information Technology for Economic and Clinical Health Act (HITECH Act): a 2009 law designed to stimulate the adoption of electronic health records (HER) in the United States

identifiable person: "one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity" (ISO/TS 25237:2008)

identification: "process of using claimed or observed attributes of an entity to single out the entity among other entities in a set of identities" (ISO/TS 25237:2008)

identified information: information that explicitly identifies an individual

identifier: "information used to claim an identity, before a potential corroboration by a corresponding authenticator" (ISO/TS 25237:2008)

indirect identifier: information that can be used to identify an individual through association with other information

inference: "refers to the ability to deduce the identity of a person associated with a set of data through "clues" contained in that information. This analysis permits determination of the individual's identity based on a combination of facts associated with that person even though specific identifiers have been removed, like name and social security number" (ASTM E1869[130])

k-anonymity: a technique "to release person-specific data such that the ability to link to other information using the quasi-identifier is limited."[131] k-anonymity achieves this through suppression of identifiers and output perturbation.

---

[130] ASTM E1869-04 (Reapproved 2014), Standard Guide for Confidentiality, Privacy, Access, and Data Security Principles for Health Information Including Electronic Health Records, ASTM International.

[131] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

l-diversity: a refinement to the k-anonymity approach which assures that groups of records specified by the same identifiers have sufficient diversity to prevent inferential disclosure[132]

likelihood: "chance of something happening" (ISO/IEC 27000:2014)

limited dataset: "A limited data set is protected health information that excludes the following direct identifiers of the individual or of relatives, employers, or household members of the individual: (i) Names; (ii) Postal address information, other than town or city, State, and zip code; (iii) Telephone numbers; (iv) Fax numbers; (v) Electronic mail addresses; (vi) Social security numbers; (vii) Medical record numbers; (viii) Health plan beneficiary numbers; (ix) Account numbers; (x) Certificate/license numbers; (xi) Vehicle identifiers and serial numbers, including license plate numbers; (xii) Device identifiers and serial numbers; (xiii) Web Universal Resource Locators (URLs); (xiv) Internet Protocol (IP) address numbers; (xv) Biometric identifiers, including finger and voice prints; and (xvi) Full face photographic images and any comparable images." (HIPAA Privacy Rule, 45 CFR 164.514 (e) (2)). Limited data sets can contain complete dates, age to the nearest hour, city, state, and complete ZIP code.

linkable information: "information about or related to an individual for which there is a possibility of logical association with other information about the individual" (SP800-122)

linked information: "information about or related to an individual that is logically associated with other information about the individual" (SP800-122)

masking: the process of systematically removing a field or replacing it with a value in a way that does not preserve the analytic utility of the value, such as replacing a phone number with asterisks or a randomly generated pseudonym[133]

non-deterministic noise: a random value that cannot be predicted

obscured data: "data that has been distorted by cryptographic or other means to hide information. It is also referred to as being masked or obfuscated." (SP800-122)

personal identifier: "information with the purpose of uniquely identifying a person within a given context" (ISO/TS 25237:2008)

personal data: "any information relating to an identified or identifiable natural person (*data subject*)" (ISO/TS 25237:2008)

personally identifiable information (PII): "Any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, social security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such

[132] Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. In Proc. 22nd Intnl. Conf. Data Engg. (ICDE), page 24, 2006.

[133] El Emam, Khaled and Luk Arbuckle, Anonymizing Health Data, O'Reilly, Cambridge, MA. 2013

as medical, educational, financial, and employment information."[134] (SP800-122)

potentially identifiable personal information (PI^2): A new category of information proposed by Robert Gellman for information that has been de-identified but can be potentially re-identified. "Potentially identifiable personal information is any personal information without overt identifiers."[135] Under Gellman's proposal, parties that wish to exchange attempts PI^2 could voluntarily subscribe to a regime that would provide for both criminal and civil penalties if the recipient of the data attempted to re-identify the data subjects.

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

protected health information (PHI): "individually identifiable health information: (1) Except as provided in paragraph (2) of this definition, that is: (i) Transmitted by electronic media; (ii) Maintained in electronic media; or (iii) Transmitted or maintained in any other form or medium. (2) *Protected health information* excludes individually identifiable health information in: (i) Education records covered by the Family Educational Rights and Privacy Act, as amended, 20 U.S.C. 1232g; (ii) Records described at 20 U.S.C. 1232g(a)(4)(B)(iv); and (iii) Employment records held by a covered entity in its role as employer." (HIPAA Privacy Rule, 45 CFR 160.103)

privacy preserving data mining (PPDM): "an [extension] of traditional data mining techniques to work with … data modified to mask sensitive information" [136]

privacy preserving data publishing (PPDP): "methods and tools for publishing data in a more hostile environment, so that the published data remains practically useful while individual privacy is preserved" [137]

process: "set of interrelated or interacting activities which transforms inputs into outputs" (ISO/IEC 27000:2014)

pseudonymization: a particular type of de-identification that both removes the association with a data subject and adds an association between a particular set of characteristics relating to the data subject and one or more pseudonyms.[138] Typically, pseudonymization is implemented by replacing direct identifiers with a pseudonym, such as a randomly generated value.

---

[134] GAO Report 08-536, Privacy: Alternatives Exist for Enhancing Protection of Personally Identifiable Information, May 2008, http://www.gao.gov/new.items/d08536.pdf

[135] Gellman, Robert; *The Deidentification Dilemma: A Legislative and Contractual Proposal*, 21 Fordham Intellectual Property, Media & Entertainment Law Journal 33 (2010), http://iplj.net/blog/wp-content/uploads/2013/09/Deidentificatin-Dilemma.pdf.

[136] Benjamin C. M. Fung, Ke Wang, Rui Chen and Philip S. Yu, Privacy-Preserving Data Publishing: A Survey on Recent Developments, Computing Surveys, June 2010.

[137] Ibid.

[138] Note: This definition is the same as the definition in ISO/TS 25237:2008, except that the word "anonymization" is replaced with the word "de-identification."

pseudonym: "personal identifier that is different from the normally used personal identifier." (ISO/TS 25237:2008)

quasi-identifier: *see indirect identifier*

recipient: "natural or legal person, public authority, agency or any other body to whom data are disclosed" (ISO/TS 25237:2008)

re-identification: general term for any process that re-establishes the relationship between identifying data and a data subject

re-identification risk: the risk that de-identified records can be re-identified. Re-identification risk is typically reported as the percentage of records in a dataset that can be re-identified.

risk: "effect of uncertainty on objectives. Note: risk is often expressed in terms of a combination of the consequences of an event (including changes in circumstances) and the associated likelihood of occurrence." (ISO/IEC 27000:2014)

synthetic data generation: a process in which seed data is used to create artificial data that has some of the statistical characteristics as the seed data

threat: "potential cause of an unwanted incident, which may result in harm to a system or organization" (ISO/IEC 27000:2014)

trusted data recipient: an entity that has limited access to the data that it receives as a result of being bound by some administrative control such as a law, regulation, or data use agreement

Appendix B Resources

B.1 Standards

- ASTM E1869-04(2014) Standard Guide for Confidentiality, Privacy, Access, and Data Security Principles for Health Information Including Electronic Health Records
- ISO/IEC 27000:2014 Information technology -- Security techniques -- Information security management systems -- Overview and vocabulary
- ISO/IEC 24760-1:2011 Information technology -- Security techniques -- A framework for identity management -- Part 1: Terminology and concepts
- ISO/TS 25237:2008(E) Health Informatics — Pseudonymization. ISO, Geneva, Switzerland. 2008. This ISO Technical Specification describes how privacy sensitive information can be de-identified using a "pseudonymization service" that replaces direct identifiers with pseudonyms. It is provides a set of terms and definitions that are considered authoritative for this document.

B.2 Official publications

AU:

- *Privacy business resource 4: De-identification of data and information*, Office of the Australian Information Commissioner, Australian Government, April 2014. http://www.oaic.gov.au/images/documents/privacy/privacy-resources/privacy-business-resources/privacy_business_resource_4.pdf

EU:

- *Opinion 05/2014 on Anonymisation Techniques,* Article 29 Data Protection Working Party, 0829/14/EN WP216, Adopted on 10 April 2014

UK:

- *Anonymisation: Managing data protection risk, Code of Practice 2012,* Information Commissioner's Office. https://ico.org.uk/media/for-organisations/documents/1061/anonymisation-code.pdf. 108 pages
- UK Anonymisation Network, http://ukanon.net/

US:

- McCallister, Erika, Tim Grance, and Karen Scarfone, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII),* Special Publication 800-122, National Institute of Standards and Technology, U.S. Department of Commerce. 2010.
- *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule,* U.S. Department of Health & Human Services, Office for Civil Rights, November 26, 2012. http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf
- *OHRP-Guidance on Research Involving Private Information or Biological Specimens*, Department of Health & Human Services, Office of Human Research Protections (OHRP), August 16, 2008. http://www.hhs.gov/ohrp/policy/cdebiol.html
- *Data De-identification: An Overview of Basic Terms*, Privacy Technical Assistance Center, U.S. Department of Education. May 2013. http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf
- *Statistical Policy Working Paper 22 (Second version, 2005),* Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, December 2005.

B.3 Law Review Articles and White Papers:

- Barth-Jones, Daniel C., The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy

Protections, Then and Now (June 4, 2012). Available at SSRN: http://ssrn.com/abstract=2076397 or http://dx.doi.org/10.2139/ssrn.2076397

- Cavoukian, Ann, and Emam, Khaled, *De-identification Protocols: Essential for Protecting Privacy*, Privacy by Design, June 25, 2014. https://www.privacybydesign.ca/content/uploads/2014/06/pbd-de-identifcation_essential.pdf
- Lagos, Yianni, and Jules Polonetsky, *Public vs. Nonpublic Data: The Benefits of Administrative Controls,* Stanford Law Review Online, 66:103, Sept. 3, 2013
- *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization* Ohm, Paul . UCLA Law Review, Vol. 57, p. 1701, 2010; U of Colorado Law Legal Studies Research Paper No. 9-12. Available at SSRN: http://ssrn.com/abstract=1450006
- *Defining Privacy and Utility in Data Sets,* Wu, Felix T. University of Colorado Law Review 84:1117 (2013).

B.4 Reports and Books:

- *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*, Committee on Strategies for Responsible Sharing of Clinical Trial Data, Board on Health Sciences Policy, Institute of Medicine of the National Academies, The National Academies Press, Washington, DC. 2015.
- P. Doyle and J. Lane, *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, North-Holland Publishing, Dec 31, 2001*
- George T. Duncan, Mark Elliot, Juan-José Salazar-Gonzalez, *Statistical Confidentiality: Principles and Practice,* Springer, 2011
- Emam, Khaled El and Luk Arbuckle, *Anonymizing Health Data,* O'Reilly, Cambridge, MA. 2013

B.5 Survey Articles

- Chris Clifton and Tamir Tassa, On Syntactic Anonymity and Differential Privacy, 2013 *Trans. Data Privacy* 6, 2 (August 2013), 161-183.
- Benjamin C. M. Fung, Ke Wang, Rui Chen and Philip S. Yu, Privacy-Preserving Data Publishing: A Survey on Recent Developments, Computing Surveys, June 2010.
- Ebaa Fayyoumi and B. John Oommen, A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases. 2010, *Software Practice and Experience.* 40, 12 (November 2010), 1161-1188. DOI=10.1002/spe.v40:12 http://dx.doi.org/10.1002/spe.v40:12