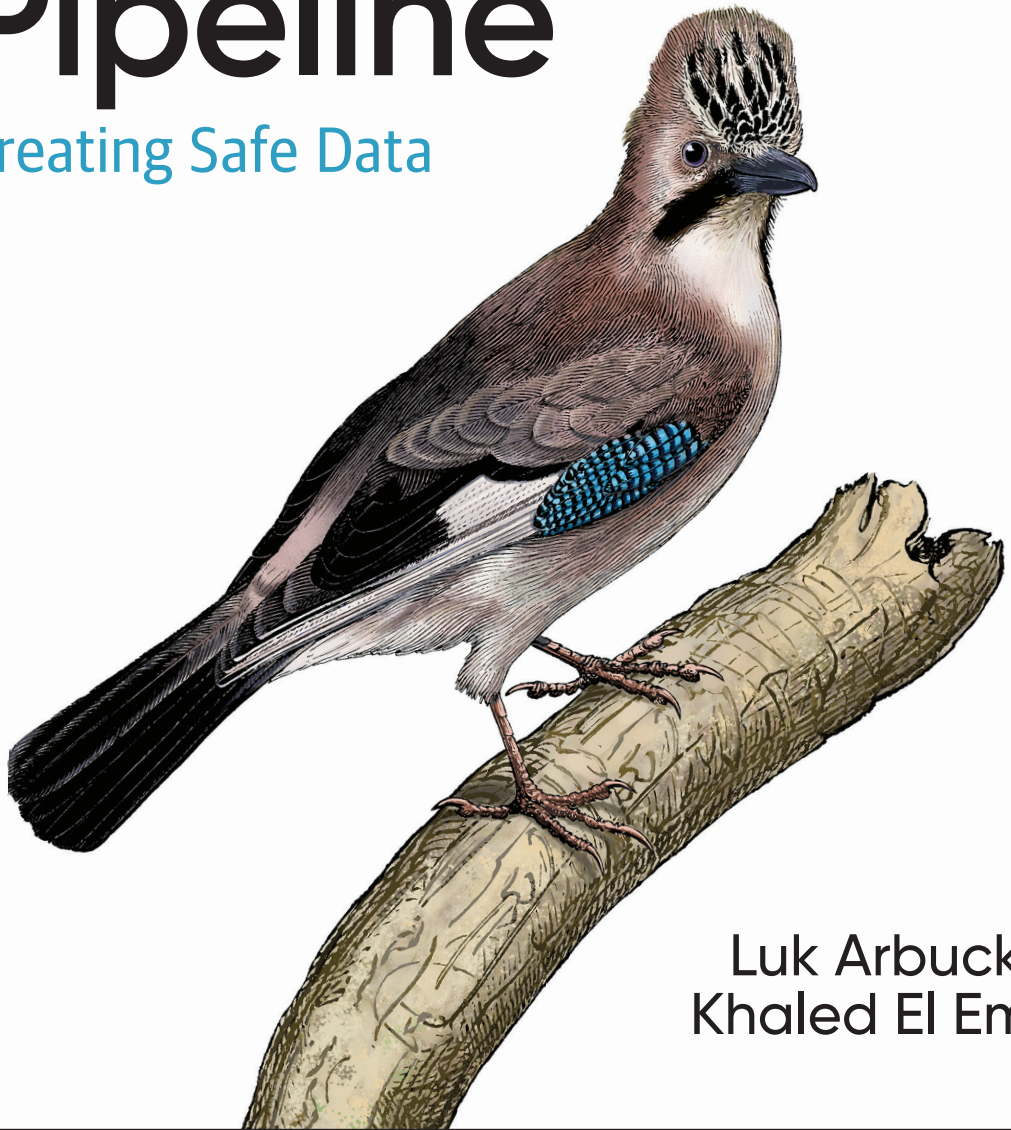
O'REILLY®

Building an Anonymization Pipeline

Creating Safe Data



Luk Arbuckle &
Khaled El Emam

O'REILLY®

Building an Anonymization Pipeline

How can you use data in a way that protects individual privacy but still provides useful and meaningful analytics? With this practical book, data architects and engineers will learn how to establish and integrate secure, repeatable anonymization processes into their data flows and analytics in a sustainable manner.

Luk Arbuckle and Khaled El Emam explore end-to-end solutions for anonymizing device and IoT data, based on collection models and use cases that address real business needs. These examples come from some of the most demanding data environments, such as healthcare, using approaches that have withstood the test of time.

- Create anonymization solutions diverse enough to cover a spectrum of use cases
- Match your solutions to the data you use, the people you share it with, and your analysis goals
- Build anonymization pipelines around various data collection models to cover different business needs
- Generate an anonymized version of original data or use an analytics platform to generate anonymized outputs
- Examine the ethical issues around the use of anonymized data

"This book provides excellent guidance and offers a solid foundation for designing and deploying a data strategy that leverages personal data to create and deliver innovation and value."

—**LeRoy A. Ruggerio**
Business Technology Leader for Hospital Medical Device R&D, Johnson & Johnson

Luk Arbuckle is chief methodologist at Privacy Analytics, where he provides strategic leadership in how to responsibly share and use data. Previously, he worked as director of technology analysis at the Office of the Privacy Commissioner of Canada.

Khaled El Emam is cofounder and CEO of Replica Analytics, a company developing data synthesis technology. He's also a professor in the Faculty of Medicine at the University of Ottawa and a senior investigator at the Children's Hospital of Eastern Ontario Research Institute.

DATA PROCESSING

US $49.99 CAN $65.99

ISBN: 978-1-492-05343-9

Twitter: @oreillymedia
facebook.com/oreilly

Building an Anonymization Pipeline

Creating Safe Data

Luk Arbuckle and Khaled El Emam

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Building an Anonymization Pipeline
by Luk Arbuckle and Khaled El Emam

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Jonathan Hassell
Development Editor: Melissa Potter
Production Editor: Christopher Faucher
Copyeditor: Sonia Saruba
Proofreader: Charles Roumeliotis

Indexer: Angela Howard
Interior Designer: David Futato
Cover Designer: Karen Montgomery
Illustrator: Rebecca Demarest

April 2020: First Edition

Revision History for the First Edition
2020-04-10: First Release

See *http://oreilly.com/catalog/errata.csp?isbn=9781492053439* for release details.

978-1-492-05343-9

[LSI]

Table of Contents

Preface

A few years ago we partnered with O'Reilly to write a book of case studies and methods for anonymizing health data, walking readers through practical methods to produce anonymized data sets in a variety of contexts.[1] Since that time, interest in anonymization, sometimes also called *de-identification*, has increased due to the growth and use of data, evolving and stricter privacy laws, and expectations of trust by privacy regulators, by private industry, and by citizens from whom data is being collected and processed.

Why We Wrote This Book

The sharing of data for the purposes of data analysis and research can have many benefits. At the same time, concerns and controversies about data ownership and data privacy elicit significant debate. O'Reilly's "Data Newsletter" on January 2, 2019, recognized that tools for secure and privacy-preserving analytics are a trend on the O'Reilly radar. Thus an idea was born: write a book that provides strategic opportunities to leverage the *spectrum of identifiability* to disassociate the personal from data in a variety of contexts to enhance privacy while providing useful data. The result is this book, in which we explore end-to-end solutions to reduce the identifiability of data. We draw on various data collection models and use cases that are enabled by real business needs, have been learned from working in some of the most demanding data environments, and are based on practical approaches that have stood the test of time.

1 Khaled El Emam and Luk Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*, (Sebastopol, CA: O'Reilly, 2014), *http://oreil.ly/anonymizing-health-data*.

The central question we are consistently asked is how to utilize data in a way that protects individual privacy, but still ensures the data is of sufficient granularity that analytics will be useful and meaningful. By incorporating anonymization methods to reduce identifiability, organizations can establish and integrate secure, repeatable anonymization processes into their data flows and analytics in a sustainable manner. We will describe different technologies that reduce identifiability by generalizing, suppressing, or randomizing data, to produce outputs of data or statistics. We will also describe how these technologies fit within the broader theme of "risk-based" methods to drive the degree of data transformations needed based on the context of data sharing.

The purpose of a risk-based approach is to replace an otherwise subjective gut check with a more guided decision-making approach that is scalable and proportionate, resulting in solutions that ensure data is useful while being sufficiently protected. Statistical estimators are used to provide objective support, with greater emphasis placed on empirical evidence to drive decision making.

We have a combined three decades of experience in data privacy, from academic research and authorship to training courses, seminars, and presentations, as well as leading highly skilled teams of researchers, data scientists, and practitioners. We've learned a great deal, and we continue to learn a great deal, about how to put privacy technology into practice. We want to share that knowledge to help drive best practice forward, demonstrating that it is possible to achieve the "win-win" of data privacy that has been championed by the likes of former privacy commissioner Dr. Ann Cavoukian in her highly influental concept of Privacy by Design.[2] There are many privacy advocates that believe that we can and should treat privacy as a societal good that is encouraged and enforced, and that there are practical ways we can achieve this while meeting the wants and needs of our modern society.

This is, however, a book of strategy, not a book of theory. Consider this book your advisor on how to plan for and use the full spectrum of anonymization tools and processes. The book will guide you in using data for purposes other than those originally intended, helping to ensure that data is not only richer but also that its use is legal and defensible. We will work through different scenarios based on three distinct classes of identifiability of the data involved, and provide details to understand some of the strategic considerations that organizations are struggling with.

2 Ann Cavoukian, "Privacy by Design: The 7 Foundational Principles," Information and Privacy Commissioner of Ontario (January 2011), *https://oreil.ly/eSQRA*.

Our aim is to help match privacy considerations to technical solutions. This book is generic, however, touching on a variety of topics relevant to anonymization. Legal interpretations are contextual, and we urge you to consult with your legal and privacy team! Materials presented in this book are for informational purposes only, and not for the purpose of providing legal advice. Okay, now that we've given our disclaimer, we can breathe easy.

Who This Book Was Written For

When conceptualizing this book, we divided the audience in two groups: those who need strategic support (our primary audience) and those who need to understand strategic decisions (our secondary audience). Whether in government or industry, it is a functional need to deliver on the promise of data. We assume that our audience is ready to do great things, beyond compliance with data privacy and data protection laws. And we assume that they are looking for data access models, to enable the safe and responsible use of data.

Primary audience (concerned with crafting a vision and ensuring the successful execution of that vision):

- Executive teams concerned with how to make the most of data, e.g., to improve efficiencies, derive new insights, and bring new products to market, all in an effort to make their services broader and better while enhancing the privacy of data subjects. They are more likely to skim this book to nail down their vision and how anonymization fits within it.
- Data architects and data engineers who need to match their problems to privacy solutions, thereby enabling secure and privacy-preserving analytics. They are more likely to home in on specific details and considerations to help support strategic decisions and figure out the specifics they need for their use cases.

Secondary audience (concerned with understanding the vision and how it will be executed):

- Data analysts and data scientists who want to understand decisions made regarding the access they have to data. As a detail-oriented group, they may have more questions than we can cover in one book! From our experience this may lead to interest in understanding privacy more broadly (certainly a good thing).
- Privacy professionals who wish to support the analytic function of an organization. They live and breathe privacy, and unless they have a technical background, they may actually want to dig into specific sections and considerations. That way

they can figure out how they can support use cases with their strong knowledge and understanding of privacy.

A core challenge with writing a book of strategy about the safe and responsible use of data is striking the right balance in terms of language and scope. This book will cover privacy, data science, and data processing. Although we attempt to introduce the reader to some basic concepts in all of these areas, we recognize that it may be challenging for some readers. We hope that the book will serve as an important reference, and encourage readers to learn more where they feel it is needed.

How This Book Is Organized

We'll provide a conceptual basis for understanding anonymization, starting with an understanding of *identifiability*, that is, providing a reasonable estimate of clustering based on identifying features in data and the likelihood of an attack. We will do this in two chapters, starting with the idea of an identifiability spectrum to understand identifiability in data in Chapter 2, and then a governance framework that explains the context of data sharing to understand threats in Chapter 3. Identifiability will be assessed in terms of both data and context, since they are intimately linked. Our identifiability spectrum will therefore evolve from the concept of data identifiability into one that encompasses both data and context.

From this conceptual basis of identifiability, we will then look at data processing steps to create different pipelines. We'll start with identified data and concepts from privacy engineering in Chapter 4, that is, how to design a system with privacy in mind, building in protections and, in particular, reducing identifiability for those novel uses of data that fall outside of the original purposes of data collection. We will also touch on the subject of having both identified and anonymized data within the same data holdings.

Once we've established the requirements related to identified data, we will consider another class of data for which direct identification has been removed, which we explained above as being pseudonymized. This is the first step to reducing identifiability, by removing names and addresses of the people in the data. In Chapter 5, we start to explicitly work toward anonymizing data. We first look at how pseudonymization fits as data protection, and introduce a first step toward anonymization. We also consider analytics technologies that can sit on top of pseudonymized data, and what that means in terms of anonymization.

Our final data pipeline is focused entirely on anonymization in Chapter 6 (so entirely about secondary uses of data). We start with the more traditional approach of pushing the anonymization at source to a recipient. But then we turn things around, considering the anonymized data as being pulled by the recipient. This way of thinking provides an interesting opportunity to leverage anonymization from a different set of

requirements, and opens up a way to build data lakes. We will do this by building on concepts introduced in other chapters, to come up with novel approaches to building a pipeline.

We finish the book in Chapter 7 with a discussion of the safe use of data, including the topics of accountability and ethics. The practical use of "deep learning" and related methods in artificial intelligence and machine learning (AIML) has introduced new concerns to the world of data privacy. Many frameworks and guiding principles have been suggested to manage these concerns, and we wish to summarize and provide some practical considerations in the context of building anonymization pipelines.

Conventions Used in This Book

The following typographical conventions are used in this book:

Italic
: Indicates new terms, URLs, email addresses, filenames, and file extensions.

`Constant width`
: Used for program listings, as well as within paragraphs to refer to program elements such as variable or function names, databases, data types, environment variables, statements, and keywords.

This element signifies a tip or suggestion.

This element signifies a general note.

This element indicates a warning or caution.

O'Reilly Online Learning

For more than 40 years, *O'Reilly Media* has provided technology and business training, knowledge, and insight to help companies succeed.

Our unique network of experts and innovators share their knowledge and expertise through books, articles, and our online learning platform. O'Reilly's online learning platform gives you on-demand access to live training courses, in-depth learning paths, interactive coding environments, and a vast collection of text and video from O'Reilly and 200+ other publishers. For more information, visit *http://oreilly.com*.

How to Contact Us

Please address comments and questions concerning this book to the publisher:

O'Reilly Media, Inc.
1005 Gravenstein Highway North
Sebastopol, CA 95472
800-998-9938 (in the United States or Canada)
707-829-0515 (international or local)
707-829-0104 (fax)

We have a web page for this book, where we list errata, examples, and any additional information. You can access this page at *https://oreil.ly/building-anonymization-pipeline*.

Email *bookquestions@oreilly.com* to comment or ask technical questions about this book.

For news and more information about our books and courses, see our website at *http://oreilly.com*.

Find us on Facebook: *http://facebook.com/oreilly*

Follow us on Twitter: *http://twitter.com/oreillymedia*

Watch us on YouTube: *http://www.youtube.com/oreillymedia*

Acknowledgments

This book would not be possible without the support of the many experts at Privacy Analytics, who work day in and day out in advisory, data and software delivery, and implementation. It's one thing to theorize solutions, it's quite another to work with organizations, large and small, to bring privacy practices and solutions to market and at scale. It's through working with clients that real-world solutions are born and grow up.

We must gush about our technical reviewers! They took the time to read the entirety of the first draft of this book and provided valuable feedback. Their varied backgrounds provided critical insights. Their feedback to the manuscript allowed us to directly address areas in need of further development. While the views and opinions expressed in this book are our own, we hope that we successfully incorporated their feedback into the final version of this book. In alphabetical order, we wish to thank: Bryan Cline, an expert in standards and risk management; Jordan Collins, an expert in real-world anonymization; Leroy Ruggerio, an expert in business technology; and Malcolm Townsend, an expert in data protection technology.

We would also like to thank Felix Ritchie for having created and promoted the adoption of the Five Safes, which served as inspiration to us! An entire chapter is dedicated to the Five Safes, and we have been fortunate to work with Felix since we drafted our first version of that chapter. We appreciated the help of Pierre Chetelat with final edits, which also served as an opportunity for him to learn about the legal and technical landscape in which we work.

Finally, we must thank O'Reilly for giving us the opportunity to write another book about anonymization in practice. And also Melissa Potter, our development editor at O'Reilly, who supported us in the writing and editing of this book. We may not have visibility behind the curtain at O'Reilly, but we also thank their team of diligent copy editors, graphic artists, technical support, and everyone else who brings books to market.

CHAPTER 1
Introduction

Data is recognized as an important driver of innovation in economic and research activities, and is used to improve services and derive new insights. Services are delivered more efficiently, at a lower cost, and with increased usability, based on an analysis of relevant data regarding how a service is provided and used. Insights improve outcomes in many facets of our lives, reducing the likelihood of fatal accidents (in travel, work, or leisure), getting us better returns from financial investments, or improving health-related outcomes by allowing us to understand disease progression and environmental influences, to name but a few examples. Sharing and using data responsibly is at the core of all these data-driven activities.

The focus of this book is on implementing and deploying solutions to reduce identifiability within a data pipeline, and it's therefore important to establish context around the technologies and data flows that will be used in production. Example applications include everything from structured data collection to Internet of Things (IoT) and device data (smart cities, telco, medical). In addition to the advantages and limitations of particular technologies, decision makers need to understand where these technologies apply within a deployed data pipeline so that they can best manage the *spectrum of identifiability*. Identifiability is more than just a black-and-white concept, as we will see when we explore a range of data transformations and disclosure contexts.

Before we delve into the concepts that will drive the selection of solutions and how they're deployed, we need to appreciate some concepts of privacy and data protection. These will help frame the scope of this book, and in particular the scope of reducing identifiability. While this is a book about anonymization, we divide the book up by different categories of identifiability that have been established by privacy and data protection laws and regulations. We will also demonstrate how to support proper anonymization through the concepts of these laws and regulations, and provide

examples of where things went wrong because proper anonymization was *not* employed. Anonymization should, in practice, involve more than just removing people's names from data.

Identifiability

Best practice recognizes that data falls on a spectrum of identifiability,[1] and that this spectrum can be leveraged to create various pipelines to anonymization. This spectrum is managed through technology-enabled processes, including security and privacy controls, but more specifically through data transformations and monitoring. We will explain how to objectively compare data sharing options for various data collection use cases to help the reader better understand how to match their problems to privacy solutions, thereby enabling secure and privacy-preserving analytics. There is a range of permutations in how to reduce identifiability, including where and when to provide useful data while meaningfully protecting privacy in light of broader benefits and needs.

While technology is an important enabler of anonymization, technology is not the end of the story. Accounting for risks in an anonymization process is critical to achieving the right level of data transformations and resulting data utility, which influences the analytic outcomes. Accordingly, to maintain usable outcomes, an organization must have efficient methods of measuring, monitoring, and assuring the controls associated with each disclosure context. Planning and documenting are also critical for any regulated area, as auditors and investigators need to review implementations to ensure the right balance is met when managing risks.

And, ultimately, anonymization can be a catalyst for responsibly using data, as it is privacy enhancing. There is a security component to responsibly using data that comes from limiting the ability to identify individuals, as well as an ethical component that comes from deriving insights that are broader than single individuals. Conceptually, we can think of this as using "statistics" (that is, numerical pieces of information) rather than single individuals, and using those statistics to leverage insights into broader populations and application areas to increase reach and impact. Let's discuss some of the other terms you'll need to know next.

1 For an excellent summary of the identifiability spectrum applied across a range of controls, see Kelsey Finch, "A Visual Guide to Practical De-Identification," *Future of Privacy Forum*, April 25, 2016, *https://oreil.ly/siE1D*.

Getting to Terms

Before we can dig in and describe anonymization in any more detail, there are some terms it would be best to introduce at the outset, for those not familiar with the privacy landscape. We will describe a variety of privacy considerations and data flows in this book based on potential data pipelines, and we will simply describe this as *data sharing*. Whether the data is released, as in a copy of the data is provided to another party, or access is granted to an external user of a repository or system internal to an organization, it's all sharing to us! Sometimes the term *disclosure* is also used for sharing data, and in a very broad sense. In an attempt to keep things simple, we will make no distinction between these terms.

We will use the terms *data custodian* to refer to the entity (meaning person or company) sharing data, and *data recipient* to refer to the entity receiving data. For internal data sharing scenarios, the data custodian is the organization as an entity, and the data recipient is a functional unit within that organization. The organization maintains oversight on the data sharing to the functional unit, and ensures that functional unit is treated as a separate unit so it can be assessed and treated as a legitimate data recipient. We will discuss this scenario in more detail later in the book.

In this book we have chosen to use the term identifiability, which pairs well with privacy laws and regulations that describe identifiable information, rather than speak of "re-identification risk." Although our measures are probabilistic, nonexperts sometimes find this approach to be both daunting and discouraging due to the focus on "risk." We hope that this change in language will set a more reasonable tone, and put the focus on more important apsects of building data pipelines that reduce identifiability and provide reasonable assurance that data is nonidentifiable.

We would struggle to describe anonymization, and privacy in general, without explaining that *personal data* is information about an identifiable individual. You may also come across the terms *personal information* (as it's referred to in Canada), *personally identifying information* (used in the US), or *protected health information* (identifiable health information defined for specific US health organizations). Personal data is probably the broadest of these terms (and due to EU privacy regulations, also of high impact globally), and since our focus is on *data* for analytics, we will use this term throughout this book. In legal documentation, the term used will depend on which law applies. For example, personally identifying information mixed with protected health information would simply be called protected health information.

When personal data is discussed, an identifiable individual is often referred to as a *data subject*. The data subject is not necesarily the "thing under study" (that is, the "unit of analysis," a term commonly used in scientific research to mean the person or thing under study). If data is collected about births, the thing under study may be the actual births, the infants, or the mothers. That is, the statistical analysis can focus on any one of these, and changing the thing under study can change how data is organized and how the statistical tools are used. For example, an analysis of mothers could be hierarchical, with infants at a different structural level. We will describe simple data structures with regard to statistical analysis in the next chapter.

For the purposes of this book, and most privacy laws and regulations, any individual represented in the data is considered a data subject. The thing under study could be households, where the adult guardians represent the individuals that are of primary interest to the study. Although the number of children a person has (as parent or guardian) is personal, children are also data subjects in their own right. That being said, laws and regulations vary, and there are exceptions. Information about professional activities may be confidential but not necessarily private. We will ignore these exceptions and instead focus on all individuals in the data as data subjects whose identity we endeavor to protect.

Laws and Regulations

Many of the terms that can help us understand anonymization are to be found in privacy laws and regulations.[2] Data protection, or privacy laws and regulations (which we will simply call laws and regulations, or privacy laws and regulations), and subsequent legal precedents, define what is meant by personal data. This isn't a book about law, and there are many laws and regulations to consider (including national, regional, sectorial, even cultural or tribal norms, depending on the country). However, there are two that are notable for our purposes, as they have influenced the field of anonymization in terms of how it is defined and its reach:

Health Insurance Portability and Accountability Act (HIPAA)
: Specific to US health data (and a subset at that),[3] HIPAA includes a Privacy Rule that provides the most descriptive definition of anonymization (called de-identification in the act). Known as Expert Determination, this approach requires someone familiar with generally accepted statistical or scientific principles and

2 Generally speaking, laws are written by a legislative assembly to codify rules, and regulations are written by admistrative agencies and departments to put these rules into practice. Both are enforceable.

3 HIPAA applies to health care providers, health care clearinghouses, and health plans (collectively known as *covered entities*), as well as their business associates. Health data that does not fall into these categories is not covered.

methods to anonymize data such that the risk of identifying an individual is "very small."[4]

General Data Protection Regulation (GDPR)
: This very comprehensive regulation of the European Union has had far-reaching effects, in part due to its extraterritorial scope (applying to residents of the EU, regardless of where that data is processed, when a service is intentionally targeting the EU), and in part due to the severity of the fines it introduced based on an organization's global revenue. The regulation is "risk based" (or contextual), with many references to risk analysis or assessments.[5]

As technology evolves, so do the emerging threats to anonymized data: more information may become publicly available, new techniques and methods become available to scrape and combine public information, and new methods emerge to launch attacks on data. Meanwhile, the technology that protects data, both cybersecurity and anonymization, will age and need updates and improvements. This means that the assessments of identifiability need periodic reviews and continous oversight to ensure the circumstances under which data was rendered nonidentifiable remain in place.[6] Similar to cybersecurity, typically the assessments need to be redone every 12 to 24 months, on top of continuous monitoring.

Since we've introduced US and EU privacy regulations, we should also clarify some of the terms used in each of these jurisdictions to refer to similar concepts. We're focusing on the two regulations mentioned above, although in truth there are also state-level privacy laws in the US (such as the California Consumer Protection Act, and LD 946 in Maine), as well as member-level privacy laws in the EU that add additional layers. For our purposes the terms in Table 1-1 should be sufficient. And, yes, you may notice that we've repeated the definition of personal data for the sake of completeness. The definitions are only basic interpretations in an attempt to bring the US and EU terms into alignment. This is only meant to provide some guidance on aligning the terms; be sure to discuss your particular situation with your legal and privacy team.

4 Details and expectations are provided by Office for 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," Department of Health and Human Services, 2015, *https://oreil.ly/OGxxa*.

5 To understand "risk based" in an EU context, see Article 29 Data Protection Working Party, "Statement on the Role of a Risk-Based Approach in Data Protection Legal Frameworks," October 4, 2017, *https://oreil.ly/A3Tpd*.

6 See, for example, the motivation described in Simson Garfinkel, "De-Identification of Personal Information," NISTIR-8053, National Institute of Standards and Technology, October 2015, *https://oreil.ly/ebsSD*.

Table 1-1. Basic definitions based on the similarities between US and EU terms

US HIPAA	EU GDPR	Common definition
Protected health information	Personal data	Information about an identifiable individual
De-identification	Anonymization	Process that removes the association between the identifying data and the data subject
Covered entity	Data controller	Entity that determines the purposes and means of processing of personal data
Business associate	Data processor	Entity that processes personal data on behalf of the data controller
Data recipient	Data processor (for personal data)	Entity to which data is disclosed by the data custodian
Limited data set	Pseudonymized data	Personal data that can no longer be attributed to a specific data subject without the use of additional information

Now that you're familiar with the terms used in these regulatory acts, let's get back to the subject of what makes personal data identifiable, and how we can interpret the term "identifiable" for the purpose of defining anonymization. Guidance from authorities is almost exclusively contextual and driven by risk assessments, attempting to balance the benefits of sharing data against an interpretation of anonymization that will sufficiently reduce identifiability to appropriately manage risks. We won't go through the various guidance documents available. Our previous work has helped influence guidance, and this book has been influenced by that guidance as well. We're all in this together! Let's consider various interpretations that have been put forward on what consitutes identifiable information, as shown in Table 1-2.

Table 1-2. Conditions on identifiability from various authorities (in alphabetical order)

Authority	Definition of identifiability
California Consumer Protection Act (US)	Directly or indirectly relates to or could reasonably be linked to a particular consumer or household
Federal Court (Canada)	Serious possibility that an individual could be identified through the use of that information, alone or in combination with other information
GDPR (EU)	Identifiability is defined by the "means reasonably likely to be used" to identify a data subject, taking into consideration objective factors (such as cost and time required to identify)
HIPAA (US)	Reasonable basis to believe identifiability, whereas not identifiable if an expert certifies that the risk of re-identification is "very small"
Illinois Supreme Court (US)	Not identifiable if it requires a highly skilled person to perform the re-identification
Office of the Privacy Commissioner of Canada	"Serious possibility" means something more than a frivolous chance and less than a balance of probabilities

As you can see from the table, authorities don't usually provide explicit measures of identifiability. It's more typical to find legal language than scientific norms in privacy

laws and regulations, even when these terms are less than clear.[7] Thankfully there is guidance and scientific norms available from experts that we draw on, some of which we will reference. We can at least divide identifiability into three well-known states.

States of Data

We mentioned the identifiability spectrum in the Preface, which is influenced by how authorities define personal data, as well as various sections in regulations and their interpretations and guidance. The identifiability spectrum is determined by accounting for:[8]

- The identity of the data recipient (so that we know who is accessing the shared data)
- Contractual controls (so that the data recipient knows their legal obligations)
- Privacy and security controls (so that limits are imposed on accessing the shared data, and on the data recipients themselves)
- Transformations of identifying information (which limit re-identifications even if the data recipients attempted to do so)

This book has been organized around a few points along the identifiability spectrum based on three main states of data: identified, pseudonymized, and anonymized. These are shown in Figure 1-1, and described in detail below.

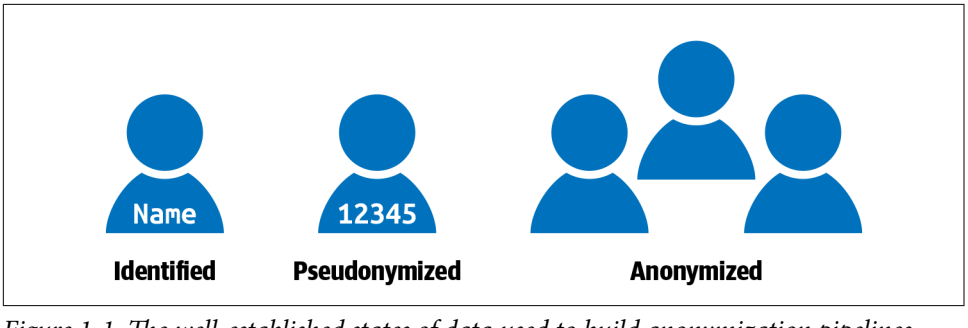


Figure 1-1. The well-established states of data used to build anonymization pipelines.

7 An interesting example of how some of these terms are interpreted is provided in Andrew Mauboussin and Michael J. Mauboussin, "If You Say Something Is *Likely,* How Likely Do People Think It Is?" *Harvard Business Review*, July 3, 2018, *https://oreil.ly/bdiIi*.

8 Many of these are described in Khaled El Emam et al., "Seven States of Data: When Is Pseudonymous Data Not Personal Information?" *Brussels Privacy Symposium: Policy and Practical Solutions for Anonymization and Pseudonymization*, 2016, *https://oreil.ly/Nn925*.

Identified

We use this term to mean that there is directly identifying information in the data, such as names or addresses. We make a slight distinction between *identified* and *identifiable*. An individual in a data set is identifiable if it is reasonable to expect that the individual *could be identified*, either with the data already immediately available or in combination with other information (external or known to the attacker). Many points along the spectrum will be considered identifiable, and therefore personal. But *identified* means the identity is known and associated with the data, which is often the case when delivering a service to an exact person. Identified data carries the most risk, and the most privacy and data protection obligations.

Pseudonymized

The term *pseudonymization* was popularized with the introduction of the GDPR. Technically speaking, when pseudonymizing, the directly identifying information doesn't need to be replaced with a pseudonym, it could just as well be a token or fake data or even suppressed entirely. The legal term pseudonymization simply means that direct identifiers have been removed in some way, as a data protection mechanism. Any additional information required to re-identify is kept separate and is subject to technical and administrative (or organizational) controls. This is how we will use the term pseudonymized in this book, while considering additional data transformations or controls that can reduce the legal obligations of working with personal data. Although pseudonymized data would therefore not be identi-*fied* data, it would still be identi-*fiable* data.

Anonymized

Anonymization is the process of removing direct and indirect identifiers for a given data sharing model, providing reasonable assurance that data is nonidentifiable. Anonymized data is therefore considered in the context of a data sharing scenario. Anonymization must be legally defensible—that is, it needs to meet the standards of current legal frameworks, and be presentable as evidence to governing bodies and regulatory authorities (i.e., data protection and privacy commissioners), to mitigate exposure and demonstrate that you have taken your responsibility toward data subjects seriously. Technically, "removing" indirect identifiers can mean various forms of generalization, suppression, or randomization, all of which will be determined by the relevant threats and prefered mitigation strategy to ensure data remains useful to the analytic needs.

The terms anonymization and de-identification are used interchangeably by some people, organizations, or even jurisdictions, but be careful as de-identification is sometimes used interchangeably with pseudonymization as well! Interpretations of each will also vary, in some cases in very substantial ways. We will consider a variety of general considerations throughout this book that should help explain most of these nuances, at least when definitions and guidance are considered more closely.

Other terms will be introduced throughout the book, but these are the ones that you need to start reading. Many of the terms we've just introduced necessarily include some discussion of regulations, so this section has served to introduce terms and regulations, at least to some degree. We will describe regulations where needed as you move through the book, in order to explain a concept or consideration. The next section will delve deeper into regulatory considerations as they relate to the process of anonymization.

Anonymization as Data Protection

There has been, and will continue to be, considerable debate around the term "anonymous," often focusing on the output of anonymization alone (i.e., can someone be reasonably identified in the data, and what's considered "reasonable").[9] We shouldn't lose sight of the fact that anonymization is a form of data protection, and thus is privacy enhancing. To be effective at enhancing privacy, anonymization needs to be used, and that means it also needs to be practical and produce useful data. Barriers that discourage or limit the use of anonymization technology will simply drive organizations to use identified data, or simply not innovate at all. There are many benefits that can be extracted from sharing and using data, so let's make sure that it's done responsibly.

We keep mentioning the need to produce "useful data" from the process of anonymization. There is a reality here that we can't escape, something we called the Goldilocks Principle in our previous book. The Goldilocks Principle is the idea that we need to balance risk against benefits, and in this case the benefits are the utility of the data and what insights may be drawn from it. It is possible to achieve a win-win situation by producing data that both serves a purpose and protects the identity of data subjects. But as data geeks, we have to be up front in saying that *there is no such thing as zero risk*. When we cross the road and look both ways, we are taking a measured risk. The risk we take when we cross the road can be quantified, but it's statistical in nature

9 For a good discussion of the debates about anonymization, and different viewpoints, we recommend everyone read Ira Rubinstein and Woodrow Hartzog, "Anonymization and Risk," *Washington Law Review* 91, no. 2 (2016): 703–60, *https://oreil.ly/Yrzj6*.

and never zero unless we never cross the road. Yet, we cross roads every day of our lives. We consider probable risks, and aim to achieve very low probabilities.

Consider the rock and hard place we are caught between. In a data sharing scenario in which we wish to achieve private data analysis, there will always be a sender (the data custodian) and a recipient (the data analyst). But the recipient is also deemed an eavesdropper or adversary (using standard security language, in this case referring to an entity or individual that may re-identify data, whether intentionally or not, thus adding risk to the process). Compare this with encryption, in which the recipient gets to decrypt and gain access to the original data shared by the sender. The recipient in the encryption example is not considered an adversary, because the intended recipient is *supposed to* decrypt the data. Not so with anonymization. The intended recipient of anonymized data should be unable to re-identify the data, and if they can, it's a problem.

Our goal in anonymizing data is to balance the needs of the recipient by providing them with useful data, while minimizing the ability of an adversary, including the recipient, to extract personal information from the data. The two roles that the recipient plays, as an eventual user of the data and also a potential adversary, is what distinguishes anonymization from encryption (in which the adversary and recipient are mutually exclusive), and what makes producing useful and safe data so challenging.

A more practical and realistic approach than striving for *zero* risk is to focus on the process of *minimizing* risk, considering anonymization as a *risk-management process*. This is the approach taken, for example, by the HITRUST Alliance, which provides a framework allowing organizations to meet the privacy requirements of multiple regulations and standards.[10] This is also the approach taken in data security, which is largely process based and contextual. We call this risk-based anonymization, which in our work has always included process- and harm-based assessments to provide a holistic approach to anonymization.[11] This approach informs the statistical estimators of identifiability and data transformations that are applied directly to data. Guidance on the topic of anonymization is almost always risk based, providing a scalable and proportionate approach to compliance.

10 "HITRUST De-Identification Framework," HITRUST Alliance, accessed March 28, 2020, *https://oreil.ly/wMxdF*.

11 Sophie Stalla-Bourdillon and Alison Knight, "Anonymous Data v. Personal Data—A False Debate: An EU Perspective on Anonymization, Pseudonymization, and Personal Data," *Wisconsin International Law Journal* 34, no. 2 (2017): 284-322, *https://oreil.ly/wctgn*.

If personal data is pseudonymized, or falls short of being considered anonymized, subsequent uses of the data must still be compatible with the original purpose for the data collection, and may require an additional legal basis for processing. Either way, pseudonymization reduces identifiability in data. We will therefore also consider methods to reduce identifiability that may fall short of anonymization, because they are both useful in their own right and are likely to build toward anonymization. We need to understand all the tools at our disposal.

We'll explore the idea of risk-based anonymization a little later in the chapter, but first we need to understand what data subject approval or consent involves and why laws or regulations don't typically require them for secondary uses of data.

Approval or Consent

As a form of data protection, anonymization itself does not normally require the approval of data subjects, although transparency is recommended and possibly required in some jurisdictions. As with other forms of data protection, anonymization is being done on behalf of data subject, to remove the association between them and the data. We use the term approval here rather than consent because under the GDPR, consent is more restrictive than in other jurisdictions (i.e., it must be "freely given, specific, informed, and unambiguous," with additional details and guidance around the interpretation).

Getting approval of data subjects can be extremely difficult and impractical. Imagine asking someone going to a hospital for treatment whether they would allow their data to be anonymized for other purposes. Is it even appropriate to be asking them when they are seeking care? Would some people feel pressured or coerced, or answer in a reactive way out of frustration or spite? It would be different in other scenarios, where the stakes aren't as high and the information not as harmful or sensitive. But timing and framing are important.

At the other extreme, approval to anonymize could be sought days, months, perhaps even years later. This could make for awkward situations when data subjects have moved on and acquaintances are asked for contact information. These acquaintances may not be on speaking terms with the data subjects or may be reluctant to share their contact information. Or the data subjects concerned may even be deceased. Contacting thousands of individuals for their approval is likely to be impractical, and unlikely to be fruitful.

But let's assume data subjects are reachable. Some privacy scholars have argued that approval can be meaningless, either because the approval request is presented in impenetrable legalese, or because data subjects don't understand the implications or simply don't want to be bothered. Depending on how the approval is structured, they

may give approval just to get access to something being offered, or elect not to be found and select the opt-out option. How is this preserving privacy?

In contrast, imagine a process in which approval is entirely voluntary and not required in exchange for a service. Government and the private sector would be forced to issue a potentially endless stream of requests to anonymize data for every use case and every service, hoping to improve operations or innovate. They would burden individuals with requests, to the point where individuals would simply ignore all requests. The concept of priming also suggests that even when cool heads prevail, people often only think about privacy when it's brought to their attention. They become sensitive to the topic because they are now thinking of it, and perhaps unnecessarily so. Opt-in would be rare, even when opting in would benefit the data subjects themselves or a broader population.

The reality is that specific sectors or use cases may see different rates of approval. Certain socioeconomic groups may be more sensitive to privacy concerns, and services and insights would become biased to specific groups. Making opt-in the default for anonymizing data, provided the process meets guidance or standards, would ensure nonpersonal data is available to improve services and derive new insights. This is why regulations offer alternatives to approval, and focus on much more than the process of reducing identifiability. Which leads us to a discussion of purpose specification, which is of critical importance to regulators.

Purpose Specification

Debate regarding anonymization usually arises when data is shared for purposes *other* than for which the data was originally collected, especially since approval by data subjects is not normally required once the data is anonymized. Although the process of anonymization is important, the uses of anonymized data are what concern people. There have been too many examples of data misuse, in which people felt discriminated against or harmed in some way, although interestingly most are probably using identified data. Anonymization will not solve data misuse, although it can help mitigate concerns.

Personal data may, for example, be collected from banking transactions, but that personal data is then anonymized and used to generate insights, e.g., to determine age groups that use a banking app versus an ATM, and at what times and on what weekdays. Such data-driven insights from nonpersonal data can improve services based on current usage patterns for different age groups. Some people may take issue with this form of targeting, even when the intent is to improve services by age group. All organizations have to make decisions to ensure the return on investment is reasonable, otherwise they will cease to exist, and this will inevitably mean making tradeoffs. However, if the targeting touches on sensitive demographic groups, it will enter

the realm of ethical considerations, even for anonymized data. This is especially true with sensitive data in general, such as health data.

If data is to be used for other purposes, for which approval of data subjects is not explicitly sought, the organization using the data should reflect carefully to ensure that its use of the data is appropriate. Specifically, harms should be considered in the broader context of ethical uses of data, which we'll discuss in more detail in later chapters. Although this may be deemed unrelated to anonymization, the reality is that it could set the tone for how a risk management approach to anonymization is evaluated. We consider framing anonymization within the broader context of data protection.

Reducing identifiability to a level in which it becomes nonpersonal is, by its very nature, technical, using a blend of statistics, computer science, and risk assessment. In order to engender trust, we must also look beyond the technical, and use best practice in privacy and data protection more broadly. Consider making the case for using anonymized data based on the purposes for which it will be used. For example, we can take a page from EU privacy regulations and consider "legitimate interests" as a way to frame anonymization as a tool to support the lawful and ethical reuse of data. That is, a data sharing scenario can consider how reusing the data (called "processing" in the regulatory language of GDPR) is legitimate, necessary, and balanced, so that it's found to be reasonable for the specified purposes.

Legitimate
: Data reuse should be something that is done now or in the very near future. The interests in reusing the data can be commercial, individual, or societal, but the reuse should avoid causing harm. It should also be possible to explain those interests clearly, and the reuse should seem reasonable in the hypothetical case explained to individuals.

Necessary
: Data reuse should be somewhat specific and targeted to the use case, and minimized to what is required to meet the objectives that have been laid out in advance. Overcollection will be frowned upon by the public, so it's best to ensure that needs are well laid out. Again, imagine the hypothetical case of trying to explain the reuse of all that data to individuals.

Balanced
: Data reuse should have well-articulated benefits that outweigh residual risks, or data protection or privacy obligations. Consider potential negative impacts and how they can be mitigated. A form of risk–benefit analysis can help inform and support the choice of mitigation strategies. Hint: reduce identifiability!

Anonymization can help address two of the three requirements listed above: it can more clearly limit the data to what is *necessary*, at least in terms of information that may be identifiable, and it can make the data more favorably *balanced* toward the beneficial side by reducing the risks of reusing the data. This leaves the legitimacy of reuse to be explained. Anonymization will help ensure that only necessary data is used and will help the benefits of reuse outweigh the potential harms. But how the anonymized data is used needs to be considered to ensure it is appropriate.

Now this isn't to say that we need to make the case for "legitimate interests" to use anonymized data, since being anonymized means that data protection laws and regulations no longer apply. What we are suggesting is that the privacy considerations above can help "legitimize" that use. We are simply drawing from some best practices to help frame the conversation and, ultimately, the reporting that takes place to explain anonymization.

Re-identification Attacks

To better understand the need for proper anonymization methods, let's consider a few well-known examples of re-identification attacks in which the anonymity of data subjects was compromised. There is a small set of such attacks that are repeated at conferences, in academic publications, and by the media, often in an attempt to raise awareness around the field of anonymization. As in any scientific discipline, these data points serve as evidence to inform and evolve the field (and where there isn't evidence, the field relies on scientific plausibility). They are what we call *demonstration attacks*, because they serve to demonstrate a potential vulnerability, although not its likelihood or impact. Demonstration attacks target the most "re-identifiable" individual to prove the possibility of re-identification. They are a risk in public data sharing, since there are no controls, and the attacker can gain notoriety for a successful attempt.

These well-known and publicized re-identification attacks were not attacks on what we consider to be anonymized data; the data would also not have been considered anonymized by experts in the field of *statistical disclosure control* (the field defined by decades of expert advice at national statistical organizations). Although the methods of statistical disclosure control have existed for decades, they were predominantly applied to national statistics and in government data sharing. Let's consider a handful of demonstration attacks and the lessons we can extract.

AOL search queries

In 2006, a team at AOL thought it would be of value to researchers in natural language processing (a field that develops algorithms in computer science to understand language) to share three months of web searches—around 20 million searches by 657,000 pseudonymous users. AOL made the data publicly available, and it can still

be found on the computers of researchers around the world and probably on peer-to-peer networks, even though AOL removed the search data from its site shortly after the release when a *New York Times* reporter published a story after having identified user 4417749.[12]

User 4417749's searches included "tea for good health," "numb fingers," "hand tremors," "dry mouth," "60 single men," "dog that urinates on everything," "landscapers in Lilburn, GA," and "homes sold in Shadow Lake subdivision Gwinnett County Georgia." Pay close attention to the last two searches. Geographic information narrows the population in a very obvious way, in this case allowing a reporter to visit the user's neighborhood and find a potential match. And this is how Thelma was found from the search queries.[13]

What's more, others claimed they were able to identify people in the search data. Many search queries contained identifying information in the form of names based on vanity searches (in which you search for yourself to see what's publicly available), or searches of friends and neighbors, place-names associated with a home or place of work, and other identifiers that could be used by pretty much anyone since the search data was public. And of course the searches also included sensitive personal information that people expected would be kept private. It's a good example of the risks associated with sharing pseudonymous data publicly.

Netflix Prize

Again in 2006, Netflix launched a data analytics competition to predict subscribers' movie ratings based on their past movie ratings. Better algorithms could, in theory, be used to provide Netflix users with targeted film recommendations so that users stay engaged and keep using the service. The competition was open to pretty much anyone, and by joining, participants would gain access to a training set of 100,480,507 ratings for 17,770 movies by 480,189 subscribers. Each rating in the training set included a pseudonym in place of the subscriber name, the movie name, the date of the rating, and the rating itself.[14]

A group of researchers demonstrated how they could match a few dozen ratings to the Internet Movie Database (IMDb), using a robust algorithm that would attempt to optimize the matches.[15] They were limited to a few dozen ratings due to a limit

12 Michael Barbaro and Tom Zeller Jr., "A Face Is Exposed for AOL Searcher No. 4417749," *The New York Times*, August 9, 2006, *https://oreil.ly/CnIBY*.

13 AOL search data leak.

14 Netflix Prize.

15 The results are described in Arvind Narayanan and Vitaly Shmatikov, "Robust De-Anonymization of Large Sparse Datasets," *Proceedings of the 2008 IEEE Symposium on Security and Privacy* (2008): 111-125, *https://oreil.ly/y0oec*.

imposed by the IMDb terms of service. They hypothesized that when Netflix users also rated movies on IMDb, the two sets of ratings would strongly agree with each other. The researchers claimed that subscribers in the Netflix dataset were unique based on a handful of ratings outside the top 500 movies and approximate rating dates (+/-1 week), and that they had found two especially strong candidates for re-identification. Based on the matching between the public IMDb movie ratings and the Netflix movie ratings, the researchers claimed to be able to infer political affiliation and religious views of these re-identification candidates by considering the non-public movies viewed and rated in the Netflix data.

Whether an adversary could know this level of detail, and confirm that their target was in the sample dataset, is debatable. However, given an appropriate database with names and overlapping information, the algorithm developed may be effective at matching datasets. It's hard to know from a demonstration attack alone if this is the case. However, in the case of mobility traces, in which geolocation points are connected to create a path, researchers found the matching algorithm to have a precision of about 20%, given their overlapping data from the same population, even though they had found that 75% of trajectories were unique from 5 data points.[16]

State Inpatient Database

Both the AOL and Netflix examples involved data sets in which pseudonyms had replaced usernames. Let's consider a different example, in which not only were names removed, but some information was also generalized, e.g., the user's date of birth was replaced by their age. For this we can turn to the Healthcare Cost and Utilization Project (HCUP), which shares databases for research and policy analysis. In 2013, the State Inpatient Database (SID) of Washington State from 2011 was subject to a demonstration attack using publicly available news reports. Privacy experts had warned that these databases required additional protection, and since the demonstration attack, multiple improvements have been introduced.

In this attack, a team searched news archives for stories about hospital encounters in Washington State. One included a 61-year-old man, Raymond, from Soap Lake, who was thrown from his motorcyle on a Saturday and hospitalized at Lincoln Hospital. Raymond was re-identified in the SID based on this publicly available information, and from this, all his other hospital encounters in the state that year were available since the database was longitudinal.

A total of 81 news reports from 2011 were collected from news archives, with the word "hospitalized" in them, and 35 patients were uniquely identified in the SID of

16 Details can be found in Huandong Wang et al., "De-Anonymization of Mobility Trajectories: Dissecting the Gaps Between Theory and Practice," *25th Annual Network and Distributed System Security Symposium* (2018), *https://oreil.ly/sq6NI*.

648,384 hospitalizations.[17] On the one hand, you could argue that 35 individuals out of 81 news reports is a significant risk, provided there's public reporting of the hospitalization; on the other hand, you could argue that 35 individuals out of 648,384 hospitalizations is a very small number for the benefits provided from sharing the data. Regardless, public sharing is challenging given the risks of a demonstration attack, whereas controls can dramatically prevent such incidents. More importantly, however, is what we learn about information that can potentially be used to identify an individual, and how this information can be used to properly measure identifiability.

Lessons learned

We need to distinguish between what is possible and what is probable, otherwise we would spend our lives crossing the street in fear of a plane being dropped on our heads (possible, but not probable). Demonstration attacks are important to understand what is possible, but they don't always scale or make sense outside of a very targeted attack on public data, where there are no controls on who has access and what they can do with the data. Our focus in this book is primarily on nonpublic data sharing, and how we can assess identifiability based on the context of that sharing.

Let's draw some lessons from these demonstration attacks.

- Pseudonymized data, in which names and other directly identifying information have been removed, are vulnerable (which is why they are considered personal data).
- Data shared publicly is at risk of demonstration attacks, the worst kind since it only takes one re-identification for attackers to claim success. Notoriety is an important motivator for attackers, leading them to publish their results.
- Contractual controls can discourage attempts at a demonstration attack (e.g., the IMDb terms of service), but will not be sufficient to eliminate all attacks. Additional controls and data transformations will be required.

With these lessons in mind, we can now make a distinction between re-identification attacks and what should constitute proper anonymization. None of the previous examples were of anonymized data in the sense that regulators use the term. We are now better positioned to discuss anonymization as it should be practiced.

17 Results are described in Latanya Sweeney, "Only You, Your Doctor, and Many Others May Know," *Technology Science*, September 29, 2015, *https://oreil.ly/0DTiH*.

Anonymization in Practice

Let's turn our attention to what we mean by the term *risk based*, since we've used this term a few times already. An evaluation of risk implicitly involves careful risk assessments, to understand more precisely where there is risk and what the impact of different mitigation strategies might be. This drives better decisions about how to prioritize and manage these risks. The process also means that risk is evaluated in an operational context, using repeatable and objective assessments to achieve our data sharing goals.

We take a very scientific approach to anonymization. Besides being evidence based, so that the approach is reasonable and adaptive to a changing threat landscape, we also determine a statistical tolerance using a threshold that is independent from how we measure identifiability, used to provide reasonable assurance that data is nonidentifiable. Based on risk assessments to evaluate the context of a data sharing scenario, we compare identifiability measures to the threshold to determine how much we need to transform identifying information until the statistical measure of identifiability meets the predefined threshold. We will describe this process in detail in the following chapters, but an overview of the process can be seen in Figure 1-2 (which can be iterative).

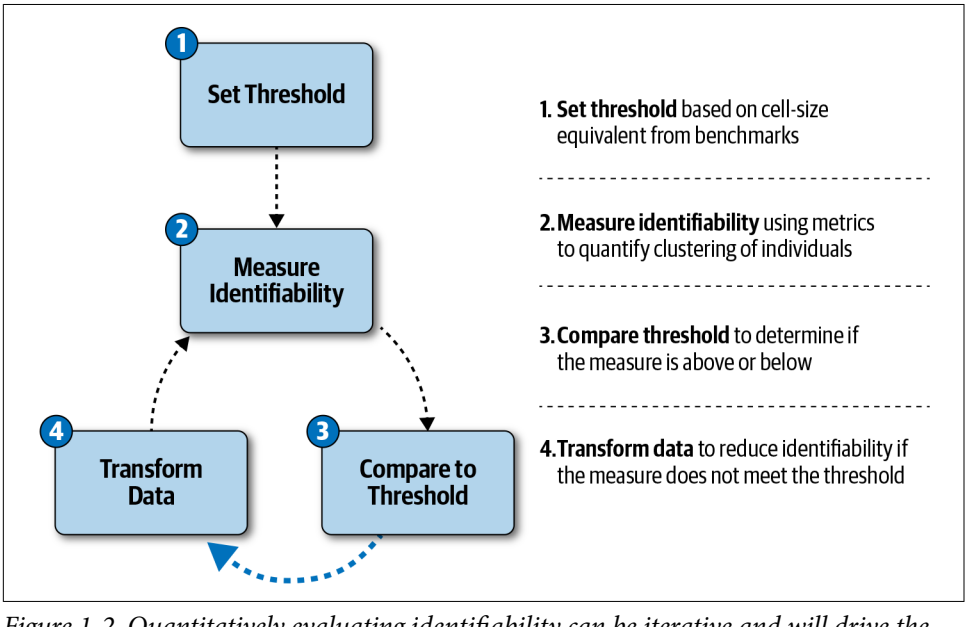


Figure 1-2. Quantitatively evaluating identifiability can be iterative and will drive the transformation of identifying data.

The threshold itself is a probability derived from benchmarks that represents the cell-size rules, which determine the minimum number of individual contributions that need to be included in an aggregation of data. Consider when the entries from identifiable categories are grouped (known as a contingency table or cross tabulation in statistics). A simple example would be age and county, where the cell-size rule could be 10 (or a probability threshold of 1/10). This would mean that there need to be at least 10 people in any one age and county in the identifying information. So for the tabulation of age of 30 and county of Shire, there would need to be at least 10 people that are aged 30 in the Shire. Our measures of identifiability are more complex than this, taking into consideration the context of data sharing and also the complexity of data, but this gives you a conceptual understanding of what we mean.

Contrast this benchmark approach with a fixed, list-based approach that invokes making required data transformations. HIPAA, mentioned earlier in the chapter, includes in its Privacy Rule a method known as Safe Harbor that uses a fixed list of 18 identifiers that need to be transformed.[18] This list includes many directly identifying pieces of information that need to be removed, such as name and Social Security number. Individual-level dates must be limited to year only, and the method also places limits on the accuracy of geographic information. Regardless of context, regardless of what data is being shared, the same approach is used.

The only saving grace to the HIPAA Safe Harbor approach is a "no actual knowledge" requirement that has been interpreted to be a catchall to verify that there are no obvious patterns in the data that could be used to identify someone, such as a rare disease. Although the Safe Harbor approach is simple, it does not provide very robust privacy protection and is only really useful for annual reporting. Also note that it's only suitable under HIPAA, as it was derived using US census information, and no other jurisdictions have provisions in their regulations to use this specific list.

Another approach to anonymization involves heuristics, which are rules of thumb derived from past experience, such as what transformations to apply based on specific data or circumstances, and fixed cell-size rules. These tend to be more complicated than simple lists, and have conditions and exceptions. Buyer beware. The devil is in the details, and it can be hard to justify heuristics without defensible evidence or metrics. Heuristics may provide a subjective gut check that things make sense, but this will be insufficient in the face of regulatory scrutiny.

The purpose of a risk-based approach is to replace an otherwise subjective gut check with a more guided decision-making approach that is scalable and proportionate, resulting in solutions that ensure that data is useful while being sufficiently protected. This is why we described risk-based anonymization as a risk management approach.

18 For details, see the de-identification guidance cited earlier by the Department of Health and Human Services.

And one of the most important ways you can reduce risk in a repeatable way is through automation, as shown in Figure 1-3.

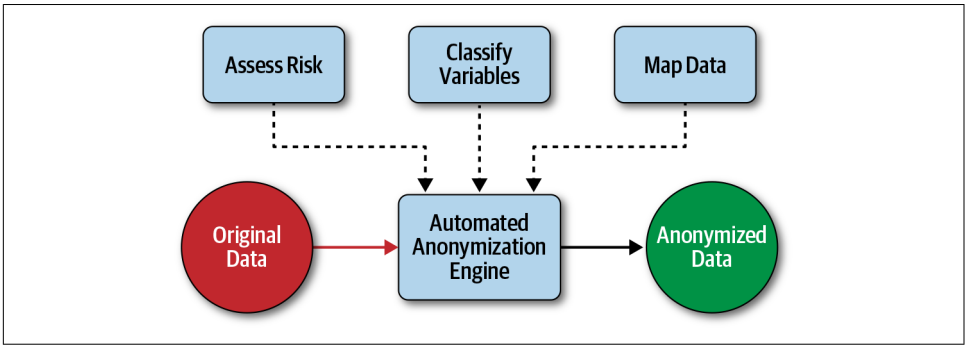


Figure 1-3. Automation means replacing a gut check with repeatable processes and auditable proof of what was done.

Creating automated risk management processes, in general, ensures that you capture all necessary information without missing anything, with auditable proof of what was done in case an issue arises that you need to correct for next time. This book will help you find areas for automation by introducing opportunities for technology-enabled processes in particular to reduce identifiability and build anonymization pipelines.

Final Thoughts

Recognizing that identifiability exists on a spectrum presents us with opportunities to meet privacy and data protection obligations in a multitude of use cases. Building an anonymization pipeline is not a linear path (pun intended!) from identified to anonymized. There are multiple points along this spectrum, as well as many criteria and constraints we need to consider to get to a solution that meets the needs of all parties and stakeholders involved. Whereas it is possible to have single-use anonymization of data to meet a specific need, this book takes a much broader view to consider how systems can be engineered with business and privacy needs in mind.

Whereas this isn't a book about privacy laws and regulations, we need to understand the basics as they relate to anonymizing personal data. Hopefully, the brief introduction in this chapter will inspire you to learn more.[19] But we will also highlight points of concern or confusion as they arise throughout the book. Importantly, we leverage three well-established states of data—identified, pseudonymized, and anonymized—to engineer anonymization pipelines.

19 An excellent resource to learn more about privacy and data protection, with newsletters, conferences, and courses, is the International Association of Privacy Professionals, *https://iapp.org*.

There are many concerns with the practice of anonymization, so it pays to remember that anonymization is privacy preserving, and understand what those concerns are so that we can address them. By being more clear about the purposes for using anonymized data, we are better positioned to ensure responsible sharing and use. Whereas there have been many reported re-identifications, such "demonstration" attacks only serve to underline the importance of using generally accepted statistical or scientific principles and methods to *properly* anonymize data. And this book will provide strategies to properly anonymize data for a range of use cases, once we've understood the identifiability spectrum in more depth, and a practical risk-management framework.

CHAPTER 2
Identifiability Spectrum

When identifiability is viewed as a spectrum, with one end signifying identified data and the other end signifying anonymized data, we find ourselves with a range of options for sharing and using data responsibly. It's therefore an opportunity that we can explore to develop a range of options depending on the use cases and data flows.

Before we dig into the details of the identifiability spectrum in this chapter, we will explore some of the questions and concerns around data sharing. From legal interpretation to practical considerations, we need to understand aspects of privacy to find creative ways to address them. And, if our objective is to understand anonymization, we need to understand how identifiability is considered from a statistical point of view. We will start by understanding how identifiability in data is estimated, before we explore risk assessments based on the context of the data sharing itself in Chapter 3. Let's start by talking about the legal landscape that informs any discussion of identifiability.

Legal Landscape

This isn't a book about privacy or data protection laws and regulations, but it's hard not to at least consider the legislative landscape when discussing anonymization. Legal interpretations change, as do the policies around them, so we're not going to spend time on this subject except to highlight a few important points as they relate to the material in this book. The concept of identifiability is important because it will determine the applicability of legislative requirements.

Most laws and regulations that aim to protect personal information define it, in one way or another, as information relating to an identifiable individual. A *reasonableness* argument is normally included to determine how information may be used, alone or in combination with other information, to determine identifiability. This provides

regulators with the necessary "wiggle room" to be technology neutral and adaptive to a changing landscape, where access to data may increase and technology will most certainly evolve. What's reasonable today may be less so tomorrow, and vice versa.

Reasonableness is judged based on what is practical and realistic at the time of processing (the operations performed on data), while excluding the highly unlikely, even impossible, or that which is prohibited by law. Ultimately, privacy laws and regulations support a reasonableness approach to identitifiability, including risk assessments to understand potential threats and vulnerabilities. At some point it becomes not only impractical, but also pointless, to re-identify if instead it would be easier to get the identified data in the first place. Anonymization introduces transaction costs that discourage attempts to re-identify.

Nonidentifiable information is therefore information that cannot be reasonably linked to an individual. As we'll see in this chapter, identifiability exists on a spectrum, and we need to define it in terms of the risk posed to data subjects in order to ensure that it can be effectively mitigated. Scholars have attempted to categorize identifiability subjectively, and this categorization is helpful for considering the range of options available to manage personal data. There are two extremes:

- Identified data will include directly identifiable information such as the data subject's name or another field of information that is unique to them, such as a Social Security number.
- Anonymized data will not include directly identifying information, and indirectly identifying information will be sufficiently transformed to ensure the remaining information is sufficiently disassociated from personal identities.

The spectrum of identifiability, and how we determine the degree of transformations and other mitigating factors necessary, are where we focus much of the discussion in this book. We'll demonstrate, in this chapter and the next, how we can minimize subjectivity, and provide objective support through identifiability measurement.

Disclosure Risk

The possibility that private or confidential information can be revealed is technically known as *disclosure risk*. Different types of variables can be used to reveal different types of information, but let's start with the two that are at the heart of anonymization. Direct identifiers are data that can essentially be used *alone* to uniquely identify individuals or their households; indirect (or quasi-) identifiers are data that can be used in *combination* with one another to identify individuals. Indirect identifiers are also referred to as quasi-identifiers in the computational disclosure control literature, and as key variables in the statistical disclosure control literature. They all mean more or less the same thing.

Legislative language includes both direct and indirect identifiers (*alone* or in *combination*) in describing personal data. Removing direct identifiers does not make data anonymized. The following figure in Figure 2-1 gives some examples of identifiers and how even indirect identifiers can be used to match records.

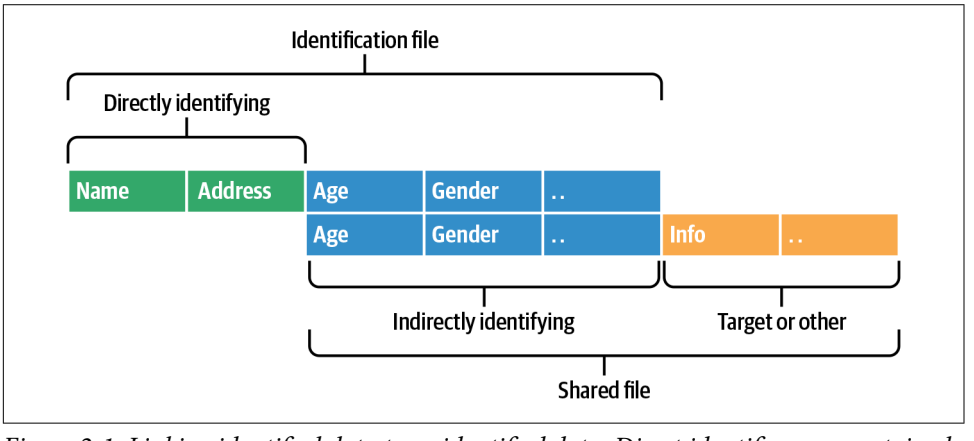


Figure 2-1. Linking identified data to unidentified data. Direct identifiers are contained in the identified data, whereas indirect (or quasi-) identifiers are contained in the unidentified data.

In discussing the spectrum of identifiability we'll focus our attention on indirectly identifying data. We'll assume any directly identifying data is properly masked (i.e., removed by field suppresion, pseudonymization, or random fake data). However, masked direct identifiers can play an important role in many anonymization pipelines by providing a means to link across data sources, so we'll revisit them in later chapters.

Types of Disclosure

Before we dive into the subject of identifiability, let's first explore three types of disclosure risk as defined in the field of statistical disclosure control.[1] Although we focus mostly on one type of disclosure risk in this book, it is important to understand disclosure risk in all its forms, since different risks require different methods and responses.

1 A classic book on the topic is George T. Duncan, Thomas B. Jabine, and Virginia A. De Wolf, *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics* (Washington, DC: National Academy Press, 1993), *https://oreil.ly/UHvV7*.

Identity disclosure
: Re-identification occurs when an individual's identity can be attached to some data, either alone or in combination with other information. For example, the name Its Me is correctly attached to a record of data about residents of Myplace.

Attribute disclosure
: Attribution occurs when sensitive information is associated with an individual or group in the data. This can occur with or without an identity disclosure. For example, the residents of Myplace *all* love chocolate.

Inferential disclosure
: Inference occurs when it's possible to learn something new about an individual or group in the data more accurately than would have otherwise been possible. For example, *some* of the residents of Myplace love cake, and so we infer all the residents probably love chocolate cake.

The difference between attribute and inferential disclosure is slight, but attribute disclosure is about making an association (no guesswork), not an inference. So if you want to avoid attribute disclosure, for the most part you can share a sample of the data or sample statistics, rather than all the data in a population. Or, if you only have a sample of data from a broader population, the most you could ever claim would be inference, and therefore attribution is no longer a concern. National statistical organizations are concerned with attribute disclosure because they collect data for entire populations, which is most often *not* the case for private industry.

Some have argued that attribute and inferential disclosure are one and the same, because it can be hard to draw a line between the two. That is, when does an inferential disclosure become so accurate that it's an attribute disclosure? It's a good point, but it can be nonetheless useful to draw a distinction between the two because some regulators and privacy advocates have argued that accurate attribution is tantamount to re-identification, and we don't want to confuse this with inference (where most organizations are focusing on data-driven insights and decision making).

Attribute disclosure is more generally considered through an impact assessment, in the context of confidentiality protection. Because attributions are made with certainty, or near certainty, they can reveal information deemed confidential. Attributions are obvious when everyone, or no one, in a group fall into a category. Everyone living at Myplace loves chocolate, or no one at Myplace loves vanilla. Substitute flavors with income levels and suddenly the impact becomes more obvious. An impact assessment would consider whether or not the benefits of sharing the information (and attributions) outweigh potential negative impacts.

Inferential disclosure is often only considered in the context of developing and analyzing disclosure control techniques. In general, if organizations equated a disclosure of personal information with inference, very little data would be shared. This is because the purpose of data analysis is to make inferences to derive insights or support evidence-based practice. If accurate inferences can't be made, there's no value in sharing data or deriving insights.

Our focus in this book will mostly be on the subject of identifiability, although we'll dip our toes into the waters of attribute and inferential disclosure as appropriate (such as a deep dive into the subject of ethics). Ultimately, though, privacy laws and regulations define personal data as being identifiable (e.g., GDPR and HIPAA).

Learning something new

Attribution or inference essentially means "learning something new" about data subjects. If re-identification takes place without learning something new, there is no actual privacy issue in practice. Imagine matching two sets of data that are identical, except one is identified and the other is not. Although technically this is an identity disclosure, the reality is that the adversary already had all the information and identities available to them, so they gain nothing from the re-identification. This is summarized in Figure 2-2.

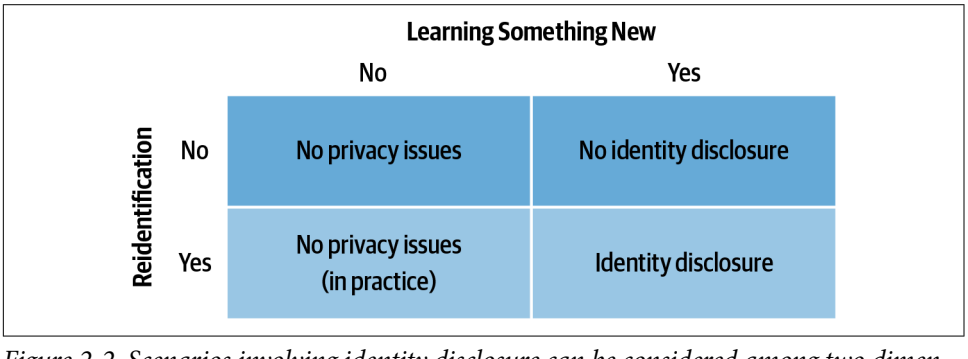


Figure 2-2. Scenarios involving identity disclosure can be considered among two dimensions: whether there's a re-identification versus whether anything new was learned from the disclosure.

Now that isn't to say that removing directly identifying information is sufficient to call data anonymous. We'll see shortly the elements that are needed to make that determination. But the point is a practical one, since identity disclosure without anything more is simply a nonissue. Ultimately, we almost always intend for data recipients to learn from the data, which is the purpose of data analysis. To help clarify what anonymization is, and what it isn't, we can consider different dimensions of data privacy.

Dimensions of Data Privacy

In discussions about data privacy and the secondary uses of data, a number of distinct concepts are oftentimes confused or treated as if they mean the same thing. There are four dimensions of data privacy that we're concerned with in this book: linkability, addressabilty, identifiability, and inference.[2] While we've discusssed identifiability and inference as they have to do with disclosure risks, we haven't discussed linkability, and addressability is a form of direct attribution (we will not discuss group attribution, because, as we've described, this is only relevant for full population data better dealt with using an impact assessment).

Sometimes all four concepts are lumped together under the "de-identification" or "anonymization" umbrella, when they actually refer to quite different things. We need to clearly divide these concepts to come up with policies and solutions that allow responsible uses of data and still achieve the societal and business benefits of data analysis.

We'll examine some of these concepts and demonstrate how they are independent of each other and therefore must be treated separately. These will be termed the dimensions of data privacy, inasmuch as it relates to anonymization. To be clear, these are not the only relevant dimensions of privacy. We want to focus on ones where more precise definitions and disentanglement would be beneficial.

To manage the four dimensions of data privacy, a different set of approaches is needed than with anonymization alone. A single approach is not going to ensure responsible data uses and disclosures, because it is likely to be a blunt hammer that misses opportunities to create useful data, or because it ignores important risks. Below are proposals reflecting some of the current thinking about managing these types of risks. Note that all four dimensions of privacy risks must be managed adequately to make a credible claim that privacy risks are managed.

Linkability

Linkability refers to the ability to link all of the events or records that belong to the same data subject together. For example, if all of the medical records that belong to the same patient have the same pseudonymized ID, it's possible to link all of them together and construct a longitudinal profile of that individual. The individual may or may not be identifiable. Identifiability is entirely independent of linkability. However, some experts have put linkability under the umbrella of anonymization because it makes anonymization more complex.

2 These dimensions were first presented in Khaled El Emam, "Setting the Record Straight on Privacy Dimensions in Big Data," *IAPP Privacy Tech* (blog), June 2, 2016, *https://oreil.ly/EVbMf*.

Treating linkability as part of anonymity prohibits, or at least discourages, transactional or longitudinal records from being protected through anonymization. This approach may have some justification in the context of open data or public data sharing, however, a blanket prohibition on anonymized longitudinal data would be extremely detrimental to a significant amount of research and analytics, such as health research, financial services, marketing, insurance, and education, to name a few.

Whereas linkability makes anonymization more complex, it is achievable in most cases. The disclosure of longitudinal data publicly (that is, open data) should be limited because it's difficult to protect that kind of detailed information within the constraints of a public release. But for nonpublic data, sharing this kind of data can be adequately anonymized. The threats are fewer when access is limited to trusted analysts with other controls in place, besides data transformations, to ensure uses of data are appropriate to the intended purposes.

Addressability

Addressability is when you have a pseudonym that can be used to target or "address" a specific individual (not necessarily an "identifiable individual"). For instance, a pseudonym can be used directly or indirectly to target advertisements to a specific individual or an individual's device. You may not know the identity of the individual, but you can address that individual anonymously. For example, an advertiser could send the pseudonym and the advertisement to an ISP that then links the pseudonym to a specific device ID and sends that advertisement to that device. The ISP already knows the identity of the consumer, and the advertiser never gets to know the identity of the consumer. In that case the pseudonym is addressable but not identifiable in the hands of the advertiser.

Alternatively, it may be possible to determine the identity of that individual, for example, if the advertiser is the ISP itself. Again, addressability is independent of identifiability. You may address someone electronically but have no capability to determine their identity, or you may be able to determine their identity and address them directly because you know their identity.

The key here is whether the pseudonym can be used to address or target a "specific individual" or an "identifiable individual." Unless there is express consent or at least some form of meaningful notice, mechanisms must ensure that addressability is performed anonymously. This can be achieved through controls on the workflows, some of which we will consider in the following chapters.

Identifiability

Identifiability is when we are able to correctly assign an event or record to an identifiable or known individual with a high probability. This is the traditional definition of identifiability, as we've already described. Anonymization standards that exist today would typically only address this specific issue of protecting against identity disclosure.

You can have a nonidentifiable, nonaddressable, and nonlinkable pseudonym, such as a random set of numbers assigned to a record (and no other record). Or a nonidentifiable, nonlinkable, but addressable pseudonym—when the pseudonym is regenerated repeatedly, but every instance of the pseudonym can be used to address an individual. In fact, there are eight possible combinations of scenarios on the three dimensions covered so far, since they're all independent of one another.

Good practices for anonymization exist, as well as standards and certification programs, and these should be followed. The purpose of this book is to describe these, and we reference other important sources you may wish to consult as the need arises.

Inference

Data analysis means building some kind of model. The model can be simple, such as being only descriptive—for example, "80% of our customers are female"—or more complex using statistical and machine learning methods to make inferences or predictions. A model can be built from identifiable data or anonymized data, from linkable data or nonlinkable data, and from addressable data or nonaddressable data. Therefore, model building is independent of the three dimensions above.

Once a model is built, it can be used for good purposes or undesirable purposes, such as discrimination against certain groups of individuals. For example, a model that predicts the likelihood of getting cancer by a certain age can be used to introduce wellness programs in high-risk communities or to deny bank loans to individuals deemed to have a high risk. Therefore, undesirable purposes are a function of the data uses rather than the inferences themselves. This means we need to introduce mechanisms to manage the risks from data uses.

In general, ethics reviews should be performed on data uses. We'll cover that in more detail in Chapter 7. Some have argued that laws or regulations should prohibit or limit certain types or classes of data uses. And questions arise regarding the extent that data uses should be compatible with the original intent of data collection, especially if the data is anonymized. Arguably, compatibility with the original collection would be substantially weaker for anonymized data than it would be for identifiable data, otherwise there would be no incentives for implementing privacy-protective mechanisms such as anonymization.

These privacy dimensions are not one and the same as identifiability, and therefore, separate considerations for building anonymization pipelines. With that in mind, we are now in a better position to understand the science of re-identification that is at the source of how we anonymize data. That is, we need to understand how data can be re-identified to understand how we will prevent re-identification.

Re-identification Science

The terms de-identification and anonymization are relatively recent terms largely introduced from the field of computer science. However, there's a long history of statistical disclosure control methods that consider the relationship between indirectly identifying information and external sources of indirectly identifying information to protect against re-identification.[3] It has long been recognized in this field that identifiability exists on a spectrum, and that there's no such thing as zero risk, because the field of statistics works in probabilities (and contrary to popular belief, a zero-probability event is almost sure not to occur but *not* impossible…cue scientist on the stand in a courtroom forced to admit that nothing is impossible).

We can think of re-identification as simply being able to associate a name to someone in data, often termed the data subject. There are many pieces we can unpack to go from sharing data to being able to successfully re-identify a data subject, as this will go to the heart of how we measure identifiability. Spoiler alert! It will require an assessment of the environment and the circumstances in which the data will be shared, and an assessment of the data itself. But we'll save that for the next chapter, when we introduce a governance framework to help think through the elements needed to assess risk.

Conceptually, we'll consider identifiability in the data in terms of how many people belong to a cluster of "similarly looking individuals," based on the indirectly identifying features that are in the data and available to an adversary.[4] These "similarly looking individuals" are sometimes called statistical twins. Take, for example, a dataset with Miss Orange, who identifies as female and, well, orange. We can consider every orange female in her population to determine the size of her identifiable cluster. Obviously this is an oversimplification, but this is how these ideas were first developed (for aggregated, or tabular, data in the field of statistical disclosure control).

3 For a recent treatment of the subject, see Thijs Benschop and Matthew Welch, "Statistical Disclosure Control for Microdata: A Theory Guide for sdcMicro," SDC Practice Guide (2019), *https://oreil.ly/sNBYl*.

4 A review of privacy metrics, related to the concept of identifiability, can be found in Isabel Wagner and David Eckhoff, "Technical Privacy Metrics: A Systematic Survey," *ACM Computing Surveys* (2018), *https://oreil.ly/7aCCt*.

Take the example in Figure 2-3. We have three females that can be clustered together. Since there are no names in this dataset, we have to randomly assign the name Miss Orange to one of the three data subjects. That implies a probability of one third for correctly attaching the name Miss Orange to her record (one Miss Orange, but three possible matches). Of course, an adversary will not know for certain if they have correctly matched the name Miss Orange to the data, but this is the starting point for theoretically developing models of identifiability. By extension, we can apply this concept to any identifiable cluster of individuals, and the probability of a correct match will be one over the size of the cluster. And this is also where we quantify the spectrum of identifiability, but we'll come back to that point later in this chapter.

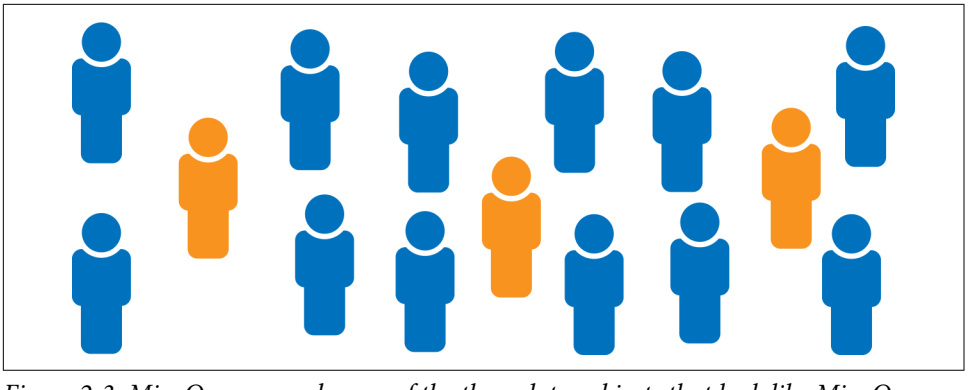


Figure 2-3. Miss Orange can be any of the three data subjects that look like Miss Orange.

Clustering is only the starting point for how we assess identifiability in data. We'll now look at some of the factors that affect identifiability.

Defined Population

We described identifiability in terms of cluster size, and in particular the cluster within a population, the emphasis being on the population of individuals that gave rise to the data itself. And this data may only be a sample or subset of that population. More formally, we can repurpose a definition of epidemiology to summarize what we mean by *re-identification science*.

Re-identification science

> The study and analysis of the distribution (who, when, and where) patterns, and determinants (or risk factors) of identifiability conditions in defined populations.

The population we're concerned with is the one that contributes to the adversary's ability to re-identify a data subject. If the data subjects are females from Europe who are over 60 years of age, our defined population is based on those three identifiable characteristics. Most data custodians will not, however, have access to data for an

entire population. But this is the population that will be defined for the purposes of measuring identifiability.

Let's revisit our example of a population with Miss Orange, but this time we consider a sample from that defined population, as shown in Figure 2-4. From the sample it would seem that there is only one possible match to Miss Orange, since there is only one orange female in the data. But that would be conflating the sample of data with the defined population. The point is that an adversary can't know who is in the sample versus who is in the population.

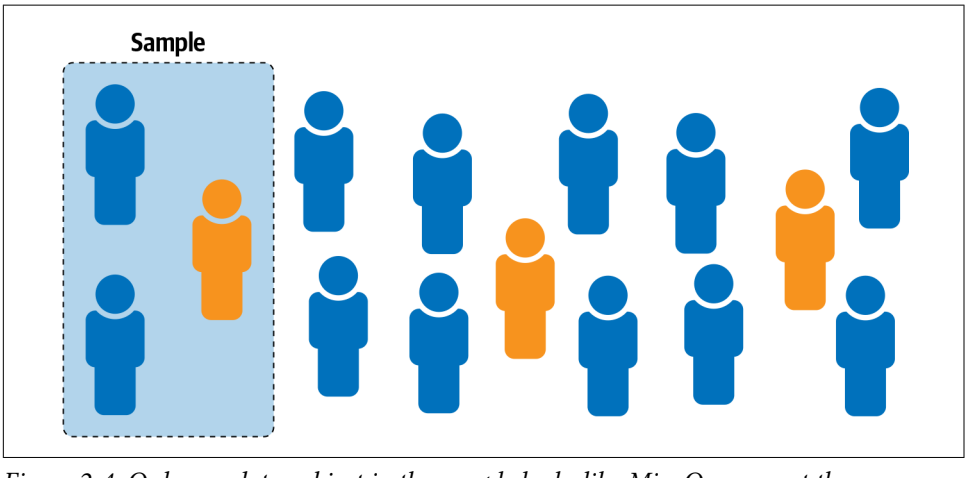


Figure 2-4. Only one data subject in the sample looks like Miss Orange, yet there are three possible matches in the population. Identifiability is evaluated at the population level.

The matching data to Miss Orange in the sample is called *sample unique*, because there is only one data subject in the sample matching Miss Orange. But we know that there are two other potential matches to Miss Orange in the population. Therefore, that record, although sample unique, is not unique in the population, called *population unique*.

This is a critical point that we need to emphasize. Sample uniques may just as well be considered a starting point in statistical estimation since they really don't, in and of themselves, represent a possible re-identification until they are considered at the population level. Population uniques represent a true vulnerability in the data. The chance of a vulnerability being found or exploited in the shared data is a separate consideration that we'll get to in the next chapter.

You can imagine that data analysts are often working with samples of data, and therefore identifiability needs to be estimated to the level of the defined population. Otherwise, identifiability will be overestimated, since all the sample uniques will be assigned an identifiability of 1, and every other cluster will also be underestimated since they will not factor in the estimated matches from the defined population. We want to provide the most useful data possible to data recipients, to ensure they can make effective use of it.

Now, if you have a sample of data from a larger population, but you don't have the data for the larger population, then you're going to need to estimate those clusters of data subjects. Estimators do, however, exist to infer the population cluster size from a sample of data. The field of statistical disclosure control has been working on these estimators for a long time, and continue to do so.[5] It can be done using only the data you have for the sample, or with help from statistics on the defined population you need to evaluate.

Demonstration Attack on Credit Card Metadata

Although we considered a few demonstration attacks in the introduction, there is another that is worth describing in the context of the defined population. Researchers were given access to a sample of financial transactions, including date, place, and amount paid. For example, a person buys a pair of shoes at their favorite store on one day, buys groceries at their local market on another day, and visits a particular coffee shop on another day, leaving a record of how much they paid for each transaction. Researchers used this to demonstrate the potential of re-identification by evaluating uniqueness, even when generalized to various levels of accuracy.

The researchers claimed that only four transactions were needed to make 90% of people unique in the data. They acknowledged that generalizing the data decreased how unique they were, but asserted that considering more transactions would make them unique again. However, the researchers evaluated uniqueness on the data that was shared with them, even though it was from a sample from a much larger population. This was not lost on experts that later commented on the work,[6] although this never gets the attention of the broader public.

Also worth mentioning, something which will be touched on at the end of this chapter and described in more detail in the next chapter, is the potential inclusion of contextual controls. Attacks need to be evaluated in the context in which data is used, considering how data is protected, and what external sources of identified data are

5 We will not delve into the mathematics of statistical disclosure control, but you can dig into it from the references we've already provided.

6 David Sánchez, Sergio Martinez, and Josep Domingo-Ferrer, "Comment on *Unique in the Shopping Mall: On the Reidentifiability of Credit Card Metadata*," *Science* 351, no. 6279 (2016): 1274, https://oreil.ly/Vm7Ou.

needed to match against. As we saw with the Netflix Prize example in the introduction, with regard to mobility traces, even matching on uniques is likely far from perfect when discrepancies between data are considered.

To summarize, the population is defined based on what the adversary can know. If the adversary knows that data subjects come from a specific region, then the region will define the population. But if the data subjects selected in the dataset are only females, this is also something the adversary can know, in which case the population is defined by females in that specific region, not just the specific region. Make sure that the population is well-defined, based on what the adversary can know.

Our defined population for the purposes of re-identification is therefore the cohort of data subjects from the geographic population that share common, knowable characteristics. These knowable characteristics are used to subset the geographic population, and are not included as indirectly identifying per se since they are not distinguishable within the data, and because they are already defined as the "denominator" for determining identifiability.

Now, with a sample of data to be shared, and a defined population in which to find individual candidates to match against, we need to consider *how* an adversary will attempt to re-identify. This will help us define threat models that capture certain nuances.

Direction of Matching

This may come as a surprise, as it's not immediately obvious, but the "direction" in which an adversary tries to match an identity to data will change the probability of the match being made. By direction of attack we mean the adversary can either start with the (unidentified) sample and try to match it to the correct individual in the population, or they can start with the (identified) population and try to match it to the correct record in the sample. We'll avoid going into the math (take a breath!), as you can find that information elsewhere.[7] But it's worth knowing because it leads to other considerations, such as the defined population we described previously, and it's required for a complete picture of identifiability. Again, we'll describe the direction of matching only conceptually.

As was just described, there are two directions in which an adversary can try to re-identify an individual in a data set: from sample to population, and from population

7 See Chapter 16 from Khaled El Emam, *Guide to the De-Identification of Personal Health Information* (Boca Raton, FL: CRC Press, 2013), *https://oreil.ly/xTMxg*.

to sample. We'll look at each in turn, recognizing that the math behind each is different, and that it will also provide us with a set of assumptions upon which to build our models of identifiability.

Sample to population (public)

A re-identification may be borne out by matching identifiable information from the shared sample back to the population, as seen in Figure 2-5. Let's assume that the adversary is a "motivated intruder," someone who is reasonably competent, but who lacks specialist skills and does not engage in criminal activity.[8] The adversary targets someone in the shared sample, and since the shared data is a random sample, they cannot be certain of who it is they have targeted. But, for whatever reason, the target data subject meets their selection criteria, and the adversary will attempt to match them to identified population data (an "identification database"), which requires the latter to be searchable in some way.

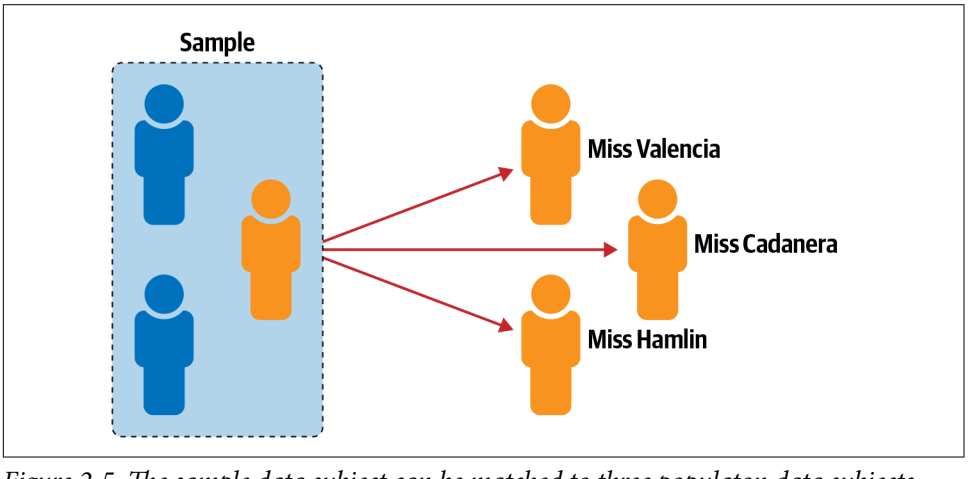


Figure 2-5. The sample data subject can be matched to three populaton data subjects. The re-identification is attempted from the sample to the population.

For example, the adversary could take a sample unique (easier) and look at who else lives in that disclosed geographic area. They only need identified population data for that geographic area, which could come from searching online or from other public sources to collect readily available information.

In matching from a sample to the population, it's therefore assumed that the adversary has background knowledge that may be available to the general public. This

8 A good example of such an exercise is Janice Branson et al., "Evaluating the Re-Identification Risk of a Clinical Study Report Anonymized Under EMA Policy 0070 and Health Canada Regulations," *Trials* 21, 200 (2020), *https://oreil.ly/zCYn5*.

information is readily available online and in public, such as demographic information, including age, gender, ethnic origin, where they live or work, languages spoken at home, income level, etc. The defined population should therefore be determined from knowable characteristics of individuals in the data based on what the public can know.

Population to sample (acquaintance)

Alternatively, a re-identification may be borne out by matching identifiable information from the population back to the shared sample, seen in Figure 2-6. In this case the adversary is targeting someone in the population, and there is no longer a need for a searchable identification database since the re-identification is on a single data subject in the population. This means that they can target someone they know, which is a very likely scenario in many re-identifications that are considered, and the identifiable information available to them will be more detailed.

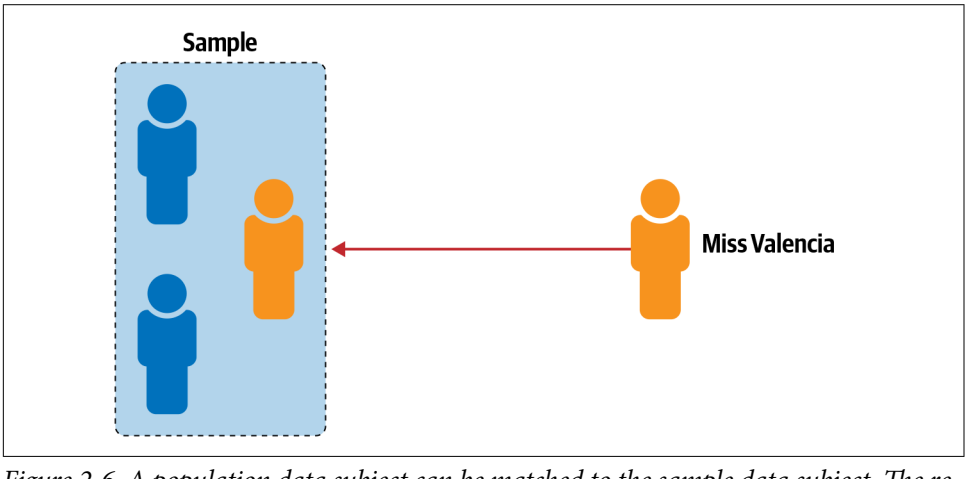


Figure 2-6. A population data subject can be matched to the sample data subject. The re-identification is attempted from the population to the sample.

In matching from a population to the shared sample, it's assumed that the adversary has the background knowledge of an acquaintance, since they are looking to match someone that they know. That is, the acquaintance may know all the information available to the general public, plus more detailed information, such as dates of events, medical diagnoses or procedures, spending habits, activities such as holidays and vacations, etc. The defined population should therefore be determined from knowable characteristics of individuals in the data based on what an acquaintance can know (we will describe acquaintances, and adversaries in general, in more detail in "Safe People" on page 49).

So, as you can see from the images, the direction of matching affects our measures of identifiability because the number of individuals considered is typically different. And we use this to build different models of attack. The adversary who is trying to re-identify an acquaintance in the data has more indirect identifiers to work with than someone who is trying to re-identify a member of the general public. This needs to be taken into account.

Having captured some of the complexities in how an adversary will attempt to re-identify data subjects, we turn our attention to the stucture of data itself so that we can define measures of identifiability. Changes in the structure of data will change our measures of identifiability.

Structure of Data

The structure of data will define specific relationships and associations between attributes of data, the *correlation structures*, that are important for analysis purposes. While we may not need to consider these exact correlation structures in how we assess identifiability, it's important to understand them so that we maintain them where they do not affect identifiability, and produce high degrees of data utility during the anonymization process.

Cross-sectional data

Here we consider multiple data subjects at a single point in time. This allows you to compare different population groups at a single point in time, or compare many different variables at the same time. Although the data is primarily comprised of attributes that are related but unordered, some attributes may be expressed as belonging to a group or transaction (transactional attributes) for the purposes of measuring identifiability.

Some examples of cross-sectional data include:

- Survey responses of data subjects taken at one point in time.
- Hospital visits without a linking variable to data subjects, so that the hospital visits become independent of one another, and the visits themselves are the thing that is analyzed. Diagnoses can be represented in a single column (as a transactional attribute), or by a fixed number of columns (e.g., an indicator variable for each possible diagnosis).

Cross-sectional data may have more indirect identifiers than we deem reasonable from the perspective of identifiability. The amount of information the adversary has is known as the power of the adversary. Adversary power can, for example, be used on non-public indirect identifiers, such as survey responses or when there are primary and secondary fields of information. These can be described as "feature sets." Therefore, we can have cross-sectional data with a feature set in which we'll apply adversary power to capture a reasonable measure of identifiability.

Time-series data

Here we consider a single data subject at multiple points in time. The repeated measures in a time-series analysis are typically in the order of at least 50+ observations, or if there are seasonal effects, in the order of 2 to 6 times the seasonality (the variations that occur at regular intervals, such as weekly or monthly). Suffice it to say there are a sufficient number of repeated measures to support accurate prediction of trends or outomes for the single data subject. Time-series data is represented by transactional attributes.

Here is an example of time-series data:

- Perceptual speed scores of an Alzheimer's patient measure how quickly the patient can compare figures or perform simple tasks related to visual perception. In order to make sense of this data, it's important to maintain the time variable and the link to the repeated measures, as this defines the (serial) correlation structure that will be used in analysis (e.g., modeling time-varying processes).

Time-series data may show trends or cyclical patterns that could be identifiable. The anonymization of time-series data is mostly new in terms of practical applications and still being developed at this time, since there hasn't been much practical demand. However, interest may increase with the introduction of the Internet of Things and device data.

Longitudinal or panel data

Longitudinal or panel data structures involve multiple subjects at multiple points in time, also described as repeated measures over time. This allows you to detect developments or changes in the characteristics of a defined population at both the group and individual level, and is more likely than cross-sectional data to suggest cause-and-effect relationships due to its scope.

Examples of longitudinal data include:

- Survey data taken at multiple points in time (e.g., at each visit, and linked to data subjects)
- Hospital visits with a linking variable to data subjects, so that each data subject has their visits linked to them.

It's important to maintain the time variable and the link to the repeated measures, as this defines the correlation structure that will be used in analysis (e.g., mixed effects models).

Longitudinal data may have repeated measures on the same day or over multiple days. We can treat the repeated measures that are deemed identifiable and the associated time variable as independent feature sets (transactional attributes), and apply the concept of adversary power (otherwise nothing new may be learned from a re-identification, eliminating most incentives to attempt a re-identification). Feature sets that are highly correlated may also be grouped so that only the most identifying feature (i.e., informative) drives statistical estimation.

Multilevel or hierarchical data

All of the above data structures may be clustered, nested, or hierarchical in nature. They may even be combined in such a fashion.

Here is an example of multilevel data:

- Survey responses of data subjects taken at each visit to a facility, for multiple facilities. A multilevel model would be repeated responses nested within respondents, with cross-sectional measures of the respondents, and the respondents nested within the facilities, with cross-sectional measures of the facilities (a three-level model of responses $\subset$ respondents $\subset$ facilities).

It may be important to maintain the multilevel nature of a dataset, at least when possible from the perspective of identifiability, since it will represent important relationships and associations. The same considerations described for previous data structures apply.

This is as far as we go in terms of describing some of the complexities with measuring identifiability from the perspective of the data alone and how it is structured. We've already provided references to resources that go into the mathematics and statistics required to build and choose estimators. However, before we close off this chapter, we do want to explain, even briefly, how an overall measure of identifiability is achieved.

Overall Identifiability

In the next chapter we'll provide a framework that highlights the elements needed to fully appreciate and assess identifiability in the context of a particular data sharing scenario. In a nutshell, we need to consider the overall level of identifiability, whereas what we described in terms of identifiability was limited to the data only. But the reasonableness described by privacy legislation includes other factors besides the data, and the guidance by regulators and data protection authorities is inherently contextual and driven by risk assessments.

As we stated earlier, in order to determine the overall level of identifiability, we require an assessment of the environment and the circumstances in which the data will be shared (context), and statistical measures from the data itself. This will capture the reasonableness aspects found in privacy legislation that was described at the start of this chapter. The overall determination of identifiability will need to be detailed, evidence based, and repeatable to provide the necessary evidence to governing bodies and regulatory authorities that there is a reasonable assurance the information is nonidentifiable, using statistical measures of identifiability to reduce subjectivity and guide decision making.

The overall level of identifiability is a function of both the data and context of sharing that data, as summarized in Figure 2-7. When expressed as probabilities, the overall measure of identifiability is the multiplication of these two factors. We'll determine what constitutes an acceptable overall estimate of identifiability, the threshold, based on benchmarks representing an assessment of factors from past data shared by reputable organizations, used to provide reasonable assurance that data is nonidentifiable.

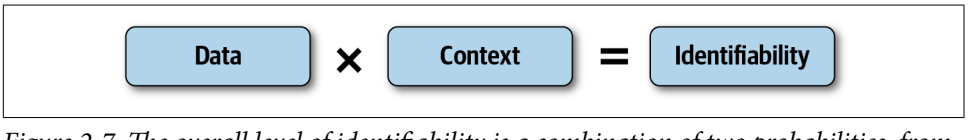


Figure 2-7. The overall level of identifiability is a combination of two probabilities, from the data and the context of sharing that data.

The overall determination of identifiability should also factor in the ability of the adversary to verify that the claimed re-identifications are accurate, or their confidence in the re-identifications they have made. It would be neither practical nor useful to treat the random assignment of names to data by the expected number of re-identifications alone, unless there's a way for the adversary to know they have correctly matched identities. This doesn't change our underlying statistical models or even our thresholds, but it does change our perception of identifiability by providing more realistic statistical estimation.

Different contexts will therefore have different levels of influence on identifiability for the very same data. We can leverage these changes in context to create anonymization pipelines in which data has different levels of identifiability from the point of collection to the point of use. This will be an important feature of this book, as we divide both context and data throughout a data pipeline, depending on the use cases and data flows.

Final Thoughts

Privacy laws and regulations apply to personal information, which is information about an identifiable individual. How identifiability is considered from a legal perspective involves a reasonableness argument regarding how likely it is that identities can be correctly linked to data. A scientific approach to evaluating identifiability involves statistical models that ultimately measure the distance between identified data and the anonymized data to be shared with data recipients.

While there are other types of disclosure risk, our focus is primarily on identity disclosures. There is a risk that some other privacy concepts, in particular those that make identifiability more complex to evaluate, are mixed with identifiability for the sake of making assessments simpler. Linkability, addressability, and inference are important in their own right, but independent of identifiability. It's important that we seperate these to be accurate in our assessments of privacy impacts.

Conceptually, the spectrum of identifiability is based on the size of clusters of similarly looking individuals. The bigger the cluster, the less identifiable the individuals in the cluster—although this implies the data is less granular, which can impact the usefulness of the data. The models that measure the size of clusters in data depend on the structure of the data, which can make those models more complex. However, reasonable assumptions can be made to ensure those measures are not overly restrictive, and useful data is available to derive meaningful insights.

We will consider the spectrum of identifiability as the range of options that can be applied in practice. We'll see examples throughout this book describing how to apply these options to develop data pipelines with various degrees of data transformations, depending on circumstances and need.

CHAPTER 3
A Practical Risk-Management Framework

While technology is an important enabler of data anonymization, technology is not the end of the story. Building an effective anonymization pipeline at an enterprise level is as much about governance as it is about technology, as we aim to deliver trust to stakeholders.[1] Accounting for risk in as an anonymization technology is critical to achieving the right level of anonymization and resulting data utility, which influences the analytic outcomes.

To maximize outcomes, an organization must have efficient methods of measuring, monitoring, and assuring the controls associated with each disclosure context. More broadly, organizations should establish a framework to manage identifiability holistically while enabling a wide range of data uses.

If you only apply technology to anonymize data, you miss out on a vital area of the overall strategy—the people and decisions behind the solution, and the processes and procedures that instill consistency. Without these elements, you miss the tenets of governance—accountability, transparency, and applicability. And you end up with less useful data.

The techniques used to achieve anonymization cannot be separated from the context in which data is shared: the exact data you're working with, the people you're sharing it with, and the goals of subsequent analysis. This is called risk-based anonymization. A framework has emerged from statistical data sharing by government agencies that is predominantly a risk-based approach—the Five Safes. We'll demonstrate how it can be operationalized in a broader setting.

1 For a good book on the basics of data governance, see Morgan Templar, *Get Governed: Building World Class Data Governance Programs* (Rescue, CA: Ivory Lady Publishing, 2017), *https://oreil.ly/PKD0h*.

Five Safes of Anonymization

Responsible data sharing requires an assessment of many factors, all of which need to be considered objectively to compare data sharing options. Only then can data custodians determine the most appropriate option for their particular circumstances, given the risks and benefits of sharing data in the first place.

One framework that has gained popularity after more than a decade of use is known as the Five Safes,[2] which is intended to capture the relevant dimensions to assess the context and results of a data sharing scenario in an effort to make sound decisions. Those dimensions are: Safe Projects, Safe People, Safe Settings, Safe Data, and Safe Outputs. The term "safe" is treated on a spectrum, as in "how safe" is it, so that this balancing can take place.

The entire premise of the Five Safes is based on the idea of risk assessment, which may be seen as subjective but with objective support through statistical estimation. Greater emphasis is then placed on empirical evidence to drive decision making.

Let's compare the Five Safes with risk-based anonymization, which requires an evaluation of the external information available to an adversary (whether a re-identification is intentional or not), and how the adversary may combine it to re-identify data. Removing personal information from data using a risk-based methodology requires an assessment of the environment and the circumstances in which the data will be shared (to know what external information will be available to an adversary), and an assessment of the data itself (to determine how the external information available to an adversary may be used to re-identify data).

With that in mind, the Five Safes can be desribed using concepts from risk-based anonymization, as shown in Figure 3-1. Our goal with this framework is the *safe use* of data, while maintaining as much granularity as possible. This is why the framework starts with defining project boundaries, and then focuses on people and settings. That way, residual risk is managed by anonymization, i.e., transforming the data to meet project needs and maintaining the highest level of data utility we can. But what we do with that data (i.e., the outputs) will still pose some concerns, which we consider last.

2 A summary of the Five Safes, by the creator himself, is given in Felix Ritchie, "The 'Five Safes': A Framework for Planning, Designing and Evaluating Data Access Solutions," paper presented at Data For Policy 2017, London (September 2017), *https://oreil.ly/7he01*.

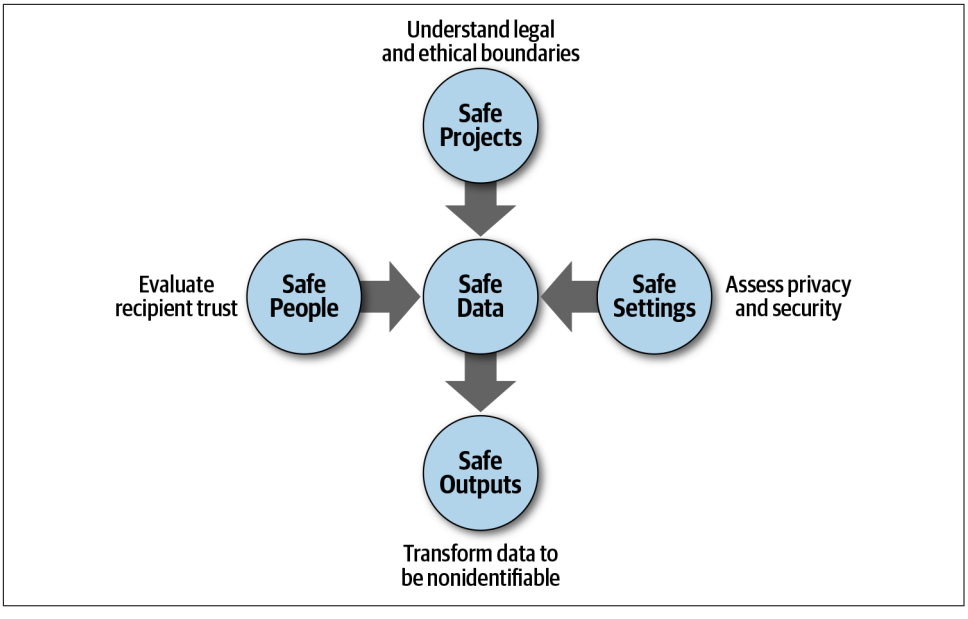


Figure 3-1. Overall risk exposure using the Five Safes, operationalized through risk-based anonymization.

We can summarize the steps of the Five Safes, operationalized through risk-based anonymization, in greater detail as follows:[3]

Safe Projects
: What are the legal and ethical boundaries of a data sharing scenario, and is anonymization needed as a privacy-protective measure?

Safe People
: Who are the anticipated data recipients, what are their motivations and capacity to re-identify, and who may they know in the data?

Safe Settings
: What are the technical and organizational controls in place to prevent a deliberate attempt to re-identify or to prevent a data breach?

Safe Data
: What is the level of identifiability, considering the people and settings of the data environment, and what threats to the data need to be managed?

3 We wrote a short practice article of this process in Luk Arbuckle and Felix Ritchie, "The Five Safes of Risk-Based Anonymization," *IEEE Security & Privacy* 17, no. 5 (October 2019): 84–89, https://oreil.ly/xi24Y.

Safe Outputs
: What are the concerns in using the anonymized data for the intended and other purposes, and what is a suitable identifiability threshold?

Imagine a health care scientist seeking access to data. In general, we may think that the use of health data will provide public benefit, and should therefore be supported. Health data can, however, be among the most sensitive data about individuals, revealing lifestyle habits, personal events that may trigger strong emotions, embarrassing information, or just things they want to keep private because, well, it's personal. Taking our responsibility toward data subjects seriously, we need to ensure the safe use of this data. Let's walk through the Five Safes in detail, keeping this example in mind.

Safe Projects

Our health care scientist is seeking access to personal data, which could be within the same institution but through a different department, or from an external organization. The intended use of the data may be of benefit to the data subjects, in their health care treatment or in some other way, or it may be more general and of potential benefit to the public at large. We need to capture all these considerations through an evaluation of data flow and primary and secondary purposes, before we decide whether to provide access and launch into the effort of ensuring the safe use of data.

Data flow
: It's important to understand the flow of data, to recognize legal and ethical boundaries and intended purposes so that we can identify the parameters needed to assess identifiability and create Safe Projects:

  - Where the collected data is coming from, who collected it, and the legal and ethical grounds for doing so
  - Where the shared data is going, who wants access, and the legal and ethical grounds for doing so
  - Whether the data is considered personal or not, and how anonymization is applied in accordance to laws and regulations

Once we've understood the flow of data, from data collection to transmission and use, we are in a position to evaluate the purposes for which the data will be used.

Primary and secondary purposes

Data is used for a primary purpose when used as intended at the time of data collection, and as understood by data subjects. It is used for a secondary purpose when used for anything else than what was originally intended. Understanding the legal

context for collection, approval mechanism,[4] and transparency will be important to determine the appropriate mechanisms for sharing data, especially for secondary purposes:

- The data custodian may have collected information for a primary purpose, such as providing care to a patient.
- Or the data custodian may have collected information explicitly for a secondary purpose, such as constructing a database of patients with diabetes for subsequent research.
- Personal information may also come indirectly through one or more data custodians, where permitted.
- Alternatively, data may come from another source claiming to be anonymized (which may need to be assessed in its own right before being used or combined with personal information).

When properly anonymized, data is no longer personal and therefore not subject to privacy laws and regulations, but ethical considerations should still be made. But first, the uses need to be understood to determine primary or secondary purposes, and determine potential legal requirements:

- An agent, acting on behalf of the data custodian, may use personal data for an authorized primary purpose (dictated by contractual agreements that define the services to be performed and the boundaries of operation).
- Depending on the jurisdiction, there may not be a legislative requirement to anonymize information that an agent uses for secondary purposes, or a requirement to obtain additional approval from data subjects for such uses. However, anonymization may be encouraged or desirable.
- The data custodian may also receive a request to share with an internal or external recipient for some secondary purpose. Sharing of personal data is sometimes mandatory, whereas at other times it may be discretionary to the data custodian. The conditions for discretionary sharing do vary.
- Other forms of data sharing that are not explicitly permitted in legislation require that either approval be obtained from the data subjects or the personal data be anonymized.

4 We use the word "approval" rather than "consent," because the latter can have very specific conditions and interpretations associated with it based on the relevant privacy laws.

Having separated primary from secondary uses, and legal requirements, we need to decide on what form of sharing will take place to determine the level of identifiability required, including nonidentifiable or anonymized.

When to anonymize

There are circumstances in which we may not anonymize the data, even when describing secondary uses. But there are also circumstances in which we may *want* to anonymize for the sake of protecting privacy, and circumstances in which we *must* do it. We divide these circumstances into four scenarios to consider in deciding how the data can be shared, and what if any anonymization is needed:

Mandatory sharing
: No approval from data subjects is required, and the data does not require anonymization because it is likely that individuals need to be identified (e.g., for law enforcement purposes). However, there may be considerable underreporting by indivuals due to privacy concerns.

Internal sharing
: It is often unnecessary for an agent to have data in identifiable form to perform their functions, even for primary purposes, and anonymization is desired to enhance privacy and avoid potential breaches.

Permitted sharing
: Approval from data subjects may be optional, under the discretion of the data custodian, for the public good (e.g., public health). There is reluctance, however, by data custodians to share personal data due to issues of individual and public trust, which anonymization can help remedy.

Other sharing
: When approval from data subjects is not possible or practical, and there are no exceptions in the legislation, the custodian must anonymize the personal data before sharing with a data recipient.

For our health care scientist, we'll assume their desired use of the data falls within the category of "other data sharing." Even if it had been a permitted sharing scenario, we would likely have wanted to anonymize the data, but given the legislative authority, our tolerance on identifiability would have been higher, meaning the release of more granular data. Since we will be anonymizing data for the health care scientist, we'll need to determine *who* will have access to the anonymized data, since this will impact the risk of potential misuse.

Safe People

Our health care scientist is unlikely to be the only one who will have access to the requested data. There may be analysts and technologists, perhaps even students, who will be working with the data. We need to understand the lab in which the data will be used, who will have access, and under what circumstances.

Data recipients are central to an assessment of context because the entity or employees may re-identify data, whether it be intentional or not. It may come as a surprise, but the anticipated recipient is also considered an adversary. This isn't to say that the anticipated recipient is malicious. Adversary is a general term meant to capture those entities that pose risks. Unintended recipients may also need to be considered, and therefore a more complete picture of all the possible recipients is warranted.

We assume that the adversary has access to the shared data, and has some background knowledge that will be used in a re-identification. The nature of that background knowledge will depend on the assumptions you are willing to make. Figure 3-2 provides some examples of the types of adversaries we consider, each of which will also have different depths of knowledge.

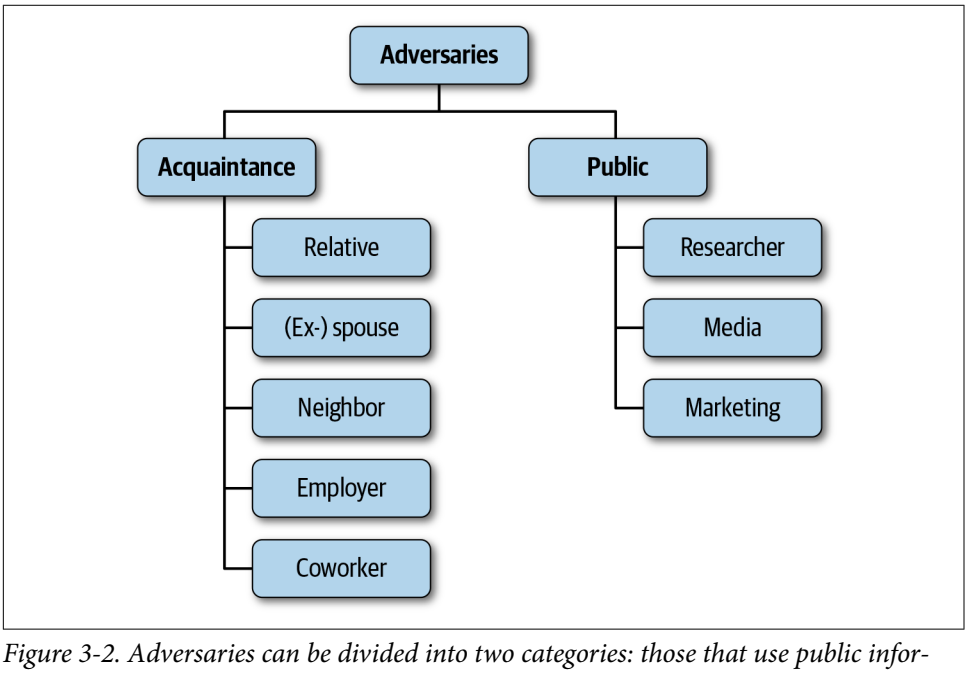


Figure 3-2. Adversaries can be divided into two categories: those that use public information to re-identify, and those that are acquaintances and have more in-depth knowledge to re-identify.

For example, the researcher, media, or marketer will use publicly available information to re-identify, which relates to the sample to population described previously in "Direction of Matching" on page 35. Whereas the relative, neighbor, and coworker will use publicly available information (since it's public, after all) *and* information that is known to them as acquaintances, which relates to the population to sample described previously in "Direction of Matching" on page 35. We therefore start to see how our concepts of re-identification science relate to scenarios that would form the basis of threat models, which will determine our degree of trust in the data recipients.

Recipient trust

The level of confidence that a data recipient will not attempt to re-identify the data can be called *recipient trust*. We can begin to evaluate the likelihood of an re-identification attempt by considering the potential adversaries in Figure 3-2. Consider the motives and capacity of the anticipated data recipient to re-identify the shared data. We assume that the data custodian is sharing data that has gone through some kind of anonymization.

When gauging recipient trust, it's important to consider both the motives and capacity of the data recipient:

Motives
: The motive to re-identify individuals in the data implies an intentional re-identification, considering issues such as conflicts of interest and the potential for financial gain from a re-identification.

Capacity
: The capacity to re-identify individuals in the data considers whether the data recipient has the skills and financial resources to re-identify the data.

Motives can be managed by having enforceable data sharing agreements or contracts with the data recipient. Such agreements will determine how likely a deliberate re-identification attempt would be. Contractual obligations need to include very specific clauses (otherwise the contract may allow for ways to re-identify a dataset):

- A prohibition on re-identification, on attempting to contact any of the patients in the data set, and on linking with other data sets without permission from the data custodian
- An audit requirement that allows the data custodian to conduct spot checks to ensure compliance with the agreement, or a requirement for regular third-party audits
- A prohibition on sharing the data with other third parties (so that the data custodian can keep track of who has the data), or a requirement to pass on the above restrictions to any other party the data is subsequently shared with

You can imagine our health care researcher being resistant to some of these clauses. Of course, the use of each of these clauses by a data custodian is optional, and to be determined by the use case and governance that a data custodian wants to have in place, based on their risk tolerance. We often hear of organizations, for example, that prefer not to have restrictions on linking with other data sets. This can be managed somewhat by adding some fine print, such as no linking to identified data, or personal data, or with data that may increase identifiability. But this won't preclude acquaintances from being potentially recognized in shared data.

Acquaintances

Recipient trust is about attempts to re-identify, but there is still a risk even when there is no attempt. Data recipients may have prior knowledge of personal information because they're acquaintances of individuals in the data (remember our list of potential adversies in Figure 3-2). This in turn may lead them to re-identify inadvertently, or spontaneously (yes, that's actually what it's sometimes called!), simply by recognizing them. It's a factor that needs to be considered when evaluating identifiability, because it relates to how safe it is to have people working with data.

Our health care scientist, and those working in their lab, provide a perfect example in the case that they are working with data in the same geography they're in or are from. In order to know intimate information about an acquaintance, those individuals working with the data would need to be some kind of friend. We can therefore incorporate in our models the probability that the adversary knows someone in the defined population covered by the data.

VIPs are also more identifiable in general, because more information about them is known publicly (we are all their acquaintance!). This would include individuals that are often in the public realm, and where there would be a media interest in writing about information that may be contained in the shared data, especially if it were unusual or pertinent to their public role. Typical VIPs would be politicians, actors and artists, and sports personalities.

A re-identification of a VIP may seem like a low likelihood event, although they are potentially more likely to be targets. A successful re-identification would, however, have a high impact, perhaps more damaging to public trust due to the increased media interest. The easiest approach to dealing with VIPs would be to remove them from shared data, rather than inflate the identifiability estimate for all data subjects. That being said, the data transformations that are planned may be sufficient if identifiers known to acquaintances are included, and if the data is a sample.

Trusting the data recipient is not the same as trusting the environment in which they will be working with the shared data. Consider how controls around the shared data may regulate the degree of responsible uses, and prevent loss or theft of data.

Safe Settings

We need to assess the data environment of our health care scientist, that is, the environment in which the shared data will be used. If anyone in an organization can walk in and use the data, we know the environment is on the low end of safe, and this will leave a significant residual risk to account for through data transformations. On the other end of the spectrum, a safer environment will mean more granular data for our health care scientist and their team. Any data localization restrictions would already be factored into the Safe Projects.

The security and privacy practices of the data recipient will have an impact on the likelihood of a rogue employee at the data recipient's site being able to re-identify the shared data. A rogue employee may not necessarily be bound by a contract unless there are strong mitigating controls in place. It also determines the likelihood of an outsider gaining access to the shared data.

An evaluation of mitigating controls needs to be detailed and evidence based, preferably mapped to existing professional, international, and government regulations, standards, and policies, including ISO/IEC 27002, where appropriate.[5] Using a standardized approach also ensures consistency, not only for a single organization that is sharing data, but across organizations, e.g., the HITRUST De-Identification Framework.[6]

There are several mitigating controls that need to be considered in dealing with personal data, and to ensure that the assessment of Safe Settings is defensible. These are considered the most basic forms of controls. Think of them as minimum standards only! We can only give you a taste of what's expected in the subsections that follow, because it's pretty detailed (although this summary covers a lot of ground).

Controlling access, disclosure, retention, and disposition of personal data:

- Only authorized staff should have access to data, and only when they need it to do their jobs.
- There should be data sharing agreements in place with collaborators and subcontractors, and all should have to sign nondisclosure or confidentiality agreements.
- There should be a data retention policy with limits on long-term use, and regular purging of data to reduce vulnerability to breaches.

5 ISO/IEC 27002, *Code of Practice for Information Security Controls* (2013), *https://oreil.ly/StVGF*.

6 We've mentioned this one before, but it's worth repeating that it can be found in HITRUST Alliance, "HITRUST De-Identification Network," *https://oreil.ly/i6t4q*.

- If any data is going to leave the relevant jurisdiction in which the data sharing is taking place, there should be enforceable data sharing agreements and policies in place to control disclosure to third parties.

Safeguarding personal data:

- It's important to respond to complaints or incidents, and that all staff receive privacy, confidentiality, and security training.
- Personnel need to be disciplined for violations of these policies and procedures, and there should be a tried and tested protocol for privacy breaches.
- Authentication measures must be in place, with logs that can be used to investigate an incident.
- Data can be accessed remotely, but that access must be secure and logged.
- On the technical side, a regularly updated program needs to be in place to prevent malicious or mobile code from being run on servers, workstations, and mobile devices, and data should be transmitted securely.
- It's also necessary to have physical security in place to protect access to computers and files, with mandatory photo ID.

Ensuring accountability and transparency in the management of personal data:

- There should be someone in a position of seniority who is accountable for the privacy, confidentiality, and security of data, and there needs to be a way to contact that person.
- Internal or external auditing and monitoring mechanisms also need to be in place.

Risk matrix

A detailed assessment of Safe Settings can be combined with our assessment of Safe People to create a standard risk matrix to capture the probability that an internal adversary may attempt to re-identify, as shown in Figure 3-3. If you've ever seen a risk matrix before, they usually contain subjective entries (e.g., low, medium, and high). The entries in our risk matrix are, however, known as expert probabilities, which have been derived from past data releases by reputable organizations and regulatory or industry guidance.[7]

7 See Chapter 18 of El Emam, *Guide to the De-Identification of Personal Health Information*, *https://oreil.ly/mlg_O*.

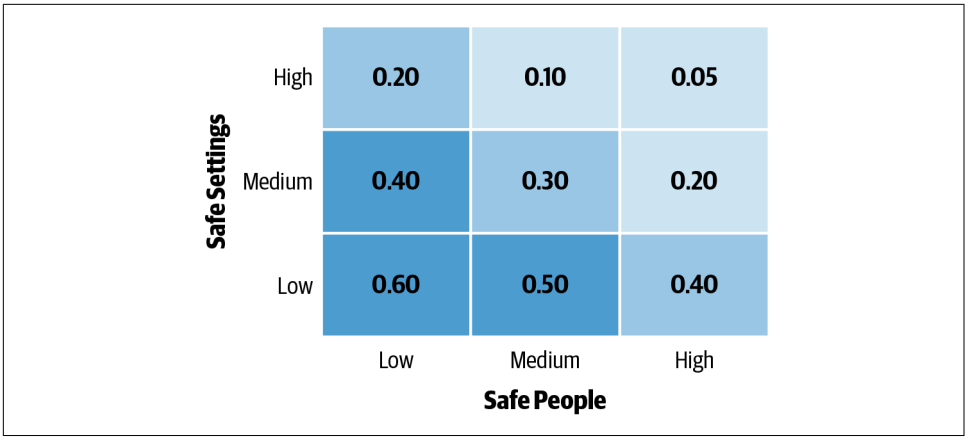


Figure 3-3. A risk matrix provides a visual demonstration of risks to assist decision making, and in this case the likelihood of an attempt to re-identify.

Having expert probabilities, instead of subjective categories of low, medium, and high, allows us to combine the entries with measures of identifiability in the data itself, which we'll see explicitly when we discuss Safe Data. This means that we can assign a probability of attempting to re-identify to our researcher and lab personnel. As can be seen from the risk matrix, the more we can trust the recipients, and the stronger the privacy and security settings, the lower the assigned probability they will attempt to re-identify data. And with that, we are now in a position to consider how safe data is.

Safe Data

At this point we've done everything we can to capture the mitigating controls in place, both technical and organizational, to evaluate our health care scientist's data environment. We are now left with reducing residual risks through data transformations.

An assessment of Safe People and Safe Settings results in an evaluation of context. A structured approach can be used to assess context and evaluate whether an attack will be realized, known as *threat modeling*. Consistent with the modeling of threat sources used in information security and risk modeling, there are three plausible attacks that can be made on data:[8]

8 See, for example, ISO/IEC 27005:2018 "Information Security Risk Management" (2018), *https://oreil.ly/AxU7r;* NIST SP 800-30 "Risk Management Guide for Information Technology Systems" (2012), *https://oreil.ly/b8LJb;* and CSE TRA-1 "Harmonized Threat and Risk Assessment (TRA) Methodology" (2007), *https://oreil.ly/dyNnD*.

Deliberate
: A targeted attempt by the data recipient as an entity, or a rogue employee due to a lack of sufficient controls, to re-identify individuals in the shared data. The risk matrix from Figure 3-3 is used to capture this probability.

Accidental (inadvertent)
: An inadvertent or unintentional re-identification, for example an individual being recognized while a recipient is working with the shared data. This probability can be estimated as that of having at least one acquaintance in the defined population.[9]

Environmental (breach)
: The data could also be lost or stolen in the case where all the controls put in place have failed to prevent a data breach. Industry-specific rates provide a means to estimate the probability of a data breach.

To produce Safe Data, the overall identifiability needs to be assessed, which is a combination of context (the probability of an attack) and data (the probability of re-identification when there is an attack).[10] As summarized in Figure 3-4, this will drive the data transformations required to reduce identifiability so that residual risks are appropriately managed. But to do this, we need to build models of identifiability.

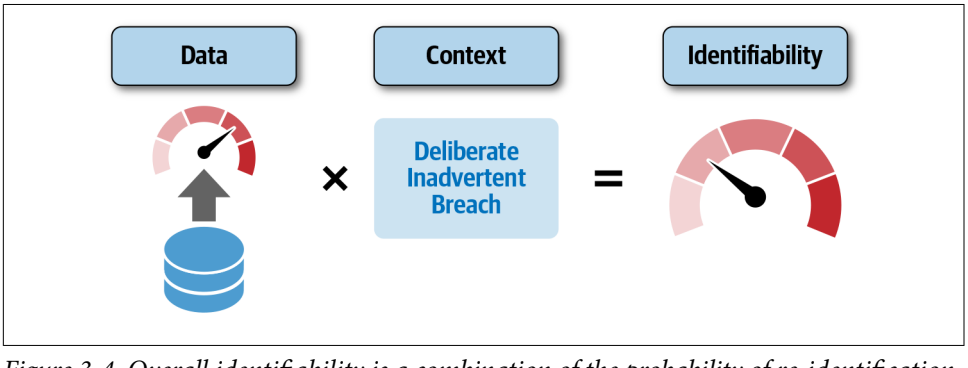


Figure 3-4. Overall identifiability is a combination of the probability of re-identification in the data given an attack times the probability of an attack in the first place (determined through threat modeling).

9 On average, people tend to have 150 friends, called the *Dunbar number*. Given the prevalence ρ of a knowable characteristic that defines the population of the data, the probability of having an acquaintance in the population can be computed in a straighforward manner using $1\text{-}(1\text{-}\rho)^{150}$. See Chapter 2 of El Emam and Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*.

10 Conceptually, this conditional probability was first presentined in Catherine Marsh et al., "The Case for Samples of Anonymized Records from the 1991 Census," *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 154, no. 2 (1991):305-340, *https://oreil.ly/ZfDKB*.

Quantifying identifiability

Because identifiability estimation invariably requires the use of statistical methods, any identifiability estimation technique will be based on a model of plausible re-identification attacks, and models make assumptions about the real world. Therefore, identifiability estimation will always imply a series of assumptions that need to be made explicit. Furthermore, because of the statistical nature of identifiability estimation, there will also be uncertainty in these estimations, and this uncertainty needs to be taken into account.

The identifiability estimation we're referring to applies to indirectly identifying data. Three kinds of attacks need to be managed, for which detailed metrics can be derived:[11]

Prosecutor attack
: The prosecutor has background information about a specific person who is known to them, and uses this background information to search for a matching record in the shared data.

Journalist attack
: The journalist doesn't know the particular individual is in the shared data, which is a subset of a larger public dataset, but does know that all the people in the data exist in a larger public dataset.

Marketer attack
: The marketer is less concerned if some of the records are misidentified. Here, the disclosure pertains to everyone in the data. The probability of disclosure from a marketer attack is always less than from prosecutor or journalist, and is therefore often ignored.

In practice, either a prosecutor or journalist attack is considered, as they represent targeted attacks, whereas a marketer attack is an average and will always be a lower probability. A prosecutor attack is considered when a target individual is known to be in the shared data, whereas a journalist attack is used when the target individual is in a larger defined population in which the shared data is only a sample. In other words, if the shared data represents the entire defined population, evaluate disclosure from a prosecutor attack; if the shared data represents a sample from the defined population, estimate disclosure from a journalist attack.

11 See Chapter 18 of El Emam, *Guide to the De-Identification of Personal Health Information*, https://oreil.ly/u0Vqf.

Prosecutor and journalist, although representing targeted attacks, are forms of average identifiability. There is still a chance that uniques remain in the data, even though on average the cluster sizes are much larger. For this reason, we advocated for what we termed *strict average* in our previous book,[12] in which a maximum identifiability metric is included that ensures there are no population uniques in the data (in matching from sample to population, as described in "Direction of Matching" on page 35).

If a population registry has information about individuals who are known to be in the shared data, an adversary may target the most identifiable data subjects. In this case, the maximum of the identifiability metric is taken across all data subjects when there are no controls in place to prevent such an attack (e.g., public data sharing). On the other hand, if an adversary will not necessarily target the identifiable data subjects, because there are controls in place to prevent such an attack, but is trying to find information about a specific individual, the identifiability metric is averaged across all data subjects since the target is random (e.g., private data sharing).

Having models in place to quantify identifiability puts us in a position to consider the outputs that will be derived from the data. In other words, we can measure identifiability, so we need to decide how identifiable the outputs should be.

Safe Outputs

Once we share data with the health care scientist, it should go without saying that they will produce models and statistics. The scientist and team want to learn from the data. The question is, what do they want to learn, and how will they use this information? We should have captured their purposes under Safe Projects, but it's possible they will find other uses that we need to keep an eye on. The anonymized data itself is an output to that health care lab, but so are the analytical results and decisions they make from the data. We want to ensure these are not disclosable in a way that would be deemed inappropriate.

Imagine that the health care scientist stated up front, in their request for data, that they wanted to study vaccination rates. Through the use of the anonymized data, the scientist finds that there is a population group that is under-vaccinated. The scientist may now take this information and launch a targeted education campaign, or publish in local media as well as in an academic journal.

Although public education seems laudable, the disclosure may result in that population group being targeted by others in the community in less than ideal ways, through shaming or being treated in a biased way. These decisions to release study

12 El Emam and Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*.

results require careful consideration. Not to mention if the results are also used in other ways, such as for marketing purposes. Although we would like to capture as many of these as possible in defining a Safe Project, we must recognize that circumstances change once the results are in hand and understood.

Ultimately, the degree of data transformation necessary to reduce identifiability to a suitable tolerance level, used to provide reasonable assurance that data is nonidentifiable, raises the question of thresholds. There are many precedents going back multiple decades for what a suitable cell size is for sharing anonymized data, with a range of options shown in Figure 3-5. Recall from "Anonymization in Practice" on page 18 that the threshold is a probability derived from cell-size rules, which determine the minimum number of individual contributions that need to be included in an aggregation of data. To decide which threshold to use, we can look at the potential *invasion of privacy*, which represents the sensitivity of the data and the approval mechanism that was in place when the data was originally collected.

Invasion of privacy

A subjective criterion can be used by the data custodian to influence the selection of an identifiability threshold to determine the potential invasion of privacy. If the invasion of privacy is deemed to be high, that should skew the decision more toward a lower threshold. On the other hand, if the invasion of privacy is deemed to be low, a higher threshold would be selected. Here are some questions to consider when evaluating for invasion of privacy:

- What are the legitimate benefits to individuals or society from analyzing the shared data?
- Is the data highly detailed, is it highly sensitive and personal in nature?
- What is the potential injury to individuals from an inappropriate processing of the data?
- What is the appropriateness of approval by data subjects for disclosing the data?

Although approval is not required of data subjects for sharing properly anonymized data, the sharing of data would not be considered as invasive of privacy when approval has been provided by data subjects compared to when no approval is sought. There are in fact multiple levels of notice and approval that can exist for the sharing of anonymized data:

- There is a court order or a provision in the relevant legislation permitting the sharing and use of the data without notice of or approval by data subjects.
- The data was unsolicited or given freely or voluntarily by the data subjects, with little expectation of it being maintained in total confidence.

- The data subjects have provided express approval that their data can be shared and used for this purpose when it was originally collected or at some point since then.
- The data custodian has consulted well-defined groups or communities regarding the sharing and use of the data, and had a positive response.
- A strategy for informing or notifying the public about potential sharing and use for the data requestor's purpose was in place when the data was collected or since then.
- Obtaining approval from data subjects at this point is inappropriate or impractical.

The practical consequence of evaluating invasion of privacy is that the suitable probability threshold (one over the cell size, or the definition of nonidentifiable) will be lower under the most invasive scenario (meaning larger average clusters, or equivalent cell sizes, are needed). This is shown in Figure 3-5. Even under the most invasive scenario, however, it is possible to share the data, but the degree of data transformation would be greater.

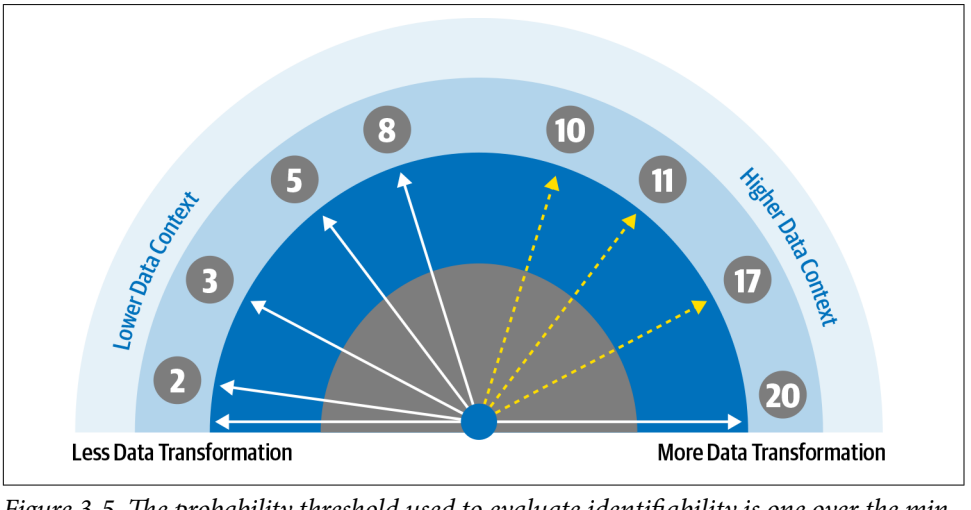


Figure 3-5. The probability threshold used to evaluate identifiability is one over the minimum cell size for the intended sharing. This threshold needs to be quantitative and defensible. The cell sizes in this diagram are benchmarks based on past precedents.

Now that you have an understanding of what the Five Safes are, in order to really understand them we need to give you an example of how they work in practice.

Five Safes in Practice

Let's pull together the information presented so far into a risk-based assessment of identifiability. We'll use our health care scientist as our example, but be more specific about the context of the data sharing.

Safe Project
: The personal data collected is from a hospital that wishes to leverage its data assets for scientific research into treatments and effects of cancer. Data will be made available to health care scientists for secondary research purposes, but only after an ethics review to ensure the uses are deemed appropriate. The environment in which the data will be used will be external to the hospital, but in the same jurisdiction.

Safe People
: The hospital has decided that only identified researchers from approved research institutions will have access to data. It has recognized that some researchers will have analysts and technologists, perhaps even students, that will be involved in working with data, and could have acquaintances in the data. For this reason, contracts will be required with the research institution to ensure there is oversight, and all staff who access data will be required to take privacy training and sign agreements regarding the ethical use of data they are being entrusted with. There should be no obvious reason to want to re-identify data (i.e., low motives and capacity).

Safe Settings
: Although the hospital would like to make the data broadly available, it is not in a position to assess the data environment on a case-by-case basis. Rather, it will only share data when the risk level is deemed low, which will be spelled out in standard data sharing contracts. These will result in a fixed data sharing context, simplifying data sharing from its perspective, and will require the institution to agree to be accountable for the research scientist's lab environment.

Safe Data
: With the previous information, the hospital is in a position to assess the data sharing context based on plausible attacks, which represent the context in which the data will be stored and used:

  Deliberate
  : The Safe People in this case have been assumed to have low motives and capacity to re-identify. The Safe Settings are fixed so that the privacy and security controls will be high. Combined, Safe People and Safe Settings are mapped to the risk matrix in Figure 3-3 to provide an expert probability of 0.05.

Accidental (inadvertent)
: The most prevalent disease in the population will be breast cancer, and the probability of knowing at least one woman with breast cancer is about 0.70. On the other hand, knowing at least one person with oral cancer, which is much less common, is about 0.054. This means the chance of having an acquaintance in the data will vary based on cancer type, although a conservative estimate would use breast cancer since it's the most common and therefore the most likely to include an acquaintance.

Environment (breach)
: Breach rates vary by the level of privacy and security controls. Previous breach rates in US health care were reported to be 0.14 for strong controls, which we can use here for this exercise, but this should be verified based on industry and jurisdiction, where breach rates are available.

Safe Outputs
: To ensure the data is appropriately transformed, based on data and context, a threshold needs to be defined. Although health data is inherently sensitive, cancer data is not itself more sensitive (compared to things like abuse, sexual orientation, etc.). Based on benchmarks, we will assume an appropriate cell size of 10, which is a probability threshold of 0.10 (or 1/10).

To operationalize the above, the hospital will measure identifiability of the data, combined with context, to determine how much the data needs to be transformed. If we assume data of oral cancer is being shared, based on the above the primary driver of risk will be from the data environment, with the risk of data being lost or stolen. In other words, we have a context probability of 0.14, and an overall threshold of 0.10, so that the data will need to be transformed so that $0.14 * data \leq 0.10$. The transformed data will therefore have an average equivalent cell size of 10 or more, and be deemed nonidentifiable in this data sharing context.

Probabilistic measures of identifiability are tools to drive decision making, providing guidance on what aspects of data sharing, be it context or data, need to be modified to ensure that identifiability is appropriately managed. The measures are based on a long history in statistical disclosure control. The modeling is useful, but as we already pointed out, there is subjectivity in all models. Our goal is to ensure the models are defensible, while capturing the broader context of data sharing to ensure our picture of identifiability is complete and reasonable. The Five Safes provides a framework to capture this context in a (hopefully) memorable way.

Final Thoughts

In many jurisdictions, demonstrating that data is nonidentifiable is a legal or regulatory requirement. Our methodology provides a basis for meeting these requirements

in a defensible, evidence-based way. We have demonstrated how the Five Safes framework can be operationalized using risk-based anonymization: each dimension is evaluated independently of the others, brought together by an overall assessment of identifiability. This allows for the evaluation of scenarios of responsible data sharing, which will be context driven given the impact different scenarios will have on the usefulness of the data.

Data utility is important for those using anonymized data, because the results of their analyses are critical for informing services provided, policy, and investment decisions. Also, the cost of getting access to data is not trivial, making it important to ensure the quality of the data received. We don't want to be wasteful, spending time and money collecting high-quality data, only to then watch that quality deteriorate through anonymization practices meant to prepare the data for secondary use.

The impact of anonymization on data utility is important, and very context driven. All stakeholders need to provide input on what is most important to them, be it data utility or privacy. It's not easy to balance the needs of everyone involved, but open communication and a commitment to producing useful data that is nonidentifiable is all that is really needed to get started. It's not an easy negotiation—and it may be iterative—but its importance cannot be underestimated. Ideally, framing that conversation around the Five Safes should help to clarify the most important points.

With the concepts of identifiability in hand, both in terms of the science and with a practical risk-management framework to capture the context of data sharing, we can turn our attention to building anonymization pipelines. To do this, we will start with identified data in the next chapter, and introduce many important concepts of privacy and data protection, using the Five Safes where it will help underscore important points.

CHAPTER 4
Identified Data

If you're in the business of working with identified data, with people's names, addresses, and other unique identifiers, you should already have the tools in place to protect that data. If you're thinking of working with identified data, well, you have a lot of standards and the like to learn. We won't be going through all of these, as that's not the focus of this book. The Five Safes of risk-based anonymization we discussed in the last chapter have provided us with the contextual elements we will need to dig into to manage risk, and in this case from the starting point of collecting identified data.

We want to provide you with some strategic privacy considerations in working with identifiable data, considerations that would fall within the realm of privacy engineering.[1] It just so happens that identifiability will play a key role in that, since it plays a key role in privacy laws and regulations in general. There are other considerations we'll explore, and we want to arm you with some basic tools and provide you with an understanding of how they interact with one another. Just being aware of them can help you in your design of systems that will process personal data (Privacy by Design!), or in updating systems to be more privacy friendly.

For many organizations, anonymization will start from their own store of identified data. Think of this as pushing data out, from identified to anonymized (with a detour through pseudonymization, but we're keeping that for another chapter). This sharing of data may be to another department within the same organization or to an entirely different organization. These pose different challenges, which we'll explore in this chapter. This should be a natural progression, from privacy engineering in general to anonymization more specifically.

1 Privacy engineering is systems engineering focused on integrating privacy objectives and privacy risk assessment into implementation requirements, in which it is understood that there is no such thing as zero risk.

Requirements Gathering

As with any engineering project, we start with requirements gathering, but, in our particular context, privacy-related requirements gathering. This will primarily involve three broad categories: use cases, data flow and data use, and data and data subjects. Evaluating these three categories will help tease out the wants and needs from a privacy perspective, and a series of probing questions can be used to better understand the details of those wants and needs. Not all these questions need be asked and answered, but they can form the baseline of what needs to be understood to gather privacy-related requirements and define privacy objectives.

Many of the privacy considerations we will work through as requirements engineering can be motived by the process of a *privacy impact assessment*, or risk assessments in general. Although often done at the end of a design cycle, the criteria found in these assessments should make their way into design as more granular expectations of privacy design. This way we leverage accepted standards and frameworks to inform the design process.

Let's work through this as though we were actually working to design a system (either from scratch or a retrofit). It could be our design, it could be someone else's, we just want to work through the steps to capture as much detail as needed in the project definition phase. We will not be delving into system design and development, and will leave the implementation to another book. Some aspects will certainly require a privacy policy or legal analysis, but we can't account for all privacy laws and regulations the world over, or the shifting privacy landscape, and we consider some of this to be material for the implementation phase.

Use Cases

For our use case, we're going to attempt to understand how a system will be used so that we can scope the privacy issues and possible solutions. It's critical to evaluate use cases in order to understand how a system will be used, so this is where we most often start gathering requirements. Thinking back to our discussion of purpose specification in "Safe Projects" on page 46, which you can think of as a concept definition in an engineering project, we should have a general idea of what sort of system we're envisioning. But at this stage we go deeper, as we want to understand the interactions with the system to determine where privacy protections could be put in place.

A use-case analysis in the area of privacy engineering can be focused around the three main objectives a system should strive for to demonstrate a desired level of trustworthiness (echoing what are known as the fair information practice principles):[2]

Predictable
: It should be possible to predict how a system will behave. This means meeting accountability requirements by ensuring that interactions and outcomes are expected. This can be achieved through purpose specification and use limitation, and a degree of transparency through forms of notice that will be provided to, or approval sought from, data subjects so that they can also predict what will happen with data about them.

Manageable
: All systems will require controls on how personal data is handled, from ingestion to internal processing and export. How manageable a system is will be determined by the granularity of administration in handling data in the defined use cases. This can be achieved by supporting alteration, deletion, or selective disclosure of personal data (and this can incorporate individual control, if desired).

Disassociated
: The different use cases supported by a system will require different levels of identifiability and data minimization. Direct interaction with data subjects will require names or other directly identifying information, whereas in other use cases they may be replaced with with tokens or pseudonyms. For analytic processing, identifiability may be further reduced to the point of being nonidentifiabile (anonymized) data.

The purpose of these objectives, summarized in Figure 4-1, is to meet the needs of more detailed privacy principles, be they enshrined in privacy laws/regulations or not, with measurable controls. We provide the above objectives to get you on your way to undestanding the basics of privacy engineering. Notice, however, that one of these principles (to disassociate individuals or groups from the data) is that of identifiability! As we said, this is a core element of privacy. Although the objectives of the system being predictable and manageable are broader, they can also be seen as supporting the objective of having data subjects disassociated from the data.

2 These objectives can be found in Sean W. Brooks et al., "An Introduction to Privacy Engineering and Risk Management in Federal Information Systems," NIST Interagency/Internal Report (NISTIR)-8062 (2017), *https://oreil.ly/bM0kS*.

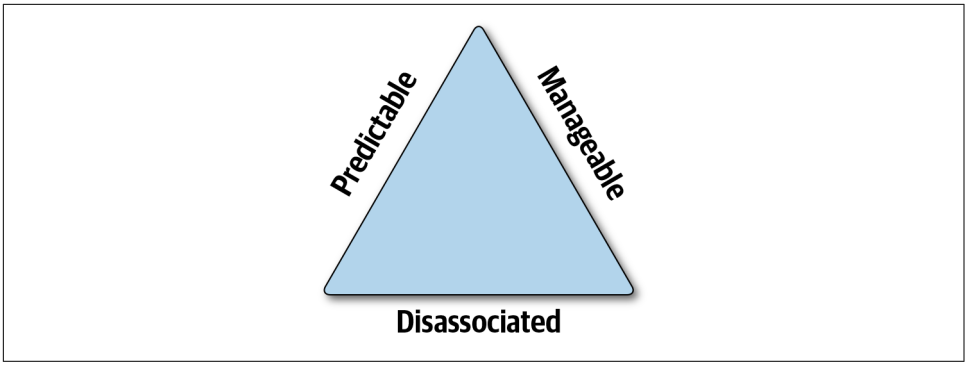


Figure 4-1. Our objectives in implementing measurable controls can be summarized by the privacy-engineering triad.

Of course, this is not enough information to drive a conversation and tease out privacy requirements, so we provide a series of "Probing Questions to Understand Use Cases". Better yet, you can think of these as opportunities to integrate privacy into your systems. Whether you're at the start or end of the design and development phase, or even revisiting a system in light of privacy and trust considerations, these questions can help you get to the bottom of which things your system truly needs to operate successfully in bringing privacy to the forefront of your operations.

Probing Questions to Understand Use Cases

- What are the interactions between the users and the supporting systems themselves (i.e., the business/system use cases)?
  - — Who are the primary and secondary actors?
  - — Are there any preconditions before a use case will take place, what is the guaranteed outcome, and what is the trigger for the use case?
  - — Provide a use case diagram, if possible, even a rough draft of plans.
- What are the assumptions that support the use case/data flow to ensure the processing of personal information is *predictable* (and thereby meets accountability requirements)?
  - — Purpose specification and use limitation?
  - — What forms of notice will be provided to, or approval sought from, data subjects (i.e., the degree of transparency)?
- What degree of control over personal information is required in the use case/data flow to ensure it is *manageable* (i.e., the granularity of administration)?
  - — Will there be mechanisms in place to support alteration, deletion, or selective disclosure of personal information?

  - — Is there a data life cycle management plan to define and automate the stages of data from cradle to grave?
- What identifiable information is operationally required for the use case/data flow, i.e., can the data be *disassociated* from an individual or group?
  - — Could the direct identifiers be replaced with a token or pseudonym, could an irreversible key be used while also eliminating the ability to single out a data subject based on public identifiability, or should the personal data be anonymized?
  - — Are there other data attributes that could be minimized to avoid disclosures of identity or associated activities, based on use cases and neeeds?
  - — What technical and administrative (organizational) controls will be in place (e.g., controlling access, disclosure, retention, and disposition of personal data, safeguarding personal data, and ensuring accountability and transparency in the management of personal data)?

The privacy-engineering objectives should each contribute in some way to enhancing privacy, but it's not all or nothing. The point is to find a balance between the objectives that is driven by the use case, as shown in Figure 4-2. In this example, the right balance was found with less need to be predictable but more need for manageable data even though the data is largely disassociated from data subjects. This is why use-case analysis is so important, to tease out wants and needs.

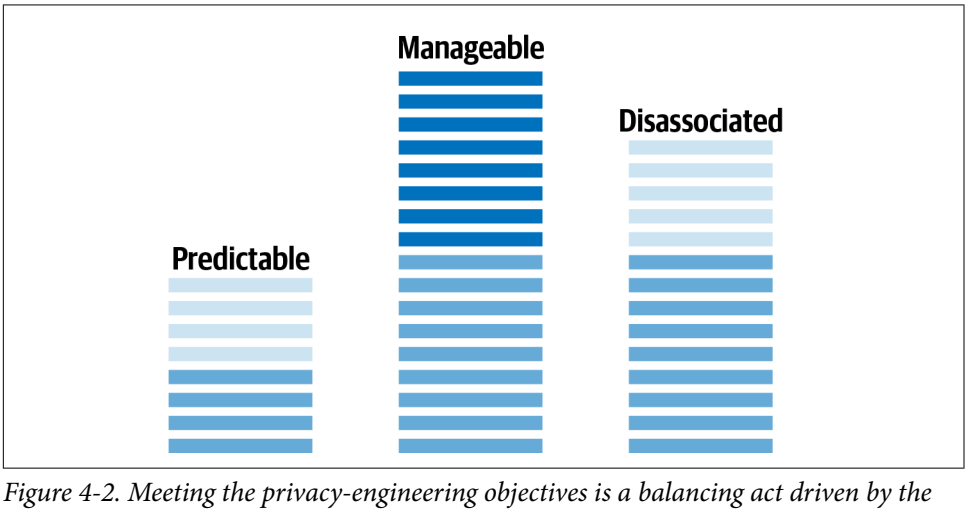


Figure 4-2. Meeting the privacy-engineering objectives is a balancing act driven by the wants and needs identified by the use case.

Say the use case was to aggregate reports that are used internally. By their very nature, these reports have limited use since the data is just a summary of information. The

aggregated information may be sufficiently disassociated from data subjects that there isn't much need for a system using these aggregate reports to be predictable, by specifying purpose and use limitation, since there is little chance they can be misused. However, aggregate information can still result in unwanted disclosures unless there are certain rules in place, such as avoiding any small aggregate counts (directly or by overlapping pieces of information). Therefore there may still be a need to ensure that a system using this information is manageable.

Data Flows

To go even deeper, and drive that conversation forward, we can discuss data flows in detail. We suggested a review of data flows in the context of the "Safe Projects" on page 46, with the purpose of identifying the data sharing scenarios as they specifically relate to identifiability. Data flows are really a continuation of our use case analysis. In our discussion of Safe Projects, we were primarily concerned with the criteria and constraints of a system, whereas now we are concerned with legal and ethical boundaries. We need to plan for the possibility of our different sharing scenarios from a legal perspective but also in order to design the appropriate sharing mechanisms:

Mandatory sharing
: If law enforcement or public health officials require access or copies of data, how will this be provided to them? There are also privacy laws and regulations that require that data subjects themselves have access to personal data, to know what is collected about them and give them rights to make corrections or amendments. Access may be interpreted broadly, and does not necessarily mean that data subjects have the ability to directly go into a system and make changes, which could be impractical and even damaging in some cases. Data subjects may also have a right to get copies of data about them in the name of data portability. The data will need to be identified in these cases.

Internal sharing
: The use cases that have been planned or developed may require various forms of internal access, or even that copies be transferred to a different department or unit of the organization. Here we are assuming that the sharing of identifiable data is permitted, as a primary purpose that supports the interactions with data subjects, or secondary purposes that privacy laws and regulations allow. We'll need to know who will have access and for what purpose to ensure the sharing is truly permitted, and whether the stated levels of identifiability described in the use case analysis are truly required, or whether greater degrees of dissociation would be acceptable given the concerns of privacy and trust.

Permitted sharing
: Just because it's allowed doesn't mean you will want to share identifiable data. If you are sharing with a third party for a secondary purpose, ask yourself if data

subjects would be surprised or upset with third-party access to information about them. This will depend on the third party, cultural norms, and, ultimately, trust (between data subjects and the third party, but also between data subjects and your organization). Transparency and anonymization will greatly improve the trust relationship.

Other sharing
: Every other scenario in which personal data sharing is not expressly permitted by privacy laws or regulations will need anonymization. Your anonymization pipeline may start with any of the above scenarios, but under this scenario we are referring to anonymization that will ensure the data is no longer identifiable, so that it can be responsibly shared. Understanding data flows is still critical here, so that we can apply a risk-based approach that ensures the most granular and useful data is made available.

We provide a list of "Probing Questions to Understand Data Flows" to help tease out the necessary details needed to understand legal and ethical boundaries. You'll notice questions involving geographic considerations, because privacy laws and regulations vary across the world, and you will need to consider cross-border data transfers and data localization laws or regulations (which require that personal data be hosted and remain within the country, unless properly anonymized to meet the highest standard of privacy protection).

Probing Questions to Understand Data Flows

- What are the data flows, from source to recipient access or use, including all data transfers and points where data transformation may occur?
  - Please provide a data flow diagram, if possible, even a rough draft of plans.
- Who owns or is the custodian of the data in question?
  - Is the client the custodian or recipient?
  - Is the custodian aware of the plan for risk mitigation?
- Where are the recipients, i.e., from what jurisdiction or geographic location will data be accessible?
  - What other data will the recipient have access to that could be associated or linked to the data that is intended to be accessible or shared?
  - How frequently will the data be accessed or shared?
  - How will the data be accessed or shared (e.g., accessed via a portal, delivered to recipient)?

Data and Data Subjects

We've figured out the use cases and data flows. Now we need to consider the data itself, and who is represented in that data. The type and structure of data may define the practicality of solutions to mitigate privacy risks, and various properties of data need to be understood to determine both identifiability and the potential invasion of privacy. This evaluation will also focus on who the data subjects are, and the expectations associated with processing of data about these subjects. Again we provide a series of "Probing Questions to Understand Data and Data Subjects" to help you through this process.

Data subjects

Consider the parameters or criteria for individuals being included in the data, and any information about other individuals that come along for the ride, such as their relatives or neighbors. Where they are from can change legal requirements, especially if they were collected intentionally. For example, if a product or service is targeting a country other than the one in which the data is actually being stored, the privacy laws and regulations of the data subject's country of residence will apply.

While it's true that it may be difficult for regulators to enforce the extraterritorial reach of privacy laws and regulations, they have and will do so as deemed necessary. Some laws and regulations have been designed with this in mind (e.g., the GDPR in the EU), while others have been interpreted as such by the courts. Either way, it's best to stay on the right side of the law and plan accordingly.

Be sure to document details of how personal data was collected, stored, protected, and used. To be auditable and defensible requires documented proof of data protections to demonstrate the reasonable measures taken to respect legal obligations and respect the expectations of data subjects. One aspect to meet those obligations is data minimization, which means understanding the data to be shared.

Structure and properties of the data

In considering the structure and properties of the data that's collected or shared, we again need to consider the stakeholder wants and needs for a system (the mantra of requirements gathering). In designing a system with privacy in mind, we need to repeatedly review and ensure that the collection and sharing supports purpose specification. Otherwise it's all too easy to slip into the habit of getting all the data that's possibly available, and one basic privacy principle is that of data minimization.

Only collect or share what you need, when you need it, for as long as you need it, and for the purposes that were specified. In the spirit of transparency, you will most likely be letting data subjects know that you're using data for a specific purpose, and your system should stick to that purpose. But even if you're not letting them know directly, it should be easy for them to understand what data a system is using based on what it does.

You may recall from the "Safe Outputs" on page 57 that we use a subjective criterion to select a threshold. This same approach can be used to consider data collection and sharing in general, since it defines identifiability tolerance. Ultimately, if you're collecting personal data, you want to reduce the potential to invade the lives of the people whose data you're entrusted with. We repeat the categories we consider in this subjective assessment of risk tolerance, and provide some additional detail to consider in defining what data you truly need. These can form part of a privacy impact assessment.

Data sensitivity
: Consider the level of detail that's needed: how many variables of information, the granularity and precision of that data, how many domains of information will be collected, and whether those domains need to be joined, etc. Also consider the sensitivity of the information collected. Certain privacy laws or regulations single out certain categories of data as being particularly sensitive, such as health information, genetic or biometric data, race or ethnicity, political opinions or religious beliefs, and a person's sexual activity or orientation.

Potential injury
: Breach notification laws or regulations can provide an indication of how regulators set the bar on potential injury to data subjects if the data is lost or stolen, or processed inappropriately. You will also want to consider how such incidents may cause direct and quantifiable damages, and measurable injury to the data subject. And consider your ability, as an organization, to enforce contracts or data sharing agreements, for internal or external data sharing.

Appropriateness of approval
: Data subjects can provide approval to participate, implicitly or explicitly, in the collection and sharing of data. They should have a basic understanding of the data collected or shared about them based on the interaction with the organization or their systems. They may even have volunteered the data, or been consulting in how the data was to be used. However, their approval is not always required, as we've seen in discussing mandatory or permitted sharing.

With that in mind, we can consider how different categories of information can affect privacy or confidentiality. What data is needed needs to be put in this context.

Categories of information

We've already mentioned directly and indirectly identifying information in the previous chapters. But now that we're working with identified data, we need to spell this out clearly and go the extra step of classifying data for the purposes of making decisions about it, so that we can determine what tools can be used to protect it accordingly. We'll revisit tools and techniques later, and focus here on the types of data that may be collected and used.

Directly identifying
: Attributes that can essentially be used alone to uniquely identify individuals or their households, such as names and known identifiers. These should only be kept for identified data, and even then you may choose to separate directly identifying attributes into a separate dataset that is linkable to the other personal data. When we want to reduce identifiability, these attributes are always removed and replaced with fake random data or with pseudonyms or tokens. The techniques used need to be robust and defensible. This is often called masking or pseudonymization (e.g., in the EU).

Indirectly identifying
: Attributes that can be used in combination with one another to identify individuals, such as known demographics and events, may need to be modified or transformed to reduce risk. These are the attributes used to measure identifiability, and are not immediately removed from the shared data because they are extremely useful for analytics. This is where all the heavy lifting takes place in terms of anonymization, because we want to minimize information loss to maintain analytic utility. We can divide these into two classes, which generally have different levels of risk:

  - Knowable to the public, such as fixed demographics
  - Knowable to an acquaintance, such as encounter dates and longitudinal characteristics or events

Confidential or target data
: Attributes that are not identifiable but would be learned from working with the data, such as behaviors and preferences. Target data may still be found in data that is anonymized, and can pose ethical considerations regarding its use. Often, when classifying personal data as identifying or not, everything that is not identifying is considered target data. Not everything is identifying, but probably all personal data will be considered a target or confidential. There are some approaches to anonymization that will try to transform confidential data, but this can have a very negative impact on data utility, as this is the information where there is a lot to learn.

Nonpersonal data
: Attributes that are not about the data subjects, such as machine data, and therefore not personal in nature. It's worth classifying nonpersonal data as it is sometimes mixed with personal data and therefore incorrectly classified as target data. However, in the context of device data, for example, it's worth separating this out (sometimes both literally and figuratively) from personal data. You are likely to want to better protect personal data in your care, given the potential impacts on trust and regulatory oversight. The nonpersonal data will still be of value, however, for analytical purposes. You just won't need all the auditing and oversight for it.

Probing Questions to Understand Data and Data Subjects

- Describe the parameters or criteria for individuals being included in the database.
  - Who are the data subjects and does any information collected about them include other individuals (e.g., relatives or household members)?
  - Where are the data subjects from, i.e., from what jurisdictions or geographic locations will data be collected?
- Describe the structure and properties of the data.
  - Wants versus needs for analysis and research using the data collected?
  - Expected data retention period?
  - Do the attributes collected support the purpose for collection and processing?
  - Is the data highly detailed, is it highly sensitive and personal in nature?
  - What is the potential injury to individuals from an inappropriate processing of the data?
  - What is the appropriateness of approval by data subjects for disclosing the data?
- Describe the identifiability of attributes, including from inferences, and what may not actually constitute personal data.
  - What is directly identifying versus indirectly identifying?
  - What is confidential, or a target, besides the identifiable data?
  - What is nonpersonal, i.e., not about a data subject?

Our focus in this chapter is on the project definition stage, namely requirements gathering and defining generic elements of a system architecture as it relates to privacy. Concept definition was captured in the Safe Projects of Chapter 3. We can use this knowledge in assisting with the transition from privacy to secondary uses of

data, which moves us from privacy requirements gathering to privacy design and development.

From Primary to Secondary Use

Now that we have scoped out our project with various privacy considerations, but have also dug into specifics related to identifiability, we're ready to plan out options for building an anonymization pipeline for secondary uses. We've touched on the differences between primary and secondary use in previous chapters. But in our experience this bears repeating.

Primary purpose
: When you offer a service, people have expectations about what data you need to collect to effectively provide the service, and they have expectations that the collected data will only be used for the direct purpose of providing that service. It's really as simply as that: a primary purpose is the main reason for the service. It defines the minimum data needed to offer the service, and the way the collected data should be used to provide that same service. You can use that data for those direct primary purposes, but not for anything else.

Secondary purpose
: Everything that is not a primary purpose is a secondary purpose. Or, put differently, secondary purposes are the indirect uses of data that were collected for a primary purpose. Some may be mandatory (e.g., reporting to law enforcement), whereas some may be permitted (e.g., for the benefit of society). *Building* analytical models from data collected from several data subjects is, for example, generally considered a secondary purpose, whereas *applying* already built analytical models to subject data for the direct purpose of delivering an expected service to that individual is a primary purpose. Reducing identifiablity is mostly applied to secondary purposes.

There are different ways to parse data from the primary to secondary purposes of collecting identified data:

- A system that operates on top of identified data, providing a primary use (a form of access control through the use of pseudonymized data)
- An analytics engine, although it might be better to have such a system operate on top of pseudonymized data
- A separate pipeline that does not affect primary use

We'll discuss pseudonymized data in Chapter 5, when we look at how direct identifiers are removed or replaced with pseudonyms. Most anonymization pipelines will start from the production environment only for the purposes of extracting data, and will not operate directly on identified data. The last thing we want is to impact a primary use of data (i.e., we don't want to impact the services provided to data subjects), or to have a leak of direct identifiers (the worst kind).

Since we will be anonymizing data starting from identified data, we'll consider direct identifiers and indirect identifiers. We separate these out because the tools we use are different and, as you will remember from previous chapters, indirect identifiers are where the magic happens in terms of measuring identifiability.

We'll also work through use cases that either start specifically with identified data, or involve identified data in some way. There's a mixed bag of complications to work through, such as controlled re-identifications (?!), mixing anonymized data with identified data, or anonymized outputs with identified data. If everything in this space was easy, we wouldn't have written a book! Hopefully, this highlights the importance of the project-definition phase we just worked through at the start of this chapter.

Dealing with Direct Identifiers

Ridding yourself of direct identifiers is the first (but far from only) step to producing anonymized data. It is far from sufficient, but in most cases you only need a linking variable to keep records and data sources connected so that you know what data belongs to what data subject (also known as *referential integrity*). This is why we will push this discussion to the chapter on collecting pseudonymized data, even though in many cases an agent of the data custodian may be engaged to produce anonymized data. However, if we are building a system from scratch, we really would prefer to anonymize from the pseudonymized data first.

But there are two use cases we need to highlight when there is a need to create realistic-looking data from direct identifiers. Wait, what?! Rest assured that we will replace the direct identifiers with fake data, but that data should represent the variety of data originally collected.

Realistic direct identifiers

A very common use case for producing anonymized data is to conduct functional and performance tests of software. Organizations developing applications that process personal data need to get that data from production environments, but this data must be anonymized before being shared with a testing group. Not only is this a secondary purpose (i.e., it is not for the purpose of delivering the service that a data

recipient expects when the data was first collected), more often than not the data environment of the test group has fewer mitigating controls in place to protect data.

Another, although less common, use case is a design jam or hackathon, in which the use cases may include writing apps or software that would otherwise use identified data when deployed. This is actually very similar to the software testing use case, although it starts from a slightly different point of motivation. The concerns are similar, though, and perhaps even more extreme depending on the circumstances in which data will be shared (for example, participants may be able to copy data to personal computers, and maintain copies at the end of the exercise).

The reason we put these use cases here, in a chapter about collecting identified data rather than pseudonymized data, is that we actually need (masked) direct identifiers to produce realistic-looking data. If the use cases envisioned are not focused on analytics, our objective will only be to ensure that properties of the data, namely data quality, are similar to allow for robust testing of applications.

So, if the names collected were stored in a 256-character string, we will want to respect that and include names there that are of a similar length to the original. We wouldn't match name length between the identified and anonymized data, as that could leak information (especially for names of rare lengths, such as very short or very long names). But somewhere in those names we will need something similar.

Masking of this sort has to be done correctly, as we do not want to leak any identifying information. One common way to break a privacy-preserving scheme is a frequency attack, in which the frequency of occurrence is used to extract information from a system or even reverse engineer results. The length of names would be one example. The distribution of name length could be used to match against external dictionaries of names by country to learn where the data was collected, or to find min/max lengths that narrow down possible names.

Dealing with Indirect Identifiers

Ridding yourself of indirect identifiers, in the same fashion as direct identifiers, would mean eliminating all risk (sounds good!) as well as all analytic utility from data rendered anonymous (oh my, that's terrible!). We've described the methods of measuring identifiability in a previous book.[3] And we've walked you through the basic concepts of measuring identifiability in Chapter 2. No matter which technological approach we use, these concepts will apply.

3 El Emam and Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*.

Rather than removing the indirect identifiers, we will transform the data/outputs to ensure the level of identifiability achieves a defensible threshold used to provide reasonable assurance that data is nonidentifiable. But we've already provided a framework for doing this in Chapter 3.

The Five Safes, operationalized through risk-based anonymization, are both a governance framework and the basis for evaluating identifiability in the context of sharing data. That's because changes to any one of the Safes can change our assessment of identifiability. They are intimately linked! Consider all the factors that affect the data-sharing context, shown in Figure 4-3.

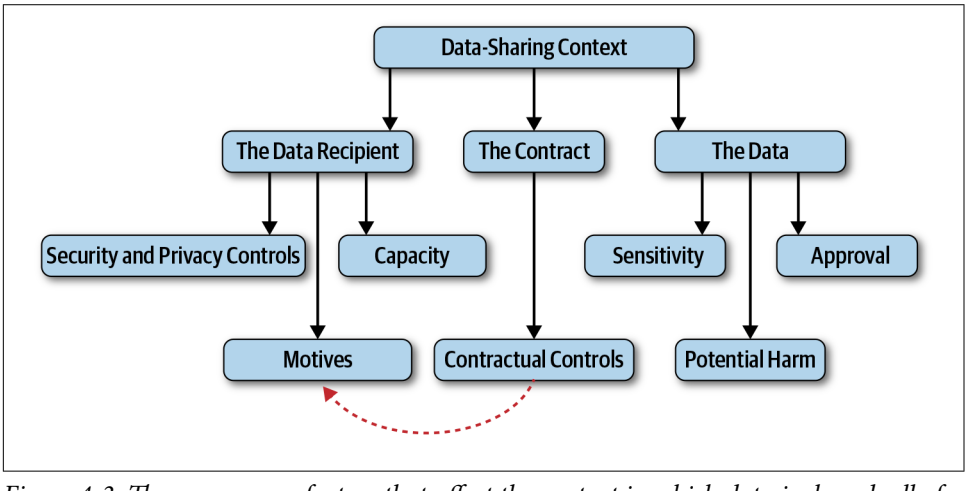


Figure 4-3. There are many factors that affect the context in which data is shared, all of which should be factored into a rigorous assessment of identifiability.

We will still transform data to achieve the defined risk tolerance, as determined from our Safe Outputs. But we are saving this for the chapter on pseudonymized data, ultimately because our current chapter is focused on working with identified data. An anonymization layer should be applied to pseudonymized data whenever possible. Secondary purposes from the original data collection should not operate directly from identified data that is in a production environment.

Having considered different types of identifiers, we can consider how we work with both identified and anonymized data, starting with how we produce anonymized data from identified data.

From Identified to Anonymized

The subject of anonymizing data may seem straightforward, in the sense that you either do it, or you don't. We will consider anonymizing for an external data recipient, who may anonymize the data, and circumstances around re-identifying for legitimate purposes. As previously described in Chapter 1, we use the term "shared" broadly to mean sharing a copy of, or sharing access to, data/outputs.

- Sharing a copy of data means that we assess identifiability when the anonymized data and outputs will be managed by another group. So it's the recipient's environment for the data that is being assessed (the Safe Settings at the recipient site), since that's where the data will be used.
- Sharing access to data means that we assess identifiability when the anonymized data/outputs will be managed by the data custodian, with controls around access by data recipients. In this case it's the custodian's environment for hosting the data that is assessed (the Safe Settings are always evaluated where the data will be hosted and used).

Anonymization should be separated from a production environment in which the primary purposes for data collection are carried out, regardless of whether you're sharing a copy or access to data and outputs. The last thing anyone wants is a failure in the anonymization to affect primary use, or a security incident in this environment. Rather, split these up. Either pipe the data out of the production environment and apply anonymization in this pipe, or pipe it into another production environment in which the anonymization will take place. In the latter case, the experts doing the anonymization will need permission to access identified (or preferably pseudonymized) data.

Once anonymized, data and outputs can be shared with the data recipients. The easy version of this is shown in Figure 4-4, in which the data recipients are external to the organization. We'll get into more complicated pipelines in subsequent chapters.

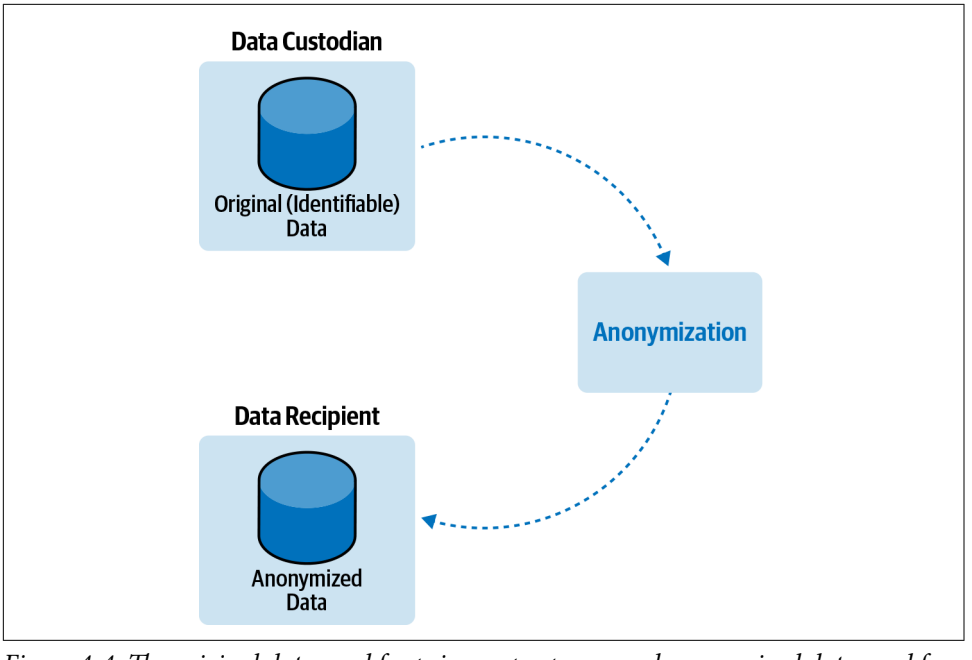


Figure 4-4. The original data used for primary purposes, and anonymized data used for secondary purposes, are managed by separate legal entities.

Data (anonymization) processors

Once data has been removed from the production environment, if it's not anonymized in the pipeline itself (through the use of automated anonymization tools, be they transforming data or outputs), it will need to be anonymized somewhere. In some cases, this is done by a data processor, an agent acting on behalf of the data custodian, and the appropriate agreements will need to be in place to ensure they have legal authority to work with (process) personal data. This can also be thought of as a pipeline, with personal data going to the processor, and anonymized data/outputs coming from the processor.

Data processing agreements are used to set up a legal relationship between the data custodian (the controller) and the data processor. The processor essentially becomes an extension of the data custodian, taking on the same responsibilities for a specified processing activity using personal data. They have no *more* rights than the custodian, but they do have requirements in that they are processing personal data. When the relationship ends, so does any use of that personal data, as it needs to be destroyed by the processor. These agreements should also specify if/how anonymized data/outputs (derived from personal data) may be used.

Tools and some training can certainly provide the means for a data custodian to anonymize data. The reason for using a data processor to anonymize data is that the expertise may not (yet) exist in-house, or anonymization may be a rare occurrence that doesn't make the business case for the cost of training and certification required to anonymize data. Or the data custodian may simply want someone else to take responsibility for both anonymization and the sharing mechanism that is decided on (including managing the feeds to different organizations).

Controlled re-identification

Imagine that a data recipient learns something of interest from the data regarding an anonymized data subject. This could be something that will affect treatment or care of a patient, fraudulent activities, or any number of things that are learned from the confidential or target data. These insights could be shared with the data custodian, who may then have a desire or need to re-identify the anonymized data. For example, the data custodian may have kept a key to a pseudonymized linking variable that would allow them to tie those specific insights back to the original data subject.

A controlled re-identification would need to be compatible with the original purpose for which the identified data was originally collected, or some form of legally permissible secondary use. It could only be done by the data custodian (who already has the original identified data), in a secure environment, by individuals with permission to access identified data.

Although a reasonableness argument is normally included in privacy laws and regulations when describing identifiability, guidance from some regulators has suggested that anonymization should be irreversible. Guidance is not law, and court rulings have described the reasonabless found in laws and regulations. However, that guidance does set expectations of what those regulators are looking for, and courts may turn to guidance if they feel it is relevant and reasonable. You can therefore consider controlled re-identifications to be a business risk in some jurisdictions, and decide how important it is to maintain the ability to reverse a pseudonym or token in your use cases, with the appropriate legal basis to support that activity.

Another, perhaps less controversial option, would be to share the analysis that led to the results of interest, in other words, the statistical methods that could be used by the data custodian on the identified data to yield the same outputs. This may not always be possible if those statistical methods, including AI/ML algorithms, are proprietary.

We started by considering how we would share anonymized data with *external* data recipients. And from that arose several considerations, namely the use of anonymization processors and controlled re-identifications. Now we consider a slightly more complex use case, in which we need to share data with *internal* data recipients.

Mixing Identified with Anonymized

Imagine a data custodian, such as an academic medical institution, that wants to share anonymized health data, collected from providing medical care to patients, with internal researchers. These researchers are not treating the patients in the health data, and the envisioned purposes are not in relation to the direct treatment of those patients. In other words they are considering secondary purposes only. This means the same organization will have identified data, used for treating patients, and anonymized data, used for research.

Regulators want to encourage the *responsible* use of data, to drive efficiencies and innovation. But some struggle with scenarios in which there is identified data used for primary purposes on one side of a Chinese wall and its anonymized counterpart used for secondary purposes on the other side of the Chinese wall. It would seem the organization has the ability to step from one side of the wall to the other whenever they please. The concern is with the separation between the identified and anonymized data, since in theory it would be possible to mix the two and render the anonymized data identified.

In theory, it would be much easier to re-identify the anonymized data since the same organization has the identified data. However, in practice, the organization has no need or desire to re-identify when it has identified data. The motives are simply not there at an organizational level, and the analogy of having the key doesn't really hold since the identified data is ever present and being used for those primary purposes. The separation between identified and anonymized data does, however, need to be real, demonstrable, and well documented with auditable proof and enforcement.

Functionally anonymized

There are advantages to the internal reuse of anonymized data, since the data custodian can in practice have more direct oversight of the controls and uses. That isn't to say there aren't risks, since making the case that data held within an organizational function is anonymized, and will remain so, means there needs to be a true separation between the anonymized and the identified data. And regulators recognize that anonymization is privacy preserving, above simply removing direct identifiers (i.e., pseudonymization).

There are obvious desires to drive efficiencies and innovate with data, while maintaining primary uses generally. But there are also (nonprivacy) regulatory requirements to maintain historical records with identified data. This serves to emphasize the point that there are practical reasons to maintain identified data and provide ways for the data custodian to serve both primary and secondary uses. For example:

- Banks are required to maintain certain records for designated periods of time, and the designated periods can vary by type of information. This can include information needed to reconstruct transactions, loan information, and evidence of compliance for any disclosures or actions taken regarding loans, savings, and fund transfers. The required retention periods can span multiple years.
- Government departments and public bodies need to comply with a variety of laws and regulations depending on the primary uses they serve. Types of data vary greatly, and can include information about civil rights, disabilities, employment, health, social services, etc. There may be requirements to maintain information for the purposes of reconstructing transactions or supporting decision making. Again, this will vary greatly based on the primary uses they serve.
- Sponsors of clinical trials are required to retain trial records for multiple years after the completion of a trial. This is to ensure accurate reporting, interpretation, and verification. Trials that are used for "marketing authorization" (the process of evaluating and granting a license for a product to be sold) have much longer retention periods, and some information needs to be retained for as long as the product is authorized.

We give data anonymized in this scenario, where identified and anonymized exist under the same legal entity, a special name:

Functionally anonymized
: Data which is transformed and protected with strong privacy, security, and contractual controls in place to ensure that identifiability is sufficiently low, within an organizational function that does not have access to the keys or additional data needed to re-identify.

There are situations in which it may be desirable, if not necessary, for an organization to work with functionally anonymized data while maintaining the keys to reverse the process (or at least the pseudonyms). The rules of engagement would need to follow the process of controlled re-identification described previously. Namely, intentional re-identification by the data custodian needs to be for a compatible purpose or a permissible secondary purpose.

Five Safes as an information barrier

To engender public and regulatory trust, we need to ensure that there is a clear separation between functionally anonymized and identified data. And this is especially true when the same organization is mixing both, as shown in Figure 4-5. We don't want anyone to think the data custodian is having their cake and eating it too.

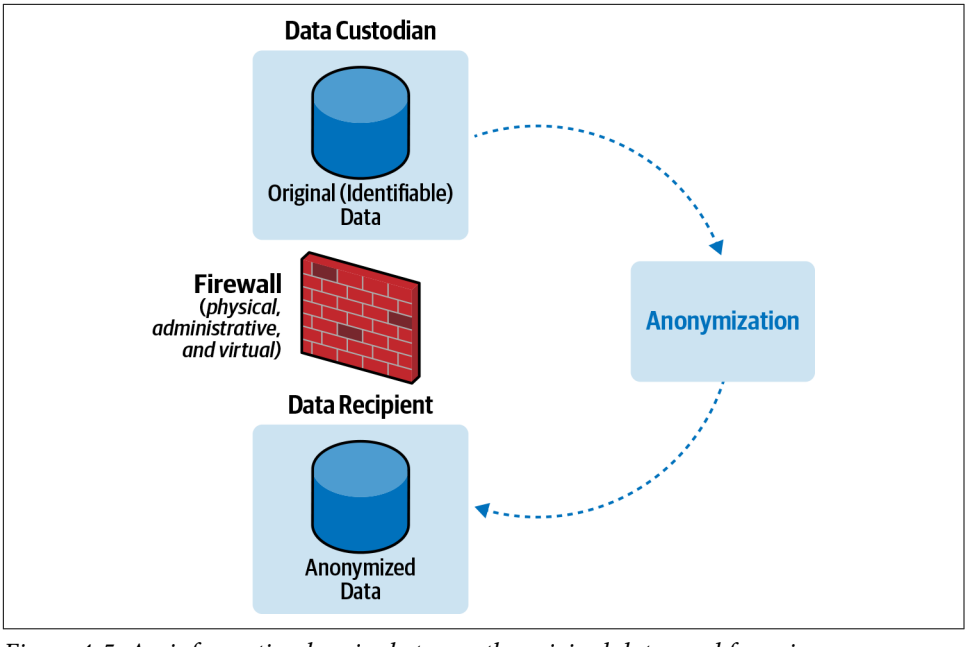


Figure 4-5. An information barrier between the original data used for primary purposes and functionally anonymized data used for secondary purposes.

Let's consider the Five Safes we presented in Chapter 3 and see how we can engender that trust:

Safe Projects
: A clear separation of purposes, and an ethics review, would certainly help set the project of creating safe use on the right path.

Safe People
: Consider that our data recipients work for the same organization. There will need to be a clear separation of those who work on the functionally anonymized data from those who work with identified data. Otherwise, the risk of them inadvertently recognizing someone would be much higher.

Safe Settings
: The data environment for the functionally anonymized data will need to be independent of the identified data, with no mixing. This implies that the data recipients, including the administrators, not have access to identified data, and even that the physical access to the functionally anonymized data be in a separate area than where other employees access identified data (i.e., to avoid accidentally looking over someone's shoulder).

Safe Data
: With the Safe People and Safe Settings clearly defined, so that we separate functionally anonymized from identified data, the usual threat modeling can take place to eliminate residual risk.

Safe Outputs
: Risk tolerance would be the same, but there would be little to no room for excuses for misusing outputs. The trust of service users would be seriously eroded if any misuse impacted those same users from which the data was derived.

This may seem like overkill to some, and it may seem to fly in the face of our framework to evaluate how safe these constraints are. But do not take this lightly, as it is a serious concern of regulators. The use of data can have many benefits, and this is recognized, but trust can only be built and maintained by having clear boundaries.

Now we can summarize the above considerations into three constraints for creating a defensible information barrier between identified and functionally anonymized data:

- Different people
- In different physical and virtual areas
- Supported by different system administrators

Some would go so far as to recommend creating separate legal entities as an option when there is a chance of mixing identified data with functionally anonymized data, to limit regulatory concerns and oversight. We have worked with organizations that have spun off new companies that would work only from anonymized data they would provide. That should tell you how serious a subject this is, but also the value that anonymized data can have (so much so that a company can turn a profit from the insights it will generate from said data, while reducing regulatory risks to ensure those profits are protected). That's serious business.

As if anonymizing data wasn't hard enough, we've now seen some of the many complicating factors to building a few, somewhat straightforward, pipelines. We've gone from identified to anonymized data, for external or internal data recipients, and considered how we can build the appropriate conditions to ensure we maintain appropriate oversight around the anonymized data and how it is used. We treated these data assets as distinct entities, completely seperate from one another. But what happens when identified and anonymized data overlap in some way?

Applying Anonymized to Identified

Regardless of the provenance of the anonymized data, there will be circumstances in which you may want to mix it with identified data, or apply model outputs from the anonymized data to the identified data. This will obviously raise eyebrows, since it may seem like a form of re-identification (even if that's not the case!), so let's consider some possibilities. To do this in a meaningful way, we need to compare populations between the anonymized data and the identified data, as shown in Figure 4-6. We assume the defined population (based on identifiability) is the same for both. We've ordered these from least to most concerning.

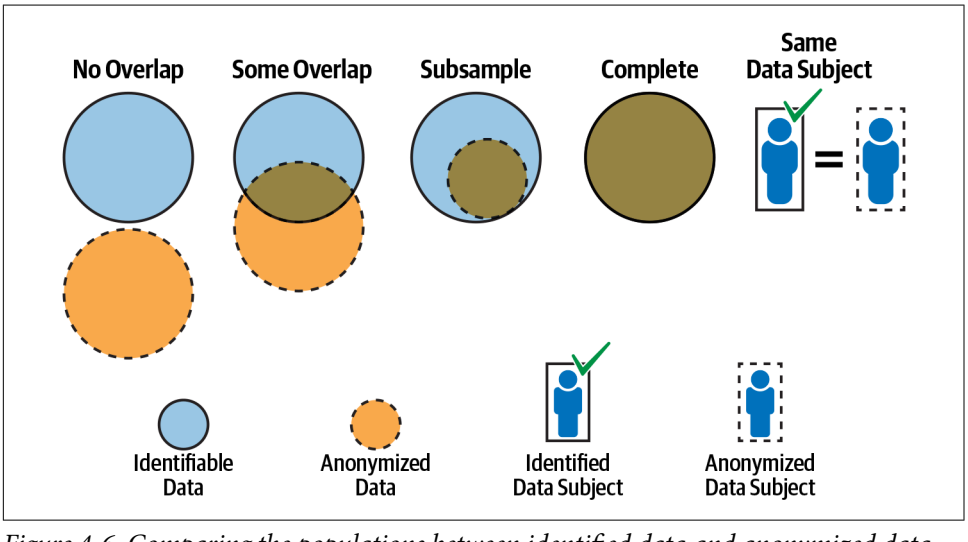


Figure 4-6. Comparing the populations between identified data and anonymized data will help us work through possible privacy pitfalls.

No overlap in populations

In this case there are no concerns, as the insights from an anonymized group are being applied to an entirely different population. You can imagine having a consumer group in one sector that provides insights into buying patterns that can be applied elsewhere. There are no risks of re-identification when the population groups don't overlap, but there are still interesting things to learn about behaviors and outcomes.

Some overlap in populations

Once we start mixing anonymized with identified data when there are data subjects that overlap, concerns may be raised about potential re-identification. In this case, however, the overlap is uncertain. We don't know which data subjects overlap, just that they share some identifiable features in common, but these have already been managed in terms of clustering based on identifiability. There

would be considerable uncertainty in attempting to re-identify, depending on the extent of the overlap between the two populations.

Subsample of populations
: When the anonymized data is a subset of the identified data, the adversary will know there are matching data subjects, but not which ones. The same is true when the identified data is a subset of the anonymized data. There is less uncertainty than the previous case of overlapping populations. Concerns would arise if there are any overlapping nonidentifiable attributes, as these would now represent a potential risk to matching between anonymized and identified, especially as the sample size increases.

Complete population
: At this point there is a significant risk of attribute disclosure, i.e., associating sensitive information to a group of individuals. The datasets must represent the defined population in its entirety, otherwise any sampling would prevent an adversary from knowing of this overlap. This is also a perfect example of a prosecutor attack (see "Safe Data" on page 54 for a refresher), since it's known who's in the anonymized data (although not which records belong to them). That means that identifiability is definitely higher than in the previous examples, and you may want to consider the ethics of how these attributions will be used.

Exact data subject
: This would occur if a linking variable was used to match anonymized data to the identified data subject. (Linking can be done in a privacy-preserving way, but that's a different subject.) This would enhance the identified profile, but would likely raise significant concerns since the anonymized data is now re-identified, possibly by someone other than the original data custodian of the personal data that was anonymized.

The overlapping and subsample cases are somewhat common when you consider census data, even in aggregate form, and inferences. Outputs on specific geographic regions can be applied to identified data to enhance analytical modeling. For example, knowing that 80% of people in a region love chocolate cake would certainly be helpful if you were modeling consumption patterns. But there's uncertainty since the populations don't perfectly match, which means at best we can make inferences.

To reduce concerns over privacy risks and impacts, it would be best if models, outputs, and insights were applied to anonymized data rather than the identified data themselves. This isn't strictly necessary, but would certainly be easier to explain to regulators. At the very least, we would advise avoiding potential attribute disclosures through the use of sampling or subsampling, and an impact assessment or ethics review.

Although the above considerations certainly seem to complicate matters, it's actually the overlap between identified and anonymized data that creates these complications. It's important to understand potential risks so that they can be mitigated, and to explain risks and mitigations to regulators. As mentioned previously, details of this nature need to be documented to ensure approaches to working with both types of data are auditable and defensible.

Final Thoughts

We started with the collection of identified data, and the concerns and considerations in designing privacy into systems that manage personal data. One of the most effective privacy tools is to disassociate data subjects from the data, or reduce identifiability, wherever and whenever possible. This chapter was meant to help you work through the project-definition phase, collecting as many requirements and concerns as possible while thinking through various use cases, starting with the collection of identified data.

There are many resources that work through the phases of privacy engineering, and this chapter was not intended to cover all of them. Our goal was to set you on the right path to building anonymization pipelines. For that we don't need to consider every aspect of privacy related to personal data, since our goal is to eliminate identities from data in a manner that is comprehensive, repeatable, and defensible.

As we described, you are likely to create a pipeline from the identified data into a new feed of pseudonymized data (since building anything that operates directly on top of the identified data would put the primary data collection and services at risk). Since this was already a hefty chapter, that was our excuse for separating the discussion of anonymization technologies from the next chapter, in which we work from the perspective of collecting pseudonymized data.

CHAPTER 5

Pseudonymized Data

Once the *identified* is removed from data, including people's names, addresses, and other unique identifiers, you are left with pseudonymized data. This term was popularized with the introduction of privacy regulations in Europe. Technically speaking, the directly identifying information does not need to be replaced with a pseudonym. It could just as well be replaced with a token or fake data or even suppressed entirely. The legal term pseudonymization simply means that direct identifiers have been removed in some way, as a data protection mechanism. And any additional information required to re-identify is kept separate and is subject to technical and administrative (or organizational) controls.

The most critical point to keep in mind is that pseudonymized data is still personal information. The untouched indirect identifiers that remain in pseudonymized data are known to pose a potential re-identification risk. Pseudonymization was introduced recognizing that identifiability exists on a spectrum, and that there are benefits to encouraging the processing of *less* identifiable data. However, we will also look at potential technologies that sit on top of pseudonymized data for secondary analyses, and where this fits in the anonymization landscape.

Data Protection and Legal Authority

Before we discuss scenarios in which pseudonymized data is further transformed into less identifiable information, it's worth spending some time considering the benefits of and opportunities for pseudonymized data. Even if pseudonymized data is still categorized as personal information, removing names and other direct identifiers will increase privacy by eliminating some immediate risks from immediately knowing who data is about. Seeing something very personal about someone, perhaps something sensitive such as health information, can help solidify their name or address in one's memory.

Imagine an analyst working with identified data, and they learn someone has had issues with a sexually transmitted disease (STD). If they don't already know this person, they could nonetheless run into them in the future and a lightbulb could go off when they realize they know this person and their sensitive diagnosis of an STD. How awkward! Even if the analyst doesn't know this person, and will never meet them, there's a creepiness factor to knowing someone out there is learning these very personal things about you. Especially when your name, address, or other direct identifiers are tied to that information. Enter pseudonymized data.

Pseudonymized Services

Pseudonymization is not perfect, but there are legitimate use cases for working with personal data. The analyst above may have been working with a hospital that is treating the person described. Whether they still suffer from an STD or not, the analyst may need to know this information for other treatments they are contemplating for a new diagnosis. We can assume the analyst has legal authority to work with personal data, since they are working on treating their ailments. But the question is whether they need to know the person's direct identifiers to run their analyses.

To be clear, it may seem disturbing to the nonhealth audience to be discussing sensitive, and perhaps "disgusting," health diagnoses. But that's the point. Health information is about real people, with real health problems. These are legitimate health problems and nothing to be embarrassed about in the context of treating health issues. But the reality is that people will be embarrassed, and there is the risk of stigma, so they will want this information kept private. Put yourself in their shoes. Whether you are open about your health information or not, the point is that we have societal norms that dictate how we treat personal and sensitive information. The purpose of this book is to arm you with tools to help respect the privacy of individuals based on those societal norms (that are largely codified by privacy laws and regulations), regardless of whether or not it is health data.

The term *pseudonymized care* has been used in ISO 25237 (*Health informatics—Pseudonymization*)[1] to describe such scenarios. That is, care is provided without having direct identifiers available, as shown in Figure 5-1. This could be our analyst, running their statistical algorithms and sending the results back to the treating physician. Or a lab technician processing biological samples and sending the results back. More likely, the results would be logged in a database, and those working from the patient's data would be alerted of the new results. Access controls would ensure that only those

1 ISO 25237, *Health Informatics—Pseudonymization* (2017), *https://oreil.ly/WfA24*.

who need access to direct identifiers would see them when they pull up the patient's file. So the analyst or lab technician sees requests and the results they produce without direct identifiers, but others may see a great deal more based on their access rights.

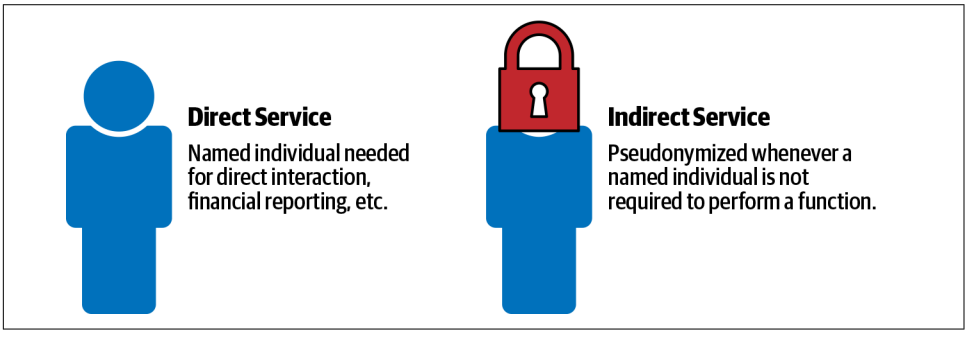


Figure 5-1. Pseudonymized data can be used whenever the data subject's name (and other directly identifying information) is not required.

Only people meeting with the patient face to face—treating physicians, nurses, administrative staff—really need to see the patient's name. (Otherwise, conversations would be awkward, referring to patients by their pseudonym! "Hello patient 666. You seem to be at a crossroads.") Of course there are other use cases in which direct identifiers will be needed, such as invoicing and billing. As much as we would like to, unfortunately we can't provide you with a list of all the possible use cases for pseudonymized data. This will require an analysis of data flows, legal boundaries, privacy policies, and working through (and justifying) business processes and user interactions. Basically all the elements of requirements engineering.

Legal Authority

Let's consider another example of a use case for pseudonymized data, one in which there are secondary purposes, but also legal authority for those secondary purposes. That is, the reasons for processing the data are different from the reasons data was originally collected. Normally, when we think of secondary uses of data, we jump to the conclusion that anonymization will be required. But that's not always the case, which is why we need to consider data flows and legal boundaries (which we explored in "Safe Projects" on page 46). In this case we review the circumstances around a maternal-child registry that has legal authority to collect and process data even though it's for secondary purposes.

The Better Outcomes Registry & Network (BORN) of Ontario integrates the data for all hospital and home births in the province.[2] It was created for improving the provision of health care, and also for research. This case study was presented in our other book, *Anonymizing Health Data*, but in the context of getting data *out* of the registry and into the hands of researchers. In that case the data is anonymized, since it leaves the protected environment of BORN. Our interest in this section is focused on another aspect, that is, the pseudonymized collection of data related to births.

BORN is a prescribed registry under Ontario's health privacy legislation and is therefore allowed to collect relevant personal data. The data is collected through a number of mechanisms, including manual data entry, and automated extraction and uploads from health record systems. The registry includes information about the infants' and mothers' health. Several sources contribute to the data collected by BORN:

- Prenatal screening labs
- Hospitals (labor, birth, and early newborn care information, including NICU admissions)
- Midwifery groups (labor, birth, and early newborn care information)
- Specialized antenatal clinics (information about congenital anomalies)
- Newborn screening labs
- Prenatal screening and newborn screening follow-up clinics
- Fertility clinics

The personal data collected by BORN is, however, pseudonymous. The registry is cross-sectional by design, in that we cannot trace mothers over time. If a mother has a baby in 2009 and another in 2011, it's simply not possible to know with any certainty that it is the same mother. This kind of data is quite common in registries and surveys. There is therefore no reason to collect the names of mothers, or medical identification numbers of any sort. The design of the registry intentionally enhances privacy by collecting pseudonymous information.

BORN has legal authority to collect the pseudonymous information it works with, and this legislation does require that safeguards be in place to ensure the data is appropriately managed through reasonable security and privacy practices. Researchers that request data must go through an approval process, and since the registry doesn't collect any direct identifiers, such as names or medical identification numbers, a researcher can only request indirect identifiers and clinical variables. Contractual obligations are agreed to with the researcher's organization, and the data received is anonymized for the identified purposes and in the context of the environment in

2 Details about BORN can be found from their website at *https://www.bornontario.ca/en/index.aspx*.

which the data will be used. But that's a different data flow than the pseudonymized collection we are presently focused on.

We have presented a couple of examples of pseudonymization that are health care specific, but the application is much broader. Hopefully this is clear. We use health care examples because the data is among the most sensitive that is collected and used, and also because we hope it will help the reader appreciate the importance of privacy-enhancing approaches to reducing identifiability. With the introduction of the legal interpretation of pseudonymization in the privacy regulations of Europe, its application is likely to find broader appeal.

Legitimate Interests

It's probably worth mentioning another area in which pseudonymization will support the secondary use of personal data. The term *legitimate interests* is from the privacy regulations of Europe, but conceptually it is applied in other jurisdictions as well. The basic idea is to provide a legal argument that justifies why data can be processed for another purpose than what it was originally collected for, when the processing is deemed "necessary" and will not outweigh the interests, rights, or freedoms of data subjects. The use of a weighing analogy is intentional here, because a balancing test is often described to consider the interests of the data custodian against the interests of the data subjects.

Avoid using the grounds of legitimate interests (or equivalent, depending on the relevant jurisdiction) if the processing will "surprise" people in some way. It needs to be something they can understand and will reasonably expect that they would agree to if it were explained to them. This is easier to do if the processing is for a current activity that is necessary for administration or compliance (e.g., governance obligations), or core to the business. Or if the processing will be of benefit in the very near future, again provided individuals would reasonably expect it. But keep in mind that you will also be required to prove the processing is necessary and cannot be done using less identifiable data.

Pseudonymization, in this context, is used to reduce the risks to the relevant privacy rights and freedoms of data subjects, and thereby support a case for relying on legitimate interests as a basis for processing. That isn't to say that pseudonymization is the only factor considered in a balancing test, but it is one aspect that can help make the case for the secondary use of data. Especially when the precise identity of an individual is not required for the processing, as in the case of the pseudonymized services described earlier.

Let's consider an example in which the processing of personal data is deemed necessary by the data custodian. In this case we have an organization that collects personal data to deliver a service of some kind to individuals (the data subjects). The organization wants to introduce a new premium feature. While they don't think testing with the personal data is exactly compatible with the original purposes of data collection, since it's an entirely new and somewhat unrelated feature, market testing has suggested that a reasonable percentage of their clients will pay extra for this new premium feature.

The organization has already tested the new feature on fake data, but they want to push it to a preproduction environment to ensure everything works as expected. They consider using personal data to be necessary because any errors could disrupt the service they provide to individuals, even those who don't sign up for the premium account (as much as they've tried to isolate it, they want to offer a free trial to all users). Whereas the nonpersonal data used was appropriate for building models and testing most use cases, they're worried there are edge cases they haven't thought of beforehand (the unknown unknowns), and these edge cases could disrupt the service.

Pseudonymization will help justify using personal data for the final, preproduction testing. They will have all their usual privacy and security controls in place for working with personal data. The engineers are a bit worried there may be elements from the direct identifiers that also need testing, so a masking tool will be used to re-create direct identifiers of the same length, complexity, and diversity found in the original data. Since the direct identifiers will be masked, the data is still considered pseudonymized, and privacy is enhanced even though personal data is being used.

We previously explained that we would assume direct identifiers were removed before we consider anonymization, which implies we're assuming the data is pseudonymized. Using pseudonymized data for secondary purposes, given legal authority or using a legitimate interests argument, is an exception. We need to move the needle along the identifiability spectrum beyond uniques, which is all we get from pseudonymization, to actual clustering of individuals based on their degree of identifiability.

A First Step to Anonymization

We've described pseudonymization as replacing or removing direct identifiers, leaving indirect identifiers in place (as well as confidential or target data, and nonpersonal data, depending on the use case). This is the first phase of reducing the identifiability of data on what we've described as a spectrum of identifiability. Now we want to introduce another point along this spectrum, beyond the basic

pseudonymization we've described but before anonymization. We will call this *strongly* pseudonymized data.[3]

The idea of strongly pseudonymized data was born out of the privacy regulations in Europe, although in truth it dates back to the earliest measures of disclosure risk through the concept of (sample and population) uniqueness (described in Chapter 2). In Europe this has been termed *singling out*, meaning that a data subject's information can be isolated as unique from the population in which the data about them was collected. This is a vulnerability we have already described, and forms the basis of the concerns with using pseudonymized data (and therefore treating it as personal).

Recall that we introduced two different classes of indirect identifiers: knowable to the public, and knowable to an acquaintance. An indirect identifier that is knowable to the public represents the greatest risk since, by definition, it's more broadly available to be used in a re-identification. We can therefore define our two classes of pseudonymization as follows:

Basic pseudonymization
: Direct identifiers are replaced or removed through masking, and any additional information required to re-identify is kept separate and is subject to technical and administrative (or organizational) controls.

Strong pseudonymization
: Direct identifiers are replaced or removed through masking, indirect identifiers that are knowable to the public are transformed to ensure that data subjects are not unique in the target population so that they can't be singled out, and any additional information required to re-identify is destroyed if required, or kept separate and is subject to technical and administrative (or organizational) controls.

If the additional information required to re-identify includes cryptographic keys, under basic pseudonymization the keys need to be stored and handled appropriately, and under strong pseudonymization the keys need to be destroyed if the legal basis for secondary use requires the process to be irreversible. This last point is subtle. We are transforming data to reduce identifiability. In some use cases, such as legitimate interests described earlier, strong pseudonymization can support the legal basis for processing even if keys are not destroyed (that is, the pseudonymization is reversible). If the intent is also to address certain data subject rights, such as access, rectification, and portability, however, the keys will need to be destroyed to make the argument

3 The term strong pseudonymization was coined in Mike Hintze and Khaled El Emam, "Comparing the Benefits of Pseudonymisation and Anonymisation Under the GDPR," *Journal of Data Protection & Privacy* 2, no. 2 (2018): 145-158, *https://oreil.ly/cA4zG*.

that it's not possible to meet these obligations (since data subjects can't be reasonably singled out).

Although the keys are destroyed with strong pseudonymization, this does not make the data anonymized. It is more difficult to associate identities to strongly pseudonymized data, in particular because indirect identifiers that are knowable to the public are transformed to ensure that data subjects are not unique in the defined population. In contrast, anonymization also includes indirect identifiers that are knowable to acquaintances.

The reason that only indirect identifiers that are knowable to the public are transformed, and only so that data subjects are not unique in the defined population, is due to a reasonableness argument. Essentially, the data custodian is not required to maintain or process additional information for the sole purpose of complying with certain data protection obligations. It is assumed that only the identifiers described in this process would be required to request certain obligations be met, otherwise the requirements to meet these requests would be overly burdensome to data subjects. And it's critical that the process of matching to identities be reliable and avoid misidentification.[4]

Table 5-1 describes obligations under Europe's privacy regulation. The legal analysis that produced these results is out of scope for us, but feel free to delve into the details of the paper if you would like to know more. The table demonstrates how strong pseudonymization has advantages over basic pseudonymization, since identifiability is reduced further and thereby enhances privacy further. Organizations get something in return for enhancing privacy beyond the very basic approach.

Table 5-1. Obligations under Europe's GDPR based on identifiability of data

Obligation	Identified	Basic pseudo	Strong pseudo	Anonymized
Notice to data subjects	Required	Required	Required	Not required
Legal basis for processing (legitimate interests, consent)	Required	Stronger case for legitimate interests	Much stronger case	Not required
Data subject rights (access, portability, rectification)	Required	Required	Not required	Not required
Right to erasure or right to be forgotten	Required	Required	May not be required	Not required
Basis for cross-border transfers	Required	Required	Required	Not required

4 This other point along the identifiability spectrum is also described in Mike Hintze, "Viewing the GDPR Through a De-Identification Lens: A Tool for Compliance, Clarification, and Consistency," *International Data Privacy Law* 8, no. 1 (February 1, 2018): 86–101, *https://oreil.ly/imgwp*.

Obligation	Identified	Basic pseudo	Strong pseudo	Anonymized
Data protection by design	Required	Partially met	Strengthens ability to meet requirement	Not required
Data security	Required	Partially met	Strengthens ability to meet requirement	Not required
Data breach notification	Required	Less likely to be required	Strengthens ability to meet requirement	Not required
Data retention limitations	Required	Required	Required	Not required
Documentation or recordkeeping obligations	Required	Required	Required	Not required
Vendor or subprocessor management	Required	Required	Required	Not required

The point of all this is that strong pseudonymization makes it harder than with basic pseudonymization to correctly attribute data to an individual in a target population. This enhances privacy further than basic pseudonymization, which can serve many benefits regardless of jurisdiction. What's good for data subjects, and society, can be good for organizations as well. However, let's not forget that we have attached the qualifier *strong* to a legal concept that is fundamentally about personal data, and so the result is still personal data.

Even though, on the surface, strong pseudonymization produces personal data, it requires eliminating population uniques so that data subjects are clustered into groups of two or more based on a subset of identifiable features. That's one small step for a data subject, one giant leap for the defined population. With this new point along the identifiability spectrum in hand, we can look again to secondary uses.

Revisiting Primary to Secondary Use

In "From Primary to Secondary Use" on page 74, we worked through some of the considerations in anonymizing data, starting from identified data. We did this in the context of moving toward secondary uses, since this full anonymization is typically applied in the context of secondary uses. We considered:

- Dealing with direct identifiers: the techniques and considerations in creating basic pseudonymized data
- Dealing with indirect identifiers: the techniques and considerations in creating anonymized data
- Identified and anonymized data: various considerations when we go from one to the other, or mix the two in one form or another

In dealing with indirect identifiers, we summarized elements of risk-based anonymization (summarized since we have already covered identifiability and the Five Safes in

previous chapters). This is where strong pseudonymization fits, since we are transforming indirect identifiers knowable to the public in order to remove the ability to single out individuals. In other words, strong pseudonymization intentionally moves the needle further along the spectrum of identifiability, progressively more privacy enhancing, but not necessarily to the point of being anonymized. Recall that with strong pseudonymization, indirect identifiers that are knowable to an acquaintance are not considered.

We will introduce pipelines that leverage the concept of strong pseudonymization to produce anonymized data. But this will be the subject of another chapter. Let's instead focus in this chapter on other ways we can leverage basic pseudonymization more generally.

Since pseudonymized data, be it basic or strong, is still personal information, the considerations we reviewed regarding identified and anonymized data in the previous section will still apply. In other words, we could replace identified with personal in that section, and the concerns would remain. That is, mixing personal with anonymized requires careful evaluation, but we've done this already, so we recommend going back to the previous chapter and section if it applies to you.

We can instead look at other tools and technologies that can be viewed through the lens of data protection, starting with different options for performing analytics on pseudonymized data and also considering generating fake data from pseudonymized data. We will also briefly consider biometric identifiers, as these pose unique challenges in terms of their complexity and uniqueness, as well as infrastructure challenges when attempting to scale.

Analytics Platforms

Research and analysis is the quintessential example of secondary uses in general. In most cases such secondary uses will require the data to be properly anonymized. Even when there is legal authority to use personal data, in all likelihood this data can and should be pseudonymized (at the very least). We can even go a step further and employ an analytics platform that can hide the underlying data, providing analytical tools without revealing the raw data on which the analytics are performed. Hiding the underlying data, however, doesn't mean the data is anonymous or that the results may not provide an avenue for identity disclosure.

As mentioned in a previous chapter, we are more likely to start from the production environment only for the purposes of extracting data and not operate directly on identified data. The last thing we want is to impact a primary use of data (i.e., we don't want to impact the services provided to data subjects), or to have a leak of direct identifiers—the worst kind. An analytics platform is no different, that is, it should

operate on pseudonymized data that is pulled from the identified data, but does not directly impact the identified data when using the platform. Analytical results could be pushed back into fields that complement the identified data, but it's unlikely they would directly alter, or operate on, identified data since what is analytically produced will be inferred.

Operating on top of personal data does pose one other challenge that we need to be aware of. Because the analytics are being derived from personal data, the system as a whole may not be considered anonymized in the usual sense. This is because the system cannot exist without the underlying personal data, since it must go back to the personal data in order to run the analytics. Depending on the use case, there may be other solutions worth exploring that avoid this potential issue, as we'll see in describing synthetic data in the next section, and anonymization pipelines in the next chapter.

Remote analysis

There is an entire branch of academic literature on the risks from analytics built from personal data.[5] Some have attempted to address these in what are called remote analysis systems, which allow someone outside an organization to run analytics without allowing them access to the underlying data itself. Risks that have been identified are managed by introducing limits in terms of what analytics are produced, to what degree of accuracy, and with what statistical tests or estimators (to verify model fit, for example).

Even for a simple linear regression (fitting a line to data so that the squared error between line and data is minimized) we may, for example, prohibit the use of overly small samples of data, and limit the outputs so that standard errors are never zero (or overly small) since that would indicate a perfect (or nearly perfect) fit, which could be misused. Say the regression is on sexual orientation for a small native population, and selection criteria are used to further narrow the population to a specific age group. Such manipulations could result in unwanted attributions, or even re-identifications when quasi-identifiers are involved, because there could be only one person who fits the selected profile.

In these more sophisticated systems, limits would normally be put in place to ensure there's always a sufficiently large cluster of similarly looking individuals in an analytical output. However, there are risks from repeatedly running analytics on data,

5 A good survey of these risks is provided in Christine M. O'Keefe and James O. Chipperfield, "A Summary of Attack Methods and Confidentiality Protection Measures for Fully Automated Remote Analysis Systems," *International Statistical Review* 81, no. 3 (December 1, 2013): 426–55, *https://oreil.ly/c5BKO*.

because the results can be differenced to extract smaller clusters, even to the point of singling out a single individual.

Let's take a very simple example using count queries. Say a query of how many people in a region within a certain economic range is run for ages 35 to 39, and there are 9 people. Another query is run for the same region and economic range for ages 35 to 40, and there are 10 people. We take the difference and we now have the economic range for a single individual of age 40. Now before we start cooking up solutions, we should keep in mind that this has been well researched already.

There are many ways to try and prevent these, and many other, types of attacks. The point for us is just to know that they exist, and therefore we have to assume that a standard analytics platform will not, in and of itself, protect personal data. Rather, we need a system that incorporates mechanisms to prevent identity disclosure.

Secure computation

An approach that is in some circumstances more suitable for accessing data for analytical purposes is secure computation, which involves computing analytic functions on encrypted data without the need to decrypt the underlying source data to run an analysis. Cryptographic primitives, or building blocks, to create secure computation protocols can come from homomorphic encryption, garbled circuits, secret sharing, or others, each with its own advantages and disadvantages.[6]

To keep things simple, we are grouping different technologies that serve more or less the same purpose under the term *secure computation*. The academic literature uses the term secure *multiparty* computation for protocols that securely and privately combine data from multiple sources. That's not necessarily what we're referring to in this section, although it certainly could be.

This approach also allows multiple sites to contribute data while providing strong privacy guarantees. This way the data can be pooled and contributors can compute analytic functions without either party knowing their inputs. Normally, in these circumstances, specific data sharing and analytic protocols are created to ensure the approach will be secure and provide strong privacy guarantees.

6 It's difficult to find a single resource that describes some of these methods in one place, however, some of the details are provided in Sophia Yakoubov et al., "A Survey of Cryptographic Approaches to Securing Big-Data Analytics in the Cloud," IEEE High Performance Extreme Computing Conference (HPEC) (2014): 1-6, *https://oreil.ly/aIQYw*.

When considering a risk-based approach to anonymization, secure computation can be formulated as "protected" pseudonymous data. The encryption ensures the "protection" in that no human is working directly with the data: instead they are only seeing the statistical results on indirect identifiers and confidential or target data. Although the source data is protected by encryption for secure computation, there would still be a need to address some of the concerns that exist with remote analysis systems in general, as previously described.

Analytics platforms using this technology are also possible, although limitations need to be imposed, similar to remote-analysis systems, in an attempt to address potential vulnerabilities from crossing analytic functions and from the outputs themselves. The additional overhead of encryption and computations over cryptographic building blocks can also pose significant computational overhead, although the technology continues to evolve with the introduction of specialized hardware. Secure computation is therefore better suited to scenarios in which there is a continuous, systematic collection and analysis of data, because the computations can be defined and optimized and then continuously applied.

Depending on the circumstances and system or user needs, analytics platforms may not be the right choice. It's often the case that analysts want to see the data they're working with to make sure they understand the data and results produced. Analytics platforms that aim to serve as data protection will not allow users to see the underlying data, and may block certain queries or functions that may reveal the underlying data. Another option may be the use of fake data.

Synthetic Data

Another approach worth exploring, especially in the face of the disclosure challenges associated with using an analytics platform, is that of fake data. We mentioned this earlier in the chapter, with the example of an organization that wants to introduce a new premium feature. Fake data can mean a few different things, but one approach that has existed for quite some time in the field of statistical disclosure control is building a model on data,[7] and using that model to produce new data, shown in Figure 5-2. The model itself could come from an academic paper, although in our context it's more likely to be built from pseudonymous data (the source in our image). Random points along the model are taken to generate fake data. As modeling has become more powerful and automated, and use cases have emerged, this idea of creating synthetic data has received more attention.

7 An important book on the topic is Jörg Drechsler, *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation* (*Lecture Notes in Statistics*) (New York: Springer-Verlag, 2011), *https://oreil.ly/8piBy*.

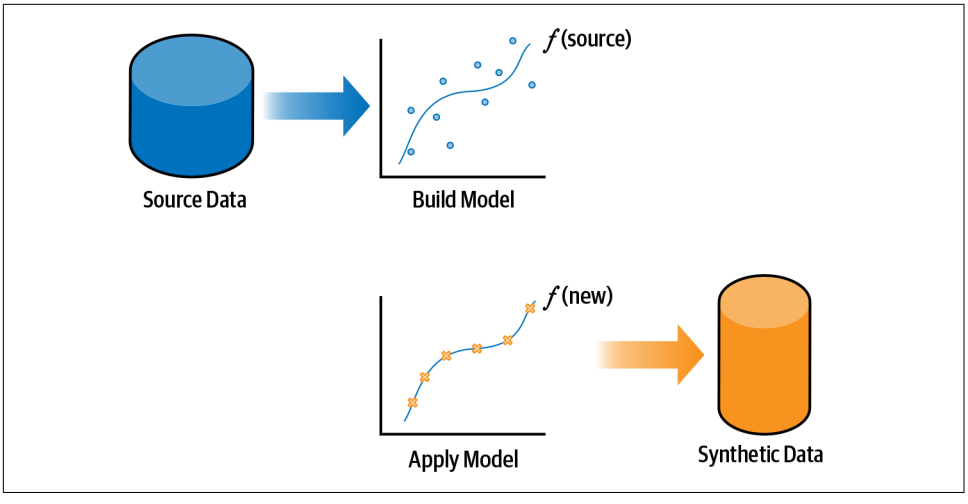


Figure 5-2. Synthetic data is generated from a model built from real data.

Synthetic data, as defined in statistical disclosure control, has not raised concerns regarding identity disclosure. Since the data, including indirect identifiers, is generated from a model, the thinking is that the synthetic data cannot reproduce the personal data from which the model is derived. The problem, however, is that the term synthetic data may be applied to systems that are different than what was originally intended, such as:

- Systems that "overfit" the original data and therefore can reproduce personal data. This may be especially true for automated systems, including advances in areas such as deep learning, where model fit is ignored or methods such as cross-validation or bootstrapping are not used to ensure the model is sufficiently general enough to produce data that varies enough from the original training data.
- Or, perhaps worse, systems that are not models in the sense in which the term synthetic data was originally conceived. For example, filters that remove outliers should not be considered as models to produce synthetic data. Filters can serve a different, albeit important, role in attempting to ensure models do not overfit the data. But the synthetic data should be derived from an appropriate model and not come from the original data.

Claims that a system produces "synthetic data" should be verified. A privacy model that considers identifiability is probably still needed to ensure the data is truly synthetic and does not reproduce personal data. Most methods described in the academic literature focus on attribute or inferential disclosure to address concerns of confidentiality. Since this is a relatively new area in practice, we make no assumptions regarding how synthetic data will be generated.

Therefore, the model used to produce synthetic data should only capture a partial structure of the original data. There is a balancing act between how accurate the model needs to be and how close the synthetic data is to the original data. It is the classic trade-off between data utility and data privacy. When done well, the synthetic data retains enough statistical properties of the original data and is nonidentifiable.[8]

Broadly speaking, there are three types of synthetic data:

Fully synthetic data
: All data, both indirectly identifying and nonidentifying, is synthetically generated from a model. This is implicitly what we've described above.

Partially synthetic data
: Only some data is generated from a model, traditionally the nonidentifying data that is deemed confidential; the rest retain the original value, or, in the case of identifying data, could be anonymized. If the identifying data is *not* anonymized, there will be limits to what can be learned in the case of a re-identification since the accompanying nonidentifying data will be synthetic. However, this would cause the privacy pro to raise eyebrows since, presumably, the synthetic data has a certain degree of utility that will still result in something being learned, and it would be hard to justify the use of this combination of identifiable and synthetic data to a data protection or privacy regulator if the intent is to produce anonymized data.

Hybrid synthetic data
: Some records are generated from a model, and some original records are mixed in. It's not clear how hybrid synthetic data would be considered nonpersonal. With original data in the mix, there will be a chance of identity disclosure. An argument for hiding in plain sight may be possible, but only if the proportion of original data is very low.

8 Summarized in Khaled El Emam and Richard Hoptroff, "The Synthetic Data Paradigm for Using and Sharing Data," Executive Update, *Data Analytics & Digital Technologies*, Cutter Consortium (2019), *https://oreil.ly/_-YNi*.

Synthetic data may be well suited when data utility is not a primary concern, when data needs to be shared very broadly, and where the overhead of signing agreements and setting up safe environments is high. This would apply to analyses that are more exploratory in nature, and when there's a need for rapid generation of nonpersonal data. Use cases in which the focus is not the analytic results may include the following (listed alphabetically):

Algorithm development
: Many of the recent advances in artificial intelligence and machine learning include algorithms that require training and testing on large pools of data that is sufficiently varied and realistic. These algorithms can be developed on synthetic data and then submitted to the data custodian for execution on real data; this brings the verified code to the data rather than sharing the data itself.

Data exploration
: Organizations that want to maximize the use of their data can make synthetic versions available for exploration and initial assessment by potential users; if the exploration yields positive results, users could then go through a process to obtain access to the anonymized data. When the desired analytics require only a handful of variables, it is also possible to use synthetic data as a proxy for real data and to produce more or less the same results.

Hackathons, data competitions, education, and training
: These require data that can be distributed widely with minimal demands on users. The focus is on ideation and experimentation with a large pool of users. That also means it would be suitable for teaching practical courses on data analysis and for software training.

Open data
: Sharing complex data publicly is challenging because of privacy concerns. A suitable approach to generating synthetic data may provide new opportunities to be explored in privacy-sensitive areas, where there is interest to share data for exploration that may not otherwise be available.

Proof-of-concept and technology evaluations
: To quickly evaluate whether a new technology works well in practice, realistic data is needed with minimal constraints.

Software testing
: Testing data-driven applications requires realistic data for functional and performance testing. Random data cannot replicate what will happen when a system goes into production. Since synthetic data is generated from a model, it would be suitable for testing in general, but it may not be sufficiently accurate for the final preproduction testing.

As we already warned, a privacy model for generating synthetic data is probably still needed to ensure the data is truly synthetic and does not reproduce personal data. We consider one such model in the next section, given its popularity in the academic literature.

Differential privacy

One model that has received a great deal of attention in the area of synthetic data is differential privacy because of the mathematical guarantees it provides.[9] In differential privacy, the "model" of privacy is based on the probability that we can tell whether or not an individual has contributed to a calculation, as shown in Figure 5-3. Take, for example, the total number of aerospace engineers with a specialization in nanotechnology of polymers. Let's say there are a dozen such engineers; the exact count is therefore 12. Take any one of these engineers out and the count is 11. In differential privacy, noise is added to hide the true count. Maybe we get a count of 11, maybe we get a count of 14; it all depends on the amount of noise deemed appropriate to ensure privacy yet still yield useful results.

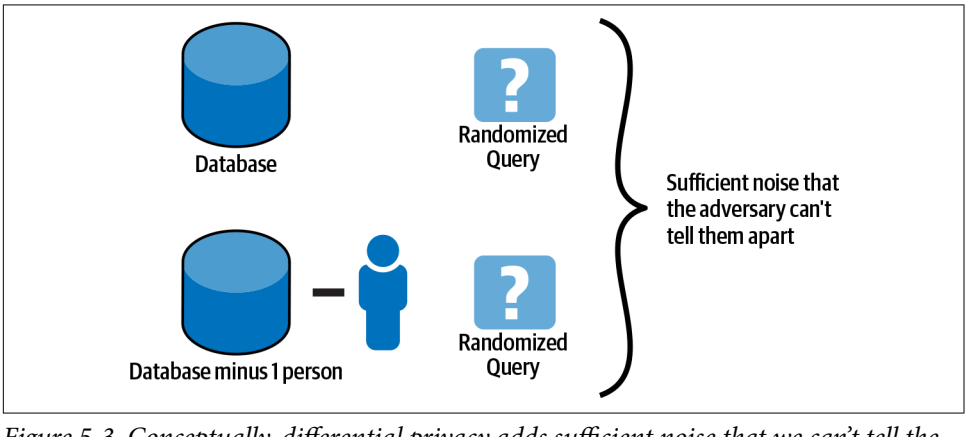


Figure 5-3. Conceptually, differential privacy adds sufficient noise that we can't tell the difference between the original database and any other version of it with one person removed.

How close we can tell whether or not an individual has contributed to a calculation is determined by an adjustable knob (the parameter ε) called the *privacy budget*. The privacy budget is a bound on a ratio of probabilities between the output distribution of a function operating over the source data, and the output distribution when any

9 For details see Cynthia Dwork and Aaron Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science* 9, no. 3–4 (August 11, 2014): 211–407, *https://oreil.ly/Cr25Z*.

one person is removed, the details of which are out of scope for our purposes. Turn the knob too high (very large ε) and very little noise is added (less privacy, more useful data); turn the knob too low (very small ε) and lots of noise is added (more privacy, less useful data). Each attempt to pull information from the data is also tracked and compared to the privacy budget, so that eventually you run out of budget and can't pull any additional information from the data.

Notice how this privacy model, at its core, says nothing about whether or not you can identify someone in the data. It does, however, provide protection against learning attributes about an individual (which doesn't require identifying them), which is why it has been classified under synthetic data generation. There's been some research to show that differential privacy can be equivalent to models that protect against identity disclosure (under a set of assumptions that we don't need concern ourselves with here). This requires combining privacy models, and is an active area of research.

There are practical considerations with systems that employ differential privacy. One consideration is the need to determine an appropriate privacy budget. The protection offered for the same privacy budget varies for different data, making it more challenging to set the parameter. Another consideration is tracking and exhausting the privacy budget. The guarantees provided by differential privacy become meaningless if the privacy budget is ignored or reset. In the case of synthetic data, this latter consideration may not be a concern for a single dataset, but it could be if there are multiple data releases from the same underlying population.

Keep in mind that meeting the definition of differential privacy is dependent on the privacy budget. Turn the knob too high (very large ε) and being "differentially private" becomes meaningless, because the probability you can tell whether or not an individual has contributed to a calculation is very close to one, meaning you can tell almost perfectly. In some cases, it can even have no impact on the calculation, providing meaningless guarantees. Turn the knob too low (very small ε) and you have no chance of telling whether or not an individual contributed to a calculation, but you also have little chance of getting useful results since there will be so much noise added to it—although this may be perfectly fine for synthetic data, depending on the use case.

There are in fact many variants to differential privacy, each an attempt to make it more useful in practice because the base definition makes the strong assumption that an adversary has infinite resources and time. These variants usually include additional parameters, further complicating how it may be used, especially as no standards exist for setting these parameters. But these could also be framed as opportunities.

With all of these options, analytics platforms or synthetic data, there are practical challenges regarding their deployment and use. There is one class of pseudonymized data that can prove quite challenging and demonstrates the need for practical solutions that can scale, which we consider next.

Biometric Identifiers

Facial images, finger and voice scans, and various other biomedical data is increasingly being considered to be biometric identifiers. Some of these are obvious because we can see the person from the data (as in the case of facial images), others because we have become accustomed to hearing about or using these technologies in our everyday lives (as in the case of finger and voice scans). Even DNA sequences are making a splash in the news because law enforcement has used samples to single out potential criminals and recently started using genealogical databases to find familial relations that point toward a suspect.[10]

Some biometric identifiers may be considered direct identifiers since they're readily available. We can easily see and compare an image of someone's face to verify their identity. We do this daily with driver's licenses and the like. There's also a plethora of online and public (or semi-public) databases of images with names that can be used for matching (e.g., social media).

Other biometric identifiers are less obvious, such as pieces of a genomic sequence, and can be thought of as pseudonymous data (which is still personal data). These would be biometric (nonfacial) images and signals that are not readily available and require specialized skills to decipher and match to databases. Such databases would have limited coverage and restricted access. This is changing, of course, and is at the heart of the "reasonable" in privacy laws and regulations. But we should not confuse researchers with law enforcement, as the latter have greater resources and access privileges in defined legal circumstances (although it can demonstrate opportunity and scale).

Novel technologies are emerging to take advantage of the uniqueness that exists in certain biomedical images and signals, notably as identifiers, which pose new challenges in the health care field that is seeking to bring new and more effective treatments to individuals. Innovation will be limited without access to relevant data to drive evidence-based insights and development. Privacy technologies will be needed to provide safe access to and use of this data.

10 For example, see Chelsea Whyte, "Police Can Now Use Millions More People's DNA to Find Criminals," *NewScientist* (October 11, 2018), *https://oreil.ly/fKroy*.

In some jurisdictions the lack of suitable, and practical, privacy technologies has led regulators to allow biomedical identifiers to be used for secondary purposes such as research under a risk-based approach, wherein the appropriate privacy and security controls are used to offset some of the risks posed from linking and re-identification. Again, this is the reasonableness in privacy laws and regulations that affords some flexibility as risks become more practical and solutions evolve and become more readily available. However, new methods are emerging that can scale much better than in past implementations.

Secure computation of genomic data

A person's entire genomic sequence is unique, and also very long and complex (e.g., there are over three billion base pairs in the human genome), meaning there are practical limiations involved in computing this data in a privacy-enhancing way.[11] This is an area ripe for development given the interest in precision medicine, meaning that treatments become tailored to a person's genomic makeup. We can summarize current approaches to protect genomic data into two camps:

Process controls
: Most of the practical implementations to protect genomic data have focused on administrative controls. This means the data is still personal and, because nongenomic identifiers are removed (e.g., name, address), it's not possible to link genomic data to other anonymized data. An incident in which data is lost or stolen would mean a loss of personal information.

Cryptographic protocols
: There has been some good academic work attempting to bring technologies in secure computation to the challenge of analyzing genomic data. Typically these approaches limit the research utility by limiting the available operations on data in order to preserve privacy, and are computationally intractable for genomic sequences.

Genomic sequences are very detailed, and analyzing a sample that will yield meaningful results requires significant computing power. It's not difficult to come up with a research protocol that would literally take years to compute on a suitable number of human genomic sequences. Until recently, only limited pilot studies have been attempted for very specific analyses or use cases.

One significant breakthrough in making the analysis of genomic sequences possible at scale comes from using tokenization, that is, replacing variants in a genomic

11 Techniques and limitations are summarized in Bonnie Berger and Cho Hyunghoon, "Emerging Technologies Towards Enhancing Privacy in Genomic Data Sharing," *Genome Biology* 20, no. 1 (2019): 128, *https://oreil.ly/6A_-y*.

sequence with a token that has no real value or meaning. The token is nothing more than a reference to the variant through a secure tokenization system.[12] The process can be described in just a few simple steps:

- Replace each genetic variant with a token, so that an individual's unordered sequence looks like noise.
- Analyze correlations between variants of interest and outcomes using tokenized data.
- De-tokenize the aggregate results, provided there is a minimum number of representative individuals in the outputs.

Using such an approach ensures that data resolution is maintained, and that data can be annotated and classified by types of variants. It also enables the use of linkage tokens with other anonymized data. The result is that institutions can share data since privacy and competitive concerns are addressed, while enabling the majority of genomic research to take place while being privacy enhancing. In other words, it helps address and bridge the divide between individual privacy and research in a practical manner.

Final Thoughts

While pseudonymized data is still personal, there are different opportunities to build data protection and privacy into the handling of this data. Laws and regulations will typically favor the use of less identifiable data, and in some cases even encourage it. This is a positive advancement in the regulatory landscape, and one organizations should make the best use of in processing personal data whenever, and wherever, possible. It will help reduce the impact of incidents by limiting what is disclosed, demonstrating to regulators and data subjects that thoughtful approaches are being taken in handling personal data.

The story doesn't end there, however. Pseudonymized data, in its most basic form, is a minimum of what can be done. Additional measures can be put in place to reduce potential disclosures from the processing of pseudonymized data. We can even introduce concepts of anonymized to pseudonymized data to make the data less identifiable, and therefore more privacy enhancing. Even when the data remains personal, there are advantages and, in some cases, regulatory carve-outs or exceptions to encourage further anonymization. Even where the laws or regulations have not anticipated such measures, the additional protection means the impact of incidents is further reduced. This means greater support from regulators and data subjects.

12 A commercial solution is described in "Unlock Genomic Data While Preserving Privacy," IQVIA, accessed March 28, 2020, *https://oreil.ly/CyCIP*.

Pseudonymized data is where we start our journey toward reducing identifiability, and in the next chapter we will use concepts introduced so far to see how we can leverage the full spectrum of identifiability. We've started by looking at anonymization in the more traditional way, from identified to anonymized as though we are pushing data through (with various considerations in working with data throughout the spectrum). In the next chapter, however, we will turn things around by thinking of it in terms of pulling anonymized data into a collection point or use.

CHAPTER 6
Anonymized Data

The data custodian of personal data, be it identified or pseudonymized, has options for the responsible sharing of data or analytic outputs. Determining the most suitable approach requires an evaluation of wants and needs, and can be complex depending on the use case. We have provided many tools and options to consider, including concepts from privacy engineering and requirements engineering to help drive these initiatives.

Oftentimes, however, it's not the data custodian that is driving the design and development of anonymization pipelines, but a partner seeking to innovate from the data. The data custodian will be compensated, perhaps through new insights or features added to products and services they receive from this partner, or also by getting access to the technology or outputs of the pipeline. Either way, both parties will want assurances that the anonymization is done properly.

In previous chapters we considered identified data, and then pseudonymized data. It therefore seems natural to build a pipeline from identified to anonymized, as if data is pushed through by the data custodian, and we will consider that in this chapter. Considering it from the other direction, as if data is pulled through the pipeline by the data recipient, introduces a perspective that can actually provide new insights and methods for creating anonymized data, and we will consider that as well.

Identifiability Spectrum Revisited

Let's start by revisiting the identifiability spectrum, a concept we first introduced in Chapter 2, from identified (with the names and addresses of data subjects) to anonymized (in which data subjects were clustered with similarly looking individuals). We can now consider four possible stages along this spectrum, with each building on the

last, and transformations driven by the need to reduce identifiability (from highest to lowest):

Direct identifiers
: To support direct interaction with data subjects, no transformations are applied to names, addresses, and other directly identifying information (although access controls can ensure that only relevant information is available for the intended purposes).

Transform direct identifiers
: To support pseudonymized services, in which there is no direct interaction with data subjects, the direct identifiers are transformed in a reversible manner (where keys to reverse the process are secured and accessible only for authorized purposes or destroyed).

Transform indirect identifiers (public)
: To remove the ability to single out data subjects based on information knowable to the public, such as fixed demographics, which is the most broadly available and therefore of greatest risk to re-identify.

Transform indirect identifiers (acquaintance)
: To remove the ability to re-identify, based on information knowable to acquaintances, given the context of the data sharing. Information here may even be used to determine the level of granularity that is appropriate for the previous identifiers.

We show these steps in Figure 6-1, a sort of stairway to anonymization. Now this isn't to say there can't be other steps on the spectrum from basic pseudonymization to anonymization, as we want to leave room for future enhancements and innovation. But these are the steps that exist based on practical considerations and influence from laws and regulations (whose interpretation is subject to evolve, of course).

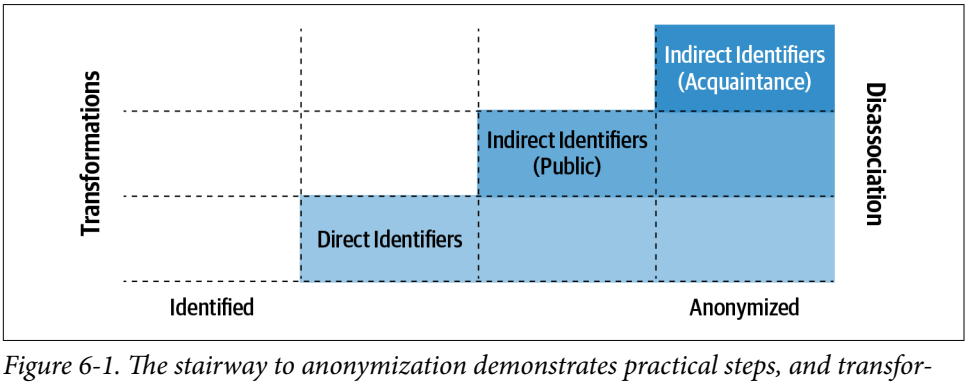


Figure 6-1. The stairway to anonymization demonstrates practical steps, and transformations of identifiable attributes, increasing disassociation between data subjects and the data (the purpose, after all, of anonymization).

Of course there is more to anonymizing data than simply "transforming" the indirect identifiers. The transformations must meaningfully change the measure of identifiability. There are also categories of information that are related or associated to the identifiers that we need to consider, which we'll do now.

Making the Connection

Recall that the more complex the structure of the data (as described in "Structure of Data" on page 38), the more complex the measures of identifiability. That means those data transformations need to address measures of identifiability specific to the structure of the data. More important still is understanding how the correlation structure and connections of the data can leak indirectly identifying information if not properly accounted for.

Highly correlated variables will cover more or less the same information and can therefore be grouped, with the most identifying variable of the group driving the identifiability measure. However, the data transformations needed to reduce identifiability, such as generalization or suppression, need to be propagated to *all* variables in the group.

Take a service date and a computer log date, which are usually one and the same except for some entries in which corrections were made to a client's records. Changing the service date without changing the log date would mean the original dates were still available through the log date. Big oops! Rather, the service date and log date should be changed together, in the very same way.

Regarding dates, there are different approaches to transform them once connected to ensure they maintain the desired level of usefulness, as described in more detail in our previous book.[1] Generalizing and selecting a random date from within the generalized interval works for cross-sectional data. For longitudinal or time-series data, a more suitable approach is date shifting. Depending on how much identifiability needs to be reduced, this shifting could be achieved by zeroing a sequence of dates to a start value and adding a random value to the sequence. Or the intervals between dates in the sequence could themselves be randomized, a bit like an accordion, increasing the level of protection.

Connecting variables to propagate data transformations can seem daunting when there are many variables or fields where it is known that identifying information could be leaked. In some cases, calculating a *leak rate* may be appropriate to

1 See Chapter 5 of El Emam and Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*.

determine if leaks of indirectly identifying information are practically significant, thereby warranting special treatment. By practically significant we mean they have meaningful impact on the identifiability measure. For example, a change of 0.1% may not be statistically or practically meaningful, and therefore not worth the effort, and can possibly be compensated for elsewhere. Leaks should also be put into perspective of how likely it is that an adversary can or will exploit them and their impact.

It may also be possible to "rebalance" the data transformations toward indirect identifiers that have fewer correlations or connections elsewhere, thereby simplifying the task. This is the advantage of having statistical measures of identifiability to drive data transformations. Indirect identifiers are not all created equal. Some will have more or fewer connections to other variables, and also impact identifiability more or less. Ultimately, it's an optimization exercise to minimize effort and minimize data transformations while maintaining the usefulness of the data where it is needed most.

We can assume these connected variables are dealt with and instead focus our attention on the indirect identifiers in the data. The first we'll consider is the more standard case of anonymization at source.

Anonymized at Source

The "source" at which data is anonymized could be the data custodian of personal data, or a partner who acts as an extension to the data custodian—a processor or business associate, acting on behalf of the data custodian and therefore under the same legal obligations in working with the personal data. In other words, personal data and anonymized data exist under the same legal entity in this pipeline.

We've discussed the Five Safes of anonymization in Chapter 3, so we won't rehash those details here. Rather, we will instead just look at the transformations of data and considerations in the process of anonymization. We assume that the data fields that are deemed necessary have already been included, while eliminating all other fields that are not required, so that our focus remains on data transformations rather than an assessment of requirements.

Masking identified data
: Remove the directly identifying elements by deletion of fields or transformations into tokens, pseudonyms, or fake data. This is basic pseudonymization.

Removing the ability to single out individuals
: Uniques based on publicly knowable elements are eliminated through data transformations.

  - This can include macro-level transformations, known as global recoding, to change a field into a less granular version of itself (e.g., date of birth into year of birth).

- Or micro-level transformations, known as local recoding, to change only the data that requires less granularity (e.g., the one unique data subject's date of birth into year of birth, and randomized to within that year range to preserve the date format).

Conducting a threat assessment
: Understand how the people and settings will determine identifiability of the data. This is the core of the Five Safes of risk-based anonymization.

  - Assuming a nonpublic use case, identifiability will be measured using an average, considering both publicly knowable and acquaintance knowable features.

Transforming indirect identifiers
: Perform data transformations required to provide reasonable assurance that the data is nonidentifiable based on established tolerances.

The reason for walking through this risk-based anonymization process is to show how the individual pieces we described in previous chapters fit together. In particular, we want to draw your attention to transforming indirect identifiers known to the public, to eliminate the ability to single out data subjects, since we will explore this further in this chapter.

The concept of *strict* average was introduced in our previous book,[2] which combines two aspects discussed above: eliminating uniqueness and average identifiability. The concept was introduced at that time to ensure that the potential vulnerability of uniqueness, or singling out, was properly addressed when an average is considered (otherwise the average could be acceptable, but include uniques). In this book we have kept the concepts of uniqueness and average identifiability separate to ensure they are properly understood, but also because the concept of singling out is valuable in its own right to meet the needs of some use cases.

The anonymized data can then be provided to a data recipient through multiple avenues:

- The anonymized data itself may be directly shared with the data recipient, after an assessment of their privacy and security practices to determine the degree of anonymization required (as in the case of the Five Safes).

2 See Chapter 2 of El Emam and Arbuckle, *Anonymizing Health Data: Case Studies and Methods to Get You Started*.

- Access to the anonymized data may be provided through some kind of secure portal in which restrictions ensure that the data remains under the control of the data custodian (or a partner) and not the recipient. This way any changes required of anonymized data can be done without relying on the data recipient to delete previous copies. Changes could be due to a data quality issue that has been identified or resolved, or even an update to the anonymization methodology due to improvements or for addressing a new vulnerability.
- Access to the anonymized data may be provided through some kind of analytics platform. The data recipient could have access to the underlying data or not if the intent is for them to only have access to outputs that they themselves generate. However, the data would be anonymized assuming the recipient could recover the underlying data.
- Analytic outputs may be provided through the use of an analytics platform that is run by the data recipient, or the outputs may be derived for them by the data custodian (or partner). The assumption here is that the recipient does not have access to the underlying anonymized data, so the analytics platform would need to have restrictions in place to prevent the reconstruction of data by crossing functions or repeated application of analytic functions.

This can be thought of as a two-step process: anonymize data and then provide data to the recipient. But what if there are other sources of presumably permissible data that we want to include? Let's consider these before we mix and match and risk getting ourselves into trouble.

Additional Sources of Data

Before we explore a multiphased approach to an anonymization pipeline, we should also consider the data sources that could be available for anonymization.

Public sources
: Data products made available by national statistical organizations, through open data or transparency initiatives.

Volunteered or authorized
: Data subjects may volunteer or authorize the dissemination of certain data, in which case the data may have some data transformations applied to reduce identifiability and protect data subjects, but at a higher tolerance.

Direct sourcing
: Nonpublic data collected through partnership with other organizations, which may or may not have been properly anonymized (putting the recipient at risk should it be determined that the data is personal).

Aggregator portals
: Environments that provide limited access to anonymized data or analytics derived from anonymized data (the primary output being aggregated data or analytics).

All of this ingested data could be used to enhance available data and products that will be anonymized. Depending on the nature of the data, these sources could increase identifiability once linked or be used to enhance data stores. Anonymization should therefore include the full data that is intended to be shared with recipients. This will likely require an assessment of data flows to identify boundaries and uses, to determine where it may be necessary to reduce identifiability.

Identifiability may also need to be assessed for the directly sourced data, at the time of ingestion. This will be especially the case if it's unclear how the data was anonymized, or there is a need to improve the anonymization based on company policies. Similarly, in the case of data from aggregators with data distribution portals, we will want to consider performing due diligence on their privacy measures.

Public data may not require an assessment of identifiability; however, it will depend on confidence in the anonymization process and jurisdiction since interpretations vary with regard to obligations to data subjects (e.g., identify the legal basis for processing and ensure use of the data is fair and justifiable). The notion of "public" data is wrought with ethical and legal nuances that should be considered.[3] Consider the purposes for which the released data is made available—public, volunteered, or authorized by data subjects. Respect the purposes for which the data was made available and ensure the use fits the *intended* uses and with ethical standards.

We've described anonymization at source, a standard affair in our world. What we would like to explore, however, is how we can leverage the intermediate steps to deliver anonymization in phases. This could have advantages in terms of developing novel use cases and driving the need for more data to fuel innovation. In particular, let's consider pooling data for the purposes of deriving richer insights than may otherwise be possible.

3 See, for example, Woodrow Hartzog, "The Public Information Fallacy," *Boston University Law Review* 99, no. 2 (2019): 459, *https://oreil.ly/UMW4P*.

Pooling Anonymized Data

Without a process and method to pool data, only localized and time-boxed outputs or insights can be established, where data was collected based on the available sample of individuals from a population at a particular interval of time. Such constrained outputs or insights can be very limited due to the size of the sample compared to the population, demographic profiles, and intentional or unintentional bias from the data collection (for example, targeting specific individuals based on demographics or attributes).

Imagine that people choose to conduct their daily banking based on the color of the bank's advertisement. Drawing insights about banking transactions from only one bank's time-boxed marketing campaign would necessarily bias the results toward specific personality characteristics, perhaps even specific demographic profiles. But the bank is likely to have different marketing campaigns, experimenting with different colors and messages. Or there may be an opportunity to pool data of several banks, each running marketing campaigns that will attract different customer profiles.

Pooling data allows for a more complete representation of a population, increasing the accuracy, applicability, and generalizability of outputs or insights drawn. Furthermore, rare events or patterns are more likely to be uncovered as the statistical power of analytics is increased. There is a real need to collect and process data in such a way that the data can be pooled to derive these insights at a broader and deeper level through advanced analytics.

It can be challenging, however, to bring various sources of data together, especially when personal information is involved. Indirect identifiers found in personal data are most often the information that is most valuable for analytical purposes. Transformations to change data into less identifying information are therefore minimized to ensure the data remains useful, while ensuring that privacy is sufficiently protected. Limiting access to indirectly identifying information is also minimized for the same reasons.

Let's consider a way to pool anonymized data, minimizing data transformations and identifiability, by considering the recipient perspective of *pulling* anonymized data into the pool and into use cases. The approach we'll describe could be a single source being tapped for the pool, or multiple sources. The key aspect will be this notion of having a pool, or an intermediate source, of data from which use cases can be tapped into.

Pros/Cons of Collecting at Source

In order to collect data from source, such as collecting data from a device or hospital, it's often desirable to anonymize personal data to protect the privacy of data subjects and reduce contractual and regulatory burdens. Contractual burdens should not be

underestimated, as every data custodian (from which there is a data source) will require a different contracting process and set of obligations, with several iterations involving many stakeholders. Regulatory burdens often involve detailed legal processes, which can be avoided when data is properly anonymized, and the technical and organizational controls are applied continuously to ensure the data remains anonymized.

In some circumstances it is possible to anonymize the indirectly identifying information at source, provided there is sufficient information to inform the process since it is statistical in nature. That is, with sufficient data, the measures of identifiability will be accurate, and the minimal set of transformations can be applied based on the context in which the data is released or shared.

Without sufficient data, however, the errors around measurement can be large, and identifiability is likely to be overestimated. For example, for small samples and with a lack of sufficient population data, there will be more sample uniques considered to be population uniques, and cluster sizes will in general be underestimated (see "Re-identification Science" on page 31 for a refresher on these concepts). In this situation the data will be overly transformed to reduce identifiability to an established level and the analytic utility of the data will be significantly reduced (even to the point of being useless).

Pooling data from multiple sources may therefore be preferred, where there is a need to leverage a more comprehensive source of data to improve measurement and more accurately anonymize data. Removing the identity of the source (e.g., removing a hospital's name) from the data also allows for a broader and more general collection of data to enable the training and use of advanced analytics. For example, modern approaches to machine learning and artificial intelligence can be trained on large pools of data.

By excluding the identity of the source (e.g., device or hospital), not only is identifiability of individuals reduced, but also the algorithms will not be biased to the source itself. Rather, the algorithms will be more general as they are trained on the indirectly identifying and other features present in the data. From a privacy perspective, the removal of the source identity has the added benefit of increasing the efficacy of anonymization techniques, which can struggle to achieve high data utility when faced with small populations.

Another significant practical challenge is getting data custodians to install complex software into their respective environments. Technologies such as secure multiparty computation can be burdensome, as they require significant computing resources, and detailed protocols must be established in advance to coordinate the secure

analytics protocols between data custodians. Although there have been advancements in this space, it's still a heavy investment on the part of data custodians and is difficult to scale to multiple, distinct data custodians. The practical challenges of secure multiparty computation have limited the available commercial offerings to only a very few.

A data collection pipeline can instead be fashioned, most likely at the expense of the organization seeking to collect data, using a lightweight anonymization engine to suitably transform that data based on the context in which the data is released or shared. No coordination is needed between data custodians, as they need not even be aware of each other's involvement. This lightweight anonymization engine is explained in the next section. It's minimal and easy to deploy, with methods that support the effective anonymization of small populations and allow for the pooling of data from multiple sources.

Methods of Collecting at Source

An intermediate anonymization step can be completed at source, such as a device or hospital, through a lightweight engine. It can be relatively easy to deploy, yet meet the modern regulatory privacy and data protection requirements of risk-based anonymization through a multistep process. This can be done using any of the following:

- Embedded code within the source system or physical device that knows how to send required data into the data pipeline
- Independent software agents or modules that are close to the source and perhaps even on the same device, in which the agents run as separate processes with no dependency concerns
- Software interfaces that receive data sent by the sources

The lightweight intermediate source anonymization pushes data to a safe environment that may even reside outside the control of the source data custodian. The intermediate anonymization performs source anonymization appropriate for the purposes of holding or pooling data in the safe environment, namely:

- Direct identifiers (e.g., name and address) are transformed (pseudonymized or removed) to ensure that data subjects are not directly identifiable with a high probability of success.
- Indirect identifiers (e.g., demographic variables) are transformed to ensure that data subjects are not unique in the population on these indirect identifiers so that they cannot be singled out.

The safe environment is only used for the purposes of holding or pooling data—no data analysis is performed on the data in this environment, turning strongly pseudonymized data into anonymized data given the context of secure pooling of data. This

environment is a holding area in which there are strong controls to prevent misuse or breaches. Eliminating the ability to single out data subjects ensures the data is anonymized for the purposes of holding or pooling data only. Any secondary uses, beyond the holding or pooling, would not be permitted on this data until the identifiability has been evaluated, as illustrated in Figure 6-2.

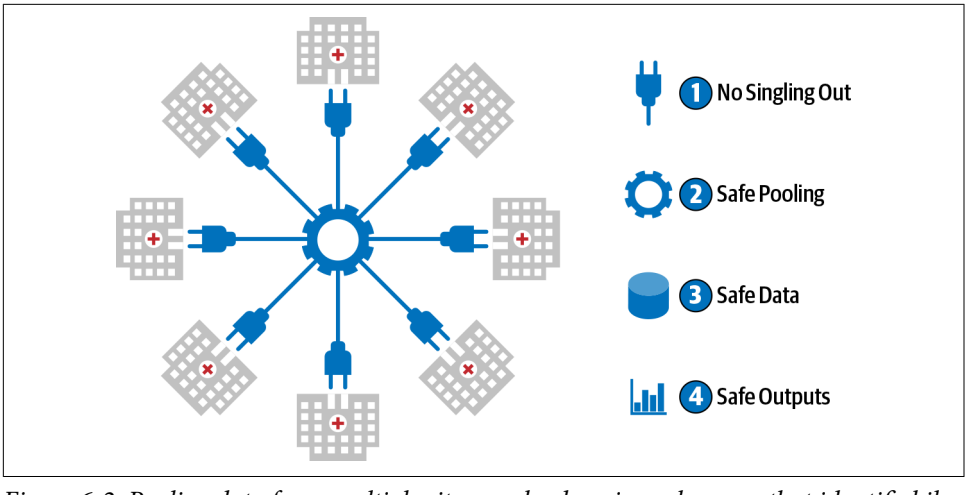


Figure 6-2. Pooling data from multiple sites can be done in such a way that identifiability is minimized.

As summarized in the figure:

1. Eliminate at source the ability to single out individuals
2. Pool this data in a secure location for further processing
3. Measure identifiability on the pooled data based on the use case
4. Anonymize to derive useful data

Appropriate controls are put in place to ensure that cryptographic keys are stored and handled appropriately—at source, by a trusted third party, or exchanged between parties in the same network such as in public key cryptography. The keys are not accessible in the safe environment. The keys are only available for the purposes of re-identifying the data at source by the data custodian. This allows any intelligence derived from the functionally anonymized data to be used for primary purposes (such as the direct care of patients, or client services, as described in "Mixing Identified with Anonymized" on page 81).

As we mentioned previously, in some jurisdictions, keeping the cryptographic keys may not be permissible for the purposes of creating nonidentifying data. In this case the cryptographic keys are destroyed, making pseudonymization irreversible.

Safe Pooling

Once the Safe Data is pooled, business-to-business integration architectures can ensure that private data flows remain unrestricted within the boundaries of the appropriate business processes. Data lakes can ensure controlled, flexible, and secure access to all data via a single data management platform; or a network structure of data collection and storage can harmonize data into a single schema prior to being stored in a data hub, which can facilitate indexing and analytics. Data sources can include devices, machines, or organizations that transfer individual or grouped collections of information to hold or pool.

Safe Outputs or insights can be drawn from pooled data with, for example, data lakes or centralized hubs, using analytical systems including, but not limited to, artificial intelligence and machine learning. Additional integration architectures can also provide mechanisms for safe access or transfer of Safe Data, outputs, or insights to appropriate recipients, organizations, or devices, with additional data transformations as deemed appropriate. Information drawn from pooled data may also be transferred to the original sources of devices or organizations to augment their data systems (including outputs or insights).

The safe environment, in which the data is pushed for holding or pooling, has strong mitigating controls in place, namely (Safe Settings):

- Access, disclosure, retention, and disposition of personal data
- Safeguarding of personal data
- Ensuring accountability and transparency in the management of personal data

For example, an IT administrator could be granted access to this safe environment for the purposes of managing the mitigating controls. Strong contracts are also in place to limit the use of the data to only that required for managing the mitigating controls. Automation may also be used to limit the administrator's role to only that which requires human intervention.

Let's draw this out further, thinking back to our threats (Safe Data):

Deliberate
: Targeted attempts are drastically minimized by strong mitigating controls and having highly trusted recipients (IT administrators, which in itself should represent a relatively small number of people).

Accidental (inadvertent)
: Inadvertent re-identification becomes unlikely when access is limited to IT administration (i.e., no data analysis is required).

Environmental (breach)
: The possibility of data being lost or stolen is always a risk, in this case minimized by having strong mitigating controls.

Recall that the ability to single out individuals has been eliminated by transforming indirect identifiers that are knowable to the public. This means that the overall identifiability for the secure pooling environment will be very low, and the impact of a potential breach incident is also mitigated.

Data holding may also be for the purposes of pooling data from single or multiple sources or connecting multiple data pools. For multiple sources, a common data model may be used to consolidate information into a data lake, without distinguishing between sources. Transactional records of data from multiple sources could result in collisions in linking identifiers, in which case these linking identifiers may be reassigned as part of the common data model.

Although the anonymization pipeline is divided into stages, the source data is considered reasonably anonymized for pooling, and the data access model to the pooled data will also ensure that data and outputs are anonymized. This should facilitate agreements with providers of the sourced data regarding access and retention of the anonymized data. Purposes will need to be defined, as well as limitations and ethical boundaries on reuse of the anonymized data. Consider how you will include benefits to the providers of source data, such as supporting their own innovation or research efforts with insights or access to anonymized data or outputs.

Linking between sources is also possible using, for example, a one-way hash, format-preserving encryption, or privacy-preserving record linkage, depending on jurisdictional requirements. This would need to be passed through the intermediate source anonymization.

Access to the Stored Data

Access to the data hold or pool for secondary purposes is only possible through risk-based anonymization. That is, a risk assessment is performed (following the approach described in the Five Safes) to determine identifiability estimates, which drive the data transformations required to access or publish a dataset that is below the identifiability threshold used to provide reasonable assurance that data is nonidentifiable. Recall that we justified the pooling with an intermediate anonymization step by limiting access to IT administration only and implementing access and audit logging of the data environment. Therefore, access and use require an additional assessment of identifiability based on the use case.

Feeding Source Anonymization

As we mentioned earlier in this chapter, it is possible to anonymize only at source, without the need to pool data in an intermediate environment. The key element is the statistics that will inform the identifiability measurement process that drives the transformations on the data. Population statistics can therefore be used to inform the source anonymization.

Most datasets are only samples from a much larger population. For example, the breast cancer patients of one hospital are likely part of a larger population of breast cancer patients in an identified region (that is, there may be multiple treatment facilities). These population statistics can come from public or nonpublic sources of information, or from the pooling of data from multiple sources.

Statistics are therefore optionally collected and fed to the intermediate source anonymization. This would improve on the level of data transformations applied to indirect identifiers that ensure that data subjects are not unique in the population. The pooled data, especially from multiple sources, can produce population statistics that are more accurate than the sample statistics generated at source.

Going a little deeper, instead of anonymizing at source using sample data, another option is to transfer frequency counts of indirect identifiers and their relationships to one another to the safe environment used for holding and pooling data. Frequency counts across indirect identifiers, per individual, could be linked by a persistent identifier generated at source. In this case, the pooling would be on the frequency statistics. This would be used to determine the intermediate source anonymization needed, with a specification file returned to the source for applying the appropriate data transformations.

A one-way hash or privacy–preserving record linkage would in this case be used to reduce the counts of indirect identifiers that do not change for a data subject. The linkages would be to frequency counts, to determine where counts should or should not be summed. Otherwise some form of secure computation could be used. The full pipeline and steps are shown in Figure 6-3.

In some cases (e.g., small samples in each of the multiple sources to be pooled), the use of frequency statistics would be augmented with measures of correlation between indirect identifiers and other measures of association. This metadata would be used to supplement and inform the risk-based anonymization strategies derived from the frequency counts.

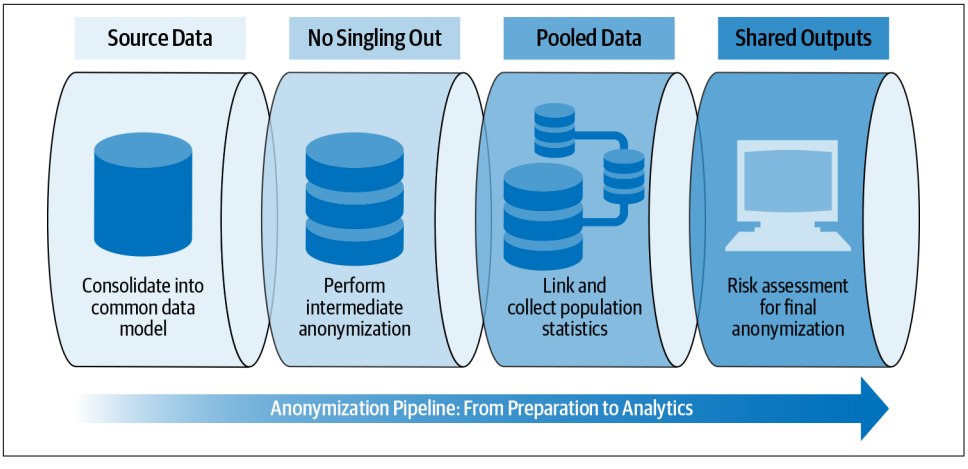


Figure 6-3. A multistage anonymization pipeline from raw source to intermediate anonymization, pooling, risk assessment, and final anonymization to share data.

Final Thoughts

Anonymization is usually considered in a single, two-step flow from identified to anonymized. Whereas the process of anonymization is described in several steps, it's still a single box in a process diagram: take identified data, feed it into the anonymization engine, and out comes anonymized data. This is certainly the most common way to anonymize data, pushed from data custodian to data recipient.

Data flows can be more complex, however, as various use cases emerge to fuel analytic engines. By rethinking the flow of data in terms of what the data recipient will envision, pulling data into their use cases provides new opportunities. Pooling data is not a new concept, with data warehouses and data lakes being well-known concepts in the industry. But it's allowed us to conceptualize new anonymization pipelines, creating phases along the identifiability spectrum with mitigating controls that ensure the data remains anonymized across these phases.

Pooling anonymized data has also solved another problem: that of generating accurate population statistics based on real-world data. These population statistics help ensure that anonymization at source produces more granular data while ensuring that the pooled data is both safe and accurate. From this we can also meet the needs of diverse use cases, as the approaches presented are flexible.

While this chapter has provided the means to anonymize data, as a single data pipeline or in a networked fashion, the responsible sharing and use of data requires us to consider the ethical uses of data and oversight mechanisms for how data is or will be used. In the next chapter we will look at these additional aspects of privacy and data protection.

CHAPTER 7
Safe Use

When we described the Five Safes of risk-based anonymization in Chapter 3, we described it as a framework for the safe use of data. Although we touched on them, we left the details of trust and ethics for this chapter. Reducing identifiability will help mitigate risks, making data reuse more balanced toward benefits. But while nonidentifiabile data is no longer personal, it can still be used in ways that are creepy or harmful.

Although we can describe trust in terms of privacy and data protection in general terms, we will also look at how these concepts are applied to analytical models and decision making. Since the world of analytics is complex, involving the fields of computer science, mathematics, and statistics, we also describe some of these technologies and their challenges with respect to their safe (and therefore responsible) use.

Governance plays a role in how an organization can develop and ensure the safe use of data. This will involve a blend of ethical oversight and monitoring the integrity and credibility of analytical models, as we'll see in this chapter. There is increasing pressure from regulators to consider data ethics, given the many ways that algorithms have failed to inspire confidence or trust. Once a framework is adopted, governance principles can be embedded in organizational tools, including technology-enabled processes, thereby getting people to acknowledge responsibility. Depending on the degree of impact and sensitivity, committees can be formed for oversight and decision making.

Foundations of Trust

Demonstrating trust is an important aspect of building respectful relationships with data subjects and regulators. A growing number of scholars are attempting to reframe the conversation of privacy in the context of trust, which fits well with our theme of

safe use. Although many of the same privacy principles emerge, thinking of them in terms of trust can provide a more positive lens to data protection, and ensure that designs are value driven (e.g., the design of anonymization pipelines). Four foundations of trust have been suggested:[1]

Honesty
: Explaining what data is collected and how it is used. The onus is on the data custodian or steward to ensure the explanation they provide is clearly articulated. Although typically described as transparency, framing it as honesty certainly changes the implied tone! Honesty requires an obligation to be clear, accurate, and forthcoming, and to ensure that any misunderstandings are corrected.

Discretion
: Treating personal data as confidential, in a broad sense, by only disclosing it in ways that data subjects would expect or generally consider reasonable. Being discreet is broader than simply not disclosing personal information; being discreet is about being measured in terms of what is shared, how it is shared, and with whom it is shared.

Protection
: Keeping data secure from threats, and keeping data safe for access. More than technical and organizational safeguards, which are clearly relevant, protection includes risk assessments, threat modeling, data minimization, etc. It includes data privacy and is driven by elements of discretion.

Loyalty
: Acting in the interests of data subjects, or at least not acting against their interests. Reusing data, to improve operations or innovate, can maintain trust by being loyal to data subjects. Be mindful that preferences, weaknesses, desires, and tendencies are exposed in personal data. Do no harm. Interests of data subjects should be considered ahead of short-term gains.

The goal is not only developing trusted relationships, but ensuring that those relationships are sustainable and long-term. As they say, it's easier to lose trust than it is to build it back up. This view provides a strong basis for viewing privacy as a societal good, avoiding the narrow view of single individuals being responsible to protect their privacy and security. It is far more helpful to develop frameworks that build trust and help organizations do the right thing. Context plays an important role here. People have expectations regarding what is collected about them, and how that information is used.

1 The foundations of trust are described in Neil Richards and Woodrow Hartzog, "Taking Trust Seriously in Privacy Law," *Stanford Technology Law Review* 19, vol. 3 (2017):431-472, *https://oreil.ly/xO0TM*.

Notice how reducing identifiability applies to many of the foundational principles of trust just described (through obfuscation, or privacy by obscurity). We can also see overlap with the concept of legitimate interests described in "Purpose Specification" on page 12. A risk-based approach to anonymization, considering all the factors described in the Five Safes of Chapter 3, will decrease risks substantially, thereby maximizing benefits and minimizing risk. It's not just the anonymization itself, though, but all the aspects that are factored into the risk-based framework we've described: Safe Projects, Safe People, Safe Settings, Safe Data, and Safe Outputs.

While anonymization will reduce privacy risks, it does not address all aspects described in the foundational principles of trust. Anonymized data can still be used to do harm through discrimination or stigmatization to groups or individuals. Regardless of whether the data represents an entire population or a sample, harm can be specific to the original data subjects from which the data was drawn. And when it is inferred to other groups, not necessarily the original data subjects, well, that doesn't make it any less egregious. It would be disloyal to the data subjects from which the data was drawn to use inferences to do harm. Again, consider privacy as a societal good.

That being said, it's not in the interests of the data subjects for an organization (public or private) to close its doors if it's providing a valued service to them or others. Loyalty does not imply that an organization can't innovate with data. Quite the contrary! We expect, if not demand, improved services, new features, new insights, and hey, new services as well. We are seeking a win-win scenario, privacy, and innovation. Organizations can consider themselves stewards of data, and use the principles of trust to help guide them in terms of safe use of that data. Let's start by considering these principles of trust in terms of the algorithms that are used to build analytical models.

Trust in Algorithms

Algorithms are trained on, and applied to, data. Algorithms also generate data through inferences and predictions. The data may not be identifiable, but it can still be misused. The misuse may not, however, be intentional. Algorithms meant to automate and derive insights have a variety of inherent technical challenges with their application in practice. We'll look at some of these a little later, after we describe different modeling approaches so that we can better frame the discussion.

Regulators have become increasingly concerned with how the insights from algorithms are being used to make decisions about individuals or influence their behaviors. For example, in 2018, data protection and privacy commissioners agreed to a declaration on ethics and data protection in artificial intelligence, endorsing several

principles that we will explore in this chapter.[2] As a result of the plethora of theoretical attacks available on black-box access models built on personal data, some have even considered that data protection laws might, in the future, consider models themselves to be personal data.[3]

"All models are wrong, but some are useful." Statistician George E.P. Box wrote this,[4] and variations on this theme, in a few academic papers. He advocated for simple models that demonstrate great insight, and a focus on what is "importantly wrong" in any model. It's worth keeping this in mind when evaluating any model that will be deployed in practice, to ensure that our trust is put in the right places.

Before we describe different modeling approaches, as well as failures of trust or principles that have emerged, let's take a moment to consider modeling more broadly. Although there have been tremendous advancements in the use of algorithms, it's still worth starting with the easy stuff first (the so-called law of parsimony). It will make everything that follows regarding responsible use easier to envision. Once you've nailed that down to ensure what is being modeled is well enough understood that you can envision what the results and boundaries on use should be, move on to the more complex models (assuming there's value in doing so).

Techniques of AIML

Others have attempted to provide a taxonony for algorithms and supporting infrastructure, so we will not delve into those details. It doesn't matter to us how you categorize and organize these things, just that you know some of the terms used and where they fit in the broader discussion of algorithms and trust. We will use the increasingly common acronym AIML, for artificial intelligence and machine learning, to capture the universe of algorithms we may consider.

Traditionally, statistics has focused on mathematical formulations of uncertainty, involving randomness described by probability distributions. Statistical modeling will therefore include formal derivations of properties (e.g., convergence of estimators,

2 See "Declaration on Ethics and Data Protection in Artificial Intelligence," *International Conference of Data Protection & Privacy Commissioners*, Brussels, Belgium (October 23, 2018), *https://oreil.ly/6VZog*.

3 For example, see Michael Veale, Reuben Binns, and Lilian Edwards, "Algorithms that Remember: Model Inversion Attacks and Data Protection Law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, no. 2133 (2018): 20180083, *https://oreil.ly/yw_lp*.

4 A classic description is provided in George E.P Box, "Robustness in the Strategy of Scientific Model Building," *Robustness in Statistics*, eds. Robert L. Launer and Graham N. Wilkinson (New York: Academic Press, 1979), 201–236, *https://oreil.ly/yqFpQ*.

unbiasedness, errors). Machine learning has focused on more empirical development of algorithms. However, the fields now borrow a great deal from one another, with a subfield called statistical learning that attempts to bridge machine learning and statistics.

Although it's important to have algorithms that "work" in practice, theory can provide a better understanding of the properties of an algorithm. This can provide greater trust in the analytical results that are derived. The topic is too big for our purposes, and likely too technical, but it's worth knowing a few of the terms that are used, and how machine learning and statistics relate. Of course there are other fields that also contribute to this space, and we will attempt to include these in our brief introduction to algorithms.

Classical machine learning

To parse out some of the techniques, we can start with "classical" approaches. We use the language of machine learning, where learning is called model fitting or estimation in statistics, and provide a bridge to the language of statistics.

Supervised learning
: Algorithms that are trained on labeled data, with both input (covariates or predictors) and output (outcome or signal). In statistics, the algorithms that fall into this category are called classification or regression.

Unsupervised learning
: Algorithms that operate without labeled data, finding patterns nonetheless. In statistics, the algorithms that fall into this category are called clustering or density estimation.

Reinforcement learning
: Algorithms that work using a form of reward system to drive how they function, assuming a basic model of behavior. The algorithms that fall into this category include dynamic programming.

These are general categories that can capture a wide range of algorithms. Many of those algorithms, such as regression, would be a first start in the spirit of building simple models that are effective (a.k.a. the law of parsimony). The above categories can also be applied to the next class of modern approaches that have inspired a great deal of attention.

Neural networks

A neural network is basically a set of regressions that includes some complex interactions between the input and output. The individual node of a neural network is basically a regression, and the network connects the nodes to create those complex interactions, an example of which is shown in Figure 7-1. To keep things simple,

think of the node as a classic linear equation with a slope and intercept. But by chaining nodes together, the model becomes considerably more complex.

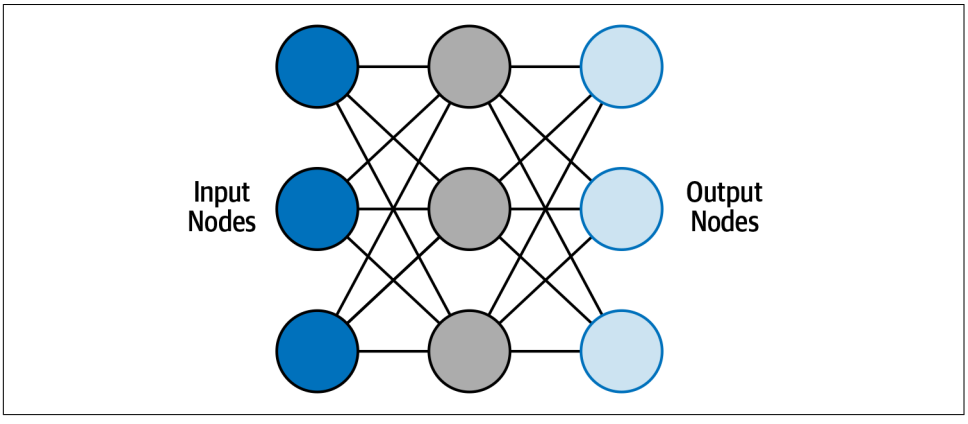


Figure 7-1. An example neural network with an input layer, a hidden layer, and an output layer. Nodes are often called neurons.

Much of the recent attention in the area of AIML is due to advancements in deep learning, which is a neural network with multiple interconnected hidden layers. Changing the type of nodes in a layer, how many nodes, how many layers, and how they are interconnected from one layer to the next changes the types of problems the algorithm is best suited to address and how (and these algorithms are given special names, e.g., deep neural network, convolution neural network, recurrent neural network). This is far outside of scope for our purposes!

These approaches often require large volumes of data to be trained, finding patterns that may otherwise be lost to simpler models. They are particularly suitable where there is uncertainty in the physicial or virtual model of how elements of a system interact (e.g., when there is a lot of noise). When underlying features and interactions are well understood, other modeling approaches can be more suitable. Regardless of approach, a knowledgeable data scientist is needed to ensure success in the training and application of algorithms, as there are many nuances to consider, as we'll see next.

Technical Challenges

Supervised algorithms are trained and validated on real data, producing a model; unsupervised algorithms usually have parameters that need be set, often chosen based on validation with real data, which we can describe as defining a model. Those models are then applied to new data, possibly even personal data. The basic process is shown in Figure 7-2. Statistical properties and validation determine the boundaries of expected operation of the model. To compare between algorithms or modeling

approaches, the same data and parameters should be used, especially those relevant to the problem at hand, which is often lacking from academic literature and can elicit significant debate between practictioners who favor one approach over another.

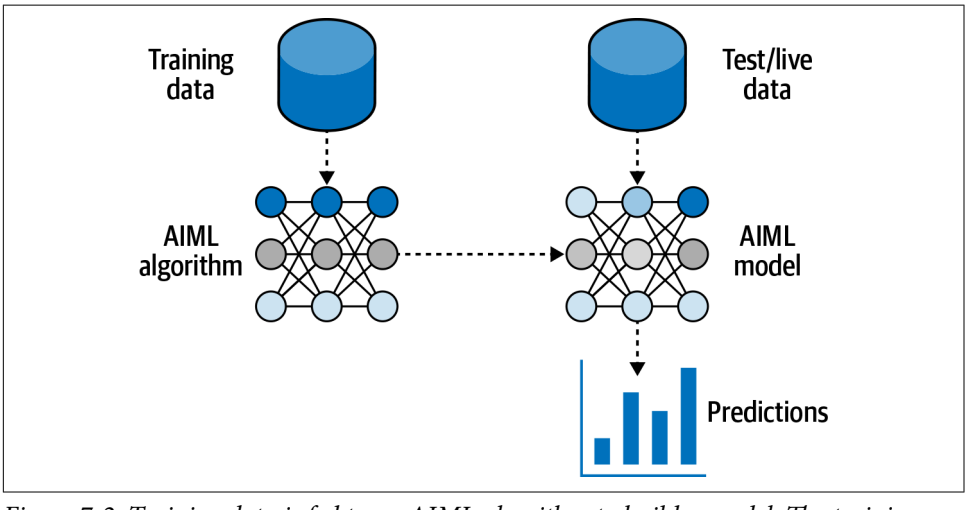


Figure 7-2. Training data is fed to an AIML algorithm to build a model. The training output can be used to tune the model. However, independent test data is fed to the model to validate its performance before live data is used.

There are many technical challenges with building or training an effective model, and ensuring that it performs as expected now and into the future. The use of models in practice also comes at a cost, sometimes refered to as *technical debt*, in terms of maintenance required to ensure that they continue to operate as intended. The release-and-forget approach is unrealistic. In other words, the job isn't done once a model is deployed. Some things to look out for (in alphabetical order):

Attacks
: Outsiders may "stress test" a model to see how it performs in extreme cases. They may also attempt to "confuse" a model by feeding it false information so that it produces incorrect or unintended outputs.

Bias
: One outcome or group of individuals may be systematically favored over another outcome or group of individuals. Statistical learning attempts to determine properties of algorithms and identify those that are unbiased estimators, but we are also referring to the broader issue of using data that is itself biased and will therefore result in biased outputs.

Confounders
: The association between input and output may be confused by some other factor that influences both input and output. Think of it as the model being focused on the wrong things that aren't really causing what is trying to be modeled.

Drift
: Variation in performance over time should be expected. Forecasting or prediction will have boundaries in which future outputs are expected to fall, based on the data that was used to train a model. Data input to a model will change over time (e.g., due to changes in data subjects) and the model's performance will therefore drift over time as well.

Generalization
: The output will be correct for the range of input that may be used. Training or validation data needs to be representative of the input data the model will be applied to, so that the output is expected. A model needs to be general enough to capture the variation that will be met when applied in practice. Models can also start as general and require an adaptation phase to become more specific for their circumstances.

  - Overtrained (or overfit) models are overly focused on the training data, capturing every variation to the point that this is all that is output correctly. Deviations result in unstable output. The degrees of freedom available to the model are likely much *higher* than required for the volume and complexity of the data. Think of it as drawing a curve through every data point provided for training, no matter how complex.
  - Undertrained (or underfit) models are lacking in their ability to capture the variation in the input, and the output is stable but off the mark. The degrees of freedom available to the model are likely much *lower* than required for the volume and complexity of the data. Think of it as drawing a straight line through a large cluster of data points.

The reality is that AIML models usually work better when they're given very specific tasks to accomplish. Oversight is needed to ensure that the training is appropriate to the task, that the models make practical and scientific sense, and that they will continue to operate effectively. Trust, but verify.

The sophisticated systems we often marvel over, such as a bipedal robot running through the snow, are actually a collection of many models working together, each performing a specific task that can be evaluated and understood (and models that depend on other models can be very complex, with hidden effects and dependencies

that make all of the above even more difficult to manage). But let's bring the conversation back to privacy and consider how algorithms can impact trust.

Algorithms Failing on Trust

The concerns that models, and deep learning models in particular, have raised are varied. They have a lot to do with the complexity of these modeling approaches and the technical challenges we described. Remember that regulators are human and they will apply a human lens to the uses of AIML. To name a few concerns that have been raised by regulators:

- The general opaqueness of many algorithms, or combinations of algorithms chained together in a production system, and being able to explain or justify a model or results (e.g., the creation of a "black box")
- Bias due to the choice of training data, or lack thereof, resulting in potentially false predictions or unfair treatment (e.g., due to demographic or socioeconomic factors)
- Their ability to distinguish complex patterns (predictive discrimination) that can seem uncanny and therefore creepy
- Dubious and poorly justified claims in areas that may be perceived as creepy (e.g., predicting sexual orientation)
- Broad application in areas that may have influence on people and their lives (e.g., making decisions on their behalf or about them)

The above concerns are not unique to neural networks or deep learning, although models from this class of algorithms have received a great deal of attention. To help understand these concerns better, we can group models into two classes:

Fixed models
: Models that produce the same output for the same input and are therefore repeatable (not withstanding manual updates to the model based on new training data or parameter adjustments)

Adaptive models
: Models that "retrain" due to an automated feedback mechanism that modifies the model based on the inputs provided, producing potentially different output for the same input it was originally validated on

Fixed or adaptive models can produce failures of trust for different reasons. A fixed model may be given input, or a series of inputs, that were unanticipated, leading to an undesirable result (fool me once, shame on you). If the evaluation was robust, undesirable results should be rare, and corrective action should be taken to ensure it doesn't happen again (fool me twice, shame on me). Failures from an adaptive model,

however, may be harder to anticipate and avoid, requiring more stringent guardrails to keep an algorithm on the right path. In both cases, monitoring of some kind is required.

Some have suggested that it is an ethical failure not to anticipate the range of input that may be provided to a model, even when an algorithm is adaptive. In cases where there is potential harm, it may even be perceived as reckless. Many ethical frameworks have been proposed as a result of these concerns.

Let's consider examples where some the above concerns have played out. We could call this section *#algorithmfail*, and there are many examples to choose from, and many complicating factors that we can only touch on.[5] Many failures of algorithms are simply funny or an inconvenience. But some will be a failure of trust, and their severity will be judged by how invasive and offensive they are. Some may consider our use of the term "fail" a bit harsh, so we should be clear that our intention is not to criticize but to inform. The failures we are describing are those of trust. Even a perfectly accurate and precise system can have failures of trust.

Rogue chatbot

In 2016, Microsoft researchers launched an adaptive chatbot on Twitter, called Tay, to see how it would interact with users. They had previous experience with an adaptive chatbot, and claimed to have stress-tested the chatbot in advance of the public launch. Context matters, though. Tay was built using training data that was public but cleaned and filtered. The well-behaved training interactions, which perhaps even included attempts to misbehave, were insufficient for the onslaught of Twitter users that fed Tay rather inappropriate input.

Tay was asked to repeat tweets that were discriminatory and rude. But Tay was also adaptive, and eventually created its own comments, based on all the inappropriate input provided by users. These comments, both repeats and self-generated, were grossly out of line with what Microsoft or anyone else would deem appropriate. While some people found this amusing, it created an unfortunate amount of bad press for Microsoft. The company had to acknowledge the failure and the need to develop appropriate guardrails in the future.[6] Ultimately, Tay was not immune from threats and did not act in the interests of all potential Microsoft users.

5 For a fun read on what can go wrong with AIML, read Janelle Shane, *You Look Like a Thing and I Love You: How Artificial Intelligence Works and Why It's Making the World a Weirder Place* (New York: Voracious, 2019), *https://oreil.ly/e1sTe*.

6 Peter Lee, "Learning From Tay's Introduction," Official Microsoft Blog, (March 25, 2016), *https://oreil.ly/ShJm0*.

To help prevent similar mishaps, Microsoft has since developed tools that it provides to developers on some of its platforms, such as tools that detect offensive text. And Microsoft introduced guidelines for developing conversational AI that the company hopes will ensure trust.[7]

There are far more concerning applications of models and algorithms than a conversational chatbot. But that such a mundane application went rogue has reminded some people of the potential pitfalls of AIML, raising concerns around its application in areas where harms could be much greater. That being said, one would hope that in areas where impacts are potentially more disastrous there would be greater control and oversight. It's the areas in between that nudge our behaviors or have unseen consequences where there's a lot of uncertainty regarding the use and impact of AIML.

Predicting criminality

One area where the use of algorithms can have serious implications to individual lives is that of predicting criminality. This is no longer limited to the realm of science fiction. On the surface it would seem beneficial to have a system that can support rational, evidence-based decision making, but the devil is in the details. And there are many details to consider, details that are often outside the realm of comprehension for most people who don't have in-depth training in both data science and its evaluation, not to mention data ethics.

One example that has received considerable attention is a commercial system that claims to predict whether a criminal will offend again. It has received considerable attention because of concerns regarding perceived accuracy and bias, such as predicting that a white male with a criminal history is less likely to offend again than a black female without a criminal history. Many experts perceive it as an algorithmic black box without appropriate external review and oversight, given the degree of impact it may have on decision making in the criminal justice system in the US.[8]

Imagine the impact this system can have in sentencing or parole. It raises serious ethical concerns that are at the heart of many of the concerns with AIML. Overreliance on an algorithm can have drastic consequences, as there will often be context that is missing or inadequately factored into a model. Experts are needed to parse out details and determine whether the algorithm is a suitable decision-support tool for the given context and to what degree. Someone needs to be accountable to answer questions, tease out additional contextual details, and ensure that an algorithm does not become

7 Lili Cheng, "Microsoft Introduces Guidelines for Developing Responsible Conversational AI," Official Microsoft Blog, November 14, 2018, *https://oreil.ly/yC8Pj*.

8 Julia Angwin et al., "Machine Bias," *ProPublica*, May 23, 2016, *https://oreil.ly/BmeNe*.

a blind "argument from authority" (i.e., a logical fallacy which, in this case, would be to believe the algorithm is right simply because it uses AIML).

This example is of a fixed model, not an adaptive one, yet there are still many concerns raised. Some progress has been made in getting external validation of the algorithms used, and some academics are working on algorithms that provide more transparency and insight into how algorithms come to their conclusions. Nonetheless, there is a lot of work to be done to provide more honest views into what factors algorithms are relying on in a given context, so that experts can evaluate *why* they produce certain results, and how they they *should* be used as a decision-support tool (including to what degree, depending on the context).

Analytical tools have also been applied to predicting crime more generally. There are many aspects that need to be considered, and corrected for, for the application of these tools to be appropriate and effective. How algorithms are used in decision support, and in what areas, is as important as how accurate and precise they are.[9]

So where does that leave us? We have complex AIML algorithms generating complex models, with many technical challenges to address, and examples of algorithms that have failed on various principles of trust that are expected of data subjects and the public at large. From these many considerations, we can hopefully draw lessons learned and conceive of principles that will help in the responsible use of AIML. Which is exactly what we do next.

Principles of Responsible AIML

There are many frameworks for big data and AIML that provide principles for responsible use. They are usually a reformulation of the privacy principles found elsewhere (such as fair information practice principles), with elements of trust included and insights specific to AIML. Rather than present you with any one of these specific frameworks, or attempt to merge them into a single framework, we will remind you of the privacy-engineering triad that was introduced in "Requirements Gathering" on page 64 and apply it to responsible AIML:

Predictable
: Expected interactions and outcomes. This is a tall order in the case of AIML, given the possibility of attacks on a model or system, potential bias from the training data, confusing associations between input and output, results drifting over time, and a model's applicability to different data and situations. Not to

9 Mara Hvistendahl, "Can 'Predictive Policing' Prevent Crime Before It Happens?," *ScienceMag*, September 28, 2016, *https://oreil.ly/9sT-i*.

mention that these interactions and outcomes need to be explainable and well understood if they are to be expected in the first place.

Manageable
: Granularity of administration. In order to manage a system built on top of AIML models, those models need to be maintainable (including possible corrections needed because they affect predictability). Simple, explainable models that can be adjusted because they are well understood and "tuneable" will be easier to manage than a black box.

Disassociated
: Level of identifiability. Anonymized data can be used for training and validating AIML models. Although it is fair to be concerned with the use of AIML to re-identify data subjects, the Five Safes described in Chapter 3 should be used to address these concerns. Remember, AIML models are usually trained for a specific task, which should be defined up front long before access is granted to data.

Combining the privacy-engineering triad with the foundations of trust described at the outset of this chapter makes for a very powerful framework for the safe use of data. Admittedly they cover much of the same content, at least conceptually, but the privacy-engineering triad is probably more technically focused.

The foundations of trust add a tone that is perhaps lacking from the somewhat more technical view of privacy, and puts more focus on privacy as a societal good. From this emerge other aspects that are found in some frameworks on responsible AIML, such as aligning on values, needs and benefits to people or society, being progressive, and making algorithms a force for good. Clearly some of these stem from concerns about where AIML may be applied in areas that could lead to significant harm to society.

While there are researchers attempting to build technology-enabled processes that they hope will enable responsible AIML, human oversight continues to be required. That means a degree of governance and oversight, as we describe next.

Governance and Oversight

A great deal of what we've described thus far requires governance and tools to ensure proper oversight. There is simply too much for a lone data scientist or even a team of data scientists to consider on their own, and they probably need both a sanity check as well as guardrails to ensure that their own experience (narrow or broad) does not bias decision making. Oversight requires coordination and ongoing review of the many aspects we've described already. The degree of oversight will vary based on the degree of impact and the potential invasion of privacy.

We can learn a lot about governance and tools from the field of health research. We'll focus our attention on two different aspects in this section, ethics and data monitoring, that can be the basis for developing resources that enable teams to safely use data, and the basis for oversight committees. They are usually described as committees or councils, but organizations need to be agile, so we want to ensure that the aspect of developing resources is also considered. Resources need to be developed or evaluated with a wide lens of expertise to ensure that they provide value and can be effectively integrated into an organization so that they're actually used.

If committees are formed, members will most likely require some training in processes, and may need protection from liability for the role they serve if the application area warrants it. Independence from the organization is often encouraged, which may be more easily justified in particularly sensitive and impactful areas, to ensure that a committee can make credible and objective recommendations. In commercial settings, independence can be difficult to achieve given commercial and competitive sensitivities. But the threat of reputational harm can often be sufficient to ensure that internal committees are provided sufficient authority to be effective.

Privacy Ethics

Ethical oversight is needed to ensure that data and insights are applied in such a way that they can meaningfully engender individual trust that data will be used safely without causing harm. Using anonymized data instead of identifiable information itself mitigates some concerns. Although ethical considerations in the use of data are largely orthogonal to the subject of anonymization, considering the potential invasion of privacy also helps mitigate some ethical concerns by ensuring that proportional measures are applied. Transferring obligations and oversight to organizations that data or models are being shared with should also be considered.

To ensure that trust is maintained, a committee of relevant experts can be formed to advise on the use of AIML models in decision-making efforts from personal data. This can help avoid a spectrum of negative legal, reputational, and regulatory actions. Members could include a layperson representing service users, a privacy professional, a lawyer, someone representing the business unit, a data analyst, an ethicist, and someone from a public relations department. The last role is important because some analytics protocols may be acceptable or not based only on the reputational or brand harm that they could cause. Depending on the size of the group, a single person can fulfill multiple roles.

Such a privacy-ethics committee would advise the business units about the risks from the development of certain AIML models or making certain decisions from models. Having this type of process also ensures that analytics objectives and approaches are documented and reviewed before they are included into products and services. While a privacy-ethics committee would have responsibility to safeguard the interests of

individuals or groups that AIML models are to be trained and applied to, its role is somewhat static once the decision to develop and apply AIML models has been made for a specific context. The committee may be informed if safe use comes into question, but that is likely to come from the monitoring of that use, and for that, another oversight mechanism is likely needed.

Data Monitoring

The primary purpose of data monitoring is to safeguard the interests of individuals or groups that AIML models are trained on and applied to, and to ensure the integrity and credibility of a service or application area based on the *continuing validity and merit* of the AIML models. This will involve reviews of accumulating data to ensure the privacy-engineering objective of predictability (including all of the technical challenges we described). Access to timely, reliable, and readily interpretable insights about emerging evidence is therefore required for the monitoring to be effective. Another way to describe this would be as AIML stewardship.

If a committee is formed, it will act in an advisory role to the leadership of the organization accountable for the AIML models. Members will need to be experts in the application area with sufficient domain knowledge of AIML to understand risks and benefits. If external experts are used, it will be critical to ensure confidentiality and protection of intellectual property, otherwise there will be no trust in providing them with operating data or insights. Internal voting members with external nonvoting members providing guidance is a good model to employ. The committee could recommend whether to continue, modify, or stop using AIML models based on the impact on individuals or groups and the privacy-engineering objective of predictability.

Experts in AIML and their many challenges are needed to provide timely and accurate monitoring data. To be effective, they will need to generate useful insights, possibly in the form of summary dashboards with critical tables and graphics that the committee can use in decision making. These experts are therefore a liaison between the data monitoring committee and the operating data itself, and are therefore members of the organization, possibly attending committee meetings and providing additional insights when requested. Even without a committee, someone will need to be accountable for decision making, and therefore the role of experts in AIML is likely to remain an independent function. AIML experts will also be needed to determine the root cause of AIML issues, and to fix or improve AIML models.

Final Thoughts

The area of AIML has rapidly evolved, gaining the attention of regulators and citizens alike. Some are fascinated by the seemingly infinite breadth and application of AIML, whereas others are terrified. The safe and responsible use of data is necessary to ensure that trust can be engendered and maintained. Trust needs to be earned and risks mitigated to ensure that trust is not lost. Of course nothing in life is free, and there's no such thing as zero risk. But there's lots that can be done through privacy engineering and the foundations of trust.

Anonymization enabled through the Five Safes is an effective tool to mitigate risks and responsibly share and use data. But the application of models derived from anonymized data can still impact people, whether they were in the orignial data used to train an AIML model or not. The application of AIML models to data collected about individuals or groups needs to be done with an eye toward ethical boundaries that are best addressed through the appropriate governance tools and oversight mechanisms.

There are many technical challenges with the application and use of AIML models, and many ways for trust to be unintentionally eroded. There is, however, tremendous potential to improve our lives through the use of technology, and therefore tremendous incentive to get it right. More than simply "do no harm," we can "be loyal" and improve lives through the safe and responsible use of AIML. As we've said already in this book, technology is an important enabler in the responsible sharing and use of data and insights, but governance is equally as important as we aim to deliver trust to stakeholders. We hope this book will help you in your journey to ethically and efficiently build your anonymization pipelines!

Index

E

F

G

H

I

J

L

M

N

O

P

R

S

T

U

V

About the Authors

Luk Arbuckle is Chief Methodologist at Privacy Analytics, providing strategic leadership in how to responsibly use and share data, and innovation into privacy-enhancing technologies and methods. He draws from an extensive background in statistics, mathematics, and engineering, as well as industry and regulatory experience. He engages senior decision makers in solving real-world problems in business and privacy engineering.

Luk was previously Director of Technology Analysis at the Office of the Privacy Commissioner of Canada, leading a highly skilled team that conducted privacy research and assisted in investigations when there was a technology component involved. Before joining the Office of the Privacy Commissioner, he developed de-identification methods and re-identification risk measurement tools, participated in the development and evaluation of secure computation protocols, and led a top-notch research and consulting team that developed and delivered data anonymization solutions.

He is coauthor of the book *Anonymizing Health Data* (O'Reilly), as well as numerous papers, guidance documents, and patents. Previously Luk did graduate and industry research in the area of digital image processing and analysis, and in the area of applied statistics. He has been awarded numerous scholarships and bursaries, including a Canada Graduate Scholarship doctoral award.

Dr. Khaled El Emam is a senior scientist at the Children's Hospital of Eastern Ontario (CHEO) Research Institute and director of the multidisciplinary Electronic Health Information Laboratory, conducting applied academic research on synthetic data generation methods and tools, and re-identification risk measurement. He is also a professor in the Faculty of Medicine (Pediatrics) at the University of Ottawa, Canada.

Khaled is the cofounder and CEO of Replica Analytics, a company focused on the development of synthetic data to drive the application of AIML in the health care industry. He is also the founder and, until the end of 2019, was the General Manager and President of Privacy Analytics, acquired by IMS Health (now IQVIA) in 2016. He currently invests, advises, and sits on the boards of technology companies developing data protection technologies and building analytics tools to support health care delivery and drug discovery.

He has been performing data analysis since the early 1990s, building statistical and machine learning models for prediction and evaluation. Since 2004, he has been developing technologies to facilitate the sharing of data for secondary analysis, from basic research on algorithms to applied solutions development, that have been

deployed globally. These technologies addressed problems in anonymization and pseudonymization, synthetic data, secure computation, and data watermarking.

He has (co-)written and (co-)edited multiple books on various privacy and software engineering topics. In 2003 and 2004, he was ranked as the top systems and software engineering scholar worldwide by the *Journal of Systems and Software* based on his research on measurement and quality evaluation and improvement.

Previously, Khaled was a Senior Research Officer at the National Research Council of Canada. He also served as the head of the Quantitative Methods Group at the Fraunhofer Institute in Kaiserslautern, Germany. He held the Canada Research Chair in Electronic Health Information at the University of Ottawa from 2005 to 2015, and has a PhD from the Department of Electrical and Electronics Engineering, King's College, at the University of London, England.

Colophon

The animal on the cover of *Building an Anonymization Pipeline* is the Eurasian jay (*Garrulus glandarius*). This bird can be found in forested areas and near human settlements across much of Europe and Asia.

The plumage of the Eurasian jay is primarily a light, reddish brown. Their crests are white with black speckles, and they have black mustache stripes that extend down from the ends of their black beaks. Their wings, which have bright blue bands and a red triangle where the wing attaches to the bird's body, are also primarily black, save for a white band visible during flight. Interestingly, the feathers of the Eurasian jay reflect UV light.

Eurasian jays primarily reside in trees and only occasionally forage on the ground. They are typically rather sedentary, though will migrate to other areas for food. Opportunistic omnivores, Eurasian jays will eat whatever food is available to them, although acorns make up the majority of their diet. These birds display complex social behaviors, including a characteristic kind of play that may be related to their fleeing from hawks, their primary predators.

Eurasian jays use a wide range of visual displays and calls for communication, and are are prodigious mimics: their song may include all the sounds they have ever heard, including crying babies, the songs of other birds, and even the sounds of lawn mowers. The current conservation status of the Eurasian jay is "Least Concern." Many of the animals on O'Reilly covers are endangered; all of them are important to the world.

The cover illustration is by Jose Marzan, based on a black and white engraving from *British Birds*. The cover fonts are Gilroy Semibold and Guardian Sans. The text font is Adobe Minion Pro; the heading font is Adobe Myriad Condensed; and the code font is Dalton Maag's Ubuntu Mono.