Privacy Enhancing Technologies:

Categories, Use Cases, and Considerations

FINTECH EDGE SPECIAL REPORT

FEDERAL RESERVE BANK
OF SAN FRANCISCO

Acronym List

Aadhaar	Unique digital identity number for Indian citizens
AML	Anti-Money Laundering
BCR	Binding Corporate Rule
CCPA	California Consumer Privacy Act
EU	European Union
FinCEN	Financial Crimes Enforcement Network
GDPR	General Data Protection Regulation
GLBA	Gramm-Leach-Bliley Act
IP	Internet Protocol
IT	Information Technology
KYC	Know-Your-Customer
MPC	Multi-party Computation
PCI SSC	Payment Card Industry Security Standards Council
PET	Privacy Enhancing Technology
PHI	Protected Heath Information
PII	Personally Identifiable Information
NIST	National Institute of Standards and Technology
NPI	Non-public Personal Information
SCC	Standard Contractual Clause
SSN	Social Security Number
ZKP	Zero Knowledge Proof

Relevant Laws and Frameworks

California Consumer Privacy Act
General Data Protection Regulation
Gramm-Leach-Bliley Act Privacy Rule
Gramm-Leach-Bliley Act Safeguards Rule
Privacy Shield Framework
United Kingdom Data Protection Act of 2018

Authors

Kaitlin Asrow, Fintech Policy Advisor
Spiro Samonas, Senior Risk Specialist

Publication Date

June 1, 2021

The views expressed in this publication are solely those of the authors and do not necessarily represent the position of the Federal Reserve Bank of San Francisco, the Board of Governors of the Federal Reserve System, or any other parts of the Federal Reserve System.

Contents

Introduction

Information, and the data[*] that underpins it, is an essential resource that provides insights to individuals, and businesses, and enables governments to operate and coordinate both locally and globally. Information can provide unique insights to improve welfare, drive business innovation, and help countries navigate crises, such as the COVID-19 global pandemic. Unfortunately, the underlying data that are used to create information and insight, can also be breached, or misused, and require significant resources to manage efficiently and securely. The global cybersecurity landscape is rife with a wide range of threats including nation states, organized crime, and malicious insiders, among many others. Compounding this, opaque data collection, processing, sales, and transfers across entities and jurisdictions may create unique, and yet known, privacy risk for individuals.

Balancing the opportunities and risks of collecting and using data is an ongoing policy debate in the United States and around the world. Although, governance and legal frameworks are being developed in the areas of cybersecurity, data rights, competition, and beyond, entities continue to use data with fragmented guidelines, and oversight, around security, privacy, and appropriate use.[1]

There are opportunities to leverage technology and innovation alongside policy to enable more secure, and demonstrably private data collection, processing, transfer, use, and storage.

The use of technical systems and tools to protect data, while still leveraging it to create value, is not new. Data anonymization and encryption are long-standing tools, but as more data are collected and data breaches increase,[2] they no longer offer the protection they once did.[3] For example, techniques and laws originally focused on protecting identified data, or information that was directly associated with an individual, such as a name or Social Security Number (SSN) (referred to as direct identifiers). However, research demonstrated that data types beyond direct identifiers, such as demographics,[4] could also be used to uniquely identify individuals. As a result, the terminology in laws and standards has shifted over time from “identified” data, to “identifiable” data. The expanded category of “identifiable” information remains challenging to define, though.[5] This evolving classification is just one example of how outdated approaches to preserving privacy, as well as the laws built around them, may fall short in offering the protection expected, and needed.

The rapid and widespread adoption of digital services that depend on data, and cater to every facet of daily life, further underscores the importance of exploring new technical approaches to preserving privacy and confidentiality in conjunction with policy evolution. This report will focus on these technical approaches; specifically, **the report will define and categorize a set of tools and techniques classified as privacy enhancing technologies (PETs) that help maintain the security and privacy of data.** The purpose of this work is to provide a panoramic view of PETs to enable regulators, policymakers, and business leaders to enhance supervision and decision-making in the areas of privacy and data protection. This report will also discuss the use cases and maturity of PETs, and any challenges or considerations around their implementation.

[*] *The terms “data” and “information” are used interchangeably in this report, but the authors acknowledge that these terms can have distinct connotations.*

The mandate of the Federal Reserve Bank of San Francisco is rooted in financial services; however, given the evolving data collection and processing activities that cross traditional sectoral boundaries, this report seeks to provide a nuanced and neutral examination of PETs for a broad policy and regulatory audience.

Overview

Privacy enhancing technologies are a group of systems, processes, and techniques that enable processing to derive value from data, while minimizing the privacy and security risk to individuals.[6] While innovations such as machine learning and quantum computing[7] are being leveraged and explored, PETs cannot be defined by a single technology. Additionally, since this is an emerging category of tools, there is no consistent definition of what constitutes a PET. Although PETs may be used to comply with privacy-focused legislation, it is important to understand that using these systems, processes, and techniques do not ensure privacy or compliance. PETs are tools that can be used in tandem with governance and operation systems.

PETs can be used by entities or directly by individuals. This report will focus primarily on systems, processes, and techniques available to entities that collect, process, and use data. The report focuses on entities instead of individuals due to their larger scope of data activities, and data management needs. The terms entities, firms and organizations will be used interchangeably. Laws like the European Union's General Data Protection Regulation (GDPR) refer to these entities as "data controllers" and "data processors."[8] Entities may be motivated to adopt PETs to support information security and confidentiality, comply with regulation, and to differentiate their products for consumers looking for increased privacy protection. It is important to distinguish between upholding privacy as an individual right, and confidentiality, which is a fundamental component of information security practices. Confidentiality is an agreement between an entity and an individual to restrict access to information to only authorized parties. The use of PETs by entities can support both broader privacy goals and confidentiality agreements between entities and their customers.[9]

While this report will not focus on privacy systems and techniques that individuals can use directly, it is important to note that broader landscape. Examples of PETs targeted at individuals are privacy-protective browsers or TOR networks[10], as well as consumer-facing dashboards that enable the implementation of individual data rights.[11]

While PETs used by entities extend beyond traditional compliance systems, recent privacy-focused laws such as GDPR and California's Consumer Privacy Act (CCPA)[12] have been important catalysts for PET uptake. GDPR created several new responsibilities around data, which drove investment in updated technical systems and business processes.[13] In addition to this, GDPR and CCPA both extended the categories of information that are deemed sensitive, and redefined the threshold that entities must meet to prevent the re-identification of anonymized data.

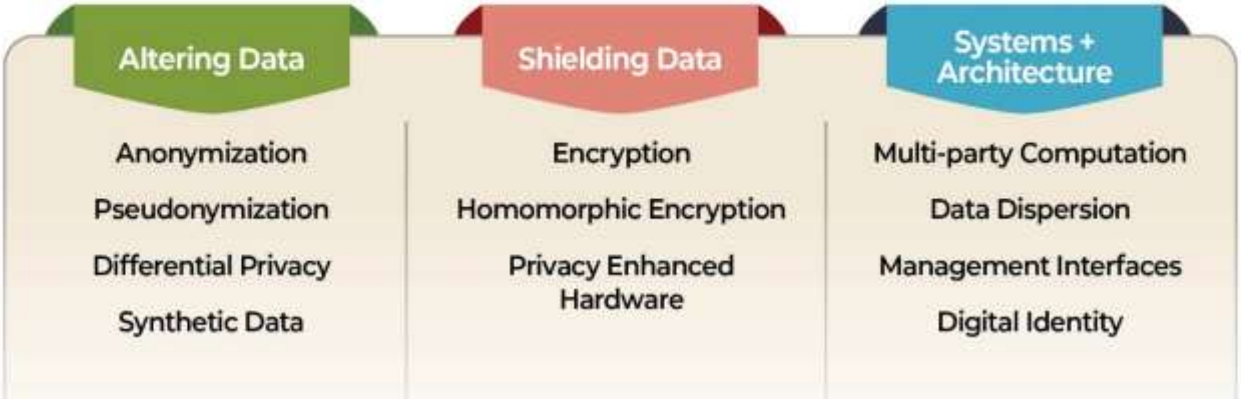
PETs are particularly important in sectors that rely on the extensive collection and use of sensitive data, such as financial services and healthcare. In financial services, requirements such as Know-Your-Customer (KYC),[14] credit reporting, money laundering detection, and fraud mitigation[15] drive the collection of data that is matched to specific individuals. Pooling this data together can help identify financial crimes and protect individuals from financial loss. Similarly, healthcare information including patient data can be pooled together for research, drug-development, and public health. The COVID-19 pandemic has highlighted the value of information in public health, from contact tracing to vaccine development. Given the importance and particularly sensitive nature of data in healthcare, this sector has been a leader in exploring the potential of PETs.[16] Combining data can also create business opportunities beyond regulatory requirements and research, such as the development of new products, tailored services, software testing, and more. The aggregation of customer data to innovate in financial services has led to the growth of many new financial technology, or fintech, firms; as more entities engage with this data it is increasingly important to consider its privacy and confidentiality while benefitting from its use.

Beyond compliance and facilitating the use of data, PETs could even help global coordination. A recent decision by the European Union (EU) supreme court invalidated an agreement between the EU and the United States (US) called the Privacy Shield, which enabled corporations to transfer data collected in the EU to the US for processing and storage.[17] The EU supreme court invalidated the Privacy Shield because it determined that the US did not have data protections equivalent to those in the EU. The invalidated Privacy Shield focused on legal and governance protections such as Standard Contractual Clauses (SCCs) and Binding Corporate Rules (BCRs), but the judgement determined that these must be supplemented with technical measures. This is particularly relevant to PETs because this group of technologies, systems, and techniques can help verify and strengthen legal and governance rules that enable global relationships.[18]

Types of Privacy Enhancing Technologies

PETs contribute to privacy and data protection in a variety of ways. The first category of PETs are tools that alter data itself. These typically seek to disrupt or break the connection between data and the individual they are associated with. Another group of PETs focuses on hiding, or shielding, data, rather than altering it. Encryption is an example of this, since it changes the format of data, but is intended to only obscure it temporarily, rather than alter it permanently. Finally, there is a broad category of PETs that represent new systems and data architectures for processing, managing, and storing data. Some of these systems break apart data for computation or storage, whereas others provide management layers to track and audit where information is flowing and for what purpose.

Figure 1: Categories of Privacy Enhancing Technologies*



Altering Data	Shielding Data	Systems + Architecture
Anonymization	Encryption	Multi-party Computation
Pseudonymization	Homomorphic Encryption	Data Dispersion
Differential Privacy	Privacy Enhanced Hardware	Management Interfaces
Synthetic Data		Digital Identity

*These categories are based on the authors' analysis of the PET space. The authors acknowledge that there are multiple ways to group these technologies, techniques, and processes.

These different categories can be used together to create layered protection. For example, data can be altered through de-identification techniques, concealed through encryption, and processed using privacy-protective systems. Following a brief discussion of the challenges that are associated with PETs, the remaining sections of the report will describe different PETs and use cases within these categories.

Challenges

The concept of leveraging innovative technological systems and processes to preserve privacy while enabling the use of data is heartening for consumer protection and legal compliance, and it enables organizations to meet unique compliance and business sector-based needs. However, there are several challenges associated with the adoption of PETs.

The first challenge is the internal capacity and expertise within entities to deploy and manage PETs. This is a common issue across technical deployments that require specialized expertise, but since many PETs are not widely used yet, they can pose unique challenges. There is variability in the configuration needed to deploy PETs. Some techniques or systems may be able to function with limited support, while others need more oversight. Some technologies may also need to be used in conjunction with business enterprise systems, and therefore, require work to integrate and maintain those connections. Firms may depend on vendors to provide the technology or expertise the firms lack, but this gives rise to third-party risk in the form of vendor management for critical Information Technology (IT) systems. Some technologies, such as homomorphic encryption, also require significant computing power, which can increase costs.[19]

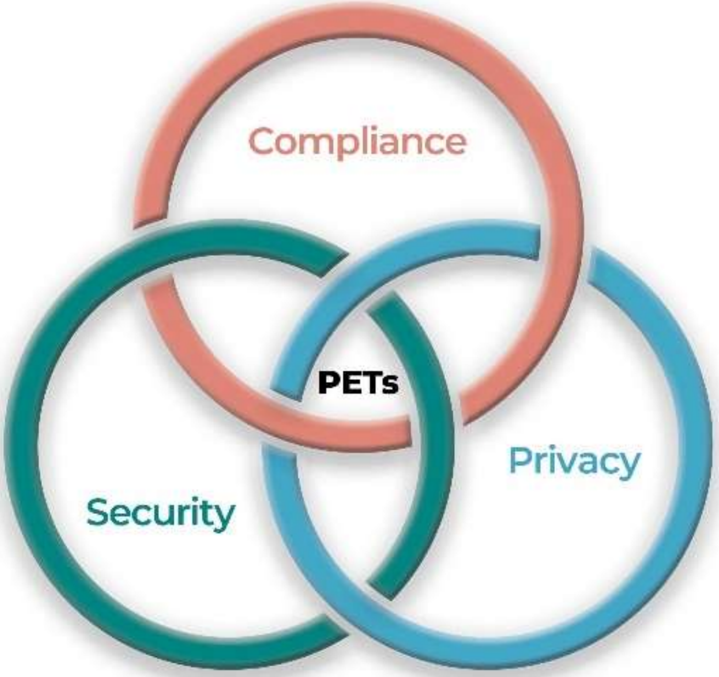
Another challenge is that PETs are in variable stages of maturity. While promising, some techniques and systems are still in early phases of development, and there is limited investment in ongoing research. Technology-driven firms with robust research and development funds are focused in this space,[20] along with pockets of academic work, but PETs as a category are not yet widely studied. This variability in maturity and research adds to the complexity of PET adoption and makes it harder for firms to determine which PETs are appropriate and what resources they need to deploy them. There are recent calls for the United States National Science Foundation to support research on this topic,[21] so research around PETs may accelerate in the future.

The use of PETs does not ensure that firms are automatically more privacy-protective, or in compliance with new laws. In many cases, even enhanced techniques can be reversed or compromised, therefore PETs still need to be treated like any technical implementation, with oversight and management around use, access, and security. There is also a risk that new approaches may be used for processing activities that would otherwise not be allowable under the law. It is important for entities to use PETs to enhance privacy and confidentiality, rather than to circumvent other requirements.

Additionally, while privacy preservation is considered positive, it may come into tension with new movements around data rights. If individuals request information about themselves, seek to port information to new providers, or would like data deleted entirely, entities need the capability to respond to those requests. Clarification has been sought around laws that provide for data rights because firms must associate data to individuals in order to comply but may prefer to use more robust and permanent de-identification techniques. Other laws make the development of these technologies challenging. For example, prohibitions against the re-identification of data make the testing of the strength of de-identification techniques difficult. Laws like the California Consumer Privacy Act (CCPA) [22] and the United Kingdom Data Protection Act of 2018[23] seek to remedy this by permitting reidentification to validate new privacy-preserving techniques.

There is also an ongoing debate about how to balance the use of PETs, such as the privacy-enhancing practice of end-to-end encryption, with the ability of law enforcement to use technical backdoors to collect data as part of criminal investigations.[24] End-to-end encryption means that as information is transmitted between devices, only the two end points will see readable data, while intermediaries, such as telecommunications providers that sit between end points, see unreadable data. Law enforcement argues that this restricts their ability to track criminals, while privacy advocates and companies feel that this is a basic step to limit unnecessary data collection and reduce privacy risks for large numbers of consumers.[25] This tension with criminal investigations also occurs in the banking sector. Detailed information is required to prosecute financial crime and there is a question as to how much granular identifiable information entities, and service providers, like cloud storage, should be able to provide.

A final challenge with certain PETs is a lack of incentive for businesses to implement them unilaterally. Technologies have been developed, but they may present an additional cost to implement or not be effective until a large portion of the market has adopted them. Examples of this include standardized formats and systems to enable multi-party computation, or networks that require collaboration such as market-wide digital identities.



These challenges highlight that despite their promise, PETs come with tradeoffs between usability and privacy. This further enforces the importance of using PETs in conjunction with policy and governance systems and frameworks.

Altering Data

Data de-identification is an overarching term that encompasses a variety of methods and tools used to obscure identifying characteristics within datasets.

Many data activities do not need to directly identify individuals or link them to associated data. For example, entities may only need to understand descriptive statistics, such as averages, or how customers are interacting with products as a group. For these use cases, there is an opportunity to permanently alter data to reduce the potential that it can be tied back to an individual. There are situations where it is necessary to link individuals to associated data; for instance, to satisfy KYC or Anti-Money Laundering (AML) requirements,[26] or to monitor for fraudulent activities. In these cases, there is a potential to use reversable de-identification or moderate how much of the data are altered. Even when direct identification is necessary, reversable de-identification may be used to increase security while data are being stored, or to enable privacy-preserving analysis to identify trends.

De-identification can also be applied to both direct identifiers and indirect identifiers. Direct data identifiers are data elements that uniquely identify an individual, such as first name, last name, home address, and Social Security Number (SSN). These types of identifiers are typically called Personal Data (under GDPR), Personally Identifiable Information (PII), Protected Heath Information (PHI) or Non-public Personal Information (NPI) in the context of different laws and regulations.[27] Indirect data identifiers are socioeconomic or demographic types of information, such as gender, race, age, religion, income, etc. Indirect data identifiers can only identify an individual when they are aggregated together. For example, there are many individuals who are age 54, but there may be only one who is also male, lives in a certain zip code, and earns a specific annual income. The treatment of indirect data identifiers varies significantly across different laws and regulations.[28] Notably, laws like GDPR and CCPA have extended the definition of PII to incorporate "identifiable data" including indirect identifiers such as Internet Protocol (IP) addresses.[29]

There are different approaches to de-identification, including removing pieces of information, replacing information, adding information, and creating synthetic or falsified versions of information.

Figure 2: De-identification Categories and Terminology

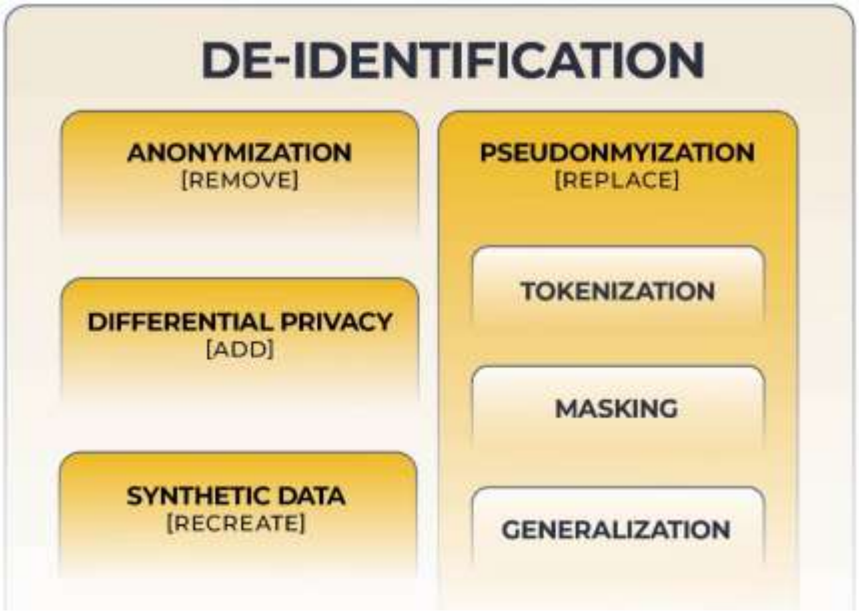


Figure 2 describes the terms used in the report for each of these processes. Each data de-identification method has different privacy and usability trade-offs, as well as a corresponding mix of regulatory and operational risks. Organizations need to understand these various methods, and assess which techniques are most appropriate to maintain the privacy and confidentiality of individual information. Important considerations include the difficulty of reverse engineering the techniques to reveal identifiable information, and how information continues to be protected, even after de-identification measures are in place.

Anonymization

The term anonymization is used broadly in privacy discussions and can mean de-identification generally, or refer to specific techniques. For the purposes of this report, anonymization is the practice of removing direct, or indirect, identifiers from a dataset. Anonymization is one of the original privacy-enhancing techniques used across industries and organizations, though it has since proved to be relatively insecure if multiple data sources are combined.

The removal of identifiers from data can be a manual or automated process. For automated processes it is important to confirm that the correct information is being removed. As discussed above, it is also important to consider when the removal of information is feasible for the use case.

The major drawback to anonymization is that information that was removed from datasets can be reconstructed by combining information from different sources. In the 1990's, Dr. Latanya Sweeney found that data that excluded direct identifiers (name, address, phone number, etc.) could still be used to identify individuals when combined with other databases, including those that are publicly available. Specifically, she found that 87% of the U.S. population could be identified using only their date of birth, gender, and zip code.[30] The ability to uniquely identify individuals by combining data from multiple sources has only increased as data generation and collection on a global scale has increased, and sensitive data sets have been exposed through data breaches.[31] New laws, such as GDPR, acknowledge this weakness and do not consider removing information a sufficiently strong form of de-identification.

Pseudonymization

Pseudonymization replaces a data element, such as an identifier, with a non-sensitive equivalent. This replacement can be reversible or irreversible. This is distinct from anonymization, which removes data elements entirely. Tokenization, masking, and generalization are different methods of pseudonymization. While these methods differ, one common characteristic is that the data used in-place of sensitive or identifiable information is fictitious, but usable. This allows for ongoing processing and analysis of the data after pseudonymization.

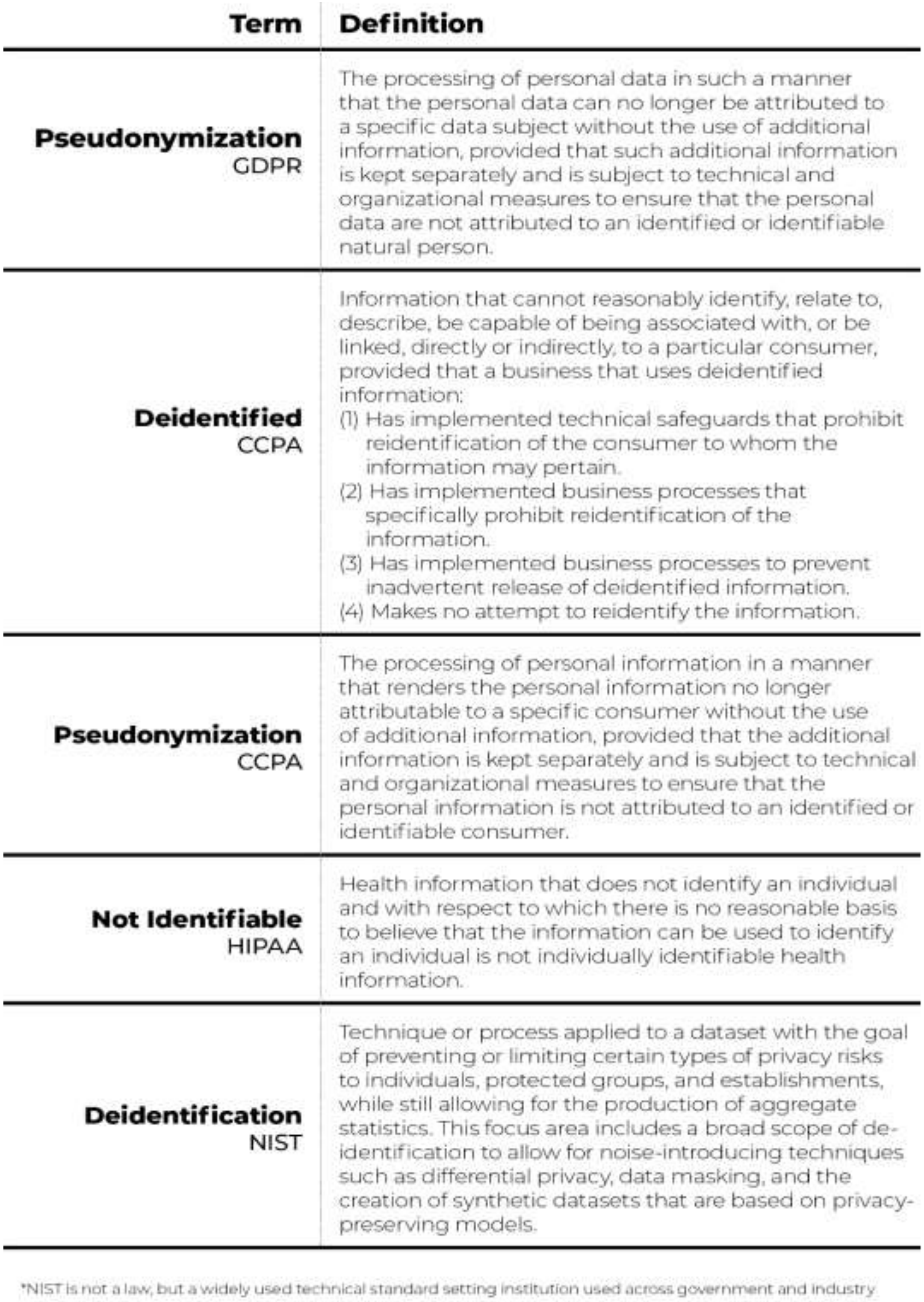
The terms deidentification and pseudonymization also appear in multiple U.S. laws and respected industry technical standards described in Figure 3.

Tokenization replaces a sensitive data element with a token. Tokens are typically random strings of numbers and letters, and in many use cases they are intended to be reversible. One entity will have a key that matches tokens to the true information they represent. A common use of tokenization is in payment systems. When a consumer swipes a credit card at a merchant, that number is replaced by a token, and only the token is stored. The card networks have the key to know which tokens are associated with an individual's true card number.[32]

Masking is like tokenization in that it replaces pieces of data with random strings of numbers and/or letters. A key difference between tokenization and masking is that masking is typically applied to data in-use and it is intended to be permanent. Data in-use represents data that one or more computer applications process on an ongoing basis. Processing may involve the creation, editing, deleting, viewing, or printing of data. There are forms of masking that are dynamic within systems, so as data are accessed, they are automatically altered before any human interaction or analysis, and there is not a key to reverse this kind of masking.

Generalization is a different technique that inputs a generalized term in place of a specific term. This is typically used on indirect identifiers. For example, instead of revealing a person's true age, a generalized database might assign each person to an age range, like "18-30" or "31-45". K-Anonymization is a related technique that sets a target for how many lines need to be generalized before no single individual can be differentiated from a large group because they all now have the same generic attributes.[33]

Figure 3: Concepts of De-identification in Laws and Standards



Term	Definition
Pseudonymization GDPR	The processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.
Deidentified CCPA	Information that cannot reasonably identify, relate to, describe, be capable of being associated with, or be linked, directly or indirectly, to a particular consumer, provided that a business that uses deidentified information: (1) Has implemented technical safeguards that prohibit reidentification of the consumer to whom the information may pertain. (2) Has implemented business processes that specifically prohibit reidentification of the information. (3) Has implemented business processes to prevent inadvertent release of deidentified information. (4) Makes no attempt to reidentify the information.
Pseudonymization CCPA	The processing of personal information in a manner that renders the personal information no longer attributable to a specific consumer without the use of additional information, provided that the additional information is kept separately and is subject to technical and organizational measures to ensure that the personal information is not attributed to an identified or identifiable consumer.
Not Identifiable HIPAA	Health information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information can be used to identify an individual is not individually identifiable health information.
Deidentification NIST	Technique or process applied to a dataset with the goal of preventing or limiting certain types of privacy risks to individuals, protected groups, and establishments, while still allowing for the production of aggregate statistics. This focus area includes a broad scope of de-identification to allow for noise-introducing techniques such as differential privacy, data masking, and the creation of synthetic datasets that are based on privacy-preserving models.

*NIST is not a law, but a widely used technical standard setting institution used across government and industry

The category of pseudonymization is very mature, and there are several market standards for these techniques. The Payment Card Industry Security Standards Council (PCI SSC), for example, considers tokenization an approved method for protecting payment card data.[34]

Unfortunately, like the challenge of anonymization described above, these methods don't automatically create privacy, even if sensitive, identifiable information such as credit card numbers are obscured. MIT researchers found that they could uniquely identify individuals based on the metadata, or descriptive details, associated with their card transactions despite the use of tokens.[35] Similarly, other reports argue that the wide use, and acceptance, of tokenization as a secure method of de-identification has created a "loophole"[36] for entities to still use data that may continue to pose a risk to individuals.

A challenge with masking and k-anonymization is that they work best with large data sets, since these techniques need enough individuals that generic attributes cannot differentiate any one person. Data that is protected using k-anonymity can still be vulnerable to attacks. For example, even if some attributes are generalized in a large data set, additional information could be added to pieces of the unaltered data to uniquely identify individuals.

Differential Privacy

Rather than removing or changing data elements to obscure identifiers, differential privacy adds random, additional data, or "noise." The goal of differential privacy is to add enough random, additional data so that real information is hidden amidst the noise. Differential privacy still allows for accurate analysis to be done on data in aggregate, because despite the added noise the combined data can provide accurate signals. One of the benefits of differential privacy is that re-identification by combining data sets is difficult because an attacker would not know which information was true. Another important element of differential privacy is its adjustability. The amount of "noise" that is added to the underlying data has a direct relationship to how difficult it is for an attacker to identify true information about individuals in the dataset. Because of this, entities can set a privacy "budget," and adjust the difficulty of re-identification. The tradeoff in inserting increasing amounts of noise into data sets is that the ability to identify true, averaged, trends or signals becomes more difficult.

Additional data, or noise can also be added to the dataset at any point. Some entities combine differential privacy with secure hardware environments, also called "on-device analysis." In these situations, noise is added before an entity ever receives the underlying information, and there is no reference to determine which data are fake. Noise can also be inserted after data are received by an entity. In these situations, entities can reverse the technique and reveal the true, identifiable data, if they have a key or reference that indicates which data were added.

Differential privacy is a mature form of deidentification and is used by market actors,[37] as well as the U.S. Census Bureau.[38]

As mentioned above, the main drawback of using differential privacy is that as more fake data are added, the accuracy of data analysis decreases. However, the "noise" is added on a per-person basis. When data from lots of people is aggregated together, the random noise becomes less of a factor. As a result, large data sets are needed to preserve the ability to obtain accurate aggregate statistics. Differential privacy is most useful for analysis and processing that does not need to be tied to individuals, or for entities that have access to those large data sets and the necessary technical resources.

Synthetic Data

Another form of altering data to protect privacy is the creation of entirely new, synthetic data. This is a step beyond pseudonymization, which replaces real data with altered data, or differential privacy, which inserts additional, fake information into real datasets. Synthetic data is commonly created through machine learning and mimics the characteristics of real-world data. The data are created by feeding real data into machine learning algorithms, which then identify characteristics and trends, and replicate those in synthetic information. Synthetic data may not always be classified as a PET because it can be used simply to increase the data available for training models, rather than an alternative to using sensitive data.

A main benefit of using synthetic data is that it can be customized to many different use cases,[39] while limiting the need to collect and store true information about individuals. Synthetic data can be used to train other models, or for testing new systems. For example, the United Kingdom's Financial Conduct Authority hosted "Tech Sprints," and provided firms with synthetic data that could be used to conduct experiments for product development.[40]

While these uses are exciting, the process of creating synthetic data is not yet mature. A lot is dependent on the quality of the algorithms used to create such data. A significant challenge is that poorly created synthetic data can be reverse engineered back to the original data that was used to produce it, therefore revealing it.

Another main drawback to synthetic data is its dependence on the quality of the original data that was used to train the machine learning systems. There may be bias in that original data, or it may not be representative for the intended use case, and synthetic data will replicate those issues. For example, synthetic data may not represent important outliers or unique characteristics of underrepresented groups.

The risks of synthetic data being reversed engineered back to original data, or to introduce bias, can be mitigated through use of high-quality models, as well as the verification and testing of those models. A final, and important issue regarding synthetic data is that it still requires the collection and processing of real data, although on a smaller scale.[41]

Shielding Data

Privacy enhancing technologies that shield data do not alter the underlying information; instead, they make data unintelligible or unusable at certain times to prevent unauthorized parties from accessing it. Conversely, when data are altered, as described in the section above, they can remain in that state indefinitely and are still intelligible, or readable, for processing.

For PETs that shield data, it is important to distinguish the three different states that data may need to be protected in: at-rest, in-use, or in-transit. There are different approaches to shielding data across these states because each state has different needs. For example, data at-rest are stored in a system at a point in time where no user is accessing or transmitting the data. Unused data can remain obscured or unintelligible. This is typically just an interim state, though, and data in-use are traditionally more difficult to shield because they need to be intelligible for processing. Advances in computing power has enabled new techniques such as homomorphic encryption, which allows for data to stay shielded through processing. Data in-transit are typically encrypted while moving; however, they may be unencrypted at both end points, and therefore more vulnerable.

Encryption

The most recognizable and common form of shielding data is encryption. Encryption is a reversible process that converts data to an unintelligible form called ciphertext; decrypting the ciphertext converts the data back into its original form (referred to as plaintext). The purpose of encryption and decryption is to allow only authorized users to access the plaintext using a key for conversion. Even if unauthorized users get access to the encrypted data, or ciphertext, they will not be able to read it without having access to the key.

Cryptographic algorithms, called ciphers, create random strings of characters to represent the underlying data. These algorithms have corresponding cryptographic keys, which are also strings of characters, and these are used together to change the underlying data into ciphertext. The longer and more complex cryptographic keys are, the harder it is for an adversary to crack the code and decipher the underlying plaintext data.[42] Encryption can use the same key to both encrypt and decrypt data, or different keys.

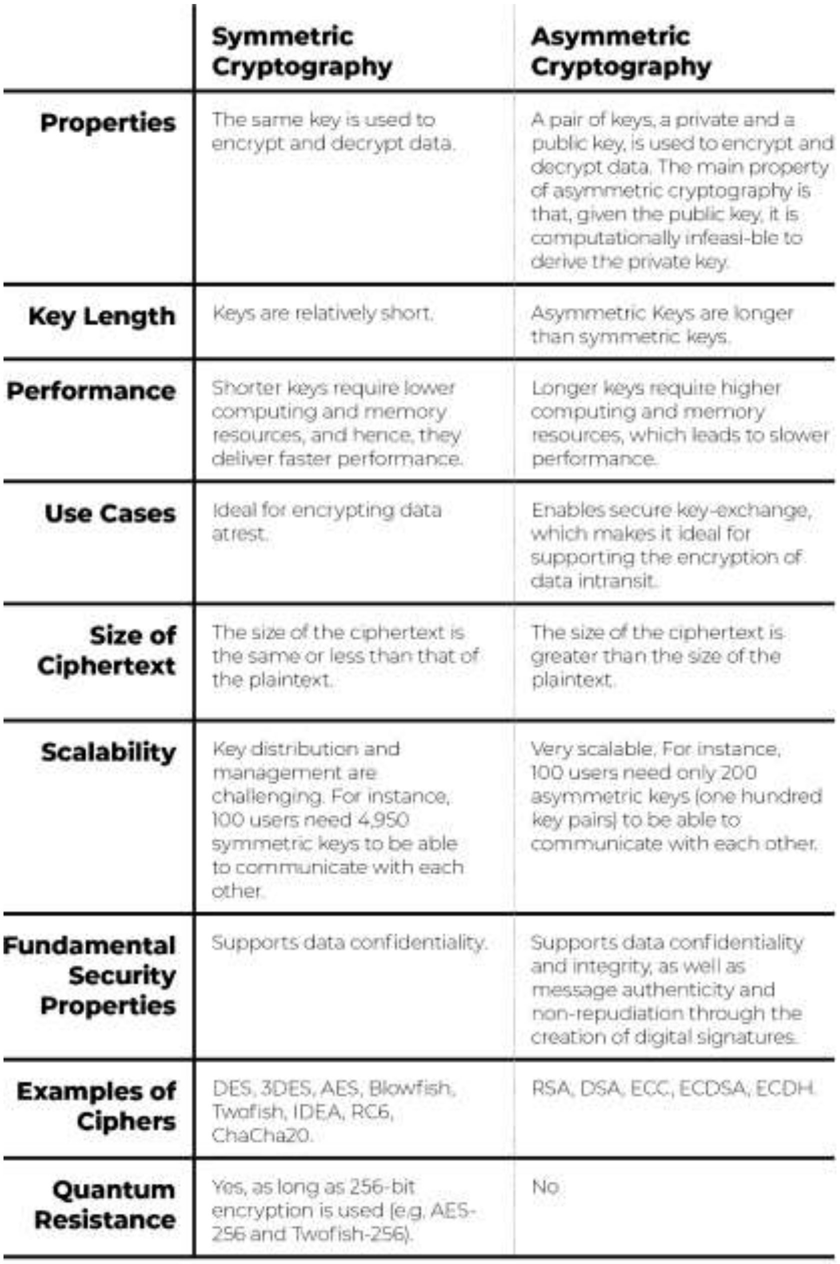
Symmetric key cryptography (also called private-key cryptography), uses the same key to both encrypt and decrypt data. Symmetric keys are relatively short, so the process of shielding and revealing, data is faster and requires fewer computing resources. It is also less resource intensive because only one secure piece of information, the symmetric key, needs to be managed. Symmetric key ciphers are typically used to encrypt data at-rest, in files and databases, because the entity storing the information is managing both sides of that process anyway. Symmetric encryption at-rest also occurs directly in computing devices to protect them from physical theft, such as disk or hard-drive encryption (see the section below on Privacy Enhanced Hardware). The best[43] at-rest symmetric key ciphers are not computationally feasible to crack with current technology.[44]

Despite its many advantages, symmetric cryptography has two important disadvantages that both pertain to key management. As described above, there is only one key that needs to be kept secret. If a user wants to share encrypted data with others, they will also have to share a copy of their one private key to decrypt the data. In this case, copies of the single key are created and distributed to authorized users. If any of these copies are lost, the data becomes vulnerable; therefore, exchanging and securing unique symmetric keys must be done carefully and it is challenging to scale.

Asymmetric cryptography, also known as public-key cryptography, is slower than symmetric cryptography because different keys are necessary to encrypt and decrypt data. However, it is more scalable,[45] since it is designed to enable secure key-exchange among multiple users. Asymmetric cryptography is based on a pair of keys that is generated for each user. One of the keys remains always private and is only known to the user, while the other is public and it is shared with any device the user would like to securely exchange data with.[46] A fundamental principle of asymmetric cryptography is that the public and private key in the key pair can both encrypt and decrypt the data. However, during a data transfer only one of the keys (either the public or private) is used to encrypt data, and the other key is used to decrypt data, and vice versa.

When a user wants to share data, they will encrypt the information using a recipient's public key. Asymmetric cryptography then ensures that only that recipient's private key, can decrypt the message. In this way, many parties can have the tools to secure data, but only one receiver can decipher the information. Because of this functionality, asymmetric cryptography is commonly used to protect data in-transit. In today's connected world, this includes extremely common use cases, such as e-mail, logging into a website, or exchanging messages on platforms, as well as digital currency applications such as sending and receiving Bitcoin.

Figure 4: Key Differences in Cryptography



	Symmetric Cryptography	Asymmetric Cryptography
Properties	The same key is used to encrypt and decrypt data.	A pair of keys, a private and a public key, is used to encrypt and decrypt data. The main property of asymmetric cryptography is that, given the public key, it is computationally infeasi-ble to derive the private key.
Key Length	Keys are relatively short.	Asymmetric Keys are longer than symmetric keys.
Performance	Shorter keys require lower computing and memory resources, and hence, they deliver faster performance.	Longer keys require higher computing and memory resources, which leads to slower performance.
Use Cases	Ideal for encrypting data atrest.	Enables secure key-exchange, which makes it ideal for supporting the encryption of data intransit.
Size of Ciphertext	The size of the ciphertext is the same or less than that of the plaintext.	The size of the ciphertext is greater than the size of the plaintext.
Scalability	Key distribution and management are challenging. For instance, 100 users need 4,950 symmetric keys to be able to communicate with each other.	Very scalable. For instance, 100 users need only 200 asymmetric keys (one hundred key pairs) to be able to communicate with each other.
Fundamental Security Properties	Supports data confidentiality.	Supports data confidentiality and integrity, as well as message authenticity and non-repudiation through the creation of digital signatures.
Examples of Ciphers	DES, 3DES, AES, Blowfish, Twofish, IDEA, RC6, ChaCha20.	RSA, DSA, ECC, ECDSA, ECDH.
Quantum Resistance	Yes, as long as 256-bit encryption is used (e.g. AES-256 and Twofish-256).	No.

The other major benefit of asymmetric encryption is the ability to verify the source of data or communications. In this case, a private key is used to encrypt information, and any entity with the corresponding public key can verify the sender of the information. Using this flow, anyone receiving data knows that the data in fact came from a specific, verified source because only that source's public key can reveal the underlying plaintext. Digital signatures and digital certificates take advantage of this feature of asymmetric cryptography to verify that data came from a trusted and verified source.

Which entities have access to encryption keys, both symmetric and asymmetric, is important for privacy laws such as GDPR. Encrypted data are not intelligible, and therefore not identifiable to an individual, or covered under the law. However, if an entity holds the key to decrypt information, they still have responsibility because it can become identifiable at any point. Under some interpretations of GDPR, if an entity receives encrypted data, but does not have the key to decrypt the information, then they do not have the same responsibilities under the law.[47]

Symmetric and asymmetric cryptography are very mature and form the backbone of the modern internet. A variety of protocols[48,49] are used to enable everything from sending messages securely, to user authentication.

HASHING

Hashing is a cryptographic function that transforms a file, folder, or even an entire disk drive, into a fixed size string of alphanumeric characters. In addition to hashes being of a fixed size, a hash function will always produce the exact same string of characters for a specific piece of data, or file. Hashing the same input at different times or on different devices should always yield the same output.

Hashing is like encryption in the sense that it transforms data into an unintelligible format; however, hashing is not typically used to shield data from unauthorized access. Instead, hashing is primarily used in integrity checks, or to ensure that files, folders, or disk drives have not been modified.

A hash is created for a file, folder, or disk, and it can be compared to the hash of the same file, folder, or disk drive at a later point in time. If the hashes match, then the information has not been altered. If the hashes do not match, then an authorized or unauthorized user has modified the data.

The transformation process for hashing is also slightly different compared to encryption. While encryption is designed to work in two directions (i.e. encryption and decryption) using keys, hashing is a one-way process and is not intended to be reversed. Some hash functions are provably difficult to reverse – meaning it is theoretically impossible (or wildly impractical) to use the output of a hash function to figure out what the input was. These hash functions are referred to as being "cryptographically secure." The ciphertext produced by weaker, non-secure hashing algorithms can be more easily linked back to the original data.

While hashing is not intended to shield data, and therefore falls outside of the scope of privacy-preserving technologies described in this report, it is an important technique for verifying the integrity of information in computer systems. For example, hashing is used to establish a chain of custody for data. Hashing is also used in conjunction with encryption to ensure message authentication and data integrity.

Homomorphic Encryption

While traditional encryption can secure data at-rest and in-transit, homomorphic encryption (HE) can shield data in-use. As described above, it is more difficult to shield data in-use because the data still needs to be intelligible for processing. Homomorphic encryption retains the usability of the data while it is shielded. This new form uses a special algebraic structure when it transforms the data, which enables arithmetic to be performed on the resulting ciphertext. There is still a key, typically asymmetric, that is used to decrypt the information, but the data can remain shielded throughout processing.[50]

This technique is still in its early stages of maturity, but it has the potential to be used widely in applications ranging from smart contracts to payment processing.[51,52] Some technology-driven firms are beginning to deploy the technique directly in their products.[53] For example, homomorphic encryption is being used to monitor whether passwords saved in browsers were ever exposed in a data breach; however, the passwords themselves remain encrypted during this analysis.[54] There are also different forms of homomorphic encryption, which vary based on the complexity of the computation that will be performed on the data. Categories of HE include partially homomorphic encryption, somewhat homomorphic encryption, and fully homomorphic encryption. If the analysis being performed on the data is limited to addition or multiplication, for example, partially homomorphic encryption is sufficient.[55]

The main challenge with homomorphic encryption is the resources required to deploy it. Encrypted data is typically much larger, and therefore takes more processing space, compared to unencrypted data. This means that more storage and processing power are needed to encrypt, store, and decrypt data not only at-rest, and in-transit, but also in-use. The advent of quantum computing will make homomorphic encryption more accessible for commercial applications, but not all entities have the necessary resources available at this point.[56]

Privacy Enhanced Hardware

Computer manufacturers are increasingly introducing off-the-shelf, privacy-enhancing features to their product lines to address business and personal use cases. Regardless of the underlying use case, this kind of hardware is deployed to shield data flowing through devices. In business use cases, these technologies can reduce the reliance on employees to follow specific privacy and security protocols, or to perform the techniques and processes themselves. These technologies are not a primary focus of this report because they are commonly targeted for consumer use; nevertheless, they are important to understand as a form of shielding data.

Examples of privacy-enhancing hardware features include:

- Privacy screens that make it difficult for strangers to observe over a user's shoulder.
- Biometric authentication, including fingerprint and/or facial recognition.
- Built-in webcam shutters.
- Kill-switches that deactivate the microphone and webcam, as well as any wireless or Bluetooth connections on a particular device.
- Drive encryption that keeps data shielded at-rest and ensures that the computer will start only when certain hardware and/or software conditions are met.
- Anti-interdiction mechanisms that detect hardware and software tampering that may occur while a device is in transit from the manufacturer's fulfillment center to the end user[57].

Many of these hardware features are available today for both businesses and individuals.

Systems + Architecture

The final category of PETs are new systems and processes for data activities. Rather than altering the data, or shielding it, systems and architectures create more secure and privacy-preserving ways for information to be handled. Some of these systems also enable greater transparency and oversight across data activities including collection, processing, transfer, use, and storage.

Multi-Party Computation

Multi-party computation is a technique that enables different entities to interact with data without revealing the complete underlying information. The technique designates data into multiple "shares," which are distributed and analyzed by different entities. Splitting up the information means that if any one entity is compromised, the full data set is not put at risk. Multi-party computation can also be combined with techniques such as homomorphic encryption, described above, so even the "shares" are not revealed during the analysis of the data.

This technique can be used on existing data sets that are then split up and provided to different service provider or analysts, or it can be used to analyze data that already exists across different organizations, collectively. The use of multi-party computation across already distributed data has the added benefit of never combining it in one central repository, and thereby reducing risk even more.

This technology is especially promising for activities that need large amounts of data, which could create additional risk if pooled together. A process known as "federated learning" trains machine learning models on data dispersed across multiple storage locations or entities. There are distinctions between these two techniques; multi-party computation is designed for privacy preservation, while federated machine learning is intended to enable a greater scale of computation across different devices.[58] Federated machine learning has vulnerabilities, in that the computations sent back to a primary server can be used to reveal the underlying data on a device.[59] Privacy-preserving techniques like multi-party computation and homomorphic encryption can be used together with federated machine learning to combine the benefits of stronger privacy and confidentiality, with greater efficiency and scale.[60,61]

A form of multi-party computation can also be done at a smaller scale through querying, or asking for confirmations from, databases without revealing underlying information. These systems can also hide the query itself from the entity holding the database or dataset to further protect confidentiality. A relevant use case could be to verify that someone holds a certain financial or physical asset, or degree from an issuing institution. A system can enable that query to be sent between entities, an analysis is performed on a database, and the binary answer is returned.

Multi-party computation is a more mature PET, and many research organizations are using it today. For example, researchers from across the world were able to use multi-party computation in 2013 to perform analysis on scientific datasets without sharing individual-level data.[62] There are also many different techniques used to accomplish multi-party computation. For example, Stanford University has developed their own approach called PRIO.[63]

While multi-party computation holds promise, like all PETs there are several challenges to its use. One of the primary hurdles is the need to standardize data structures across the entities that conduct the data analysis. If different stakeholders have unique standards or systems, this can disrupt the ability of other systems to perform analysis on data from those entities, or to interpret results provided back.

There are also certain checks and verifications that cannot be done during analysis because data are not visible to the entity performing the computations. For example, testing for representativeness in a dataset may be important for predictive machine learning so that biases are not introduced, but that may not be possible if only a portion of the data can be viewed.

Finally, multi-party computation relies heavily on the capability and security of the dispersed servers that are conducting the analysis. This means that all the parties involved in a multi-party computation need to have strong cybersecurity controls and capabilities.

Data Dispersion

Data dispersion refers to a process where data are broken into smaller pieces and maintained across a distributed storage infrastructure that, typically, spans multiple geographic locations. In this process, software is used to break data fields up in a random way. For example, if a piece of data is a social security number, the software will break apart the 9 digits in random chunks and store them in different places.

Data dispersion can provide data security and enhanced privacy because even if a storage location is breached, or those files accessed, the information will not be complete, or comprehensible, without the remaining pieces. Information could still be compromised, but multiple locations would need to be targeted, along with the underlying software that broke the data into smaller pieces.

Data dispersion can also improve scalability and performance of systems because smaller pieces of information are being stored, and it can be used in tandem with processes that create redundancy and backup storage. Data can be replicated, and then broken down into smaller pieces and stored across multiple devices. This is called storage slicing, and it is a concept similar to the mature Redundant Array of Inexpensive Disks (RAID) technology. RAID helps ensure data can be made available and reassembled even if some storage devices, or locations are compromised, or otherwise unavailable.

Data dispersion can be used with other PETs such as encryption. The small chunks of information can subsequently be encrypted in storage, in a process known as database sharding, or "microsharding."[64]

Data dispersion is becoming much more common with the increasing use of Cloud services. Distributing storage through the Cloud has significantly reduced the cost and administrative burden associated with maintaining multiple storage locations required for dispersion. However, the distribution of data across multiple geographical locations can also increase compliance and availability risk. Since data can be dispersed across several geographical regions, and even across several Cloud service providers, outages can occur and make pieces of data inaccessible. Additionally, data may be subject to the laws and regulations of different jurisdictions, complicating regulatory compliance.[65]

Management Interfaces

As firms gather data, business systems are needed to make information accessible and actionable. Entities may want to centralize data, or link systems to make data usable across business units, while also preserving the confidentiality of that information. Management interfaces are software systems that sit between datasets or databases and the employees or entities that access those datasets or databases.

These systems can perform several different functions to help preserve privacy and confidentiality. Management systems can create manual and automated access controls for databases. These access controls can be based on authorizations that only some employees have. Other access controls can be purpose-based, in which case, they restrict the use of information to activities that an individual has consented to. These systems can also perform security checks and maintain audit trails of any user who accessed information, along with the purpose of accessing the information. This audit information can be fed into a centralized view to facilitate compliance monitoring.

An important element of these kinds of systems is their ability to identify data types, tag information, or add metadata that describes certain characteristics of the data, such as sensitivity. For example, if information is identified as sensitive, systems can perform other privacy-enhancing techniques automatically and without human intervention, such as altering data, discussed earlier in the report. Additionally, sensitive information can be tagged to indicate that it belongs in a more restricted database. Data tagging can also be used to implement individual data rights, such as consent. Information that are collected can be tagged with the consent given, making it easier to identify any inappropriate use of that information.

Management systems and data tagging can also be used across entities. Tags can be added during collection and processing that indicate what entities were involved in storing and analyzing information, which creates a lineage of responsibility. If errors are introduced or a breach occurs, it is, then, possible to identify the party at fault.

The use cases for management interfaces and data tagging are very broad. Systems like this were initially used to manage electronic copyright laws, but today they can be used by almost any entity who collects, processes, or uses information. There is also the potential to blend data dispersion concepts with systems like this to give individuals themselves more insight into, and control over, data while keeping it protected. An example of this is the DigiLocker project in India, which uses a system of Application Programming Interfaces (API) to visualize data from many different government databases in one platform. The system does not pull data into a central repository; rather, it creates a representation of that information stored in other locations.[66]

Like all types of PETs, there are a variety of challenges to management interfaces and data tagging. There are many different types of interfaces and entities offering data management systems. This makes it important to consider the quality of these systems, as well as whether a particular system is appropriate for an entity's goals. A consideration around management systems is how granular the control over data is. Sometimes rules need to be set for an entire data set, or it can be done at a row, column, or cell level. Another variable are the different forms of access restriction. As discussed above, restrictions can be attribute-based, purpose-based, or specialized for an entity's rules or business arrangements. Another important consideration is how management systems interact with storage systems, and other PETs that may be deployed. For example, data stored in a centralized location, or dispersed may impact how management systems work.

A specific challenge with data tagging, or the addition of metadata, is that adding information with the intent to more securely track and process information, can create a new avenue to reidentify and tie data back to specific individuals. An important challenge to management interfaces is how they interact with changing law at the state, federal, and international level. Privacy- and data-related laws are evolving quickly, therefore many of these systems still need humans to monitor these changes and update processes. Finally, management interfaces represent significant security risks as single points of failure. Compromised management interfaces may allow unauthorized entities to access data, and/or move laterally across other systems that contain sensitive data.

Digital Identity

As activities become increasingly digital, it is important to be able to verify that individuals are who they claim to be online. Today, vast amounts of data are collected and passed between many different organizations to perform digital identity verification. New, more comprehensive, systems have been proposed and implemented in some countries that could reduce the need to collect and share sensitive identifying information, and thereby preserve privacy.

There are many techniques for identifying individuals physically and digitally that occur across a multitude of use cases. Different entities have specialized in performing narrow verifications for individuals. For example, the government's Department of Motor Vehicles provides a physical certificate that proves your ability to drive legally, and it is also used as a source of broader identity verification. A multitude of other firms specialize in verification for authorizing payments, for Know Your Customer onboarding in financial services, for performing background checks, and much more.[67]

Adding to this complexity is the need for both initial verification of an individual's identity at the start of a relationship, and then authenticating that they are the same person each time they return. In the physical world, if someone shows identification to a security guard, that is the initial verification, but once the guard knows that persons face, they can just let them through with a glance. The digital world is not as simple.

The initial step of verifying identities is typically done by collecting information from an individual, and then validating that with trusted sources, such as a government entity, like the DMV. This initial check can be wide ranging, and it is common to piece together information from multiple sources and make sure they match. Additional data may also be collected at the verification step to compare to subsequent authentications, such as digital behavior (typing style) and device information.

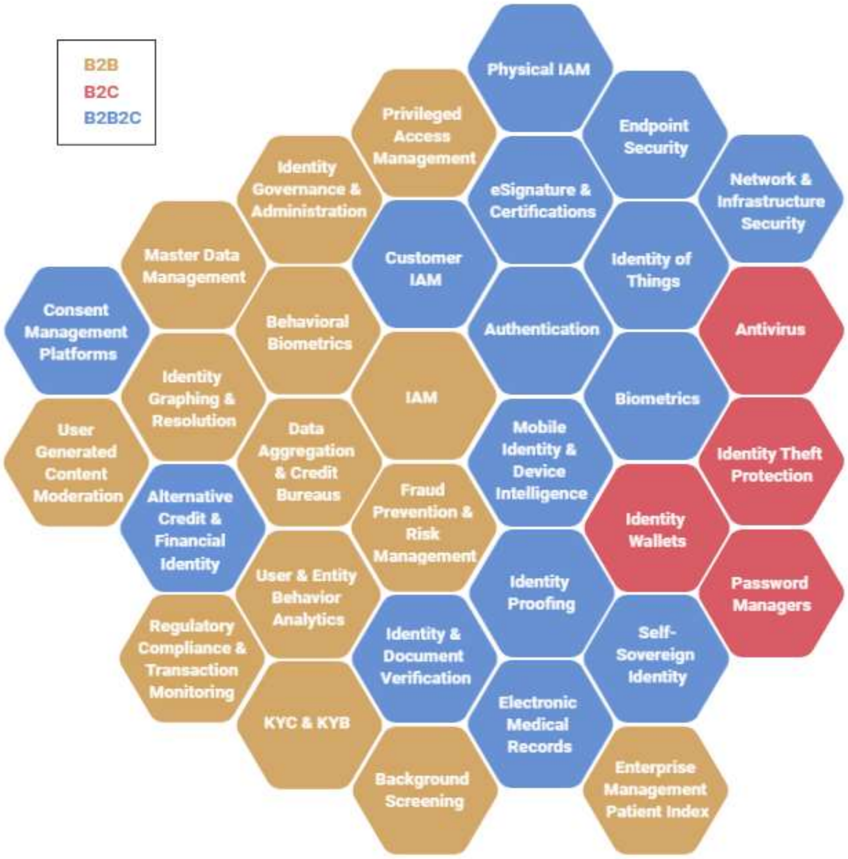
Once an individual has been verified, there are different ways to authenticate them when they return. These are typically categorized as asking for (1) something they know, (2) something they have, or (3) something they are. The most common example of the first category are usernames and passwords, or security questions. Something someone has, or the second category, refers to the possession of authentication tokens. These can be VPN tokens that plug into computers, or an authenticator application on a cell phone. An example of this is when text messages are sent to cell phones to authenticate users. The final category, of "something people are" can be biometrics, like a thumbprint or even a heartbeat,[68] or a new category of behavioral biometrics that is based on unique patterns we each create when we type or navigate webpages. [69]

When entities are authenticating returning individuals, they can use multiple forms of authentication, i.e. something that they know, have, and/or are, this is called multi-factor authentication (MFA). Any additional information collected at the earlier verification stage may also be used as a double check. For example, whether the same IP address is being used.

Software providers create digital certificates (discussed in the sections above) which are transferred between their services, so individuals do not need to be constantly authenticated. When unique digital signatures, or traces for verified individuals, work across a single provider, this is called Single-Sign On (SSO). Digital certificates can also be used across different websites and providers; an example of this is Federated Identity Management (FIM).[70] There are also common standards for digital verification and authentication used across markets, such as the NIST 800-63, FIDO, and W3C.

While these systems are used heavily today, they create several privacy concerns. The first concern is the amount of data that are being collected for initial verification and subsequent authentication. Much of this data, from location to biometrics, are highly sensitive. Another concern is the security of these systems. As more data are revealed in breaches, bad actors can easily use that information to steal identities. Furthermore, systems that only use a single form of authentication, like a username and password, which are both something the user knows, have become notoriously weak. The Financial Crimes Enforcement Network (FinCEN) estimates that there are billions of usernames and passwords, as well as sensitive personal information, currently exposed to fraudulent actors.[71]

Figure 5: Digital Identity Use Cases


B2B
B2C
B2B2C
Physical IAM
Privileged Access Management
Endpoint Security
Identity Governance & Administration
eSignature & Certifications
Network & Infrastructure Security
Master Data Management
Customer IAM
Identity of Things
Consent Management Platforms
Behavioral Biometrics
Authentication
Antivirus
Identity Graphing & Resolution
IAM
Biometrics
User Generated Content Moderation
Data Aggregation & Credit Bureaus
Mobile Identity & Device Intelligence
Identity Theft Protection
Alternative Credit & Financial Identity
Fraud Prevention & Risk Management
Identity Wallets
User & Entity Behavior Analytics
Identity Proofing
Password Managers
Regulatory Compliance & Transaction Monitoring
Identity & Document Verification
Self-Sovereign Identity
KYC & KYB
Electronic Medical Records
Background Screening
Enterprise Management Patient Index

Graphic from OneWorldIdentity Digital Identity Landscape Report

An additional challenge is the diversity of entities handling this data across different use cases, and their commercial incentives. While these entities may be very secure, there are simply more points of potential weakness where private information could be accessed or misused. The commercial nature of many of these verification and authentication services can also make oversight for issues like privacy and equitable access challenging. Many of the largest systems for online verification are run by technology-focused firms that fall outside of regulated markets like financial services. Because of this, it can be challenging to review these systems to make sure they are preserving privacy and treating different groups equitably. The concerns across privacy, security, and oversight in online identification are even more important as more and more services become digital.

New proposals have been put forth, and implemented in different countries, that strive to create a more unified approach to digital identity. Two examples are decentralized systems utilizing blockchain technology,[72] and centralized systems, typically run by governments or large trusted actors.

Digital identity systems using blockchain technology create a shared, but distributed and immutable record across different entities, or nodes. Entities can add pieces of information, such as a degree, or confirm attestations, such as a government validating an online identity. A primary benefit of this kind of system is that individuals can reveal limited, but relevant information, such as their age, or simply a trusted confirmation, without providing additional, associated data. While blockchain is generally considered secure because you need the confirmation of multiple parties and once recorded the information it is immutable, recent evidence suggests that blockchains can be successfully attacked. If digital identities are managed on this kind of system, it is important to be conscious of potential software insecurities at the nodes that supply data and potential flaws in the underlying cryptography.[73]

While these ideas are still early in their development, several private entities[74] and consortiums[75] have been established. One of the main challenges with the distributed approach to digital identity is the need for large-scale adoption to be effective. A large enough cohort of entities need to both provide identify verification and accept attestations for the system to work for individuals.

Another unified system for digital identity that is being explored is centralization in one trusted entity, the most well-known being India's Aadhaar program. The program was created by gathering demographic and biometric data from every Indian citizen, and then assigning them a unique 12-digit number.[76] Estonia also has a nationwide digital identity system, and other countries such as Denmark, Finland, and Singapore[77] have targeted identity utilities for certain sectors, such as financial services. Other examples of government-based systems include digital driver's licenses, which have been explored in U.S. states.[78] The value of centralizing this kind of identification with governments is that governments typically provide and manage physical forms of resident or citizen identification already, and the incentives between governments and individuals are, hopefully, aligned around keeping the underlying information secure and confidential.

There are also private companies exploring how to provide and use, a unified, centralized digital identity beyond the single-sign on and federated systems described above.[79] Like distributed systems, centralized forms of digital identity provision need scale to be effective, and it is not clear whether private entities can reach the same scale as government providers. While there are large entities with that potential, there is a related concern that they may not be incentivized to include marginalized populations. Additionally, if there are several different private entities competing, one single system may not be able to achieve enough market dominance to make it widely usable.

ZERO-KNOWLEDGE PROOFS

Zero-knowledge proofs are a cross-cutting concept that means information can be confirmed without revealing the underlying data the confirmation is based on. Zero-knowledge proofs occur across different types of PETs.

In the digital identity space current and proposed systems can transfer a binary verification that someone is who they say they are without passing on a social security or driver's license number.

Asymmetric encryption can also provide zero-knowledge proof. Someone receiving asymmetrically encrypted data knows that the data originates from a specific source because only the sender's corresponding public key works to decrypt the information. The existence of that unique pair of keys means additional information does not need to be gathered to verify the data source.

Hashing can also provide zero-knowledge proof because it demonstrates that information has not been altered without revisiting the underlying data.

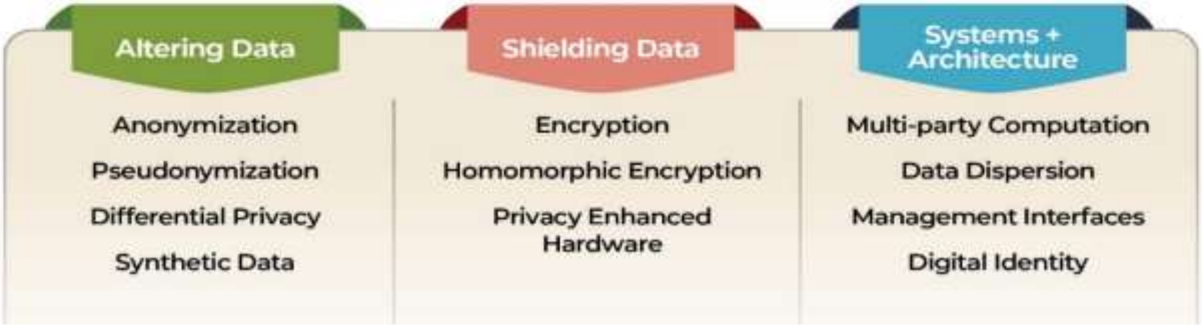
Major challenges to centralized systems include security and accountability. Centralized systems create a single point of failure and can be vulnerable to attacks and/or corruption. For example, India's Aadhar system has been a target for hackers and has been breached several times.[80] Additionally, one system for digital identity is extremely important for systemic stability and the digital economy, therefore it is essential to consider how oversight and accountability would work for either a government or private provider.

Conclusion

Privacy enhancing technologies are a fascinating and exciting set of tools that can help capture the value of data while keeping it secure, confidential, and private.[81] Despite this potential, PETs are not standalone solutions to privacy and security concerns, and must be used in tandem with robust policy and governance systems.

As regulators, policymakers, and businesses explore this space it is important to understand the diversity of techniques and systems that make up PETs, their different strengths, and goals.

Figure 1: Categories of Privacy Enhancing Technologies*



Altering Data	Shielding Data	Systems + Architecture
Anonymization	Encryption	Multi-party Computation
Pseudonymization	Homomorphic Encryption	Data Dispersion
Differential Privacy	Privacy Enhanced Hardware	Management Interfaces
Synthetic Data		Digital Identity

**These categories are based on the authors' analysis of the PET space. The authors acknowledge that there are multiple ways to group these technologies, techniques, and processes.*

Methods of altering data, such as anonymization, are intended to be permanent, but can be reversed leading to privacy and confidentiality risk for entities and individuals. Techniques and systems that are intended to shield data are typically temporary, and they protect data across states of storage, transit, and use. A main risk to these PETs is the fact that they are designed to be reversed, and therefore it is important to know which entities have that capability. Systems and architectures are a broad category of PETs that can help both manage privacy and enable active individual data rights by tracking concepts like individual consent. This category also includes the potential for market-wide systems like digital identity which could reduce our overall dependence on data to verify individuals in online spaces. Finally, it is important for decision-makers to understand that these categories of PETs can, and many times should, be used together to increase privacy protection.

The development and use of PETs is still in early stages, and while much of it is driven by sector-specific requirements and new regulation, there is a larger social shift towards "privacy by design"[82] and minimizing data collection and use where possible. As this space evolves there may be a greater convergence around research, and the development of accepted standards for some of these techniques and processes.

Privacy enhancing technologies are a promising set of tools, techniques, and systems that can help keep data secure and private while still leveraging it to create value. PETs have the potential to enable both greater security and confidentiality of data across use cases, and to enable greater individual control over data by providing transparency, choice, and auditability within data systems.

Given the nuances and ongoing evolution of this space, there are several important concepts and questions to consider moving forward:

- PETs do not make activities inherently compliant with privacy regulations, and they should not be used to circumvent consumer protections.
- PETs may be reversed; therefore, regulatory requirements around altering or shielding data should be focused on how difficult it is to reveal the underlying information, not whether it can be done at all.
- More work is needed to determine whether there is a threshold when information can no longer be 'reasonably' associated with an individual.
- PETs have the potential to create technical barriers to intentional or unintentional misuse of data within companies.
- How PETs intersect with data rights, and whether individuals should retain rights around de-identified information.
- How business practices could evolve to minimize the collection and use of data. [83]
- How PETs may be used in wider digital public infrastructure for cybersecurity and cloud storage.
- How PETs can be made affordable and accessible to all entities.

While there is clearly more work to be done, many public[84],[85] and private[86] entities are exploring the topic of PETs and their potential impact on financial services and beyond.

The Federal Reserve Bank of San Francisco looks forward to continuing to participate in research and dialogue around privacy enhancing technologies and their applications.

References

[1] Mathews, Anna Wilde. “Major Hospitals Form Company to Capitalize on Their Troves of Health Data.” The Wall Street Journal, February 11, 2021.

[2] “2019 ‘Worst Year on Record’ for Data Breaches.” Ensighten, September 11, 2019.

[3] Ohm, Paul. “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization.” UCLA Law Review. UCLA School of Law, Accessed March 2020.

[4] Sweeney, Latanya. “Simple Demographics Often Identify People Uniquely.” Data Privacy Lab, January 1, 2000.

[5]Acosta, Nefi. “Are IP addresses 'personal information' under CCPA?” Internet Association of Privacy Professionals, April 28, 2020.

[6] Kenny, Steve. “An Introduction to Privacy Enhancing Technologies.” Internet Association of Privacy Professionals, May 1, 2008.

[7] Press, Gil. “Intel Advances on The Road to Quantum Practicality.” Forbes.com, August 17, 2020.

[8] “Data Controllers and Processors.” GDRPEU.org, accessed March 2021.

[9] “What is the difference between privacy, confidentiality and security.” Techopedia, July 2, 2020.

[10] “Privacy Enhancing Technologies – A Review of Tools and Techniques.” Office of the Privacy Commissioner of Canada, November 2017.

[11] “Control TowerSM.” Wells Fargo, Accessed March 19, 2021.

[12] “California Consumer Privacy Act of 2018.” California Legislative Information, California Civil Code, Accessed March 2021.

[13] Rehm, Rainer. “One year with GDPR: Greater data hygiene and security, higher bureaucracy, and more uncertainty.” Zscaler, May 24, 2019.

[14] “Know Your Customer.” Board of Governors of the Federal Reserve System. Section 601.0, Accessed March 2021.

[15] “Assessing Compliance with BSA Regulatory Requirements.” FFIEC BSA/AML Manual, Accessed March 2021.

[16] Electronic Health Information Laboratory. Accessed May 2021.

[17] Fennessy, Caitlin. “The 'Schrems II' decision: EU-US data transfers in question.” Internet Association of Privacy Professionals, July 16, 2020.

[18] Kang, Sunny Seon. “Post-Schrems II, Privacy-Enhancing Technologies for Cross-Border Data Transfers.” Jurist, January 25, 2021.

[19] Sadler, Chris. “Homomorphic Encryption Could Fix The Gaps In Our Data Security.” New America, September 1, 2020.

[20] “Security, Privacy and Abuse Prevention.” Google Research, Accessed April 2021.

[21] “Promoting Digital Privacy Technologies Act.” 117th Congress, Introduced February 4, 2021.

[22] See (12)

[23] “Data Protection Act of 2018.” Part 6 Enforcement, Offences relating to personal data Section 172. UK Government, Accessed May 2021.

[24] Schneier, Bruce. “Former FBI General Counsel Jim Baker Chooses Encryption Over Backdoors.” Schneier on Security, October 38, 2019.

[25] Barrett, Brian. “The Apple-FBI Fight Isn't About Privacy vs. Security. Don't Be Misled.” Wired, February 24, 2016.

[26] “Bank Secrecy Act / Office of Foreign Assets Control.” Board of Governors of the Federal Reserve System, Accessed May 2021.

[27] The Gramm-Leach-Bliley Act (GLBA, 15 USC § 6801- 6827) refers to PII collected by financial institutions as “Non-public Private Information (NPI)”, whereas the Health Insurance Portability and Accountability Act (HIPAA, 42 USC § 1320d) refers to PII collected by “covered entities” (i.e. health plans, clearinghouses, and certain health care providers) as Protected Health Information (PHI).

[28] The University of Massachusetts Medical School maintains a comprehensive list of direct and indirect identifiers as it pertains to the Health Insurance Portability and Accountability Act (HIPAA):

[29] MAC and IP addresses are both essential in computer networking; they are unique identifiers that allow computing devices (i.e. computers, smartphones, tablets, and other network-enabled devices) to communicate on a local network or the Internet.

[30] “Policy and Law: Identifiability of de-identified data.” Research Accomplishments of Latanya Sweeney, PH.D, Accessed March 2020.

[31] Bradbury, Danny. "Data Breach Numbers Skyrocket in 2019." InfoSecurity, August 6, 2012.
[32] "Tokenization: Everything You Need to Know." CardConnect, July 20, 2020.
[33] Samarati, P. and L. Sweeney. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression." Semantic Scholar, 1998.
[34] "PCI Security Standards Council Releases PCI DSS Tokenization Guidelines." Payment Card Industry Security Standards Council, August 12, 2011.
[35] Yves-Alexandre de Montjoye et al. "Unique in the shopping mall: On the re-identifiability of credit card metadata." Science Magazine, January 30, 2015.
[36] Helm, Burt. "Credit card companies are tracking shoppers like never before: Inside the next phase of surveillance capitalism." FastCompany, May 12, 2019.
[37] "Differential Privacy Overview." Apple, Accessed May 2021.
[38] "Differential Privacy for Census Data Explained." National Conference of State Legislators, March 15, 2021.
[39] "The Ultimate Guide to Synthetic Data in 2021." AI Multiple, February 2, 2021.
[40] "2019 Global AML and Financial Crime TechSprint." UK Financial Conduct Authority, Accessed May 2021.
[41] Benedetti, Marcelo. "The Advantages and Limitations of Synthetic Data." Sama, January 24, 2018.
[42] Each additional character in a key, doubles the number of possible keys. This means that, if an adversary attempts to guess the key, they will need more processing power and double the time to do so.
[43] The current gold standard in symmetric cryptography is the Advanced Encryption Standard (AES) cipher, which was selected by the National Institute of Standards and Technology (NIST) following a competition in 2001.
[44] Gallagher, Ryan. "A Swiss Company Says It Found Weakness That Imperils Encryption." MSN, February 7, 2021.
[45] Scalability is supported by the fact that if n users want to securely exchange messages with each other, the total number of asymmetric keys required is $n \times 2$, since each user will need to have one private and one public key. This means that 10 users will need 20 asymmetric keys (as opposed to 45 symmetric keys), and 100 users will need 200 asymmetric keys (as opposed to 4,950 symmetric keys). The difference in the number of required keys between symmetric and asymmetric cryptography is substantial.
[46] Please refer to this article for an explanation of the difference between public and private keys in asymmetric cryptography: https://medium.com/@vrypan/explaining-public-key-cryptography-to-non-geeks-f0994b3c2d5.
[47] Gerald Spindler and Philipp Schmechel. "Personal Data and Encryption in the European General Data Protection Regulation." 2016.
[48]The Internet is a massive networking infrastructure: a network of networks. The Worldwide Web, also known as the Web, is the system we use to access the Internet and connect with other devices to exchange data.
[49] A prominent use of TLS can be found in the security of Hyper Text Transfer Protocol (HTTP) communications Another major application of TLS is in the encryption of Domain Name System traffic. DNS is akin to a phone book that contains a list of domain names (e.g. federalreserve.gov, frbsf.org) and their corresponding IP addresses. The reason is that computers are, unsurprisingly, unable to understand the domain names and website names that humans use and can only understand IP addresses. For instance, each time a user wants to visit a website, DNS ensures that the user is directed to the right IP address that corresponds to the website they want to visit. DNS queries are sent in plaintext, which means that they are unencrypted. The problem is that DNS queries are public, and DNS servers maintain a record of the IPs of devices that requested to visit a particular website, thus potentially violating the privacy of Internet users.
[50] "UN Privacy Preserving Techniques Handbook." BigData Un Global Working Group, August 1, 2019.
[51] Sheridan, Kelly. "Major Brazilian Bank Tests Homomorphic Encryption on financial Data." DarkReading, January 10, 2010.
[52] Maass, Eric. "Fully Homomorphic Encryption: Unlocking the Value of Sensitive Data While Preserving Privacy." SecurityIntelligence, December 17, 2020.
[53] "Microsoft SEAL." Microsoft Research, Accessed May 2021.
[54] Kannepalli, Sreekanth. "Password Monitor: Safeguarding Passwords on Mircosoft Edge." Microsoft Research Blog, January 21, 2021.
[55] Messina, Graeme. "What is homomorphic encryption." InfoSec Institute, April 27, 2021.
[56] Leprince-Ringuet, Daphne. "A quantum computer just solved a decades-old problem three million times faster than a classical computer." ZDNet, February 23, 2021.
[57]Wilhelm, Alex. "The NSA, Cisco, and the Issue of Interdiction." TechCrunch, May 18, 2014.
[58] Hong, Cheng. "Federated Machined Learning and MPC." Medium. January 16, 2020.

[59] "The Privacy Risk Right Under Our Nose in Federated Learning (Part 1)." Inpher, February 23, 2021.
[60] Mugunthan et al. "SMPAI: Secure Multi-Party Computation for Federated Learning." 33rd Conference on Neural Information Processing Systems, 2019.
[61] Ma, Xu et al. "Privacy preserving multi-party computation delegation for deep learning in cloud computing." Information Sciences, Volume 459, August 2018.
[62] Doiron et al. "Data harmonization and federated analysis of population-based studies: the BioSHaRE project." National Library of Medicine, November 21, 2013.
[63] "Prio." Stanford University, Accessed April 2021.
[64] Meher, Easha. "Database Sharding: A Comprehensive Guide." Hevo, March 17, 2021.
[65] Pollet, Mathieu. "France under fire for use of Amazon-hosted Doctolib for jab bookings." Euractiv, March 2, 2021.
[66] "DigiLocker." Government of India, Accessed May 2021.
[67] Jarae, Travis. "OWI Digital Identity Landscape." One World Identity, January 26, 2021.
68 Matchar, Emily. "Using Your Heartbeat as a Password." Smithsonian Magazine, January 30, 2017.
[69] Korolov, Maria. "What is biometrics? 10 Physical and behaviroal identifiers that can be used for authentication." CSO Online, February 12, 2019.
[70] Piazza, Dan. "Authentication, Authorization, Single Sign-On, & Federated Identity Explained." Stealthbits, November 19, 2021.
[71] "Prepared Remarks of FinCEN Director Kenneth A. Blanco, Delivered at the Federal Identity (FedID) Forum and Exposition." Financial Crimes Enforcement Network, September 24, 2019.
[72] Spring Labs. Accessed March 2020.
[73] Orcutt, Mike. "Once hailed as unhackable, blockchains are now getting hacked." MIT Technology Review, February 19, 2019.
[74] See (71)
[75] Sovrin. Accessed March 2020.
[76] "What Is Aadhaar - Unique Identification Authority of India: Government of India." Unique Identification Authority of India. Government of India, Accessed March 30, 2020.
[77] "MAS to Roll out National KYC Utility for Singapore." Finextra, March 24, 2017.
[78] Whitney, Lance. "The Driver's License of the Future Is Coming to Your Smartphone." CNET, March 21, 2015.
[79] Page, Rosalyn. "Mastercard to Pilot New Digital ID System." CMO Australia, December 17, 2019.
[80] Doshi, Vidhi. "A security breach in India has left a billion people at risk of identity theft." The Washington Post, January 4, 2018.
[81] Polonetsky, Jules & Elizabeth Renieris. "10 Privacy Risks and 10 Privacy Enhancing Technologies to Watch in the Next Decade. Future of Privacy Forum." Future of Privacy Forum, 2020.
[82] Cavoukian, Ann. "Privacy by Design The 7 Foundational Principles." Internet Association of Privacy Professionals, Accessed May 2021.
[83] Schechner, Sam and Keach Hagey. "Google to Stop Selling Ads Based on Your Specific Web Browsing." Wall Street Journal, March 3, 2021.
[84] "Privacy enhancing technologies." European Union Agency for Cybersecurity, Accessed May 2021.
[85] "Privacy Enhancing Technologies – A Review of Tools and Techniques." Office of the Privacy Commissioner of Canada, November 2017.
[86] "The Alan Turing Institute partners to research security and privacy technologies in digital identity systems." Alan Turing Institute, April 29, 2020.