Privacy-Preserving Techniques for Trustworthy Data Sharing: Opportunities and Challenges for Future Research

Lidia Dutkiewicz, Yuliya Miadzvetskaya, Hosea Ofe, Alan Barnett, Lukas Helminger, Stefanie Lindstaedt, and Andreas Trügler

Abstract One of the foundations of data sharing in the European Union (EU) is trust, especially in view of the advancing digitalization and recent developments with respect to European Data Spaces. In this chapter, we argue that privacy-preserving techniques, such as multi-party computation and fully homomorphic encryption, can play a positive role in enhancing trust in data sharing transactions. We therefore focus on an interdisciplinary perspective on how privacy-preserving techniques can facilitate trustworthy data sharing. We start with introducing the legal landscape of data sharing in the EU. Then, we discuss the different functions of third-party intermediaries, namely, data marketplaces. Before giving a legal perspective on privacy-preserving techniques for enhancing trust in data sharing, we briefly touch upon the Data Governance Act (DGA) proposal with relation to trust and its intersection with the General Data Protection Regulation (GDPR). We continue with an overview on the technical aspects of privacy-preserving methods in the later part, where we focus on methods based on cryptography (such as homomorphic encryption, multi-party computation, private set intersection) and link

L. Dutkiewicz · Y. Miadzvetskaya
KU Leuven Centre for IT & IP Law – imec, Leuven, Belgium

H. Ofe
TU Delft, Faculty of Technology, Policy and Management, Delft, The Netherlands

A. Barnett
OCTO Research Office, Dell Technologies, Ovens, County Cork, Ireland

L. Helminger · S. Lindstaedt
Graz University of Technology, Graz, Austria

Know-Center GmbH, Graz, Austria

A. Trügler (✉)
Know-Center GmbH, Graz, Austria

Graz University of Technology, Graz, Austria
e-mail: atruegler@know-center.at

E. Curry et al. (eds.), *Data Spaces*, https://doi.org/10.1007/978-3-030-98636-0_15

them to smart contracts. We discuss the main principles behind these methods and highlight the open challenges with respect to privacy, performance bottlenecks, and a more widespread application of privacy-preserving analytics. Finally, we suggest directions for future research by highlighting that the mutual understanding of legal frameworks and technical capabilities will form an essential building block of sustainable and secure data sharing in the future

Keywords Data law · Data sharing · Trust · Data Governance Act · Privacy-enhancing techniques · Homomorphic encryption · Multi-party computation · Cryptography · Private set intersection · Federated learning · GDPR · Data marketplace · Data Governance · Smart contracts · Secure enclave

1 Introduction

One of the backbones of data sharing intermediaries and European Data Spaces is privacy, especially in view of the advancing digitalization and global economic and socioeconomic developments. New research breakthroughs and the possibilities of privacy-preserving technologies have to comply with data protection laws to enable a secure and sustainable data economy.

In this chapter, we therefore focus on an interdisciplinary perspective on how privacy-preserving techniques can facilitate trustworthy data sharing. We start with introducing the legal landscape of data sharing in the European Union and give an overview on the technical aspects of privacy-preserving methods in the later part. We discuss the main principles behind these methods and highlight the open challenges with respect to privacy and suggestions for future research for data platforms.

The chapter relates to the technical priorities of data processing architecture of the European Big Data Value Strategic Research and Innovation Agenda [1]. It addresses the horizontal concern of data protection of the BDV Technical Reference Model, and it addresses the vertical concerns of Marketplaces, Industrial Data Platforms, and Personal Data Platforms.

The chapter relates to the Knowledge and Learning, Reasoning, and Decision-Making enablers of the AI, Data and Robotics Strategic Research, Innovation and Deployment Agenda [2].

1.1 Data Sharing Now: A Legal Patchwork

Advances in ICT have had and continue to have fundamental impacts on society. A vital aspect of this trend is the vast amount of data collected and used as data-related technologies impact the socioeconomic life of companies and individuals. Data is often referred to as a new oil, new resource, new infrastructure, and the fifth freedom

of the EU internal market. This trend toward treating data as an economic asset just like goods, capital, and services is known as a "commodification of data."

An estimated amount of 33 zettabytes of data was generated worldwide in 2018, and according to the European Data Strategy, this amount of data is expected to rise to 175 zettabytes in 2025. The EU's data economy value is estimated to reach 550 billion euros by 2025 [3]. The free movement of personal and non-personal data is therefore of strategic importance for fostering the EU data-driven economy. However, one of the main difficulties for this economic opportunity to materialize resides in the fact that data transactions are regulated in the EU by a legal patchwork. The intersections between those legal instruments are often a subject of controversies.

First of all, there is the General Data Protection Regulation (GDPR)[1] that applies since 25 May 2018 and constitutes the cornerstone of the EU personal data-related framework. The GDPR touches upon a few data protection-related questions particularly relevant to data market ecosystems such as this of TRUSTS.[2] These include, e.g., the determination of controllership and the ensuing allocation of data protection responsibilities and the legal basis for processing personal data [4].

Second, the Regulation on the free flow of non-personal data[3] is another building block of the EU data-related legal patchwork. According to its Article 1, the Regulation ensures the free flow of data other than personal data within the Union. The Regulation aims at removing obstacles to the free movement of non-personal data across the EU, notably data localization requirements, unless they are justified on grounds of public security (Article 4 of the Regulation) and vendor lock-in practices in the private sector.

At the same time, it remains unclear how to delineate what qualifies as personal data and what remains outside the scope of the personal data protection regime. In accordance with Article 4 of the GDPR, the notion of personal data is rather broad and encompasses "any information relating to an identified or identifiable natural person." It is not excluded that technological developments will make it possible to turn anonymized data into personal data and vice versa.[4] Thus, it is always safer to treat any data as personal.

Another difficulty concerns a mixed data set composed of both personal and non-personal data. The Regulation on the free flow of non-personal data applies only to the non-personal data part of the data set. Where data sets are inextricably linked, the GDPR shall prevail in accordance with Article 2(2) of the Regulation. The Commission also published informative guidance on the interaction between

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

[2] Trusted Secure Data Sharing Space, Horizon 2020, https://www.trusts-data.eu/

[3] Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union, OJ 2018 L 303/59.

[4] GDPR, rec. 9.

the Regulation on the free flow of non-personal data and the GDPR where it clarified which rules to follow when processing mixed data sets and explained the concept of data sets "being inextricably linked" [5].

The Open Data Directive[5] in force since 2019 is another building block of the EU data-related framework. Its main aim is to allow free re-use of data that are held by national public sector bodies. This is meant to foster the emergence of new businesses that offer digital products and services. The Directive aims at increased re-use of data held by public sector bodies and certain public undertakings. However, the Open Data Directive does not apply to documents for which third parties hold intellectual property rights or that constitute commercial secrets. The Open Data Directive does not prevail over the GDPR in accordance with its Art. 1(4) and only applies to data that is not personal.

Moreover, there is a vast amount of EU legislation indirectly applicable to data sharing consisting of general and horizontal legislation (e.g., Database Directive, Copyright DSM Directive, Trade Secrets Directive, Software Directive, Regulation of B2B unfair commercial practices) and sector-specific rules (e.g., the PSD2 and the AML). For absence of a horizontal legal framework regulating B2B data sharing, the EU has been active in elaborating soft law guidelines for businesses [6].

Up to this date, the legal patchwork for data transactions does not sufficiently address the commodification of data and leaves some uncertainties when it comes to applicable rules.

However, recently, the EU has shifted its focus to other ways of regulating data transactions, notably data sharing, data re-use, and making the data available. In the European Data Strategy, the European Commission emphasized that the development of data marketplaces is a key policy instrument to revitalize the full potential of the value of data generated across member states [4]. The broad aim of the strategy is to "create a genuine single market for data, where personal and non-personal data, including confidential and sensitive data, are secure and where businesses and the public sector have easy access to huge amounts of high-quality data to create and innovate" [4].

1.2 Data Marketplaces

In spite of the economic potential data is suggested to have, data sharing between companies has not taken off at sufficient scale. This is, among others, due to a "lack of trust between economic operators that the data will be used in line with contractual agreements, imbalances in negotiating power, the fear of misappropriation of the data by third parties, and a lack of legal clarity on who can do what with the data" [4].

[5] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, OJ 2019 L 172/56.

To address these challenges, the trusted third-party intermediaries (e.g., data marketplaces) come into play. Data marketplaces are defined as platforms that provide services for buying and selling of data products [7]. They bring data suppliers and data users together to exchange data in a secure online platform. Based on the matching function they perform, data marketplaces can range from one to one, one to many, many to one, and many to many [6]. For example, one-to-one data marketplaces enable bilateral exchanges between two parties, while many-to-many are multi-lateral marketplaces [6].

Data marketplaces can also be characterized based on the functions they perform. As indicated by the European Commission, a data marketplace is a specific type of intermediary which may have the following functions [6]:

Match-Making Between Potential Data Supplier and Data Buyer
In that scenario, the platform matches the supply and demand between the potential suppliers and potential buyers and facilitates data sharing between the parties. From an economic perspective, it lowers transaction costs through combining different data sources [9].

The Actual Transfer of the Data and Trust Creation
For businesses, data trading is quite sensitive since they become vulnerable to competitors or adverse effects. Platforms may therefore rely on the usage of privacy-preserving technologies, perform screening of data sharing partners, supervise and protocol the individual transactions, as well as enforce usage constraints.

Provider of the Technical Infrastructure
Data marketplaces may be defined as an "architecture allowing programmability and reuse of content and data, typically through API, and organizing modularity between a stable core and variable components" [10].

Data intermediaries can also provide additional services and functionalities such as model contract clauses or (pseudo)anonymization services (if personal or confidential data are exchanged), privacy-preserving data analytics, etc.

The variety of data marketplaces and the functions they can perform raise the question of how to regulate the activities of data sharing intermediaries.

1.3 Data Governance Act ("DGA")

In November 2020, the European Commission put forward a proposal for a regulation on European Data Governance[6] (Data Governance Act, "DGA") that provides for the rules aimed at facilitating the re-use of publicly held data, regulating the activities of data sharing intermediaries, fostering data altruism, and preventing international access to EU-based data by foreign governments and

[6] Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act) COM/2020/767 final, Brussels, 25.11.2020.

entities. According to the Impact Assessment of the European Commission, the overall objective of the DGA proposal is to set the conditions for the development of common European Data Spaces and strengthen trust in data sharing and in data intermediaries.

With the DGA proposal, in order to increase trust in such data sharing services, the EC aims to create an EU-wide regulatory framework, which would set out highly harmonized requirements related to the trustworthy provision of data sharing services. According to the proposal, a key element to bring trust and more control for data holder and data users in data sharing services is the neutrality of intermediaries—data sharing service providers.[7] The Regulation proposes a number of measures to increase trust in data sharing, including the structural separation between the data sharing service and any other services provided and a notification regime for data sharing providers.

Moreover, the intersection between the GDPR and DGA raises a number of questions. First of all, data processing principles, enshrined in the GDPR, such as purpose limitation and data minimization, are difficultly compatible with the objective of stimulating data sharing in the EU. Secondly, the sharing of personal data by data subjects requires trust in data controllers and data users to prevent any cases of misuse of personal data for different purposes than those communicated at the moment of data collection or sharing.

Finally, the DGA provides for techniques enabling privacy-friendly analyses where personal data are involved, such as anonymization, pseudonymization, differential privacy, generalization, or suppression and randomization. The application of these privacy-enhancing technologies and compliance with the GDPR are meant to ensure the safe re-use of personal data and commercially confidential business data for research, innovation, and statistical purposes.[8] Against this background, this chapter argues that privacy-preserving techniques, such as multi-party computation and fully homomorphic encryption, can play a positive role as enablers of trust in data sharing in compliance with fundamental rights to privacy and data protection. In the next section, we will provide a legal perspective on different privacy-preserving techniques and their impact on leveraging trust for data transactions.

2 Legal Perspective on Privacy-Preserving Techniques for Enhancing Trust in Data Sharing

2.1 What Is Trust?

Trust is a fundamental aspect of social interactions. It is generally understood as a relationship in which an agent (the trustor) decides to depend on another agent's (the

[7] DGA, rec. 26.

[8] DGA, rec. 6.

trustee) foreseeable behavior in order to fulfil his expectations [11]. Trust is a much-discussed concept in ethics of digital technologies. In recent years, the concept of trust in digital contexts—known as e-trust—has come to the fore [12]. According to Taddeo, "e-trust occurs in environments where direct and physical contacts do not take place, where moral and social pressures can be differently perceived, and where interactions are mediated by digital devices." However, it is beyond the scope of this chapter to further elaborate on this concept. Our objective is to explore the relations between trust and data markets and how trust could be put into effect in the data markets.

2.2 The Role of Trust in Data Markets

A study on data sharing between companies in Europe identified key characteristics of a thriving data-driven economy. They include, among others, the availability of data sets from actors across the economy and the necessary infrastructure, knowledge, and skills within companies that would make possible to engage in data sharing and re-use. Other features included the existence of trust between independent economic operators, appropriate cybersecurity measures, and the development of common standards for technologies and data interoperability [13].

Trust between data suppliers and data users is one of the success factors for data sharing between companies (ibid., 83). There are different visions to successfully build trust, such as high security levels, enabling communication between data suppliers and users, and providing clarity with respect to what will be ultimately done with users' data (ibid.). Other ways include "empowering data suppliers and giving them full control over their datasets" and providing "comprehensive licensing agreements outlining data usage conditions and restrictions" (ibid.). Finally, informing data users about the origin of the data and lawfulness of data sharing activities have also been identified as key in building trust (ibid.).

In the context of data marketplace, enhancing trust requires a trusted third-party intermediary who brings data suppliers and data users together to exchange data in a secure online platform. TRUSTS goal is to create such a secure and trustworthy European data market. Against this background, how can one ensure that a data marketplace fulfils its role of the "trustworthy" intermediary?

2.3 Privacy-Preserving Techniques as a Means to Bring More Trust in Data Sharing

Privacy-preserving techniques play a crucial role for bringing trust to data markets and ensuring that personal data remains under the control of data subjects and is further shared with no harm on fundamental rights and freedoms of individuals.

Traditionally, the applicable legal regime will depend on the nature of data (personal/non-personal) at stake. In order to assess whether data on which privacy-preserving or re-identification techniques have been performed are considered as personal data or as anonymous information, the following criteria shall be used. First, the personal or non-personal character of the data depends on the identifiability of the individual (the data subject). The identifiable natural person is an individual who can be identified, directly or indirectly, in particular by reference to an identifier, inter alia a name, an identification number, location data, or an online identifier.[9]

Second, identifiability also depends on the capacity of actors to reverse an anonymization process with a decryption key or direct identifiers.[10] The identifiability is a dynamic concept. While it may not be possible to identify someone today with all the available means, it may happen at a later stage due to a technological progress. To determine whether an individual is identifiable, Recital 26 of the GDPR underlines that account should be taken of all the means reasonably likely to be used to identify the natural person directly or indirectly. This includes all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing.[11]

Furthermore, according to the CJEU, the abovementioned concept of "means reasonably likely to be used" does not imply that all the information enabling the identification of the data subject is in the hands of one person, i.e., the data controller.[12] Where additional data are required to identify the individual, what matters is the means reasonably likely to be used in order to access and combine such additional data (ibid.). As an illustration, dynamic IP addresses constitute personal data for online media service providers that can legally obtain required additional information held by internet service providers to identify an individual behind a dynamic IP address at a specific moment of time (ibid. para 47–48).

On the one hand, there is an absolute approach supporting that data on which privacy-preserving techniques have been applied will almost always remain personal as long as it is possible to reverse the process and identify the individual. Furthermore, it is also claimed that no technique is "perfect" and endurable against future technological developments [14]. On the other hand, a relative, risk-based approach builds on the criterion of "means that are reasonably likely to be used" in order to identify an individual.[13] Following the latter, privacy-preserving techniques provide for different degrees of re-identification taking into account contextual

[9] GDPR, Art. 4 (1).

[10] Article 29 Working Party, 'Opinion 4/2007 on the concept of personal data' (WP 136, 20 June 2007) p. 19–20.

[11] GDPR, Rec. 26.

[12] CJEU 19 October 2016 C582/14 Patrick Breyer v Bundesrepublik Deutschland ECLI:EU:C:2016:779 ('Breyer case') para 43–45.

[13] Article 29 Working Party, 'Statement on the role of a risk-based approach in data protection legal framework' (WP 218, 30 May 2014).

elements, such as the technical process, the safeguards restricting access to the data, and the overall realistic risk of re-identification. In other words, if excessive effort, in technical, organizational, and financial terms, is required for reversing privacy-enhancing techniques, the re-identification of the natural person may not be considered as likely.

Anonymization, for instance, is considered to provide for different levels of re-identification. If we apply the absolute approach, only data that have been irreversibly anonymized and whose original raw data set has been deleted may be considered as data that are no longer personal.[14]

When it comes to encryption, the GDPR does not define "encrypted data" or "encryption" but refers to encryption in several provisions as a risk mitigation measure. Encryption is listed as one of the "appropriate safeguards" of Article 6(4)(e) GDPR and is mentioned as an appropriate technical and organizational measure to ensure the security of processing.[15]

Since the GDPR does not define "encrypted data," it has to be examined whether encrypted data are anonymous or pseudonymous data. As it has been mentioned above, the answer to this question depends on whether an absolute or a relative approach regarding the identifiability of a data subject is applied. When personal data are encrypted, the data will always remain personal to the holders or to the authorized users of the decryption key. However, encrypted data may even be considered as personal if there are means reasonably likely to be used by others for decrypting them [15]. If encryption prevents an unauthorized party from having access to data, then the data in question no longer refer to an identified or identifiable person [14]. Consequently, it has to be examined which level of encryption is sufficient for the encrypted personal data to be considered as anonymous. Such an evaluation of the encryption method should take account of objective factors. These include the level of security of encrypted data and decryption prevention, such as the strength of the encryption algorithm used, the length of the encryption key, and the security of the key management [15].

Importantly, we have to distinguish between encrypted transfer of data (e.g., via end-to-end encryption) and encrypted storing of data (e.g., in a cloud) [14]. Processing of stored encrypted data is possible by using fully homomorphic encryption (FHE) or secure multi-party computation (MPC). In such a scenario, for the processing of the data, no decryption and thus no knowledge of the private key is needed. Moreover, the result of the processing is encrypted and can only be decrypted by the user and not by the cloud provider. The cloud provider will never see the data in plaintext. Thus, when processing personal data with the use of FHE, the GDPR is not applicable to the cloud provider which consequently does not process personal data (ibid.).

[14] Article 29 Working Party, 'Opinion 05/2014 on Anonymisation Techniques' (WP 216 10 April 214).

[15] GDPR, Article 32 (1)(a).

Therefore, encrypted personal data will be anonymous data, when it would require an excessively high effort or cost or it would cause serious disadvantages to reverse the process and re-identify the individual. It has to be considered whether there are reasonably likely means which could give a third party a potential possibility of obtaining the key. For instance, MPC allows data to be shared in a secret form (i.e., encrypted), while at the same time meaningful computations are performed on these data. Once the data have been divided into the shares, it is stored on different servers. At no point in this process, parties involved in data sharing and computing on the data—other than the data controller—can have access to the data [16].

Spindler et al. rightly argue that when applying an *absolute* approach on the identifiability of data subjects, these data shares would have to be considered as personal data. It is theoretically possible that all fragments of the data are gathered and put together; however, in practice, this is highly unlikely (ibid.). This unreasonable chance of collusion may lead to ruling out the applicability of the GDPR.

In addition to these concepts, the GDPR has introduced the notion and definition of "pseudonymization." More specifically, pseudonymization refers to 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. Furthermore, such additional information shall be kept separately and shall be subject to technical and organizational measures preventing the identifiability of a natural person.[16] Pseudonymization is commonly perceived as a data security measure that reduces linkability by replacing any identifying characteristic or attribute by another identifier, a pseudonym.[17] According to the GDPR, pseudonymized data are personal data.[18] Thus, data could be considered pseudonymized, and hence personal, insofar as the technical process they have undergone is reversible.

Nevertheless, it remains questionable whether reversibly anonymized, encrypted, and split data will be considered as personal, pseudonymized data or whether they will be referred to as anonymous toward the parties that cannot access the additional information, reverse the technical process, and identify the individual [14].

In the next section, we will provide a detailed technical description of these privacy-preserving techniques.

[16] GDPR, Art. 4 (5).

[17] Article 29 Working Party, 'Opinion 05/2014 on Anonymisation Techniques' (WP 216 10 April 2014).

[18] GDPR, Rec. 26.

3 Methods for Privacy-Preserving Analytics

Throughout the centuries, cryptographic ciphers have been designed to protect stored data or, with the emergence of modern information transmission, also to protect data in transmission. These scenarios usually follow an all-or-nothing principle where, e.g., two parties can access full information and outsiders nothing or where only the data owner has full information and nobody else. In reality, trust relationships are often a lot more complicated and diverse of course as we have seen in the previous sections, especially when it comes to outsourcing computations or accessing pre-trained machine learning models. Some of the very successful cryptosystems like RSA, for example, also have a special and usually unwanted property that allows to do limited calculations on the encrypted ciphertexts while preserving structure (called homomorphic property) to the unencrypted data. This means adding two ciphertexts yields the encrypted version of the plaintext sum, for example. These partial homomorphic properties led to a quest for new cryptosystems which turn the unwanted side effect into an advantage and allow unlimited manipulations and calculations on encrypted data. This opened up a new era of cryptography that allows to evaluate functions on encrypted, unknown data and to anchor cryptographic privacy-preserving methods in modern data analytics. The applications of such privacy-preserving techniques are widespread and range from evaluations of medical data [17, 18], over data mining [19] to applications in finance [20]. In this section, we give an overview of two main cryptographic protocols and primitives, FHE [21] and MPC [22], and discuss their links to data platforms and data sharing spaces. Additionally, we also introduce private set intersection (PSI) as a special MPC case.

3.1 Homomorphic Encryption

The introduction of "A fully homomorphic encryption system" by Craig Gentry [21] is regarded as one of the biggest advances in modern cryptography. Since then, many variations and improvements of (fully) homomorphic encryption have been developed. The main principle behind FHE is to start from a Somewhat Homomorphic Encryption (SHE) scheme that allows a limited number of operations. Gentry then introduced a technique called *bootstrapping* to refresh the ciphertexts to allow additional operations. Repeating the process opened the door for unlimited operations resulting in the change from somewhat to fully homomorphic encryption.

The starting point of all cryptographic protocols are mathematical problems that are very hard to solve (at least given appropriate constraints regarding time or computational power). The modern versions of FHE are based on such hard problems called Learning with Errors or an optimized variant thereof, which are formulated on mathematical lattices [22]. The security in these schemes comes from the introduction of random noise into the ciphertexts, which is removed

again during the decryption process. The main bottleneck of such approaches is that this noise starts to grow for each computed operation, e.g., adding two ciphertexts results roughly in doubling the original noise. Once a certain threshold has been reached, the resulting ciphertext cannot be decrypted anymore because the randomness prevails over the actual encrypted information. Before this point is reached, the bootstrapping process comes into play and allows to start over with a fresh noise budget by re-encrypting the original ciphertext into a new ciphertext with lower noise. This leads to a high-performance overhead for bootstrapping, and in several libraries, this functionality is therefore not even implemented at the moment. Instead, SHE is much more efficient and already sufficient for typical encrypted evaluations. Very complex evaluations cannot be realized with SHE because the number of calculations is limited.

In general, one of the main advantages of homomorphic encryption is the ability to outsource computation without giving up any privacy. Sensitive data can be homomorphically evaluated on a data platform or cloud, and only the data owners can decrypt computed results. Suppose you want to benefit from the evaluation of a machine learning model from a service provider, but you don't want to share your data with anyone outside your company. Setting up an FHE framework will allow you to do this without having to trust the service provider since they are not able to access the actual content of your data. An example of such a platform for medical data has been developed by researchers and engineers from the École Polytechnique Fédérale de Lausanne and the Lausanne University Hospital, for example [24]. They also use multi-party computation which we discuss in the next section. Another example of the advantages of FHE is the connection of human mobility to infectious diseases, where typically sensitive and private data have to be jointly evaluated to link these two fields. An efficient FHE implementation of a protocol where two parties can securely compute a Covid heatmap without revealing sensitive data was recently published [25, 26].

3.2 Secure Multi-Party Computation

Secure multi-party computation is a subfield of cryptography that enables privacy-preserving computations between multiple participants. It first appeared in computer science literature around 1980. In recent years, secure multi-party computation has become practical due to extensive ongoing research and exponential growth in computing power. Every traditional computation involving two or more participants can be made privacy-preserving through secure multi-party computation. However, this transformation's computational overhead varies depending on the underlying computation and sometimes can be prohibitive. To illustrate the privacy and confidentiality guarantees offered by secure multi-party computation, we consider the case of anti-money laundering. As with most anti-fraud activities, anti-money laundering benefits from collaboration. However, financial institutions are reluctant to share data because of competition and data protection regulations.

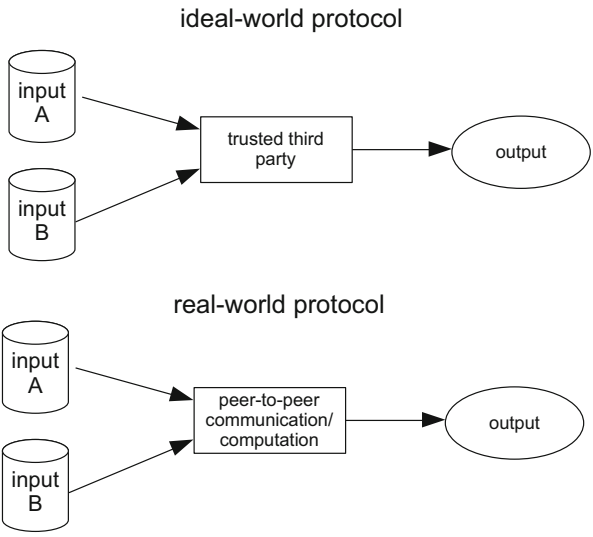

Fig. 1 Secure multi-party computation

A secure multi-party anti-money laundering computation would flag suspicious transactions without revealing any other information. To understand what this means, imagine an ideal world where there exists a hypothetical trusted third party. In this ideal world, every institution sends its data to the trusted third party which performs the anti-money laundering computation and reports back to the institutions about any detected suspicious behavior. Because the trusted third party cannot be corrupted, nothing except the output of the computation gets shared between institutions.

Secure multi-party computation provides similar confidentiality and privacy in the real world, where one cannot fully trust third parties. Therefore, what can be achieved in the ideal world can also be done by applying secure multi-party computation, as illustrated in Fig. 1.

3.2.1 Private Set Intersection

Private set intersection is a special-purpose secure multi-party computation. It allows two participants to compute the intersection of their data sets. Thereby, neither participant learns information from the protocol execution, except for the data entries in the intersection. For instance, private set intersection enables two companies to find out common customers privately—information that can subsequently be used for a joint advertising campaign. Note that, in Fig. 2, the output of the protocol is John, but company A would not know about company B's customers Marlene and Elsie. Private set intersection is the most mature secure multi-party protocol, and computational overhead is small. Therefore, when parties

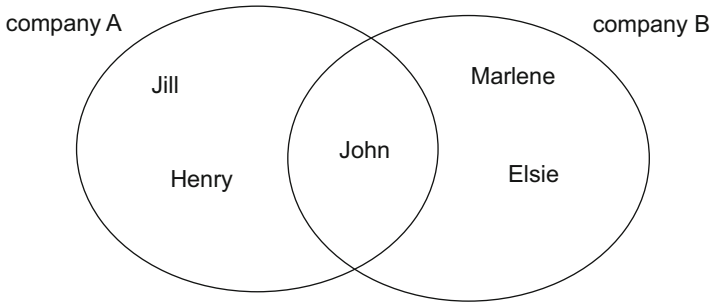

Fig. 2 Basic principle of private set intersection

engage in a private set intersection protocol, they do not have to expect significant performance issues.

4 Privacy-Preserving Technologies for Smart Contracts

Smart contracts are another example of where privacy-preserving techniques can be applied. They enact digital transactions that in a manner are similar to a physical transaction underpinned by a legal contract. Smart contract applications in a blockchain environment function within the context of the blockchain. Blockchains were not originally designed for preserving privacy; their original purpose was to verify integrity and legitimacy via transaction chains rooted in cryptographic hashes. In a public blockchain, data is available to all participants in unencrypted form – a problematic design for privacy preservation; off-chain smart contracts with hashes stored on-chain for verification purposes are a notable solution to this design problem [27].

Some blockchain variants can mitigate privacy concerns. Private and consortium blockchains utilize one or many managing authorities, and only approved authority members can access the blockchain data, but these infrastructures are typically much smaller than their public counterparts. The large, decentralized nature of public blockchains typically offers stronger security and integrity while foregoing the privacy and confidentiality controls of private and consortium blockchain variants [28].

4.1 Encrypted On-Chain Data with Homomorphic Encryption

This approach stores personal data on-chain in encrypted form. Applications cannot typically process encrypted data, and all participants on the blockchain will have visibility of any decryption operation, revealing both data and cryptographic keys. Homomorphic encryption, described earlier, enables operations on encrypted

data, preserving the privacy of on-chain smart contracts. However, the mentioned performance bottlenecks of FHE are currently a limiting factor for enterprise-level blockchain scenarios, and more research is needed in this regard.

4.2 Smart Contracts Based on Multi-party Computation

MPC splits personal data into specific subsets, ensuring that each subset is meaningless individually. The data owner sends each subset to a separate actor for processing. Processing only one data subset renders each processing actor unable to infer any further understanding of the source data, but the data owner can recombine the computational results from each actor into a complete output. MPCs are theoretically highly collusion resistant as every actor must collude to infer the source data's meaning. Personal smart contract data could, as such, be safely computed using MPC.

4.3 Secure Enclaves

Secure enclaves, or SEs, conceal program state and segregate enclaved code from external access. SEs are provided by trusted execution environments (TEEs)—secure CPU sections supported on several modern CPUs. Coupling SEs and asymmetric-key cryptography enables encryption of smart contracts using an SEs' public key, with the private key held in the SE; thus, the smart contract ciphertext can only be decrypted within that SE.

A chief issue with SEs is certain companies dominating the TEE hardware space, which creates a reliance on a less diverse set of chip architectures; this increases the possible impact of any security flaw found in one such widely adopted architecture—further compounded by past practical attacks, such as "Meltdown" and "Spectre," targeting such architectures. Another argument against TEEs purports that commercial TEE implementations are not necessarily publicly visible and, in these cases, can't be as rigorously analyzed as, say, public specifications from the Trusted Compute Group on which such implementations are based [29].

5 Conclusion: Opportunities and Future Challenges

The notion of enhancing trust in data sharing is present in various European Commission's documents, including the European strategy for data and the proposal for the Data Governance Act. The Commission intends to continue its work on the setting up of common rules for EU-wide common interoperable Data Spaces which would address issues of *trust*. First, clear and trustworthy rules for data sharing and

Data Governance are needed. However, it remains to be seen whether the DGA and other Commission's initiatives will fulfil its promise to "increase trust in data sharing services."

Second, data transaction involving personal data would benefit from further explanation in the text of the DGA on how privacy-preserving techniques could increase the level of trust and control of data holders over their personal data in their personal Data Spaces.

Regarding the technical aspects of privacy-preserving methods, future research should address the current performance bottlenecks to allow efficient and secure computations also for complex scenarios. This will enable also a more widespread application of privacy-preserving analytics for data sharing spaces and beyond. With the possible rise of quantum computers, there is also a growing need for long-term secure systems; methods like FHE that rely on lattice-based problems are already regarded as quantum-secure.

In general, the mutual understanding of legal frameworks, the benefits of data sharing spaces, and the corresponding technical capabilities will form an essential building block of a sustainable and secure European economy in the future.

Acknowledgments This work was supported by EU's Horizon 2020 project TRUSTS under grant agreement n°871481 and by the "DDAI" COMET Module within the COMET (Competence Centers for Excellent Technologies) Programme, funded by the Austrian Federal Ministry for Transport, Innovation and Technology (bmvit), the Austrian Federal Ministry for Digital and Economic Affairs (bmdw), the Austrian Research Promotion Agency (FFG), the province of Styria (SFG), and partners from industry and academia. The COMET Programme is managed by FFG.

References

1. Zillner, S., Curry, E., Metzger, A., Auer, S., & Seidl, R. (Eds.). (2017). *European big data value strategic research & innovation agenda*. Big Data Value Association.
2. Zillner, S., Bisset, D., Milano, M., Curry, E., García Robles, A., Hahn, T., Irgens, M., Lafrenz, R., Liepert, B., O'Sullivan, B. and Smeulders, A. (Eds.) (2020, September). *Strategic research, innovation and deployment agenda—AI, data and robotics partnership. Third release.* BDVA, euRobotics, ELLIS, EurAI and CLAIRE.
3. Glennon, M., et al. (2020). *The European data market monitoring tool*. EU Publications.
4. European Commission. (2020). *A European strategy for data COM/2020/66 final.* s.l.:s.n.
5. European Commission. (2019). *Communication from the Commission to the European Parliament and the Council. Guidance on the Regulation on a framework for the free flow of non-personal data in the European Union COM/2019/250 final.* s.l.:s.n.
6. European Commission. (2018). *Guidance on sharing private sector data in the European data economy, Accompanying the document "Towards a common European data space"*. s.l.:s.n.
7. Spiekermann, M. (2019). Data marketplaces: Trends and monetisation of data goods. *Intereconomics, 54*(7), 208–216.
8. Koutroumpis, P., Leiponen, A. & Thomas, L. D. (2017). The (unfulfilled) potential of data marketplaces (No. 53). *ETLA Working Papers.*
9. Richter, H., & Slowinski, P. R. (2018). The data sharing economy: On the emergence of new intermediaries. *IIC - International Review of Intellectual Property and Competition Law, 50*(12), 4–29.

10. Plantin, J.-C., Lagoze, C., & Edwards, P. N. (2018). Re-integrating scholarly infrastructure: The ambiguous role of data sharing platforms. *Big Data & Society, 5*(1), 205395171875668.
11. Gambetta, D. (1988). Can we trust trust? In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (pp. 213–237). Blackwell.
12. Taddeo, M. (2009). Defining trust and E-trust. *International Journal of Technology and Human Interaction, 5*(4), 23–35.
13. Arnaut, C. et al. (2018). *Study on data sharing between companies in Europe.* s.l.:s.n.
14. Spindler, G. & Schmechel, P. (2016). *Personal data and encryption in the european general data protection regulation.* s.l.:s.n.
15. Hon, W. K., Millard, C., & Walden, I. (2011). The problem of \textquotesinglePersonal Data\textquotesingle in cloud computing - what information is regulated? The cloud of unknowing, Part 1. *SSRN Electronic Journal*.
16. Roman, D., & Vu, K. (2019). Enabling data markets using smart contracts and multi-party computation. In *Business Information Systems Workshops* (pp. 258–263). Springer International Publishing.
17. Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence, 2*(6), 305–311.
18. Thapa, C., & Camtepe, S. (2021). Precision health data: Requirements, challenges and existing techniques for data security and privacy. *Computers in Biology and Medicine, 129*, 104130.
19. Lindell, Y., & Pinkas, B. (2000). *Privacy preserving data mining* (pp. 36–54). Springer.
20. Masters, O. et al. (2019). *Towards a homomorphic machine learning big data pipeline for the financial services sector.* s.l.:s.n.
21. Gentry, C. (2009). *Fully homomorphic encryption using ideal lattices* (pp. 169–178). Association for Computing Machinery.
22. Yao, A. C. (1986). *How to generate and exchange secrets* (pp. 162–167). s.l.: s.n.
23. Regev, O. (2009). On lattices, learning with errors, random linear codes, and cryptography. *Journal of the ACM, 56*(9).
24. MedCO. (2020). https://medco.epfl.ch. s.l.:s.n.
25. Bampoulidis, A. et al. (2020). *Privately connecting mobility to infectious diseases via applied cryptography.* s.l.:s.n.
26. Covid-Heatmap. (2021). https://covid-heatmap.iaik.tugraz.at/en/. s.l.:s.n.
27. Neubauer, M. & Goebel, A. (2018). *Blockchain for off-chain smart contracts.* s.l.:s.n.
28. Sharma, T. K. (2020). *Types of blockchains explained.* s.l.:s.n.
29. Anderson, R. (2003). *"Trusted computing" Frequently asked questions.* s.l.:s.n.