D3.1 Overview of Existing Technologies

e-SIDES

Ethical and Societal Implications of Data Sciences

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731873

e-SIDES – Ethical and Societal Implications of Data Sciences

Data-driven innovation is deeply transforming society and the economy. Although there are potentially enormous economic and social benefits this innovation also brings new challenges for individual and collective privacy, security, as well as democracy and participation. The main objective of the CSA e-SIDES is to complement the research on privacy-preserving big data technologies, by analyzing, mapping and clearly identifying the main societal and ethical challenges emerging from the adoption of big data technologies, conforming to the principles of responsible research and innovation; setting up and organizing a sustainable dialogue between industry, research and social actors, as well as networking with the main Research and Innovation Actions and Large Scale Pilots and other framework program projects interested in these issues. It will investigate stakeholders' concerns, and collect their input, framing these results in a clear conceptual framework showing the potential trade-offs between conflicting needs and providing a basis to validate privacy-preserving technologies. It will prepare and widely disseminate community shared conclusions and recommendations highlighting the best way to ultimately build confidence of citizens and businesses towards big data and the data economy.

This document reflects the view of its authors only.

The European Commission is not responsible for any use that may be made of the information this document contains.

Use of any materials from this document should be referenced and is at the user's own risk.

D3.1 Overview of existing technologies

Work package	WP 3 – Review of existing technologies
Lead author	Daniel Bachlechner (Fraunhofer ISI)
Contributing authors	Michael Friedewald (Fraunhofer ISI) Jana Weitkamp (Fraunhofer ISI) Nicholas Martin (Fraunhofer ISI)
Internal review	Bart Custers (Leiden University) Karolina La Fors (Leiden University) Richard Stevens (IDC)
Due Date	M12 (December 2017)
Date	18 January 2018
Version	1.0
Type	Report
Dissemination level	Public

This document is Deliverable 3.1 of Work Package 3 of the e-SIDES project on Ethical and Societal Implications of Data Science. e-SIDES is an EU funded Coordination and Support Action (CSA) that complements Research and Innovation Actions (RIAs) on privacy-preserving big data technologies by exploring the societal and ethical implications of big data technologies and providing a broad basis and wider context to validate privacy-preserving technologies. All interested stakeholders are invited to look for further information about the e-SIDES results and initiatives at www.e-sides.eu.

Executive Summary

This report provides an overview of existing approaches, methods and technologies that may have the potential to address ethical, legal, societal and economic issues raised by big data applications. Among the issues are threats to privacy and self-determination, strong interdependencies, limited trustworthiness and lack of accountability. While numerous issues are taken into account, threats to privacy receive particular attention. Based on a review of the comprehensive body of literature, technologies considered as privacy enhancing or privacy-preserving were identified and assigned to eleven classes. The classes are anonymisation, sanitisation, encryption, multi-party computation, access control, policy enforcement, accountability, data provenance, transparency, access and portability, and user control.

Anonymisation is performed by encrypting or removing personally identifiable information from datasets. Traditional anonymisation techniques fail in the context of big data applications because there are too many data points for a single individual. Sanitisation is done by encrypting or removing sensitive information from datasets. Anonymisation is a type of sanitisation. Encryption is the encoding of information so that only authorised parties can access it. In the context of big data applications, it is necessary to go beyond the "encrypt all or nothing" model. Multi-party computation is a field of cryptography that relies on the distribution of data and processing tasks over multiple parties. Although it was proven to be theoretically plausible, there are still no practical solutions.

Access control describes the selective restriction of access to places or resources. Big data applications typically require fine-grained access control and traditional approaches increasingly become unmanageable. Policy enforcement focuses on the enforcement of rules for the use and handling of resources. Automated policy enforcement mechanisms are considered particularly important in the big data era. Accountability requires the evaluation of compliance with policies and the provision of evidence. A cornerstone of accountability in the context of big data applications is the provision of automated and scalable control and auditing processes. Data provenance relies on being able to attest the origin and authenticity of information. The aim is to provide a record of the processing history of pieces of data.

Transparency is the providing intelligible and easily accessible information regarding the approach and algorithms for data collection and processing. It may be achieved by providing purely textual information, through multichannel and layered approaches, or standardized icons and pictograms. Access and portability facilitates the use and handling of data in different contexts. Having access to data means that data subjects can look through and check the data stored. Portability gives data subjects the possibility to change service providers without losing their data. User control refers to the specification and enforcement of rules for data use and handling. Means that allows reaching user control include consent mechanisms, privacy preferences, sticky policies and personal data stores.

As the project aims to complement other research on privacy-preserving big data technologies and data-driven innovation funded by the European Union under the H2020 programme, the role the classes of technologies play in such projects was also studied. For this purpose, project representatives were interviewed personally or asked to provide relevant information per e-mail. Moreover, the websites as well as already published deliverables of relevant projects were analysed. It seems that the emergence of big data changes the protection of privacy as well as the relevance of other related issues significantly. Finding an adequate balance between the protection of privacy and maintaining the utility of big data seems to be challenging. The amounts of attention that other projects funded by the European Union devote to privacy and related issues vary significantly.

Contents

Figures

Tables

Abbreviations

ABAC	Attribute based access control
ABE	Attribute based encryption
ACL	Access control list
AES	Advanced Encryption Standard
bABE	Broadcast Ciphertext Policy Attribute Based Encryption
C-PRE	Conditional PRE
CP-ABE	Ciphertext Policy Attribute Based Encryption
CSA	Coordination and Support Action
DAG	Directed Acyclic Graph
DOB	Date of birth
DTE	Deterministic Encryption
ENISA	European Union Agency for Network and Information Security
EU	European Union

FHE	Fully Homomorphic Encryption
GDPR	General Data Protection Regulation
HIBE	Hierarchical IBE
HPE	Hierarchical Predicate Encryption
IA	Innovation Action
IBE	Identity-Based Encryption
ICT	Information and Communication Technology
IoT	Internet of Things
KP-ABE	Key Policy Attribute Based Encryption
MA-ABE	Multiauthority Attribute Based Encryption
MPC	Multi-Party Computation
OPE	Order-Preserving Encryption
PBAC	Policy-Based access control
PEKS	Public Key Encryption with Keyword Search
PET	Privacy Enhancing Tool
PDP	Provable Data Processing
PIR	Private Information Retrieval
PKE	Public Key Encryption
POR	Poofs of retrievability
PPDM	Privacy-Preserving data mining
PPDP	Privacy-Preserving data publishing
PPE	Property-Preserving Encryption
PPP	Private Public Partnership
PRE	Proxy re-encryption
RBAC	Role-based Access Control
RIA	Research and Innovation Action
SKE	Symmetric Key Encryption
SSE	Symmetric Searchable Encryption
TPA	Third Party Auditor
TTP	Trusted Third Party
XACML	Extensible Access Control Mark-up Language

1. Introduction

This chapter outlines the background, the methodology and the structure of this document.

1.1. Background

This report is Deliverable 3.1 of the e-SIDES project. In this project, the ethical, legal, societal and economic implications of big data applications are examined in order to complement the research on privacy-preserving big data technologies (mainly carried out by ICT-18-2016 projects) and data-driven innovation (carried out, for instance, by ICT-14-2016-2017 and ICT-15-2016-2017 projects).

The first step in the project was to identify important ethical, legal, societal and economic issues that are raised by big data applications[1]. The second step, which is the focus of this report, is to provide an overview of existing technologies that may have the potential to address some of the issues. The third step will be to assess the potential the technologies actually have in this regard.

Threats to privacy is only one of the issues identified, but one that has received particular attention in the context of big data applications. Not less important issues include, for instance, threats to self-determination, strong interdependencies, limited trustworthiness and lack of accountability. Among the reasons for the particular interest in privacy are that its protection has a long tradition in Europe and beyond and that the emergence of new information and communication technologies (ICTs) is known to have posed threats to privacy before.

Due to the particular relevance of privacy issues in the context of new ICTs, it has been tried to address the issues, among others, by means of technological measures. One way is to create new technologies that address issues of the previous technologies. A typical example of such a technological solution are anonymisers. These are technological tools to anonymise personal data and aim to mitigate privacy issues. Such technological solutions are called **privacy enhancing tools** (PETs).

Although such technological solutions may be helpful in addressing ethical and societal issues, the e-SIDES project does not focus on these types of technological solutions but rather on a new approach, in which new technologies are designed in ways that they actually are already privacy-preserving when they are ready for their first time use. The idea is to distil design requirements from ethical, legal, societal and economic perspectives and to build new technologies that take these requirements into account. In other words, privacy requirements are taken into account throughout the entire engineering process. This is what is called **Privacy by Design**[2] and becomes mandatory through Art. 25 of the new European Union (EU) General Data Protection Regulation (GDPR).

Obviously in the term privacy by design, can be transformed into other equivalents, such as ethics by design, legitimacy by design, trust by design and equality by design. To avoid a plethora of terminology, we will use the term privacy by design as a term in which technologies are designed in such a way that

[1] See Deliverables D2.1 and Deliverable D2.2 available at www.esides.eu

[2] Ann Cavoukian and Jeff Jonas, "Privacy by Design in the Age of Big Data," (Information and Privacy Commissioner, Ontario, Canada, 2012), https://jeffjonas.typepad.com/Privacy-by-Design-in-the-Era-of-Big-Data.pdf (accessed December 14, 2017)

they take into account all kinds of ethical, legal and societal issues. Spiekermann[3], for instance, discusses aspects of such technology design extensively.

A typical example of a privacy by design approach is so-called **privacy-preserving data mining** (PPDM). When processing big data, data mining is a set of analysis tools that performs automatic analysis of data using mathematical algorithms, in order to find new patterns and relations in data.[4] These tools can be designed in a way in which such data mining preserves privacy. Hence, PPDM refers to data mining technologies with built-in requirements (based on privacy by design) to preserve privacy.[5]

However, since privacy by design, PPDM and related approaches and technologies are still in their infancy, a useful starting point in mapping existing technologies is to start with PETs, which are technological measures that protect privacy by eliminating or reducing personal data or by preventing unnecessary or undesired processing of personal data.[6] According to the PETs division of the German Informatics Society, privacy may be protected by:[7]

- data avoidance and data minimisation (i.e., the reduction of personal data that is collected, stored and processed);
- system data protection (i.e., the implementation of technical and organisational measures to protect personal data by those who collect, store and process data);
- self-data protection (i.e., the implementation of technical and organisational measures to protect personal data by those whose data is collected, stored and processed); and
- transparency and other confidence-building measures.

The e-SIDES project focuses in particular on technological measures of system data protection that are relevant in the context of big data. Technological measures of self-data protection such as anti-tracking tools and anonymising networks, although they have also received significant attention since the emergence of big data applications, are not in the focus of e-SIDES. Moreover, the project does not pay particular attention to non-technological (or organisational) measures to protect personal data. This means that organisational measures are addressed only if they play a key role as complements of technological measures.

The measures that aim at protecting privacy may also allow addressing ethical and societal issues other than privacy to some extent. Furthermore, there may also be technological measures, taken, for instance, from the field of cloud computing, that are not necessarily considered privacy-preserving technologies but that are useful to address one or several of the issues studied in the project. They are relevant for e-

[3] Sarah Spiekermann, *Ethical IT innovation: A value-based system design approach* (Boca Raton, London, New York: CRC Press, 2016)

[4] Toon Calders and Bart Custers, "What Is Data Mining and How Does It Work?," in *Discrimination and Privacy in the Information Society*, vol. 3, ed. Bart Custers et al., 27–42 3 (Berlin, Heidelberg: Springer, 2013)

[5] Bart Custers et al., eds., *Discrimination and Privacy in the Information Society* 3 (Berlin, Heidelberg: Springer, 2013)

[6] John J. Borking and Charles D. Raab, "Laws, PETs and Other Technologies for Privacy Protection," *Journal of Information Law & Technology*, no. 1 (2001), https://www2.warwick.ac.uk/fac/soc/law/elj/jilt/2001_1/borking (accessed September 26, 2017)

[7] https://fg-pet.gi.de/themen.html#c36

SIDES but as the range of potentially relevant technologies is extremely broad, they are not within the scope of the research conducted as the basis of this report.

Privacy by Design is highly relevant for e-SIDES as it is much more efficient to develop privacy-preserving big data solutions right away than to add PETs to them at a later stage.

Algorithmic accountability is a concept closely related to ethical and societal implications of big data applications and thus also relevant from the e-SIDES perspective. The concept has received considerable attention recently. Obviously, apart from the data itself also the algorithms used to search for, structure and deliver data must be taken into account when studying the implications of big data. As the product of humans, algorithms can have issues resulting from human bias or oversight. Algorithmic accountability is promoted as a way to help such issues be recognized and corrected. The concept goes hand in hand with algorithmic transparency, which requires companies to be open about the purpose, structure and underlying actions of the algorithms used.

1.2. Methodology

The basis of the presented technologies and classifications is a review of literature focusing on technological measures to protect personal data in the context of big data. The structure of the overview of technologies extends the one used in a report published by the European Union Agency for Network and Information Security (ENISA). The respective ENISA report authored by D'Acquisto et al.[8] was published in 2015.

Around 200 articles on privacy-preserving big data technologies were reviewed in the second half of 2017 to make sure the report includes a comprehensive overview of existing technologies. The articles focus on technologies with different levels of development. While some are already used within the scope of big data solutions others still have a long way to go before they can be effectively used in the context of big data.

The technologies described in the literature were assigned to 11 technology classes (i.e., anonymisation, sanitisation, encryption, multi-party computation (MPC), access control, policy enforcement, accountability, data provenance, transparency, access and portability, and user control). The classes extend the classification proposed by D'Acquisto et al. with a few additional classes. Moreover, subclasses were introduced where useful. Referring to the top-level classes only, the technologies could be well discussed with representatives of related projects focusing on privacy-preserving big data technologies on data-driven innovation.

The discussions allowed checking the completeness of the overview of technologies and understanding the relevance individual classes of technologies have for related projects. The technologies were discussed with representatives of related projects face-to-face at the 2017 European Big Data Value Forum in Versailles and per e-mail. Additionally, the websites of the projects turned out to be useful to gain insight

[8] Giuseppe D'Acquisto et al., "Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics," (ENISA, 2015), https://www.enisa.europa.eu/publications/big-data-protection/at_download/fullReport (accessed September 26, 2017)

into the relationship between the project and the technologies. The ICT-18-2016 projects are of particular importance for e-SIDES.

1.3. Structure

This deliverable is structured as follows:

- Chapter 1 outlines the background, the methodology and the structure of the deliverable.
- Chapter 2 discusses approaches, methods and technologies to classify privacy-preserving big data technologies and proposes a classification to be used in e-SIDES.
- Chapter 3 describes classes of technologies in detail.
- Chapter 4 provides and overview of related projects and outlines the relevance of the classes of technologies for these projects.
- Chapter 5 concludes the deliverable.

2. Classifications

This chapter discusses approaches to classify technologies that are considered privacy-preserving or privacy enhancing. While some of the classifications focus on privacy-preserving big data technologies, others do not focus on the big data context explicitly. The discussed approaches use big data lifecycle phases, user roles involved in data mining applications, dimensions relevant in the context of privacy-preserving technologies and their application, priorities set in literature and different strategies as a basis.

2.1. Lifecycle

Jain et al.[9] as well as Mehmood et al.[10] classify privacy-preserving big data technologies according to their relevance for specific phases of the big data lifecycle. Figure 1 shows the lifecycle consisting of the phases data generation, data storage and data processing.

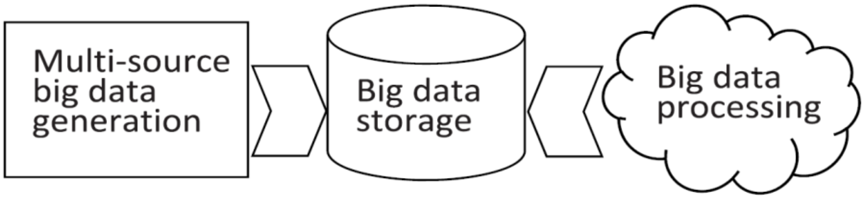


Figure 1 Big data lifecycle as depicted by Mehmood et al.

Data Generation can be classified into active data generation and passive data generation. Active data generation means that the data subject is willing to provide data to a third party, while passive data generation refers to situations in which data is generated, for instance, by the data subject's online activity. In this case, the data subject may not be aware of the data collection. Key challenges for data subjects are to protect personal and sensitive information and to keep control over their data. The risk of privacy violations during data generation can be minimised by restricting access or by falsifying data (data obfuscation). While access restriction techniques try to limit the access to private data, falsifying data techniques alter the original data before it is released to an untrusted party.

While storing high volume data is not a big challenge, securing the data can be very challenging. If the **Data Storage** system is compromised, it can be harmful as personal data may be disclosed. Therefore, it needs to be ensured that the stored data is protected against such threats. The conventional security mechanisms to protect data can be divided into four categories. They are file-level data security schemes, database-level data security schemes, media-level security schemes and application-level encryption schemes. The conventional mechanisms, which were optimised for existing storage architectures may not be applicable to big data, though. One promising technology to address the particular requirements of big data is storage virtualisation. However, using a cloud service means that data will be transmitted to a third party.

[9] Priyank Jain, Manasi Gyanchandani and Nilay Khare, “Big data privacy: A technological perspective and review,” *Journal of Big Data* 3, no. 1 (2016)
[10] Abid Mehmood et al., “Protection of Big Data Privacy,” *IEEE Access* 4 (2016)

According to Mehmood et al., the approaches to privacy protection in the data storage phase are mainly based on encryption techniques. Encryption-based techniques can, according to them, be further divided into attribute-based encryption (ABE), identity-based encryption (IBE) and storage path encryption (see the next chapter for more detailed explanations). In addition, to protect sensitive data, hybrid clouds are used where sensitive data are stored in a private cloud. With respect to integrity verification schemes, provable data processing (PDP), proofs of retrievability (POR) and public auditing are mentioned by Mehmood et al.

Privacy protection in **Data Processing** can be divided into two phases. In the first phase, the goal is to safeguard information from unsolicited disclosure because the collected data may contain sensitive information about the data subject. Privacy-preserving data publishing (PPDP) is a relevant concept for this phase. In the second phase, the goal is to extract meaningful information from the data without violating the privacy. This phase thus calls for PPDM.[11]

Mehmood et al. state that in PPDP, anonymisation techniques such as generalisation and suppression are used to protect privacy. Ensuring the utility of the data while preserving privacy is a great challenge in this context. In the knowledge extracting process, there exist several mechanisms to extract useful information from large-scale and complex data. These mechanisms can be further divided into clustering, classification and association-rule-mining-based techniques. While clustering and classification split the input data into different groups, association-rule-mining-based techniques find useful relationships and trends in the input data.

Jain et al. differentiate between traditional and recent methods. With respect to traditional methods, de-identification, HybrEx,[12] privacy-preserving aggregation and operations over encrypted data are discussed. The recent methods include differential privacy, identity-based anonymisation, a privacy-preserving a priori algorithm for the MapReduce framework, privacy-preserving big data publishing and fast anonymisation of big data streams. Moreover, Jain et al. discuss the specifics of privacy and security in healthcare big data applications in this context.

2.2. User roles

Xu et al.[13] classify privacy-preserving big data technologies based on their relevance for four user roles. The user roles used by Xu et al. are data provider, data collector, data miner and decision maker. Figure 2 shows the relationship between the different roles. The data provider is the originator of data that is desired for data mining. The data collector collects data from data providers and publishes it to the data miner. The data miner performs the data mining. Finally, the decision maker makes decisions based on the data mining results to achieve certain goals.

[11] Stan Matwin, "Privacy-Preserving Data Mining Techniques: Survey and Challenges," in *Discrimination and Privacy in the Information Society*, vol. 3, ed. Bart Custers et al., 209–21 3 (Berlin, Heidelberg: Springer, 2013)

[12] HybrEx stands for Hybrid Execution and is an execution model for confidentiality and privacy in cloud computing.

[13] Lei Xu et al., "Information Security in Big Data: Privacy and Data Mining," *IEEE Access* 2 (2014)

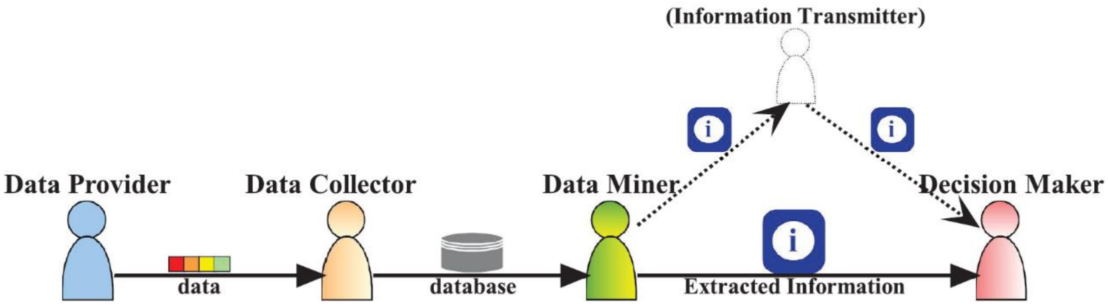


Figure 2 Data mining application scenario as specified by Xu et al.

Xu et al. investigate the technologies using a game-theoretical approach. The rationality is that, in data mining, each user pursues high self-interests in terms of privacy preservation or data utility, and the interests of different users are correlated. Hence, Xu et al. proceed from the assumption that the interactions among different users can be modelled as a game.

The major concern of a **data provider** is whether the data provided to others can be controlled. On the one hand, the provider should be able to make private data inaccessible to the data collector. On the other hand, if the provider has to provide some data to the data collector, sensitive information should be hidden as much as possible and enough compensation should be received for the possible loss in privacy. According to Xu et al., a data provider can limit the access to data, trade privacy for benefit and provide false data. [14] The first and the last option are closely related to those proposed by Jain et al.[15] and Mehmood et al.[16] for the data generation phase. The second option is somehow between the other two. Based on the benefit, a data provider can decide to provide no data, fuzzy data or accurate data.

The data collected from data providers may contain sensitive information. Directly releasing collected data to the data miner could thus violate the data providers' privacy, hence data modification is required. However, the data must still be useful after modification. Therefore, the major concern of the **data collector** is to guarantee that the modified data contains no sensitive information but still preserves high utility. PPDP is thus particularly relevant for data collectors. Before being published to others, the data needs to be anonymised or, more generally, sanitised. Xu et al. list the same anonymisation operations that are also mentioned, for instance, by Mehmood et al. including operations such as generalisation and suppression. Xu et al. go beyond what many other authors do by detailing how privacy-preserving publishing could look like in the context of social network data and trajectory data. Similarly, Jain et al. discuss big data privacy and security aspects in the context of healthcare.

A **data miner** applies mining algorithms to the data provided by a data collector, and wishes to extract useful information from the data in a privacy-preserving manner. PPDM covers two types of protections, namely the protection of the sensitive data itself and the protection of sensitive mining results. The user-role-based methodology proposed by Xu et al., proceeds from the assumption that the data collector should take the major responsibility of protecting sensitive data, while the data miner should focus on how to hide the sensitive mining results from untrusted parties. Xu et al. describe privacy-preserving

[14] Ibid.

[15] Jain, Gyanchandani and Khare, “Big data privacy”

[16] Mehmood et al., “Protection of Big Data Privacy”

association rule mining, privacy-preserving classification and privacy-preserving clustering in detail. These approaches are in line with what Mehmood et al. describe in the context of the knowledge extracting process.

A **decision maker** usually gets the data mining results directly from the data miner. Xu et al. also mention an information transmitter that may be involved. If involved, it is possible that the information transmitter changes the mining results intentionally or unintentionally, which could cause serious loss to the decision maker. Therefore, what the decision maker concerns is whether the mining results are credible. Data provenance and web information credibility are thus key aspects for the decision maker. For more details, see Xu et al.

2.3. Dimensions

Heurix et al.[17] propose a taxonomy for technologies that are considered privacy-preserving or privacy enhancing. Their taxonomy is intended to serve as a tool for the systematic comparison of different technologies. Such a comparison is expected to help identifying limitations of existing technologies, complementary technologies and potential research directions, and could thus also be valuable for e-SIDES. Heurix et al. do not explicitly focus on privacy-preserving big data technologies, though. As shown in Figure 3, they classify technologies based on seven primary dimensions.

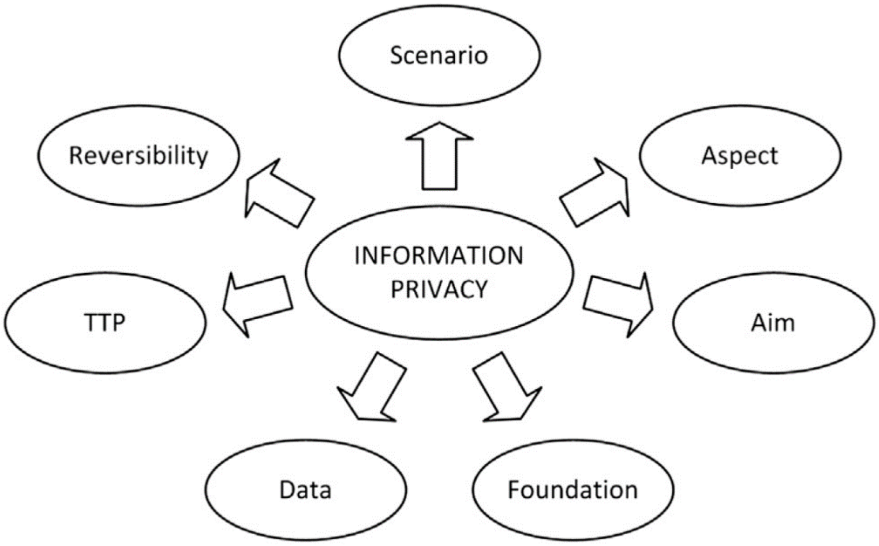


Figure 3 Primary dimensions used by Heurix et al.

The dimension **Scenario** defines the primary untrusted actor and potential risk source in a privacy-sensitive information exchange operation. Heurix et al. consider four scenarios. In the first scenario, the user of a service is regarded as not trustworthy and a potential risk source. The second scenario refers to a situation where the service provider aims to gain more information about the service user than necessary. In the third scenario, both actors do not trust each other. This scenario is also used for typical

[17] Johannes Heurix et al., "A taxonomy for privacy enhancing technologies," *Computers & Security* 53 (2015)

communication scenarios of equal partners. The last scenario refers to a situation where an external agent threatens the primary actors' privacy.

The privacy aspect addressed by a technology is defined by the **Aspect** dimension. In this context, Heurix et al. distinguish between identity, content and behaviour. Identity is the primary aspect of privacy and refers to hiding or masking the identity of involved persons. Identity protection can be enforced by either anonymity or pseudonymity. Content refers to hiding or masking the data content of a service and corresponds with the classic security attribute of confidentiality. Finally, behaviour refers to hiding the behaviour of actors, which includes, for instance, the access pattern of a user to a specific service.

A technology's purpose is defined by the **Aim** dimension. Heurix et al. describe four aims. The first one is indistinguishability, which makes it impossible to unambiguously distinguish one entity from another. This refers to hiding an individual within a specified set of individuals. Unlinkability, which is the second aim described by Heurix et al., means that entities cannot be related to others, where the entities are not necessarily of the same class (e.g., a person and corresponding medical data). The third aim is deniability. It refers to the ability to plausibly deny a fact, possession or transaction. Finally, confidentiality refers to the secrecy of a data fragment's content. It is usually achieved by encryption.

The dimension **Foundation** defines the underlying security model and cryptographic primitive. The security model may be an information-theoretic or a computational one. Heurix et al. stress, however, that most modern cryptographic algorithms are based on the computational security model. With respect to cryptographic primitives, Heurix et al. differentiate between symmetric, asymmetric and unkeyed algorithms such as hash algorithms. Moreover, algorithms may be non-cryptographic. Specific technologies may rely on one cryptographic primitive or a combination of multiple primitives.

The state of data that is addressed by a technology is defined by the *Data* dimension. With respect to the state of data, Heurix et al. differentiate between stored data (data-at-rest), transmitted data (data-in-motion) and processed data.

The necessity of a trusted third party (TTP) and its relevance for a technology is defined in the **Trusted Third Party** dimension. A TTP's task is to act as a broker or mediator between two or more parties to provide security- and privacy-critical support functions. The frequency attribute denotes how often a technology interacts with the involved parties. Such interactions may occur permanently, situational or never. The phase attribute describes in which specific phases of technology protocols a TTP is involved. The involvement may happen during regular operation, in the setup phase or not at all. Finally, the task attribute determines a potential TTP involvement by describing which task the TTP fulfils. According to Heurix et al., the involvement may concern a regular task, a validation task or no task.

The *Reversibility* dimension defines if and under which circumstances an operation of a technology is reversible. With respect to reversibility, Heurix et al. focus on the requirement of cooperation of the data originator and the degree to which an operation can be reversed. It may be that an operation can be reversed fully, partially or not at all. Moreover, an operation can be deniable, which means that the data originator is able to deniably reverse an operation.

Heurix et al. applied their taxonomy on several technologies including k-anonymity, l-diversity, t-closeness, 1-server and n-server private information retrieval (PIR), and search encryption (see the next chapter for more detailed explanations). The common goal of technologies that are considered privacy-preserving or privacy enhancing is to protect the privacy of subjects by protecting personal data or masking a subjects' behaviour, or otherwise ensure that sensitive information is not disclosed to unintended parties. The heterogeneity of the aims and properties of specific technologies makes it difficult to effectively categorise and compare them. The taxonomy proposed by Heurix et al. may prove useful for e-SIDES. Particularly, when the relevance of technologies for specific application contexts is discussed.

2.4. Trends in research

Nelson & Olovsson[18] quantitatively and qualitatively analysed 82 papers that focus on security and privacy in the context of big data and were published in proceedings of highest-ranking conferences between 2012 and 2015. In their analysis, the authors paid particular attention to confidentiality, data integrity, privacy, data analysis, visualisation, stream processing and data format in the context of big data. Nelson & Olovsson do not suggest a classification of technologies but they provide an overview of the attention certain technologies received over the last few years. This overview allows e-SIDES to better understand related trends in research.

With respect to privacy, Nelson & Olovsson came across several privacy models, such as k-anonymity, l-diversity, t-closeness and differential privacy, which can be used to anonymise data (see chapter 3.1). The first three are techniques for releasing entire sets of data through PPDP, whereas differential privacy is used for PPDM. Thus, differential privacy is obtained without processing the entire dataset. Therefore, anonymising larger datasets can be difficult from an efficiency perspective. However, larger datasets have greater potential to hide individual data points within the set.

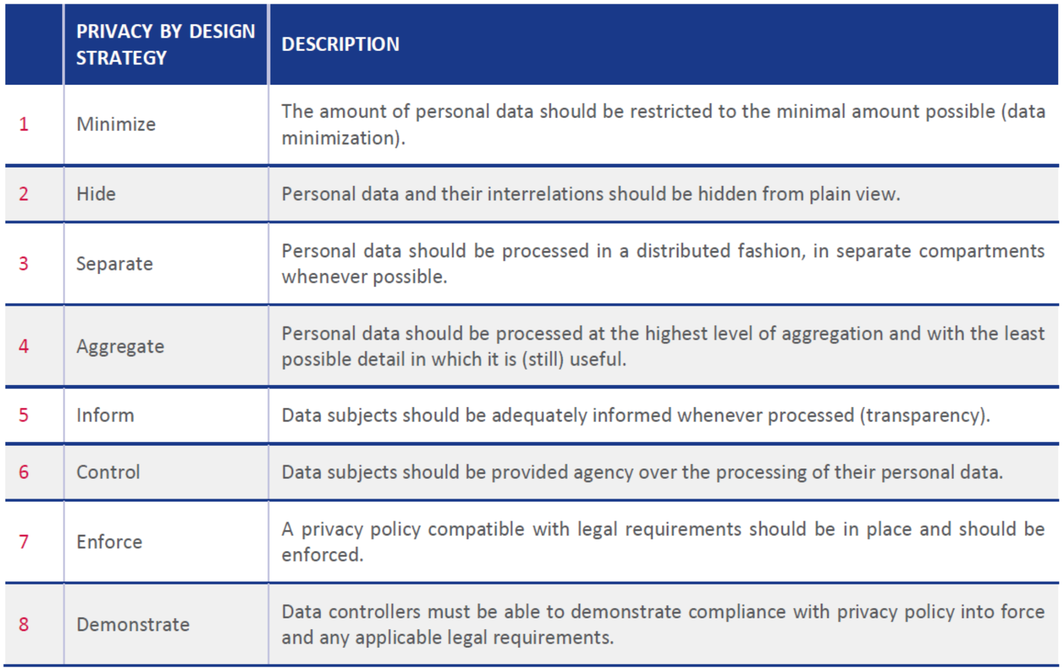
Out of a total of 61 privacy-oriented papers, one paper uses k-anonymity, and another paper uses l-diversity and t-closeness but also differential privacy to anonymise data. Another paper introduces a successor to t-closeness called β-likeness. In comparison, a large share, 46 papers, of the privacy-oriented papers focuses only on differential privacy as their privacy model. Most of them propose methods for releasing differentially private data structures.

An observation Nelson & Olovsson considered particularly interesting is that differential privacy can have a large impact on the accuracy of results. In one article, accuracy loss of 15% to 30% was the result of enforcing differential privacy on a telecommunications platform. In fact, guaranteeing differential privacy while maintaining high utility of the data is not trivial. From the reviewed papers, 15 investigated utility in combination with differential privacy.

2.5. Strategies

ENISA[19] proposes eight privacy-by-design strategies as the basis for the discussion of privacy-preserving big data technologies. A list of the strategies is provided in Figure 4.

[18] Boel Nelson and Tomas Olovsson, "Security and privacy for big data: A systematic literature review," in *2016 IEEE International Conference on Big Data: Dec 05-Dec 08, 2015, Washington D.C., USA : proceedings*, ed. James Joshi, 3693–702 (Piscataway, NJ: IEEE, 2016)
[19] D'Acquisto et al., "Privacy by design in big data"



	PRIVACY BY DESIGN STRATEGY	DESCRIPTION
1	Minimize	The amount of personal data should be restricted to the minimal amount possible (data minimization).
2	Hide	Personal data and their interrelations should be hidden from plain view.
3	Separate	Personal data should be processed in a distributed fashion, in separate compartments whenever possible.
4	Aggregate	Personal data should be processed at the highest level of aggregation and with the least possible detail in which it is (still) useful.
5	Inform	Data subjects should be adequately informed whenever processed (transparency).
6	Control	Data subjects should be provided agency over the processing of their personal data.
7	Enforce	A privacy policy compatible with legal requirements should be in place and should be enforced.
8	Demonstrate	Data controllers must be able to demonstrate compliance with privacy policy into force and any applicable legal requirements.

Figure 4 Privacy-by-design strategies proposed by ENISA

According to ENISA, these strategies are relevant in different phases of the big data analytics value chain. ENISA links privacy-enhancing big data technologies with strategies and strategies with big data value chain phases. The phases described by ENISA are related to the big data lifecycle and the data mining user roles introduced previously. The ENISA big data analytics value chain includes the four phases, namely, data acquisition/collection, data analysis and data curation, data storage and data use. Strategies can be assigned to more than one phase.

With respect to **data acquisition/collection**, the strategies *Minimize*, *Aggregate*, *Hide*, *Inform* and *Control* are considered relevant. To restrict the amount of personal data, ENISA stresses the importance of both defining what data is needed before starting with the data collection and carrying out a privacy impact assessment. To aggregate data, local anonymisation should be carried out according to ENISA. Privacy enhancing tools for individuals, which are not in the focus of e-SIDES, are relevant for hiding personal data from plain view. In this context, ENISA mentions anti-tracking tools, encryption tools, identity masking tools and secure file sharing tools. Transparency mechanisms need to be put in place to ensure that data subjects are adequately informed when their data is processed. Finally, it is considered important to empower data subjects so that they have control over the processing of their data. For this purpose, ENISA suggests mechanisms for expressing consent and privacy preferences, opt-out mechanisms, sticky policies and personal data stores.

For **data analysis and data curation**, *Aggregate* and *Hide* are the most relevant strategies. According to ENISA, anonymisation techniques such as the k-anonymity family or differential privacy can be applied. For hiding personal data from plain view, searchable encryption and privacy-preserving computation are mentioned by ENISA.

Data storage benefits particularly from the strategies *Hide* and *Separate*. Authentication and access control mechanisms as well as encryption of data-at-rest are considered useful to hide personal data. To process data in a distributed fashion, distributed or de-centralised storage and analytics facilities are relevant according to ENISA.

Finally, concerning **data use**, *Aggregate* is the most relevant strategy. Anonymisation techniques are considered relevant by ENISA during data use. Moreover, data quality and data provenance are key aspects.

The strategies *Enforce* and *Demonstrate* are important for all elements of the value chain. According to ENISA, automated policy definition and enforcement are essential in the context of big data applications. Accountability is a key aspect and compliance tools[20] may be useful to achieve it.

The structure and content of chapter 3 is based on previous work on privacy-preserving big data technologies carried out by ENISA to a considerable extent.

[20] Compliance tools help getting and remaining compliant, or verifying compliance with relevant laws, policies and regulations.

3. Technologies

This chapter describes the most relevant classes of technologies in detail. In addition to technologies also approaches and methods are taken into account. The ENISA report “Privacy by design in big data”[21] provides the basis of this chapter. Figure 5 gives an overview of the different classes of technologies identified. The classes are anonymisation, sanitisation, encryption, MPC, access control, policy enforcement, accountability, data provenance, transparency, access and portability, and user control. As some of the classes are closely related, some borders between classes are blurred and some classes even overlap to a certain degree, not every class is discussed in a separate subchapter. Despite the constraints, the classification proved to be immensely useful to discuss the technologies both internally and with people outside the project.

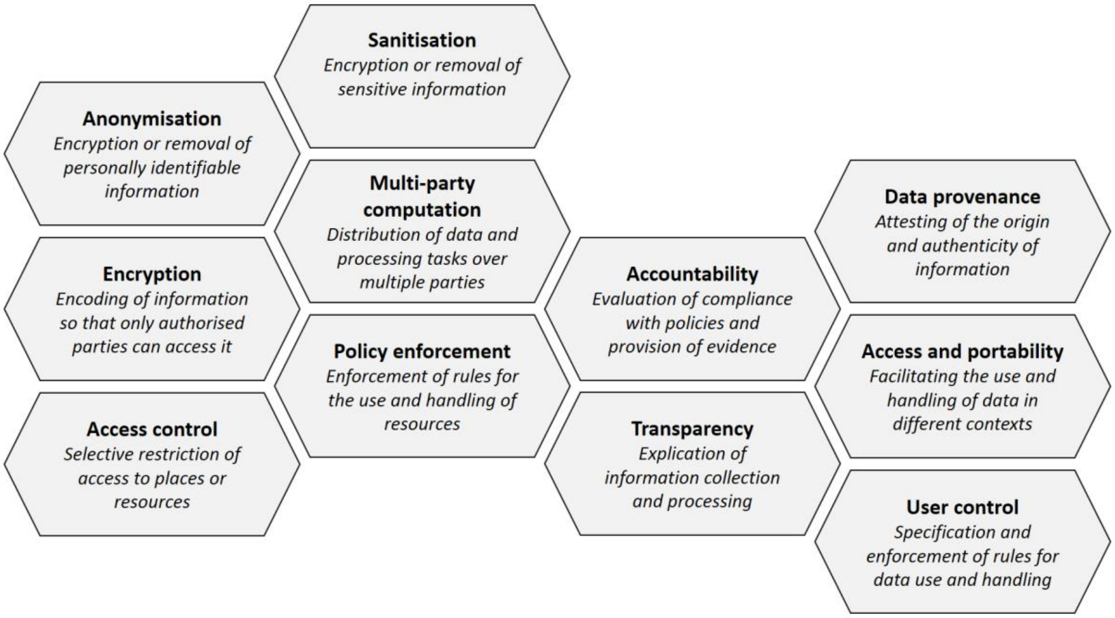

Sanitisation
Encryption or removal of sensitive information
Anonymisation
Encryption or removal of personally identifiable information
Multi-party computation
Distribution of data and processing tasks over multiple parties
Data provenance
Attesting of the origin and authenticity of information
Encryption
Encoding of information so that only authorised parties can access it
Accountability
Evaluation of compliance with policies and provision of evidence
Policy enforcement
Enforcement of rules for the use and handling of resources
Access and portability
Facilitating the use and handling of data in different contexts
Access control
Selective restriction of access to places or resources
Transparency
Explication of information collection and processing
User control
Specification and enforcement of rules for data use and handling

Figure 5 Overview of the classes of technologies described

3.1. Anonymisation

Anonymisation means changing personal data so that factual details can no longer be linked to a specific person, or only with disproportionate effort.[22] From a more formal (or technical perspective) being

[21] Ibid.

[22] Anonymization has thus two distinct subgoals: (1) to avoid the disclosure of the identity of an individual whose data is in a dataset and (2) to avoid the disclosure of (sensitive) attributes for a specific individual; see, e.g., Stan Matwin, “Privacy-Preserving Data Mining Techniques” in *Discrimination and Privacy in the Information Society*

anonymous means a person's indistinguishability from others in a given group.[23] Anonymity depends on the distribution of people with different attributes within the group and grows with the size of the group.

Typical datasets (see Table 1) usually contain different elements or attributes, one of which is usually considered a key attribute that uniquely identifies a person. Apart from this it often contains simple demographic data such as zip code, birth date or gender that are not themselves unique identifiers but, in combination, can be linked with external information to re-identify (some of) the subjects to whom (some of) the records in the original dataset refer. These are called quasi-identifiers. Finally the dataset contains one or more sensitive and non-sensitive attributes.

Key attribute/identifier	***Quasi-identifiers***			***Sensitive attribute***
Name	**DOB**	**Gender**	**Zip Code**	**Disease**
Andre Anderson	17/4/65	Male	79098	Heart Disease
Brian Barlow	31/7/65	Male	79096	Hepatitis
Carl Christians	17/1/65	Male	79098	Bronchitis
Dan Dalton	4/7/83	Male	79331	Broken Arm
Emil Edwards	31/12/81	Male	79336	Flu
Frances Farmer	6/7/83	Female	79338	Diabetes
Grace Gardner	31/10/83	Female	79331	Gastritis

Table 1 Example of a dataset with key attribute/identifier, quasi-identifiers and sensitive attribute

It has been shown that simply removing the identifier (e.g., the name of the person concerned) is not sufficient to anonymise the dataset and to effectively protect privacy. For instance, Date of Birth (DOB), Gender and Zip Code is sufficient to unambiguously identify 87% of the individuals in the United States.[24] Apart from this research suggests that when linked with other data, most anonymised data can be de-anonymised – that is, the identifying information can be reconstructed.

As a consequence complete anonymity has become an obsolete notion, and would in many cases prevent any useful communication or transaction. More realistic to achieve is “factual anonymity”, where finding a link between factual data and a specific person is not a priori impossible, but is “costly” in terms of time and required computing resources. The usual result of the anonymisation procedures described below is “factual anonymity”.

Anonymisation has to be distinguished from pseudonymisation. Pseudonymisation means that one (usually identifying) attribute in a dataset is replaced by another (the pseudonym). While anonymisation aims to irreversibly destroy any way of identifying the data subject, pseudonymisation substitutes the identity of the data subject in such a way that additional information is required to re-identify him or her.

[23] Andreas Pfitzmann and Marit Hansen, “A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management,” (2010), http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf

[24] Latanya Sweeney, “k-Anonymity: A model for protecting privacy,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10, no. 5 (2002)

Although pseudonymisation has many uses, it only provides a limited protection for the identity of data subjects and in many cases as it still allows identification using indirect means.[25]

There are basically three different privacy risks that more sophisticated anonymisation techniques are addressing:

- *singling out* some or all records in a dataset that identify the data subject,
- *linkability* of two or more records of the same data subject (in the same or in different database) which allows assigning an individual to a group without necessarily identifying him/her and
- the possibility to deduce with significant probability the value of an attribute (*interference*).[26]

These three risks occur in any type of data collection and processing, not just big data. For big data, however, there are some specific aspects to be considered:[27]

- Firstly, "linkability" is the core element of many big data analyses. Therefore, it is a challenge to make big data anonymisation compatible with the requirements of the applications (functionality).[28]
- Secondly, anonymisation techniques that are effective for individual sets should also work when data is composed from different sources.
- Thirdly, techniques need to be effective not only for static data (i.e., traditional databases) but also for dynamic data and dynamic data with a high throughput (e.g., streaming data). This is a question of computability (necessary computing power, valid and efficient theoretical frameworks).
- Finally effective approaches need to take into account that in big data there is seldom one central data controller but that data storage and data processing are usually decentralized (i.e., with several data controllers and data processors, possibly even under different legal systems.

Anonymisation techniques generally follow two basic approaches: randomisation and generalisation.[29] In their review article, Nelson and Olovsson mention three common methods based on generalisation (see chapter 3.1.2): (1) k-anonymity, (2) l-diversity, (3) t-closeness (a successor is β-likeness) and one based on randomisation: (4) differential privacy. The first three are mathematical privacy models for releasing entire datasets through PPDP. Differential privacy, which is obtained without modifying the entire dataset, is used for PPDM.[30]

[25] There is the additional issue how a pseudonym can be created, so that it cannot be easily traced back to the identity of the data subject.
[26] Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques,", http://ec.europa.eu /justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (accessed December 14, 2017)
[27] Jordi Soria-Comas and Josep Domingo-Ferrer, "Big Data Privacy: Challenges to Privacy Principles and Models," *Data Science and Engineering* 1, no. 1 (2016)
[28] D'Acquisto et al., "Privacy by design in big data" discuss different approaches to find a balance between privacy and functionality but have to admit that there is no panacea and that a balance of interests must be made in each individual case.
[29] Mehmood et al., "Protection of Big Data Privacy"
[30] Boel Nelson and Tomas Olovsson, "Security and privacy for big data: A systematic literature review" in *2016 IEEE International Conference on Big Data*

3.1.1. Randomisation and differential privacy

The aim of randomisation that changes the data in a dataset is to eliminate strong links between attributes and the identity of the data subject. If the data is sufficiently disturbed (e.g., by adding noise as detailed below), it can no longer be related to a specific person. While randomisation does not reduce the veracity of each record since it still contains the true attributes of the data subject it protects against inference risks.

One possible way to randomise data is to **add "noise"** to the data in order to mask the true attribute values of a dataset.[31] If the added noise is sufficiently large the individual record values cannot be recovered. When the overall distribution is retained, one can assume that the results of processing the dataset will remain accurate. The level and type of noise will depend on the kind of processing that is foreseen. The extent and type of noise depends on the level of accuracy required for each type of processing.[32]

Another randomization technique is **permutation**, which consists of shuffling the values of attributes in a table so that some of them are artificially linked to different data objects. This is useful when it is important to maintain the exact distribution of each attribute within the dataset. Permutation can be understood as a special case of noise addition, in which the attribute values are not changed by a random value, but can only assume the values that are given in the dataset. Permutation ensures that the range and distribution of values remain the same while the links between individuals and values are destroyed.

Techniques such as noise addition and permutation usually cannot guarantee complete anonymisation alone and thus needs to be combined with other techniques such as the removal of obvious attributes or quasi-identifiers.

Differential privacy is a technique based on randomisation and goes beyond the pure anonymisation of datasets. Differential privacy is a privacy model that seeks to limit the impact of any individual subject's contribution to the outcome of the analysis. This means that a person's privacy is protected if it does not make a difference in a query whether or not the data from the respective person is included in the dataset. In this way, only the results of a query are anonymised and not the dataset itself. This is usually achieved by randomly distorting the data after each query by add a little bit of random noise to the true result of the query. Differential privacy allows you to determine how much and what type of noise has to be added to guarantee a certain degree of privacy. The error resulting from the distortion is less than or equal to the error caused by the deletion or addition of a single data record. The result cannot be traced back to a single data record (ε-differential privacy).[33] The data controller, however, remains able to identify individuals in query results, since the original dataset has not been altered.

[31] Kato Mivule, "Utilizing Noise Addition for Data Privacy, an Overview," (2013), https://arxiv.org/ftp/arxiv/papers/1309/1309.3958.pdf (accessed January 15, 2018), for instance, discusses the use of noise addition in more detail.

[32] Charu C. Aggarwal and Philip S. Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms," in *Privacy-Preserving Data Mining: Models and Algorithms*, vol. 34, ed. Charu C. Aggarwal and Philip S. Yu, 11–52, Advances in Database Systems, 34 v. v. 34 (Boston, MA: Springer US, 2008) and Benjamin C. M. Fung et al., "Privacy-preserving data publishing," *ACM Computing Surveys* 42, no. 4 (2010)

[33] Cynthia Dwork and Aaron Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends® in Theoretical Computer Science* 9, 3-4 (2014)

Several proposals have been made how to generate differentially private datasets, basically following two approaches: (1) add noise to mask the value(s) of the original record or (2) create dummy data that meets the requirements of differential privacy. With respect to the popular ε-differential privacy, D'Acquisto et al. mention the variants crowd-blending privacy and BlowFish.[34]

For statistical databases these methods provide a formally verifiable and quantifiable privacy guarantee. Changes of the database are not problematic as long as the distribution of the attributes does not change. However, differential privacy also has some disadvantages: for more complex queries the privacy guarantee is difficult to verify and quantify. Beyond that, the gain in privacy comes at the expense of the accuracy of the results.[35] And finally Haeberlen et al. have shown that at least some implementations of differential privacy methods are vulnerable for covert channel attacks.[36] In general differential privacy may provide good protection against singling out individuals. By using multiple requests, however, it is still possible to link datasets or infer information about individuals. Therefore, search queries must be monitored to identify activities that might compromise anonymity.[37]

3.1.2. Generalisation: k-anonymity, l-diversity and t-closeness

Generalisation is the second class of anonymisation techniques. The basic idea is to generalise (or dilute) the attributes of the data subjects (esp. quasi-identifiers) by changing their scale or magnitude (e.g., decade rather than year; region rather than city). Generalisation can be effective to present singling out, but it does not always allow an effective anonymisation. For this reason, some specific and sophisticated techniques (such as k-anonymity and its successors) have been developed, which can also avoid linkability and inference risks.

Unique identifier	***Quasi-identifiers***			***Sensitive attribute***
Name	**DOB**	**Gender**	**Zip Code**	**Disease**
*	'65	Male	7909*	Heart Disease
*	'65	Male	7909*	Hepatitis
*	'65	Male	7909*	Bronchitis
*	'81-'83	Male	7933*	Broken Arm
*	'81-'83	Male	7933*	Flu
*	'81-'83	Female	7933*	Diabetes
*	'81-'83	Female	7933*	Gastritis

Table 2 Anonymised table with suppressed unique identifier and generalised quasi-identifiers

[34] See the references given by D'Acquisto et al., "Privacy by design in big data", 31

[35] Ibid., 32 and Rathindra Sarathy and Krishnamurty Muralidhar, "Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data," *Transactions on Data Privacy* 4, no. 1 (2011)

[36] Andreas Haeberlen, Benjamin C. Pierce and Arjun Narayan, "Differential privacy under fire," in *Proceedings of the 20th USENIX Conference on Security*, 33 (USENIX Association Berkeley, 2011); for a detailed critique, see, e.g., Jane R. Bambauer, Krish Muralidhar and Rathindra Sarathy, "Fool's Gold: an Illustrated Critique of Differential Privacy," *Vanderbilt Journal of Entertainment & Technology Law* 16, no. 4 (2014)

[37] Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques"

k-Anonymity is a formal data protection model that allows making statements about anonymised datasets. It was developed in 1998 by Sweeney with the aim to publish anonymised sensitive data.[38] It shall ensure that the persons from whom the data was collected cannot be re-identified, while the data itself remains useful for the intended scientific purpose. It is a compromise between a higher level of data protection on the one hand, and a loss of data accuracy on the other.

k-Anonymity relies on preventing two datasets from being joined through generalization of quasi-identifiers[39] or suppression of sensitive attributes in order to avoid re-identification.[40] Typical methods to achieve this is perturbation, e.g. the addition of dummy datasets that contain no useful information ("noise") or the deletion of attributes or datasets that are not necessary for the specific analysis.

A dataset is said to satisfy *k-anonymity* for k>1 if, for each combination of quasi-identifier attribute values, at least k records exist in the dataset sharing that combination.[41] k-Anonymity is able to prevent identity disclosure, that is, a record in the k-anonymised dataset cannot be mapped back to the corresponding record in the original dataset.

k-Anonymity is, however, a relatively simple method to ensure privacy that has specific weaknesses that can be exploited:[42]

1. When every member of a given combination of identity-revealing traits has the same sensitive value, the sensitive value for the set of k records may be exactly predicted (homogeneity attack).
2. Additional background information about the association between one or more quasi-identifier attributes with the sensitive attribute can be used to the set of possible values for the sensitive attribute (background knowledge attack).

To overcome these weaknesses of k-anonymity Machanavajjhala et al. have proposed *l-diversity* as an extension of k-anonymity with reduced granularity of a data representation, i.e. sensitive attributes must be "diverse" within each quasi-identifier equivalence class. The core idea is to model background knowledge as a probability fuction over the attributes in such a way that the difference between prior belief in a sentive attribute is very different the posterior belief after seeing the published generalised data. More formally, a dataset is said to satisfy l-diversity if, for each group of records sharing a combination of quasi-identifier attribute values, there are at least l *well-represented* values for each confidential attribute.[43] There are several definitions of what *well-represented* means.[44] L-diversity gives stronger privacy guarantees than k-anonymity. Nevertheless l-diversity may be difficult and unnecessary to achieve.

[38] Sweeney, "k-Anonymity: A model for protecting privacy"
[39] Generalisation means replacing quasi-identifiers with less specific, but semantically consistent values (e.g., clustering all ages between 20 and 29 into one 2*.
[40] Ibid. and Fung et al., "Privacy-preserving data publishing": 17ff
[41] D'Acquisto et al., "Privacy by design in big data"
[42] Ashwin Machanavajjhala et al., "L -diversity," *ACM Transactions on Knowledge Discovery from Data* 1, no. 1 (2007)
[43] D'Acquisto et al., "Privacy by design in big data"
[44] Machanavajjhala et al., "L -diversity": 17ff.

Since only the unambiguous assignment of a sensitive attributes to a person is reflected in the degree of protection, this concept is insufficient to prevent attribute disclosure since it does not guarantee that an attacker cannot gain *general* information about the sensitive attributes: (1) When the sensitive attribute values in an equivalence class are distinct but semantically similar, an attacker can learn important information (similarity attack) and (2) when the overall distribution is skewed, l-diversity does not prevent attribute disclosure (skewness attack). Furthermore l-diversity does not prevent probabilistic inference attacks.[45]

As a reaction, Li et al. have proposed *t-closeness* as a refinement of l-diversity.[46] The idea behind t-closeness is that the distribution of sensitive attributes within each group of quasi-identifiers should be "close" to their distribution in the entire original database. A dataset is said to satisfy t-closeness if, for each group of records sharing a combination of quasi-identifier attribute values, the distance between the distribution of the confidential attribute in the group and the distribution of the attribute in the whole dataset is no more than a threshold *t*.[47]

Further extensions of this family of methods are, for instance p-sensitive k-anonymity, (n,t)-closeness and β-likeness.[48]

All these techniques can ensure that records about an individual cannot be singled out in a database. l-diversity and t-closeness does no longer allow inference attacks with a 100% confidence. In particular, the possibility of linking data, as with any generalisation approach, cannot be ruled out.[49]

In practice these methods are often the method of choice for data publishing because the approach does not distort the data: even the generalized data is "true", i.e. it represents true (even though possibly imprecise) statements about the original data.

3.1.3. Assessment

One can conclude that none of the both approaches is really superior to the other. First there is a fundamental difference between the two: k-anonymity (and derived techniques) aim to anonymise the dataset before it is release for further analysis while differential privacy aims to anonymise the results of the data analysis and does not change the dataset itself. Regarding the specific requirements of big data they state that k-anonymity offers linkability but is not composable while differential privacy provides

[45] Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," in *Proceedings of the IEEE 23rd International Conference on Data Engineering*, 106–15 (IEEE, 2007)
[46] Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian, "Closeness: A New Privacy Measure for Data Publishing," *IEEE Transactions on Knowledge and Data Engineering* 22, no. 7 (2010)
[47] D'Acquisto et al., "Privacy by design in big data"
[48] Boel Nelson and Tomas Olovsson, "Security and privacy for big data: A systematic literature review" in *2016 IEEE International Conference on Big Data*
[49] D'Acquisto et al., "Privacy by design in big data" and Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques"

strong composability but no linkability. This is the background for attempts to combine k-anonymity with elements of differential privacy (e.g., (k,ε)-anonymity).[50]

For both models the complexity of the computation (and thus necessary computing resources) depend very much on the concrete implementation.

Table 3 gives an overview of the strengths and weaknesses of the techniques considered in this chapter with regard to the three basic privacy risks.

	Is … still a risk?		
	Singling out	**Linkability**	**Inference**
Pseudonymisation	Yes	Yes	Yes
Noise addition, permutation	Yes	May not	May not
Differential privacy	May not	May not	May not
k-anonymity	No	Yes	Yes
l-diversity; t-closeness	No	Yes	May not

Source: Article 29 Data Protection Working Party[51]

Table 3 Strengths and weaknesses of anonymisation techniques

3.2. Encryption

Encryption transforms data in a way that only authorised parties can read it. It is a fundamental security technique and a strong protection measure for personal data. Its role can be integral in big data, as long as it is performed using suitable encryption algorithms and key sizes, and the encryption keys are adequately secured.[52]

3.2.1. Database encryption

Locally encrypted storage on disk or files system level is widely offered in big data and cloud environments. Symmetric key encryption (SKE) schemes based on the Advanced Encryption Standard (AES) are efficient and secure. However, there are concerns related to secure and scalable key management, and to the possibility to perform certain functionalities without disclosing the secret key. Public (or asymmetric) key encryption (PKE) schemes such as RSA[53] allow addressing these concerns but are more demanding in terms of computing resources. Therefore, hybrid schemes are predominant. They combine the advantages of PKE in scalability and key management with the speed and space advantages of symmetric encryption.

[50] Naoise Holohan et al., "(k,ε)-Anonymity: k-Anonymity with ε-Differential Privacy," arXiv:1710.01615 (2017), https://arxiv.org/pdf/1710.01615.pdf

[51] Article 29 Data Protection Working Party, "Opinion 05/2014 on Anonymisation Techniques"

[52] D'Acquisto et al., "Privacy by design in big data"

[53] RSA is one of the first practical public-key cryptosystems. It is widely used for secure data transmission.

However, going beyond the "encrypt all or nothing" model and offering finer grained data sharing policies is necessary in the context of big data.[54] Such policies allow different users to access different parts of the data, share information and perform the analytics needed.

Abbas and Khan[55], Mehmood et al.[56] and D'Acquisto et al.[57] describe such alternative cryptographic primitives that go beyond SKE-based approaches and PKE-based hybrid approaches. Among them are

- attribute-based encryption (ABE)
- identity-based encryption (IBE)/hierarchical IBE (HIBE)
- proxy re-encryption (PRE)
- functional encryption
- predicate encryption/hierarchical predicate encryption (HPE)
- storage-path encryption

The above-mentioned techniques are aimed at allowing more flexibility regarding access and retrieval of encrypted data. Some of the techniques are considered particularly relevant in cloud environments. Abbas and Khan[58], for instance, focus on e-health cloud privacy-preserving approaches.

ABE allows sharing data among different user groups, while preserving users' privacy.[59] ABE combines access control and PKE in a way that the secret key used for the encryption and the ciphertext depend upon certain attributes (e.g., the individual's country, job or habit). In this way, the decryption of the ciphertext can be performed only if the presented set of attributes matches the attributes of the ciphertext. Variants of ABE are ciphertext policy attribute-based encryption (CP-ABE), key policy attribute-based encryption (KP-ABE), multiauthority attribute-based encryption (MA-ABE) and broadcast ciphertext-policy attribute-based encryption (bABE).[60]

IBE is the simplest implementation of ABE. IBE uses any string, for instance, a name or an e-mail address as the public key and the corresponding decryption keys are issued by a trusted party. [61] A variant of IBE called **HIBE** allows multiple private key generators arranged in a hierarchical form to easily handle the task of private key generation.[62]

PRE is a cryptographic primitive that allows a semi-trusted proxy to convert the ciphertext encrypted under the public key of one user into a ciphertext that can be decrypted through another user's private

[54] Ibid.

[55] Assad Abbas and Samee U. Khan, "A Review on the State-of-the-Art Privacy-Preserving Approaches in the e-Health Clouds," *IEEE Journal of Biomedical and Health Informatics* 18, no. 4 (2014)

[56] Mehmood et al., "Protection of Big Data Privacy"

[57] D'Acquisto et al., "Privacy by design in big data"

[58] Abbas and Khan, "A Review on the State-of-the-Art Privacy-Preserving Approaches in the e-Health Clouds"

[59] D'Acquisto et al., "Privacy by design in big data" and Vipul Goyal et al., "Attribute-based encryption for fine-grained access control of encrypted data," in *Proceedings of the 13th ACM Conference on Computer and Communications Security*, 89–98 (2006)

[60] Abbas and Khan, "A Review on the State-of-the-Art Privacy-Preserving Approaches in the e-Health Clouds"

[61] Ibid.

[62] Ibid.

key. Conditional proxy re-encryption (C-PRE) allows enforcing fine-grained access.[63] In the **C-PRE,** as compared to the PRE, the delegator categorizes plaintexts into portions and also the permissions to decrypt each portion are delegated through a proxy under the same pair of keys. The approach assumes the presence of multiple trusted authorities in the PHR system. The trusted authorities ensure the enforcement of the sticky policies besides authorizing the users to get the decryption keys for read and write operations.

Functional encryption is an advancement of ABE where a user with certain attributes has a key that enables him/her to access a particular function of the encrypted data.[64] Functional encryption is particularly interesting in cases where everyone can see a set of ciphertext, but only a specific portion of it can be decrypted and for certain processing that is mandated in the secret key itself. However, due to performance issues, practical implementation is still in its infancy.

Predicate encryption is a PKE-based paradigm used to offer fine-grained access control over encrypted data. [65] In predicate encryption, the secret keys correspond to the predicates and these secret keys are used to decrypt the ciphertext associated with the attributes corresponding to the predicate. HPE is a cryptographic primitive that facilitates the delegation of the search capabilities. However, the delegated users have more restrictive capabilities as compared to the delegating user. The HPE-based schemes can be used to realise searchable encryption.

Storage path encryption means that instead of encrypting the data, only the storage path is encrypted.[66]

3.2.2. Encrypted search

Search is one the most important computer operations and a core area of databases and information retrieval. Encrypted search can be a very powerful tool, especially for big data analytics, allowing full search functionality without the need to disclose any personal data.

With respect to encrypted search, ENISA mentions

- property-preserving encryption (PPE)
- structured encryption

PPE is one of the best techniques when maximising efficiency and query expressiveness. [67] PPE is based on the idea of encrypting data in a way that some property is preserved. The simplest form is deterministic encryption (DTE) that preserves equity. More complex forms include order-preserving encryption (OPE) and orthogonality preserving encryption.

Structured encryption is the basis of Boolean keyword search on encrypted data, which can be a good option when maximizing privacy and efficiency.[68] The limitation approaches based on structured encryption have in common is the lack of query expressiveness, which makes them more suitable for

[63] Ibid. and Mehmood et al., “Protection of Big Data Privacy”
[64] D'Acquisto et al., “Privacy by design in big data”
[65] Abbas and Khan, “A Review on the State-of-the-Art Privacy-Preserving Approaches in the e-Health Clouds”
[66] Mehmood et al., “Protection of Big Data Privacy”
[67] D'Acquisto et al., “Privacy by design in big data”
[68] Ibid.

searches in unstructured data. Symmetric searchable encryption (SSE) encrypts data using a symmetric encryption algorithm and allows for later research matching a given keyword. Public key encryption with keyword search (PEKS) encrypts the database using public key encryption and allows keyword search.

3.2.3. Encrypted computation

When maximizing privacy and query expressiveness the best solutions are fully homomorphic encryption (FHE) and oblivious RAM (ORAM).

FHE enables an unlimited number of computations but with great efficiency loss. For FHE schemes, the slow running speed, which results from computationally intensive calculations, and the large ciphertext are the main challenges. Efficient implementations are therefore as critical as methods to compress the ciphertext.[69]

Homomorphic encryption is a particular type of encryption that permits computations on ciphertexts and also results are obtained in an encrypted form.[70]

Oblivious RAM algorithms allow a client to store large amounts of data (e.g., in the cloud) while hiding the identities of the entities being accessed.

Secure MPC is a field of cryptography aimed at enabling different parties to compute a function over their inputs without disclosing the individual inputs. Due to the particular relevance of the topic for the ICT-18-2016 Research and Innovation Actions (RIAs), MPC is discussed separately in chapter 3.3.

3.3. Multi-party computation

Multi-party computation (MPC), which actually is a field of encryption, plays a key role in two of the ICT-18-2016 RIAs, namely, SPECIAL and SODA. As SODA only recently completed and published a comprehensive deliverable on the state-of-the-art of MPC-based big data analytics[71], there is no need to discuss MPC at length here. In this chapter, the key points of the SODA deliverable are summarised paying particular attention to the essentials as well as concrete examples. Lindell and Pinkas[72], for instance, provide a tutorial-like introduction to MPC-based PPDM with an emphasis on formal security analysis. Moreover, the two authors discuss specialized constructions for some primitives in the two-party setting. Du and Atallah list open issues of MPC in the context of big data[73].

MPC is relevant in the context of PPDM. The common aim of all PPDM approaches is to protect sensitive data used in data mining algorithms. In traditional PPDM, there is one data owner, and another party

[69] Yin Hu, "Improving the Efficiency of Homomorphic Encryption Schemes,", PhD thesis (Worcester Polytechnic Institute, 2013), https://web.wpi.edu/Pubs/ETD/Available/etd-042513-154859/unrestricted/YHu.pdf (accessed January 15, 2018)

[70] Abbas and Khan, "A Review on the State-of-the-Art Privacy-Preserving Approaches in the e-Health Clouds"

[71] "D2.1 State of the Art Analysis of MPC-Based Big Data Analytics," (SODA, 2017), https://www.soda-project.eu/wp-content/uploads/2017/02/SODA-D2.1-WP2-State-of-the-art.pdf

[72] Yehuda Lindell and Benny Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining," *The Journal of Privacy and Confidentiality* 1, no. 1 (2009)

[73] Wenliang Du and Mikhail J. Atallah, "Secure multi-party computation problems and their applications: a review and open problems," in *Proceedings of the 2001 Workshop on New Security Paradigms*, 13–22 (2001)

performs the data mining. In distributed PPDM, there are multiple data owners, whose combined data is mined in a collaborative fashion, either by the owners themselves or by one or more third parties. The MPC techniques developed within the scope of SODA fall into this category. Privacy-preserving querying takes place if the data owner does the mining itself.

Distributed PPDM addresses the problem to perform data mining on sensitive datasets from multiple parties when there is no single party that is trusted to hold all of the data. Based on who performs the mining, three deployment models are described:

- A single miner collecting data from many clients, where each holds an individual record about itself.
- A relatively small number of parties each holding a sub-database, where joint data mining is performed and each party contributes computing power.
- A larger number of parties holding a sub-database but the joint data mining is outsourced to a smaller number of computing parties that do not necessarily provide input themselves.

In each case, the data can be horizontally partitioned, meaning that each data provider holds all attribute values for one or more records or vertically partitioned, meaning that there is a common set of records for which each data provider holds a number of attribute values.

In theory, every algorithm can be performed with MPC to fully protect the privacy of sensitive inputs. However, because of the large performance penalty this induces, in practice, often a hybrid approach is used in which the data providers combine MPC with local computation. Such approaches range from ones where the local computation is complex and the MPC aggregation easy to ones where the local computation is easy and for the bulk of the work MPC is used.

Due to efficiency reasons, most approaches are from the end of the spectrum where local computation is used for most of the work. An example for such approaches is the protocol for deep learning by gradient descent recently proposed by Bonawitz et al.[74] An example from the other end of the spectrum is the linear programming solver proposed by de Hoogh[75].

Several common patterns of distributed PPDM can be identified in the literature. Instead of using MPC to aggregate local computation results, it is also possible to add noise to or mask the input data or local computation results before the aggregation is performed. Differential privacy can be used to avoid the disclosure of sensitive information from opening intermediate computation results. Many variations on this idea exist. For instance, it is possible to securely perform aggregation and add noise prior to opening the aggregate.

The SODA project recently surveyed the state-of-the-art in the MPC approach to PPDM. Articles were classified by means of a manual clustering according to the main data-mining-related topics addressed. SODA's categorization distinguishes between supervised methods (e.g., classification, regression), unsupervised methods (e.g., clustering, pattern mining), search, querying and matching problems (e.g.,

[74] Keith Bonawitz et al., “Practical Secure Aggregation for Privacy-preserving Machine Learning,”, Cryptology ePrint Archive 2017/281 (Google, 2017), https://eprint.iacr.org/2017/281

[75] Sebastiaan de Hoogh, “Design of large scale applications of secure multiparty computation: secure linear programming,”, PhD thesis (Eindhoven University of Technology, 2012)

biometric matching, genomic sequence matching), recommendation techniques (e.g., matrix factorisation) and auxiliary methods from statistics (e.g., computing the mean, median, statistical tests). The authors did their best to cover most of the applied research directions in the field of MPC-based PPDM by means of discussing exemplary papers.

3.4. Access control and policy enforcement

ENISA states that access control is essential for protecting personal data in databases, and one of the most fundamental security measures that is applicable to any application, ensuring that only authorised processes and procedures can gain access to data.[76] Access control is also the most common approach to ensure policy enforcement: by only permitting access to users who are policy compliant, policy can be enforced by default.

Access controls usually consist of two steps: identification and authentication. Identification is usually based on identity information. Authentication methods typically rely on passwords, access tokens, biometrics or combinations of the three. Methods involving two independent ways for verifying identity are referred to as two-factor authentication methods.

One of the goals of the emerging identity management systems is, according to Wang and Kobsa[77], to allow users to have more than one digital identity and be able to freely choose which identity to use. In a multi-level security system, the accessibility of information depends on the authorisation level of a user. In closed systems only explicitly authorised accesses are allowed, whereas in open systems accesses that are not explicitly forbidden are allowed.[78]

As Moreno et al.[79] note, big data environments traditionally do not prioritise security, and often only provide basic forms of access control. For example, Colombo and Ferrari[80] examine eight of the main big data platforms like Hadoop and MongoDB, and find that the level of access control offered varies considerably, with only Apache Hive providing fine-grained access control. Moreover, none support privacy policies. Colombo and Ferrari argue that the absence, to date, of a standard query language and data model for big data platforms makes the development of a general privacy-aware access control enforcement solution difficult. Therefore, they focus their efforts specifically on MongoDB, one of the most popular NoSQL data stores, for which they design a framework for integrating privacy-aware access control functionalities.[81]

[76] D'Acquisto et al., "Privacy by design in big data"

[77] Yang Wang and Alfred Kobsa, "Privacy-Enhancing Technologies," in *Handbook of research on social and organizational liabilities in information security*, ed. Manish Gupta and Raj Sharman (Hershey, PA: Information Science Reference, 2009)

[78] Silvana Castano et al., *Database security* (ACM Press, 1995)

[79] Julio Moreno, Manuel Serrano and Eduardo Fernández-Medina, "Main Issues in Big Data Security," *Future Internet* 8, no. 3 (2016)

[80] Pietro Colombo and Elena Ferrari, "Privacy Aware Access Control for Big Data: A Research Roadmap," *Big Data Research* 2, no. 4 (2015)

[81] Ibid. and Pietro Colombo and Elena Ferrari, "Complementing MongoDB with Advanced Access Control Features: Concepts and Research Challenges," in *Proceedings of the 23rd Italian Symposium on Advanced Database Systems*, ed. Domenico Lembo, Andrea Marrella and Riccardo Torlone, 343–50 (Curran Associates, 2015)

While ENISA argues that traditional approaches to access control such as **role-based access control** (RBAC) or user-based access control are quickly becoming unmanageable in the context of big data, these approaches still remain in use. For instance, both the MongoDB and Apache Cassandra database systems utilise RBAC.[82] RBAC determines access to resources based on the user's role within the organization. Each role carries a specific set of privileges associated with it. User-based access control is often implemented through access control lists (ACLs). ACLs are tables that list each user and determine its access to a resource.

A further set of approaches such is **attribute-based access control** (ABAC) that can conceptually support fine grained access control policies in big data based on attributes that are evaluated at run-time. ABAC determines access to resources based on policies taking attributes of the user, the resource and the environment state into account. ABAC is sometimes referred to as policy-based access control (PBAC).

Extensible access control markup language[83] (XACML) is a well acknowledged open standard that supports the creation of access control policies and comparative evaluation of access requests according to predefined policy rules. XACML defines an architecture, a policy language, and a request/response scheme. It does not handle attribute management, which is left to traditional identity and access management tools, databases and directories. As a published standard specification, one of the goals of XACML is to promote common terminology and interoperability between access control implementations by multiple vendors.

Further policy languages are Common Policy[84] and the Platform for Privacy Preferences Project (P3P)[85]. Common Policy defines a framework for authorisation policies controlling access to application-specific data. The Common Policy framework is enhanced by domain-specific policy documents. In contract, P3P enables websites to express their privacy practices in a standard format that can be retrieved automatically and interpreted easily by users. Websites implementing such policies make their practises explicit and thus open them to public scrutiny.

Samarati and De Capitani di Vimercati[86] review research directions pursued in the context of data protection in data outsourcing scenarios. The authors state that the use of selective encryption has been proposed already a while ago. Different portions of the data can be encrypted with different keys that are then distributed to users according to their access privileges. According to Samarati and De Capitani di Vimercat, the problems related to the definition, management and evolution of the authorisation policies and therefore of the corresponding encryption have not been solved satisfactorily. Promising proposals integrate access control and encryption. The data to be outsourced is encrypted with different keys depending on the authorisations to be enforced on the data. Such a policy is then translated into an equivalent encryption policy regulating which data are encrypted with which key and regulating key

[82] Colombo and Ferrari, "Privacy Aware Access Control for Big Data"

[83] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

[84] http://www.faqs.org/rfcs/rfc4745.html

[85] http://www.w3.org/P3P/

[86] Pierangela Samarati and Sabrina D. C. Di Vimercati, "Data protection in outsourcing scenarios: Issues and directions," in *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security* (2010)

release to users. Another approach along these lines is that adopted by Li et al.[87] Data are encrypted, and sticky policies are used to regulate access to the keys: they are only issued to users who satisfy the policy constraints.

One important challenge in access control for big data is implementing sufficiently **fine-grained access controls**, so as to enable differential access depending on the sensitivity of the data and the authorisation level of the users. According to Ulusoy et al.[88] especially access-control approaches for MapReduce -- one of the most widely used big data programming models for processing and storing large datasets -- mostly have an all-or-nothing quality; either permitting access to the entire dataset, or no access at all. To overcome this problem, they propose a modular framework named **GuardMR** that enforces fine-grained security policies at the key-value level (instead of the higher-level file level) of MapReduce. Based on the organisational role (security clearance) of the users, GuardMR permits different levels of access to the data.

A related challenge, especially in systems with many different users and data subjects facing different contexts, is to decentralise decisions over access and privacy policy setting and enforcement sufficiently, while still guaranteeing security. Writing about big-data usage in health care and social networks, Samuel et al.[89] argue that when the number of users is large, it is often more computationally complex to compose a single privacy policy and access-control regime for all users than to permit users to make decentralized, contextually-aware decisions about privacy setting and access themselves. Therefore, they design a so-called "Secure User Data Repository System" as a hybrid approach, wherein traditional RBAC are supplemented by a component allowing users to decide the circumstances under which data is disclosed and access granted. A prototype of their system, called Intelligent Privacy Manager (iPM) has been implemented and can be accessed online[90].

Another approach to enhancing user controls over access to their data while maintaining security is that by Zhou et al.[91], who propose a random-walk-based privacy-preserving access control for online social networks. The approach employs random walking to form profiles for users. Based on the profiles, users can carry out access control according to the secure computation of closeness. Furthermore, users can set the permissible threshold independently according to their access policy. In this way, the leakage of privacy as it exists in traditional ABAC can be removed. Zhou et al. state that experimental results show that the proposed scheme is reasonable and practical. An issues that the authors intend to address in future work, is the limited efficiency of the computation of closeness.

Big data contexts often involve large-scale, distributed networks of systems ("systems of systems") with many users and owners who belong to different organisations. An example are multinational

[87] Shuyu Li et al., "A Sticky Policy Framework for Big Data Security," in *Proceedings of the IEEE 1st International Conference on Big Data Computing Service and Applications*, 130–7 (IEEE, 2015), http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7184873 (accessed December 14, 2017)

[88] Huseyin Ulusoy et al., "GuardMR: Fine-grained Security Policy Enforcement for MapReduce Systems," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security - ASIA CCS '15*, ed. Feng Bao et al., 285–96 (New York, New York, USA: ACM Press, 2015)

[89] Arjmand Samuel et al., "A Framework for Composition and Enforcement of Privacy-Aware and Context-Driven Authorization Mechanism for Multimedia Big Data," *IEEE Transactions on Multimedia* 17, no. 9 (2015)

[90] https://engineering.purdue.edu/dmultlab/

[91] You-sheng Zhou, En-wei Peng and Cheng-qing Guo, "A Random-Walk Based Privacy-Preserving Access Control for Online Social Networks," *International Journal of Advanced Computer Science and Applications* 7, no. 2 (2016)

environmental sensor networks like GEOSS, in which 61 nations are involved.[92] A key challenge in systems like these where there may be no trust between owners and users is to simultaneously achieve access control, privacy-preservation (users may want to hide their identity and transactions on the data), and accountability (users who misbehave should nevertheless be identified). To solve this problem, He et al.[93] have designed a **privacy-preserving and accountable access control protocol** named APAC, which they apply to wireless sensor networks. APAC exploits an adapted ring-signature technique and the separation of duties principle to not only offer strict access control and privacy protection but also allows auditing and pinpointing misbehaving users and owners. Moreover, to further increase security and privacy, APAC does not rely on a trusted third party auditor (TPA). Instead, the supervisory functions conventionally allotted to a TPA are divided between group managers and a "law-enforcement authority". Unique users can only be identified when the group managers and the law-enforcement authority cooperate.

One limitation of access control as a strategy for policy enforcement is its "once-and-for-all"-character and application specificity. This can become a problem in particular in the Internet of Things (IoT) context. As Pasquier et al.[94] note, decisions over whether users get access to some data are, by definition, taken before they gain access, but *thereafter* access-control systems foresee no further action (monitoring, enforcement, etc.). Moreover, traditionally, access-control policies tend to be application-specific. This is particularly problematic in large-scale distributed systems like IoT, where many heterogeneous users and applications interact. Under these circumstances, application-specific (rather than system-wide) access policies are liable to lead to policy conflicts. As an enhancement of access control for policy enforcement (and audit) in the IoT they therefore propose **information flow control**, to enable continuous, data-centric, cross-application and end-to-end control of data flows. Concretely, two-component tags (a form of metadata) are attached to each data item, and each application. Data are then only allowed flow to apps with matching labels.

3.5. Accountability and audit mechanisms

While accountability is a key concept in data protection law whose importance will be further reinforced when the EU GDPR comes into force in May 2018, the prospect of a fully accountable information processing system remains, in the words of Druschel et al.[95], a "longterm vision". As ENISA[96] note, a cornerstone of accountable information systems is the provision of automated and scalable control and auditing processes that can evaluate the level of compliance with privacy policies against predefined machine-readable rules. Due to the rapid spread of cloud-based information processing systems, data integrity verification – proving that the data is intact, uncorrupted and untampered with – is emerging as a further important area of research.

[92] https://www.earthobservations.org/geoss.php

[93] Daojing He, Sammy Chan and Mohsen Guizani, "Accountable and Privacy-Enhanced Access Control in Wireless Sensor Networks," *IEEE Transactions on Wireless Communications* 14, no. 1 (2015)

[94] Thomas F.J.M. Pasquier et al., "Managing Big Data with Information Flow Control," in *Proceedings of the IEEE 8th International Conference on Cloud Computing*, 524–31 (IEEE, 2015), https://www.cl.cam.ac.uk/research/srg/opera/publications/papers/2015ieeeCloud.pdf (accessed December 14, 2017)

[95] Peter Druschel et al., "Towards Accountable Information Systems," (2014), http://dig.csail.mit.edu/2014/AccountableSystems2014/abs/druschel-ext-abs.pdf (accessed December 14, 2013)

[96] D'Acquisto et al., "Privacy by design in big data"

According to both Sen et al.[97] and Upadhyaya et al.[98], even in many information and communication technology (ICT) companies efforts to comply with privacy and other policies have so far remained heavily based on manual review and employee training, which is time consuming, sometimes unreliable, and does not scale well. As policy compliance in many organisations is however handled by lawyers and other staff with little technical training, a key challenge when developing automated compliance tools is to design them thus that they can be used by people with limited training in software coding. To meet these challenges, Sen et al. [99] have developed a **workflow for privacy compliance in MapReduce-like big data systems**. The workflow consists of two components; a formal language called **LEGALESE**, and **GROK**, a data-inventory mapper for MapReduce-like systems. The prototype system has been applied to the backend data-analytics pipeline of the search engine Bing. Bootstrapped by a small team, it scales to the needs of several thousand developers and checks compliance daily of millions of lines of source code.

While the LEGALESE/GROK workflow only checks compliance, but does not enforce it, the system developed by Upadhyaya et al.[100], called **DataLawyer**, also provides continuous policy enforcement. It is a middleware layer to be installed on top of relational database management systems. It allows users to run normal SQL queries, but checks each query against specified policies before letting it execute. Non-compliant queries are rejected and the user informed about the violation. However, like LEGALESE/GROK, DataLawyer does not protect against malicious users. While the LEGALESE/GROK workflow was developed primarily for in-house use, to help organisations facilitate compliance by their own employees, DataLawyer is targeted at both the suppliers of datasets and data bases, and their clients.

According to Pasquier et al.[101], the IoT poses particular challenges for auditing compliance with privacy policies. This is partly because of its vast scale, but especially because given the IoT's nature as a network of interconnected "things", traditional application (thing)-centric approaches to auditing are liable to constitute an obstacle to understanding system-wide behaviour. They therefore propose an **information-centric audit mechanism built around provenance data** represented as a directed acyclic graph (DAG). Pasquier and Eyers[102] apply this use of provenance data to the case of a cloud-connected smart home system, to demonstrate compliance with the privacy-policy recommendations of the French Data Protection Authority.

The spread of cloud computing has created further requirements for audit, namely **verifying the integrity of data stored in the cloud**. This poses two closely linked challenges: on the one hand, designing

[97] Shayak Sen et al., “Bootstrapping Privacy Compliance in Big Data Systems,”, https://www.andrew.cmu.edu/user/danupam/sen-guha-datta-oakland14.pdf (accessed December 14, 2017)

[98] Prasang Upadhyaya, Magdalena Balazinska and Dan Suciu, “Automatic Enforcement of Data Use Policies with DataLawyer,” in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 213–25 (ACM, 2015), http://cloud-data-pricing.cs.washington.edu/datalawyer.pdf (accessed December 14, 2017)

[99] Shayak Sen et al., “Bootstrapping Privacy Compliance in Big Data Systems”

[100] Prasang Upadhyaya, Magdalena Balazinska and Dan Suciu, “Automatic Enforcement of Data Use Policies with DataLawyer” in *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*

[101] Thomas Pasquier et al., “Data provenance to audit compliance with privacy policy in the Internet of Things,” *Personal and Ubiquitous Computing* 8, no. 12 (2017), https://link.springer.com/content/pdf/10.1007/s00779-017-1067-4.pdf (accessed December 14, 2017)

[102] Thomas F. J.-M. Pasquier and David Eyers, “Information Flow Audit for Transparency and Compliance in the Handling of Personal Data,” in *Proceedings of the 2016 IEEE International Conference on Cloud Engineering: Workshops*, 112–7 (IEEE, 2016)

mechanisms to audit cloud-stored data efficiently (that is, with minimal computational overhead); on the other hand, designing audit processes so as to protect the privacy of the cloud users (that is, prevent unintended data disclosure occurring due the audit). Traditional approaches to cloud auditing involve retrieving the entire stored data from the cloud and then verifying this, e.g., by checking signatures or hash values. However, this is computationally costly, especially in the context of big data.[103] Therefore, researchers have proposed outsourcing the audit process to a third party, and sought to develop ways of verifying the data's integrity without retrieving all, or even any of it. Avoidance of data retrieval is motivated by security/data protection considerations as much as computational efficiency: if retrieval is performed by a third party, privacy is threatened.

According to Liu et al.[104] PDP and POR are the two main approaches to cloud data-integrity verification without retrieval. The basic idea of both is to split the data file stored in the cloud into blocks, and then attach some metadata to the data, in form of either a homomorphic linear authenticator or a homomorphic verifiable tag. This metadata is then used to verify the underlying data blocks. Of the two schemes, PDP is said to be the safer and more efficient.[105] Subsequent work by Erway et al.[106] and Liu et al.[107], among others, has sought to develop PDP further to facilitate the verification of dynamic data (such as is the norm in big data contexts).

Very extensive further work has also been done on the problem of ensuring privacy preservation when auditing is performed by a third party. For example, Wang et al.[108] exploit ring signatures to protect the privacy of users with shared data. While conventional PDP/POR approaches to auditing shared data could still allow a third-party auditor to discover significant confidential information (e.g., which user or which block is the more valuable target), their scheme preserves identity privacy. A different approach is taken by Shen et al.[109], who seek to dispense with the conventional third-party auditor entirely, replacing this with a “third party medium” and blinding data during the upload and data-auditing phases.

3.6. Data provenance

Provenance provides a **record of the sources and transformations** (processing history) of a piece of data. It aims to answer the questions: Where do data come from? Who manipulated the data? What transformations were applied?[110] Provenance is commonly represented in the form of a DAG showing the interactions between the data items, processes and agents in question. Provenance data is critical for

[103] Boyang Wang, Baochun Li and Hui Li, “Oruta: Privacy-preserving public auditing for shared data in the cloud,” *IEEE Transactions on Cloud Computing* 2, no. 1 (2014)

[104] Chang Liu et al., “External integrity verification for outsourced big data in cloud and IoT: A big picture,” *Future Generation Computer Systems* 49 (2015)

[105] Ibid. and Mehmood et al., “Protection of Big Data Privacy”

[106] Chris Erway et al., “Dynamic provable data possession,” in *Proceedings of the 16th ACM Conference on Computer and Communications Security*, 213–22 (ACM, 2009)

[107] Chang Liu et al., “Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-Grained Updates,” *IEEE Transactions on Parallel and Distributed Systems* 25, no. 9 (2014)

[108] Wang, Li and Li, “Oruta”

[109] Wenting Shen et al., “Light-weight and privacy-preserving secure cloud auditing scheme for group users via the third party medium,” *Journal of Network and Computer Applications* 82 (2017)

[110] Pasquier et al., “Data provenance to audit compliance with privacy policy in the Internet of Things” and D'Acquisto et al., “Privacy by design in big data”

numerous capabilities. It can enable the re-use and reproduction of experiments, debugging, process optimisation and performance prediction.[111] By providing transparency, attesting data origin and authenticity, it facilitates judgements about the trustworthiness of data and thus helps improve decision-making.[112] Scholars have also proposed using provenance for audit and accountability[113], and as a tool for access control.[114]

However, in particular – albeit not only – in a big data context, the collection and utilisation of provenance information also faces a number of challenges requiring further research and development work. These challenges concern both the question of security and privacy-protection, and the very collection and processing of provenance information given the particular characteristics of big data (volume, velocity, etc.)

Davidson et al.[115], among others, have argued that there is an inherent **trade-off between the utility provided by provenance information, and the level of privacy/security guarantees** maintained. More utility tends to equal lower privacy/security guarantees, and vice-versa. D'Acquisto et al.[116] emphasise that unlike in statistical analysis, with provenance no aggregation takes place and the information is shown exactly as it is, potentially allowing the identification of individuals. If the provenance graph contains sensitive information – which it may – this is a serious problem.

A number of approaches have been proposed to try to solve this problem. According to Pasquier et al.[117], the most common solution suggested in the literature is to **abstract the provenance graph** so that sensitive information is hidden but the semantic information necessary for provenance analysis nevertheless conserved. In a wide-ranging review article, Bertino et al.[118] also discuss this approach, under the rubric of sanitising DAGs. In their view, sanitisation confronts two fundamental challenges. On the one hand, there is the risk of "sanitising away" so much information that the utility of the remaining (sanitised) DAG is impaired. To address this challenge, Bertino et al. call for research to define policy languages that better specify what provenance information, exactly, is needed for the policy in question (e.g., accountability) to be performed. Policy-aware sanitisation algorithms that would search the possible sanitisation domain to find sanitisations that would both satisfy the policy (e.g., accountability) and achieve the desired privacy/security level are a further research area they suggest.

[111] Jianwu Wang et al., “Big data provenance: Challenges, state of the art and opportunities,” in *Proceedings of the 2015 IEEE International Conference on Big Data*, 2509–16 (IEEE, 2015), http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7364047 (accessed December 14, 2017)

[112] Elisa Bertino et al., “A roadmap for privacy-enhanced secure data provenance,” *Journal of Intelligent Information Systems* 43, no. 3 (2014) and D'Acquisto et al., “Privacy by design in big data”

[113] Pasquier et al., “Data provenance to audit compliance with privacy policy in the Internet of Things”

[114] Jaehong Park, Dang Nguyen and Ravi Sandhu, “A provenance-based access control model,” in *Proceedings of the 10th Annual International Conference on Privacy, Security and Trust*, 137–44 (IEEE, 2012)

[115] Susan B. Davidson et al., “On provenance and privacy,” in *Proceedings of the 14th International Conference on Database Theory*, ed. Tova Milo, 3–10 (New York, New York, USA: ACM Press, 2011)

[116] D'Acquisto et al., “Privacy by design in big data”

[117] Pasquier et al., “Data provenance to audit compliance with privacy policy in the Internet of Things”

[118] Bertino et al., “A roadmap for privacy-enhanced secure data provenance”

The other challenge for sanitisation that Bertino et al. see is the potential for inference attacks. If background information from other sources is available to an attacker, sanitisation may be insufficient, as inferring sensitive information could be possible by combining *sanitised* provenance with other background information. They recommend exploring graph mining techniques as a way of helping researchers better understand which parts of a DAG to sanitise and whether sanitisation has been sufficient.

Access controls are a further approach for achieving privacy and security goals in provenance, which is suggested for instance by Davidson et al.[119] In general, implementing fine-grained access controls requires clear definitions of which parts of the provenance graph need to be protected from whom. According to Bertino et al.[120] this demands defining new access control languages and mechanisms specific to provenance and DAGs. A last approach, that seems to have been little-explored to date, is **extending cryptographic techniques** to support queries on encrypted provenance. [121]

While research efforts to find technical solutions to minimise the trade-off between provenance utility and the maintenance of privacy and security guarantees is thus ongoing, it seems likely that this trade-off will remain to some extent inherent. Therefore, Bertino et al. argue that the potential of **risk-management approaches** should be explored. Concretely, this would require developing methods to estimate both the potential utility and the potential risks created by the release of provenance data.

Privacy and security are not, though, the only challenge posed by provenance in a big data context. Another significant issue is **scalability** and the **size of the provenance data** collected. Wang et al.[122] note that obtaining a fine-grained provenance tracking of a big data workflow can easily lead to provenance data that is several times larger than the original dataset. Pasquier et al.[123] canvass several approaches to dealing with these challenges, including **stream processing** (wherein queries are applied to provenance data as they are generated), **storing and processing at rest**, **reducing the amount of provenance collected** in the first place ("take only what you need") – an approach that may also help achieve privacy/security objectives – and **graph compression**.

Dealing with the problem of scalability may also require the development of new, **decentralised provenance systems**. This is the claim of Tas et al.[124], who evaluate the performance and scalability of several state-of-the-art centralised provenance systems (specifically, PReServ, Karma and Komadu). They find that while these systems work well even for relatively large provenance graphs with up to 4000 "social operations", once the graphs exceed this size performance and scalability deteriorates.

[119] Susan B. Davidson et al., "On provenance and privacy" in *Proceedings of the 14th International Conference on Database Theory*

[120] Bertino et al., "A roadmap for privacy-enhanced secure data provenance"

[121] Ibid.

[122] Jianwu Wang et al., "Big data provenance" in *Proceedings of the 2015 IEEE International Conference on Big Data*

[123] Pasquier et al., "Data provenance to audit compliance with privacy policy in the Internet of Things"

[124] Yucel Tas, Mohamed J. Baeth and Mehmet S. Aktas, "An Approach to Standalone Provenance Systems for Big Social Provenance Data," in *Proceedings of the 12th International Conference on Semantics, Knowledge and Grids*, 9–16 (IEEE, 2016)

A somewhat related problem is addressed by Korolev and Joshi.[125] While big data MapReduce workflows commonly run on distributed clusters of the Hadoop type, most of the widely used scientific workflow systems used to represent and execute computation experiments (e.g., VisTrails, NyPipe and Taverna) are designed to be used on a single (rather than a distributed) workstation, and thus struggle to track provenance for big data MapReduce workflows. Moreover, these systems are not always designed to permit controlled sharing of workflows on sensitive and restricted data. Korolev and Joshi have therefore proposed the **tool PROB to track provenance in big data** experiments. PROB utilises Git-Annex to store only the hash values of datasets, not the datasets themselves, thereby enabling sharing of workflows without having to share (possibly restricted) datasets.

More broadly, Wang et al.[126] argue that **existing provenance models** and standards like OPM[127] need to be **extended for big data**, to include additional information like data quality, data compression, and execution environment information. Thus while current approaches to provenance focus mainly on recording information about "internal state changes" of the data, for big data "external provenance" information should also be recorded, in particular the parameter configurations of the big data engines used (e.g., Hadoop) on a given execution. Hadoop for instance has more than 200 parameters.

3.7. Transparency

Proper information and transparency is a key issue in any data processing, so as to allow individuals to understand how their data are being processed and to make relevant informed choices.[128]

Transparency in big data may be achieved by

- purely textual information
- multichannel and layered approaches
- standardized icons and pictograms

Purely textual information does not seem to cope with the evolution of services and to comprehensively inform users on the processing of data occurring in the complex big data value chain.

To improve the effectiveness of information, multichannel and layered approaches have been suggested.

Standardised icons and pictograms (e.g., Disconnect Privacy Icons, Mozilla privacy icons) is another promising emerging approach for transparency in big data. However, measures need to be taken that people are able to understand the graphic scheme as well as the pictographic parts of icons.[129]

[125] Vlad Korolev and Anupam Joshi, "PROB: A tool for tracking provenance and reproducibility of big data experiments," in *Proceedings of the 20th IEEE International Symposium on High Performance Computer Architecture* (Orlando, Florida, USA: IEEE, 2014), http://ebiquity.umbc.edu/_file_directory_/papers/693.pdf (accessed December 14, 2017)

[126] Jianwu Wang et al., "Big data provenance" in *Proceedings of the 2015 IEEE International Conference on Big Data*

[127] Open Provenance Model, http://openprovenance.org/

[128] D'Acquisto et al., "Privacy by design in big data"

[129] John S. Pettersson, "A brief evaluation of icons suggested for use in standardised information policies: Referring to the Annex in the first reading of the European Parliament on COM (2012) 0011,", Working paper (2014), https://www.diva-portal.org/smash/get/diva2:720798/FULLTEXT01.pdf (accessed January 15, 2018)

3.8. Access and data portability

Providing access to users on their data is an important privacy condition as well as an obligation of data controllers.

Data portability is an additional and important tool for users, giving the possibility to change service providers without losing their data.

There are already some interesting initiatives in this respect. Relevant projects are the Midata UK initiative[130], providing access in transactions and consumption for the energy, finance, telecommunications and retail sectors, and the French MesInfos platform[131] for access to financial, communication, health, insurance and energy data.

3.9. User control

User control is critical in the context of big data applications. Users must be able to determine the compliance of controllers and processors with certain rules. User control can be reached by means of different approaches. The use of consent mechanisms is probably the most prominent one.

3.9.1. Consent mechanisms

Consent is a fundamental element of data protection. However, traditional consent mechanisms are in conflict with the use of data for multiple purposes, where some might not be known at the time of data and consent collection. New usable and practical mechanisms to collect consent are needed that do not constitute a barrier for the usability of services.

Engineered banner solutions, for instance, provide a basis for user-friendly consent mechanisms. They are increasingly used to give users of websites and apps information about the use of cookies and other forms of local storage. In the big data context, a higher degree of automation in the collection and withdrawal of consent is considered important.[132]

Software agents providing consent on the behalf of the user based on the properties of certain applications could be a topic to explore in the future. Moreover, positive user actions such as specific gestures or motions could be used to constitute consent.

3.9.2. Privacy preferences and sticky policies

Sticky policies can provide a mechanism for attaching privacy preferences to specific datasets and accordingly drive data processing decisions. Privacy policies and requirements cannot only be expressed by data subjects but also by other parties. For example, acceptable purposes, allowed recipients or deletion periods can be indicated. If potential data recipients formalise their commitments concerning purposes and conditions of data processing in privacy policy documents, formal statements of both parties can be compared prior to the collection, sharing or use of personal data. Several techniques have been proposed that support automated negotiation procedures between the data subject and third parties.[133]

[130] https://www.gov.uk/government/news/the-midata-vision-of-consumer-empowerment

[131] http://mesinfos.fing.org/

[132] D'Acquisto et al., "Privacy by design in big data"

[133] Nicholas R. Jennings et al., "Automated Negotiation: Prospects, Methods and Challenges," *International Journal of Group Decision and Negotiation* 10, no. 2 (2001)

Through different negotiation steps, a commonly acceptable solution is reached. The result, a fine-grained privacy contract, governs the use of personal data.

The enforcement of privacy preferences and sticky policies is very important for their practical adoption and implementation. In principle, enforcement is weak, as the individual would need to trust that all parties involved in the processing of personal data would respect his/her preferences.

3.9.3. Personal data stores

A range of technical solutions have been proposed to give data subjects' increased control over their data through transition from distributed data models to user-centric models. Such solutions (aka personal data vaults, personal data lockers, personal clouds, personal data stores) enable individuals to gather, store, update, correct, analyse and/or share personal data.

4. Related projects

This chapter describes related research projects and outlines links between the classes of technologies introduced and the projects. The focus lies on research projects funded by the EU. The chapter provide a preliminary overview only but already allow e-SIDES to adapt its activities to the projects' requirements. The e-SDIES project aims to deepen the cooperation with related project in the future. Moreover, measures are taken to extend the focus to industrial companies.

4.1. ICT-18-2016 RIAs

The focus of this chapter is on the RIAs SPECIAL, SODA and MDMH. Just like the Coordination and Support Action (CSA) e-SIDES, these projects are funded under ICT-18-2016 (Big data PPP: Privacy-preserving big data technologies). The three RIAs are tasked to advance technologies for data access, processing and analysis to better protect personal data in line with existing and future EU rules. With their explicit focus on the further development of privacy-preserving big data technologies, the projects are of particular relevance for e-SIDES.

4.1.1. SPECIAL

The **SPECIAL**[134] (Scalable Policy-aware Linked Data Architecture for Privacy, Transparency and Compliance) project aims to address the contradiction between big data innovation and data protection by proposing a technical solution that makes both of these goals more realistic. SPECIAL develops technology that

- supports the acquisition of user consent at collection time and the recording of both data and metadata according to policies prescribed by the law or the user;
- caters for privacy-aware, secure workflows that include usage/access control, transparency and compliance verification;
- demonstrates robustness in terms of performance, scalability and security; and
- provides a dashboard with feedback and control features that make privacy in big data comprehensible and manageable for data subjects, controllers and processors.

Big data architectures are combined and extended to handle linked data, harness them with sticky policies as well as scalable queryable encryption, and develop advanced user interaction and control features. The results are validated via real world use cases. SPECIAL exploits results of the EU projects Big Data Europe and PrimeLife.

SPECIAL has already completed several deliverables. D1.2 focuses on legal requirements for privacy-enhancing big data and thus is particularly relevant for e-SIDES. The deliverables identifies and analyses the legal frame conditions in the context of big data across the EU. Another particularly relevant deliverable is D3.1. This deliverable describes the initial setup of the policy-aware linked data architecture and engine (also referred to as SPECIAL platform) developed within the scope of the project.

[134] https://www.specialprivacy.eu/

With respect to the groups of technologies introduced in chapter 3, representatives of the SPECIAL project provided e-SIDES with comprehensive information.[135] Particularly relevant for SPECIAL are encryption, MPC, access control, policy enforcement, accountability, data provenance, transparency and, last but not least, user control. Regarding user control, the representatives referred to a transparency and compliance dashboard that is developed in a manner that tackles the users' cognitive limitations. Key functions of the dashboard include presenting data processing and sharing events in an easily digestible manner, and enabling the user to understand the implications of existing and future consent for processing and sharing. Concerning accountability and transparency, it was stated that SPECIAL focuses on transparency and compliance checking of personal data processing and sharing with respect to usage policies and legal obligations. With respect to data provenance, the representatives mentioned that all data sharing and processing is stored in a log. Moreover, it is ensured that the W3C provenance framework[136] fits seamlessly into the SPECIAL approach. The project focuses on the enforcement of three types of policies: usage policies, legislative obligations and business rules. Regarding access control, it was pointed out that the SPECIAL platform as well as transparency and compliance APIs are restricted to those who are authorised. With respect to MPC, SPECIAL builds on the results of the project Big Data Europe and the SANSA stack[137]. Encryption is particularly relevant for SPECIAL in the context of linked data. Anonymisation, and access and data portability are no major foci of SPECIAL but they will most likely be touched within the scope of the project. Sanitisation is not on the project's roadmap. The SPECIAL consortium has already published several scientific articles describing aspects of the project's development activities in detail.

The 3-year project started in January 2017 and will end in December 2019.

4.1.2. SODA

The **SODA**[138] (Scalable Oblivious Data Analytics) project has two main aims. The first aim is to enable practical privacy-preserving analytics of information from multiple data assets using MPC techniques. The main technological challenge that SODA faces is to make MPC, where data is only made available for encrypted processing, scale to big data. Within the scope of the project, MPC is embedded into a comprehensive privacy approach. Therefore, the second aim is to combine MPC with a multidisciplinary approach towards privacy. Legal analyses performed in a feedback loop with technical development are expected to ensure compliance with EU data protection regulation. The results will at least be validated in a medical demonstrator. The techniques will be subjected to public hacking challenges and technical innovations released as open-source improvements to the FRESCO MPC framework[139].

So far, four deliverables have been made available by SODA. The most relevant one for e-SIDES is D2.1, which describes the state of the art of MPC-based big data analysis. The deliverable provides an overview of PPDM primitives, (plain-text) data mining, streaming algorithms and reviews previous works on PPDM.

[135] The information was provided by e-mail.
[136] http://www.w3.org/standards/techs/provenance#w3c_all
[137] http://sansa-stack.net/
[138] https://www.soda-project.eu/
[139] https://github.com/aicis/fresco

Representatives of the SODA project provided e-SIDES with a detailed discussion of SODA in the light of the technologies introduced in chapter 3.[140] They emphasised once again that the main task and challenge of the SODA project is to make MPC practical for analytics on big data. To enable MPC on big volumes of data the scalability need to be improved and performance bottlenecks solved. In order to do so, a use case-driven approach is used to bring together theoretical as well as practical expertise. Concerning technology development, SODA focused on:

- MPC (this is the primary technology SODA uses to enable privacy-preserving big data analytics);
- Anonymisation and sanitisation (after determining the result of a data analytics task using MPC, post-processing is performed to make sure that this result does not leak unintended personal/sensitive information; for this purpose, anonymisation and sanitisation approaches such as differential privacy are used).

However, many of the other groups of technologies are important in the context of the SODA project as well and will surely be a part of any viable solution, in particular:

- Encryption (used as a building block in MPC solutions but not to achieve a security end by itself);
- User control, access control and policy enforcement (when personal data is used in MPC, it needs to be ensured that the data subject is in charge of what happens to their personal data);
- Transparency (data subjects should have insight on how their personal information is used).

SODA also aims to provide a thorough legal analysis of the current privacy law in the EU, with emphasis on the GDPR. The building blocks of the legal evaluation are:

- To assess what anonymous data mean under the GDPR, and whether encrypted or otherwise de-identified data can be regarded as anonymous data.
- To delineate the lawfulness of processing of personal data as well as the special conditions of the processing of special categories of personal data (e.g., health data).
- To outline the legal obligations placed on the controller and the processor by the GDPR.
- To evaluate the rights of the data subject.

Access and data portability is addressed under point 2 of the legal evaluation. SODA aims to demonstrate, inter alia, whether – and under which circumstances – a change in purpose and re-use of data is allowed, especially for scientific research. From the perspective of the individual, the right of access and the right of data portability will be discussed primarily under point 4. However, since every right of a data subject is on the other hand also an obligation for the controller, the issue will be addressed under point 3 as well.

Accountability, transparency and user control are also relevant for SODA's legal evaluation, mainly in relation to points 3 and 4. The principle of transparency, according to the GDPR, places an overall duty on the controllers and processors to carry out the processing activities in a clear and comprehensible way. Accountability means, that controllers shall be responsible for and able to demonstrate compliance with the GDPR, and serves thereby as a core principle when discussing the question of liability.

[140] The information was provided by e-mail.

The 3-year project started in January 2017 and will end in December 2019.

4.1.3. MHMD

The **MHMD**[141] (My Health - My Data) project aims to build and test new models of privacy and data protection that meet the requirements of the biomedical sector, in which issues of data subjects' privacy and data security represent a crucial challenge. The current ICT landscape in the sector, according to MHMD, shows a myriad of isolated, locally hosted patient data repositories, managed by clinical centres and other organisations. Even massive data breaches are thus not uncommon and patients generally lack a clear understanding of who uses their personal information and for what purposes. The project works on an open biomedical information network also referred to as MHMD platform. Within the scope of the MHMD project, it is aimed to guide the implementation of data and identity protection systems, assess de-identification and encryption technologies, allow advanced analytics, and evaluate the overall reliability of a generic multi-modular architecture.

From the e-SIDES point of view, apart from the work on the platform, the envisaged analysis of users' behavioural patterns, and ethical and cultural orientations is of particular relevance. MHMD aims to identify hidden dynamics in the interactions between humans and complex information services, to improve the design of data-driven platforms and to foster the development of a true information marketplace, in which individuals are able to exercise full control on their personal data and leverage their value.

So far, one deliverable that focuses on requirements has been published. The deliverable does not only describe requirements for sharing personal and health data but also outlines user stories and specifies features to be implemented by the MHMD platform. It is expected that several issues relevant for e-SIDES will emerge over the course of the MHMD project (in particular with respect to anonymisation, encryption, accountability and user control). Consequently, the progress of the project will be monitored closely.

The 3-year project started already in November 2016 and ends in October 2019.

4.2. Other projects

In the following, we focus on the Innovation Actions (IAs) funded under ICT-14-2016-2017 (Big data PPP: Cross-sectorial and cross-lingual data integration and experimentation) and ICT-15-2016-2017 (Big data PPP: Large Scale Pilot actions in sectors best benefitting from data-driven innovation). The IAs are tasked to address data challenges in cross-domain setups and to carry out large scale sectorial demonstrations, respectively. The e-SIDES project aims to support them in the development of solutions that protect personal data and address related issues not only in line with existing EU rules but also anticipating the expectations of more confident data subjects in the future.

The relevant ICT-14-2016-2017 projects are SLIPO, AEGIS, EW-Shopp, Data Pitch, QROWD, euBusinessGraph, FashionBrain and BigDataOcean.

[141] http://www.myhealthmydata.eu/

The **SLIPO**[142] (Scalable Linking and Integration of scalable Big POI data) project aims to deliver technologies to address the data integration challenges of POI data in terms of coverage, timeliness, accuracy and richness. In previous projects such as the EU-funded project GeoKnow linked data technologies have been developed and applied to extract value from open, crowd-sourced and proprietary data. SLIPO now focuses on the transfer of the research output of these projects to POI data and the introduction of validated and cost-effective innovations across their value chain.

So far, no deliverables have been published. As the project does not use personal data, privacy is not a key issue for SLIPO. However, trustworthiness of data is a central aspect, according to a project representative.[143] SLIPO has to deal intensively with data quality issues when addressing data integration challenges. Another relevant aspect is the interdependency between and the dominance of specific actors in the field.

Regarding the groups of technologies introduced in chapter 3, those addressing data provenance, and access and portability are most relevant for SLIPO. They are not the main focus of the project but they are relevant.

The 3-year project started in January 2017 and will end in December 2019.

The **AEGIS**[144] (Advanced Big Data Value Chain for Public Safety and Personal Security) project aims to create a curated, semantically enhanced, interlinked and multilingual repository for safety-related data that allows organisations in the public safety and personal security sector to provide better services to their customers. Among the main areas to which AEGIS will contribute are the representation of knowledge, intelligence extraction and application development, business models for the data sharing economy, crypto currency algorithms to validate transactions and to handle IPRs, data quality and data privacy issues.

Several deliverables have already been completed by the project. D1.2, which describes the project's methodology as well as high-level application scenarios, is particularly relevant for e-SIDES. Among others, the deliverable describes ethical, privacy and IPR considerations and, based on them, proposes a strategy that takes relevant requirements and regulations fully into account. It further specifies that AEGIS combines the Privacy-by-Design and the Data Protection Goals method, and defines 17 concrete requirements ranging from legitimate aim and purpose limitation to adequate mechanisms and tools for safeguarding IPRs on data artefacts and data usage. Another deliverable that is relevant for e-SIDES is D1.1. The deliverable reviews the domain landscape and defines the data value chain. The AEGIS consortium points out clearly that it considers privacy a critical aspect in the context of data use.

Early in 2018 a deliverable will be completed that describes the final set of pilots. These pilots will have gone through an ethics review. Moreover, a mock-up of the data platform will be available at that time.

AEGIS has an Ethics Advisory Board that works closely with the project consortium to tackle ethical and privacy issues and makes sure respective requirements are met. The project uses personal data but makes

[142] http://www.slipo.eu/
[143] A project representative was interviewed during the European Big Data Value Forum.
[144] https://www.aegis-bigdata.eu/

sure the data is anonymised. A project representative stated that the actual way in which the anonymised or non-personal data is used should be up to the user of the platform.[145]

With respect to the technologies introduced in chapter 3, a project representative explained that anonymisation and sanitisation, access control, data provenance and access and portability are most relevant for AEGIS. With respect to anonymisation and sanitisation, AEGIS enhances and uses a pseudonymisation tool that was developed within the scope of a previous project. A block-chain-based approach is applied to ensure data provenance.

The 2.5-year project started in January 2017 and will end in June 2019.

The **EW-Shopp**[146] (Supporting Event and Weather-based Data Analytics and Marketing along the Shopper Journey) project aims to deploy and host a data integration platform to ease data integration tasks in the e-commerce, retail and marketing domain. This will be achieved by embedding shared data models, robust data management techniques and semantic reconciliation methods. The project integrates contextual variables (e.g., weather conditions, calendar events, holidays) into the analysis of consumer behaviour.

So far, three general deliverables have been published. These deliverables describe the project's data management plan, the website setup and the exploitation and dissemination strategy. There does not seem to be a direct relation to privacy or other societal and economic issues. However, e-SIDES will carefully monitor the future development of EW-Shopp as it is very likely that privacy related issues would come up as the project proceeds.

The 3-year project started in January 2017 and will end in December 2019.

The **Data Pitch**[147] (Accelerating data to market) project is an open innovation programme that aims at bringing together corporate and public-sector organisations that have data with start-ups and SMEs that work with data. It raises a competition with several tracks, which describe challenges set by the data-provisioning organisations. An example for such a track is Health/Wellness: It is asked how data can be used to help people improve their health and wellness and/or make health services more efficient and inclusive, while ensuring that patient and citizen privacy is respected in the use and analysis of this data.

Currently, no deliverables of the project are publicly available. As the first call was closed just recently, it is likely that this will change soon. We will keep track of the upcoming developments of Data Pitch as there are many areas that seem to be of relevance for e-SIDES.

The 3-year project started in January 2017 and will end in December 2019.

The **QROWD**[148] (Because Big Data Integration is Humanly Possible) project aims to offer innovative solutions to improve mobility, reduce traffic congestion and make navigation safer and more efficient. To

[145] A project representative was interviewed during the European Big Data Value Forum.
[146] http://www.ew-shopp.eu/
[147] https://datapitch.eu/
[148] http://qrowd-project.eu/

achieve that, QROWD integrates geographic, transport, meteorological, cross-domain and news data with the goal of maximizing the value of big data in planning and managing urban traffic and mobility.

Five deliverables have already been published. They include a data catalogue as well as descriptions of requirements and the architecture, the project's online presence and brand guidelines, and the data management plan. Moreover, there is a deliverable focusing on the protection of personal data (POPD) and ethics. This deliverable states that ethical and privacy-protective treatment of the data is crucial to the development of the project. Therefore, we will continue to monitor the project’s outcomes in the future as it is likely that many issues of interest to e-SIDES will emerge.

The 3-year project started in December 2016 and will end in November 2019.

The **euBusinessGraph**[149] (Enabling the European Business Graph for Innovative Data Products and Services) project aims to create a business graph, which is understood as a highly interconnected graph of Europe-wide company-related information both from authoritative and non-authoritative sources (including data from both the public and the private sector). By doing so, the project strives to provide a data marketplace that enables the creation of a set of data-driven products and services via a set of six business cases.

So far, no deliverables have been published. We will continue to monitor the project’s development to see if a relation to privacy or other ethical or societal issues emerges.

The 2.5-year project started in January 2017 and will end in June 2019.

The **FashionBrain**[150] (Understanding Europe’s Fashion Data Universe) project aims at combining data from different sources to support different fashion industry players (i.e., the retailers and the customers) by predicting upcoming fashion trends from social media as well as by providing personalised recommendations and advanced fashion item search to customers.

So far, no deliverables have been published. We will continue to monitor the project’s development to see if a relation to privacy emerges.

The 3-year project started in January 2017 and will end in December 2019.

The **BigDataOcean**[151] (Exploiting Oceans of Data for Maritime Applications) project aims to enable maritime big data scenarios through a multi-segment platform that combines data of different velocity, variety and volume under an interlinked, trusted and multilingual engine. The project capitalises on existing big data technologies but rolls out a new value chain of interrelated data streams that is expected to transform the way maritime-related industries work. The infrastructure will be combined with four pilots that will allow developing a large maritime database. The pilots focus on fault prediction and proactive maintenance, protection from oil spills, maritime security and anomaly detection, and clean energy from wave power.

[149] http://eubusinessgraph.eu/
[150] https://fashionbrain-project.eu/
[151] http://www.bigdataocean.eu/

BigDataOcean has already completed several deliverables of which some are available on the project website. Two deliverables focusing on ethical requirements have not (yet) been made available. Among the deliverables that are available, D2.1, which describes an analysis of big data components, tools and methodologies, is the most relevant one from the perspective of e-SIDES. Within the scope of the discussion of the questions related to data collection, among others, data access control, privacy and security are addressed. Technology options that are discussed include encryption and MPC. It is not detailed, however, what role privacy plays in the context of BigDataOcean.

A representative of the project stated that privacy as well as other ethical and societal aspects do not play a central role for BigDataOcean.[152] Nevertheless, the representative pointed out that the project is very interested in a closer collaboration with e-SIDES. It was suggested that members of the e-SIDES team get access to an early version of the data platform. It could then discuss possible ethical and societal issues with the developers. This is a great opportunity for e-SIDES to better understand the challenges developers face when trying to address societal and ethical issues in the context of big data solutions.

The 2.5-year project started in January 2017 and will end in June 2019.

The relevant ICT-15-2016-2017 IAs are TT and DataBio.

The **TT**[153] (Transforming Transport) project aims to demonstrate the transformations that big data will bring to the mobility and logistics market in terms of, for instance, increased operational efficiency, improved customer experience and new business models. The project focuses on how big data reshapes transport processes and services using pilots from seven domains that are particularly important for the sector. These domains include smart highways, sustainable vehicle fleets, proactive rail infrastructures, ports as intelligent logistics hubs, efficient air transport, multi-modal urban mobility, and dynamic supply chains.

The deliverables the project has completed so far mostly focus on the individual pilot designs. From the perspective of e-SIDES, D1.3 and D3.2 are particularly relevant. D1.3 discusses the project's IPR and data management approach. Among others, the deliverable describe problems with data assets and possible solutions. One of the aspects discussed in this context is the anonymisation of personal data. Location data/floating car data, social media data, passenger data and e-commerce customer data are described as data asset types that require particular attention. D3.2 describes the project's data management plan. For each pilot, the deliverable not only describes which data assets are used and if they are shared but also if ethical aspects need to be considered. The chapters on ethical aspects mostly focus on the question if personal data is used or not and, if personal data is used, how the data subjects' privacy is protected. The pilot design deliverables address, among others, the data assets used as well as technical aspects of the pilot deployment.

TT offered e-SIDES to add a few questions to a survey that will be completed by the teams working on the pilots.[154] The survey is a great opportunity for e-SIDES to better understand the role privacy (and related aspects) plays in real-world settings. Most likely, e-SIDES will formulate questions focusing on the

[152] A project representative was interviewed during the European Big Data Value Forum.
[153] https://transformingtransport.eu/
[154] A project representative was interviewed during the European Big Data Value Forum.

potential gap between what existing privacy-preserving big data technologies offer and what the needs in practice are. Moreover, the perceived ability to understand existing technologies, and to determine and express concrete needs may be addressed. Additionally, TT provided e-SIDES with an internal document detailing privacy-related difficulties the project faces with respect to location data.

The 2.5-year project started in January 2017 and will end in June 2019.

The **DataBio**[155] (Data-Driven Bioeconomy) project aims to demonstrate the benefits of big data technologies in terms of sustainable improvement and productivity of bioeconomy industry raw materials. For this purpose, a data platform is built based on existing software components that is suitable for different industries and user profiles to ensure effective utilisation of existing datasets, effective participation of the ICT industry and easy setup of new multivendor applications. The project deploys pilots in the fields of agriculture, forestry and fishery. Among the main areas to be addressed by the platform are data acquisition and curation, data variety management, predictive analytics and machine learning, real-time analytics and stream processing, and advanced visualisation and customised technological feedback to the user. Privacy aspects are addressed in the context of data acquisition and curation.

The deliverables the project has completed so far mostly focus on the pilot definitions. For e-SIDES, the data management plan detailed in D6.2 is of particular relevance as it described the dataset used in the project. Moreover, it addresses how privacy is protected and what ethical issues need to be taken into account in DataBio. However, the guidelines mostly focus on how information is exchanged between members of the consortium. With respect to the pilots, it's the lead partner that is responsible. The pilot definition deliverables address, among others, the data assets used for the individual pilots.

A representative of the DataBio project stated that with respect to privacy, the project hopes to be able to implement guidelines provided by the BDVA.[156] The representative also suggested that e-SIDES might cooperate with the BDVA in this regard.

The 3-year project started in January 2017 and will end in December 2019.

Apart from the ICT-14-2016-2017 and ICT-15-2016-2017 IAs, the links between the classes of technologies and the ICT17-2016-2017 (Big data PPP: Support, industrial skills, benchmarking and evaluation) CSA BDVe and the ICT35-2016 (Enabling responsible ICT-related research and innovation) RIA K-PLEX are relevant for e-SIDES.

The **BDVe**[157] (Big Data Value ecosystem) project aims at supporting the Big Data Value Public-Private Partnership (PPP) in realising a vibrant data-driven EU economy. The major priorities for the project are various. BDVe strives to be accurately informed about the most important facts in big data as a solid basis to support the decision-making process in the PPP and develop a lively community. Furthermore, the implementation of the PPP from an operational point of view is supported as well as the development of a European network of infrastructures and centres of excellence around big data. Additionally, a

[155] https://www.databio.eu/
[156] A project representative was interviewed during the European Big Data Value Forum.
[157] http://www.big-data-value.eu/

professional communications strategy is being set up as well as a framework that is expected to support the acceleration of data-driven businesses.

The 4-year project started in January 2017 and will end in December 2020.

The **K-PLEX**[158] (Knowledge Complexity) project investigates the strategies humanities and cultural researchers have developed to deal with their typically unstructured data. It aims to use humanities knowledge to explore bias in big data approaches to knowledge creation.

So far, no deliverables have been published. Although privacy does not seem to among the project's priorities, we will continue to monitor the project's progress.

The 15-month project started in January 2017 and will end already in March 2018.

Last but not least, one of the project representatives interviewed in connection with one of the ongoing research project also mentioned the recently completed research project DAIAD. The project was funded under ICT-2013-11 (ICT for water resources management) and is relevant for e-SIDES in several respects.

The **DAIAD**[159] (Open Water Management - From Droplets of Participation to Streams of Knowledge) project started from the assumption that the empowerment of consumers is a potentially groundbreaking approach for efficient water use and reuse. Therefore, the project aimed to develop innovative low cost, inclusive technologies for real-time, high-granularity water monitoring and knowledge extraction. The consortium devised multi-modal feedback interfaces, recommendation and analysis services to communicate knowledge and incur behavioural changes to consumers in residential settings. Moreover, big data management and analysis technologies were applied to provide efficient management and analysis of real-time water consumption data as well as multiple relevant data sources.

The project representative the e-SIDES team interviewed explained that privacy was relevant for DAIAD insofar as consent had to be obtained from trial participants.[160] Issues and values that were most relevant for DAIAD are trustworthiness, interdependency and welfare.

With respect to the technologies introduced in chapter 3, anonymisation and user control are most relevant for the project. Anonymisation was not the main focus of the project but it was relevant. Regarding user control, consent mechanisms were particularly relevant.

The 3.5-year Collaborative Project (CP) started in March 2014 and ended in August 2017.

[158] https://kplex-project.com/

[159] http://daiad.eu/

[160] A project representative was interviewed during the European Big Data Value Forum.

5. Summary and conclusion

This report provides an overview of existing approaches, methods and technologies that may have the potential to address ethical, legal, societal and economic issues that are raised by big data applications. The issues including not only threats to privacy but also, for instance, threats to self-determination, strong interdependencies, limited trustworthiness and lack of accountability were identified in a previous phase of the project (see Deliverable 2.2 of the project).

Based on a comprehensive literature review, technologies considered as privacy-enhancing or privacy-preserving were identified and assigned to eleven classes:

1) **Anonymisation** is performed by encrypting or removing personally identifiable information from datasets. Traditional anonymisation techniques fail in the context of big data applications because there are hundreds of data points for a single individual. A full de-identification cannot be achieved. Privacy models that may be used when anonymizing data include k-anonymity, l-diversity, t-closeness and differential privacy.
2) **Sanitisation** is done by encrypting or removing sensitive information from datasets. Anonymisation is a type of sanitisation aiming at privacy protection. Sanitisation techniques other than encryption and removal of columns include masking data, substitution, shuffling and number variance. In the big data era, for instance, it can be difficult to find substitution data in large quantities.
3) **Encryption** is the encoding of information so that only authorised parties can access it. In the context of big data applications it is not enough to combine the advantages of public key encryption in scalability and key management with the speed and space advantages of symmetric encryption. Fine grade sharing policies are necessary that go beyond the *encrypt all or nothing* model. Relevant cryptographic primitives include ABE, IBE, PRE and functional encryption.
4) **Multi-party computation** (MPC) relies on the distribution of data and processing tasks over multiple parties. MPC is a field of cryptography with the aim to allow securely computing the result of any function without revealing the input data. Although MPC was proven to be theoretically plausible, there are still no practical solutions. Key challenges in the big data context are utility, performance and ease of use.
5) **Access control** describes the selective restriction of access to places or resources. Big data applications typically require fine-grained access control. Traditional approaches such as RBAC and user-based access control are becoming less and less manageable. ABAC in an example for a set of approaches that can conceptually support fine grained access control policies in big data based on attributes that are evaluated at run-time.
6) **Policy enforcement** focuses on the enforcement of rules for the use and handling of resources. Automated policy enforcement mechanisms are particularly important in the big data era as policies get easily lost or neglected in the course of data being transferred between different systems. Data expiration policies, for instance, are already enforced by some big data solutions.

7) **Accountability** requires the evaluation of compliance with policies and the provision of evidence. A cornerstone of accountability in the context of big data applications is the provision of automated and scalable control and auditing processes that can evaluate the level of compliance with policies. PDP and POR are among the main approaches to cloud data-integrity verification without retrieval.
8) **Data provenance** relies on being able to attest the origin and authenticity of information. The aim is to provide a record of the processing history of pieces of data. Fine-grained provenance is difficult to achieve because big data is typically highly heterogeneous. Additionally, the use of many different analytics and storage solutions may result in prohibitively large amount of provenance information to be transferred between systems.
9) **Transparency** calls for the explication of information collection and processing. In the context of big data applications, transparency may be achieved by purely textual information, multichannel and layered approaches, or standardized icons and pictograms. Transparency is considered critical to allow data subjects informed choices.
10) **Access and portability** facilitates the use and handling of data in different contexts. Having access to data means that data subjects can look through and check the data stored. Portability gives data subjects the possibility to change service providers without losing their data.
11) **Users control** refers to the specification and enforcement of rules for data use and handling. Consent mechanisms are one means that allows reaching user control, others are privacy preferences, sticky policies and personal data stores.

As the project aims to complement other research on privacy-preserving big data technologies and data-driven innovation funded by the EU under the H2020 programme, the role the classes of technologies play in such projects was also studied. For this purpose, project representatives were interviewed personally or asked to provide relevant information per e-mail. Moreover, the websites as well as already publicly available deliverables of relevant projects, mostly those funded under ICT-18-2016, ICT-14-2016-2017 and ICT-15-2016-2017, were analysed.

The potential that the classes of technologies actually have to address the issues will be assessed in the next phase of the e-SIDES project.

The main conclusions of the work leading to this report are:

- There is a **considerable body of literature** on technologies considered as privacy enhancing or privacy-preserving. The foundations for many of the classes of technologies have been laid already decades ago, particularly as far as classes such as anonymisation, encryption, access control and accountability are concerned. The difficulties related to big data came to the fore more recently.
- The **pool of relevant technologies is highly diverse**. Approaches, methods and technologies, although they often have a common basis, regularly differ in a few details. Consequently, if is neither possible to provide an exhaustive overview nor to indicate what developments are superior to others and for what reason. There is n.
- The literature does not provide insight into the **relevance of the technologies in practice**. It remains unclear to what extent the approaches, methods and technologies discussed are actually

part of commercial products or publicly available solutions used in the context of big data applications. Getting such insight requires to approach the developers directly.

- The **assessment of the potential** that the classes of technologies actually have to address the issues needs to be carried out in close consultation with stakeholder actually involved in big data applications. The literature alone does not provide a suitable basis for a holistic assessment. It appears useful to consider aspects such as big data lifecycle phases and user roles during the assessment.
- The emergence of **big data changes the protection of privacy** significantly. The same holds for the approaches pursued to address other related issues that become more relevant in the context of big data applications or more difficult to address. A full de-identification of data is not achievable. Moreover, it is challenging to find an adequate balance between the protection of privacy and the utility of big data.
- Other projects on data-driven innovation funded by the EU devote **different amounts of attention to privacy** and related issues. While some tackle ethical and societal issues comprehensively and take measures to make sure related respective requirements are met, others seem to turn a blind eye to such issues. Some of them stated explicitly that they hope the Big Data Value Association or other projects such as e-SIDES will provide them with the guidance they might need.
- The **recognition of responsibility for tasks** related to addressing ethical and societal implications of the data platforms they develop and the services they provide varies significantly from one project to another. While some of the EU-funded projects assume the full responsibility of their developments and actively search for solutions themselves, others attempt to delegate a significant part of the responsibility to the users of their platforms and services. They take the measures prescribed by law but leave everything else entirely to the user.

Bibliography

Abbas, Assad, and Samee U. Khan. “A Review on the State-of-the-Art Privacy-Preserving Approaches in the e-Health Clouds.” *IEEE Journal of Biomedical and Health Informatics* 18, no. 4 (2014): 1431–1441.

Aggarwal, Charu C., and Philip S. Yu. “A General Survey of Privacy-Preserving Data Mining Models and Algorithms.” In *Privacy-Preserving Data Mining: Models and Algorithms.* vol. 34. Edited by Charu C. Aggarwal and Philip S. Yu, 11–52. Advances in Database Systems, 34 v. v. 34. Boston, MA: Springer US, 2008.

Article 29 Data Protection Working Party. “Opinion 05/2014 on Anonymisation Techniques.”. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (accessed December 14, 2017).

Bambauer, Jane R., Krish Muralidhar, and Rathindra Sarathy. “Fool's Gold: an Illustrated Critique of Differential Privacy.” *Vanderbilt Journal of Entertainment & Technology Law* 16, no. 4 (2014): 701–755.

Bertino, Elisa, Gabriel Ghinita, Murat Kantarcioglu, Dang Nguyen, Jae Park, Ravi Sandhu, Salmin Sultana, Bhavani Thuraisingham, and Shouhuai Xu. “A roadmap for privacy-enhanced secure data provenance.” *Journal of Intelligent Information Systems* 43, no. 3 (2014): 481–501.

Bonawitz, Keith, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. B. McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. “Practical Secure Aggregation for Privacy-preserving Machine Learning.”. Cryptology ePrint Archive 2017/281. https://eprint.iacr.org/2017/281.

Borking, John J., and Charles D. Raab. “Laws, PETs and Other Technologies for Privacy Protection.” *Journal of Information Law & Technology*, no. 1 (2001). https://www2.warwick.ac.uk/fac/soc/law/elj/jilt/2001_1/borking (accessed September 26, 2017).

Calders, Toon, and Bart Custers. “What Is Data Mining and How Does It Work?” In *Discrimination and Privacy in the Information Society.* vol. 3. Edited by Bart Custers et al., 27–42 3. Berlin, Heidelberg: Springer, 2013.

Castano, Silvana, Mariagrazia Fugini, Giancarlo Martella, and Pierangela Samarati. *Database security*. ACM Press, 1995.

Cavoukian, Ann, and Jeff Jonas. “Privacy by Design in the Age of Big Data.”. https://jeffjonas.typepad.com/Privacy-by-Design-in-the-Era-of-Big-Data.pdf (accessed December 14, 2017).

Colombo, Pietro, and Elena Ferrari. “Complementing MongoDB with Advanced Access Control Features: Concepts and Research Challenges.” In *Proceedings of the 23rd Italian Symposium on Advanced Database Systems*. Edited by Domenico Lembo, Andrea Marrella and Riccardo Torlone, 343–50. Curran Associates, 2015.

——— . “Privacy Aware Access Control for Big Data: A Research Roadmap.” *Big Data Research* 2, no. 4 (2015): 145–154.

Custers, Bart, Toon Calders, Bart Schermer, and Tal Zarsky, eds. *Discrimination and Privacy in the Information Society* 3. Berlin, Heidelberg: Springer, 2013.

“D2.1 State of the Art Analysis of MPC-Based Big Data Analytics.”. https://www.soda-project.eu/wp-content/uploads/2017/02/SODA-D2.1-WP2-State-of-the-art.pdf.

D'Acquisto, Giuseppe, Josep Domingo-Ferrer, Panayiotis Kikiras, Vicenç Torra, Yves-Alexandre de Montjoye, and Athena Bourka. “Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics.”. https://www.enisa.europa.eu/publications/big-data-protection/at_download/fullReport (accessed September 26, 2017).

Davidson, Susan B., Sanjeev Khanna, Sudeepa Roy, Julia Stoyanovich, Val Tannen, and Yi Chen. “On provenance and privacy.” In *Proceedings of the 14th International Conference on Database Theory*. Edited by Tova Milo, 3–10. New York, New York, USA: ACM Press, 2011.

Druschel, Peter, Eslam Elnikety, Deepak Gar, Aastha Mehta, and Anjo Vahldiek. “Towards Accountable Information Systems.” (2014). http://dig.csail.mit.edu/2014/AccountableSystems2014/abs/druschel-ext-abs.pdf (accessed December 14, 2013).

Du, Wenliang, and Mikhail J. Atallah. “Secure multi-party computation problems and their applications: a review and open problems.” In *Proceedings of the 2001 Workshop on New Security Paradigms*, 13–22. 2001.

Dwork, Cynthia, and Aaron Roth. “The Algorithmic Foundations of Differential Privacy.” *Foundations and Trends® in Theoretical Computer Science* 9, 3-4 (2014): 211–407.

Erway, Chris, Alptekin Küpçü, Charalampos Papamanthou, and Roberto Tamassia. “Dynamic provable data possession.” In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, 213–22. ACM, 2009.

Fung, Benjamin C. M., Ke Wang, Rui Chen, and Philip S. Yu. “Privacy-preserving data publishing.” *ACM Computing Surveys* 42, no. 4 (2010): 1–53.

Goyal, Vipul, Omkant Pandey, Amit Sahai, and Brent Waters. “Attribute-based encryption for fine-grained access control of encrypted data.” In *Proceedings of the 13th ACM Conference on Computer and Communications Security*, 89–98. 2006.

Haeberlen, Andreas, Benjamin C. Pierce, and Arjun Narayan. “Differential privacy under fire.” In *Proceedings of the 20th USENIX Conference on Security*, 33. USENIX Association Berkeley, 2011.

He, Daojing, Sammy Chan, and Mohsen Guizani. “Accountable and Privacy-Enhanced Access Control in Wireless Sensor Networks.” *IEEE Transactions on Wireless Communications* 14, no. 1 (2015): 389–398.

Heurix, Johannes, Peter Zimmermann, Thomas Neubauer, and Stefan Fenz. “A taxonomy for privacy enhancing technologies.” *Computers & Security* 53 (2015): 1–17.

Holohan, Naoise, Spiros Antonatos, Stefano Braghin, and Pól M. Aonghusa. “(k,ε)-Anonymity: k-Anonymity with ε-Differential Privacy.” arXiv:1710.01615. https://arxiv.org/pdf/1710.01615.pdf.

Hoogh, Sebastiaan de. “Design of large scale applications of secure multiparty computation: secure linear programming.”. PhD thesis.

Hu, Yin. “Improving the Efficiency of Homomorphic Encryption Schemes.”. PhD thesis. https://web.wpi.edu/Pubs/ETD/Available/etd-042513-154859/unrestricted/YHu.pdf (accessed January 15, 2018).

Jain, Priyank, Manasi Gyanchandani, and Nilay Khare. “Big data privacy: A technological perspective and review.” *Journal of Big Data* 3, no. 1 (2016): 120.

Jennings, Nicholas R., Peyman Faratin, Alessio R. Lomuscio, Simon Parsons, Carles Sierra, and Michael Wooldridge. “Automated Negotiation: Prospects, Methods and Challenges.” *International Journal of Group Decision and Negotiation* 10, no. 2 (2001): 199–215.

Korolev, Vlad, and Anupam Joshi. “PROB: A tool for tracking provenance and reproducibility of big data experiments.” In *Proceedings of the 20th IEEE International Symposium on High Performance*

Computer Architecture. Orlando, Florida, USA: IEEE, 2014. http://ebiquity.umbc.edu/_file_directory_/papers/693.pdf (accessed December 14, 2017).

Li, Ninghui, Tiancheng Li, and Suresh Venkatasubramanian. “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity.” In *Proceedings of the IEEE 23rd International Conference on Data Engineering*, 106–15. IEEE, 2007.

——— . “Closeness: A New Privacy Measure for Data Publishing.” *IEEE Transactions on Knowledge and Data Engineering* 22, no. 7 (2010): 943–956.

Li, Shuyu, Tao Zhang, Jerry Gao, and Younghee Park. “A Sticky Policy Framework for Big Data Security.” In *Proceedings of the IEEE 1st International Conference on Big Data Computing Service and Applications*, 130–7. IEEE, 2015. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7184873 (accessed December 14, 2017).

Lindell, Yehuda, and Benny Pinkas. “Secure Multiparty Computation for Privacy-Preserving Data Mining.” *The Journal of Privacy and Confidentiality* 1, no. 1 (2009): 59–98.

Liu, Chang, Jinjun Chen, Laurence T. Yang, Xuyun Zhang, Chi Yang, Rajiv Ranjan, and Kotagiri Rao. “Authorized Public Auditing of Dynamic Big Data Storage on Cloud with Efficient Verifiable Fine-Grained Updates.” *IEEE Transactions on Parallel and Distributed Systems* 25, no. 9 (2014): 2234–2244.

Liu, Chang, Chi Yang, Xuyun Zhang, and Jinjun Chen. “External integrity verification for outsourced big data in cloud and IoT: A big picture.” *Future Generation Computer Systems* 49 (2015): 58–67.

Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. “L -diversity.” *ACM Transactions on Knowledge Discovery from Data* 1, no. 1 (2007): 3-es.

Matwin, Stan. “Privacy-Preserving Data Mining Techniques: Survey and Challenges.” In *Discrimination and Privacy in the Information Society.* vol. 3. Edited by Bart Custers et al., 209–21 3. Berlin, Heidelberg: Springer, 2013.

Mehmood, Abid, Iynkaran Natgunanathan, Yong Xiang, Guang Hua, and Song Guo. “Protection of Big Data Privacy.” *IEEE Access* 4 (2016): 1821–1834.

Mivule, Kato. “Utilizing Noise Addition for Data Privacy, an Overview.”. https://arxiv.org/ftp/arxiv/papers/1309/1309.3958.pdf (accessed January 15, 2018).

Moreno, Julio, Manuel Serrano, and Eduardo Fernández-Medina. “Main Issues in Big Data Security.” *Future Internet* 8, no. 3 (2016): 44.

Nelson, Boel, and Tomas Olovsson. “Security and privacy for big data: A systematic literature review.” In *2016 IEEE International Conference on Big Data: Dec 05-Dec 08, 2015, Washington D.C., USA : proceedings*. Edited by James Joshi, 3693–702. Piscataway, NJ: IEEE, 2016.

Park, Jaehong, Dang Nguyen, and Ravi Sandhu. “A provenance-based access control model.” In *Proceedings of the 10th Annual International Conference on Privacy, Security and Trust*, 137–44. IEEE, 2012.

Pasquier, Thomas, Jatinder Singh, Julia Powles, David Eyers, Margo Seltzer, and Jean Bacon. “Data provenance to audit compliance with privacy policy in the Internet of Things.” *Personal and Ubiquitous Computing* 8, no. 12 (2017). https://link.springer.com/content/pdf/10.1007/s00779-017-1067-4.pdf (accessed December 14, 2017).

Pasquier, Thomas F. J.-M., and David Eyers. “Information Flow Audit for Transparency and Compliance in the Handling of Personal Data.” In *Proceedings of the 2016 IEEE International Conference on Cloud Engineering: Workshops*, 112–7. IEEE, 2016.

Pasquier, Thomas F.J.M., Jatinder Singh, Jean Bacon, and Olivier Hermant. “Managing Big Data with Information Flow Control.” In *Proceedings of the IEEE 8th International Conference on Cloud Computing*, 524–31. IEEE, 2015. https://www.cl.cam.ac.uk/research/srg/opera/publications/papers/2015ieeeCloud.pdf (accessed December 14, 2017).

Pettersson, John S. “A brief evaluation of icons suggested for use in standardised information policies: Referring to the Annex in the first reading of the European Parliament on COM (2012) 0011.”. Working paper. https://www.diva-portal.org/smash/get/diva2:720798/FULLTEXT01.pdf (accessed January 15, 2018).

Pfitzmann, Andreas, and Marit Hansen. “A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management.”. http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf.

Samarati, Pierangela, and Sabrina D. C. Di Vimercati. “Data protection in outsourcing scenarios: Issues and directions.” In *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security.* 2010.

Samuel, Arjmand, Muhammad I. Sarfraz, Hammad Haseeb, Saleh Basalamah, and Arif Ghafoor. “A Framework for Composition and Enforcement of Privacy-Aware and Context-Driven Authorization Mechanism for Multimedia Big Data.” *IEEE Transactions on Multimedia* 17, no. 9 (2015): 1484–1494.

Sarathy, Rathindra, and Krishnamurty Muralidhar. “Evaluating Laplace Noise Addition to Satisfy Differential Privacy for Numeric Data.” *Transactions on Data Privacy* 4, no. 1 (2011).

Sen, Shayak, Saikat Guha, Anupam Datta, Sriram K. Rajamani, Janice Tsai, Wing, and Jeannette M. “Bootstrapping Privacy Compliance in Big Data Systems.”. https://www.andrew.cmu.edu/user/danupam/sen-guha-datta-oakland14.pdf (accessed December 14, 2017).

Shen, Wenting, Jia Yu, Hui Xia, Hanlin Zhang, Xiuqing Lu, and Rong Hao. “Light-weight and privacy-preserving secure cloud auditing scheme for group users via the third party medium.” *Journal of Network and Computer Applications* 82 (2017): 56–64.

Soria-Comas, Jordi, and Josep Domingo-Ferrer. “Big Data Privacy: Challenges to Privacy Principles and Models.” *Data Science and Engineering* 1, no. 1 (2016): 21–28.

Spiekermann, Sarah. *Ethical IT innovation: A value-based system design approach*. Boca Raton, London, New York: CRC Press, 2016.

Sweeney, Latanya. “k-Anonymity: A model for protecting privacy.” *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems* 10, no. 5 (2002): 557–570.

Tas, Yucel, Mohamed J. Baeth, and Mehmet S. Aktas. “An Approach to Standalone Provenance Systems for Big Social Provenance Data.” In *Proceedings of the 12th International Conference on Semantics, Knowledge and Grids*, 9–16. IEEE, 2016.

Ulusoy, Huseyin, Pietro Colombo, Elena Ferrari, Murat Kantarcioglu, and Erman Pattuk. “GuardMR: Fine-grained Security Policy Enforcement for MapReduce Systems.” In *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security - ASIA CCS '15*. Edited by Feng Bao et al., 285–96. New York, New York, USA: ACM Press, 2015.

Upadhyaya, Prasang, Magdalena Balazinska, and Dan Suciu. “Automatic Enforcement of Data Use Policies with DataLawyer.” In *Proceedings of the 2015 ACM SIGMOD International Conference on*

Management of Data, 213–25. ACM, 2015. http://cloud-data-pricing.cs.washington.edu/datalawyer.pdf (accessed December 14, 2017).

Wang, Boyang, Baochun Li, and Hui Li. “Oruta: Privacy-preserving public auditing for shared data in the cloud.” *IEEE Transactions on Cloud Computing* 2, no. 1 (2014): 43–56.

Wang, Jianwu, Daniel Crawl, Shweta Purawat, Mai Nguyen, and Ilkay Altintas. “Big data provenance: Challenges, state of the art and opportunities.” In *Proceedings of the 2015 IEEE International Conference on Big Data*, 2509–16. IEEE, 2015. http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7364047 (accessed December 14, 2017).

Wang, Yang, and Alfred Kobsa. “Privacy-Enhancing Technologies.” In *Handbook of research on social and organizational liabilities in information security*. Edited by Manish Gupta and Raj Sharman. Hershey, PA: Information Science Reference, 2009.

Xu, Lei, Chunxiao Jiang, Jian Wang, Jian Yuan, and Yong Ren. “Information Security in Big Data: Privacy and Data Mining.” *IEEE Access* 2 (2014): 1149–1176.

Zhou, You-sheng, En-wei Peng, and Cheng-qing Guo. “A Random-Walk Based Privacy-Preserving Access Control for Online Social Networks.” *International Journal of Advanced Computer Science and Applications* 7, no. 2 (2016): 74–79.

Appendix

This appendix provides the summary of a Big Data Value PPP networking session on privacy-preserving big data technologies held as part of the 2017 European Big Data Value Forum. The 90-minute session was held on 22 November 2017 and organised jointly by BDVe and e-SIDES.

Data-driven innovation is deeply transforming society and the economy. Although this transformation brings huge opportunities and enormous economic and social benefits for individuals and businesses, a greater access and use of data creates new challenges in many domains. The main implications affect privacy, security, as well as democracy and participation. To ensure a fair balance between data protection and data-driven innovation, legislators are pushing the research and business actors to consider and address these challenges emerging from the development and use of big data and analytics. To draw the attention of the big data community, and discuss how big data technologies may help overcoming them and preventing loss of privacy we organized a networking session as a platform for discussion.

The session consisted of a keynote speech and a panel discussion centring around on three main questions. The audience was involved through a real-time voting on the three questions as well as the opportunity to direct questions to the panellists.

Ernesto Damiani gave the keynote speech. Damiani is research professor at Khalifa University and chair of the university's Information Security Group and Program, and EBTIC. He is on extended leave from the Department of Computer Science of the Università degli Studi di Milano, where he leads the SESAR research lab and is the head of the PhD Program in Computer Science.

Ernesto Damiani provided a **keynote speech** regarding big data and privacy perspectives. As an overall framework, he stressed the importance of a paradigm shift. According to his vision, "big data is not just a technological advance but represents a paradigm shift in extracting value from complex multi-party processes". He demonstrated this paradigm shift by pointing out the diminishing role of classic data warehouses through a chart that showed a tendency in big data aggregation from structured data towards unstructured data and from batch to real time data aggregation. In his speech, he highlighted that nowadays data comes from a large number of sources and claimed that "by the time you open data you need it for another purpose". Therefore, he argued that there is a move towards a multi-party vision in computation, which results in instances where big data is released via different ways but also via formats. According to him big data can predict discrete attributes, predict continuous attributes, determine groups and predict influencers. In his speech, he identified three main problems:

1) Data representations for big data applications are inherently anti-privacy. Data relations are made in current MPC networks explicit.
2) Where to anonymise? Ernesto Damiani suggested that, for instance, within the context of online monitoring for stream analytics anonymisation is difficult; within the context of online monitoring for batch analytics anonymisation would be possible based on premises made for the user. Furthermore, in the contexts of offline mining classical anonymisation techniques are available and could be useful to learn from.

3) How will privacy-preservation affect analytics? He considered this as almost a *contradictio in terminus*, because big data analytics is about linking variables for predictions, but three common pitfalls can be distinguished: clustering after hashing, reducing dimensionality after anonymisation and deep-learning encrypted/anonymised data.

Six panellists participated in the panel discussion. Names, projects and organisations of the panellists are shown in Table 4. The panellists were first asked to provide short position statements with respect to privacy-preserving big data technologies. Most of panellists used the position statement to explain how their project is related to big data and privacy. Subsequently, they were asked to briefly comment on three questions.

Name	Project	Organisation
Sabrina Kirrane	SPECIAL	Vienna University of Economics and Business
Anna Zsofia Horvath	SODA	University of Göttingen
Edwin Morley-Fletcher	MHMD	Lynkeus
Gianluca Ripa	TT	Cefriel
Mike Priddy	K-PLEX	DANS
Philip Carnelley		IDC

Table 4 Names, projects and organisations of the panellists

Philip Carnelley introduced himself as representing the industrial point of view. In his statement he recapped some of the reasons why big data is important: because it brings intelligent applications, better services, it increases customer satisfaction and thus brings great benefits for individuals, governments and businesses in general.

Mike Priddy described K-PLEX as a project that focuses on the complexity of knowledge within the humanities. It centres upon the importance of inventorising narratives that result in big data, the importance of hidden data, machine translation data, mapping inequalities in language and on black-box problems. K-PLEX also has the objective to answer how big data can be used to give advice for these problems. Whereas the project thus is not directly about privacy and security issues, it touches many related aspects while collecting, managing and using data.

Gianluca Ripa explained the TT project as one that is run based on 13 pilots that are small digital ecosystems. They first of all describe data so that they can show how public data is shared or managed by private companies and what the challenges are in data management within these public-private relationships.

Edwin Morley-Fletcher introduced MHMD as a project that develops a model for data sharing between hospitals and patients. It aims at securing health data in blockchain systems and at developing trusted third-party-based systems while tackling GDPR challenges regarding data within the healthcare sector.

Anna Zsofia Horvath introduced SODA as project that is centred on highly sensitive (healthcare) data. She stressed that SODA aims to overcome challenges that emerge from three specific case studies. One case focuses on data within hospitals. The other two cases are inter-sectorial: one assesses big data combination throughout hospital and insurance companies' relationship and a second throughout hospital-to-hospital relationships. As a primary task of the project, she also mentioned to evaluate the legal implications of the GDPR.

Sabrina Kirrane described SPECIAL as a project that aims at assessing how to anonymise data, how to give meaning to consent, by investigating what we know about processing in light of the challenges of the GDPR. According to her statement, a key question within this is how to balance transparency for users with respect to their data and still offer rich and useful environments (e.g., recommendations).

The **first question** focused on the compatibility of big data exploitation and privacy. The panellists were asked whether or not there is a trade-off between big data exploitation/data driven innovation and privacy from their point of view. Also, the potential effects of a trade-off are of interest.

To summarize the various aspects, most panellists agreed that there is some kind of a trade-off and that it is difficult to keep data driven innovation and privacy balanced.

In Sabrina Kirrane's view the two can co-exist, although it might be difficult to achieve. While user studies have shown that a lot of people actually want useful recommendations, there is a huge lack of awareness that has to be addressed. Still it is difficult to get consent as the legislation is very subjective. She additionally stressed the importance of two aspects; i.e. transparency and compliance that ground the informed consent that is used in the system.

Anna Zsofia Horvath agreed that there is a trade-off, but that it is complicated to achieve. While the enormous value of big data is out of question, the GDPR as a legal framework is not concrete enough. Furthermore, while consent seems to be the best but not the only way, scientific secondary privileged purposes are not explained well and are not secure enough.

Edwin Morley-Fletcher seconded that there is a trade-off that is complicated. Dynamic consent and smart contracts are used in MHMD. Still some institution want to move beyond consent. Patient data for example is used sometimes for other purposes. For example: patient associations provided a lot of complaints regarding areas of encryption, and the Art. 29 WP recommended that secure MPC, homomorphic encryption could be useful for some of the problems, but not all. There are lot of remaining gaps there.

Gianluca Ripa also sees the difficulty in balancing the two. While there are algorithms that can provide data depending on the contexts the problem of the costs of the solutions becomes more and more important. So data use is not only context- but also increasingly cost-dependent.

In Mike Priddy's opinion big data is about individual identity in a rich way. Big data registration and aggregation about individuals affects them in multiple ways. Therefore the right not to know that should also be considered. Furthermore he stresses that privacy is broader than the individual. Confidential data exists, but the self is relational it produces social facts and has consequences, and the current legislation is individual oriented, whereas we have built big networks. Where is trust? How about misused data? A big challenge remains between transparency and accountability, but also regarding accountability versus responsibility.

According to Philip Carnelley theoretically there is no trade-off, but practically there is one. From an industry perspective, big data brings great benefits but so does privacy. By increasing the trust in what they are doing businesses can actually be more competitive. Because if consumers would trust your company more regarding what is done with their data than those companies, which do not have that, they would choose your company versus others.

The **second question** focused on the maturity of developments in the field. The panellists were asked to assess the current level of development of privacy-preserving big data technologies. Also, the challenges that need to be met and how they can be addressed were of interest. To summarize, all of the participants agreed that the need for development is still enormous.

Philip Carnelley for instance thought that the level of maturity is rudimentary. A big problem is that although the GDPR does apply in the US and the EU, most of the US mind-set about privacy is different and US companies need to change their mind-set (especially because they use EU citizens' data and they have to simply comply with EU rules). Once people realize that maybe development will speed up, in the US as well as in Europe.

In Mike Priddy's perspective the humanities need to be in the centre of the design. While dealing with tools that will manage and edit data there often is a lack of understanding what the narrative in the context is, what the challenges are, where data comes from. In his opinion, there are lot more problems than privacy that need to be tackled.

Gianluca Ripa thinks that a lot of problems stem from private companies managing public infrastructures. These enormous amounts of data need to be appropriately managed as normally they are not for sharing per se.

Edwin Morley-Fletcher is happy that he is a European as through the GDPR, transparency is also more available to healthcare. In his view, routine data must be made shareable and comparable. He also mentioned the concept of qualified anonymity: data can be anonymous from outside protected by a key – but this could allow for much things to happen.

Anna Zsofia Horvath emphasizes that when we talk about privacy-preservation, we talk about the basic human right, and all that also comes from the GDPR. Even data that is technically possible to re-identify can at the end of the day be seen as anonymous. So, that technicality could be less of a burden from the perspective of the GDPR.

Sabrina Kirrane tackled the problem that many of the computer scientists have not learnt ethics as it is not a basis in all sciences (although in her opinion it should be). In data science they should think about ethics, but this is a challenge. Data scientists need to comprehend that they cannot build technology without ethics requirements. In her experience, vocabulary turned out to be an issue, e.g. the definition of personal data and privacy. Still people want utility, so templates are needed, e.g. you protect either too much your information, or you can give richer recommendation how to use it.

The **third question** focused on differences between application areas. The panellists were asked for application areas in which meeting the privacy requirements is particularly difficult.

Sabrina Kirrane mentioned telecom operators. All companies' businesses depend on data-driven services, financial regulation, and the importance of the ePrivacy directive kicks in. Their demos show how they can improve city infrastructures. But transparency and compliance in a generic way can go through all domains and what cannot should be vertical and rather context-dependent.

Anna Zsofia Horvath stated that when someone works with healthcare data, general difficulties with compliance with the GDPR emerge. First, there are a lot of complicated issues, for instance, with respect to informed consent or concerning the terms used to refer to patients. Second, do not overload patients with data. Third, how do I recalibrate data for different purposes? How can I prove that the latter is necessary for improvement? She stressed that the contextual element is important.

Edwin Morley-Fletcher mentioned the different values of big data: The predictive value, the economic value and there is something more. No catch-22 mechanisms are needed. For instance, in one hospital he used synthetic data (by def. anonymous data), which is logically very difficult. Virtual patients (synthetic data) are currently used by pharmaceuticals for prognosis and other types of analysis.

Gianluca Ripa sees a problem that at the moment, 'real forecasts images' cannot be provided by data analysts.

Mike Priddy had a problem with how the question is formulated. In his view, technology is moving faster than society. Privacy is a public good and should be regarded as the right to not to be recorded in the first place. Most of all people need to understand how data is being used, and what and how you give it away; we need an informed community out there.

Philip Carnelley added that from an industry point of view sales and marketing are leading. Why do you use AI on the top of big data? Most reasons were marketing and sales. But, when using AI how do you explain what you exactly did and what you did not with data? Lots of issues remain there.

Questions from the audience touched upon issues, such as:

- Why is the US mind-set towards privacy bad/not constructive, why is the European one better for big data value preservation? Some panellists indicate the EU perspective better addresses the information and power asymmetry between individuals and companies.
- Follow-on effects and societal effects of privacy. It is noted that privacy is not only an issue for individuals. For instance, when DNA data is used by law enforcement, this also contains data regarding relatives. Another example is posting group pictures on social media on which are also other people that may not have consented to this.
- The privacy friendliness of future technology and the extent to which future legislation will require this. One of the panellists noted that blockchain technology may contribute to this, as this may be used to enhance transparency and combined with other tools/technologies may result in more privacy.

The results of the real-time voting are shown in Table 5. Of approximately 50 participants in the networking session, a total number of 28 cast their votes. The participants were asked to rate three statement on a Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

Statement	Average score
There is an apparent trade-off between big data exploitation and privacy.	3.6
Regarding the level of development, privacy-preserving big data technologies still have a long way to go.	4.1
Individual application areas differ considerably with regard to the difficulty of meeting the privacy requirements.	3.9

Table 5 Statements and average scores of the real-time voting