Privacy-Preserving Technologies for Automating GDPR Compliance

The General Data Protection Regulation (GDPR) is a legal framework that aims to protect personal data privacy and ensure secure data handling within the European Union. With the growing popularity of machine learning models and the increasing need for data privacy, there is a pressing need to automate GDPR compliance in data-driven applications. This summary systematic scientific literature overview on privacy-preserving technologies that can achieve GDPR compliance, distinguishing between user-centric data protection, automated compliance, and learning from big data analysis in a privacy-preserving manner, highlights the best practices and key gaps in up-to-date research, informing the contribution of TITAN to GDPR-compliant secure data sharing and processing in the EOSC.

The widespread use of Machine Learning and AI models for data analysis and personalised risks assessment and prediction (in health/well-being and other fields) has recently been raising significant and increasing privacy concerns. Zentrix has been working on developing digitally blind evaluation systems that eliminate proxies and enable privacy-preserving data analytics in the related ongoing and previously started initiatives, such as other projects in the European AI Security Network (EASiNet) initiative, like the HARPOCRATES project.

Three main categories of current approaches for achieving GDPR compliance can be distinguished: user-centric data protection, automated compliance, and learning from big data analysis in a privacy-preserving manner.

User-centric data protection

User-centric data protection approaches emphasize the privacy and control of users over their personal data while enabling secure processing of such data. These approaches often involve the use of advanced cryptographic techniques to protect data from unauthorized access and processing while still allowing specific authorized computations. We discuss two prominent cryptographic schemes in the context of user-centric data protection: functional encryption and hybrid homomorphic encryption.

- Functional Encryption (FE) is a powerful cryptographic primitive that enables the decryption of encrypted data to reveal a specific function of the underlying plaintext without revealing any additional information (Sahai & Waters 2005). FE allows data owners to delegate specific computations to untrusted parties without compromising data privacy. Several studies have explored the application of FE to GDPR compliance, particularly in the context of access control and secure data sharing. However, FE schemes are often computationally expensive and may introduce scalability challenges, especially for large-scale applications.
- Hybrid Homomorphic Encryption (HHE) is another cryptographic approach to user-centric data protection that combines the advantages of both partial and fully homomorphic encryption schemes (Clear & McGoldrick 2014). HHE enables secure computations on encrypted data without requiring the data to be decrypted, thereby ensuring data privacy. The use of HHE has been proposed for various privacy-preserving applications, such as secure data aggregation and privacy-preserving machine learning (Gilad-Bachrach, Dowlin et al. 2016). HHE is particularly relevant to GDPR compliance as it allows organizations to process user data without exposing sensitive information, thus fulfilling the data protection requirements mandated by the regulation.

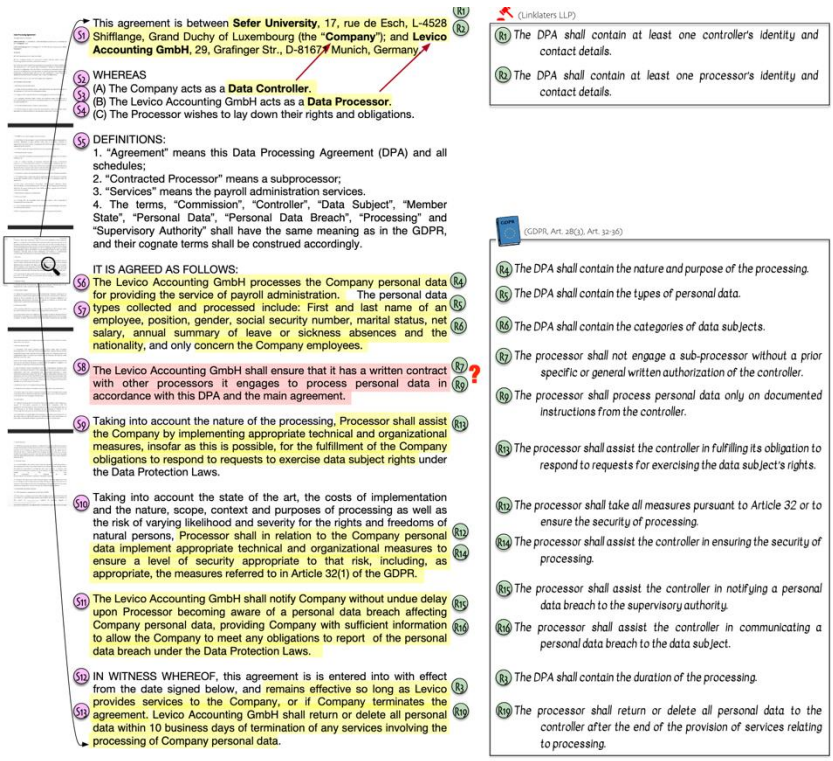
Despite the potential of these cryptographic schemes in achieving user-centric data protection, there are still limitations and challenges that need to be addressed. For instance, the practical implementation of these schemes can be complex and computationally intensive, potentially hindering their adoption in real-world applications. Further research is needed to improve the efficiency and scalability of these cryptographic techniques to meet the growing demands of GDPR compliance in privacy-preserving data processing.

Automated compliance

Automated compliance approaches seek to streamline the GDPR compliance process by leveraging machine learning algorithms and artificial intelligence. These approaches primarily focus on the automated identification of personally identifiable information (PII) in datasets, ensuring data subject rights are met, and monitoring compliance through privacy risk assessments.

Automated PII identification is a critical aspect of GDPR compliance, as organizations need to be aware of the sensitive information they store and process. Several studies have proposed machine learning-based methods for detecting PII in structured and unstructured data - however, these techniques suffer from limited accuracy and require large, representative training datasets to achieve optimal performance.

Data subject rights, such as the right to be forgotten and the right to data portability, are essential components of GDPR compliance. Automated compliance solutions in this area focus on developing tools and frameworks that enable organizations to efficiently manage and respond to data subject requests. One approach involves the use of blockchain technology to create immutable records of data processing activities, which can be easily audited and shared with data subjects as needed (Mahindrakar & Joshi 2020).Another approach entails the development of natural language processing algorithms to automatically interpret and process data subject requests submitted in free text format (Amaral, Azeem et al. 2021), as exemplified on the image on the right.



Privacy risk assessments are necessary for organizations to understand and manage the risks associated with data processing activities. Several automated compliance methods have been proposed to facilitate the assessment and monitoring of privacy risks, including the use of decision support systems and artificial intelligence techniques to predict and

evaluate potential privacy violations (Lore, Basile et al. 2023). These methods can help organizations proactively identify potential GDPR non-compliance issues and take corrective actions to mitigate risks.

Despite the advancements in automated compliance approaches, there remain challenges in terms of accuracy, scalability, and adaptability to diverse organizational contexts. Further research is needed to enhance the effectiveness of these methods and develop comprehensive solutions that address the multifaceted requirements of GDPR compliance.

Learning from big data analysis in a privacy-preserving manner

This category explores methods like differential privacy, federated learning, and secure multi-party computation to learn from big data while preserving privacy. These approaches enable organizations to draw insights and learn from large-scale datasets without exposing sensitive information, thus adhering to the GDPR requirements and compliance.

Differential privacy (DP) is a mathematical framework for quantifying and managing the privacy risks associated with the release of statistical information derived from sensitive data. It provides strong privacy guarantees by adding carefully calibrated noise to the output of data analysis algorithms, ensuring that individual data records cannot be distinguished or re-identified. Recent studies have explored the application of DP in various domains, including machine learning, data mining, and statistical analysis, to achieve GDPR compliance (Cummings & Desai 2018).

Federated learning (FL) is a distributed machine learning paradigm that enables multiple data owners to collaboratively train a shared model without disclosing their raw data (Bonawitz, Kairouz et al. 2021). FL allows organizations to extract valuable insights from decentralized datasets while preserving data privacy and complying with GDPR requirements. Several works have investigated the application of FL in domains such as healthcare, finance, and smart cities, demonstrating its potential for privacy-preserving data analysis (Truong, Sun et al. 2021).

Secure multi-party computation (SMPC) is a cryptographic technique that allows multiple parties to jointly compute a function over their private inputs without revealing the inputs themselves. SMPC enables privacy-preserving data processing and analytics in collaborative environments, making it a relevant approach for GDPR compliance. Recent research has explored the use of SMPC in various applications, including secure data sharing, privacy-preserving data mining, and encrypted data analysis (Veeningen, Chatterjea et al. 2018).

Though all the described approaches have shown promise in enabling privacy-preserving automated GDPR compliance, substantial challenges remain, as all the approaches introduce trade-offs between privacy, data utility, learning from big data, and computational efficiency. Further research is needed to address these limitations and develop more efficient and scalable privacy-preserving techniques, optimize the trade-offs, and investigate the practical applicability of these techniques in real-world settings, key challenges facing the TITAN work on potential implementation and enriching of EOSC with the support for automated intelligent cross-organizational and cross-border sensitive data sharing compliant with GDPR.