Methods of allowing third parties to process data efficiently with security and privacy guarantees: A Systematic Literature Review

James Eggers

October 18, 2022

Abstract

Our world is full of information of varying degrees of sensitivity. Some of our most sensitive data, such as financial data and passwords, are leaked and stolen–often to disastrous consequences like eroding individuals' trust in third parties using their data for essential tasks, such as applying for a loan. This literature review will review the state of modern research into pipelines that allow for the efficient processing of sensitive data with security and privacy guarantees for the data subject.

1 Introduction

In today's world, companies and governments often use our most sensitive data in various ways, such as reviewing our bank statements when applying for a loan or hoarding our browsing history to show us more targeted advertisements.

These tasks often have good outcomes for us; we may get the loan or be shown more relevant ads to us that may be helpful but along the way, the possibility always exists that the data could end up in the wrong hands. Due to this, the public is becoming warier about sharing their data which is difficult for the individual because they have additional security measures to consider when sharing their data. Processing this data is difficult for companies as they must have very high levels of security to process this data and not leak it, and sometimes this happens regardless.

This literature review will explore ways of sharing sensitive data so that the entity who shared it can be confident that only the desired parties have access to the data.

2 Methodology

In this paper, we perform a systematic literature review (SLR) to document methods of guaranteeing the security and privacy of shared sensitive data in

quantitative and qualitative ways. Our research questions that this paper should answer and discuss are as follows:

2.1 Research questions

- What methods of reliably handling data that third parties can process in a way that guarantees security and privacy exist?
- Which methods are efficient enough to use in a natural setting?

2.2 Search Queries

- data AND (privacy OR security) AND (guarantee OR assurance)

2.3 Inclusion and exclusion criteria, and quality assessment criteria

The query is generally limited to the paper title due to the large body of research on this topic. We wanted to focus on privacy guarantees and exclude papers that do not focus on this. The search is also generally limited to research published from 2013 to 2023 to that relatively modern research can be reviewed. Due to their high quality and subject matter relevance, we included some papers from outside the 2013 - 2023 time period.

Research that fell into the following categories was excluded:

- Research not overly focused on our research area.
- Research that did not present a novel method for guaranteeing data privacy or security
- Research that was from a poor-quality journal.

2.4 Selection and validation process

Google Scholar was used to get a broad spectrum of research; in total, the query returned 49 results, of which 27 we excluded, leaving 22 pieces of relevant research for this literature review. Figure 1 is a diagram that shows the overall process that was undertaken to select relevant papers. Figure 2 shows a graph of the number of papers gathered per category, and Figure 3 shows a graph of the number of papers gathered per source.

3 Research Categorisation

In this section, we organise our papers into categories depending on the methods of guaranteeing privacy they describe.

In table 1 we list each category we sorted the research into and the corresponding number of papers in that category. Table 2 lists our sources with the number of papers from that source.

Figure 1: Methodology Diagram

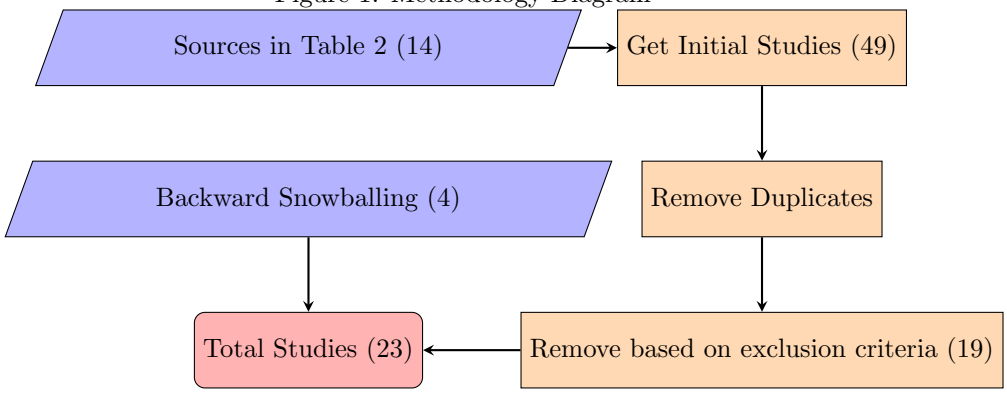

Figure 2: Volume of papers per category

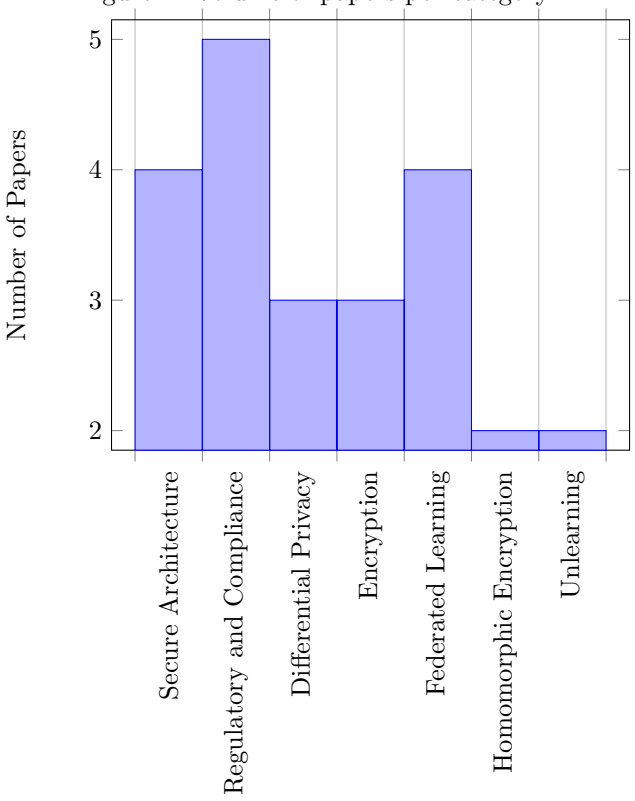

Figure 3: Volume of papers per source

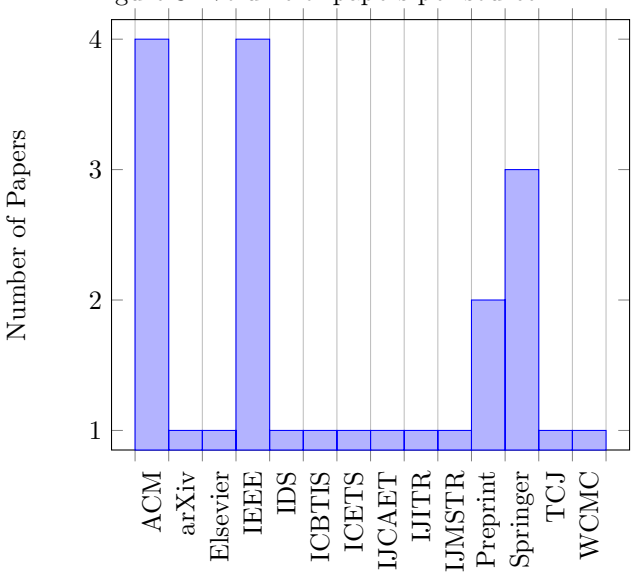

Table 1: Research Categories

Category	Number of Papers
Secure Architecture	4
Regulatory and Compliance	5
Differential Privacy	3
Encryption	3
Federated Learning	4
Homomorphic Encryption	2
Unlearning	2

Table 2: Research Sources

Category	Number of Papers
ACM	4
arXiv	1
Elsevier	1
IEEE	4
Information and Documentation Services	1
International Conference on Blockchain Technology and Information Security	1
International Congress on Educational and Technology in Sciences	1
International Journal of Computer Aided Engineering and Technology	1
International Journal of Innovative Technology and Research	1
International Journal of Monitoring and Surveillance Technologies Research	1
Preprint	2
Springer	3
The Computer Journal	1
Wireless Communications and Mobile Computing	1

4 Result Synthesis

In this section, we analyse each category to understand how it can contribute to our research questions.

4.1 Secure Architecture

Saavedra et al. (2020) proposed a technique guaranteeing sensitive patient data privacy. This method works by masking sensitive data through several steps and then passing it along to hospital administrators, who can analyse it without seeing the irrelevant sensitive data belonging to the patient, severely reducing the risk to patient privacy. Saavedra et al. (2020) describes high compliance rates with the model, resulting in an efficient system that guarantees patient privacy without adding complex overheads like encryption.

De la Rosa Algarín & Demurjian (2014) describes a system that uses role-based-access-control (RBAC), whereby certain people are assigned roles, and based on their role; they can see certain information they require. Peng et al. (2022) describes a system for improved access control where "the improved access control system has the more efficient guarantee ability in education information security, which is about 47.3% higher than the traditional access control system.".

Another critical aspect of a privacy-focused architecture is to empower its

users to use it in a way that limits the spread of their data while still achieving their primary goals. In Naher & Hashem (2019) a system for location sharing is described that leans into this goal by telling the user when they share their location if an adversary may be able to deduce if it is a sensitive location to the user, such as their home.

4.2 Regulatory and Compliance

We can think of the legal system as a way to ensure data privacy. By making a law, we try to direct the behaviour of various entities. In Lingbin (2022), new regulations are described for cross-border sharing, hoping to increase user privacy. Fajar Marta (2016) investigates using privacy-minded design principles for open government data portals, resulting in a set of guidelines a government can use to incorporate privacy into their data portals. However, these are guidelines and not regulations, resulting in a lower bar for compliance and, thus, not as effective a privacy measure. We can see other situations where privacy and security guidelines are created, Sen & Madria (2018) discusses methods of augmenting preexisting guidelines. It is an interesting exercise to think about how new regulations or guidelines may guarantee a user's privacy. In Gharib (2020), a model of citizens sharing their data only with entities they expressly consent to while being legally guaranteed of this by various privacy laws.

There are, of course, various standards (such as the ISO/IEC 27000 Family) an entity such as a company or government can be externally audited for and eventually get a certification to say they comply with it. These standards can help provide a level of confidence to a data subject that their data will be treated with a measure of protection. Tamunobarafiri et al. (2017) describes how in 2016 alone, over 16 million healthcare records were breached; a new assessment tool is proposed in this paper which can be used to assess and choose cloud vendor solution. This assessment tool is likened in Tamunobarafiri et al. (2017) to other frameworks such as those from the Cloud Security Alliance.

4.3 Federated Learning

Originally proposed by Google in 2016 in Konečný et al. (2016), Federated Learning is the idea that we can train a centralised model over several decentralised nodes without the data leaving those nodes. Since the data never leaves the node presumably owned by the owner of the data, we can be assured that privacy can be guaranteed, assuming the receiver cannot reverse engineer the data from the model, a point made in Jiang et al. (2022).

Building on this, Yang et al. (2019) acknowledges the difficulties involved in the AI industry today, one of which is noted as privacy and security. The paper proposes a comprehensive, secure federated learning framework which builds on Konečný et al. (2016).

Although Federated Learning does not involve sending large amounts of data to a centralised server for processing, Jiang et al. (2022) makes the point that frequent model updates could result in communication model bottlenecks–and

it will be essential to improve efficiency. However, Federated Learning has been put to actual and practical use in the following situations, showing it can be used efficiently to preserve at least a measure of privacy:

- In Jiang et al. (2022), a system for optimising how mobile networks regulate transmission power using Federated Learning is proposed.
- Google, in Konečný et al. (2016), describe several experiments using Federated Learning, such as next-word prediction in a Reddit data set.
- In 2017, Google also proposed using Federated Learning for mobile phone model updates in McMahan et al. (2016).

4.4 Differential Privacy

Li et al. (2015) discusses the release of sensitive medical data, such as portions of the population who are HIV positive. For some people, this is a sensitive piece of information they do not want to be general knowledge, so when this data is released, it could be anonymised to remove identifying information such as the person's name. However, if their age, sex, and location are released, and the sample size is relatively small, it may be possible to narrow down whom the person is by inferring it from the age, sex, and location data.

In Dwork et al. (2006), a system for adding noise to the results of a query on a database is described. As the size of the database increases, less noise needs to be added to a query. However, smaller databases will need more noise; this idea would later become differential privacy. Differential privacy can therefore help us share information about data sets patterns without compromising the privacy of the individuals involved in sharing the data. In Li et al. (2015), we see how this can be applied to HIV statistics by efficiently adding noise to the returned data and presenting it in a histogram without compromising patient privacy.

In Studer (2013), a technique called knowledge base distortion is described where we distort a data set to such an extent that it is impossible to leak private information about a person, which could be an efficient way of processing data with privacy guarantees. However, it is only suitable for a specific problem domain where the actual data does not need to be sent.

4.5 Encryption

Encryption-based techniques for reliably handling data and processing it with privacy guarantees are perhaps the most promising. Abiodun et al. (2021) notes that Internet-of-Things (IoT) devices have become increasingly common and that they generate a lot of sensitive data (for example, a wifi-enabled home security camera stores and streams sensitive video footage). The paper proposes a method of protecting the data using the Triple-DES cryptographic algorithm. At the same time, this is not particularly novel; it shows that we can protect this data using cryptography efficiently and securely.

In Sivakami & Umadevi (2022), an improved, novel Rivest Shamir Adleman (ERSA) algorithm is described. The algorithm works in three phases:

- Firstly, digital signatures are utilised for authentication.
- The Fuzzy Inference System (FIS) is used to ensure the required level of security
- The improved ERSA algorithm is then used to encrypt the data. The paper's authors explain that they improved upon RSA by generating primes using the Sieve of Atkins (SoA) algorithm or using a non-prime factor.

The level of security this algorithm provides is debatable due to being new; often, security holes can be found later. However, broadly, symmetric Encryption like Tripe-DES mentioned in Abiodun et al. (2021) is efficient and secure (for now). If substituted for a modern alternative like AES, we can securely and efficiently encrypt data and provide a privacy guarantee. However, our research question is if we can allow third parties to process our data with privacy guarantees. With AES / DES, a third party cannot process the data meaningfully without first decrypting it, so this solution is likely not a very good candidate to fulfil this objective.

KAVITHA (n.d.) describes a novel addition to public-key encryption to resolve some security vulnerabilities they note in the paper. Public key encryption utilises asymmetric encryption to share data between two parties without ever sharing the private key that can be used for decryption; this is very advantageous as it allows us to efficiently and privately transfer data between two parties. However, if we do not entirely trust the receiving party, this method is not as helpful as it only protects the data from outside observers.

4.6 Homomorphic Encryption

Homomorphic encryption allows us to encrypt a piece of data and for a third party to apply certain mathematical functions (such as addition, subtraction, multiplication, or division) to the data. At the same time, it is still encrypted–once the result is decrypted, it is the correct answer. Naehrig et al. (2011) describes the efficiency of Fully Homomorphic Schemes as the "elephant in the room" for this type of technology. When we talk about Homomorphic Encryption, we usually refer to two kinds, Fully Homomorphic Encryption (FHE) and Somewhat Homomorphic Encryption (SHE). FHE implements addition, subtraction, multiplication, and division, and SHE implements a subset of those. While Fully Homomorphic Encryption schemes allow third parties to meaningfully process our data while the third party never actually has direct access to it, it is too slow to be efficient for many applications, as noted in Naehrig et al. (2011). However, this paper does list some concrete examples of workable use cases for Homomorphic Encryption:

- In a medical setting, automated monitoring devices could constantly stream numerical data to an FHE-based cloud system which could operate on the encrypted data
- In the financial world, FHE-based cloud systems could be used to operate on encrypted data such as confidential corporate financials or stock prices.

It is all well and good to theorise about potential uses. However, the interesting aspect here is efficiency. In Naehrig et al. (2011), the paper describes that they could use Somewhat-Homomorphic Encryption, and the sum of 100 128-bit numbers can be computed in 20 milliseconds on a consumer-grade laptop. These are encouraging results, obviously, for large amounts of data like image or video analysis, the amount of data is considerably larger, and processing times will be slower. However, for processes which require operating on small quantities of numerical data with privacy guarantees, Homomorphic Encryption Schemes seem promising.

Xu et al. (2022) presents a "data-privacy preserving" scheme based on the blockchain and homomorphic encryption; the paper says their scheme can "carry out effective data encryption transmission".

4.7 Unlearning

If a company has been asked to delete specific data, it may be required to remove it from any machine learning models trained on that data. This situation is described in Cong & Mahdavi (2022), and they proposed an exciting novel method of removing relevant data from models; this is called "unlearning". We can use unlearning to create a more secure data processing pipeline; for example, we cannot purport to have "data privacy guarantees" if we do not remove the data we said we would.

Aldaghri et al. (2021) presents a method of splitting input data for a model up into shards and also presents an unlearning protocol; they say that their method shows promising experimental results when compared with the baseline performance.

Unlearning provides an important solution to a difficult problem, machine learning models can be nebulous and tools to retrain them with certain data removed can significantly improve the privacy guarantee of a system–after all you cannot lose data you do not have,

5 Discussion and Insights

To re-cap, the research questions we want to answer are as follows:

- What methods of reliably handling data that third parties can process in a way that guarantees security and privacy exist?
- Which methods are efficient enough to use in a natural setting?

Table 3: Abbreviation Legend

Abbreviation	Description
LD	Large amount of data
SD	Small amount of data
VO	Need to view original data
NVO	No need to view original data

Table 4: Methods of processing data with privacy and security guarantees in an efficient way

Context / Section	4.1	4.2	4.3	4.4	4.5	4.6	4.7
LD & NVO	✓	✓	✓	✓	✓		✓
SD & NVO	✓	✓	✓	✓	✓	✓	✓
LD & VO	✓	✓			✓		✓
SD & VO	✓	✓			✓		✓

In the course of our literature review, we found six potential methods of handling data that third parties can process in a way that can guarantee security and privacy. Of those, all could be efficient, but not all could be used efficiently in all situations. For example, creating a securely architected role-based access system could provide privacy guarantees by not allowing certain people access to sensitive data, and it could do this efficiently. However, homomorphic encryption could arguably provide a better privacy guarantee but only in situations where the quantity of data to be analysed is small and numerical. Given this consideration, we conclude that we must consider the context and shape of the data we want to protect when trying to create a secure data processing pipeline with privacy guarantees.

In table 4, we outline different situations in which different methods could be used efficiently to answer our second research question. Table 3 is a legend for the abbreviations used in table 3.

To answer our first research question, we can say that several methods can allow us to process data with privacy and security guarantees. However, the method we choose is contingent upon the context in which we process the data. As well as this, some methods provide varying degrees of guarantee. For example, with a secure role-based architecture that only allows authorised personnel to access sensitive health data, the possibility of a bug in the authorisation exists, allowing unauthorised personnel to view specific data. Compared with homomorphic encryption (4.6), encrypted data cannot be decrypted without the decryption key, meaning even in the case of an authorisation bug, the data remains private.

Our second research question looks at the efficiency of the various methods detailed in this literature review. The context of the data must again be taken into account; some ways will provide a high privacy guarantee but very low efficiency, while others may do the opposite. Table 4 is a matrix showing which

methods would be efficient with certain data types and contexts.

5.1 Research Gaps

In the course of writing this literature review, several research gaps were identified and are listed below:

- Homomorphic Encryption (4.6) presents a holy-grail type answer to our research questions. Research has been ongoing for many decades; however, efficient fully homomorphic encryption (FHE) that can act on a significant amount of data remains elusive and more Research would benefit this area.
- Federated Learning (4.3) also presents a very promising avenue for processing data with privacy guarantees. However, more Research is needed to understand how we can accomplish this without leaking data in the corresponding model by reverse-engineering it.
- Research into designing a framework for choosing the most beneficial method for guaranteeing data privacy would be valuable. Specifically, it would allow organisations to select an appropriate privacy framework easily.

6 Conclusion

This research covered seven important potential methods of reliably handling data that third parties can process in a way that guarantees security and privacy. In our literature review, we reviewed 22 pieces of relevant research. From this body of research, we were able to conclude that to process data with privacy and security guarantees; we must take the context of the data into account–it is not one size fits all. Homomorphic Encryption presents a tantalising holy-grail-type solution to our research questions; however, it is not possible to use it in all situations because it is too inefficient (relevant to our second research question). Given this, we must be cognizant that other methods are available and, depending on their implementation, can provide good privacy and security guarantees if implemented correctly.

The current state of the literature on this subject covers each specific method in good detail. However, more research is needed to combine these methods in a secure framework. For example, it is not obvious how to choose the correct way for people who are not experts in the field. As mentioned in 5.1, further study into this would help guide how people utilise these methods; using the proper method for the right task could have significant positive implications for protecting the data involved.

References

Abiodun, M. K., Awotunde, J. B., Ogundokun, R. O., Adeniyi, E. A. & Arowolo, M. O. (2021), Security and information assurance for iot-based big data, *in* 'Artificial Intelligence for Cyber Security: Methods, Issues and Possible Horizons or Opportunities', Springer, pp. 189–211.

Aldaghri, N., Mahdavifar, H. & Beirami, A. (2021), 'Coded machine unlearning', *IEEE Access* **9**, 88137–88150.

Cong, W. & Mahdavi, M. (2022), 'Privacy matters! efficient graph representation unlearning with data removal guarantee'.

De la Rosa Algarín, A. & Demurjian, S. A. (2014), An approach to facilitate security assurance for information sharing and exchange in big-data applications, *in* 'Emerging trends in ICT security', Elsevier, pp. 65–83.

Dwork, C., McSherry, F., Nissim, K. & Smith, A. (2006), Calibrating noise to sensitivity in private data analysis, *in* S. Halevi & T. Rabin, eds, 'Theory of Cryptography', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 265–284.

Fajar Marta, R. (2016), 'Open government data portal design principles: implementing transparency, privacy, and information quality assurance by design'.

Gharib, M. (2020), Citizens as data donors: Maximizing participation through privacy assurance and behavioral change, *in* 'Data Privacy Management, Cryptocurrencies and Blockchain Technology', Springer, pp. 229–239.

Jiang, J., Han, K., Du, Y., Zhu, G., Wang, Z. & Cui, S. (2022), 'Optimized power control for over-the-air federated averaging with data privacy guarantee', *IEEE Transactions on Vehicular Technology* .

KAVITHA, B. (n.d.), 'A novel and capable scheme assurance data privacy of encryption category'.

Konečný, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T. & Bacon, D. (2016), Federated learning: Strategies for improving communication efficiency, *in* 'NIPS Workshop on Private Multi-Party Machine Learning'.
URL: *https://arxiv.org/abs/1610.05492*

Li, H., Dai, Y. & Lin, X. (2015), Efficient e-health data release with consistency guarantee under differential privacy, *in* '2015 17th International Conference on E-health Networking, Application & Services (HealthCom)', IEEE, pp. 602–608.

Lingbin, D. (2022), 'New regulations on japan's cross-border data flow and china's path: An analysis based on the perspective of data security assurance', *Information and Documentation Services* **43**(1), 52–60.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. y. (2016), 'Communication-efficient learning of deep networks from decentralized data'. **URL:** *https://arxiv.org/abs/1602.05629*

Naehrig, M., Lauter, K. & Vaikuntanathan, V. (2011), Can homomorphic encryption be practical?, *in* 'Proceedings of the 3rd ACM Workshop on Cloud Computing Security Workshop', CCSW '11, Association for Computing Machinery, New York, NY, USA, p. 113–124.
URL: *https://doi.org/10.1145/2046660.2046682*

Naher, N. & Hashem, T. (2019), 'Think ahead: Enabling continuous sharing of location data in real-time with privacy guarantee', *The Computer Journal* **62**(1), 1–19.

Peng, Z., Liang, F. & Mu, L. (2022), 'Big data-based access control system in educational information security assurance', *Wireless Communications and Mobile Computing* **2022**.

Saavedra, S., Llatas, J. & Armas-Aguirre, J. (2020), 'Process mining model to guarantee the privacy of personal data in the healthcare sector'.

Sen, A. & Madria, S. (2018), Data analysis of cloud security alliance's security, trust & assurance registry, *in* 'Proceedings of the 19th International Conference on Distributed Computing and Networking', pp. 1–10.

Sivakami, K. & Umadevi, V. (2022), 'Ersa: enhanced rsa cryptography algorithm to guarantee high security level for data in cloud environment', *International Journal of Computer Aided Engineering and Technology* **16**(2), 170–193.

Studer, T. (2013), 'A universal approach to guarantee data privacy', *Logica universalis* **7**(2), 195–209.

Tamunobarafiri, A., Aghili, S. & Butakov, S. (2017), 'Data security and privacy assurance considerations in cloud computing for health insurance providers', *International Journal of Monitoring and Surveillance Technologies Research (IJMSTR)* **5**(4), 1–22.

Xu, G., Zhang, J. & Wang, L. (2022), An edge computing data privacy-preserving scheme based on blockchain and homomorphic encryption, *in* '2022 International Conference on Blockchain Technology and Information Security (ICBCTIS)', pp. 156–159.

Yang, Q., Liu, Y., Chen, T. & Tong, Y. (2019), 'Federated machine learning: Concept and applications', *ACM Transactions on Intelligent Systems and Technology (TIST)* **10**(2), 1–19.