Federated Learning as an Analytical Framework for Personal Data Management – a proposition paper

Maciej Zuziak [1] and Salvatore Rinzivillo [2]

[1] *Consiglio Nazionale delle Ricerche, Via Giuseppe Moruzzi, 1, Pisa, Italy*
[2] *Consiglio Nazionale delle Ricerche, Via Giuseppe Moruzzi, 1, Pisa, Italy*

Abstract

Data minimisation and storage limitation are two principles incorporated in the GDPR aimed to increase personal data subjects' control over their own data and put restrictions on the amount of information that may be extracted from them in the data mining process. Implementation of those two principles has always been a challenging task, as their interpretation is discretional and current legislative measures may not necessarily protect data subjects adequately. In this paper, we introduce the concept of distributed learning as a viable tool for implementing data minimisation and storage limitation principles and argue that perhaps it could be appropriate to consider a branch of distributed learning, namely the concept of federated learning, as an analytical measure for guaranteeing data limitation and minimisation. To further support this thesis, we discuss how Federated Learning may be used in geospatial data analysis while the final outcomes of the experiments are yet to be published.

Keywords

Federated Learning, General Data Protection Regulation, Data Management

1. Introduction

This article lays down a proposal for the extended study of Federated Learning as a viable tool for implementing the data minimisation and storage limitation principles incorporated in Article 5 of the General Data Protection Regulation of 2016. In this paper, we argue several theses:

1. Data minimisation and storage limitation principles are two discretionary measures that rely heavily on the technological framework applied for data collection & data processing;
2. Distributed learning in general, and Federated Learning especially, are viable tools for the implementation of the data minimisation & storage limitation principles, thus giving data subjects increased control over their personal data;
3. Distributed Learning, especially federated learning (which we consider a subbranch of the former), should be further assessed in the context of minimisation-aware systems for digital services-oriented datamining.
4. It could be beneficial to the user awareness to implement architecture based on the idea of distributed learning and provide them with a backchannel about the training process and the use of their data. That way, we can not only raise awareness about the value of personal data but also possibly mitigate the adverse effects of using the users' devices to handle part of the model training.

In connection to thesis no. 2, multiple authors have published on the topic of privacy-preserving federated learning systems [1-8]. We propose putting federated learning in the context of data minimisation and storage limitation principles of the General Data Protection Regulation and the upcoming legal framework for data protection and AI regulation. The data protection by design as

1st International Workshop on Imagining the AI Landscape After the AI Act, May, 2022, Amsterdam, Netherlands;
EMAIL: maciejkrzysztof.zuziak@isti.cnr.it (A. 1); rinzivillo@isti.cnr.it (A. 2);
ORCID: 0000-0003-4297-4973 (A. 1); 0000-0003-4404-4147 (A. 2);

presented by S. Rosello et al. [9] in 2021 may be a viable solution to many complicated issues arising from the need to be compliant with an increasing amount of regulations, and the federated learning as a tool for that is gaining increased attention [10]. We want to contribute to the ongoing discussion by putting the distributed learning systems in the context of two specific principles, namely, the principle of data minimisation & storage limitation. In connection to thesis no. 3 and no. 4, we present here an outline of a possible experiment that may be carried out to assess the performance of the proposed measures. We also briefly argue why the low-user engagement methods such as federated learning may be the best choice for the latter.

2. Data Minimisation and Storage Limitation Principles in the GDPR

Data minimisation and data limitation are two terms that belong to the broader set of principles that refers to data quality. Together with the 1) lawfulness, fairness and transparency, 2) purpose limitation, 3) accuracy, and 4) integrity and confidentiality, they are shaping the way personal data should be controlled, processed and discarded throughout the whole knowledge discovery cycle [11]. According to the Article 5 of the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [12]:

> *Article 5*
>
> *Principles relating to processing of personal data*
>
> *Personal data shall be:*
>
> *(c) adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation');*
>
> *(e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed; personal data may be stored for longer periods insofar as the personal data will be processed solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) subject to implementation of the appropriate technical and organisational measures required by this Regulation in order to safeguard the rights and freedoms of the data subject ('storage limitation');*
>
> *2. The controller shall be responsible for, and be able to demonstrate compliance with, paragraph 1 ('accountability').*[2]

The data minimisation principle is not a new measure, as it was already incorporated in the Article 3(1)(c) of Regulation (Ec) No 45/2001 Of The European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data [13], and the wording of that principles was almost the same as the one incorporated in the GDPR. It primarily concerns which type (and what amount) of data is targeted for extraction, while the storage limitation generally specifies how long and under what condition the personal data may be stored.

In line with the data minimisation and storage limitation is the principle of purpose limitation, according to which personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes [14]–[16]. Although essential

[2] The following principles are also elaborated on in recital no. 39 of the Regulation (EU) 2016/679.

in its nature, it must be highlighted that it is generally described as different and independent from the data minimisation and storage limitation principles that are described in this article.

According to the Information Commissioner's Office (ICO)[3], the principle of data minimisation requires that the processed personal *data should be sufficient to fulfil the stated purpose [adequate] properly, have a rational link to that purpose [relevant] and was not held in an amount that exceeds what is strictly necessary for that purpose [limitation to what is necessary]* [17].

This matter could be analysed from a solely legal or technological side, depending on the chosen aspect. Providing an example, the adequateness and relevance could be seen as a primarily legal issue connected to the stated purpose of the personal data processing and establishing a rational link between processing and that purpose, while the aspect of data storage limitation is more technical measures, that relies heavily on how we store, preprocess and analyse the collected data. It can be argued that it is almost impossible to implement regulatory measures that could possibly guarantee that the amount of collected data is adequate, as the data subjects have minimal insight into how much of their data is collected and what types of data are extracted. Once the raw data is transferred beyond the users' device into the Date Warehouses and Analytical Databases [18], it is almost impossible to guarantee any level of data storage limitation principle, as the users themselves would have to control and oversee the whole data lifecycle, with multiple inquires and requests issued towards the data controller and data processor – to ensure, that they do comply with the biding regulations.

The necessity to enforce better privacy standards while seeking beyond purely legal remedies has inspired some researchers to reconsider their approach to data protection compliance. In 2019 the European Union Agency for Cybersecurity (ENISA) had published Recommendations on shaping technology according to GDPR provisions - Exploring the notion of data protection by default [19], while in the upcoming years, the concept of Data protection by design has caught the attention of some authors [9]. It is worth noting that federated learning was one of those technologies that were distinguished most commonly in the context of GDPR-compliant technology [10], [20], while some of the publications pointed out also the risk associated with that method of distributed learning [21][4].

One more thing must be highlighted in the context of the data minimisation and storage limitation principles. While we assume that protecting their personal data is in the best interest of the data subjects, it is the sole responsibility of the data processors and controllers to comply with any possible obligations arising from those principles. At this point, we would also like to briefly elaborate on the choice of relevant technology, as this stays in particular connection with the mentioned above. Over the recent years, many proposals regarding decentralised learning and analytics have been raised, and some of them, such as Personal Data Stores, attracted widespread attention [22], [23]. In our opinion, in the current data economy paradigm, the architecture for decentralised learning should be unintrusive, secure, and effortless from the user's point of view. It is an essential notion that forcing users to opt-out for more time and knowledge consuming solutions could be shifting the burden of data minimisation and storage limitation principles towards the data subject – which would be unacceptable from the axiological point of view. While the shift towards a more decentralised ecosystem may result in the adoption of more user-centric methods (such as PDS/PBS), we present here a "one-step-approach" where the data subjects gain more control over their data without directly shifting the paradigm of current data processing right and obligations under the European legislation. In the context of the unintrusive-secure-effortless paradigm, we firmly believe that distributed learning is a viable choice.

[3] In this article we are referring to both, guidelines and explanations of the European Data Protection Supervisor (EDPS) as well as those provided by the Information Commissioner's Office (ICO). If they are any discrepancies between GDPR and UK GDPR we are raising and explaining them in advance. We also use those references to put the overviewed principles in a slightly broadly concept.

[4] Due to the conceptual nature of this article, we will not go into much detail regarding the privacy issues that may be found in FL. However, it is worth highlighting, that FL-based systems may be more prone to some types of attacks that can infringe the participant's privacy. For a short synapsis on that issues see: Inpher: *The Privacy Risk Right Under Our Nose in Federated Learning (Part 1)*, 23 February 2021; and for more detailed analysis see especially: Nguyen Truong et al.: *Privacy Preservation in Federated Learning: An insightful survey from the GDPR Perspective*, 18 March 2021; or L. Melis et al.: *Exploiting Unintended Feature Leakage in Collaborative Learning*, 1 November 2018.

3. Distributed Learning and Federated Learning

Federated learning originated from the idea of training the model on the dataset distributed over a wide area. Generally, the federated optimisation was proposed to handle the data that is:

- Massively Distributed (data points are stored across a vast number of nodes);
- Not Independent and Identically Distributed (data on each node may be drawn from different distributions);
- Unbalanced (Different nodes may vary by many orders of magnitude in the number of training examples they hold) [24].

The experiment performed by J. Konecny et al. in 2015 was conducted under the following assumptions:

- The data stored on multiple nodes may be privacy-sensitive, so the key objective should be to train the model on local nodes but not to transfer the data to one central node;
- Some of the nodes connected to the network (or all of them) may not necessarily have stable access to the network, so in real-life circumstances, it will be crucial to minimise the round of communications;
- The data is not independent and identically distributed [24]

A few years later, after the proposed experiment, Federated Learning has gained popularity amongst the Data Science community, with much work centered on privacy-related issues. According to the current state-of-the-art, Federated Learning could be defined *as a machine learning setting where many clients (e.g. mobile devices or whole organisations) collaboratively train a model under the orchestration of a central server (e.g. service provider) while keeping the training data decentralised.*

Formally, the problem was defined as a minimalisation of the objective function:
$F(x) = E_{i\sim P}[F_i(x)]$, where $F_i(x) = E_{\varepsilon\sim D_i}[f_i(x, \varepsilon)]$ where:

- $x \in \mathbb{R}^d$ represents the parameters for the global model;
- $F_i: \mathbb{R}^d \to \mathbb{R}$ denotes the local objective function at client *i*;
- P denotes the distribution of the population of clients [25].[5]

4. Distributed Learning and Federated Learning

In the previous sections, we have placed federated learning in the context of data minimisation and storage limitation principles of GDPR. In this chapter, we want to propose a specific application scenario that could be carried out regarding the assessed framework. Before overviewing the proposition of the experiment, we would like to formulate a few key marks on the characteristics of the system that should be favorable to the implementation of the data minimisation & storage limitation principles. Namely:

- The system should minimise the amount of raw data transferred beyond users' devices (beyond the realm of the clients). The system should also explicitly allow users to choose whether they want to participate in the training loop while clearly indicating that it may be beneficial but not necessary to participate in the model's training.
- If the users consent to participation in the training, the system should explicitly declare that they can withdraw from it at any time they deem appropriate.
- Users should have a range of information on how the system is trained, what type of data is processed on their local devices, and the hyperparameters of the general model that are being updated in the current (or upcoming) training iterations.

[5] This definition presented in Jianyu Wang et al. - A Field Guide to Federated Optimization is based on the overview of the existing work on the problem – it may be worthwhile noting, that different authors approach the same problem quite differently when putting it in the formal manner.

In accordance with that, we want to realise an analytical framework where collaborative data computation is possible on spatio-temporal data. In particular, we focus on the analysis of individual-based contributed GPS data collected during the movement of personal vehicles.

The analytical framework will have several capabilities: computation of aggregation-based indicators (i.e. the radius of gyration, CO2 emission estimation); collective patterns (i.e. aggregated traffic flows and models for description and prediction; profiling of user behaviour; sustainability compatible behaviour estimation); global models (i.e. temporal footprint of traffic evolution, learning of predictive models for traffic forecasting, etc.). These three dimensions need to be instantiated into a distributed/federated setting, where several computational challenges need to be addressed:

- the individual-based choice for participation, managing a range of levels for the collaboration, ranging from full data disclosure to avoid any type of participation, passing through different levels of data perturbation and obfuscation;
- implementing a one-against-all framework, where the client may share only a local learned model that can be compared with the global one, to give the user feedback and raise self-awareness;
- designing mechanisms for allowing opt-out of a client, eventually refreshing the existing models already learned.

Apart from performing the experiment, many mixed technological and legal issues arise in connection with the distributed learning environment (and federated learning especially). Those problems were not yet thoroughly researched, and they may possess a tremendous challenge when discussing the distributed data processing environment. A few exemplary questions in that regard:

- What are the legal consequences of opting out by a user who participated in the original training of the model?
- If the user has opted out of the model – what are the measures that can be taken to rectify the model and possibly delete any traces of personal data from that model?
- How can we communicate and explain the training process to the users of edge devices? How can we avoid disconnecting them from the network or opting out of the training?

The successful implementation of the said experiment will allow us to contribute to the discussion on federated learning as well as further explore the concept of using distributed learning as a tool for the implementation of the data minimisation and storage limitation principles.

Thus far, many experiments have been conducted, and the federated learning was tested in different settings and circumstances. The federated learning was used to deliver experiments on, among other things: recommendation systems [1], [2], [26], meta-learning systems for fraudulent credit card detection [27] or learning systems for mobile keyboard prediction [28].

One advantage of working on that particular technology is the wide range of tools that may be used to perform simulation of a decentralised environment – allowing researchers to focus on a particular problem rather than on implementing the technological framework from scratch – Tensor Flow Federated (TFF) [29], FedML [30], PySyft [31], PyVertical [32], Leaf [33] are just a few examples of tools that can be used to work with the concept of distributed learning while conducting experiments.

5. Closing Remarks

Unquestionably, the decentralisation of data collection and processing is a promising concept that possibly can shift the paradigm toward a more equitable and engaging future of collaborative data science. The idea of distributed learning was born from a strict necessity – with the growing amount of data that must be processed, it is even harder to rely on centralised methods that would require constant expansion of the storage (and computational) resources. However, the major challenge may not necessarily arise from the optimisation problems but from reaching a specific level of compliance with the current legislation and sustaining a high level of collaboration with the users of edge devices.

We presented our view on the development of that technology, where the strong emphasis is placed on the principle of data minimisation and storage limitations. It is crucial to present users with a clear and well-defined trade-off – without that, the cost of sacrificing (some of them) their devices'

computational power may deter them from such decentralised frameworks. It must be stated that here we have taken into account primarily only one approach to distributed learning, namely, federated learning. In our opinion, it could suffice the unintrusive-secure-effortless paradigm that we shared earlier. Notwithstanding any other benefits coming from such an approach, much research may be conducted to shape that technology fully compliant and user-friendly.

6. Acknowledgements

The research is part of the Legality Attentive Data Scientist project. The Project has received funding from the European Union's Horizon Marie Skłodowska-Curie Actions (MSCA) 2020 Innovative Training Networks (ITN). Grant Agreement ID: 956562

This Word template was created by Aleksandr Ometov, TAU, Finland. The template is made available under a Creative Commons License Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

7. References

[1] Tao Qi, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, and Xing Xie, Privacy-Preserving News Recommendation Model Learning, Findings of the Association for Computational Linguistics: EMNLP 2020, Online, Nov. 2020, pp. 1423–1432. doi: 10.18653/v1/2020.findings-emnlp.128.

[2] Yangjie Qin, Ming Li, and Jia Zhu, Privacy-preserving federated learning framework in multi-media courses recommendation, Wirel. Netw., Jan. 2022, doi: 10.1007/s11276-021-02854-1.

[3] S. R. Pokhrel and J. Choi, Federated Learning With Blockchain for Autonomous Vehicles: Analysis and Design Challenges, IEEE Trans. Commun., vol. 68, no. 8, pp. 4734–4746, Aug. 2020, doi: 10.1109/TCOMM.2020.2990686.

[4] M. Ammad-ud-din et al., Federated Collaborative Filtering for Privacy-Preserving Personalized Recommendation System, ArXiv190109888 Cs Stat, Jan. 2019, Accessed: 10 February, 2022. [Online]. Available: http://arxiv.org/abs/1901.09888

[5] Waqar Ali, Rajesh Kumar, Zhiyi Deng, Yanshong Wang, and Jie Shao, A Federated Learning Approach for Privacy Protection in Context-Aware Recommender Systems, Comput. J., vol. 64, no. 7, pp. 1016–1027, Jul. 2021, doi: 10.1093/comjnl/bxab025.

[6] Yuzheng Li, Chuan Chen, Nan Liu, Huawei Huang, Zibin Zheng, and Qiang Yan, A Blockchain-Based Decentralised Federated Learning Framework with Committee Consensus, IEEE Netw., vol. 35, no. 1, pp. 234–241, Jan. 2021, doi: 10.1109/MNET.011.2000263.

[7] P. Kairouz, Ziyu Liu, and T. Steinke, The Distributed Discrete Gaussian Mechanism for Federated Learning with Secure Aggregation, in Proceedings of the 38th International Conference on Machine Learning, Jul. 2021, pp. 5201–5212. Accessed: 03 March, 2022. [Online]. Available: https://proceedings.mlr.press/v139/kairouz21a.html

[8] K. Bonawitz, P. Kairouz, B. McMahan, D. Ramage, and Google, Federated Learning and Privacy -Building privacy-preserving systems for machine learning and data science on decentralised data, ACM Queue, vol. 19, no. 5, Nov. 2021, [Online]. Available: https://queue.acm.org/detail.cfm?id=3501293

[9] R. Stephanie, Data protection by design in AI? The case of federated learning, Social Science Research Network, Rochester, NY, SSRN Scholarly Paper 3879613, May 2021. Accessed: 11 April, 2022. [Online]. Available: https://papers.ssrn.com/abstract=3879613

[10] N. Truong, Kai Sun, Siyao Wang, F. Guitton, and Yike Guo, Privacy preservation in federated learning: An insightful survey from the GDPR perspective, Comput. Secur., vol. 110, p. 102402, Nov. 2021, doi: 10.1016/j.cose.2021.102402.

[11] European Data Protection Supervisor, Data Protection Glossary, An official website of the European Union, 2022. https://edps.europa.eu/data-protection/data-protection/glossary/d_en (accessed 07 April, 2022).

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

Regulation) (Text with EEA relevance). 2016. Accessed: 08 February, 2022. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj
[13] Regulation (EC) No 45/2001 of the European Parliament and of the Council of 18 December 2000 on the protection of individuals with regard to the processing of personal data by the Community institutions and bodies and on the free movement of such data, vol. 008. 2000. Accessed: 07 April, 2022. [Online]. Available: http://data.europa.eu/eli/reg/2001/45/oj/eng
[14] M. von Grafenstein, The function of the principle of purpose limitation in light of Article 8 ECFR and further fundamental rights, in The Principle of Purpose Limitation in Data Protection Laws, 1st ed., Nomos Verlagsgesellschaft mbH, 2018, pp. 109–596. Accessed: 07 April, 2022. [Online]. Available: https://www.jstor.org/stable/j.ctv941v5w.5
[15] Information Commissioner's Office, GDPR: Principle (b): Purpose limitation', Information Commissioner's Office Website, 17 January, 2022. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/purpose-limitation/ (accessed 07 April, 2022).
[16] Dataguise, Inc, GDPR Purpose Limitation Principle - GDPR Knowledge Center, Dataguise Website, 2022. https://www.dataguise.com/gdpr-knowledge-center/purpose-limitation-principle/ (accessed 07 April, 2022).
[17] Information Commissioner's Office, GDPR Principle (c): Data minimisation', Information Commissioner's Office Website, 11 February, 2021. https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/principles/data-minimisation/ (accessed 07 April, 2022).
[18] S. Suh, Practical Applications of Data Mining. Jones & Bartlett Publishers, 2012.
[19] ENISA, Recommendations on shaping technology according to GDPR provisions - Exploring the notion of data protection by default, ENISA website. https://www.enisa.europa.eu/publications/recommendations-on-shaping-technology-according-to-gdpr-provisions-part-2 (accessed 07 April, 2022).
[20] Musketeer, Benefits and challenges of federated learning under the GDPR, Medium, 04 October, 2021. https://h2020musketeer.medium.com/benefits-and-challenges-of-federated-learning-under-the-gdpr-15c89ff76d9c (accessed 11 April, 2022).
[21] Impher, The Privacy Risk Right Under Our Nose in Federated Learning (Part 1)' Inpher Website, 23 February, 2021. https://inpher.io/journal-blog/the-privacy-risk-right-under-our-nose-in-federated-learning-part-1/ (accessed 11 April, 2022).
[22] Y.-A. de Montjoye, E. Shmueli, Samuel S. Wang, and A. S. Pentland, openPDS: Protecting the Privacy of Metadata through SafeAnswers, PLOS ONE, vol. 9, no. 7, p. e98790, Jul. 2014, doi: 10.1371/journal.pone.0098790.
[23] M. Vescovi, C. Perentis, C. Leonardi, B. Lepri, and C. Moiso, My data store: toward user awareness and control on personal data', in Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, New York, NY, USA, Sep. 2014, pp. 179–182. doi: 10.1145/2638728.2638745.
[24] J. Konečný, B. McMahan, and D. Ramage, Federated Optimization: Distributed Optimization Beyond the Datacenter, ArXiv151103575 Cs Math, Nov. 2015, Accessed: 11 February, 2022. [Online]. Available: http://arxiv.org/abs/1511.03575
[25] Jianyu Wang et al., A Field Guide to Federated Optimization, ArXiv210706917 Cs, Jul. 2021, Accessed: 12 May, 2022. [Online]. Available: http://arxiv.org/abs/2107.06917
[26] Feng Liang, Weike Pan, and Zhong Ming, FedRec++: Lossless Federated Recommendation with Explicit Feedback, Proc. AAAI Conf. Artif. Intell., vol. 35, no. 5, Art. no. 5, May 2021.
[27] Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang, Federated Meta-Learning for Fraudulent Credit Card Detection, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan, Jul. 2020, pp. 4654–4660. doi: 10.24963/ijcai.2020/642.
[28] A. Hard et al., Federated Learning for Mobile Keyboard Prediction, ArXiv181103604 Cs, Feb. 2019, Accessed: 10 February, 2022. [Online]. Available: http://arxiv.org/abs/1811.03604
[29] Google Brain Team, TensorFlow Federated, tensorflow.org, 2022. https://www.tensorflow.org/federated (accessed 11 February, 2022).
[30] FedAI, FedML - The Federated Learning/Analytics and Edge AI Platform, FedML Official Website, 2022. https://fedml.ai/ (accessed 11 February, 2022).

[31] A. Ziller et al., PySyft: A Library for Easy Federated Learning, in Federated Learning Systems: Towards Next-Generation AI, M. H. ur Rehman and M. M. Gaber, Eds. Cham: Springer International Publishing, 2021, pp. 111–139. doi: 10.1007/978-3-030-70604-3_5.
[32] D. Romanini et al., PyVertical: A Vertical Federated Learning Framework for Multi-headed SplitNN, ArXiv210400489 Cs, Apr. 2021, Accessed: 11 February, 2022. [Online]. Available: http://arxiv.org/abs/2104.00489
[33] S. Caldas et al., 'LEAF: A Benchmark for Federated Settings', ArXiv181201097 Cs Stat, Dec. 2019, Accessed: 12 April, 2022. [Online]. Available: http://arxiv.org/abs/1812.01097