The Data Minimization Principle in Machine Learning

Prakhar Ganesh [1,2] Cuong Tran [3,4] Reza Shokri [5] Ferdinando Fioretto [3]

[1]McGill University [2]Mila – Quebec AI Institute [3]University of Virginia [4]Dyania Health [5]National University of Singapore

SCAN ME

The Data Minimization Principle

General Data Protection Regulation (GDPR), Europe	gdpr-info.eu/
Article 4(1): *"personal data" means any information relating to an identified or identifiable natural person ("data subject") [...];* **Article 5(1)(b):** *Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes;* **Article 5(1)(c):** *Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.*	
California Privacy Rights Act (CPRA), USA	cppa.ca.gov/
Section 1798.100 (a)(1) & (a)(2): *[...] A business shall not collect additional categories of (sensitive) personal information or use (sensitive) personal information collected for additional purposes that are incompatible with the disclosed purpose for which the (sensitive) personal information was collected without providing the consumer with notice consistent with this section.* **Section 1798.140 (v)(1):** *"Personal information" means information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household.*	
General Personal Data Protection Law (LGPD), Brazil	lgpd-brazil.info/
Article 5(I): *personal data: information regarding an identified or identifiable natural person;* **Article 6:** *Activities of processing of personal data shall be subject to the following principles,* **I:** *processing done for legitimate, specific and explicit purposes of which the data subject is informed, with no possibility of subsequent processing that is incompatible with these purposes;* **II:** *compatibility of the processing with the purposes communicated to the data subject, in accordance with the context of the processing;* **III:** *limitation of the processing to the minimum necessary to achieve its purposes, covering data that are relevant, proportional and non-excessive in relation to purposes of the data processing [...];*	
Protection of Personal Information Act (POPIA), South Africa	popia.co.za/
Section 1: *"personal information" means information relating to an identifiable, living, natural person, [...]* **Section 10:** *Personal information may only be processed if, given the purpose of processing, it is adequate, relevant and not excessive* **Section 13(1):** *Personal information must be collected for a specific, explicitly defined and lawful purpose [...]*	
Consumer Data Rights (CDR), Australia	www.legislation.gov.au/Details/F2023C00735
Rule 1.8(a): *An accredited person complies with the data minimisation principle if when making a consumer data request on behalf of a CDR consumer, it does not seek to collect: (i) more CDR data than is reasonably needed; [...]* **Rule 4.11(1)(a):** *When asking a CDR consumer to give a consent, an accredited person must allow the CDR consumer to choose the types of CDR data to which the consent will apply by enabling the CDR consumer to actively select or otherwise clearly indicate: (i)..; and (ii) in the case of a use consent— the specific uses of collected data to which they are consenting; [...]*	
Personal Information Protection Act (PIPA), South Korea	www.pipc.go.kr/eng/index.do
Article 2(1): *The term "personal information" means any of the following information relating to a living individual: (a) Information that identifies a particular individual [...]; (b) Information which, even if it by itself does not identify a particular individual, may be easily combined with other information to identify a particular individual [...];* **Article 3(1):** *The personal information controller shall specify explicitly the purposes for which personal information is processed; and shall collect personal information lawfully and fairly to the minimum extent necessary for such purposes.*	

Purpose limitation

Data collectors must define a clear, legal objective before data collection and use the data solely for that objective. In ML, this *purpose* translates to collecting data solely for training models to achieve optimal performance, for a clearly defined performance metric.

Data relevance

Data minimization aims to remove data that does not serve the purpose. In ML contexts, this translates to retaining only data that contributes to the performance of the model. We interpret this as an optimization, minimizing data while maintaining performance.

Data Privacy

Data minimization aims to avoid the collection of any unnecessary information that could compromise an individual's privacy, thus creating an implicit expectation of privacy from minimization.

How do we formalize data minimization in ML?

Data Minimization as Optimization

The goal of *data minimization* can be modelled as minimizing the size of the dataset D by removing entries from the features $\boldsymbol{x}_i$ (data relevance) while also retaining similar performance as it would have on the complete dataset (purpose).

$$\underset{\boldsymbol{B}\in\{\perp,1\}^{n\times p}}{\text{Minimize}} \quad \|\boldsymbol{B}\|_1 \qquad s.t.: \quad J(\hat{\theta};\mathbf{X},\mathbf{Y}) - J(\check{\theta};\mathbf{X},\mathbf{Y}) \leq \alpha$$

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} \ell\left(f_\theta(\boldsymbol{x}_i \odot \boldsymbol{B}_i), y_i\right)$$

EU AI Act: Recital 69

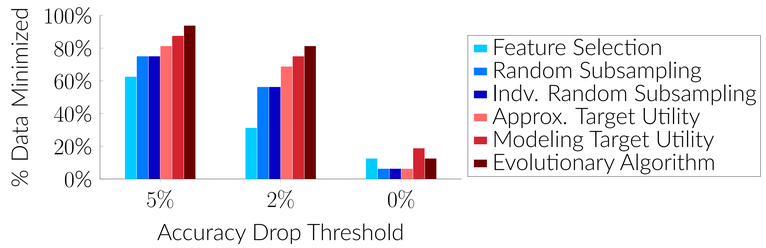
The right to privacy and to protection of personal data must be guaranteed throughout the entire lifecycle of the AI system. In this regard, the principles of data minimisation and data protection by design and by default, as set out in Union data protection law, are applicable when personal data are processed.

Do current data minimization requirements in regulations genuinely meet privacy expectations? **Unfortunately, No!**

Empirical Analysis

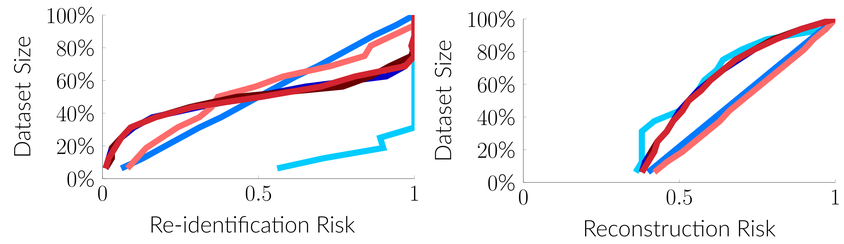
Utility

A substantial amount of data can often be removed without sacrificing utility, suggesting that much of the collected data is superfluous in the datasets analyzed.



Privacy

There exists a fundamental misalignment between the goals of data minimization and actual privacy.



Open Challenges

- How to more efficiently solve the optimization challenges that come with minimization?
- Can other formalisms be created to better operationalize data minimization that incorporates nuances like active minimization during data collection, instead of data removal?
- What ethical and fairness considerations of data minimization are needed to make sure we don't leave the minority more vulnerable to privacy risks?

prakhar.ganesh@mila.quebec