Examining Data Compartmentalization for AI Governance

Anonymous Authors[1]

1. Introduction

Typical ML pipelines do not exploit any structure or hierarchy of pre-training data – all sources are mixed and consumed by training algorithms that are agnostic to their structure. As a result, information from all data sources is fused in the model parameters. This poses a challenge for AI governance, as legal and policy restrictions may not apply uniformly to the entire pre-training data corpus.

Because data usage constraints can be time-, place-, and context-dependent, there is a need for training and/or serving models in a way that is aware of these dependencies. A conservative approach to address any required change to the pre-training data corpus is to retrain the model, even if the majority of data is unchanged. While foolproof, retraining billion-parameter models from scratch is costly, inefficient, and impractical. This has made relevant ML techniques that leverage *compartmentalized data*: data which is partitioned into semantically meaningful groups.

Principle characteristics of data compartmentalization include granularity (fine or coarse partitioning), uniformity (distribution of group sizes), interrelation (degree to which groups relate), rigidity (fixed or variable partitioning), fluctuation (frequency of any group's inclusion or exclusion), exactness (strict separability of groups), and temporality (known *a priori* or *posterior* to training). Each group might refer to the data owner (e.g., an individual or an institution), to a particular domain, or (more generally) to consistent usage constraints (e.g., as dictated by access policies).

We unify a number of existing techniques at the modeling, algorithmic, and inference levels under the framework of enabling data compartmentalization. Despite their varied motivations and settings, all offer mechanisms for providing finer-grained control over subsets of data. We examine their effectiveness, practicality, and relevance to governance, and present open questions to better align policy motivations with technical approaches and inform future work.

2. Strategies for Data Compartmentalization

Various existing techniques across the ML pipeline can be seen as facilitating the use of compartmentalized data. These strategies may be help tease apart model dependencies from data dependencies to address AI governance concerns.

Model architectures. Model architectures that take into account data compartmentalization tend to be *modular*, that is, composed of specialized sub-networks, each responsible for a specific subtask or functionality; see (Pfeiffer et al., 2024) for a survey. These modules can be trained, fine-tuned, or even swapped out independently without affecting the entire model. A simple example is a flat Mixture of Experts architecture, where each expert is trained on a different group of data (Jacobs et al., 1991).

Training algorithms. *Federated learning* (FL) (McMahan et al., 2017) limits data sharing by training across siloed data in a distributed manner. Prototypical FL algorithms bake information across clients into a shared model parametrically through iterative averaging (Reddi et al., 2020), which is not suitable for traceability or exclusion of some data source. Frequently used with FL, *group-level differential privacy* (DP) (Dwork et al., 2006) extends DP to *groups* of examples, where a grouping refers to all examples attributed to an individual, institution, domain, or source. By operating on compartmentalized data, group-level DP bounds the influence of any group on the model, treating all data groups as sensitive. By contrast, *machine unlearning* (MU) removes (the influence of) a specified subset of training data (the "forget set") from models (Nguyen et al., 2022).

Retrieval and inference. Non-parametric access of data sources through *retrieval* allows for maintaining full separability of those sources from model weights. The approach presented in SILO (Min et al., 2023) provides such an example. Recently, there has been work advocating for retrieval augmentation in FL (Muhamed et al., 2024), where clients maintain private data stores accessed only at inference. The merging of FL's ownership-based data partitioning with retrieval yields an approach for owner-based selection of data sources divorced from shared model weights.

3. Opportunities for AI Governance

Growing interest in better controlling large models has spurred research and led to voluntary commitments and nascent regulatory frameworks (Bommasani et al., 2022; Shevlane et al., 2023). Some motivations stem from practical constraints on data access (e.g., regulatory and licensing compliance), while others relate to risks of AI (e.g., bias, harms, and privacy). By strategically partitioning and managing data within AI systems, practitioners may be better equipped to align their models with overarching principles of responsible development and deployment (UK Department for Science, Innovation and Technology, 2024).

Enhancing traceability of model outputs. Attributing model outputs to the sources that were most influential is needed for interpretability, grounding, factuality, and mitigating harms. Data compartmentalization can make it easier to identify, isolate, and address subsets of the data that are found to be erroneous or problematic. When paired with influence functions (Koh et al., 2019), data compartmentalization may provide a path towards credit assignment.

Allowing efficient data deletion. When subsets of data have been identified as problematic (either due to explicit labeling, or as a result of measuring influence), one may want to remove this data from the model. Approaches that enable data compartmentalization may support more efficient deletion from trained models, compared to naively retraining a monolithic model from scratch.

Enabling domain-specific models for regulatory compliance. Compartmentalizing sensitive domain-specific data may facilitate compliance with regulations on model use in particular contexts. To comply with securities regulations, a model trained on financial data could exclude insider information when used for investment recommendations. This would prevent the misuse of privileged information while still allowing for the use of other relevant data for analysis.

Facilitating compliance with licensing terms. Maintaining data source separability will allow for using each source according to its associated license, rather than using the most restrictive terms among all data sources in the mix. Though efforts to attribute licenses to data are underway (Longpre et al., 2023), their feasibility is uncertain given the evolving nature of licensing terms and data interdependencies.

Fostering collaborative model development. Organizations could contribute to a joint model without revealing their data, by training separate modules that are combined only at inference time based on access policies. This could enable extensible models trained on data from multiple organizations in a privacy-preserving way (Rieke et al., 2020).

4. Hypotheses and Open Questions

We have reviewed several strategies that leverage data compartmentalization and offer finer-grained control of data. None of them is a panacea; they have different strengths and weaknesses that make them suitable to different settings.

Formulating hypotheses on suitability. Several considerations influence the suitability of each data compartmentalization strategy for different applications, including tradeoffs in computational complexity, model performance, and application-specific priorities. Here, we formulate hypotheses based on key characteristics we identify in Section 1 that we invite research to investigate: 1) Modularity may be most applicable to settings in which data compartmentalization is coarse, fixed, infrequently fluctuating, exact, and known *a priori* to training. 2) DP and FL are also well-suited to address risks associated with fixed groupings known *a priori*, though at a finer granularity (e.g., user-level) without strict separability and tolerant of frequent inclusion/exclusion. DP in particular uniformly bounds the risk of all groups at the expense of utility. 3) By contrast, MU and retrieval may be better poised to address specific and limited data groupings that are not fixed, frequently fluctuate, and may be defined *posterior* to training. Retrieval offers strict separability of groups, while MU may yield approximate separation.

A number of open questions remain surrounding the use of these strategies to address the needs of AI governance.

Technical considerations.

- *Suitability*: What is a robust set of principles to inform the choice of strategy for a particular application? Research is needed to investigate the hypotheses we make, and compile criteria for assessing the relevance of each strategy.

- *Composability*: How can these techniques be effectively combined to achieve multiple goals simultaneously? For instance, can FL be used in conjunction with MU to remove data from specific clients while preserving the model?

- *Evaluation*: How can we rigorously evaluate the effectiveness of these techniques, particularly for MU, where defining and measuring "forgetting" is a challenge?

Legal and policy alignment.

- *Targeting the right intervention*: Given a specific policy goal (e.g., mitigating bias), what part of the ML pipeline should be targeted for data compartmentalization? Are methods that only process outputs of models sufficient?

- *Metrics of success*: What are the appropriate metrics for measuring the success of data compartmentalization in achieving legal and ethical objectives? How can we balance these metrics with traditional model performance metrics?

- *Considering alternatives*: How do data compartmentalization techniques compare to alternative strategies (e.g., careful data curation, output filtering, representation engineering) in terms of optimality, efficiency, and effectiveness across different AI governance challenges and contexts?

Outlook. As the field of AI continues to evolve, so too will the legal and ethical landscape surrounding data usage. The above strategies provide a flexible framework for addressing these evolving needs. By engaging with open questions through interdisciplinary dialogue, we pave the way for the development of responsible and compliant AI systems.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the opportunities and risks of foundation models, 2022.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proc. of the Third Conf. on Theory of Cryptography (TCC)*, pp. 265–284, 2006.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Koh, P. W. W., Ang, K.-S., Teo, H., and Liang, P. S. On the accuracy of influence functions for measuring group effects. *Advances in neural information processing systems*, 32, 2019.

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., Wu, X., Shippole, E., Bollacker, K., Wu, T., Villa, L., Pentland, S., and Hooker, S. The data provenance initiative: A large scale audit of dataset licensing attribution in ai, 2023.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Min, S., Gururangan, S., Wallace, E., Hajishirzi, H., Smith, N. A., and Zettlemoyer, L. Silo language models: Isolating legal risk in a nonparametric datastore, 2023.

Muhamed, A., Thaker, P., Diab, M. T., and Smith, V. Cache me if you can: The case for retrieval augmentation in federated learning. In *Privacy Regulation and Protection in Machine Learning*, 2024. URL https://openreview.net/forum?id=MKd1SkDbbz.

Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Pfeiffer, J., Ruder, S., Vulić, I., and Ponti, E. M. Modular deep learning, 2024.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. Adaptive federated optimization. In *International Conference on Learning Representations*, 2020.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., and Dafoe, A. Model evaluation for extreme risks, 2023.

UK Department for Science, Innovation and Technology. Frontier AI Safety Commitments. *AI Seoul Summit*, 2024.