# Building Legal Datasets

**Jerrold Soh Tsin Howe**
Centre for Computational Law, Yong Pung How School of Law,
Singapore Management University
jerroldsoh@smu.edu.sg

## Abstract

Data-centric AI calls for better, not just bigger, datasets. As data protection laws with extra-territorial reach proliferate worldwide, ensuring that datasets are *legal* is an increasingly crucial yet overlooked component of "better". To help dataset builders become more willing and able to navigate this complex legal space, this paper reviews key legal obligations surrounding ML datasets, examines the practical impact of data laws on ML pipelines, and offers a framework for building legal datasets.

## 1 Introduction

Data-centric AI is about making better datasets. But what does "better" mean? Conventionally it has meant cheaper. That is, easier to crowdsource [14], generate [4], augment [5], or broadly to collect [22]. *Bigger* is also often better, as the rise of large language models suggest [12, 33]. To statisticians, better typically means *unbiased*, though "bias" is used differently from in the bias-variance tradeoff [8], or in algorithmic bias [7]. The growing "responsible AI" literature emphasizes that datasets are better when they are ethically and fairly sourced [22, 13, 24]. This paper underscores *legality* as one desideratum for "better". To this end, it reviews key legal obligations on data collection and use, examines the practical impact of data laws on ML pipelines, and offers a framework for thinking about data legality.

## 2 When are datasets legal?

Legal datasets may be understood broadly as datasets which are legally collected, retained, processed, *and* disseminated. This fourfold categorization builds off Solove's classic taxonomy of privacy [27], and finds expression in a range of relatively new legislation worldwide. This notably includes the European Union's (EU's) *General Data Protection Regulation* (GDPR) which came into force in 2018. Parallel to the GDPR are *national* data laws, such as South Korea's *Personal Information Protection Act* (passed in 2011), Singapore's *Personal Data Protection Act* (passed in 2014) and, most recently, China's *Personal Information Protection Law* (PIPL, August 2021). While the US does not presently have data legislation at the federal level, states like California, New York, and Massachusetts have passed data privacy acts. Further, legal scholars and courts have increasingly considered how pre-existing laws, such as copyright and anti-discrimination law, affect ML datasets [25, 19, 10]. As there are too many countries and variations to cover, I use the GDPR, PIPL, and California's *Consumer Privacy Act* (CCPA, 2018) as case studies.

Although one jurisdiction's laws generally do not apply in another, modern data laws tend to have extra-territorial effect. Both the GDPR and PIPL apply as long as any personal data about persons in the EU/China is processed for any commercial or behavioral monitoring purposes (GDPR, Art 3; PIPL, Art 3). Likewise, Art 2 of the EU's proposed *Artificial Intelligence Act* (AIA) expressly covers AI systems deployed in, or whose outputs are used, in the EU, regardless of where the providers

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

and users of the system are. By contrast, the CCPA applies primarily to large businesses which "do business in" the state (CCPA, §1798.140). Thus, ML researchers and practitioners worldwide are now subject to foreign, and increasingly complex [17], data laws. Below I non-exhaustively review key legal obligations they impose. Note that "legal" here refers only to formal *law*. This distinguishes my scope from (no less important) work on "ethical", "fair" or "responsible" AI (e.g. [22, 13, 24]). Despite clear overlaps, neither is a subset of the other. To illustrate, for some in certain states abortion is ethical yet illegal, for others elsewhere it is unethical yet legal.

### 2.1 Collection

Most centrally, data protection laws require informed consent before "personal" data may be obtained (GDPR, Arts 6–11; PIPL, Arts 13–17). The CCPA does not expressly require "consent", but businesses must inform consumers of the scope and purposes of data collected before collection (CCPA, §1798.100). The legality of numerous facial recognition datasets has been challenged for lack of consent [1, 22, 21]. Facial recognition clearly involves personal data because the task is to *identify*. "Personal data" is, however, wider. Article 4 GDPR defines it as "any information relating to an identified or identifiable natural person". One is "identifiable" when they may be identified directly (i.e. by name) or indirectly. Names are not necessary; zip codes, gender, race, etc, could collectively identify. Indeed, Wong suggests that the EU's definition of personal data "appears to be capable of encompassing all information in its ambit", as EU courts have taken "personal" to include not only data *about* a person, but also data which *affects* them [31].

The breadth of data laws explains why although most of ImageNet's [6] label classes do not target persons, its caretakers recently blurred out all human faces in the data, citing privacy concerns [16]. A similar fate appears to have befallen the new Meta's facial recognition systems [21]. While most legal scrutiny has been on images, text, sound, and other modalities can also be "personal". One's forum posts, even if pseudonymous, could reveal much of their background. As such, dataset builders should be deliberate about obtaining consent even (or especially) when it is not obvious if the data is "personal".

### 2.2 Retention

A standard feature of legal consent is that consent may be withdrawn at any time. Data subjects may request to correct or erase their data (GDPR, Arts 16–17; PIPL, Arts 15, 16, 44–47; CCPA, §1798.105–106). Beyond consent, data controllers are also obliged to keep data in personally-identifiable form for no longer than necessary for its stated purposes (GDPR, Art 5(1)(e); CCPA, §1798.100(3)). Data that has served these purposes (say, the model has been trained) must be deleted or anonymized. However, given that data previously collected for one purpose can turn out useful for another, deletion may be quite undesirable for ML engineers. Anonymization is not much better, since preventing re-identification may require destroying most of a dataset's informative signals [23, 32].

As such, prior thought should be given to delineating, and communicating, what the data will be used for. Conveying a specific purpose such as "training ML models" may not cover *maintaining* or *updating* the model post-deployment. Too general a purpose, such as "for ML processing" invites user suspicion and may fall outside the legal requirement that consent must be given in respect of "specific" purposes (GDPR, Art 6(1)(a); PIPL, Art 6).

### 2.3 Processing

Consent obligations surrounding data collection apply equally to data use. Further, data subjects have a right to be informed of and object to decisions "based solely on automated processing" (GDPR, Art 22; see also PIPL, Art 24). This legally advantages human-*in-the-loop* systems. Beyond data protection laws, anti-discrimination laws in certain jurisdictions (e.g. US disparate treatment/impact laws; UK's *Equality Act* 2010) may prohibit the use of protected attributes like race and gender for profiling [11]. This restricts the feature set which can legally be used for training ML models. Features highly co-linear with protected attributes may be indirectly prohibited as well.

While the obligations above may be more relevant to *models* than to *datasets*, laws can target the latter directly. Most prominently, the proposed Art 10 AIA stipulates that "[t]raining, validation and testing data sets" to be used in what the Act identifies as "high-risk AI systems" shall be "relevant,

representative, free of errors and complete". This extends to having "appropriate statistical properties" regarding the system's target persons, and considering characteristics "particular to the ... setting within which the high-risk AI system is intended to be used". The draft AIA is in early stages and may take years and numerous amendments to come into force (if it does). Should it become law as is, it may effectively render data-centricity *legally mandatory* for "high-risk" AI. Examples of high-risk AI enumerated in Annex III AIA non-exhaustively include systems for biometric identification, educational assessments, recruitment, credit scoring, law enforcement, and judicial decisions.

### 2.4 Sharing and disclosure

As data sharing or disclosure also constitutes processing, unauthorized disclosure is also a breach (GDPR, Arts 4(2); PIPL, Art 25). Thus, datasets with potentially personal information cannot be open-sourced without proper anonymization, even for research purposes. Another concern particular to large neural networks is the possibility that the network may memorize and leak personal information in the training data [20, 3]. Personal information in training datasets may therefore need to be removed in advance.

### 2.5 Research exemptions

The obligations above may be subject to limited research exemptions whose scope differs across jurisdictions [18]. For instance, both the GDPR and the CCPA regard subsequent scientific or statistical research as compatible with the initial purposes of the data collection for which consent was presumably obtained. This allows research to proceed without needing to ask for additional consent, subject to appropriate safeguards (GDPR, Arts 1(b) & 89(1); CCPA, §1798.140(s)). On its face, this *may* have been sufficient to cover research applications of the ImageNet data (discussed above) without requiring anonymization. China's PIPL, however, does not have such an exemption.

## 3 Implications on ML pipelines

There is, in short, an growing range of legal constraints on when and how data may be used. This has obvious implications for the ML community. Since *legal* data is necessarily a subset of *all* data, prioritizing legality seems to require sacrificing model performance. But less data is not always worse, especially if it also means less noise. More formally, if we think of ML broadly as seeking $argmax_H g(H, y)$, where the hypothesis $H(\theta; X, D)$ takes weights $\theta$ learned from features $X$ in dataset $D$, and $g(H, y)$ is a performance metric measured against (holdout) truth labels $y$, then legality constraints might be understood as follows:

$$argmax_H g(H, y) \quad \mathbf{s.t.} \quad X \subseteq X_{legal}; D \subseteq D_{legal}$$

with $X_{legal}$ and $D_{legal}$ respectively denoting the legally-permissible set of features and datasets.

Formally framing the problem as such identifies three situations where legal constraints may not necessarily limit model performance. For brevity, we illustrate this with $D$, though similar logic applies with $X$. First, if $D_{legal} = D_\Omega$ (all data relevant to a given task). That is, data laws have no practical effect on datasets in that area. For example, the task involves only cat detection and never implicates personal data. Second, if $D^*$, the theoretically optimal dataset, happens to be perfectly legal and so $D^* = D_{legal}$. Third, and least obviously, a legally-constrained optimization problem produce might produce *better* performance than its unconstrained variant. This counter-intuitive result may occur in the real world because the law may force one to exclude noisy data (or features) that one would otherwise have included. In this sense, data laws, like other optimization constraints, may turn out to have a useful regularizing effect.

Moreover, in practice $g(H, y)$ is not the only metric to be optimized. Even assuming we care solely about economic value, profits, while presumably correlated with F1 score-type performance metrics, turn also on variables like user adoption and trust [28, 9, 15]. A perfectly accurate classifier that is never used generates no marginal benefit. Fines also reduce profit. In this way, ignoring legality can be seen as another source of "hidden debt" in ML pipelines [26]. Early investments in processes and practices for making legal datasets could yield better *real world* performance, particularly in the

**Before Collection — Which Law(s) Apply?**

1. Will directly or indirectly personally-identifying data ("personal data") be collected? Note the wide scope of "personal data" (see Part 2.1).
2. If personal data is collected, which countries are we (the data *processors*) in? What data laws do these countries have?
3. What countries are the data *subjects* in? Do these countries have *extra-territorial* data laws?
4. What is the data for? Are these purposes (e.g. research) subject to any exemptions?

**During Collection — Obtaining Consent**

1. Are subjects adequately informed of the purposes of and consent to data collection? Adequate "consent" varies by country, but refer generally to GDPR, Arts 6(1)(a) & 7.
2. Ensure that communicated purposes truly cover all intended data uses. Purposes may be stated more broadly to leave room for future uses, but should not be framed too widely either.

**After Collection — Data Use and Model Training**

1. Ensure that subjects who withdraw consent are removed from the dataset. Data pipelines should be robust to consent withdrawals. Already-generated artefacts should be re-generated.
2. Check that protected attributes (and strong correlates) are excluded from the feature set.

**After Training —- Deployment**

1. If the system is entirely end-to-end with no humans-in-the-loop, subjects must be informed.
2. If the communicated purposes of data collection have been spent, the data may need to be anonymized or deleted.
3. Does the trained model leak personal data? If so, either tweak the model and its outputs, or obtain consent for data *disclosure*.

Figure 1: Suggested framework for dataset legality

long-run (where legal enforcement becomes more feasible). Apart from any obligation to follow the law just because it is law, there are practical reasons why the ML community ought to do so.

## 4 Building legal datasets

Complying with the intricate and growing web of data laws is non-trivial. The challenge is how we might turn motherhood calls for "multi-disciplinary collaboration" into actionable steps for ML researchers. The rise of ethics guidelines and responsible AI checklists [34, 24] offers one solution. In a sense, this involves scholars from ethics, sociology, and other (typically qualitative) disciplines reducing complex, open-ended obligations into simpler, close-ended compliance heuristics for computer scientists to follow. Furthering this trend, Figure 1 offers a framework for thinking about dataset legality. This builds on existing work that already incorporates some legal principles (e.g. [24]) but differs in two ways. First, the framework focuses solely on *legality* and thus *complements* responsible/ethical AI work. Second, while checklists and impact statements are generally backward-looking, encouraging researchers to justify choices already made, these considerations are forward-looking, encouraging researchers to think about legality at each stage of the ML process. This is crucial because legal errors, especially the need to obtain informed consent for processing, are expensive to rectify *post facto* if not avoided *ex ante*.

## 5 Limitations

All heuristics are wrong, but some are useful [2]. This paper does not cover all the legal obligations, duties, and exemptions affecting ML datasets. Nor can following the proposed framework completely guarantee legality (nor fairness or morality). Indeed, the corpus of legislation affecting ML datasets is set to grow, amid concerns that current laws offer insufficient safeguards [30, 29]. The draft AIA, if and when passed, would significantly alter the AI and data governance landscape; other jurisdictions may follow suit with their own Acts. The minutiae of dataset legality should be fleshed out in future, lengthier work. The primary aim here is to spark discussion on when and why legal data is better

data. Data-centric AI presents an opportunity for the ML community to build better datasets — in all the technical, statistical, ethical, and legal senses of the word.

## Acknowledgments and Disclosure of Funding

The authors disclose no funding sources nor competing interests.

## References

[1] Legality of collecting faces online challenged, May 2021. URL https://www.bbc.com/news/technology-57268121.

[2] G. Box. Science and Statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976. URL http://www-sop.inria.fr/members/Ian.Jermyn/philosophy/writings/Boxonmaths.pdf.

[3] C. Chen, B. Wu, M. Qiu, L. Wang, and J. Zhou. A Comprehensive Analysis of Information Leakage in Deep Transfer Learning. *arXiv:2009.01989 [cs]*, Sept. 2020. URL http://arxiv.org/abs/2009.01989. arXiv: 2009.01989.

[4] B. Collins, J. Deng, K. Li, and L. Fei-Fei. Towards Scalable Dataset Construction: An Active Learning Approach. In D. Forsyth, P. Torr, and A. Zisserman, editors, *Computer Vision – ECCV 2008*, volume 5302, pages 86–98. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-88681-5 978-3-540-88682-2. doi: 10.1007/978-3-540-88682-2_8. URL http://link.springer.com/10.1007/978-3-540-88682-2_8. Series Title: Lecture Notes in Computer Science.

[5] T. Dao, A. Gu, A. J. Ratner, V. Smith, C. D. Sa, and C. Re. A Kernel Theory of Modern Data Augmentation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 1528–1537, 2019.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[7] B. Friedman and H. Nissenbaum. Bias in Computer Systems. *ACM Transactions on Information Systems*, 14(3):330–347, July 1996.

[8] S. Geman, E. Bienenstock, and R. Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4:1–58, 1992.

[9] O. Gillath, T. Ai, M. S. Branicky, S. Keshmiri, R. B. Davison, and R. Spaulding. Attachment and trust in artificial intelligence. *Computers in Human Behavior*, 115:106607, Feb. 2021. ISSN 07475632. doi: 10.1016/j.chb.2020.106607. URL https://linkinghub.elsevier.com/retrieve/pii/S074756322030354X.

[10] T. B. Gillis and J. L. Spiess. Big Data and Discrimination. *The University of Chicago Law Review*, 86(2): 459, 2019.

[11] D. Hellman. Measuring Algorithmic Fairness. *Virginia Law Review*, 106(4):811, 2020.

[12] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song. Pretrained Transformers Improve Out-of-Distribution Robustness. pages 2744–2751, July 2020. doi: 10.18653/v1/2020.acl-main.244. URL https://aclanthology.org/2020.acl-main.244.

[13] B. Hutchinson, A. Smart, A. Hanna, E. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 560–575, Virtual Event Canada, Mar. 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445918. URL https://dl.acm.org/doi/10.1145/3442188.3445918.

[14] L. C. Irani and M. S. Silberman. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 611–620, Paris France, Apr. 2013. ACM. ISBN 978-1-4503-1899-0. doi: 10.1145/2470654.2470742. URL https://dl.acm.org/doi/10.1145/2470654.2470742.

[15] C. X. Kerasidou, A. Kerasidou, M. Buscher, and S. Wilkinson. Before and beyond trust: reliance in medical AI. *Journal of Medical Ethics*, Aug. 2021. doi: 10.1136/medethics-2020-107095.

[16] W. Knight. Researchers Blur Faces That Launched a Thousand Algorithms. *Wired*, Mar. 2021. ISSN 1059-1028. URL https://www.wired.com/story/researchers-blur-faces-launched-thousand-algorithms/. Section: tags.

[17] B.-J. Koops. The trouble with European data protection law. *International Data Privacy Law*, 4(4): 250–261, Nov. 2014. ISSN 2044-3994, 2044-4001. doi: 10.1093/idpl/ipu023. URL https://academic.oup.com/idpl/article-lookup/doi/10.1093/idpl/ipu023.

[18] C. Mabel and S. Tara. The Research Exemption Carve Out: Understanding Research Participants Rights Under Gdpr and U.s. Data Privacy Laws. *Jurimetrics*, 60:125, 2019.

[19] S. G. Mayson. Bias In, Bias Out. *Yale Law Journal*, 128:2218, 2019. URL https://www.yalelawjournal.org/pdf/Mayson_p5g2tz2m.pdf.

[20] M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753, May 2019. doi: 10.1109/SP.2019.00065. arXiv: 1812.00910.

[21] M. O'Brien and B. Ortutay. Facebook to shut down face-recognition system, delete data. *Washington Post*, Nov. 2021. ISSN 0190-8286. URL https://www.washingtonpost.com/business/facebook-to-shut-down-face-recognition-system-delete-data/2021/11/02/79c0b782-3c02-11ec-bd6f-da376f47304e_story.html.

[22] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. In *Proceedings of the Workshop on ML-Retrospectives, Surveys & Meta-Analyses @ NeurIPS 2020*, Dec. 2020.

[23] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1), Dec. 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10933-3. URL http://www.nature.com/articles/s41467-019-10933-3.

[24] A. Rogers, T. Baldwin, and K. Leins. Just What do You Think You're Doing, Dave?' A Checklist for Responsible Data Use in NLP. Sept. 2021. URL http://arxiv.org/abs/2109.06598.

[25] M. Sag. The New Legal Landscape for Text Mining and Machine Learning. *Journal of the Copyright Society of the USA*, 66:291, 2019. doi: 10.2139/ssrn.3331606.

[26] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden Technical Debt in Machine Learning Systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 2, pages 2503–2511, 2015.

[27] D. J. Solove. A Taxonomy of Privacy. *University of Pennsylvania Law Review*, 154(3):477, Jan. 2006. ISSN 00419907. doi: 10.2307/40041279. URL https://www.jstor.org/stable/10.2307/40041279?origin=crossref.

[28] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel. The relationship between trust in ai and trustworthy machine learning technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 272–283, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372834. URL https://doi.org/10.1145/3351095.3372834.

[29] United Nations High Commissioner for Human Rights. The right to privacy in the digital age. Technical Report A/HRC/48/31, Sept. 2021.

[30] S. Wachter. Data protection in the age of big data. *Nature Electronics*, 2(1):6–7, Jan. 2019. ISSN 2520-1131. doi: 10.1038/s41928-018-0193-y. URL https://www.nature.com/articles/s41928-018-0193-y. Institution: Nature Publishing Group Number: 1.

[31] B. Wong. Delimiting the concept of personal data after the GDPR. *Legal Studies*, 39(3):517–532, Sept. 2019. ISSN 0261-3875, 1748-121X. doi: 10.1017/lst.2018.52. URL https://www.cambridge.org/core/product/identifier/S0261387518000521/type/journal_article.

[32] H. Xu and N. Zhang. Implications of Data Anonymization on the Statistical Evidence of Disparity. *Management Science*, pages 1–19, 2021. ISSN 1556-5068. doi: 10.2139/ssrn.3662612.

[33] R. Zhong, D. Ghosh, D. Klein, and J. Steinhardt. Are Larger Pretrained Language Models Uniformly Better? Comparing Performance at the Instance Level. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3813–3827, Aug. 2021. doi: 10.18653/v1/2021.findings-acl.334. URL https://aclanthology.org/2021.findings-acl.334.

[34] M. Zook, S. Barocas, D. Boyd, K. Crawford, E. Keller, S. P. Gangadharan, A. Goodman, R. Hollander, B. A. Koenig, J. Metcalf, A. Narayanan, A. Nelson, and F. Pasquale. Ten simple rules for responsible big data research. *PLOS Computational Biology*, 13(3):e1005399, Mar. 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005399. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005399.