How Data-Related AI Research can Support Technical Solutions for Regulatory Compliance

Anonymous Author(s)
Affiliation
Address
email

Abstract

Ensuring high-quality, representative, and secure datasets is critical for compliance with emerging regulatory frameworks such as the EU AI Act (Art. 10). In this paper, we survey five key data-centric challenges: intrinsic and context-dependent data quality, availability, variability, and security, and link each challenge to established and emerging methods and research. We then propose a workflow that integrates best practices from machine learning research with regulatory requirements, illustrating how each step can be operationalized to meet the "relevant, representative, error-free" criteria. Our analysis highlights opportunities for regulators to refine their mandates by incorporating advances in ML research.

1 Introduction

High-risk AI systems depend critically on training data that is "relevant, representative, error-free, and complete" (EU AI Act Art. 10). Yet, practitioners often satisfy these mandates with coarse governance checklists and simple metrics (e.g. label-error rate, missing-value fraction), leaving a wide gulf between regulatory intent and technical practice. Concurrently, AI research has produced methods such as noisy label detection, core-sets, data attribution, and valuation that can directly address legal requirements but remain underused in compliance workflows. In this paper, we bridge this gap by (i) proposing a regulation-aligned taxonomy of five challenges in ML data (inherent data quality, context-dependent data quality, availability, variability and security); (ii) mapping established and emerging research and methods to this taxonomy; (iii) sketching an exemplary workflow for *regulatable ML*. Our analysis highlights opportunities for regulators and practitioners to narrow the gap between policy and practice by leveraging recent advances in data-centric ML.

2 Related Work

Data quality in machine learning encompasses regulatory requirements, documentation, quantitative metrics, and other methodologies. We review major contributions in each area and highlight gaps.

Regulatory Frameworks The EU AI Act [18] mandates that training, validation, and test sets be "relevant, representative, error-free, and complete" (Art. 10(3)), but leaves the technical implementation open. These criteria are further detailed in standards, such as those from the ISO/IEC JTC 1/SC 42 committee for AI. Most notably ISO/IEC 5259 [33], specifies 24 data-quality metrics for analytics and machine learning. ISO/IEC 5259 organizes these metrics along two dimensions (inherent quality and context-dependent quality) and operationalizes the notions of relevance, representativeness, accuracy, and completeness, albeit at a relatively high level of abstraction. Detailed definitions of each metric are provided in Appendix C.

Submitted to Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Challenge	Focus	Regulatory Source
Inherent Data Quality	Accuracy, label errors, duplicates	AI Act 10; GDPR5(1)(d)
Context-Dependent Quality	Bias, imbalance, feature relevance	AI Act 10
Data Availability	Sufficiency, edge cases, labeling budget	GDPR Art. 5(1)(c); 15
Data Variability	Distribution shift, drift detection	AI Act 72; GDPR 5(1)(d)
Data Security	Privacy, poisoning, breach protection	AI Act 15; GDPR 32

Table 1: Regulation-aligned taxonomy of ML data challenges.

Data Documentation and Transparency Standardized documentation frameworks, such as Data Cards [56] and Datasheets for Datasets [24], record a dataset's origin, composition, labeling process, and limitations. While these efforts support regulatory compliance by improving transparency, they do not engage directly with legal mandates or prescribe quantitative metrics for quality.

Data Quality Metrics Surveys on data quality metrics [78, 10, 48, 27, 55] collate measures, often aligned with ISO/IEC 5259, that quantify dimensions such as completeness, consistency, and timeliness. Although taxonomies vary, a common dichotomy separates inherent and context-dependent dimensions. Few works explicitly tie metric selection to regulatory requirements, leaving a gap between theoretical measures and compliance.

Data-Centric AI Summaries and Recommendations Jakubik et al. [34] distinguish between dataset extension (collection) and refinement (quality improvement), illustrating commercial tooling for each. Zha et al. [74] organize the data lifecycle into training-data development, inference-data development, and maintenance, providing resources and tools but omitting deeper treatment of emerging topics like data valuation (Section 6.6). Hammoudeh and Lowd [32] survey methods for training-data influence (e.g., poisoning, backdoor attacks, data reduction) with a technical focus but limited practical guidance, and Yu et al. [73] review dataset distillation techniques, highlighting applications in continual and federated learning, privacy, and robustness. Practitioner-oriented recommendations by Lones [43], Orr and Crawford [53], and Zhao et al. [77] advocate for dataset diversification, rigorous quality checks, and transparent documentation. However, these standards rarely provide explicit mappings to regulatory criteria or concrete metrics, leaving practitioners to interpret abstract requirements without clear technical guidance.

2.1 Interim Conclusion

While prior work addresses legal mandates, documentation standards, metric surveys, and data-centric methodologies, these threads remain largely siloed rather than integrated into a unified compliance framework. In the next section, we introduce key terminology and an analytical lens to bridge this gap.

3 Taxonomy of ML Data Challenges

Scientific literature and regulatory frameworks use varied terminology for ML data challenges. In this work we use the terms: inherent data quality, context-dependent data quality, data availability, data variability, and data security to broadly categorize data problems. Table 1 summarizes these, with examples and regulatory references.

Inherent Data Quality Inherent data quality encompasses all metrics that can be computed with respect to the available data without any further knowledge. Key challenges include ensuring label and annotation accuracy, identifying and correcting noise and errors such as incorrect values, inconsistencies, and duplicate data [51, 9]. These issues can degrade model accuracy, skew the training process, and inflate dataset size.

Relationship to Regulation GDPR Article 5(1)(d) mandates data accuracy [17]. The AI Act Article 10 requires error-free datasets [18].

Context-Dependent Data Quality Context-dependent data quality involves properties requiring external assumptions about the deployment environment. Challenges include class imbalances causing biased outcomes [11], bias/fairness issues perpetuating societal inequities [47], improper data splits leading to overfitting, and irrelevant features reducing model relevance. Each of these issues rests on implicit assumptions about real-world conditions and thus undermines true dataset representativeness.

Relationship to Regulation AI Act Article 10 mandates relevant, representative datasets [18].

Data Availability Data availability addresses collecting sufficient data and ensuring safe access. Challenges include limited data due to labeling costs, class imbalances, or privacy constraints, and excessive data causing storage/processing issues [12]. Techniques like transfer learning [79], and active learning [60] maximize data utility.

Relationship to Regulation GDPR Article 5(1)(c) mandates data minimization, limiting collection to necessary data [17]. Article 15 grants individuals the right to access their personal data, impacting availability.

Data Variability Data variability describes changes in data over time, such as evolving user behaviors, shifting environments, or updated collection methods, causing drift between training and real-world data. Static models may fail in dynamic settings, reducing reliability. Drift detection [6] and adaptive methods address this.

Relationship to Regulation AI Act Article 72 requires post-market monitoring for high-risk systems [18]. GDPR Article 5(1)(d) mandates that personal data must be kept up-to-date to maintain its relevance and integrity [17].

Data Security AI systems are prone to both classical security aspects related to data access and novel attack schemes such as data poisoning [66] or membership inference attacks [61]. Classical privacy concerns involve safeguarding sensitive information from unauthorized access and ensuring compliance with regulations such as GDPR [17]. Security, on the other hand, focuses on protecting data from breaches, malicious attacks, and unauthorized modifications. Encryption, secure data storage, and access controls are measures to maintain data integrity and security. From the machine learning (ML) side, techniques like data anonymization [49], differential privacy [4] and measures against data poisoning [66] can help protect individual privacy while maintaining data utility.

Relationship to Regulation AI Act Article 15 mandates robustness and cybersecurity for high-risk systems [18]. GDPR Article 32 requires secure processing [17].

4 Methodology

The aim of this paper is to map emerging and established research efforts onto these dimensions. To this end, we conducted a structured literature review. First, we queried Google Scholar, arXiv, and major AI/ML conference proceedings for studies on data-centric practices and regulatory requirements, supplementing these results with recommendations generated by large-language models. Second, we extracted methods catalogued in foundational surveys and overviews [43, 74, 34] and designated them as *established*. Finally, we identified additional techniques absent from these summaries, labeled them as *emerging*, and mapped every method to our taxonomy of data challenges.

5 Mapping Established Methods to Regulatory Needs

We now show how common techniques directly address regulatory requirements. Table 2 (top) summarizes the mapping, followed by brief explanations. For each method, we illustrate *opportunities* to address the identified data quality issues. We also discuss potential *challenges* associated with implementing these methods in practice.

5.1 Data Validation Techniques and Metrics

Data validation ensures that training datasets are accurate and error-free. Statistical techniques can detect and correct label mistakes [51, 52], while tools like TensorFlow Data Validation (TFDV) flag

	INHERENT	CONTEXT	AVAILABILITY	VARIABILITY	SECURITY
Data Validation	✓				
Drift Detection				✓	
Feature Selection	✓	✓			✓
Active Learning		✓	✓		
Core-Sets	✓		✓		
Data Augmentation		✓	✓		✓
Syntehtic Data	✓	✓	✓	✓	✓
Data Difficulty	✓		✓		
Data Distillation			✓		✓
Memorization	✓		✓		✓
Explainable AI	✓	✓	✓		✓
Data Attribution	✓	✓			✓
Data Valuation	✓	✓	✓		✓
Sample Size	✓	✓	✓		

Table 2: Data Challenges and corresponding methods and research areas. The top methods are more established whereas the bottom ones require more research.

anomalies and produce descriptive statistics [9]. Beyond error detection, several works introduce metrics for data quality. Mitchell et al. [48] reviews measures, e.g. Euclidean distance, KL-divergence, but without prescribing thresholds [48], and Zhao et al. [76] analyze over 100 ML datasets to propose a framework for evaluating reliability, validity, and diversity. Other studies examine dimensions such as accuracy, completeness, and timeliness in line with ISO 5259; see Appendix C for a detailed comparison [29, 78, 10].

Opportunities: Data validation techniques play a crucial role in detecting irregularities within datasets, which can significantly impact model performance [51]. By identifying and correcting these errors, data validation enhances dataset accuracy and reliability, ultimately leading to improved model performance and increased trustworthiness. This directly addresses INHERENT data quality challenges.

Challenges: One critical challenge in data validation is distinguishing between errors and edge cases, particularly when automatic corrections are applied. For instance, balancing distributions may not always be desirable [64]. Automatically fixing errors can inadvertently alter data instances that were originally correct or overlook complex errors that require human judgment.

5.2 Drift Detection

Drift detection refers to the recognition of shifts between the data a model was trained on and real-world data. Identifying such drifts can be achieved by tracking performance changes over time or, from a data perspective, by comparing the distribution of the training data to that of the real-world data [26].

Opportunities: Drift detection directly addresses data VARIABILITY by identifying shifts in live data. This information can be utilized to take various actions, such as stopping the model, triggering a fail-safe mode, or requiring human intervention to reassess or retrain the model.

Challenges: Although drift detection is crucial, implementing it effectively can be challenging. Identifying the right metrics for detecting drift and establishing thresholds for action can be complex.

5.3 Feature Selection

Feature Selection is an important aspect of machine learning and has been well addressed in the literature [65]. Jakubik et al. [34] summarize it as one of the two major components of data-centric

AI. They distinguish between methods aimed at improving feature quality, which include removing irrelevant features, and methods for creating or acquiring new relevant features.

Opportunities: Feature selection can reduce the size of the dataset while also enhancing its representativity. It addresses both INHERENT and CONTEXT-DEPENDENT data quality; removing unnecessary features increases the inherent value of the data, while certain features may only be essential for specific tasks, thereby improving the contextual relevance of the dataset. Furthermore, the choice of features may present SECURITY issues, as they might reveal sensitive information.

Challenges: While feature selection is vital, it can be challenging to determine which features are truly irrelevant or redundant. Additionally, the process may require domain expertise to ensure that important features relevant to specific tasks are not inadvertently discarded. This balance is necessary to maintain the overall predictive power of the model.

5.4 Active Learning

In active learning, the algorithm has access to a small labeled dataset D and a large unlabeled dataset U. The algorithm can query an oracle (e.g., a human labeler) for labels up to a certain budget b. The objective is to optimize model performance within the constraints of this budget [60]. A common method for achieving this is by selecting data instances with high predictive uncertainty.

Opportunities: Active learning efficiently utilizes labeling resources by focusing on the most informative data instances, thereby addressing the challenge of AVAILABILITY. Furthermore, active learning supports the selection of the most relevant and diverse data instances for a specific task, enhancing the representativity of the dataset for that application. This, in turn, addresses CONTEXT-DEPENDENT data quality, ensuring that the model is trained on data that accurately reflects the target domain.

Challenges: Implementing active learning can be computationally intensive, as it requires iterative model training and evaluation. Additionally, the effectiveness of active learning heavily depends on the choice of the query strategy, which may not be universally optimal for all types of data and tasks. Moreover, the need for an oracle to provide labels can introduce delays and inconsistencies, especially if human labelers are used.

5.5 Core-Sets

A core-set is a small subset $S \subseteq D$ of a full dataset D such that the learning algorithm achieves similar performance when trained on this subset as it would on the entire dataset [20]. A popular method for finding core-sets is the use of clustering techniques, where representative samples are selected based on their distances to cluster centroids.

Opportunities: Core-sets significantly reduce the volume of data required for training without compromising model performance. This is particularly beneficial for large datasets, as it minimizes storage and computational needs while retaining the essential characteristics of the data. By ensuring that the dataset remains manageable and efficient, core-sets address the challenge of data AVAILABILITY. Conversely, if a smaller dataset can achieve similar performance to a larger one, the smaller dataset should be favored.The data size can be considered an INHERENT property.

Challenges: Identifying an optimal core-set can be both challenging and computationally expensive, especially for complex datasets. The process often involves sophisticated algorithms and heuristics, which may not be straightforward to implement. Furthermore, core-sets might not capture all the nuances of the original data, particularly in scenarios where rare events or minority classes are important. This limitation can lead to reduced performance in applications where such events are critical, underscoring the need for careful consideration when employing core-sets.

5.6 Data Augmentation

Data augmentation methods enhance the quality and diversity of training data by artificially increasing the size of datasets. Techniques like Synthetic Minority Over-sampling Technique (SMOTE) [11] can be effective in addressing data representativity issues, particularly class imbalances.

Opportunities: Data augmentation techniques contribute to creating a more diverse and representative dataset. By artificially increasing the data size, these methods effectively address class imbalances and

enhance the model's generalizability, leading to improved performance and fairness. While data augmentation primarily addresses data AVAILABILITY, it can also improve CONTEXT-DEPENDENT data quality for specific tasks and may provide partial defenses against overfitting and model vulnerability to SECURITY breaches by diversifying the training data.

Challenges: A recent article titled *The Good, the Bad, and the Ugly Sides of Data Augmentation* summarizes the challenges well [42]. Artificially modified or balanced data may not represent real-world scenarios, potentially introducing noise or artifacts that could negatively affect model performance [16]. Additionally, the effectiveness of different augmentation techniques varies greatly depending on the specific dataset and task, necessitating careful experimentation and tuning.

5.7 Synthetic Data

Synthetic data refers to the creation of artificial data for machine learning. This approach is particularly advantageous in scenarios where data collection is difficult, expensive, or where privacy concerns are significant [5, 44]. Typically, techniques such as Generative Adversarial Networks (GANs) or stable diffusion are employed to generate this data.

Opportunities: Synthetic data enables the generation of large, diverse datasets without collecting sensitive real-world samples, thereby reducing cost and time while safeguarding privacy and ensuring regulatory compliance. By filling gaps, e.g. underrepresented classes or rare scenarios, and simulating varied conditions, it boosts model robustness (addressing INHERENT and CONTEXT-DEPENDENT data quality, and VARIABILITY) and mitigates SECURITY risks tied to real data.

Challenges: Generating synthetic data can be challenging, especially for images, because it often requires training large models like GANs [28] or diffusion models [59]. Additionally, synthetic data can introduce another layer of bias and a content gap [70]. Obtaining performance comparable to or superior to real data remains a significant hurdle in many applications [5].

6 Mapping Emerging Methods to Regulatory Needs

Next, we map emerging directions onto our taxonomy. We classify a method as "emerging" if it was identified in our review but is not yet covered by existing surveys. Table 2 (bottom) provides an overview of these methods and the primary data challenges they address.

6.1 Data Difficulty

Meding et al. [46] describe dichotomous data difficulty and show that many datasets, such as ImageNet, suffer from imbalanced data difficulty. There are many data instances in the test set that are never classified correctly (called *impossible*) and many that are always classified correctly (called *trivial*). They show that models can be better compared on the remaining instances. A similar conclusion can be drawn for label errors. Northcutt et al. [51] show that larger models tend to be favored on datasets with label errors, while smaller models might actually outperform them when evaluated on a dataset without errors.

Opportunities: Understanding dichotomous data difficulty enhances the precision of model performance evaluation. Performance across these distinct samples should be assessed specifically, potentially through extending labels with metadata that indicates the difficulty level of individual samples. This approach could also help identify areas of the dataset that are underrepresented. If certain concepts that tend to be more difficult can be identified, it may suggest the need for gathering additional training data. Thus, this understanding can serve as both an INHERENT data quality metric and a means to address data AVAILABILITY.

Challenges: Addressing dichotomous data difficulty necessitates training multiple models, which can be computationally intensive. Moreover, while it may function as a data quality metric, its validation outside of academic settings remains uncertain and requires further investigation.

6.2 Data Distillation

Data Distillation is a relatively new field in machine learning, first introduced by Wang et al. [71]. Conceptually related to core-sets, the goal of data distillation is to condense a large dataset into a

smaller one that maintains similar performance. The key distinction is that distilled data consists of synthetic images, which are often not recognizable by humans. A popular method for creating these synthetic images is gradient matching [75].

Opportunities: Similar to core-sets and synthetic data, data distillation can reduce the required dataset size, thereby addressing the data AVAILABILITY issue. Furthermore, as Yu et al. [73] point out, it can also contribute to resolving privacy, security, and robustness challenges, thus addressing SECURITY challenges.

Challenges: Although data distillation holds significant potential and has been evaluated on real-world examples, such as medical data [41], its practical applications, particularly in meeting regulatory demands, remain underexplored.

6.3 Memorization

Memorization occurs when a model heavily relies on unique training instances to make predictions, akin to a student memorizing a rare fact for an exam. For example, in a facial recognition dataset, a single image of a person with a distinctive tattoo may be memorized, improving accuracy for that individual but risking privacy through membership inference attacks [72]. Feldman [21] defines a training instance as unique if its removal reduces the model's ability to classify it correctly. Estimating memorization involves training models with and without the instance, though efficient methods exist [21, 36]. Jiang et al. [36] further use memorization as a measure to categorize the structure of a dataset and show that mislabeled instances are harder to memorize.

Opportunities: Memorization scores identify underrepresented regions (e.g., rare faces) and label errors, addressing data AVAILABILITY and INHERENT data quality. They also flag privacy risks, enhancing SECURITY. Additionally, memorization has proven useful in data pruning [63].

Challenges: A primary issue is the difficulty of translating memorization scores into actionable data collection strategies. Memorization scores identify unique training samples critical for generalization but do not inherently specify what additional data to collect.

6.4 Explainable AI

Explainable AI (XAI) techniques aim to make machine learning models transparent by revealing which inputs drive predictions. Early methods such as LIME [58] and SHAP [45] highlight influential features or regions in individual samples. Recent work evaluates XAI's role in debugging models [2, 1], detecting bias [13, 19], certifying AI systems [23], and building user trust [67, 68]. Crucially, XAI can uncover data issues, for example, Ribeiro et al. showed that a wolf detector was focusing on snowy backgrounds rather than the animal itself, leading to targeted data augmentation [58].

Opportunities: XAI pinpoints spurious correlations and gaps in data coverage, guiding selective data collection or relabeling to improve both INHERENT and CONTEXT-DEPENDENT data quality. By exposing underrepresented scenarios, it addresses AVAILABILITY concerns, and by revealing memorized examples, it highlights potential privacy and SECURITY risks [22].

Challenges: Interpretations generated by XAI methods are often ambiguous or misleading, requiring expert judgment to translate insights into actionable data-quality improvements [3, 23]. Overreliance on these explanations can introduce automation bias, and there is a lack of standardized, scalable workflows for applying XAI to data-quality assurance. Practical adoption of XAI-driven remains rare, underscoring its status as an emerging method.

6.5 Data Attribution

Data attribution traces a model's predictions back to specific training samples. For each test prediction, it assigns scores to training samples based on their influence [32]. A high attribution value may indicate that removing sample i from the training set would likely result in j being misclassified. Conversely, a low (or negative) value implies that the presence of i contributes to the misclassification of j. Thus, data attribution can be interpreted as a matrix M, where rows represent training instances and columns test instances. Since such a matrix can become very large and interpretation challenging, it is common to display the top five samples with the highest positive influence and the top five

with the highest negative influence (see Figure 1 of the Appendix). Tools like TRAK [54] make the computation efficient by reducing the number of models that need to be retrained.

Opportunities: Data attribution detects label errors, data leakage, and underrepresented classes, addressing INHERENT and CONTEXT-DEPENDENT quality issues. It also highlights privacy risks, thereby addressing SECURITY concerns. Typically, mislabeled training samples show strong negative influence, while test instances with few positive influencers reveal underrepresented regions and potential security vulnerabilities.

Challenges: While the applicability of XAI has been widely discussed in ML research, data attribution remains underexplored and is not well known among researchers [50]. Although the applications mentioned seem plausible, they have not been systematically analyzed. Furthermore, interpreting the attribution scores can be challenging due to the complexity of the relationships between training and test samples.

6.6 Data Valuation

Data valuation is closely related to data attribution. In data valuation, each training instance i is assigned a scalar value v_i, indicating its influence on the model performance. The value of instance i can be interpreted as the average attribution score $v_i = \sum_{k=0}^{m} M_{i,k}$, where M is the attribution matrix from Section 6.5. Although data values can be computed from an attribution matrix, e.g., created by TRAK, it is currently more common to rely on sampling-based methods. Ghorbani and Zou [25] apply the Shapley value from cooperative game theory to data valuation. In recent years, many methods have been introduced to speed up the computation of data values, most of which aim to find better and faster approximations for the Shapley value [69, 39, 40, 35, 62].

Opportunities: By assigning each training example a value that reflects its impact on model performance, data valuation enables targeted dataset pruning, removing low-value or noisy samples to boost INHERENT data quality, and supports domain transfer by selecting high-value instances for context-specific tasks, thereby improving CONTEXT-DEPENDENT quality. It also underpins emerging data marketplaces by quantifying availability without exposing raw data, and highlights unique, high-value examples that may pose privacy or SECURITY risks.

Challenges: While data valuation addresses numerous challenges, more specialized methods often exist, and the computational costs are high. Efficiently calculating Shapley values, especially for large datasets, remains difficult.

6.7 Sample-size Estimation

Determining appropriate sample sizes for both test and training sets is crucial for reliable model evaluation and efficient data collection.

Several theoretical approaches have been proposed for test-set sizing. Guyon [31] suggest that the optimal fraction r reserved for validation (or test) should scale inversely with the square root of the number of model parameters $|\theta|$: $r = \frac{1}{\sqrt{|\theta|}}$.

An alternative formulation asks what absolute number of test samples n is needed to estimate error rates with statistical significance. Guyon et al. [30] gives the rule of thumb $n \approx \frac{100}{p}$, where p is the expected error rate of the best recognizer (e.g., a human). Additional formulations are discussed in Appendix D.

When test sets are small, high variance in performance estimates can undermine confidence. Bouthillier et al. [8] recommend running multiple evaluation trials, varying data order, initialization, etc., to stabilize metrics. Although the standard deviation decreases with larger sets, repeated runs can compensate when gathering more data is infeasible (see Figure 2 of the Appendix for an example).

For training-set sizing, simple heuristics (e.g. the "one-in-ten" rule of ten examples per parameter) lack solid justification. Instead, practitioners often plot learning curves, model performance versus fraction of training data, which typically exhibit logarithmic gains (Figueroa et al. 2012; Viering et al. 2021). Early additions yield large improvements, while returns diminish beyond a certain point (e.g., performance surpasses 90% at 50% of data, with less than a 5% gain from doubling in Appendix B).

Opportunities: Determining the appropriate test-set size ensures that performance estimates are both statistically sound and reflective of real-world conditions, thereby addressing CONTEXT-DEPENDENT quality. Estimating the amount of training data helps avoid unnecessary collection and annotation costs, addressing data AVAILABILITY. Finally, the total dataset size constitutes an INHERENT quality attribute of the data.

Challenges: In domains with scarce or costly data, gathering an ideal number of samples may simply be infeasible. Even when test or training sets are sufficiently large, models can still pick up spurious correlations. Finally, despite regulatory calls such as EU AI Act Art. 10, systematic and widely adopted methods for sample-size estimation remain underrepresented in practice.

7 An Exemplary Workflow for Regulated ML

Below we outline a concise five-step process that operationalizes the methods from before. A full treatment is beyond this paper's scope, but this sketch shows how established and emerging methods can fit into a practical pipeline.

1. **Define Objectives & Scope** Work with domain/compliance experts to enumerate use-case edge cases and data obligations and document the outcomes.
2. **Plan Acquisition & Splits** Estimate test-set size (e.g. $n \approx 100/p$ or $r = 1/\sqrt{|\theta|}$), collect roughly $5n$ samples, and stratify into train ($3n$), test (n), and holdout (n) sets.
3. **Validate Inherent Quality & Document** Run automated checks (e.g. CleanLab for label errors, schema validation). Generate a Data Card recording provenance, quality metrics, and corrections.
4. **Advanced Assessment & Remediation** On the holdout set, apply drift detection and memorization/attribution analyses (Sec. 6) to find coverage gaps. Remediate via targeted re-labeling, augmentation, synthetic data, or distillation.
5. **Final Audit & Continuous Monitoring** Merge the holdout back into training, finalize documentation, and archive an audit trail. Deploy live drift monitors and schedule periodic re-validation to maintain ongoing compliance.

8 Discussion & Outlook

In this paper, we introduced a regulation-aligned taxonomy of five ML data challenges, systematically mapped both established and emerging data-centric methods to these challenges, and sketched a five-step workflow for "regulatable ML." While these methods offer powerful techniques to enhance the quality, safety, and regulatory compliance of AI systems, they require further empirical validation and integration into practical toolchains and workflows. Future work should focus on validating these approaches in real-world settings and embedding them into end-to-end compliance processes. Finally, closer collaboration between regulators and the ML research community will be essential to refine regulatory mandates and accelerate the adoption of data-centric best practices.

References

[1] Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. From attribution maps to human-understandable explanations through concept relevance propagation, 2023. URL http://arxiv.org/pdf/2206.03208.

[2] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[3] Julius Adebayo, Michael Muelly, Hal Abelson, and Been Kim. Post hoc explanations may be ineffective for detecting unknown spurious correlation, 2022. URL https://arxiv.org/abs/2212.04629.

[4] Samah Baraheem and Zhongmei Yao. A survey on differential privacy with machine learning and future outlook, 2022. URL https://arxiv.org/abs/2211.10708.

[5] André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. Comprehensive exploration of synthetic data generation: A survey, 2024. URL https://arxiv.org/abs/2401.02524.

[6] Firas Bayram, Bestoun S. Ahmed, and Andreas Kassler. From concept drift to model degradation: An overview on performance-aware drift detectors. *Knowledge-Based Systems*, 245: 108632, 2022. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2022.108632. URL https://www.sciencedirect.com/science/article/pii/S0950705122002854.

[7] Claudia Beleites, Ute Neugebauer, Thomas Bocklitz, Christoph Krafft, and Jürgen Popp. Sample size planning for classification models, 1 2013. ISSN 00032670.

[8] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. Accounting for variance in machine learning benchmarks, 2021.

[9] Eric Breck, Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, and Martin Zinkevich. Data validation for machine learning, 2019. URL https://research.google/pubs/pub47967/.

[10] Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. The effects of data quality on machine learning performance. 7 2022. URL http://arxiv.org/abs/2207.14529.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://dx.doi.org/10.1613/jair.953.

[12] Xue-Wen Chen and Xiaotong Lin. Big data deep learning: Challenges and perspectives. *IEEE Access*, 2:514–525, 2014. doi: 10.1109/ACCESS.2014.2325029.

[13] Julien Colin, Thomas FEL, Remi Cadene, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 2832–2845. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/13113e938f2957891c0c5e8df811dd01-Paper-Conference.pdf.

[14] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms, oct 1998. ISSN 0899-7667. URL https://doi.org/10.1162/089976698300017197.

[15] Alexander Dubbs. Test set sizing via random matrix theory, 2024. URL https://arxiv.org/abs/2112.05977.

[16] Yotam Elor and Hadar Averbuch-Elor. To smote, or not to smote?, 2022. URL https://arxiv.org/abs/2201.08528.

[17] European Parliament and Council. Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016. URL https://data.europa.eu/eli/reg/2016/679/oj.

[18] European Parliament and Council. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act) (text with eea relevance), 2024. ISSN 52021PC0206. URL http://data.europa.eu/eli/reg/2024/1689/oj.

[19] Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2711–2721, June 2023.

[20] Dan Feldman. Core-sets: An updated survey, 2020. ISSN 19424795. - Name core set dates back to 2005.

[21] Vitaly Feldman. Does learning require memorization? a short tale about a long tail, 2020. ISSN 07378017. URL https://arxiv.org/abs/1906.05271. - Long Tail examples
- Mixture distribution.

[22] Benjamin Fresz, Elena Dubovitskaya, Danilo Brajovic, Marco Huber, and Christian Horz. How should ai decisions be explained? requirements for explanations from the perspective of european law, 2024.

[23] Benjamin Fresz, Vincent Philipp Göbels, Safa Omri, Danilo Brajovic, Andreas Aichele, Janika Kutz, Jens Neuhüttler, and Marco F. Huber. The contribution of xai for the safe development and certification of ai: An expert-based analysis, 2024. URL https://arxiv.org/abs/2408.02379.

[24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets, 2018. URL http://arxiv.org/abs/1803.09010.

[25] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning, 2019.

[26] Igor Goldenberg and Geoffrey I. Webb. Survey of distance measures for quantifying concept drift and shift in numeric data. *Knowledge and Information Systems*, 60:591–615, 8 2019. ISSN 02193116. doi: 10.1007/s10115-018-1257-z.

[27] Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, and Lingzhong Meng. A survey on dataset quality in machine learning. *Information and Software Technology*, 162, 10 2023. ISSN 09505849. doi: 10.1016/j.infsof.2023.107268.

[28] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

[29] Nitin Gupta, Hima Patel, Shazia Afzal, Naveen Panwar, Ruhi Sharma Mittal, Shanmukha Guttula, Abhinav Jain, Lokesh Nagalapatti, Sameep Mehta, Sandeep Hans, Pranay Lohia, Aniya Aggarwal, and Diptikalyan Saha. Data quality toolkit: Automatic assessment of data quality and remediation for machine learning datasets. 8 2021. URL http://arxiv.org/abs/2108.05935.

[30] I. Guyon, J. Makhoul, R. Schwartz, and V. Vapnik. What size test set gives good error rate estimates? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):52–64, 1998. doi: 10.1109/34.655649.

[31] Isabelle M Guyon. A scaling law for the validation-set training-set size ratio. 1997. URL https://api.semanticscholar.org/CorpusID:16194090.

[32] Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: a survey. *Machine Learning*, 113(5):2351–2403, March 2024. ISSN 1573-0565. doi: 10.1007/s10994-023-06495-7. URL http://dx.doi.org/10.1007/s10994-023-06495-7.

[33] ISO/IEC 5259. Artificial intelligence — data quality for analytics and machine learning (ml). Standard, International Organization for Standardization, Geneva, CH, 2023. URL https://www.iso.org/standard/81088.html.

[34] Johannes Jakubik, Michael Vössing, Niklas Kühl, Jannis Walk, and Gerhard Satzger. Data-centric artificial intelligence, 2024. URL https://arxiv.org/abs/2212.11854.

[35] Kevin Fu Jiang, Weixin Liang, James Zou, and Yongchan Kwon. Opendataval: a unified benchmark for data valuation, 2023. URL https://arxiv.org/abs/2306.10577.

[36] Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C. Mozer. Characterizing structural regularities of labeled data in overparameterized models, 2020. ISSN 2331-8422. URL http://arxiv.org/abs/2002.03206.

[37] V. Roshan Joseph. Optimal ratio for data splitting, 8 2022. ISSN 19321872.

[38] Frank Konietschke, Karima Schwab, and Markus Pauly. Small sample sizes: A big data problem in high-dimensional data analysis, 3 2021. ISSN 14770334.

[39] Yongchan Kwon and James Zou. Beta shapley: a unified and noise-reduced data valuation framework for machine learning, 2022. URL https://arxiv.org/abs/2110.14049.

[40] Yongchan Kwon and James Zou. Data-oob: Out-of-bag estimate as a simple and efficient data value, 2023. URL https://arxiv.org/abs/2304.07718.

[41] Guang Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Soft-label anonymous gastric x-ray image distillation. In *2020 IEEE International Conference on Image Processing (ICIP)*, page 305–309. IEEE, October 2020. doi: 10.1109/icip40778.2020.9191357. URL http://dx.doi.org/10.1109/ICIP40778.2020.9191357.

[42] Chi-Heng Lin, Chiraag Kaushik, Eva L. Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective, 2024. URL https://arxiv.org/abs/2210.05021.

[43] Michael A. Lones. How to avoid machine learning pitfalls: a guide for academic researchers, 2021. URL https://arxiv.org/abs/2108.02497.

[44] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. Machine learning for synthetic data generation: A review, 2024. URL https://arxiv.org/abs/2302.04062.

[45] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[46] Kristof Meding, Luca M. Schulze Buschoff, Robert Geirhos, and Felix A. Wichmann. Trivial or impossible – dichotomous data difficulty masks model differences (on imagenet and beyond), 2021. URL http://arxiv.org/abs/2110.05922.

[47] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning, 2022. URL https://arxiv.org/abs/1908.09635.

[48] Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. Measuring data, 12 2022. URL http://arxiv.org/abs/2212.05129.

[49] Stéphane Monteiro, Diogo Oliveira, João António, Filipe Sá, Cristina Wanzeller, Pedro Martins, and Maryam Abbasi. Data anonymization: Techniques and models. In José Luís Reis, Marisa Del Rio Araujo, Luís Paulo Reis, and José Paulo Marques dos Santos, editors, *Marketing and Smart Technologies*, pages 73–84, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-99-0333-7.

[50] Elisa Nguyen, Evgenii Kortukov, Jean Y. Song, and Seong Joon Oh. Exploring practitioner perspectives on training data attribution explanations, 2023. URL https://arxiv.org/abs/2310.20477.

[51] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 3 2021. URL http://arxiv.org/abs/2103.14749.

[52] Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels, 2022. URL https://arxiv.org/abs/1911.00068.

[53] Will Orr and Kate Crawford. Building better datasets: Seven recommendations for responsible design from dataset creators. *Journal of Data-centric Machine Learning Research*, 2024. URL https://openreview.net/forum?id=6bd8BrRKTW.

[54] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale, 2023. URL https://arxiv.org/abs/2303.14186.

[55] Maria Priestley, Fionntán O'Donnell, and Elena Simperl. A survey of data quality requirements that matter in ml development pipelines. *Journal of Data and Information Quality*, 15, 6 2023. ISSN 19361963. doi: 10.1145/3592616.

[56] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai, 4 2022. URL http://arxiv.org/abs/2204.01075.

[57] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners and open problems. volume 1, pages 417–423. Publ by IEEE, 1990. ISBN 0818620625. doi: 10.1109/icpr.1990.118138.

[58] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[60] Burr Settles. Active learning literature survey, 2010. ISSN 00483931.

[61] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models, 2017. URL https://arxiv.org/abs/1610.05820.

[62] Rachael Hwee Ling Sim, Xinyi Xu, and Kian Hsiang Low. Data valuation in machine learning: "ingredients", strategies, and open challenges. In *International Joint Conference on Artificial Intelligence*, 2022. URL https://api.semanticscholar.org/CorpusID:249319573.

[63] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 6 2022. URL http://arxiv.org/abs/2206.14486.

[64] Ahmad S. Tarawneh, Ahmad B. Hassanat, Ghada Awad Altarawneh, and Abdullah Almuhaimeed. Stop oversampling for class imbalance learning: A review. *IEEE Access*, 10: 47643–47660, 2022. doi: 10.1109/ACCESS.2022.3169512.

[65] Dipti Theng and Kishor K. Bhoyar. Feature selection techniques for machine learning: a survey of more than two decades of research, 3 2024. ISSN 02193116.

[66] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Comput. Surv.*, 55(8), dec 2022. ISSN 0360-0300. doi: 10.1145/3551636.

[67] Amy Turner, Meenakshi Kaushik, Mu-Ti Huang, and Srikar Varanasi. Calibrating trust in ai-assisted decision making, 2020. URL https://www.ischool.berkeley.edu/projects/2020/calibrating-trust-ai-assisted-decision-making.

[68] Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. The effects of explanations on automation bias. *Artificial Intelligence*, 322:103952, 2023. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2023.103952. URL https://www.sciencedirect.com/science/article/pii/S000437022300098X.

[69] Jiachen T. Wang and Ruoxi Jia. Data banzhaf: A robust data valuation framework for machine learning, 2023. URL https://arxiv.org/abs/2205.15466.

[70] Ruyu Wang, Sabrina Schmedding, and Marco F. Huber. Improving the effectiveness of deep generative data, 2023. URL https://arxiv.org/abs/2311.03959.

[71] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation, 2020. URL https://arxiv.org/abs/1811.10959.

[72] Jiaheng Wei, Yanjun Zhang, Leo Yu Zhang, Ming Ding, Chao Chen, Kok-Leong Ong, Jun Zhang, and Yang Xiang. Memorization in deep learning: A survey, 2024. URL https://arxiv.org/abs/2406.03880.

[73] Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review, 2023. URL https://arxiv.org/abs/2301.07014.

[74] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey, 2023. URL https://arxiv.org/abs/2303.10158.

[75] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching, 2021. URL https://arxiv.org/abs/2006.05929.

[76] Dora Zhao, Jerone T A Andrews, Orestis Papakyriakopoulos, and Alice Xiang. Position: Measure dataset diversity, don't just claim it, 2024.

[77] Dora Zhao, Jerone T. A. Andrews, Orestis Papakyriakopoulos, and Alice Xiang. Position: Measure dataset diversity, don't just claim it, 2024. URL https://arxiv.org/abs/2407.08188.

[78] Yuhan Zhou, Fengjiao Tu, Kewei Sha, Junhua Ding, and Haihua Chen. A survey on data quality dimensions and tools for machine learning. 6 2024. URL http://arxiv.org/abs/2406.19614.

[79] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning, 2020. URL https://arxiv.org/abs/1911.02685.

Appendix

A Data Attribution Example

Figure 1: Example of data attributions computed with TRAK. The leftmost image is the test sample (a cat). Above it are the five training images with the highest positive influence: adding them to the training set increases the model's confidence in correctly classifying the test image. Below are the five images with the strongest negative influence: their presence tends to reduce classification confidence. Some attributions can be counterintuitive (e.g. a dog image showing positive influence on the cat sample), highlighting challenges in interpreting influence scores.

B Removing Train and Test Data

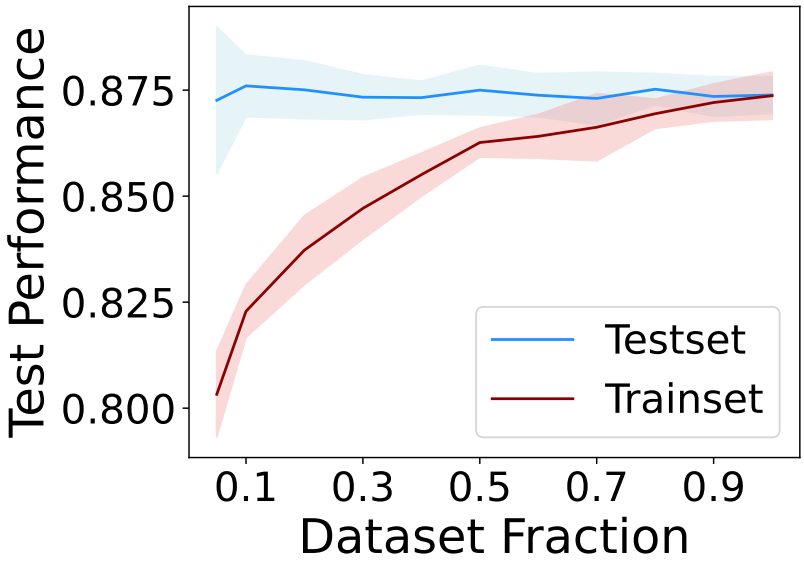

Figure 2: **Blue:** Test accuracy and deviation (y-axis) when training 10 models on full training data with increasing test size (x-axis). **Red:** Test performance and deviation (y-axis) with increasing size of training data (x-axis) on full test set.

C Data Quality Metrics from Research

Table 3: Examples of different data measurements from [48].

	DISTANCE	DENSITY	DIVERSITY	TENDENCY	ASSOCIATION
Physical Sciences	Length	Mass-per-volume	Biodiversity	Mean, Median, Mode	Correlation
General Data Measures	Euclidean Distance	Data Density	Gini Diversity	Burstiness	
	Cosine Similarity	KNN Density	Vendi Score		
	Earth Mover's Distance				
	Kullback-Leibler Divergence				
Modality-Specific Data Measures	Word Mover's Distance (language)	Information Density (language)	Text Diversity (language)	Perplexity (language)	Pointwise Mutual Information
	Levenshtein Distance (language)	Idea Density (language)	Lexical Diversity (language)	Fit to Zipf's Law (language)	
	Inception Distance (vision)	The Inception score (vision)	Image Diversity (vision)		
			Subset Diversity (vision)		

Table 4: Data Quality dimensions, metrics, descriptions, and examples from [78].

DIMENSION	METRICS	DESCRIPTION	EXAMPLES
Intrinsic	Correctness	A record in a dataset is free of errors.	Before starting a mailing campaign, the correctness of the attributes “postal code" shall be evaluated, and even small deviations shall be penalized because a deviation of only 1% (the postal codes 80000 and 79200) hinders the delivery of a mailing.
		Data is correctly labeled if it is a labeled record.	In the medical domain, an informal phrase of ‘’lack of feeling" should be labeled as “numbness".
	Duplication	Measures if the same instances repeat in the dataset, especially in both the training and test datasets.	If a record in a medical concept training dataset is “Hunger – don’t want to eat", and there is exactly the same record in a test dataset, then the record is considered as an overlapped record in the two datasets.
	Trustworthiness	Defines how factual the source that provides the information is. It can be subjectively evaluated, such as indicating the level on a scale, or the data can go through fact-check algorithms.	For a medical concept dataset, it should be obtained directly from the hospital’s system, which undergoes regular data quality checks and is maintained according to industry standards.
Contextual	Class imbalance	Evaluates if the distribution of examples across the known classes is biased or skewed.	Most of the contemporary works on class imbalance fall into the imbalance ratios ranging from 1:4 up to 1:100. The imbalance ratio may range from 1:1000 up to 1:5000 for extreme class imbalance problems.
	Completeness	A complete dataset should include as few missing values as possible.	A medical insurance dataset must include a customer’s birthdate, otherwise the medical consumption forecast model performance will be hindered.
	Comprehensiveness	A dataset contains all representative samples from the population.	In a medical text classification task, the training dataset should contain sufficient labeled medical texts covering all the conditions, symptoms, and treatments.
	Unbiasedness	Refers to whether the data used for machine learning training has a distribution bias or historical bias.	Photo recognition software does not recognize the facial expressions of ethnic minorities, or electronic soap dispensers that do not respond to darker skin tones because the training image datasets have an insufficient representation of some geographic regions.
	Variety	Requires each validation dataset and the test dataset to contain a significant amount of new data compared to the corresponding training dataset.	The percentage of the overlapped data between a test/validation dataset and its corresponding training dataset should be as low as possible, such as less than 10%.
Representational	Conformity	Measures how much the data conforms to the conventions for capturing information in a certain manner, including machine-readable data structures and formats for capturing specific attributes.	In a text classification task, a dataset of textual documents is labeled with sentiment (positive, negative, neutral). The labels should be encoded following a standardized set of categories, and all data processing, such as removing punctuation, converting text to lowercase, and tokenizing sentences, should be made to the whole dataset.
	Consistency	Requires data to be presented in the same format and to be compatible with previous data.	In an image classification task, if one dataset uses pixel values in the range [0, 255], while another dataset scales pixel values to the range [0, 1], this will cause inconsistency in model training and predictions.
Accessibility	Availability	High data availability ensures that data is readily accessible with defined user permissions for access and modifications.	In a healthcare ML application for diagnosing diseases from medical images, user entitlements are managed through strict access controls, allowing only authorized medical professionals and data scientists to access the images.

Table 5: Data Quality Metrics from ISO/IEC 5259 [33]. The standard distinguishes between inherent and system-dependent data quality metrics. Inherent metrics assess the intrinsic properties of the data itself, such as accuracy, completeness, and consistency. System-dependent metrics evaluate the data's quality within the context of its use in a specific system and involve availability, portability, and recoverability. A few metrics belong to both categories. Descriptions are generated with ChatGPT.

METRIC	DESCRIPTION	CATEGORY
Accuracy	The degree to which data correctly describes the "real world" object or event.	Inherent
Completeness	The extent to which data has no missing values.	Inherent
Consistency	Ensuring data is consistent and not contradictory across different datasets and systems.	Inherent
Credibility	The degree to which data is trustworthy and believable, often based on its source.	Inherent
Currentness	How up-to-date data is, depending on the intended use.	Inherent
Accessibility	The ease with which data can be accessed and retrieved.	Both
Compliance	Adherence to relevant standards, policies, and regulations.	Both
Confidentiality	Not provided in the standard.	Both
Efficiency	The extent to which data provides the expected level of performance.	Both
Precision	The level of detail and exactness of the data (e.g., decimal places in numerical values).	Both
Traceability	Not provided in the standard.	Both
Understandability	The ease with which data can be comprehended and used by stakeholders.	Both
Availability	Not provided in the standard.	System-Dependent
Portability	The ease with which data can be transferred and used across different systems.	System-Dependent
Recoverability	Not provided in the standard.	System-Dependent
Auditability	Part of data that has undergone an audit or is available for it.	NA
Identifiability	The capability to identify personally identifiable information in the dataset.	NA
Effectiveness	The degree to which data contributes to achieving the desired outcome or objective.	NA
Balance	Ensuring that the dataset is evenly distributed and representative of various groups.	NA
Diversity	The difference between the samples in the dataset.	NA
Relevance	The extent to which data is applicable and useful for the intended purpose.	NA
Representativeness	The degree to which data accurately reflects the broader population or phenomenon.	NA
Similarity	The extent to which data instances are similar to each other in terms of specified criteria.	NA
Timeliness	The latency between when the data is used and when it is available.	NA

Table 6: Recommendations for Reliability and Validity from Zhao et al. [76].

TOPIC	RECOMMENDATION	DESCRIPTION
Reliability	Inter-annotator agreement	An established method for assessing reliability, particularly in crowdsourcing, is through inter-annotator agreement. This method often entails multiple annotators labeling an instance, with the final label determined by a majority vote. Another method to gauge inter-annotator reliability is by employing statistical measures of agreement. We find that some text datasets provide quantitative metrics to quantify inter-annotator agreement, such as Fleiss's κ or Cohen's κ. While consensus methods are employed in both text and image datasets, quantitative metrics for inter-annotator agreement are reported exclusively in text datasets. We recommend that image dataset curators also incorporate these statistical measures when evaluating crowdsourced labels.
	Test-retest reliability	Another approach that dataset collectors can adopt is the test-retest method. In education, this method involves administering the same test twice over a period, with consistent results indicating reliability. This principle is particularly relevant when assessing the reliability of collection methods like web scraping. For instance, curators can reapply the same methodology to recollect instances, validating whether the recollected dataset maintains the same diversity properties. Nonetheless, a lack of reliability from these tests does not necessarily imply that the collection methodology inadequately captures diversity. Changes in the underlying data distribution over time can influence the results. For example, when evaluating linguistic diversity using data scraped from Reddit, major societal events, such as elections, can unexpectedly alter the distribution. Even in such cases, measuring test-retest reliability remains valuable for gaining insights into potential shifts in data distributions.
Validity	Convergent validity: Cross-dataset generalization	Commonly employed to evaluate "dataset bias," cross-dataset generalization enables researchers to compare datasets. By utilizing existing datasets with similar structures (e.g., label taxonomy, modality) and constructs of diversity, collectors can train on their dataset and test on existing datasets or vice versa, comparing relevant metrics such as accuracy. Model performance can also be assessed against standard train-test splits from the same dataset. If the models perform similarly in both cross-dataset and same-dataset scenarios, it suggests that the datasets have similar distributions for the target variable, indicating correlated constructs of diversity. Model performance can also be assessed against standard train-test splits from the same dataset. If the models perform similarly in both cross-dataset and same-dataset scenarios, it suggests that the datasets have similar distributions for the target variable, indicating correlated constructs of diversity. However, a constraint of employing cross-dataset generalization is the necessity for congruent taxonomies (for the target variable) and comparable distributions across datasets.
	Convergent validity: Comparing existing diversity metrics	Dataset collectors can leverage established metrics for measuring data diversity. For instance, the Vendi Score, drawing inspiration from ecology and quantum statistical mechanics, has been introduced as a measure of diversity within image and text dataset categories. Curators can demonstrate how their collection process aligns with such recognized diversity metrics. Given that diversity metrics depend on the embedding space employed, datasets should be benchmarked across a multiplicity of spaces optimized for the definition of diversity selected by the dataset curators.
	Discriminant validity	Discriminant validity assesses whether measurements for theoretically unrelated constructs yield unrelated results. Consider the initial Visual Question Answer (VQA) dataset, which aimed to collect diverse and interesting questions and answers, encompassing question types such as "What is ...", "How many ...", and "Do you see a ...". If diversity is defined by the types of questions asked, it should have no relation to other factors, such as gender distribution. Prior works identified language biases in how questions and answers are formulated in the VQA dataset. For instance, based on the dataset construction, a model predicting "Yes" whenever the question begins with "Do you see a ..." can achieve high accuracy without considering the image in question. This suggests potential low discriminant validity for the given measure, highlighting the importance of applying discriminant validity to mitigate construction biases during dataset creation.

D Suggested Split Rations for Train-Test Splits

Regarding split ratios, several works attempt to find theoretical justifications for optimal splits. Recently, Dubbs [15] derived ideal splits for linear regression models and independent Gaussian distributions. Guyon [31] suggest that *the fraction of patterns reserved for the validation set should be inversely proportional to the square root of the number of free adjustable parameters*. Let $|\theta|$ be the number of adjustable parameters in the model, then the optimal test ratio r is given by:

$$r = \frac{1}{\sqrt{|\theta|}} \ .$$

A similar formula is suggested by Joseph [37]. Instead of the free parameters of the model, they suggest using the number of parameters in a linear regression model that explains the data well. Let $|\beta|$ be the number of parameters in such a linear regression model, then the optimal ratio r is given by:

$$r = \frac{1}{\sqrt{|\beta|} + 1} \ .$$

Another approach is to reverse the question and ask *what size test set gives good error rate estimates* [30]. Guyon et al. [30] propose various formulas for this. The simplest form suggests that the number of test samples n should be approximately:

$$n \approx \frac{100}{p}$$

where p is the *expected error rate of the best recognizer*, i.e., a human. This approach is related to sample size estimation based on statistical significance, commonly used in medical and psychological research. For example, to evaluate the effect of medication against a certain disease and the target is an error margin of ϵ with a confidence level of $\sigma = 0.99$, the necessary sample size n is given by:

$$n \approx \frac{z_{0.99}^2 \times p(1-p)}{\epsilon^2} \ .$$

Here, p is the a-priori known occurrence rate of the disease, and $z_{0.99}$ is the value for the selected confidence interval derived from the normal distribution. This method can also be applied to machine learning and has been discussed in several works from the medical domain [7, 14, 57, 38]. In the context of ML, the interpretation could be as follows: if a model is trained to predict the disease with a randomly drawn test set of size n, there is a 99% confidence that the real-world model performance is within 1% of the error rate on the test set (for $\epsilon = 0.01$ and $\sigma = 0.99$).