[Data] Quality Lies In The Eyes Of The Beholder

Xavier Pleimling
Virginia Tech
Blacksburg, Virginia, USA
xavierp7@vt.edu

Vedant Shah
Virginia Tech
Blacksburg, Virginia, USA
vedant02@vt.edu

Ismini Lourentzou
Virginia Tech
Blacksburg, Virginia, USA
ilourentzou@vt.edu

ABSTRACT

As large-scale machine learning models become more prevalent in assistive and pervasive technologies, the research community has started examining limitations and challenges that arise from training data, *e.g.*, fairness, bias, and interpretability issues. To this end, data-centric approaches are increasingly prevailing over time, showing that high-quality data is a critical component in many applications. Several studies explore methods to define and improve data quality, however, no uniform definition exists. In this work, we present an empirical analysis of the multifaceted problem of evaluating data quality. Our work aims at identifying data quality challenges that are most commonly observed by data users and practitioners. Inspired by the need for generally applicable methods, we select a representative set of quality indicators, that covers a broad spectrum of issues, and investigate the utility of these indicators on a broad range of datasets through inter-annotator agreement analysis. Our work provides insights and presents open challenges in designing improved data life cycles.

CCS CONCEPTS

• **Theory of computation** → **Incomplete, inconsistent, and uncertain databases**; • **Information systems** → *Data analytics.*

KEYWORDS

datasets, data quality, data quality metrics, data utility, data annotation, incomplete data, inconsistent data, duplicate data, incorrect data, user survey

ACM Reference Format:

Xavier Pleimling, Vedant Shah, and Ismini Lourentzou. 2022. [Data] Quality Lies In The Eyes Of The Beholder. In *The 15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22), June 29-July 1, 2022, Corfu, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3529190.3529222

1 INTRODUCTION

Advances in hardware resources and the availability of large data quantities have allowed predictive modeling methodologies to be applied more broadly. The importance of Machine Learning (ML) and Artificial Intelligence (AI) has been intensified through the past years, especially after the recent breakthroughs in healthcare, genomics, robotics, climate change, *etc.* [Bhardwaj et al. 2017; Nambiappan et al. 2021; Rolnick et al. 2022; Volk et al. 2020]. As a result, the role of data has been growing rapidly as the foundational basis for training and evaluating machine learning models.

To produce reliable solutions, supervised machine learning typically depends on three data-related basic elements: volume, quality input and a good set of labels for the task at hand. Unsupervised methods can alleviate some problems related to data quality, however, they cannot match the effectiveness of supervised methods in all cases [Kim et al. 2020; Patacchiola and Storkey 2020]. There is a direct multiplicative relationship between input quality, label quality and training procedures. Any errors in the input or label space directly affect the learned model and consequently any derived insights. For example, without appropriate labels, ML models are unable to capture the task characteristics. In addition, high variance, data scarcity and noise render accurate modeling challenging and thus deteriorate the learned model's predictive performance. Most importantly, such challenges are highly intertwined with data curation, annotation and sharing, all of which can inform the kinds of problems data scientists and researchers often face in model development.

Data volume has been the driving force of the exponential growth of AI/ML in academic research. Collecting, cleaning and extracting useful features from data, however, is a time-consuming process that typically spreads over several years in order for models to reach or surpass human performance. For example, ImageNet [Deng et al. 2009] contains more than 1 million examples and spans over 1000 classes, while current language models are trained on terabytes of text data [Brown et al. 2020]. It is highly unlikely that these large data quantities can be reached in real-world scenarios in pervasive technologies, especially in domains with high expertise such as healthcare and human-robot manipulation where data annotation can quickly become very costly, or in new domains that are often faced with a cold-start problem. Even in the best case that unlimited and unrestricted data annotation is indeed possible, the issue may remain. Ensuring that labels and input variables are appropriate for the task at hand, such that incorrect decisions that can cause high monetary costs and other critical issues are avoided, remains of critical importance.

There are several research directions that address problems related - but not necessarily explicitly specific - to data quality and can be largely categorized on: (1) methods that improve quality and trustworthiness on already existing datasets [Asudeh et al. 2019; Bolukbasi et al. 2016; Kohler and Link 2021; Yoon et al. 2018], (2) methods that deal with data acquisition and either target the creation of (weak) annotations, or circumvent the need for labels by automatically labeling datasets or by utilizing unlabeled data [Ratner et al. 2017; Sheng et al. 2008; Tae and Whang 2021; Van Engelen and Hoos 2020], (3) methods that focus on improving the model or

PETRA '22, June 29-July 1, 2022, Corfu, Greece

ACM ISBN 978-1-4503-9631-8/22/06…$15.00
https://doi.org/10.1145/3529190.3529222

model training in a variety of scenarios, for example when there is class imbalance or data distribution mismatch between training and test data [Fuchs et al. 2021; Hendrycks et al. 2019; Seiffert et al. 2010; Tan et al. 2018], (4) works that identify good practices and data curation frameworks [Gebru et al. 2021; Sambasivan et al. 2021; Xin et al. 2018], and (5) methods that design formal quantitative metrics of data quality [Daimler and Wisnesky 2020; Heinrich et al. 2018; Jiang et al. 2009; Mishra et al. 2020; Raviv et al. 2020].

Despite the growing research on related areas, data quality remains an ill-posed concept. Most notably, no uniform definition of data quality or quality criteria exists. This is largely attributed to the fact that what is considered a high-quality dataset is highly subjective, and each dataset may be appropriate for specific use-cases but its quality may be insufficient for other purposes. The problem of ensuring data quality is therefore exacerbated by the multiplicity of potential problems with data. If a universal data quality metric existed, it would allow for empirical evaluation of data sources along several dimensions, *e.g.*, informativeness, bias, trustworthiness, information veracity, diversity, *etc.*. Such quality index would enable robust and generalizable evaluation of data pipelines, and hence greatly improve the the quality of the AI-based assistive technologies, reducing the prevalence of undesirable outcomes that arise from data bias, security and privacy issues.

Due to the importance of data quality in downstream applications, and the fact that data can impact model predictions in critical social and healthcare domains, *e.g.*, cancer treatment, stroke rehabilitation, law enforcement and surveillance, we present a study on data quality issues often encountered in practice. Our work is focused on understanding data users and identifying representative quality indicators that cover a broad spectrum of data quality issues, with the least possible assumptions. To this end, we define, identify, and present empirical evidence on the multifaceted problem of evaluating data quality. Moreover, inspired by the need for generally applicable methods to address data bottlenecks, we ponder whether there exists one universal metric of data quality and whether machine learning techniques can be leveraged to learn such metrics on a data-driven per-case basis. We hypothesize that this can only be achieved for data quality indicators that subject matter experts exhibit high inter-annotator agreement, *i.e.*, tasks with low subjectivity, and present an analogous study. Finally, we discuss future directions and opportunities in designing improved data life cycles. The contributions of our work can be summarized as follows:

- We present a qualitative study on data quality factors, that aims to uncover which issues are more frequently observed by data practitioners and what kind of properties high-quality datasets are expected to possess.
- Based on these observations, we define a set of data quality annotation dimensions that are distributed alongside a list of diverse datasets. This second part of the study aims at investigating which dimensions are highly subjective and to determine whether learning an aggregated data quality metric based on these annotations is indeed possible.
- Our experimental analysis shows that practitioners have a good sense of the most important data quality dimensions, but the beliefs as to whether a specific dataset is of high quality heavily depend on the data user and their perception of value.

2 RELATED WORK

There has been a significant amount of existing literature on research areas related to data quality. Below we review tangent areas of focus, as well as data quality directions, largely divided into user studies and proposed formulations of quantitative metrics for data quality.

Good data is key to good model development [Sambasivan et al. 2021]. Several recent studies, and the tech industry in general, have increased attention on data quantity as a pivotal factor in a model's projected success. The emphasis on data quantity, often referred to as "data volume", is in line with the notion that more abundant labeled data relates to a higher likelihood of learning diverse phenomena, which in turn leads to models that can generalize better [Swayamdipta et al. 2020]. However, data volume requirements have made it difficult to assess data quality [Cai and Zhu 2015; Swayamdipta et al. 2020]. Thus, data quality has become one of the most undervalued components of AI.

Early work deals with data cleaning and imputation for removing duplicates and substituting missing values [Lakshminarayan et al. 1996; Winkler 2004]. Many works target data valuation on a per-example basis [Ghorbani and Zou 2019]. More specifically, under the premise that not all examples in a dataset contribute equally towards the learning process of a model, related work designs data filtering or importance sampling strategies for AI training [Elvira et al. 2019; Katharopoulos and Fleuret 2018; Lourentzou et al. 2021; Ren et al. 2020; Robinson et al. 2021; Wang et al. 2021]. In addition, there has been a longstanding line of research on data annotation, in particular in active learning and crowdsourcing [Chang et al. 2017; Gal et al. 2017; Ho et al. 2015; Lourentzou et al. 2018; Settles 2010; Zhang et al. 2016]. Several studies also focus on data practices and pipelines for AI practitioners [Kandel et al. 2012; Xin et al. 2018]. Data documentation is another well-established area of research in the data management community [Bhardwaj et al. 2014; Buneman et al. 2001], that has recently attracted interest in machine learning as a means to produce data and model standards [Gebru et al. 2021; Hutchinson et al. 2021; Mitchell et al. 2019].

Applying AI/ML models in high-stakes domains such as loan allocation, healthcare, *etc.* requires that the model be built on quality data, due to the very nature of decisions that are made based on the outcomes produced by these models [Sambasivan et al. 2021]. As model performance heavily depends on the quality of the dataset, it is imperative that academia and industry start focusing on data quality as a key factor of a model's projected success and its significant impact on the effectiveness of a model built for real-world applications. While some work has been done in this area, there is a lot of work yet to be done and many questions yet to be answered. Research papers in this direction highlight certain aspects of data quality issues and provide some heuristics on how some of these issues can be solved using various statistical and non-statistical approaches [Cai and Zhu 2015; Pipino et al. 2002; Sambasivan et al. 2021; Swayamdipta et al. 2020]. In particular, Cai and Zhu [2015] discuss data quality challenges, identify common good practices and devise hierarchical quality standards for Big Data based on multifaceted quality indicators. Sambasivan et al. [2021] define data cascades as compounding events causing negative, downstream effects from data issues, resulting in technical debt over time, and explain how data cascades can have both short-term and long-term negative impacts.

In [Swayamdipta et al. 2020], the authors call attention to the problem that a large number of data models tend to fit the dataset distribution rather than the task, and introduce data maps (a model-based tool to characterize and diagnose datasets) as an attempt to resolve this issue. The authors categorize data points into three main regions/groups, *i.e.*, ambiguous, easy, and hard, that are observed from the data maps obtained based on model-dependent measures such as confidence and variability, and show that such regional data selection improves model generalization and can potentially speed up training. Moreover, Fenza et al. [2021] start with the assumption that the performance of an ML model heavily depends on the quality of the training dataset, which in turn relies on the consistency of labels assigned to similar items, and authors attempt to define a training data consistency measure for ranking problems, based on the consensus value introduced in group decision making. The main idea is to measure the consistency among similar input features with respect to their output and group data based on input characteristics to determine how coherent the outputs are. The authors also identify a statistical relationship between training data quality and the effectiveness of the resulting model.

In terms of designing metrics for assessing data quality, Mishra et al. [2020] implement data quality indices for natural language processing tasks and show how the proposed components and data visualizations can mitigate spurious correlations during data creation. Moreover, the authors showcase how the proposed data creation framework can improve data quality in a dynamic setting where new instances are added to a pre-existing set of samples. Schelter et al. [2018] present a data quality verification system that enables users to design 'unit tests' for data and combine them with readily available quality constraints. In addition, the authors present machine learning approaches for enhancing constraint recommendations, estimating column predictability and detecting anomalies in historic data quality time series. Other works try to mathematically define data quality and formally verify that data integrity is preserved during data transformations [Daimler and Wisnesky 2020; Jiang et al. 2009; Raviv et al. 2020]. However, most data quality measures are developed for ad hoc task-dependent settings.

In summary, existing studies focus on aspects of data quality in specific areas, such as NLP, Big Data, AI/ML, *etc.*, with a focus on understanding the challenges that practitioners face via interviews and surveys. Our work differs in that we ask the question of whether a general data quality indicator exists or whether such a metric can be learned. Albeit this research question lacks research attention, it can potentially establish general-purpose data quality assessment methods, *if* agreement on quality indicators could be achieved. We also highlight that, to the best of our knowledge, only a couple of works have focused on data quality in pervasive technologies [Hernández et al. 2017; Udoh 2020]. Our work extracts quality challenges from a broad set of datasets with diverse modalities. These datasets are often used for training computational models for critical technologies and applications, from IoT sensors, face recognition and object detectors, to chatbots, natural language understanding, *etc.*

3 METHODOLOGY

Between August and December 2021, we conducted semi-structured surveys with a total of 48 academic (student) practitioners located in the US (33 male and 15 female). All surveys are focused on defining and qualitatively measuring data quality aspects, and personally identifiable information was omitted when collecting responses. The study involves a cascade of two steps: (i) **`selection`** step, with a first questionnaire that determines the data quality challenges commonly faced in data analytics, and (ii) **`annotation`** step, in which participants provide per-dataset annotations for each of the identified data quality challenges.

For the **`selection`** step, participants are given a set of questions to determine what, in their opinion, are examples of high-quality datasets and common dataset quality problems. They are also asked whether they believe there can exist a universal metric for data quality, as well as why they believe so. In total, three specific questions are asked: **(1) What are some examples of data quality problems?**, **(2) What kind of properties does a high-quality dataset have?**, and **(3) Is there one universal metric of data quality?** The responses to each question are then analyzed, grouped or aggregated, and visualized. From this analysis, the most frequent data quality dimensions and challenges are identified, which are used later for the second part of the study, *i.e.*, the annotation step.

Participants are also asked to suggest datasets for the next step of this study. In total, 18 datasets are selected with varying sizes, formats and modalities. These include datasets that are suggested by the participants, such as the Waste Classification dataset [Sekar 2019], as well as datasets that are commonly used in machine learning research, such as the CIFAR-100 [Krizhevsky 2009], CelebA [Liu et al. 2015], and UCI Adult [Dua and Graff 2017] datasets. Concerning the modalities, the selected datasets include both image-based datasets and text-based datasets. All datasets are listed in Table 2. The selected datasets are distributed among the 48 participants, with each participant assigned two datasets. Consequently, each dataset is distributed to five or six unique participants for annotation.

In terms of the **`annotation`**, a set of questions is created for the participants to answer about their assigned datasets. The chosen questions are the outcome of the response analysis from the selection step. We design questions that are mostly objective and mainly focus on to what extent the participant agrees or disagrees with the most common data quality problems observed on the given data sets. This set of questions is answered on a scale with four options: (1) "Disagree", (2) "Mostly Disagree", (3) "Mostly Agree", and (4) "Agree". In summary, each participant is asked to answer 10 questions for each assigned dataset, and answers are provided on a 4-range scale. Results are then collected, visualized and analyzed via inter-annotator agreement, to evaluate which dimension related to data quality is easier to determine with respect to a specific dataset and whether an aggregated metric of data quality can be designed and learned. Intuitively, the higher the agreement between annotators, the more likely data quality can be approximated with a learnable function, and modeled with machine learning. In contrast,

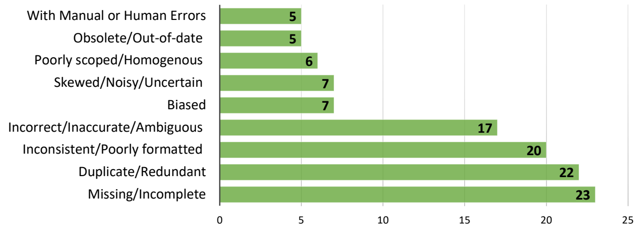

Figure 1: Bar chart presenting the most frequently encountered data quality problems.

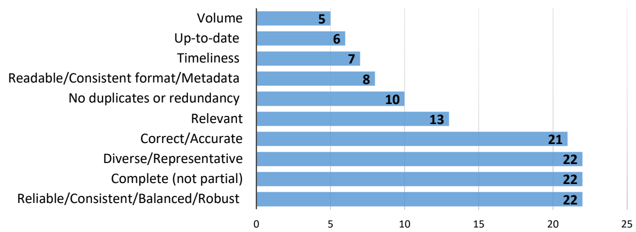

Figure 2: Bar chart presenting the most commonly suggested properties of high-quality data.

the lower agreement indicates high subjectivity and increased difficulty in learning a data quality function based on the aggregated annotations.

4 RESULTS

The initial survey reveals that the most commonly reported examples of data quality problems include missing and incomplete data, redundancy, and inconsistent data, as indicated by responses that are provided by at least 5 different participants (presented in Figure 1). Other common responses include poorly formatted data, ambiguity, bias, skewed distributions, noise, high uncertainty, obsolete data, inaccurate or unclear data, and human errors. In particular, problems related to missing, duplicate, inconsistent, and inaccurate data are frequently reported, with at least 17 participants reporting them. The initial survey (`selection` step) also reveals the most commonly reported properties that high quality data exhibit, as identified by at least 5 different participants. As shown in Figure 2, balanced, diverse, and reliable data, with complete information that represents the whole data population are considered in general of high quality by participants. Other properties involve data that have a consistent format and rich metadata available, no duplicates, as well as large-scale (at a proper volume) data that are up-to-date and become available in a timely fashion, corresponding to the "freshness" of data (otherwise termed in related fields as "age of information" [Yates et al. 2021]. Specifically, properties such as reliability, completeness, diversity, and accuracy are most commonly noted among participants, with at least 21 noting each of these properties.

In terms of whether a universal data quality metric can exist, more than half of the participants believe that data quality can

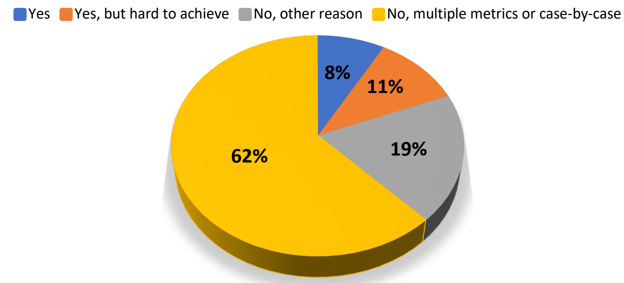

Figure 3: Pie chart representing participant answers to whether there can exist one universal metric of data quality.

be defined by multifaceted metrics that are used on a case-by-case basis (Figure 3). Some participants indicated that it may be possible to have a universal metric of data quality, but that it is also hard to achieve. The above results are used to create questions for the annotation of the selected datasets, presented in Table 1 (`annotation` step).

The responses (labels on a 4-range scale) from the annotation step are aggregated and an inter-annotator agreement analysis is performed. Krippendorff's Alpha [Krippendorff 2011] is used to determine the overall agreement for each dataset. Krippendorff's alpha can be computed as follows:

$$\alpha = 1 - \frac{\frac{1}{n}\sum_{c\in R}\sum_{k\in R}\delta^2_{ck}o_{ck}}{\frac{1}{n(n-1)}\sum_{c\in R}\sum_{k\in R}\delta^2_{ck}n_c n_k}, \tag{1}$$

where R is the set of all possible responses, δ denotes a metric function, typically $\delta_{ck} = \mathbb{1}(c = k)$ for nominal data, n denotes the total number of distinct ratings, $c \in R$ and $k \in R$ each denote the c^{th} and k^{th} distinct ratings, o_{ck} denotes the number of observed (c, k) pairs, and n_c and n_k denote the number of c and k values, respectively. An observed agreement metric is utilized to measure agreement for each question separately, computed as follows [McHugh et al. 2012]:

$$p_o = \sum_{i=1}^{4}\frac{r_i(r_i-1)}{r(r-1)}, \tag{2}$$

where r is the total number of raters and r_i is the number of raters that assigned the i-th rating out of the four possible ratings. Table 2 presents results for each of the questions found in Table 1. Each cell is color coded based on the typical interpretation cut-offs, *i.e.*, slight $(0 - 0.2)$, fair $(0.21 - 0.40)$, moderate $(0.41 - 0.60)$, substantial $(0.61 - 0.80)$ and perfect $(0.81 - 1)$ agreement.

Performing the inter-annotator analysis reveals that the per-dataset agreement was generally marked as fair. The datasets with the most agreed-upon responses were the Iris [Fisher 1936], the Lexnorm2015 [Baldwin et al. 2015], and the Wikipedia Toxicity [AI 2018] datasets, with alpha values of roughly 0.49, 0.48, and 0.43, respectively. Yet, the majority of the alpha values range between 0.1 to 0.3 which indicates rather poor agreement. In terms of individual questions, very few questions per dataset have a substantial agreement, such as Q9 for the Iris dataset. On average, Q5 has the

Table 1: Data Quality Annotation Questions

Index	Question
Q1	The dataset contains missing attributes, metadata, labels, etc.
Q2	There are significantly many duplicates.
Q3	There exists significant bias in the data.
Q4	The dataset can be used for modern machine learning problems and tasks (*i.e.*, the dataset is not outdated).
Q5	The dataset is easily accessible and usable (easy to download, easy to parse, standardized format, good organization).
Q6	The dataset is diversified with an appropriate scope (covering all cases).
Q7	The dataset is imbalanced or skewed.
Q8	The dataset is ethical (*i.e.*, cannot be used for malicious purposes, lack of privacy, etc.).
Q9	The dataset is properly annotated and does not contain human errors.
Q10	The dataset is versatile and useful for many downstream applications.

Table 2: Inter-annotator Agreement. First column shows the per-dataset Krippendorff's Alpha score, while the per-question observed agreement is presented in columns Q1-Q10. Cells are color coded based on the typical interpretation guidelines.

Dataset	α	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Avg.
Caltech101 [Li et al. 2003]	−0.03	0.20	0.20	0.20	0.20	0.10	0.20	0.60	0.40	0.10	0.30	0.25
Caltech256 [Griffin et al. 2006]	0.26	0.30	0.20	0.30	0.20	0.40	0.40	0.60	0.40	0.30	0.20	0.33
CelebA [Liu et al. 2015]	0.09	0.00	0.50	0.17	0.57	0.50	0.33	0.17	0.17	0.50	0.50	0.33
CIFAR-10 [Krizhevsky 2009]	0.13	0.17	0.17	0.00	0.50	0.50	0.17	0.00	0.50	0.17	0.50	0.27
CIFAR-100 [Krizhevsky 2009]	0.14	0.00	0.33	0.17	0.33	0.50	0.50	0.17	0.33	0.33	0.17	0.28
CORD-19 [Wang et al. 2020]	0.19	0.50	0.00	0.17	0.50	1.00	0.50	0.17	0.50	0.17	0.17	0.37
dsprites [Matthey et al. 2017]	0.12	0.40	0.13	0.20	0.20	0.27	0.40	0.20	0.27	0.13	0.27	0.25
FaceMask [Vrigkas et al. 2022]	0.25	0.13	0.67	0.20	0.40	0.27	0.40	0.20	0.13	0.20	0.20	0.28
IMDB-wiki [Rothe et al. 2018]	0.20	0.20	0.47	0.27	0.40	0.27	0.40	0.13	0.13	0.27	0.47	0.30
IOT-Temp [Purohit 2019]	0.27	0.40	0.13	0.20	0.13	1.00	0.13	0.13	0.67	0.27	0.20	0.33
Iris [Fisher 1936]	0.49	0.30	0.40	0.30	0.20	0.60	0.30	0.60	0.60	0.40	0.20	0.39
Lexnorm2015 [Baldwin et al. 2015]	0.48	0.27	0.27	0.47	0.27	0.27	0.40	0.20	0.40	1.00	0.27	0.38
MNIST [Deng 2012]	0.32	0.40	0.27	0.27	0.20	0.20	0.67	0.27	0.40	0.40	0.20	0.33
Tokio Olympics [Sarkhel 2021]	0.23	0.10	0.20	0.60	0.30	0.40	0.40	0.60	0.20	0.10	0.10	0.30
Superstore [Sahoo 2020]	0.17	0.27	0.40	0.27	0.13	0.40	0.27	0.27	0.27	0.27	0.13	0.27
Toxicity [AI 2018]	0.43	0.17	0.50	0.17	0.17	0.33	0.33	0.50	0.00	0.33	0.17	0.27
UCIAdult [Dua and Graff 2017]	0.22	0.40	0.30	0.20	0.30	0.10	0.20	1.00	0.40	0.20	0.30	0.34
Waste [Sekar 2019]	0.30	0.40	0.27	0.13	0.27	0.27	0.20	0.20	0.67	0.13	0.27	0.28
Average	–	**0.26**	**0.30**	**0.24**	**0.29**	**0.41**	**0.34**	**0.33**	**0.36**	**0.29**	**0.26**	–

highest agreement, and this could be attributed to the fact that accessibility of data is generally faster to determine as downloading and loading data is the first step before any data analysis, *e.g.*, Q2 or Q7. Most questions, however, have observed values of 0.5 or less, and the average observed agreement for each question, averaging across all datasets, ranges between 0.24 and 0.39, *i.e.*, fair agreement. The inter-annotator agreement results reveal that determining the quality of a dataset depends not only on the actual data but also on the dataset user and how they define the value of each dimension.

Overall, assessing data quality is highly subjective and relies on the perception and role of the data user. The survey responses indicate that participants are well-aware of a broad set of data quality problems. Nevertheless, our observations also show that developing a predictive model for each of the selected data quality dimensions, let alone learning a universal data quality metric, would be a challenging research direction, and any quantitative quality metrics may not necessarily align with the end-user perceptions of data quality. Despite the high subjectivity of the user study, our results are useful as a step towards formally defining metrics and best practices for data quality.

5 CONCLUSION

In this paper, we present the first effort toward investigating the learnability of data quality metrics. Through our study, we present a multitude of identified indicators and important data quality dimensions. Our analysis shows that most indicators remain subjective with respect to the user and task at hand. Thus, we conjecture that designing universal data quality metrics is a rather challenging task that would require multi-disciplinary approaches to integrate fundamental principles, and hope that the research community will target further work in this direction. We note that our study on data quality dimensions is not exhaustive; further research is required to include a comprehensive set of properties that pertain to data quality. In the future, we hope to design data quality metrics for healthcare domains and formally define the relationship between data quality, explainability and several types of data bias. Future research is needed to investigate how to design data quality metrics that align with the highly subjective data user perceptions of quality in pervasive computing.

REFERENCES

Jigsaw/Conversation AI. 2018. Toxic Comment Classification Challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

Abolfazl Asudeh, Zhongjun Jin, and H. V. Jagadish. 2019. Assessing and Remedying Coverage for a Given Dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. 554–565.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared Tasks of the 2015 Workshop on Noisy User-generated Text: Twitter Lexical Normalization and Named Entity Recognition. In *Proceedings of the Workshop on Noisy User-generated Text*. Association for Computational Linguistics, Beijing, China, 126–135.

Anant Bhardwaj, Souvik Bhattacherjee, Amit Chavan, Amol Deshpande, Aaron J Elmore, Samuel Madden, and Aditya G Parameswaran. 2014. DataHub: Collaborative Data Science & Dataset Version Management at Scale. *arXiv preprint arXiv:1409.0798* (2014).

Rohan Bhardwaj, Ankita R Nambiar, and Debojyoti Dutta. 2017. A Study of Machine Learning in Healthcare. In *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 2. IEEE, 236–241.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems* 29 (2016), 4349–4357.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-shot Learners. In *Advances in Neural Information Processing Systems*, Vol. 33. 1877–1901.

Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. 2001. Why and Where: A Characterization of Data Provenance. In *International Conference on Database Theory*. Springer, 316–330.

Li Cai and Yangyong Zhu. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data science journal* 14 (2015).

Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.

Eric Daimler and Ryan Wisnesky. 2020. Informal Data Transformation Considered Harmful. *arXiv preprint arXiv:2001.00338* (2020).

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-scale Hierarchical Image Database. In *2009 IEEE conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

Li Deng. 2012. The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Processing Magazine* 29, 6 (2012), 141–142.

Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

Víctor Elvira, Luca Martino, David Luengo, and Mónica F Bugallo. 2019. Generalized Multiple Importance Sampling. *Statist. Sci.* 34, 1 (2019), 129–155.

Giuseppe Fenza, Mariacristina Gallo, Vincenzo Loia, Francesco Orciuoli, and Enrique Herrera-Viedma. 2021. Data set quality in Machine Learning: Consistency measure based on Group Decision Making. *Applied Soft Computing* 106 (2021), 107366.

Ronald A Fisher. 1936. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 7, 2 (1936), 179–188.

Alexander Fuchs, Christian Knoll, and Franz Pernkopf. 2021. Distribution Mismatch Correction for Improved Robustness in Deep Neural Networks. arXiv:2110.01955 [cs.LG]

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian Active Learning with Image Data. In *ICML*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92.

Amirata Ghorbani and James Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *International Conference on Machine Learning*. PMLR, 2242–2251.

Greg Griffin, Alex Holub, and Pietro Perona. 2006. Caltech256 Image Dataset. (2006). http://www.vision.caltech.edu/Image_Datasets/Caltech256/

Bernd Heinrich, Diana Hristova, Mathias Klier, Alexander Schiller, and Michael Szubartowicz. 2018. Requirements for Data Quality Metrics. *J. Data and Information Quality* 9, 2, Article 12 (jan 2018), 32 pages.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. 2019. Using Pre-Training Can Improve Model Robustness and Uncertainty. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2712–2721.

Netzahualcóyotl Hernández, Luis A Castro, Jesús Favela, Layla Michán, and Bert Arnrich. 2017. Data Quality in Mobile Sensing Datasets for Pervasive Healthcare. In *Handbook of Large-Scale Distributed Computing in Smart Healthcare*. Springer, 217–238.

Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing High Quality Crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*. 419–429.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 560–575.

Lei Jiang, Daniele Barone, Alex Borgida, and John Mylopoulos. 2009. Measuring and Comparing Effectiveness of Data Quality Techniques. In *International Conference on Advanced Information Systems Engineering*. Springer, 171–185.

Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Enterprise Data Analysis and Visualization: An Interview Study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2917–2926.

Angelos Katharopoulos and François Fleuret. 2018. Not All Samples Are Created Equal: Deep Learning with Importance Sampling. In *International Conference on Machine Learning*. PMLR, 2525–2534.

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. 2020. Adversarial Self-supervised Contrastive Learning. In *Thirty-fourth Conference on Neural Information Processing Systems, NeurIPS 2020*. NeurIPS.

Henning Kohler and Sebastian Link. 2021. Possibilistic Data Cleaning. *IEEE Transactions on Knowledge and Data Engineering* (2021).

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. (2011).

A Krizhevsky. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront* (2009).

Kamakshi Lakshminarayan, Steven A Harp, Robert P Goldman, Tariq Samad, et al. 1996. Imputation of Missing Data Using Machine Learning Techniques. In *KDD*, Vol. 96.

Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato. 2003. Caltech101 Image Dataset. (2003). http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Ismini Lourentzou, Daniel Gruhl, Alfredo Alba, Anna Lisa Gentile, Petar Ristoski, Chad Deluca, Steven R Welch, and Chengxiang Zhai. 2021. AdaReNet: Adaptive Reweighted Semi-supervised Active Learning to Accelerate Label Acquisition. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*. 431–438.

Ismini Lourentzou, Daniel Gruhl, and Steve Welch. 2018. Exploring the Efficiency of Batch Active Learning for Human-in-the-Loop Relation Extraction. In *Companion Proceedings of the The Web Conference 2018*. 1131–1138.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. 2017. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/

Marry L McHugh et al. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica* 22, 3 (2012), 276–282.

Swaroop Mishra, Anjana Arunkumar, Bhavdeep Sachdeva, Chris Bryan, and Chitta Baral. 2020. DQI: Measuring Data Quality in NLP. *arXiv preprint arXiv:2005.00816* (2020).

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 220–229.

Harish Ram Nambiappan, Krishna Chaitanya Kodur, Maria Kyrarini, Fillia Makedon, and Nicholas Gans. 2021. MINA: A Multitasking Intelligent Nurse Aid Robot. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*. 266–267.

Massimiliano Patacchiola and Amos J Storkey. 2020. Self-Supervised Relational Reasoning for Representation Learning. In *NeurIPS*.

Leo L Pipino, Yang W Lee, and Richard Y Wang. 2002. Data Quality Assessment. *Commun. ACM* 45, 4 (2002), 211–218.

Tarun Purohit. 2019. Temperature Readings: IOT Devices. Relational Dataset from IOT Devices to Record Temperature Readings. https://www.kaggle.com/atulanandjha/temperature-readings-iot-devices

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid Training Data Creation with Weak Supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, Vol. 11. NIH Public Access, 269.

Netanel Raviv, Siddharth Jain, and Jehoshua Bruck. 2020. What is the Value of Data? On Mathematical Methods for Data Quality Estimation. In *2020 IEEE International*

Symposium on Information Theory (ISIT). IEEE, 2825–2830.
Zhongzheng Ren, Raymond A Yeh, and Alexander G Schwing. 2020. Not All Unlabeled Data are Equal: Learning to Weight Data in Semi-supervised Learning. *NeurIPS* (2020).
Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive Learning with Hard Negative Samples. In *ICLR*.
David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. 2022. Tackling Climate Change with Machine Learning. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–96.
Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2018. Deep Expectation of Real and Apparent Age from a Single Image Without Facial Landmarks. *International Journal of Computer Vision* 126, 2-4 (2018), 144–157.
Rohit Sahoo. 2020. Superstore Sales Dataset: Predict Sales Using Time Series. https://www.kaggle.com/rohitsahoo/sales-forecasting
Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone Wants to do the Model Work, not the Data Work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
Arjun Prasad Sarkhel. 2021. 2021 Olympics in Tokyo: Data about Athletes, Teams, Coaches, Events. https://www.kaggle.com/arjunprasadsarkhel/2021-olympics-in-tokyo/metadata
Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. 2018. Automating Large-Scale Data Quality Verification. *Proceedings of the VLDB Endowment* 11, 12 (2018), 1781–1794.
Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2010. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 40, 1 (2010), 185–197.
Sashaank Sekar. 2019. Waste Classification. https://www.kaggle.com/techsash/waste-classification-data/
Burr Settles. 2010. Active Learning Literature Survey. *University of Wisconsin, Madison* 52, 55-66 (2010), 11.
Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get Another Label? Improving Data Quality and Data Mining using Multiple, Noisy Labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, NV, USA) *(KDD '08)*. Association for Computing Machinery, New York, NY, USA, 614–622.
Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9275–9293.
Ki Hyun Tae and Steven Euijong Whang. 2021. *Slice Tuner: A Selective Data Acquisition Framework for Accurate and Fair Machine Learning Models*. Association for Computing Machinery, New York, NY, USA, 1771–1783.
Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A Survey on Deep Transfer Learning. In *International Conference on Artificial Neural Networks*. Springer, 270–279.
Emmanuel Sebastian Udoh. 2020. Is the data fair? An assessment of the data quality of algorithmic policing systems. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*. 1–7.
Jesper E Van Engelen and Holger H Hoos. 2020. A survey on semi-supervised learning. *Machine Learning* 109, 2 (2020), 373–440.
Michael Jeffrey Volk, Ismini Lourentzou, Shekhar Mishra, Lam Tung Vo, Chengxiang Zhai, and Huimin Zhao. 2020. Biosystems Design by Machine Learning. *ACS Synthetic Biology* 9, 7 (2020), 1514–1533.
Michalis Vrigkas, Evangelia-Andriana Kourfalidou, Marina E Plissiti, and Christophoros Nikou. 2022. FaceMask: A New Image Dataset for the Automated Identification of People Wearing Masks in the Wild. *Sensors* 22, 3 (2022), 896.
Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 Open Research Dataset.
Tianhao Wang, Yi Zeng, Ming Jin, and Ruoxi Jia. 2021. A Unified Framework for Task-Driven Data Quality Management. *arXiv preprint arXiv:2106.05484* (2021).
William E Winkler. 2004. Methods for evaluating and creating data quality. *Information Systems* 29, 7 (2004), 531–550.
Doris Xin, Litian Ma, Shuchen Song, and Aditya Parameswaran. 2018. How Developers Iterate on Machine Learning Workflows: A Survey of the Applied Machine Learning Literature. *arXiv preprint arXiv:1803.10311* (2018).
Roy D Yates, Yin Sun, D Richard Brown, Sanjit K Kaul, Eytan Modiano, and Sennur Ulukus. 2021. Age of Information: An Introduction and Survey. *IEEE Journal on Selected Areas in Communications* 39, 5 (2021), 1183–1210.
Jinsung Yoon, James Jordon, and Mihaela Schaar. 2018. GAIN: Missing Data Imputation Using Generative Adversarial Nets. In *International Conference on Machine Learning*. PMLR, 5689–5698.
Jing Zhang, Xindong Wu, and Victor S Sheng. 2016. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review* 46, 4 (2016), 543–576.