engineering proceedings

MDPI

Proceeding Paper

Quality of Labeled Data in Machine Learning: Common Sense and the Controversial Effect for User Behavior Models †

Maxim Bakaev * and Vladimir Khvorostov

Faculty of Automation and Computer Engineering, Novosibirsk State Technical University, Pr. K. Marksa 20, 630073 Novosibirsk, Russia; xvorostov@corp.nstu.ru

* Correspondence: bakaev@corp.nstu.ru

† Presented at the 15th International Conference “Intelligent Systems” (INTELS’22), Moscow, Russia, 14–16 December 2022.

Abstract: Intelligent systems today are increasingly required to predict or imitate human perception and behavior. In this, feature-based Machine Learning (ML) models are still common, since collecting appropriate training data from human subjects for the data-hungry Deep Learning models is costly. Considerable effort is put into ensuring data quality, particularly in crowd-annotation platforms (e.g., Amazon MTurk), where fees of top workers can be several times higher than the median. The common knowledge is that quality of input data is beneficial for the end quality of ML models, though quantitative estimations of the effect are rare. In our study, we investigate how labeled data quality affects the accuracy of models that predict users’ subjective impressions—per the scales of Complexity, Aesthetics and Orderliness assessed by 70 subjects. The material, about 500 web page screenshots, was also labeled by 11 workers of varying diligence, whose work quality was validated by another 20 verifiers. Unexpectedly, we found significant *negative* correlations between the workers’ precision and R^2s of the models, for two out of the three scales ($r_{11} = -0.768$ for Aesthetics, $r_{11} = -0.644$ for Orderliness). We speculate that the controversial effect might be explained by a bias in the indiligent labelers’ output that corresponds to subjectivity in human perception of visual objects.

Keywords: web interfaces; intelligent systems; machine learning; image recognition

check for updates

Citation: Bakaev, M.; Khvorostov, V. Quality of Labeled Data in Machine Learning: Common Sense and the Controversial Effect for User Behavior Models. *Eng. Proc.* **2023**, *33*, 3. https://doi.org/10.3390/engproc2023033003

Academic Editors: Askhat Diveev, Ivan Zelinka, Arutun Avetisyan and Alexander Ilin

Published: 9 May 2023

1. Introduction

One of the implicit assumptions in Machine Learning (ML) is that the data that get through the preliminary screenings and tweaks to the model training stage are appropriate. As for ML models that seek to predict or simulate human behavior, such as user behavior models (UBMs) in the field of Human–Computer Interaction (HCI), the situation is rather more sophisticated. The actual interaction-related data, which are generally the input of the predictive UBMs [1], arguably cannot be “bad”, as long as they reflect the human “imperfection”. However, there are also increasingly important subjective dimensions, from perceptional “how pleasant is our website design” in HCI to “how likely is it that you would recommend our service to a friend” in marketing. By definition, the subjective impressions are usually directly provisioned by human subjects—although indirect methods do exist, e.g., facial emotion recognition. Correspondingly, Deep Learning is slow to take off in this field, and an ample share of the models are feature based and rely on labeled data and the subjective assessments.

There is a general consensus that inaccurately annotated data are a hindrance and that the labeled data quality does not come for free. In micro-task platforms, such as Amazon Mechanical Turk (MTurk), filtering of crowdworkers can be carried out by a reputation that is principally based on the Approval Rate supplied by task requesters [2]. The fees charged by higher-paid workers are about four times above the *median* ones in MTurk [3], even though it has been shown that even top workers can be indiligent [4].

Reputation might have seemed an easy solution to crowd-labeled data quality a decade ago [2], but the arsenal of methods and tools has been rapidly expanding since then [5], as we subsequently outline in Section 2.1. The currently mainstream data quality control methods are *majority/group consensus* and *ground truth*, which necessarily imply redundancy (several workers performing the same task), wasting up to 33% of the output.

Even if data labeling work is carried out by volunteers and is technically free, their limited effort should be used efficiently too. Although volunteers generally have higher motivation than crowdworkers, redundancy might still be necessary to reach the certainty thresholds [6]. Setting the latter is actually a major problem for a requester, which we believe is not adequately covered in existing studies. Similarly to software debugging, more is always better, and there is no hard threshold to improving the quality of the data, only the one advised by practicability. Many developments to improve input data for UBMs, e.g., the enhanced version of the robust *Aalto Interface Metrics* (https://github.com/aalto-ui/aim, accessed on 1 June 2022) [7], are underway with the best intentions. Unfortunately, estimating the concrete "return on investment" in data quality remains problematic, as quantitative studies of its end effect in ML are scarce.

In our paper, we explore the relation between the completeness and precision of the input data produced by 11 human labelers and the quality of the ensuing 33 user behavior models built for 487 web page screenshots assessed by another 70 participants. Rather unexpectedly, we find that the significant correlation between the labelers' precision and the quality of the models constructed for the subjective scales of aesthetics and orderliness is **negative**. We attribute this preposterous result to the bias in indiligent labelers that brings their output closer to some subjective dimensions of human visual perception. We did not find any significant correlations for the labeling of completeness—even for complexity, which is known to be affected by the number of visual elements. Our results question the traditional data quality measures' applicability for human-related data, although further research is necessary.

The outcome has been preliminary reported and discussed at the *2021 Fall Conference of Ergonomic Society of Korea (ESK)*. In the current paper, we present the extended version of our results, referencing some of our previous related publications, such as [8,9]. In Section 2, we briefly review the research relevant to human behavior data quality in ML and describe our experiment. In Section 3, we construct the models and analyze the effects of the input data on their quality. In the final section, we discuss the findings and their possible causes, and outline directions for further research.

2. Method and Related Work

2.1. Data Quality Control in ML

As noted by philosophers long ago, the concept of *good* is very subjective. In relation to ML, it was recently demonstrated that the understanding of "good data" varies considerably for different stakeholders [10]. The concept of *quality*, though more objective and operational, is domain specific [11] and multi-dimensional [12]. With respect to data, it commonly involves the aspects of completeness, consistency, lack of duplicates, accuracy, timeliness for the purpose, and so on—some researchers identify as many as 20 dimensions. Since ML is predominantly concerned with *precision* and *recall* of the models, it associates data quality for the most part with *completeness* and *accuracy*.

The importance of these two dimensions in data quality was well recognized even before the ML era, and the related methodologies were classified as the ones helping "selection, customization, and application of data quality assessment and improvement techniques" [13]. Currently, the data quality control incorporates techniques for data collection planning, cleaning, profiling, evaluation, monitoring, etc. About a decade ago, strong focus in the field was established concerning crowd data, due to rapid advancement of crowdworking platforms such as Amazon MTurk (2005), microworkers.com (2009), Yandex.Toloka (2014), etc. The whole family of related meta-tools dedicated to data quality control emerged, such as CDAS (2012), Crowd Truth (2014), iCrowd (2015), DOCS for

AMT (2015), and others [9]. A comprehensive review of quality control in crowdsourcing can be found in [5], where the methods are organized into three major groups: individual, group and computation based. The former two generally imply involvement of humans in the assessment of the annotators or of the tasks' output.

It should be noted though that there has been a certain decrease in research enthusiasm towards crowd data since then, as the involved disadvantages had been acknowledged [4]. ML and Intelligent Systems came to rely more on unstructured and uncontrolled data sources [14], see Big Data [11,15] and data scrapped from the web [16]. A recent related publication carefully catalogs the software tools for data quality measurement and monitoring, listing a whopping 667 of them [17]. All in all, the quantitative engineering of data quality is better developed in the fields where data generation is easier to control. A recent example of such a field is IoT (see review in [18]), while the most established one is industrial data, where datasets are well structured and plentiful. Researchers in the field of industrial data quality already formulate it as a *dataset selection problem* and propose, e.g., the criteria of *estimated relative return improvement* and *estimated action stochasticity* [19]. However, those working with human-related data more often than not have no luxury of choosing between several datasets relevant to their specific problem.

2.2. Human Factor in Data Quality

The comparative rarity of reusing human behavior-related data outside of reproducibility and meta-analysis studies (e.g., [20] using the dataset from [8]) is partially due to its high value. The latter mainly comes from costly human time needed to generate or label the data, but its potential economic value may be involved too—think of social networks users' behavior data. Another reason that decreases the chance for finding appropriate data for a specific problem is that human data are a task too and context-dependent, and there is never a perfect match of factors and conditions. Moreover, their quality is arguably less formalizable on the scale from "good" to "bad", and the emerging concept of "fitness-for-use" [21] might prove to be more appropriate than "quality".

In the dawn of the AI/ML era, the human factor in data was rather considered a nuisance (cf. user needs in the era of mainframes). For instance, *10 reasons for bad data quality* comprehensively listed by Lee et al. in 2006, include "subjective judgments during data generation" [22]. Lately though, there is more recognition that human-related data are special, and specific quality dimensions are introduced, such as ethical ones [23]. The latter are arguably a response to the recently highlighted "inappropriate" behavior of trained AIs, who started to demonstrate "racist", "sexist", or "offensive" behavior [24]—just in accordance with the patterns they found in human-generated training data.

Still, the urgency of ML methods to describe human behavior is widely recognized, and UBMs that both incorporate domain knowledge and are trained on practical data is a popular implementation. The models' output are certain key performance characteristics—the examples in HCI field are success rate, time to complete a task, dimensions of subjective satisfaction, etc. The corresponding input data generally would need to specify the characteristics of the target users and the parameters of a candidate UI [1]. The techniques for parameterizing the UI can rely on manual labeling, on automated design mining algorithms (see in [25]), or on their combination [9]. Indeed, the algorithms for calculating the features are already numerous, and the developers put in a considerable effort in improving them [7]). However, the data quality studies in the field rather focus on adherence to "best practices" [26] and the reasons leading to "bad" models [27]. Quantitative studies of the data effect are rare, if any.

So, we undertook the following experimental study to relate the measured dimensions of the input data quality and the quality parameters for some simple UBMs.

2.3. The Experiment Description

2.3.1. Material

The material in our experiment was screenshots of website homepages belonging to universities and colleges from all over the world (but only their English versions). First, we automatically collected 10,639 screenshots in PNG format using a dedicated Python script crawling through various catalogs, DBPedia, etc. Then, we manually selected 497 of them for the experiment (see [8] for more detail)—hereafter, references as the UIs. To ensure better diversity of UI elements, the screenshots were made for full web pages, not just of the part above the fold or of a fixed size.

2.3.2. Procedure

UI Assessment

In a dedicated online survey (see details in [8]), the participants provided the subjective assessments of their impressions for each UI, per the three visual perception scales that we employed:

- How visually complex is the UI: *Complexity*;
- How aesthetically pleasant is the UI: *Aesthetics*;
- How orderly is the UI: *Orderliness*.

Complexity and aesthetics were elected as arguably the most popular dimensions in studies of subjective visual perception [20]. Orderliness was added mostly for the purpose of validity control of the assessments, as most studies in HCI are uniform about the positive correlation of UI regularity with aesthetics and the negative one with complexity. For each of three the scales, Likert ratings were used (1—lowest, 7—highest). The participants were instructed to provide their honest subjective assessments and were told that there are no right or wrong answers. The screenshots were randomly assigned to each participant successively, and the completeness of the assessment for all the 3 scales per UI was mandatory and controlled by the survey software.

UI Labeling

The labelers used LabelImg (Version 1.8.1, from https://github.com/tzutalin/labelImg, accessed on 1 June 2022), a third-party dedicated software tool that saves the output as XML files in PASCAL VOC format. They were asked to draw bounding rectangles around UI elements in the screenshots, as precisely as possible, and to choose one of the 20 pre-defined classes for the element: *image, background image, text, textinput, link, button, etc.* (see the complete list in [9]). The participants were provided with the written instruction on UI labeling and on technical usage of LabelImg and asked to process as many UI elements in each UI as possible. The screenshots were distributed among them near evenly, but no random assignment was performed.

The Labeling Verification

For each UI element in each screenshot, the verifiers could specify the labeling as *correct* or *incorrect*. In addition, for each UI the they were asked to subjectively assess completeness, i.e., if all the visible UI elements had been labeled, on the scale from 1 (very few elements) to 100 (all elements). The verifiers had the written instruction with recommendations for making the *correct/incorrect* decision, based on the UI elements' bounding box precision and the correct specification of its class. To support the verification procedure, we have developed a custom web-based software. The previously labeled UIs were distributed among the verifiers near evenly, but without a random assignment.

2.3.3. Subjects

There were 3 groups of human participants, mostly students of Novosibirsk State Technical University (NSTU), who performed the aforementioned activities:

1. *The UI assessment* was performed by 70 participants (43 females, 27 males), whose age ranged from 18 to 29 (mean 20.86, SD = 1.75).
2. *The UI labeling* was performed a few months later by another 11 participants (6 male, 5 female), with the age ranging from 20 to 24 (mean = 20.5, SD = 0.74).
3. *The verification* of the labelers' output was performed a few months later by yet another 20 participants (10 male, 10 female), whose ages ranged from 20 to 22 (mean = 21.1, SD = 0.45).

All the participants took part in the study voluntarily, and informed consent was obtained. They had normal or corrected to normal vision and reasonably high experience in the general usage of IT.

2.3.4. Design

The mean UI assessment ratings per the screenshots on the three scales of complexity, aesthetics, and orderliness (ScaleC, ScaleA, and ScaleO, respectively) were used as the *output variables* for the 3 × 11 = 33 user behavior models that we would construct for each scale and each labeler.

The *input data* for the models were 8 factors, whose values we automatically calculated for each UI from the labeling data, using our dedicated Python script:

1. number of all UI elements,
2. number of text elements,
3. share of the text elements' area in the screenshot,
4. number of image elements,
5. share of the image elements' area,
6. number of background image elements,
7. share of the background image elements' area,
8. share of whitespace (the screenshot area minus all the other labeled elements).

From the 20 labeled classes, we deliberately chose the most visually prominent ones: *text*, *image* and *background image*, since our experiment implied visual perception of the material, but no interaction with the UIs—hence, no *link*, *radiobutton*, *selectbox*, *textinput*, and so on.

So, our experiment had between-subject design. The main independent variables were *subjective completeness (SC)* and *Precision*, averaged for each of the 11 labelers:

$$Precision = \frac{correct}{correct + incorrect}. \tag{1}$$

The (derived) dependent variables were the quality parameters (R^2s) of the user behavior models. We also controlled for another derived variable, the number of screenshots processed by each labeler (UI).

Our **hypothesis** was that higher SC and precision, corresponding to better quality of the labeling data, should result in the better quality of the models (R^2s).

3. Results

3.1. Descriptive Statistics

In total, we collected 12705 assessments for the 497 UIs. Further, the 11 labelers specified 42,716 elements in 495 UIs (see [Table 1] in [9]), and the quality of their work was evaluated by 20 verifiers. Some UIs had technical problems or incomplete evaluations, so, we remained with 487 valid UIs (98.0%), for which the descriptive statistics are presented in Table 1. The first and second names of the labelers are abbreviated in the IDs.

Table 1. The descriptive statistics per the labelers (M ± SD).

UI Labeling			UI Assessment		
ID	**UIs**	**Elements**	**ScaleC**	**ScaleA**	**ScaleO**
AA	54	4802	3.67 ± 0.55	4.20 ± 0.86	4.43 ± 0.63
GD	44	3520	3.55 ± 0.56	4.07 ± 0.74	4.47 ± 0.53
KK	44	3927	3.33 ± 0.59	4.32 ± 0.71	4.59 ± 0.57
MA	44	5349	3.60 ± 0.63	4.02 ± 0.76	4.36 ± 0.56
NE	44	4994	3.57 ± 0.65	3.97 ± 0.90	4.34 ± 0.64
PV	43	4544	3.69 ± 0.74	4.34 ± 0.74	4.66 ± 0.58
PE	42	2569	3.69 ± 0.64	3.79 ± 1.07	4.16 ± 0.80
SV	43	3737	3.54 ± 0.63	4.22 ± 0.90	4.46 ± 0.68
ShM	41	1675	3.55 ± 0.71	4.05 ± 0.88	4.43 ± 0.56
SoM	45	3266	3.62 ± 0.73	4.25 ± 0.91	4.44 ± 0.68
VY	43	3630	3.47 ± 0.61	4.07 ± 0.83	4.52 ± 0.67
Total	**487**	**42,013**	**3.57 ± 0.64**	**4.12 ± 0.86**	**4.44 ± 0.64**

To check for the homogeneity of the UI assessments per the 11 labelers, we ran ANOVA tests for all three scales. We found a barely significant effect of ID only on ScaleO ($F_{10,476} = 1.87$, $p = 0.047$), but not on ScaleC ($F_{10,476} = 1.21$, $p = 0.284$) or ScaleA ($F_{10,476} = 1.63$, $p = 0.096$). The post-hoc test for ScaleO (Tukey HSD, since there were many levels of the independent variables) found significant difference (at $\alpha = 0.05$) only between labelers PV and PE ($p = 0.012$). The variances were not different ($p = 0.372$), so the ANOVA assumptions were met. Pearson correlations for the assessments per UIs were highly significant between ScaleA and ScaleO ($r_{487} = 0.771, p < 0.001$), as well as between ScaleC and ScaleO ($r_{487} = -0.145$, $p = 0.001$), but not between ScaleC and ScaleA.

In the verification, 37,053 labeled elements were specified as correct and 4967 as incorrect, and the mean Precision per labelers was 88.7%, which indicates a reasonably good work quality. The Pearson correlation between Precision and SC per labelers was not significant ($p = 0.727$), which suggests that these two aspects of UI labeling quality are distinct. The correlation between SC and the average number of correct objects was significant ($r_{11} = 0.622$, $p = 0.041$), unlike for the number of all labeled objects ($r_{11} = 0.170$, $p = 0.618$), which reinforces the meaningfulness of the verification.

3.2. The Effect of the Input Data Quality in the Models

To construct the UBMs, we relied on simple linear regression, since we only had a limited number of data samples (41–54) for each labeler. So, we built 33 models, each having the same 8 factors calculated from each labeler's output. The R^2s obtained for the models are presented in Table 2, together with the mean labelers' quality parameters obtained from the UI's verifications.

Since the number of screenshots processed by each labeler (UI) was not exactly the same (see in Table 1), we checked its correlations with R^2s for each of the three scales. We found that neither of the Pearson correlations was significant at $\alpha = 0.05$, so treating all labelers' models universally is justified.

The subsequent Pearson correlations analysis revealed that the SC did not have a significant correlation (at $\alpha = 0.05$) with the models' quality parameter (R^2) for either of the scales. Even for ScaleC, the correlation was $r_{11} = -0.062$ $(p = 0.856)$, whereas the visual complexity of a user interfaces is known to be influenced by the number of elements [25]. For the sake of checking the conceptual validity of our SC variable, we also checked the association between the factual *average number of elements per UI* for each labeler and the R^2s. Again, neither of the Pearson correlations were significant (at $\alpha = 0.05$), the correlation for ScaleC being $r_{11} = 0.274$ $(p = 0.415)$.

Table 2. The labelers' and the models' quality.

UI Labeling Quality			Models' Quality (R^2s)		
ID	**SC**	**Precision**	**ScaleC**	**ScaleA**	**ScaleO**
AA	73.0%	89.0%	0.108	0.149	0.114
GD	84.3%	89.9%	0.261	0.345	0.222
KK	82.5%	95.5%	0.261	0.252	0.152
MA	75.1%	72.0%	0.362	0.486	0.295
NE	78.3%	85.1%	0.316	0.488	0.416
PV	81.7%	91.6%	0.363	0.289	0.199
PE	72.0%	77.9%	0.165	0.568	0.611
SV	80.4%	97.4%	0.277	0.176	0.213
ShM	77.5%	89.5%	0.337	0.324	0.215
SoM	56.0%	95.9%	0.304	0.309	0.198
VY	95.5%	92.8%	0.204	0.110	0.169
Avg.	**77.8%**	**88.7%**	**0.269**	**0.318**	**0.255**

For precision, we found significant **negative** correlations with the R^2s for ScaleA ($r_{11} = -0.768, p = 0.006$) and ScaleO ($r_{11} = -0.644$, $p = 0.032$), but not for ScaleC ($r_{11} = -0.051$, $p = 0.883$). Recognizing the possible inaccuracy of our quality measures, we tried treating R^2 and the precision as ordinal variables—this is rather practical, since task requesters are often interested in only accepting the output from the best labelers. However, the results did not change very much for Kendall's tau-b correlation measure: $\tau_{11} = -0.491, p = 0.036$ for ScaleA and $\tau_{11} = -0.418, p = 0.073$ for ScaleO.

4. Discussion and Conclusions

Seeking to explore the effect of input data quality, we undertook an experimental study with 101 human participants and 497 web UIs. Our assumption was that better quality of the UI labeling should result in better quality of UBMs.

Contrary to our expectations, we found significant *negative* correlations between the labeling quality parameters and the resulting models' quality (see Table 2) for the subjective impression dimensions of aesthetics ($r_{11} = -0.768$) and orderliness ($r_{11} = -0.644$). Before deciding to report the negative research results in the current paper, we revisited the possible biases. However, the following considerations re-enforce the validity of our findings:

1. *Invalid UI assessment:* there was almost no significant difference in the distribution of the ratings per the labelers.
2. *Invalid UI labeling:* dimensions of precision (88.7%) and SC (77.8%) indicated high work quality and were distinct.
3. *Invalid Verification:* SC was correlated ($r_{11} = 0.622$) with the number of correct objects, but not with the number of all objects.
4. *Invalid subjective impressions scales:* as expected, ScaleA and ScaleO had significant positive correlation ($r_{487} = 0.771$), while ScaleC and ScaleO had significant negative correlation ($r_{487} = -0.145$). The relation between ScaleA and ScaleC was more controversial, as known from the literature [20], and we did not find a significant correlation.
5. *Imperfection in quality measurement:* we tried the objective measure for SC (elements per UI) and ordinal scale correlation (Kendall's tau-b) for Precision, but there were no major changes in the outcomes.
6. *Uncontrolled differences in the models:* the sample sizes varied from 41 to 45 (and even to 54 for one of the labelers), but there was no correlation between UI and the models' R^2.

The discovered negative correlations between the labelers' precision and the quality of the resulting models are not entirely clear to us, and we do not yet have a convincing explanation. We would like to note that the effect was found for the scale of aesthetics and the related scale of orderliness, but not for the less subjective scale of complexity. It is believed that aesthetics judgements for visual objects are rather high level, involving the factors of layout, visual hierarchy, colors, etc. Individual elements are grouped according to Gestalt principles, and imprecisions and omissions might even contribute to that—think of an Impressionist painting. Correspondingly, we might speculate that the indilligent workers would have a bias towards picking the UI elements and labeling them in a way matching the actual human perception. However, a much closer look at their output would be required before making any justified conclusions.

Among the limitations of our study, we see the relative minimalism of the linear regression UBMs. We only employed eight factors, and often they would not even be significant in the models. Correspondingly, the absolute quality levels of some models were rather modest, while the average R^2 per the 33 models turned out to be 0.281. The latter is arguably acceptable for our small-scale study that deliberately incorporated the potentially low-quality input data. For instance, in our another study with the same set of university websites screenshots, R^2s ranged from 0.105 to 0.248 (similarly, aesthetics had the highest R^2 of the three scales) [8]. However, recognizably, there the number of factors was smaller, and the number of samples was higher. In any case, in the current study we were interested in the relative values and never intended to use the models in production.

Our further research prospects involve experimentation with more labelers and a more diverse set of the web UIs. Having collected more data, we plan to employ artificial neural network (ANN) models, instead of the simple linear regression ones. ANNs are known as universal approximators and can naturally handle systematic bias in data.

Author Contributions: Conceptualization, M.B.; methodology, M.B.; software, V.K.; validation, M.B. and V.K.; formal analysis, M.B.; investigation, M.B.; resources, M.B. and V.K.; data curation, M.B. and V.K.; writing—original draft preparation, M.B.; writing—review and editing, M.B.; visualization, M.B.; supervision, M.B.; project administration, M.B.; funding acquisition, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by RFBR, grant number 19-29-01017.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of Faculty of Humanities of Novosibirsk State Technical University (protocol code 7_02_2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from M.B. (the first author). The data are not publicly available due to privacy reasons.

Acknowledgments: We would like to thank those who contributed to the project. Sebastian Heil, Martin Gaedke, Anna Stepanova and Galina Hamgushkeeva; as well as Alyona Bakaeva, who provided inspiration for this work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Oulasvirta, A. User interface design with combinatorial optimization. *Computer* **2017**, *50*, 40–47. [CrossRef]
2. Peer, E.; Vosgerau, J.; Acquisti, A. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behav. Res. Methods* **2014**, *46*, 1023–1031. [CrossRef] [PubMed]
3. Hara, K.; Adams, A.; Milland, K.; Savage, S.; Callison-Burch, Ch.; Bigham, J.P. A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–14.
4. Saravanos, A.; Zervoudakis, S.; Zheng, D.; Stott, N.; Hawryluk, B.; Delfino, D. The hidden cost of using Amazon Mechanical Turk for research. In *International Conference on Human-Computer Interaction*; Springer International Publishing: New York, NY, USA, 2021; pp. 147–164.
5. Daniel, F.; Kucherbaev, P.; Cappiello, C.; Benatallah, B.; Allahbakhsh, M. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–40. [CrossRef]
6. Salk, C.; Moltchanova, E.; See, L.; Sturn, T.; McCallum, I.; Fritz, S. How many people need to classify the same image? A method for optimizing volunteer contributions in binary geographical classifications. *PLoS ONE* **2022**, *17*, e0267114. [CrossRef] [PubMed]

7. Oulasvirta, A.; De Pascale, S.; Koch, J.; Langerak, T.; Jokinen, J.; Todi, K.; Laine, M.; Kristhombuge, M.; Zhu, Y.; Miniukovich, A.; et al. Aalto Interface Metrics (AIM): A service and codebase for computational GUI evaluation. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings, Berlin, Germany, 14–17 October 2018; pp. 16–19.
8. Boychuk, E.; Bakaev, M. Entropy and compression based analysis of web user interfaces. In *International Conference on Web Engineering*; Springer International Publishing: New York, NY, USA, 2019; pp. 253–261.
9. Heil, S.; Bakaev, M.; Gaedke, M. Assessing completeness in training data for image-based analysis of web user interfaces. *CEUR Workshop Proc.* **2019**, *2500*, 17.
10. Thakkar, D.; Ismail, A.; Kumar, P.; Hanna, A.; Sambasivan, N.; Kumar, N. When is Machine Learning Data Good? Valuing in Public Health Datafication. In Proceedings of the CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April–5 May 2022; pp. 1–16.
11. Gudivada, V.; Apon, A.; Ding, J. Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *Int. J. Adv. Softw.* **2017**, *10*, 1–20.
12. Wiemer, H.; Dementyev, A.; Ihlenfeldt, S. A Holistic Quality Assurance Approach for Machine Learning Applications in Cyber-Physical Production Systems. *Appl. Sci.* **2021**, *11*, 9590. [CrossRef]
13. Batini, C.; Cappiello, C.; Francalanci, C.; Maurino, A. Methodologies for data quality assessment and improvement. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–52. [CrossRef]
14. Bakaev, M.; Avdeenko, T. Intelligent information system to support decision-making based on unstructured web data. *ICIC Express Lett.* **2015**, *9*, 1017–1023.
15. Taleb, I.; Serhani, M.A.; Dssouli, R. Big data quality: A survey. In Proceedings of the IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, USA, 2–7 July 2018; pp. 166–173.
16. Bakaev, M.; Khvorostov, V.; Heil, S.; Gaedke, M. Web intelligence linked open data for website design reuse. In *International Conference on Web Engineering*; Springer International Publishing: New York, NY, USA, 2017; pp. 370–377.
17. Ehrlinger, L.; Wöß, W. A survey of data quality measurement and monitoring tools. *Front. Big Data* **2022**, *5*, 850611. [CrossRef] [PubMed]
18. Alwan, A.A.; Ciupala, M.A.; Brimicombe, A.J.; Ghorashi, S.A.; Baravalle, A.; Falcarin, P. Data quality challenges in large-scale cyber-physical systems: A systematic review. *Inf. Syst.* **2022**, *105*, 101951. [CrossRef]
19. Swazinna, P.; Udluft, S.; Runkler, T. Measuring Data Quality for Dataset Selection in Offline Reinforcement Learning. In Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, 5–7 December 2021; pp. 1–8.
20. Miniukovich, A.; Marchese, M. Relationship between visual complexity and aesthetics of webpages. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–13.
21. Jonietz, D. A concept for fitness-for-use evaluation in Machine Learning pipelines. Presented at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia, 6–14 December 2021.
22. Lee, Y.W.; Pipino, L.L.; Funk, J.D.; Wang, R.Y. *Journey to Data Quality*; The MIT Press: Cambridge, MA, USA, 2006.
23. Hagendorff, T. Linking Human And Machine Behavior: A New Approach to Evaluate Training Data Quality for Beneficial Machine Learning. *Minds Mach.* **2021**, *31*, 563–593. [CrossRef] [PubMed]
24. Ciarochi, J. Racist robots: Eradicating algorithmic bias. *Triplebyte Compil. Blog.* **2020**. Available online: https://triplebyte.com/blog/racist-robots-detecting-bias-in-ai-systems (accessed on 1 June 2022).
25. Bakaev, M.; Heil, S.; Khvorostov, V.; Gaedke, M. Auto-extraction and integration of metrics for web user interfaces. *J. Web Eng.* **2018**, *17*, 561–590. [CrossRef]
26. Geiger, R.S.; Cope, D.; Ip, J.; Lotosh, M.; Shah, A.; Weng, J.; Tang, R. "Garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *Quant. Sci. Stud.* **2021**, *2*, 795–827. [CrossRef]
27. Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; Aroyo, L.M. "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, Yokohama, Japan, 8–13 May 2021; pp. 1–15.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.