Training-Induced Class Imbalance in Crowdsourced Data

Shawn Ogunseye[1], Jeffrey Parsons[2] and Doyinsola Afolabi[3]

[1]Bentley University, Waltham, MA, USA
[2]Memorial University of Newfoundland and Labrador, St. John's, NL, Canada
[3] University of Lagos, Akoka, Lagos State, Nigeria

Abstract

In this paper, we examine how the design of data-collection systems can lead to imbalanced data. Specifically, we scrutinize how training affects the imbalance of data in a data crowdsourcing experiment. We randomly assigned contributors to explicitly trained, implicitly trained, and untrained (control) groups and asked them to report artificial insect sightings in a simulated crowdsourcing task. We posit that training contributors can lead them to selectively pay attention to and report specific aspects of observations while ignoring others. In the experiment, explicitly trained contributors reported less balanced data than untrained and implicitly trained contributors did. We then explored the effect of training-induced imbalance on an unsupervised classification task and found that the purity of classes formed was lower for explicitly trained contributors than for the other two types of contributors. We conclude by discussing the implications of artificial imbalance for the usefulness and insightfulness of crowdsourced data.

Keywords

Data Imbalance, Crowdsourcing Design, Crowd Knowledge, Data-driven insight

1. Introduction

Data crowdsourcing is one effective way that organizations can access information about a phenomenon of interest from willing human contributors. This method has been widely used to collect data across diverse domains, ranging from monitoring invasive species of mosquitoes that transmit diseases, such as dengue fever and the Zika virus (The Invasive Mosquito Project, 2017), providing guidance for consumer purchase decisions [1], and keeping track of outpatient health of seniors [2]. In data-crowdsourcing projects, information is collected from an undefined sample population, which is a source of concern for data consumers – organizations and individuals who gather input via crowdsourcing platforms – who wish to have the best data possible. Research has therefore sought to guide the design of data-crowdsourcing systems to ensure high-quality data is collected.

However, scholarship on data quality has centered chiefly on the representational quality of data – its capacity to adequately represent observed real-world phenomena. The consideration for what constitutes high-quality data (data fit for specific use) is limited, usually focusing on accuracy (the extent to which the information correctly represents observed real-world phenomena) and completeness (the extent to which it contains all the attributes of a phenomenon of interest) [3]. But data is increasingly repurposed to answer questions not anticipated when it was collected. There is, therefore, a need to better understand how to improve the quality and quantity of insights that data can provide [3], [4]. This makes understanding other intrinsic properties of data that can affect the usefulness of crowdsourced data – the quality of information derivable from an analyzed dataset – a critical success factor for data consumers.

VLDB 2021 Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale, August 20, 2021, Copenhagen, Denmark
EMAIL: sogunseye@bentley.edu (S. Ogunseye); jeffreyp@mun.ca (J. Parsons); dogunbiyi@unilag.edu.ng (D. Afolabi)
ORCID: 0000-0001-5774-4965 (S. Ogunseye); 0000-0002-4819-2801 (J. Parsons); 0000-0001-8442-7367 (D. Afolabi)

One intrinsic property of crowdsourced data that can affect its usefulness is ***class imbalance*** (also called ***data imbalance***) [5]–[7]. A balanced dataset is one in which the attributes needed to classify all instances of classes reported in the dataset are sufficiently proportional to allow classification algorithms to differentiate these instances into recognizable classes. When data is balanced, the instances that make up different classes in the dataset are evenly or equitably represented. The classification of the data is not skewed towards some instances to the detriment of others. Data imbalance is a pervasive problem usually originating from the world from which data is collected. The attributes of entities are not uniformly distributed in nature; some occur more frequently than others. For example, a dataset of fraud markers in financial transactions will show an imbalance because fraudulent transactions are much rarer than legitimate transactions. Likewise, markers in intrusion-detection datasets and in cancer-detection data will be imbalanced for the same reason – rareness of the target attribute or class.

Imbalance in data is a persistent problem in analytics [8]–[11], negatively affecting the accuracy of models in terms of regression [12], [13], classification (supervised classification) [14], clustering (unsupervised classification) [11], and artificial neural networks [15]. In many cases, the rarely occurring instances are deemed important, and their sparse occurrence implies algorithms may misclassify them into more prevalent classes. This is because machine-learning algorithms identify members of classes using models made from the broadest set of similar attributes of instances found in data. That is, they show a "maximum-generality bias" [16, p. 201], assuming the data has enough attributes of the relevant states of a phenomenon of interest. The use of the most commonly occurring attributes to form classes constrains the ability of classification algorithms to form classes from rarely occurring attributes. Instances that would have formed a minority class get subsumed into a majority class [9], [16].

Imbalance is assumed to be solely caused by the inherent nature of the observed phenomenon and mainly addressed after data has been collected using algorithmic and data-level strategies [15], [17] (referred to here as ***after-the-fact measures***). But imbalance may also be caused "artificially" [9, p. 1]. The design choices that data consumers make about data crowdsourcing projects can bias contributors to provide more or less balanced data. These decisions include how the crowdsourcing system is designed, the task(s) assigned to the sample population, who is recruited, and what motivates contributors [18].

Preventing imbalance, when possible, would benefit data consumers who must trust the effectiveness of after-the-fact measures when they do not have sufficient domain knowledge to know what classes are expected in their datasets. Without domain knowledge, or when data consumers do not know what insights may lie in a dataset, the potential for discoveries may be lost or significantly reduced as minority classes get subsumed [21]. Moreover, after-the-fact strategies introduce new problems that degrade a machine-learning model's performance, such as random over-sampling, which increases the likelihood of over-fitting as several exact replications of the minority class are added to the initial dataset [19], [20]. The resulting learning process takes more time as the dataset becomes larger**.** Algorithmic approaches may not accurately detect all the minority classes and their instances [10]. Increasing our understanding of how to prevent or mitigate artificial imbalance can improve the usefulness of crowdsourced data when machine learning is applied to them.

In this paper, we consider a design decision that can affect the imbalance of data in crowdsourcing projects – the decision to recruit knowledgeable contributors. Data consumers prefer to recruit knowledgeable contributors, but when these are scarce, they train novices to become more proficient in a crowdsourcing task. We investigate the effect of this design choice on the balance of contributed data. Through training, we induce the acquisition of task knowledge in contributors in a simulated crowdsourcing experiment and compare the level of imbalance in the data they provided. The study provides evidence that trained contributors report more imbalanced data than untrained contributors.

2. Knowledge and Data Contribution

Training contributors is a common design decision that data consumers make to improve the quality of crowdsourced data. However, this decision may affect the balance of data contributors' reports. Consider that humans are overloaded with sensory information every second. In a reporting task, such

as identifying an entity, we learn about objects by paying selective attention to the relevant features that aid in identification. Consequently, irrelevant features (those not helpful for determining class membership) are safely ignored. Selective attention is the cognitive process of attending to one or more sensory stimuli while ignoring others considered irrelevant to a task [21]. Although selective attention leads to efficient learning, especially when making connections between instances with few similar features, it comes with costs. The direct cost of selective attention is a learned inattention to features that are not relevant to a particular data-reporting task [22]–[24]. These features, however, may be critical for classification in another context, so a failure to capture them precludes the possibility of using these features in a different context [25].

Learning leads trained contributors to focus on relevant diagnostic features (i.e., for a specific task such as species identification), making them less likely to attend to non-diagnostic attributes than novices. We consider two forms of learning from the literature [23]: supervised learning – engendered by some form of explicit training (e.g., by a teacher) with sufficient feedback to improve the classifier's skill – and unsupervised learning – without explicit training (self-taught). Unsupervised or implicit learning may involve less rule-based processing and, consequently, more attentiveness to attributes, while supervised learning or explicit learning leads to a sharper focus on the acquisition of rules. Trained contributors will tend to selectively attend to only attributes they have been exposed to and learned to prioritize in training. Trained contributors will selectively attend to the few attributes they have learned, leading to disproportionate distribution of attributes in a dataset. Implicitly trained contributors may form different inclusion rules involving different attributes of the entity. The attributes they attend to and report will be influenced by the salience of features more so than the attributes reported by explicitly trained contributors. Meanwhile, explicitly trained contributors have a uniform set of attributes that they have been taught. They will focus mainly on those attributes or entities they have learned about and report attributes and entities that deviate from their existing knowledge to a lesser extent. As a result, explicitly trained contributors will report the most imbalanced data. Additionally, untrained contributors will attend to more attributes and consistently report these attributes about the entities they observe. We, therefore, predict that data from untrained crowds would be more balanced than data from trained crowds.

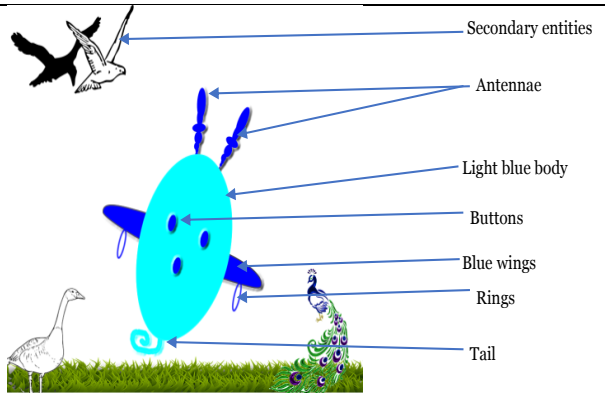
Proposition 1: *Untrained contributors will report more balanced data than implicitly or explicitly trained contributors.*

The consequence of imbalanced data is reduced ability to infer additional classes from data. We, therefore, predict that the design decision to train will negatively affect the usefulness of crowdsourced data, most significantly for the explicitly trained group and, to the least extent, the untrained group.

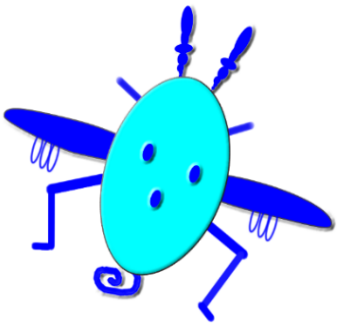
Proposition 2: *Data from untrained contributors will be more accurately classified using classification-based machine learning algorithms than data from trained contributors.*

3. Research Method

To test these propositions, we designed an experiment to simulate a data crowdsourcing task. These types of projects often seek knowledgeable contributors [26] and sometimes provide training to ensure they can provide the type and quality of data needed by scientists [27]–[29]. In addition, many data crowdsourcing projects are interested in discoveries [33]. Data crowdsourcing is, therefore, a proper context to test the impact of training on data imbalance. Following the example of [31], we designed an experiment with two classes of artificial insects: tyrans and nontyrans. We used artificial creatures as primary entities of interest to limit the effect of contributor prior knowledge on the study. We defined tyran as a class (species) of artificial insects whose members meet a classification rule consisting of five requirements: *(1) short tail, (2) light blue bodies, (3) two or three buttons on their light blue bodies, (4) blue wings, and (5) either one or two rings on each blue wing.* Similar artificial stimuli that do not satisfy this classification rule (i.e., they meet some, but not all, of the conditions) are nontyrans. Figure 1 shows a sample tyran and a sample nontyran used in the experiment.



Tyran with parts labeled



Nontyran: It has three rings on each wing. Note that the number of legs is not diagnostic.

Figure 1: Sample Tyran and Nontyran Images

The experiment consisted of twenty images, each presented on a separate slide. Sixteen slides (the test images) showed a mixture of tyrans and nontyrans. Four images containing unrelated items were placed intermittently within the sequence of test images to check whether participants paid attention to the task. These items were differently shaped/colored stimuli that were not insects (e.g., a triangle), and each participant was expected to report these stimuli correctly. The slides were presented in a nonrandomized order to all three groups, with images 5, 10, 15, and 20 showing items not related to the actual task. Each image had one primary entity and 13 of the 16 slides also included secondary entities. Secondary entities presented in the images are everyday objects, such as birds, insects, and fences. We displayed images of the entities in separate PowerPoint slides and asked contributors a nonleading question: "What do you see?"[2]

3.1. Task

We asked participants to imagine that designers of a game, similar to Pokémon Go, required their assistance in designing an aspect of the game. For example, the game requires players to interact with artificial insects. Specific insects called Tyrans are harmful and can kill a player's character, while similar insects called nontyrans, which lack some defining features, can provide energy to a player's character in the game. The designers, therefore, needed to test if the participants can report data about these insects to help improve their game. The participants were further informed that the goal of the experiment was to examine how people report the entities they observe.

Participants were issued data entry booklets in which to write down their observations. For each participant, individually numbered pages within the booklet had spaces to record their observation about a correspondingly numbered image slide that was projected on a screen. The prompt on each page of the booklet was, "What do you see?" Participants were also required to complete a demographic section after the study.

3.2. Participants

The 93 students who participated in this experiment were undergraduate students at a Canadian university. Students chose to receive either course credit or a donation to their class graduation, and each participant was entered into a draw for a campus bookstore gift card. After screening for completeness and the attentiveness of the contributor using embedded 'catch' items, responses from 84 participants were analyzed. Submissions from the remaining nine participants were excluded due to

[2] This is similar to the prompt used by eBird, a popular data crowdsourcing platform (www.ebird.org).

illegible writing, failure to report at least 3 out of 4 catch items correctly, and incomplete reports. Thirty-six participants identified as male and 48 as female.
We randomly assigned participants to three groups: (1) explicitly trained, (2) implicitly trained, and (3) untrained. The explicitly trained group members were taught the classification rule introduced above for identifying the primary entities as tyrans or nontyrans. To increase their familiarity with the task, participants in the explicitly trained group were also shown five sample tyrans, asked if they were tyrans, and given feedback on why these entities qualified as tyrans. We only showed participants images of tyrans because there are unlimited ways the attributes of a primary entity may violate the classification rules (here, we also follow [31], [32]). We briefed participants in the implicitly trained group on the task they would perform and showed them the same five target stimuli used to teach the explicitly trained group, one at a time, to allow them to elicit classification criteria. The participants were allowed to study each image; however, we did not provide explicit rules to members of this group, nor did we give them feedback on their ability to determine whether an entity is a tyran or not. Members of the untrained group were not shown any sample images. However, like those of the other groups, they were informed that we were interested in examining how people report information.

3.3. Measures

We developed a coding scheme that accounts for attributes of both the primary and secondary entities reported by participants. Two of the authors coded the first ten reports to establish consensus and conformance with the coding scheme. The first author coded the remaining reports, while the second author reviewed the coded data at different stages of the coding process. The variables coded for are presented in Table 1.

Table 1.
Variables Coded in the Contributed Data

Codes for Attribute types	Description
Behavior Attribute	The number of attributes describing the behavior of the primary entity
Mutual Attribute (primary entity)	The number of primary entity mutual attributes (a class of attributes that show an interaction between the primary entities and secondary entities)
Diagnostic Attribute	The number of attributes intrinsic to the primary entity that can be used to identify the primary entity
Non-diagnostic Attribute	The number of attributes intrinsic to the primary entity that cannot be used to identify the primary entity
Secondary Entity	The number of secondary entities reported
Diagnostic Attributes (Secondary Entity)	The number of attributes that can be used to identify the secondary entity
Mutual Attribute (Secondary Entity)	The number of attributes describing an interaction between the secondary entity and other entities
Behavior Attribute (Secondary Entity)	The number of attributes about the behavior of the secondary entity

Using the coded data about each attribute class, we analyzed the image data to see their attribute compositions. Eight classes of attributes are present in different proportions in the dataset. Our goal is to understand the probability they will be classified correctly using a classification type machine-learning algorithm.

3.4. Manipulation Check

Before testing for imbalance, we confirmed that the trained contributors exhibited selective attention to the primary entity and its attributes more than untrained contributors. Since secondary entities in the images were familiar objects (such as birds and fences), we examined the degree to which each group reported their presence as evidence of a differing degree of selective attention to primary entities. Applying one-way ANOVA to the data of the 13 images that included secondary entities, we found that the untrained contributors reported more secondary entities than did trained contributors. The implicitly trained group reported more secondary entities than did the explicitly trained group. The results indicate that those we expected to show more selective attention (i.e., the explicitly trained group) did so by reporting the least number of secondary entities (Table 2). This shows that our training was effective and selective attention indeed occurs at different levels in our groups.

Table 2.
Differences in the Reporting of Secondary Entities

A	B	mean(A)	mean(B)	Mean Diff.	Std. Err	T	p-value
E	I	1.036	1.544	−0.508	0.109	−4.674	0.001
E	U	1.036	2.289	−1.253	0.109	−11.521	0.001
I	U	1.544	2.289	−0.745	0.109	−6.847	0.001

E=Explicitly trained group, I= Implicitly trained group, U= Untrained group

3.5. Comparing Imbalance in Attribute Classes Between Datasets

We used only images containing secondary entities in our analysis. This is because some of the attributes coded for included Behavioral and Mutual Attributes for both the primary and secondary entities. Images with other entities provided a reference frame for contributors to report behavior and attribute types that involve more than one entity. We assume a case where mutual and behavioral attributes can give more insight into the observed entity than only attributes intrinsic to the entity. To compare the level of imbalance in the datasets collected by the three groups of contributors, we used the Shannon diversity index (H) – a mathematical measure of variability and can be used to compare entities in a specific space. Two different aspects contribute to the measurement of diversity: richness and evenness [30]. Richness is the total number of unique classes of a thing (attribute classes) in the data, and evenness is the distribution of the number of instances (attributes) for the available classes in the dataset.

The Shannon diversity index is denoted by the formula below:

$$H = \sum_{s}^{i=1} [(p_i) \times \ln(p_i)]$$

Where p_i = proportion of total sample represented by class *i* (divide number of instances belonging to each class *i* by the total number of instances). *ln* is the natural logarithm, $\sum$ is the sum of the calculations, and s is the number of classes.

Although H is sometimes used as an indicator of imbalance, it is most sensitive to the number of classes in an observation, so it is usually biased towards measuring class richness. Evenness (E_H) is a better measure of imbalance because it focuses on the distribution of instances (attributes) for the available classes, regardless of the number of classes available. $E_H = H / H_{max} = H / \ln S$ is measured as a number between 0 and 1, with 1 denoting perfect evenness.

- S = number of attribute types, = class richness
- H_{max}=ln(S) = Maximum diversity possible
- E_H = Evenness = H/ Hmax

In our experiment, each group observed and reported on the same thirteen images. We, therefore, compared the evenness in the number of attributes reported for each attribute class we expected. We calculated values of Shannon's equitability index (E_H) of each group across each of the 13 images. Using ANOVA, we compared the E_H for each group for all the images. There was no significant difference between the untrained and implicitly trained groups (U and I) with a p-value of 0.7989. However, there was a significant difference in the E_H values for explicitly trained contributors and the other groups. The explicitly trained contributors reported more imbalanced data than the implicitly trained and untrained contributors (see Table 3).

Table 3.
Comparison of the level of imbalance in the datasets

A	B	Mean A	Mean B	Mean Diff	F	p-value
E	I	0.594	0.755	-0.161	17.922	0.000
E	U	0.594	0.748	-0.154	15.914	0.000
I	U	0.755	0.748	0.007	0.066	0.799

E=Explicitly trained group, I= Implicitly trained group, U= Untrained group

Even though the Shannon equitability index gives insight into which dataset should be more balanced in theory, there is little empirical evidence to show that a high or low Shannon equitability index translates to more or less balance in real classification situations. There is a need to therefore examine the effect of the Shannon index score on actual classification of data. Given that we have different datasets about the same observation from the same number of people and under the same conditions, we have a unique opportunity to better understand how evenness translates to classification quality.

3.6. Relating Shannon Index to Classification Quality

We applied unsupervised classification to the data from the three groups. We clustered the data reported for each image using affinity propagation, an algorithm that automatically determines the number of clusters suitable for a dataset by optimizing the fit function [34]. Using the same model, the dataset from the explicitly trained group generated 122 clusters, while the dataset from the implicitly trained group generated 166 clusters. The dataset from the untrained group generated 171 clusters. Again, since these datasets were provided under the same conditions and within the same time limits, the key differentiating factor in determining the number of clusters formed was the level of training provided to contributors. The higher number of clusters may indicate a higher volume of data supplied by the contributors in the untrained group or a higher capacity for the model to cluster the data for the untrained group.

To better understand the effect of training on the quality of classification, we calculated the *purity* of the clusters formed from each dataset. Purity evaluation estimates the homogeneity of members of a cluster using human judgment and, therefore, estimates the degree of imbalance in a dataset [35]. To compute purity scores, we calculate the percentage of objects that are correctly and wrongly classified in a cluster based on "ground truths" or knowledge we have about the observed entities. Two people examined and coded attributes in the clusters for the number of right and wrong class members in each cluster. The coders included one of the authors and a student who was not briefed on the purpose or context of the study but presented the clusters and was asked to judge the suitability of the members of each cluster for class membership. The two coders achieved an interrater reliability score (Cohen's Kappa) of 0.86 without a need for discussion among the coders or any resolution. Then we applied the purity formula:

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j|c_i \cap t_j|$$

Where N = number of objects (data points), k = number of clusters, ci is a cluster in C, and tj is the classification which has the maximum count for cluster ci. We found that the clusters formed from data

from the untrained group were purer than clusters formed from the data for the trained groups. Also, the data from the implicitly trained group was as impure as the data from the explicitly trained group. Table 4 shows the purity scores compared using ANOVA.

Table 4:

Comparing the Purity of Classification of Resulting Datasets

A	B	mean(A)	mean(B)	diff	SE	T	p-value
E	I	0.789	0.830	-0.040	0.025	-1.638	0.232
E	U	0.789	0.895	-0.106	0.025	-4.265	0.001
I	U	0.830	0.895	-0.065	0.025	-2.627	0.025

E=Explicitly trained group, I= Implicitly trained group, U= Untrained group

From the purity evaluation, we found that attribute instances that would have formed minority classes are subsumed into majority classes more often in datasets provided by trained contributors than was the case for the dataset provided by untrained contributors.

4. Discussion

Decisions made in the design of data crowdsourcing systems can affect the balance of data. We examine one such design decision – training contributors in a data collection task. Training contributors can induce an imbalance in crowdsourced datasets because it leads contributors to focus mainly on relevant features they have learned. Trained contributors will selectively attend to attributes to which they have been exposed and have learned to prioritize in training. We see this in the results of our manipulation check in Table 2. Implicitly trained contributors reported more secondary entities than explicitly trained contributors, and both trained groups show more selective attention than the untrained group.

Training contributors has consequences for the balance of crowdsourced data reported. Untrained and implicitly trained contributors have a higher propensity to report balanced data than explicitly trained contributors. Training contributors with explicit rules rather than allowing them to learn on their own therefore has the most adverse effect on the balance of data. Data from the explicitly trained contributors also resulted in the most impure classifications, potentially limiting the insights that can be gathered from data.

Focusing on the effect of training on imbalance is unlike the majority of research on data imbalance which seek to address imbalance after data has been collected. Our approach is preventative, seeking to proactively design crowdsourcing systems to collect balanced data. One key value of this approach is that it emphasizes that imbalance is not solely inherent in phenomena but can be caused by our designs of data crowdsourcing systems and that some imbalance, particularly design-induced imbalance, is preventable.

Nonetheless, there are limits to the generalizability of our results. Indexes such as the Shannon Equitability Index give us an insight into how a design choice can affect the balance of data, but they may not accurately predict how a machine-learning algorithm will process data. Also, our hypotheses have not been tested in other conditions beyond this experiment.

Beyond these limits, emphasizing that imbalance can be artificial should prompt further research into preventing data imbalance where possible. Generally, future research on identifying how to design crowdsourcing systems so that they do not inadvertently promote the collection of imbalanced data would benefit data consumers. More specific to training, it would be helpful to understand how to mitigate the impact of selective attention on the balance of data when contributors need to be trained. Also, it would be interesting to know if contributors could be trained to report balanced data.

5. Conclusion

Imbalance is usually considered an inherent consequence of collecting data about things in the real world and has been mainly addressed after data has been collected. In this paper, we showed that

imbalance could also be caused artificially by the design choices made for data-crowdsourcing systems. The paper emphasizes the effect of training on the balance of data and the resulting usefulness of collected data. Design decisions, such as the choice to train contributors, can have negative consequences for the insightfulness of data because they encourage cognitive biases that limit the data contributed. Data requirements can change at several stages of a decision-making process – during collection or even after the initial analytics results come in. Thus, imbalanced data may support present known uses of data but fail to support emergent uses. How we design our data-crowdsourcing systems should therefore be determined by the priority we place on the insightfulness of our crowdsourced data, now and in the future.

6. References

[1] D. C. Edelman, "Branding in the digital age," *Harvard business review*, vol. 88, no. 12, pp. 62–69, 2010.

[2] S. Ogunseye, S. X. Komiak, and P. Komiak, "The Impact of Senior-Friendliness Guidelines on Seniors' Use of Personal Health Records," in *2015 International Conference on Healthcare Informatics*, 2015, pp. 597–602.

[3] S. Ogunseye and J. Parsons, "Designing for Information Quality in the Era of Repurposable Crowdsourced User-Generated Content," in *International Conference on Advanced Information Systems Engineering*, 2018, pp. 180–185.

[4] W. A. Günther, M. H. R. Mehrizi, M. Huysman, and F. Feldberg, "Debating big data: A literature review on realizing value from big data," *The Journal of Strategic Information Systems*, 2017.

[5] B. J. Hecht and M. Stephens, "A Tale of Cities: Urban Biases in Volunteered Geographic Information.," *ICWSM*, vol. 14, no. 14, pp. 197–205, 2014.

[6] J.-X. Liu, Y.-D. Ji, W.-F. Lv, and K. Xu, "Budget-aware dynamic incentive mechanism in spatial crowdsourcing," *Journal of Computer Science and Technology*, vol. 32, no. 5, pp. 890–904, 2017.

[7] Q. Xu, J. Xiong, Q. Huang, and Y. Yao, "Robust evaluation for quality of experience in crowdsourcing," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 43–52.

[8] R. Caruana, "Learning from imbalanced data: Rank metrics and extra tasks," in *Proc. Am. Assoc. for Artificial Intelligence (AAAI) Conf*, 2000, pp. 51–57.

[9] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: special issue on learning from imbalanced data sets," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, Jun. 2004, DOI: 10.1145/1007730.1007733.

[10] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, p. 27, 2019.

[11] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Prog Artif Intell*, vol. 5, no. 4, pp. 221–232, Nov. 2016, DOI: 10.1007/s13748-016-0094-0.

[12] T. Oommen, L. G. Baise, and R. M. Vogel, "Sampling bias and class imbalance in maximum-likelihood logistic regression," *Mathematical Geosciences*, vol. 43, no. 1, pp. 99–120, 2011.

[13] J. M. Snyder Jr, O. Folke, and S. Hirano, "Partisan imbalance in regression discontinuity studies based on electoral thresholds," *Political Science Research and Methods*, vol. 3, no. 2, p. 169, 2015.

[14] F. Thabtah, S. Hammoud, F. Kamalov, and A. Gonsalves, "Data imbalance in classification: Experimental evaluation," *Information Sciences*, vol. 513, pp. 429–441, 2020.

[15] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[16] Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358–3378, Dec. 2007, DOI: 10.1016/j.patcog.2007.04.009.

[17] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[18] T. W. Malone, R. Laubacher, and C. Dellarocas, "The collective intelligence genome," *MIT*

Sloan Management Review, vol. 51, no. 3, p. 21, 2010.
[19] N. V. Chawla, "Data mining for imbalanced datasets: An overview," *Data mining and knowledge discovery handbook*, pp. 875–886, 2009.
[20] D. D. Margineantu and T. G. Dietterich, "Bootstrap methods for the cost-sensitive evaluation of classifiers," 2000.
[21] G. Murphy and C. M. Greene, "Perceptual Load Affects Eyewitness Accuracy and Susceptibility to Leading Questions," *Front. Psychol.*, vol. 7, 2016, DOI: 10.3389/fpsyg.2016.01322.
[22] B. Colner and B. Rehder, "A new theory of classification and feature inference learning: An exemplar fragment model," in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2009, pp. 371–376. Accessed: Jan. 12, 2017.
[23] A. B. Hoffman and B. Rehder, "The costs of supervised classification: The effect of learning task on conceptual flexibility.," *Journal of Experimental Psychology: General*, vol. 139, no. 2, p. 319, 2010.
[24] S. Ogunseye, J. Parsons, and R. Lukyanenko, "Do Crowds Go Stale? Exploring the Effects of Crowd Reuse on Data Diversity," Workshop on Information Technology and Systems, Seoul, Korea 2017.
[25] O. S. Ogunseye, "Understanding information diversity in the era of repurposable crowdsourced data," Ph.D. Thesis, Memorial University of Newfoundland, 2020.
[26] A. Wiggins, G. Newman, R. D. Stevenson, and K. Crowston, "Mechanisms for Data Quality and Validation in Citizen Science," in *2011 IEEE Seventh International Conference on e-Science Workshops*, Dec. 2011, pp. 14–19. DOI: 10.1109/eScienceW.2011.27.
[27] U. Gadiraju, B. Fetahu, and R. Kawase, "Training Workers for Improving Performance in Crowdsourcing Microtasks," in *Design for Teaching and Learning in a Networked World*, vol. 9307, G. Conole, T. Klobučar, C. Rensing, J. Konert, and E. Lavoué, Eds. Cham: Springer International Publishing, 2015, pp. 100–114. DOI: 10.1007/978-3-319-24258-3_8.
[28] G. Newman *et al.*, "Teaching citizen science skills online: Implications for invasive species training programs," *Appl. Environ. Educ. Commun.*, vol. 9, no. 4, pp. 276–286, 2010, DOI: 10.1080/1533015X.2010.530896.
[29] F. I. Paez Wulff, "Recruitment, Training, and Social Dynamics in Geo-Crowdsourcing for Accessibility," 2014. Accessed: May 04, 2017. [Online]. Available: http://digilib.gmu.edu/jspui/handle/1920/9042
[30] R. Lukyanenko, J. Parsons, Y. F. Wiersma, and M. Maddah, "Expecting the unexpected: effects of data collection design choices on the quality of crowdsourced user-generated content," *MIS Quarterly*, vol. 43, no. 2, pp. 623–647, 2019.
[31] H. Kloos and V. M. Sloutsky, "What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories.," *Journal of Experimental Psychology: General*, vol. 137, no. 1, p. 52, 2008.
[32] S. Ogunseye, J. Parsons, and R. Lukyanenko, "To Train or Not to Train? How Training Affects the Diversity of Crowdsourced Data," International Conference on Information Systems, India, 2020.
[33] A. J. Daly, J. M. Baetens, and B. De Baets, "Ecological diversity: Measuring the unmeasurable," *Mathematics*, vol. 6, no. 7, 2018, DOI: 10.3390/math6070119.
[34] D. Dueck, *Affinity propagation: clustering data by passing messages*. Citeseer, 2009.
[35] W. Prachuabsupakij and N. Soonthornphisaj, "Cluster-based sampling of multiclass imbalanced data," *Intelligent Data Analysis*, vol. 18, no. 6, pp. 1109–1135, 2014.