Crowdsourcing for Repurposable Data: What We Lose When We Train Our Crowds

Shawn Ogunseye
University of British Columbia
shawn.ogunseye@sauder.ubc.ca

Jeffrey Parsons
Memorial University of Newfoundland
jeffreyp@mun.ca

Roman Lukyanenko
HEC Montréal
roman.lukyanenko@hec.ca

ABSTRACT

Users of crowdsourced data expect that knowledge of the domain of a data crowdsourcing task will positively affect the data that their contributors provide, so they train potential participants on the crowdsourcing task to be performed. We carried out an experiment to test how training affects data quality and data repurposability – the capacity for data to flexibly accommodate both anticipated and unanticipated uses. Eighty-four contributors trained explicitly (using rules), implicitly (using exemplars), and untrained, report the sighting of artificial insects and other entities in a simulated citizen science project. We find that there are no information quality or data repurposability advantages to training contributors. Trained contributors reported fewer differentiating attributes of entities and fewer total attributes of the entities they observed. Trained contributors are therefore less likely to report data that can lead to discoveries. We discuss the implications of our findings to the design of inclusive data crowdsourcing systems.

Keywords

Data Crowdsourcing, Information Diversity, Data Repurposability, Crowd Knowledge

INTRODUCTION

Crowdsourcing is a popular way of outsourcing tasks, normally done by an organization or by professionals, to an undefined, frequently online, group with varying levels of motivation and skill. One form of crowdsourcing is ***data crowdsourcing*** – engaging people to provide data, such as product reviews or wildlife observations (Lukyanenko and Parsons 2019; Ogunseye and Parsons 2018). Data crowdsourcing has been successfully used in many domains to, for example, understand customers, develop new products, improve service quality, and support scientific research (Castriotta and Di Guardo 2011; Hosseini et al. 2014; Tarrell et al. 2013; Tripathi et al. 2014).

As the composition of a crowd is often outside the control of those wishing to use crowdsourced data, there is limited ability to manage the data collection process and ensure the quality of data collected. One popular strategy to improve data quality in crowdsourcing is to recruit contributors who have the knowledge necessary to perform the data collection task (Surowiecki, 2005; Wiggins & He, 2016; Wiggins et al., 2011). When knowledgeable contributors are scarce or cannot be readily identified or targeted, this scarcity can be mitigated by training potential contributors to attain the desired level of proficiency (Yang et al. 2018). Training teaches contributors to provide information that is accurate and complete enough to be used for the immediate purpose for which a data collection task was designed.

However, crowdsourced data are also often used in ways not envisioned when the data were collected (e.g., Ballesteros et al., 2014; Bollen et al., 2011; Harrison et al., 2014). Data provided by trained contributors might not be readily ***repurposable –*** that is, usable for tasks other than those for which the data collection process was designed (Parsons and Wand 2014). One important characteristic that makes data repurposable is its ***diversity*** – the extent to which records in a data set contain information about different features of the observed phenomena (Ogunseye & Parsons 2018).

Diverse data contains multiple points of view, allowing it to meet the requirements of different users and different uses, even lead to discoveries (Ghasemaghaei & Calic, 2019; Parsons & Wand, 2014; Woodall, 2017). It might describe entities of interest in more detail or describe entities in addition to those at the focus of a data collection task. More diverse data is more repurposable than less diverse data. Repurposing data – which is a core element of data analytics (Woodall and Wainman 2015) – is widespread (Ransbotham and Kiron 2017). It is, therefore, beneficial to users of crowdsourced data to better understand how training affects the diversity of crowdsourced data.

Data crowdsourcing can lead to discoveries and new insights (Lukyanenko, Wiggins, et al. 2019). For example, new species of beetles were discovered by citizen scientists (Schilthuizen et al. 2017). Data crowdsourcing also led to the rediscovery of a butterfly species thought to be extinct (Lawrence 2015). In data crowdsourcing, contributors are the first to observe entities and choose what to report to data users; contributors determine the extent to which discoveries can be made. In this paper, we examine the effect of training contributors on the diversity of data collected in a crowdsourcing task. We conducted a lab experiment asking people to report sightings of artificial entities to test the effect of explicit (rule-based) training and implicit (exemplar-based) training on data diversity. We found that the diversity of data collected depends on whether and how we train crowd members.

Next, we describe the theoretical foundations for our study and propose several hypotheses about the effects of training on diversity. We then describe our experiment, present the results, and discuss the implications of our findings.

THEORETICAL FOUNDATION AND HYPOTHESES

To understand how training affects the diversity of crowdsourced data, it is important to understand how people learn, as learning is the desired outcome of training. An important element of human learning is categorization – the process of grouping instances based on their similarity. According to categorization theory, when humans seek to identify an entity, we observe its attributes and compare them with those of categories we already know (Goldstone and Kersten 2003; Harnad 2005; Piaget and Inhelder 1969; Rosch 1973). If the attributes of the new entity to match those of a known category, we classify the new entity as a member of that category. In general, this requires focusing on a subset of the observable attributes of the entity – namely, those needed to match those of a known category. When we are unable to match the observed attributes of an entity to those of a known category, we might create a new category for the entity and pay attention to the attributes of the entity that distinguish it from known categories (Katsuki and Constantinidis 2014; Wolfe 1994). Research on how infants, young children, and adults learn supports this reasoning. For example, infants (six to eight months) and young children who lack prior knowledge tend to pay attention to more of an entity's attributes than adults. Adults tend to pay attention to a few specific attributes of an entity needed to categorize it (Best et al. 2013; Gelman and Markman 1986; Kloos and Sloutsky 2008).

When observing an entity in the world, many attributes of both the entity and its surroundings compete for our attention at any one time. Given our limited cognitive resources, we tend to focus on a few critical attributes that help in categorizing the focal entity (Bjorklund and Harnishfeger 1990); in doing so, we tend to ignore attributes (and any other entities in our perceptual field) that are irrelevant to categorizing it (Prat-Ortega and de la Rocha 2018). This phenomenon is called ***selective attention***: "the differential processing of simultaneous sources of information" (Johnston and Dark 1986, p. 44). When we seek to classify an entity into a known category, our selective attention is knowledge-driven and allocated in a top-down manner (Katsuki and Constantinidis 2014), "derived from knowledge about the current task" (Buschman and Miller 2007, p. 1860). Our knowledge informs us about what attributes are important to categorize an entity and what attributes to look for (Wickens and McCarley 2008).

On the other hand, if adults encounter an unfamiliar entity (one not similar to known categories), we tend to notice attributes that stand out (Gopnik 2009). We do not have a basis to selectively attend to attributes and, thus, our attentional allocation is more bottom-up or stimulus driven. In bottom-up attentional allocation, "target stimuli 'pop out' if they differ sufficiently from their background in terms of features such as color or orientation" (Katsuki & Constantinidis, 2014, p 509). Bottom-up attention is driven by the salient attributes inherent in stimuli (Buschman & Miller, 2007), and by the cognitive effort required to search for attributes, rather than a predetermined strategy to focus on attributes relevant to a particular category. Salient attributes are attributes that are prominent in a contributor's visual space. Attributes of stimuli, such as their color, size, and shape, affect their capacity to attract an observer's attention (Theeuwes 2010) and are the default attention capture mechanism when the contributor has no prior knowledge or insufficient prior knowledge guiding their attention allocation (Buschman and Miller 2007; Katsuki and Constantinidis 2014).

Selective attention provides a suitable foundation for understanding the impact of training on data diversity in crowdsourcing. Trained contributors have the knowledge to categorize entities based on their attributes and are expected to apply selective attention in focusing on attributes relevant to categorization. Untrained contributors lack this knowledge and are expected to report observed attributes of entities, whether or not these are germane to a particular categorization.

Explicit training – teaching contributors the rules needed to identify or categorize an entity – enables contributors to apply top-down attentional control. It provides contributors with diagnostic attributes; that is, attributes that must be individually present for an entity to belong to that category and are collectively sufficient to categorize the entity. It leads contributors to look for specific attributes of an entity to be present when they observe it and leads them to focus on these expected attributes when categorizing the entity (Hoffman and Rehder 2010). In the absence of explicit training, learning can also take place by observation (implicit learning). Exposure to instances of a category allows contributors to infer diagnostic attributes by observing the similarities in attributes between members of the category to which they have been exposed (Rosch 1973). When learning implicitly, the learner attends to as many salient attributes as possible, which can lead to more attributes than necessary being learned, including some that are not diagnostic for the category.

Trained contributors are expected to selectively attend to the diagnostic attributes of a primary entity and will report more of these attributes than untrained contributors. However, untrained contributors are not expected to selectively attend to diagnostic attributes or a target entity. They are instead expected to report more attributes in aggregate about all entities in their visual field than are explicitly or implicitly trained contributors, implying greater diversity in the data they report. Implicitly trained contributors, who have been exposed to entities and their attributes, are expected to exhibit less selectively in attending to diagnostic attributes and are expected to report more diverse data than explicitly trained contributors.

> ***H1:*** *Untrained contributors will report more total attributes of observed entities than will implicitly trained contributors, who, in turn, will report more total attributes than will explicitly trained contributors.*

Training leads contributors to focus on the diagnostic (for categorization) attributes of entities. Knowledge of the diagnostic attributes to categorize an entity helps reduce the cognitive resources expended on identification tasks by reducing the attention allocated to attributes that are irrelevant to identifying the entity. These irrelevant, nondiagnostic attributes can be attributes inherent in the entity. There may also be attributes that inform us of the entity's actions (behavioral attributes) or its state in relation to other entities in its environment (mutual attributes). It is more cognitively economical for an explicitly trained contributor to focus on diagnostic attributes when observing an entity, and to ignore nondiagnostic attributes, adhering to the rules they have been taught. Implicitly trained contributors attempt to determine which attributes of the target entity are diagnostic by observing attributes common to all instances of a category to which they are exposed. Doing so is possible because people can learn to classify entities in an unsupervised way by studying the statistical frequency of entity attributes from repeated exposure to stimuli (Barlow 1989; Kloos and Sloutsky 2008); however, it is less efficient and less effective than explicit training. For implicitly trained contributors, the process of identifying diagnostic attributes entails first paying attention to the prominent attributes of the entity and then revising the mental list of relevant attributes with each exposure to an entity until they are confident about the valuable attributes and those that are irrelevant to a task. Implicitly trained contributors, who use a bottom-up approach to infer diagnostic attributes, are expected to attend to more attributes of a target entity while learning about the entity than explicitly trained contributors, including non-diagnostic attributes.

Implicitly trained contributors and untrained contributors use bottom-up attentional allocation, but implicit learners are already sensitized to the target entity via exposure to it during training. Implicitly trained contributors are expected to focus on the target entity they have learned about through exemplars when other entities are present in their visual fields. In comparison, untrained contributors are not expected to selectively attend to diagnostic attributes useful for categorizing an instance; instead, they will attend to the attributes of both primary and any other secondary entities in their visual field based on the salience of these attributes (and therefore the salient entities). Unlike implicitly trained contributors, untrained contributors are not sensitized to any one entity. We expect their attention to be more dispersed across their visual field and not limited to the target entity. Untrained contributors will, therefore, be less likely than implicitly trained contributors to report nondiagnostic attributes of a singular target entity when there are other entities present in their visual field. However, untrained contributors will report more nondiagnostic attributes than explicitly trained contributors who focus on diagnostic attributes corresponding to the rules they learned.

> ***H2a:*** *Implicitly trained contributors will report more nondiagnostic attributes of a target entity than will Untrained contributors who, in turn, will report more nondiagnostic attributes of a target entity than will Explicitly trained contributors.*

Since training helps implicitly trained contributors screen out attributes of an entity considered irrelevant to the data crowdsourcing tasks and prioritize attributes that frequently occur in observed exemplars, implicitly trained contributors are expected to report fewer non-diagnostic attributes that are not consistent across the exemplars or an inherent part of the exemplars of the target entity observed during training. These include attributes that describe the entity's behavior and state in relation to other entities and the entity's environment. Similarly, since attributes extrinsic to an entity, like its environment, cannot always be predicted, they may not be included in the rules taught to explicitly trained contributors, and hence would not be expected to be reported by explicitly trained contributors. However, untrained contributors, because of their lack of tendency to selectively attend to any specific attributes, are more likely to report the mutual and behavioral nondiagnostic attributes of a target entity.

> ***H2b:*** *Untrained contributors will report more mutual and behavioral attributes of a target entity than will trained contributors.*

Trained contributors who already know the target entity will be more likely to pay attention to it during a data reporting task, allocating less attention to other stimuli in their visual field. Because trained contributors are familiar with the target entity and its diagnostic attributes and view the tasks as identifying whether or not a focal entity belongs to a particular category, they will commit their cognitive resources to determine whether the target entity possesses these diagnostic attributes. When target entities are interacting with secondary entities in their environments, trained contributors will pay less attention to such interactions than will untrained contributors. For example, some mosquitoes are known to live in containers (Ritchie et al., 2003; Sprenger, 1987). In a reporting task about mosquitoes, trained contributors might focus on identifying the entity, and not mention the container. However, such information would be important to understanding if the mosquito is a Zika virus transmitting type[1], should there be a need to repurpose the data for such use.

Training will lead to learned inattention to secondary entities (Hoffman and Rehder 2010). Explicitly trained contributors know the "valuable" attributes of the entity and will concentrate on verifying their presence while ignoring other stimuli in their visual fields. Implicitly trained contributors who have learned diagnostic attributes through exemplars will also pay more attention to target entities they have been exposed to during training than to new entities. However, because they do not have to adhere to any classification rules, unlike explicitly trained contributors, their attention is expected to be more distributed over the visual space. Therefore, they are expected to report more information about secondary entities than explicitly trained contributors. Untrained contributors are expected to have the greatest tendency to pay attention to other entities when they are present in their field of vision because they are less task-directed and are more salience-driven than implicitly trained contributors. We expect that if any other entities in the contributor's visual field stand out, then an untrained contributor who has not been primed to focus on any entity will report information about those entities than trained contributors.

> ***H3:*** *Untrained contributors will report more data about secondary stimuli and their attributes than will implicitly trained contributors, who, in turn, will report more of these stimuli and attributes than will explicitly trained contributors.*

When contributors are trained using exemplars or rules, they expect the instances they encounter in the real world to either meet the rules or violate them. These rules typically consist of attributes and their values (e.g., red tail, where red is the value of the color attribute of the entity's tail). It is possible that not all differences in attribute values of instances of the phenomenon are accounted for in the rules they were taught or the exemplars to which they were exposed during training. The potential existence of instances whose attribute values (e.g., their color, number, shape or size) deviate from the values to which they were exposed in training is most prevalent when crowdsourcing for information about living organisms or phenomena where human knowledge is limited. For example, organisms (including plants and animals) are still undergoing speciation - forming new and distinct species due to evolution (Levin 2019; Ritchie and Immonen 2010; Wilkerson et al. 2015), and there is still so much we do not know about the universe. While classifying galaxies from images, a data contributor to the GalaxyZoo project – Hanny van Arkel – helped identify a "brand new type of astronomical object previously unknown to scientists" because she flagged some of its attributes as atypical; "it appeared as a blue squiggle" ("Hanny's Voorwerp – History of a Mystery" 2013). How training affects a contributor's ability to report attribute values that may lead to differentiation is pertinent in data crowdsourcing where discoveries are possible and welcomed. It is also important when incorrect classifications can have immense consequences, such as in ecological classifications (Wang et al., 2008).

Selective attention from training can limit trained contributors' reporting of differentiated attribute values in attributes for which they exhibit learned inattention or did not prioritize. Contributors, because of their training, may not attend to some attributes that are irrelevant to identifying an entity and thus not notice if that attribute is missing in a future instance. Also, contributors may have attended to an attribute but have not considered it pertinent to their task, and have not used it in their classification rule; thus, they might notice if the attribute is missing in a new instance of the entity, but not notice if the attribute is different (e.g., has a short tail, when all previous examples had a long tail) (Simons and Rensink 2005). When contributors encounter entities during a data crowdsourcing project, some attributes of these entities might possess unfamiliar attribute values. For example, two observed instances of a species of insect may deviate from a contributor's previously observed instances for the following reasons: (1) an instance does not have antennae;

[1] The Aedes aegypti mosquito, which transmits the Zika virus, is a "container-breeding mosquito" ("Zika, Mosquitoes, and Standing Water | | Blogs | CDC" 2016)

and (2) an instance has blue antennae instead of black. If the contributor did not attend to the antenna when they learned about the insect, they may not notice antenna are missing in case 1 and would not report it as a differentiated attribute value for that instance. If the contributor, however, attended to antennas but deciphered that it is an impertinent attribute for their task, they may not recall the color and would not report the blue color in case 2 as a differentiated attribute value.

Trained contributors selectively attend to a specific set of attributes, and their focus on these few specific attributes is expected to improve their capacity to notice and report differentiated attribute values when they occur in an observed instance. This effect will be stronger for explicitly trained contributors than for implicitly trained contributors. We expect that explicitly and implicitly trained contributors will report more differentiated attribute values affecting diagnostic attributes than will untrained contributors, because, unlike untrained contributors, they will allocate their attention mainly to the diagnostic attributes of the target entity.

> ***H4:** Explicitly trained contributors will report more differentiated attribute values involving the diagnostic attributes of a target entity than will implicitly trained contributors, who, in turn, will report more differentiated attribute values than will untrained contributors.*

Explicitly trained contributors will report fewer occurrences of differentiated attribute values than untrained contributors when the attributes affected are nondiagnostic. Unlike explicitly trained contributors, implicitly trained contributors will report more differentiated attribute values in nondiagnostic attributes, because they have learned these attributes while developing their classification rules. Untrained contributors will report fewer differentiated attribute values than implicitly trained contributors because they have not learned which attributes are consistent across instances and their attention is more dispersed across the visual field. Thus, they do not know what attributes of a target entity are usual or atypical.

> ***H5:** Implicitly trained contributors will report more differentiated attribute values involving the nondiagnostic attributes of a target entity than will explicitly trained contributors and untrained contributors.*

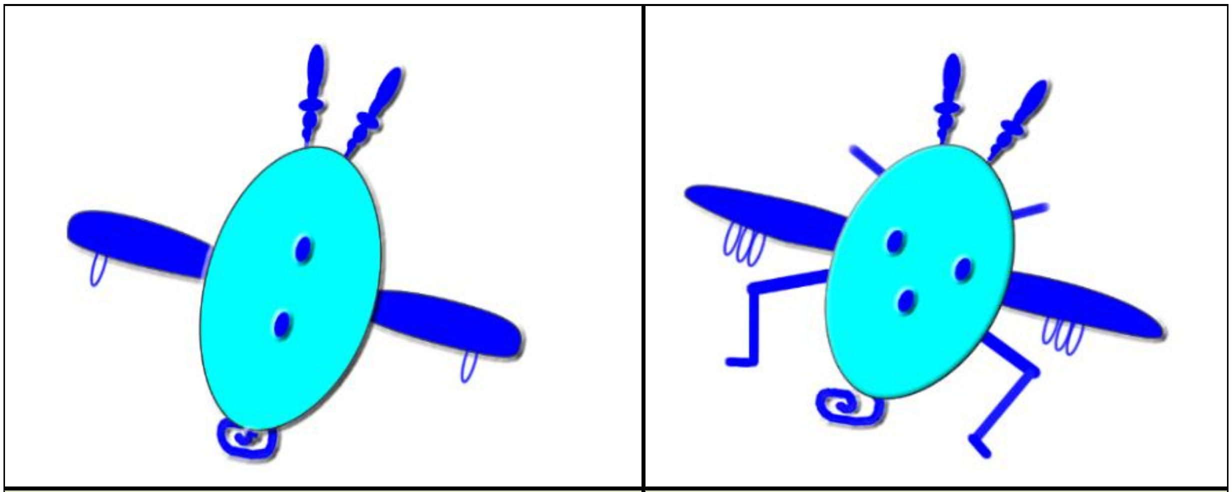
STUDY DESIGN

We designed an experiment in the context of citizen science, a type of crowdsourcing in which citizens contribute to data collection and/or analysis, defining the research question, or even designing a study while gaining scientific knowledge through their involvement in the research. Citizen science projects often seek knowledgeable contributors; “most projects show greater concern over the lack of contributor expertise than the lack of analysis methods suited to the type of data generated in citizen science” Wiggins et al. (2011, p. 17). Many citizen science projects are interested in discoveries (Lukyanenko, Parsons, et al. 2019). It is, therefore, a suitable context to test the impact of training on information diversity.

The target artificial stimuli used in this study are called tyrans. These stimuli were designed following Kloos and Sloutsky’s (2008) artificial stimuli. Tyrans are defined as a class (species) of artificial insects whose members meet a classification rule (a set of attributes and values of these attributes). Stimuli that do not meet this rule are nontyrans. The classification rule consists of five requirements: *tyrans have (1) a 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.* Nontyrans may share various attributes with tyrans but will fail to meet at least one of the predetermined diagnostic requirements. Each image was presented to participants in a separate PowerPoint slide. Figure 1 shows a sample tyran and a sample nontyran used in the experiment.

We tested the experimental materials with 12 ecology students who are familiar with observing and reporting organisms in the field. We tested for the suitability of the prompt to elicit unbiased responses from contributors. We found that asking contributors a nonleading question, i.e., “what do you see?”[2] was less biasing than asking them to identify the entity they observed. We also tested for the complexity of the task and the ease of learning the classification rule to ensure that the number of attributes in the classification rule did not make the task too complex for the participants. From the results, we simplified the classification rule to consist of five attributes of the target entity. We conducted another pretest to examine the effect of the changes made based on our initial pretest. Fifteen business students participated in the second pretest, which identified the appropriate duration of the reporting tasks and confirmed the effectiveness of the changes made due to the first pretest. All participants recorded their sightings on an answer sheet. Based on our findings from the pretests, we set the display time for each image presented to the participants to 40 seconds.

[2] This is the prompt used by eBird, a popular citizen science platform (www.ebird.org).



Tyran. It follows the classification rule: two blue wings, a short tail, a light blue body, two or three buttons on the light blue body, and one or two rings on each blue wing.	Nontyran because it has three rings on each wing. The number of legs is not diagnostic.

Figure 1. Sample Tyran and Nontyran Images

Various images of tyrans and nontyrans were created to test our hypotheses. Figure 2 shows some of the diagnostic and nondiagnostic attributes of the target entity. It is an example of a manipulated Tyran image with shorter wings and a differently colored tail. In Figure 2, a change to a diagnostic attribute has occurred.

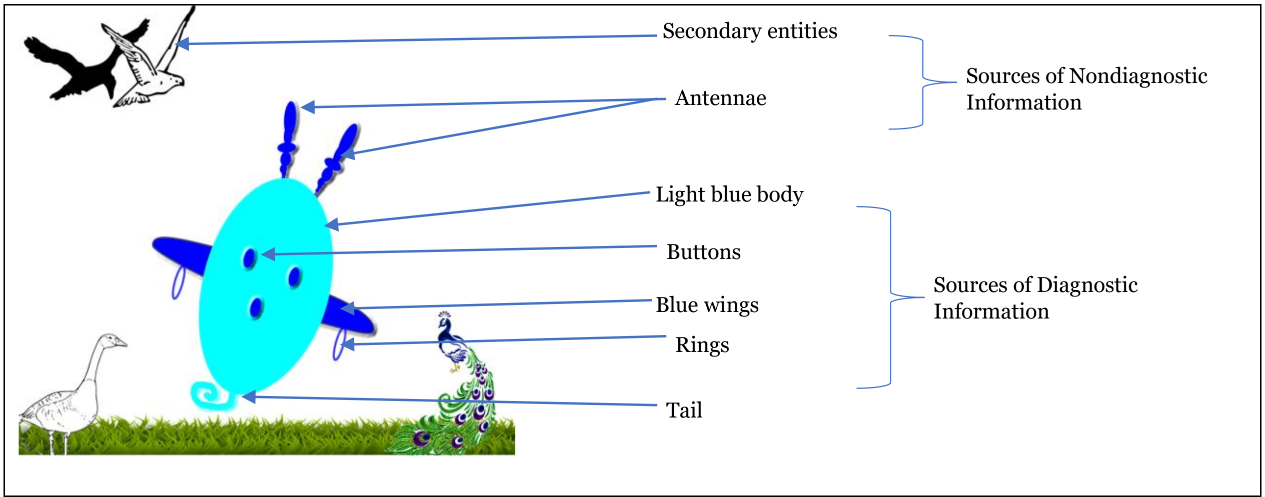

Secondary entities
Antennae
Sources of Nondiagnostic Information
Light blue body
Buttons
Blue wings
Rings
Tail
Sources of Diagnostic Information

Figure 2. Manipulations of Diagnostic and Nondiagnostic Attributes

We presented sixteen images (a mixture of tyrans and nontyrans) to the participants, and all sixteen images test the capacity of contributors to report accurate and diverse information. However, six images included manipulations of the diagnostic attributes and nondiagnostic attributes of the target entity, i.e., three for each attribute type. For example, even though nondiagnostic, the antennae are shorter in some of the tyrans images presented than in those presented in the training phase of the experiment. The presence of patterns on the wings of some tyrans, the number and shape of antennae, and the number of legs on the tyrans are differentiated attribute values we included in the images.

Four slides containing catch items were placed intermittently among the test item slides (tyran and nontyran insects) to check whether the participants were paying attention throughout the experiment. The catch items were differently shaped/colored stimuli that were not insects, and each participant was expected to correctly report them as nontyrans or to provide specific descriptions of their attributes. The slides were presented in a nonrandomized order to all three groups, with image 5, 10, 15, and 20 containing catch items not related to the actual task.

Participants in the explicitly trained group were provided with the classification rule introduced above for classifying stimuli as either tyrans or nontyrans. They went through a training phase in which they were taught the rule and were shown five sample tyrans to allow them to become familiar with applying the criteria in the classification task. Following the Kloos and Sloutsky (2008) study, participants were not shown images of the distractor stimuli or nontyrans, as we assumed participants could encounter infinitely many types of nontyrans when classifying in the field, making it unrealistic to show all possible non-tyrans. The participants were also tested on their knowledge and received feedback on their ability to identify tyrans. We presented participants with images and asked whether they thought each image was a tyran or not and why. After they provided their answers, we showed them the correct responses and explained how they satisfied the classification rule.

Those in the implicitly trained group were briefed on the task to be performed and were shown the same five target stimuli used to teach the explicitly trained group to allow them to try to determine the classification criteria. 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. Additionally, we did not show the untrained group any sample images. However, like the other groups, members of this group were informed that we were interested in examining how people report information.

Responses from eighty-four participants, who were randomly assigned to the three groups (untrained, implicitly trained, and explicitly trained), were processed after screening for completeness and the attentiveness of the contributor. Each group had 28 participants, who were all undergraduate university students majoring in business. Thirty-six of the participants were male and forty-eight were female.

RESULTS

We developed a coding scheme that accounts for attributes of the target entity and attributes of other stimuli reported by participants. The authors agreed on a coding scheme, and two of the authors participated in coding the first ten reports to establish consensus and conformance with the coding scheme. The first author coded the remaining reports with the other authors reviewing the coded data at different stages of the coding process.

We counted the number of attributes of the stimuli in the presented images reported (see sample in Figure 3). Additionally, we used a one-way analysis of variance (ANOVA) and Tukey's HSD[3] test for post hoc comparison of the group averages (excluding the catch item images used for screening purposes only) to compare the variables described in Table 1 below across the groups.

The components of information diversity are the attributes of target entities and secondary entities present in the visual field (presented image) of the contributor. Each image has one target entity (except in catch images), and either no secondary entity or one or more secondary entities. Attributes of a target entity that can be used to classify it as a tyran or non-tyran are diagnostic attributes. Other attributes of an entity that are not important for classification are non-diagnostic attributes. Nondiagnostic attributes include all mutual and behavior attributes. Mutual attributes are attributes that describe the relation of one entity to another. Behavior attributes are attributes that tell us about the current state of the target entity.

Secondary entities presented in the images are common organisms or objects such as birds, insects, and fences. All attributes of secondary entities are nondiagnostic in the context of this study as they do not help identify the target entity. We, therefore, have two categories of non-diagnostic attributes – those inherent in the target entity (simply referred to as Nondiagnostic Attributes) and those not inherent in the entity (e.g., behavioral attributes, and attributes describing secondary entities). To ensure the results of our analyses are not due to the inherent differences in the images presented, we standardized the data for each variable across the presented images using the Robust Scaler. The Robust Scaler is a standardization and variance scaling technique provided in the Scikit-learn machine learning package of Python, and it is the most accommodating of outliers since it uses data in the 1st and 3rd quartiles to center and scale the entire data set; extremely high values do not have any effect on the results (www.scikit-learn.org).

To determine the difference in diversity between participants who have received different types of training about the entity, we compared the aggregate number of attributes reported by each group. Information diversity – the number of unique attributes reported by contributors about an entity of interest, information diversity is the sum of all the attributes reported (see Table 1), which is as follows:

[3] Tukey's honestly significant difference (HSD) corrects for multiple comparisons (Homack, 2001).

Information Diversity = Diagnostic Attributes + Nondiagnostic Attributes + Behavior Attributes + Mutual Attributes + Secondary Entity Attributes + Secondary Entity Mutual + Secondary Entity Behavior + Diagnostic Differentiation + Nondiagnostic Differentiation

Codes	Description
Accuracy	The accuracy of identification of the target entity (tyran or nontyran; trained groups only)
Diagnostic Attributes	The number of target entity diagnostic attributes mentioned
Diagnostic Values	The number of values reported for each diagnostic attribute of the target entity., i.e., the amount of information reported for each attribute; for example, the values for the diagnostic attribute blue wings may be "short"
Nondiagnostic Attributes	The number of target entity nondiagnostic attributes mentioned
Nondiagnostic Values	The number of attribute values for nondiagnostic attributes (e.g., the color of the tail, where the presence of a tail is a diagnostic attribute, the length of the tail is a diagnostic value, but the color of the tail is a nondiagnostic attribute value even though the tail is diagnostic)
Behavior Attributes	Entity behavior: descriptions provided for the behavior or perceived activity of the entity
Mutual Attributes	Entity mutual attribute: descriptions provided for the relationship of the target entity in terms of other entities, including its environment
Secondary Entities	The number of secondary entities reported
Secondary Entity Attributes	Secondary entity attribute (attributes of secondary entities)
Secondary Entity Mutual	Secondary entity mutual attribute: description of the relationship between secondary entities
Secondary Entity Behavior	The number of descriptors of secondary entity behavior reported
Diagnostic Differentiations	The number of differentiated attribute values in the diagnostic attributes reported
Nondiagnostic Differentiations	The number of differentiated attribute values in the nondiagnostic attributes reported

Table 1. Variables Coded in the Contributed Data

Information diversity is significantly different across groups, with $F(2,81) = 85.967$, $p < 0.001$. Table 2 shows that for information diversity, the group mean of the untrained group is significantly higher than those of the explicitly trained and implicitly trained groups.

In addition, the group average of the implicitly trained group is significantly greater than that of the explicitly trained group. This supports Hypothesis 1.

Information Diversity	A	B	mean(A)	mean(B)	Mean Diff.	Std. Err	T	p-value	η^2
	E	I	4.070	7.433	-3.363	0.383	-8.777	0.001	0.079
	E	U	4.070	8.984	-4.914	0.383	-12.825	0.001	0.155
	I	U	7.433	8.984	-1.551	0.383	-4.048	0.001	0.018

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

Table 2. Results for the Effect of Training on Information Diversity

For the specific attribute types that make up the information diversity aggregate, we compared the number of Nondiagnostic Attributes of the target entity between the treatment conditions. For each image presented to the participants, we also compared other nondiagnostic attributes, such as attributes describing the state of the target entity (Behavior Attributes) and attributes describing the target entity’s interaction with other entities or with its environment (Mutual Attributes). The results are presented in Table 3. The mean value of Nondiagnostic Attributes is significantly different across the three groups, with *F* (2,1341) = 60.405, $p = 0.001$.

The average number of nondiagnostic attributes reported is highest for the implicitly trained group and lowest for the explicitly trained group. The number of attributes reported that describe the target entity’s behavior (Behavior Attributes) is also significantly different across groups. The group averages for the explicitly trained and implicitly trained groups are significantly lower than that for the untrained group. However, the explicitly trained and implicitly trained groups are not significantly different. The number of mutual attributes is also significantly different across groups. For mutual attributes (Mutual Attributes), we again observe that the group means of the explicitly trained and implicitly trained groups are significantly lower than that of the untrained group but the same for the explicitly trained and implicitly trained groups. These results support Hypotheses 2a and 2b. Beyond intrinsic nondiagnostic attributes, untrained contributors were better than trained contributors at reporting the behavior (Behavior Attributes) and state (Mutual Attributes) of the entities.

	A	B	mean(A)	mean(B)	Mean Diff.	Std. Err	T	p-value	η^2
Nondiagnostic Attributes	E	I	0.580	2.801	-2.221	0.205	-10.844	0.001	0.116
	E	U	0.580	2.009	-1.429	0.205	-6.975	0.001	0.052
	I	U	2.801	2.009	0.792	0.205	3.869	0.001	0.016
Behavioral Attributes	E	I	0.045	0.022	0.022	0.062	0.362	0.900	0.000
	E	U	0.045	0.491	-0.446	0.062	-7.243	0.001	0.055
	I	U	0.022	0.491	-0.469	0.062	-7.605	0.001	0.061
Mutual Attributes	E	I	0.681	0.725	-0.045	0.151	-0.295	0.900	0.000
	E	U	0.681	1.763	-1.083	0.151	-7.162	0.001	0.054
	I	U	0.725	1.763	-1.038	0.151	-6.867	0.001	0.050
Secondary Entities	E	I	1.036	1.544	-0.508	0.109	-4.674	0.001	0.029
	E	U	1.036	2.289	-1.253	0.109	-11.521	0.001	0.154
	I	U	1.544	2.289	-0.745	0.109	-6.847	0.001	0.060
Secondary Entity Attributes	E	I	0.246	1.261	-1.016	0.162	-6.274	0.001	0.042
	E	U	0.246	1.730	-1.484	0.162	-9.170	0.001	0.086
	I	U	1.261	1.730	-0.469	0.162	-2.896	0.011	0.009
Secondary Entity Behavior	E	I	0.056	0.257	-0.201	0.085	-2.368	0.047	0.006
	E	U	0.056	0.491	-0.435	0.085	-5.131	0.001	0.029
	I	U	0.257	0.491	-0.234	0.085	-2.763	0.016	0.008
Secondary Entity Mutual	E	I	0.826	0.547	0.279	0.134	2.078	0.094	0.005
	E	U	0.826	1.105	-0.279	0.134	-2.078	0.094	0.005

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

Table 3. Results for the Effect of Training on Attribute Types

For the secondary entities, we analyze the variables Secondary Entities, Secondary Entity Attributes, Secondary Entity Behavior, and Secondary Entity Mutual. As predicted, the untrained group reported more secondary entities than the trained groups, and the implicitly trained contributors reported more than the explicitly trained. The mean for Secondary Entity Attributes is significantly different across groups: the untrained group is the highest, and the explicitly trained group is the lowest. Secondary Entity Behavior is also significantly different across the three groups. The average Secondary Entity Behavior for the untrained group is the highest, significantly higher than that of the explicitly trained group, which is the lowest. However, there is no statistically significant difference between the number of Secondary Entity Behavior reported by the untrained and implicitly trained groups. Secondary Entity Mutual is also significantly different across groups. For Secondary Entity Mutual, the mean of the untrained group is significantly higher than those of the explicitly trained and implicitly trained groups; however, there is no significant difference between the explicitly trained and implicitly trained groups. These results support Hypothesis 3. Untrained contributors are better than trained contributors at reporting secondary entities and their states.

The number of differentiated attribute values in the target entity is measured using the variables Diagnostic Differentiation and Nondiagnostic Differentiation. To compare the difference in these variables across the groups, we used one-way ANOVA. The results of the post hoc comparison of the group means are shown in Table 4.

	A	B	mean(A)	mean(B)	Mean Diff.	Std. Err	T	p-value	η^2
Diagnostic Differentiations	E	I	0.112	0.100	0.011	0.044	0.252	0.900	0.000
	E	U	0.112	0.056	0.056	0.044	1.261	0.419	0.002
	I	U	0.100	0.056	0.045	0.044	1.008	0.664	0.001
Nondiagnostic Differentiations	E	I	0.011	0.279	-0.268	0.052	-5.150	0.001	0.029
	E	U	0.011	0.067	-0.056	0.052	-1.073	0.610	0.001
	I	U	0.279	0.067	0.212	0.052	4.077	0.001	0.018

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

Table 4. Results for the Effect of Training on differentiated attribute values

From the ANOVA results, the average number of Diagnostic Differentiations reported is not significantly different across the three groups ($F(2,221) = 0.154, p = 0.064$) at the 5% level of significance. The results of the post hoc tests presented in Table 4 also show that there are no significant differences in the pairwise group means. This result does not support Hypothesis 4. Trained contributors do not do better than untrained contributors at reporting differentiated attribute values for diagnostic attributes.

However, group comparisons show that the average for Nondiagnostic Differentiations is significantly different, with $F(2,249) = 18.196, p < 0.0001$. From Table 4, the implicitly trained group reported significantly more attribute differentiation values for nondiagnostic attributes than the untrained and explicitly trained groups. There is no significant difference in the group means of the untrained and explicitly trained groups. While all groups reported equal numbers of differentiated attribute values in diagnostic attributes, the implicitly trained group reported more differentiated attribute values in the nondiagnostic attributes. This result supports Hypothesis 5.

The primary reason why contributors are trained is to ensure they provide high-quality data. Notwithstanding the above results, data users might reasonably worry that untrained contributors may not report information that will be useful for identifying target entities. We therefore compared the extent to which trained and untrained contributors reported diagnostic attributes of the target entity. We found that the number of diagnostic attributes reported is not significantly different across groups, with

F (2,1341) = 0.92, *p* = 0.399 at the 5% level of significance. Post-hoc analysis also reveals significant similarity between all the number of diagnostic attributes reported by all groups (see Table 5).

	A	B	mean(A)	mean(B)	Mean Diff.	Std. Err	T	p-value	η^2
Diagnostic Attributes	E	I	1.514	1.440	0.074	0.183	0.407	0.900	0.000
	E	U	1.514	1.272	0.242	0.183	1.324	0.383	0.002
	I	U	1.440	1.272	0.167	0.183	0.917	0.718	0.001

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

Table 5. Results for the Effect of Training on Diagnostic Attributes

Furthermore, since data users will benefit from understanding if they will be trading off traditional information quality for information diversity, we also tested for the effect of training on information quality. The two dimensions of information quality most pertinent to information users are accuracy and completeness (Wang & Strong, 1995). Accuracy is an operation on the attributes of an entity to *identify the entity correctly* (Wand and Wang 1996). Contributors perceive the attributes of an entity and analyze those attributes matching it to diagnostic attributes in their memory to classify it (Harnad, 2005) correctly. Accuracy is correct identification and classification (Harnad 2005; Wand and Wang 1996). Training provides the knowledge needed to classify or identify entities accurately. Contributors who show the greatest tendency for selective attention – explicitly trained contributors – will be better at accurately identifying entities than contributors with more flexible attention allocation. Explicitly trained contributors have reliable rules to guide their classification and exclusion of an entity from a target category. Implicitly trained contributors arrive at a classification rule guided by the salience of the entity's attributes. They may or may not elicit the correct rule or the complete set of diagnostic attributes needed to classify an entity and will, therefore, be less accurate than explicitly trained contributors. Untrained contributors cannot classify entities themselves, as they have no knowledge to guide such a classification.

We analyzed the data from our experiment to understand the relationship between training and accuracy. Accuracy is valued as one or zero or for each target entity presented, depending on whether the contributor correctly identifies the target entity as either a tyran or a non-tyran (1) or not (0). We compare the proportion of Accuracy $= 1$ in the Explicitly Trained Group and the Implicitly Trained Group using a chi-square test.

	Accuracy		
Group	0	1	Total
E	62 (13.8%)	386 (86.2%)	448 (100%)
I	162 (36.2%)	286 (63.8%)	448 (100%)
Total	224	672	895

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

Table 6. Differences in Accuracy for Explicitly and Implicitly Trained Groups

From Table 6, the chi-squared statistic of independence is 58.339, with a p-value < 0.0001. The proportion of accuracy in the Explicitly Trained Group is 0.862, and the proportion accuracy in the Implicitly trained group is 0.638. Thus, the explicitly trained group is significantly more accurate than the implicitly trained group.

Completeness is the presence of information about an entity that is sufficient for a particular use (Nelson et al. 2005). Completeness includes the breadth and depth of information (or attributes) reported about an entity (Wang & Strong, 1996). Breadth refers to the number of unique attributes reported about an entity, while the depth refers to the amount of information provided about each attribute. A contributor may mention a bird's wings and tail (breadth of attributes) and also mention that there are two wings and a green tail (the attribute values two and green gives more depth of information to each attribute). Completeness is contextual (Nelson et al. 2005). In the context of this study, we define complete information as that which is sufficient to identify the observed instance and its state. Therefore, information that includes all the possible diagnostic and nondiagnostic attributes (intrinsic to the entity) is the most complete for identifying the entity. Completeness (breadth) is derived by summing the Diagnostic Attributes and Nondiagnostic Attributes reported by each contributor for a target entity. Similarly, Completeness (Depth) is derived by aggregating the diagnostic attribute values (Diagnostic Values) and nondiagnostic attribute values (Nondiagnostic Values) reported for the target entity

	A	B	mean(A)	mean(B)	Mean Diff.	Std. Err	T	p-value	η^2
Completeness (breadth)	E	I	2.094	4.241	-2.147	0.322	-6.670	0.001	0.047
	E	U	2.094	3.281	-1.187	0.322	-3.688	0.001	0.015
	I	U	4.241	3.281	0.960	0.322	2.982	0.008	0.010
Completeness (depth)	E	I	1.016	2.773	-1.758	0.289	-6.077	0.001	0.040
	E	U	1.016	2.835	-1.819	0.289	-6.289	0.001	0.042
	I	U	2.773	2.835	-0.061	0.289	-0.212	0.900	0.000

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

Table 7. Differences in Completeness

The breadth of attributes reported is significantly different across the groups with $F(2,1341) = 22.327$, $p < 0.001$. The post hoc test results from Table 7 shows that the explicitly trained group reported significantly fewer attributes about the target entity than did the implicitly trained group and the untrained group. The untrained group reported fewer attributes about the target entity than did the implicitly trained group. Depth is also significantly different across the groups with $F(2,1341) = 25.507$, $p < 0.0001$. The post-hoc test results show that the average depth for the untrained group is significantly greater than the average depth for the explicitly trained group. The implicitly trained group also has a mean that is greater than that of the explicitly trained group but significantly different from that of the untrained group.

Explicitly trained contributors focus on the diagnostic attributes to which they have been introduced and ignore attributes that are not diagnostic, providing incomplete information about the observed entity. Implicitly trained contributors use a bottom-up approach to learn attributes during training; thus, they have attended to non-diagnostic attributes as well as diagnostic attributes of the target entity. Implicitly trained contributors, therefore, report more complete data (breadth and depth) about a target entity than explicitly trained contributors. While, in comparison, untrained contributors distribute their attention broadly across all salient entities in the visual field, including the salient attributes of secondary entities, attempting to report as much breadth and depth of information as possible. They, therefore, report less breadth of attributes about the target entity than implicitly trained contributors as they trade-off focusing on the target entity alone for focusing on all the entities in their visual field, but more depth of values for the attributes they deem salient enough to report about.

DISCUSSION

Potential contributors to data crowdsourcing projects are frequently trained based on the assumption that contributor knowledge is necessary for obtaining high-quality crowdsourced data. However, our study raises important questions about this assumption. We found that training does not affect the accuracy of attributes that crowds report. Both trained and untrained contributors report diagnostic attributes with similar levels of accuracy. In addition, training does not affect the reporting of diagnostic attributes. Contributors, whether trained or untrained, can report the attributes needed to accurately identify an entity. However, training affects the contributors' ability to report diverse data. Untrained contributors report more diverse data than trained contributors (i.e., attributes beyond those required for a classification task). Some crowdsourcing tasks involve collecting data about fleeting phenomena, so there is no opportunity to observe an entity again later. Therefore, organizations might want to capture as much detail about a phenomenon as they can the first time; untrained contributors might be best suited to such ventures. Moreover, attributes that are unimportant in one data use context can become important in another (Hoffman and Rehder 2010; Ogunseye and Parsons 2016). In such cases, untrained contributors are more likely to provide the most repurposable data.

Differentiated attribute values noticed in instances can convey useful information, such as the existence of a new subclass of an entity (or a new class of entities) or lead to other discoveries about an entity. If the goal of a data crowdsourcing task includes making discoveries by collecting differentiated attribute values for salient diagnostic attributes, then training offers no benefits, but instead unnecessarily limits the inclusiveness of the data crowdsourcing project. However, since implicitly trained contributors are better than explicitly trained and untrained contributors at reporting differentiated attribute values in nondiagnostic attributes, contributors should be trained implicitly in projects where the differentiated values for nondiagnostic attributes can also lead to discoveries.

Information quality is tied to a specific use context (Nelson et al. 2005; Wang et al. 1995). Therefore, data collected with attention to traditional information quality dimensions such as completeness and accuracy might not be useful when information needs to be repurposed, requiring resource-intensive changes to the crowdsourcing project or repeating the crowdsourcing tasks. The completeness of attributes reported in crowdsourced information about a target entity is also affected by training. All contributors, trained or untrained, reported attributes; however, explicitly trained contributors reported the least complete data and implicitly trained contributors reported the most. Explicitly trained contributors selectively attended to diagnostic attributes to the detriment of other attributes of an entity.

Implicitly trained contributors and untrained contributors provide higher quality and more diverse data than explicitly trained contributors when they are not required to classify entities but accurately report attributes. When high-quality repurposable data is the goal of a data crowdsourcing project, data users are better off not training contributors at all, or training contributors implicitly.

Finally, there are several limitations to the generalizability of our findings. First, in our experiment, we used only five exemplars in the implicit training condition, which is not in general adequate to learn all rules for classifying entities. Second, we assume organizations and individuals that own and use data crowdsourcing platforms can classify entities after data has been collected, based on the attribute data provided by contributors (Lukyanenko et al., 2019). However, this might not be easy or realistic in every case. Third, a comparison of trained and untrained contributors works in the context of our study because we can understand the descriptions provided by untrained contributors about the target artificial entity and the images used in the experiment are readily accessible to us for confirmation. This is not the case in some real-world scenarios, and data users might not be able to understand descriptions provided by untrained contributors in the field.

CONCLUSION

Diverse data is more repurposable, able to meet emergent data requirements, and yield more insights to phenomena. Diversity would not be important if we know exactly the uses of crowdsourced data and if those uses will not change. However, this is rarely the case, in part due to the widespread use of analytics and machine learning to seek insights from data. Many data crowdsourcing projects can therefore benefit from the flexibility that diverse data affords (Lukyanenko et al. 2016; Ogunseye and Parsons 2018; Parsons and Wand 2014).

Users of data crowdsourcing tacitly assume that allowing unknown novice data contributors to provide unrestricted data is antithetical to the collection of high-quality data. Underlying this tacit assumption is the belief that trained contributors will provide better quality data than untrained contributors and a generally narrow view of information quality, whereby organizations focus only on collecting accurate (and sometimes complete) data. Training intends to promote accuracy, but this study shows it can inhibit data diversity. Furthermore, training does not affect the capacity of crowds to report diagnostic attributes accurately. Both untrained and trained contributors accurately report diagnostic attributes, which can be used by humans or machines to determine the class of an entity. However, explicit training can constrain data-driven insight and discoveries.

Data users want data crowdsourcing projects that provide insightful data and can lead to discoveries. Achieving this requires adjusting or even abandoning some long-held notions of what high-quality data is and how to acquire it – especially if we want to repurpose our data.

ACKNOWLEDGMENTS

This research was carried out during the first author's (Shawn Ogunseye) PhD program at Memorial University of Newfoundland. The research is supported by a SSHRC grant and has benefited from input from prior SIGSAND workshops.

REFERENCES

1. Ballesteros, J., Carbunar, B., Rahman, M., Rishe, N., and Iyengar, S. S. 2014. "Towards Safe Cities: A Mobile and Social Networking Approach," *IEEE Transactions on Parallel and Distributed Systems* (25:9), pp. 2451–2462.
2. Barlow, H. B. 1989. "Unsupervised Learning," *Neural Computation* (1:3), pp. 295–311.
3. Best, C. A., Yim, H., and Sloutsky, V. M. 2013. "The Cost of Selective Attention in Category Learning: Developmental Differences between Adults and Infants," *Journal of Experimental Child Psychology* (116:2), pp. 105–119.
4. Bjorklund, D. F., and Harnishfeger, K. K. 1990. "The Resources Construct in Cognitive Development: Diverse Sources of Evidence and a Theory of Inefficient Inhibition," *Developmental Review* (10:1), Elsevier, pp. 48–71.

5. Bollen, J., Mao, H., and Zeng, X. 2011. "Twitter Mood Predicts the Stock Market," *Journal of Computational Science* (2:1), pp. 1–8.

6. Buschman, T. J., and Miller, E. K. 2007. "Top-down versus Bottom-up Control of Attention in the Prefrontal and Posterior Parietal Cortices," *Science* (315:5820), pp. 1860–1862.

7. Castriotta, M., and Di Guardo, M. C. 2011. "Open Innovation and Crowdsourcing: The Case of Mulino Bianco," in *Information Technology and Innovation Trends in Organizations*, Springer, pp. 407–414. (http://link.springer.com/chapter/10.1007/978-3-7908-2632-6_46).

8. Gelman, S. A., and Markman, E. M. 1986. "Categories and Induction in Young Children," *Cognition* (23:3), pp. 183–209.

9. Ghasemaghaei, M., and Calic, G. 2019. "Can Big Data Improve Firm Decision Quality? The Role of Data Quality and Data Diagnosticity," *Decision Support Systems* (120), pp. 38–49.

10. Goldstone, R. L., and Kersten, A. 2003. "Concepts and Categorization," *Handbook of Psychology*.

11. Gopnik, A. 2009. "How to Think Like a Baby - Big Think." (https://bigthink.com/videos/how-to-think-like-a-baby, accessed May 18, 2019).

12. "Hanny's Voorwerp – History of a Mystery." 2013. *Daily Zooniverse*, , September 24. (https://daily.zooniverse.org/2013/09/24/hannys-voorwerp/, accessed April 25, 2020).

13. Harnad, S. 2005. "To Cognize Is to Categorize: Cognition Is Categorization," *Handbook of Categorization in Cognitive Science*, pp. 20–45.

14. Harrison, C., Jorder, M., Stern, H., Stavinsky, F., Reddy, V., Hanson, H., Waechter, H., Lowe, L., Gravano, L., and Balter, S. 2014. "Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness—New York City, 2012–2013," *MMWR. Morbidity and Mortality Weekly Report* (63:20), p. 441.

15. Hoffman, A. B., and Rehder, B. 2010. "The Costs of Supervised Classification: The Effect of Learning Task on Conceptual Flexibility.," *Journal of Experimental Psychology: General* (139:2), p. 319.

16. Hosseini, M., Phalp, K. T., Taylor, J., and Ali, R. 2014. *Towards Crowdsourcing for Requirements Engineering*.

17. Johnston, W. A., and Dark, V. J. 1986. "Selective Attention," *Annual Review of Psychology* (37:1), pp. 43–75.

18. Katsuki, F., and Constantinidis, C. 2014. "Bottom-Up and Top-Down Attention: Different Processes and Overlapping Neural Systems," *The Neuroscientist* (20:5), pp. 509–521.

19. Kloos, H., and Sloutsky, V. M. 2008. "What's behind Different Kinds of Kinds: Effects of Statistical Density on Learning and Representation of Categories.," *Journal of Experimental Psychology: General* (137:1), p. 52.

20. Lawrence, J. M. 2015. "Rediscovery of the Threatened Stoffberg Widow Butterfly, Dingana Fraterna: The Value of Citizen Scientists for African Conservation," *Journal of Insect Conservation* (19:4), Springer, pp. 801–803.

21. Levin, D. A. 2019. "Plant Speciation in the Age of Climate Change," *Annals of Botany* (124:5), Oxford University Press US, pp. 769–775.

22. Lukyanenko, R., and Parsons, J. 2019. "Beyond Micro-Tasks: Research Opportunities in Observational Crowdsourcing," in *Crowdsourcing: Concepts, Methodologies, Tools, and Applications*, IGI Global, pp. 1510–1535.

23. Lukyanenko, R., Parsons, J., and Wiersma, Y. F. 2016. "Emerging Problems of Data Quality in Citizen Science," *Conservation Biology* (30:3), pp. 447–449. (https://doi.org/10.1111/cobi.12706).

24. Lukyanenko, R., Parsons, J., Wiersma, Y. F., and Maddah, M. 2019. "Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content," *MIS Quarterly* (43:2), pp. 623–647.

25. Lukyanenko, R., Wiggins, A., and Rosser, H. K. 2019. "Citizen Science: An Information Quality Research Frontier," *Information Systems Frontiers*, Springer, pp. 1–23.

26. Nelson, R. R., Todd, P. A., and Wixom, B. H. 2005. "Antecedents of Information and System Quality: An Empirical Examination within the Context of Data Warehousing," *Journal of Management Information Systems* (21:4), pp. 199–235.

27. Ogunseye, S., and Parsons, J. 2016. "Can Expertise Impair the Quality of Crowdsourced Data?," *SIGOPEN Developmental Workshop at ICIS 2016*.

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

29. Parsons, J., and Wand, Y. 2014. "A Foundation for Open Information Environments," *Proceedings of the European Conference on Information Systems (ECIS) 2014, Tel Aviv, Israel, June 9-11, 2014, ISBN 978-0-9915567-0-0*

30. Piaget, J., and Inhelder, B. 1969. *The Psychology of the Child*, (Vol. 5001), Basic books.

31. Prat-Ortega, G., and de la Rocha, J. 2018. "Selective Attention: A Plausible Mechanism Underlying

Confirmation Bias," *Current Biology* (28:19), Elsevier, pp. R1151–R1154.

32. Ransbotham, S., and Kiron, D. 2017. "Analytics as a Source of Business Innovation," *MIT Sloan Management Review; Cambridge* (58:3), N/a-0.
33. Ritchie, M. G., and Immonen, E. 2010. "Speciation: Mosquitoes Singing in Harmony," *Current Biology* (20:2), pp. R58–R60.
34. Ritchie, S. A., Long, S., Hart, A., Webb, C. E., and Russell, R. C. 2003. "An Adulticidal Sticky Ovitrap for Sampling Container-Breeding Mosquitoes," *Journal of the American Mosquito Control Association* (19:3), c1985-, pp. 235–242.
35. Rosch, E. H. 1973. "Natural Categories," *Cognitive Psychology* (4:3), pp. 328–350.
36. Schilthuizen, M., Seip, L. A., Otani, S., Suhaimi, J., and Njunjić, I. 2017. "Three New Minute Leaf Litter Beetles Discovered by Citizen Scientists in Maliau Basin, Malaysian Borneo (Coleoptera: Leiodidae, Chrysomelidae)," *Biodiversity Data Journal* (5), Pensoft Publishers.
37. Simons, D. J., and Rensink, R. A. 2005. "Change Blindness: Past, Present, and Future," *Trends in Cognitive Sciences* (9:1), pp. 16–20.
38. Sprenger, P. R. D. 1987. "The Used Tire Trade: A Mechanism for the Worldwide Dispersal of Container Breeding Mosquitoes," *J. Am. Mosq. Control. Assoc* (3), p. 494.
39. Surowiecki, J. 2005. *The Wisdom of Crowds*, Anchor.
40. Tarrell, A., Tahmasbi, N., Kocsis, D., Tripathi, A., Pedersen, J., Xiong, J., Oh, O., and de Vreede, G.-J. 2013. *Crowdsourcing: A Snapshot of Published Research*.
41. Theeuwes, J. 2010. "Top–down and Bottom–up Control of Visual Selection," *Acta Psychologica* (135:2), pp. 77–99.
42. Tripathi, A., Tahmasbi, N., Khazanchi, D., and Najjar, L. 2014. "Crowdsourcing Typology: A Review of Is Research and Organizations," *Proceedings of the Midwest Association for Information Systems (MWAIS)*.
43. Wand, Y., and Wang, R. Y. 1996. "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM* (39:11), pp. 86–96.
44. Wang, J. Y., Frasier, T. R., Yang, S. C., and White, B. N. 2008. "Detecting Recent Speciation Events: The Case of the Finless Porpoise (Genus Neophocaena)," *Heredity* (101:2), Nature Publishing Group, pp. 145–155. (https://doi.org/10.1038/hdy.2008.40).
45. Wang, R. Y., Storey, V. C., and Firth, C. P. 1995. "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge & Data Engineering* (4), pp. 623–640.
46. Wang, R. Y., and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), pp. 5–33.
47. Wickens, C. D., and McCarley, J. M. 2008. *Applied Attention Theory*, Boca Raton, FL: CRC Press.
48. Wiggins, A., and He, Y. 2016. "Community-Based Data Validation Practices in Citizen Science," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW* (Vol. 27), Association for Computing Machinery, pp. 1548–1559. (https://doi.org/10.1145/2818048.2820063).
49. Wiggins, A., Newman, G., Stevenson, R. D., and Crowston, K. 2011. "Mechanisms for Data Quality and Validation in Citizen Science," in *2011 IEEE Seventh International Conference on E-Science Workshops*, , December, pp. 14–19.
50. Wilkerson, R. C., Linton, Y.-M., Fonseca, D. M., Schultz, T. R., Price, D. C., and Strickman, D. A. 2015. "Making Mosquito Taxonomy Useful: A Stable Classification of Tribe Aedini That Balances Utility with Current Knowledge of Evolutionary Relationships," *PLoS ONE* (10:7). (https://doi.org/10.1371/journal.pone.0133602).
51. Wolfe, J. M. 1994. "Guided Search 2.0 a Revised Model of Visual Search," *Psychonomic Bulletin & Review* (1:2), pp. 202–238.
52. Woodall, P. 2017. "The Data Repurposing Challenge: New Pressures from Data Analytics," *Journal of Data and Information Quality (JDIQ)* (8:3–4), p. 11.
53. Woodall, P., and Wainman, A. 2015. *Data Quality in Analytics: Key Problems Arising from the Repurposing of Manufacturing Data*.
54. Yang, R., Xue, Y., and Gomes, C. 2018. *Pedagogical Value-Aligned Crowdsourcing: Inspiring the Wisdom of Crowds via Interactive Teaching*.
55. "Zika, Mosquitoes, and Standing Water | | Blogs | CDC." 2016. *Zika, Mosquitoes, and Standing Water*, March.(https://blogs.cdc.gov/publichealthmatters/2016/03/zikaandwater/, accessed May 1, 2020).