Finer Granularity Means Better Data: a Crowdsourcing Lab Experiment

Ryan J. A. Murphy[1], Jeffrey Parsons[1]

[1] *Memorial University of Newfoundland, 230 Elizabeth Avenue, St. John's, NL, Canada*

Abstract

Data granularity is the level of direct correspondence between data in an Information System (IS) and the real-world things represented by the data. It determines the amount of detail that may be captured, stored, and used by contributors and consumers of information in an IS. We present a between-groups lab experiment in which we manipulated the granularity of a crowdsourcing project's interface to assess the impact of granularity on data completeness, data correctness, and overall contributor participation. We found that contributors using a finer-grained data collection interface contributed more complete data, while contributors using a coarser-grained data collection interface contributed more incorrect data. Moreover, the level of granularity did not influence the degree of participation. These findings suggest that granularity is an important issue in the design of data crowdsourcing projects.

Keywords

Data crowdsourcing, observational crowdsourcing, granularity, conceptual model

1. Introduction

How does the design of data crowdsourcing projects influence the level of detail contributors are able and willing to contribute? How does the level of detail contributors are able to contribute influence their ability to contribute complete and correct data? In this paper, we explore how granularity (the level of direct correspondence between data in an Information System (IS) and the real-world things represented by that data [1]) influences the level of completeness and correctness in data captured by a crowd in a data crowdsourcing project.

Data crowdsourcing is a phenomenon in which a crowd is mobilized to collect or analyze large volumes, varieties, and/or velocities of data [2]—of potentially-questionable veracity [3]. Observational crowdsourcing is a kind of data crowdsourcing in which contributors capture observations about some domain (e.g., wildlife) of the real world over a continuous period [4, 5].

An important opportunity for data crowdsourcing projects is the unanticipated use and reuse of collected data [6, 7]. Yet, once data has been collected, it can be very difficult or even impossible to return to the observation that was the object of that data and capture more detail from it. For this reason, data granularity is an important issue for data crowdsourcing (and especially for observational crowdsourcing). Granularity is an important factor in the ability to use and reuse data. If data is captured at finer-grained levels of detail (e.g., features and descriptions of observed wildlife), it may be possible to combine the collected details in useful ways. However, if data is collected at coarse-grained levels (e.g., classes, such as type of animal), then potentially important details about the observation cannot be captured and may be lost forever [8].

However, before we can leverage the granularity in the design of data crowdsourcing projects, we need to guarantee that fine-grained data collected from data crowdsourcing actually does contain more detail than coarse-grained data. Moreover, we must make sure that this detail is useful (e.g., does it facilitate more complete representation of the real-world phenomena the observer is capturing?) and

VLDB 2021 Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale, August 20, 2021, Copenhagen, Denmark.
EMAIL: rmurphy@mun.ca (A. 1); jeffreyp@mun.ca (A. 2)
ORCID: 0000-0003-2428-299X (A. 1); 0000-0002-4819-2801 (A. 2)

that it is correct (e.g., finer-grained data capture does not introduce errors). So, in the present research, we explore three related research questions: (1) Are observations captured at finer levels of granularity more or less complete than observations captured at coarser levels of granularity? (2) Are observations captured at finer levels of granularity more or less correct than those captured at coarser levels of granularity? (i.e., does finer-grained data cause contributors to introduce more errors?) (3) As contributing fine-grained data may require more effort than coarse-grained data, do contributors providing fine-grained observations contribute fewer overall observations?

2. Experimenting with granularity in crowdsourced data collection

One way to expose the conceptual model of an information system to data contributors is through the interface used to submit data. The interfaces data contributors use shape the data they contribute, as prompts, input boxes and so on shape and frame the structure, type, and content of contributions. Therefore, in this experiment, we use different degrees of granulation in data collection interfaces to represent different levels of granularity in the hypothetical project's conceptual model.

2.1. Hypotheses

2.1.1. Data completeness

Wand and Wang [9] define completeness, an intrinsic dimension of data quality, as "the ability of an information system to represent every meaningful state of the represented real world system" (p. 93). Granularity clearly plays a role in enabling this ability in information systems: if the conceptual model of the IS is insufficiently granulated to represent some details of the real world, the data captured within that IS (determined by data collection and storage decisions) cannot include those details. An IS with a more finely-granulated conceptual model will be more able to completely represent a given domain than an IS with a less granulated conceptual model. At the same time, it may be unnecessary to maximally granulate a conceptual model in order to benefit from the effects of granularity. A coarse-grained model may facilitate a similar level of completeness as a fine-grained model. While both would encourage contributors to break down their contributions into more detail, at a certain threshold that level of detail may be tedious, leading to diminishing returns on increasing levels of granularity. We are therefore interested in the degree of benefit fine granularity and coarse granularity each provide over an ungranulated alternative. This gives us our first set of hypotheses:

- H(1a): An IS with a fine-grained data collection interface will generate more complete data than an IS with an ungranulated data collection interface.
- H(1b): An IS with a coarse-grained conceptual model will generate more complete data than an IS with an ungranulated data collection interface.

Note that we do not hypothesize a difference between coarse-grained and fine-grained interfaces in terms of the completeness of data collected.

2.1.2. Data correctness

We follow Wand and Wang's [9] definition of data correctness: the degree to which an information system's data represents valid states of its real-world domain. The more finely granulated the conceptual model of an IS is, the more specifically it can represent the target domain. Coarsely granulated conceptual models create data that depend on user inference to fill in details. In turn, we propose that less granulated conceptual models lead to data that is more prone to incorrectness. This gives us our second set of hypotheses (Note that we do not hypothesize a difference between coarse-grained and fine-grained interfaces in terms of the correctness of data collected):

- H(2a): An IS with a fine-grained data collection interface will generate less incorrect data than an IS with an ungranulated data collection interface.
- H(2b): An IS with a coarse-grained data collection interface will generate less incorrect data than an IS with an ungranulated data collection interface.

2.1.3. Granularity and cognitive effort

As discussed above, we expect granulated data to be more complete and more correct than ungranulated data. However, capturing this degree of completeness and correctness will be more cognitively demanding on users. Consequently, we expect that contributors will make fewer contributions when data collection is granulated than when it is ungranulated. This gives us two final hypotheses:

- H(3a): Users contributing data to an information system with an ungranulated data collection interface will produce more contributions than users contributing data to an information system with a fine-grained data collection interface.
- H(3b): Users contributing data to an information system with an ungranulated data collection interface will produce more contributions than users contributing data to an information system with a coarse-grained data collection interface.

2.2. Methodology

To explore how granularity influences data completeness, we ran a between-groups experiment in which we manipulated the data collection interface of a data crowdsourcing system to instantiate three different levels of conceptual model granularity. Participants were randomly assigned to one of the three conditions. After reviewing instructions for the task, completing a simple comprehension test, and responding to some background questions (e.g., self-report measures of wildlife expertise), participants were presented a set of up to 20 photos of wildlife, one at a time, in random order. Their task was to describe each photo using the system interface. Each participant completed a minimum of five such observations, after which they could choose to end their participation at any time. After participants opted to exit, or after they completed all 20 observations, they were asked a few questions about their experience in the before concluding their participation in the experiment. The experimental manipulation was never revealed to participants—i.e., participants in one condition were not told about the existence of alternative interfaces. More details about the experiment's participants, materials, and resulting measures are described below.

2.2.1. Participants

One hundred participants were recruited via Amazon Mechanical Turk (MTurk), a crowdsourcing platform in which people receive money for completing micro-tasks. Participants received $2 USD as a reward for completely participating in the experiment. 41 participants were randomly assigned to the "ungranulated" condition, 32 to the "coarse granularity" condition, and 27 to the "fine granularity" condition. Several participants (six from the "ungranulated" condition, five from the "coarse granularity" condition, and two from the "fine granularity" condition) seemed to complete each observation by using the sample photo to complete a reverse-image search on Google, then copying and pasting information from the first result into the task text boxes. These results were discarded, leaving 87 participants: 35 in the "ungranulated" condition, 27 in the "coarse granularity" condition, and 25 in the "fine granularity" condition.

2.2.2. Materials

In the MTurk interface, participants were given a brief overview of the experiment (essentially explaining that they would be helping researchers describe photos of wildlife), an informed consent form, a hyperlink to the task, and a text box in which to paste a completion code received after successfully completing the task.

The hyperlink in the MTurk interface brought participants to the experimental task materials, developed using Qualtrics survey software. Upon arrival, participants were presented with the task instructions. After reviewing the instructions, participants needed to successfully answer two questions testing their basic comprehension of the task. They were then asked four questions about their background before beginning the task itself. These materials are provided in Appendix 1.

2.2.3. Task

After completing the background questions, participants began the experimental task. Each participant completed between five and 20 observations in which they described the contents of a photo of wildlife. We allowed participants to exit after a minimum of five completed observations in order to assess whether our granularity manipulation influenced the number of contributions a participant was willing to make. This design allowed us to evaluate the influence of the experimental conditions in a setting without an extrinsic motivator. If participants completed fewer observations in any given condition (e.g., if they dropped out before completing 20 observations), it would indicate that the condition involved a more difficult or aversive task compared to a condition where participants completed more observations. The sample photos were taken from NL Nature (www.nlnature.com), a crowdsourcing platform in which users contribute sightings of wildlife in Newfoundland and Labrador. Up to twenty different photos of wildlife were presented to each participant in random order. For each observation, we measured the number of seconds between when the observation was first loaded and when the participant submitted the observation to exit the task or to move on to the next observation.

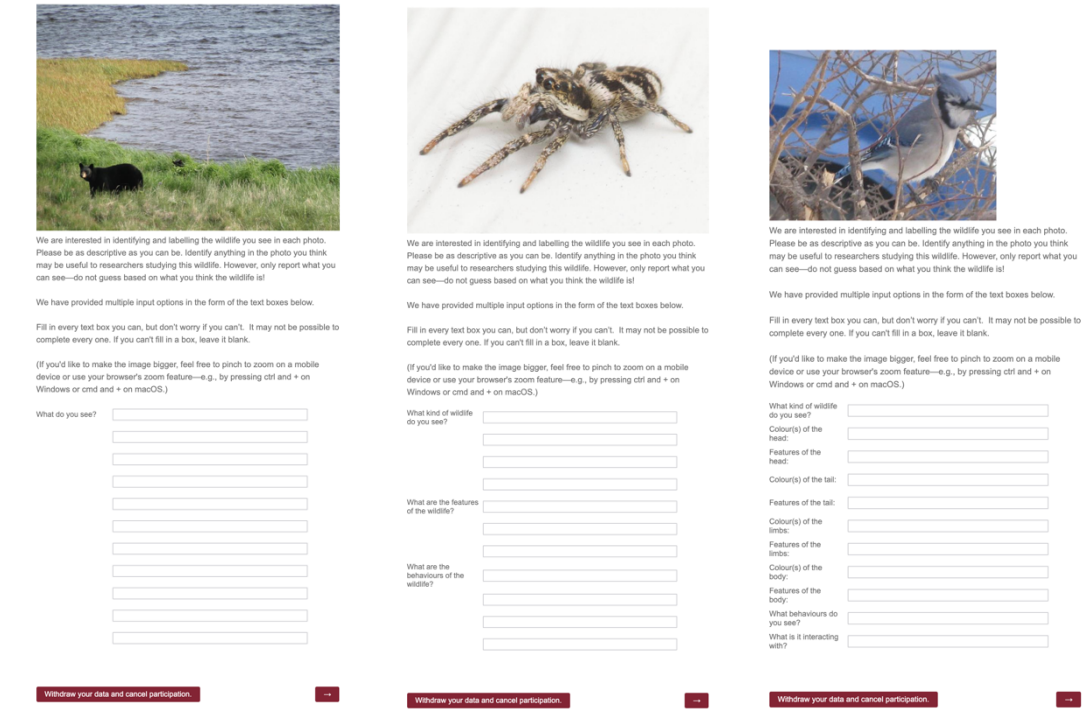

Figure 2: From left to right: ungranulated, coarse granularity, and fine granularity interfaces for the experimental task

For each observation, depending on the condition to which they were assigned, participants were presented with a screen like one of the three shown in Figure 2. While each participant had to complete the same task—describing the wildlife in the photo—we modified the ways in which participants had to provide their description across each condition. In the "Ungranulated" condition, participants were asked "What do you see?", seeking to emulate similar interfaces in real-world data crowdsourcing interfaces. In the "Coarse granularity" condition, participants were instead asked "What kind of wildlife do you see?", "What are the features of the wildlife?", and "What are the behaviours of the wildlife?"—

these three questions granulate the question "what do you see?" into three major possible groups of observations. Last, in the "Fine granularity" condition, participants were asked "What kind of wildlife do you see?", "Colour(s) of the head:", "Features of the head:", "Colour(s) of the tail:", "Features of the tail:", "Colour(s) of the limbs:", "Features of the limbs:", "Colour(s) of the body:", "Features of the body:", "What behaviours do you see?", and "What is it interacting with?", further granulating the questions in the coarse-grained category. To maintain structural equivalence, each condition included only 11 text boxes of identical size, exactly as depicted in the figure. Participants were instructed only to report what they could actually see, not to guess based on what kind of wildlife they thought they were looking at. Participants were also instructed not to fill in text boxes if it wasn't possible to answer a given prompt.

Table 1
Participant background information.

Condition	n	Expertise in wildlife (5-point scale)		Education in biology/ecology (4-point scale)		Self-identification as a citizen scientist (5-point scale)		Hours spent outdoors (5-point scale)	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Ungranulated	35	2.23	1.33	2.14	0.77	2.4	1.29	1.63	0.91
Coarse granularity	27	1.93	1.04	2.22	0.85	2.04	1.29	1.78	1.28
Fine granularity	25	1.96	0.84	2.24	0.83	2.4	1.26	1.56	0.96

2.3. Results

Analysis was completed with SPSS Statistics version 27. As seen in Table 2, participants in the Ungranulated condition completed an average of 9.17 (std. dev. = .86) observations (e.g., they looked at and described 9.17 photos of wildlife, on average). Participants in the Coarse Granularity condition completed an average of 9 (std. dev. = .89) observations, and participants in the Fine Granularity condition completed an average of 7.68 (std. dev. = .94) observations. The results presented here therefore include analysis of 756 total observations from 87 participants across the three conditions.

Table 2
Observations completed per condition

Condition	n	Observations completed				
		Total	Mean	Std. Dev.	Skewness (std. dev.)	Kurtosis (std. dev.)
Ungranulated	35	384	9.17	0.86	1.02 (.40)	-.40 (.78)
Coarse granularity	27	275	9	0.89	1.25 (.45)	.345 (.87)
Fine granularity	25	216	7.68	0.94	1.77 (.46)	1.73 (.90)

Using a one-way ANOVA (Table 2), we found no significant difference between groups' self-reported biology/ecology education ($F(2, 84) = .692$, $p = .882$), identification as citizen scientists ($F(2, 84) = .750$, $p = .475$), or hours spent outdoors ($F(2, 84) = .297$, $p = .744$). However, we found that participants' self-report of wildlife expertise violated Levene's test of homogeneity of variances, $F(2, 84) = 6.576$, $p = .002$. Therefore we used a one-way ANOVA with Welch's adjusted F ratio to test for significance, again finding none, $F(2, 55.683) = .597$, $p = .554$.

To explore potential differences in task completion time (Table 3), we analyzed the difference between the average completion time per-observation in each condition. These measures violated Levene's test for equality of variances, $F(2,84) = 3.648$, $p = .030$. So, we conducted a one-way ANOVA with Welch's adjusted F ratio, finding a statistically significant difference in mean observation completion time per-participant between conditions, $F(2, 49.918) = 3.619$, $p = .034$. Because group sizes were uneven and variances were unequal, we used Games-Howell's post-hoc procedure for multiple comparisons, finding only a significant mean difference between the fine-grained and coarse-grained conditions (mean difference = 48.68, p (.035) < α (.05).)

Table 3
Task experience measures

Condition	n	Completion time per-observation		Confidence in the task (5-point scale)		Ease of the task (5-point scale)		Enjoyability of the task (5-point scale)	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Ungranulated	35	86.40	73.39	3.91	0.98	3.77	0.84	3.89	0.99
Coarse granularity	27	88.41	41.22	3.78	0.93	3.70	1.10	4.30	0.72
Fine granularity	25	137.09	84.72	3.88	0.78	3.64	1.08	3.96	1.02

Finally, we checked to see if participants in any condition had more or less confidence, found the task easier or more difficult, or more or less enjoyable than the other conditions. To do this, we conducted a one-way ANOVA on each measure, finding no statistically significant difference in participant confidence ($F(2, 84) = .177$, $p = .838$), ease ($F(2, 84) = .128$, $p = .838$), or enjoyment ($F(2, 84) = 1.609$, $p = .206$) across the three conditions (Table 3).

2.3.1. Measuring completeness and correctness

To compare the differences between our experimental conditions, each observation for every participant was coded by one of the authors into three measures: a "total feature count," "total correct feature count," and "total nonconforming feature count" per observation.

Each observation was randomly ordered with the condition hidden, such that the coder could not tell which observations were submitted under which conditions. The coder reviewed the content of each submission by comparing it to the photo, counting the number of distinct pieces of information the participant contributed about the wildlife in the photo. This process resulted in a "total feature count" per observation. An example of this coding process, including the sample photo used, is provided in Appendix 2.

Next, the coder tallied the "total correct feature count" by comparing the text of each feature against the photo. If the content of the observation described something that did not conform with the photo (e.g., "furry body" for a bird, or "eight legs" for a spider with two legs obscured in the photo), completely inferential ("looking for food" or "flies", referring to a bird standing in a meadow), opinionated ("nice" or "attractive") or otherwise meaningless ("it's difficult to say."), it was not tallied. All other submitted features were summed in a "total correct feature count" per observation. When in doubt about a given feature's uniqueness or its correctness, the coder always opted to include the uncertain feature in the count. Last, we subtracted the total correct feature count from the total feature count for each observation, resulting in a "total nonconforming feature count" measure for each observation: a tally of features that were not evident by the photo alone.

Does granularity influence data completeness?

To test Hypotheses 1a and 1b, we evaluated the difference between conditions in terms of the total correct feature count. Observations-per-condition sample sizes were low (ranging from 6 to 22)—recall that participants had to complete at least five observations, but could quit anytime between five and 20. We believe that some images were less likely to be seen than others due to chance distribution (images were presented in random order). Because of these small sample sizes, we first tested the assumption of normality using Shapiro-Wilk, finding that normality is violated for 21 cases within the 60 condition-sample pairs (tables available upon request). Since our data does not adhere to a normal distribution, we used the Kruskal-Wallis *H*-test [10] to assess if the number of reported features is significantly different between conditions. If the Kruskal-Wallis test is significant, it indicates that at least one of the samples in a comparison is dominant over the other samples. If a given sample showed significance according to Kruskal-Wallis, we then used pairwise comparisons with Bonferroni corrections to identify which condition-pairs were significantly different from one another. Table 4 briefly presents the results of these analyses.

Table 4
Results of the differences in data completeness between groups across all observations as resulting from the Kruskal-Wallas *H*-test. *: $p < .05$.

Observation	Asymptotic Significance *p*-values	Significant pairwise comparisons (Bonferroni-corrected Mann-Whitney U test *p* value)
White-throated Sparrow	.095	N/A
Rabbit	.021*	Ungranulated–Fine granularity (.17)
Purple Finch	.012*	Coarse granularity–Fine granularity (.036) Ungranulated–Fine granularity (.10)
Moose	.002*	Ungranulated–Fine granularity (.002) Coarse granularity–Fine granularity (.016)
Kittiwake	.016*	Ungranulated–Fine granularity (.014)
Jumping Spider	.003*	Ungranulated–Fine granularity (.002) Coarse granularity–Fine granularity (.029)
Trout	.004*	Coarse granularity–Fine granularity (.008) Ungranulated–Fine granularity (.012)
Mourning Cloak Butterfly	.270	N/A
Spotted Sandpiper	.005*	Ungranulated–Fine granularity (.005)
Whale	.108	N/A
Harbor Seal	.002*	Ungranulated–Fine granularity (.001)
Weasel	.008*	Ungranulated–Fine granularity (.007) Coarse granularity–Fine granularity (.036)
Hummingbird	.035*	Coarse granularity–Fine granularity (.048)
Fox	.025*	Ungranulated–Fine granularity (.033)
Crab	.012*	Ungranulated–Fine granularity (.016) Coarse granularity–Fine granularity (.032)
Bat	.051	N/A
Blue Jay	.000*	Ungranulated–Fine granularity (.000)
Lynx	.000*	Ungranulated–Fine granularity (.000) Coarse granularity–Fine granularity (.021)
Black Bear	.016*	Ungranulated–Fine granularity (.020) Coarse granularity–Fine granularity (.049)
American Crow	.043*	Ungranulated–Fine granularity (.036)

As illustrated by the results in the table, we found differences between the conditions in 16/20 of the images. In all 16 of these cases, fine granularity was consistently dominant. Moreover, in eight of these 16 cases, fine granularity observations dominated both coarse and ungranulated observations. This evidence leads us to accept hypothesis (1a). However, our results do not show that coarse granularity leads to a more complete dataset than ungranulated, so we reject hypothesis (1b).

Does granularity influence data correctness?

To test hypotheses (2a) and (2b), we examined the difference between conditions in terms of the total nonconforming feature count for each observation. Again, we began by testing the assumption of normality with Shapiro-Wilk (Appendix 3), and again, very few (only four) image-condition pairs followed normal distribution. Therefore, we used the Kruskal-Wallis *H*-test on this data to test for statistically significant differences between conditions in each sample. When significant differences were found, we used Mann-Whitney pairwise comparisons with Bonferroni corrections to identify which condition-pairs were significantly different from one another.

Table 5
Results of the differences in data correctness between groups across all observations as resulting from the Kruskal-Wallas *H*-test. *: p < .05

Observation	Asymptotic Significance *p*-values	Significant pairwise comparisons (Bonferroni-corrected Mann-Whitney U test significance value)
White-throated Sparrow	.111	N/A
Rabbit	.009*	Ungranulated–coarse granularity (.006)
Purple Finch	.106	N/A
Moose	.001*	Fine granularity–coarse granularity (.001) Ungranulated–coarse granularity (.034)
Kittiwake	.001*	Fine granularity–coarse granularity (.001) Ungranulated–coarse granularity (.009)
Jumping Spider	.002*	Fine granularity–coarse granularity (.010) Ungranulated–coarse granularity (.006)
Trout	.000*	Fine granularity–coarse granularity (.001) Ungranulated–coarse granularity (.001)
Mourning Cloak Butterfly	.005*	Ungranulated–coarse granularity (.004)
Spotted Sandpiper	.001*	Fine granularity–coarse granularity (.002) Ungranulated–coarse granularity (.011)
Whale	.256	N/A
Harbor Seal	.056	N/A
Weasel	.209	N/A
Hummingbird	.004*	Ungranulated–coarse granularity (.004)
Fox	.028*	Ungranulated–coarse granularity (.028)
Crab	.007*	Ungranulated–coarse granularity (.007)
Bat	.000*	Ungranulated–coarse granularity (.000)
Blue Jay	.039*	Fine granularity–coarse granularity (.041)
Lynx	.000*	Fine granularity–coarse granularity (.001) Ungranulated–coarse granularity (.001)
Black Bear	.051	N/A
American Crow	.026*	Ungranulated–coarse granularity (.023)

As can be seen in Table 5, 14/20 images had significant differences between conditions in terms of the number of nonconforming features contributed by participants. However, contrary to our hypotheses, it was not the ungranulated condition that produced the most nonconforming data. Instead, observations in the coarse granularity condition were consistently more nonconforming. This leads us to reject hypotheses (2a) and (2b).

Does granularity influence contribution quantity?

Hypotheses (3a) and (3b) concern the number of contributions made by each user. We expect that participants in ungranulated condition will provide fewer contributions than those in the granulated conditions. To test this hypothesis, we used a one-way ANOVA to compare the means of the number of completed observations across participants in the three conditions. Surprisingly, we found no statistically significant difference between the groups ($F(2, 84) = .773$, $p = .465$). Therefore, we reject hypotheses (3a) and (3b).

2.4. Discussion

In all but four images, the number of correct features described by participants in the fine-grained condition was substantially greater than those described by participants in the ungranulated condition. Two of the four exceptions were photos of a bat and a whale. In both photos, the wildlife was relatively obscure, and many participants struggled to identify what it was. The obscurity may have limited what participants could comment on, even when asked many detailed questions about the wildlife (as in the fine-grained condition.) In the other two exceptions, the cause of the discrepancy is less obvious. One sample was a close-up photo of a Mourning cloak butterfly, the other of a White-throated sparrow. Perhaps the level of detail and variety of patterns and colours visible on both subjects facilitated more descriptive contributions.

Otherwise, however, the evidence overwhelmingly supports the acceptance of hypothesis 1a. Fine-grained data collection interfaces drastically changed the degree of completeness of description provided by participants. Moreover, the fine-grained data collected here was often specific to the exact animal in the photo participants were observing (e.g., “small antlers” vs. “antlers,” in the case of the Moose). Additionally, while it was beyond the scope of the present study to explore the usefulness of this data, anecdotally the data includes many examples where finer-grained data may have helped recover otherwise-bad contributions. For two particularly illustrative examples, when participants identified a Moose as a “forest donkey” and a Harbor seal as a “sea cow”), the other features they provided may still have been useful: according to these participants, the “forest donkey” had strong, long legs, while the “sea cow” lived in snowy water, and in the North. In other words, when asked to contribute granulated data participants seemed to describe the instance, not just the class, of what they were observing [8].

Surprisingly, fine-grained data was substantially more detailed than coarse-grained data, while the difference between coarse-grained data and the ungranulated condition was not statistically significant (i.e., hypothesis 1b was rejected). It could be that there is a threshold for the benefits of fine granularity in data collection interface designs: only after participants are encouraged to contribute a certain degree of detail does the effect on completeness start to show. On the other hand, granulation in data collection interface designs could have a linear relationship with completeness. A future experiment could test the nature of this relationship with more fine-grained manipulations of granularity in a data collection interface.

Our second set of hypotheses was not supported by the evidence in our experiment: more nonconforming data was introduced by participants in the coarse-grained condition, not in the ungranulated condition as we had expected. To restate, we used Wand and Wang’s [9] notion of data correctness to code participants’ descriptions, which notes that data that is incorrect is that which “does not conform to [the real-world things] *used to create the data*” (emphasis added). If a participant stated that the moose has “large antlers,” but the moose in the photo had short, stubby antlers, we coded this as nonconforming. Yet, obviously, the participant was basing their contribution on their understanding of moose in general, not *the* moose in the photo. The interpretation of this data as “incorrect” therefore

may be overly strict. Recoding the data to differentiate between "errors" (e.g., describing the moose as a "hippopotamus," as one participant did) and these generalized inferences may lead to new insights about the relationship between granularity and data correctness.

Still, the finding that participants using a coarse granularity interface performed the worst with this definition of correctness is worth discussion. Perhaps these participants felt tasked to provide information, but without the specific, detailed scope of a fine-grained interface, they weren't sure what else to add—thus, they provided information that wasn't supported by the photo.

It is also worth noting that there were three images in which the fine-grained condition participants did not provide *any* nonconforming information—the nonconforming feature total was 0 for all participants. The sample size of these three groups was small: 6, 7, and 10 participants. This means that fine-grained data collection interfaces both increase data completeness (hypothesis 1a) while maximizing data correctness (at least, defined as data that conforms to the real-world it was created from).

We found no statistically significant difference between groups in the number of observations participants completed. As per hypotheses 3a and 3b, we expected that more fine-grained data collection would be more demanding of participants, leading to fewer observations per participant in the coarse-grained and fine-grained conditions. Participants in the fine-grained condition generally produced more features than their counterparts in the other conditions, but perhaps fine-grained data collection made this task easier for them to do, balancing the amount of detail against lower cognitive effort per feature. Put another way, it may be easier to answer simple questions about the specific colours and features of wildlife than to answer more ambiguous questions about what a photo contains. Note, however, that we also found that participants in the fine-grained condition spent longer at the task than those in the coarse-grained condition. Managing crowd motivation is a crucial issue [11]: asking too much of contributors could have caused disengagement. Yet, in our experiment, participants using a fine-grained data collection interface submitted more information, more correctly, and spent longer doing so than their peers.

3. Implications

3.1. Contributions

The key contribution of this paper is considering data granularity as a key consideration for designers of data crowdsourcing projects. In many contexts in which the observed phenomena are fleeting, data simply cannot be re-collected. We conducted an empirical study on the effects of granularity on data collection via crowdsourcing, finding evidence suggesting that fine-grained data facilitates the collection of more complete and potentially more correct information, while having no effect on the number of contributions participants were willing to make. To underscore this point: finer-grained data may provide a more complete and correct representation of contributor observations, with no effect on level of participation. We encourage designers of crowdsourcing platforms to strive to collect more fine-grained data when possible, as data captured in this form may be more valuable than more coarse-grained forms.

3.2. Limitations

While we asked participants to self-report several important background characteristics, such as biology/ecology education and degree of expertise in wildlife, we did not constrain these measures to any particular geography, and we did not ask about participants' geographical backgrounds. Participants may have been participating from all over the world. It is possible that participants in certain conditions were more or less familiar with wildlife in Canada's East coast than others; future experiments should account for this geographical factor.

There are several possible limitations about our encoding of features from participant contributions. First, only one coder encoded the collected data. To make our results more robust, we are in the process of engaging a second coder to establish interrater reliability as part of our coding process. As previously

discussed, we may have had too strict a definition of correctness: some descriptors offered by participants were true of the wildlife sample in general, but not evident in the photo. Another question is whether a class-based descriptor should count as one feature, or if instead it implies many features. Participants in the fine granularity condition still reported classes but added many features. Also, reporting the species was not strictly the task: describing the individual animal itself was. What if researchers were searching this labelled dataset for instances of sickly-looking animals, to surveil for potential zoonoses? A fine-grained dataset would likely be more useful in this case (several participants described wildlife in this experiment as "healthy looking" or similar). Moreover, this effect further illustrates the benefits of fine-grained data when classes are mistakenly reported. In many cases where a species is mistakenly reported but other details are provided, the other details are still useful information.

Finally, the experimental design provides only a very crude operationalization of granularity. In practice, there could be many levels or degrees of granularity and we do not have insight into how varying levels of granularity might have unanticipated effects on other outcomes, such as contributor motivation or engagement.

3.3. Future directions

The development of data science has been characterized in terms of three movements: business intelligence and analytics 1.0, 2.0, and 3.0 [12]. Data science 3.0 includes increased use of mobile sensor data, more individualized and contextual analysis, and more human-centered and mobile data reporting (e.g., visualization; [12], see Table 1, p. 1169). To this end, is there a fourth wave of business intelligence and analytics? The 4.0 movement might involve recognizing the important role data contributors play in a data-driven world. To take advantage of this movement, data consumers and analysts should account for data producers in the design of their information systems. This 4.0 wave might therefore be characterized by design-centric data models calibrated to the ontology of the world a given data project aims to represent. This means tuning for appropriate granulations—as a corollary, other dimensions may be open to tuning as well.

The guidelines in [6] include a stipulation for mechanisms that automatically reconcile the instance-based data collected in the project with the coarse-grained features of a Target Organizational Model for the project sponsor's needs. Machine learning techniques such as supervised classifiers [13] may be useful here. Such a technique might be used as an automatic reconciliation system that treats every new contribution of sets of attributes as raw data and, simultaneously, as training data for an instance. A recent study, for example, demonstrates the potential of machine learning classification by classifying fine-grained crowdsourced data into more useful coarse-grained data with reasonable accuracy [7]. Further explorations of how to use similar artificial intelligence tools to enhance the utility of crowdsourced data is a potent area for future research.

4. Conclusion

The Internet, big data technologies, and other trends are rapidly unlocking new possibilities for massive, directed collaboration: data crowdsourcing. These methods can allow data consumers to collect data at unprecedented scales. However, if the data these activities generate is poorly captured, it limits their potential value. We have provided a better understanding of the effect of finer-grained data capture on data collection in crowdsourcing projects. Projects that enable their contributors to provide finer-grained data may be better suited to leverage data at big data scales.

5. Acknowledgements

This research was partially supported by grants from The Natural Sciences and Engineering Research Council of Canada (NSERC) and The Social Sciences and Humanities Research Council of Canada (SSHRC).

6. References

[1] R. Murphy and J. Parsons, Capturing the Forest or the Trees: Designing for Granularity in Data Crowdsourcing, in: Proceedings of the 53rd Hawaii International Conference on System Sciences, 2020, pp. 395-404.

[2] M. Stonebraker, What Does 'Big Data' Mean, Blog@CACM, 2012. URL: https://cacm.acm.org/blogs/blog-cacm/155468-what-does-big-data-mean/fulltext

[3] The Four V's of Big Data, IBM Big Data & Analytics Hub, n.d. URL: https://www.ibmbigdatahub.com/infographic/four-vs-big-data

[4] A. Castellanos, R. Lukyanenko, V. Storey, Modeling Observational Crowdsourcing, in: ER Forum, Demo and Poster 2020, 2020. URL: http://ceur-ws.org/Vol-2716/paper11.pdf

[5] R. Lukyanenko, J. Parsons, Beyond Micro-Tasks: Research Opportunities in Observational Crowdsourcing, Journal of Database Management 29.1 (2018): 1-22.

[6] R. Lukyanenko, Y. Wiersma, B. Huber, J. Parsons, G. Wachinger, R. Meldt, Representing Crowd Knowledge: Guidelines for Conceptual Modeling of User-generated Content, Journal of the Association for Information Systems 18.4 (2017): 297-339.

[7] R. Lukyanenko, J. Parsons, Y. Wiersma, M. Maddah, Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content, Management Information Systems Quarterly 43.2 (2019): 634-647.

[8] J. Parsons, Y. Wand, Emancipating Instances from the Tyranny of Classes in Information Modeling, ACM Transactions on Database Systems 25.2 (2000): 228-268.

[9] Y. Wand, R. Y. Wang, Anchoring Data Quality Dimensions in Ontological Foundations, Communications of the ACM, 39.11 (1996): 86-95.

[10] W. J. Conover, Practical Nonparametric Statistics, 3rd. ed., New York, NY, John Wiley & Sons, 1998.

[11] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, M. Vukovic, An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets, in: Proceedings of the 5th International Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media, 2011, pp. 321-328.

[12] H. Chen, R. H. L. Chiang, V. Storey, Business Intelligence and Analytics: From Big Data to Big Impact, MIS Quarterly 36.4 (2012), pp. 1165-1188.

[13] F. Provost, T. Fawcett, Introduction to Predictive Modeling: From Correlation to Supervised Segmentation, USA, O'Reilly Media, 2013.

7. Appendix 1

7.1. General Task Instructions

For the remainder of your participation in this study, you will be asked to report what you observe about a series of photos of wildlife. We are interested in identifying and labelling the flora and fauna you see in each photo. Please be as descriptive as you can be—fill in everything you can. Identify anything in the photo you think may be useful to researchers studying this wildlife.

To qualify as having completed this task, you must complete at least **five** such observations. That is, you must look at five different photos and report what you see for each. At any point before you complete the fifth observation, you may choose to exit the study by closing the window. If you exit the study in this way you will not be counted as having participated in the study and your data will not be used.

You may continue beyond five observations to complete as many as you'd like; more observations are helpful for our research.

7.2. Task Comprehension Check

Before you continue, we must check that you understand the task. Please demonstrate your comprehension by responding to the following questions:

1. In this study, you are reporting what you observe about:
    - Space
    - Wildlife
    - People
    - Architecture
2. In this study, you must complete a minimum of *how many* observations to participate?
    - 2
    - 10
    - 1
    - 5

*Note for reviewers: only participants who answer **b. Wildlife** for question 1 and **d. 5** for question 2 will continue to participate in the study.*

7.3. Participant Background Questionnaire

Please respond to the following questions:

1. I am an expert in wildlife.
    - Strongly disagree
    - Somewhat disagree
    - Neither agree nor disagree
    - Somewhat agree
    - Strongly agree
2. At what level of education have you studied wildlife, ecology, or biology?
    - I have never studied wildlife, ecology, or biology
    - I have some high school education in wildlife, ecology, or biology
    - I have some college or university education in wildlife, ecology, or biology
    - I have a college or university degree in wildlife, ecology, or biology
3. I consider myself a citizen scientist.
    - Strongly disagree
    - Somewhat disagree
    - Neither agree nor disagree
    - Somewhat agree
    - Strongly agree
4. Approximately how many hours per week do you spend outdoors?
    - 0-5
    - 5-10
    - 10-15
    - 15-20
    - 20+

8. Appendix 2

8.1. Data Coding Instructions

We have collected observations—descriptions of photos of wildlife—from participants using an experimental interface. The interface instantiates conceptual model granularity into three levels: ungranulated, coarse-grained, and fine-grained data. In our experiment, participants were randomly assigned to one of these levels of granularity. They completed between five and 20 observations.

To compare the differences between our experimental conditions, we are coding each observation for every participant. Each observation is randomly ordered with the condition hidden, such that coders cannot tell which observations were submitted under which conditions.

The result of this coding process will be three measures of each observation: a "total feature count," "total correct feature count," and "total nonconforming feature count" per observation.

"Total feature count" is the total number of distinct features described in the text of each observation. "Wing" is one feature; "Brown wing" is two: the animal has a "wing," and the wing is "brown."

To code the "total correct feature count", compare the text of each observation with the corresponding photo. Total correct feature count is the total number of features that are concretely, visibly present in the photo. "Four legs" would count as two correct features (the observed animal has legs, and it has four of them) *if and only if* all four legs are visible in the photo. The contents of the observation should not be counted when they describe:

- something that does not conform with the photo (e.g., "furry body" for a bird, or "eight legs" for a spider with two legs obscured in the photo),
- is an assumption of the observer ("looking for food" or "flies", referring to a bird standing in a meadow),
- was opinionated ("nice" or "attractive"), or
- was otherwise meaningless ("it's difficult to say.")

When in doubt about a given feature's uniqueness or correctness, *always* opt to include the uncertain feature in the count.

The total nonconforming feature count is calculated by subtracting the total correct feature count from the total feature count.

To summarize, the coding algorithm is:

1. Review the contents of each part of the observation.
2. Count the number of distinct features described in the part. This is the "total feature count"—write it in the coding sheet.
    - When uncertain about a given feature's uniqueness, always opt to include the uncertain feature in the count.
3. Count the number of features that directly correspond to what is observable in the photo. This is the "total correct feature count"—write it in the coding sheet.
    - Do not count non-conforming observations, assumptions, opinions, or otherwise meaningless information.
    - When uncertain about a given feature's correctness, always opt to include the uncertain feature in the count.
4. Subtract the "total correct feature count" from the "total feature count." This is the "total nonconforming feature count"—write it in the coding sheet.
5. Repeat steps 1–4 for each part of the observation.
6. Repeat steps 1–5 for each observation.

For demonstration purposes, an example of this process is on the next page. Notes on incorrect features are provided for explanation only. You do <u>not</u> need to record your own notes when encoding the data.

9. Appendix 3

Figure 1: The sample photo of a moose used by participants in the study

Table 7
A demonstration of how features submitted by a participant were encoded for statistical analysis

Data collected from participant	Individual features identified by the coder, separated by commas	Total feature count	Total correct feature count (Comment on why a feature was not correct)	Total nonconforming feature count
moose	Moose (a type of animal)	1	1	0
large deer like mammal	The moose is large, The moose is deer-like, The moose is a mammal	3	2 (The fact that the moose is a mammal cannot be observed by the photo alone)	1
male has very large antlers	Male moose have antlers, moose antlers are very large	2	1 (The antlers in the photo are quite stubby)	1
long face; hanging skin under their chin	The moose has a long face, the moose has skin hanging from underneath the chin	2	2	0
eat grass and bushes	Moose eat grass, moose eat bushes	2	0 (The moose is not eating anything in the photo)	2
often live near water	Moose often live near water	1	0 (The moose is not near water in the photo)	1
can swim	Moose can swim	1	0 (The moose is not swimming in the photo)	1
they make loud bellowing sounds sometimes	Moose sometimes make bellowing sounds, these sounds are loud	2	0 (Cannot tell if the moose is bellowing in the photo, much less at what volume)	2
	Total:	12	6	6