Capturing the Forest or the Trees: Designing for Granularity in Data Crowdsourcing

Ryan Murphy
Memorial University of Newfoundland
rmurphy@mun.ca

Jeffrey Parsons
Memorial University of Newfoundland
jeffreyp@mun.ca

Abstract

Crowdsourcing is a method of completing a task by engaging a large group of heterogeneous contributors. Data crowdsourcing is crowdsourcing of data collection. In this paper, we demonstrate how data crowdsourcing projects can be differentiated along five dimensions: (1) the extent to which tasks are well-defined; (2) the duration of the task; (3) the type of value generated by the consumers of crowdsourcing data; (4) the variety of contribution allowed when completing the task; and (5) the relative value of each contribution. We argue that the quality of information created by a crowd depends on the granularity of contributions contributors are able to make. Finally, we propose a set of principles for designing crowdsourcing system to align the level of granularity of contributions with project objectives.

1. Introduction

Data granularity is the degree of detail at which data is captured, stored, and used by an information system (IS). Data granularity can therefore be thought of as the level of direct correspondence between data in an IS and the real-world things represented by the data represents. Information about the world is reflected in the level of resolution in captured data: that is, real-world phenomena represented in the system may be represented at varying levels of abstraction depending on the data's granularity [1].

When the world is represented in data, different features of the world are made salient by different granularities of the data. To extend a classic cliché: if you're looking to navigate through the woods, a map would likely be more useful if the forest itself was displayed in aggregate, rather than if each individual tree was labelled. On the other hand, if your goal is to find a particular aspen, a map that showed only the shape of the forest would be useless. Thus, when modeling the woods, a modeler must ask if they want to see the forest or the trees. The answer, of course, depends on how the map will be used.

Data granularity has a significant effect on the quality of data generated by ***data crowdsourcing systems***: systems that mobilize (large numbers of) people outside of traditional organizational structures to contribute data to a common project [2]. For example, different instances of road hazards in the real world might be captured as a single coarse-grained (or low resolution) "road hazard" class in a municipal problem-reporting app. In this case, these reports share a similarity: they are all road hazards, at least according to the perspectives of data contributors. The system takes advantage of this similarity to hide the *fine-grained* details of their peculiarities using a coarser granule. Aggregating fine-grained data into coarser classes simplifies data collection and storage [1]. Classification can also ease the cognitive burden on contributors [3-4]. However, data captured at coarser granularity can obscure important and useful finer-grained details that otherwise might have been collected, such as whether a road hazard is a pothole (which would need a repair crew to resolve it) or loose garbage (which would require a garbage collector). To use this detail in an application, data in a crowdsourcing system must be captured in as fine-grained detail as possible, thereby separating instances into more fine-grained components [1].

Selecting the appropriate granularity to *present* data may be easy. The presentation of information usually has a clear purpose with a particular audience. However, selecting the appropriate granularity for *collecting* data is not necessarily as simple, but as the example at the outset illustrates, it is particularly important in data crowdsourcing. Crowdsourcing projects often involve the general public in collecting data [5]. Contributors may therefore have limited training and motivation [6-10]. Crowdsourcing projects can also be quite large, operating at big data scales. Galaxy Zoo, for example, is a crowdsourcing platform launched in 2007 in which contributors analyze space imaging data. More than 400,000 volunteers have participated over its four iterations thus far, completing over 11 million classifications [11-12].

URI: https://hdl.handle.net/10125/63788
978-0-9981331-3-3

Like many big data projects, crowdsourced data has great potential for reuse [13-14]. This means that data should be useful for multiple purposes, some unknown at the time of collection [15]. Yet, in data crowdsourcing projects, contributors typically have varying degrees of expertise [15-17] and motivation [6-10]. It can therefore be difficult to guarantee the quality of contributor data at scale. This underscores the importance of getting data collection right the first time. If data is poorly granulated at the collection stage, it can be difficult for contributors to effectively report an observation [16]. It may also be discouraging to participate in the platform [7]. Finally, captured data may be difficult to reuse, especially for unexpected purposes [15]. Each of these challenges has significant implications for the design of a crowdsourcing IS.

In this paper we show how data collection granularity can affect the quality of information generated by crowdsourcing systems. We then present a set of principles to help design for granularity by optimizing data quality and crowdsourcing system performance, given an ideal level of granularity for a particular project. We achieve these objectives first by reviewing relevant literature on crowdsourcing, data science, data quality, and granularity. Then, we use these sources to synthesize a novel taxonomy of crowdsourcing approaches. Second, we extend research on conceptual modeling in information systems and particularly on crowdsourcing to develop a theory of problems with data collection granularity. Finally, we synthesize and describe three principles to help design crowdsourcing projects with effective granularity for crowd data collection.

2. A typology of crowdsourcing projects

Crowdsourcing differs from conventional (e.g., organizational) data collection approaches in that the design of the project should account for data contributors (crowd members contributing to the project) as much as data consumers (persons or organizations using the data) [16-17]. Conventional approaches to data quality—manufacturing, marketing, and service approaches [18-20]—emphasize the data consumer as the ultimate arbiter or data quality [21]. In other words, the quality of data (and the information systems that manufacture, market, and service it) are judged against a data-centric fitness-for-use paradigm: whether data is good or not depends on (for instance) how accurate, complete, and timely it is for a given consumer and use case. An alternative approach is a design-centric view of data quality that emphasizes fundamental principles of conceptual modeling [21]. In this view, the data quality of an information system is judged by the quality of the conceptual model the data in the system adheres to. In particular, ontology and cognitive psychology theories suggest that data is deficient when systems use predetermined and fixed classification in data modeling [21]. Classification affords important cognitive benefits, including cognitive economy (storage and processing efficiency achieved by classifying instances [3]) and inference (the ability to infer unobserved details about an instance based on the class(es) to which it belongs [4]).

However, classification can also cause information loss. Information loss occurs whenever a fine-grained instance is stored as a coarse-grained class: some of the details of the instance that are not represented by the selected class are lost (Parsons, 1996). From our perspective, a design-centric view of data quality suggests that higher data quality is attained when the fit between the real world and a project's data model minimizes information loss [21].

To develop principles to design for granularity in data crowdsourcing, it may be important to consider different types of crowdsourcing projects. Below we discuss two typologies. These typologies of crowdsourcing are important as they imply differences in approach to data collection, analysis, and/or use. In other words, these categories provide different characterizations of crowdsourcing projects from which to derive an approach to data granularity.

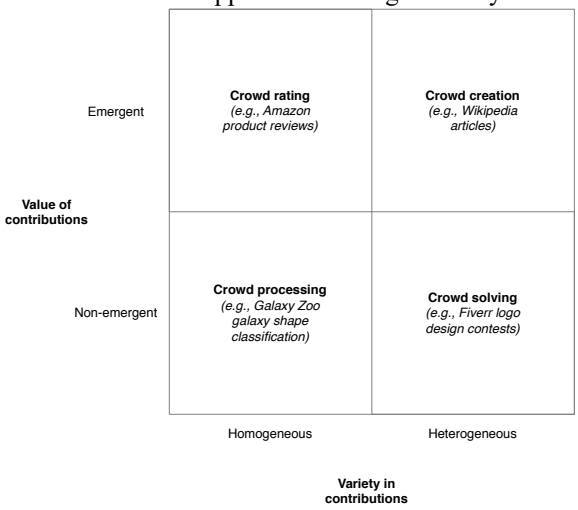

Figure 1. Emergent value vs. variety in contributions, adapted from [22, p. 6].

The first typology characterizes different crowdsourcing projects along two dimensions: how value is derived from crowd contributions (emergent, in which contributions are combined; or non-emergent, in which contributions are useful individually) and how uniform each contribution might be (homogeneous, in which each contribution has equivalent structure—e.g., ratings—or

heterogeneous, in which different contributions may be more or less valuable depending on their content—e.g., creative problem solving, in which some contributions may be more effective solutions than others) [22]. These two dimensions are combined to create a 2x2 matrix defining four types of crowdsourcing (see Figure 1).

The second typology differentiates two types of data crowdsourcing [23]. The first is task-based or micro-task projects, in which participants work on well-defined tasks (e.g., Amazon's Mechanical Turk). Participants in task-based crowdsourcing might only contribute to a given project for a brief period (as in Mechanical Turk tasks) and are often incentivized through pay or other remuneration. Task-based crowdsourcing projects are typically targeted and, when oriented towards a research objective, are hypothesis-driven and deductive. Data collected in these projects fit a closed world predetermined by the designer of the crowdsourcing tasks [24]. Task-based crowdsourcing may be contrasted with observational crowdsourcing, in which the crowd completes more open-ended tasks continuously over a long period of time [23]. These projects are typically performed out in the world, and contributors are usually volunteers.

This second typology presents a simple duality between observational and task-based projects [23]. This typology can be extended to account for an extra level of variance in these two paradigms. The typology collapses long-term collection and open-ended problems into observational projects, and short-term well-defined problems into task-based projects. However, there exists examples of short-term open-ended projects and long-term well-defined projects that defy this typology. Consider ratings in a mobile operating system app store as an example. Contributions are collected for a well-defined problem: how good an app is on a 1-5 scale, averaged over multiple contributors. Yet ratings are collected over a long time: some apps exist on the app store for a decade, and some users return to update their reviews as functionality in the app increases or decreases their satisfaction. With this argument, we extend this second typology into two dimensions: contribution definition and contribution limits.

These typologies neither compete nor explain the same features of different crowdsourcing projects. Therefore, we have modified some of the concepts from each of these predecessor typologies and aligned them in a new multi-dimensional *crowdsourcing contribution typology* (see Table 1). This typology provides a contributor-centered set of dimensions that help characterize projects based on the approach the project takes to the contributions it solicits from participants. Each of these dimensions describes a spectrum with two polarities. Most projects fall somewhere in between the extremes—based on our knowledge of existing crowdsourcing projects, "pure" examples of each polarity are rare. Next we define each dimension of the typology.

Contribution *Definition* describes how a project defines a successful contribution. Open-ended projects do not provide a clear objective. *Open-ended* crowdsourcing projects may be viewed as platforms within which people make contributions entirely of their own intrinsic motivations and interests. In contrast, well-defined problems explicate what successful contributions look like. An example of a somewhat open-ended crowdsourcing project is the Zooniverse Project Builder (zooniverse.org/lab). This platform crowdsources crowdsourcing projects, providing a platform for contributors—in this case, researchers acting as crowdsourcing project coordinators—to build their own Galaxy Zoo-like projects. While the Zooniverse Project Builder provides guidelines and limits on what constitutes a suitable project, contributors are able to submit anything that fulfills those guidelines: there is no "correct" project. In contrast, when contributions are *well-defined*, there are explicit parameters on what constitutes effective participation. Compare the Zooniverse Project Builder with the earlier-mentioned Galaxy Zoo, which asks contributors to help analyze astronomic images by classifying the shape of photographed galaxies in order to help researchers understand how galaxies form. Galaxy Zoo tasks usually feature right and wrong answers.

Contribution *Limits* define the limits a project places on achieving its objective. This is usually a duration: *short-term* projects must be completed within a given period, while *long-term* projects may accept contributions indefinitely. An example of a short-term project is the Audubon Christmas Bird Count, in which contributed data must be collected within a three-week period at the end of the year (https://www.audubon.org/conservation/join-christmas-bird-count). In contrast, the eBird platform (ebird.org) accepts contributions continuously.

Contribution *Emergence* describes the extent to which individual contributions compound and transform one another. In a project with *holistic* emergence, the sum of all contributions is different from each part taken separately. Quirky (quirky.com) provides a platform for inventors to propose and build new consumer products. Contributions to a given project build on one another, and the resulting product is continually transformed as a result. In contrast, projects with *discrete* contributions make progress as they aggregate each contribution.

Contribution *Variety* describes the extent to which participants may vary the form or content of their contributions. Contributions may feature *differing* variety, in which case contributions that diverge from one another in form and content may be equally valuable. Observational crowdsourcing platforms like iNaturalist (inaturalist.org) ask participants to submit observations of fauna or flora, but these observations may be of rare or common sightings and may include a variety of types of media. In contrast, a project may expect all contributions to be the *same*. A project like Galaxy Zoo asks every participant to submit the same contributions—classifications, based on answers to multiple choice questions—on different images.

Contribution *Value* is the extent to which one contribution may be more valuable than others. Some projects may feature *unique* contribution value. In extreme cases, such as a crowdsourced search and rescue operation (e.g., crowdsourcerescue.com), there may be successful contributions—a report of a missing person—and unsuccessful ones, such as those that mislead people in the field. In other projects, contributions may be *equally* valuable. Voting systems such as the SXSW PanelPicker (panelpicker.sxsw.com) make collective decisions through the equally-valuable contributions of votes on the composition of an effective panel.

This framework may be used taxonomically, to categorize existing crowdsourcing collaboration projects. It may also be used in crowdsourcing system development, to aid the design of effective crowdsourcing systems for a given purpose.

3. Data science and data quality in crowdsourcing

The significance of data granularity is underscored by recent attention on data science and big data. Trends in technology such as new and cheaper sensors, computers, Internet access, and analytical algorithms have unlocked exponential growth in data collection and data use [25]. As a result, data has rapidly shifted from being scarce to overabundant [26]. The big data phenomenon is exemplified by the "four V's": volume, velocity, variety, and veracity [27]. Respectively, recent advances in technology have enabled the collection and analysis of data at unprecedented scales, speeds, variety of forms, and levels of uncertainty. As a result, conventional approaches to data have been challenged, requiring changes from hardware [28] and engineering [27] to governance [14] and management [30-31].

Unsurprisingly, the value of data is gaining increasing recognition [29-32]. Yet the value of data is directly tied to its quality. It is not enough to arbitrarily collect and use data if that data is somehow of poor quality. However, data of high quality might hold value well beyond initial purposes if it supports flexible reuse [13]. But how do we ensure high data quality in crowdsourcing, especially for flexibility and reuse?

As discussed earlier, conventional approaches to data quality emphasize the role of the data consumer [19, 33, 36]. These approaches focus on data quality dimensions such as completeness ("The extent to which data are of sufficient breadth, depth, and scope for the task at hand", [19, p. 32], relevance ("The extent to which data are applicable and helpful for the task at hand", [19, p. 31], and timeliness ("The extent to which the age of the data is appropriate for task at hand", [19, p. 32]. As can be seen from these definitions, however, data quality assessment is typically tied to a particular use.

This use-/consumer-centric approach is dominant in data science. Perhaps the most dominant process model of data mining, the CRoss-Industry Standard Process for Data Mining (CRISP-DM), follows a use-centred approach [34]. Similarly, getting started on a data analytics project begins with selecting an appropriate challenge, then identifying the data,

Table 1. The crowdsourcing contribution typology of crowdsourcing projects.

Definition	Limits	Emergence	Variety	Value
Open-ended: Participants are given system features, but not direction, on what constitutes a valuable contribution.	Short-term: The project will accept contributions only within a certain timeframe.	Holistic: Contributions accumulate and transform one another, becoming different from the raw sum of the parts.	Different: Participants may vary the form and/or content of their contributions.	Unique: Each contribution provides unique value to the overall project, and some may be drastically more valuable than others.
Well-defined: Successful contributions are well-defined and communicated to participants.	Long-term: The project accepts contributions continuously, with no explicit end date.	Discrete: Contributions are accepted and valued in parallel with one another.	Same: Successful contributions are all generally the same in form and content.	Equal: Each contribution is equally valuable.

models, processes, and analytics that can help make progress on that challenge [35]. In the same theme, [36] suggests defining the quality of information as the potential of a data set to achieve a specific goal. Granted, most of these interpretations acknowledge the cyclical nature of data-driven projects (e.g., see the CRISP-DM life cycle [34, p. 10]). Still, each interpretation is objective-centered. The stipulation of a data model at the beginning of a project means establishing data granularity at this stage, too. Too coarse, and useful details might disappear into higher level aggregates. Too fine, and useful insights can become difficult to discern (or data will need to be processed extensively before used). While an effective analyst will set appropriate granularity levels for a given objective, the conceptual model articulated at this stage nonetheless anchors future data collection to the objective. If poorly conceived, this will limit extension and reuse for other, unanticipated objectives.

A previous study attempted to address this problem [15]. That work highlights four modeling challenges in crowdsourcing: (1) Representing the diverse views of information contributors; (2) Representing instances of classes unknown to a data model and attributes unknown to the data model of instances; (3) Supporting unanticipated uses of data; and (4) Ensuring that contributors can provide useful data [15]. The objective- and data consumer-oriented approaches above are traditional modeling approaches, based on specialized abstractions of the real world predetermined to be useful to the task at hand. Traditional approaches fully address the fourth challenge, but only partially address challenge 3 and struggle with challenges 1-2. In contrast, emerging approaches such as those based on predetermined abstractions or on flexibility fully address challenges 1-2, partially address challenge 3, and struggle with challenge 4. To help reconcile traditional and emerging approaches to address all four challenges, the authors propose six conceptual modeling guidelines [15, p. 306-308].

These guidelines provide a robust way forward [15]. If completely adopted, conceptual models that follow them would allow a contributor to collect data at the finest possible granularity. Unfortunately, that same flexibility also suggests there is still work to do. The task at hand is to determine the appropriate level of granularity for data collection in a crowdsourcing project. This task assumes the crowdsourcing project has a particular purpose, and therefore a Target Organizational Model must exist: a conceptual model that represents the project coordinator's view of the phenomena to be captured via data crowdsourcing and how this crowdsourced data is intended to be used [15]. Thus, the question of how to design effective systems for appropriate data collection granularity remains.

4. Data granularity problems

To understand potential challenges with data granularity, it is crucial to understand that information granules do not exist in the world. We create data granules to simplify and manage information that represents the world [1]. Granulation therefore supports cognitive economy and inference [37]: by collapsing details into higher-order abstractions, data is more efficiently stored while still allowing users to infer details about the world. It is nonetheless important to note the trade-offs between coarse- and fine-grained data.

Chunking fine details into coarser granules—abstraction—leads to information loss. To explain this further, we invoke ontology [38]. An instance is a thing [39]: the most elemental construct represented by an information system. Things possess attributes [38]: attributes can have different values that dictate the attribute's current state for that thing. Classes describe groups of instances with common attributes. Given that instances are unique, no class can (or should) be a perfect fit for an instance. Therefore, if classification is used to capture an instance, the resulting data does not capture the unique quality of that instance: information loss occurs [21].

For a concrete—if crude—example, imagine a medical error reporting system at a local hospital. To save time, as hospital staff are very busy, the system uses a form to capture reports on causality whenever an error occurs. This form has pre-specified options that are quite robust, accounting for any kind of issue that might precipitate a mistake. Note, however, the use of the word "kind". Whenever medical errors are reported, the staff reporting them must provide a coarse-grained representation of the actual events. Perhaps one of these options is "Staff were distracted". This coarse-grained data could be useful—maybe management will notice that a particular unit is frequently distracted and stage an intervention. Yet intervening on the entire unit might be uncalled for. The reporting system does not provide the staff reporting the error with more detailed options, and therefore the staff are unable to report the cause of the distraction: the presence of management.

It is worth considering problems in the opposite direction, too. If classes provide cognitive economy and inferential abilities at a cost of potential information loss, then fine-grained data presents an opportunity for information gain at an economic and inferential cost. At large scales and speeds—and with

insufficient analytical capacity—it may be impossible to derive insights from fine-grained data when classes otherwise might have been useful. In an ideal world, data is collected at the finest possible level of granularity while the crowdsourcing system dynamically aggregates high resolution data into coarser granules for different purposes and contexts. The challenge is obviously dynamic, automatic analysis and aggregation. More data means more processing power is required to analyze it [28]. More data-driven products and services may also place more demands on data users, in the form of requiring new skills and knowledge to work with the data, increased scrutiny about the objectivity and accuracy of the data, questioning of the equivalences of contributions at an instance level, and ethical concerns [43]. Fine-grain data will be less structured—and there will ultimately be more of it than if it were collected at coarser resolutions—requiring additional work to prepare it for use [40-41]. Data collected at too-fine levels may also be subject to redundant or spurious findings and overfitting [29, 46]. In other words, making data bigger is not necessarily better.

All of this is to say that for a given project or purpose, there is likely an ideal level of granularity in data collection. This level of "optimal" granularity balances information loss due to classification against the analysis required to use fine-grained data. If there is an ideal level of data collection granularity, then there can be ways of ensuring that a crowdsourcing project encourages the collection of data at that ideal level. This is the subject of the next section.

5. Designing for granularity: three principles for data crowdsourcing systems

How should crowdsourcing system designers set appropriate levels of data granularity? The work on conceptual modeling in crowdsourcing projects suggests that data should be collected at the finest level possible. However, coarse levels of granularity might be sufficient for certain types of crowdsourcing projects. Further, contributor expertise and motivation complicate matters: not all contributors will be willing to contribute high quantities and qualities of data all of the time, nor are all contributors able to contribute at the same proficiency. For instance, it may frustrate contributors who provide fine-grained critiques of a product only to aggregate them into a reductive 5-star review.

These challenges highlight the need for principles for crowdsourcing data models with which to design for granularity in data collection. Recall the conceptualization of data quality as a gradient fit between the real world and the data that represents that world [21]. This conceptualization acknowledges the data contributor's role in data quality and suggests that data quality is determined by adherence to the project's conceptual model of the data it uses. Given that many data projects are nonetheless purpose-driven, principles that help design systems for granularity should help establish a compromise between data that is coarse enough to satisfy the data consumer while being fine-grained enough to minimize data loss and maximize potential reuse. These design principles should also help maximize participant contribution ability while minimizing contributor burnout (in the form of reduced contribution rate) and dropout (in the form of disengaging with the project). Likewise, these principles may help designers find symmetry between the analysis demands of fine-grained data and the information loss of coarse-grained data.

To respond to these tensions and based on the work on crowdsourcing, data science, and data quality presented above, we propose three design principles for granularity. Taking these principles into account should help system designers build data models with appropriate data fit inclusive of both the goals of the project sponsor and the unanticipated needs of potential new uses. These principles assume a project using these principles already adheres to the guidelines in [15]. This implies two important corollaries: (1) The conceptual model of the project was designed with a data-first model-after paradigm, as opposed to conventional model-first data-after paradigm [43]. This means that principles that emphasize this direction of conceptual modeling are not necessary; (2) The conceptual model of the project includes a Target Organizational Model that specifies the needs of the data consumer, and the system includes a mechanism for automatically reconciling the instance-based data collected in the project with the coarse-grained features of this Target Organizational Model. In combination with (1), this means that while the crowdsourcing project at hand can involve the collection of extremely fine-grained data because the Target Organizational Model provides available mechanisms for fitting that data to sponsor purposes. Further, the key mechanisms by which data granularity is influenced are Target Organizational Model-based cueing via examples, instructions, and other aspects of the project's contributor user experience.

5.1. Principle 1: Design for extensibility

The first principle acknowledges the potential unanticipated uses of crowdsourced data [13, 15]. Does the project exist for a specific goal, or might

support to operate the project extend beyond the initial vision? Use of this principle depends on the project sponsor. If support and resources—including time—may continue beyond the initial project focus, then designers should encourage data collection at finer granularities than the Target Organizational Model suggests. If not, the system's cues (e.g., the instructions provided to contributors, user interface designs, or the types of data contributors are allowed to input) should encourage contributors to provide data that precisely matches the Target Organizational Model. Alternatively, system designers may develop features for mixed granularity, allowing contributors to capture data at the level of specificity they prefer (and encouraging them to report details on as fine a level as possible).

The purpose, crowd, and type of crowdsourcing project may shift over time. Exciting and unanticipated results may expand the project's purpose. The project may be more popular than expected and therefore attract contributors of much greater diversity than it was designed for. Because of these shifts, the project type itself may change along any of the four dimensions of the crowdsourcing contribution typology. As these shifts occur, the level of collected granularity should shift as well. To that end, system designers should plan an iterative granularity evaluation cycle, designating points along the project's lifetime at which these principles should be reapplied to the system's design.

5.2. Principle 2: Design for the crowd

Some projects leverage established audiences, while others may be designed to recruit (or only allow) contributors with a certain level of expertise about the project goals. What kinds of contributors make up the project's crowd? For an extreme example, consider the Wikipedia article on Jimmy Wales, the founder of Wikipedia. There is a substantial difference in the motivations, expertise levels, and even perspectives of the general public, established biographers, senior editors, and Jimmy Wales himself. Yet all the above have an equal opportunity to contribute to the article, and probably have [44, see section Jimmy Wales].

Potential information loss due to coarse-grained data is particularly exacerbated when a system asks contributors to complete tasks at levels either below or beyond their capabilities [21]. Another stream of research on crowdsourcing that relates to data granularity examines data contributors—the crowd. Research in this arena explores two related factors: crowd expertise and crowd motivation.

The former tackles how to facilitate high-quality participation from contributors with low levels of domain expertise. This problem has been addressed in a way that directly relates to data granularity by developing a contributor-centered crowdsourcing system [16-17]. The system described in [16-17] enables more contributions (and more accurate contributions) using basic classes and attributes instead of asking contributors to report species-genus level classifications (e.g., Rusty Blackbird). An incidental finding was that contributors might add unexpected attributes to instance reports (e.g., "beautiful"). These bonus fine-grained data points further illustrate the potential of high-resolution data capture for data reuse and extension. Other platforms address contributor expertise through standards and training. In some of these platforms (e.g., Galaxy Zoo [11]), contributors have ready access to help documentation while contributing. Others (e.g., Stardust@home) mandate this training and even test their contributors before they are allowed to contribute. Providing guidance or requiring training and testing allows data consumers to be more certain that contributors have met certain standards of accuracy before contributing to the project [20].

A variety of researchers have also examined the impact of contributor motivation on data quality in crowdsourcing projects. Broadly, motivational studies find that intrinsic motivation is more important than extrinsic motivators in determining contribution quality of crowdsourcing data producers [6-10].

There are therefore three considerations to consider when designing for granularity:

1. *Expertise*. Contributors with limited literacy for the domain of the crowdsourcing project may find reporting specific kinds of data (such as species) difficult, leading to mistakes or disengagement [16-17, 21]. Therefore, if a project involves contributors with varying levels of expertise and data accuracy, the system's cues should encourage contributions at a finer granularity.
2. *Project motivational model*. Intrinsic motivation is a core factor in determining the quality of contributions [6-10]. If the crowdsourcing project at hand is driven by a mission that will be meaningful to contributors, those contributors will be motivated to provide higher-quality contributions [6]. To that end, the project may demand more of its contributors. Interpreted differently, providing options for fine-grained data collection may allow participants to contribute more data that is more meaningful to them, increasing motivation to contribute.
3. *Task variety, flexibility, and autonomy*. Another result of several studies of crowd motivation suggests that projects will be more motivating if contributors are able to complete a variety of tasks

with flexibility and autonomy [9-10]. This suggests that, where possible, fine-grained data collection will be more motivating to the project's crowd as they can flexibly contribute in a variety of ways. Finer granularity also fosters greater autonomy as contributors more accurately and easily report attribute-based data [15-17].

5.3. Principle 3: Design for the project type

The typology presented earlier gives system designers more options in considering the various dimensions of a potential crowdsourcing project. Each of the dimensions of the typology suggests different considerations for data granularity:

1. *Contribution definition*. Open-ended contributions have less defined purposes, and therefore their Target Organizational Model should be less stringent. This implies that crowdsourcing platforms developed to address open-ended problems should collect more fine-grained data than closed-ended problems.
2. *Contribution limits*. Short-term projects, such as Fiverr design contests, provide less time for rich fine-grained data reporting (and likely have less time for analysis of such rich data, too). Short-term projects are therefore likely to require more coarse-grained data collection to be effective.
3. *Contribution emergence*. Projects that combine contributions into a gestalt will benefit from fine-grained data collection, as the fine-grained data can be combined and recombined more flexibly than coarse-grained data.
4. *Contribution variety*. Projects that seek uniform contributions should granulate contributors' options as much as possible by providing templates that allow the data consumer to clearly and accurately infer details about the real-world objects the data represents. By considering data consumers' goals (e.g., the Target Organizational Model), project coordinators may collect more effective—yet still uniform—data by ensuring that the options available to contributors match the level of detail data consumers require. Conversely, projects with diverse contribution variety should maximize data granularity to facilitate richer data capture in contributions of differing form and content.
5. *Contribution value*. The more valuable individual contributions may be, relative to one another, the more a project may benefit from higher degrees of granularity. Allowing contributors to granulate contributions to a higher degree facilitates contribution richness, making it more likely that a valuable piece of information related to that contribution will not be lost.

6. Discussion

6.1. Contributions

This paper includes several contributions to scholarship at the intersection of crowdsourcing and data science. First, we combine two typologies of crowdsourcing systems to provide a multidimensional contribution-centered typology to characterize different crowdsourcing systems. We suggest how these dimensions might aid in the design of new crowdsourcing systems for a given purpose. Second, we extend work on conceptual modeling in crowdsourcing and user-generated content to argue for the significance of data collection granularity in crowdsourcing. Using theory and examples, we illustrate the challenges of inappropriate data granularity and suggest some consequences of leaving these challenges unchecked. Third, we propose a set of design principles for granularity in crowdsourcing data collection. The three principles are based in theory on crowdsourcing systems, the motivations of their contributors, and on the typology of crowdsourcing systems. These principles can be used by crowdsourcing system designers to make judgments about the level of granularity they should cue their participants to collect.

6.3. Future directions

The primary future direction we propose is an experimental study of the effects of granularity on the quality of crowdsourced information. Nonetheless, there are some other interesting observations that may provide fodder for future research.

As research on crowdsourcing expands, examples of crowdsourcing applications continue to proliferate. The multidimensional crowdsourcing contribution typology encapsulates a breadth of projects that might benefit from research on crowdsourcing. To that end, these projects seem increasingly less suited to conventional definitions of crowdsourcing [2]. It seems increasingly appropriate to consider crowdsourcing projects as a part of a broader subset of mass collaboration projects. This conceptualization both extends the applications that may benefit from crowdsourcing research and allows the inclusion of additional activities that could inform crowdsourcing projects. Important lessons may be learned from these activities for crowdsourcing projects (and vice versa).

The development of data science has been characterized in terms of three movements: business

intelligence and analytics 1.0, 2.0, and 3.0 [45]. 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 [45, 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 [15] 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 [46] 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 [47]. Further explorations of how to use similar artificial intelligence tools to enhance the utility of crowdsourced data is a potent area for future research.

6.4. Conclusion

Crowdsourcing is a vibrant field. The Internet, big data technologies, and other trends are rapidly unlocking new possibilities for massive, directed collaboration. Yet important issues such as data granularity remain and may stand in the way of effective use of these systems until they are resolved. In particular, the challenges of data granularity blur whether crowdsourcing systems should ask their contributors to map the forest or to identify the trees. This paper proposes a simple approach to resolving this tension in the form of design principles for granularity. It also presents a novel typology that may enrich comparisons and, therefore, future study of crowdsourcing projects.

A key limitation of this paper is that our contributions draw solely from theory and experience. A clear next step is experimental study to assess the evidence for the impact of granularity and the effectiveness of the proposed principles.

7. References

[1] Bargiela, A., & Pedrycz, W. (2003). Granular computing as an emerging paradigm of information processing. In *Granular computing: an introduction* (pp. 1–18). New York: Springer Science+Business Media, LLC.

[2] Howe, J. (2006, June 2). Crowdsourcing: A Definition. Retrieved September 20, 2018, from http://www.crowdsourcing.com/cs/2006/06/crowdsourcing_a.html

[3] Rosch, E. (1978). Principles of Categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27-48). Hillsdale, New Jersey, USA: Lawrence Erlbaum Associates, Publishers.

[4] Parsons, J. (1996). An information model based on classification theory. *Management Science*, *42*(10), 1437.

[5] Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the World-Wide Web. *Communications of the ACM*, *54*(4), 86.

[6] Chandler, D., & Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, *90*, 123–133. https://doi.org/10.1016/j.jebo.2013.03.003

[7] Nov, O., Arazy, O., & Anderson, D. (2014). Scientists@Home: What Drives the Quantity and Quality of Online Citizen Science Participation? *PLOS ONE*, *9*(4), e90375. https://doi.org/10.1371/journal.pone.0090375

[8] Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., & Vukovic, M. (2011). An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 8.

[9] Zheng, H., Li, D., & Hou, W. (2011). Task Design, Motivation, and Participation in Crowdsourcing Contests. *International Journal of Electronic Commerce*, *15*(4), 57–88. https://doi.org/10.2753/JEC1086-4415150402

[10] Kaufmann, N., Schulze, T., & Veit, D. (2011, August). More than fun and money. Worker Motivation in Crowdsourcing - A Study on Mechanical Turk. *Proceedings of AMCIS.* 11, 1-11.

[11] Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., … Vandenberg, J. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, *389*(3), 1179–1189. https://doi.org/10.1111/j.1365-2966.2008.13689.x

[12] Raddick, M. J., Prather, E. E., & Wallace, C. S. (2019). Galaxy zoo: Science content knowledge of citizen scientists. *Public Understanding of Science*, *28*(6), 636-651. doi:10.1177/0963662519840222

[13] Faniel, I. M., & Zimmerman, A. (2011). Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. *International Journal of Digital Curation*, *6*(1), 58–69. https://doi.org/10.2218/ijdc.v6i1.172

[14] Lynch, C. (2008). Big data: How do your data grow? *Nature*, *455*, 28–29. https://doi.org/10.1038/455028a

[15] Lukyanenko, R., Wiersma, Y., Huber, B., Parsons, J., Wachinger, G., & Meldt, R. (2017). Representing Crowd Knowledge: Guidelines for Conceptual Modeling of User-generated Content. *Journal of the Association for Information Systems; Atlanta*, *18*(4), 297–339.
[16] Lukyanenko, R., Parsons, J., & Wiersma, Y. (2014a). The Impact of Conceptual Modeling on Dataset Completeness: A Field Experiment. *ICIS 2014 Proceedings*.
[17] Lukyanenko, R., Parsons, J., & Wiersma, Y. F. (2014b). The IQ of the Crowd: Understanding and Improving Information Quality in Structured User-Generated Content. *Information Systems Research*, *25*(4), 669–689. https://doi.org/10.1287/isre.2014.0537
[18] Ballou, D., Wang, R., Pazer, H., & Tayi, G. K. (1998). Modeling information manufacturing systems to determine information product quality. *Management Science*, *44*(4), 462–484.
[18] Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems, 12*(4), 5–33.
[20] Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information Quality Benchmarks: Product and Service Performance. *Commun. ACM*, *45*(4), 184–192.
[21] Parsons, J., & Lukyanenko, R. (2011). Rethinking Data Quality as an Outcome of Conceptual Modeling Choices. In *ICIQ 2011 - Proceedings of the 16th International Conference on Information Quality*.
[23] [22] Geiger, D., Rosemann, M., Fielt, E., & Schader, M. (2012). Crowdsourcing Information Systems - Definition, Typology, and Design. *33rd International Conference on Information Systems, Orlando 2012*, 11.
[23] Lukyanenko, R., & Parsons, J. (2018). Beyond Micro-Tasks: Research Opportunities in Observational Crowdsourcing. *Journal of Database Management*, *29*(1), 1–22. https://doi.org/10.4018/JDM.2018010101
[24] Reiter, R. (1981). On Closed World Data Bases. *Readings in Artificial Intelligence.*, 119–140.
[25] Lohr, S. (2012, February 11). Big Data's Impact in the World. *The New York Times*.
[26] Cukier, K. (2010, February 25). Data, data everywhere. *The Economist*.
[27] The Four V's of Big Data. (n.d.). Retrieved from https://www.ibmbigdatahub.com/infographic/four-vs-big-data
[28] Jacobs, A. (2009). The pathologies of big data. *Communications of the ACM*, *52*(8), 36.
[29] Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, *28*(2), 3–28. https://doi.org/10.1257/jep.28.2.3
[30] McAfee, A., & Brynjolfsson, E. (2012). Big Data: The Management Revolution. *Harvard Business Review*, 9.
[31] Palmer, M. (2006). Data is the New Oil [Blog]. Retrieved from https://ana.blogs.com/maestros/2006/11/data_is_the_new.html
[32] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.
[33] Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data Quality in Context. *Communications of the ACM*, *40*(5), 103–110. https://doi.org/10.1145/253769.253804
[34] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide* (p. 76). The CRISP-DM Consortium.
[35] LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., & Kruschwitz, N. (2011). Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, *52*(2), 21.
[36] Kenett, R. S., & Shmueli, G. (2014). On information quality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *177*(1), 3–38.
[37] Parsons, J. & Wand. Y. (2008). Using Cognitive Principles to Guide Classification in Information Systems Modeling. *MIS Quarterly*, *32*(4), 839.
[38] Wand, Y., & Weber, R. (1990). An ontological model of an information system. *IEEE Transactions on Software Engineering*, *16*(11), 1282–1292.
[39] Parsons, J., & Wand, Y. (2000). Emancipating Instances from the Tyranny of Classes in Information Modeling. *ACM Trans. Database Syst.*, *25*(2), 228–268.
[40] Halevy, A., Rajaraman, A., Corp, K., & Ordille, J. (2006). Data Integration: The Teenage Years (p. 8). Presented at the VLDB '06, Seoul, Korea: ACM.
[41] Negash, S. (2004). Business Intelligence. *Communications of the Association for Information Systems*, *13*, 20.
[42] boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.
[43] Parsons, J. (2018, October). *Is there a Role for Conceptual Modeling in the Age of Big Data?* Presented at the 37th International Conference on Conceptual Modeling (ER 2018), Xi'an, China.
[44] Conflict-of-interest editing on Wikipedia. (2018). In *Wikipedia*. Retrieved from https://en.wikipedia.org/w/index.php?title=Conflict-of-interest_editing_on_Wikipedia&oldid=872043959
[45] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188.
[46] Provost, F., & Fawcett, T. (2013). Introduction to Predictive Modeling: From Correlation to Supervised Segmentation. In *Data Science for Business* (pp. 43–80). USA: O'Reilly Media.
[47] Lukyanenko, R., Parsons, J., Wiersma, Y. F., & Maddah, M. (2019). Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content. *MIS Quarterly*, *43*(2), 623–647.