

The Challenges of Knowledge Combination in ML-based Crowdsourcing – The ODF Killer Shrimp Challenge Using ML and Kaggle

Adrian Bumann
Chalmers University of Technology
adrian.bumann@chalmers.se

Robin Teigland
Chalmers University of Technology
robin.teigland@chalmers.se

Abstract

Organizations are increasingly using digital technologies, such as crowdsourcing platforms and machine learning, to tackle innovation challenges. These technologies often require the combination of heterogeneous technical and domain-specific knowledge from diverse actors to achieve the organization's innovation goals. While research has focused on knowledge combination for relatively simple tasks on crowdsourcing platforms and within ML-based innovation, we know little about how knowledge is combined in emerging innovation approaches incorporating ML and crowdsourcing to solve domain-specific innovation challenges. Thus, this paper investigates the following: What are the challenges to knowledge combination in domain-specific ML-based crowdsourcing? We conducted a case study of an environmental challenge – how to use ML to predict the spread of a marine invasive species, led by the Swedish consortium Ocean Data Factory Sweden using the crowdsourcing platform Kaggle. After discussing our results, we end the paper with recommendations on how to integrate crowdsourcing into domain-specific digital innovation processes.

1. Introduction

As both machine learning (ML) and crowdsourcing have gained popularity in the last years, there are increasing attempts to combine these innovation approaches through leveraging the crowd for developing ML-based solutions. For example, scholars have explored how the crowd can be employed in data annotation [1] using crowd work platforms like Amazon's Mechanical Turk. While being an important task since ML models are only as good as their data, it is also often mundane and requires little knowledge, such as labeling cars in aerial photos [2]. A slightly different approach is taken by citizen science platforms like Zooniverse where users are asked to differentiate various types of animals or plants. These tasks are more knowledge-intensive and the platforms are typically

frequented by users eager to learn and moderated by domain experts [3]. Crowdsourcing is also increasingly used to evaluate or debug existing ML models. While some tasks require contextual knowledge, such as when evaluating word clusters from newspaper articles, they are typically aimed towards a purely technically orientated crowd [4].

Research on hybrid intelligence systems investigates how human and machine intelligence can complement each other to improve algorithmic predictions. As example, Gaur et al. [5] developed an approach where users could correct mistakes made by a speech recognition software in real time. The user would read the transcript and listen to the corresponding audio, typing corrections which would then be integrated automatically. Further, several scholars have explored dynamics on ML-themed crowdsourcing platforms such as Kaggle or CrowdAI where users are asked to develop predictive ML models in competitions. These platforms emphasize both collaboration and competition among members.

While both ML-based technology and crowdsourcing benefit from the “rapid and pervasive digitization of innovation processes” [6], they also create new challenges for organizations to retrieve and combine relevant knowledge. When applying digital technologies like ML to domain-specific issues, organizations require diverse knowledge and expertise to translate domain context and logics into technical specifications and vice versa. Thus, following calls to investigate how digital innovation unfolds as dynamic problem-solution processes [6] [7] as well as how knowledge is recombined within and outside of crowdsourcing platforms [8], we developed the following research question:

RQ: What are the challenges to knowledge combination in domain-specific ML-based crowdsourcing?

To answer our research question, we investigated how one innovation consortium, Ocean Data Factory Sweden (ODF) approached a digital innovation process in the context of marine biology by integrating ML and

the knowledge of the crowd. More specifically, we conducted an in-depth single case study [9] of one challenge – how to predict the spread of an invasive species using ML and the crowdsourcing platform Kaggle. Below we provide an overview of insights from relevant literature before presenting how we structured and analyzed our case study. Subsequently, we present our results, followed by a discussion of theoretical and practical implications of our findings.

2. Background

The role of knowledge as an organizational resource for successful innovation outcomes is well acknowledged [10]. Digital technologies change this notion twofold. First, they facilitate the acquisition of external knowledge by drastically decreasing costs for information processing and communication [11]. This has led to more organizations opening their innovation process to exploit external knowledge [12] [13]. Second, as digital technologies are characterized as generative and convergent, digital innovation processes have become less bounded and more distributed across disciplinary, organizational and other boundaries [7]. For example, the application of ML has enabled benefits across industries, ranging from simply increasing operational efficiency to transforming business models. However, depending on the use case, innovators require specialized technical knowledge to apply those tools. Both a computer vision algorithm to analyze drone images in agriculture and the Netflix recommendation algorithm are based on similar statistical methods used in ML but may require very different data sources, software, and knowledge to develop them. As a result, the actors, capabilities and thus the knowledge required to successfully innovate using digital technologies have become more heterogeneous [6] [14].

In the following, we discuss knowledge combination in crowdsourced innovation contests and AI/ML-based innovation processes, with a specific focus on related challenges.

2.1. Knowledge Combination in Crowdsourcing

Crowdsourcing is a broad term and closely related to concepts such as citizen science, open democracy or crowd work. We focus on crowdsourcing in the form of innovation contests that involve an organization (the seeker) broadcasting a problem statement to a large group of self-selecting participants (the solvers) and proposes some form of reward for solutions meeting pre-defined criteria [15], [16]. Often, this involves an innovation intermediary that hosts the challenge on an

online platform and maintains an existing network of problem solvers. While such contests have been traditionally used to solve “technical, innovation-related” problems specific to a single organization, such as developing a recommendation algorithm for Netflix [15], they are increasingly employed to tackle complex societal challenges, such as environmental protection or public health [16]. Such challenges involve many different stakeholders and rarely hold a simple solution. Thus, it is difficult to know in advance which knowledge will be required. Innovation contests allow access to a broad variety of actors which, in many cases, are encouraged to collaboratively develop and combine their ideas on innovative solutions [17] which makes them an increasingly popular choice among organizations tackling these challenges. This also allows organizations to gather ideas much faster than possible in traditional organizational settings involving only people inside the organization [18]. The most recent example includes the “Global Hack” initiative, a series of innovation contests in 53 countries which called upon the public to develop novel solutions in the face of the Covid-19 pandemic. The initiative attracted participants with vastly different backgrounds and produced over 5000 ideas, ranging from 3-D printed face shields to tracking apps [19].

Prior to the innovation contest, the seeker organization needs to identify an issue and generate a problem statement, i.e. a description of the problem and criteria for the desired solutions [20]. This initial problem formulation already requires different internal actors to collaborate and to exchange problem-related knowledge [21]. Problem-related knowledge depends on the technical, physical and social context of the problem and is described as tacit or “sticky” knowledge [10]. Thus, the problem-related knowledge gained by individuals in one context cannot be shared with others in a different context without effort, e.g. discussions, experiments or the use of boundary objects [10] [22]. This leads to two challenges for seeker organizations when generating a problem statement. First, they need to encode their own tacit knowledge in a written problem statement. Second, they need to make additional problem-related knowledge available to external solvers with different backgrounds [21]. The way a problem is formulated influences what solutions the solvers can provide [23]. Boudreau et al. [17] found that problems that are loosely defined and require the input of multiple knowledge domains attract more competitors and increase the chance of finding valuable solutions while the opposite is true for single-domain problems.

While online platforms for innovation contests usually offer various technical functions to retrieve and share problem-relevant knowledge, there are various

factors that enable effective knowledge transfer amongst participants. Especially when the crowd has diverse expertise, integrating and translating ideas can be enabled through boundary-spanning practices [24]. Many platforms have discussion boards where participants can exchange task-relevant information. The role of the seeker organization then is to moderate and foster a collaborative and supportive culture, as well as to give clear guidance on desired outcomes [16] [25].

2.2 Knowledge Combination in ML-based Innovation

In organizational sciences, ML has been named as one of many tools on a broader level how organizations work with and gain potentially valuable insights from big data [1]. Data are often described as the raw resource which can be transformed into information and, when contextualized, knowledge [26]. What knowledge is produced from data depends however on those analyzing the data, i.e. data scientists. While data scientists are trained in statistics and analytical skills, their acquired domain knowledge and personal mindset influence how they approach and make sense of data [27]. Bholat [28] describes how inductive and deductive analytical approaches to the same datasets can produce very different knowledges. As data are often used for different purposes than they were originally collected for [29], data scientists need to not only understand the data but also the context of data collection to make sense of it, and they need to make educated assumptions when conditions are unknown [30]. In a study of the role of business intelligence in knowledge creation, industry professionals agreed that data should be supplemented with common sense, human intelligence and domain knowledge which are difficult to capture through data [31]. In a similar stance, Jagadish et al. [32] argue that processes for data analysis “will be designed explicitly to have a human in the loop”.

Organizations have employed different approaches to incorporate specialized data science knowledge in their innovation processes, with many organizations setting up internal, multidisciplinary teams with diverse sets of competencies [33]. As AI and ML have become increasingly popular in the last decade, some have noted exaggerated expectations towards the potential value those technologies can bring similar to previous IT innovations which were deemed the “latest fashion” [34]. In the case of IBM in healthcare, IBM’s CEO announced that AI “can change almost everything about health care,” and speculated about a medical “golden age”. However, after seven years, only five out of 24 projects produced any lasting results. Most projects failed to apply IBM’s technical capabilities in the complex domain environment of a medical ward [35].

Thus, multidisciplinary teams engaged in developing ML-based innovations need not only to translate a domain-specific issue into technical terms, but also understand the technical limitations. For that reason, Anderson [36] describes key skills for data scientists not being limited to technical knowledge but also “to learn on the fly and to communicate well in order to answer business questions, explaining complex results to nontechnical stakeholders”.

Central to this paper is the notion that developing ML models for domain-specific problems requires the collaboration of domain and data experts. While there is little research on this specific phenomenon, there is a large existing body of literature on cross-domain collaboration and its challenges, focusing on organizational dynamics and emerging tensions in blending diverse domains of expertise [37] [38]. Domain specialists often possess complex, ambiguous, and tacit knowledge which is difficult to transmit and requires mutual understanding [39]. Therefore, such collaboration does not only require knowledge transmission but also knowledge translation and transformation [10]. Domain specialists also often have diverging logics, taxonomies, and beliefs based on their professional background. The term “noise” can have completely different meanings to an engineer or a data scientist. Dougherty [40] describes this as distinct “thought worlds” which she defines as “a community of persons engaged in a certain domain of activity who have a shared understanding about that activity”. These thought worlds are coherent and make sense for someone within that domain but may be difficult to decipher for others. Traversing these thought worlds becomes even more difficult when domain experts innovate using novel and complex technologies like ML. It is rarely clear which knowledge is required to solve complex problems [16] and the intangible nature of digital tools complicates the use of tangible artifacts that can serve as boundary-spanning tools [41].

As more and more organizations take advantage of the potential provided by crowdsourcing and ML, the knowledge required to integrate them into organizational innovation processes becomes more diverse. We have shown that there is a vast body of literature on the challenges of knowledge combination in crowdsourced innovation contests and developing ML-based solutions. We argue however that current literature does not show how these two aspects of digital innovation are integrated and how knowledge is combined in emerging innovation approaches incorporating ML and crowdsourcing to solve domain-specific innovation challenges. Thus, we aim to answer following research question:

RQ: What are the challenges to knowledge combination in domain-specific ML-based crowdsourcing?

3. Methodology

To investigate our research question, we employ a single-case study research design [9]. More specifically we conducted a study of the Killer Shrimp Challenge on Kaggle by Ocean Data Factory Sweden (ODF). ODF is a Swedish triple-helix consortium with the aim to “enable Sweden to be a global leader in sustainability and innovation in the global digital blue economy” (see also: www.oceandatafactory.se). The goal is to tackle environmental challenges using open data and machine learning. Founded in August 2019, ODF is one of 16 ongoing data-driven innovation projects, dubbed “data labs”, funded by Sweden’s government agency for research and innovation Vinnova. ODF describes itself comprised of actors from the public, private and academic sectors with backgrounds ranging from organizational, marine, maritime and data science.

Kaggle is an online community with a focus on data science and machine learning. Founded in 2010 and acquired by Google in 2017, it has grown to >1 million registered users. A central function of the platform is that users, alone or in teams, can solve challenges provided by external organizations. Performance is evaluated based on the accuracy of the users’ prediction algorithm. Users can climb an internal ranking depending on their performance and some challenges include monetary rewards. An example challenge is the Deepfake Detection Challenge by AWS, Facebook, Microsoft and academics with 33,007 entries by 2281 teams competing for \$1 million in prizes (see also: www.kaggle.com).

To achieve triangulation, we used multiple data sources [42]:

- In-depth, semi-structured interviews with 13 team members of ODF, ranging from 0,5 to 1 hour each (October 2019 – May 2020)
- In-depth, semi-structured interviews with six participants of the crowdsourcing challenge, ranging from 0,5 to 1 hour each (May 2020)
- Analysis of the websites of ODF (August – June 2020) and Kaggle (March – June 2020)
- Participant observation and meeting notes of biweekly internal ODF meetings (August 2019 – June 2020)
- Internal mail and chat communication of ODF

Interview guides for ODF members were based on four principal themes – individual background, expertise, involvement within ODF, and perspectives on

organizational and technical processes within ODF. Interview guides for Kaggle participants focused on individual background and expertise, motivation to join the challenge, individual problem-solving process and resources used during the process. All interviews were recorded, transcribed using the transcription service otter.ai and edited for accuracy. A thematic analysis [43] of the material was performed to explore how teams with diverse domain expertise and participants on a crowdsourcing platform work and collaborate in developing ML-based solutions for environmental issues in marine science. Data analysis started with a thorough reading of the interview transcripts in order to obtain familiarity with the data. Initial codes rooted in the data were constructed and compared to each other for repeated patterns and then grouped together based on shared meaningful content. These clusters formed the basis for the identified themes.

4. The Killer Shrimp Challenge

4.1 Initial Situation

The digitization in environmental sciences has led to large amounts of open data available on public data portals. However, many organizations struggle to make use of the data due to low data quality or lack in data science expertise [44]. ODF took this as starting point to find a use case that was acute, promising, and where there were enough suitable data available. Eventually, it was decided to build an ML algorithm to predict the spread of a certain invasive species, dubbed the “Killer Shrimp”, that caused severe environmental damage in the Baltic Sea. This was based on the input of project members and an early assessment of available data. This was followed by a phase where ODF members continuously explored for more open data and discussed amongst each other to translate the domain problem into technical terms. Some crucial data were found missing for which the team made workarounds or educated assumptions. The final result was an aggregated dataset of approximately 2.8 million data points and an ML model which predicted the species spread under different conditions.

4.2 The Kaggle Challenge

While the original project goal was developing the ML model internally, ODF members noted in January 2020 that the aggregated data set could be a valuable resource for other digital innovators. Thus, it was decided to design an innovation contest on Kaggle. The goals were three-fold: first, a broad public, including marine and data scientists, should be given the

opportunity to learn about the problem of invasive species and develop digital solutions. Second, results could produce additional technical knowledge in the form of new ML models. Third, this could provide valuable insights for other data labs on how to include crowdsourcing in digital innovation processes. ODF launched the competition in March 2020 (Figure 1) with a duration of three months. The “Killer Shrimp Invasion” challenge included a description of both the marine context and the data and encouraged explicitly marine and data scientists to join. The objective was to develop predictive ML models based on the data, with a reward of 150€ for the best submission. Data were provided in the form of a training and a test set in .csv format as well as an example submission. The challenge was promoted mainly via social media accounts of the involved organizations and in Kaggle-themed communities on the social news website reddit. Eventually, 30 users participated, uploading 221 submissions in total.

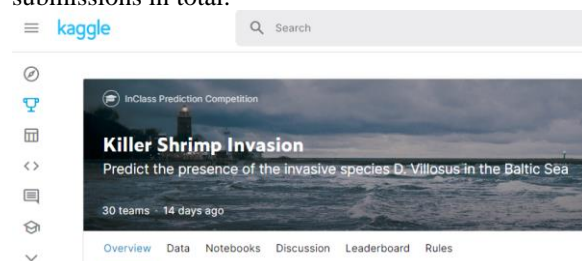


Figure 1. Kaggle challenge header

5. Findings and Insights

In this section, we present the core findings and insights gained from this case regarding to how the knowledge is combined throughout the phases of ML-based innovation employing crowdsourcing. First, we describe how ODF translated the domain problem of invasive species into a technical problem. Second, we give an overview of the dynamics during the crowdsourcing process. Third, we analyze the interplay between ODF and the challenge participants. Finally, we derive recommendations on how to integrate crowdsourcing into domain-specific digital innovation processes.

5.1 Problem Formulation and Translation

The initial phase was characterized by an exploratory attitude where ODF members combined their knowledge to simultaneously explore current issues in the marine environment and whether internal capabilities and available data were adequate to produce a suitable solution. The decision to focus on invasive species in the Baltic Sea was largely guided by a few

team members who had worked with the topic before and saw it as both technically feasible and interesting to external stakeholders like government agencies. The team also followed two technical goals defined by ODF leadership in order to demonstrate the benefits of open data and allow others to reproduce the results: 1) the final result should include AI or ML, and 2) the process should adhere to the FAIR principles, i.e. any used data should be findable, accessible, interoperable and reusable. The following phase of problem definition was characterized by intense and close discussions on how to translate the overall issue – certain invasive species causing harm in the Baltic Sea – into a specific problem solvable with an algorithm. The team went through iterative cycles of identifying parameters that were relevant in describing the context (such as physical conditions that made an area suitable for the species) and open data that represented these parameters. Team members noted the importance of having and exchanging diverse knowledge during this phase. Commenting on the final outcome, an ODF data scientist noted *“what I think works well is that we have this group of people from diverse backgrounds. There's a constant knowledge transfer happening, and I think we've done really well to incorporate everyone's expertise and their ideas in the final products.”*

This problem formulation process did not follow a clearly defined path, but resembled what Nambisan [6] describes as “dynamic problem-solution design pairing”, going through multiple problem-solution constellations until a clear goal was set. During an internal meeting, a team member who had ties to an environmental government agency described how a dashboard visualizing different data types could help their operation. This was discussed as a viable use case until a data scientist remarked that this would not necessarily require the use of ML, upon which the team searched for technologically more advanced applications. Interestingly, although the issue of invasive species was rooted in marine science, the translation into technical specifications also required knowledge from other domains. When ODF wanted to represent in their model how the species would spread across the Baltic Sea they noted that most invasive species spread via the ballast water carried onboard merchant ships. However, finding suitable data of relevant shipping routes and locations of ballast water exchange proved more difficult than thought. Eventually, this data limitation was overcome with the help of one team member who was doing academic research in the shipping sector by leveraging his knowledge on ports in the Baltic Sea. Instead of shipping data, the model now included a list of relevant ports and mapped inland presence data of the species to those ports. While the team acknowledged this as a

potentially unrealistic assumption, it was deemed suitable enough for ODF's purpose, with a data scientist noting it would be easier to include this assumption in the model and implement suitable data once it emerged than to keep on searching.

This reflected a general trend that ODF members had different perceptions of what the team could realistically achieve with the available data. This in turn influenced how they approached the process of translating the environmental problem in a technical problem. Non-data scientists tended to be more practically orientated, thinking about developing solutions which would be highly accurate and ready to use for external stakeholders. Data scientists on the other side advocated for a more iterative, experimental approach to understand what solutions were feasible. One data scientist described it as *"with a little bit of data that we had initially, we already started to test some models. Then, as [more] data came in, we started to adjust the way that we thought about our models so we could still keep working and make progress without having too much of a dependency on that 'perfect data' that we're waiting for. Because [...] in the end you're never going to have this [perfect data]."* Similarly, a non-data scientist reflected in the end of the project that he underestimated the technical challenges and that one of his personal learnings was that *"we didn't have to revolutionize everything. [instead] we narrowed [the solution] down, and the result was still very useful"*.

Eventually, the team aggregated the data they collected and built a ML model. When presenting the ML model in the form of a python notebook to team members who were not directly involved in the development process, most struggled to understand its function. The understanding process was then helped when a data scientist acted as knowledge broker and built a simple visualization app using the platform Heroku. Broken down into two selection options and a chart of the Baltic Sea showing the presence of the species (Figure 2), the visualization app served as a boundary object to illustrate the mechanics and outcomes of the ML model. As one marine scientist put it, *"this really helps me understand what you were doing"*. The same person remarked that the visualization helped also to convey ODF's activities to external marine science organizations, stating that *"this is something I can show [to others]"* and reactions were positive, partially because it felt *"more substantial"* than comparable projects which only contained technical descriptions.

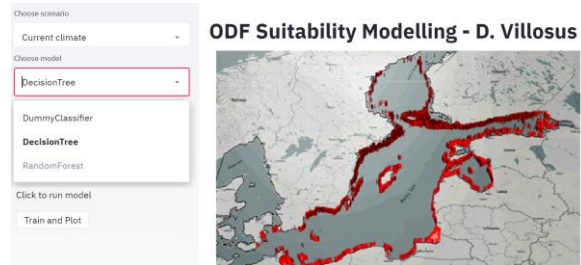


Figure 2. Visualization app of ODF's ML model functionality

5.2 Analysis of the General Crowdsourcing Process

Even though the public promotion of the crowdsourcing challenge was aimed at a broad audience, with the specific goal to attract both data and marine scientists as well as anyone without specific knowledge in either and just "generally passionate about the ocean" (challenge description), participants were overwhelmingly data scientists. Of the 30 participants, 23 had a specific data science background, and all 30 had been active on Kaggle before. Although no dedicated interviews were done with external marine scientists who did not join the challenge, ODF members without data science knowledge indicated they were intimidated by the technical difficulty.

Those who participated stated they became aware of the challenge either by browsing on Kaggle or through posts in Kaggle-themed communities on reddit. While all interviewees stated being generally in favor of environmental protection, none saw it as the primary motivation. Instead, most were drawn to the challenge from the fact that its practical, domain-specific goal stood out from other challenges on the platform which most described as too hypothetical or profit-driven in comparison. This was further aided by the challenge title "Killer Shrimp Invasion" which participants described as catchy and capturing their attention while browsing other challenges. The monetary reward of 150€ for the challenge winner was seen as negligible by participants.

"the biggest influence was: is it a real-world problem? Are we trying to differentiate cats from dogs or cats from dogs? Differentiation is an interesting problem [but] it doesn't have a lot of relevance for the real world. But here you have a nonprofit trying to solve a real problem, where the solution is going to have a real impact." (Kaggle participant)

"I'm interested in these more academic, more environmental kind of challenges, [as opposed to] boring challenges like credit card fraud." (Kaggle participant)

Other participants noted that this challenge was an interesting opportunity to work with other types of data

than they normally worked with. “[in my job, I’m working with] customer behaviors data.[...]This competition sounded very academic, which I like, so I was quite excited.[...]this kind of Killer Shrimp, it’s data regarding the nature. The purpose of the challenge is not to make money, but to get insight into that phenomenon.” (Kaggle participant).

Interestingly, the domain context, while being a motivating factor and entirely unrelated to the participants’ professional backgrounds, had little impact on how the participants approached the problem. Participants gathered enough background information to understand why the issue was relevant and what solution was required. But when they started the solving process, they mainly explored the data to make sense of the technical problem and how to approach it, applying methods they had used before in other contexts. The visualization app in Figure 2 was included in the challenge description. Some participants stated that they used it out of curiosity, and it helped them to understand “what [ODF] is after”. But in contrast to marine scientists, participants did not need it to make sense of the technical challenge. Some participants acknowledged that the physical nature of the data made it easier to picture the context and stay motivated, but it did not influence their work process. “Of course, reading parameters like salinity, wave exposure, it’s more interesting than working with X1, X2. You know what those are. But [it] doesn’t change what methods I use.” (Kaggle participant)

When asked why the domain context had such little influence on their work, participants named the way challenges on Kaggle were structured as a major reason. Participants only can use the already compiled datasets they are provided, and the outcome is automatically evaluated by a quantitative measure. A participant noted that this type of challenge, even though rooted in a real-world context, did not equate to solving a real-world problem: “It’s a simple process. I’m just strictly modeling all the data which is already aggregated. It’s not like the competition is not [set] in the real world. But to me, it’s a modeling challenge, not solving a complete problem.”

Participants had mixed opinions on whether they liked this structure. Some expressed that they would prefer a more open challenge which would include participants in the initial problem generation and data aggregation. For one, it was seen as opportunity to foster collaboration with domain specialists which in turn would lead to more useful results: “It would be a good experience to work with [those] who actually understand the problem.[...]Like this, how do I know my model is actually going to be useful?” The process of aggregating data and understanding which data was relevant in reflecting the real-life problem was also

acknowledged as crucial skill for data scientists. As for most, the motivation to participate was to train their own skills, not including this preparatory phase was described as a missed learning opportunity. “80% of the work is cleaning, preparing, labeling the data.[...]You don’t learn that on Kaggle”. However, all participants acknowledged that such a broader challenge would require more time and resources both for participants and organizers. Subsequently, several participants stated they would not have joined such an open challenge as it would have taken too much time and their main focus was on improving their skills, not on becoming part of a larger endeavor.

As there is the possibility to work in teams on Kaggle, ODF had initially expected that participants with different backgrounds would team up to solve the challenge using their combined knowledge. However, all participants ended up working on their own. This was both due to the problem statement and the platform design. Participants claimed they perceived the challenge simple enough that they could solve it on their own, given that the main focus was on building a model which would result in a high prediction score. Further, participants described the process of finding potential team mates as unclear. Although public user profiles on Kaggle display one’s experiences and scores in previous challenges and users can include job title or skills, participants found it difficult to know who of the other participants would have suitable skills to complement their own. Instead, knowledge transfer among participants happened mainly on the challenge-internal discussion board and focused on what techniques were most appropriate to use for the data set.

5.3 Interplay of Seeker and Solvers

During the challenge, one of ODF’s data scientists moderated the challenge-internal discussion board. Participants noted this as generally positive and motivating as this allowed them to clarify uncertainties. Those interactions also revealed some of the challenges the participants encountered arising from the characteristics from the data. Instead of using the domain-specific context, participants used the provided data sets to make sense of the problem and plan their solving process. Most stated that their typical work process on Kaggle started with “playing with the data”, using exploratory data analysis to produce visual graphs and understand its main characteristics. This gave them an understanding of what they were working with and what methods would be suitable. However, the datasets provided in the Killer Shrimp challenge had a strong imbalance which some participants found unusual. Some participants asked ODF how the data was collected and how certain parameters like wave

exposure were calculated in order understand this imbalance. In that sense, it was less the context of the physical marine environment and more the context of the marine data which was necessary to help participants make sense of their results. In the words of a participant: *“Understanding the intent and the science behind data is just as important as understanding the data”*. Accordingly, much of the discussion on the board revolved around finding suitable techniques for this dataset. Some expressed also frustration as they saw this imbalance as a barrier to produce useful results.

As all of the challenge participants had previous experience in Kaggle challenges, that experience also allowed them to detect characteristics in the data which ODF had not considered before. In the early phase of the challenge, participants found an exploit which allowed them to achieve a perfect score on the leaderboard with little effort. As the initial competition rules stated that submissions would be rated only on the leaderboard score, participants voiced concerns on the discussion board that users might submit impractical ML models only to receive the prize money. ODF reacted by introducing an additional rule that all participants had to submit their code and that the challenge winner would be determined both by their quantitative score and a qualitative evaluation of their technical approach. This was met with favorable feedback. In interviews, participants highlighted their motivation to try and learn new technical methods and to produce results that would be helpful to ODF and other organizations, and this rule change would help to achieve these goals.

“I think it's absolutely fair. Participants should stand by their work. [...] The result should be useful.” (Kaggle participant)

“I enjoy learning about all of this. So [the rule change] is good.” (Kaggle participant)

During and after the challenge, ODF team members used leaderboard scores and submitted notebooks with code to make sense of how they could utilize the results. Even though the Kaggle platform is designed to automatically identify the best results through its quantitative ranking, ODF realized it required technical knowledge to evaluate the usefulness of the submissions. When one participant submitted a notebook exploiting the data imbalance as mentioned above, one ODF member without a data science background wrote enthusiastically in the internal ODF chat *“New Killer Shrimp leader today, from Stockholm, score 1.00000! Looks like someone we should invite [to share his insights]”*, assuming that a high score equated to a useful solution. Even without the exploit, a ODF data scientist noted the benefit of a more qualitative evaluation to identify not only the most technically advanced ML solution but also the one which provided the most practical value for ODF. This however required

good understanding of both the data science methods used in the ML models and the organizational goals of ODF. There again, one of the data scientists acted as a knowledge broker, examining and communicating the usefulness of the submissions to other ODF members.

5.4 Derived Recommendations for Integrating Crowdsourcing into Digital Innovation Processes

Based on the identified challenges, we derive several key recommendations on how to integrate crowdsourcing into domain-specific digital innovation processes. To do so, we draw on measures communicated as successful by involved actors in the Killer Shrimp challenge as well as on the analysis of our collected data. First and foremost, a thorough internal problem formulation process is crucial in understanding the domain problem and translating it into technical terms. This phase does not necessarily follow a clear line and it is not obvious in advance what expertise will be required. Thus, organizations can benefit from including actors with highly diverse knowledge and adopting an agile, exploratory work approach to explore multiple opportunities. The knowledge combination is aided by intense and close collaboration and the ability of single actors to broker knowledge across domain boundaries. Second, the crowdsourcing platform and its crowd should suit the formulated problem and desired outcomes. In the case of Kaggle, the domain-specific problem helped attract talent and create intrinsic motivation, while the platform design encouraged technical solutions and reduced the need for contextual domain knowledge but also restricted the solution space. This may be different on platforms that attract a more diverse crowd and promote more open challenges. Third, moderation during the challenge helps motivation and to reduce uncertainties on technical questions or the challenge objectives. Fourth, given the abstract nature of digital technologies, boundary objects in the form of data visualization and apps can help actors to make sense of data and ML models. Finally, organizations can benefit from evaluating the quality of ML-based solutions not only based on quantitative measures but also using an in-depth qualitative evaluation. This again requires a diverse set of knowledge, and it can be beneficial to include challenge participants in presenting their findings themselves.

6. Discussion & Conclusion

In this paper, we described how heterogenous knowledge was combined in a process to develop ML-based solutions, involving both diverse experts and

participants in a crowdsourced innovation contest. Using the case of the innovation consortium ODF Sweden and the crowdsourcing platform Kaggle, we explored in this paper challenges arising from drawing upon the wisdom of the crowd in domain-specific digital innovation processes. To derive proper recommendations, we explored the problem formulation and translation process within ODF, how Kaggle users approached and made sense of the posed challenge, the dynamics between ODF and Kaggle users, and subsequently delivered key recommendations on how to integrate crowdsourcing into domain-specific digital innovation processes for both organizations that would like to use a similar innovation process and platform operators. With this paper, we extend current knowledge in the realms of IS, knowledge management, and platform ecosystems by shedding light on a phenomenon that has not been in the main focus of research in these areas yet. We also contribute to practice as organizations developing domain-specific ML-based solutions gain insights on how to potentially implement crowdsourcing in their processes.

As every research, our work also faces limitations. While a single-case study can provide an-depth examination of a phenomenon, it is limited in its external validity [43] and we encourage further research on diverse knowledge combination in crowd-enabled, AI/ML-based innovation in other settings. For example, cases set in healthcare or agriculture might provide interesting new insights. As shown earlier, the platform Kaggle allows challenges to be posed and evaluated in a streamlined way. Further, the sample size of interviewed participants was relatively small and participants had similar backgrounds in data science. It would be interesting to test the findings of this paper in an innovation contest with a more open problem statement and a more diverse crowd. Similarly, future researchers might investigate similar challenges in physical innovation contests like hackathons.

7. References

- [1] V. S. Sheng and J. Zhang, "Machine Learning with Crowdsourcing: A Brief Summary of the Past Research and Future Directions," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, Art. no. 01, Jul. 2019, doi: 10.1609/aaai.v33i01.33019837.
- [2] E. Strickland, "In the Coming Automated Economy, People Will Work for AI," *IEEE Spectrum: Technology, Engineering, and Science News*, 2018. <https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-the-coming-automated-economy-people-will-work-for-ai> (accessed Jul. 13, 2020).
- [3] J. Reed, M. J. Raddick, A. Lardner, and K. Carney, "An Exploratory Factor Analysis of Motivations for Participating in Zooniverse, a Collection of Virtual Citizen Science Projects," in *2013 46th Hawaii International Conference on System Sciences*, Jan. 2013, pp. 610–619, doi: 10.1109/HICSS.2013.85.
- [4] J. W. Vaughan, "Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research," *Journal of Machine Learning Research*, p. 46, 2018.
- [5] Y. Gaur, F. Metze, and J. P. Bigham, "Manipulating Word Lattices to Incorporate Human Corrections," Sep. 2016, pp. 3062–3065, doi: 10.21437/Interspeech.2016-660.
- [6] S. Nambisan, K. Lyytinen, A. Majchrzak, and M. Song, "Digital Innovation Management: Reinventing Innovation Management Research in a Digital World," *MISQ*, vol. 41, no. 1, pp. 223–238, Jan. 2017, doi: 10.25300/MISQ/2017/41:1.03.
- [7] Y. Yoo, R. Boland, K. Lyytinen, and A. Majchrzak, "Organizing for Innovation in the Digitized World," *Organization Science*, vol. 23, pp. 1398–1408, Sep. 2012, doi: 10.1287/orsc.1120.0771.
- [8] C. L. Tucci, A. Afuah, and G. Viscusi, Eds., *Creating and capturing value through crowdsourcing*, First edition. Oxford: Oxford University Press, 2018.
- [9] R. K. Yin, *Case study research and applications: design and methods*, Sixth edition. London: SAGE, 2018.
- [10] P. R. Carlile, "A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development," *Organization Science*, vol. 13, no. 4, pp. 442–455, Aug. 2002, doi: 10.1287/orsc.13.4.442.2953.
- [11] K. Trantopoulos, G. Von Krogh, M. W. Wallin, and M. Woerter, "External knowledge and information technology: Implications for process innovation performance," *MIS Quarterly: Management Information Systems*, vol. 41, no. 1, pp. 287–300, 2017, doi: 10.25300/MISQ/2017/41.1.15.
- [12] H. W. Chesbrough, *Open innovation: the new imperative for creating and profiting from technology*. Boston, Mass: Harvard Business School Press, 2003.
- [13] E. Von Hippel, *Democratizing Innovation*. Cambridge, MA: MIT Press, 2005.
- [14] D. Dougherty and D. D. Dunne, "Digital Science and Knowledge Boundaries in Complex Innovation," *Organization Science*, vol. 23, no. 5, pp. 1467–1484, Oct. 2011, doi: 10.1287/orsc.1110.0700.
- [15] A. King and K. R. Lakhani, "Using Open Innovation to Identify the Best Ideas," *MIT Sloan Management Review*, p. 11, 2013.
- [16] A. Majchrzak and A. Malhotra, "Unleashing the Crowd: Overcoming the Managerial Challenges," in *Unleashing the Crowd: Collaborative Solutions to Wicked Business and Societal Problems*, A. Majchrzak and A. Malhotra, Eds. Cham: Springer International Publishing, 2020, pp. 265–293.
- [17] K. J. Boudreau, N. Lacetera, and K. R. Lakhani, "Incentives and Problem Uncertainty in Innovation Contests: An Empirical Analysis," *Management*

- Science*, vol. 57, no. 5, pp. 843–863, Apr. 2011, doi: 10.1287/mnsc.1110.1322.
- [18] M. Bogers, H. Chesbrough, S. Heaton, and D. J. Teece, “Strategic Management of Open Innovation: A Dynamic Capabilities Perspective,” *California Management Review*, vol. 62, no. 1, pp. 77–94, Nov. 2019, doi: 10.1177/0008125619885150.
- [19] H. Buisman, “Global Hack: Coming together to find solutions,” *Observatory of Public Sector Innovation*, Apr. 08, 2020. <https://www.oecd-opsi.org/globalhack/> (accessed Jul. 15, 2020).
- [20] K. Lakhani, L. Jeppesen, P. Lohse, and J. Panetta, “The Value of Openness in Scientific Problem Solving,” *Harvard Business School WP, No. 07-050*, 2007, Jan. 2007.
- [21] M. W. Wallin, G. von Krogh, and J. H. Sieg, “A problem in the making: How firms formulate sharable problems for open innovation contests,” in *Creating and capturing value through crowdsourcing*, Oxford University Press (OUP), 2018, pp. 127–153.
- [22] K. D. Knorr-Cetina, *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Elsevier, 2013.
- [23] D. C. Brabham, “Crowdsourcing as a model for problem solving: An introduction and cases,” *Convergence*, vol. 14, no. 1, pp. 75–90, 2008.
- [24] N. Levina and A.-L. Fayard, *Tapping into Diversity Through Open Innovation Platforms: The Emergence of Boundary-Spanning Practices*. Oxford University Press.
- [25] L. B. Jeppesen and K. R. Lakhani, “Marginality and Problem-Solving Effectiveness in Broadcast Search,” *Organization Science*, vol. 21, no. 5, pp. 1016–1033, Oct. 2010, doi: 10.1287/orsc.1090.0491.
- [26] N. L. Henry, “Knowledge Management: A New Concern for Public Administration,” *Public Administration Review*, vol. 34, no. 3, pp. 189–196, 1974, doi: 10.2307/974902.
- [27] R. Sharma, S. Mithas, and A. Kankanhalli, “Transforming decision-making processes: a research agenda for understanding the impact of business analytics on organisations,” *European Journal of Information Systems*, vol. 23, no. 4, pp. 433–441, Jul. 2014, doi: 10.1057/ejis.2014.17.
- [28] D. Bholat, “Big Data and central banks,” *Big Data & Society*, Apr. 2015, doi: 10.1177/2053951715579469.
- [29] I. D. Constantiou and J. Kallinikos, “New Games, New Rules: Big Data and the Changing Context of Strategy,” *Journal of Information Technology*, Mar. 2015, doi: 10.1057/jit.2014.17.
- [30] N. Kshetri, “The emerging role of Big Data in key development issues: Opportunities, challenges, and concerns,” *Big Data & Society*, Dec. 2014, doi: 10.1177/2053951714564227.
- [31] A. Shollo and R. D. Galliers, “Towards an understanding of the role of business intelligence systems in organisational knowing,” *Information Systems Journal*, vol. 26, no. 4, pp. 339–367, 2016, doi: 10.1111/isj.12071.
- [32] H. V. Jagadish *et al.*, “Big data and its technical challenges,” *Commun. ACM*, vol. 57, no. 7, pp. 86–94, Jul. 2014, doi: 10.1145/2611567.
- [33] F. Svahn, L. Mathiassen, and R. Lindgren, “Embracing Digital Innovation in Incumbent Firms: How Volvo Cars Managed Competing Concerns,” *MIS Quarterly*, vol. 41, no. 1, pp. 239–254, Mar. 2017.
- [34] P. Wang, “Chasing the Hottest IT: Effects of Information Technology Fashion on Organizations,” *MIS Quarterly*, vol. 34, no. 1, pp. 63–85, 2010, doi: 10.2307/20721415.
- [35] E. Strickland, “How IBM Watson Overpromised and Underdelivered on AI Health Care - IEEE Spectrum,” *IEEE Spectrum: Technology, Engineering, and Science News*, 2019. <https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care> (accessed Jun. 06, 2020).
- [36] H. Bowne-Anderson, “What Data Scientists Really Do, According to 35 Data Scientists,” *Harvard Business Review*, Aug. 15, 2018.
- [37] A. Majchrzak, P. H. B. More, and S. Faraj, “Transcending Knowledge Differences in Cross-Functional Teams,” *Organization Science*, vol. 23, no. 4, pp. 951–970, Jul. 2011, doi: 10.1287/orsc.1110.0677.
- [38] J. Mengis, D. Nicolini, and J. Swan, “Integrating knowledge in the face of epistemic uncertainty: Dialogically drawing distinctions,” *Management Learning*, Sep. 2018, doi: 10.1177/1350507618797216.
- [39] B. Kogut and U. Zander, “Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology,” *Organization Science*, vol. 3, no. 3, pp. 383–397, 1992.
- [40] D. Dougherty, “Interpretive Barriers to Successful Product Innovation in Large Firms,” *Organization Science*, vol. 3, no. 2, pp. 179–202, May 1992, doi: 10.1287/orsc.3.2.179.
- [41] D. Dougherty and D. D. Dunne, “Digital Science and Knowledge Boundaries in Complex Innovation,” *Organization Science*, vol. 23, no. 5, pp. 1467–1484, Oct. 2012, doi: 10.1287/orsc.1110.0700.
- [42] K. M. Eisenhardt, “Building Theories from Case Study Research,” *The Academy of Management Review*, vol. 14, no. 4, p. 532, Oct. 1989, doi: 10.2307/258557.
- [43] U. Flick, *The SAGE Handbook of Qualitative Data Analysis*. 1 Oliver’s Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd, 2014.
- [44] The Economist, “Out of the box,” *The Economist*, 2015.