Crowdsourcing Data Science: A Qualitative Analysis of Organizations' Usage of Kaggle Competitions

Christoph Tauchert
TU Darmstadt
tauchert@is.tu-darmstadt.de

Peter Buxmann
TU Darmstadt
buxmann@is.tu-darmstadt.de

Jannis Lambinus
TU Darmstadt
jannis.lambinus@web.de

Abstract

In light of the ongoing digitization, companies accumulate data, which they want to transform into value. However, data scientists are rare and organizations are struggling to acquire talents. At the same time, individuals who are interested in machine learning are participating in competitions on data science internet platforms. To investigate if companies can tackle their data science challenges by hosting data science competitions on internet platforms, we conducted ten interviews with data scientists. While there are various perceived benefits, such as discussing with participants and learning new, state of the art approaches, these competitions can only cover a fraction of tasks that typically occur during data science projects. We identified 12 factors within three categories that influence an organization's perceived success when hosting a data science competition.

1. Introduction

"Data is just like crude. It is valuable, but if unrefined it cannot really be used. […] So must data be broken down, analyzed for it to have value" [30]. When companies want to refine their valuable data treasures they face various questions such as: How to deal with large amounts of data? How to extract valuable insights from the data? How can the business benefit most from the utilization of data? To create value from data, companies employ data scientists who analyze the data that the company holds.

According to the 2019 Gartner CIO report, companies are struggling with an acute shortage of talents when it comes to their efforts in implementing artificial intelligence [8]. Since data science is heavily related to machine learning and therefore artificial intelligence, this shortage also affects the companies' efforts to turn their data into value.

One theoretical possibility to deal with the scarce resource of data scientists could be to leverage the concept of crowdsourcing. The method to draw on the so-called wisdom of the crowd for problem-solving has been established in various domains for several years. Since data science is a fairly new domain, the use of crowdsourcing has not been adopted largely, yet. One platform that enables companies to seek help from a wide range of data scientists is Kaggle.com. The website's focus is hosting machine learning competitions, organized by the respective companies, for which participants try to build prediction models.

While there generally has been a lot of research done for crowdsourcing, there is, after an extensive investigation, almost no research available addressing the combination of both, crowdsourcing and data science. The overall objective of this study is to provide an overview of crowdsourcing in data science, with a special focus on factors that influence the organization's perceived success of a data science competition. To facilitate the achievement of this objective the study uses expert interviews that are conducted with data scientists from different industries. The interview data is enriched with data that is crawled directly from the data science platform Kaggle.

The research questions this study attempts to answer are as follows:

(1) For what purpose do organizations host data science competitions?
(2) Which factors influence the organizations' perceived success when hosting a data science competition?

The remainder of this manuscript is structured as follows: To begin with, we provide a brief overview of the theoretical background and related research to mark off the research area before the qualitative study design is presented. After introducing our study sample comprising ten interviewees, we derive the results. Finally, we conclude the manuscript by pointing out the limitations of our study and providing specific avenues for future research.

URI: https://hdl.handle.net/10125/63768
978-0-9981331-3-3

2. Theoretical Background

2.1. Data Science and Kaggle Competitions

In recent years, the term data science has become a buzzword that is surrounded by a lot of hype. An article of the Harvard Business Review even designated data scientist as “the Sexiest Job of the 21st Century” [9]. On the other hand, there are voices, who have criticized the closeness of the definitions of the terms data (or business) analytics and data science, but due to new types of data, new methods and new questions a change in the wording is accepted [4, 11]. Van Der Aalst defines data science as follows: ‘‘Data science is an interdisciplinary field aiming to turn data into real value. […]. The value may be provided in the form of predictions, automated decisions, models learned from data, or any type of data visualization delivering insights. Data science includes data extraction, data preparation, data exploration, data transformation, storage and retrieval, computing infrastructures, various types of mining and learning, presentation of explanations and predictions, and the exploitation of results taking into account ethical, social, legal, and business aspects’’[1].

A fundamental concept of data science is to systematically extract useful knowledge from data to solve business problems [33]. A widely accepted codification of this process is the CRISP-DM (CRoss Industry Standard Process for Data Mining) framework. The entire process is described by six phases on a highly aggregated level [6]:

(1) Business Understanding: The purpose of the initial phase is to understand the customer’s needs, determine major factors that have to be considered and formulate business objectives.

(2) Data Understanding: The second step consists of data collection, description, exploration, and verification.

(3) Data Preparation: This phase continues with the handling of data by “cleaning” it to be suitable for later analysis.

(4) Modeling: The fourth phase of CRISP-DM starts with the actual selection of the modeling technique. A subset of the data has to be selected for training, testing, and evaluation of the model. Afterward, one or more models are built with varying parameters whose output can be evaluated. The evaluation is based upon the domain knowledge, the data mining goals chosen in phase one and the test design.

(5) Evaluation: This phase deals with the evaluation of the model with regard to the set business objectives. At this point, it has to be decided whether the model satisfies all requirements.

(6) Deployment: The final phase of the framework addresses the issue of actually deploying the model as well as how to maintain and monitor the outcomes of the project in the long run if used in daily business.

Our study focuses on Kaggle, which is the world’s largest online platform for data science with more than 1,000,000 members. While the platform is a large repository for public datasets and a place to exchange for data scientists through discussion forums and public Jupyter notebooks, its main feature is hosting machine learning competitions for various organizations [29].

The general concept of a Kaggle competition requires participants to develop a prediction model for a precisely defined problem from given data. The submitted models are evaluated in real-time and the respective prediction score is shown in a leaderboard, which creates a competitive environment. However, the final ranking is calculated based on a separate non-public subset of test data. Afterward, the participants that created the highest-ranked submissions receive the prize money, often in return for the intellectual property of the solution [23].

However, when comparing the tasks of data scientists and the concept of Kaggle competitions, it seems that these competitions do not allow to crowdsource all activities related to data science but only a subset. While data science is also about understanding the business, identifying fields of application as well as required and available data, the scope of the competitions only covers tasks closely related to machine learning, like data cleaning and model building.

2.2. Crowdsourcing

The idea behind crowdsourcing is that an organization proposes the voluntary processing of a task that is presented in an open call to an undefined group of individuals or teams [13]. A strength of crowdsourcing lies in the open call to the broader public which can serve as a means to obtain new ideas and approaches from people outside the usual domain and boundaries [2]. Crowdsourcing can be collaborative or competitive. The former encourages participants to collectively work towards a common solution while the last one aims at the collection of various independent solutions out of which the crowdsourcer can select the winning solutions [2]. Competitive crowdsourcing initiatives often result in a financial or non-financial compensation of winning participants [40]. The Kaggle competitions described above fall into the category of competitive (or tournament-based) crowdsourcing.

The broad adoption of crowdsourcing led to a large number of scientific papers examining this topic with various different foci. However, since a crowdsourcing task's success and thereby likewise the overall success of the hosting platform itself, is significantly dependent on the number of individuals participating at a given task, research has focused on the user's perspective of crowdsourcing.

Studies addressing the users' motivation to participate in crowdsourcing usually consider two distinct kinds of motivation, i.e., extrinsic and intrinsic motivation, drawing on the self-determination theory [10]. Extrinsic motivation refers to performing an action to attain an external result. In other words, the incentive is coming from an outside source. Intrinsic motivation, in contrast, is independent of some outcome but arises from the pure fun and joy of doing something [10]. The incentive is to be satisfied and the task itself is central instead of a promised reward [34]. Factors of motivation that were identified include: task autonomy and skill variety as factors of fun and delight [23, 31, 41], financial compensation [23, 25, 31, 41], social motivation (i.e., reputation) [18, 23, 31], tacitness [41], learning [18, 25], self-marketing [25], meaningfulness / impact of the task [5], complexity [12, 37, 41], event duration [37], number of events [37].

From an organization's perspective, crowdsourcing is designed to get others to solve problems by using knowledge that the organization may not normally have access to [22]. And therefore, the main reason for organizations to initiate a crowdsourcing campaign is to get the result of a given task or the resolution of a problem [13]. Often crowdsourcing is associated with innovation processes such as new product development or product improvements [32]. In this case, companies get creative ideas, that might be commercially exploitable [24]. This approach is supported by studies that show that many of those user innovations are characterized by high commercial attractiveness [19]. Besides concrete innovations, companies also try to create any type of added value by crowdsourcing through value creation or increased profits [39]. Another goal that organizations might pursue through crowdsourcing is to obtain knowledge and especially talent from the crowd by using crowdsourcing campaigns as an employee recruitment tool [20].

We found one study that used Kaggle as a context [17]. It assessed how participants' engagement is related to their solutions' creativity. The results show that higher cognitive and emotional engagement is associated with more creative output. Further emphasis is put on the willingness to share obtained knowledge. The data shows that the need for versatile problem-solving skills makes a competition intrinsically inspiring, which in turn strengthens the desire to share a promising solution with others.

To summarize, so far a lot of research on crowdsourcing has focused on the motivation of users to participate in and companies to host crowdsourcing events. The present study aims to give insights into the organizations' perspective on the success of data science competitions. Therefore, it provides a basis to fill the research gap that currently exists in this area.

4. Method

The goal of our study was to expand the current stage of IS research concerning the crowdsourcing of data science projects. Since the amount of companies that are conducting data science challenges on platforms such as Kaggle is low and the field has not been extensively explored, an explorative approach using interviews with experts seems appropriate to investigate the problems occurring in this particular context [15]. According to Weber [38], content analysis is an appropriate approach to assess open-ended questions and therefore, it is suitable for the evaluation of the collected qualitative data. The interviews were transcribed, coded and analyzed taking into account relevant publicly available data through triangulation [21]. Therefore, we collected data from Kaggle using a self-written crawler and conducted explorative data analysis. The data was crawled mid-December 2018. We decided to use Kaggle as a context since it is by far the largest (most registered and active users) and most open (commercial and non-commercial) platform for data science competitions. The two alternative platforms, Codalab and RAMP, are intended for research problems only.

4.1. Research Design

Our main information source was in-depth expert interviews, which were conducted in a semi-structured way. Following the guiding principles of Sarker et al. [36], we prepared an interview protocol and acquired key informants in different companies using professional social networks (i.e., LinkedIn, XING). During the interviews, we kept our questions open in order to enable participants to speak freely.

The interview guide comprised five different sections: The first part comprised general questions about the interview partner and the company he/she works for and introduced the context of the interview. The second section tackled the topic of how data science is used in the company in general. In the third section, we focused on data science competitions and asked the interview partner about their experiences and opinions about data science challenges on internet platforms. In the fourth section, we wanted to know how data science platforms, in general, are perceived by the experts. In the last section, the informants had the chance to comment openly on the topic and add remarks.

Due to the semi-structured approach, questions were gradually adjusted in order to account for the interview partners' individual situation.

4.2. Sample and Data Collection

Table 1 provides an overview of the participants. The interviews were conducted over a three-month period and took place between November 2018 and January 2019. In total ten interviews with highly involved participants were conducted of whom all were data scientists. After the tenth interview, data collection was discontinued since no new, previously unmentioned aspects were mentioned [15].

Table 1. Sample Description

ID	Industry	Total Employees	Data Scientists	Revenue [bn. €]
A	Telco	10,000 - 100,000	10 – 50	> 20
B	Research	< 10,000	-	-
C	Government	< 10,000	< 10	-
D	Financial Services	> 100,000	10 – 50	> 20
E	Chemical	10,000 - 100,000	> 50	5 - 20
F	Research	-	-	-
G	Software	10,000 - 100,000	> 50	> 20
H	Price Comparison	< 10,000	< 10	< 5
J	Automotive	> 100,000	> 50	> 20
K	Financial Services	-	< 10	-

The average duration of the interviews was approx. 30 minutes and the interviews were mostly held via telephone due to geographical distance.

We used a conventional approach to content analysis, which aims to describe a phenomenon to allow new insights to emerge [21]. This is also described as inductive category development [28]. Subsequently, the transcripts were assessed by using the MAXQDA software and by conducting two coding cycles as recommended by Saldaña [35]. The first coding cycle comprised a mixture of attribute coding and descriptive coding. The former was performed to obtain essential insights about the data and its descriptive information. The latter was used to extract additional aspects, key thoughts, and concepts from the interview data. In a second cycle, the formerly created codes were combined into a smaller number of sets using pattern coding [35]. By discussing and assessing the coding process with a group of three IS researchers and students, an investigator triangulation helped to ensure rigor and trustworthiness. Furthermore, the crawled data from the Kaggle platform was used for data triangulation [15].

5. Results and Discussion

5.1. Information about Competitions

Competitions constitute the most important aspect of Kaggle since it is the service it started with and it is still mainly what they are known for.

Until the date of data collection, 309 competitions have been hosted at the platform, including the still-active ones. 50.5 % of competitions were hosted by companies or organizations, which provided an explicitly defined problem to be solved and offered a reward (mostly price money). 24.6 % (76) competitions were categorized as *research*. The entities behind those competitions are usually non-commercial institutions with some scientific background. They are thus often not able to provide as much prize money as commercial companies. To facilitate research competitions Kaggle offers to sponsor them by providing $25,000 as prize money. 16 competitions (5.2 %) were in the category "recruitment". In general, these competitions do not differ from the aforementioned competitions except that they offer job interviews for the highest-ranking participants. The other competitions belonged to the categories *playground*, *getting started*, *masters* and *analytics*.

Companies planning to host a competition have to compete with other active competitions for the attention of users. One factor that can be directly influenced by the firms and that might increase Kagglers' motivation to participate is the prize money rewarded to the highest-ranking participants. Table 2 shows the statistics for rewards and participants for competitions that offered any prize money (> $0).

Table 2. Rewards' and Participants' Structure of Competitions

Statistic	Reward		Teams	
	Featured	Research	Featured	Research
Count	154	61	154	61
Mean	49,219	7,051	1,128	320
Max.	1,200,000	25,000	7,198	1,386

The numbers represent 215 competitions in total, thereof 154 featured and 61 research competitions. It can be seen that the mean prize money for category featured ($49,219) is seven times the amount for research competitions ($7,051). Regarding the number of participating teams per competition, we see that featured competitions (1,128) have about 3.5 times as many teams as those with a research label (320). One reason for this might be the in average significantly lower prize money. Another reason might be that the Kaggle community is more interested in industry competitions than in research competitions.

The USA has hosted the majority of competitions, accounting for 65 % of all competitions with 138 hosted competitions. About 34 competitions are coming from companies and research institutions in Europe, mainly from the United Kingdom (9), France (8), Spain (7) and Germany (6). Asian countries with participating companies are mainly Russia (5), Japan (5), Israel (3), Taiwan (2) and China (2).

5.2. Results

While coding the transcribed interviews, we noticed that codes could be categorized in: (1) platform-related, (2) organization-related and (3) outcome-related factors.

Platform-related factors

Community. The capabilities of the data scientists on Kaggle are considered to be very high. Hence, several experts (A, C, J, and K) see a good chance to obtain high-quality models from a competition. With an average of 1,128 teams that are participating in such competitions the potential for great ideas and solutions is relatively high. Experts C and J say that Kaggle competitions attract some of the best data scientists in the world, such as for example Tianqi Chen, the lead developer of the popular XGBoost framework, participated at eight competitions. In addition, experts F and K perceive their respective competitions as successful, even though their competitions are not finished at the time of the interviews. They are both largely satisfied with the number of participating teams, as it means potentially a lot of new ideas (for expert F) as well as many people being aware of the company, who wants to increase brand popularity (for expert K). With having two to three times more people than expected and reaching the targeted number of 1,000 teams within the first week, respectively, it is apparently relatively easy to attract a lot of people and motivate them to participate. These statements correspond with the data retrieved from Kaggle showing an average of about 1,100 participating teams per competition. A possible reason for such a high number might be that the number of new competitions is not steadily growing, as one might suspect, but is instead staying at a relatively constant level of about three new competitions per month. Expert E mentions that an ambitious participation at a competition is accompanied by an expenditure of time close to full time. Therefore, it can be assumed that Kagglers, in general, do not participate in multiple competitions simultaneously. More simultaneously active competitions would thus reduce the average number of participants per competition, which would be counterproductive as companies try to attract as many Kagglers as possible. A study of Shao et al. from 2012 supports this presumption. The study’s findings suggest that a higher competition intensity in a crowdsourcing context is associated with a significant decrease in participating users [28].

Infrastructure. By providing data storage capacities for data sets and computing power for machine learning models, Kaggle is removing barriers that would otherwise hamper companies to organize such data science competitions. Companies struggle enough with the collection and preparing of data and therefore are happy that they do not have to worry about technical infrastructure.

Regulations. While Kaggle is trying to have a low technical barrier, they do have other barriers in place. The minimal amount of prize money for *featured* competitions is $25,000. Depending on the company size that might be a lot of money to spend on an unknown outcome. Especially small and medium-sized companies, who could really benefit from this approach, could be scared off for this reason. Another restriction Kaggle imposes on the hosting organization is, that only supervised machine learning problems are allowed. Companies whose field of activity is in an area where unsupervised or reinforcement learning approaches are necessary cannot host a competition. Expert F and her team started the first competition with the intention to have it as the first of a whole series of competitions. Since her team is especially interested in unsupervised learning problems such as anomaly detection, they are reconsidering whether they complete the series of competitions.

Organization-related factors

Marketing. A further reason for experts H and K to host competitions is to do public relations or brand building. The experts say that the proper utilization of a company's data is getting more and more crucial to stay competitive in the business. The market for data scientists, however, is very small. It is therefore important that data scientists, as potential employees, know the company and recognize the brand. By hosting a machine learning competition, the firms try to increase their attractiveness towards the data science community. Another means of marketing are hackathons. Those originally from software development coming short-term events have been named by several experts (A, B, D, and H) when they have been asked if they have hosted a machine learning competition so far. The association between a machine learning competition on Kaggle and a hackathon can be seen as reasonable, as hackathons usually do not create fully-fledged solutions but rather partially usable prototypes and furthermore are intended to increase the brand awareness among possible employees or customers [16]. This coincides with what the experts think about the results of Kaggle competitions, as mentioned above, and also with the aim to engage in brand building. One obvious difference, however, is that hackathons are local in general, while a Kaggle competition reaches out to a worldwide distributed audience. It can, therefore, be assumed that Kaggle is a new means of marketing to reach out to the data science community and complements the established practice of hackathons.

Recruiting. One of the incentive types on Kaggle is the prospect of a job interview at the hosting company. As there is a high demand for data scientists, it seems to make sense to draw on a data science community as large as Kaggle to get in touch with potential employees. However, the data analysis shows that the competition category *recruitment* has only been chosen 16 times, with the last appearance in the first quarter of 2017. These findings suggest that companies do not like this option, maybe because it did not prove to be successful. The three interviewed hosting companies (B, F, and K) are in line with this development and do not focus on recruitment through Kaggle. Expert F, who is working for a research institution, says that for an academically career other skills are higher valued than those skills that can be shown at a competition. In addition, expert K, who is working for a commercial company, states that recruitment would be a nice side effect but not of special interest. The experts E and G have used the Kaggle job board successfully in the past, which is not directly linked to the competitions but presents regular job advertisements. Expert K additionally mentions that an advantage of a featured or research competition is the participation of a worldwide-distributed audience. Although a recruitment competition is in general free for everyone to join, she might be right because a certain proportion of potential users might not be motivated to participate, assuming a job offer is unappealing for participants not looking for a job. A lower participation rate, however, would have a negative impact on the important objective of obtaining new ideas and innovative approaches from submitted solutions. This statement is in line with expert J, saying that about 70 % of a data scientist's actual work is not required on Kaggle. In the remaining 30 %, however, participants can excel and obtain excellent knowledge, according to him. Expert K emphasizes that the participants' aim on Kaggle is always to get a high final score, i.e. to maximize the accuracy of the model, whereas in a real-world problem other aspects such as the speed of a prediction or interpretability might play a major role.

Data. Seven out of the ten experts that have been interviewed are working for companies that have not been hosting a competition on Kaggle yet but are considering it (experts A, C, D, E, G, H, and J). When asked for reasons that might justify this, often their first answer was the apparent need to publish sensitive data. For most companies, a problem that theoretically would be suitable to be solved through crowdsourcing, would contain some type of sensitive data, be it internal data about the company and its projects or customer data, which would potentially allow identification of those customers. Although there are possibilities to anonymize data (e.g. k-Anonymity [3] and L-diversity [27]) the companies apparently shy away from putting the sometimes considerable amount of effort into it. As those methods also cannot fully guarantee that any identification can be ruled out [26], they might not want to take the risk of having a public data scandal. Experts E and G mention that their companies' conservative attitude towards sensitive data-related projects in public is typical for German companies. Research has shown that there are differences in the innovation and risk culture between for instance the United States of America and Europe, with European cultures being more reserved [14]. The experts' opinion corresponds with the findings of the data analysis regarding hosted competitions, as about 65 % of all competitions are hosted by US-based companies or institutions, even when competitions hosted by Kaggle and Google itself are excluded.

Top Management Support. Only expert E states that the decision-makers in his company presumably do not know about the possibilities of crowdsourcing for data science projects. However, according to him, this would be the most significant factor why no competition has been hosted so far. Therefore, it seems that the awareness of data science platforms, in particular Kaggle, is fairly high among decision-makers working in the realm of data science.

Use Case. Additionally, expert E as well as expert A say that they did not have any problem that they wanted to get solved by the crowd. At this point, it remains unclear whether they have all the necessary resources to solve the problems internally to a satisfying extent or whether they do not have problems suitable for a Kaggle competition, which are only supervised learning problems so far. However, it seems to be unlikely for companies of their size (both 10,000 to 100,000 employees and revenue of at least $5 bn. per year) to not have any business problem linked to supervised learning.

Lack of Resources. However, for expert B the further usage of Kaggle is less dependent on the features provided by the platform but more on how the competitions are organized within his institution. He, as well as expert K, states that they did all the work of hosting the competitions in parallel with their regular full-time job. The expert, therefore, would prefer to have a dedicated team working on the task of organizing, conducting and evaluating the whole competition to increase efficiency, which so far is not the case.

Outcome-related factors

Innovation. Independent whether their company has been active on Kaggle or not, all experts do name the innovative power behind the competitions as a decisive reason. The two words "new ideas" spring up regularly during the interviews, although no question specifically asks for it. The capabilities of the data scientists on Kaggle are considered to be very high. Hence, several experts (A, C, J, and K) see a good chance to obtain high-quality models from a competition. As stated before, Kaggle competitions attract some of the best data scientists in the world.

Incompleteness. Interestingly, none of the three experts working for a hosting company expects to receive a fully completed machine learning model. Although expert K hopes for a high-quality model, she does not take it for granted and expresses herself cautious about the upcoming results. The two other experts (B and F) do not even expect a solution, which is able to solve the respective problem. Instead, their plan is to closely examine submissions for different approaches on how to tackle their problems. They hope to see approaches that their team did not think of but that show promising results. This way of thinking is presumably found rarely on other established crowdsourcing platforms, e.g. Amazon Mechanical Turk or 99designs, where actually usable and finished results are expected in general. However, the differences in the complexity between the tasks on those platforms compared to tasks on Kaggle are considerably high, making a direct comparison difficult. The concept of using the community for solution finding is closely related to "open innovation", where companies integrate external sources into the usually internal innovation process. The external sources get reached via an open call to a large, unknown crowd [7]. This is very similar to the definition of crowdsourcing. Open innovation is not intended to replace but to complement the traditional innovation process [7] which is in line with the statement of expert B, saying that crowdsourcing in data science is not used to replace the internal process but used as an additional channel. Kaggle, therefore, seems not so much to be about actually solving a problem directly but to support the organizing company at ultimately achieving a complete solution.

Learning. Expert F sees high value in monitoring the progress of participants through closely following the discussions on the competition forum and in answering those questions. As most user presumably do not have the same domain/business background as the organizing team, they approach the problem unbiased, which includes interesting information for the team of expert F. The data analysis verifies that there are indeed a lot of discussions during a competition with an average of 101 threads per competition. Considering that the average competition lasts for 78 days, this means more than one new thread per competition and day. The expert's statement shows that the crowdsourcing process on Kaggle is not just done by providing a relevant problem with subsequent waiting for a fitting solution, but that it is more a constant, interactive and collaborative process with learnings on both sides. The assessment of the overall success of a competition is therefore not solely dependent on the best final solutions but also on the process to reach them.

Figure 1 depicts all identified factors and their influence on the organization's perceived success of data science competitions.

6. Conclusion

The combination of crowdsourcing and data science is a relatively new concept, which has not been exhaustively researched. Therefore, this study creates a basis for further studies in this context. We enriched the qualitative interview data with data that we crawled directly from the Kaggle platform. This approach allows for a broad overview of different interesting aspects and data triangulation.

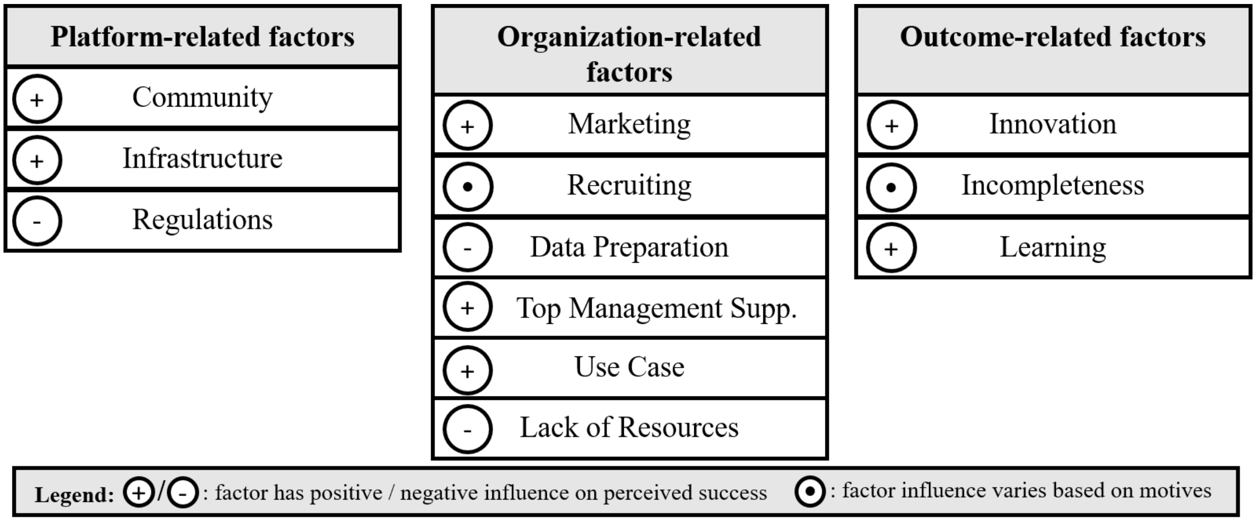

Figure 1. Factors Influencing the Success of Data Science Competitions

The data shows that so far 32 companies and research institutions have hosted at least two competitions, some of them up to four. It, therefore, seems that for some companies crowdsourcing in data science might indeed work and deliver good solutions.

The interviews show that companies highly value the innovative power of the community of data scientists on Kaggle but see problems in dealing with sensitive data in a public context. Brand building and partially recruiting are seen as positive aspects.

Crowdsourcing proved to be a valuable concept for companies to leverage the wisdom of a heterogeneous crowd. Data science is currently rapidly expanding and still in a relatively early stage. The combination of both fields promises a lot of potential. As more companies and interested people get in touch with data science, platforms like Kaggle might emerge creating a competitive market. A lot of research needs to be done to obtain further insights into this new market comprising the combination of crowdsourcing and data science.

The present study examines the relatively new combination of crowdsourcing with data science. So far there has been almost no research conducted in this specific context. This explorative study aims to serve as a basis for further studies in the context of crowdsourcing in data science. The main reason for companies for hosting a machine learning competition is the innovative power inherent in the wisdom of the crowd. It is important to obtain insights, whether the solutions, especially the winning ones, actually deliver the desired innovation. Therefore, further research should, among other things, focus on companies that have hosted competitions in the past, which also means that another cultural context will have to be considered. The other part of companies, i.e. those who do not host competitions, see the biggest problem in the publishing of sensitive data. It is important to know how rational this justification actually is, and how well common anonymization techniques can be utilized to make datasets suitable for those competitions. Thereby, companies could better assess the risk related to hosting. As marketing reasons are also named by the experts, research should get insights about the actual perception of companies in the community. Ultimately, it needs to be examined whether machine learning competitions are indeed an appropriate marketing tool to increase brand awareness in the data science community. Furthermore, it is essential for companies to know how to design a competition, e.g. in terms of prize money, duration as well as topic and problem description, respectively. Therefore, a closer comparison between more and less successful competitions is needed.

The results of this study indicate that crowdsourcing and data science can be combined in a successful manner. However, companies, which plan to host a machine learning competition, should bear in mind that the circumstances are appropriate. Firstly, in most cases, Kaggle is presumably not a way to get a given problem solved by others for cheap money in a short time. Rather, the crowd should be seen as a means to enrich the internal data science process. Permanent communication and collaboration between participants and the host are most likely to be the best way to achieve promising results. To ensure such a process, companies should provide a dedicated team of internal employees to organize and supervise the competition instead of doing it next to daily work. If companies have a well-designed backend system for their data, which allows for easy preparation of datasets, Kaggle is more likely to serve as a good platform to use the wisdom of the crowd for problem-solving. Otherwise, composing a well-suited dataset can be a difficult and time-consuming task.

As every study, also the present study and its results are to be seen and interpreted in consideration of certain limitations. Since this study is based on a relatively small sample of only ten interviews, we cannot draw confident conclusions. Furthermore, this study aims to provide broad oversight of the subject matter. Therefore, the different aspects are examined at a very high level and are only scratched on the surface. The experts' answers in the interviews are naturally at least partially subjective and should not be seen as a matter of fact. Lastly, with only three experts working for hosting companies, the generalizability of their answers needs to be evaluated carefully.

7. References

[1] Van Der Aalst, W.M.P., *Process Mining : Data Science in Action.*, Springer, Heidelberg, 2018.

[2] Afuah, A., and C.L. Tucci, "Crowdsourcing As a Solution to Distant Search", *Academy of Management Review 37*(3), 2012, pp. 355–375.

[3] Bayardo, R.J., and R. Agrawal, "Data Privacy Through Optimal k-Anonymization", *Proceedings of the 21st International Conference on Data Engineering (ICDE 2005)*, IEEE (2005).

[4] Bichler, M., A. Heinzl, • Wil, and M.P. Van Der Aalst, "Business Analytics and Data Science: Once Again?", *Business & Information Systems Engineering 59*(2), 2017, pp. 77–79.

[5] Chandler, D., and A. Kapelner, "Breaking monotony with meaning: Motivation in crowdsourcing markets", *Journal of Economic Behavior & Organization 90*, 2013, pp. 123–133.

[6] Chapman, P., J. Clinton, R. Kerber, et al., "CRISP-DM 1.0 Step-by-step data mining guide", 2000.

[7] Chesbrough, H., W. Vanhaverbeke, and J. West, eds., *Open Innovation: Researching a New Paradigm - Google Books*, Oxford University Press on Demand, 2006.

[8] Costello, K., "Gartner Survey Shows 37 Percent of Organizations Have Implemented AI in Some Form", 2019. https://www.gartner.com/en/newsroom/press-releases/2019-01-21-gartner-survey-shows-37-percent-of-organizations-have

[9] Davenport, T.H., and D.J. Patil, "Data Scientist: The Sexiest Job of the 21st Century", *Harvard Business Review 90*(5), 2012, pp. 70–76.

[10] Deci, E.L., and R.M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior*, Springer US, Boston, MA, 1985.

[11] Dhar, V., "Data science and prediction", *Communications of the ACM 56*(12), 2013, pp. 64–73.

[12] Eickhoff, C., "Crowd-powered experts: helping surgeons interpret breast cancer images", *Proceedings of the First International Workshop on Gamification for Information Retrieval - GamifIR '14*, ACM Press (2014), 53–56.

[13] Estellés-Arolas, E., and F. González-Ladrón-de-Guevara, "Towards an integrated crowdsourcing definition", *Journal of Information Science 38*(2), 2012, pp. 189–200.

[14] Ezell, S., and P. Marxgut, *Comparing American and European Innovation Cultures*, 2015.

[15] Flick, U., "No Triangulation in Qualitative Research", In U. Flick, E. von Kardorff and I. Steinke, eds., *A Companion to Qualitative Research*. Sage, London, 2004, 178–183.

[16] Frey, F.J., and M. Luks, "The innovation-driven hackathon", *Proceedings of the 21st European Conference on Pattern Languages of Programs - EuroPlop '16*, ACM Press (2016), 1–11.

[17] Garcia Martinez, M., "Solver engagement in knowledge sharing in crowdsourcing communities: Exploring the link to creativity", *Research Policy 44*(8), 2015, pp. 1419–1430.

[18] Hertel, G., S. Niedner, and S. Herrmann, "Motivation of software developers in Open Source projects: an Internet-based survey of contributors to the Linux kernel", *Research Policy 32*(7), 2003, pp. 1159–1177.

[19] von Hippel, E., *Democratizing innovation: Users take center stage*, MIT Press, Cambridge, MA, 2005.

[20] Howe, J., "The rise of crowdsourcing", *Wired magazine 14*(6), 2006, pp. 1–4.

[21] Hsieh, H.-F., and S.E. Shannon, "Three Approaches to Qualitative Content Analysis", *Qualitative Health Research 15*(9), 2005, pp. 1277–1288.

[22] Jeppesen, L.B., and K.R. Lakhani, "Marginality and problem-solving effectiveness in broadcast search", *Organization Science 21*(5), 2010, pp. 1016–1033.

[23] Kaufmann, N., T. Schulze, and D. Veit, "More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk", *Proceedings of the Seventeenth Americas Conference on Information Systems*, (2011).

[24] Kleemann, F., G.G. Voß, and K. Rieder, "Un (der) paid innovators: The commercial utilization of consumer work through crowdsourcing", *Science, technology & innovation studies 4*(1), 2008, pp. 5–26.

[25] Leimeister, J.M., M. Huber, U. Bretschneider, and H. Krcmar, “Leveraging Crowdsourcing: Activation-Supporting Components for IT-Based Ideas Competition”, *Journal of Management Information Systems 26*(1), 2009, pp. 197–224.

[26] Li, N., T. Li, and S. Venkatasubramanian, “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity”, *2007 IEEE 23rd International Conference on Data Engineering*, IEEE (2007), 106–115.

[27] Machanavajjhala, A., J. Gehrke, D. Kifer, and M. Venkitasubramaniam, “L-diversity: privacy beyond k-anonymity”, *22nd International Conference on Data Engineering (ICDE’06)*, IEEE (2006), 24–24.

[28] Mayring, P., “Qualitative content analysis”, *A companion to qualitative research 1*, 2004, pp. 159–176.

[29] Metz, R., “A startup called Kaggle tries to bring smart people to knotty problems.”, *MIT Technology Review 116*(5), 2013, pp. 51.

[30] Palmer, M., “Data is the New Oil”, 2006. https://ana.blogs.com/maestros/2006/11/data_is_the_new.html

[31] Pilz, D., and H. Gewald, “Does Money Matter? Motivational Factors for Participation in Paid- and Non-Profit-Crowdsourcing Communities”, *Wirtschaftsinformatik Proceedings 2013*, 2013.

[32] Poetz, M.K., and M. Schreier, “The value of crowdsourcing: Can users really compete with professionals in generating new product ideas?”, *Journal of Product Innovation Management 29*(2), 2012, pp. 245–256.

[33] Provost, F., and T. Fawcett, *Data Science for Business: What you need to know about data mining and data-analytic thinking*, “ O’Reilly Media, Inc.”, 2013.

[34] Ryan, R.M., and E.L. Deci, “Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions”, *Contemporary Educational Psychology 25*(1), 2000, pp. 54–67.

[35] Saldaña, J., *The coding manual for qualitative researchers*, Sage, 2015.

[36] Sarker, S., X. Xiao, and T. Beaulieu, “Guest Editorial: Qualitative Studies in Information Systems: A Critical Review and Some Guiding Principles”, *MIS Quarterly 37*(4), 2013.

[37] Shao, B., L. Shi, B. Xu, and L. Liu, “Factors affecting participation of solvers in crowdsourcing: an empirical study from China”, *Electronic Markets 22*(2), 2012, pp. 73–82.

[38] Weber, R.P., *Basic Content Analysis*, Sage, Newbury Park, CA, 1990.

[39] Yang, J., L.A. Adamic, and M.S. Ackerman, “Crowdsourcing and knowledge sharing: strategic user behavior on taskcn”, *Proceedings of the 9th ACM conference on Electronic commerce*, ACM (2008), 246–255.

[40] Zhao, Y., and Q. Zhu, “Evaluation on crowdsourcing research: Current status and future direction”, *Information Systems Frontiers 16*(3), 2014, pp. 417–434.

[41] Zheng, H., D. Li, and W. Hou, “Task Design, Motivation, and Participation in Crowdsourcing Contests”, *International Journal of Electronic Commerce 15*(4), 2011, pp. 57–88.