# Data Analysis Competition Platform for Educational Purposes: Lessons Learned and Future Challenges

**Yukino Baba**
Kyoto University; RIKEN Center for AIP
baba@i.kyoto-u.ac.jp

**Tomoumi Takase, Kyohei Atarashi**
Hokkaido University
{takase_t, atarashi_k}@complex.ist.hokudai.ac.jp

**Satoshi Oyama**
Hokkaido University; RIKEN Center for AIP
oyama@ist.hokudai.ac.jp

**Hisashi Kashima**
Kyoto University; RIKEN Center for AIP
kashima@i.kyoto-u.ac.jp

**Abstract**

Data analysis education plays an important role in accelerating the efficient use of data analysis technologies in various domains. Not only the knowledge of statistics and machine learning, but also practical skills of deploying machine learning and data analysis techniques, are required for conducting data analysis projects in the real world. Data analysis competitions, such as Kaggle, have been considered as an efficient system for learning such skills by addressing real data analysis problems. However, current data analysis competitions are not designed for educational purposes and it is not well studied how data analysis competition platforms should be designed for enhancing educational effectiveness. To answer this research question, we built, and subsequently operated an educational data analysis competition platform called *University of Big Data* for several years. In this paper, we present our approaches for supporting and motivating learners and the results of our case studies. We found that providing a tutorial article is beneficial for encouraging active participation of learners, and a leaderboard system allowing an unlimited number of submissions can motivate the efforts of learners. We further discuss future directions of educational data analysis competitions.

## 1 Introduction

Data analysis education plays an important role in accelerating the efficient use of data analysis technologies in various domains. Not only knowledge of statistics and machine learning, but also practical skills of deploying machine learning and data analysis techniques, are required for conducting data analysis projects in the real world. In addition to classroom lectures, hands-on training with a variety of real datasets is necessary for learning such skills.

Data analysis competitions provide opportunities for data scientists to try and promote their practical data analysis skills. In each data analysis competition, participants apply machine learning algorithms to the given dataset, and compete to build the best predictive models. In conjunction with data mining and machine learning conferences, data analysis competitions such as the KDD Cup (Kohavi et al. 2000) have been held recently. Netflix, Inc. organized the Netflix Prize from 2006 to 2010 to develop the best movie recommendation algorithm by utilizing the power of the crowds of data scientists (Töscher, Jahrer, and Bell 2009).[1] Data analysis competition platforms, such as Kaggle[2] or CrowdANALYTIX,[3] have been launched as well. These platforms generalize the Netflix approach and enable companies and organizations to conduct open competitions for seeking the best predictive model for their own data and problems by leveraging data scientists from all over the world (Baba et al. 2014).

Data analysis competitions have been considered an efficient platform for learning such skills by addressing real data analysis problems. For example, Kaggle offers a tool for running data science competitions in class, called Kaggle in Class.[4] This tool enables lectures to use the features of Kaggle, such as its evaluation system, leaderboard, or forum. Kaggle in Class has been used for data analysis education in hundreds of universities around the world. This fact indicates that data analysis competitions have been recognized as an efficient tool for training practical data analysis skills; a competition system wherein a real dataset is provided for participants and their outcomes are evaluated right after submission would be beneficial for encouraging self-study, and competing against others may improve the motivation for learning.

Although the educational efficiency of data analysis competitions has been noticed, current data analysis competitions are not designed for the educational purpose, and it is not well studied how data analysis competition platforms should be designed for enhancing educational effectiveness. The suitable design of a data analysis competition for learn-



[1] http://www.netflixprize.com/
[2] https://www.kaggle.com
[3] https://www.crowdanalytix.com
[4] https://inclass.kaggle.com

ers would be different to that for professionals. For example, the professional competitions typically offer prize money as an incentive for participation; however, such prize money may not be applicable or appropriate for educational purposes and thus we need to seek other mechanisms for motivating participants. In addition, as we expect the participation of more beginners in educational competitions than in professional competitions, approaches for supporting such beginners are required. Because the goal of professional competitions is to let professionals compete to build the best predictive model, they are designed in an anti-collaborative way; collaboration mechanisms can be introduced into educational competitions so that learners collaboratively develop their skills.

To investigate an efficient platform design of educational data analysis competitions, we created our own platform, called *University of Big Data*,[5] on which we have operated several competitions. We deployed several features for supporting beginners and motivating participants. In this paper, we report our approaches and the results of our case studies. We found that providing a tutorial article is beneficial for encouraging active participation of learners, and a leaderboard system allowing an unlimited number of submissions can motivate the performance of learners. We further discuss the future directions of educational data analysis competitions.

## 2 Overview of Our Platform

University of Big Data is built as a Web application and publicly available for everyone. The platform was launched in 2014, and approximately 700 users are registered as participants as of September 2017. University of Big Data has been used in lectures in universities and companies. Competitions on University of Big Data are held using actual datasets, mainly provided from industries. We offer practical data analysis challenges, such as product recommendation for an online market, or character recognition for business cards.

Users participate in a competition as follows: they are first asked to create an account for the platform and are given an access to the datasets for a target competition. Then, by using the datasets, they train a predictive model which outputs a prediction for a given set of test samples. Right after submitting the prediction to the platform, a public score of the submission is displayed on the leaderboard. The public score is calculated by using a subset of test samples. Users are allowed to submit their predictions multiple times until the competition finishes; however, in most of the competitions, the maximum number of submissions per day is limited to three. When the competition finishes, the private scores are revealed. The private scores are calculated by using the remaining portion of the test samples. Final ranking is determined based on the private scores. In contrast to common data analysis competitions, we do not offer prize money to the winners.

[5]http://universityofbigdata.net

## 3 Approaches for Supporting and Motivating Learners

University of Big Data accommodates several features for enhancing learning effectiveness. We introduce our approaches for supporting and motivating learners.

### 3.1 Tutorial

The target of University of Big Data includes beginners in data analysis. We observed that there was a bottleneck for such beginners to participate in competitions because they were likely to have trouble in the first trial for building a predictive model. For supporting their first step, each competition in University of Big Data provides a tutorial article. The tutorial articles provide step-by-step introductions for preparing the first submission. These articles explain how to import given datasets, perform preprocessing, prepare and train a predictive model, and output a prediction for test samples. Example code is provided in the tutorials as well, and a leaner easily prepares a submission just by following the tutorial.

### 3.2 Leaderboard

The public scores of all players are displayed on a leaderboard during the competition period. The leaderboard enables players to see their current ranking to encourage rivalry between players, which may motivate their active involvement. We also provide a time-series leaderboard, as shown in Figure 1. This leaderboard presents the changes of scores over time so that players can check their progress. In common data analysis competitions, there is a limitation on the number of submissions per day for avoiding unfair information leakage from a leaderboard, such as the ground truth labels of a subset of test samples. This limitation discourages the motivated learners from incorporating a large amount of trial and error. University of Big Data provides a leaderboard system using the Ladder algorithm (Blum and Hardt 2015). While an ordinary leaderboard updates players' scores every time they make new submissions, a Ladder leaderboard updates scores when the score of a new submission is statistically significantly better than that of the player's previous submission. Thus, a Ladder leaderboard is robust against attacks for information leakage and we can remove the limitation of the number of submissions.

### 3.3 Rating

Tutorial articles are provided for supporting beginners in preparing their first submissions in a competition, and Ladder leaderboards encourage players to make many submissions in a competition. To motivate users to participate in multiple competitions, University of Big Data implements a player rating system, which is based on the algorithm applied for the TopCoder Marathon Match.[6] Concisely, after each competition completes, the algorithm improves the ratings of players when they outperform the other players who had higher ratings at the beginning of the competition. The

[6]https://community.topcoder.com/longcontest/?module=Static&d1=support&d2=ratings

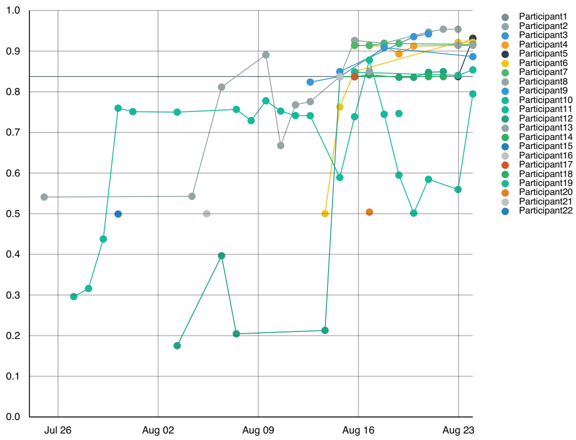



Figure 1: Time-series visualization of the leaderboard, which is for presenting the progress of each participant.

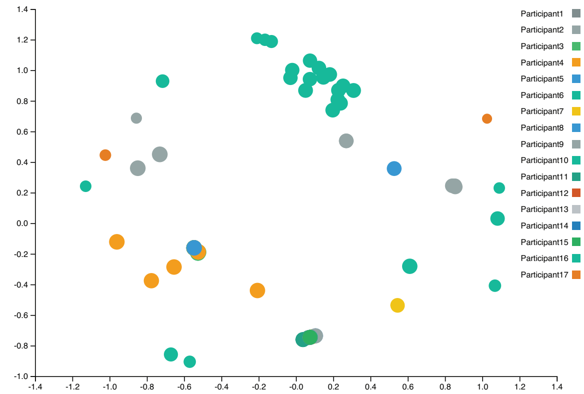



Figure 2: Similarity visualization of submissions, for supporting offline discussions among participants.

ratings are publicly available on the platform so that highly rated players can promote their skills.

### 3.4 Winner report

Similar to the rating system, University of Big Data highlights highly skilled players via the winner reports. In the reports, the winners of a competition present their approaches for predictive modeling, preprocessing datasets, and feature engineering. These reports provide good feedback for the participants of the same competition by comparing their approaches to the winners' approaches. Additionally, the reports can be learning materials for users who participate in future competitions.

### 3.5 Similarity visualization

For motivating offline discussion between the participants of a competition, we provide a similarity map between submissions, as shown in Figure 2. This visualization is created by using Multidimensional scaling (MDS), representing each submission as an $n$-dimensional vector, and the size of each circle indicates the public score (larger is better). We expect that this similarity map would be a trigger for discussion and knowledge sharing between participants.

## 4 Case Studies

We conducted two case studies to investigate the efficiency of our approaches for motivating leaners: tutorial and leaderboard.

### 4.1 Tutorial

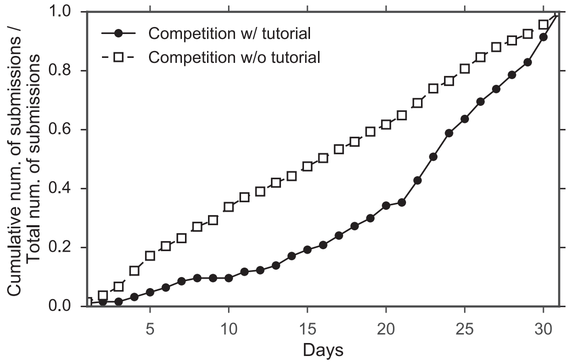
We conducted two competitions to investigate whether a tutorial encouraged participation. One competition was about purchase prediction in an online market, while the other was about the prediction of solar energy production. We provided a tutorial article only for the online market competition, and checked how the number of submissions and that of participants changed before and after posting the tutorial.

The two competitions targeted participants of the machine learning summer school in 2015, and were held during the same period, from July 24th to August 23rd, 2015. There were 22 and 45 participants in the online market competition and the solar energy competition, respectively; 17 people participated in both competitions. We had 187 submissions for the online market competition, and 669 for the solar energy competition. The simpler data in the latter competition resulted in more participants and submissions. The online market competition asked participants to predict which items a customer would buy in an online market by using a browsing behavior log and the attributes of users and items; the participants in this competition were required to combine multiple datasets to solve the problem.

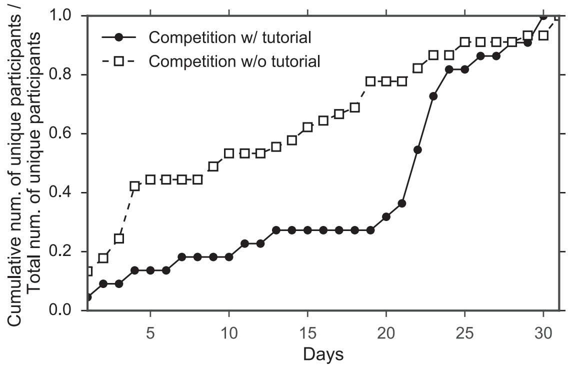
The tutorial for the online market competition was made available on the 21st day after the competition started. We did not provide a tutorial for the solar energy competition. Figure 3a shows the cumulative number of submissions of both competitions. We see that the number of submissions increased after the 21st day of the competition with the tutorial (i.e., the online market competition). Although we had 66 submissions in the first 21 days and the average number of submissions per day was 3.1 in this period, there were 121 submissions in the ten days after the release of the tutorial and the average number of submissions per day increased to 12.1. Note that seasonal effects do not explain this increase, as the number of participants and submissions was stable in the competition without tutorial (i.e., the solar energy competition) before and after the 21st day. Figure 3b shows the number of unique participants up to the date. In the competition with the tutorial, although we only had eight participants before the release of the tutorial, we obtained 14 new participants after the release. The new participants submitted multiple times and we conclude that the tutorial supported the first-step of participation and motivated active participations in the competition.

### 4.2 Leaderboard

Using another pair of competitions, we examine whether a Ladder leaderboard encourages the participants to make

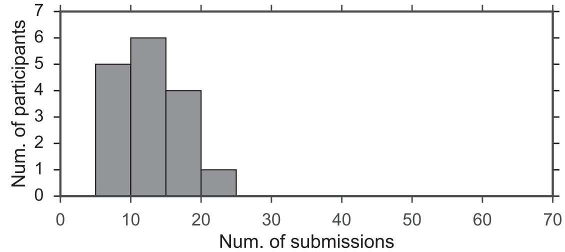



(a) Number of submissions

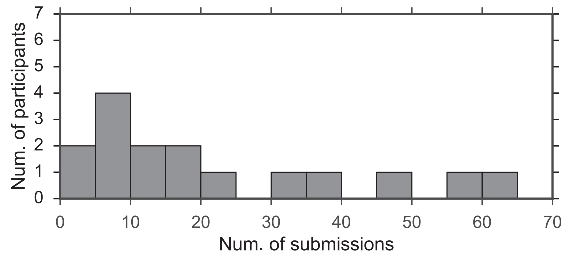



(b) Number of participants

Figure 3: Number of submissions and number of unique participants up to each date. In the competition with a tutorial, after the release of the tutorial (on the 21st day), both the number of submissions and that of participants greatly increased.




(a) Redular-leaderboard competition




(b) Ladder-leaderboard competition

Figure 4: Distribution of the number of submissions per participant. Participants in the Ladder-leaderboard competition tended to make a lot of submissions.

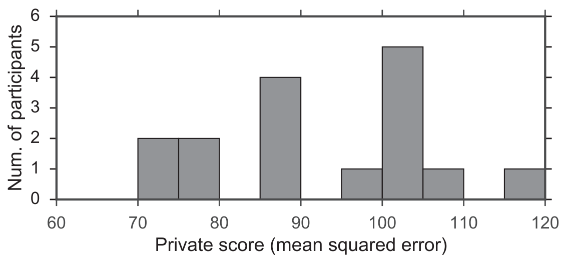
more submissions. A Ladder leaderboard allows us to remove the limitation of the number of submissions per day. The two competitions in this experiment were based on the same prediction problem using the same dataset, and were held in the same period. We employed a regular leaderboard system in one competition and a Ladder leaderboard in the other. The participants in the former competition were limited to submitting their outputs up to three times per day, and those in the latter competition did not have such a limitation. Participants took the same course at university; we randomly separated them into two groups, and a group was assigned to each competition. Each group consisted of 16 students. The competitions were about a regression problem, and they were held from June 23rd to July 27th, 2017.

Figure 4 demonstrates the distribution of the number of submissions in the redular-leaderboard competition, and that in the Ladder-leaderboard competition. The maximum number was 22 in the redular-leaderboard competition. In contrast, there were six participants who submitted more than 20 times in the Ladder-leaderboard competition. This fact indicates that the unlimited number of submissions with the Ladder leaderboard encourages active participations by the students. There were $191$ submissions in the redular-leaderboard competition, and $352$ in the Ladder-leaderboard competition; this number was $80\%$ more than the redular-leaderboard competition.

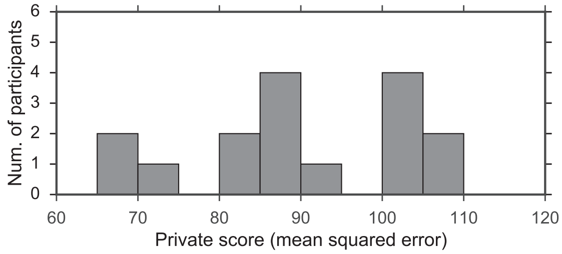
Table 1 shows the results of the top three participants in the competitions. Note that we used the same set of test samples for calculating the private scores in both competitions. Each of the scores was the mean squared error of the prediction; a lower score indicates a more accurate prediction. We observe that the top three players in the Ladder-leaderboard competition outperformed the best player in the redular-leaderboard competition. The first and the second winners of the Ladder-leaderboard competition submitted $34$ and $61$ times, respectively, and a lot of trial and error would bring higher performance. Figure 5 shows the distribution of the private scores in each competition, which demonstrates that the participations in the Ladder-leaderboard competition tended to build better prediction models than those in the redular-leaderboard competition. Figure 6 plots the relationship between the number of submissions and the private scores. In the Ladder-leaderboard competition, there is a clear pattern where players who submitted more times achieved better prediction performance. These results conclude that an unlimited number of submissions with a Ladder leaderboard motivates participants to use a lot of trial and error to continuously improve their predictive models.

## 5 Future Directions

Through the operation of the competition platform for learning data analysis for several years, we identify future research directions for enhancing educational effects.




(a) Redular-leaderboard competition




(b) Ladder-leaderboard competition

Figure 5: Distribution of private scores of each participant. The average score of the Ladder-leaderboard competition is better than the redular-leaderboard competition.

## 5.1 Collaborative learning

Data analysis competitions are designed to make individual players (or teams) compete for reward. Competition is a useful form for motivating learners; however, we should additionally consider encouraging *collaboration* between players for providing efficient learning opportunities. In our case studies, we observed that a player tried a prediction model (multi-layer perceptron) and mentioned its efficiency to the friends; the friends then improved their scores by using the model and a few of them outperformed the player who first tried the model. If we intensify such knowledge sharing on online platforms, we can accomplish collaborative learning among classes or organizations. Assessing the contribution of active knowledge sharing as an *assist point* would be helpful for encouraging collaboration. An incentive mechanism for collaboration in data analysis competitions has been proposed (Abernethy and Frongillo 2011). This mechanism is based on a competition system where each submitted model is open to all participants and they can submit an improved version of the current model. The final reward for a participant is determined based on the degree of improvement in each step. Such a systematic mechanism can be applied in educational data analysis competitions for supporting collaborative learning.

## 5.2 Skill visualization

Performing data analysis projects in practical situations requires a variety of skills; for example, a data analysis project consists of several phases, such as preprocessing, feature engineering, and predictive modeling, which need different types of skills. There are various problem settings including regression, classification, recommendation, and anomaly

Table 1: Ranking of the top three players. The top three players in the Ladder- leaderboard competition outperformed the best player in the redular-leaderboard competition.

(a) Redular-leaderboard competition

| Rank | Private score (Mean squared error) | Num. of submissions |
|---|---|---|
| 1 | 72.70 | 19 |
| 2 | 72.89 | 7 |
| 3 | 76.91 | 11 |

(b) Ladder-leaderboard competition

| Rank | Private score (Mean squared error) | Num. of submissions |
|---|---|---|
| 1 | 67.57 | 34 |
| 2 | 67.79 | 61 |
| 3 | 71.64 | 10 |

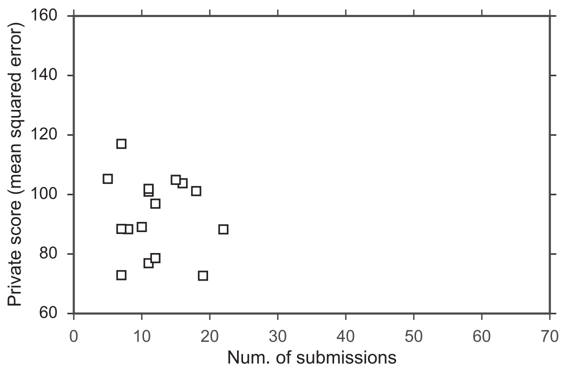
detection, and data types such as texts, images, and spatial-temporal data, and learners have different knowledge and experiences with these. Visualization of strengths and weaknesses of each participant would be helpful for learning data analysis skills. A method for skill visualization based on winning information of crowdsourcing contests (Baba, Kinoshita, and Kashima 2016) can be applicable for this purpose.

## 5.3 Competitions for exploratory data analysis

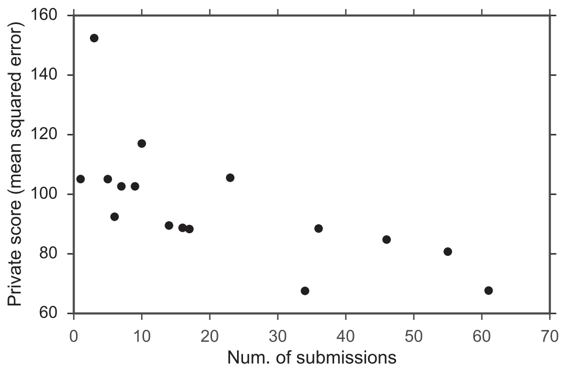
In addition to machine learning or predictive modeling, there is another type of data analysis called *exploratory data analysis* (Tukey 1977), which aims to understand data through visualization and summarization to gain insight or knowledge from it. Particular efforts are made for exploratory data analysis in some data analysis projects, and training this type of data analysis skills is important in practical situations. There have been a few competitions of exploratory data analysis. In KDD Cup 2000, for instance, participants were asked to answer the question, "Given a set of purchases over a period of time, characterize visitors who spend more than $12 on an average order at the site?", by analyzing the dataset. Kaggle offered a competition asking participants to find insights from the US census data. American Statistical Association (ASA) DataFest[7] is a hackathon event where teams of undergraduate students are asked to solve a exploratory data analysis problem.

Submissions in predictive modeling competitions are quantitatively evaluated by using prediction accuracies; however, it is a crucial issue to prepare a suitable evaluation method for the results of exploratory data analysis. In the KDD Cup 2000, submissions were graded by experts according to presentation quality, correctness, and importance of each insight. This approach has a problem of scalability when the number of submissions increases. A promising solution is peer-grading, which has been applied in MOOCs (Piech et al. 2013; Raman and Joachims 2014), where each student evaluates the submissions from other

[7] http://ww2.amstat.org/education/datafest/index.cfm




(a) Redular-leaderboard competition




(b) Ladder-leaderboard competition

Figure 6: Number of submissions and private score of each participant. There is a clear pattern that the participants who submitted more times achieved better prediction performance in the Ladder-leaderboard competition.

students. This approach would be beneficial for learning by inspecting the submissions from others. Another solution is to ask crowdsourcing workers to evaluate the submissions (Baba and Kashima 2013; Sunahase, Baba, and Kashima 2017). It was reported that crowdsourcing workers can reliably assess the results of exploratory data analysis (Baba and Kashima 2015). Designing a suitable evaluation workflow of peer-grading or crowd-grading is important for implementing competition platforms for exploratory data analysis.

## 6 Conclusion

We investigated an efficient platform design of educational data analysis competitions by building our own competition platform and operating several competitions. We introduced our approaches for supporting beginners and motivating participants, such as, tutorial articles, a leaderboard mechanism for an unlimited number of submissions, a user rating system, winner reports, and similarity visualization. Through case studies, we examined the efficiency of tutorial articles and the leaderboard mechanism. We observed that the tutorial articles contributed to encouraging the participation, especially for beginners, and the Ladder leaderboard motivated participants to perform more trial and error and helped to motivate the efforts from the participants. There are possible future works including a mechanism for enhancing collaborative learning, skill visualization, and competition design for exploratory data analysis.

## References

Abernethy, J. D., and Frongillo, R. M. 2011. A collaborative mechanism for crowdsourcing prediction problems. In *Advances in Neural Information Processing Systems 24 (NIPS)*.

Baba, Y., and Kashima, H. 2013. Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

Baba, Y., and Kashima, H. 2015. Crowdsourcing data understanding: a case study using open government data. In *An adjunct to the proceedings of the 3rd AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Baba, Y.; Nori, N.; Saito, S.; and Kashima, H. 2014. Crowdsourced data analytics: a case study of a predictive modeling competition. In *Proceedings of the 2014 International Conference on Data Science and Advanced Analytics (DSAA)*.

Baba, Y.; Kinoshita, K.; and Kashima, H. 2016. Participation recommendation system for crowdsourcing contests. *Expert Systems with Applications* 58.

Blum, A., and Hardt, M. 2015. The ladder: a reliable leaderboard for machine learning competitions. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*.

Kohavi, R.; Brodley, C. E.; Frasca, B.; Mason, L.; and Zheng, Z. 2000. KDD-Cup 2000 organizers' report: peeling the onion. *ACM SIGKDD Explorations Newsletter* 2(2).

Piech, C.; Huang, J.; Chen, Z.; Do, C.; Ng, A.; and Koller, D. 2013. Tuned models of peer assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM)*.

Raman, K., and Joachims, T. 2014. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*.

Sunahase, T.; Baba, Y.; and Kashima, H. 2017. Pairwise hits: Quality estimation from pairwise comparisons in creator-evaluator crowdsourcing process. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence (AAAI)*.

Töscher, A.; Jahrer, M.; and Bell, R. M. 2009. The BigChaos solution to the Netflix Grand Prize. http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Addison-Wesley.