Boosting performance in data science competition using topic-driven analytics: evidence from recommendation system design on Kaggle

***Abstract*—**Research developments in the recommendation system and electronic commerce literature present more accurate and comprehensive recommendation system solutions. However, while these developments add new features to the recommendation systems, the question of whether a novel solution would excel in practice remains. Open innovation and crowdsourcing platforms are becoming an arena for designers to test their solutions in business competitions. We show how structural topical modeling identifies topical themes that improve contestant performance using forum message data during the competition period. Our topic modeling analysis identifies technological and business issues that emerge in recommendation system development. An econometric framework further investigates the link between topic distribution and performance. The multi-period difference-in-differences estimator reports no significant statistical relation when linking all message communications to the performance. However, topic-dominant and topic-dispersed messages are both found to positively and significantly impact performance. Our result shows that structural topical modeling has an essential role to critically examine the most valuable message links to boost performance. Stakeholders may prioritize the messages with specific topics and/or a mixture of topics. We provide research and practical implications for researchers, business analysts, developers, and managers to improve their experiences when engaging in recommendation system design on platforms.

***Index terms*—Recommendation systems; Structural topic modeling; Knowledge sharing; Decision support; Difference-in-differences**

I. INTRODUCTION

It is increasingly acknowledged that a recommendation system can be vital to the success of a company. For example, 80% of the television programs that people watch on Netflix are discovered by its recommendation system. Those programs recommended to Netflix customers are decided by machine learning algorithms using datasets that include customer profile, TV content, and so on [1]. In the meantime, companies are facing constant challenges in using recommendation systems. Netflix's recommendation system collects datasets from its interaction with customers; for instance, the timestamp, the device, and the duration of customer views. The changing nature of these customer behavior patterns poses challenges to the decision support system to constantly evolve to re-train the machine learning model in order to provide robust and accurate results [2]. The availability of machine learning technology offers a wide range of options that companies can strategically use in their businesses [3], [4]. However, it can be a difficult task to design and deploy recommendation systems in different business contexts due to the demanding efforts required to design proper algorithms and to deploy them as the information systems that best meet business needs.

Data scientists and business practitioners discuss these challenges in their business projects and try to come up with solutions to provide customers with the ultimate experience using data science technology. For instance, data scientists often use online platforms to share their design learning experiences. Profound discussions on novel algorithms, experimentation setup, dataset usage, and evaluation metrics are shared on blogs, websites, and social media. These platforms allow experts to publicly share knowledge on data science. Companies host competitions to challenge the public to create innovative solutions to data science challenges.

Very often [5], the designers encounter issues and discuss these with each other. The website forum hosts discussions on different competitions concerning various subjects, e.g., design of the system and business interpretation of the analysis results. This paper uses the

forum discussions to uncover the challenges and potential links to performance in recommendation system design when participating in Kaggle competition. Therefore, we address the following research questions:

- What are the universal challenges of recommendation system design discussed in the competitions?
- What are the categories of those challenges (e.g., data science-related challenges, managerial challenges)?
- How do forum communication messages influence performance?
- What types of message topic influence the performance?

We deploy a structural topic modeling approach to identify topics from the forum discussions. We quantify the weight of different topics – namely, topic proportions – in topic modeling. We present keywords associated with the topics and then refer to the text to better understand the discussions over potentially thousands of messages and millions of words. The topic distribution drawn from topic proportions is integrated into the econometric framework where we try to explain and predict designer performance using message communication,

Our work advances the research on the Kaggle platform, data science innovation coopetition, and recommendation system design in multiple ways. First, we use structural topic modeling to identify the design challenges discussed in the competition forum. Our finding is based on an empirical setting where companies post their real-world challenges. We also link messages to team performance in specific competitions to predict increase in performance. We run difference-in-differences (DID) estimators to test the statistical effect of message exchange on performance over multiple time periods. Three types of treatment measure are considered: all types of messages, messages labeled with the most statistically dominant topic by the structural topic model, and messages labeled with a wide spread of weights on all topics. To the best of our knowledge, this is the first study to critically investigate design challenges in

competition discussions using text analytics and causal inference with a quasi-experimental design. Prior studies primarily focus on the performance of the competitors without exploring the discussions [6], [7]; even though the studies are motivated by the discussions [6]. Our work enhances the research in innovation and performance improvement in platform economy, as we show the value of Kaggle competition communities to improve project team performance. Second, our work uses a novel research method – the structural topic modeling and advanced difference-in-differences technique [8] – to extract topics from discussions, which helps us to better understand the dynamics behind message communication and performance, in contrast to studies that focus on quantitative data, qualitative data using interviews, or solely topic modeling [6], [9]–[13]. This would allow companies to better assess the design issues and take effective actions, e.g., to improve interdepartmental communication, provide professional skill training for staff, and invest in data science artifacts/talents. The third contribution of our work is that we conduct a holistic review of literature from distinctive backgrounds in computer science, engineering, business, and social science. We suggest that future recommendation system research should consider mixed paradigms (machine learning vs. behavioral sciences) and mixed methods (qualitative vs. quantitative). Our research results support platform owners, business innovators, analysts, data scientists, and developers who are involved in recommendation system businesses.

In the next section, we undertake a literature review of recommendation system design. Following this, we introduce the modeling technique and data collection. We highlight topic modeling techniques and compare the state-of-the-art technique – structural topic modeling [14] – to conventional methods such as Latent Dirichlet allocation [15] to show the methodological advances in our research. An econometric framework assesses the impact of topical distribution from message communication over performance. Following the methodology section on modeling techniques, we present our case studies using datasets

collected from five real-life competitions with a total of 188,334 words, 510 discussion threads, and 3265 messages. Capitalizing on the extracted topics, we test the message's effect on competition team performance using difference-in-differences estimation. In the results and discussion section, we discuss our findings using the discovered topics and validate them with the original discussion text. We further discuss the statistical results using topic distribution as a predictor of performance. We identify some of the challenges in the implication sections and discuss the importance of our findings and potential contributions. We conclude the paper with limitations and suggestions for future research.

II. LITERATURE REVIEW

In the literature review section, we first introduce how data science technologies such as recommendation systems are relevant to platform businesses like Kaggle competition. Then we review how platform economy is essential to business success, with its potential network effects and community structure.

The recommendation system is an essential part of data science technology in electronic commerce and companies are interested in using such systems to improve their businesses to better engage customers with their products and services. Prior research makes an effort to develop the business functions of the recommendation systems – for instance, employing data-driven approaches to use context-related metrics and social media data to achieve more accurate recommendations [16], [17]. The customers' decision-making process behind the recommendation system is of interest to researchers as well [18], [19].

Novel algorithms and statistical inference techniques are used for technological advancements. Novel recommendation system applications are proposed in many different areas such as hotel and tourism management [17], social media content recommendation [16], and news article recommendation [20], [21]. Novel data science algorithms – e.g., deep

learning – are systems used to enhance the performance [20]–[22]. While these technological advances are inspiring, a recent study shows that we might need to have second thoughts before accepting some of the reported successes [23]. Following a survey of previous studies that propose novel recommendation systems, researchers critically examined the methods suggested by these studies. Only seven out of 20 selected algorithms were proved to be reproducible with reasonable effort while the others unfortunately were not. Six of these seven methods are outperformed by other traditional simpler methods, so the contribution that is claimed might be questionable. The algorithm might only perform well on some datasets [24] in specific domains such as movie recommendation but may not perform well in other domains such as job recommendation. This suggested that business owners who are interested in adopting such technology might be interested to test these solutions publicly before production, such as using a platform like Kaggle to host competitions.

Prior literature argues that companies' internal business processes are critical to support data science innovation using the Kaggle platform [25]. User behavior captured from business processes adds more insights, as prior work suggests that using community detection and association rules in this setting would provide improved recommendation results [26]. Other works consider user behavior such as social trust and bias [27], [28] to improve the decision making that takes place in recommendation systems. User experience and customer review data can be useful to understand recommendation system use [18], [19], [29], [30]. Experiments show that positive opinions and recommendation output jointly presented to customers are more likely to be accepted, but in the case that the previous customer's recommendation is not consistent with the system recommendation, there may be negative consequences [19].

Apart from the efforts to engage with recommendation system users, researchers are also interested in understanding the innovation process of generating novel data science solutions. In the innovation context, a social development process is observed where companies

obtain innovative solutions from a crowd [31], [32]. Researchers reveal the complex dynamics behind the competition, involving how competitors compete, support, and cooperate with each other [9], [33]. More specifically, the current support mechanism, which aims to facilitate the innovation, is investigated [7] and challenged [6]. The information-seeking and information-sharing activities are widely observed in the innovation context [7]. They could be vital resources to provide useful feedback to competitors [6]. While there are many works published in innovation research, there is limited evidence to show whether the findings could be relevant to performance in data science competitions hosted by companies [7], [31], [32].

Business innovators often turn to platforms in the hope to achieve business success [34]. Platforms like Kaggle connect businesses with individuals and teams to collaborate on data science challenges. Prior literature interviewed data science experts in industry and concluded that platform connects business and crowdsourcing expertise successfully [25]. The study also emphasized that it is vital to create permanent communication and collaboration via Kaggle competitions through channels such as discussion forum rather than seeking for something "quick and cheap" [25]. This is in line with some of the past works which studied the benefit of the network effect from platform economy [35], [36]. Prior literature explored the ecosystem behind the success of the platform business model, a component of which is how external forces – for example, resources and talents – could have a significant impact [37]. While research recognizes the impact of the community and network effect, earlier research also argues that understanding regarding distinctive mechanisms originating from platform designs is absent [38], [39]. More fundamentally, research attempted to understand innovation diffusion in a network context, where novel information is measured via modeling topics from email text corpus [13]. Topical distributions – e.g., focused and diverse topics – are used to measure the novelty of the information, which is generally considered to be valuable in the research context to understand social capital and workplace performance [13]. From such

literature we are eager to know about an open platform setting where the entire community has access to message exchanges, and how individuals and teams perform in such a context. While information technology enables a connected world, research lacks an understanding of the content flow in a networked economy and its non-static nature in a dynamic business environment [40]. We learn from the literature that there are significant gaps between data science design practices, as well as a lack of opportunities to investigate the challenges and overcome them through more robust theory and practice. We, therefore, aim to identify the impact of interaction within the competition forum on design performance. Particularly, we are interested to know how content within the community such as focused and diverse topics would impact performances on the Kaggle platform over time during competition.

III. METHODOLOGY

A. Structural Topic Modeling

The literature discussed above mainly focuses on structured data in numerical forms – e.g., quantitative survey data and performance score indicators. We explore, in greater depth, the unstructured data context where most of the information is available in the form of text. Prior research shows that, if participants focus too much on the score feedback (e.g., predictive accuracy score) that is given solely by the competition platform, they tend to underperform [6]. In forum discussions, some participants share their competition experience and point out that, if they only use the score system from the Kaggle website (https://www.kaggle.com/kaggle/meta-kaggle), this may lead to a biased solution [6]. Designers may develop better solutions by taking other contestants' suggestions into account. Hence, we posit that the text data from discussions can add value to the competition.

Obviously, we cannot read all text from the document files by ourselves and it takes a long time to summarize them without any help. Topic modeling is a statistical technique which

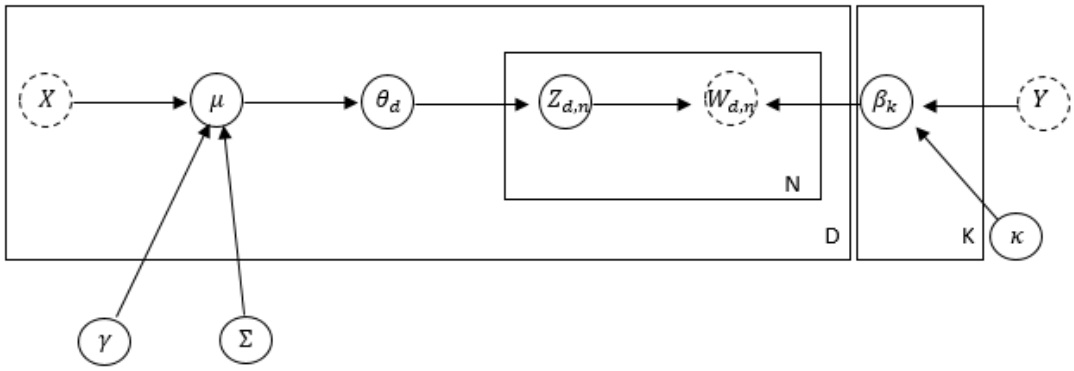
allows us to extract topics and key words from the text data. We are in favor of this approach because it allows us to summarize the main topics using extracted keywords from the text and we can search the specific paragraphs for further detail.

There are a number of different ways to model text data using topic modeling; one of the most popular techniques is Latent Dirichlet Allocation (LDA) [9] . A piece of text, or, more specifically, a document is considered as a combination of different topics, while in each topic there are different keywords. Using the observed words in the documents, LDA tries to infer the probability distribution of the topics in the documents, and the keyword distribution in the topics [12], [42]. Please refer to the **APPENDIX** for further details on LDA.

While the LDA is interesting, it has some limitations. The LDA only works with words and does not take other types of variables into account. Documents may have different characteristics and sometimes it might not be appropriate to treat them the same way [43], [44]. Structural topic modeling (STM) [45] extends the LDA with its ability to take into account categorical or numerical variables during the model inferences. For example, a user might discuss the recommendation system design in several posts at different stages of the competition; however, what they say might vary based on their progress. STM is able to use the information of the author of the posts and the time stamp of the post in its modeling process. This allows us to address the heterogeneity of the datasets (document type, author profile) [14] and behavior change over time.

In structural topic modeling, additional variables are included to take topical prevalence and topic content into account, as shown in **TABLE I.** Extra statistical inference steps will take place to estimate the distribution parameters from various sources – e.g., Gamma, normal, Laplace, and exponential distributions.

TABLE I
ADDITIONAL PARAMETER NOTATIONS OF STRUCTURAL TOPIC MODEL



X	Document-specific variable(s) used to build topical prevalence
Y	Document-specific variable(s) used to build topical content
γ	Distribution parameter for topical prevalence
Σ	Distribution parameter to generate topic distribution θ_d
κ	Distribution parameter for topical content

Topical prevalence:

1. Generate distribution parameter $\sigma_k \sim$ Gamma (s_γ, r_γ)
2. Generate distribution parameter $\gamma_k \sim N\ (0, {\sigma_k}^2)$
3. Generate distribution parameter $\mu_{d,k} = X_d \gamma_k$

Topical content:

1. Generate distribution parameter $\tau_k \sim$ Gamma (s_k, r_k)
2. Generate distribution parameter $\kappa_k \sim$ Laplace $(0, \tau_k)$
3. Generate distribution parameter $\beta_k \propto$ Exponential (κ_k)

The language model:

1. Generate distribution parameter $\theta_d \sim$ Logistic Normal (μ_d, Σ)
2. Generate topic $Z_{d,n} \sim$ Multinomial (θ_d)
3. Generate word $W_{d,n} \sim$ Multinomial $(\beta_{d,k=Z_{d,n}})$

The main differences in the two techniques are summarized in **TABLE II** below. Prior study of STM affords a simple and straightforward comparison of the techniques [46]. STM is able to use the variables such as "topical prevalence" and "topical content" during the statistical inference to estimate topic and keyword distributions. We omit the technical details and algebraic notations of LDA and STM, and refer interested readers to Blei, Ng and Jordan [15], [46].

TABLE II
DIFFERENCES AND BENEFITS OF USING THE STRUCTURAL TOPIC MODEL

	Topic model (Latent Dirichlet allocation)	**The Structural Topic Model (STM)**	**Benefits using STM**
Topic distribution within the document is:	A random variable from one fixed Dirichlet distribution.	A random variable drawn from a Lognormal distribution that is based on document-level data.	Account for the differences between different documents.
Word distribution within the topic is:	Common across the corpus.	Based on topic, document-variable data, and topic-variable interactions.	Adjust the word distribution based on variables.

B. The Econometric Framework to Predict Performance

Message-level topic distribution provides a statistical trend of topics within each message. Using such information, we operationalize the message-level measure: topical dominant and dispersion. This allows us to identify unique types of "treatment". While a dominant topic shows a clear theme within a post, dispersion could suggest a range of topics that are significant within a post text[47]. For each competition, statistical inference is drawn to estimate the STM with hyperparameter that finds the best trade-off between *semantic coherence* and *exclusivity*. Each message i in competition c is then labeled with $Z_{i,c}$, with the most dominant topic given the topic weight $\theta_{i,c}$ within that message.

$$Z_{i,c} = arg \max_{i \in \{1,2,..|c|\}} \theta_{i,c} \, . \quad \textbf{(1)}$$

At a given moment t the specific message i is posted. One can indicate when a message communication appears:

$$G_{i,t}^{1} = \begin{cases} 1 \; if \; a \; message \; is \; posted \; by \; team \; i \; at \; time \; t \\ 0 \; else \end{cases} . \quad \textbf{(2)}$$

Alternatively, one could only consider the message with a dominant topic:

$$G_{i,t}^{2} = \begin{cases} 1 \; if \; Z_{i,c} = Z_c \\ 0 \; else \end{cases}, \quad \textbf{(3)}$$

where $Z_c = arg \max_{i \in \{1,2,..|c|\}} \widetilde{\theta_{i,c}}$ is the most dominant topic among competition c. This is a "filtering" mechanism to retain only the most statistically "dominant" messages for further analysis.

We are also interested to measure the dispersion of the topic weights using standard deviation of topic weights. Prior literature is interested to measure messages that have focused or diverse topics [13]. The more dispersed the measure is, the more topic-related content there is, than just one or a few dominating ones [47]. To the best of our knowledge, we are the first to adjust this in the Kaggle context and use it as a treatment.

We use the median value of all topic weight standard deviations $\sqrt{E(\theta_{i,c}^2) - E(\theta_{i,c})^2}$ as the threshold to filter the most "topical dispersed" messages. This allows us to extract the most diverse topics despite the topical distribution.

$$G_{i,t}^{3} = \begin{cases} 1 \; if \; \sqrt{E(\theta_{i,c}^2) - E(\theta_{i,c})^2} < median(\sqrt{E(\theta_c^2) - E(\theta_c)^2}) \\ 0 \; else \end{cases} . \quad \textbf{(4)}$$

Together with the performance score obtained from submission data, we could test whether the topic distribution information would have an impact on the performance, as shown in **Fig. 1**.

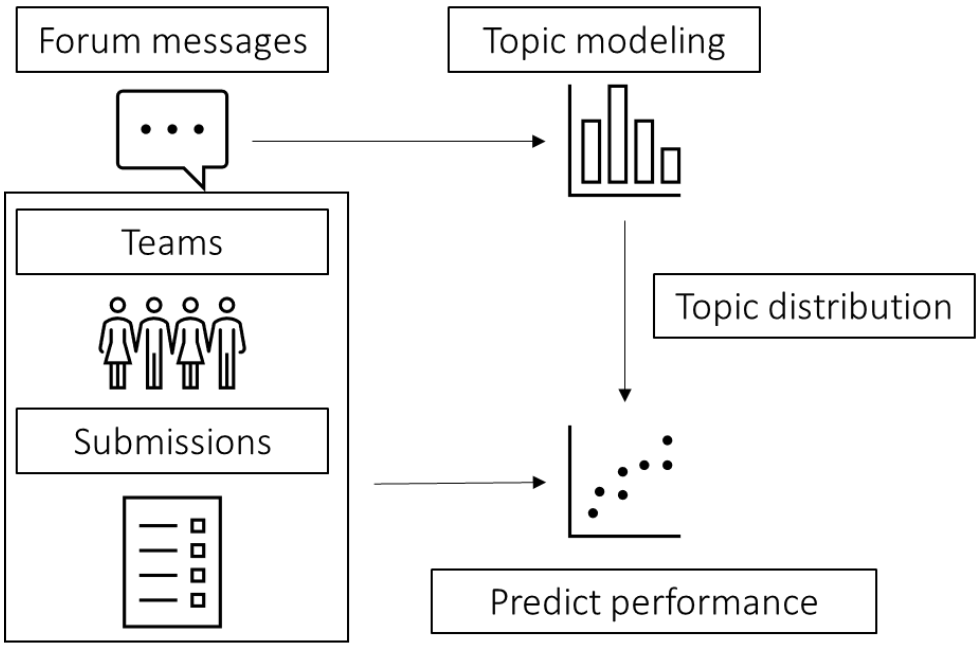
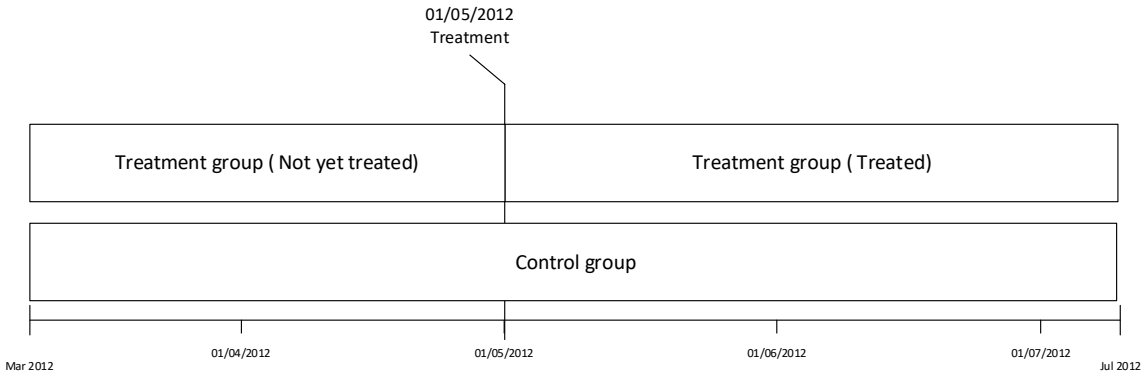
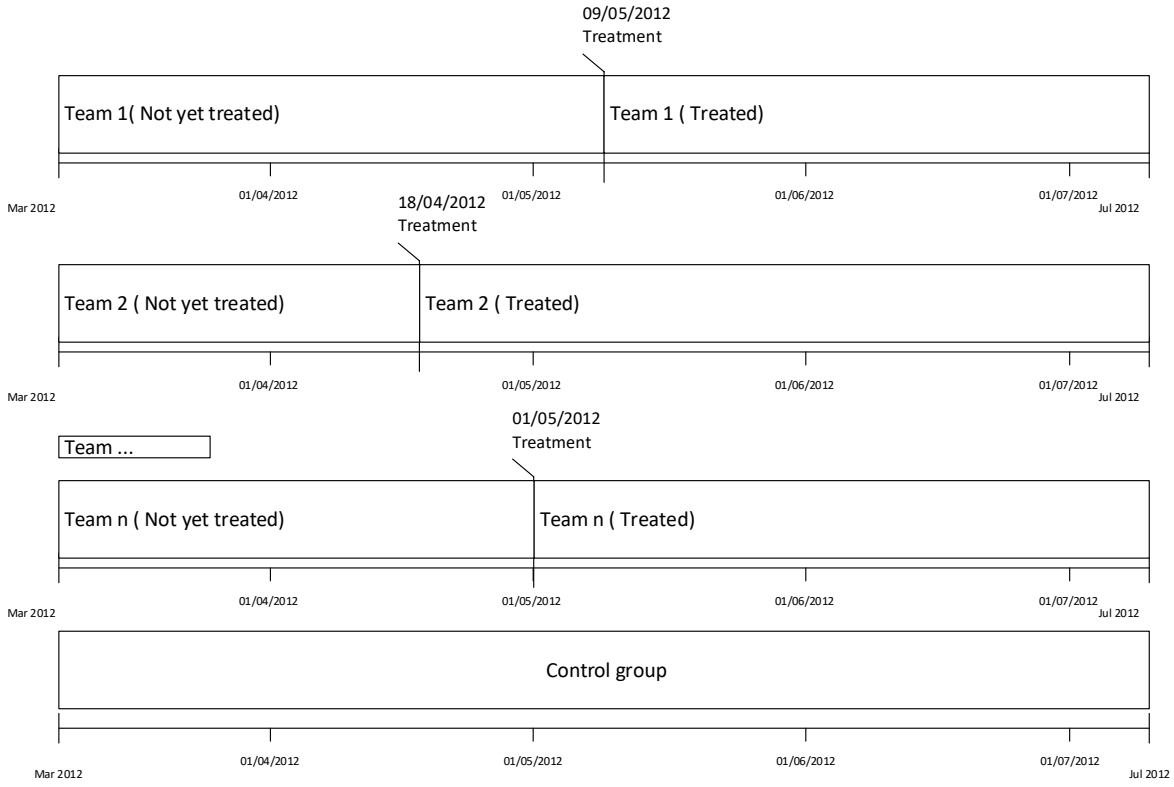


Fig. 1. An overview of the main research method.

The traditional difference-in-differences estimator operates in a treatment versus control group setting while, in this paper, the treatment could come into effect at different time periods. For instance, different teams might message and respond at different points in time, without a common timeline to receive treatment. This is due to the nature of the activity/process in that a team could submit a solution any time during the competition period, as indicated in **Fig. 2**.



Traditional difference-in-differences estimator setup



Difference-in-differences estimator setup with multiple time periods

Fig. 2. Difference-in-differences estimator setup – an illustration.

To formally define the econometric framework that estimates the effect of the treatment, we denote the performance outcome as follows.

$$Y_{i,t} = Y_{i,t}(0) + \sum_{g=2}^{\tau} \left(Y_{i,t}(g) - Y_{i,t}(0) \right) \cdot G_{i,g}. \quad \textbf{(5)}$$

Within a period of time $t \in \{2,3 \dots, \tau\}$, $Y_{i,t}$ is the outcome for observation i at time t. G is the first time period when a team gets treated, and G_g is a binary variable $G_g = 1$ if treated in period g.

The **a**verage **t**reatment effect for **t**reated samples (ATT) is estimated using the difference in outcome regarding the treatment group at specific period g.

$$ATT(g,t) = E(Y_t(g) - Y_t(0) | G_g = 1) \quad \textbf{(6)}$$

The ATT gives a picture of the treatment effect on outcome, taking the effect across multiple periods into account. The summary of the ATT makes use of aggregated weights $w(g,t)$, so that the aggregated scheme takes the form

$$\theta = \sum_{g \epsilon G} \sum_{t=2}^{\tau} w(g,t) \cdot ATT(g,t). \quad \textbf{(7)}$$

Weights could be determined by empirical data such as the size of treatment group and length of time period exposed to the treatment [8]. We measure the performance outcome using the average performance every two weeks.

In the DID analysis, we deploy three types of treatment; these are all messages $G_{i,t}^1$, topic-dominant $G_{i,t}^2$, and topic-dispersion $G_{i,t}^3$. This allows us to refine our hypotheses:

Hypothesis 1. (Just any) Message exchange will not increase team performance.

Hypothesis 2. Topic-dominant message exchange will not increase team performance.

Hypothesis 3. Topic-dispersed message exchange will not increase team performance.

IV. CASE STUDY – DATA COLLECTION

The raw data are obtained from Meta Kaggle, Kaggle's public data repository on competitions, team members, submission scores, and kernels. Kaggle is a website that holds different competitions for many types of data challenge [9]. Datasets from Kaggle competitions have been used in many research studies [48]–[50]. We investigate the recommendation competitions in terms of their recency, and popularity in Kaggle. We also consider the application domain of the competition to be unique and applicable to everyday life. After initial screening, we find five different recommendation competitions and use these for our research. These competitions target data science individuals and teams who want to challenge the existing best practice. The contexts of the competitions differ; these could be music, jobs, events, hotels, and products to recommend. The details of these competitions are summarized in **TABLE III** below. Details about the competition objectives can be found in the **APPENDIX**.

TABLE III
A SUMMARY OF THE COMPETITION CHALLENGE

Title	Enabled	Deadline	# Teams	# Competitors	# Submissions	Average performance
WSDM - KKBox's Music Recommendation	9/27/2017	12/17/2017	1081	1253	15555	0.649
Job Recommendation	08/03/2012	10/07/2012	81	95	687	0.094
Event Recommendation	01/11/2013	2/20/2013	223	285	3021	0.340
Expedia Hotel Recommendations	4/15/2016	06/10/2016	1974	2209	22713	0.391
Santander Product Recommendation	10/26/2016	12/21/2016	1787	2084	28772	0.023

Overall, after they are enabled, the competitions last about 50 to 90 days, except for the event recommendation competition which lasts about 40 days. Although it is a relatively short period of time for a competition, it still attracts more competitors compared to the job recommendation competition (285 against 95). Possibly, because the job recommendation challenge was hosted early in 2012, fewer people were aware of it. We also notice that the

contents of the competitions are not the same so the desired outcome may differ based on distinctive objectives.

To compile the datasets for topic modeling, we extract the message content, the message topic, the poster ID, the time of the post, and the competition they posted to. A data-cleaning procedure is undertaken to address the data quality issue: we remove some messages with empty text, and remove some of the HTML tags when necessary. For instance, a tag like <quote> or
 is used to quote the previous message in a discussion and can appear many times. As a result, we remove tags to avoid potential bias when counting word frequencies. In total, we observe 510 discussion threads and 3265 messages with 188,334 words. There are some observed differences in terms of the volume of the discussions, as we summarize in **TABLE IV**. While the hotel and product recommendation competitions contribute 30% or more to the total, the job recommendation competition only contributes about 4% of the total messages. Public Score with Full Precision from the submission data is selected as the performance measure over time.

TABLE IV
DESCRIPTIVE STATISTICS FOR TOPICS

A total of 188,334 words, 510 topics, 3,265 messages				
Event names	# messages	Per cent of messages	# topics	Per cent of topics
7162 KKBox's Music Recommendation Challenge	519	15.90%	65	12.75%
3046 CareerBuilder “Job Recommendation Engine Challenge”	130	3.98%	27	5.29%
3288 Event Recommendation Engine Challenge	277	8.48%	64	12.55%
5056 Expedia Hotel Recommendations	1213	37.15%	169	33.14%
5558 Santander Product Recommendation	1126	34.49%	185	36.27%

V. RESULTS AND DISCUSSION

A. Structural Topic Model Setup – Topic Numbers

Prior work uses humans to determine the number of topics [51].While there are no fixed rules in selecting the number of topics, statistical metrics quantify how well a solution with a specific number of topics fits the data. Often there is a trade-off between the number of topics and statistical fit. Several metrics can be used to evaluate a topic model with a given number of topics; one of the most widely used is semantic coherence [52]. We denote $D(v_i, v_j)$ as the frequency count of word v_i and word v_j appearing in the same document. For a topic model that has k topics containing a list of M words, its semantic coherence is computed as

$$C_k = \sum_{i=2}^{M} \sum_{j=1}^{i-1} log\ (\frac{D(v_i, v_j)+1}{D(v_j)}). \quad \textbf{(8)}$$

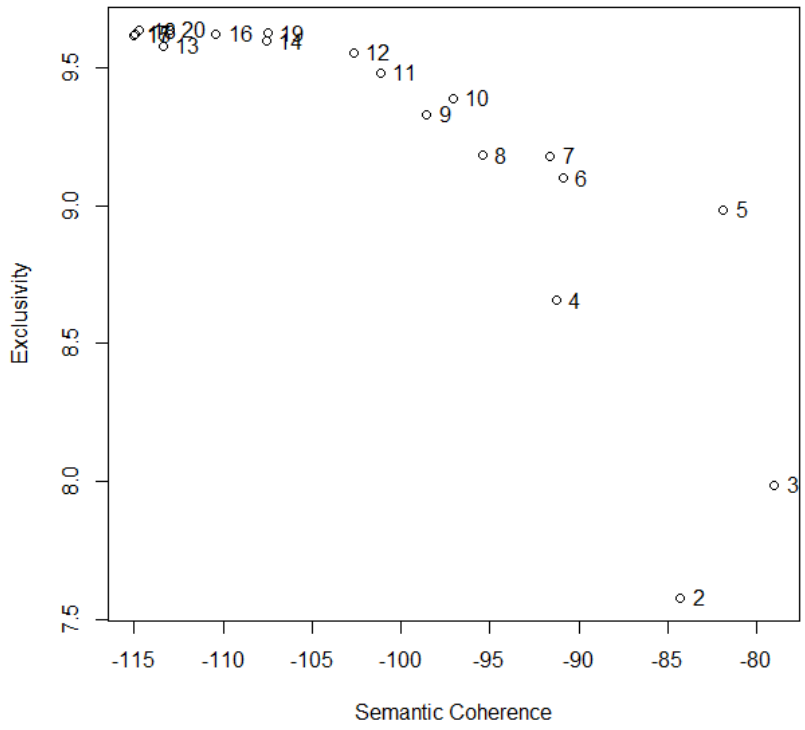
Semantic coherence reaches its maximum value when the most probable words from the same topic co-occur most frequently.

While semantic coherence is widely adopted in the literature, it can sometimes introduce bias [46]. When there are a few topics dominated by a few very frequently used words, the semantic coherence score can be very high without providing topics that are distinctive in meaning. Thus, alternative measures should also be considered. In practice, researchers often use some other measures such as the held-out likelihood, or they conduct residual analysis to accompany the evaluation of a topic model. The most frequently used measure is *exclusivity*. The rationale for this is that if a word is commonly observed in a topic, it might also be important to know whether this specific word is commonly seen in other topics as well or whether it is relatively exclusive to the specific topic [53]. Exclusivity measures how words are used differently across different topics [54]. The exclusivity score for a word v in

topic k is computed as the weighted harmonic mean of the word's rank in terms of exclusivity and frequency. The weight w is used to adjust the importance between exclusivity and frequency. The empirical cumulative distribution function is used to compute word proportion in topic $\beta_{k,v}$.

$$FREX_{k,v} = \left(\frac{w}{ECDF\left(\frac{\beta_{k,v}}{\sum_{j=1}^{K}\beta_{j,v}}\right)} + \frac{1-w}{ECDF(\beta_{k,v})}\right)^{-1}. \quad \textbf{(9)}$$

To find the proper k value for the number of topics, we run different alternatives from two to 20 topics. We can see the curve for the trade-off between them. We select the solution $k = 7$ in **Fig. 3** as it maintains a good balance between the semantic coherence and exclusivity, compared with other solutions such as $k = 6$ or $k = 8$ [55].



Statistics used to select number of topics
Semantic coherence = 9.178040
Exclusivity = -91.62606

Fig. 3. Structural topic model fit.

B. Structural Topic Model Topics

After setting the number of topics, we use competition ID and time stamp of the posts as topical prevalence in our STM. This allows us to control the heterogeneity among different competitions and model the moment when candidates submit their work. Topics are likely to be different in each competition because the objectives are different. A prior study further shows that candidates allocate different amounts of effort in the early and late stages of the competition [56]. Another study also shows that participants' behavior on Kaggle changes over time as they first focus on learning and then shift to submission [6]. We take these variables into account during the statistical inference of the topic model and conduct a robustness check by adding user ID to the model. This allows us to check whether user heterogeneity has an impact on the results. The statistical metrics are nearly identical, so our result is robust with or without modeling user heterogeneity explicitly.

TABLE V
TOPIC KEYWORDS AND PROPORTION

Labels	Topics	(Top) keywords	Proportions
Model training	Topic 1	Dataset train	0.142
Hotel and event recommendation	Topic 2	User hotel event	0.119
Data manipulation	Topic 3	Use file tri(ed)	0.156
Product recommendation	Topic 4	product month custom(er)	0.104
Model selection	Topic 5	feature use model	0.098
Result submission	Topic 6	Score submission result	0.154
Competition participation	Topic 7	Thank competition use	0.226

We present the keywords in the seven topics extracted from the documents in **TABLE V**. Each topic from 1 to 6 accounts for about 10-15% of the total topic proportion and topic 7 weights slightly more with about 22.6%. Based on the content we observe from the text, labels are given to different topics for meaningful interpretations. The full list of keywords per topic is available in the **APPENDIX**.

We use the keywords to trace back to the original text so that we can gain a better understanding of the topics. **TABLE XI** uses a few paragraphs of discussions to present some concrete ideas about the topics. This also allows us to qualitatively validate whether the topics we labeled from the topic model are consistent with what the text says. Interested readers can refer to the **APPENDIX** for a detailed overview.

We look at the topic proportion distributions in the discussions from **TABLE XI** and visualize the proportions in **Fig. 4**. This allows us to observe the topic distributions in different discussions – e.g., in *discussion 1* the focus is technical (topic 1 model training and topic 5 model selection). The other discussions are more focused on business such as *discussion 2*, as a high proportion of topic 2 is observed (Hotel and event recommendation). We also observe in *discussion 3* that relatively large topic proportions go to topic 2 (Hotel and event recommendation) and topic 3 (Data manipulation), so there is a mix between the business problem and the technical problem.

This raises another question about how often the topic proportions in the discussions are mixed. We answer this question below in the section entitled Topic Correlation Analysis. Last, it is also interesting to note that although topic 7 (Competition participation) overall has a large topic proportion, it is not always a dominant topic in discussions. A possible reason for this is that, although it is almost universal to talk about competition participation and thank those who provided their advice, the core discussion has to be about a specific problem – either a technical problem or a business problem. The competition participation and interaction are widely observed in different posted discussions but the participation and interaction in the competition are linked to problem solving.

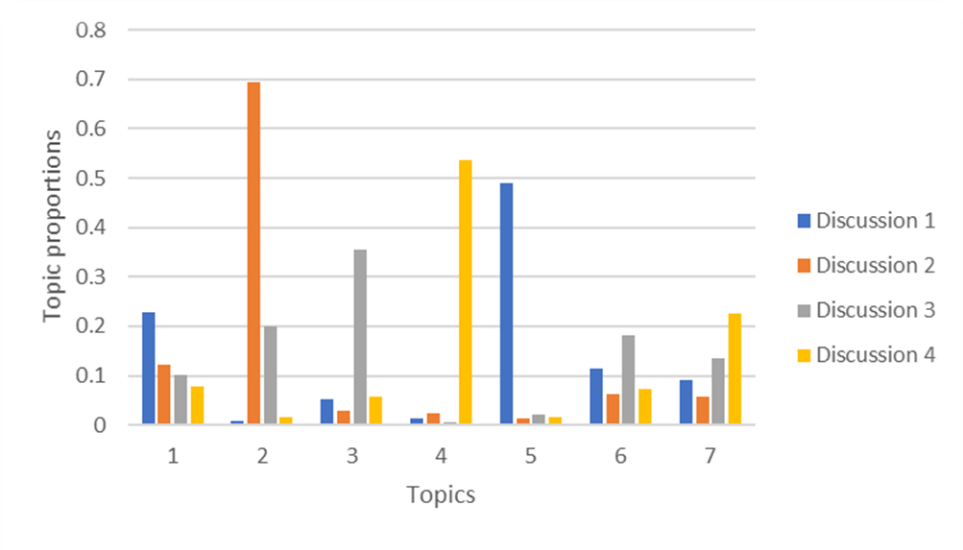


Fig. 4. Topic proportion distributions in discussions.

C. Topic Correlation Analysis

Above, we observe that a given text document will have a high score regarding a number of topics but a low value in other topic proportions, as we show in **TABLE VI**. We are also interested to know whether topics are positively related to each other. We want to conduct an analysis that covers all discussions in order to quantify the co-occurrences of the topics beyond the four discussions we covered in **Fig. 4**. The topic correlation measures the co-occurrence of the different topics in the text. Positive correlations among topics show that these topics are likely to appear together in the same document.

We observe no strong correlations among topics in **TABLE VI** (no correlation values > 0.5 or <-0.5). Participants try to concentrate on one topic or a few specific topics per post. The chance that all topics appear in the same post is low. Competition usage (CU) has a medium level of negative correlation values (from -0.151 to -0.344), meaning that when participants discuss their usage on the competition platform, it is likely that they are new to the system and thus unfamiliar with it. Since they are still new and trying to adapt to the system, they are less likely to be involved with other subjects in their posts.

TABLE VI
TOPIC CORRELATION TABLE.

	MT	HER	DM	PR	FE	RS	CU
MT	1.000	0.030	-0.119	-0.057	-0.035	0.000	-0.344
HER	0.030	1.000	-0.255	-0.189	-0.245	-0.205	-0.279
DM	-0.119	-0.255	1.000	-0.261	-0.042	-0.096	-0.174
PR	-0.057	-0.189	-0.261	1.000	-0.191	-0.123	-0.257
FE	-0.035	-0.245	-0.042	-0.191	1.000	-0.142	-0.151
RS	0.000	-0.205	-0.096	-0.123	-0.142	1.000	-0.165
CU	-0.344	-0.279	-0.174	-0.257	-0.151	-0.165	1.000

Correlation = 1	
Correlation = 0	
Correlation = -1	

MT	Model training	Topic 1
HER	Hotel and event recommendation	Topic 2
DM	Data manipulation	Topic 3
PR	Product recommendation	Topic 4
FE	Feature engineering	Topic 5
RS	Result submission	Topic 6
CU	Competition usage	Topic 7

We summarize our results from our analysis which shows a holistic understanding of the challenge in recommendation system design. Structural topic modeling results reveal the support-seeking and information-sharing behavior. We identify that topics such as model building, testing, and variable selection are the most frequent data science challenges discussed by the designers. We also discover that designers face major difficulties when they are unsure about the data-generation mechanism, so they need to understand the business context in order to acquire a better picture of the data-generation mechanism behind the scenes. In-depth investigation reveals the exchange between the host and the contestants in *discussion* 2. Through this type of discussion, contestants become aware of the user behavior which might not be obvious from a designer's perspective. This also allows the designer to gain an understanding about the hosting company's data management and data collection strategy.

D. Difference-in-differences Estimation On Performance

Further, at each competition, we investigate how the extracted topics would influence the performance. Topic distributions is used as a predictor to predict the performance.

TABLE VII
PRE-TEST OF PARALLEL TRENDS ASSUMPTION

	Message of all types	Topic dominant	Topic dispersion
7162	0.999	<0.01	0.999
3046	0.127	0.849	<0.01
3288	0.771	0.326	0.994
5558	0.988	0.332	0.453
5056	0.950	0.925	<0.01

p-value for pre-test of parallel trends assumption.
High p-value implies failed to reject the hypothesis that parallel assumption is violated.

The pre-test of parallel trends assumption gives a robust check to see if the data provided respect the underlying "parallel assumption" that difference-in-differences estimation holds. A statistically significant result (p-value < 0.05 in **TABLE VIII**) suggests that the data could not be used for further testing as the assumption is violated. Four out of five cases used in testing topic-dominant treatment effect satisfy the assumption and three out of five cases satisfy in testing topic dispersion. Propensity score matching is used to investigate further with the cases where the assumption is violated, as suggested in the literature[57], [58].

TABLE VIII
OVERALL TREATMENT EFFECT

	Message of all types	Topic dominant	Topic dispersion
7162	0.0039 (0.0062)	-0.1159 (0.0755)	0.1975 (0.0527) ***
3046	0.0208 (0.0195)	0.0795 (0.0373) **	*0.0591 (0.022)* ***
3288	-0.0129 (0.0279)	0.6560 (0.0021) ***	0.1803 (0.0816) ***
5558	-8.00E-04 (7.00E-04)	0.0060 (0.0018) ***	0.0083 (0.0012) ***
5056	-0.0286 (0.0177)	0.2298 (0.0367) ***	*0.0939 (0.0216)* ***

Overall treatment effect (Standard error)
Significance level p-value < 0.1 *, <0.05 **, <0.01 ***

The DID result shows that:

- Performance does not improve with just any type of message communication.
- Performance improves when a message with a dominant topic is exchanged.
- Performance improves when a message with a dispersion of topic is exchanged.

While one might think that exchange information would help to boost the performance, obviously only those exchanges with a distinctive topical feature do. Topic-dominant communication often implies a strong and easy-to-understand message, while topic dispersion shows heterogeneity and diversity of the idea. When tested in an aggregated setting, **TABLE VI** shows that topics are mostly exclusive to each other; however, this is somewhat different if one looks at a fine granularity within each specific competition. Given the relatively low topic correlation, the construction of the topic dispersion takes into account those with values that exceed the median.

Alongside the pre-test of parallel trends assumption, we test different alternative settings to see if we could obtain a robust estimation result. Besides the bi-weekly average, we also test weekly performance. Results show that the shorter the time period in which we take the weekly average, the less likely it is that the parallel assumption would be satisfied due to the statistical fluctuation and data sparsity. It seems to be more reasonable to use two-week intervals to aggregate more submission performance data. While “simple” weighting is used with a strategy to aggregate the ATT, an alternative model using “dynamic” weighting yields similar results.

In the next section, we discuss the implications for stakeholders in the research society and the business world.

VI. IMPLICATIONS FOR RESEARCH AND PRACTICE

Topic modeling advances our understanding of the innovation process in data science – e.g., recommendation system design. The extracted topics provide us with a holistic approach mixing qualitative and quantitative data to cover distinctive factors discussed in the Kaggle competitions. Our result identifies various technical and business subjects from the discussion data during the information seeking on the platform. Topical distributions in messages through time are indicators that highlight specific communications and interactions that critically affect performance when Kagglers engage with the platform. Topical distributions exhibit varying patterns between the aggregated level and the competition level. Heterogeneity among competition communication thus makes it difficult to learn from one competition in order to excel in the next one, showing the distinctive network effects within each forum's community. Text analytics such as topic modeling could potentially support the communication and learning. Research should also not overlook the direct interaction between host and contestants [9], and should take into account new indicators from topic modeling. Data science competition hosts could be the agents to bridge the gap between data science talents and business as they could deploy text analytics such as topic modeling to support the innovation process.

Our research suggests several managerial implications. A data science development process should not only emphasize the technological topics; topics in the context, such as the customer behavior in the data, may also play a significant role in improving the project outcome. Therefore, dialogs between the designer communities are important to gain a better understanding of the mixed topics. To help designers effectively navigate through mass communications, a "taxonomy" system on the message text record could be deployed, benefiting from the topic modeling results. Companies should also think about what mixture of topics would lead to a boost in performance during the innovation process. Communication within one dominant theme topic could be useful as it highlights the major issue of the question

under investigation within the discussion forum. One should also not deny that a dispersed mixture of the topics is often the nature of innovation challenge, as is the case with topic modeling that shows mixed results in topic presentation. Organizations that make good use of topical information and run effective data analytics could lead innovation competition by acknowledging what significantly drives performance increase.

Our results can also be informative for platforms such as Kaggle and the companies that participate. We show evidence of what positive changes in a networked platform business model using data (text) analytics could be, in relation to possible business success [25]. It is advisable to extract useful feedback and provide it to participants, the platform, and companies. In additional to numerical performance indicators such as predictive accuracy, qualitative data – e.g., discussions, suggestions, and solutions – could also offer helpful feedback to contestants and eventually support the design process. Since not just any type of message from the mass communication would directly improve performance, stakeholders need to think carefully about the content flow in the networked economy before taking any further advice and engaging with message exchange.

Our finding is relevant for organizations seeking to improve their data science expertise. Open-source communities, such as Github (https://github.com/) and Stackoverflow (https://stackoverflow.com/), among many others [59], lead the way in data science innovation as a central hub to share open software and support. Our research suggests that the innovation platform could also be a valuable open source to enhance data science design, given that one understands the novel and valuable content within the platform community. With the hundreds and thousands of discussions available, key findings that help boost performance using topic modeling become essential building blocks of the knowledge discovery process in an organization's roadmap for innovation. Data science-driven technological change generates

complementary as well as different answers to an organization's enquiries relating to technological innovation.

VII. LIMITATIONS AND FUTURE WORK

While we use Kaggle competition discussion data on recommendation systems, we are aware that there are many other platforms that can hold discussions on machine learning model design, which may further inspire our research. For instance, Stack Overflow also has numerous discussions on relevant topics [60], [61]. Many technology companies also have their own online forums to discuss these subjects – e.g., Microsoft Azure, Matlab, and so on. The data source can never be exhaustive. In the future, comparisons should be drawn to gain additional insights from different discussion sources, and from different types of modeling task.

We are also aware that many current discussion strands are about elementary and fundamental issues of the subject matter. Advanced topics may not appear in the competition as the number of people with advanced knowledge is limited and they may be reluctant to share it. This may introduce potential bias in our research. Obviously, the sparsity of discussions on advanced topics make it harder to understand the more difficult challenges.

Prior literature mentions limitations of text analytics when subject to linguistic noise data and ambiguity [62], [63]. Behavioral constraints such as the manner of speech and writing style vary among different people. Bias can occur in text analytics research such as sentiment analysis, where negative wording can be replaced with positive comments as a tone of speech in a professional context or for sarcasm. While structural topic modeling does not directly rely on the polarity of the sentiments, we do consider that the writing style may vary. Hence it is difficult to completely translate the text data into numerical insights [62]. The fact that the text is written in the cyber space may also introduce uncertainty to the writing style.

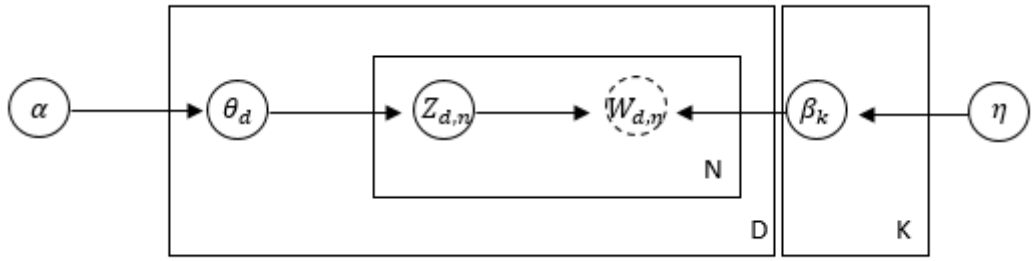
VIII. CONCLUSION

We demonstrate an effective knowledge discovery process from competition forum discussions on Kaggle. Companies developing data science projects could learn from the technical discussions on data science technology in this paper in order to better understand the business challenges behind the design process. Managers should realize that discussions, questions, and answers relating to these challenges are valuable data sources for future reference when developing their data science business. Using such knowledge, companies can tackle some of the challenges they face in technological design and business, add value to their data science talents, and innovate their data-driven business.

APPENDIX

LDA

TABLE IX
LDA PARAMETER NOTATIONS



N	Number of words
D	Number of documents
K	Number of topics
α	Dirichlet parameter prior of topic distribution in documents
θ_d	Topic distribution in document d
$Z_{d,n}$	Topic for word n in document d
$W_{d,n}$	The observed word n in document d
β_k	Word distribution of topic k
η	Dirichlet parameter prior of word distribution in topics

In LDA, the observed words are used to draw statistical inferences to find the distribution (from Dirichlet or multinomial distributions) that generates the topics and words. A generative process is used to describe this inference procedure in **TABLE IX**:

1. Generate distribution parameter θ_d~ Dirichlet(α)
2. Generate distribution parameter β_k~ Dirichlet (η)
3. Generate topic $Z_{d,n}$~ Multinomial (θ_d)
4. Generate word $W_{d,n}$~ Multinomial (β_k)

TABLE X
COMPETITION RULES AND OBJECTIVES

The 11th ACM International Conference on Web Search and Data Mining (WSDM 2018) – KKBox's Music Recommendation Challenge **Competition tasks**: How would an algorithm know if listeners will like a new song or a new artist? How would it know what songs to recommend to brand new users? **Company objectives**: The dataset is from KKBOX, Asia's leading music streaming service, holding the world's most comprehensive Asia-Pop music library with over 30 million tracks. They currently use a collaborative filtering-based algorithm with matrix factorization and word embedding in their recommendation system but believe new techniques could lead to better results.
CareerBuilder "Job Recommendation Engine Challenge" **Competition tasks**: To predict what jobs its users will apply for based on their previous applications, demographic information, and work history. **Competition objective**: The objective of the Competition is to develop an algorithm that uses available job seeker data to predict what job opportunities job seekers are most likely to apply for. Once that information is known, the sponsor will be able to recommend those job opportunities to job seekers.
Event Recommendation Engine Challenge **Competition tasks**: To predict what events our users will be interested in based on events they've responded to in the past, user demographic information, and what events they've seen and clicked on in our app. **Competition objective**: The objective of the Competition is to develop an algorithm that uses available user and event data to predict user interest in events.
Expedia Hotel Recommendations **Competition tasks**: Expedia wants to take the proverbial rabbit hole out of hotel search by providing personalized hotel recommendations to their users. **Competition objective**: Which hotel type will an Expedia customer book? Expedia is challenging Kagglers to contextualize customer data and predict the likelihood that a user will stay at 100 different hotel groups.
Santander Product Recommendation **Competition tasks**: Santander is challenging Kagglers to predict which products their existing customers will use in the next month based on their past behavior and that of similar customers. **Competition objective**: With a more effective recommendation system in place, Santander can better meet the individual needs of all customers and ensure their satisfaction no matter where they are in life.

Extracted full list of topics

Topic 1 Top Words:
Highest Prob: data, set, train, test, predict, can, will FREX: leakag, data, test, set, miss, extern, sampl Lift: memori, anonym, argu, codeisbookingcod, codeorigdestinationdistancecod, dplyr, eventidl Score: data, train, test, set, height, leakag, sampl
Topic 2 Top Words:
Highest Prob: user, hotel, event, cluster, book, citi, distanc FREX: hotel, event, cluster, book, citi, distanc, destin Lift: -interest, arc, arnold, asi, attende, blank, boston Score: hotel, event, user, distanc, cluster, book, citi
Topic 3 Top Words:
Highest Prob: use, file, tri, xgboost, can, script, get FREX: ram, memori, load, panda, return, gerhard, logist Lift: def, kwds, makeengineself, multimap, numpi, vik, xgboostclassifi Score: python, memori, ram, xgboost, codepr, file, run
Topic 4 Top Words:
Highest Prob: product, month, custom, new, account, june, predict FREX: product, month, custom, account, june, lag, buy Lift: deposit, -minut, -month, abcd, acquir, age-rang, ahor Score: product, custom, month, scroll, june, transpar, romanserifspan
Topic 5 Top Words:
Highest Prob: featur, use, model, time, can, song, user

FREX: song, split, listen, categor, layer, embed, lightgbm

Lift: bid, booster, chines, ci-dt, clang, cmake, co-ci

Score: song, featur, listen, embed, ensembl, nns, user-song

Topic 6 Top Words:

Highest Prob: score, submiss, result, get, model, probabl, tri

FREX: leaderboard, posit, submiss, metric, public, vote, xgb

Lift: ala, bob, constraint, epiplus, fwiw, gert, guerrero

Score: score, leak, valid, submiss, leaderboard, public, vote

Topic 7 Top Words:

Highest Prob: thank, competit, use, know, will, work, share

FREX: team, particip, idlespecul, spark, winner, form, sourc

Lift: abhijit, algo, bet, blog, cup, decent, fetch

Score: thank, competit, team, kaggl, share, particip, congrat

Top topic keywords

Highest Prob: Highest probability

FREX: A frequency measure accounts for exclusivity and frequency in

$$\text{FREX}_{k,v} = \left(\frac{w}{\text{ECDF}\left(\frac{\beta_{k,v}}{\sum_{j=1}^{K}\beta_{j,v}}\right)} + \frac{1-w}{\text{ECDF}(\beta_{k,v})}\right)-1.$$

Lift: The frequency of a word divided by its frequency in other topics

Score: The log frequency of a word in one topic divided by the log frequency of this word in other topics

Topic validation via text document

Topics 1, 3 and 5 are technical topics related to model training, data files, and feature variables. For instance, in *discussion 1* of **TABLE XI**, participants mention training and testing their model solutions using hold-out or cross-validation techniques, involving topic 1. This allows them to evaluate the accuracy of the system to compete for higher ranks. The codes used to split the dataset into training and testing are presented, to allow participants to select the model (topic 5). Obviously, not all discussions are technically focused; for example, in *discussion 2*, the main focus is to understand the hotel recommendation problem (topic 2). In *discussion 3*, participants have to decide from among a list of machine learning and statistical techniques which ones to use in their experiments and how these techniques would scale up to the large datasets. They also discuss how to use the 1.1 gigabyte of data resource (topic 3) and variables about events (topic 2). Therefore, in *discussion 3*, the focus is balanced between technical and business problems. Topic 4 covers a discussion on product recommendation. Buying habits and trends are important aspects to take into account when modeling the data. Topic 7 is related to the usage of the competition website for competitors to submit their work and seek support. Participants discuss their submission results, the usage of the website, and competition rules. In *discussion 4*, topics 4 and 7 are identified as the main topics.

TABLE XI
DISCUSSIONS

Discussion 1
“When I trained my lgbm model with the whole train data set, I found that when I got around 0.84 local AUC, my online performance would be the best. Anybody would share some idea about this?”
“The are two validation strategy. One is using the 'train test split ' to randomly split the train data into two size, such as the kernal1. And this validation strategy seems a good result. The other is use the orderly data, just as the author discuss in this kernal2. I have tried both the validation strategy, found that the first one performs good in result. I am very confused now. Any one could explain this?”
“I think your local cv maybe will overfit if you randomly split train and validation sets as the train/test split is based on time.”
Discussion 2
"Hello guys I have an ambiguity in producing submission result and it's that in each record in my result set I must have only one correct hotel_cluster or it could be more than one. Suppose that in my first record I predict 5 hotel_cluster. does one of them only true and if the answer is yes, is the position of the correct branch_cluster effect on the point in that record or not? thanks"
"About is_booking: is_booking = 1 if a given hotel was booked and is_booking = 0 if a given hotel was clicked (i.e. a user clicked a link to see hotel details on a hotel infosite page). This column is omitted from the holdout data because all events in the holdout data are bookings. About cnt: Basically, it's the # of clicks on a given hotel inforsite page in the context of a user session. A user session is defined with a 30 min of inactivity. It happens rarely that a user books the same hotel more than once in the same session, hence usually cnt = 1 if is_booking = 1. Intuitively, the higher the cnt the more a user is interested in a given hotel. Again, this column is omitted from the holdout data because all events in the holdout data are bookings. About d1-d149: This is a latent description of hotel reviews that are related to a given search destination. These columns correspond to different facets (e.g. beach, ski, etc.) and values are (log) probabilities that a customer would endorse a hotel in the destination for a specific facet. Adam"
"Is there an hotel id? Thanks, Asi"
"Hi, About cnt: does it mean that if cnt;1 & is_booking=0, the timestamp is the timestamp of the first click, and that all subsequent clicks are not logged ?"

TABLE XI
DISCUSSIONS CONTINUED

Discussion 3
"The contest was indeed exciting.. If anything I learnt the perils of overfitting in this contest. Our code is in a bit of a mess.. :) .. I will put it up along with a blog post after we clean it, but here is a summary of what we did. We used regression (random forest, grad boost regressor in scikit) to score each (user, event) pair. A target of 1 if interested and 0 if not. (Funnily everytime we tried to use the 'not interested' column my score decreased hence we ignored it). Being amateurs in programming and python, we didn't know how to handle the 3 million events, but we noticed only 30k odd events featured in any other data file, hence we pruned the rest of the events and made a 13MB file out of the 1.1GB file and only worked with that.. :) Also we didn't do any clustering, (user or event) we just put all the event details also in the feature vector for the (user, event) pair. The feature vector had three parts : The user part (containing age, sex, locale, no. of events attended etc.) , the event part (no. of attendees, word freq count etc) and the 'User-Event' part. The main components of User-Event part is detailed below: # and fraction of friends who attended the event; friendship with event creator; if event city is a substring of user location; No of `similar' events attended by the user; No. of similar events attended by the friends of the user; Time between event start and event seen by the user; After a cross validation and some careful weighting of the regressors, we managed 0.727 (3rd) by the time public leaderboard closed. In the last one week, we added more RFs/GBRs with different parameters and also added dolaameng's regression results to the already significant number of sub-learners. (Thanks to dolaameng for his code) Managed to get 0.707 and 6th place in the final result. Best, Harishgp."
"@Harishgp (Funnily everytime we tried to use the 'not interested' column my score decreased hence we ignored it). I was puzzled by this one too. Seem to remember doing a query and finding that after taking the funny business about time stamps into account, there weren't any not-interesteds in the remaining data. @Andrei I didn't manage to get anything out of user age and gender. I'm still wondering if (and how) that info can be used in some useful way.; FWIW, I managed to get something out of that by adding up the keyword vectors of all people of a particular gender in a locale's attendances and interests, and taking the cosine similarity with the event in question. Worried now about over fitting (lost 20 places in the final cut perhaps from this and another feature.) Also my GLM results showed older people were a little less likely to be interested."
"@Andrei. Excellent solution! your blog says that; I chose a Random Forest (again.. scikit-learn), because it was able to work with missing values.; I use sklearn, too. Could you illustrate how to work with missing values in sklearn? thanks in advance."

TABLE XI
DISCUSSIONS CONTINUED.

Discussion 4
"@_/@Y _/@|)|-|@\/ Hello, I am exploring this new field that is Recommendation. Till date I was working on Classification, Clustering, text Mining. Just want to know that anyone using Recommendation Model / Collaborative Filtering? If yes can anyone just explain me theoretically Hi, there's a public script (kernel) using CF. You might check that out. " “I would argue that while collaborative filtering works quite well on this data, it is not the best approach given the relatively low number of products. While it might be hard for Amazon to fit a classification method to each of its products and the data might be way to sparse, it is still possible here.” "@Maximilian Hahn Can you please elaborate following sentence - While it might be hard for Amazon to fit a classification method to each of its products and the data might be way to sparse I didnt (didn't) got that" "Maybe it is worth to start thinking about an example: Let's say you have an online shop selling 20 products. Now you have collected some data after selling say 10000 products total. If the probability of buying a product is the same for each product, this means you have roughly 500 purchases per product. You could now fit a logistic regression model for each product. This would mean to fit 20 logistic regression models based on 500 purchases each. This is a feasible task and the models will be comparable since the number of purchases for each product is the same. Each model then produces a probability for a new customer and a ranking of products can be made. But now suppose you have 1000000 products. You would have to fit 1000000 logistic regressions if you wish to take the same approach. Moreover, the distribution of purchases over the products might not be uniform. There are some products with less purchases than others, so the models for these will be worse in predictive quality. What does that mean for the ranking of the predicted probability for a new customer? Maybe you get my point if you think along these lines. "

Propensity score matching

We use R package "MatchIt" to enhance the econometric estimator. Specifically, propensity score matching is used to investigate the cases where the pre-test of parallel trends assumption is violated. The PSM performs pairing and outputs a subset of the data considering treatment variable and covariates of the observations.

The multiple period dataset is organized in such a way that covariates are coded into periods in weeks. We consider three main types of covariates – namely, *weekly submission performance*, *weekly number of messages*, and *the week number receiving the first message*.

Control units versus treated units ratio could have an impact on the matched subset sample size. With the original ratio in mind from the raw datasets, we test the ratio with varying values and the estimator results are consistent while keeping the matching ratio close to the original dataset.

REFERENCES

[1] L. Plummer, “This is how Netflix’s top-secret recommendation system works,” 2017. [Online]. Available: https://www.wired.co.uk/article/how-do-netflixs-algorithms-work-machine-learning-helps-to-predict-what-viewers-will-like. [Accessed: 07-Sep-2019].

[2] Netflix, “How Netflix’s Recommendations System Works,” 2019. [Online]. Available: https://help.netflix.com/en/node/100639. [Accessed: 07-Sep-2019].

[3] M. Lévesque and N. Joglekar, “Guest Editorial Resource, Routine, Reputation, or Regulation Shortages: Can Data-and Analytics-Driven Capabilities Inform Tech Entrepreneur Decisions,” *IEEE Trans. Eng. Manag.*, vol. 65, no. 4, pp. 537–544, 2018.

[4] M. Gorgoglione, U. Panniello, and A. Tuzhilin, “Recommendation strategies in personalization applications,” *Inf. Manag.*, 2019.

[5] R. Abdalkareem, E. Shihab, and J. Rilling, “What do developers use the crowd for? a study using stack overflow,” *IEEE Softw.*, vol. 34, no. 2, pp. 53–60, 2017.

[6] H. C. B. Lee, S. Ba, X. Li, and J. Stallaert, “Salience Bias in Crowdsourcing Contests,” *Inf. Syst. Res.*, vol. 29, no. 2, pp. 401–418, Mar. 2018, doi: 10.1287/isre.2018.0775.

[7] D. E. O’Leary, “An empirical analysis of information search and information sharing in crowdsourcing data analytic contests,” *Decis. Support Syst.*, vol. 120, pp. 1–13, 2019, doi: https://doi.org/10.1016/j.dss.2019.03.003.

[8] B. Callaway and P. H. C. Sant’Anna, “Difference-in-differences with multiple time periods,” *J. Econom.*, 2020.

[9] D. Renard and J. G. Davis, “Social interdependence on crowdsourcing platforms,” *J. Bus. Res.*, vol. 103, pp. 186–194, 2019, doi:

https://doi.org/10.1016/j.jbusres.2019.06.033.

[10] X. Wang, Y. Qiao, Y. Hou, S. Zhang, and X. Han, “Measuring Technology Complementarity Between Enterprises With an hLDA Topic Model,” *IEEE Trans. Eng. Manag.*, 2019.

[11] Z. Li, H. Tang, X. Xu, and Q. Chen, “Knowledge Topic-Structure Exploration for Online Innovative Knowledge Acquisition,” *IEEE Trans. Eng. Manag.*, 2019.

[12] H. Chen, X. Wang, S. Pan, and F. Xiong, “Identify topic relations in scientific literature using topic modeling,” *IEEE Trans. Eng. Manag.*, 2019.

[13] S. Aral and M. Van Alstyne, “The diversity-bandwidth trade-off,” *Am. J. Sociol.*, vol. 117, no. 1, pp. 90–171, 2011.

[14] N. C. Lindstedt, “Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005–2017,” *Soc. Curr.*, p. 2329496519846505, 2019.

[15] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[16] T. Bogers and A. van den Bosch, “Fusing Recommendations for Social Bookmarking Web Sites,” *Int. J. Electron. Commer.*, vol. 15, no. 3, pp. 31–72, Apr. 2011, doi: 10.2753/JEC1086-4415150303.

[17] L. Liu, N. Mehandjiev, and D.-L. Xu, “Context Similarity Metric for Multidimensional Service Recommendation,” *Int. J. Electron. Commer.*, vol. 18, no. 1, pp. 73–104, Oct. 2013, doi: 10.2753/JEC1086-4415180103.

[18] U. Gretzel and D. R. Fesenmaier, “Persuasion in Recommender Systems,” *Int. J. Electron. Commer.*, vol. 11, no. 2, pp. 81–100, 2006.

[19] D. Baum and M. Spann, “The Interplay Between Online Consumer Reviews and Recommender Systems: An Experimental Analysis,” *Int. J. Electron. Commer.*, vol. 19, no. 1, pp. 129–162, Oct. 2014, doi: 10.2753/JEC1086-4415190104.

[20] J. Ren, J. Long, and Z. Xu, “Financial news recommendation based on graph embeddings,” *Decis. Support Syst.*, vol. 125, p. 113115, 2019, doi: https://doi.org/10.1016/j.dss.2019.113115.

[21] M. Zihayat, A. Ayanso, X. Zhao, H. Davoudi, and A. An, “A utility-based news recommendation system,” *Decis. Support Syst.*, vol. 117, pp. 14–27, 2019, doi: https://doi.org/10.1016/j.dss.2018.12.001.

[22] Y. Guan, Q. Wei, and G. Chen, “Deep learning based personalized recommendation with multi-view information integration,” *Decis. Support Syst.*, vol. 118, pp. 58–69, 2019, doi: https://doi.org/10.1016/j.dss.2019.01.003.

[23] M. Ferrari Dacrema, P. Cremonesi, and D. Jannach, “Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches,” *Thirteenth ACM Conference on Recommender Systems (RecSys ’19)*. ACM, New York, NY, USA, Copenhagen, Denmark, p. 10, 2019.

[24] F. M. Harper and J. A. Konstan, “The movielens datasets: History and context,” *Acm Trans. Interact. Intell. Syst.*, vol. 5, no. 4, p. 19, 2016.

[25] C. Tauchert, P. Buxmann, and J. Lambinus, “Crowdsourcing Data Science: A Qualitative Analysis of Organizations’ Usage of Kaggle Competitions,” in *Proceedings of the 53rd Hawaii international conference on system sciences*, 2020.

[26] H. Feng, J. Tian, H. J. Wang, and M. Li, “Personalized recommendations based on time-weighted overlapping community detection,” *Inf. Manag.*, vol. 52, no. 7, pp. 789–800,

2015, doi: https://doi.org/10.1016/j.im.2015.02.004.

[27] D. L.R. and N. Pervin, “Towards generating scalable personalized recommendations: Integrating social trust, social bias, and geo-spatial clustering,” *Decis. Support Syst.*, vol. 122, p. 113066, 2019, doi: https://doi.org/10.1016/j.dss.2019.05.006.

[28] T. Yu, J. Guo, W. Li, H. J. Wang, and L. Fan, “Recommendation with diversity: An adaptive trust-aware model,” *Decis. Support Syst.*, vol. 123, p. 113073, 2019, doi: https://doi.org/10.1016/j.dss.2019.113073.

[29] B. Xiao and I. Benbasat, “An empirical examination of the influence of biased personalized product recommendations on consumers’ decision making outcomes,” *Decis. Support Syst.*, vol. 110, pp. 46–57, 2018, doi: https://doi.org/10.1016/j.dss.2018.03.005.

[30] D. S. Chatterjee, “Explaining customer ratings and recommendations by combining qualitative and quantitative user generated contents,” *Decis. Support Syst.*, vol. 119, pp. 14–22, 2019, doi: https://doi.org/10.1016/j.dss.2019.02.008.

[31] M. Yang and P. Jiang, “Improved Bayesian Causal Map Approach for Community-Based Product Design Project Feasibility Analysis,” *IEEE Trans. Eng. Manag.*, 2019.

[32] S. Liu, F. Xia, B. Gao, G. Jiang, and J. Zhang, “Hybrid Influences of Social Subsystem and Technical Subsystem Risks in the Crowdsourcing Marketplace,” *IEEE Trans. Eng. Manag.*, pp. 1–15, 2019, doi: 10.1109/TEM.2019.2902446.

[33] H. Zhang, Z. Wang, S. Chen, and C. Guo, “Product recommendation in online social networking communities: An empirical study of antecedents and a mediator,” *Inf. Manag.*, vol. 56, no. 2, pp. 185–195, 2019, doi: https://doi.org/10.1016/j.im.2018.05.001.

[34] G. Parker and M. Van Alstyne, “Innovation, openness, and platform control,” *Manage. Sci.*, vol. 64, no. 7, pp. 3015–3032, 2018.

[35] H. J. Ye and A. Kankanhalli, “Solvers’ participation in crowdsourcing platforms: Examining the impacts of trust, and benefit and cost factors,” *J. Strateg. Inf. Syst.*, vol. 26, no. 2, pp. 101–117, 2017.

[36] M. W. Van Alstyne, A. Di Fiore, and S. Schneider, “4 mistakes that kill crowdsourcing efforts,” *Harv. Bus. Rev.*, 2017.

[37] M. W. Van Alstyne, G. G. Parker, and S. P. Choudary, “Pipelines, platforms, and the new rules of strategy,” *Harv. Bus. Rev.*, vol. 94, no. 4, pp. 54–62, 2016.

[38] P. Song, L. Xue, A. Rai, and C. Zhang, “The ecosystem of software platform: A study of asymmetric cross-side network effects and platform governance,” *Mis Q.*, vol. 42, no. 1, pp. 121–142, 2018.

[39] M. F. Niculescu, D. J. Wu, and L. Xu, “Strategic intellectual property sharing: Competition on an open technology platform under network effects,” *Inf. Syst. Res.*, vol. 29, no. 2, pp. 498–519, 2018.

[40] A. Sundararajan, F. Provost, G. Oestreicher-Singer, and S. Aral, “Research commentary—information in digital, economic, and social networks,” *Inf. Syst. Res.*, vol. 24, no. 4, pp. 883–905, 2013.

[41] S. M. C. Loureiro, J. Guerreiro, S. Eloy, D. Langaro, and P. Panchapakesan, “Understanding the use of Virtual Reality in Marketing: A text mining-based review,” *J. Bus. Res.*, vol. 100, pp. 514–530, 2019, doi: https://doi.org/10.1016/j.jbusres.2018.10.055.

[42] I. Park, B. Yoon, S. Kim, and H. Seol, “Technological Opportunities Discovery for

Safety Through Topic Modeling and Opinion Mining in the Fourth Industrial Revolution: The Case of Artificial Intelligence," *IEEE Trans. Eng. Manag.*, 2019.

[43] M. García Lozano, J. Schreiber, and J. Brynielsson, "Tracking geographical locations using a geo-aware topic model for analyzing social media data," *Decis. Support Syst.*, vol. 99, pp. 18–29, 2017, doi: https://doi.org/10.1016/j.dss.2017.05.006.

[44] R. Gruss, A. S. Abrahams, W. Fan, and G. A. Wang, "By the numbers: The magic of numerical intelligence in text analytic systems," *Decis. Support Syst.*, vol. 113, pp. 86–98, 2018, doi: https://doi.org/10.1016/j.dss.2018.07.004.

[45] M. E. Roberts, B. M. Stewart, and D. Tingley, "stm: R package for structural topic models," *J. Stat. Softw.*, vol. 10, no. 2, pp. 1–40, 2014.

[46] M. E. Roberts, B. M. Stewart, D. Tingley, and E. M. Airoldi, "The structural topic model and applied social science," 2013.

[47] G. M. Lee, S. He, J. Lee, and A. B. Whinston, "Matching mobile applications for cross-promotion," *Inf. Syst. Res.*, vol. 31, no. 3, pp. 865–891, 2020.

[48] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis. Support Syst.*, vol. 115, pp. 24–35, 2018, doi: https://doi.org/10.1016/j.dss.2018.09.002.

[49] N. Kozodoi, S. Lessmann, K. Papakonstantinou, Y. Gatsoulis, and B. Baesens, "A multi-objective approach for profit-driven feature selection in credit scoring," *Decis. Support Syst.*, vol. 120, pp. 106–117, 2019, doi: https://doi.org/10.1016/j.dss.2019.03.011.

[50] A. Khalemsky and R. Gelbard, "A dynamic classification unit for online segmentation of big data via small data buffers," *Decis. Support Syst.*, vol. 128, p. 113157, 2020, doi: https://doi.org/10.1016/j.dss.2019.113157.

[51] N. Pröllochs and S. Feuerriegel, “Business analytics for strategic management: Identifying and assessing corporate challenges via topic modeling,” *Inf. Manag.*, p. 103070, 2018, doi: https://doi.org/10.1016/j.im.2018.05.003.

[52] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, “Optimizing semantic coherence in topic models,” in *Proceedings of the conference on empirical methods in natural language processing*, 2011, pp. 262–272.

[53] J. M. Bischof and E. M. Airoldi, *Summarizing topical content with word frequency and exclusivity*. Edinburgh, Scotland: Omnipress, 2012.

[54] E. M. Airoldi and J. M. Bischof, “Improving and evaluating topic models and other models of text,” *J. Am. Stat. Assoc.*, vol. 111, no. 516, pp. 1381–1403, 2016.

[55] K. D. Kuhn, “Using structural topic modeling to identify latent topics and trends in aviation incident reports,” *Transp. Res. Part C Emerg. Technol.*, vol. 87, pp. 105–122, 2018.

[56] I. Dissanayake, J. Zhang, M. Yasar, and S. P. Nerur, “Strategic effort allocation in online innovation tournaments,” *Inf. Manag.*, vol. 55, no. 3, pp. 396–406, 2018.

[57] J. Foerderer, N. Lueker, and A. Heinzl, “And the Winner Is…? The Desirable and Undesirable Effects of Platform Awards,” *Inf. Syst. Res.*, 2021.

[58] E. A. Stuart, G. King, K. Imai, and D. Ho, “MatchIt: nonparametric preprocessing for parametric causal inference,” *J. Stat. Softw.*, 2011.

[59] J. Elliott, “Benchmarks Provided by Automated ML Tools!,” 2019. [Online]. Available: https://www.kaggle.com/c/ieee-fraud-detection/discussion/99983. [Accessed: 07-Dec-2019].

[60] B. Vasilescu, V. Filkov, and A. Serebrenik, “Stackoverflow and github: Associations

between software development and crowdsourced knowledge," in *2013 International Conference on Social Computing*, 2013, pp. 188–195.

[61] S. Wang, D. Lo, and L. Jiang, "An empirical study on developer interactions in StackOverflow," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, 2013, pp. 1019–1024.

[62] J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang, "Understanding the limiting factors of topic modeling via posterior contraction analysis," in *International Conference on Machine Learning*, 2014, pp. 190–198.

[63] C. Ponsiglione, L. Cannavacciuolo, S. Primario, I. Quinto, and G. Zollo, "The ambiguity of natural language as resource for organizational design: A computational analysis," *J. Bus. Res.*, 2019, doi: https://doi.org/10.1016/j.jbusres.2019.11.052.