Technological Forecasting & Social Change 198 (2024) 122936

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

journal homepage: www.elsevier.com/locate/techfore

The impact of forum content on data science open innovation performance: A system dynamics-based causal machine learning approach

Libo Li *, Huan Yu, Martin Kunc

Southampton Business School, University of Southampton, 12 University Rd, Southampton SO17 1BJ, United Kingdom

ARTICLE INFO

Keywords:
System dynamics
Open innovation
Topic modelling
Causal machine learning

ABSTRACT

Open innovation in data science generally takes the form of public competitions where teams exchange messages and solutions by competing and collaborating simultaneously. Team behaviours are widely heterogeneous in terms of the performance of their solutions and the participation in knowledge creation. We present a novel research framework for open innovation by integrating system dynamics and structural topic modelling to extract open factors and adopting a machine learning-based difference-in-differences estimator to understand the impact of team behaviour on their performance using data from Kaggle's competition. Our results identify four team behaviour categories—active, learner, lurker, and passive— in data science open innovation competitions which depend on the performance of their solutions and actions related to posting and reading messages in the forum. Furthermore, the activities of model evaluation, community support, and business understanding are the top three most positive and significant factors affecting team performance. Our research contributes to the literature by highlighting the value of forum feedback and exploring the data science activities in the forum discussion, in relation to innovation performance, to enrich the empirical understanding of open innovation. Research implications for researchers and practitioners participating in, organising, and supporting data science open innovation activities are provided.

1. Introduction

Open Innovation and crowdsourcing are two of the most popular topics in the field of innovation management and are attracting significant interest. The growing acceptance of external cooperation in data sciences facilitates the open innovation as a new paradigm. Crowdsourcing, as a mode of open innovation, attracts a large crowd of people to submit innovative data science ideas in response to calls to solve complex data analytics tasks. Crowdsourcing can be collaborative, competitive, or coopetitive. Being open to external participants who can bring new ideas and solutions (Afuah and Tucci, 2012), collective intelligence generated from crowdsourcing coopetition brings competitive advantage for the companies that adopt the open innovation strategy. User-generated content (UGC) is a main element of collective intelligence. It has been shown that the quality of UGC affects the performance of innovation in online innovation communities from the perspective of the hosting organisation (Ye et al., 2012). Compared with the entire organisation innovation, the performance of individual participants and teams is also influenced by the sharing UGC. Individual teams in crowdsourcing coopetition have different motives and incentives to participate; therefore, they generate diverse content. Four motives—learning, direct compensation, self-marketing, and social motives—have been explored to explain the motivation of participants in open innovation projects (Leimeister et al., 2009). In addition, the performance of participating teams is associated with the creativity that they employ in the solution which could be inspired by the sharing UGC. Researchers have traditionally employed mixed research methods (e.g., experiments and surveys) to explore the underlying factors but these may not be suitable for understanding dynamic interactions and drivers between teams during the knowledge creation and sharing processes (Wang et al., 2019). Thus, the dynamic effect of UGC on team performance remains under-investigated.

One of the main platforms employed for crowdsourcing coopetition in data science is Kaggle, which was established in 2010 and hosts competitive crowdsourcing activities (Athanasopoulos and Hyndman, 2011). Kaggle has a large community of data scientists from over 100 countries (Bojer and Meldgaard, 2021) and attracts, on average, >1000 teams per month with some teams being involved almost full time on the

* Corresponding author.
E-mail addresses: Libo.li@soton.ac.uk (L. Li), huan.yu@soton.ac.uk (H. Yu), m.h.kunc@soton.ac.uk (M. Kunc).

https://doi.org/10.1016/j.techfore.2023.122936
Received 15 December 2022; Received in revised form 15 October 2023; Accepted 17 October 2023

platform (Tauchert et al., 2020). Kaggle has a discussion session that allows the participants to share their UGC which consists of ideas, code, and solutions on the forums and online notebook (Bojer and Meldgaard, 2021). Based on the forum, a virtual community has been constructed to facilitate knowledge generation, sharing, and exchange. Virtual communities have multiple effects; some positive, such as ease of interacting online, and some negative, such as being influenced by other participants through their messages (Li et al., 2022). The availability of rich UGC from the Kaggle platform allows us to analyse the heterogeneity of team behaviour and explore the impact of UGC on team performance. This motivates us to address the following research questions:

- How can different team behaviours be identified during open innovation in crowdsourcing coopetition?
- What factors from UGC drive team performance over time?

To answer the two research questions, we propose a novel open innovation research framework that focuses on team behaviour in data science open innovation competitions. The performance of open innovation depends on the competing teams' solutions and the actions related to posting and reading messages in the forum that involve sharing and acquiring knowledge. The research framework combines the system dynamics (SD) modelling and causal machine learning (ML) methodology to understand team behaviour and open innovation performance. The research framework has been applied to extract factors that capture the dynamic and volatile behaviour patterns in a real-life open innovation context. Using data from a popular Kaggle competition, SD models reflect researchers' diverse theoretical understanding of the endogenous feedback behaviour responsible for the performance observed. ML methodology, such as topic modelling, has been applied to understand the evolution of UGC for open innovation (Saura et al., 2023). However, to the best of our knowledge, there is limited work in combining SD modelling and ML methodology to provide a more holistic view of how innovation processes and UGC dynamically affect the team performance. Hence, we identify this as a research gap.

The paper makes both theoretical and empirical contributions to data science open innovation. We develop a causal ML research framework using the SD model to understand complex coopetition behaviour in data science open innovation. Our novel approach also makes use of structural topic modelling (STM) results to assess the impact of topic content from forum messages on team performance. Our work departs from prior theoretical work which mainly focused on perceived factors within a time-invariant context (Garcia Martinez, 2015; Garcia Martinez, 2017; Shao et al., 2012) and investigates the role of feedback on learning and knowledge sharing in open innovation. We assess the team performance and knowledge acquisition behaviour and identify four team categories—active, learner, lurker, and passive—considering the role of feedback on learning and knowledge sharing. Our research results show a significant positive impact of UGC on team performance. Our research also provides a statistical validation of the casual relationship using the text data in a longitudinal setting.

The rest of the paper is organised as follows. Section 2 reviews the recent relevant literature and Section 3 describes the data that we collected from a specific Kaggle competition. In Section 4, the research framework, which combines the analytical methodology of STM, SD modelling, and the Double/debiased Machine Learning for difference in differences framework (DMLDiD) estimator, is proposed to extract factors from team behaviour and the knowledge shared in the virtual community as well as the casual relationship between these factors and team performance. Section 5 elaborates on the data analysis and the results of its application followed by a discussion of research and business implications in Section 6. Finally, a conclusion and directions for further research are presented.

2. Literature review

Crowdsourcing collective intelligence through coopetition plays a vital role in data science open innovation. Our research is related to two broad streams of literature: crowdsourcing coopetition in open innovation and collective intelligence generation through team knowledge management. We first review the relevant literature of these two streams in Sections 2.1 and 2.2. In addition, the analytical methodology of SD which has been used to extract the team behaviour processes in innovation is reviewed in Section 2.3.

2.1. Crowdsourcing coopetition in open innovation

Crowdsourcing coopetition is one of the most important sources for open innovation. Kaggle, the most popular crowdsourcing platform, hosts data science competitions with cash prizes sponsored by companies and organisations. Tauchert et al. (2020) identified several factors within three categories—platform-related, organisation-related, and outcome-related—that influence an organisation's perceived success when hosting a data science competition, such as building a community which attracts the best data scientists to participate in and promote discussions on the forum as a way of interactive learning. In addition, the factor of competition intensity is associated with significantly decreased participants (Shao et al., 2012).

The majority of the crowdsourcing platforms provide a function of a discussion forum which creates a virtual community of practice during the competition. Participants can learn from each other through effective knowledge collaboration; for example, sharing through messages and transforming them into new knowledge, even without a monetary reward (Faraj et al., 2011). It is suggested that discussion forums are valuable sources of information to refine ideas, but the impact of comments depends on their quality (Javadi Khasraghi and Hirschheim, 2022). Some platforms also use a voting system to identify the quality of a comment, which allows teams to have some perceptions of the usefulness of their solutions. However, the connections between the competing teams can also be weak since they can join and abandon competitions easily (Faraj et al., 2011). One study (Otto and Simon, 2008) found that online communities need structural control to maintain the platform's attractiveness, credibility, and content value.

Apart from monetary rewards, crowdsourcing platforms use gamification and social dominance-based faultlines (e.g., leader board rank) to motivate and help teams to improve the performance in the competitions (Cao et al., 2022). Rank also helps team members to understand the quality of their solution and further improve it before submitting a revised version. Another measure of the quality of a team's solution is the visibility of their kernels, which are entities used to share part or all the code on Kaggle, in terms of votes. Using the above-mentioned measurements, researchers (Javadi Khasraghi and Hirschheim, 2022) found a significant effect on team performance from cross-team collaboration.

Prior research in open innovation mainly focuses on conceptualising these perceived factors of the competition, highlighting competitor experience and linking it to creativity outcomes. In addition, perceived task variety, complexity, specialisation, and autonomy in the competition have been found to be correlated with engagement and motivation to compete (Amabile et al., 1994; Morgeson and Humphrey, 2006; Shao et al., 2012; Garcia Martinez, 2015; Garcia Martinez, 2017). User interactions, such as learning and knowledge sharing from online communities, have been identified to have an impact on open innovation, such as team collaboration and innovation performance (Faraj et al., 2011; Jin et al., 2021; Javadi Khasraghi and Hirschheim, 2022). However, the use of forum messages—which are time-variant and of a publicly open nature— as communication and feedback for competitors has not been fully investigated (Javadi Khasraghi and Hirschheim, 2022). To investigate the impact of such feedback on team performance via incremental learning and knowledge sharing in open innovation

other than perceived factors, we propose a novel theoretical research framework in which the specific UGC data collected from the discussion forum are analysed to understand the participants' dynamic behaviour and performance. The techniques of topic modelling and difference-in-differences estimator (DiD) are reviewed below.

To extract the valuable patterns from the discussion forum, topic modelling in text mining has been widely adopted to automatically summarise the topics (e.g., activities) using topic assignments and topic keywords. Prior research explored the application of topic modelling in solar cell technology, healthcare, and other contexts (Erzurumlu and Pachamanova, 2020; Ma et al., 2021; Xu et al., 2021; Zhu and Cunningham, 2022) from the classical probabilistic topic models, such as Latent Dirichlet Allocation (Blei et al., 2003). However, the STM is becoming more prevalent in empirical research (Gao et al., 2022; Kumar and Srivastava, 2022; Rose et al., 2022). In particular, STM discovers topics through not only the text document itself but also from the metadata (defined as attributes) associated with the documents. This is particularly favourable for social science research because the outputs of STM can be used to conduct hypothesis testing about these relationships between topics and the other types of data (Roberts et al., 2019). For instance, prior work makes use of STM to understand feedback and challenges from online audiences (Doldor et al., 2019; Tonidandel et al., 2022).

To assess the relationships between the activities and team performance, the DiD, which is widely used in economics and quasi-experiments, estimates treatment effects on observational longitudinal data (Angrist and Pischke, 2008). Along with many research efforts in the causal inference literature (Abadie, 2005; Pearl, 2019; Schumann et al., 2021), recent works on DMLDiD show that embedding ML techniques into the DiD estimator advances the estimator of causal inferences in the presence of high-dimensional cases (Chernozhukov et al., 2018; Chang, 2020). This is specifically relevant to this study when working with high-dimension data involving text rather than traditional longitudinal/panel data.

In addition to participation, other factors, the motivation of team behaviour, such as the effectiveness of knowledge creation processes as well as the motivational factors driving teams to share knowledge with other teams (Roberts et al., 2006), are reviewed in the next sub-section.

2.2. Team knowledge management processes

Crowdsourcing collective intelligence, which can be generated through team knowledge management processes, is another potential factor which affects the performance. During competitions, team knowledge management comprises two main processes: knowledge generation and knowledge sharing. In terms of knowledge generation, one factor to consider is the motivation driving the behaviour of teams in virtual communities. There is a positive relationship between intrinsic motivation, e.g., need for achievement[1] and need for affiliation,[2] and knowledge collaboration (Garcia Martinez, 2017). While need for achievement increases initially and stabilises, the need for affiliation grows strongly as the community grows. In terms of extrinsic motivation, knowledge collaboration behaviour is driven by reciprocity and use value (e.g., value of the information). The information value increases at the beginning of the community but stabilises over time. Finally, motivation factors based on the community, such as sense of belonging and sense of satisfaction, grow over time as its members feel integrated into the community (Wang et al., 2019). Scholarly work (Garcia Martinez, 2017) found that problem solving, among other factors, had an impact on the intrinsic motivation of participants in Kaggle competitions. The challenge of the problem with respect to the knowledge of the participants is another key motivation. In particular, the high level of autonomy that they hold to define their own methodologies also supports their motivation during Kaggle competitions. Finally, participants' positive attitudes towards knowledge contribution generate better submissions and participation in competitions.

On the other hand, knowledge sharing does not improve everybody's performance because it depends on the quality, volume, and generativity of the knowledge exchange. Knowledge quality and volume are two areas covered extensively in previous research related to online communities. Generativity is the capability of shared knowledge to further develop into additional knowledge through a derivation process—i.e., derivative knowledge (Jin et al., 2021). It is also proposed that the ability for teams to represent knowledge affects knowledge transfer (Shi et al., 2022). Simultaneously, the receiving team needs to have the willingness and absorptive capacity to use the shared knowledge.

2.3. SD in innovation

Open innovation processes are strongly driven by feedback processes occurring between teams sharing ideas and solutions while increasing their stocks of knowledge over time through learning from other teams. Therefore, SD models, which portray information feedback processes and stock accumulation, are suitable to study the behaviour of teams in open innovation. For example, Wu and Gong (2019) developed a model of open innovation communities drawn from three areas: governance mechanisms, knowledge management, and community user behaviour. Governance mechanisms consist of the arrangements made by organisations to control the community and invest in resources to transform the interactions in the community into organisational knowledge. Knowledge management involves the recognition, acquisition, transformation, and application of knowledge through the coordination of governance mechanisms at organisational and technological levels. Community user behaviour involves processes of innovation and interaction between users and the organisation. Innovation performance, which depends on the knowledge management process, attracts users who, through their interactive and innovative behaviours, generate new innovations, thus creating a positive feedback loop. The research model considered users' online posts, views, and comments as a source of knowledge for the company. These are then transformed into innovation performance, which is measured through patents by using knowledge management-related capabilities.

SD has also been used to study virtual communities. A previous study by Diker (2004) evaluated membership dynamics in an open online system. Interestingly, Diker found that the most effective process occurs when experienced members support new members through coaching and revising/editing their work. Otto and Simon (2008) studied how to achieve effective online community networks and concluded that the sustainability of the network needs constant monitoring and rules and regulations. Mao et al. (2007) investigated the dynamic impact of the motivation mechanisms on an online collaborative system. They observed that sustainable communities need encouragement to share knowledge of good quality, so the use of ratings is a valid approach to measure the quality of the knowledge and the intensity of contribution. Other research studied the dynamic interactions between knowledge collaboration motivations and community behaviour in virtual communities of practice (Wang et al., 2019). Using data from Wikipedia, Wang et al. (2019) evaluated intrinsic, extrinsic, and community drivers of knowledge collaborations.

In summary, SD can model innovation processes at diverse levels of analysis, industry, product, organisation, and process (Kunc, 2012) due to its strengths in identifying causal mechanisms underpinned by information feedback processes. This current study contributes to SD models portraying innovation processes by adding a new level of analysis, the team level, which is an intermediate level between process and organisation.

[1] Need for achievement is the aspiration to become better by improving personal skills when interacting with others (McClelland et al., 1953).

[2] Need for affiliation is the need to generate personal and social relationships within a community (McClelland et al., 1953).

3. Data

To test our research model, we employed a longitudinal dataset from teams that participated in *The Home Credit Default Risk* competition on the Kaggle platform across a period of three months. Kaggle is one of the most popular platforms employed for crowdsourcing coopetition in data science. It hosts various real-life data science challenges, which are widely adopted by companies. The data collection from Kaggle follows the same practice of the existing open innovation literature (Garcia Martinez, 2015; Tauchert et al., 2020; Javadi Khasraghi and Hirschheim, 2022). The details of this competition are published on the Kaggle website at https://www.kaggle.com/c/home-credit-default-risk. The home credit default risk competition is the ideal entity to use for this study because it is one of the most popular Kaggle competitions with a reward of US$70,000.

We combine several sources of data to form a rich dataset. First, we collect team performance scores from the leaderboard as the outcome of the dependent variable. The competition was held from 17 May 2018 to 29 August 2018. During the competition, any team could submit their solutions at any time within the competition timeline. Once submitted, the solution is evaluated by the sponsor using an accuracy metric: the area under the ROC curve (AUROC). The rank of the performance can be observed in the relevant section of the leaderboard. There are two scores—public score and private score. The public score comes from the public leaderboard where only a proportion of the test results are reported while the private score is tested on the full test data which are used for the final ranking.

A total of 132,097 submitted solutions from 7198 teams are collected. The average performance of AUROC is 0.7604. Fig. 1 shows the daily average of team performance (left) and the daily total number of submissions (right). During the competition, a trajectory of slight improvement in average performance over time can be observed. In addition, the number of submissions holds steady over time apart from two surges in the middle and before the deadline. The difference of the final score between the top 25 teams and the winner is not high, as Fig. 2 shows, suggesting a fierce competition.

Moreover, teams were able to communicate with each other and with the organiser by sending messages or sharing solutions through the forum. A forum contains multiple discussions or threads, and each discussion contains several posts which are individual messages written by team members. The left part of Fig. 3 shows that the number of created posts begins to increase in the first half of the competition, falls slightly in the middle, and then grows dramatically as the deadline approaches. The great majority of the posts are sent by a few teams shown on the right of Fig. 3. The observed pattern is consistent with the Pareto principle in general and quite common in online communities (Johnson et al., 2014). In the recent M5 competition on Kaggle (Li et al., 2022), it is also observed that a high concentration of posts was generated by a few participants. A few teams acted as the main information producers through posting and answering in the competition while the majority of the participants in Kaggle lurked around the competition, studying the data and learning analysis methods (Hayashi et al., 2021). Particularly for the top teams, an average of 10.3 solutions per team were submitted and 40.5 messages were generated among the top 25 teams.

In terms of the observed pattern, most participating teams are idle with a low number of postings and submissions, so we include only the top 1000 teams in the private ranking, without loss of generality. This is because the lower rank teams (e.g., private rank >1500) have a very limited number of posts, with <15 % of the posts. In addition, teams in the lower rank do not generate competitive solutions. The selection criteria of the top teams are also consistent with the prior literature in open innovation. For example, the top 25 solutions and around 300 teams have been selected in the study (Garcia Martinez, 2017; Bojer and Meldgaard, 2021). This is also in line with the prior finding that, on average, there are around 1000 teams participating in a Kaggle competition (Tauchert et al., 2020). To further check the robustness of our results, we compare the performance by varying sample sizes of 1000 ± 250 teams. Table A5 shows that the results are consistent.

4. Methodology

To understand how UGC in the forum affects the team performance in the online competition, we first employ STM to extract the "topics" that occur in a collection of posts in a virtual discussion forum. Next, we construct a SD simulation model to produce the factors of team behaviour throughout the entire period of the online competition. Further, we implement a DMLDiD estimator to evaluate the effect of topic content on team performance in an online competition controlled by team behaviour factors. The related research methodologies are described below.

4.1. STM

STM is an extension of standard topic modelling specifically designed to support social science research. Topic modelling is a type of

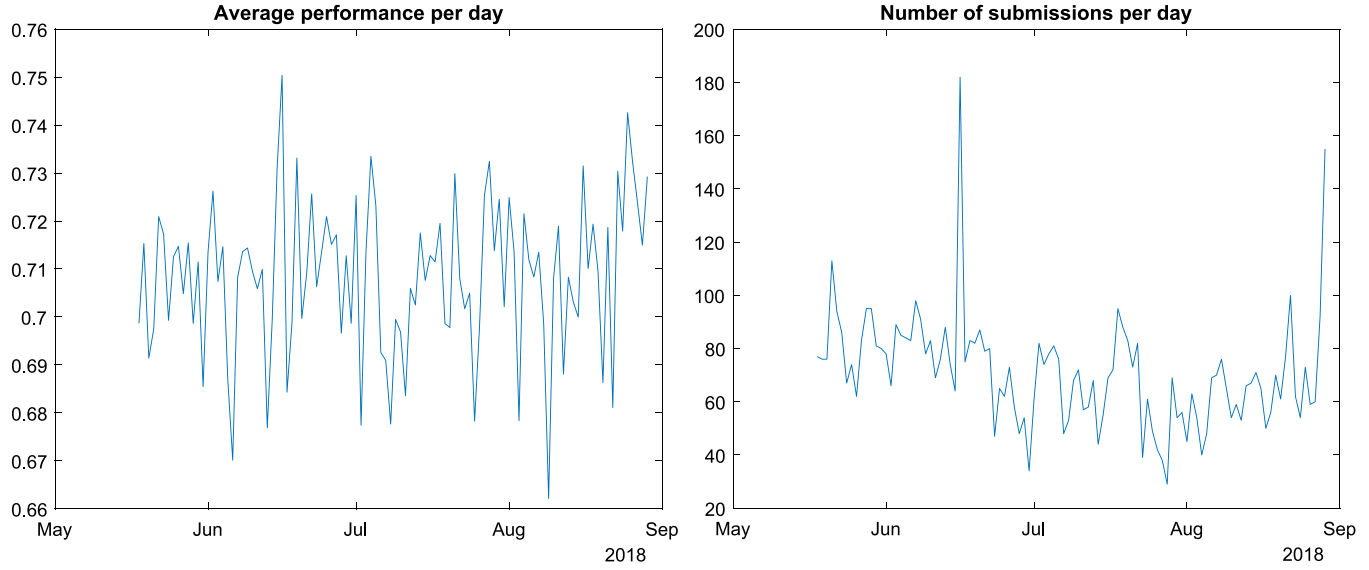


Fig. 1. Average performance (AUROC) per day (left) and number of submissions per day (right).

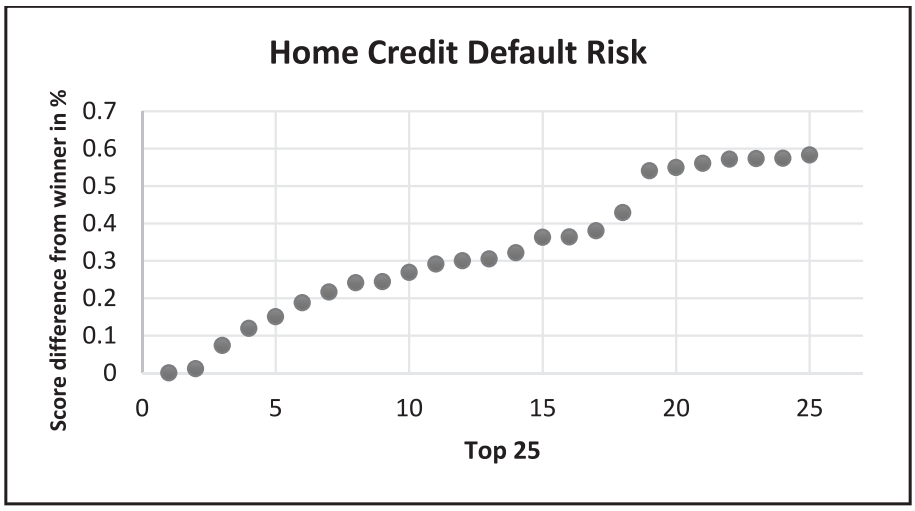


Fig. 2. The difference of final solution performance between the winner and the top 25 teams.

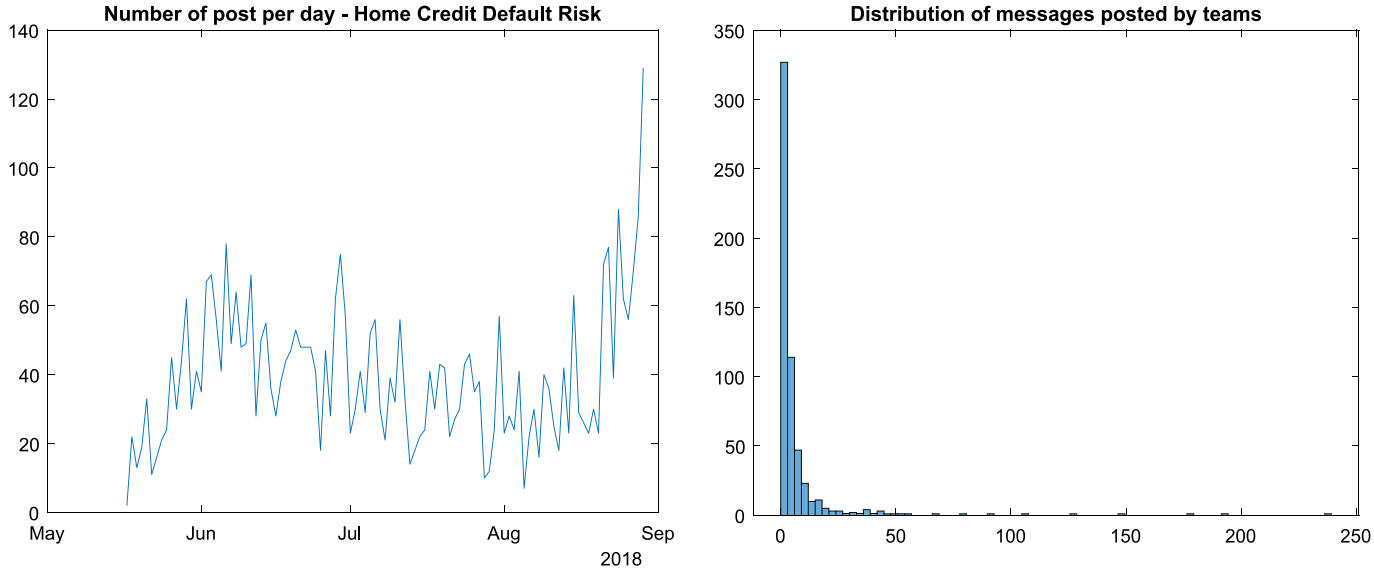


Fig. 3. Number of posts per day (left) and distribution of messages posted by teams (right).

unsupervised ML model which aims to solve two tasks (see Fig. 4): (i) topic prevalence to mine topics from multiple text data and (ii) topic content to figure out the coverage probability of each document for each topic.

To incorporate the metadata which affect both topic prevalence and topic content in the corpus structure, STM relaxes the restrictive assumptions of independent stationary topics in the latent Dirichlet allocation model and specifies the priors as generalised linear models through message-level covariates. The plate diagram for the STM applied in this study is shown in Fig. A3 in the Appendix.

The basic premise of topic modelling is to model messages as a distribution of topics (topic prevalence) and topics as a distribution of words (topic content). Formally, we define a word as the basic unit of data from a vocabulary, a post is a sequence of words, and a corpus is a collection of messages. STM is a generative probabilistic model of a corpus, and the generative process for each message in the forum can be found in Fig. A4 in the Appendix (Roberts et al., 2019). Specifically, step 1 and 2 are the generative process of topic prevalence and topic content, respectively.

We use the same post metadata which consist of three covariates of team ID, date duration, and thread ID to generate both topic prevalence and topic content. The prior of topic prevalence is sampled from a logistic-normal distribution and the prior of topic content is the combination of three effects of topic covariates which consist of team ID, date duration, thread ID, and the topic–covariate interaction deviated from a baseline word probability which represents the log-transformed rate of any given word across the corpus. The regularising priors for the generalised linear model coefficients of topic prevalence and topic content are sampled from the Normal-Gamma prior and the Gamma-Lasso prior, respectively. The posterior inference of STM could be estimated through a fast variant of the semi-collapsed nonconjugate variational expectation–maximisation algorithm (Blei et al., 2003).

4.2. SD model

The SD model represents an endogenous perspective on the behaviour of teams based on information feedback loops related to the performance of their submissions and knowledge stock accumulation over time, as discussed in the literature review. In more detail, there are two stocks: current knowledge team and message processed by team. The current knowledge stock grows through the acquisition of knowledge from each message observed. The more knowledge there is, the stronger

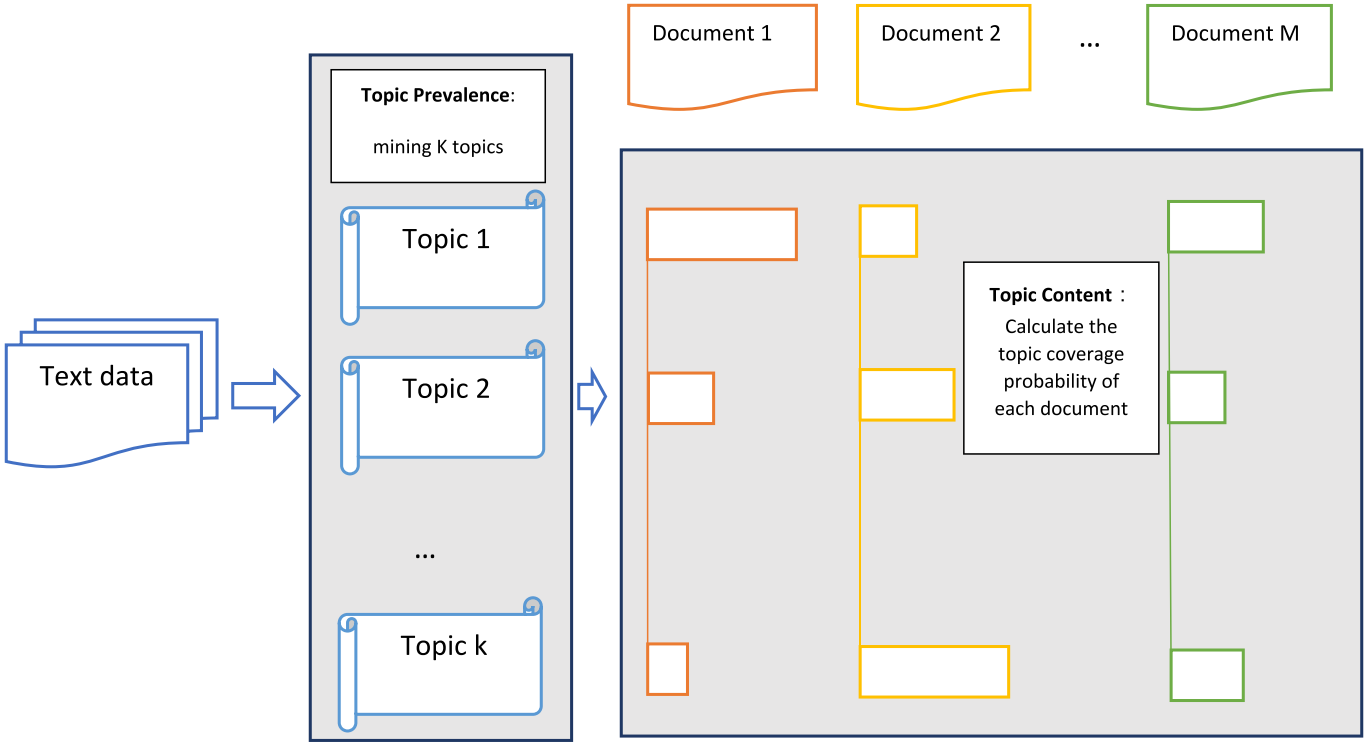


Fig. 4. Topic modelling.

the ability to acquire knowledge is, which generates a reinforcing feedback loop (R1 in Fig. 5). The current knowledge stock helps to obtain a certain solution to the problem through a model, defined as variable "Model result". The solution is evaluated by Kaggle competition organisers and a value is assigned. The value is compared with the average of competing teams' results to determine a position in the leaderboard. If the position is above the average, then the impact of the results does not reduce the stock of knowledge, 'results impact', and 'knowledge lost' variables, leading to a reinforcing feedback loop (R2 in Fig. 5). However, if the position is below the average, the team will

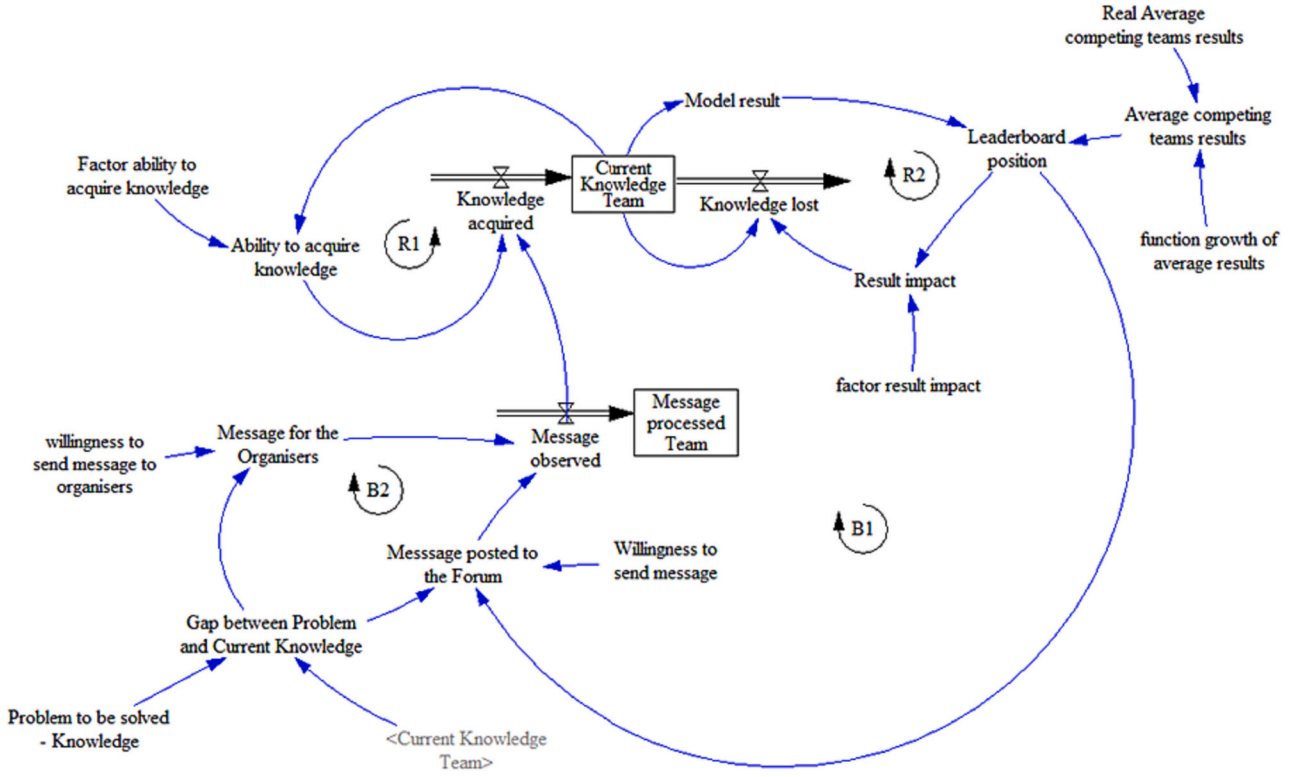


Fig. 5. Stock and flow diagram of the SD model.

discard some of its knowledge defined by 'factor result impact'. The position in the leaderboard generates messages that will be posted in the forum, given a certain willingness to send messages. After the messages are observed, the messages generate new knowledge and increase the stock of knowledge, leading to better model results and a higher leaderboard position. This feedback loop is balancing (B1 in Fig. 5). Finally, another balancing feedback loop (B2 in Fig. 5) originates from the action to reduce the gap between the current knowledge and the knowledge required to solve the problem. The gap between the current knowledge and the team knowledge generates messages for the organisers. Table A1 in the Appendix contains the related equations and variables.

The SD model validation involves examining the model on both structural and behavioural grounds (Morecroft, 2015). To validate our model structure, we compared the SD model with knowledge about the Kaggle system, as described in the literature review. In terms of behaviour, the model was calibrated using optimisation with real data from the competition (Dangerfield and Roberts, 2018). Real data for total messages and model performance were the target for the optimisation process. The variables to be calibrated were factors associated with knowledge acquisition, knowledge lost, and willingness to send messages. Table A2 shows the results of the calibration process, and Fig. A1 in the Appendix shows a good visual fit between the simulated and real data. The statistical tests confirm an adequate calibration (see Table A3 in the Appendix).

4.3. DMLDiD

Causal ML is a type of research method that uses observational data to understand the data generation mechanism behind it (Pearl, 2019). Known as a predictive tool (Bertsimas and Kallus, 2020), the ML model works well with complex and often non-linear data patterns compared with its statistical model counterpart (Breiman, 2001; Shmueli and Koppius, 2011).

The difference-in-differences estimator has been used widely in empirical research to evaluate the causal effects when there exists a natural experiment with a treated group and a control (untreated) group. The DiD estimator is known to be robust against statistical bias such as self-selection (Bertrand et al., 2004; Zhao, 2004; Wooldridge, 2010). By applying the DiD estimator, we address endogeneity issues in this context, consistent with prior work (Francis et al., 2015). The traditional DiD estimator depends on a parallel trend assumption that, in the absence of treatment, the difference between the treated and control groups is observed to be constant over time. However, this assumption may not hold in the case where individual participant characteristics might be associated with the variations of the team performance. The semiparametric DiD estimator has been proposed to address the violation of the parallel trend assumption. It proposes a non-parametrical weighting scheme on the propensity score which could be used to estimate the average effect of the treatment. However, when researchers apply ML methods to estimate the propensity score on a high-dimensional dataset in the first-step estimation, the semiparametric DiD model might fail. Based on the semiparametric DiD, DMLDiD applies Neyman orthogonality to obtain valid inferences, which overcomes the problem of the bias caused by the semiparametric DiD. Specifically, DMLDiD develops a new Neyman-orthogonal score function which adds an adjustment term on the original score function in the semiparametric DiD.

To evaluate the effect of topic content on team performance in an online competition, we first transform the message-level topic into a team-level topic over a specified period so that both content and behaviour variables are in the same team level. For each message sent by a team, the topic content distribution can be obtained. We sum the probability of each topic content from messages sent by the same team within a specific time period and choose the topic with the highest probability score as the main topic that the specific team discussed during that specific time period.

Next, we construct a vector of time-invariant control variables that affect the performance across various team behaviours. We use SD model to infer the parameters of the related behaviour factors; for example, ability to acquire knowledge, result impact, willingness to send message, and willingness to send message to organisers.

Twenty teams are used as seed teams to make the parameter inference for the other teams. These representative teams are chosen on the basis of two measurements—performance score and number of posts. According to the pairwise combination of the two levels of each measurement, four categories of teams are identified: (i) active team whose score is high and who posts often; (ii) passive team whose score is low and who seldom posts; (iii) learner team whose score is low but posts often; and (iv) lurker team whose score is high but who seldom posts (see Table 1).

Among each team category, five representative teams are chosen with the highest rank of the sum of the two measurements. The parameter inference for any non-seed team is based on a similarity metric which is defined as a Euclidean distance-based SoftMax weighting scheme for the 20 seed teams. Ten variables are used to calculate the similarity of non-seed teams weighted on seed teams; these are total number of posts, public score rank, private score rank, first day of submission, last day of submission, number of views, comments and votes on Kernel, number of posts sent to organisers, and number of posts sent to other teams.

Last, the implemented DMLDiD model can be estimated in a simple two-step framework. Here, we define the treatment as a particular topic being discussed. In the first-step estimation, we apply the regularised ML model, such as lasso logistic regression, ridge regression model, and elastic net regression model, to estimate the propensity score which is the probability of a specific topic being discussed conditional on observed individual participant characteristics. The logistic regression model assumes that the log-odds have a linear relationship with the covariates used in DMLDiD. The lasso, ridge regression, and elastic net methods estimate the coefficients. The related formulae and objective function are shown in the DMLDiD section in the Appendix. In the second-step estimation, the average treatment effect (ATT) on the treated topic is identified in an S-fold cross-fitting. The specific estimation of ATT is also shown in the DMLDiD section from the Appendix.

5. Analysis and results

5.1. Topic modelling results

Our primary independent variables of treatment are team-level content indicators from which we obtained topics from the messages posted by team members in the forum of the home credit default risk competition on the Kaggle platform. The content discussed by participants is obtained as topics from STM of text mining. The total number of topics has been chosen by comparing two statistical metrics—semantic coherence and exclusivity. Semantic coherence is a metric that assesses the topic quality in terms of its coherence keywords within the same topic. Exclusivity is another aspect of the topic quality since a good topic model should infer topical words that are frequent and exclusive. From Fig. 6, we choose the number of topics as 10 since this shows a relatively good balance between semantic coherence and exclusivity.

The keywords in each topic cluster are listed in the second column of Table 2. The extracted words span across the specific activities on

Table 1
List of the keywords and topics from STM.

		Forum post	
		High post	Low post
Private score	High rank	Active	Lurker
	Low rank	Learner	Passive

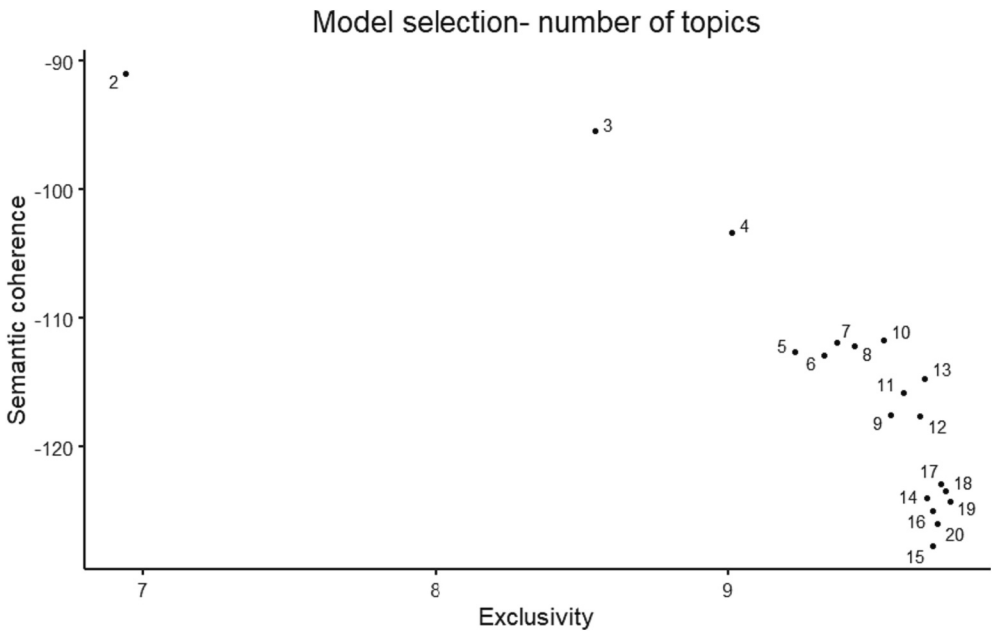


Fig. 6. Selection of number of topics: semantic coherence versus exclusivity Home Credit Default Risk.

Table 2
List of the keywords and topics from STM.

No.	Keywords	Topic
1	competition, team, Kaggle, blockquote, people, will, medal	Submission
2	will, like, number, public, reality, probability, look	Business understanding
3	score, differ, fold, predict, improve, one, parameter	Model inference
4	can, value, help, way, see, share, find	Community support
5	train, set, AUC, test, data, valid, nan	Model evaluation
6	run, lightgbm, use, code, error, file, codepr	Model building
7	thank, tried, now, new, local, will, know	Knowledge sharing
8	model, time, much, kernel, work, better, got	Computational cost
9	feature, mean, credit, loan, application, engine, month	Data preparation
10	use, think, data, also, make, import, one	Data understanding

Kaggle, such as Kaggle competition, user profile, and Kaggle Kernel on which users share their code script/notebook, to the interaction with the virtual community, such as information sharing and support. Taking into account the subject of the competition of credit risk and the nature of the competition in data analytics, we use the cross-industry standard process for data mining (CRISP-DM) to formulate the topics from the mined bag of words related to data science (Martínez-Plumed et al., 2021) and formally summarise the topics as data preparation, business understanding, data understanding, model building, and evaluation, shown in the second column of Table 2. Some non-CRISP-DM topics are also shown in Table 2, such as submission, community support, knowledge sharing, and computational cost.

We periodise the entire duration of the competition into 11 discrete blocks of time by the equal interval binning method. Each block consists of 10 days. Fig. 7 visualises the weight change of each topic over time. It shows that, at the beginning of the competition, the post content is more focused on topic 9 of data preparation. In the middle, model building (topic 6) and knowledge sharing (topic 7) become more promising. When approaching the deadline, topic 1 of submission dominates the other topics. Fig. 8 shows the correlation between each pair of topics. If two topics have a positive correlation, they are more likely to be discussed in the same message. On the contrary, when the correlation is

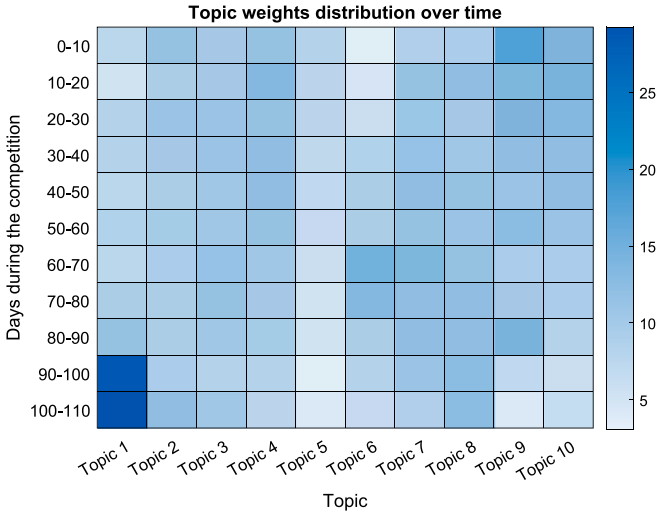


Fig. 7. Heatmap of topic weight distribution.

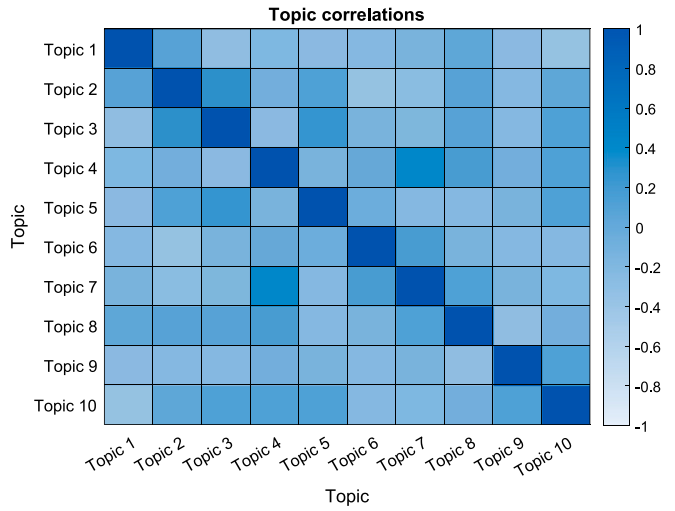


Fig. 8. Topic correlations.

negative, they are more likely to be discussed in different messages. From Fig. 8, topics 4 and 7 are clearly correlated since community support and knowledge sharing are often occurring simultaneously.

The top 25 teams have different behaviours in terms of topics discussed. Most teams posted messages related to topics 1 to 4 and 7 to 9 at similar percentages. Notable differences can be observed in team 3, which focused on topic 9, and team 13 which focused on topic 1, as shown in Fig. 9.

5.2. SD-based causal machine learning

We use a DMLDiD estimator to establish the effect of 10 topics discussed through the messages of the virtual forum on team performance of the online competition. We run separate DiD estimations for each topic and run several robust tests to validate the results. The results are consistent, and we report the analysis below and the robustness check in the Appendix.

To investigate the impact of topics discussed through the virtual forum on team performance of online competition, we employ the following DiD models:

Model 1: $Y_{it} = \beta_j D_{ijt} + \theta_t T_t + \alpha_i X_i + \varepsilon_{it}$

Model 2: $Y_{it} = \beta_j D_{ijt} + \theta_t T_t + \varepsilon_{it}$

Model 3: $Y_{it} = \beta_j D_{ijt} + \varepsilon_{it}$

where i is the index for teams, j is the index for topics, and t is the index for time periods. The whole duration of the competition has been divided into 11 discrete time periods, t, each of which consists of 10 days. The dependent variable Y_{it} is the performance score of team i in time period t. The treatment indicator of D_{ijt} is the binary variable for the main topic content j. Let t_{ij}^* be the time period in which the main topic j was first discussed by team i; $D_{ijt} = 1, if\ t \geq t_{ij}^*$ for treatment and post-treatment time period; $D_{ijt} = 0\ if\ t < t_{ij}^*$ for pre-treatment time period; and zero for all the time periods if topic j is not the main topic discussed by team i which is $D_{ij} = 0$ for all t. X_i is the fixed effect of SD variables for team i and T_t is the fixed effect for time period t. ε_{it} is the error term. The coefficient β_j is the parameter of interest which captures the effect of topic j. θ_t and α_i are constant parameters for the fixed effect of time period and the team, respectively.

To reduce the potential selection bias, we use propensity score matching (PSM) to construct a matching group based on observable measures to compare the team representatives with no significant difference between the team which discusses the specific topic and the team which never discusses that topic. Under the PSM scheme, as attributes to be matched upon, we use the total number of posts, public score rank, private score rank, first day of submission, last day of submission, number of views, comments and votes on Kernel, and number of posts sent to the organiser and other teams. We adopt nearest-neighbour matching with generalised linear model as the distance measure for PSM. We also adopt a variable ratio of $n_0/2n_1$ and a control

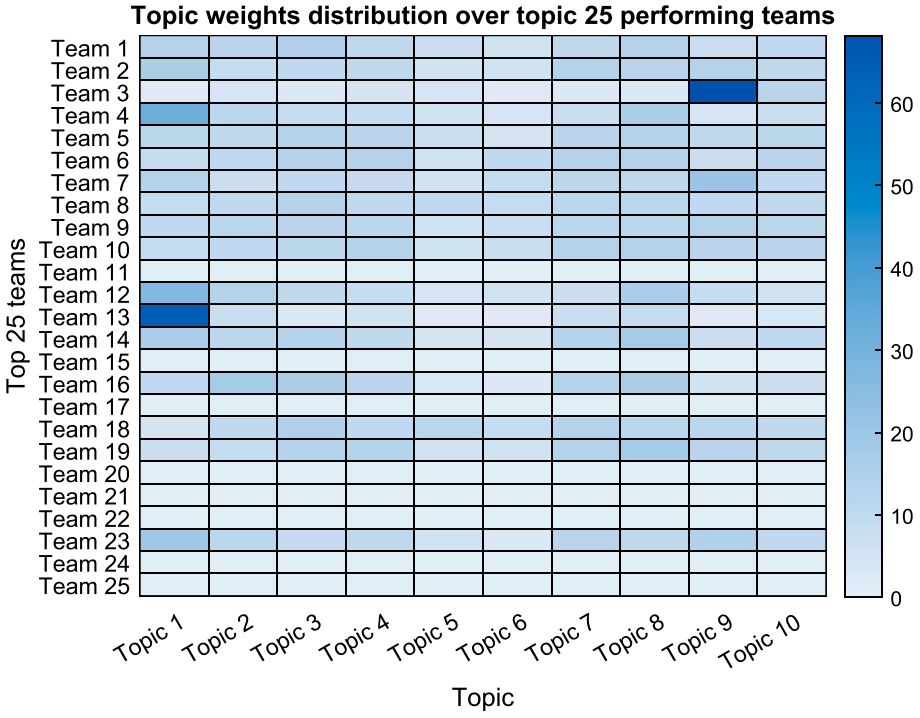


Fig. 9. Topic posting by the top 25 teams.

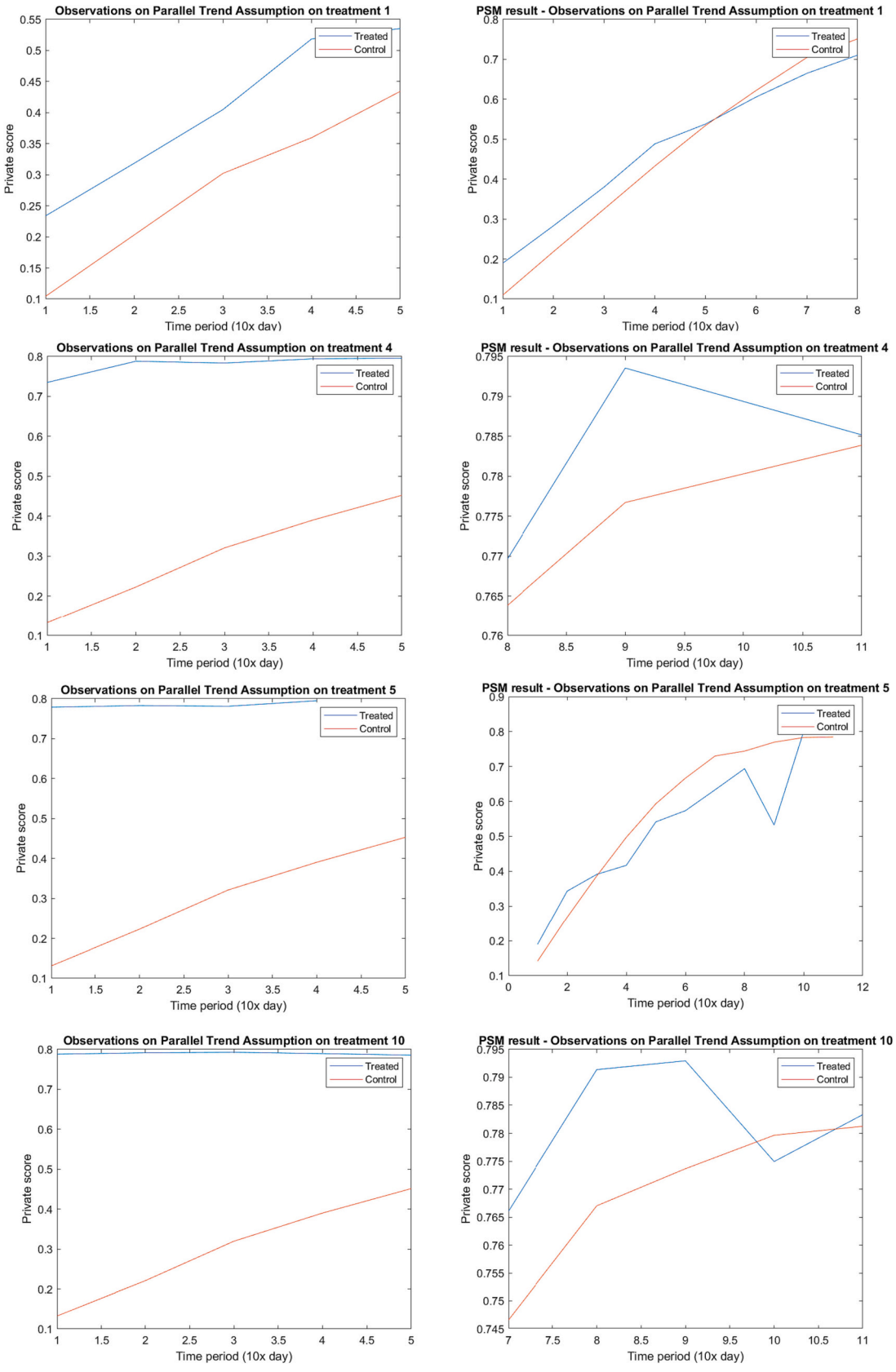


Fig. 10. Parallel assumption test on traditional DiD (left) and DiD with PSM (right).

Table 3
Average treatment effect on treated group.

Topic	1	2	3	4	5	6	7	8	9	10
Model 1	0.365*	5.801***	2.346***	6.843***	9.087***	2.787***	3.369***	3.782***	1.294***	5.031***
	(0.023)	(0.389)	(0.137)	(0.481)	(0.768)	(0.155)	(0.192)	(0.229)	(0.074)	(0.34)
$\psi 1$	0.587	−5.045	−2.333	−1.727	−4.039	−1.026	−1.708	−30.838	−0.792	−4.012
Model 2	2.823***	5.473***	6.19***	2.714***	7.646***	10.751***	3.522***	3.783***	3.694***	1.799***
	(0.052)	(0.32)	(0.414)	(0.13)	(0.513)	(0.827)	(0.174)	(0.199)	(0.197)	(0.079)
$\psi 2$	2.415	−3.742	0.902	2.948	8.347	1.725	1.094	−26.13	1.427	2.482
Model 3	3.045***	11.318***	4.042***	13.477***	18.119***	5.738***	6.189***	5.901***	2.64***	9.144***
	(0.061)	(0.739)	(0.199)	(0.893)	(1.382)	(0.28)	(0.323)	(0.311)	(0.118)	(0.531)
$\psi 3$	2.373	2.359	2.971	10.087	19.762	4.306	3.405	−38.026	2.605	7.314

bound of $[1, n_0/n_1]$ where n_0 and n_1 are the number of control and treated teams, respectively (Ming and Rosenbaum, 2000).

Fig. 10 shows the treatment and control groups' performance score over time which could be used to test parallel assumption on original data (left) and matching data with PSM (right) for some selected topics (full results on all topics are reported in Fig. A2 in the Appendix). Obviously, it exhibits the non-parallel trend for both original data and PSM matching data and a violation of parallel trend assumption in the DiD setting. Therefore, we use a DMLDiD estimator to establish the effect of 10 topics discussed through the post of the virtual forum on team performance of online competition. We use the LASSO regression as our first step of propensity score matching estimation. The results are reported in Table 3.

The casual effect of topic content on team performance can be obtained as the average effect on treated ATT of the 10 topics. A positive and significant value means that a topic has a positive influence on team performance score, while a negative and significant value means that the topic has a negative impact on team performance score. The propensity score ψ shows the performance of matching—a value of 0 denotes the perfect matching. The smaller the absolute value of propensity score, the better the performance of matching. From Table 4, the positive significant coefficients of the 10 topics show that all 10 topics discussed in the forum positively improve performance. Moreover, the value of ATT varies among different topics. From Table 4, we observe that, while all topics' themes play a significant role, model evaluation, community support, and data understanding are the top three topics which have the highest values of ATT. In addition, topic 1 of submission has the smallest coefficient compared with the rest of the topics and shows a relatively weak impact of submission discussion on the team performance in the final score.

Based on model 1, we formulate model 2 without controlling for the team's SD features and model 3 without controlling for SD features and time heterogeneity. Table 3 shows that the performance of matching models 2 and 3 decreases compared with model 1. Without controlling for the team's SD features and time heterogeneity, the result of all the coefficients seems to be overestimated. Without controlling for the team's SD features, the effects of topic of submission, model inference, model building, and data preparation seem to be overestimated while the effects of topic of community support, model evaluation, and data understanding seem to be underestimated. Therefore, involving the team's SD features and the time index does improve the performance of our analysis, and the unobserved heterogeneity at the team level and time period level seems to be responsible for the differences in our results for the three models.

The other two variations of regularised logistic regression with Ridge ($\alpha = 0$) and Elastic net ($\alpha = 0.5$) have also been tested as additional sensitivity analysis. Each model is estimated 100 times and averaged. The result is consistent, as reported in Table A4. We further check the robustness of the results by varying the number of teams ±250. As reported in Table A5, the results are consistent.

To summarise, we started evaluating the content of messages using topic modelling. This information helped us to understand the focus of the teams over time during the competition. However, this information was at a broadly high level and did not provide specific information about the impact of specific topics on individual team behaviour. We addressed this gap using an SD-based causal ML model, which can be considered an AI model enhanced with SD. SD contributed to the model with seeded information to infer the behaviour of representative teams more accurately (see comparison between models 1, 2, and 3).

6. Discussion and implications

6.1. Research implications for open innovation

Our research results enrich theory by exploring aspects which influence innovation performance in data science open innovation. We found that business/data understanding, model building, evaluation, community support, and the usage of a submission system platform are the main knowledge topics exchanged by teams. The results also showed that some teams contribute to the creation of knowledge through posting to support other teams while other teams are only absorbing knowledge by reading the messages and not contributing to the forum. In terms of team behaviour, we segmented all the teams into four categories—active, learner, lurker, and passive—and chose the same number of teams in each category as representative teams.

Our work contributes to open innovation theory in other ways. The DiD results assess the significance of each treatment variable and reflect on the theory in data science and innovation management (Martínez-Plumed et al., 2021). While much has been discovered from the perceived value and cognitive aspects (Antikainen et al., 2010; Garcia

Table 4
"Topic importance" based on estimated treatment effect.

Topic	Keywords	Topic	ATT
5	train, set, AUC, test, data, valid, nan	Model evaluation	9.087
4	can, value, help, way, see, share, find	Community support	6.844
2	will, like, number, public, reality, probability, look	Business understanding	5.802
10	use, think, data, also, make, import, one	Data understanding	5.032
8	model, time, much, kernel, work, better, got	Computational cost	3.782
7	thank, tri, now, new, local, will, know	Knowledge sharing	3.370
6	run, lightgbm, use, code, error, file, codepr	Model building	2.787
3	score, differ, fold, predict, improve, one, parameter	Model inference	2.346
9	feature, mean, credit, loan, application, engine, month	Data preparation	1.294
1	competition, team, Kaggle, blockquote, people, will, medal	Submission	0.365

Martinez, 2017), only limited examination has been undertaken in the scope of the task force and workflow in the data science context. Our work bridges this gap and reveals the relationship between types of activities and performance of the solution using extracted topics from discussion forum content.

Our findings support Bojer and Meldgaard's (2021) conclusion that communities learned from the feedback obtained through the leaderboard and messages. Different from prior work (Wu and Gong, 2019), our model focuses on the innovation performance at the team level rather than at the company level. At the team level, innovation behaviour is driven by balancing feedback loops associated with exogeneous goals such as comparative performance through the competition's leaderboard and the level of difficulty of the problem for teams. Only limited studies have explored how different levels of collaboration affect team performance in crowdsourcing contests (for an exception, see Javadi Khasraghi and Hirschheim (2022)).

6.2. Research implication for system dynamics and data science

From a behaviour perspective, using the SD model, we have the following reflections. First, we found situations in our data where the performance of the model cannot replicate the behaviour observed since lack of consistency in the data. The use of data analytics can help to identify boundary conditions for SD models when optimisation is used for calibration. Second, calibrating SD models using empirical data is very useful to detect patterns across different groups. Definitively, SD modelling can help us to understand endogenous behaviours in clusters as well as in individual entities. These clusters can be used to 'seed' ML models. To summarise our methodology, we suggest starting with a simple SD model that provides an endogenous feedback theory of the behaviour being observed in the individual entities captured in the data and calibrating it using some selected examples. The next step is to test the model using other cases identified by classification algorithms to find boundary conditions. Finally, researchers can expand the model, or develop a new model, to incorporate those boundary conditions for further data processing and inference.

Our work also contributes to the Knowledge Discovery in Databases (KDD) process. Prior literature proposed innovative KDD processes based on different data science scenarios driven by the user case (Martínez-Plumed et al., 2021) as well as the design (Singh et al., 2022). Comparing our results to early-day knowledge discovery and data-driven processes, researchers should recognise the shift in the research paradigm from data mining to a data science trajectory. While CRISP-DM variables such as business understanding, data understanding, and model evaluation still play important roles, our result also reflects other important aspects such as platform characteristics (community support and knowledge sharing) and model building in detail (issues in model inference and computational cost). This potentially suggests that the KDD process model should be extended to include these dimensions when applied in the open innovation context.

Additionally, our work contributes to research methodology by combining SD and ML. We develop a framework that combines a SD model with casual ML and natural language processing. The result of the SD model is also consistent with the categorisation of team behaviour in terms of feedback information processes and dynamic accumulation of knowledge stocks. The insights and results from the SD model allow us to control for causal inference and reduce potential bias in the estimation. Furthermore, we integrate the SD model results into the DMLDiD framework to enhance our understanding about innovation performance based on learning, knowledge sharing, and feedback in teams. STM allows us to investigate post messages using topic assignment in order to complement the insights from the previous methods with a categorisation of the type of messages shared. This creates an opportunity to conduct fine granularity testing of treatment effects conditioning on a specific topic type, which might link to a specific type of theory and action behind these topic themes including, but not limited to, the top discovery, team characteristics, and its community at large (Saura et al., 2021).

Finally, our study shows a possible alternative where SD/ML methods could add value to the current research focus on survey-based research in descriptive settings (Garcia Martinez, 2015; Garcia Martinez, 2017). Researchers should endeavour to include empirical evidence of behaviours such as unstructured data and observation of participants' behaviour in a longitudinal setting whenever possible. This could potentially motivate multidisciplinary research and a distinctive research paradigm in qualitative and quantitative settings.

6.3. Practical implications

It is evident that companies recognise the usefulness of data science open innovation (Tauchert et al., 2020; Javadi Khasraghi and Hirschheim, 2022). However, given its complex nature and the interaction between team players, organisers, and businesses, the main driving force behind performance increase is still largely unknown. Our work demonstrates that user-generated content in the forum is a valuable resource for companies to organise data science open innovation competitions.

Platforms could focus on design functions to enable knowledge sharing and foster community support. Companies should identify the area(s) where participants most need the knowledge. During the competition journey, it may also be valuable for the forum to make recommendations to specific threads that might be relevant to specific contestants given their prior post. From our identified topics, competition organisers could develop activities related to these specific topics for participants, such as data preparation and model inference, which potentially foster discussion that might help improve participants' performance. Technological support could be provided to participants who experience computational issues—e.g., high-performance computing services.

We observe significantly different keywords in topic clusters which might be related to various aspects of data science innovation, such as data understanding, model building, and evaluation. In the future, companies setting up competitions on Kaggle may think about exploring multiple goals rather than just modelling performance using a statistical metric. It might be beneficial to explore external data sources and data value (Martínez-Plumed et al., 2021). While some of the explorations already happen, as Kaggle has made it possible to specify in the data policy that external data sources could be used or excluded, there are more potential activities that a company could choose to innovate in their business. For instance, some businesses would be more interested in developing a novel analytical model, while others have a limited choice of models (e.g., due to a need for model comprehensibility) but would rather evaluate each model more extensively using distinctive metrics and experiment setups.

7. Conclusion and limitations

In this paper, we integrated causal ML with a SD model and STM to gain insights about team performance in data science open innovation. We employed data from Kaggle, which is a well-known source of data for this type of research. We deployed STM to understand UGC in the forum and explore the dynamic behaviour in open innovation using a SD model. Our results show that UGC has a positive effect in improving performance over time. Our results also show that business understanding, model evaluation, and community support are the types of

message content that teams mostly share in data science open innovation coopetition contexts, and they impact team performance. Hopefully, our paper will inspire more complementary research between SD modelling and ML/AI modelling through the combination of insights to generate more robust predictive models.

We identify several limitations in our research. We only applied one competition while other researchers have experimented with more competitions, e.g., brand design, creative writing, and others (Shi et al., 2022). Testing in more competitions will help to identify potential differences in team behaviour given the changes in the context—e.g., level of complexity of the task. We believe that our model is generic, so further replications in different contexts can be useful to confirm the generalisability of our approach.

Other sources of knowledge—e.g., online question-and-answer sites such as Stack Overflow (https://stackoverflow.com/)—might be useful to complement the knowledge exchange during the competition which we have not explored in this study, as it is also an important source for content creators (Gómez et al., 2013). Often known as a platform for dissemination of innovation, such a source might also contribute to the stock of knowledge in the learning and feedback process (Barua et al., 2014).

Another limitation of our approach is the lack of discrimination between messages to answer queries from other participants and messages to ask for information from other participants, such as in Li et al. (2022). We assume that teams can also learn from the answers to the queries from other participants when they reply to them. However, further extension of the approach can be the use of network analysis to identify the linkages between teams and their flows of information.

In terms of data, we did not use Kaggle skill points to measure their initial stock of knowledge, which has been employed in previous research (Garcia Martinez, 2017), since the stock of knowledge from previous competitions may be unrelated to the problem that they are facing for the competition. Finally, we did not use solution-sharing measures—e.g., Kernel votes—to evaluate team performance and knowledge-sharing processes because it is not possible to identify the teams using this approach. Potential future research may try to identify the teams using the solutions and team networks.

CRediT authorship contribution statement

Libo Li: Conceptualization, Data curation, Methodology, Formal analysis, Writing – review & editing. **Huan Yu:** Conceptualization, Formal analysis, Visualization, Validation, Writing – review & editing. **Martin Kunc:** Conceptualization, Methodology, Formal analysis, Writing – review & editing.

Declaration of competing interest

None.

Data availability

Data will be made available on request.

Acknowledgement

We thank editors for handling our manuscript. We also thank peer reviewers for their invaluable suggestions.

Appendix A

The SD model – theoretical background

Table A1
Equations of the SD model.

Concept	Type of component	Equation	Explanation	Source
Current knowledge	Stock	Knowledge = Knowledge acquired – Knowledge lost	Stock of existing knowledge per team, which reflects the knowledge acquired from messages and the knowledge discarded due to relatively poor performance.	Private performance of the team during the Kaggle competition. Javadi Khasraghi and Hirschheim (2022)
Knowledge acquired	Flow	Message observed*Ability to acquire knowledge	Absorption of the knowledge from a message.	Wu and Gong (2019); Javadi Khasraghi and Hirschheim (2022).
Ability to acquire knowledge	Auxiliary	Current knowledge team/factor ability to acquire knowledge	It reflects the ability of a team to transform messages into incremental knowledge. It is based on the current knowledge and the ability to use its knowledge to understand messages (see factor ability to learn new knowledge).	Wu and Gong (2019); Javadi Khasraghi and Hirschheim (2022).
Factor ability to learn new knowledge	Constant	See results in Table A2	This variable captures the capacity to transform messages into new knowledge. An automatic calibration process generates its value. See calibration.	Wu and Gong (2019); Javadi Khasraghi and Hirschheim (2022).
Knowledge lost	Flow	Knowledge*result impact	Outflow of existing knowledge.	Otto and Simon (2008).
Model results	Auxiliary	Current knowledge	The results obtained from their analytics solution (model) are the same as the level of knowledge.	Wu and Gong (2019); Javadi Khasraghi and Hirschheim (2022).
Leaderboard position	Auxiliary	Model result/Average competing teams' results	This equation indicates how the results obtained by a team are above or below the current results obtained by the rest of the teams.	Cao et al. (2022); Javadi Khasraghi and Hirschheim (2022).
Average competing teams results	Constant	Real average competing teams results	Time series of the average results obtained by the teams in the competition.	Private performance of the team during the Kaggle competition.
Result impact	Auxiliary	IF (Leaderboard position>5 % above average) THEN (0) ELSE (Leaderboard position/factor result impact)	Indicates the consideration of the team with respect to its leaderboard position. If the performance is above the rest of the teams, there will not be any change in the knowledge. If the performance is	Cao et al. (2022).

(*continued on next page*)

Table A1 (*continued*)

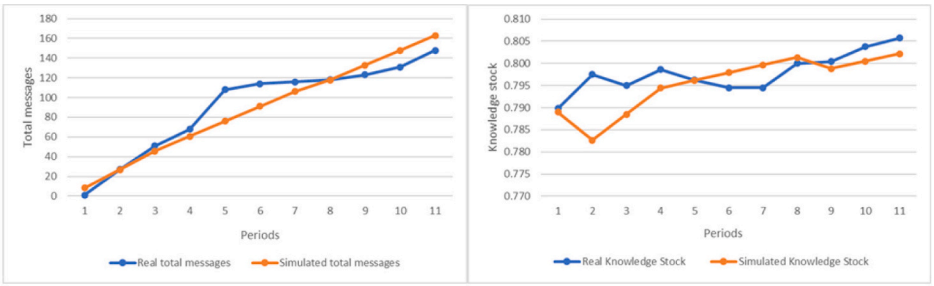
Concept	Type of component	Equation	Explanation	Source
			below 5 %, the team will eliminate some of its knowledge based on a parameter (see factor result impact).	
Factor result impact	Constant	See results in Table A2	It represents the importance given to the position in the leaderboard. An automatic calibration process generates the value. See calibration.	Cao et al. (2022).
Message processed	Stock	Message processed = Message observed	Accumulation of messages sent by the team during the competition.	Number posted by a team during the Kaggle competition. Javadi Khasraghi and Hirschheim (2022).
Message observed	Flow	Message posted to the Forum+Message for the Organisers	All messages posted to either the forum for other teams or the organisers are observed.	Wu and Gong (2019); Javadi Khasraghi and Hirschheim (2022).
Message posted to the Forum	Auxiliary	Gap between Problem and Current Knowledge*Willingness to send message + IF (Leaderboard position is below than the average) THEN (Leaderboard position*Willingness to send message) ELSE (0)	There are two types of messages posted in the Forum: messages to learn more about the problem and messages due to poor performance. Both actions are controlled by a factor called Willingness to send message (see next row).	Wang et al. (2019); Wu and Gong (2019); Javadi Khasraghi and Hirschheim (2022).
Willingness to send message	Constant	See results in Table A2.	It indicates a behavioural predisposition to send a message to obtain help based on the relative performance of the team. An automatic calibration process generates the value. See calibration.	Wu and Gong (2019); Javadi Khasraghi and Hirschheim (2022).
Problem to be solved – Knowledge	Constant	Maximum value to obtain from the competition	It indicates the complexity of the problem for the teams competing.	
Gap between problem and current knowledge.	Auxiliary	Problem to be solved – Knowledge/Current Knowledge Team	Relationship between the stock of knowledge and the knowledge required for the problem complexity.	Cao et al. (2022).
Message for the organisers	Auxiliary	IF (Gap between Problem and Current Knowledge >1) THEN (Gap between Problem and Current Knowledge*willingness to send message to organisers) ELSE (0)	Message to the organisers due to the gap between current knowledge and knowledge needed to solve the problem.	Wang et al. (2019); Wu and Gong (2019).
Willingness to send message to organisers	Constant	See results in Table A2.	It indicates a behavioural predisposition to send a message to obtain help from the organisers. An automatic calibration process generates the value. See calibration.	Wu and Gong (2019).

Sample output of the system dynamics model

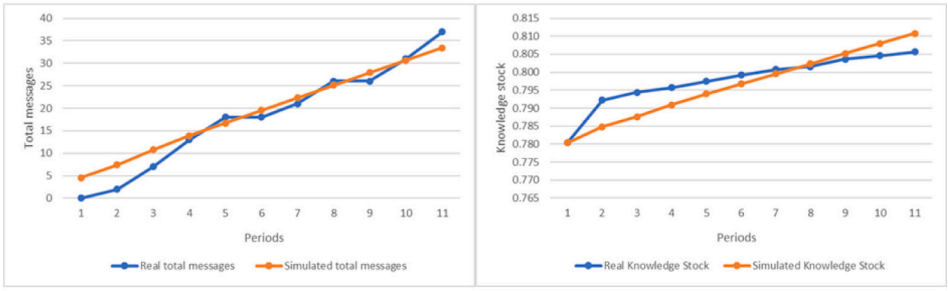
Table A2
Five cases of teams with high ranks and high number of posts.

Variable	Team 1	Team 2	Team 5	Team 9	Team 23
Team performance: beginning/end	0.7898 / 0.8057	0.7804 / 0.8056	0.7931 / 0.8045	0.7867 / 0.8039	0.7749 / 0.8012
Total number of messages	148	37	127	239	16
Factor ability to acquire knowledge	672.69	800	801.78	790.34	342.15
Factor result impact	493.76	819.43	635.99	363.04	12,143.00
Willingness to send message	5.62	0.45	4.82	4.79	0.12
Willingness to send message to organisers	0.83	1.79	2.57	1.74	0.00
The number of the team is the ranking in private score.					

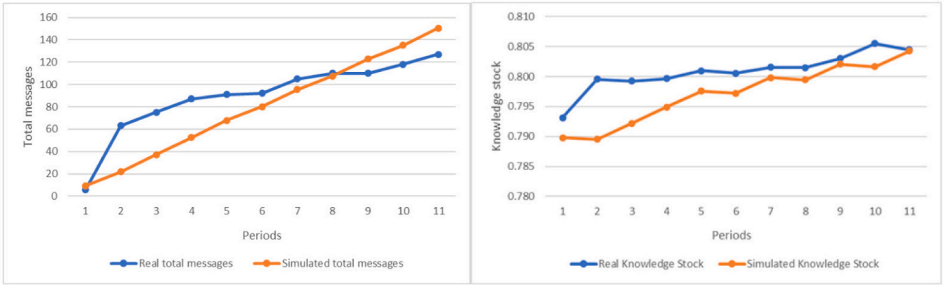
Team 1



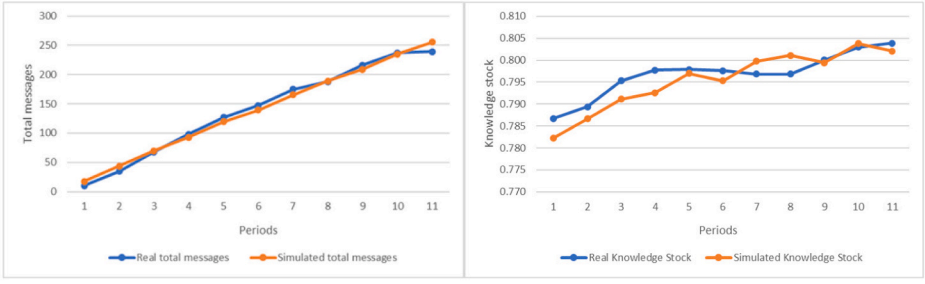
Team 2



Team 5



Team 9



Team 23

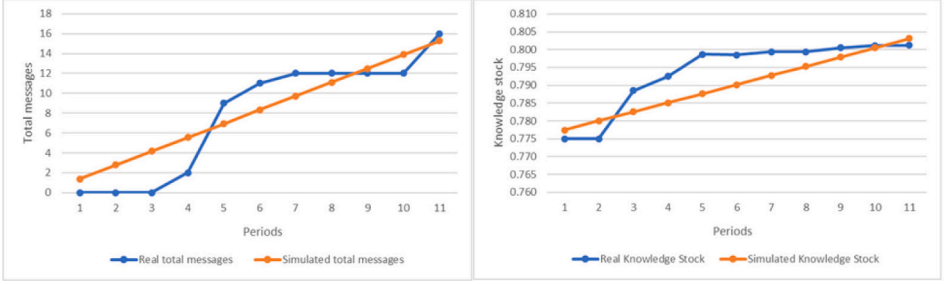


Fig. A1. Time series for Simulated and Real Data for four teams.

Table A3
Mean absolute percentage error results.

Variable	Team 1	Team 2	Team 5	Team 9	Team 23
Total messages	75.6 %	437.4 %	27.8 %	12.1 %	755.8 %
	(10 %)	(10 %)	(18 %)	(7 %)	(4 %)
Knowledge stock	0.5 %	0.4 %	0.5 %	0.3 %	0.6 %

The number between brackets is the result for the last period.

A test of parallel assumption

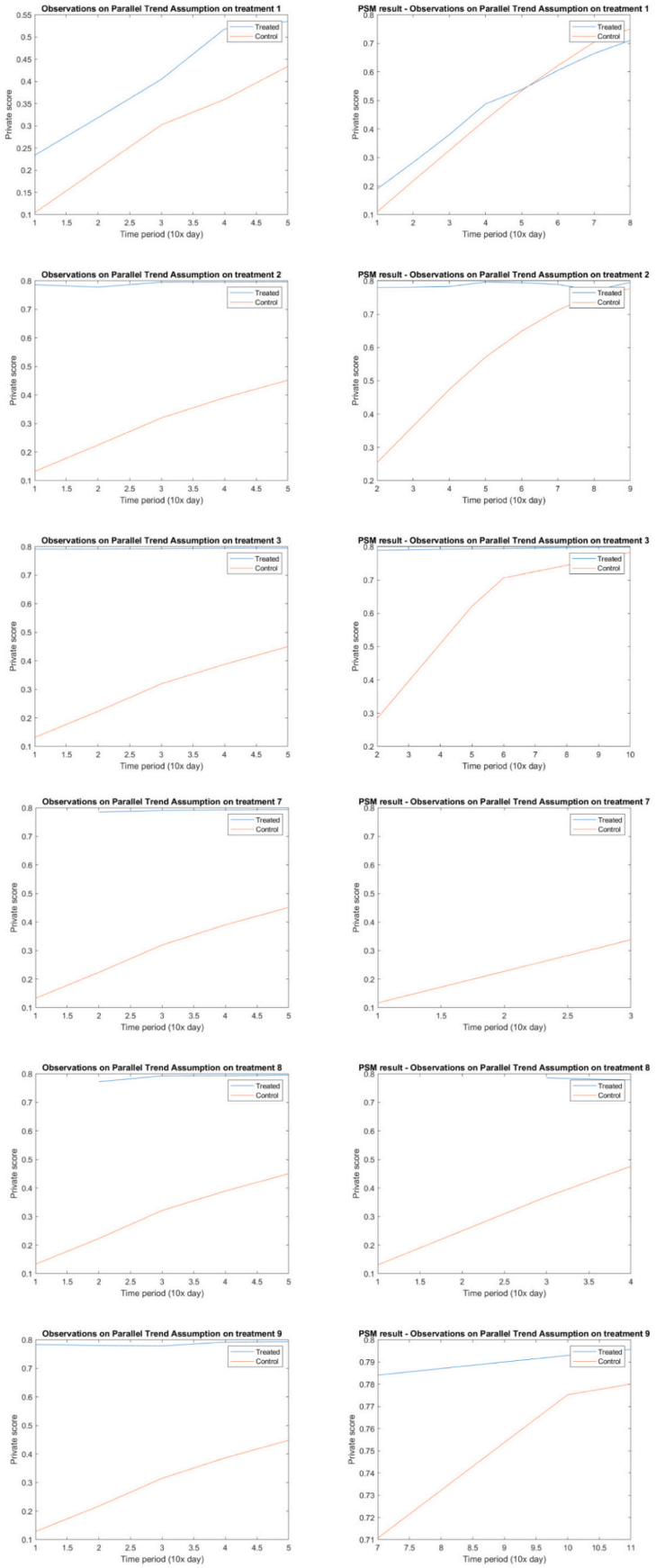


Fig. A2. A test of parallel assumption on raw data (left) and propensity score-matched data (right).

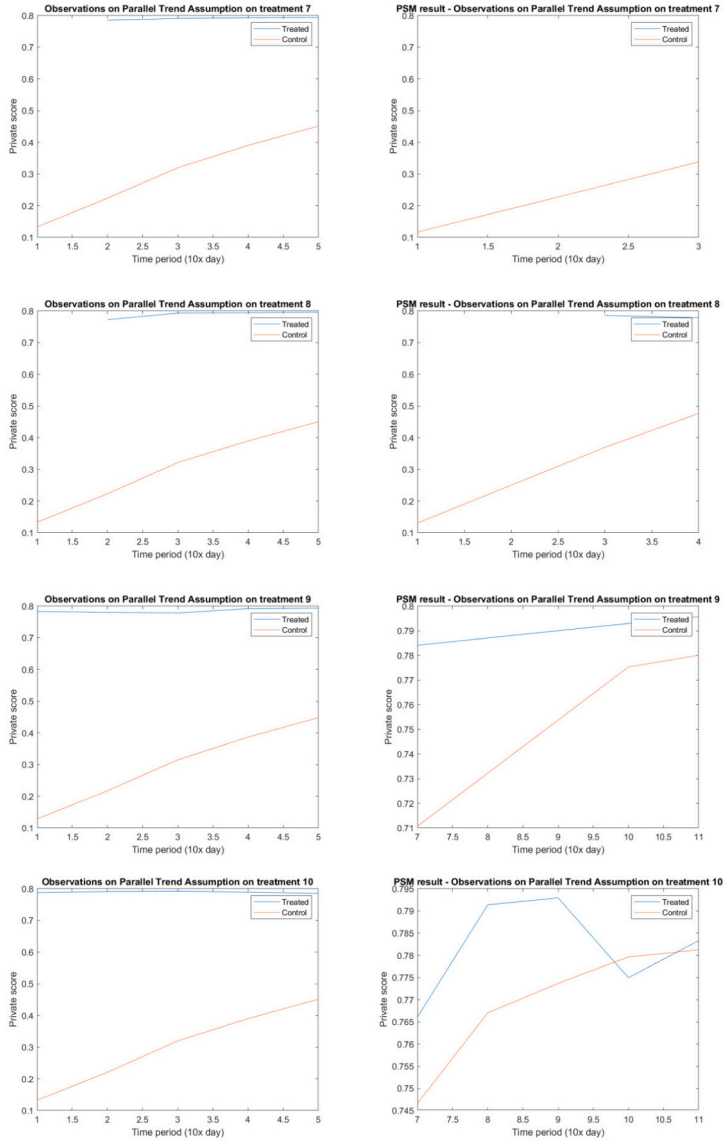


Fig. A2. (*continued*).

STM

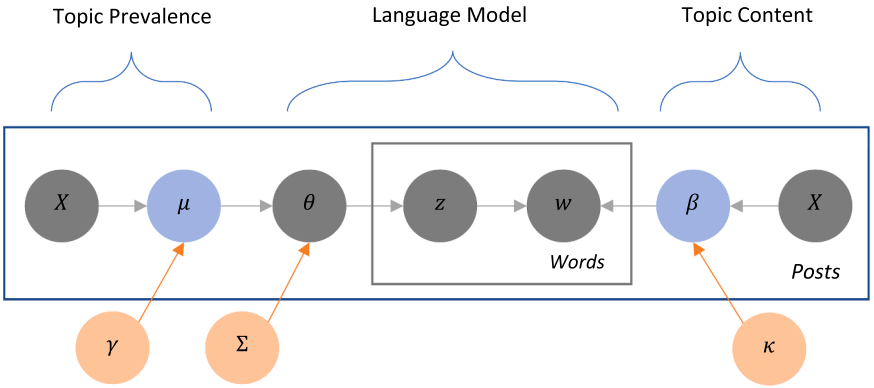


Fig. A3. Plate diagram for the STM applied in Kaggle's posts.

Step 1: Choose $\theta_d \sim LogisticNormal(\mu_d, \Sigma)$;

$\mu_{d,k} = X_d \gamma_k$;

$\gamma_k \sim Normal(0, \sigma_k^2)$;

$\sigma_k^2 \sim Gamma(s^\gamma, r^\gamma)$.

Step 2: For each of N words w_n:

a. Choose a topic $z_{d,n} \sim Multinominal(\theta_d)$;

b. Choose a word w_n from a multinominal probability conditioned on the topic $z_{d,n}$, $w_{d,n} \sim Multinominal(\beta_d^{k=z_{d,n}})$;

$\beta_{d,v}^k \propto \exp(m_v + \kappa_v^{',k} + \kappa_v^{x,'} + \kappa_v^{x,k})$;

$\kappa_v^{x,k} \sim Laplace(0, \tau_v^{x,k})$;

$\tau_v^{x,k} \sim Gamma(s^\kappa, r^\kappa)$.

Fig. A4. Generative process of STM.

DMLDiD

Logistic regression:

$$log\left(\frac{p(z^{(k)}|X^{DiD})}{1-p(z^{(k)}|X^{DiD})}\right) = b_0 + b_1 X_1^{DiD} + \cdots + b_j X_j^{DiD}$$

Objective function of ridge, lasso, and elastic net logistic regression:

$$\hat{b}^{Ridge} = \frac{argmin}{b} \sum_{d=1}^{M} \left(y_d - b_0 - \sum_{j=1}^{J} b_j X_j^{DiD}\right)^2 + \lambda \sum_{j=1}^{J} b_j^2$$

$$\hat{b}^{Lasso} = \frac{argmin}{b} \sum_{d=1}^{M} \left(y_d - b_0 - \sum_{j=1}^{J} b_j X_j^{DiD}\right)^2 + \lambda \sum_{j=1}^{J} |b_j|$$

$$\hat{b}^{Elastic\ net} = \frac{argmin}{b} \sum_{d=1}^{M} \left(y_d - b_0 - \sum_{j=1}^{J} b_j X_j^{DiD}\right)^2 + \lambda_1 \sum_{j=1}^{J} b_j^2 + \lambda_2 \sum_{j=1}^{J} |b_j|$$

Estimation of ATT

More specifically, for an S-fold random partitioning, we denote $\{I_s\}_{s=1}^{S}$ of the M post observations of $\{w_1, w_2, \cdots, w_M\}$ and define the auxiliary

sample $I_s^c \equiv \{w_1, w_2, \cdots, w_M\} \backslash I_s$. For each S-fold random partitioning, we construct the intermediate ATT estimator as:

$$\widetilde{\alpha}_s = \frac{1}{n} \sum_{i \in I_s} \frac{D_i - \widehat{g}_s(X_i)}{\widehat{p}_s \widehat{\lambda}_s (1 - \widehat{\lambda}_s)(1 - \widehat{g}_s(X_i))} \times ((T_i - \widehat{\lambda}_s) Y_i - \widehat{l}_s(X_i)),$$

where $\widehat{p}_s = \frac{1}{M} \sum_{i \in I_s^c} D_i$, $\widehat{\lambda}_s = \frac{1}{M} \sum_{i \in I_s^c} T_i$ and $(\widehat{g}_s, \widehat{l}_s)$ are the estimators trained from I_k^c; $\widehat{g}$ are the estimated propensity scores from the related regularised ML model; and $\widehat{l}$ is the estimator obtained from the propensity of the control group. The final ATT estimation is constructed on the averages of all folds, $\widetilde{\alpha} = \frac{1}{|S|} \sum_{s=1}^{S} \widetilde{\alpha}_s$. The fold s is set to be 2 in this paper, consistent with prior literature (Chang, 2020).

Robustness check – ML model choice

Three regularised regression results averaged over 100 runs are summarised below in Table A4. We test Lasso ($\alpha = 1$), the Elastic net ($\alpha = 0.5$), and Ridge ($\alpha = 0$). The penalisation parameter λ is tuned over 3-fold cross validation. A forward slash means that the results are not available due to unsuccessful model building resulting from noisy data.

Table A4
Robustness checks —machine learning model choice.

Topic	Lasso	Elastic net $\alpha = 0.5$	Ridge
1	0.365*	0.367**	0.94***
	(0.023)	(0.02)	(0.043)
2	5.801***	5.848***	6.848***
	(0.389)	(0.392)	(0.454)
3	2.346***	2.369***	2.636***
	(0.137)	(0.136)	(0.15)
4	6.843***	/	8.015***
	(0.481)		(0.551)
5	9.087***	9.498***	11.094***
	(0.768)	(0.788)	(0.888)
6	2.787***	2.817***	3.383***
	(0.155)	(0.156)	(0.181)
7	3.369***	3.301***	3.631***
	(0.192)	(0.188)	(0.209)
8	3.782***	3.736***	/
	(0.229)	(0.228)	
9	1.294***	1.28***	1.23***
	(0.074)	(0.073)	(0.075)
10	5.031***	4.934***	5.487***
	(0.34)	(0.334)	(0.37)

Robustness check – number of teams included in the analysis

Table A5 shows that a sample size of 750 reported similar results to the main paper results in terms of model estimate and significance (except topic 1, whereas its estimate is insignificant anyway). When the sample size increases to 1250, the model estimates are mostly stable as well.

Table A5
Robustness checks – number of teams included.

Topic	750	1000 (reported in the paper)	1250
1	0.033	0.365*	0.36*
	(0.016)	(0.023)	(0.024)
2	5.31***	5.801***	6.442***
	(0.392)	(0.389)	(0.408)
3	1.449***	2.346***	3.176***
	(0.108)	(0.137)	(0.171)
4	6.186***	6.843***	/
	(0.457)	(0.481)	
5	6.962***	9.087***	9.844***
	(0.634)	(0.768)	(0.775)
6	1.789***	2.787***	3.435***
	(0.118)	(0.155)	(0.175)
7	3.127***	3.369***	3.775***
	(0.203)	(0.192)	(0.201)
8	2.732***	3.782***	4.72***
	(0.184)	(0.229)	(0.27)
9	0.794**	1.294***	1.837***
	(0.06)	(0.074)	(0.093)
10	3.672***	5.031***	6.262***
	(0.282)	(0.34)	(0.383)

References

Abadie, A., 2005. Semiparametric difference-in-differences estimators. Rev. Econ. Stud. 72, 1–19. https://doi.org/10.1111/0034-6527.00321.

Afuah, A., Tucci, C.L., 2012. Crowdsourcing as a solution to distant search. Acad. Manag. Rev. 37, 355–375. https://doi.org/10.5465/amr.2010.0146.

Amabile, T.M., Hill, K.G., Hennessey, B.A., Tighe, E.M., 1994. The work preference inventory: assessing intrinsic and extrinsic motivational orientations. J. Pers. Soc. Psychol. 66, 950.

Angrist, J.D., Pischke, J.-S., 2008. Mostly Harmless Econometrics. Princeton university press.

Antikainen, M., Mäkipää, M., Ahonen, M., 2010. Motivating and supporting collaboration in open innovation. Eur. J. Innov. Manag. 13, 100–119.

Athanasopoulos, G., Hyndman, R.J., 2011. The value of feedback in forecasting competitions. Int. J. Forecast. 27, 845–849. https://doi.org/10.1016/j.ijforecast.2011.03.002.

Barua, A., Thomas, S.W., Hassan, A.E., 2014. What are developers talking about? An analysis of topics and trends in Stack Overflow. Empir. Softw. Eng. 19, 619–654. https://doi.org/10.1007/s10664-012-9231-y.

Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? Q. J. Econ. 119, 249–275.

Bertsimas, D., Kallus, N., 2020. From predictive to prescriptive analytics. Manag. Sci. 66, 1025–1044.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Bojer, C.S., Meldgaard, J.P., 2021. Kaggle forecasting competitions: an overlooked learning opportunity. Int. J. Forecast. 37, 587–603. https://doi.org/10.1016/j.ijforecast.2020.07.007.

Breiman, L., 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. 16, 199–231.

Cao, F., Wang, W., Lim, E., Liu, X., Tan, C.-W., 2022. Do social dominance-based Faultlines help or hurt team performance in crowdsourcing tournaments? J. Manag. Inf. Syst. 39, 247–275.

Chang, N.-C., 2020. Double/debiased machine learning for difference-in-differences models. Econ. J. 23, 177–191. https://doi.org/10.1093/ectj/utaa001.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econ. J. 21, C1–C68. https://doi.org/10.1111/ectj.12097.

Dangerfield, B., Roberts, C., 2018. An overview of strategy and tactics in system dynamics optimization. Syst. Dyn. 165–196.

Diker, V., 2004. A Dynamic Feedback Framework for Studying Growth Policies in Open Online Collaboration Communities. AMCIS 2004 Proc. 328.

Doldor, E., Wyatt, M., Silvester, J., 2019. Statesmen or cheerleaders? Using topic modeling to examine gendered messages in narrative developmental feedback for leaders. Leadersh. Q. 30, 101308 https://doi.org/10.1016/j.leaqua.2019.101308.

Erzurumlu, S.S., Pachamanova, D., 2020. Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations. Technol. Forecast. Soc. Change 156, 120041. https://doi.org/10.1016/j.techfore.2020.120041.

Faraj, S., Jarvenpaa, S.L., Majchrzak, A., 2011. Knowledge collaboration in online communities. Organ. Sci. 22, 1224–1239.

Francis, B., Hasan, I., Park, J.C., Wu, Q., 2015. Gender differences in financial reporting decision making: evidence from accounting conservatism. Contemp. Account. Res. 32, 1285–1318. https://doi.org/10.1111/1911-3846.12098.

Gao, B., Zhu, M., Liu, S., Jiang, M., 2022. Different voices between Airbnb and hotel customers: an integrated analysis of online reviews using structural topic model. J. Hosp. Tour. Manag. 51, 119–131. https://doi.org/10.1016/j.jhtm.2022.03.004.

Garcia Martinez, M., 2015. Solver engagement in knowledge sharing in crowdsourcing communities: exploring the link to creativity. Res. Policy 44, 1419–1430. https://doi.org/10.1016/j.respol.2015.05.010.

Garcia Martinez, M., 2017. Inspiring crowdsourcing communities to create novel solutions: competition design and the mediating role of trust. Technol. Forecast. Soc. Change 117, 296–304. https://doi.org/10.1016/j.techfore.2016.11.015.

Gómez, C., Cleary, B., Singer, L., 2013. A study of innovation diffusion through link sharing on stack overflow. In: 2013 10th Working Conference on Mining Software Repositories (MSR). IEEE, pp. 81–84.

Hayashi, T., Shimizu, T., Fukami, Y., 2021. Collaborative Problem Solving on a Data Platform Kaggle. ArXiv Prepr. ArXiv210711929.

Javadi Khasraghi, H., Hirschheim, R., 2022. Collaboration in crowdsourcing contests: how different levels of collaboration affect team performance. Behav. Inform. Technol. 41, 1566–1582.

Jin, Y., Lee, H.C.B., Ba, S., Stallaert, J., 2021. Winning by learning? Effect of knowledge sharing in crowdsourcing contests. Inf. Syst. Res. 32, 836–859.

Johnson, S.L., Faraj, S., Kudaravalli, S., 2014. Formation of power law distributions in online communities. MIS Q. 38, 795–808.

Kumar, V., Srivastava, A., 2022. Trends in the thematic landscape of corporate social responsibility research: a structural topic modeling approach. J. Bus. Res. 150, 26–37. https://doi.org/10.1016/j.jbusres.2022.05.075.

Kunc, M., 2012. System dynamics and innovation: a complex problem with multiple levels of analysis. In: The 30th International Conference of the System Dynamics Society.

Leimeister, J.M., Huber, M., Bretschneider, U., Krcmar, H., 2009. Leveraging crowdsourcing: activation-supporting components for IT-based ideas competition. J. Manag. Inf. Syst. 26, 197–224. https://doi.org/10.2753/MIS0742-1222260108.

Li, X., Bai, Y., Kang, Y., 2022. Exploring the social influence of the Kaggle virtual community on the M5 competition. Int. J. Forecast. 38, 1507–1518.

Ma, T., Zhou, X., Liu, J., Lou, Z., Hua, Z., Wang, R., 2021. Combining topic modeling and SAO semantic analysis to identify technological opportunities of emerging technologies. Technol. Forecast. Soc. Change 173, 121159. https://doi.org/10.1016/j.techfore.2021.121159.

Mao, Y., Vassileva, J., Grassmann, W., 2007. A system dynamics approach to study virtual communities. In: 2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07). IEEE, p. 178a.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M.J., Flach, P., 2021. CRISP-DM twenty years later: from data mining processes to data science trajectories. IEEE Trans. Knowl. Data Eng. 33, 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680.

McClelland, D.C., Atkinson, J.W., Clark, R.A., Lowell, E.L., 1953. The Achievement Motive, Century Psychology Series. Appleton-Century-Crofts, East Norwalk, CT, US. https://doi.org/10.1037/11144-000.

Ming, K., Rosenbaum, P.R., 2000. Substantial gains in bias reduction from matching with a variable number of controls. Biometrics 56, 118–124.

Morecroft, J.D.W., 2015. Strategic Modelling and Business Dynamics: A Feedback Systems Approach. John Wiley & Sons.

Morgeson, F.P., Humphrey, S.E., 2006. The Work Design Questionnaire (WDQ): developing and validating a comprehensive measure for assessing job design and the nature of work. J. Appl. Psychol. 91, 1321.

Otto, P., Simon, M., 2008. Dynamic perspectives on social characteristics and sustainability in online community networks. Syst. Dyn. Rev. J. Syst. Dyn. Soc. 24, 321–347.

Pearl, J., 2019. The seven tools of causal inference, with reflections on machine learning. Commun. ACM 62, 54–60.

Roberts, J.A., Hann, I.-H., Slaughter, S.A., 2006. Understanding the motivations, participation, and performance of open source software developers: a longitudinal study of the Apache projects. Manag. Sci. 52, 984–999.

Roberts, M.E., Stewart, B.M., Tingley, D., 2019. Stm: An R package for structural topic models. J. Stat. Softw. 91, 1–40.

Rose, R.L., Puranik, T.G., Mavris, D.N., Rao, A.H., 2022. Application of structural topic modeling to aviation safety data. Reliab. Eng. Syst. Saf. 224, 108522 https://doi.org/10.1016/j.ress.2022.108522.

Saura, J.R., Ribeiro-Soriano, D., Palacios-Marqués, D., 2021. From user-generated data to data-driven innovation: a research agenda to understand user privacy in digital markets. Int. J. Inf. Manag. 60, 102331 https://doi.org/10.1016/j.ijinfomgt.2021.102331.

Saura, J.R., Palacios-Marqués, D., Ribeiro-Soriano, D., 2023. Exploring the boundaries of open innovation: evidence from social media mining. Technovation 119, 102447. https://doi.org/10.1016/j.technovation.2021.102447.

Schumann, M., Severini, T.A., Tripathi, G., 2021. Integrated likelihood based inference for nonlinear panel data models with unobserved effects. J. Econ. 223, 73–95. https://doi.org/10.1016/j.jeconom.2020.10.001.

Shao, B., Shi, L., Xu, B., Liu, L., 2012. Factors affecting participation of solvers in crowdsourcing: an empirical study from China. Electron. Mark. 22, 73–82. https://doi.org/10.1007/s12525-012-0093-3.

Shi, X., Evans, R., Shan, W., 2022. Solver engagement in online crowdsourcing communities: the roles of perceived interactivity, relationship quality and psychological ownership. Technol. Forecast. Soc. Change 175, 121389.

Shmueli, G., Koppius, O.R., 2011. Predictive analytics in information systems research. MIS Q. 553–572.

Singh, V., Singh, A., Joshi, K., 2022. Fair CRISP-DM: Embedding Fairness in Machine Learning (ML) Development Life Cycle. HICSS, pp. 1–10.

Tauchert, C., Buxmann, P., Lambinus, J., 2020. Crowdsourcing data science: a qualitative analysis of organizations' usage of Kaggle competitions. In: Proceedings of the 53rd Hawaii International Conference on System Sciences.

Tonidandel, S., Summerville, K.M., Gentry, W.A., Young, S.F., 2022. Using structural topic modeling to gain insight into challenges faced by leaders. Leadersh. Q. 33, 101576 https://doi.org/10.1016/j.leaqua.2021.101576.

Wang, J., Zhang, R., Hao, J.-X., Chen, X., 2019. Motivation factors of knowledge collaboration in virtual communities of practice: a perspective from system dynamics. J. Knowl. Manag. 23, 466–488.

Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT press.

Wu, B., Gong, C., 2019. Impact of open innovation communities on enterprise innovation performance: a system dynamics perspective. Sustainability 11, 4794.

Xu, S., Hao, L., Yang, G., Lu, K., An, X., 2021. A topic models based framework for detecting and forecasting emerging technologies. Technol. Forecast. Soc. Change 162, 120366. https://doi.org/10.1016/j.techfore.2020.120366.

Ye, H., Kankanhalli, A., Huber, M.J., Bretschneider, U., Blohm, I., Goswami, S., Leimeister, J.M., Krcmar, H., 2012. Collaboration and the Quality of User Generated Ideas in Online Innovation Communities. Academy of Management Annual Meeting. https://doi.org/10.5465/AMBPP.2012.16803abstract.

Zhao, Z., 2004. Using matching to estimate treatment effects: data requirements, matching metrics, and Monte Carlo evidence. Rev. Econ. Stat. 86, 91–107.

Zhu, L., Cunningham, S.W., 2022. Unveiling the knowledge structure of technological forecasting and social change (1969–2020) through an NMF-based hierarchical topic model. Technol. Forecast. Soc. Change 174, 121277. https://doi.org/10.1016/j.techfore.2021.121277.

Libo Li is a Lecturer in Business analytics at Southampton Business School, University of Southampton, United Kingdom. Libo received his Ph.D. in Business economics from Katholieke Universiteit Leuven in Belgium. Libo's research interests are social network analytics, data mining, machine learning, and statistical modelling. His research appears in academic outlets and conferences including Computer networks, European Journal of

Operational Research, IEEE Transactions on Engineering Management and International Conference on Information Systems.

Huan Yu is a Lecturer in Business Analytics within Southampton Business School at the University of Southampton. Huan completed her PhD in Management Science and Engineering from School of Management, University of Science and Technology of China. During her PhD, she was a visiting research fellow at IESEG school of Management in Paris and National University of Singapore. Huan's research interests are customer learning, demand modelling and decision-making optimization under uncertainty.

Martin Kunc is Professor of Management Science at Southampton Business School, University of Southampton, UK. He serves as editor-in-chief of the Journal of the Operational Research Society and associate editor for the Journal of Simulation and Journal of Business Analytics. He has published >50 papers in journals such as SMJ, Omega, JORS and TFSCh. His current research interests include simulation, scenario planning and behavioural operations.