informs

https://pubsonline.informs.org/doi/abs/10.1287/mksc.2023.1434

MARKETING SCIENCE
Articles in Advance, pp. 1–19
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Can Artificial Intelligence Help in Ideation? A Theory-Based Model for Idea Screening in Crowdsourcing Contests

J. Jason Bell,[a,*] Christian Pescher,[b] Gerard J. Tellis,[c] Johann Fueller[d]

[a] Saïd Business School, University of Oxford, Oxford OX1 1HP, United Kingdom; [b] Universidad de los Andes, Bogotá, 110311, Colombia;
[c] Marshall School of Business, University of Southern California, Los Angeles, California 90089; [d] University of Innsbruck, 6020 Innsbruck, Austria

*Corresponding author

Contact: jason.bell@sbs.ox.ac.uk, https://orcid.org/0000-0001-7780-2368 (JJB); christian.pescher@fau.de, https://orcid.org/0000-0002-5132-9178 (CP); tellis@usc.edu, https://orcid.org/0000-0002-2677-3840 (GJT); johann.fueller@uibk.ac.at (JF)

Received: August 13, 2020
Revised: July 28, 2021; September 27, 2022
Accepted: January 10, 2023
Published Online in Articles in Advance:
■■■■ ■■, 2023

https://doi.org/10.1287/mksc.2023.1434

Abstract. Crowdsourcing generates up to thousands of ideas per contest. The selection of best ideas is costly because of the limited number, objectivity, and attention of experts. Using a data set of 21 crowdsourcing contests that include 4,191 ideas, we test how artificial intelli- gence can assist experts in screening ideas. The authors have three major findings. First, whereas even the best previously published theory-based models cannot mimic human experts in choosing the best ideas, a simple model using the least average shrinkage and selection operator can efficiently screen out ideas considered bad by experts. In an additional 22nd hold-out contest with internal and external experts, the simple model does better than external experts in predicting the ideas selected by internal experts. Second, the authors develop an idea screening efficiency curve that trades off the false negative rate against the total ideas screened. Managers can choose the desired point on this curve given their loss function. The best model specification can screen out 44% of ideas, sacrificing only 14% of good ideas. Alternatively, for those unwilling to lose any winners, a novel two-step approach screens out 21% of ideas without sacrificing a single first place winner. Third, a new predictor, word atypical- ity, is simple and efficient in screening. Theoretically, this predictor screens out atypical ideas and keeps inclusive and rich ideas.

History: Olivier Toubia served as the senior editor.
Supplemental Material: The web appendix and data are available at https://doi.org/10.1287/mksc.2023. 1434.

Keywords: creativity • ideation • crowdsourcing • prototypicality • word atypicality • LASSO • random forest • RuleFit • AI • natural language processing

1. Introduction

Artificial intelligence (AI) faces exciting prospects. It is transforming existing business tasks to be done faster, cheaper, and with higher quality. Managers consider AI to be the most important general-purpose technology of our times (Brynjolfsson and McAfee 2017). AI has the potential to change industries just as the internet did 30 years ago or electricity did 100 years ago (Fü ller et al. 2022). In innovation, AI challenges what has been taken for granted (Cockburn et al. 2019). Currently, innova- tion managers see the huge potential of AI-assisted methods (Fü ller et al. 2022) but are uncertain how it can help in ideation and idea screening.

Idea generation and screening are fundamental to marketing success because they are the start of a new product (Toubia and Flores 2007). They belong to the "fuzzy front end," a key point of leverage in the strategy of the firm (Dahan and Hauser 2001, Eling et al. 2014). Crowdsourcing taps diverse information sources and generates a high volume of ideas at low cost (Terwiesch and Ulrich 2009). Commercial crowdsourcing platforms offer ideation-related services (Luo and Toubia 2015).

The large group of ideas includes many redundant or poor ideas. Thus, crowdsourcing presents a new challenge in ideation: screening ideas to identify the best. This screening process can be performed by users, that is, contest participants, Amazon Mechanical Turk work- ers, or experts. Relying solely on contest participants for screening can be problematic because they may act strategically. For example, participants may vote down good ideas that compete with their own. Using Amazon Mechanical Turk workers can be problematic because of their potentially low expertise for specialized contests. The more "cutting edge" a product is, the more likely experts are required (O'Quin and Besemer 1999). The use of experts is very costly because of their limited number, cognitive capability (Toubia 2006), attention span, or objectivity. A solution to this problem may be to let many experts each screen few ideas. Such work division may be effective (Toubia and Flores 2007).

However, this approach may still be insufficient if the number of ideas is very large, which is common in crowdsourcing.

Any idea judged to be high quality by experts is taken to the next stage of development. Costs increase exponentially as ideas progress through the new product funnel from generation and screening to development, prototyping, market testing, and commercialization. Errors in screening can end up being very costly for firms. Urban and Katz (1983) gave the following justification for the cost of ASSESSOR, a tool they created for screening new products: "[It is] a screening device intended to eliminate product failures at a low cost ($50,000) rather than carrying them on to test market where they would at a high cost ($1-2M)." Thus, idea screening is critical to reducing large future costs of new products.

Besides the limited number of experts, their capacity to judge is also limited. When faced with many ideas, tedium sets in, and the judgment of experts may fluctu- ate or deteriorate. In this context, idea screening can be valuable as it allows experts to focus on a smaller num- ber of the best ideas. Any mechanism used to screen ideas carries type I and II errors. A type I error means wrongly selecting a bad idea (potential loser), whereas a type II error means wrongly screening out a good idea (potential winner). In companies, managers have no choice but to accept some errors. In the words of Urban and Katz (1983), "The manager's task, therefore, is to set GO/NO cutoff values that balance these errors and maximize the firm's expected profit." This implies that tools that can aid in the setting of such cutoff values are of substantial value to managers.

AI models may aid in screening at the idea stage of the new product development process when the new product funnel is widest. AI models (Goodfellow et al. 2016) have several advantages over human experts. First, once developed, AI models are relatively low cost to operate. Second, they do not share internal biases or succumb to adverse incentives. Third, they are private, so firms can use them as decision aids without disclosing sensitive intellectual property to third parties. Fourth, they do not tire. Fifth, theory-based AI models are transparent and not black boxes.

An entirely different approach is to use multiarmed bandits (e.g., Jamieson et al. 2015, Jain et al. 2017, Sievert et al. 2017, Katariya et al. 2018). This study is an early attempt in AI to evaluate ideas based on theory-based models. Theory allows for generalizability of findings to other data sets because it provides an understanding of why ideas are good or not. The models tested also have low data requirements: most need only the text of the ideas to work. In sum, the theory-based models have at least four advantages over bandits: they are instantaneous, generalizable, low cost, and suffer no confidentiality concerns because managers and researchers do not have to show the ideas to outside experts for judgments. The last advantage is important, for example, when companies search for innovations of high confidentiality and/or high strategic performance.

Screening is the first of three levels at which AI can help in ideation. Ideators are still needed for idea gener- ation and experts for idea selection. The second level is selecting the best ideas, thus bypassing experts altogether. Ideators are still needed for idea generation, but machines can replace humans for idea selection. The third level is generating the best ideas. This third level, if automated, would eliminate the need for ideators and make crowdsourcing obsolete.

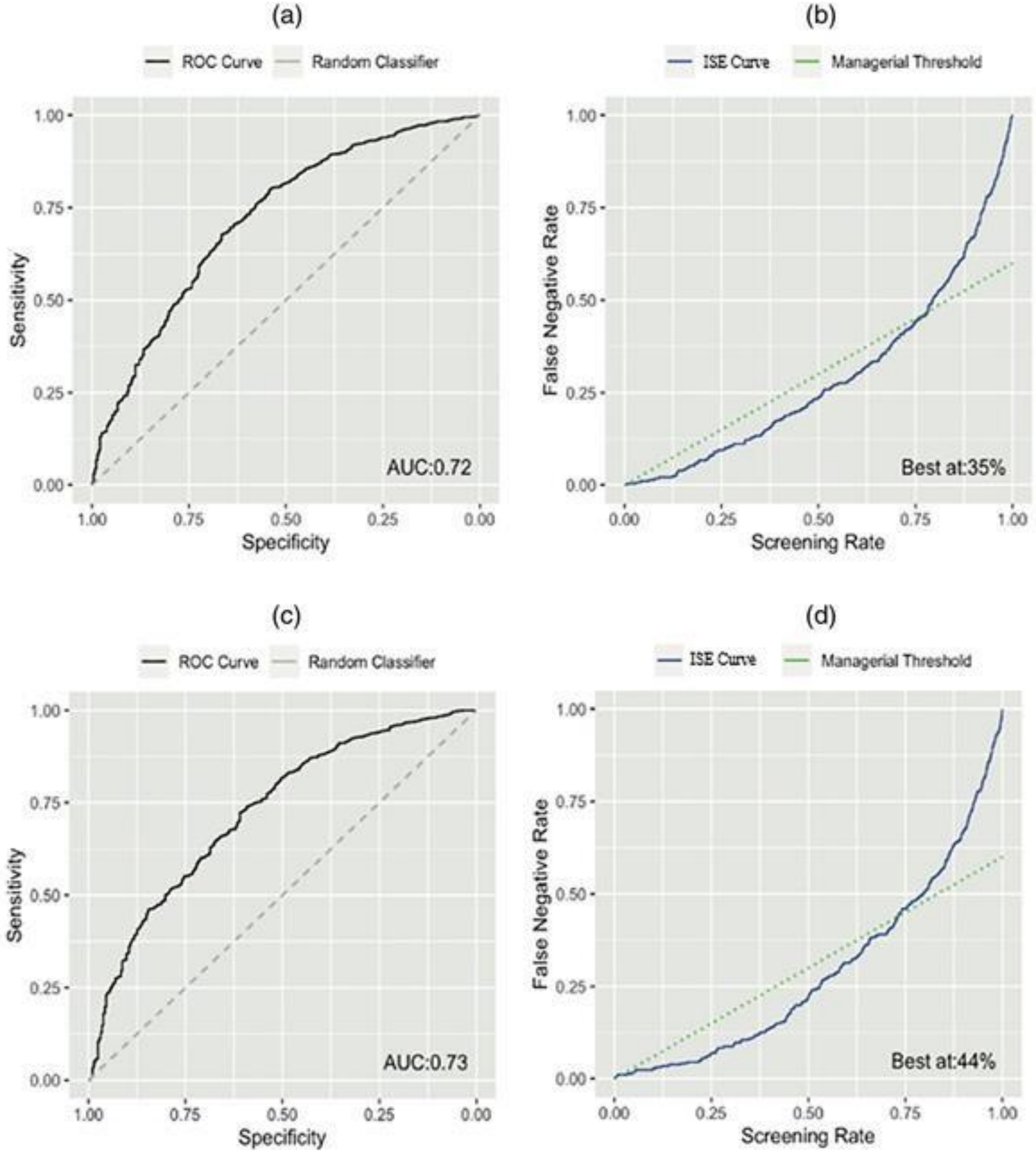
The data for our study come from Hyve, an innova- tion company that runs a crowdsourcing platform for idea generation and selection. We asked the crowdsour- cing platform's director to specify a threshold of accu- racy that would satisfy Hyve's clients, that is, for a useful cost function. He gave us two thresholds of accu- racy: screen out 25% of all ideas without sacrificing more than 15% of good ideas or screen out 50% of all ideas without sacrificing more than 30% of good ideas. We use Hyve's two criteria to construct a reference line we then compare with our proposed idea screening effi- ciency (ISE) curve. This curve plots the false negative rate (good ideas wrongly sacrificed) against the percent- age of all ideas screened out (see Figure 1(b) and (d)). The maximal distance between both is the optimal screening rate.

To identify an AI model for idea screening, this paper tests the out-of-sample performance of eight model specifications for idea selection, including those from three previously published theory-based models: word colocation (Toubia and Netzer 2017), topic atypicality (Berger and Packard 2018), and inspiration redundancy (Stephen et al. 2016). Here, we provide a brief intuition for each theory-based model and detail them in the "theory" section. The intuition of word colocation is that good ideas balance novelty and familiarity. The intuition of topic atypicality (designed for song lyrics) is that good ideas differ from other ideas (the typical) in the same contest. The intuition of inspiration redundancy is that good ideas come from ideators with diverse connections who provide less redundant sources of inspiration.

We test the models with their original predictors and new ones we develop. Most importantly, we test these models out of sample, whereas in their original exposition, they were tested in sample. In-sample testing may exploit sample idiosyncrasies rather than underlying patterns of creativity. We use four methods for testing the models: the least average shrinkage and selection operator (LASSO), Bayesian stacking, random forest, and RuleFit. We also tested multiarm bandits (see Web Appendices 1 and 2).

The specific goals of this paper are the following. (1) Asses how AI can assist managers in idea screening by

Figure 1. (Color online)



Sensitivity = True Positives / All Positives
Specificity = True Negatives / All Negatives
False Negative Rate = False Negatives / All Positives = 1 - Sensitivity
Screening rate = Number of Ideas Dropped / Number of All Ideas

Notes. (a) Model with three theoretical predictors: ROC curve. (b) Model with three theoretical predictors: idea screening efficiency curve. (c) Model with all predictors from Table 2: ROC curve. (d) Model with all predictors from Table 2: idea screening efficiency curve.

studying to what extent one can replace expert evaluations with models. In particular, we compare the performance of three published theory-based models using out-of-sample prediction. (2) Identify simple models and predictors for idea screening if any. (3) Compare out-of-sample prediction performance of four methods: LASSO, Bayesian stacking, random forest, and RuleFit for testing models. We test the models′ performance on 21 different real-world crowdsourcing contests con- ducted for large firms. The pooled data contains 4,191 ideas from 1,467 ideators. We also test on a 22nd hold- out contest with internal and external experts.

This study has three major findings. First, whereas even the best previously published theory-based mod- els cannot mimic human experts in choosing the best ideas, a simple model using LASSO can efficiently screen out ideas considered bad by experts. In an addi- tional 22nd hold-out contest with internal and external experts, the simple model does better than external ex- perts in predicting the ideas selected by internal experts. Second, the authors develop an idea screening efficiency curve that trades off the false negative rate against the total ideas screened. Managers can choose the desired point on this curve given their loss function. The best model specification can screen out 44% of ideas, sacrific- ing only 14% of good ideas. Alternatively, for those unwilling to lose any winners, a novel two-step app- roach screens out 21% of ideas without sacrificing a sin- gle first place winner. Third, a new predictor, word atypicality, is simple and efficient in screening. Theoreti- cally, this predictor screens out atypical ideas and keeps inclusive and rich ideas. These three findings provide methodological, substantive, and managerial contributions, respectively, to the literature on ideation. The rest of the paper is in the following seven sections: litera- ture, model, data, method, results, analysis of extended data set, and discussion.

2. Literature

Our work relates to the literature on crowdsourcing (Stephen et al. 2016, Toubia and Netzer 2017, Allen et al. 2018), in which screening large numbers of ideas can provide huge efficiencies. The three theoretical models of interest are word colocation (Toubia and Netzer 2017), topic atypicality (Berger and Packard 2018), and inspiration redundancy (Stephen et al. 2016). Table 1 provides an overview of these original models, our extensions of them, and their respective intuitions. We first review the theory underlying these models to guide our research.

2.1. Word Colocation

The internet contains a large amount of freely accessible text. One important contribution of Toubia and Netzer (2017) is to show how publicly available data can be used to potentially automate idea evaluation. Their metrics access "global" information (i.e., information not specific to the evaluation context) to assess idea quality. To apply this information, for example, to eval- uate ideas, individuals need to categorize this informa- tion. Prototype theory (e.g., Mervis and Rosch 1981), which draws on the concept of atypicality in semantic categories (Rosch et al. 1976), provides a good categori- zation approach.

2.1.1. Novelty vs. Familiarity. Many new ideas and concepts are the outcome of a process of combination and reorganization of existing ideas and concepts (Mobley et al. 1992). Innovativeness consists of reassembling elements from existing knowledge bases in a novel fashion (Dahl and Moreau 2002). Thereby, within an idea, a moderate level of incongruity between the concepts that make up the idea can be beneficial (Finke et al. 1992). Research has identified a theory about optimal levels of incongruity: the concept of familiarity vs. novelty (Tou- bia and Netzer 2017) draws on the cognitive perspective of innovativeness (Dahl et al. 1999, Goldenberg and Mazursky 2002), which asserts that evaluators rely on information stored in their memories to judge ideas (Toubia and Netzer 2017). When individuals perceive stimuli related to their knowledge, the stimuli activate their domain-specific schemas (Bilalic´ et al. 2008). Because of schema activation, Toubia and Netzer (2017) argue that, if an idea is too novel, its evaluation takes place largely in a vacuum, and the evaluator will not know how to judge it; if an idea is too familiar, it seems to be rather incremental—not new or interesting at all. Thus, experts rank highest those ideas that optimally balance novelty and familiarity based on their current knowledge.

With modern text-mining methods, word colocation networks can be constructed in seconds. In a word colocation network, the vertices are words (or, technically, word stems or word lemmas), and edges indicate co-occurrence. Words that appear together more frequently have higher edge weights and are, therefore, "closer" to each other (Netzer et al. 2012). Toubia and Netzer (2017) use the group of edge weights in an idea to measure its balance of novelty compared with familiarity. A key point is that it is not the words that directly determine novelty or familiarity, but instead, the combinations of words within an idea. Thus, for Toubia and Netzer (2017), an idea is novel to the degree it contains words that typically do not appear together. It is familiar to the degree that it contains words that frequently appear together.

Toubia and Netzer (2017) type the problem description of the contest in Google and use Google Search results to construct a word colocation network. This approach implicitly assumes that Google Search results represent the popularity of the respective website and its content among the crowds because the search results are rank-ordered based on the activity of a very large group of users. Thus, a word colocation network com- puted from averaging over high ranking (roughly top 50) results, provides information on whether ideas that use certain pairwise word combinations are thought to con- tain a desirable balance between novelty and familiarity.

2.2. Topic Atypicality

Atypicality is a construct that deals with the uniqueness of an idea relative to a set of ideas (Berger and Packard 2018). Besides innovativeness, the communication of an idea is crucial for success (Runco 1995, Kilgour et al.

Table 1. Intuition of Original and Extended Models

	Intuition of original approach	Original approach	Intuition of extension	Extension
Word colocation (Toubia and Netzer 2017)	Semantic network approach: Good ideas are prototypical, that is, they balance novelty and familiarity (represented by Google Search results)	Reference corpus 1. Enter contest title and description in Google Search. 2. Take first 50 pages and read html code of these pages. 3. Screen out stopwords from code (e.g., "the" and "and"). 4. Lemmatize the text (only words remain). 5. Build semantic network for each contest. a. Nodes: number of pages, on which respective word was used. b. Edges: number of pages, on which nodes connected by the edge occur jointly. Idea 1. Idea's semantic network uses idea's words and edges from the reference corpus. 2. Calculation of metrics from Table 2. 3. Toubia and Netzer's (2017) key metric rototypicality: a.Computes ECDFs for each document of reference corpus. b. Takes the average of those results. c. Compares ECDF of given idea to average using KS distance (KS distance = maximum absolute difference between both vectors).		
Topic atypicality (Berger and Packard 2018)	Semantic network approach: Ideas are better if they are different from other ideas submitted to the contest	Implemented analogously to Berger and Packard (2018). 1. LDA generates a topic model from corpus of ideas. 2. LDA model output: set of linear equations, one for each topic, equation terms represent words; word terms' coefficients indicate importance of word to topic. 3. Topic atypicality $A_{TA} C(, i :)$distance between idea and location of overall corpus in topic space. For topic l, idea i, reference corpus C $A(C, l, i)_{TA} = \frac{1}{1 - \lvert l(C) - l(i) \rvert = (l(C) + l(i) + :001)}:$ Total topic atypicality, $A_{TA}(C, i)$ is calculated by taking the average of each $A_{TA}(C, l, i)$ over topics l.	LDA extracts popular (common) topics (dimensions), for example, word bundles. LDA may miss unique words as "errors"; successful new product ideas tend to be novel or unique. Metrics that capture word atypicality may be superior.	1 — Jaccard distance between word sets; $W_C =$ set of words in C; W_i $=$set of words in i; word atypicality: $A_{WA}(C, i) = 1 - \frac{W_C \cap W_i}{W_C \cup W_i}:$
Inspiration redundancy (Stephen et al. 2016)	Social network approach: Ideators with access to diverse information, that is, with contacts that don't talk with each other, submit better ideas.	a. Original version of clustering coefficient.Metric calculated on undirected network of comments of ideator n at end of the contest. - Advantage: makes use of all information at the end of the contest to evaluate ideas. - Clustering coefficient for node n in an undirected graph: number of	The Stephen et al. (2016) metric for inspiration redundancy only considers network structure of first degree contacts; operationalization by Burt (2004) allows us to	- a_{nm} indicates existence of link between n and m (one if link present, zero otherwise). - $p_{nm}(t) = a_{nm} = N_n(t)$; $N_n(t)$is neighborhood of n at time t.

Table 1. (Continued)

	Intuition of original approach	Original approach	Intuition of extension	Extension
		edges among n`s neighbors that exist at the end of the contest divided by the number that could exist. - If N_n is neighborhood of n at the end of the contest, m and p are neighbors of n, e_{mp} is an edge between m and p, v_m and v_p are nodes for m and p, and node n has k_n total neighbors, the clustering coefficient CC_n is $CC(n, N_n) = \frac{2 \mid \{e_{mp} : v_m, v_p \in N_n\} \mid}{k_n(k_n - 1)}$: b. Modified version of clustering coefficient.Metric calculated on undirected network of comments of ideator at the point of idea submission. - Advantage: only considers information available at point of idea submission: no endogeneity. - Clustering coefficient for node n in an undirected graph at time t: number of edges among n`s neighbors that exist at the time of idea submission divided by the number that could exist. - If $N_n\ (t)$is neighborhood of n at time t, m and p are neighbors of n, e_{mp} is an edge between m and p, v_m and v_p are nodes for m and p, and node n has $k_n(t)$total neighbors at time t, the modified clustering coefficient CC_n at time t is $MCC(n, N_n(t), t) = \frac{2 \mid \{e_{mp} : v_m, v_p \in N_n(t)\} \mid}{k_n(t)(k_n(t) - 1)}$:	consider second degree contacts.	- Constraint metric $C(i, N_n(t), t) = \sum_{m \neq n} \left(\rho_{nm}(t) + \sum_{p \neq n; p \neq m} \rho_{np}(t) \rho_{pm}(t)\right)^2$. - measure sums across neighbors m of a node n.

2020). Poor communication of an idea makes it hard for external experts to see the idea's merit (Simonton 1999).

The question is whether atypicality is positively or negatively related to idea quality. The literature pro- vides evidence for both. On the one hand, some research suggests that, in music, a creative context, songs with atypical lyrics, that is, lyrics that diverge in content from a genre average, are more likely to become successful (Berger and Packard 2018) because novelty or atypical- ity can increase attention, evaluation, and liking (Ber- lyne 1970, Berger and Packard 2018). If we apply this logic to our setting of idea screening, experts may prefer ideas that are atypical or differentiated from others. On the other hand, other research suggests that genre-typical creative content tends to have a higher quality (Ritchie 2001, Lamb et al. 2015). Evaluations are better if the com- munication is clear and complete (Dean et al. 2006); includes a lot of details (Durand and vanHuss 1992); or is elaborated, that is, understandable, complete, and contains many elements (Besemer and Treffinger 1981). The link is that completeness assists comprehension, which leads to higher judgments of the idea (Sukhov 2018). Research in ideation finds that, if ideators independently come up with similar ideas to a given problem, these typical ideas tend to be better. The reason is that these less atypical, that is, more common, ideas may indicate a widely held need, which indicates market acceptance of the innova- tion; this leads to the fact that more typical ideas tend to have a higher value (Kornish and Ulrich 2011).

2.3. Inspiration Redundancy

The key idea of interconnectivity is to consider the influence on the ideator through a network. If ideas submitted to crowdsourcing contests are visible to other participants, it can inspire and potentially influence them (Wooten and Ulrich 2019). The network structure that surrounds ideators can provide information about the redundancy of their inspirations, which, in turn, influences the quality of the ideas they submit (Stephen et al. 2016).

If the ideator's network neighbors are not connected to one another, ideators receive independent inspira- tions. In contrast, if the network neighbors are con- nected, they also influence each other, and ideators receive similar, redundant inspirations (Burt 2004). Ste- phen et al. (2016) name the following reasons why higher redundancy leads to lower quality ideas: (1) a decreasing size of the set of neighbors' ideas that serve as inspirations when ideating may lead to decreasing innovativeness, (2) idea redundancy could stifle individ- ual innovativeness because it interferes with psychologi- cal mechanisms such as fixation (Bayus 2013) involved in processing others' ideas, and (3) the recurrence of an idea operates as a proof signal. These mechanisms may lead to similar ideas and a decrease in variance of idea quality in the contest. However, a high variance of idea quality is desirable because it increases the likelihood of finding a few outstanding ideas (Terwiesch and Ulrich 2009). Thus, high interconnectivity may relate negatively to idea quality.

3. Theoretical Models

This section describes how we operationalize the three theoretical models: word colocation, topic atypicality, and inspiration redundancy.[1] Two of the theoretical models apply to the text of ideas, and the third one is based on the ideator's commenting network. Each theo- retical model provides a key metric: the Kolmogorov– Smirnov (KS) distance (to measure word colocation; Toubia and Netzer 2017), peer deviation latent Dirich- let allocation (LDA) (to measure topic atypicality; Berger and Packard 2018), and clustering coefficient (to measure inspiration redundancy; Stephen et al. 2016).

3.1. Text Mining

Text mining (Netzer et al. 2012, 2019; Berger et al. 2020) has become increasingly popular as a tool because it helps to detect patterns in large unstructured text corpora, by which one can generate knowledge about consumers (Wedel and Kannan 2016, Matz and Netzer 2017).

3.1.1. Text Preprocessing. In each case, we prepare the text by eliminating "stopwords" or words that appear extremely commonly (e.g., "the" and "and"). We run a process called lemmatization, which stands in place of the traditional approach of stemming. Whereas stemming simply truncates words to reduce duplicates of the same word, lemmatization attempts to remove inflectional end- ings and reduce words to a base dictionary word. Word stems are not always words, whereas word lemmas always are. We chose lemmatization over stemming mainly for convenience in working with the text because it is easier to make sense of word lemmas than word stems. The impact on downstream performance is likely to be very small, but lemmatization may be slightly better (Balakrisnan and Lloyd-Yemoh 2014).

3.1.2. Word Colocation. We apply the work of Toubia and Netzer (2017), which uses several metrics computed from the words used to describe an idea. The main building blocks for these various metrics are word fre- quencies and Jaccard indices. Both word frequencies and Jaccard indices require a reference corpus and a word colocation network to be computed. As men- tioned, Toubia and Netzer (2017) introduce a novel ref- erence corpus consisting of the first 50 Google results when the ideation topic is entered as a search term. The word frequencies are simply the number of times each word in the idea appears in the reference corpus. The Jaccard index, computed on a word pair, is the intersec- tion over the union of documents containing the respec- tive words in the pair. Here, the word "document" refers generally to a body of text. In practice, the researcher specifies the documents. If j and k are words and D_j and D_k are sets of documents containing them, respectively, then the Jaccard index between j and k is

$$J(j, k) = \frac{D_j \cap D_k}{D_j \cup D_k}: \tag{1}$$

For example, if the documents are "one two three," "one two," and "one," then the Jaccard index between the words "one" and "three" is 1/3 because one docu- ment contains both, whereas all three contain at least one of the two. With the Jaccard indices and the word frequencies from each idea, Toubia and Netzer (2017) construct several metrics: the average, max, and min word frequency of an idea; the average max and min Jaccard indices from an idea; the coefficient of variation of word frequencies; and the coefficient of variation of Jaccard indices. A key metric is the KS distance, which is created by first computing empirical cumulative distri- bution functions (ECDFs) for each document of a refer- ence corpus, taking the average of those results, and then comparing the ECDF of a given idea to the average using the KS distance. The KS distance is the maximum absolute difference between two vectors. Toubia and Netzer (2017) show that the KS distance, their metric to balance novelty and familiarity, relates negatively to idea quality even after controlling for other word- derived metrics. Because Jaccard indices measure how often word pairs are collocated, we view the KS distance as capturing the idea's word colocation.

Using each reference corpus, we compute the metrics used in Toubia and Netzer (2017): the mean, min, max, and coefficient of variation for both Jaccard indices and word frequencies of each idea and the KS distance for each idea.

3.1.3. Topic Atypicality.

This model is developed in the spirit of Berger and Packard's (2018) model of content atypicality. Their application is further removed from our ideation context compared with the other two theory-based models: they study innovativeness in a music setting. The dependent variable is song popularity. Berger and Packard (2018) use LDA to generate a topic model from a corpus of song lyrics. LDA "is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities" (Blei et al. 2003). In other words, LDA views topics as having a relationship with individual words: some words are likely or unlikely to appear for certain topics. A common representation for LDA model outputs is a set of linear equations, one for each topic, in which the equation terms represent words. Each word term has a coefficient indicating the importance of that word to the topic. These equations can be inputs into operations. Specifically, a topic score can be computed for a document by applying the model weights against each word in the document. Thinking of the topics as coordinates, each document can be imagined as having a "location" somewhere in topic space. In our case, we create an LDA model for the entire set of ideas in a contest and compute its topic location by applying the LDA weights to its vocabulary. Then, for each idea, we compute the distance between it and the location of the overall corpus in topic space.

Berger and Packard (2018) use Ireland and Pennebaker's (2010) style-matching formula to measure dis- tance, and it is sensitive to the degree of deviation within each topic. Define the "topic function," l, as a mapping from a collection of words into a particular topic. For LDA models, each topic function is simply a weighted average of indicator variables for specific words. Ideas are represented by i with reference corpus C. Then, for a single topic, l, the topic atypicality $A_{TA}(C, l, i)$ is given by

$$A_{TA}(C, l, i) = \frac{1}{1 - |l(C) - l(i)| = (l(C) + l(i) + :001)}: \quad (2)$$

If W_i is the set of words in i and n_l indexes topics, then the topic function $l(i)$ is given by

$$l(i) = \frac{\gamma_{n_l}(W_i)}{\sum_n \gamma_n(W_i)}: \quad (3)$$

Each $\gamma_n(W_i)$ is a posterior parameter measuring one topic probability conditional on the set of words W_i. The posterior distribution of γ_n is conditional on W_i, so it depends not only on the different words in W_i, but also how frequently each occurs.[2] For details, see Blei et al. (2003) and Hoffman et al. (2010). The total topic atypicality, $A_{TA}(C, i)$, is calculated by taking the average of each $A_{TA}(C, l, i)$ over topics l. Mathematically, if N_l is the number of topics,

$$A_{TA}(C, i) = \frac{\sum_l A_{TA}(C, l, i)}{N_l}: \quad (4)$$

In addition to topic atypicality, we develop a new metric, word atypicality A_{WA}. This metric counts the number of words in common between an idea and a reference corpus C divided by the total vocabulary size of C and deducts this value from one. The intuition of word atypicality is that it identifies ideas with unique words; ideas with many unique words have higher values of word atypicality. Ideas that do not contain unique words have a score of zero. Mathematically, word atypicality is

$$A_{WA}(C, i) = 1 - \frac{\underline{W_C \cap W_i}}{W_C \cup W_i}, \quad (5)$$

which is simply one minus the Jaccard index between the word sets. In realistic settings, W_c is a much larger set than W_i, which tends to make the second term small. This means that word atypicality, in turn, is often near one.

Two relevant differences distinguish word atypicality from topic atypicality. First, word atypicality is at the level of words, whereas topic atypicality is specified at the topic level. Topics in LDA come from general pat- terns in the corpus and so are less sensitive or insensitive to unique words. Unique words may well fall into less meaningful topics or be screened out altogether from the analysis. Word atypicality, on the other hand, is exclusively sensitive to unique words. The second dif- ference is that word atypicality measures deviations as binary; either a word overlaps or does not. Topic atypi- cality measures the degree of deviations as a continuous distance. Topic atypicality is sensitive to the number of times a particular word is used, whereas word atypical- ity is not. This feature may mean that topic atypicality is susceptible to noisy outliers, which use certain words with high LDA-topic weights, whereas word atypicality would not be. As word atypicality ends up featuring centrally in the important results, Web Appendix 3 pro- vides some examples that facilitate its understanding.

3.2. Network Metrics

Recent work uses social network data to study innovativeness in ideation (Stephen et al. 2016, Godart and Galunic 2019). We explain the two metrics developed using networks.

3.2.1. Inspiration Redundancy. This model is inspired by Stephen et al. (2016). For each ideator, we create a network in which nodes are ideators and links are com- ments. If one ideator makes a comment on another idea- tor's idea, those two ideators are connected. Otherwise,

they are not connected. Using this network, we compute the clustering coefficient (Watts and Strogatz 1998) for each ideator. The clustering coefficient for a node n in an undirected graph is the number of edges among n's neighbors that exist divided by the number that could exist. If N_n is the neighborhood of n, m and p are neighbors of n, e_{mp} is an edge between m and p, v_m and v_p are nodes for m and p, and node n has k_n total neighbors, the original clustering coefficient CC_n is

$$CC(n, N_n) = \frac{2\,|\{e_{mp} : v_m, v_p \in N_n\}|}{k_n(k_n - 1)} : \qquad (6)$$

Because the unit of analysis is ideas, we then apply the clustering coefficient of each ideator to all the ideator's ideas. This clustering coefficient is computed for the group of comments at the very end of each contest. We view the clustering coefficient as measuring the ideator's inspiration redundancy because it captures the degree of connectivity of each ideator's subcommunity. Because comments reflect the information present in the network surrounding an ideator, the clustering coefficient measures the diversity of information around the ideator. Stephen et al. (2016) show in their context that higher clustering coefficients relate negatively to idea innovativeness as judged by consumers. The reason is that high clustering indicates that the inspiration sources in that area of the network are less diverse. This original version of the clustering coefficient is calculated at the end of the contest but prior to idea evaluation by the experts. Thereby, it makes full use of all information available at the point of evaluating the ideas at the end of the contest. From a managerial perspective, it is feasible to implement (in fact, it is the simplest approach), and it has the potential to predict well.

This original approach to calculating the clustering coefficient may be subject to endogeneity because it considers information that is available after idea submission.[3] To address this limitation, we implement a modified version of the clustering coefficient, which considers information available only until the time of submission and only uses outdegree. If $N_n(t)$ is the neighborhood of n at time t, m and p are neighbors of n, e_{mp} is an (outgoing) edge between m and p, v_m and v_p are nodes for m and p, and node n has $k_n(t)$ total neighbors at time t, the modified clustering coefficient MCC_n at time t is

$$MCC(n, N_n(t), t) = \frac{2\,|\{e_{mp} : v_m, v_p \in N_n(t)\}|}{k_n(t)(k_n(t) - 1)} : \qquad (7)$$

We also use the constraints metric (Burt 2009). This metric is used to measure the popularity of cultural elements in fashion (Godart and Galunic 2019). The clustering coefficient only considers information from the direct neighbors, but the constraints metric also considers second order neighbors (i.e., neighbors of neighbors), which is particularly helpful in the context of sparse networks early in the contests. The constraints metric reflects the value of a particular node as a "bridge" between nodes that are otherwise rarely bridged. In the context of innovativeness, connections between concepts, people, or facts that are not typically connected have higher potential for novelty.

Let a_{nm} indicate the existence of an outgoing link from n to m, so it equals one if ideator n commented on m's idea and zero otherwise. Then, define $\rho_{nm}(t) = a_{nm} = N_n(t)$, where $N_n(t)$ is the neighborhood of n at time t just as earlier. Then, the constraints metric is

$$C(i, N_n(t), t) = \sum_{m \neq n} \left(\rho_{nm}(t) + \sum_{p \neq n; p \neq m} \rho_{np}(t)\rho_{pm}(t) \right)^2 : \qquad (8)$$

This metric sums across neighbors of a node n. For each (outgoing) neighbor m, the constraints metric for node n is larger if n and m have many neighbors in common. The metric is also larger if n has fewer connections. We expect this metric to be negatively related to idea quality as larger values indicate that node n has fewer connections and connects fewer groups of otherwise unconnected nodes.

4. Data

4.1. Data Source and Context

Our data consists of 21 different crowdsourcing contests that a crowdsourcing platform, Hyve, conducted for large corporate clients. Web Appendix 4 provides an overview of the contests. The client firms were Frankfurt Airport, Lufthansa, MasterCard, Deloitte, Telekom, Vodafone, Zeiss, Volkswagen, and DHL. Typically, both Hyve and its clients recruit ideators by public announcements and privately contacting past ideators. The contest usually answers one specific question and runs for a limited time, between 30 and 80 days, with an announced deadline. All contests are idea-generation contests that search for innovative ideas about future products, services, or business models. We only use data from contests in which the ideas are verbally described. This excludes another popular form of contest, design contests (cf. Allen et al. 2018). The contests run on the same platform and offer social networking functions: ideators can explore, evaluate, and comment on the ideas of others. The platform does not allow formal collaboration. The idea evaluation occurs in three stages: experts rate all ideas, experts select a shortlist of ideas to be presented to the jury of client’s managers, jury selects winners.

4.1.1. Expert Ratings. After the contest is over, experts on the topic, usually from the client company, evaluate all ideas on a five-point scale.

4.1.2. Experts’ Shortlist of Ideas. Next, based on their own evaluations, the experts build up a shortlist of 12 to 30 ideas. Usually, the experts’ top-rated ideas make the

shortlist. However, if one expert likes a specific idea well, the expert can discuss it with other experts; if those do not oppose, such an idea can additionally make the shortlist. All shortlisted ideas usually receive at least a small prize or some formal recognition.

4.1.3. Jury’s Selection of Winners.

Finally, the shortlist of ideas as well as a contest overview is presented to a jury of 5 to 10 members, which usually consists of the client’s top executives, experienced innovators, professors, or consultants on the topic of the contest. The jury selects winners in one of two ways. Either the jury members individually vote on the ideas using a scorecard tailored to specific innovation criteria and a simple aggregation determines the winners or the jury jointly selects the winners in a discussion. In addition to the ideas proposed by the experts, jury members have the option to pick and evaluate any idea from the contest in a discussion session.

4.2. Dependent Variable and Predictors

This section provides a description of the dependent and independent variables.

4.2.1. Dependent Variable: Success.

The dependent variable is success. We consider an idea a success if it makes the experts’ shortlist. We choose this shortlist as the success metric for three reasons:

First, every shortlisted idea receives some prize or formal recognition, which means a reward. We consider receiving rewards to be an indication of success.

Second, in private discussions, managers of Hyve stated that they have a high degree of confidence in the shortlist but not in the winner.

Third, the juries select very few winners relative to the many ideas. Ten contests had a total of 12 winners; some contests had more than one winner. Such a rare event in the dependent variable results in a low variance, which models typically cannot capture in a meaningful way. In contrast, the experts’ shortlists consist of up to 30 ideas and, thus, are a richer dependent variable than the jury’s winners. So we code shortlisted ideas as one and other ideas as zero.

4.2.2. Predictors.

Table 2 lists the independent variables, which we explain as follows.

The three theoretical models have sets of metrics, which we include as predictors in our model specifications. Word colocation yields these predictors: max, min, mean, and coefficient of variation for Jaccard indices and node frequencies along with the Kolmogorov−Smirnov distance. Based on topic atypicality, we develop a predictor, word atypicality. The reference corpus for both is the set of other ideas in the same contest as the focal idea. We also create a variant of word atypicality in which the reference corpus is the Google Search results. The inspiration redundancy model yields these predictors: ideator degree, clustering coefficient, and Burt’s constraints metric.

Our goal is to screen ideas out of sample, in which out of sample refers to estimating on 20 contests and predicting on the 21st. Because each out-of-sample prediction task involves a single contest, any contest-level variables are constant across ideas. In other words, contest-level predictors cannot distinguish between ideas within a contest. Therefore, we do not include contest-level variables. Because we need to control for differing numbers of shortlisted ideas and total ideas across contests, we include the ratio of shortlisted ideas to total ideas in each contest as control. This variable is not absorbed in the global intercept, and it is usable for out-of-sampleprediction.

Despite spending months comparing numerous models and methods, we only present the coefficients of two models in Section 6. First, for consistency with prior literature, we present the results of a model that only contains the three original predictors based on the literature. Second, we present the results of a model that contains all predictors from Table 2.

5. Method

Inspired by these theoretical models, we test eight model specifications, using 14 predictors and four methods, on 4,191 ideas from 21 contests. The eight model specifications include predictors from each of the theoretical models, both alone and separately, along with some new predictors developed here.

Table 2. Comprehensive List of Predictor Variables in Various Models

Source	Variables	Computed from which data?
TN	Mean, min, max, and coef. of variation of Jaccard indices between word pairs	Google Search
TN	Mean, min, max, and coef. of variation of node frequencies	Google Search
TN	Kolmogorov−Smirnov distance from Toubia and Netzer (2017)	Google Search
BP	Word atypicality	All ideas from the same contest
BP	Topic atypicality	All ideas from the same contest
SZG	Degree	Comments network
SZG	Modified clustering coefficient (a.k.a. transitivity)	Comments network
SZG	Constraints (Burt’s metric)	Comments network

Notes. TN, Toubia and Netzer (2017); SZG, Stephen et al. (2016); BP, Berger and Packard (2018).

The three models discussed earlier each have one central predictor. Word colocation has the Kolmogorov–Smirnov distance (Toubia and Netzer 2017), inspiration redundancy has the clustering coefficient (Stephen et al. 2016), and topic atypicality has the latent Dirichlet allocation topic distance from Berger and Packard (2018). We develop a new predictor called word atypicality, inspired by topic atypicality. We use word count (number of words in an idea, including stop words) as a naïve bench- mark (Kornish and Jones 2021). We use additional pre- dictors that were either originally included as controls in the papers publishing the three theoretical models or extensions of those theoretical models. Fourteen predictors appear in at least one model specification. Table 2 shows all 14 predictors. The eight model specifications are composed of various

combinations of predictors. Three specifications have one predictor, which is the central predictor from the three theoretical models. A separate model specification uses all three. Word atypicality with two different reference corpora (ideas from the same contest and Google Search) and word count add three more model specifications. The final model specification includes all 14 predictors.

Overall, we train (fit) each model specification on 20 contests and determine performance on the one contest held out. So we have 21 iterations for each model specification and method. This testing amounts to 588 runs (21 contests × 7 specifications × 4 methods). For predictive rigor, all our testing is out of sample and cross-validated.

5.1. Model Specification, Fitting, and Prediction

For all model specifications, the dependent variable is binary whether an idea is on the shortlist or not:

$$y_i = \begin{matrix} 1 \text{ if idea } i \text{ is shortlisted} \\ 0 \text{ otherwise:} \end{matrix} \tag{9}$$

The models vary in specification, depending on which set of the 14 predictors from Table 2 are used. The predictors are characteristics of ideas, ideators, and contests. We do in-sample fitting to estimate the relative standardized coefficients and out-of-sample fitting to ascertain relative performance of models.

To find the most parsimonious set of predictors, we use LASSO, a statistical and AI method (Tibshirani 1996; see also Rafieian and Yoganarasimhan 2021). LASSO has robust performance across many entirely different set- tings (Abadie and Kasy 2019). This feature makes LASSO appealing for our context because we need predictor variables that are robust across contests on varying topics for entirely different clients with distinct judging panels. Web Appendix 5 presents details on LASSO.

Here, we briefly summarize the procedure used to fit the models. We use outer and inner cross-validation loops. The outer loop consists of three steps: designate one contest as the holdout, train the model on the remain- inner loop uses cross-validation to set the LASSO penalty parameter, which controls parsimony. This inner loop cross-validation operates on 20 contests. The inner and outer loops connect through the tuning parameter: for each holdout contest in the outer loop, a corresponding inner loop is used to find the best setting of the tuning parameter (on the nonholdout contests) and make predictions for the holdout. Web Appendix 6 describes this cross-validation procedure step by step.

5.2. Receiver Operating Characteristic (ROC) Curve and Idea Screening Efficiency Curve

5.2.1. ROC Curve and Area Under the ROC Curve.

Removing "bad" ideas requires a high degree of sensitivity (in the technical sense: $\frac{TP}{TP+FN}$; $TP =$ true positives, $FN =$ false negatives) in order not to accidently remove "good" ideas. The main criterion for predictive accuracy out of sample is the ROC curve. The ROC curve is com- monly used in the computer science and information systems literatures. It plots the false positive rate (x-axis) versus the true positive rate (y-axis) for values between zero and one. (See Figure 1(a) and (c)). Our main criterion of predictive accuracy out of sample is the area under the ROC curve (AUC). It provides information on the good- ness of fit of the model, whereby 0.5 indicates not better than random, whereas higher values indicate increasing levels of the model's goodness of fit. AUCs between 0.7 and 0.8 indicate acceptable fit, AUCs above 0.8 excellent fit (Hosmer and Lemeshow 2013).

5.2.2. Idea Screening Efficiency Curve.

In idea screening, Hyve's clients previously had to choose ex ante whether they wanted to minimize false negatives (retain all good ideas), which comes at the cost of high screening effort (potentially very high) or to screen out false posi- ing 20 contests, make predictions for the holdout. The tives (bad ideas), which may come at the sacrifice of eliminating some good ideas. Thereby, the exact preference differed between clients. As an additional flexible criterion, we develop the idea screening efficiency curve, which we plot against the threshold of acceptable performance provided by Hyve's director. This was done to screen out 25% (50%) of bad ideas without sacrificing more than 15% (30%) of good ideas. The ISE curve is a plot of the percentage of all ideas screened out on the *x*- axis against the false negative rate on the *y*-axis (see Figure 1(b) and (d)). The ISE curve has the same shape as the ROC curve (though rotated 180°) but is presented with axes that are more directly interpretable within the context of our problem. The false negative rate is

$$1 - sensitivity = \frac{FN}{FN + TP}: \qquad (10)$$

In words, it is the percentage of shortlisted ideas that are falsely predicted to be nonshortlisted ideas.

The ISE curve is a flexible tool that provides information about all possible trade-offs between any given reduction in screening effort (ideas) versus the respective sacrifice of good ideas. This stands in contrast to methods such as rare events logistic regression (King and Zeng 2001), which applies alternate cutoffs for binary classification when the distribution of positives and negatives is unbalanced. Because the ISE curve plots the results of all possible cutoffs, any decision maker can choose the cutoff that optimizes the trade-off between false negatives and all ideas screened. Beyond crowdsourcing, this ISE curve is useful for any predictive exercise with high imbalance between positives and negatives, in which decision makers differ in their loss functions. It provides a simple elegant visual to trade off false negatives and positives. Such prospects could be ideas, potential consumers, potential new products, or proposals.

5.3. Reference Methods

We also test LASSO against three other methods proposed by or inspired by the reviewers: random forest, RuleFit, and Bayesian stacking. As with Lasso, random forest (Breiman 2001) generally performs well on tabu- lar data and resists overfitting. RuleFit (Friedman and Popescu 2008) is a combination of random forest and LASSO. Bayesian stacking (Yao et al. 2018) constructs a weighted average across different models. Details of each of these methods are in Web Appendices 7–9.

6. Results

Our summary results are the following. First, current models are unable to replace humans in selecting the best ideas. Second, however, these models do an excellent job in screening bad ideas, reducing experts' tedium and enabling their focus on the best ideas. Third, the authors develop an idea screening efficiency curve that relates the false negative rate of good ideas screened out with the rate of ideas screened. Managers can choose the desired point on this curve for optimal idea screening. For example, the best model specification can screen out as much as 44% of bad ideas, sacrificing only 14% of good ideas. A two-step model screens out 21% of the worst ideas without sacrificing a single first place winner. Fourth, a new predictor, word atypicality, is simple and efficient in such screening. Theoretically, this predictor screens out atypical ideas and keeps inclusive and rich ideas that experts rated high. Word count is simpler but does not perform as well, is easy to game, and may be misleading as a metric (Kornish and Jones 2021).

Detailed results follow.

6.1. AUC: Results

6.1.1. In-Sample Estimates of Coefficients. To appreciate effect sizes of coefficients, we first test in sample the specification that includes only three theory-based predictors, pooling all 21 contests. (Out-of-sample tests yield 21 sets of coefficients, one for each contest held out.) To control for contest heterogeneity when pooling contests, we also include the percentage of shortlisted ideas of each contest (% *Shortlisted*). This control is different for each contest but the same for each idea within a contest.

Table 3 shows the standardized coefficients of the LASSO logistic regression estimated in sample for the pooled 21 contests. The model specification includes the three theory-based predictors. LASSO retains all three. This result means the three are complementary, each capturing one unique dimension of the innovative- ness of an idea. The largest standardized coefficient is on the clustering coefficient with a value of —0.19. The next largest is on the Kolmogorov–Smirnov distance calculated on the Google reference network with a value of —0.08. The third is on topic atypicality with a value of —0.07. Notably, the signs of the first two coefficients are in the direction predicted by their original theory. However, the sign for the third, topic atypicality, is opposite to that in the original application (music). The reason may be due to different dependent variables (attention-getting versus creative) and contexts (music versus ideation). In music, novelty is most attention-getting, and so the most atypical song is rated highest (sign is positive). In ideation, the most detailed and comprehensive idea (most inclusive or "typical") is rated highest.

6.1.2. Out-of-Sample Predictive Performance. For predictive rigor, we test the same model specification with the three theory-based predictors out of sample with cross-validation. Web Appendix 6 explains our method for out-of-sample predictions. Figure 1(a) shows that the AUC of the ROC curve has a value of 0.72. Thus, the three original predictors from the theory-based models jointly reach a threshold that is generally considered to be acceptable.

Next, we test the model specification with all 14 predictors from Table 2. Importantly, LASSO retains only word atypicality as a predictor in all 21 contests, sometimes complemented by another predictor. Web Appendix 10 contains some additional information about the performance of word atypicality. Figure 1(c) shows the out-of-sample ROC curve. The AUC is 0.73, which is a

Table 3. Variables Retained by LASSO (in Sample)

Source	Variable name	Standardized coefficient
	Intercept	—3.50
SZG	Original clustering coefficient TN	—0.19
	Kolmogorov–Smirnov (Google)	—0.08
BP	Topic atypicality	—0.07
Control	Percentage shortlisted (contest level)	0.10

Notes. Input: Variables inspired by original models. DV: Shortlisted (yes/no). TN, Toubia and Netzer (2017); SZG, Stephen et al. (2016); BP, Berger and Packard (2018).

little over the value above of 0.72 even though it (usually) contains only one predictor. This result has two important implications. One, that the new predictor we develop, word atypicality, is more powerful in prediction than any other predictor, singly or in combination. Two, word atypicality encompasses the predictor capacity of all three theory-based predictors.

6.2. LASSO's Comparison with Other Methods

We next compare these out-of-sample results with LASSO to results using three other methods, RuleFit, random forest, and Bayesian stacking, for the model specification that includes all 14 predictors. Web Appendices 7–9 provide some additional information on modeling and key results. Random forest has an AUC of 0.69, RuleFit of 0.70, and Bayesian stacking of 0.72. Overall, LASSO does better than the other three methods. Also, the optima of the idea screening efficiency curves for random forest, RuleFit, and Bayesian stack- ing exceed the performance threshold from Hyve by smaller amounts.

Recall that Bayesian stacking is an ensemble that creates a weighted average of models. Bayesian stacking finds the weights of each model by using leave-one-out cross-validation. In the case of Bayesian stacking, we tried creating ensembles from many different models formed by selecting random subsets of predictors. The best configuration we find is an ensemble over three model specifications. Model specification 1 uses the KS distance from word colocation and the clustering coeffi- cient from inspiration redundancy as predictors, model specification 2 uses word count alone, and model speci- fication 3 uses topic and word atypicality. The combination of the predictions from these three model specifications via Bayesian stacking reaches an AUC of 0.72. These results compare with LASSO's AUC of 0.73.

6.3. Optimal Screening Rate on the Idea Screening Efficiency Curve

Figure 1(b) shows the ISE curve Figure 1(b) shows it for the model specification with the three theory-based predictors. Figure 1(d) shows it for the model specification with all predictors from Table 2. The green dotted line shows the managerial threshold given by Hyve: screening 25% of all ideas without losing more than 15% of good ideas or screening 50% of ideas without losing more than 30% of good ideas. In Figure 1(b) and (d), Hyve's standard is met anywhere the solid curve falls below the dotted line, and the optimum point is when the curve is maximally below the dotted line.

Table 4 shows the optimal screening rate for all eight model specifications. For our data and the model specification with all predictors, the best performance screens out 44% of all ideas at the cost of sacrificing only 14% of good ideas. This result exceeds Hyve's standard and indicates that our model can provide a substantial reduction in experts' workload.

The code to run this model is in Web Appendix 13.

6.4. Theory-Based Predictors and Word Atypicality: Substitutes or Complements?

Preliminary results based on Section 6.1 indicate that the three theory-based predictors tend to be complementary, but that word atypicality tends to encompass all three. We analyze the intersection of sets of ideas predicted by various predictors at the top and the bottom to further explore complementarity. For this purpose, we use a method, new to marketing, called the super exact test for efficient testing of multiset interactions (Wang et al. 2015). This method allows for testing the statistical significance of the size of overlap between multiple sets. For each pre- dictor, that is, the three theory-based predictors and word

Table 4. Out of Sample: Eight Model Specifications

Dependent variable Predictors in various models	Shortlist Optimal screening rate/% of good ideas screened out at optimal screening rate	Winner Percent screened before losing winner, %[a]
Word colocation only (TN original)	28/08	15
Topic atypicality only (BP original)	29/11	12
Inspiration redundancy only (SZG original)	24/09	10
Word atypicality only (this study)	40/13	12
Word atypicality (Google)	40/14	15
Word count only (naïve)	40/15	13
Model with three theoretical predictors (word colocation, topic atypicality,	35/13	14

inspiration redundancy)		
Model with all predictors (see Table 2)	44/14	12

Note. TN, Toubia and Netzer (2017); SZG, Stephen et al. (2016); BP, Berger and Packard (2018).

[a]After accounting for percentage of winners in contests.

**AUC of all models > 0.7; even small differences in AUC translate to substantial differences in screening rate, which is the managers’ prime concern.

Table 5. Overlap Between Model's Predictions of Top 25% and Screening of Bottom 25%; Expected Overlapby Chance: 6.25%

	Top 25% of ideas predicted by each model/bottom 25% of ideas predicted by each model			
	1. Word colocation	2. Topic atypicality	3. Inspiration redundancy	4. Word atypicality
1. Word colocation				
2. Topic atypicality	7.4*/6.6			
3. Inspiration redundancy	5.4/7.4*	5.3/5.4		
4. Word atypicality	10.4*/12.6*	9.3*/7.5*	7.4*/8.0*	

Note. Bold indicates that both metrics significantly overlap at the top and bottom.
*$p < 0.05$.

atypicality, we divide the predicted ideas into three categories: top 25% of predicted ideas (to retain), bottom 25% of predicted ideas (to screen out), remaining predicted ideas. Then, we compare the intersections between the top and bottom 25% of predicted ideas by all four predictors.

Table 5 contains the results for overlap at the top (select) and bottom (screen). By chance, the pairwise overlap in sets classified by any two predictors would be 25% × 25% =6.25%. The pairwise overlap between the three theory-based predictors, word colocation, inspiration redundancy, and topic atypicality, indicates that no pair significantly overlaps at the top and bottom. This additional result confirms that the three theory-based predictors are unique, each capturing different dimensions of what experts consider to be good ideas. So they are complementary. On the other hand, the ideas classified by word atypicality overlap significantly with each of the three original models at the top and bottom. This result confirms that, even though word atypicality is parsimonious, it partly captures dimensions that are unique to the three theory-based predictors. Thus, it is a substi- tute to the theory-based predictors.

7. Managerial Relevance

The key to applicability of a model is its relevance for managers. To address this issue, we carry out four additional analyses: screening ideas without losing winners (Section 7.1), a new 22nd contest with multiple internal and external ratings (Section 7.2), and the relationship between theory-based predictors and managerial ratings (Section 7.3). Details of each of these analyses follow.

7.1. Two-Step Approach

The goal of the two-step approach is to screen out the worst ideas without losing a winner. Step 1 scores each idea within each contest on a simple heuristic and screens out ideas that score the lowest. Step 2 runs the best predictive model from Table 4 on the reduced cor- pus of ideas.

7.1.1. Summary of Step 1. Details can be found in Web Appendix 11. We rank-order all ideas on one or more predictors. We test one predictor at a time or combinations of two or three predictors from the 14 predictors in Table 2. We then screen out ideas that fall below a threshold to reduce noise in the corpus of ideas. In the case of one predictor, we screen out the worst ideas on that predictor. In the case of two predictors, we screen out the worst ideas on the rankings of both predictors (i.e., the bottom intersection of two rankings of ideas). In the case of three predictors, we screen out the worst ideas on the rankings of all three predictors (i.e., the bottom intersection of three rankings of ideas). Testing one predictor at a time requires 14 runs. Testing two predictors at a time requires 91 runs. Testing three predictors at a time requires 364 runs. In total, this ranking exercise requires 469 runs. We also test various thresholds for screening out from 5% to 35% of ideas in increments of 5%. This exercise then grows to 3,283 runs (7 thresholds × 469 runs).

7.1.2. Summary of Step 2. On this reduced corpus of ideas, we run our standard out-of-sample LASSO regression from Section 5. For this step, we test all eight model specifications in Table 4.

7.1.3. Results. Table 6 shows the results of the two-step approach. After extensive testing, pairs of predictors work better than single predictors or three predictors. For pairs, the best results occur when using word count and

Table 6. Best Performing Model in Two-Step Analysis

First step (best of 105 possible pairs of first five models from Table 5)	Second step (best of all eight models from Table 5)	Percentage screened before losing winner
Topic atypicality and word count	Word colocation	21

Notes. First step: screen out ideas in bottom 25% according to both predictors. Second step: LASSO with predictors listed in second column.

topic atypicality and a threshold of 25%. Thus, these two predictors reveal enough information about the content of ideas to screen out those that merely add noise. Prior to this, managers knew that the corpus of ideas contained a lot of junk but did not have a simple way to screen that out. Step 1 does not sacrifice any winners when screening out the bottom ranked 8% of ideas. By comparison, word count as a naïve benchmark, sacrifices two winners when screening out the bottom 8% of ideas. Thus, whereas word count is simpler, in this context, it per- forms less well for managers whose cost function for sacrificing winners is steep. In step 2, word colocation performs best of the eight model specifications in Table 4. Step 2 screens out the bottom ranked 13% of ideas in addition to those screened out in step 1. Both steps together yield a screening rate of approximately 21% without sacrificing a single winner.

The code to run the two-step approach is in Web Appendix 14.

7.2. Relationship of Theory-Based Predictors to Managerial Ratings

Theory-based models are most relevant for managers if they relate to managers′ own ratings of ideas. We draw on additional information available in the corpus of ideas′ ratings to identify any such relationship. In each contest, experts rated ideas on various dimensions that differ among contests on a five-point scale from very low to very high. For 11 contests (see Web Appendix 4 for details), we possess information on ideas′ ratings on these dimensions. Managers consider three dimensions relevant in six or more contests: innovativeness of idea (643 ideas), communication of idea (483 ideas), and sales potential (641 ideas).

The cells in Table 7 show Pearson correlation coefficients between each of the theory-based models plus word count in rows and ratings of managers in columns on selected dimensions. Word colocation correlates highest with innovativeness of an idea (*r*:—0.12, $p < 0.05$). Word atypicality correlates highest with the communication of idea (*r*: −0.31, $p < 0.05$) and sales potential (*r*:—0.19, $p <$ 0.05). Inspiration redundancy correlates with innovativeness (*r*:—0.08, p<0.05) and sales potential (*r*:—0.08, $p < 0.05$). The naïve word count correlates with each dimension (*r* (innovativeness of idea):—0.07, $p < 0.05$; *r* (communication of idea): 0.15, $p < 0.05$; *r* (sales potential): 0.10, $p < 0.05$). Overall, selected dimensions of managerial ratings correlate almost twice as highly with theory-based predictors than with naïve word count.

7.3. Application to a New Client

A new, large consumer goods client of Hyve provided a new contest of 2,947 ideas for evaluation. Internal experts from the company evaluate all ideas in an extensive process and shortlist 54 ideas. From these 54 ideas they select five ideas for funding, which we call winners.

After the internal evaluation, the division head feels that the company could make even better use of the contest′s ideas. The company pays Hyve to reevaluate all ideas and to select any number of ideas the experts consider to be good. Hyve′s experts shortlist 1,125 ideas. Hyve′s experts had no knowledge of the ideas selected by the internal experts, so both evaluations are independent.

We apply our AI models to screen ideas for this contest. We test all eight model specifications from Table 4 out of sample. We report the AUC, optimal screening rate, and percentage of ideas screened before sacrificing a winner.

Table 8 shows the results. The AUC for the model with all predictors is 0.72, as with that obtained from the pooled 21 contests. The model specification with only word atypicality screens out 61% of all ideas, sacrificing 24% of shortlisted ideas. The same model specification also screens out 62% of all ideas without sacrificing any one of the five winners. Screening out 62% of ideas with- out losing a winner in this 22nd contest is much higher than is possible for the prior pooled 21 contests (Table 4). This improvement in performance is perhaps because the number of ideas in this 22nd contest is much higher

Table 7. Managerial Relevance of the Theory′s Metrics

		Managerial dimensions		
		Innovativeness of idea (643 ideas)	Communication quality of idea (483 ideas)	Sales potential (641 ideas)
Source theory	Metric	Pearson correlation coefficients		
Word colocation (TN)	Kolmogorov−Smirnov	—0.12*	—0.22*	—0.04
Topic atypicality (BP)	Peer deviation LDA	0.03	—0.03	0.02
Word atypicality (SZG)	Peer deviation Jaccard	—0.07*	—0.31*	—0.19*
Naïve model	Word count	0.07*	0.15*	0.10*
Inspiration redundancy	Clustering coefficient	—0.08*	0.03	—0.08*

Note. TN, Toubia and Netzer (2017); SZG, Stephen et al. (2016); BP, Berger and Packard (2018). *p < 0.05.

Table 8. Out-of-Sample: Predictors Included in Models: Internal Contest

Dependent variable Models	Shortlist Optimal screening rate/% of good ideas screened out at optimal screening rate	Winner Percent screened before sacrificing a winner, %[a]
Word colocation only (TN original)	37/15	44
Topic atypicality only (BP original)	11/4	5
Inspiration redundancy only (SZG original)	Unavailable	—
Word atypicality only (this study)	61/24	62
Word atypicality (Google corpus)	59/22	69
Word count only (naïve)	58/24	67
Model with three theoretical predictors (word colocation, topic atypicality, inspiration redundancy)	11/4	8
Model with all predictors (see Table 2)	61/24	62

Note. TN, Toubia and Netzer (2017); SZG = Stephen et al. (2016); BP, Berger and Packard (2018).

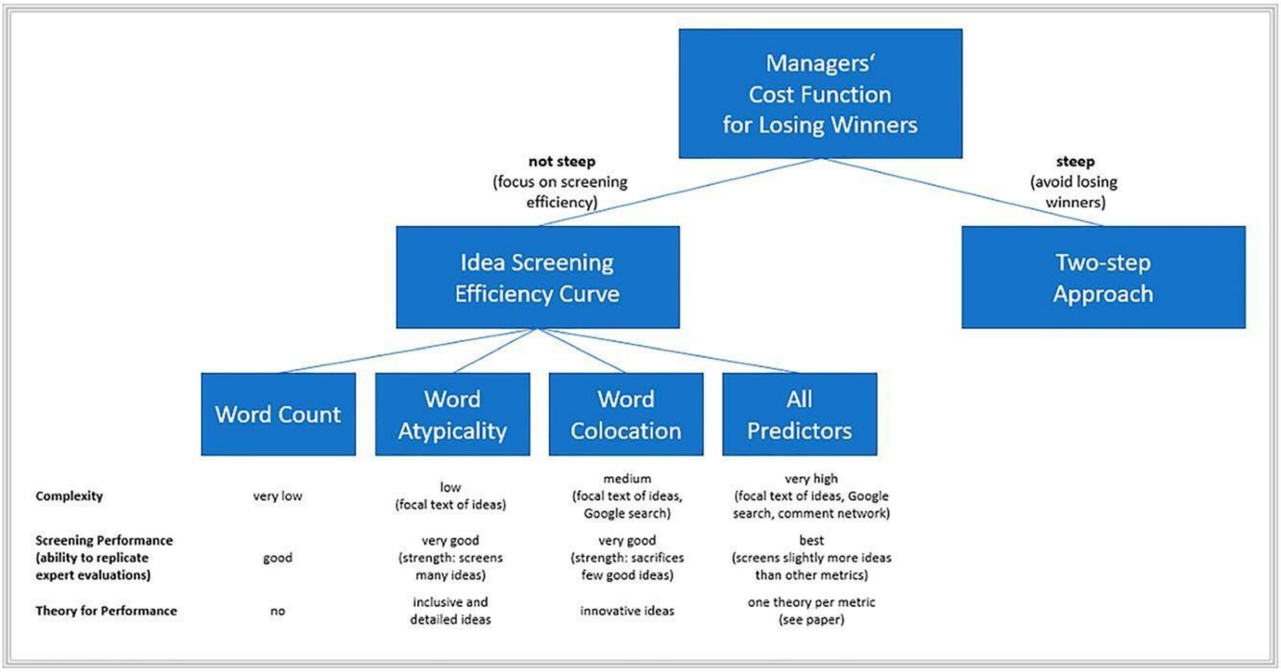
and the percentage of winners much lower than in any prior contest.

Importantly, in comparison with our models, Hyve's experts screen out 62% of all ideas, sacrificing four of five winners. This application provides preliminary evidence that the proposed model not only does well, but it does better than external experts. If subsequent research supports this finding, this means that innovation managers and scientists soon have a low-cost tool that can replace work that nowadays is often outsourced to externals such as Amazon MTurk workers or that cannot be done properly at all, for example, because of confidentiality reasons. Because its predictors come from theory and are clearly defined, the results of the model are more transparent and explainable than are ratings of experts.

7.4. Recommendation Scheme for Managers

Figure 2 provides an overview of our recommendations to managers. If managers seek to avoid losing winners in return for lower screening efficiency, they should use the two-step approach (see Web Appendix 14 for code to run the two-step approach). If, instead, they focus on decreasing effort evaluating numerous ideas, they should use LASSO based on various metrics to generate the ISE curve (see Web Appendices 13 and 15 for code to do the LASSO and the ISE curve, respectively). The simplest predictor is word count, which has almost the same accuracy as other predictors. It screens out ideas that are short. However, word count is easy to game. If ideators know that is the rule, they can write ideas that are long and wordy. If managers are willing to invest a little extra effort for interpretable results, they should use one of two

Figure 2. (Color online) Recommendation Scheme for Managers to Screen Ideas


Managers' Cost Function for Losing Winners
not steep
(focus on screening efficiency)
steep
(avoid losing winners)
Idea Screening Efficiency Curve
Two-step Approach
Word Count
Word Atypicality
Word Colocation
All Predictors
Complexity
very low
low
(focal text of ideas)
medium
(focal text of ideas, Google search)
very high
(focal text of ideas, Google search, comment network)
Screening Performance
(ability to replicate expert evaluations)
good
very good
(strength: screens many ideas)
very good
(strength: sacrifices few good ideas)
best
(screens slightly more ideas than other metrics)
Theory for Performance
no
inclusive and detailed ideas
innovative ideas
one theory per metric
(see paper)

predictors: (1) word atypicality, which screens out atypical ideas and keeps inclusive and rich ideas, or (2) word colocation, which retains ideas that use novel words. If managers want the best screening to reduce the time and effort of experts, they should implement a model with all 14 predictors. We presented our results to innovation managers who typically preferred more screening efficiency (word atypicality); Web Appendix 12 contains procedure and results.

8. Discussion

Crowdsourcing generates up to thousands of ideas per contest. The selection of best ideas is costly because of the limited number, objectivity, and attention of experts. Using a data set of 21 crowdsourcing contests that include 4,191 ideas, we test how AI can assist experts in screening ideas.

8.1. Summary of Results and Contribution

This study has three major findings. First, whereas even the best previously published theory-based models cannot mimic human experts in choosing the best ideas, a simple model using LASSO can efficiently screen out ideas considered bad by experts. In an additional 22nd hold-out contest with internal and external experts, the simple model does better than external experts in predicting the ideas selected by internal experts. Second, the authors develop an idea screening efficiency curve that trades off the false negative rate against the total ideas screened. Managers can choose the desired point on this curve given their loss function. The best model specification can screen out 44% of ideas, sacrificing only 14% of good ideas. Alternatively, for those unwilling to lose any winners, a novel two-step approach screens out 21% of ideas without sacrificing a single first place winner. Third, a new predictor, word atypicality, is simple and efficient in screening. Theoretically, this predictor screens out atypical ideas and keeps inclusive and rich ideas. These three findings provide methodological, substantive, and managerial contributions respectively to the literature on ideation.

8.2. Questions and Answers

We now answer questions raised by this research.

First, why does AI do better in screening bad ideas than selecting winners? We suggest three possible reasons. One, our contests are blessed with rich data consisting of many long ideas. Word atypicality works by screening out short, poorly developed ideas with unique words that may not relate to the client’s problem. Two, we derive the semantic network based on Google Search results of the first 50 results pages when typing in the contest topic (Toubia and Netzer 2017). Words from the contest description may have central positions in the semantic network. Thus, typical ideas tend to be more “on topic” than atypical ideas. Three, experts prefer complete, elaborated, and detailed idea descriptions (Besemer and Treffinger 1981, Dean et al. 2006) potentially because details facilitate knowledgeable experts to assess idea quality (Sukhov et al. 2021).

Second, can AI replace humans in ideation? At the current stage of assessing ideas, AI cannot fully or even partially substitute for human experts. However, AI can assist humans, such as the platform’s managers, by screening out bad ideas (level 1 of AI in ideation), thereby reducing humans’ cognitive load and helping them focus on the good ideas. The current study is a first step to narrow the ideation funnel early on. Yet, as research advances and additional models bring in unique and helpful perspectives, research may develop models to select the best ideas (level 2 of AI in ideation) or even to automatically generate outstanding ideas (level 3 of AI in ideation).

Third, how does LASSO compare with multiarmed bandits? Prior methods from computer science deal with distinguishing good from bad ideas based on model fitting on massive data. One important example is the use of multiarmed bandits (e.g., Jain and Jamieson 2018, Jamieson and Jain 2018, Fiez et al. 2019, Auer 2002), which deals with efficiently allocating evaluations across ideas (arms) to get the most learning done. Whereas evaluations are easy and cheap in settings in which lay people can form opinions quickly, for example, liking of designs, they are costly in settings with scarce experts. In such settings, our approach uses theory to understand the underlying patterns of expert judgments. In fact, the strength of theory-based predictors is to screen out bad ideas. So contest managers can first use our theory-based predictors to reduce the number of ideas for experts to judge.

Fourth, why does AI fare much better in the context of chess, face recognition, and even music (e.g., Berger and Packard 2018) than in ideation? Chess has fixed precise rules for movement of pieces, very clear payoffs, and one goal, which enables models to find the best move by elimination of alternatives. Face recognition inputs millions of faces, which lets models identify a few patterns. Music has a finite number of characteristics and patterns that models can easily pick up. In contrast, ideation has a few winners characterized by a vast number of opaque dimensions.

8.3. Limitations and Future Research

This study has limitations that future research may fruitfully address. First, as with all such contests, our contests suffer from survival bias: firms do not commercialize ideas that were not shortlisted. So one never knows their market success. However, using real expert evaluators from companies goes further than prior studies (e.g., Stephen et al. 2016, Toubia and Netzer 2017). Second, there could be a common bias among experts and AI. For example, both might be biased by ideas that are described more clearly, more eloquently, and with love for detail. We cannot tell based on the data at hand. Third, because we use AI-based approaches based on text complemented by inspiration redundancy (Stephen et al. 2016), we do not include additional information, for example, from images (we exclude design contests), behavioral data, or background information about the ideators, in our analysis. Fourth, during the screening process, experts often refine ideas. Whereas the raw idea might be unattractive per se, the refined ones might be innovative. Our screening approach is performed at the end of the ideation contests. We do not know and can certainly not rule out whether ideas that were screened out by the experts during the screening process could have been refined to make them innovative in later stages of the new product development process. Fifth, in line with some studies from Toubia and Netzer (2017), we assume ideas to be good if experts from Hyve's clients selected them to be presented to a jury of decision makers. As such, our approach mimics the unknown decision criteria these human experts implicitly apply. Whereas using data from various contests shields our results to some degree against idiosyncrasies of a specific data set, we cannot rule out that our models and the experts might be biased in the same way. Therefore, our models can assist human experts in what they would be doing otherwise themselves, but we cannot make the claim that our models necessarily identify the best or most innovative ideas. Ideally, the best approach to identify the best ideas would be to develop each idea into a new product and to test its success in real markets. However, in the context of thousands of ideas, many of which are poorly developed and/or similar to others, this approach is prohibitively expensive and practically infeasible.

Endnotes

[1] We attempt to operationalize the theoretical models in the same way as the published research when possible, but in some cases, our implementation differs slightly because of context.

[2] The LDA inference algorithm we use is from Hoffman et al. (2010) through the Python package Gensim (Rehurek 2011). This algorithm uses variational Bayes, in which the $\gamma_n(W_i)$s parameterize the distribution approximating the posterior over "per-document topic weights."

[3] This concern is not relevant in the experimental approach by Stephen et al. (2016) in which the ideators' network position was induced by the experimental design, but it may become problematic in our study, which uses field data when the comment network evolves over time. Additionally, we use comments as a proxy for exposure, whereas the Stephen et al. (2016) paper uses real links between individuals.

References

Abadie A, Kasy M (2019) The risk of machine learning. *Rev. Econom. Statist.* 101(5):743–762.

Allen BJ, Chandrasekaran D, Basuroy S (2018) Design crowdsourcing: The impact on new product performance of crowdsourcing design solutions from the "crowd." *J. Marketing* 82(2):106–123.

Amatriain X, Lathia N, Pujol JM, Kwak H, Oliver N (2009) The wisdom of the few: A collaborative filtering approach based on expert opinions from the web. *Proc. 32nd Internat. SCM SIGIR Conf.*, 532–539.

Auer P (2002) Using confidence bounds for exploitation-exploration trade-offs. *J. Machine Learn. Res.* 3:397–422.

Bayus BL (2013) Crowdsourcing new product ideas over time: An analysis of the Dell Ideastorm community. *Management Sci.* 59(1):226–244.

Berger J, Humphreys A, Ludwig S, Moe WW, Netzer O, Schweidel DA (2020) Uniting the tribes: Using text for marketing insight. *J. Marketing* 84(1):1–25.

Berger J, Packard G (2018) Are atypical things more popular? *Psych. Sci.* 29(7):1178–1184.

Berlyne DE (1970) Novelty, complexity, and hedonic value. *Perception Psychophysics* 8:279–286.

Besemer SP, Treffinger DJ (1981) Analysis of creative products: Review and synthesis. *J. Creative Behav.* 15(3):158–178.

Bilalić M, McLeod P, Gobet F (2008) Inflexibility of experts—Reality or myth? Quantifying the Einstellung effect in chess masters. *Cognitive Psych.* 56(2):73–102.

Bishop CM (2006) *Pattern Recognition and Machine Learning* (Springer).

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. *J. Machine Learn. Res.* 3:993–1022.

Breiman L (2001) Random forests. *Machine Learn.* 45(1):5–32.

Brynjolfsson E, McAfee A (2017) The business of artificial intelligence: What it can—and cannot—do for your organization. *Harvard Business Review Online* (July 18), https://hbr.org/2017/07/the-business-of-artificial-intelligence.

Burt RS (2004) Structural holes and good ideas. *Amer. J. Sociol.* 110(2):349–399.

Burt RS (2009) *Structural Holes: The Social Structure of Competition* (Harvard University Press, Boston).

Cao W, Li J, Tao Y, Li Z (2015) On top-k selection in multi-armed bandits and hidden bipartite graphs. *NIPS* 8(1):3–32.

Cockburn I, Henderson R, Stern S (2019) The impact of artificial intelligence on innovation: An exploratory analysis. Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence* (University of Chicago Press, Chicago), 115–148.

Dahan E, Hauser JR (2002) Product development—Managing a dispersed process. Weitz B, Wensley R, eds. *Handbook of Marketing* (Sage Publication, New York), 179–222.

Dahl DW, Chattopadhyay A, Gorn GJ (1999) The use of visual mental imagery in new product design. *J. Marketing Res.* 36(1):18–28.

Dahl DW, Moreau CP (2002) The influence and value of analogical thinking during new product ideation. *J. Marketing Res.* 39(1):47–60.

Dean DL, Hender JM, Rodgers TL, Santanen E (2006) Identifying good ideas: Constructs and scales for idea evaluation. *J. Assoc. Inform. Systems* 7(10):646–699.

Eling K, Griffin A, Langerak F (2014) Using intuition in fuzzy front-end decision-making: A conceptual framework. *J. Product Innovation Management* 31(5):956–972.

Fiez T, Jain L, Jamieson K, Ratliff L (2019) Sequential experimental design for transductive linear bandits. *33rd Conf. Neural Inform. Processing Systems*, 10667–10777.

Friedman J, Hastie T, Tibshirani R (2001) *The Elements of Statistical Learning*. Springer Series in Statistics, vol. 1, no. 10 (New York).

Friedman JH, Popescu BE (2008) Predictive learning via rule ensembles. *Ann. Appl. Statist.* 2(3):916–954.

Füller J, Hutter K, Wahl J, Bilgram V, Tekic R (2022) How AI revolutionizes innovation management—Perceptions and implementation preferences of AI-based innovators. *Tech. Forecasting Soc. Change* 178:121598.

Giora R (2003) *On Our Mind: Salience, Context, and Figurative Language* (Oxford University Press, New York).

Godart FC, Galunic C (2019) Explaining the popularity of cultural elements: Networks, culture, and the structural embeddedness of high fashion trends. *Organ. Sci.* 30(1):151–168.

Goldenberg J, Mazursky S (2002) *Creativity in Product Innovation* (Cambridge University Press, Cambridge, MA).

Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning* (MIT Press, Cambridge, MA).

Hill S, Ready-Campbell N (2011) The wisdom of (experts in) crowds. *Internat. J. Electronic Commerce* 15(3):73−101.

Hoffman M, Bach F, Blei D (2010) Online learning for latent Dirich- let allocation. *Adv. Neural Inform. Processing Systems* 23.

Hosmer DW, Lemeshow S (2013) *Applied Logistic Regression*, 2nd ed. (John Wiley & Sons, New York).

Ireland ME, Pennebaker JW (2010) Language style matching in writ- ing: Synchrony in essays, correspondence, and poetry. *J. Person- ality Soc. Psych.* 99(3):549−571.

Jain L, Jamieson K (2018) Firing bandits: Optimizing crowdfunding. *Proc. 35th Internat. Conf. Machine Learn.*, 2206−2214.

Jain L, Mason B, Nowak R (2017) Learning low-dimensional metrics. *31st Conf. Neural Inform. Processing Systems*, 4139−4147.

Jamieson K, Jain L (2018) A bandit approach to multiple testing with false discovery control. *32nd Conf. Neural Inform. Processing Systems*, 1−11.

Jamieson K, Jain L, Fernandez C, Glattard N, Nowak R (2015) NEXT: A system for real-world development, evaluation, and application of active learning. *Adv. Neural Inform. Processing Sys- tems* 28:2656−2664.

Katariya S, Jain L, Sengupta N, Evans J, Nowak R (2018) Adaptive sampling for coarse ranking. *Proc. 21st Internat. Conf. Artificial Intelligence Statist.*, 1−16.

Kilgour M, Koslow S, O′Connor H (2020) Why do great creative ideas get rejected? *J. Advertising Res.* 60(1):12−27.

King G, Zeng L (2001) Logistic regression in rare events data. *Politi- cal Anal.* 9(2):137−163.

Kornish LJ, Jones SM (2021) Raw ideas in the fuzzy front end: Verbos- ity increases perceived creativity. *Marketing Sci.* 40(6):1106−1122.

Kornish LJ, Ulrich KT (2011) Opportunity spaces in innovation: Empirical analysis of large samples of ideas. *Management Sci.* 57(1):107−128.

Lamb C, Brown DG, Clarke C (2015) Human competence in creativ- ity evaluation. Toivonen H, Colton S, Cook M, Ventura D, eds. *Proc. Sixth Internat. Conf. Comput. Creativity*, 102−109.

Luo L, Toubia O (2015) Improving online idea generation platforms and customizing the task structure on the basis of consumers′ domain-specific knowledge. *J. Marketing* 79(5):100−114.

Matz SC, Netzer O (2017) Big data as a window into consumers′ psychology. *Current Opinion Behav. Sci.* 18:7−12.

Mervis CV, Rosch E (1981) Categorization of natural objects. *Annual Rev. Psych.* 32(1):89−115.

Mobley MI, Doares LM, Mumford MD (1992) Process analytic mod- els of creative capacities: Evidence for the combination and reorganization process. *Creativity Res. J.* 5(2):125−155.

Netzer O, Feldman R, Goldenberg J, Fresko M (2012) Mine your own business: Market-structure surveillance through text min- ing. *Marketing Sci.* 31(3):521−543.

Netzer O, Lemaire A, Herzenstein M (2019) When words sweat: Written words can predict loan default in the text of loan appli- cations. *J. Marketing* 56(6):1–81.

O′Quin K, Besemer SP (1999) Creative products. *Encyclopedia of Crea- tivity*, vol. 1, 413−422.

Rafieian O, Yoganarasimhan H (2021) Targeting and privacy in mobile advertising. *Marketing Sci.* 40(2):193−218.

Rehurek R, Sojka P (2011) *Gensim−Python Framework for Vector Space Modelling* (NLP Centre, Faculty of Informatics, Masaryk Univer- sity, Brno, Czech Republic).

Ritchie G (2001) Assessing creativity. Wiggins GA, ed. *Proc. AISB′01 Sympos. Artificial Intelligence Creativity Arts Sci.*, 3−11.

Rosch E, Simpson C, Miller RS (1976) Structural bases of typicality effects. *J. Experiment. Psych.* 2(4):491−502.

Runco MA (1995) Insight for creativity, expression for impact. *Crea- tivity Res. J.* 8(4):377−390.

Sievert S, Ross D, Jain L, Jamieson K, Nowak R, Mankoff R (2017) NEXT: A system to easily connect crowdsourcing adaptive data collection. *Proc. 16th Python Sci. Conf.*, 113−119.

Simonton BK (1999) Creativity from a historiometric perspective. Sternberg, RJ, ed. *Handbook of Creativity* (Cambridge University Press, Cambridge, UK), 116−136.

Stephen A, Zubcsek PP, Goldenberg J (2016) Lower connectivity is better: The effects of network structure on redundancy of ideas and customer innovativeness in interdependent ideation tasks. *J. Marketing Res.* 53(2):263−279.

Sukhov A (2018) The role of perceived comprehension in idea eval- uation. *Creativity Innovation Management* 27(2):183−195.

Sukhov A, Sihvonen A, Netz J, Magnusson P, Olsson L (2021) How experts screen ideas: The complex interplay of intuition, analy- sis, and sensemaking. *J. Product Innovation Management* 38(2): 248−270.

Terwiesch C, Ulrich KT (2009) *Creating and Selecting Exceptional Opportunities* (Harvard University Press, Cambridge, MA).

Tibshirani R (1996) Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. B* 58(1):267−288.

Toubia O (2006) Idea generation, creativity, and incentives. *Market- ing Sci.* 25(5):411−425.

Toubia O, Flores L (2007) Adaptive idea screening using consumers. *Marketing Sci.* 26(3):342−360.

Toubia O, Netzer O (2017) Idea generation, creativity, and prototy- picality. *Marketing Sci.* 36(1):1−20.

Urban GL, Katz GM (1983) Pre-test-market models: Validation and managerial implications. *J. Marketing Res.* 20(3):221−234.

Wang M, Zhao Y, Zhang B (2015) Collective dynamics of ‵small- world′ networks. *Sci. Rep.* 5(1):1−12.

Ward TB (1995) What′s old about new ideas?″ Smith SM, Ward TB, Fiske RA, eds. *The Creative Cognition Approach* (MIT Press, Cam- bridge, MA), 157−178.

Watts DJ, Strogatz SH (1998) Collective dynamics of ‵small-world′ networks. *Nature* 393(6684):440−442.

Wedel M, Kannan PK (2016) Marketing analytics for data-rich envir- onments. *J. Marketing* 80(6):97−121.

Wessling KS, Huber J, Netzer O (2017) MTurk character misrepre- sentation: Assessment and solutions. *J. Consumer Res.* 44(1): 211− 230.

Wooten JO, Ulrich KT (2019) The impact of visibility in innovation tournaments: Evidence from field experiments. Working paper.

NUMBER 1 OF 4

AUTHOR QUERIES

DATE 3/7/2023 JOB

NAME MKSC

ARTICLE 20231434

QUERIES FOR AUTHOR Bell et al.

THIS QUERY FORM MUST BE RETURNED WITH ALL PROOFS FOR CORRECTIONS

Q: 1_Please provide short title for running head. Please note that entire running head, includ- ing author names, must be , 81 characters/spaces.

Q: 2_Please provide all funding information, if applicable, including the names of the institutions as well as any grant numbers associated with the funding, or confirm that no fund- ing was received.

Q: 3_Your color figures appear in color in your page proof to simulate their presentation in color online, but they will be converted to grayscale in the print version of your article because you have not requested color processing for print. Please (a) check your figures to confirm that they will be sufficiently clear in grayscale presentation in the print version of your article and (b) modify your figure legends as necessary to remove any references
to specific colors in the figure.

Q: 4_Please confirm that the article title, author names, affiliations, and email addresses are set correctly. If applicable, please provide author ORCID numbers.

Q: 5_Please confirm that corresponding author designation in author line is correct.

Q: 6_Per journal style, ‟while″ has been changed to ‟whereas″ to clarify a causal rather than a temporal relationship. Please confirm throughout.

Q: 7_Please confirm that keywords are correct as set.

Q: 8_Please confirm that heading levels are correct as set.

Q: 9_Please verify that all displayed equations and in-text math notations are set correctly.

Q: 10_Per INFORMS style, built-up fractions included within paragraph text should be converted to their in-line form, i.e., with a solidus rule, as in $[a\ 1\ (b/x)]^{1/2}$, whenever possi- ble. Long or complex in-line formulas (within paragraph text) should be set as display equations. Please review all in-text equations and modify as necessary to be represented accurately as either display or in-line equations.

Q: 11_Per INFORMS style, all variables should be italic and all vectors should be bold. Please confirm that all terms have been formatted properly throughout.

Q: 12_The in-text citation ‟Dahan and Hauser 2001″ is not in the reference list. Please correct the citation, add the reference to the list, or delete the citation.

Q: 13_Please ensure that all direct quotes include a page number citation.

Q: 14_The in-text citation ‶Finke et al. 1992″ is not in the reference list. Please correct the cita- tion, add the reference to the list, or delete the citation.

Q: 15_The in-text citation ‶Durand and vanHuss 1992″ is not in the reference list. Please correct the citation, add the reference to the list, or delete the citation.

Q: 16_Per journal style, ‶since″ has been changed to ‶because″ to clarify a causal rather than a temporal relationship. Please confirm throughout.

Q: 17_The in-text citation ‶Balakrisnan and Lloyd-Yemoh 2014″ is not in the reference list. Please correct the citation, add the reference to the list, or delete the citation.

Q: 18_The in-text citation ‶Yao et al. 2018″ is not in the reference list. Please correct the citation, add the reference to the list, or delete the citation.

Q: 19_Because figures appear in color online only, please revise the text related to figures to eliminate any reference to color.

Q: 20_Per journal style, ‶due to″ has been changed to ‶because of″ per its use as an adverb in this context to conform with INFORMS style. Please confirm throughout.

Q: 21_Please include any necessary acknowledgments here.

Q: 22_The in-text citation ‶Rehurek, 2011″ is not in the reference list. Please correct the citation, add the reference to the list, or delete the citation.

Q: 23_Reference ‶Amatriain, Lathia, Pujol, Kwak, Oliver, 2009″ is not cited in the text. Please add an in-text citation or delete the reference.

Q: 24_Please ensure that all conference proceedings entries on the reference list include the proceedings editors, the correct proceedings title, the publisher, the city of publication, and the page range.

Q: 25_Please ensure that all book entries on the reference list include the city of publication as well as the publisher.

Q: 26_Reference ‶Bishop, 2006″ is not cited in the text. Please add an in-text citation or delete the reference.

Q: 27_Reference ‶Cao, Li, Tao, Li, 2015″ is not cited in the text. Please add an in-text citation or delete the reference.

Q: 28_Reference ‶Dahan, Hauser, 2002″ is not cited in the text. Please add an in-text citation or delete the reference.

Q: 29_Reference ‶Friedman, Hastie, Tibshirani, 2001″ is not cited in the text. Please add an in- text citation or delete the reference.

Q: 30_Reference ‶Giora, 2003″ is not cited in the text. Please add an in-text citation or delete the reference.

Q: 31_For the Goldenberg and Mazursky reference, please confirm the publisher and publish- ing location.

Q: 32_Reference "Hill, Ready-Campbell, 2011" is not cited in the text. Please add an in-text cita- tion or delete the reference.

Q: 33_Please ensure that all journal entries on the reference list include the volume and issue numbers and the page range. Please confirm all that have been added or updated.

Q: 34_Please ensure that all book chapter entries on the reference list include the chapter and book titles, the book editors, the publisher and city of publication, and the page range of the cited chapter.

Q: 35_Reference "Rehurek, Sojka, 2011" is not cited in the text. Please add an in-text citation or delete the reference.

Q: 36_Reference "Ward, 1995" is not cited in the text. Please add an in-text citation or delete the reference.

Q: 37_Reference "Wessling, Huber, Netzer, 2017" is not cited in the text. Please add an in-text citation or delete the reference.

Q: 38_Please update all "Working paper" references if possible. If still unpublished, please ensure the institution and location of the first author is provided in each case.

Q: 39_Please include a short, overall descriptive caption for Figure 1.

Q: 40_Please check that all tables and figures, including their titles and notes (which may con- tain edits) are set correctly.

Q: 41_Table footnote with "**" linking symbol does not have a matching callout in the table.

Q: 42_Please revise the heading/notes for Tables 5 and 7 to reflect that color will not appear in print.

Q: 43_For Table 8, please include a note corresponding to footnote indicator "^{a}".

Q: 44_Per journal style, author bios must be fewer than 500 characters in length. Please adjust accordingly (current: 798). Please ensure that all retained abbreviations are spelled out. Gerard J. Tellis (PhD Michigan) is the Neely Chaired Professor of American Enterprise, Director of the Institute for Outlier Research in Business, and Director of the Center for Global Innovation at the Marshall School of Business. Dr. Tellis is one of the world's

leading experts in effective advertising, virality on social media, radical innovation, and diffusion of innovations. He has published seven books

and more than 200 papers (http://www.gtellis.net) that have earned more than 30,000 citations. His publications have won more than 25awards.

Q: 45_Per journal style, author bios must be fewer than 500 characters in length. Please adjust accordingly (current: 1189).

Johann Feuller holds the Chair for Innovation and Entrepreneurship at Innsbruck

University. Before his professorship, Johann was a fellow at the National Aeronautics and Space Administration Tournament Laboratory-Research at Harvard University and visiting scholar and research affiliate at the Massachusetts Institute of Technology Sloan School of Management. His research interests are in artificial intelligence, open innova- tion, crowdsourcing, corporate entrepreneurship, incubation, and digitalization. Johann published in journals such as *California Management Review*, *Information Systems Research*, *Marketing Science*, *Management Information Systems Quarterly*, *Research Policy*, and others. He won several awards, such as the best paper award in 2015 of *Journal of Interactive Marketing* or the Case Centre Award 2015 for his *Harvard Business Review* case study on open innovation @ Siemens. He holds several advisory board member positions at start- ups such as Icaros, Tawny, and FdG. Johann is also a member of the scientific board of University of Vaasa. As chief executive officer of Hyve AG, Johann empowers international corporations to innovate and to become more entrepreneurial.