IdeaHound: Improving Large-scale Collaborative Ideation with Crowd-powered Real-time Semantic Modeling

Pao Siangliulue[1] **Joel Chan**[2] **Steven P. Dow**[3] **Krzysztof Z. Gajos**[1]

[1]SEAS, Harvard University
Cambridge, MA USA
{paopow, kgajos}@seas.harvard.edu

[2]Carnegie Mellon University
Pittsburgh, PA USA
{joelchuc}@cs.cmu.edu

[3]Design Lab, UC San Diego
La Jolla, CA 92093
{spdow}@ucsd.edu

ABSTRACT

Prior work on creativity support tools demonstrates how a computational semantic model of a solution space can enable interventions that substantially improve the number, quality and diversity of ideas. However, automated semantic modeling often falls short when people contribute short text snippets or sketches. Innovation platforms can employ humans to provide semantic judgments to construct a semantic model, but this relies on external workers completing a large number of tedious micro tasks. This requirement threatens both accuracy (external workers may lack expertise and context to make accurate semantic judgments) and scalability (external workers are costly). In this paper, we introduce IDEAHOUND, an ideation system that seamlessly integrates the task of defining semantic relationships among ideas into the primary task of idea generation. The system combines implicit human actions with machine learning to create a computational semantic model of the emerging solution space. The integrated nature of these judgments allows IDEAHOUND to leverage the expertise and efforts of participants who are already motivated to contribute to idea generation, overcoming the issues of scalability inherent to existing approaches. Our results show that participants were equally willing to use (and just as productive using) IDEAHOUND compared to a conventional platform that did not require organizing ideas. Our integrated crowdsourcing approach also creates a more accurate semantic model than an existing crowdsourced approach (performed by external crowds). We demonstrate how this model enables helpful creative interventions: providing diverse inspirational examples, providing similar ideas for a given idea and providing a visual overview of the solution space.

Author Keywords

Idea generation, collaborative innovation

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

UIST '16, October 16 - 19, 2016, Tokyo, Japan
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4189-9/16/10 $15.00
DOI: http://dx.doi.org/10.1145/2984511.2984578

INTRODUCTION

Large creative online communities will transform the way our society innovates. Existing communities, like OpenIDEO (openideo.com), where people propose solutions to social problems, and platforms, like coUrbanize (courbanize.com), where cities gather ideas from their citizens, attract large numbers of users, many of whom contribute ideas or designs. The promise of these online communities is that participants will benefit from exposure to ideas of others and, thus inspired, will generate better ideas than they would have otherwise. In practice, however, crowd innovation challenges result in large quantities of simple, mundane and repetitive ideas [2, 22, 40]. Consequently, many organizations have come to see crowd innovation platforms more as marketing gimmicks that energize their customers or constituents, rather than real sources of innovation. Meanwhile, numerous creativity-enhancing interventions targeted at individuals and small groups exist, and many of these interventions have been demonstrated to measurably improve the creative outcomes. How might we build on these successes to improve the quality and diversity of ideas contributed on large scale collaborative ideation platforms?

Many creativity-enhancing interventions leverage corpora of relevant design examples *and* a computational insight into the structure of the solution space revealed by those examples. For example, Design Gallery for 3D modeling [30] and other similar systems [27, 47] help users gain a quick intuition of the solution space and facilitate recombination of disparate ideas [33] by showing them multiple *diverse* alternatives. ReflectionSpace [43] and Freed [32] support reflection in the design process by presenting users' designs in the context of other related artifacts. Adaptive Ideas web design tool [27] and DesignScape [37] promote broad exploration of the solution space during the divergent phase of idea generation by showing a diverse set of examples and design alternatives. They also support refinement by allowing users to explore sets of closely related ideas, all of which pursue the same general approach, but in subtly different ways [27, 37].

All of these systems leveraged some computational representation that made it possible to tell which ideas were similar to each other and which were different. They either leveraged the fact that the design space was parameterized to begin with (e.g., 3D models in [30]) or they used some mechanism to automatically compute descriptive features of the artifacts (e.g., [27, 47]). On existing large scale collaborative ideation

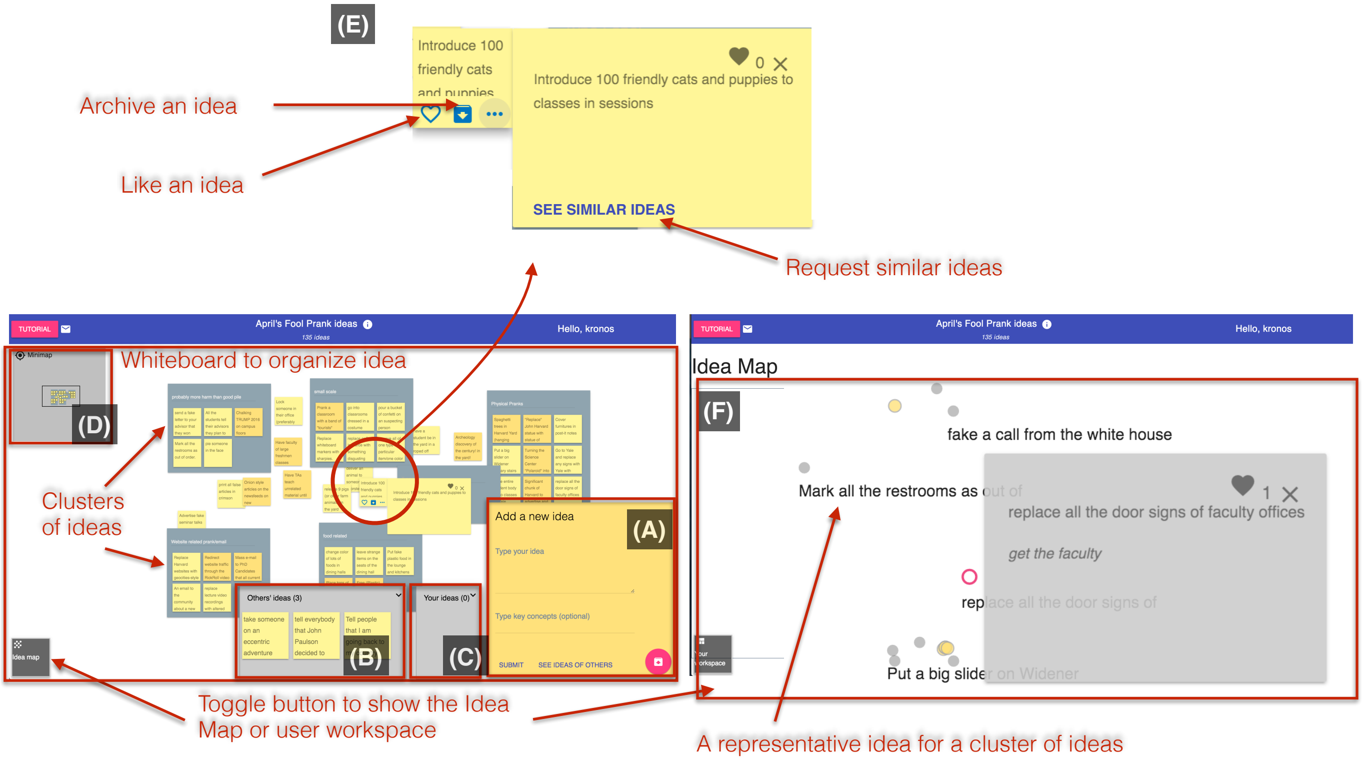

(E)
Introduce 100 friendly cats and puppies
Archive an idea
Like an idea
0
Introduce 100 friendly cats and puppies to classes in sessions
SEE SIMILAR IDEAS
Request similar ideas
TUTORIAL
April's Fool Prank ideas
135 ideas
Hello, kronos
Minimap
Whiteboard to organize idea
(D)
Clusters of ideas
Others' ideas (3)
take someone on an eccentric adventure
tell everybody that John Paulson decided to
Tell people that I am
(B)
Your ideas (0)
(C)
Add a new idea
(A)
Type your idea
Type key concepts (optional)
SUBMIT
SEE IDEAS OF OTHERS
Idea map
Idea Map
(F)
fake a call from the white house
Mark all the restrooms as out of
1
replace all the door signs of faculty offices
get the faculty
replace all the door signs of
Put a big slide on Widener
Your workspace
Toggle button to show the Idea Map or user workspace
A representative idea for a cluster of ideas

Figure 1. IdeaHound interface. (A) A box where users can type and submit their ideas; (B) When users request ideas from others, they appear on the Others' Ideas pane, (C) When users submit an idea, it first appears on the Your Ideas pane; Users can move ideas from (B) and (C) to organize on the whiteboard area. When they place ideas close to each other, a cluster will form around the ideas. (D) A minimap of the workspace. Users can pan and zoom the whiteboard or control the zoom from the minimap view. (E) When they hover over an idea, a control panel allows users to like the idea, remove the idea from the workspace, or open up a Details pane for that idea. On the Details pane, users can click "See similar ideas" to request ideas of others that are similar to that idea. (F) The idea map visualization is a 2D map that gives an overview of the solution space. Each dot represents an idea. The user's own ideas are in orange while the ideas from others are in yellow. A label for each cluster of ideas on the idea map visualization shows a sampled idea from that cluster.

platforms, people tend to communicate their initial ideas in the form of short text snippets or sketches. Thus, no *a priori* parametrization of the solution space is available. Furthermore, feature discovery mechanisms such as probabilistic topic modeling [3] do not perform well with such representations [8, 11].

Crowd-powered systems offer a possible solution: by judiciously combining human judgement and machine learning, it is possible to discover useful structure in collections of arbitrary artifacts [1, 10, 44, 48]. However, existing crowd-powered approaches have a crucial limitation on their applicability to large-scale collaborative ideation platforms: they depend on people completing a large number of tedious and repetitive micro-tasks. This requirement means platforms that seek to leverage such approaches must employ large numbers of external workers (e.g., from online labor markets such as Amazon Mechanical Turk or UpWork). This is not a desirable approach, for two main reasons. First, employing large numbers of workers is expensive, which limits the ability of these systems to scale to very large innovation platforms. Secondly, even if cost was not a concern, many online creative communities assume some amount of shared knowledge (e.g., local knowledge among contributors to a municipal participatory budgeting platform), which would not be available to workers hired outside the community. Thus, the human judgements on semantic relationships among ideas should come from the creative community itself.

This paper's first study suggests that it is infeasible to expect unpaid, intrinsically-motivated participants to complete a secondary task of judging ideas of others in addition to the primary task of generating ideas. We recruited unpaid, intrinsically motivated participants to generate ideas and we then asked them to evaluate ideas generated by other members of the community (rate similarity between ideas and idea quality). When we required participants to complete these evaluation tasks, they found the tasks to be tedious and repetitive; when the completion of the tasks was voluntary, participants did not complete enough of those tasks to inform the creation of a reliable computational model.

In response to this challenge, we designed IdeaHound, a self-sustainable system for supporting creative ideation at scale. A crucial, novel component of IdeaHound is an *integrated crowdsourcing* approach that seamlessly integrates the potentially tedious secondary task of analyzing semantic relationships among ideas with the more intrinsically-motivated primary task of idea generation. Our integrated approach leverages the insight that people naturally tend to spatially organize their inspirational material (including their own ideas) such that ideas and inspirations that share something in common are grouped together. IDEAHOUND thus presents users with a prominent affordance for spatially organizing their own ideas and ideas of others. IDEAHOUND continuously monitors the evolving spatial organizations created by all members of the community and creates a global model capturing relative similarities and differences among ideas. This model can help the community accomplish tasks both during idea generation (e.g., finding inspirations and gaining overview of solution space) and after idea generation (e.g., organizing ideas and selecting ideas). Figure 1 illustrates the main features of IDEAHOUND.

Our empirical studies demonstrate the viability of this approach. In Studies 2 and 3, participants implicitly defined semantic relationships among ideas by spatially organizing their own ideas and those of their peers while they were generating novel ideas. The results of these studies demonstrate that, even though participants were not explicitly asked to spatially organize ideas, they naturally did so frequently and thoughtfully enough to create an accurate computational model of the semantic relationships among ideas. The resulting model agreed with standard (and more expensive) human judgements more closely than a computational model created using a conventional outsourcing approach [44], where a separate crowd (of equally qualified participants) completed stand-alone semantic judgment tasks. Further, participants generated as many ideas (despite doing the extra work arranging ideas) and were as satisfied with the *integrated crowdsourcing* interface as they were with an equivalent conventional interface that required no additional work besides submitting their own ideas and browsing the ideas of others.

We demonstrate how the resulting semantic model can be used to enable three creativity-enhancing interventions in IDEAHOUND: sampling diverse inspirational examples, exploring similar ideas, and providing a visual overview of the emerging solution space. In Study 4, we conducted a preliminary end-to-end evaluation of IDEAHOUND. The results show that people found the suggested diverse sets of ideas helpful for their idea generation. They also found that the map visualization provided them with a quick and useful overview of the evolving solution space.

This paper makes the following contributions:

1. A crowdsourcing approach that integrates the potentially tedious task of evaluating creative ideas with the more exciting task of idea generation, such that contributors, who are intrinsically motivated only to contribute to idea generation, perform both tasks.

2. An end-to-end system, called IDEAHOUND, which uses crowd-contributed spatial arrangements of ideas to construct a robust model of semantic relationship among ideas. IDEAHOUND uses this model to enable three creativity-enhancing interventions: sampling diverse inspirational examples, exploring similar ideas, and providing a visual overview of the emerging solution space. While similar interventions were previously used to enhance the performance of individuals and small groups, IDEAHOUND makes it possible to support creative communities of hundreds or thousands of contributors.

3. Empirical studies that demonstrate the need for and the viability of an integrated crowdsourcing approach for supporting enhanced collective ideation at scale.

RELATED WORK

Creativity Enhancing Interventions

As noted in the introduction, large-scale ideation systems typically do not live up to their promise in practice: they tend to collect large numbers of redundant and shallow ideas of variable quality [2, 22, 40]. The emerging literature on creative cognition and creativity support tools has identified a number of creativity-enhancing interventions that can significantly improve the performance of large-scale ideation systems by improving individual creativity and/or enhancing collaboration capabilities.

For example, while exposure to mundane examples may hinder creativity [19, 23], individuals can come up with more diverse and/or creative ideas if they have access to diverse and high quality inspirational examples [6, 31, 36, 44, 45, 46]. The net effect of increasing individual creativity is that the community can converge on novel, high quality solutions more quickly than if all participants simply saw their own ideas [4]. Inspirational examples can be drawn from peers' ideas for the same problem [36, 44, 45], or from external sources [6, 18, 27]. It is important that example sets be relatively small, because participants have limited time and cognitive resources [20, 29]. If people have to process large numbers of examples, they can resort to effort-saving but suboptimal strategies, such as merely referring to (instead of deeply building on) other ideas [20]. Further, the content of the ideas people see also matters. Prior research has found that examples are most inspirational if they are diverse [18, 44] and/or appropriate to their current context [18]. Examples can also increase creativity by supporting exploration of iterations and variations on a solution approach [7, 27, 46], which can lead to not just higher quality [13], but also more novel ideas [35, 41]. In contrast, poorly chosen examples can even harm ideation, by inducing distraction [36] or fixation [19, 23].

Individuals and communities can also achieve better creative outcomes if they have access to a "map" of the solution space that shows the kinds of solutions that have been explored by the community so far and/or in prior/external efforts to solve the problem, and how they relate to each other semantically. The map's higher-level view of the solution space can enable deeper insights into the solution space [14, 30, 47], and the abstract solution "schemas" that might describe clusters of related ideas [54]). These deeper insights have been shown to facilitate more effective recombination of ideas than with raw example ideas [28, 54]. These maps can also improve iteration on ideas by enabling people to discover and explore many closely related solution alternatives [9, 17, 27, 38].

At the community level, maps have also been shown to help the community keep track of their exploration of the solution space [33, 34]. The map can give participants an overall sense of what ideas have already been conceived, what "gaps" might exist, and where to focus their efforts. Participants can then make the best use of their limited time to make contributions that are most valuable to the community, avoiding redundant effort. This coordination benefit is supported by simulation studies [51, 53], as well as an empirical study of collaborative ideation on programming problems [4]. These maps greatly reduce the costs of manual coordination across collaborators, which can be extremely high in large-scale collaboration systems [21]. However, these benefits have heretofore largely been realized in systems that engaged a small number of dedicated leaders to manually construct such maps.

The common thread behind these interventions is that they depend on having access to both a large corpus of solutions (whether generated externally or by peers in the same community) *and* a semantic model that specifies the structure of the solution space (e.g., how solutions relate to each other).

Uncovering Semantic Relationships In Large Corpora

Some automated mechanisms that extract information about the emerging solution space from a collection of ideas exist for domains where ideation artifacts are created with well-defined structures, such as webpages or geometric shapes in 3D modeling [14, 27, 30, 47]. In contrast, our aim is to improve systems that address a wide variety of problems, where ideation inputs are most commonly in the form of unstructured short text snippets or sketches. Fully automated topic analysis approaches exist for analyzing large collections of unstructured text [3]. However, these approaches often miss key nuances in the data [8, 11], and struggle with short text snippets. They also cannot handle unstructured sketches without some initial segmentation of sketches into reasonable "units" (analogous to words in topic modeling of texts).

A promising alternative leverages human computation, whether exclusively or in a hybrid system with machine intelligence [1, 10, 44]. Human computation approaches have been successfully applied to organize artifacts in various domains such as 3D modeling [9, 47], graphic designs [37, 38] and music composition [17]. These approaches all require considerable number of inputs from humans to discover the design space of ideas; some of these inputs are extracted from users' interactions with the system [17, 47], but most of these inputs are from small, explicit human computation tasks, such as clustering subsets of items [1, 10], completing similarity comparisons between items [44], or identifying attributes of items [9, 37, 38].

One key disadvantage of human computation micro tasks is that they tend to be uninteresting and repetitive. The specific activities (tagging, judgements of relative similarity) take time, do not directly contribute to the ideation process, and are often perceived as tedious. Contributors to the online communities generally avoid doing tedious maintenance tasks (in this case providing information about ideas) to do more interesting tasks (generating ideas) [26]. One could argue that these activities could be performed by external crowds hired specifically for the purpose (which is what some existing systems do [44]). However, even if cost was not an issue, this approach can be challenging for those creative tasks where specialized domain knowledge is required. Further, outsourcing is typically done in batches, but some of the key creativity-enhancing interventions of collaborative ideation systems (e.g., coordination via a solution space overview) require (near) real-time continuous updates to the model. In this work, we seek to increase the feasibility of human computation approaches by exploring ways to integrate the semantic organization tasks into participants' primary activities.

DESIGN GOALS

The end goal of this work is to improve large-scale collaboration with creativity-enhancing interventions, such as providing diverse inspirational examples, enabling exploration of similar ideas (for iteration), and providing a real-time "map" or overview of the solution space. As we have seen in the review of prior work, these interventions depend on having access to a semantic model that captures the structure of the solution space (e.g., how solutions relate to each other). However, none of the existing solutions for constructing such models are adequate: the completely automated approaches are unlikely to work well with short text snippets and sketches, while the crowd-powered solutions depend on large numbers of external workers completing many tedious/repetitive semantic judgment tasks.

Therefore, the technical focus of this work is to create an approach for semantic modeling of solution spaces that meets two main requirements:

1. *Nearly Real-time.* The approach should be able to provide a nearly real-time model of the solution space.
2. *Self-sustainable.* The approach should not depend primarily on external labor.

Our general approach is to combine methods from crowdsourcing and machine learning research. Specifically, similarly to [44], we rely on a modest number of human judgements regarding relative similarities of pairs of ideas and we then use machine learning techniques to efficiently combine those human judgements into a consistent and comprehensive model of the emerging solution space. Unlike the prior work, however, we seek to engage the members of the creative community themselves in the process of constructing the semantic model instead of outsourcing the task to external crowds. In the following sections, we describe the rationale, design, technical details, and evaluation of our approach. We also demonstrate how this approach allowed us to build IDEAHOUND, an end-to-end self-sustainable system that enables three creativity interventions for enhancing collective ideation at scale.

STUDY 1: SEPARATE TASKS TO COLLECT SEMANTIC RELATIONSHIPS AMONG IDEAS

A straightforward approach to solicit the necessary human judgements of semantic relationships among ideas is to explicitly ask the members of the community to contribute these judgements. We tested this approach in a study conducted on the LABINTHEWILD.ORG platform [39], which attracts intrinsically motivated, unpaid online participants who take part in studies in return for informative feedback on their performance. We recruited 2,061 participants to generate ideas for birthday messages for a 50-year-old female firefighter. The study had four parts: 1) participants generated as many ideas as they could in 4 minutes, 2) after they finished generating ideas, they were asked to provide a small set (5–10) of human judgments of semantic similarity between ideas (using the same mechanism as [44]), 3) they were presented with a results page, and 4) they ranked ideas of others based on their quality. The last part of the experiment was optional and participants could skip this part at any time.

141 of the 2,061 participants who finished generating ideas in part 1 dropped out before completing the semantic judgments in part 2. Further, fewer than half of the participants (743 out of the remaining 1,920) finished the optional ranking task in part 3. Some participants noted in their post-study open-ended comments that the semantic judgement and ranking tasks were repetitive, unappealing and took too much time. One participant almost gave up on the semantic judgment task because it was "boring and cruel". This suggests that when given a choice to optionally complete these extra human judgment tasks, few participants on these platforms will choose to do so.

INTEGRATED CROWDSOURCING OF CREATIVE IDEAS AND SEMANTIC RELATIONSHIPS

Instead of asking users to provide insights into a solution space by doing tedious tasks that detract from generating ideas, we sought to design an interaction that seamlessly integrated subjective judgement tasks with idea generation. Figure 1 shows the main interface for the final prototype.

In designing this solution, we drew inspiration from several existing systems, which require diverse kinds of work to be accomplished, but whose users are intrinsically motivated to do only a subset of those tasks. For example, Duolingo integrated the potentially tedious secondary task of translating real world text with the intrinsically valuable primary activity of learning a new language. In the CROWDY system [52], people who want to learn specific skills (e.g., web programming) improve video tutorials for future learners (the secondary task) as a byproduct of learning from those tutorials (the primary task). The users of the American Sign Language (ASL) flashcard quiz [5] improve the feature-based indexing of the signs for the new ASL dictionary (the secondary task) as a byproduct of practicing the signs (the primary task). In another system [24], students generate formative feedback on each other's assignments (the secondary task) as a byproduct of studying for an exam (the primary task). While all of these prior systems leveraged the users' desire to learn, we believe the approach of integrating a valuable but potentially tedious secondary task into an intrinsically motivating task generalizes to other settings where users have different intrinsic motivations.

In the rest of this section, we describe the iterative development of our integrated crowdsourcing approach through a series of formative prototypes.

Initial Design: Continuous Spatial Arrangement

We based our design on the insight that people naturally spatially organize their inspirational material. The key feature of our design is a whiteboard space where users can arrange their own ideas or ideas of others. Because the spaces where users' own ideas and the inspirational examples first appear are very small, users naturally tend to drag ideas (their own and those of others) onto the canvas and organize them spatially. Because the whiteboard naturally affords continuous spatial arrangements (placing ideas close or far from each other), we initially built on the SpAM approach of collecting similarity information from people's spatial arrangements [15,16]. From each spatial arrangement generated by a user, our system extracted similarity scores from relative distances between pairs of ideas. Then the system aggregated these implicit similarity judgements from users' whiteboards using multidimensional scaling (MDS) algorithm to generate an aggregate semantic model.

Our pilot study with this version of the prototype showed that people naturally organized ideas spatially without instructions to do so, but not in the way that the system was designed for: Instead of organizing ideas such that physical distances among them would represent the degree of dissimilarity, as SpAM assumes, people tended to aggregate ideas into discrete clusters. Open-ended comments from participants revealed that such discrete clustering (rather than continuous spatial arrangement) gave them a better sense of the emerging themes and provided a more readable "big picture" of the possible approaches to the creative challenge at hand.

Revised Design: Explicit Clustering

In our second design we made cluster-forming actions explicit. Whenever a user brings two ideas into close proximity, an outline is drawn around both ideas to indicate that they are now grouped into a cluster. Cluster management is fluid yet explicit: when a user brings an idea close to an existing cluster, the cluster automatically expands; when the user drags an idea away from a cluster, the idea is removed from the cluster.

According to feedback from our pilot studies, this approach was intuitive and matched users' expectations well. However, they reported that they sometimes forgot what concept they had intended to capture with each cluster. This was particularly frustrating to the users when ideation was performed over the course of several days: when they returned to the task after a day's break, they had a difficult time remembering the organizational structure they had been working to create.

Also, when we analyzed clusters created as part of several studies, it became clear that not all clusters were used to capture semantic similarity. Instead, some clusters were used to store “other” ideas or user’s own ideas regardless of their semantics. This was problematic because it created a mismatch between the actual semantics of some of the clusters and the assumptions made by our algorithm.

Final Design: Explicit Clustering with Labels

To address these two issues, in our final prototype we introduced a clear affordance to add optional textual labels to clusters. This design turned out to be very effective. Not only did it help participants remember better what each cluster was intended to capture, it also substantially reduced the number of clusters that did not capture semantic similarity. Thus, this design choice simultaneously made the spatial organization capability more useful to the users and made the user-generated clusters a more valuable source of data for the machine learning algorithm.

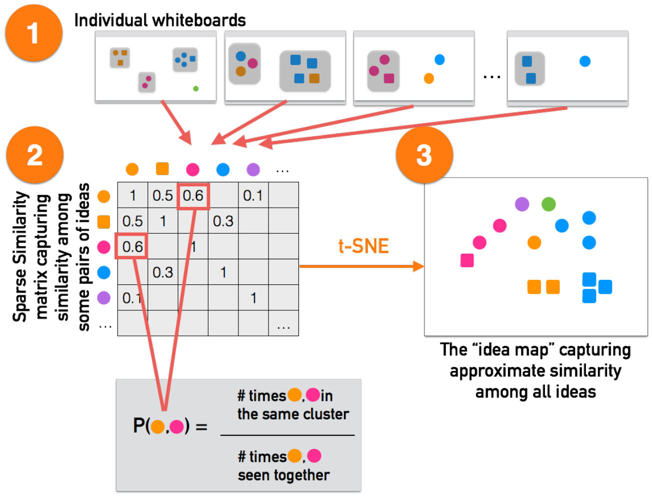

1
Individual whiteboards
...
2
Sparse Similarity matrix capturing similarity among some pairs of ideas
1 0.5 0.6 0.1
0.5 1 0.3
0.6 1
0.3 1
0.1 1
...
t-SNE
3
The “idea map” capturing approximate similarity among all ideas
P(●,●) =
# times ●,● in the same cluster
# times ●,● seen together

Figure 2. Computational model generation process. The system 1) aggregates grouping information from all users’ whiteboard organization, 2) constructs a sparse similarity matrix from aggregated grouping, and 3) generates an “idea map” that puts similar ideas closer to each other and keeps dissimilar ideas far from each other according to similarity matrix in 2).

Computational Model

As illustrated in Figure 2, to compute a global computational representation of how similar or different the collected ideas are from each other, our system initially constructs a similarity matrix from clusters across the users. Here, the similarity between two ideas is the empirical probability that the two ideas will be in the same cluster if they are both placed on the same whiteboard. This similarity matrix is sparse, however: not all pairs of ideas appear on the same whiteboards, so not all pairwise similarities are estimated. Therefore, the system computes an approximate idea similarity matrix (but one that estimates all pairwise similarities) using the t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm [49]. Following [44], we refer to this embedding as an *idea map*. This embedding provides an approximate estimate of similarities among all pairs of ideas, for which at least some similarity data are available.

EVALUATION OF THE TECHNICAL APPROACH

The central goal of our approach is that the potentially tedious secondary task of organizing ideas be integrated seamlessly into the intrinsically motivating primary task of idea generation. We evaluated our approach in two ways. In Study 2, we evaluated the experience and creative output of the users who used the system with our integrated crowdsourcing approach, compared to users who used a conventional interface. In Study 3, we evaluated the accuracy of the semantic model created using our integrated approach by comparing it to a model generated using a previously-validated method [44] that relies on outsourced crowd workers.

Study 2: User Experience and Creative Output with the Integrated Crowdsourcing Approach

In this study, we compared the experience and creative output of the users who used the integrated crowdsourcing approach, to users who used a conventional interface. We hypothesized that there would be no difference in experience and creative output between those who ideated with the integrated crowdsourcing interface and those who used a conventional interface.

Design

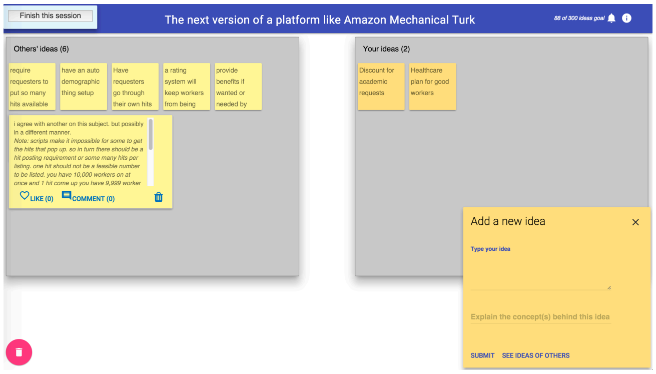
We used a between-subjects design with one factor with two conditions:

- *Integrated*: Participants used the integrated crowdsourcing interface like the one shown in Figure 1 to generate ideas. They could request to see ideas of others by clicking on the “SEE IDEAS OF OTHERS” button (Box A of Figure 1). The system then presented a set of up to three ideas. From ideas for which the system had information, the system sampled the first and the second idea. The first idea was selected randomly and the second idea was the idea that was predicted to be the most different from the first idea; the third idea was sampled randomly from ideas for which the system had no information. If there were no more unseen ideas, the system asked the user to request ideas again later. Participants could organize their own ideas and ideas of others together on the whiteboard. Unlike the interface in Figure 1, the participants could not request to see similar ideas to an idea or look at an idea map visualization.

- *Single-task*: Participants used a more conventional system without an integrated whiteboard (Figure 3). They could request to see ideas of others by clicking on a “SEE IDEAS OF OTHERS” button (bottom right of Figure 3). The system then presented a set of three ideas sampled randomly. As with the other design, if there were no more unseen ideas, the system asked the user to request ideas again later.

Task

Participants, who were recruited via Amazon Mechanical Turk (MTurk), generated ideas in asynchronous groups of 6–12. Each group was prompted to generate ideas for one of two prompts: 1) features for the next version of a micro task market platform like MTurk (*New features*), and 2) new tasks that can be posted on a microtask market (*New tasks*). We designed these tasks such that our participants would have the relevant domain expertise and the motivation to generate



require requesters to put so many hits available

have an auto demographic thing setup

Have requesters go through their own hits

a rating system will keep workers from being

provide benefits if wanted or needed by

i agree with another on this subject. but possibly in a different manner.
Note: scripts make it impossible for some to get the hits that pop up. so in turn there should be a hit posting requirement or some many hits per listing. one hit should not be a feasible number to be listed. you have 10,000 workers on at once and 1 hit come up you have 9,999 worker

Discount for academic requests

Healthcare plan for good workers

Figure 3. The interface for the *Single-task* condition of Study 2. The requested ideas of others are automatically placed on the left pane while participant's own ideas appear on the right pane. Participants could not move ideas around by dragging.

novel and valuable ideas. We gave participants freedom to choose when to start their idea generation session so not all participants had to generate ideas at the same time. Participants could complete one, two or three idea generation sessions with a mandatory break of at least 15 minutes between sessions. We set up the task this way to simulate real collaborative asynchronous idea generation platforms where contributors may revisit the platform to contribute more ideas at different time. Early arrivers might have had a different experience from those who started later because they saw different compositions of ideas.

Procedure

Before starting the first idea generation session, participants answered a demographics survey. Then participants went over the tutorial of the system and completed a practice task. Following insights from prior UI evaluations on MTurk [25], the practice task required participants to use each major feature of the system at least once before they could proceed to the main task. For each idea generation session, participants spent at least 12 minutes on generating ideas. At the end of the session, they answered survey questions about their experience in that session. They were required to wait at least 15 minutes before starting another session. If they chose to do the next session, the system would bring them back to the saved workspace where they ended the prior session.

Participants

We recruited 80 participants via MTurk to generate ideas. We limited recruitment to workers who had completed at least 1,000 HITs with approval rate greater than 95%. After seeing some participants' comments on grammatical errors of submitted ideas in the first two groups, we limited recruitment to U.S. residents (54 participants) for the rest of the experiment. Participants were paid \$2.00, \$3.50 or \$5.00 depending on whether they completed one, two or three ideation sessions.

Out of 80 recruited participants, 55 participants finished at least one ideation session; 23 participants dropped out of the experiment during the tutorial session, and 2 participants started but did not finish the first session. We only included the participants who finished at least one session in our analysis. The participants were randomly assigned to six different groups as summarized in Table 1.

27 participants (87%) in the *Single-task* condition finished all three sessions, 2 participants finished only two sessions and 2 participants finished only one session. 16 participants (67%) in the *Integrated* condition finished all three sessions, 2 participants finished only two sessions and 6 participants finished only one session. On average participants in the *Single-task* condition completed 1.81 sessions compared to 1.42 sessions in the *Integrated* condition. This difference was marginally significant ($\chi^2(1, N = 55) = 3.632, p = 0.0567$)

Measures and analysis

We compared the creative output of participants in the two conditions on the following measures.

- *Number of submitted ideas per participant*
- *Diversity of submitted ideas*: We used the same diversity measure as in [44]. Specifically, for each group, we randomly sampled 50 pairs of submitted ideas (300 pairs for 6 groups). We recruited 58 independent MTurk workers to rate similarity of pairs of ideas on a scale of 1 (not at all similar) to 7 (very similar). Each rater rated 25 pairs of ideas from the experiment. To ensure that the workers understood the task, they also rated 4 practice pairs that were rated as very similar or very different by one of the authors. Each pair of ideas was rated by 3–4 raters. We normalized (i.e., converted to z-scores) the ratings—including those of practice pairs—within each rater prior to aggregating the results. We flipped the sign of z-scored similarity ratings to derive diversity scores of a pair.
- *Creativity of submitted ideas*: We used the same creativity measure as in [44]. For each group, we randomly sampled 50 submitted ideas (300 ideas for 6 groups). We recruited 58 independent MTurk workers to rate ideas on two scales: novelty (1= not at all novel, 7 = very novel) and value (1 = not at all valuable, 7 = very valuable). Each rater rated 25 ideas. Each idea was rated by 4–5 raters. As before, we converted worker ratings into z-scores prior to analysis.

To compare user experience between the two systems, we collected participants' subjective responses (reported on a 7-point Likert scale) to questions that related to the following three aspects of their ideation experience:

- *Perception of helpfulness of ideas of others as selected by the system* (4 questions)
- *Perception of helpfulness of the system* (3 questions)
- *Mental effort and task difficulty* (2 questions)

We list the actual survey questions in Table 2. Noting that most participants finished either just one or all three sessions, we report the survey results after the first and the third sessions.

We also asked the participants in the *Integrated* condition to answer a separate set of 7-point Likert-scale questions related to their experience of organizing ideas on the whiteboard;

Task	Group	Condition	Number of participants	Number of sessions	# of generated ideas
Features for AMT	G1	Integrated	8	22	58
	G2	Single-task	12	35	217
	G3	Integrated	6	16	91
	G4	Single-task	9	26	95
New types of HIT	G5	Integrated	10	20	160
	G6	Single-task	10	26	143

Table 1. Number of participants, sessions and submitted ideas in each group.

Q10: *"Organizing ideas on the whiteboard helped me generate ideas. (1 Strongly disagree - 7 Strongly agree)"* and Q11: *"Organizing ideas on the whiteboard got in the way of generating ideas. (1 Strongly disagree - 7 Strongly agree)"*

We used analysis of variance for analyses involving Number of submitted ideas, Diversity of submitted ideas and Creativity of submitted ideas. We used ordinal regression for all analyses involving Likert-scale responses. We also used ordinal regression to compare the number of sessions completed under the two conditions.

A lack of statistically significant result does not constitute valid evidence for the lack of actual difference. Because we wish to demonstrate a lack of *substantial* differences in the quality of the experience between *Integrated* and *Single-task* conditions, we also computed effect sizes (Cohen's d) for all subjective measures and some of the performance measures. As is customary, we interpret effect sizes between 0.2 and 0.49 as small, between 0.5 and 0.79 as moderate, and those larger than 0.8 as large [12]. If our goal were to demonstrate the presence of statistically significant differences, we would have adjusted the p-values to account for the fact that we conducted multiple statistical comparisons based on data from a single experiment [42]. Given that our goal is the opposite, we report raw p-values throughout.

Results

No substantial difference in the number and diversity of examples seen. On average, the *Single-task* participants requested 33.8 ideas (SD=26.77), while the *Integrated* participants requested 21.6 ideas (SD=19.97). This difference is not significant ($F(1, 53) = 3.4691$, $p = 0.0681$). The average diversity scores of seen example sets were 0.074 (SD=0.46) in the *Single-task* conditions and -0.055 (SD=0.55) in the *Integrated* condition. This difference was small ($d = 0.26$) and not statistically significant $F(1, 58) = 0.9586$, $p = 0.3316$).

To derive the diversity score of examples we used a method analogous to the one used to compute the diversity of submitted ideas: We first randomly sampled 10 sets of seen examples from each group (60 sets for 6 groups). We recruited 30 independent MTurk workers to rate similarity of the 177 pairs of ideas on a scale of 1 (not at all similar) to 7 (very similar). Each rater rated up to 30 pairs of ideas. Each pair of ideas was rated by 5 raters. We normalized (i.e., converted to z-scores) the ratings within each rater prior to aggregating the results. We flipped the sign of z-scored similarity ratings to derive diversity scores of a pair. For each example set, we calculated the diversity score of an example set as the averaged pairwise diversity scores of ideas in that set.

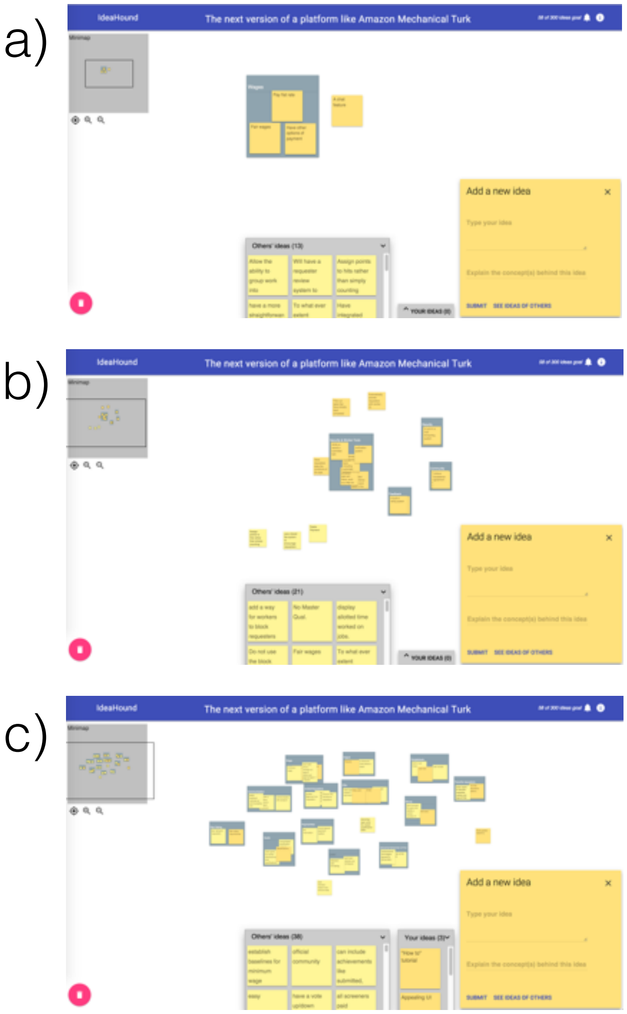


Figure 4. Participants engaged in organizing ideas to varying extent, ranging from making hardly any clusters (a: *P5*, *G1*), to moderate organization (b: *P6*, *G1*), to extensive organization (c: *P31*, *G3*).

No substantial difference in productivity. The average number of ideas submitted per session by a participant in the *Single-task* condition was 5.34, while the average number in the *Integrated* condition was 5.30. This difference was neither substantial ($d = 0.013$) nor significant ($F(1, 53) = 0.0022$, $p = 0.9626$).

No substantial difference in the diversity of submitted ideas. The average diversity score of submitted ideas in the *Single-task* condition was 0.132, while the average diversity score of submitted ideas in the *Integrated* condition was 0.105. This difference was neither substantial ($d = 0.042$) nor significant ($F(1, 299) = 0.1311$, $p = 0.7176$).

No substantial difference in the creativity of submitted ideas. The average novelty score of submitted ideas in the *Single-task* condition was 0.030, while the average novelty score of submitted ideas in the *Integrated* condition was -

0.030. This difference was neither substantial ($d = 0.099$) nor significant ($F(1, 299) = 0.7309$, $p = 0.3933$).

The average value score of submitted ideas in the *Single-task* condition was 0.028, while the average value score of submitted ideas in the *Integrated* condition was -0.028. This difference was neither substantial ($d = 0.096$) nor significant ($F(1, 299) = 0.6843$, $p = 0.4088$).

Participants in both conditions perceived the system-selected ideas of others as similarly helpful. Questions Q1 to Q4 in Table 2 measured the participants' perception of the usefulness of the ideas of others selected by the system. We found no significant differences in perception of helpfulness of ideas of others between the *Single-task* and the *Integrated* condition and none of the effect sizes was larger than small. We reported the p-values and effects sizes in Table 2.

Participants in both conditions perceived the system as similarly helpful. Question Q5 to Q7 in Table 2 measured the participants' perception of the usefulness of the ideation system. We found no significant difference in perception of helpfulness of system between the *Single-task* and the *Integrated* condition and none of the effect sizes was larger than small.

The whiteboard interface initially demands more mental effort. Question Q8 and Q9 in Table 2 measured the participants' perception of mental effort required to do the task and the difficulty of the task. We found no significant difference of task difficulty between the *Single-task* and the *Integrated* condition and none of the effect sizes was larger than small.

However, after completing the first session, participants in the *Integrated* condition reported significantly higher mental effort than in the *Singled-task* condition ($p = 0.0452$) and this difference was moderate in magnitude ($d = 0.6058$). However, this difference was no longer present after session 3, suggesting that the system became easier to use once participants gained some practice with it.

Organizing ideas on the whiteboard helps in generating ideas and does not get in the way. The level of organization varied across participants in the *Integrated* condition (Figure 4). On average, a participant put 21.2 ideas on the board (SD=15.35) and formed 4.79 clusters (SD=3.52).

The responses to the 7-point Likert scale questions for participants in the *Integrated* condition show participants found that organizing ideas helped them generate ideas (Q10, session 1: M=5.17, SD = 1.69, session 3: M=4.69, SD=2.21) and that it did not get in the way of generating ideas (Q11, session 1: M=2.50, SD = 1.82, session 3: M=3.13, SD=2.31).

When further prompted to explain how organizing ideas helped them generate ideas, participants stated that organizing ideas helped them "avoid repetition, and build off of previous ideas" [P25] and "[give] a clear picture of how things were grouped and [help] brainstorm more [ideas] based on grouping" [P51]. When further prompted how organizing ideas got in the way of generating ideas, most participants either did not provide a response or stated that the activity didn't get in the way of idea generation. One participant commented that she "did spend a bit of time organizing things instead of generating ideas. But it still helped in other ways" [P31].

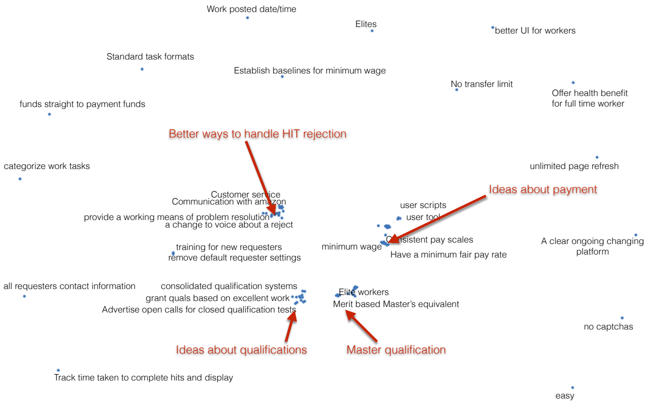


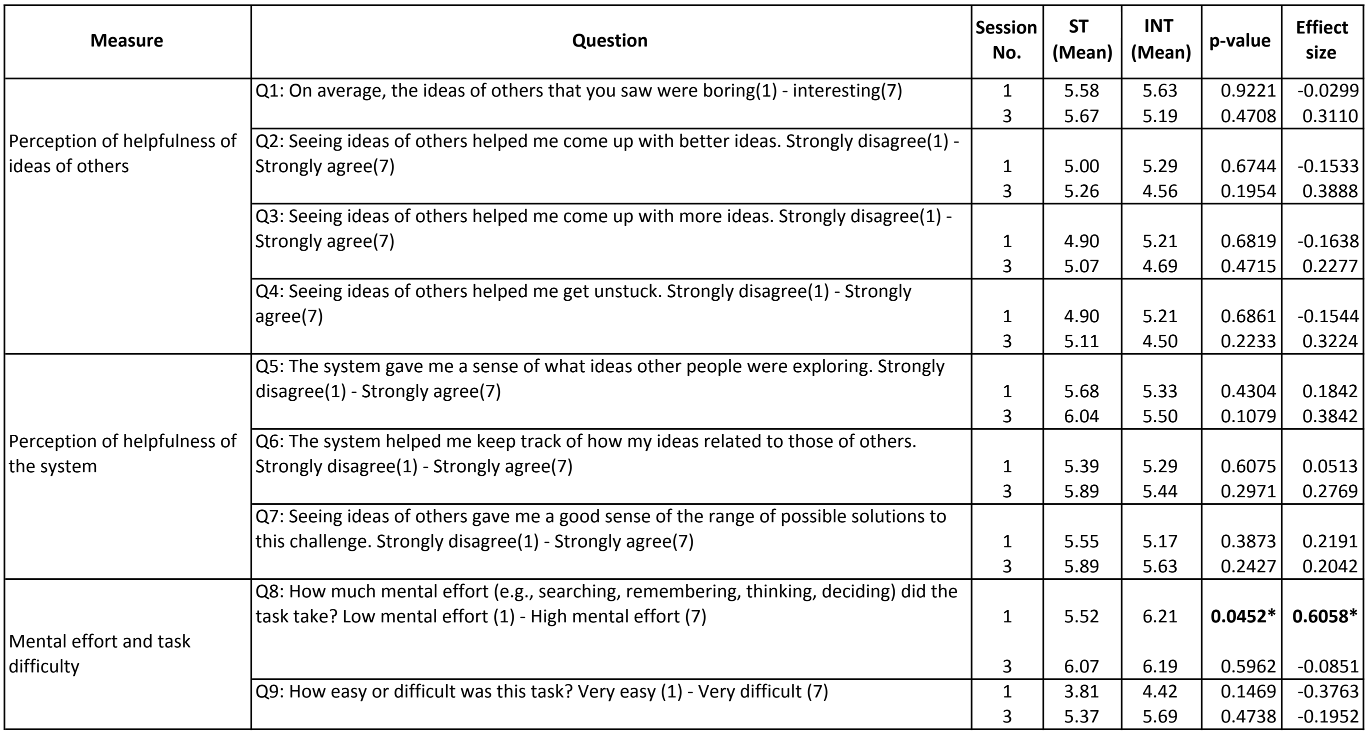
Figure 5. An idea map of *G3* in the *Integrated* condition, showing clusters of ideas around different topics. Isolated ideas around the edge are the ideas that either are different from other ideas or are the ideas that the system does not know much about yet.

The sparsity of similarity matrix (from direct human judgement) increases with the number of generated ideas. Participants in the *Integrated* condition organized both ideas of others (10.6 ideas on average) and their own (11.5 ideas on average). The proportion of idea pairs that actually received human judgements in the *Integrated* group varied from 0.28 (*G5*, 160 ideas) to 0.53 (*G1*, 58 ideas). The median of number of human judgements for each pair was 1 for all groups in the *Integrated* condition. The sparsity naturally increases with the number of ideas as the size of the similarity matrix grows quadratically with the number of ideas. As we will see in the next section, this sparsity does not significantly impact the quality of the resulting semantic model.

Study 3: Evaluating Model Quality Using Data from the Integrated Crowdsourcing Approach

Study 2 demonstrated that the whiteboard organization successfully integrated the secondary task of semantic judgment into the primary task of idea generation in a seamless fashion. But are these semantic judgments sufficient for building an accurate semantic model? In this study, we evaluated the accuracy of the integrated semantic modeling approach by comparing an *Integrated idea map* (Figure 5) generated by the system for one of the *Integrated* groups from Study 2, to one generated using a previously-validated method [44] that relies on outsourced crowd workers. We will refer to this comparison semantic model as the *Outsourced idea map*.

To generate the *Outsourced idea map*, we followed the procedure described in [44]. Specifically, for the 91 ideas generated by participants in the selected group, we posted 40 MTurk tasks for workers who had not done the idea generation task to collect 1,000 responses about similarity relationship between ideas. Each worker completed a series of triplet simi-



Measure	Question	Session No.	ST (Mean)	INT (Mean)	p-value	Effiect size
Perception of helpfulness of ideas of others	Q1: On average, the ideas of others that you saw were boring(1) - interesting(7)	1	5.58	5.63	0.9221	-0.0299
		3	5.67	5.19	0.4708	0.3110
	Q2: Seeing ideas of others helped me come up with better ideas. Strongly disagree(1) - Strongly agree(7)	1	5.00	5.29	0.6744	-0.1533
		3	5.26	4.56	0.1954	0.3888
	Q3: Seeing ideas of others helped me come up with more ideas. Strongly disagree(1) - Strongly agree(7)	1	4.90	5.21	0.6819	-0.1638
		3	5.07	4.69	0.4715	0.2277
	Q4: Seeing ideas of others helped me get unstuck. Strongly disagree(1) - Strongly agree(7)	1	4.90	5.21	0.6861	-0.1544
		3	5.11	4.50	0.2233	0.3224
Perception of helpfulness of the system	Q5: The system gave me a sense of what ideas other people were exploring. Strongly disagree(1) - Strongly agree(7)	1	5.68	5.33	0.4304	0.1842
		3	6.04	5.50	0.1079	0.3842
	Q6: The system helped me keep track of how my ideas related to those of others. Strongly disagree(1) - Strongly agree(7)	1	5.39	5.29	0.6075	0.0513
		3	5.89	5.44	0.2971	0.2769
	Q7: Seeing ideas of others gave me a good sense of the range of possible solutions to this challenge. Strongly disagree(1) - Strongly agree(7)	1	5.55	5.17	0.3873	0.2191
		3	5.89	5.63	0.2427	0.2042
Mental effort and task difficulty	Q8: How much mental effort (e.g., searching, remembering, thinking, deciding) did the task take? Low mental effort (1) - High mental effort (7)	1	5.52	6.21	**0.0452***	**0.6058***
		3	6.07	6.19	0.5962	-0.0851
	Q9: How easy or difficult was this task? Very easy (1) - Very difficult (7)	1	3.81	4.42	0.1469	-0.3763
		3	5.37	5.69	0.4738	-0.1952

Table 2. Summary of subjective responses after session 1 and after session 3. ST stands for *Single-task* and INT stands for *Integrated*. Participants rated the INT condition as demanding significantly more mental effort than the ST condition. We used Cohen's d to capture effect size.

larity comparison tasks: "is idea A more similar to idea B or C?" [48]. We used an active learning heuristic to sample the questions to ask to maximize expected information gain per question [48]. We then used t-Distributed Stochastic Triplet Embedding (t-STE) [50] to generate an *Outsourced idea map* from these responses. Although the *Integrated* and the *Outsourced* idea maps were generated from different forms of human input (spatial arrangements in the *Integrated* condition and triplet comparisons in the *Outsourced* condition), the algorithms used to aggregate the results (t-SNE [49] in the *Integrated* condition and t-STE [50] in the *Outsourced condition*) are both based on the same mathematical insights and should yield results with closely comparable characteristics. Thus, the key question at hand is whether collecting implicit semantic judgments from an integrated secondary task yields data of sufficient coverage and quality to build a semantic model that is at least as accurate as building a model from explicit semantic judgments collected from the external workers.

Measures and analysis

To compare the two idea maps, we measured each map against a standard baseline for comparison, which is a set of pairwise similarity ratings between ideas generated by independent human judges [13]. This similarity rating method yields accurate assessments of pairwise similarity among ideas and serves as an excellent gold standard. It is not a scalable mechanism for constructing semantic models in the first place, however, because the number of pairwise comparisons it requires grows quadratically with the number of ideas.

To obtain these independent similarity ratings, we posted 66 MTurk tasks to recruit workers (who have not previously participated in any of our other studies) to rate similarity of 550 pairs of ideas, randomly sampled across all participants, on a scale from 1 (not at all similar) to 7 (very similar). We provided a rubric with example pairs of ideas and their desired ratings. Each rater assessed 29 pairs of ideas, four of which were examples of pairs of ideas that we showed in the rubric (so that we could see if they paid attention to the instructions). Each pair of ideas was rated by at least three raters. We standardized (i.e., converted to z-scores) the ratings for each rater prior to aggregating the results. After excluding 2 workers whose answers to the rubric questions indicated that they were not paying close attention to the task, we were left with 1,725 similarity ratings.

We then computed the correlations between the human similarity ratings and the pairwise distances among ideas from each idea map. To test for potential statistical difference between the two correlations, we transformed the correlations into z-scores using Fisher's r-to-z transformation.

Results

We found a significant correlation (Spearman correlation, $\rho = -0.4848$, $p < .0001$) between distances from the *Integrated idea map* and the human similarity ratings, and a significant correlation (Spearman correlation, $\rho = -0.3878$, $p < .0001$) between distances from the *Outsourced idea map* and the human similarity ratings. Note that map distances capture *differences* among ideas while the participants were asked to assess similarity, so the negative correlation coefficient is the desirable outcome.

After transforming the correlations using Fisher's r-to-z transformation, we found the correlation between the *Integrated idea map* and human ratings to be significantly larger in magnitude than the correlation between the *Outsourced idea map* and human ratings ($z = 1.99, p = 0.046$). In other words, our proposed approach resulted in an idea map that better modeled the actual semantic relationships among the idea than the previous method [44] that relied on mass outsourced human computation tasks.

IDEAHOUND: CREATIVITY INTERVENTIONS ENABLED BY REAL-TIME SEMANTIC MODELING OF GENERATED IDEAS

Equipped with the capability to derive a computational model of semantic distances among contributed ideas, we have built IDEAHOUND, a system for collaborative ideation at scale. IDEAHOUND serves as a step towards our end goal of improved large-scale collaborative ideation. IDEAHOUND includes three creativity interventions enabled by the availability of a semantic model of generated ideas. These interventions are illustrated in Figure 6 and described here:

Diverse Inspirational Examples

When a user requests to see ideas of others, the system consults the global idea map and selects a set of three ideas that the user has not seen before. The requested ideas appear in the Others' ideas pane in the workspace (Figure 1B). Two of these ideas are substantially different according to the idea map (i.e., the distance between the two ideas on the map has to be greater than a specified threshold). The third idea is selected randomly from a pool of ideas that has been placed on none or the whiteboards. This procedure balances the need to collect judgements on newly contributed ideas and the need to present the users with ideas that are known to be substantially different from each other.

Similar Ideas Lookup

A user can request ideas similar to a particular idea by clicking on a request for similar idea button for that idea (Figure 1E). The system then consults the map to locate up to three ideas that are close to that idea (i.e., the distances between the ideas and the query idea do not exceed a specified threshold). The set of selected similar ideas will appear next to the query idea on the whiteboard (as in Figure 6b).

Visualization of the Solution Space

IDEAHOUND provides users with a visualization of an idea map (Figure 1F). The visualization shows dots on the map, each dot representing an idea. Ideas that are rendered close to each other are judged to be similar to each other. The system clusters ideas and shows a short text for the ideas that are centers of clusters to give user a quick overview of the space without cluttering the display with too many labels. Users can infer how much each part of the solution space has been explored by looking at the number of ideas in that area. They can zoom in to get a closer look at a particular region or zoom out to see an overview. The ideas submitted by the user are rendered in a different color from ideas by others to help the user see their contributions in context and decide on which direction to pursue next.

STUDY 4: INITIAL EVALUATION OF IDEAHOUND

To gauge the effectiveness of introduced interventions (and by extension, the usefulness of the semantic model produced by our integrated crowdsourcing approach), we ran a preliminary qualitative study to investigate how people use IDEAHOUND. This study is complementary to Study 3. While Study 3 verified the accuracy of the semantic model, Study 4 aims to provide a proof-of-concept demonstration that the semantic model generated with our approach can in fact support beneficial creativity interventions. The focus of this study is on users' experience and perception of the creativity interventions supported by the semantic model. Because we designed the study to simulate the early stages of an ideation process, we disabled looking up of similar ideas—an intervention that we hypothesized to be particularly useful in later stages of the ideation process. Thus, the focus of Study 4 is on the *Diverse inspirational examples* and *Map visualization of ideas in a solution space* interventions.

Participants

We recruited 7 participants (4 female) aged 18 to 32 through a call for participation sent to Harvard University students' mailing-lists. Participants were compensated $15 for taking part in the study.

Task

Participants generated ideas for April Fools pranks for their university. All participants worked as part of the same team. That is, they could see each others' ideas on IDEAHOUND.

Procedure

Each participant was given a link to access their workspace for the prank ideation task on IDEAHOUND. They then used IDEAHOUND to generate prank ideas in two 10-minute sessions at their own pace over the course of two days before the scheduled time for their individual in-person interviews. During the interview session, participants generated a few more ideas while thinking aloud for 5–10 minutes. They then filled out a short survey on their experience, and talked with the researcher about their experience and creative process. The entire interview session lasted about 30 minutes.

Results

Participants generated 115 ideas. On average, participants found the system somewhat helpful in helping them find inspirations from ideas of others (M=4.71, SD=1.74; 1 = not helpful and 7 = very helpful) and come up with ideas (M=4.57, SD=1.99; 1 = not helpful and 7 = very helpful).

Organizing ideas on the whiteboard

Six of the seven participants used the whiteboard to organize ideas. In the survey, five participants reported that they found the virtual whiteboard to be the most useful aspect of the system. Although the degrees of idea organization varied, participants who did not organize ideas as much reported that

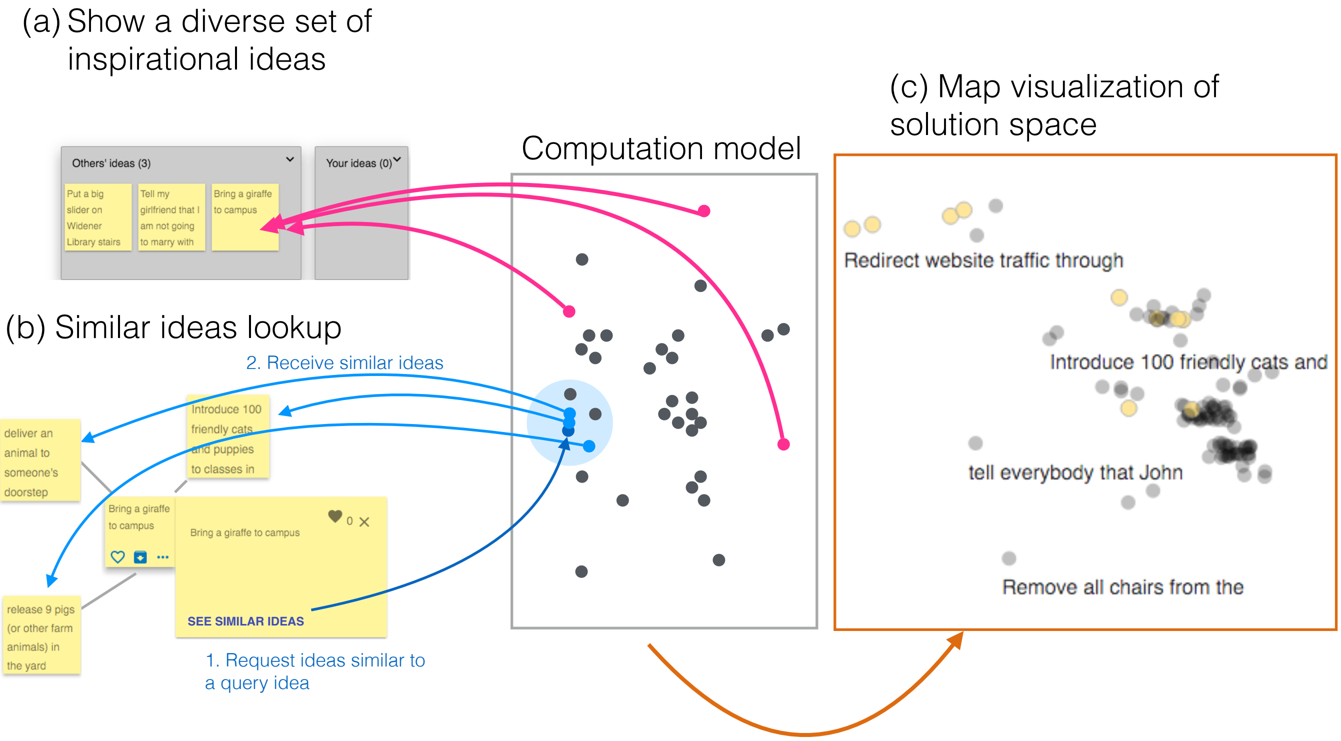

(a) Show a diverse set of inspirational ideas
Others' ideas (3)
Your ideas (0)
Put a big slider on Widener Library stairs
Tell my girlfriend that I am not going to marry with
Bring a giraffe to campus
Computation model
(c) Map visualization of solution space
Redirect website traffic through
Introduce 100 friendly cats and
tell everybody that John
Remove all chairs from the
(b) Similar ideas lookup
2. Receive similar ideas
deliver an animal to someone's doorstep
Introduce 100 friendly cats and puppies to classes in
Bring a giraffe to campus
Bring a giraffe to campus
0
SEE SIMILAR IDEAS
release 9 pigs (or other farm animals) in the yard
1. Request ideas similar to a query idea

Figure 6. Proposed interventions to improve experience and output of idea generation task as implemented by IDEAHOUND . The computational model box represents show the shape of the solution space through idea instances and their relationships. (a) When a user requests to see ideas of others, the system selects a set of diverse ideas (instead of sample randomly). (b) A user can ask to see a set of ideas that are similar to a certain idea. (c) A user can get a quick overview of the solution space through an map visualization that shows their ideas and ideas of others in the solution space.

they would have organized ideas more if they had been more invested in the task and had had more time.

Participants reported organizing ideas on the whiteboard as a way to "construct [their] mind map" [P2] and establish landmarks to come back to later [P4]. Organizing ideas on whiteboard helped them see relationships between ideas (n=4), detect patterns of emerging ideas (n=4), and to kill time while thinking about new ideas (n=2). One participant [P1] reported that they did not use the whiteboard to organize ideas because looking at others' ideas or his old ideas distracted him.

Getting inspired by seeing diverse ideas sampled from the computational model

Most participants found seeing ideas of others helpful in their idea generation process. They reported building on the ideas of others and they liked to "look at others' ideas for inspiration" [P3]. P2 commented that seeing ideas of others was especially useful when he ran out of ideas. None of the participants found provided ideas of others repetitive.

However, not everyone found seeing other people's ideas helpful. One participant [P1] did not use the example request features because he likes to start generating ideas from a "blank slate" without external influences.

Reading the idea map visualization

Participants had mixed reactions to the idea map visualization. Participants used the idea map visualization to get a quick overview of ideas submitted by others (n=5), to explore many different alternatives proposed by others (n=2), and to kill time while thinking about new ideas (n=4). They also used the idea map visualization to detect patterns of the solution space [P1] and discover underexplored part of the solution space for both individual [P2, P4, P5]. P2 commented that he tried to "look for space where his ideas are not located". Similarly, P4 and P5 stated that they tend to look at the area of the map with few ideas.

However, participants also pointed out limitations of the current version of the idea map visualization. Participants sometimes had a hard time seeing the connections between ideas that were placed close to each other on the map and expressed interest in getting an explanation of the relationships between ideas. They also mentioned that it was tricky to select the ideas that were not centers of the clusters because the size of the dots and they wished the idea map visualization would allow them to open the detail windows for more than one idea at a time.

DISCUSSION AND FUTURE WORK

Integrating Idea Generation and Organization Into a Single Activity

Results of Study 1 demonstrated that members of volunteer communities may not always be motivated to perform work that is necessary for the good of the community, but which is perceived as tedious and as detracting from the primary interest of the community. In our case, people who were intrinsically interested in contributing novel ideas were not motivated to evaluate ideas generated by others. We have thus created an alternative interface for idea generation, one that seamlessly integrated evaluation of ideas with the primary task of idea generation.

The results from Studies 2 and 3 show that our integrated approach can model semantic relationships between ideas more accurately than a previously validated crowdsourced approach [44] with minimal impact on users' ideation experience. Although participants initially reported exerting higher mental effort with the *Integrated* system than with the conventional *Single-task* system, as the participants acclimated to the novel interface over subsequent sessions, the difference in mental effort nearly disappeared. Additionally, participants in the *Integrated* condition did not think that organizing ideas detracted from their primary task of generating ideas. This is in contrast to the results from Study 1, which demonstrated that people were not willing to evaluate ideas of others if they perceived it as an additional task. Consistent with our initial formative studies, the results from Study 4 also suggested that organizing ideas on the whiteboard *helped* idea generation by encouraging the users to make sense of the solution space upfront. A longer study could help verify whether this is the case.

Meaningfulness of clusters

We expected the clusters that users generate to be meaningful because users organize ideas in IDEAHOUND only when it is helpful to them. However, during our formative studies, we observed that not all clusters were of equal quality. In some clusters, it was unclear why the ideas were grouped together. Introducing an affordance for adding explicit labels to clusters helped reduce this problem. Yet, a small fraction of clusters in Study 2 were not labeled. When a cluster was unlabeled, it was not always immediately clear how to derive meaning from it. Excluding all unlabeled clusters from the input to our model might improve the quality of the models by reducing noise, but it might also decrease the accuracy of the models by taking away data. Our initial experiments, in which we manually flagged and removed "noisy" clusters, did not substantially impact the quality of the resulting models. However, we plan to more systematically investigate mechanisms that can help identify and filter non-meaningful clusters to further improve the quality of the resulting idea map.

Scalability

While we only showed the viability our approach in small groups of 6–10 people, this approach should also be applicable for larger ideation groups. The amount of human input required by our approach to create a semantic model of the solution space grows linearly in the number of ideas (as explained in [44]). In our method, because ideators are also organizers, the amount of input provided to organize ideas grows at the same rate as the number of ideas, so we expect no computational barriers for our system to scale. Two of our interventions (diverse inspirational examples and ability to lookup similar ideas) will not be affected negatively by the size of the community, but the idea map visualization will need to be revised so that it is still readable even if thousands of ideas are present.

Creativity Interventions

One might wonder why the *Integrated* intervention in Study 2 did not improve creative performance, as the results of prior work [44] would predict. A closer inspection of the example sets presented to the participants in the *Integrated* condition reveals that they were not significantly more diverse than those in the *Single-task* condition. The difference may be attributable to the way we sampled the pairs of "diverse" examples from an idea map: our algorithm first picked an idea at random from the idea map and then searched for another idea that the model predicted to be maximally dissimilar to the first. But because mundane ideas are, by definition, substantially more prevalent than unusual ones, the first randomly selected idea was almost always fairly mundane. Given that the third idea was chosen at random from among the most recently-generated ideas, this sampling approach resulted in sets of inspirational ideas that were not substantially different from those picked entirely at random. A better approach, will be to first randomly select distinct regions on the idea map (independently of the density of ideas in those regions, thus not privileging common ideas) and then sample an idea from each of these regions.

The most novel intervention we tested was the idea map visualization, which presented a succinct synthesis of the ideas explored by the community so far. Participants in Study 4 found this idea map visualization useful in providing an overview of the solution space. Some participants [P4, P5] used the visualization to identify underexplored parts of the solution space and to decide how best to contribute to the group effort. Thus, an idea map visualization can act as a guide to coordinate group ideation effort by directing people to explore different parts of the map to avoid redundant work.

The results of Study 4 also suggest ways to improve the interface of the idea map visualization. Specifically, participants sometimes did not understand why certain ideas appeared close to each other on the idea map visualization and would have liked to see explanations of the semantic relationships implied by the visualization. This finding is likely explained by the fact that different people appeared to construct different mental models of the emerging solution space. In Studies 2 and 4 we repeatedly observed that different participants grouped the same ideas differently. This observation is consistent with prior findings [1]. While this still allowed the algorithm to create a computational model that captured meaningful semantic relationships among ideas, it suggests that a grouping that is intuitive to one participant may be surprising to another. In the future, we will leverage the labels participants attach to the clusters they create. As these labels explicitly reveal shared semantics among a group of ideas, they may be the right vocabulary with which to communicate the rationale behind different clusters on the idea map visualization.

CONCLUSION

Prior work on creative cognition and creativity support tools demonstrated that having a computational semantic model of a solution space can enable a number of interventions that demonstrably improve the number, quality and diversity of ideas people generate. In large-scale online innovation platforms, where people contribute ideas in the form of short text snippets or sketches, no prior feasible mechanism existed for creating such computational models. We contribute such a mechanism: it combines human judgements with machine learning to estimate similarity among all ideas contributed by a community. Because people were not willing to contribute subjective judgements of idea similarity when they perceived this to be a separate task unrelated to the primary activity of idea generation, we developed a novel system, called IdeaHound, which seamlessly integrates the secondary task of providing feedback on semantic relationships among ideas into the primary task of idea generation.

The results of our studies demonstrate the viability of our approach. We found that people were as willing to use IdeaHound to simultaneously generate and organize ideas as they were a conventional design that did not require organizing ideas. Furthermore, the subjective judgements implicitly collected through IdeaHound resulted in a more accurate computational model of semantic relationships among ideas than an existing approach [44], which relied on outsourcing the task to an external crowd.

We also show how this computational model can support creative interventions that users find useful, specifically, providing diverse inspirational examples, and providing an overview of the solution space in the form of an idea map visualization.

In future work, we plan to explore different interactions that extract other useful information that helps the ideation process—such as idea quality and semantic attributes—and other creative interventions that improve users' experience and creative output such as coordinating group effort. We also intend to explore in more depth how access to a shared representation of the emerging solution space can impact coordination of collaborative ideation. We envision an intelligent self-sustainable ideation platform that integrates these techniques in all parts of the innovation pipeline, from problem structuring, to ideation, to feedback, selection and prototyping. Such a framework can enhance creative experience for users at both individual and community level.

ACKNOWLEDGEMENTS

The authors would like to thank Kenneth Arnold and Ofra Amir for valuable discussions. This work was funded in part by awards from the National Science Foundation (IIS-1208382, and IIS-1122206).

REFERENCES

1. André, P., Kittur, A., and Dow, S. P. Crowd synthesis: Extracting categories and clusters from complex data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work*, CSCW '14, ACM (New York, NY, USA, 2014), 989–998.
2. Bjelland, O. M., and Wood, R. C. An inside view of IBM's' innovation jam'. *MIT Sloan management review 50*, 1 (2008), 32–40.
3. Blei, D. M. Probabilistic topic models. *Communications of the ACM 55*, 4 (Apr. 2012).
4. Boudreau, K. J., and Lakhani, K. R. Open disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology. *Research Policy 44*, 1 (2015), 4–19.
5. Bragg, D., Rector, K., and Ladner, R. E. A user-powered american sign language dictionary. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work &#38; Social Computing*, CSCW '15, ACM (New York, NY, USA, 2015), 1837–1848.
6. Chan, J., Fu, K., Schunn, C. D., Cagan, J., Wood, K. L., and Kotovsky, K. On the benefits and pitfalls of analogies for innovative design: Ideation performance based on analogical distance, commonness, and modality of examples. *Journal of Mechanical Design 133* (2011), 081004.
7. Chan, J., and Schunn, C. The impact of analogies on creative concept generation: Lessons from an in vivo study in engineering design. *Cognitive Science 39*, 1 (2015), 126–155.
8. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems* (2009), 288–296.
9. Chaudhuri, S., Kalogerakis, E., Giguere, S., and Funkhouser, T. Attribit: content creation with semantic attributes. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, ACM (2013), 193–202.
10. Chilton, L. B., Kim, J., André, P., Cordeiro, F., Landay, J. A., Weld, D. S., Dow, S. P., Miller, R. C., and Zhang, H. Frenzy: Collaborative data organization for creating conference sessions. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, ACM (New York, NY, USA, 2014), 1255–1264.
11. Chuang, J., Ramage, D., Manning, C., and Heer, J. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012), 443–452.
12. Cohen, J. *Statistical power analysis for the behavioral sciences (rev*. Lawrence Erlbaum Associates, Inc, 1977.
13. Dow, S., Glassco, A., Kass, J., Schwarz, M., Schwartz, D., and Klemmer, S. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *Transactions on Computer-Human Interaction (TOCHI 17*, 4 (2010).

14. Gerber, D. J., Lin, S.-H. E., Pan, B. P., and Solmaz, A. S. Design optioneering: multi-disciplinary design optimization through parameterization, domain integration and automation of a genetic algorithm. In *Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design*, Society for Computer Simulation International (2012), 11.

15. Goldstone, R. An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers 26*, 4 (1994), 381–386.

16. Hout, M. C., Goldinger, S. D., and Ferguson, R. W. The versatility of spam: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General 142*, 1 (2013), 256.

17. Huang, C.-Z. A., Duvenaud, D., Arnold, K. C., Partridge, B., Oberholtzer, J. W., and Gajos, K. Z. Active learning of intuitive control knobs for synthesizers using gaussian processes. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, ACM (New York, NY, USA, 2014), 115–124.

18. Huang, C.-Z. A., Duvenaud, D., and Gajos, K. Z. Chordripple: Recommending chords to help novice composers go beyond the ordinary. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, IUI '16, ACM (New York, NY, USA, 2016), 241–250.

19. Jansson, D. G., and Smith, S. M. Design fixation. *Design Studies 12*, 1 (1991), 3–11.

20. Javadi, E., Mahoney, J., and Gebauer, J. The impact of user interface design on idea integration in electronic brainstorming: an attention-based view. *Journal of the Association for Information Systems 14*, 1 (2013), 1–21.

21. Kittur, A., Lee, B., and Kraut, R. E. Coordination in Collective Intelligence: The Role of Team Structure and Task Interdependence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, ACM (New York, NY, USA, 2009), 1495–1504.

22. Klein, M., and Garcia, A. C. B. High-speed idea filtering with the bag of lemons. *Available at SSRN 2501787* (2014).

23. Kohn, N. W., and Smith, S. M. Collaborative fixation: Effects of others' ideas on brainstorming. *Applied Cognitive Psychology 25*, 3 (2011), 359–371.

24. Komarov, S., and Gajos, K. Z. Organic peer assessment. In *Proceedings of the CHI 2014 Learning Innovation at Scale workshop* (2014).

25. Komarov, S., Reinecke, K., and Gajos, K. Z. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, ACM (New York, NY, USA, 2013), 207–216.

26. Kraut, R. E., Resnick, P., Kiesler, S., Burke, M., Chen, Y., Kittur, N., Konstan, J., Ren, Y., and Riedl, J. *Building successful online communities: Evidence-based social design*. Mit Press, 2012.

27. Lee, B., Srivastava, S., Kumar, R., Brafman, R., and Klemmer, S. R. Designing with interactive example galleries. In *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM Request Permissions (Apr. 2010).

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

29. Majchrzak, A., and Malhotra, A. Towards an information systems perspective and research agenda on crowdsourcing for innovation. *The Journal of Strategic Information Systems 22*, 4 (Dec. 2013), 257–268.

30. Marks, J., Andalman, B., Beardsley, P. A., Freeman, W., Gibson, S., Hodgins, J., Kang, T., Mirtich, B., Pfister, H., Ruml, W., et al. Design galleries: A general approach to setting parameters for computer graphics and animation. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co. (1997), 389–400.

31. Marsh, R. L., Landau, J. D., and Hicks, J. L. How examples may (and may not) constrain creativity. *Memory & Cognition 24*, 5 (1996), 669–680.

32. Mendels, P., Frens, J., and Overbeeke, K. Freed: a system for creating multiple views of a digital collection during the design process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2011), 1481–1490.

33. Nickerson, J. V., Corter, J. E., Tversky, B., Zahner, D., and Rho, Y. J. The spatial nature of thought: Understanding information systems design through diagrams. In *in Boland, R., Limayem, M., and Pentland B.,(eds), Proceedings of the 29th International Conference on Information Systems*, Citeseer (2008).

34. Nickerson, J. V., and Yu, L. L. Going Meta: Design Space and Evaluation Space in Software Design. In *Software Designers in Action: A Human-Centric Look at Design Work*, A. van der Hoek and M. Petre, Eds. Chapman and Hall/CRC, Rochester, NY, Sept. 2012, 323–344.

35. Nijstad, B. A., De Dreu, C. K. W., Rietzschel, E. F., and Baas, M. The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology 21* (2010), 34–77.

36. Nijstad, B. A., Stroebe, W., and Lodewijkx, H. F. Cognitive stimulation and interference in groups: Exposure effects in an idea generation task. *Journal of experimental social psychology 38*, 6 (2002), 535–544.

37. O'Donovan, P., Agarwala, A., and Hertzmann, A. Designscape: Design with interactive layout suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM (2015), 1221–1224.

38. O'Donovan, P., Lībeks, J., Agarwala, A., and Hertzmann, A. Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 92.

39. Reinecke, K., and Gajos, K. Z. Labinthewild: Conducting large-scale online experiments with uncompensated samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, ACM (New York, NY, USA, 2015), 1364–1378.

40. Riedl, C., Blohm, I., Leimeister, J. M., and Krcmar, H. Rating scales for collective intelligence in innovation communities: Why quick and easy decision making does not get it right. In *Proceedings of Thirty First International Conference on Information Systems* (2010).

41. Rietzschel, E. F., Nijstad, B. A., and Stroebe, W. Relative accessibility of domain knowledge and creativity: The effects of knowledge activation on the quantity and originality of generated ideas. *Journal of Experimental Social Psychology 43*, 6 (2007), 933–946.

42. Shaffer, J. P. Multiple hypothesis-testing. *Annual Review of Psychology 46* (1995), 561–584.

43. Sharmin, M., and Bailey, B. P. Reflectionspace: an interactive visualization tool for supporting reflection-on-action in design. In *Proceedings of the 9th ACM Conference on Creativity & Cognition*, ACM (2013), 83–92.

44. Siangliulue, P., Arnold, K. C., Gajos, K. Z., and Dow, S. P. Toward collaborative ideation at scale—leveraging ideas from others to generate more creative and diverse ideas. In *Proceedings of CSCW'15* (2015).

45. Siangliulue, P., Chan, J., Gajos, K., and Dow, S. P. Providing timely examples improves the quantity and quality of generated ideas. In *Proceedings of the ACM Conference on Creativity and Cognition* (2015).

46. Sio, U. N., Kotovsky, K., and Cagan, J. Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies 39* (July 2015), 70–99.

47. Talton, J. O., Gibson, D., Yang, L., Hanrahan, P., and Koltun, V. Exploratory modeling with collaborative design spaces. *ACM Transactions on Graphics-TOG 28*, 5 (2009), 167.

48. Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. T. Adaptively Learning the Crowd Kernel. *arXiv.org* (May 2011).

49. Van Der Maaten, L. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research 15*, 1 (2014), 3221–3245.

50. Van der Maaten, L., and Weinberger, K. Stochastic triplet embedding. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, IEEE (2012), 1–6.

51. Vuculescu, O., and Bergenholtz, C. How to Solve Problems with Crowds: A Computer-Based Simulation Model. *Creativity and Innovation Management 23*, 2 (June 2014), 121–136.

52. Weir, S., Kim, J., Gajos, K. Z., and Miller, R. C. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, ACM (New York, NY, USA, 2015), 405–416.

53. Wisdom, T. N., and Goldstone, R. L. Innovation, Imitation, and Problem Solving in a Networked Group. *Nonlinear Dynamics-Psychology and Life Sciences 15*, 2 (2011), 229.

54. Yu, L., Kittur, A., and Kraut, R. E. Distributed analogical idea generation: inventing with crowds. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, ACM (2014), 1245–1254.