Comparing Different Sensemaking Approaches for Large-Scale Ideation

ABSTRACT

Platforms that support large-scale idea generation often expose ideators to previous solutions. However, research suggests people generate better ideas if they see abstracted solution paths—generated through human sensemaking processes (e.g., descriptions of solution approaches)—rather than being inundated with all prior ideas. Automated (and semi-automated) methods can also offer interpretations of earlier solutions. To explore relative benefits of different sensemaking approaches, we conducted an online study where 245 participants generated ideas for two problems in one of five conditions: 1) no stimuli, 2) exposure to all prior ideas, or solution paths extracted from prior ideas using 3) a fully automated workflow, 4) a hybrid human-machine approach, and 5) a fully manual approach. The results show, contrary to expectations, that human-generated solution paths do not improve ideation (as measured by fluency and breadth of ideation) over simply showing all ideas. Machine-generated solution paths sometimes significantly improve fluency and breadth of ideation over no ideas (although at some cost to idea quality). These findings suggest that automated sensemaking can improve idea generation, but we need more research to understand the value of human sensemaking for crowd ideation.

Author Keywords

Creativity; crowdsourcing; brainstorming; sensemaking

ACM Classification Keywords

Human-centered computing~Collaborative and social computing

INTRODUCTION

Large-scale ideation platforms have shown much potential for solving difficult problems by leveraging the scale and diversity of online crowds [3,6,32,53]. Current approaches excel at collecting many solutions, especially in a competitive paradigm where crowd innovators work in isolation from each other. However, a significant concern with this approach is that it often results in redundant ideation and wasted effort as newcomers retrace obvious but suboptimal solution paths [32]. For example, in their recent 10 to the 100^{th} crowdsourced innovation project, Google had to recruit 3,000 of their employees to prune the 150,000 ideas received from the crowd, pushing the project nine months behind schedule.

Cognitive research on creative ideation suggests that people can produce better ideas if they are able to learn from the efforts of others, recombining ideas into new ideas and iterating on new ideas to improve them [17,19,26,48,49]. However, at crowd scale, simply exposing all ideas to everyone may not be the most effective strategy. Ideators may not have sufficient time or cognitive resources to sift through potentially hundreds to thousands of other ideas, or only superficially process and build on ideas rather than leveraging them to generate new insights [30,31]. The presence of superficial details in raw ideas might also lead to cognitive fixation [29,39,50]. For these reasons, abstracted solution paths — which distill essential solution approaches from a number of different ideas while avoiding superficial details — may provide a better way for ideators to interact with prior ideas.

Available sensemaking strategies for abstracting solution paths can be placed on a hypothesized cost-quality tradeoff continuum, where the quality (especially its benefit for ideation) is a function of the strategy's cost. At the high end of the continuum (high cost, hypothesized high quality), there is a mature set of design synthesis strategies (e.g., affinity diagramming) employed effectively by design teams to make sense of a solution space [4,25,33]; these strategies require significant manual human effort, and therefore may be hard to scale to crowd ideation. However, these sensemaking outputs should hypothetically have high value for ideation. At the low end (low cost, hypothesized low quality) sit a number of automated approaches (e.g., unsupervised machine learning methods like Latent Semantic Indexing (LSI) [16] and K-means clustering) that can extract semantic themes in text-based idea sets very quickly and efficiently. However, the intrinsic quality of these themes (e.g., as measured by correspondence with gold standard human clusters/categories) is often relatively low (certainly lower than human-produced themes). Thus, these sensemaking outputs should hypothetically have relatively low value for ideation. In between sit hybrid approaches (medium cost, medium quality) that combine human and

Submitted to CHI2016

machine effort [28,34,52,56,57] Some of these approaches may require optimization of complex machine learning parameters to be successful, limiting their accessibility to large scale ideation platforms. But, relatively simple and accessible workflows could also be devised, where humans label initial machine clusterings formed using algorithms like K-means). The range of options for synthesizing existing ideas into abstracted solutions paths motivates us to explore two main research questions:

1) Do abstracted solution paths inspire more and better ideas compared to seeing all raw ideas or no ideas at crowd scale?
2) If so, how does the sensemaking strategy for abstract solution paths—from fully automated to hybrid human-machine to fully manual—affect the quantity and quality of generated ideas?

In an online study, 245 participants from Amazon Mechanical Turk (mTurk) generated ideas for two innovation problems (generating ideas for a novel fabric technology and for improving the mobile experience mTurk) in one of five conditions: 1) no stimulation (**control**), 2) viewing all prior ideas for the problem (**all ideas**), 3) with solution paths extracted from human-labeled paths generated for human-generated clusters of ideas (**human-human**), 4) human-labeled paths generated for machine-generated clusters (**machine-human**), or 5) machine-labeled paths generated for machine-generated clusters (**machine-machine**).

Our results show that, contrary to expectations, human-generated solution paths do not improve ideation over simply showing all ideas, as measured by fluency (total number of ideas) and breadth of ideation (mean pairwise distance in LSI representation of the solution space). Moreover, machine-generated solution paths sometimes improve fluency and breadth of ideation over no ideas (although at a slight cost to idea quality). These findings suggest that large scale ideation could benefit from automated sensemaking, but more research is needed to better understand the value of human-generated sensemaking in a crowd setting.

This paper contributes:

1) Empirical findings on the relative value of different sensemaking approaches for ideation
2) Evidence that a simple, easy to implement approach to automated sensemaking (using LSA and k-means clustering) can improve fluency and breadth of ideation.

RELATED WORK

Effective Collaborative Ideation

There is consensus in the literature on creative cognition seeing what others have thought and/or are currently thinking about the problem increases people's ability to generate creative ideas (i.e., ideas that are both novel and of high quality) [17,19,26,48,49]. For example, people can generate more creative ideas when they have access to example designs [36,38,46–48], draw analogies to past experiences [10,15,27,58], or see some ideas from others who are working on the same problem [7,17]. Experiments with artificial problem-solving tasks (e.g., solving picture puzzles) also suggest that building on others' ideas not only improves individual creativity, but also maximizes the *community's* ability to reach an optimal solution [55].

However, these benefits of collaborative ideation can be challenging to realize at crowd scale (typically hundreds to thousands of ideas). At this scale, interesting but statistically rare ideas may be less likely to be noticed and built upon, leading the crowd to focus on common ideas. From a cognitive perspective, effective collaborative ideation depends on being able to attend to and deeply process other ideas [30,48]. But people are limited in their capacity to process information [14,42], and the number of ideas produced at crowd scale certainly exceeds this capacity. When given too many ideas as potential inspiration, people may stop attending to them, or only build on them in superficial ways [30,48]. Ideally, we would like crowd ideators to focus on the primary task of generating ideas, rather than expending most of their effort making sense of a large volume of ideas in order to extract useful inspiration. Another issue is that the presentation of the ideas in their "raw form", which includes many superficial details, can unduly constrain ideators' search to ideas closely related to the idea [29,39,50]. Therefore, we hypothesize that simply exposing all ideas to all contributors, while simple to implement, is likely to be an ineffective solution.

Higher-level abstracted solution paths that distill essential solution approaches shared by a number of different raw ideas may be a good alternative to raw ideas. A classic example of a solution path in the cognitive science literature is the convergence solution schema (successfully attack a single target by converging from multiple points if a single, focused attack is not feasible), abstracted from a variety of situations (e.g., generals attacking a fortress via mined bridges radiating from the fortress, a doctor destroying a tumor with radiation rays without also destroying healthy tissue) [22]. Abstractions such as categories can be a powerful way to compress large volumes of information into more manageable chunks [11,41], greatly reducing the number of "bits" that ideators need to process. This should increase the probability that ideators will actually be able to attend to and benefit from them. Also, by representing primarily the "essential" information that helps to organize ideas, such as in schemas [22,58], abstractions could help reduce fixation on superficial details.

Method	Fabric Display	Improve Turk
Machine-Machine	advertisements billboards create shirts/pants cuffs companies	time timer site assigned loading notification
Machine-Human	use for new locations for advertising	send timer alerts
Human-Human	advanced display of advertisements	Provide notifications for time-sensitive activity

Table 1. Example solution paths extract by each method for both problems. For comparability, we select examples that are close in semantic meaning.

Some studies have shown that provide one or a few manually generated and selected abstractions can lead to better creative performance compared to showing ideas in their raw form [40,54,58]. However, this manual selection may not always be feasible at crowd scale. In this paper, we extend this prior work by examining the value viewing many (~15-20) abstractions rather than just one or two.

Automated Sensemaking Strategies

Numerous automated sensemaking methods exist, ranging from relatively *mature* (e.g., well-studied, shown to be robust across a range of settings) and *simple* (i.e., produces reasonable results without requiring significant tuning of many complex model parameters) vector-space models like Term-Frequency Inverse-Document Frequency (TF-IDF) and LSI [16] models and clustering algorithms like K-means and agglomerative hierarchical clustering, to more sophisticated methods like probabilistic topic models [5].

Research on these methods has largely focused on improving correspondence with "gold standard" human models of the items being structured (e.g., gold standard human-generated categories [8]), or matching human performance on various benchmark tasks (e.g., simple A:B, C:D word analogies [43]). Some work has examined the value of automated sensemaking outputs for complex tasks like ideation [18,21], intelligence analysis [51], and conducting scientific literature reviews [12]. However, relatively little work has compared the value of automated sensemaking outputs to human-produced outputs for complex tasks in general. One notable exception is a study by André et al [2], who found that simple TF-IDF sensemaking over academic papers could provide suggestions for papers to attendees at an academic conference that were rated at comparable levels of relevance as suggestions provided from a human-generated sensemaking model (via partial clustering). We are not aware of any such in the context of ideation. Such research is needed to adequately reason about cost-quality tradeoffs involved in selecting automated vs. manual sensemaking approaches.

SOLUTION PATH EXTRACTION

Our goal is to explore the cost-quality curve for current "off-the-shelf" methods that would be readily available to crowd innovation platforms. Here we describe in more detail each of the three points on the cost-quality curve that we select for comparison: 1) manual human-human (highest cost), 2) machine-human (medium cost), and 3) machine-machine (lowest cost). Each of these specifies a different combination of how ideas are clustered into solution paths (cluster phase)**,** and then how those solution paths get described (label phase), i.e., whether humans or machines complete each phase.

Datasets

We extracted solution paths for ideas previously collected for two problems: 1) new product ideas for a novel "fabric display" (fabric display problem), and 2) ideas for improving workers' experience performing HITs on mTurk on mobile devices (improve mTurk problem). These problems are representative of problems typically addressed in crowd innovation platforms.

Both idea datasets were assembled from idea datasets collected in prior studies ([47] for the fabric display problem, and [35] for the improve Turk problem). The ideas in those datasets were collected from mTurk workers, and the authors shared the data with us. We randomly sampled 120 ideas for each problem. Examples of solution paths extracted by each method are shown in Table 1.

Sensemaking Methods

Machine-Machine Solution Path Extraction

For this method of extracting solution paths, machines complete both the **cluster** and **label** phases of solution extraction. Our goal is to design an automated workflow that closely approximates realistic crowd ideation scenarios. In realistic scenarios, one very rarely has "gold standard" data available to assist with evaluation and tuning of complex model parameters. Therefore, we would like to choose relatively mature models with well-known properties that can work relatively well "out-of-the-box" without significant parametric tuning. To meet these constraints, we chose to do unsupervised feature identification using a combination of TF-IDF (widely used in information retrieval; no parameter settings required) and LSA (widely used in information retrieval; only choose number of dimensions), and clustering using the standard K-means algorithm (widely applicable; only choose number of clusters). In our workflow, we first identify semantic features in the set of ideas using TF-IDF and LSA, and then use these features to achieve unsupervised clustering of the ideas with K-means. We then automatically obtain labels

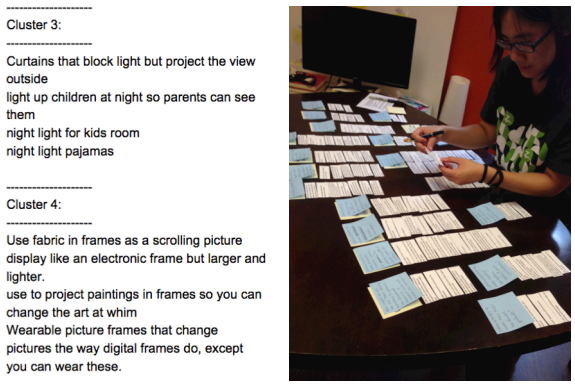

Figure 1. Screenshot of human labeling task (left) physical setup of human cluster-label task (right)

for the clusters by finding the most informative keywords within each cluster.

In the **cluster** phase, ideas are first tokenized, removing uninformative (stop) words, and then weighted using TF-IDF. Each idea then is re-represented as a vector of *w* TF-IDF weights (where *w* is the number words in the corpus of ideas). We then use LSA to reduce the dimensionality of the vectors. Based on prior experience clustering brainstorming ideas, our intuition is that somewhere between 15 to 25 patterns is often sufficient to adequately describe emerging solution patterns. Therefore, we set our LSA model to retain 15, 20 and 25 dimensions. For each of these parameter settings, we then identify solution paths by finding the same number of clusters of ideas (i.e., 15, 20 or 25) with the K-means clustering algorithm. We use the LSA dimension weights as features for the ideas. Rather than report the running time (which would vary by machine, dataset, and implementation of the algorithms), we leave it to the reader to examine the well-studied computational complexity properties of LSA and K-means and extrapolate accordingly to their desired settings.

For the **label** phase, we again select a method that would not require significant parameter tuning. To do this, we treat each cluster as a document, and compute TF-IDF weights for all words. We then choose the top *n* words with the highest TF-IDF weights within each cluster. *N* is determined by the average length of human labels (minus stopwords). The intuition behind this method is to identify words that distinguish between the clusters.

Machine-Human Solution Path Extraction

For this method, the machines complete the **cluster** phase, but humans are recruited for the **label** phase. The **cluster** phase is the same as the machine-machine method (i.e., clusters generated by combination of LSA and K-means). In the **label** phase, the labelers received each of the *M* clusters of ideas identified in the machine clustering phase. The labeling task was completed in a Google Document, which contained 1) instructions for labeling, 2) a description of the problem, and 2) each of the *M* clusters as an unlabeled list, with a blank header. Labelers described each cluster by typing a description in the blank header (see Fig. 1, left panel).

Labelers received the following instructions: *"Here is a set of 120 ideas for the problem, divided into clusters of ideas. Each cluster of ideas is meant to represent a solution pattern for the problem. Recall that we would like to provide these patterns as inspiration for future brainstormers who will also work on this problem. Your goal is to write a meaningful description for each cluster. Good descriptions capture the essential shared theme of a group of ideas in an appropriately specific fashion (i.e., contain enough detail to be useful to future brainstormers, but are not simple restatements of each idea). Ideally, they capture a theme shared by all ideas in their cluster; in some (rare) cases, you may find it difficult to discern the common theme: in this case, please do your best to describe a pattern shared by at least 2-3 ideas in the cluster. As a good test, good descriptions can fit well into the template* 'How can we _____ to solve the problem?'.

A different set of 3 research assistants — also advanced HCI design students — served as labelers. Labelers on average took approximately 28 minutes to complete labeling for each problem.

This approach represents a midpoint on the cost-quality tradeoff curve, and also allows us to potentially tease apart the value provided by clustering (which influences the semantics of the labels produced) and labeling (which vary in such features as coherence, specificity, and phrasing) provided by machines and humans.

Human-Human Solution Path Extraction

For this method, humans performed both the **cluster** and **label** phases of solution path extraction in a single task. The clusterer-labelers received all 120 ideas in a single stack. Each idea was printed out on a slip of paper. Clusterer-labelers then sorted ideas into clusters on a 45" by 45" table, and labeled clusters by writing descriptions on Post-It notes. Figure 1 (right panel) shows the physical setup of the task.

Clusterer-labelers received a description of the problem as well as the following instructions: *"Here is a set of 120 ideas for the problem. Please identify solution patterns in the set of ideas. We would like to provide these patterns as inspiration for future brainstormers who will also work on this problem. Do this by grouping ideas on the whiteboard into clusters that define patterns and writing descriptions of those patterns. Good descriptions capture the essential shared theme of a group of ideas in an appropriately specific fashion (i.e., contain enough detail to be useful to future brainstormers, but are not simple restatements of each idea). As a good test, good descriptions can fit well into the template* 'How can we _____ to solve the problem?' *You may identify clusters as large or as small as you like (even singleton clusters). Please do not create miscellaneous clusters or sort ideas by quality."*

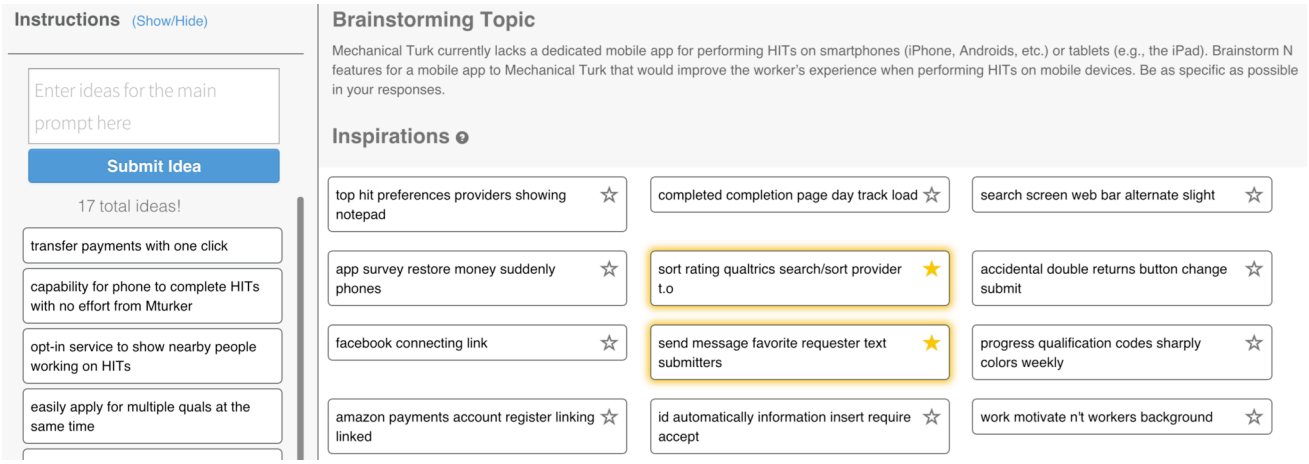

Figure 2. Screenshot of ideation interface. Actual inspirations from the machine-machine condition are shown. Participants enter ideas on the left, and view inspirations on the right. Participants can star inspirations they find useful.

Three research assistants — advanced HCI design students — performed this task. On average, clustering and labeling a single problem took approximately 52 minutes, significantly longer than just labeling. An average of 22 solution paths were identified for the fabric display problem, and 16 for the improve turk problem. Note that this task could be performed using digital tools; however, we chose to implement these steps on paper to have as close a match as possible to existing "off-the-shelf" human sensemaking approaches (e.g., affinity diagramming).

IDEATION EXPERIMENT

Overview

With extracted solution paths, we then conducted an online ideation experiment to compare the relative ideation value of the different sensemaking approaches, compared to simple exposure to all prior ideas, or no stimulation.

Method

Participants

We recruited 139 mTurk workers (42% female, mean age = 32.6 years, *SD* = 9.6) for this study. To ensure quality data, all participants had to have approval rates of at least 95% with at least 100 completed HITs.

Study Design

Participants were randomly assigned to one of 5 conditions: 1) unconstrained (**control,** N=21), 2) viewing all 120 prior ideas for the problem (**all-ideas,** N=24), viewing 3) machine-labeled solution paths generated for machine-generated clusters (**machine-machine**: least costly**,** N=32), 4) human-labeled paths generated for machine-generated clusters (**machine-human**: somewhat costly**,** N=34), or 5) with human-labeled paths generated for human-generated clusterings of ideas (**human-human**: most costly**,** N=28).

Within the path conditions, participants were randomly assigned one of the 3 path sets generated using that solution path extraction method.

Brainstorming Task

Participants generated ideas for both the fabric display and improve Turk problems.

Brainstorming Interface

Participants generated ideas using a simple ideation interface. Inspirations (whether ideas or solution paths) were provided to participants in an "inspiration feed" in the right panel of their interface (see Figure 2). Participants could "bookmark" particular inspirations that they found helpful. No sorting and filtering options were provided. The limited sorting/filtering available in the **all-ideas** condition might be considered primitive, but it is actually similar to many existing platforms that might provide rudimentary filtering by very broad, pre-defined categories (that may be reused across problem). Participants in the **control** condition generated ideas with a simpler ideation interface that removed the inspiration feed.

Procedure

After providing informed consent, participants experienced a brief tutorial to familiarize themselves with the interface. Embedded within the tutorial was an alternative uses task (where participants were asked to think of as many alternative uses as possible for a bowling pin). Participants then generated ideas for both the fabric display and improve Turk problems. Participants were given 8 minutes to work on each problem, and the order of the problems was randomized across participants. Participants in the inspiration conditions received the following instructions regarding the inspirations: *"Below are some inspirations to boost your creativity. Feel free to create variations on them, elaborate on them, recombine them into new ideas, or simply use them to stimulate your thinking. If you find an inspiration to be helpful for your thinking, please let us know by clicking on the star button! This will help us provide better inspirations to future brainstormers."* After completing both problems, participants then completed a brief survey with

	High Quality	Low Quality
High novelty	(1.04, 1.51) Allow demographic info to be filled out automatically so it doesn't have to be done each time (1.52, 1.01) Virtual keyboard that allows you to have a row of buttons for only the keys that are used in the task, instead of finding them on the normal keyboard	(1.22, -1.09) Fingerprint Captchas are possible. (1.10, -1.74) music to focus worker
Low novelty	(-0.87, 1.04) Grey out HITS that have been done before and cannot be repeated (-1.00, 1.16) make it easy to find good hits	(-1.11, -1.38) Be able to complete more hits (-1.58, -1.83) Have direct contact information such as a phone number clearly listed for MTurk workers.

Table 2. Example ideas at each combination of low and high novelty and quality for the improve Turk problem.

questions about demographics and the participants' experiences during the task.

Measures

We measure key aspects of participants' ideation process (fluency, breadth of search and degree of iteration) as well as creative outcomes (novelty and quality of ideas).

Fluency: Number of Ideas

Fluency was operationalized as the number of ideas generated by a participant for a given problem.

Breadth of Search in Solution Space

We used LSI to characterize the nature of participants' search through the solution space. To maximize the accuracy of the model, for each problem we enriched the model with the full set of ideas previously collected from mTurk workers for each of the two problems (1,354 for Fabric Display, and 2,287 for Improve Turk). Thus, the total sizes of the training corpora for the LSI models were 2,354 ideas for Fabric Display (prior ideas plus 1,000 ideas from our experiment) and 3,403 for Improve Turk (prior ideas plus 1,116 ideas from our experiment).

We used the same procedure to build the LSI space as with the first step of machine clustering (i.e., weight words with TF-IDF before estimating LSI), except that we retained 200 dimensions (in keeping with prior rules of thumb for larger datasets [37]).

Breadth was operationalized as the mean pairwise distance between a given participant's ideas. Higher mean pairwise distance indicates that participants' ideas are sampled from very diverse regions of the solution space. Distances were calculated by subtracting pairwise cosines from 1, yielding distance scores between 0 (semantically identical) and 1 (semantically very different).

Novelty and Quality of Ideas

To explore impact on novelty and quality, we obtained ratings of novelty and quality for ideas for the improve Turk problem. We recruited 414 workers from mTurk to rate the ideas for novelty and quality. Novelty of an idea was operationalized as the degree to which it was novel, ranging from a scale of 1 (Extremely Obvious) to 7 (Extremely Novel). Quality of an idea was operationalized as the degree to which it would be useful for solving the problem, given that it was actually implemented (i.e., separating out concerns over feasibility), ranging from a scale of 1 (Not Useful at All) to 7 (Extremely Useful). Each worker rated a random sample of approximately 20 ideas.

mTurk workers are suitable judges, given that they have first-hand expertise in the domain of worker experience on mTurk, and are also potential users of mTurk mobile products. While the raw inter-rater agreement was relatively low (Krippendorff's alpha = .23 and .25, respectively), the overall aggregate measure had acceptable correspondence with each judges' intuitions. Computing correlations between each judges' ratings and the overall aggregate score yielded average correlations of .52 for novelty and .58 for quality. To deal with potential differences in usage of the rating scale across raters (e.g., some might only use the upper end of the scale), we normalized scores within raters (i.e., difference between rating and mean rating provide by rater, divided by the standard deviation of the raters' ratings). Table 2 shows examples of low and high novelty and quality ideas in the data.

Inspiration Use: Number of Bookmarked Inspirations

We also measured the degree to which participants found inspirations to be useful by counting the number of inspirations that were bookmarked by each participant.

Control Measures

Our primary control measure is participants' performance for the baseline fluency task (i.e., number of bowling pin alternative uses generated). This measure captures aspects of participants' base level of creative fluency (as a proxy for individual creativity [24]), as well as aspects of motivation and conscientiousness; all of these factors are expected to influence creative performance, and are therefore accounted for in our analyses by including baseline fluency as a covariate predictor in our statistical models.

Results

Participants generated a total of 2,116 ideas across the two problems, across conditions (1,000 ideas for Fabric Display and 1,116 for Improve Turk).

	Number of ideas	
	Fabric display	**Improve Turk**
Condition	**M *(SE)***	**M *(SE)***
Control	6.46 *(0.79)*	7.13 (*0.93)*
All-Ideas	8.40 *(0.77)* ^{m}	8.32 *(0.91)*
Machine-Machine	8.29 *(0.67)* ^{m}	9.59 *(0.78)* *
Machine-Human	7.56 *(0.63)*	7.43 *(0.75)*
Human-Human	7.30 *(0.68)*	8.13 *(0.81)*

** p < .05 vs. control, ^{m}p < .10 vs control*

Table 3. Machine-generated paths increase number of ideas relative to control condition. Means are adjusted for baseline fluency.

	Breadth of Search	
	Fabric display	**Improve Turk**
Condition	**M *(SE)***	**M *(SE)***
Control	0.89 *(0.02)*	0.81 *(0.03)*
All-Ideas	0.93 *(0.02)*	0.89 *(0.03)**
Machine-Machine	0.95 *(0.02)***	0.89 *(0.02)**
Machine-Human	0.95 *(0.02)**	0.86 *(0.02)*
Human-Human	0.94 *(0.02)* ^{m}	0.86 *(0.03)*

** p < .05 vs. control, ^{m}p < .10 vs control*

Table 4. External stimulation increases breadth of search by relative to control condition, for both problems.

Machine-Generated Paths Stimulate More Ideas

Table 3 shows the means and standard errors for each of the conditions. We first estimated separate ANCOVAs with condition as a main effect and baseline fluency as a control covariate for each of the two problems. For Fabric Display, there was no statistically significant overall effect of condition, F(4,133)=1.11, *p*=0.35. However, planned contrasts suggest that both the **all-ideas** (model beta=1.9, *p*=0.08) and **machine-machine** conditions (beta=1.8, *p*=0.08) trend towards more ideas than **control** (see Table 3, left column). For Improve Turk, the overall effect of condition was also not statistically significant, F(4,139)=1.40, *p*=0.23. However, the **machine-machine** condition did display a significant trend towards more ideas than control, beta=2.5, *p*=0.04 (see Table 3, right column).

Machine-Generated Paths Increase Breadth of Search

Table 4 shows the means and standard errors for each of the conditions. Baseline fluency was not significantly correlated with breadth of search in either problem (r=.06, *p*=.47 for Fabric Display, and r=.03, *p*=.76 for Improve Turk). Therefore, we estimated separate ANOVAs with condition as the only factor for each of the two problems. For the fabric display problem, there was a marginally significant overall effect of condition, F(4,133)=2.12, *p*=0.08. Planned contrasts suggest that both the **machine-machine** (beta=0.06, *p*=.01) and **machine-human** paths (beta=0.02, *p*=.02) lead to significantly more breadth of search than control, while **human-human** paths are marginally significantly better than control (beta=0.05, *p*=.05). For the Improve Turk problem, the overall main effect of condition is not statistically significant, F(4,139)=1.53, *p*=0.20). However, planned contrasts suggest that both all ideas (beta=0.07, *p*=.04) and **machine-machine** conditions (beta=0.06, *p*=.04) are better than **control**.

Solution Paths Do not Impact Novelty of Ideas

Table 5 shows the means and standard errors for each of the conditions. Baseline fluency was not significantly correlated with mean novelty. Therefore, we estimated an ANOVA with condition as main effect for mean novelty of ideas. There was no main effect of condition on mean novelty of ideas, F(4,133)=2.12,

Machine-Generated Paths Reduce Quality of Ideas

Table 5 shows the means and standard errors for each of the conditions. Baseline fluency was not significantly correlated with mean quality. Therefore, we estimated an ANOVA with condition as main effect for mean quality of ideas. There was a marginally significant main effect of condition on mean quality of ideas, F(4,129)=2.15, *p*=.08. Planned contrasts suggested that the mean quality of ideas in the **machine-machine** (M=–0.08, *SE*=0.06) and **machine-human conditions** (M=–0.02, *SE*=0.05) were of significantly *lower* quality than in the **control** condition (M=0.17, *SE*=.07), beta=-.25, *p*=.01, and beta=–.18, *p*=.04. Ideas in the **human-human** condition were of marginally *lower* quality (M=0.01, *SE*=0.06) than in the **control** condition (beta=-.16, *p*=.09).

Equal Number of Good Ideas Across Conditions

Given the reduction in quality associated with machine labels, we wondered about the impact of those paths on aggregate *creativity* (a combination of both novelty and

	Mean Novelty	**Mean Quality**
Condition	**M *(SE)***	**M *(SE)***
Control	0.04 *(0.07)*	0.17 *(0.07)*
All-Ideas	0.02 *(0.07)*	0.07 *(0.07)*
Machine-Machine	0.06 *(0.06)*	-0.09 *(0.06)***
Machine-Human	0.03 *(0.06)*	-0.02 *(0.06)**
Human-Human	-0.09 *(0.07)*	0.01 *(0.07)* ^{m}

** p < .05 vs. control, ^{m}p < .10 vs control*

Table 5. External stimulation does not impact novelty of ideas, but machine-generated paths reduces quality relative to control condition.

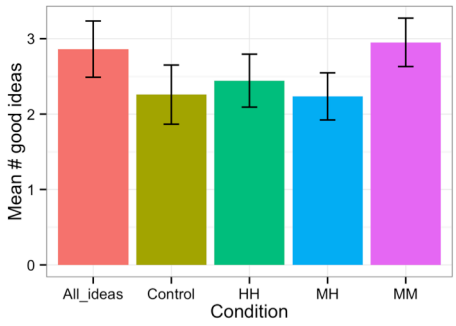

Table 3. Participants generate similar numbers of good ideas across conditions, adjusted for baseline fluency.

quality). Collective innovation platforms might accept a slight drop in mean quality as long as overall creativity does not suffer. Here, we follow Reinig et al [45] to measure creative output as the number of good ideas, where a good idea is defined as an idea with a novelty and quality scores that are both above the mean. Examples of such ideas are in the upper right quadrant of Table 2.

An ANCOVA controlling for baseline fluency showed no main effect of condition, $F(4,128)=0.98$, $p=.42$; further, the trends approximately track the statistical patterns for total number of ideas, suggesting that participants across conditions are generating good ideas at a fairly constant rate. Indeed, when examining the proportion of good ideas as a dependent measure, an ANCOVA again shows no significant differences across conditions, $F(4,128)=.61$, $p=.66$.

DISCUSSION

In this research, we examine the relative value of different approaches to inspiring crowd ideators with prior ideas. We empirically test whether abstracted solution paths (generated by a range of sensemaking methods) can enable ideators to better benefit from prior ideas (as measured by impact on fluency, breadth, and novelty and quality of ideas). We also empirically explore how the ideation value of solution paths varies with the cost of the sensemaking strategy that produced them. Our study yields two main sets of findings, which we discuss in turn.

No Consistent Benefit of Solution Paths

The first main set of findings is that solution paths do not consistently improve ideation more than simply showing all raw ideas to ideators. With respect to fluency, for the Improve Turk problem, machine-machine paths (but not machine-human, human-human, or all-ideas) improved fluency over control; however, for the Fabric Display problem, **all-ideas** and **machine-machine** paths both improved fluency. With respect to breadth of search, solution paths (regardless of source) improved breadth of search over control for the Fabric Display problem; however, for the Improve Turk problem, only **all-ideas** and **machine-machine** paths improved breadth over control. None of the solution paths conditions improved novelty over control, and some (**human-machine** and **machine-machine**) even *reduced* quality relative to control, while all-ideas did not.

There is a range of possible explanations for why solution paths did not improve ideation in our experiment. First, perhaps the solution paths reduced quality because they were potentially derived from both bad and good ideas (recall that we randomly sampled the initial set of 120 ideas from the prior datasets). However, it seems unlikely to have been the main driving factor. While one participant who saw all ideas did complain that there were many bad ideas, nobody voiced this complaint in the paths conditions. Further, one participant (in the **all-ideas** condition) specifically noted that, *"a lot of people had some good ideas about different features for a Mturk app."* Nevertheless, it is possible that the presence of even one or two bad ideas might be sufficient to contribute to fixation. Second, it could be that that there were too many inspirations to sift through. However, 15-25 inspirations seems like a manageable enough number of items to be able to relatively quickly scan and find interesting ideas to build on. We also specifically designed the system to be able to fit all of the inspirations on one screen in the paths conditions (so that there would not need to be much scrolling). Indeed, a number of participants in the paths conditions appreciated that they could just glance over to see inspirations, while many participants in the **all-ideas** condition complained that there were too many inspirations to consider at once.

A more interesting alternative explanation might be that, rather than providing bad solutions, the stimuli were steering participants away from good solutions, due to a desire to not repeat ideas (despite our instructions saying it was ok to build on or recombine inspirations). Indeed, some participants said in the survey that they were influenced by the inspirations in this way. For example, one participant (in the **machine-human** condition) said, *"I could just read them and see if an idea I had had already been provided and if it was I would move on to the next idea."* One participant (in the same condition) even said, *"I would come up with an idea and then notice it was already listed on the inspiration list, so I felt discouraged about that."* Another participant (in the **machine-machine** condition) said that it was challenging *"trying not to repeat the inspirations. The first inspiration was about a movie theater, which coincidentally, was also similar to my first idea."* If this happened often enough, participants might have searched more broadly, but away from any good ideas that happened to be represented in the solution paths. In fact, perhaps the relative ease of gaining a broad sense of the solution space with solution paths—rather than with the 120 raw ideas—could explain why quality was reduced only in the paths conditions. Relatedly, participants might have felt overwhelmed by the prospect of adding something new to the existing solution space. One participant in the **machine-human** condition said, *"There were so many inspirations*

for the second brainstorming task that it was hard to come up with anything new to add".

Overall, these findings suggest that it may not be straightforward to translate the benefits of abstracted solution paths from individual and group settings into a crowd setting. For example, if providing tens of abstracted solution paths at once as potential inspirations could be overwhelming, alternative delivery mechanisms, such as on-demand delivery of individual inspirations [9,47], or focusing ideators on a single theme [20,40].

Automated Sensemaking Can Improve Ideation

The second main set of findings concerns the value of machine-generated paths. In our data, the only form of stimulation that provided consistent benefits was machine-generated solution paths, which improved fluency and breadth across both problems. However, this increase in fluency and breadth came at the cost of reduced quality of ideas. We did also find, however, that the reduction in quality did not hamper the production of good ideas (i.e., ideas that are both novel and of high quality).

One explanation for why these solution paths (though hypothesized to be of low intrinsic quality), might have benefited ideation is that participants were essentially leveraging them as keyword clouds. For example, one participant in the Machine-Machine condition said, *"The inspirations seemed nonsensical when read as a whole, but one or two words would catch my attention and spark an idea. For example, I think one of the ones for the fabric said "bathroom" or "towel" or something like that. That gave me an image of a shower curtain and having video play on the shower curtain. The inspirations were jumping off points."* Another participant said, *"I would list as many ideas as possible and then browse quickly through key words and run with whatever came to mind. Often I would combine words and that would help me think of an original idea."*. In theory, the keywords generated using our machine-machine workflow should identify words that can best discriminate between major patterns of solutions in the set of ideas. It would be valuable to test whether these paths benefit ideators more than simpler ways of generating keyword clouds (e.g., frequency-based).

This set of findings suggests that the relationship between cost and quality (in terms of value for ideation) might not be simply linear, monotonic, or even positive. From a practical standpoint, large scale ideation platforms could gain value from employing very simple automated sensemaking methods (such as the workflow we used in this study). These methods might provide the most value in the early stages of the innovation process, where the focus is on quickly exploring as much of the space as possible before focusing in on more promising solution approaches, and some reduction in quality might be acceptable in exchange.

FUTURE WORK

Our finding that solution paths did not consistently benefit ideators (over providing all ideas or no ideas) runs counter to theory, and suggests caution when using solution paths in a crowd setting. It will be important for future research to test the potential explanations we explored for the lack of (and even slightly negative) effect of solution paths, particularly since many of these factors are likely to be present in collective innovation settings (e.g., large number of potential inspirations, presence of some bad ideas, pressure to come up with original ideas rather than build on ideas).

It would also likely be useful to investigate task designs and workflows that could overcome some of those potential challenges. For example, could rephrasing inspirations as questions (e.g., "How can we…?") mitigate participants' tendency to move away from solution paths? Could we also intelligently divide up solution paths to different portions of the crowd such that they can explore a smaller set of solution path in depth and would this help mitigate potential cognitive load issues from seeing many inspirations? Would it be useful to integrate some form of evaluation into the sensemaking process, so as to more quickly weed out bad ideas? Future efforts to enable more collaborative large scale ideation could benefit from examination of these questions.

Finally, in this study we used a fully manual approach to human sensemaking. But, methods exist for sensemaking using distributed human computation [1,2,13,23]. Future work might fruitfully explore how the value of human sensemaking for crowd ideation might not only depend on task design and workflow factors, but whether such distributed workflows could provide similar benefits.

CONCLUSION

In this paper, we examined the ideation value of different approaches to sensemaking over prior ideas, using an online ideation experiment that compares ideation under no stimulation, exposure to all ideas, or abstracted solution paths from fully automated, machine-human hybrid, or fully manual sensemaking approaches. Our results suggest that simple automated sensemaking methods can provide some value (e.g., increased fluency, breadth of search) to large-scale ideation platforms. Our results also motivate further research on how to best enable crowd ideators to benefit from (human) sensemaking outputs.

REFERENCES

1. Paul André, Aniket Kittur, and Steven P. Dow. 2014. Crowd Synthesis: Extracting Categories and Clusters from Complex Data. *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ACM, 989–998. http://doi.org/10.1145/2531602.2531653
2. Paul André, Haoqi Zhang, Juho Kim, Lydia Chilton, Steven P. Dow, and Robert C. Miller. 2013. Community clustering: Leveraging an academic crowd to form

coherent conference sessions. *First AAAI Conference on Human Computation and Crowdsourcing*.

3. Brian P. Bailey and Eric Horvitz. 2010. What's Your Idea?: A Case Study of a Grassroots Innovation Pipeline Within a Large Software Company. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2065–2074. http://doi.org/10.1145/1753326.1753641
4. Hugh Beyer and Karen Holtzblatt. 1997. *Contextual design: defining customer-centered systems*. Elsevier.
5. D. M. Blei. 2012. Probabilistic topic models. *Communications of the ACM* 55, 4: 77–84.
6. Kevin J. Boudreau and Karim R. Lakhani. 2013. Using the crowd as an innovation partner. *Harvard business review* 91, 4: 60–69.
7. Kevin J. Boudreau and Karim R. Lakhani. 2015. 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: 4–19. http://doi.org/10.1016/j.respol.2014.08.001
8. Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 288–296.
9. Joel Chan, Steven C. Dang, and Steven P. Dow. 2016. Improving Crowd Innovation with Expert Facilitation. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*.
10. Joel Chan and Christian Schunn. 2015. The impact of analogies on creative concept generation: Lessons from an in vivo study in engineering design. *Cognitive Science* 39, 1: 126–155.
11. W. G. Chase and H. A. Simon. 1973. The mind's eye in chess. In *Visual Information Processing*, W. G. Chase (ed.). New York, NY, 215–281.
12. Duen Horng Chau, Aniket Kittur, Jason I. Hong, and Christos Faloutsos. 2011. Apolo: Making Sense of Large Network Data by Combining Rich User Interaction and Machine Learning. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 167–176. http://doi.org/10.1145/1978942.1978967
13. Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing taxonomy creation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1999–2008.
14. N. Cowan. 2000. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences* 24: 87–185.
15. D. W. Dahl and P. Moreau. 2002. The Influence and value of analogical thinking during new product ideation. *Journal of Marketing Research* 39, 1: 47–60.
16. S. Deerwester, S. T. Dumais, G. W. Furnas, and T. K. Landauer. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41, 6: 1990.
17. Darleen M. DeRosa, Carter L. Smith, and Donald A. Hantula. 2007. The medium matters: Mining the long-promised merit of group interaction in creative idea generation tasks in a meta-analysis of the electronic group brainstorming literature. *Computers in Human Behavior* 23, 3: 1549–1581. http://doi.org/10.1016/j.chb.2005.07.003
18. J. R. Duflou and P. Verhaegen. 2011. Systematic innovation through patent based product aspect analysis. *CIRP Annals - Manufacturing Technology* 60, 1: 203–206.
19. C. Eckert and M. Stacey. 1998. Fortune Favours Only the Prepared Mind: Why Sources of Inspiration are Essential for Continuing Creativity. *Creativity and Innovation Management* 7, 1: 1–12.
20. Antonio Ferreira, Valeria Herskovic, and Pedro Antunes. 2008. Attention-Based Management of Information Flows in Synchronous Electronic Brainstorming. In *Groupware: Design, Implementation, and Use*, Robert O. Briggs, Gert-Jan -. J. Vreede and Aaron S. Read (eds.). Springer-Verlag, Berlin, Heidelberg, 1–16. Retrieved from http://dx.doi.org/10.1007/978-3-540-92831-7_1
21. Katherine Fu, Joel Chan, Jonathan Cagan, Kenneth Kotovsky, Christian Schunn, and Kristin Wood. 2013. The Meaning of Near and Far: The Impact of Structuring Design Databases and the Effect of Distance of Analogy on Design Output. *Journal of Mechanical Design* 135, 2: 021007. http://doi.org/10.1115/1.4023158
22. M. L. Gick and K. J. Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology* 15, 1: 1–38.
23. Ryan G. Gomes, Peter Welinder, Andreas Krause, and Pietro Perona. 2011. Crowdclustering. *Advances in Neural Information Processing Systems 24*, Curran Associates, Inc., 558–566. Retrieved June 9, 2015 from http://papers.nips.cc/paper/4187-crowdclustering.pdf
24. J. P. Guilford. 1950. Creativity. *American Psychologist* 5: 444–454.
25. Raja Gumienny, Steven Dow, Matthias Wenzel, Lutz Gericke, and Christoph Meinel. 2015. Tagging User Research Data: How to Support the Synthesis of Information in Design Teams. In *Design Thinking Research*, Hasso Plattner, Christoph Meinel and Larry Leifer (eds.). Springer International Publishing, 169–191. Retrieved July 6, 2015 from http://link.springer.com/chapter/10.1007/978-3-319-06823-7_10
26. S. R. Herring, C. C. Chang, J. Krantzler, and B. P. Bailey. 2009. Getting inspired!: understanding how and why examples are used in creative design practice. *Proceedings of the 27th international conference on Human factors in computing systems*, ACM, 87–96.

27. K. J. Holyoak and P. Thagard. 1996. *Mental leaps: Analogy in creative thought*. Cambridge, MA.
28. James Y Zou, Kamalika Chaudhuri, and Adam Tauman Kalai. Crowdsourcing feature discovery via adaptively chosen comparisons.
29. David G. Jansson and Steven M. Smith. 1991. Design fixation. *Design Studies* 12, 1: 3–11.
30. Elahe Javadi and Wai-Tat -. T. Fu. 2011. Idea Visibility, Information Diversity, and Idea Integration in Electronic Brainstorming. *Proceedings of the 6th International Conference on Foundations of Augmented Cognition: Directing the Future of Adaptive Systems*, Springer-Verlag, 517–524. Retrieved from http://dl.acm.org/citation.cfm?id=2021773.2021837
31. Elahe Javadi, Joseph Mahoney, and Judith Gebauer. 2013. 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: 1–21.
32. Mark Klein and Gregorio Convertino. 2015. A Roadmap for Open Innovation Systems. *Journal of Social Media for Organizations* 2, 1: 1.
33. Jon Kolko. 2011. *Exposing the magic of design: A practitioner's guide to the methods and theory of synthesis*. Oxford University Press.
34. Adriana Kovashka and Kristen Grauman. 2014. Discovering Shades of Attribute Meaning with the Crowd. *Third International Workshop on Parts and Attributes*.
35. Filip Krynicki. 2014. Methods and models for quantitative analysis of crowd brainstorming.
36. Chinmay Kulkarni, Steven P. Dow, and Scott R. Klemmer. 2012. Early and repeated exposure to examples improves creative work. *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*.
37. T. K. Landauer, P. W. Foltz, and D. Laham. 1998. An introduction to latent semantic analysis. *Discourse Processes* 25, 2: 259–284.
38. Brian Lee, Savil Srivastava, Ranjitha Kumar, Ronen Brafman, and Scott R. Klemmer. 2010. Designing with Interactive Example Galleries. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2257–2266. http://doi.org/10.1145/1753326.1753667
39. J. S. Linsey, I. Tseng, K. Fu, J. Cagan, K. L. Wood, and C. D. Schunn. 2010. A Study of Design Fixation, Its Mitigation and Perception in Engineering Design Faculty. *Journal of Mechanical Design* 132, 4: 041003.
40. Lan Luo and Olivier Toubia. 2015. Improving Online Idea Generation Platforms and Customizing the Task Structure Based on Consumers' Domain Specific Knowledge. *Journal of Marketing*. http://doi.org/10.1509/jm.13.0212
41. D. L. Medin. 1989. Concepts and conceptual structure. *American Psychologist* 44, 12: 1469–1481.
42. George A. Miller. 1956. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63, 2: 81–97. http://doi.org/10.1037/h0043158
43. Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* 12: 1532–1543.
44. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation.
45. Bruce A. Reinig, Robert O. Briggs, and Jay F. Nunamaker. 2007. On the Measurement of Ideation Quality. *Journal of Management Information Systems* 23, 4: 143–161. http://doi.org/10.2753/MIS0742-1222230407
46. Pao Siangliulue, Kenneth C. Arnold, Krzysztof Z. Gajos, and Steven P. Dow. 2015. Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas. *Proceedings of CSCW'15*.
47. Pao Siangliulue, Joel Chan, Kzryzstof Gajos, and Steven P. Dow. 2015. Providing timely examples improves the quantity and quality of generated ideas. *Proceedings of the ACM Conference on Creativity and Cognition*.
48. Ut Na Sio, Kenneth Kotovsky, and Jonathan Cagan. 2015. Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies* 39: 70–99. http://doi.org/10.1016/j.destud.2015.04.004
49. S. M. Smith, N. W. Kohn, and J. Shah. 2008. What you see is what you get: Effects of provocative stimuli in creative invention. *Proceedings of NSF International Workshop on Studying Design Creativity*.
50. S. M. Smith, T. B. Ward, and J. S. Schumacher. 1993. Constraining effects of examples in a creative generation task. *Memory & Cognition* 21, 6: 837–45.
51. John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. *Information Visualization* 7, 2: 118–132. http://doi.org/10.1057/palgrave.ivs.9500180
52. Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. 2011. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*.
53. Christian Terwiesch and Yi Xu. 2008. Innovation contests, open innovation, and multiagent problem solving. *Management science* 54, 9: 1529–1543.
54. Thomas B. Ward, Merryl J. Patterson, and Cynthia M. Sifonis. 2004. The Role of Specificity and Abstraction in Creative Idea Generation. *Creativity Research Journal* 16, 1: 1–9.
55. Thomas N. Wisdom and Robert L. Goldstone. 2011. Innovation, Imitation, and Problem Solving in a Networked Group. *Nonlinear Dynamics-Psychology and Life Sciences* 15, 2: 229.

56. Jinfeng Yi, Rong Jin, Shaili Jain, Tianbao Yang, and Anil K. Jain. 2012. Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning. *Advances in Neural Information Processing Systems*, 1772–1780.
57. Yisong Yue, Chong Wang, Khalid El-Arini, and Carlos Guestrin. 2014. Personalized Collaborative Clustering. *Proceedings of the 23rd International Conference on World Wide Web*, ACM, 75–84. http://doi.org/10.1145/2566486.2567991
58. Lixiu Yu, Aniket Kittur, and Robert E. Kraut. 2014. Distributed Analogical Idea Generation: Inventing with Crowds. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1245–1254. http://doi.org/10.1145/2556288.2557371