An Evaluation of Semantically Grouped Word Cloud Designs

Marti A. Hearst, Emily Pedersen, Lekha Patil, Elsie Lee, Paul Laskowski, and Steven Franconeri

Abstract—Word clouds continue to be a popular tool for summarizing textual information, despite their well-documented deficiencies for analytic tasks. Much of their popularity rests on their playful visual appeal. In this paper, we present the results of a series of controlled experiments that show that layouts in which words are arranged into semantically and visually distinct zones are more effective for understanding the underlying topics than standard word cloud layouts. White space separators and/or spatially grouped color coding led to significantly stronger understanding of the underlying topics compared to a standard Wordle layout, while simultaneously scoring higher on measures of aesthetic appeal. This work is an advance on prior research on semantic layouts for word clouds because that prior work has either not ensured that the different semantic groupings are visually or semantically distinct, or has not performed usability studies. An additional contribution of this work is the development of a dataset for a semantic category identification task that can be used for replication of these results or future evaluations of word cloud designs.

Index Terms—Information visualization, word clouds, text analysis, data analytics, evaluation

1 Introduction

Textual information is inherently difficult to visualize quantitatively, due to its nominal nature [16]. For at least a decade, a common method for showing textual information visually for lay people and scientists alike has been some variation of a word cloud. Standard word cloud designs such as the popular Wordle are styled for entertainment, surprise, and self-expression, but are not optimally designed for data analysis tasks [9]. Nonetheless, word cloud layouts are used widely today both in public forums and in analysis systems for tasks to which they are not fully suited.

In this work, we seek a balance between the aesthetic appeal of Wordle-style word clouds and the functionality and comprehensibility needed for data analysis tasks. We claim that analytic task performance would substantially improve if word cloud designs visually grouped semantically related words, either spatially or with color (or both). We present a series of controlled experiments that explore part of the design space of layouts that group words into zones according to semantics, color, and spatial positioning. Visually organizing words according to category structure led to markedly better understanding of the underlying topics compared to the standard word cloud layout, with the caveat that the categories themselves should be coherent.

- *M. Hearst, E. Pedersen, L. Patil, and P. Laskowski are with UC Berkeley, Berkeley, CA 94720. E-mail: {hearst, epedersen, lppatil}@berkeley.edu, , paul@ischool.berkeley.edu.*
- *E. Lee and S. Franconeri are with Northwestern University, Evanston, IL 60208. E-mail: {elsie.lee, franconeri}@northwestern.edu.*

Manuscript received 7 Sept. 2018; revised 26 Jan. 2019; accepted 8 Feb. 2019. Date of publication 12 Mar. 2019; date of current version 4 Aug. 2020. (Corresponding author: Marti A. Hearst.) Recommended for acceptance by S. C North. Digital Object Identifier no. 10.1109/TVCG.2019.2904683

Our studies focus on assessing the use of word clouds for analysis tasks like summarizing, gisting, and understanding the topics of underlying documents. To this end, we have developed a task (similar to the game of Taboo) and a corresponding dataset for which the correct answers are known. This dataset has high accuracy when categories are tested individually, and is straightforward to apply to the task of assessing word layout algorithms. This dataset should allow other researchers to replicate this work and extend it to new evaluations.

There are existing examples of the type of semantically organized layouts that we test here, but they have not been evaluated with usability tests. For instance, in a series of blog posts, Clark [5] clusters words from speeches and books based on how often they co-occur near one another, and groups those words spatially near each other and in the same color, placing the largest groups near one another in the word cloud, as shown in Fig. 1a. The design in Fig. 1b is from Topicrama [25], a text analysis system that includes a word cloud view. This display dynamically updates to show words from one to four different topics, depending on other interactions with an interface. Note that both of these designs exclusively use horizontal layout of words, and only minor font size variation.

We claim that layouts using semantic groupings like these designs should be more widely adopted, and speculate that they have not because:

- Published usability studies of word clouds have not convincingly shown the advantages of semantic groupings, and
- Automated tools that *easily and reliably* produce coherently semantically grouped word clouds in an aesthetically pleasing way are not widely available or known.

Fig. 1. Examples of the word clouds that are organized by color and topic. (a) Cloud derived from *Origin of Species* by Charles Darwin from the Neoformix blog by Clark [5]. (b) Topics from the TopicPanorama interface from Wang et al. [36].

Our goal is to develop and evaluate designs for word clouds that improve performance in topic recognition tasks, while at the same time retaining their aesthetic appeal. To this end, after reviewing the literature, we present a series of controlled experiments with human participants that explore the design elements that aid in the recognition of underlying meaning, while maintaining an aesthetically pleasing visual design.

In summary, our key findings are:

1) Visually grouped layouts are more effective in time-constrained category understanding tasks, compared to ungrouped layouts.
2) Visual grouping can be done by separating categories via white space or by color distinction, or both together.
3) Organized layouts with gaps defined by white space tend to be preferred over more tightly packed, less organized looking layouts for analytic tasks.
4) These results hold for semantically distinct categories; in this work, studies have not been attempted with those whose semantics overlap.

This work does *not* contribute algorithms for semantic grouping of text or automatically generating layouts. Instead, grouping is done manually, and layout is done either with simple programs or by making use of existing tools.

The remainder of the paper is organized as follows. Section 2 summarizes related work. Section 3 describes the development of the category stimuli and experimental task used for Experiments 1-3. Section 4 describes Experiment 1, which examines the role of white space and font size variation in word cloud layouts. Four different monochrome designs are compared, showing that Wordles produce far lower performance compared to Column layouts. In Section 5, Experiment 2 adds color to the designs, showing that semantically-coded color can improve even the performance of a Wordle, but not to the degree of a semantically organized layout. Section 6 describes Experiment 3, in which the white space requirement is relaxed, finding that color coding and spatial proximity can perform nearly as well as coding with white space gaps. The final experiment appears in Section 7 in which subjective responses are obtained for four different designs, finding that participants preferred word clouds with a more organized layout over more typical Wordle style designs. Finally, Section 8 discusses the ramifications of these findings and future work, and Section 9 draws conclusions.

2 Related Work

2.1 Usability Studies of Word Cloud Properties

When word clouds, initially known as tag clouds, first became popular on the web, they were used to show the frequency of user-created tags on blog posts and photo sharing sites. Although some authors claimed that tag clouds were useful for navigating and searching web collections and understanding their underlying statistics [17], two early formal usability studies found that the spatial layouts and irregular font sizes of tag clouds were detrimental to such understanding compared to a simple alphabetical list of words [12], [30]. The initial negative usability results with word clouds led some researchers to conclude that their main value was their fun, engaging nature, and as an indicator of social interaction [17], [33].

This general finding has since been replicated several times. For instance, both Lohmann et al. [26] and Schrammel et al. [32] showed that for finding a specific tag, sequential layout with alphabetical ordering and no font variation worked better than spatial layouts or layouts with font size variation. Studies by Heimerl et al. [19], Sinclair et al. [33], and Kuo et al. [23] found that participants preferred a search box for the task of finding a word over looking up words in word clouds.

More recently, Felix et al. [10] studied the outcome of varying numerous factors in the design of word clouds. They concluded that the best layout is determined by the underlying task: if the goal is to extract main concepts, independent of frequency, linearly presented lists were best. If searching for particular words, font size can guide visual search, even though it is unrelated to the task. If recognizing the frequency of values is important, then performance would likely be better for designs that use visual indicators such as bar charts, row layouts, or column layouts of words, rather than the font size variations of word clouds.

These results suggest that word clouds are not effective ways to help a viewer look up words or compare word frequencies. Instead, word cloud designs may be better suited for more complex analytic tasks such as summarizing and understanding topics.

2.2 Word Clouds as Tools for Creative Expression

The Wordle tool is an enormously popular visualization tool, used by millions of people for a wide range of tasks [9]. Feinberg, the designer of Wordle, recognized that standard

tag cloud layout was ineffective for navigation or other interactive tasks. Once freed from these constraints, he was able to create a design intended to be "typographically lively" and "useful primarily for pleasure" [9].

Feinberg and colleagues conducted a survey of 4,306 Wordle users [34], finding that they were not used primarily for analytic tasks. In fact, roughly 50 percent of survey participants turned out not to understand what font size means in Wordles, and a significant portion thought that font color had a meaning despite it being assigned arbitrarily. Rather than exploring new data, a large majority of users uploaded content that they wrote themselves or that is deeply familiar to them, and use the tool primarily for creative personal expression. A secondary major usage is by teachers to excite students about subject matter, often by presenting words as a puzzle to capture students' attention and interest. Feinberg also suspects that some of Wordle's success is attributable to its ease of use, or as it puts it, "its one-paste/one-click instant gratification."

Building on the idea of word clouds as tools for individual expression and creativity, several researchers have developed direct manipulation interfaces to help users modify Wordle-style layouts, including ManiWordle [22] and WordlePlus [20]. Of particular interest is the recent EdWordle system [37] which recognizes the importance of keeping groups of words close together as the user manipulates the visualization (while simultaneously reducing white space between words). One domain expert user of EdWordle produced, after several hours of work, a compelling arrangement of news consisting of five different themes, each in a separate color and a separate spatial group, which she referred to as a "storytelling cloud." However, most study participants were not as skilled with the tool.

Our view is that word clouds should support this kind of semantically coherent arrangement by default.

2.3 White Space in Word Clouds

An important feature of the solutions we examine in this paper is the separation of themes within word clouds by white space. By contrast, most research on word layout has a goal of compactness and reduction of white space.

The exception is for systems that allow comparisons between word clouds. Castella et al.'s [3] Word Storms system presents multiple whole word clouds side-by-side to allow comparison of documents' contents. Parallel Tag Clouds by Collins et al. [7] show vertically separated columns of tags, one per document, with lines connecting words across columns. These can be used for examining differences across documents and across time, with an emphasis on how the same words and themes change over time. The WordBridge [21] creates a node-link diagram and places a separate word cloud at each node.

2.4 Semantic Layout in Word Clouds

2.4.1 Semantic Layout Algorithms

Semantics in word clouds has been investigated extensively. Early work by Hasan-Montero and Herrero-Solana [14], cluster tags by their co-occurrence (how often they were simultaneously assigned to the same resource on a social tagging platform). The clusters are visualized as early-style tag clouds with one cluster's words shown per horizontal line. Fujimura et al. [11] use euclidean distance determined by a cosine similarity measure between tags from blogs and topographic map layout to show tags at high resolution.

Several approaches [1], [8], [28], [38], [39] create a vector representation of words or tags, weight the terms, typically with some variant of TF-IDF weighting, define a similarity metric between the vectors, usually a variant of cosine similarity, use multidimensional scaling to reduce the vector representations to a two-dimensional point in space to obtain a layout in which similar words are close together, and then use a force-directed model to adjust the final layout.

Cui et al. [8] take temporal information into account, creating different word clouds for documents as they change over time, and adding a highlight color behind words to indicate if they have disappeared from one time period to the next. Paulovich et al. [28] design an algorithm which can place words into distinct polygonal shapes while retaining semantic positioning. Barth et al. [1] introduce three alternatives to the force-directed layout whose goal is to optimize the adjacencies between edge weights in the graph of semantically related words. Xu et al. [39] use word embeddings [27] instead of standard word vector representations to create the clusters, and assign colors to clusters. Wu et al. [38] introduce the seam carving algorithm into the layout, finding more stable results, and use Bubble Sets [6] to emphasize the groupings with background colors (see Fig. 2d).

2.4.2 Usability Assessments of Semantic Layouts

The semantics-positioned methods described above have not been assessed with usability studies. However, some researchers have assessed semantic layouts. In most cases, these are manually arranged groupings. This subsection summarizes this work.

Schrammel et al. [32] compared four different layouts in a usability study with 24 participants. Tags were randomly drawn from a photo sharing service and four different tag clouds consisting of 7 horizontal lines, each containing 11 words were created from a fixed set of words (see Fig. 2a); this is a layout similar to that of [14], which was among the earliest work in semantically arranging tag clouds. The ordering of the words in the layouts were: (a) alphabetical, (b) random, (c) based on co-occurrence, and (d) based on WordNet similarity. This study overall did not find evidence in support of semantic organization of tag clouds, but did suggest that users may be interested in such a layout if it can be designed more successfully.

Lohman et al. [26] compare four different tag cloud layouts in a controlled study with 12 participants. The data were hand-crafted, consisting of neutral terms (in German) from common knowledge areas such as furniture and animals that were manually arranged into equal-sized rectangular bounding boxes. The layouts were (i) sequential horizontal, alphabetical order (ii) circular, with the most popular tags radiating from the center (iii) thematic clusters, loosely organized and distributed across the four quadrants of the rectangle (see Fig. 2b), and (iv) baseline (sequential, alphabetical sorting, no weighting of fonts). As mentioned above, for search tasks, alphabetic sorting without font variation was fastest.

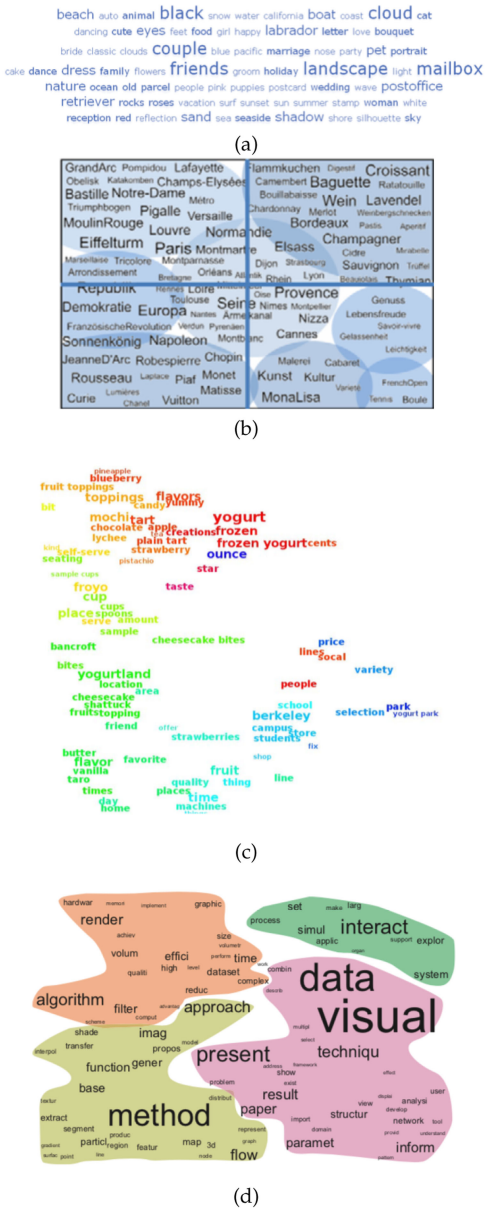


Fig. 2. Examples of semantic layouts from the research literature. (a) Co-occurring tags from Schrammel et al. [32]; (b) Manually created theme-based layout from Lohmann et al. [26]; note that the boundaries between themes are not visually demarcated (the blue backgrounds were not visible to participants); (c) Recloud layout of restaurant reviews from Wang et al. [35]; (d) Seam carving view of the top words from the Eurovis 2005 conference by Wu et al., with Bubble Sets [38].

However, for one of the tasks examined — finding tags on a given topic — the thematically clustered tag cloud performed best in terms of time taken, although not significantly better than the circular condition. The authors noted that "not all participants seem to recognize its ordering principle; some even selected tags from topic areas that were not asked for in the task," suggesting that the layout may not have indicated the semantic organization successfully. An examination of Fig. 2b shows that the themes are not distinguished visually either by color or white space. The work presented here focuses on providing a more successful thematic view.

Lohman et al. [26] also performed an eye tracking analysis which showed the tendency of participants to look at the upper left quadrant for three out of the four designs, but that this tendency was mostly mitigated for the thematically organized view. This may suggest that the semantic structure, in addition to default reading patterns, influenced where people looked more often in the thematically organized word clouds.

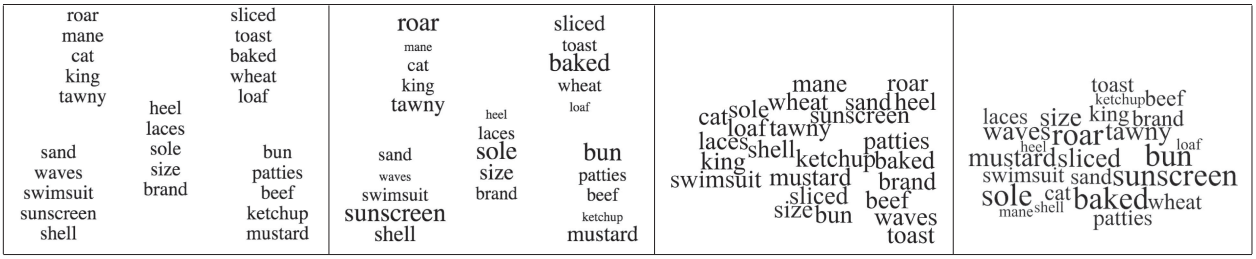
Chen et al. [4] created tag cloud visualizations called TagClusters containing semantic groups with labels (e.g., Rock, Indie, Punk) in capital letters, and surrounding category words in lower case (e.g., soft rock, progressive rock). Each group was shown with translucent overlapping colored backgrounds behind the words. In this case, grouping of words was accomplished by phrases sharing a common word or by being synonyms of one another. Placement of groups was determined by tag co-occurrence within a music sharing application. A usability study comparing the TagClusters to tag clouds found that the 12 participants were more effective with the TagClusters at uncovering group structure, finding relations between tags, and judging similarity across groups. Participants also preferred the groups to the tag clouds. Our work can be seen as validating and furthering these results.

Wang et al. [35] present the ReCloud system which creates a graph structure from dependency parses over restaurant review text and runs a force-directed layout algorithm over the resulting graph. The algorithm also clusters nodes, and assigns different colors to each cluster (see Fig. 2c). The final layout is modified according to an algorithm similar to Wordle's [9]. ReCloud is intended to be used for decision support, and was evaluated in the case of understanding restaurant reviews. The authors compared a color-coded, semantically grouped word cloud to a black-and-white, randomly organized baseline as well as the standard text reviews of the restaurants. They found the semantically grouped word cloud was more effective when comparing closely similar reviews, but found no difference for obviously dissimilar reviews and found no preference difference between the two layouts. They did find the semantically organized cloud had lower mental demand (as measured by the NASA TLX questionnaire [13]) in a time-constrained feature identification task.

Finally, research on text analysis and visualization comparing human-understandable categories to automatically generated clusters has shown that unsupervised clusters are more difficult to understand than categories [15], [16], [29]. The disadvantages of unsupervised clustering include their lack of predictability, their conflation of many dimensions simultaneously, the difficulty of labeling the groups, and the counter-intuitiveness of cluster subhierarchies.

3 Task and Category Stimuli

Word clouds are described in the literature as being useful for analytic tasks, including finding the gist of the underlying document, or summarizing the topics within a text collection. We wanted to use a highly reproducible task that reflects the goals of using word clouds for analytic tasks. However, asking participants to state summaries can result in high variation in acceptable responses and make studies difficult to reproduce.



(a) Column, Single Font (b) Column, Multi Font (c) Wordle, Single Font (d) Wordle, Multi Font

Fig. 3. Designs for Experiment 1; all are monochrome. The leftmost two are in Column layout; the rightmost two are in Wordle layout.

We devised a task that effectively reflects this goal while allowing for rigorous controlled experimental study. This task is similar to the game of Taboo: five words are shown that are clues for a category, and the participant's task is to guess the unnamed category. For example, presented with the words "menu waiter dishes tablecloth bill", the participant is expected to guess "restaurant" (the "taboo" words are the five clues, "restaurant" is the unnamed target word.) Because word clouds usually consist of sets of topics derived from underlying documents, this task reflects the implicit information extraction goal for the viewers of word clouds. It also goes beyond the standard task of simply stating what the largest or most important word in the image is, because the participant is not allowed to name any word that is visible in the image.

In the stimuli used in the study, 5 categories' clue words are presented together in each visual stimulus, for a total of 25 words and 5 categories to be guessed. We use the term *category* in this task to distinguish from finding the true underlying topics from a document or collection.

We developed and tested a set of 60 categories that are highly guessable when shown in isolation to fluent speakers of English. Similar to Lohmann et al. [26], we carefully developed this set of stimuli to ensure they are familiar to people from the target culture, and attempted to minimize bias or distress caused by emotionally powerful or politically charged terms. We iteratively tested the cue words until determining they were highly guessable. The process of selection of categories was iterative, and included consulting children's vocabulary lists and consideration of basic level categories [31]. Many potential categories are not unambiguously guessable from a few cue words or not generally familiar enough.

The stimuli was assessed with fluent English speaking Mechanical Turk crowd workers, who were each presented with the five clue words and asked to guess the Taboo category word. If a category word was not guessed correctly by at least 90 percent of workers, it was discarded from the set. The final set was tested on 36 workers each, achieving 94 percent accuracy on average (s.d. 1.8) on the target word.[1] These categories are themselves a contribution of this work and are available for other researchers to use; details appear in the supplemental materials, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TVCG.2019.2904683.

1. The singular and plural were both considered correct, in a few cases misspellings of at most 2 characters were corrected, and for 7 targets, one alternative is allowed, e.g.: bike/bicycle.

4 Exp 1: Grouping By Categories

Unless stated otherwise, the experiments described in this paper engage participants from the Amazon Mechanical Turk crowd sourcing platform. All participants are requested to be fluent English speakers. For Experiments 2 and 3, participants were required to pass a color vision check. Participants were paid $9/hour (.75 cents US for 5 minutes) and were not allowed to repeat any task, nor overlap among experiments. Material was presented via Qualtrics survey software.

4.1 Hypotheses and Image Stimuli

As mentioned in Section 2, word cloud layout designs assume the importance of reducing white space. In this study we break that assumption by laying out words according to semantic similarity, and allowing white space to appear between categories. We also allow variation in font sizes of words to increase the playful aspect of the design.

In order to reduce the potentially confounding factor of color, the layouts consisted of words in black font against a white background. To represent word clouds, a layout from the Wordle website[2] was generated, constrained to show horizontal texts.

Four views were compared against one another; two versions of Column views and two versions of Wordle views. In both cases, a single font version and a multiple font version was created (see Fig. 3). For Column layouts, each column corresponded to the words of one category. The text of the columns was center-justified. For both Column and Wordle layouts, font size variation was assigned via a random mapping among the five words in a category and weights of 1.4, 1.2, 1.0, 1.0, and 0.7, with one weight per word per category. These were used consistently across visual layouts. For the column layout, the weight was multiplied by the default font size (which was 40pt, with line spacing of 43.5). For the Wordle condition, an input file was generated which contained each word duplicated [weight*10] times. The Wordle layout was selected that produces all words in a horizontal layout. (Note that different runs produced different outputs in the online Wordle tool, even in the all-horizontal layout condition.) The Scheherazade font was used for all images.

We hypothesized that both Column views would outperform both Wordle views, and that single font would perform better than variable font for reaching correct answers under a time constrained condition.

2. www.wordle.net

TABLE 1
Design for Experiment 1; C Refers to Column Layout, W refers to Wordle Layout, SFont is Single Font, and MFont is Multiple Font Layout

PG	N	Sets 0-2	Sets 3-5	Sets 6-8	Sets 9-11
1	14	C SFont	C MFont	W SFont	W MFont
2	14	C MFont	W SFont	W MFont	C SFont
3	14	W SFont	W MFont	C SFont	C MFont
4	15	W MFont	C SFont	C MFont	W SFont

Sets refer to grouping of visualizations, e.g., Sets 0-2 correspond to visualizations 0-2. Each participant group (PG) consisting of N participants performed within-participant trials with the layouts shown. Stimuli were presented in randomized order within participant groups.

4.2 Experiment Design

We use time-constrained tasks in order to tease out distinctions between layouts, and to better reflect the claim that is often made that visualizations like word clouds give insight at a glance.

To create the image stimuli, the sixty categories were randomly grouped into 12 combinations of five categories each. Each combination is made into a distinct visualization consisting of 25 words (five cue words per category). The task was to identify as many of the category names as possible given the five cue words, within a time limit, while viewing the visualization. The study was a 36 person within-participants design as shown in Table 1; within each participant group, the images were shown in a randomized order.

After initial testing in a laboratory setting, the study was conducted with crowd workers on Amazon Mechanical Turk which has been shown to have results that are comparable to in-person laboratory studies for similar types of information visualization tasks [18].

After agreeing to a consent form and indicating that they were fluent English speakers, participants were trained on two practice questions. The first showed the five cue words for a single category in a row and the second showed the 25 cue words for five categories in a random spatial layout. In both cases, participants were told their task was to find the categories associated with the words shown and not to name any of the visible words. After the time was up, the participants were shown what the correct answers were.

After the practice tasks were completed, workers had 15 seconds to view each visualization and, while viewing the design, type as many categories as they could within the time limit. They were allowed to rest between tasks and continue when ready. After completing all 12 tasks they were shown examples of the four designs in a randomized order and asked to answer two subjective questions: "Which of these four image types do you prefer for this task?" and "Which of these four image types do you find most visually pleasing?" The first question is meant to assess the functionality of the design, and the second the aesthetics. If a participant did not achieve at least an average score of 0.8, their results were discarded.

4.3 Analysis

We evaluate our hypotheses using both linear and binomial mixed effect models. Our first specification (model 1) is a parsimonious ordinary least squares regression. It assumes that the score that a participant s gets for categories c is given by:

$$Score = \beta_0 + \beta_1 \text{Wordle} + \beta_2 \text{MFont} + \beta_3 \text{Wordle} \cdot \text{MFont} + D_c + u_{s,c}.$$

Here, Wordle is an indicator variable for Wordle format. MFont is an indicator for multiple font sizes. Both of these variables are demeaned. For example, Wordle is equal to 0.5 for Wordle layouts and -0.5 for Column layouts. This is a convenient specification, since it allows for direct interpretation of the coefficient of the variable as a main effect. In other words, the coefficient on Wordle does not change if we remove the interaction term and the same is true for the coefficient on MFont.

We include a fixed effect D_c for set of categories, to account for the possibility that some categories are inherently more difficult to guess than others. Finally, the term $u_{s,c}$ is the idiosyncratic error for the specific participant s in combination with specific categories c. We assume that each observation of $u_{s,c}$ is independent and identically distributed and the error is uncorrelated with all explanatory variables.

Our second specification (model 2) builds on model 1, but it is a linear mixed model with a random effects term representing the specific participant s. This accounts for the possibility that some participants are better at identifying categories than others. Guided by an exploratory analysis, we further added dummy variables to indicate whether a participant was attempting their very first or second visualization task. This allows the model to adjust for relatively low scores while participants are learning.

Finally, we also fit a mixed effect binomial model to the data (model 3). This technique directly models the probability that a participant gives the correct answer for each category. The final score is the number of successful answers out of five attempts. A benefit of this type of model is that the outcome is naturally constrained to fit within an interval. Such a model may be expected to provide a better fit to data, particularly as scores get close to 0 or close to 5.

4.4 Results

The overall mean score and standard error for each design in Experiment 1 is shown in the left panel of Fig. 4 (and for comparison, across Experiments 2 and 3 as well). Both Wordle designs perform substantially worse than the Column designs. Table 2 shows the estimated coefficients for each model. As expected, model 1 provides the worst fit to the data (log likelihood of -1098), model 2 provides a better fit (log likelihood of -1033), and model 3 provides the best fit (log likelihood of -955). We focus below on model 2, noting that our results are highly consistent across all specifications.

We find that the average score for a Wordle image is 1.37 points lower than for an average column layout, and this difference is statistically significant ($p < .001$) and robust across all specifications. We also calculate Cohen's d for this effect to be -0.67. Cohen's d is interpreted as the ratio of the score difference over the standard deviation within a single layout. The results further indicate a practically significant effect. Thus, as hypothesized, this experiment provides strong evidence that word clouds as designed in Wordles are difficult

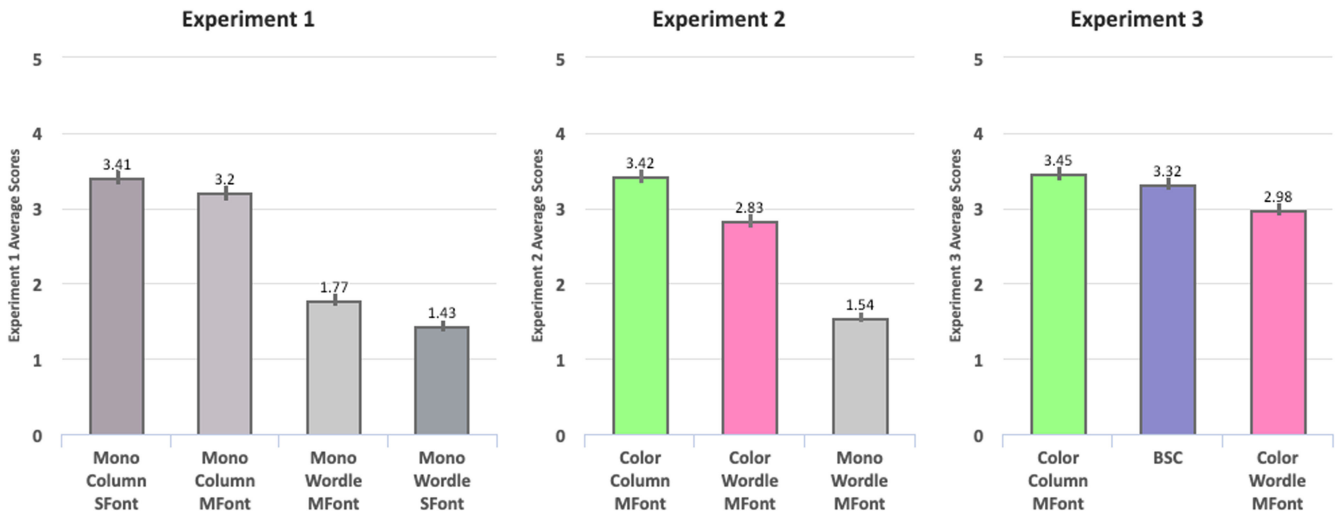


Fig. 4. Average scores and standard errors for the conditions of Experiments 1, 2, and 3. Bar colors indicate shared conditions across experiments.

for extracting semantic theme information, especially as compared to a simple design that can achieve the same goal.

The main effect of MFont is not significant in our models. On the other hand, we found a significant interaction term, suggesting that, for example, the effect of font size is different within Wordle layouts, and within Column layouts. Because of this, we proceed with post hoc estimates of the simple effects of these variables. All p-values include a Bonferroni correction.

Wordles perform significantly worse than Column layouts whether or not multiple fonts are present. However, the difference is largest for images with multiple font sizes: Wordles score 1.71 points lower than column formats within this group, and the difference is statistically significant ($p < .001$). Within single font size images, Wordles score 1.03 points lower than Column formats ($p < .001$).

Looking just at Column layouts, we found that multiple font sizes score 0.43 points higher than single font sizes ($p < .001$). On the other hand, we did not find evidence that multiple font sizes affect scores within Wordles. While not significant, our point estimate is actually negative, providing suggestive evidence that multiple font sizes may even detract from the usability of Wordles.

Participants were asked two subjective questions, about which layout type they preferred for the task, which they found most visually pleasing. For functionality, the Column layouts were preferred over the Wordle layouts by 93% (52/56) of participants who responded, and for aesthetics, Column layouts were preferred by 83% (45/54). (1/57 participants did not respond to the functionality question, and 3/57 did not respond to the aesthetics question.)

TABLE 2
Results of Experiment 1

	Dependent variable: Score		
	Linear		Binomial
	model 1	model 2	model 3
Wordle	-1.39^{***}	-1.37^{***}	-1.37^{***}
	(0.10)	(0.08)	(0.08)
MFont	0.10	0.09	0.15
	(0.10)	(0.08)	(0.08)
Wordle:MFont	-0.68^{***}	-0.68^{***}	-0.17^{***}
	(0.19)	(0.16)	(0.16)
Image Fixed Effects	Yes	Yes	Yes
Order Fixed Effects	No	Yes	Yes
Subject Random Effects	No	Yes	Yes
Constant	2.16^{***}	2.26^{***}	-0.22
	(0.17)	(0.17)	(0.17)
Table of Simple Effects			
MFont−SFont in Wordle	-0.24	-0.25	-0.20
	(0.14)	(0.11)	(0.11)
MFont−SFont in Columns	0.44^{**}	0.43^{***}	0.50^{***}
	(0.14)	(0.11)	(0.11)
Wordle−Column in SFont	-1.05^{***}	-1.03^{***}	-1.02^{***}
	(0.14)	(0.11)	(0.11)
Wordle−Column in MFont	-1.73^{***}	-1.71^{***}	-1.72^{***}
	(0.14)	(0.11)	(0.12)
Log Likelihood	−1,098	−1,033	−955

Note: *Standard errors in parentheses* $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$ *Simple effects are Bonferroni corrected.*

4.5 Eyetracking

An eyetracking study was conducted on a similar task with ten participants, 6 female and 4 male, who were already familiar with word clouds. Four were undergraduates participating for class credit; the rest were graduate students or post doctoral researchers recruited through word of mouth. All were native English speakers. Layout types shown were Column SFont, Column MFont, and Wordle Mfont; Wordle SFont was excluded.

Eye movements were recorded by an Eyelink 1000 Tower Mount eyetracker, sampling eye position monocularly at 1000 Hz. Participants first viewed instructions on a laptop and were able to ask any questions about the experiment before they began. After a calibration procedure, they viewed 12 images for 10 seconds each, four from each layout type, while they verbally identified categories. Fixation reports were computed by SR-Research's DataViewer software. Fixations (mean duration 274 ms) were categorized as being on a given word if its position was within the word's bounding box, with a 5 pixel buffer. Bounding boxes were computed with MATLABs Optical Character Recognition routines, followed by hand-correction. Because the OCR was not able to accurately find bounding boxes for the

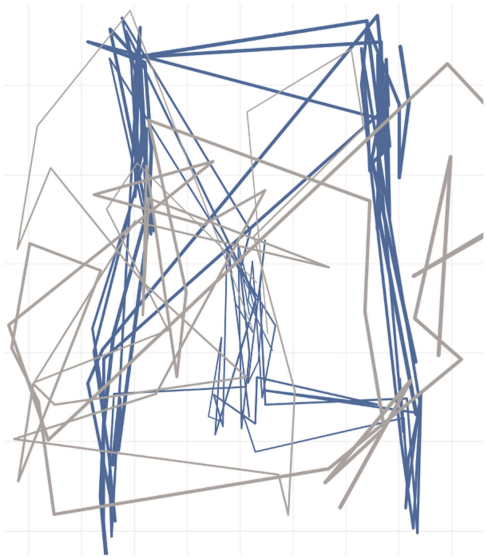


Fig. 5. Twelve trials for a single participant, showing eye traces of the Column SFont and Column MFont conditions (combined, in dark blue) and Wordle MFont condition (light grey). The blue traces show streaks of fixations within a single category, while the grey reveal haphazard inspection.

Wordle images, these were manually identified. Many bounding boxes overlapped such that a fixation's intended word was ambiguous. We analyzed fixation patterns in two ways: (a) assume that the fixated word is the word with the lowest arbitrary ID number in the word database, which should provide an unbiased sample for the analysis, but also (b) repeat the analysis for each of the possible words, and average together the results. Both methods show the same qualitative pattern, and we report method (b) here.

Behavioral performance was similar to the previous experiments: equally high in the Column SFont (mean score = 4.0, stddev = 0.7) and Column MFont (mean score = 3.9, stddev = 0.7) conditions, and both conditions showed higher performance (both $p < .001$) than the Wordle MFont condition (mean score = 2.1, stddev = 0.5).

Within the 10 seconds, participants made an average of 17.5 fixations across different words (excluding consecutive fixations within the same word). We predicted that in Column SFont and Column MFont conditions, participant fixations would be temporally clustered within the same category of words, compared to the Wordle MFont condition, where fixations would be relatively interleaved across categories. As illustrated by Fig. 5, the ratio of same-category to different-category fixations in the Column SFont condition was 2.3:1 (stddev = 0.59), which was statistically equivalent to the ratio of 2.1:1 (stddev = 0.69) in the Column MFont condition. For all ten participants, both conditions had a higher ratio than the Wordle MFont condition, which showed far fewer fixations within categories compared to across categories (0.24:1; stddev = 0.06, $t(9) < 10, p < .001$).

These results suggest the reason for the performance advantage for Column designs over the Wordle design is they spatially cluster words of the same category, which encourages the viewer to consider one category at a time. In contrast, the Wordle design, which intermixes words of different categories, fails to present each category's words in temporal proximity to each other.

5 Exp 2: Semantically Mapped Color

5.1 Hypotheses and Image Stimuli

In order to create a version of word clouds that will be accepted generally, color must be assigned in some manner. Wordles assign colors without considering a word's topic, using arbitrary color mappings only to delineate word boundaries and improve aesthetics [9]. We hypothesize that the best use of color is to associate it with meaning. Assigning all words in the same spatial grouping the same color further reinforces that grouping visually. Thus, for Experiment 2, all the cue words for a given category were assigned the same color. (In a real-world visualization, the designer might try to associate colors with the underlying meaning of the category, e.g., the color blue with the category ocean [24], but colors were assigned arbitrarily in this study.)

For the Wordle layout, we hypothesize that assigning colors according to semantic category will improve the results on the category guessing task, but not as much as the column layout should. We compared this to a monochrome baseline of the same Wordle layout, to be comparable to Experiment 1. We did not compare to a randomly assigned color Wordle, as we assumed that would perform similarly to (or worse than) the monochrome Wordle.

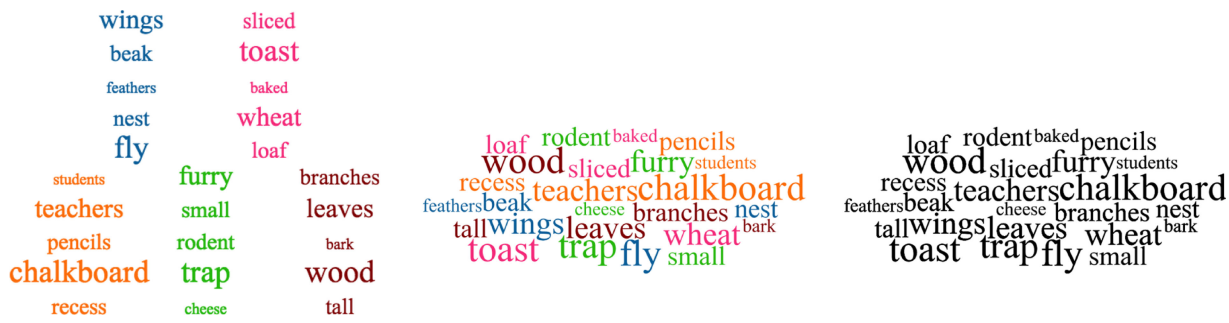
Experiment 2 uses the Column layout with multiple fonts. Although this layout does not score as well as the version with single font, we suspect it will score higher in terms of subjective impressions (i.e., fun or engaging) in less high-stakes settings. We did not compare to a monochrome Column baseline since the monochrome Wordle can serve as the Experiment 1 comparison, and it is unlikely that users would prefer it subjectively over a color Column layout.

5.2 Experiment Design

Three types of layout were contrasted in Experiment 2. Fig. 6a: Column layout, with one color per category (and column), Fig. 6b: Wordle layout with one color per category, and Fig. 6c: monochrome Wordle layout to act as a baseline which was identical to the color Wordle case except for this attribute. Each word was assigned the same color in the two color conditions. Colors were drawn from Bartram et al.'s "exciting" category [2] and chosen for high mutual contrast. As before, the Wordle layout was selected that produces all words in a horizontal layout and the Scheherazade font was used for all images.

As in Experiment 1, the sixty categories were randomly grouped into 12 combinations of five categories each. Each combination is made into a distinct visualization consisting of 25 words (five cue words per category). The task was to name as many of the category names as possible given the five cue words, within a time limit, while viewing the visualization. The study is a within-participants design as shown in Table 3.

The study was conducted with crowd workers on Amazon Mechanical Turk with fluent English speakers who asserted they had no color vision deficiencies, and who were able to pass a simple color deficiency test. After agreeing to a consent form and reading instructions, participants were told their task was to find the five categories associated with the 25 words shown. Participants were shown a series of practice tasks, which were more instructional in nature



(a) Color Column (b) Color Wordle (c) Mono Wordle

Fig. 6. Designs for Experiment 2: (a) Column layout with colors semantically assigned, (b) Wordles, with colors semantically assigned, (c) Wordles, with same layout and monochrome, acting as a baseline.

than for Experiment 1. First they were shown one set of cue words in a single color arranged spatially on a field and asked to name the category. Then a second example was shown which included two sets of cue words, in two sets of colors, and the participants were told to name the two categories present. The same was repeated with 5 sets of cue words and 5 colors. In essence, they were shown how to "decode" a semantically color-coded word cloud. Finally, they were shown same image in black letters on a white field and told to expect examples of this sort as well. Correct answers were shown in all cases.

After the practice session, workers had 15 seconds to view each of 12 visualizations and, while viewing the design, type as many of the 5 categories as they could within the time limit. They were allowed to rest between tasks and continue when ready. After completing all 12 tasks they were asked to indicate their subjective preference among the three designs they had been exposed to. If a participant did not achieve at least an average score of 0.8, their results were discarded. The ordering of tasks to participants is as shown in Table 3.

5.3 Analysis

The analysis of Experiment 2 is similar to that of Experiment 1. Model 1 assumes that the score that a participant s receives on categories c is given by:

$$Score = \beta_0 + \beta_1 \text{Color} + \beta_2 \text{Columns} + D_c + u_{s,c}.$$

Here, Color is an indicator variable that equals 1 for both color Wordles and color Column formats. Columns is an indicator that equals 1 only for color Column formats. As before, D_c represents the effect of a specific category, which may be inherently easier or harder than other categories. $u_{s,c}$ is the idiosyncratic error associated with participant s in combination with categories c. We assume that each error is independently and identically distributed, and uncorrelated with all explanatory variables.

The coefficients have the following interpretation: β_1 represents the difference in mean scores between monochrome Wordles and color Wordles. β_2 represents the additional difference in moving from color Wordles to the color Column layout. The difference between monochrome Wordles and color Columns may be written as $\beta_1 + \beta_2$.

As before, we provide additional specifications as a robustness check. Model 2 is a linear mixed model incorporating subject-level random effects and model 3 is a mixed effects binomial model. These models are analogous to the corresponding specifications from Experiment 1.

5.4 Results

The center panel of Fig. 4 shows the raw mean score for each design tested in Experiment 2. Consistent with our hypotheses, the color Column view with multiple fonts scores the highest (3.42), followed by the Wordle with semantically assigned color (2.83), and by the monochrome Wordle (1.54).

Table 4 provides the estimated coefficients for Experiment 2. The results were robust across all specifications and we focus our narrative on model 2. Adding semantically-coded color to a Wordle is associated with a 1.28 increase in mean score ($p < .001$). Cohen's d can be computed as 0.58,

TABLE 3
Design for Experiment 2; Sets Refer to Grouping of Visualizations, e.g., Sets 0-3 Correspond to Visualizations 0-3

PG	N	Sets 0-3	Sets 4-7	Sets 8-11
1	18	Color Columns	Color Wordles	Mono Wordles
2	17	Color Wordles	Mono Wordle	Color Columns
3	18	Mono Wordles	Color Columns	Color Wordles

Each participant group (PG) performed within-participant trials with the layouts shown. Stimuli order was shown in randomized order within participant groups.

TABLE 4
Results of Experiment 2

	Dependent variable: Score		
	Linear		Binomial
	model 1	model 2	model 3
Color	1.28^{***}	1.28^{***}	1.22^{***}
	(0.12)	(0.09)	(0.10)
Columns	0.60^{***}	0.60^{***}	0.58^{***}
	(0.12)	(0.09)	(0.10)
Image Fixed Effects	Yes	Yes	Yes
Subject Random Effects	No	Yes	Yes
Constant	1.40^{***}	1.40^{***}	-1.07^{***}
	(0.18)	(0.18)	(0.19)
Log Likelihood	−951	−855	−819

Note: $^{*}p<0.05$; $^{**}p<0.01$; $^{***}p<0.001$.

Fig. 7. Designs for Experiment 3 are identical to Experiment 2 except the monochrome Wordles of Fig. 6c are replaced with this BSC layout, which is organized both spatially and by color.

suggesting that this effect is large compared to the natural variation with each layout.

The effect of moving from a semantically-colored Wordle to a color Column format is less dramatic than the effect of adding color, but still substantial. We estimate an increase in score to be 0.60 ($p < .001$). Cohen's d can be computed as .27, which indicates a substantial effect size, though smaller than we found for adding color.

We can additionally compare monochrome Wordles directly against color Column layouts in a post hoc manner. For these groups, we see a score difference of 1.88 (stderr = 0.09, $p < .001$). Cohen's d is 0.85, further indicating a practically large effect.

We asked for a single combined preference among the three designs in this experiment. Across conditions, 88% (47/53) of participants preferred the color Column layout. 9% (5/53) preferred the color Wordles, and 1 preferred the monochrome Wordles.

6 Exp 3: Semantic Grouping

6.1 Hypotheses and Image Stimuli

Because we suspect that people continue to use word clouds because their spatial layout is visually appealing, we were interested in seeing if a more organized spatial layout could come close to the accuracy of a column-style layout. Therefore, we attempted to generate word cloud layouts that grouped semantically related words together, and simultaneously coded them with the same color, to see if participants would be able to perform as well with these as with the column view, and if they would prefer these layouts over the column view.

Experiment 3 examined the effects of removing white space as a grouping cue. It maintains the general layout of a word cloud with horizontal words, but imposed wider variation in font size than the designs studied in the earlier two experiments. Fig. 7 shows an example, which is generated from a website that implements variations of a wide range of layout algorithms from the literature.[3] The authors implement a variation of Wu et al.'s seam carving algorithm [38] that colors words within a semantic group. The algorithm successfully grouped category cue words together spatially, suggesting that good selection of words prior to grouping them within word clouds is key to the creation of semantic layouts. Because this implementation of seam carving differs from the original, we refer to this layout style as the Barth et al. Seam Carving algorithm, or BSC for short, below.

3. http://wordcloud.cs.arizona.edu/. The parameters chosen to generate the stimuli are: Layout: Seam Carving, Similarity: Cosine Coefficient, Ranking: TF/ICF - Brown Corpus, Font: Crimson Serif Color: ColorBrewer 2 (the generated svg was converted to conform to the color palette and Scheherazade font to match the other stimuli of Experiment 2), Aspect Radio: 4:3.

TABLE 5
Design for Experiment 3 is Identical to Experiment 2 Except Mono Color Wordles are Replaced with Semantically Organized Groupings

PG	N	Sets 0-3	Sets 4-7	Sets 8-11
1	17	Color Columns	Color Wordles	BSC
2	16	Color Wordles	BSC	Color Columns
3	15	BSC	Color Columns	Color Wordles

We hypothesized that the Column layout would perform better than the BSC layout, which in turn would perform better than the semantically color-coded Wordle layout.

The experiment design is identical to that of Experiment 2, with the BSC layout replacing monochrome Wordles, see Table 5. Crowd workers acted as participants in the same manner as Experiment 2, although the training was changed to remove the monochrome stimuli. As before, a result was dropped if the average score was less than 0.8.

6.2 Analysis

The analysis of Experiment 3 is similar to that of Experiment 2. Model 1 assumes that the score that a participant s receives on categories c is given by:

$$Score = \beta_0 + \beta_1 \text{Semantic} + \beta_2 \text{Columns} + D_c + u_{s,c}.$$

Here, Semantic is an indicator variable that equals 1 for both color Columns and BSC. Columns is an indicator that equals 1 only for the Column format. As before, D_c represents the effect of a specific category, which may be inherently easier or harder than other categories. $u_{s,c}$ is the idiosyncratic error associated with participant s in combination with categories c. We assume that each error is independently and identically distributed, and uncorrelated with all explanatory variables.

Our coefficients have the following interpretation: β_1 represents the difference in mean scores between BSC and the standard Wordles. β_2 represents the additional difference in moving from semantically grouped word clouds to the column format. The difference between semantic word clouds and color columns may be written as $\beta_1 + \beta_2$.

Following an analogous procedure to the previous two experiments, we include a linear mixed model (2) and a mixed effects binomial model (3) to understand the robustness of our estimates.

6.3 Results

The overall mean score for each design in Experiment 3 is shown in the right panel of Fig. 4. As anticipated, workers on average scored best with the Column layout (3.45), second best with the BSC semantically grouped layout (3.32), and third best with the semantically colored, but not spatially grouped Wordles (2.98). Note that the grouping of categories into stimuli images was the same in Experiments 2 and 3, but different in Experiment 1, which could effect comparison across experiments.

Our estimated effects were highly significant across specifications and we focus below on model 2. We estimate that moving from semantically color coded Wordles to

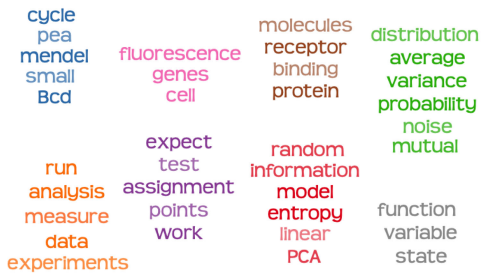
TABLE 6
Results of Experiment 3

	Dependent variable: Score		
	Linear		Binomial
	model 1	model 2	model 3
Semantic	0.35**	0.35***	0.35***
	(0.13)	(0.10)	(0.10)
Columns	0.12	0.12	0.12
	(0.13)	(0.10)	(0.10)
Image Fixed Effects	Yes	Yes	Yes
Subject Random Effects	No	Yes	Yes
Constant	2.74***	2.74***	0.21
	(0.19)	(0.19)	(0.20)
Log Likelihood	−912	−822	−780

Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

semantically grouped BSC layouts results in a 0.35 increase in score, which is statistically significant ($p < .001$). Cohen's d for Semantic is 0.15, which would typically be considered small. Post hoc, we additionally compare color coded Wordles directly against Column layouts, finding a score difference of 0.47 (stderr = 0.10, $p < .001$) Framing these results another way, the act of semantically assigning a color Wordle provides 75 percent of the benefit associated with rearranging words into columns. On the other hand, the difference between BSC and Column designs was estimated at just 0.12 and we could not reject the null hypothesis that the true difference was equal to zero. A 95 percent confidence interval for this effect ranges from −0.07 points to 0.31 points. Framing these results another way, the estimated difference between BSC and Column designs was only one third as large as the estimated difference between Wordle and BSC designs. Overall, our results are supportive of the idea that white space in column designs is not overly important for performance. Detailed results are shown in Table 6.

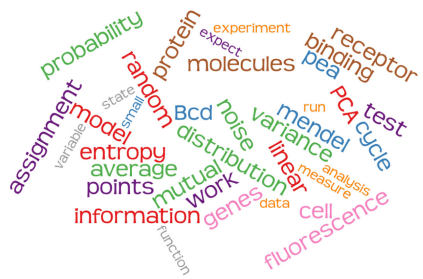
The subjective preference question was the same as for Experiment 1. 90 percent of workers preferred the Column layout for the task. The results were more split for the question about aesthetics, with 56% (27/48) choosing the Column layout, 24% (12/48) choosing the color-coded Wordle, and 18% (9/48) choosing the BSC layout. The close split between the two spatial layouts may reflect a tension between the improved grouping of the semantic layout but the drawbacks caused by the increased variation in font size.



(a) Column Layout

(b) Radial Layout

(c) BSC Layout



(d) Davies Layout

Fig. 8. Designs for Experiment 4 showing the contents of a course. Note there are eight categories instead of the five of Experiments 1-3.

7 Exp 4: Subjective Assessment

To assess the subjective preferences of the different layouts in a setting in which workers were not being tested on their ability to perform a task, we posed a subjective evaluation question as a distractor task in a study being performed by a different researcher in the lab of the co-authors.

This experiment was motivated by an email that one of the co-authors received, an advertisement for a Biology course, with key words jumbled within a word cloud. We asked workers to imagine that they were considering taking a course, and asked them to rate four different presentations of the course's content. We selected the most informative words from the original word cloud provided, omitting those that seemed unnecessary, and manually arranged them into groups. We created two views manually, a Column layout shown in Fig. 8a and a Radial layout shown in Fig. 8b. We ran the BSC algorithm of Experiment 3 on the selected words and found that the choices largely aligned with our divisions, shown in Fig. 8c. We also created a fourth view shown in Fig. 8d that mimicked to some degree the original emailed design using the relevant website, which for convenience here we call the Davies design after

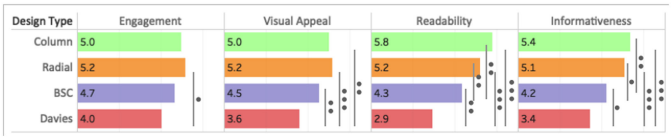


Fig. 9. Experiment 4 average ratings, (7 pt. scale, 1 = poor, 7 = excellent). Bars indicate pairwise significant differences; 1 dot corresponds to $p < 0.05$, 2 dots to $p < 0.01$, and 3 dots to $p < 0.001$.

the author of the code.[4] Colors were programmatically modified to ensure that the same palette was used across the designs. The number of font sizes varied across the designs; the teen font was used for all designs. The workers were asked to consider both the aesthetic and the functional aspects of the design; specifically, they were asked to:

> Imagine you were considering taking an online course and you visited the website for the course to see what it was about. Below you see four different views of the contents of the course. Rate these four designs from 1 - 7, where 1 is poor and 7 is excellent. Consider that the design is appearing on the course home page, and the goal is both to entice you to take the course and to inform you about its contents. Please scan all four views before rating any of them.

Participants were then asked to rate each of the designs along four dimensions: Readability, Informativeness, Visual Appeal, and Engagement. The first two ratings were designed to capture functional aspects, and the latter two to capture aesthetic aspects, although this was hypothetical as participants were not asked to actually perform a task.

Fig. 9 shows the average scores of participant preferences, by question type. Results of individual scales were pairwise compared with a Wilcox test and a Bonferroni correction, multiplying p-values by 6 to account for 6 comparisons with each variable (within Readability and within Informativeness, etc). Informativeness and Readability ratings were highly correlated with each other (R = 0.78), and the rest of the inter correlations were moderate (mean R = 0.59, min = 0.54, max = 0.66).

Participants rejected the Davies design across dimensions, with the exception of Engagement, for which it was virtually indistinguishable from the other designs. Interestingly, it scored significantly more poorly for Visual Appeal than all other designs. The Column and Radial view were indistinguishable across dimensions, suggesting that they may be mutually substitutable from a subjective viewpoint. The BSC view was rated significantly differently from all three other views on dimensions of Visual Appeal, Readability, and Informativeness, with an average score around 4.5 in most cases, suggesting that for communication purposes, people may prefer a layout separating categories with white space.

Participants were asked to briefly explain the reasons for their ratings. A typical comment favoring the Column and the Radial view was "The ones which are cleaner and better structured I gave better ratings, while the ones which are messy, cramped, and/or chaotic I gave worse ratings." A comment that gave a high score for engagement for the Davies layout (but scored it low otherwise) wrote "I based all my ratings on what I find visually appealing. I find well organized, easy to read text very appealing so I rated the images I was presented according to that criteria."

4. https://www.jasondavies.com/wordcloud/

8 Discussion

The study results show that word cloud-style visualizations that visually group semantic categories are more effective in our simulated analytic task, compared to the relatively disorganized layout of the typical design. Preliminary evidence also suggests that these organized layouts are preferred, at least for people engaged in analytic tasks. The results of this work leave several questions open for future investigation.

8.1 Semantically Coherent Groups

In the present experiments, analytic ground truth was defined by categorization across clearly distinguished sets of words. For each Taboo-style category, each cue word helps to indicate what the secret category is, without the added complexity of extraneous words. The categories are also semantically distinct from one another, except in a few cases in which the randomization placed two similar categories (e.g., butterfly and bee) in the same image. Category identification would be more difficult for semantically overlapping groups, though we would predict that visually grouped designs would still outperform non-grouped designs.

Future work should explore the combination of visually grouped word cloud designs with state-of-the-art text summarization and keyword-selection techniques. Word cloud tools could generate candidate groups of words that represent the text's underlying meaning in a coherent and non-overlapping way. These could also include two-word phrases when appropriate.

8.2 Variation within Layouts

The studies showed a clear advantage for views that group semantically related words together over those that intermix them for the task of comprehending underlying concepts. An eye-tracking analysis provided support for the hypothesis that viewers consider the words for one category at a time in the Column view, making it easier to interpret than the typical word cloud.

Some questions remain unanswered in this work about the role of font size, white space, word alignment, and word angle variation. We have only touched on a few points in the design space for white space variation, and there are likely interactions between font size, number of words in a group, and spatial placement that we have not investigated. When visually separating word categories, we tested separation types including spatial (e.g., the Column format), color (e.g., the color Wordle), and combinations of the two (e.g., color column), though did not contrast the effectiveness of those formats directly in a single experiment. This would be worthwhile to test in future work, because these types of visual grouping vary in strength (spatial grouping tends to be stronger than color) and mental processing (spatial groups seem to be carved out in parallel, while color groupings may be constructed

one at a time by the visual system, which could increase the type of serial within-group processing that facilitates performance in this task) [40], [41].

We also have not explored the role of larger words within each category. Those words might lead viewers to notice, or fail to notice, other words in the category. If the "Taboo" category name word were displayed in a large boldface font among smaller depictions of the corresponding cue words, the viewer might ignore the smaller words. What kind of translucency levels are helpful for differentiating words within a visual grouping? Are vertical or other non-horizontal word alignments acceptable in these layouts? How many words can be in a semantic category, and how many categories can be present within a representation in order for the coherence effects to be retained?

8.3 Subjective Responses

In the subjective responses obtained in the experiments so far, most participants preferred layouts that organized words into groups separated by both white space and color. In Experiments 1 and 2, alternatives to Wordles were strongly preferred. In Experiment 3, 90 percent of participants preferred the Column layout for the task, but for aesthetics, 56 percent preferred the Column layout over both the semantically-color coded Wordle and the BSC semantically organized word cloud. The final experiment asked participants to balance their preference between form and function on a hypothetical task — that of assessing their response to an advertisement for a course, both in terms of how well the design engaged them with the content, and how well it informed them about that content. This task had a larger number of categories than the earlier experiments (eight instead of five), and the number of words in each category varied from three to five. Preferences were in favor of designs in which the words were separated into visually distinguished zones, with the Column and Radial view rated highly by most participants, and the most word cloud-like style layout rated below neutral in most cases.

Note however that participants were engaged in an analytic task when asked to complete these ratings. If instead they had been engaged in a more whimsical task, their ratings may have differed. But given that the goal of this work is to find recommendations for layouts to be used for analytic tasks, these ratings are appropriate. Future work should compare ratings across a wider range of people engaged in more diverse tasks.

9 Conclusions

To depict the underlying categories, or topics of a document collection via a set of words, this work suggests that it is better to organize those words into zones of meaning, and display those zones in visually distinct groups, via spatial or color grouping. Despite being less playful than a typical word cloud, this display is rated as visually appealing by viewers performing analytic tasks. Additionally, the Taboo task can serve as an experimental test that simulates analytic tasks in text-based interfaces that involve understanding categories of words.

Acknowledgments

The authors thank the anonymous reviewers and the editor for their thoughtful and extensive comments, which greatly improved this work. Patil was supported in part by a University of California LEADS scholarship.

References

[1] L. Barth, S. G. Kobourov, and S. Pupyrev, "Experimental comparison of semantic word clouds," in *Proc. Int. Symp. Exp. Algorithms*, 2014, pp. 247–258.

[2] L. Bartram, A. Patra, and M. Stone, "Affective color in visualization," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2017, pp. 1364–1374.

[3] Q. Castella and C. Sutton, "Word storms: Multiples of word clouds for visual comparison of documents," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 665–676.

[4] Y.-X. Chen, R. Santamaría, A. Butz, and R. Therón, "Tagclusters: Semantic aggregation of collaborative tags beyond tagclouds," in *Proc. Int. Symp. Smart Graph.*, 2009, pp. 56–67.

[5] J. Clark, "Clustered word clouds," *Neoformix Blog*, 2009. [Online]. Available: http://www.neoformix.com/2008/ClusteredWordClouds.html

[6] C. Collins, G. Penn, and S. Carpendale, "Bubble sets: Revealing set relations with isocontours over existing visualizations," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1009–1016, Dec. 2009.

[7] C. Collins, F. B. Viegas, and M. Wattenberg, "Parallel tag clouds to explore and analyze faceted text corpora," in *Proc. IEEE Symp. Visual Analytics Sci. Technol.*, 2009, pp. 91–98.

[8] W. Cui, Y. Wu, S. Liu, F. Wei, M. X. Zhou, and H. Qu, "Context preserving dynamic word cloud visualization," in *Proc. Vis. Symp. (PacificVis)*, 2010, pp. 121–128.

[9] J. Feinberg, "Wordle," in *Beautiful Visualization: Looking at Data Through the Eyes of Experts*. Newton, MA, USA: O'Reilly Media, Inc., 2010, ch. 3.

[10] C. Felix, S. Franconeri, and E. Bertini, "Taking word clouds apart: An empirical investigation of the design space for keyword summaries," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 657–666, Jan. 2018.

[11] K. Fujimura, S. Fujimura, T. Matsubayashi, T. Yamada, and H. Okuda, "Topigraphy: Visualization for large-scale tag clouds," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 1087–1088.

[12] M. J. Halvey and M. T. Keane, "An assessment of tag presentation techniques," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 1313–1314.

[13] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in Psychology*, vol. 1988. Amsterdam, The Netherlands: Elsevier, 1988, pp. 139–183.

[14] Y. Hassan-Montero and V. Herrero-Solana, "Improving tag-clouds as visual information retrieval interfaces," in *Proc. Int. Conf. Multidisciplinary Inf. Sci. Technol.*, 2006, pp. 25–28.

[15] M. A. Hearst, "Clustering versus faceted categories for information exploration," *Commun. ACM*, vol. 49, no. 4, pp. 59–61, 2006.

[16] M. A. Hearst, *Search User Interfaces*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[17] M. A. Hearst and D. Rosner, "Tag clouds: Data analysis tool or social signaller?" in *Proc. 41st Annu. Hawaii Int. Conf. Syst. Sci.*, 2008, pp. 160–160.

[18] J. Heer and M. Bostock, "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2010, pp. 203–212.

[19] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," in *Proc. 47th Hawaii Int. Conf. Syst. Sci.*, 2014, pp. 1833–1842.

[20] J. Jo, B. Lee, and J. Seo, "Wordleplus: Expanding wordle's use through natural interaction and animation," *IEEE Comput. Graph. Appl.*, vol. 35, no. 6, pp. 20–28, Nov./Dec. 2015.

[21] K. Kim, S. Ko, N. Elmqvist, and D. S. Ebert, "Wordbridge: Using composite tag clouds in node-link diagrams for visualizing content and relations in text corpora," in *Proc. 44th Hawaii Int. Conf. Syst. Sci.*, 2011, pp. 1–8.

[22] K. Koh, B. Lee, B. Kim, and J. Seo, "Maniwordle: Providing flexible control over wordle," *IEEE Trans. Vis. Comput. Graph.*, vol. 16, no. 6, pp. 1190–1197, Nov./Dec. 2010.

[23] B. Y. Kuo, T. Hentrich, B. M. Good, and M. D. Wilkinson, "Tag clouds for summarizing web search results," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 1203–1204.
[24] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer, "Selecting semantically-resonant colors for data visualization," in *Comput. Graph. Forum*, vol. 32, no. 3pt4, 2013, pp. 401–410.
[25] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo, "Topicpanorama: A full picture of relevant topics," in *Proc. Conf. Visual Analytics Sci. Technol.*, 2014, pp. 183–192.
[26] S. Lohmann, J. Ziegler, and L. Tetzlaff, "Comparison of tag cloud layouts: Task-related performance and visual exploration," in *Proc. 12th IFIP TC 13 Int. Conf. Human-Comput. Interaction: Part I*, 2009, pp. 392–404.
[27] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. - Vol. 2*, 2013, pp. 3111–3119.
[28] F. V. Paulovich, F. Toledo, G. P. Telles, R. Minghim, and L. G. Nonato, "Semantic wordification of document collections," *Comput. Graph. Forum*, vol. 31, no. 3pt3, pp. 1145–1153, 2012.
[29] W. Pratt, M. A. Hearst, and L. M. Fagan, "A knowledge-based approach to organizing retrieved documents," in *Proc. 16th Nat. Conf. Artif. Intell.*, 1999, pp. 80–85.
[30] A. W. Rivadeneira, D. M. Gruen, M. J. Muller, and D. R. Millen, "Getting our head in the clouds: Toward evaluation studies of tagclouds," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2007, pp. 995–998.
[31] E. Rosch, "Principles of categorization," in *Concepts: Core Readings*, E. Margolis and S. Laurence, Eds. Cambridge, MA, USA: MIT Press, 1999.
[32] J. Schrammel, M. Leitner, and M. Tscheligi, "Semantically structured tag clouds: An empirical evaluation of clustered presentation approaches," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, 2009, pp. 2037–2040.
[33] J. Sinclair and M. Cardew-Hall, "The folksonomy tag cloud: When is it useful?" *J. Inf. Sci.*, vol. 34, no. 1, pp. 15–29, 2008.
[34] F. B. Viegas, M. Wattenberg, and J. Feinberg, "Participatory visualization with wordle," *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1137–1144, Nov. 2009.
[35] J. Wang, J. Zhao, S. Guo, C. North, and N. Ramakrishnan, "Recloud: Semantics-based word cloud visualization of user reviews," in *Proc. Graph. Interface*, 2014, pp. 151–158.
[36] X. Wang, S. Liu, J. Liu, J. Chen, J. Zhu, and B. Guo, "Topicpanorama: A full picture of relevant topics," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 12, pp. 2508–2521, Dec. 2016.
[37] Y. Wang, X. Chu, C. Bao, L. Zhu, O. Deussen, B. Chen, and M. Sedlmair, "Edwordle: Consistency-preserving word cloud editing," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 647–656, Jan. 2018.
[38] Y. Wu, T. Provan, F. Wei, S. Liu, and K.-L. Ma, "Semantic-preserving word clouds by seam carving," *Comput. Graph. Forum*, vol. 30, no. 3, pp. 741–750, 2011.
[39] J. Xu, Y. Tao, and H. Lin, "Semantic word cloud generation based on word embeddings," in *Proc. Pacific Vis. Symp*, 2016, pp. 239–243.
[40] D. Yu, D. Tam, and S. L. Franconeri, "Gestalt similarity groupings are not constructed in parallel," *Cognition*, vol. 182, pp. 8–13, 2019.
[41] D. Yu, X. Xiao, D. K. Bemis, and S. L. Franconeri, "Similarity grouping as feature-based selection," *Psychological Sci.*, vol. 30, no. 3, pp. 376–385, 2019.

Marti Hearst is a professor with the School of Information and Computer Science, UC Berkeley. Her research interests include information visualization, search user interfaces, computational linguistics, and improving education.

Emily Pedersen received the dual BA degrees in computer science and in cognitive science, both from UC Berkeley, in Spring 2018. She is working toward the MS degree in computer science at UC Berkeley, and will receive the degree in Spring 2019. Her research interests include data visualization and user experience.

Lekha Patil received the BA degree in applied mathematics from UC Berkeley with a concentration in data science. She will receive the degree in Spring 2019. She is a UC LEADS scholar and her research interests include data visualization and natural language processing.

Elsie Lee received the BA degree from Rutgers University, New Brunswick as a double major in psychology and information technology & informatics, in May 2017. Her research interests include human computer interaction, decision making, and information visualization.

Paul Laskowski is an adjunct assistant professor with the School of Information, UC Berkeley. His research interests include network architecture, digital privacy, and information economics. His work draws on a variety of techniques from economics, statistics, and the analysis of algorithms.

Steven Franconeri is a professor of psychology with Northwestern University, and director of the Northwestern Cognitive Science Program. He studies visual thinking and visual communication, across psychology, education, and information visualization.