Taking Word Clouds Apart: An Empirical Investigation of the Design Space for Keyword Summaries

Cristian Felix, Steven Franconeri, *Member, IEEE*, and Enrico Bertini, *Member, IEEE*

Abstract—In this paper we present a set of four user studies aimed at exploring the visual design space of what we call *keyword summaries*: lists of words with associated quantitative values used to help people derive an intuition of what information a given document collection (or part of it) may contain. We seek to systematically study how different visual representations may affect people's performance in extracting information out of keyword summaries. To this purpose, we first create a design space of possible visual representations and compare the possible solutions in this design space through a variety of representative tasks and performance metrics. Other researchers have, in the past, studied some aspects of effectiveness with word clouds, however, the existing literature is somewhat scattered and do not seem to address the problem in a sufficiently systematic and holistic manner. The results of our studies showed a strong dependency on the tasks users are performing. In this paper we present details of our methodology, the results, as well as, guidelines on how to design effective keyword summaries based in our discoveries.

Index Terms—Word Clouds, Tag Clouds, Text Visualization, Keyword Summaries

1 Introduction

In this paper we present a set of four user studies aimed at exploring the visual design space of what we call *keyword summaries*: lists of words with associated quantitative values used to help people derive an intuition of what information a given document collection (or part of it) may contain.

Such summaries are often generated through a set of natural language processing steps aimed at extracting the most relevant words and are very often represented as *word clouds*, that is, collections of words organized in space-optimized compact layouts in which font size encodes the frequency (or other relevance) value.

In this work, we seek to systematically study how different visual representations may affect people's performance in extracting information out of keyword summaries. To this purpose, we first create a design space of possible visual representations and then compare the possible solutions in this design space through a variety of representative tasks and performance metrics.

Even though word clouds are often used more as an emotional experience than an analytical tool [26], our focus in studying keyword summaries is on their use in exploratory data analysis, that is, when visual representations of a set of keywords and their frequency (or other value) is used to help an analyst generate questions, hypotheses and insights on the underlying data set.

This set of studies is motivated by our recent work on developing applications to analyze large sets of opinion data collected as sets of comments [9]. In this context, users generate keyword summaries of comments retrieved using a specific query and use the summary to get a sense of people's opinions. For instance, in summarizing reviews of restaurants that receive negative (1 star) reviews, one can identify trends and major issues that consumers mention in their reviews.

Similar problems are faced by a multitude of communicators, data analysts and developers when deciding what is the most appropriate form to utilize when visualizing sets of keywords that summarize a given set of documents (e.g., in social media, humanities, journalism, marketing).

- *Cristian Felix is with New York University. E-mail: cristian.felix@nyu.edu.*
- *Steven Franconeri is with Northwestern University. E-mail: franconeri@northwestern.edu.*
- *Enrico Bertini is with New York University. E-mail: enrico.bertini@nyu.edu.*

Manuscript received 31 Mar. 2017; accepted 1 Aug. 2017.
Date of publication 28 Aug. 2017; date of current version 1 Oct. 2017.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2017.2746018

This study is also motivated by the indiscriminate use of *word clouds* as the default solution to visualize keyword summaries. On the one hand, word clouds are extremely popular and do not seem to have generated major concerns in their users. On the other hand, several researchers and practitioners have voiced their dissatisfaction with them due to numerous shortcomings they seem to have [3, 12, 13, 18, 20, 21], namely: (1) the lack of natural reading order in how words are laid out; (2) the use of font size to communicate quantitative information, which is believed to be sub-optimal compared to other visual channels; and (3) the variation in word size due to word length rather than value.

Other researchers have in the past studied some aspects of effectiveness with word clouds. As we will describe in the next section, however, the existing literature is somewhat scattered and does not seem to address the problem in a sufficiently systematic and holistic manner. More specifically, the existing works, while useful, seem to focus on particular solutions or situations with a limited effort to place them in a framework of possible design choices.

In this work, we propose to study the visual design space of *keywords summaries* more systematically. More precisely, our systematic approach derives from two factors. First, we define a design space based on *spatial layout* and *value encoding*, two key visual parameters linked to performance, and include all meaningful combinations in the study. Second, we study these combinations across a set of representative tasks aimed at spanning a wide spectrum of task granularity: low-level tasks, to address low-level perceptual issues, and high-level tasks, to deal with tasks in which cognition plays a more prominent role.

To the best of our knowledge this approach is novel and will be a significant contribution to both theory and practice. We conjecture that organizing the work around a well-defined design space of possible solutions and a variety of tasks can not only lead to useful practical and theoretical insights, but also help researchers with a foundation to use for future studies in this area.

In the following section we describe existing studies and place them in a descriptive framework. Such framework is going to help us understand how the existing studies relate to one to another and where major gaps exist in the literature. We then move on to describing our study rationale, which includes the design space we devised for keyword summaries and the tasks we included in our studies. In Section 4 we describe the four studies in detail, providing details on their design, execution and results. Finally, in Section 6 we discuss the implications of the studies and in Section 7 we describe potential future work.

2 Background and Related Research

In this section we provide background information about how word clouds are used, visualization techniques employed and empirical re-

search performed to better understand how humans extract information out of them.

2.1 Word Clouds Origin and Use

At least two names have been commonly used to denote a collection of words that depict the content of a collection, *tag clouds* and *word clouds*. Historically, the term *tag cloud* derives from how blogs and web sites have used lists of textual tags with associated frequencies to provide a visual index to their content. The term *word clouds*, on the other hand, seems to originate from the idea of generating a document summary by extracting its most frequent words. The two terms however have been used interchangeably over the years referring to different goals, data extraction methods and different ways to depict information visually. In this study we will only use the term *word clouds*, referring to all possible uses found in the literature, and we will focus on their use as a method to automatically extract and visualize keywords from a document collection with the purpose of summarizing its content.

In this context, a word cloud visualization method receives as an input a list of words, each with an associated frequency value, and creates a visual depiction of this information. Several visual representations have been devised for word clouds throughout the years. By far the most common representation used is the one in which words are positioned in an unordered layout, to optimize the use of space, and frequency values are mapped to font size. Several variants however exist.

One common variant is to assign different colors to the words to create more aesthetically pleasing clouds or to encode an additional data attribute. Another very popular variant is *Wordle*, a word cloud in which words can be positioned horizontally or vertically (or even other orientations) and smaller keywords can also be nested inside the empty spaces of large letters in order to maximize space usage. This kind of representation has been used with the intent to maximize aesthetic impact and social interactions [14].

Sometimes, word clouds are also organized in more structured layouts. For instance the words can be organized row-wise or column-wise to increase their legibility and ease their scanning. Word cloud variants have also been devised to convey additional information available in the data. As an example, *parallel tag clouds* uses parallel lists of words connected by lines to allow comparison between subsets of documents (e.g., how they change over time) [7]. *SparkClouds* adds small line charts below each word to convey information on how the relevance of each word changes over time [17]. *BirdVis* uses geo-located word clouds to summarize comments generated by groups of bird watchers positioned in different geographical locations [10]. The *keyphrases* method proposed by Chuang et al., extends the techniques to n-grams to show more of the context of each word [4]. *Semantic-Preserving Word Clouds* position the words so that semantically related words are positioned together [27]. And *Phrase Nets* organizes the extracted words in a network in which words are connected by user-defined relationships [24].

In this paper, we focus on the case in which the information to be conveyed consists of a list of words with an associated numerical value.

2.2 Summary of Empirical Research

In preparation for this research we performed a literature review and found a total of 8 research papers reporting on studies aimed at studying word clouds empirically.

Rivadeneira et al. [20] produced one of the first works evaluating tag clouds. In their work they propose a set of 4 groups of tasks that can be used for evaluation: *search*, *browsing*, *impression formation (gist)* and *matching*. *Search* consists of finding a specific term or concept in the set of words displayed in the cloud. *Browsing* consists of inspecting the clouds to see if any term catches the user's interest. *Impression formation (gist)* consists of extracting the overall set of topics or concepts present in the cloud (as opposed to single terms). *Matching* consist of specifying which, among a list of predefined topics, is most related to the cloud presented to the participant.

In this study, using words extracted from psychology datasets and news reports, they conducted two experiments. In the first one, they used classic word clouds, with an unordered layout and font size encoding frequency, and asked the participants to report which words they could recall after a 20 seconds exposure. The results showed that participants recalled the words with larger font size (value) more often.

In the second experiment, they compared four different word cloud designs in which one was a simple list of terms and the rest used font size to encode frequency and the following positioning strategies: row layout ordered by frequency, row layout ordered alphabetically, spatial/unordered. The main tasks tested were *impression formation (gist)* and *matching*. For the gist task, the participants saw a word cloud for 30 seconds and were then asked to specify concepts/topics they contained. For the matching task, the participants had to match the word cloud's content to a set of predefined target concepts. The list design was found to be the most effective for the gist task and an effect of font size was found in the matching task, that is, the participants performed better when the target was presented with larger fonts.

Alexander & Gleicher [2] also studied the gist task in a recent new study. The study is based on topics generated from an automated topic extraction procedure (LDA) applied to news articles. It asked the participants to provide a topic name for the word cloud and to identify words not present in the word cloud as belonging or not to the topic expressed by the cloud. The study compared two main designs: spatial/unordered word clouds with font size encoding frequency and simple lists sorted in descending order of frequency and all fonts set to the same size. The results showed no effect of visual design on the results.

Halvey & Keane [12] asked people to find a country name in a list of countries using 3 different layouts, vertical list, horizontal list and tag cloud, each presented in two versions, ordered alphabetically and random. The results showed that ordering and font size play a major role in the search task, with alphabetical order outperforming unordered designs and words with a bigger font being easier to detect.

Lohmann et al. [18] asked participants to perform three type of tasks: find a tag by name, find tag by size and find tags belonging to a topic. They compared 4 different designs, sequential (row-wise), circular (with words organized in concentric circles and most important words in the center) and clustered (with semantically close words placed close together) using font size to encode frequency, and a sequential version without font encoding. The results showed that: alphabetically ordered tag clouds do not show frequency visually, perform better for word search tasks; the circular layout ordered by frequency performs better for the value search tasks; and thematically clustered layouts performed better for the search by topic task.

Schrammel et al [22] conducted a study to investigate the effects of different ordering of words, all using a row-wise sequential layout strategy and frequency mapped to font size. For all experiments they used a set of tags extracted from Flickr and 4 ordering strategy: *alphabetic*, *random*, *folksonomy-based*, and *linguistic-based*. In the folksonomy-based solution, words were placed according to their relatedness computed using Flickr's related tags feature. In the linguistic-based solution, words were placed according to their semantic relatedness computed using WordNet.

In the first experiment researchers asked participants to find a specific tag on the tag cloud and found that the alphabetic ordering was faster, followed by the folksonomy-based ordering. In the second experiment, participants were asked to find to find a tag belonging to an assigned topic and found no effect of ordering or font size. In the third experiment, they asked the participants to scan a tag cloud for 30 seconds and report on which keywords they could recall. The results showed an effect of font size, that is, words with bigger fonts were more likely to be recalled, and no effect of ordering.

Word clouds have also been tested using an eye-tracking methodology, mostly confirming the results reported in previous experiments. Lohmann et al. [18] confirmed the stronger eye fixation on the top-left quadrant of the cloud but also that eye fixation areas are affected by the layout used. Schrammel et al. [21] confirmed that people mostly scan rather than read the words one by one jumping from one location of the visual space to another. They also confirmed that bigger fonts take most of the user's attention and that the top-left quadrant is the one that

receives most attention.

Another interesting study is the one performed by Bateman et al. [3] in which they tested 8 different font features (size, weight, color, intensity, number of pixels, tag width, number of characters, tag area) to verify which one is more appropriate to represent importance. The study showed that font size, font weight, color hue and intensity are the best features for this purpose.

Focusing on font characteristics, the work of Alexander et al. [1] also provides interesting insights. In their work, they study whether word length affects magnitude perception. The results showed that there is a very small effect and that it can only be noticed when the difference in size between the words is of one single pixel.

Finally, the work of Hoeber et al. [16] is the only one we found in which the frequency value is encoded with an additional mark (a horizontal bar) rather than font size. They presented a search interface in which the results are summarized as a keyword summary in which frequency is presented as a horizontal bar chart. In the study participants were asked to search for documents related to a topic (e.g., "new hydroelectric projects") and rate the quality of the result with and without the help of the summary. The study however also showed that bar charts ordered by frequency demanded less time for the user to make decisions and it was also ranked as the favorite.

2.2.1 Main Findings

Reviewing the works presented above we can summarize some of the main findings that recur across multiple studies. Regarding position of words, using simple ordered lists, helps people find words more quickly than using solutions where the keywords are arranged randomly and unordered, and it also shows good performance for "gisting" tasks. Ordering keyword summaries lexicographically improves search by word tasks, since the user can go straight to the region where the target word is located. Ordering by value also improves performance on task based on value as the user can quickly read the top words. People scan keyword summaries, starting from top-left and going to bottom-right, this has an effect on how quickly people find target words, with words on the bottom-right corner taking longer to be found.

Visual encoding has an effect on drawing the attention of the user. Font size has a strong attractive power, with bigger fonts being remembered more easily, also people intuitively assume that bigger fonts refer to more important words. This attractive effect of font presents some trade-offs, for example, searching for smaller words in situations where no encoding is applied is sometimes faster, as when using font size, the bigger words distracts the user from the target word that is small.

In Figure 1 we present a summary of the tasks and treatments that have been studied in previous work. There are questions that are unanswered in previous work, leaving gaps in the existing literature base. For example, "How would using an additional mark instead of font feature improve performance in different tasks" or "How well people can decode values from font size".

2.2.2 Gaps and Main Focus of Our Study

From the related work we can also identify some relevant gaps. One problem we identify is a lack of systematic breakdown of the design space into relevant components. Most of the studies focus on a few conditions that vary across multiple parameters. Consequently, it is not always easy to understand what is the root cause of an observed effect. For instance, in studies that compare spatial/unordered word clouds to sorted lists it is not possible to understand if the observed effect is due to the ordering property of lists or the way words are arranged in the visual space. Similarly, some potentially interesting designs are never included in the studies we reviewed. In particular we identified two major design space gaps: *parallel lists* and *additional marks*. One interesting positioning strategy is to arrange the words in a set of *parallel lists*, rather than one single list extending exclusively in the vertical direction. This design solution is particularly interesting because it allows designers to use an even aspect ratio rather than one that grows only vertically. Also, as we have seen above, lists tend to perform well when tested with some tasks. Therefore, it seems natural to evaluate how designs that leverage the list arrangement perform. The use of *additional marks* is a way to convey magnitude information through marks and visual channels that are, at least theoretically, better than font size. It is surprising to notice how few studies include this specific kind of solution in their comparisons. Related to this last point, we also notice only one single study including a task based on extracting the magnitude value out of the visual representation. Yet, if magnitude is used in visual encoding, it seems natural to verify how accurately people can extract information out of it.

3 Study Rationale

The main goal of our study is to break down the design space of word clouds in meaningful components and then test their performance against a variety of relevant tasks. In the following section, we describe the design space we have devised and the sets of tasks we used to run our studies.

3.1 Proposed Design Space

We will assume the only information available to build such summaries is the list of keywords returned by an automated keyword extraction method (we provide details on which methods we use in our studies below) and the frequency of each keyword in the collection. In this respect, the study is orthogonal to studies that address the problem of improving the keyword extraction process and the use of additional metrics in place of, or in addition to, frequency. In our study we focus exclusively on how to visually encode this information and how different solutions may lead to different performance outcomes.

The main visual encoding problem therefore consists in deciding how to visually represent a list of words/labels with an associated quantitative value. We identify two main visual parameters that can be freely varied and combined: **layout** and **magnitude encoding**. The layout strategy consists in deciding how to position the words in the spatial substrate available for the visualization. The magnitude encoding strategy consists in deciding how to visually encode the quantitative value associated to each word.

We expect these two factors to have major effects on performance with different tasks and also to have, to some degree, some interaction effects. We expect the position of the words to affect the order in which an observer decides to scan the list of words. Similarly, we expect different magnitude encoding strategies to affect the precision with which values can be compared and extracted. Furthermore, we expect different magnitude encoding strategies to have an effect on guiding the viewer's attention and, as such, to also play a role, together with layout, in determining the order in which the viewer scans and reads the words. Figure 1 provides a summary of the elements we include in our design space. [1]

3.1.1 Layouts: Word Positioning

There are multiple ways keywords can be positioned in the visual space. In our design space we include three main options: *horizontal*, *vertical*, *spatial*. We choose these three main layouts because they are representative of the three main reading directions they promote: horizontal, vertical and spatial.

Horizontal: In this layout keywords are placed one after the other in a row until there is no more space available, then a new line is created. It starts from the top-left corner following the reading order from left to right, the standard used in Western countries [2]. This layout is one of the most popular: it was popularized by tools like TagCrowd [23] and it is also used in sites like Google Books [11]. This layout has also been included several of the studies we mentioned above [3,12,16,18,20,21].

Vertical: In this layout keywords are positioned one after the other in a vertical direction. When no additional vertical space is available a new column is created, thus leading to a multi-column design. One special case of this layout are single-column lists. In this case the list grows only vertically leading to the use of either an uneven aspect ratio of the visual space or the use of a scroll bar. Since we do not want these two aspects to work as confounding factors in our studies, they

[1] more examples and details at: https://nyuvis.github.io/word-cloud

[2] the order can be reversed for viewers who use a different convention

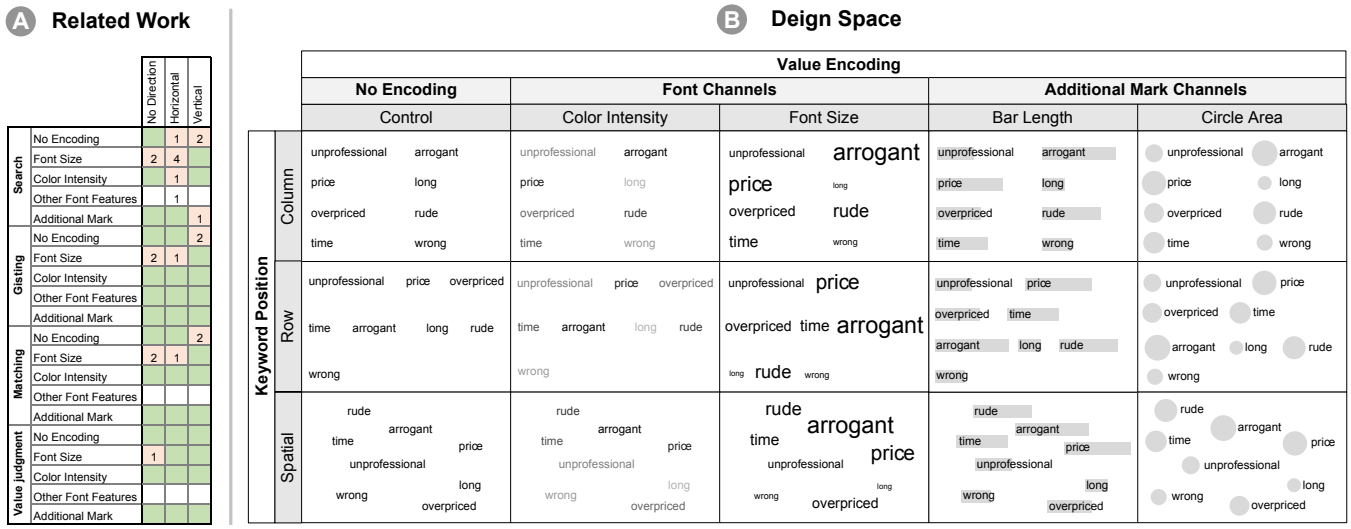

Fig. 1. Design space summary showing examples of visualizations generated by the intersection of different visual encodings and layouts for 8 keywords (note: the actual studies presented below used summaries with a larger number of words).

are not included in our experiments. As mentioned above, this type of solution is surprisingly rare and underutilized. One example of its use can be found in our previous work on TextTile, an interactive text data analysis system we built [9].

Spatial: In this layout there is no specific order used to position the keywords in the visual space. Many different algorithms have been devised to position the words. The large majority of them however have been designed with the major intent of optimizing space use rather than reading performance. One interesting exception is the use of "semantic layouts": methods that aim at placing words that are semantically related in adjacent positions [22, 27]. While interesting, we do not include this option in our studies because they rely heavily on the quality of methods that compute the "semantic relatedness" of the words and are also not easy to implement uniformly across the spatial layouts we want to test. The spatial layout strategy is by far the most popular and the one that includes many variants. In our studies we use the spiral placement layout proposed in Wordle [26] and modify it to avoid having word nested inside each other and rotated keywords.

3.1.2 Marks and Channels: Magnitude Encoding

A common way to describe the visual encoding process of a quantitative variable is to select a *mark*, a geometric primitive that represents the object, and a visual channel, a visual property that encodes a value associated to the object [19].

In popular word cloud designs such as Wordle, the mark used is actually the word itself and the channel used is the size of the font (typically encoded through a linear mapping between data value and font size). This specific choice however does not cover all sets of possible combinations.

There are two main additional options that have potentially interesting applications. The first one, is to use other visual properties of fonts to encode quantitative information (e.g., font color intensity). The second one, is to introduce additional marks (e.g., bars or circles) with the purpose of conveying quantitative information and let words function exclusively as labels attached to the marks.

The advantage of using font properties such as size and color intensity is that they do not require the potentially distracting effect and complexity of encoding related information in two separate visual objects. The advantage of using separate marks, is that it enables the use of more effective encoding strategies for the communication of quantitative information. Font properties, in fact, do not permit the use of some of the most effective visual channels such as position and length [19]. Furthermore, words are also influenced by word length which may interfere with other useful spatial visual properties. [1]

In order to study these two main strategies, we selected two representative solutions for each category: for solutions based on font properties we included font size and color intensity, whereas for solutions based on additional marks we included bars, encoding magnitude with bar position+length, and circles, encoding magnitude with area size.

Font Property: Size: This is the most commonly used channel in word clouds. It maps magnitude to the height of the font. Previous research suggests it is the font channel that best conveys the meaning of importance [3]. The visualization literature and theory, suggest that font size is also not an optimal channel to convey quantitative information [14, 25] because of the irregular shape and not direct relationship between height, width and painting area.

Font Property: Color Intensity: According to Beateman [3], after discarding color hue that only works for categorical values and font weight that has to little resolution to be effective, color intensity is the next best font channel to convey importance of a keyword and supports quantitative information.

Additional Mark: Bars Length: Bars are one of the most used marks in visualization, it allows the use of very strong channels like length and when aligned they can double encode the value using position, the strongest channel for quantitative visualization [19]. In our experiments we use horizontal bars, since they align with the word reading order. We also place them beneath the keyword to save space and to facilitate visual matching between the label and the mark.

Additional Mark: Circle Area: Circles are often used in scatter plots as bubbles to convey a third value or in other visualizations in which the spatial properties are already used to encode other information (e.g., bubble maps). Since the diameter of circles change together with area size, it is not possible to find a unique strategy to overlap circles with labels the same way we do with bars. For this reason, in our experiments labels are placed on the left side of each circle.

3.2 Benchmark Tasks

In selecting benchmark tasks for our studies we aim at two main goals. First, we aim at tasks that are representative of the goals pursued when keyword summaries are used in data analysis settings, that is, when the main goal of the user to identify interesting patterns that may lead to useful hypotheses and discoveries. Second, we aim at tasks able to capture performance at various levels of granularity: from low-level perceptual tasks to more high-level ones that require more complex cognitive efforts.

Our main intent is not only to verify which design elements work best in each of these tasks, but also to connect performance across tasks and see if results observed on lower level task translates into observed improvements in higher level tasks.

Our experiments are therefore organized around the following tasks: (1) **Magnitude Judgment**, to study performance in extracting quantitative information out of the encoded magnitudes; (2) **Keyword Search**, to capture performance in searching for a specific word; (3) **Topic Matching**, to capture performance in matching a word cloud to a set of predefined topic classes; and (4) **Topic Discovery**, to capture performance in extracting useful topics out of an assigned word cloud.

We chose tasks 2 to 4 based on the study presented to Rivadeneira et al. [20], where they can be mapped to search, matching and gisting respectively. Task 1 was chosen based on the classic study of Cleveland and McGill [6]. In the following section, we provide full details on our studies, including information on the specific setup used to study each of these tasks.

4 Studies

In this section we describe the series of studies we conducted in order to evaluate each combination of task, layout, and visual encoding we described above. Each subsection covers one task type and each task type was tested using cross combinations of layout and visual encoding. The studies were conducted on-line using Amazon Mechanical Turk. Previous research has shown the reliability of this platform to support visual perception studies [15], allowing us to achieve a high number and diversity of subjects. All participants recruited in our studies were from the United States and had a task acceptance rate of at least 99% and levels of education ranging from primary education to doctorate, with at least 60% of them having an undergraduate degree or higher. The participants' age across the studies always ranged between 18 and 71 years. The studies were conducted independently from each other and each participant was limited to participation in one study. All results are presented using effect sizes and confidence intervals (using bootstrap 95% confidence intervals) as suggested in [8]. All analyses reported in the studies were pre-planned before gathering the data of each experiment.

4.1 Study 1: Magnitude Judgment

This first study focuses on understanding how accurate are people in decoding quantitative values from the keyword summary. Being able to gather this information accurately is important because one of the main goals of a keyword summary is to gain an understanding of how frequency or relevance distribute across the words it displays. More precisely, our goal is to understand how different layouts and encoding strategies may affect magnitude estimation and comparison.

To this purpose, we model our experiment after the classic graphical perception experiment conducted by Cleveland and McGill [6] in which subjects are presented with different visual encodings and asked to judge proportions between pairs of values. In our study we use a similar set up, with the main difference of replacing the conditions used in the original experiment with the ones described in our design space.

4.1.1 Method

The first study used a mixed 3×4 design, with the 3 layouts (row, column, spatial) assigned between subjects and the 4 visual encodings (font size, color intensity, bar length and circle area) assigned within subjects. This choice was made to find a balance between the duration of a trial for each participants and the number of participants needed to test all possible combinations of the design space.

We recruited a total of 60 participants and split them randomly into 3 groups of 20 participants for each condition. Each participant performed a total of 48 trials split into 4 sections of 12 trials, one for each visual encoding tested. The order of the visual encodings, as well as the list of words used in each trial, were randomized to avoid learning effects. Each keyword summary depicted a total of 24 keywords in an area of 610×410 pixels.

The data were generated from 3 different data sets: a set of health care reviews, an email collection, and a collection of surveys on humanitarian issues. For each dataset we generated multiple thematic summaries by extracting the top 24 keywords most relevant to a specific category, e.g., keywords related to the *dentists* category in the healthcare reviews dataset. Relevance was computed using the Normalized Google Distance [5] between the category and each word.

The magnitude of the value associated to each keyword was generated using the following equation $w_i = 10 \times 10^{\frac{i-1}{24}}$, where w_i is the ith keyword in the summary. This formula produces a smooth exponential distributions, in which the proportion between any two words is constant and limits the proportions in our study to just 10 possible values, making the data easier to generate and analyze. Such distribution also mimics real-world data sets in which frequency is often distributed exponentially.

At the beginning of each experiment we presented a consent form. After the participant agreed with the terms, we collected demographic data and presented instructions to describe the task to be performed. After that, the participant went through a training phase to familiarize with the conditions and the task. After the training phase, a screen was shown with final instructions for the actual study and an option to opt out or perform the training phase again.

For each trial in the study, two words were randomly highlighted using a red triangle below each one. The participants were then instructed to select the smaller of the two and to provide an estimate of how much bigger was the larger compared to the smaller.

For each trial, we collected the proportion estimated by the participant and measured the absolute error using the formula $error = |percentage_{true} - percentage_{estimated}|$, where *error* is the absolute error, $percentage_{true}$ is the true proportion between the values and $percentage_{estimated}$ is the estimated proportion. In order to minimize the effect of outliers we a applied a logarithmic function $log\text{-}error = log_2(error + 1/8)$, where the $1/8$ term is used to avoid an indefinite result when the error is 0.

4.1.2 Results

In Figure 2 we present the results of the study. We consider, in order, first the effect of encoding, then layout, and finally the combination between these two factors. For each case we present interval estimates of the *log-error* calculated as described above. The estimates have been computed by first calculating the mid-mean of *log-error* for each participant and then computing the 95% bootstrapped confidence intervals sampling from the pool of participants.

Figure 2(A) shows the effect of visual encoding. Bar and circle marks perform better than font size and color intensity, although there are some overlapping results. A clear view of the difference between using font properties and additional mark is shown below in Figure 2(B) where contrast confidence intervals have been computed aggregating these two categories.

Figure 2(C) shows the performance of each layout: while the column layout interval has some overlap with row, it presents almost no overlap with the spatial layout condition. One possible explanation is that in spatial layout the marks are not aligned and as such they are harder to compare. The column and row layouts, in contrast, align marks vertically or horizontally and, as such, make comparison between some of the channels easier.

The interaction between marks and alignment is clearer in Figure 2(D), which shows the performance of each combination of layout and visual encoding. It is possible to see a consistent effect of layout across all visual encodings, with the combination *bar+column* being the most effective. Interestingly, the same advantage of aligned layouts shows up even with the color encoding, a channel that is not expected to benefit from alignment.

Overall, keyword summaries benefit from using additional marks when the task involves judgment of values. Bars show a small advantage over circles; a result that is corroborated by previous research stating that humans are better at judging length then area [19]. Bars benefit even more from aligned layouts, allowing comparison on aligned scales.

4.2 Study 2: Keyword Search

In this section we focus on how different encodings and layouts affect the time it takes to find a keyword in a summary. Searching for a keyword is a common low-level task when the user wants to confirm the presence of a specific word he or she has in mind. In previous work the same task has been investigated but always including, among the conditions, one or more layouts in which the words were ordered alphabetically.

Sorting a keyword summary alphabetically reduces the time to find a word [12] and as such it makes orderable layouts (e.g., column layout) more effective than un-orderable layouts (e.g., spatial layout). Since this advantage of orderable layouts is established, we designed our studies in a way to avoid this factor and decided to randomize the

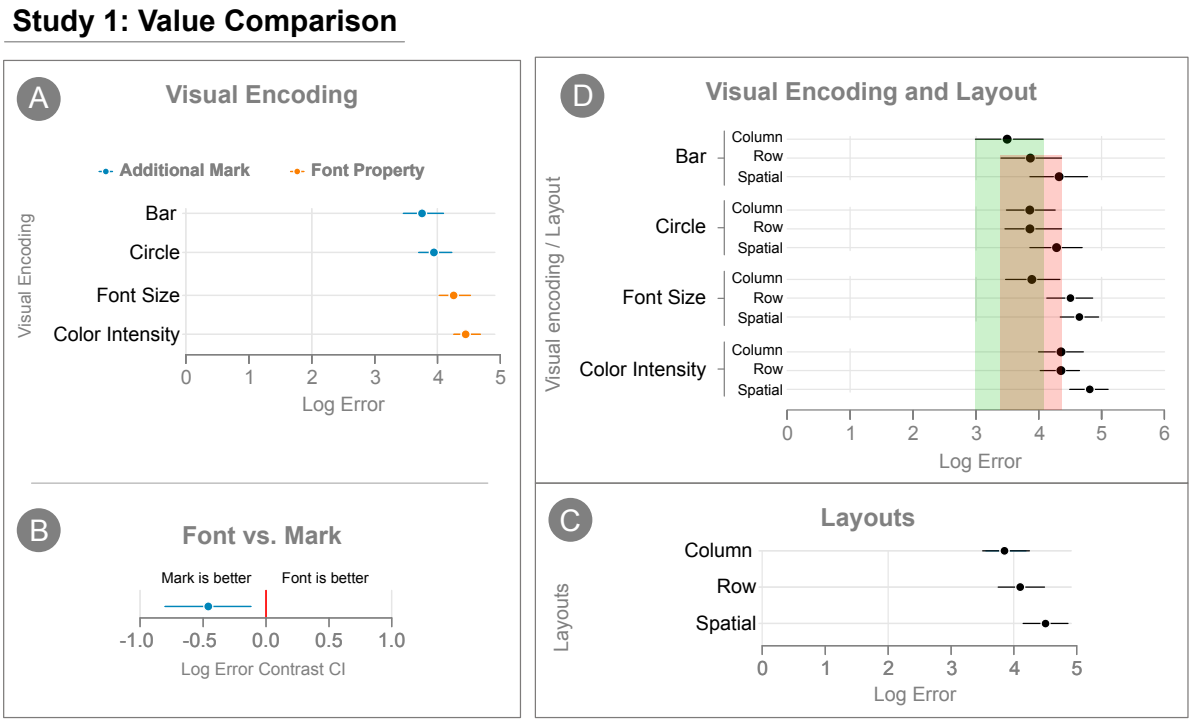

Fig. 2. **Magnitude Judgment (Study 1).** Comparison of mean log error for: (A) visual encoding strategy; (B) encoding with font properties vs. additional marks; (C) layout strategy; and (D) combinations of layout + visual encoding strategies.

order of the words in every condition. In turn, this has also the benefit of comparing layouts independently from their ordering properties, including the case in which a column or row layout is sorted according to a different strategy (e.g., sorted by magnitude).

4.2.1 Method

In this study we used the same mixed 3×4 design used in Study 1, with the 3 layouts assigned between subjects and the 4 visual encodings assigned within subjects. For this task, however, we controlled for two additional parameters: *target quadrant* and *target magnitude*, which respectively represent in which quadrant the target word is and which magnitude the target word is associated to. The target can be in one out of 5 possible quadrants (top-left, top-right, bottom-left, bottom-right and center) and can have 3 possible associated magnitudes (small, medium and large). We hypothesized that searching for a specific word would be influenced by these two additional parameters because: (a) people tend to read words in reading order, from top to bottom and from left to right (albeit limited to Western countries) and (b) because visually encoding the target with high magnitudes can speed up search by making the target stand out from the rest.

For each participant we generated a total of 60 trials, which came from all possible combinations of 5 target quadrants, 3 target magnitudes, and 4 visual encodings. All the trials were pre-computed and assigned in random order to each participant to avoid learning effects. Each keyword summary we generated was made of 500×400 pixels, each depicting a list of 50 keywords. We recruited 60 participants from Amazon Mechanical Turk, split them into 3 groups of 20 participants and each group was assigned to one of the layouts.

Following the same procedure of our first study, we presented the participants first with instructions on how to perform the tasks, then with a training phase to check their understanding and level of proficiency, and then, when the participants confirmed they were ready, they could start the actual test.

In each trial, we instructed the participants to find a given target keyword in a maximum amount of 15 seconds (we found this to be more than enough time through a pilot study). The target word was positioned, during the whole duration of the experiment, right above the keyword summary together with a timer to convey information about elapsed time. The participant could mark the target as found by first clicking on it and then submitting with a submit button. The trial was marked as unfinished when the participant did not click on any word by the end of the time allocated.

The main metric used to evaluate this task is the time necessary to find a keyword. More precisely we calculated the time interval between the time each summary was presented to the participant and the time the user clicked on the submit button.

4.2.2 Results

The results are analyzed using mean time (calculated for each participant) as the main effect size and bootstrap 95% confidence intervals calculated as in Study 1. In Figure 3(A) we show the main effect of visual encoding. Font size and color clearly outperform bar and circle with a difference between the estimated means of about 1.18 seconds. In Figure 3(B) we show the effect of target magnitude parameter in relation to the encoding parameter. As expected, larger magnitudes lead to faster target selection. We can also see that the difference between encodings is more prominent with large and medium magnitudes and tends to disappear with small ones.

In Figure 3(C) we show the main effect of layout. The spatial and column layouts have very similar performance and seem to have a somewhat better performance than the row layout. Given the amount of overlap between the confidence intervals the evidence for performance differences is very weak.

In Figure 3(D) we show all the conditions at once organized according to the effect of mark first and then layout. As one can see, the effect of layout is stable across all conditions, with spatial and column layout outperforming the row layout virtually in every condition. We refrain to comment further on individual comparisons as they may be the result of spurious trends due to the high number of conditions presented in the chart.

Finally, Figure 3(E) shows the effect of quadrant on performance. As we expected, our study replicates the same finding reported by Halvey & Keane [12] and Schramel et al. [21], that is, we can see a progressive decrease in performance going from top to bottom and left to right, with top-left being clearly better than bottom-right with a difference between the estimated means of about 1.6 seconds. The figure also shows that the effect of quadrant is also slightly modulated by layout, with the row layout being affected more prominently than the others.

In contrast to the magnitude judgment study, search tasks are negatively affected by the addition of marks. Adding marks seems to increase the time it takes to spot the target keyword, possibly because they interfere with the search task. The effect of layout, similarly to

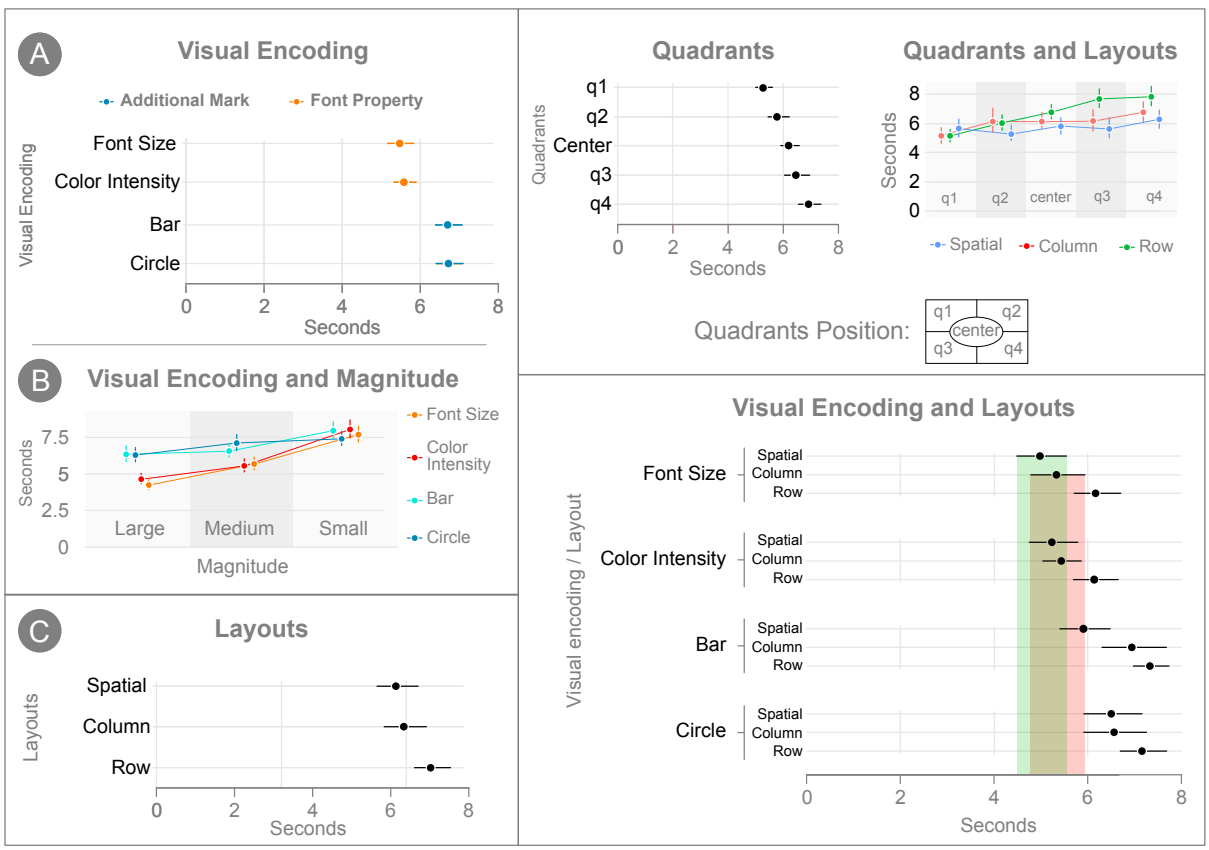

Fig. 3. **Keyword Search (Study 2).** Comparison of mean task completion time for: (A) visual encoding strategies; (B) magnitude of the target and visual encoding strategy used; (C) layout strategies; (D) visual encoding + layout strategies; (E) target quadrant and target quadrant + layout.

Study 1, is not particularly prominent but it seems to favor in this case the spatial and column layout.

4.3 Study 3: Topic Matching

In our third experiment we study topic matching, that is, the ability of viewers to identify topics the keyword summaries describe. This study enables us to test the conditions with a more realistic higher level task which requires participants to seek and integrate information across multiple keywords rather than from a few selected ones.

4.3.1 Method

In this study we used a 3×5 mixed-design with order inverted compared to the first two studies, that is, using layout as within subjects factor and visual encoding as between subjects factor. This choice was made due to the increased complexity of the topic matching task. To reduce the cognitive strain required to perform the whole set of trials, we decided to use layout, which has only 3 distinct values, as the within subjects factor. All trials used keyword summaries with 24 keywords, displayed in a 610×410 pixels canvas. In addition to the 4 visual encodings described in the design space, we also included a control condition where no visual encoding was used to visualize the magnitude value; the values were just written in front of each respective keyword as numbers.

We recruited 150 participants from Amazon Mechanical Turk, split them into 5 groups of 30 participants, one for each value encoding strategy. Each participant performed 30 trials, 10 for each layout assigned in random order. The topics for each trial were also selected in random order without replacement.

For this test we used exclusively the data set of medical reviews described in Study 1 and generated keyword summaries using 30 different medical specialties contained in the data set as categories. We built each keyword summary using the 24 most discriminative keywords of its own category using the same procedure described in Section 4.1.1.

For each keyword summary, we generated 4 possible labels: one corresponding to its associated category and 3 additional ones to use as alternative options. To generate the alternative options we created a similarity function that estimates the similarity between any two categories as the cosine distance between word vectors that describe each category. Such function was then used to generate for a given category three alternative labels at a small, medium and high distance. In order to avoid an excessively easy recognition of a category from its keywords, we removed from all trials keywords that were too similar to the labels used to describe the categories.

Each participant was presented with a consent form, instructions, and a training step to verify proficiency with the task (using a data set not included in the actual test). The participant was admitted to the actual study only after having performed the three preliminary training tasks correctly.

After the training phase, the study began and for each trial the participant was given 10 seconds to select which, among the displayed 4 options corresponded to the actual category. The 4 options were positioned below the keyword summary during the entire time of the trial and an option could be selected with a mouse click and submitted by clicking on a submit button. The time limit was introduced to emphasize the difference between the conditions we tested. In a preliminary pilot study we found that given unlimited time, the participants would be able to perform the matching task with high accuracy across all conditions. For each study we measure accuracy as the percentage of topics the participant selected correctly.

4.3.2 Results

The results are reported using mean accuracy as the main metric and bootstrap 95% confidence intervals as uncertainty estimates. Figure 4(A) shows the mean accuracy for each visual encoding. *Bar* and *color intensity* show slightly better performance, but the extensive overlap between different encodings prevents us to conclude anything strong

regarding the effect of encoding.

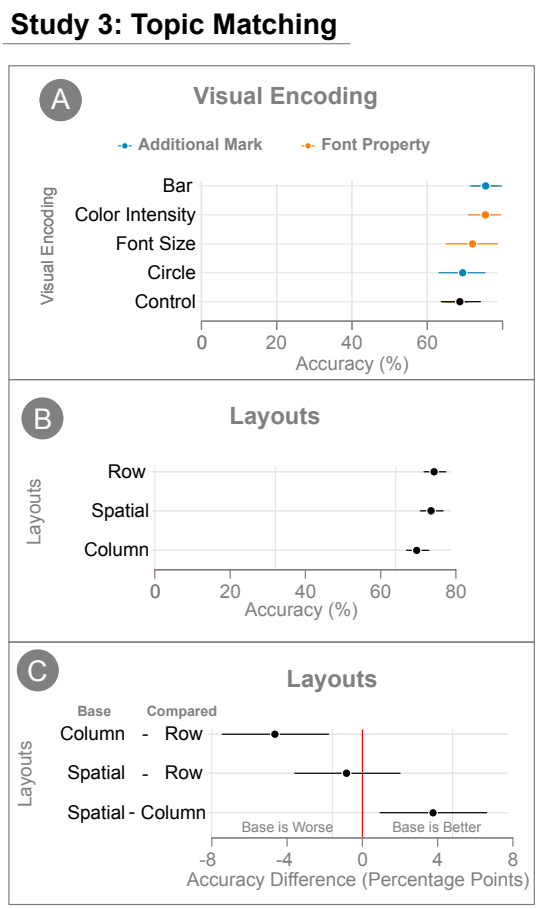

Fig. 4. **Topic Matching (Study 3).** Comparison of mean topic matching accuracy for: (A) visual encoding strategies; (B) layout strategies; (C) layout strategies as pairs of contrast confidence intervals.

Figure 4(B) and (C) show the accuracy results by layout. Here as well the results do not show any particularly strong trend. The lack of strong difference between conditions seems to indicate that as a task increases in complexity, the difference between the conditions becomes smaller or disappears.

4.4 Study 4: Topic Discovery

In this final study we examine how different design solutions affect the performance of a topic discovery task. In this experiment, we simulated the scenario in which the reader does not have a specific target to search for but he or she is rather exploring the keyword summary to extract and identify potential topics of interest.

4.4.1 Method

Similar to Study 3, we designed the study as a 3 layouts by 5 encodings (4 visual encodings + one conditions without any value encoding depicting just the keyword) mixed design. We recruited 150 participants from Amazon Mechanical Turk and assigned all the conditions using a between-subjects design method, with 10 participants assigned randomly to each combination of *layout* + *visual encoding*. We generated a keyword summary with 24 words placed in a canvas of 610×410 pixels according to the given layout and visual encoding.

To generate the topics, we decided to use the data set of medical reviews described above and focus on topics capturing issues patients have with their doctors. To this purpose, we sampled reviews with negative scores and manually extracted 4 main topics: *manners*, *mistakes*, *waiting time* and *financial issues*. For each of these topics we manually selected 4 topic-matching words, 4 words with more generic meaning (for example the word "*hospital*" was considered generic since all reviews are about hospitals), and a set of 4 unrelated words to use as noise. The magnitude of each word was selected using the same exponential distribution described in section 4.1 for Study 1, so that each of the 6 groups of words (4 topics, 1 generic, 1 noise) had its keywords associated with values of similar magnitude but in random order for each participant.

To verify that the keywords selected were a good match with the topics, we conducted a small survey with 10 Master and PhD students gathered from our lab. Each participant received the list of words and was asked to match topics with the most related words. We then compared the answers to verify the level of agreement and found only a few words with high degree of disagreement. Those words were discarded and replaced with more specific words, based on participants' feedback. As a final step, we also added 4 stop words to function as additional noise.

We presented the study as a fictitious scenario. In the scenario the participants were instructed to imagine being a data analyst in an insurance company analyzing reviews for a hospital and that their goal was to identify *as many issues as possible* using the keyword summary. The study started with instructions describing the scenario, followed by a training task, after which the user performed the main task. The main task consisted of 2 steps. First the keyword summary was presented to the user for 30 seconds, in the second step, we removed the keyword summary and asked the participant to provide short sentences describing the issues they identified on the keyword summary.

The answers provided by the participants were then coded by two of the authors, assigning matching topics to each sentence submitted. After performing the coding, we found an agreement value between the two coders of 91% (and a Kappa coefficient of inter-rater reliability of 0.81). Given the high level of agreement, inconsistencies between the two coders were simply resolved by randomly sampling between their two sets of results. We then computed two metrics: *accuracy*, representing the percentage of the topics the participant reported correctly out of all those they reported, and *coverage*, representing the percentage of topics identified out of all the available topics.

4.4.2 Results

The results are reported using mean accuracy and coverage as the main metrics and bootstrap 95% confidence intervals as uncertainty estimates. Figure 5(A) shows accuracy and coverage results for the 5 possible visual encoding strategies. As one can see, accuracy is high for all conditions, whereas coverage ranges between 55%-75%. All intervals overlap considerably, preventing us to conclude anything substantial on the tested conditions. Interestingly, the performance of the control condition, with no visual encoding of the magnitude value, is equal, if not better, than the other conditions. We obtain similar results in Figure 5(B), when comparing *layouts*: all intervals overlap considerably and the point estimates are all close one to another.

An unplanned analysis we performed is the analysis of the effect of target magnitude on topic identification. To this purpose we segmented the trials according to the magnitudes used to display the topic words (remember that trials with all possible sizes have been included for every topic tested) and calculated the mean coverage value and 95% boostrap confidence intervals. Figure 5(C) shows the results. As one can see, the topics encoded with larger magnitudes are identified more often than topics with smaller magnitude words, corroborating the results found in Study 2 (see Figure 3 (B)).

5 Summary of Results

Overall the results showed a strong dependency on tasks, that is, there is no condition that clearly outperforms all the others across all tested tasks. In the *magnitude judgment* task, having additional marks improved accuracy, and using a spatial layout led to lower performance. In the *keyword search* task, we find almost opposite results: with font properties reducing time to find a word and spatial layout being the best option, together with the column layout. In the *topic matching* task, we did not observe any major differences. The only condition that seems to perform worse than the others is the column layout. Finally, in the *topic discovery* task the difference between layouts disappears and

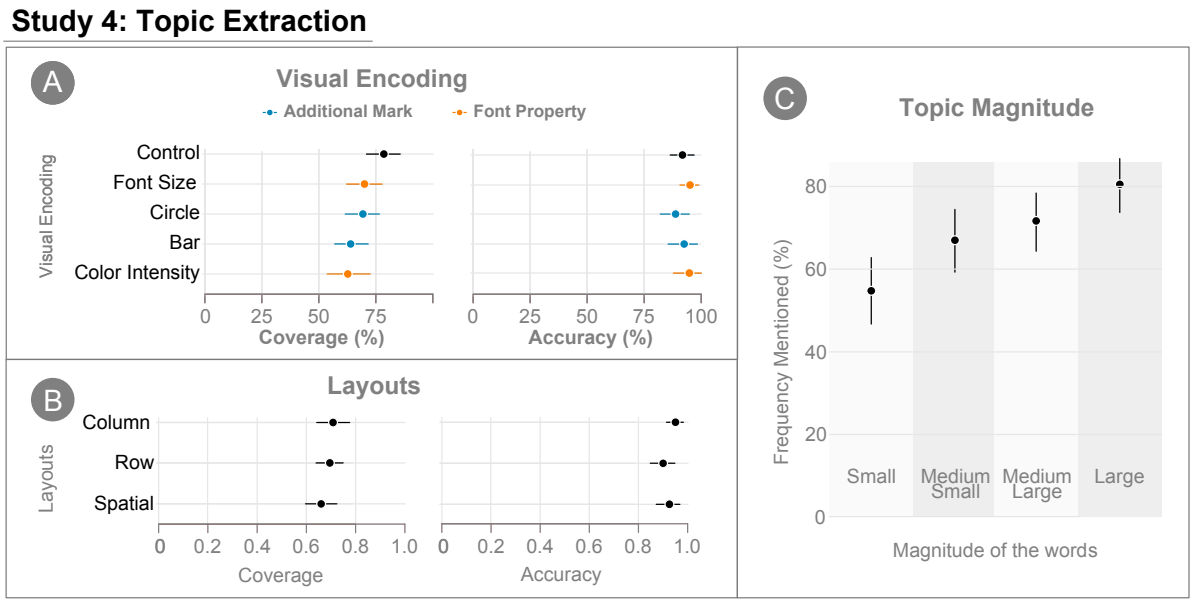

Fig. 5. **Topic Discovery (Study 4).** Comparison of mean topic detection accuracy and coverage for: (A) visual encoding strategies; (B) layout strategies; (C) magnitude of target topic.

color intensity is the only encoding clearly worse than the top ranked *control*. The study also, somewhat surprisingly, shows that the control condition does not perform any worse than the others.

The studies also led to a number of additional findings. In the *magnitude judgment* task, the ranking of visual variables is in accordance with results reported in previous studies [6, 15], but it also extends them by confirming the poor performance of font size and color intensity. In the *word search* task, we confirm the previous finding [12, 21] that the quadrant in which a target word is located has an effect on time. Finally, we also find that search time depends substantially on the magnitude value associated to the target word (Study 2 and Study 4).

6 Discussion and Guidelines

The results we summarized above lead to a number of interesting observations. One of the most relevant findings is the dependence of performance on task. As noted above, there is no condition that clearly outperforms the others across all tasks. Furthermore, as we move from lower level to higher level tasks, any difference observed between the conditions seems to vanish.

Regarding the shortcomings often mentioned regarding traditional word clouds, namely, the lack of natural reading order and the use of font size to encode quantitative information we find that these negative effects seem to be circumscribed to specific tasks. More precisely, we find that font encoding is an issue when comparing magnitude values (Study 1) but it can also speed up search time when the target happens to have a large magnitude associated to it (Study 2). As for the effect of spatial layout, we did not observe a strong effect. In the magnitude judgment task we do find a slight decrease of performance (Study 1) but no strong effect is found in the keyword search task (Study 2). Whether these effects have an impact on higher level tasks such as those that we tested is not clear, further research is needed.

"Which solution should be used then?". The answer seems to depend on the specific use one wants to make of a keyword summary. If the goal is to get a general sense of the main concepts contained in the summary, disregarding frequency or relevance, simple lists seem to be a powerful solution. This is corroborated by the observation that as soon as marks or fonts encode frequency values, the reading order of the observer can be influenced and thus generate potentially harmful biases. If searching for specific words is an important task, adding marks rather than using font encoding seems to be detrimental to the search. It is important to keep in mind that this advantage is highly dependent on whether a target word is associated with a large value or not. We conjecture that an increased performance in searching a high frequency word can also signal an excessive influence of visual encoding on attention. If high frequency words attract the attention of the reader too strongly, they may lead to sub-optimal scanning paths and to neglecting potentially interesting terms and concepts; an effect that may be relevant in time-critical situations and in data analysis more in general.

If extracting frequency or relevance values associated to the words is important, then using additional marks such as bars (using length/position) or circles (using area) seems to be a good choice, as well as using a column or row layout. If a compromise between search and value encoding is needed then the column layout with bars seems to be a good solution because: a) the column layout is the only one that performs well both with magnitude judgment and keyword search and b) the bar encoding does not seem to have a too strong effect on attention. More research is needed to verify this further.

Before concluding, we want to briefly mention other important aspects we have not tested in our work. First of all, our studies did not investigate how the elements of our design space affect aesthetics judgments. Since visualization is sometimes used in contexts where aesthetics is one of the possible relevant factors, knowing that effect would have some practical relevance. Furthermore, our design space does not exhaust all possible variations one may want to study in a keyword summary. Some other potentially relevant dimensions include: word orientation, number of words, use of color to encode a secondary value, font type and word length. The analysis of these conditions may be addressed in future studies.

7 Future Work

One of the most important aspects to address in future work is the effect of using different keyword extraction methods to generate summaries. While in this study we focused exclusively on aspects that pertain to visual encoding, it is important to keep in mind that the quality of a summary may heavily depend on which terms are selected in the first place. Another important aspect to study is the effect of the number of words shown in a summary. In our study we kept the number of words fixed but it would be useful to know how performance relates to keyword cardinality. Finally, many interesting effects we found in our studies may benefit from further investigation based on eye-tracking analysis. Especially, understanding how a viewer directs attention to the words as layouts and marks change. We are particularly interested in further investigating whether some visual encoding may excessively attract attention to some regions, leading the viewer to neglect potentially useful words.

8 Acknowledgments

We thank Pierre Dragicevic and Marti Hearst for their invaluable feedback and suggestions. We also thank Yelp for providing the healthcare reviews dataset. This work is supported in part by CAPES Foundation, Ministry of Education of Brazil - process: BEX 13235/13-3.

REFERENCES

[1] E. Alexander, C. Chang, M. Shimabukuro, S. Franconeri, C. Collins, and M. Gleicher. The Biasing Effect of Word Length in Font Size Encodings. In *Poster Proceedings of the IEEE Visualization Conference*. IEEE, 2016.

[2] E. Alexander and M. Gleicher. Assessing topic representations for gist-forming. In *Proc. of the International Working Conference on Advanced Visual Interfaces (AVI)*. ACM, 2016.

[3] S. Bateman, C. Gutwin, and M. Nacenta. Seeing things in the clouds: The effect of visual features on tag cloud selections. In *Proc. of the ACM Conference on Hypertext and Hypermedia (HT)*. ACM, 2008.

[4] J. Chuang, C. D. Manning, and J. Heer. without the clutter of unimportant words: Descriptive keyphrases for text visualization. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(3):19, 2012.

[5] R. L. Cilibrasi and P. M. Vitanyi. The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3), 2007.

[6] W. S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[7] C. Collins, F. B. Viegas, and M. Wattenberg. Parallel tag clouds to explore and analyze faceted text corpora. In *Proc. of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*. IEEE, 2009.

[8] P. Dragicevic. Fair statistical communication in hci. In *Modern Statistical Methods for HCI*, pp. 291–330. Springer, 2016.

[9] C. Felix, A. V. Pandey, and E. Bertini. TextTile: An Interactive Visualization Tool for Seamless Exploratory Analysis of Structured Data and Unstructured Text. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):161–170, 2017.

[10] N. Ferreira, L. Lins, D. Fink, S. Kelling, C. Wood, J. Freire, and C. Silva. Birdvis: Visualizing and understanding bird populations. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2374–2383, 2011.

[11] Google. Google books. `http://books.google.com`. Accessed: 2017-02-11.

[12] M. J. Halvey and M. T. Keane. An assessment of tag presentation techniques. In *Proc. of the International Conference on World Wide Web (WWW)*. ACM, 2007.

[13] M. A. Hearst. Whats Up with Tag Clouds? *Visual Business Intelligence Newsletter*, 2008.

[14] M. A. Hearst and D. Rosner. Tag clouds: Data analysis tool or social signaller? In *Proc. of the Annual Hawaii International Conference on System Sciences (HICSS)*, 2008.

[15] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pp. 203–212. ACM, 2010.

[16] O. Hoeber and H. Liu. Comparing tag clouds, term histograms, and term lists for enhancing personalized web search. In *Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 2010.

[17] B. Lee, N. H. Riche, A. K. Karlson, and S. Carpendale. Sparkclouds: Visualizing trends in tag clouds. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1182–1189, 2010.

[18] S. Lohmann, J. Ziegler, and L. Tetzlaff. *Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration*. Springer Berlin Heidelberg, 2009.

[19] T. Munzner. Visualization analysis and design, 2014.

[20] 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. of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2007.

[21] J. Schrammel, S. Deutsch, and M. Tscheligi. Visual search strategies of tag clouds - results from an eyetracking study. In *Proc. of the IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT)*. Springer-Verlag, 2009.

[22] J. Schrammel, M. Leitner, and M. Tscheligi. Semantically structured tag clouds: An empirical evaluation of clustered presentation approaches. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2009.

[23] D. Steinbock. Tagcrowd. `http://tagcrowd.com`. Accessed: 2017-02-11.

[24] F. Van Ham, M. Wattenberg, and F. B. Viégas. Mapping text with phrase nets. *IEEE transactions on Visualization and Computer Graphics*, 15(6), 2009.

[25] F. B. Viégas and M. Wattenberg. Timelines: Tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, 2008.

[26] F. B. Viegas, M. Wattenberg, and J. Feinberg. Participatory visualization with wordle. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1137–1144, Nov. 2009.

[27] Y. Wu, T. Provan, F. Wei, S. Liu, and K.-L. Ma. Semantic-preserving word clouds by seam carving. In *Proc. of the EG/IEEE VGTC Conference on Visualization (EuroVis)*. Blackwell Publishing Ltd, 2011.