Evaluating the Effect of Style in Information Visualization

Andrew Vande Moere, Martin Tomitsch, Christoph Wimmer,
Christoph Boesch, and Thomas Grechenig

Abstract—This paper reports on a between-subject, comparative online study of three information visualization demonstrators that each displayed the same dataset by way of an identical scatterplot technique, yet were different in style in terms of visual and interactive embellishment. We validated stylistic adherence and integrity through a separate experiment in which a small cohort of participants assigned our three demonstrators to predefined groups of stylistic examples, after which they described the styles with their own words. From the online study, we discovered significant differences in how participants execute specific interaction operations, and the types of insights that followed from them. However, in spite of significant differences in apparent usability, enjoyability and usefulness between the style demonstrators, no variation was found on the self-reported depth, expert-rated depth, confidence or difficulty of the resulting insights. Three different methods of insight analysis have been applied, revealing how style impacts the creation of insights, ranging from higher-level pattern seeking to a more reflective and interpretative engagement with content, which is what underlies the patterns. As this study only forms the first step in determining how the impact of style in information visualization could be best evaluated, we propose several guidelines and tips on how to gather, compare and categorize insights through an online evaluation study, particularly in terms of analyzing the concise, yet wide variety of insights and observations in a trustworthy and reproducable manner.

Index Terms—Visualization, design, style, aesthetics, evaluation, online study, user experience.

1 Introduction

Information visualization is concerned with exploiting the cognitive capabilities of human visual perception in order to convey meaningful patterns and trends hidden in abstract datasets. As data has steadily become more complex in terms of its size, dimensionality and time-variance, the field has been challenged to create new techniques that are more sophisticated, and to develop objective evaluation methods that are able to benchmark these different techniques against each other. Because of its strong historical roots in scientific reasoning, research in information visualization has mainly focused on optimizing performance measures for typical data exploration and analysis tasks, and particularly the aspects of usability and utility. The relevance whether visualizations might benefit – or suffer – from the use of visual or interactive embellishments, has therefore been relatively neglected, especially in terms of empirical studies. Inspired by Norman's famous mantra "*attractive things work better*" [17], such research typically aims to discover gains in task efficiency or long-term recall, to discover how embellishments can be purposefully exploited to make future visualizations even more effective.

Driven by ever more user-friendly and sophisticated visualization toolkits, the rising availability of publicly accessible and socially relevant datasets, and the emergence of educational practices that reward the merging of technical virtuosity and visual creativity, an increasing number of artists, designers and journalists are now applying information visualization principles as a powerful way of visual expression [15, 21, 32]. This online practice seems to purposefully use striking visual styles, for instance to attract the attention of a sizable audience, to compel potential users to engage with the visualization, or to share the visualization experience with others. Although many of these visualizations are based on well-proven data mapping techniques, it is still relatively unknown whether the use of expressive stylization impacts their performance, for instance in the generation of insights. Moreover, some explicit cases of extreme stylization also reveal the boundaries of the information visualization practice, in particular at which utility, usability and even usefulness play a considerably less crucial role [12].

Our research hypothesizes that the use of visual style in information visualization has a measurable effect on the kinds of insights that people discover, and on how people perceive their own discovered insights. For instance, anecdotal evidence exists on how an embellished visualization might lead to more 'shallow' insights, or that people might find these insights less trustworthy than when discovered via a less-embellished counterpart. Yet, these 'shallow' insights might lead to more subjective interpretation or personal reflection, in which the meaning of a data pattern becomes more important than its factual basis. Our study therefore did not focus on aspects that relate to task performance, but instead aimed to measure how style, made apparent visually as well as through interactive features, impacts the characteristics of the resulting insights. Is it true, for instance, that a 'traditional' scatter plot representation leads to more 'deep' insights than a stylized counterpart that conveys the exact same data?

Inspired by the hypothesis that "*casual visualizations … provide other kinds of insight that complement … [analytical insights]*" [21], this paper aims to measure of what these "*other kinds of insight*" might consist of. Therefore, this paper presents the results of a between-subject comparative study, in which three different interactive information visualization demonstrators were benchmarked against each other. The style of each demonstrator was based on the visual characteristics of a predefined collection of good practice exemplars, and their stylistic resemblance was validated by a separate categorization study. In an attempt to achieve a sufficient number of participants that counterbalance the various subjective factors (e.g. culture, gender, experience, age) that are typically involved in measuring subjective aspects such as style, and to situate the evaluation within the context of intended use [10], the comparative study was accomplished online.

- *Andrew Vande Moere is with KU Leuven. E-Mail: andrew.vandemoere@asro.kuleuven.be.*
- *Martin Tomitsch is with The University of Sydney. E-Mail: martin.tomitsch@sydney.edu.au*
- *Christoph Wimmer is with T.U.Wien. E-Mail: christoph.wimmer@inso.tuwien.ac.at*
- *Christoph Boesch was with T.U.Wien.*
- *Thomas Grechenig is with T.U.Wien. E-Mail: thomas.grechenig@inso.tuwien.ac.at*

Manuscript received 31 March 2012; accepted 1 August 2012; posted online 14 October 2012; mailed on 5 October 2012.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org

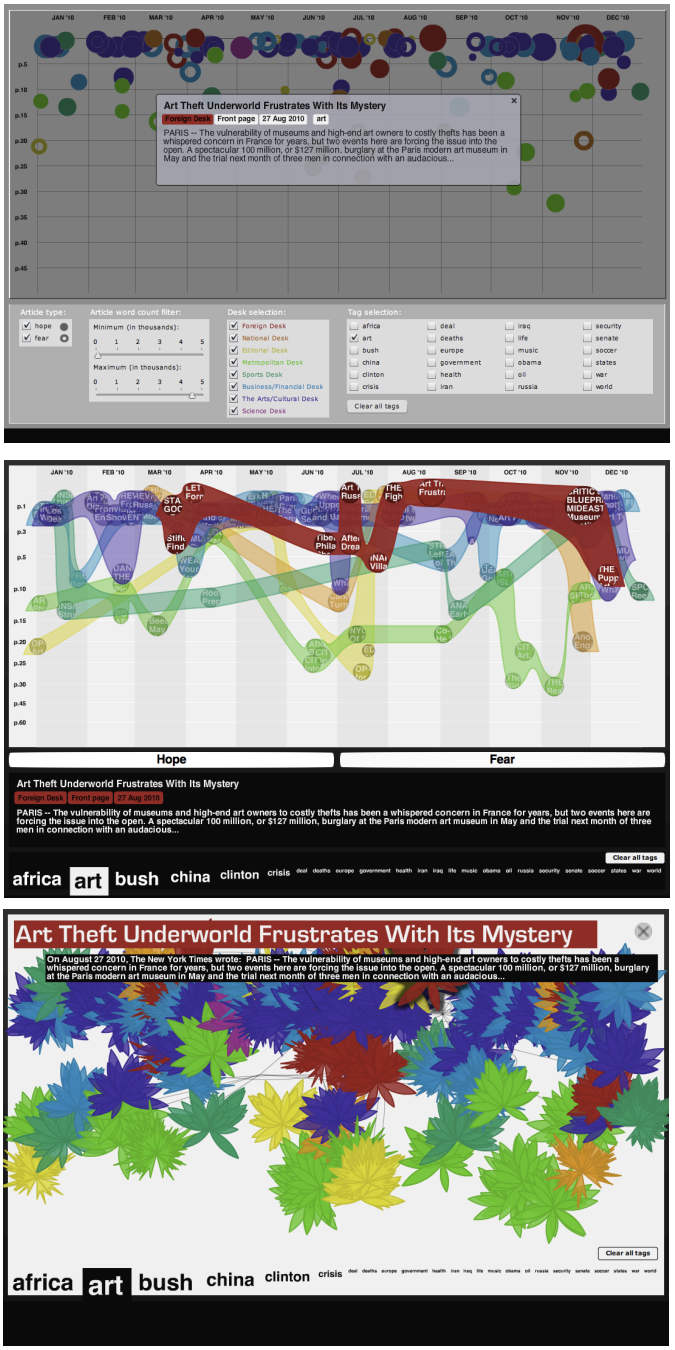

Fig. 1. The three different style demonstrators showing an identical view. In this view, the user has selected a specific news story about an art theft. Note the differences in visual treatment of the scatter plot technique and the graphical integration of the news article title, date, abstract and tags in the screen layout. Top: Analytical Style (ANA). Middle: Magazine Style (MAG). Bottom: Artistic Style (ART).

2 Background

2.1 Style

Style is an abstract concept that relates to how an artefact – such as a visualization – can be recognized, and be potentially grouped in a specific category. By choosing a specific way how a visualization is given an externally recognizable form – visually as well as in its interactive features – a developer consciously or unconsciously establishes a set of 'rules'. If other developers consider these rules inspirational for their own approaches, they might also apply identical, or very similar, characteristics, so that the according visualizations then take over that specific 'style'. Some empirical evidence exists that style plays an important role in the perception of users, as it is often the only 'way' to make a product stand out [28]. However, it is often the "social circumstances" surrounding the design of an information presentation that determines the choice of a style, which more often than not tends to "*differ from those described by the rational approach*" [30]: developers deliberately tend to adopt different stylistic preferences (e.g. the use of depth), depending on whether they aim to create a favorable impression versus providing information for optimal decision-making.

While some people fear the danger of perceiving style as more important than substance, style has become a ubiquitous phenomenon of which the positive effects should not be ignored. Although the use of style does not overcome evident issues of bad usability or reliability of a particular system, it tends to matter when all else is equal [20]. For instance, the main motivation of applying an 'artistic' style in visualization has been linked to the aim to convey insights that are neither objective nor connected to productivity metrics, but instead have a forceful or actionable meaning [32], to provide insights into mundane activities [21], or to create the awareness that "*the data exists at all*" [12]. On the other hand, aesthetics, one particular aspect of style, can reach well beyond the experiential or the superficial, as it has been shown to positively influence task performance [13, 29]. For instance, latency in task abandonment and erroneous response time are correlated to a visualization's perceived beauty [3], search task efficiency improves with a more "classical" layout of visual objects [26], and non-utilitarian "visual embellishments" do not seem to affect interpretation accuracy, and positively influence long-term recall in the case of simple infographic charts [1].

2.2 Insight Reports

Information visualization research has dedicated an increasing amount of attention to develop objective evaluation methodologies. One direction focuses on how visualization amplifies analytical reasoning by measuring its ultimate purpose, that of *conveying insight* [18, 23]. Although a commonly accepted definition of insight has yet to emerge in the community, some early classifications [4, 5] and insight-acquiring processes [33] have already been proposed. In our study, we have compared how the use of style in visualization impacts the generation of insight, in order to "*enable the direct comparison of visualization design alternatives*" [18]. To the best of our knowledge, few studies exist that deployed an insight analysis methodology to benchmark different visualization approaches against each other, and those that did were accomplished in a controlled lab environment applying the talk-aloud method to record the insights [19], focused on comparing analytical methodologies [23] or determined the impact of a particular design approach [8]

3 Demonstrator Design

The first phase of our study involved the design of three visualization demonstrators that differed in stylistic approach.

3.1 The Dataset

Each demonstrator was based on an identical dataset, in order to guarantee their comparability in terms of the insights that they could potentially generate. The dataset was chosen to be agnostic to a specific stylistic approach, in that some datasets inherently carry a style metaphor. For instance, people might expect data about dance music to be shown through a rather 'experimental' style, while cancer statistics might require a more 'scientific' style. Therefore, each demonstrator displays the same collection of historical news stories gathered from the U.S. newspaper The New York Times. The topic of news was chosen because it forms a common subject of many existing popular visualizations online, and because it has a natural affinity to science as well as art. News is 'scientific' in terms of being quantifiable, such as in terms of an article's word count or its date of publication; and categorical, in its thematic focus.

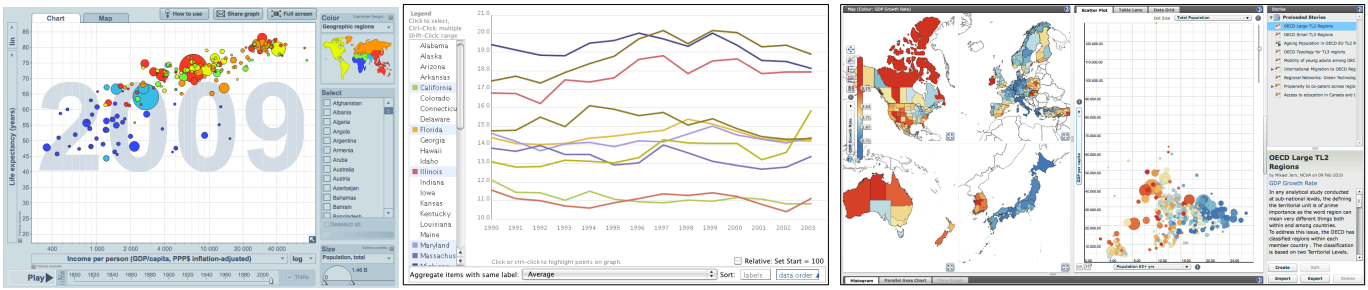

Fig. 2. Analytical visualization style exemplars. Left: Gapminder [25]; middle: Many Eyes [31]; right: OECD eXplorer [11].

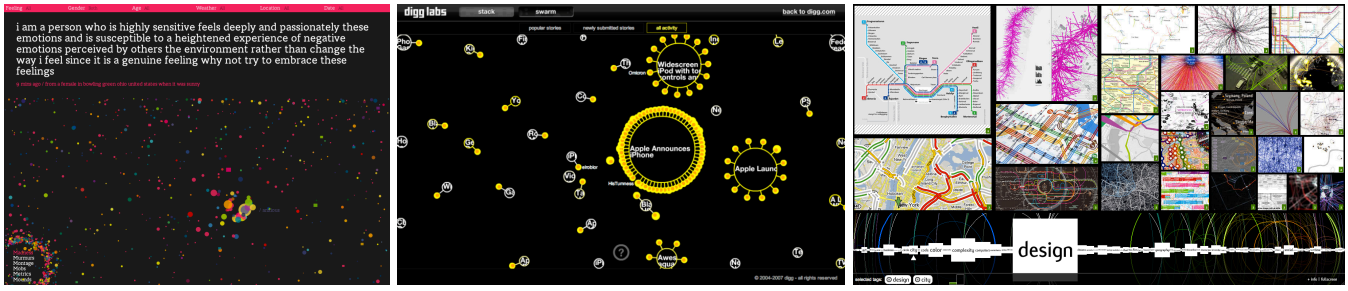

Fig. 3. Magazine visualization style exemplars. Left: We Feel Fine [9]. Middle: Digg Labs [7]. Right: remap [2].

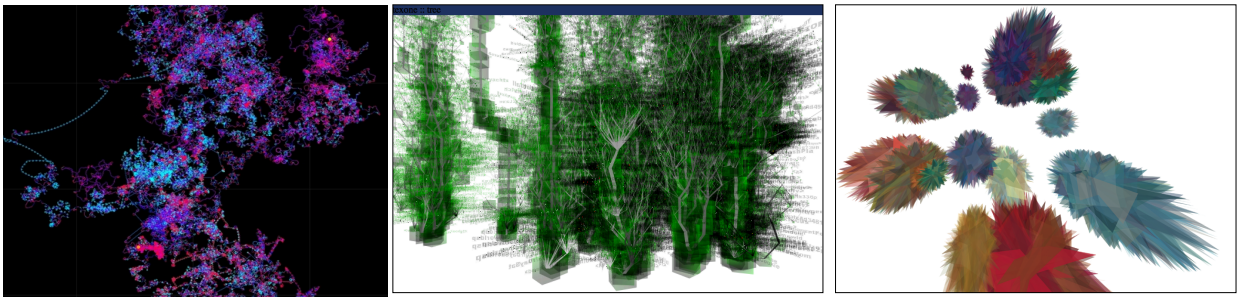

Fig. 4. Artistic visualization style exemplars. Left: Bitalizer [22]. Middle: Texone [24]. Right: Poetry on the Road 2004 [16].

News can be 'subjective' in terms of its implicit meaning or its personal interpretation. Since early 2009, the NYTimes offers an Article Search API that aims to make the discovery and exploration of news content easier [27]. Using this service, we generated a dataset that contained 4644 unique news stories that featured the terms 'hope' or 'fear' and were published between 1 January and 31 December 2010. These two terms were also used as filters in order to limit the dataset size, which in turn influences the performance and technical complexity of the demonstrators. In addition, the orthogonal meaning of these two terms was meant to facilitate different avenues of personalized data exploration. Each news story consisted of a title, a short abstract, the publication date, the page number and its news desk. In addition, a set of 24 keywords (tags) was derived by ranking and filtering the most frequent words within all the collected news articles.

Each demonstrator was based on the traditional scatter plot approach: each unique news story was mapped in terms of time (X-axis) and page number (Y-axis). The size of each mapped visual element corresponded to the word count of the corresponding news article. The technical implementation was accomplished based on the Adobe Flex 3.0 framework and the Flare ActionScript library [14].

3.2 The Demonstrators

Our design process focused on varying the visual and interaction styles of three demonstrators so that they could be independently recognized to belong to a specific stylistic direction, while keeping all other aspects as constant as possible. While such a design brief seems relatively simple, its execution proved to be far more complex. While each demonstrator should be representative of a specific style, none should 'stand out' from the others, neither positively nor negatively. However, a developer typically does not have the same affinity for different styles, while multiple developers working on separate styles have difficulty to adhere to universalized design constraints.

The stylistic differences were grounded on the visual and interactive qualities observed in nine best-practice exemplars, which we grouped in three distinct styles (see Figures 2-4). The nine exemplars were selected based on the findings of the "information aesthetics" model [15]. This two-axis model captures how the visualization practice balances the communication of data patterns (*intrinsic* in terms of conveying facts and trends) versus meaning (*extrinsic* in conveying what underlies the data patterns) through the use of *direct* (e.g. reversible in terms of recognizing data values from the representation) versus *interpretative* (e.g. irreversible) mapping techniques. These nine exemplars were clustered in three groups in the belief that two groups demonstrated two extremes of the model – i.e. Analytical is intrinsic and direct, Artistic is extrinsic and interpretative – while Magazine forms a 'middle ground'.

3.2.1 Analytical Style Demonstrator (ANA)

The design of the ANA demonstrator (see Figure 1, top) was based on a shortlist of existing scatter plots that facilitate the analysis of statistical data for lay users, and are relatively popular in the online visualization practice, such as "Gapminder" [25], "Many Eyes" [31], and "OECD explorer" [11] (see Figure 2). The design aspects that were isolated and then incorporated in this demonstrator include: dedicated screen space for user interface elements, such as a list of checkboxes; a background grid and prominent text labelling; and value-specific categorization (i.e. color) and mapping (i.e. scaling of

bubbles). ANA offered a task-specific filter that allowed sorting news stories by their word count. The ANA demonstrator also copied how the 'graph' becomes separated from the 'content': while the visual elements could be hovered to receive summary information, the news article blurb appeared (after user selection) in a separate light-box screen that overlaid the actual scatterplot graph, which was then darkened.

3.2.2 Magazine Style Demonstrator (MAG)

"We Feel Fine" [9] (see Figure 3) demonstrates smooth, interactive animations, a lack of traditional menu items, and a tight integration of content and graph as more detailed information appeared directly above the visualization, without overlaying it. "Digg Labs" [7] was taken as an example of how textual and visual elements can be tightly integrated, such as how a story title is cropped inside a circular element, appearing only in full after hovering the mouse. "Remap" [2] demonstrates an alternative approach to the common checkbox list filtering, as it utilizes an animated 'fisheye' scaling of keywords. We also took inspiration in the apparently useful yet quite aesthetic "Bubble Set" technique [6], which uses continuous and concave iso-contours to delineate the membership of multiple stories to the same news desk: the changing thickness and swerving nature of these shapes were meant to better highlight the varying but continuous nature of thematic news importance over time. Notably, MAG featured no color legend, as it was intended that users gained the news desk category solely through paying attention to the news article blurb pane. The selection of articles was accentuated by a 'swoosh' sound and a smooth animation of the selected 'bubble' floating towards the article blurb pane at the bottom. The graph always displayed the articles of both 'fear' and 'hope', but the ones of the inactive category were blurred in the background. Black lines connected articles with similar keywords. The Y-axis was logarithmic, to dedicate more space to the first 10 pages, which were most densely populated.

3.2.3 Artistic Style Demonstrator (ART)

"Bitalizer" [22], "Texone" [24] and "Poetry on the Road" [16], all shown in Figure 4, demonstrate how different sorts of data – digital files, HTML structure and poetry text, respectively – can be interpreted as purely numerical parameters that create compelling visual forms by way of clever data-to-shape generation algorithms. Accordingly, our demonstrator attempted to mimic this approach by depicting individual articles as flowers: 'hope' articles were depicted by boat-shaped petals, while 'fear' articles had petals with spikes. Articles with common tags were connected by organic black lines. In contrast to the other two demonstrators, ART featured no mouse hover preview prior to article selection. The selection of individual articles triggered a visual and audio typewriter-like effect to reveal the article blurb, which was more elaborately visually treated and appeared on the top of the graph. Like MAG, ART did not include a color-to-news desk legend and had no axis labelling whatsoever, in order to encourage users to 'decipher' the visual mapping by actively relating the visual attributes to the content. ART also featured an ambient background sound track.

3.2.4 Style Consistency Validation Experiment

In an attempt to stay as close as possible to the given style exemplars, the three demonstrators contained various elements (listed in Table 1) in terms of interactivity, sound and visual prominence that might not be strictly recognized as "stylistic" features. Subsequently to the development process, we therefore validated our adherence to the three predefined stylistic approaches by querying 8 students and Faculty staff members (5 male, 3 female) originating from disciplines related to design. None of them were previously involved in the study, and all had little knowledge of information visualization. Participants were presented with printouts of the selected nine exemplars (i.e. Figures 2-4) arranged into the three stylistic clusters, and printed screenshots of our three demonstrators (i.e. Figure 1).

Table 1. Stylistic & non-stylistic differences among demonstrators.

	ANA	MAG	ART
Hover Preview	Summary information	Summary information	Not available
Available Filters	Hope vs. Fear Word count News desk Keywords	Hope vs. Fear Keywords	Keywords
Filter Controls	Checkbox list Range Slider	Liquid keyword list Hope/Fear buttons Bubble graph	Liquid keyword list
Available Legend	X axis Y axis (linear) News desk color	X axis Y axis (logarithmic)	Not shown
# Words	Bubble size	Bubble size	Flower size
Hope vs. Fear	Circle vs. doughnut	Color shades	Spiky vs. rounded petals
Article View Position	Light-box on top of graph	Below graph	Overlaying on top of 'graph'
Audio	Not used	Swoosh sound	Typewriter effect Background music

We asked them to assign each of the demonstrators to one of the clusters. Through thinking-aloud and follow-up questions we asked participants to describe each category with adjectives and why they placed each demonstrator in a cluster. Six of the participants assigned all demonstrators to the same cluster of exemplars that we used as design inspiration for the respective demonstrator. The remaining two people assigned the MAG demonstrator to the ANA cluster, as they considered MAG to resemble a traditional scatter plot representation. As they oversaw the prominent interface controls, they focused on the colors and circular shapes to determine their choice.

All participants were very confident when talking about the clusters as being different styles. The analysis of the think-aloud protocol and recorded answers showed that all participants described the ANA cluster with quantitative adjectives, such as *analytical, scientific, structured*, and *technical*. The ART cluster was described with terms such as *abstract, artistic, arty*, and *beautiful*. Participants stated that they would expect to find this type of visualization in an art gallery, while they thought that ANA was used in an accounting or news environment. Participants seemed to find it less straightforward to come up with descriptive adjectives for the MAG stack, but five of them thought it was very *designed, creative*, or *aesthetic*. Two participants explicitly stated that they would expect to find MAG in magazines or the public sector. One participant pointed out that the round shapes in the MAG demonstrator were playful and "*like something you want to touch*", while none of the participants mentioned a similar emotional affordance for the ART demonstrator.

Supported by the relatively high overlap in the participants sorting the demonstrators and confidently describing their respective styles during the validation experiment, we decided not to make changes to the design of the demonstrators.

4 Evaluation Methodology and Results

This study aims to measure how style, in terms of its visual and interactive features, influences the kind of insights people generate.

4.1 Evaluation Study Setup

The evaluation study occurred online in order to reach a sufficiently large participant audience, while the online medium also mimics the real-world communication channel [10] of today's popular visualization practice. Participants were recruited through a call on a visualization-focused blog, via messages on several mailing lists on the topic of visualization and human-computer interaction, and by (re)posting the link on various electronic social networks.

The evaluation study consisted of a between-subject user experiment, in order to minimize learning effects, to avoid the cross-fertilization of insights between each demonstrator, and to limit the

required time and effort to participate in the study. Each participant was allowed to only partake in the study once, as a browser cookie blocked any recurrent access attempt. Naturally, there exist ways to circumvent this restriction, though the between-subject design aspect was always kept hidden for all participants at all times. The study was designed to require between 15-20 minutes, and participation was fully anonymous and without any reward.

The landing page contained an introduction, stated the time required to complete the study and the technical requirements (e.g. browser plug-in, screen size). The study launched in a dedicated browser window, fixed to 1440x900 pixels. While this resolution excluded some users with smaller screens, it was essential to assure readability and the continuous presence of the insight report form.

4.1.1 Pre-Study: Introduction Stage

The online study consisted of three distinct stages, of which the first displayed a short, narrated tutorial video. The video format was chosen in favor of a textual or graphical explanation, to assure a high rate of compliance. The purpose of this introductory video was to: 1) provide a brief explanation of the chosen dataset (e.g. NYTimes news data filtered by 'hope' and 'fear'); 2) give a brief overview of the study's purpose, i.e. collecting insights, together with a succinct definition of what an insight constitutes; 3) explain the demonstrator, including its purpose, its visual structure and its interaction features; and 4) demonstrate how an insight could be discovered and subsequently recorded with the web form. As each participant was presented with a video that explained the demonstrator they would interact with, three different videos had to be created. While each video had the same duration (i.e. 2m20s), and demonstrated the same insight discovery process, some visuals and terms were swapped to correspond to the respective demonstrator. To convey some idea of how much time and effort was expected, a message at the start of the study encouraged participants to discover about 3 different insights.

4.1.2 Study: Insight Recording Stage

In the second stage, participants were presented with the main study. On the right side of the demonstrator, a narrow web form was displayed that allowed participants to enter a single insight, which could then be recorded without having to refresh the visualization. Participants were asked to: describe the insight; rate their confidence in, and deepness of, that insight; explain how they came about this insight; and estimate how difficult it was to generate this insight. Each question featured an additional, brief 'help tip' that reiterated the issue in a more descriptive way. For instance, the help tip for insight depth read: "*Rate in how far the insight brings about new knowledge or creates further interesting questions*." A general 'help' link on the top of the screen showed a textual explanation of the dataset, how it was translated into visual form, and reiterated the available interaction features. The help description was identical for the three demonstrators, except where some demonstrator-specific terms were replaced for clarity reasons. Participants were free to finish at any time, even without recording any insight, by selecting a 'Finish' button. Participants were warned they would not be able to go back before proceeding.

4.1.3 Post-Study: Survey Stage

The last stage consisted of a single-page survey form. First, participants had to rate eleven different qualitative properties on a five-point semantic differential scale. The order of the labels (i.e. implied negative versus positive connotation) was randomized to prevent implicit value judgements in the form. Participants were also presented with four open questions: what they liked about the visualization, what they disliked, which problems they experienced, and what they would use the visualization for. Lastly, a questionnaire queried for the participants' age, gender, their expertise regarding information visualization, and their literacy regarding news. Participants could choose to receive a report about the study results by leaving their email address. The email data was never associated with the study data in order to maintain full anonymity. All fields in this questionnaire were compulsory except of birth date and gender.

4.1.4 Data Logging

Each visit to the study was stored in a unique cookie, ensuring the anonymity of participants while allowing us to prevent returning visitors to break the between-subject design. Next to the web forms, the system also recorded the time spent, the number of steps completed, and all user interactions. For each interaction, the timestamp and the interface element that was clicked, as well as the overall state, such as the keywords that were already selected, was recorded. The evaluation study framework was custom-developed as a PHP web application, logging data as CSV files directly to disk.

4.2 User Participation Analysis

Each participant was assigned to one of the three conditions upon first accessing the study through round-robin. In total, 4192 people visited the study website over the course of four weeks. A total of 762 people interacted with the demonstrators in some way: ANA (N=224), MAG (N=302), ART (N=236). 142 of these completed the study: ANA (N=45), MAG (N=53), and ART (N=44). A study entry was considered as completed, if the survey stage was successfully submitted. Successful participants spent on average 14m14s to finish the study (SD=13m22): ANA (M=18m09s, SD=17m53s), MAG (M=12m49s, SD=10m23s), ART (M=11m55s SD=10m12s). Although the analytical style counter-intuitively led to the longest engagement, these durations were influenced by the suggestion that the study would take up to 20 minutes, and are thus of limited value as a measure of user engagement. We applied a ln-transform and ANOVA with post-hoc Games-Howell tests, as they were not normally distributed. This analysis revealed a significant difference between ANA and MAG, as well as ANA and ART at $p<.05$, but no significant difference in duration between MAG and ART.

We discarded 4 results: three (ANA:1, MAG:2) because of insufficient activity and one (MAG) because of technical problems described in the post-test questionnaire. This left us with 138 valid submissions: ANA (N=44), MAG (N=50), and ART (N=44). 21 participants completed the study without recording any insight: ANA (N=1), MAG (N=11), and ART (N=9). 32 participants (23.2%) were female, 104 (75.4%) were male. Two (1.4%) chose to not state their gender. The age of participants ranged between 19 and 67 (M=34.36, SD=9.36), while the self-reported expertise with visualization was relatively high (M=3.05, SD=.69) on a scale from 1 (no experience) to 4 (expert): ANA (M=2.89, SD=.63), MAG (M=3.16, SD=.71), ART (M=3.09, SD=.71). We did not find any significant difference between the three conditions in terms of the self-reported demographic criteria.

4.3 User Interaction Analysis

We analyzed the types of interaction patterns performed for each demonstrator by tracking each individual operation, where an operation is defined as an interaction that results in a state change of the visualization. A total of 762 people performed at least one operation with one of the demonstrators. All subsequent analysis is limited to the 138 participants who successfully completed the study. Participants performed twice as many operations in ANA (M=181.0, SD=205.32) than both MAG (M=87.7, SD=83.89) and ART (M=88.9, SD=96.63). As the number of operations was not normally distributed, we applied a log-transform and performed an ANOVA with post hoc Games-Howell tests, revealing a significant difference between ANA and MAG as well as ANA and ART at $p<.05$, but no significant difference between MAG and ART. This interaction behavior might be best explained because MAG and ART were conceptually similar in their interaction features, while ANA offered specific data filtering operations. There were considerable differences in terms of the mean number of articles clicked to reveal more detailed information: ANA (M=1.09, SD=2.13), MAG (M=7.62, SD=13.73), ART (M=36.32, SD=38.73), suggesting fundamentally

different interaction patterns. Where ANA users were interacting on a higher pattern-seeking level, MAG and ART users were more inclined to learn about the 'content' of the news articles. While the large difference for ART might be due to the lack of a 'mouse-hover' preview of news headlines, the considerable difference between MAG and ANA cannot be explained by the article preview or select features, as they were essentially similar. Therefore, the difference might be better explained by the tight visual integration of the detailed news pane. Tag operations were only different between ANA (M=91.2, SD=117.5) versus MAG (M=38.5, SD=43.8) and ART (M=44.6, SD=62.0), as the last two featured the same interactive fisheye keyword menu.

4.4 Insight Analysis

From the 138 participants, we collected 315 valid insights. 4 insights explicitly discussed usability issues, and were not considered for further analysis: ANA: 107 insights (M=2.43 per participant, SD=1.26); MAG: 112 insights (M=2.24, SD=1.93); ART: 92 insights (M=2.09, SD=1.95). An ANOVA did not show a significant effect on number of insights per participant, probably as the study explicitly encouraged the submission of about 3 different insights.

4.4.1 Insight Typology Analysis

Fact Typology. Two researchers classified each insight based on Chen et al.'s [5] fact taxonomy. It was chosen not to add more rating experts, as this approach only led to an inter-coder agreement of 34.4%. For instance, the relatively simple insight "*There seem to be fewer articles in the middle of the year*" (ART), can potentially be classified as 'distribution' ("skewed distribution"), outlier ("density difference"), 'cluster' (e.g. "dissimilarity between this and other clusters"), or, when other information such as *"… than the last part of the year"* is implied, as 'difference' (e.g. "distribution between elements"). Each researcher revisited all insights, which led to an agreement of 89.4%. The final classification was consolidated by deciding upon each conflicting rating in mutual agreement.

Meta Fact. A 'meta fact' (ANA: 6% (6), MAG: 12% (13), ART: 29% (27)) typically contains a comment on the user interface or study setup, instead of describing a fact grounded by the graph, and varied between *"there should be a … OR instead of AND Boolean-operator option*" (ANA), *"This chart is terribly, terribly confusing. I say this as a data visualization professional"* (MAG) and *"…the piling up of symbols seems intentionally designed to make it hard to read them…"* (ART). The relatively large number of 'meta facts' identified for ART can most likely be related to its visual ambiguity and the lack of any direct legend or explanation (except of the 'Help' feature). As a result, several participants reported the deciphering of the data mapping as an insight, such as the meaning of petals and colors: *"Green leaves are from the sport section"* (ART). Few participants described insights by directly referring the visual representation: *"On the front page, art is treated chiefly by the blue department. On the back pages, it is covered by the green department"* (ART). Others associated alternative meanings to the data mapping: *"Iraq itself is very bad (very red)"* (ART), and *"The orange petals have the most interesting stories. They cover various topics and they seem to have more personal outlooks"* (ART). Some took the opportunity to be humorous: "*All stories in the New York times are correlated with spooky music*" (ART). The clear distribution of 'meta facts' among the three demonstrators most probably demonstrates the difference in clearness, intuitiveness and usability. However, it also skews the relative occurrence of different fact types in favor of comments, so that we decided to not include 'meta facts' in our further analysis.

Other Facts. Table 2 reports on the performance of each demonstrator in terms of the discovery of fact-based insights. Especially with ANA, many participants identified 'clusters' (22%), like *"Obama's on the front page a lot"* (ANA). While the number of insights relating to 'value' facts was generally low, the vast majority originated from ANA (6 out of 7), such as *"Articles with tag 'Russia' contain about 40% fears"* (ANA).

Table 2. Insights by fact type [5], in relative and (absolute numbers). (*) 'Meaning' category added by authors.

	ANA	MAG	ART
Difference	24% (24)	26% (26)	17% (11)
Cluster	22% (22)	15% (15)	9% (6)
Distribution	11% (11)	12% (12)	17% (11)
Compound	9% (9)	14% (14)	11% (7)
Trend	8% (8)	4% (4)	8% (5)
Outliers	6% (6)	10% (10)	15% (10)
Value	6% (6)	1% (1)	0% (0)
Association	5% (5)	3% (3)	6% (4)
Meaning (*)	3% (3)	4% (4)	14% (9)
Extreme	4% (4)	6% (6)	0% (0)
Categories	2% (2)	1% (1)	0% (0)
Rank	1% (1)	2% (2)	3% (2)

Insights indicating 'extremes', which typically resulted from participants deliberately looking for extreme values: *"There is only one article that mentions 'hope' along with Clinton, Bush and Obama all together"* (MAG), were not recorded for ART. However, ART was ideal in terms of conveying 'outliers', such as "*There are only few reports on music from Iraq or Iran*" (ART). ANA (8%) and ART (8%) led to more insights regarding 'trends', such as *"Reactions to the Gulf oil spill caused a significant spike in coverage of 'oil', however interest dropped off within a few months"* (ART). In spite of the iso-contour shapes highlighting semi-continuous movements over time, 'trends' performed the worst for MAG (4%). 'Difference' insights were prevalent in MAG (26%) and ANA (23%) when compared to ART (17%), such as *"Despite a lot of fear at the beginning in Obama's health care revolution plan, no more fear after the bill were passed*" (ANA). ART performed best in terms of discovering 'distribution'. Surprisingly, 'association' insights were evenly distributed over all 3 demonstrators.

4.4.2 Insight Meaning Analysis

Meaning. To better acknowledge the characteristics of insights, we decided to add a separate class to Chen et al.'s taxonomy, which we coin as 'meaning'. Meaning insights contain obvious connotations to the content, which is discovered through exploring the graph, and do not directly relate to the graph, such as: "*The science desk has relatively nothing to say about hope or fear until it comes to health or life matters. Why has science nothing much to say about hope or fear in other topics?"* (ANA), "*The Togo national soccer team were machine-gunned by terrorists in 2010*" (ART), or "*The NYt's coverage of the Senate gives a false sense of hope*" (MAG). The highest relative number of 'meaning' insights were found for ART (14%), while ANA (3%) and MAG (4%) performed almost equally. As the limited interaction features of ART encouraged reading article blurbs, participants might felt forced to pay more attention to content instead of hunting for visual patterns.

Fact with Meaning. Some insights included some form of meaning to contextualize or explain the driving principles behind a reported fact: *"…few [articles of Obama] … with the tags 'senate' and/or 'government'. This means Obama is treated in an isolated way, not to say out of context."* (ANA), "*In June-July 2010, most of the sport articles talk about hope. This might be due to quoting what coaches or players said before a football match of the 2010 Football WC*" (MAG) and *"The visualization can reveal the activity of a politician. While Obama and Clinton are still active, Bush retires"* (ART). Although all "fact with meaning" insights were coded on the basis of their fact in Table 1, they are still worthwhile to report on separately: ANA:12% (12), MAG:8% (8), and ART:3% (2). This result shows again how ANA prefers to utilize meaning to explain a fact, whereas ART highlights content independently from the graph.

Depth. Participants self-rated each insight regarding how they perceived their confidence, its depth and the difficulty to discover the insight, on a five-point semantic differential scale (see Table 3). An ANOVA showed a significant effect of demonstrator on depth ratings ($F(2,262)=6.56$, $p<.01$).

Table 3. Mean values and (standard deviation) of self-reported and post-study depth ratings. Significant differences highlighted in bold.

Rating (1 - 5)	ANA	MAG	ART
Uncertain - confident	4.10 (1.11)	4.21 (0.87)	4.17 (0.95)
shallow - deep	**3.18 (1.10)**	2.93 (1.08)	**2.54 (1.17)**
difficult - easy	3.78 (1.17)	3.63 (1.29)	4.00 (1.24)
shallow – deep (expert rating)	2.44 (0.78)	2.36 (0.70)	2.28 (0.64)

Post-hoc Bonferroni tests showed significant differences between ANA and ART (p<.01), but not between ANA and MAG, or MAG and ART. Participants estimated insights significantly deeper for ANA than ART, although they might have considered the complexity to discover the insight, rather than solely the insight itself. ANOVA showed no significant effect of demonstrator on both confidence and difficulty ratings. Using a custom-built rating application that anonymized and randomized all insights, all insights were rated on depth by three researchers, individually. The results ranged from a deep "*While editorials about Bush frequently used the terms 'hope' and 'fear' both, the terms were avoided in other types of articles. By contrast, 'hope' and 'fear' occurred in news articles about Clinton at rates almost equal to their prevalence in editorials*" (MAG, rating: 2.7) to a shallow "*Lots of hopeful news on sport*" (ANA, rating: 1.1). Although the post-study blind expert ratings indicate a similar trend than the self-reported ratings, ANOVA showed no significant differences between demonstrators.

4.4.3 Insight Categorization Analysis

Table 4. Coding categories and distribution from the card sorting classification, in relative and (absolute numbers).

	ANA	MAG	ART
Emotional	5% (5)	11% (12)	11% (10)
Rational	7% (8)	6% (7)	7% (6)
Analytical	48% (51)	34% (38)	28% (26)
Plain	34% (36)	36% (40)	33% (30)
Technical	6% (6)	7% (8)	17% (16)
Interface	1% (1)	6% (7)	4% (4)

Given the low inter-coder agreement with the typology analysis (see Section 4.4.1), we decided to use an open coding strategy. We selected two groups of two participants (all male), one of PhD students, the other of student interns. All participants were not involved in the study before, and had no prior knowledge of the insight report methodology. Each group was provided with a set of cards that each contained the individual insights in randomized order, and without any annotation to which style they belonged. Participants were instructed to group insights into similar categories that dealt with the 'type' of insight. The two groups, together with two senior researchers then jointly decided upon a common categorization scheme, by revising some of the original codings and renaming one of the categories. Notably, it is possible that other analysts following the same process may have grouped the insights differently. However, this approach gave us a more workable alternative, as it better captured the tacit differences in insights. The categories we decided upon were:

Rational. An observation that contains some reasoning, such as 'why' it occurred (e.g. "*X is more than Y, because of...*").

Technical. An observation that is based on deciphering a visual result, often through describing 'exact' filter settings (e.g. "*there are X articles tagged 'Y' with more than 'Z' words*") .

Emotional. An observation that contains a subjective interpretation (e.g. "*it is strange that…*", "*it seems that…*").

Plain. A broad, general observation with no reasoning and few filters ("*most X are Y*").

Analytical is based on a visual pattern, such as a similarity, a trend, or a comparison, typically involving a series of observations.

Interface. Comments that related to perceived problems in the interface, such as the lack of a legend.

Table 4 shows the according distribution over the three styles, demonstrating a wide range of insights that are not necessary factual. ANA led to almost twice as many 'analytical' insights than the other demonstrators. ART led to a higher number of 'technical' insights, potentially because it was difficult to observe more intricate patterns, or relate them with other trends, and participants instead tried to understand the visual mapping technique that was used. The large number of 'interface' insights for MAG and ART indicate the frustration of some participants when they approached these demonstrators with a goal-focused mindset. Insights categorized as 'rational' and 'emotional' largely correspond to what we earlier described as "meaning" and "fact+meaning". MAG and ART resulted in a larger number of 'emotional' insights, likely due to the lack of filters, which might have encouraged participants to make more spontaneous or ambiguous observations. Notably, the total number of observations from these categories is similar across all styles, while Table 2 shows that most of "meaning" observations resulted from ART. Thus, it seems that ART led participants to record a meaning or interpretation, of a finding. While in ANA and MAG, participants recorded a fact, followed by its interpretation.

4.5 Style Preference Ratings

Table 5. Mean preference ratings and (standard deviation). Significant differences are highlighted in bold.

Rating (1 - 5)	ANA	MAG	ART
ugly - beautiful	3.48 (0.85)	3.08 (1.03)	3.11 (1.02)
ambiguous - clear	**3.39 (1.17)**	**1.98 (0.89)**	**2.00 (0.86)**
boring - engaging	**3.43 (0.93)**	3.10 (0.95)	**2.80 (1.00)**
difficult - easy to understand	**3.55 (1.04)**	**2.08 (1.07)**	**2.14 (1.07)**
intended inform – express	**2.80 (1.15)**	**3.54 (1.18)**	**3.66 (1.06)**
useless - useful	**3.61 (0.95)**	**2.70 (1.09)**	**2.45 (0.90)**
frustrating - enjoyable	**3.43 (1.00)**	**2.54 (1.16)**	**2.34 (1.06)**
unusable - usable	**3.77 (0.91)**	**2.78 (1.13)**	**2.64 (1.12)**
obtrusive - fluid	3.27 (0.95)	3.08 (1.01)	2.80 (1.00)
non-functional - functional	**3.93 (0.82)**	**2.80 (1.18)**	**2.50 (1.13)**
tool - art	**2.30 (1.07)**	**3.32 (1.19)**	**3.68 (0.93)**

An ANOVA test of the self-reported preference ratings showed a significant effect of demonstrator on clearness (F(2,135)=30.51, p<.01), engagement (F(2,135)=4.84, p<.01), ease of understanding (F(2, 135)=27.64, p<.01), intention to express or inform (F(2, 135) = 7.60, p<.01), usefulness (F(2,135)=16.96, p<.01), enjoyability (F(2,135)=12.89, p<.01), usability (F(2,135)=15.12, p<.01), functionality (F(2,135)=22.51, p<.01) and categorization as art or tool (F(2,135)=19.86, p<.01). Post-hoc Bonferroni tests showed significant differences between ANA and MAG on clearness, ease of understanding, intention to express or inform, usefulness, enjoyability, usability, functionality and categorization as art or tool, all with (p<.01). Post-hoc Bonferroni tests showed significant differences between ANA and ART on clearness, engagement, ease of understanding, intention to express or inform, usefulness, enjoyability, usability, functionality and categorization as art or tool, all with (p<.01). Post-hoc Bonferroni tests revealed no significant differences between MAG and ART in any preference rating whatsoever. An ANOVA revealed no significant effect of demonstrator on beauty and fluidity. For utilitarian and functional characteristics, but also in terms of enjoyability, ANA thus scored significantly higher than both MAG and ART. MAG and ART were considered more as works of art with an intention to express, whereas participants rated ANA more as a tool with an intention to inform. Overall, it seems typically utilitarian and functional characteristics led to the perception of enjoyability. From a usability standpoint, the analytical style (ANA) is the clear winner.

4.6 Subjective Assessment

We used affinity diagramming for grouping the open comments from the post-study survey. Each comment was divided into individual parts, which were clustered into groups, resulting in 20 groups for

'likes', and 16 groups for 'dislikes'. Overall, ANA received the highest number of positive comments (1.41 per participant, MAG:1.3, ART:1.18), while MAG received the highest number of negative comments (MAG:1.98, ART:1.75, ANA:1.55). There were a number of positive comments regarding visual design and appearance (ANA:14%, MAG:20%, ART:18%) and layout (ANA:5%, MAG:2%, ART:2%) across all demonstrators. However, comments that the demonstrator was "*pretty*", "*beautiful*" or "*aesthetically pleasing*" were more common for MAG (16%) than ANA (5%) and ART (7%). Positive comments regarding usability aspects such as ease of use (ANA:14%, MAG:4%, ART:5%), learnability (ANA:7%, MAG:4%, ART:5%) and clearness (ANA:20%, MAG:4%, ART:5%) were most pronounced for ANA. A number of participants mentioned that they liked the choice and size of the data set across all three demonstrators (ANA:9%, MAG:8%, ART:11%). No one mentioned that the ANA demonstrator was fun or enjoyable to use, while a few people did say so about MAG (2%) and ART (5%). Only 1 participant had nothing positive to say about ANA, but 6 for MAG and 7 for ART.

For each demonstrator, the highest number of negative comments regarded limitations of the representation (ANA:21%, MAG:29%, ART:34%). This included visual clutter in ANA ("*difficult to distinguish hope and fear if there are many articles displayed*"), dislike of color ribbons or clarity of the lines connecting articles in MAG, and flower size or connecting lines in ART. In ANA, several participants commented on the limitations of the available data (16%), such as its short timeframe. A few asked for more quantitative tools (ANA:6%; MAG:2%, ART:3%). However, the number of participants commenting on missing control elements (e.g. zoom) and limitations of present control elements was higher in ANA (22%) than in MAG (14%), and ART (8%): seemingly, the more interaction features one offers to users, the more they notice the lack of other controls. There were very few (1) general expressions of dislike for ANA, compared to MAG (15) and ART (13), such as: "*irritating, too much eye-candy, data-ink-ratio*" (ART) or "*disorientation, frustration*" (ART). ART received the most comments regarding the lack of descriptions (12), followed by MAG (7) and ANA (3), probably due to the absence of a legend in both.

5 Discussion

Benchmarking visualizations against each other based on the analysis of their insights is challenging, and several critical observations regarding this study and its methodology need to be made.

5.1.1 Study Methodology

Participant Motivation. Executing a comparative experiment via the online medium came with several disadvantages that made the subsequent insight analysis relatively complex. The most obvious observation is that the reported insights were recorded in a brief manner (M=17,86 words, SD=11,74, total of 5662 words), so that their subsequent analysis resembled more a semantic analysis exercise, rather than a reproducible grouping of insight characteristics. As a result, the categorization of an insight often depended on a single noun or verb in a very short sentence.

Participant Expectations. We believe that the preconceptions of participants might explain some of the skewed results (e.g. no significant difference in self-reported confidence or difficulty of insights, as well as no significant difference in insight depth reported by experts). This phenomenon is also exemplified by the apparent differences in submissions without any insight (ANA:1, MAG:11, ART:9) and the open-ended feedback. Motivated by the wish to fulfil the study brief as good as possible, most participants were able to overcome the apparent incomprehensibility of the ART demonstrator without any significant effect on the quality of insights, even while spending the least amount of time on the study.

Insight Analysis. No unified methodology seems to exist that is able to capture the rich typology of insights. While Chen et al.'s [5] fact taxonomy still seems useful to categorize analytical insights, it proved too untrustworthy to reach acceptable levels of agreement between individual researchers for any insight that was not clearly analytical. As a result, we recommend two alternatives for comparative insight analysis. First, we propose to consider classifying insights by their analytical value as well as any *meaning* that is derived from it by the user (See 4.4.2) or a combination of both. Accordingly, we have discovered how the artistic style (ART) led to more insights that were based on 'reflecting' on the content, which often were not grounded in a perceived visual pattern. Secondly, we propose to use new methods of insight categorization, such as card sorting or affinity diagramming. These methods proved to be more reliable since they allowed for an iterative process of classification and negotiation. Unfortunately, these methods suffer from a low rate of reproducibility, and must be used within the same comparative study to make meaningful conclusions.

Open-Ended Web Forms. Some participants were inclined to use the insight report text box to complain about apparent usability issues, while others more correctly treated the solving of these issues as insights. This phenomenon is intriguing, in particular when the study was specifically set up to detect usability inefficacies through discovering any detrimental differences in insight characteristics (e.g. a less usable visualization style should lead to less deep insights). While usability reporting was included in the last stage of the study, it proved too late for many participants. Therefore, we propose to provide a separate entry box, dedicated to usability issues, in parallel to the insight report form.

Participant Cohort. The self-reported visualization proficiency of the participants that finished the study was remarkably high (M=3.05, SD=.69). Therefore, some 'expert' bias in terms of preference ratings might have occurred against specific styles (e.g. MAG and ART), as more experienced participants might have expected something more akin to "classical" information visualization tools (i.e. ANA). The more experienced participant cohort might also have been more motivated to perform well in the study, which could have lead to the reporting of more deep insights regardless of the efficiency of the demonstrator. In the extreme, the study might reflect the beliefs of motivated members of the visualization community instead of that of the masses.

Controlling Style. Because of the open-ended nature of the concept of style, and its implications on an overall design concept – which even includes interactive features and non-graphical elements – it is unclear the extent to which particular design decisions, such as the choice of visual metaphor of ART, or the nature of interactive features in MAG and ART, might have affected the measures that were recorded.

5.1.2 Impact of Style on Visualization

We investigated the following hypotheses, among others:

Factual Insight. ANA is better than MAG and ART for identifying *analytical* insights, versus *meaning*-based insights.

User Interaction. Interaction in ANA encourages more *fact*-finding insights, while ART focuses on the exploration of insights that deal with *content*. MAG facilitates both.

Insight Depth. Insights originating from ANA are *deeper* than MAG. Insights from MAG are deeper than ART.

Similarity of Styles. We discovered very few significant differences between the MAG and ART styles. Although we believed that the Bubble Set-inspired scatter plot technique (MAG) was sufficiently different from the use of abstract, overlapping flowers (ART), the two styles performed very similarly for almost all of the study results. Significant differences only appeared when comparing ANA and MAG, and ANA and ART. As a first indication, there seems to be a significantly stronger difference in insight generation between an analytical style and one that has been embellished, and this regardless of the embellishment style.

Visual Quality. No significant difference was found in terms of beauty between the three stylistic approaches. While MAG and ART were being considered more artistic and intended to express, they did not convince as being more beautiful than its analytical style

counterpart (ANA). In fact, a standard scatter plot with default check box buttons was deemed more beautiful than one containing flowers, animation effects and background music. While we wonder whether a within-subject experimental setup would result in similar findings, this phenomenon might be best explained by a flawed design process, in which craftsmanship and the current style zeitgeist could have played a more prominent role.

Insight Depth. In spite of being different in usability as well as interaction styles, there was no difference on the confidence and the difficulty in finding the insights, which is a relatively remarkable finding. The more embellished styles seem to only influence self-reported insight depth, which was however not acknowledged by the post-study rankings of external raters.

Interaction versus Meaning. ANA has lead to significantly more interaction operations, which was geared towards discovering patterns and largely ignored the exploration of content. This might be due to the segregation of visualization and content in the user interface, and the advanced parameters that were made readily available. On the other hand, removing interactive filters and explanatory labeling (i.e. ART) will still engage people to explore the data, but on a different level, forcing them to dive into the content instead of discovering and giving meaning to visual patterns. This study thus shows that by limiting simple interaction capabilities (e.g. mouse hovering to preview a title), people can be steered towards specific insight creation behaviors.

6 Conclusion

This study reported on a range of findings after comparing three different stylistic scatter plot visualizations of the same dataset. Next to interpreting the results in the context of stylistic impact, we also propose a critical reflection of using an online comparative test with insight reports as a feasible evaluation methodology.

The impact of style on usability. Overall, the specific stylistic approach played no significant role, in that few differences were discovered between the usability of the magazine and the artistic style. This stands in contrast to the fact that many usability differences were found between the analytical (non-embellished) style and the others. Although the analytical style required significantly more clicks and more time to create an insight, it was considered significantly more clear, engaging, easy to understand, informative, useful, enjoyable, usable, functional and tool-like than its embellished counterparts. The basic message here is: *do not embellish a visualization with visual or interactive features when usability is an important concern*.

The impact of style on insight depth. In spite of extreme visual difficulties (e.g. no color legend, no axis labels, overlapping elements), the insights from the artistic style were not considered more difficult or less confident in comparison to the other two styles, yet people considered them to be more shallow. However, the post-study blind expert ratings of insights could not acknowledge any significant differences in depth between the three styles. Regardless of their usability performance, all stylistic approaches had the ability to create the same depth, confidence and difficulty of insights. *Users consider the insights from an analytical style as deeper, but still seem to be able to overcome visual and interaction difficulties to discover as deep insights as with an analytical style*.

The impact of style on the kind of insights. We did discover differences in the kind of insights that were generated, in that the more embellished styles lead to more insights that contained some form of reasoning, reflection or interpretation. This is best explained by how their embellishments tended to 'hide' visual patterns, hereby encouraging participants to engage with the content instead. *By limiting simple interaction capabilities (e.g. mouse hovering to preview a title), people can be steered towards specific insight creation behaviors, from higher pattern-seeking level to a more reflective engagement with concepts that underlie the patterns, i.e. content.* The ideal, of course, is where pattern-finding meets the discovery of the principles that drive the patterns.

Developing style demonstrators. While independent participants could fairly accurately recognize and name the styles, our demonstrators might still not be the best representative samples that are possible: it is hard to accomplish a convincing visualization in one particular style, let alone in multiple styles, simultaneously. Hence, the two embellished styles performed too similarly to reach measurable differences in usability and insight categorization. *Designing representations of style might require more craftsmanship and more extensive iterative user testing than our naïve approach of 'skinning' a common scatterplot technique.*

Benchmarking visualizations through analyzing insights. The categorization of insights proved to be a complex and subjective process with a low inter-coder agreement rate, as it resembled more a semantic text analysis. Unfortunately, several methodological circumstances played an influential role in classifying insights in a meaningful way, such as the lack of more descriptive insight recordings, or the motivation of participants to deliver meaningful results even when being offered inefficient tools. It might therefore be useful for future online insight report studies to:

1. Make an explicit distinction between the *analytical characteristics of an insight and its meaning (e.g. reflection, interpretation)* that provides its context. Notably, some reported insights were not grounded in the perception of any graphical stimulus, while others were so intrinsically related to a complex visual pattern that the coding of the insight became difficult.
2. Explicitly request the *reporting of meaning*, next to the description of the factual or analytical basis of an insight.
3. Motivate participants to *report their insights in a more expansive way*, potentially even encouraging them to *categorize* their own insights (similar to a heuristic evaluation method).
4. Allow the *reporting of usability issues in parallel with the insight reporting*, and make them equally important, in order to avoid that participants treat the reporting of issues as insights and to limit participant frustration when being confronted with less efficient visualization techniques.
5. Consider *richer analysis options for insight categorization*, such as the methods of open coding, card sorting or affinity diagramming, in order to discover more intrinsic and tacit differences in the kinds of insights that were reported.

Naturally, an alternative approach is to conduct insight report studies in a controlled lab environment, which allows participants to report more expansive insights via the talk-aloud methods (e.g. [19]), and gives researchers the chance to provide more precise instructions or request clarifications where needed.

While our results were not as clear-cut as originally expected, we consider this study as a crucial first step towards a better understanding of the impact of style in information visualization in the online medium. We hope our findings, in addition to the provided methodological tips and guidelines, will benefit the future evaluation of visualization techniques that aim beyond measuring commonly agreed usability metrics.

References

[1] S. Bateman, R. Mandryk, C. Gutwin, A. Genest, D. McDine and C. Brooks, "Useful Junk? The Effects of Visual Embellishment on Comprehension and Memorability of Charts," *Conference on Human Factors in Computing Systems (CHI'10)*, ACM, 2010, pp. 2573-2582.

[2] Bestiario, "reMap: a project by Bestiario based on visualcomplexity.com," 2009; http://www.bestiario.org/research/remap/.

[3] N. Cawthon and A. Vande Moere, "The Effect of Aesthetic on the Usability of Data Visualization," *IEEE Conference on Information Visualization (IV'07)*, IEEE, 2007, pp. 637-648.

[4] R. Chang, C. Ziemkiewicz and T.M.R. Green, W., "Defining Insight for Visual Analytics," *IEEE Computer Graphics and Applications* vol. 29, no. 2, 2009, pp. 14-17.

[5] Y. Chen, J. Yang and W. Ribarsky, "Toward Effective Insight Management in Visual Analytics Systems," *IEEE Pacific Visualization Symposium (PacificVis'09)*, IEEE, 2009, pp. 49-56.

[6] C. Collins, G. Penn and S. Carpendale, "Bubble Sets: Revealing Set Relations with Isocontours over Existing Visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, no. 6, 2009, pp. 1009-1016.
[7] Stamen Design, "Digg Labs," 2007; http://labs.digg.com.
[8] N. Diakopoulos, F. Kivran-Swaine and M. Naaman, "Playable Data: Characterizing the Design Space of Game-y Infographics," *Conference on Human Factors in Computing Systems (CHI'11)* ACM, 2011.
[9] J. Harris and S. Kamvar, "We Feel Fine - An Exploration of Human Emotion, in Six Movements," 2006; http://www.wefeelfine.org/.
[10] P. Isenberg, T. Zuk, C. Collins and S. Carpendale, "Grounded Evaluation of Information Visualizations," *Conference on BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV '08)*, ACM, 2008.
[11] M. Jern, "Collaborative Web-Enabled Geoanalytics Applied to OECD Regional Data," *Cooperative Design, Visualization, and Engineering (CDVE'09)*5738/2009, Springer, 2009, pp. 379-394.
[12] R. Kosara, "Visualization Criticism - The Missing Link Between Information Visualization and Art," *Conference on Information Visualization (IV'07)*, IEEE, 2007, pp. 631-636.
[13] M. Kurosu and K. Kashimura, "Apparent Usability vs. Inherent Usability: Experimental Analysis on the Determinants of the Apparent Usability," *Conference on Human factors in Computing Systems (CHI'95)*, ACM, 1995, pp. 292-293.
[14] UC Berkeley Visualization Lab, "Flare - Data Visualization for the Web," 2008; http://flare.prefuse.org/.
[15] A. Lau and A. Vande Moere, "Towards a Model of Information Aesthetic Visualization," *International Conference on Information Visualisation (IV'07)*, IEEE, 2007, pp. 87-92.
[16] B. Mueller, "Poetry on the Road," 2004; http://www.esono.com/boris/projects/poetry04/.
[17] D. Norman, "Emotion & Design: Attractive Things Work Better," *interactions*, vol. 9, no. 4, 2002, pp. 36-42.
[18] C. North, "Toward Measuring Visualization Insight," *IEEE Computer Graphics*, vol. 26, no. 3, 2006, pp. 6-9.
[19] C. North, P. Saraiya and K. Duca, "A Comparison of Benchmark Task and Insight Evaluation Methods for Information Visualization," *Information Visualization*, vol. 10, no. 3, 2011, pp. 162-181.
[20] V. Postrel, *The Substance of Style: How the Rise of Aesthetic Value is Remaking Commerce, Culture and Consciousness*, Harper Collins, 2003.
[21] Z. Pousman, J. Stasko and M. Mateas, "Casual Information Visualization: Depictions of Data in Everyday Life," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, 2007, pp. 1145-1152.
[22] B. Reavis, "Bitalizer - Bending Bits into Structure," 2008; http://thirdroute.com/projects/bitalizer/.
[23] M. Rester, M. Pohl, S. Wiltner, K. Hinum, S. Miksch, C. Popow and S. Ohmann, "Evaluating an Infovis Technique using Insight Reports," *International Conference on Information Visualization (IV'07)*, IEEE, 2007, pp. 693-700.
[24] C. Riekoff, "Texone Code Tree," 2005; http://texone.org.
[25] H. Rosling, "Visual Technology Unveils the Beauty of Statistics and Swaps Policy from Dissemination to Access," *Journal of the International Association for Official Statistics*, vol. 24, no. 1-2, 2007, pp. 103-104.
[26] C. Salimun, H. Purchase, D.R. Simmons and S. Brewster, "The Effect of Aesthetically Pleasing Composition on Visual Search Performance," *Nordic Conference on Human-Computer Interaction (NordiCHI'10)*, ACM, 2010, pp. 422-431.
[27] The New York Times, "Times Developer Network," 2009; http://developer.nytimes.com/docs/article_search_api/.
[28] N. Tractinsky, "Towards the Study of Aesthetics in Information Technology," *Conference on Information Systems*, 2004, pp. 771-780.
[29] N. Tractinsky, S.-K. A. and D. Ikar, "What is Beautiful is Usable " *Interacting with Computers*, vol. 13, no. 2, 2000, pp. 127-145.
[30] N. Tractinsky and J. Meyer, "Chartjunk or Goldgraph? Effects of Presentation Objectives and Content Desirability on Information Presentation," *MIS Quarterly*, vol. 23, no. 3, 1999, pp. 397-420.
[31] F. Viégas, M. Wattenberg, F. van Ham, J. Kriss and M. McKeon, "Many Eyes: a Site for Visualization at Internet Scale," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, 2007, pp. 1121-1128.
[32] F. Viégas and M. Wattenberg, "Artistic Data Visualization: Beyond Visual Analytic," *Lecture Notes in Computer Science*, vol. 4564, no. 15, 2007, pp. 182-191.
[33] J.S. Yi, Y.-a. Kang, J.T. Stasko and J.A. Jacko, "Understanding and Characterizing Insights: How Do People Gain Insights using Information Visualization?," *Conference on BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV '08)*, ACM, 2008, pp. 1-6.