Untangling High-Dimensional Data in Web Design

Victoria Flores, Maxine Lim, and Cesar Torres

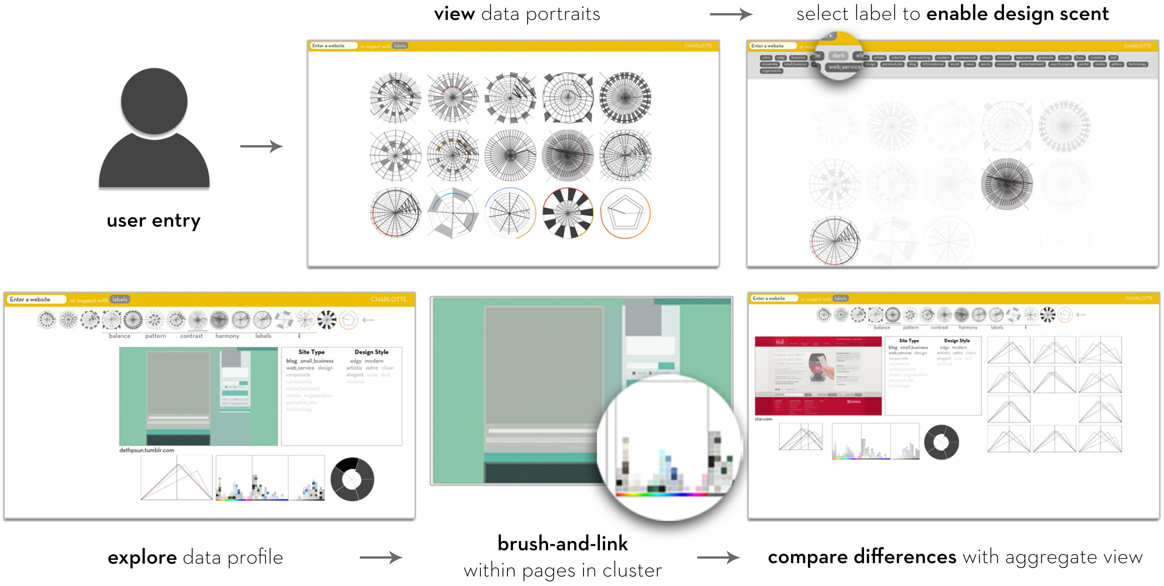

Fig. 1. The exploratory flow for Charlotte. User begins by viewing data portraits of Web page clusters, inspecting them visually, by URL, or page labels. The next level uses a data profile to allow for deeper examination of a grouping. The profile consists of visualizations capturing design principles across all the pages in the cluster. Brushing and linking allows for individual data points to be put in context, and the grouping itself can be seen in context using small multiples.

Abstract— Exploring a design space must generally be done by manual inspection of many design examples. Visualizing design data in an aggregate way can make this process more efficient, but as design data lies in a high-dimensional feature space, selecting the important elements to view is challenging. This work presents Charlotte, a system that enables exploration of Web designs represented in 1,713 dimensions by applying the concept of data portraits and generating visualizations that capture groups of pages with respect to a selected set of design principles. Charlotte demonstrates that meaningful patterns and trends among the design data can explored by using these principles to inform data-driven portraits.

Index Terms—high-dimensional data, design

1 Introduction

Investigating current design practices can make designers more visually fluent, allowing them to draw from examples and learn from errors. But getting a sense for what a design space looks like is nontrivial and may require years of experience. To explore a design space, people usually must browse through many individual examples, which after an extended period of time might only compose a small fraction of existing work. The barrier to comprehensively characterizing a design space is high, which may prohibit many from attempting it. What if there was a way to browse design in aggregate such that hundreds of examples could be inspected at once?

Related work has used data portraits to create compact visualizations of complex subjects, namely people. We applied this concept to the field of Web design, enabling aggregate visualization of design with respect to a set of design principles. However, the problem with aggregating information for complex subjects, whose data may lie in thousand-dimensional feature spaces, is selecting which features to visualize.

This paper's primary contribution is Charlotte, a system that uses data portraits and aggregate visualizations to allow exploration of complex, high-dimensional subjects, in this case, Web designs. Charlotte enables aggregate visualization of Web design with respect to a selection of design principles. We enable exploration on the level of page groupings using a Web abstraction and multiple visualizations of design elements, as well as on the level of individual pages. The views and interactions that we implement allow users to form hypotheses about the design space, then verify these hypotheses by finer-grained inspection of pages.

In the following sections we first describe related work on which our system builds and the features of our data set. We proceed to outline the methods we used to form high-level groupings and select and encode design principles, as well as the implementation of our system. Finally we present some findings from our own investigations

- *Victoria Flores is with Stanford University, email: vflores3@stanford.edu.*
- *Maxine Lim is with Stanford University, email: maxinel@stanford.edu.*
- *Cesar Torres is with Stanford University, email: ctorres7@stanford.edu.*

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 e-mail to: tvcg@computer.org.

using Charlotte and discuss future work.

2 Related Work

This system builds off of related work on high-dimensional feature spaces, clustering data, and data portraits. With the increasing amount of data being generated, working with high-dimensional feature spaces has become more relevant than ever. From principal component analysis to k-means clustering, many methods are currently being explored to help make sense of large data sets [6]. Instead of using advanced techniques, we begin by creating informal groupings of our data. We then leverage information unique to Web design, namely URL namespace, descriptive labels, and design principles, to inform explorations.

Previous work has also shown how clustering objects using different visualization techniques can produce different results [5]. Although we realize our grouping method, if uninformed, may limit subsequent explorations, we see our work as a first step in investigating how to effectively group Web designs.

Our work also draw from research on data portraits, which represent people in online communities using the data they generate [1]. Data portraits are based on the notion that users can better interact with users if they have easy access to information about them. They represent subjects using metaphors that encode important data. For example in an online forum, a person may be depicted as a flower whose number of petals represents his amount of recent activity [9]. This approach, a hybrid of subjective art and objective information, aims to succinctly communicate data in a way that is easy for humans to interpret and use for comparison.

Our system adapts the data portrait concept for Web design by representing Web pages using a spider web metaphor. We use them to provide an approximate characterization of Web designs for a high-level comparison. In contrast to previous work we encode significantly more information in our data portraits, though the precision to which the data is encoded is approximate rather than exact.

3 Data Set

Our data set consists of 1,713-dimensional feature vectors and two classes of crowdsourced labels for 3,218 pages. It is drawn from the Webzeitgeist, a platform for machine learning on Web design that provides a repository of Web pages along with their computed visual segmentations, feature vectors, and crowdsourced labels [7]. The visual segmentations restructure the underlying DOM (Document Object Model) of the Web page such that each element is a visually salient region of the page. The feature vectors capture DOM-related and computer vision features for each Web page. DOM features include information such as color, size, tag name, etc., while computer vision features include those from the Gist graphics library. For each page, we also have a set of crowdsourced labels describing its site type (such as “blog or “news) and design style (such as “minimal or “colorful).

4 Method

We built an interactive system to exploring Web design leveraging a rich data set. To allow for high-level explorations, we group Web pages into clusters based on their raw feature vectors. We enable comparisons using data portraits and querying by URL and page labels. To enable deeper comparisons among clusters, we create data profiles for each one, summarizing aggregate statistics drawn from design principles and page labels. Interactions that allow comparisons among clusters and inspection of individual pages empowers users to observe and infer design patterns and trends. This exploratory flow is shown in Figure .

4.1 Clustering Web Pages

Navigating through a set of thousands of pages can be overwhelming, so before enabling in-depth exploration, we first grouped pages by similarity. We applied an unsupervised machine learning technique called K-means clustering to our raw feature vectors, forming fifteen distinct clusters of pages.

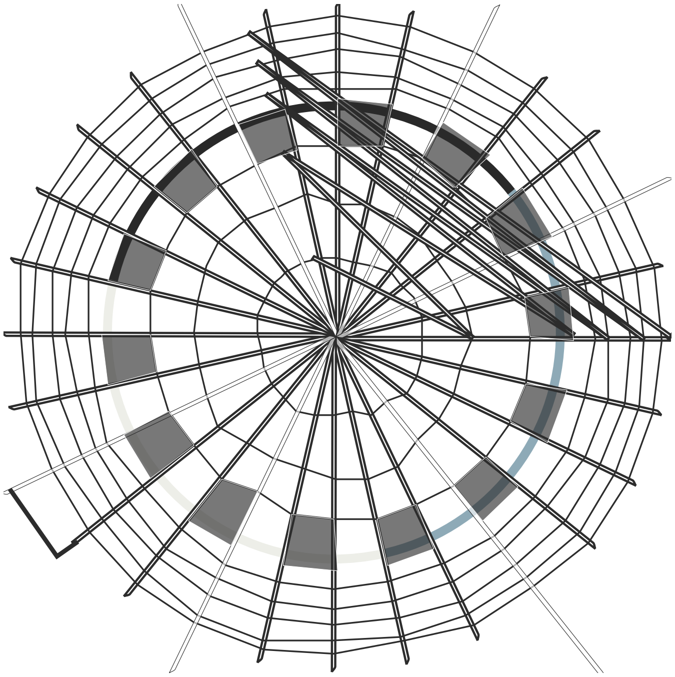

Fig. 2. A data portrait of a group of Web pages that encodes design principles such as balance and harmony.

4.1.1 Webs as Data Portraits

To visualize the distinguishing features of each cluster, we created data portraits for them using a spider “web” abstraction. Based on the notion that a data portrait can efficiently encode large amounts of disparate data, we used Adobe Illustrator to hand-craft web representations of the clusters using seven principle encodings: balance, color harmony, contrast, emphasis, movement, pattern, and rhythm. A subset of these encodings were further used to develop visualizations for the clusters’ data profiles. The web concept is fitting because it invokes the complexity and richness reflected in our data.

To encode these principles into the webs, we computed aggregate statistics for the target cluster using features such as spatial distribution of page elements, dominant color, largest leaf element or average number of siblings in the DOM tree, etc. We then translated these statistics to correspond to visual features on the webs. For example balance was represented by the proximity of radial rings on the web. More rings closer to the center indicates a page that is skewed to the right. Color is encoded simply by applying the three dominant colors to a ring on the web. Contrast is separately encoded using opacity of a checkerboard pattern in one ring of the web. For example, higher contrast will produce a black and white pattern, while lower contrast tends toward greys. Emphasis is represented by a bolder ring, whose radius represents the amount of emphasis on a page. Movement is encoded using a zig-zagging path across the web. Pattern is represented by the number of elements on a page; more complex pages will have more cells in the web. Rhythm represented using a series of extended web strings. Besides drawing attention, these webs enable high-level comparisons of clusters from a “zoomed-out” view. Figure 2 shows an example of one of the webs.

4.1.2 Scenting by URL and Labels

Information scenting by URL and page labels allowed these high-level groupings to serve as an entry point to deeper investigation. While our data portraits sketch the characteristics of each cluster of pages, selecting which group to explore based on abstracted principles may be challenging. To encourage exploration of individual clusters, we allowed users to identify clusters containing a particular URL. Webs containing the URL remain, and others fade away. We also allow clusters to be identified by site type and style labels. Upon selection of a label such as ”dark,” clusters containing pages labeled ”dark” remain opaque. The clusters’ opacities are adjusted according to how many relevant pages they contain, with the darkest webs indicating clusters with the highest number of relevant pages. Scenting allows users to associate clusters with more familiar characteristics, guiding them to more explore more deeply.

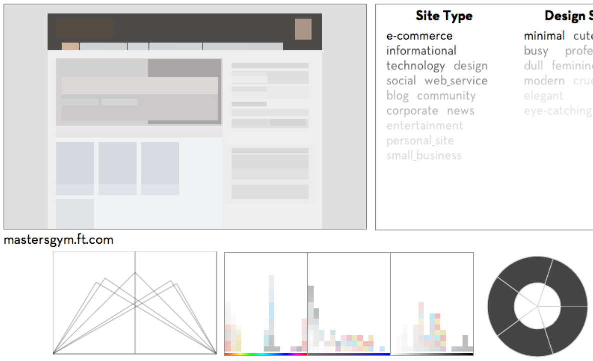

Fig. 3. An example of a data profile for one group of pages.

4.2 Constructing a Data Profile

For every cluster, we constructed a data profile to visualize all its pages with respect to different design principles. There are many important principles, and there may be multiple ways of quantifying each one. In this work we do not claim to define the best set of principles nor create optimal metrics for assessing them. Instead, we aim to select a set of principles that may be of interest to designers, vary across different Web designs, and can be approximated using our feature set. The aspects we chose to visualize are derived from a set of commonly accepted design principles and feedback from students in our visualization class [4, 8]. They include balance, movement, contrast, harmony, and unity. An example of a data profile is shown in Figure 3

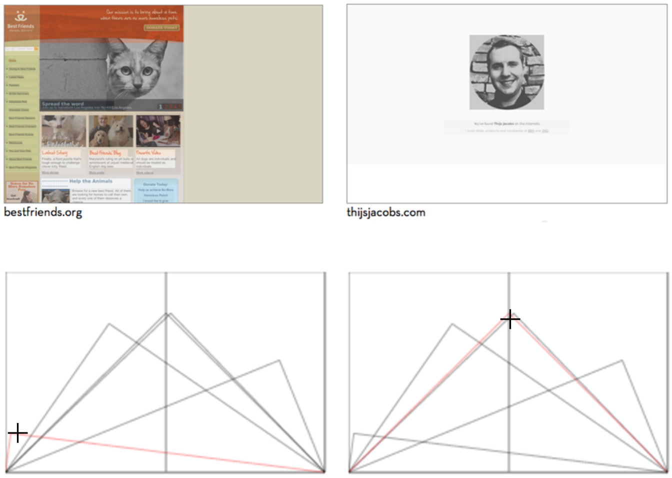

Fig. 4. The left-most point (left) corresponds to a page with its elements heavily concentrated on the left side of the page, while the center point (right) refers to a symmetric page.

4.2.1 Balance

We computed the balance metric for each Web page by summing up the vertical edges of all elements that fell on the left and right sides of the page. The resulting ratios were then normalized with respect to page area to enable comparisons among different pages. A straightforward way to view these ratios might be to plot them in a histogram. However, balance is about finding the center, where structural forces meet [2]. By creating a point whose angle reflects this ratio, our skewed lines representation helps identify a center of balance and gives a sense of in what direction and to what extent the page is skewed. Figure 4 shows how this method contrasts the balance for a left-oriented page compared to one that is perfectly symmetric.

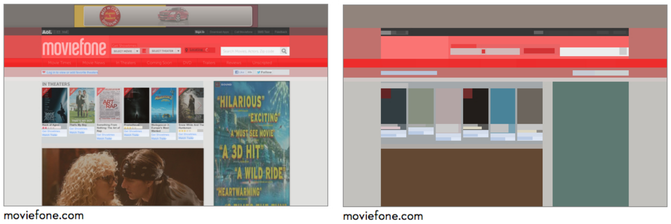

Fig. 5. Page screenshot (left) and composition representation with text and images removed (right).

4.2.2 Movement

To view movement, the path that the viewer’s eye follows on the design, we show an alternate view of page layout as shown in Figure 5. Images and text can distract from the underlying layout of the page, so site content was omitted to create emphasis on how elements were arranged on a page. For the profile of each page cluster, we used the layout of its centroid, which can be considered a representative example of it.

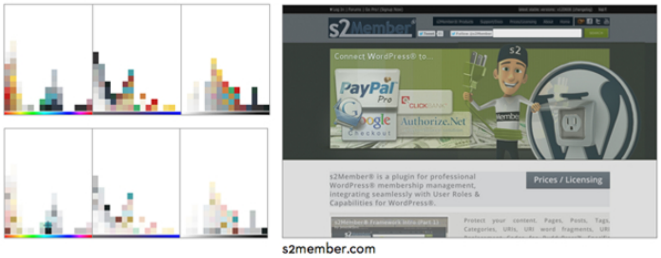

Fig. 6. Histograms for hue, saturation, and lightness (left). Hovering over a swatch highlights the color palette for the corresponding page (right).

4.2.3 Contrast and Harmony

Contrast and harmony were captured using page color. For every page in the cluster we selected a palette of the dominant color along with nine others generated by median cut quantization [3]. Swatches for each color were plotted in histograms and grouped by hue, saturation, and lightness as shown in Figure 6. These histograms show the distribution of colors in this cluster across these three measures.

4.2.4 Unity

We used negative space to loosely approximate unity by reasoning that pages with more negative space would have more distinct groupings of elements and thus be more unified. To compute negative space, we took the background color of each page and computed the area of foreground elements based on a certain threshold. We visualized the resulting ratios using the ratio of white to black area in a segmented pie chart as shown in Figure 7. In contrast with other graph types, this form gives a a better sense for the aggregate negative space ratio for the entire cluster.

4.2.5 Labels

Although site type and style labels are not design principles, these labels are useful because they allow users to associate more explicit descriptors with the Web pages they are viewing. Therefore for each

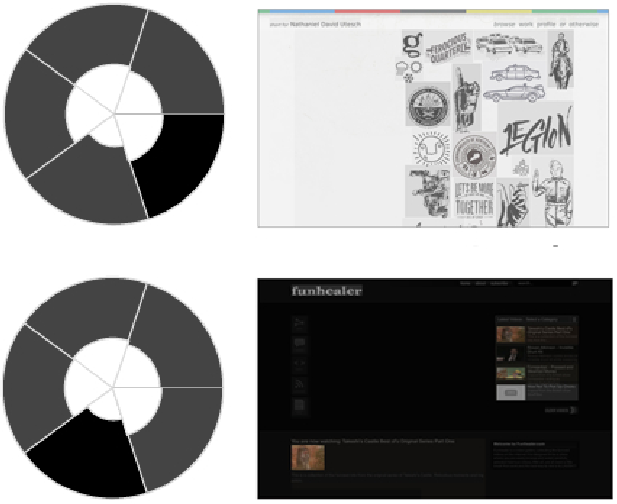

Fig. 7. Negative space charts. Hovering over a swatch (left) highlights the color palette for the corresponding page (right). The bottom right wedge on the wheel corresponds to a page with a moderate amount of positive space (top), while the bottom left wedge refers to one with very little negative space (bottom). These ratios are reflected in the respective wedges.

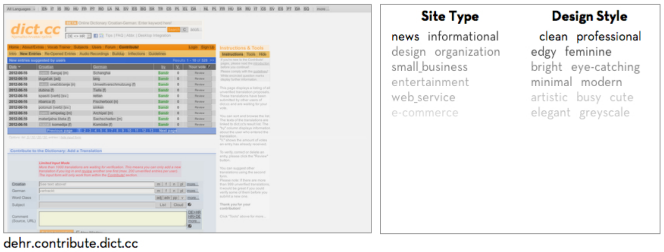

Fig. 8. Site type and style labels for a given page. Frequency is encoded in the label's opacity.

cluster we also show the labels applied to pages in that grouping as simple text lists as in Figure 8. The opacity of each label encodes their relative frequency.

4.3 Interactions

We implemented several interaction techniques that allowed users to identify design patterns and more closely examine Web design characteristics. The initial view for an individual cluster consists mainly of the data profile for that cluster, displaying all the design principles in their own visualizations. Brushing over any data point in each visualization highlights the corresponding point in the others. Clicking on a data point shows the screenshot of the page in the central area.

To compare principles across clusters, users can click on a principle of interest, e.g., balance. Small multiples of the visualizations appear on the right of the data profile, allowing users to see how the current cluster compares to others in the design space. Hovering over any of the multiples highlights the corresponding web beneath the toolbar. Some of these interactions are depicted in Figure .

4.4 Implementation

The Charlotte system is constructed using the Rails Web application framework and is comprised of four components: a data preprocessor, a SQLite database, an application controller, and a data cacher. A diagram is shown in 9.

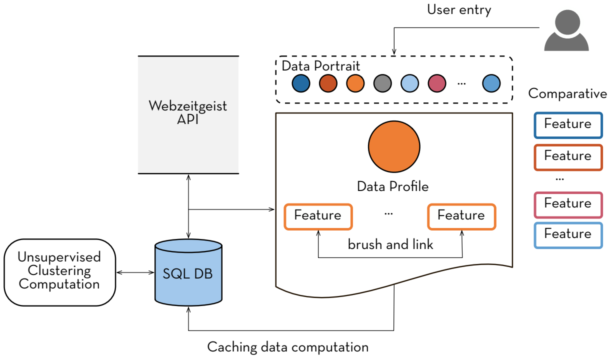

Fig. 9. The architecture of the Charlotte system. The database draws from Webzeitgeist and results from a Matlab process to provide relevant information to the application.

4.4.1 Data Preparation

The subset of pages used in our system from the Webzeitgeist repository contains over 3,000 Web pages and over 300,000 visual blocks. In order to utilize this data and scale to support the full corpus of 100,000 pages, we used the Webzeitgeist API, which provides access to page level queries (screenshot, visual segmentation, page feature vectors) and general batch queries (page identifiers). We selected a subset of the pages that had either of two types of crowdsourced labels: site type and style.

4.4.2 Preprocessor

The preprocessor seeds the database with the initial subset of labeled pages, label frequencies, and the mappings of each URI to a host. A Matlab process runs in the background and computes the k-means clusters based on this subset of pages. The cluster assignments and centroids are stored in a SQLite table.

4.4.3 Application controller

To handle the outward facing application, the application controller handles the primary data transfer of cluster information. All data calls thereafter are handled by the data cacher.

4.4.4 Data Cacher

Since we are calculating aggregate statistics for a group of pages, the amount of client-side processing is computationally intensive and does not scale to large data sets. In order to alleviate this problem, one common solution is to compute these features on the server-side as a preprocess. However, process becomes time-intensive. Our data cacher instead distributes the calculation of aggregate statistics in smaller chunks. To form the chunks it stochastically selects a subset of x uncomputed nodes and n computed nodes where x is derived experimentally based on client-side computational constraints. These x nodes are then sent to the client to compute and subsequently added to the database for faster access. This pipeline enabled finding the clusters' aggregate statistic without a costly preprocess.

5 Evaluation

To evaluate our system, we discuss design patterns and trends identified using our tool.

5.1 Findings

For high-level exploration of the design space, we noted some relationships labels and clusters. For example, we saw that clusters prominent with respect to the "dark" style label were almost exclusive from the groups for "cute" This exclusivity indicates that ”dark" and ”cute" pages may lie in separate sets. In addition, inspecting the cluster that remains opaque for both of these labels may show in the intersection

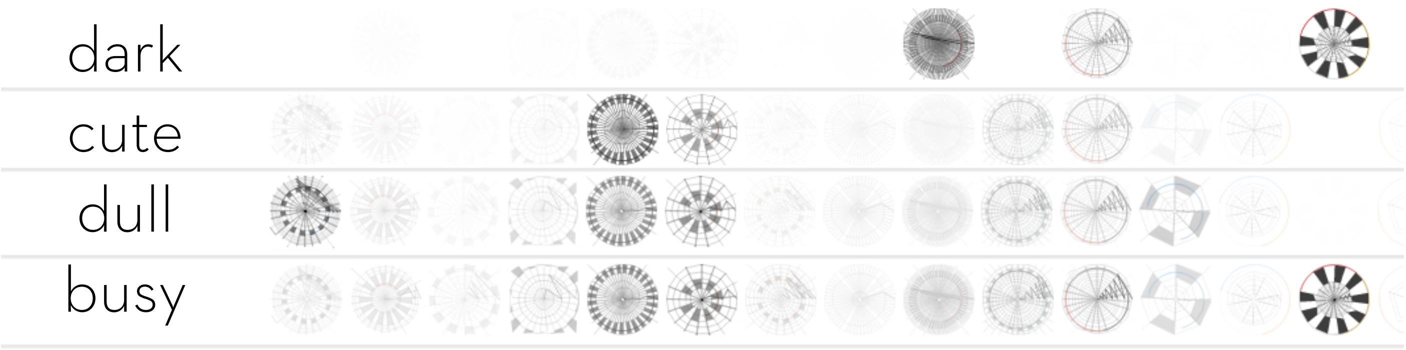

Fig. 10. Caption.

of these spaces. In contrast, we see that "cute," "dull," and "busy" all highlight a similar set of webs, suggesting that there are many pages that have a combination of such labels. However outliers that are illuminated strongly for "dull" and "busy" but not the rest shows the presence of page groupings that are unique with respect to these labelings. Figure 10 depicts these findings.

A similar finding can be observed with a combination of URLs and site type labels. For example the set of clusters prominent when searching for "nytimes" is almost the same as the set for "news," suggesting that the Web site for the New York Times is a typical example of a news site.

These high-level explorations can be used to form hypotheses about the exclusivity and similarities of Web sites for target URLs and page labels. It can also help identify outlier groups and illustrate where individual Web sites lie in the design space. Allowing for users to form hypothesis at this zoomed-out view of the design space can provide a starting point for more in-depth exploration.

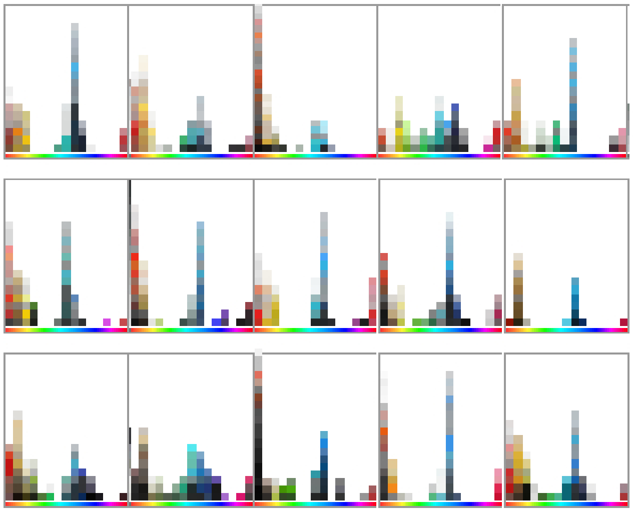

Fig. 11. Small multiples of color histograms, allowing for comparison color distributions among different clusters.

At the data profile level, users can use the aggregate data portraits to get a sense for the characteristics of pages in a cluster. Clusters might be distinguished by an acute lack of symmetry, an unusual color distribution, or a high amount of negative space. Users can then explore how a given cluster compares to others via the expanded small multiple view.

Using the small multiple view shown in Figure 11, we can inspect the color distributions across all clusters with respect to hue, saturation, and lightness. In inspecting hue, page colors tend to be heavily concentrated in the red and blue area of the spectrum, with the yellow area relatively neglected. We can also notice that page colors tend to be relatively unsaturated, with the high saturation bins generally empty. Designers can leverage this information to either conform or rebel against current trends in Web design. For example in selecting a color for Charlotte's toolbar, we decided to use the underrepresented color yellow in order to stand out from all the blue- and red-based Web sites.

6 Future Work

There are a number of ways that extending our system could allow for richer exploration. First, an extended feature set could give us more data to visualize. Second, more visualizations for different design principles and elements could be implemented. Third, allowing for better scalability of our visualizations and interaction techniques would help users explore more deeply. Enabling interactions that allow users to track individual designs or select a subset of clusters or pages to focus on might help people use this system as a design tool. Another useful evaluation would be to test this system by running a user study. Given time constraints we did not have the means to conduct one in a principled way, but gathering user feedback and observing how people use the system could provide valuable insights on its strengths and weaknesses. Finally we could work to apply our system methodology to other domains with high-dimensional subjects, such as biology.

7 Conclusion

In this paper we presented Charlotte, a system for visualizing high dimensional data at scale using the concept of aggregate data portraits based on important design principles. We described Charlotte's features and interactions as well as provided examples of exploratory workflows in which users can engage to discover trends in the data. With the ever-growing repository of data on the Web, leveraging data in different forms, such as in art or design, has the potential to provide powerful design tools. Addressing the high-dimensional representations of designs will be a nontrivial obstacle.

Acknowledgments

The authors wish to thank the CS448B staff for their helpful input, as well as the students in the class who gave feedback on this project.

References

[1] *Data portraits*, 2010.
[2] R. Arnheim. *Art and visual perception: a psychology of the creative eye*. University of California Press, 1954.
[3] D. Bloomberg. Color quantization. 2008.
[4] S. Bradley. The 7 components of design. December 2012.
[5] M. A. Hearst, D. R. Karger, and J. O. Pedersen. Scatter/gather as a tool for the navigation of retrieval results. In *AAAI Fall Symposium on AI Applications in Knowledge Navigation and Retrieval*, 1995.

[6] H.-P. Kriegel, P. Kroger, and A. Zimek. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, 2009.

[7] R. Kumar, A. Satyanarayan, C. Torres, M. Lim, S. Ahmad, S. R. Klemmer, and J. O. Talton. Webzeitgeist: Design mining the web. In *CHI '13: Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2013.

[8] J. Lovett. The principles of design. 1999.

[9] R. Xiong and J. Donath. Peoplegarden: creating data portraits for users. In *Proceedings of the 12th annual ACM symposium on User interface software and technology*, 1999.