Estimating Color-Concept Associations from Image Statistics

Ragini Rathore, Zachary Leggon, Laurent Lessard, and Karen B. Schloss

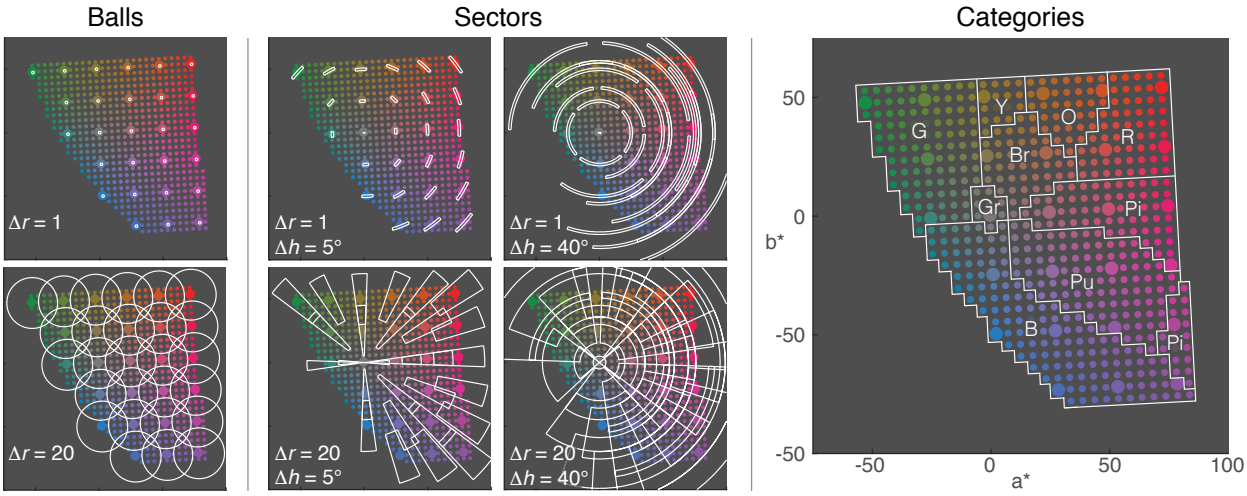

Figure 1. We constructed models that estimate human color-concept associations using color distributions extracted from images of relevant concepts. We compared methods for extracting color distributions by defining different kinds of color tolerance regions (white outlines) around each target color (regularly spaced large dots) in CIELAB space. Subplots show a planar view of CIELAB space at $L^* = 50$, with color tolerance regions defined as balls (left column; radius Δr), cylindrical sectors (middle columns; radius Δr and hue angle Δh), and category boundaries around each target color (right column; **R**ed, **O**range, **Y**ellow, **G**reen, **B**lue, **Pu**rple, **Pi**nk, **Br**own, **Gr**ay; white and black not shown). Each target color is counted as "present" in the image each time any color in its tolerance region is observed. This has a smoothing effect, which enables the inclusion of colors that are not present in the image but similar to colors that are. A model that includes two sector features and a category feature best approximated human color-concept associations for unseen concepts and images (see text for details).

Abstract—To interpret the meanings of colors in visualizations of categorical information, people must determine how distinct colors correspond to different concepts. This process is easier when assignments between colors and concepts in visualizations match people's expectations, making color palettes *semantically interpretable*. Efforts have been underway to optimize color palette design for semantic interpretablity, but this requires having good estimates of human color-concept associations. Obtaining these data from humans is costly, which motivates the need for automated methods. We developed and evaluated a new method for automatically estimating color-concept associations in a way that strongly correlates with human ratings. Building on prior studies using Google Images, our approach operates directly on Google Image search results without the need for humans in the loop. Specifically, we evaluated several methods for extracting raw pixel content of the images in order to best estimate color-concept associations obtained from human ratings. The most effective method extracted colors using a combination of cylindrical sectors and color categories in color space. We demonstrate that our approach can accurately estimate average human color-concept associations for different fruits using only a small set of images. The approach also generalizes moderately well to more complicated recycling-related concepts of objects that can appear in any color.

Index Terms—Visual Reasoning, Visual Communication, Visual Encoding, Color Perception, Color Cognition, Color Categories

1 Introduction

In visualizations of categorical information (e.g., graphs, maps, and diagrams), designers encode categories using visual properties (e.g., colors, sizes, shapes, and textures) [6]. Color is especially useful for encoding categories for two main reasons. First, cognitive representations of color have strong categorical structure [5, 38, 52], which naturally maps to categories of data [8, 15]. Second, people have rich semantic associations with colors called *color-concept associations* (e.g., a particular red associated with strawberries, roses, and anger) [17, 30, 33], which they use to interpret meanings of colors in visualizations [22, 23, 40, 41, 43]. Indeed, it is easier to interpret visualizations if semantic encoding between colors and concepts (referred to as *color-concept assignments)* match people's expectations derived from their color-concept associations [23, 40, 41].

Recent research has investigated how to optimize color palette design to produce color-concept assignments that are easy to interpret [23, 41, 43]. Methods typically involve quantifying associations between each color and concept of interest, and then using those data to calculate optimal color-concept assignments for the visualization [4, 23, 41, 43]. It may seem that the best approach would be to assign concepts to their most strongly associated colors, but that is not always the case.

- *Ragini Rathore, Computer Sciences and Wisconsin Institute for Discovery (WID), University of Wisconsin–Madison, Email: rrathore3@wisc.edu.*
- *Zachary Leggon, Biology and WID, University of Wisconisn–Madison, Email: zleggon@wisc.edu.*
- *Laurent Lessard, Electrical and Computer Engineering and WID, University of Wisconsin–Madison. Email: laurent.lessard@wisc.edu.*
- *Karen B. Schloss, Psychology and WID, University of Wisconsin–Madison. Email: kschloss@wisc.edu.*

Manuscript received 31 Mar. 2019; accepted 1 Aug. 2019.
Date of publication 16 Aug. 2019; date of current version 20 Oct. 2019.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2019.2934536

Sometimes, it is better to assign concepts to weakly associated colors to avoid confusions that can arise when multiple concepts are associated with similar colors (see Section 2.2) [41]. Thus, to leverage these optimization methods for visualization design, it is necessary to have an effective and efficient way to quantify human color-concept associations over a large range of colors. Only knowing the top, or even the top few strongest associated colors with each concept may provide insufficient data for optimal assignment.

One way to quantify human color-concept associations is with human judgments, but collecting such data requires time and effort. A more efficient alternative is to automatically estimate color-concept associations using image or language databases. Previous studies laid important groundwork for how to do so as part of end-to-end methods for palette design [14, 18, 24, 25, 43]. However, without directly comparing estimated color-concept associations to human judgments, it is unclear how well they match. Further, questions remain about how best to extract information from these databases to match human judgments.

The goal of our study was to understand how to effectively and efficiently estimate color-concept associations that match human judgments. These estimates can serve as input for palette design, both for creating visualizations and for creating stimuli to use for visual reasoning studies on how people interpret visualizations.

Contributions. Our main contribution is a new method for automatically estimating color-concept associations in a way that strongly correlates with human ratings. Our method operates directly on Google Image search results, without the need for humans in the loop. Creating an accurate model requires fine-tuning the way in which color information is extracted from images. We found that color extraction was most effective when it used features aligned with perceptual dimensions in color space and cognitive representations of color categories.

To test the different extraction methods, we used a systematic approach starting with simple geometry in color space and building toward methods more grounded in color perception and cognition. We used cross-validation with a set of human ratings to train and validate the model. Once generated, the model can be used to estimate associations for concepts and colors not seen in the training process without requiring additional human data. We demonstrated the effectiveness of this process by training the model using human association ratings between 12 fruit concepts and 58 colors, and testing it on a dataset of 6 recycling-themed concepts and 37 different colors.

2 Related Work

Several factors are relevant when designing color palettes for visualizing categorical information. First and foremost, colors that represent different categories must appear different; *perceptually discriminable* [8, 15, 45, 46]. Other considerations include selecting colors that have distinct names [16], are aesthetically preferable [12], or evoke particular emotions [4]. Most relevant to the present work, it is desirable to select *semantically interpretable* color palettes to help people interpret the meanings of colors in visualizations [23, 41, 43]. This can be achieved by selecting "semantically resonant" colors, which are colors that evoke particular concepts [23]. It can also be achieved when only a subset of the colors are semantically resonant if conditions support people's ability to infer the other assignments [41], see Section 2.2.

2.1 Creating semantically interpretable color palettes

Many approaches exist for creating semantically interpretable color palettes [23, 41, 43], but they generally involve the same two stages:

1. Quantifying color-concept associations.
2. Assigning colors to concepts in visualizations, using the color-concept associations from stage 1.

Assigning colors to concepts in visualizations (stage 2) relies on input from color-concept associations (stage 1), which suggests assignments are only as good as the association data used to generate them. Color-concept association data are good when they match human judgments.

2.1.1 Quantifying color-concept associations

A direct way to quantify human color-concept associations is with humans judgments. Methods include having participants rate association strengths between colors and concepts [30, 41], select colors that best fit concepts [9, 31, 53], or name concepts associated with colors [29, 33]. However, collecting these data is time- and resource-intensive.

An alternative approach is to estimate human color-concept associations from large databases, such as tagged images [4, 23–25, 43], color naming data sets [14, 24, 43], semantic networks [14], and natural language corpora [43]. Each type of database enables linking colors with concepts but has strengths and weaknesses, so automated methods often combine information from multiple databases [14, 24, 25, 43].

Extracting colors from tagged images (e.g., Flickr, Google Images) provides detailed color information for a given concept because of the large range of colors within images [4, 23–25, 43]. Methods must specify how to represent colors from images and what parts of the image to include. Linder et al. [24, 25] obtained a color histogram with bin sizes of $15 \times 15 \times 15$ units in CIELAB space from all pixels in an image. Similarly, Lin et al. [23] calculated color histograms using smaller $5 \times 5 \times 5$ bins, together with heuristics to smooth the histogram and remove black or white backgrounds. In a different approach, Setlur and Stone [43] and Bartram et al. [4] used clustering algorithms to determine dominant colors in images. Clustering can be effective for finding the top colors in an image, but does not provide information on the full color range.

Image extraction methods must also specify how to query image databases to obtain images for each concept. Lin et al. [23] used queries of each concept word alone (e.g., "apple") and queries with "clipart" appended to the concept word (e.g.,"apple clipart"). They reasoned that for some concepts, humanmade illustrations would better capture people's associations (e.g., associations between "money" and green might be missed from photographs of US dollars, which are more grayish than green). Setlur and Stone [43] also used "clipart", and filtered by color using Google's dominant color filter.

Language-based databases provide a different approach, linking colors and concepts through naming databases (XKCD color survey [29] used in [14, 25, 43]), concept networks (ConceptNet [26, 44] used in [14]), or linguistic corpora (Google Ngram viewer used in [43]). Naming databases provide information about color-concept pairs that were spontaneously named by participants. These data can be sparse if they they lack information about concepts that are moderately associated with a color but not strongly associated enough to elicit a color name. Concept networks and linguistic corpora link concepts to color words (not coordinates in a color space), so these methods tend to be used in conjunction with naming [14] and image [43] databases to link color words to color coordinates for use in visualizations.

2.1.2 Assigning colors to concepts in visualizations

Once color-concept associations are quantified, they can be used to compute color-concept assignments for visualizations. Various methods have been explored for computing assignments. Lin et al. [23] computed *affinity scores* for each color-concept pair using an entropy-based metric, and then solved the assignment problem (a linear program) to find the color-concept assignment that maximized the sum of affinity scores. They found that participants were better at interpreting charts of fictitious fruit sales with color palettes generated from their algorithm, compared with the Tableau 10 standard order (Fig. 2A). In another approach, Setlur and Stone [43] applied k-means clustering to quantize input colors into visually discriminable clusters using CIELAB Euclidean distance, and iteratively reassigned colors until all color-concept conflicts were resolved.

In a third approach, Schloss et al. [41] solved an assignment problem similar to [23], except they computed merit functions (affinity scores) differently. They compared three merit functions: (1) *isolated*, assigning each concept to its most associated color while avoiding conflicts, (2) *balanced*, maximizing association strength while minimizing confusability, and (3) *baseline*, maximizing confusability (Fig. 2B). They found that participants were able to accurately interpret color meanings for unlabeled recycling bins using the balanced color set, were less

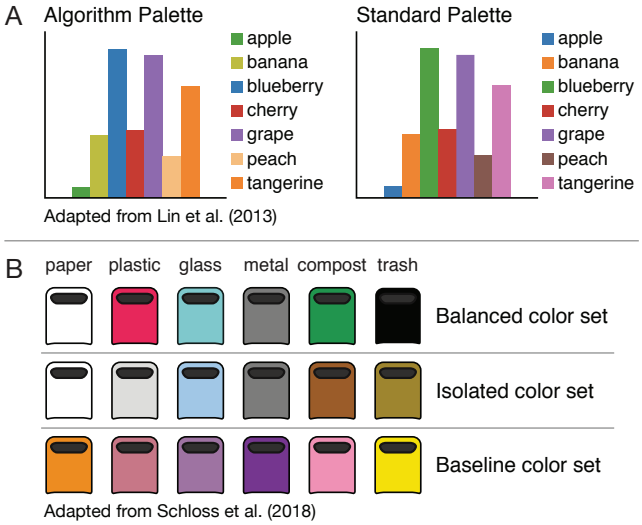

Figure 2. Examples of designs that use color-concept associations to automatically assign colors to concepts. In (A), data visualizations showed fictitious fruit sales and participants interpreted the colors using legends [23]. In (B), visualizations showed recycling bins and participants interpreted colors without legends or labels [41].

accurate for bins using the isolated color set, and were at chance for bins using the baseline color set.

2.2 Interpreting visualizations of categorical information, and implications for palette design

When people interpret the meanings of colors in visualizations, they use a process called *assignment inference* [41]. In assignment inference, people infer assignments between colors and concepts that would optimize the association strengths over all paired colors and concepts. Sometimes, that means inferring that concepts are assigned to their strongest associated color (e.g., paper to white in Fig. 2B). But, other times it means inferring that concepts are assigned to weaker associates, even when there are stronger associates in the visualization (e.g., plastic to red and glass to blue-green, even though both concepts are more strongly associated with white) [41]. This implies that people can interpret meanings of colors that are not semantically resonant (i.e., strongly associated with the concepts they represent) if there is sufficient context to solve the assignment problem (though what constitutes "sufficient context" is the subject of ongoing research).

People's capacity for assignment inference suggests that for any set of concepts, it is possible to construct many semantically interpretable color palettes. This flexibility will enable designers to navigate tradeoffs between semantics and other design objectives—discriminabilty, nameablity, aesthetics, and emotional connotation, which are sometimes conflicting [4, 12]. To create palette designs that account for these objectives, it is necessary to have good estimates of human color-concept associations over a broad range of colors.

3 General Method

In this section, we describe our approach to training and testing our algorithm for automatically estimating human color-concept associations (illustrated in Fig. 3). We began by collecting color-concept association data from human participants to use as ground truth. Then, we used Google Images to query each of the concepts and retrieve relevant images. We tested over 180 different methods (*features*) across Experiments 1A-C for extracting color distributions from images. We selected features (how many and which ones to use) by applying sparse regression with cross-validation to avoid over-fitting and produce estimates that generalized well. Model weights for each feature were chosen by ordinary linear regression.

For training and testing in Experiments 1 and 2, we used a set of 58 colors uniformly sampled in CIELAB color space ($\Delta E = 25$), which we call the UW-58 colors (see Supplementary Material Section S.1 and Supplementary Table S.1). The concepts were 12 fruits: *avocado*, *blueberry*, *cantaloupe*, *grapefruit*, *honeydew*, *lemon*, *lime*, *mango*, *orange*, *raspberry*, *strawberry*, and *watermelon*. We chose fruits, as in prior work [23,43], because fruits have characteristic, directly observable colors (high color diagnosticity [47]). We sought to establish our method for simple cases like fruit where we believed there was sufficient color information within images to estimate human color-concept associations. In future work it will be possible to identify edge cases where the method is less effective and address those limitations. In Experiment 3, we tested how our trained algorithm generalized to a different set of colors and concepts using color-concept association ratings for recycling concepts from [41]. In the remainder of this section, we describe our methods in detail.

3.1 Human ratings of color-concept associations

To obtain ground truth for training and testing our models, we had participants rate association strengths between each of the UW-58 colors and 12 fruits.

Participants. We tested 55 undergraduates (mean age = 18.52, 31 females, 24 males) at UW–Madison. Data was missing from one participant (technical error). All had normal color vision (screened with HRR Pseudoisochromatic Plates [13]), gave informed consent, and received partial course credit. The UW–Madison Internal Review Board approved the protocol.

Design, displays, and procedure. Participants rated how much they associated each of the UW-58 colors with each of 12 fruit concepts. Displays on each trial contained a concept name at the top of the screen in black text, a colored square centered below the name (100×100 pixels), and a line-mark slider scale centered below the colored square. The left end point of the scale was labeled "not at all", the right end point labeled "very much", and the center point was marked with a vertical divider. Participants made their ratings by sliding the cursor along the response scale and clicking to record their response. Displays remained on the screen until response. Trials were separated by a 500 ms inter-trial interval. All 58 colors were rated for a given concept before going onto the next concept. The order of the concepts and the order of colors within each concept were randomly generated for each participant.

Before beginning, participants completed an anchoring procedure so they knew what it meant to associate "not at all" and "very much" in the context of these concepts and colors [34]. While viewing a display showing all colors and a list of all concepts, they pointed to the color they associated most/least with each concept. They were told to rate those colors near the endpoints of the scale during the task.

Displays were presented on a 24.1 in ASUS ProArt PA249Q monitor (1920×1200 resolution), viewed from a distance of about 60 cm. The background was gray (CIE Illuminant D65, x = .3127, y = .3290, Y = 10 cd/m^2). The task was run using Presentation (`www.neurobs.com`). We used a Photo Research PR-655 SpectraScan spectroradiometer to calibrate the monitor and verify accurate presentation of the colors. The deviance between the measured colors and target colors in CIE 1931 xyY coordinates was $< .01$ for x and y, and < 1 cd/m^2 for Y.

The mean ratings for all fruit-color pairs averaged over all participants are shown in Supplementary Figs. S.1 and S.2 and Supplementary Table S.3.

3.2 Training color extraction and testing performance

Our goal was to learn a method for extracting color distributions from images such that the extracted profiles were reliable estimates of human color-concept association ratings from Section 3.1. We will describe our approach here using generic parameter names. The parameter values we actually used in the experiments are specified in the relevant experiment descriptions (Section 4).

For each of the n_{con} concepts, we queried Google Images using the name of the concept and downloaded the top n_{img} results. We then compiled a list of n_{feat} *features*. A feature is a function $f : (\text{image}, \text{color}) \rightarrow \mathbb{R}$ that quantifies the presence of a given target color in a given image. For example, a feature could be "the proportion

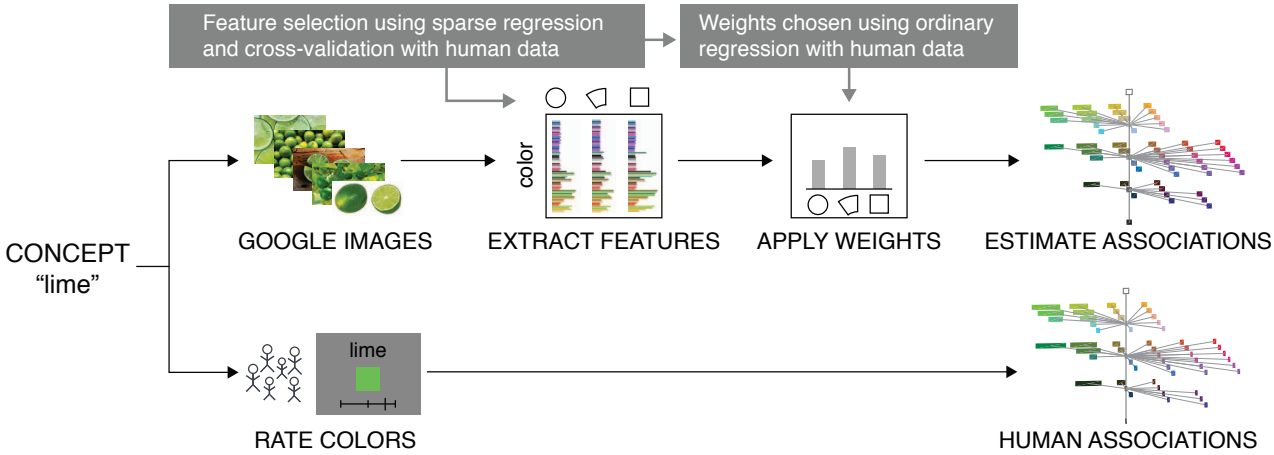

Figure 3. Illustration of our pipeline for automatically extracting color distributions from images. The bottom flow (concepts to color ratings to human associations) describes the slow yet reliable direct approach using human experiments to determine ground-truth associations. The top flow involves querying Google Images, extracting colors using a variety of different methods (*features*), then weighting those features appropriately to obtain estimated associations. Deciding which features to use and how to weight them is learned from human association data using sparse regression and cross-validation. Once the model is trained, color-concept associations can be quickly estimated for new concepts without additional human data.

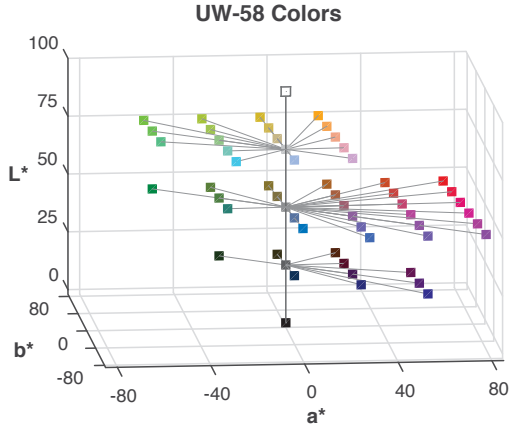

Figure 4. UW-58 colors used in Experiment 1 and 2, plotted in CIELAB color space. Exact coordinates are given in Supplementary Table S.1.

of pixels in the middle 20% of the image that are within $\Delta r = 40$ of the target color". For each of the n_{img} images and each of the n_{col} colors, we evaluated each of the n_{feat} features. This resulted in a matrix $X \in \mathbb{R}^{n_{\text{con}} n_{\text{img}} n_{\text{col}} \times n_{\text{feat}}}$, where each row was a $(\text{concept}, \text{image}, \text{color})$ triplet and each column was a different feature. We then constructed a vector $y \in \mathbb{R}^{n_{\text{con}} n_{\text{img}} n_{\text{col}} \times 1}$ such that the i^{th} element of y contains the average color-concept rating from the human experiments, for the concept and color used in the i^{th} row of X. Note that each rating in y was repeated n_{img} times because for each $(\text{concept}, \text{color})$ pair, there are n_{img} images. We used the data (X, y) in two ways: (1) to select how many features to use and (2) to choose feature weights.

Feature selection. We used sparse regression (lasso) with leave-one-out cross-validation to select features. Specifically, we selected a concept, partitioned (X, y) by rows into test data $(X_{\text{test}}, y_{\text{test}})$, containing the rows pertaining to the selected concept, and training data $(X_{\text{train}}, y_{\text{train}})$, containing the remaining concepts. We then used sparse regression on the training data with a sweep of the regularization parameter. This was repeated with every concept and we plotted the average test error versus the number of nonzero weights (Fig. 5). We examined the plot and found that $k = 4$ features provided a good trade-off between error and model complexity, so we used $k = 4$ features for all subsequent experiments. To select which features to use, we ran one final sparse regression using the full data (X, y) and chose the regularization parameter such that $k = 4$ features emerged.

Using cross-validation for model selection is standard practice in modern data science workflows. By validating the model against data that was unseen during the training phase, we are protected against overfitting and we help ensure that our model will generalize to unseen concepts and colors or different training images.

Choosing feature weights. Once the best k features were identified, we selected the weights by performing an ordinary linear regression with these k features. We ensured that human data used to compute the model weights were always different from human data used test model performance. In Experiment 1 and 2, when we trained and tested on fruit concepts, we chose feature weights using 11 fruits and tested on the 12th fruit (repeated for each fruit). In Experiment 3, we used a model trained on all 12 fruit concepts to test how well it predicted data for recycling concepts (Section 4.5).

Testing the model on new data. Once the model has been trained, it can be used to estimate color-concept associations for new concepts *and* new colors without the need to gather any new human ratings. The concept is queried in Google Images, the k chosen features are extracted from the images for the desired colors, and the trained model weights are applied to the features to obtain the estimates.

Reproducible experiments. We used cross-validation in order to ensure our results hold more broadly beyond our chosen concepts, colors, and images. We developed code in Python for all the experiments and made use of the `scikit-image` [48] and `scikit-learn` [36] libraries to perform all the image processing and regression tasks. Our code is available at `https://github.com/Raginii/Color-Concept-Associations-using-Google-Images`. This repository can be downloaded to a local machine to replicate the entire study or adapt it to test new concepts and colors. The repository also contains a write-up with detailed instructions.

4 Experiments

In Experiments 1A–1C, we systematically tested methods for extracting colors from images and assessed model fits with human color-concept association ratings. In Experiment 2, we examined model performance using different image types (top 50 Google Image downloads, cartoons, and photographs). In Experiment 3, we tested how well our best model from Experiment 1 generalized to a different set of concepts and colors.

4.1 Experiment 1A: Balls in Cartesian coordinates

Perhaps the simplest approach for extracting colors from images would be to (1) define a set of target colors of interest in a color space (e.g., CIELAB), (2) query a concept in Google Images (e.g., "blueberry"), (3) download images returned for that concept, and (4) count the number of pixels in each image that has each target color within the set. However, that level of precision would exclude many colors in images, including some that are perceptually indistinguishable from the target colors.

Moreover, not all pixels in the image may be relevant. For example, when people take pictures of objects, they tend to put the object near the center of the frame [32], so it is sensible that images ranked highly in Google Images for particular query terms will have the most relevant content near the center of the frame.

In this experiment, we varied *color tolerance*: allowing colors that are not perfect matches with targets to still be counted, and *spatial windows*: including subsets of the pixels that may be more likely to contain relevant colors for the concept. Our goal was to determine which combination of color tolerance and spatial window best captured human color-concept associations.

4.1.1 Methods

In this section, we explain how we downloaded and processed the images, constructed features, and selected features for our model.

Images. We used the same set of testing and training images throughout Experiment 1, downloaded for each fruit using Vasa's [49] Google Images Download Python script available on GitHub. We used the top $n_{\text{img}} = 50$ images returned for each fruit that were in `.jpg` format, but our results are robust to using a different number of images (see Supplementary Section S.4). We did not use the Google Image Search API used in prior work [23, 43] because it has since been deprecated. We made the images uniform size by re-scaling them to 100×100 pixels, which often changed the aspect ratio.

We converted RGB values in the images to CIELAB coordinates using the `rgb2lab` function in `scikit-image`. This conversion makes assumptions about the monitor white point and only approximates true CIELAB coordinates. It is standard to use this approximation in visualization research given that in practice, people view visualizations on unstandardized monitors under uncontrolled conditions [12, 16, 45, 46]. Although our model estimations are constructed based on approximations of CIELAB coordinates, our human data were collected on a calibrated monitor with true CIELAB coordinates (see Section 3.1). Thus, the fit between human ratings and model approximations speaks to the robustness of our approach.

Features. We constructed 30 features from all possible combinations of 5 color tolerances and 6 spatial windows as described below.

Color tolerances. When looking for a target color within a set of pixels in an image (set defined by the spatial window), we counted the fraction of all pixels under consideration whose color fell within a ball of radius Δr in CIELAB space of the target color. We tested balls with five possible values of Δr: 1, 10, 20, 30, and 40.

Spatial windows. We varied spatial windows in six ways. The first five extracted pixels from the center 20%, 40%, 60%, 80%, and 100% of the image, measured as a proportion of the total area. The 6$^{\text{th}}$ way used a figure-ground segmentation algorithm to select figural, "object-like" regions from the image, and exclude the background. Here "window" is not rectangular, but rather the shape of the figural region(s) as estimated by the active contour algorithm [28, 50], *Snakes*. This uses an initial contour binary mask and iteratively moves to find the object boundaries. We used the MATLAB implementation `activecontour` from the Image Processing Toolbox for 500 iterations, setting the initial contour as the image boundary. Although prior work suggested that figure-ground segmentation might not provide a benefit beyond eliminating borders [23], we aimed to test its effects in the context of our other color tolerance and spatial window parameters.

Feature selection. We applied the feature selection method described in Section 3.2 using the 30 features described above, as well as a constant offset term, which is standard when using regression. Using an offset is equivalent to adding one more feature equal to the constant function 1. Fig. 5 shows the mean squared error (MSE) for each number of features, averaged across all $n_{\text{con}} = 12$ fruits, using the $n_{\text{col}} = 58$ UW-58 colors, and using $n_{\text{img}} = 50$ Google Image query results for each fruit.

Based on this plot, we decided to use 4 features, which yields a good trade-off between model complexity and error reduction. We then used the full dataset to select the best 4 features, as described in Section 3.2.

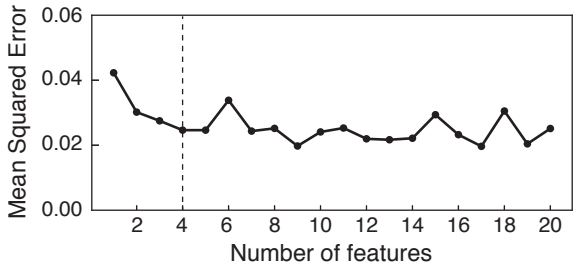

Figure 5. Mean squared test error (MSE) as a function of the number of features selected using sparse regression (lasso). MSE is averaged over 12 models obtained using leave-one-out cross-validation, using each fruit category as different test set. We selected models with four features.

The best 4 features were (1) constant offset, (2) 20% window with $\Delta r = 40$ (positive weight), (3) 100% window with $\Delta r = 40$ (negative weight), and (4) segmented figure with $\Delta r = 40$ (positive weight), see Table 1. The positive weight on the center of the image and negative weight on 100% of the image can be interpreted as a crude form of background suppression.

Table 1. Top 4 features selected using sparse regression as more features were made available in Experiments 1A to 1C. Ball features were not selected when sector or category features became available.

Model description	Features selected
Ball model	constant offset
(Experiment 1A)	Ball: $\Delta r = 40$; 20% window
Features available:	Ball: $\Delta r = 40$; 100% window
Ball only	Ball: $\Delta r = 40$; segmented
Sector model	constant offset
(Experiment 1B)	Sector: $\Delta r = 40$, $\Delta h = 40°$; 20% window
Features available:	Sector: $\Delta r = 40$, $\Delta h = 30°$; 40% window
Ball, Sector	Sector: $\Delta r = 40$, $\Delta h = 40°$; segmented
Sector+Cat model	constant offset
(Experiment 1C)	Sector: $\Delta r = 40$, $\Delta h = 40°$; 20% window
Features available:	Sector: $\Delta r = 40$, $\Delta h = 40°$; segmented
Ball, Sector, Category	Category; 20% window

4.1.2 Results and discussion

We tested the model on each of the fruits using the leave-one-out cross-validation procedure described in Section 3.2. We trained model weights using the $11 \times 50 = 550$ training images and averaged the model estimates across the 50 test images. We tested the effectiveness of the Ball model by correlating its estimates with mean human ratings over all 12 fruits $\times$ 58 colors and found a moderate correlation of .65 (Table 2). Fig. 6 shows the correlations separately for each fruit (light gray points), with fruits sorted along the x-axis from highest to lowest r value for the Ball model. There is wide variability in the model fits, ranging from $r = .93$ for *orange* to $r = .27$ for *blueberry*.

To understand why the model performed poorly for some fruits, we plotted estimated ratings for each color as a function of human ratings. Fig. 7 highlights a subset of three fruits with high, medium, and low correlations. The full set of plots are shown in Supplementary Fig. S.3. Error in performance seemed to arise from underestimating the association strength for colors that did not appear in the images but were associated with the concepts. This was particularly apparent for *blueberry*, where participants strongly associated a variety of blues that were more saturated and purplish than the blues that appeared in images of blueberries. Model estimates for those blues were as low as model estimates for oranges and greens, which were clearly *not* associated with *blueberry*. The model also overestimated values for grays and purples that were not associated with *blueberry*. These results suggested that different kinds of features would be necessary for capturing human color-concept associations.

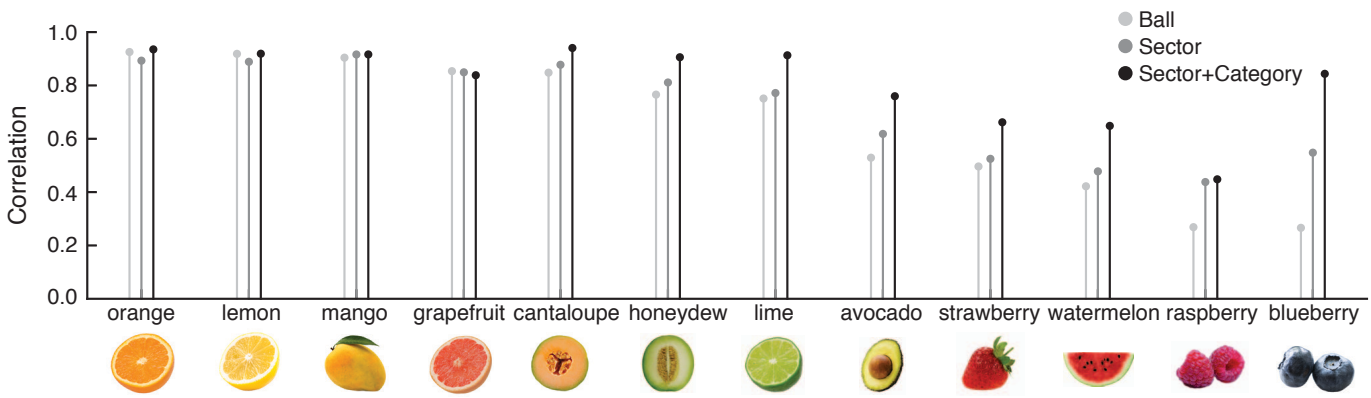

Figure 6. Correlations between model estimations and human ratings across each of the UW-58 colors for each fruit from the best 4-feature model in Experiment 1A (Ball model), Experiment 1B (Sector model), and Experiment 1C (Sector+Category model). The Sector+Category model performed best, followed by the Sector model, then the Ball model, see Table 2 and text for statistics.

Table 2. Correlations (r) between mean human color-concept association ratings and estimated associations (12 fruits $\times$ 58 colors $=$ 696 items) for each model in Experiments 1 and 2 (also shown Fig.8). All $p < .001$.

Model	Top 50	Photo	Cartoon
Ball	.65	.62	.72
Sector	.72	.69	.72
Sector+Category	.81	.80	.80

4.2 Experiment 1B: Sectors in cylindrical coordinates

A potential limitation of the ball features in Experiment 1A is that varying the size of the ball has different perceptual consequences depending on the location in color space. This is because perceptual dimensions of color are cylindrical (angle: hue, radius: chroma, height: lightness) rather than Cartesian. Balls of a fixed size that are closer to the central L* axis will span a greater range of hue angles than balls that are farther away (i.e., higher chroma). In the extreme, a ball centered on a* $= 0$ and b* $= 0$ (e.g., a shade of gray) will include colors of all hues. Thus, if we wanted a ball that was large enough to subsume all the high chroma blues (i.e., colors strongly associated with blueberries), that same large ball placed near the achromatic axis would subsume all hues of sizable chroma (see Figure 1).

To have independent control over hue and chroma variability, we defined new features with tolerance regions as cylindrical sectors around the target colors according to hue angle and chroma (Fig. 1). We tested whether color extraction using sector features, more aligned with perceptual dimensions of color space, would better estimate human color-concept associations. We used the same images as in Experiment 1A, and confined the number of features to four, using the same sparse regression approach as in Experiment 1A for feature selection. We assessed whether the best model included any of these new features, and if so, if that model's estimates fit human ratings significantly better than the Ball model did in Experiment 1A.

4.2.1 Methods

We included the same 30 features from Experiment 1A, plus 150 new features: 25 new color tolerance regions $\times$ 6 spatial windows (same spatial windows as Experiment 1A) for a total of 180 features. The 25 new color tolerance regions were defined in cylindrical coordinates in CIELch space using all combinations of 5 hue angle tolerances (Δh: 5°, 10°, 20°, 30°, 40°) and 5 chroma/lightness tolerances (Δr: 1, 10, 20, 30, 40) around each target color, see Fig. 1. The tolerances for chroma and lightness co-varied, so $\Delta r = 10$ means that both chroma and lightness had a tolerance of 10. Note that CIELch space is the same as CIELAB space except it uses cylindrical rather than Cartesian coordinates.

4.2.2 Results and discussion

As in Section 4.1.1, we first extracted the best 4 features from the pool of 180 features using sparse regression. The 4 top features only included sector features and no ball features (Table 1), so we refer to the model from Experiment 1B as the "Sector" model.

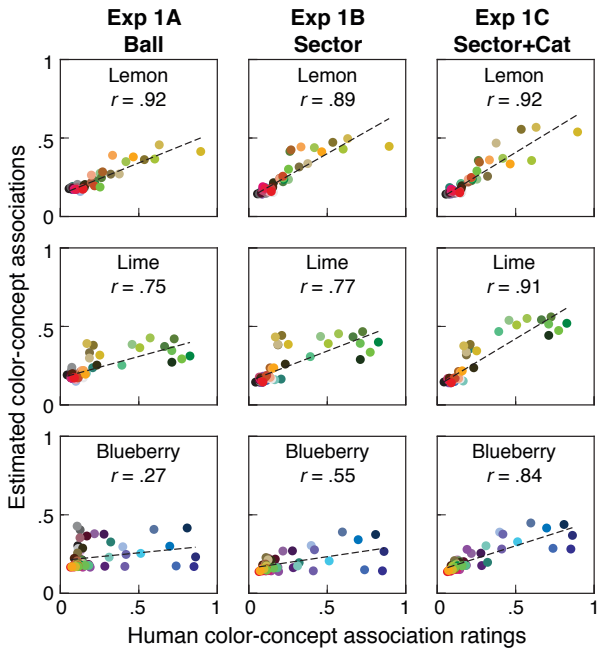

Figure 7. Scatter plots showing relationships between model estimates and human ratings for *lemon*, *lime*, and *blueberry* using models from Experiments 1A–1C. Marks represent each of the UW-58 colors, dashed line represent best-fit regression lines, and r values indicate correlations within each plot. Adding more perceptually relevant (sector) and cognitively relevant (category) features improved fit for fruits where ball features performed poorly.

We tested the effectiveness of the Sector model by correlating its estimates with mean human ratings over all 12 fruits $\times$ 58 colors. This correlation (r = .72) was stronger for the Sector model than for the Ball model (Table 2), and that difference was significant ($z = 2.46$, $p = .014$). Fig. 6 (medium gray points) shows that the Sector model improved performance for fruits that had the weakest correlations in Experiment 1A, but there is still room for improvement. The scatter plots in Fig. 7 show that the model still under-predicts ratings for blues that are strongly associated with blueberries but are not found in the images. A similar problem can be observed for limes, where several greens are associated with limes, but are not extracted from the images. The full set of scatter plots is in Supplementary Fig. S.4.

These results suggest that when people form color-concept associations, they might extrapolate to colors that are not directly observed from visual input. We address this possibility in Experiment 1C.

4.3 Experiment 1C: Color categories

In Experiment 1C, we examined the possibility that human color-concept associations extrapolate to colors that are not directly observed from visual input. One way that extrapolation might occur is through color categorization. Although colors exist in a continuous space, humans partition this space into discrete categories. English speakers use 11 color categories with the basic color terms red, green, blue, yellow, black, white, gray, orange, purple, brown, and pink [5]. The number of basic color terms varies across languages [5, 11, 20, 37, 52], but there are regularities in the locus of categories in color space [1, 20].

We propose that when people form color-concept associations from visual input, they extrapolate to other colors that are not in the visual input but share the same category (*category extrapolation hypothesis)*. To test this hypothesis, it was necessary to first identify the categories of each color within images. We did so using a method provided by Parraga and Akbarinia [35], which used psychophysical data to determine category boundaries for each of the 11 color terms. Their algorithm enables efficient lookup and categorization of each pixel within and image. We then constructed a new type of feature that represents the proportion of pixels in the image that share the color category of each of the UW-58 colors. For example, if .60 of the pixels in the image are categorized as "blue" (using [35]), then all UW-58 colors that are also categorized as "blue" will receive a feature value of .60, regardless of how much of those UW-58 colors were in the image. We assessed whether the best model included any category new features, and if so, whether that model's estimates fit human rating significantly better than the Sector model did in Experiment 1B.

4.3.1 Methods

We included the 30 ball and 150 sector features in Experiments 1A and 1B, plus 6 new color category features for a total of 186 features. We generated category features using Parraga and Akbarinia's [35] method to obtain the color categories of our UW-58 colors and the categories of each pixel within our images. We used the functions available through their Github repository [2] to convert RGB color coordinates to color categories. Specifically, we converted the $100 \times 100 \times 3$ image arrays in RGB to 100×100 arrays, where each element in the array represents the pixel's color category. For each UW-58 color, we defined the features to be the fraction of pixels in the spatial window that belonged to the same color category as the UW-58 color. We repeated the above procedure with the 6 spatial windows as before.

4.3.2 Results and discussion

Similar to Experiments 1A and 1B, we used sparse regression to extract the best 4 features among 186 total features. The model selected the constant offset and two of the same sector features from Experiment 1B, plus one new category feature (no ball features), see Table 1). Thus, we refer to this new model as the "Sector+Category" model. We obtained the model weights via linear regression as detailed in Section 3.2.

We tested the effectiveness of the Sector+Category model by correlating its estimates with mean human ratings over all 12 fruits $\times$ 58 colors. This correlation was stronger for the Sector+Category model than for the Ball model or Sector model (Table 2), and those differences were significant ($z = 6.55$, $p < .001$; $z = 4.08$, $p < .001$, respectively). Fig. 6 shows that the Sector+Category model (black points) further improved fits for the fruits that had weaker fits using the other two models. As seen in Fig. 7, by including category extrapolation, this model increased the estimated values for colors that were strongly associated with limes and blueberries (greens and blues, respectively) that were under-predicted by the previous models because those colors were not in the images. The full set of scatter plots is in Supplementary Fig. S.5.

4.4 Experiment 2: Comparing image types

As described in Section 2.1.1, Lin et al. [23] queried concept words alone and concept words appended with "clipart", reasoning that humanmade illustrations might better capture associations for some types of concepts. We propose that humans produce clipart illustrations based on their color-concept associations, not solely based on real-world color input. If color-concept associations are already incorporated into clipart, that would explain why clipart is useful for estimating color-concept associations, especially when natural images fall short. However, if a model contains features that effectively estimate human color-concept associations, it may have sufficient information to do as well for natural images as it does for humanmade illustrations, such as clipart. To examine this hypothesis, we tested our models from Experiments 1A-C on two new image types: cartoons (humanmade illustrations) and photographs (not illustrations).

In addition, we used the approach of Lin et. al. (mentioned in Section 2.1.2) to compute probabilities, a precursor to their affinity score that most corresponds to color-concept associations. We then compared those probabilities with our human ratings.

4.4.1 Methods

We downloaded two new sets of images, by querying each fruit name appended with "cartoon" or "photo". We queried "cartoon" rather than "clipart" because clipart sometimes contained parts of photographs with the background deleted, and we wanted to constraint this image set to humanmade illustrations. Unlike Experiment 1 where we used the first 50 images returned by Google Images, we manually curated the photo and cartoon image sets to ensure (a) they were photos for the photo set and cartoons in the cartoon set, (b) they included an observable image of the queried fruit somewhere in the image, and (c) they were not images of cartoon characters (e.g., "Strawberry Shortcake", a character in an animated children's TV show that first aired in 2003). This resulted in 50 images in each set.

We trained and tested using the same top 4 features from the Ball, Sector, and Sector+Category models in Experiment 1, except we substituted the training images for the manually curated sets of cartoon images or photo images. This yielded three sets of model weights for the different image types. We then compared each model's performance for the three image types.

4.4.2 Results and discussion

Fig. 8 and Table 2 show the overall correlations between color-concept associations and model estimates across all 12 fruits $\times$ 58 colors. The correlations for the top 50 images are those previously reported in Experiment 1A–1C, and included here as a baseline. Fig. 8 also shows overall correlations with the probabilities computed using the method of Lin et. al. [23] as another baseline.

The results suggest there is a benefit to using human-illustrated cartoons for the Ball model (which does not effectively capture human color-concept associations), but the benefit diminishes for the Sector and Sector+Category models (which better capture human color-concept associations). Specifically, the Ball model using cartoons was significantly more correlated with human ratings than the Ball model using top 50 images ($z = 2.46$, $p = .014$) with no difference between cartoon and top 50 images for the other two models (Sector: $z = 0.0$,

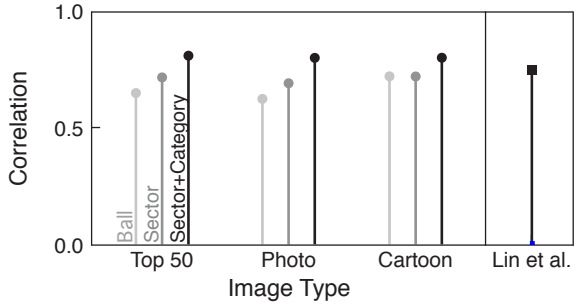

Figure 8. Correlations for top 50 images, photo images, and cartoon images using the Ball, Sector, and Sector+Category models. The Sector+Category model performed best and was similar across all image types. The Ball model was worst for top 50 and photo images, but less poor for cartoon images. Estimates based on Lin et al. [23] were strongly correlated with human ratings, but less so than the Sector+Category model (see text for statistics and explanation).

$p = 1.0$; Sector+Category: $z = .53$, $p = .596$). There were no significant differences in fits using photo vs. top 50 images for any of the three model types (Ball: $z = 0.94$, $p = .347$; Sector: $z = 1.11$, $p = .267$; Sector+Category: $z = 0.53$, $p = .596$).

To further understand these differences, we tested for interactions between image type and feature type, using linear mixed-effect regression (R version 3.4.1, lme4 1.1-13, see [7]). The dependent measure was model error for each fruit (sum of the squared errors across colors for each fruit). We included fixed effects for Model, Image, and their interactions, and random slopes and intercepts for fruit type within each Model contrast and Image contrast. We initially tried including random slopes and intercepts for interactions, but the model became too large and the solver did not converge. We tested two contrasts for the Model factor. The first was Category+Sector vs. average of Ball and Sector, which enabled us to test whether Category+Sector was overall better than the other two models. The second was Sector vs. Ball, which enabled us to test whether the Sector model was better than the Ball model. We tested two contrasts for the Image factor. The first was cartoon vs. average of top 50 and photo, which enabled us to test whether cartoons were overall better than the other two images types. The second contrast was top 50 vs. photo, which enabled us to test whether top 50 images (which included some cartoons) were better than photos. Reported beta and t-values are absolute values.

The results for the Model contrasts showed that Sector+Category model preformed best, and the Sector model performed better than the Ball model. That is, there was less error for Sector+Category than the combination of Ball and Sector ($\beta = 0.10, t(11) = 4.42, p = .001$), and less error for Sector than for Ball ($\beta = 0.083, t(11) = 7.90, p < .001$). The contrasts for Image were not significant ($ts < 1$), indicating no overall benefit for human made cartoons.

However, the first Image contrast comparing cartoons vs. the average of top 50 and photo interacted with both Model contrasts. Looking at the interaction with the first Model contrast (Sector+Category vs. average of Ball and Sector), the degree to which Sector+Category model outperformed the other models was greater for top 50 and photo images than for cartoons ($\beta = 0.01, t(44) = 4.38, p < .001$), see Fig. 8. Looking at the interaction with the second Model contrast (Sector vs. Ball), the degree to which Sector model outperformed the Ball model was greater for the top 50 and photo images than the cartoons ($\beta = 0.02$, $t(44) = 6.65, p < .001$). No other interactions were significant ($ts < 1$).

In this experiment, we also evaluate how Lin et al.'s estimates of color-concept associations [23] match our human ratings. Their estimates come from a hybrid of color distributions extracted from top image downloads and clipart, so we provide their model with our top 50 images and cartoons as input. As shown in Fig. 8, the correlation for all fruits and colors was $r = .74$, which is similar to our Sector models ($r = .69$ to .72 depending on image type) and less strong than our Sector+Category models ($r = .80$ to .81 depending on image type). The difference in correlation for the Lin et al. model and our Sector+Category model for the top 50 images was significant ($z = 3.29$, $p < .001$). We note that these models are not directly comparable because our models used either top 50 images or cartoons, not both at the same time (except if cartoons happened to appear in the top 50 images).

In summary, using cartoon images instead of other image types helped compensate for the poor performance of our Ball model. However, image type made no difference for our more effective Sector and Sector+Category models. We interpret these results as showing that cartoons help the Ball model compensate for poorer performance because humans make cartoons in a way that builds in aspects of human color-concept associations. For example, visual inspection suggests that cartoon blueberries tend to contain saturated blues that are highly associated with blueberries yet not present in photographs of blueberries. However, the benefit of humanmade illustrations is reduced if model features are better able to capture human color-concept associations. This suggests that using our Sector+Category models on the top Google Image downloads is sufficient for estimating human color-concept associations without further need to curate the image set, at least for the concepts tested here.

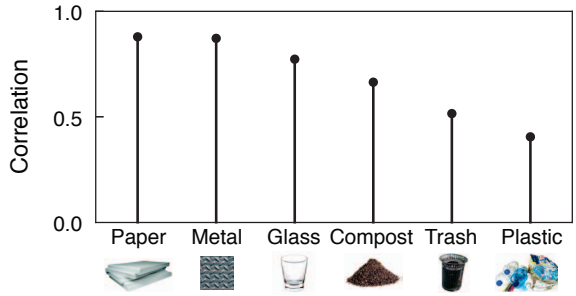

Figure 9. Correlations between human ratings and estimated ratings across all colors for each recycling-related concept using the Sector+Category model. The range of correlations is similar to fruits (Fig. 6.)

4.5 Experiment 3: Generalizing beyond fruit

Experiment 3 tested how well our model trained on fruit generalized to a new set of concepts and colors using the recycling color-concept association dataset from [41]. The concepts were: *paper*, *plastic*, *glass*, *metal*, *compost*, and *trash*, and colors were the BCP-37 (see Supplementary Table S.2). Unlike fruits, which have characteristic colors, recyclables and trash can come in any color. Still, human ratings show systematic color-concept associations for these concepts, and we aimed to see how well our model could estimate those ratings.

4.5.1 Methods

As in Experiment 1, we downloaded the top 50 images from Google Images for each recycling-related concept. To estimate color-concept associations, we used our Sector+Category model from Experiment 1C with feature weights determined from a single linear regression using all 12 fruits, as described in Section 4.3.

4.5.2 Results and discussion

We tested the effectiveness of the Sector+Category model by correlating its estimates with mean human ratings over all 6 recycling-related concepts $\times$ 37 colors. The correlation was $r = .68$, $p < .001$, moderately strong, but significantly weaker than the corresponding correlation of .81 for fruit concepts in Experiment 1C ($z = 3.84$, $p < .001$). Fig. 9 shows the correlations separately for each recycling concept. The fits range from .88 for paper to .40 for plastic, similar to the range for fruits in Experiment 1C (.94 to .49) (see Supplementary Fig. S.6).

5 General Discussion

Creating color palettes that are semantically interpretable involves two main steps, (1) quantifying color-concept associations and (2) using those color-concept associations to generate unique assignments of colors to concepts for visualization. Our study focused on this first step, with the goal of understanding how to automatically estimate human color-concept associations from image statistics.

5.1 Practical and theoretical applications

We built on approaches from prior work [4, 23–25, 43] and harnessed perceptual and cognitive structure in color space to develop a new method for effectively estimating human color-concept associations. Our method can be used to create the input for various approaches to assigning colors to concepts for visualizations [4, 14, 23, 41, 43]. By estimating full distributions of color-concept associations over color space that approximate human judgments (as opposed to identifying only the top associated colors), it should be possible to use assignment methods to define multiple candidate color palettes that are semantically interpretable for a given visualization. This flexibility will enable balancing semantics with other important factors in design, including perceptual discriminability [15, 45, 46], name difference [16], aesthetics [12], and emotional connotation [4].

Our method can also be used to design stimuli for studies on visual reasoning. For example, evidence suggests people use assignment inference to interpret visualizations [41] (Section 2.1.2), but little is known about how assignment inference works. Studying this processes

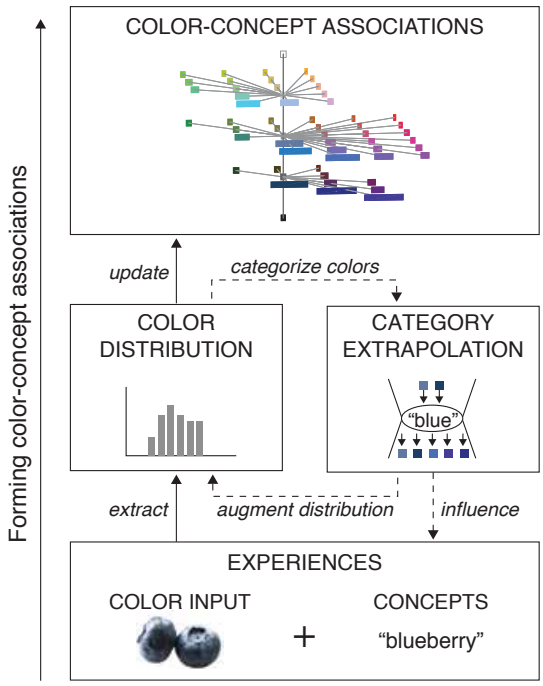

Figure 10. Process model for how color-concept associations are formed.

requires the ability to carefully manipulate color-concept association strengths within visualizations, which requires having good estimates of human color-concept associations.

5.2 Forming color-concept associations

In addition to providing a new method for estimating color-concept associations, this study sparked new insights into how people might form color-concept associations in the first place. In Fig. 10, the path with solid arrows illustrates our initial premise for how people learn color concept associations from their environment. When they experience co-occurrences between colors and concepts, they extract color distributions and use them to update color-concept associations [39]. In Fig. 10, color-concept association strength is indicated using marker width (e.g., *blueberry* is highly associated with certain shades of blue). However, in the present study we found that it was insufficient to only extract colors (or nearby colors) from images to estimate human color-concept associations (Experiments 1A and 1B). We needed to extrapolate to other colors that shared the same category as colors within the image to produce more accurate estimates (Experiment 1C).

Based on these results, we propose there is another part of the process—category extrapolation—illustrated by the path with dashed arrows in Fig. 10. While extracting the color distribution from color input, people categorize colors using basic color terms (e.g., "blue"). This categorization process extrapolates to colors that are not in the visual input, but are within the same color category (e.g., extrapolating to all colors categorized as "blue" upon seeing a blueberry, even though only a subset of blues are found in the image). We believe that extrapolated colors augment the color distribution extracted from color input, which in turn further updates color-concept associations. Given that categorization can influence color perception [10, 19, 37, 51] and memory [3, 21] (see [52] for a review), category extrapolation may also feedback to influence color experiences.

5.3 Open questions and limitations

Generalizability to other concepts. In this study, we focused on fruit—concrete objects with directly observable colors—so we could study different methods of extracting and extrapolating colors from images where the colors would be systematic. We assessed generalizability to other, recycling-related concepts that are less color diagnostic [47] than fruit (e.g., paper, plastic, and glass can be any color), but recyclables are still concrete objects. However, there is concern that image-based methods may not be effective for estimating color-concept associations for abstract concepts that do not have directly observable colors [23, 43], though see [4]. Nonetheless, people do have systematic associations between colors and abstract concepts [30, 31, 53]. Future work will be needed to asses the boundary conditions of image based methods and further explore incorporating other, possibly language-based methods [14, 43] for estimating color-concept associations for abstract concepts.

Image segmentation. Our models might also be limited in their ability to generalize for concepts that refer to backgrounds rather than objects (e.g., "sky") [23]. All three models included a feature that extracted colors from figural regions and segmented away the backgrounds. Further research is needed to evaluate performance for background-related concepts, but limitations might be mitigated using semantic segmentation, in which particular regions of images are tagged with semantic labels [27].

Cultural differences. Our category extrapolation hypothesis implies that color-concept associations could differ between cultures whose languages have different color terms. Different languages partition color space in different ways [5, 11, 20, 37, 52]—e.g., some languages have separate color terms for blues and greens, whereas others have one term for both blues and greens. If a language has separate terms for blues and greens, experiencing blue objects like blueberries should result in color-concept associations that extrapolate only to other blues, not greens. But, if a language has one term for blues *and* greens, experiencing blueberries should result in associations that extrapolate to blues *and* greens. This is an exciting area for future research.

Structure of color categories. Our model defined color categories using a boundary approach—either a color was in a given category or not, with no distinction among category members. However, color categories have more complex structure, including a prototype, or best example, and varying levels of membership surrounding the prototype [38]. A model that accounts for these complexities in category structure may improve on the fit to human color-concept associations.

6 Conclusion

The goal of this study was to assess methods for automatically estimating color-concept associations from images. We tested different color extraction features that varied in color tolerance and spatial window, different kinds of images, and different concept sets. The most effective model used features that were relevant to human perception and cognition—features aligned with perceptual dimensions of color space and a feature that extrapolated to all colors within a color category. This model performed similarly well across the top 50 images from Google Images, curated photographs, and curated cartoon images. The model also generalized reasonably well to a different set of colors and concepts without changing any parameters. Through this study, we produced a method trained and validated on human data for automatically estimating color-concept associations, while generating new hypotheses about how color input and category extrapolation work together to produce human color-concept associations.

Acknowledgments

The authors thank Christoph Witzel, Brian Yin, John Curtain, Joris Roos, Anna Bartel, and Emily Ward for their thoughtful feedback on this work, and Melissa Schoenlein, Shannon Sibrel, Autumn Wickman, Yuke Liang, and Marin Murack for their help with data collection. This work was supported in part by the Office of the Vice Chancellor for Research and Graduate Education at UW–Madison and the Wisconsin Alumni Research Foundation. The funding bodies played no role in designing the study, collecting, analyzing, or interpreting the data, or writing the manuscript.

REFERENCES

[1] J. T. Abbott, T. L. Griffiths, and T. Regier. Focal colors across languages are representative members of color categories. *PNAS*, 113(40):11178–11183, 2016.

[2] A. Akbarinia. Color categorisation. `https://github.com/ArashAkbarinia/ColourCategorisation`, 2017.

[3] G.-Y. Bae, M. Olkkonen, S. R. Allred, and J. I. Flombaum. Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, 144(4):744, 2015.

[4] L. Bartram, A. Patra, and M. Stone. Affective color in visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1364–1374. ACM, 2017.

[5] B. Berlin and P. Kay. *Basic color terms: Their universality and evolution*. University of California Press, 1969.

[6] J. Bertin. *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin Press, Madison, 1983.

[7] M. Brauer and J. J. Curtin. Linear mixed-effects models and the analysis of nonindependent data: A unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. *Psychological Methods*, 23(3):389–411, 2018.

[8] C. A. Brewer. Color use guidelines for mapping and visualization. In A. M. MacEachren and D. R. F. Taylor, editors, *Visualization in Modern Cartography*, pages 123–148. Elsevier Science Inc., Tarrytown, 1994.

[9] R. D'Andrade and M. Egan. The colors of emotion. *American Ethnologist*, 1(1):49–63, 1974.

[10] L. Forder and G. Lupyan. Hearing words changes color perception: Facilitation of color discrimination by verbal and visual cues. *Journal of Experimental Psychology: General*, 148(7):1105, 2019.

[11] E. Gibson, R. Futrell, J. Jara-Ettinger, K. Mahowald, L. Bergen, S. Ratnasingam, M. Gibson, S. T. Piantadosi, and B. R. Conway. Color naming across languages reflects color use. *PNAS*, 114(40):10785–10790, 2017.

[12] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss. Colorgorical: Creating discriminable and preferable color palettes for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):521–530, 2017.

[13] L. H. Hardy, G. Rand, M. C. Rittler, J. Neitz, and J. Bailey. *HRR pseudoisochromatic plates*. Richmond Products, 2002.

[14] C. Havasi, R. Speer, and J. Holmgren. Automated color selection using semantic knowledge. In *2010 AAAI Fall Symposium Series*, 2010.

[15] C. G. Healey. Choosing effective colours for data visualization. In *Proceedings of the 7th Conference on Visualization'96*, pages 263–ff. IEEE Computer Society Press, 1996.

[16] J. Heer and M. Stone. Color naming models for color selection, image editing and palette design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1007–1016. ACM, 2012.

[17] N. Humphrey. The colour currency of nature. *Colour for Architecture*, 5:95–98, 1976.

[18] A. Jahanian, S. Keshvari, S. Vishwanathan, and J. P. Allebach. Colors–messengers of concepts: Visual design mining for learning color semantics. *ACM Transactions on Computer-Human Interaction*, 24(1):2, 2017.

[19] P. Kay and W. Kempton. What is the Sapir–Whorf hypothesis? *American Anthropologist*, 86(1):65–79, 1984.

[20] P. Kay and T. Regier. Resolving the question of color naming universals. *PNAS*, 100(15):9085–9089, 2003.

[21] L. J. Kelly and E. Heit. Recognition memory for hue: Prototypical bias and the role of labeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(6):955, 2017.

[22] M. Kinateder, W. H. Warren, and K. B. Schloss. What color are emergency exit signs? Egress behavior differs from verbal report. *Applied Ergonomics*, 75:155–160, 2019.

[23] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer. Selecting semantically-resonant colors for data visualization. In *Computer Graphics Forum*, volume 32, pages 401–410. Wiley Online Library, 2013.

[24] A. Lindner, N. Bonnier, and S. Süsstrunk. What is the color of chocolate?–extracting color values of semantic expressions. In *Conference on Colour in Graphics, Imaging, and Vision*, volume 2012, pages 355–361. Society for Imaging Science and Technology, 2012.

[25] A. Lindner, B. Z. Li, N. Bonnier, and S. Süsstrunk. A large-scale multilingual color thesaurus. In *Color and Imaging Conference*, volume 2012, pages 30–35. Society for Imaging Science and Technology, 2012.

[26] H. Liu and P. Singh. ConceptNeta practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226, 2004.

[27] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[28] A. W. Michael Kass and D. Terzopoulo. Snakes: Active contour models. *International Journal of Computer Vision*, 1:321–331, 1998.

[29] R. Munroe. xkcd color survey results, 2010.

[30] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The measurement of meaning*. University of Illinois press, 1957.

[31] L.-C. Ou, M. R. Luo, A. Woodcock, and A. Wright. A study of colour emotion and colour preference. Part I: Colour emotions for single colours. *Color Research & Application*, 29(3):232–240, 2004.

[32] S. E. Palmer, J. S. Gardner, and T. D. Wickens. Aesthetic issues in spatial composition: Effects of position and direction on framing single objects. *Spatial Vision*, 21(3):421–450, 2008.

[33] S. E. Palmer and K. B. Schloss. An ecological valence theory of human color preference. *PNAS*, 107(19):8877–8882, 2010.

[34] S. E. Palmer, K. B. Schloss, and J. Sammartino. Visual aesthetics and human preference. *Annual Review of Psychology*, 64:77–107, 2013.

[35] C. A. Parraga and A. Akbarinia. Nice: A computational solution to close the gap from colour perception to colour categorization. *PloS one*, 11(3):1–32, 2016.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[37] D. Roberson, J. Davidoff, I. R. Davies, and L. R. Shapiro. Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, 50(4):378–411, 2005.

[38] E. H. Rosch. Natural categories. *Cognitive Psychology*, 4(3):328–350, 1973.

[39] K. B. Schloss. A color inference framework. In . G. V. P. L. MacDonald, C. P. Biggam, editor, *Progress in Colour Studies: Cognition, Language, and Beyond*. John Benjamins, Amsterdam, 2018.

[40] K. B. Schloss, C. C. Gramazio, A. T. Silverman, M. L. Parker, and A. S. Wang. Mapping color to meaning in colormap data visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):810–819, 2019.

[41] K. B. Schloss, L. Lessard, C. S. Walmsley, and K. Foley. Color inference in visual communication: the meaning of colors in recycling. *Cognitive Research: Principles and Implications*, 3(1):5, 2018.

[42] K. B. Schloss, E. D. Strauss, and S. E. Palmer. Object color preferences. *Color Research & Application*, 38(6):393–411, 2013.

[43] V. Setlur and M. C. Stone. A linguistic approach to categorical color assignment for data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):698–707, 2016.

[44] R. Speer, J. Chin, and C. Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[45] M. Stone, D. A. Szafir, and V. Setlur. An engineering model for color difference as a function of size. In *Color and Imaging Conference*, volume 2014, pages 253–258. Society for Imaging Science and Technology, 2014.

[46] D. A. Szafir. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401, 2018.

[47] J. W. Tanaka and L. M. Presnell. Color diagnosticity in object recognition. *Perception & Psychophysics*, 61(6):1140–1153, 1999.

[48] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[49] H. Vasa. Google images download. `https://github.com/hardikvasa/google-images-download`, 2018.

[50] R. T. Whitaker. A level-set approach to 3d reconstruction from range data. *International journal of computer vision*, 29(3):203–231, 1998.

[51] J. Winawer, N. Witthoft, M. C. Frank, L. Wu, A. R. Wade, and L. Boroditsky. Russian blues reveal effects of language on color discrimination. *PNAS*, 104(19):7780–7785, 2007.

[52] C. Witzel and K. R. Gegenfurtner. Color perception: Objects, constancy, and categories. *Annual Review of Vision Science*, 4:475–499, 2018.

[53] B. Wright and L. Rainwater. The meanings of color. *The Journal of General Psychology*, 67(1):89–99, 1962.