A Classification of Infographics

Helen C. Purchase[1], Katherine Isaacs[2], Thomas Bueti[2], Ben Hastings[2], Aadam Kassam[2], Allen Kim[2], Steffan van Hoesen[2]

[1] School of Computing Science, University of Glasgow, UK
[2] Department of Computer Science, University of Arizona, USA

helen.purchase@glasgow.ac.uk

Abstract. Classifications are useful for describing existing phenomena and guiding further investigation. Several classifications of diagrams have been proposed, typically based on analytical rather than empirical methodologies. A notable exception is the work of Lohse and his colleagues, published in Communications of the ACM in December 1994. The classification of diagrams that Lohse proposed was derived from bottom-up grouping data collected from sixteen participants and based on 60 diagrams. Mean values on ten Likert-scales were used to predict diagram class. We follow a similar methodology to Lohse, using real-world infographics (i.e. embellished data charts) as our stimuli. We propose a structural classification of infographics, and determine whether infographics class can be predicted from values on Likert scales.

Keywords: Infographics, Classification, Empirical Studies

1 Introduction

Infographics present quantitative data (like that in bar charts or scatterplots), and are typically embellished with graphic elements or pictures. Infographics can increasingly be found in popular media, online, in public presentations and organisations' brochures, making data more visible, engaging, and memorable. Several researchers investigate the effect of using embellishments in data presentation by conducting empirical studies, the stimuli sometimes "real" (sourced from media publications) and sometimes "fabricated" (created by researchers for the purposes of their experiment).

With increasing infographics research, classification is useful. "A carefully designed classification can serve to show not only the full range of available possibilities but also the relationships between these, and … acts more as an instrument rather than simply as a 'filing cabinet'" (Rankin [1]). Kwasnik [2] explores the relationship between classifications and knowledge discovery: "Classification is a way of seeing. Phenomena of interest are represented in a context of relationships that, at their best, function as theories by providing description, explanation, prediction, heuristics, and the generation of new questions."

Classifications can be generated by thorough and systematic analysis of a range of stimuli [3,4,5], or by soliciting the views of human participants (Lohse et al. [6]). The research reported in this paper takes the latter approach: we conducted an empirical study to create a classification of infographics, based on "real" stimuli.

2 Prior Research

Garcia and Cox [4] considered diagrams in the UK National School Curriculum, classifying them into 20 types, and discussing them with respect to children's "graphical readiness" to interpret diagrams. Purchase [3] analysed diagrams from the proceedings of the first seven conferences on the Theory and Application of Diagrams: her primary classification is abstract vs concrete and embellishments are defined as 'additional visual elements'. Novic [7] classes scientific diagrams as "iconic", "charts and graphs" and "schematic diagrams". Blackwell and Engelhardt [8] surveyed several diagram taxonomies, noting differences according to the nature of the visual elements used, their positioning, their semantics, and context of use. Rankin [1] commented on the diversity of classification criteria used by different researchers, distinguishing between two types of diagrammatic classification: functional (focusing on purpose) and structural (focusing on form). Our motivator is the CACM article by Lohse et al. (1994) [6], who presented the first structural classification of diagrams based on empirical data, collected from 16 participants.

The term 'infographic' is defined in many different ways. Saleh et al [11] write: "Infographics are complex graphic designs integrating text, images, charts and sketches". Albers [5] writes: "an infographic takes a large amount of information in text or numerical form and condenses it into a combination of text images and with a goal of making the information presentable." We wished to focus on the metaphorical use of graphical elements (e.g. pictures of coins, cakes, monkeys, suitcases, wine glasses) as a means of depicting data: that is, if these graphical elements were removed from the image, then this would remove the representation of the data. So, a bar chart with a picture of the moon in the background is not an infographic; a bar chart where each bar is represented by a picture of a space shuttle of a different height is. Haroz et al [12] discovered that superfluous images not used for representing data were distracting, and so we insist that any graphics items directly depict data values.

Albers [5] used an 'open-ended card sort' method on 25 infographics to devise four categories: bullet list equivalent, snapshot with graphic needs, flat information with graphic needs, and information flow/process – a categorisation formed from the author's personal view. Borkin et al [10] do not describe how they created the 12 categories in their 'visualisation taxonomy'; Saleh et al [11] investigate the 'stylistic similarity' of infographics, but do not explicitly identify or name different 'styles'. Popular websites (e.g. excelcharts [13], juiceanalytics [14]) propose classifications of data charts, but do not include charts with graphical embellishments.

3 Methodology

We follow the empirical and data analysis methodology of Lohse et al [6] closely, our objectives being to create a hierarchical taxonomy of different types of infographics, and devise a means of predicting the class of an infographic, based on the responses to ten Likert scales. Our empirically-derived classification structure can inform further empirical research on infographics.

We use the following 'infographic' definition: "An image that presents a data set, where the data quantities are depicted using pictures of recognisable common items."

3.1 Materials

We used existing data sets: Saleh et al's [11] set of 19,594 infographics, and Borkin et al's [9] 5,693. In addition, we looked at 55 infographics from the Times Higher Education magazine and a set of 174 infographics previously gathered from a range of sources. Most images were eliminated quickly because they presented more than one data set, were of poor resolution, were duplicates, had an extreme aspect ratio, had text not in English, were photographs, or were data charts not embellished with images. We eliminated those where the images or pictures used to embellish the data chart were not integral to the presentation of data. We chose 60 infographics to ensure data presentation method variety (see www.dcs.gla.ac.uk/~hcp/infographics).

Our starting point for devising our Likert scales was Lohse et al's original ten [6], although we also drew from those used by Quispel [15], Loroco et al [16], and Harrison et al [17]. Our scales are: spatial/non-spatial; non-temporal/temporal; hard to understand/ easy to understand; concrete/ abstract; attractive/ unattractive; emphasizes the whole/ emphasizes parts; informative/ uninformative; minimal/ cluttered; shows patterns/ does not show patterns; literal/ metaphorical.

3.2 Experimental Procedure

Twenty participants took part (10 female, mean age=33, 9 students, 3 high school graduates, 8 university graduates). Three were studying computer science, and the rest were a mixture of a variety of subjects (e.g. Law, Social Work, Business); none were studying visualization, graphic design or art. Each experiment was conducted one-on-one, and took approximately 90 minutes.

Table 1 shows how our procedure differs from that of Lohse et al [6]. Each participant was given the 60 infographics in a pile, in a different random order for each participant, and asked to describe briefly, aloud, what each infographic was about. They then laid all the infographics out on the table and grouped them according to "visual design." If participants were not sure what was meant by the phrase "visual design", this was explained to them using phrases like "the way in which the graphic has been designed", or "the overall visual design of the infographic." They could have as many groups as they liked, as many infographics in each group as they liked, and could take as long as they wished. They then explained their rationale behind each group. After a break, the participants rated each infographic on the ten Likert scales.

3.3 Data Analysis

We follow the data analysis procedure of Lohse et al.

(1) Outlier pruning. We calculate the distance between pairs of participants using Jaccard coefficients: the distance between participants Pi and Pj is 1 - A / (N - B) where N is the number of infographic pairs (60 * 59 / 2 = 1770), A is the number of infographic pairs that appear together in both Pi and Pj's groupings, and B is the number of infographic pairs that appear in separate groups in both Pi and Pj's groupings. Complete linkage hierarchical clustering on the matrix of Jaccard coefficients produced a tree: participants on singleton branches until final mergings are considered outliers.

(2) Classification of Infographics. We derive a hierarchical cluster tree of infographics using complete linkage hierarchical clustering. The similarity matrix comprises similarity scores for infographics pairs: the number of participants who put the pair in the same group. We normalized the similarity and subtracted from one to convert to distances for clustering. The existence of ties in distance scores leads to different hierarchical clusterings based on the ordering of infographics in the matrix. Following Lohse et al, we computed six hierarchical clusterings, permuting the matrix each time.

(3) Predicting the classification. We use average Likert scores for each infographic. We perform a principal components analysis (PCA) on the rating scales to determine if any scales should be removed due to explaining little of the variance. With the remaining scales, we then build two classifiers, one using classification and regression trees (CART) and one using linear discriminant analysis (LDA). Per the requirements of the scikit-learn library and following Lohse et al, we input cluster priors through the 'class_weight' parameter for CART and as a passed parameter for the LDA. Both the CART and LDA were evaluated using 11-fold cross-validation (as in Lohse et al). We used the default Gini index as the splitting criterion for the CART analysis.

Table 1: Experimental procedure.

	Lohse et al [6]	Our experiment
Stimuli	60 diagrams, chosen to be "representative...within the domain of static, two-dimensional graphic representations."	Infographics with primary aim of presenting quantitative data, embellished with images.
Familiarisation	Participants named each diagram (step 1).	Participants described what each infographic is "about" (step 1).
Rating	Participants rated each diagram on ten nine-point Likert scales (step 2).	Participants rated each diagram on ten nine-point Likert scales (step 4)
Grouping	Participants performed a bottom-up sorting task on randomly laid out diagrams, grouping items with respect to "similarity" (step 3).	Participants grouped the infographics with respect to 'visual design' (step 2).
Explanation	Participants gave the rationale for their grouping (step 4)	Participants gave the rationale for their groups (step 3)

4 Results

Outliers are participants on singleton branches until the final stages of merging; our clustering yielded one such participant who grouped by subject matter rather than design. Further analysis of the reasons participants gave for their grouping indicated three others focused on attributes other than 'visual design' (e.g. colour, semantics, audience). We removed these four participants' data, leaving 16 valid data sets.

We set the similarity distance threshold to 0.9, resulting in seven to eight clusters for each of the six cluster trees. We inspected these six clusterings to form a metaclustering by grouping infographics that appeared in the same cluster in the majority. Our classification analysis revealed six top-level categories, two of which are comprised of two second-level sub-categories. Two infographics appeared with similar frequency in two categories (in the 'area-as-quantity' and 'single circle' classes): they both presented two sets of data. We had attempted to ensure that each infographic only presented one data set – these two had slipped through the net of our filtering process so were removed from further analysis. Two other multiply-classified infographics were both based on flags – we therefore created a separate 'flag' category for them, the seventh top-level category. The seven categories are:

- **Bar Charts** (16). A bar chart is the main data presentation form.
- **Geographical** (4). The primary shape is a geographical map.
- **Units** (6). The quantity of the data is represented by several small graphic images, each representing an amount of data.
- **Area-as-Quantity**. Different data quantities are represented by the areas of shapes. In some cases, these are **Familiar Shapes** (e.g. circles, triangles) (9); in others **Uncommon Shapes** are used (e.g. dinosaurs, mail boxes) (5).
- **Single Circle** (5). Data is represented within a singular circular form.
- **Proportion-as-Quantity**. The data quantities are shown as proportions of a larger object. Divisions of **Rectangular Shapes** are most common (6), although **Irregular Shapes** (e.g. banana, wine glass) are also used (5).
- **Flags** (2). The primary shape used is that of national flags.

The first three principal components accounted for 91.1% of the variance. Each Likert scale had a squared factor loading >10% in at least one of the first three principal components. Thus, we chose to keep all of the scales. To avoid overfitting the CART tree, we set the maximum number of leaves to 10, similar in detail to Lohse et al (11). The resulting tree correctly classifies 55.2% of the infographics with a cross validation mean accuracy of 28%. Examining the CART tree and the distribution of average Likert values for all of the infographics, we observed there is a high degree of variance within many of the clusters for each Likert. For example, paired bar charts often represent before and after, giving them a higher temporal score than the non-paired bar charts in the bar chart group. The LDA resulted in a slightly more accurate classifier (63.8%, with cross validation mean accuracy of 38%).

5 Discussion

Some specific infographics produced surprising results. A line chart (i03, see website in Section 3.1) was consistently grouped as a bar chart; its source was The Times Higher magazine, as was the case for several bar charts – perhaps there is a common generic 'Times Higher' visual style that led it to be grouped with others from the same source? Alternatively, since this was the only infographic based on a 'line chart', it may have been grouped with bar charts so as to not be a singleton group. The cartogram (i19) was the extreme on several Likert scales, and was not classified as 'Geographic'. We believe that some participants did not recognise it as representing a world map. An infographic which represented money as piles of poker chips (i57) was not classified as a bar chart; however, since the individual piles of chips have no meaning, and it is the comparative area of the two piles that is important, 'Area-as-

Quantity' is indeed the best classification for it. The map of Africa showing how its area compared to that of other countries (i23) was predominantly classified as Geographical, although it might also reasonably be in the Proportion-as-Quantity (Irregular Shapes) or Area-as-Quantity (Uncommon Shapes) categories. i58 might have better been classified in the Area-as-Quantity class (Familiar Shapes) – we believe that the highly rectangular nature of the items depicted in it led it to be grouped with the other Rectangular Shapes as part of Proportion-as-Quantity. We deliberately included an infographic that depicted a single data point (i26) as an extreme example; it was classified as Proportion-as-Quantity since, we believe, the range was implicitly interpreted as [-40°F, 140°F], the common range of thermometers of that design.

Our classification is richer than those of Albers [5] and Borkin et al [10], which are based on popular categories of data charts (e.g. donut chart, stacked area chart, line, scatter plot, tree [10]) or are vague (e.g. "flat information with graphic needs" [5]). Some of our empirically-derived classifications are similar to common data charts (e.g Bar Charts, Geographical), but they also include categories based on how the space on the page is used to depict data (e.g Proportion-as-Quantity). Unlike other classifications, our results show that participants were not only aware of how data was being depicted (e.g. using proportions to show quantity), but were also highly sensitive to the types of shapes used – familiar, uncommon, rectangular, irregular. No other classification considers the form or shape of the graphical embellishments used.

There is a strong prevalence of infographics that rely on area comparisons to show difference in data values: 14 Area-as-Quantity and 11 Proportion-as-Quantity. It is well known that perception of area is less accurate than perception of length or position [18]. This phenomenon might actually serve infographics designers who wish mislead readers: Tufte [19, pp69-70] gives examples where perception of area rather than length can easily lead to incorrect inferences.

The Likert scales were poor predictors of class, in contrast to Lohse et al's results. The data indicates that the Likert scales are orthogonal to the classifications – that is, their values bear little relation to the groupings created by the participants. Thus, whether an infographic is attractive or not, or easy to understand, or temporal etc. does not reflect its visual form. In many ways, this is reassuring news for infographics designers – they are not obliged to use any of the nine specified categories if they wish to emphasise any of these Likert properties. In addition, Lohse et al suggested that their successful predictions might have been a result of participants doing the Likert scales before the grouping task, and then implicitly using these scales in their grouping. Our participants completed the Likert scales after the grouping task, so as to mitigate against this possibility. Having the two tasks done by different participants of similar demographic profile might be a more reliable way of testing the predictions: that way, there would be no cross-contamination between tasks.

Any empirical study is subject to limitations. Our classification results are bound by the scope of the 60 infographics we chose (from a total set of 25,516), and our prediction results by our choice of Likert Scales. The demographic profile of our participants is reasonably well-spread, although slightly skewed towards younger ages. Future work can validate our hierarchy with other infographics and participants.

6 Conclusion

The prevalence of infographics in the popular media, advertising, public notices and organizational brochures makes them a rich source for diagrammatic research. There is still a great deal of empirical work to be done in this area: what makes infographics memorable or engaging? Do graphical embellishments inhibit interpretation – both of individual data points or the overall message? How can deliberately misleading messages be presented without being obvious? Classifications provide frameworks for research, and are particularly useful if based on real-world examples and created through human experimentation. Our novel classification of infographics provides an empirically derived basis for researchers in this area – who no longer need to create their own analytical classifications.

Acknowledgements

This research was conducted while the first author was visiting the University of Arizona. We are grateful to all the participants, and to Babak Saleh and Michelle Borkin who shared their image data sets. Ethical approval was given by the University of Arizona Institutional Review Board (ref: 1711982836).

References

1. Rankin, R.A.: A taxonomy of graph types. Information Design Journal, vol. 6(2), pp. 147-159. (1990).
2. Kwasnik, B.: The role of classification in knowledge representation and discovery. Library Trends, vol. 48, pp. 22–47. (1999).
3. Purchase, H.C.: Twelve years of diagrams research. Journal of Visual Languages and Computing, vol. 25, pp57-75. (2014).
4. Garcia Garcia, G., Cox, R.: Diagrams in the UK National Curriculum. In: Stapleton, G., Howse, J., Lee, J. (eds.) Proceedings of the Fifth International Conference on the Theory & Application of Diagrams. Lecture Notes in Computer Science, LNAI, 5223. pp. 360–363. Berlin (2008).
5. Albers, M.J.: Infographics: Horrid chartjunk or quality communication. IPCC, IEEE, pp. 13-15. (2014).
6. Lohse, G.L., Biolsi, K., Rueter, H.H.: A Classification of Visual Representations. Communications of the ACM, 37(12), 36-49 (1994).
7. Novick, L.R.: The importance of both diagrammatic conventions and domain-specific knowledge for diagram literacy in science. In: Barker-Plummer, D., Cox, R., Swoboda, N. (eds.) Diagrams 2006, LNAI 4045, pp. 1–11, Springer-Verlag (2006).
8. Blackwell, A.F., Engelhardt, Y.: A meta-taxonomy for diagram research. In: Anderson, M., Meyer, B., Olivier, P., (eds.) Diagrammatic Representation and Reasoning, pp.47-64. London: Springer-Verlag (2002)
9. Borkin, M.A., Bylinskii, Z., Kim, N.W., Bainbridge, C.M., Yeh, C.S., Borkin, D., Pfister, H., Oliva, A.: Beyond memorability: Visualization recognition and recall. Visualization and Computer Graphics, IEEE Trans., vol. 22(1). pp. 519–528. (2016)
10. Borkin, M.A., Vo, A.A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., Pfister, H.: What makes a visualization memorable? IEEE Trans. Vis. Comput. Graph., vol. 19(12), pp. 2306–2315. (2013)

11. Saleh, B., Dontcheva, M., Hertzmann, A., Liu, Z.: Learning style similarity for searching infographics. In: Proceedings of the 41st Graphics Interface Conference (GI’15), pp. 59–64. Canadian Information Processing Society, Toronto (2015).
12. Haroz, S., Kosara, R., Franconeri, S.L.: ISOTYPE Visualization: Working Memory, Performance, and Engagement with Pictographs. Proceedings CHI '15, pp. 1191–1200. ACM, New York, (2015).
13. A classification of chart types, https://excelcharts.com/classification-chart-types/, last accessed 2017/11/15
14. JuiceAnalytics Chart Chooser, http://labs.juiceanalytics.com/chartchooser/index.html, last accessed 2017/11/15
15. Quispel, A.: Data for all: How professionals and non-professionals in design use and evaluate information visualizations. PhD Thesis, Tilburg University (2016)
16. Locoro, A., Cabitza, F., Actis-Grosso, R., Batini, C.: Static and interactive infographics in daily tasks: a value-in use and quality of interaction user study. Computers in Human Behavior, vol. 71, pp. 240-257. (2017)
17. Harrison, L., Reinecke, K., Chang, R.: Infographic Aesthetics: Designing for the First Impression. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI ’15, pp.1187–1190. ACM, New York (2015).
18. Cleveland, W.S. and McGill, R.: Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. Journal of the American Statistical Association, vol. 79:387, pp. 531-554 (1984).
19. Tufte, E.R. The Visual Display of Quantitative Information, Graphics Press (1983)

Appendix: The infographics classification tree