Hunches and Sketches: rapid interactive exploration of large datasets through approximate visualisations

Advait Sarkar, Alan F Blackwell, Mateja Jamnik, Martin Spott

Computer Laboratory, University of Cambridge, UK
BT Research and Technology, Ipswich, UK
{advait.sarkar, alan.blackwell, mateja.jamnik}@cl.cam.ac.uk
martin.spott@bt.com

Abstract. Information visualisation presents powerful techniques for data analytics. However, rendering visualisations of big datasets is impractical on commodity hardware. There is increasing interest in approaches where data sampling and probabilistic algorithms are used to support faster processing of large datasets. This approach to approximate computation has not yet paid close attention to the way that approximate visualisations are perceived and employed by human users, as a specific variety of diagrammatic convention. Our intent is to apply this understanding of approximate visualisations as a diagrammatic class to mainstream data science and information visualisation research.

Keywords: visualising uncertainty, information visualisation, exploratory data analysis, approximate inference, big data, sketches.

1 Visual analysis, large datasets, and uncertainty

The utility of visualisation for data analysis cannot be understated. The power of the human perceptual system paired with the visualisation capabilities of modern software tools allows for the rapid detection of trends, outliers, and comparisons of quantities – even by those without statistical expertise. Visualisations are also useful for those with deeper analytical skill. The space of analytical questions one can ask of a particular dataset is infinite, but only some questions yield interesting answers. Consequently, exploratory data analysis is divided between two approaches: the "top-down", hypothesis-testing approach wherein a specific statistical technique is used to answer a specific statistical question (e.g. *Is there a significant difference between these groups?*, or *How does a change in X affect Y?*), and the "bottom-up", hypothesis-generation approach where the interesting questions are identified and formulated (e.g. *Should I investigate the relationship between variables X and Y?*).

Cognitive task analysis of this sensemaking process suggests that experienced analysts often invoke the two processes in an opportunistic mix [1]. Visualisations can help rapidly prune the space of interesting hypotheses, and it can help verify

many of these hypotheses [2], which spares the analyst the effort of conducting a more elaborate statistical investigation of a question that in hindsight turns out to be uninteresting, or the wrong question to ask.

Shneiderman claims that a combined approach would enable more effective exploration whilst giving users a greater sense of control over the direction their exploration takes [3]. Bertini and Lalanne call for researchers to identify which aspects of analytical problems can be best solved using the human perceptual system, which are best solved using machine learning techniques, and then design for this blend of strengths [4]. Keim et al. refer to tools which embrace the idea of human-machine collaboration to solve analytical problems as "advanced visual analytics interfaces" [5].

There are many such advocates of increased integration between sophisticated statistical techniques and information visualisation tools. While this is an attractive idea, recent increases in the sheer volume of data can make visual techniques inaccessible to those attempting to perform analysis on commodity hardware. For instance, a scatterplot of 10,000 data points renders relatively quickly in Microsoft Excel or R on a commodity desktop computer. However, as of this writing it is grindingly slow to render a scatterplot of 10,000,000 points. This is the situation we often find ourselves in today. It is not conducive at all to rapid interactive exploration, and defeats the benefits of visualisation. This problem is unlikely to be alleviated by advances in hardware, as the growth in data volumes is facilitated in part by improved processing capacities. Advances in distributed computing are similarly a double-edged sword, potentially improving the computing power available for rendering visualisations but also facilitating data volume growth.

One solution to this problem is to not interact with the entire dataset, but to first reduce or transform it. For instance, a small representative sample could help generate/eliminate many of the same candidate hypotheses as if one were operating on the entire dataset. Besides sampling [7], a number of approximation techniques have been developed in the past few decades that allow for fast processing of large datasets in exchange for small, quantifiable error bounds, including *sketches* and *online aggregation* [8, 9]. These advances have led to the development of database tools that can perform fast approximate queries [10].

An important note about terminology: the aforementioned "sketches" are in fact simply data structures and algorithms. They are only sketches in the sense that they are approximations of the original dataset; they are otherwise unrelated to the normal use of the word "sketch", i.e. they are not intrinsically visual entities. For instance, the Bloom filter [11] is a data structure that represents a mathematical set and supports fast approximate membership querying. It does not represent the set exactly, but rather hashes items into a compact bit vector that approximates, or sketches, the original set. While such techniques do not use the word "sketch" in more than a metaphorical manner, the idea that these approaches could be augmented with visualisations appears to be an interesting avenue for exploration.

2 From sketches of large datasets to hunches

Our first main proposition is that data summarisation techniques can be used to interactively render approximate, exploratory visualisations of large datasets. For instance, in Figure 1, the plots on the left are of relatively large datasets. They render slowly and are therefore difficult to interact with. The plots on the right use samples or sketches of those datasets, and render much faster.

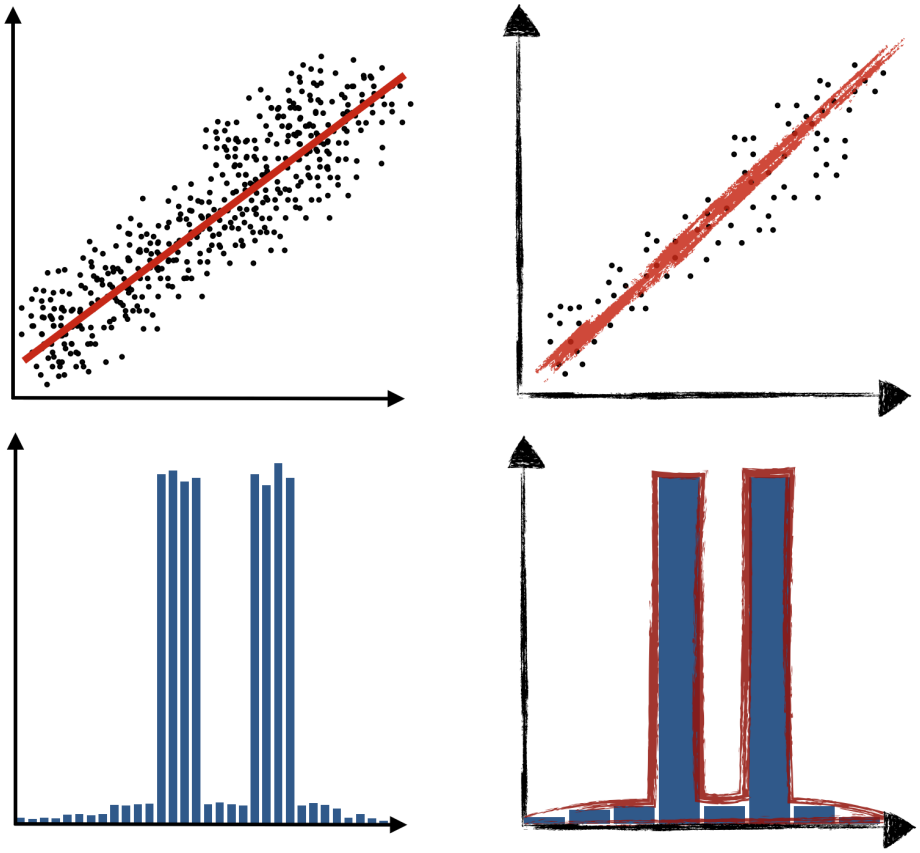

Fig. 1. Exact and approximate visualisations.

Our second main proposition is that depicting summaries of large datasets is a potentially useful application for techniques for visualising uncertainty. There are many such techniques, perhaps the most familiar being the use of error bars in bar charts and histograms. The literature discusses a variety of other techniques [12–14], including the use of transparency, blurring, painterly rendering [15, 16], and animation [17]. In particular the use of informal, sketch-like visualisations is thought to influence willingness to interact with and question the visualisation [18]. As noted by Eckert et al. [19], sketches are not simply degraded versions of a canonically accurate visual representation, but support specific cognitive and social functions.

Visualisations of uncertainty emphasise that these summaries will not support exact inference, but instead facilitate rapid informal reasoning and the formation of "hunches" – approximate hypotheses and heuristics for exploring the hypothesis space. Hunch-driven reasoning yields informal answers to open-ended questions an analyst might have (e.g. *Does this look like signal or noise? Does there appear to be cluster structure in the data? What is the general shape of the distribution? Is there an inflection point in the time series?*) before formulating specific statistical questions. These hunches may be produced on a mixed-initiative basis, i.e. collaboratively by the user and the system, thus providing a new interaction metaphor for "intelligent discovery assistants" [20].

The upper right graph in Figure 1 shows a reduced dataset which is much faster to render than the full dataset to its left. While the slope of the trend line may differ from the true slope of the trend line for the entire dataset, and the confidence intervals of any regression analysis might be wider, the reduced dataset is sufficient for the analyst to form the hunch (or informal hypothesis) of a linear relationship. The approximate nature of this hypothesis is expressed through its informal rendering, emphasising that it is not the regression coefficients that are important, but rather that a linear model may be viable. Similarly, the histogram in the lower right may have been created using a fast approximate cardinality estimator such as the linear counting algorithm [21]. It is an imperfect representation of the dataset to its left, however, the important observation is that a bimodal distribution exists, not the specific frequencies being represented.

Going forward, it will be important to study and identify several common types of these visual insights. While it would be worthwhile to demonstrate that certain transformations of the original dataset through sketching and sampling techniques will necessarily preserve these insights, it is also important to consider how we might visualise transformations that make no such guarantees or have probabilistic error bounds, which would greatly expand the range of techniques available for these interactive visualisations.

3 Conclusion

We have presented a vision for a programme of research into new tools for the interactive analysis of large datasets through approximate visualisations. These combine fast approximation techniques and techniques for visualising uncertainty, yielding new approaches to interacting with approximate visual hypotheses, or "hunches". These approaches have the potential to afford rapid interaction with large datasets through conventional, accessible modern tools for information visualisation, running on commodity hardware.

Acknowledgements

Advait Sarkar is funded through an EPSRC Industrial CASE studentship sponsored by BT Research and Technology, and also through a Premium Studentship from the University of Cambridge Computer Laboratory.

References

1. Pirolli, P., Card, S.: The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis. Proceedings of International Conference on Intelligence Analysis, 5, 2-4 (2005).
2. Keim, D.A: Visual Exploration of Large Data Sets. CACM 44(8), pp. 38-44 (2001)
3. Shneiderman, B.: Inventing discovery tools: combining information visualization with data mining. Information Visualization, vol. 1, no. 1, pp. 5-12, (2002)
4. Bertini, E., Lalanne, D.: Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. ACM SIGKDD Explorations Newsletter, 11(2), 9 (2010)
5. Keim, D. A., Bak, P., Bertini, E., Oelke, D., Spretke, D., Ziegler, H.: Advanced visual analytics interfaces. Proc. AVI '10, 3 (2010)
6. Blackwell, A. F., Church, L., Plimmer, B. Gray, D.: Formality in Sketches and Visual Representation: Some Informal Reflections. VLHCC workshop 11-18 (2008)
7. Chaudhuri, S., Das, G., Narasayya, V.: Optimized stratified sampling for approximate query processing. ACM Transactions on Database Systems, 32(2), 9 (2007)
8. Cormode, G.: Sketch techniques for massive data. Synposes for Massive Data: Samples, Histograms, Wavelets and Sketches, 1-3 (2011)
9. Hellerstein, J. M., Haas, P. J., Wang, H. J.: Online aggregation. ACM SIGMOD Record, 26(2), 171-182 (1997)
10. Agarwal, S., Mozafari, B., Panda, A., Milner, H., Madden, S., Stoica, I.: BlinkDB: queries with bounded errors and bounded response times on very large data. Proc. 8th ACM European Conference on Computer Systems (pp. 29-42) (2013)
11. Bloom, B. H.: Space/time trade-offs in hash coding with allowable errors. Communications of the ACM, 13(7), 422-426 (1970)
12. Johnson, C. R., Sanderson, A.: A next step: Visualizing errors and uncertainty. Computer Graphics and Applications, IEEE, 23(5), 6-10 (2003).
13. Zuk, T., Carpendale, S.: Theoretical analysis of uncertainty visualizations. Proc. SPIE 6060, Visualization and Data Analysis 2006, 606007 (2006)
14. Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M., Pavel, M.: A typology for visualizing uncertainty. In Electronic Imaging 2005 (pp. 146-157). International Society for Optics and Photonics (2005)
15. Boukhelifa, N., Bezerianos, A., Isenberg, T., Fekete, J.: Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty. IEEE Transactions on Visualization and Computer Graphics (TVCG), 18(12), 2769-2778 (2012)
16. Wood, J., Isenberg, P., Isenberg, T., Dykes, J., Boukhelifa, N., Slingsby, A.: Sketchy rendering for information visualization. IEEE TVCG, 18(12), 2749-2758 (2012)
17. Ehlschlaeger, C. R., Shortridge, A. M., Goodchild, M. F.: Visualizing spatial data uncertainty using animation. Computers & Geosciences, 23(4), 387-395 (1997)
18. Bresciani, S., Blackwell, A. F., Eppler, M.: A Collaborative Dimensions Framework: Understanding the mediating role of conceptual visualizations in collaborative knowledge work. Proc. 41st HICSS (pp. 364-364). IEEE (2008)
19. Eckert, C., Blackwell, A., Stacey, M., Earl, C., Church, L.: Sketching across design domains: Roles and formalities. Artificial Intelligence for Engineering Design, Analysis and Manufacturing, 26(3), 245-266 (2012)
20. Serban, F., Vanschoren, J., Kietz, J.-U., Bernstein, A.: A survey of intelligent assistants for data analysis. ACM Computing Surveys, 45(3), 1-35. (2013)
21. Whang, K. Y., Vander-Zanden, B. T., Taylor, H. M.: A linear-time probabilistic counting algorithm for database applications. ACM TODS, 15(2), 208-229 (1990)