Simple is Good: Observations of Visualization Use Amongst the Big Data Digerati

Daniel M. Russell

Google

1600 Amphitheatre Parkway
Mountain View, CA 94043, USA

1-650-253-2862

drussell@Google.com

ABSTRACT

While modern information visualization (IV) has been around for several decades, the inventions of IV seem to be peripheral to the everyday work in companies that would seem to be the most likely to use these inventions. In this case study, Google uses very few IV tools, relying mostly on more traditional ways of looking at data and data relationships. What has brought about this state of affairs? An analysis shows that the basic causes of low adoption are (a) difficulty of data wrangling and sharing the work products of analysis, (b) the need to share a common visual language literacy across different parts of the organization, (c) problems in using IV tools to communicate and present complex data analyses. At the same time, IV technology is found to be more useful in the investigation phase of research, rather than for communication and presentation reasons.

CCS Concepts

• *Human-centered computing~HCI design and evaluation methods* • *Human-centered computing~Field studies*

Keywords

Information Visualization; Evaluation; Field studies.

1. INTRODUCTION

Google is a classic big data company; it has huge information analysis tasks that encompass a wide variety of data types, with many different kinds of relationships and properties. As such, you might expect that Google, like other companies that have big data analysis tasks, would be an avid user of Information Visualization tools and systems. After all, one of the basic ideas behind all of the research that has been done on IV has been to simplify the analytical tasks of complex data. [19]

Yet, as one attends meetings, reads analysis documents, and talks with Google analysts, you can't help but be surprised by the relative lack of advanced IV tools. What's going on? Why does Google not use IV systems constantly?

Interviews with analysts and a careful examination of analyses, and presentations within big data companies suggest that while IV technology is used, it's not used nearly as broadly as expected,

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

Copyright is held by the owner/author(s).

AVI '16, June 07-10, 2016, Bari, Italy

ACM 978-1-4503-4131-8/16/06.

<http://dx.doi.org/10.1145/2909132.2933287>

and in particular, that it is used almost exclusively by a relatively small number of analysts. Why is this so?

2. STUDIES OF IV USES

While there have been a number of studies of IV use, many of them [13] have focused on the analysts who use the IV technology. These studies seek to “encapsulate the current practices in the information visualization research community” and are often field studies that are primarily about the uses that IV experts (developers, IV users, data analysts) make of their tools. These studies focus on how well the visualization functions to bring out insights through visual analytics. Occasionally, case studies will be done, but again, they focus primarily on the expert users—people who are already invested in the IV tool use.

In this study, I look at how the products of IV are used throughout the corporate setting more broadly. Experts have expert tools that are feature-rich, deep, and often difficult-to-use; power-tools for power-users. Is this the case for IV technology generally? How much does the research of the IV field influence decision-making in a place like Google?

2.1 Interviews with analysts

In early 2016, I conducted a series of 10 interviews with Google analysts—that is, the people whose job it is to examine data from Google products and services, and then answer specific data-based questions or to do research into particular behaviors of human and system behaviors. These are the people who use all of the traditional tools of big data science; large file systems (Google File System, GFS, [7]), MapReduce to run analysis tasks [5], web-scale query systems such as Dremel [15], TensorFlow for machine learning [1], and Spanner for globally replicated databases [3].

As a part of their day-to-day jobs, they deal with massive data sets, posing questions about the data they see, analyzing it for normal and special-case behaviors. Typical questions might include overall system performance, the relative use-patterns of different versions of a UI, timing questions, global distribution of subsystem use, and so on. These are, in many ways, the people for whom IV tools are ideal.

The interviews took the form of semi-structured interviews, asking questions about their day-to-day use of different IV systems, what issues they had with them, and how they communicated their findings to the stakeholders, peers, and others within the company.

The thing that struck me during these conversations was the relative *lack* of IV system use across the participants. There certainly are IV systems in use, but according to their reports, such tools are used to gain personal (or at most, small team) insights; occasionally they produce a graph or two, and then they move on.

Unlike database systems, whose *raison d'être* is to be a back-end service that is used to generate reports showing how data satisfies a given query, IV systems are designed to be used to explore data relationships and produce engaging, useful, and insightful visual representations of that data.

The analyst comments led me to do a more extensive analysis of work products, trying to understand where IV technology is used, and where it shows up in the overall work and life of the company. I summarize the issues that came in common across the analysts in Section 3 (Issues).

2.2 IV Use in Analysis Reports

Motivated by this disconnect, we did an analysis of 100 randomly selected (internal) presentations that are used to convey the results of analysis. This set of presentations covered a broad range of analytic topics, from system performance to the results of A/B UI testing and even new product uptake and adoption. These 100 reports were all longer than 15 pages in length and were the final key deliverable for each of the research projects covered. (For comparison, I also scanned 25 written reports, with essentially the same results.)

Each report was scanned to find all IV outputs (both static and interactive). Each such graph, diagram, or illustration was then identified as to IV type. The results are below in this graph.

Figure 1. The most common chart types in Google's analysis reports and presentations.

Of the top 16 IV diagram types, only 5—Session Viewer diagrams, heatmaps, Sankey, sunburst, and force vector diagrams—might be thought of as more sophisticated visualizations. (Session Viewer charts are discussed in section 4, Case Study).

As you can see, traditional, easy-to-use visualization methods such as line graphs, histograms, and tables dominate. (And yes, it's correct: In this set of 100 analyses, each report averages 13.25 charts / report, most of which are line graphs. This is very data-presentation heavy work.)

Out of this collection of ~1700 pages of reports, there was one 3D scatterplot, and one 3D heatmap showing a complex configuration space.

In practice, these reports are often presented as-is: What's in the report is what's presented to the stakeholders. On occasion, the meeting would launch the interactive version of the analysis, but the analysts reported that this was rare (with <10% of all meetings), and when an interactive data visualization was shown in an analysis meeting, 90% of those were either tables (e.g.,

spreadsheets that could be filtered or sorted interactively) or line graphs (where the variables displayed could be changed, added, removed, or rescaled). The majority of such interactive use in meetings was from dashboards, the vast majority of which are multiples of line graphs, histograms, and arc-node diagrams.

To be sure, analysts can and do use more sophisticated IV tools to explore data (and Google has created and supports many of these, e.g., [6, 8, 22], but the use cases of advanced IV are often limited to the investigation phase, rather than the reportage phase. But even those uses are more limited than one might expect.

Why?

3. ISSUES THAT PREVENT WIDE IV USE

In the interviews, one of the key questions asked was “What factors prevent sophisticated information visualizations from being more widely used?” This question arose from the observation in Figure 1 that, despite decades of research in IV methods, the most commonly used visualizations are *not* the most innovative or novel, but the most aged and common.

The most frequently mentioned reasons about non-adoption given by our analyst participants in our interviews are issues that seem to be outside of the realm of information visualization research, but which are crucial to IV's wide-spread use and uptake. Each of these issues was mentioned by at least 4 different interviewees. While all of these issues have little to do with the research agenda of the IV community, each is seen as critical to IV use in a corporate context.

3.1 Sharing

A common issue for the analysts was the need to share analyses, both data sets and analyses in the middle of the analysis process. Often, an analyst would want to share and discuss their ongoing work—a crucial aspect of any IV tool is that it allow easy sharing of data + visualization. (most preferred: sharing a single URL with embedded parameters that could be shared via email or instant message).

3.2 Repeatability

The second key concern for the analysts is the ability to repeat their analysis, both within their own analysis context, but also within another analyst's context. This almost always requires sharing common data sets (and getting the shared data access privileges set correctly) AND having all of the customization parameters shared as well. As the visualizations (and the amount of state kept within the tool) grow in complexity and sophistication, simply sharing visualizations + data becomes more complex—especially if the analysts want to have the same visualization more than once.

3.3 Data Wrangling

Since the beginning of IV as a data-centered enterprise, getting the data into the right form—collecting, filtering, cleaning, validating—all of those “wrangling” steps take a substantial amount of time. [2, 10]

Although every IV practitioner knows that data wrangling is a major part of the IV work process, the connections between the IV tool and “getting the data right” so it can be visualized is a significant part of the struggle.

Consequently, IV tools that tightly integrate data access, versioning, and management are preferred over those that have a more hands-off relationship, and putting all of the data management steps purely outside of the IV system.

3.4 Bespoke Visualizations

Among the analysts, few have the time to spend doing IV research: their day-job is data analysis. While some problems and analyses require more subtle and sophisticated visual analytics, the majority of the job is working with data and extracting insights. Their jobs frequently simply do not require *bespoke visualizations* that offer only small improvements over what's already available in the IV tools they have at hand.

To be sure, sometimes data relationships ARE intricate and complex. If the analysis need is large enough, the analysts can (and do) create bespoke visualizations by building new interactive visualizations (most frequently in some combination of R, D3 [4], Google's Visualization Toolkit [9], and Python—but this is clearly not a preferred path. (For example, the Sunburst interactive visualization from D3 has become relatively well-known within Google as a good way to explore hierarchical containment—not just because it's revealing during use, but also because the presentation form is understandable as a static image. More on this in Section 5.)

3.5 Defaults are Simple and Good

As one analyst said, “simple is good...” which embodies a key lesson for these analysts: They prefer to use what's easily available, usually as options in the current data analysis and visualization toolkit. Again, this is a kind of time/task tradeoff optimization issue. In a data analysis environment where time to analysis completion is intense, using an easily available charting choice that is integrated is often preferred over another option that might offer only marginal improvement in insights/minute [16]. We know that defaults are rarely changed by tool users. [14] As a consequence, it seems that the vast majority of charts and visualizations are created with standard tools (spreadsheets and their accompanying visualization types), or with standard charting libraries.

3.6 Problem of Visualization Literacy

While career analysts might well have and use IV tools to help them with their work, perhaps the biggest single issue that our participants raised is the problem of having to explain what a given visualization means.

“When I started to explain the visualization diagram, I could see the execs eyes glaze over.” one analyst said in our discussion, “and you don't want to see that happen when you get to the discussion about your data.”

While stakeholders deeply literate people, they often don't have the broad ranging visualization literacy that's required to understand both how the IV is presenting the underlying data, AND what this data is saying to them with respect to decision-making. To repeat this theme: the day-job of the stakeholders is decision-making, not visualization—they also prefer visualization that are easy to understand and to use for immediate data insights. The problem of getting a new visual pattern into common use is tricky: it has to be a coordinated effort between members of the IV community and the users, who need to see the long-term benefit of spending time learning how to read a new visualization deeply enough to have insights into the data and confident enough to use those insights for decision-making.

4. CASE STUDY: Session Viewer

Session Viewer is an IV system for exploring session logs. Used primarily by Google search analysts, it let you look at a set of

anonymized search sessions side-by-side for contrast-and-compare purposes [12].

Session Viewer was originally built as part of Lam's 2007 PhD research [11] as a Java application. An analyst would create a dataset (typically a CSV created by using other tools to extract a time series from search logs), then import that data into SessionView, and then begin to explore the data by changing filters, views, selecting subranges, tagging events, etc. It was an excellent example of a visualization tool that was well-matched the *research* needs of the analysts (In fact, a part of Lam's thesis research was doing contextual inquiry to understand what analytical needs the user population had).

However, by being a Java application (which required setting up a Java environment), and requiring that data extraction and loading was decoupled from the investigation, and that there was no way to effectively share a running instance of the application, Session Viewer met limited use.

It (as so many IV systems do) *demoed well, but failed to catch on*.

The field of information visualization is full of one-off instances of visualization methods (see, for instance, the Periodic Table of Visualizations [17] for a sampling).

For Session Viewer, there was a small, but annoying amount of overhead in both launching the Java environment (which was out of the regular workflow for most analysts) and the special preparation of a dataset for loading and visualization. As Lam points out [13], the visualization and user interaction must satisfy a real information need on the part of the user, presenting data in a visually understandable form that maps from visuals onto underlying data properties. From the field studies in Lam's thesis work [11], it was clear that the tool satisfied those needs quite well.

But it had these small user disruptions that together drove analyst use fairly low.

4.1 Reimplementation leads to adoption

However, in a rewrite of the system during 2012, Lam reimplemented Session Viewer as a JavaScript-based system, meaning that it could be use through a regular browser (just as nearly all of the Google analyst tools are presented), *and* it was coupled to a simple backend system that would quickly extract and load the data based on a pattern specification. In addition, a number of sophisticated query tools were removed from the system (e.g., to do pattern-matching of particular kinds of subsequences in the data), making it a much simpler system to use. As mentioned in Issue 3.1 (above), all of the Session Viewer state was encoded in the URL, vastly simplifying the problem of sharing a Session Viewer visualization.

Once these small disruptions were removed, Session Viewer started to pick up additional users, and has become an important tool in the work of analysts at Google. This is especially true now that the Session Viewer visualization language of query sequences shown side-by-side has become part of the shared visual language within the company (See Figure 2). This was accomplished by a proactive effort to evangelize its use, and share the skills of how to use Session Viewer, and the knowledge of how to read its diagrams.

Figure 2: Session Viewer shows several different search sequences side-by-side (each row is one sequence). Colors indicate different kinds of actions, and all of the sequences are aligned on the common event type at time 0. The visual language and interpretation of Session Viewer charts takes some time to percolate throughout the organization.

5. PRESENTATION vs. INVESTIGATION

It's clear that while sophisticated IVs are used by big data analysts, they are most often as *investigation* tools, rather than as *presentation* tools.

As with the stakeholders, when analysts create bespoke IV tools, visualizations, or entire systems, the investment has to be worth it. We rarely see IV tools being developed for one-off uses. If the investment in development time (and novel IV invention effort) can be shared across a number of users, then the cost can be justified. As seen in Figure 1, the number of different *presentation* IV forms is relatively limited, with line graphs, often in collections of multiple graphs showing different aspects of the data side-by-side.

5.1 When are IVs used?

On the whole, the analysts report that they most often use IVs (especially interactive IVs) when *exploring* and coming to understand their data. These sophisticated explorations often require “coming to know the data,” and “spending time with the data” to understand and verify expected relationships and values, as well as discovering unexpected variances in the data, or finding unexpected relationships that become visible only through visualization.

However, as the analysts explore their data sets, much of the experience is interactive—that is, as they brush, rotate, zoom, filter, and subset over the data, their understandings often grow out of those perceptual experiences. While useful for discovery, these interactions are then converted to “presentation graphics” that are quickly and easily understood. There is frequently a non-trivial amount of work in shifting from insights gained through visual analytics to a set of insights that can be presented to the intended audience (usually the stakeholders). While it works well to using interactive IVs to share insights between analysts, the conversion to presentation is a skilled step as well, one that frequently involves a translation to more traditional presentation forms.

The Sunburst visualization (Figure 3) is a good for exploration, mostly used as a way to discover sequences and proportions. But the insights are gained through actually *interacting* with the slices and the revealed structure within the data [20].

While Sunbursts were initially a tool for data analysts, it has, over the past couple of years, become part of the shared visual language, largely as the result of a small group of analysts who invested the time and energy to make this particular IV into an accepted *presentation* format. It did not happen overnight nor accidentally, but over time with a growing socially shared understanding of what this diagram represented and how that data could be used.

5.2 Social aspects of using an IV

As the Sunburst IV shows us, IVs are used for exploration and sometimes as presentations—but growing the visualization literacy of an organization is a slow process.

Sequences sunburst

Figure 3: A Sunburst diagram is used interactively for research and exploration. Brushing across different arcs opens up subfractions of data that share a common ancestry. A Sunburst is useful both as an investigation tool and a presentation tool. [20]

The value proposition must be clear (what does this reveal in the data that isn't possible via another visualization?), and the

organization as a whole must understand how it represents the data (i.e., in general the consumers understand what does an arc mean? what data is being excluded? how does the proportion change?).

Very often, large data sets with many columns and relationships are examined in teams; that is, to understand extremely large data sets, a social understanding process is at work. Tables such as

Figure 4 are usually the target of a group of analysts working together to interpret what's going on.

In essence, for large data sets, there are multiple, simultaneous views onto what the data is saying; simply understanding what the relationships are is complicated (this is especially true in a system that has involved data provenance, with many different data sources being integrated and analyzed as a whole.).

Experiment	Variable 1	Variable 1 (%)	reside	link-over	cone-true	Width-size	HeightMismatch	HeightMismatch (%)	HeightMismatch (%)	VervMajor
20492	15.9005	14.0665	0.9574	4.5047	5.3148	3.2896	0.0000	33.0412	17.5758	2.5293
	-19.34%	-9.90%	-17.89%	-13.95%	-2.14%	-12.98%	-100.00%	-12.34%	150.47%	-18.59%
20492	-37.97, -0.70] %	-16.74, -3.07] %	-01.18, 15.40] %	-24.52, -3.37] %	-13.76, 9.47] %	-26.75, 0.78] %	-100.00, -100.00] %	-16.87, -7.81] %	120.61, 180.34] %	-34.32, -2.86] %
20492	20.4018	1.8658	0.2925	0.2995	0.8477	0.4261	5.6037	4.6867	1.9057	0.1558
20492	20.0329	1.8263	0.2776	0.2854	0.8357	0.4277	5.3849	4.6886	1.9100	0.1496
	-1.81%	-2.12%	-5.11%	-4.71%	-1.42%	0.37%	-3.90%	0.04%	0.23%	-3.99%
20492	-7.96, -4.29] %	-3.47, -0.70] %	-10.29, 0.08] %	-9.29, -0.14] %	-3.67, 0.64] %	-2.91, 3.65] %	-7.41, -0.40] %	-0.42, 0.49] %	-0.96, 1.41] %	-7.15, -0.82] %
20492	9.9747	15.8509	0.5838	2.5298	9.9003	2.8460	5.5937	29.0082	9.4038	1.0421
20492	8.5455	14.6667	0.6303	2.0121	9.6485	2.3758	0.0000	27.9909	18.6530	1.1187
	-14.33%	-7.52%	7.97%	-20.46%	-2.54%	-36.89, 3.84] %	-100.00%	-3.51%	96.47%	7.35%
20492	-40.26, 11.60] %	-16.86, 1.81] %		-46.07, 5.74] %	-13.49, 8.40] %			-9.63, 2.61] %	71.26, 121.69] %	
20492	45.7305	2.7236	0.6758	0.7490	0.8832	0.4155	9.6486	7.6944	0.4748	0.6969
20492	46.7347	2.7739	0.6741	0.7709	0.9096	0.4192	9.5654	7.7542	0.4861	0.7037
	2.20%	1.85%	-0.25%	2.93%	2.99%	0.87%	-0.86%	0.78%	2.38%	0.98%
20492	-1.67, 6.08] %	-1.99, 5.68] %	-5.73, 5.22] %	-1.72, 7.58] %	-2.15, 8.13] %		-6.63, 3.30] %	-0.10, 1.65] %	-1.25, 6.01] %	-1.56, 3.51] %
20492	17.6663	28.8416	1.2158	15.1976	0.3377	12.0905	14.4178	44.3088	0.3299	2.7054
20492	16.6107	24.9335	1.0638	12.7660	0.3989	10.7048	0.0251	36.0285	12.8719	2.6520
	-5.98%	-13.55%	-12.50%	-16.00%	18.12%	-11.46%	-99.83%	-18.68%	3.801.48] %	-1.97%
20492	-36.16, 24.21] %	-27.80, 0.80] %		-38.25, 6.29] %		-33.98, 10.66] %	-100.42, -99.23] %	-27.10, -10.27] %		-47.04, 43.10] %
20492	52.1440	3.8135	1.1890	1.6162	0.6522	0.3561	2.7555	3.5381	4.6973	0.9662
20492	53.8552	4.0143	1.2943	1.6676	0.6779	0.3545	3.1561	3.6689	4.8580	1.0323
	3.28%	5.28%	8.85%	4.42%	3.94%	-0.45%	14.61%	3.70%	3.42%	6.85%
20492	1.75, 4.78] %	4.07, 6.46] %	7.08, 10.62] %	2.51, 6.33] %	2.11, 5.77] %	-2.81, 1.91] %	8.88, 15.34] %	3.00, 4.39] %	2.86, 3.88] %	5.38, 6.31] %
20492	57.1321	7.5509	2.0736	3.8146	1.1933	0.4695	11.4079	47.8435	6.7698	8.1529
20492	41.1024	6.5925	1.4297	3.6537	1.1517	0.3574	0.0000	39.4171	18.5064	5.9745
	-29.06%	-12.69%	-31.05%	-4.22%	-3.48%		-100.00%	-17.61%	173.37%	-26.72%
20492	-67.35, 1.23] %	-31.60, 6.21] %	-77.95, 15.85] %	-23.58, 15.14] %			-100.00, -100.00] %	-21.74, -13.49] %	119.36, 227.37] %	-45.57, -7.86] %

Figure 4: Simple data tables, often in large collections extending for many screenfuls both horizontally and vertically with some secondary color coding, dominate the IV landscape. (This chart is visually identical to actual tables used, but the data is synthetic).

In such teams, not only do the data experts have to merge their interpretations, but they also need to be able to talk (and interact with) a visual presentation of the data. Thus the data table can be both an IV *investigation* tool and a *presentation* tool. But an important requirement is that everyone in the social network of analysts and stakeholders do NOT see the visualization, but rather, they see only the data, its structure, and the relationships that are operationally important. Anything that gets in the way of that immediate understanding is something to remove. Information Visualization is really a social activity, with a shared basis of visual literacy as a common requirement.

6. SUMMARY

It's clear that IV is a shared activity, and one that changes over time. The Session Viewer story shows that small changes in the cost of use can have an important effect on the uptake and use, with small variances in the cost structure of use making large changes in sensemaking behaviors. [18, 21]

The SessionView and Sunburst stories also shows that the culture can change to incorporate additional visualization basic display patterns over time. Sankey diagrams are starting to make this transition into a broadly understood visual language, although other kinds of visualizations (for instance, dendrograms or spider diagrams) are not yet part of a shared visual literacy.

This analysis shows the frequency of IV use within a particular domain (big data analysis task within a large technology company); of course these chart types will shift with the domain. A GIS company will doubtlessly use more maps, just as analyses in eyetracking will tend to use more heatmaps. Each analytical culture has its own ways of analyzing, and presenting their data.

Overall, analyzing large amounts of data requires more than just pulling together a cleverly designed visualization—the problem is more complex than that. Not only do the cultures of analysis and decision-making have to agree on a visual language, but the overall costs of inventing a new and novel IV (vs. using an existing IV) and creating a specialty presentation that can be used and understood need to be considered as well.

It's not that visual analytics and sophisticated IV systems aren't used, the big point is that they're not used as much as you might expect across the analytic landscape.

Innovations, especially ones that need to move a culture by changing visualization literacy and visualization creation skills may take a long time to make the leap from research into day-to-day production use.

7. ACKNOWLEDGMENTS

Thanks to all of the Google analysts who graciously gave of their time and insights about their patterns and practices in working with data. I'm especially grateful to my early readers (you know who you are) who gave valuable feedback.

8. REFERENCES

- [1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., & Ghemawat, S. (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv preprint arXiv:1603.04467.
- [2] Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. Knowledge and information systems, 1(1), 5-32.

- [3] Corbett, J. C., Dean, J., Epstein, M., Fikes, A., Frost, C., Furman, J. J., & Hsieh, W. (2013). Spanner: Google's globally distributed database. *ACM Transactions on Computer Systems (TOCS)*, 31(3), 8.
- [4] D3. <https://github.com/mbostock/d3/wiki> Last access on May 1, 2016.
- [5] Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107-113.
- [6] Gesmann, M., & de Castillo, D. (2011). Using the Google visualisation API with R. *The R Journal*, 3(2), 40-44.
- [7] Ghemawat, S., Gobioff, H., & Leung, S. T. (2003, October). The Google file system. In *ACM SIGOPS operating systems review* (Vol. 37, No. 5, pp. 29-43). ACM.
- [8] GPDE. Google Public Data Explorer. <https://www.google.com/publicdata/directory> (last accessed: May 1, 2016) For the interactive Bubble Chart of , see: <https://goo.gl/K4fvSw>
- [9] GViz. Google Visualization Toolkit API (2015) https://developers.google.com/chart/interactive/docs/reference#constructor_1 Last accessed on May 1, 2016.
- [10] Hernández, M. A., & Stolfo, S. J. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery*, 2(1), 9-37.
- [11] Lam, H. (2007) Visualization of large data sets to support exploratory data analysis. Ph.D Thesis, Univ. of British Columbia, Canada.
- [12] Lam, H., Russell, D., Tang, D., & Munzner, T. (2007). Session viewer: Visual exploratory analysis of web session logs. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on* (pp. 147-154). IEEE.
- [13] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., & Carpendale, S. (2012). Empirical studies in information visualization: Seven scenarios. *Visualization and Computer Graphics, IEEE Transactions on*, 18(9), 1520-1536
- [14] Mackay, Wendy E. "Triggers and barriers to customizing software." (1991) *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM.
- [15] Melnik, S., Gubarev, A., Long, J. J., Romer, G., Shivakumar, S., Tolton, M., & Vassilakis, T. (2010). Dremel: interactive analysis of web-scale datasets. *Proceedings of the VLDB Endowment*, 3(1-2), 330-339. Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.
- [16] North, C., Saraiya, P., & Duca, K. (2011). A comparison of benchmark task and insight evaluation methods for information visualization. *Information Visualization*, 1473871611415989.
- [17] Periodic Table of Visualizations. http://www.visual-literacy.org/periodic_table/periodic_table.html Last accessed on May 1, 2016.
- [18] Pirolli, P., & Card, S. (2005, May). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis* (Vol. 5, pp. 2-4).
- [19] Plaisant, Catherine. "The challenge of information visualization evaluation." *Proceedings of the working conference on Advanced visual interfaces*. ACM, 2004.
- [20] Rodden, K. (2014). Applying a Sunburst Visualization to Summarize User Navigation Sequences. *Computer Graphics and Applications, IEEE*, 34(5), 36-40.
- [21] Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993, May). The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems* (pp. 269-276). ACM.
- [22] Viégas, Fernanda, Martin Wattenberg, Jack Hebert, Geoffrey Borggaard, Alison Cichowlas, Jonathan Feinberg, Jon Orwant, and Christopher Wren. "Google+ ripples: A native visualization of information flow." In *Proceedings of the 22nd international conference on World Wide Web*, pp. 1389-1398. International World Wide Web Conferences Steering Committee, 2013.