Big Challenges for Visual Analytics

Assisting sensemaking of Big Data with Visual analytics

Mikkel Blytt
Department of Product Design
Norwegian University of Science and Technology

ABSTRACT

Big data rises with the growth of data and information. It refers to the practice of storing and analysing the vast amounts of data that is forever being generated, most of it from the Internet, computers or sensors. Sensemaking is the process of making sense of the data and gaining valuable insight; this is a multifaceted challenge. Visual analytics is based on visualizing data so that interesting relationships can be discovered. It is a promising method for sensemaking of big data, but there are big challenges it needs to address. This paper reviews literature from both fields to identify these challenges, and what answers have emerged to deal with them.

KEYWORDS: Big Data, Visual Analytics, Sensemaking, Enabling technologies

1. INTRODUCTION

“Big Data” is the name given to the fairly fuzzy technological trend of processing incredibly large amounts of data. This data is made accessible and often produced with today’s technology, but it is very challenging to capture, store and use it productively from both a technological and human perspective. On the Gartner hype cycle, which represents the hype versus maturity of new technologies, big data is at the very top, called the “peak of inflated expectations”, and on the verge of falling down into the “trough of disillusionment”, as it is being adapted without the necessary expertise or considerations to the many challenges it brings with it (Gartner, 2013).

Sensemaking is one of the big challenges. Many methods are used to attempt to draw insight from the large and complex data. These methods include advanced data mining algorithms, data visualizations, and statistical analysis. But our capacity to collect and store data is outgrowing the capacity to analyse it (Keim et al., 2006). This paper examines visual analytics as a method for sensemaking of big data. This is the practice of visualizing data in such a way that interesting relationships can be discovered (Begoli and Horey, 2012), this augments our cognitive and visual skills with the strengths of automated computing. It is a promising method, but it has to adapt to the challenges posed by big data, which will require new approaches (Gould, 2013).

This paper reviews literature dealing with the human challenges arising at the intersection of the two fields, big data and visual analytics. As such it takes the "human perspective” on big data and visual analytics, and does not emphasize the many technological challenges. The goal of this paper is to identify how human barriers to sensemaking of big data can be breached with visual analytics. Can visual analytics improve our capacity to analyze big data, and thereby help create understanding and value? Are both fields in agreement of what constitutes the main challenges, and how are they being solved?

1.1 A note on definitions

Big data is not a precisely defined term, there is no set boundary for what constitutes big as opposed to "normal" data. The primary definition of big data, reflected by Wikipedia's selected definition, is Gartner's (2013) definition. They define big data as:

> *"[…] high-volume, high-velocity, high-variety information assets that demand cost-effective innovative forms of information processing for enhanced insight and decision making."*

Some papers eschew a formal definition of the term, but they all capture the three V's (volume, velocity, variety) as a central part of big data. The V's will be covered in more detail later as they are what sets big data apart. The second important definition is that of sensemaking. This is defined as understanding information or gaining insight, and it is the goal of data analysis. To put it in context, it is the point where one's understanding of a dataset enables better decision-making (MacGregor, Scott and Wallace, 2013). To reach this stage, especially for big data, an analyst needs tools, like visual analytics, that support sensemaking.

2. METHOD

This paper is a review article. The findings are based on reviewing a subset of the literature touching on the two topics of big data and visual analytics. Relevant literature was found through the database Scopus, which helped to get an overview and understanding of big data, visual analytics and the intersection of the two. Other media, such as blogs, business whitepapers (IBM, Tableau, SAS, Accenture), and the social media site reddit.com (from the bigdata, dataisbeautiful, visualization, and datascience subreddits) was used in the early stage of research to get a basic understanding of the topics. The method of research contributed to the shape and focus of this paper. The method can be loosely defined into four steps:

1. Review of big data on blogs, social and other media for a basic understanding and overview of the topic.
2. Meta-analysis of the search term "Big Data" in the Scopus database and Google trends.
3. Further searches and search refinement based on the results of the meta-analysis.
4. Reviewing the found literature. This composes the bulk of this paper.

2.1 Meta-analysis of "Big Data" in Scopus and Google Trends

Using search refinement techniques in Scopus helped uncover the overlap between the topics, but the method is dependent on articles being adequately tagged. Based on a personal observation this was not always the case. Often the title, abstract and keywords would not be enough to identify the content of the items, and excluding items dealing with overly technical topics proved difficult. This means the results of the analysis does not align with the author's personal impression (which is that the overlap is of greater significance than shown here). All findings were at the time of writing, and could be subject to change.

Using the search term "Big Data" in Scopus reveals 1525 results, of which 1387 are published after 2011. A majority of these articles deal with computer science topics, often of a very technical nature. Refining the search reveals that "analysis" and "analytics" are keywords in 850 and 352 items respectively (with 231 overlapping both). Google Trends (Google, 2013), which shows search trends over time, also identifies an explosion in the term "Big Data" at the beginning of 2012 and that the top related search was "big data analytics". For the big data articles with "analysis" or "analytics" as keywords, 89 and 123 are also keyworded "visual" and "visualization" (with 49 overlapping both). For reference, a search for "Visual Analytics" reveals 1194 items, with 580 being published after 2011. A search for items with both “Visual Analytics” and “Big Data” reveals 29 items.

The meta-analysis refined the focus of this paper, angling it towards visual analytics. The impression from the literature is that it is an important method for sensemaking of big data.

2.2 Search refining and reviewed papers

After the meta-analysis, additional search terms were added. A final list of search terms, in order of relevance, reads: Big data, Visual analytics, Visualization, Sensemaking, interaction, design, decision-making, information retrieval, information science, clusters, network analysis, multidimensional and multivariate.

A total of 94 articles were considered and classified by relevance, in addition to social media, blogs and industry whitepapers dealing with the same topics. Of these, not all are cited in this paper, but only a subset that seemed to best fit into the general view of the topics (as to avoid too much overlap on areas that are not subject to much discussion). In the following sections, the findings of the literature are presented. The two topics, Big Data and Visual Analytics are explained, with emphasis given to the sensemaking challenges that arise where the two intersect. Furthermore, the findings are also discussed in order to understand the state of the field; how the challenges are related and how they are and aren’t being solved. Again, it is important to stress that these are the “human” challenges, and computational issues, of which there are a lot (as revealed by a Scopus search for “Big Data”), are not covered in this paper.

3. BIG DATA

Data has always been an asset, for uncovering relationships or driving decisions. As such, several methods for making sense of data exist already. So in order to understand why sensemaking of big data is challenging it is important to understand what makes big data different from “small” data, and why it is seen as potentially valuable.

The Gartner definition of big data (Gartner, 2013) helps explain the term, but it doesn’t convey the sense of scale and impact that big data potentially has. Even though a recent trend, big data has ramifications for government, business, and both “soft” and “hard” sciences (Boyd and Crawford, 2012). It has emerged as a natural consequence from the simple fact that there is a whole lot more data available today than ever before (Cukier and Mayer-Schoenberger, 2013). With ubiquitous computing, the growth of data has exploded. In 2000, a quarter of all stored information was nondigital. Today, this number is less than two percent (Cukier and Mayer-Schoenberger, 2013), and data generation does not stop, but increases with technological advancements.

Data Set / Domain	**Description**
Large Hadron Collider (CERN)	13-15 petabytes in 2010
Internet Communications (Cisco)	667 exabytes in 2013
Social Media	12+ terabytes of tweets every day and growing.
Human Digital Universe	1.7 zettabytes (2011) to grow to 7.9 zettabytes in 2015
British Library UK Website Crawl	110 terabytes per domain crawl to be archived
Other	RFIDS, smart electric meters, 4.6 billion camera phones w/ GPS
Reference scale in increments by 10^3: kilo, mega, giga, tera, peta, exa, zetta, yotta	

Table 1: Example of Big Data sources, adapted from Armour et al. (2012).

Some big data sources from Armour et al. (2012) are shown in table 1. Additional examples such as transaction data, clickstream data, genome sequencing, weather data and a variety of scientific simulations are also potential sources of big data (Boyd and Crawford, 2012). As can be seen, several of the examples are types of data that has been used to inform decisions and generate value before big data emerged as a term, and the change from now to then is simply one of scale. This illustrates a central tenet in the literature and practice of big data, here exemplified by a quote from Cukier and Mayer-Schoenberger (2013):

> *"[…] the idea is that we can learn from a large body of information things we could not comprehend when we used only smaller amounts."*

If one ignores the computational difficulties behind capturing and storing the data, big data is about our capability to analyse these complex and large datasets and extract value from them (Boyd and Crawford, 2012). On the surface this is similar to "small" data, but differs in that these high-value insights are now hidden in huge amounts of low-value, scattered data (Liu et al., 2013), or as explained by Czerwinski et al. (2012): distilling terabytes of low-value data into even just a single bit of high-value data.

One change from "small" data is that with big data, the correlations that are drawn across the data are valuable because of the sheer size of the sample. This is view is taken to the point where some advocate ignoring the cause and just look at the correlations (Cuckier and Mayer-Schoenberger, 2013). However, it is the author's impression that this view, which is more sensationalistic, is more popular on popular and social media than the scientific literature.

3.1 The use of big data

There exists no set manual for how to use big data, and it is given from the many fields that have an interest or potential value in using big data that this will vary greatly. What follows are some examples of big data in practice.

- The vice president of Zynga, the company behind the game Farmville, is quoted calling his company an analytics company disguised as a gaming company (Czerwinski et al., 2012). Their use of data analysis to continuously improve their games has resulted in over 40 million active players for Farmville 2 as revealed by a Google search.
- Predicting flu trends. Google uses collected search data to predict flu trends around the world (Google, 2013).
- Government surveillance, by agencies such as NSA constitutes big data, where the data is used to predict citizen behaviour and thereby prevent crime (NSA, 2013)

From these and other examples it is the author's impression that the use of big data could be crudely split into two groups: Targeted use of data to inform decisions, where the data is collected and analysed to find the answer to a specific problem, and exploratory use, where data is collected and explored with the hope that it will yield something interesting or of value.

4. THREE V'S OR MORE

The Gartner (2013) definition touches on three inherent attributes of big data, which frequently referred to as the three V's. These three, volume, velocity and variety, are repeated in nearly all of the literature, and while some papers propose additional descriptors these aren't as pervasive. Each of the V's represents a challenge on its own for sensemaking of big data, but they are also intertwined, which further sets big data apart from "small" data.

4.1 Volume

The volume is amount of data, in bytes, or just the quantity of records, transactions or files in a data set (SAS, 2012). This aligns with the aforementioned idea that by processing large amounts of data, you can identify what you otherwise couldn't. More data means better models, but it comes at cost. For analytics, or sensemaking, the keyword is scalability (Lycett, 2013). Sophisticated technology and techniques are required to analyse such large volumes, but it might not be that the method scales when data is increased by an order or even orders of magnitude (Armour et al., 2012). Visualizations can become cluttered, the efficiency of algorithms peak and even deteriorate, and accuracy to even tenths of a percent could potentially identify far more “wrong” answers than “right” ones as the number of records grow (Armour et al., 2012).

Another issue is that the process of loading data from the database to the analytical software, and then performing operations on it takes a long time. With big data today, the interactivity and fast response expected from “small” data is no longer there (Czerwinski et al., 2012). This means that the analytical workflow has to be different than it would have been with traditional means to reduce the time to extract value from the data.

4.2 Velocity

Velocity refers to the speed of data creation. Data today is rapidly being generated, and data streams (e.g. twitter messages) can be captured continuously. This requires consideration and potential trade-offs as all data does not have an identical lifespan. Lycett (2013) identifies that the granularity of the data, and the acceptable latency are important considerations for managing the velocity of big data. Which data can be discarded, what data has the most critical (short) latency, and where can accuracy be traded for speed?

With constant generation of new data, one also needs to understand the feedback loops that arise (Lycett, 2013). In a nutshell, this is because data is used to inform decision-making, which impacts the new data being created, and the cycle is repeated. This makes real-time analysis of data streams a valuable possibility, but is made difficult because of the large volumes of data that constantly ticks in.

4.3 Variety

Data variety is stressed as the biggest barrier to sensemaking of big data by Armour et al. Big data has high variance, because it combines data from multiple sources, often in multiple formats (SAS, 2012). These sources can often be classified into three broad subcategories: Structured data, which is the type of data that fits into a relational scheme, such as a spreadsheet. This is the simplest type of data to deal with, as it is easy to transform and query. Semi-structured data is similar to structured data, but is considered semi-structured when the data is incomplete or irregular. The last type is unstructured data. This is all data that doesn’t fit well into relational schemes. Social media data, pictures, videos, text and web traffic are only some examples. (SAS, 2012)

As it is, a lot of data is unstructured, but still desirable for analysis. This data can be riddled with errors or inconsistencies, and there is no standardized format for capturing it. Nor is all data is trustworthy, or of value, but this can be hard to differentiate until later in the analytical process. This means that clean up and integration, which is a part of the data value chain that will be touched on later, is important. From a sensemaking perspective, it means dealing with large heterogeneous data sets, where one has to consider an information loss from clean-up, missing or incomplete data, incompatible formats, semantic considerations and inconsistencies (Armour et al., 2012, Lycett, 2013, and Hung, Kasim and Li, 2012).

4.4 Additional V’s

The three V’s are inherent attributes of big data, and they are at the core of why big data presents unique challenges for sensemaking. Conventional methods, and high certitude approaches are often inappropriate (Barth, Bean and Davenport, 2012), and the trade-offs in data quality, response time, and even result quality must be handled explicitly in a way that is not the case with “small” data (Gopalkrishnan et al., 2012).

Additional V's that crop up in the literature are far less frequent than the main three. As an observational note they appear more commonly in big data blogs or business whitepapers. Criticism of additional V's (veracity, validity, volatility or value) is based on the fact that they are simply extensions of the three V's, or not definitional attributes of big data itself.

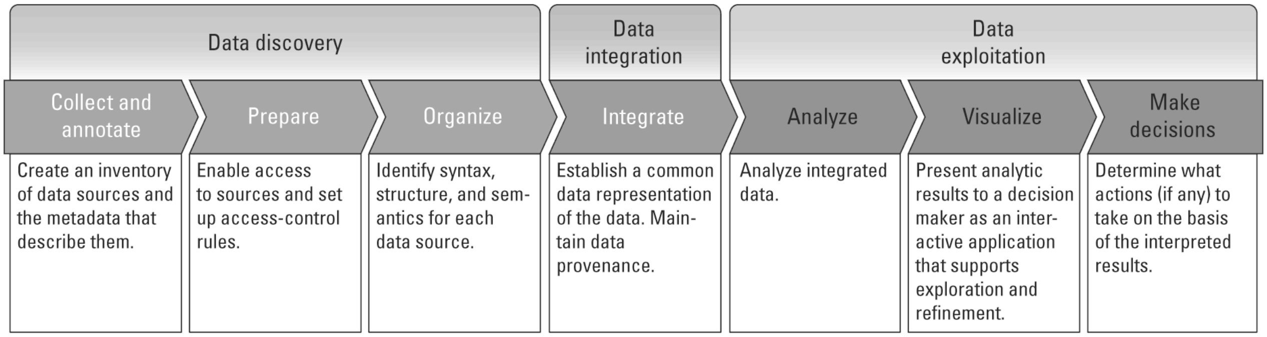
Data discovery
Data integration
Data exploitation
Collect and annotate
Prepare
Organize
Integrate
Analyze
Visualize
Make decisions
Create an inventory of data sources and the metadata that describe them.
Enable access to sources and set up access-control rules.
Identify syntax, structure, and semantics for each data source.
Establish a common data representation of the data. Maintain data provenance.
Analyze integrated data.
Present analytic results to a decision maker as an interactive application that supports exploration and refinement.
Determine what actions (if any) to take on the basis of the interpreted results.

Figure 1: Big Data Value Chain, adapted from Miller and Mork (2013). It is the author's impression that the value chain is widely accepted in the literature.

5. BIG DATA VALUE CHAIN

Sensemaking of big data is the crux. But in terms of working with data, there are many steps that must be taken in order to extract value. Data acquisition, clean up, transformation, analysis and reporting are all important steps that make up the data value chain. The value-chain concept, by Michael E. Porter describes a series of activities that create and build value (Porter, 1985). In figure 1, the data value chain from Miller and Mork (2013) is presented. This illustrates where sensemaking fits into the whole process of working with big data and the context in which analysis takes place. The main takeaways from this model, is that there is an implicit feedback loop at play due to the cyclical nature of the process, and that all three stages are faced with the challenges the three V's bring along. The right data needs to be collected, and it needs to be integrated and organized in order for it to be of any value. The importance of this stressed in the literature (Miller and Mork, 2013, Hung, Kasim and Li, 2012, Armour et al., 2012, Czerwinski et al., and Begoli and Horey, 2012), and is deserving of broader attention than it is being given here. After all, the sensemaking process is of little value if the data one attempts to use isn't capable of yielding the insights or understanding one seeks.

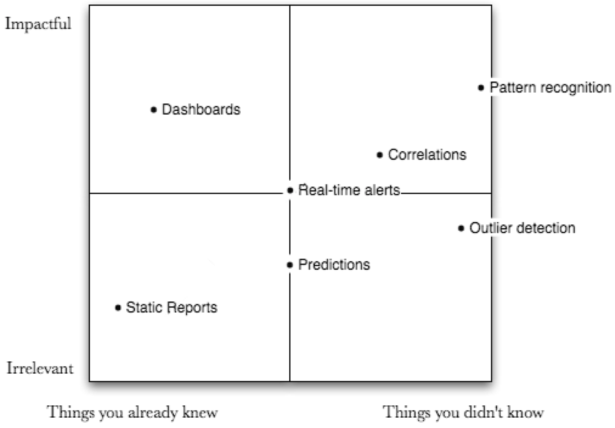

Figure 2: Chart adapted from Begoli and Horey (2012). Discoveries from data can be labelled within the chart. The top-right quadrant is the most valuable.

6. BIG DATA ANALYTICS

As can be seen from the Google trend analysis (Google, 2013) and as an observation from the literature, big data analytics and big data are intertwined. As such, the problems are extensively covered in the literature, often alongside potential solutions. These are often technical in nature, but not all. Some are inherent sensemaking problems requiring consideration for any platform dealing with big data. The goal for any such platform is to enable sensemaking, and thereby extract valuable insights from the data. Figure 2 by Liu et al. (2012) illustrates what one can learn from data, with the goal being to find the connections in the top-right quadrant of figure 2 (impactful things you didn't know).

6.1 Datafication

Transforming the original data into a useful format for analysis is necessary. Variety (4.3) explains how data comes in multiple formats, and not all of it fits in a traditional row and column format. This is the process of datafication, explained by Normann (2001), and is divided into three: dematerialisation, liquidity and density. Summarised, this is our ability to separate the information from the physical world or an asset (dematerialisation), so it can be manipulated and moved around (liquidity), and then recombined for another context (density). From the big data value chain perspective, these three come into play in data collection and integration (dematerialisation), then in analysis (liquidity and density), and decision-making (density)(Miller and Mork, 2013 and Lycett, 2013). This explains what happens to the data in each step of the sensemaking process. Hung, Kasim and Li (2012) also supports that this is central to sensemaking of big data, pointing to the difficulty in transforming large heterogeneous data sets to something that can even begin to be analysed. Armour et al. (2012) calls the concept for "quantitative qualitative computation", and explains that there is a need for better dematrialisation techniques for unstructured text, image and audio information.

6.2 Types of analytics

Big data is multidisciplinary, and is used in several contexts. Manufacturing, retail, healthcare, government, insurance, telecom and energy industries are some examples reaping benefits from big data (SAS, 2012), and these require different and often several, types of analytics in order to extract the desired knowledge from data. Begoli and Horey (2012) and Armour et al. (2012) stress the importance of enabling several types of analysis methods for sensemaking. Statistical analysis, data mining, machine learning and visual analytics all have advantages and disadvantages for sensemaking of data. Depending on how they are used, in terms of descriptive, estimative, predictive and prescriptive analysis they can lead to very different decisions (Armour et al., 2012). This is important to keep in mind as this paper only covers visual analytics. It would be wrong to assume that this is the superior or sole analytical technique; it can only be seen as one of many valuable tools for sensemaking (Keim et al., 2006).

6.4 Enabling users

As we have from Keim et al. (2006), our capacity to capture and store data is outgrowing our capability to analyse it. A similar concern is expressed frequently in the literature, often in relation to another problem, namely the required expertise to deal with big data analytics. Data scientists and data analysts today rely on deep knowledge of computer science, programming and statistics for sensemaking of data (Barth, Bean and Davenport, 2012). Not only is this skill set in short supply, but can also come at a cost in domain understanding. Domain understanding can also be called "specialist knowledge" (e.g. a biologists understanding of the genome). Begoli and Horey (2012) claim that domain understanding is one of the three main elements in sensemaking. It is crucial that analysts have a deep understanding of what they are analysing, and not just the technical proficiency required to perform the analysis as their expertise can't be supplanted by data (Ransbotham, 2012).

However, tools today have yet to reach the point where no specialist knowledge is necessary to perform rigorous analysis of big data. (SAS, 2012) As put by Czerwinski et al. (2012), to enable the "typical excel user" to perform data analysis, has the potential to be of great value. This would, on the face of it, make it possible for the people with domain knowledge to also be able to perform analysis within their respective domains. Though it then remains to see if the insights of non-specialists can be accurate or valuable, especially when the data might be of varying quality (Gopalkrishnan et al., 2012) and of the complexity represented by big data.

6.5 The creative analyst (users, interface and storytelling)

Since big data is not self-explanatory (Boyd and Crawford, 2012), we have a context where people have to process data through an interface. These people need to be able to examine and analyse, and immerse themselves in the data, which is difficult due to the many complexities of big data. This means that the interface is crucial, in terms of how it allows the users to interact with the data it presents. With a good interface and some creativity, the analyst can find and create stories with the data (Dombrowski, Kim and Lund, 2013). This creativity is necessary precisely because traditional sensemaking methods don't scale, and because big data is abstract and hard to grasp. The natural inclination to tell stories helps to contextualize the data, compressing the vast amounts of information into high-value insights (Dombrowski, Kim and Lund, 2013, and Czerwinski et al., 2012). As such, the design of any such interface and its interactive capability is of great importance.

This covers the general challenges for sensemaking of big data from a human perspective. Next it is explained how these challenges transfer and otherwise relate to visual analytics, and also what techniques have emerged to deal with them.

6.6 A note on visualization

The astute reader will have seen visualization in its own category in the big data value chain (figure 1). Visual analytics doesn't overlap the type of data visualization referred to there. Visual analytics covers the use of visualization in the sensemaking process, but visualization in the big data value chain refers to making the results of said analysis accessible and understandable, often for the decision makers in the organization. In other words, communicating the insight gained from the analysis. (Miller and Mork, 2013)

7. VISUAL ANALYTICS

Visual Analytic stems from the field of information visualization. An accepted definition is that visual analytics is "the science of analytical reasoning facilitated by interactive visual interfaces" (Keim et al., 2006). It is a tool that enables sensemaking of information. Other attempts to define visual analytics typically expand on the aforementioned definition, pointing out that visual analytics combines automated techniques along with the interactive visualizations (Keim et al., 2011). It is not the purpose of this paper to extensively cover the techniques and theories it is built on beyond the very basics.

The main strength of visual analytics as a sensemaking tool is that it engages the whole of our perceptual and cognitive capabilities to the analytical process, and augments it with advanced computation. By representing the information visually the cognitive load required to perform analysis is reduced, this enables the user to process and make sense of massive data sets that would otherwise not have been feasible (Keim et al., 2006). As such visual analytics is an interdisciplinary field, much like big data, and it draws from multiple analytical techniques (such as data mining and statistical analysis), as well as interaction design, and cognitive and perceptual science (Keim et al., 2006). None of these tenets are challenged in the literature, and it is because of them that visual analytics is identified as an accepted and promising method for sensemaking of big data (Czerwinski et al., 2012, Li and Qin, 2013, Armour et al., 2012, Begoli and Horey, 2012, Dombrowski, Kim and Lund, 2013, and Gophalkrishnan et al., 2012). Li and Qin (2013) go as far as claiming visualization to be a basic function for both normal and expert analysts, precisely because it can make data to speak to the user at an intuitive level.

7.1 Visual Analytics for Big Data

Though frequently brought up as a valuable sensemaking tool for big data, both fields (visual analytics and big data) identifies that the intersection is not without difficulties. Enabling sensemaking of big data is identified as the main challenge visual analytics is facing as a field, not just from a human perspective but also a technical one (which is not covered here) (Fekete, 2013). Dealing with such complex and vast amounts of data strains both our cognitive abilities and the underlying software. A keyword of special importance is scalability (Keim et al., 2006). As this paper has already covered, it is a central issue with big data that traditional methods for sensemaking haven't scaled to deal with the three V's, making it hard to extract value. Based on the research in the field of visual analytics this is expressed in terms of novel demands on the interactive capabilities, the level of automation, the workflow, and the visual representation of data (Fekete, 2013, Gould, 2013, Begoli and Horey, 2012, Banks et al., and Keim and Schneidewind, 2005). Big data requires that the information seeking mantra "overview first, zoom/filter, details on demand" proposed by Ben Shneiderman in 1996 be updated to better suit today's reality (Keim and Schneidewind, 2005, and Keim et al., 2006 and 2011).

7.2 Working with visual analytics

The concept of datafication is closely related to visual analytics and datafication is also well exemplified by it. Data is taken from its context and placed into a visual framework, where hopefully new insights can be extracted. This process is not trivial, and an abstract outline of the process is shown in figure 3 (Keim et al., 2011). From the figure it can be seen that the process is cyclical as with big data, and builds on iteration. The user needs to interact and engage with the visual analytics tool in order to reach the desired endpoint. If one were to put it into the big data value chain previously show in figure 1 (Miller and Mork, 2013), it would fit in the broad category of data exploitation, whilst also touching on data integration (Miller and Mork, 2013). Within the framework of the visual analytics there is room for several types of workflows, but as the end goal is always sensemaking they can be judged by this as a common metric (Aurelio, 2013).

The art of selecting the right type of visual representation and the visual literacy that is required to extract value from them are topics of research on their own within the field of information visualization (Yau, 2013), but this is closely related to the process of working with visual analytics. In particular with big data being so varied and used in a myriad of fields, there is a need for several types of visualizations. It is not the purpose of this paper to cover them, but instead uncover the challenges they are all facing. What these are, and how they are connected is explored in more detail next.

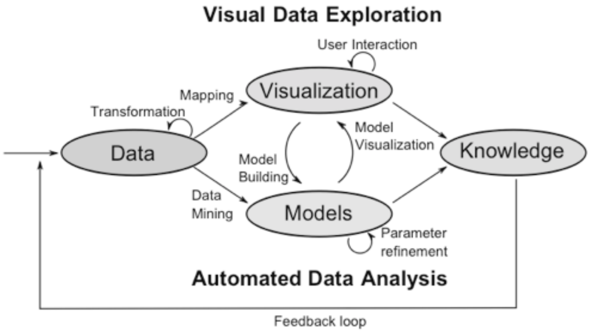

Figure 3: The visual analytics process, adapted from Aurelio (2013). It builds on a cycle interaction between the user, the data and the visual representations.

8. CHALLENGES FOR VISUAL ANALYTICS

Big data is different, and the three V's and other sensemaking challenges presented earlier also apply to visual analytics. As are the demands for big data analysis, as covered in Chapter 6: Big data analysis. The underlying premise of big data means that the potential value would disintegrate if one had to break up the data and analyse it all in separate. (Boyd and Crawford, 2012).

The issue of scalability has been mentioned, but to understand what it actually means we also have to consider the three V's in the context of visual analytics. The following areas are those in which the field of visual analytics have identified sensemaking challenges for dealing with big data: information space, visual representations, interpretability, multidimensional data, network data, workflow and real time analysis. All of the challenges and how they relate to the three V's of big data are summarized in table 3.

8.1 Space and visual representations

Keim and Schneidewind (2005) conclude that several visual representations don't scale well. Scatterplots are a prime example of a representation that quickly loses value as data exceeds a certain threshold. The two limiting factors for our ability to process visual representations are the human perception and the available display area. Shneiderman (Czerwinski et al., 2012), whom is frequently cited within the field of visual analytics, defines big data as the amount of data that would exceed the available pixels of the display at a 1:1 representation (one item of data per pixel). This relates directly to the volume characteristic of big data, and it is the case that the available data can easily surpass the available space in orders of magnitude as shown in table 1.

This strains the aforementioned information seeking mantra, where the overview becomes difficult to grasp as the display of all the information exceeds the visual area. Scaling the visualisation down would represent a loss of information, and showing parts at the time strains our cognitive ability. Finding patterns over several visualisations is almost impossible (Keim and Schneidewind, 2005). Even if one were to scale up the available area, it might be that the perceptual ability still can't handle the full representation of data (Fekete, 2013). Figure 4 is a sample image taken from a blog outlining the scalability problems of the scatterplot. In this case, density mapping could solve the tangle but might come at the cost of hiding outliers in the data if not done with care. Similar issues and trade-offs are also to be found in other types of visualizations, and is why multiple visualizations are often required which again impacts the workflow.

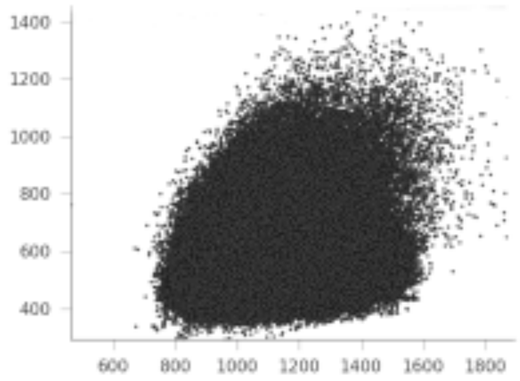

Figure 4: A Scatterplot with more than 2 million points, an example of scalability challenges for traditional visual representations. (micropore.wordpress.com).

8.2 Interpretability

Related to the growth of unstructured data, and the often-varying quality of the data, visual representations need to take into account new measures that previously have been unnecessary. Datasets with data that is often missing, faulty or inconsistent (Armour et al., 2012, Lycett, 2013, and Hung, Kasim and Li, 2012) represent a challenge to visual analytics. Representing uncertainty in the visual structures will be necessary to reach the right conclusions. The guidelines from prominent people within the field of information visualization, Edward Tufte and Manuel Lima in particular, do not address this potential for missing or faulty data (Gould, 2013, and Keim et al., 2006). If not handled by the visual analytics platform, the user will be strained and have to divide the attention available for analysis to correct for such uncertainties.

8.3 Multidimensional data and network data

Looking at single dimensions is often no longer sufficient, but visualizing multiple dimensions rapidly increases in difficulty. The volume and variety characteristics of big data do however mean that this is a requirement, and as such can't be neglected. (Armour et al., 2012). Breaking data down into fewer dimensions per visualization results a similar problem as described under space and visual representations (8.1), namely losing the ability to find hidden patterns across the data (Fernstad, Johansson and Shaw, 2012). This particular issue is not solely a big data issue, and as such several techniques for dimensionality reduction do exist. This again compromises the "overview first" part of the information seeking mantra, as it removes structures of potential importance.

Network data is also a growing challenge. Techniques for representation of relational or network data, relying on graph theory, make up a notable subset of the literature that combines visual analytics and big data. The Internet and network applications create enormous dynamic networks with incredibly complex structures. (Ma and Muelder, 2013) These quickly become unintelligible "hairballs" without careful consideration of the visual representation. These visualizations do however rely heavily on complex algorithms to be generated. This means they represent a very technical challenge, and go beyond the scope of this article. It must be stated that in spite of this it is clear from the literature that networks are an important area of visual analytics for big data.

8.4 Workflow and real-time analysis

The visual analytics process has room for several workflows, and the workflow must be chosen depending on the problem faced by the user. (Aurelio, 2013) The constant generation of new data, or velocity, often implies that the acceptable time it takes to extract value from data is variable (Lycett, 2013). This is something that must be considered in the workflow, as is the implicit potential for real-time analysis (Barth, Bean and Davenport, 2012, and Dombrowski, Kim and Lund, 2013). The representation of continuously changing data streams in a way that is graspable presents a difficult challenge for visual analytics, especially if one is to retain the interactive capabilities that are necessary for sensemaking. The next segment examines commercial examples and other potential solutions from the literature.

Visual Analytics Challenge	**Three V's of Big Data**	*Notes*
System scalability	Volume, variety, velocity	*Underlying challenges for analysis*
Available Information Space	Volume	*Data exceeds available pixels in orders of magnitude*
Visual Representations	Volume and variety	*Traditional visualizations, like scatterplots become unintelligible*
Interpretability	Variety	*Faulty, missing or inconsistent data needs to be considered to avoid errors during analysis*
Multidimensional data	Variety	*Drawing insights across many dimensions requires new techniques for dimensionality reduction*
Network or relational data	Variety and volume	*Graph drawing quickly become unintelligible hairballs*
Workflow and real time analysis	Velocity	*Dynamic data, how to create representations of changing and new data*

Table 3: The author's take on big data challenges for visual analytics, summarized for convenience.

9. PROPOSED SOLUTIONS IN LITERATURE

Both the literature and existing commercial solutions for visual analytics offer potential solutions for sensemaking of big data. However, it is the author's impression that the field is divergent with a multitude of different approaches, and no unified direction exists at the time of writing. Workflow, interaction, novel view hierarchies, novel dimensionaliy reduction techniques and collaborative analysis are some of the areas covered in the literature for enabling sensemaking of big data with visual analytics. By looking at these and the commercial solutions that are available the current state of the field becomes clearer.

9.1 Workflow

The workflow proposed by the information seeking mantra is challenged in the literature. Fekete (2013) criticises its use in a big data context because it was for being built around smaller amounts of data. Using the same approach for big data makes the time it takes to analyse the larger volumes very impractical. To counteract this, it suggests a revised workflow. They argue that it should be designed around the available interactions (e.g. operations such as brushing, linking, filtering or summarizing) to ensure that the analysis is done in the most efficient manner (Fekete, 2013). Liu et al., (2013) suggests another approach. They criticise the current visual analytics platforms for being built around the data as opposed to the end user. This is in line with the suggestions from Dombrowski, Kim and Lund (2013), where the analyst is at the centre of the workflow, and not the data. Neither paper presents a firm guideline, but instead encourage further research into a more user centric workflow for visual analysis. Here again it is also mentioned that this approach has the potential to enable non-specialists to do their own analysis (Liu et al., 2013).

9.2 Interaction

Interaction is closely linked with workflow. There are two types of interaction in the context of data analysis: the operations or controls available to the user and the more abstract interaction with the user and the problem space. (Arias-Hernández, Green and Wakkary, 2011) Emphasis purely on the visual representation of data, at the neglect of interactive capabilities hurts the ability to explore and extract value from data (Arias-Hernández, Green and Wakkary, 2011). Instead both of these also emphasise the need for user-centred research. Jacko et al. (2007) combines the user's intent and the relevant operation. This results in seven broad categories of operations, based on the user's intent. The categories are: select, explore, reconfigure, encode, abstract, filter and connect; these make up the required interactive capability of a visual analytics platform.

9.3 Information space and multidimensional data

The ability to identify patterns across data that is broken up is so diminished; this has resulted in several suggestions for displaying large volumes of data while still maintaining an overview. Keim and Schneidewind (2005) present a scalable multi-resolution visualization. By presenting a hierarchy of views, where the data is presented at different levels of detail, the overview of the whole dataset is partially maintained while promising areas can be shown in greater detail. This is in accordance with the information seeking mantra (details on demand), but avoids removing the visual representation from the context, the overall data structure. An example of this approach is shown in figure 5.

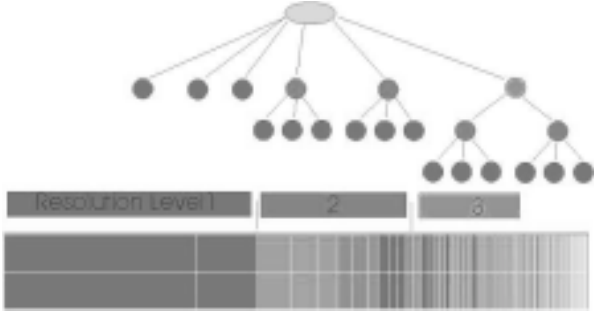

Figure 5: The multi-resolution approach showcased (Keim and Schneidewind, 2005). Overview of the data is maintained even as the user increases the detail of a subset of the data.

In order to avoid a similar issue, Fernstad, Johansson and Shaw (2012) provide a novel method dimensionality reduction. The method is based on displaying an overview of the multidimensional data, and then allowing the user to interactively filter or rank the variables before generating a more detailed visualization. This removes less valuable dimensions or undesirable clutter, and avoids the problem where automated processes could remove structures of potential importance or value (Fernstad, Johansson and Shaw, 2012).

9.4 Collaborative analysis

In papers dealing with interaction in the context of visual analytics, the need for collaborative analysis is touched on. Big data has such inherent complexity that it is difficult to for any single analyst to grasp the whole picture. As such, facilitating communication through annotation and sharing of visual structures becomes an important consideration. Such sharing capabilities can enable sensemaking at a greater rate and also of a higher quality (Leftheriotis, 2012, Chang et al., 2009, and Agrawala and Heer, 2007).

9.5 State of the field

The field of visual analytics has accepted that big data represents a challenge, one that requires innovation to solve. (Fekete, 2013, Gould, 2013, Keim and Schneidenwind, 2005, and Keim et al., 2006) It is the author's impression that the implications of big data are explored and that the challenges are well understood, but that few papers deal with explicit solutions. Instead it is more frequently called for research, which can serve as a basis for these solutions, in particular research into user-centred approaches (Chang et al., 2009, Fekete, 2013, Liu et al., 2013 and Jacko et al., 2007). In this same context there are also mentions of the concepts explored here under enabling users and the creative analyst (6.4 and 6.5). The emphasis seems to be on the growing need for more analytical capacity, and a greater focus on the visual analytics workflow rather than the visual representations of data.

9.6 Commercial review

Big data is also big business, and there are several commercial actors providing visual analytics for it. Behrisch et al. (2012) reviews several state of the art commercial systems and examines their capabilities. By joining the results of this literature review with the commercial solutions we can see if the conducted research is aligned with the commercial solutions or if they diverge. Behrisch et al. (2012) concludes from the review that the following issues are still challenging for the commercial systems: handling semi- and unstructured data, advanced visualizations (innovative visual representations), customizable visualizations, real time analysis and predictive analysis. This suggests that there is indeed some degree of unity between the research within the field of visual analytics and the commercial state of the field. All of the mentioned topics, with the exception of predictive analysis, are areas with significant coverage in the literature.

10. CONCLUSIONS

This paper has examined the emergent technological trend big data and looked at how it challenges the traditional methods for sensemaking from a human perspective. Furthermore it explores visual analytics, a promising method for sensemaking of big data, and examines how the challenges apply to it and how they are being dealt with. The paper summarizes and explores literature dealing with both fields and their intersection.

10.1 Big data

Big data is characterised by three V's, volume, variety and velocity and represents a shift in how it is possible to acquire and process information. Ubiquitous computing, the Internet and sensor technology results in a constant generation of huge volumes of varied data, and capturing and analysing this data has led to valuable insights. Sensemaking of this data requires new innovative approaches, and because of this our capacity to store the data is greater than our capacity to analyse it. To solve the challenges propositions in the literature include enabling non-specialists to do analysis (by creating accessible analytical platforms) and to enable new and creative ways of analysis. For this purpose, the field also identifies that the ability to visualize the data, for both analysis and presentation, can be a key player.

10.2 Visual Analytics

Literature within the field of visual analytics acknowledges the sensemaking challenges for dealing with big data. It also identifies how the three V's of big data translate to the field. An important discovery is that several of the theories and guidelines are and have been central to visual analytics may not scale to the levels required by big data. As a response, the field is expanding from its roots, namely information visualization, to also include a more user-centred focus and go beyond how to create visual data representations. Despite being identified, several key challenges remain unsolved or have multiple proposed solutions pulling in separate directions. There is a call for research to deal with these gaps, and the paper also finds that the existing commercial platforms are facing similar challenges.

10.3 In closing

The literature stresses that big data is here to stay and will only grow in scope. What this means is that the ability to analyse and extract value from the data will similarly grow in importance. The technological challenges for the fields have not been covered in this paper, but they only represent one potential bottleneck for sensemaking of big data. As automation only goes so far, human barriers to sensemaking are just as important to understand and overcome.

REFERENCES

Agrawala, M., Heer J., *Design Considerations for Collaborative Visual Analytics*, IEEE, November 2007

Arias-Hernández, R., Green, T. M., Wakkary, R., *Expanding the Scope: Interaction Design Perspectives for Visual Analytics,* IEEE, 2011

Armour, F., Espinosa, J. A., Kaisler, S., Money, W., *Big Data: Issues and Challenges Moving Forward,* IEEE, 2012

Aurelio, D., *Visualizing Information Associated With Architectural Design Variations and Simulations,* HCI International, July 2013

Banks, D., Bryson, S., Haimes, R., Liere, R., Uselton., S, *Automation or Interaction: What's Best for Big Data,* NASA Ames Research Center, Panelist debate, no date given

Barth, P., Bean, R., Davenport, T.H., *How 'Big Data' is Different,* MIT Sloan Management Review, 2012

Begoli, E. and Horey, H., *Design Principles for Effective Discovery from Big Data,* IEEE, 2012

Behrisch, M., Keim, D., Last, H., Mittelstädt, S., Pompl, R., Schreck, T., Stoffel, A., Weber, S., Zhang, L., *Visual Analytics for the Big Data Era – A Comparative Review of State-of-the-Art Commercial Systems,* IEEE, October 2012

Boyd, D. and Crawford, K., *Critical Questions for Big Data,* Information, Communication & Society, May 2012

Chang, R., O'Connel, T. A., Pike, W.A., Stasko, J., *The Science of Interaction,* Sage Publications, 2009

Cukier, K. and Mayer-Schoenberger, V., *The Rise of Big Data: How it's Changing the Way we Think About the World,* ProQuest, October 2013

Czerwinski, M., DeLine, R., Drucker, S., Fisher, D., *Interactions with Big Data Analytics,* 2012

Dombrowski, C., Kim, J., Lund, A., *Telling the Story in Big Data,* June 2013

Fekete, J., *Visual Analytics Infrastructures: From Data Management to Exploration,* IEEE, 2013

Fernstad, S. J., Johansson, J., Shaw, J., *Quality-based Guidance for Exploratory Dimensionality Reduction,* Sage Publications, October 2012

Gartner, online resource, Big Data definition: http://www.gartner.com/it-glossary/big-data (retrieved November 2013) and the Hype-Cycle: http://www.gartner.com/newsroom/id/2575515 (retrieved September 2013)

Google, online resource, Google Trends: http://www.google.com/trends/ (retrieved November 2013) and Google Flu Prediction: http://www.google.org/flutrends (retrieved December 2013)

Gopalkrishnan, V., Guszcza, J., Lewis, H., Steier, D., *Big Data, Big Business: Bridging the Gap,* Deloitte Analytics, 2012

Gould, E. W., *Dot, Line, Network: Helping Individuals Make Sense of "New Data",* HCI International, July 2013

Hung, T., Kasim, H., Li, X., *Data Value Chain as a Service Framework: For Enabling Data Handling, Data Security and Data Analysis in the Cloud,* IEEE, 2012

Jacko, J. A., Kang, Y. A., Stasko, J. T., Yi, H. S., *Toward a Deeper Understanding of the Role of Interaction in Information Visualization,* IEEEE, November 2007

Keim, D. A., Kohlhammer, J., Pohl, M., Santucci, G., *Solving Problems with Visual Analytics,* Procedia Computer Science, 2011

Keim, D. A., Mansmann, F., Schneidewind, J., Ziegler, H., *Challenges in Visual Data Analysis,* IEEE, July 2006

Keim, D. A. and Schneidewind, J., *Scalable Visual Data Exploration of Large Data Sets via MultiResolution,* Journal of Universal Computer Science, 2005

Leftheriotis, I., *Scalable Interaction Design for Colaborative Visual Analytics,* IEEE, November 2007

Li, Z. H. and Qin, H. F., *Research on the Method of Big Data Analysis,* Asian Network, 2013

Liu, Q., Madhavan, K. P. C., McKenna, A. F., Vorvoreanu, M., *Designing Discovery Experience for Big Data Interaction: A case of Web-Based Knowledge Mining and Interactive Visualization Platform,* HCI International, July 2013

Lycett, M., *'Datafication': Making Sense of (Big) Data in a Complex World,* European Journal of Information Systems, 2013

Ma, K., Muelder, C. W., *Large-scale Graph Visualization and Analytics,* IEEE, 2013

MacGregor, C. G., Scott, S. D., Wallace, J. R., *Collaborative Sensemaking on a Digital Tabletop and Personal Tablets: Prioritization, Comparisons and Tableaux,* CHI, April 2013

Micropore blog, online resource, available at http://micropore.wordpress.com/2011/10/01/2d-density-plot-or-2d-histogram/ (retrieved October 2013)

Miller, H. G. and Mork, P., *From Data to Decisions: A Value Chain for Big Data,* IEEE, 2013

Norman, R., *"Reframing Business: When the Map Changes the Landscape",* John Wiley & Sons, 2001

National Security Agency (NSA), online resource, accessible at: http://www.nsa.gov (retrieved December 2013)

Porter, M. E., *"The Competitive Advantage: Creating and Sustaining Superior Performance", NY: Free Press, 1985*

Ransbotham, S., *Why Detailed Data is as Important as Big Data,* interview by David Kiron, MIT Sloan Management Review, 2012

Reddit, social media, online resource, accessed in the period September – December 2013, available at http://reddit.com/r/bigdata and also /r/dataisbeautiful, /r/visualization and /r/datascience

SAS (author(s) not listed), *Data Equity: Unlocking the Value of Big Data,* 2012, copy available at http://www.sas.com/reg/gen/uk/data-equity

Yau, N., *"Visualize This: The Flowing Data Guide to Design, Visualization, and Statistics",* Majorie Weiser Milas Fund, 2011