Visualizations for Real Time Big Data

Studienarbeit

Borja Nicolau ([2928973])

TECHNISCHE UNIVERSITÄT DARMSTADT

ise. Information Systems and Electronic Services

Technische Universität Darmstadt

Fachbereich Rechts- und Wirtschaftswissenschaften

Fachgebiet Wirtschaftsinformatik - Information Systems & Electronic Services

Prof. Dr. Alexander Benlian

Betreuer: Elena Davcheva

Studienarbeit zu dem Thema:

Visualizations for Real Time Big Data

Bearbeitet von: Borja Nicolau

Matr.-Nr.: [2928973]

Studiengang: Wirtschafstingenieurwesen

Eingereicht am: 22.02.2017

Förmliche Erklärung

Hiermit erkläre ich, Borja Nicolau, geboren am 12.07.1992, an Eides statt, dass ich die vorliegende Master Thesis ohne fremde Hilfe und nur unter Verwendung der zulässigen Mittel sowie der angegebenen Literatur angefertigt habe.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Darmstadt, den 22.02.2017

_____

(Unterschrift)

Abstract

We live in a Big Data era in which data are generated every minute at astonishing rates: internet cookies, social networks, all kind of sensors, climate measurements, smartphone GPS systems, etc.

While in the last years the value has moved from possessing the data to knowing how to interpret them, now, with all these amounts of data being generated every minute, the value reside in the ability of interpreting them in real time. This can give a competitive advantage to those able to achieve it, providing them the ability of taking advantage of real time opportunities. Visual analytics, a science combining the strengths of machines with those of humans, can be the solution to that.

This work gives a review of existing literature about visualizations of real time Big Data by discussing the main parameters that must be taken into consideration, examining some approaches to perform effective and efficiently and identifying guidelines to evaluate how good a visualization tool is.

Results have shown that, although this science is still in a non-mature phase and little is written about this concrete case, the implementation of existing strategies – some of them are even obtained from a more generic case –, can be of great help for successfully visualizing real time Big Data.

Table of contents

List of figures

1 Introduction

Nowadays data are produced at unprecedented rates. More and more data are generated in real-time on the Internet – streams, social networks, weblogs, cookies – and by modern equipment or devices – sensors, GPS systems, satellite cameras –. Nevertheless, while the capacity to collect and store new data grows tremendously, the ability to analyse them increases at much lower rates (Keim et. al 2008).

If analysis is applied appropriately these data can provide very rich information and therefore, improving analytical capability to handle such data is a development opportunity in current business (Zhang et. al, 2012).

Handling these amounts of data in real time, meaning handling them at the same time that they are being produced or that they are constantly changing, poses a major challenge to Big Data science, as it implies making the analysis and the decision-taking process in a continuous changing environment and with a time constraint. Sometimes processing data in real time is the unique option of treating them in a useful way, as the answers or decisions should be taken at the same moment. Moreover, it can give companies a crucial competitive advantage. In this special environment, traditional tools may not be capable enough to perform efficiently and get useful information out of a complicated data set (Wang and Meister, 2010).

Visual analytics, a science to visually analyse large amounts of data, is presented as a possible solution for that. It merges the strengths of computers in processing data mathematically and statistically, with the ones of humans in quickly gain insight through a visual analysis. However, while a lot has been written about Big Data and also about Visualizations or Visual Analytics, very little is written about the concrete case of applying visual analytics to interpret Big Data in real time.

This work pretends to shed light on visualizations of real time Big Data by examining previous literature about the topic and also reasoning what could be useful from existing literature about more general areas, such as Big Data or Visual Analytics alone.

The structure of this article is organised as follows:

After this introduction, chapter two introduces the basic concepts for this work: definition and main problems of Big Data and visual analytics are given, to then introduce the concrete case of visualizing real time Big Data and give some brief introductory examples.

Succeeding that, the important visualization parameters are discussed. It will be treated how factors such as movement, transitions, speed, interaction, screen or prediction can influence in the analysis process and it will be suggested which things have to be taken into account to successfully handle with them.

Important visualization parameters' chapter is followed by an introduction to some abstraction techniques – techniques that pursue offering simplified representation of the whole to improve the analysis –.

Chapter 5 offers some concrete approaches that can be useful to visualizing real time Big Data, such as lowering computing precision, event based visualizations or how to maximize the data set analysed.

Then, some concrete visualization methods are introduced and some clarifications about them when representing dynamic data are made.

At last, chapter 7 is dedicated to discussing how to evaluate a visualization tool through a series of evaluation scenarios, in order to know not only how good is it performing, but also what is user's perception about or which things could make the tool even better.

2 Basis

2.1 Big Data

"Computers teach you something important: it does not make sense to remember everything, what is useful is being able to find things on it."
Douglas Couplant

Big Data has become one of the major topics in the last years. What a few years ago seemed to be a specific and technical issue is now a subject about which everyone, at least, has heard.

Looking at Google trends one can observe the evolution of the term "Big Data" searches in the whole world since 2004 through the biggest search's portal.

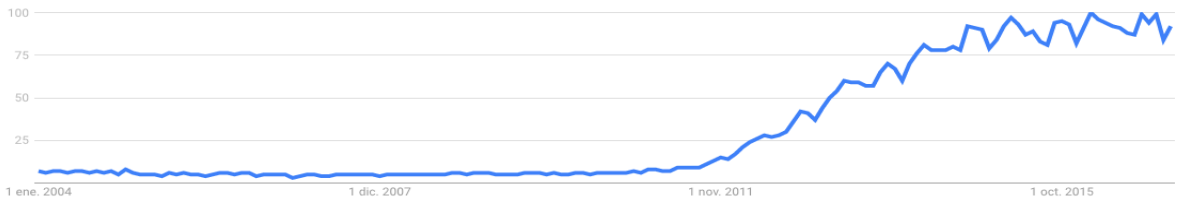

Figure 1: Evolution of "Big Data" searches in Google since 2004

Figure 1 shows the popularity of the term between 0-100, meaning 100 the maximum popularity in a concrete region and time. As it can be seen, the growth has been almost exponential, reaching the top of popularity in the last years.

A data creation explosion occurred in the beginnings of this decade. According to IBM, in 2012 2.5 quintillion bytes of data were generated every day. Moreover, 90% of existing data in 2012 had been created in the previous two years (Zhang et. al, 2012).

A later report by OBS (2014) confirmed this trend as well as showed some astonishing numbers: on the Internet in 2014, 6 articles were published in Wikipedia every minute; 204 million emails were sent; 47,000 smartphone and tablet applications were downloaded; more than 100 new accounts were opened on LinkedIn and 320 on Twitter, where about 100,000 tweets were written every minute; 277,000 logins were made on Facebook; 30 hours of video were uploaded to YouTube and 1.3 million videos were viewed.

Furthermore, the report estimated the trend would continue in growth, saying that in the next four years data traffic would grow 63% in smartphones, 87% in tablets, 30% in notebooks and 113% in M2M devices.

For all that it is said we are in the Big Data era (Keim et. al, 2013; Zhang et. al, 2012), an era in which data is generated at an incredible speed everywhere — satellite images, online transactions, scientific experiments, smartphones applications, GPS signals, social media posts, etc.

In the business world a good data culture is something becoming more and more important every day, as it has great importance in companies in the race to gain competitive advantages and exploit its possibilities in favour of an increase in revenue, a reduction of costs or a greater satisfaction of the user who acquires its products or services (Gallego, 2013).

Big Data analysis pretends to see the big picture and extract conclusions. It's a process that can condense terabytes of low-value data - for example, all financial data of a company in the last decade - intro a single bit of high-value data - for example, should company X acquire company Y? (Fisher et. al, 2012).

But what is exactly Big Data, how is it defined and which are the main factors that characterised it?

2.1.1 Definition

Big Data was defined by Yurevich and Vasilevich (2013, page 1) as *"a large data set, with volume growing exponentially, which can be too large, to raw, or too much unstructured for classical data processing methods."*

While their definition of Big Data spotlights the characteristics of the set, Gallego (2013, page 1) defines it as a science: *"the science that focuses on the treatment of large volumes of information with mathematical and computer techniques and that allows data collection, processing and visualization, obtaining a great velocity in the analysis, being able to anticipate trends, with the objective of understanding and optimizing certain services depending on user's behaviour, to satisfy needs in real time, and to develop first-order strategies in a given sector."*

Many other definitions of Big Data have been done, as it can be seen in Ward and Barker (2013) and De Mauro et. al's (2015) compilations of Big Data definitions.

Oracle gives a definition focused upon infrastructure: "Big data is the inclusion of additional data sources to augment existing operations" and Microsoft treats it as a process "of applying serious computing power - the latest in machine learning and artificial intelligence - to seriously massive and often highly complex sets of information".

The survey by Ward and Baker also shows how even small but really complex quantity of data can be considered Big Data, as the Method for an Integrated Knowledge Environment (MIKE2.0) project did: "Big Data can be very small and not all large datasets are big". This is an argument in favour of complexity and not size as the dominant factor.

After reviewing lots of definitions, De Mauro et. al (page 103) conclude that the consensual definition could be: *"Big Data represents the information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value."*

2.1.2 Main attributes

The last definition given included terms such as volume, velocity and variety. The main attributes of Big Data have also been discussed in previous literature an one of the most agreed visions is precisely the three V's one, first mentioned by Gartner (Yurevich and Vasilevich, 2013). These are volume, variety and velocity.

- *Volume* refers to the quantity of data. As the term "big" per se is not quantitative, a number should be defined to express which volume should data reach to be considered Big Data. Previous researchers have regularly used one million or more data cases as a threshold (Liu et. al 2013)
- Multi-source or multi-format data obtainment make up the *variety* attribute.
- *Velocity* refers to the increasing rate at which data are produced and the need of High Data Processing Speed.

This classification has had good acceptation. However, others have included more attributes such as Value, Veracity, Complexity or Unstructuredness (De Mauro et. al, 2015).

IBM included veracity as a fourth V (Ward and Barker, 2013) related to questions of trust and uncertainty regarding the data. Ward and Barker, after collating their reviewed definitions, proposed a wider classification through three other factors:

- *Size,* referring to the volume and therefore being comparable to Volume.
- *Complexity,* referring to structure, behaviour and permutations. This second factor could include both Variety and Velocity.
- *Technologies,* encompassing the tools and techniques used to process a sizable or complex dataset.

As it can be seen, two first factors encompass the three V's, while the third one extends Big Data characteristics.

2.1.3 Main problems

The main problems of Big Data are resumed in work done by Cox and Ellsworth (1997) and some of them exposed below. In this section only the problems strictly due to Big Data will be treated. Problems more related with visualizations and problems that appear due to the real time constraint will be treated in further sections.

Before introducing Big Data main problems, a distinction between Big Data Sets and Big Data Collections should be made:

- Big Data Sets, also called Big Data Objects, are *"single data sets that are too large to be processed by standard algorithms and software on the hardware one has available. Clearly, a data object too large for one installation may be manageable at another. Big Data Sets typically are the result of large-scale simulations."* (Cox and Ellsworth 1997, pages 1-3)

- The aggregation of Big Data Sets gives as a result Big Data Collections.

Below are main Big Data problems. As it will be seen, most of them are strongly related with the factors exposed above (volume or size, complexity and technologies):

- *Size and memory:* Data may be too big to fit in main memory. Sometimes, even to big to fit in local disk. With this large, it is not possible to rely on virtual memory to manage the difference between data set size and physical memory size. Thus, special actions must be taken.

- *Bandwidth and latency:* Due to data large and to the requirement to find alternatives to operating system virtual memory, the bandwidths and latencies between data store and main memory must be managed carefully.

- *Technologies and data models*: sometimes the multi-dimensional data structures required are not adequately supported by actual database technologies. In addition, sometimes there are no standardized models of the data structures required, or inversely, there are so many standards to choose from.

- *Variety and incompatibility:* Data are generally distributed among multiple sites, many times in heterogeneous databases. In addition, there are incompatible data interfaces and representations. Fox and Hendler (2011, page 706), talking about this as the data-scaling problem, added that the data are not only coming from different places, but also linked between them: *"the challenge is that many of the major scientific problems facing our world are becoming critically linked to the interdependence and interrelatedness of data from multiple instruments, fields, and sources.*"

- *Poor metadata:* Data are generally not self-describing. The metadata, which facilitate discovery and use with information such as where and when were the data collected, what calibration was applied, what are the units, etc., are often not stored with the data. Often there is no independent definition of the data types in the underlying data and the relationship between them, which would facilitate the construction of higher-level tools to use the data. These problems make difficult the data location. Keim et. al (2008) refer to that as "data provenance".

2.2 Visualizations

"Vision is the art of seeing what is invisible to others".
Jonathan Swift.

Vision is our dominant sense, with almost a quarter of our brain devoted to processing visual stimuli (Khairi et. al, 2013). In a few tenths of a second, humans can recognize features in megapixel displays, recall related images and identify anomalies (Ahlberg and Shneiderman, 1994).

Because of this human ability to quickly gain insight through a visual analysis, and because of the exponential increase in data complexity, interest in visualizations and visual analytics has increased during the last years (Choo and Park, 2013). A Google search on "data visualization" led, in 2010, to 1,220,000 links (Wang and Meisner, 2010); nowadays it leads to 8.610.000 results.

In this section, first visualizations will be defined. Then, the concept visual analytics will be introduced and some of its classifications and problems will be exposed.

2.2.1 Definition

Visualization is defined by Manovich (2008, page 127) as *"the situations in which quantified data, not visual itself (for example, data captured by meteorological sensors or the set of addresses describing the trajectory of a message over a computer network) are transformed in visual representations"*

Kornblitt et. al (2000, page 14) use a similar description but refers to it as a variety of techniques instead of as a situation: *"data visualization (sometimes referred to as scientific visualization, or just visualization) is a term applied to a variety of techniques and processes for the representation of, or transformation of, data or information into images— including graphs, pictures, or other graphical forms."*

From both definitions it can be said that visualization, interpreted as a situation or as a process, is the representation or transformation of data to images.

Lam et. al worked in a classification of kinds of evaluations for information visualizations that will be treated in 7th chapter. Interest for this section relies on the stages they proposed for data visualization, which can be considered the phases of visualization. In addition, they suggest which should be the main goal of each one (2012, page 2):

- *Pre-design:* understand potential users' work environment and workflow.
- *Design:* arrange a visual encoding and interaction space based on human perception.

- *Prototype:* check if a visualization tool has achieved its design goals and compare the prototype with the current state-of-the-art systems or techniques.
- *Deployment:* see how a visualization performs - how influences workflow and its supported processes – and evaluate its effectiveness and uses in the field.
- *Re-design:* improve a current design by identifying usability problems.

2.2.2 Visual Analytics

Data analysis is about turning empirical information into conclusions, to reason about data content and show causality. It should be used to describe things and learn about the world (Tufte, 2016). The way of doing it through visualizations is visual analytics, abbreviated as VA.

The field of visual analytics focuses on the treatment of massive, heterogeneous, and dynamic volumes of information by integrating human judgement through visual representations and interaction techniques in the analysis process. Visual analytics combines the strengths of machines with those of humans. On the one hand, statistics and mathematics are the protagonists on the automatic analysis side; on the other hand, human capabilities help in the attempt to perceive, to relate, to gain insight, to draw conclusions and to make better decisions (Keim et. al, 2008).

For all that, it can be said that visual analytics is at the confluence of statistics, machine learning, information visualization, human computer interaction, data management, and memory optimization (Zhang et. al, 2012).

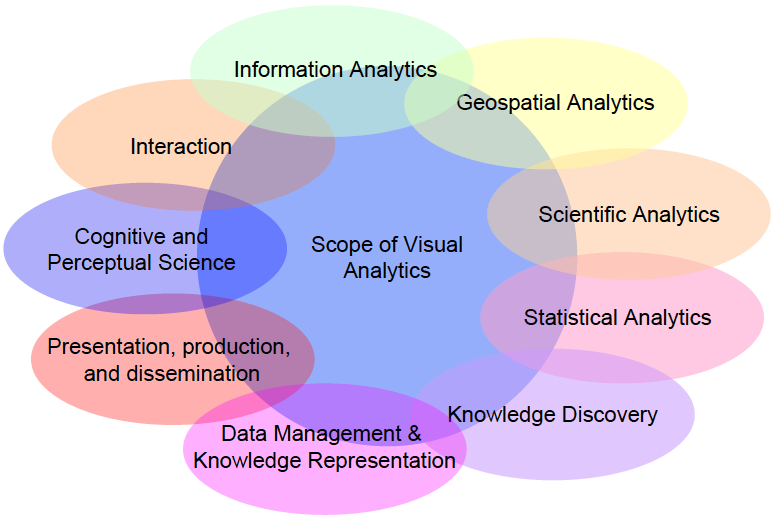

Figure 2: Scope of Visual Analytics (taken from Keim et. al, 2008)

The aim to use visualizations is the need to abstract, concentrate or transform a large and difficult amount of data to manageable, but meaningful, proportions (Kornblitt et. al, 2000). With that purpose, visualization tools can be used to extract information rapidly from files with complicated data structures (Wang and Meisner, 2010). For that, visual properties such as position, size, shape and colour can help in users attempt to discern and interpret patterns in data (Heer and Shneiderman, 2012).

Keim et. al (2008) present a graphic description of the visual analytics iterative process, involving information gathering, data preprocessing, knowledge representation, interaction and decision making.

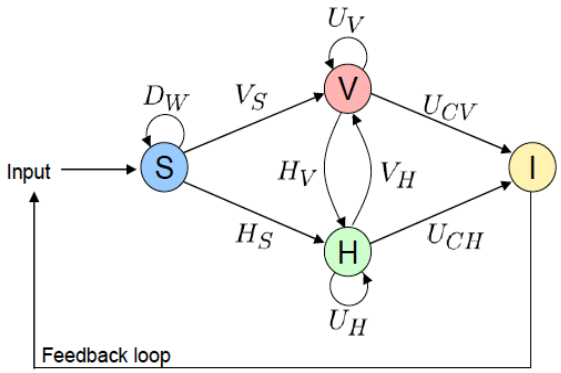

Figure 3: Process of visual analytics Font: (taken from Keim et. al 2008)

S: *Data sets*
V: *Visualizations*
H: *Hyptohesis*
I: *Insight, meaning output of the process.*
Dx: *Basic data pre-processing functionalities, such as transformations (Dt), cleaning (Dc), selection (Dsl) and integration (Di).*
Vx*: visualization functions, which could be visualizing data (Vs) or hypothesis (Vh)*
Hx*: hypothesis generation process. Hypothesis can come from data (Hs) or visualizations (Hv)*
Ux*: User interactions. They can affect only visualizations (Uv) or only hypothesis (Uh). When the user interaction is leading to an insight, it can be thanks to visualizations (Ucv) or to hypothesis (Uch).*

2.2.3 Classifications

2.2.3.1 End Product vs. Exploratory Tool

The first classification is about if visualization should be treated as an end product or as an exploratory tool. End product refers to a visualization as a simple way to represent some information. On the other hand, the exploratory tool vision expands its purposes to the analysis of information, arguing that visualization has proven effective for not only presenting essential information in vas amounts of data, but also driving complex analyses (Keim et. al, 2013).

If the second vision encompasses also the first one, making it wider, no doubts should be about it and visualization should be considered an exploratory tool that also enables user to represent information as a final step. For this work, in which the purpose of visualizations is helping in Real Time Data Analysis, they will be treated as an exploratory tool.

In that way, Doleisch (2007) says that visualization tools can be used to explore, to analyse and to present. In practice, these three goals are no easy to separate and effective visualization tools should offer the three of them.

However, visualization still becomes too often an end product of scientific analysis. This happens, according to Fox and Hendler (2011) because visualization analysis has become a bottleneck due to continuous decrease in price of new technologies for data generation (in terms of cost per data generated), whereas visualization costs are falling much more slowly. However, they pointed that new database technologies, coupled with emerging Web-based technologies, may hold the key to lower the cost of visualization generation and allow it to become a more integral part of the scientific process.

2.2.3.2 Visuospatial vs. Not visuospatial

This differentiation, made by Tversky and Morrisson (2002), says that graphic displays can be divided in two bands: on the one hand, those that portray things that are essentially visuospatial, like maps, molecules and architectural drawings; on the other hand, those that represent things that are not inherently visual, like organization charts or flow diagrams. The first ones have a clear and obvious advantage over others because of the use of the same language between data and representation. This differentiation suggests that it is possible that, depending on data's nature, data visualization would be more difficult to understand. Nevertheless, this natural correspondence does not mean that all this kinds of graphics are immediately understood.

2.2.4 Main Problems

In this section some visualization problems related with Big Data are presented. Some of them will be more widely treated in upcoming chapters.

- *Incompatibility:* Some traditional data visualization tools are inadequate to handle Big Data (Liu et. al, 2013)
- *Visual clutter and human perception:* The high number of data points often makes visualizations too dense to be useful. Getting a larger screen to avoid visual clutter can be inefficient due to human perception constraints (Yurevich and Vasilevich, 2013; Choo and Park, 2013; Khairi et. al, 2013).

- *Information loss:* when using some methods, such as filtering, to avoid visual clutter or to focus in a concrete area for a visualization, it is possible that some important information gets lost during the process (Yurevich and Vasilevich, 2013).

- *Poor adaptability:* Some expert systems have been successfully built for specialized fields, but then only perform reasonably within a limited scope (Keim et. al, 2008). It's challenging to construct integrative visualizations that can simultaneously and effectively work for a variety of data sources (Khairi et. al, 2013).

- *High Performance Requirements:* To treat with Big Data visualizatons special kills are required (Yurevich and Vasilevich, 2013). This becomes even more significant in dynamic visualizations, where data are constantly changing.

- *Too robust designs:* Computational modules in visual analytics can be difficult to understand, depending on the analyst knowledge they can be sometimes even more difficult to understand than the original raw data. Moreover, many visual analytics systems choose a specific computational method, treat it as a black box, and focus on analysis of its output. But if the analysts don't understand correctly the algorithm and its parameters, a computational method might not perform well enough to analyse data properly (Choo & Park, 2013).

- *User acceptability:* many novel visualization techniques have been presented without success, primarily due to the users' refusal to change their working routines (Keim et. al, 2008).

2.3 Visualizations for Real Time Big Data

After introducing the two main concepts of this work - Big Data and Visualizations - this section will introduce the concrete case of visualizing this Big Data in real time, meaning visualizing it at the same time that they are being generated.

2.3.1 Introduction

During the last quarter of the 20th century data visualization has grown into a mature, vibrant and multi-disciplinary research area. From the early 70s to middle 80s, many of the advances in statistical graphics concerned static graphs for multidimensional quantitative data, designed to allow the analyst to see relations in progressively higher dimensions. In the 90s, many ideas were brought together to provide more general systems for dynamic, interactive graphics, combined with data manipulation and analysis. Since then, the developments have greatly increased computer processing speed and capacity, allowing computationally intensive methods to access to massive data problems and real-time streaming data (Friendly 2006).

Many examples are found in literature about dynamic data or dynamic visualizations. Dynamic visualizations are defined as time-dependent visualizations, meaning that they change dynamically over time (Müller and Schumann, 2003). However, this usually refers to data previously accumulated through a period of time and analysed a posteriori. The extra requirement to meet the exact characteristics of this work would be that this change occurs in real time, but both cases can be compared to find similitudes in the main characteristics to consider or the approaches proposed.

Real Time, therefore, added to the terms Big Data and Visualizations, refers to the way of visualizing Big Data at the same time that these are produced or change. That means that data will be constantly changing and consequently the visualization will also change.

Nowadays data streams are producing new data at astonishing pace - internet and social media, monitoring sensors, smartphone's location data, transaction files, etc. – and in their analysis time plays a specially important role because data visualization lifespan can be short due to different things: in many cases detailed information is abundant and in the long-term storage capacities do not suffice to log all data (Keim et. al, 2008); other times, the change frequency itself makes the data visualization time really short. Thus, the value is not anymore the possession of big quantities of information, rather to know what to do with them and how to obtain useful insights from them at the same moment as they are being produced. Accordingly, the ability to make timely decisions based on available data is crucial to business success (Keim et. al, 2013).

When this happens, traditional tools and code writing may not be the best way to extract useful information out of a complicated data set. (Wang and Meister, 2010)

In recent years, books and software packages have picked up the pace to provide users with new platforms for dynamic data visualization (Wang and Meister, 2010). But visualizing time-oriented data is not easy, as it is very difficult to consider all aspects involved. In fact, most methods are highly customized for this reason (Aigner et. al, 2007). Even more, most visual analytics' tools or techniques don't properly accommodate Big Data, and even more don't accommodate for Real Time Big Data (Choo and Park 2013).

One example of problems with visualizations for Real Time Big Data is that in most cases, once the visualization is created, it is no longer tied to the data, so that it becomes an immutable information product. The challenge in these cases is that the visualizations are linked to the underlying data and can change dynamically as the data changes (Fox and Hendler, 2011).

Moreover, time-varying information visualization and analysis often differs considerably from traditional static data representations of fixed datasets. Often, users are not concerned about exact data values (for example, the number of total tweets of the day related with a company

or with an advertising campaign), but rather are interested in comparisons such as how data values evolve in time (for example, how has evolved during the last 3 hours or if there is more activity than yesterday), or in the context of the whole dataset (for example, which share quote has it or how better is it performing compared to competence's advertising campaign). Only a few approaches exist that are capable of representing such time-varying characteristics alterations over time (Vande, 2004).

For all that, visualizations for real time Big Data are still far away from a mature period. The research done for this work has shown the existing scarcity in this concrete field. While there is a lot written about Big Data, about Visualizations and even about Big Data Visualizations, very few articles refer to the concrete case of Real Time Big Data Visualizations. Therefore, in many cases the research method for this work has been reading about a more general aspect, such as Big Data or Visualizations, and extracting from there the useful information for the more concrete case of Real Time Big Data Visualizations.

Before stepping to the next section, one concept that will after appear should be clarified: the distinction between Real Time Big Data and Real Time Interaction. While in Real Time Big Data the representation changes because of changes in the data, in Real Time Interaction the visualization changes due to an action done by the user. Many articles mention "Real Time" referring to the second one and this can lead into a misunderstanding.

2.3.2 Brief examples

One of the most classical examples for referring to visualizations of real time Big Data is social media. Social networks importance in our every day life have become bigger and bigger during the last decade, even changing some of our behavioural modes, and the amount of data generated very minute is astonishing.

One of the changes, for example, is that many people use to see television at the same time as commenting what they see through twitter. In USA 38% of people who have a smartphone admits that usually use social networks while watching TV. This has lead into the birth of "Social TV Analytics", a new discipline specialized in the measurement and analysis of conversations in social networks about TV topics. This gives TV programs real time information about their emissions, allowing them to evaluate the feedback, interact with audience and create an extra value for active viewers. In the last years Twitter has been acquiring data analysis companies to the creation of a standard metric that can offer a real numbers of audience and link TV and Twitter. With these acquisitions twitter has become a SocialTVAnalytics-Company with big projection in the Big Data world (Gallego ,2013).

Another clear example is the use of cookies in internet. Nowadays they provide a huge amount of real time information whose analysis, specially for eCommerce, can be crucial in the business success.

Visualization of real time Big Data cannot only be useful for social media analysis and Internet, but also for many other areas:

The financial market with its thousands of different stocks, bonds, futures, commodities, market indices and currencies generates a lot of data every second. Progresses in the analysis capabilities of these big amounts are really promising (Keim et. al, 2008).

Smart cities are also potential “users” of visualization of real time Big Data. It could help, for example, in monitoring, interpreting and optimizing traffic and parking areas. Beck (2003) talks about Real Time Visualization of Big 3D City Models.

Environmental monitoring can help in improving weather and climate observations and predictions. In extreme situations, new technologies could also being used to deal with disaster and emergency management can help determining the on-going process of an emergency: quantifying the amount of damage, assigning priorities and providing effective coordination in the help. (Keim et. al, 2008).

At last, one concrete example of use is the invention described by Kornblitt et. al (2000) to visualize dynamic data in the field of semiconductor manufacturing processes.

3 Important Visualization Parameters

Neuromarketing lessons from Benartzi & Lehrer (2015) show how nowadays there is so much information at our disposal that sometimes overcomes us. Big amount of information causes inattention blindness, which leads to a decrease comprehension. Therefore, what is useful is not giving lot of information, but getting our attention. The ability of organising or representing the information in the right way, so the user can profit the most from it, has become the core item. And sometimes the way to do it is making it simpler.

A similitude with Big Data visualizations could be made: sometimes there is so many data that people just feel overwhelmed and don't know what to do with it. Even more when Real Time Big Data appears, as they constantly change.

Visualizations can generate optic effects to make the comprehension easier: for example, topographic lines help to visualize 3D; an image with different intensity in the colours can represent difference in concentrations; also representing trajectories in a static image can help in the understanding of a dynamic action (Tufte, 2016).

Visuospatial characteristics of a dynamic visualization such as size, shape, colour and arrangement of its component entities, and also temporal properties such as speed, direction and continuity can affect the relative perceptual salience of displayed information (Fischer, Lowe and Schwan, 2007). Consequently, dynamic visualizations designed in the right way can be strongly effective in data analysis.

Wang and Meisner (2010) used animation and other tools to help discerning clusters and outliers, helping in pinpointing a specific thing or group, or simply breaking down a huge quantity of information into small pieces that are more manageable.

Visualizations can help in all that but to do it in the right way there are key parameters that play an important role. Most of times these parameters appear in literature as visualizations key factors only. The aim of this work is to evaluate them in real time Big Data visualizations. Sometimes what it is said is also valid for the concrete case of study but sometimes not. Therefore, when needed, some clarifications will be made.

The parameters presented in this section are: movement and transitions, speed, interaction, screen and prediction.

3.1 Movement and Transitions

One of the main implications from analysing Big Data in Real Time is that this data will be constantly changing. Thus, visualizations of this data will have movement. In this chapter some studies about benefits and difficulties of animation will be presented, as well as giving some advices about transitions in order to make them the best understandable as possible.

Most of the articles presented below talk about movement and transitions in designed situations, named animations or motions. This is not the identic case of visualizing Real Time Big Data, because the changes will be imposed by the data and not designed by a designer in order to permit a better understanding of the representation. Still, it is important to have these findings in mind because they can help in the process of visualizing real time Big Data, and of course they can be also applied in some cases.

3.1.1 Motion as a filter or brushing tool

Previous research has shown that animation may increase viewer's attention, facilitate learning, decision-making, and increase levels of engagement. Motion can give rise to perceptions of causality, helping in the discovering of cause-and-effect relationships (Heer and Robertson, 2007).

Bartra, Ware and Calvert (2001) studied the requirements for how motion can be usefully applied to visualizations with multiple groups of data objects, in order to be used as a filtering or brushing tool. To do that, they conducted an experiment that investigated the effectiveness of different motions in assisting a visual search task.

Results showed that motion can be effectively used to group different visual elements through a mental process of filtering and brushing but also showed that to be effective, it requires coherence between the elements. Coherence, in this case, meant common frequency and phase.

In all cases, subjects who participated in the experiment said that once the motion started the static icons fell out from the elements to be searched, automatically restricting the search to the moving groups. Contrarily, unrelated moving objects with close timing and similar paths will be erroneously visually associated.

This hints that moving or appearing objects will be the ones that human's eye first notices and relates. Therefore, analysts should have this in mind in order to avoid making wrong impulsive relationships, because random combination of moving icons can cause false perceptual grouping.

3.1.2 Designing effective transitions

This section will be based on Heer and Robertson work (2007), in which they proposed design principles for creating effective transitions and conducted two experiments finding that animated transitions can significantly improve graphical perception, as well as on Simon's and Rensink's article "Change blindness" (2005). At the same time, some nuances will be introduced from the incredulous point of view of Tversky and Morrisson's (2008) work "Does animation facilitate?"

3.1.2.1 Transitions taxonomy

First of all, Heer and Robertson propose a taxonomy of transitions and some recommendations to how to do them. It must be noticed that most of them are due to user interaction, so the change is not directly produced by a change of data, rather by an action from user. Anyhow, this is something that will also happen during the analysis needed while visualizing real time Big Data: visualizations will change automatically because of data changes, but also can change due to user's interaction (as it will be seen in 3.3).

The transition types in Heer and Robertson's taxonomy are:

- *View transformation*. A change in viewpoint such as zooming or panning.
- *Filtering. S*pecifying which elements should be visible and which not. For this transition, they recommend fading items in and out using alpha blending, rather than other more aggressive techniques such as making new points suddenly appear or appear like falling from sky.
- *Visualization change. C*hanges to the visual mappings applied to the data. For example, changing from a histogram to a streamgraph.
- *Data schema change. T*he data dimensions being visualized change. For example, starting from a univariate data chart one might wish to visualize an additional data column, resulting in a number of possible bivariate graphs.
- *Ordering. S*patially rearranging ordinal data dimensions. For example, sorting on attribute values.
- *Time-step. T*emporal changes to data values. For example, a transition between data from the current and previous year. This one is the one more directly related with real time data visualizations. It is possible that time-step transitions require axis rescaling.
- *Substrate transformation. C*hanges to the spatial substrate in which marks are embedded, such as the previous mentioned axis rescaling or log transformations. Heer and Robertson experiments showed that axis rescaling makes change estimation difficult, as they increased overall error and unknown responses. However, the use of animation tempered these effects. To make such changes in a clear way, they recommend axis labels and gridlines move to depict scale changes, again through a fading in and out. For example, when changing from a quantitative to an ordinal scale, old labels and gridlines first fade out and then new ones fade in.

3.1.2.2 Transitions design principles

These are the design principles one should follow and concepts that should be taken in account when building a visualization tool, based on Heer and Robertson (2007) and Simon's and Rensink (2005).

- *Use simple transitions.* Complicated transformations with unpredictable motion paths or multiple simultaneous changes result in increased cognitive load and difficult analysis. Although attention can be distributed to 4-5 items at a time, only a single change can be seen at any moment.

  However, simplicity is not easy to achieve when it comes to Big Data representations. And what is worse, even when motion is simplified perception of motion may not be accurate. Paths of moving objects, for example, are perceived as closer to horizontal or vertical than they actually are (Tversky and Morrisson, 2002).

- *Minimize occlusion.* If objects occlude each other during a transition, they will be more difficult to track, potentially harming perception. This is also difficult to achieve because, as it will be seen, cluttering is one of the big problems of Big Data Visualizations.

- *Use staging for complex transitions.* When transitions are too complex and can't be simplified, one can break up the transition into smaller subtransitions. This allows multiple changes to be easily observed through stages. For example, separating axis rescaling from value changes may help. Heer and Robertson experiments showed that staged animation was significantly preferred to animation and had lower error rates for object tracking. Again, this staging should be made as simple as possible: the results further discourage the use of complex multi-stage transitions.

  The problem in staging is that then animation may be wrongly comprehended discretely (Tversky and Morrisson, 2002)

- *Mainain the area of interest in the center.* Changes to central items are detected faster than changes elsewhere. More concretely, large, fast-moving entities near the centre are more likely to attract attention than small, slow-moving entities near its periphery (Fischer, Lower and Schwan, 2007). To achieve this, automatic view transformations could be applied to the visualization.

- *Maintain valid data graphics during transitions.* The objective of this is to avoid wrong attributions to the data. For discrete time steps that are regular in time and close enough one to another, it is common to assume an underlying data model with continuous time. If this is not possible, it can be done through interpolation. Sometimes linear interpolation is not valid and therefore special interpolation techniques have been proposed to overcome this problem (Müller and Schumann, 2003).

- *Group similar transitions. "The Gestalt principle of Common Fate states that objects that undergo similar visual changes are more likely to be perceptually grouped, helping viewers to understand that elements are simultaneously undergoing the same operation."* (Heer and Morrisson 2007, page 1243)

- *Respect semantic correspondence.* For example, marks representing specific data points should not be reused to depict different data points across a transition.

- *Avoid ambiguity and randomization.* For example, with colors, because studies using suggest that items of similar color can be grouped into a single memory structure. Therefore, if there's no reason, it is better to not represent different data items with same color.

- *Make transitions as long as needed, but no longer.* Previous research recommend transition times around 1 second, although transitions with few changes or short movements can be even faster.

To all that, it must be added that when someone is expert in something tends to detect better changes (for example, american football experts tend to better detect meaninguful changes in football scenes). Therefore, a Big Data Analyst will be much better trained for interpreting each transition in a right way.

The experiments conducted by Heer and Robertson provided strong evidence that, with careful design, animated transitions can improve graphical perception of changes between statistical data graphics.

Tversky and Morrison (2002), however, question animation benefits because they say that the benefits observed can't be directed attributed to animation. They say that, in order to know if animation per se is facilitatory, animated graphics must be compared to informationally equivalent static graphics. Consequently, if animations show more details than static graphics, they are not comparable because there is an advantage for one of them.

They also express incredulity about the benefits from animation pointing out that they are due to other facts, such as interactivity or prediction, which report in benefits in learning by themselves. Both interactivity and prediction, actions that according to Tversky and Morrisson may overcome the disadvantages of complex and discrete-perceived animations, will be treated in upcoming sections.

Once again, it should be considered that all these recommendations are difficult to implement when visualizing real time Big Data because the changes are imposed by data and not decided by the analyst or the visualization tool. Nevertheless, they could be programmed as preferences to, when possible, be applied in data transitions.

3.2 Speed

Movement and transitions have been shown to be an important parameter for visualizations. Of course, the speed at which this movements occur plays also an important role in human perception and understanding of data.

Like in the previous chapter, most studies presented analyse animations in which the speed could be chosen, which, a priori, would not be the case of visualizing Real Time Big Data. Nevertheless, the analyst could chose a transition speed time in order to get a better understanding of what is represented, even knowing that he might be missing some changes in between.

Tversky and Morrisson (2002, page 258) assert that to accord with the Principle of Apprehension (the principle states that the structure and content of the external representation should be readily and accurately perceived and comprehended), *"animations must be slow and clear enough for observers to perceive movements, changes, and their timing, and to understand the changes in relations between the parts and the sequence of events."*

But how much is "slow enough?" Animations that are too slow may prove boring or degrade task times, while those that are too fast may result in increased error (Heer and Robertson, 2007).

Fischer, Lowe and Schwan (2007) conducted an experiment to look into how speed could affect in the understanding of a clock mechanism. They showed participants two animations of the clock's mechanical operation: one at normal speed and another one speeded up. Results showed that speed affects distribution of attention, being higher and more precise in speeded up version. This led in a better understanding of the information and surprisingly, results showed no cognitive load difference.

However, to achieve these advantages through an increase of speed, it is fundamental that the key functional components are well represented, with some visuospatial characteristics (such as the ones mentioned in the previous section) being considered. This, again, is a condition quite difficult to achieve in visualizing real time Big Data, as the analyst can't decide the distribution of data.

Fischer, Lowe and Schwan also proved that speed can be used to raise the perceptual salience of thematically relevant aspects of the display. In the concrete case if this work, in which the position of relevant data portions can't be decided and most of times even can't be known before visualizing it, this simply implies that depending on the speed it is possible that attention focuses on different places.

Deeper on that, they distinguished two kind of attention behaviours. Top-down attention and processing was influenced by background knowledge about mechanical pendulum clocks, while bottom-up attention and processing was largely influenced by perceptual attributes of the animated display. This means that human's automatic and non-controlled attention will be more attracted by perceptual attributes and more affected by these speed or visuospatial changes. This makes even more difficult the analysis and decision-making in visualizations for real time Big Data, because many times those should be made under a time constraint and under a non-chosen and even non-known visuospatial distribution.

Although an increase in speed can help in the understanding of changes, it is obvious that more changes will be appreciated if speed is slower. Newtson and Rinder (1979) confirmed that through an experiment in which they showed subjects a filmed problem solving sequence in slow-motion, normal speed or fast motion, requesting them to count how many actions they perceive in the sequence. People perception of the actions was different depending on the speed, being higher in slow-motion.

Results of both previous studies show that a sequence can be mentally divided in more steps when speed goes slower, but this doesn't mean that the understanding of the sequence will be higher.

After all, it is clear that it doesn't exist a general ideal speed, as it will depend on other factors such as information display, goals of the analysis or analyst skills. Another important conclusion of this chapter is that high speed visualizations can be good interpreted and analysed when information is well distributed and analysts have a good knowledge background about the subject. This sheds light to visualizations of real time Big Data, assuming that with experts in the analysis and good data treatment before visualizing them, the fact of having short time to view them can stop being a major obstacle.

3.3 Interaction

Previous chapters showed how changes in data are difficult to interpret due to changes in position and the speed at which they occur. Notwithstanding, thanks to computer progresses in the last decades, we have the possibility of storing each representation to later allow a rewind. Change blindness, referred to the changes apparently not noticed by analyst, can somehow been stored in our brain and noticed when viewed for second time (Simons and Rensnik, 2005). This throws us into another main visualization characteristic: Interaction.

Interaction is not only offering the possibility to rewind and go forward in visualizations, but something much more complex and something crucial for the analysis process in visual analytics (Aigner et. al, 2007). It is a way to manually parameterize the visualization and analysis tools, but it is not trivial for users and developers to set all those parameters. (Aigner et. al, 2008).

The importance of interaction in visual analytics has grown to the point that some researchers consider that it is through the interactive manipulation of a visual interface that knowledge is constructed, tested, refined and shared. Consequently, there is a claim that interaction should not be an afterthought – a set of controls in visual display that permit modifications – but the first thing considered in the development of an analysis system (Pike et. al, 2009)

Interaction permits a direct manipulation over the visualization that is essential to treat with the data. It is needed, for example, to filter the data, to drill down to a subset of the dimensions or to link and brush different views to see the data from different perspectives. It is also crucial for zooming and changing the focus of view: analysts usually want to see a partial and more detailed data representation from the area of interest. (Zhang et. al, 2012).

In addition, and more concrete to the topic treated in this work, interaction in dynamic data has been proven to improve learning, because it can help to overcome the difficulties of perception and comprehension. Tversky and Morrison (2002) studied the effects of interaction in animation and concluded that if learners are in control of the speed of animation and can view and review, stop and start, zoom in and out, then perception problems can be reduced.

3.3.1 The two levels of interaction

The information visualization community distinguishes between two levels of interaction: low-level and high-level interactions (Pike et. al 2009).

Low-level interactions are considered those between user and software interface, in which user's goal is simply changing the representation to uncover relationships and trends. High-level interactions are those between the user and the information space, in which user's goal is to generate understanding. Figure 4 shows deeper information about each level characteristics.

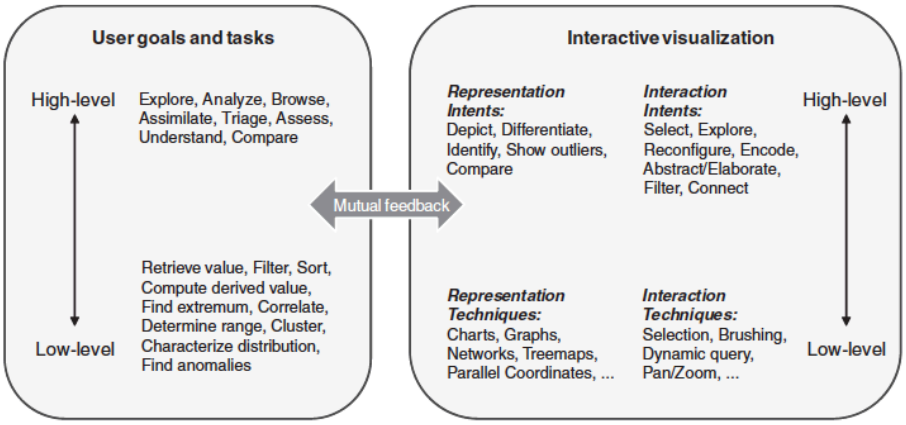

Figure 4: Usergoals, tasks and interactive visualization's characteristics of low and high interaction's level. (taken from Pike et. al, 2009)

3.3.2 Interaction Design Principles

As mentioned above, it is not easy to set all interaction parameters in the right way. It is an intangible concept that is difficult to design, quantify and evaluate (Elmqvist et. al, 2011). Confusing widgets, complex dialog boxes, hidden operations, incomprehensible displays, or slow response times can limit the range and depth of information internalized by analysts, may delay the decision taking process or induce to errors. Therefore, to be most effective, visual analytics tools must support the fluent and flexible use and interaction with the pace of human thought (Heer and Shneiderman, 2012).

This section pretends to give some guidance found in previous research about how to design interaction for visualizations.

At first, the principles of direct manipulation are, according to Ahlberg and Shneiderman (1994) and Elmqvist et. al (2011):

- *Visual and continuous representation of the world of action,* including both objects and actions.
- *Quick, incremental and reversible actions*, whose impact on the object of interest is immediately visible. Visual analysis tools should have the ability of recording and visualize analysts' interaction histories in order to, at least, provide basic undo and redo support. (Heer and Shneiderman, 2012)
- *Physical actions instead of complex syntax.* For example, selection by pointing and not by typing. Control panels separated from the visualizations should be avoid.
- *Immediate and continuous display of results.* The immediate feedback should be given for every key press or mouse motion, not just for major events. That means that the visualization must be able to respond in real time. If this is not possible, precomputing should be considered. Real time interaction will be extended in the next section, as it is probably the most important interaction parameter for Visualizations of Real Time Big Data.
- *Layered or spiral approach to learning* that permits usage with minimal knowledge.

These principles were later extended to few others. Below there's a compilation of them, extracted from Aigner et. al (2008), Pike et. al (2009), Elmqvist et. al (2011), Heer and Shneiderman (2012) and Zhang et. al (2012):

- *Reward interaction.* Interaction design should have as objective encouraging users to interact, as a dialogue that should be initiated and maintained. When they do it is recommended to give them a reward, because rewards are effects that trigger a positive user. The purpose of that is to keep the user stimulated while exploring.

- *Capture user intentionality*. Describes the need to recognize what the user is trying to achieve through the interaction. Knowing why a user is following the steps that he is following is crucial to a visual analytics tool's ability to modify its presentation, suggest alternatives or identify additional information for the user. In order to do that, some visual analytics systems offer users the possibility of annotating their thought processes as they work.

- *Ensure that interaction never 'ends.'* It should always be possible to continue exploring the data; the user should never reach a dead end in which there is no possibility of proceeding.

- *Reinforce a clear conceptual model*. Means that the user should always have a clear idea of the state of the visualization and interactive operations should be reversible, allowing the user to go back to a previous state.

- *Avoid explicit mode changes.* It is recommended to integrate all operations in the same mode instead of introducing different modes, as it is proven that mode changes may break user's flow.

- *Allow collaboration.* In visual analytics work, rarely happens that a single person undertakes the whole analysis process. Therefore, collaboration, in forms such as shared interactive displays, appears as a crucial feature.

- *Assure coordination.* Many analysis problems require multiple views that permit analysts a vision of data from different perspectives and also to facilitate comparison. For example, Tufte advocates the use of "small multiples:" a collection of visualizations placed in spatial proximity and typically using the same measures and scales. Providing multiple views simultaneously connected by linking and brushing functionality is one of the most effective approaches and a major strength of some tools. Coordination in interaction is basic, understood as the propagation of interaction originated from one view to all other. But the coordination should not only be present in the visualization step, but also in all the process. For example, *"huge investments in time and money are often lost, because of the lack of possibilities to properly interact with databases."* (Keim et. al 2008, page 76)

3.3.3 Real Time Interaction

Most of the tools analysed by Zhang et. al (2012) supported interactive actions such as filtering and zooming as well as distortion of views such as changing into logarithmic scale. However, Liu et. al (2013) point out that one of visual analytics problems is that most visual analytics tools are not designed to support a real time interaction at large data scale.

Here, it must be remarked that this real time interaction refers to the possibility of interacting with data or visualization and getting a response at the same moment. It should not be confused with interacting with dynamic data in real time.

According to Liu et. al (2013), this is one of the biggest problems for interaction with Big Data visualizations. Querying large data stores can lead into high latency, disrupting fluent interaction. Even with data reduction methods can be too large to process in real-time.

In Visualizations for Real Time Big Data the area of interest is not static and can dynamically change during research process. Consequently, real time interaction becomes even a more important requirement. If interaction carries latency and data is constantly changing, it is possible that through the process of generating the interaction required, data changes and impossibilities it.

3.3.4 More: Touchscreens vs. Touchpads or mouse-mice

An interesting group of studies were conducted by Brasel and Gips (2013), showing that touchscreens create stronger psychological ownership over chosen products in online shopping scenarios when compared to touchpads or mice. This increases the endowment effect, an effect saying that people use to give more value to something that they own.

Obviously, the results of these studies can't be directly related to the topic in treatment in this work. Nevertheless, it can be adverted as an interesting future research challenge, asking if can also touchscreens generate benefits in data analysis through visualizations, compared to other interactive ways.

3.4 Screen

Visualizing data implies a screen in which these data are projected. Many of the problems in visual analytics are related with the size or resolution of the screen. The disposition of data elements across the screen can also be a differentiating factor in the analysis. Hence, screen is also an important visualization parameter that must be studied.

In this section, first of all size and resolution problems will be exposed; after that, a novel visualization tool is introduced: hybrid reality environments; and at last, a study about different kinds of data landscapes and representations will be remarked.

3.4.1 Size and resolution

"People are familiar with spatial concepts such as distance and height as part of their everyday life. Spatialization takes advantage of this knowledge by using a spatial metaphor to display abstract data, promoting understanding of high dimensional relationships by enabling users to easily see similarities, clusters, and outliers." (Tory et. al 2007, page 1262)

Typical resolution of conventional displays oscillates between 1 and 3 million pixels, sometimes even 4 million pixels (Liu et. al, 2013; Khairi et. al, 2013), and this is most of times insufficient to visualize big datasets at their native resolution and to represent them with the spatialization needed.

A main challenge in big-data visualization is avoiding visual clutter. A limited screen space, a too large amount of data represented in it, or both of them together, often make visualization too dense to be useful. Instead of seeing each data point human eye sees only one big spot, in a phenomenon also called visual noise (Choo and Park, 2013; Yurevich and Vasilevich, 2013). Cluttered visualizations increase the cognitive workflow of users, making it difficult to read variables, compare elements, or recognize trends in the data (Khairi et. al, 2013).

The most obvious solution that comes to mind is getting a bigger screen, in order to have enough space to represent all data. Nonetheless, this sometimes is not a good solution because of human perception capacity. When the number of visualized objects becomes large, humans loose the ability to acquire useful information and can have difficulties in extracting meaningful information. What is more, cognitive load increases (Yurevich and Vasilevich, 2013; Choo and Park, 2013).

Most typical representations strategies are, according to Liu et. al (2013):

- Pixel-oriented visualization techniques, which plot data points as single pixels to maximize screen utilization.
- Spatial displacement techniques, with the pro that reduce occlusion but the contra that do not preserve spatial information. Examples are jittering and topological distortion.
- Parallel coordinates and scatterplot matrices, which can also reduce clutter thanks to dimensions reorder.
- Alpha blending (transparency), often used to encode density and avoid over-plotting.

Still, according to the authors of this classification, each of these techniques requires drawing every data record, which imposes inherent scalability limits (Liu et. al, 2013). Instead, other researchers defend that the amount of data to be analysed should not be a constraint for that and that visualizations scalability should be limited by the chosen resolution of the visualized data, not by the number of records (Cox and Ellsworth, 1997; Liu et. al, 2013)

3.4.2 Hybrid Reality Environments

One solution to improve performance of traditional desktop monitors in complex analysis scenarios with large amounts of data is the Hybrid Reality environment, exposed by Khairi et. al (2013) and Febretti et. al (2013) in their repectives articles.

Hybrid Reality environments are large display walls, constructed by tiling LCD monitors to form a contiguous surface and to allow scientists juxtaposing a variety of datasets for analysis and correlation.

Although it has been previously said that also large screens present difficulties for the understanding do to human perception, this Hybrid Reality environments have many interesting features that facilitate the analysis through visualizations.

One of the examples explained is CAVE 2 (Cave Automatic Virtual Environment), *"an hybrid-reality environment employing commercial thin-bezel, stereo-capable LCD panels. CAVE2 comprises 72 panels (18 columns by 4 rows) arranged on a 24-foot-diameter cylinder with a total resolution of 72 Mpixels (36-Mpixel stereo resolution). It uses micropolarization technology for stereoscopic depth."* (Khairi et. al 2013, page 39)

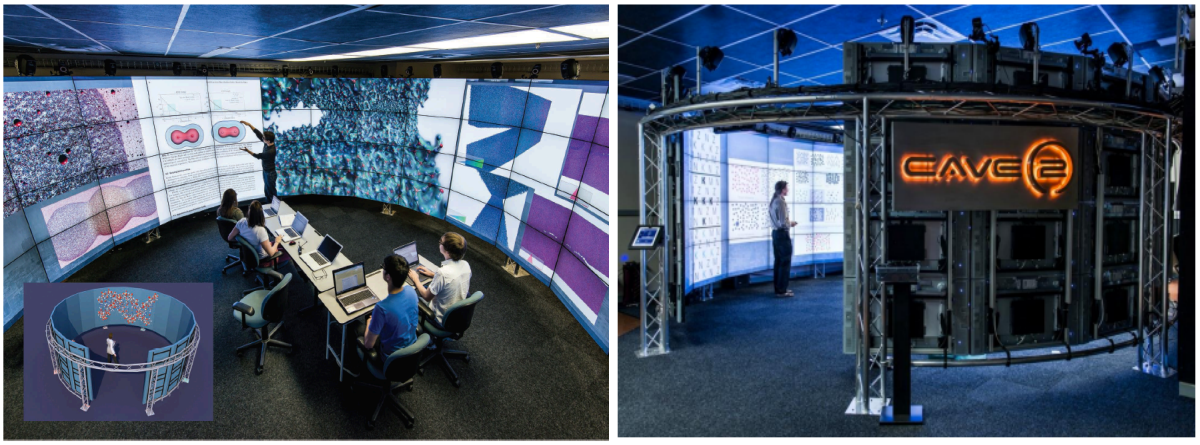

Figure 5: CAVE 2 images (taken from Khairi et. al, 2013 and Febretti et. al, 2013)

These are the main characteristics of Hybrid Reality environments:

- *A large, high-resolution display with a high pixel density.* The resolution should come close to matching human visual acuity.
- *Support for stereoscopic depth.* A 2D display aggravates clutter by forcing designers to represent all the elements on a single 2D plane. This reduction of visual dimensionality, which dumps human's ability to perceive and interpret 3D spaces, impedes visualizations' scalability. The hybrid reality's 3D display, on the other hand, allows designers the freedom of layering information in 3D to reduce clutter and potentially improve comprehension and performance. Moreover, the benefits of stereoscopy aren't limited to 3D datasets, as 2D representations can also benefit from stereoscopic depth.

- *Immersion - Support for naturalistic interactions.* Although 3D displays are often difficult to interact with (Tory et. al 2007), one of Hybrid Reality environments' main features is transforming traditional tiled display walls into highly immersive systems, incorporating head tracking, six-degree-of-freedom input devices, and in some cases a panoramic field of view. This permits a great visualization interaction and immersion experience, and there's solid empirical evidence about the benefit of immersion in data exploration and analysis scenarios.

  About the concrete characteristics of this immersion, head tracking coupled with motion parallax cues helps analysts understanding spatial relationships by leveraging our natural capacity for spatial cognition. Head tracking allows users to experience a viewer centred perspective, enabling them to use embodied interaction. This lets them make sense of complex 3D datasets, giving them a better understanding of the spatial relationships and structures in those datasets.

  Using stereoscopic depth time can be represented in the third dimension by rendering elements at varying depths. This can be a really interesting feature for visualizations of Real Time Big Data.

- *A space to encourage multiple collocated individuals to collaborate.* Traditional virtual-reality environments tend to be enclosed and somewhat isolating. Hybrid Reality environments, on the other hand, provide large and open spaces, allowing for a greater degree of physical navigation, such as walking up to the display surface to see details.

  Solving complex problems involving Big Data often requires a variety of scientific expertise. Therefore hybrid reality environments provide a space where scientists from different disciplines can comfortably sit together to analyse and interpret data. Up to 15 individuals can comfortably stand in CAVE2.

- *A software layer that leverages the hardware to simultaneously display multiple related datasets and utilize hybrid 2D-3D visualization and interaction modalities.*

Hybrid reality environments already count on some success examples:

Some years ago NASA created the Endurance (Environmentally Nondisturbing Under-ice Robotic Antarctic Explorer), an autonomous underwater vehicle designed to map Lake Bonney's geometry, geochemistry, and biology in 3D. For bathymetry reconstruction, the source data consisted of approximately 200 million distinct sonar range returns. With it, scientists designed a visualization tool that mixes 2D scatterplots of chemical properties in the water column and 3D views of the lake's bathymetry data. An hybrid reality environment allowed researchers to visualize the 3D bathymetry model in high detail, eliminating the need for excessive virtual navigation such as zooming and panning.

Another experiment was carried on with the aim of understanding terrestrial insects navigational strategies. To de that, insects movement patterns should be tracked and analysed, and the large number of trajectories collected during experimentation makes it really difficult to visualize it on traditional displays. The visualization in hybrid reality environments employed a multiple layouts to simultaneously show approximately 500 insect trajectories collected under various experimental conditions. This layout let the researcher divide the screen into configurable bins to group related trajectories. Moreover, the use of stereoscopy to encode time made insect movement's temporality visually evident. This was of paramount importance to the researcher, who was interested in understanding the insect's decision-making process over time.

None of both experiments is exactly the case of visualizations for real time Big Data, although both are visualizations of Big Data and show the benefits of doing it in such environments. Nevertheless, second one shows the possibility of using stereoscopy to capture temporality, which could be really useful in representing real time Big Data. More features such as interaction or possibility to collaborate could make the difference in the decision-making process for real time Big Data, where it becomes more and more important to analyse and decide in a short period of time.

3.4.3 Kinds of landscapes

Until now it has been talked about screen size and how this can affect to data representation and human understanding of them. But, imagining that data could be well represented without neither cluttering nor suffering from human perception constraints: which kind of representation – or spatialization – would be better?

To analyse it, the following lines will be based on a study carried out by Tory et. al (2007).

They made the classification based on three characteristics:

- Graphical mark used: points, when only points are showed; landscapes, when a surface has been fitted to the set of underlying points. In landscapes points may also appear and for this study in concrete, all landscapes had points.
- Dimensionality: 2D o 3D
- Colouring method: colours or only grey. Different colours are differentiated as "target level". Colour legends in figure 6 show the different target levels.

They built seven representations based on those characteristics (see an example in figure 6) and asked participants to identify in which section of the picture there was major quantity of points from category X (Category X refers to a concrete cell from the grid). Then, they measured time in response and also accurateness in the responses.

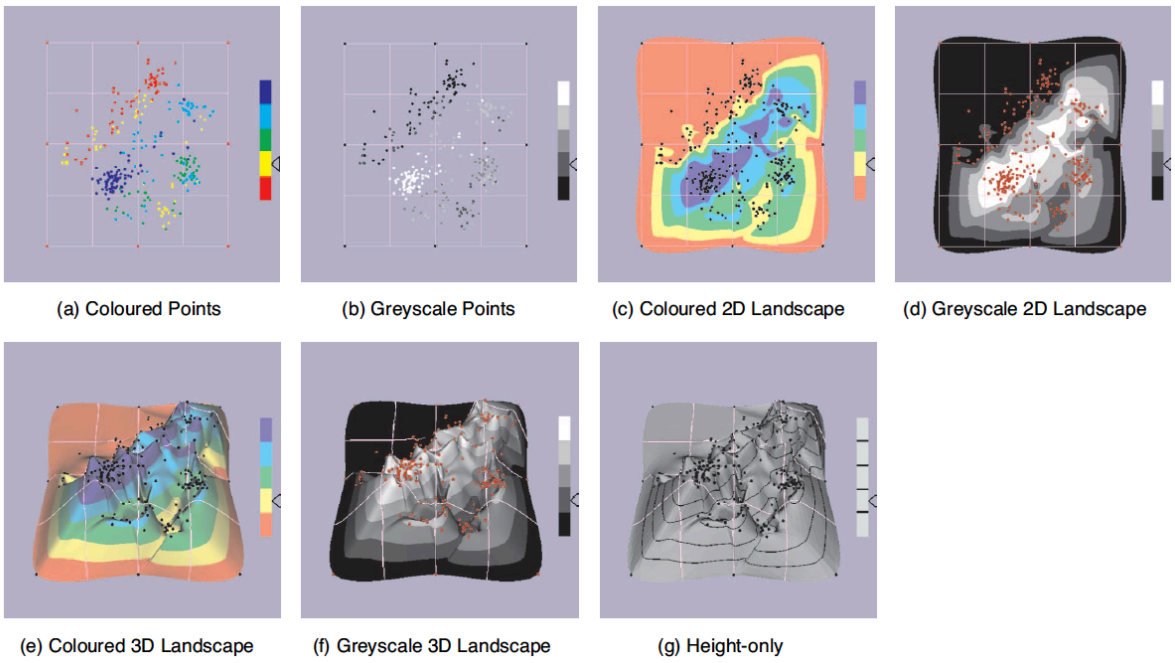

Figure 6: 7 different types of landscapes (taken from Tory et. al, 2007)

The results showed the next conclusions:

- Coloured points supported the best performance and were substantially better than landscapes.
- 2D coloured landscapes performed second best and may therefore be suitable to some applications.
- Landscapes that encoded the data using both colour and height (3D) were slower than landscapes using colour alone (2D), with no difference in accuracy.
- Height-only was least effective.
- Colour would be faster and more accurate than greyscale for all display types.
- Low and high target levels (different colors) would be faster and more accurate than middle target levels.
- Higher data complexity would increase errors and response time.

Participants were also asked to rate the seven displays on 3 criteria: distracting, how distracting was the interface to complete the task; find dots, how easy was to find dots; and overall, how easy made it the interface to complete the task overall.

Coloured points were most preferred and were rated significantly better than all other displays, with the exception of comparison with 2D colour in “Find Dots” and “Overall”, in which both got similar results. The height-only display was least preferred, and was rated significantly worse than quite all the rest. The remaining display types were not rated significantly different.

It should be said that this study has some limitations. On the one hand, the participants weren’t told what the data represented or how to interpret the displays. Thus, results could be different for users experienced in visual analytics. On the other hand, the experiment was conducted with a relatively small numbers of points, so results can’t be directly taken for other scenarios such as when the number of points exceeds the display resolution and they overlap. Moreover, the question was about identifying where was the big quantity of points, which seems a relatively easy task compared with the tasks that analysts should perform when analysing Big Data.

3.4.4 More: Screen vs. paper

Similar to what has been done before with interaction between touchscreens or mice, here will be exposed an interesting article by Anne Mangen et. al in order to reflect on the results and suggest possible research lines.

Mangen et. al (2013) work presents a review of studies and also its own study about the disadvantages of presenting information in a screen rather than in paper. The results in their study showed that reading linear narrative and expository texts on a computer screen leads to poorer reading comprehension than reading the same texts on paper. These results, however, cannot be direct extrapolated to the case of Visualizations for Real Time Big Data. First of all because data is not about neither narrative nor expository texts. Moreover, it is hard to imagine how to visualize data in print when this data constantly changes, as it happens in this case. Nonetheless, it makes it clear that interpreting things in computer screens can impose barriers and difficulties.

Studies with similar results are also presented in Mangen’s article:

Some concluded that hypertext structure tends to increase cognitive demands of decision-making. Hypertext means digital text, which can be only text or text with images or graphics. Related to that, a study by Kerr and Symons revealed that this phenomenon occurs only when time constraints are presents. Thus, if people are given enough time, may be able to comprehend equal amounts of information from paper and computer. Unfortunately, time is one of the lacks in real time visual analytics, as data is constantly changing.

Finally, another peculiarity for reading texts in computer screens: in previous sections it has been proven that interaction helps in visual analytics. In the case of text reading, however, it

seems that interactive tools associated with computer screens can produce the opposite effect. Many researchers have shown that scrolling is known to hinder the process of reading by imposing a spatial instability. This may affect the reader's mental representation in a negative way and decrease its comprehension.

3.5 Prediction

"Science is prediction, not explanation."
Fred Hoyle

At last in this chapter, a short reference to prediction in analysis will be done.

Tufte (2016) puts special emphasis on pointing out how important is to know in which direction one is going to work with the data before having it. He says that Big Data, due to the fact of being so large, can be treated and modified until obtaining a result almost in every possible direction. Therefore, it is so important to know previously the direction in which the analysis should be conducted, in order to conduct the analysis efficiently. He calls this the future of confirmatory analysis of data.

Moreover, prediction alone is known to facilitate learning (Tverky and Morrisson, 2002). In their studies, when participants were asked to anticipate results, and after that were shown graphics or animations to check if their predictions were right or not, the learning was higher.

For this reason prediction becomes important in visualizing real time Big Data. Data will be visualized in big quantities and with short decision timings; therefore, having a predictive background knowledge about the data, about the possible results and about the possible actions to be taken depending on these results can be decisive in the race to successfully obtain the insights needed. Knowing the commonly given results or trends can also be useful to identify the anomalous behaviours.

4 Data Abstraction

"Simplicity is the ultimate sophistication."
Da Vinci

As it has been previously seen, when dealing with large volumes of data visualization problems use to appear. The large amounts of data represented can lead into overcrowded screens and, consequently, difficult the analysis. Data abstraction can be a solution for that.
Data abstraction is the reduction of a particular amount of data to a simplified representation of the whole, a strategy consisting in reducing Big Data to smaller. The main idea is to create an abstraction that conveys key ideas while suppressing irrelevant details (Aigner et. al 2008).

The application of these techniques facilitate the exploration of even huge data sets by starting with a compact overview image, which avoids overlapping data, and then adding more details interactively. It also can be done conversely by starting from the whole set and applying abstractions until needed.

Data abstractions and interaction are two concepts strongly related. Specially in the case of starting from the whole set and applying abstractions until needed, because the user or analyst will specify this abstractions through interactive actions. In the first case, starting directly from a compact image and the adding more details, interaction is present in the add-part. However, data abstraction can be a previous step done through computation and in that case interaction would not be involved.

In this section some data abstraction techniques are introduced.

4.1 Filtering

Filtering is the process of taking away or ignoring irrelevant data objects or data objects that are unnecessary in a specific moment (Bartra, Ware & Calvert, 2001).

Yurevich and Vasilevich (2013) describe three kind of filtering approaches. :

- *Dynamic Query Filters.* Consist in the implementation of some behaviour patterns to be recognised when they appear. Then, the analyst has a direct access to them in order to ease some routine actions. So, now the analyst only has to press on one of the user elements to achieve the desired results and probably that result would be enough to make a decision or to make the area of search a bit smaller. It is a kind of pre-work that later allows analysts to save time in the filtering actions. Dynamic query filters are strong related with the event based visualization approach, that will be explained in chapter 5.

- *Starfield Display*. This approach is based on the idea that the whole data set is always visible. At the first level, in order to not be fully cluttered, some data is aggregated and the analyst sees only some grouped information. But then analysts can make detailed requests, such as zooming actions, through which each selected group collapses into more and more detailed data.

- *Tight Coupling*. With tight coupling some data elements are directly and robustly linked to each other, which prevents analysts from making mistakes of moving in a wrong direction. Thus, for example, when pressing a data element of the group, the other would also be selected or lighted.

Filtering in real time Big Data visualizations is a constant process, as data is constantly changing or being added and then new filter actions will probably be required. It can be a difficult and confusing process when, at the same time, data is changing because of the filtering action required and because of data changes. Then, analyst can doubt if the changes are due to one or another. For that, dynamic query filters can be a really good help because pre-defined conditions can avoid the appearance of not desired or not useful data.

4.2 Sampling

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population.

In simple random sampling, every data point has the same probability of being selected. It is easier and quicker to program but the sample resulting may not be representative and can miss important data. Systematic sampling sorts data points in a particular order and selects data points at regular intervals with a random start.

Stratified sampling divides a data set into disjoint subgroups (groups no element in common) also named “strata” and then applies simple or systematic sampling within each stratum. However, this method requires that specific dimensions be chosen before, demanding prior knowledge and often costly pre-processing. (Liu et. al, 2013)

4.3 Principal Component Analysis

Principal component analysis, abbreviated as PCA, is a data abstraction technique used to reduce the number of variables and to detect structure in multivariate data sets. PCA results in a transformation of the original data space into a different domain—the principal component space. That is, the original data space is transformed in such a way that the first principal component appears like most of the original data set’s variance; then, the second one resembles most of the remaining variance, and so on. The use of principal component leads to a more compressed description of correlations in the data and allows a better understanding of trends.

Important to know is that, in principle, PCA does not distinguish between independent and dependent variables. This often raises problems in the context of time oriented data because the temporal context gets lost and the interpretation gets hampered. Therefore, it is preferable to exclude the independent variable "time" from PCA (Aigner et. al, 2008).

For that, it would not be recommended for visualizations in which the temporal evolution is a crucial factor of study, rather for those in which the important factor is exclusively the data represented in each moment.

4.4 Qualitative Data Abstraction

Named "temporal data abstractions" by Aigner et. al (2008), the technique consists on a reduction from quantitative values to qualitative values, which are much easier to understand. One example is, in values from 1 to 10, grouping 1 to 3 as "low", 4 to 7 as "neutral" and 8 to 10 as "high". This kind of abstractions most of the times require knowledge in the field, in order to be able to make good and meaningful categorisations.

As these categorisations are done before showing the data, the visualization then shows information grouped, which can avoid cluttering and help in gaining time through the analysis.

4.5 Model Fitting

Another reduction strategy is to describe data in terms of mathematical models or statistical summaries. For example, one might fit a model and visualize the resulting parameters or theoretical density. For scatter plots one can use regression models to fit trend lines; examples for time series data include moving averages and auto-regressive models (Liu et. al, 2013).

This approach is the one less directly related to visual analytics, as the mathematical models or statistical analysis themselves give some results and visualization part gets less responsibility in the analysis process. However, using them, for example, to first draw a prediction line and then continuously check how new real time data fits to it, could be a useful way of combining model fitting and posterior visualization.

4.6 Binned Aggregation

"Binning aggregates data and visualizes density by counting the number of data points falling within each predefined bin." (Liu et. al 2013, page 3). For a numeric variable, one can define bins as contiguous intervals. For categorical variables, one can simply treat each value as a bin.

Binned aggregation can be compared with - or even equal considered - clustering methods, defined by Aigner et. al (2008, page 54) as methods to *"reduce the number of data tuples by finding expressive representatives for groups of tuples."*

Liu et. al designed a strategy to enable real time interaction with binned plots, consisting in four steps: data cube queries to support interaction, from data cubes to multivariate tiles, dense versus sparse data tile storage and parallel query processing. Their results allowed a performance of 50 frames-per-second brushing and linking while enabling analysts to interactively examine billions of data records in real time.

5 Approaches

In this chapter some approaches to the Visualizations of Real Time Big Data are presented. As said in the basis chapter, really few is written for the concrete case of study of this work. Therefore, some of the approaches that follow are not introduced in research as concrete solutions to the visualization of real time Big Data, but rather to Big Data or Visualization problems. Nonetheless, they are considered interesting and suitable for the concrete case of this work because some of the conclusions extracted from them can be applicable.

5.1 Low precision and partial iterations computation

Throughout this work it has been repeated on several occasions that response time plays a crucial role in visualization for real time Big Data due to the importance of being agile in the decision making process when data is changing constantly. In the attempt to make the whole process quicker, Choo and Park (2013) suggest a way to permit interaction in real time for computation methods.

Computational methods for the treatment of large amounts of data require significant time, which can obstruct real-time visualization and successive interaction. Therefore, the main objective of their work was making them fast enough to ensure these both things.

They found out that computational methods are too slow because they tend to compute excessively precise results compared to what humans can perceive. Removing this excessive computation would permit achieving the main objectives without pauperizing the results presented to the analyst, as he would not notice the difference.

For doing it, precision and convergence are the key concepts in this approach:

Most computational methods work with the precision of modern CPUs, which is typically double precision. This precision gives at least 10 significant decimal digits, and usually gives 15 to 17 digits. So many decimals are not needed because it is exceeds human's perception capabilities. Moreover, it is possible that it even won't make difference in the representation on the screen, because of resolution or size.

Besides that, computational methods have become so complex that they often have no closed-form solution. Instead, many of them iteratively refine the solution until it converges to a final solution. With that, the notion of convergence becomes critical in determining when to finish the iterations. Algorithms have their own stopping criteria but it has been observed that the major refinement of the solution typically occurs in early iterations, whereas only minor changes occur in later iterations. Figure 7 shows an example of that.

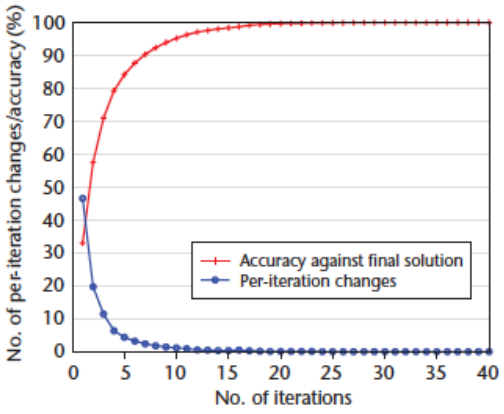

Figure 7: Example of convergence (taken from Choo & Park, 2013)

At first, and related to precision, they reviewed literature focused on adopting lower precision. One example showed that the change from double to single precision only supposed a two pixel-wise displacement between the two cases, while taking much less time. They pointed out that calculations about exactly how much precision is required for a given resolution of the screen space could be conducted before starting the computational process.

Assuming the benefits of lowering precision, the solution proposed by Choo and Park focused on taking advantage of what has been observed about convergence. They solution consisted in iteration-level interactive visualization and after that, only if needed, iterative refinement of computational results.

Iteration-level interactive visualization consists in the possibility of visualizing intermediate results at various iterations and also letting users interact with those results in real time. Visualizing the intermediate results can provide information about the entire data, although certain individual data items might not be as precise in final iterations. Additionally, computational methods can become much more responsive thanks to the interaction permission, because interactions can be reflected in later iterations. That means, for example, if users make a specific change, such as a parameter change, to a computational method, they won't need to wait through a complete run of iterations to see the change's results.

Although this solution implies a clear improvement into visual analytics, it could also bring with it a little bit of confusion due to the constant changes: ones due to the different iterative results shown, others due to data changes. Therefore, some kind of distinction should be easily appreciated between both of them to make it easier for analysts.

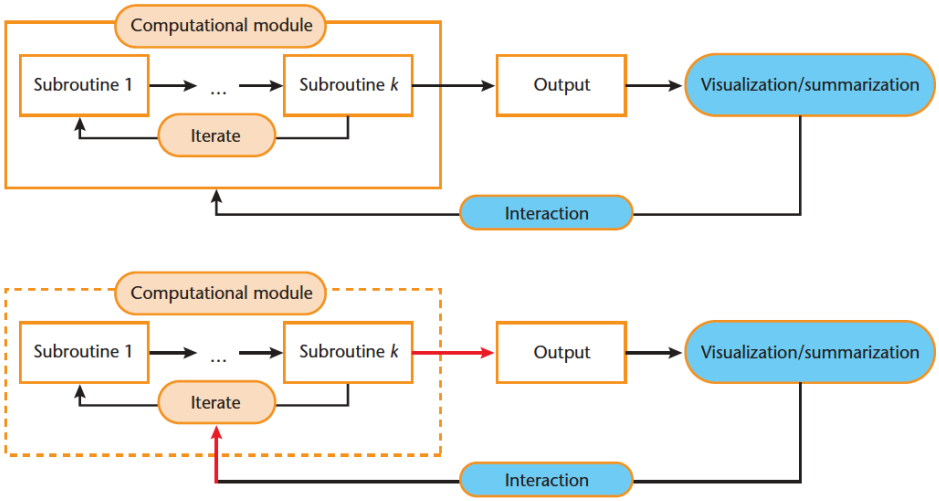

Figure 8: Representation of computational methods in standard approach, above, and in iteration-level interactive visualization, below. (taken from Choo and Park ,2013).

When applying iteration-level interactive visualization, it is possible that sometimes more precise information is required. In that case, they propose an iterative refinement of computational results. This would require further computation, but it would only be executed if needed instead of as a default action.

Data Scale Confinement

At last, Choo and Park propose another solution: Data scale confinement. As it is known, sometimes processing the entire dataset doesn't make sense because there's no screen or resolution enough to visualize all the elements. Having a fixed number of available pixels acts like a bottle neck in algorithm efficiency: while with enough resolution or screen it would be O(n), which assumes that every "n" data element is processed at least once, the "m" fixed number of pixels reduce it to O(m). Therefore, it is useless to process the whole data set and the solution proposed is processing only a part of it.

One of the easiest ways to select this data subset is random sampling, although other more carefully designed sampling methods can be adopted. After that, if some user interaction such as zooming requires the computational results for data items that haven't yet been processed, it can be solved through a different kind of efficient computation. For, when there is a large-scale dataset for which only a certain subset of the data has been clustered, a simple classification method based on the already computed clusters can be applied in order to process the other ones needed.

With all that, and knowing that these approximated approaches sometimes can't give the exact same results as those generated by using, full precision, a complete iterative processes the or the entire data from the beginning, they're a viable and really interesting way to ensure real time for visual analytics for Big Data.

5.2 Maximizing data set analysed

The approaches by Choo and Park tried to obtain the best interaction and performance possible, even though sometimes it meant reducing the set size or precision. Another approach is trying to maximize first the data set size that can be analysed and second maximize performance. Cox and Ellsworth (1997) use this second strategy in their work "Managing Big Data for Scientific Visualization". Two different proposals are exposed here below in order to solve main common problem of Big Data and maximize the data set size that can be analysed:

5.2.1 4 layers model

The layered model of data visualization is inspired in a previous solution developed by Lloyd A. Treinish, but adding a new layer: data management.

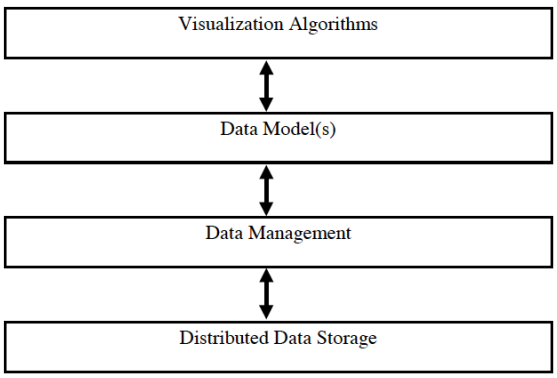

Figure 9: Layered architecture of visualization and data management (taken from Cox & Ellsworth, 1997).

The functions of each layer are:

- *Visualization Algorithms layer:* it contains visualization algorithms.
- *Data model layer:* it is responsible for presenting an API that supports data types and data values in an independent way from platform and machine. It is also responsible for translations and reformatting necessary to support the common API representation.
- *Data management layer:* it is responsible for managing the flow of data into and out of main memory, either from local disk or possibly from remote disk (or even tape).
- *Distributed data storage layer:* it is responsible for providing an API (or APIs) that move data among distributed machines, disks, and potentially tapes.

"The main advantage of this approach is that it allows data management algorithms to be decoupled from the issues of data model, data representation, and also decoupled from the precise implementation of distributed data movement and storage." (Cox and Ellswort 1997, page 5)

5.2.2 Moving the computation to the data

When data is too large to fit in local memory or bandwidth for data transfer is insufficient, Cox and Ellsworth propose moving the computation to the data, which can be done in two ways:

- One, finding or buying a computer that can hold the whole data on the disk and in core. Use this large computer to do the computation. Figure 10 represents this second option, which has de disadvantages of a high cost in case of acquisition of the super-computer; in case of not buying it, availability can generate problems.
- Other, *"partitioning the visualization so that traversal is done on a machine with sufficient disk and memory to hold the data, calculate synthetic geometry on that machine, and download this geometry over a fast local network to the local workstation for visualization"* (Cox and Ellsworth, page 6).

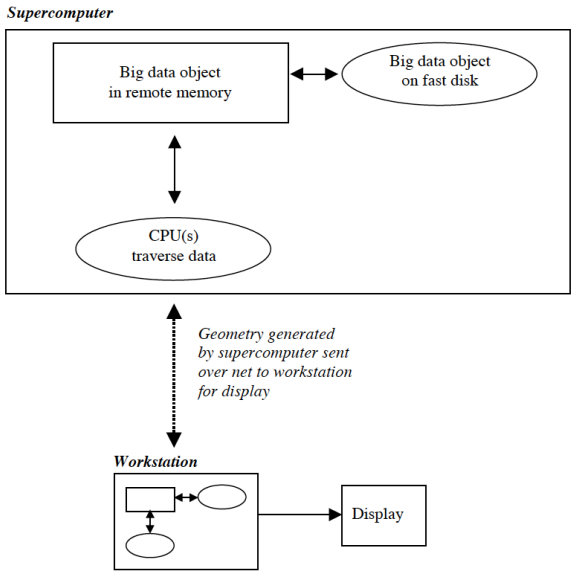

Figure 10: Moving computation to data schema (taken from Cox & Ellsworth, 1997).

5.2.3 More

Other strategies mentioned by Cox & Ellsworth to achieve the objectives previously mentioned are segmentation and paging. Specially, application-controlled segments, which they proved to be a successful technique for visualizing Big Data objects. Also application-controlled demand-paged segments were proposed and confirmed to be a possible strategy for visualizing large objects out of core.

A restriction of the studies carried by Cox and Ellsworth is that the visualization algorithms they studied had only sparsely traversed the data.

As said in the introduction of this section, the strategies proposed by them are focused on maximizing the data set analysed. Although these approaches are less directly related with the visualization part – they are more centred in previous steps than in the analysis and representations parts themselves –, they can bring benefits to the whole process, so they must be considered. Probably for the case of visualizing real time Big Data, most of times, maximizing the data set analysed would not be the first concern of analysts; but once the analysis process problems have been solved and visualization tools have been effectively designed, considering these approaches can lead into extra benefits.

5.3 Event-based visualization

At last, another approach that results promising and very useful for visualizing real time Big Data: event-based visualization (Aigner et. al, 2008; Tominski, 2006). The approach is born with the idea of giving a special support in steering visual analysis with a kind of automatic parameterization of visual representations.

Events are defined as situations that occur if conditions previously defined by users, which are expressed with respect to entities of a data set, become true. The process is as follows:

1. *User interests*. Users specify their interests with encapsulations of conditions, named even types, through event formulas that are developed for that. These formulas make use of elements of predicate logic, including variables, predicates, functions, aggregate functions, logical operators and quantifiers. Sequence events are also supported and they enable users to specify conditions of interest regarding temporally ordered sequences of tuples. For the concrete case of visualizing real time Big Data, a very useful criterion regarding which events can be classified is reoccurrence.

2. *Relevant Data Portions*. This step determines whether the interests defined as event types are present in the data set under consideration. If dynamic data have to be considered, detection efficiency becomes crucial. Here incremental detection methods, methods that operate on a differential data set rather than on the whole data, can help. For incremental detection methods the conditions are not evaluated with respect to the whole dataset, but with respect to a differential dataset that contains only the changes that have occurred since the last evaluation of the conditions. It is obvious that the differential dataset will generally be much smaller and therefore much quicker to execute. However, incremental methods also impose restrictions on possible event types.

3. *Application of user interests in Visual Representations*. For this purpose, three requirements have to be accomplished:

    - Communicate the fact that something interesting has been found.
    - Emphasize interesting data among the rest of the data.
    - Convey what makes the data interesting.

    For that, it is essential that visual representation clearly reflect that something interesting is contained in the data. To meet this requirement, easy to perceive visual cues must be used.

Event-based visualization not only allows to present the data for effective analysis but also to interact with the simulation in real-time: parameters can be adapted during the process by adding or removing condition elements interactively (Müller and Schumann, 2003).

This approach becomes really useful for the concrete case of visualizing real time Big Data because it gives the option of pre configuring some predictable behaviours or some remarkable possible trends, and then, while data is being added or when data change, detection will be automatic and this interesting events will be emphasized in the visual representation. With that, some manual repetitive steps can be skipped and once again, the time gained through these actions can be decisive in the analysis and decision-making process.

For dynamic data the event detection must be performed repeatedly, every time that data changes. Consequently, the efficiency of the event detection becomes crucial. When data change very frequently, the event detection must be able to handle all changes in time and this requirement can be fulfilled only up to a certain frequency of data changes. Beyond this frequency, it is impossible to detect event instances in a timely manner. In this case, a possible solution would be lowering the change frequency even knowing that some changes can be missed.

6 Visualization Methods

Until now it has been talked about important parameters for visualizing real time Big Data, including some crucial requirements or typical problems, and also about some approaches or strategies that can overcome Big Data and visualization problems in order to make real time Big Data visualizations easier.

Now, in this section, some concrete visualization methods that can suit to real time Big Data will be introduced. Nonetheless, before that it is important to discuss and clarify a few important concepts.

6.1 Previous clarifications

6.1.1 Frequency of change and data represented

When talking about real time Big Data visualizations, there are different possibilities of change rate and also different possibilities of what to represent. Figure 11 shows a basic schema for that.

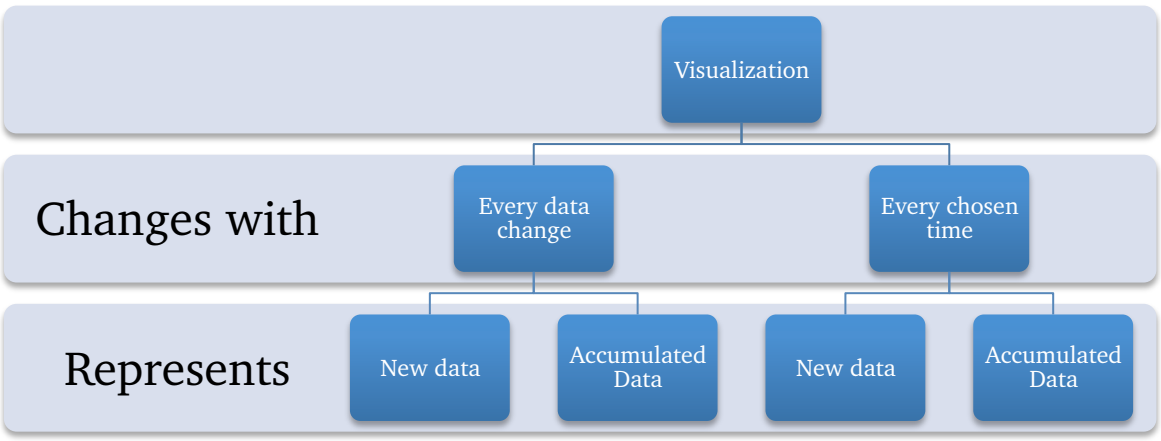

Figure 11: Possible changes and data representations

A priori, real time Big Data visualizations should change with every and each data change. Nevertheless, sometimes this could extremely difficult the computation and the analysis, it could be unnecessary for an effective analysis or simply it could not make sense at all. Therefore, sometimes there's the possibility of choosing a frequency of change that permits the analysts to suppose that visualization is still in real time at the same time as facilitates computation and analysis.

Moreover, there is another decision factor about what to represent: on the on hand, it can be represented only new data compared with the last data change. On the other hand, there is the possibility of representing an accumulation of data; that is, adding new data to the one before.

To put an example, imagine the representation of the amounts of tweets written in the whole world related to some hashtags:

- The first case would mean changing the representation every time a tweet is written and representing only the amount of new tweets (1). Therefore, this strategy makes no sense unless the data represented change slow enough to be perceived and interest is only focused on where this change occur. Even so, this is not a very interesting approach.

- Second option implies a real time modification of the visualization every time that a tweet is written, but simply adding this amount to the one previously shown through an accumulation. This would be the natural "real time Big Data" approach, but as said before, in some cases the rate of change can be so high that computation or representation problems appear.

- The third combination entails choosing a frequency of change and representing only new tweets written in this period. This is a really interesting option when the interest focuses on studying a certain period of time (for example, how has evolved a hashtag popularity in the last ten minutes). Nevertheless, it looks more like an interesting extra feature than like an interesting default mode.

- Last option appears as a solution for when second option is not feasible for too high frequency reasons. It implies, as the second one, representing a continuous aggregation of data, but changing only every chosen time.

For all these options interaction plays an important role, giving the possibility of resetting the accumulation, choosing a new change frequency rate, changing from one mode to another, etc.

6.1.2 Time axis

Another concept that should inevitably be mentioned with dynamic data is time axis. It is virtually impossible to design effective analysis methods without knowing the characteristics of the time axis. For this purpose, Frank (1998) introduced a taxonomy that was posteriorly mentioned and explained by Aigner et. al (2007) and Müller and Schumann (2003). Two main categories are introduced in this taxonomy: time primitives and time structure.

- The time primitives of the time axis can be discrete time points or intervals.

    - Discrete time points are instants in time with no duration. If data are given on a discrete time points axis, then particular data values are valid only at certain points and the space between those points it's not representable. Therefore, this is a kind of a time abstraction.

- Intervals are defined as a temporal primitive with an extent. It can be specified with two time points or a point plus a duration.

When time is involved in data visualizations, the question of whether time points or time intervals are considered is decisive. Depending on the one chosen, different relations will be possible and therefore, different analysis tasks or goals can be accomplished. Most of the known visualization techniques that represent time-oriented data consider time points (Aigner et. al, 2008).

- The structure of the time axis is divided into linear time, cyclic time or branching time:
  - Linear time assumes a starting point and corresponds to time as collection of temporal primitives ordered from past to future.
  - Cyclic time axis is composed of a finite number of temporal primitives in which a temporal primitive X is always preceded and succeeded at the same time by another temporal primitive B (February comes after January but at the same time January succeeds February). Many processes are cyclic and a cyclic time axis can be crucial to see and interpret the cycles. For that, choosing the right parameterization is basic. One example is spiral graph visualization, showed in Figure 12.

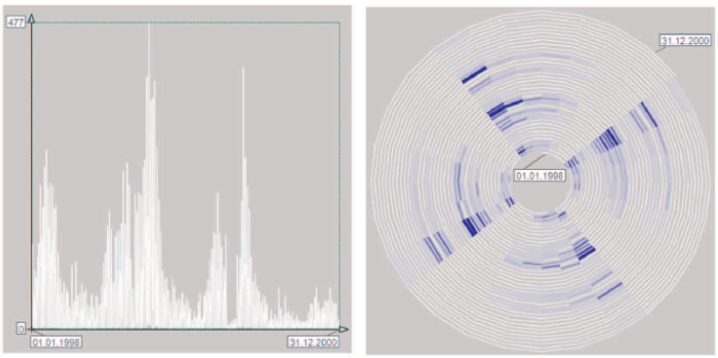

Figure 12: Linear vs. cyclic visualization (taken from Aigner et. al ,2008).

  - Branching time is used for scenarios where sequences of actions are foreseen. Branching time axes are modelled as graph, but methods for analysing branching time are still rare.

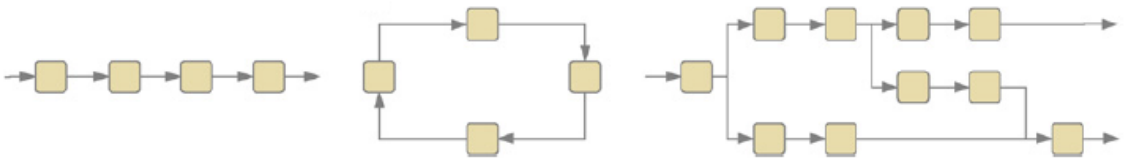

Figure 13: Time structures: linear, cyclic or brunching (taken from Aigner et. al 2007)

This taxonomy characterises the time axis. Nevertheless, it must be said that time axis is not mandatory for real time Big Data analysis. Although time is a really important variable, sometimes the analysis can be carried on without need of time axis. Notwithstanding, it is important to know about this taxonomy for the cases in which time axis is needed.

6.1.3 Data and representation classifications

Apart of time, when talking about dynamic data two other main things have to be considered before choosing one technique or another one: data and representation. Aigner et. al (2007) proposes another taxonomy for that. Most relevant parameters for this work are presented below.

- For data category, the things to consider are number of dependent variables and level of data abstraction.
    - Data can be univariate or multivariate. Most of times Big Data implies multivariate data and handling larger number of variables is one of visual analytic challenges.
    - Level of data abstraction refers to the difference between when all data is represented or some reductions are made, called abstractions and previously introduced in section 4. In case of Big Data representing all data becomes practically impossible. However, only few visualizations support temporal data abstractions.
- Time dependency and dimensionality are the two subcategories in representation's category:
    - Time dependency encompasses static and dynamic representations. Static do not change automatically over the time while dynamic do, and the interest of this work is on dynamic representations. Note that the presence of interaction is not the fact that makes the difference. A representation can change because of interactions but not be considered dynamic, because it doesn't change automatically over the time.
    - Visualization techniques usually show data in 2D or 3D. The superiority of 3D representations over 2D ones is still not clear. While some researchers argue that two dimensions are enough to provide effective data analysis and third dimension involves unnecessary difficulties like occlusion and lost information on back faces, others mark out the third dimension as a possibility to represent more information and voice that possible drawbacks of third dimension can be overcome with advanced interaction techniques or additional visual cues.

6.2 Techniques

"Very often, there are many different ways to represent the data under consideration and it is unclear which representation is the best one (Keim et. al, 2008)."

Down below are introduced some visualization techniques that can represent Big Data and that can be useful for the case of real time Big Data visualizations. There are a large number of visualization techniques and each one can fit better for a concrete case. However, *"there exists no visualization framework that can handle all types of times and data."* (Aigner et. al, 2007)

6.2.1 Hierarchical techniques

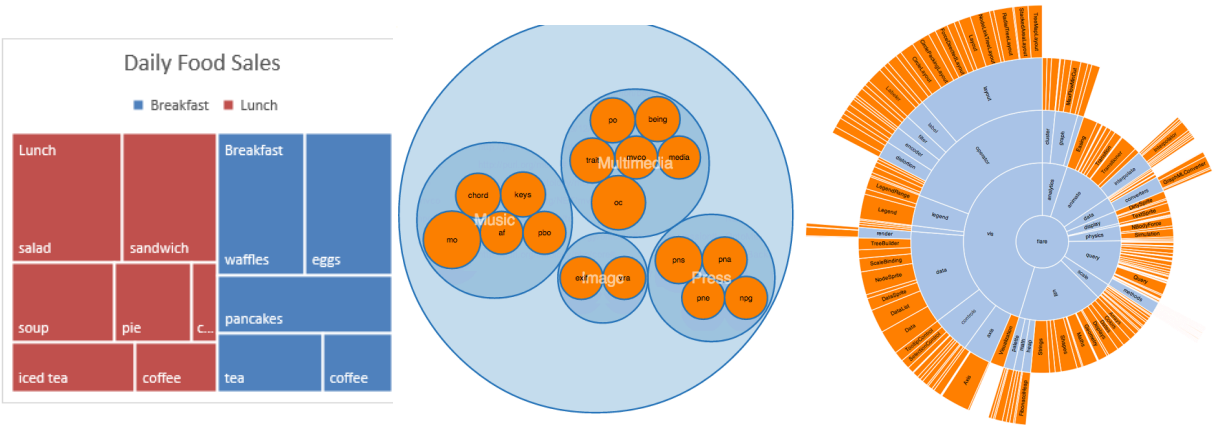

Figure 14: Tree Map, Circle Packing and Sunburst examples.[1]

When data has a hierarchical structure, visualization techniques such as tree map, circle mapping and sunburst are commonly used.

The Treemap is represented by a root rectangle, divided into groups, also represented by the smaller rectangles, which correspond to data objects from a set. Circle packing is an alternative with the difference that it uses circles instead of rectangles, which implies a better space-efficiency. At last, sunburst is similar to the two previous but with the use polar coordinates. Consequently, the main variables are radius and arc length instead of width and height. It has the advantage that it is usually easily perceptible by most humans. (Yurevich and Vasilevich, 2013).

These three techniques have the restriction that they can only show two data factors: the first one is the factor represented through volume and the second is a colour used for grouping the shapes.

[1] Sources: goo.gl/XYS4S0 , goo.gl/IlHABl and goo.gl/Agd16C

In addition, they only show data at one concrete moment in time and therefore they are not recommended to represent data in which the interest relies on identifying historical trends and time patterns. For visualizing data in real time, as representations is able only to show data at a concrete moment and the evolution through time is not shown, changes should be efficiently highlighted in the updates.

6.2.2 Circular network diagram

Circular network diagram is a technique in which relationships between different data objects are represented. For that, data objects are placed around a circle and linked by curves. The different line width or colour saturation is used to represent the intensity of their relativeness. As a difference with the previous ones, circular network diagram can represent as many factors as needed.

Nevertheless, it has the big disadvantage that when facing Big Data representation can become imperceptible and may need regrouping. This problem becomes higher with the use of circular network diagram for visualizing real time Big Data, because the constant changes could make even more confuse the whole representation. Nevertheless, it is one of the most useful techniques when the analysis focus is put on relationships between factors.

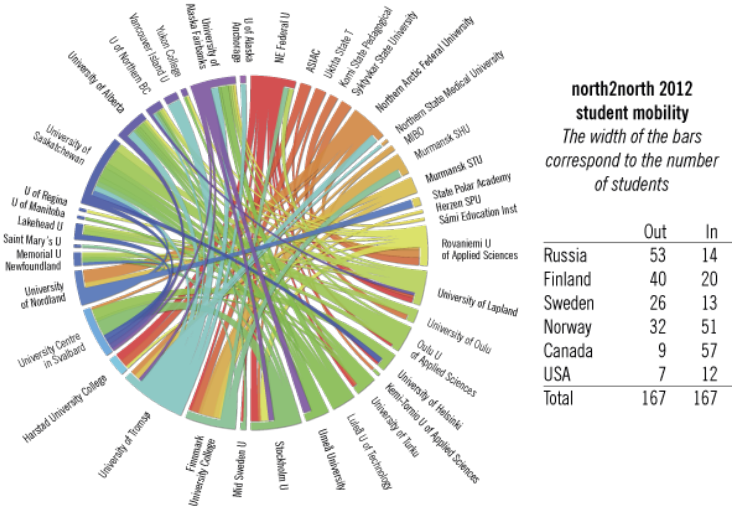

Figure 15: Example of Circular Network Diagram.[2]

6.2.3 Parallel Coordinates

In parallel coordinates visualizations, every data factor to be analysed is placed on one of the axis, and the corresponding values of data object in relative scale are placed on the other. Then, a line represents the value of each data object for each data factor. One useful characteristic of this method is the possibility of scaling each axis according to the pertinent factor, which can help in space optimization and cluttering avoidance.

[2] Source: goo.gl/n9H4ck

Representations in 3D are possible, which allow including more variables in the analysis.

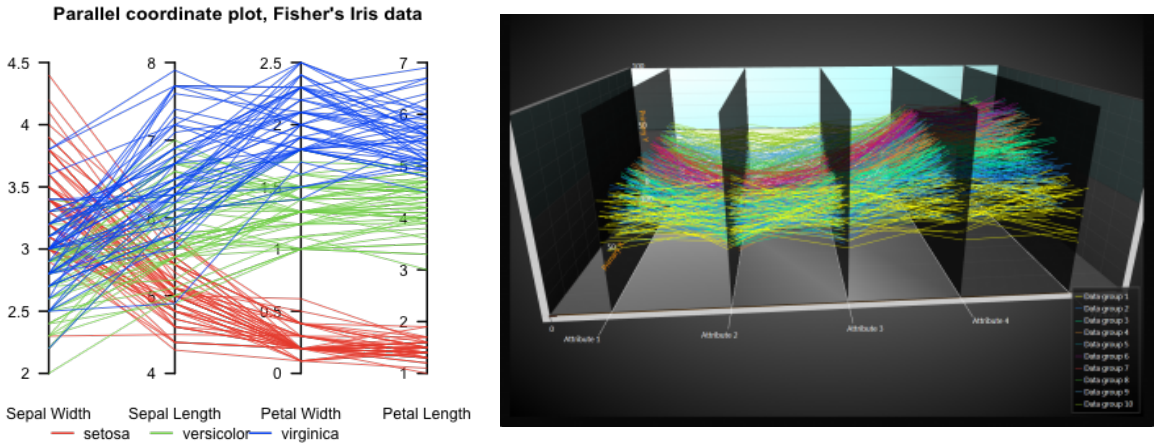

Figure 16: Examples of 2D and 3D Parallel Coordinates.[3]

6.2.4 Streamgraph

Streamgraph is a type of a stacked area graph, in which different data categories are represented around a central time axis, resulting in flowing and organic shape. It can show many individual time series, while also conveying their sum.

It is suitable to visualize data evolution in time and it is ideal to discover trends and patterns over time. It can be used to visualize real time Big Data, by simply moving the visualization to the left and adding new part in the right part while data changes.

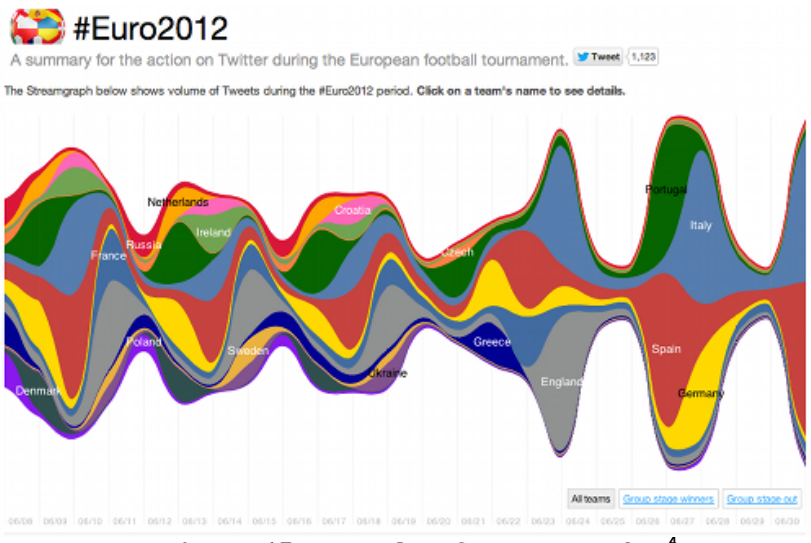

Figure 17: Example of Streamgraph. [4]

[3] Sources: goo.gl/7fWQdG and goo.gl/2FMVRY
[4] Source: goo.gl/774BJ4

One drawback of this this method is that it only works with one data-dimension, and another one is that it usually suffers from legibility issues with large datasets. The categories with smaller values are often drowned out to make way for categories with much larger values, making it impossible to see all the data. Therefore, Streamgraphs are recommended for giving a more general view of the data rather than to spend much time working with it.

6.2.5 Flow visualizations

The fact that flow visualizations are specially designed to represent dynamic data made them attractive to the case of visualizing real time Big Data. An extensive revision of flow visualization techniques can be found in Laramee et. al (2004), showing that different kinds with different forms and figures can be used to better adapt to each data set.

Two concrete examples for that are image based flow visualization, a technique for the visualization of 2D vector fields in general and fluid flow fields in particular (Van Wijk, 2002), and feature and event flow visualization (Reinders et. al, 2001), an ideal way of representing data when the approach of event based detection is the one chosen. Figure 18 shows a representation of each one

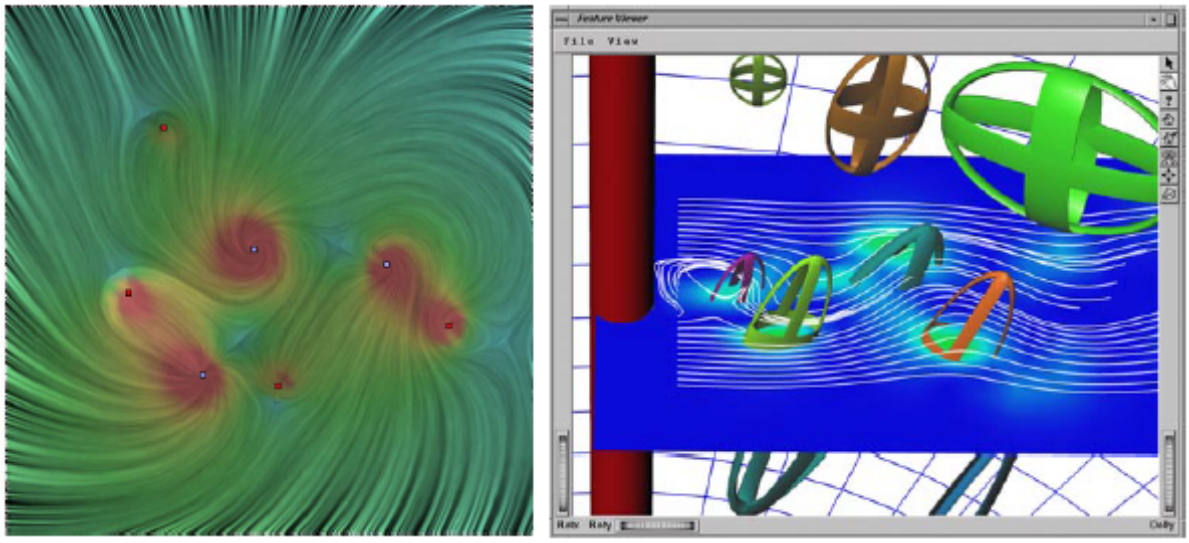

Figure 18: Examples of image based flow visualization and feature and event flow visualization (taken from Van Wijk, 2002, and Reinders et. al ,2001, respectively).

In image flow visualization each image is the result of warping the previous image, followed by blending with some background image, which makes transitions less sharp.

Feature and event flow visualization is designed to explore time-dependent data using an event graph viewer (shows the event graph in an abstract 2D way) at the same time as a 3D feature viewer (shows icons representing the features in 3D-space). Together they provide an excellent way for visualization, give users a guidance of the tracking process and allow good exploration of the time-dependent phenomena.

While the first one represents data in 2D, the second one does it in 3D. In addition, feature and event flow visualization permits data abstraction, while image based flow visualization does not.

6.2.6 Flocking boids

The term flocking boids comes from the parallelism with the flocking behaviour of birds. The method is capable of clustering groups of boids that experience similar data alterations, while the ones experiencing a behaviour different to the main group are represented like leaving the group.

The left representation in figure 19 shows an example of an exchange market's flocking boid visualization taken from Vande (2004). The dominant red colour shows that the majority of the stock market was exposed to negatively changing stock market prices at that moment. Sudden significant events are depicted by expulsed individual boids that leave the main flock. In this case, those companies that have a significantly different price change become separated into two directions: those winning (in white and on the right), and those losing (in red and on the left). The parallel directions of the lines denote a similar price change for that timeframe.

There is also the possibility of representing data separated between different time periods. In figure 19, on the right part, one can see the representation of four different days.

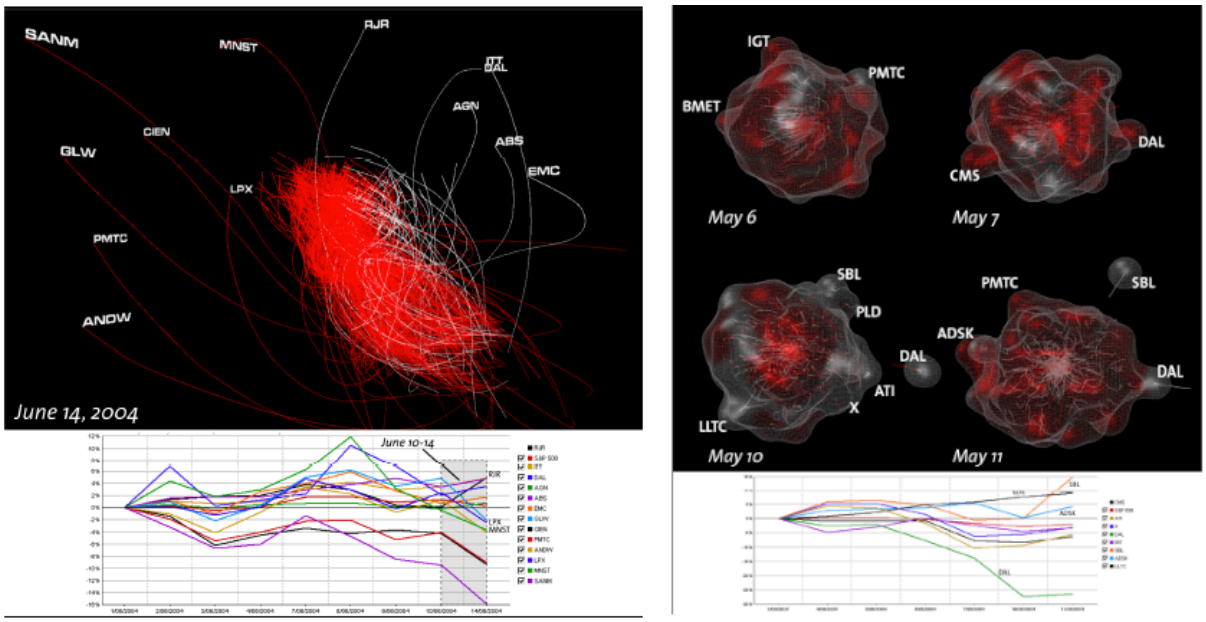

Figure 19: Examples of Flocking Boid (taken from Vande, 2004).

One constraint of this method is that it usually needs a specific time to calculate the time-varying data similarities and evolution. Therefore, it does not offer an immediate representation of a dataset at a certain point in time.

7 Evaluation

"A truthful evaluation of yourself gives feedback for growth and success."
Brenda Johnson

After discussing the important parameters for Real Time Big Data Visualizations, some data abstractions, concrete approaches to effectively visualize Big Data in real time and some visualization techniques, it is interesting to know how to evaluate how good is a visualization tool. To do that Lam et. al's work (2012) will be reviewed in this section, highlighting most relevant aspects.

"Evaluation in information visualization is complex since, for a thorough understanding of a tool, it not only involves assessing the visualizations themselves, but also the complex processes that a tool is meant to support. Examples of such processes are exploratory data analysis and reasoning, communication through visualization, or collaborative data analysis (Lam et. al 2012, page 1)"

Their work focused on evaluating different information visualization aspects, each one called a scenario. For each scenario, being seven in total, they gave a definition, identified the main goals and outputs, suggested evaluation questions and gave some examples.

The classification of the most important seven scenarios was done as follows: they examined 850 previous articles, from which 361 included things about visualizations. In those 361 papers they identified the most mentioned tags about evaluation strategies. After a reduction exercise, they got a set of 17 tags that were grouped into 7 evaluation scenarios in order to represent main distinguishable evaluation questions and goals (see Figure 20).

Paper Tags	Scenario
Process	
1. People's workflow, work practices	UWP
2. Data analysis	VDAR
3. Decision making	VDAR
4. Knowledge management	VDAR
5. Knowledge discovery	VDAR
6. Communication, learning, teaching, publishing	CTV
7. Casual information acquisition	CTV
8. Collaboration	CDA
Visualization	
9. Visualization-analytical operation	UP
10. Perception and cognition	UP
11. Usability/effectiveness	UP&UE
12. Potential usage	UE
13. Adoption	UE
14. Algorithm performance	VA
15. Algorithm quality	VA
Not included in scenarios	
16. Proposed evaluation methodologies	-
17. Evaluation metric development	-

Figure 20: Table with 17 most mentioned evaluation tags and their categorisation into each one of he 7 scenarios (taken from Lam et. al, 2012).

Within the seven scenarios, two main groups should be distinguished: on the one hand, the process group, in which the main goal is to understand the underlying process and the roles played by visualizations; the first 4 scenarios make up this category. On the other hand, the visualization group, where the goal is to test design decisions, to explore a design space, to compare with existing systems, or to discover usability issues; the last 3 scenarios make up this category.

The seven scenarios are introduced and discussed here below.

7.1 Understanding Environments and Work Practices (UWP)

This evaluation is centred on studying people and their task processes, as it is something rarely done. Determining the usability of visual interfaces means not only measuring user preferences or performance, but also understanding what users need to do their job effectively (Pike et. al, 2009). To achieve this objective, current tasks, work environments, work practices and workflows should be studied.

Main questions that should be done are:

- What is the context of use of visualizations?
- What are the characteristics of the identified user group and work environments?
- In which daily activities should the visualization tool be integrated?
- What types of analyses should the visualization tool support?

What is special for Visualizations of Real Time Big Data in this scenario is that work practices can be very dynamic and therefore it can be difficult to find common patrons. As there is no much written about how to face this concrete situation, there are not standardized work practices and workflows. However, work practices and workflows from more general visualization strategies can be taken as a start point, and then customize it to the concrete features of the case in study.

Another interesting point is the characteristics of the user group. Usually, Big Data analysis is carried on by data scientists with experience in the field. One of the cases with more weight of real time Big Data is social media, where enormous amount of data are generated every minute. This data turns out to be interesting not only for companies with experts in their workforce, but also for individuals, as there is a growing trend of individuals taking advantage of social network's impact. Consequently, while developing a tool for analysing social media's data, it would make sense to do it both thinking in experts and non-experts as potential users. This distinction could lead into two quite different tools.

7.2 Evaluating Visual Data Analysis and Reasoning (VDAR)

Second scenario evaluates tool's ability to support visual analysis and reasoning about the data. What is important is that this evaluation is made from the whole tool, rather than from a concrete aspect. Unlike some other evaluation scenarios, in this one is quite easy to get quantifiable outputs: for example, insights obtained during an analysis. It can also be evaluated expressing the quality of the data analysis experience.

Questions that could be done:

- How many insights were obtained in a period of time in a concrete analysis with this tool and how many with another one in the same period of time?
- How good does it support hypothesis generation, the schematization of information, interactive examination or analysis processes such as searching, filtering or zooming?
- How good would the analyst rate the analysis experience with this tool?

7.3 Evaluating Communication through Visualization (CTV)

As the scenario's name suggests, the intention of this evaluation is to study how good the tool communicates the information needed. This is not related with the way of presenting the data, rather than with how messages are transmitted to the user.

In the case of event-based visualization, the approach introduced in section 5.3, this would be one of the most important evaluations. The facility to select kinds of events, formulate them and then the way of being noticed when this events occur are crucial in this approach; for that, a good communication is essential.

Besides, in real time Big Data analysis, with time playing such a decisive role in the decision-making process, communication should be as effective as possible to avoid delays in secondary processes (such as understanding the messages given by visualization tool) and allow analysts to focus most of time in the data analysis.

Questions that could be done:

- Is the tool helpful in explaining and communicating concepts to third parties?
- How fast are the messages understood? Could them be simpler, and therefore, easier and faster to understand?
- Which is the interaction form used (finger touch, writing, mice…)? Can the user choose it or it is predefined?

7.4 Evaluating Collaborative Data Analysis (CDA)

"Researchers in visual analytics often focus on the perceptual and cognitive processes of a single analyst. In practice, real-world analysis is also a social process that may involve multiple interpretations, discussion, and dissemination of results." (Heer and Shneiderman 2012, page 18)

Fourth scenario evaluates how a data visualization tool supports collaborative analysis and collaborative decision making processes. At the same time, it should ensure that the allowance of collaborative analysis does not imply a decrease in performance for doing the analysis, which would counter the benefits of allowing a group analysis.

This would be one of the strengths of Hybrid Reality Environments, explained in section 3.4.2, because of the possibility to allocate more than one person in the place of analysis. Nevertheless, there are some other collaborative ways that do not imply being physically together.

A good assistance to collaboration is view-sharing via application bookmarking, which enables analysts to take up an exploration where their collaborators left off. (Heer and Shneiderman, 2012)

Effective and efficient tools should support social interaction. At least they must be able to export views or data subsets to share or revise them, but it is desirable the capability of exporting the settings for the control panels. This allows other analysts to see and operate the same visualization. In case of dynamic data, the tool should not only permit access to the actual analysis, but also to the historic of steps done, interactions occurred and of course, to previous data.

Few possible questions to evaluate tool's performance in this scenario are:

- What is the process of collaborative analysis and what are users' requirements for that: Is there a need for all users to work at the same time (in real time Big Data analysis, most of times there is)? Will analysts be in the same place or distributed among different ones?
- Does the tool support effective and efficient collaborative data analysis?
- Does it stimulate the group analysis?

Answering these questions will lead to a concrete tool that supports the specifications needed. Each of these configurations may require specialized strategies that consider the division of work, access control, presence indicators, and activity awareness (Heer and Shneiderman, 2012)

7.5 Evaluating User Performance (UP)

From now on, the three scenarios left are the ones related directly with the visualization aspects, rather than with the process. Despite that, sometimes the differences between one scenario and another one is due to small nuances. For example, this scenarios is quite similar to the second one "Evaluating Visual Data Analysis and Reasoning", but while second scenario tried to evaluate how good was the process of analysis, this one is more strictly focused on the outputs, without caring about the process.

In this one, user performance is predominantly measured in terms of objectively measurable metrics such task completion time and task accuracy. Outputs are generally numerical values analysed and modelled through statistical methods. It is also possible to measure subjective performance such as work quality as long, but a requirement for that is that the metrics used can be objectively.

Some possible questions to objectively measure user performance are:

- Which is mean time that a user, or a group of them, needs to complete the task? How accurate are the results obtained?
- How does one visualization or interaction technique compare to another as measured by human performance?

In this section some indirect aspects of human performance can be treated, such as human's perception or cognitive constraints, which play an important role in Big Data visualization analysis. This has been specially shown in sections 3 and 4. A good question to evaluate them would be "What are the limits of human visual perception and cognition for specific kinds of visual encoding or interaction techniques?" This would help in the design of the tool, knowing that this limitations act like a bottleneck in user's performance.

The study of perceptual limits under diverse circumstances, such as different data set sizes or display formats, is really important to explore the scalability of visualization techniques.

7.6 Evaluating User Experience (UE)

First of all, to avoid confusions, it has to be clarified that evaluating user experience refers to user's feeling not about user's previous knowledge about the field.

It is probably the most subjective evaluation. To get a complete evaluation, not only opinions solicited should be taken into account, but also those opinions that spontaneously appear during or after the process. The goal is to know user's opinion about the tool and to see, from the point of view of users, to what extent the visualization tool supports the intended tasks.

Questions to be done:

- What features are seen as useful and what features are missing? To the ones useful, how could they be improved?
- Is the tool understandable?
- How can features be reworked to improve the supported work processes?
- Are there limitations of the current system that would difficult its adoption?

In a certain way, this is the sum of all previous evaluations but seen from the user's mind. Comparing the results of this one with the others can give the right view about if the user perceives the same as what the empirical results show.

7.7 Evaluating Visualization Algorithms (VA)

A visualization algorithm defined as *"a procedure that optimizes the visual display of information according to a given visualization goal."* (Lam et. al 2012, page 11)

In this scenario performance and quality of visualization algorithms are studied by judging generated outputs quantitatively. With difference to other scenarios in which outputs were also evaluated (VDAR and UP), this one only evaluates the computational aspects, without evaluating the posterior user interpretation of them.

Two ways of quantifying how good an algorithm works are are comparing the scores of solutions achieved by different algorithms and exploring limits and behaviour of the algorithm according to data size, complexity and special cases.

A few questions usually raised are:

- Is the algorithm faster than other? Under what circumstances? How does it scale to different data sizes and complexities? How does it work in extreme cases?
- Which algorithm shows the patterns of interest better or gives a more truthful representation of the data?
- Which algorithm produces the least cluttered view?

7.8 Evaluation scenarios' discussion

After briefly introducing each scenario, some common things have to be added.

First, it is difficult to completely pull apart one scenario with the others. Therefore, some scenarios overlap with others in concrete aspects. For example, the quantity of insights obtained during an analysis is a metric that could be used to evaluate both Visual Data Analysis and Reasoning, and User Performance.

Second, some concrete ways are mentioned in order to obtain answers from the questions formerly raised. These are case studies, controlled experiments, field observation, interviews, heuristic evaluation or log analysis.

Third, it should be pointed out that research has been strongly focused on evaluations in the last three scenarios, the ones belonging to visualization itself and not to process. As Lam et. al show, most mentioned tags are those that pertain to Evaluating User Performance–UP (33%), Evaluating User Experience–UE (34%), and Evaluating Visualization Algorithms–VA (22%). Together, these three scenarios contribute to 85% of all evaluation papers over the years. The other 15% is for the process group.

Therefore, there is an interesting research gap in that way, as improving the process will consequently improve the results. The questions related to that, such as how visualization tools can be integrated and used in everyday work environments, how does a visualization tool support communication and knowledge transfer or how does a visualization tool support collaborative analysis, are strongly associated with concepts such as interaction, collaboration, integration or coordination, which have repeatedly appeared in this work as key factors for visualizing real time Big Data.

8 Conclusions

The aim of this work was to do a review of existing literature in visualizations of real time Big Data – and also similar areas from which interesting insights could be taken from – to discuss which where the main parameters to consider and how to face them, which existing strategies or approaches could be useful to treat with the characteristic difficulties of the case and identifying useful evaluation forms for that. It was also an objective to give some examples of visualization methods and abstractions.

The study showed, at first, the scarcity of literature directly related with the concrete case of visualizing Big Data in real time. Nevertheless, with the existing literature and an addition of concepts of Big Data and Visual Analytics that could be applied to the visualization of real time Big Data, this work has achieved the goals set leading to the following conclusions.

8.1 Main visualization parameters

The parameters that have been analysed in this work are movement and transitions, speed, interaction, screen and prediction.

First one have shown how the fact that in real time visualizations the data represented will be in constant movement can affect to human's interpretation of them. For example, most of times moving elements are the ones first perceived. In order to handle these changes as good as possible, an agglutinate of recommendations to design effective transitions has been introduced. This advice should be taken having in consideration that when visualizing real time Big Data the changes would be imposed by data and therefore, these design principles can not always be applied.

Related to speed, it is clear that it doesn't exist a general ideal speed, as it will depend on other factors such as information display, goals of the analysis or analyst skills. An important conclusion of this work is that high-speed visualizations can be good interpreted and analysed when information is well distributed and analysts have a good knowledge background about the subject. This sheds light to visualizations of real time Big Data, assuming that with experts in the analysis and good data treatment before visualizing them, the fact of having high-speed changes can be a less critical obstacle.

Interaction has been pointed out as one of the most important visualization parameters, highlighting that it is through the interactive manipulation of a visual interface that knowledge is constructed, tested, refined and shared. The most salient need related to interaction is the need of real time interaction, allowing users to immediately be reflected their actions and avoiding delays that could lead them to a big confusion in a such changing environment as visualizing real time big data. Interaction's section also includes a description design principles to support the fluent, flexible and effective use of interaction.

Insufficient screen size and screen resolution are two of the most common problems when visualizing Big Data. Consequently, they are also present in the concrete case of real time. In this work Hybrid Reality environments have been introduced as a possible solution for these problems that also offer solutions for other problems through really interesting features such as offering stereoscopic view, allowing collaborative analysis, supporting naturalistic interactions and giving an immersion experience.

Also in screen's chapter a study about kinds of landscapes is presented, showing that coloured points are the best option – both considering user's performance and preferences - even when they compete against 3D landscapes. Nevertheless, the study was carried out with a relatively small number of points, which implies that the results are valid when Big Data is already reduced to smaller quantities of data, but are still not proven when large amounts of data are visualized.

At last, a short chapter has been dedicated to highlight the importance having a predictive background knowledge about the data, in order to be able to identify expected or anomalous behaviours while visualizing them in real time.

8.2 Data abstractions

Data abstractions such as filtering, principal component analysis or binned aggregation have been introduced as strategies that can help in the visualization of real time Big Data. It has been proven how they can be very useful, both as a previous step to filter irrelevant data and visualize only the important parts and also as part of the visual analysis, through an interactive process to continuously focus on the desired data portion.

8.3 Approaches

The approaches introduced have showed that, in a task in which quick response is so important such as visualizing real time Big Data, the strategies that can speed up the process allowing real time analysis, interaction and decision-making become crucial.

In that way, it has been seen that sometimes reducing computing precision can give these benefits without harming the results accuracy. Similar to that, visualizing partial of intermediate iterative process results instead of final ones can lead to the same benefits.

Event based visualizations have been pointed out as one of the most fruitful approaches, as they permit an automatic detection of concrete interests that will be highlighted while data are interpreted in real time. However, a previous work has to be made in order to correctly specify these events.

At last, and focusing more in the data processing part rather than in the visualization one, some strategies have been introduced to maximize the data set that can be processed: the 4

layers model, moving computing to the data, segmentation and paging were some of the topics proposed to achieve that.

8.4 Visualization methods

Visualizations methods' chapter has first shown the importance of knowing some characteristics of the time axis and also considering the possible change rates and what to represent with every change. After that, a series of concrete visualization methods have been introduced, proving that depending on various factors one or another can be more suitable for visualizing data.

8.5 Evaluations

At last, it has been proven the importance of evaluations, as they are not only a way of checking if something is working correctly, but also an ideal exercise to detect what could be improved or which innovations are demanded.

The seven evaluation scenarios reviewed have give a complete framework of what has to be considered in a good evaluation, as well as which questions should be made. It is a great way of checking if all the concepts previously mentioned have been successfully applied. A finding in this chapter has been the untapped potential of evaluations focused in the process, such as understanding environments and work practices. Future work on that may lead to a better design of visualization tools.

Bibliography

Ahlberg, Christopher & Shneiderman, Ben (1994): *Visual information seeking:tight coupling of dynamic query filters with starfield displays,* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 313-317.

Aigner, Wolfgang et. al (2007): *Visualizing time-oriented data: a systematic review.* Science Direct, Computer & Graphics, 31, 401-409.

Aigner, Wolfgang et. al (2008): *Visual Method for analysing Time-Oriented Data.* IEEE Transacions on visualizations and computer graphics. Vol 14 (1), 47-60.

Bartram, Lyn; Ware, Colin & Calvert, Tom (2001): "*Filtering and Integrating Visual Information with Motion*". Journal of Information Visualization, Vol 1, 66-79.

Beck, Michael (2003): *Real Time Visualization of big 3D City Models.* International archives of the photogrametry remote sensing and spatial information sciences, Vol 35 (5), 1-6.

Benartzi, Shlomo & Lehrer, Jonah (2015): *The Smarter Screen.*

Big Data en números (2014). 15.12.2016. https://goo.gl/KofvQ1

Brasel, S.Adam & Gips, James (2014): *Tablets, touchscreens, and touchpads: How varying touch interfaces triggerpsychological ownership and endowment*. Journal of Consumer Psychology, Vol 24 (2) 226–233.

Choo, Jaegul & Park, Haesun (2013): *Customizing Computational Methods for Visual Analytics with Big Data.* IEEE Computer Society, July/Augus 2013, 22-28.

Cox, Michael & Ellsworth, David (1997): *Managing Big Data for Scientific Visualization.*

De Mauro, Andrea; Greco, Marco & Grimaldi, Michelle: *What is Big Data? A consensual definition and a review of key research topics.* AIP Conference Proceedings 1644 , 97-104.

Doleisch, Helmut (2007): *SimVis: Interactive Visual Analysis of large and time-dependent 3D simulation data.* Proceedings of the 39th conference on Winter simulation, 712-720

Elmqvist, Niklas et. al (2011): "*Fluid interaction for information visualization*". Information Visualization, Vol. 10 (4), 327-340.

Fischer, Sebastian; K.Lowe, Richard & Schwan, Stephan (2007): *Effects of Presentation Speed of a Dynamic Visualization on the Understanding of a Mechanical System.* Journal of Applied Cognitive Psychology.

Fisher, Danyel et. al (2012): *Interactions with Big Data Analytics.* Magazine Interactions, Vol.19 (3), 50-57.

Febretti et. al (2013): *"CAVE2: A Hybrid Reality Environment for Immersive Simulation and Information Analysis"*. The Engineering Reality of Virtual Reality.

Fox, Peter & Hendler, James (2011): *Changing the Equation on Scientific Data Visualization.* Science, Vol 331, 705-708.

Frank, A.U. (1998): *Different Types of "Times" in GIS.* Spatial and Temporal Reasoning in Geographic Information Systems, 40-62.

Friendly, Michael (2006): *A brief history of data visualization.* Handbook.

Gallego, Francisco (2013): The Role of Analytical Management of Social Audiences. Revista TELOS, 1-8.

Heer, Jeffrey & Robertson, G. George (2007): *Animated Transitions in Statistical Data Graphics.* IEEE Transactions on Visualization and Computer Graphics, Vol.13 (6), 1240-1247.

Heer, Jeffrey and Shneiderman, Ben (2012): *A Taxonomy of Tools that Support the Fluent and Flexible use of Visualizations.* Interactive Dynamic for Visual Analytics, Vol.10 (2), 1-26.

Leutner, Detlev; Leopold, Claudia & Sumfleth, Elke (2009): *Cognitive load and science text comprehension: effects of drawing and mentally imagining text context.* Computers in Human behavior 25, 284-289.

Keim, Daniel et. al (2008): *Visual Analytics: Scope and Challenges.* Lecture notes in computer science, No. 4404 , 76-90.

Keim, Daniel; Qu, Huamim & Ma, Kwan-Liu (2013). *Big Data Visualization.* IEE Computer Society, 50-51.

Kornblit, Avinoam et. al (2000): *Dynamic Data Visualization.* United States Patent.

Lam, Heidi et. al (2012): *Empirical Studies in Information Visualization: Seven Scenarios.* IEEE Transactions on Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers, Vol. 18 (9), 1520- 1536.

Laramee, Robert et. al (2004): *The State of the Art in Flow Visualization: dense and texture-based techniques.* Computer Graphics Forum, Vol. 22 (2), 203-221.

Leutner, Detlev; Leopold, Claudia & Sumfleth, Elke (2009): *Cognitive load and science text comprehension: Effects of drawingand mentally imagining text content* Computers in Human Behavior 25 (2009) 284–289

Liu, Zhicheng; Jiang, Biye & Heer, Jeffrey (2013): *im Mens: Real-Time visual Querying of Big Data.* Eurographic conference on visualization (EuroVis), 32 (3).

Newtson, Darreb & Rindner, Richard (1979): *Variation in behavior perception and ability attribution.* Journal of Personality and Social Psychology 1979, Vol 37 (10), 1847-1858.

Mangen, Anne; Walgermo, Bente R. & Bronnick, Kolbjorn (2013): *Reading linear texts on paper versus computer screen: Effects on reading comprehension.* International Journal of Educational Research, Vol. 58, 61–68.

Manovich, Lev (2008): *La visualización de datos como nueva abstracción y antisublime.* Revista estudios visuales,5. 125-135.

Müller, Wolfgang & Schumann, Heidrun (2003): *Visualization Methods for Time-Dependent Data - An Overwiev.* In: Proceedings of winter simulation conference, New Orleans, USA. 737-744.

Pike, William A. et. al (2009): *The science of interaction.* Information Visualization, Vol. 8 (4), 263-274.

Reda, Kahiri et. al: Visualizing Large, Heterogeneous Data in Hybrid-Reality Environments. IEEE Computer Society, July/Augus 2013, 38-47.

Reinders, Freek; Post, Frits & Spoelder, Hans (2001): *Visualization of Time-Dependent Data using Feature Tracking and Event Detection.* The Visual Computer, Vol.17 (1), 55-71

Simons, Daniel & Rensik, Ronald (2005): *Change Blindness: past, present and future.* Trends in Cognitive Sciences, Vol 9 (1), 16-20.

Tominski, Christian (2006): *Event-Based Visualization for User-Centered Visual Analysis.* Dissertation zur Erlangung des akademischen Grades Doktor-Ingenieur (Dr.-Ing.) der Fakultät für Informatik und Elektrotechnik der Universit at Rostock.

Tory, Melanie et. al (2007): *Spazialization Design: Comparing points and landscapes.* IEEE Transacions on visualizations and computer graphics. Vol 13 (6), 1262-1269.

Tufte, Edward (2016): *The future of data analysis.* Keynote session. goo.gl/x1HJxi

Tversky, Barbara & Morrison, Julie Bauer (2002): *Animation: Can it facilitate?* Int. J. Human Computer Studies, Vol 57, 247-262.

Van Wijk, Jarke (2002): *Image based flow visualizations.* ACM Transactions on Graphics. Vol. 21 (3), 745-754.

Vande, Andrew (2004): *Time-Varying Data Visualization using Information Flocking Boids.* IEEE Information Visualization. 10.1109/INFVIS.2004.65.

Wang, Chamont & Meisner, Michele (2010): *Dynamic Data Visualization.* CS-BIGS 4 (1), 9-22.

Ward, Stuart Jonathan & Barker, Adam (2013): *Undefined by Data: A survey of Big Data definitions.* Computer Sciences Databases. arXiv:1309.5821v1

Yurevich, Evgeniy & Vasilevich, Vasiliy (2013): *Analytical Review of Data Visualization Methods in Application to Big Data.* Journal of Electrical and Computer Engineering, Vol.2013.

Zhang, Leishi et. al (2012): *Visual Analytics for Big Data Era - A Comparative State of the Art Commercial Systems.* IEEE, 173-182.