Big Data Visualization

Ouafae Ben Ali and Anass El Haddadi
National School of Applied Sciences Al-Hoceima, Morocco

Abstract—Nowadays, the data is a golden capital for any business organization that wants to improve their business. The big organizations and the most reputed ones do not only think to collect data, but they make continuous efforts to use this data for efficient decision making.

Data Visualization plays a crucial importance in big data analysis.

The existence of various data visualization methods can be confusing for users to choose the most appropriate.

In this document we give a guide for data visualization methods choosing, and how the traditional methods are improved to meet the big data visualizations need.

The next topic highlights the importance of visual perception in data visualization design and gives some rules for the color choosing.

Index Terms—Data visualization; big data visualization, visual analytics; methods of data visualization; color

I. INTRODUCTION

The visual human system is very powerful; it can catch information immediately and intuitively. Human beings are innately visual creatures; half of human brain is devoted to visual processing: It processes images 60 000 times more quickly than text. Thus the use of visual perception to interpret and understand complex data is not a new concept; it has deep root.

Data visualization is a graphical representation of information. The abstraction of data allows to the user to pick up the information easier and faster. The data visualization is practiced in many disciplines. It communicates information clearly and efficiently. As the authors cited in their paper [1], the purpose of visualization is to improve the clarity and the elegance of the abstract information to allow better understanding of big data and give better interaction and communication.

A successful visualization must be [2]

- Informative: Provide access to information and let user gain knowledge.
- Efficient: Access to this information should be as straightforward as possible, without sacrificing any necessary, relevant complexity.
- Aesthetic: The use of graphical construction as axes; layout, shape, colors, lines... must be harmonic to guide the user, communicate meaning, reveal relationships and highlight conclusions.

Recently, the data visualization became one of the primary interests of data analyses. It can tell many stories that the data cannot tell directly. In their paper [3], the authors show the benefits of data visualization according to the respondent percentages of a survey (Table1).

Table 1: Benefits of data visualization tools [3]

<table>
<thead>
<tr>
<th>Benefits</th>
<th>Percentages (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improved decision-making</td>
<td>77</td>
</tr>
<tr>
<td>Better ad-hoc data analysis</td>
<td>43</td>
</tr>
<tr>
<td>Improved collaboration/information sharing</td>
<td>41</td>
</tr>
<tr>
<td>Provide self-service capabilities to end users</td>
<td>36</td>
</tr>
<tr>
<td>Increased return on investment (ROI)</td>
<td>34</td>
</tr>
<tr>
<td>Time savings</td>
<td>20</td>
</tr>
<tr>
<td>Reduced burden on IT</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 1: Benefits of data visualization tools [3]

Throughout visualization history, there have been many prominent events; the most important that revolutionize the data visualization is computer development. The figure1 [1] schematizes the evolution of data visualization over the years.

II. OVERVIEW OF DATA VISUALIZATION METHODS

Data Visualization is an analysis tool that enables to users to explore data and find stories that cannot be found with formal statistical tests. The primary step is to know what to look for and what questions to ask based on the available data. Nathan [6] describes the process to follow for getting a good data visualization (figure2).

Many visualization options exist, it can be hard to figure out what graph or chart suits the data best. The first necessary step is to understand and learn about the data. Nathan Yau [5] divides visualization to many areas to make a choice for one particular type of data easier.

A. Visualizing Patterns over Time:

Data is changed over time; temporal data can be categorized as discrete or continuous recording to the author [5].

Figure 1: The evolution of visualization methodology [1]
**Discrete Points in Time:** to show changes over time for several variables, it is possible to use:
- Bars
- Stack the Bars
- Points

**Continuous Data:** it is similar to the previous case. The difference is that continuous data represents constantly changing phenomena.
- Time Series Chart
- Step Chart

**B. Visualizing Proportions:**

Many methods that use size or area to show differences or similarities between values or to a whole.
- Pie
- donut:
- Stacked Bar Chart
- Treemap

**Proportions over Time:**
- Stacked area chart

**Point-by-Point**
- Line plot

**C. Visualizing Relationships**

To look for relationship nature between variables: causal or correlative relationship, and prove it graphically it is possible to use:
- Correlations are shown in scatterplots
- Scatterplot Matrix
- Bubbles sized by area
- Bubbles sized by radius
- Distribution

**D. Spotting Differences**

When the user doesn’t know what question he should ask, there are some methods to visualize an overview about the data like:
- Heatmap
- Star Charts
- Parallel coordinates
- Dot plot
- Boxplot

**E. Visualizing Spatial Relationships**

Nathan [5] explains the maps are a sensitive and delicate type of visualization. The user must be focused on the data as well attentive to the geography dimension

---

![Figure 2: The iterative data exploration process [6]](image)
The interoperability with visual imagery and data analysis techniques is the third factor. Visualization methods should extract the essence of data by the most appropriate visual forms. Diversify visualization angles to allow the user to have different visual points of view that must be considered. In this way, the author enumerated several notions:

- Dynamic projection [12]
- Interactive filtering [13]
- Scaling images [14]
- Interactive distortion [15]
- Interactive combination [16, 17]

III. BIG DATA AND VISUALIZATION

The first section summarizes the most used data visualizations methods.

In this section, we highlight a trend term that is big data, and we discuss on how the visualization can be helpful to analyze a large amount of data.

A. Big Data Overview

Day after day, the Data in our world grows exponentially; human behavior, astounding information, variety details are captured, collected, and stored to be analyzed. The large volume of data; whether structured or unstructured that inundate the world is the phenomenon of big data. However, the quantity is not the most important; it is what organizations do with data to be developed and enable its success.

According to S. SAGIROGLU and D. SINANC [18], Big data is a term for massive data sets having ample, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results.

The next part describes some processing methods for big data analysis and the importance of data visualization in data analysis.

B. Big Data Visualization

In his paper [1], he listed the big data processing methods including applied mathematics, statistics, computer science and economics. According to the authors [1]: "those are the basis for data analysis techniques such as Data Mining, Neural Networks, Machine Learning," Signal Processing and Visualization Methods. Those methods are interconnected and firmly connected to each other as presented in figure 4[1].

The visualization is an important part of Big Data Analytics. Intuitively, the visual representation is more likely to be accepted by a human in comparison with unstructured textual information. However many companies and open source projects see the future of big data analytics via visualization [1].

According to Mordor Intelligence [19], the data visualization applications market “is currently valued at USD 4.12 billion and is expected to grow at a CAGR of 9.21%, to reach USD 6.99 billion by the end of 2022". According to the IDG Research study that listed in the SAS white paper [20] “98% of the companies that using effective big data use data visualization for analysis."

In his book [21], Simon shows that Amazon, Twitter, Apple, Facebook, and Google use data visualization to improve decision into their businesses.

For this reason, it is necessary to develop big data visualization methods to better understand and present complex data.

Face to this undeniable challenge; many reputed groups, designers, are focused on how to improve visualization tools and establishing new interactive platforms. In his paper [22], Husain provides a list of recently developed and powerful visualization tools, platforms, and API.

Many Eyes [23] enables users to upload their multivariate data, generate graphical displays and engage a broader audience. It is an IBM project.

Socrata [24] is an open source project that enables the servicing and sharing of dynamic data from public, government and private organizations.

D3 [25] is a modern JavaScript library ubiquitously used for developing dynamic data visualizations.

D3 [26] is short for “Data Driven Documents for JavaScript” is an open source library that can render amazing charts out of various data sources using HTML, SVG, and CSS. This library responds advanced visualization need with complex data sets and allows smooth interaction and sharing.

Cytoscape [27] is an open source software platform. It is most commonly used for bioinformatics, but it can be used to visualize and analyze network graphs based on the nodes and edges. Cytoscape is available as a stand-alone desktop application and a web application.

Tableau [28] is commercial data visualization platform for interrogating complex, structured/unstructured, multi-source data. Tableau [28] offers many advantages as drag and drop and real-time interaction with different types of visualizations (graphs, charts, maps, etc...). It also allows web embedding and team collaboration.

International Business Machines (IBM) Software [29], Microsoft [30], Amazon [31] and Google [32] are commercial Big Data platforms;

As this topic is not a main subject of the paper, this list is not exhaustive.
C. Big data visualization challenges

The graphical representation of a large amount of data is not trivial, as his survey [33]; the author cited some problems that can be meted in Big data visualization:

1) Visual Noise: presenting a whole of data sets that are related to each other can be a mess on the screen, the data visibility can be lost face to this phenomena.

2) Large Image Perception: To solve the above problem; an approach of distributing data above large screen: this generated another problem which is large image perception. The human perception is limited. As seen the graphical representation is not only limited by device screen; but also by the human perception.

3) Information Loss: the data aggregation and filtration are another approaches proposed to reduce the number of visualized points. This way can be useful to solve the above problem, but it can mislead the user by hiding impressive sets.

4) High-Performance Requirements: The listed problems become harder when the visualization is dynamic. Display a whole of data can be Costly in terms of time and resources.

5) High Rate of Image Change: the number of data changes and its intensity on display cannot be controlled by the user.

The authors [34] cited more challenges of big data visualization such as perceptual and interactive scalability.

In summary Big Data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges.

D. Big data visualization methods

The data amount and complexity are in a continuous increasing, the need to accompany this change also growing: the traditional data visualization methods are inadequate to present big data. The development, improvement, and optimization of the tools and methods that can solve the challenges of the big data visualization can bring enormous benefits to analysts.

Representing data as patterns and forms is of paramount importance for Big Data interpretation, as already discussed the visual perception is more likely to be accepted by a human in comparison with unstructured textual information. However, the perception is limited, especially when it is presented as a significant amount of numerical or text data.

Big data visualizations approaches can be performed through[33]:

- More Than One View per Representation Display
- Dynamical Changes in Number of Factors
- Filtering: such as dynamic query filters, starfield display, and tight coupling.

In this section, we list some big data visualization methods as described and classified in [34] considering the following data criteria: (1) large data volume, (2) data variety, and (3) data dynamics.
large amounts of data dynamics using animation. Data object are placed around a circle and linked by curves based on the rate of their relativeness. Color can be used to group the data into different categories, which aids in making comparisons and distinguishing groups.

So, this method directly links several objects and shows how relative it is. It is an elegant and compact way to show networks of relations between items such as products, individuals or groups.

**Parallel Coordinates**: as defined in the paper [38] “is a widely used visualization technique for multivariate data and high-dimensional geometry.” It has been a popular visualization technique for multivariate data [39].

This method allows the presentation high-dimensional geometry of data. The first criterion is met.

2D parallel coordinates method only allows the identification of relationships between adjacent axes [40] that is why many approaches have been proposed (figure 14):

- Methods that keep 2D parallel coordinates and arrange the axes to show the relationships or to reduce clutter.
- Techniques that extend parallel coordinates from 2D to 3D to show many relationships simultaneously.
- **Streamgraph** as defined in [34] is a “type of a stacked area graph, which is displaced around a central axis, resulting in flowing and organic shape.”

Series of similar events are displayed in the timeline. The first world war is abstracted using this method by Abi-Haidar[41]. Unstructured text is supported by this method.

This method supports one datadimension, but it can be applied to large data [33].

**Circular Network Diagram** (chord diagram visualization) [37]: this chart visualizes the inter-relationships between entities (figure 13). Data object are placed around a circle and linked with curves based on the rate of their relativeness. Color can be used to group the data into different categories, which aids in making comparisons and distinguishing groups.

TreeMap [35]: is represented by a root rectangle, divided into groups, also represented by the smaller rectangles (figure 10), which correspond to data objects from a set [35]. This method of visualization is used for hierarchical data two-dimensional.

The treemap method can be applied to large data volumes; iteratively representing data layers for each level of the hierarchy. This method satisfies the large data volume criterion. However, the method can only show two data dimensions presented by size and color shapes. And the data representation appears at one moment in time. So the criterion data variety and dynamicity are not met in this method.

**Circle Packing**: it is an alternative to Treemap that uses circles instead of rectangles.

The Primitive shape is a circle; which can include circles as presented in the (figure 11). The most advantage of this method is the possibility to place and perceive a lot of objects with many levels of hierarchy. The area of each circle presents an attribute such as quantity. Color may be used to present the second fact. This method looks more beautiful, but it is not as space-efficient as a Treemap, as there is a lot of empty spaces within the circles.

**Sunburst**: This method is a directive of treemap: it converted to a polar coordinate system (figure 12). It is more flexible and allows repainting the whole diagram by changing the radius and arc length. Thanks to this property; this method can quickly show data dynamics using animation. It allows understanding large amounts of data using efficient and intuitive graphic [36].

**E. Other Approaches**

The new approaches to big data visualization tend towards simplification and improvement in forms of images, diagrams or animation [1]. In this vision, the author presents techniques to better present big data.

1) **Tag cloud**: also called word cloud, this technique is used to visualize text analysis: word, size, color and position indicate characteristics of the word: frequency or prominence. The perception of the most prominent terms in the text is faster using this method (figure 15).

2) **Clustergram**: this graph is used in cluster analysis for non-hierarchical clustering algorithms like k-means and hierarchical cluster algorithms when the number of observations is large [42].

3) **Motion charts**: it is a dynamic bubble chart which allows active exploration of large and multivariate data and interacts with it (figure 17). These tools: Google [32], amCharts and IBM Many Eyes provide motion chart.

4) **Dashboard** [43]: This technique allows visualizing server logs in real time (figure 18). Log files can be of various formats. The dashboard consists of three layers: data (raw data), analysis (includes formulas and imported data from data layer to tables) and presentation (graphical representation based on the analysis layer).
IV. DATA VISUALIZATION DESIGN

The first section provides an overview of Data Visualization and looks at currently used methods for presenting different types of Data. The second section indicates the main challenges and issues in Big Data Visualization and provides some approaches to master them.

In further section: a brief background of the importance of visual perception is given. Color importance and techniques of choosing color are discussed.

A. Visual Perceptions:

The information is very dependent on the way in which it is presented[44]. Human being tends to comprehend visual information quicker than raw data is very true, but this does not mean that all visualizations are understandable with the same degree [45].

On the other hand, Christopher g. Healey[46] explains that we can improve the quality and quantity of displayed information by perception understanding. Authors in their paper[47] suggest that hard efforts are made in recent years to benefit from human visual perception properties in visualization design.

On the other hand; many research works have shown that “a limited set of visual properties that are processed preattentively”[48]. Preattentive properties or features offer various benefits with little effort.

Healey[49] cites some advantages of preattentive features as -speed: preattentive tasks can be performed in 200 milliseconds -independence of the size: time needed to carry out preattentive task is not dependent on display size.

The following table (Table 2) lists some preattentive features and provides references that describe the tasks that can be performed using these features [49].

<table>
<thead>
<tr>
<th>Feature</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>line (blob) orientation</td>
<td>Julez &amp; Bergen (1983); Wolfel (1992)</td>
</tr>
<tr>
<td>length</td>
<td>Triesman &amp; Gorman (1988)</td>
</tr>
<tr>
<td>width</td>
<td>Julez (1984)</td>
</tr>
<tr>
<td>size</td>
<td>Triesman &amp; Gelade (1980)</td>
</tr>
<tr>
<td>curvature</td>
<td>Triesman &amp; Gorman (1988)</td>
</tr>
<tr>
<td>number</td>
<td>Julez (1983); Trick &amp; Pylyshyn (1994)</td>
</tr>
<tr>
<td>terminators</td>
<td>Julez &amp; Bergen (1983)</td>
</tr>
<tr>
<td>intersection</td>
<td>Julez &amp; Bergen (1983)</td>
</tr>
<tr>
<td>closure</td>
<td>Enns (1986); Triesman &amp; Souther (1986)</td>
</tr>
<tr>
<td>color (hue)</td>
<td>Triesman &amp; Gorman (1988); Nagy &amp; Sanchez (1990); D’Zmura (1991)</td>
</tr>
<tr>
<td>intensity</td>
<td>Beck et al. (1983); Triesman &amp; Gorman (1988)</td>
</tr>
<tr>
<td>flicker</td>
<td>Julez (1971)</td>
</tr>
<tr>
<td>direction of motion</td>
<td>Nakayama &amp; Silverman (1986); Driver &amp; McLeod (1992)</td>
</tr>
<tr>
<td>binocular lustre</td>
<td>Wolfe &amp; Frazel (1988)</td>
</tr>
<tr>
<td>stereoscopic depth</td>
<td>Nakayama &amp; Silverman (1986)</td>
</tr>
<tr>
<td>3-D depth cues</td>
<td>Enns (1990)</td>
</tr>
<tr>
<td>lighting direction</td>
<td>Enns (1990)</td>
</tr>
</tbody>
</table>

Table 2: Lists of preattentive features and references that describe the tasks that can be carried out using these features.

The next topic discusses a color feature and how it can be used to perform data visualizations shapes.

B. Color: Overview and importance

Color plays a significant role in the daily life; people make a hard effort to surround their entourage with the colors [50]. It impacts emotions and impacts performance [51].

Color enchants the life: it used in many disciplines like science and engineering not only for aesthetics reasons but also for effectiveness.

In the data visualization; Color is a powerful and often-used visual feature [50].

Color can be an illustrative and informative in data visualization. For this reason, the specialists are quite cognizant of the importance to work on algorithms and rules that orient designer for efficient design [47].

Each method of the data visualization already listed in the previous parts uses colors to become more ready, faster to understand and to memorize.

Netflix [52] is a highly successful organization with 48 million subscribers and a market cap of nearly $26 billion. It has utilized color analysis in digital video/image to catch more and more users (figure 19).
Complement colors are placed directly opposite each other on this wheel. Two balanced colors combined will become neutral or gray [55].

Figure 22: Tints, shades, and tones of five different colors [51]

Another term is involved when we are speaking about data visualization design: the legibility. Legibility of data visualization is the amelioration of the quality of being bright enough to read. The issue of legibility comes in data visualization when complex small shapes are used [56]. Many factors impact the legibility in data visualization as a font text, font contact, and spacing [56]. The difference in value between the symbol (text, line, etc.) and its background determines the legibility [54]. Though Hue and Chroma do not contribute in the legibility, while the luminance contrast does.

C. Color selection

1) Color Brewer

Many color palettes and algorithms for color choosing are developed recently. This topic receives attention increasingly because as cited by Stone [51]: "A color used well can enhance and clarify a presentation. A color used poorly will obscure, muddle and confuse".

To select the right combination of a palette; users refer to many available palettes as ColorBrewer. ColorBrewer [57] is an online tool that helps users to choose the appropriate color. It contains 35 color sets divided into three groups: qualitative, sequential and diverging [57].

Sequential Color Schemes (figure 23-a): it can be presented by one single hue (6 available schemes) or multi-hue (12 6 available schemes).

Diverging color schemes (figure 23-b): as explained in his paper [57] this type “should be used when a critical data class or breakpoint needs to be emphasized.” Diverging schemes are always multi-hue sequences.

Qualitative Color Schemes (figure 23-c): for a successful presentation of qualitative schemes it must vary hue and keep saturation and lightness constant.

Figure 23: ColorBrewer contains 35 color scheme sets. (a) sequential, (b) diverging and (c) qualitative [57].

In this section, we define most common terms used in color design.

The color is specified by three dimensions:
Hue: is the color’s name, such as red, green or orange. We distinguish three types of colors: primary, secondary, tertiary.
Value: is perceived lightness or darkness of the color.
Chroma: describes its colorfulness; High Chroma colors are vivid or saturated, little Chroma colors are grayish or muted [54].

The hue dimension is circular, typically drawn as a hue circle (Figure 20).

In color design we distinguish two different terms:
Contrast: contrasting colors are different, and opposites and contrasting hues are on the opposite side of the hue circle.
Analogy: analogous colors are similar; and Analogous hues are close together, most simply variations of the same color name. According to [54], Contrast draws attention, analogy groups.

As an example, the author shows He shows in figure 21 that the red squares catch the attention and stand out from blue-green (analogous field) squares ones.

Figure 21: Contrast and analogy. The red squares contrast with the analogous blue-green ones.

On the other hand; the term tint applies to a color that has been lightened and desaturated by mixing it with white. A tone is a color grayed and darkened by mixing it with black.
The purpose of this paper is to prepare a background to develop successful data visualization methods. The right choice of these methods must go far to let users better understanding and complete view of the data. In this paper, we have invested to classify the data visualization methods according to many factors. As we disused many methods are improved, and some others are developed to respond to big data challenges. Moreover, the study extends to obtain relevant color rules used to improve the big data visualization. The right choice of these corresponding elements will make a successful visualization. The purpose of this paper is to prepare a background to develop a datavisualization tool for the XEW [59]platform and extend XEWgraph[60].

APPENDIX

REFERENCES

[18] SAGIROGLU and D.SINANC; ”big data: a review” 2013
[23] Fernanda B. Viegas; Martin Wattenberg; Frank van Ham; Jesse Kriss; Matt McKeon; ManyEyes: a Site for Visualization at Internet Scale;2007
[26] www.cytoscape.org(last checked January 2017)
[33] Lindong Wang1, Guanghui Wang2, Cheryl Ann Alexander3, Big Data and Visualization: Methods, Challenges and Technology Progress;2015
[41] Matthias Schonlau; Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams;2004
[50] Andrew J. Elliot & Markus A. Maier; “Color Psychology: Effects of Perceiving Color on Psychological Functioning in Humans”; 2014
[54] STONE, Maureen C. A survey of color for computer graphics. Course at SIGGRAPH, 2001, p. 8-17
[56] M Harrower, CA Brewer; “ColorBrewer.org: an online tool for selecting color schemes for maps” 2003