Overview of Data Visualization

O. Ben Ali and M. El Haddad, *National School of Applied Sciences Tangier* ,Z. Boulouard and L. Koutti, *Faculty of Sciences Agadir*, F. Bouhafer and A. El Haddadi, *National School of Applied Sciences Al-Hoceima*

***Abstract* — Data visualization can be defined as data transformation into interactive visual representations. It is very important since it allows users to have an insightful vision on a subject that might interest thems. The Big Data phenomenon has urged scientists to develop and dedicate an entire research field to data visualization since it allows the user to easily have an idea on the content provided by his different data sources, based on his visual abilities. In this paper, we will present an overview of the literature related to this topic starting by its definition then moving to its challenges and later, presenting its methods and comparing some of its most used tools**

***Index Terms* — Big Data, Data, Data Classification, Data Extraction, Data visualization tools, Database, Decision Making, Multimedia Database, Relational Database, Semi-Structured Data, Structured Data, Unstructured Data, XML.**

I. Introduction

With over 2,5 million terabytes of data created on a daily basis nowadays, the human mind has become overwhelmed with the information flood generated from every source we can imagine, from climate sensors to social media posts and scientific publications. In order to overcome this problem, the best solution is to make a better use of our eyes.

Data visualization is a tool that allows us to see the patterns and connections that matter and then designing that information so that it would make more sens or tell a story or help us focus only on the most relevant information.

Data can tell us amazing stories and data visualization is a powerful means to, on the hand, discover and understand these stories, and on the other hand, to share them with others. Whether it concerns politics, sales, sports, etc…, statistical information, usually presented with words and numbers, remains abstract. Yet, we can still display it visually, and in order to do this, we must give it a form.

For example, we would like to discover the evolution of the minimum hourly wage in Morocco over the last decade [1], the numbers in table 1 can certainly describe it but we need to focus on them in order to find a pattern, while the chart in figure 1 can display that same pattern more quickly and more accuratly.

TABLE I. A Table Displaying the Evolution of the Minimum Hourly Wage in Morocco from 1994 to 2015 [1]

Year	Minimum Wage (MAD per Hour)
1994	7,26
1996	7,98
2000	8,78
2004	9,22
2004	9,66
2008	10,14
2009	10,64
2011	11,7
2013	12,24
2014	12,85
2015	13,46

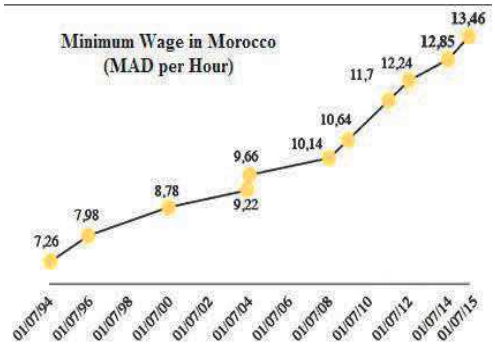

Fig. 1. A Chart Displaying the Evolution of the Minimum Hourly Wage in Morocco from 1994 to 2015 [1]

Another similar example, let's say we want to find out which Hollywood movies belong to the same genre and which of them made more income or had more budget. We definitely can answer this query by going through the tables and numbers from IMDb, but we easily get overwhelmed by the information glut. Another way to do it, is to translate these qualitative and quantitative data into shapes, colors, sizes, etc… So if we go back to our example, the movies can be shaped as round bubbles colored according to their genre and sized according to their income, the budget along with more useful information such as title, producers, etc… can be displayed in a tip that may show whenever we hover the movie's bubble.

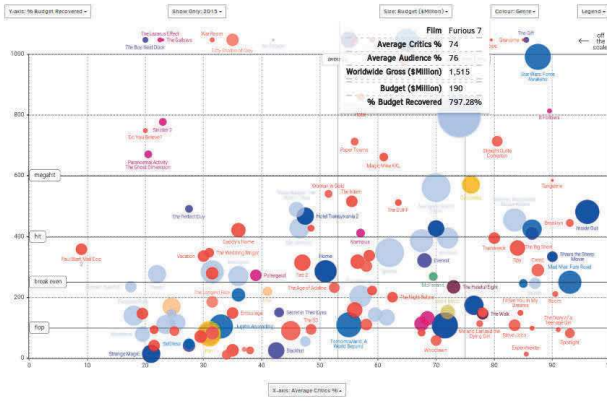

Fig. 2. Movies Explorer [2]

In the following paragraphs of this survey, we will define the roles of data visualization as well as its challenges, then we will have an overview over its methods and most used tools and technologies

II. Data Visualization: Roles and Challenges

A. *Roles of Data Visualization*

According to Mao [3], the iterative process of questioning and extracting knowledge out of data is called "Exploratory Data Analysis (EDA)". EDA requires heavy data manipulation and as such, the analyst can easily be misleaded. For this reason, EDA relies mostly on efficient and expressive data visualization. This urged Munzner [4] to define rules of effectiveness and expressiveness of data translation. These rules were used by Mao [3] as guidelines in order to ease the EDA process.

Graphs for example use the human's perceptual system in order to ease the decision process as stated by Larkin and Simon [5]. It is done by grouping together all relevant information in more efficient way than pure text and thus, avoiding a large number of search of non-relevant information.

B. *Challenges of Data Visualization*

With the data being more overwhelming, a good data visualization needs to be precise and at the same time able to provide an overall picture of the data it represents.

1) Expressiveness and Effectiveness

According to Munzner [4], a good data visualization needs to have two essential components, marks and channels.

Marks are basic geometric elements that depict data observations. Figure 3-1 is an example for marks that include points, lines and areas.

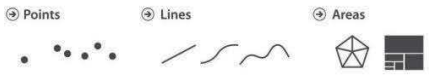

Fig. 3-1. Marks [4]

Channels control the marks positions, sizes, colors etc… as described in Figure 3-2.

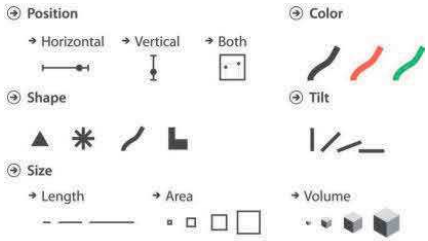

Fig. 3-2. Channels [4]

Munzner's guideline [4] in Figure 3-3 has categorized channels into identity channels and magnitude channels. Identity channels describe what the object is and how it is designed, they are more suitable for qualitative variables. Magnitude channels describe how much of an object there is, so they are more suitable for quantitative variables. The effectiveness of the channels is ranked from "most effective" at the top to "least effective" at the bottom.

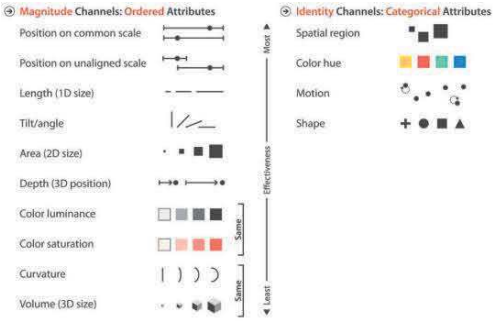

Fig. 3-3. Munzner's guideline [4]

2) Human Perception and Screen Limits

According to Agrawal et al. [6], human eyes cannot easily get meaningful information when the data exceedes a certain limit, and very few systems available nowadays are able to display it in a manner that would ease the human perception.

Frölich and Plate [7] argued that the growth of data visualization becomes challenging when it gets limited by the screen dimensions or resolution. A statement confirmed by Travel [8] who said "Given the resolution of conventional displays, visualizing every data point can lead to over-plotting, overlapping, and may overwhelm user's perceptional and cognitive capacities".

3) Performance

Gorodov and Gubarev [9] have stated several problems that Big Data Visualization can face, especially when it comes to performance. These challenges are described as follows:

- Visual Noise: It happens when the data is too connected to be easily extracted or separated on the screen.
- Information Loss: Can be caused by the reduction methods used to extract the most relevant data.
- Large Image Perception: Lower resolution screens can limit the produced visualizations.

4) Data Visualization VS Big Data's 4Vs

One of the most common definitions of Big Data is the one proposed by IBM's data scientists [10] who defined it as data able to respond to four criteria or axes. These axes are also

known as "The 4Vs of Big Data".

Intel [11] has turned big data's 4Vs into the following data visualization opportunities:

- Volume: The data visualization must be able to derive meaning from large volumes of data.
- Variety: The data visualization must be able to combine as many data sources as needed.
- Velocity: The data visualization must enable businesses to replace batch processing with real-time stream processing.
- Value: The data visualization must enable users to create business value by gaining insights from big data.

III. Data Visualization Methods

In this section, we will first talk about the data visualization pipeline, then move on to presenting conventional data visualization methods along with the extension of some of them to handling the Big Data challenges introduced earlier.

A. Data Visualization Pipeline

Introduced by Liu et al. [12], it describes the process of data visualization and presents it as a pipeline of five modules:

- Data Transformation and Analysis: Extracts relevant data out of the input data.
- Filtering: Selects the data to be visualized.
- Mapping: Matches the filtered data with geometric shapes (points, lines, etc…) and their attributes (color, size, etc…).
- Rendering: Transforms geometric data into image data.
- UI Controls: Ease the user's interaction with the data and its comprehension.

This pipeline is described in Figure 4.

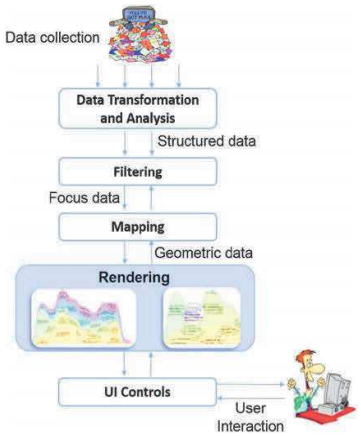

Fig. 4. Data Visualization Pipeline [12]

B. Data Visualization Methods

1) Conventional Methods:

Khan et al. [13] have cited many conventional data visualization methods that are often used, such as tables, histograms, scatter plots, line charts, bar charts, pie charts, area charts, flow charts, bubble charts, multiple data series or combination of charts, time line, Venn diagrams, data flow diagrams, and entity relationship diagram, etc…

They have also mentioned some data visualization methods that have been used although they are less known compared to the other precited ones, such as parallel coordinates, treemaps, cone trees, and semantic networks.

2) Extension to Big Data:

According to Shull [14], an optimized Big Data Visualization should start by giving a general overview, then allowing filters, zooms and any other detailed information on demand.

Gorodov and Gubarev [9] have proposed several approaches to visualize big data, such as multiple views per display, dynamic factors alteration and dynamic filtering, etc… These methods were later categorized according to Big Data's 3Vs (Volume, Variety, Velocity). The result of their classification is as follows:

TABLE II. Classification of Visualization Methods [9]

Method Name	Big Data Class
Treemap	Can only be applied to hierarchical data
Circle Packing	Can only be applied to hierarchical data
Sunburst	Volume + Velocity
Parallel Coordinates	Volume + Velocity + Variety
Streamgraph	Volume + Velocity
Circular Network Diagram	Volume + Velocity

The material's performance has an important impact on the quality of a data visualization method. As such, many attempts to adopt state of the art technologies to these methods were adopted.

Wang et al. [15] have stated that many big data visualization tools were based on Hadoop architecture. These tools provide efficient data analysis, yet, they lack adequate visualization.

Liu et al. [16] have proposed a web-based data analysis tool based on WebGL in order to process data using GPU.

IV. Data Visualization Tools

The literature provides us with with many Data Visualization tools, and in this section, we will try to present the most popular ones.

Suda and Hampton-Smith [16] brought a review identifying 38 data visualization tools that are, according to them, the most used in 2016. Among these tools we will choose to cite the ones that were most cited in the recent reviews.

Sharma [17] has classified these tools into simple user oriented and developer oriented.

A. Simple User Oriented Data Visualization Tools

1) Tableau:

Tableau [18] is a commercial data visualization tool that allows users to easily drag and drop data into the system and have a real-time interaction with different types of visualizations (graphs, charts, maps, etc…). It also allows web embedding and team collaboration.

2) RAW:

Raw [19] is an open source customizable tool that lets users create vector-based data visualizations. It allows Data uploads from different sources such as apps and computers. The visualizations are later exported as SVG or PNG and embedded to the webpage.

3) TimelineJS:

TimelineJS [20] is an open source tool that allows users to create rich and interactive timelines based on data sources provided as Google spreadsheets.

B. Developer Oriented Data Visualization Tools

1) D3JS:

D3JS [21], short for "Data Driven Documents for JavaScript" is an open source JavaScript library that allows rendering amazing charts out of varied data sources using HTML, SVG, and CSS. This library, is capable of some seriously advanced visualizations with complex data sets and allows easy interaction and sharing.

2) FusionCharts:

FusionCharts Suite XT [22] comes with more than 90 chart types, over 965 data-driven maps, and ready-made business dashboards and demos. It is based on a JavaScript API that enables integration with any AJAX application or JavaScript framework. FusionCharts is highly interactive, customizable and multiplatform.

3) Google Charts:

Google Charts [23] is a chart visualization solution used by a majority of users throughout the web. It is highly flexible and provides an excellent set of developer tools behind it. It becomes very handy when it comes to specialist visualizations such as geocharts and gauges. It also includes built-in animation and user interaction controls.

4) HighCharts:

HighCharts [24] is a JavaScript charting library that provides the user with a wide range of charts types he might use. The output rendering adapts to the user's browser since it provides SVG rendering for modern browsers and VML rendering for Internet Explorer. The charts are beautifully animated and the framework supports live data streams. It is free to download and non-commercial use (and licensable for commercial use).

V. A COMPARATIVE REVIEW OF DATA VISUALIZATIONS TOOLS

With the important number of visualization tools available, it can be confusing to choose the right tool for the right purpose. This section presents some of the most popular open source JavaScript libraries and the possibilities offered by each one of them.

These Data visualizations tools can be provided as services, platforms, widgets or libraries.

A. Data visualization tools types

Leishi Zhang [25] differentiates two types of tools:

1) Open Source Toolkits

A number of open-source Visual Analytics toolkits exist; each covers a specific set of functionalities for visualization, analysis and interaction.

The functionality of these toolkits is compared according to three criteria: (1) visualization functions, (2) analysis capabilities, and (3) supported development environment [25].

2) Commercial Visual Analytic Systems

An alternative is to resort to software suites which integrate required functionality in software systems that work either standalone, or integrate, more or less seamlessly, into an existing information infrastructure. Example systems include Tableau [18], Spotfire [26] , and QlikView [27]. Commercial toolkits typically require no or only limited configurations or program adjustments, to become operational.

The principal goal of the survey is to provide to developers a reference for choosing the adapted library to given problem.

We will do it by highlighting the visualizations techniques provided by each of these tools. We divide visualization techniques into graphical representations of data and interaction techniques and we move to compare systems and architectures supported for each tools.

B. Functional comparison:

According to Leishi Zhang [25], we divide visualization techniques into graphical representations of data and interaction techniques to analyze the visualization functionality of each system.

The visual part is related to the form which the data is displayed like: bar chart, line chart, pie chart.

Interaction technique like zooming, panning, etc..

1) Graphical representation:

a. Numerical data

There are several methods to present the relationship between Numerical data, the most common are: bar chart, line chart, pie chart and scatter plots these methods are especially used for data with low dimensions. However, for data with higher dimensionality, some other methods are used like: parallel coordinates, heat maps and scatter plot matrix.

b. Text web:

Word cloud or tag cloud is a way to visualize text data. It's the most used way to present the qualitative data because it makes it easy to spot word frequencies.

The word that appears in bold in front is the most used in an interview, a document or a text. Usually Tags text is single words shown with font size or color.

This method helps to collect quickly the most important and interesting terms.

It is appreciated for being more captivating for the audience and make a quick impact, Moreover it is simple, easy to understand and easy to share.

c. Geo related data

Geo-visualization short for Geographic Visualization, is the creation and use of visual representations to facilitate thinking, understanding, and knowledge construction about human and physical environments, at geographic scales of measurement[28].

The important stories that numbers have to tell often involve location [29].So it will be easier for the analyst that they be represented on a map.

Visualizations tools are challenged to provide geographical projections methods to respond to user's need.

TABLE III. VISUALIZATION AVAIBILTY OF VISULAIZATION TECHNIQUES INSOME DATA VISULAIZATION TOOLS

	Numerica data					Geo related data	Netowrk data		Word clouds
	bar-line-pie-chart-histogram	scatterplot	heatmaps	Parallel coordinates	scatterplot Matrix	Projection on Map	tree map	Others graphs	
Highcharts	✔	✔	✔	✔		✔	✔	✔	-
D3	✔	✔	✔	✔	✔	✔	✔		✔
google charts	✔	✔	-	✔	-	✔	✔	✔	✔
chartjs	✔	✔	-	-	-	-	-	-	-

d. Network data (graph):

Graphs and networks visualization is an important part of data visualization because it is used in many applications to visualize interactions within many networks in different fields including biology (biological network analysis), finance, web sociology (social media networks, friendship and acquaintance networks, collaboration..) network theory, etc..

Networking graph can be based on hierarchical components i.e parent-child relations between data items as well as non-hierarchical components [30].

In his paper, Danny Holten [30] proposes some Tree Visualization Techniques to present graph in case of hierarchical structure such as treemap and the hyperbolic view.

2) Interactivity:

Having a data visualization interface of allowing the user interaction is a critical need.

Muzammil and Sarwar describe on their paper [31] that users can interact with interfaces or visualization in different ways by means of mouse over, single click, double click, or can add multiple interactive options by mouse right button click. There are many interactive techniques available to interact with charts or graphical representation to understand the drill down details.

Card et al introduce the interactive mechanism of visualization in 1999. This mechanism describes human interaction indifferent phases as shown in the following diagram.

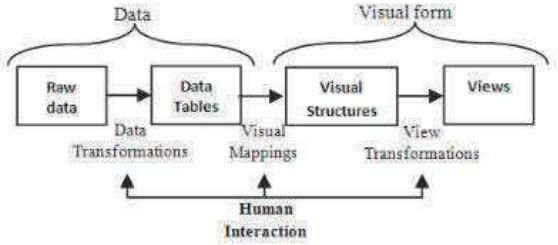

Fig. 5. Interactive Mechanism [31]

This table summarizes some interactivity action for each tool:

TABLE IV. A TABLE DISPLAYING SOME INTERACTIVITY ACTIONS FOR JS TOOLS [32]

	Ability to zoom in and out of charts	Annotations on the chart	Combination of charts	Interactive (responds to mouse hover/click)
Highcharts	✔	✔	✔	✔
D3	✔	-	✔	✔
google charts	-	✔	✔	✔

VI. CONCLUSION

Data visualization changes the way people experience information and the way we live since it allows us to have an in-depth exploration regarding several issues: mobility, social media, geolocalization, data mashups, big data, globalization and science communication.

In this paper, we have defined the roles of Data Visualization and then dove into its challenges from human perception to performance and to handling big data issues. We have also presented a survey on state of the art Data Visualization methodologies and tools.

As shown, data visualization methods were categorized into conventional and big data oriented, while the tools were classified into simple user oriented and developer oriented.

REFERENCES

[1] http://www.leconomiste.com/article/940681-la-hausse-du-smig-en-deux-temps (last checked 12/04/2016).

[2] http://www.informationisbeautiful.net (last checked 12/04/2016).

[3] Y. Mao "Data Visualization in Exploratory Data Analysis: An Overview of Methods and Technologies", University of Texas at Arlington, 2015.

[4] T. Munzner, "Visualization Analysis & Design", Boca Raton, FL: CRC Press, 2014.

[5] J. H. Larkin and H. A. Simon, "Why a Diagram is (Sometimes) Worth Ten Thousand Words," Cognitive Science, pp. 65-100, Jan. 1987.

[6] R. Agrawal, A. Kadadi, X. Dai, F. Andres, "Challenges and Opportunities with Big Data Visualization", 7th International Conference on Management of computational and collective intelligence in Digital EcoSystems, ACM Conference on, Caraguatatuba, Brazil, 2015, pp. 169-173.

[7] B. Frölich, J. Plate, "The cubic mouse: A new device for three-dimensional input", Human Factors in Computing System, Proceedings

of the CHI '00 SIGCHI Conference on, The Hague, The Netherlands, 2000, pp. 526-531.
[8] P. Travel, "Modeling and Simulation Design", AK Peters Ltd. MA., 2007.
[9] E. Y. Gorodov, V. V. Gubarev, "Analytical Review of Data Visualization Methods in Application to Big Data", Electrical and Computer Engineering Journal of., Article ID 969458, 2013, pp. 1-7.
[10] http://www.ibmbigdatahub.com/infographic/four-vs-big-data (last checked 16/04/2016).
[11] Intel IT Center, Big Data Visualization: "Turning Big Data Into Big Insights", White Paper, March 2013, pp. 1-14.
[12] S. Liu, W. Cui, Y. Wu, M. Liu, "A Survey on Information Visualization: Recent Advances and Challenges", The Visual Computer, vol. 30, 2014, pp. 1373-1393.
[13] M. Khan, S.S. Khan, "Data and Information Visualization Methods and Interactive Mechanisms: A Survey", International Journal of Computer Applications, 2011, pp. 1-14.
[14] F. Shull, "Getting an Intuition for Big Data", IEEE Software, July/August 2012, pp 1-5.
[15] L. Wang, G. Wang, C. A. Alexander, "Big Data Visualization: Methods, Challenges and Technology Progess", Digital Technologies, vol. 1, Science and Education Publishing, 2015, pp. 33-38.
[16] http://www.creativebloq.com/design-tools/data-visualization-712402 (last checked 25/04/2016).
[17] http://thenextweb.com/dd/2015/04/21/the-14-best-data-visualization-tools/ (last checked 25/04/2016).
[18] https://public.tableau.com/s/ (last checked 25/04/2016).
[19] http://raw.densitydesign.org/ (last checked 25/04/2016).
[20] http://timeline.knightlab.com/ (last checked 25/04/2016).
[21] https://d3js.org/ (last checked 25/04/2016).
[22] http://www.fusioncharts.com/ (last checked 25/04/2016).
[23] https://developers.google.com/chart/interactive/docs/ (last checked 25/04/2016).
[24] http://www.highcharts.com/ (last checked 25/04/2016).

[25] Zhang, Leishi et al. "Visual analytics for the big data era—A comparative review of state-of-the-art commercial systems." Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on 14 Oct. 2012: 173-182
[26] http://spotfire.tibco.com/ (last checked 01/06/2016).
[27] http://www.qlik.com/ (last checked 01/06/2016).
[28] http://map.sdsu.edu/geog104/unit-9.html (last checked 10/06/2016).
[29] Few, S. "Introduction to Geographical Data Visualization - Perceptual Edge." 2009.
[30] Holten, Danny. "Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data." Visualization and Computer Graphics, IEEE Transactions on 12.5 (2006): 741-748.
[31] Muzammil Khan. Data and Information Visualization Methods, and Interactive Mechanisms: A Survey.International Journal of Computer Applications (0975 – 8887).Volume 34– No.1, November 2011
[32] http://socialcompare.com/en/comparison/javascript-graphs-and-charts-libraries (last checked 10/06/2016).