Big Data Visualization Tools and Techniques Survey

Pramath Parashar
pparasha@kent.edu

Abstract

Before the emergence of the concept of big data visualization, data visualization has been widely used. From population data to student performance statistics, can be visualized to display and explore the rules. Big data can now be visualized in a variety of ways, each with a different focus. The purpose of Big data visualization is to "let data speak", to use graphics to tell the story of data. Visualization is a way of representing data, an abstract representation of the real world. Like words, it tells us all kinds of stories. Visualization as a medium has evolved into a great way to tell stories. This survey report on big data visualization summarizes the Techniques and strategies of big data analysis and the tools of big data visualization mentioned in a large number of academic papers.

SECTION I :Introduction

Big data visualization is a very broad concept, mainly including scientific visualization, information visualization and visual analysis. The integration of these three branches forms a new discipline 'big data Visualization', which is a new starting point in the field of visualization research.

Data visualization in the broad sense involves information technology, natural science, statistical analysis, graphics, interaction, geographic information and other disciplines.

Among them, scientific visualization is an interdisciplinary research and application field in science. It mainly focuses on the visualization of three-dimensional phenomena, such as various systems in architecture, meteorology, medicine or biology. The important thing is the realistic rendering of the body, surface and light source, etc. The purpose is to illustrate the scientific basis in a graphical way, so that scientists can understand, explain and collect rules from the data.

Information visualization is the study of interactive visual representation of abstract data to enhance human cognition. Abstract data includes both digital and non-digital data. Such as geographic information and text. Bar charts, trend charts, flow charts, and tree charts are all part of information visualization. The design of these graphs transforms abstract concepts into visual information.

Visual analysis is a new field formed with the development of scientific visualization and information visualization, focusing on analysis and reasoning through interactive visual interface.

Why do we need big data visualization? Humans use vision to capture more information than any other organ, and big data visualization is about harnessing innate human skills to enhance the efficiency of data processing and organization. Big data visualization can help us process more complex information and enhance memory.

Finally, how do we achieve big data visualization? That's the focus of our survey. In this survey report on big data visualization, we sorted out the techniques and strategies of big data visualization from 41 papers, introduced the tools mentioned in the paper to help us achieve big data visualization, and proposed the current challenges of big data visualization and feasible solutions in the future.

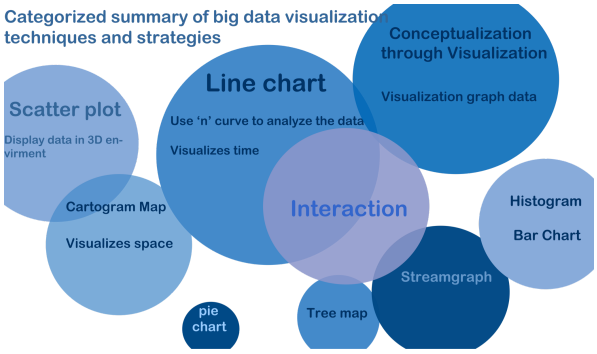


Fig1 : categorized summary of big data visualization techniques and strategies.

SECTION II :Background

Due to the overwhelming size of the big data visualization field, the background of big data visualization can be organized in 2 different aspects. The brief fields in data visualization can be seen in Fig.1.

The first aspect of big data visualization is the visualization tools. The basic chart of the visualization as well known as plot, scatter, histogram and so on. Not only the basic format of the data visualization, There are various types of data such as time series, graph, image, video and so on. To handle those data types to visualize in a more creative and convenient way, data visualization tools are developed in different environments.

The second aspect of big data visualization is the visualization techniques. Mainly the technique is about managing the huge data sources in more efficient ways such as indexing and map-reduce algorithms. Secondly, to display the data more distinctive and understandable, methods on transforming data structure, classification between data relationship and data prediction are used. In addition, machine learning technology is aimed here to solve these research questions more efficiently. Moreover, user interaction on the data visualization is one of the methods to enhance the understanding and useability.

From the following sections, we state challenges of the current research questions in big data visualization and introduce the related paper on each technique in this survey.

SECTION III : Challenges

Big Data mining presents a plethora of appealing possibilities. However, while exploring Big Data sets and extracting value and knowledge from such information mines, researchers and professionals face a number of problems.

Due to the volume, variety, and velocity of data, big data visualization is difficult. The most difficult aspect of working with big data is navigating massive data volumes and efficiently displaying meaningful and usable data visualization and analysis results. Data capture, storage, searching, sharing, analysis, management, and visualization are just a few of the challenges. In addition, there are concerns about security and privacy, particularly in distributed data-driven applications. Often, the avalanche of data and scattered streams outstrips our ability to manage them. While the size of Big Data continues to grow at an exponential rate, the existing technological capacity to manage and analyze Big Data sets is limited to petabytes, exabytes, and zettabytes of data.

New mechanisms must be developed to examine data in order to assist decision makers in gaining understanding from it in a simple and straightforward manner using graphs and charts. Traditional visualization techniques are incapable of dealing with extremely huge data sets. The visualization tool should be able to provide us with visuals with the least amount of latency as feasible. Parallelization is also essential to deal with such a large volume of data, which is a hurdle in visualization.

Big data visualization's main purpose is to find intriguing patterns in large amounts of data. Data dimensions must be properly selected for pattern

mining. If we only utilize a few dimensions, our visualization will be limited, and many intriguing patterns will be lost; similarly, if we use all of the dimensions, our visualization will be thick, and the users will not find it useful. "Given the resolution of standard displays (1.3 million pixels), showing every data point might lead to over-plotting, overlapping, and may overload user's perceptual and cognitive capacities," according to one study. The majority of today's visualization technologies perform poorly in terms of scalability, usefulness, and reaction time.
Based on this , we can categorize the big data challenges.**[11][13][16][18][20]**

Big Data cleaning
Those five stages (cleaning, aggregation, encoding, storage, and access) are not novel in traditional data management and are well-known. The issue with Big Data is figuring out how to deal with the complexity of Big Data's nature (speed, volume, and variety) while processing it in a distributed environment with a range of applications. In reality, verifying the legitimacy of sources and data quality before engaging resources is critical for valid analytical outcomes. Data sources, on the other hand, may contain noise, inaccuracies, or incomplete data. The problem is determining how to clean such large data sets and determining whether data is accurate and helpful.

Big Data aggregation
Another problem is synchronizing external data sources and distributed Big Data platforms (such as applications, repositories, sensors, networks, and so on) with an organization's internal infrastructures. The majority of the time, analyzing data generated within businesses is insufficient. It's critical to go a step further and combine internal and external data sources in order to extract significant insight and knowledge. Third-party sources, market fluctuations, weather forecasting, and traffic conditions, data from social networks, consumer remarks, and citizen feedback are all examples of external data. This can help, for example, to improve the prediction model strength used in analytics.

Imbalanced systems capacities
The architecture and capacity of computers is a significant topic. Indeed, Moore's Law states that the performance of CPUs doubles every 18 months, and that the performance of disk drives doubles at the same pace. The I/O operations, on the other hand, do not follow the same performance pattern. (For example, random I/O speeds have improved slightly, whereas sequential I/O speeds have increased slowly as density has increased) (Chen and Zhang, 2014). As a result of the uneven system capacities, data access may be slowed, affecting the performance and scalability of Big Data applications. From a different perspective, we may see the various device capacities on a network (e.i, sensors, disks, memories). This may cause the system to slow down.

Imbalanced Big Data
Another difficulty is classifying datasets that are unbalanced. In recent years, this subject has gotten a lot of attention. In reality, real-world applications may yield classes with varying distributions.

The first sort of class is one that has a small number of instances and is underrepresented (known as the minority or positive class).
The second class, which has a large number of instances, is (known as the majority or negative class). Identifying minority groups is critical in a variety of industries, including medical diagnosis and software flaw identification.

Big Data analytics
Big data has enormous prospects and transformative potential for a variety of industries; nevertheless, it also poses unprecedented hurdles in terms of utilizing such massive and growing volumes of data. Advanced relationships between features and data exploration For example, data analysis allows a company to get useful insight and track patterns that could have a beneficial or bad impact on the company. Other data-driven applications, such as navigation, social networks, finance, healthcare, astronomy, and intelligent transportation systems, require real-time analysis as well. To achieve reliable results, monitor changes in diverse sectors, and predict future observations, innovative algorithms and effective data mining methods are required.

Big Data machine learning
Machine learning's goal is to uncover knowledge and make informed decisions. Many real-world

applications employ it, including recommendation engines, recognition systems, informatics and data mining, and autonomous control systems. In general, supervised learning, unsupervised learning, and reinforcement learning are the three subdomains of Machine Learning (ML).

SECTION IV:
Big Data Visualization Tools

Tableau

Tableau is a business big data visualization platform that lets you generate charts, graphs, maps, and a variety of other visualizations. It's been snipped off for charts and graphs. It is based on a desktop application that may be used to create visuals analytics. A server edition is available in addition to the desktop edition.The user can view reports online with this solution on a smartphone app A cloud-hosted service is used in this scenario. Moreover, there is an option that allows the buyer to install the on-premises solution. Interactive dashboards with a variety of visualizations displayed in various formats provided succinct ways to describe the covid-19 pandemic's progression**[18].**

Infogram

This utility includes a number of interactive charts and maps to assist users in visualizing data in a pleasing manner. For chart objects such as column, bar, pie, or word cloud, the tools are disabled.**[14]** It falls under the genre of infographic software because the user can even add a map to her infographic, resulting in an attractive report. Infogram offers team accounts for journalists and media enterprises, as well as branded designs for businesses and classroom accounts for educational purposes.

ChartBlocks

ChartBlocks is an online tool that creates visualizations from spreadsheets, databases, and live feeds without requiring any scripting. The JavaScript library D3.js is used to build the charts behind the scenes in HTML5. This application makes charts and widgets that are compatible with any screen size and device because it is web-based. Charts can also be embedded in any web page and shared on Twitter and Facebook.**[14][19]** This category also includes libraries or modules for building chart or graphical widgets, which are typically used inside web applications and use Javascript objects and functions, such as:

Plottly

Plottly is a program that allows you to create clean and sleek charts from a simple spreadsheet. Plottly is used by a number of well-known firms, including Google, the US Air Force, Goji, and New York University. It is mostly a web application, but it also has an API for a variety of languages, including JavaScript and Python.

D3.js

D3.js**[8] [9] [19] [24] [25]** is a library written in the JavaScript language for generating graphics or visualizations that manipulate and bring data to life using HTML, Scalable Vector Graphics (SVG), and cascading style sheets (CSS) techniques. It emphasizes current Web standards, which use all the capabilities of modern Internet browsers without using proprietary standards.

Ember Charts

It is built on the Ember.js framework and makes use of D3.js. Time series, bar, pie, and scatter charts are all available in Ember Charts. It's simple to extend, offers best practices and interaction, and is resilient when faced with faulty data.

Charts from Google

Google Charts is a Java library that uses HTML5 and SVG to provide 100% cross-browser compatibility, including support for older Internet Explorer versions through VML.All of the charts are interactive, and some of them may even be zoomed.[19] Google Charts is quite user-friendly, with a very attractive and extensive gallery where users can see the kind of visualizations and interactions they require.

SAS

SAS analytics is a Web-based tool that allows different users to access large amounts of updated data as well as information stored in an in-memory server. These servers work on parallel networks, which are simply shifting workloads from one computer to another, giving users safer and faster

access to data. In the **[6]** by Battineni G et al.,SAS tools are used by a company called Bio Axis Medical Services, which is a major healthcare provider in Greece providing higher quality medical care to its patients and clients.

BioVis Explorer

BioVis Explorer is a web-based visualization tool that provides an interactive and intuitive evaluation of published visualization methods, including faceted browsing and relationships with related methods.In the **[17]** Scientists are convinced that, based on the initial expert evaluation and their favorable experiences with this web-based approach, it will have a positive impact on the increasingly data-driven studies and subsequent conclusions in various fields of systems biology.

Qlik

Qlik provides powerful business intelligence, analytics, and enterprise reporting capabilities. It is free for individual use, has a strong community, and there are plenty of third-party resources available to help new users understand how to integrate it into their projects. In the **[7]** by Li Ket al. , they developed a dashboard for visualizing student learning activities using Qlik Sense Desktop.

Ontodia3D

Razdiakonov Daniil et al mentioned new tools Ontodia3D in **[10]** is a web-based extensible tool that builds on the most recent technology stack. Its key features: visual customization based on different graphical component types, and 3D exploration of small and large graphical structures.The 3D representation of graphics offers multiple benefits and opportunities compared to 2D. First, we have another spatial dimension to represent and group data. In 3D space, the depth of the space allows for more meaningful placement and grouping of objects, as well as shapes and textures of 3D objects.

SECTION V: Big Data Visualization Techniques

1.Display data in a 3D environment

Displaying data in a 3D environment helps users explore an area of interest that is more natural than traditional maps.

Examples of presenting data to 3D scenes are mentioned in the **[1]**. Based on multi-temporal satellite data, a general model for estimating production potential area is established in FOODIE platform. LANDSAT satellite image ESPA library is used as the main data source to provide surface reflectance products, main vegetation index (NDVI, EVI) and CFmask algorithm to identify clouds. Scenes from Rost Nice farm area in the past 8 years were selected to collect cloud-free data in the second half of vegetation period. According to the relationship between each pixel and the mean value of the whole field, the estimated output of the independent scene is calculated. Such visualizations can help farmers better understand their fields. Farmers can explore crop estimates in relation to topography, slope, orientation and topography.

Another example of display data in the 3D visualization platform is provided in the **[2]** of Olga et al. The 3D visualization platform shows the patient data flow for each individual procedure or operation (Fig.2, Fig.3, Fig.4). The three axes correspond to length of stay (blue number), wait time for investigation procedures or surgery (green number), and patient age (red number).

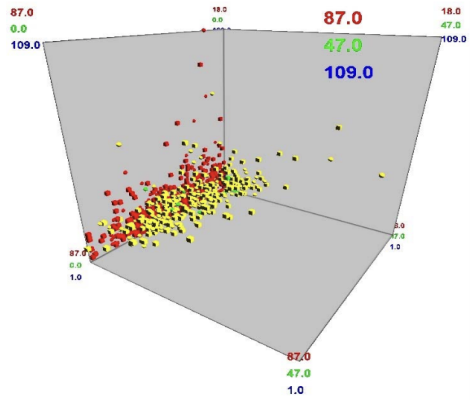


Fig.2. Echocardiography. Marker-cube – women, marker-ball – men. Red marker – 0, yellow – 1, green

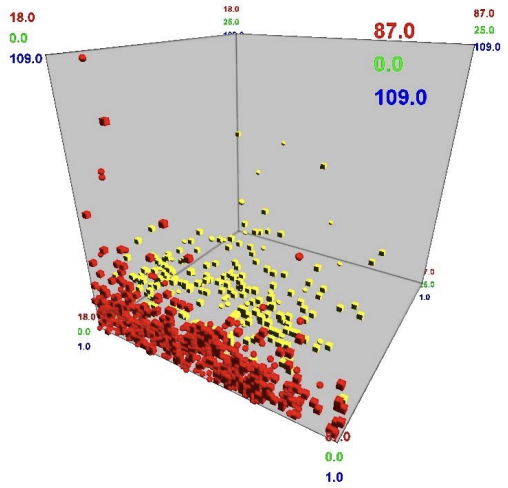


Fig.3. Thyroid Ultrasound. Marker-cube – women, marker-ball – men. Redmarker – 0, yellow – 1, green – 2 investigation procedures.

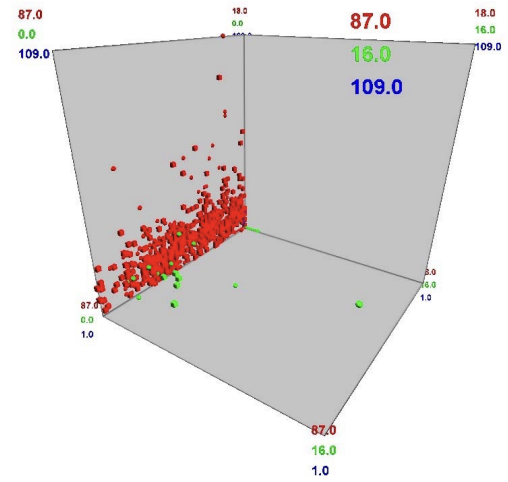


Fig.4. Coronary Angiography. Marker-cube – women, marker-ball – men. Red marker – 0, green – 1 investigation procedure.

3d visualization of patient data streams helps to plan the hospital's time workload. We can see that echocardiography, thyroid ultrasound are prevalent in patients with type 1 diabetes; Coronary angiography is mainly performed in elderly patients. This type of big data visualization reveals sociomedical differences in patients and helps plan inpatient and outpatient time management.

2.Use 'n' curve to analyze the data

The N-curve mentioned by M. Pi et al. **[3]** is a graph representing the accumulated number of outbound vehicles and arriving vehicles in a certain period of time in the traffic flow theory. Traffic engineering scholars use the N-curve to analyze the traffic flow in congested areas and the causes of traffic congestion, and the N-curve can analyze traffic time, location, and cause of congestion, as well as congestion damage including time delays and the number of delayed vehicles.

In Fig.5, the total number of vehicles under traffic congestion is *V2 'v1*, and the total delay time is *TD*. The time of traffic congestion is represented as *TTdelay*, and *TTff* can be obtained from the time from *X1* to *X2*.

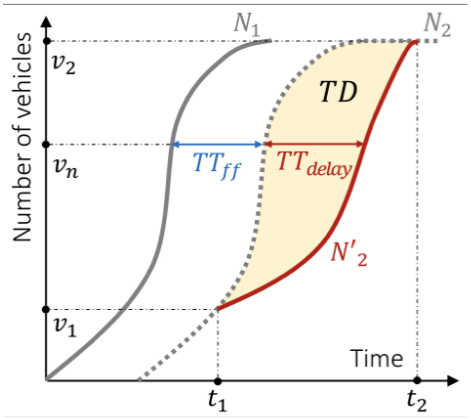


Fig.5. N-curve shows the cumulative number of vehicles that pass a certain location from X1 to X2 at time t1 to t2.

3. Visualizes space-time (Time series)

When the chart has regional information and needs to be highlighted, the map can be used to visualize the space, and the map serves as the main background to present all information points.

To view the changes in the index values over time, usually by adding a time axis, which is a common trend chart.

Another good example of space-time visualization was presented in the **[4]** of Carson K et al.

Contour maps use shading, coloring, or differences in symbol placement within predefined areas to represent the average of attributes or quantities within those areas. For example, Fig.6 shows a contour map of COVID-19 epidemiological data, where shaded differences indicate the total number of confirmed cases. The darker the color, the more severe the outbreak, the higher the number of confirmed cases.

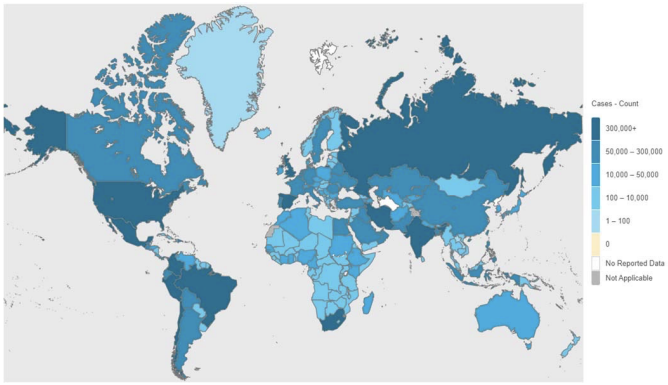


Fig.6. A snapshot of a bubble map showing the total number of confirmed cases among different countries in the world as of August 07, 2020.

In addition to spatial information, temporal information from COVID-19 epidemiological data is also important as it shows trends. Temporal information (such as daily or cumulative new cases, confirmed cases, and deaths) is usually represented by broken lines, bars (or stacked bars), and areas under a curve (or stacked areas under a curve). Fig.7 shows an example.

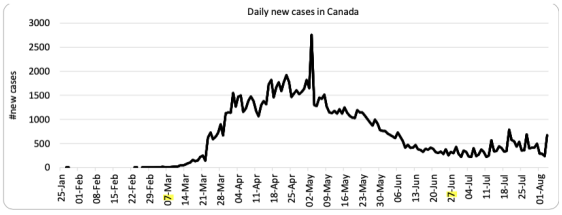


Fig.7. A line graph showing the daily new confirmed cases in Canada fromJanuary 25 (for the first confirmed case in Canada) to August 04, 2020.

Another good example of time-series data visualization is stated in this paper [32]. They proposed an end-to-end deep-learning framework for multi-horizon time series forecasting, with novel structures to better capture temporal patterns on future horizons. first propagate information of future input variables in both forward and backward directions with a bi-directional LSTM decoder, then at each future time step they used the decoder hidden state to attend to several different periods of the history and generate attention vector individually. use a fully connected layer to emit quantile predictions based on the temporal context feature and the entire framework can be trained end-to-end and deployed with standard deep-learning platform Tensorflow and TyTorch.

They used LSTM based model called Single-Attention(h=1 and Multimodal-Attention(h=3) They used two e-commercial data called CEFCom and JD50K. The multimodal-Attention(h=3) showed the minimum mean square error loss 33.78 and 38.89 for CEFCom and JD50k respectively.

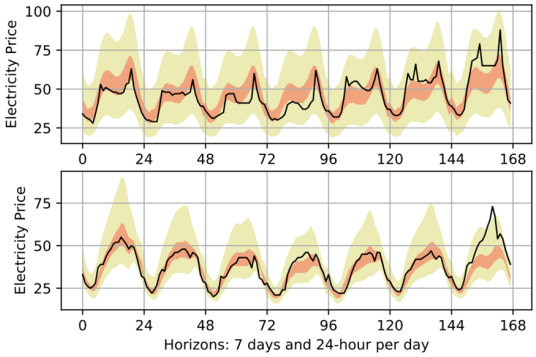


Fig.8. :time series electricity price in 7 days

The figure Fig.8 shows the multi-quantile forecasts provided by Multimodal-Attention(h=3) on two evaluation weeks of distinctive patterns. The upper series have two modalities within 24 hours while lower series have only one. By observing the quantile predictions of 0.25 and 0.75, we can see that the model is able to capture these distinct temporal patterns on future horizons by attending to the history.

There was an interesting approach that [39] showed time-series information in multiple texts by converting them to graph based visualization. First of all, they transform the time series news data to event information graph. Then they separated to two different graphs which are information change graph and abstract information graph. The keywords in both graphs are demonstrated the period of time by different colors in the node. An abstract information graph can be shown in c-x. from the topic “Japanese army”.

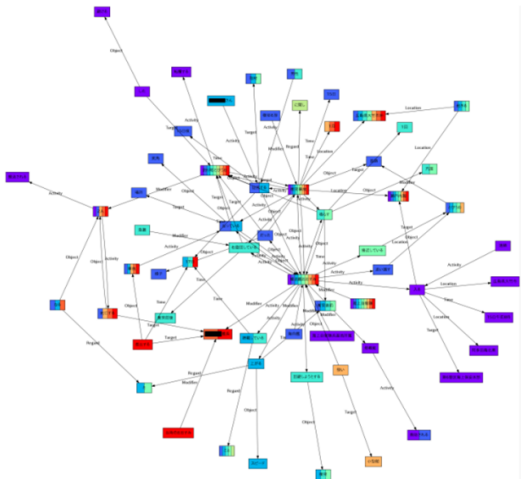


Fig.9. A subgraph showing the abstract information of texts.

Another example of data visualization on time series is proposed in [40]. In this paper, the stock market with a complex nonlinear system and time series prediction model based on XGBoost are presented. Mainly they use deep neural network model to train and predict the time series stock data. The 3 time series visualization with the original and prediction are shown in Fig.9. the stock 000400.SZ showed the shortest frechet distance (0.908) among all prediction.

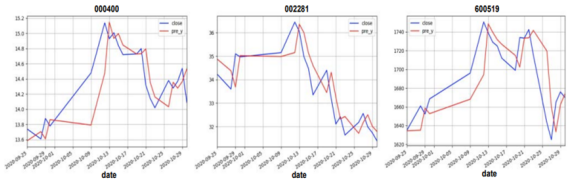


Fig.10. Visualization of predicted and original stock data.

Another example of time series data is in the financial sector as the two tasks of monitoring bank transactions and credit control usually involve time oriented and multivariate aspects. **R.A. Leite et al. in paper [24]** presents an approach to identify fraudulent events in the financial sector using a tool EVA (Event Detection with Visual Analytics). It is developed as a web application by using Angular and D3 technologies. In this approach a system is set up by creating a customer profile for each customer account based on the accounts transaction history, this is to classify the customer and profile construction. The visual analytics approach is based on a scoring system for financial fraud detection. The scoring system compares new transactions with the customer's profile of past transactions in order to compute a score that flags this transaction as either 'possible fraud' (suspicious) or not suspicious. The higher the score, the more suspicious the transaction. After score calculation, transaction scores that exceed threshold are manually scrutinized by an investigator. Once the investigator assesses the case and decides that a suspicious transaction is possibly fraudulent, they call the account owner to check the transaction's veracity. In lieu of the account owner's response the bank stops the transaction in case the account didn't authorize it.

Currently the prototype of EVA only detects possible fraud events related to "unauthorized transaction" and it is scalable for other types of frauds. For visual analytics to determine the elements that represent suspicious transactions bar charts, line charts and scatter plots (considering transaction amount and frequency as major attribute) have been developed Amount v/s Overall Score, Amount x Time, Top receiver x Amount, Top receiver x Frequency, Amount(s) x Frequency and Score Construction.

4. Visualizes graph data

The various ways of empirical user evaluation of graph visualization are presented in [33]. For the evaluation of graph interpretation, first, the visual design is evaluated by several aspects which are visual properties of vertices and edges, metaphorical graph representations and visual encoding of a graph sequence. Second evaluation point of the interpretation is the components in layout which are layout algorithms, aesthetics criteria and clutter reduction techniques. Third, special properties such as graph classes and topology and edge properties are evaluated. Finally, enriched graph visualizations and graph interactions are evaluated. Another category for the evaluation is graph memorability. The memorability in graph dynamics, memorability of graph structure are mainly evaluated in this field. Finally graph expression and creation are evaluated in two different aspects which are the graph visualization based on user intervention and working with graph drawing tools.

The paper [35] shows the visualization technique for analyzing phrases and Sentiments using directed graph data structure called EmojiText which is created in the D3 framework. They put text or emotion data models to preprocessing steps.

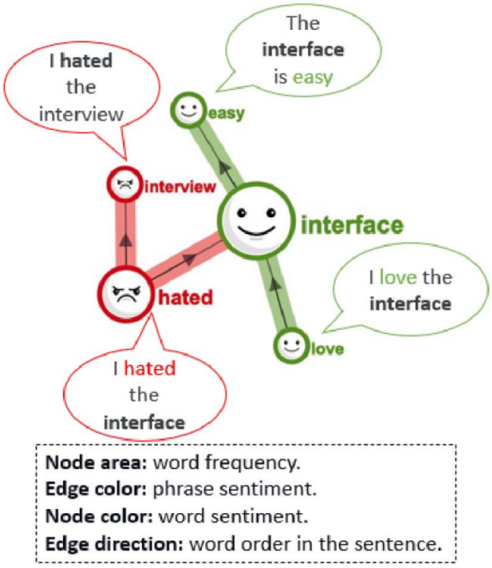


Fig.11. graph visualization of text and emoji.

In pre-processing, word identification, frequency attributes, emotion polarity identification and time identification are proceeded. Then set filter and color to visualize them in emojiText using directed graphs which can be seen from Fig.11..

There is another approach in the paper [36]. They implemented a graph data visualization of news with virtual reality using HTC Vive which is a head mounted display. They first analyze the different papers and design guidelines and interactive tools to apply in the project. Then they developed the application in a VR environment and showed the graphic data as shown in Fig.12.

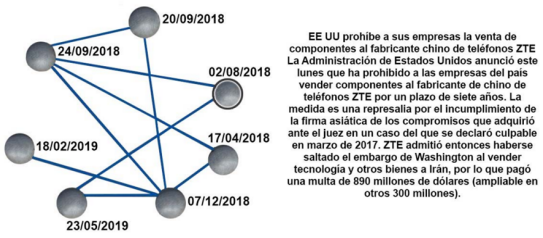


Fig.12.: Overview of the application

Finally, they evaluated the results of the understanding of graphs in an immersive environment. Their qualitative result showed that 90% of the users made correct answers from visualized graph data in the VR environment.

In this paper [37], the authors present OPT+graph which is a web-based program visualization tool to support learning programming graph data structure. They detected techniques of graph data structure for visualization based on representation of adjacency matrix, array of edges, and array of adjacency list. In their experiments, about 52.38% respondents said the platform is efficient and effective. 66.67% of respondents successfully solved problems in their experiment. The platform is shown in Fig.13..

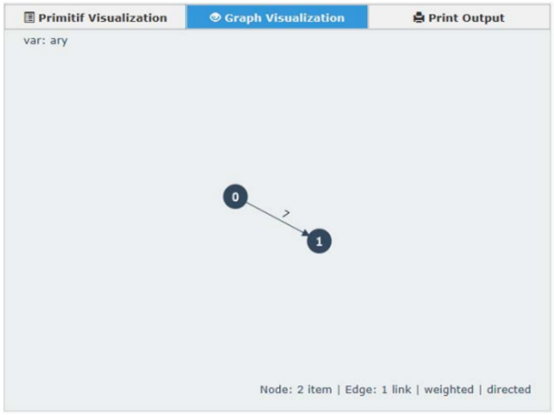


Fig.13. visualization of graph data structure.

There is another great approach in [38]. They proposed a graph based information block detection in an infographic. First of all, they extract features using gestalt feature extractor to obtain the spatial and chromatic features of infographic elements. Then, they group the information block and finally identify narrative sequence. This process can be seen in the Fig.14.

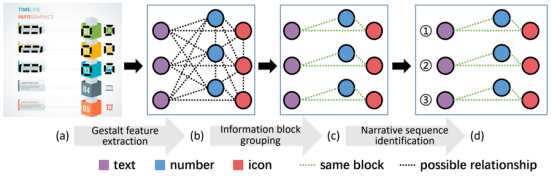


Fig.14. the whole process of block detection.

They used 4 different datasets (Visually29k, InfoVIF, Timeline, InfoBlock) to evaluate their model. Their models were based on R-CNN method. IBD-Graph full model showed the best accuracy(50%) among all models.

Another application by **C. Chen et al. in paper [22]** presents "InfDetect" , a large graph based fraud detection system for E-commerce insurance. InfDetect provides an interface for commonly used graphs, standard data processing procedure and a uniform graph learning platform. This paper focuses on security deposit insurance and return-freight insurance and to detect fraud; transaction fraud, device sharing fraud and friendship graph are constructed to reveal pattern for fraud classification and build a buyer-seller graph to identify fraudulent orders. Dataset is sampled from claim history and one transaction is generated for each day. For an insurance product each claim involves two parties, so the claim representation is obtained by concatenating the involved user embeddings.

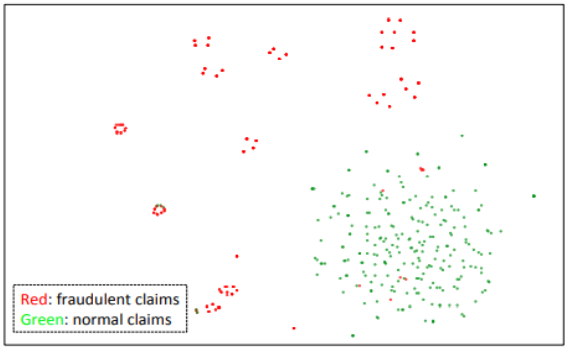


Figure 15 above shows Claim Level embedding.

Figure 15.above shows claims level embedding. Red dots represents fraudulent claims while green dots refers to normal claims. The fraudulent claims form different small clusters demonstrating a gang behavior in fraudulent claims.

Similarly, figure 16. below shows User Level Embedding where red colour marks the fraudulent user who initiated the fraud and green color to mark normal in the embedding space.

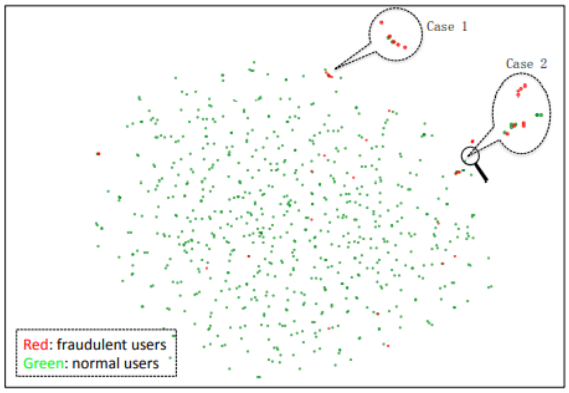
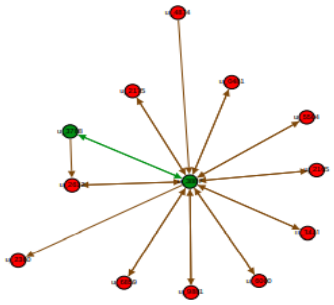


Figure 16: User Level Embedding



(a) Case 1

Fig 17: Visualization of fraudulent claims related user over the transaction graphs

Above figure 17 shows Visualization of fraudulent claims related user over the transaction graphs and helps examine two clusters of fraudulent users and their behavior on the transaction graph. In the case 1, the fraudulent users (in red) exchange funds through a normal user (in green). This is a typical pattern where fraudulent users do not directly contact, instead, they find a "normal" user with a clean record to do so to cover their fraudulent behaviors/monetary traces. Thereby, user embeddings learned using transaction graph are helpful for discovering fraudulent users and claims.

5. Interaction

After the data is graphically completed, it can be transformed into dynamic and controllable charts based on the actual situation, so that users can better perceive the change process of data and improve their experience during manipulation.

Dynamic is usually achieved in two ways: interaction and animation.
Interactions include mouse movements, clicks, linkage responses to multiple charts, and so on.
In **[5]** presents an interactive Internet of Things data visualization approach for business intelligence to analyze and minimize adoption and self-service data resource issues.

Another example of interactive data visualization is stated in [34]. They created a visual analytics framework for human-in-the-loop, example based graph pattern search via graph representation learning. They used a novel graph neural network for node-alignment called NeuroAlign and a vosia; alysis system called GraphQ which provides a visual query interface with a query editor and a multi-scale visualization of the result and a user feedback mechanism for refining the results with additional constraints.

6. Visualization and Trend Forecasting for multi-sourced data using data mining and deep learning.

In order to analyze heterogeneous and multi sourced data several data mining and deep learning techniques are adopted to study potential data patterns and trend prediction. Analysis of such big data enables systems to keep track of occurred events, identify similarities from incidents, ensure control and preventative strategies for decision making on future events. The following papers present big data visualization and trend prediction techniques for multi sourced data which are extracted through integrated systems.

In paper [21], M. Feng et al., has used big data analytics to study criminal data of San Francisco, Chicago and Philadelphia locations to identify crime patterns and how they are related with time. Considering the geographic nature of the crime incidents, an interactive map based on Google map is used for data visualization, where crime incidents are clustered according to their latitude/longitude information. For data analysis and visualization, 13 featured attributes are identified from crime incident dataset records, and implementing any algorithms on the datasets preprocessing steps are performed like time is discretized into columns, for missing coordinate attributes, impute random values sampled from non-missing value, compute their mean and then replace the missing ones, timestamp incident occurrence into year, month, date, hour and minute.

(a) (b)

Fig.18. Time series plot for (a) San-Francisco (b) Chicago

Fig.18 summarized crime incidents in each year for the cities. In San-Francisco, the number of crime incidents seems to soar since 2012 and reaches its peak in 2013, whilst the numbers for the other city tend to decrease yet following certain patterns.

(a) (b)

Fig.19. Wordcloud of crime description in (a) San Francisco (b) Chicago

Fig.19. shows wordcloud plots to illustrate the significance of different categories of crime. San-Francisco has theft and property related crimes and Chicago has domestic crimes.

Similarly, the hourly trend of crime in each city, bubble plot of crime analysis, box plot of monthly statistics of crime in each city is analysed and it was found that the monthly crime rate is likely linked to local climate. A time series is a sequence of numeric data points successively indexed or listed/graphed in time order. Time series exhibit a variety of patterns and to show crime evolved over time it is helped to decompose the crime time series.

For crime trend forecasting using The Prophet model, Neural Network model and LSTM model in terms of RMSE and spearman correlation is studied and the model is trained with 3 years training data and the visualization trend findings showed that the Prophet and LSTM model performed better that traditional

neural network model as the neural network has lower RSME and low correlation between predicted values and real ones. The final findings through this prediction analysis will help police departments and law enforcement organizations comprehend crime issues and provide insight to anticipate likelihood of events, resource deployment and optimize decision making.

In paper [25] A. Sanchez and W. Rivera presents visualization techniques for data related to weather conditions, twitter activity and energy consumption for a Smart Grid and data analysis is performed using Apache Spark cluster framework. The availability of sensors attached to grid equipment is used to help diagnose grid problems and prevent outages, allowing crews to pinpoint the exact location of faulty hardware. Also, in service provider domain data from smart meters, sensors and phase measurement units can be used to estimate energy process and consumer behavior. It helps visualise interdependency between the energy demand and prices enhancing the decision making process. The data size goes upto 8TB per year which helps design and implement Demand Side Energy Management strategies and also to set Demand response mechanisms.

Data Sources	Data Analytics
Energy Resources & Sensors Electricity Market Weather Consumer Load & Behavior Social Interaction	Demand Response Control Operational Planning Dynamic Pricing Models
Data Cleaning	Online Learning
Trustworthiness & Privacy	

Fig 20: Data Analytics Challenges in Smart Grid

Fig 20. above summarizes the major data sources, the data analytics task that can be achieved and major challenges experienced while working on smart grid data.

Data mining algorithms are applied on these datasets to extract patterns using Python Flask web development framework to visualize energy data, real time data from different sources in Puerto Rico. The application used is Oasis dashboard-Fig 21 where the jQuery is used for content display manipulation and to handle Ajax requests to the server. The D3 library is utilized for visualizing content in an interactive manner.

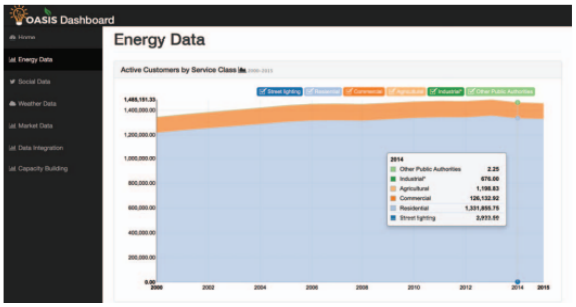


Fig: 21. Energy Generation in Puerto Rico.

.

The weather data is collected using the Forecastio-py library from Dark Sky API and the dashboards related to precipitation probability, humidity percentage, wind speed and direction, cloud coverage percentage and temperature. Using time series analysis and classification techniques the energy demand and service requirement prediction are done and using sentimental analysis algorithm the classification of energy related algorithm is done.

Another interesting application of Big data is in Wind power sector where big data can be used for wind power prediction and in designing a decision support system for wind power production. **Paper [26] by A.Zhu. et al**. presents big data processing using Convolutional neural network; with historical data of wind power from a wind farm as input, this paper sets the parameters, trains the CNN model in MATLAB and then predicts wind power. CNN which is a feed forward neural network, extracts features from a two dimensional image and uses the back propagation algorithm to solve the unknown parameters in the network. Then the classification of regression is done to obtain an output. In the pre-processing stage CNN processes the one-dimensional array of input data into a two-dimensional matrix. In this study Elia wind farm transmission system data is used to train the CNN model and for regression analysis. Further, a property file which is CNN input and a response file for comparison of predicted data is generated.

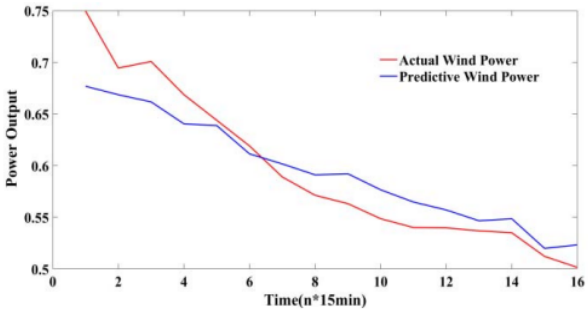


Fig.22. Time series data visualization.

As shown in fig.22, the wind farm records wind power every 15 min. Each line of the property file and response file make a sample, thereby, with 400 wind power historical data as inputs, 16 regression prediction outputs are obtained i.e output power in the proceeding 4 hours. The red line represents the actual wind power and blue line represents the Predicted wind power. The results of prediction are verified using MSE i.e. mean squared error and it shows that the predicted value and the actual value coincides proving the feasibility of CNN applied to regression prediction. Similarly in **paper [27] by R. Donida Labati et al.** presents a decision support system using big data techniques that can predict electric power production, estimate a variability index for prediction and analyze wind farm production characteristics. Since the prediction is solely based on weather forecasts, neural network and on technique for calibrating and thresholding the weather forecast based on the distinctive characteristics of the WF orography, the proposed system is suitable for Wind farm that cannot collect or manage real time data acquired by the sensors.

In this study Numerical Weather Prediction (NWP) data is used to learn the orography information and Wind farm characteristics. The NWP data of two wind farms located at different orography for a period of two years is collected and is used for pre-processing where data harmonization and feature extraction is performed on data. The data is computed into different sets of numerical features and from NWP data and at configuration stage it is calibrated to improve wind forecasts of NWP data. The threshold system detects cases in which the wind energy is insufficient to activate the wind farm turbines. The neural prediction variability index of power prediction is estimated by computing statistics from the results of the neural network that compose the neural prediction module. To verify the accuracy of weather forecast the data obtained from anemometer measurement and NWP are compared.

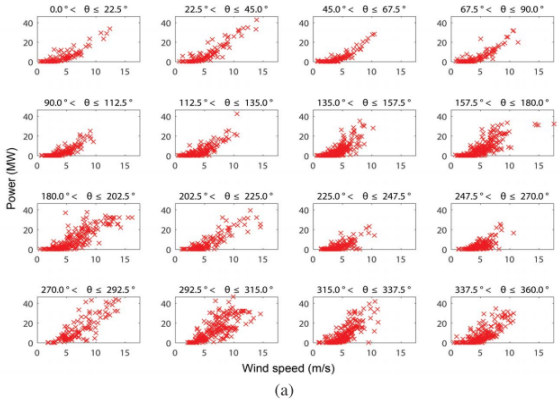


Fig.23. Analysis performed for two wind farms.

Fig.23 above shows an example of this analysis performed for two WFs. (Wind farms) Fig. (a) shows the ratios of power produced to wind speed for discrete sets of wind angles at the two WFs.

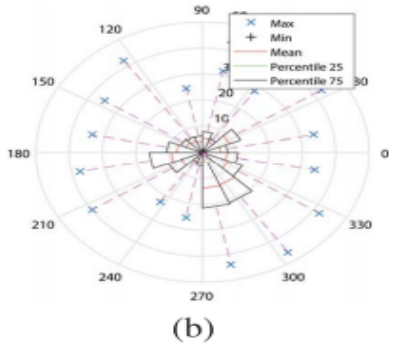


Fig.24. Polar boxplot of two wind farms.

Fig.24 shows “Polar boxplots” of the produced power for each set of wind angles at the two WFs,

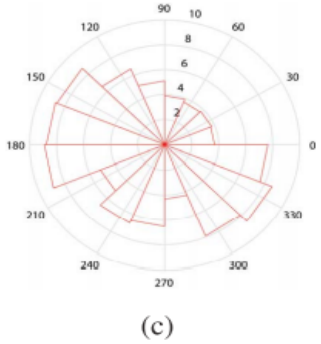


Fig.25. The frequency of wind angles

and Fig.25 represents the frequency at which the wind angle pertains to each set of wind angles. This figure shows that the two analyzed WFs exhibit strong differences in terms of orography and production characteristics.

In paper [28] M. Li and Q. Zhou, explains a case study using flight data recordings to discover the factors affecting the Airplane Fuel efficiency using big data visualization technique during different phases of flight (like Taxi, Takeoff, Climb, Cruise, Approach, Rollout). For study a dataset of 33 flight tail recordings is provided. Each flight tail has a different aircraft and has 25 flying instances. R package is employed to facilitate data preparation, data analysis and visualization and to efficiently reduce the programming and execution time.

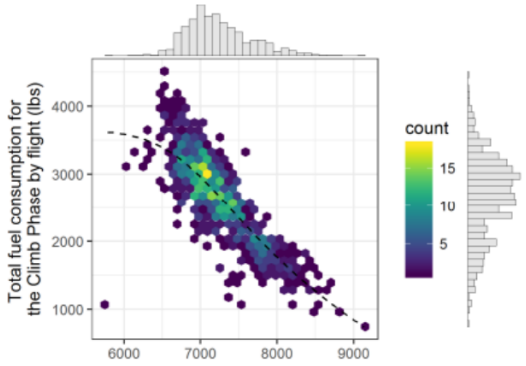


Fig.26. The mean fuel flow during the Climb Phase by flight instances during the Climb Phase.

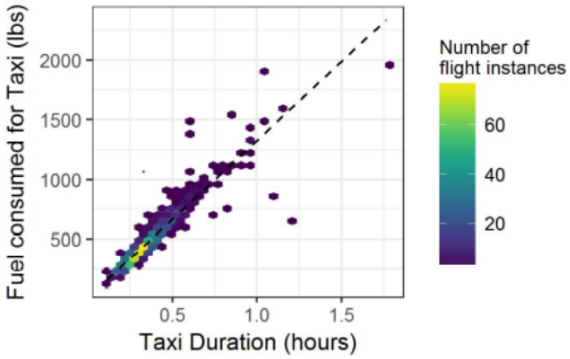


Figure.27. The time and fuel consumed in the Taxi flight phase

As per fig.26 and fig.27 above the study explains that in taxi phase, the taxiing time and fuel consumption exhibit linear correlation, in climb phase and approach phase multidimensional visualization revealed different trajectory patterns overtime,considering altitude, distance and fuel consumption which could be useful reference for path planning. In the Cruise phase, the cruising altitudes were set to different altitudes discreetly. Visualizations from the large and dense data set show that the fuel flow rate has a negative relationship with the altitude. Therefore, the Taxi, Climb, Cruise and Approach phases have higher fuel efficiency potentials.

Another example for multidimensional data visualization is presented by Fan Du et al. in paper [30]. This paper demonstrates a workflow for an event sequence recommendation system implemented in a tool named Event Action based on personal history record data. It is an interactive prescriptive analytics system and user interface to assist user in making action plan and to raise users confidence in the action plan. The typical workflow starts from selecting a seed record and the first step is to find a group of similar records. After submitting the similar records, a recommendation model is computed and users can review a recommended action plan. Then, users can further refine the plan by directly editing the plan using the activities of similar records as a reference or refining the similar records to generate an updated recommendation. The system backend consist of a data pipeline for finding a similar record and an automatic algorithm for generating recommendation.

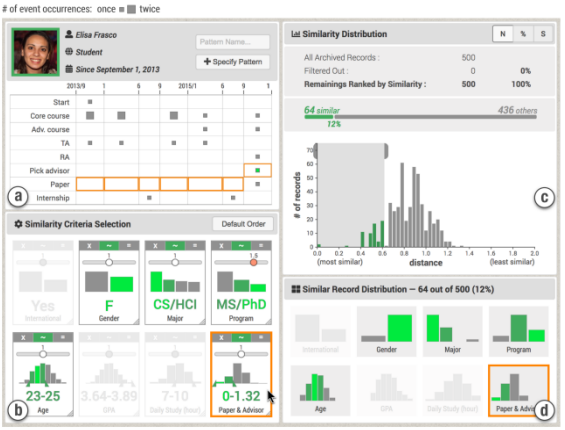


Figure 28 above shows the visualization dashboard of Event Action recommendations system.

In paper[12], they suggest a method for improving the evidence-gathering process in decision-making in the context of Smart Cities by using visualization techniques for Big Data. Proposed artificial intelligence combined the methodology with the ability to include real-time data in the context of Smart Cities to help users define their goals and generate the optimal type of visualization.
Using existing information and Artificial Intelligence to give evidence for city officials to make decisions that promote sustainability through city growth and resource management. The Artificial Intelligence approach is a new topic in our overall and broader methodology, which we introduce in this paper for the first time. Furthermore ,(ii) improving work by incorporating the necessary steps and solutions to process data in real-time, and (iii) make context information more accessible to users, allowing them to better understand the output of (iv) an Artificial Intelligence algorithm we trained for this case study.
These visualizations depict real-time data from IoT sensors, allowing users to obtain the information they need to make strategic and tactical decisions.

In paper[16], we'll look at how visualization techniques can be used with open government data to assist tackle this challenge. They looked at previously published articles on Open Government Data Visualization to get a sense of how visualization techniques are being used on Open Government Data and what the most prevalent issues are when dealing with it. We discovered, among other things, that datasets connected to transportation are the most commonly utilized, and that Map is the most commonly used visualization approach.

7. Visualization using Heat map data:
Visualization analysis has proved to be beneficial in various domains across industry and in a real world shopping environment, the shopping activity of an individual also generates huge big data that describes consumer behavior and spending patterns. **Paper [23] Peter ChunYu Yau et al.** presents a proof of concept methodology and shows that “holding time” and “the frequency of spots” have a certain relationship to the purchase decision made by the consumer. A real time analysis is conducted using a smart space laboratory and by collecting biometric data like body temperature, real time location in store, user activity in both physical and digital interaction.

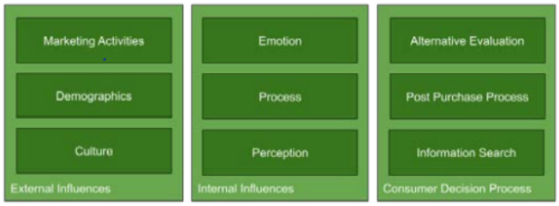


Fig.29. Simplified Consumer Behavior Model with Selected Components

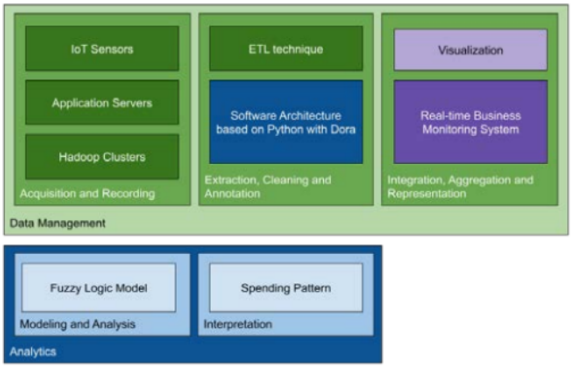


Fig.30. Processes for extracting insights from big data in this research with detail components.

Fig. 29 and Fig 30: Data Analysis is done using Dora and to power the big data storage and computation ability, Hadoop is selected to be the database in this study. For heat map visualization heatmap.js package is used to track customer position relationship with purchase decision. To provide higher accuracy on data transfer and enhance data security, infrared thermal detectors work with motion detectors to record room and subject (users) body temperature at every 2 sec, static and motion time are recorded accordingly and various IOT devices work together with mobile devices.

Fig.31. Heap map visualization of Users real time location

Fig.31 shows heat map visualization of Users real time location. In real-time per their current location user dataset is captured every 5 min interval from t:0 to t:35 and the result showed that the longer the hold (static) time, user stayed on a single spot, the higher the chances s/he will make certain purchase activity before the end of shopping journey.

8. Data visualization comparison for effectiveness. DeepEye is presented in the paper [41] as a novel system for automatic data visualization that is trained with a binary classifier to decide the visualization quality, and using supervised learning method in machine learning technique to decide better visualization among two of them. Also, the deepEye uses the expert's knowledge to specify partial orders and rules to make decisions. The Fig.28 shows the overview of DeepEye system.

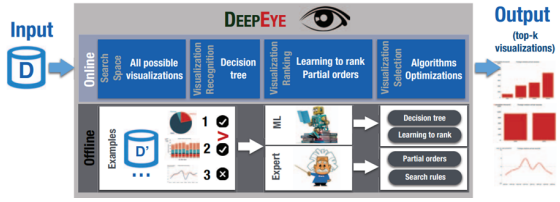


Fig.32. whole process of DeepEye

They proceeded to experiment with 10 different testing datasets and 9 real use case with data and visualizations to compare. The average effectiveness of the system is 93.2% to 99.5% among all datasets with bar, line, pie and scatter charts.

SECTION VI: Big Data Storage Technologies:

The volume and complexity of information generated every day throughout the world has increased tremendously in the last decade, to the point where many software applications are simply unable to handle volumes of data. **In paper [29], D. Deibe et.al.** particularly explains a case study for adoption of big data storage solutions in a web-based LiDAR visualization environment.. LiDAR being one of the most growing technologies has received an extra boost due to the rise and popularization of the unmanned aerial vehicles. This technology aims to scan large terrain extensions through laser pulses in order to obtain detailed 3D models in the form of point clouds which can be visualized and analyzed later.

Due to the massive LiDAR datasets that are being gathered nowadays, and that will keep growing in forthcoming years, disk space required to store some of the most dense and detailed datasets can easily overpass terabytes in size being not able to be stored in a single hard disk device. Though traditional client server structure ref Fig. 33 have been relying on common server software solutions like Apache HTTP server, it stills lacks many of the advantages presented in most big data solutions, such as low latency and high throughput, horizontal scalability, high availability though data replication, data distribution across several machines or integration with big data computing frameworks:

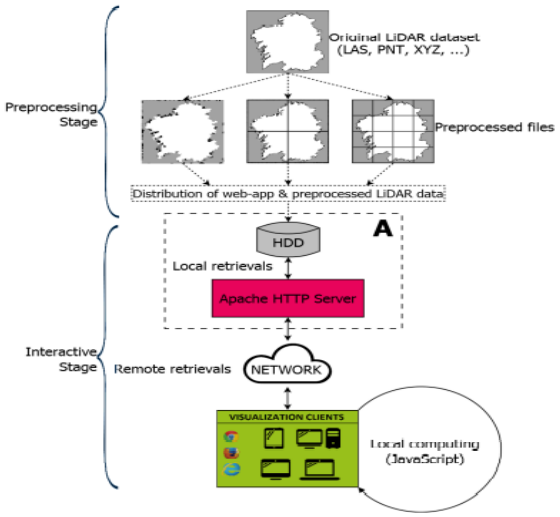


Fig.33. Traditional deployment of web application forLiDAR

Considering the massive volumes of LiDAR data that could have to be stored, a stand-alone server might not be enough for storing all LiDAR datasets and, if so, the throughput of the system could not be enough for handling very high levels of network traffic, which could be a problem for offering real-time interaction for all connected clients. Additionally, Apache stands as a single point of failure, meaning that any malfunction on the server could cause a temporary shutdown of the service or the permanent loss of data. In order to migrate from a non-big-data oriented approach to a big-data oriented one, the box A in fig above needs to be replaced by components like Gateways or proxies which work as

intermediaries to grant access to storage facilities, Data servers which are incharge of all read/write operations of data and Metadata servers and Query routers which are in charge of receiving and serving data queries. **In paper [29]** a deployment analysis is conducted on four big data storage technologies to determine which is best replacement for Apache HTTP server

HDFS is a distributed file system used by big data computing technologies such as Hadoop, Spark or Flink. It provides fault tolerance and data replication. Commonly, it operates over files of more than 1 gigabyte or 1 terabyte in size. These files are automatically distributed among the nodes of the cluster. Before being distributed, data is divided into chunks of a fixed size. Although the files stored in HDFS are commonly accessed by Hadoop, Spark or Flink, for computational purposes it implements a REST API (WebHDFS) allowing any kind of application to directly retrieve data from the system. The emphasis of HDFS is on high throughput data access rather than low latency data access.

MongoDB is a document store where data is handled in the form of JSON-like documents. Aside from implementing the same common features like data distribution and replication, fault tolerance and high scalability. MongoDB stands out for its automatic data balancing mechanism. The balancing mechanism ensures an even distribution of the data across all nodes of a cluster regardless of its initial distribution. The main components of a MongoDB cluster are the shards, the configuration servers and the query routers.

Cassandra is a wide column store, a class of database where data is stored in records with the capability of holding very large numbers of dynamic columns. It provides features such as fault tolerance, data distribution and data replication. It is highly scalable, being used by some of the companies, for Cassandra every node in the cluster is identical, being able to store data and resolve client/application queries. It has a fast and easy deployment process.

Redis is an in-memory, key-value structure store with all the traditional advantages of big data technologies, data distribution and replication, fault tolerance and high scalability. Redis achieves especially good performance during data readings as it has an in-memory design, which allows it to serve all data queries directly from main memory, unlike other solutions, where data are served from disk and only most recently used data are served from main memory. In a Redis cluster, two types of nodes can be found, masters and slaves. All data stored in Redis are divided among the master nodes presented in the cluster. All master nodes have zero or more slaves associated to them and each slave is in charge of storing replicas of the data from its corresponding master node. In Redis, it is mandatory to deploy at least three master nodes, meanwhile the number of slaves depends on the replication factor. **In the paper [29]** all four technologies have been analysed and it has been concluded that, in the case study, Cassandra and MongoDB are the most suitable options as Cassandra offers lower latency, higher throughput and more storage capacity. On the other hand, MongoDB offers great versatility in data modelling and unique features (such as the geo-spatial queries) that could be very useful in the near future for using in a LiDAR data context. MongoDB also offers quite good levels of latency, throughput and storage capacity, therefore, in the current state of our application, both storage technologies could be considered as equally suitable, although this situation could change in the future depending on the evolution in the development and requirements of the application.

SECTION VII
Big Data Visualization Best Practices

Effective data representation strategies are used in data visualizations to integrate, unify, and standardize data from many sources. Visual representations that are well-designed allow visual access to large volumes of data in easily digestible portions. This encourages increasing usage and learning, as well as improved work performance.[31]

In paper[15], it depicts that small sample size investigations, better data visualization approaches are required to ensure transparency. The tactics and tools presented in this review are intended to increase transparency by enhancing the quality of figures, as well as to aid scientists, academic journals, and

funding agencies in making long-term improvements to data visualization in scientific publications.

Many journals and publishers have recently implemented policies encouraging or mandating writers to use more informative figures that illustrate data points or distribution.

This primer provides a detailed overview of strategies for dealing with this problem by (1) outlining techniques for selecting the appropriate type of figure based on the study design, sample size, and type of outcome variable; (2) examining techniques for creating effective dot plots, box plots, and violin plots; and (3) demonstrating how to avoid sending mixed messages by aligning the figure structure with the study design and analysis goals. Other typical visualization difficulties identified in the systematic review are also addressed.

Figures that demonstrate the data distribution should be used instead of bar graphs.

Bar graphs were employed 47.7% of the time to depict continuous data in articles with data figures, which is typical for small datasets. Fig.34.1 shows that dot plots should be used instead of bar graphs in most cases.

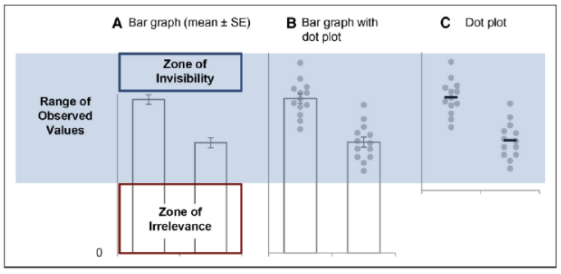


Fig.34.1.

Dots could be added to box plots. The smallest and largest groupings in a box plot had median sample sizes of 20 (interquartile range, 11.5–56) and 66 (interquartile range, 17.5–302.5), respectively. There are no clear rules for when a dataset is large enough to employ box plots; yet, these data suggest that box plots are occasionally used with tiny samples.

Now , the below (Fig.34.2) elaborated chart explains the best practices for most commonly used figure types.

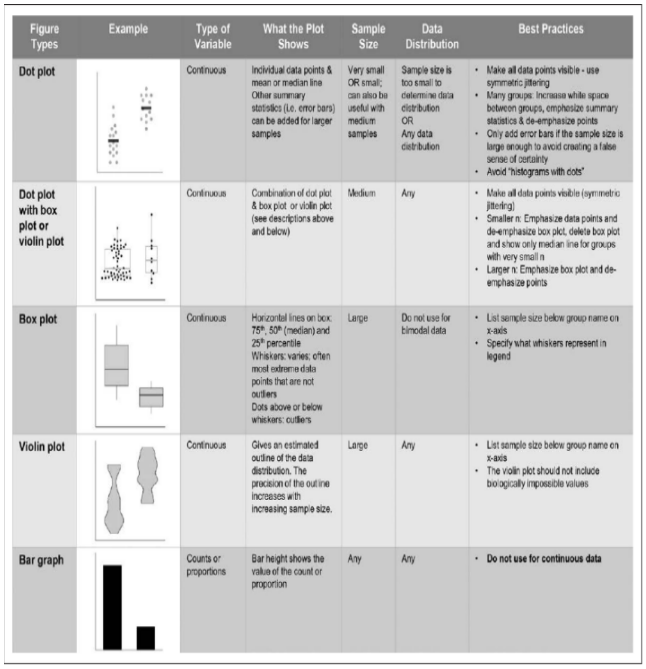


Figure Types	Example	Type of Variable	What the Plot Shows	Sample Size	Data Distribution	Best Practices
Dot plot		Continuous	Individual data points & mean or median line Other summary statistics (i.e. error bars) can be added for larger samples	Very small OR small; can also be useful with medium samples	Sample size is too small to determine data distribution OR Any data distribution	• Make all data points visible - use symmetric jittering • Many groups: Increase white space between groups, emphasize summary statistics & de-emphasize points • Only add error bars if the sample size is large enough to avoid creating a false sense of certainty • Avoid "histograms with dots"
Dot plot with box plot or violin plot		Continuous	Combination of dot plot & box plot or violin plot (see descriptions above and below)	Medium	Any	• Make all data points visible (symmetric jittering) • Smaller n: Emphasize data points and de-emphasize box plot, delete box plot and show only median line for groups with very small n • Larger n: Emphasize box plot and de-emphasize points
Box plot		Continuous	Horizontal lines on box: 75th, 50th (median) and 25th percentile Whiskers: varies; often most extreme data points that are not outliers Dots above or below whiskers: outliers	Large	Do not use for bimodal data	• List sample size below group name on x-axis • Specify what whiskers represent in legend
Violin plot		Continuous	Gives an estimated outline of the data distribution. The precision of the outline increases with increasing sample size.	Large	Any	• List sample size below group name on x-axis • The violin plot should not include biologically impossible values
Bar graph		Counts or proportions	Bar height shows the value of the count or proportion	Any	Any	• **Do not use for continuous data**

Fig.34.2.

To make overlapping spots apparent in scatter plots and flow cytometry figures, use semi-transparency or show gradients. Generally, scatterplots have overlapping points.Techniques such as semi transparency, shaded color gradients, or gradient lines should be used to make overlapping points visible.Small datasets with few overlapping points benefit from semi transparency. For huge datasets with many overlapping points, such as flow cytometry plots, gradients should be displayed.

Use color maps that are colorblind-friendly. Up to 8% of males and 0.5 percent of women of northern European heritage suffer from the most frequent kind of colorblindness.

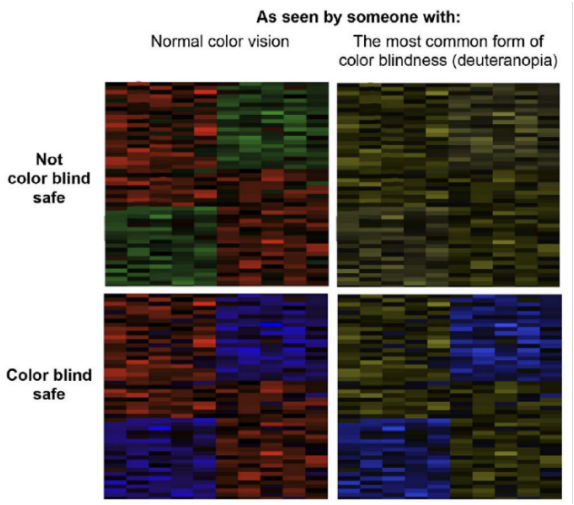


Fig.34.3.

Only 15.4 percent of these publications (4/26) used color maps with key features visible to someone with deuteranopia, despite the fact that 12.6 percent of papers (26 of 206) used a color map on a graph or clinical image. Color palettes that were not colorblind friendly were utilized in the majority of studies featuring heat maps for example check Fig.34.3. Colorblind-safe alternatives should be used instead of non-colorblind-safe color mappings (e.g., rainbow, red/green). As many rainbow color maps appeared to have been generated by flow cytometry or ultrasound software, this may need engaging with device manufacturers.[15]

To improve transparency in small sample size investigations, better data visualization approaches are required. The tactics and tools presented in this review are aimed at increasing transparency by enhancing the quality of figures, as well as supporting scientists, academic journals, and funding agencies in making long-term improvements to data visualization in scientific publications.

SECTION VIII : Conclusion

Various processing, analytical tools, and dynamic visualization are all provided by today's Big Data platforms. Such platforms enable knowledge and value to be extracted from a complex dynamic environment. They also aid decision-making by providing recommendations and detecting anomalies, odd behavior, or new trends automatically.In this survey, we looked at the characteristics of Big Data and discussed the issues that Big Data computing systems provide.

Furthermore, we have discussed the importance of Big Data mining in a variety of contexts. In addition, we've concentrated on the components and technologies that make up each tier of Big Data platforms. The capabilities, advantages, and limitations of various technologies and distributions have also been compared. We've also divided Big Data systems into categories based on the capabilities and services they deliver to end customers. The Big data storage solutions and representation strategies adopted as best practices are being discussed. As a result, it presents a thorough examination of the methods, and practices currently in use in Big Data computing. Despite significant advancements in the Big Data industry, we can see from our comparison of various systems that there are numerous flaws. The majority of the time, they are linked to the architectures and methodologies that have been employed.

In order to build next-generation Big Data infrastructures, more effort is needed in various areas, including data organization, domain-specific tools, and platform tools. As a result, technological challenges in various Big Data fields can be researched further and serve as a valuable study topic.

SECTION IX: References

[1] *K. Jedlička and K. Charvát, "Visualisation of Big Data in Agriculture and Rural Development," 2018 IST-Africa Week Conference (IST-Africa), 2018, pp. Page 1 of 8-Page 8 of 8.*

[2] *O. Kolesnichenko et al., "Big Data Analytics of Inpatients Flow with Diabetes Mellitus type 1 : Revealing new awareness with Advanced Visualization of Medical Information System Data," 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2019, pp.*

[3] *M. Pi, H. Yeon, H. Son and Y. Jang, "Visual Cause Analytics for Traffic Congestion," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 3, pp. 2186-2201, 1 March 2021, doi: 10.1109 /TVCG.2019.2940580.*

191-196, doi: 10.1109 /CONFLUENCE.2019.8776910.

[4] *C. K. Leung, Y. Chen, C. S. H. Hoi, S. Shang, Y. Wen and A. Cuzzocrea, "Big Data Visualization and Visual Analytics of COVID-19 Data," 2020 24th International Conference Information Visualisation (IV), 2020, pp. 415- 420, doi: 10.1109/IV51561.2020.00073.*

[5]*Zhang, L., Vinodhini, B. & Maragatham, T. Interactive IoT Data Visualization for Decision Making in Business Intelligence. Arab J Sci Eng (2021). https://doi.org/10.1007/s13369-021-05889-w*

[6] *Battineni G., Mittal M., Jain S. (2021) Data Visualization in the Transformation of Healthcare Industries. In: Roy S., Goyal L.M.,*

Mittal M. (eds) Advanced Prognostic Predictive Modelling in Healthcare Data Analytics. Lecture Notes on Data Engineering and Communications Technologies, vol 64. Springer, Singapore. https://doi.org/10.1007/978-981-16-0538-3_1

***[7]** Li K. (2019) Visualization of Learning Activities in Classroom Blended with e-Learning System. In: Uskov V., Howlett R., Jain L. (eds) Smart Education and e-Learning 2019. Smart Innovation, Systems and Technologies, vol 144. Springer, Singapore. https://doi.org/10.1007/978-981-13-8260-4_13*

***[8]** de Camargo L.F., Moraes A., Dias D.R.C., Brega J.R.F. (2020) Information Visualization Applied to Computer Network Security. In: Gervasi O. et al. (eds) Computational Science and Its Applications – ICCSA 2020. ICCSA 2020. Lecture Notes in Computer Science, vol 12250. Springer, Cham. https://doi.org/10.1007/978-3-030-58802-1_4]*

***[9]**Kahil M.S., Bouramoul A., Derdour M. (2020) Big Data and Interactive Visualization: Overview on Challenges, Techniques and Tools. In: Ezziyyani M. (eds) Advanced Intelligent Systems for Sustainable Development (AI2SD'2019). AI2SD 2019. Advances in Intelligent Systems and Computing, vol 1105. Springer, Cham. https://doi.org/10.1007/978-3-030-36674-2_17*

***[10]** Daniil R., Wohlgenannt G., Pavlov D., Emelyanov Y., Mouromtsev D. (2019) A New Tool for Linked Data Visualization and Exploration in 3D/VR Space. In: Hitzler P. et al. (eds) The Semantic Web: ESWC 2019 Satellite Events. ESWC 2019. Lecture Notes in Computer Science, vol 11762. Springer, Cham. https://doi.org/10.1007/978-3-030-32327-1_33*

***[11]**Chawla, G., Bamal, S., & Khatana, R. (2018). Big data analytics for data visualization: Review of techniques. International Journal of Computer Applications, 182(21), 37-40.*

***[12]**Lavalle, A., Teruel, M. A., Maté, A., & Trujillo, J. (2020). Improving sustainability of smart cities through visualization techniques for big data from iot devices. Sustainability, 12(14), 5595.*

***[13]**Andrienko, G., Andrienko, N., Drucker, S., Fekete, J. D., Fisher, D., Idreos, S., ... & Sharaf, M. (2020, March). Big data visualization and analytics: Future research challenges and emerging applications. In BigVis 2020-3rd International Workshop on Big Data Visual Exploration and Analytics.*

***[14]**Liu, J., Tang, T., Wang, W., Xu, B., Kong, X., & Xia, F. (2018). A survey of scholarly data visualization. Ieee Access, 6, 19205-19221.*

***[15]**Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., ... & Milic, N. M. (2019). Reveal, don't conceal: transforming data visualization to improve transparency. Circulation, 140(18), 1506-1518.*

***[16]**Eberhardt, A., & Silveira, M. S. (2018, May). Show me the data! A systematic mapping on open government data visualization. In Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age (pp. 1-10).*

***[17]**Kerren, A., Kucher, K., Li, Y. F., & Schreiber, F. (2017). BioVis Explorer: A visual guide for biological data visualization techniques. PLoS One, 12(11), e0187341.*

***[18]**Comba, J. L. (2020). Data visualization for the understanding of COVID-19. Computing in Science & Engineering, 22(6), 81-86.*

***[19]**Ceccarini, C., Mirri, S., Salomoni, P., & Prandi, C. (2021). On exploiting data visualization and IoT for increasing sustainability and safety in a smart campus. Mobile Networks and Applications, 26(5), 2066-2075.*

***[20]**Hajirahimova, M., & Ismayilova, M. (2018). Big data visualization: Existing approaches and problems. Problems of Information Technology, 9, 72-83.*

***[21]** M. Feng et al., "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," in IEEE Access, vol. 7, pp. 106111-106123, 2019, doi: 10.1109/ACCESS.2019.2930410.*

***[22]** C. Chen et al., "InfDetect: a Large Scale Graph-based Fraud Detection System for E-Commerce Insurance," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 1765-1773, doi: 10.1109/BigData47090.2019.9006115.*

***[23]** Peter ChunYu Yau, Dennis Wong, Woo Hok Luen, and Joseph Leung. 2020. Understanding Consumer Behavior by Big Data Visualization in the Smart Space Laboratory. In Proceedings of the 2020 5th International Conference on Big Data and Computing (ICBDC 2020). Association for Computing Machinery, New York, NY, USA, 13–17. DOI:https://doi.org/10.1145/3404687.3404705*

***[24]** R. A. Leite et al., "EVA: Visual Analytics to Identify Fraudulent Events," in IEEE Transactions on Visualization and Computer Graphics, vol. 24, no. 1, pp. 330-339, Jan. 2018, doi: 10.1109/TVCG.2017.2744758.*

***[25]** A. Sanchez and W. Rivera, "Big Data Analysis and Visualization for the Smart Grid," 2017 IEEE International Congress on Big Data (BigData Congress), 2017, pp. 414-418, doi: 10.1109/BigDataCongress.2017.59.*

***[26]** A. Zhu, X. Li, Z. Mo and R. Wu, "Wind power prediction based on a convolutional neural network," 2017 International Conference on Circuits, Devices and Systems (ICCDS), 2017, pp. 131-135, doi: 10.1109/ICCDS.2017.8120465.*

***[27]** R. Donida Labati, A. Genovese, V. Piuri, F. Scotti and G. Sforza, "A Decision Support System for Wind Power Production," in IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 50, no. 1, pp. 290-304, Jan. 2020, doi: 10.1109/TSMC.2017.2783681.*

***[28]** M. Li and Q. Zhou, "Industrial Big Data Visualization: A Case Study Using Flight Data Recordings to Discover the Factors Affecting the Airplane Fuel Efficiency," 2017 IEEE Trustcom/BigDataSE/ICESS, 2017, pp. 853-858, doi: 10.1109/Trustcom/BigDataSE/ICESS.2017.322.*

***[29]** D. Deibe, M. Amor and R. Doallo, "Big data storage technologies: a case study for web-based LiDAR visualization," 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 3831-3840, doi: 10.1109/BigData.2018.8622589.*

***[30]** Fan Du, Catherine Plaisant, Neil Spring, Kenyon Crowley, and Ben Shneiderman. 2019. EventAction: A Visual Analytics Approach to Explainable Recommendation for Event Sequences. ACM Trans. Interact. Intell. Syst. 9, 4, Article 21 (December 2019), 31 pages. DOI:https://doi.org/10.1145/3301402*

***[31]**Naidoo, J., & Campbell, K. (2016, October). Best practices for data visualization. In 2016 IEEE International Professional Communication Conference (IPCC) (pp. 1-3). IEEE.*

***[32]**Fan, C., Zhang, Y., Pan, Y., Li, X., Zhang, C., Yuan, R., ... & Huang, H. (2019, July). Multi-horizon time series forecasting with temporal attention learning. In Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery & data mining (pp. 2527-2535).*

***[33]**Burch, M., Huang, W., Wakefield, M., Purchase, H. C., Weiskopf, D., & Hua, J. (2020). The state of the art in empirical user evaluation of graph visualizations. IEEE Access, 9, 4173-4198.*

[34]*Song, H., Dai, Z., Xu, P., & Ren, L. (2021). Interactive Visual Pattern Search on Graph Data via Graph Representation Learning. IEEE Transactions on Visualization and Computer Graphics, 28(1), 335-345.*

[35]*Costa, I., Lima, R., dos Santos, C. G. R., Meiguins, B. S., Soares, A. G. M., & da SilvaFranco, R. Y. (2021, July). EmojiText: An Information Visualization Technique for Analyzing Phrases and Sentiments. In 2021 25th International Conference Information Visualisation (IV) (pp. 114-119). IEEE.*

[36]*Pachas-Baños, A., De La Cruz-Leyva, J., & Shiguihara-Juárez, P. (2019, November). Effectiveness of graph data visualization of news using VR experience. In 2019 IEEE Sciences and Humanities International Research Conference (SHIRCON) (pp. 1-4). IEEE.*

[37]*Dien, H. E., & Asnar, Y. D. W. (2018, November). OPT+ Graph: Detection of Graph Data Structure on Program Visualization Tool to Support Learning. In 2018 5th International Conference on Data and Software Engineering (ICoDSE) (pp. 1-6). IEEE.*

[38]*Lin, J., Cai, Y., Wu, X., & Lu, J. (2021). Graph-Based Information Block Detection in Infographic with Gestalt Organization Principles. IEEE Transactions on Visualization and Computer Graphics.*

[39]*Kawada, H., Akaishi, M., & Hosobe, H. (2018, July). A Graph-Based Visualization of Time-Series Information in Multiple Texts. In 2018 22nd International Conference Information Visualisation (IV) (pp. 44-49). IEEE.*

[40]*Er, X., & Sun, Y. (2021, July). Visualization Analysis of Stock Data and Intelligent Time Series Stock Price Prediction Based on Extreme Gradient Boosting. In 2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE) (pp. 272-279). IEEE.*

[41]*Luo, Y., Qin, X., Tang, N., & Li, G. (2018, April). Deepeye: Towards automatic data visualization. In 2018 IEEE 34th international conference on data engineering (ICDE) (pp. 101-112). IEEE.*