**PhD Research Topic:**

*DEVELOPING AN EFFECTIVE FRAMEWORK FOR BIG DATA VISUALISATION WITH FOCUS ON METHODS APPLICABLE TO THE BIG DATA V*3

***1.1 General Outline***

General Outline

The growing reliance of the 21st century has resulted in large sets of data being processed and stored on daily basis. Much of this information contains data from Bank Records to Medical Records, from Financial Credit Rating to Political Party Alliance, from Criminal Records to Shopping Habits, from Social Network Data to e-government applications, and so forth. This amount of data when used wisely has vast resources that can be used for strategic, tactical and operational decision in every work of life.

The current digital space comprises of loose pieces of data about a single entity in different formats, variations, volumes and structures. These pieces of information when they meet the criterion of either “Volume-Velocity Class, Volume Variety Class, Velocity Variety Class and Volume Velocity Variety class, are what make a big data.

Big data according to (Evgeniy & Vasiliy, 2013) is a phenomenon with no boundaries consisting of infinite data accumulation that can be presented in various ways. The name itself implies a large sets of data in various formats, it can be structured, unstructured and has 3 main properties, namely Volume, Velocity and Variety.

*The scope of the proposed research is to design and develop a visualisation framework over big data to allow easy and interpretable data analytics in order to derive intelligence and extract useful knowledge.*

Although visualisation tools and methods such as the Tree maps developed by (B. Johnson & B. Shneiderman) and Circle packing method developed by (Huang and Chen, 2006) have been deeply investigated recently, the problem of extending the methods into a framework that will allow Analysis and Response based on a user centred approach has not been investigated.

Other notable frameworks research in this subject area is the (Komlodi, John, & Wayne, 2004)visualisation framework for intrusion detection.

Inspired by this main motivation is the research aim, which will design and develop a user centred visualisation framework over big data to allow easy and interpretable data analyses in order to derive intelligence and extract useful knowledge.

Some of the notable shortcomings of other similar research frameworks that the proposed research intends overcome include lack of scalability over time, Visual Noise and lack of user centred approach that allows user entry and response, as noted by (Sifei, Xiaorong, & Henry, 2015), many methods out there only concentrate on visualisation and not response from the system.

The decision for which visualisation method to be applied to volume and variety class of big data would involve the analysis of methods based on the following factors; **being applicable to large volumes of data**, **likelihood of data visualisation**, **data presentation in different formats** and **data presentation performance**.

***1.2 Big Data Visualisation Methods***

The below visualisation methods can be used to support decision making on which method is applicable to a big data class with further analysis into applicable challenges and approaches. Each description argues the classification to a specific big data class with assumption of the following data criteria’s;

* Large Volume Data
* Data Variety
* Data dynamics

*1.2.2 TreeMap: (Volume)*

Commonly used for displaying hierarchical information using nested tables, tree map as described by (Evgeniy & Vasiliy, 2013) has a strict requirement for application to data-data object filling which has to be linked together hierarchically. A common example of a tree map visualisation method is the free spaces on computer hard rives.

Tree maps developed by (B. Johnson & B. Shneiderman) are effective for visualisation of hierarchical and quantitative information the techniques has been used by Stock Markets Electronic Product Catalogues to Query and Visualise Data (Baehrecke, 2004).

Tree Map method can be applied to large volumes of data and it is scalable in the sense that when spaces are filled the analyst can move to the next block or go further down the lower level hierarchy, hence making it suitable for developing s scalable visualisation framework for Bid Data

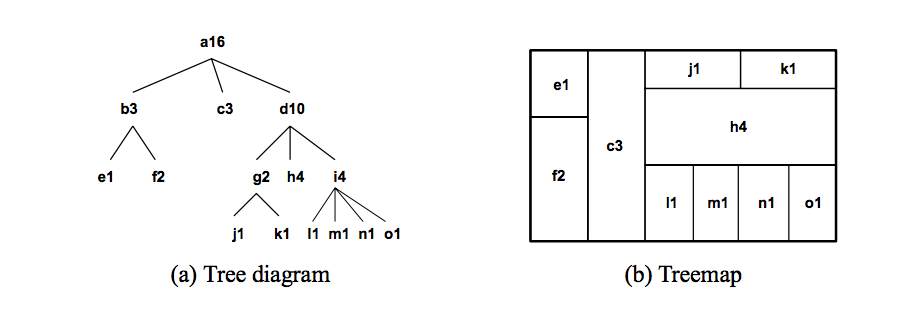


Figure 1. Tree Diagram (https://graphics.ethz.ch/teaching/scivis\_common/Literature/squarifiedTreeMaps.pdf)

*Pros*

1. Grouping hierarchically outlines a clear data relationship
2. Special Colours make extreme out layers clearly visible

*Cons*

1. Tree maps require data to be hierarchical
2. Tree maps are not effective for querying and visualising historical trends pattern
3. There are no negative values for factors in size calculation

*1.2.1 Circle Packing*

Circle method is the direct opposite of Treemap, they are commonly used for large amounts of hierarchical structured data. Circle packing method was developed by (Huang and Chen, 2006). Circle packing method is based on tree map method hence inheriting properties from tree map that includes large volume data criteria.

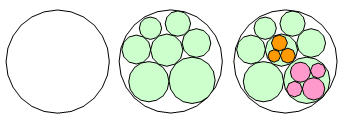


Figure 2. Packing Circle with random radii (http://www.infovis-wiki.net/index.php?title=Circle\_Packing)

In Circle Packing, Nodes are packed into smaller nodes as presented in the figure above, the root nodes holds the first level of nodes, child nodes are further packed into each node in the root and if a node has children they are packed into it. Although circle method also has its differences from tree map as below

*Pros*

1. Circle Packing methods are space efficient compared to tree maps
2. Circle packing method is suitable for large and hierarchical data types

*Cons*

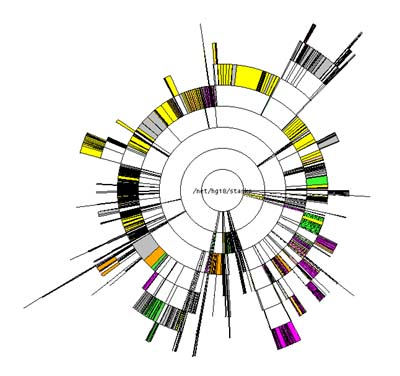
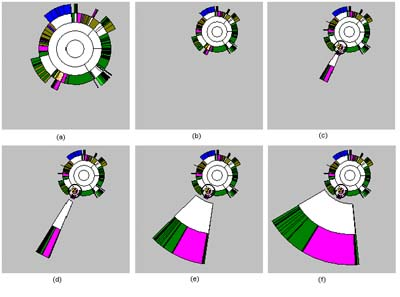
1. Circle Packing share similar disadvantages to tree map method including its suitability for hierarchical data and no negative values factor in size calculation

1.2.3 *Sunburst:*

As the name implies is a radial tree map like method that uses polar coordinate systems. The Sunburst uses Radius and Arc length as oppose to width and height used by tree map, this particular feature of Sunburst allows effective repaint of only the side affected by data change as oppose to recreating the whole view (Evgeniy & Vasiliy, 2013).

The Sunburst’s unique feature makes it suitable for data with variable velocity such as live streaming data hence updating ONLY the side of the diagram affected by data change as oppose to recreating the whole diagram simply by changing its radius.

Sunbursts bursting structure supports scalability of data visualisation simply by continuously growing out in radius and length accommodating more with the help of global views and sequence of angular details on the diagram.

Global view (left) and detailed view (right) of 20,000 files on a hard drive (http://www.cc.gatech.edu/gvu/ii/sunburst/)

*Pros*

1. Supports data velocity such as live data streaming and batches of data
2. Sunburst methods are easily perceptible by humans with the help of its clear views and structures

*Cons*

1. Sunburst views are prone to rapid changes due there support for live data
2. As the same with tree map, requires data to be hierarchical with top of the hierarchy at the centre and deeper levels farther away from the centre (http://www.cc.gatech.edu/gvu/ii/sunburst/)

1.2.4 *StreamGraph*

Steamgraph graph also known as Theme River is a method that involves stacked area graph running across an x-axis depicting how data has changed over time.

Perhaps the best known example of a stream graph was published in the New York Times in February 2008 depicting the ([Ebb and Flow of Movies: Box Office Receipts Over Past 20 Years](http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html)).

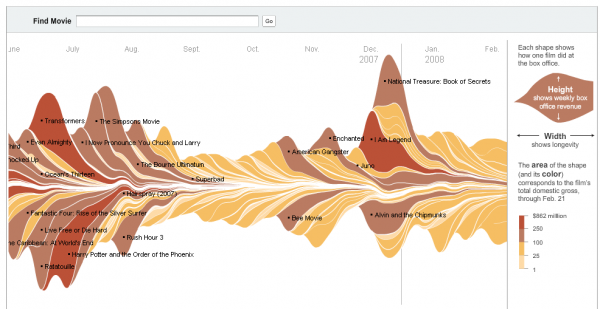


Figure 3. Films from the summer 2007 (http://leebyron.com/streamgraph/stackedgraphs\_byron\_wattenberg.pdf)

Steamgraph graph also known as Theme River is a method that involves stacked area graph running across an x-axis depicting how data has changed over time

Steam graph is effective for showing the many individual time series while also displaying their quantitative variables. It combines the height of individual layers with the height of the overall graph to create steam like graph structure, however some trade-offs include no space in-between the layer as this would affect the sum and change on middle layers will cause wiggles on the other layers even when the data has no relationship with the changes data (Byron & Martin , 2012).

Steam graphs methods are not effective for big data classes with Variety and Velocity, but can still be used for Volume. Although (Evgeniy & Vasiliy, 2013) further explains that when new data is added to steam graphs, the subsequent diagrams can be used to live stream data but only data related to quantity. Common example uses of steam graphs include Cinema Trends and Musical Genre.

*Pros*

1. They are very effective for visualising trends over time
2. The method is effective for large data volume with emphasis on

*Cons*

1. A single factor is displayed during data representation
2. The method is majorly dependent on data layers

1.2.5 *Parallel Coordinates*

Parallel coordinates are one of the most famous visualization techniques, and among the most common subjects of academic papers in visualization. While initially confusing, they are powerful tool for understanding multi-dimensional numerical datasets.

In Parallel Coordinates, all the data factors that need to be analyzed are placed on one axis and the corresponding values of data in relative scale are placed on the other side (Evgeniy & Vasiliy, 2013). Linked cross lines represent each data object displaying its position in the context of other objects.

The Parallel Coordinates method supports all three criteria of big data, Volume Velocity and Variety. The method can handle large objects per single screen for several factors which satisfies **Variety**, it depends on maximum and minimum values for each factor and recreating the diagram while only affecting the side where data has changed hence, making it suitable for **Velocity** in big data, and also support large data sets in **Volumes** (Malakar and Vishwanath, 2015).

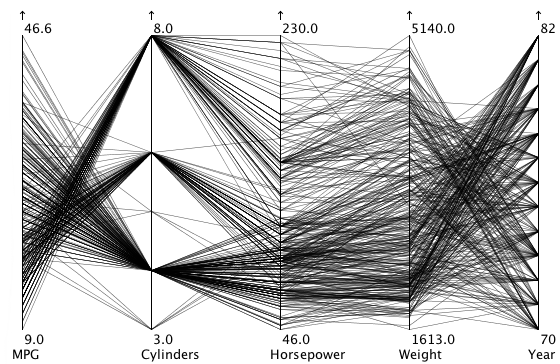


Figure 4. Parallel Coordinates (https://eagereyes.org/techniques/parallel-coordinates)

*Pros*

1. Ordering of different factors will not influence entire diagram hence accommodating varieties of data to b e visualised
2. Allows analysis of both whole data and individual data at the same time.

*Cons*

1. There is limitation to the number of factors that can be shown at a time
2. Visualisation of live data can enhance the diagram completely by rapid changes to the view.

As a result of the above analyses of the big data visualisation methods, which are; Tree Map, Circle Packing, Steam Graph, Sunburst and Parallel Coordinates, with there *pros* and *cons,* we will be showing which method is suitable for large volume data (Volume), method that can handle various types of data (Variety) and Methods that can handle change such as live data streams (Velocity).

Tree Map methods can only be applied to hierarchical data same with Circle Packing, Sunburst can be applied to data with Volume and Velocity same with Steam Graph which can also can be applied to Volume and Velocity.

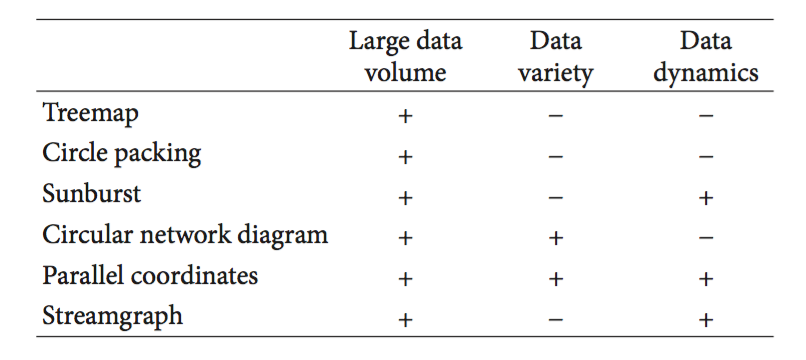


Figure 5. Classification of Visualisation methods based on there properties (Evgeniy & Vasiliy, 2013)

As a result the above table can be used to categorise visualisation methods based on there applicability to the big data classes and there major impact on building a framework for visualisation which will guide the project as to what method would suite each data classes.

***1.3 Research Aim***

This research will design and develop a visualisation framework over big data to allow easy and interpretable data analytics in order to derive intelligence and extract useful knowledge. The research aims to extend the extensive work done on Parallel Coordinates method of big data visualisation*.*

***1.4 Research Objectives***

* To Design and Develop a scalable framework for big data visualization
* To integrate existing Parallel Coordinate visualisation method with common open source data analytics tools such Hadoop and Rapid miner
* To explore interactive visualisation and user centred tools for Volume, Velocity and Variety focused Data
* To evaluate the framework with a real world showcase
* To discover insight and knowledge and provide suggestions for what visualisation method is applicable to specific big data classes.

**2. Related Studies**

# *2.1* An Information Visualization Framework for Intrusion Detection

# Various investigations have been conducted in the area of information visualization framework with this specific research by (Komlodi, John, & Wayne, 2004) carried out on a framework for intrusion detection. The author identified three main task phases that are Monitoring, Analysis and Response after interviews and evaluations with security experts.

# The author noted that the most important information feature should always be emphasized and the visualization should make information drill down possible

# Most importantly in the research, the author demonstrated a user-centered research and requirement gathering visualization framework development efforts, the author also noted that many intrusion detection visualization systems out there only addresses the monitoring phase with little regard for the other two phases.

# In relation to the proposed research, after visualization and analysis, the proposed framework could allow a more complex response based on user defined selection, filter, clustering, drill down and zooming. This could improve the complete framework for large data visualization as proposed by the research.

# *2.2* A framework for Cloud-Based large scale Data-Analytics and Visualization

# Another related research conducted by (Sifei, Xiaorong, & Henry, 2015) is about developing a cloud based data visualization and analytics framework for dengue fever analytics. The author identified R, Rapidminder and Hadoop for effective and high performance cloud based data analytics.

# Part of the research involved a real world data analytics for dengue fever incident located in Singapore with emphasis on analyzing time series data, special temporal data and harmonization visualization in public private cloud.

# 

Figure 6. Cloud Based Visualisation for Rain fall (Sifei, Xiaorong, & Henry, 2015)

# Some of the key features identified by the author includes application and framework architecture which was based on Ecalyptus and Public cloud Amazon, Linux visual machines where created with RapidAnalystics and R, Hadoop.

# Other key features of the framework include the modified interface for Rapidminer, R and Hadoop and interface data visualization.

# In relation to the proposed framework, further development could be carried out using real world data analytics with emphasis on scheduled algorithm to predict future trends after initial data execution with user centered interface designs.

# 2.3 ****Analytics over large-scale multidimensional data: the big data revolution****

# This specific research by (Alfredo, Il-Yeol, & Karen, 2011) emphasis on the Large Scale analytics over multidirectional Data. The author focuses on research issues and achievement in the field of analytics over big data.

# Some of the key tools for big data as discussed by the author includes is Hadoop, with design of complex Hadoop integrated multidimensional data playing a paramount role in the research. Some of the key issues identified by the author includes;

# Combining the benefits of RDBMS and NoSQL databases systems for flexibility of data analysis over big data portions.

# High scalability as a primary feature for any Big Data analysis system

# Data source heterogeneously with emphasis on web, scientific and legacy system data.

# The research provided critical discussion on the state-of-art achievement in analytics over big-data, emerging context of analytics over multidimensional data and actual research trends such as designing complex analytics over hadoop integrated multidimensional data.

# In relation to the proposed framework, further development could be carried out in the filed of integrating Hadoop as big data analytics tool as part of the visualization methods, due to Hadoop capability to adapts its engine to all data evolutions in big data (Alfredo, Il-Yeol, & Karen, 2011).

**3. Proposed research methodologies**

In achieving the aim of this research, which is designing and developing a visualisation framework over big data to allow easy and interpretable data analytics in order to derive intelligence and extract useful knowledge with user’s preferences and needs, method(s) of research has to be adopted during the process. First and foremost, list of user’s preference and needs will have to be gathered and this tends to be achieved using primary means, but we need to first describe it in a more practical way.

**3.1 Methodological Background**

At first, conceptual investigation needs to be conducted on what needs and preferences are and how it can be represented so that they can support effective decision-making and response. Representing these will be conducted with using quantitative research approach.

The combination of a qualitative and quantitative research would be used to guide the proposed study. Mixed research methodology as the name implies involves the use of both Quantitative and Qualitative research approach to carry out research, it allows combination of data collected from Quantitative and Qualitative (M.K, 2006).

The Combining quantitative and qualitative research for the visualisation framework has certain advantages, both research methods would provide a more complete understanding of User needs which will leverages the inductive social nature of qualitative approach and data driven nature of quantitative to successfully carry out the research.

**4. List of sources (or kinds of sources) that will be consulted**

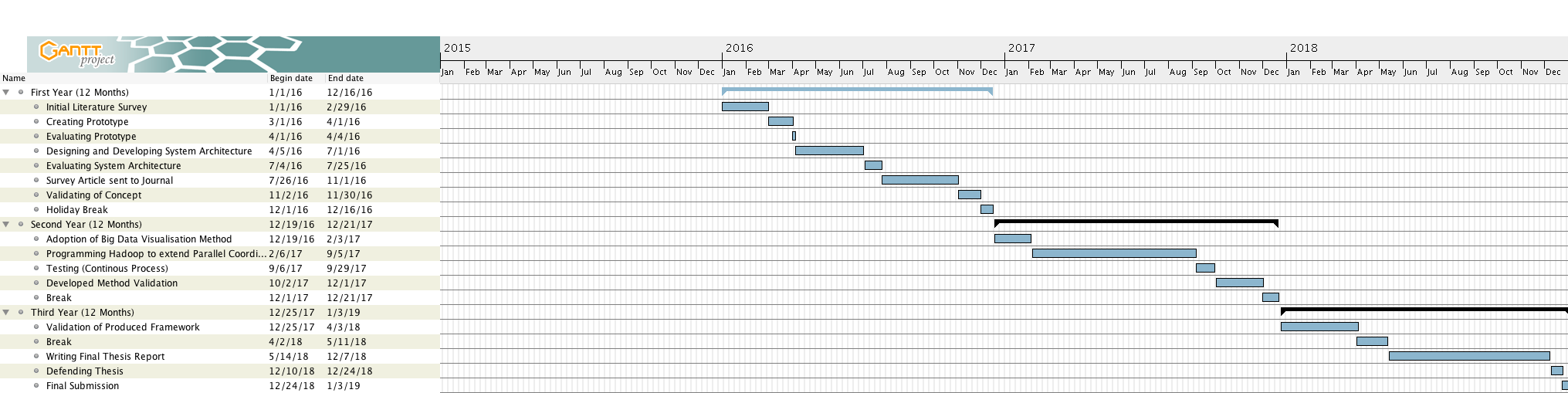
During the process of the research, various types of sources will be consulted to achieve the aim of the research and these sources include:

* Journal articles
* Academic and Trade Journals
* Books and eBooks
* Newspaper Articles
* Academic Websites
* Conference Proceedings
* Published Reports
* Magazine Articles
* Workshop Papers

Other lists of sources that will be consulted can be found at the bibliography of this report.

**5. Timescale (Gantt chart)**

This research is projected to be completed within 36months starting from January 2016 to January 2019. Below is a Gantt chart showing estimate of deliverables:



**Figure 7:** *Gantt-Chat illustrating time plan from the duration of the research life circle (This subject to change)*

**Validation of work:** As an important part of this research, this study intends to conduct a validation process in order to measure the effectiveness of the produced system. A formal experimentation plan will be develop during the process of the research, as experiments will be conducted at early stages to at first identify user’s preferences, before testing the technology with the users to gather their opinion and advice. This validation will be done twice within the research life circle.

**6. Conclusion**

In this proposal, we have identified an emerging problem of visualisation within the filed of Big data with the rapid creation of data. We also identified some challenges associated with visualisation methods such the methods suitable for a particular big data class.

Despite the wide variety of methods and approaches researched in the field of big data visualisation, there is no concrete framework that can be used to guide a data analyst on what method or approach to use based on Data Volume, Variety and Velocity, hence the importance of the proposed research which aims to develop an effective framework to guide through the process of visualisation while avoiding some of the challenges. With the rapid rate of data creation, the research has the potential to make a lasting mark on how data is processed and consumed by the present and future data analysts.

Future work in this filed can be carried out on applying the proposed frameworks to other data driven study areas such as digital forensics where the visualisation framework can be used to identify key information of importance from large sets of information.

# References

Alfredo, C., Il-Yeol, S., & Karen, C. D. (2011). Analytics over Large-Scale Multidimensional Data: The Big Data Revolution. *DOLAP’11* .

Byron, L., & Martin , W. (2012). Stacked Graphs – Geometry & Aesthetics.

Baehrecke, E. H. (2004). Visualization and analysis of microarray and gene ontology data with treemaps. *BMC Bioinformatics 2004* .

Evgeniy , G. Y., & Vasiliy, G. V. (2013). Analytical Review of Data Visualization Methods in Application to Big Data. *Journal of Electrical and Computer Engineering , 2013*.

Huang, Wenqi, and Mao Chen. 'Note On: An Improved Algorithm For The Packing Of Unequal Circles Within A Larger Containing Circle'. *Computers & Industrial Engineering* 50.3 (2006): 338-344. Web.

Komlodi, A., John, R. G., & Wayne, G. L. (2004, April 24–29 ). An Information Visualization Framework for Intrusion Detection. *Department of Information Systems, UMBC* , 113-703 .

Malakar, Preeti, and Venkatram Vishwanath. 'Hierarchical Read–Write Optimizations For Scientific Applications With Multi-Variable Structured Datasets'. *Int J Parallel Prog* (2015): n. pag. Web.

Sifei, L., Xiaorong, L., & Henry, P. (2015). Cloud-Based Data Analytic and Visualization Framework for Dengue Fever Analytics. *Institute of High Performance Computing* .

Shneiderman, Ben. 'Tree Visualization With Tree-Maps: 2-D Space-Filling Approach'. *TOG* 11.1 (1992): 92-99. Web.