

## Proposal Information

Proposal Title: **AutoVis - Automated Visualisation using Machine Learning**

### Student Information:

Name : Muhammad Abdullah Akmal

Email: b7031482@my.shu.ac.uk - agfa.94@gmail.com

Telephone: +44 7568922477

Module: MSc Big Data Analytics

### Supervisor Information:

Name : Dr. Christopher R Roast

Email: c.r.roast@shu.ac.uk

Telephone: +44 1142256845



## Contents

<b>1</b>	<b>Problem Statement</b>	<b>3</b>
<b>2</b>	<b>Objectives</b>	<b>3</b>
<b>3</b>	<b>Introduction</b>	<b>4</b>
<b>4</b>	<b>Motivation</b>	<b>4</b>
<b>5</b>	<b>Literature Review</b>	<b>4</b>
5.1	Related Work . . . . .	6
<b>6</b>	<b>Dataset</b>	<b>6</b>
<b>7</b>	<b>Research Methodology</b>	<b>7</b>
<b>8</b>	<b>Timeline</b>	<b>7</b>
<b>9</b>	<b>Potential Outcome</b>	<b>8</b>
<b>10</b>	<b>Advantages</b>	<b>9</b>
<b>11</b>	<b>Challenges</b>	<b>9</b>
<b>12</b>	<b>Ethics and Conduct</b>	<b>9</b>
	<b>References</b>	<b>10</b>

## List of Figures

1	Project Diagram . . . . .	3
2	Flow of visualisation from speed to expressiveness . . . . .	5
3	Cycle of adopted Research Methodology . . . . .	7
4	Task divisions and Temporary Deadlines . . . . .	8
5	Gantt Chart . . . . .	8

# 1 Problem Statement

”Automatic generation of graphics/visuals from the raw dataset by identifying the structure of dataset using machine learning techniques.”

To explain further, this problem is divided two folds:

1. Collection of datasets that machine learning could be used upon identifying the structure.
2. Then use of machine learning to train a model that can learn the rules of visualisation (mapping from decision matrix to visualisation language).

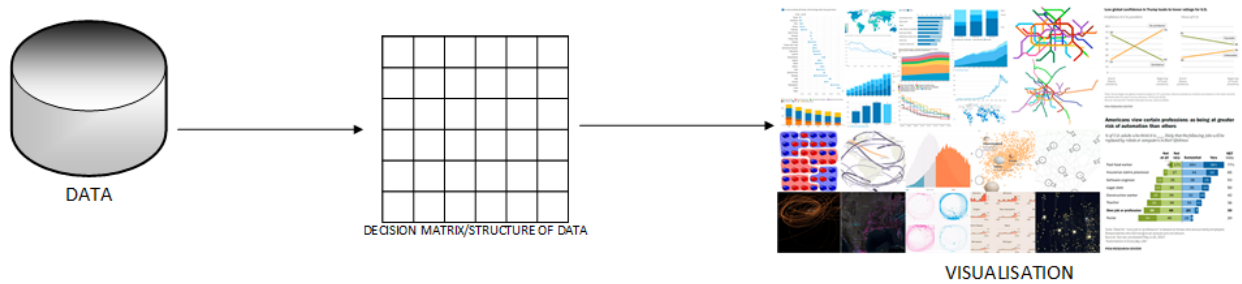


Figure 1: Project Diagram

# 2 Objectives

Following are the main objectives intended to achieve during this project:

1. Collection of datasets which machine learning algorithms can used to train an algorithm.
2. Creation of Decision matrix depending upon the structure of datasets using Neural network. This matrix will contain the information about the what kind of visualisation might be useful for the specific kind of data.
3. Creation of Model that automatically select the subset of fields for visualisation (generally datasets have several fields which cannot be concurrently visualised). This Model would help to identify the differences between different data types. Data can be of any type namely string, numeric, temporal, ordinal, categorical etc. Moreover, applying transformations depending upon the data type e.g. aggregate transform function can be applied to numeric data but cannot to string data. Finally, the model must translate the decision matrix to a visualisation language like vega-lite (Satyanarayan, Moritz, Wongsuphasawat, & Heer, 2017) etc.

### 3 Introduction

Visualisation is divided into two sub fields, scientific visualisation and information visualisation. Scientific visualisation deals with the scientific data which involves spatial component e.g. 3D medical imagery etc. while information visualisation deals with the data that doesn't involve the spatial factor e.g. weather forecast, document data etc. (Tory & Muller, 2004).

Information visualisation is defined by Card, Mackinlay and Shneiderman (Card, Mackinlay, and Shneiderman, 1999) as the tool of visualisation for amplification of human cognitive abilities. Information visualisation is referred to as Visual Data Mining (Frenay & Dumas, 2016), since humans are particularly good at identifying outliers and trends via visualisation (Treisman, 1985).

Machine learning and information visualisation both somehow deal with the better understanding for user via visualisation and analysis of dataset. Machine learning is basically used for finding pattern in large datasets (Frenay & Dumas, 2016). Combining both fields can help in Computationally enhanced visualisation (Rayar, Barrat, Bouali, & Venturini, 2016), visually enhanced mining techniques and Integrated Visualisation and mining possibilities.

We mainly will deal in this project with the information visualisation and its automation using the machine learning.

### 4 Motivation

Majority of the advanced automated visualisation are based on heuristics. Heuristics are set of basic rules which are used to make decisions and judgments about specific data. It would be great if they can learn the patterns by themselves and follow them to improve upon the previous ones. Best visualisation to the user can increase the sense of understanding regarding that specific data. Most of the time in data mining and analysis process is spent in data visualisation and understanding. If this step is done automatically, it will save a lot of process time and resources consumption. Intent of this project is in relation with effective visualisation, automation in visualisation and deep learning neural network. The idea behind this research is to come up with a method which reduces the user role in the data visualisation. Machine learning will play its role by reducing the user only to select data descriptions and may be to some extent identify the algorithms and automatically everything is done by computer. Later, it may be users may be given a set of visualisation from which specific selection can be made.

### 5 Literature Review

Generally, before diving into the data for specific usage, analysts use data visualisation techniques to understand the data. For that they use different range of tools starting from the

completely abstract tools, easy to learn and fast to make visualisations e.g. Microsoft excel, Google Sheets are easy to use but have limited functionalities. While, some of them require expertise and give more appropriate results e.g. HTML canvas and OPENGL gives the better results but require programming expertise to achieve it (Dibia & Demiralp, 2018) as shown in figure 2.

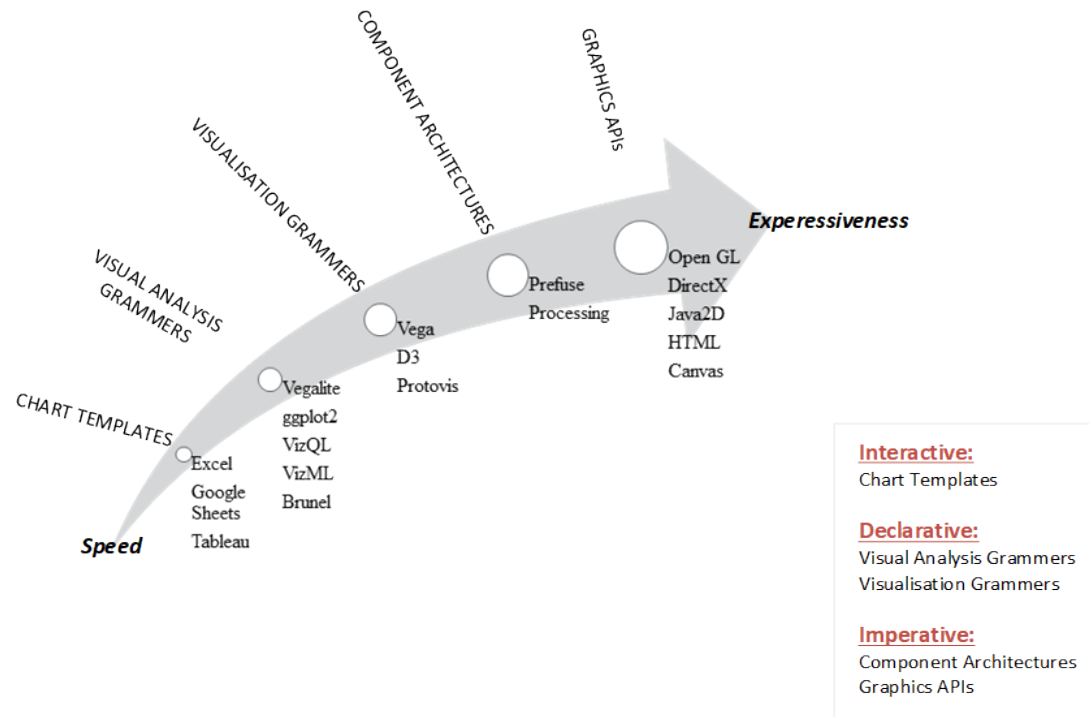


Figure 2: Flow of visualisation from speed to expressiveness

Declarative Languages basically forms a logical and methodological framework for program and system. It uses propositional techniques based on rational concepts for specifying the properties and objects. (Broy, 1991). They are like a tradeoff between others. You get enough speed and expressiveness (Wickham, 2007). But they might be difficult to understand the syntax and cope with the level of abstraction adopted. Plus, they might have some re-usability issues e.g. for non-R (the programming language) users it might be difficult to understand ggplot2 etc.

One of the key issues in the visualisation is the scalability. Humans cannot perceive the knowledge when more than a few combined features are put together. Moreover, there perception is limited in between 2 and 3 dimensions. In real time, its really difficult to compute large datasets interactively. This is where machine learning comes in. ML provides with the ability to automatically summarize or compress the big data solutions via clustering and projection

methods (Keim, Munzner, Rossi, & Verleysen, 2015). Connecting ML with IV(Information Visualisation), we can somehow resolve the human perception problem and interactivity issues.

DeepDive, Yago and EntityCube are some of the examples of corpus building which extracts the data and by the help of Deep NLP tools from the web and then create knowledge base systems (Niu, Zhang, & Shavlik, 2012).

Just to gain more insight the next subsection will go through the related work done.

## 5.1 Related Work

BOZ is an automated visualisation tool used for designing of the task-oriented graphics and presentation of these graphics. BOZ allows the users to draw the logical conclusion from the set of graphics and help in streamlining this information depending upon the search of the users (M.Casner, 1991). It only uses analytical approach that combines set of heuristics to reach to the required searched information.

VizDeck (Key, Howe, Perry, & Aragon, 2012) is a self-organising dashboard for visual analytics. By looking at the statistical properties of the data, it recommends the appropriate visual analytics. A prototype card games adopts these recommendations and organises the interactive visual dashboard in no time without any programming.

APT (A Presentation Tool) is a prototype model create for automated designing of the graphical presentation by clearly defining graphical language that explains the syntactic and semantic properties of the graphical presentation. (Mackinlay, 1986). AI was used in implementation of this prototype and most of the design is generated using the compositional algebra which includes the compositional operators and primitive graphical languages. It deals with the 2D static presentations automations like bar chart, scatter plot, connected graphs etc.

SAGE (Roth, Kolojejchick, Mattis, & Goldstein, 1994), an automated presentation designing system, which takes as an input the data characteristics and primitive knowledge about the visualisation intentions. By incorporating SageBrush (graphics are construct using the primitives about the design or partial design) and SageBook (browser for retrieving previously created images). SAGE inherit functionalities from other systems like APT, BOZ and ANDD (Automated Network Diagram designer, which takes the network model and set of design directives as an input and produces network diagram (Marks, 1991)).

Data2Vis (a web-based prototype), uses LSTM-based neural translation model which formulate the seq2seq translation, train it and then generate visualisation (Dibia and Demiralp, 2018).

## 6 Dataset

R - Dataset repository (R-DataSet, 2018) would be main source of data. While, different kind of data will be collected depending upon the structure, after initial training and testing of the

model to improve the model.

## 7 Research Methodology

Top-Down(Deductive) approach will be utilised in this project implementation as shown in the figure 3.

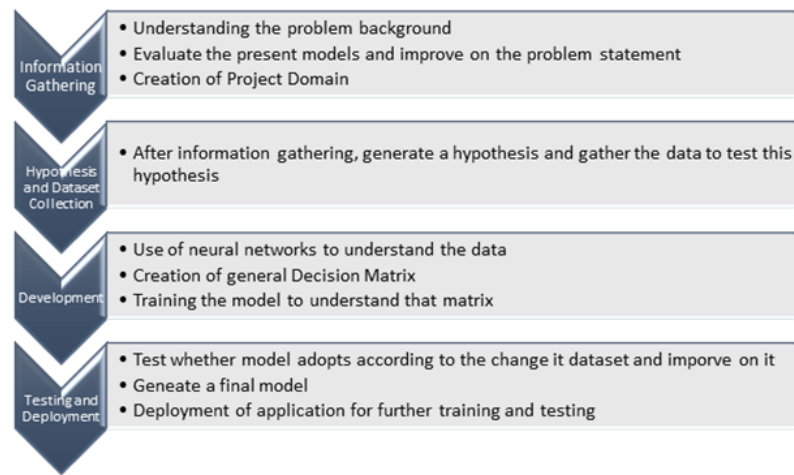


Figure 3: Cycle of adopted Research Methodology

Step 3 and Step 4 are recursive depending upon the improvements required. While, step 2 remain interchangeable according to the requirement.

Intention is to start in maturing the problem statement and go through more research understanding the implementation of the current system discussed in the section 4.1. R-Dataset is primary source while we are keen to create or find more data resources which have patterns and make sense to machine learning algorithms.

We basically rely on the Deep Learning algorithms of machine learning to create decision matrix and then built on that to understand and visualise the data. Literature shows the seq2seq, LSTM neural translation accompanying with RNN (Recurrent Neural Network) and Vega-lite may be an approach to this problem. Generating this model, training it and deploying to validation and further improvements will be last and final step.

## 8 Timeline

Following Gant Chart shows the initial workplan to complete the project and division into different tasks as shown in the figures 4 and 5.

		Name	Duration	Start	Finish	Predecessors
1		Literature Survey	13 days	7/25/18 8:00 AM	8/10/18 5:00 PM	
2		Getting Started with Training Neural Network	14 days	8/1/18 8:00 AM	8/20/18 5:00 PM	
3		Understanding the Vegalite	4 days	8/15/18 8:00 AM	8/20/18 5:00 PM	
4		Dataset Gathering	20 days	8/13/18 8:00 AM	9/7/18 5:00 PM	1
5		Project Implementation	60 days	9/10/18 8:00 AM	11/30/18 5:00 PM	1;4
6		Projec Report/Final Thesis	66 days	9/20/18 8:00 AM	12/20/18 5:00 PM	1;2;3;4
7		Project Testing and Improvements	14 days	12/3/18 8:00 AM	12/20/18 5:00 PM	5
8		Paper Writing	45 days	12/25/18 8:00 AM	2/25/19 5:00 PM	6

Figure 4: Task divisions and Temporary Deadlines

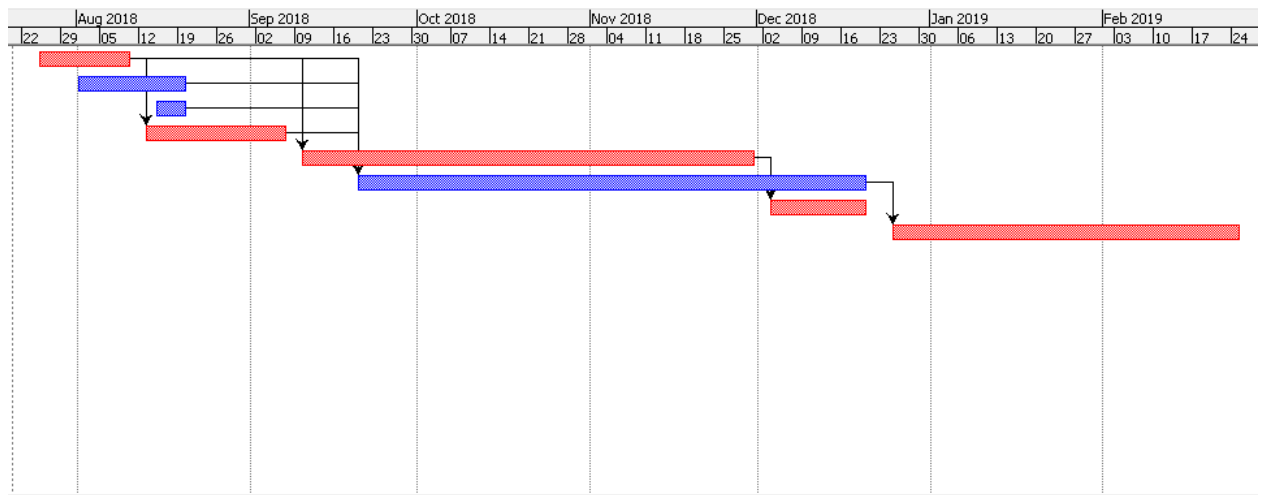


Figure 5: Gantt Chart

## 9 Potential Outcome

We intend to come up with deep neural network approach like RNN multi sequential with long-short term memory, building on the previous knowledge. This would be able to successfully convert the set of data into a knowledgeable matrix which can be further used to take visualisation decisions. Moreover, for testing/validation we intend to create a prototype web or software-based application that would be able to perform three operations:

- Importing the data (potentially in JSON)
- Generate Visualisation
- Update Visualisation (applying different transformation functions)



## 10 Advantages

This research project can help in following lines:

- It will enable user to create insightful visualisation with less or no programming.
- It will help in handy, fast and escalate the visualisation capabilities of users.
- Analysts can use it for initial understanding of the data and then adopting the algorithms according to the results, speeding up the process for them.
- It will help in understanding the structure of data which is unusual or unknown.
- It can help in exploring the data which is complex, saving the effort of going through applying different visualisation techniques and then reaching to the one that makes sense.

## 11 Challenges

Following are the likely challenges in this project:

- Tackling the unstructured data, training, modelling and visualisation must be one of the challenges.
- Collection of data and making the machine learning algorithms to understand this data is a big challenge.
- Automatic selection of attributes would be one of the challenges
- Incorporation of different data visualisation techniques at one place may be something achievable but might prove to be tedious work.

## 12 Ethics and Conduct

Meaningful visualisation results in increase of knowledge and understanding in the relevant problem. This can help in future prediction which can help in more improve decision in that field. Since, visualisation is the cognitive process which is under research, so it is difficult to come up with ethical guide to visualisation but some of the ethics that should be considered by the designers while creating graphics:

- Visualisations are intended to bring attention to relevant matter.
- Visualisations are based on thorough analysis of information.

- Visualisation are built in a way that are easy to comprehend.
- Selection of meaningful, clear, efficient and in-depth informative graphic that makes sense to the viewer rather than selecting one which a designer likes or easy to implement (Cairo, 2014).
- Things like hierarchy of visual properties and appropriate labelling must be kept in mind while designing (Skau, 2012).
- Depict the data and analysis in accurate way.
- Clearly exposing to different visualisation techniques and remain open to criticism.
- Visualisation shouldnt be intentionally used for hiding or confusing the truth. It shall not misguide the uninformed.
- Designer shall remain fully responsible for virtual and actual meaning portrayed by the graphics.

## References

- Broy, M. (1991). *Declarative specification and declarative programming*. Retrieved from <https://dl-acm-org.lcproxy.shu.ac.uk/citation.cfm?id=952788>
- Cairo, A. (2014). *Ethical infographics: In data visualization, journalism meets engineering*.
- Dibia, V., & Demiralp, C. (2018). *Data2vis: Automatic generation of data visualizations using sequence to sequence recurrent neural networks*. Retrieved from <http://arxiv.org/abs/1804.03126>
- Frenay, B., & Dumas, B. (2016). *Information visualisation and machine learning: Characteristics, convergence and perspective*. Retrieved from <https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2016-18.pdf>
- Keim, D. A., Munzner, T., Rossi, F., & Verleysen, M. (2015). *Bridging information visualisation with machine learning*. Retrieved from <https://www.dagstuhl.de/en/program/>
- Key, A., Howe, B., Perry, D., & Aragon, C. (2012). *Vizdeck: Self-organizing dashboards for visual analytics*. Retrieved from <https://doi.org/10.1145/2213836.2213931>
- Mackinlay, J. (1986). *Vizdeck: Self-organizing dashboards for visual analytics*. Retrieved from <https://doi.org/10.1145/22949.22950>
- Marks, J. W. (1991). *Automating the design of network diagrams*. Retrieved from <https://dl.acm.org/citation.cfm?id=124656>
- M.Casner, S. (1991). *A task analytic approach to the automated design of graphic presentation*. Retrieved from <https://doi.org/10.1145/108360.108361>

- Niu, F., Zhang, C., & Shavlik, J. W. (2012). *Deepdive: Web-scale knowledge-base construction using statistical learning and inference* (Vol. 12(1)).
- Rayar, F., Barrat, S., Bouali, F., & Venturini, G. (2016). *Incremental hierarchical indexing and visualisation of large image collections*. Retrieved from <https://hal.archives-ouvertes.fr/hal-01315650/document>
- R-DataSet. (2018). *R-dataset repository information*. Retrieved from <https://vincentarelbundock.github.io/Rdatasets/datasets.html>
- Roth, S. F., Kolojejchick, J., Mattis, J., & Goldstein, J. (1994). *Interactive graphic design using automatic presentation knowledge*. Retrieved from <https://doi.org/10.1145/191666.191719>
- Satyanarayan, A., Moritz, D., Wongsuphasawat, K., & Heer, J. (2017). *Vega-lite: A grammar of interactive graphics*. Retrieved from <https://doi.org/10.1109/TVCG.2016.2599030>
- Skau, D. (2012). *A code of ethics for data visualization professionals*. Retrieved from <https://visual.ly/blog/a-code-of-ethics-for-data-visualization-professionals/>
- Tory, M., & Muller, T. (2004). *Rethinking visualization: A high-level taxonomy*. Retrieved from <https://doi.org/10.1109/INFVIS.2004.59>
- Treisman, A. (1985). *Preattentive processing in vision*. Retrieved from [https://doi.org/10.1016/S0734-189X\(85\)80004-9](https://doi.org/10.1016/S0734-189X(85)80004-9)
- Wickham, H. (2007). *A layered grammar of graphics*. Retrieved from <https://doi.org/10.1198/jcgs.2009.07098>