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

List of Figures

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).

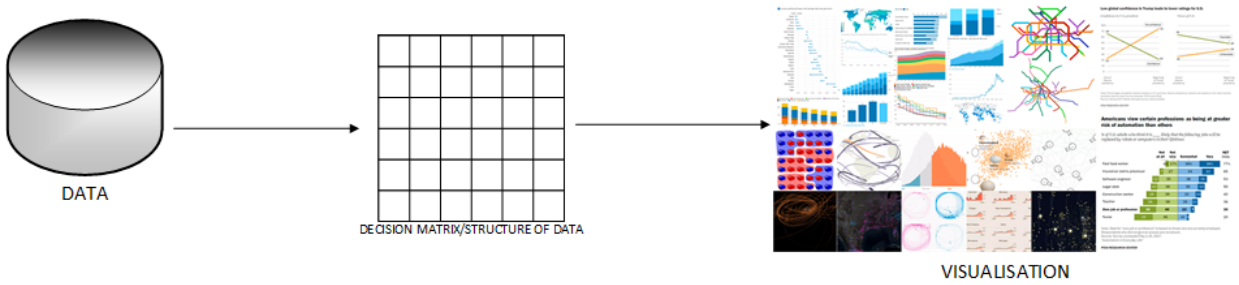

Figure 1: Project Diagram

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, Mackingly and Shneiderman (Card, Mackinglay, and Shneiderman, 1999) as the tool of visualisation for amplification of human cognitive abilities. Information visualisation is referred 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 use to make decisions and judgments about specific It would be great if they can learn the patterns by itself 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 spend in data visualisation and understanding. If this step is done automatically, it will save alot 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, analysists 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.

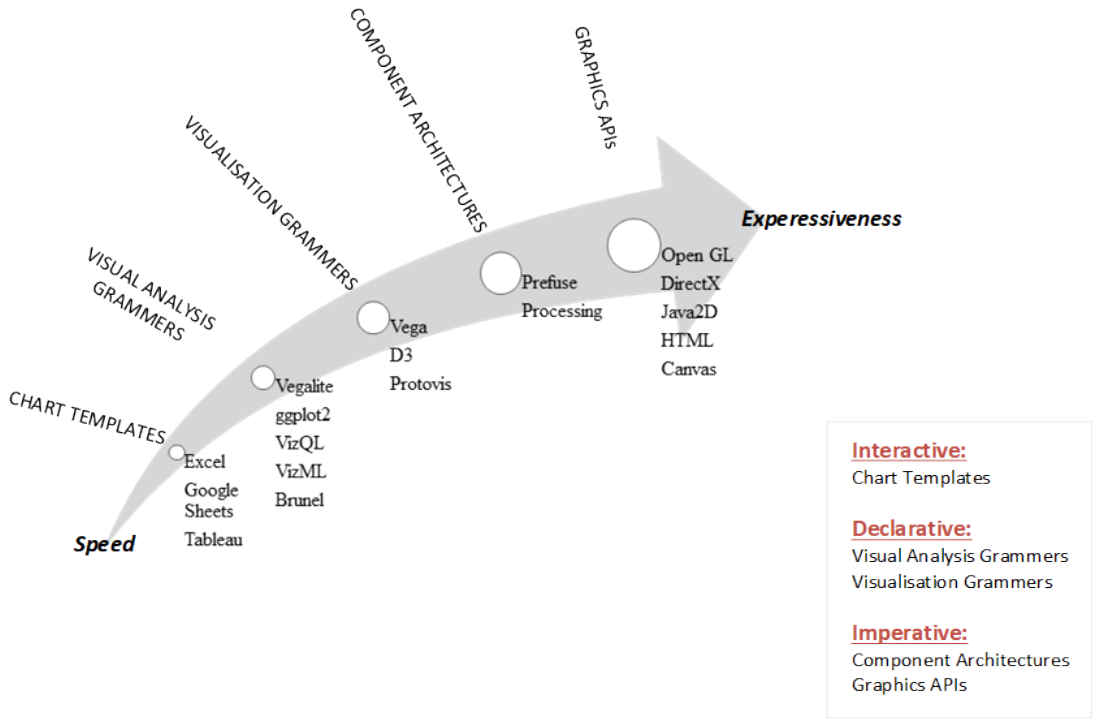

Figure 2: Flow of visualisation from speed to expressiveness

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.

Artificial Neural Networks(ANN) are computing system which are inspired from the biological neurological system of animals. These system have an ability to learn from its past experiences, gain on it progressively and help in automatic decision making process using the improved knowledge (Dawson, 2016). It consists of neurons. These are divided into different layers and they can send signals to each others in the form of 0's and 1's.

If there are multiple layers between input and output layer then such ANN is called Deep Neural Networks (DNN). They are usually deepforward networks, i.e. they move from input to output, with different weights attached to neurons. If they can move in any direction then such networks are called Recurrent Neural Networds (RNN), they consist of memory Long-short term memory (LSTM), which allows it to store information while propagation (Hochreiter & Schmidhuber, 1997).

DNN's surpasses the machine translation system performance, as compared to phrase based approach. DNN's used large datasets which help in appropriate model fitting. Different approaches already have been employed in industry to translate the domain specific languages and programming languages, try to make system learn to write imporoved code (Balog, Gaunt, & Brockschmidt, 2017). Data2Vis (Dibia & Demiralp, 2018), used the RNN seq2seq to directly translate the source and target. (Poco & Heer, 2017) used the CNN (Convolutional Neural Networks) for the classification of type of graphical markers (e.g. bars, lines, areas, points etc.). This classification was further used to encode data in the chart. This encoding is used to recover visual encoding specification. They used OCR alongside it also.

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. These set of heurisics can help us to understand what rules are used to identify the data. Which ultimately can help in building the decision matrix.

Interactive visualisation is really important when we deal with data exploration. To acheive that Vizdeck model can prove to be helpful.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. To bring on the automation, the working of APT can be really helpful in the current scenario. APT (A Presentation Tool) is a porotype 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 & Demiralp, 2018). Data2Vis gives the understanding how deep learning models can be used to identify the structure via visualisation languages.

6 Dataset

(Poco & Heer, 2017) generated the dataset using the Vega Visualisation grammer by automatically generating the charts, collecting the charts from the Quartz (a news website) and RDataset. We will try to follow the same methodoly. RDataset (R-DataSet, 2018) is available automatically, vega files `https://github.com/victordibia/data2vis/tree/master/bin/tools`. It consists of the 1147 datasets distributed in the statistical software and software packages. By using the compass recommendation engine (Wongsuphasawat et al., 2016) on 11 different dataset vega specification charts will be created and by using the python we will extract different charts from the Quartz website. This is an approximate estimation of chart corpus that would be available to us after the processing table 1

	Vega Charts	**Quartz**
Area Charts	477	0
Bar Charts	1358	191
Line Charts	360	283
Scatter Plots	2123	1
Total	**4,318**	**475**

Table 1: Approximately Estimated Chart Corpus

7 Research Methodology

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

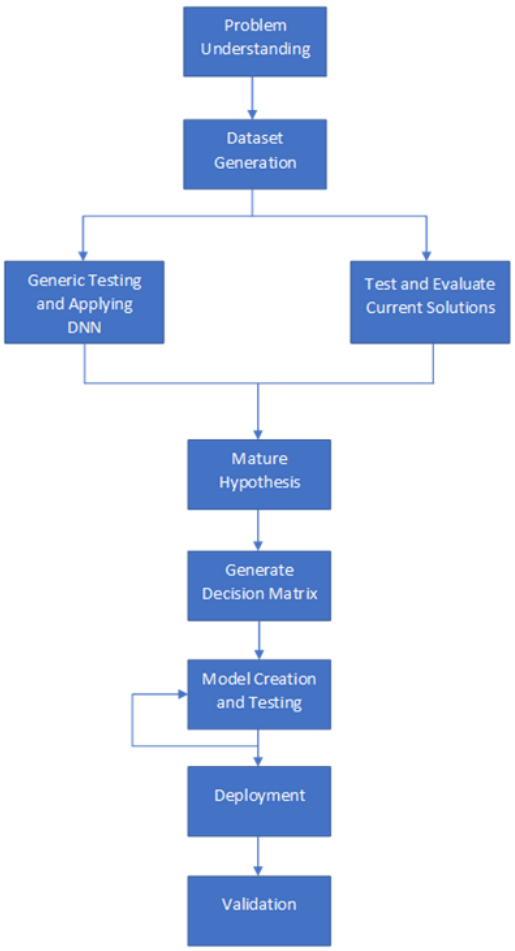

Figure 3: Cycle of adopted Research Methodology

Following are the steps:

1. Understand the problem background
2. Generate Dataset
3. Start generic Testing to understand the data and applying different machine learning techniques

4. Test and Evaluate the present solution
5. Mature hypothesis
6. Generate Decision Matrix
7. Model Creation and Testing
8. Deployment
9. Validation with Unknown Dataset via deployed application

Intention is to start in maturing the problem statement, alongside generating the dataset and go through more research understanding the implementation of the current system discussed in the section 5.1. After that Mature the hypothesis, generate the decision matrix which contains information about the visualisation of data. This decision matrix will further be used to fit a model. Using cross validation, it would be tested and this step is repeatable to achieve better results. After that we will move towards the deployment and testing of the final system.

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	Problem Understanding & Literature Survey	10 days	8/19/18 8:00 AM	8/31/18 5:00 PM	
2	Dataset Generation	10 days	8/24/18 8:00 AM	9/6/18 5:00 PM	
3	Generic Testing and Applying DNN's	10 days	9/6/18 8:00 AM	9/19/18 5:00 PM	
4	Test and Evaluate Current Solutions	12 days	9/6/18 8:00 AM	9/21/18 5:00 PM	
5	Understanding the Vegalite and Mature Hypothe	4 days	9/19/18 8:00 AM	9/24/18 5:00 PM	2
6	Project Implementation - Phase 1 (Generate Dec	20 days	9/25/18 8:00 AM	10/22/18 5:00 PM	1;2;5
7	Project Implementation - Phase 2 (Initial Model F	20 days	10/23/18 8:00 AM	11/19/18 5:00 PM	1;2;6
8	Project Implementation - Phase 3 (Final Model Ge	20 days	11/20/18 8:00 AM	12/17/18 5:00 PM	1;2;7
9	Projec Report/Final Thesis	41 days	10/23/18 8:00 AM	12/18/18 5:00 PM	1;2;3;4;5;6
10	Paper Writing	30 days	12/19/18 8:00 AM	1/29/19 5:00 PM	9

Figure 4: Task divisions and Temporary Deadlines

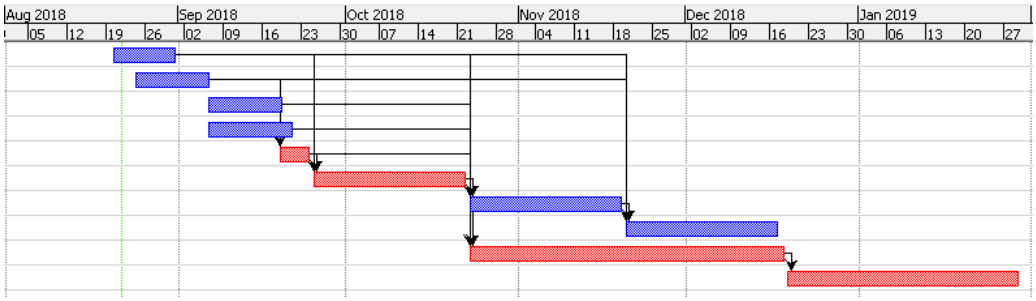

Figure 5: Gantt Chart

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.
- Analysists 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.
- Automatically 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 shouldn't 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

Balog, M., Gaunt, A. L., & Brockschmidt, M. (2017). *Deepcoder: Learning to write programs.* Retrieved from https://arxiv.org/pdf/1611.01989.pdf

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.*

Dawson, C. (2016). *Applied neural networds.* Retrieved from http://www.mdpi.com/books/pdfview/book/236

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

Hochreiter, S., & Schmidhuber, J. (1997). *Long short-term memory* (Vol. 9) (No. 8). Retrieved from https://doi.org/10.1162/neco.1997.9.8.1735 doi: 10.1162/neco.1997.9.8.1735

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

Poco, J., & Heer, J. (2017). *Reverse-engineering visualizations: Recovering visual encodings from chart images* (Vol. 36(3)). Retrieved from https://idl.cs.washington.edu/files/2017-ReverseEngineeringVis-EuroVis.pdf

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

Wickham, H. (2007). *A layered grammar of graphics*. Retrieved from https://doi.org/10.1198/jcgs.2009.07098

Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., & Heer, J. (2016). *Voyager: Exploratory analysis via faceted browsing of visualization recommendations* (Vol. 22) (No. 1). doi: 10.1109/TVCG.2015.2467191