USING MACHINE LEARNING TO SUPPORT BETTER AND INTELLIGENT VISUALISATION FOR GENOMIC DATA

Zhonglin Qu 18885806

Master of Philosophy

Supervisor: Dr Quang Vinh Nguyen

Co-Supervisor: Dr Yi Zhou

Co-Supervisor: Assoc. Professor Daniel R. Catchpoole

Submitted in fulfilment of the requirements for the degree of
Master of Philosophy

School of Computing, Engineering and Mathematics
Western Sydney University
2019

Keywords

Machine Learning, Visualisation, Artificial Intelligence, Virtual Reality, Genomic Data, Cancer Data, Visualisation Tools, Information Visualisation, Intelligent Visualisation, Visual Analytic.

Abstract

Massive amounts of genomic data are created for the advent of *Next Generation Sequencing* technologies. Great technological advances in methods of characterising the human diseases, including genetic and environmental factors, make it a great opportunity to understand the diseases and to find new diagnoses and treatments. Translating medical data becomes more and more rich and challenging. Visualisation can greatly aid the processing and integration of complex data. Genomic data visual analytics is rapidly evolving alongside with advances in high-throughput technologies such as Artificial Intelligence (AI), and Virtual Reality (VR). Personalised medicine requires new genomic visualisation tools, which can efficiently extract knowledge from the genomic data effectively and speed up expert decisions about the best treatment of an individual patient’s needs. However, meaningful visual analysis of such large genomic data remains a serious challenge.

Visualising these complex genomic data requires not only simply plotting of data but should also lead to better decisions. Machine learning has the ability to make prediction and aid in decision-making. Machine learning and visualisation are both effective ways to deal with big data, but they focus on different purposes. Machine learning applies statistical learning techniques to automatically identify patterns in data to make highly accurate prediction, while visualisation can leverage the human perceptual system to interpret and uncover hidden patterns in big data. Clinicians, experts and researchers intend to use both visualisation and machine learning to analyse their complex genomic data, but it is a serious challenge for them to understand and trust machine learning models in the serious medical industry.

The main goal of this thesis is to study the feasibility of intelligent and interactive visualisation which combined with machine learning algorithms for medical data analysis. A prototype has also been developed to illustrate the concept that visualising genomics data from childhood cancers in meaningful and dynamic ways could lead to better decisions. Machine learning algorithms are used and illustrated during visualising the cancer genomic data in order to provide highly accurate predictions. This research could open a new and exciting path to discovery for disease diagnostics and therapies.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

If appropriate, list any abbreviations used in the thesis.

AI: Artificial Intelligence
VR: Virtual Reality
AR: Augmented Reality
GBM: Glioblastoma Multiforme
LGG: Lower Grade Glioma
DNA: Deoxyribonucleic Acid
RNA: Ribonucleic Acid
TCGA: The Cancer Genome Atlas
SNA: Special Nucleic Acids
MNV: Murine Norovirus
InDels: Insertion and Deletion
GO: Gene Ontology
GIAB: Genome in a Bottle
UCSC: University of California Santa Crus
AML: Acute Myeloid Leukemia
ALL: Acute Lymphoblastic Leukaemia
ARMS: Alveolar Rhabdomyosarcoma
ERMS: Embryonal Rhabdomyosarcoma
3D: Three Dimensions
ID3: Iterative Dichotomiser 3
BMT: Bone Marrow Transplantation

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: [redacted]

Date: 06/08/2019

Acknowledgements

I cannot believe I am finally here. This thesis would not have been possible without the ongoing support of my supervisors, family, friends, and the academic staff at Western Sydney University.

In particular, I would like to acknowledge the dedication and guidance I received from my supervisors: Dr Quang Vinh Nguyen, Dr Yi Zhou and Assoc. Professor Daniel R. Catchpoole. Thank you all for your support and for sharing your extensive knowledge with me. Vinh, thank you for warmly welcoming me to the world of visual analytics and sharing your knowledge with me. I would not have made it to this point without your patience, persistence and care. Yi, thank you for helping me find and understand all the AI and machine learning knowledge. Dan, thank you for introducing me to the medical industry and arousing my curiosity about genomic data.

To my family: my husband Johnny, my children Zihan and Collins, thank you for your enduring support throughout this process. I would not have made it this far without your love and support.

Finally, to my amazing friends in the HICI group, thank you all for making this academic journal slightly more manageable and enjoyable.

Chapter 1: Introduction

In recent years, Artificial Intelligence (AI) has started to be used for big data visualisation including multivariate genomic data for the development of new technologies. Machine learning, as one branch of the AI field, is a way of solving problems without explicitly codifying the solution and a way of building systems that improve themselves over time. AI or machine learning in specific has been applied in genomics for analysing genome sequencing, gene editing, clinical workflow and direct-to-consumer genomics. Future applications of machine learning in the field of genomics are diverse and may potentially contribute to the development of patient or population-specific pharmaceutical drugs. Although machine learning has extraordinary predictive abilities, the machine learning models and the algorithms are hard to understand and maybe even harder to trust, especially in serious industries such as the medical industry. Visualising machine learning models and predictive results in a meaningful way can interpret the complex algorithms and help clinicians, researchers and experts understand and trust the predictive results.

This thesis proposes a novel visualisation prototype that can illustrate the machine learning model and real-time predictive results along with conventional visualisation methods. The visualisation integrates a machine learning model and gives real-time predictions to assist researchers or clinicians' decisions. The process of machine learning prediction is illustrated in the visualisation as well. The new visualisation tool can interpret the machine learning model for the domain experts who may not be familiar in predictive mathematics algorithms, and it can make the genomic data visualisation and decision-making procedure more reliable for them.

This chapter outlines the genomic data background (Section 1.1), the context of visualisation (Section 1.2), visualisation for genomic data (Section 1.3), Artificial Intelligence (AI) and machine learning for genomic data (Section 1.4). Section 1.5 describes the research aim and research questions. Finally, Section 1.6 includes an outline of the remaining chapters of the thesis.

1.1 GENOMIC DATA BACKGROUND

After computers and the Internet entered almost every arena of human society, large amounts of digital data are generated and collected in the different format, including medical and genomic data, which is also advancing at a dramatic pace. We are entering an era of big data – datasets that are characterised by high volume, velocity, variety, resolution and indexicality, relationality and flexibility (Khushboo Wadhwani, 2017). Large datasets have become very important sources for discovering insights and ultimately helps to make more precise decisions. However, big data brings in some challenges such as volume, variety, combining multiple data sets, velocity, veracity, data quality, data availability, data discovery, data quality, data extensiveness, personally recognisable information, data assertiveness, quantifiability, data processing, and data management (Khushboo Wadhwani, 2017).

Genomic is a recent convergence of multiple science disciplines including genetics, molecular biology, biochemistry, statistics and computer sciences. Since Gregor Mendel, known as the "father of modern genetics", discovered the basic principles of heredity which became the foundation of modern genetics and leading to the study of heredity (Biography, 2017), huge amounts of genomic data have been collected around the world by different organisations. For example, one of the world's largest pharmaceutical companies AstraZeneca has launched a massive effort to compile genome sequences and health records from two million people (Ledford, 2016). The company and its collaborators hoped to unearth rare genetic sequences that are associated with disease and with responses to treatment (Ledford, 2016). Meanwhile, Human Genome Project had successfully completed the ambitious goal of collecting sequence code covering three billion base pairs in the human genome, two years ahead of the previous projects (Francis S. Collins, 2012). From Figure 1, we can see the exponential pace of genomic data growth.

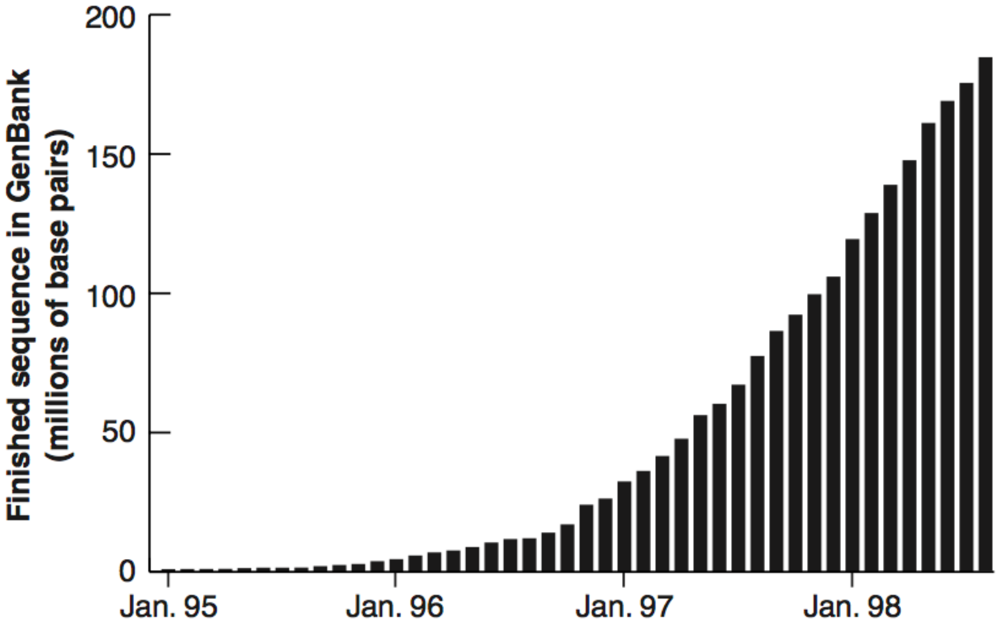


Figure 1 Worldwide human genome sequencing progress

(measured as base pairs of finished sequence deposited with GenBank) (Francis S. Collins, 2012).

With the development of new technologies, genomic data can be collected and stored in a short period of time and the cost has been dramatically reduced as well. Importantly, the technologies also lead to the age of individual genome sequencing which supports an era of personalised medicine (McClean, 2011). Personalised cancer medicine based on the molecular characteristics of a tumour from an individual patient has great potential in the therapy of many types of cancer (Wistuba, Gelovani, Jacoby, Davis, & Herbst, 2011).

In Figure 2, we can find that DNA sequencing capacities have grown rapidly since 2015 and this trend will likely continue in the future. If the growth continues at the current rate by doubling every seven months, then we should reach more than one Exabytes (10^{18}) of sequence per year in the next five years and the approach one Zettabytes (10^{21}) of sequence per year by 2025 (Stephens et al., 2015). In human health, the major needs driven by the big data are how to interpret genomic sequences and how to find patterns over very large collections in very high dimensions.

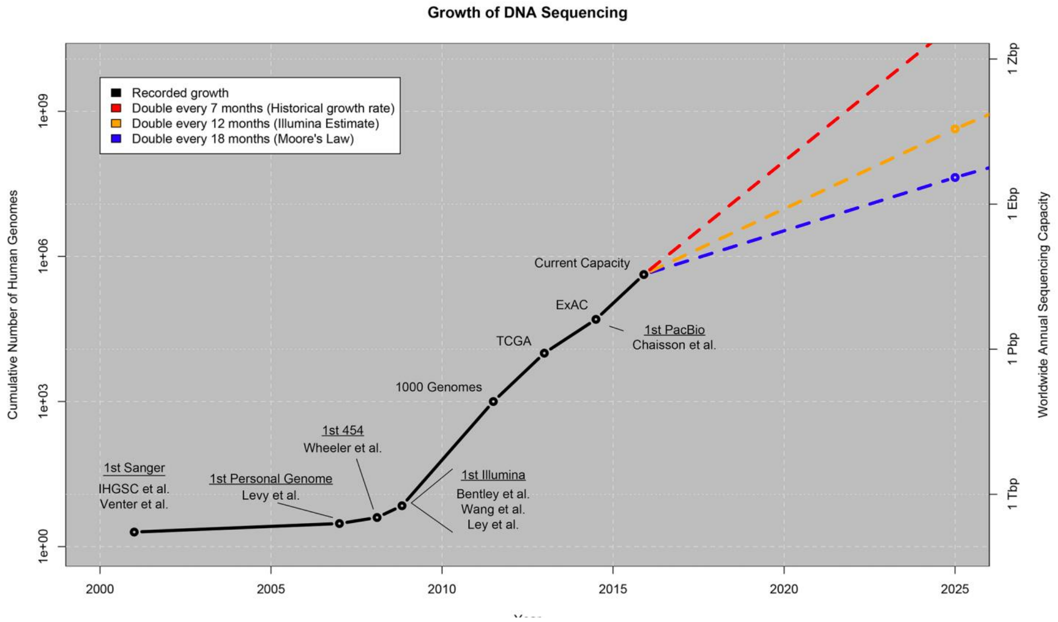


Figure 2 Growth of DNA sequencing.

The plot shows the growth of DNA sequencing both in the total number of human genomes sequenced (left axis) as well as the worldwide annual sequencing capacity (right axis: Tera-basepairs (Tbp), Peta-basepairs (Pbp), Exa-basepa (Stephens et al., 2015).

1.2 VISUALISATION

"A picture is worth a thousand words." This is an adage especially for life science which is one of the biggest generators of enormous datasets as a result of recent and rapid technological advances. Visualisation is a quick, easy way to convey large amounts of complex data in a universal manner to help humans finding the potential values in their big data. Visualisation is becoming an increasingly important part of cognitive systems which can provide the highest bandwidth channel from the computer to the human. The term visualisation, in the past, meant constructing a visual image in the mind and now comes to mean something more like a graphical representation of data or concepts. The visualising way can be functioned as a cognitive tool which has the following advantages: providing an ability to comprehend huge amounts of data; allowing the perception of emergent properties that were not anticipated; enabling problems with the data to become immediately apparent; facilitating understanding of both large-scale and small-scale features of the data; and facilitating hypothesis formation (Green, Ribarsky, & Fisher, 2008; Keahey, 2013; Ware, 2013). Some intuitive visualisation tools are used to visualise multidimensional cancer genomics data which integrate different types of alterations with clinical data for extraction of useful knowledge from the vast amount of data generated by high-throughput technologies (Q. V. Nguyen, Qian, Huang, & Zhang, 2013; Schroeder, Gonzalez-Perez, & Lopez-Bigas, 2013).

There are hundreds of visualisation methods in the research community. Different visualisation techniques may suit with different applications and datasets with different sizes and properties. As shown in Figure 3, Pie chart, bar chart, line chart, and bubble plot are classic visualisation techniques. Pie charts can show 3-10 data items, bar charts can show fewer than 50 data items, line charts can show fewer than 500 data items, bubble plots can show fewer than 500 data items and scatterplots can show fewer than 10,000 data items.

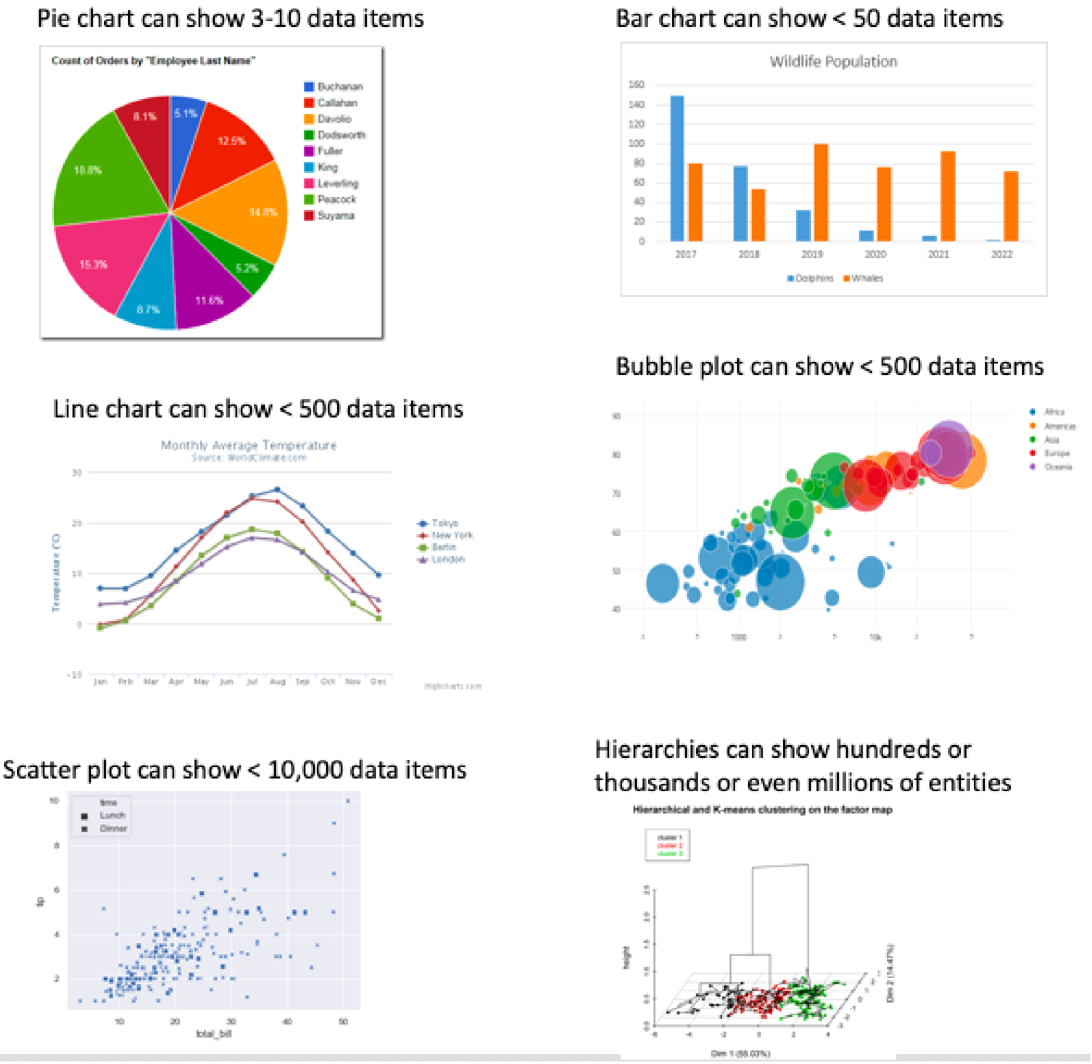


Figure 3 Different visualisation techniques suit with different size of datasets.

New visualisation techniques are used for big data such as hierarchies, which are very popular in data analytics and are powerful data abstractions for aggregating information into broader categories. Hierarchies are often referred as “tree” and some of them change to tree map. Many tree visualisation methods finely tuned for specific types of data such as genome sequencing, large social graphs and tournament matches, which can show hundreds or thousands or even millions of entities, are arranged in a hierarchical structure (Keahey, 2013). For example, Figure 4 shows a treemap visualisation of a collection of choices for streaming music and video tracks by a social network community, that a media service could find useful when designing personalised offers of music and videos for download.

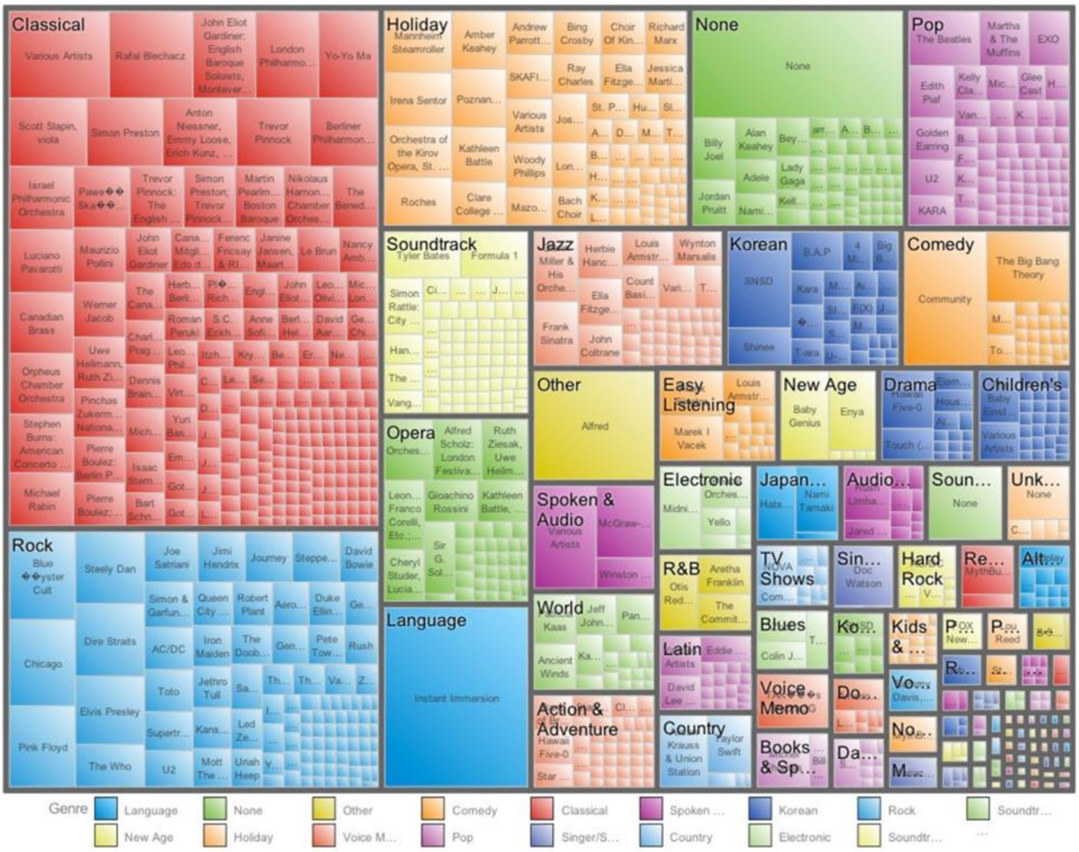


Figure 4 Treemap view of a social network’s track selections from a streaming media service network.

Colour represents the genres of the selected tracks, with each genre subdivided into rectangles for each artist. Size of rectangle for both genre and artist represents the number of track plays in that category (Keahey, 2013).

1.3 VISUALISATION FOR GENOMIC DATA

The complexity of the genomic data makes these datasets incomprehensible without effective visualisation methods. Genomic data visualisation is a rapidly evolving field that has achieved in many areas such as hardware acceleration, standardised exchangeable file formats, dimensionality reduction, visual feature selection, multivariate data analyses, interoperability, 3D rendering and visualisation of complex data at different resolutions, especially in image processing combined with artificial intelligence-based pattern recognition (Pavlopoulos et al., 2015).

Figure 5 shows the trends of multivariate data analyses and visualisation including A) Timeline of the emergence of relevant technologies and concepts, and B) Visualisation of k-means partitional clustering algorithm, C) 3D visualisation of a principal component analysis, D) Visualisation of gene-expression measures across time using parallel coordinates, E) Visualisation of gene-expression clustering across time, F) 2D hierarchical clustering to visualise gene expressions against several time points or conditions, G) Hypothetical integration of analyses and expression heatmaps and the control of objects by VR devices (Pavlopoulos et al., 2015).

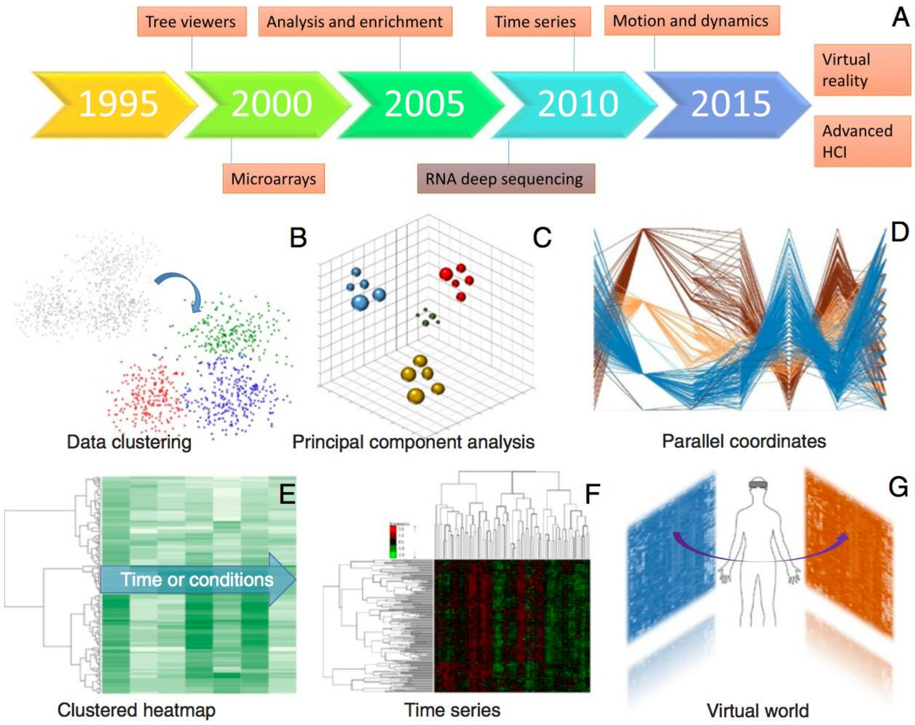


Figure 5 Trends of multivariate data analyses.

Visualisation corresponding to the timeline of relevant technologies and methods (Pavlopoulos et al., 2015).

1.4 AI FOR GENOMIC DATA VISUALISATION

Artificial Intelligence (AI) is already part of our everyday lives and has been heralded as the key to our civilization's brightest future (Mills, 2016). Machine learning, as an approach to achieve artificial intelligence, is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world (Copeland, 2016). AI and machine learning boost the next generation of visualisation -- intelligent visualisation. Intelligent visualisation could remove the need for a human user to handle tedious or repetitive tasks by learning from previous sessions and input data. Intelligent visualisation combines machine learning algorithms to make high-level, goal-oriented decisions, which makes data visualisation technology directly accessible to a wide range of application scientists (Fuchs, Waser, & Groller, 2009; Ma, 2007).

Some modern data visualisation tools use AI technology, modern 3D plot(s), mobile device(s) and VR technique(s) to tell the full story of genomic data. 3D and VR techniques immerse the user in a digitally created space and simulate movement in three dimensions to greatly increase the bandwidth of data available to our brains (Leung, Delong, Alipanahi, & Frey, 2016; Q. V. Nguyen et al., 2016; Shilling, 2017). Most of the visualisation tools allow users to interact with the data in a way that is far more natural such as reaching out to manipulate objects with our hands, moving around them to view them from a clearer perspective and highlighting objects of interest with a point of the finger.

Machine learning combined with data visualisation should have three stages: developing an algorithm, applying genomic data to the algorithm, and predicting new unlabelled data (Libbrecht & Noble, 2015). Figure 6 shows a canonical example of a machine learning application with these three stages. A training set of DNA sequences is provided as input to a learning procedure, along with binary labels indicating whether each sequence is centred on a transcription start site (TSS) or not. The learning algorithm then produces a model that can then be subsequently used, in conjunction with a prediction algorithm, to assign predicted labels (such as 'TSS' or 'not TSS') to unlabelled test sequences. In the figure, the red-blue gradient might represent, for example, the scores of various motif models (one per column) against the DNA sequence.

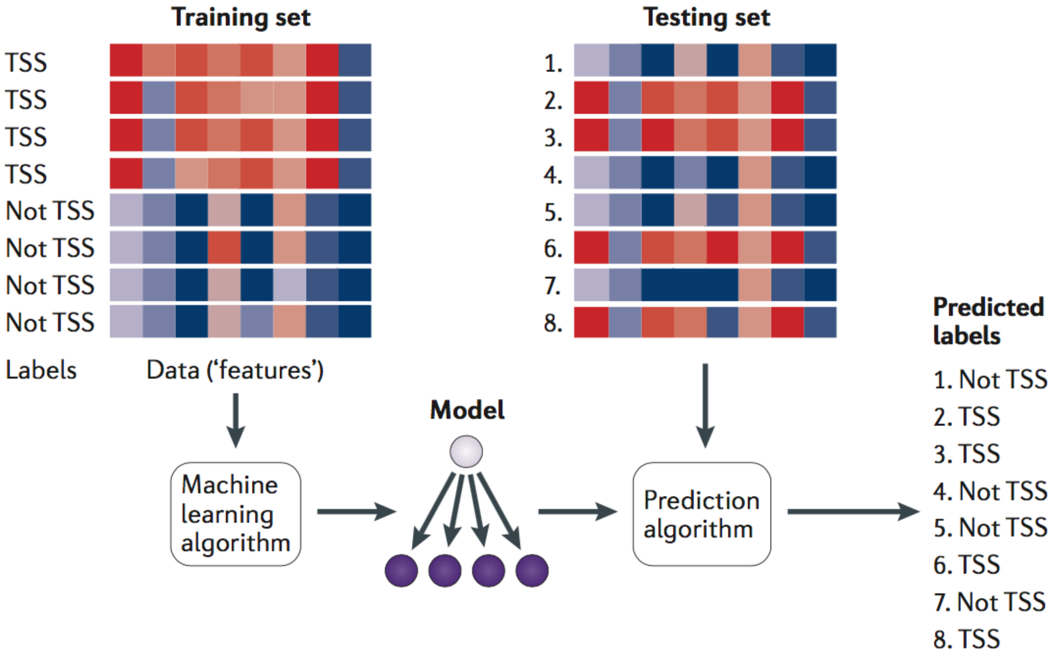


Figure 6 A canonical example of a machine learning application with these three stages for DNA sequences.

(Libbrecht & Noble, 2015).

1.5 RESEARCH AND QUESTIONS AND AIM

Machine learning methods have become more important to genomic data. Gnomic visualisation tools combined with machine learning algorithms would be a new trend in the genomic visualisation evolution in the future. Machine learning is able to address important problems in genomic medicine, for example, creating a predictive model to determine how variations in the DNA of individuals can affect the risk of different disease and to find causal explanations, so that targeted therapies can be designed (Leung et al., 2016). Intelligent data visualisation can provide support to find the relationship between genomic data and diseases and then cure the disease with targeting personalised therapy (Quang Vinh Nguyen et al., 2011). In the analysis of genomic data, the current statistical analysis methods are not enough for achieving data insight, meanwhile, the application of machine learning and data visualisation has become more attractive. Although machine learning has extraordinary predictive abilities, the machine learning models and the algorithms are hard to understand and maybe even harder to trust, especially in serious industries such as the medical industry. Visualising machine learning models and predictive results in a meaningful way can help interpret the complex algorithms and help clinicians, researchers and experts understand and trust the predictive results.

The aim of this thesis is to develop a meaningful intelligent visualisation phototype to show large and complex genomic data processed by machine learning models in order to find clues amongst childhood genomic cancer data. Decision tree, which is a machine learning algorithm, will be used in this intelligent visualisation phototype to provide predictive choices for clinicians to assist their decisions. The visualisation prototype not only visualises the patients' data with the traditional techniques, but also illustrates the machine learning model and prediction process. The success of this research will aid the clinicians to tailor the treatment to the most efficacious for each individual and access to complex genomics data in meaningful and predictive ways.

The following is my research questions:

1. *How to use the decision tree model to support effective and intelligent visualisation for genomic data?*

The research aims to choose and apply a suitable machine learning algorithm to process genomic data and then visualise intelligently the structured data in the cohort. In this stage, we will use the decision tree model do predictions and help clinicians and medical researchers make better decisions. The model will create effective and intelligent scatterplot visualisations where the axis and attribute mappings and visual properties can be selected intelligently based on the user preference and the nature of the data.

2. *How to develop a prototype to illustrate the effectiveness of the model to support better and intelligent visualisations by using scatterplots?*

The thesis will develop an application and visualisation methods to present genomic data interactively and intelligently. The visualisation methods should be suitable for visualising big data and suitable for interpreting the machine learning model to the users as well.

Before we developed our prototype, we carried out a usability study to get users' requirements and includes i) a systematically review about popular genomic data visualisation tools, and ii) a preliminary study for the qualitative review. In the systematically review, we provide a comprehensive comparison of the tools in both aspects i) the visualisation methods in genomic and cancer data fields and ii) the trends of visualisation in genomic analytics fields from year 2000s. We reviewed the situation of current genomic and cancer data, the potential application to personalised medicine, and methods for genomic data visualisation. Here we assess the units of traditional approaches such as scatter plots, heatmaps, coordinates, networks and clustering, as well as emerging technologies involving AI and VR. We also review the evolution of genomic data visualisation tools from the speed of technology development, effective interactions, current tool status, tool integrations and new features. In the qualitative review, we interviewed five domain experts to collect feedback of three visualisation tools in order to gain a better understanding of the domain users' preferences and expectations for the new genomic data visualisation tools.

The research starts with a literature review which presents the related work. Then it follows with three steps as following: i) a usability study to get users' requirements which are stated in Chapter 3: Structured and Qualitative Studies on Genomic Visual Analytics. ii) a prototype implementation of my research which stated in Chapter 4: Research Design, and iii) case studies on the two datasets which stated in Chapter 5: Case Studies.

1.6 THESIS OUTLINE

The thesis includes Chapter 2: Literature Review; Chapter 3: Structured and Qualitative Studies on Genomic Visual Analytics ; Chapter 4: Research Design; Chapter 5: Case Studies; and Chapter 6: Discussion, Conclusions and Future Work.

Chapter 2: Literature Review describes the related work in the genomic data visualisation field and the existing research that relevant to this thesis. It focuses on the genomic visualisation, intelligent visualisation, artificial intelligence, visualisation methods and trends for genomic data, and the research implications.

Chapter 3: Structured and Qualitative Studies on Genomic Visual Analytics describes a structured review and a preliminary study from the usability study which is designed for the feedback analysis. The purpose of this chapter is to collect the requirements from both the existed tools and the end users who are clinicians or researchers.

Chapter 4: Research Design describes the design process from methodology, how to choose algorithms and selected tools for developing demos and applications.

The research design process is based on the lessons obtained through the study in Chapter 3.

Chapter 5: Results describes two case studies based on two datasets which are RMS and ALL patients' genomic data. The two case studies use the prototype system to execute the requirements that are collected in the usability study in Chapter 3. The design methodology and models in Chapter 4 are also applied in the two case studies.

Chapter 6: Discussion, Conclusions and Future Work describes the discussion, research analysis and the future work arising from these studies. It also includes the publication that this thesis contributed.

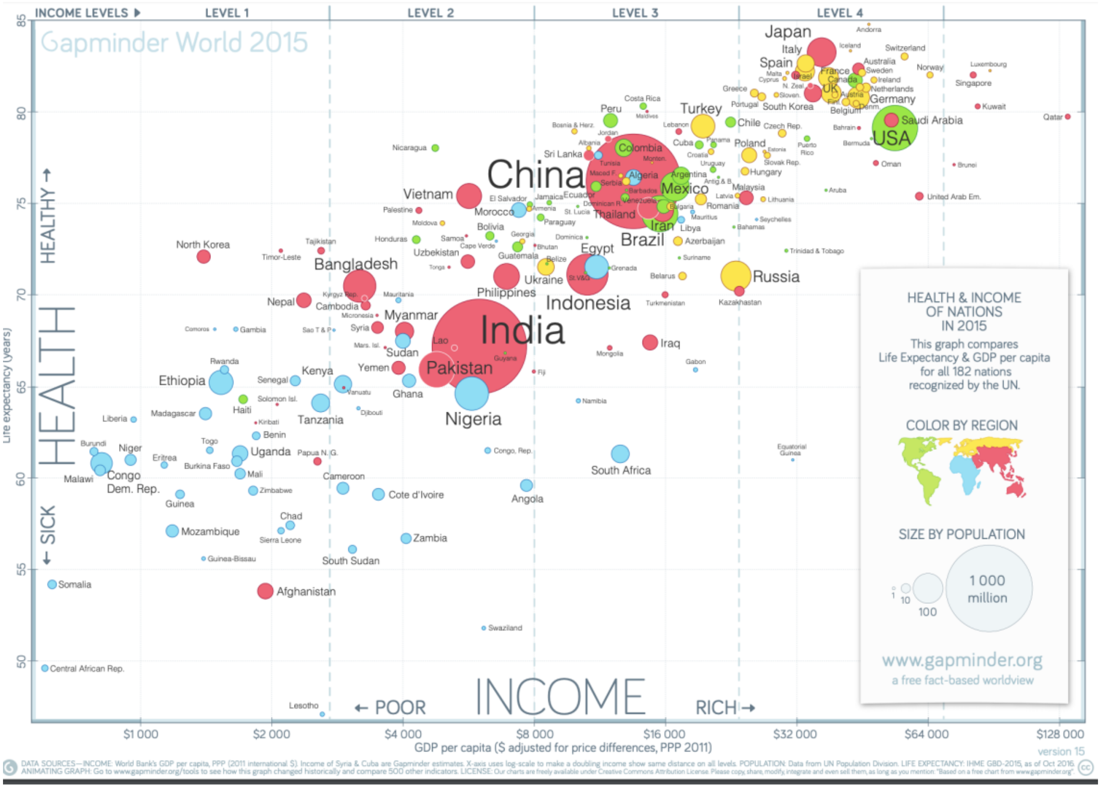
Chapter 2: Literature Review

This chapter begins with the Conception of Visualisation (Section 2.1); The Conception of Machine Learning (Section 2.2) and reviews literature on the following topics: [topic 1] (Section 2.3) [Intelligent Visualisation]; [topic 2] (Section 2.4) [Methods of Genomic Data Analytics and Visualisation]; and [topic 3] (Section 2.5) [Trends of Genomic Data Analytics and Visualisation]. Section 2.6 highlights the summary and implications from the literature and develops the conceptual framework for the study.

2.1 VISUALISATION

Visualisation is an essential tool for the human to understand information and uncover insights hidden in their data. The human short-term memory is capable of holding 3 – 7 items in place simultaneously, which means that people can only juggle a few items in their head before they start to lose track of them. Visual process interprets data into visual channels which externalizes the data and enables people to think about and manipulate the data at a higher level. The human visual system is by far the richest, most immediate, highest bandwidth pipeline into the human mind, which is estimated to process about nine megabits of information per second, that corresponds to close to one million letters of text per second (Keahey, 2013). Visualisation is designed to maximise the complementary cognitive strengths of both humans and computers. Humans have perceptual abilities, earliest reasoning skills such as adaptation and accommodation while the computer has superior working memory and can process information without cognitive biases. Human cognition model is used in visual analytics to create and analyse hypotheses which are initiated by the human, but the computer plays a significant role in shortening the process and neutralising biases, as contributing to a more solid conclusion through use of its strengths (Green et al., 2008).

Visual analytics usually use interactive visual interfaces to engage the fast-visual circuitry of the human brain to quickly find relations in complex data, trigger creative thoughts, and use these elements to steer the underlying computational analysis process which can extract new information for further insight (Garg, Nam, Ramakrishnan, & Mueller, 2008). For example, the visual representation of quantitative data makes possible to obtain a fast understanding of the displayed reality. It may help to the patterns systematisation of complex relationship among various data types (Ortega & Aguillo, 2013). Figure 7 shows the Life Expectancy and Income of 182 nations in the year 2015 (Gapminder, 2015). Each bubble indicates a country. Size indicates the population. Colour indicates region or continents. It's clear in this chart that India and China as Asia countries have more population and medium health level, most of the countries in Africa have less population and lower level health, and most countries in Europe and North America have less population but very high health level.



DATA SOURCES—INCOME: World Bank's GDP per capita, PPP (2011 international $). Income of Syria & Cuba are Gapminder estimates. X-axis uses log-scale to make a doubling income show same distance on all levels. POPULATION: Data from UN Population Division. LIFE EXPECTANCY: IHME GBD-2015, as of Oct 2016. ANIMATING GRAPH: Go to www.gapminder.org/tools to see how this graph changed historically and compare 500 other indicators. LICENSE: Our charts are freely available under Creative Commons Attribution License. Please copy, share, modify, integrate and even sell them, as long as you mention: "Based on a free chart from www.gapminder.org".

Figure 7 the Life Expectancy and Income of 182 nations in the year 2015.

(Gapminder, 2015).

2.1.1 Usability and Interactive Dynamics

Universal usability for visualisation remains a formidable challenge as we need to address the needs of the different user who might have different network speed, different screen size, or different general knowledge. Designers need to choose rapid and high-resolution colour displays to present and manipulate big data in compact and user-controlled ways. Visualisation is used to provide compact graphical presentations and user interfaces for interactively manipulating a large number of data items, which usually are extracted from far larger datasets. It also uses the enormous visual bandwidth and the remarkable human visual system to drive users to find potential information, make decisions, or propose explanations for patterns, groups of items, or individual items. Perceptual psychologists, statisticians, and graphic designers provide valuable advice about presenting data information, but user-interface designers still have challenges in processor speed, graphics devices and dynamic displays (Catherine Plaisant, 2005). Usability has three important criticisms: focus on well-defined tasks and goals, emphasis on efficient and effective, and satisfaction. Dimensions of usability in defining the conversation and driving the process are defined include learnability, efficiency, memorability, error tolerant and satisfaction (Quesenbery, 2003). Figure 8 shows a design process with consideration of usability from the very beginning of a project.

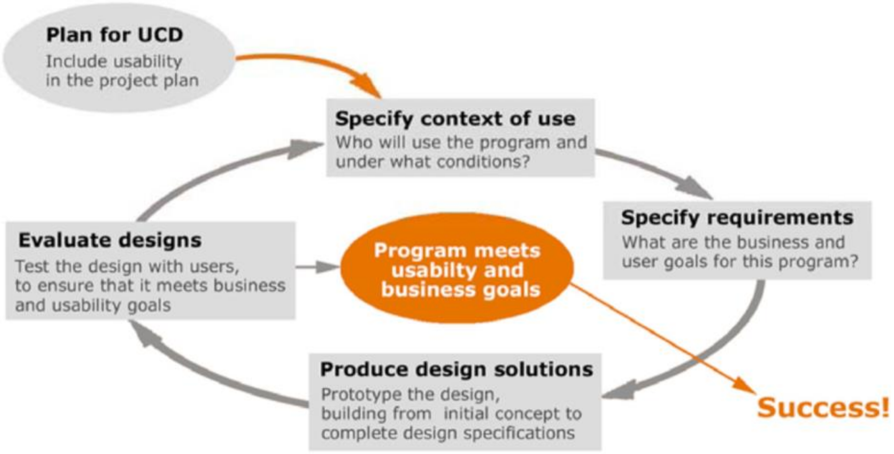

Plan for UCD
Include usability in the project plan
Specify context of use
Who will use the program and under what conditions?
Specify requirements
What are the business and user goals for this program?
Evaluate designs
Test the design with users, to ensure that it meets business and usability goals
Program meets usabilty and business goals
Produce design solutions
Prototype the design, building from initial concept to complete design specifications
Success!

Figure 8 User-centred design (UCD).

This process is often called user-centred design (UCD) and comes with its own research tradition and international standards (Larry Goldberg, Trisha O’Connell, & Ben Shneiderman, 2011).

Interactive dynamics is also very important for visual analysis because a single image is not enough to provide a powerful means of making sense of data. Effective visual analytics tools must support the fluent and flexible use of visualisations at rates resonant with the pace of human thought and with basic interactive actions: filter, sort, derive and view manipulation (Jeffrey Heer, 2012). For example, Figure 9 shows a map which can be zoomed, filtered and scrolled. Real-time interactivity is very essential for showing meaningful dataset in a visualisation method.

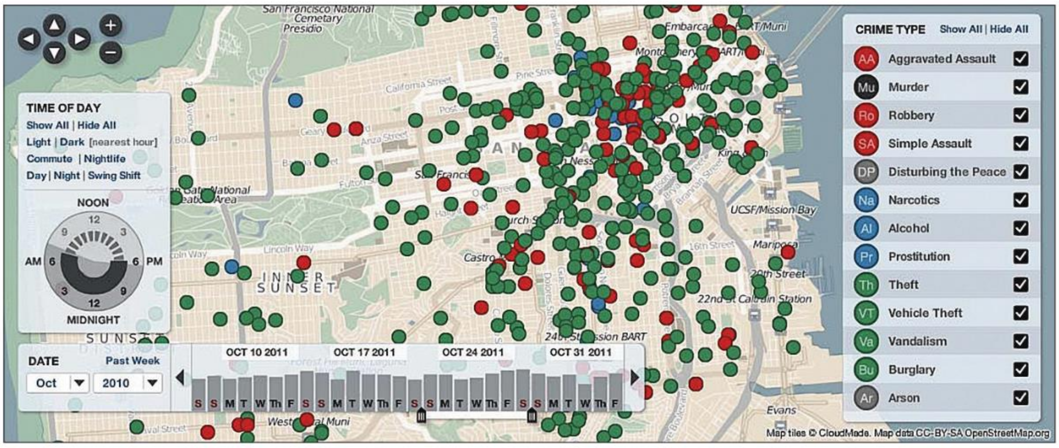


Figure 9 Zoomable map from crimeSpotting.org.

(Jeffrey Heer, 2012).

2.1.2 Multidimensional Data Visualisation

High dimensionality is one of the major challenges for data visualisation because parameter optimisation problems require an understanding of the behaviour of the objective function in the *n*-dimensional space around the optimum and it is not an easy process to convert the high-dimensional data to low-dimensional geometry for display. Understanding the relationship between attributes in the large datasets is essential to extract information subject to constraints on their position or value of dimensional choice for display (Selan dos Santos, 2004).

The challenge of mapping high-dimensional data to lower-dimensional visual representations for large complex information is to find an insightful mapping method. The methods focus on the goals of generating representations that best show phenomena contained in the high-dimensional data like clusters and global or local correlations. Scatterplots and Parallel Coordinates are both commonly used visualisation technique to deal with multivariate datasets (Tatu et al., 2009).

Figure 10 is an example of scatterplots which used to visually inspect the clustering of individuals breeds to their assigned breeds. The first two eigenvectors of the genomic relationship matrix of all individuals in the dataset were used to visually inspect the clustering of individuals according to their assigned breeds. The eigenvector clustering shows significant overlap between breeds (e.g. Hanoverians and Trakehner), while sub-clusters within breeds were apparent (Claas Heuer, 2016).

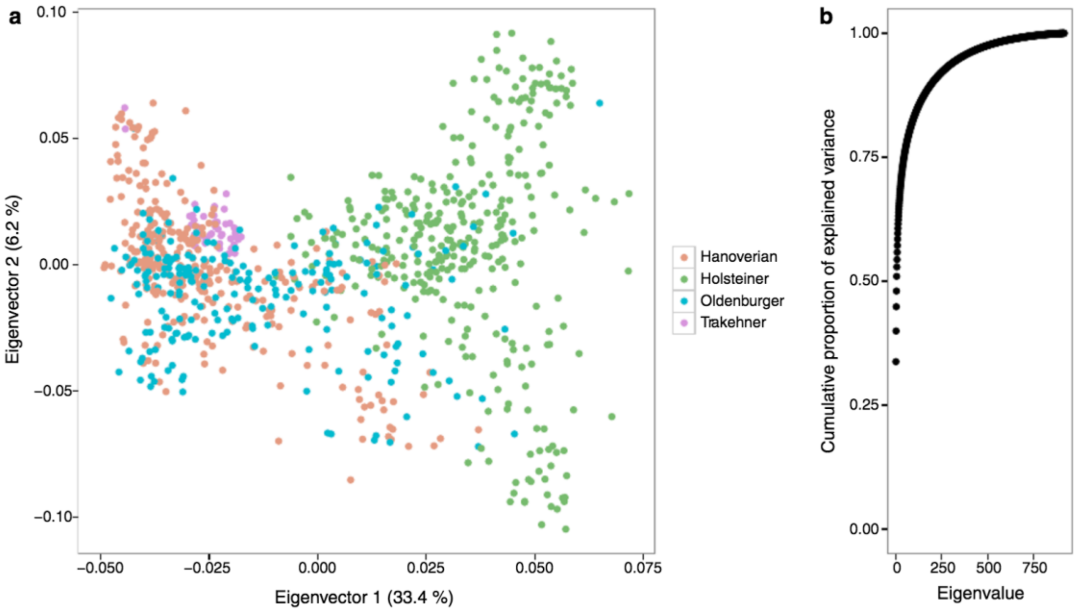


Figure 10 Scatterplot.

Scatterplot of the first two eigenvectors of the genomic relationship matrix (a) and cumulative proportion of explained variance by eigenvalues in decreasing order (b) (Claas Heuer, 2016).

Figure 11 is an example of Parallel Coordinates which shows the coordinate view of all cells and nine selected genes. It extrudes the coordinate axes into the third dimension and order the data lines which represents one cell-back-to-front according to their position along the AP- or DV axis of the embryo. Spatial and gene expression information are clearly separated while the basic character of spatial gene expression patterns is preserved in one dimension (Viswebmaster, 2009).

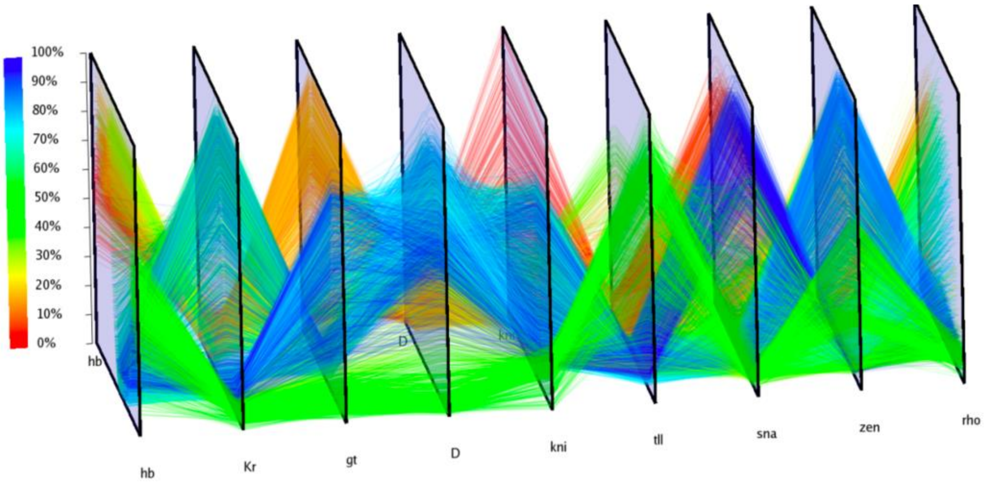


Figure 11 3D parallel coordinate.

It is used to view of all cells and nine selected genes (Viswebmaster, 2009)

Another method to visualise high dimensional data is to visualise the data in a 3-D space by dividing the high dimension data into several groups of lower dimensional data first. Then using different icons to represent the different sets of data in the form, such as line, point, polygon, etc. (R. Jayabrabu, 2012).

3-D visualisation creatively uses colour, size, the combination of space and time, and advanced computer graphics to show multidimensional data. For instance, neuroscientists Emmanuelle Tognoli and Scott Kelso developed a five-dimensional model known as the 5-D colourimetric technique, that provides a dynamic and comprehensive view of brain activity the through spatiotemporal display and colour coding. Another example is Microsoft's Holograph, an interactive 3-D platform (Figure 12) that can render static and dynamic images above or below a plane for more natural exploration and manipulation of complex data. And commentary from team members Curtis Wong and David Brown posted on Microsoft News suggests that Holograph may one day allow users to actually reach inside a visual and interact with it (Towler, 2015).

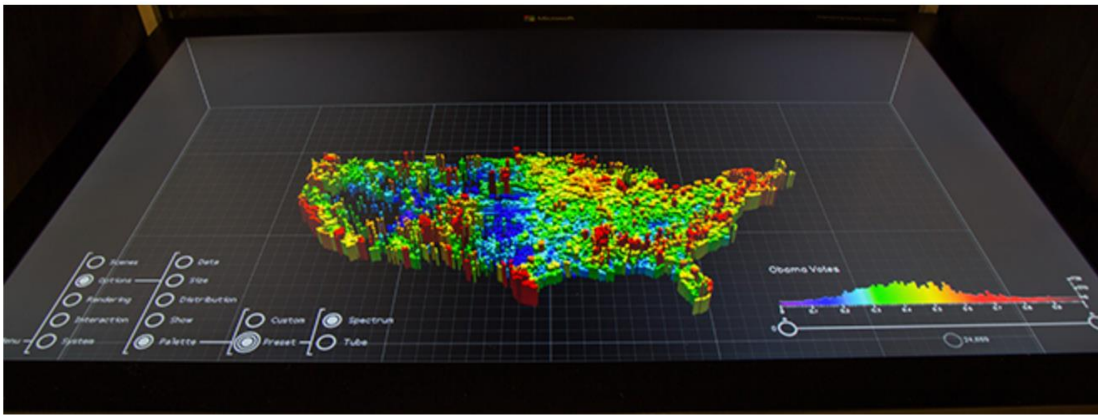


Figure 12 Microsoft's Holograph, an interactive 3-D platform.

(Towler, 2015)

2.2 ARTIFICIAL INTELLIGENCE

2.2.1 AI and Machine Learning

Artificial Intelligence (AI) is a term of cognitive technologies and a big forest of academic and commercial work around the science and engineering intelligent machines. AI has many branches with many significant connections and commonalities among them. Figure 13 shows the most active AI branches (Mills, 2016). AI has been used in many industries and got more and more achievements. The most dramatic outcome of Artificial Intelligence research is Google's AlphaGo program which won the world's best Go player Ke Jie three-match series (Russel, 2017).

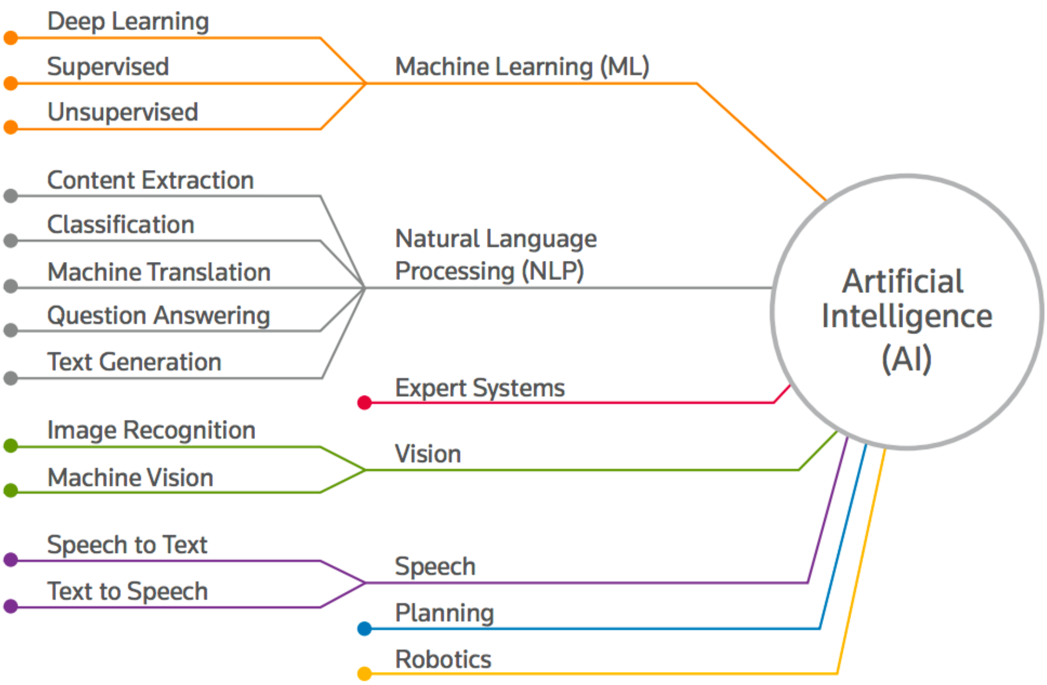


Figure 13 Hierarchies in AI research.

(Mills, 2016)

Machine learning has broad potential across industries and uses cases as shown in Figure 14. McKinsey identified 120 potential use cases of machine learning in 12 industries and surveyed more than 600 industry experts on their potential impact. They found an extraordinary breadth of potential applications for machine learning. Each of the use cases was identified as being one of the top three in an industry by at least one expert in that industry. McKinsey plotted the top 120 use cases below, with the y-axis shows the volume of available data (encompassing its breadth and frequency), while the x-axis shows the potential impact, based on surveys of more than 600 industry experts. Size of the bubble indicates the variety of data (number of data types), and the colour of the bubble indicates different industries (Columbus, 2017). For the Healthcare industry, predicting personalised health outcome and diagnosing diseases are both in the higher potential areas.

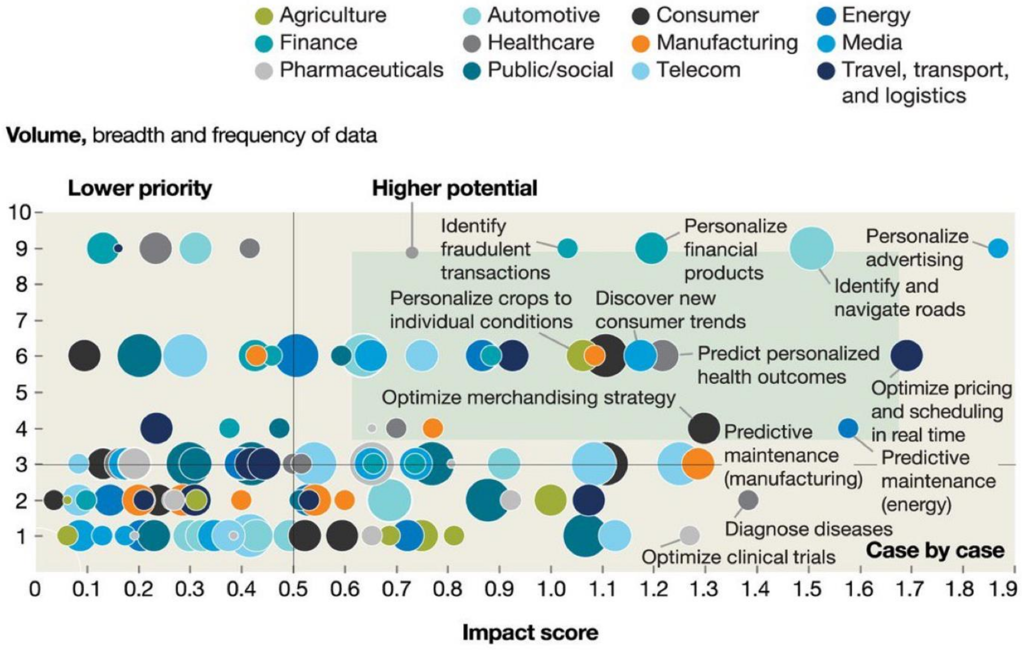


Figure 14 Machine Learning potential industries.

The y-axis shows the volume of available data (encompassing its breadth and frequency), while the x-axis shows the potential impact, based on surveys of more than 600 industry experts. Size of the bubble indicates variety of data (number of data types), and the colour of the bubble indicates different industries (James Manyika, 2017).

2.2.2 AI Use in Medical Clinical Practice

AI started to be used in medical clinical practice from the 1980s and in March 2000, a monthly magazine titled Medical Device & Diagnostic Industry published an article claiming that “the medical device industry is seeing an emergence of computer-based intelligent decision support system (DSSs) and expert system, the current system, the current success of which reflects a maturation of artificial intelligence (AI) technology.” Which mentioned several AI-infused devices such as Agilent Acute Cardiac Ischemia Time-Insensitive Predictive Instrument, Intelligent electro-cardiagram (ECG) device that predicts the probability of acute cardiac ischemia (ACI) and General Electric MAC 5000 Resting Test System. Dr. Paul Kligfield, Division of Cardiology at Cornell University stated: “Digital electrocardiographs of all major manufacturers now are capable of providing automated diagnostic statements that can help the physician.” (Nilsson, 2009).

As shown in Figure 13, machine learning is one branch of the field of artificial intelligence, a way of solving problems without explicitly codifying the solution and a way of building systems that improve themselves over time. The machine learning goal is typically to build predictive or descriptive models from characteristic features of a dataset and then use those features to draw conclusions from other similar datasets. For example, in cancer detection, diagnosis, and management, machine learning helps identify significant factors in high-dimensional datasets of genomic, proteomic, chemical or clinical data that can be used to understand of predicate underlying diseases, in addition to providing possible insights into effective disease management strategies. Machine learning classifiers do best when the number of dimensions is small (less than 100) and the number of data points is larger (greater than 1000). A most significant challenge in the application of machine learning to biological data is the problem of validation, or the task of determining the expected error rate from classifier when applied to a new dataset. 10-folder cross-validation and 10- folder validation idea is used as validation techniques (McCarthy et al., 2004).

Machine learning algorithms tend to create the non-linear, non-monotonic, non-polynomial, and even non-continuous functions that approximate the relationship between independent and dependent variables in a dataset. Some industries, such as serious legal mandate in the regulated verticals of banking, insurance, and medicine, need trusting machine learning models. Many organisations and individuals start to embrace machine learning algorithms for predictive modelling task, but it is still a challenge to the widespread practical use of data interpretation. A unique conundrum of banking, insurance, and other similar industries is to find ways to make more and more accurate predictions, but keep their models and modelling process transparent and interpretable (Patrick Hall, 2017).

2.2.3 Traditional Machine Learning and Modern Machine Learning Process

Traditional analytical lifecycle process can be augmented with machine learning techniques leading to potentially more accurate predictions from regulator-approved linear, monotonic models (Patrick Hall, 2017). Figure 15 outlines three possible scenarios in which analytical processes can be augmented with machine learning: introduce complex predictors into traditional, linear models; use multiple gated linear models, and predict linear model degradation.

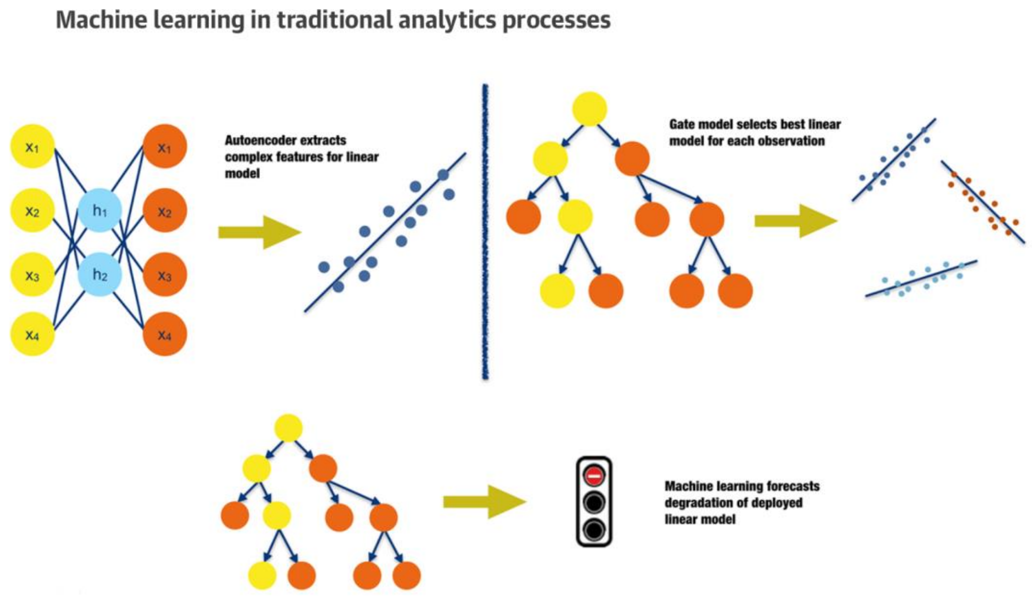

Machine learning in traditional analytics processes
x1
x2
x3
x4
h1
h2
x1
x2
x3
x4
Autoencoder extracts complex features for linear model
Gate model selects best linear model for each observation
Machine learning forecasts degradation of deployed linear model

Figure 15 Diagrams of several potential uses for machine learning in traditional analytical process.

(Patrick Hall, 2017)

Figure 16 is an illustration of cross-validated predictions from two decision trees and a linear regression being combined by another decision tree in a stacked ensemble. It is a more rigorous way to combine model predictions. This model incorporates machine learning models into traditional analytical processes in order to use linear, understandable models more efficiently and accurately (Patrick Hall, 2017).

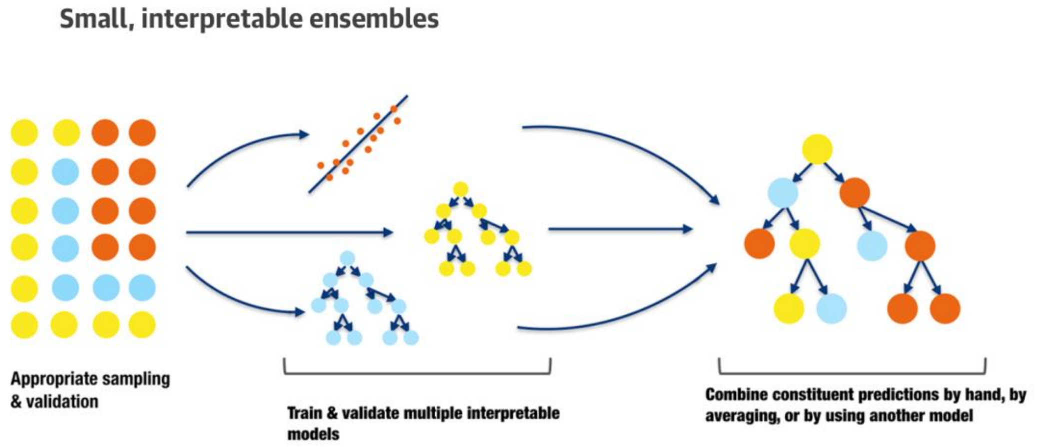

Small, interpretable ensembles
Appropriate sampling & validation
Train & validate multiple interpretable models
Combine constituent predictions by hand, by averaging, or by using another model

Figure 16 A diagram of a small, stacked ensemble.

(Patrick Hall, 2017).

2.2.4 Predicting Genetic Diseases with Decision Tree Model

Decision tree is a predictive model that uses a set of binary rules to calculate a targeted value which can be used for classification, and simple hierarchical structure to identify the class that belong objects from some descriptive traits. Decision tree can be utilised in a wide range of human activities and particularly in automated decision making (Badr Hssina, 2014). In the medical field, there are many applications of decision tree models which can aid in the diagnosis and identification of treatment protocols such as CART, random forest models. In molecular biology, decision trees are used to analyse amino acid sequences in the human genome project because it is simple to understand and interpret, and can help determine worst, best and expected values for different scenarios as well (kane, 2015). Figure 17 is an example of a decision tree model for a map. Training data is used to build the model. The tree generator needs to determine which variable to split at a node and the value of the split, make a terminal note, and assign terminal nodes to a class.

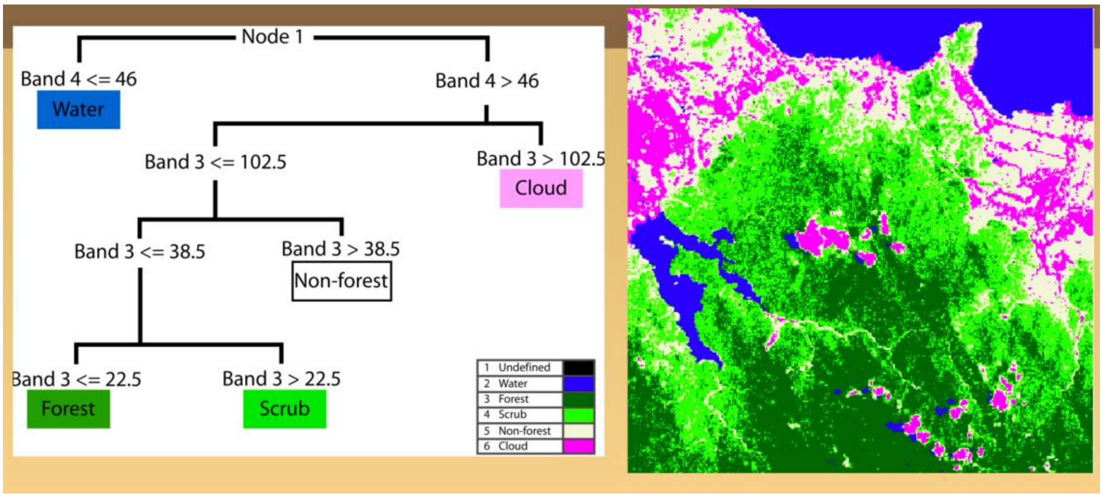


Figure 17 Example of decision tree.

(Horning, 2015)

Random forest is one of the decision tree models which uses many decision tree models to classify or regress data. At Mendelics in Brazil, random forests have been used for almost three years in the field of genetics (Mario, 2016). The human genome is over 3 billion nucleotides long and every person has thousands of mutations. Unfortunately, some of these mutations, instead of changing the colour of your eyes, cause diseases. It is not straightforward to determine which mutations cause diseases in the middle of this sea of mutations. Researchers in genetics have used machine learning models built with random forests to solve this problem (Mario, 2016).

2.3 INTELLIGENT VISUALISATION

2.3.1 What is Intelligent Visualisation and Why It Matters?

The objective of intelligent visualisation can be explained as “give everybody the right information at the right time and in the right way” which includes two aspects: the first refers to the problem of selecting the relevant information, depending on the situation and the needs, goals, and characteristics of the user; another aspect is that the information should be presented in a way promoting its rapid perception, proper understanding, and effective use which means effective preparation, organisation, and representation of the information (Natalia Andrienko, 2007).

The digital universe is doubling in size every two years (Oracle, 2015), and the amount of data that crosses the Internet every second is greater than all the data stored in the Internet just 20 years ago, which amounts to exabytes of data being created on a daily basis (Polsky, 2017). It is impossible for the human brain to process more than one value at a time let alone hundreds, thousands, millions or billions. Visualisations is the single easiest way for our brains to receive and interpret a large amount of information. Data visualisation represents data in a pictorial or graphical format which can simplify data values, promote the understanding of them, and communicate important concepts and ideas (Polsky, 2017). Advanced data visualisations not only support more in-deep and complex analytics to get insight into what is happened, but also can forecast what might happen with machine learning algorithms.

Figure 18 shows a traditional electronic spreadsheet which limits what you can see, but data visualisation can easily interpret, saving time and energy. As we can see from the left-side image, the spreadsheet cannot display a lot of data in the table and it extremely perceives a large number of numerical numbers. The right image uses charts and heatmaps illustrates the number of units that correspond to each age (represented by the colour gradient) as well as the reliability as the age of a unite increase. In a matter of seconds, we can see the units approaching 20 years of age are approximately 40 percent reliable. This visual simplifies the totality of the data, instantly clarifying what is happening with the reliability of the cell phone motors values, trends and the property of the information in a much better and clear way.

Figure 18 Traditional spreadsheet and visualisation.

A traditional spreadsheet limits what you can see, meanwhile, the right visualisation simplified the totality of the data and easy to get information instantly (Polsky, 2017)

2.3.2 Background on Intelligent Visualisation

The history of statistical graphics and data visualisation started from the earliest map-making and visual depiction, and then thematic cartography, statistics and statistical graphics, with applications and innovations in many fields of medicine and science that are often intertwined with each other (Friendly, 2006). Using pictures to understand data has been around for centuries. Maps and graphs started in the 17th century, then the pie chart was invented in the early 1800s, and then several decades later, one of the most cited examples of statistical graphics happened when Charles Minard mapped Napoleon’s invasion of Russia. The statistical graphics depicted the size of the army as well as the path of Napoleon’s retreat from Moscow, and tied that information to temperature and timescales for a more in-depth understanding of the event (SAS, 2017).

Technology speeds up data visualisation and lets it become a rapidly evolving blend of science and art as computers made it possible to process a large amount of data at lightning-fast speeds. There is no “one-size-fits-all” technique to visual data because every task and dataset has its own unique properties. Visualisation systems can integrate a large number of intelligent algorithms to automatically compose or recommend effective visualisation given a user’s task context. There are three categories in existing systems: task-based systems which use formal task descriptions as input to construct appropriate visual presentation; data property-based systems which focus on the dataset being visualised and using features of the data itself as input to the visualisation recommendation or composition algorithm; and hybrid systems which use a combination of both data properties and explicit representation of user intent to determine a proper visualisation. Behaviour-driven visualisation consists of two distinct phases: i) pattern detection which analyses user behaviour dynamically to find semantically meaningful interaction patterns by using a library of pattern definitions developed through observation of real-world visual analytic activity; and ii) visualisation recommendation which uses intelligent algorithms to detected patterns to infer a user’s intended visual task, and then automatically suggests alternative visualisations that support the inferred visual task more directly than the user’s current visualisation (Gotz & Wen, 2009).

Harvest is a behaviour-driven visualisation intelligent visual analytics system which builds a graph-based representation of interconnected trails to represent the user’s visual exploration behaviour. When users save their work via the bookmark, Harvest preserves both the state of the visualisation as well as the automatically recorded analytic trail. When a bookmark is later restored, the trail is restored as well. This allows a user to review the exploration recommendation, and it quickly led users to proper visualisation for their tasks (Gotz et al., 2010). Figure 19 shows key Harvest technologies: smart visual analytic widgets, dynamic visualisation recommendation, and semantics-based capture of insight provenance.

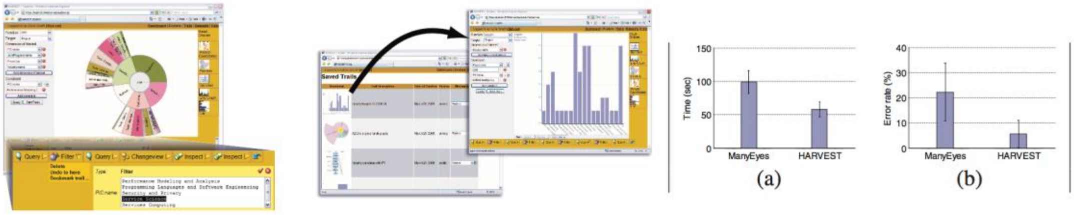


Figure 19 Behaviour-driven visualisation.

The left shows the history panel displays the unfolding analytics trail; the middle shows users can restore saved trails to re-use past analyses; the right shows mean and 95% confidence interval of task completion time and task error (Gotz et al., 2010)

2.3.3 Machine Learning and Intelligent Visualisation

Machine learning methods use statistical learning and computers to make predictions by finding patterns and unearthing boundaries in data (Stephanie, 2015). Machine learning algorithms have been used to assist in the exploration process in the past and now are used to depict various attributes using multiple views and to allow the engineer to interactively select a subset of the data in these views. The interactive visual analysis combined with machine learning algorithms can find hidden relations between multi attributes in the tedious multidimensional dataset (Fuchs et al., 2009). The future of big data visual exploration will involve the tight integration of visualisation tools with traditional techniques from such disciplines as statistics, machine learning, operations research, and simulation. Visual exploration also needs to combine fast automatic data mining algorithms with the intuitive power of the human mind which can improve the quality and speed of the data exploration process (Keim, 2001).

The fields of visualisation and machine learning have been addressing big data analysis from different perspectives and advances in both communities and need to be leveraged and in order to make progress. Machine learning has proposed algorithms and techniques that can process large volumes of data, enabling visualisation to scale, while information visualisation can leverage the human perceptual system to interpret and uncover hidden patterns in these datasets. Visualisation benefits from machine learning in exploratory procedures such as feature selection, dimensionality reduction and clustering (Daniel A. Keim, 2015).

Figure 20 interplays between machine learning and data visualisation. The core dataset (top) stores the information from the data stream which is pre-processed for binary hashing and core sets discovery. Pre-processing enables index-based data retrieval, selection of representative data instances, and fast distance computation. Multi-view visualisation initially displays data. The core set also supports the user in digging deeper and retrieving data from neighbourhood, time, location or concept-specific spaces. Data-related semantic concepts are retrieved from related databases and organized in ontology or network. Visualisations are interlinked: any change in selection in one view updates the information in all other views. Machine learning algorithms for clustering, assessment of concept enrichment, outlier detection and classification of uncharacterised data instances are triggered on the fly. User's interactions are recorded and modelled and provide means of predicting them and executing the most likely data-intensive operations that the user can trigger in the future before they are actually needed. The user can change the attributes or position of data instances in any visualisation, thus visually changing the objective function that is optimized in the visualisations. Change of objective function is followed by repositioning of data elements in the visualisations.

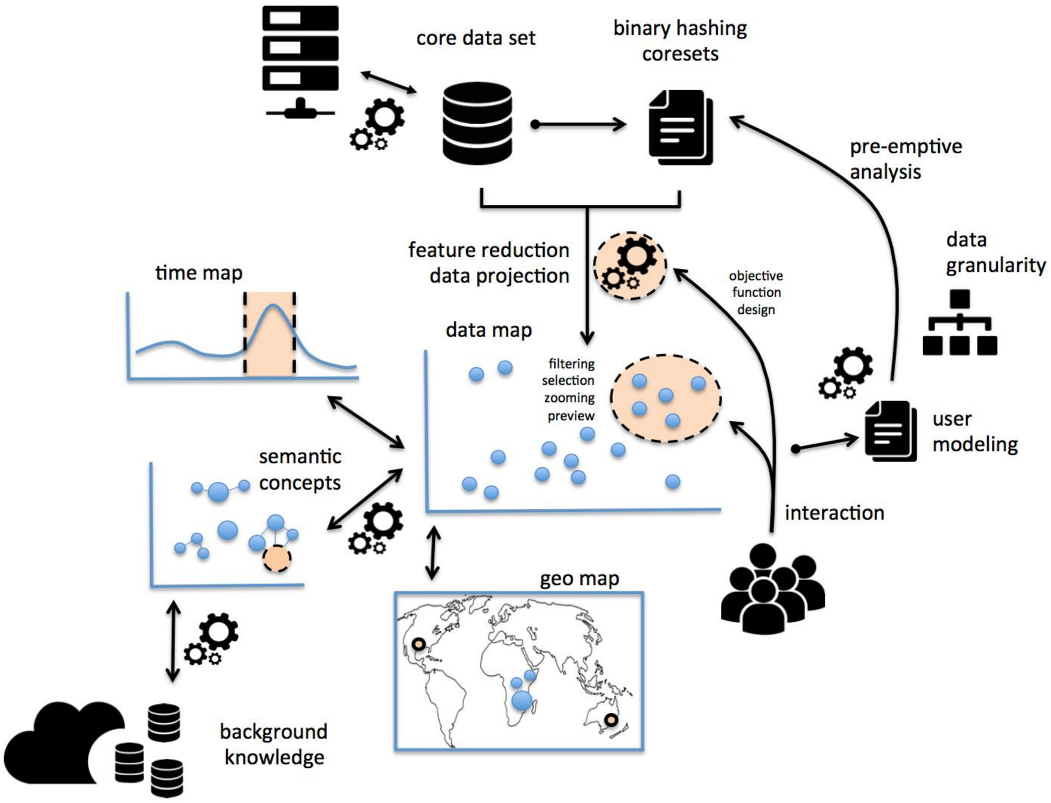


Figure 20 interplay between machine learning and data visualisation.

(Daniel A. Keim, 2015).

Combining visualisation and machine learning is a challenge for big data analysis. Machine learning should not only be used for mining data but also be used to data visualisation to address the variety and span of data sources to improve human perception.

2.4 METHODS OF GENOMIC DATA ANALYTICS AND VISUALISATION

2.4.1 Genome Data Collected and Stored

Entire genome sequences are getting completion which drives biology goes to the midst of an intellectual and experimental sea change. Huge data have been collected from the fortuitous confluence of technological advances in protein and DNA analysis as well as imaging advances in cell biology. In the late 1980s, the international Human Genome Project started to collect human genome sequence which stimulated developments both in high-throughput DNA sequencing, which were essential for the success of the project, and in powerful computational tools for sequence analysis. Figure 21 shows the components of studies in Saccharomyces cerevisiae where more and more genes had been studied from 1996 to 2000. In 1996, it was estimated that 30% known genes and 30% were unrecognizable. The situation is even more striking in multicellular organisms. In 1998, it was reported that only 7% had been studied previously, although 42% of the genes had some match to proteins and sequences of random complementary DNAs (expressed sequence tags). These matches can often be clues to the function of previously unstudied genes. By 2000, the number of completely novel genes with no match to anything previously encountered in DNA sequence was reduced to 17% of the 13,600 Drosophila genes in the fly genome (Vukmirovic & Tilghman, 2000).

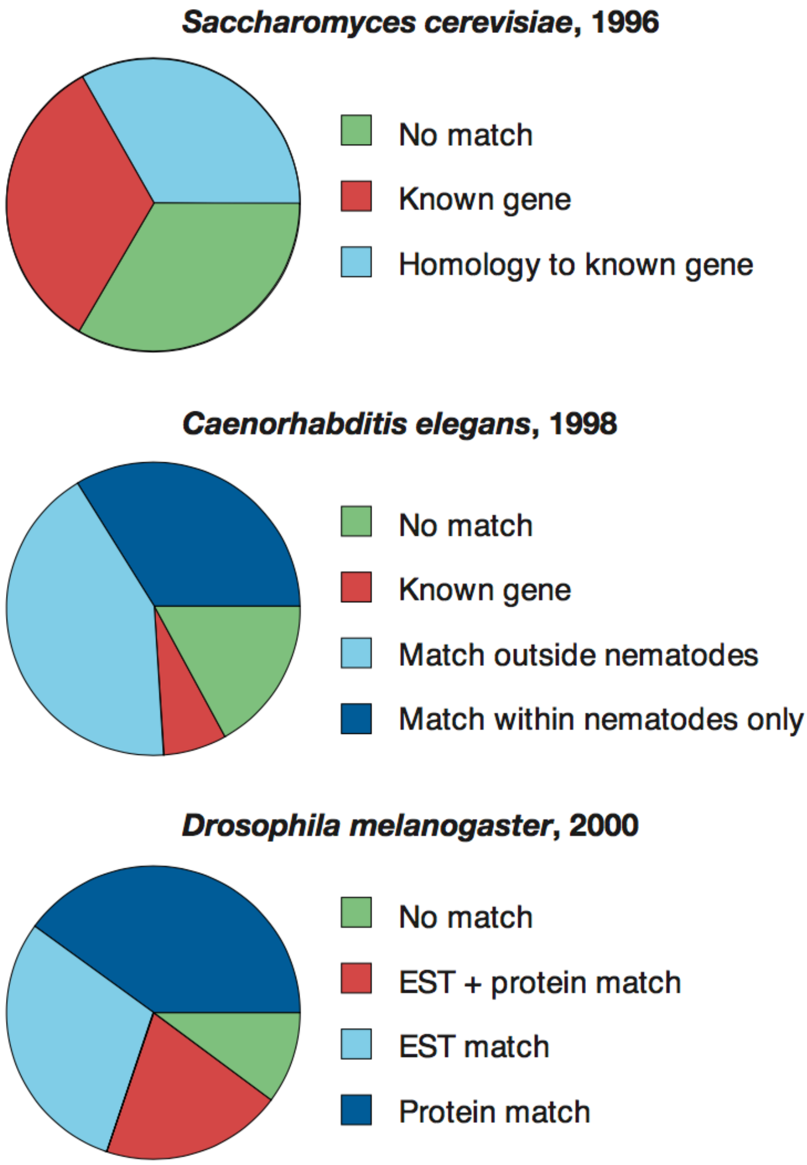
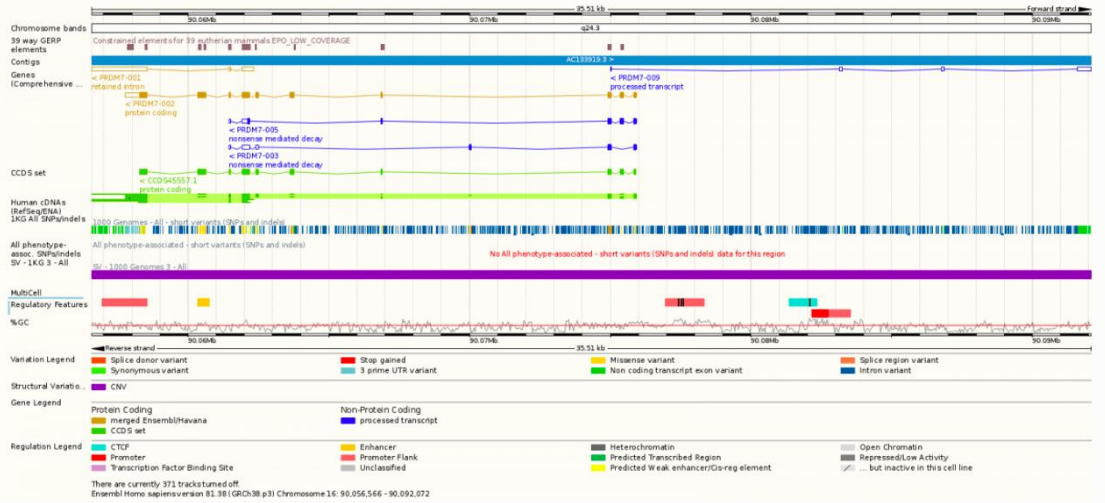


Figure 21 The distribution of genes in eukaryotic genomes.

Shown for three organisms are the relative number of genes that were previously identified, to be known, and that had no match in any sequence database at the time of completion of the genome sequence (Vukmirovic & Tilghman, 2000)

Huge genome data has been collected in the past years and started to be shared by different groups such as scientist, bioinformaticians, clinicians and related researchers with various channels, for example, genome browser. Genome browser, which can make data openly accessible to support and progress scientific research across the globe, is an online graphical interface used to display genomic data. The genomic data is from Ensemble based in European, UCSC based at the University of California Santa Crus, and National Centre for Biotechnology Information base in

Maryland in the USA(Genome, 2016). Figure 22 shows the basic structure of the display on many genome browsers which present the genome sequence horizontally across the screen. Elements in the sequence are presented in specific colours and shapes according to a key.



No All phenotype-associated - short variants (SNPs and indels) data for this region

There are currently 371 tracks turned off.
Ensembl Homo sapiens version 81.38 (GRCh38.p3) Chromosome 16: 90,056,566 - 90,092,072

Figure 22 Screenshot taken from the Ensembl genome browser.

It is showing the visualisation of the genes and other features of interest on human chromosome 16 (Genome, 2016)

2.4.2 Visualisation Tools for Analysing Genomic Data

Massive genomic datasets are generated by different projects which are stored and shared by different groups of professionals. Some basic analysis tools are developed such as Genome Analysis Toolkit and X:Map. Genome Analysis Toolkit (GATK) is a structured programming framework designed to ease the development of efficient and robust analysis tools for next-generation DNA sequencers using the functional programming philosophy of MapReduce (McKenna et al., 2010). X:Map is a tool which designed specifically for high-density microarrays that are required to show for each gene, transcript and exon the probe sets that match it, their specificity and for each probe, their locations of potential hybridization and for each individual exon, its sequence (Yates, Okoniewski, & Miller, 2008). It is essential to use visualisation of multidimensional oncogenomics data to extract useful knowledge from the vast amount of data generated by high-throughput technologies.

By using computational and statistical methodologies, effective visualisation is crucial to successful extraction of knowledge from oncogenomics data for domain experts. High-throughput technologies allow the comparison of the genomic sequences, epigenomics profiles, and transcriptomes of tumour cells with those of normal cells. Visualisation techniques and tools can integrate different type of alterations with clinical experience to show the vast amount of multidimensional oncogenomics data in different types of plots such as heatmaps, genomic coordinates, and networks (Bhojwani et al., 2008; Rebeiz & Posakony, 2004; Schroeder et al., 2013). Figure 23 shows examples of three visualisation methods: matrix heatmaps, genomic coordinates and networks that are frequently used in cancer genomics research. Each of the three visualisation methods - matrix heatmaps (from Gitools), genomic coordinates (from UCSC/Cancer Genetics Browser, IGV and Savant) and networks (from CircleMap, Regulome explorer, Caleydo/StratomeX, and Cytoscape) - is associated with a vertex in the triangle. Tools that are placed at a vertex indicate the main visualisation method; those placed in between the vertices use a mixed-model visualisation method (Albuquerque et al., 2017; Schroeder et al., 2013).

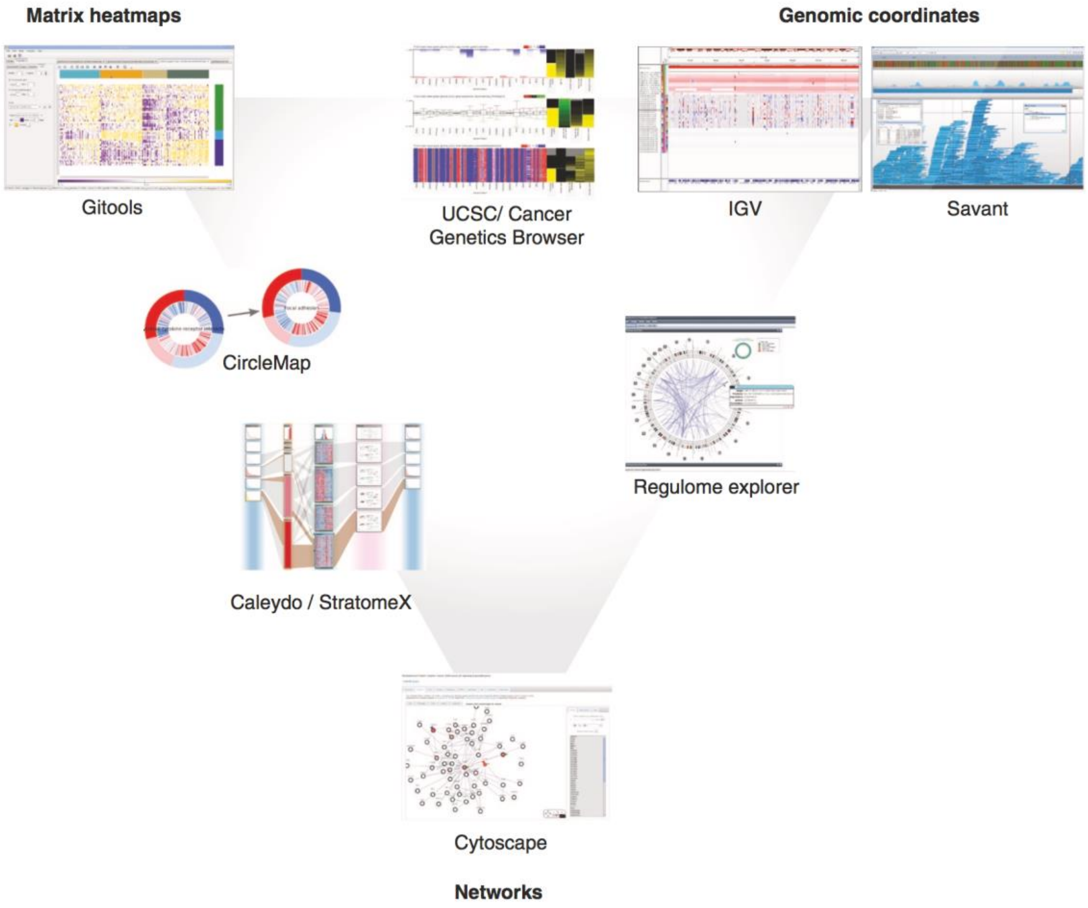


Figure 23 Three visualisation methods.

Matrix heat maps, genomic coordinates and networks. (Schroeder et al., 2013)

With such big accumulated genomic data, the current analysis tools may not be sufficient for genomic data generation, distribution, and visualisation. Personalised medicine presents a unique challenge for new tools which can efficiently extract knowledge from the data, explore the multiple relationships between the data, and speed up expert's decisions. Along with personalised cancer medicine's development, cancer genomics data visualisation in the clinical setting is likely to become a key topic in the near future. Efficient tools, that support the visual stratification of the tumour genomic profiles and that highlight their relationships to know drugs or treatments, will be more useful than the existing research-oriented tools (Schroeder et al., 2013). In order to help downstream analysts to access and manipulate the massive sequencing datasets in a programmatic way, new feature-rich, efficient, and robust analysis tools are developed to process data and answer specific scientific questions (Chittaro, 2006; McKenna et al., 2010). Further efforts are required to develop new tools to meet the new demands and challenges in the field.

2.5 TRENDS OF GENOMIC DATA ANALYTICS AND VISUALISATION

2.5.1 AI, Machine Learning and Big Data in Healthcare Industry

New technology breaks the barriers such as cost, computing power to implement artificial intelligence and big data to the healthcare industry. Nowadays, sequencing of individual genomes and then comparing them to a vast database allow doctors to predict the probability of a particular disease and choose the best ways to treat those diseases when they appear. Google, Apple, Samsung, and other companies are investing billions in developing new biometric sensors. Combined with big data, the information from these sensors could help to prevent disease and extend lifespans (Marr, 2016a).

Computer scientists at Stanford created an artificial intelligent diagnosis algorithm for skin cancer and it performed with inspiring accuracy (Andre Esteva, 2017). They made a database of 129,450 skin disease images and trained their algorithms to visually diagnose potential cancer. For skin cancer, early detection is critical and mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 and can therefore potentially provide low-cost universal access to vital diagnostic care. Neural network, as a deep learning algorithm, allow medical practitioners and patients to proactively track skin lesions and detect cancer earlier. This fast, scalable method is deployable on mobile devices and holds the potential for substantial clinical impact, including broadening the scope of primary care practice and augmenting clinical decision-making for dermatology specialists and other cancer specialists (Andre Esteva, 2017). Figure 24 shows the skin cancer classification technique which shows data flow from left to right. The 757 training classes are defined using a novel taxonomy of skin disease and a partitioning algorithm that maps diseases into training classes (Andre Esteva, 2017).

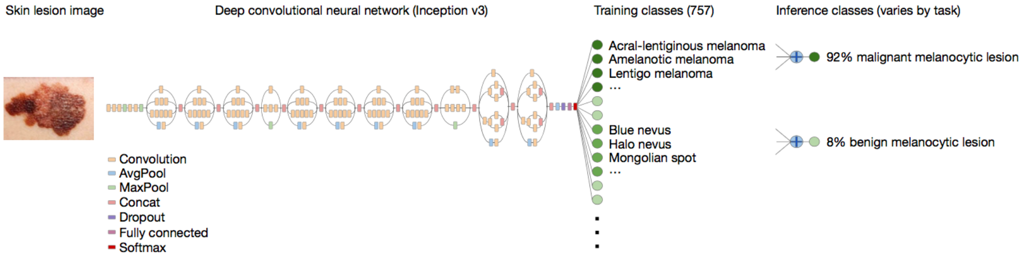


Figure 24 Skin cancer classification technique.

(Andre Esteva, 2017)

2.5.2 Personalised Medicine is Effective and Precise

Mark Zuckerberg said in his presentation in Harvard “How about curing all diseases and getting people involved by asking volunteers to share their health data – track their health data and share their genomes? You know, today our society spends more than 50 times as much treating people who are sick, as we invest in finding cures, so people don’t get sick in the first place.” (Zucerberg, 2017). Personalised medicine presents the unique challenge for new tools which can efficiently extract knowledge from the data, explore the multiple relationships between the data, and speed up expert’s decisions. Personalised medicine is the tailors of medical treatment to the individual characteristics, needs and preferences of each patient. Patients can be treated and monitored more precisely and effectively and in better ways to meet their individual needs. This benefits from the advancement in a wide range of fields from genomics to medical imaging to regenerative medicine, increasing of computational power, and the advent of mobile and wireless capability and other technologies (Cordeiro, 2014; Margaret A. Hamburg, 2013; Savoia et al., 2017; Vogenberg, Isaacson Barash, & Pursel, 2010).

Personal health data are exploding with the increasing number of mobile health applications. Mobile health has grown exponentially over the last decades and it is expected to reach $20.7 billion market worth by 2018, with nearly 96 million users (Juniper, 2018). Thousands of applications and more are being developed are used to collect personal health and lifestyle data, which make personalised health more approachable than ever imagined. Data analytical tools can be used to visualise data from the population level to a more personalised approach, and from a reactive method to proactive methods focus on prevention, wellness, and most importantly –the individual (Boudreaux et al., 2014; Krisa D. Tailor, 2014).

In 2012, a new therapy with the drug Kalydeco for cystic fibrosis (CF), a serious inherited disease that impairs the lungs and digestive system, was approved for patients with a specific genetic mutation- in a gene that is important for regulating the transport of salt and water in the body (Margaret A. Hamburg, 2013). Kalydeco and more other cancer drugs have been approved for use in patients whose tumours have specific genetic characteristics that are identified by a companion diagnostic test which point to the emergence of a new era of personalised medicine (Margaret A. Hamburg, 2013).

More and more knowledge about associations between genomic factors and disease has rapidly accumulated as shown in Figure 25. Genomic analyses have provided new biological insights into the pathogenesis and classification of diseases and determinants of success and failure of therapies. This leads to the development of new analytical approaches that use multidimensional datasets and embrace the complexity of genomic data for personalised medicine (Procter et al., 2010; Sikic, Tibshirani, & Lacayo, 2008).

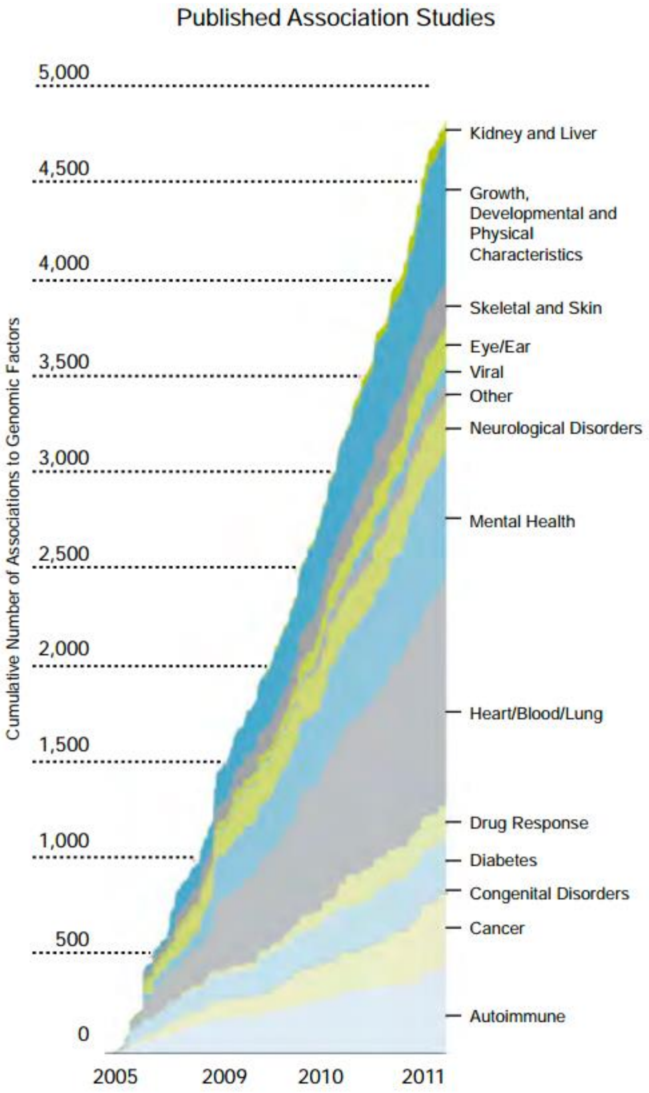


Figure 25 Knowledge about associations between genomic factors and disease has rapidly accumulated.

(Raskin, 2011)

2.5.3 Interactive Visualisation for Genome Comparison

Interactive visualisation of complex genomic data is an effective way to bring the insight of information and discover the relationships, non-trivial structures, and irregularities that may pertain to the disease course of the patient. Basic statistics and visualisations without effective interaction and capabilities to control the visual data mining process are often insufficient for the analysis and exploration process. Intelligent visualisation can focus on patient-to-patient comparisons through the biological data and display the multi-dimensional data in cooperation with the automated analysis (Q. V. Nguyen, Nelmes, Huang, Simoff, & Catchpoole, 2014).

Intelligent genomic visualisation can support experts in the process of hypotheses generation concerning the roles of genes in diseases and find the complex interdependencies between genes by bringing gene expressions into context with pathways (Lex, Streit, Kruijff, & Schmalstieg, 2010).

2.6 SUMMARY AND IMPLICATIONS

Machine learning is a modern way to address important problems in genomic medicine, for example, creating a predictive model to determine how variations in the DNA of individuals can affect the risk of different disease and to find causal explanations so that targeted therapies can be designed (Leung et al., 2016).

Intelligent visualisation is an effective way to analyse large multidimensional genomic data to extract knowledge and intelligently map the data to the most appropriate graphical representation to each community of participants. Combining machine learning algorithms with intelligent visualisation is a challenge to show complex genomic data in a meaningful way and give predictive choices.

In this thesis, we focus on applying machine learning model to genomic data visualisation tool. We started the research with a usability study to collect requirements from the existed tools and from the end users who are researchers and clinicians. We then designed our visualisation tool based on the requirements we collect from the usability. This included choosing machine learning algorithms, designing visualisation techniques with different tools such as R, Unity3D. And last step, we developed a visualisation prototype with two case studies on two childhood genomic cancer datasets. The two case studies use the prototype system to execute the lessons that are collected in the usability study and the design methodology and models in research design.

Chapter 3: Structured and Qualitative Studies on Genomic Visual Analytics

This chapter describes the usability study before the research design that systematically reviews popular genomic data visualisation tools and collects user's preference for a new genomic data visualisation tool. The systematic review contributes to a review paper that has been accepted by the Sage journal Cancer Informatics. The structured review is stated in (Section 3.1). The small preliminary group study for the quallitative review is stated in (Section 3.2). The last part (Section 3.3) discusses why new visualisation tools are needed. The purpose of this usability study is to collect requirements for the research design in Chapter 4 and case studies in Chapter 5.

3.1 STRUCTURED REVIEW

3.1.1 Genomic Data Visualisation Methods Review

Nowadays, new visualisation tools and methods such as cluster analysis, AI, and VR are introduced by different groups of people including designers, software developers and scientists. They try to combine existing visualisation tools with new technological opportunities especially AI and VR to maximise human knowledge and intuition (García-Hernández, Anthes, Wiedemann, & Kranzlmüller, 2016; Golestan Hashemi et al., 2017; Olshannikova, Ometov, Koucheryavy, & Olsson, 2015). Figure 26 shows genomic visualisation methods used in recent years, including scatter plots, cluster, matrix heatmaps, genomic coordinates, networks, AI and VR from screenshots of tools that are frequently used in cancer genomics research distributed according to their visualisation principles. Scatter plots, network, heatmap, and coordinates are four traditional methods for visualising genomic data which have been used in most popular visual analytics tools. Clustering methods are used to support all the above visualisation methods to enhance the classification. AI algorithms support visualisation by automatically identify patterns and making highly accurate prediction while visualisation methods can aid or interpret AI by framing predictive modelling problem and evaluating model. VR, AR, Immersive, and mobile devices are the new environments for data visualisation to make the interactions with data in a more natural or easier way. We list all the visualisation methods and their descriptions in Table 1. We now explain and evaluate each visualisation method with example tools in the following paragraphs. We also analyse the combinations between these methods and how to use them in research and clinical fields.

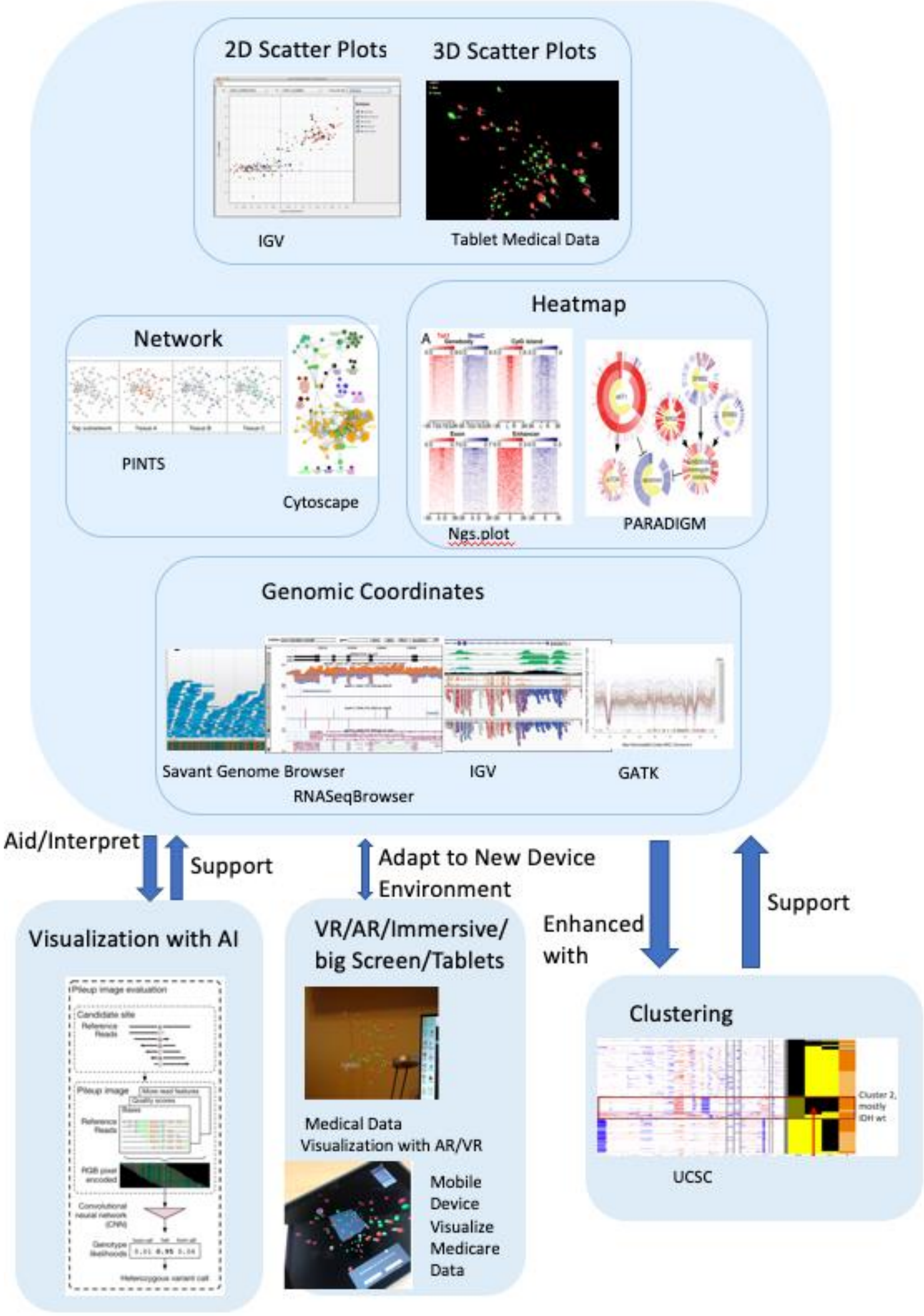


Figure 26 Genomic Data Visualisation Methods: Scatter plots, Cluster, Heatmap, Network, Genomic Coordinates, AI and VR for visualisation.

Methods	Description	Example Visualisation Tools
2D Scatter Plot	The scatter diagram graphs pairs of numerical data, with one variable on each axis, to look for a relationship between them. If the variables are correlated, the points will fall along a line or curve. The better the correlation, the tighter the points will hug the line (ASQ, 2018).	IGV (IGV, 2018), UCSC (Mary Goldman, 2017)
3D Scatter Plot	3D scatter plots are used to plot data points on three axes in the attempt to show the relationship between three variables. Each row in the data table is represented by a marker whose position depends on its values in the columns set on the X, Y, and Z axes. A fourth variable can be set to correspond to the colour or size of the markers, thus adding yet another dimension to the plot (Tibco, 2018).	Medical Data Visualisation(Q. V. Nguyen et al., 2014).
Heatmap	A heatmap is a graphical representation of data that uses a system of colour-coding to represent different values. A common method of visualising gene expression data is to display it as a heatmap. In heatmaps, the data is displayed in a grid where each row represents a gene and each column represents a sample. The colour and intensity of the boxes are used to represent changes in gene expression (EMBL-EBI, 2018).	Ngs.plot (Shen, Shao, Liu, & Nestler, 2014), Gitools (Perez-Llamas & Lopez-Bigas, 2011), PARADIGM (Vaske et al., 2010).
Clustering	A cluster is a group of similar elements. Each cluster can be represented by a profile, either a summary measure such as a cluster means or one of the elements itself, which is called a medoid or centroid (K. S. Pollard 2003).	Medical Data Visualisation(Q. V. Nguyen et al., 2014), UCSC (M. Goldman et al., 2015).
Network	A network graph uses information from both the link and node data sets to generate a graphical depiction of the network. The nodes and links in a network graph can be arranged in a variety of layout patterns(SAS, 2018).	Cytoscape (Shannon et al., 2003).
Genomic Coordinate	Genomic Coordinate can visualise single-nucleotide polymorphism(SNP) including their physical location relative to their host gene, and the structure of the relevant transcripts to provide intuitive supplements to the understanding of their functions (Zhang et al., 2015).	UCSC (M. Goldman et al., 2015), IGV (IGV, 2018), RNASeqBrowser (An et al., 2015),GATK (McKenna et al., 2010), Savant Genome (M. Fiume, Williams, Brook, & Brudno, 2010).
AI (Artificial Intelligence)	Artificial Intelligence (AI) is a term of cognitive technologies and a big forest of academic and commercial work around the science and engineering intelligent machines. AI has many branches with many significant connections and commonalities among them, Machine Learning is one of the branches (Mills, 2016).	DeepVariant (Google, 2017), GDC DAVE (NIH, 2017b).

VR	VR is by immersing the user in a digitally created space with a 360-degree field of vision and simulated movement in three dimensions, it should be possible to greatly increase the bandwidth of data available to our brains (Marr, 2016b).	UWS Microsoft HoloLens Visualisation (Lau, Nguyen, Qu, Simoff, & Catchpoole, 2019)

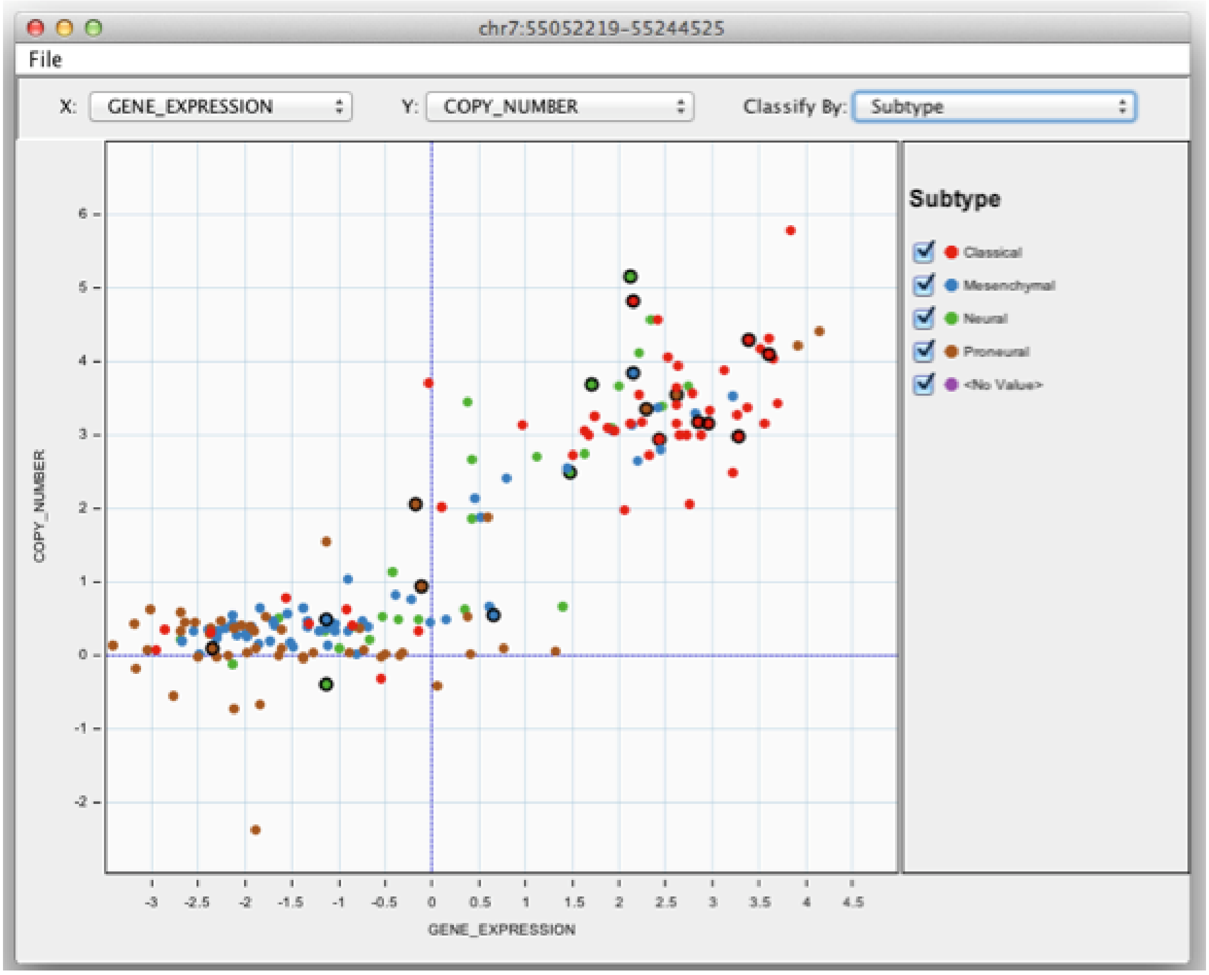
Table 1 The list of visualisation methods, their descriptions and examples

Combine Different Visualisation Methods

Researchers and doctors usually combine different visualisation methods in a typical analysis procedure to assist their work. For example, they need first to normalise experimental and batch differences between samples and then to identify up and down regulated genes based on a fold-change level when comparing across samples, such as between a healthy and a non-healthy tissue. In this procedure, principal component analysis or partitioned clustering algorithms (Ciaramella et al., 2008; Pollard & van der Laan, 2005) can be used to group together genes with similar behaviour patterns, scatter-plotting is the typical visualisation to represent such groupings. Then, categorising genes with similar behaviour patterns across time, hierarchical clustering based on expression correlation can be performed with clustering heatmaps which can allow data from distant genome loci to be grouped and visualised together for comparison (Eisen, Spellman, Brown, & Botstein, 1998; Huang da, Sherman, & Lempicki, 2009). As shown in Figure 26, scatter plots, network, heatmaps, genomic coordinates are the four classic genomic and cancer data visualisation methods and the clustering method can be used to enhance the above four methods. AI is a new technology and has started supporting traditional visualisation methods recently. Genomic and cancer visualisations have increasingly supported VR, AR, Immersive big screen and mobile devices to enlarge human’s perception (Matte-Tailliez, Toffano-Nioche, Ferey, Kepes, & Gherbi, 2006).

Scatter Plots

The scatter diagram graphs pairs of numerical data, with one variable on each axis, to look for a relationship between them. If the variables are correlated, the points will fall along a line or curve. The better the correlation, the tighter the points will hug the line (ASQ, 2018). A scatter plot is a simple way to visualise differentially expressed genes, for example, An IGV scatter plot, as shown in Figure 27, displays the relationship between two sets of sample tracks of continuous-valued data in a genomic region. Supported data types include gene expression, copy number data, and methylation data (IGV, 2018).



X: GENE_EXPRESSION   Y: COPY_NUMBER   Classify By: Subtype

Figure 27 IGV scatter plot displays the relationship

between two sets of sample tracks of continuous-value data in a genomic region (IGV, 2018).

UCSC Cancer Genomics Browser is a web-based application for hosting, visualising, and analysing cancer genomics datasets which include 575 public datasets from genome-wide analyses of over 227,000 samples. The browser can display genome-wide experimental measurements for multiple samples, which can originate from multiple datasets alongside their associated colour-coded clinical information. The browser provides interactive views of data from genomic regions to annotated biological pathways and user-contributed collections of genes. Integrated statistical tools provide quantitative analysis within all available datasets. Users can easily discover and share their research observations by exploring the relationship between genomic alterations and phenotypes with multidimensional visualisations (M. Goldman et al., 2015).

A UCSC scatter plots are used to quickly and easily see the relationship between any two variables or columns of data such as Glioblastoma Multiforme (GBM) and Lower Grade Glioma (LGG) samples in Figure 28 (Mary Goldman, 2017). The x-axis shows copy number variation in chromosome 19q and the y-axis shows copy number variation in chromosome 1p. Samples are coloured by primary disease. We can see here that there is a subset of samples in LGG that have a strong correlation between a deletion of chr19q and chr1p. GBM samples do not show this relationship (Mary Goldman, 2017).

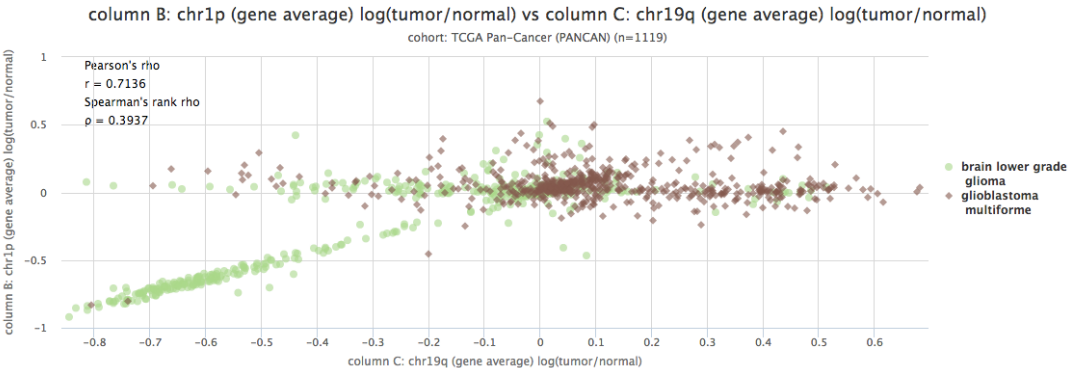


Figure 28 UCSC scatter plots.

It is for Glioblastoma Multiforme (GBM) and Lower Grade Glioma (LGG) samples in TCGA (Mary Goldman, 2017)

3D Scatter Plot is used to discover relationships between three variables at the same time and is boosted by the recent widespread use of virtual reality devices. Even though Virtual Reality (VR) has been in development for decades, only recently are industries dedication sizable resources, both money and time, into producing compelling experiences for VR. Virtual Reality reveals spatially complex structures behind 3D data in an easy visualisation way. 3D scatter plots can solve the problematic issues on common 2D scatter plots such as the overlapping of data and the absence of depth perception (Gray, 2016). Some genomic and cancer data visualisation tools such as Medical Data Visualisation (Lau et al., 2019; Q. V. Nguyen et al., 2014) start to use 3D scatter plots and support mixed reality devices such as Microsoft HoloLens.

Heatmaps

Heatmaps are 2D graphical false-colour image representations of data which makes use of a predefined colour scheme and different colours display different values and variations in a data matrix. Heatmap plot is a fundamental method in genomic data visualisation and is broadly used to unravel patterns hidden in genomic data, especially for gene expression analysis and methylation profiling (Gu, Eils, & Schlesner, 2016). Heatmaps may also be combined with clustering methods which group markers together based on the similarity of their patterns. Many genomic visualisation tools provide heatmap plots, such as Ngs.plot (Shen et al., 2014), Gitools (Perez-Llamas & Lopez-Bigas, 2011) and PARADIGM (Vaske et al., 2010). Figure 29 shows an example of a heatmap visualisation that compares gene of interests between the selected patients ALL92, ALL129, ALL321 and ALL323.

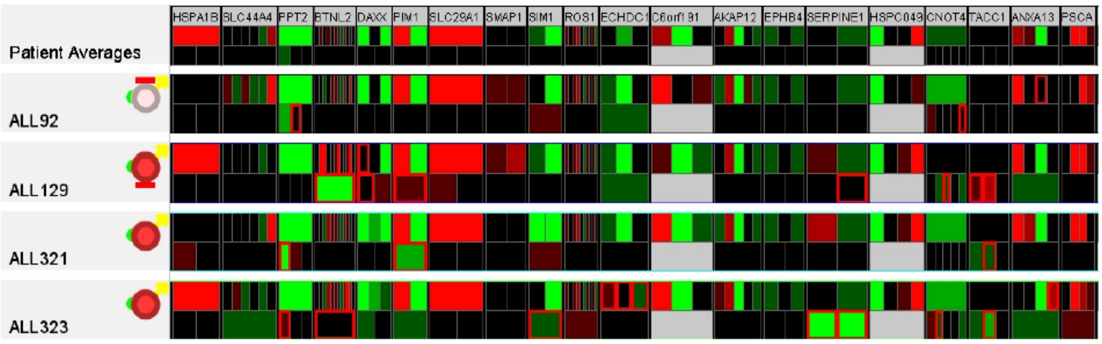


Figure 29 A heatmap for comparing genes between different patients.

It is for comparing genes of interests between patients ALL92, ALL129, ALL321 and ALL323 which were chosen by users. (Q. V. Nguyen et al., 2014)

Heatmaps are very handy for large, multi-dimensional datasets visualisation. High-throughput gene expression data are often displayed using heat maps which data are displayed in a grid where each row represents a gene and each column represents a sample. Colours and intensity of each box represent variations of gene expression. Scientists often use green-black-red heat maps to visualise gene expression data from microarrays (Genomics, 2017).

Most heatmap representations are also combined with clustering methods to group genes or samples based on their expression patterns. Each gene is represented as a row and is colour-coded to represent the intensity of its variation, such as positive or negative, relative to a reference value, and biological samples are represented as columns in the grid (Levin, 2017).

Genomic coordinates

Genomic coordinates plot is a common way to visualise oncogenomics data to show alterations tied to their genomic loci. UCSC, IGV, RNASeqBrowser, GATK, Savant Genome provide genomic coordinates. The different tools may have different focuses but most of them can display genomic topography of alterations in each tumour samples as genomic tracks to inspect particular genome loci.

Integrative Genomics Viewer (IGV) is a lightweight visualisation tool for interactive exploration of integrated genomics datasets and it makes use of efficient, multi-resolution file formats to enable intuitive real-time exploration of diverse, large-scale genomic datasets on standard desktop computers. IGV can handle large heterogeneous data set to provide a smooth and intuitive user experience at all levels of genome resolution. It uses special data tiling technique which is a pyramidal data structure to support interactive exploration of large-scale genomic data sets on standard desktop computers (Thorvaldsdottir, Robinson, & Mesirov, 2013). In IGV, all tracks can be annotated with a coordinates application colour-coded sample and clinical information and genomic regions can be annotated with text labels (Robinson et al., 2011). Figure 30 shows an IGV attribute panel which displays a colour-coded matrix of phenotypic and clinical data. Just below the command bar is a header panel with an ideogram representation of the currently viewed chromosome, along with a genome coordinate ruler that indicates the size of the region in view. The remainder of the window is divided into one or more data panels and an attribute panel. Data are mapped to the genomic coordinates of the reference genome and are displayed in the data panels as horizontal rows called ‘tracks’. Each track typically represents one sample, experiment or genomic annotation. If any sample or track attributes have been loaded, they are displayed as a colour-coded matrix in the attribute panel. Each column in the matrix corresponds to an attribute, and a track’s attribute values are displayed as a row of coloured cells adjacent to the track (Thorvaldsdottir et al., 2013).

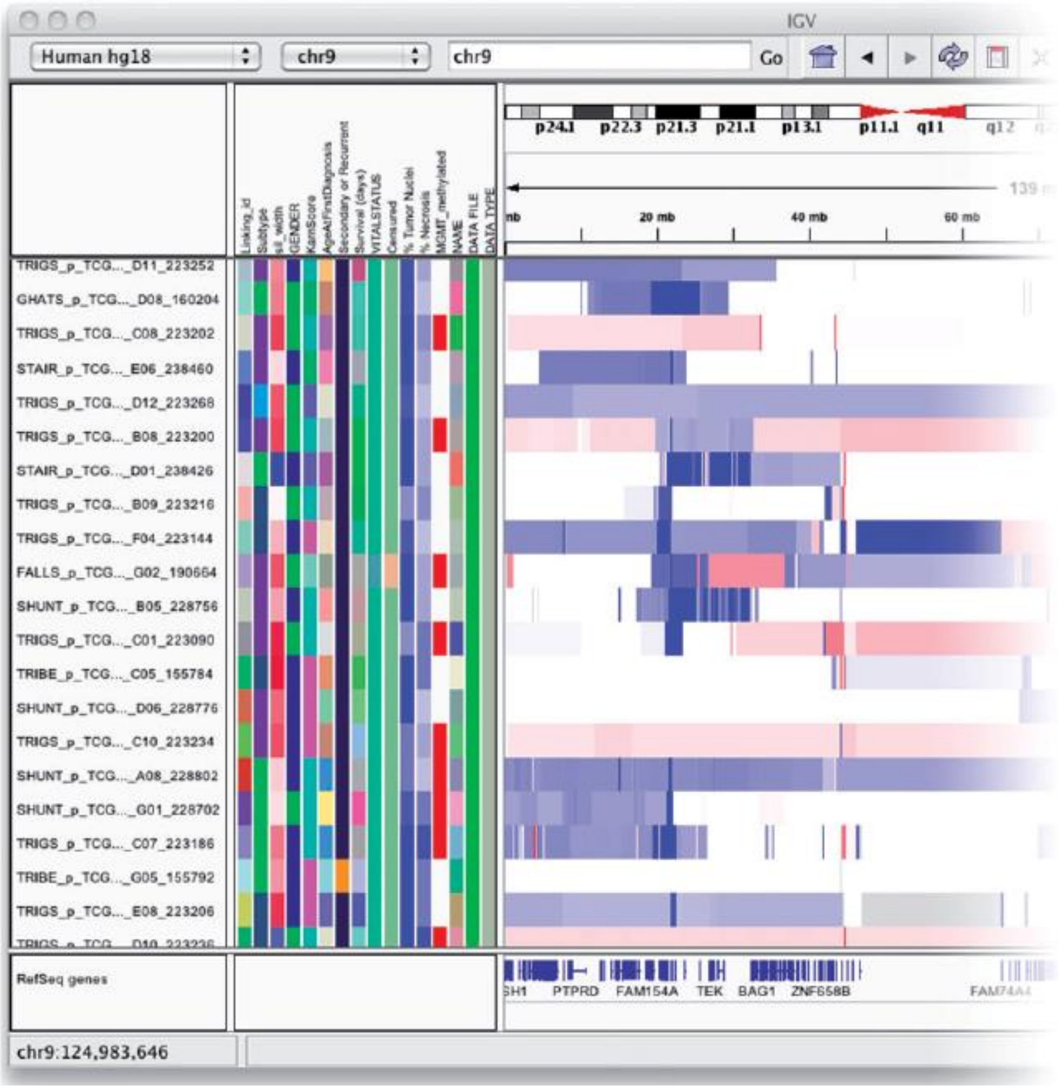

IGV
Human hg18
chr9
chr9
Go
p24.1
p22.3
p21.3
p21.1
p13.1
p11.1
q11
q12
139 m
mb
20 mb
40 mb
60 mb
Linking_id
Subtype
sil_width
GENDER
KarnScore
AgeAtFirstDiagnosis
Secondary or Recurrent
Survival (days)
VITALSTATUS
Censored
% Tumor Nuclei
% Necrosis
MGMT_methylated
NAME
DATA FILE
DATA TYPE
TRIGS_p_TCG..._D11_223252
GHATS_p_TCG..._D08_160204
TRIGS_p_TCG..._C08_223202
STAIR_p_TCG..._E06_238460
TRIGS_p_TCG..._D12_223268
TRIGS_p_TCG..._B08_223200
STAIR_p_TCG..._D01_238426
TRIGS_p_TCG..._B09_223216
TRIGS_p_TCG..._F04_223144
FALLS_p_TCG..._G02_190664
SHUNT_p_TCG..._B05_228756
TRIGS_p_TCG..._C01_223090
TRIBE_p_TCG..._C05_155784
SHUNT_p_TCG..._D06_228776
TRIGS_p_TCG..._C10_223234
SHUNT_p_TCG..._A08_228802
SHUNT_p_TCG..._G01_228702
TRIGS_p_TCG..._C07_223186
TRIBE_p_TCG..._G05_155792
TRIGS_p_TCG..._E08_223206
RefSeq genes
SH1
PTPRD
FAM154A
TEK
BAG1
ZNF658B
FAM74A4
chr9:124,983,646

Figure 30 IGV genomic coordinates.

It shows a colour-coded matrix of phenotypic and clinical data (Thorvaldsdottir et al., 2013)

Networks

Networks can show functional relationships between different genomic entities to allow the researchers to visually explore clusters of nodes representing highly interconnected altered genes that can constitute driver pathways or subnetworks. An example is Cytoscape which provides network visualisation in genomic research.

Cytoscape is an open source software for visualising complex networks and integrating these with any type of attribute Networks Desktop data such as genomics data and clinical patient information. Cytoscape is most powerful when used in conjunction with large databases of protein-protein, protein-DNA, and genetic interactions that are increasingly available for humans and model organisms. The software is extensible through a straightforward plug-in architecture, allowing rapid development of additional computational analyses and features (Shannon et al., 2003).

Figure 31 shows breast cancer genomic data visualisation with network method from Cytoscape v3.4.0. The upper network is the GO analysis based on the biological process of the 513 DEGs and the bottom network shows the KEGG pathway analysis of the 513 DEGs (Liu et al., 2018).

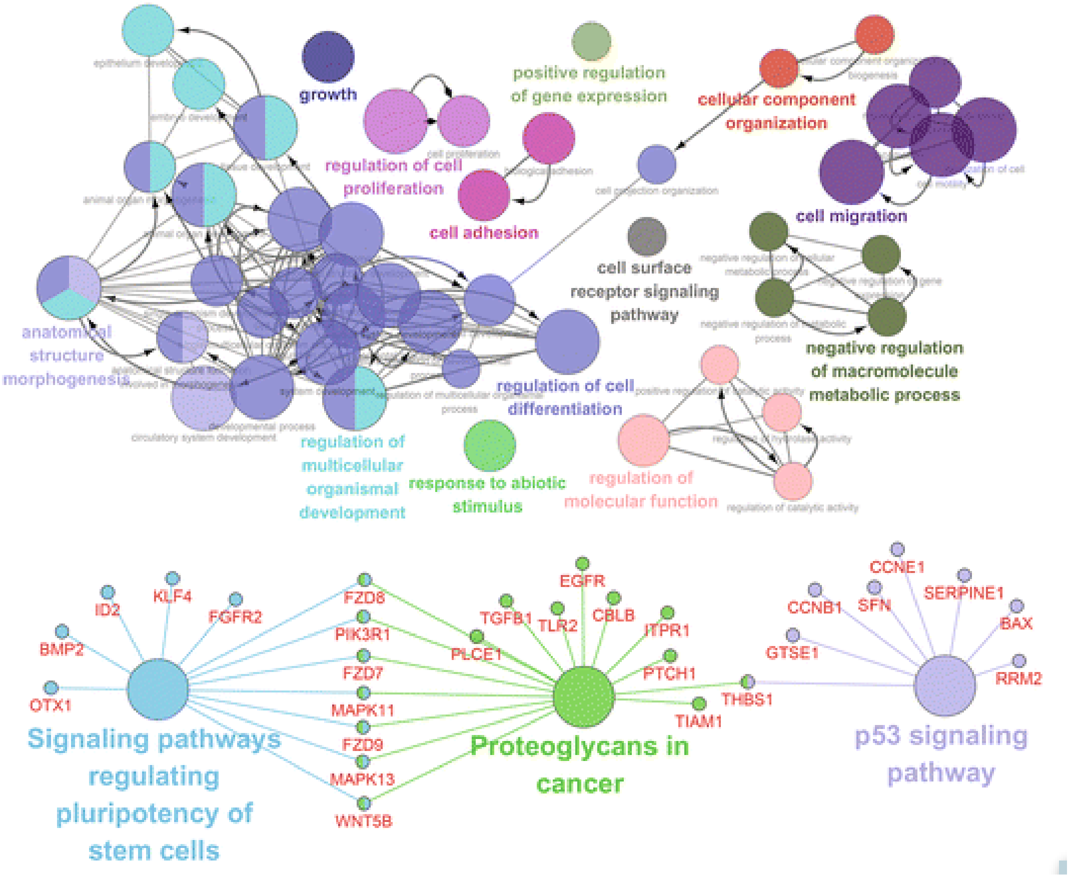

growth
positive regulation of gene expression
cellular component organization
regulation of cell proliferation
cell adhesion
cell migration
cell surface receptor signaling pathway
anatomical structure morphogenesis
regulation of cell differentiation
negative regulation of macromolecule metabolic process
regulation of multicellular organismal development
response to abiotic stimulus
regulation of molecular function
ID2
KLF4
FGFR2
BMP2
OTX1
FZD8
PIK3R1
FZD7
MAPK11
FZD9
MAPK13
WNT5B
EGFR
TGFB1
TLR2
CBLB
ITPR1
PLCE1
PTCH1
THBS1
TIAM1
CCNE1
CCNB1
SFN
SERPINE1
BAX
GTSE1
RRM2
Signaling pathways regulating pluripotency of stem cells
Proteoglycans in cancer
p53 signaling pathway

Figure 31 Network visualisation from Cytoscape v3.4.0.

The upper network is the GO analysis based on the biological process of the 513 DEGs and the bottom network shows the KEGG pathway analysis of the 513 DEGs (Liu et al., 2018)

Cluster

The cluster is a strategy that used to combine other visualisation methods such as scatter plots, heatmaps, networks. For example, medical data visualisation uses scatter plot cluster while UCSC uses heat map cluster. A cluster is usually a group of similar elements that can be represented by a profile, either a summary measure such as a cluster means or one of the elements itself.

Clustering combined with heatmaps enable grouping of genes or samples which can be obtained through high-throughput sequencing methods such as RNAseq or DNA microarray studies together. Clustering is useful in visualising similarity of gene expression pattern (Genomics, 2017). Figure 32 shows using the UCSC Cancer Genomics Browser to explore relationships between somatic mutation profiles, genomic subtypes and survival. In the Figure 32 a) shows somatic mutations for the most-significantly mutated genes in TCGA Acute Myeloid Leukemia (AML) tumour samples3. Samples are arranged in rows and genes in columns. Red indicates that the tumour sample harbours non-synonymous coding mutations in the corresponding gene while white indicated that such mutations were not detected. (b) Column 1 represents the miRNA expression clusters3, Column 2 represents the DNA methylation clusters, and Column 3 represents cytogenetic risk category for the AML cohort (Peter Laird, Personal Communication). For each column, each cluster or category was assigned a distinct colour from the D3 colour map, with five clusters for miRNA expression (cluster 1–5) and nine for DNA methylation (cluster 1–9), and three for cytogenetic risk category (favourable, intermediate, poor). A strong concordance is observed between miRNA cluster 3 (orange), DNA methylation cluster 3 (also orange) and intermediate cytogenetic risk (light blue); and between miRNA cluster 5 (green), DNA methylation cluster 5 (also green) and favourable cytogenetic risk (dark blue) (Cline et al., 2013).

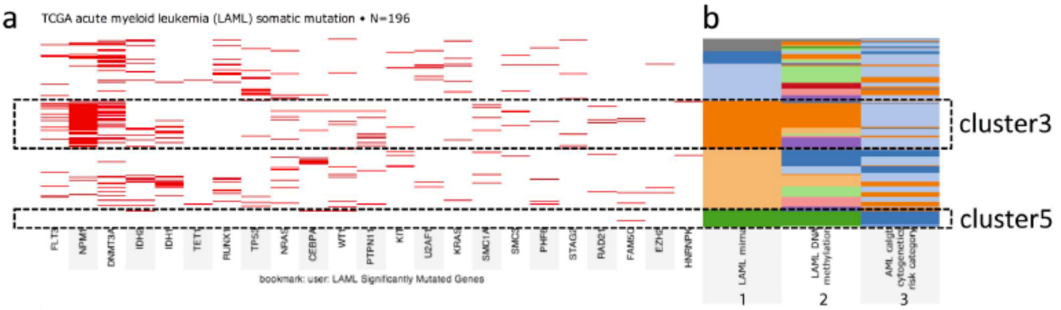


Figure 32 UCSC shows clustering heatmaps.

It is to explore relationships between somatic mutation profiles, genomic subtypes and survival (Cline et al., 2013).

Clustering method also supports scatter plots, network, genomic coordinates methods to show a group of similar elements. Clustering data can identify a subset of representative examples to process sensory signals and detect patterns in data. Clustering data based on a measure of similarity is a critical step in scientific data analysis and in engineering systems. A common approach is to use data to learn a set of centres such that the sum of squared error between data points and their nearest centres is small (Frey & Dueck, 2007).

Artificial Intelligence

In recent years, Artificial Intelligence (AI) starts to use big data visualisations including multivariate genomic data for the development of quick hardware (Nilsson, 2009). DeepVariant (Knight, 2017) is a tool that uses the latest AI techniques to build a more accurate picture of a person's genome from sequencing data. The tool is fed the data which is from millions of high-throughputs reads and fully sequenced genomes from the Genome in a Bottle (GIAB) project, a public-private effort to promote genomic sequencing tools and techniques, to a deep-learning system and painstakingly tweaked the parameters of the model until it learned to interpret sequenced data with a high level of accuracy (Knight, 2017). DeepVariant is a genomic variant caller which uses deep neural networks to call genetic variants in germline genomes. It is originally developed by Google Brain and Verily Life Science and it won the 2016 PrecisionFDA Truth Challenge award for Highest SNP Performance (Google, 2017).

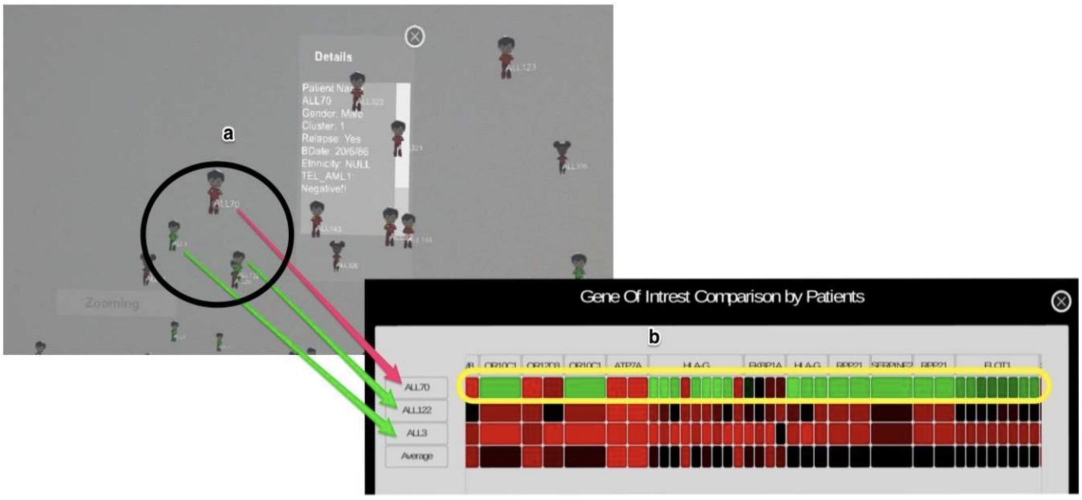
Emerging Platforms

Virtual Reality (VR) enables the psychophysical immersive experience in an artificial computer-generated virtual environment (Simpson, LaViola, Laidlaw, Forsberg, & van Dam, 2000). Augmented Reality (AR), usually, is built upon VR in integrating and overlaying the virtual environment into the user's real-world allowing the user to interact with the virtual objects in the context of his actual surroundings (Chang, Peng Xu, & Wang, 2013; Shan, Doyle, Samavi, & Al-Rei, 2017). Special equipment such as a Head-Mounted Display (HMD) or Cave Automatic Virtual Environment (CAVE) system is required for the use of VR/AR technologies. The sensor and camera on the equipment will help the system to determine and track the user moment and move the point of view accordingly.

Shan et al. (2017) developed an AR visualisation which runs on the mobile platform to deliver real-time 3D brain tumour volume rendering. It allows the clinician to visualise and communicate with the patients on their tumours size and location. The visualisation uses the facial features of the patient as the tracking point to project the reconstructed brain tumour model onto the same location as the subject's actual anatomy.

Chang Y, Xu WP and Wang L (Chang et al., 2013; Shan et al., 2017) have created a 3D AR visualisation for archaeological purposes. It uses the ARToolKit in rendering the objects. The purpose of the visualisation is to create a platform for underground cultural heritage protection and research. VR has been used in big data visualisations including multivariate genomic data for modern VR devices. A number of analysts even believe that the applications of AI to VR enables important possibilities such as AI-based continuous image recognition reporting results in a VR display (Dooley, 2017). One of the biggest challenges of big data is extracting information in a way that the human mind can comprehend and the human mind is just too feeble to comprehend such vast amounts of information. Immersive environments can measure people's reactions of large data sets to understand the subconscious process of the human brain to determine the optimum amount of information to display. VR either simplifies the visualisations so as to reduce the cognitive load, thus keeping the user less stressed and more able to focus. Or it will guide the person to areas of the data representation that are not as heavy in information (Stolk et al., 2002; Verma, 2017).

Interactive visualisation tools have also been used in childhood cancer research that can show the whole group of patients' data with a 3D scatter plot as well as to individual patient's details, zoom and rotate the visualisation plot, compare gene among several patients and interact with users and shows the comparison visualisation between selected patients (Q. V. Nguyen et al., 2016). The tool supports different operating systems including mixed reality devices. Figure 33 shows a 3D scatter plot from the tool run in Microsoft HoloLens, which is a pair of mixed reality smart glasses developed and manufactured by Microsoft. HoloLens gained popularity for being one of the first computers running the Windows Mixed Reality platform under the Windows 10 operating system and it can trace its lineage to Kinect, an add-on for Microsoft's Xbox gaming console that was introduced in 2010 (Microsoft, 2018).



Details

Patient Na
ALL70
Gender: Male
Cluster: 1
Relapse: Yes
BDate: 20/6/86
Ethnicity: NULL
TEL_AML1
Negative!!

Figure 33 A Visualisation tool for childhood cancer research that runs in Microsoft HoloLens.

(Lau et al., 2019).

3.1.2 Trends of Genomic Data Analytics

Figure 34 shows tools for visual analytics of genomic and cancer data with the around years they started to be developed or the related paper to be written. We can see around 2000 to 2015, most genomic data visualisation tools only use some traditional methods such as scatter plots, heatmaps, genomic coordinates, networks and clustering. Since 2016, AI has been integrated with visualisation tools such as machine learning algorithms for predictions and personalised medicine. Some visualisation tools support new environments such as mobile devices, VR, AR, and large and high-resolution screens. Some tools were used on for a short time such as X:map and GenomeCom, while some tools were developed very early before 2010, but are kept maintenance and adding new features until now such as GATK and Cytoscape, which are still very popular genomic data visualisation tools now. Integration is also a way to keep a tool lasting for a longer time. For example, Epiviz can obtain annotation data from the UCSC, Gitools can get heatmaps from IGV, and RNASeqBrowser is compatible with UCSC as shown in Figure 34 with purple arrows.

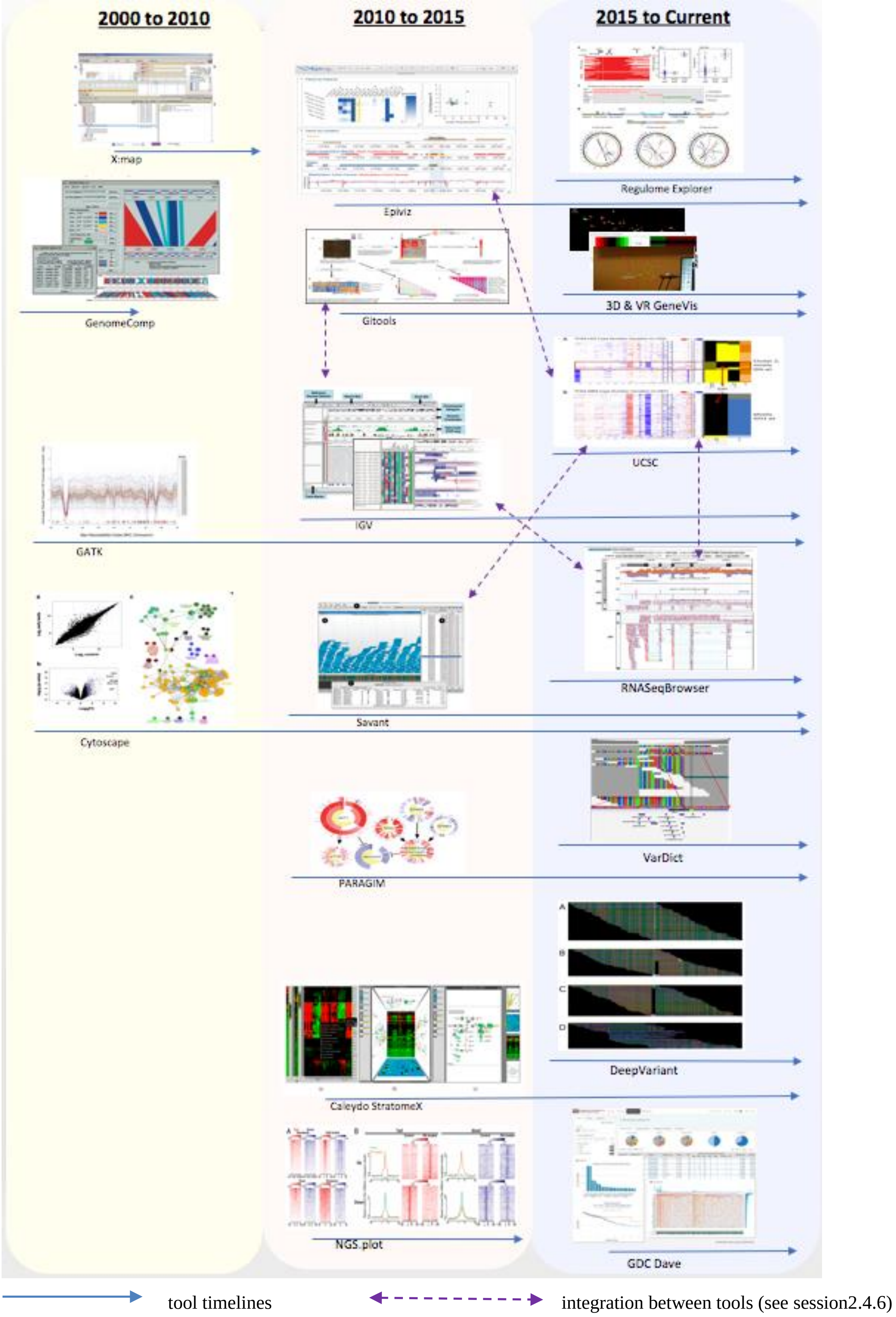


Figure 34 Tools timeline and integration, the blue arrows stand for the timeline and the green arrows stand for integration.

Current Status of Genomic and Cancer Data Visualisation Tools

This thesis classifies the tools for visual analytics of genomic and cancer data in Table 2. Some tools have not been updated recently such as GenomeComp, X:map, PARAGIM, and NGS.plot while most tools are still getting maintenance, or upgraded with new technologies such as IGV and other tools (see Figure 34). Some non-updated tools are still used and can be downloaded. GenomeComp is a visualisation tool which is implemented as a stand-alone program that can compare, parse and visualise large genomic sequences, especially closely related genomes such as interspecies or interstrain (Jian Yanga, 2003). It was developed by Laboratory of Bioinformatics, Institute of Biophysics Beijing and use Perl/TK that can run in Linux, Unix, Mac OS X and Microsoft Windows operating systems. The final version was updated in 2004 (Center, 2004).

X:Map is a tool designed specifically for high-density microarrays that are required to show for each gene, transcript and exon the probe sets that match it, their specificity and for each probe, their locations of potential hybridisation and for each individual exon, its sequence (Yates et al., 2008). X:Map is a Genome annotation database browser developed by University of Manchester in the UK around 2008. PARAGIM is a tool which focuses on inferring patient-specific genetic activities incorporating curated pathway interactions among genes and can predict the degree to which a pathway’s activities are altered in the patient using probabilistic inference.

CircleMap is one of the PARADIGM visualisation methods that produce heatmaps with a circular layout. Different data sets coming from the same samples can be plotted as different layered circles that form a node. The data layers are plotted application maintaining the sample order, which can be adjusted by the user. CircleMap visualisation can be used to display multiple datasets centred around each gene in a pathway (Vaske et al., 2010). The tool is a factor graph framework for pathway inference on high-throughput genomic data and was developed by Charles Vaske and Steve Benz from the Regents of the University of California, Santa Cruz in around 2010.

Ngs.plot is a tool to help understand the relationship between the millions of functional DNA elements and their protein regulators and demonstrate how they work in conjunction to manifest diverse phenotypes, which is a key to know the mammalian genome. The tool visualises massive datasets and genomic information based on big sequencing data and it can produce one billion sequencing reads in a few days. Ngs.plot uses two steps to quickly mine and visualise genome samples, the first step is to define a region of interest and the second step is to plot something meaningful (Shen et al., 2014). It is a quick mining and visualisation tool for NGS data, platform independent, and programming language is R and Python. It was developed by Peter Briggs from the University of Manchester in around 2014, supported by the Friedman Brain Institute and the National Institutes of Health.

New Visualisation Methods are Applied to Tools.

The active tools usually have a commercial website and keep being updated with new methods. For example, GATK can do deep learning with modern AI technology by using variants and annotations encoded as tensors, which carry the precise read and reference sequences, read flags, as well as base and mapping quality (Samwell, 2017). Genome Analysis Toolkit (GATK) is a structured programming framework designed to process exomes and whole genomes generated with illumine sequencing technology and also can be adapted to handle a variety of other technologies and experimental designs. This toolkit focuses on the variant discovery and also includes many utilities to perform related tasks such as processing and quality control of high-throughput sequencing data (GATK, 2017). The GATK provides a small but rich set of data access patterns that encompass the majority of analysis tool needs and it can separate specific analysis calculations from common data management infrastructure for correctness, stability and efficiency (McKenna et al., 2010).

DeepVariant is also a visualisation tool that uses machine learning technique. It mainly uses AI to identify all the mutations that an individual inherits from their parents and modelled loosely on the networks of neurons in the human brain (Moleten, 2017). DeepVariant also applies the latest AI techniques to build a more accurate picture of a person’s genome from sequencing data. The tool is fed with the data from millions of high-throughputs reads and fully sequenced genomes to a deep-learning system. The tool also tweaks the parameters of the model until it learned to interpret sequenced data with a high level of accuracy (Knight, 2017).

Verdict tool (Lai et al., 2016) uses PCR technology to amplify genes before submitting them to sequencing. VarDict's abilities to detect PCR artefacts, such as amplicon bias and mispaired primers, together with the linear scalability to depth, make it desirable in reducing both false positives and false negatives. VarDict is highlighted as a unique variant caller of high value in cancer translational research by the properties, and the value to detect otherwise missed variants in cancer samples demonstrated. The algorithm is open source and freely available for public use. VarDict is a novel and versatile variant caller for both DNA- and RNA-sequencing data and it simultaneously calls SNA, MNV, InDels, complex and structural variants, expanding the detected genetic driver landscape of tumours. VarDict has more accurate allele frequency estimation by performing local realignments on the fly. VarDict has three main features including i) performing scales linearly to sequencing depth, enabling ultra-deep sequencing used to explore tumour evolution or detect tumour DNA circulating in blood; ii) preforming amplicon aware variant calling for polymerase chain reaction (PCR)-based targeted sequencing which is often used in diagnostic setting; and iii) detecting differences in somatic and loss of heterozygosity variants between paired samples. VarDict uses data from the Cancer Genome Atlas (TCGA) Lung Adenocarcinoma dataset to call known driver mutations in KRAS, EGFR, BRAF, PIK3CA, and MET in 16% more patients than previously published variant calls (Lai et al., 2016).

Some visualisation tools start to support VR/AR/Immersive big screen and mobile devices such as Children Cancer Data Visualisation tools. It can show the whole group of patients' data with a 3D scatter plot and check a single patient's details, zoom and rotate the visualisation plot, compare Gene among several patients and interact with users and shows the Comparison Visualisation between selected patients (Q. V. Nguyen et al., 2016)

Tools are Integrated with Each Other

Some visualisation tools can be integrated to be used together for better analysis outputs. For example, Epiviz can obtain annotation data from the UCSC genome browser (Chelaru, Smith, Goldstein, & Bravo, 2014). Epiviz is a genomic information visualisation tool which can quickly and easily visualise and compare large amounts of genomic information resulting from high-throughput sequencing experiments. As the first system to provide tight integration between a state-of-the-art analytics platform and a modern, powerful, integrative visualisation system for functional genomics, Epiviz can interactively support a number of widely used, state-of-the-art methods for i) ChIPseq where iterative visualisation of data and results of peak-calling algorithms is necessary; ii) RNA-seq analyse where both location-based coverage and feature-based expression levels are required; iii) methylation analyses using where location-based analysis at multiple genomic scales (Chelaru et al., 2014).

Gitools can get heatmaps from IGV through load command and then send locate commands for selected rows in the heatmaps to IGV via IGV logo in the Gitools toolbar, which makes it easy to spot and compare genes of interest within IGV (Gitools, 2018). Gitools is a desktop application for analysis and visualisation of matrices using interactive heatmaps which contain multiple dimensions. It has interactive capabilities to allow the user to filter, sort, move, and hide rows and columns in the heatmaps. Gitools is especially useful for cancer genomic analysis as it includes several methods for integrating data sources as well as the ability of importing data directly from some other tools and sources. Gitools can be used by researchers without advanced knowledge on bioinformatics as well as experienced users who usually perform complex operations and analyses using the command line interfaces (Perez-Llamas & Lopez-Bigas, 2011).

Savant also allows users to automatically download annotation tracks from various public resources such as the UCSC Genome Browser (M. Fiume et al., 2010). Savant Genome Browser is a sequence annotation, desktop visualisation and analysis browser for genomics data. This tool was primarily developed for the effective visualisation of large sets of high-throughput sequencing data. Multiple visualisation modes enable the exploration of genome-based sequence, points, intervals, or continuous datasets. Plugins, such as WikiPathways plugin are available, which aids the navigation of the data by the integration of pathways (M. Fiume et al., 2010).

RNASeqBrowser can be compatible with UCSC files and extend the functionality over IGV. RNASeqBrowser is a visualisation tool that adds several new types of tracks to show NGS data such as individual raw reads, SNPs and InDel. The tool can dynamically generate RNA secondary structure which is useful for identifying non-coding RNA such as miRNA, and it overlays NGS wiggle data to display differential expression. Paired reads are also connected in the browser to enable easier identification of novel exon/intron borders and chimaeric transcripts. Strand specific RNAseq data is also supported by RNASeqBrowser that displays reads above (positive strand transcript) or below (negative strand transcripts) a central line (An et al., 2015).

Interaction Tools

The good tools usually allow users to interact intuitively with data and choose multi visualisation methods to support different research purpose. For example, GDC (Genomic Data Commons) DAVE (Data Analysis, Visualisation, and Exploration) Tools usually use scatter plots to visualise mutations and their frequency across cases that are mapped to a graphical visualisation of protein-coding regions. They use heatmaps to visualise the top mutated genes across projects and the number of cases affected. Web interface of GDC DAVE Tools can analyse cancer genomic data, in real time and online, without the need to download or process the data. Users can navigate from project cohorts to individual patients, to specific genes and mutations of interest. DAVE tools normally use specialised graphs to visualise genomic signatures of cancer and identify potential drivers of disease. They also visualise patient survival curves and identify the molecular consequence of a mutation on resultant protein (Staudt, 2017). DAVE Tools allow users to interact intuitively with GDC (Genomic Data Commons) data and promote the development of a true cancer genomics knowledge base, which includes the following key features: i) view most frequently mutated genes, ii) plot high impact mutations using oncoGrid, iii) perform survival analysis, iv) visualise mutations for protein-coding regions, v) view cancer distribution, view top mutated genes across projects, vi) view genes annotated by COCMIC, vii) build and compare custom cohorts, and viii) perform set operations (NIH, 2017b).

We summarise the visual analytics tools for genomic and cancer data as discussed above in Table 2.

Tool Name/Website	Description	Visualisation Methods	Developer/Year	Tool Type
X:Map http://xmap.picr.man.ac.uk .	X:Map is a tool which designed specifically for high-density microarrays that are required to show for each gene, transcript and exon the probe sets that match it, their specificity and for each probe, their locations of potential hybridization and for each individual exon, its sequence(Yates et al., 2008).	Heatmap, Genomic coordinates	University of Manchester in the UK 2008	Genome annotation database browser
GenomeComp http://www.mgc.ac.cn/GenomeComp/	GenomeComp is a visualisation tool which is implemented as a stand-alone program that can compare, parse and visualise large genomic sequences, especially closely related genomes such as interspecies or interstrain (Jian Yanga, 2003).	Genomic coordinates	Laboratory of Bioinformatics, Institute of Biophysics Beijing 2002--2004	Use Perl/TK, run in Linux, Unix, Mac OS X and Microsoft Windows
Epiviz http://epiviz.cbcb.umd.edu/	Epiviz is a genomic information visualisation tool which can quickly and easily visualise and compare large amounts of genomic information resulting from high-throughput sequencing experiments. It is the first system to provide tight integration between a state-of-the-art analytics platform and a modern, powerful, integrative visualisation system for functional genomics (Chelaru et al., 2014).	Heatmaps, 2D scatter plot, Genomic coordinates	University of Maryland 2014 --now	Web-based genome browsing application
Gitools http://www.gitools.org/	Gitools is a desktop application for analysis and visualisation of matrices using interactive heatmaps which contain multiple dimensions. It has interactive capabilities to allow the user to filter, sort, move, and hide rows and columns in the Heatmaps. Gitools is especially useful for cancer genomic analysis as it includes all the methods implemented for some Integrative sources, and can import data directly from some other tools(Perez-Llamas & Lopez-Bigas, 2011).	Heatmaps	Biomedical Genomics Group located in Barcelona at the Biomedical Research Park in Barcelona 2011 - Current	Desktop application
UCSC https://genome-cancer.ucsc.edu/	UCSC Cancer Genomics Browser is a web-based application for hosting, visualising, and analysing cancer genomics datasets. The browser provides interactive views of data from genomic regions to annotated biological pathways and user-contributed collections of genes. (M. Goldman et al., 2015).	Heatmap, Cluster	UC Santa Cruz in the University of California system 2015 - Current	Web-based application
Integrative Genomics Viewer (IGV)	Integrative Genomics Viewer (IGV) is a lightweight visualisation tool for interactive exploration of integrated genomics datasets and it supports a wide range of genomic data including aligned sequence reads, mutations, copy number, RNAi screen, gene expression,	Heatmap, Genomic coordinates,	Broad Institute, the University of California 2013 - Current	Visualisation tool for integrated

http://software.broadinstitute.org/software/igv/	methylation, and genomic annotations(Robinson et al., 2011).	Cluster, 2D scatter plot		genomics datasets
Savant Genome Browser http://www.genomesavant.com/p/home/index/	Savant Genome Browser is a Sequence annotation, desktop visualisation and analysis browser for genomics data. This tool was primarily developed for the effective visualisation of large sets of high-throughput sequencing data. Multiple visualisation modes enable the exploration of genome-based sequence, points, intervals, or continuous datasets. Plugins are available, amongst which is the WikiPathways plugin, which aids the navigation of the data by the integration of pathways(M. Fiume et al., 2010).	Genomic coordinates, Heatmap, Cluster	The Computational Biology Lab at the University of Toronto (Marc Fiume, 2017). 2010 - Current	Desktop visualisation and analysis browser for genomics data
PARADIGM http://sbenz.github.io/Paradigm/	PARADIGM is a tool which focuses on inferring patient-specific genetic activities incorporating curated pathway interactions among genes and can predict the degree to which a pathway's activities are altered in the patient using probabilistic inference. CircleMap is one of the PARADIGM visualisation methods that produce heatmaps with a circular layout (Vaske et al., 2010).	Heatmap	Charles Vaske, Steve Benz, University of California, Santa Cruz 2010	A factor graph framework for pathway inference on high-throughput genomic data
Caleydo StratomeX http://caleydo.org/tools/stratomex/	Caleydo StratomeX is a visual analytics framework prepared for the visualisation of interdependencies between multiple datasets. It allows exploration of relationships between multiple groupings and different datasets. It can cluster genomics data of different alterations and represents them as matrix heatmaps. The different groupings are connected by ribbons whose width corresponds to the number of samples shared by the connected clusters. Clinical data and pathway maps can be integrated to characterise the clusters (Lex et al., 2012).	Heatmap, Cluster,	Marc Streit, Linz, Alexander Lex, Nils Gehlenborg, Christian Partl, Samuel Gratzl, Hanspeter pfister, Dieter Schmalstieg, and Peter J. Park (Caleydo, 2017). 2012 - Current	StratomeX is a visual analytics framework for the analysis of multiple stratified datasets.
Regulome Explorer http://explorer.cancerregulome.org	Regulome Explorer is a tool for the visualisation options include circular and linear genomic coordinates and networks(TCGA, 2012). TCGA takes an integrated approach toward a systems level understanding of regulatory disruptions in cancer which are intertwined within complex dynamical networks through a multitude of interactions among different types of molecules(GDAC, 2016).	Heatmap, Genomic coordinates	Institute for Systems Biology and MD Anderson Cancer Centre 2016 - Current	A tool for the integrative exploration of associations between clinical and molecular features of data
Cytoscape	Cytoscape is an open source software for visualising complex networks and integrating these with any type of	Networks	U.S. National Institute of General Medical Sciences (NIGMS)	An open source software platform for

http://www.cytoscape.org	attribute Networks Desktop data such as genomics data and clinical patient information. (Shannon et al., 2003).		and National Resource for Network Biology (NRNB). 2003 - Current	visualising complex networks
Ngs.plot https://code.google.com/p/ngsplot	Ngs.plot is a tool to help understand the relationship between the millions of functional DNA elements and their protein regulators and demonstrate how they work in conjunction to manifest diverse phenotypes. Ngs.plot uses two steps to quickly mine and visualise genome samples, the first step is to define a region of interest and the second step is to plot something meaningful(Shen et al., 2014).	Heatmap	Peter Briggs from the University of Manchester supported by the Friedman Brain Institute; and the National Institutes of Health 2014	a quick mining and visualisation tool for NGS data Programming language is R and Python
GDC DAVE (Genomic Data Commons Data Analysis, Visualisation, and Exploration) https://gdc.cancer.gov/analyze-data/gdc-dave-tools	GDC DAVE Tools allow users to interact intuitively with GDC (Genomic Data Commons) data and promote the development of a true cancer genomics knowledge base, which including the following key features: view most frequently mutated genes, plot high impact mutations using oncoGrid, perform survival analysis, visualise mutations for protein-coding regions, view cancer distribution, view top mutated genes across projects, view genes annotated by COCMIC, build and compare custom cohorts, and perform set operations (NIH, 2017b).	Heatmap, 2D Scatter Plot, Cluster	the National Cancer Institute (NCI) Centre for Cancer Genomics (CCG) from Maryland, United States. 2016 - Current	GDC Data Portal
VarDict https://github.com/AstraZeneca-NGS/VarDict	VarDict is a novel and versatile variant caller for both DNA- and RNA-sequencing data and it simultaneously calls SNA, MNV, InDels, complex and structural variants, expanding the detected genetic driver landscape of tumours (Lai et al., 2016).	Heatmap, Genomic coordinates	AstraZeneca which is in United States. 2016 - Current	VarDict is implemented in Perl
DeepVariant https://github.com/google/deepvariant	DeepVariant is a tool that uses the latest AI techniques to build a more accurate picture of a person’s genome from sequencing data. The tool is fed the data which is from millions of high-throughputs reads and fully sequenced genomes to a deep-learning system and painstakingly tweaked the parameters of the model until it learned to interpret sequenced data with a high level of accuracy (Knight, 2017).	AI, Genomic coordinates, Heatmap	Google Brain and Verily Life Science. 2016 - Current	Deep neural networks to call genetic variants in germline genomes.
RNASeqBrowser	RNASeqBrowser is a visualisation tool that incorporates and extend the function of the UCSC genome browser and NGS visualisation tools such as IGV (An et al., 2015).	Genomic coordinates, Cluster	JA, Australian Government Department of Health;	A visualisation tool that incorporates and extend the

			2015 - Current	function UCSC and IGV
Children Cancer Data Visualisation tool	Children Cancer Data Visualisation tool can show the whole group of patients' data with a 3D scatter plot and check a single patient's details, zoom and rotate the visualisation plot, compare Gene among several patients and interact with users and shows the Comparison Visualisation between selected patients (Q. V. Nguyen et al., 2016)	3D Scatter Plot, Heatmap, Cluster, VR	Western Sydney University 2016 - Current	Developed by Java, Unity 3D

Table 2 Tools for Visual Analytics of Genomic and Cancer Data

3.2 QUALITATIVE STUDY ON DOMAIN EXPERTS

To gain a better understanding of the domain user's preferences and expectation, we carried out a pilot study on genomic visualisation tools with the end users who were cancer researchers and medical doctors. As the participants were the end user of the genomic data visualisation tools, they could give feedback on the effectiveness and usefulness of the tools. Most of the participants were very familiar with disease datasets and they really needed such visualisation tools in their work. This is a qualitative analysis where all the feedback, comments and discovery are collected and analysed through one-to-one interviews. Particularly, the usefulness of the software and how the tools assist with genomic analysis are evaluated.

3.2.1 Prepared Tool Introduction

The three medical visualisation tools, developed at Western Sydney University, can visualise cohorts of patients with childhood Acute Lymphoblastic Leukaemia (ALL) and individual patient's details. The genomic data is extracted from bone marrow or peripheral blood specimens ethically obtained from ALL patients. Using various data mining strategies such as random forest deep learning algorithm, the combined genomic features, that best characterises the cohort of patients into chosen classes, are captured.

The three tools run in the different platform but use the same dataset, which was the expression and genomic SNP profiles of 100 paediatric B-cell ALL patients that were generated using Affymetrix expression microarrays (U133A, U133A 2.0 and U133 Plus 2.0) and Illumina NS12 SNP microarrays, respectively. Affymetrix expression microarrays generate 22,277 attributes, while each Illumina SNP microarray 13,917 attributes for each patient sample. Each attribute was mapped to a probe of DNA (or a gene), and the value for each attribute corresponded to the expression levels or genotype for the gene.

For complex problems, often many 100s of genomic features are required to build models that distinguish the ALL patients across possibilities. Visualising these results often requires 2D or 3D graphical representations of the cohort which presents as a 'cloud of spots,' each spot representing an individual patient in 'similarity space,' its location to other spots (patients) being a measure of overall similarity between the set of genomic features used to build the model as shown in Figure 35.

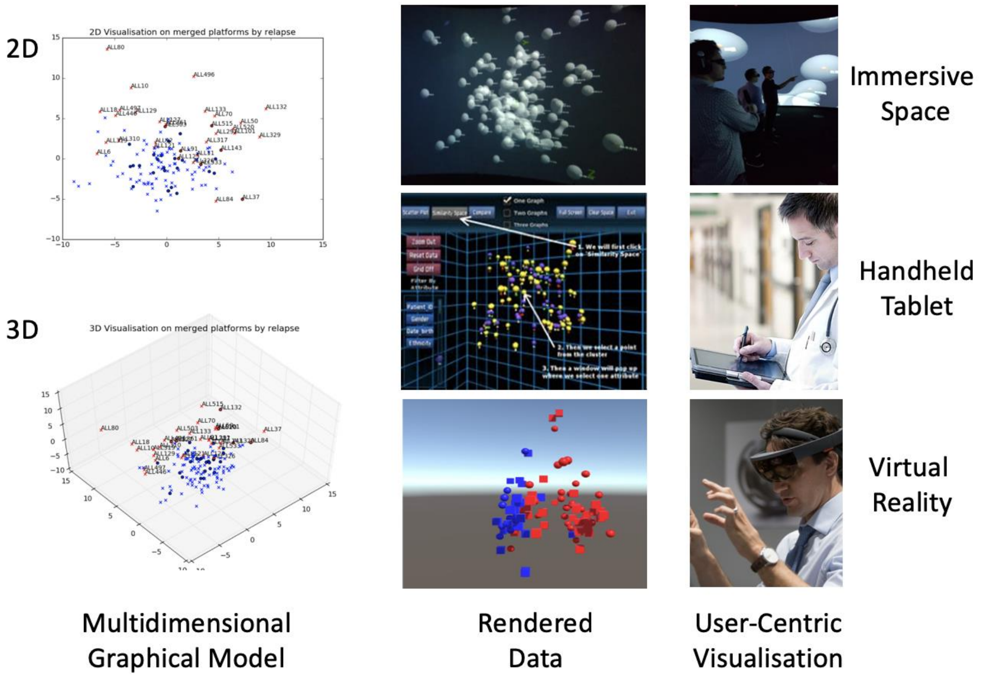

2D
2D Visualisation on merged platforms by relapse
3D
3D Visualisation on merged platforms by relapse
Immersive Space
Handheld Tablet
Virtual Reality
Multidimensional Graphical Model
Rendered Data
User-Centric Visualisation

Figure 35 Genomic Data Visualisation Models.

The three novel tools allow investigators to interrogate this graphical cloud and identify a single individual or cluster of individuals. Further, the location of the selected patient(s) in the cloud will inform how the genomic features differentiate them from others in the cloud. The visualisation tools allow for selection of the individual patient genomic features and to perform the patient-to-patient comparison. Depending on the model criteria, patients with similar genomics as defined by a closer location within similarity space but who have variant clinical characteristics (e.g. pathology, treatment outcome, adverse reactions etc) may be identified, providing the analyst with insights for how patient with similar genomics may be best treated in the future.

The investigation scenario is that the analysts need to choose a better treatment method for a new patient.

First, the analysts visualise the entire patient population in the similarity space to see an overview of the genetic similarity of the patient cohort where the closer patients are hypothetically genetically similar. The analysts might move rapidly to any location in the space, indicate the position of new patients in the genetic similarity space, customise visualisation via interaction, controllable attributes and filtering, and extract and picture specific features and patients within the similarity space. Then it is time to do patient-to-patient comparison. The tool displays concurrently the total gene expression and SNP data generated for each patient. Each probe set in the gene is represented as a dot point on the horizontal axis while the vertical position shows the order of the gene sorted by chromosomal order. The table includes all of the biological data associated with the patient.

From the overview of the entire genetic and biomedical information, the analysts can identify patterns and abnormality before exploring further. We also provide semantic zooming to enlarge the area of focus. The level of detail is updated automatically upon the information and available space. We believe understanding the biological differences within individual patients may influence clinical management decisions for those patients.

Medical visualisation tool developed by Nguyen et al from Western Sydney University can run on regular personal computers, including the main features:

- Show the whole group of patients' data with a 3D scatter plot. The main view shows patients in cohort with patient's labels on in a 3D scatterplot See Figure 36 A. The distances between patients in this space were indicative of genetic similarity.
- The patient separation found did not agree with clinical markers (e.g., white blood cell count cytogenetics) that were used for prognostication and risk stratification.
- In contrast, similarity spaces constructed with either the expression data or SNP data alone did not recover such a clear distinction.

- Patient-to-patient comparison view and comparison of the genes of interest (Think about the way you use this if you want to see how a new patient who was treated on a recent treatment protocol compared to a genomically similar patient who was treated on an older protocol)
- Check a patient's detail (double click on one of the patient bubble). Click to open a patient detail shows a patient's Affymetrix Gene Expression and Illumina SNPs detail.
- Set Microarray or SNP difference for one patient (Open a patient detail and find the menu under Action)
- Zoom and rotate the visualisation plot (Alt + left mouse to zoom, Ctrl + left mouse to move, left mouse to rotate)
- Turn the label on/off (the menu is under Draw Property). Rich graphical attributes, such as labels, axes, colours, size, shapes and visual bars, were also used to present clinical and background properties. The presentation can be adjusted by the users
- Genes of interest's visualisation to compare gene among several patients and interact with users (choose each patient you want to compare and add to Gene Comparison Visualisation, and then find Analysis->Show colour Map). This window shows the Comparison Visualisation between selected patients
- Modify visualisation by changing colours, size, icons for important attributes

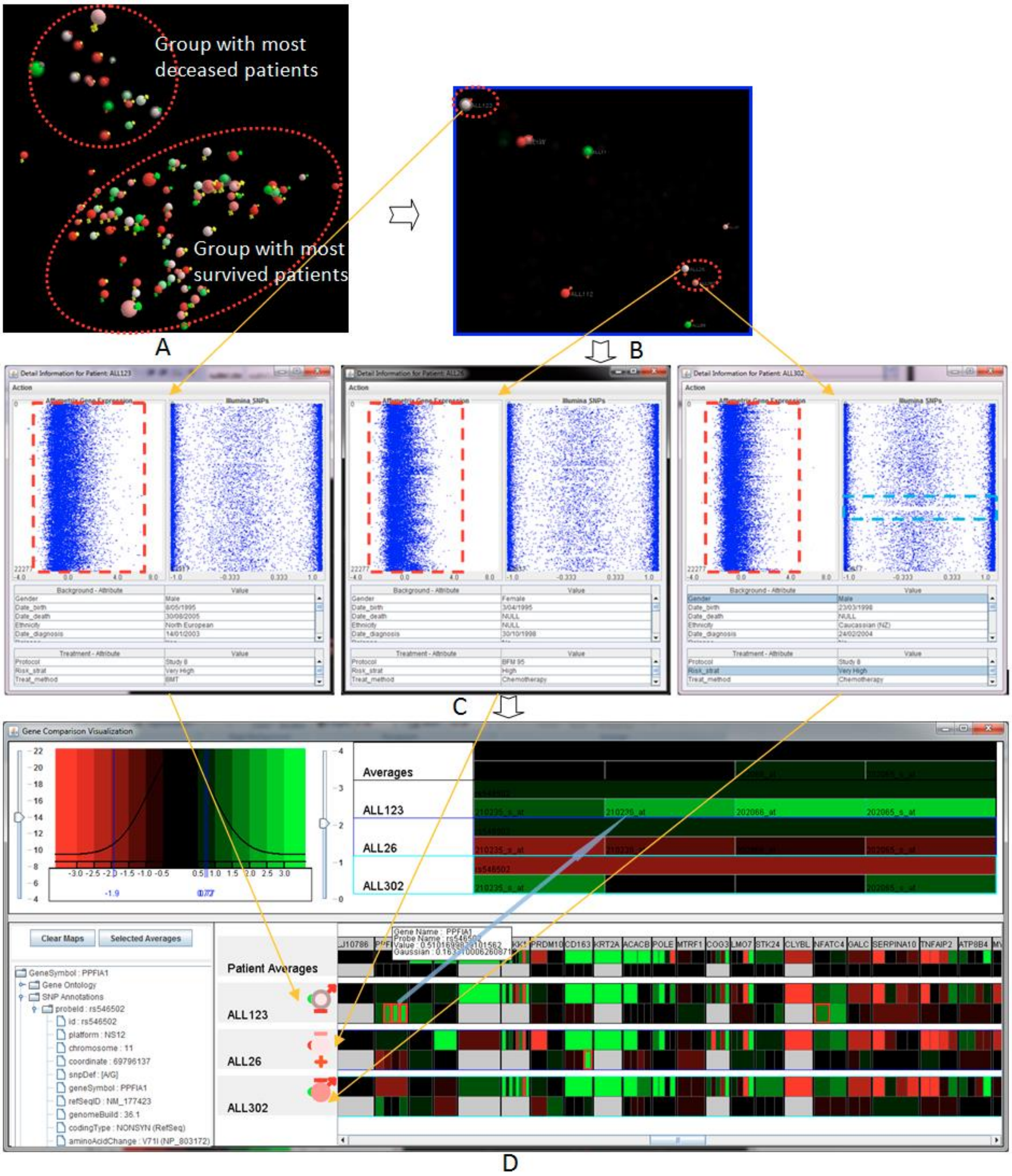

Group with most deceased patients
Group with most survived patients
A
B
C
D
Affymetrix Gene Expression
Illumina SNPs
Gene Comparison Visualization
Averages
ALL123
ALL26
ALL302
Clear Maps
Selected Averages
Patient Averages

Figure 36 The seamless visual analysis.

From the overview of the entire patient population in the similarity space (A); focused patients at a navigational stage (B); to patient-to-patient comparison view of raw data outputs from a data collection step, (left) Affymetrix gene expression and (right) Illumina SNP, both ordered to chromosomal location (C) and, finally, to the genes of interest view (D) (Q. V. Nguyen et al., 2014).

3D ScatterPlot Data Visualisation (Lau et al., 2019; Q. Vinh Nguyen et al., 2018) on mobile devices was developed in Unity3D by Lau et al. that can run in tablets such as IPAD or Android System such as Samsung tablet (Figure 37), including main Features:

- Show and compare selected patients' genomic data with heatmap plot
- Select a group of patients whose data are similar and compare them.
- Zoom the visualisation plots to check the patterns of a big group of patients or a small group of patients.
- Show all the group of patients in a 3D scatter plot with the different colour based on the attribute that the user chooses. Figure 37 shows a scatter plot with a different colour based on the patient's gender.
- The users can choose different patients one by one or a small group of patients at a time to compare their genomic information. Figure 3 shows how to choose a small group of patients
- After users choose some patients, then they can choose gene to compare genomic information as shown in Figure 37b
- The users can interact with the visualisation by zooming, rotating, finding a patient, choosing patients, updating gene list, and hiding or showing patients' label.
- The users can also change the attribute to show visualisation based on the different attribute. The default attribute is Gender.

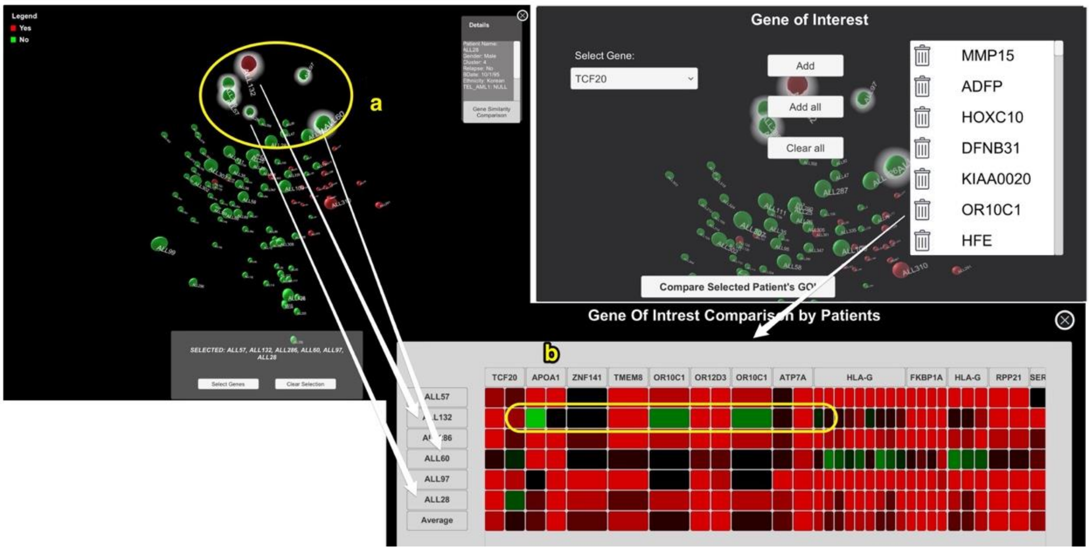


Figure 37 Visualisation tool that run on mobile devices.

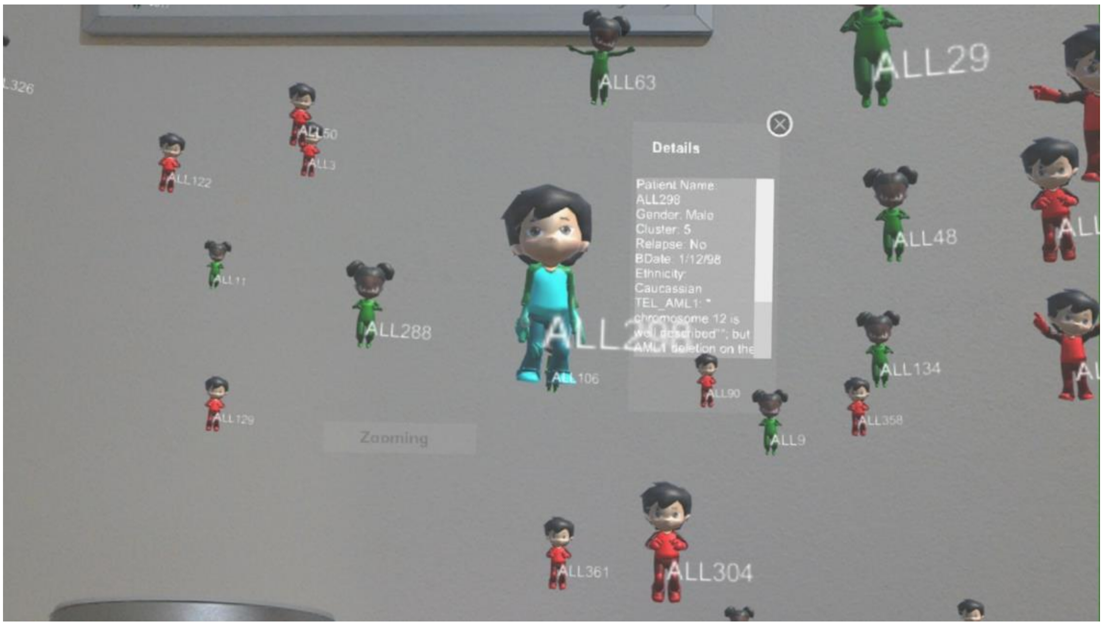
An example showing the patients in the similarity space. The left figure shows a group of patients (circled in yellow) that are mostly relapsed (shown with glowing border). The user can conduct the detailed gene study of selected a non-relapsed patients ALL132 and other relapsed ones (ALL 57, ALL386, ALL60, ALL97, ALL28), and then select all the genes of interest. (b) The 2D Heatmaps illustrates the differences in gene expression value of ALL132 in comparison of others.

VR Visualisation Tool that runs on Microsoft Hololens developed by Lau et al. is very similar to the one that runs in tablet tool (Lau et al., 2019).

Microsoft HoloLens as shown in Figure 38 is used in this tool, which is a pair of mixed reality smart glasses developed by manufactured by Microsoft. HoloLens gained popularity for being one of the first computers running the Windows Mixed Reality platform under the Windows 10 operation system and it can trace its lineage to Kinect, and add-on for Microsoft's Xbox gaming console that was introduced in 2010. Main Features:

- Zoom – Hand movement to move the hologram
- Normal – To rotate the hologram
- Showing ID – Display all the patient ID
- Hiding ID – Hide all the patient ID
- Patient On – Details information screen for patient visible
- Patient Off – Details information screen for patient invisible

- Big Data – Big data set (1000 nodes)
- Medium Data – Medium size data set (500 nodes)
- Small Data – Original data of 100 nodes.
- Goodbye – To exit the program.



Details

Patient Name: ALL298
Gender: Male
Cluster: 5
Relapse: No
BDate: 1/12/98
Ethnicity: Caucassian
TEL_AML1: "chromosome 12 is well described"", but AML1 deletion on the

Figure 38 A Visualisation on mixed reality Microsoft HoloLens.

3.2.2 Interview Feedback

This section summarises the qualitative feedback from the five domain experts. All identity information of the participants has been removed excluding their genders. The interview form is shown in the Appendix section.

Participant 1:

Participant 1 tried our three tools and he thought that our visualisation tools were useful in capturing similarities and individual difference among patient genetics. Participant 1's favourite tool is the 3D Scatter Plot Data Visualisation tool that runs on mobile devices because it can check information from distance. He said that the Medical Visualisation tool on Windows is very useful and the VR Data Visualisation tool on Microsoft HoloLens is potentially useful and worth to do it.

Participant 1 also gave some comments about how to make our tools better, which are: i) the UI in windows version is too complex and text need to be bigger; ii) in the windows version and mobile device version of tools, similarity window should also show all the data of each patient; iii) the VR version of the tool relies on the environment but has a potential future.

Participant 2

Participant 2 works on personalised medicine. She needs to compare one gene among different patients to see how a targeted drug designed for the specific gene affects the gene or not in her work. They do not have a much big cohort, that is why they need to compare their RNA data with the database from the Saint Jude hospital in America. And she is not familiar with disease dataset because her work does not focus on genomic data.

Participant 2 tried our three tools and she thought that our visualisation tools were useful in capturing similarities and individual difference among patient genetics. She would like to use the tool to compare genes with our Heatmap graphs that are very popular in the medical industry in her work. She tried her favourite gene named ROS1 and tried to compare the gene among several patients whose positions are far or close with each other to see the colourful Heatmaps she is familiar with. She hoped to see the patients' label information in the 3D scatter plot to recognise a specific patient, and we do have this feature to satisfy her.

Participant 2 also gave some comments about how to make our tools more friendly, which are: i) the gene should be ordered by alphabet order for finding a specific gene easily; ii) after choosing a gene, she hoped that the gene's location should be memorized when she went back to the scroll menu again, in a way that she could continue to choose from the same location; iii) she is curious about the location of the patients and hopes to have such information or explanation in the tool.

Participant 3

Participant 3 is a 1st-year PhD student with a biological background and works on medical datasets. His research focused on medical data mining and statistics, and he will also do some machine learning algorithms such as random forest and SVM. He is very familiar with disease dataset.

Participant 3 tried our three tools and he thought our visualisation tools were useful in capturing similarities and individual difference among patient genetics. But he would not use the tools in his work because his research topic is different. He mentioned that besides general overview of the cohort, the user might be interested in other features since there are a lot of attributes for each patient, and the tools do show to the user each patient details by clicking a patient bubble. He is satisfied with this feature. He agreed that our visualisation tools would be useful to enable personalised medicine in the way to compare and identify patient among a cohort.

Participant 3 also gave a comment about how to make our tools better, which is: The specific patient group, that the users might be interested in, should be highlighted to be found easily, for example, the closest patient, the furthest patient.

Participant 4

Participant 4 is a clinician and on training to be a paediatric oncologist and he is very familiar with disease datasets. He needs to investigate one treatment method affects some genes or not, if not, another treatment applies and the related genes are investigated again.

Participant 4 tried our three tools and he thought our visualisation tools were useful in capturing similarities and individual difference among patient genetics. He thought our tools would be useful to compare pre-treatment to post-treatment genomic data for the same patient to document molecular remission. He said our tools are definitely potential in the future and he would like to consider using the tools when they are more mature. For the personalised medicine, he thought grouping patients based on genomic data and finding a suitable treatment accordingly in the right strategy would help in treatment, and also help to move to other regimens to spare futile therapy of toxicity. It is always a better idea to understand abstract genetic information in visual format.

Participant 4 also gave some comments about how to make our tools more user-friendly, which are: i) put all the visualisation graphs on one screen to make the tool more user-friendly; ii) he hopes he can compare one patient’s genomic data in different stages to see different effects from different treatment methods. But our current tools can only compare one result for one patient.

Participant 5

Participant 5 is a data scientist and bioinformatician. And he is involved in projects that attempt to solve biological questions in the cancer biology field. He deals with a variety of research data such as genomics, epi-genomics, transcriptomics, proteomics, pharmacoomics, and cell imaging. He is very familiar with the disease data.

Participant 5 tried our three tools and he thought our visualisation tools were useful in capturing similarities and individual difference among patient genetics. He thought that the tools abstracted important variables that could predict the outcome of treatment. Visualising the similarities is valuable for giving more deeper insight as to the reasons. By zooming in and out, users can visualise the distances between clusters. For the potential using the tools in the future, he thought our tools take out the complexity in the data and simplify the patterns, which is very useful. He also thought our tools enable clinical information to ask fundamental questions related to a patient’s treatment options. As an informatician to delivery and make a decision in a group, he mentioned our mobile device tool might be useful to carry information during travel and VR tool might be very useful to present information in a group.

Participant 5 also gave some comments about how to make our tools more user-friendly, which are: i) VR device—HoloLens is too heavy to make arms get tired from over-exploring the 3D space; ii) Ipad version needs to work on improving depth perception (the cube concept is a good idea to choose in 3D space); iii) the Ipad screen is slightly too small for him.

3.3 DISCUSSION

We analysed the feedback from the small pilot group study on genomic visualisation tools as described in the above section. One of them mentioned that genomic visualisation tools take out the complexity of the data and simplify the patterns, which is very useful for the current work. For the targeted medicine and personalised treatment research, genomic and cancer data visualisation tools have advantages of being able to see and explore the patients’ data in the cohort and to assist better decision making accordingly.

The participants also indicated the further features they preferred to have in the genomic and cancer data visualisation tools. First, they would like to have all information on only a single window when necessary instead of jumping among different windows because they would like to see all the information in one screen. Second, more interactions are needed because users would like to add new data easily and see more detailed information in a few interactions. Third, AI is very attractive for the prediction feature. Analysis based on AI algorithms is a trend. Fourth, VR is very useful to present information in a group to assist with team decision. Fifth, visualisation tools running in mobile devices are useful but limited on the small screen. And last, the different role in the medical industry uses tools in the different way, some like scatter plots while others like heatmaps; some like to compare only one gene among different patient while others like to compare several genes for only one patient in different treatment stage.

In summary, genomic and cancer data visualisation tools are essential to facilitate decision-making for treatment methods or targeted medicine. New technologies have been used in recent years to create visualisation tools that can explore complex genomic data. Further efforts are needed to develop new tools to meet the changing needs of the field.

Chapter 4: Research Design

Based on the knowledge and insights obtained from the structured review and the qualitative study that is stated in Chapter 3, we design the intelligent visualisation model, including choosing the suitable machine learning algorithm and the development tools to execute the design. This chapter describes the design adopted by this research to achieve the aims and objectives stated in Section 4.1 of Chapter 4. Section 4.2 discusses the methodology used in the study, the stages by which the methodology was implemented, and the research design. Section 4.3 describes the development process.

4.1 AIMS AND OBJECTIVES

For the multidimensional genomic cancer data, it is a challenge to explore them in an effective and meaningful way, especially combined them with machine learning algorithms to find the insight knowledge and get a reasonable prediction. Although machine learning has extraordinary predictive abilities, the machine learning models and the algorithms are hard to be understood and maybe even harder to be trusted, especially in serious industries such as the medical industry (Patrick Hall, 2017). Visualising machine learning models and predictive results in a meaningful way can interpret the complex algorithms and help clinicians, researchers and experts understand and trust the predictive results. This research will focus on using machine learning to support intelligent visualisation for the current genomic data. Machine learning can help to visualise data in a cohort to assist doctor’s decision, and then improve the lives of people facing similar genetic problems.

The overall research aim is derived from the specific research question about how to utilise machine learning algorithms for improving data visualisation to present genomic data intelligently. The ultimate research goal is to develop an intelligent visualisation prototype to utilise machine learning algorithms for presenting genomic data and assisting decision making effectively as well as improve the human trust on machine learning outputs. The visualisation will illustrate data interactively, insightfully and predictively. The research will choose data to train the machine learning model and get reasonable predictions for selected attributes, and then visualise them in a cohort to assist doctors' decision.

4.2 METHODOLOGY

This research is to analyse and visualise the childhood cancer genomic data with machine learning methods to guide personalised treatment decisions and illustrate the visualisations.

To achieve the proposed goals, a task-by-task, three tasks are implemented. The first task is to create a model to bridge intelligent visualisation and the machine learning algorithms. The second task is to create a machine learning model to apply decision tree algorithm to genomic data. The last task is to develop multi-view scatterplot visualisation combined with a machine learning algorithm to illustrate information effectively and predictively.

4.2.1 Bridging Data Visualisation with Machine Learning Framework

A framework is developed in the first stage to bridge genomic data visualisation and machine learning algorithms. Machine learning can assist the process of data visualisation, meanwhile, visualisation can drive machine learning processes. 3D scatter plot is used for the visualisation. Scatterplot graph has quite a lot of benefits to present data because it allows visualising the multi-attributes with different visual features.

Machine learning combined with data visualisation should have three stages: developing an algorithm, applying genomic data to the algorithm, and predicting new unlabelled data (Libbrecht & Noble, 2015). Figure 39 shows the framework of an intelligent visualisation and machine learning combined model. It has four parts: data, machine learning model, visualisations and users. In this thesis, the users are usually researchers or clinicians. A training genomic dataset is used to train a machine learning model, the machine learning model uses Iterative Dichotomiser 3 (ID3) decision tree algorithm (Wilson, 2008). A statistic 3D scatterplot (Tibco, 2018) is also developed based on the dataset, the scatter plot can interact with both users and the machine learning model.

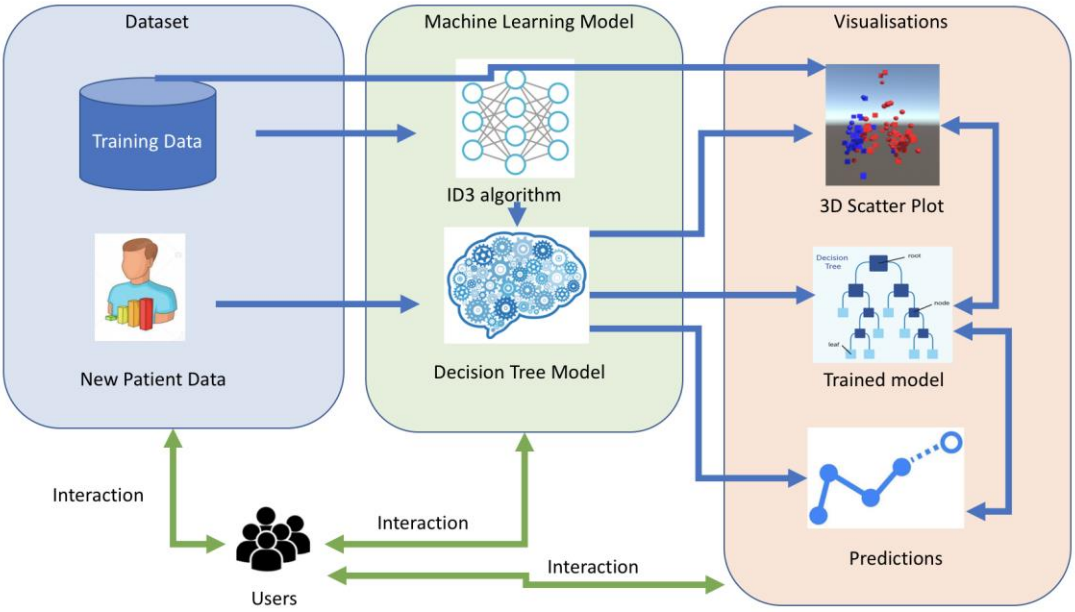


Figure 39 Intelligent visualisation and machine learning framework.

The training data is used to train a machine learning model and is visualised in a 3D scatter plot. The users can input new patient data to machine learning model and get a real-time visualised prediction to assist their decisions.

The data part phase includes two different types of data: training data and new patient data. Training data is used for training machine learning model with the ID3 algorithm and drawing 3D scatter plot, while the new data is inputted to trained machine learning model and the model will give real-time predictions. The users interact with the data by inputting new patient data for prediction and add the new patient data to the training dataset for the future training process.

The machine learning model uses decision tree ID3 as an algorithm and can be interacted with the visualisation part and users. Users can choose, expand, and fold the trained tree model. The trained tree model can be visualised by a tree plot and a 3D scatter plot, and the prediction process and results are also visualised as a tree plot. The machine learning model will be introduced in more detail in the machine learning model section.

The visualisation part illustrates all the patients' data in a 3D scatter plot, the machine learning model in a tree plot and the real-time prediction results in another tree plot. The tree visualisations are used to interpret the prediction process and illustrate the possible choices after implementing machine learning solution. The machine learning model can be interacted with the 3D scatter plot by choosing different branches and highlighting the related group in the 3D scatter plot. The statistic scatter plot shows all the patients' data in the cohort by default, and if you choose a branch of the tree model, the related group of patients will start to spin. The user, for example, a clinician, can input new patient data to the machine learning model to get prediction results in real time. The trained model and predicted results are both illustrated by tree plots based on the selected attributes.

4.2.2 Machine Learning Algorithm Model

A predictive machine learning model for the genomic data has been developed in the second stage. Decision tree algorithms are used to create the machine learning model. The model could be trained with existed data and predict the likely future possibilities. The decision tree models are written in C# programming language.

Figure 40 shows a machine learning model for genomic data. The machine learning model for genomic data includes two parts: a batch training part and a real-time prediction part. In the batch training part, the existing data are chosen to train the machine learning model. In the first stage, the ID3 (Jearanaitanakij, 2005) decision tree algorithm is used because it is suitable for the current genomic data prediction and easy to interpret (Badr Hssina, 2014). Other models will be used in the future if new structured data is used or new predictions are needed. The historical data is used to trained decision tree model. After the model is trained and a new patient data has arrived, real-time predictions occur and become illustrated. The prediction results are shown in a tree view which is part of the decision tree model to show the decision process. The new patient data can be added to the trained data to make the trained model more accurate in the future.

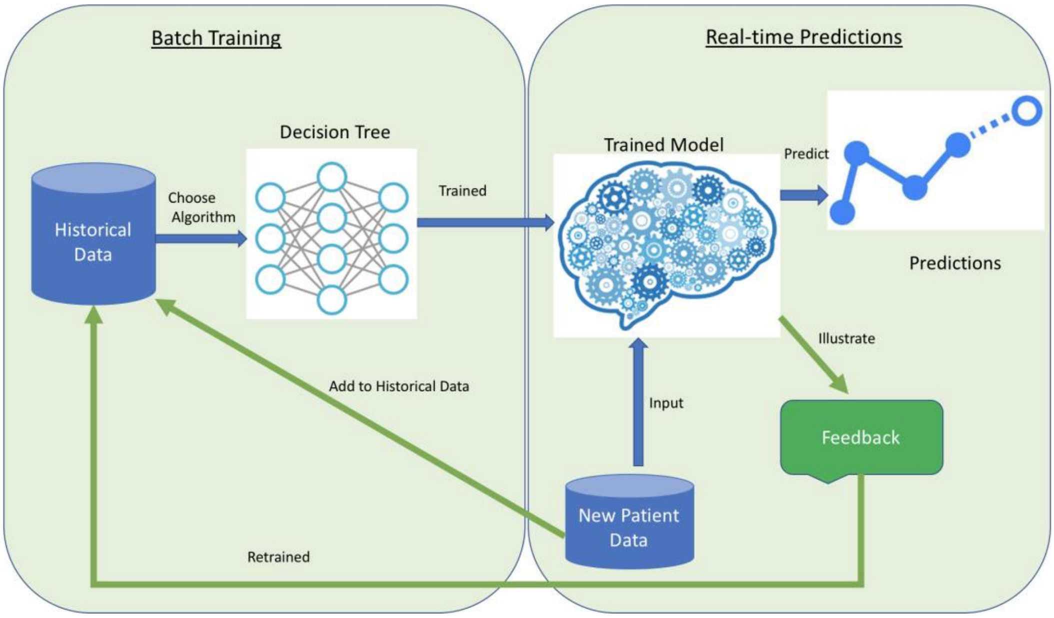


Figure 40 Machine learning model.

Decision tree algorithm transforms raw data into rule-based decision-making trees. It is a tree structure in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree is commonly used for gaining information for the purpose of decision-making. It starts with a root node, which is for users to take actions. From this node, users split each node recursively according to the decision tree learning algorithm. The final result is a decision tree structure in which each branch represents a possible scenario of the decision and its outcome (Song & Lu, 2015).

Decision tree learning algorithm has been successfully used in the expert systems in capturing knowledge. The main task performing in these systems is using inductive methods to the given values of attributes of an unknown object to determine appropriate classification according to decision tree rules. ID3 is a simple decision tree learning algorithm developed by Ross Quinlan(1983) (J. R. Quinlan, 1986; R. Quinlan, 2018). The basic idea of the ID3 algorithm is to construct the decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. In order to select the attribute that is most useful for classifying the given sets, ID3 use a metric—information gain (Li, Lei, Zhao, Zhang, & Han, 2013).

The ID3 algorithm begins with the original set *S* as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates the entropy *H(S)* (or information gain *IG(S)* of that attribute. It then selects the attribute which has the smallest entropy (or largest information gain) value. The set *S* is then split by the selected attribute to produce subsets of the data. The algorithm continues to recur on each subset, considering only the attributes never selected before. Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch.

Entropy *H(S)* is a measure of the amount of uncertainty in the (data) set S (i.e. entropy characterises the dataset S).

$$H(S) = \sum_{x \in X} - p(x) \log_2 p(x)$$

where *S* is the current (data) set for which entropy is being calculated (changes every iteration of the ID3 algorithm). *X* is the set of classes in *S*. *p*(*x*) is the proportion of the number of elements in class *x* to the number of elements in *S*. When *H*(*S*)=*0*, the set *S* is perfectly classified (i.e. all elements in *S* are of the same class). In ID3, entropy is calculated for each remaining attribute. The attribute with the smallest entropy is used to split the set S on this iteration. The higher the entropy, the higher the potential to improve the classification here.

Information gain *IG*(*A*) is the measure of the difference in entropy from before to after the set *S* is split on an attribute *A*. In other words, how much uncertainty in *S* was reduced after the splitting set *S* on attribute *A*.

$$IG(A,S) = H(S) - \sum_{t \in T} p(t) H(t)$$

where *H(S)* is the Entropy of set *S*, *T* is the subsets created from the splitting set *S* by attribute *A* such that

$$S = \bigcup_{t \in T} t \quad ,$$

where *p(t)* is the proportion of the number of elements in *t* to the number of elements in set *S*, *H(t)* is the entropy of subset. In ID3, information gain can be calculated (instead of entropy) for each remaining attribute. The attribute with the largest information gain is used to split the set S on this iteration

4.2.3 Visualisation Prototype

Interactive visualisation prototype is developed in the third stage to link the machine learning with visual analytics. Scatter plots are used in this stage to illustrate the data in this implementation. Visualising genomic data exploration process is incorporated into machine learning techniques to enable users to steer and drive the computational algorithms. User interactions with the system are designed and implemented as mechanisms by which users can augment the visualisation parameters, filter data, and other direct changes to the application.

Unity3D is used to implement the prototype in this stage, and C# is chosen in the Unity3D to develop machine learning algorithms. Unity3D is a tool to create games originally, and now are widely used to visualise big datasets for its ability to effectively visualise more than two or three dimensions and virtual reality technologies. At the moment, Unity 3D appears to be the most used integrated development environment (IDE) in hyper-dimensional data exploration, the nature of virtual reality data visualisation, and the ability to export application to mobile devices. C# is one of the unity 3D scripting languages which also includes basic AI libraries.

The development process should include the following steps. First, the visualisation incorporates a usability study to evaluate the effectiveness and feasibility of the proposed framework, within a real-world application. Second, the visualisation prototype uses machine learning methods to enhance data visualisation that can intelligently display the genomic information predictively and on screens effectively. Third, the visualisation prototype also includes the interaction to allow transitioning between various views on the data. By looking at the information in various perspectives, we can gain better understand the relationships within the patients and to help improve the understanding and information conveyed. The details of the development process are described in Section 4.3.

4.3 DEVELOPMENT PROCESS

4.3.1 Design Evolution

Before Unity3D is used to develop our prototype, R is used to develop our demo to analyse the feasibility. R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and

MacOS. A structured set of 101 patient data Rhabdomyosarcoma (RMS) dataset is used as input to our demo.

Step 1 - Use R to show basic 3D scatter plot:

The library “plotly” in R (https://plot.ly/r/) is used to draw a 3D scatter plot as shown in Figure 41. The left scatter plot used colour to show different “Histology” which are “ARMS” and “ERMS”. The right scatter plot used colour to show patient “status” which are “alive” and “dead from disease”. From the scatter plot, we can see that the attribute “Histology” divided patients in very clear two groups. As a result, “Histology” is chosen as one of our decision tree attributes.

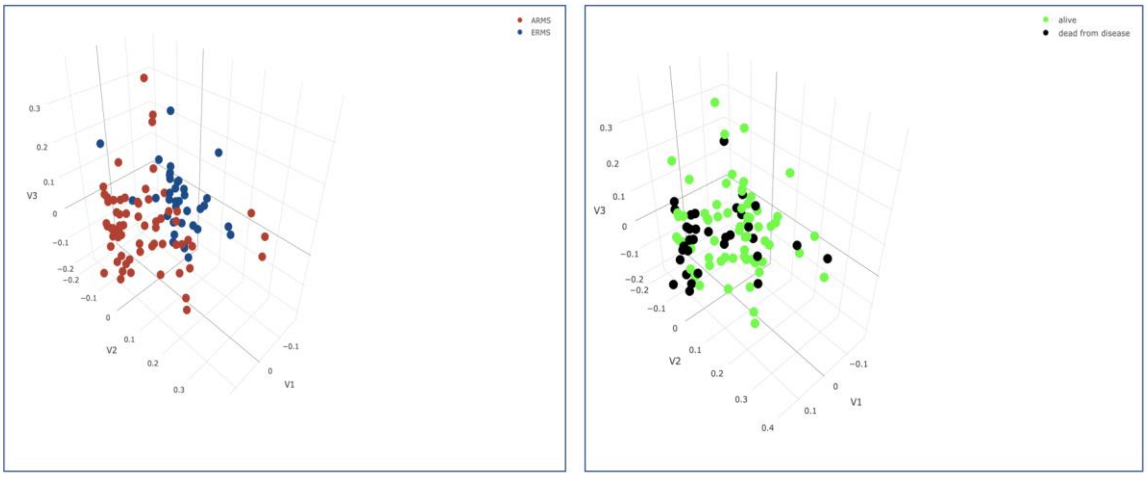


Figure 41 Scatter Plot with R.

The left scatter plot used colour to show different “Histology” which are “ARMS” and “ERMS”. The right scatter plot used colour to show patient “status” which are “alive” and “dead from disease”

Step 2 - Draw decision tree in R and by manual:

The library “rpart” in R to draw our decision tree. “Status” is chosen as the prediction attribute and “Histology”, “Sex”, “AgeStatus” are chosen as input attrubutes to draw a decision tree as shown in Figure 42. And then another decision tree is manually drawn as shown in Figure 43. The two decision trees are then compared and the numbers on each node are same.

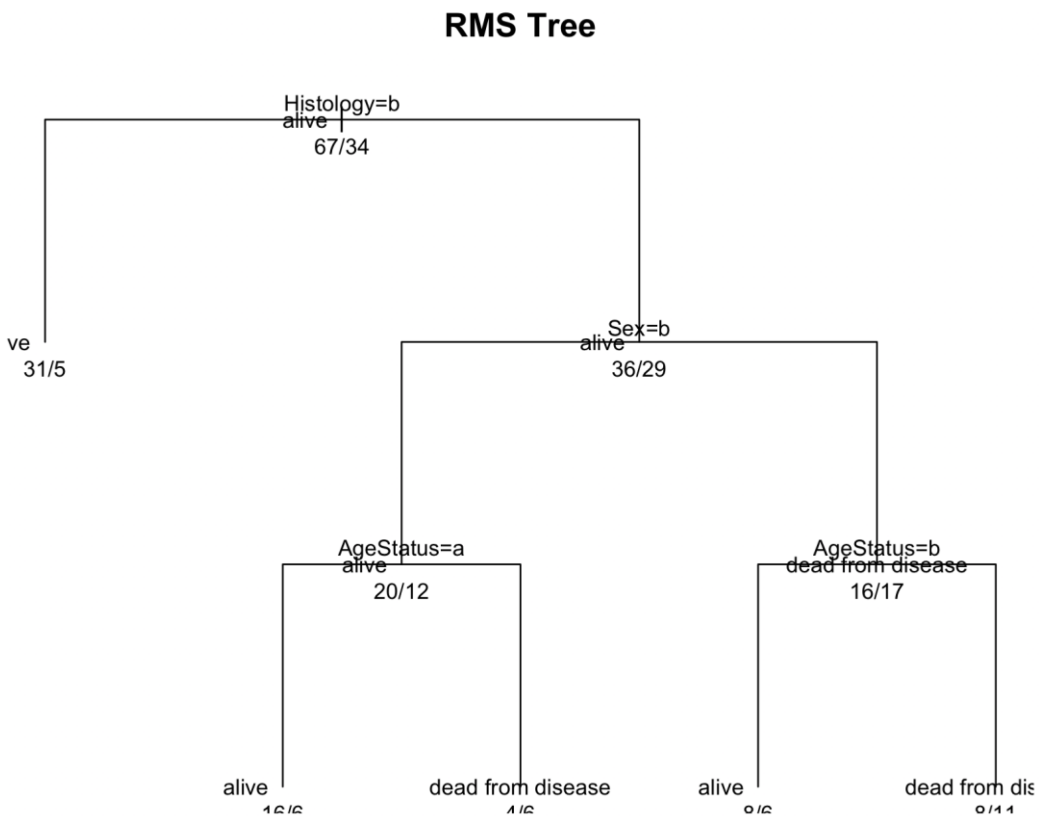


Figure 42 Decision tree drawn in R

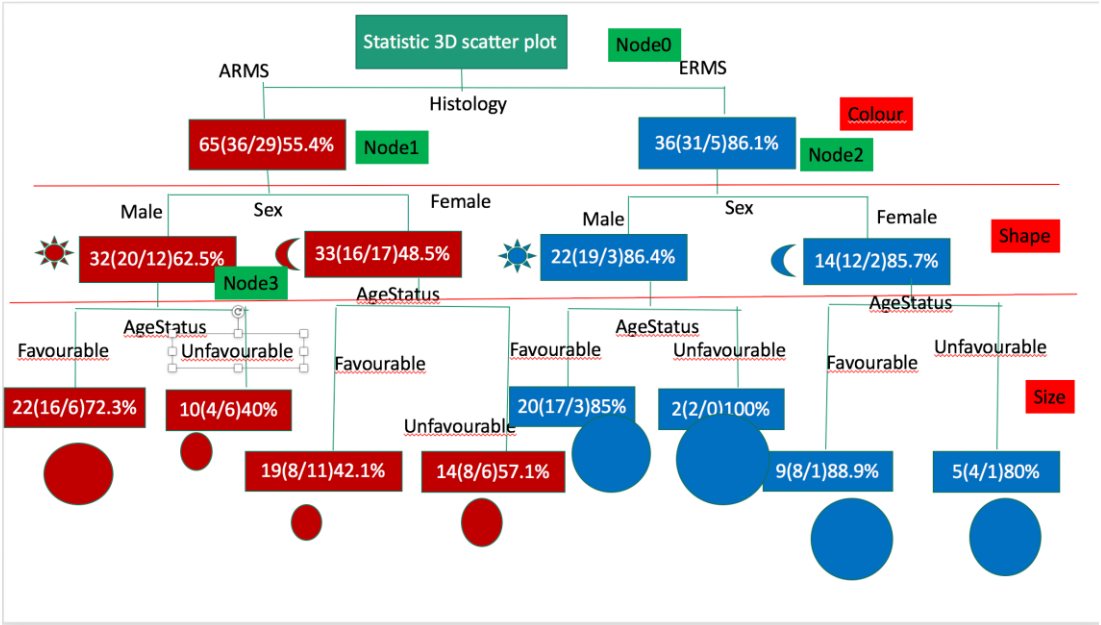


Figure 43 Decision tree drawn by manual

The users can choose different attributes to draw the tree, the Figure 44 is another tree that uses Age as the node to split the dataset.

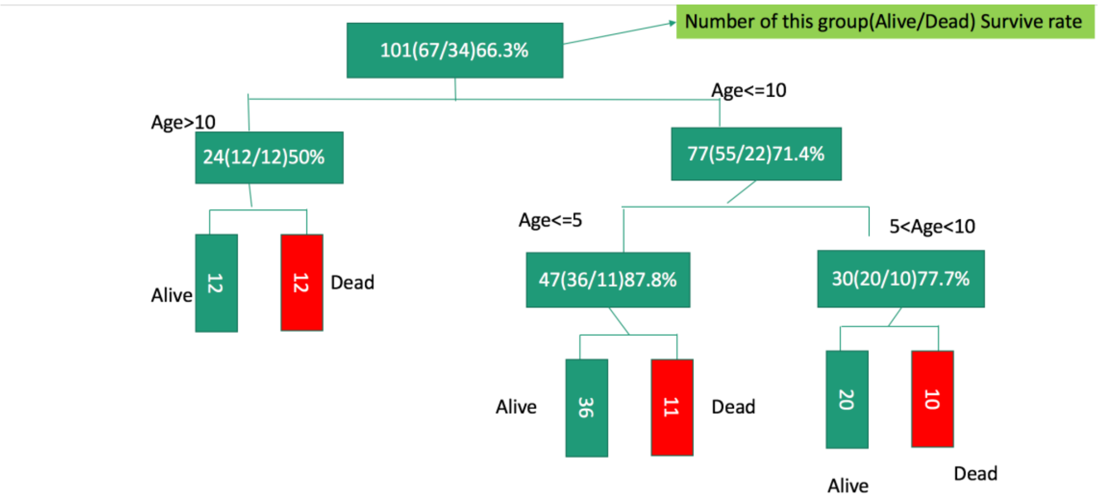


Figure 44 Tree used Age as node

Step 3 - Draw an animated 3D scatter plot in R:

The library "plot_ly" is used to draw another animated 3D scatter plot as shown in Figure 45. The colour red and blue is used to stand for "Histology" which are "ARMS" and "ERMS". The shapes square and circle are used to stand for "Sex" which are "male" and "female".

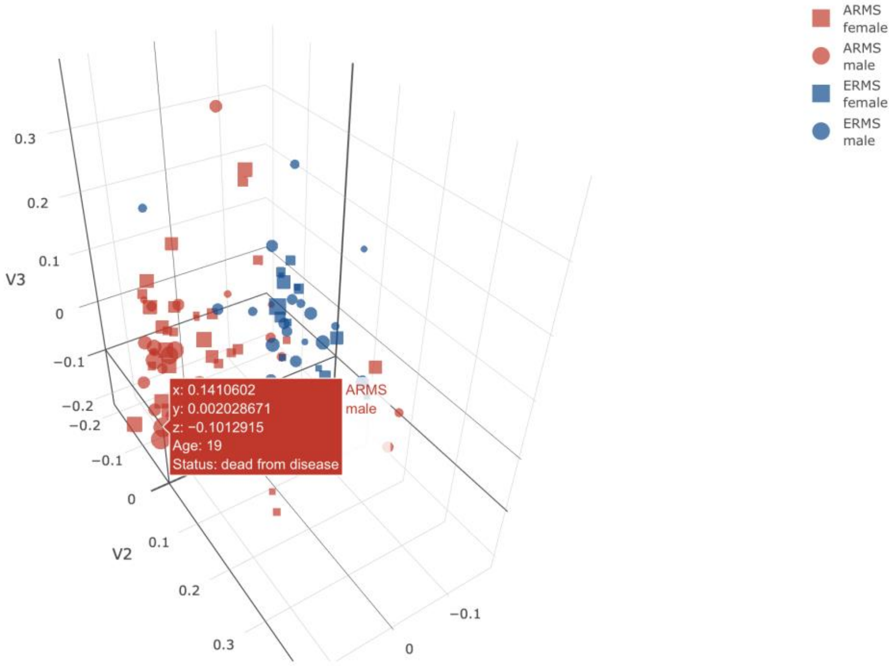


Figure 45 animated 3D scatter plot drawn by R

Step 4 - Draw 3D scatter plot in Unity 3D:

At last, Unity3D is tried to draw the 3D scatter plot as shown in Figure 46. Unity 3D is a cross-platform game engine developed by Unity Technologies, and it is the creator of the world's most widely-used real-time 3D (RT3D) development platform, giving content creators around the world the tools to create rich, interactive 2D, 3D, VR and AR experiences. In our prototype, the colour red and blue is used to stand for "Histology" which are "ARMS" and "ERMS". The shape cube and sphere are used to stand for "Sex" which are "male" and "female".

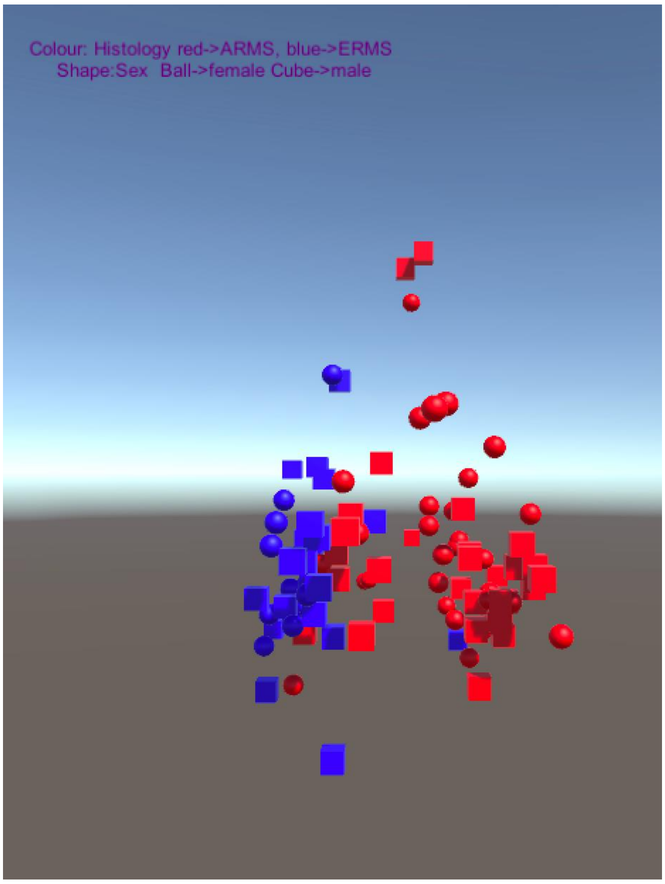
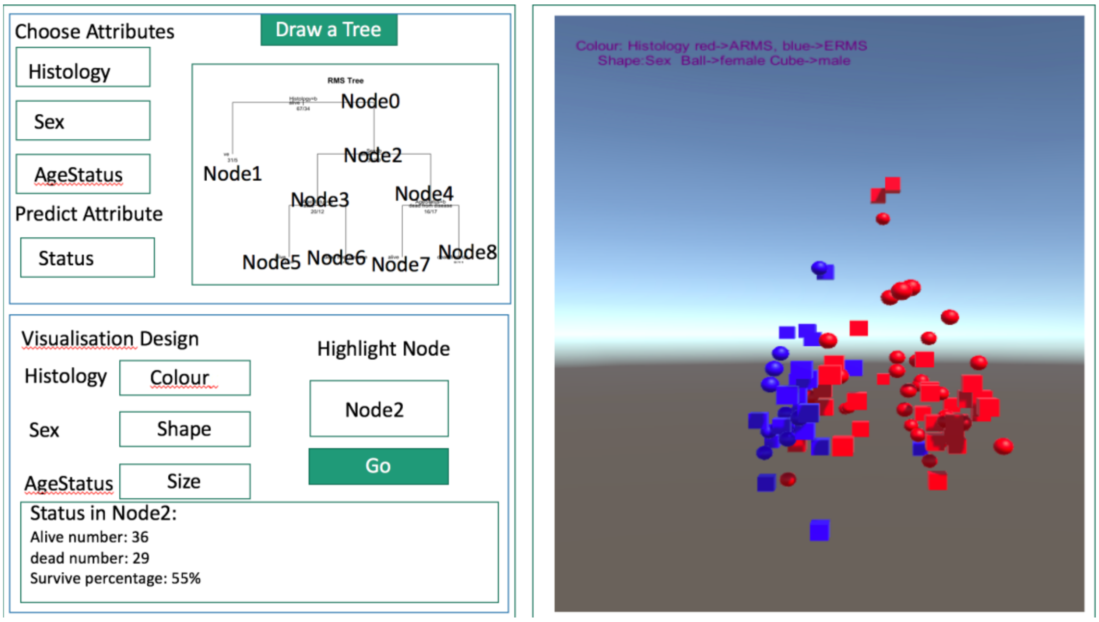

Colour: Histology red->ARMS, blue->ERMS
Shape:Sex Ball->female Cube->male

Figure 46 3D scatter plot drawn by Unity3D

Step 5 - Design User Interface (UI)

From step 1 to 4, it can be found that there are clear patterns in the data, but the boundaries for delineating them are not obvious. Finding patterns in data is where machine learning comes in, machine learning methods use statistical learning to identify boundaries. A User Interface (UI) is drawn by manual as shown in Figure 47. The visualisation prototype is able to change their views at any given point to display the information in a more relevant approach for the application as needed. Scatter plots will be used in the visualisation prototype. User interactions are designed and implemented in this visualisation prototype to augment the visualisation parameters, and filter data and other direct changes to the application. In addition, user interactions for machine learning are also designed to adapt the predictive machine learning model.



Status in Node2:
Alive number: 36
dead number: 29
Survive percentage: 55%

Figure 47 An example of manual UI design

For the visualisation, users can choose a decision tree node to show the specific group of data in highlighted. And also choose a visualisation feature to show them such as colour, size, shape, etc. The visualisation prototype process is shown in Figure 48, and the final version is Figure 39.

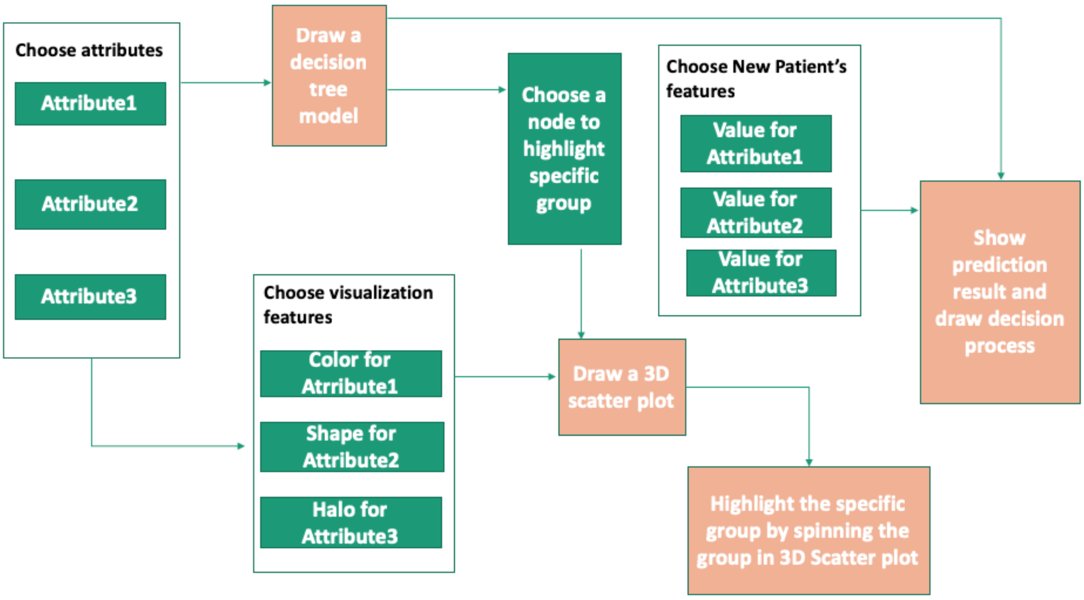

Choose attributes
Attribute1
Attribute2
Attribute3
Draw a decision tree model
Choose a node to highlight specific group
Choose New Patient's features
Value for Attribute1
Value for Attribute2
Value for Attribute3
Show prediction result and draw decision process
Choose visualization features
Color for Atrribute1
Shape for Attribute2
Halo for Attribute3
Draw a 3D scatter plot
Highlight the specific group by spinning the group in 3D Scatter plot

Figure 48 A visualisation process, the final version is shown in Figure 39

4.3.2 ID3 - A Decision Tree Algorithm

The machine learning model will use the decision tree as the algorithm in the first stage because the decision tree as tree-based algorithm empowers the predictive model with high accuracy, stability and ease of interpretation. The machine learning model is trained by the existed genomic data independently of all assumption and finds out patterns hidden in data, and then, when new data comes, the model will give predictive results. Predictive power is the key of the machine learning model which is different to traditional statistical models.

For inductive learning, decision tree learning is attractive for three reasons: i) good generalisation for unobserved instance, ii) efficient in computation, and iii) rendering the classification process self-evident. The training data may contain errors. This can be dealt with pruning techniques that this thesis does not cover. This is because the used datasets are accurate in our implementation.

The three widely used decision tree learning algorithms are ID3 (J. R. Quinlan, 1986; R. Quinlan, 2018), CART (a Classification And Regression Tree) (Breiman, Friedman, Olshen, & Stone, 1984) and C4.5 (J. R. Quinlan, 1986; R. Quinlan, 2018). They have slight differences as shown in Table 3 (Sonia Singh, 2014). ID3 is chosen as our decision tree model because pruning is not needed as all the data is accurate and no missing values are in our datasets.

	Splitting Criteria	**Attribute type**	**Missing Values**	**Pruning Strategy**	**Outlier Detection**
ID3	Information Gain	Handles only Categorical value	Do not handle missing values	No pruning is done	Susceptible to outliers
CART	Towing Criteria	Handles both Categorical & Numeric value	Handle missing values	Cost-Complexity pruning is used	Can handle Outliers
C4.5	Gain Ratio	Handles both Categorical & Numeric vale	Handle missing vales	Error Based pruning is used	Susceptible to outliers

Table 3 Differences of decision tree model

(Sonia Singh, 2014).

Chapter 5: Case Studies

We developed our intelligent visualisation prototype based on the structured and qualitative studies stated in Chapter 3, especially the domain users' feedback on the genomic data visualisation tool preliminary study. This chapter also illustrates the models and the intelligent visualisations in Chapter 4 through the two case studies. In this prototype, we focused on visualising, framing and evaluating machine learning model and the prediction process. Intelligent data visualisation tools are needed to find the relationship between genomic data and diseases and aid in the process of targeted and personalised therapy. The current statistical analysis methods are not enough for achieving better data insights. Application of machine learning and data visualisation has become more attractive in genomic data analytics. Intelligent visualisation combined with machine learning algorithms for genomic data is a big challenge and is becoming a new trend in the genomic visualisation evolution. Our prototype illustrates not only traditional genomic data visualisation but also the machine learning model and the prediction process. We put all the visualisations on only one screen and added interactions among different visualisations.

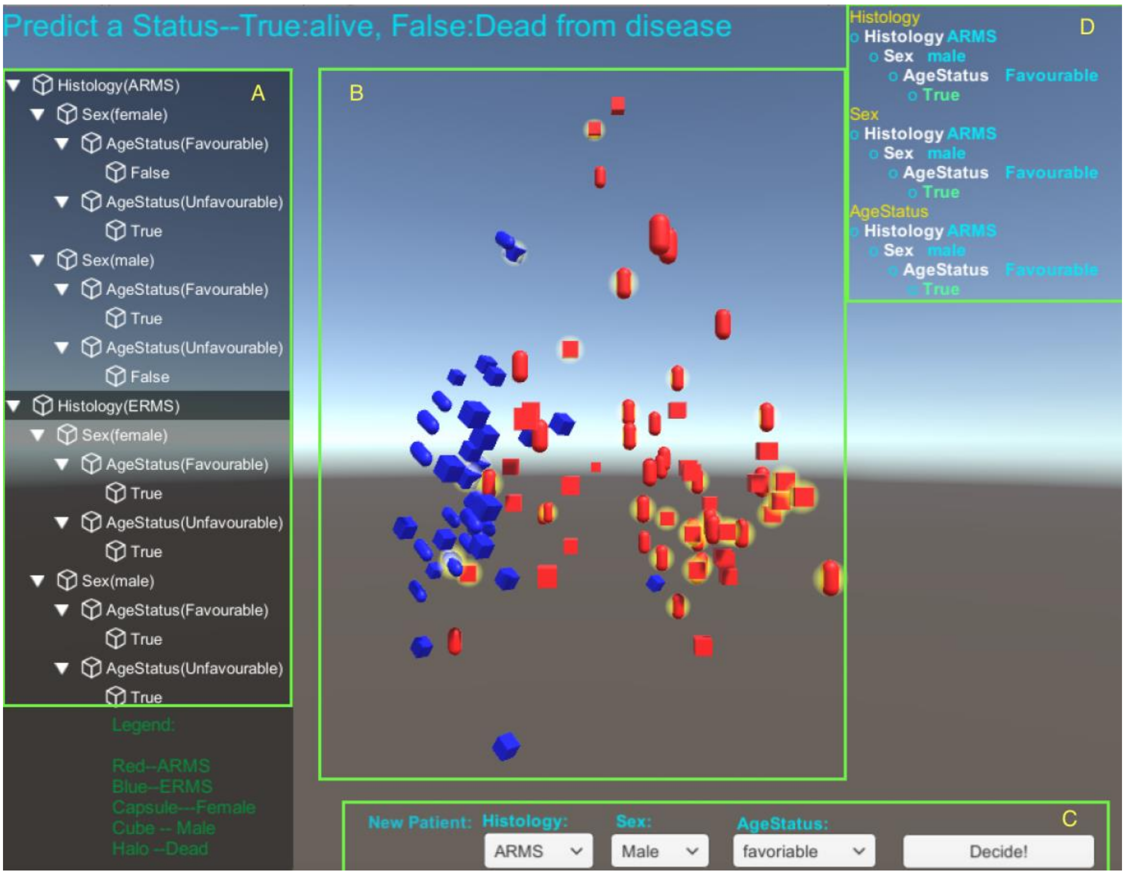
We applied two genomic cancer datasets to our prototype and stated in the following two case studies: RMS dataset visualisation in (section 5.1) and ALL dataset visualisation in (Section 5.2).

5.1 CASE STUDY 1—RMS DATASET

We used a structured set of 101 patients' data Rhabdomyosarcoma (RMS) dataset from the Westmead Children Hospital. RMS is the most common soft tissue childhood sarcoma with an incidence rate of 17 new cases per year in Australia (Wachtel et al., 2006). The two major histological subtypes of RMS are alveolar (ARMS) and embryonal (ERMS). ERMS patients have a more positive prognosis. This difference in prognosis has led researchers to use molecular markers with the aim of developing more accurate classifiers of RMS subtypes. The prototype is developed with Unity3D (Technologies, 2018) and C# programming language is chosen in Unity3D to develop machine learning algorithms. We use SQLite (SQLite, 2018) to manage the training data. As we only have 101 patient data and they are certain to be accurate, we used all of them to train our decision tree model and then show all of them in a 3D scatter plot. For example, when a new patient genomic data comes, we wanted to predict the patient status as either alive or dead based on three attributes: Histology, Sex, and AgeStatus. In

Figure 49，we visualised the trained machine learning model in a tree-plot on the left panel, a 3D scatter plot for all RMS patients in the middle, and the real-time prediction results on the top right in another tree structure. The users (e.g. clinicians or researchers) can interact between the tree model and scatter plot, input new patient data on the bottom fields and get the prediction results in another tree plot on the top right. We have not included an operation of adding new patient data to re-train the machine learning model in this prototype yet.

In the 3D scatter plot, the red colour patients are ARMS patients while the blue ones are ERMS patients. The capsules stand for female patients while the cube shape ones stand for male patients. The patients with the yellow halo are dead while the patients without halo are alive. This 3D scatter plot is connected with the machine learning model tree on the left. When the user chooses a branch of the machine learning tree model, the related group of patients' shapes spin. For example, if the user wants to highlight the group of patients with Histology as "ERMS", Sex as "male", and AgeStatus as "Favouriable", then the user clicks the related branch in the tree plot (in part A), a group of blue cubes are spinning in the 3D scatter plot. If the user chooses a branch in the tree plot with Histology as "ARMS", Sex as "female", and AgeStatus as "Unfavouriable", then a group of red capsules will spin in the 3D scatter plot. The user can also choose the father branch ERMS to spin all the blue colour patients.



Predict a Status--True:alive, False:Dead from disease

- Histology(ARMS)
  - Sex(female)
    - AgeStatus(Favourable)
      - False
    - AgeStatus(Unfavourable)
      - True
  - Sex(male)
    - AgeStatus(Favourable)
      - True
    - AgeStatus(Unfavourable)
      - False
- Histology(ERMS)
  - Sex(female)
    - AgeStatus(Favourable)
      - True
    - AgeStatus(Unfavourable)
      - True
  - Sex(male)
    - AgeStatus(Favourable)
      - True
    - AgeStatus(Unfavourable)
      - True

Legend:

Red--ARMS
Blue--ERMS
Capsule---Female
Cube -- Male
Halo --Dead

Histology
- Histology ARMS
  - Sex male
    - AgeStatus Favourable
      - True

Sex
- Histology ARMS
  - Sex male
    - AgeStatus Favourable
      - True

AgeStatus
- Histology ARMS
  - Sex male
    - AgeStatus Favourable
      - True

New Patient: Histology: ARMS Sex: Male AgeStatus: favoriable

Figure 49 Visualisation Illustrates the machine learning model, 3D scatter plot, input fields and the real-time prediction result for RMS patient dataset.

A → Machine learning model tree. When a branch is selected, the related group of patients in B will spin. B → 3D scatter plot for 101 RMS patients. The Red colour patients are ARMS patients while the blue ones are ERMS patients. The capsule shapes stand for female patients while the cube shape ones stand for male patients. The patients with a yellow halo are dead while the one without halo are survived. C→ New patient data input fields. We choose three attributes which are Histology, Sex, and AgeStatus. Choose the values as the new patient data such as "ARMS", "Male", "favoriable", when the button "Decide!" is clicked, the real-time result is shown on D, in this case, the real-time prediction result is green indicating "True" which means the patient would survive.

For the machine learning prediction process and results, we illustrate them on the right top section D part. For example, in this prototype part C, we choose the values "ARMS" for Histology, "Male" for Sex, and "favoriable" for AgeStatus as the new patient data attributes. when the button "Decide!" is clicked, the real-time result is shown. In this case, the real-time prediction result is green indicating "True" which means the patient would survive. If you chose the values “ARMS” for Histology, “Male” for Sex, and “unfavoriable” for AgeStatus as the new patient data the predictive result is red indicating “False”, then the patient would die. The users can choose to show the patient ID or not. In this case study

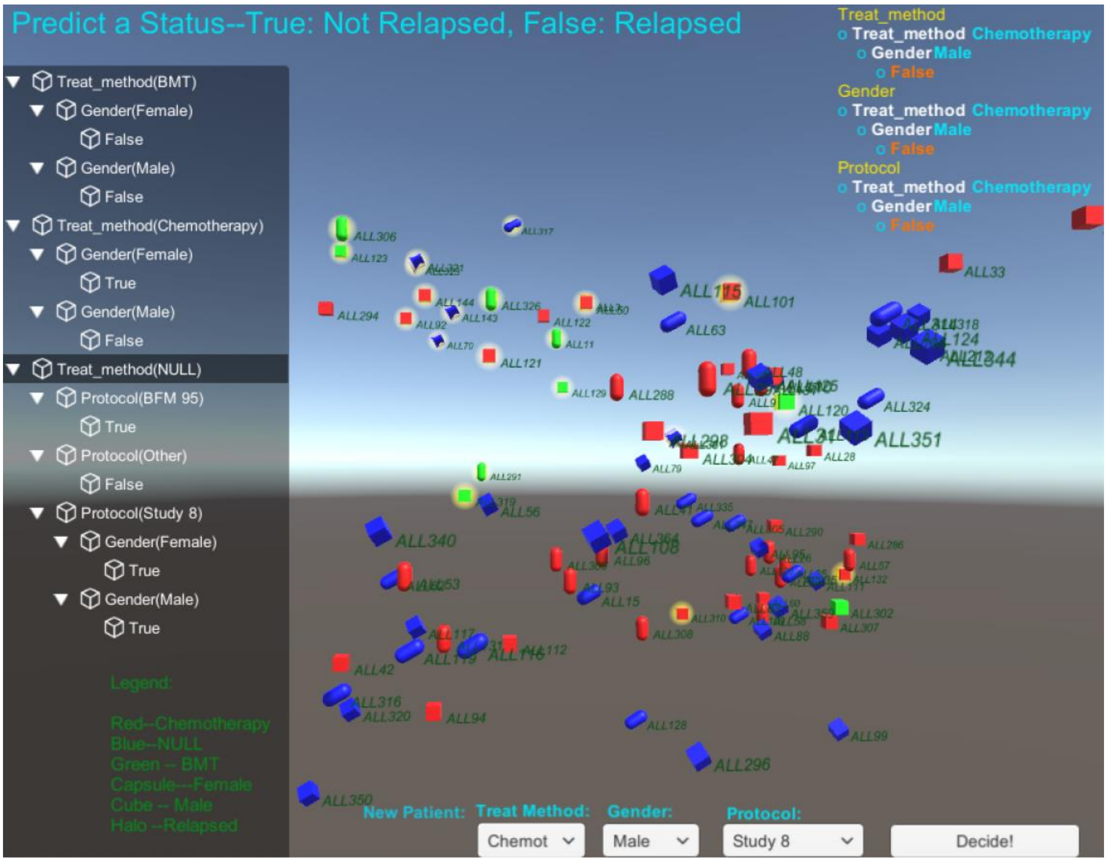
Figure 49, we chose not to show the patient ID, and we will choose to show patient ID in the next case study.

In this case study, the decision tree shown in part A was trained by all the patient data. They are all real, accurate and can all fit into the machine learning model. Each branch node represents a choice between a number of alternatives and each leaf node represents a decision. This decision tree fitted all the training examples and is fully grown to give 100% accuracy on that data. But when we checked the decision tree on another dataset, not 100% data was fit to the decision tree as happened in case study2.

5.2 CASE STUDY 2—ALL DATASET

We applied another structured dataset of the genomic expression and genomic profiles of paediatric B-cell ALL patients treated at the Children’s Hospital at Westmead to our visualisation prototype. The expression and genomic SNP profiles of pediatric B-cell ALL patients were generated using Affymetrix expression microarrays (U133A, U133A 2.0, and U133 Plus 2.0) and Illumina NS12 SNP microarrays, respectively (Q. V. Nguyen et al., 2014).

For example, in a scenario, as shown in Figure 50， when a new patient genomic data comes, we want to predict whether the patient status Relapsed or Not Relapsed based on three attributes: Treatment, Gender, and Protocol. We visualised the trained machine learning model in a tree plot on the left, 3D scatter plot for ALL patients in the middle with patient ID label (or identification) on, and the real-time prediction results on the top right in another tree plot. The users can interact between the tree model and scatter plot, input new patient data on the bottom fields and get the real-time prediction results in another tree plot on the top right.



Predict a Status--True: Not Relapsed, False: Relapsed

- ▼ Treat_method(BMT)
  - ▼ Gender(Female)
    - False
  - ▼ Gender(Male)
    - False
- ▼ Treat_method(Chemotherapy)
  - ▼ Gender(Female)
    - True
  - ▼ Gender(Male)
    - False
- ▼ Treat_method(NULL)
  - ▼ Protocol(BFM 95)
    - True
  - ▼ Protocol(Other)
    - False
  - ▼ Protocol(Study 8)
    - ▼ Gender(Female)
      - True
    - ▼ Gender(Male)
      - True

Legend:

Red--Chemotherapy
Blue--NULL
Green -- BMT
Capsule---Female
Cube -- Male
Halo --Relapsed

Treat_method
- Treat_method Chemotherapy
  - Gender Male
    - False

Gender
- Treat_method Chemotherapy
  - Gender Male
    - False

Protocol
- Treat_method Chemotherapy
  - Gender Male
    - False

New Patient: Treat Method: Chemot Gender: Male Protocol: Study 8

Figure 50 Visualisation Illustrates the machine learning model, 3D scatter plot, input fields and the real-time prediction result for ALL patient dataset.

The red colour patients used treat method “Chemotherapy”, the blue ones used treat method “NULL” and the green ones used treat method “BMT”. The capsule shapes stand for female patients while the cube shape ones stand for male patients. The patient with a yellow halo is relapsed while the one without halo is not relapsed. When a branch in the tree on the left side is chosen, the related group of patients in the 3Dscatter plot will spin. For prediction, we choose three attributes which are Treat Method, Gender, and Protocol. Choose the values “Chemotherapy”, “Male”, and “Protocol” as the new patient data, when the button “Decide!” is clicked, the real-time result is shown on the top right tree, in this case, the real-time prediction result is red “False” which means the patient would relapse.

In the 3D scatter plot, the red colour patients used treatment method “Chemotherapy”, the blue colour patients used treatment method “NULL”, and the green ones used Bone Marrow Transplantation (BMT) treatment method. The capsule shapes stand for female patients while the cube shape ones stand for male patients. The patients with the yellow halo are relapsed and the ones without halo are not relapsed. We also showed patient label beside each patient in this visualisation. This 3D scatter plot is connected with the machine learning model tree on the left. When the user chooses a branch of the machine learning tree model, the related group of patients' shapes spin. For example, if the user wants to highlight the group of patients with Treatment Method as "NULL", Gender as "female", and Protocol as "Study8", then the user clicks the related branch in the tree plot, a group of blue cubes will spin in the 3D scatter plot. If the user chooses a branch in the tree plot with the treat method as "NULL", Protocol as "Study 8", then a group of blue patients including cubes and capsules are spinning in the 3D scatter plot. The users can identify that there are only several relapsed (with halo) patients in this group. When the users select the treatment method as "Chemotherapy" branch in the left tree plot, the related red colour patients would spin. In this group, more patients were relapsed indicating with the yellow halo. The users then can choose the child branch such as the female group to be highlighted as being spun to find the pattern in this group.

For the machine learning prediction process and results, we illustrate them on the top right part with another tree plot. For example, in the bottom part of this prototype, we chose the values "BMT" for Treat Method, "Male" for Gender, and "study8" for Protocol as the new patient data values. When the button "Decide!" is clicked, the real-time result gets shown. In this case, the real-time prediction result is red indicating "False", signifying the patient would relapse. If you chose the values "Chemotherapy" for Treat Method, "Female" for Gender, and "study8" for Protocol as the new patient data values, the predictive result is green indicating "True", signifying the patient would not relapse. In this case study we choose to show patient ID in green colour beside the patient cube or capsule 3D shapes.

In this case study, the decision tree is trained by all the patient data. The data are all real and accurate, but some data is not fit to the machine learning model, which ends up less accurate decision tree. The overfitting data is caused by two major situations which are the presence of noise and lack of representative instances. In this case, the decision tree avoided the overfitting by pruning sections of the tree that provide little power to classify instances. Pruning reduces the complexity of the final classifier and improves predictive accuracy by reduction of overfitting. Moreover, the complexity of the model structure for all the patient data will decide the machine learning model tree's structure and make the tree plot different.

Chapter 6: Discussion, Conclusion and Future Work

6.1 DISCUSSION

Genomic research is critical to progress against cancer. By the study of cancer genomes, the abnormalities in genes has been revealed and drive the development and growth of many types of cancer. Genomic and cancer data visualisation tools can assist to improve our understanding of the biology of cancer and lead to new methods of diagnosing and treating the disease. Over the past decade, large-scale research projects have begun to survey and catalogue the genomic changes associated with a number of types of cancer which have revealed unexpected genetic similarities across different types of tumours. For instance, mutations in the HER2 gene (distinct from amplifications of this gene, for which therapies have been developed for breast, esophageal, and gastric cancers) have been found in a number of cancers, including breast, bladder, pancreatic, and ovarian (NIH, 2017a).

The cancer genomics research field is rapidly evolving in parallel with advances in high-throughput genomics technologies. This evolution of the field requires continuous advancement in visualisation techniques and tools. As this rapid scientific evolution continues, cancer researchers are highly dependent on computational management, analysis and visualisation of data. The conventional genomic and cancer data visualisation tools are two-dimensional and present data by changing with the creative use of colour and size, combination of space and time, and advanced computer graphics. 2D scatter plot, networks, heatmaps, and genomic coordinates are the traditional visualisation graphs used for genomic and cancer data. Most visualisation tools have these four visualisation methods, for example, IGV supports all the four visualisation methods.

Genomic and cancer data visualisation is entering a new era with emerging sources of artificial intelligence and new visual environment equipment such as VR/AR/Immersive big screen and mobile devices. New technologies and evolving cognitive framework are opening new horizons to let data visualisation getting more accurate and contextual. VR and related technologies have been adopted in the healthcare industry. Medical researchers have been exploring ways to create 3D models of patients' internal organs using VR since the 1990s. Recently, VR and related technologies are used to plan complex operations, reduce anxiety in cancer patients, and help patients overcome balance and mobility problems resulting from stroke or head injury. Virtual reality and augmented reality are primarily considered a medium for delivering entertainment to offer intriguing possibility of letting us step inside the data. 3D vision instantly broadens the available canvas and interaction become more intuitive as we can reach out to touch and manipulate what is shown to us. VR environment is expected to bring a revolution in genomic data visualisation as one could integrate meta-genomic data in virtual worlds. Approaching the problem from a different angle, VR devices such as Google Glass, HoloLens and Magic Leap offer an augmented reality experience which can facilitate the learning process of the biological systems because it builds on exploratory learning.

Genomic and cancer data visualisation tools are essential to facilitate decision-making for the treatment methods or targeted medicine. New technologies have been used in recent years to create visualisation tools that can explore complex genomic data. Further efforts are needed to develop new tools to meet the changing needs of the field.

6.2 CONCLUSION

Personalised medicine refers to diagnosis and treatment based on a person's entire DNA sequence. Variants in the DNA sequence determine the differences between individuals and differences between types of cells such as tumour cells and non-tumour cells. Genomic and cancer data visualisation tools can assist in improving our understanding of the study of cancer and lead to new methods of diagnosing and treating the disease. Personalised genomic cancer medicine uses the latest genome sequencing to look at the genetics of cancer rather than treating it based on location to allow us to understand inherited cancer risk and find more effective treatments for people with cancer (Stevens & Rodriguez, 2015).

AI is playing an integral role in the evolution of the field of genomics. Genomics is closely related to precision medicine whose market size projected to reach $87 billion by 2023 (Insights, 2016), the field of personalised medicine is an approach to patient care that encompasses genetics, behaviours and environment with a goal of implementing a patient or population specific treatment intervention in contrast to a one-size-fits-all approach. AI and machine learning have applied in genomics for analysing genome sequencing, gene editing, clinical workflow and direct-to-consumer genomics. Future applications of machine learning in the field of genomics are diverse and many potentially contribution to the development of patient or population-specific pharmaceutical drugs to look at the role of genetics in the context of how an individual responds to drugs (Sennaar, 2018). While the field is still quite new, there is evidence of research involving machine learning. For example, Tacrolimus is regarded as the first study that applied machine learning models in renal transplant patients. Tacrolimus is commonly administered to patients following a solid organ transplantation to prevent "acute rejection" of the new organ (Tang et al., 2017).

This thesis combines AI and visualisation together to assist personalised genomic data analysis, enabling by i) a systematic review of the visualisation tools, ii) a qualitative review with a group of domain experts and iii) an intelligent visualisation prototype.

We reviewed methods for genomic data visualisation including traditional approaches such as scatter plots, heatmaps, coordinates and network, as well as emerging technologies such as AI and VR; we also compare genomic data visualisation tools by time and analyse the evolution of visualising genomic data. We carried out an expert evaluation and analysed the experts' feedback about the usability of genomic data visualisation tools as well. The preliminary qualitative evaluation with domain experts is for evaluating the effectiveness and domain view-points of three genomic visualisation tools.

We have described our new visualisation prototype that not only shows the entire patient population in traditional 3D scatter plot but also illustrates, frames and evaluates a machine learning model. The visualisation links the machine learning model with the 3D scatter plot and gives real-time predictions to assist researchers or clinicians' decisions. The new visualisation tool can interpret the machine learning model to researchers or clinicians who are not experts in predictive mathematics algorithms, which makes the genomic data visualisation and decision-making procedure more reliable for them. Genomic and cancer data visualisation tools are essential to facilitate decision-making for the treatment methods or targeted medicine.

Our prototype contributes a new way for visual analytics to visualise, understand, evaluate, frame machine learning models and the prediction results.

6.3 FUTURE WORK

In the future, we will expand the machine learning algorithm to others types such as neural network and random forest based on the dataset features, and then illustrate these machine learning models in visualisation tools to interpret, frame and evaluate the complex machine learning models. As a result, the clinicians and researchers would read their dataset visualisation and predictive decision model at the same time to make the visualisation tools more reliable and trustable. We have not included adding new patient data to re-train the machine learning model yet, and we will add this feature in the future. The systematic analysis and a formal evaluation will also be carried out in the future.

6.4 PUBLICATIONS

This thesis has contributed to the following peer-reviewed journal and conference publications

Z Qu, CW Lau, QV Nguyen, Y Zhou, DR Catchpoole (2019). Visual Analytics of Genomic and Cancer Data: A Systematic Review. *Cancer Informatics*, vol 18, SAGE, pp. 1-18, doi: 10.1177/1176935119835546.

Z Qu, Y Zhou, QV Nguyen, DR Catchpoole (2019). Using Visualization to Illustrate Machine Learning Models for Genomic Data. In Proceedings of *the Australasian Computer Science Week Multi-conference*, Sydney, NSW, Australia, ACM, doi: 10.1145/3290688.3290719

CW Lau, QV Nguyen, **Z Qu**, S Simoff, D Catchpoole (2019). Immersive Intelligence Genomic Data Visualisation. In Proceedings of *the Australasian Computer Science Week Multi-conference*, Sydney, NSW, Australia, ACM, doi: 10.1145/3290688.3290722

QV Nguyen, **Z Qu**, ML Huang, CW Lau, S Simoff, DR Catchpoole (2018). A Mobile Tool for Interactive Visualisation of Genomics Data. In Proceedings of *the 9th International Conference on Information Technology in Medicine and Education*, IEEE, pp. 688-697.

Bibliography

Albuquerque, M. A., Grande, B. M., Ritch, E. J., Pararajalingam, P., Jessa, S., Krzywinski, M., . . . Morin, R. D. (2017). Enhancing knowledge discovery from cancer genomics data with Galaxy. *GigaScience, 6*(5), 1-13. doi:10.1093/gigascience/gix015

An, J., Lai, J., Wood, D. L., Sajjanhar, A., Wang, C., Tevz, G., . . . Nelson, C. C. (2015). RNASeqBrowser: a genome browser for simultaneous visualization of raw strand specific RNAseq reads and UCSC genome browser custom tracks. *BMC Genomics, 16*, 145. doi:10.1186/s12864-015-1346-2

Andre Esteva, B. K., Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. blau, Sebastian Thrun. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Macmillan Publishers Limited part of Springer Nature*.

ASQ. (2018). Scatter Diagram. Retrieved from http://asq.org/learn-about-quality/cause-analysis-tools/overview/scatter.html

Badr Hssina, A. M., Hanane Ezzikuri, Mohammed Erritali. (2014). A comparative study of decision tree ID3 and C4.5. *IJACSA*.

Bhojwani, D., Kang, H., Menezes, R. X., Yang, W., Sather, H., Moskowitz, N. P., . . . German Cooperative Study Group for Childhood Acute Lymphoblastic, L. (2008). Gene expression signatures predictive of early response and outcome in high-risk childhood acute lymphoblastic leukemia: A Children's Oncology Group Study [corrected]. *J Clin Oncol, 26*(27), 4376-4384. doi:10.1200/JCO.2007.14.4519

Biography. (2017). Gregor Mendel Biography.com. *The Biography.com website*.

Boudreaux, E. D., Waring, M. E., Hayes, R. B., Sadasivam, R. S., Mullen, S., & Pagoto, S. (2014). Evaluating and selecting mobile health apps: strategies for healthcare providers and healthcare organizations. *Transl Behav Med, 4*(4), 363-371. doi:10.1007/s13142-014-0293-9

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. . Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Caleydo. (2017). Integrative visualization of stratified heterogeneous data for disease subtype analysis. Retrieved from http://caleydo.org/tools/stratomex/

Catherine Plaisant, J. D., A.M. MacEachren, M.-J. Kraak. (2005). Information Visualization and the Challenge of Universal Usability *Exploring Geovisualization*.

Center, C. N. H. G. (2004). GenomeComp: A whole Genome Comparision and Visualization Tool. Retrieved from http://www.mgc.ac.cn/GenomeComp/

Chang, Y., Peng Xu, W., & Wang, L. (2013). *Research on 3D Visualization of Underground Antique Tomb Based on Augmented Reality* (Vol. 336-338).

Chelaru, F., Smith, L., Goldstein, N., & Bravo, H. C. (2014). Epiviz: interactive visual analytics for functional genomics data. *Nat Methods, 11*(9), 938-940. doi:10.1038/nmeth.3038

Chittaro, L. (2006). Visualization of Patient Data at Different Temporal Granularities on Mobile Devices. *HCI Lab, Dept. of Math and Computer Science, University of Udine*.

Ciaramella, A., Cocozza, S., Iorio, F., Miele, G., Napolitano, F., Pinelli, M., . . . Tagliaferri, R. (2008). Interactive data analysis and clustering of genomic data. *Neural Netw, 21*(2-3), 368-378. doi:10.1016/j.neunet.2007.12.026

Claas Heuer, C. S., Jens Tetens, Christa Kühn, Georg Thaller. (2016). Genomic prediction of unordered categorical traits: an application to subpopulation assignment in German Warmblood horses. *Genetics Selection Evolution*. doi:10.1186/s12711-016-0192-2

Cline, M. S., Craft, B., Swatloski, T., Goldman, M., Ma, S., Haussler, D., & Zhu, J. C. (2013). Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser. *Scientific Reports, 3*. doi:ARTN 2652

10.1038/srep02652

Columbus, L. (2017). McKinsey's 2016 Analytics Study Defines The Future Of Machine Learning. *Featured Posts, Technology / Software*.

Copeland, M. (2016). What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?

Cordeiro, J. V. (2014). Ethical and legal challenges of personalized medicine: Paradigmatic examples of research, prevention, diagnosis and treatment. *Revista Portuguesa de Saúde Pública, 32*(2), 164-180. doi:10.1016/j.rpsp.2014.10.002

Daniel A. Keim, T. M., Fabrice Rossi, Michel Verleysen. (2015). *Bridging Information Visualization with Machine Learning*. Retrieved from http://drops.dagstuhl.de/opus/volltexte/2015/5266/pdf/dagrep_v005_i003_p001_s15101.pdf

Dooley, B. J. (2017). Why AI with Augmented and Virtual Reality Will Be the Next Big Thing. Retrieved from https://tdwi.org/articles/2017/04/04/ai-with-augmented-and-virtual-reality-next-big-thing.aspx

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A, 95*(25), 14863-14868.

EMBL-EBI. (2018). Biological interpretation of gene expression data. Retrieved from https://www.ebi.ac.uk/training/online/course/functional-genomics-ii-common-technologies-and-data-analysis-methods/biological-0

Fiume, M. (2017). Genome Savant. Retrieved from http://www.genomesavant.com/p/home/index

Fiume, M., Williams, V., Brook, A., & Brudno, M. (2010). Savant: genome browser for high-throughput sequencing data. *Bioinformatics, 26*(16), 1938-1944. doi:10.1093/bioinformatics/btq332

Francis S. Collins, A. P., Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, and the members of the DOE and NIH planning groups. (2012). New Goals for the U.S. Human Genome Project: 1998-2003. *Science*.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science, 315*(5814), 972-976. doi:10.1126/science.1136800

Friendly, M. (2006). A Brief History of Data Visualization [Press release]. Retrieved from http://www.datavis.ca/papers/hbook.pdf

Fuchs, R., Waser, J., & Groller, M. E. (2009). Visual human+machine learning. *IEEE Trans Vis Comput Graph, 15*(6), 1327-1334. doi:10.1109/TVCG.2009.199

Gapminder. (2015). Updated Gapminder World Poster 2015! Retrieved from https://www.gapminder.org/downloads/updated-gapminder-world-poster-2015/

García-Hernández, R. J., Anthes, C., Wiedemann, M., & Kranzlmüller, D. (2016, 5-12 March 2016). *Perspectives for using virtual reality to extend visual data mining in information visualization.* Paper presented at the 2016 IEEE Aerospace Conference.

Garg, S., Nam, J. E., Ramakrishnan, I. V., & Mueller, K. (2008, 19-24 Oct. 2008). *Model-driven Visual Analytics.* Paper presented at the 2008 IEEE Symposium on Visual Analytics Science and Technology.

GATK. (2017). Genome Analysis Toolkit.

GDAC. (2016). TCGA Genome Data Analysis Center (GDAC) for Systems Analysis of the Cancer Regulome Retrieved from http://www.cancerregulome.org

Genome, Y. (2016). How are sequenced genomes stored and shared? Retrieved from http://www.yourgenome.org/facts/how-are-sequenced-genomes-stored-and-shared

Genomics, F. (2017). Why data visualization is so important in biology. Retrieved from https://www.fiosgenomics.com/data-visualization-and-data-analysis/

Gitools. (2018). Toot to tool communication. *Gitools 2.3.0 documentation.* Retrieved from http://www.gitools.org/docs/UserGuide_ToolCommunication.html

Goldman, M. (2017). UCSC Xena: Box Plots & Scatter Plots. *UCSC Xena.* Retrieved from http://xena.ucsc.edu/bar-graph-scatter-plot/

Goldman, M., Craft, B., Swatloski, T., Cline, M., Morozova, O., Diekhans, M., . . . Zhu, J. (2015). The UCSC Cancer Genomics Browser: update 2015. *Nucleic Acids Res, 43*(Database issue), D812-817. doi:10.1093/nar/gku1073

Golestan Hashemi, F. S., Razi Ismail, M., Rafii Yusop, M., Golestan Hashemi, M. S., Nadimi Shahraki, M. H., Rastegari, H., . . . Aslani, F. (2017). Intelligent mining of large-scale bio-data: Bioinformatics applications. *Biotechnology & Biotechnological Equipment, 32*(1), 10-29. doi:10.1080/13102818.2017.1364977

Google. (2017). Running DeepVariant on Google Cloud Platform. Retrieved from https://cloud.google.com/genomics/deepvariant

Gotz, D., & Wen, Z. (2009). *Behavior-driven visualization recommendation*. Paper presented at the Proceedings of the 14th international conference on Intelligent user interfaces, Sanibel Island, Florida, USA.

Gotz, D., When, Z., Lu, J., Kissa, P., Cao, N., Qian, W. H., . . . Zhou, M. X. (2010). *HARVEST: an intelligent visual analytic tool for the masses*. Paper presented at the Proceedings of the first international workshop on Intelligent visual interfaces for text analysis, Hong Kong, China.

Gray, G. E. (2016). Navigating 3d Scatter Plots in Immersive Virtual Reality. *University of Washington*.

Green, T. M., Ribarsky, W., & Fisher, B. (2008, 19-24 Oct. 2008). *Visual analytics for complex concepts using a human cognition model.* Paper presented at the 2008 IEEE Symposium on Visual Analytics Science and Technology.

Gu, Z. G., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics, 32*(18), 2847-2849. doi:10.1093/bioinformatics/btw313

Horning, N. (2015). Introduction to decision trees and random forests *American Museum of Natural History's Center for Biodiversity and Conservation*.

Huang da, W., Sherman, B. T., & Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res, 37*(1), 1-13. doi:10.1093/nar/gkn923

IGV. (2018). Scatter Plots. Retrieved from http://software.broadinstitute.org/software/igv/ScatterPlots

Insights, G. M. (2016). Precision Medicine Market Size to Exceed $87 Billion by 2023: Global Market Insights Inc. Retrieved from https://www.prnewswire.com/news-releases/precision-medicine-market-size-to-exceed-87-billion-by-2023-global-market-insights-inc-599454691.html

James Manyika, M. C., Anu Madgavkar, Susan Lund. (2017). What's now and next in analytics, AI, and Automation. *McKinsey & Company*.

Jearanaitanakij, K. (2005). Classifying Continuous Data Set by ID3 Algorithm. In (pp. 1048-1051).

Jeffrey Heer, B. S. (2012). Interactive dynamics for visual analysis. *Communications of the ACM, 55*(4). doi:10.1145/2133806.2133821

Jian Yanga, J. W., Zhi-Jian Yaob, Qi Jinc, Yan Shenb, Runsheng Chena. (2003). GenomeComp: a visualization tool for microbial genome comparison. *Journal of Microbiological Methods, 54*.

Juniper. (2018). Digital Health: Vendor Analysis, Emerging Technologies & Market Forecasts 2017-2022.

K. S. Pollard , M. J. v. d. L. (2003). Cluster Analysis of Genomic Data. *Center for Bioinformatics and Computational Biology*.

kane, D. (Writer). (2015). Data Science - Part V- Decision Trees & Random. In.

Keahey, T. A. (2013). Using visualization to understand big data. *Advanced visualization*.

Keim, D. A. (2001). Visual Exploration of Large Data Sets. *Daniel A. Keim, 44*(8).

Khushboo Wadhwani, Y. W. (2017). Big Data Challenges and Solutions. doi:10.13140/RG.2.2.16548.88961

Knight, W. (2017). Google Has Released an AI Tool That Makes Sense of Your Genome. Retrieved from https://www.technologyreview.com/s/609647/google-has-released-an-ai-tool-that-makes-sense-of-your-genome/

Krisa D. Tailor, S. I. (2014). Data Visualization in Health Care: Optimizing the Utility of Claims Data through Visual Analysis.

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., . . . Dry, J. R. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res, 44*(11), e108. doi:10.1093/nar/gkw227

Larry Goldberg, B., Bettijoyce Lide, MS, Svetlana Lowry, PhD, Holly A. Massett, PhD,, Trisha O'Connell, M., Jennifer Preece, PhD, Whitney Quesenbery, BA,, & Ben Shneiderman, P. (2011). Usability and Accessibility in Consumer Health Informatics Current Trends and Future Challenges. *Journal of Preventive Medicine, 40(5S2),* S187-S197.

Lau, C. W., Nguyen, Q. V., Qu, Z., Simoff, S., & Catchpoole, D. (2019). Immersive Intelligence Genomic Data Visualisation. *ACM*. doi:10.1145/3290688.3290722

Ledford, H. (2016). AstraZeneca launches project to sequence 2 million genomes. *Nature, 532*(7600), 427. doi:10.1038/nature.2016.19797

Leung, M. K. K., Delong, A., Alipanahi, B., & Frey, B. J. (2016). Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets. *Proceedings of the IEEE, 104*(1), 176-197. doi:10.1109/jproc.2015.2494198

Levin, C. (2017). Your top 3 heatmap generation tools. Retrieved from https://blog.omictools.com/your-top-3-heatmap-generation-tools/

Lex, A., Streit, M., Kruijff, E., & Schmalstieg, D. (2010, 2-5 March 2010). *Caleydo: Design and evaluation of a visual analysis framework for gene expression data in its biological context.* Paper presented at the 2010 IEEE Pacific Visualization Symposium (PacificVis).

Lex, A., Streit, M., Schulz, H. J., Partl, C., Schmalstieg, D., Park, P. J., & Gehlenborg, N. (2012). StratomeX: Visual Analysis of Large-Scale Heterogeneous Genomics Data for Cancer Subtype Characterization. *Comput Graph Forum, 31*(33), 1175-1184. doi:10.1111/j.1467-8659.2012.03110.x

Li, J., Lei, J., Zhao, X., Zhang, C., & Han, X. (2013). An Improved ID3 Algorithm. *Applied Mechanics and Materials, 444-445*, 723. doi:10.4028/www.scientific.net/AMM.444-445.723

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat Rev Genet, 16*(6), 321-332. doi:10.1038/nrg3920

Liu, M. S., Liu, Y., Deng, L., Wang, D., He, X. Y., Zhou, L., . . . Liu, S. L. (2018). Transcriptional profiles of different states of cancer stem cells in triple-negative breast cancer. *Molecular Cancer, 17*. doi:ARTN 65

10.1186/s12943-018-0809-x

Ma, K. L. (2007). Machine learning to boost the next generation of visualization technology. *IEEE Comput Graph Appl, 27*(5), 6-9. doi:10.1109/MCG.2007.129

Margaret A. Hamburg, M. D. (2013). Paving the Way for Personalized Medicine FDA's Role in a New Era of Medical Product Development *FDA*.

Mario, V. D. (2016). Predicting genetic diseases with CloudForest. *GopherAcademy*.

Marr, B. (2016a). 3 Industries That Will Be Transformed By AI, Machine Learning And Big Data In The Next Decade. Retrieved from https://www.forbes.com/sites/bernardmarr/2016/09/27/3-industries-that-will-be-transformed-by-ai-machine-learning-and-big-data-in-the-next-decade/#770b87b183ed

Marr, B. (2016b). How VR Will Revolutionize Big Data Visualizations. *Tech BigData*. Retrieved from https://www.forbes.com/sites/bernardmarr/2016/05/04/how-vr-will-revolutionize-big-data-visualizations/#2f50d104e151

Matte-Tailliez, O., Toffano-Nioche, C., Ferey, N., Kepes, F., & Gherbi, R. (2006, 24-28 April 2006). *Immersive Visualization for Genome Exploration and Analysis.* Paper presented at the 2006 2nd International Conference on Information & Communication Technologies.

McCarthy, J. F., Marx, K. A., Hoffman, P. E., Gee, A. G., O'Neil, P., Ujwal, M. L., & Hotchkiss, J. (2004). Applications of Machine Learning and High-Dimensional Visualization in Cancer Detection, Diagnosis, and Management. *Annals of the New York Academy of Sciences, 1020*(1), 239-262. doi:10.1196/annals.1310.020

McClean, P. (2011). A History of Genetics and Genomics

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Microsoft. (2018). Microsoft HoloLens. Retrieved from https://www.microsoft.com/en-au/hololens

Mills, M. (2016). ARTIFICIAL INTELLIGENCE IN LAW: THE STATE OF PLAY 2016 *Thomson Reuters S031401/3-16.*

Moleten, M. (2017). Google is giving away AI that can build your genome sequence. Retrieved from https://www.wired.com/story/google-is-giving-away-ai-that-can-build-your-genome-sequence/

Natalia Andrienko, G. A. (2007). Intelligent Visualisation and Information Presentation for Civil Crisis Management. *Transactions in GIS, 11*(6), 11. doi:10.1111/j.1467-9671.2007.01078.x

Nguyen, Q. V., Gleeson, A., Ho, N., Huang, M. L., Simoff, S., & Catchpoole, D. (2011). Visual Analytics of Clinical and Genetic Datasets of Acute Lymphoblastic Leukaemia. In B.-L. Lu, L. Zhang, & J. Kwok (Eds.), *Neural Information Processing: 18th International Conference, ICONIP 2011, Shanghai, China, November 13-17, 2011, Proceedings, Part I* (pp. 113-120). Berlin, Heidelberg: Springer Berlin Heidelberg.

Nguyen, Q. V., Khalifa, N. H., Alzamora, P., Gleeson, A., Catchpoole, D., Kennedy, P. J., & Simoff, S. (2016). Visual Analytics of Complex Genomics Data to Guide Effective Treatment Decisions. *Journal of Imaging, 2*(4), 29. doi:UNSP 29

10.3390/jimaging2040029

Nguyen, Q. V., Nelmes, G., Huang, M. L., Simoff, S., & Catchpoole, D. (2014). Interactive Visualization for Patient-to-Patient Comparison. *Genomics Inform, 12*(1), 21-34. doi:10.5808/GI.2014.12.1.21

Nguyen, Q. V., Qian, Y., Huang, M. L., & Zhang, J. W. (2013). TabuVis: A tool for visual analytics multidimensional datasets. *Science China-Information Sciences, 56*(5), 1-12. doi:ARTN 052105

10.1007/s11432-013-4870-1

Nguyen, Q. V., Qu, Z., Huang, M. L., Lau, C. W., Simoff, S., & Catchpoole, D. R. (2018, 19-21 Oct. 2018). *A Mobile Tool for Interactive Visualisation of Genomics Data.* Paper presented at the 2018 9th International Conference on Information Technology in Medicine and Education (ITME).

NIH. (2017a). Cancer Genomic Research. Retrieved from https://www.cancer.gov/research/areas/genomics

NIH. (2017b). GDC Dave Tools. Retrieved from https://gdc.cancer.gov/analyze-data/gdc-dave-tools

Nilsson, N. J. (2009). THE QUEST FOR ARTIFICIAL INTELLIGENCE A HISTORY OF IDEAS AND ACHIEVEMENTS.

Olshannikova, E., Ometov, A., Koucheryavy, Y., & Olsson, T. (2015). Visualizing Big Data with augmented and virtual reality: challenges and research agenda. *Journal of Big Data, 2*(1). doi:10.1186/s40537-015-0031-2

Oracle. (2015). See the Signals Oracle Data Visualization Cloud Service. *Oracle Data Visualization*.

Ortega, J., & Aguillo, I. (2013). *Network visualisation as a way to the web usage analysis* (Vol. 65).

Patrick Hall, W. P., SriSatish Ambati. (2017). Ideas on interpreting machine learning. *O'Reilly*.

Pavlopoulos, G. A., Malliarakis, D., Papanikolaou, N., Theodosiou, T., Enright, A. J., & Iliopoulos, I. (2015). Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *GigaScience, 4*, 38. doi:10.1186/s13742-015-0077-2

Perez-Llamas, C., & Lopez-Bigas, N. (2011). Gitools: analysis and visualisation of genomic data using interactive heat-maps. *PloS One, 6*(5), e19541. doi:10.1371/journal.pone.0019541

Pollard, K. S., & van der Laan, M. J. (2005). Cluster Analysis of Genomic Data. In R. Gentleman, V. J. Carey, W. Huber, R. A. Irizarry, & S. Dudoit (Eds.), *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (pp. 209-228). New York, NY: Springer New York.

Polsky, A. (2017). Why your brain needs data visualization. Retrieved from https://www.sas.com/en_us/insights/articles/analytics/why-your-brain-needs-data-visualization.html

Procter, J. B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F., & Barton, G. J. (2010). Visualization of multiple alignments, phylogenies and gene family evolution. *Nat Methods, 7*(3 Suppl), S16-25. doi:10.1038/nmeth.1434

Quesenbery, W. (2003). Dimensions of Usability:

Defining the Conversation, Driving the Process. *Proceedings of the UPA 2003 Conference*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*(1), 81-106. doi:10.1007/BF00116251

Quinlan, R. (2018). Ross Quinlan. Retrieved from http://www.rulequest.com/Personal/

R. Jayabrabu, V. S., K. Vivekanandan. (2012). A framework: Cluster detection and multidimensional visualization of automated data mining using intelligent agents. *International Journal of Artificial Intelligence & Applications (IJAIA), 3*, 15.

Raskin, A. C., E. (2011). The Transformation of Medicine and Its Consequences for Investors. In *The Dawn of Molecular Medicine*. New York, NY: Alliance Bernstein.

Rebeiz, M., & Posakony, J. W. (2004). GenePalette: a universal software tool for genome sequence visualization and analysis. *Dev Biol, 271*(2), 431-438. doi:10.1016/j.ydbio.2004.04.011

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol, 29*(1), 24-26. doi:10.1038/nbt.1754

Russel, J. (2017). Google's AlphaGo AI wins three-match series against the world's best Go player. *TC Sessions Robotics*.

Samwell. (2017). Deep learning in GATK4. Retrieved from https://software.broadinstitute.org/gatk/blog?id=10996

SAS. (2017). Data Visualization what it is and why it matters. Retrieved from https://www.sas.com/en_us/insights/big-data/data-visualization.html

SAS. (2018). Network Visualization Workshop2.1 User's Guide. Retrieved from http://support.sas.com/documentation/cdl/en/grnvwug/62918/HTML/default/viewer.htm#p0q343kxjyj36jn1e2z6lulkda3j.htm

Savoia, C., Volpe, M., Grassi, G., Borghi, C., Agabiti Rosei, E., & Touyz, R. M. (2017). Personalized medicine-a modern approach for the diagnosis and management of hypertension. *Clin Sci (Lond), 131*(22), 2671-2685. doi:10.1042/CS20160407

Schroeder, M. P., Gonzalez-Perez, A., & Lopez-Bigas, N. (2013). Visualizing multidimensional cancer genomics data. *Genome Med, 5*(1), 9. doi:10.1186/gm413

Selan dos Santos, K. B. (2004). Gaining understanding of multivariate and multidimensional data through visualization. *Computers & Graphics, 28*(3).

Sennaar, K. (2018). Machine Learning in Genomics - Current Efforts and Future Applications. Retrieved from https://www.techemergence.com/machine-learning-in-genomics-applications/

Shan, Q., Doyle, T. E., Samavi, R., & Al-Rei, M. (2017). Augmented Reality Based Brain Tumor 3D Visualization. *8th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (Euspn 2017) / 7th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (Icth-2017) / Affiliated Workshops, 113*, 400-407. doi:10.1016/j.procs.2017.08.356

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., . . . Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res, 13*(11), 2498-2504. doi:10.1101/gr.1239303

Shen, L., Shao, N. Y., Liu, X. C., & Nestler, E. (2014). ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics, 15*. doi:Artn 284

10.1186/1471-2164-15-284

Shilling, C. (2017). How Augmented Reality will change Data Visualization. Retrieved from http://blog.i2econsulting.com/how-augmented-reality-will-change-data-visualization/

Sikic, B. I., Tibshirani, R., & Lacayo, N. J. (2008). Genomics of childhood leukemias: the virtue of complexity. *J Clin Oncol, 26*(27), 4367-4368. doi:10.1200/JCO.2008.16.4285

Simpson, R. M., LaViola, J. J., Laidlaw, D. H., Forsberg, A. S., & van Dam, A. (2000). Immersive VR for scientific visualization: a progress report. *IEEE Computer Graphics and Applications, 20*(6), 26-52. doi:10.1109/38.888006

Song, Y.-y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction.(Biostatistics in psychiatry (26))(Report). *Shanghai Archives of Psychiatry, 27*(2), 130. doi:10.11919/j.issn.1002-0829.215044

Sonia Singh, P. G. (2014). COMPARATIVE STUDY ID3, CART AND C4.5 DECISION TREE ALGORITHM: A SURVEY. *International Journal of Advanced Information Science and Technology (IJAIST), 27*.

SQLite. (2018). SQLite. Retrieved from https://www.sqlite.org/index.html

Staudt, l. (2017). Introducing DAVE: Online Analysis Tools for the Genomic Data Commons. Retrieved from https://www.cancer.gov/news-events/cancer-currents-blog/2017/gdc-dave-tools

Stephanie, T. (2015). A visual introduction to machine learning. Retrieved from http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., . . . Robinson, G. E. (2015). Big Data: Astronomical or Genomical? *PLoS Biol, 13*(7), e1002195. doi:10.1371/journal.pbio.1002195

Stevens, E. A., & Rodriguez, C. P. (2015). Genomic medicine and targeted therapy for solid tumors. In (Vol. 111, pp. 38-42).

Stolk, B., Abdoelrahman, F., Koning, A., Wielinga, P., Neefs, J.-M., Stubbs, A., . . . Van der Spek, P. (2002). *Mining the human genome using virtual reality*.

Tang, J., Liu, R., Zhang, Y. L., Liu, M. Z., Hu, Y. F., Shao, M. J., . . . Zhang, W. (2017). Application of Machine-Learning Models to Predict Tacrolimus Stable Dose in Renal Transplant Recipients. *Sci Rep, 7*, 42192. doi:10.1038/srep42192

Tatu, A., Albuquerque, G., Eisemann, M., Schneidewind, J., Theisel, H., Magnork, M., & Keim, D. (2009, 12-13 Oct. 2009). *Combining automated analysis and* *visualization techniques for effective exploration of high-dimensional data.* Paper presented at the 2009 IEEE Symposium on Visual Analytics Science and Technology.

TCGA, N. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature, 487*(7407), 330-337. doi:http://www.nature.com/nature/journal/v487/n7407/abs/nature11252.html#supplementary-information

Technologies, U. (2018). Unity. Retrieved from https://unity3d.com/

Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform, 14*(2), 178-192. doi:10.1093/bib/bbs017

Tibco. (2018). What is a 3D Scatter Plot? Retrieved from https://docs.tibco.com/pub/spotfire/6.5.1/doc/html/3d_scat/3d_scat_what_is_a_3d_scatter_plot.htm

Towler, W. (2015). Data Visualization: The future of data visualization. Retrieved from http://analytics-magazine.org/data-visualization-the-future-of-data-visualization/

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., . . . Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics, 26*(12), i237-245. doi:10.1093/bioinformatics/btq182

Verma, P. (2017). When Virtual Reality Meets Big Data. Retrieved from https://www.ge.com/digital/blog/when-virtual-reality-meets-big-data

Viswebmaster. (2009). Drosophila Gene Expression Data Exploration and Visualization. Retrieved from http://vis.lbl.gov/Events/SC07/Drosophila/

Vogenberg, F. R., Isaacson Barash, C., & Pursel, M. (2010). Personalized Medicine: Part 1: Evolution and Development into Theranostics. *Pharmacy and Therapeutics, 35*(10), 560-576.

Vukmirovic, O. G., & Tilghman, S. M. (2000). Exploring genome space. *Nature, 405*(6788), 820-822.

Wachtel, M., Runge, T., Leuschner, I., Stegmaier, S., Koscielniak, E., Treuner, J., . . . Schafer, B. (2006). Subtype and prognostic classification of rhabdomyosarcoma by immunohistochemistry. *Journal of Clinical Oncology, 24*(5), 816-822. doi:10.1200/JCO.2005.03.4934

Ware, C. (2013). Information Visualization perception for design.

Wilson, B. (2008). Induction of Decision Trees. Retrieved from http://www.cse.unsw.edu.au/~billw/cs9414/notes/ml/06prop/id3/id3.html

Wistuba, II, Gelovani, J. G., Jacoby, J. J., Davis, S. E., & Herbst, R. S. (2011). Methodological and practical challenges for personalized cancer therapies. *Nat Rev Clin Oncol, 8*(3), 135-141. doi:10.1038/nrclinonc.2011.2

Yates, T., Okoniewski, M. J., & Miller, C. J. (2008). X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis. *Nucleic Acids Res, 36*(Database issue), D780-786. doi:10.1093/nar/gkm779

Zhang, F., Xu, Y., Cao, H., Jin, C., Cheng, Z., Wang, G., & Shugart, Y. Y. (2015). Mapsnp: an R package to plot a genomic map for single nucleotide polymorphisms. *PloS One, 10*(4), e0123609. doi:10.1371/journal.pone.0123609

Zhonglin Qu, Y. Z., Quang Vinh Nguyen, Daniel R. Catchpoole. (2019). Using Visualization to Illustrate Machine Learning Models for Genomic Data. *ACM*. doi:10.1145/3290688.3290719

Zucerberg, M. (2017). Facebook CEO Mark Zuckerberg's Harvard Commencement Speech (Full Transcript). Retrieved from https://singjupost.com/facebook-ceo-mark-zuckerbergs-harvard-commencement-speech-full-transcript/2/

Appendices

Appendix A: Interview Form

Data Visualisation Tools Feed back

Your Name: Interview Date:

-

1. Can you please give a little background about your research work. Are you familiar with the disease dataset?

2. Are the visualisation tools useful in capturing similarities between patient genetics? Can you please give more information on your opinion.

3. Are the visualisation tools useful in capturing individual differences and making patient to patient comparisons? Would you like to give more feedback on the scenarios of the application in your work?

4. What do you think about the potential/usefulness of each tool? Would you apply the visualisation tools into your work or research to support your decisions?

5. In a bigger ambition, would you think the visualisation tool is useful to enable personalised medicine? In which way? How would the visualisation tools help researchers/medical doctors to make more sense of the data and make better decisions?

6. Are there any other comments on the visualisation tools?