Understanding Pandemic Outbreaks through Data Visualisation - An Assessment of Current Tools and Techniques

Robert Howe

Department of Computer Science, University of British Columbia

1. INTRODUCTION

1.1 Pandemic Outbreak Visualisation

There has been a significant amount of work in the area of epidemic and pandemic disease visualisation, particularly as this domain may be crucial to the survival of the human species. A pandemic disease is one that spreads to be prevalent over a whole country or the world, as opposed to an epidemic disease which is more local - on a city scale or smaller. Naturally there is much overlap in understanding these two types of disease spread, however this paper will focus more on the larger scale pandemic spreads.

Examples of notable pandemic diseases throughout history include: the bubonic plague in 14th-15th century Europe, killing approximately 25 million people[1]; tuberculosis, which currently infects one third of the world's current population with new infections occurring at a rate of approximately one person per second killing in 50% of cases[2]; and malaria, infecting approximately 350-500 million people worldwide[3]. These diseases present an immense hurdle for humanity, infecting and killing large numbers of the global population, however all of these diseases started with a single infection. Understanding how this single infection can spread to become a worldwide pandemic will help us to understand not only how to aid in reducing numbers of current pandemic infections, but also to prevent future pandemic outbreaks.

The question then becomes what is the best way to study and understand the data we have available on pandemic diseases in order to make meaningful conclusions from it? One of the most intuitive ways to do this is through information visualisation. Visualisation of the data allows us to much more easily recognise patterns in it and pick out key areas of interest for further study. Once these patterns have been identified and further analysis has been carried out, the information gained may be used to help form health policy to aid in the prevention of future pandemic outbreaks.

In line with this emphasis placed on understanding pandemic outbreaks, the VAST Challenge (a major annual information visualisation competition) has included several challenges involving visualising epidemic and pandemic diseases, in particular characterising where they originate and how they spread. The focus of this paper is part of the VAST Challenge from 2010, which involved analysing synthetic data consisting of hospitalisation and death records in order to characterise the spread of a fictitious outbreak of drafa fever[4].

1.2. Domain, Task, and Data

1.2.1. Domain

Visualising pandemic outbreak information will typically fall into the domains of health policy and medical research, both of which will have heavily overlapping goals. The health policy side will be aiming to understand the spread of the disease in order to determine the best method for containing it, and future pandemic outbreaks. The medical side will be aiming to understand many characteristics of the disease, including spread, in order to determine which treatments will be effective at containing the outbreak. This analysis may either be carried out after an outbreak has occurred, or while one is actually happening.

The visualisations constructed during this process will not only be used with the intention of helping researchers to understand the disease, but also to enable them to easily present their findings to policy makers who may have limited time to dedicate to hearing about this research.

1.2.2 Task

The task at hand is to analyse past entries to the VAST Challenge 2010, as well as the general literature available for pandemic disease visualisation, to find out which approaches have been the most effective in understanding and characterising the spread of these diseases. Analysing these previous approaches that used hospitalisation and death records will involve assessing them on a number of different criteria, including: ability to isolate pandemic disease symptoms from general illnesses; ability to locate the origin (location) of a pandemic disease; ability to characterise how the disease spreads (including variants); and ability to identify anomalies in the data. If addressing a current outbreak, these goals should be able to be completed relatively quickly.

Once this analysis of available literature has been completed, several current software packages will be used to create visualisations based on the VAST Challenge 2010 data. These tools will be compared and evaluated for their suitability in terms of creating pandemic visualisations for this domain. They will be assessed using the same criteria used for evaluation of the currently available literature in this area to hopefully provide some insight into future possibilities in this area, and whether current tools are sufficiently meeting the demands of the relevant domains.

1.2.3 Data

The data provided for the VAST Challenge 2010 is typical of the kind of data available in pandemic disease studies. This data consists of tables of hospitalisation and death records for affected countries, occasionally with more supplemental information, however in this paper we mostly consider those two types of records. These records typically contain some sort of patient identifier, along with date of hospital admission or death, and any relevant symptoms for the hospitalisation records.

When the sample size is large enough (ie there are enough records) this data will typically be accurate enough to carry out the pandemic disease analysis. Unfortunately however, symptoms in hospitalisation records are recorded using a wide variety of terminology, recorded by many different doctors, and so this data will often need to be cleaned to a certain extent before any meaningful analysis can be carried out on it.

2. Related Work

The related work in this task falls into two broad categories: previous VAST Challenge 2010 entries; and general literature available on pandemic visualisation. These two categories will be considered separately, and then a conclusion will be made including both categories and determining what has been the most effective approach in this domain. The papers addressed here will be summarised according to the analysis framework used in Munzner 2015[5].

2.1. Past VAST Challenge Entries

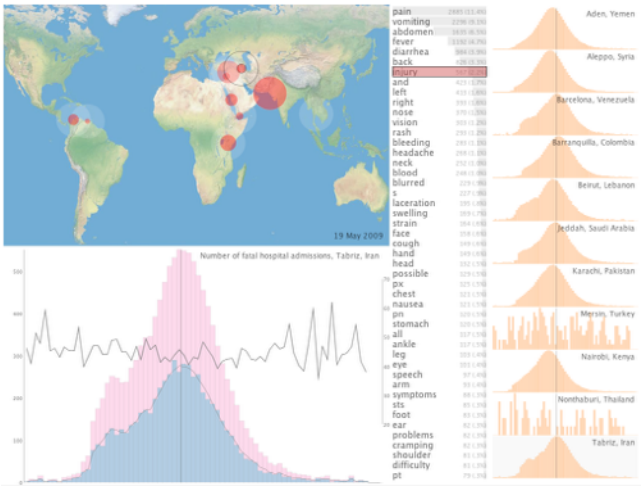
2.1.1. giCentre - PandemView

The giCentre is based in the Computer Science department at the City University London and specialises in developing techniques for visualising data. Their entry into the VAST Challenge 2010 (including their application "PandemView") was awarded with "Good overall design and analysis" and so is a good candidate for studying visualisation effectiveness[4][6].

For this entry they broke their analysis down into five separate components: identifying drafa fever symptoms; spread of the disease; timing of the outbreaks; disease variants; and anomalies; each requiring their own visualisations to aid in understanding the disease.

The first component (identifying drafa fever symptoms) was approached using an alphabetical tag cloud for each country involved in the outbreak. These tag clouds contained symptoms found in hospitalisation records and so easily allowed for the identification of major contributors to hospitalisation. As these symptoms were for all hospitalisations - not those specifically for drafa fever admissions - they were compared with observed fatal symptom frequencies in order to provide drafa symptom candidates as fatal admissions were more likely to be drafa cases than not. This use of a tag cloud clearly and quickly shows those symptoms most likely to be involved with the disease in a way that creates a useful association in the human mind. While a bar graph showing numbered frequencies of different symptoms would do an adequate job of identifying the same relevant symptoms, the tag cloud takes advantage of the natural human visual tendency to place a higher emphasis on more obvious (in this case, larger and coloured red) elements. This visualisation thus most likely makes it easier for researchers to remember these symptoms for future reference as they stick out more clearly in the mind. If we were concerned with exact frequencies of symptoms then using font size as a channel would be inadequate as the varying word lengths and sizes would create confusion, however when we are concerned with a threshold value - whether the symptom is involved in the disease or not - the tag cloud is more than sufficient, and individual frequencies may be studied later. The tag cloud however does not provide a detailed indication of when these symptoms are occurring, particularly as the disease progresses, which may be important information for controlling the outbreak.

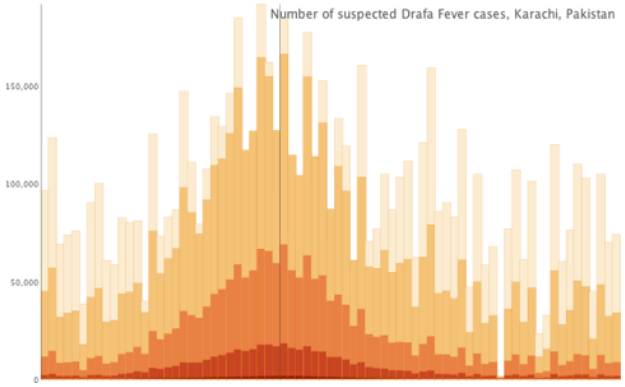
After this initial step, the authors move on to characterising the spread of the disease. For this step they developed their PandemView application using Processing - "a set of Java libraries for rapid development of graphical and visualization sketches". The PandemView interface consists of four separate views as shown in figure 1, and these views help to create a full picture of the spread of drafa fever. The bottom left view gives an idea of how age or gender may influence infection rates. For gender, blue and pink stacked bar charts are used changing over



1. The PandemView application showing four separate views: the top left view shows a geographic map of the world, with number of hospitalisation or death records for a specific date (selected by the user) shown overlaid as red circles with area representing number of records; the bottom left view shows a detailed bar graph of hospitalisation or death records over time for one city or country (selected by the user), with options to show a stacked bar graph using gender (blue representing male and pink representing female) and a black line graph overlaid representing age; the middle view shows symptoms and their frequencies for the selected record as a horizontal bar graph; the right view shows vertically aligned vertical bar graphs representing number of hospital or death records for all of the different countries.

time as the pandemic goes on. Normally stacked bar charts are not useful for this kind of analysis, as they do not allow direct comparison of the heights of segments within one bar, however in this case a 50% male/female line was added to show where the male bar should finish and the female bar should begin if there is no correlation with gender, which made the comparison easier. The stacked bar charts were most likely used as a way to provide a sense a continuity with the histograms on the right, however for this more detailed analysis of gender and age, separate rather than stacked bar charts may prove more useful. This view in general may be somewhat unnecessary, due to the fact that this kind of analysis would typically happen at a later stage in more detail. The line graph here represents the average age of patients, which is not a good measure to use as often it may be only younger and older persons affected by the disease, as turned out to actually be the case in this disease, and this is not be reflected by an average. This adds to the point that this type of analysis should probably be carried out in more detail at a later stage.

In order to determine the origin of the disease, certain key values from the generated histograms (seen in figure 1) were used. For each city dates were recorded for the earliest appearance of the disease, the peak of the disease, and how long it took to make a full recovery. Based on this data, and helped by overlaying this onto the map (figure 1 top left), they were able to characterise the spread of the disease as it originated and spread across several countries in a matter of weeks. In addition to this, they were able to show which cities made the fastest recovery, a potential point for further analysis in helping to stop the spread of the disease.



2. Stacked bar chart view available as a replacement of the lower left view in figure 1. Here the darkest colour represents patients with at least 3 drafa fever symptoms, the next lightest shade at least 2, the next lightest shade at least 1, and the lightest shade is general admissions.

Another feature of PandemView is its ability to show hospital admissions with number of drafa fever symptoms overlaid as stacked bar charts as seen in figure 2. This allows for identifying various time patterns of the outbreak against the background noise of typical hospital admissions, and is hugely useful in confirming hypotheses on timing of the outbreak. It appears very effective at isolating disease cases from general illnesses.

Finally, in terms of detecting anomalies PandemView relies largely on the histograms of death records over time in order to complete this task. Any obvious anomalies will be detected such as the two examples on the right of figure 1 that have highly variable death rates over time, unlike the rest of the histograms which have an obvious peak. More subtle anomalies however may not be able to be detected, however this may not be of particular significance as it is most likely the major anomalies that will be of most interest to researchers.

In summary the techniques developed in this paper were able to: somewhat effectively isolate pandemic disease cases from general illnesses using the alphabetical tag cloud and stacked bar charts for drafa symptoms (figure 2); determine the origin of the disease based on data and insights gained from generated and aligned histograms; characterise the spread of the disease, particularly against the background noise of general hospital admissions and deaths, using the stacked bar charts of drafa symptoms (figure 2); and effectively determine major anomalies which may be useful for further study. This paper is an example of a thorough execution of simple visualisation methods such as bar charts, histograms, and map overlays, leading to an application that serves as a good starting point for pandemic disease analysis and getting a quick overview, but does not provide tools for a thorough analysis to gain a deep understanding of the disease.

System	PandemView
What: Data	Table of hospitalisation and death records for several major cities and countries
Why: Tasks	Locate origin of the disease and characterise its spread
How: Encode	Map overlaid with hospitalisation/death data; histograms of hospitalisation/death data over time for each area; larger histogram for one selected area also showing gender (stacked bars) and age (line graph); symptom occurrences ordered by frequency; symptom occurrences displayed in tag cloud

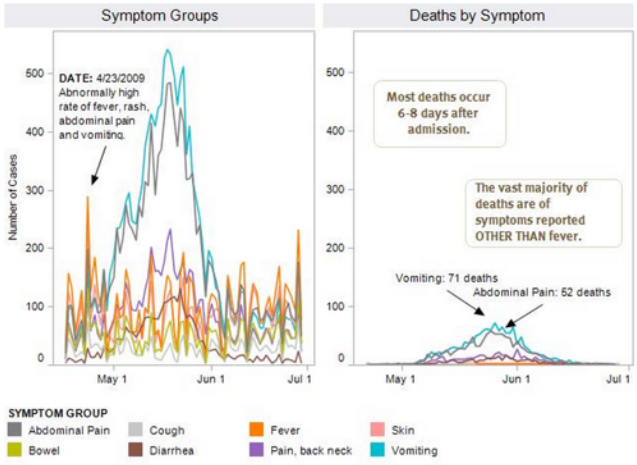
Table 1. Summary of the PandemView system.

2.1.2 Periscopic - Aggregate Symptoms Visualisation

Periscopic is a commercial data visualisation firm based in Portland, OR, USA. They specialise in "socially-conscious" data visualisations that help companies and organisations to promote information transparency and public awareness. Their entry into the VAST Challenge 2010[4] was awarded with "Effective visualisation of symptoms" and so warrants closer examination[7].

The Periscopic entry is not as detailed and formal in their analysis as was the giCentre entry, as their primary goal appears to have been related to understanding the symptoms of the disease rather than a full analysis of its ability to spread. They noted early on that the symptom data needed to be cleaned, with some symptoms having as many as 60 variations (for example, "abdominal pain" and "upper right abdominal pain"), with many of these being able to be grouped into the same category. Certain hospitalisation or death records could also be ignored, such as those due to dog bites and car accidents, which are most likely unrelated to this specific pandemic. Preliminary analysis should however be carried out before this data is excluded, in case some insight due to an anomaly is noticed (for example if the disease is spreading through dog bites), but once this analysis has been completed the relevant data may be excluded.

Once the data had been appropriately cleaned, the observed symptom frequencies were plotted as a line graph over the course of the epidemic for each country. Initially all symptoms (ungrouped) were plotted, which created an indistinguishable mess apart from a few peaks, however the general bell curve of frequency expected for a pandemic was seen. At the expected

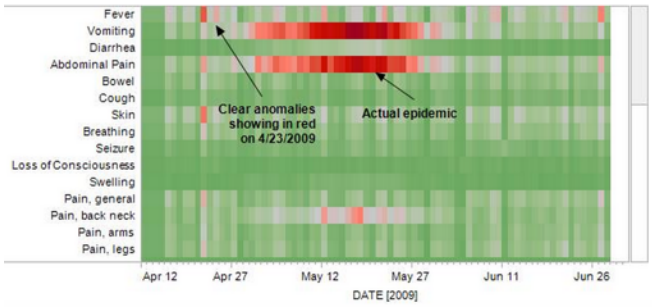


3. Symptoms grouped into 8 major categories and then plotted over time based on frequency (left) and then associated deaths by major symptom group (right)

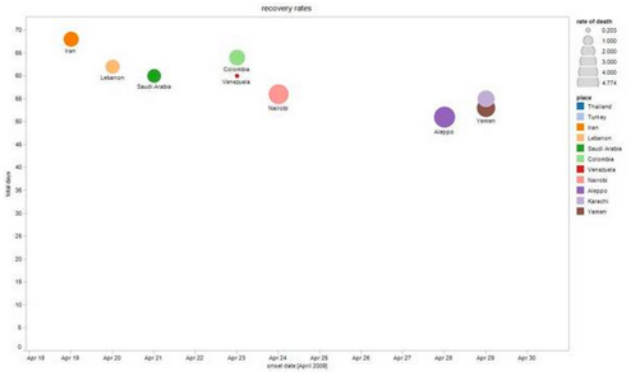
peak (at the centre of the bell curve), the visualisation of all symptoms showed a large number of cases of vomiting and fever, but also showed a slight preliminary peak of dizziness and blurred vision.

As this visualisation of all symptoms was somewhat ineffective (due to an incredibly cluttered view making pattern recognition very difficult), instead the symptoms were grouped into 8 major groups, and the observed frequencies of these were then plotted over time (figure 3). As can be seen from this figure, there is a clear preliminary spike of fever, rash, abdominal pain, and vomiting several weeks before the peak of the pandemic. While this may or may not be linked to the pandemic itself, detection of these kinds of anomalies is hugely important as they may give health workers in the affected countries an early warning sign to look out for, making control of the pandemic significantly easier and more effective. Further to this they looked at number of death records grouped by associated symptoms (figure 3) and found that most deaths were occurring 6-8 days after admission, and that the vast majority of deaths were not related to fever. This again is strange given the earlier information about the preliminary spike involving fever, however the main objective of this tool is not to determine the exact pathology of the disease, but rather to give suggestions based on the available data. This visualisation approach using line graphs is very simple but very effective, demonstrating that visualisation really is not about what looks flashy and impressive, but what will most obviously show patterns in the data for its intended purpose.

Further to this line graph approach, Periscopic constructed a heat matrix of hospital admission records based on major groups of symptoms (figure 4). This heat matrix appears as equally effective as, if not more effective than, the line graph of observed symptom frequencies in terms of identifying key warning and peak symptoms. While the line graph gets somewhat cluttered, particularly around the preliminary peak, the heat matrix is able to clearly identify the preliminary peak as well as the main peak while also clearly showing which individual symptom groups are involved and remaining uncluttered. This reduction in clutter may be significant to researchers in terms of reducing errors that may be caused by using the line graph alone, and so the heat matrix becomes very important. The line graph is still strong in showing general trends more intuitively however, and so when used together the two make an effective combination in terms of



4. Heat matrix for the major symptom groups over time, clearly showing the preliminary peak (at April 23) as well as the main peak (in May). Symptoms most likely not involved in the pandemic can be clearly seen in green, and possible involved symptoms such as diarrhoea can be further examined.



5. Scatter plot showing observed death rates for different cities involved in the outbreak. X axis shows onset date and Y axis shows total days of the outbreak. Different cities are represented by different circles, each with a unique colour. Circle area represents death rate for that particular city and the scale is shown in the upper right.

understanding how observed symptom frequencies have changed and fluctuated over the course of the pandemic.

In addition to their detailed symptom analysis, Periscopic then went on to plot the recovery rates of different cities based on data obtained in the previous steps, which they represented as a simple scatterplot (figure 5). This scatterplot implemented as is is not a very effective visualisation for several reasons. Firstly the scale of the graph on both axes is very spread out, with very few data points and a lot of white space, which suggests the space is not being used very effectively. Even with this large spread, there are still two data points overlapping, suggesting circle area is not a good choice to encode death rate here. Circle area may also not have been a good choice as humans are not as good at translating two dimensional changes (such as circle area difference) into real difference in data as they are for one dimensional changes (such as bar charts), particularly when the data points are as far away as they are in this plot. The scale is also not very useful, where on the right different circle area is shown to represent a certain death rate, but translating this into the circle areas in the plot will require significant cognitive load when a histogram or bar chart may well have shown this information much more simply with less cognitive load.

Finally in order to locate the origin of the disease, Periscopic overlaid the number of deaths over time for each city onto a map of the relevant area, and created an animation beginning with the first recorded death to observe where the outbreak began. Using only this animation with no reference to other data such as first possible drafa case (extrapolated based on symptoms) an incorrect conclusion was reached and the wrong country of origin was suspected. This shows that a combination of data may be needed in order to get a full picture, and while the map data is definitely helpful for understanding the spread of the disease, determining the origin may be easier with a simple table of key dates for each country.

In summary the Periscopic entry was: effectively able to isolate pandemic symptoms from general illness using the line graphs and heat matrix; unable to correctly locate the origin of the outbreak based solely on its map animation; somewhat able to characterise the spread of the disease using their map animation; and able to identify some key anomalies in the data, such as the

System	Aggregate Symptoms Visualisation
What: Data	Table of hospitalisation and death records for several major cities and countries
What: Derived	Several major symptom groups based on cleaned hospitalisation records
Why: Tasks	Locate origin of the disease and characterise its spread
How: Encode	Line graph of derived major symptom groups' observed frequencies over time; heat matrix of derived major symptom groups over time; scatterplot showing observed death rates

Table 2. Summary of the Periscopic visualisation.

preliminary peak in symptom frequencies, and it was able to do this using simple and quickly understood tools.

2.1.3 VAST Challenge Entries Conclusions

The two entries analysed here differ quite significantly in their approach to tackling this problem, and they complement each other quite well. The giCentre entry was strong in its overview of the data, using histograms and map data effectively to determine the origin and characteristics of the spread of the disease. However it was not as strong in identifying symptom relevance in terms of the overall pandemic. The Periscopic entry on the other hand was somewhat weak in terms of determining the origin and spread of the disease, but managed to provide significant insight into the importance of different symptoms as well as the timings at which they occurred. If the strength of the two solutions could be combined into one application then this would be quite a strong overall tool for use in pandemic visualisation.

There were of course many other entries to the VAST Challenge 2010 and in general they followed the trend of using histograms to analyse how the hospitalisation and death records changed over time, with some incorporation of map data. The main variance (with the two most unique examples given here) was in identifying disease symptoms as separate from general illnesses.

2.2. Pandemic Visualisation Literature

In addition to specific visualisation applications such as those found in the VAST Challenge entries, there exist a number of pandemic visualisation tools that exist in the general literature with real world examples. These tools have been created professionally within the domain of health policy and medical research and so warrant close analysis, particularly as they may provide insight into visualisation trends and techniques that exist in real world applications.

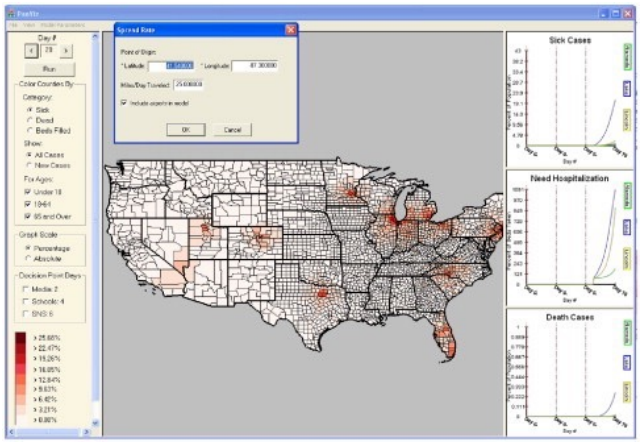
2.2.1 A Pandemic Influenza Modeling and Visualisation Tool

The National Strategy for Pandemic Influenza in the US outlines plans for community responses to pandemic outbreaks of the influenza virus. These plans require a certain level of understanding of a disease outbreak in its early stages, and so researchers at the University of Purdue developed the PanViz tool to aid in this understanding, and hopefully to aid in containing and controlling these outbreaks[8].

The PanViz application is intended to be used by public health officials and decision makers for two main purposes: firstly to model a pandemic outbreak as it is happening and so provide guidance and suggested courses of action; but secondly to model potential pandemic outbreaks based on previous data, and allow US counties to measure their preparedness and adjust their emergency responses in case of an actual outbreak.

For the first purpose of modelling a pandemic outbreak while it is happening, the application provides a host of modelling tools incorporating as much available data as possible to give a clear picture of how the outbreak is progressing. The interface, shown in figure 6, is typical of pandemic visualisation tools with a main view showing a map of the affected area, and smaller related views showing more detailed statistics of how the outbreak is progressing over time. The authors claim that the data leading into the application can be updated across the country in near real-time as actual data is obtained over the course of the pandemic, making it potentially a very powerful tool for public health officials when an outbreak is occurring.

The primary choropleth map in the centre of the interface can be overlaid with a range of relevant data such as number of infected patients, or number of available hospital beds. There are a number of filtering options on the left to better analyse a particular subset of data, and this data can be stepped through day by day, with the map updating correspondingly. Individual states can be looked at in more detail by isolating them and zooming in on them if required. The map uses thick black outlines for each US county, and thicker ones for each state, which somewhat confuses the data being represented. The data is on a colour scale from white - typically no infected patients or a lot of free beds, points of less interest - to deep red/maroon - typically many infected patients or few free beds, points of higher interest. As can be seen in figure 6, areas of the country with more counties in the same amount of area will look inherently darker due to the high density of black lines marking county borders, with a similar affect for areas with smaller and more dense states. This may confuse the data being shown as the most affected regions will look quite dark, and so having this high number of black borders may lead to perception errors. Instead it may be more effective to use grey borders with a high transparency so they are less noticeable, or even to have an option to switch them off altogether as the strict county borders



6. The main interface for the PanViz application. In the centre is a choropleth map of all US states and counties, displaying data selected on the left. The right shows several line graphs of important changes in the outbreak over time.

will not always be a primary point of interest when looking at this data. As a possible extension to this, it could be helpful to place this map in the context of a world map, even without relevant data for other countries, simply for a sense of context as at the moment the background grey space is fulfilling no purpose.

The line graphs on the right of the interface are important to gain a more detailed understanding of how the outbreak is changing over time, what stage it is currently in, and thus what actions need to be taken in response. They can either display only the data currently available, or instead they can show projections based on mathematical models of typical pandemic outbreaks, taking into account the available data for this particular outbreak. They are a simple and effective way to quickly gain more insight into how the outbreak is progressing.

As mentioned earlier, these main tools can be used to display information on an outbreak as it progresses, or they can show a model of a potential outbreak. This modelling process may be equally important as it allows a county to measure its preparedness in the case of an outbreak. The model incorporates a wide variety of parameters, many of which can be modified by health officials wishing to play out a specific scenario, such as incorporating early vaccination and social distancing. This also lets them make theoretical decisions and see how they play out by stepping day by day through the model of the outbreak and observing how the disease progresses.

As a visualisation tool, PanViz seems well catered to modelling outbreaks and thus testing preparedness, but is not quite as full featured as could be useful for a tool being used during an actual outbreak. For example, like in the Periscopic VAST Challenge entry, major symptom groups could be plotted over time along with major indicators such as number of deaths in order to provide early warning signs which would greatly aid in the decision making process around public health, assuming this data is available.

The authors of the paper mention that this is a somewhat difficult tool to validate as they are unable to test it on a real pandemic, and quantifying preparedness for an outbreak is very difficult so it is hard to validate its benefits through modelling. It is however being implemented in the Indiana State Department of Health and so real results may be apparent soon.

In summary this tool appears useful for public health officials and decision makers by allowing them to run through various potential scenarios, significantly speeding up the decision making process, however may not be as useful to medical researchers or those health workers looking to catch all possible early warning signs. It does however provide a good basis to work on, and if the system can be implemented with real-time updates as a pandemic progresses this would surely benefit those health officials and decision makers greatly.

System	PanViz
What: Data	Number of infected patients, number of deaths, number of available hospital beds for different US counties. Parameters in simulation may be altered by user
Why: Tasks	Understand spread of disease; model spread of disease to test for different counties' preparedness
How: Encode	Choropleth map overlaid with available data; line graphs of sick, dead, and hospital beds over time

Table 3. Summary of PanViz application.

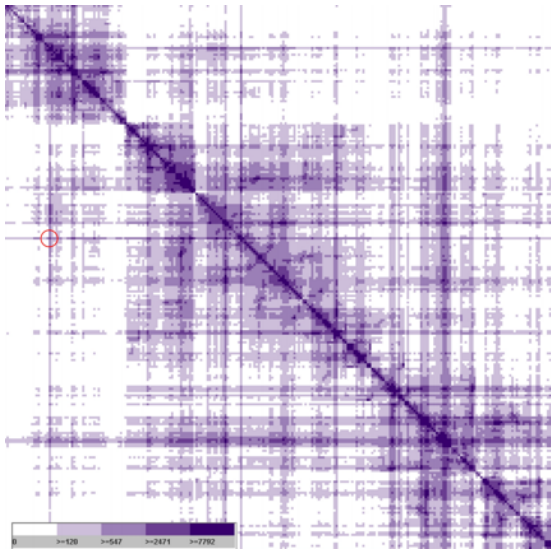
2.2.2 Visual Analytics of Spatial Interaction Patterns for Pandemic Decision Support

A large part of understanding and stopping pandemic outbreaks is understanding how they spread, and characterising this spread in some way. At a high level this can be achieved through observing choropleth maps over time (as have already been seen in this paper), or through flow maps showing general movement patterns. These techniques do little to help understand the spread of the disease at a low level however, and while there is often research into the disease itself and what capability it has to spread, there is typically less research on how a population's behaviour and movement patterns may contribute to the spread of a disease. This is most likely due to data for this being hard to acquire, as well as the huge scope of data that would need to be analysed in order to gain meaningful insight. Understanding these population dynamics however would be hugely useful for understanding how a pandemic disease spreads, particularly within a city, and would lead to vastly improved health policy surrounding response to pandemic outbreaks in their early stages.

As a result of this, a researcher at the University of South Carolina has been working on visualisations to explore the possible trends in social interaction data for individual-based activities. The data being examined here is based on complex simulation systems to generate "near-realistic" and very large data sets that indicate individuals' daily activities and social interactions for city-wide or even nation-wide areas[9][10].

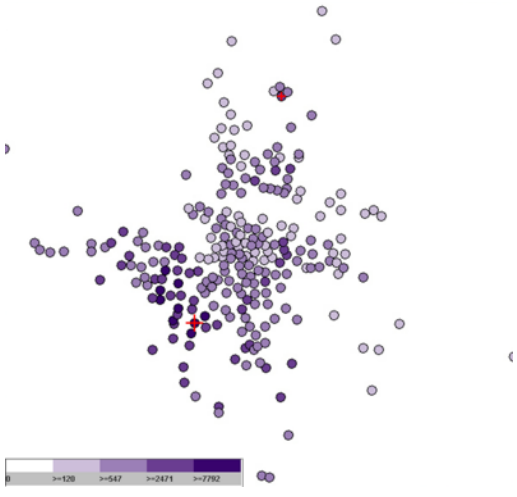
The first part of the data generated was for interactions between hundreds of thousands of different locations within the city of Portland, OR, and the second part of the data was for interactions for a simulated influenza outbreak. The interactions in these two datasets were considered as spatial interaction problems, with an interaction consisting of either a person visiting both locations, or the virus being transmitted from one location to another. Those locations with higher numbers of interactions were observed more closely, and groups of locations were clustered together breaking the large initial graph into several spatially contiguous subgraphs. The interaction between any two locations was given a quantifiable interaction strength, and this data was then plotted into an adjacency matrix for the Portland city data, and the influenza data, as seen in figure 7.

The adjacency matrix generated is reordered from a standard ordering (eg location 1 through 100), with locations that share the same groups of visitors being placed next to each other. This reordering is intended to allow researchers to better see trends in the data, however it may also promote visualising trends where there are none. This is naturally a very fine line and is sometimes unavoidable when dealing with large, effectively unordered sets of data, and the main objective is finding trends and patterns. The data needs to be ordered in such a way that will allow these trends and patterns to be identified, but it also should not be modified too much in order to promote unwanted pattern finding. The adjacency matrices used here appear to do a good job however, particularly as subsections of the matrix can be more closely examined for further analysis if a pattern or trend is identified.



7. Spatial interaction matrix (248 x 248) for the Portland city data, a similar matrix is generated for the simulated pandemic data. Each cell in the matrix represents the interaction strength between two locations. The red circle marks a selected row which can be analysed in further detail (figure 8).

This closer examination is done by selecting an individual cell in the adjacency matrix, and this will generate the same interaction strength data but instead using circles in an area representing real spatial distances. This type of view is much more suited to understanding the geographical implications for the spread of the disease, once an area of interest has been identified by the adjacency matrix. While the research involved in this paper is still at an early stage, so naturally visualisations are somewhat still in development, it seems like it would be very useful for the sake of context to have this map superimposed onto a map of the relevant area. While not necessary, it would help in the user's ability to quickly digest and understand the information they are being presented with. The paper refers to this view as a flow map, however it appears more as a map for a single temporal reference,



8. Interaction strength data on a geographically accurate area showing interactions between one location and all others.

representing the interaction strength between each location on that day. Time can be stepped through however, with the map or adjacency matrix updating itself based on the particular day chosen. While it is useful to have this time stepping ability, the ability to gain an overview of the data over time would be much more useful, but potentially not feasible at this stage as is mentioned in the paper.

In summary, the work being carried out here is less of an overview like most broad-scale pandemic visualisation tools, and more of a detailed and specific look at one aspect of pandemic outbreaks. This one aspect however is a crucial part of understanding how these diseases spread, and could provide very useful information to public health officials and decision makers, such as which areas they need to focus on, and how they might go about preventing high numbers of interactions leading to increased spread of the disease. The author of the paper has mentioned himself however that the visualisations at their current stage are somewhat unintuitive, particularly due to the unintuitive nature of spatial interaction patterns. This information is hugely important though, and so this area of study definitely warrants further research.

System	Spatial Interaction Patterns
What: Data	Simulated individual human interaction data for major metropolitan areas
What: Derived	Clustered areas creating several spatially contiguous subgraphs
Why: Tasks	Understand how a disease may spread through analysis of human interaction patterns
How: Encode	Adjacency matrix showing cluster interactions; linked view showing interactions for one cluster with all others in spatially accurate area

Table 4. Summary of spatial interaction visualisation.

2.2.3 Genomic Analysis and Geographic Visualisation of the Spread of Avian Influenza (H5N1)

Several strains of influenza have resulted in worldwide pandemics, with a notable recent example being "bird flu" or H5N1 influenza. Wild birds can carry all strains of influenza and so any of these may be the source of the next major outbreak. It therefore seems worthwhile to study how these diseases mutate so that they may end up being transmitted to humans, and also the patterns in bird movements and behaviours that may cause this transmission from birds to humans. Researchers in the Department of Biomedical Informatics at the Ohio State University have come up with a visualisation tool to help understand how these strains of influenza mutate, and how they spread geographically[11]. At this stage it is currently unknown which avian hosts spread H5N1 and there are several competing hypotheses.

The visualisation is intended to identify previously unrecognised patterns, and then to go on and analyse these patterns for statistical significance. It was created using Google Earth and can be seen in detail in figure 9. The main aspect of the visualisation is the phylogenetic tree branching towards the Earth.

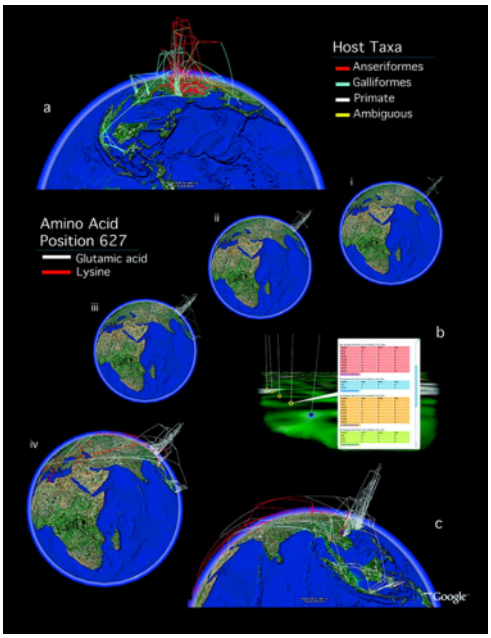
The root of the tree is the ancestral virus genome being looked at currently, with branches representing genetic mutations over time. Terminal branches represent genomes which have been at those locations where they terminate, and thus represent the current spread of mutations for a particular virus being observed. Different branches of the tree can be hovered over to get a "tool tip" type description box appearing with key mutation information and links to more detailed online data such as that found in GenBank.

The phylogenetic tree can be time stepped through as well, with different steps representing different points in the evolution of the virus as represented by the phylogenetic tree. In this way, key mutations can be looked at and how they change over time, allowing researchers to see how important these mutations are. If they promote the spread of a virus and are carried on then they should be looked at more closely for potential impact and spread to humans, however if they decrease in retention over time then they may not pose such a threat.

Included in this visualisation is the ability to compare several different host taxa at once for the same type of virus, and so this can provide significant insight into which animals may need to be monitored more closely to prevent future outbreaks. It will also provide insight into questions such as whether it is primarily the trade of poultry that is promoting the worldwide spread of these viruses, or if it is instead wild bird migrations. Colour is used here to encode the different host taxa, or to encode a specific mutation depending on the data being studied. This use of colour is clear and effective as it is showing a qualitative difference between only a few categories, making it obvious to the user which parts of the phylogenetic tree belong to which category. Colour is also used for the globe as part of the Google Earth software package which may cause some conflict in interpreting channels. It is unlikely that there would be serious confusion due to this conflict, but this visualisation may be somewhat clearer if the globe were simplified to either two colours (one for land and one for ocean), or a grayscale model again using different shades for land and ocean.

Despite several potential visual links being discovered for further analysis, no statistically significant conclusions could be reached due to the quality of data being used. There is a significant lack of metadata associated with much genomic data being stored, as this generally requires manual editing and annotation. Unfortunately this means that relevant avian data such as behaviour and migrations is often missing from genomic virus data, and so makes analysis more difficult. Additionally, it is suspected that much of the trade surrounding poultry happens illegally, with several major cases confirmed, and so any data on this is naturally somewhat unreliable. Finally, data from sources such as poultry farmers is often not made publicly available due to fear of forced culling of infected stock without compensation, and so often the data being looked at is an incomplete picture.

The tool itself however appears to be a sound and novel approach to visualising the migration of different influenza viruses associated with different host taxa. Once the data in this domain becomes more freely available, or the metadata associated with virus genomic data improves in quality, this visualisation tool may well be helpful to public health officials and decision makers in terms of monitoring animal migration patterns, and how this may lead to a pandemic outbreak.



9. Visualisation using Google Earth and a phylogenetic tree branching towards the Earth. The top shows a phylogenetic tree for a virus colour-coded for different host taxa. The middle succession of four images (starting middle right, finishing bottom left) shows a phylogenetic tree progressing over time. The bottom right shows a description box with mutation information.

System	Genomic Analysis and Geographic Visualisation of Bird Flu
What: Data	Host organism movement patterns; various strains of bird flu including their host organism, location they were discovered, and mutation history
Why: Tasks	Monitor host organism movements and virus mutation and movements to predict outbreak occurence
How: Encode	Phylogenetic tree overlaid onto a 3D globe with the root outwards and branches touching down in relevant locations

Table 5. Summary of bird flu visualisation

2.2.4 Visualisation Literature Conclusions

There are many examples of expected pandemic visualisations in the literature, following the approach of overlaying infection data on a map and showing its progression over time. This approach appears to be very effective at locating the origin of an outbreak when combined with more detailed statistics, potentially in the form of line graphs or histograms, about infection rates for different areas over time as well as involved symptoms. The interesting cases to examine then are those that are trying to provide a deeper insight, allowing more effective responses to outbreaks by showing trends in spatial interaction data or influenza mutation patterns for example. The expected pandemic

visualisations are a very important and strong base to begin pandemic disease analysis, however this deeper research will hopefully leave us with the ability to understand the intricate details of these diseases and how we might better respond to them.

3. Task and Data Discussion

Based on the analysis of the VAST Challenge entries and the general literature surrounding pandemic disease visualisation, there are many important facts to be learned for those looking to create their own visualisations. The rest of this paper will be about creating a visualisation incorporating lessons learned from those past attempts in order to assess two major software packages aimed at creating visualisations - Tableau and Lyra. Firstly a final design will be decided on in terms of what needs to be created in order for these visualisations to achieve their required tasks. Once this has been completed, Tableau and Lyra will be assessed individually and comparatively in order to determine their suitability for creating visualisations in this domain, and recommendations will be made based on these tools and what else is currently available.

3.1. Design Justification

First we must define what these visualisations need to achieve and why. The visualisation should be able to:

1. Locate the origin of a pandemic outbreak. This is an important step in terms of understanding how the disease originated, for example which animal it might have come from. This is very important in terms of understanding the physiology and pathology of the disease, and thus finding a cure or vaccine. It is also important to find the origin point as this will then aid in understanding how the disease has spread since then.
2. Characterise the spread of the disease. It is important to understand how the disease spreads so that we can implement better strategies to control it. This may also aid in understanding the pathology of the disease.
3. Provide symptom analysis. This is helpful in understanding what health workers need to look out for, and may provide as an early warning sign that the disease is spreading. Isolating pandemic cases from general illnesses will be hugely important to this.
4. Within each of the previous three, identify anomalies that may contribute to a better understanding of the disease and its spread.

Any visualisation that can complete these four tasks effectively will be of help in terms of effectively understanding and controlling pandemic disease outbreaks. The application generated may be tailored to more scientific work, with an increased focus on discovery, or for public health officials and decision makers for either personal use or presentation, in order to aid in the decision making process regarding controlling outbreaks.

The data being used here will typically be for hospitalisation and death records for several different cities and countries. This data will need to be cleaned before being loaded into the visualisation tool, so the tool will not need to clean the data itself.

3.2. Description and Analysis of Solution Idiom

In order to address the tasks outlined above, the solution idiom will consist of several components.

Firstly, a map view. This map will preferably be a standard rectangular map of the world, with infection data overlaid on top of it. The infection data will be represented on the world map as circles of varying area, representing suspected number of pandemic cases. While circle area is not an ideal channel for accurate data analysis, when overlaying onto a world map it is not possible to use a one dimensional length method, and this analysis is more about trends than exact figures so circle area should suffice. The user should be able to time-step through the data, showing the relevant infection data for any given day. This map view will help in achieving tasks 1 and 2 outlined above.

In complement to this, line graphs of other relevant data such as hospitalisation rates for all given cities should be viewable at the same time as map data. This will also aid in completing tasks 1 and 2 outlined above. These graphs should be vertically aligned so that trends may be more easily spotted and anomalies more easily detected (task 4 outlined above). The user should also be able to select any one of these graphs and enlarge it to perform a more detailed analysis.

Separate to the above views, line graphs of major symptom groups over time should be implemented so as to allow users to perform detailed symptom analysis (task 3 outlined above). This need not be in the same view as the map view as the two are not relevantly linked, however it would be useful to have symptom data for various countries, all vertically aligned, so that anomalies may be detected (task 4).

For all of the above views, the user should be able to filter data as they see fit so that only relevant data or points of interest are shown if they wish to study these in further detail.

It should be noted here as well however that these details are for the construction of a very general pandemic visualisation intended to be applicable to a broad number of cases. There are many cases however where the user would want something different, such as if they're only looking at one county or state within the US, circle area becomes a poor choice for encoding infection data as a choropleth map would give a much better overview of the information. This is because the circle area is condensing a spread of information into one point, whereas a choropleth map can somewhat accurately represent a spread of data, if separated into chunks. Various minor adjustments like these may be made, however the aim of this exercise is determine the suitability of different software packages to visualisations in this domain, and so minor adjustments will not make a major impact on this. It is assumed that if they can create a visualisation as outlined above, then they will also be able to make any visualisation with minor adjustments.

In addition to all of the above components, the software package should allow for extensibility. This does not necessarily mean adding more views or information into the already created visualisation, but new research on pandemic disease visualisation is continually being carried out, and new insights can be gained from these new approaches. For this reason it may be crucial for these software packages to create visualisations such as those based on spatial interaction, and so they will be somewhat assessed on extensibility.

4. Results

The results obtained here are for pre-existing software packages involving minimal coding (beyond running scripts to clean and format the data appropriately), due to insufficient time and also to determine what options there are for public health officials or decision makers to create their own visualisations with a small amount of training on the relevant software package. Recommendations will also briefly be made on possible programming options for this domain as well.

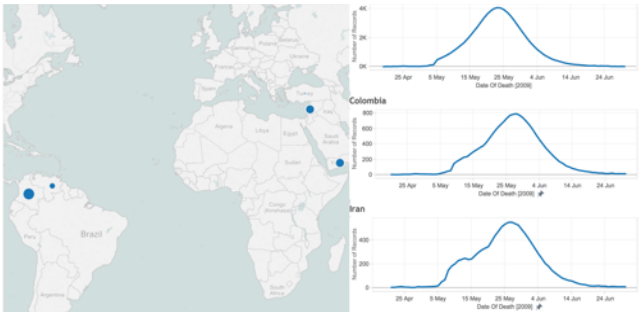
4.1 Tableau

Tableau Desktop[12] (made by Tableau Software) is a very extensive "do-it-yourself" type software package for making visualisations, and at the time of writing it is widely considered to be one of the most powerful non-programming approaches to visualising data. It was primarily developed to aid in business analytics, and this remains its primary use, however solutions have been developed using Tableau across a wide range of domains such as government analytics, healthcare analytics, and education reporting. As most of the tasks required by pandemic visualisation tools can be completed using relatively basic visualisations (such as line graphs, histograms, and geographic data) which are supported by Tableau, it seems an obvious choice to test for suitability in pandemic visualisations.

As can be seen from figure 10, a rough framework for a pandemic was able to be created using Tableau, and this was completed in roughly three hours. The death records for each country/city consisted of a column of dates, and a column of patient identifiers (up to a 6 digit number). This data was correctly interpreted by Tableau and was quickly able to be placed into line graphs or histograms, with relevant calculations such as sum of records grouped by date.

Trying to load geographic data quickly shows the limitations of Tableau. The data needs to be formatted in a way that Tableau expects it and there is little room for error. In this instance geographic data needed to be explicitly listed as a column, rather than being able to group data points and assign them a geographic location. For someone with a knowledge of shell scripting this is not a big problem, but if this program is to be used by public health officials who probably do not have experience with this, the application becomes somewhat less feasible. Once the data had been properly formatted however, it was relatively straightforward to get a map overlay of the information. After this an animation was created based on records for different days of the course of the infection which could be stepped through at will. Again this animation was very simple to set up once the data had been properly formatted, however getting to that stage required knowledge of shell scripting. The animation could also not be viewed at the same time as the line graph of death records, a single point for the map could be shown but the animation could not be stepped through.

Symptom analysis was easily carried out as well with the creation of line graphs based on symptom groups. This could be done creating custom groups in Tableau but this would be



10. Rough framework for a pandemic visualisation created using Tableau Desktop. Static map on the left showing one particular time point, with line graphs of deaths over time for different countries on the right

significantly slower than writing a shell script that searches for all relevant terms.

Linked highlighting was available in a limited form in some data views, such as selecting a data point within two linked data sets, however this took a significant amount of configuration to achieve. In terms of extensibility, Tableau offers quite a customisable interface which is well suited for future updates when customising a visualisation to include new techniques. The software package itself is also under constant development, and so incorporates new trends in visualisation as they arise.

4.2 Lyra and Other Web Packages

Lyra[13] is a web based Visualisation Design Environment (VDE) currently in beta, developed by researchers in the computer science department at the University of Washington. Web based visualisation tools are particularly interesting as the internet continues to grow and web tools become more powerful. In this case the quick delivery of a fully interactive and customisable visualisation for pandemic information could be a very useful tool when dealing with pressing decisions involving an outbreak. Despite being in beta, it looks to be one of the most full-featured web based visualisation tools available and so was chosen for further examination here.

Unfortunately not all tasks could be completed using Lyra, which is most likely as a result of its stage in development. Line graphs and bar charts could be generated from given data, although with some issues. Data import options worked well, however there were not as many immediate calculations (such as number of records grouped by date) as in Tableau. Once the data had been imported and selected for use in a line graph, the line graph would occasionally have unpredictable behaviour such as displaying upside down (with 0 on the y axis at the top of the screen), and sometimes would not display at all and the application needed to be reloaded. Similar results were obtained when attempting to perform symptom analysis, with occasional success when the same procedure was repeated several times.

For geographic data the visualisation would not load at all, despite explicitly selecting latitude and longitude fields. This was a general trend throughout the system, items disappearing and appearing at random, and items with very bizarre behaviour. This is naturally excusable as it's currently still in development in the beta stage, however it means that a functional visualisation was not achieved.

In terms of extensibility, as a tool still in development it holds promise for the future. As a non-commercial product it is difficult to tell how much further the project will be taken, but if all of the existing features can be made to work well without bugs then it should be a considerably powerful tool.

The promise of a web based tool for pandemic visualisation involving an interface where any public health official or decision maker can make edits as they see fit is an enticing one. This process is something that should be strived for, however unfortunately at this stage there is nothing up to the task. Several other web based tools were examined, such as IBM's ManyEyes[14], Google Charts[15], and CartoDB[16]. Unfortunately none of these have all of the features required for a sufficient pandemic visualisation, and are overall too simplistic. One notable exception to this was GeoCommons[17], a free visualisation tool created by a commercial firm called Esri that was examined briefly. This tool provided intuitive mapping abilities along with the option to create subviews showing more detailed line graphs

or bar charts within the main view. It was unable however to provide more advanced features such as linked highlighting, and was somewhat restricted in terms of ability to customise the look of the visualisation, and how the data was interpreted.

5. Discussion and Future Work

On the whole, it can be said with some confidence that none of the tools here are particularly suitable to the pandemic disease visualisation domain. Tableau can certainly be used to create a visualisation that allows for a quick high-level overview of pandemic disease information such as hospitalisation and death records. It was able to create a map with overlaid data that could be time-stepped through, along with subviews showing more detailed graph data, which are important for getting a quick understanding of the disease. It is however a very expensive software package, and the training tools associated with it are again very expensive. These training tools would most likely be required as public health officials may not be particularly technologically literate, and these tools may not even be enough as the results for this paper required the writing of specific scripts, which would most likely require hiring developers. It then becomes hard to argue for funds to go into licenses and training for Tableau, rather than towards development of a custom application that can be modified in the event of a new outbreak.

The web packages that were studied were on the whole either too simplistic or too early in development to create a sufficient visualisation. ManyEyes, Google Charts, and CartoDB could all perform one part of the required analysis well, but none could provide a complete package. GeoCommons came the closest to providing this comprehensive package, but fell short in areas involving linked highlighting, and customisation of the look of the application. As mentioned previously, Lyra is currently in too early a stage of development - particularly with regards to mapping tools - to create an effective visualisation.

This then begs the question of which tools will be effective for visualisations in this domain. As no programming options have been assessed, formal recommendations for these cannot be made, however there are some options which show particular promise for this domain. It would be useful here to have a web-based tool, as this would allow for quick distribution of the visualisation which could be very helpful in emergency scenarios where the general public needs to be informed of potential dangers. A web-based tool would also be easily updated as new information arises which is again useful in an emergency situation.

One prominent example is D3.js[18] which is currently widely used for web-based visualisations due to its thorough documentation and availability of online support. The feature set available allows for creation of interactive maps, with subviews showing more detailed information as well as the capability for more advanced features such as linked highlighting. P5.js[19] is another good example of a web-based tool based on the Processing approach, however is currently not as widely used as D3. For non web-based tools, Processing[20] is an example of a currently widely used language, built on Java and is again thoroughly documented with a large amount of online support. Not enough time was available to test these tools in detail, which could be a basis for future work.

There is currently a large amount of valuable research available for creating visualisations aimed at getting a high-level overview of data such as hospitalisation and death records[6][7][8], and further research here may be of limited value. Competitions such as the VAST Challenge promote progress in these areas, and the PandemView application developed at the City University London is a good example of how accomplished many of the tools in this area are. The goal for future pandemic visualisations then will be to aid understanding so that pandemics can be controlled more quickly, or that they may be prevented from happening entirely. This kind of goal may be achieved by visualisations of new kinds of data, such as those monitoring the movements of host organisms as well as mutations in the relevant infective organisms, or spatial interaction data for humans in metropolitan areas where diseases can spread quickly.

6. Conclusion

While current research into the area is sound and sufficient for gaining a quick high-level understanding of pandemic disease spread, there is still a lot to be learned about the dynamics of pandemic diseases so that they may be stopped more quickly, or prevented from happening entirely. These more advanced methods of control will only be achieved through analysis of different types and sources of data, which may currently be too unreliable for meaningful insights to be realised. Continued research into visualising new sources of data should yield improved results regarding control and prevention of pandemic outbreaks.

References

1. S. Haensch; R. Bianucci, M. Signoli, M. Rajerison, M. Schultz, S. Kacki, M. Vermunt, D.A. Weston, D. Hurst, M. Achtman, E. Carniel, B. Bramanti, Besansky, Nora J., ed. "Distinct Clones of Yersinia pestis Caused the Black Death". *PLoS Pathogens* 6 (10): e1001134, September 2010.
2. "Tuberculosis Fact sheet N°104". *World Health Organization*. November 2010
3. "Malaria Facts". *Centers for Disease Control and Prevention*. Retrieved November 2014
4. "Vast Challenge 2010". http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202010/challenges/MC2%20-%20Characterization%20of%20Pandemic%20Spread/ *Visual Analytics Benchmark Repository*. Retrieved October 2014
5. T. Munzner. "Visualization Analysis and Design" *CRC Press*. 2015. ISBN: 978-1-4665-0891-0
6. J. Wood, J. Dykes, A. Slingsb. "Hospitalization Records - Characterization of Pandemic Spread". *Vast Challenge 2010*. (2010) http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202010/challenges/MC2%20-%20Characterization%20of%20Pandemic%20Spread/entries/giCentre,%20City%20University%20London/
7. K. Rees. "Hospitalization Records - Characterization of Pandemic Spread". *VAST Challenge 2010*. (2010) http://hcil2.cs.umd.edu/newvarepository/VAST%20Challenge%202010/challenges/MC2%20-%20Characterization%20of%20Pandemic%20Spread/entries/Periscopic/
8. R. Maciejewski, P. Livengood, S. Rudolph, T. F. Collins, D. S. Ebert, R. T. Brigantic, C.D. Corley, G. A. Muller, S. W. Sanders. "A pandemic influenza modeling and visualization tool", *Journal of Visual Languages & Computing*, Volume 22, Issue 4, August 2011, Pages 268-278, ISSN 1045-926X
9. D. Guo. "Visual analytics of spatial interaction patterns for pandemic decision support." *International Journal of Geographical Information Science* 21.8 (2007): 859-877.
10. D. Guo. "Spatio-temporal visual analytics for pandemic response and decision support", *Workshop on Visualization, Analytics and Spatial Decision Support*. 2006.
11. D. Janies, A. W. Hill, R. Guralnick, F. Habib, E. Waltari, W. C. Wheeler. "Genomic analysis and geographic visualization of the

spread of avian influenza (H5N1)". *Systematic Biology*, (2007) 56(2), 321-329.
12. http://www.tableausoftware.com/ Retrieved November 2014
13. http://idl.cs.washington.edu/projects/lyra/ Retrieved November 2014
14. http://www-969.ibm.com/software/analytics/manyeyes/#/ Retrieved November 2014
15. https://developers.google.com/chart/ Retrieved November 2014
16. http://cartodb.com/ Retrieved November 2014
17. http://geocommons.com/ Retrieved November 2014
18. http://d3js.org/ Retrieved November 2014
19. http://p5js.org/ Retrieved November 2014
20. https://processing.org/ Retrieved November 2014