Video Script

This video presents our approach for solving the VAST Challenge 2011. We demonstrate the visualizations and explain the analytics used in order to solve each of the mini challenges.

MC1

1. To solve the mini challenge 1, we hav**e created this dashboard**
2. There are over a million tweets in the dataset, so th**e first step is to filter these twee**ts. The user can enter the search terms in th**e search bar** to filter the tweets on. The symptoms mentioned in the challenge description are given as the search terms by default.
3. We then use **Word2vec model to find the words similar** to these search terms. We did this because, there may be other words which are not mentioned in the search terms list but are still relevant for us. For ex. Tweets talking about headache, cramps are important and need to be considered for analysis even though these exact words not are included in our list .
4. Out of more than a million tweets, we extracted around 55000 tweets for the default search terms.
5. Once the tweets are filtered, we plot the filtered tweets on the **city map.**
6. To see the **trend in tweets**, We also found the number of relevant tweets seen each hour and plotted it across time. With the help of this line graph, we observed that there is a **spike in the number of cases on 18th** and the number of tweets are significantly high over the next few days. Thus we were able to find the time period when the outbreak started to spread.
7. Since now we have the affected area and date when the outbreak started, we try to narrow down the cause. Using the play button, we observed the rising spread of the disease in the affected regions. We see t**wo distinct regions of disease spread**. One is the regions Plainville and Smogtown which are along the river and the other is near Downtown, Uptown and Eastside. The intersection of these regions could be the possible ground zero location.
8. We select the approximate area of origin on the map and search for all the tweets around the date when the outbreak started. We also remove the search terms.
9. Now on this subset of tweets, we perform **part of speech tagging**. To find suspicious events and causes, we extracted the **most frequent nouns** that occur in these tweets. These pop up in the related events word cloud.

10. **Ignoring trivial words like day, it etc, there are 3 major events in the area in the given time period.** We investigated each of these events further. One of them turns out to be a bomb drill in downtown. The second was a car accident on Highway 270. The third is a truck accident on highway 610. After looking at the tweets for the three incidents above and searching for relevant words and events for each of these three events, we inferred that the truck accident is the likely cause. We **filtered the tweets further, now using the keywords like "truck", "accident", while still filtering on time and location**. We get a list of approximately 500 tweets which are about the truck accident. The location of the accident can be narrowed down to I-610 Bridge over the VAST River.

For part 2, of the mini-challenge we have to find **the mode of transmissio**n. We select the two regions that we have identified earlier one by one and observe the symptoms of that region.
For region 1, we can see the sympto**ms are mostly flu-like.** We can also see that the **wind direction corresponds** to the direction in which the disease is spreading
For region 2, we saw that the symptoms are related to **water-borne disease**. This region is also near the **vastopolis river.** Going back and looking at the truck accident related tweets, we found that the cargo from the truck was spilled in the river which may have caused the disease to spread.

The trend graph also plots the number of new and old users posting the tweets. We can observe that the **trend of new users reporting of any kind of symptoms is reducing day** by day.
From this we can conclude that the **outbreak is contained** and additional measures to control it is not necessary at the present moment.
Our approach is generalizable and the user can enter any search term, select any region and time period and get the related events. Using the same approach we were able to find solutions to various problems for this challenge.

MC2

For solving mc2, we have used Heatmaps as they work well in providing q**uick information at a glance.**

To find out the events of interest, we first parsed the different network log files that were provided and converted the data in a consistent format. Next, for each of the given network log file, we counted the entries for **every minute for all 3 days**. This would give us an idea about the **network traffic per minute**. Finally, we plotted 3 heat maps that show the network traffic for Firewall, IDS and security log files. Using these heatmaps, the CNO team member can very easily trace any anomalies in the company's network traffic. The team member can select the date of interest and then very quickly observe the entire computer network behavior for the entire day at a glance.

F**or day 1,** we saw very high rise in traffic in the Firewall heatmap. It started from 11:39 and lasted till 12:51. IDS heatmap also showed some anomalies during the same time. The IDS heatmap also showed some suspicious event from 11:15 - 11:41. On closer look at the Firewall and IDS logs for these timestamps, we figured out that there was a Denial of Service Attack and Port Scan in these time frames.

F**or day 2, f**rom the IDS and Firewall heatmaps, we found suspicious activities between 09:01 - 09:27 and 10:56 - 12:28. On drilling down further, these were identified as port scans.

We could not find any event of interest fo**r day 3**.

For the time duration of 3 days, we found interesting results using the Firewall and IDS logs but the **same approach did not work for the given security logs**. This is because the frequency of entries in the security logs **does not give enough information** to identify a potential security threat and additional filtering/preprocessing is required to better analyze the security log data.

Re**garding part 2, timeliness of the system,**

The interface shows data for the entire day in a glance and since the traffic is plotted per minute, the team member can identify the occurrence of **an interesting event as early as the next minute.** This approach can very easily be extended to accommodate real-time streaming data.

MC3

1. For mini-challenge 3, we were given 4474 news report and we had to find the news reports related to potential threats.
2. To tackle this challenge we performed topic modeling using Latent Dirichlet Allocation Algorithm to classify the news reports into different topics. After experimenting with various numbers, we found out that 100 gave a good distribution of topics over the dataset.
3. We are able to identify the topic from the word distribution in the data table. For example, topic 13 looks like it talks about stock market.
4. In this scatter plot, X-axis is the topic number and y-axis is the probability of the news article belonging to that particular topic. User can zoom over a particular topic or hover over the points to get the news article headline
5. We can select news articles from the scatter plot by the lasso select tool or the box select tool. The selected articles are then plotted as connected graphs; the nodes represent the news articles and the length of the edges denote the similarity between the end nodes. We have used cosine similarity to measure the similarity between different news articles. The slider can be used to set the desired threshold for similarity.

6. The user can then zoom in or select different nodes in this plot to view the news articles of interest which are listed in the datatable next to it.
7. We observe that some topics have a high probability of belonging to a particular topic. For example, on closely observing the news articles with high probability of belonging to topic 2 we are able to see that all these reports contain sports match scores. If we are not interested in this topic, we can simply select these news reports and click on “remove news items”.
8. If we find clusters that are of interest then it can be selected and added to another datatable that will contain all articles that require further analysis and investigation. To do this, simple select the articles and click on “add news item to selected list”
9. There is an option to save the progress made in the analysis so that we do not have to repeat these steps each and every time. Instead we can load the previous work and pick up from where we had left.
10. After performing all these steps, we were able to narrow down to around 50 news articles related to threats as shown in the table here.
11. On going through these 50 reports, The bioterror attack planned by the group ‘Paramurderers of Chaos’ stands out. Other threats include bomb threat and isolated incidents like missing military equipment, radioactive cargo, etc.
12. Just like the previous two mini challenge approaches, this approach is also generalizable. The user can use this interface to find out news reports related to any topic and is not constrained to find out threat related reports only.

END

Using the results of each of the 3 mini challenges, we are able to hypothesize the bioterror plot by the terrorist group called paramurderes of chaos, and the epidemic caused by the unintentional truck accident. This project demonstrated that analytics powered by visualizations can help solve complex real world problems.