Detecting Trends in Malicious URLs

by

Arick Althoff

DATA 230 - Project Report

5/23/2021

Abstract

Criminals are abusing the internet more and more for personal gain, and one weapon they use are links to websites that attempt to download viruses, steal your information, or redirect you to other malicious websites. This project revolves around helping the audience detect the trends in a malicious URL for study or personal knowledge.

The presentation informs the audience of the many tricks and trends a malicious URL is constructed from for the purpose of making others more knowledgeable about browsing the internet safely. URL length, certain symbols, certain words in a URL, and its website type may all be indicators that can be analyzed to determine a malicious website. This presentation attempts to apply the visualization tools and information that is used to create a comprehensive story, dashboard, and graphs that conveys information in a compelling and factual manner.

At the end of the presentation, an interactive dashboard is given to convey more compelling trends as well as effectively display information. The visual tools utilized in this project include proper use of color, data-ink ratios, structure, and many more.

1. Introduction

A malicious URL can be a link that redirects a phone or computer to another unwanted website. It could try to download software onto a person's phone, computer, or browser. Or, it could be a phishing website to attempt to steal information. Browsing the internet requires users to have some sense of safety as malicious links and websites steal personal information or spread viruses. The number of users subjected to malicious attacks increase every year and will keep rising as social media websites and more avenues of attack become more and more available. One of the best ways to protect yourself, is to become familiar with the ways these attackers try to attack you.

This project's objective is to familiarize the more common ways to recognize whether a URL is malicious in nature. Of course, the URL that you can see when previewing a website is not always the URL you want to go to as redirecting exists. However, those URLs may have malicious identifiers of their own. This project attempts to analyze the behaviors or trends of just the URLs to see if there are any identifiers that may make some links a higher chance to be malicious than others.

2. Use Cases

Pattern Detection

The first use case is merely an educational one to protect users from becoming victims of phishing attempts or malicious downloads. Using this information, users can pre-scan a URL to determine which parts of it may be deemed malicious. For example, if a certain file type has a higher percent chance to be malicious, they may attempt to avoid them in the future.

This form of pattern detection may not be sufficient in completely protecting yourself from the malicious parts of the internet. However, this information may be great introductory

lessons to someone who is just being introduced. There are many avenues someone can take to educating themselves on this topic, but this is at least one method that may be useful for those who are not technologically inclined.

Feature Importance for Machine Learning

Using this method, students or researchers aiming to build a prediction model with not much knowledge about URL's or their construction can determine which features may be more important than others. This method attempts to extract more features than necessary to analyze all of them and them with others in that category.

Feature Importance for Implementation

This project may reveal insights about the important features that can be used as an early detection model for an internet browser extension. Most of the protection browsers provide is proactive or reactive, proactive is where they go forth and crawl through all the pages in the web to detect malicious pages. Or reactive, where they use blacksites that report URLs to warn or prevent users from accessing those sites.

Neither protection scheme offers protection in a manner that combats these new methods that fraudsters are using, nor do they quell the actions of more curious internet users. Social media has become a platform where clicking a link gives a malicious user access to your page to post the link to your friends, all the while stealing information and private messages. Knowing what to look out for before clicking one of these links may assist people in the future.

3. Intended Audience

As mentioned above, the intended audience are those who may have just started to access the internet freely. It could help active users realize some more potential of malicious behavior

by revealing traits they weren't aware of. On top of that, it could reaffirm knowledge that experts may already be aware of.

Researchers may be able to use this information to determine the features or attributes of a URL that could be deemed malicious. This presentation is a comprehensive exploration of the data that may guide the process of students or researchers using this information. As mentioned above, this information could be used to create classification models to implement as a web crawler to classify thousands of sites each hour.

4. Data Source

The only requirement needed for compiling this dataset were URL strings and the classification for those strings whether they are benign or malicious. In total, there were two data sources on Kaggle.com used to construct the dataset. The first, was created in 2019 by Siddarth Kumar. This dataset was created from various sources that will be extremely useful for the visualization of this project. The first source was a couple blacklist sites which are sites that monitor and list reported malicious websites and whether or not they are actually malicious or not. The second source was a popular website list according to website traffic (Siddharth Kumar, 2019).

Another dataset created in 2020 by Ak Singh was compiled using a web crawler that classified websites using the Google Safe Browsing API (Singh, 2020, Content). This will assist in supplementing the 2019 dataset with more rows as well as add more unique rows that were generated in a different manner. Although, this dataset had 1,530,687 rows, so an equal sized sample was taken from this to let the visualization program run smoothly while still conveying the same message. The total number of combined unique rows is 896,229 URLs and classifications.

Additional features from the URL strings were created using Python programming on the data. In total, 28 features were generated including counts for certain symbols, binary checks for https, and domain features.

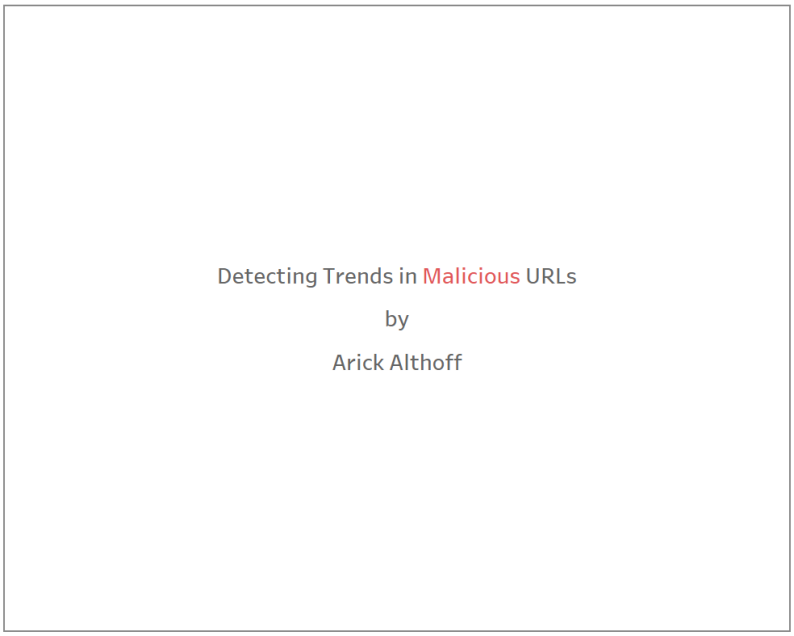
5. Presentation

The entirety of the presentation covers 9 slides and 2 dashboards. The slides are organized to reveal trends of malicious URLs and take a “factors” data story approach. In a factors approach, this presentation attempts to divide the main goal of detecting trends into different categories. Each item along the way contributes to the goal. The dashboards summarize those items as well as introduce a final point.

Before starting, most of the default values were changed to values that would be more in line of better visualization practices. For instance, the graph lines would be removed when not needed. Almost all text was changed from a 9 size font to an 11 or 12 sized font. And, attention was paid to the colors, axis titles, and axis marks.

Slide 1

Caption: Introduction



The first slide is very simple. As for color, it begins to introduce a recurring color theme where malicious is red. This will translate later into being representative of bars turning red, where the audience can immediately associate that as a malicious link. This will take advantage of pre-attentive processing later on where the audience will link the color symbolism quickly.

This mostly empty slide also allows the presenter to talk through and introduce the topic in general. This is the script used to talk over this slide:

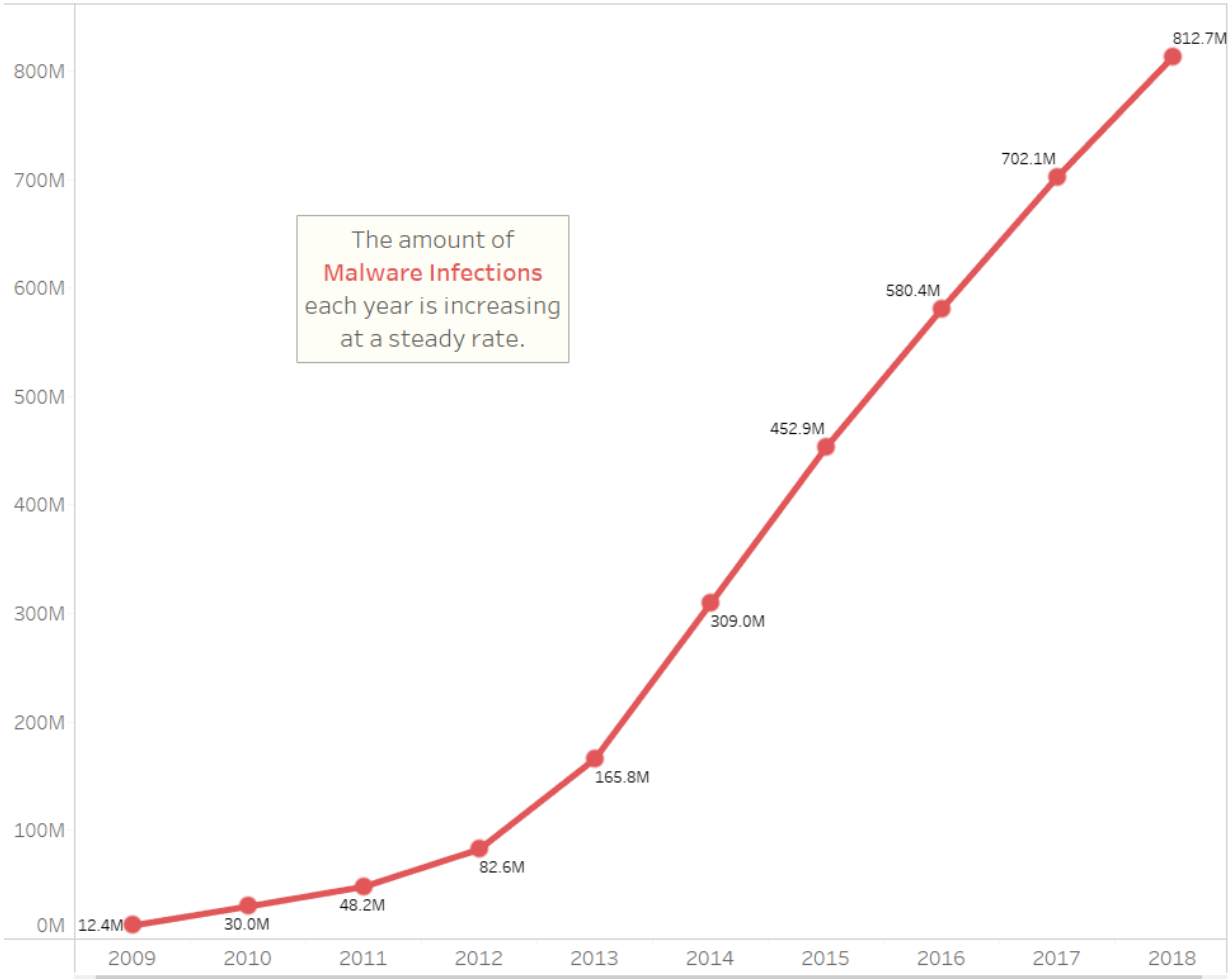
> "Hello, my name is Arick Althoff, and I'm going to present a story on detecting trends in malicious URLs. After a short presentation, I'll introduce a couple dashboards I worked on related to this topic. A malicious URL can be a link that redirects you to another unwanted website, it could try to download software onto your phone, computer, or browser, or it could be a phishing website to attempt to steal your information. This

presentation revolves around educating people on how to browse the internet safely, but browsing safely is more than what just some graphs can show you. You need to be vigilant and learn to be safe in whatever environment, whether it's browsing facebook, opening emails, or trying to find school textbooks online. Some of this information is pretty basic and may not be very useful to every student in the class, but this presentation may be used to educate a wider audience."

This script announces the goal that the presentation is trying to accomplish as well as introduces the dangers of malicious behavior. The description of what a malicious URL can do, may help gain the attention of more of the audience.

Slide 2

Caption: Malware infections increase each year.



As mentioned earlier, the red introduced in the first slide is the primary thing the audience will see. They will also notice a sharply increasing line as part of the pre-attentive processing.

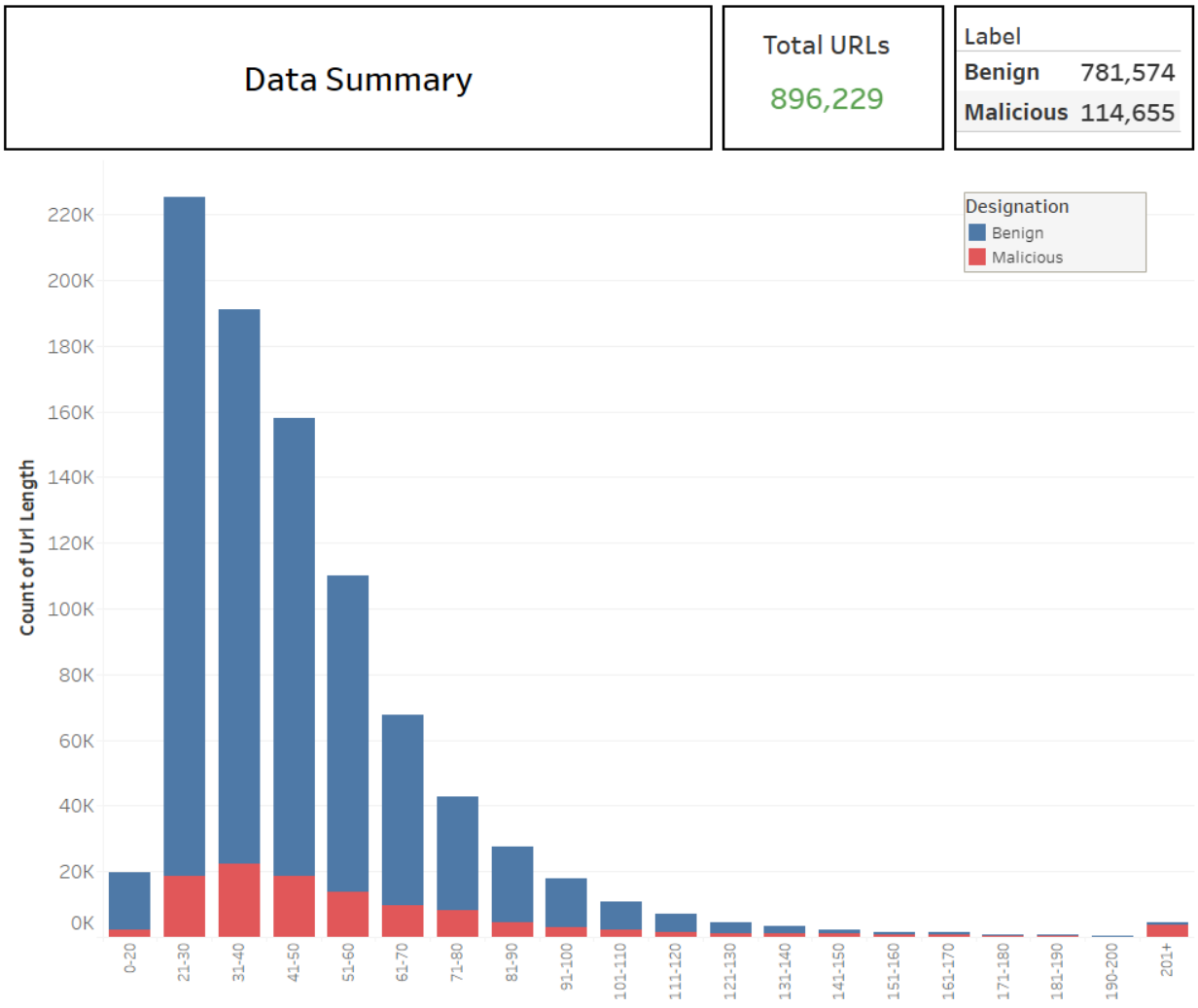
The annotation basically summarized what is going on in the slide and what the line represents. This slide is meant to pull the audience into the story and be engaged throughout. Some of the final notes that the audience may notice as they analyze the information is that the

numbers start at just 12.4 million and end all the way at 812.7 million, which is a very powerful message to leave the audience.

The grid was removed because there are already many indicators of the values in each year. The axis title of number of URL counts and year was also removed as the annotation makes it very apparent what the data is trying to convey. This increases the data-ink ratio.

Slide 3

Caption: About the data

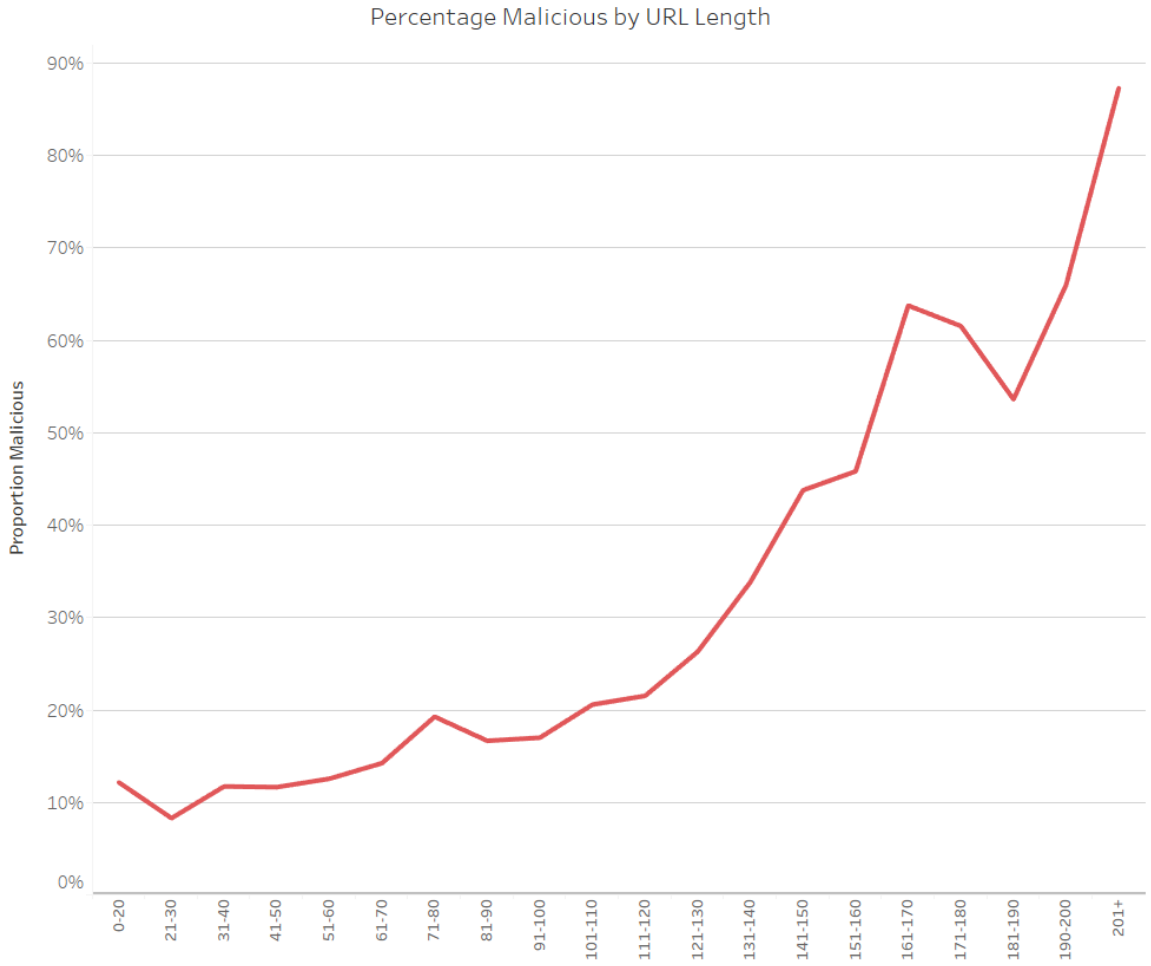


The slide takes into account visualization structure where the basic summary statistics are covered at the top as well as a large title. Again, users can easily distinguish the red malicious URLs in each bar of blue. And, this is the first time blue is associated with the benign label. Because of this, a legend is used to clarify what each color means. Notice that the chart lines are reintegrated at each twenty thousand mark to reshow the different counts of URL lengths.

This slide generally has the speaker talking about the data sources while the audience looks over the general distribution of the data across all URL lengths. The speaker then discusses the distribution of the URL lengths where certain lengths are more popular than others. Then, this discussion transitions into the next slide where the first discussion of malicious URL trends begins.

Slide 4

Caption: URL length may be an indicator of maliciousness

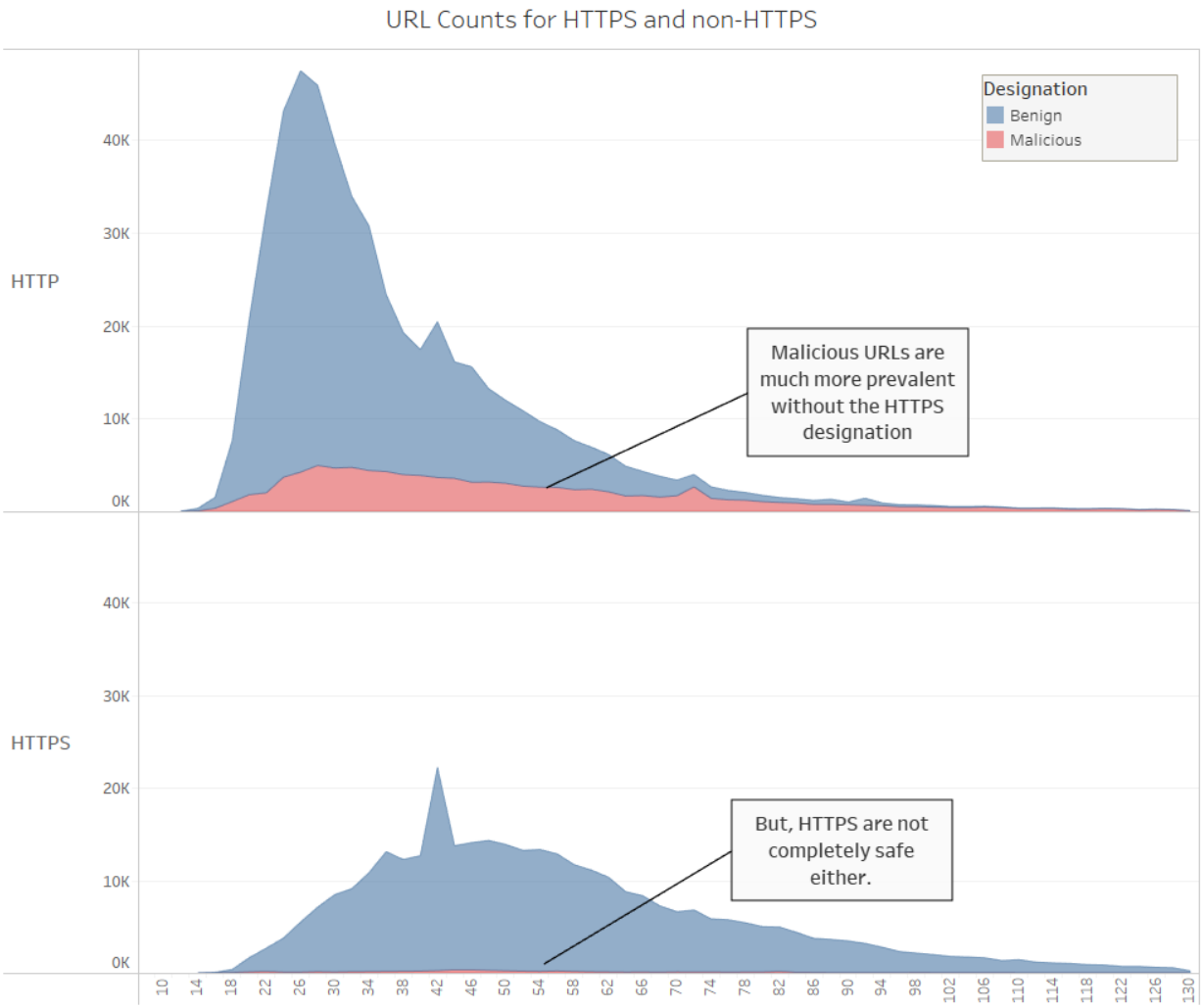


Since this chart is meant to display a trend, there is no real importance on the exact values for each URL length bin. Because of this, points are not highlighted and additional vertical lines are not placed. The values are also binned to smooth the jaggedness of the line. Also, since this is addressing the proportion of values that are malicious the line is red.

The speaker is responsible for talking about how the values begin at 12% which is the same percent of values that are malicious for the whole dataset. Then, the values decrease in popularity, but increase greatly in the percentage of maliciousness.

Slide 5

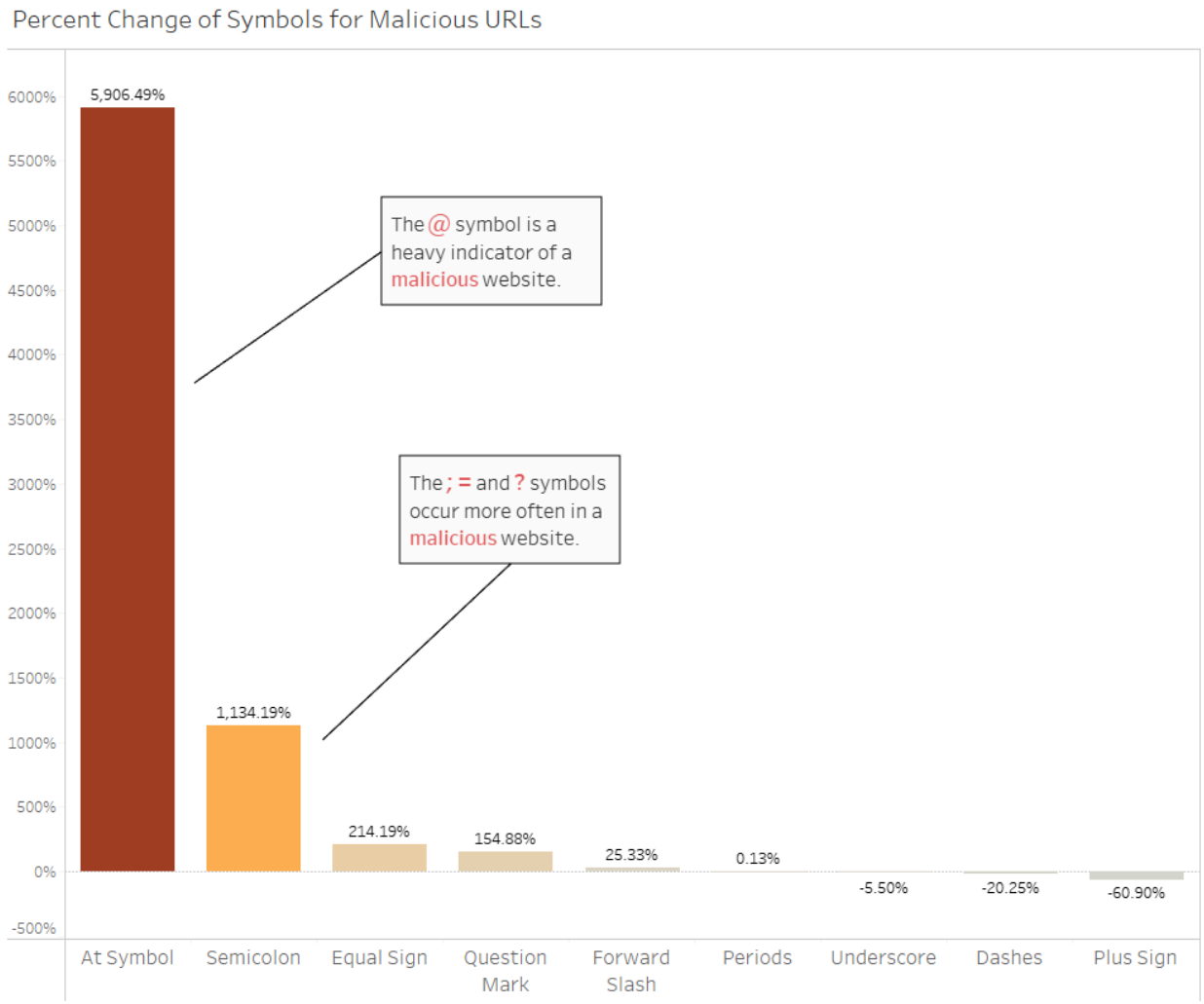
Caption: The HTTPS protocol protects users.



This is a visualization that is guided by annotations. The two sections are divided by HTTP and HTTPS protocols and charts the amounts, separated by URL length. This visualization could have easily been shown using a pie chart, but this iteration felt more enticing. The large chunk of red in comparison really translates to convey the message that HTTP protocols are much more dangerous. Such a small sliver appears in the HTTPS chart.

Slide 6

Caption: Number of symbols may be an indicator of malicious websites.



This graph is a culmination of multiple variables. Each bar is a formula for the percent change between the average number of symbols. The original version of this graph was a stacked bar plot, where each bar would represent malicious and benign average amounts. However, that method did not follow the Law of Simplicity, and the signals of which symbols were most significant would have needed to be closely analyzed. In this updated version, it is very apparent which symbols are most malicious in behavior.

The formula used to calculate each bar is the average number of symbols in total minus the number that are benign; all divided by the amount that is benign. Finally, that value is subtracted by one. This is an example of what the calculation looks like when measured in Tableau.

```
(ZN(AVG([Num @])) - LOOKUP(ZN(AVG([Num @])), -1)) /
ABS(LOOKUP(ZN(AVG([Num @])), -1))
```

One of the drawbacks of using this chart is that it may not be apparent how the bars were calculated or what they even represent. This can be curbed by the speaker as they introduce this chart. For example, the script used to introduce it this time:

> "This chart is examining the percent change in average number of certain symbols between benign and malicious URLs. So for example, the average malicious URL has almost 6000% more at symbols than a benign one."

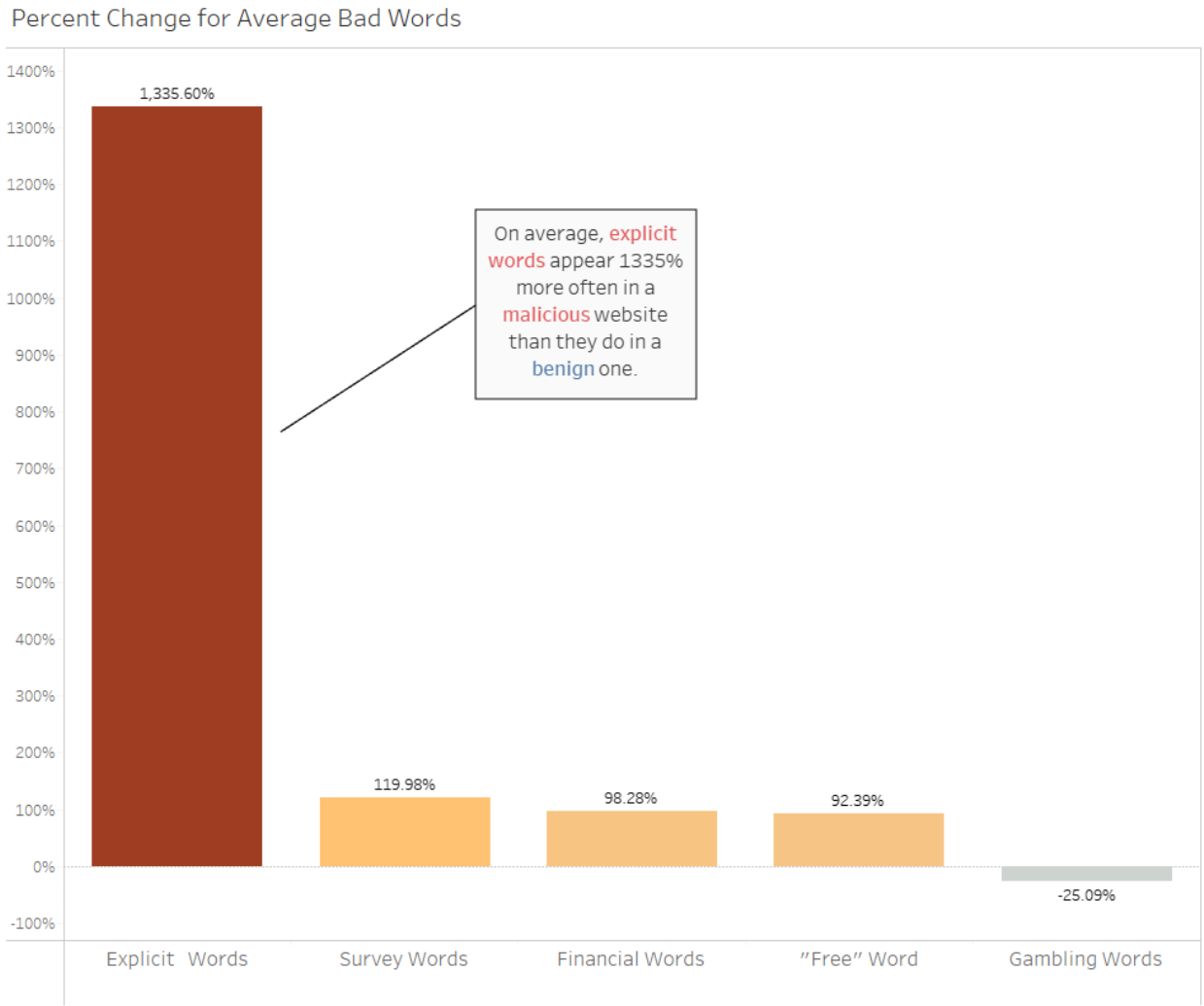
The first sentence immediately tells the audience what the meaning behind the bars are and what they represent. This knowledge will allow the audience to gain a deeper understanding when they fully process the chart.

In terms of color, this range breaks away from the traditional color scheme set by the previous slides for malicious values. The reasoning behind this is that each bar is a formula of benign and malicious values, and the color being used to represent red for maliciousness in previous slides would need to exist on a scale. In testing, using the light red wouldn't be able to capture the malicious behavior of the semicolon, as that would be a very light color. In this form, maliciousness still is represented by a dark red, and those dark red values are the most stand out

values. These dark red areas represent huge differences in the visualization, so by using the Law of Focal Point, attention is given to these values as they are most important. However, as the values move down in magnitude, the dark red turns orange and dark yellow for the values that still possess some threat.

Slide 7

Caption: Language in a URL may indicate a malicious link.

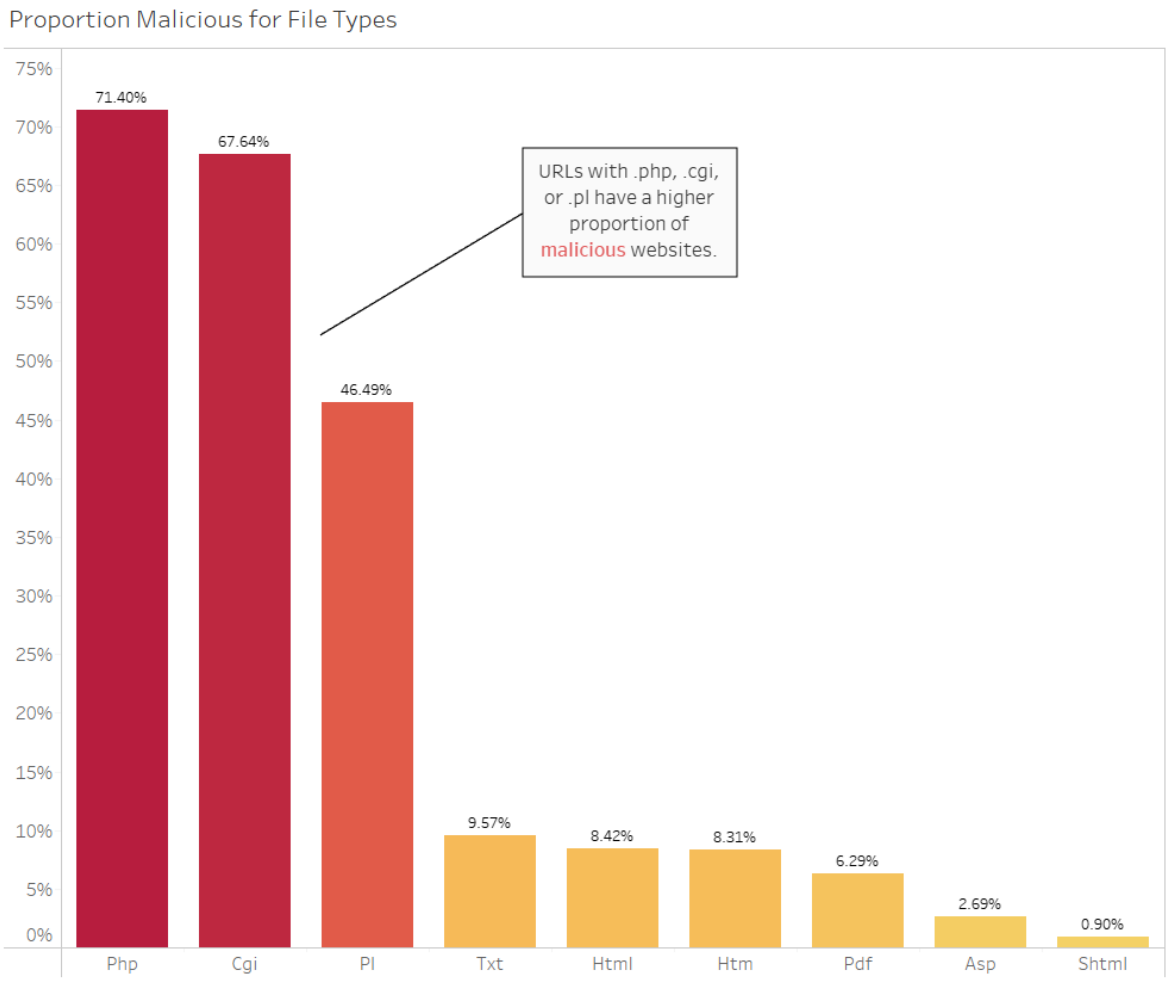


This slide uses the same coloring scheme as slide 6 and well as the same mathematically derived formula. These slides are purposely arranged together so the audience can easily decipher what each bar represents. The annotation is meant to further describe what the data is

trying to convey. In terms of color, there is again a stand out value in dark red with three other bars that are yellow. Because this presentation isn't focused on trying to find a benign URL, not much importance is given to analyzing "gambling words" by both the speaker and the coloring scheme. Instead of giving it a blue color for benign, a gray color was used to really represent unimportant information. This will keep the audience's attention on the values that are important.

Slide 8

Caption: File types online can be a method for malicious websites



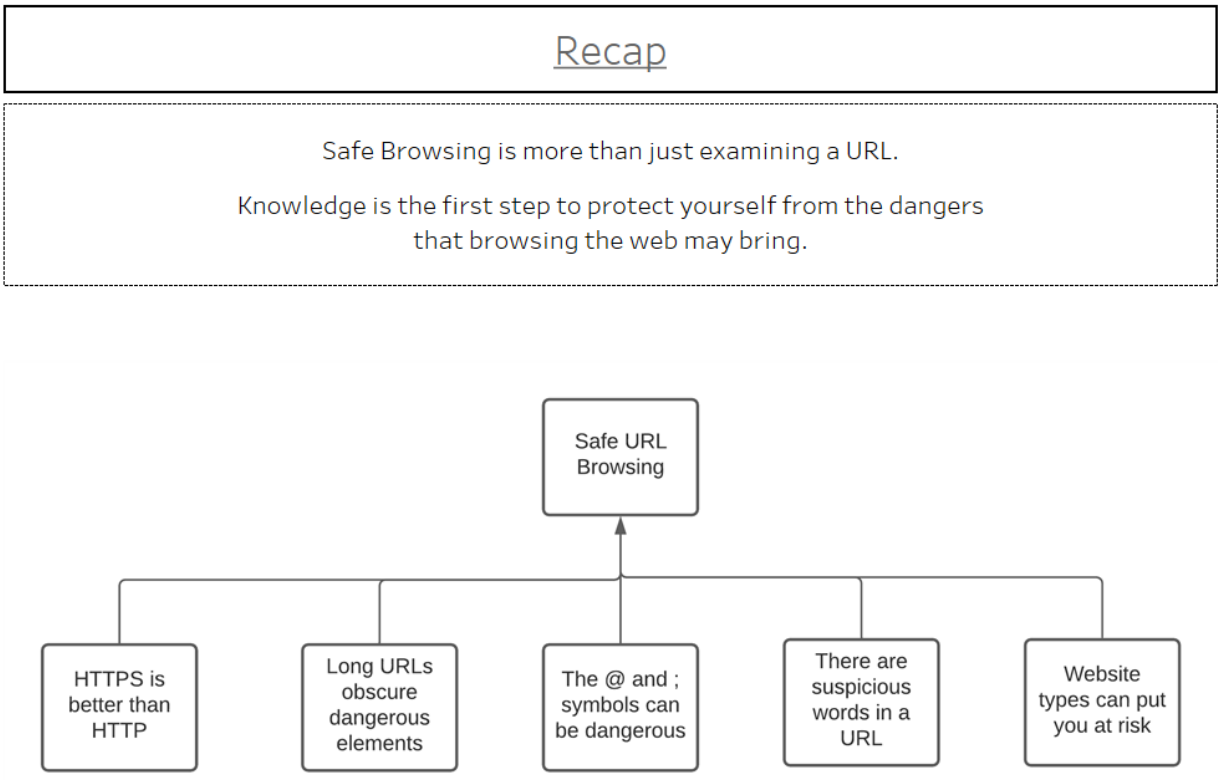
While this slide possesses the same coloring scheme as the previous two slides, it was not calculated the same way. The formula is the percent of values that are malicious for all values

that are a certain file type. This was calculated using Tableau's quick table calculation as a percent of total. The final formula appears as SUM([Is Php]) / TOTAL(SUM([Is Php])) for the "php" bar. However, the quick calculation is what makes the bars accurately portray the information.

Slides 6, 7, and 8 each have had their grids removed as each bar has the percent related to their values displayed at the top of the bars. This makes it easy for the audience to compare the values of each bar.

Slide 9

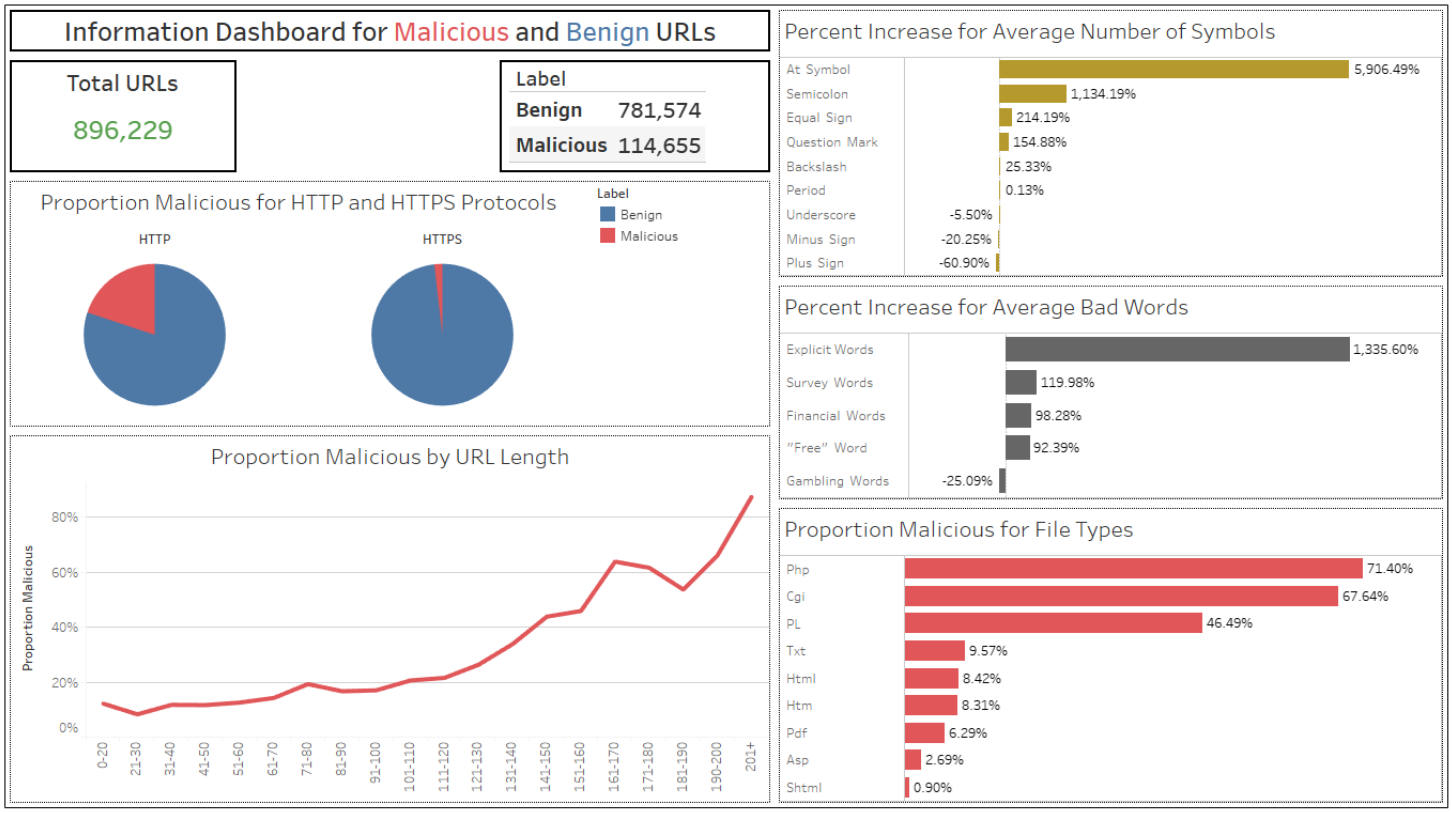
Caption: Recap



This slide is meant to summarize the information from the previous slides. This is more meant to prepare the audience for a switch in a presentation format to a dashboard. It does not possess any significant information but stands to remind the user the purpose of the presentation

as well as the trends observed. The chart at the bottom shows all the factors that were discussed in this story type.

Dashboard 1

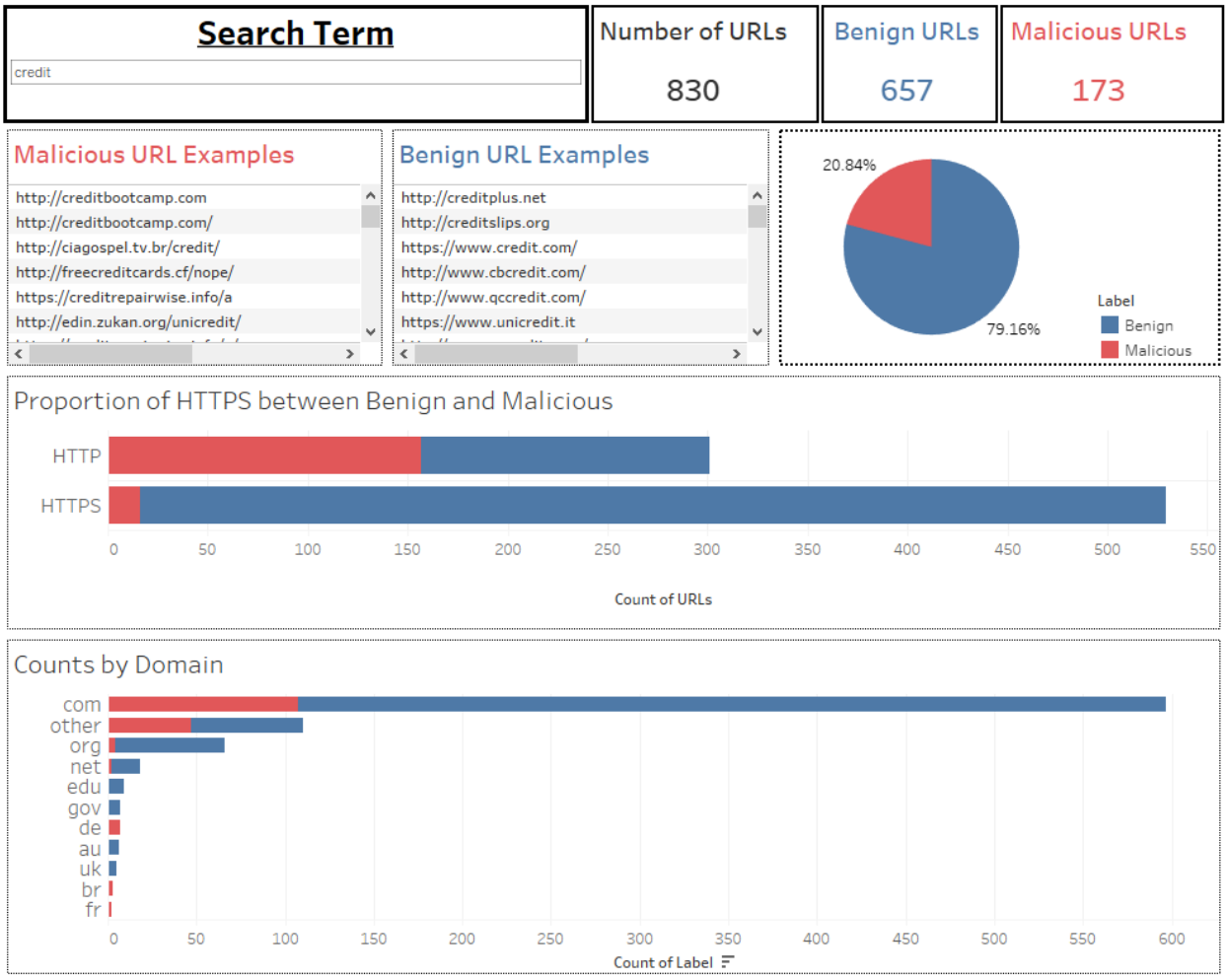


Dashboard 1 is a summary of all the trends observed in the presentation. The visualization structure of the dashboard is similar to the one in slide 3. This structure is similar to a book reading from left to right. As you move down the book, the information becomes more and more involved. The purpose of this dashboard is merely a summary, as such, no more insights would be taken from this dashboard.

The graph on HTTPS is represented by a simpler pie chart in this format. While the format in slide 5 still shows the same information, this format doesn't convey the exact severity as it seems like a small slice.

As for color, since this is an informational dashboard, the dark red and yellow tones in slides 6, 7, 8 were replaced by solid colors. Furthermore, each of these graphs use horizontal lines instead of the original vertical ones. This fits in with the structure of the dashboard much better than any representation of vertical bars.

Dashboard 2 - Search Term 1

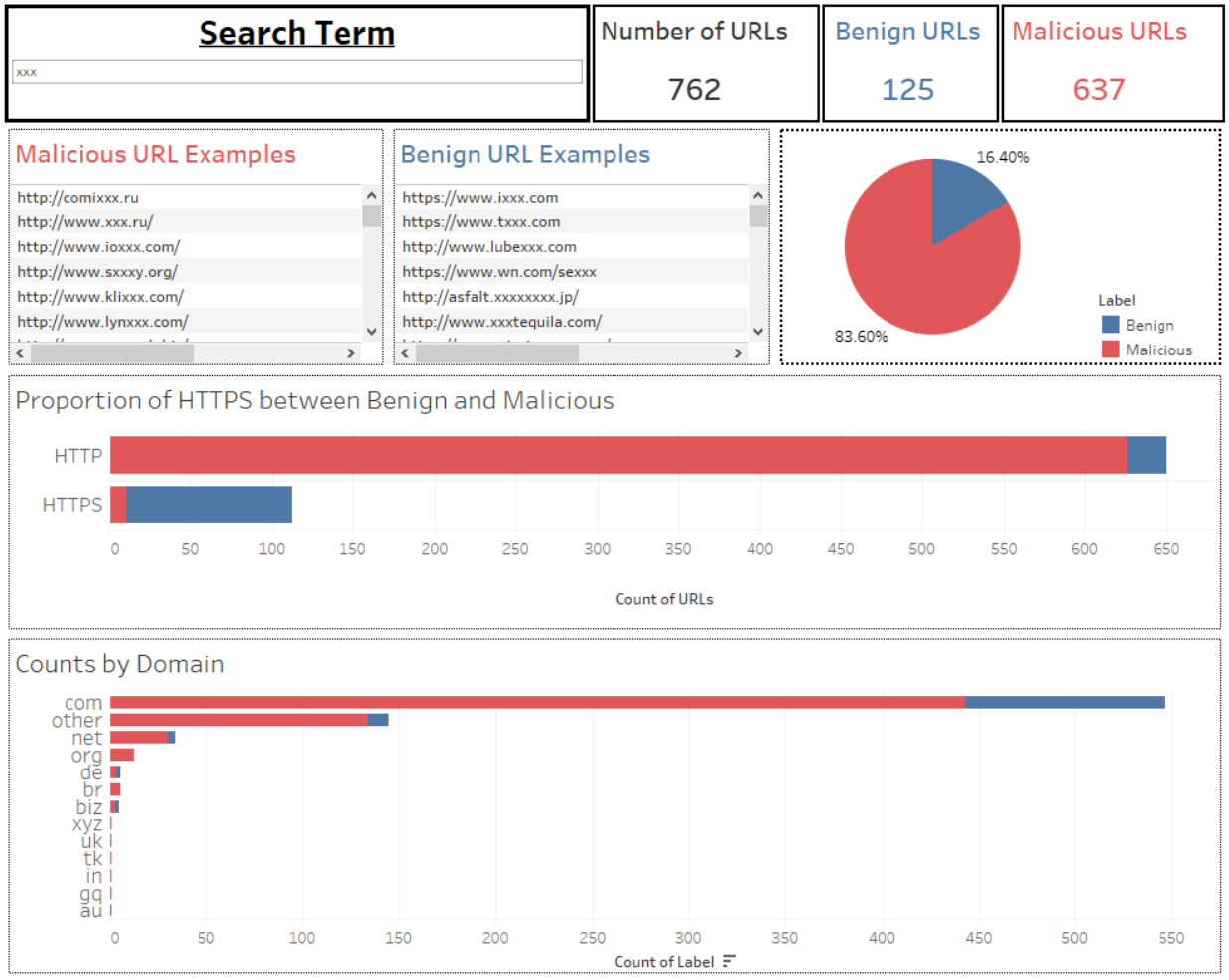


This is the dashboard that allows the user to interact with the data. The user enters a search term and Tableau searches through all URLs that have that search term inside them. In the example above, the term “credit” is being searched.

The structure of the dashboard is top-down, where the search bar and summary statistics are the most basic pieces of information. So, they reside at the top. Just below, lists of example malicious and benign URLs that actually contain the search term are given. And, to the right of that, a simple pie chart with the proportions listed are also shown. Then, HTTPS proportions are provided below. Instead of utilizing pie charts to display this information, a stacked bar chart can also display the magnitude of the counts of HTTP and HTTPS type protocols as well as their proportions. Finally, the same stacked bar chart is used to display maliciousness in domains as well. This could be useful in examining which websites are most popular, and which are completely malicious or benign.

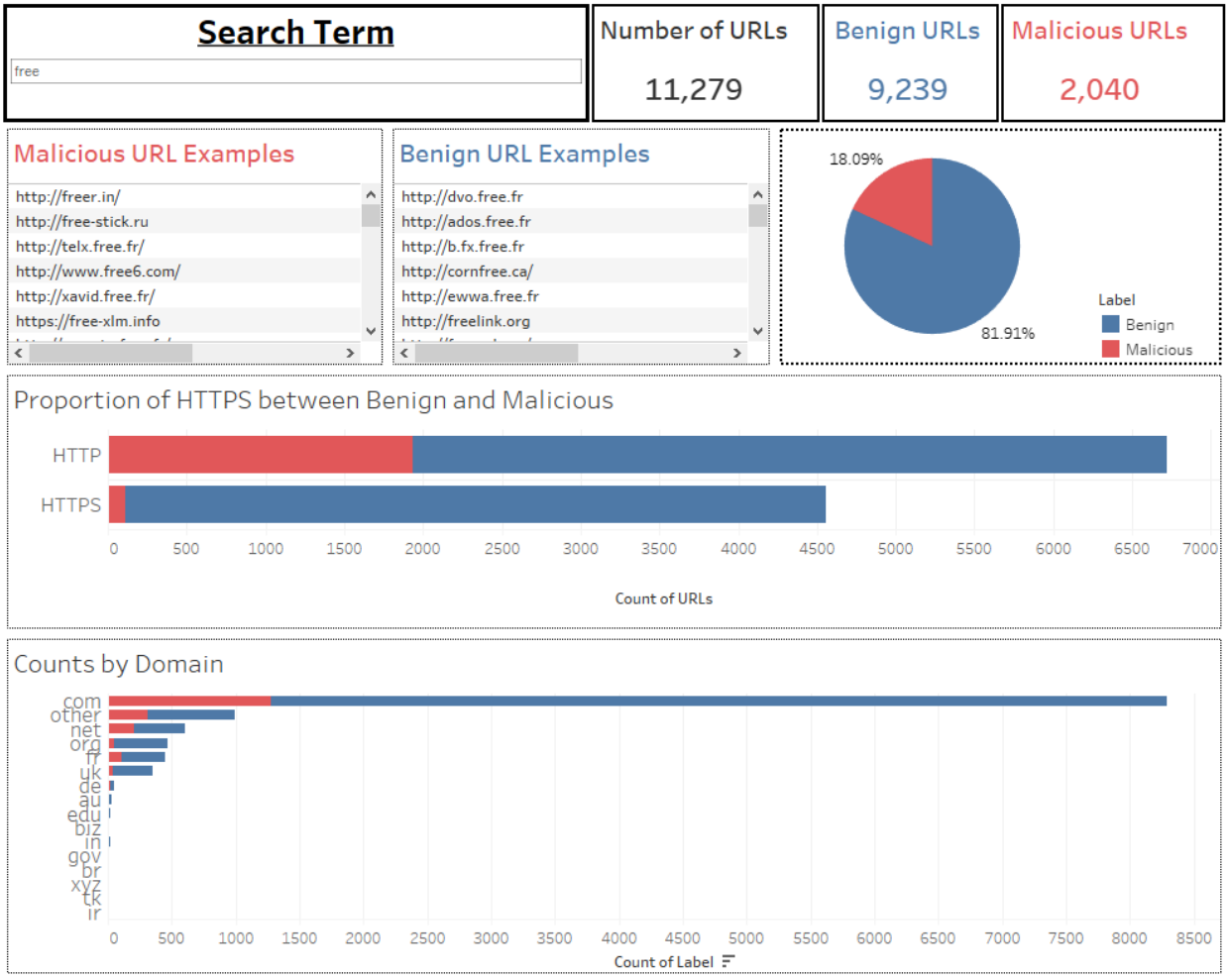
In terms of a presentation, the term “credit” is used to showcase the dashboard’s useability. It is a word that may be associated with malicious URLs, but it isn’t highly concentrated.

Dashboard 2 - Search Term 2



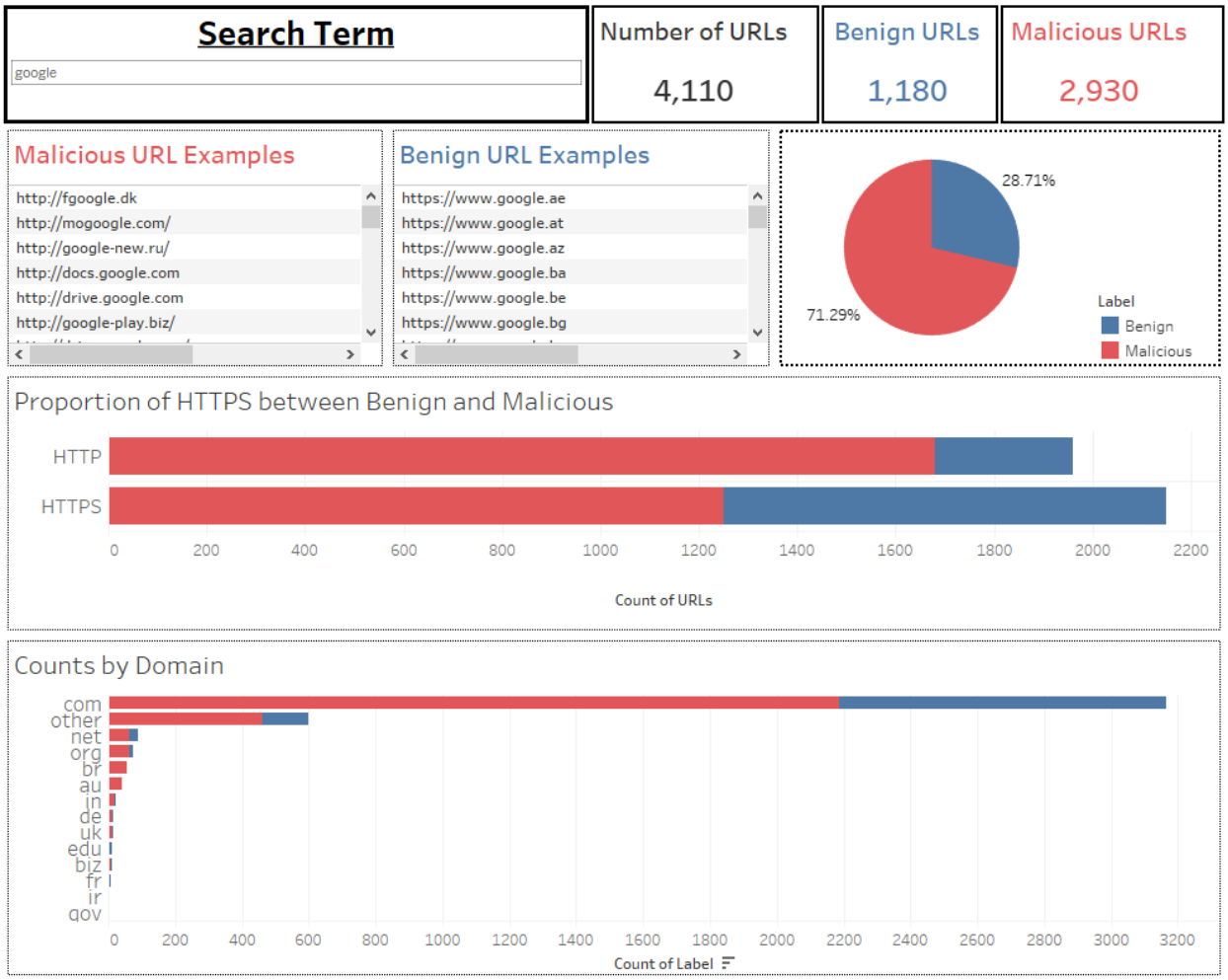
The next term “xxx” is used to display the interactiveness as the dashboard completely changes from mostly blue to mostly red. It is a term discussed during a previous slide, where explicit words were highly correlated with malicious behavior. This is a good demonstration of the dashboard’s capabilities. But, the next words are meant to reveal another trend of malicious behavior.

Dashboard 2 - Search Term 3



To revert the dashboard into a more predominantly blue display, the word “free” is used. This shows that not all suspicious words that you can think of may be highly malicious.

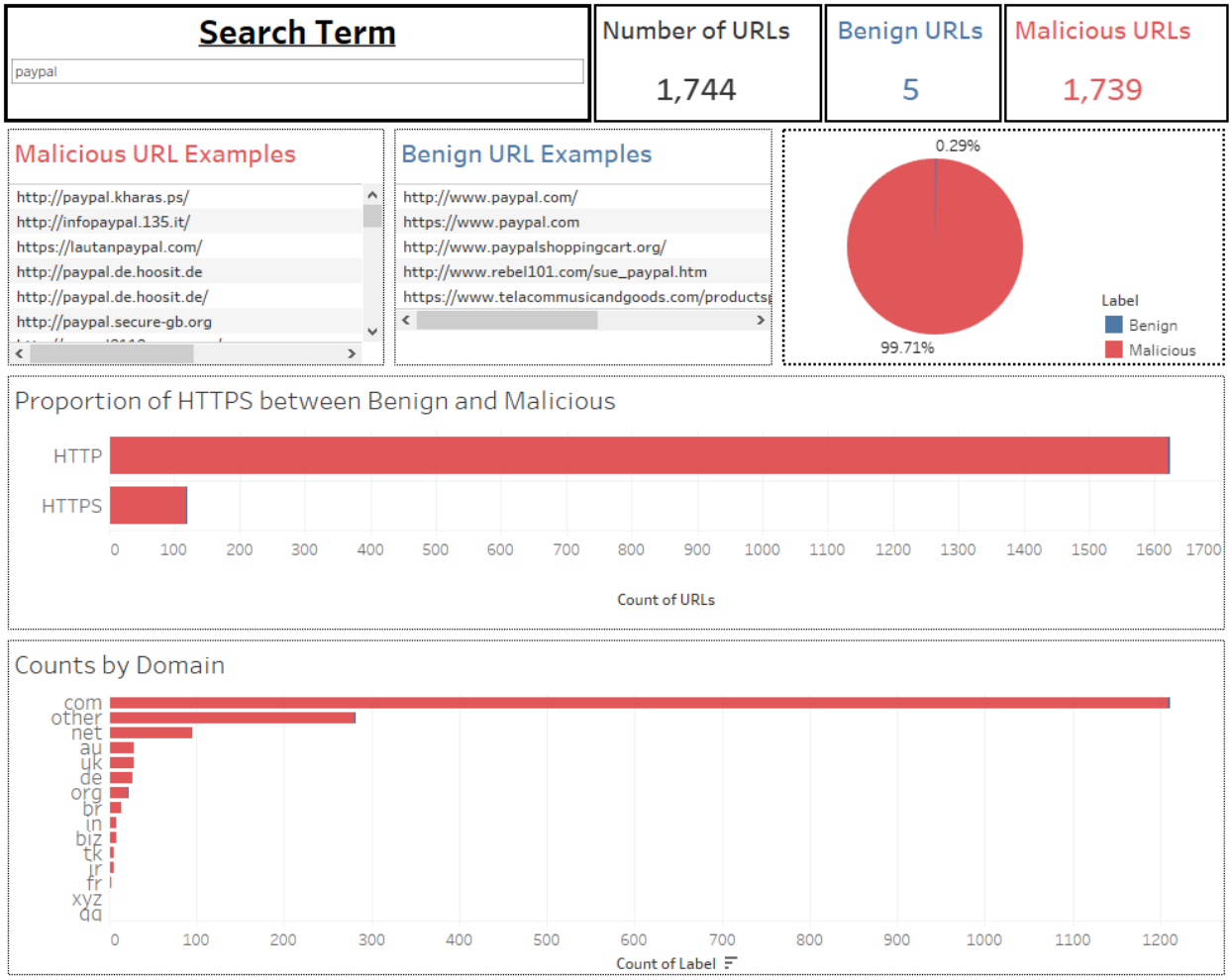
Dashboard 2 - Search Term 3



After setting the dashboard to a mostly blue word, the speaker introduces an unexpected result for searching a term like “google”. This will show the high contrast, and amplify the effects of noticing how much an innocent term like google could be malicious.

The speaker points out under the examples that there is a http version of docs.google.com and drive.google.com that are malicious. This shows a new trend that malicious URLs attempt to trick you into thinking you are on the correct website or are clicking on the correct website.

Dashboard 2 - Search Term 4



As for the final slide, and the slide that will remain in the mind of the audience, the search term “paypal” is used. Paypal, unlike Google, has only a handful of websites that are benign. But, it still has a high amount of mimicry and malicious behavior. Likely, this is due to the financial nature of a website like Paypal which is directly related to financial fraud.

Changing the term from “google” to “paypal” makes the chart even more red than what “google” had already shown. This last search term amplifies the point the previous search term had made. The speaker should repeat the purpose of the presentation at this point that

summarizes all of the trends that were observed as well as this new mimicry trend that has been observed.

Data Sources and Bibliography

AK Singh. (03/04/2020). Dataset of Malicious and Benign Webpages. Retrieved from https://www.kaggle.com/aksingh2411/dataset-of-malicious-and-benign-webpages

Siddharth Kumar. (05/31/2019). Malicious And Benign URLs. Retrieved from https://www.kaggle.com/siddharthkumar25/malicious-and-benign-urls