**Assignment 2: Data Analysis at Home and on the Web**

This lab is to be done in groups of two only.

**Part 1: From Data to Graphs**

**Preamble**

Read Chapter 4, entitled "From Data to Graphs", from the book, "Applied Security Visualization", by Raffael Marty.

The book is available in electronic form through Safari, a service provided through Seneca Libraries. If you prefer a hard copy of the chapter, I will provide one for you. The book is full of great material. Although the code used in the book is s bit out of date, the theory and case studies are excellent, and I highly recommend it.

The chapter describes what we do with our data once we have collected it. It describes how we often need to take data from different sources and combine it to get meaning, and how we are often faced with dealing with data that seems difficult to understand. In the I.T. security industry, it is very important to be able to deal with such situations.

As an example in the chapter, the author looks at his own laptop and does an analysis of secure browsing network traffic on that laptop, looking at web traffic. differentiating TOR and web proxy traffic from web traffic that does not take this secure route. He does so by analyzing metadata.

**Requirements**

I would like you to do a similar analysis on a computer of your own. You may emulate Marty's exercise, or you may build an example of your own. Your example must meet the following criteria:

* you must follow the six steps of the Information Visualization Process as outlined in the book.
* you must define a problem you want to solve, a question you want to answer
* you must choose a problem with a reasonable degree of complexity. It must compare data from at least two different sources (e.g. different log files, different browsers, etc. )

**Examples**

Here is a list of examples of the kinds of things that might be interesting:

* The author's example is very interesting, and you may choose to duplicate it.
* What does your computer do while it is sitting idle?
* What kind of requests are generated when you access your favourite web sites? Cookies? Other domains? Weird requests?
* What kind of requests reach your host in various environments? Is time of day a factor?
* any activity within your computer that can be monitored and compared, like processes vs. CPU usage.

**Deliverable**

Your submission must be a report (pdf) that includes the following:

* abstract: states your question,
* Description of environment
* data dictionary: describes your data,
* code segments: includes your code, with explanation,
* the results of your code,
* Observations: your interpretation of the results.
* Conclusions: What you learned, a description of its usefulness

You must use python and/or R, but may also use any code of your choice (sed, awk, etc). If you use exclusively R, your best choice is to create your file with knitr.

**Part 2: Interacting with APIs**

"Threat intelligence, also known as cyber threat intelligence (CTI), is organized, analyzed and refined information about potential or current attacks that threaten an organization".[(citation)](http://whatis.techtarget.com/definition/threat-intelligence-cyber-threat-intelligence) A threat intelligence platform is one tool we can use proactively to increase awareness about the threats facing our environment. However, we can quickly become overwhelmed by the amopunt of data provided by a threat intelligence tool. Read the Dark Reading article, "[Threat Intelligence Platforms: The Next 'Must-Have' For Harried Security Operations Teams](http://www.darkreading.com/threat-intelligence-platforms-the-next-must-have-for-harried-security-operations-teams/d/d-id/1320671)".

One of our most successful alumni, Roy Firestein, who is head of research at ESentire (as of 2016), has created [Cymon](https://cymon.io/), what he calls "the largest tracker of open-source security reports about phishing, malware, botnets and other malicious activities".

Also please familiarize yourself with "[a curated list of Awesome Threat Intelligence resources](https://github.com/hslatman/awesome-threat-intelligence)"

Read through [Cymon's API documentation](http://docs.cymon.io/). Write a report which:

* describes the data provided by Cymon through its API - both its content and its format (you may choose another resource from the curated list, but you **MUST** write code that interacts through an API)
* describes how one can use the API to extract useful data.
* uses R **and** Python to perform some exploratory data analysis on data provided through the Cymon API. Come up with some interesting questions and try to answer them.

**Deliverables:**

**Reports:**

Create one report for each part of this assignment. Embed interesting charts as graphics in your reports. Make sure you include an explanation with each chart.

Upload the reports to this site at Blackboard. If the combined size of your reports are larger than the upload limit, let me know and I will create a second upload link.

**Code:**

I would like you to develop your code as a group using GitHub. Store your code on your private account (which you should have by now) on GitHub. In your reports, for each chart Include a link to the code, along with information about how it can be accessed.