RSAConference2015

Singapore | 22-24 July | Marina Bay Sands

CHANGE

Challenge today's security thinking

SESSION ID: TTA-F04

High dimensional visualization of malware families

Matt Wolff

Chief Data Scientist

Cylance, Inc.
@cylanceinc

Glenn Chisholm

Chief Technology Officer

Cylance, Inc.
@cylanceinc

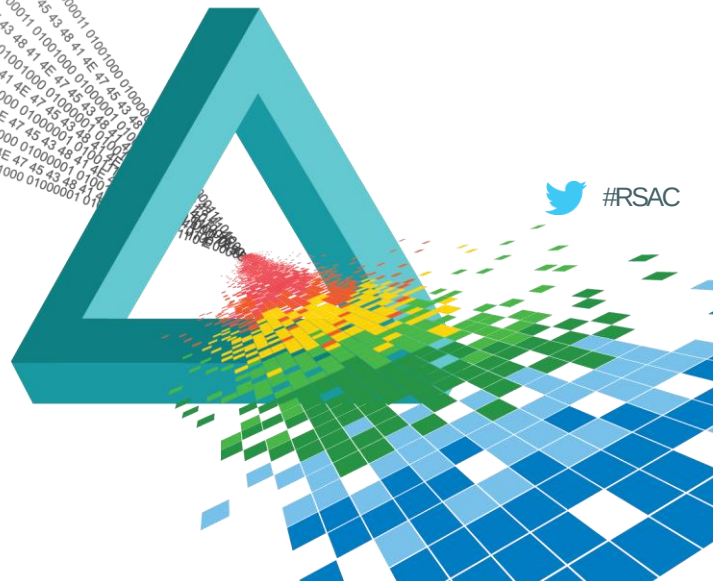
Visualization of Malware Families

- Having an understanding of the relationships between individual pieces of malware provides insight to the analyst.
- This can be acomplished through detailed reverse engineering or reliance on industry naming and identification.
- We seek to utilize automated feature extraction to develop a feature set and capability to visualize relationships between sample sets of malware.
- Ultimately we represent a sub set of malware using Chernoff Faces and explore the relationships

RSAConference2015

Singapore | 22-24 July | Marina Bay Sands

Naming and Identification of Malware



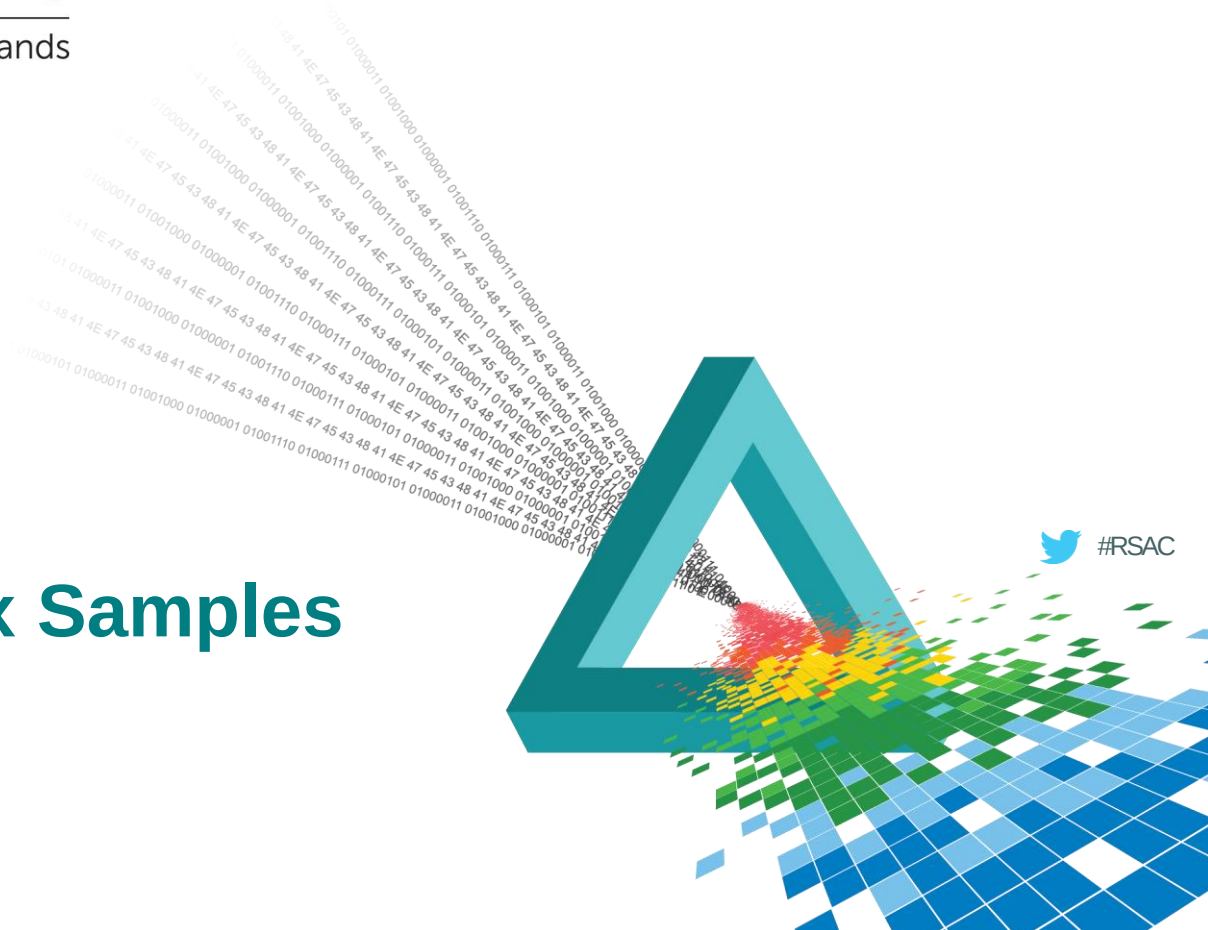
#RSAC

Malware Families

- Malware is commonly named to allow easy identification of function and purpose (Also for marketing value).
- Traditional Anti Virus companies have used one of:
  - Generic designations such as Trojan, Password Stealer or the ever helpful Generic.
  - Specific familial names to identify a specific branch of malware, this is common with known attack kits such as Zeus.
  - An attack specific name where a group of tools used in an attack are named together despite not having any relationship or in some cases being malicious.

Malware Families

- Understanding the reliability of these names is important.
- An incorrectly named sample may lead to poor decisions by a security team.
  - Not every organisation has access to reverse engineers
  - If you don't truly know its purpose, responding is impossible
- Can we trust the names we get for a set of samples?
- Can we begin to see subsets of samples?



The 65,000 Dridex Samples

Dridex

- We were presented with this set by some financial sector researchers who called these out as a set of Dridex samples.
- Dridex is a set of malware samples that targets end users to steal financial information.
- Once resident it monitors for specific financial institutions identified through a configuration file.
- Interesting in its wide distribution and well structured campaigns.

Null :→

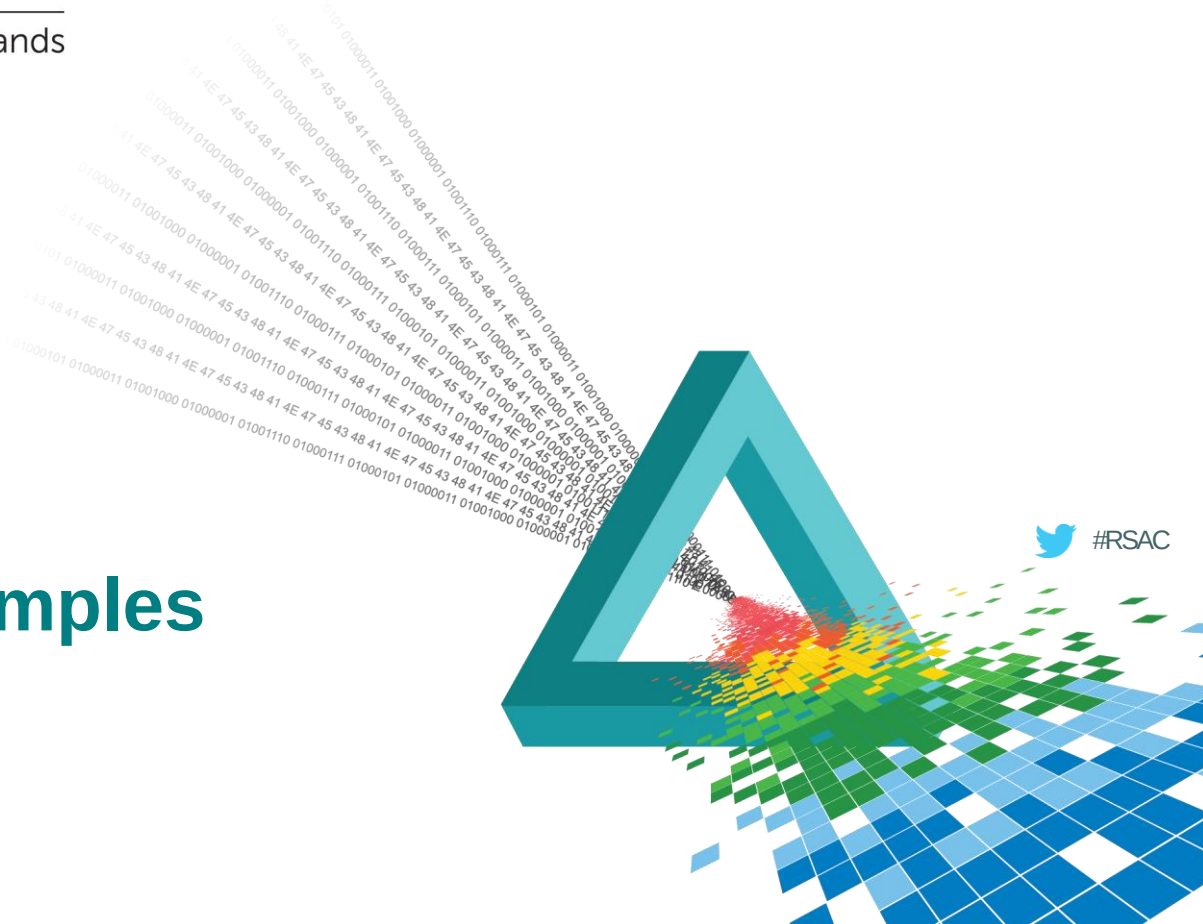
What does the AV industry think?

Literally hundreds of names for a group of samples, can we start to understand which samples are closely related?

Dridex

- Are any of these samples infact Dridex?
- If so what are the names for the actual Dridex samples?
- The vast majority of these names are of no actual value, yet people rely upon them.
  - Some imply these are ransomware
  - Some are banking trojans (Maybe Dridex?)
- Can we begin to find populations of samples?

Clustering the Samples

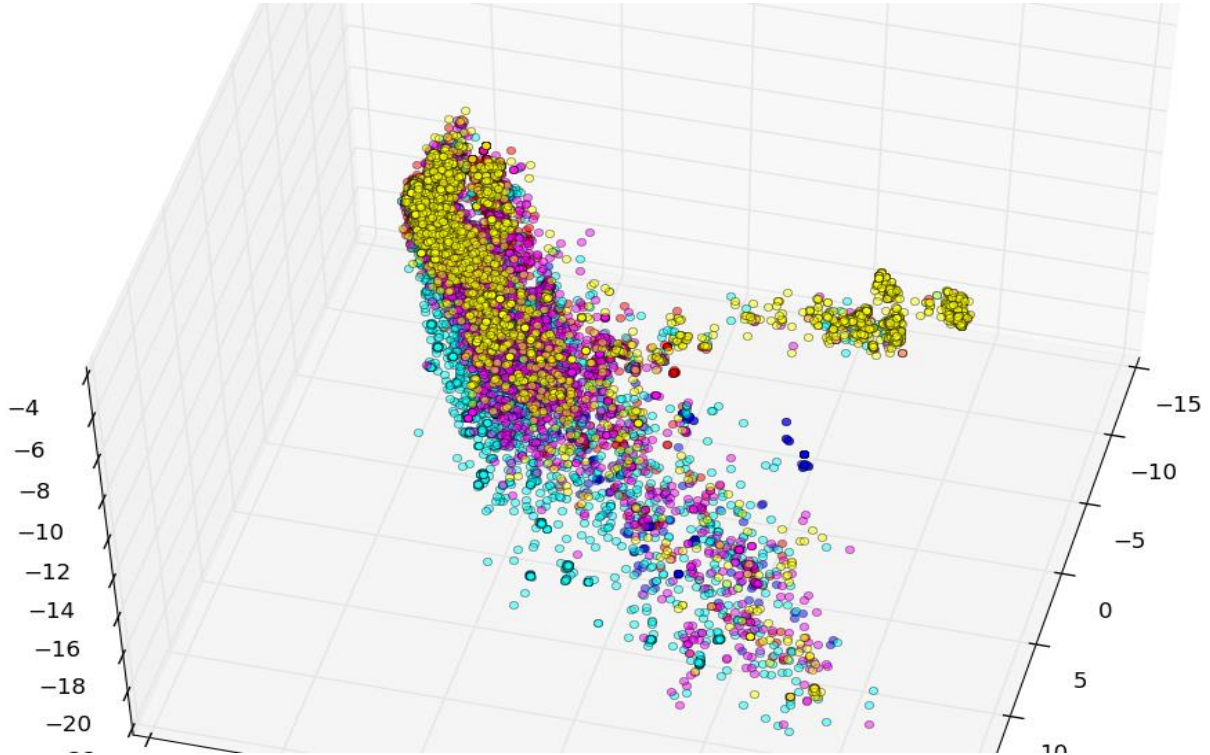


Fetaure Extraction

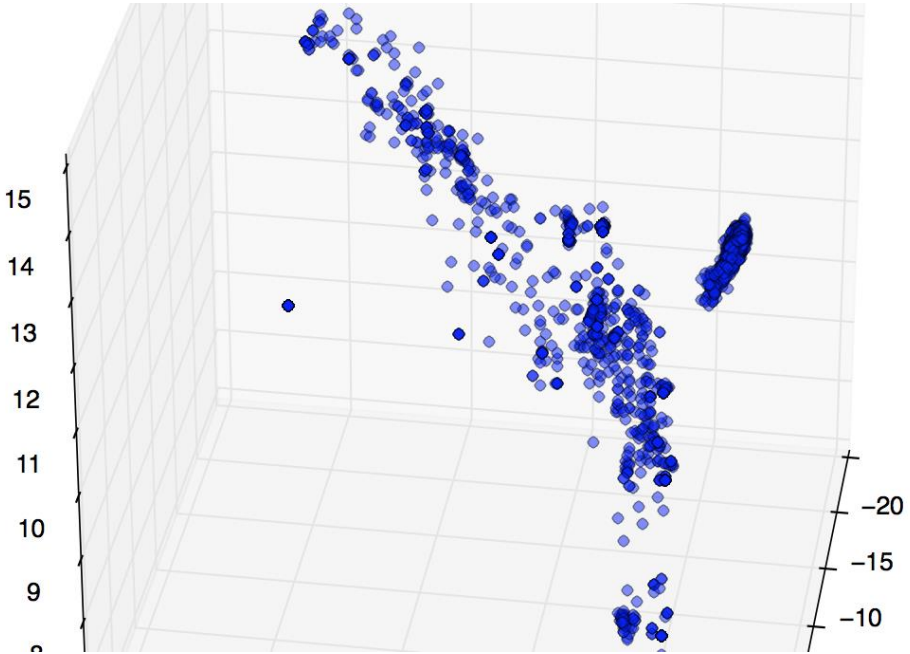
- Extracting a large number of features from the samples:
  - Structural
    - Headers
    - Sections
  - Content:
    - Strings
    - Resources
    - Code
- Can we establish relationships between samples?

3 Dimensional visualization

- With the extracted features, we then use techniques to reduce the dimensionality down to 3
  - SVD
  - T-SNE
- What does the data look like in 3 dimensions?



General malware. Reduced using SVD.



Clustering in 3 Dimensions

“Dridex” samples. Reduced using SVD.

Notes

- Visually, it is difficult to differentiate between different malware groups in 3 dimensions.
- Even the Dridex samples do not appear as a single cluster, but as multiple points in space.
- WordCloud method on Dridex clusters
  - Group Dridex into small clusters of close samples
  - Visualize the different names applied to samples in the cluster

Dridex Clustering Analysis

- Applying clustering techniques we extract 236 separate clusters from the initial set of 64,971 samples.
- The majority of clusters are smaller than 10 samples.
- 1661 of the samples did not fall into any cluster.
- However one cluster has 45,426 of the samples within it.
- Starting with this single cluster our goal will be to randomly select out samples for detailed analysis and determine the efficacy of the technique.

AV Names for Cluster 1

Picking the biggest cluster what does AV think of these samples

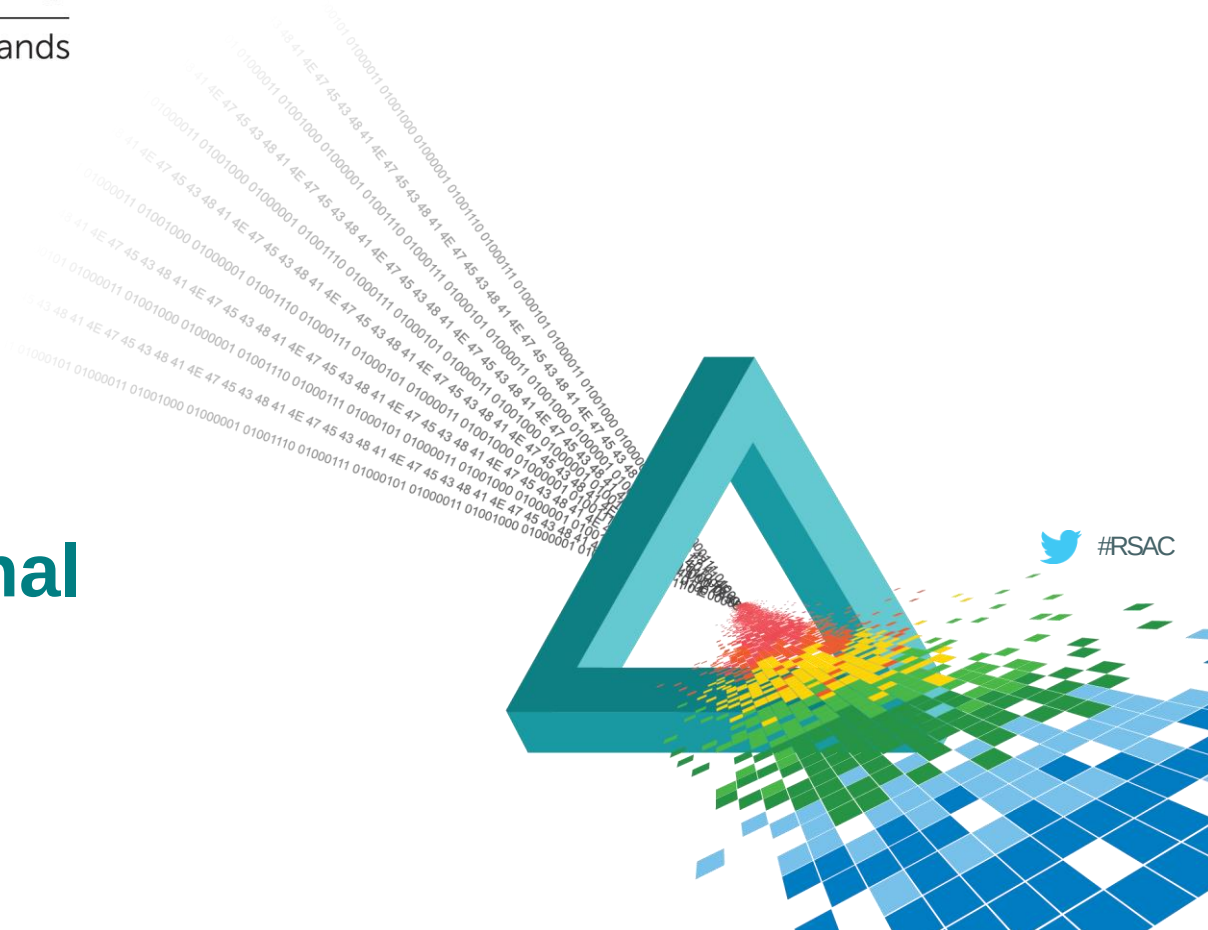
AV Names for Cluster 80

These samples show a strong trend to being ransomware. However some are called CRIDEX or Infostealer

RSAConference2015

Singapore | 22-24 July | Marina Bay Sands

Higher Dimensional Visualization

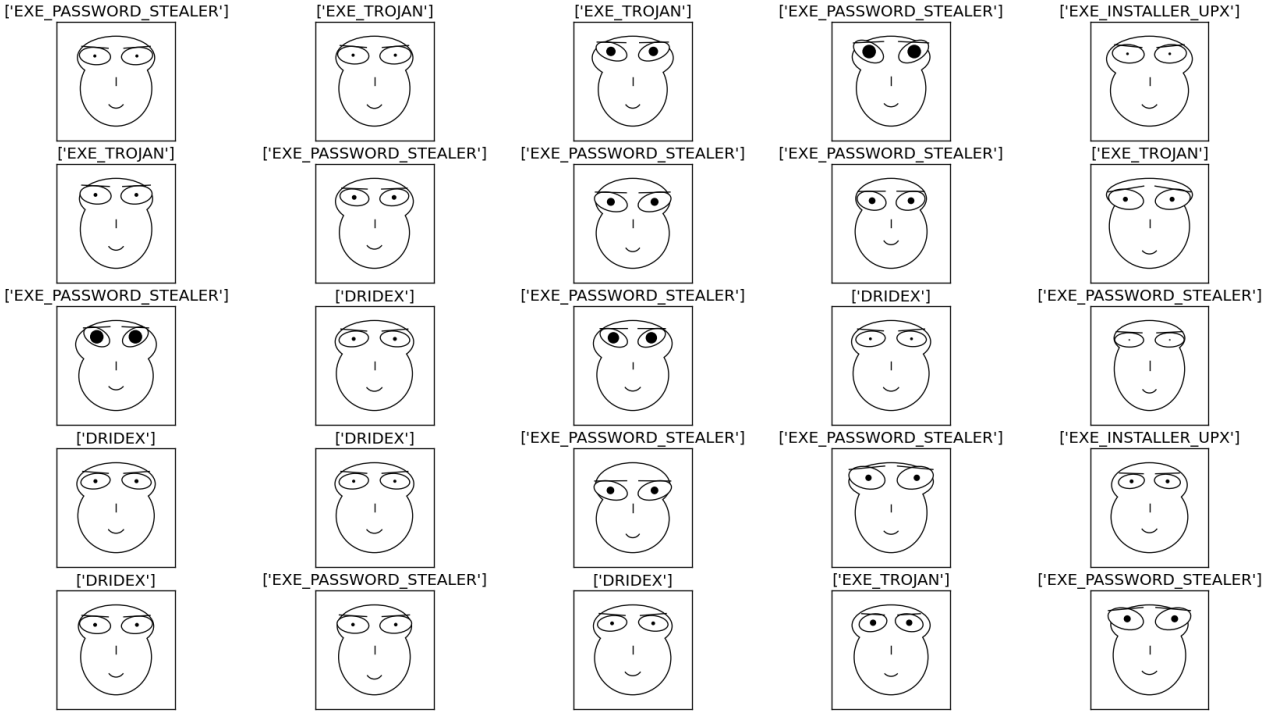


#RSAC

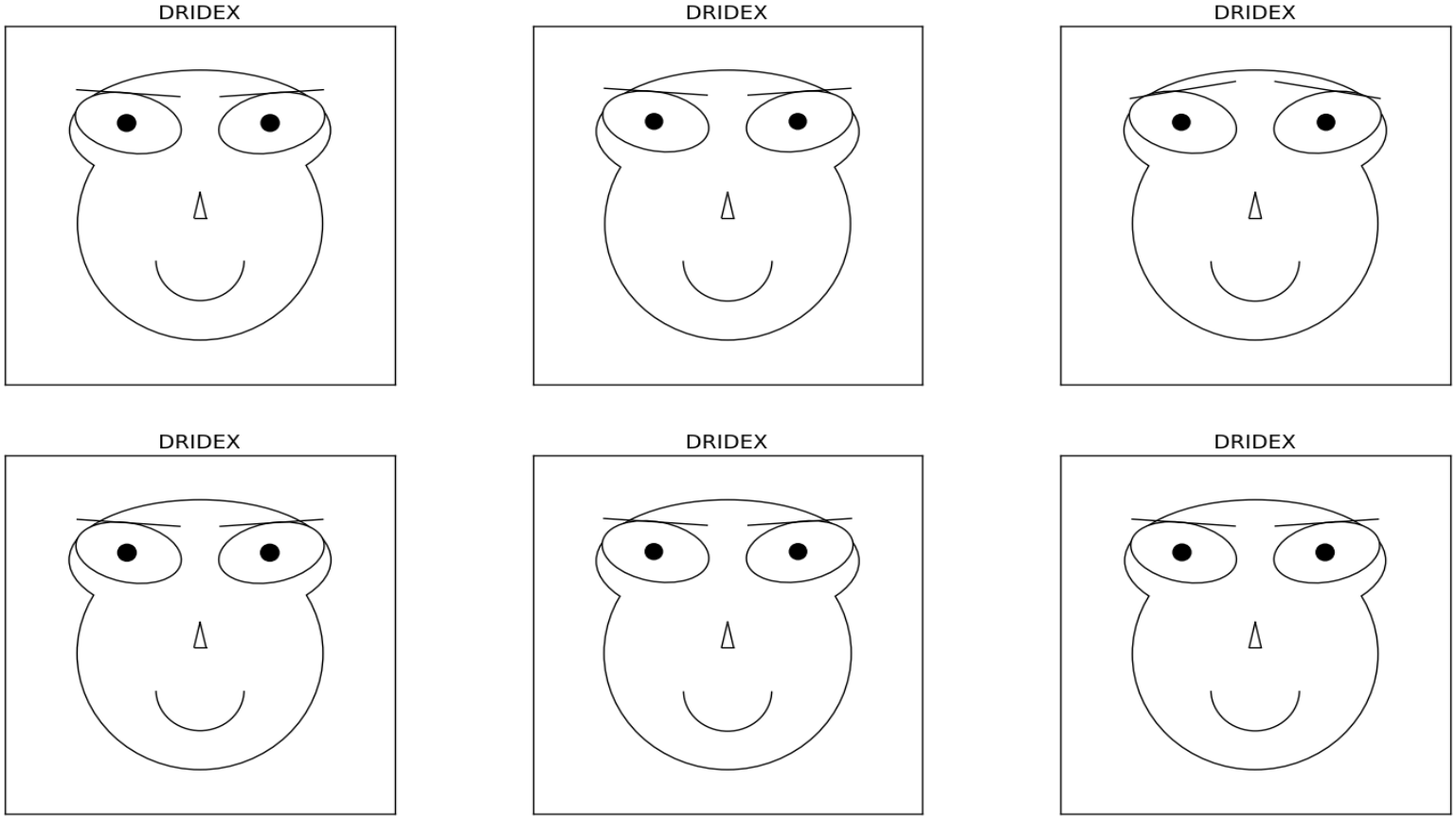
Chernoff Face Visualization

- Human faces provide a variety of characteristics that can be easily interpreted.
- The shape, size, placement and orientation of face features such as eyes, ears, mouth, etc. represent values of variables
- A visualization technique, known as Chernoff Faces, leverages this data to generate high dimensional visualizations of data
- First, we take a look at the similarity of various types of malware samples using 17 dimensions

Chernoff Face Visualization: General Malware



Chernoff Face Visualization: DRIDEX

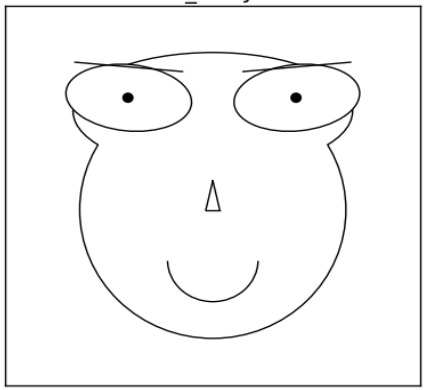


Chernoff Face Visualization: DRIDEX

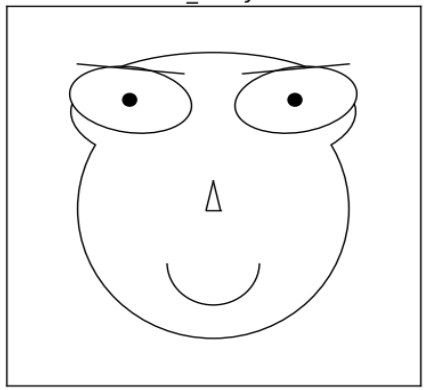
- At a higher dimension, many of the DRIDEX clusters look similar.
- This is a strong indication that many of these samples are in fact related, and the DRIDEX name is applied to samples that are similar

Chernoff Face Visualization: TROJAN

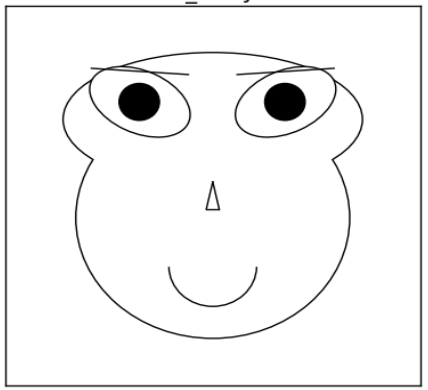
EXE_TROJAN



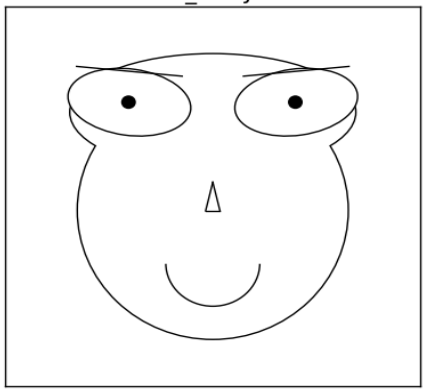
EXE_TROJAN



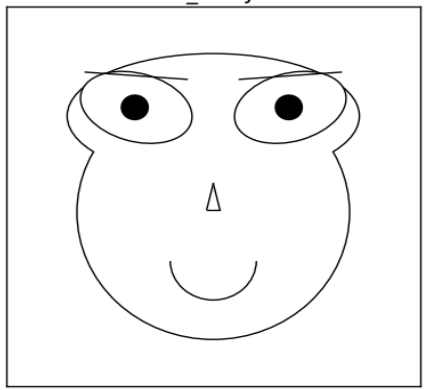
EXE_TROJAN



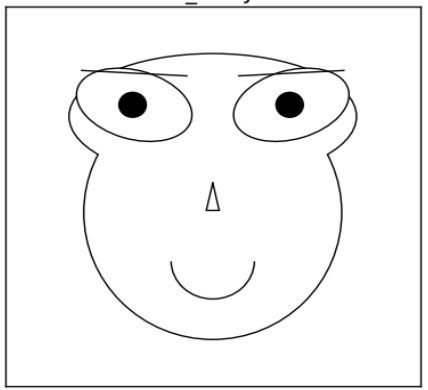
EXE_TROJAN



EXE_TROJAN



EXE_TROJAN

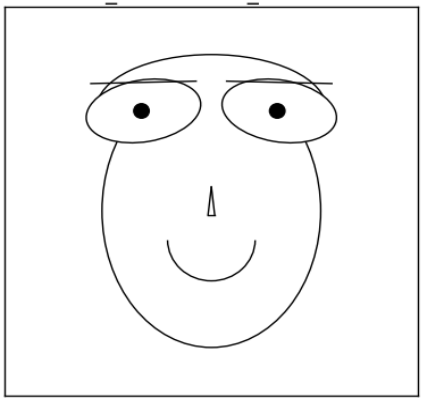


Chernoff Face Visualization: TROJAN

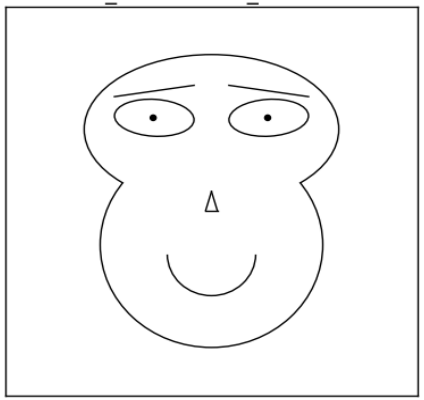
- With the more generic name Trojan, we are starting to see some variety in the dataset
- With these clusters, significant differences in the eyes indiciate that the variance between trojan clusters is more noticable than with DRIDEX

Chernoff Face Visualization: STEALER

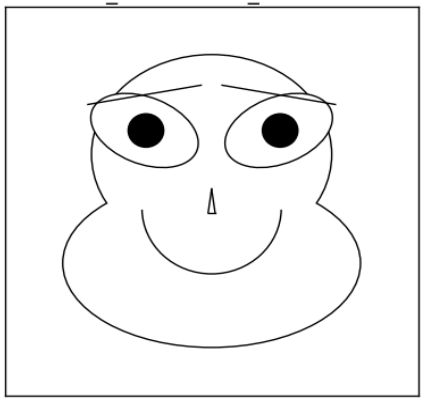
EXE_PASSWORD_STEALER



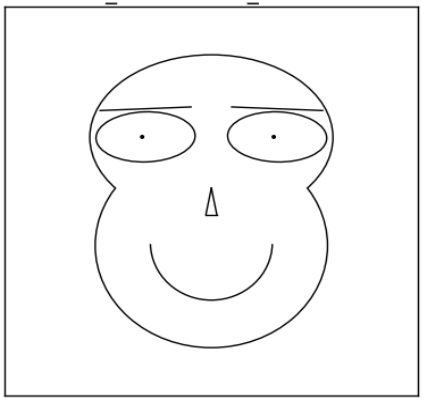
EXE_PASSWORD_STEALER



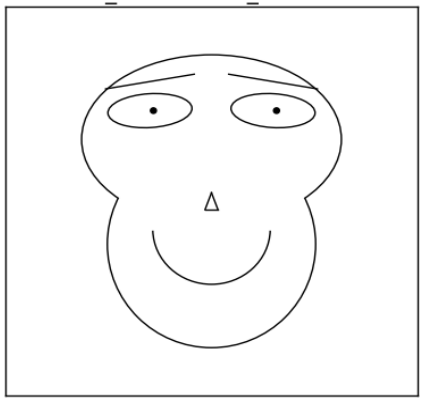
EXE_PASSWORD_STEALER



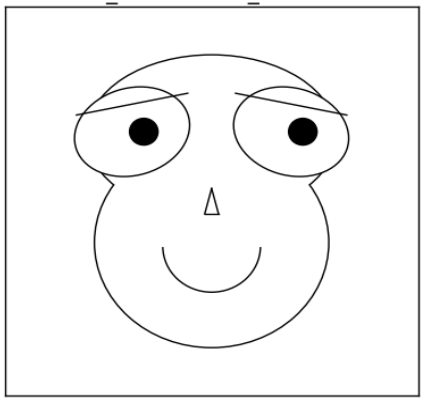
EXE_PASSWORD_STEALER



EXE_PASSWORD_STEALER



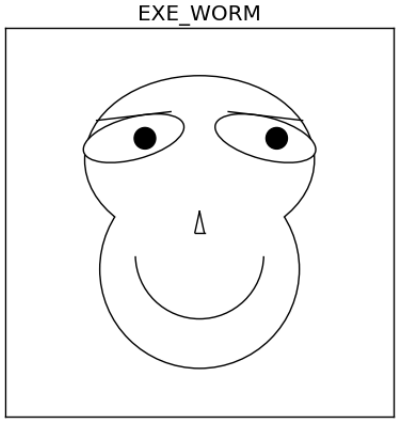
EXE_PASSWORD_STEALER

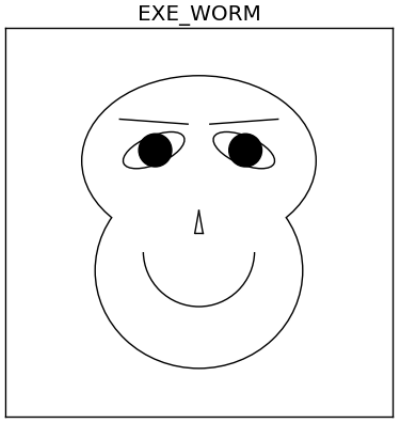


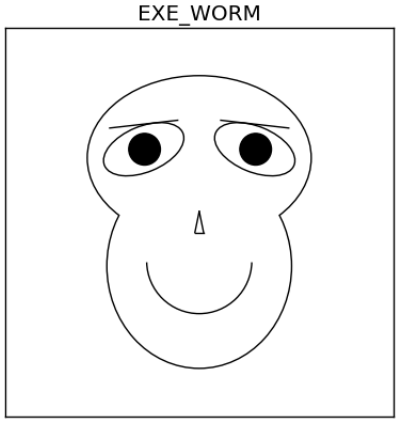
Chernoff Face Visualization: STEALER

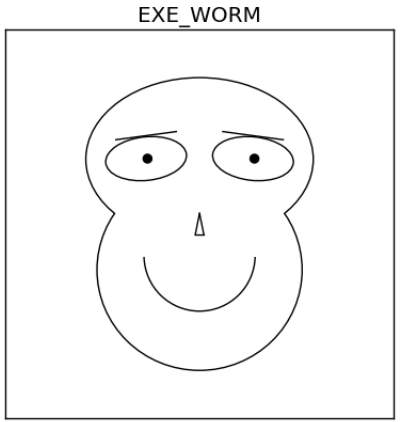
- Another “generic” name, STEALER
- Here we see a large variance in the different clusters. It would be reasonable to state at this point that many of these clusters are in fact not related at all in terms of their features.
- While the malware in this group may steal information, the methods used and functionality contained in the samples varies greatly.

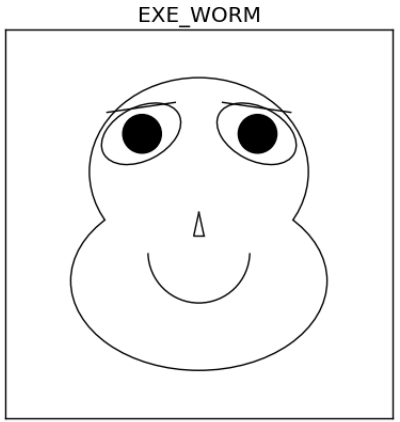
Chernoff Face Visualization: WORM

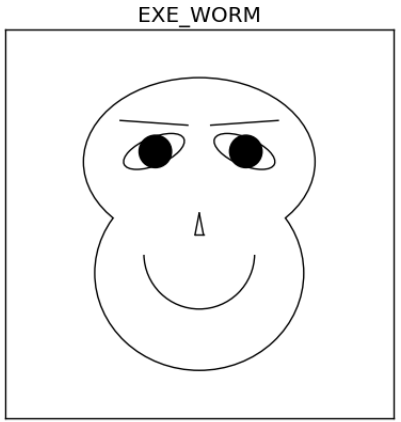










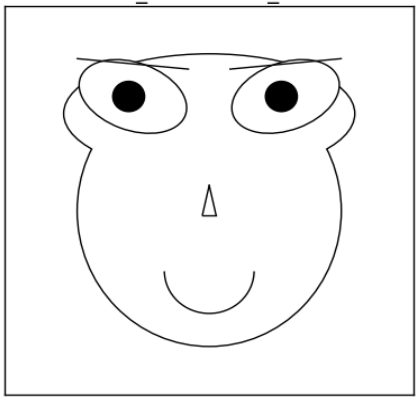


Chernoff Face Visualization: WORM

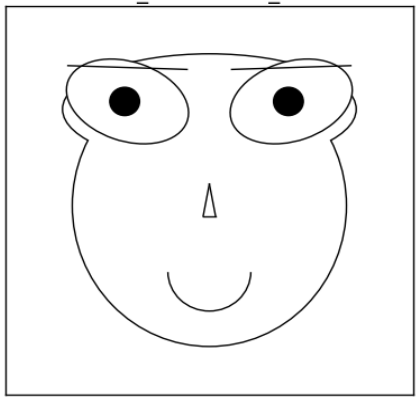
- Similar to what we saw with STEALER, large variance in what is considered WORM as well

Chernoff Face Visualization: INSTALLER

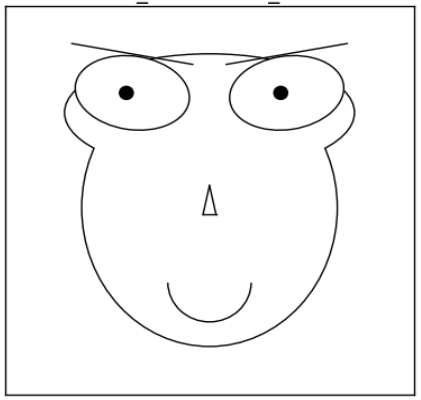
EXE_INSTALLER_UPX



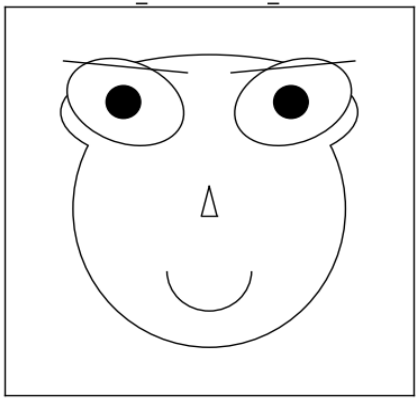
EXE_INSTALLER_UPX



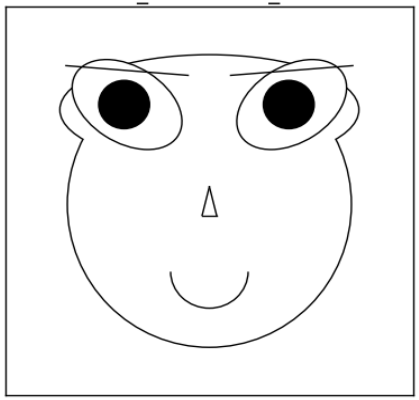
EXE_INSTALLER_UPX



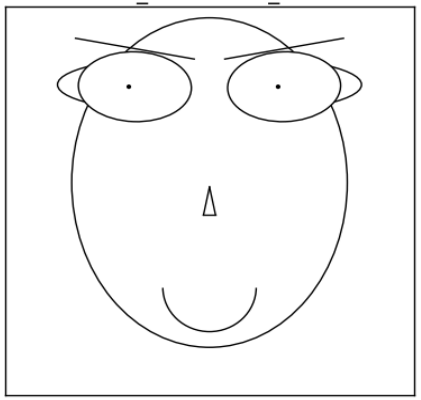
EXE_INSTALLER_UPX



EXE_INSTALLER_UPX



EXE_INSTALLER_UPX



Chernoff Face Visualization: INSTALLER

- Collection of UPX packed malware samples that work as droppers
- While there is still a good amount of variance in the clusters, not as significant as we saw with STEALER and WORK
- The reduction in variance compared to previous groups can be attributed to the use of packing

Analysis

- Higher dimensional visualization helped to identify similarities and variance in different malware naming groups
- Sample like Dridex, and probably other well known malware campaigns, often have high similarity between samples
- Other malware names, such as worm or password stealer often have weak similarity between samples

Apply Slide

- Higher dimensional visualization can be an effective tool to when looking for similarities between or in groups
- Is some cases, malware names may have minimal correlation between samples of the same name
- A single sample of malware can have a wide range of different names, making it difficult to identify what the sample is

Additional info

- Questions?
- Contact us: datascience@cylance.com
- Special thanks to Xuan Zhou for her contributions