Using Machine Learning to Find Hackers and Malware

By Sam Triolo

What are we looking for and where are they?

* Who are we looking for?
  * The Bad Guys
    * Hackers
    * Malware
* How do we find them?
  * Unsupervised K-means
    * Outliers
    * Unusual group membership
  * Supervised K-means to tune and alert going forward

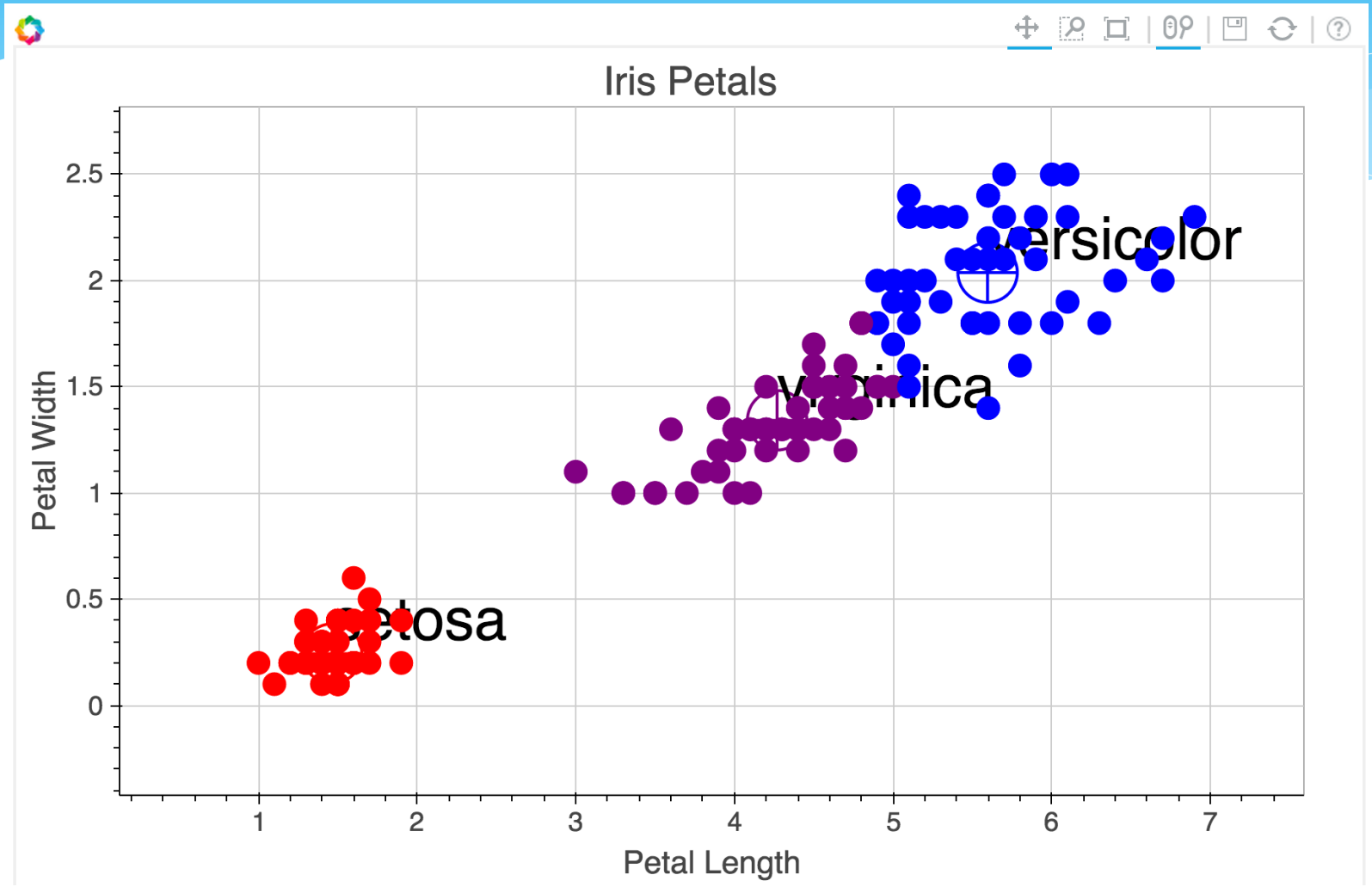
Tools Used / Data Collected

* Vbscript
  * Microsoft Active Directory Login Successful (4624) and Login Denied (4625) event data
* Python
  * ETL above data
* MongoDB
* Scikit-learn - machine learning – K-means
* Ipython – data visualization
  * Bokeh – remote, interactive visualization

Features / Analysis

* Six features were used for analysis
  * unique destination logins (i.e. a set)
  * total logins (any login to any host of any kind)
  * w2s (workstation to server) logins, s2w logins, w2w logins, and s2s logins (determined by IP and hostname conventions)

Data Visualization



Conclusion

* By grouping users based on usage patterns, users of a certain type (e.g. a non-technical user) who's usage patterns more closely grouped them with a different type (e.g. a system administrator) would recommend further investigation
* Users who were outliers within their own group (i.e. unusual behavior within that group) would also recommend further investigation

References

* Simple k-means example:
http://mnemstudio.org/clustering-k-means-example-1.htm

* Windows event reference:
https://www.ultimatewindowssecurity.com/securitylog/encyclopedia/default.aspx?i=j

* Statistics courses:
https://onlinecourses.science.psu.edu/statprogram/programs