**To do :**

* Trim the data occurs
* Pick important features (research)
* Design questions around features
* Test and write solutions
* Convert info\_traffic\_peer\_particpiants to number values

**Introduction:**

This lab will cover different tools and methods regarding machine learning to classify automated systems.

Use the link below to download the database: (CSV)  
<http://data.caida.org/datasets/peeringdb-v1/>

Highlighted portion is optional.

In order to use the next section, here are some tools that might be helpful:

* Table plus (recommended) trial version (link)
* SQL developer
* Oracle database manager

This database is in SQL, here is an example code to extract

SELECT TableA.\*, TableB.\*, TableC.\*, TableD.\*

FROM TableA

JOIN TableB

ON TableB.aID = TableA.aID

JOIN TableC

ON TableC.cID = TableB.cID

JOIN TableD

ON TableD.dID = TableA.dID

Research SQL query and obtain methods to use in future questions.

Question 1: Use SQL query to obtain the 5 tuples in the 2nd table, what is the information within the 5 tuples?

Question 2: use SQL query to obtain the rest of the tables within the database, so you can get an understanding of the data.

Question 3: convert the SQL dataset into a CSV, concatenate and import the data into python for further analysis.

Question (): how many unique AS are in the data?

Question (): write a python program that takes an input AS number and outputs the peer participant, name of the management facility, the city, state and zipcode.

Question 4, 5,6: draw different diagrams to show the layout of the data, use entropy?

Question (): draw a table of the count of all the labels

In this lab we will focus on these 3 lables: …

Question (): plot a bar graph of the number of IP prefixes vs type of peer participants. For the 3 labels

Question (), drop column info\_traffic\_peer\_particpiants and info\_prefixes\_peer\_particpiants, what are your results now? Explain why.

Questions 7,8,9: use different classification methods to solve for accuracy precision and recall: SVM, GNB, Random forest, NN.

Question 10: From the different classifiers which one was the highest in terms of accuracy? Take one of the AS from that classifier and input it into the local directory? Was it correct?

Additional questions: outliers? Test outliers in local directories? Results, why?

**Challenges:**

* Data was in SQL, sqlite
* Multiple tables
* No labels
* Picking features
* Table plus not export correct columns (need to do it in python or find and export from legit source)