Data Analysis, Machine Learning, Bro and You!

Together again like never before...
Presenter

Brian Wylie
Working at Kitware Inc.
Background in Information Security and Vis
Likes open source and mixed Corgis
What’s the point of this talk?

Provide software classes and examples that make the *path* from Bro Network data to the popular data analysis and machine learning libraries *easy*.

When you say *easy*, what do you mean?

```
# Create a Pandas dataframe from a Bro log
bro_df = LogToDataFrame('/path/to/dns.log')
```

*One line of code:*

**Bro Log → Pandas DataFrame**

Pandas DataFrame with all the right types and timestamp as index
What’s the intended audience?

- People who like **Python**
- Interested in **Pandas**, **scikit-learn**, **Spark**, **Parquet**
- **Hate** seeing examples on **Iris** data or **TF-IDF**
- **Frustrated** when trying to use your own data
- Want **easy** examples using **Bro!**
Are you going to show super scalable blah?

- Presentation will talk about **Pandas**, **Scikit-Learn**
- We also have classes/notebooks on:
  - Kafka
  - Parquet
  - Spark
- We’ll show a some of this stuff…

Please see tomorrow’s great Talk 😊

3:30 p.m. **Spark and Bro: When Bro-Cut Won’t Cut It**
Eric Dull, Joseph Mosby, & Brian Sacash; Deloitte & Touche
Talk Outline

- **Big Picture**
- **Software Bridges**
  - Bro to Python
  - Bro to Pandas
  - Bro to Scikit-Learn

- **Example: Anomaly Detection**
  - Bro DNS and HTTP logs
  - Categorical and Numeric Data
  - Clustering
  - Isolation Forests

What is the best way to do data science on Bro Network data?

I’m not sure… Ahhh!!!
Security Data → Data Analysis and Machine Learning

Data flow diagram of how Pandas and Scikit-Learn are used.

- **DataFrame** = Pandas
- **Numpy array** = Scikit-Learn

- **JSON**
- **Agents**
- **Packets**
- **Logs**
- **Bro IDS**

**DataFrame**

- **Stats**
- **Filtering**
- **Grouping**
- **Vis/Plots**

**numpy array**

- **Clustering**
- **Anomaly**
- **Stats**
- **ML**
Talk Outline

● Big Picture
● **Software Bridges (BAT)**
  ○ Bro to Python
  ○ Bro to Pandas
  ○ Bro to Scikit-Learn

● Example: Anomaly Detection
  ○ Bro DNS and HTTP logs
  ○ Categorical and Numeric Data
  ○ Clustering
  ○ Isolation Forests

You guys haven't seen my rabbit have you?
Bro Analysis Tools

$ pip install bat

What is BAT?

A simple to use Python Module that makes getting Bro data into popular data analysis and ML package super easy!

https://github.com/Kitware/bat

Who’s Kitware?

- ~130 people, offices around the world
- Developing and supporting open source software for 25 years
- New information security program
- Summer Internships available 😊
Talk Outline

● Big Picture
● Software Bridges
  ○ Bro to Python
  ○ Bro to Pandas
  ○ Bro to Scikit-Learn

● Example: Anomaly Detection
  ○ Bro DNS and HTTP logs
  ○ Categorical and Numeric Data
  ○ Clustering
  ○ Isolation Forests

You guys haven't seen my rabbit have you?
Hello World

Step 1: $ pip install bat
Step 2: Write a few lines of code
Step 3: There is no step 3...

Output: Streaming (generator) of Python dictionaries with the proper type conversions.

```python
from pprint import pprint
from bat import bro_log_reader

# Run the bro reader on a given log file
reader = bro_log_reader.BroLogReader('dhcp.log')
for row in reader.readrows():
    pprint(row)

<<< Output >>>
{'assigned_ip': '192.168.84.10',
'id.orig_h': '192.168.84.10',
'id.orig_p': 68,
'id.resp_h': '192.168.84.1',
'id.resp_p': 67,
'lease_time': datetime.timedelta(49710, 23000),
'mac': '00:20:18:eb:ca:54',
'trans_id': 495764278,
'ts': datetime.datetime(2012, 7, 20, 3, 14, 12, 219654),
'uid': 'CJsdG95nCNF1RXuN5'}
```
Talk Outline

● Big Picture
● Software Bridges
  ○ Bro to Python
  ○ Bro to Pandas
  ○ Pandas to Scikit-Learn
● Example: Anomaly Detection
  ○ Bro DNS and HTTP logs
  ○ Categorical and Numeric Data
  ○ Clustering
  ○ Isolation Forests
Pandas DataFrames

“Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with relational or labeled data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.”

Demo: Bro To Pandas
Talk Outline

- **Big Picture**
- **Software Bridges**
  - Bro to Python
  - Python to Pandas
  - Pandas to Scikit-Learn

- **Example: Anomaly Detection**
  - Bro DNS and HTTP logs
  - Categorical and Numeric Data
  - Clustering
  - Isolation Forests
Scikit-Learn

“Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.”

- We create numpy ndarrays with proper handling of both categorical and numeric types. Our DataFrameToMatrix class supports fit, fit_transform, and transform methods.
- Internal maps for categorical ‘one-hot’ encoding and numerical normalization means that serialization and train/evaluate use cases are supported.

Demo: Bro To Scikit
Talk Outline

● Big Picture
● Software Bridges
  ○ Bro to Python
  ○ Python to Pandas
  ○ Pandas to Scikit-Learn

● Example: Anomaly Detection
  ○ Bro DNS and HTTP logs
  ○ Categorical and Numeric Data
  ○ Clustering
  ○ Isolation Forests

One fish is red. You don’t need machine learning for that!
Anomaly Detection

Popular Mental Images

Popular Misconception: It’s going to show me ‘bad’ stuff
Anomaly Detection

Just gets you to base camp...

~.01%: Possibly Malicious (Recommender System)

~1%: Interesting traffic (Organization + User Feedback)

~5%: Anomalous traffic (Anomaly Detection)

~95%: Normal network traffic that can be filtered out early in the pipeline

100%: All Traffic (unknown mix)
Normal to Anomalous

Anomaly Detection

Challenges:
- Streaming Data
- Data Volume
- Categorical and Numerical Types
- Efficient DataFrame/Matrix conversions

Output:
- 1-5% of data
- Uncommon (by def)
- Good Base Camp

Example: 1M HTTP Logs to 10k anomalous rows *

* http://github.com/Kitware/bat/blob/master/notebooks/Anomaly_Detection.ipynb
Isolation Forests: Anomaly Detection

9 Divisions (not anomalous)

4 Divisions (anomalous)

https://github.com/Kitware/bat/blob/master/notebooks/Anomaly_Detection.ipynb
Anomalous to Interesting

Organization + User Feedback

Challenges:
- Streaming Data
- Organization and Clustering
- Engaging the Human
- User Interface and Feedback*

Output:
- Fraction of 1%-5%
- Clustered/organized
- Ready for Feedback*

Example: 10k rows clustered and organized for displayed to user *

* Feedback will be used in the next phase of the pipeline
* http://github.com/Kitware/bat/blob/master/notebooks/Anomaly_Detection.ipynb
Demo: Anomaly Detection

https://github.com/Kitware/bat/blob/master/notebooks/Bro_to_Scikit.ipynb
https://github.com/Kitware/bat/blob/master/notebooks/Anomaly_Detection.ipynb
Demo: Bro to Kafka to Spark

https://github.com/Kitware/bat/blob/master/notebooks/Bro_to_Kafka_to_Spark.ipynb
Demo: Bro to Parquet to Spark

https://github.com/Kitware/bat/blob/master/notebooks/Bro_to_Parquet_to_Spark.ipynb
Questions/Comments?