

Data Science for Cybersecurity Hand-on Lab

September 25, 2024

Copyright 2024 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific entity, product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute nor of Carnegie Mellon University - Software Engineering Institute by any such named or represented entity.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

CERT® and Carnegie Mellon® are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM24-1042



Clarence Worrell, PhD, PE
Senior Data Scientist
Software Engineering Institute
Carnegie Mellon University



Thomas Scanlon, PhD
Technical Manager, CERT Data Science
Software Engineering Institute
Carnegie Mellon University

Data Science Tools

Data Science Tools

- **No-code and low-code**

- Excel (spreadsheets and macros)
- Orange (drag-and-drop machine learning)

- **Code-based tools**

- Python (general language, many data science libraries)
- R (statistics)
- MATLAB (engineering)

- **Many other options**

The BETH Dataset

BETH Dataset

BETH* is a real cybersecurity dataset published in 2021 as a benchmark for anomaly detection researchers

- 8 million records, generated by 23 hosts, during 5 discontinuous hours
- Each host includes benign traffic as well as at most one single attack
- Each record is labeled as to whether it is “benign” or “malicious”

*Highnam, K., Arulkumaran, K., Hanif, Z., & Jennings, N. R. (2021). "BETH dataset: Real Cybersecurity Data for Anomaly Detection Research." ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.

<http://www.gatsby.ucl.ac.uk/~balaji/udl2021/accepted-papers/UDL2021-paper-033.pdf>

BETH Dataset (cont.)

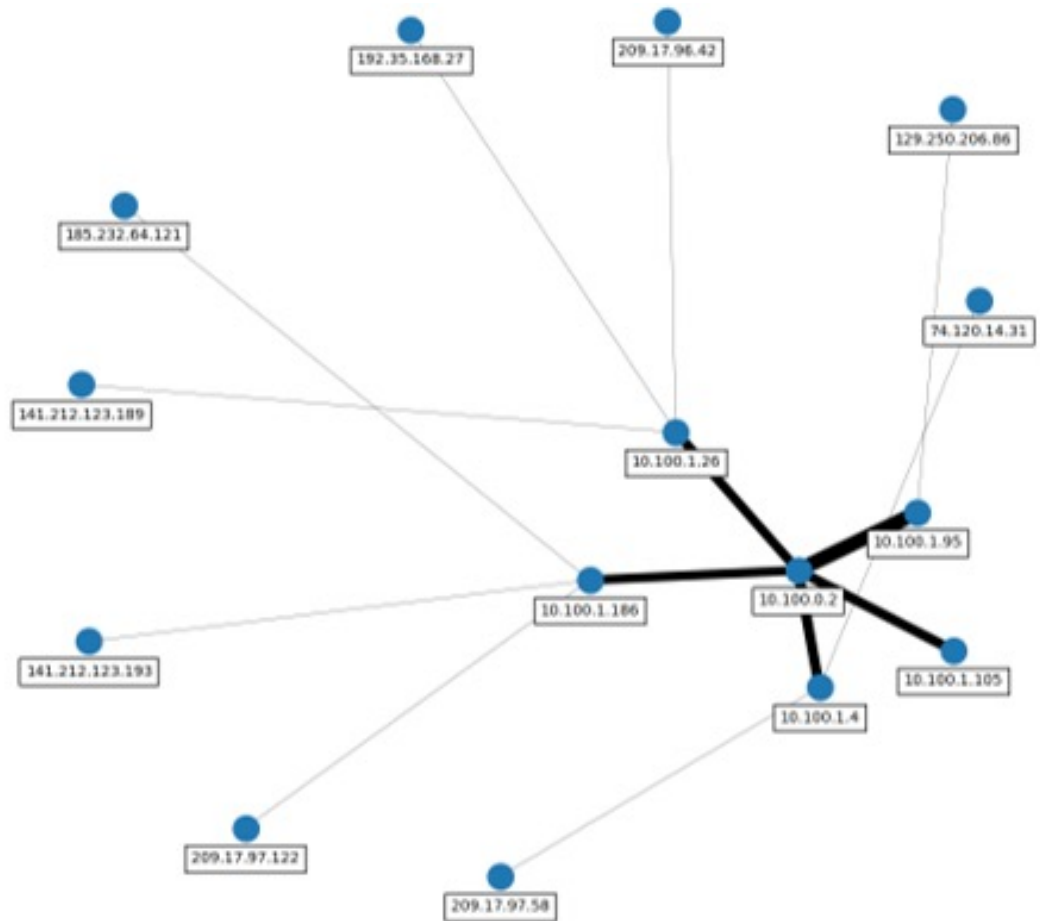
- System log files

| timestamp | processId | threadId | parentProcessId | userId | mountNamespace | processName | hostName | eventId | eventName | sockAddress | argsNum | returnValue | args | sus | evil |
|------------|-----------|----------|-----------------|--------|----------------|-----------------|-----------------|---------|--------------------|--------------|---------|-------------|---------------|-----|------|
| 129.050634 | 382 | 382 | 1 | 101 | 4026532232 | systemd-resolve | ip-10-100-1-217 | 41 | socket | [1401591956] | 3 | 15 | {{'name': 'dd | 0 | 0 |
| 129.051238 | 379 | 379 | 1 | 100 | 4026532231 | systemd-network | ip-10-100-1-217 | 41 | socket | [1398532280] | 3 | 15 | {{'name': 'dd | 0 | 0 |
| 129.051434 | 1 | 1 | 0 | 0 | 4026531840 | systemd | ip-10-100-1-217 | 1005 | security_file_open | [1403628671] | 4 | 0 | {{'name': 'pa | 0 | 0 |
| 129.051481 | 1 | 1 | 0 | 0 | 4026531840 | systemd | ip-10-100-1-217 | 257 | openat | [] | 4 | 17 | {{'name': 'di | 0 | 0 |
| 129.051522 | 1 | 1 | 0 | 0 | 4026531840 | systemd | ip-10-100-1-217 | 5 | fstat | [1403628671] | 2 | 0 | {{'name': 'fd | 0 | 0 |
| 129.051635 | 1 | 1 | 0 | 0 | 4026531840 | systemd | ip-10-100-1-217 | 3 | close | [1403628672] | 1 | 0 | {{'name': 'fd | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

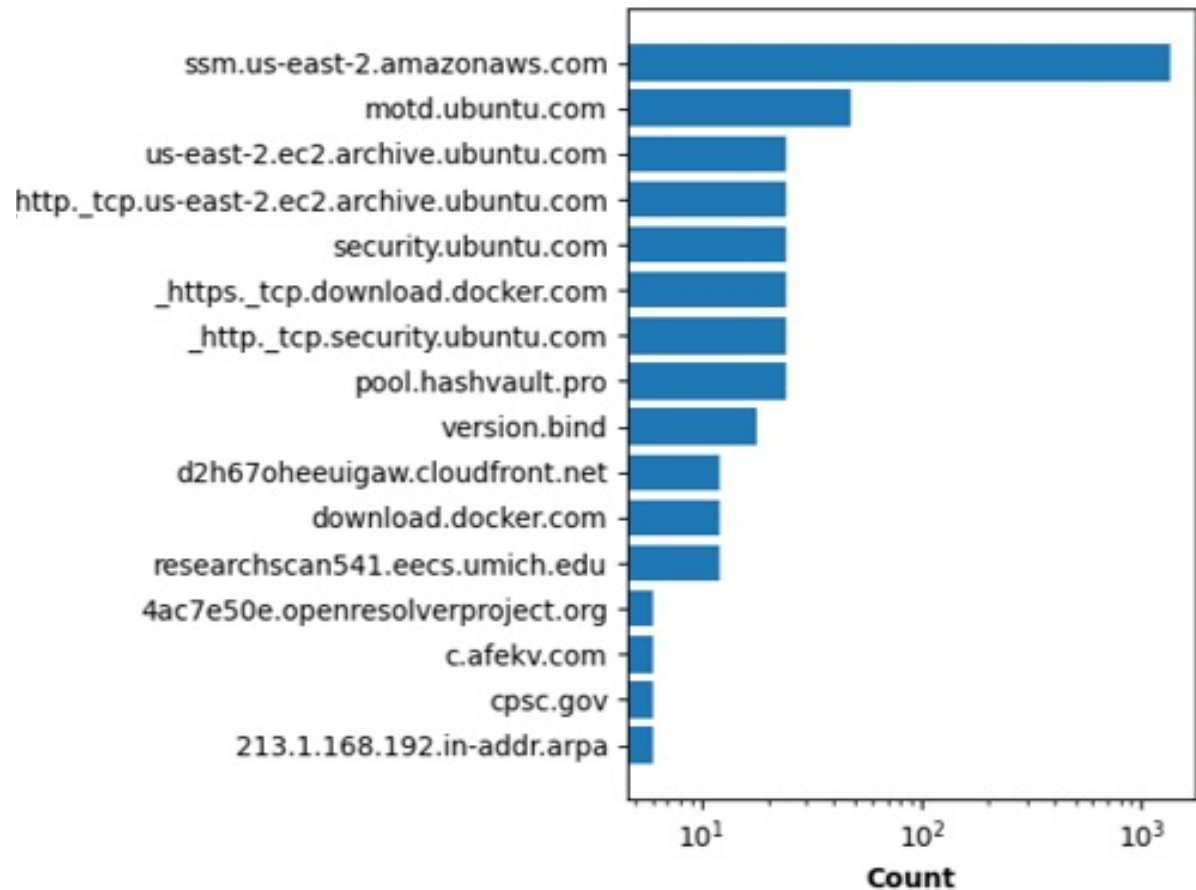
- DNS query log files

| Timestamp | SourceIP | DestinationIP | DnsQuery | DnsAnswer | DnsAnswerTTL | DnsQueryNames | DnsQueryClass | DnsQueryType | NumberOfAnswers | DnsResponseCode | DnsOpCode | SensorId | sus | evil |
|----------------------|--------------|---------------|-----------------------------------|-----------|--------------|------------------|---------------|--------------|-----------------|-----------------|-----------|-----------------|-----|------|
| 2021-05-16T17:13:14Z | 10.100.1.95 | 10.100.0.2 | ssm.us-east-2.amazonaws.com | | | ssm.us-east-2.am | ['IN'] | ['A'] | 0 | 0 | 0 | ip-10-100-1-95 | 0 | 0 |
| 2021-05-16T17:13:14Z | 10.100.0.2 | 10.100.1.95 | ssm.us-east-2.am[["52.95.19.240"] | ['17'] | | ssm.us-east-2.am | ['IN'] | ['A'] | 1 | 0 | 0 | ip-10-100-1-95 | 0 | 0 |
| 2021-05-16T17:13:14Z | 10.100.1.95 | 10.100.0.2 | ssm.us-east-2.amazonaws.com | | | ssm.us-east-2.am | ['IN'] | ['AAAA'] | 0 | 0 | 0 | ip-10-100-1-95 | 0 | 0 |
| 2021-05-16T17:13:14Z | 10.100.0.2 | 10.100.1.95 | ssm.us-east-2.amazonaws.com | | | ssm.us-east-2.am | ['IN'] | ['AAAA'] | 0 | 0 | 0 | ip-10-100-1-95 | 0 | 0 |
| 2021-05-16T17:13:16Z | 10.100.1.186 | 10.100.0.2 | ssm.us-east-2.amazonaws.com | | | ssm.us-east-2.am | ['IN'] | ['A'] | 0 | 0 | 0 | ip-10-100-1-186 | 0 | 0 |
| 2021-05-16T17:13:16Z | 10.100.0.2 | 10.100.1.186 | ssm.us-east-2.am[["52.95.21.209"] | ['41'] | | ssm.us-east-2.am | ['IN'] | ['A'] | 1 | 0 | 0 | ip-10-100-1-186 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

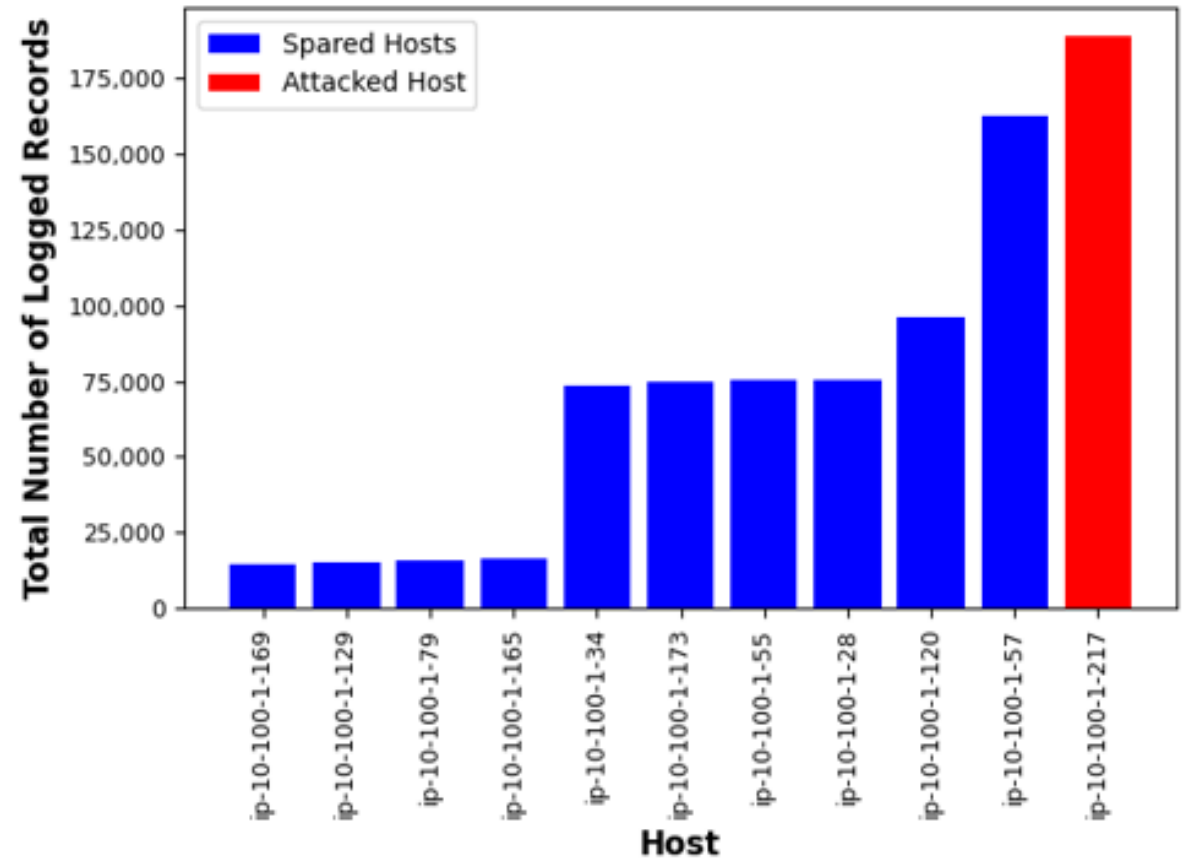
DNS Query Traffic between IP Addresses



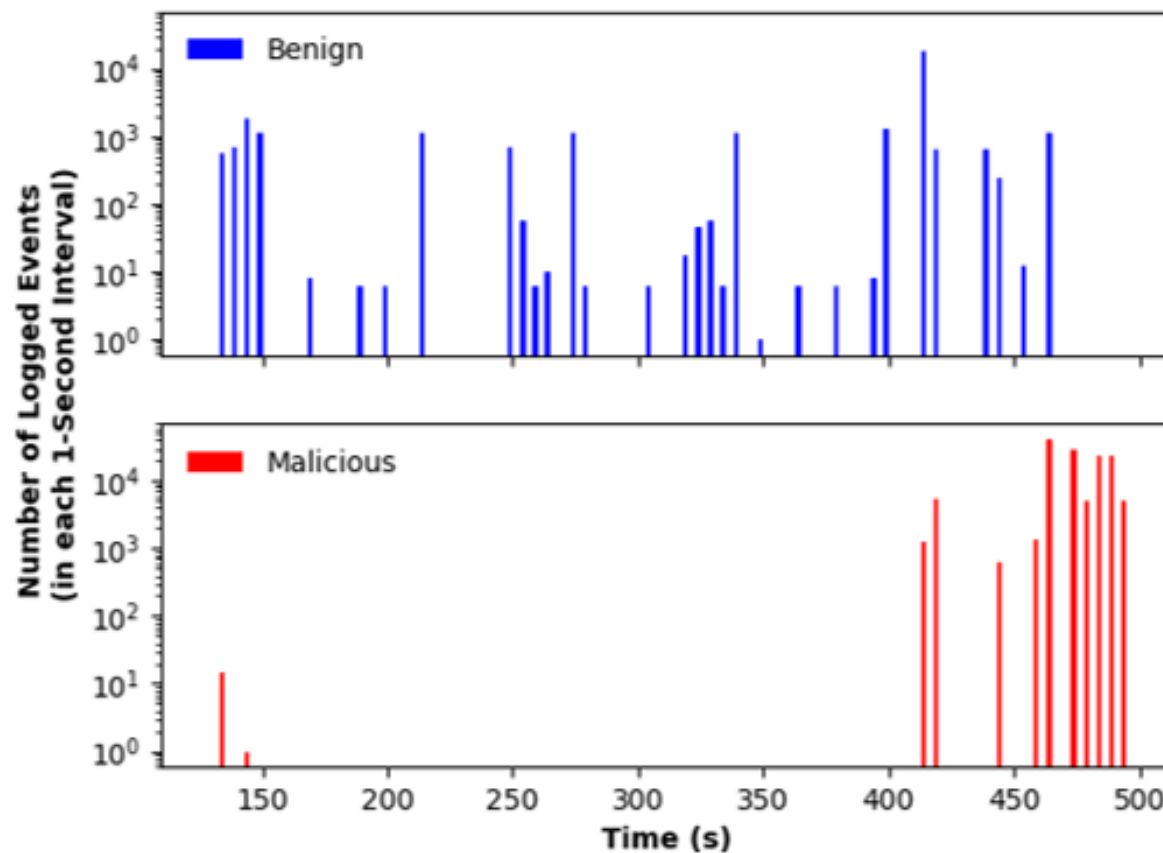
DNS Query Volume by Domain



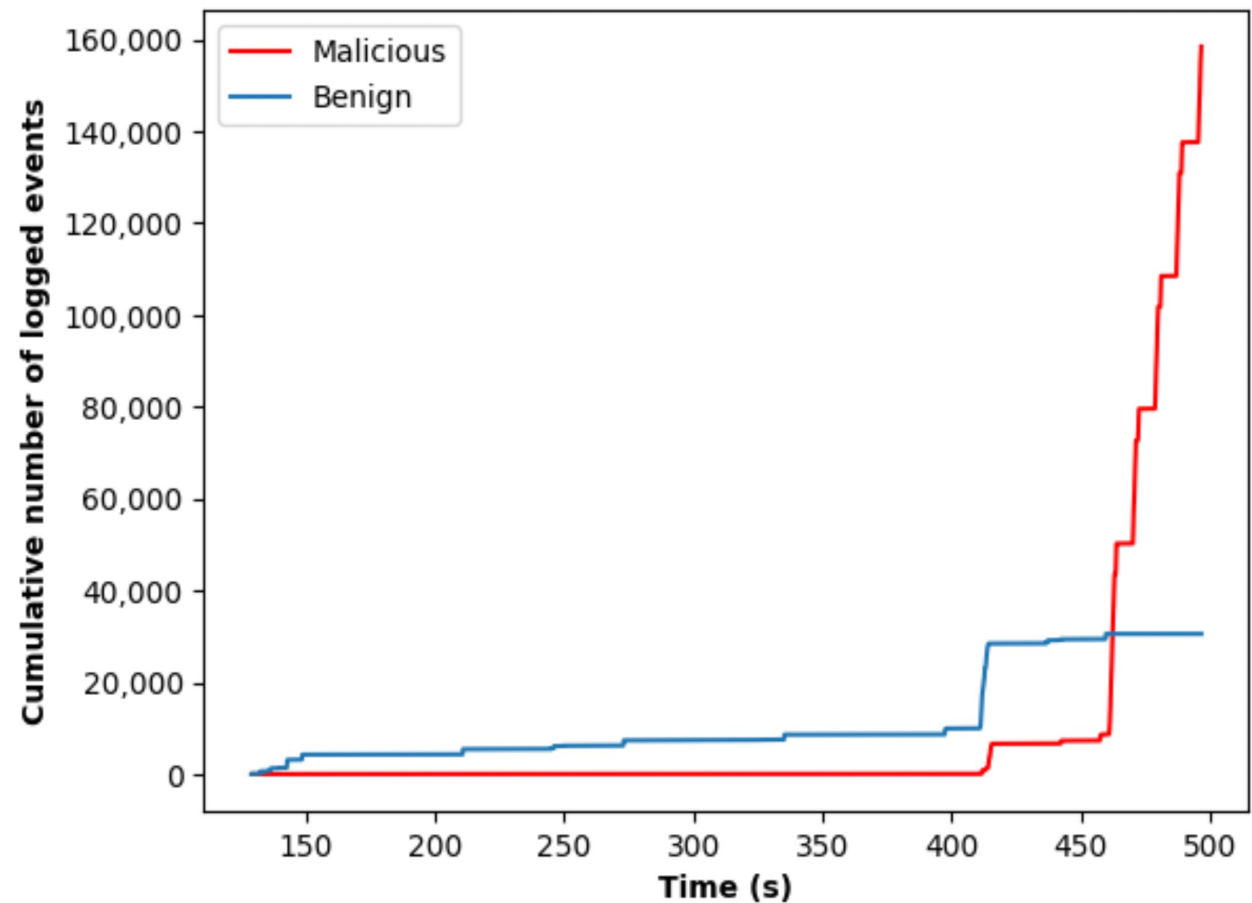
Logged Events by Host



Logged Events in Time on the Attacked Host



Logged Events in Time on the Attacked Host



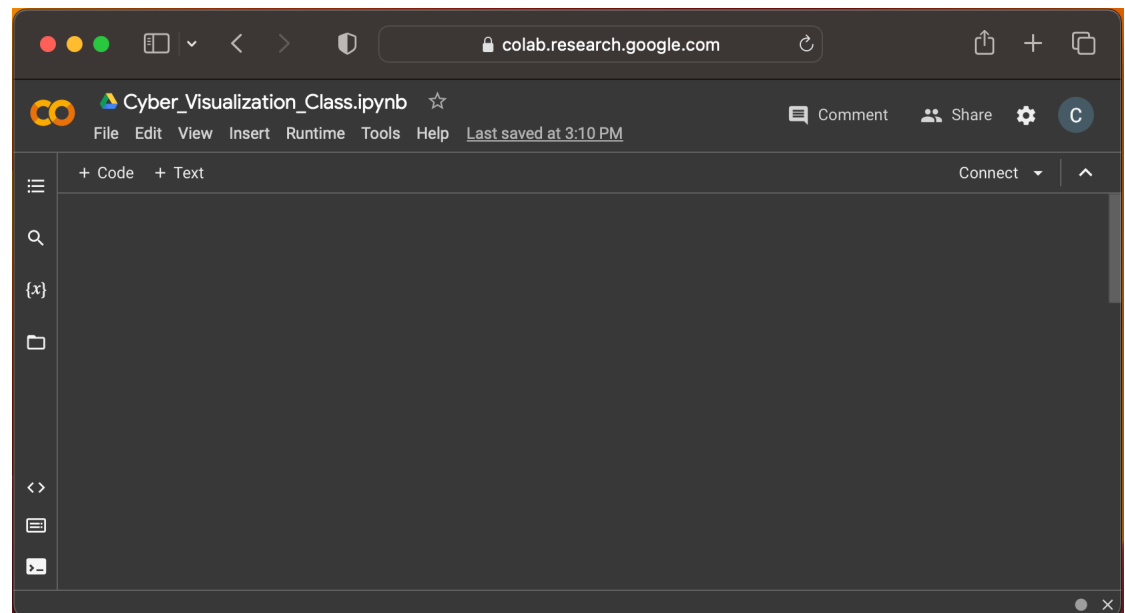
Lab Environment Setup

Environment Options

- **Option 1: Google Colab (recommended)**
 - Write and execute code in a browser
 - No installation required
 - Requires gmail account
 - <https://colab.research.google.com>
- **Option 2: Student-preferred python environment**

Getting Started

- Sign into your gmail account in your browser
- Go to <https://colab.research.google.com>
- Click “File → New Notebook”



Tour of Google Colab Functionality

- Add and execute code cells
- Add formatted text cells
- Import data

Hands-on Exercise

Install and Import Libraries

```
!pip install umap-learn  
import umap
```

Read .csv into Pandas Dataframe

```
df = pandas.read_csv('data.csv')  
df.dtypes
```


View First 5 Records

```
df.head()
```

Histograms of the Raw Data

```
df.hist(figsize=(12, 10))  
plt.tight_layout()
```

Histograms of the Engineered Features

```
df_eng, X, y = preprocess(df)
df_eng.hist(figsize=(12, 10))
plt.tight_layout()
```

Correlations Plot

```
correlations = df_eng.corr()  
seaborn.heatmap(correlations, cmap='vlag')  
plt.show()
```

UMAP Dimensionality Reduction

```
manifold = umap.UMAP().fit(X)  
X_reduced = manifold.transform(X)
```


Fit Anomaly Detection Model

```
model = sklearn.ensemble.IsolationForest().fit(X)
```



THANK YOU!