```
wget https://data.hpc.imperial.ac.uk/resolve/\?doi\=9422\&file\=4\&access\= -O full_BETH_dataset.zip
```

Kaggle Dataset

Workshop Paper

&

https://www.camlis.org/2021/schedule

BETH Dataset

Real Cybersecurity Data for Anomaly Detection Research

Kate Highnam, Kai Arulkumaran, Zachary Hanif, Nicholas R. Jennings

Imperial College London
ARAYA
University of Maryland
Loughborough University

TL;DR

- New cybersecurity dataset for anomaly detection benchmarking
  - Over 8 million data points
  - Modern host activity and attacks (in the Cloud!)
  - Fully labelled (by hand)
  - Each host contains benign activity and at most a single attack
    - Constrained vulnerability during data collection for accessibility and control over noise
    - Ideal for behavioral analysis and other research tasks
  - Further data is currently being collected
- Benchmarking conducted using:
  - **Robust Covariance** [Rousseuw, 1984]
  - **One-Class SVM** [Schölkopf et al., 2001]
  - **Isolation Forest** [Liu et al., 2008]
  - **VAE** [Kingma & Welling, 2013] **+ DoSE-SVM** [Morningstar, et al. 2021]

The Problem

Unsupervised Anomaly Detection

Defining the problem

GOAL

Identify the unexpected within the data

RULES

Unsupervised, ideally labels available for verification

RESULTS

Anomaly?

Outlier?

Changing Distribution?

Unsupervised Anomaly Detection

Defining the problem

GOAL

Identify the unexpected within the data

RULES

Unsupervised, ideally labels available for verification

RESULTS

Anomaly?

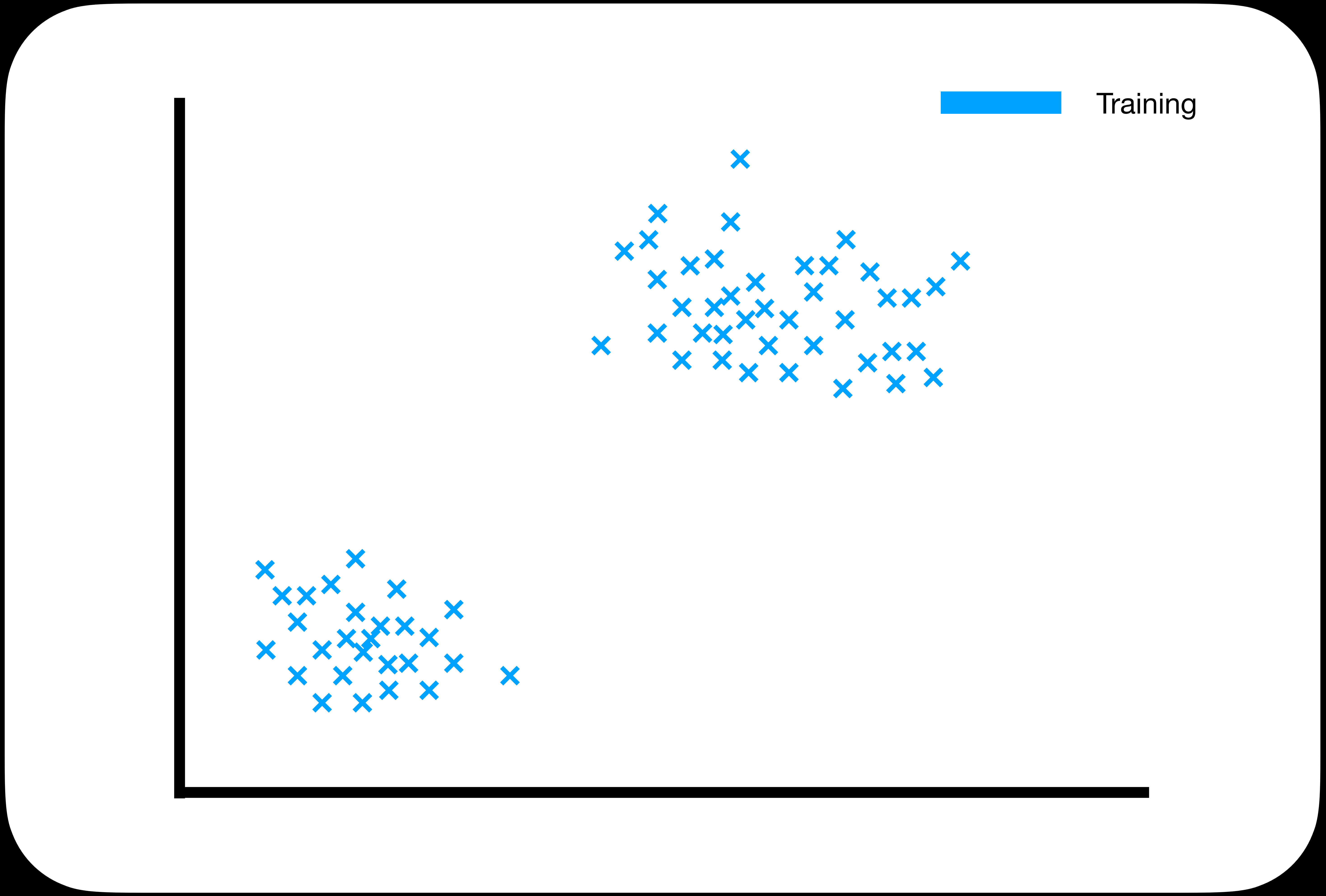
Outlier?

Changing Distribution?

Robust Systems

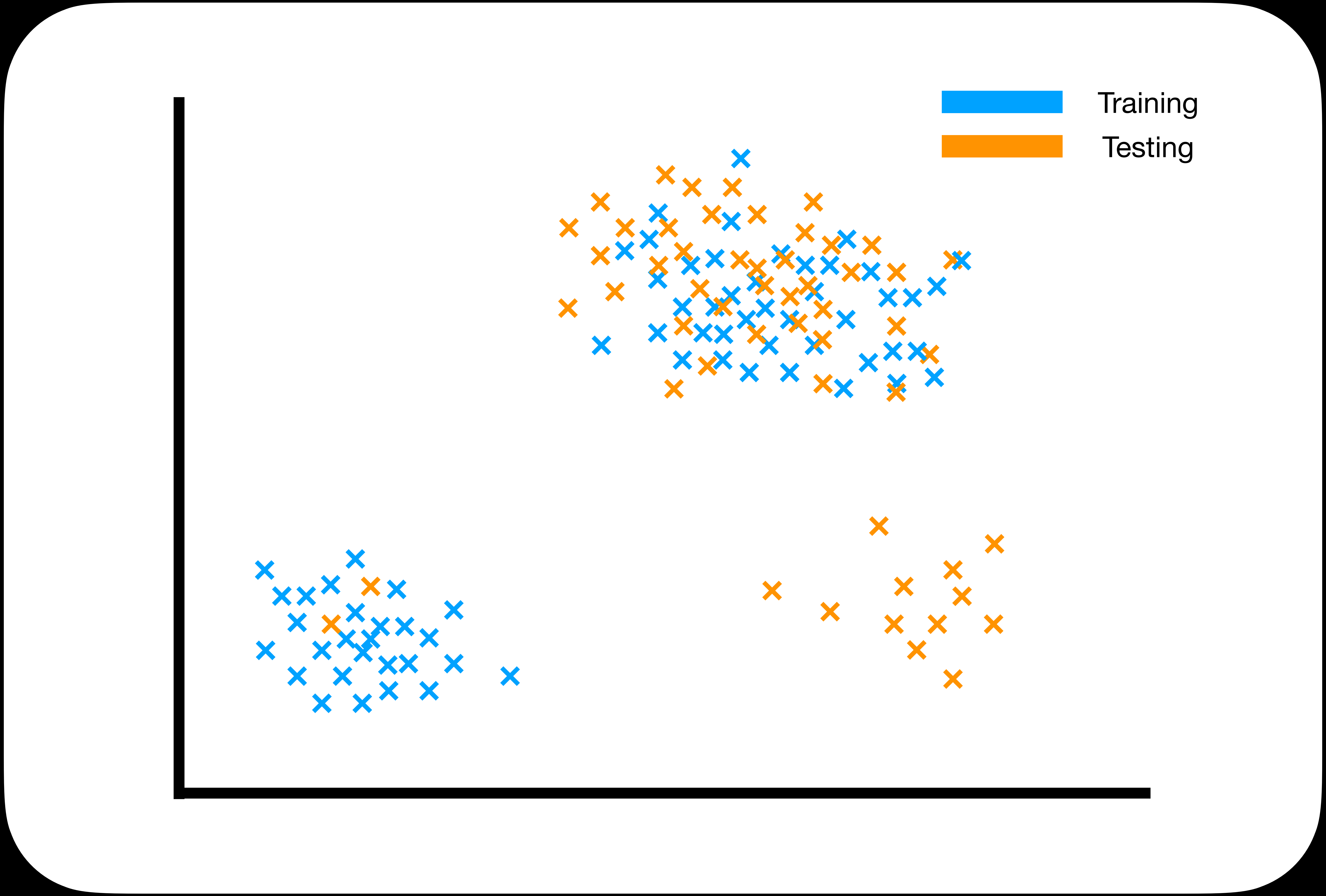
Machine Learning Datasets

For Unsupervised Anomaly Detection



Machine Learning Datasets

For Unsupervised Anomaly Detection



Machine Learning Datasets

For (Unsupervised) Anomaly Detection

- MNIST
- EMNIST
- FashionMNIST
- CIFAR10
- SVHN (Digits in Natural Images)
- CelebA
- ImageNet
- STL-10
- Reuters
- 20newsgroup

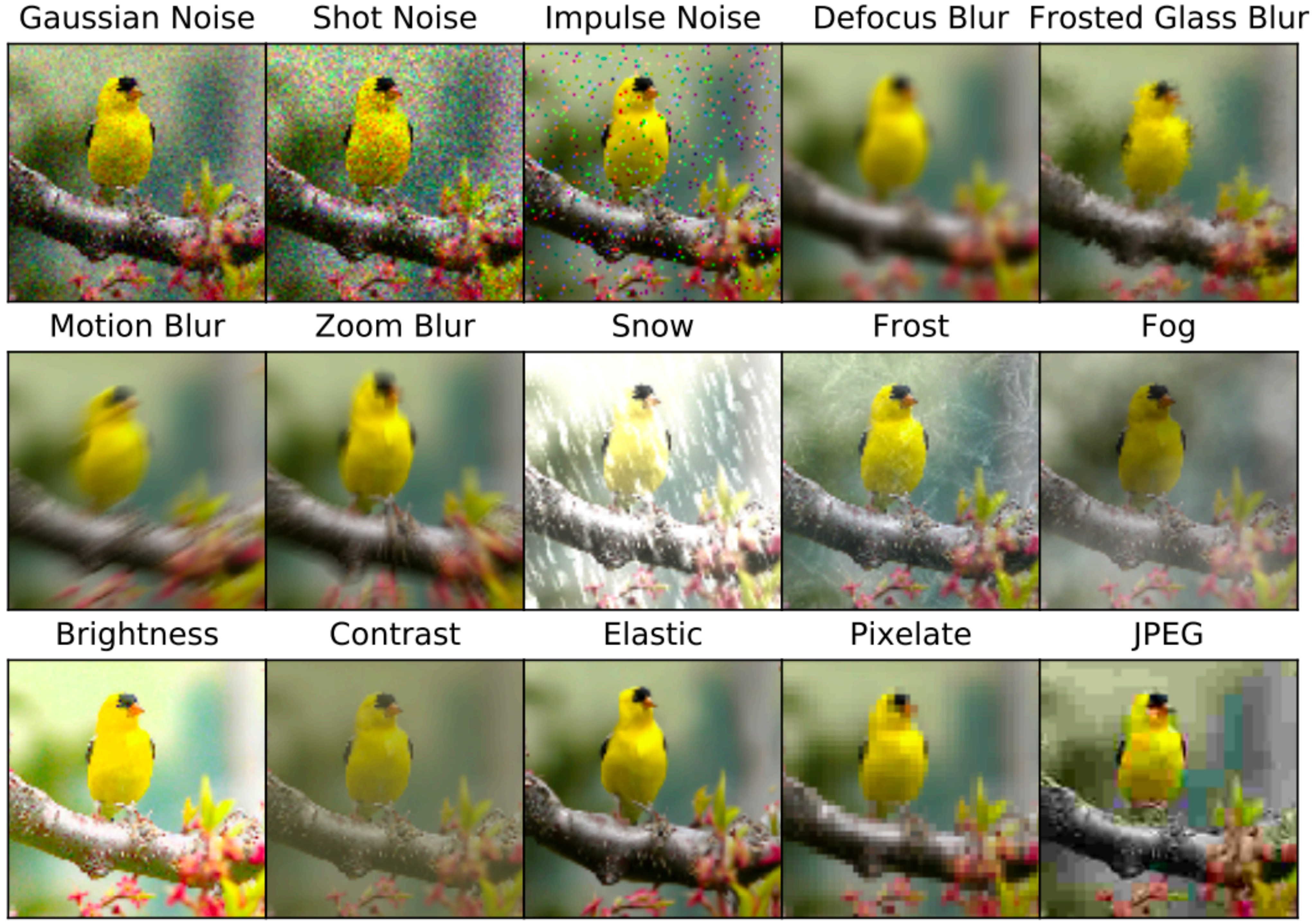


Figure 1: Our IMAGENET-C dataset consists of 15 types of algorithmically generated corruptions from noise, blur, weather, and digital categories. Each type of corruption has five levels of severity, resulting in 75 distinct corruptions. See different severity levels in Appendix B.

Hendrycks & Dietterich, 2019

Machine Learning Datasets

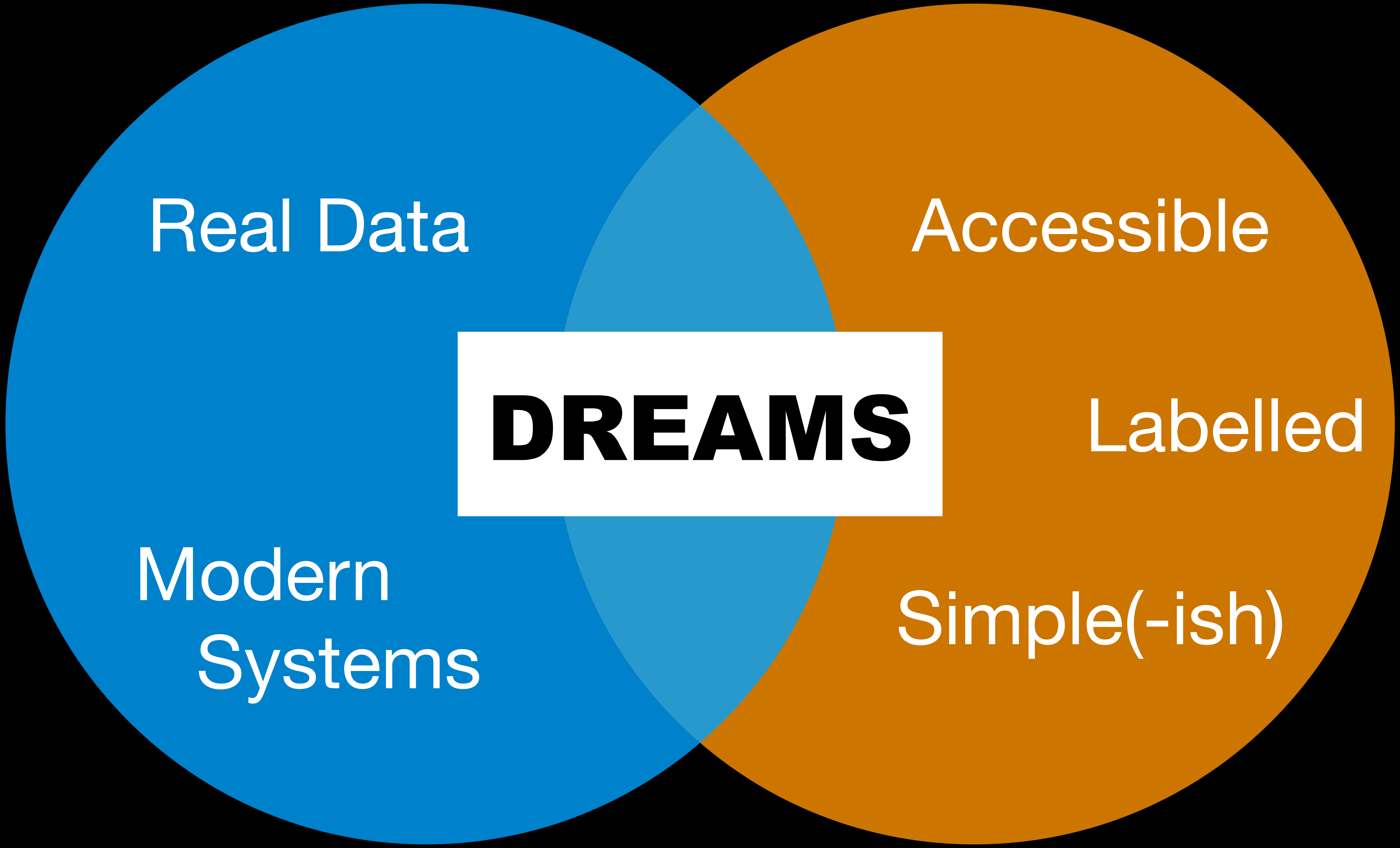
For (Unsupervised) Anomaly Detection *in cyber security*

- MNIST
- EMNIST
- FashionMNIST
- CIFAR10
- SVHN (Digits in Natural Images)
- CelebA
- ImageNet
- STL-10
- Reuters
- 20newsgroup

- DARPA 1998/1999
- KDD 1999
- NSL-KDD (2009)
- ISCX IDS 2012

Machine Learning Datasets

Ideal for ML experts while relevant for cyber security



Machine Learning Datasets

for cyber security

	Size	Includes Kernel Traffic	Real Live Traffic	Limited User Activity	Simple Network Environment	Cloud
DARPA 1998/1999	Not Stated	○	✕	✕	✕	✕
KDD 1999	7+ million records	✕	✕	✕	✕	✕
NSL-KDD (2009)	~2 million records	✕	✕	✕	✕	✕
ISCX IDS 2012	~2 million and ~81.1G of pcaps	✕	✕	✕	✕	✕

We need new data.

What is a Honeypot?

A pleasant looking trap for unpleasant people

What is a Honeypot?

A pleasant looking trap for unpleasant people

BETH

BPF-Extended Tracking Honeypot

BETH

BPF-Extended Tracking Honeypot

To be continued…

Honeypot Tracking

Tracking inside and out

Kernel-level
Process Calls

Network Activity

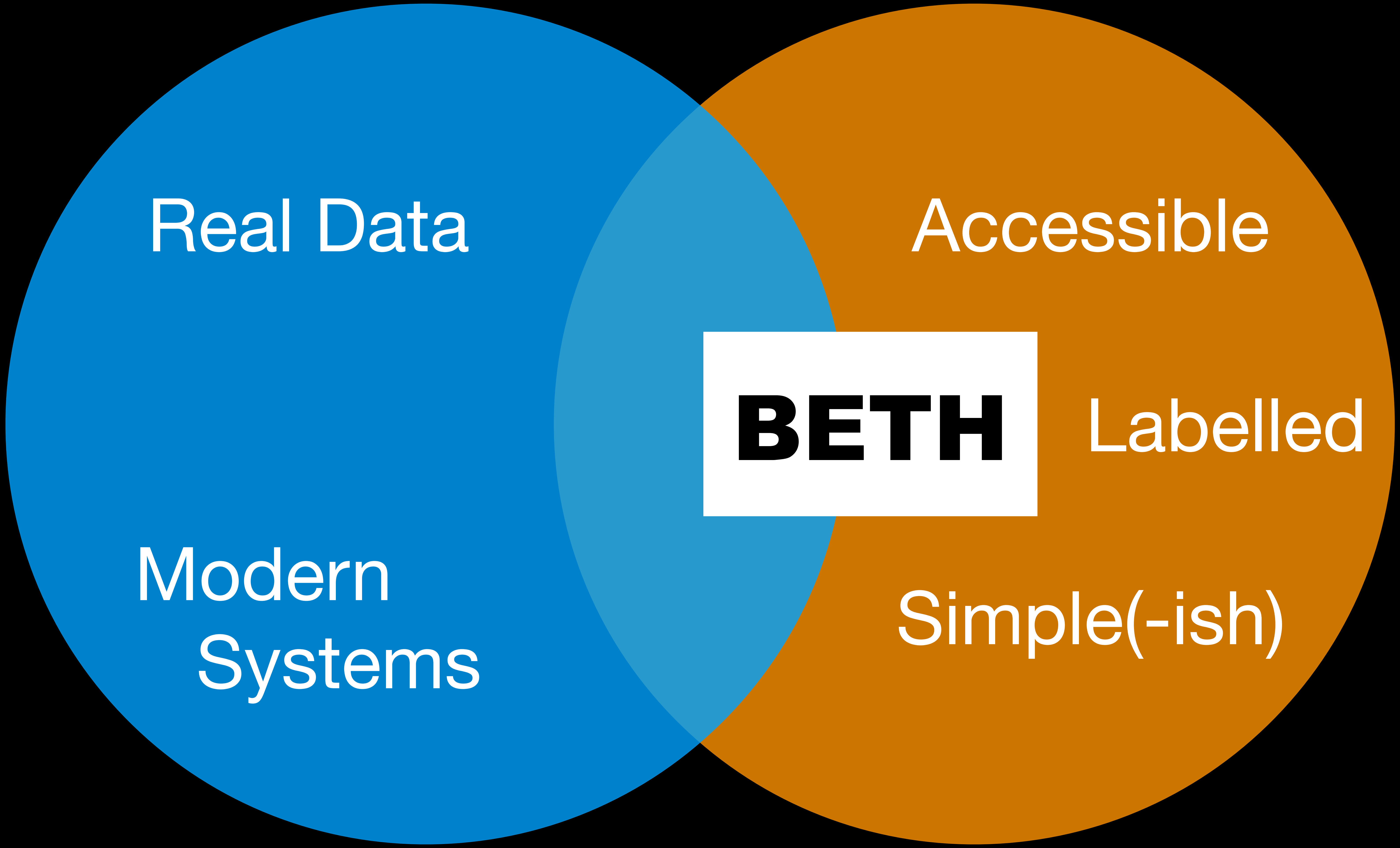
Honeypot Tracking

Single instances in the Cloud

The Data

Machine Learning Datasets

Ideal for ML experts while relevant for cyber security



Machine Learning Datasets

For Anomaly Detection *for cyber security*

	Size	Includes Kernel Traffic	Real Live Traffic	Limited User Activity	Simple Network Environment	Cloud
DARPA 1998/1999	Not Stated	O	X	X	X	X
KDD 1999	7+ million records	X	X	X	X	X
NSL-KDD (2009)	~2 million records	X	X	X	X	X
ISCX IDS 2012	~2 million and ~81.1G of pcaps	X	X	X	X	X
BETH	8+ million records*	O	O	O	O	O

*We are currently recording more data from our honeypots and will add them to the dataset for public use

BETH Dataset

Logs from the BETH, kernel and network… but mostly kernel

Feature	Type	Description
timestamp	float	Seconds since system boot
processId*	int	Integer label for the process spawning this log
threadId	int	Integer label for the thread spawning this log
parentProcessId*	int	Parent's integer label for the process spawning this log
userId*	int	Login integer ID of user spawning this log
mountNamespace*	int (long)	Set mounting restrictions this process log works within
processName	String	String command executed
hostName	String	Name of host server
eventId*	int	ID for the event generating this log
eventName	String	Name of the event generating this log
argsNum*	int	Length of args
returnValue*	int	Value returned from this event log (usually 0)
stackAddresses	list of int	memory values relevant to the process
args	list of dictionaries	List of arguments passed to this process
sus	int (0 or 1)	Binary label as a suspicious event (1 is suspicious, 0 is not)
evil	int (0 or 1)	Binary as a known malicious event (0 is benign, 1 is not)

Benign data = not evil, maybe sus

Malicious data = evil and sus

BETH Dataset

DNS (Network) logs

⚠️ Not all process call logs have corresponding DNS logs… ⚠️

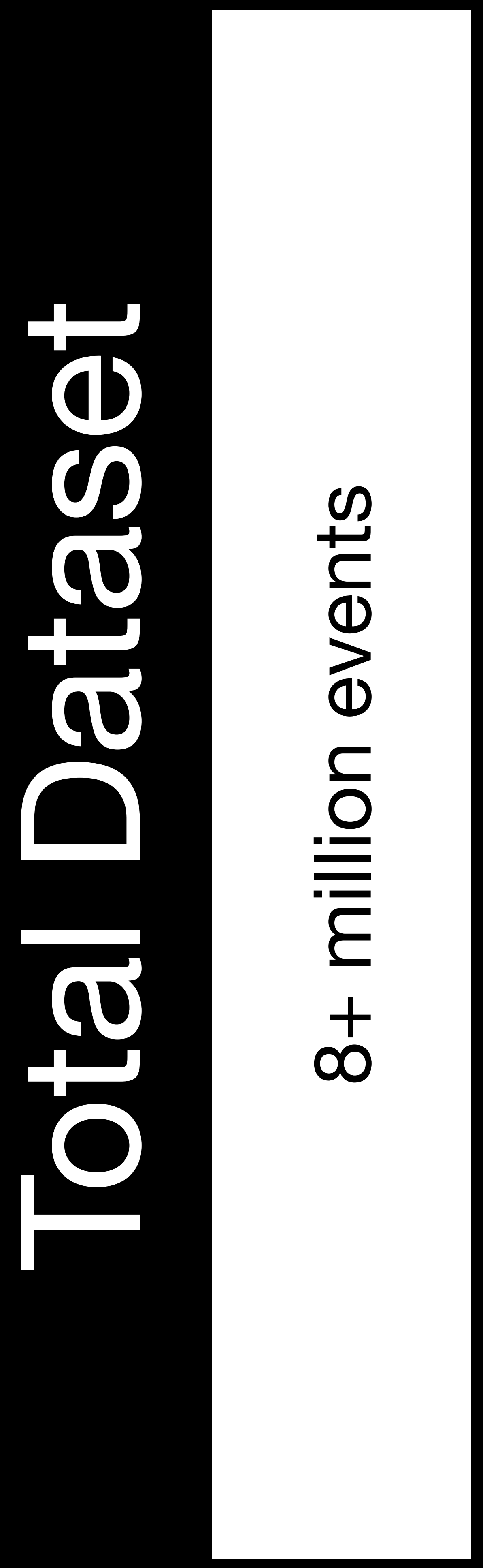
FEATURE	TYPE	DESCRIPTION
TIMESTAMP	STRING	DATE AND TIME IN THE FORMAT "YYYY-MM-DDTHH:MM:SSZ" FOR WHEN THE PACKET WAS SENT OR RECEIVED
SOURCEIP	STRING	SOURCE IP ADDRESS OF THE PACKET
DESTINATIONIP	STRING	DESTINATION IP ADDRESS OF THE PACKET
DNSQUERY	STRING	THE SENT DNS QUERY (E.G. THE URL SUBMITTED - "GOOGLE.COM")
DNSANSWER	LIST OF STRINGS	DNS RESPONSE; CAN BE NULL
DNSANSWERTTL	LIST OF STRINGS (INT)	LIST OF INTEGERS SENT AS STRINGS, CAN BE NULL; THE TIME TO LIVE OF THE DNS ANSWER
DNSQUERYNAMES	LIST OF STRINGS	NAME OF THE REQUESTED RESOURCE
DNSQUERYCLASS	LIST OF STRINGS	CLASS CODE FOR THE RESOURCE QUERY
DNSQUERYTYPE	LIST OF STRINGS	TYPE OF RESOURCE RECORD (A, AAAA, MX, TXT, ETC.)
NUMBEROFANSWERS	STRING (INT)	NUMBER OF ANSWER HEADERS IN THE PACKET
DNSOPCODE	STRING (INT)	HEADER INFORMATION REGARDING WHICH OPERATION THIS PACKET WAS SENT (E.G. STANDARD QUERY IS 0)
SENSORID	STRING	SAME AS THE HOSTNAME IN THE PROCESS RECORDS; NAME OF HOST SERVER
SUS	INT (0 OR 1)	BINARY LABEL AS A SUSPICIOUS EVENT (1 IS SUSPICIOUS, 0 IS NOT)
EVIL	INT (0 OR 1)	BINARY AS A KNOWN MALICIOUS EVENT (0 IS BENIGN, 1 IS NOT)

Currently extending this to provide full PCAP :)

Benchmarks

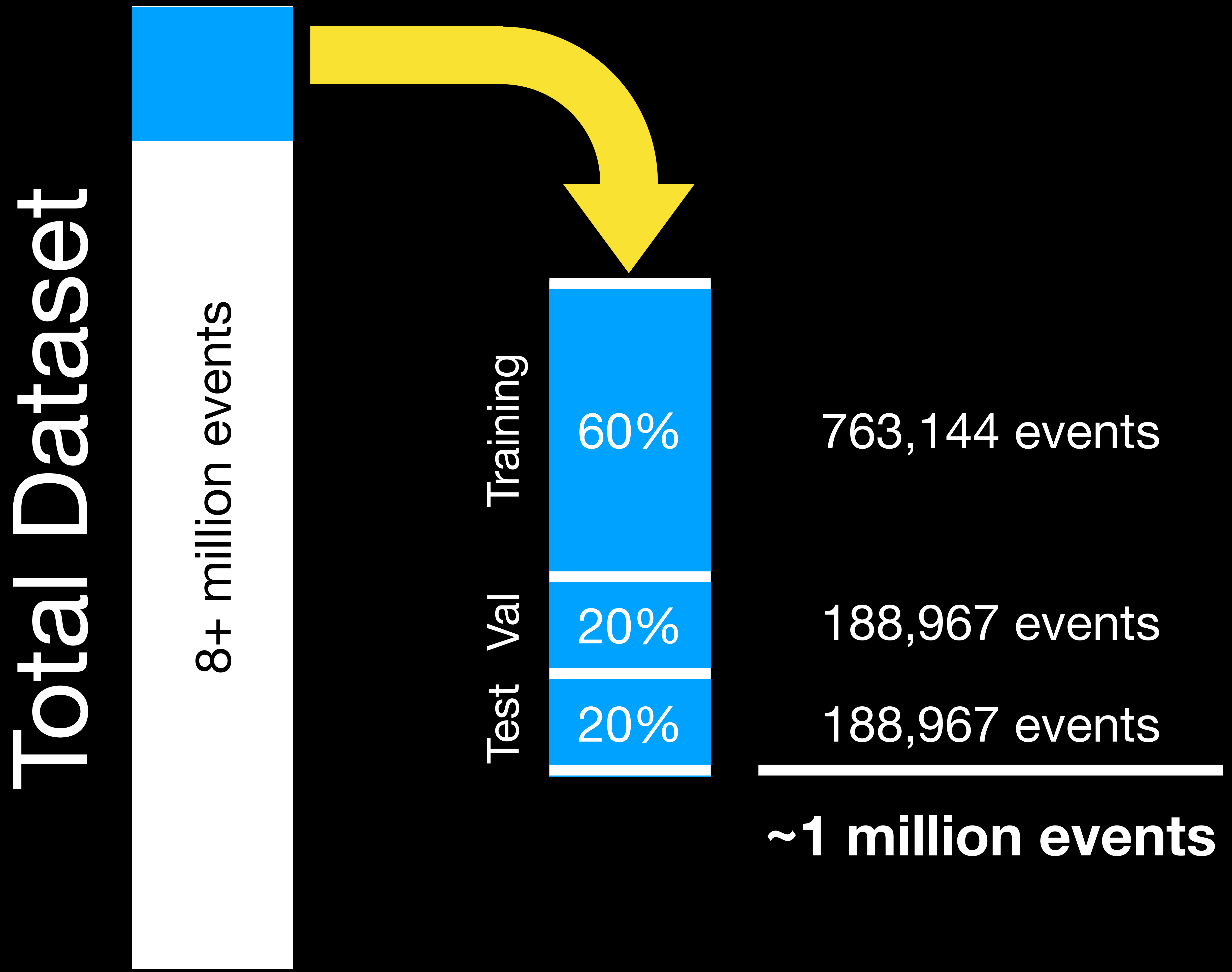
BETH Dataset Statistics

Training, Validation, and Testing Benchmarks



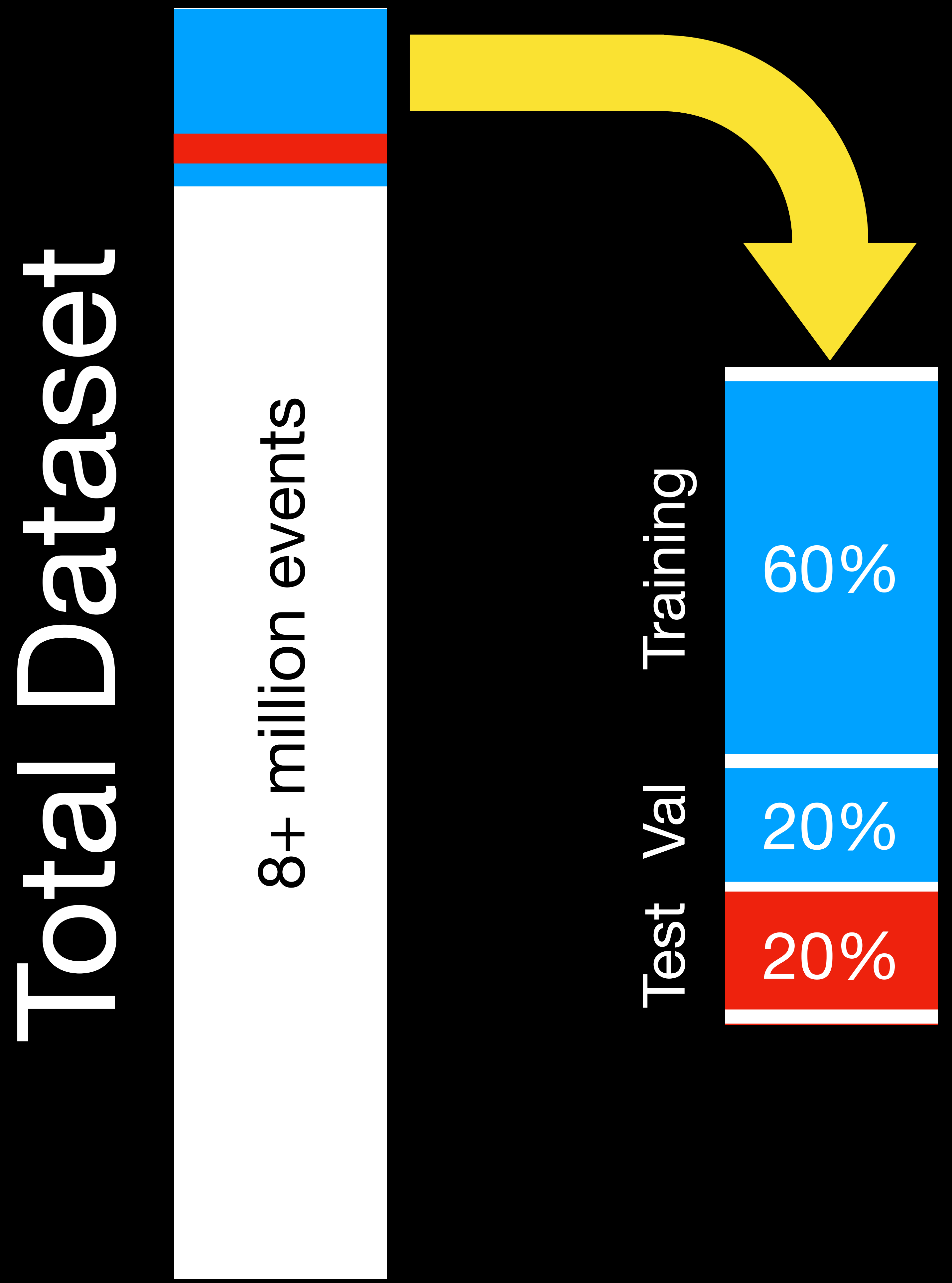
BETH Dataset Statistics

Training, Validation, and Testing Benchmarks



BETH Dataset Statistics

Training, Validation, and Testing Benchmarks



763,144 events

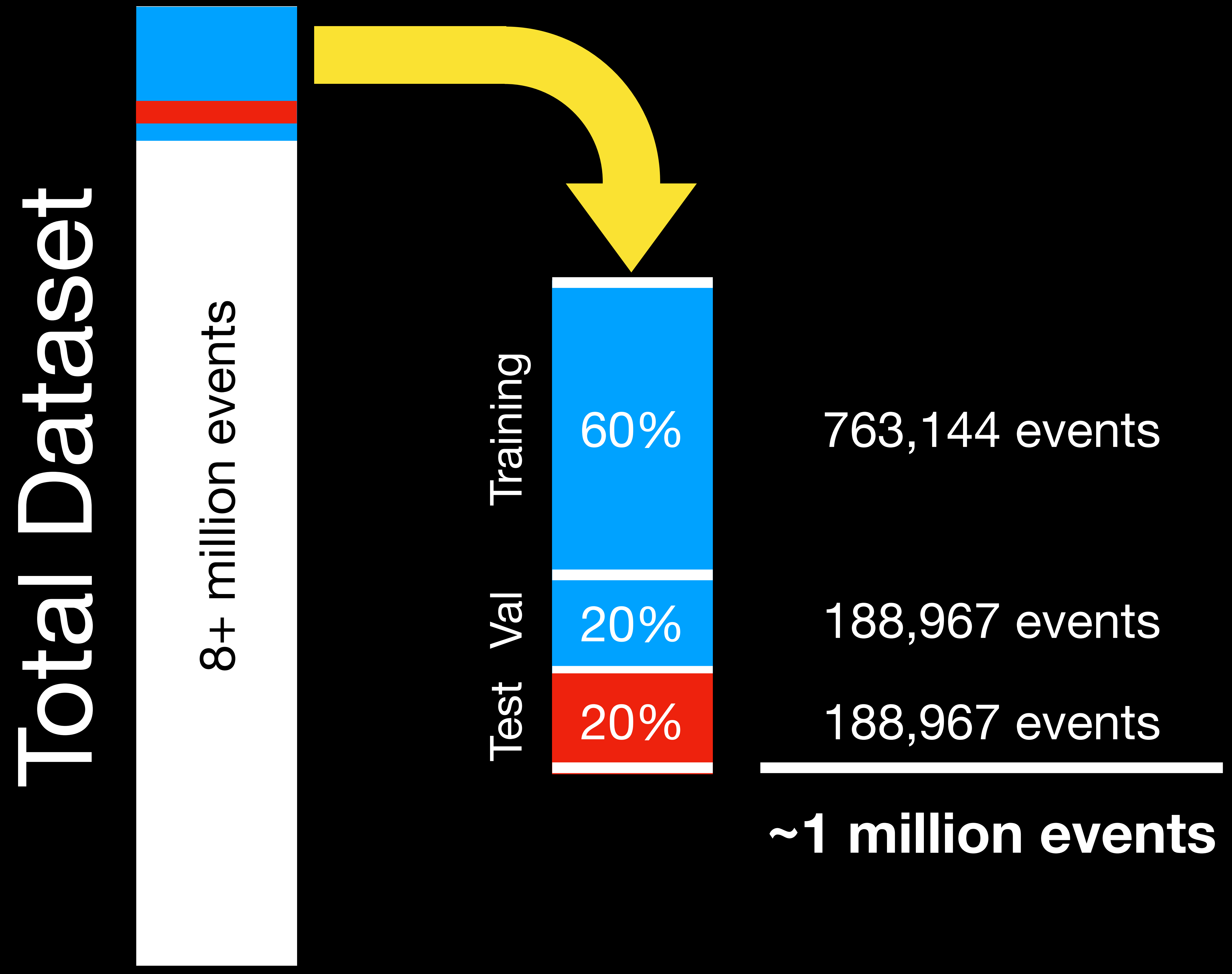
188,967 events

188,967 events

~1 million events

BETH Dataset Statistics

Training, Validation, and Testing Benchmarks



Dataset	sus=0, evil=0	sus=1, evil=0	sus=1, evil=1
Training	761875 (99.8%)	1269 (0.02%)	0 (0.00%)
Validation	188181 (99.6%)	786 (0.04%)	0 (0.00%)
Testing	17508 (9.27%)	13027 (6.89%)	158432 (83.84%)

Benchmarks

Unsupervised Anomaly Detection Methods

Robust Covariance

One-Class SVM

Isolation Forest

VAE + DoSE (SVM)

Benchmarks

Unsupervised Anomaly Detection Methods

METHOD	AUROC
ROBUST COVARIANCE	0.519
ONE-CLASS SVM	0.605
IFOREST	**0.850**
VAE + DOSE (SVM)	0.698

Future Work

For me and you!

More data!

Full PCAPS!

Profile attacker/malware’s behavior

Benchmark more unsupervised anomaly detection methods

Fingerprint Analysis

Time series analysis of execution sequences to profile process names

Graph analysis of process relationships to find malicious cliques?

References

Datasets

- Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. Emnist: Extending mnist to handwritten letters. In 2017 International Joint Conference on Neural Networks (IJCNN), pp. 2921–2926. IEEE, 2017.
- Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. Proceedings of the International Conference on Learning Representations, 2019.
- Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-ofdistribution robustness. arXiv preprint arXiv:2004.06100, 2020.
- Hettich, S. and Bay, S. The uci kdd archive [http://kdd. ics. uci. edu]. irvine, ca: University of california. Department of Information and Computer Science, 152, 1999
- Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). URL http://www. cs.toronto.edu/˜kriz/cifar.html.
- Labs, M. L. 1998 darpa intrusion detection evaluation dataset, 1998. URL https: //www.ll.mit.edu/r-d/datasets/ 1998-darpa-intrusion-detection-evaluation-dataset
- Lang, K. Newsweeder: Learning to filter netnews. In Machine Learning Proceedings 1995, pp. 331–339. Elsevier, 1995.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradientbased learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.
- Lewis, D. D. Reuters-21578 text categorization collection data set, 1997
- Lippmann, R. P., Fried, D. J., Graf, I., Haines, J. W., Kendall, K. R., McClung, D., Weber, D., Webster, S. E., Wyschogrod, D., Cunningham, R. K., et al. Evaluating intrusion detection systems: The 1998 darpa off-line intrusion detection evaluation. In Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00, volume 2, pp. 12–26. IEEE, 2000
- Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.
- Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. arXiv preprint arXiv:1906.02530, 2019.
- Ring, Markus, et al. "A survey of network-based intrusion detection data sets." *Computers & Security* 86 (2019): 147-167.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- Shiravi, A., Shiravi, H., Tavallaee, M., and Ghorbani, A. A. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. Computers & Security, 31(3):357–374, 2012. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2011.12. 012. URL https://www.sciencedirect.com/ science/article/pii/S0167404811001672
- Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. A detailed analysis of the kdd cup 99 data set. In 2009 IEEE symposium on computational intelligence for security and defense applications, pp. 1–6. IEEE, 2009.
- Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

References

Benchmarks

- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.
- Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In 2008 eighth ieee international conference on data mining, pp. 413–422. IEEE, 2008.
- Morningstar, W., Ham, C., Gallagher, A., Lakshminarayanan, B., Alemi, A., and Dillon, J. Density of states estimation for out of distribution detection. In International Conference on Artificial Intelligence and Statistics, pp. 3232–3240. PMLR, 2021
- Rousseeuw, P. J. Least median of squares regression. Journal of the American statistical association, 79(388):871– 880, 1984.
- Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., ¨ and Williamson, R. C. Estimating the support of a highdimensional distribution. Neural computation, 13(7): 1443–1471, 2001.

```
wget https://data.hpc.imperial.ac.uk/resolve/\?doi\=9422\&file\=4\&access\= -O full_BETH_dataset.zip
```

Kaggle Dataset

Workshop Paper

&

https://www.camlis.org/2021/schedule

BETH Dataset

Real Cybersecurity Data for Anomaly Detection Research

Kate Highnam, Kai Arulkumaran, Zachary Hanif, Nicholas R. Jennings

Imperial College London

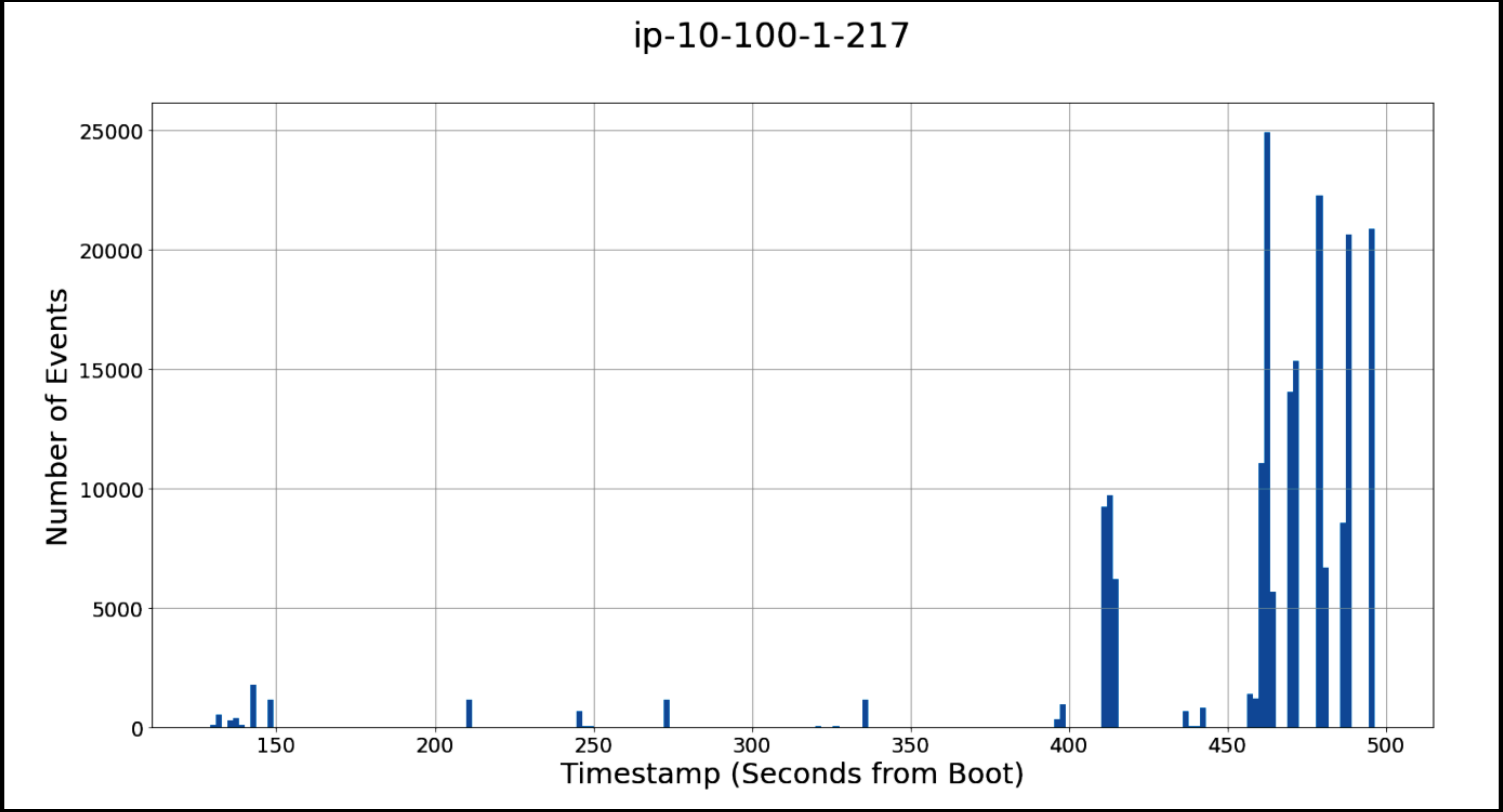
ARAYA

UNIVERSITY OF MARYLAND

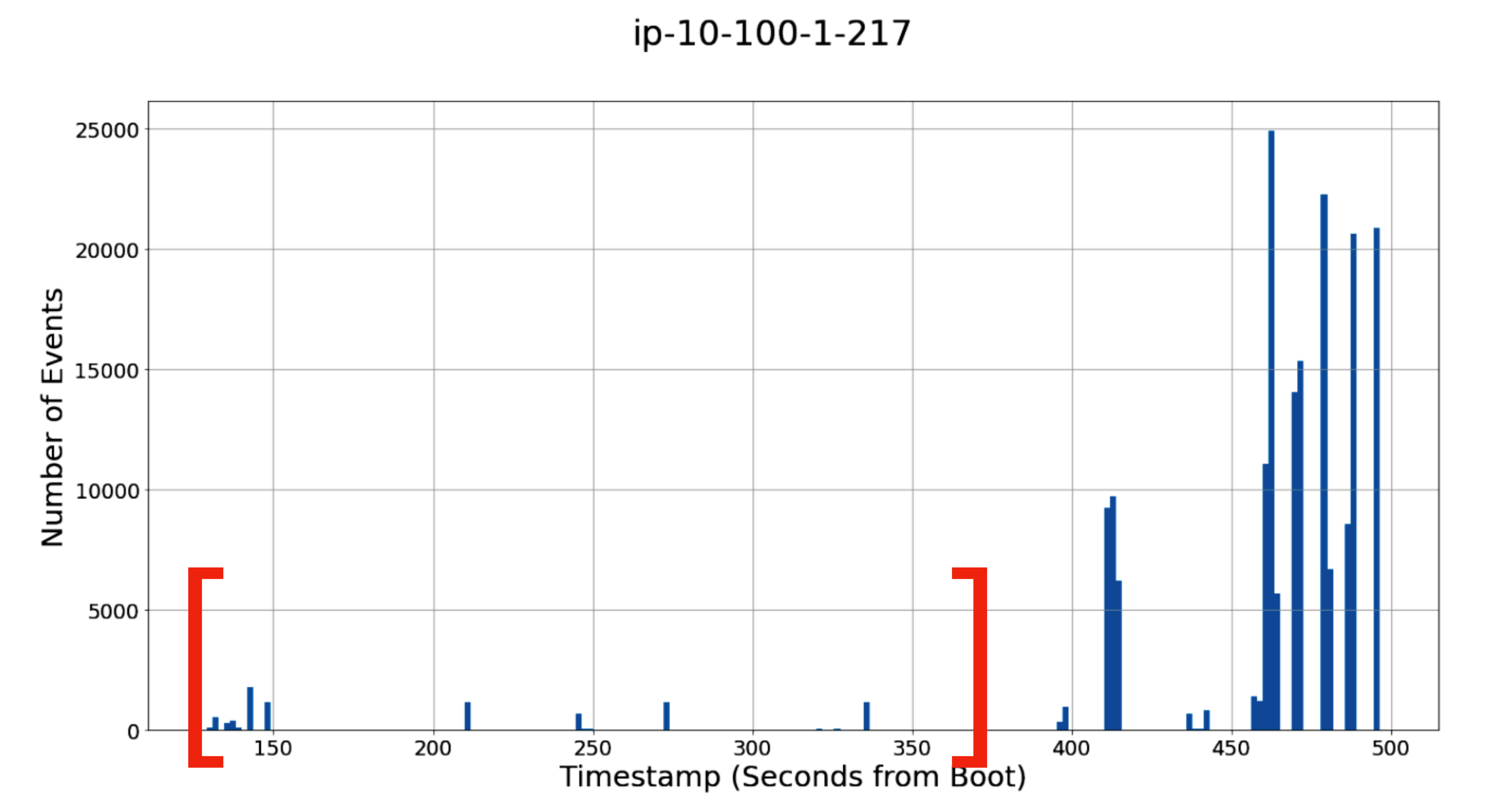
Loughborough University

Appendix

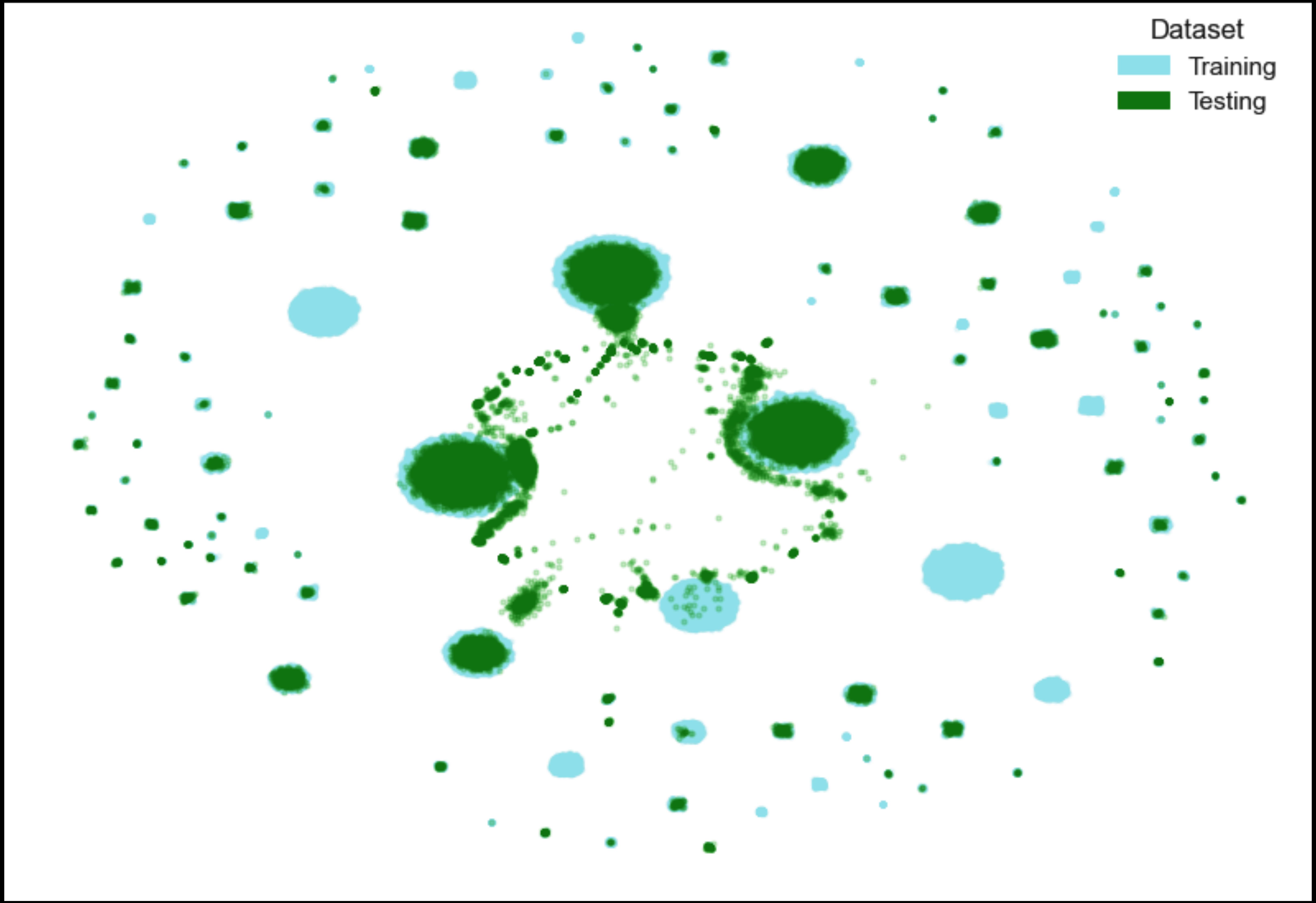
Timeline

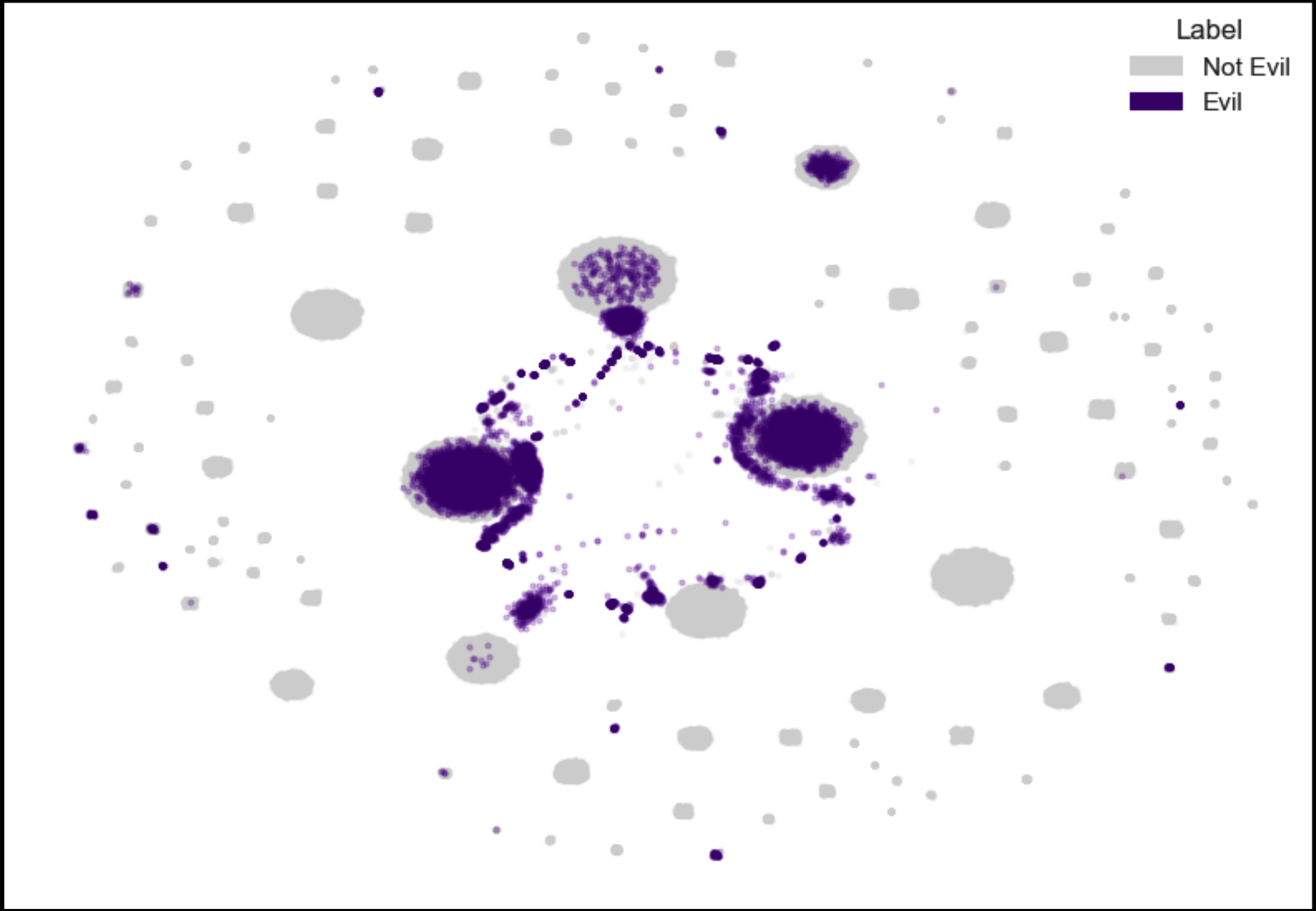

ip-10-100-1-217
Number of Events
25000
20000
15000
10000
5000
0
150
200
250
300
350
400
450
500
Timestamp (Seconds from Boot)

Timeline



ip-10-100-1-217
Number of Events
25000
20000
15000
10000
5000
0
150
200
250
300
350
400
450
500
Timestamp (Seconds from Boot)


Dataset
Training
Testing


Label
Not Evil
Evil

BETH Dataset

Logs from the BETH, kernel and network… but mostly kernel


```
timestamp,processId,parentProcessId,userId,processName,hostName,eventId,eventName,argsNum,returnValue,args,sus,evil
126.233491,1,0,0,systemd,ip-10-100-1-105,1005,security_file_open,4,0,"[{'name': 'pathname', 'type': 'const char*', 'value': '/proc/384/cgroup'}, {'name': 'flags', 'type': 'unsigned long', 'value': 'O_RDONLY|O_LARGEFILE'}, {'name': 'dev', 'type': 'dev_t', 'value': 5}, {'name': 'inode', 'type': 'unsigned long', 'value': 39481}]",0,0
126.233165,384,1,101,systemd-resolve,ip-10-100-1-105,41,socket,3,15,"[{'name': 'domain', 'type': 'int', 'value': 'AF_UNIX'}, {'name': 'type', 'type': 'int', 'value': 'SOCK_DGRAM|SOCK_CLOEXEC'}, {'name': 'protocol', 'type': 'int', 'value': 0}]",0,0
126.233559,1,0,0,systemd,ip-10-100-1-105,5,fstat,2,0,"[{'name': 'fd', 'type': 'int', 'value': 18}, {'name': 'statbuf', 'type': 'struct stat*', 'value': '0x7FFF1D8D98F0'}]",0,0
126.233681,1,0,0,systemd,ip-10-100-1-105,3,close,1,0,"[{'name': 'fd', 'type': 'int', 'value': 18}]",0,0
126.233796,384,1,101,systemd-resolve,ip-10-100-1-105,3,close,1,0,"[{'name': 'fd', 'type': 'int', 'value': 15}]",0,0
126.23353,1,0,0,systemd,ip-10-100-1-105,257,openat,4,18,"[{'name': 'dirfd', 'type': 'int', 'value': -100}, {'name': 'pathname', 'type': 'const char*', 'value': '/proc/384/cgroup'}, {'name': 'flags', 'type': 'unsigned long', 'value': 'O_RDONLY|O_CLOEXEC'}, {'name': 'mode', 'type': 'int*', 'value': 1223040804}]",0,0
126.23389,384,1,101,systemd-resolve,ip-10-100-1-105,1005,security_file_open,4,0,"[{'name': 'pathname', 'type': 'const char*', 'value': '/run/systemd/netif/links/5'}, {'name': 'flags', 'type': 'unsigned long', 'value': 'O_RDONLY|O_LARGEFILE'}, {'name': 'dev', 'type': 'dev_t', 'value': 25}, {'name': 'inode', 'type': 'unsigned long', 'value': 527}]",0,0
126.233959,384,1,101,systemd-resolve,ip-10-100-1-105,257,openat,4,15,"[{'name': 'dirfd', 'type': 'int', 'value': -100}, {'name': 'pathname', 'type': 'const char*', 'value': '/run/systemd/netif/links/5'}, {'name': 'flags', 'type': 'unsigned long', 'value': 'O_RDONLY|O_CLOEXEC'}, {'name': 'mode', 'type': 'int*', 'value': 964865707}]",0,0
126.233996,384,1,101,systemd-resolve,ip-10-100-1-105,5,fstat,2,0,"[{'name': 'fd', 'type': 'int', 'value': 15}, {'name': 'statbuf', 'type': 'struct stat*', 'value': '0x7FFFB77D84D0'}]",0,0
```

BETH Dataset

DNS (Network) logs

```
Timestamp,SourceIP,DestinationIP,DnsQuery,DnsAnswer,DnsAnswerTTL,DnsQueryNames,DnsQueryCl
ass,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.amazonaws.com,['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.amazonaws.com,['52.95.19.240'],
['17'],ssm.us-east-2.amazonaws.com,['IN'],['A'],1,0,0,ip-10-100-1-95,0,0
2021-05-16T21:38:54Z,10.100.0.2,10.100.1.95,download.docker.com,"['99.86.61.59',
'99.86.61.79', '99.86.61.24', '99.86.61.58']","['267', '45', '45', '45',
'45']",download.docker.com,['IN'],['A'],5,0,0,ip-10-100-1-95,1,0
2021-05-16T21:02:51Z,10.100.0.2,10.100.1.4,motd.ubuntu.com,"['2a05:d018:91c:3200:2846:99f
b:81b6:1e11', '2a05:d018:91c:3200:c887:2f22:290f:a7c']","['210', '210']",motd.ubuntu.com,
['IN'],['AAAA'],2,0,0,ip-10-100-1-4,0,0
2021-05-16T21:08:29Z,10.100.1.105,10.100.0.2,pool.hashvault.pro,,,pool.hashvault.pro,
['IN'],['A'],0,0,0,ip-10-100-1-105,1,1
2021-05-16T21:38:54Z,10.100.1.95,10.100.0.2,security.ubuntu.com,,,security.ubuntu.com,
['IN'],['A'],0,0,0,ip-10-100-1-95,0,0
```


Currently extending this to provide full PCAP :)