AWS Service Baseline

How should we format the data?

- Looking at past projects including the Beth Data Set:

Table 2. The description and type of each feature within the kernel-level process logs, tracking every create, clone, and kill process call. Starred features were included in the model baselines and converted as described in Appendix A.

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)

- **"**Thus, for our benchmarking, we converted several fields to binary variables based on field expertise, as described in Appendix A. Each record was manually labeled suspicious (sus) or evil to assist analysis. Logs marked suspicious indicate unusual activity or outliers in the data distribution, such as an external userId with a systemd process3 , infrequent daemon process calls (e.g., "acpid" or "accounts-daemon"), or calls to close processes that we did not observe as being started. Evil4 indicates a malicious external presence not inherent to the system, such as a bash execution call to list the computer's memory information, remove other users' ssh access, or un-tar an added file. Events marked evil are considered "out of distribution," as they are generated from a data distribution not seen during training**"**

Services that help in labeling the data coming in and out of the model:

Amazon SageMaker Data Labeling
https://aws.amazon.com/sagemaker/data-labeling/?sagemaker-data-wrangler-whats-new.sort-by=item.additionalFields.postDateTime&sagemaker-data-wrangler-whats-new.sort-order=desc

- Amazon SageMaker Ground Truth, if you want the flexibility to build and manage your own data labeling workflows and workforce, you can use SageMaker Ground Truth. SageMaker Ground Truth is a data labeling service that makes it easy to label data and gives you the option to use human annotators through Amazon Mechanical Turk, third-party vendors, or your own private workforce.
    - As part of the AWS Free Tier, you can get started with SageMaker Ground Truth for free. For the first two months after the first use of Amazon SageMaker, your first 500 objects labeled per month are free (excluding any additional costs incurred by using a labeling vendor, Amazon Mechanical Turk, or synthetic data).
- Amazon SageMaker Ground Truth Plus, with SageMaker Ground Truth Plus, you can create high-quality training datasets without having to build labeling applications or manage labeling workforces on your own. SageMaker Ground Truth Plus helps reduce data labeling costs by up to 40%. SageMaker Ground Truth Plus provides an expert workforce that is trained on ML tasks and can help meet your data security, privacy, and compliance requirements. You upload your data, and then SageMaker Ground Truth Plus creates and manages data labeling workflows and the workforce on your behalf.
    - To get your customized quote, fill out the project requirement form. Object pricing details You are charged for the number of dataset objects that are reviewed. A dataset object is defined as an atomic unit of data across all modalities.
- Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. This could include anything from conducting simple data validation and research to more subjective tasks like survey participation, content moderation, and more. MTurk enables companies to harness the collective intelligence, skills, and insights from a global workforce to streamline business processes, augment data collection and analysis, and accelerate machine learning development.

Object pricing details

You are charged for the number of dataset objects that are reviewed. A dataset object is defined as an atomic unit of data across all modalities.

Reviewed objects (images, video frames, text documents, audio files, etc.)

Number of reviewed objects per month	Price per reviewed object
Less than 50,000 objects	$0.08
50,000 to 1,000,000 objects	$0.04
Greater than 1,000,000 objects	$0.02

3D point clouds

Frame	Price per frame
Single frame	$3.00
Sequence of frames	$3.00 for 1st frame / $1.50 for 2nd frame onwards
Custom Pricing	Contact us

Labor pricing details

Built-in workflow with Amazon Mechanical Turk

If you use Amazon Mechanical Turk for labeling, you are charged per object per review instance. We recommend that you use multiple labelers per object to improve label accuracy.

Workflow	Suggested price per labeler
Image classification	$0.012
Text classification	$0.012
Named Entity Recognition (NER)	$0.024
Bounding box	$0.036
Semantic Segmentation	$0.84

Vendor

If you use a vendor, the cost per label is set by the vendor. You can see each vendor's pricing details in AWS Marketplace.