



# データセットの これまでとこれから

TrendMicro株式会社

東 結香



データセットを

どうやって探していますか？

どうやって作っていますか？

# 自然言語処理や画像分野だと

papers with code

- SOTAも論文も探しやすい

<https://paperswithcode.com/>

The screenshot shows the Datasets website interface. At the top, there is a search bar and navigation links for "Browse State-of-the-Art", "Datasets", "Methods", and "More". The main heading is "Datasets" with a subtitle "8,345 machine learning datasets". Below this, there is a notification to "Share your dataset with the ML community!". The main content area displays "8345 dataset results" and lists several popular datasets with their descriptions and associated papers/benchmarks:

- CIFAR-10**: The CIFAR-10 dataset (Canadian Institute for Advanced Research, 10 classes) is a subset of the Tiny Images dataset and consists of 60000 32x32 color images. The images are labelled with... 11,692 PAPERS • 71 BENCHMARKS
- ImageNet**: The ImageNet dataset contains 14,197,122 annotated images according to the WordNet hierarchy. Since 2010 the dataset is used in the ImageNet Large Scale Visual Recognition Challenge... 11,218 PAPERS • 106 BENCHMARKS
- COCO (Microsoft Common Objects in Context)**: The MS COCO (Microsoft Common Objects in Context) dataset is a large-scale object detection, segmentation, key-point detection, and captioning dataset. The dataset consists of 328K... 8,106 PAPERS • 82 BENCHMARKS
- MNIST**: The MNIST database (Modified National Institute of Standards and Technology database) is a large collection of handwritten digits. It has a training set of 60,000 examples, and a test set of... 6,279 PAPERS • 50 BENCHMARKS
- CIFAR-100**: The CIFAR-100 dataset (Canadian Institute for Advanced Research, 100 classes) is a subset of the Tiny Images dataset and consists of 60000 32x32 color images. The 100 classes in the... 6,035 PAPERS • 44 BENCHMARKS
- Cityscapes**: Cityscapes is a large-scale database which focuses on semantic understanding of urban street scenes. It provides semantic, instance-wise, and dense pixel annotations for 30 classes grouped... 2,810 PAPERS • 41 BENCHMARKS
- KITTI**: KITTI (Karlsruhe Institute of Technology and Toyota Technological Institute) is one of the most popular datasets for use in mobile robotics and autonomous driving. It consists of hours of... 2,702 PAPERS • 122 BENCHMARKS
- SVHN (Street View House Numbers)**

On the left side, there are filters for "Filter by Modality" (Images: 2400, Texts: 2233, Videos: 780, Audio: 481, Medical: 297, 3D: 263, Graphs: 206), "Filter by Task" (Question Answering: 341, Semantic Segmentation: 272, Object Detection: 236, Image Classification: 221, Speech Recognition: 197, Language Modelling: 142, Text Generation: 130), and "Filter by Language" (English: 2344).

# セキュリティ分野だと？

- Kaggle等のコンペティション
- 大学・研究機関
- Github
- 論文から読み解く

The screenshot shows the Kaggle Datasets page with a search for 'malware'. The results list several datasets:

- Microsoft Malware Classification Challenge**: The Microsoft Malware Classification Challenge of a huge dataset of nearly 0.5 terabytes, containing 27 papers and 1 benchmark.
- AndroZoo**: ...It currently contains 15,097,876 different APKs, each of which has been (or will be) analysed by tens of different AntiVirus products to know which applications are detected as Malware. 8 papers, no benchmarks yet.
- SOREL-20M (Sophos/ReversingLabs-20 Million)**: ...is a large-scale dataset consisting of nearly 20 million files with pre-extracted features and metadata, high-quality labels derived from multiple sources, information about vendor detection. 8 papers, no benchmarks yet.
- Maling**: The Maling Dataset contains 9,339 malware byteplot images from 25 different families. 6 papers, 1 benchmark.
- MaleX**: MaleX is a curated dataset of malware and benign Windows executable samples for malware.

The screenshot shows the Kaggle Datasets page with a search for 'malware'. The results list several datasets:

- Malware Timestamps**: Chris Deotte - Updated 4 years ago. Usability 6.3 - 426 kB.
- Malware Executable Detection**: PIYUSH RUMAO - Updated 3 years ago. Usability 9.7 - 1 File (CSV) - 28 kB.
- Android Malware Detection**: Cyber Cop - Updated 5 months ago. Usability 9.4 - 1 File (CSV) - 47 MB.

The screenshot shows the Canadian Institute for Cybersecurity (CIC) Datasets page. The page lists available datasets:

- Ground-truth dataset real/fake
- IoT dataset
- Dark web
- DNS datasets
- IDS datasets
- ISCX datasets, 2009-2016
- Malware
- Operational technology

他の分野のデータセットを見て思うこと

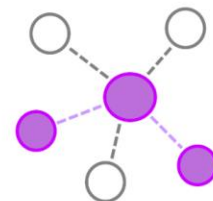
# セキュリティ分野におけるタスクの概念

- 自然言語処理では様々なタスクが用意されている
  - 常識があるか？
    - 小学生レベルの問題が解ける
  - 意味を理解しているか？
    - 代名詞が何を指しているか判断できる
  - 文脈を理解しているか？
    - 物語の結末を予測できる

▶ セキュリティ分野でのタスクは？

# 見つけたいものの毎のデータセット

- 検出したい攻撃手法ごとに生成・整理されたデータセット



Security Datasets

Search this book...

## HOW-TO

Create Datasets

Consume Datasets

## ATOMIC DATASETS

aws

initial\_access

persistence

privilege\_escalation

defense\_evasion

collection

discovery

linux

windows

## COMPOUND DATASETS

Golden SAML AD FS Mail Access

Log4Shell



Contents

Introduction

Goals

Projects Using Security Datasets

Authors

Contributing

License: MIT

## Introduction

launch binder License MIT Follow Open Threat Research Community Open Source

The **Security Datasets** project is an open-source initiative that contributes malicious and benign datasets, from different platforms, to the infosec community to expedite data analysis and threat research.

## Goals

- Provide open portable datasets to expedite the development of data analytics.
- Facilitate and expedite adversary techniques simulation.
- Allow security analysts around the world to test their skills with real data.
- Improve the testing and validation of detection analytics in an easier, practical, modular and more affordable way.
- Enable data scientists to have labeled and unlabeled data for initial research and features development.
- Help the community map datasets to other open source projects such as Sigma, Atomic Red Team, Threat Hunter Playbook (Jupyter Notebooks) and MITRE ATT&CK.
- Provide datasets for other social/community events such as Capture The Flags (CTFs) or hackathons to encourage collaboration.

## Projects Using Security Datasets

- [ThreatHunter-Playbook](#)

## Authors

- Roberto Rodriguez @Cyb3rWard0g
- Jose Luis Rodriguez @Cyb3rPandaH

## Contributing

Help us build the largest library of datasets for the InfoSec community!. Learn more about how you could do it [here!](#)

## License: MIT

[Security Datasets's MIT License](#)

<https://securitydatasets.com/>

Public







# 参考URL

- papers with code

<https://paperswithcode.com/>

- Canadian Institute for Cybersecurity datasets

<https://www.unb.ca/cic/datasets/index.html>

- Kaggle Dataset

<https://www.kaggle.com/datasets>

- Security Datasets project

<https://securitydatasets.com/>