データセットの
これまでとこれから

TrendMicro株式会社

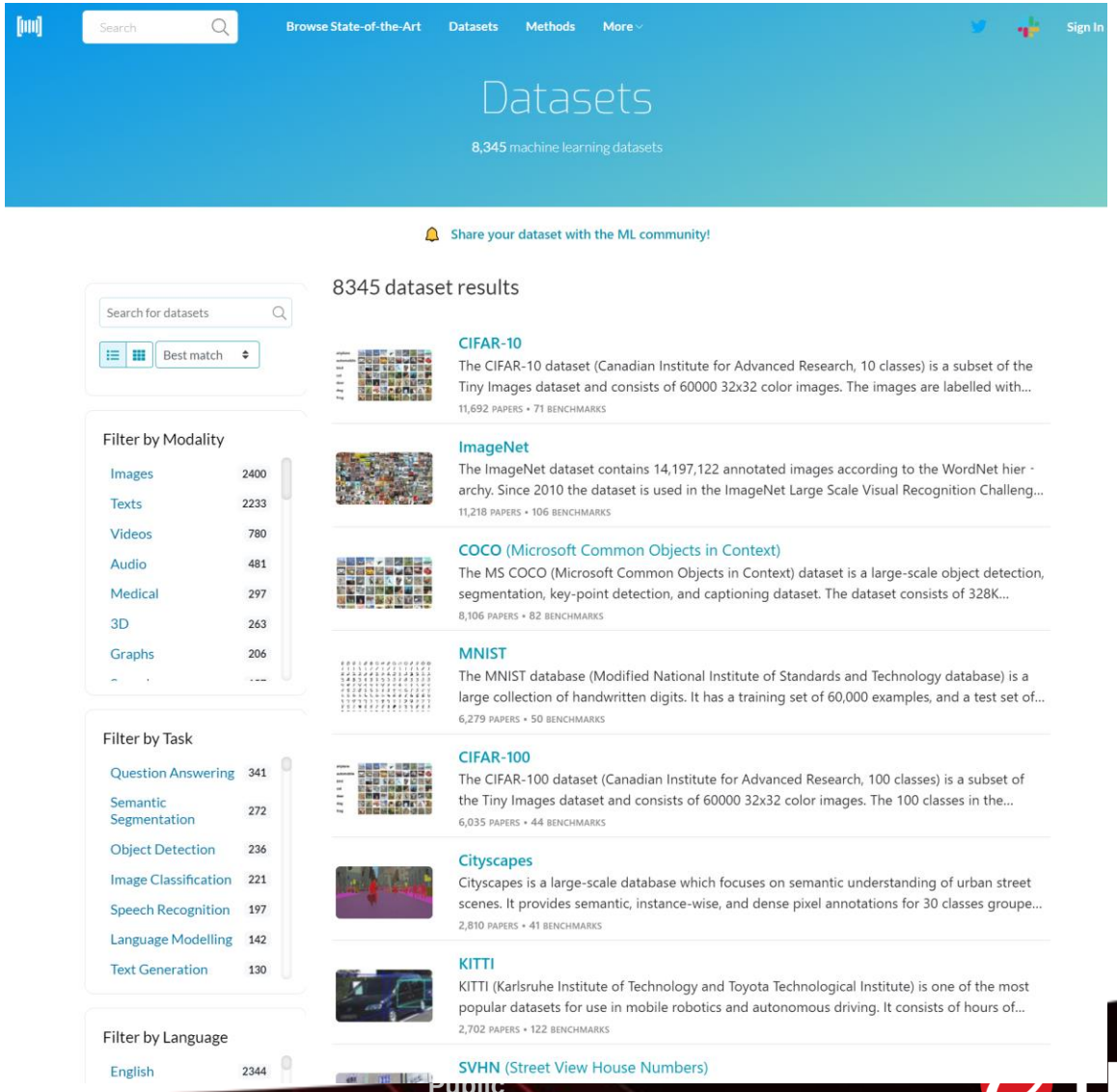
東　結香

データセットを
どうやって探してますか？
どうやって作ってますか？

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/

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

他の分野のデータセットを見て思うこと

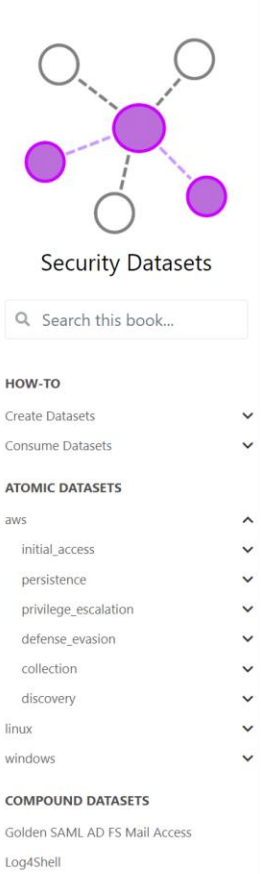
セキュリティ分野におけるタスクの概念

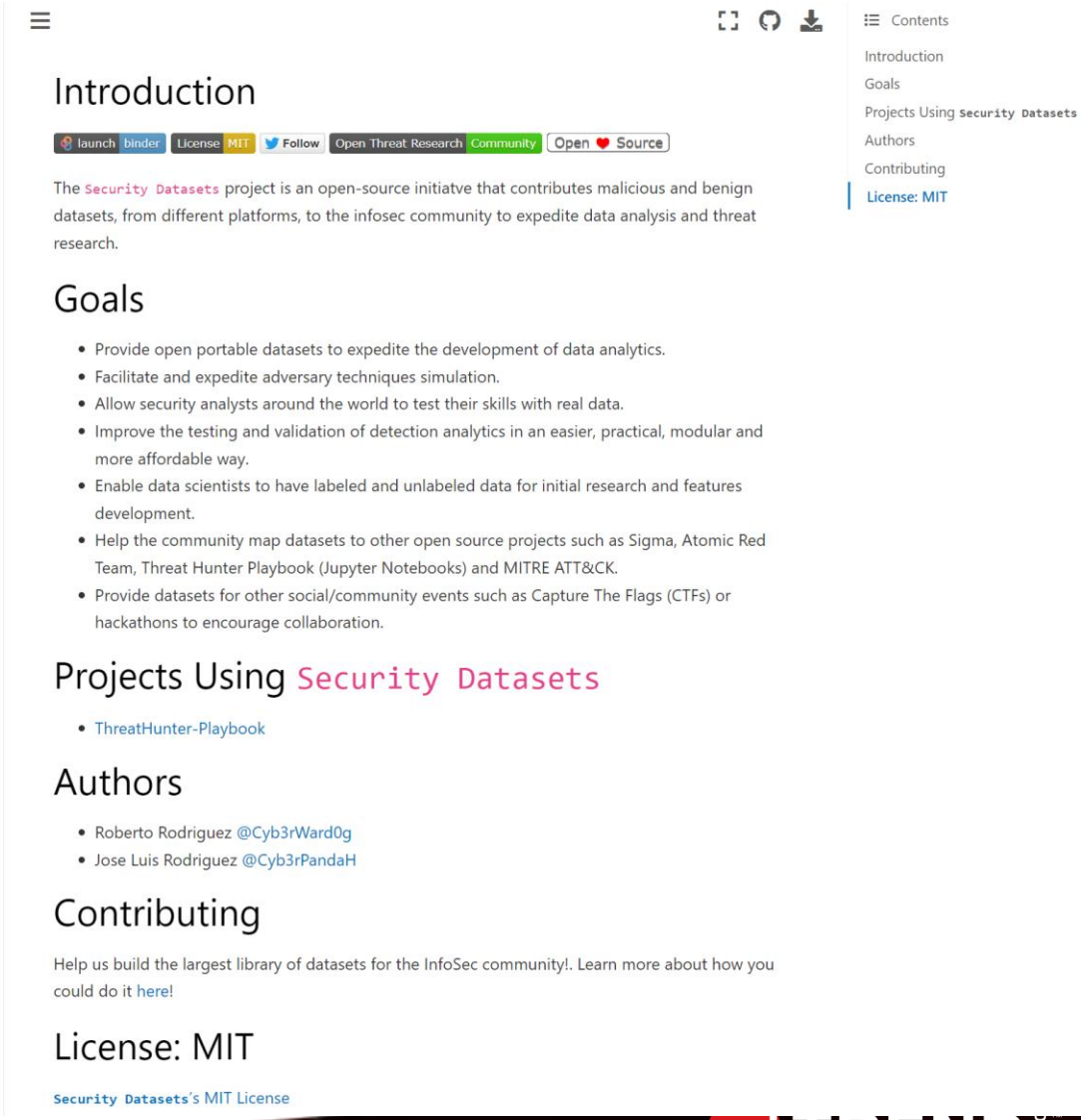
- 自然言語処理では様々なタスクが用意されている
  - 常識があるか？
    - 小学生レベルの問題が解ける
  - 意味を理解しているか？
    - 代名詞が何を指しているか判断できる
  - 文脈を理解しているか？
    - 物語の結末を予測できる

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





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/

