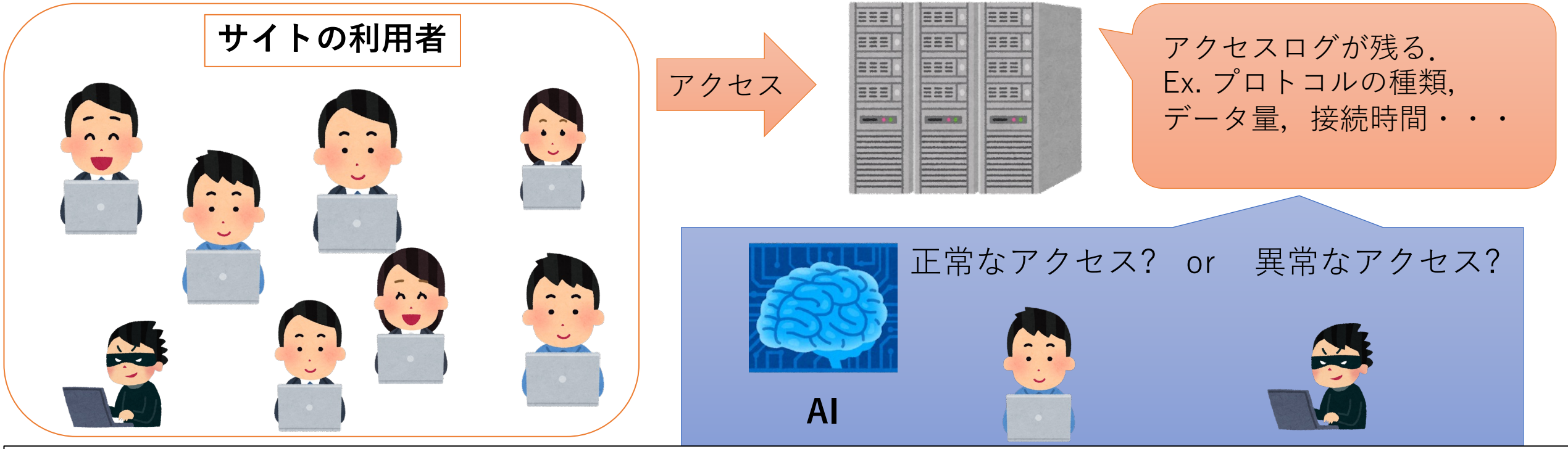
研究テーマ

ネットワーク侵入検知データセットにおける半教師あり学習を用いた学習データの少ないデータセットでの機械学習精度向上

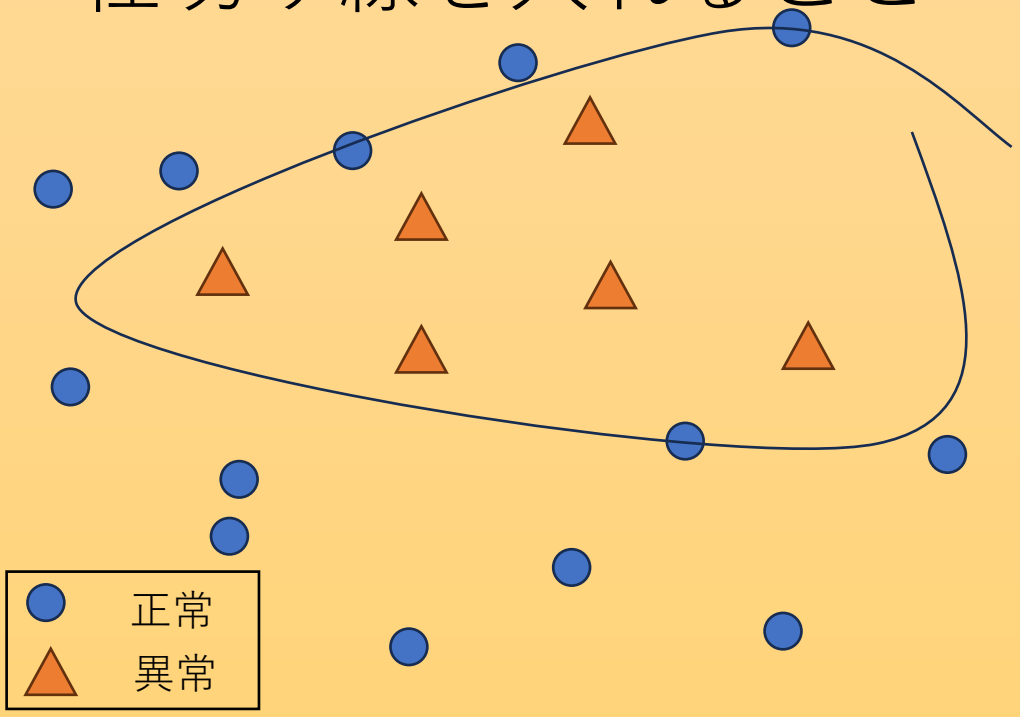


キーワード：機械学習, AI, ニューラルネットワーク, オートエンコーダ, 勾配ブースティング決定木, 勾配降下法, アンサンブル学習, 決定木, 特徴量エンジニアリング, オーバーサンプリング, 不均衡データセット, TensorFlow(Keras), LightGBM, SMOTE

機械学習とは

学習(分類問題の場合)

＝機械学習モデルを作る
＝仕切り線を入れること



機械学習モデルとは？

データx(x1, x2, x3…)

機械学習モデル

y=0(正常)

y=1(異常)

重要!

データxを入れたら，予測yを返す
関数y=f(x)を作ること

学習の流れ

境界線の関数y=f(x)のパラメータ(a1, a2…)を変えていくよー
Ex. f(x) = a1x1+a2x2+…

学習前

正常と異常を分類できていない

正常
異常

学習後

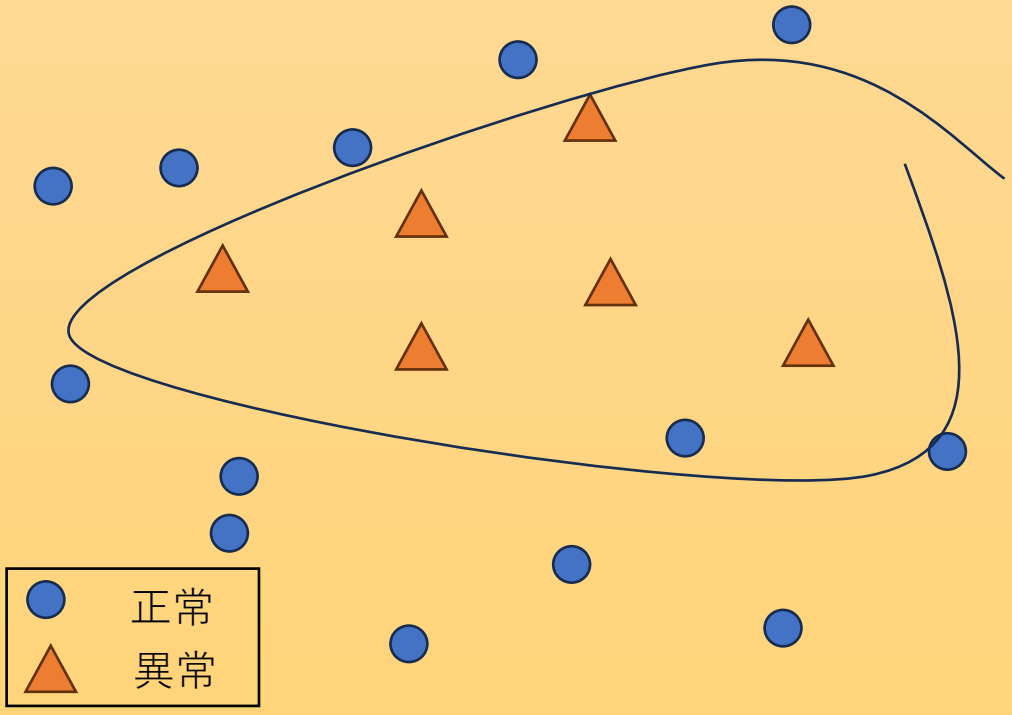
ほぼ分類できている

異常か正常かまだ分からないデータ

正常
異常

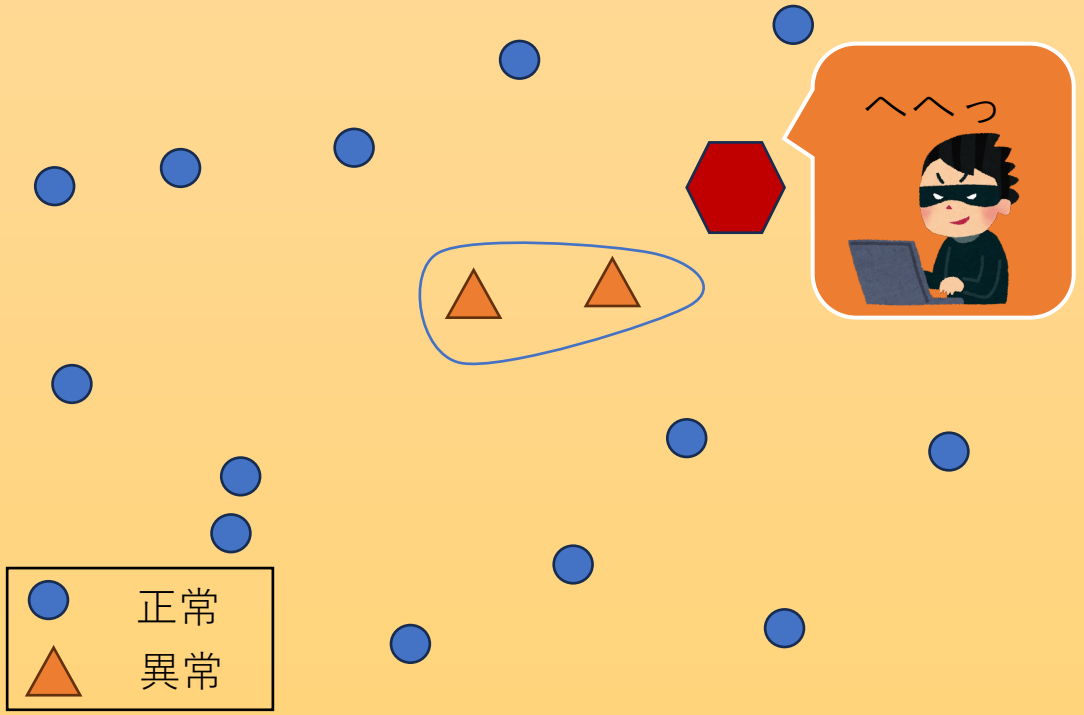
教師データが少ないとどうなるか

学習データが多い場合



学習データが少ない場合

どこに境界線引いたら良いか
分からないよ～



データが少なくても精度を上げるために

方法1: 不正なデータの特徴を掴んだデータを作る

※例

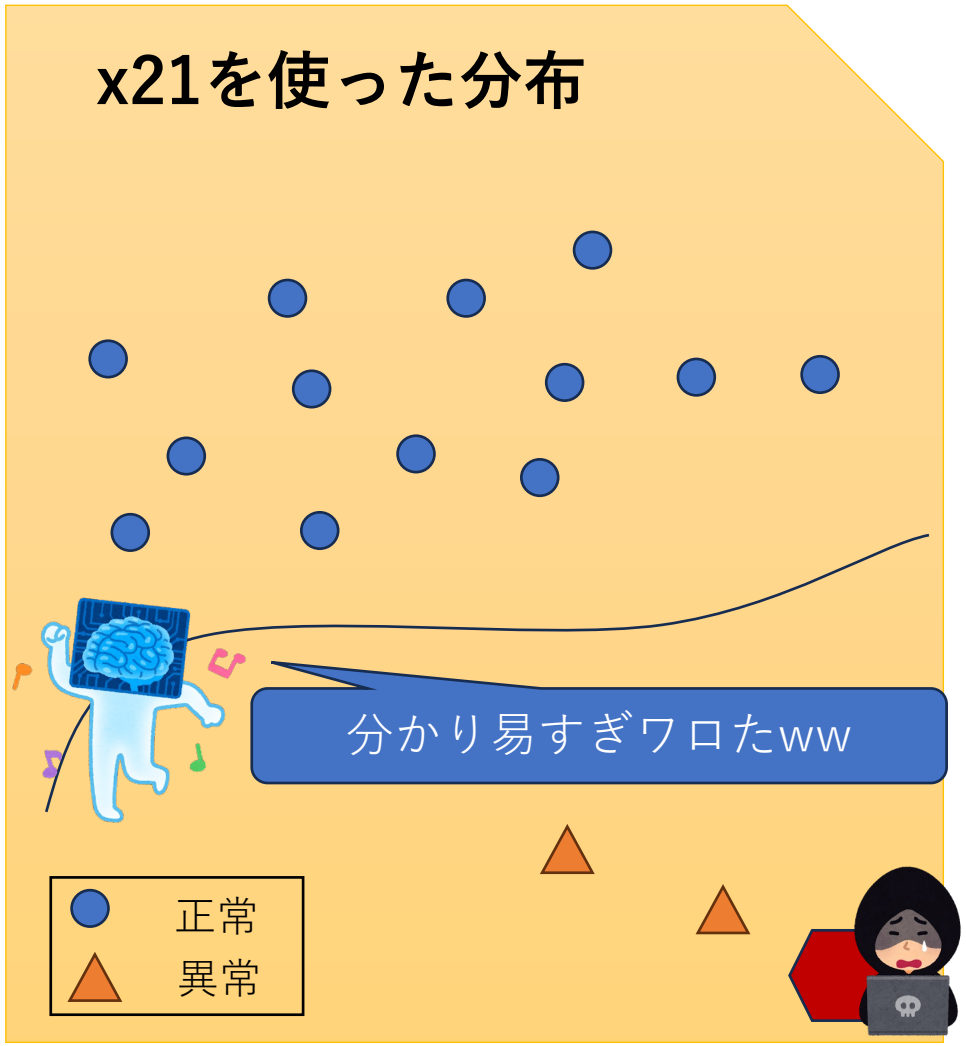
データx（x1,x2,x3…,x20）

不正なデータは，x14と，x18
の値が大きい傾向がある．

新しい特徴量x21を作成

$$x21 = 100 \times x14 + 100 \times x18$$

キーワード：特徴量エンジニアリング，
オートエンコーダによる特徴量抽出



データが少なくても精度を上げるために

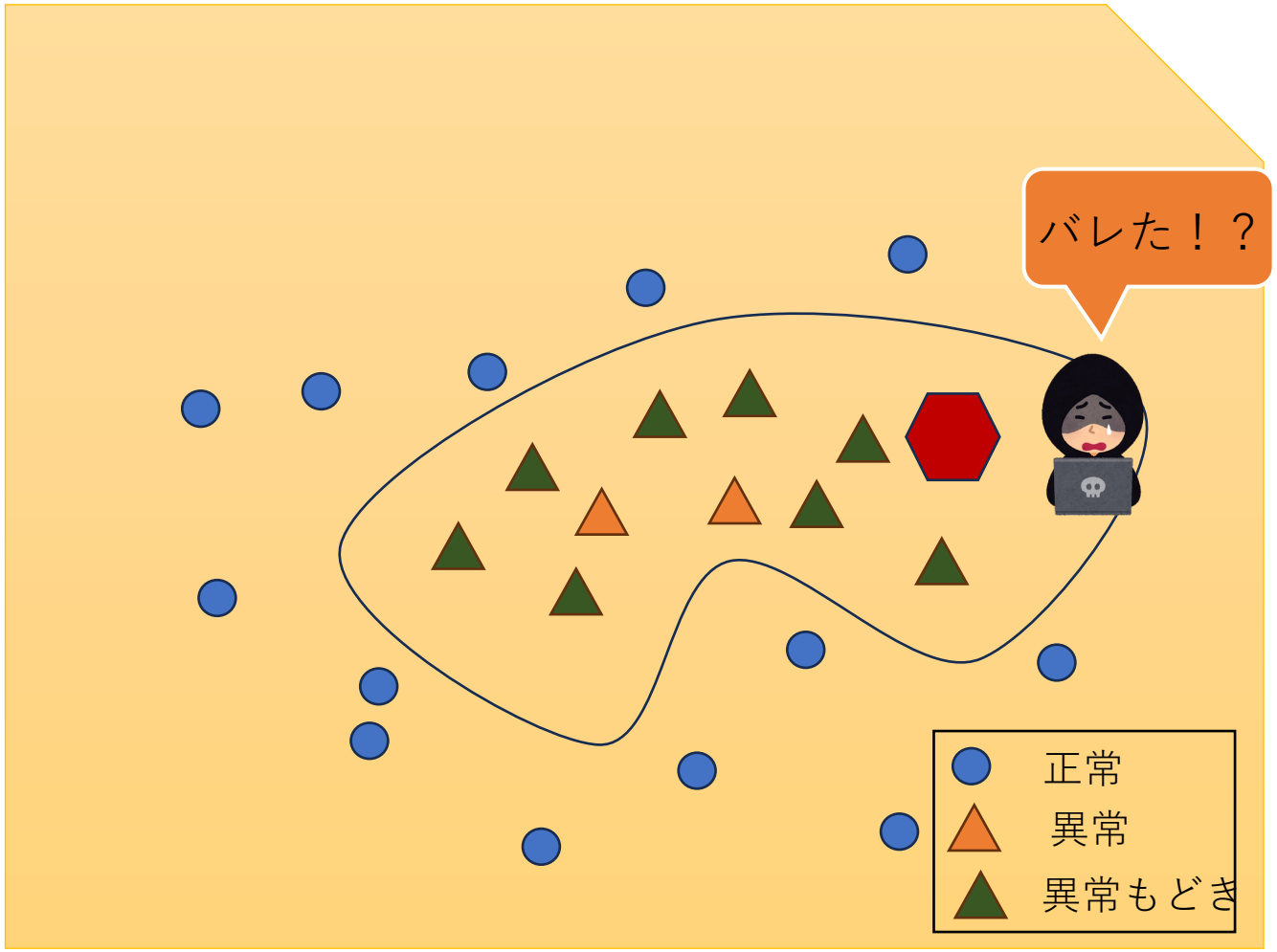
方法2: 不正なデータもどきを沢山生成する.

異常データに近いデータ沢山作るよ

ここら辺に線を引けばいいかな～

精度向上

キーワード：オーバーサンプリング，SMOTE



半教師あり学習とは？

教師あり学習

正常と異常の境界線を引くこと

問題

学習データが少ないと精度が低い
（上手く境界線が引けない）

教師なし学習

データの特徴ごとにまとめること

問題

正常と異常を分類できない

半教師あり学習

ポイント

正常と異常の特徴を捉えて，学習データが少なくても
精度良く境界線を引くことができる！かも？

実際，研究で書くプログラムの例

```
from keras.layers import Dense
新規 *
def main():
    model = keras.Sequential([
        Dense(units=19, activation='relu', input_dim=38, name='encoder1'),
        Dense(units=10, activation='relu', name='encoder2'),
        Dense(units=19, activation='relu'),
        Dense(units=38, activation='relu'),
    ])
    model.compile(optimizer='adam', loss='mean_squared_error', metrics=['accuracy'])
    model.fit(x_train, x_train,
              epochs=1,  # データセットを使って学習する回数
              batch_size=32,
              validation_data=(x_train, x_train),  # 評価用データ（検証データ）の指定
              )

    x_pred = model.predict(x_test) # モデルを使って実際に，予測


if __name__ == '__main__':
    main()
```

モデルの設定を定義

学習

未知のデータに対して予測