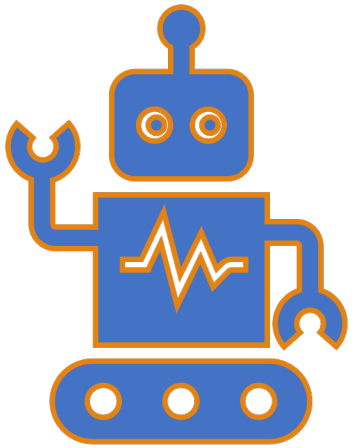
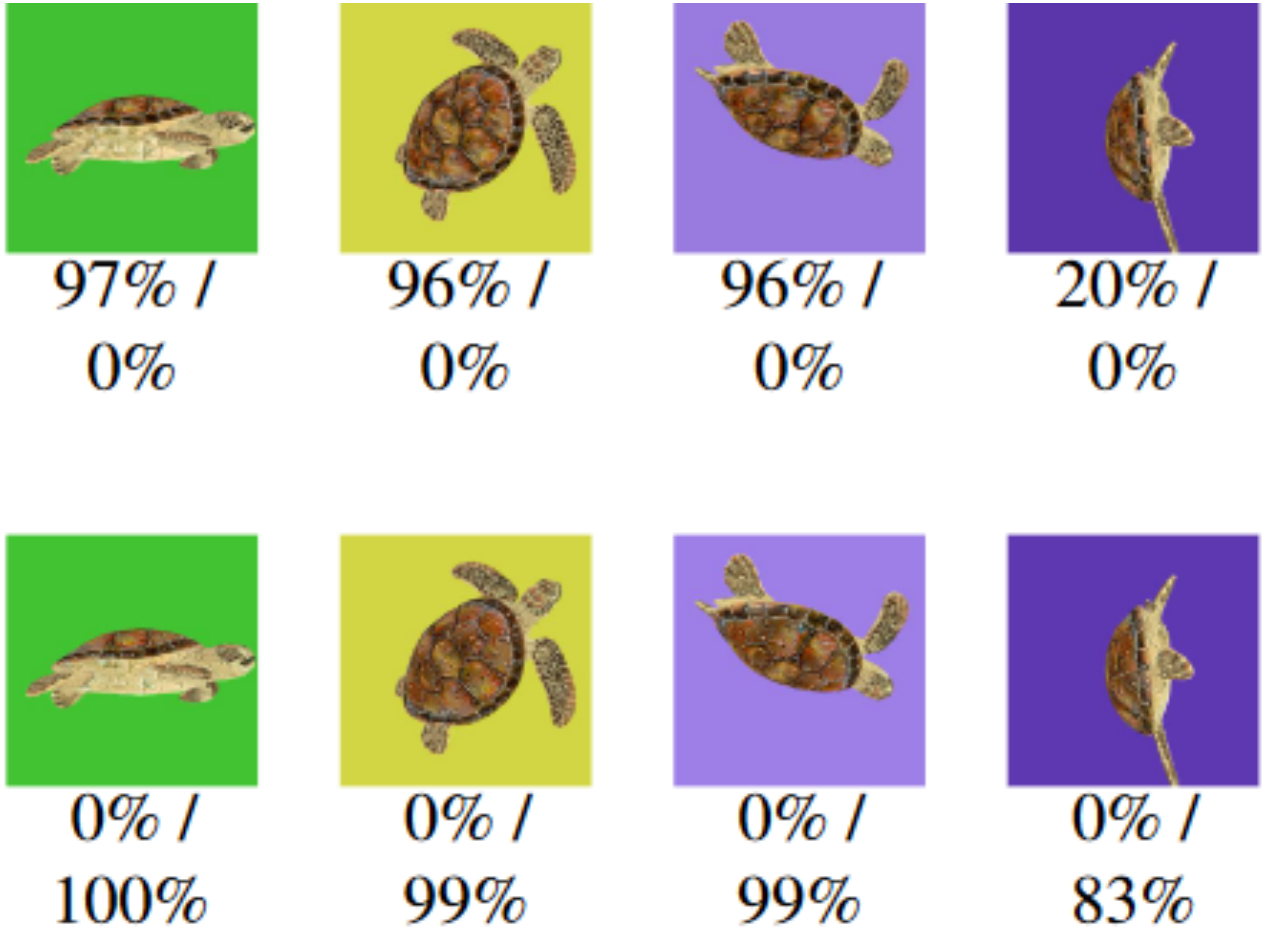


実は簡単!?
AIを攻撃してみよう

KOTOKAZE

この違いは？というお話です



亀を
ジグソーパズル
と分類している

出典: https://arxiv.org/pdf/1707.07397.pdf

そもそもAIって？

学習

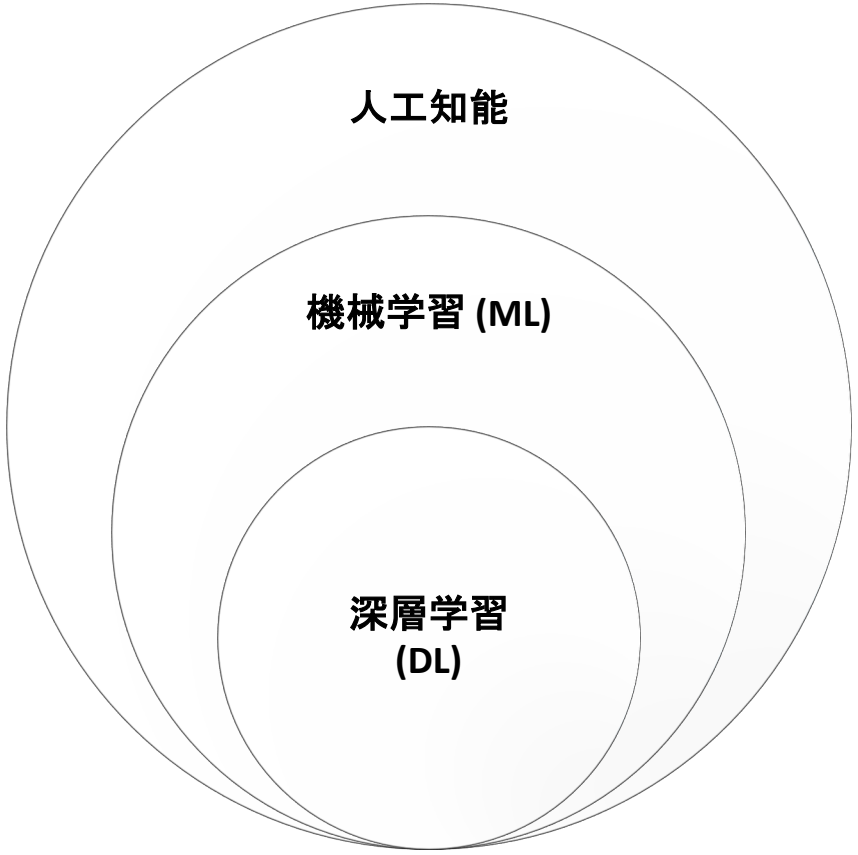
推論

認識

記憶

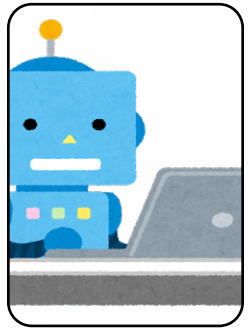
の４つをコンピュータ上で行うもの

AI の分類



ML

- 要設定



DL

- 自動チューニング

機械学習 (ML) の分類

機械学習

教師あり

- 分類
- 回帰

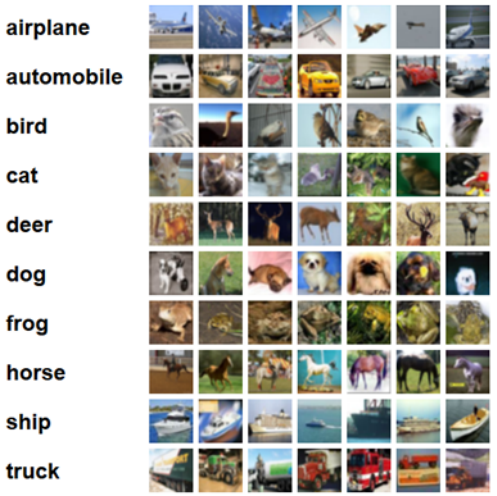
教師なし

強化学習

教師あり学習

分類

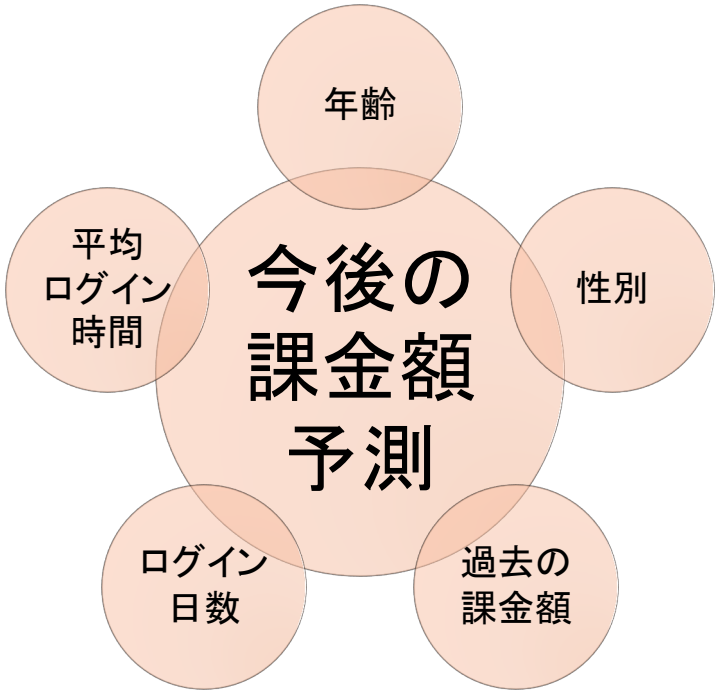
学習データ



判定したいもの

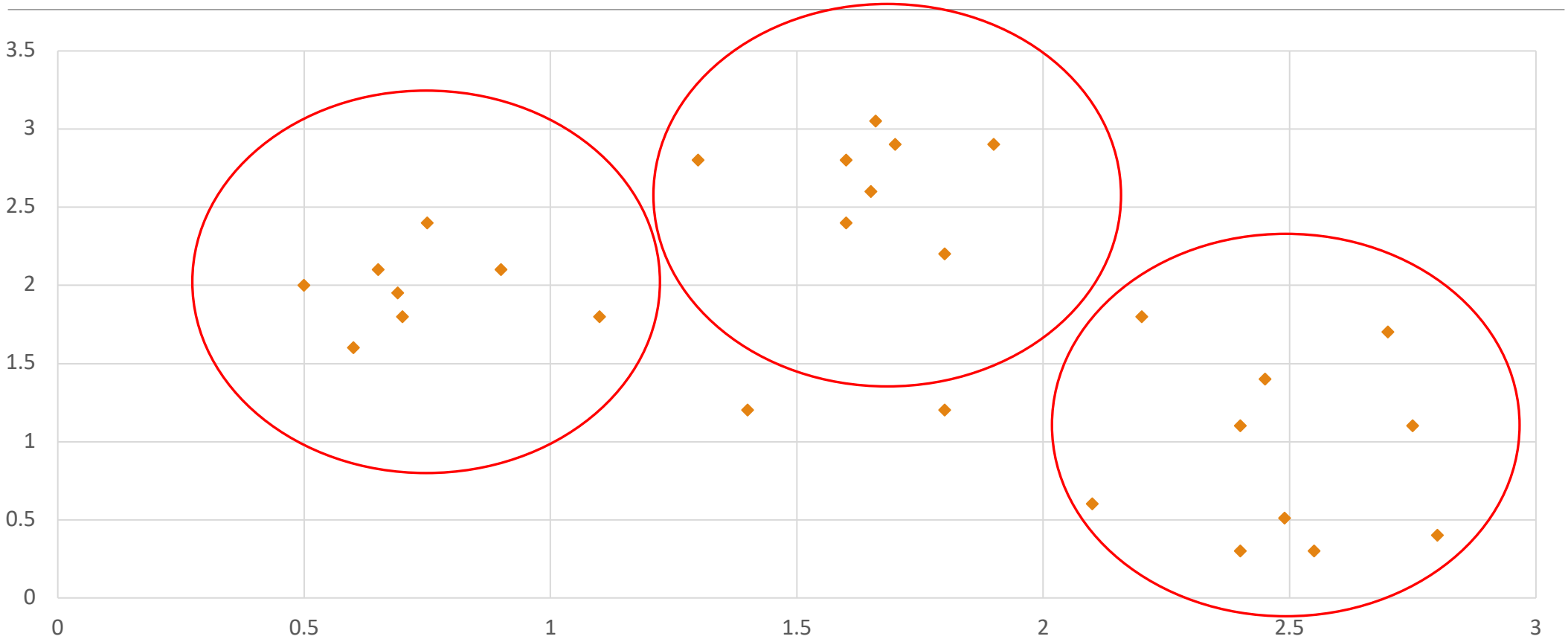
ラベリング済み画像

回帰

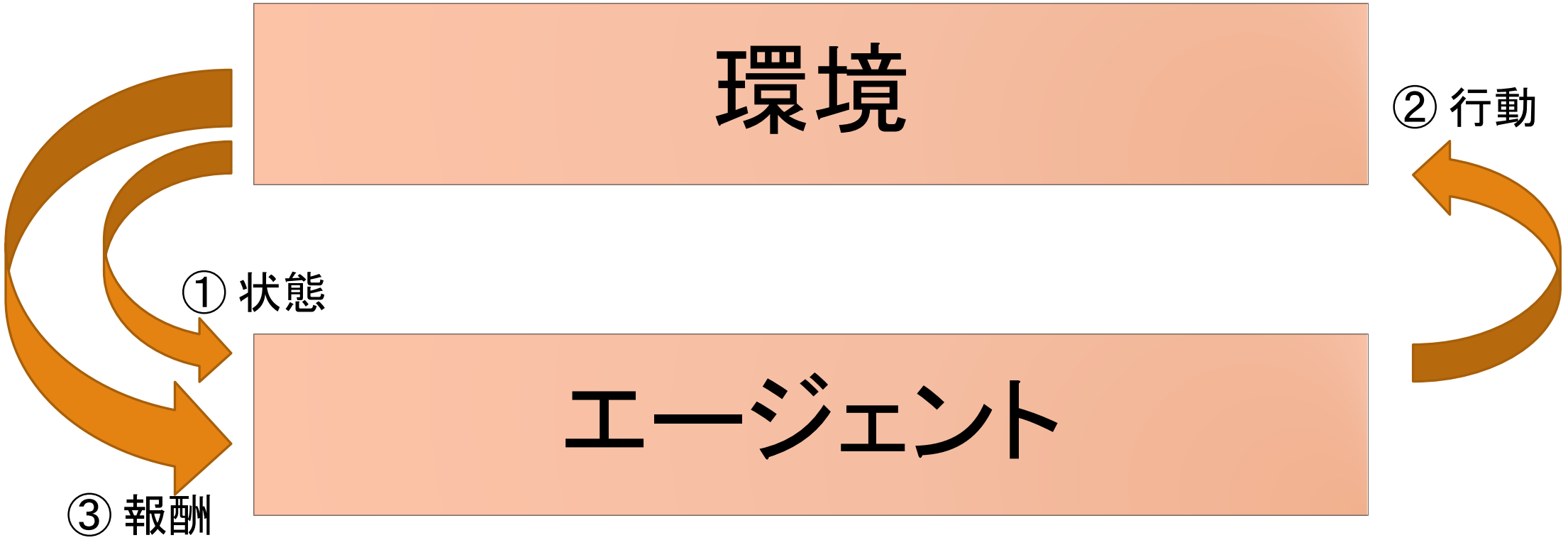


出典: http://www.cs.toronto.edu/~kriz/cifar.html

教師なし学習: データの自動分類



強化学習

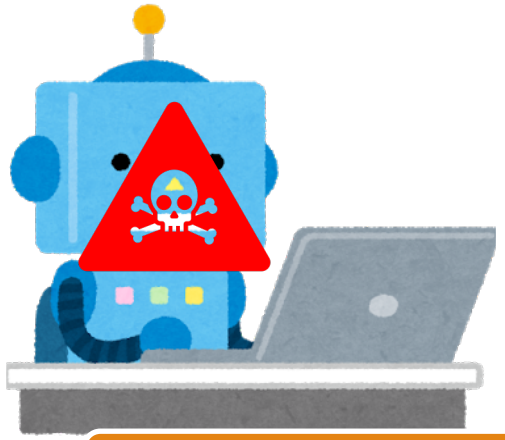


モデルへの攻撃の余地

学習データ汚染

学習モデル汚染

クエリ

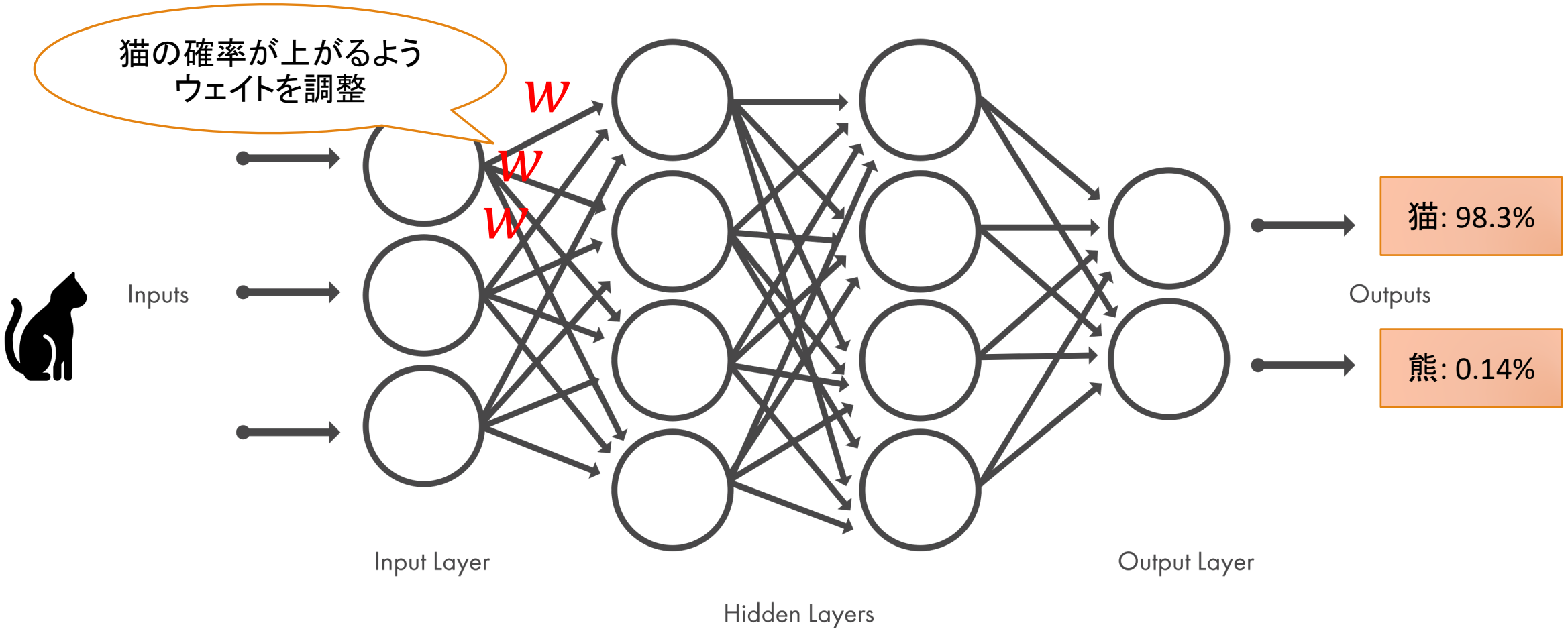


学習データの
収集・作成

モデルの学習

推論

画像分類器 (CNN) の学習の仕組み



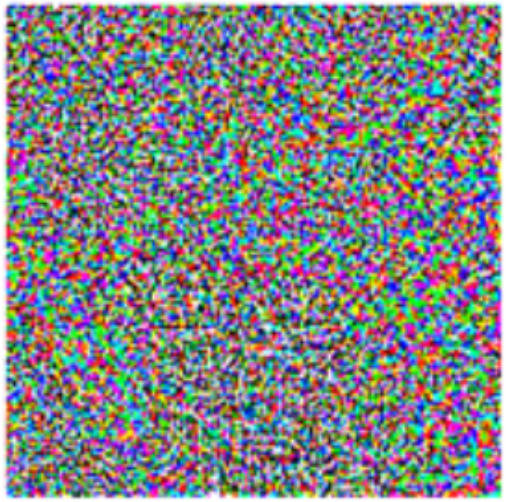
出典: https://pytorch.org/tutorials/beginner/blitz/neural_networks_tutorial.html

敵対的サンプル (Adversarial Examples)

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

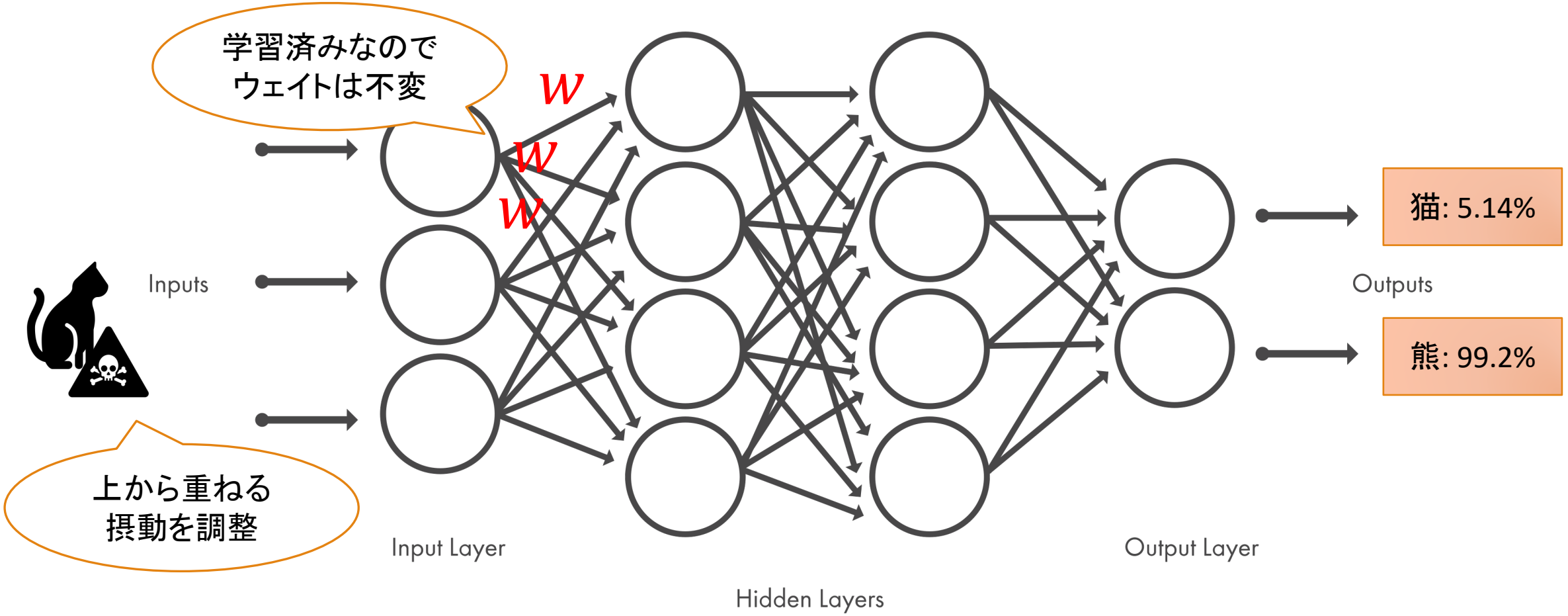
$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

出典: http://www.iro.umontreal.ca/~memisevr/dlss2015/goodfellow_adv.pdf

AEs 攻撃の仕組み



出典: https://pytorch.org/tutorials/beginner/blitz/neural_networks_tutorial.html