AIを悪意のある攻撃から守れ
～敵対的攻撃の検出手法に関する研究～

広島市立大学大学院情報科学研究科知能工学専攻
データ科学講座・データ工学グループ・敵対的攻撃検出チーム

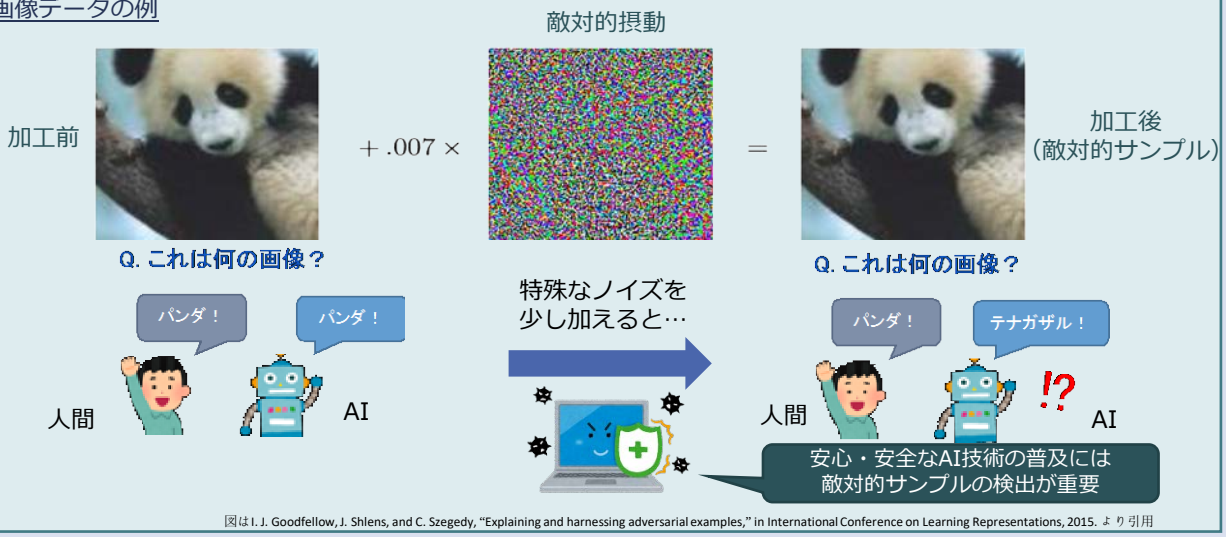
深層学習の実社会への応用が期待されるなか、人工知能（AI）を騙す手法への対策が重要となっています。
人工知能を騙す手法のひとつである敵対的サンプルを用いた敵的攻撃の検出手法を研究しています。

敵対的サンプル

敵対的サンプル （AE: Adversarial Example）

➡ 人間には認識できないような微小なノイズでAI（深層学習モデル）の誤認識を引き起こすデータ

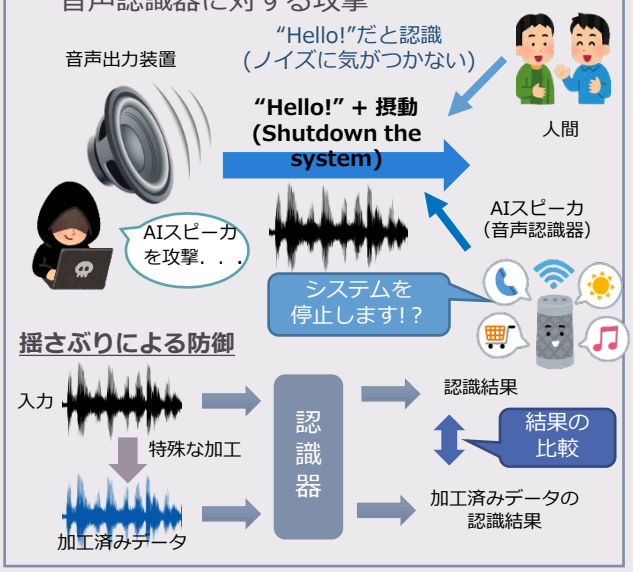
画像データの例



図はI. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in International Conference on Learning Representations, 2015. より引用

音声データ

AAE:　Audio Adversarial Example
音声認識器に対する攻撃

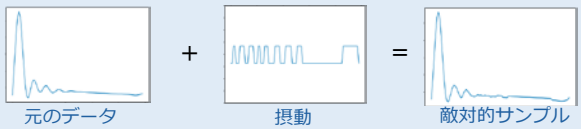


揺さぶりによる防御

時系列データ

時系列データ分類器に対する敵対的サンプル

心電図検査，地震検知，食品検査などに使用



検出モデルによる防御

敵対的サンプルかどうかを判定するAI（=検出モデル）を使用