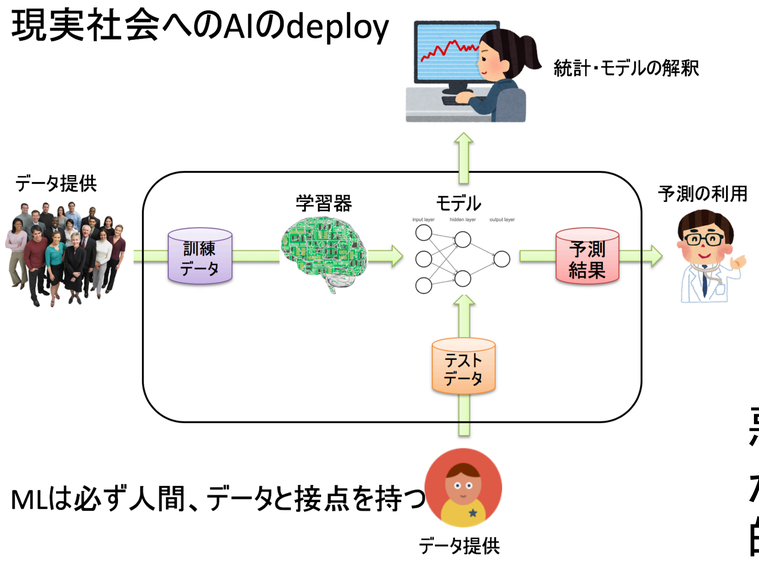
AIセキュリティ・プライバシーチーム

AI Security & Privacy Team
佐久間 淳
Jun Sakuma

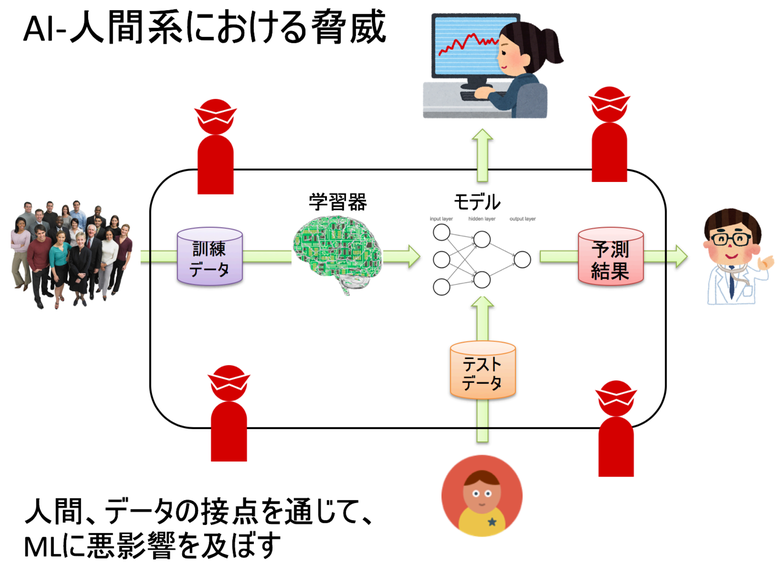
1. 背景

現実社会へのAIのdeploy



悪意を持つ者が紛れる敵対的環境では…

AI-人間系における脅威

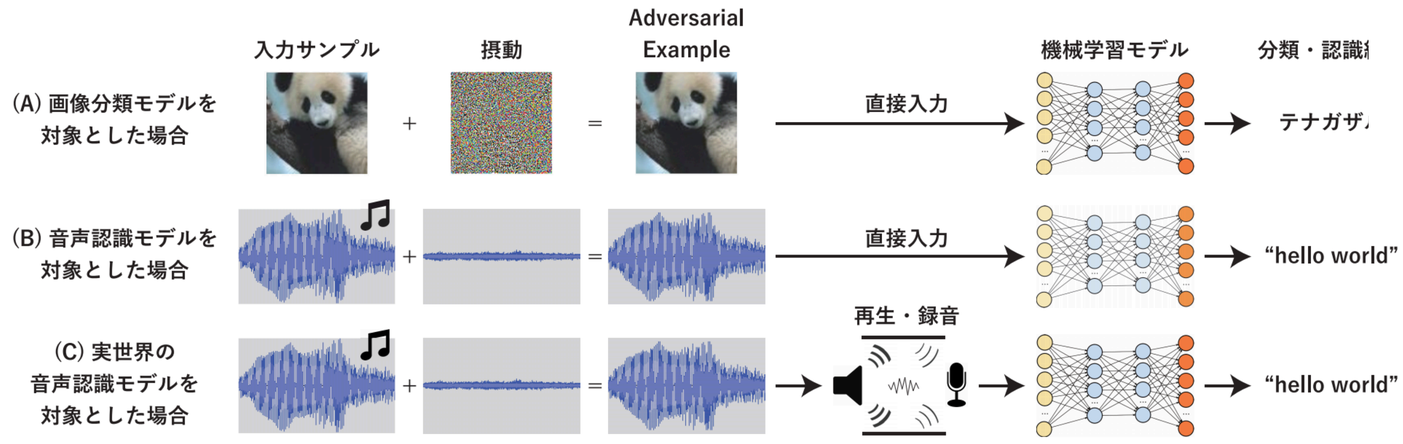


人間、データの接点を通じて、MLに悪影響を及ぼす

2. ミッション

- 敵対的環境において生じうるAIへの悪影響を明らかに
- 敵対的環境でAIが正しく安定に動作することを保証
- 「正しく安定」=セキュリティ、プライバシー、公正

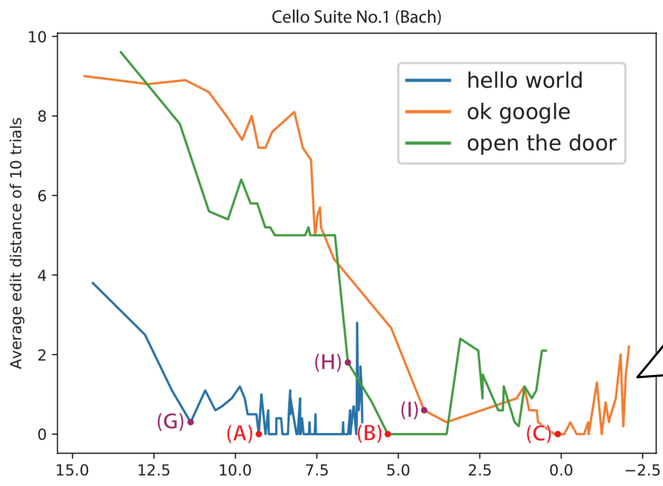
成果1 音声認識AIへのover-the-air攻撃[IJCAI'19]



- 音声AIへの攻撃=人間には無害に聞こえる (e.g, 音楽)が、音声認識機には特定のコマンド(e.g., Unlock door)と認識される音声ファイル
- Over-the-air攻撃=スピーカーから発生し、マイクで取得した音声による攻撃
- 放送、スピーカー、インターネット動画などを通じて多数のデバイスに影響を及ぼす

	Input sample	Target phrase	SNR
(A)	Bach	hello world	9.3dB
(B)	Bach	open the door	5.3dB
(C)	Bach	ok google	0.2dB
(D)	Owl City	hello world	11.8dB
(E)	Owl City	open the door	13.4dB
(F)	Owl City	ok google	2.6dB

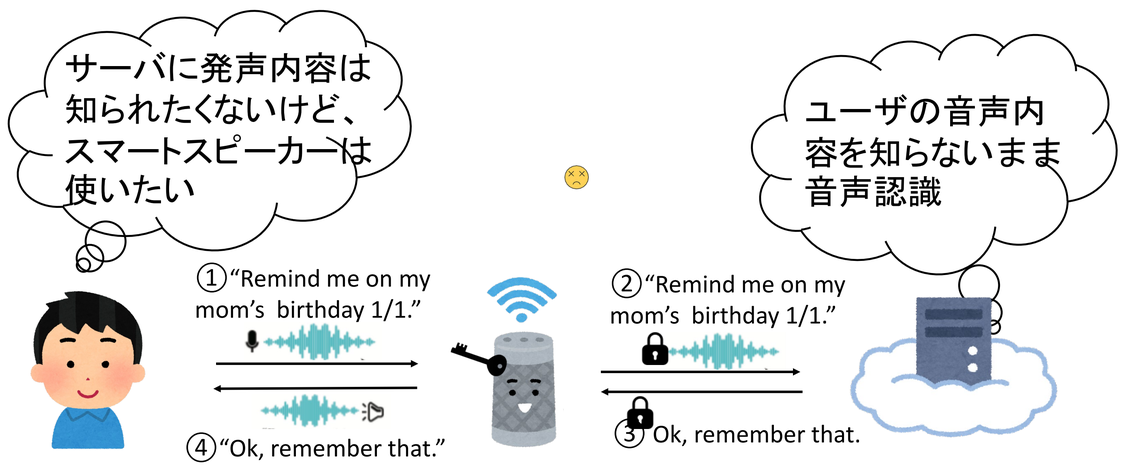
100%攻撃に成功する敵対的サンプルのSNR



SNR vs 認識された文字列と正解の編集距離

比較的小さいノイズで攻撃達成

成果2 リアルタイム音声認識秘密計算に向けて　RNN評価に必要な行列乗算秘密計算の高速化



リアルタイムプライバシ保護音声認識
=n秒の音声認識をn秒の秘密計算で完了

アルゴリズム

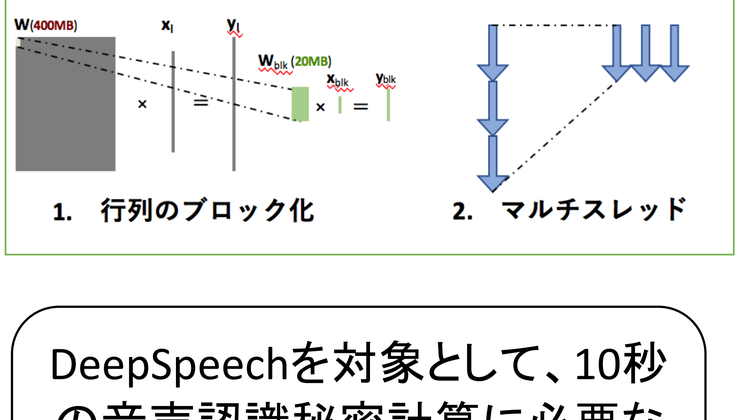
$y_1=Wx_1, y_2=Wx_2, \ldots, v_L=Wx_L$に対して、

既存法
$\langle U_1, v_1, z_1\rangle$
$\langle U_2, v_2, z_2\rangle$
…
$\langle U_L, v_L, z_L\rangle$

Uの生成がL-1回の節約

提案法
$\langle U, v_1, z_1\rangle$
$\langle U, v_2, z_2\rangle$
…
$\langle U, v_L, z_L\rangle$

実装の最適化



音声認識のMT生成時間[秒]

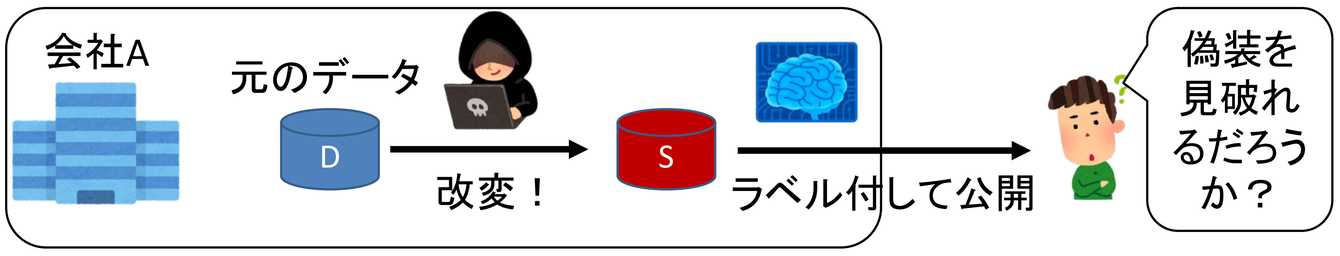
音声の長さ	既存法	提案法	最適化した提案法
1秒	206.11	71.11	1.23
5秒	1029.51	347.81	3.02
10秒	2057.69	708.81	5.74

DeepSpeechを対象として、10秒の音声認識秘密計算に必要な行列乗算を10秒以下の計算で達成

成果3 機械学習モデルの公平性の偽装 [AAAI'20]

- **公平性**: 機械学習モデルが人種や性別に偏らないようにしたい
- **公平性の宣伝**: 企業が提供する機械学習ベースのサービスが公平であると宣伝するために**公平性の証拠を公開**
- **公平性の偽装**: モデルが不公平でも公平に見えるように**証拠を改変**

シナリオ: 公平性の証拠としてラベル付したベンチマークデータを公開



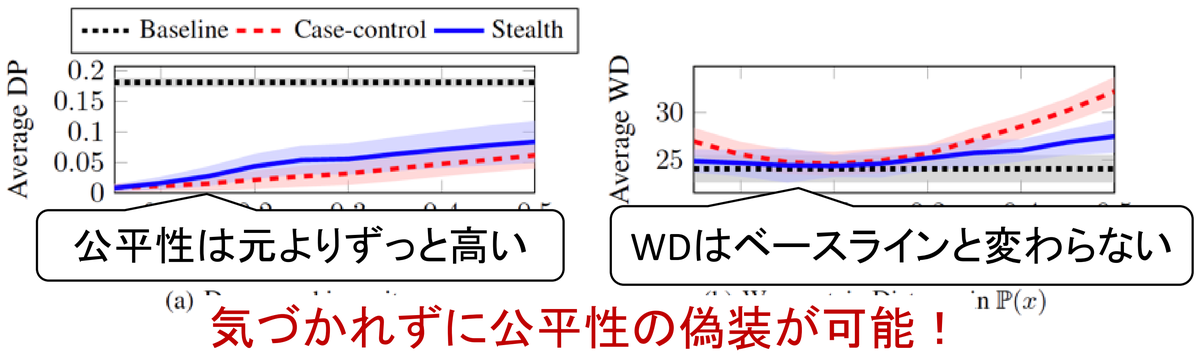
攻撃アルゴリズム

$$\min W(S,D) \text{ sub to } C(S)=C_T$$

Wasserstein距離最小化
SとDの見分けがつかない

分割表を指定の形に
Sが公平に見えるように

Adultデータセットでの実験結果



公平性は元よりずっと高い

WDはベースラインと変わらない

気づかれずに公平性の偽装が可能！