Vasisht Duddu⁺, Sebastian Szyller⁺, N. Asokan^{+,†}

SoK: Unintended Interactions among Machine Learning Defenses and Risks

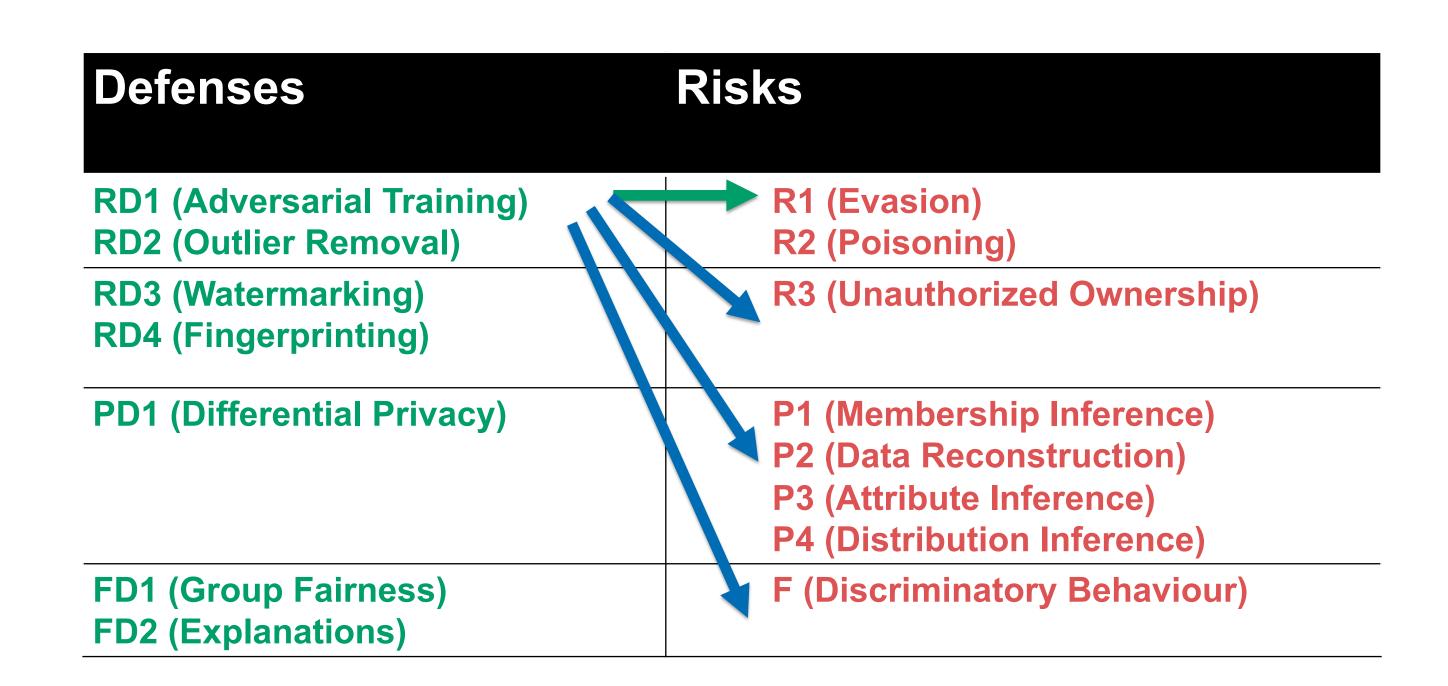
Motivation

- ML models susceptible to different risks to security, privacy, and fairness
- Defenses designed against specific risks

May increase or decrease unrelated risks
Unintended interactions

No systematic framework to understand them

Unintended interactions



Conjectured causes: overfitting, memorization

Framework: Underlying causes

Overfitting: Difference in train and test accuracy

Factors: Trainset size (D1); Model capacity (M1)

Memorization: Difference in model prediction on data record w/ and w/o it in training dataset

Influencing factors:

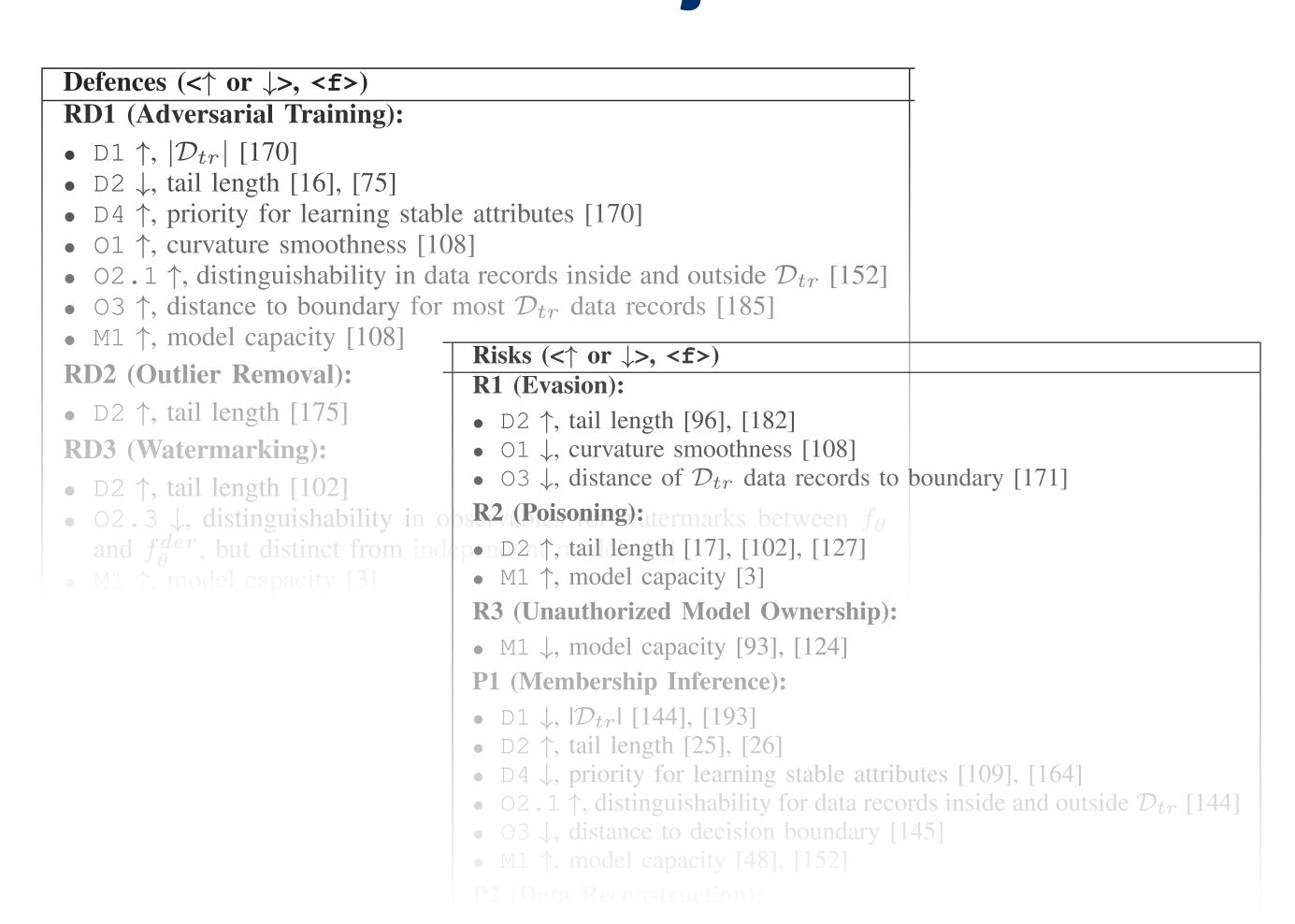
- Dataset: Tail length of distribution (D2);
 number of attributes (D3); priority of learning stable attributes (D4)
- Objective function: curvature smoothness (O1); distinguishability of observables across datasets (O2.1), subgroups (O2.2), models (O2.3); distance to decision boundary (O3)
- Model: same as M1

Situating prior work in framework

- Risk increases→●, decreases→●
- Interaction unexplored→
- Factors evaluated: empirical→●, theoretical→⊙,
 conjectured→○

Defenses	Risks	I	OVFT D1	D2	M D3	l emorizati D4	on 01	02	Both 03 M1	References
RD1 (Adversarial Training)	R1 (Evasion) R2 (Poisoning) R3 (Unauthorized Model Ownership) P1 (Membership Inference) P2 (Data Reconstruction) P3 (Attribute Inference) P4 (Distribution Inference) F (Discriminatory Behaviour)		○ ⊙, ●	· · · · · · · · · · · · · · · · · · ·			•	1: •	•	[96], [108], [182], [202] [161], [179] [90] ([101]: •) [71], [152] [117], [204] [156] [16], [38], [75], [105]
RD2 (Outlier Removal)	R1 (Evasion) R2 (Poisoning) R3 (Unauthorized Model Ownership)									[63] [163]
	P1 (Membership Inference) P2 (Data Reconstruction)			•						[26], [47]
	P3 (Attribute Inference) P4 (Distribution Inference) F (Discriminatory Behaviour)		•							[142]

Guideline for conjectures



Effectiveness of defense correlates with factor Change in factor (<f>) correlates with risk

- †: positive correlation; |: negative correlation
 Use arrows for <defense, f> and <f, risk>:
- If (\uparrow,\uparrow) or (\downarrow,\downarrow) \rightarrow \bullet ; else (\uparrow,\downarrow) or (\downarrow,\uparrow) \rightarrow \bullet

Conjecture is:

- unanimous if all factors agree
- determined by dominant factor (O1, O2, O3)

Non-common factors may affect interaction





