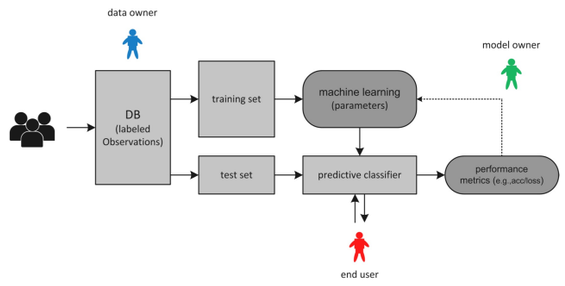
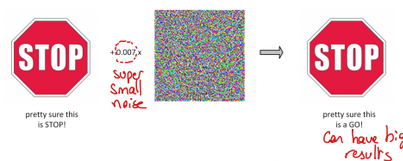
ML Security

General model:



Adversarial example

↳ Modify input on the user side, indistinguishable from naked eye



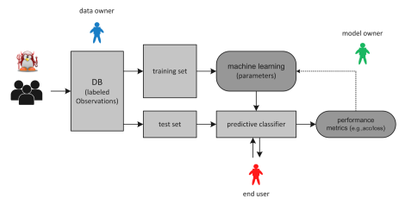
↳ We can send a lot of modified images to the model and use an optimization algorithm (gradient descent) to find the right noise

Countermeasures:

↳ Preprocessing on input
↳ Detection of suspicious input
↳ Adversarial training (ie generate adversarial random input and inject in dataset)

Data Poisoning

If the model is "self learning" from user input:



Counter measures:

→ Outlier detection
→ Compute distance between models
→ Distributed Data Source (everything cannot come from 1 ip)

Model Extraction

↳ Idea: extract model from queries!

↳ Much easier than model learning (ie creating the model)!

	Model creation	extracting f(x)
function	complex / noisy	deterministic/simple
Labeling	done by hand by humans for training and testing	on query of x

General idea:

- Logistic regression on data needs $|data| \approx 10 \cdot |features|$
- Model extract° on LR $|data| \approx |features| + c$

Membership inference

↳ Goal: Adversary would like to know if "you" were used in the training dataset

↳ How?: Error (Training) < Error (testing)

↳ Solutions: → Differencial Privacy (ie noisy output) but loses functionality
→ Well estimated model

Fairness

↳ General Idea: W/ humans discrimination comes from intent. No intention in computers (except if desined with bias => algorithm transparency, trojan, ...)

↳ Group x Individual: Group fairness: hire 50/50 M/F. Seems fair but might not be for certain males who were more qualified

Individual fairness: Similar qualif should be treated the same

↳ Formal definitions: For a model with A (= Sensitive attribute (gender)) R (= Ai's score prediction), Y (= the truth):

Independance $R \perp\!\!\!\perp A$:
("equality of outcome")
- Probability of Getting high score must be the same for everyone!
- Might lead to inequality (female was more qualified but we needed fairness so we hired a male)

Separation: $R \perp\!\!\!\perp A \mid Y$
("equality of opportunity")
→ Given 2 qualified ppl, their score should be ⊥ of A.

Lema: If
- Y binary (yes/no)
- $A \not\perp\!\!\!\perp R$
- $R \not\perp\!\!\!\perp Y$

Then both separation and independance cannot hold