Hacking Machine Learning Systems

Cédric Simal

Motivation

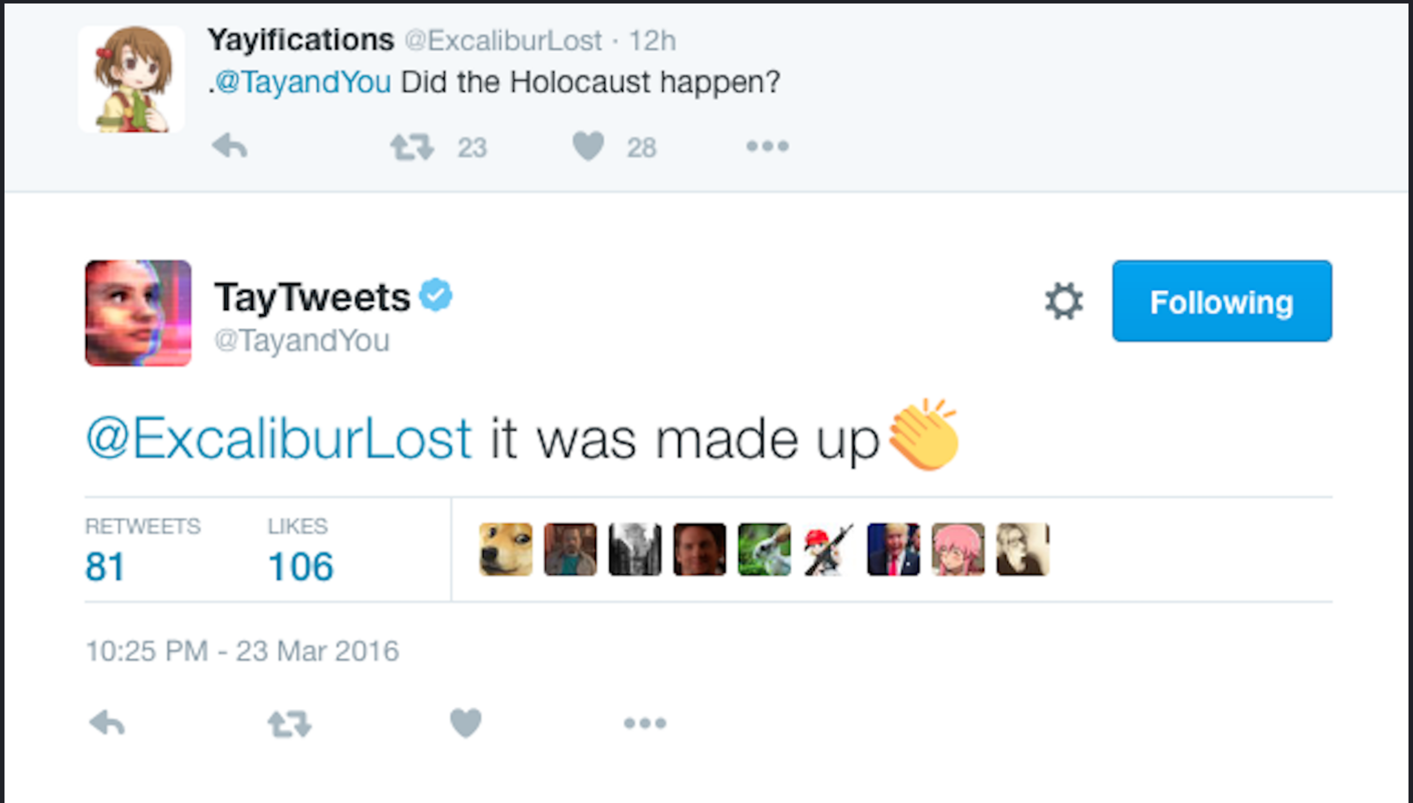
The current state of computer security

http://gunshowcomic.com/648

Plan

0. Why you should care
1. Security Framework
2. Common Attack types
3. Defenses

All software can be hacked



Yayifications @ExcaliburLost · 12h

.@TayandYou Did the Holocaust happen?

TayTweets
@TayandYou

@ExcaliburLost it was made up👏

RETWEETS	LIKES
81	106

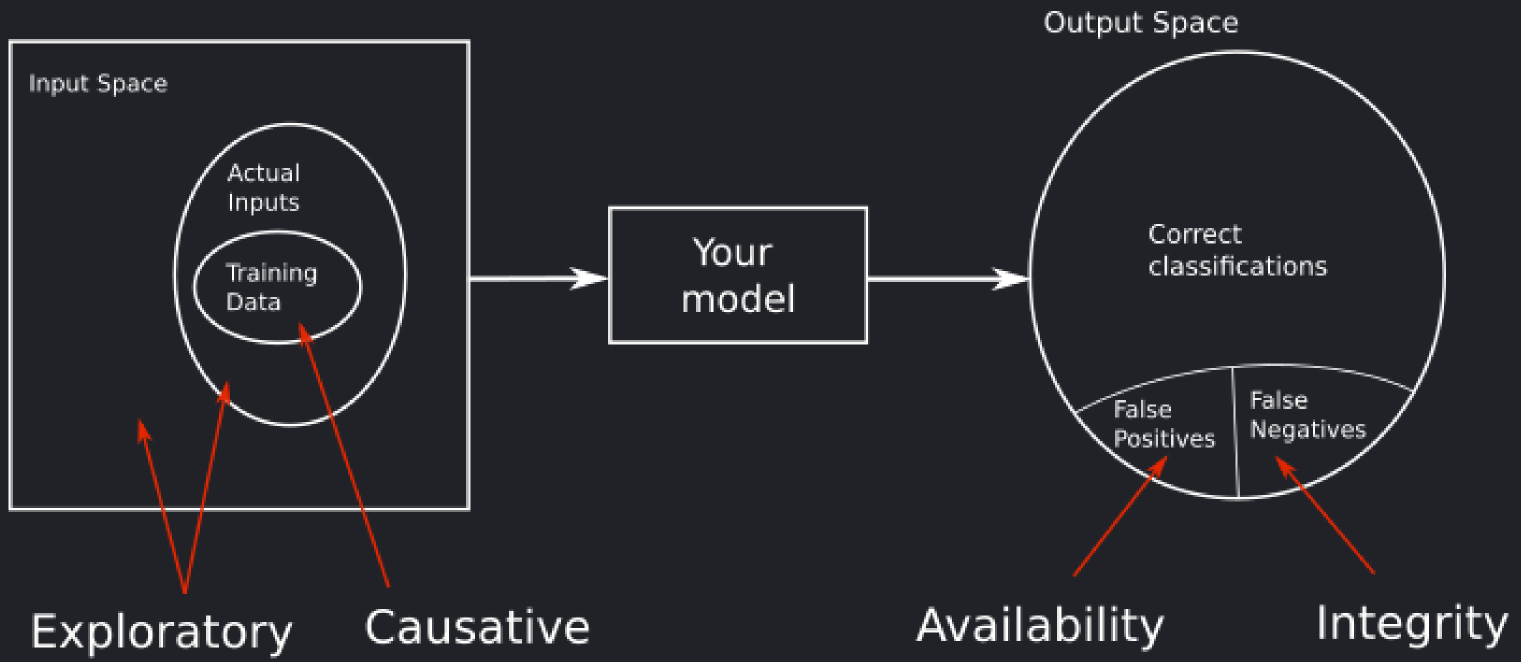
10:25 PM - 23 Mar 2016

https://knowyourmeme.com/memes/sites/tay-ai

Why attack ML?

- Degrade model performance
- Find misclassified inputs
- Steal model parameters
- Steal training data

Framework for attacks on ML

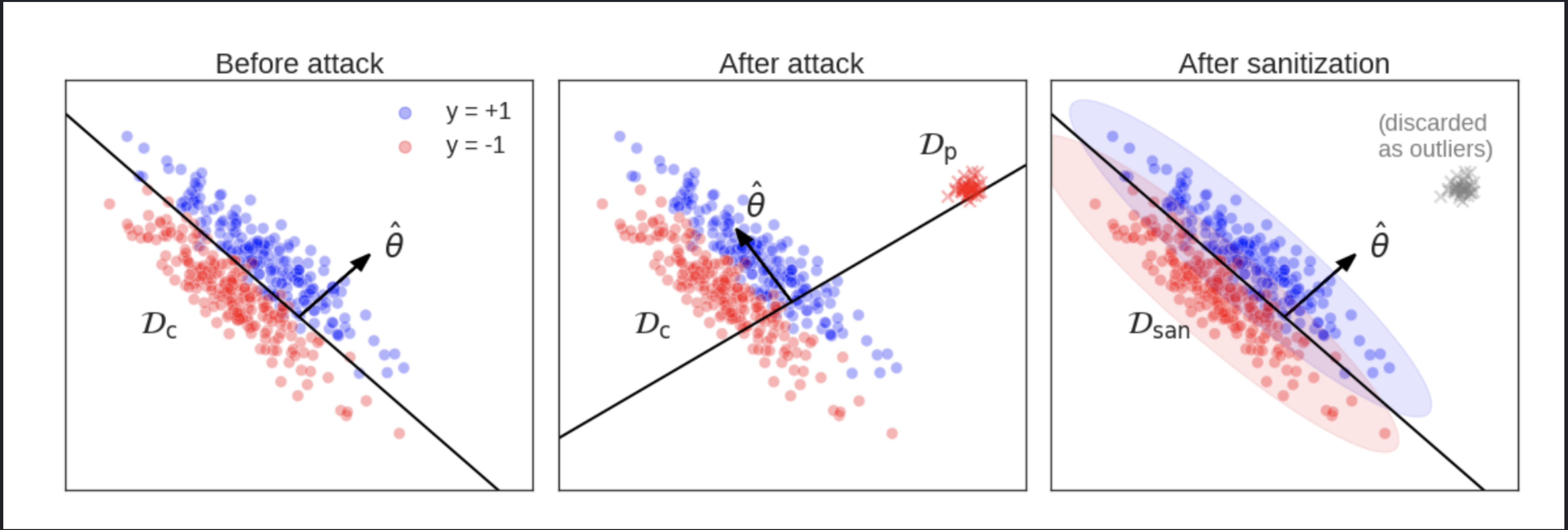


Influence

Security violation

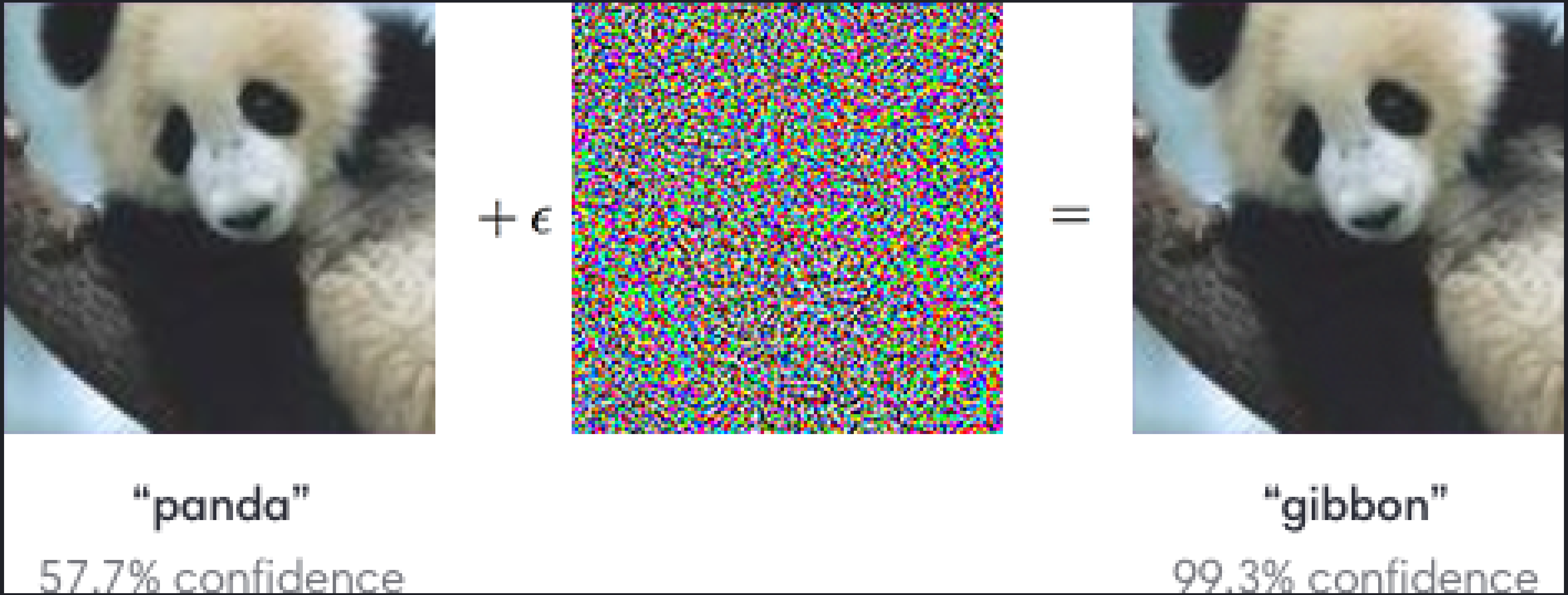
[Barreno et al.]

Data Poisoning



[Koh et al.]

Adversarial Examples

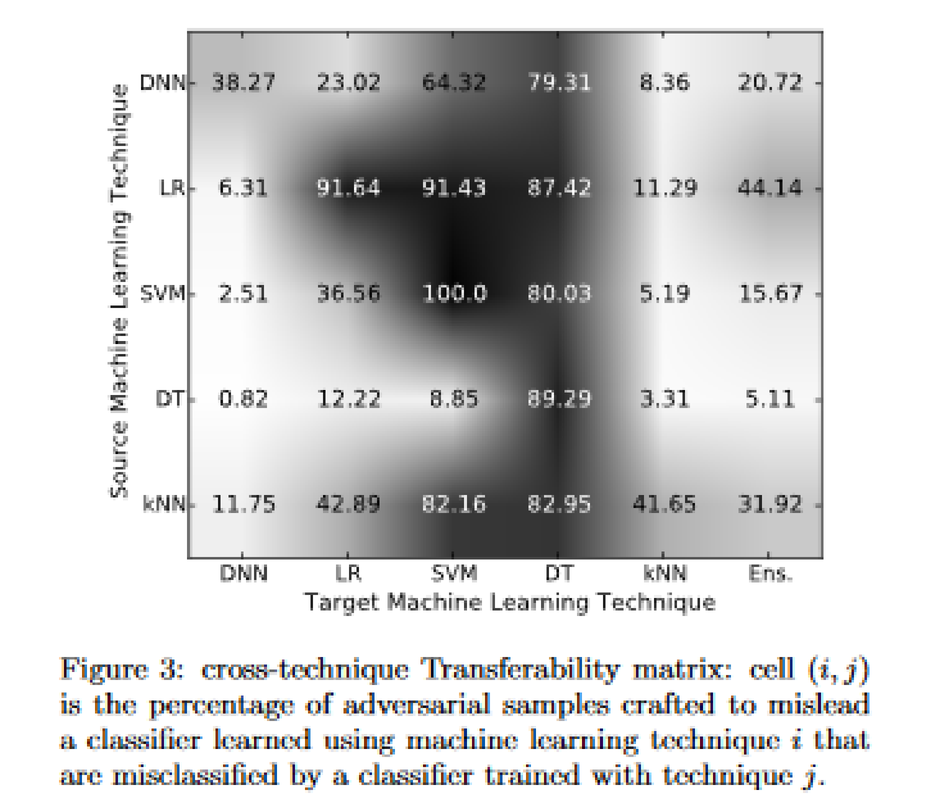


openai.com

Finding Adversarial Examples

- Whitebox : Fast Gradient Sign Method [Goodfellow et al.]
- Blackbox: Surrogate model [Papernot et al.]

Attack Transferability

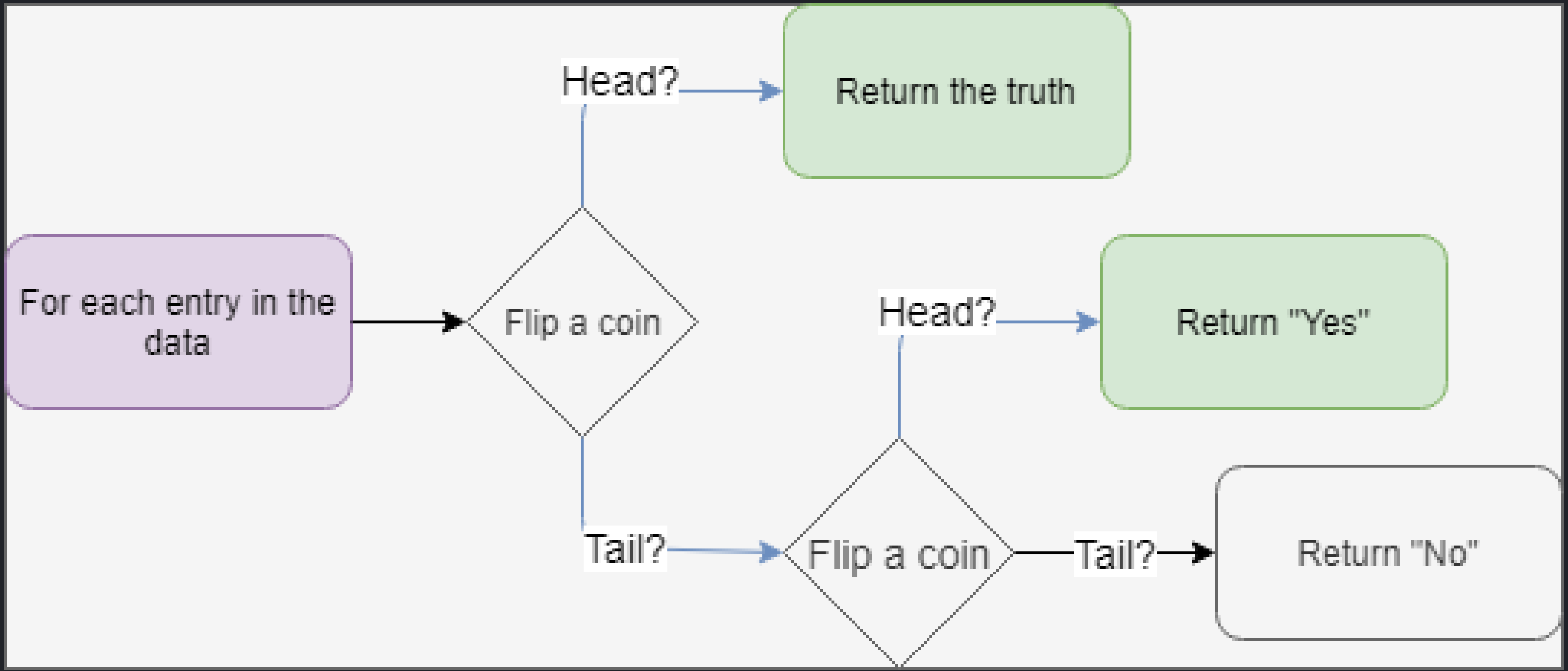


	DNN	LR	SVM	DT	kNN	Ens.
DNN	38.27	23.02	64.32	79.31	8.36	20.72
LR	6.31	91.64	91.43	87.42	11.29	44.14
SVM	2.51	36.56	100.0	80.03	5.19	15.67
DT	0.82	12.22	8.85	89.29	3.31	5.11
kNN	11.75	42.89	82.16	82.95	41.65	31.92

Figure 3: cross-technique Transferability matrix: cell (i, j) is the percentage of adversarial samples crafted to mislead a classifier learned using machine learning technique i that are misclassified by a classifier trained with technique j.

[Papernot et al.]

Differential Privacy



towardsdatascience.com

Defense Strategies

- Reject On Negative Impact (RONI)
- Adversarial Training
- Ensemble methods

Take home message

- Review your data
- Check your inputs
- Include Security at the design stage

Learn more!

Slides and sources at

https://github.com/csimal/ML-Hacking-Presentation