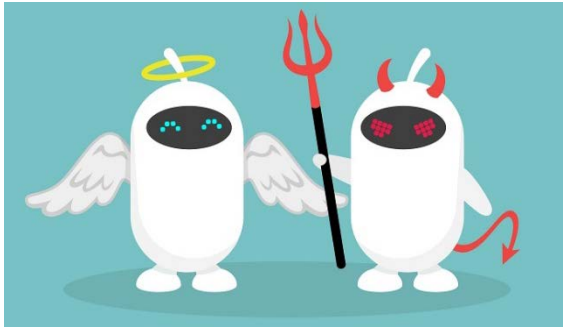


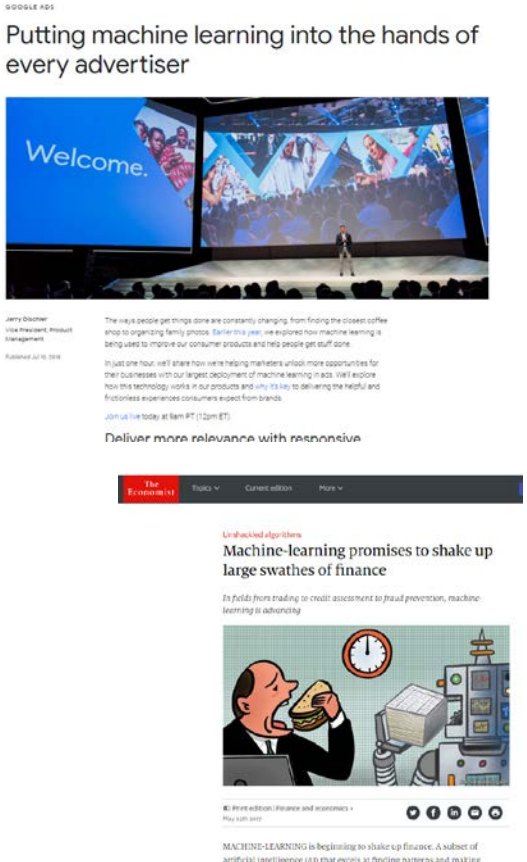
When Foes are Friends: A privacy perspective on adversarial examples

Carmela Troncoso
Security and Privacy Engineering Lab (EPFL)
carmela.troncoso@epfl.ch

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

SPRING
SECURITY AND PRIVACY ENGINEERING LABORATORY

The machine learning revolution


GOOGLE ADS

Putting machine learning into the hands of every advertiser

Jerry Dischler
Vice President, Product Management

Published Jul 10, 2018

The ways people get things done are constantly changing, from finding the closest coffee shop to organizing family photos. Earlier this year, we explored how machine learning is being used to improve our consumer products and help people get stuff done.

In just one hour, we'll share how we're helping marketers unlock more opportunities for their businesses with our largest deployment of machine learning in ads. We'll explore how this technology works in our products and why it's key to delivering the helpful and frictionless experiences consumers expect from brands.

Join us live today at 9am PT (12pm ET).

Deliver more relevance with responsive


Forbes

Billionaires Innovation Leadership Money Consumer Industry

44,707 views | Jun 11, 2018, 12:42am

10 Ways Machine Learning Is Revolutionizing Supply Chain Management

Louis Columbus Contributor

Bottom line: Machine learning makes it possible to discover patterns in supply chain data by relying on algorithms that quickly pinpoint the most influential factors to a supply networks' success, while constantly learning in the process.

The Economist

Topics Current edition More

Unshackled algorithms

Machine-learning promises to shake up large swathes of finance

In fields from trading to credit assessment to fraud prevention, machine learning is advancing

Print edition | Finance and economics

MACHINE-LEARNING is beginning to shake up finance. A subset of


nature biomedical engineering

Search Current Submit Login

Machine learning in healthcare

Collection home Research News & Comment

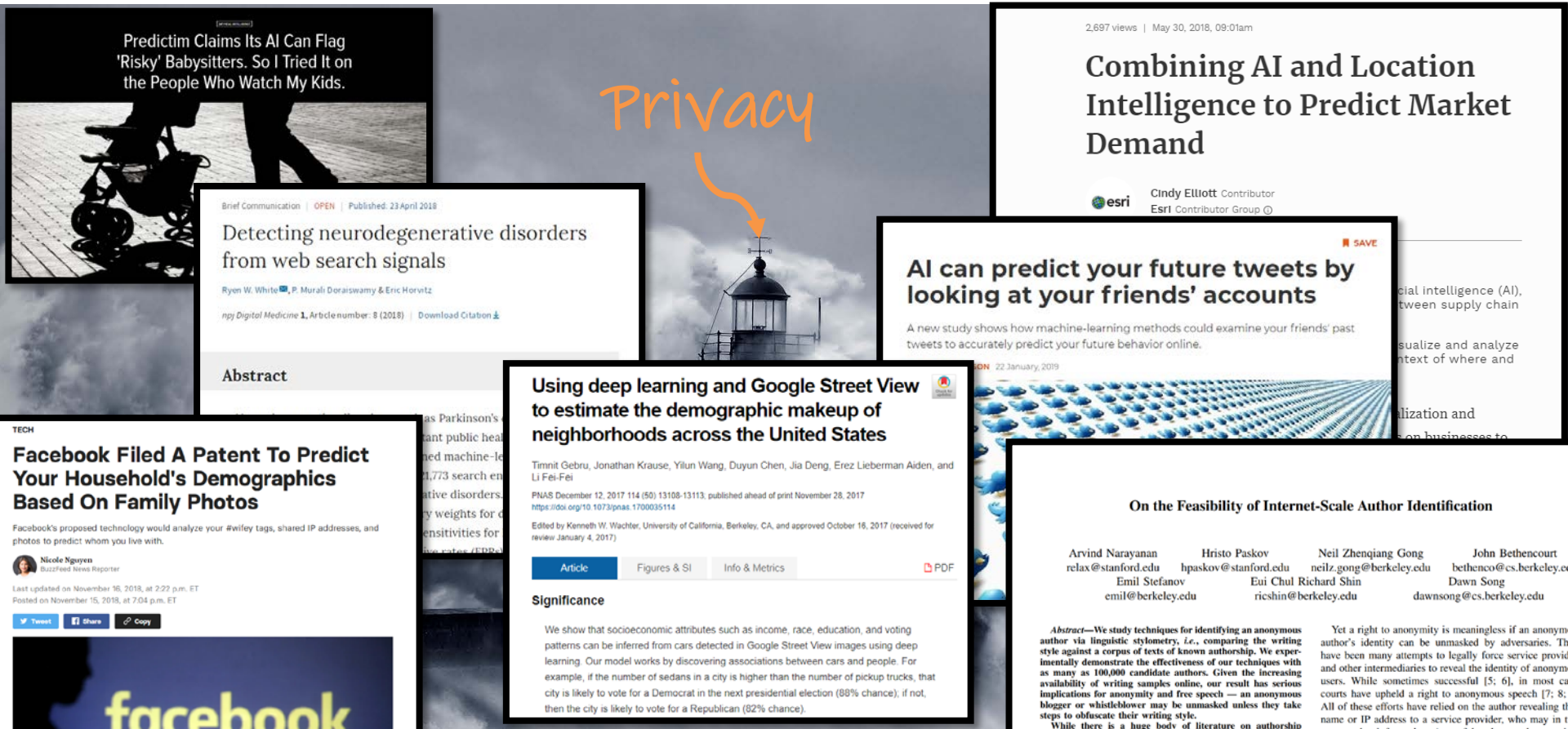
The accelerating power of machine learning in diagnosing disease and in sorting and classifying health data will empower physicians and speed up decision making in the clinic.

This Collection is updated when relevant new content is published. Content appears in reverse chronological order. See all Collections from Nature Biomedical Engineering.

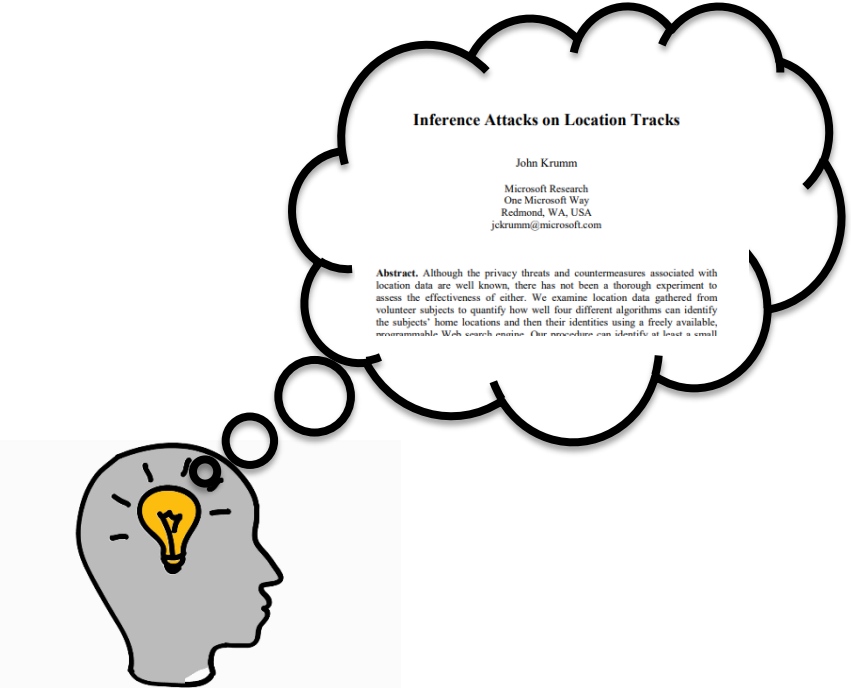
Research

The machine learning tsunami

The machine learning tsunami



Attacks are not new… but the adversary is



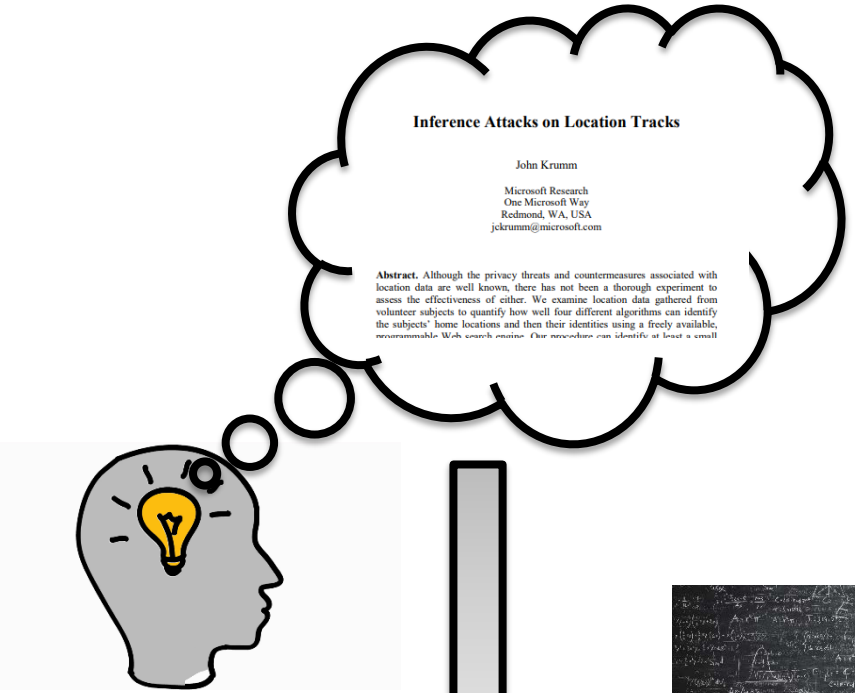
Inference Attacks on Location Tracks

John Krumm

Microsoft Research
One Microsoft Way
Redmond, WA, USA
jckrumm@microsoft.com

Abstract. Although the privacy threats and countermeasures associated with location data are well known, there has not been a thorough experiment to assess the effectiveness of either. We examine location data gathered from volunteer subjects to quantify how well four different algorithms can identify the subjects' home locations and then their identities using a freely available, programmable Web search engine. Our procedure can identify at least a small

Attacks are not new… but the adversary is

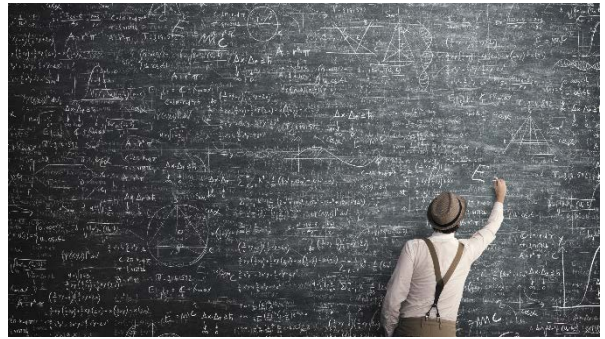


Inference Attacks on Location Tracks

John Krumm

Microsoft Research
One Microsoft Way
Redmond, WA, USA
jckrumm@microsoft.com

Abstract. Although the privacy threats and countermeasures associated with location data are well known, there has not been a thorough experiment to assess the effectiveness of either. We examine location data gathered from volunteer subjects to quantify how well four different algorithms can identify the subjects' home locations and then their identities using a freely available,



Protecting Location Privacy:
Optimal Strategy against Localization Attacks

Reza Shokri[†], George Theodorakopoulos[‡], Carmela Troncoso[⋆],
Jean-Pierre Hubaux[†], and Jean-Yves Le Boudec[†]

[†]LCA, EPFL, Lausanne, Switzerland,
[⋆]ESAT/COSIC, K.U.Leuven, Leuven-Heverlee, Belgium,
[‡]School of Computer Science and Informatics, Cardiff University, Cardiff, UK
[†]firstname.lastname@epfl.ch, [‡]g.theodorakopoulos@cs.cardiff.ac.uk,
[⋆]carmela.troncoso@esat.kuleuven.be

ABSTRACT

The mainstream approach to protecting the location-privacy of mobile users in location-based services (LBSs) is to alter the users' actual locations in order to reduce the location information exposed to the service provider. The location obfuscation algorithm behind an effective location-privacy preserving mechanism (LPPM) must consider three fundamen-

1. INTRODUCTION

The widespread use of smart mobile devices with continuous connection to the Internet has fostered the development of a variety of successful location-based services (LBSs). Even though LBSs can be very useful, these benefits come at a cost of users' privacy. The whereabouts users' disclose to the service provider expose aspects of their private life

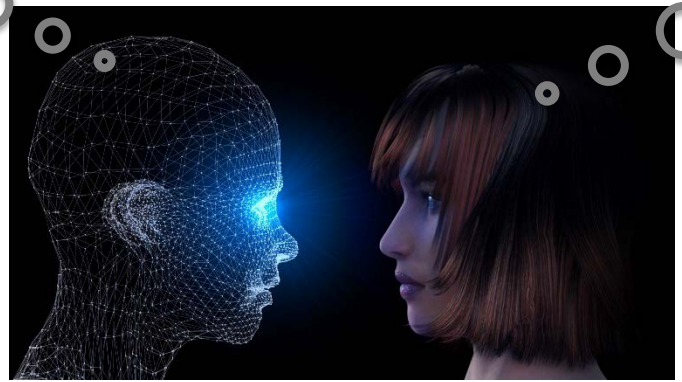
Attacks are not new… but the adversary is

Brief Communication | OPEN | Published: 23 April 2018

Detecting neurodegenerative disorders from web search signals

Ryen W. White, P. Murali Doraiswamy & Eric Horvitz

npj Digital Medicine **1**, Article number: 8 (2018) | Download Citation



?

Attacks are not new… but the adversary is

Brief Communication | OPEN | Published: 23 April 2018

Detecting neurodegenerative disorders from web search signals

Ryen W. White, P. Murali Doraiswamy & Eric Horvitz

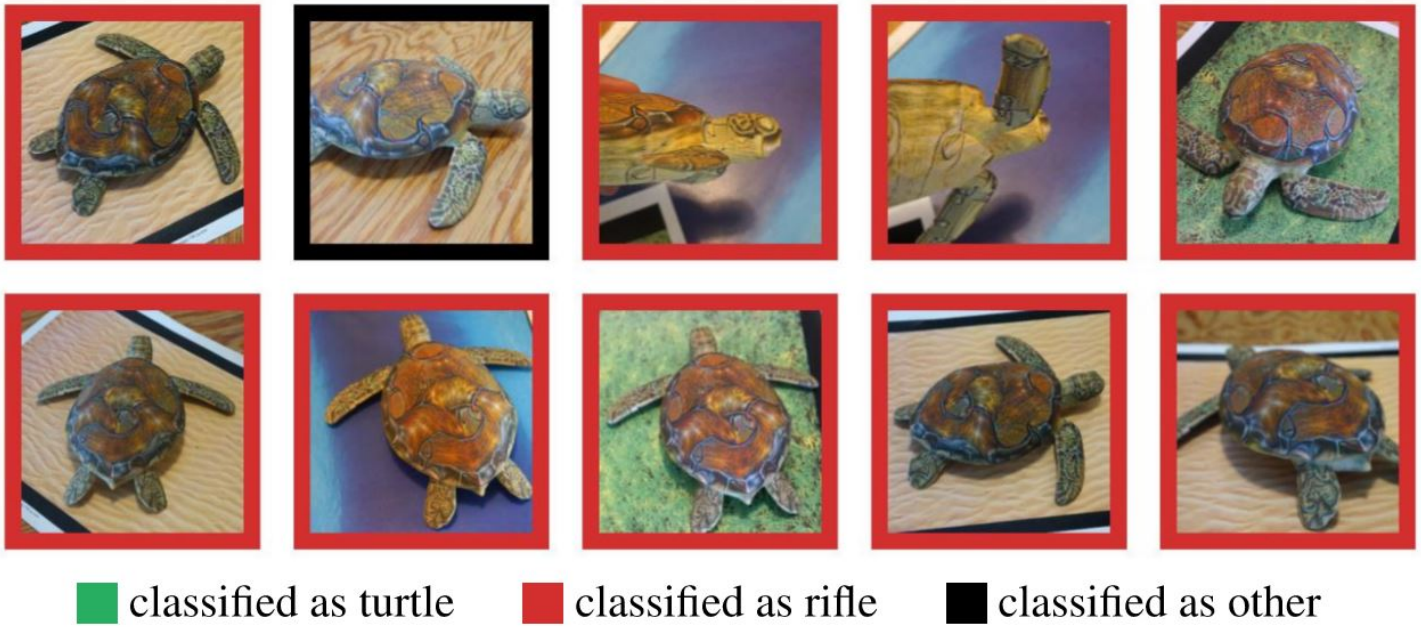
npj Digital Medicine **1**, Article number: 8 (2018) | Download Citation

?

Feature name	Class	Brief description	Weight
TimeBetweenRepeatQueries	Repetition	AVG time between repeat queries	1.000000
FractionOfQueriesAreRepeats	Repetition	% of all queries that are repeat queries	0.971182
NumberOfTremorEvents	Motor	# of tremor events[a]	0.715004
AverageTremorFrequency	Motor	AVG tremor frequency in hertz (# of oscillations/time)	0.595772
FractionOfQueriesHaveSymptoms	Symptom	% of all queries with 1+ symptoms	0.457336
AgeIs50To85	Risk Factors	Inferred searcher age is 50–85 years	0.432355
FractionOfClicksAreRepeats	Repetition	% of result clicks that are repeat clicks on same result	0.341164
FractionOfQueriesHaveRiskFactors	Risk Factors	% of all queries with 1+ risk factors	0.329801
GenderIsFemale	Risk Factors	Inferred gender is female	0.313425
TotalTimeCursorMoving	Motor	Total time mouse cursor is actively moving	0.297699
NumberOfScrollEvents	Motor	# of scroll events	0.259432
NumberOfScrollEventsDownward	Motor	# of scroll events downward	0.256692
AverageScrolVelocity	Motor	AVG scrolling velocity	0.249454
MinimumCursorYCoordinate	Motor	MIN *y*-coordinate of mouse cursor (top of page *y* is 0)	0.247770
FractionOfCursorTransitionsAreDirectionChanges	Motor	% of mouse cursor transitions with direction changes[b]	0.243873
AverageCursorAcceleration	Motor	AVG acceleration of mouse cursor	0.239814
NumberOfHyperlinkClicks	Motor	# of hyperlink clicks	0.239568
AverageCursorVelocity	Motor	AVG velocity of mouse cursor	0.232418
NumberOfCursorTransitionsAreDirectedUpward	Motor	# of transitions directed upward	0.232311
TotalDistanceScrolled	Motor	Total distance scrolled	0.215000
AverageCursorXCoordinate	Motor	AVG *x*-coordinate of mouse cursor (left of page *x* is 0)	0.214955
FractionOfCursorTransitionsWhitespace	Motor	% of time mouse cursor in whitespace[c]	0.211926

The goal is not to understand, it is to beat!

The ML community has build the path: **ADVERSARIAL EXAMPLES**!



Adversarial examples are only “adversarial” when you are the algorithm!

Are we done? Can we save privacy?

Are we done? Can we save privacy?

Privacy defenses are constrained

Limited eligible transformations for…

Transformations are high cost…

- hiding demographics in text → changes in meaning/readability
- hiding interests from searches → undesired projected image
- hiding network traffic patterns → $$$$$$
- …

Current approaches are of limited utility…

Our approach

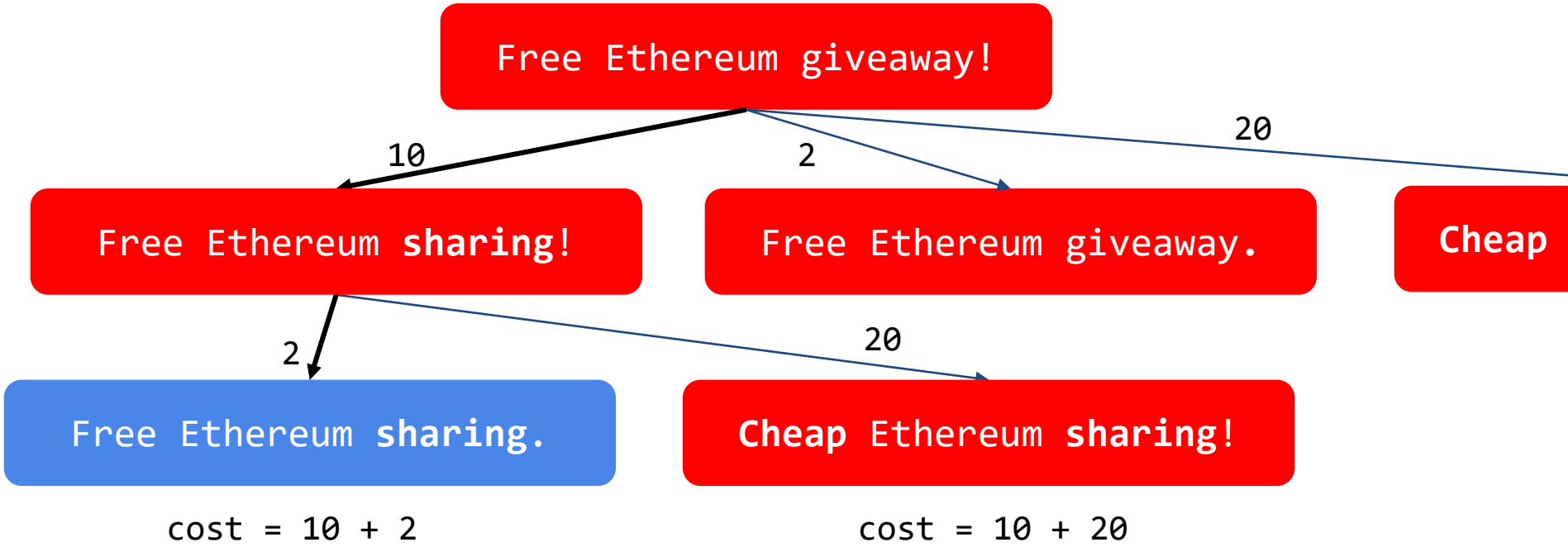
Limited eligible transformations for…

Model the search of adversarial examples as a graph search

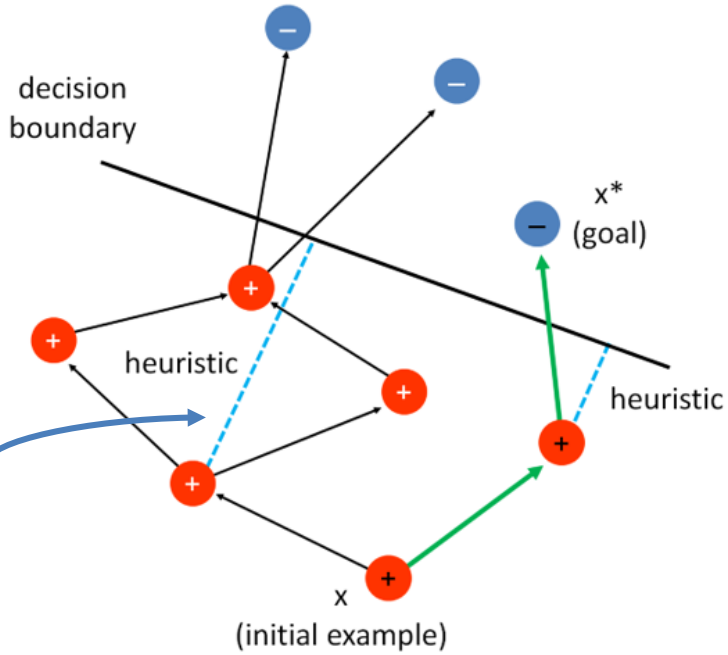
Transformations are high cost…

Use A* as search algorithm to ensure minimal cost

Transformation graph as a capability model



A* : provable minimal cost adversarial example



Heuristic: pointwise robustness

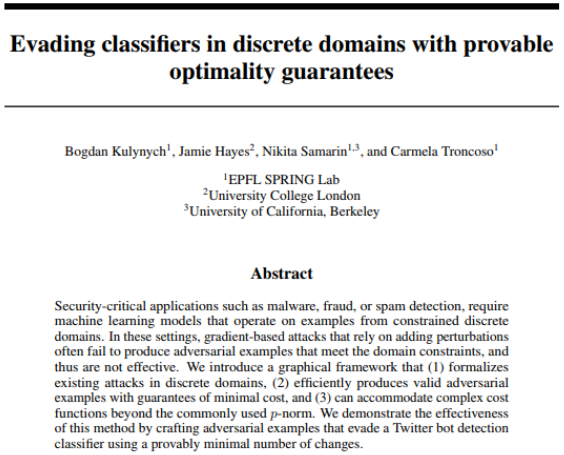
Where we stand now

- Adversarial examples are a basis to systematically build privacy defenses against machine learning

- Graph search: a promising avenue to defenses
  - Encode defender capabilities
  - Encode cost

- Added value
  - Minimal cost = security guarantee
  - Baseline for other defenses

Thanks!



Evading classifiers in discrete domains with provable optimality guarantees

Bogdan Kulynych[1], Jamie Hayes[2], Nikita Samarin[1,3], and Carmela Troncoso[1]

[1]EPFL SPRING Lab
[2]University College London
[3]University of California, Berkeley

Abstract

Security-critical applications such as malware, fraud, or spam detection, require machine learning models that operate on examples from constrained discrete domains. In these settings, gradient-based attacks that rely on adding perturbations often fail to produce adversarial examples that meet the domain constraints, and thus are not effective. We introduce a graphical framework that (1) formalizes existing attacks in discrete domains, (2) efficiently produces valid adversarial examples with guarantees of minimal cost, and (3) can accommodate complex cost functions beyond the commonly used p-norm. We demonstrate the effectiveness of this method by crafting adversarial examples that evade a Twitter bot detection classifier using a provably minimal number of changes.

https://arxiv.org/abs/1810.10939

Library for attacking machine learning in discrete domains using graph search

https://github.com/spring-epfl/trickster

SPRING
SECURITY AND PRIVACY ENGINEERING LABORATORY

https://spring.epfl.ch/