



When Foes are Friends: A privacy perspective on adversarial examples

Carmela Troncoso

Security and Privacy Engineering Lab (EPFL)

carmela.troncoso@epfl.ch

The machine learning revolution

GOOGLE ADS

Putting machine learning into the hands of every advertiser



Jerry DiShier
Vice President, Product
Management

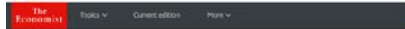
Published Jul 16, 2016

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 latest deployment of machine learning in ads. We'll explore how this technology works in our products and give you a key to delivering the helpful and frictionless experiences consumers expect from brands.

[Join us live today at 8am PT \(12am ET\)](#)

Deliver more relevance with responsive



Unhacked 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 • May 2017

MACHINE-LEARNING is beginning to shake up finance. A subset of artificial intelligence (AI) that excels at finding patterns and making



Forbes

Billionaires Innovation Leadership Money Consumer Industry

44,767 views | Jun 15, 2016, 12:42pm

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.

nature biomedical engineering

Collection | 18 October 2016

Machine learning in healthcare

Collection home Search 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

Privacy



Predictim Claims Its AI Can Flag 'Risky' Babysitters. So I Tried It on the People Who Watch My Kids.

2,697 views | May 30, 2018, 09:01am

Combining AI and Location Intelligence to Predict Market Demand

esri Cindy Elliott Contributor
Esri Contributor Group

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

Abstract

AI can predict your future tweets by looking at your friends' accounts

A new study shows how machine-learning methods could examine your friends' past tweets to accurately predict your future behavior online.

22 January, 2019

Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States

Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei

PNAS December 12, 2017 114 (50) 13108-13113, published ahead of print November 28, 2017
<https://doi.org/10.1073/pnas.1700035114>

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved October 16, 2017 (received for review January 4, 2017)

Article

Figures & SI

Info & Metrics

PDF

Significance

We show that socioeconomic attributes such as income, race, education, and voting patterns can be inferred from cars detected in Google Street View images using deep learning. Our model works by discovering associations between cars and people. For example, if the number of sedans in a city is higher than the number of pickup trucks, that city is likely to vote for a Democrat in the next presidential election (89% chance); if not, then the city is likely to vote for a Republican (82% chance).

On the Feasibility of Internet-Scale Author Identification

Arvind Narayanan
relax@stanford.edu
Email Stefanov
emil@berkeley.edu

Hristo Paskov
hpaskov@stanford.edu
Eui Chul Richard Shin
ricshin@berkeley.edu

Neil Zhenqiang Gong
neil.gong@berkeley.edu
Dawn Song
dawnsong@cs.berkeley.edu

John Bethencourt
bethenco@cs.berkeley.edu
Dawn Song
dawnsong@cs.berkeley.edu

Abstract—We study techniques for identifying an anonymous author via linguistic stylometry, *i.e.*, comparing the writing style against a corpus of texts of known authorship. We experimentally demonstrate the effectiveness of our techniques with as many as 100,000 candidate authors. Given the increasing availability of writing samples online, our result has serious implications for anonymity and free speech — an anonymous blogger or whistleblower may be unmasked unless they take steps to obfuscate their writing style.

While there is a huge body of literature on authorship identification, yet a right to anonymity is meaningless if an anonymous author's identity can be unmasked by adversaries. There have been many attempts to legally force service providers and other intermediaries to reveal the identity of anonymous users. While sometimes successful [5; 6], in most cases courts have upheld a right to anonymous speech [7; 8]. All of these efforts have relied on the author revealing their name or IP address to a service provider, who may in turn

cial intelligence (AI),
between supply chain
visualize and analyze
context of where and
alization and
on businesses to

TECH

Facebook Filed A Patent To Predict Your Household's Demographics Based On Family Photos

Facebook's proposed technology would analyze your #wifey tags, shared IP addresses, and photos to predict whom you live with.

Nicole Nguyen
BuzzFeed News Reporter

Last updated on November 16, 2018, at 2:27 p.m. ET
Posted on November 15, 2018, at 7:04 p.m. ET

Tweet Share Copy

facebook

Attacks are not new... but the adversary is

Inference Attacks on Location Tracks

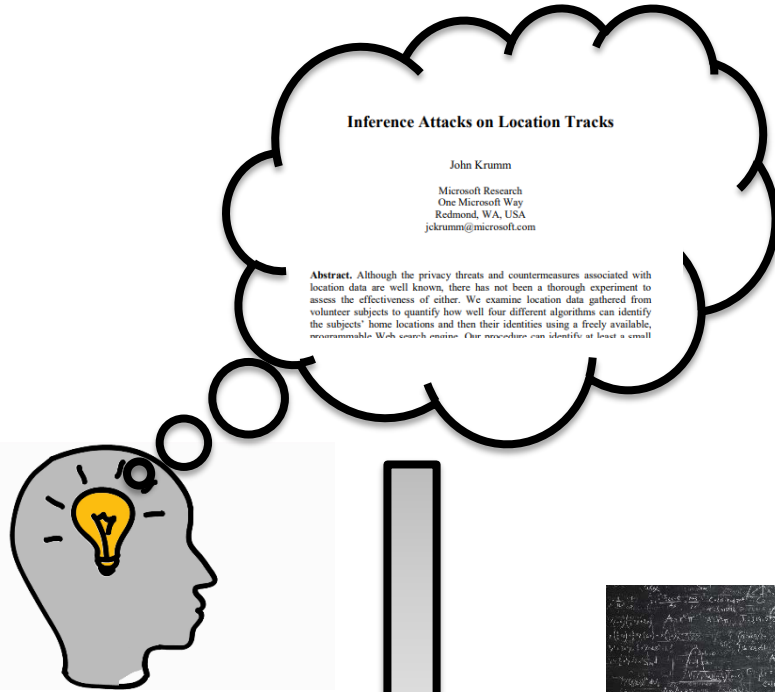
John Krumm

Microsoft Research
One Microsoft Way
Redmond, WA, USA
jkrumm@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, searchable Web search engine. Our procedure can identify at least a small



Attacks are not new... but the adversary is

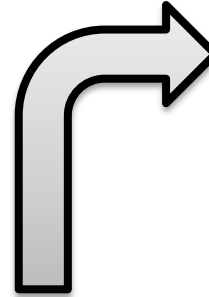
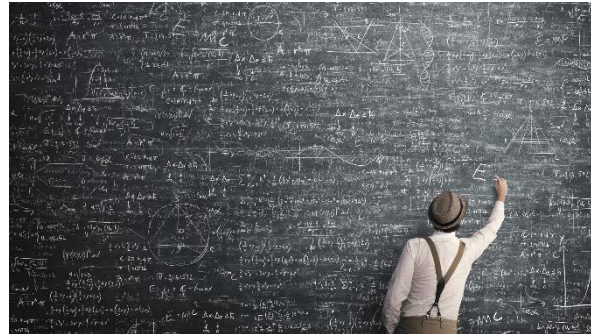
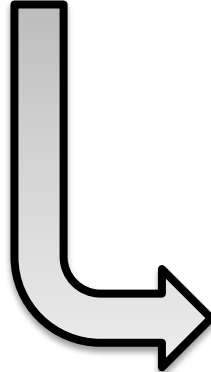


Inference Attacks on Location Tracks

John Krumm

Microsoft Research
One Microsoft Way
Redmond, WA, USA
jkrumm@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, non-rememberable Web search engine. Our procedure can identify at least a small



Protecting Location Privacy: Optimal Strategy against Localization Attacks

Reza Shokri¹, George Theodorakopoulos¹, Carmela Troncoso¹,
Jean-Pierre Hubaux¹, and Jean-Yves Le Boudec¹

¹LCA, EPFL, Lausanne, Switzerland,
²ESAT/COSIC, K.U.Leuven, Leuven-Heverlee, Belgium,
³School of Computer Science and Informatics, Cardiff University, Cardiff, UK,
⁴firstname.lastname@epfl.ch, ⁵g.theodorakopoulos@cs.cardiff.ac.uk,
⁶carmela.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 that is not apparent at first, but can be inferred from the

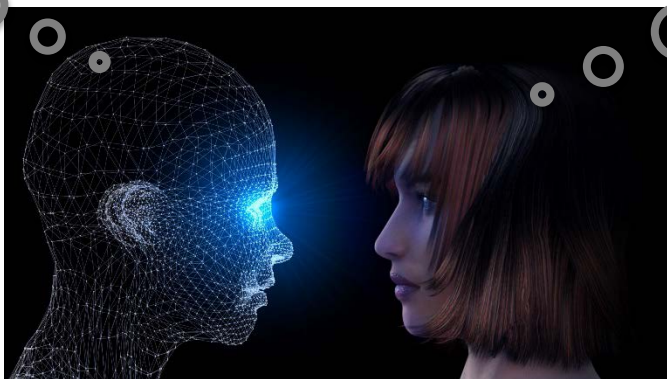
Attacks are not new... but the adversary is

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

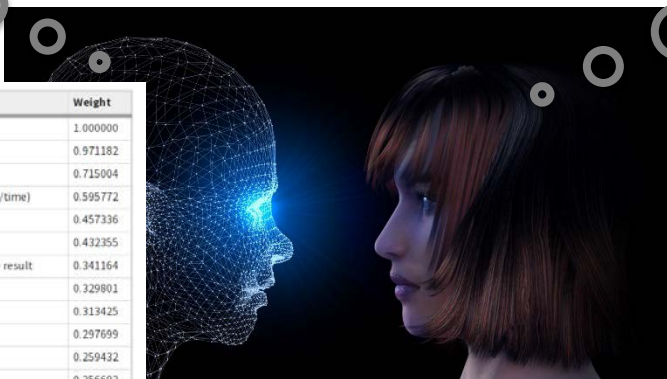
Detecting neurodegenerative disorders from web search signals

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

npj Digital Medicine 1, Article number: 8 (2018) | [Download Citation](#)

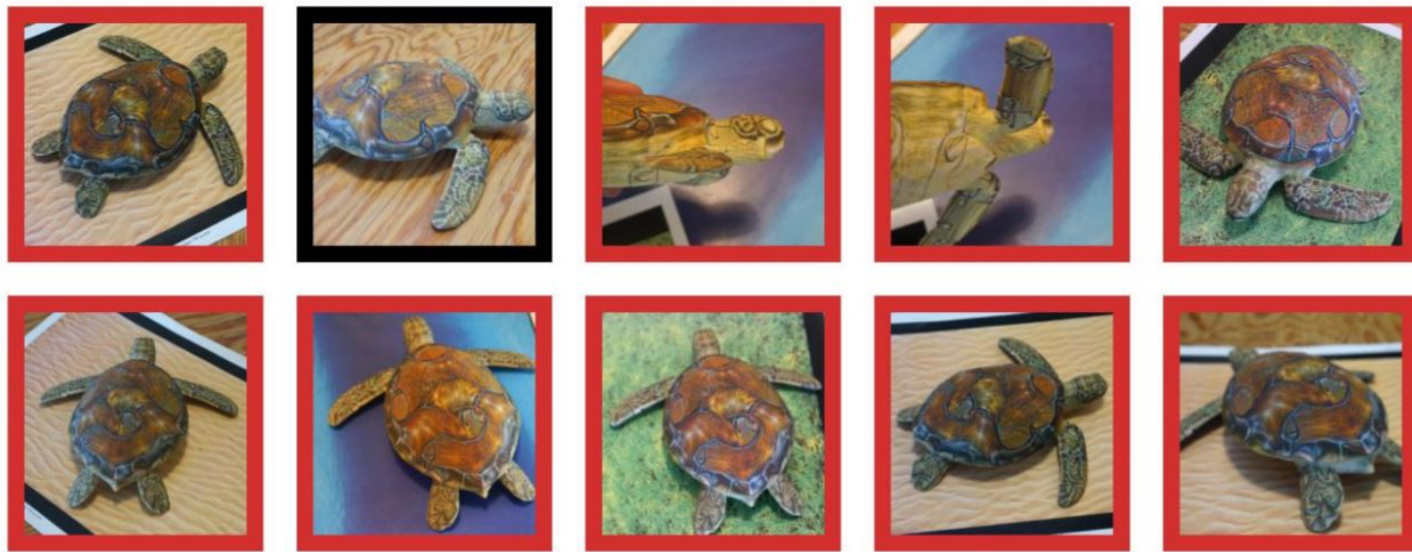


Feature name	Class	Brief description	Weight
TimeBetweenRepeatQueries	Repetition	AVG time between repeat queries	1.00000
FractionOfQueriesAreRepeats	Repetition	% of all queries that are repeat queries	0.971182
NumberOfTremorEvents	Motor	# of tremor events ^a	0.715004
AverageTremorFrequency	Motor	AVG tremor 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
AverageScrollVelocity	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



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

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



■ classified as turtle

■ classified as rifle

■ classified as other

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...



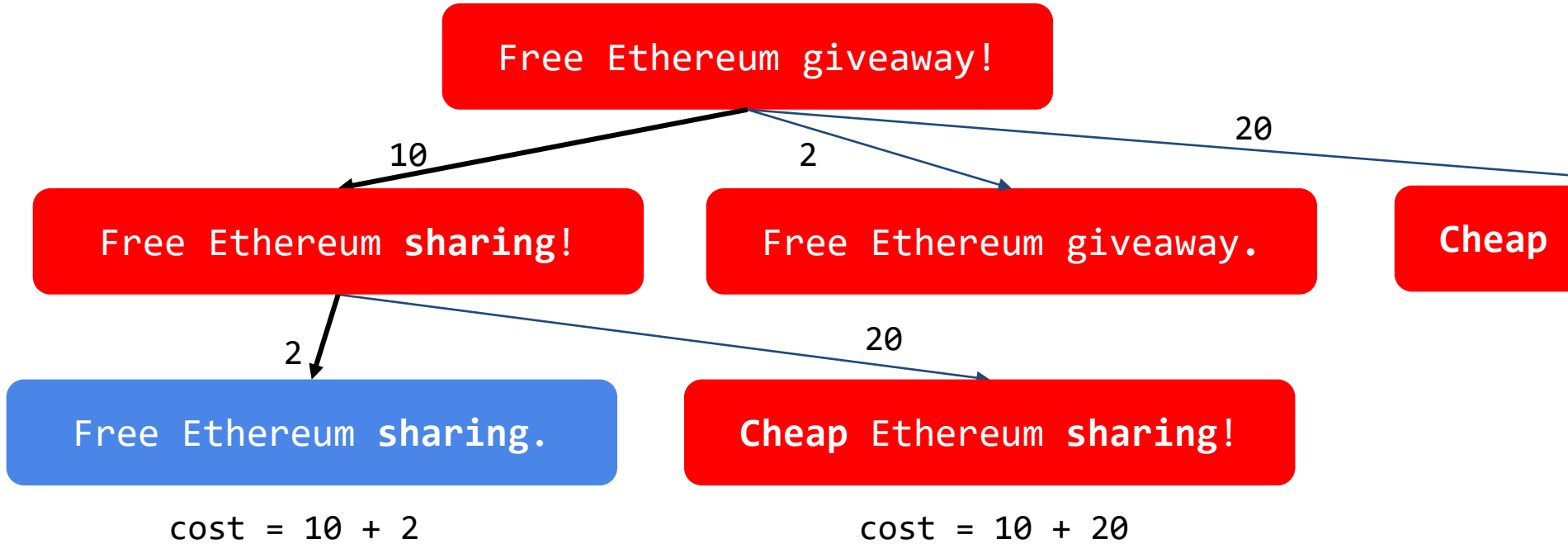
Model the search of adversarial
examples as a graph search

Transformations are
high cost...

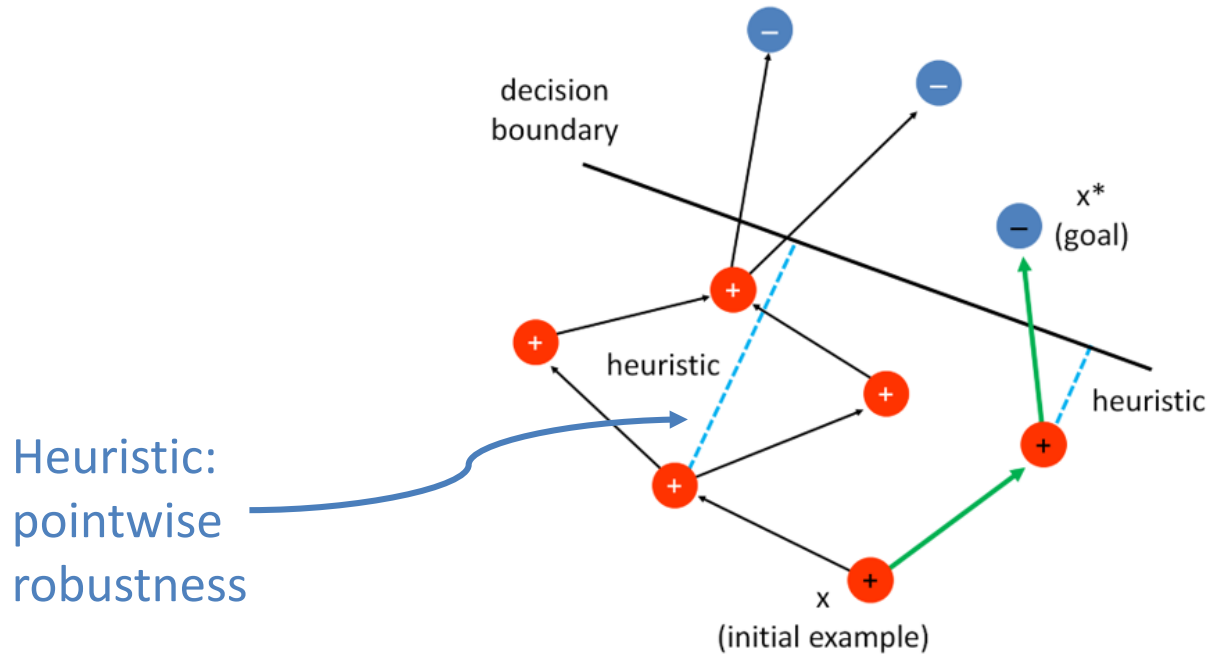


Use A^* as search algorithm to
ensure minimal cost

Transformation graph as a capability model



A^* : provable minimal cost adversarial example



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 Kulynych¹, Jamie Hayes², Nikita Samarin^{1,3}, and Carmela Troncoso¹

¹EPFL SPRING Lab

²University College London

³University 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.



Library for attacking machine learning in discrete domains using graph search

<https://arxiv.org/abs/1810.10939>

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



<https://spring.epfl.ch/>