PETs, POTs, and Pitfalls

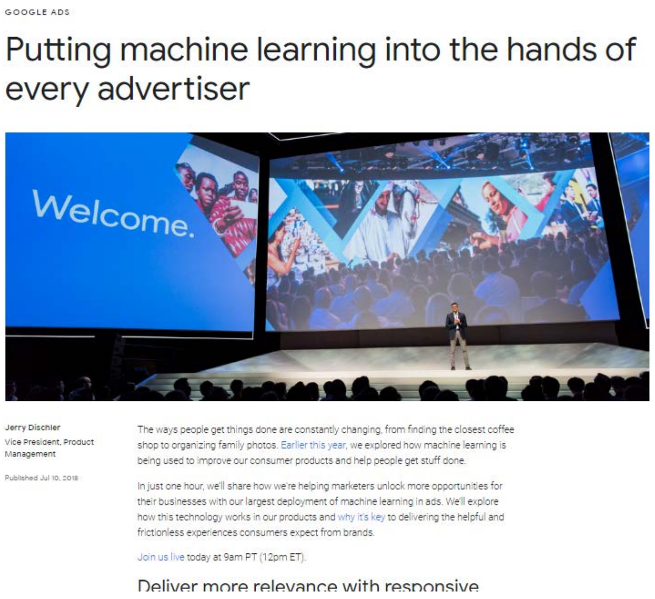
Rethinking the Protection of Users against Machine Learning

Carmela Troncoso

@carmelatroncoso

Security and Privacy Engineering Lab

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

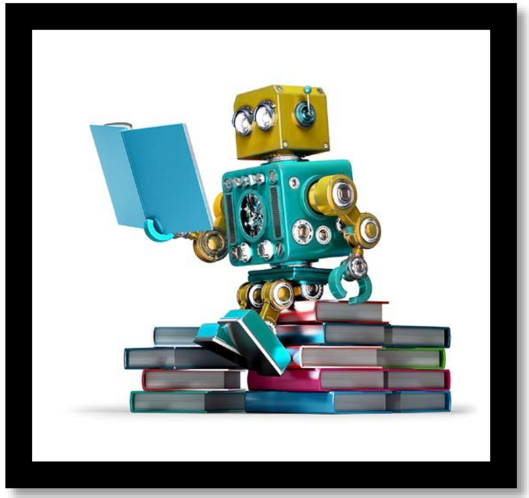
Unshackled algorithms

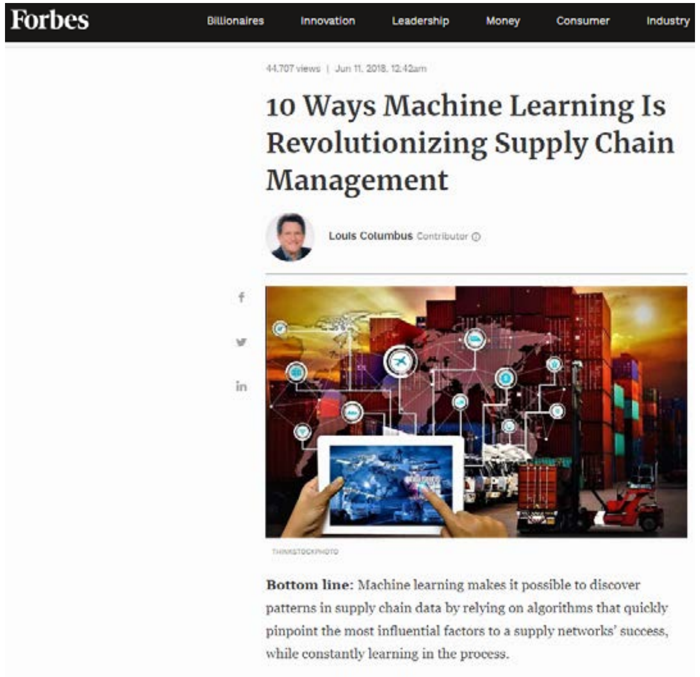
Machine-learning promises to shake up large swathes of finance

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

Print edition | Finance and economics ›
May 25th 2017

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





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

10 Ways Machine Learning Is Revolutionizing Supply Chain Management

Louis Columbus Contributor

THINKSTOCKPHOTO

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.

Collection | 10 October 2018

Machine learning in healthcare

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*.

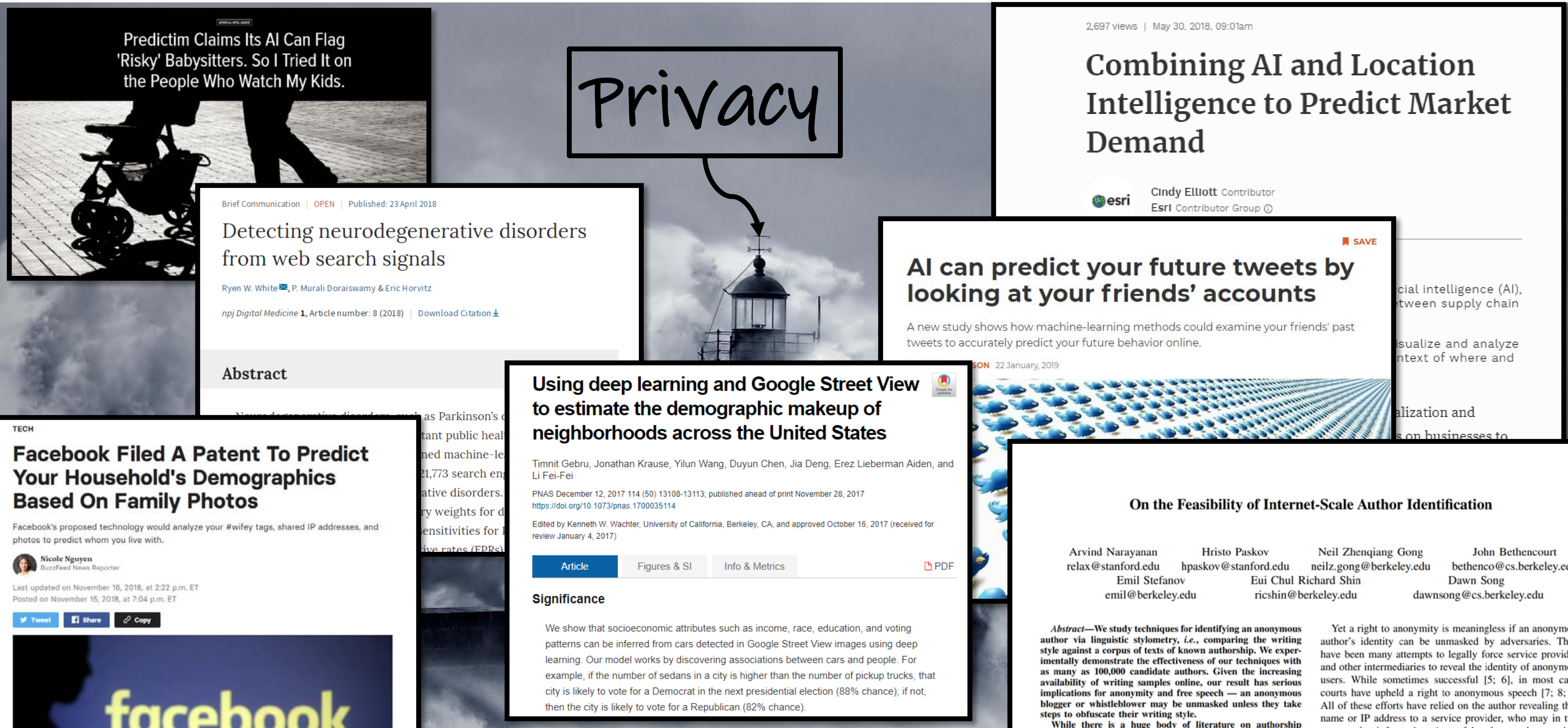
Research

The machine learning tsunami

Privacy

Social Justice

The ML tsunami on privacy



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

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

as Parkinson's d
tant public heal
ned machine-le
21,773 search eng
ative disorders.
ry weights for d
ensitivities for F
ive rates (FPRs)

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

Combining AI and Location Intelligence to Predict Market Demand

Cindy Elliott Contributor
Esri Contributor Group

cial intelligence (AI),
tween supply chain
sualize and analyze
ntext of where and
alization and
s on businesses to

SAVE

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.

SON 22 January, 2019

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 2:22 p.m. ET
Posted on November 15, 2018, at 7:04 p.m. ET

Tweet | Share | Copy

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 (88% chance); if not, then the city is likely to vote for a Republican (82% chance).

On the Feasibility of Internet-Scale Author Identification

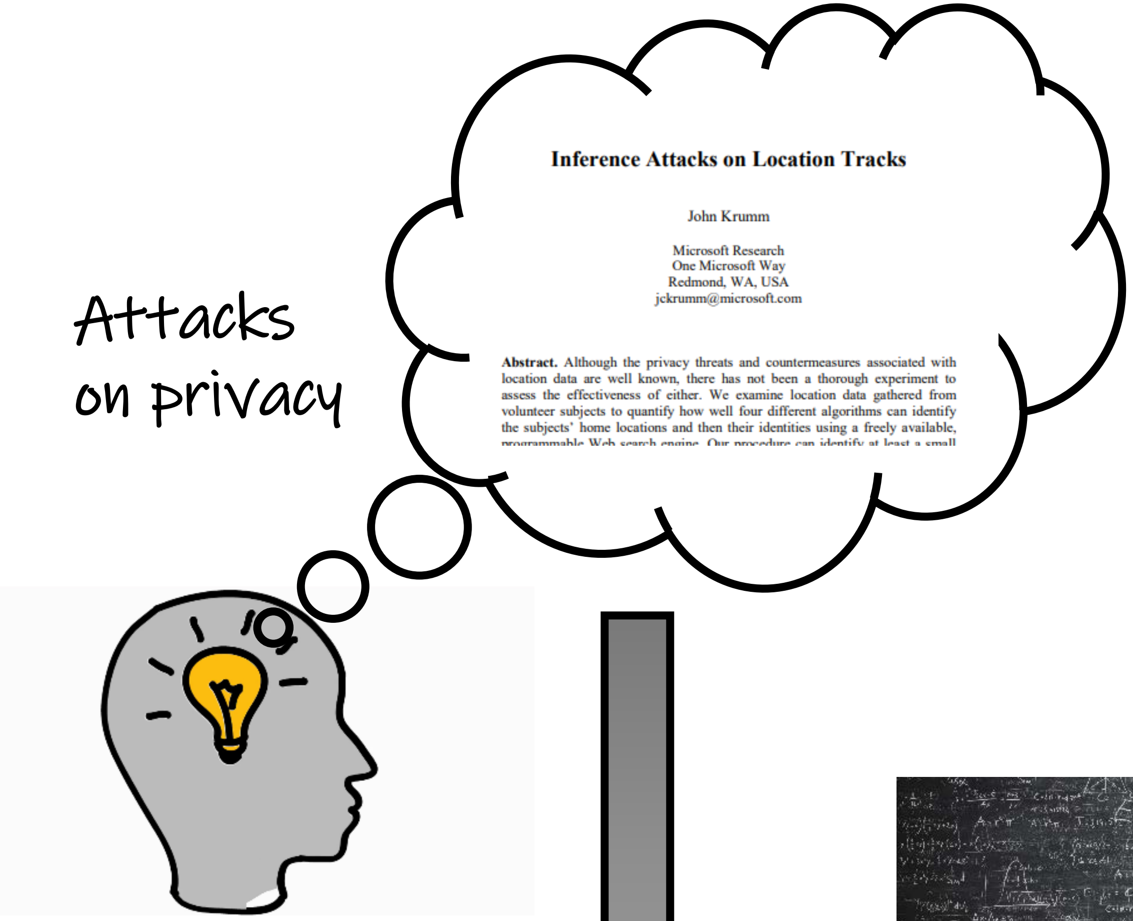
Arvind Narayanan relax@stanford.edu
Hristo Paskov hpaskov@stanford.edu
Neil Zhenqiang Gong neilz.gong@berkeley.edu
John Bethencourt bethenco@cs.berkeley.e
Emil Stefanov emil@berkeley.edu
Eui Chul Richard Shin ricshin@berkeley.edu
Dawn Song dawnsong@cs.berkeley.edu

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

Yet a right to anonymity is meaningless if an anonym
author's identity can be unmasked by adversaries. Th
have been many attempts to legally force service provid
and other intermediaries to reveal the identity of anonym
users. While sometimes successful [5; 6], in most ca
courts have upheld a right to anonymous speech [7; 8;
All of these efforts have relied on the author revealing th
name or IP address to a service provider, who may in t

Attacks are not new... but the adversary is



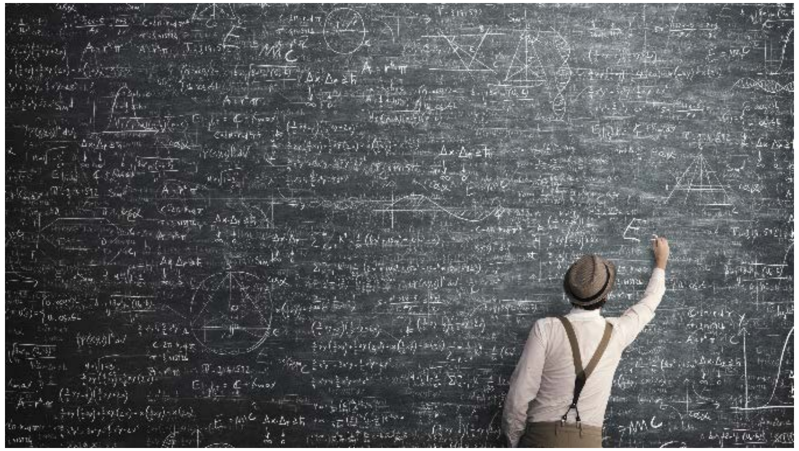
Inference Attacks on Location Tracks

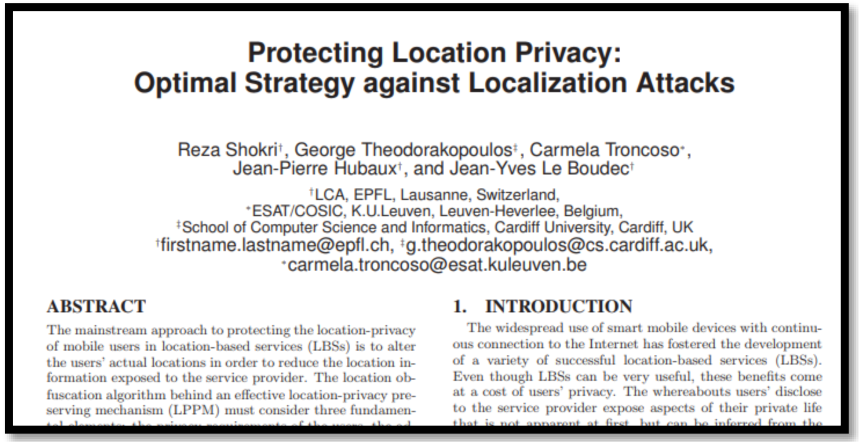
John Krumm

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 on privacy





**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 fundamental elements: the privacy requirements of the users, the ad-

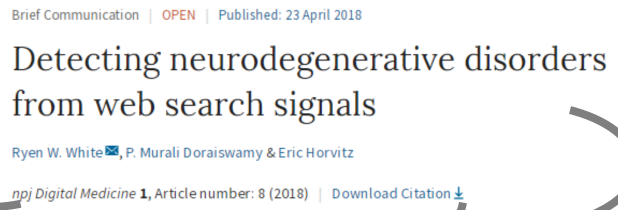
1. INTRODUCTION

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

Privacy Enhancing Technologies

PETs

Attacks are not new… but the adversary is



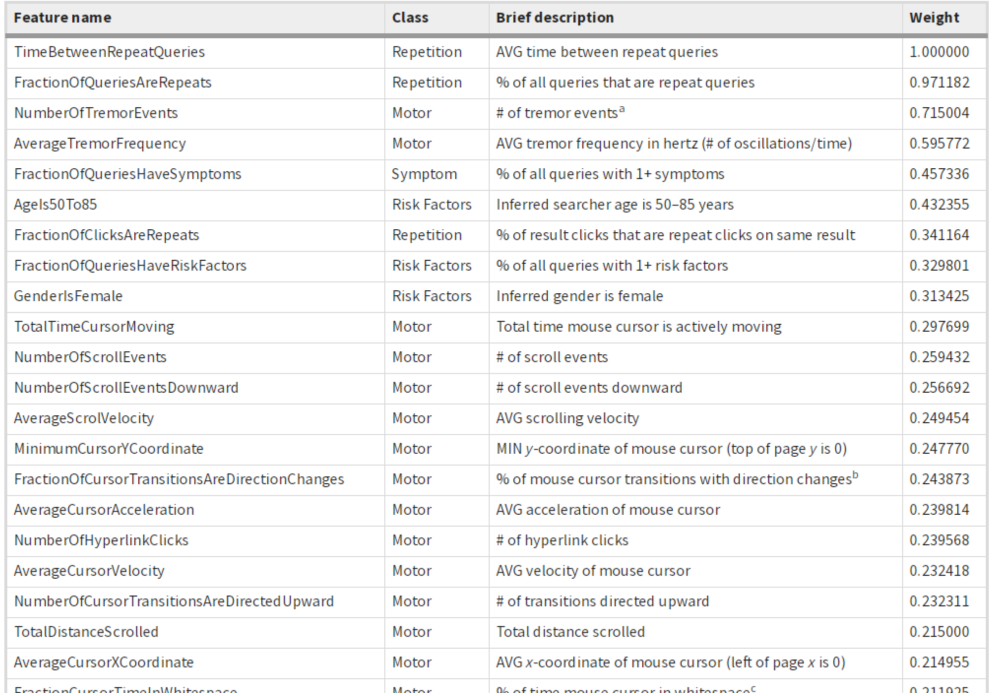
Brief Communication | OPEN | Published: 23 April 2018

Detecting neurodegenerative disorders from web search signals

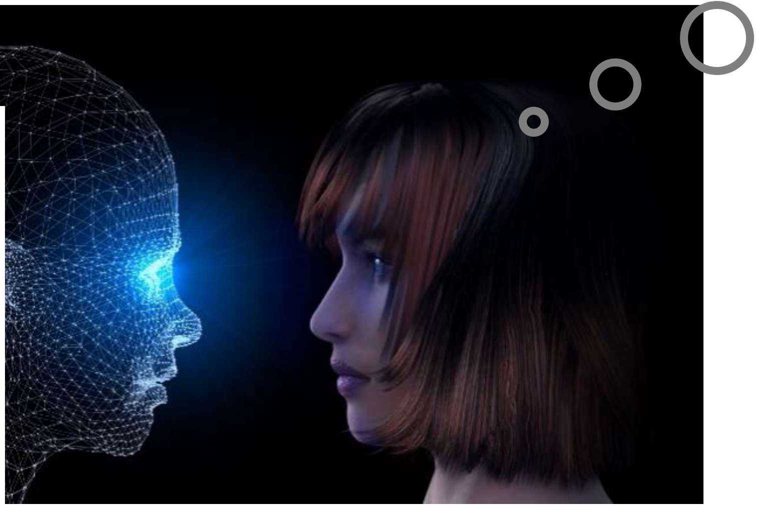
Ryen W. White, P. Murali Doraiswamy & Eric Horvitz

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

Attacks



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
FractionCursorTimeInWhitespace	Motor	% of time mouse cursor in whitespace[c]	0.211925



?

PETs??

Machine Learning

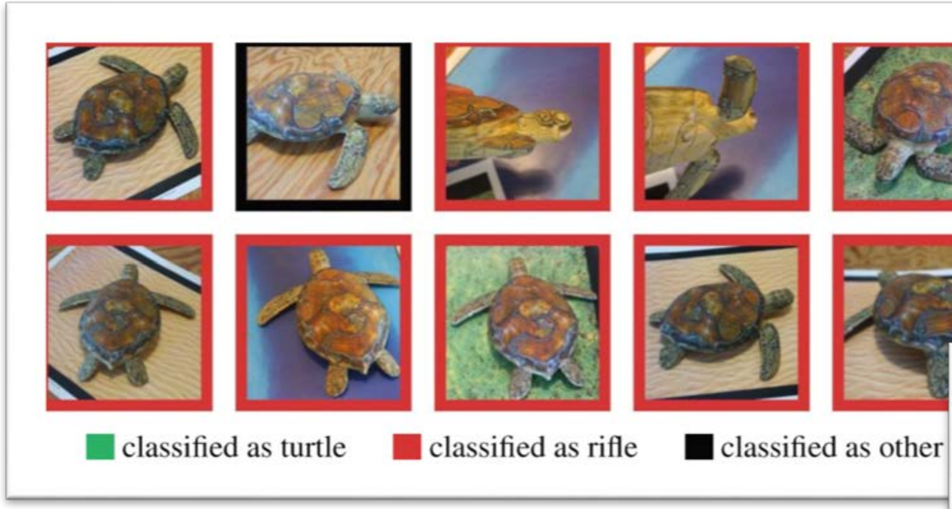
GAME OF ~~THRONES~~

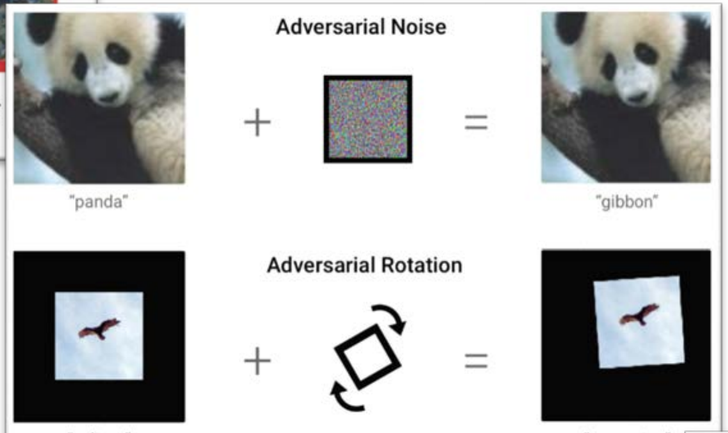
WINTER IS COMING

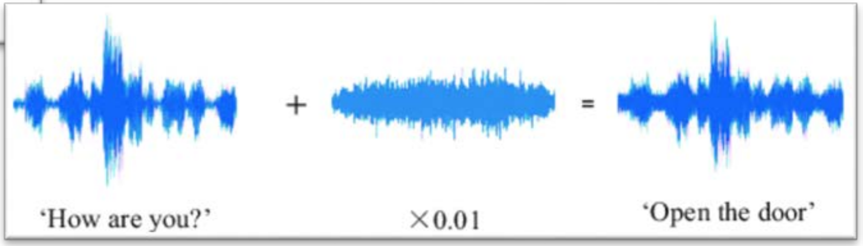
for privacy

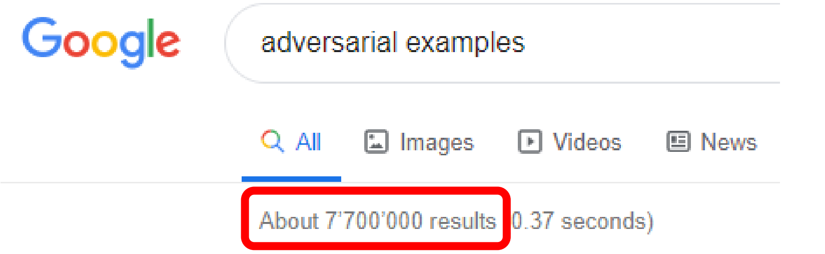
The goal is not to understand, it is to beat!

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



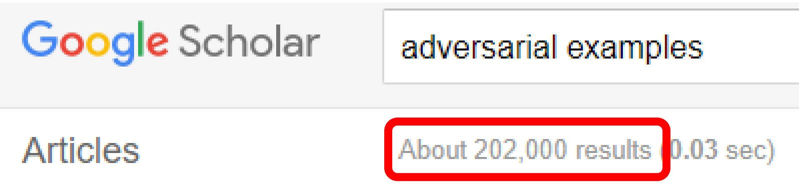






adversarial examples

About 7'700'000 results (0.37 seconds)



adversarial examples

About 202,000 results (0.03 sec)

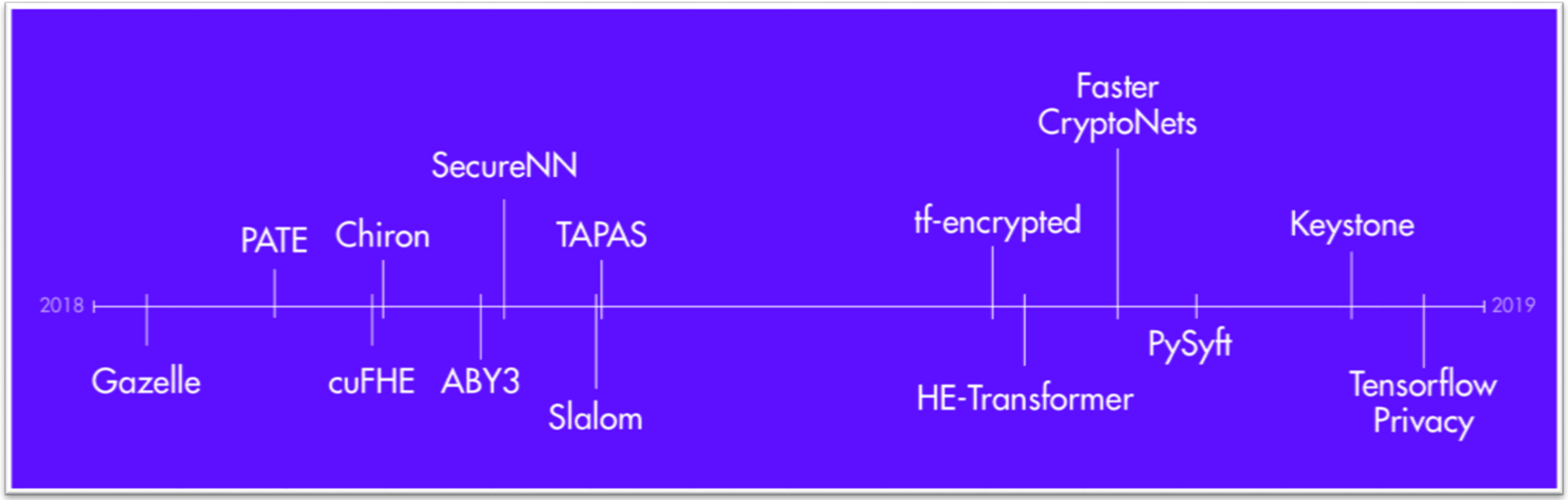
Adversarial examples are only **adversarial** when you are the owner of the algorithm!

Adversarial examples are only **adversarial** when you are the owner of the algorithm!

PETs!!

ML models

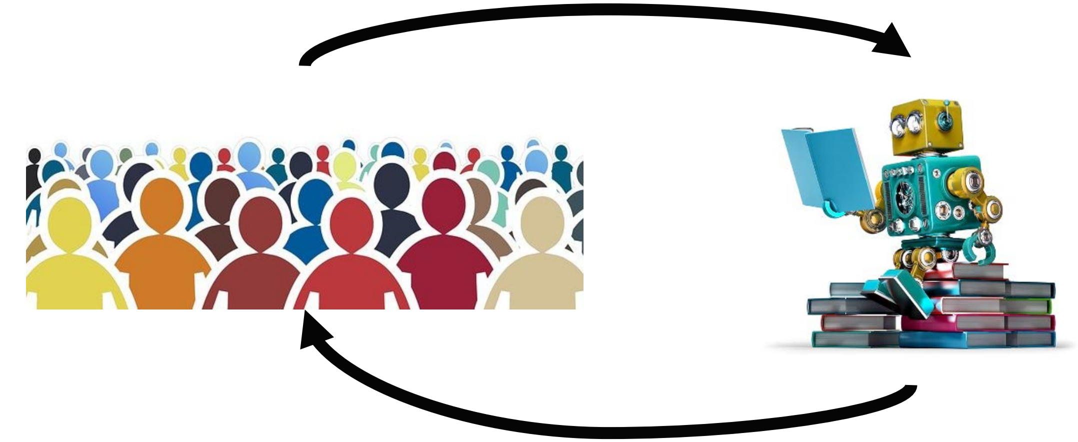
Wait! Why do we need adversarial examples if we have privacy-preserving ML!!



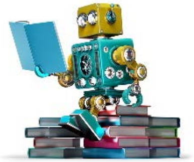
Jason Mancuso, Ben DeCoste and Gavin Uhma.
https://medium.com/dropoutlabs/privacy-preserving-machine-learning-2018-a-year-in-review-b6345a95ae0f

Machine learning as a privacy adversary

ML Privacy-oriented Literature

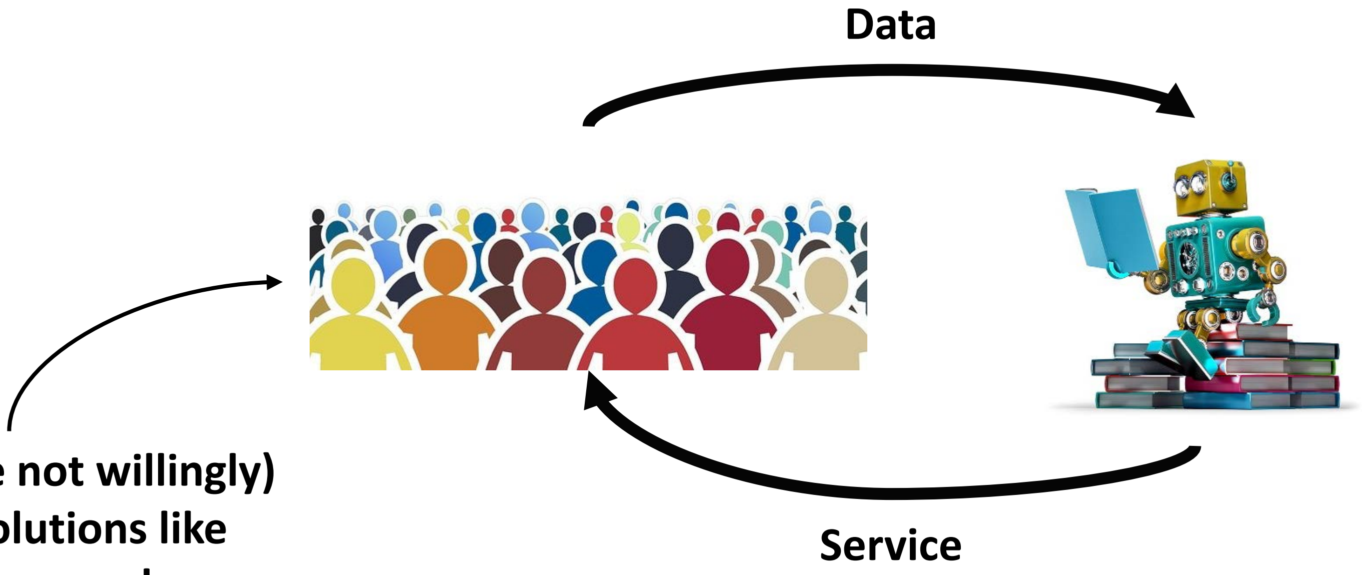


Actively (maybe not willingly) provide data. Solutions like Differential privacy and Encryption are suitable

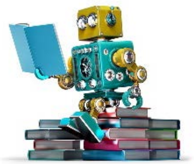
Avoid that  learns about data

Machine learning as a privacy adversary

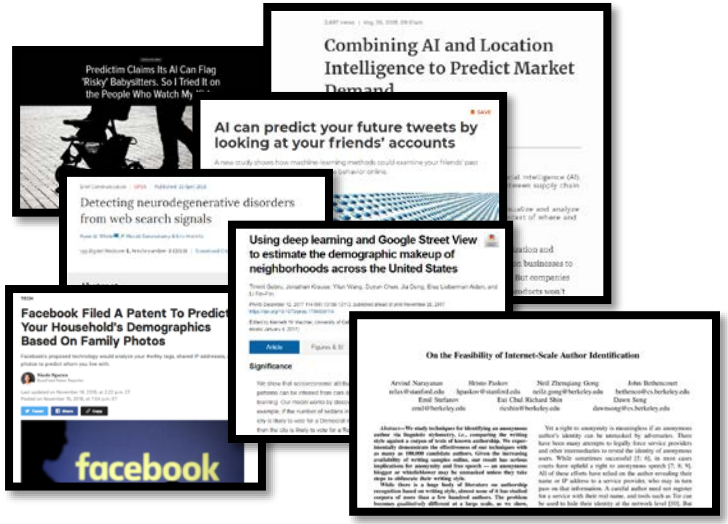
ML Privacy-oriented Literature



Actively (maybe not willingly) provide data. Solutions like Differential privacy and Encryption are suitable

Avoid that  **learns about data**

In this talk



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

Combining AI and Location Intelligence to Predict Market Demand

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

Detecting neurodegenerative disorders from web search signals

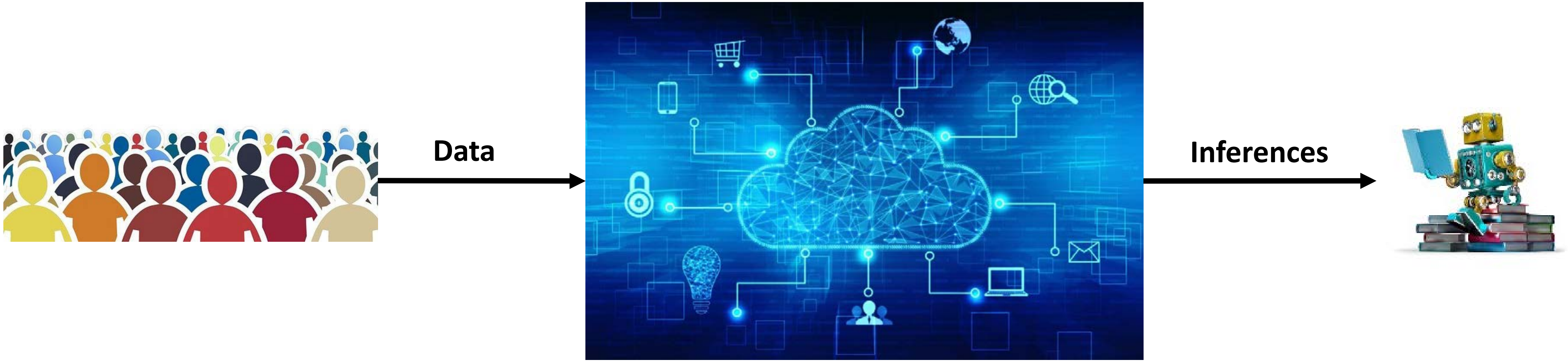
Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States

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

On the Feasibility of Internet-Scale Author Identification

No active sharing!
Cannot count on

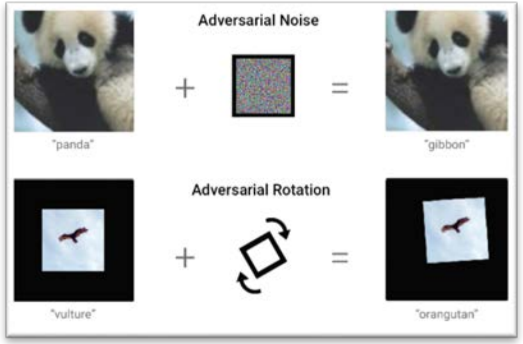
Adversarial examples as privacy defenses



Data

Inferences

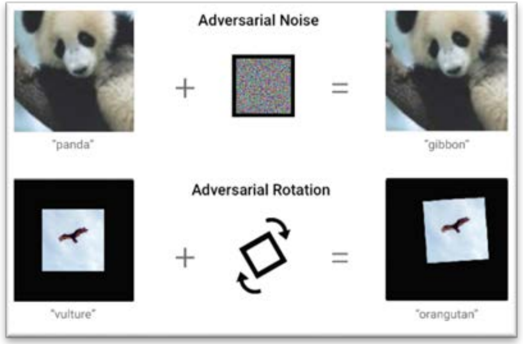
Use ML adversarial example techniques to transform data!



Adversarial examples as privacy defenses

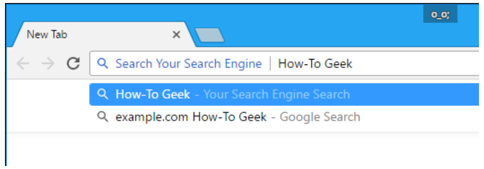
Can this solve all privacy problems?

Use ML adversarial example techniques to transform data!



Can this solve all privacy problems?

Protect web searches from inferences



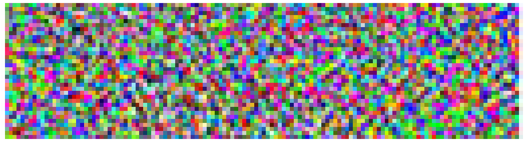
New Tab

Search Your Search Engine | How-To Geek

How-To Geek - Your Search Engine Search

example.com How-To Geek - Google Search

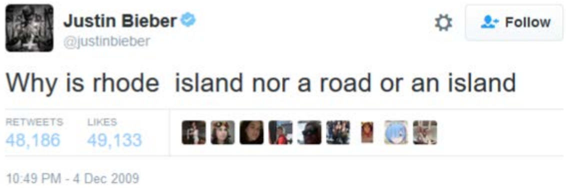
+



=

???

Protect tweets from inferences



Justin Bieber @justinbieber

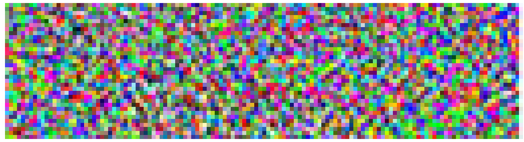
Follow

Why is rhode island nor a road or an island

RETWEETS 48,186 LIKES 49,133

10:49 PM - 4 Dec 2009

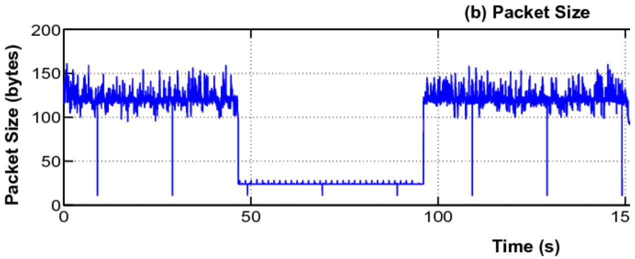
+



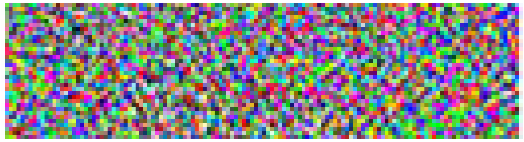
=

???

Protect traffic patterns



+

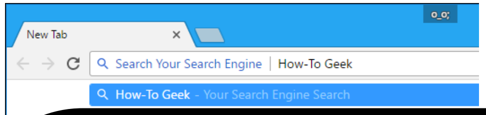


=

???

Can this solve all privacy problems?

Protect web searches from inferences

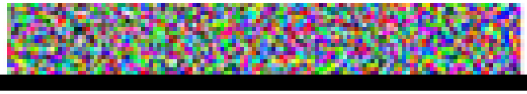


New Tab

Search Your Search Engine | How-To Geek

How-To Geek - Your Search Engine Search

+



=

???

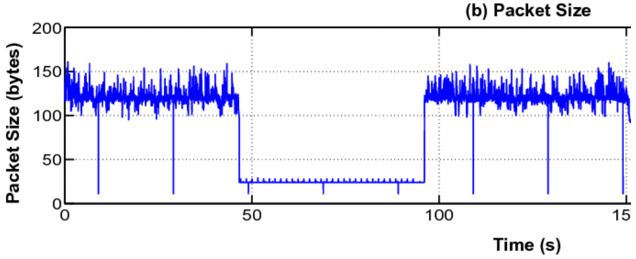
In **privacy problems** adversarial examples belong to a **DISCRETE** and **CONSTRAINED** domain

FEASIBILITY

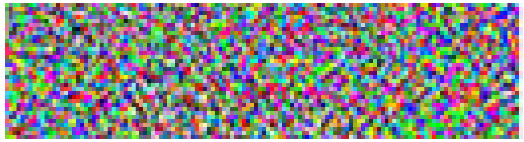
COST

Protect tweets from inferences

Protect traffic patterns



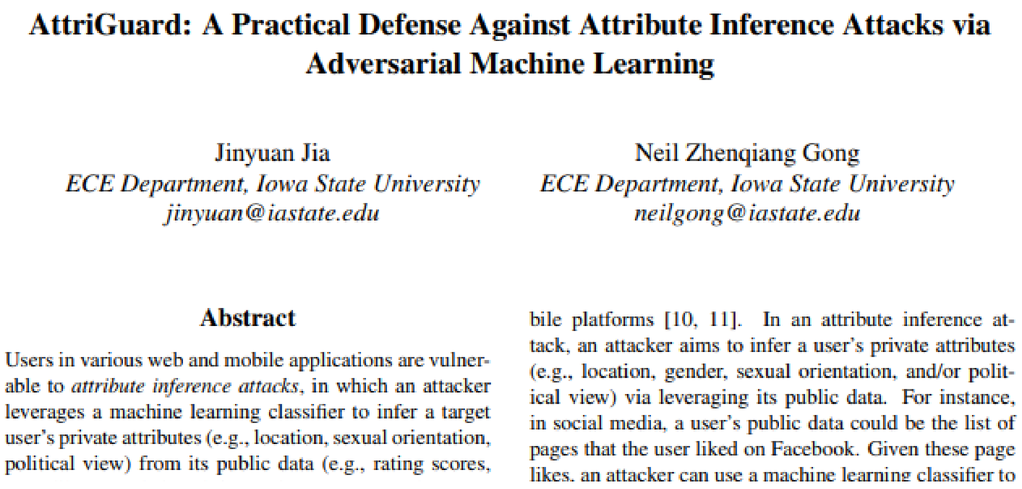
+



=

???

Nobody has thought of this?



AttriGuard: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning

Jinyuan Jia
ECE Department, Iowa State University
jinyuan@iastate.edu

Neil Zhenqiang Gong
ECE Department, Iowa State University
neilgong@iastate.edu

Abstract

Users in various web and mobile applications are vulnerable to *attribute inference attacks*, in which an attacker leverages a machine learning classifier to infer a target user's private attributes (e.g., location, sexual orientation, political view) from its public data (e.g., rating scores,

bile platforms [10, 11]. In an attribute inference attack, an attacker aims to infer a user's private attributes (e.g., location, gender, sexual orientation, and/or political view) via leveraging its public data. For instance, in social media, a user's public data could be the list of pages that the user liked on Facebook. Given these page likes, an attacker can use a machine learning classifier to

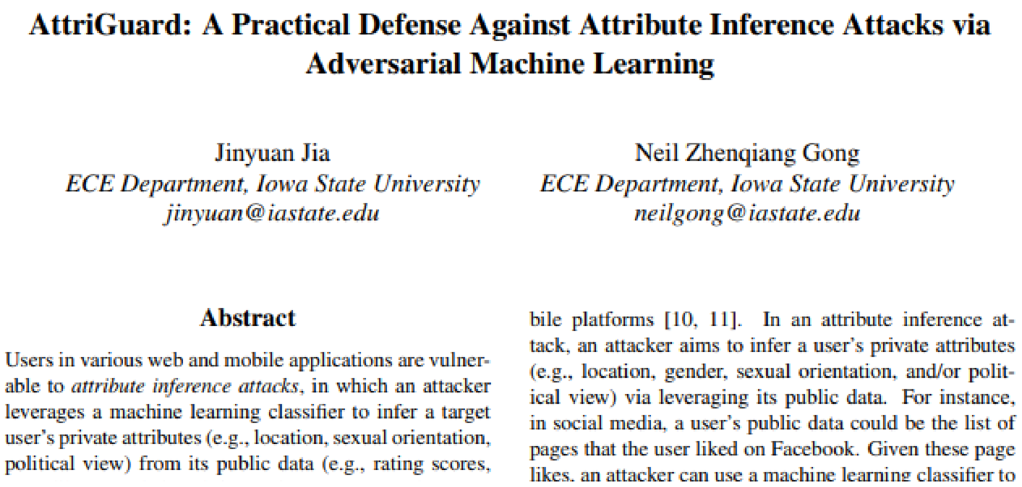
Usenix Security Symposium - 2018

Modify social network attributes to avoid inferences

Use adversarial examples (evasion attacks) to keep utility

Use a version of Jacobian-based Saliency Map Attack (JSMA)
“aware of policies” = only do feasible transformations

Nobody has thought of this?



AttriGuard: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning

Jinyuan Jia
ECE Department, Iowa State University
jinyuan@iastate.edu

Neil Zhenqiang Gong
ECE Department, Iowa State University
neilgong@iastate.edu

Abstract

Users in various web and mobile applications are vulnerable to *attribute inference attacks*, in which an attacker leverages a machine learning classifier to infer a target user's private attributes (e.g., location, sexual orientation, political view) from its public data (e.g., rating scores,

bile platforms [10, 11]. In an attribute inference attack, an attacker aims to infer a user's private attributes (e.g., location, gender, sexual orientation, and/or political view) via leveraging its public data. For instance, in social media, a user's public data could be the list of pages that the user liked on Facebook. Given these page likes, an attacker can use a machine learning classifier to

Usenix Security Symposium - 2018

Modify social network attributes to avoid inferences

Use adversarial examples (evasion attacks) to keep utility

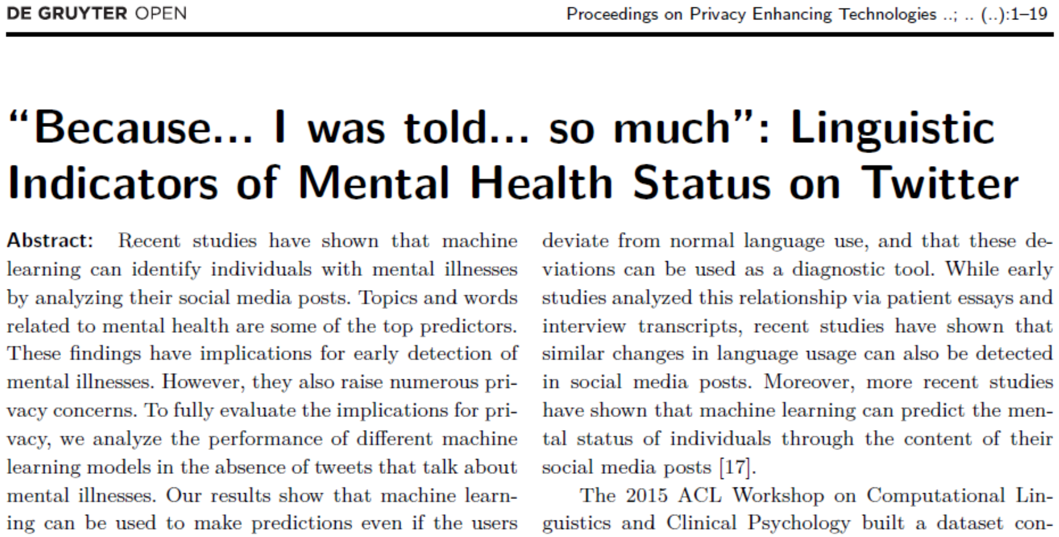
Use a version of Jacobian-based Saliency Map Attack (JSMA)
"aware of policies" = only do feasible transformations

PoPETS - 2019

Modify Twitter line to avoid inferences

Add, remove, replace tweets

Greedy search by importance for classifier



DE GRUYTER OPEN Proceedings on Privacy Enhancing Technologies ..; .. (..):1–19

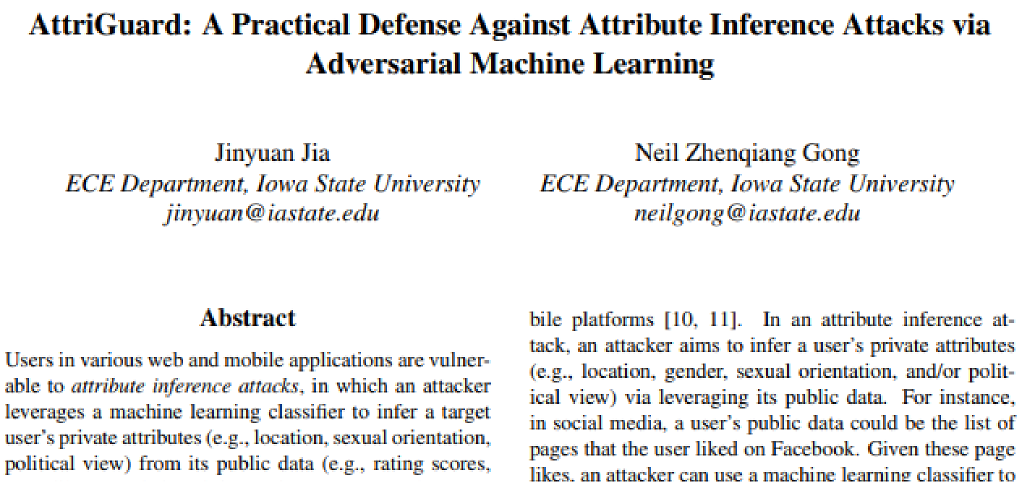
"Because... I was told... so much": Linguistic Indicators of Mental Health Status on Twitter

Abstract: Recent studies have shown that machine learning can identify individuals with mental illnesses by analyzing their social media posts. Topics and words related to mental health are some of the top predictors. These findings have implications for early detection of mental illnesses. However, they also raise numerous privacy concerns. To fully evaluate the implications for privacy, we analyze the performance of different machine learning models in the absence of tweets that talk about mental illnesses. Our results show that machine learning can be used to make predictions even if the users

deviate from normal language use, and that these deviations can be used as a diagnostic tool. While early studies analyzed this relationship via patient essays and interview transcripts, recent studies have shown that similar changes in language usage can also be detected in social media posts. Moreover, more recent studies have shown that machine learning can predict the mental status of individuals through the content of their social media posts [17].

The 2015 ACL Workshop on Computational Linguistics and Clinical Psychology built a dataset con-

Nobody has thought of this?



AttriGuard: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning

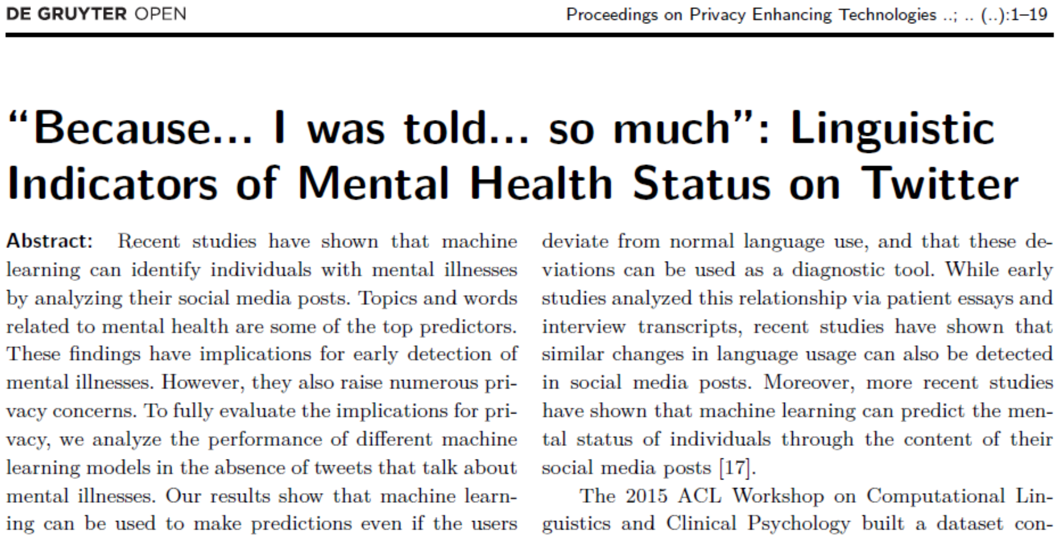
Jinyuan Jia
ECE Department, Iowa State University
jinyuan@iastate.edu

Neil Zhenqiang Gong
ECE Department, Iowa State University
neilgong@iastate.edu

Abstract

Users in various web and mobile applications are vulnerable to *attribute inference attacks*, in which an attacker leverages a machine learning classifier to infer a target user's private attributes (e.g., location, sexual orientation, political view) from its public data (e.g., rating scores,

bile platforms [10, 11]. In an attribute inference attack, an attacker aims to infer a user's private attributes (e.g., location, gender, sexual orientation, and/or political view) via leveraging its public data. For instance, in social media, a user's public data could be the list of pages that the user liked on Facebook. Given these page likes, an attacker can use a machine learning classifier to



DE GRUYTER OPEN Proceedings on Privacy Enhancing Technologies ..; .. (..):1–19

"Because... I was told... so much": Linguistic Indicators of Mental Health Status on Twitter

Abstract: Recent studies have shown that machine learning can identify individuals with mental illnesses by analyzing their social media posts. Topics and words related to mental health are some of the top predictors. These findings have implications for early detection of mental illnesses. However, they also raise numerous privacy concerns. To fully evaluate the implications for privacy, we analyze the performance of different machine learning models in the absence of tweets that talk about mental illnesses. Our results show that machine learning can be used to make predictions even if the users

deviate from normal language use, and that these deviations can be used as a diagnostic tool. While early studies analyzed this relationship via patient essays and interview transcripts, recent studies have shown that similar changes in language usage can also be detected in social media posts. Moreover, more recent studies have shown that machine learning can predict the mental status of individuals through the content of their social media posts [17].

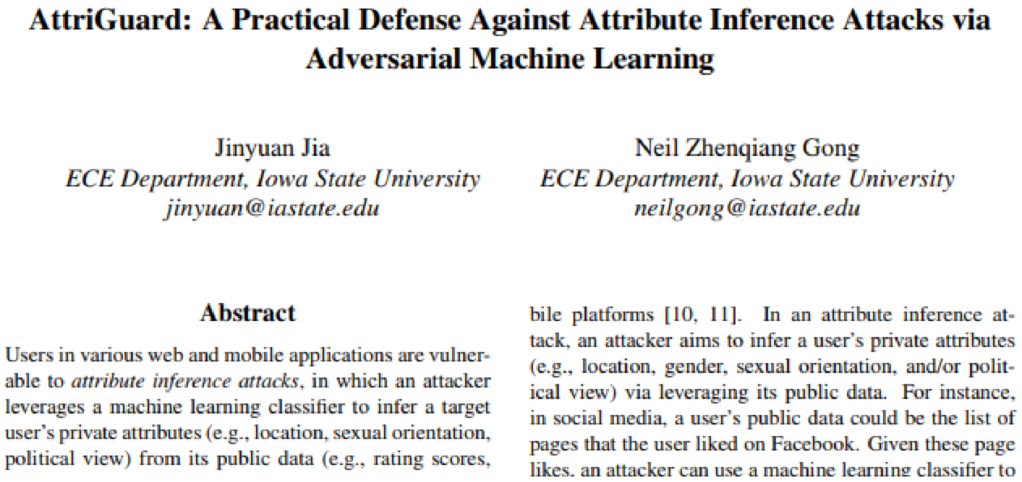
The 2015 ACL Workshop on Computational Linguistics and Clinical Psychology built a dataset con-

Non-privacy constrained applications

Text:
Goal: change classification (positive to negative sentiment,
change inferred topic for a post)

Malware:
Goal: change classification (from malicious to benign)

Nobody has thought of this?



AttriGuard: A Practical Defense Against Attribute Inference Attacks via Adversarial Machine Learning

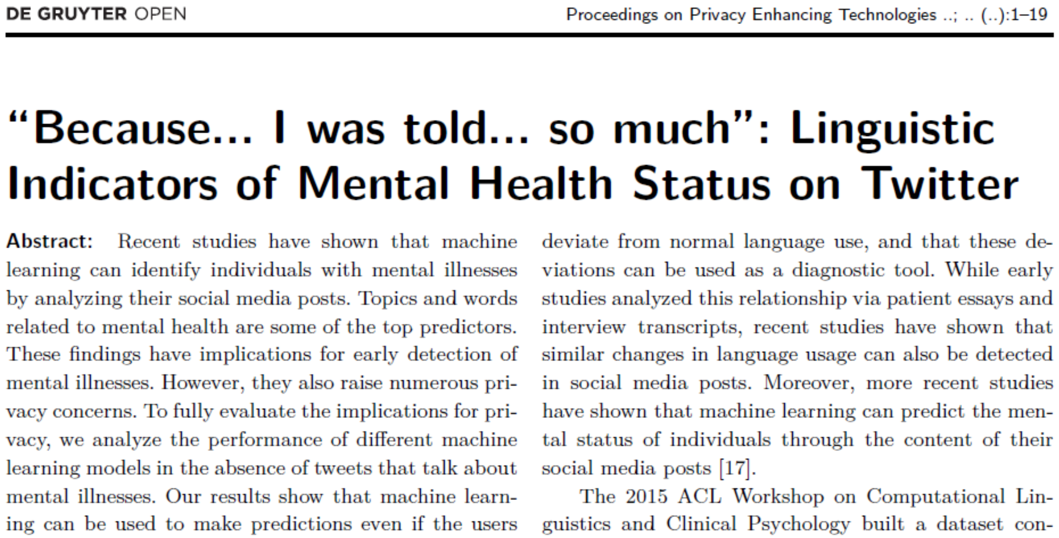
Jinyuan Jia
ECE Department, Iowa State University
jinyuan@iastate.edu

Neil Zhenqiang Gong
ECE Department, Iowa State University
neilgong@iastate.edu

Abstract

Users in various web and mobile applications are vulnerable to *attribute inference attacks*, in which an attacker leverages a machine learning classifier to infer a target user's private attributes (e.g., location, sexual orientation, political view) from its public data (e.g., rating scores,

bile platforms [10, 11]. In an attribute inference attack, an attacker aims to infer a user's private attributes (e.g., location, gender, sexual orientation, and/or political view) via leveraging its public data. For instance, in social media, a user's public data could be the list of pages that the user liked on Facebook. Given these page likes, an attacker can use a machine learning classifier to



DE GRUYTER OPEN   Proceedings on Privacy Enhancing Technologies ..; .. (..):1–19

"Because... I was told... so much": Linguistic Indicators of Mental Health Status on Twitter

Abstract: Recent studies have shown that machine learning can identify individuals with mental illnesses by analyzing their social media posts. Topics and words related to mental health are some of the top predictors. These findings have implications for early detection of mental illnesses. However, they also raise numerous privacy concerns. To fully evaluate the implications for privacy, we analyze the performance of different machine learning models in the absence of tweets that talk about mental illnesses. Our results show that machine learning can be used to make predictions even if the users

deviate from normal language use, and that these deviations can be used as a diagnostic tool. While early studies analyzed this relationship via patient essays and interview transcripts, recent studies have shown that similar changes in language usage can also be detected in social media posts. Moreover, more recent studies have shown that machine learning can predict the mental status of individuals through the content of their social media posts [17].

The 2015 ACL Workshop on Computational Linguistics and Clinical Psychology built a dataset con-

Non-privacy constrained applications

Text:
Goal: change classification (positive to negative sentiment, change inferred topic for a post)

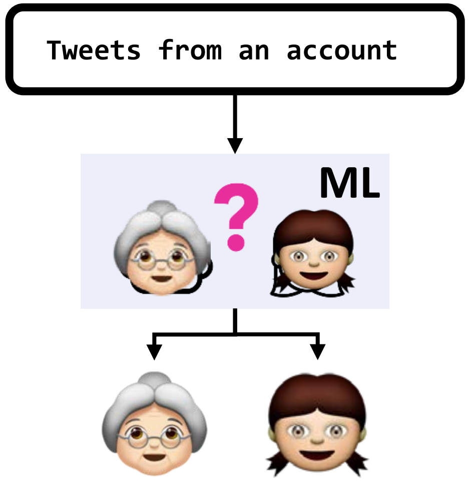
Malware:
Goal: change classification (from malicious to benign)

Repeated patterns:

- **Model transformation**
- **Find new search algorithm**
  e.g., Hill climbing, beam search
- **Evaluate & compare performance**

But NO systematic design method ☹

Our proposal: Evasion as a graph



Protecting users from demographic inferences

Goal change Twitter line classification regarding age

Transformations

Use synonyms ←

Introduce typos

Change punctuation

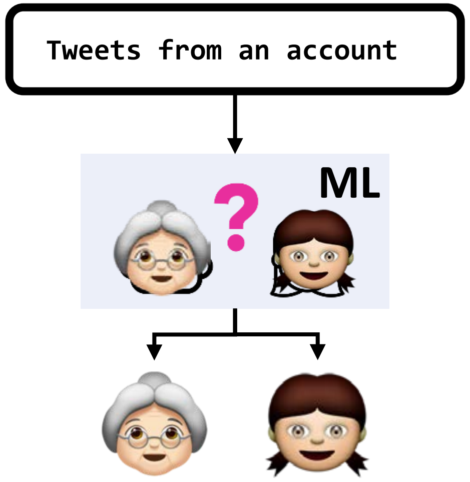
Cost

Keep the meaning!

Our proposal: Evasion as a graph

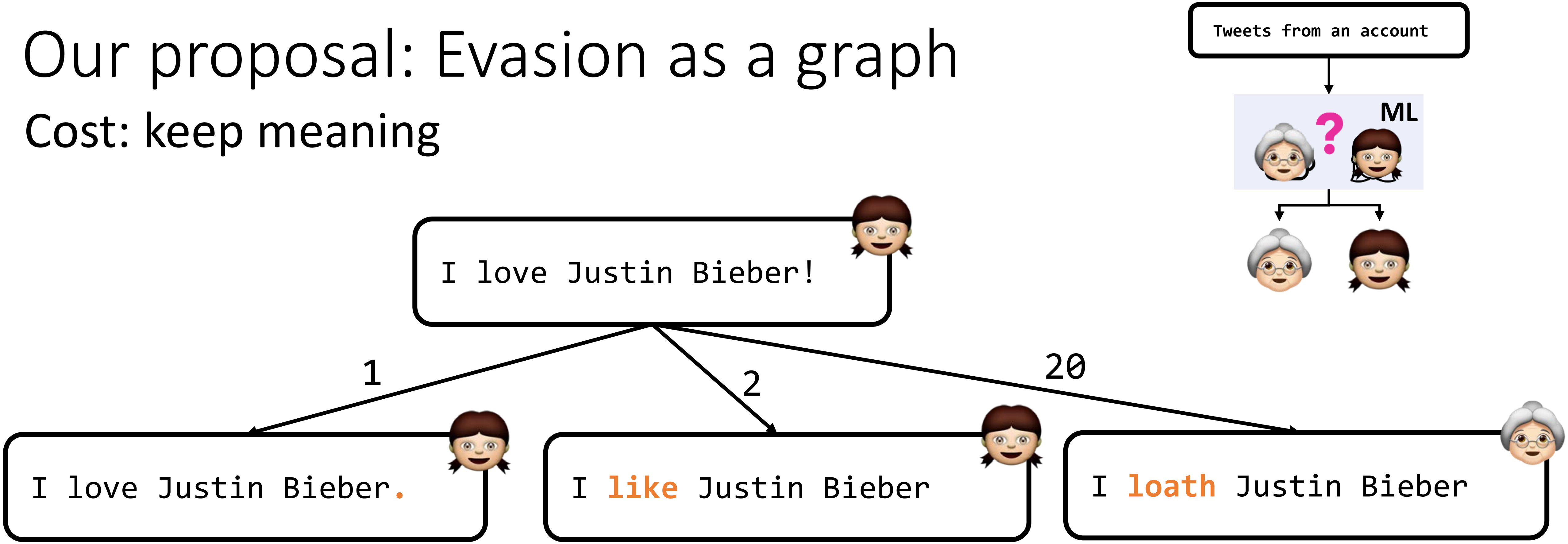
Cost: keep meaning

I love Justin Bieber!



Our proposal: Evasion as a graph

Cost: keep meaning



Tweets from an account

I love Justin Bieber!

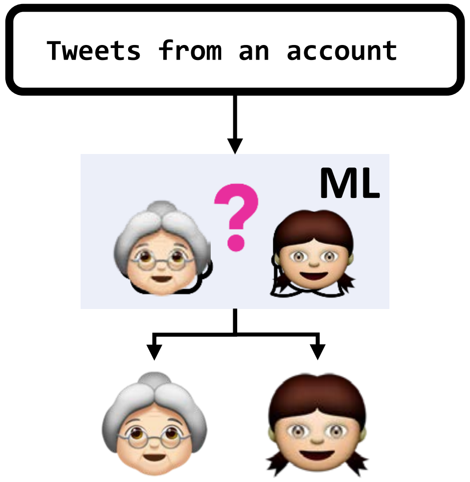
I love Justin Bieber.

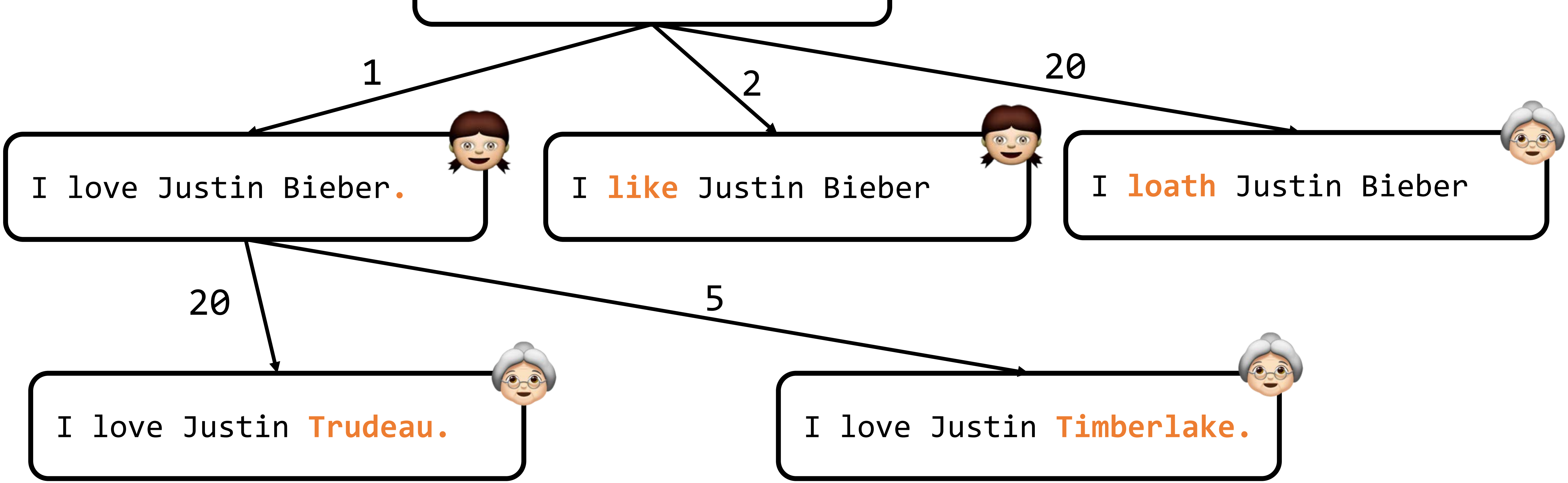
I like Justin Bieber

I loath Justin Bieber

Our proposal: Evasion as a graph

Cost: keep meaning



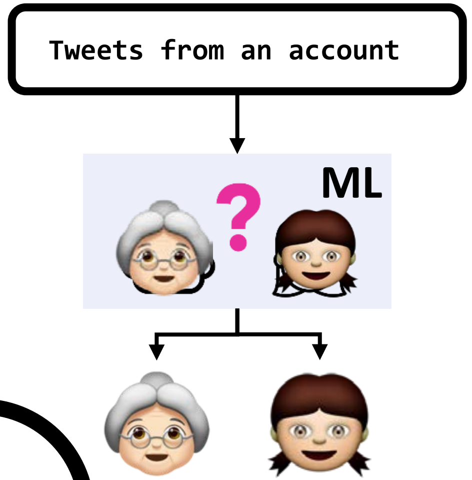


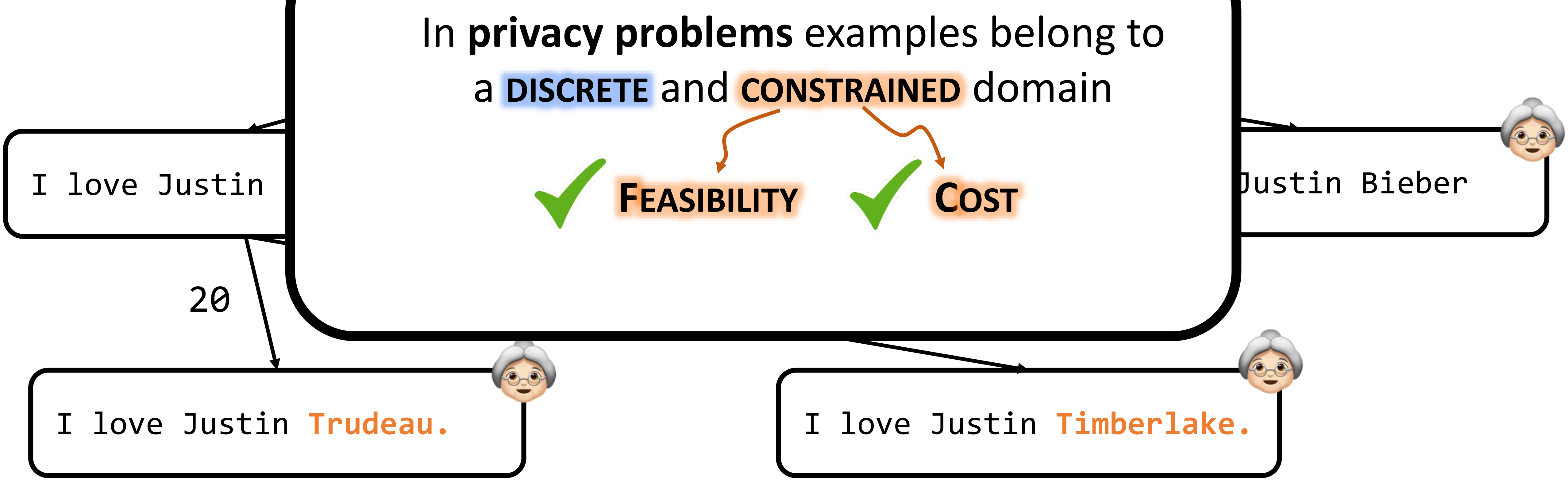
cost = 1 + 20 = 21

cost = 1 + 5 = 6

Our proposal: Evasion as a graph

Cost: keep meaning





In **privacy problems** examples belong to a **DISCRETE** and **CONSTRAINED** domain

✓ **FEASIBILITY** ✓ **COST**

cost = 1 + 20 = 21

cost = 1 + 5 = 6

The graph approach comes with advantages

✓ **Enables the use of graph theory to**

EFFICIENTLY find adversarial examples (A*, beam search, hill climbing, etc)

CAPTURES most attacks in the literature! (comparison base)

✓ **Finds provable MINIMAL COST adversarial examples (A*) if**

- The discrete domain is a subset of R^m

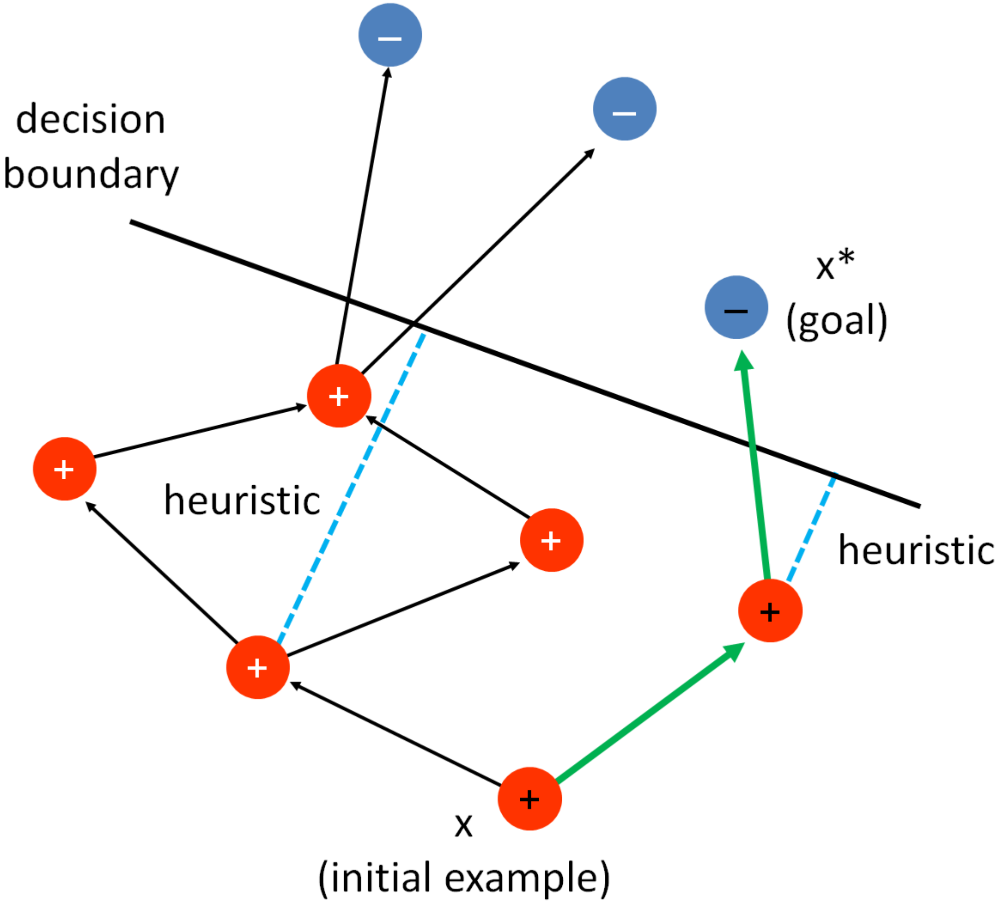
For example, categorical one-hot encoded features: [0 1 0 0]

- Cost of each single transformation is L^p

For example, L^∞([0 1 0 0], [1 0 0 0]) = 1

- We can compute pointwise robustness for the target classifier over R^m

Finding minimal cost adv. examples: the concept



$$x^* = \arg\min_{x' \in \mathbb{X}} C(x, x') \text{ s.t. } \mathsf{goal}(x') = \top,$$

$$\mathsf{goal}(x') = \begin{cases} \top, & t = 1 \text{ and } \sigma(f(x')) > l \\ \top, & t = 0 \text{ and } \sigma(f(x')) \leq 1 - l \\ \bot, & \text{otherwise} \end{cases}$$

Confidence of the example

Adversarial examples for privacy

✓ **Provide privacy in settings where the ML model is adversarial and not cooperative**

✓ **Privacy is CONSTRAINED , a graphical approach can be used to**

EFFICIENTLY find FEASIBLE adversarial examples

find MINIMAL COST adversarial examples

✓ **Even if they cannot be deployed in practice, this approach provides a BASELINE to compare defenses' efficiency**

Bonus: applicable to security problems!

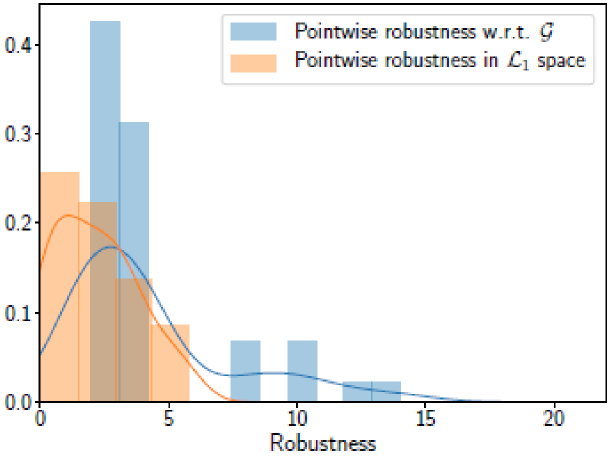
MINIMAL COST adversarial examples can become security metrics!

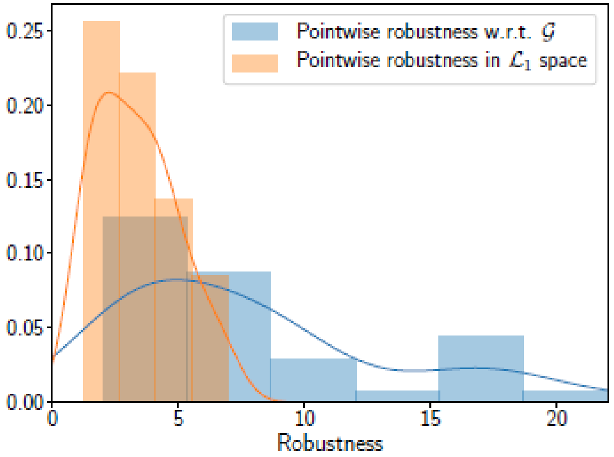
Cost can be associated with RISK

Cannot stop attacks, but can we ensure they are expensive?

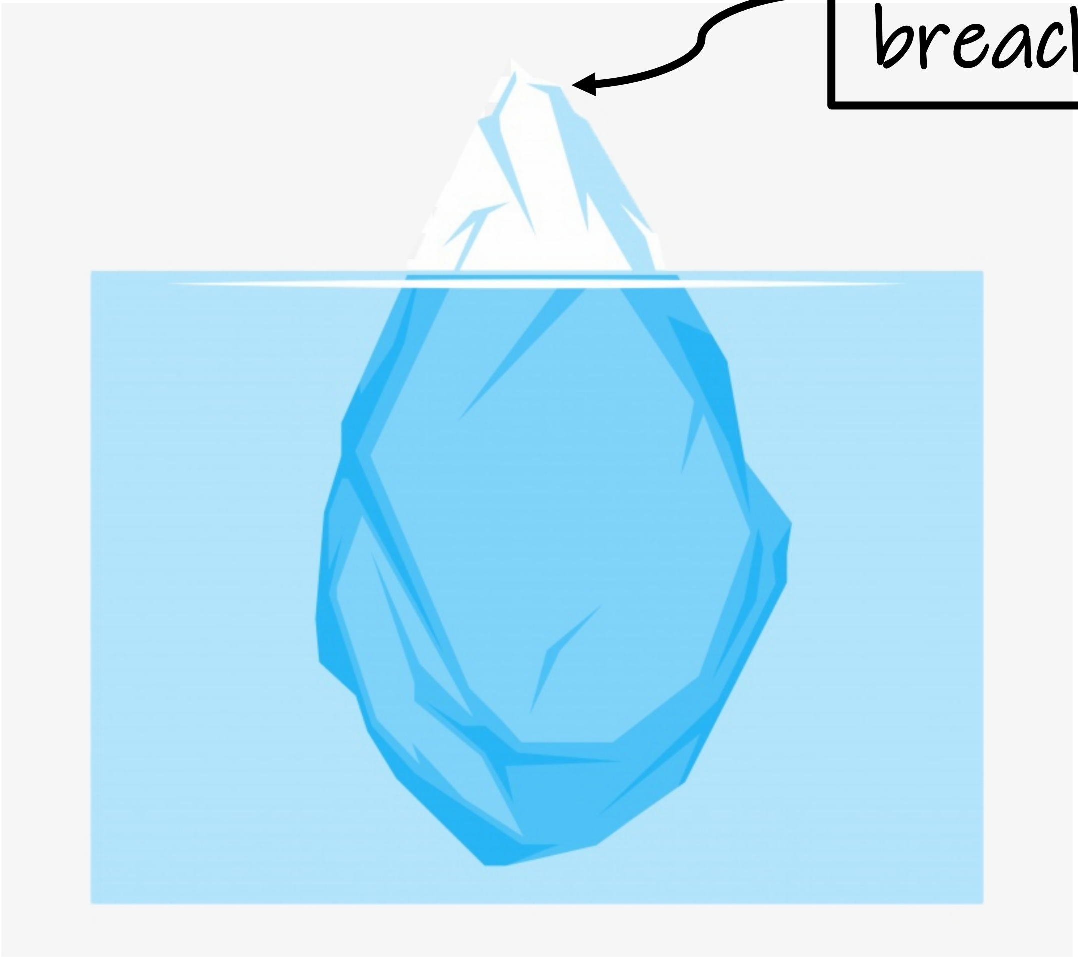
Constrained domains security

Continuous-domains approaches can be very conservative!



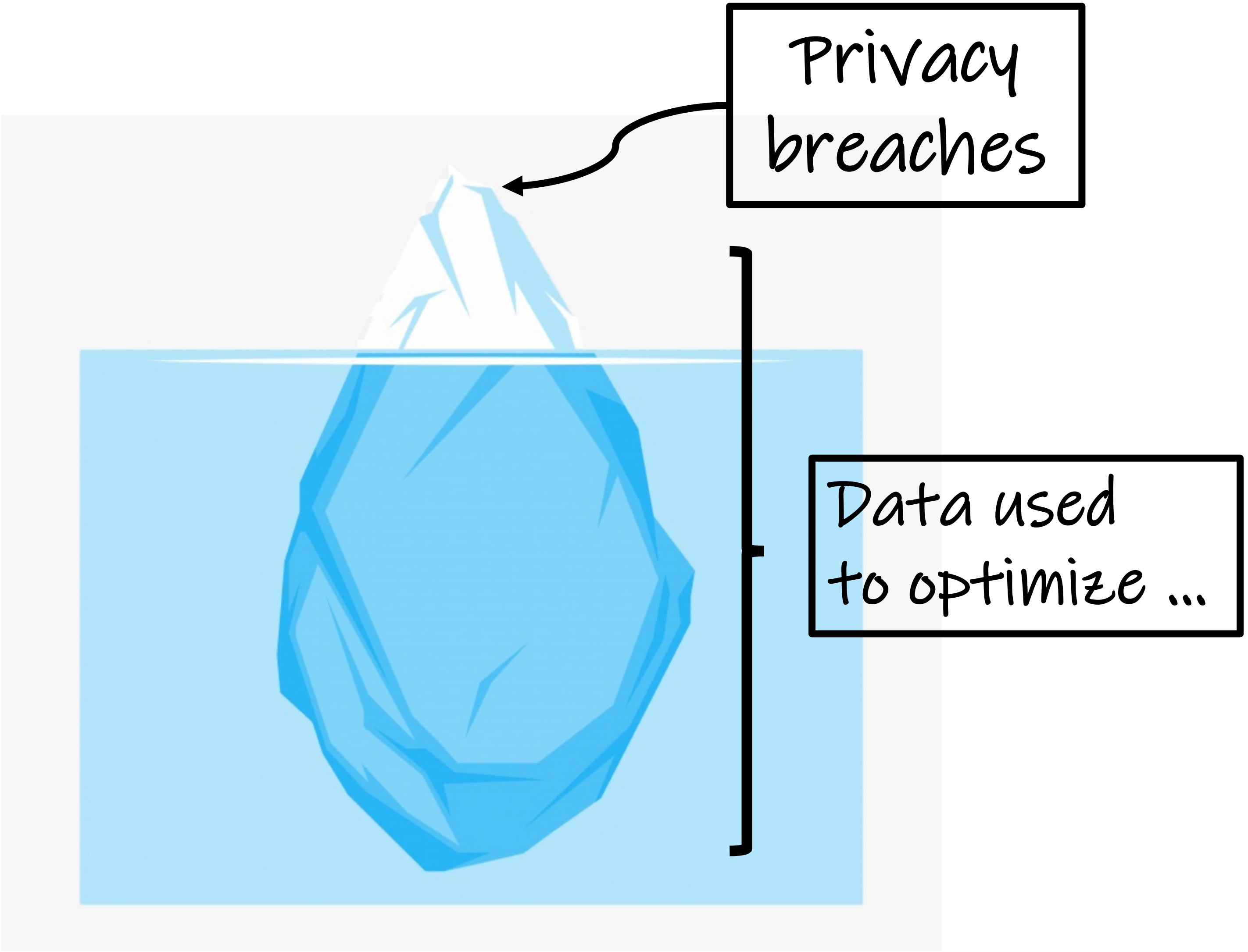


Only privacy is at stake?



Privacy breaches

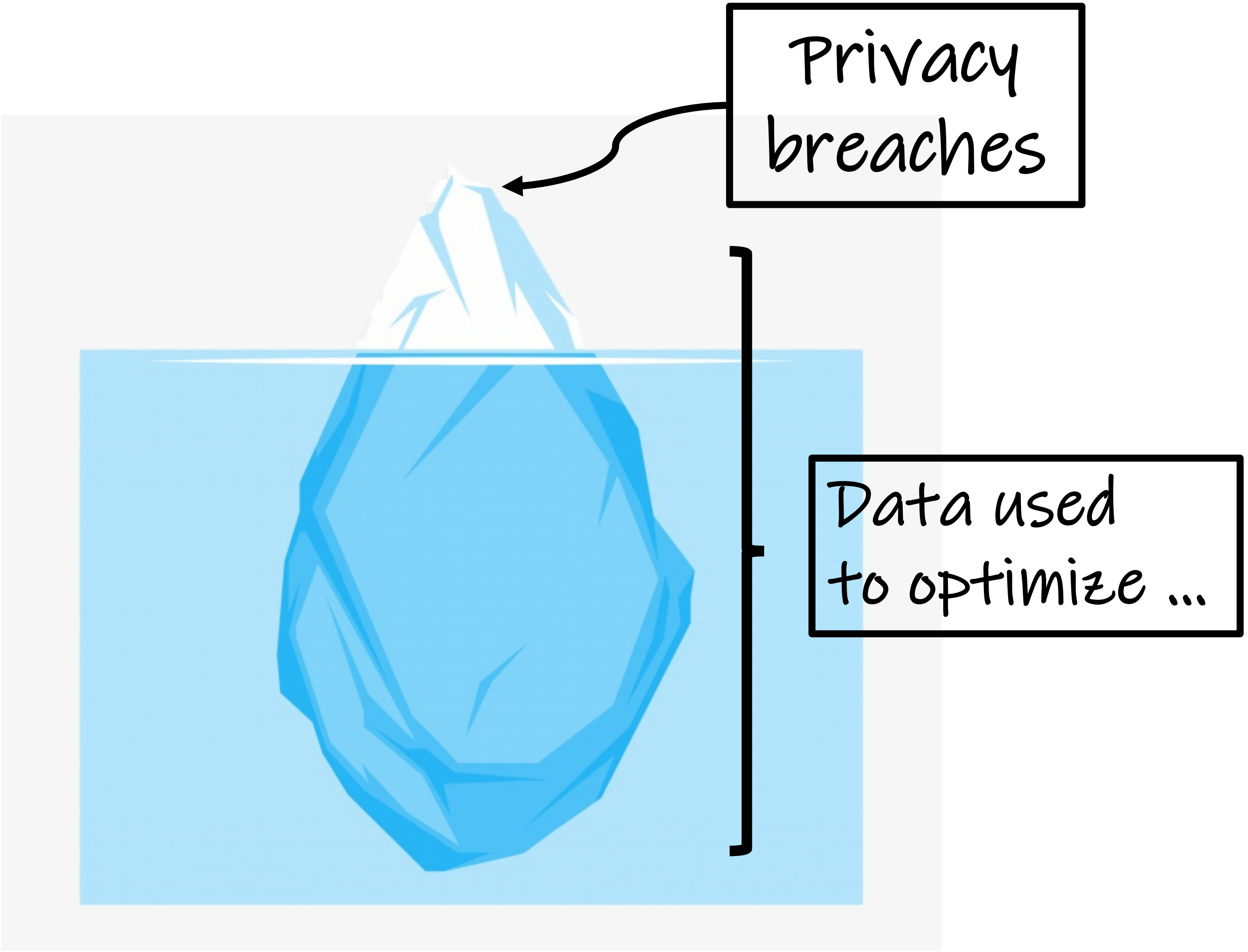
Only privacy is at stake?



Privacy breaches

Data used to optimize …

Only privacy is at stake?



Privacy breaches

Data used to optimize ...

Prevalent use of optimization algorithms to extract maximum economic value from the manipulation of people's activities and their environment

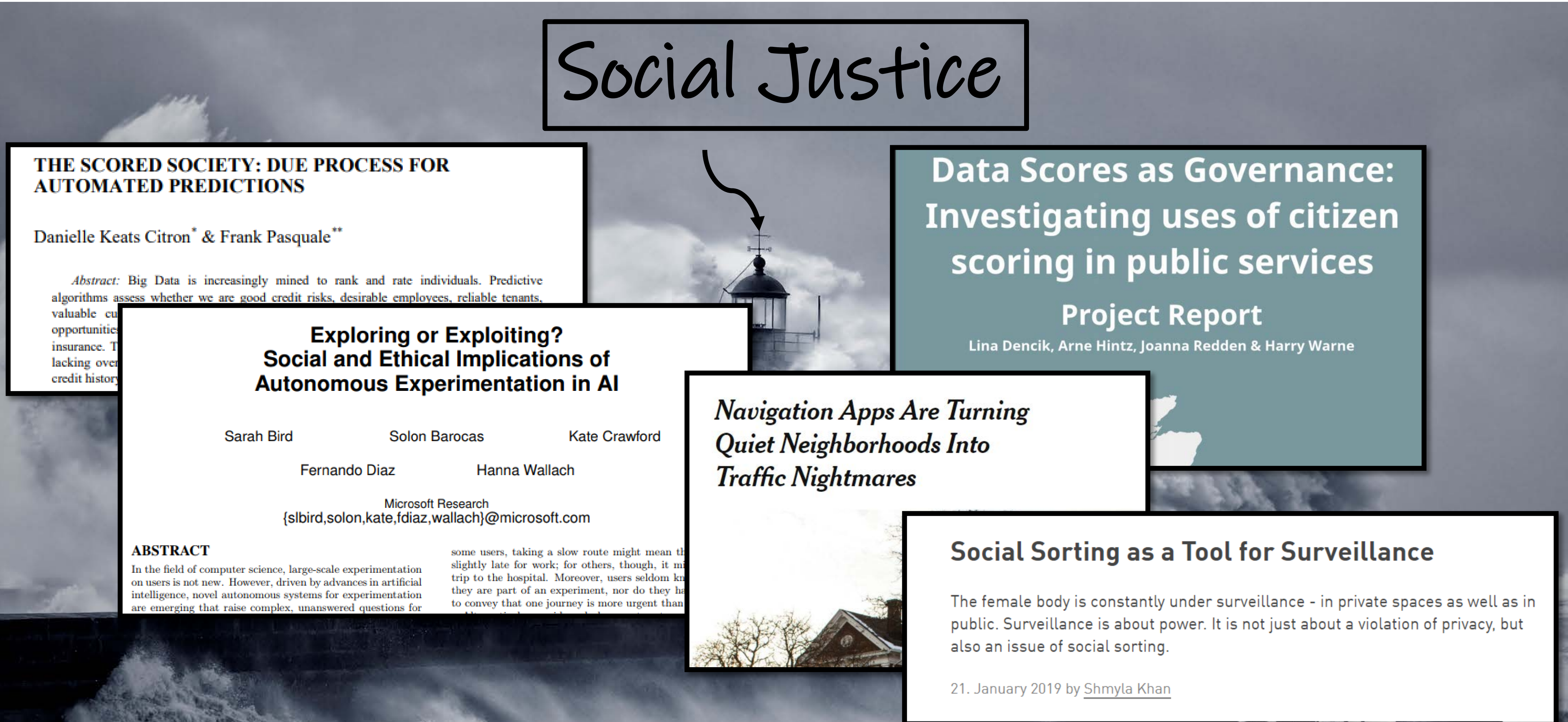
Advertisement
(e.g., Facebook ads)

Routing
(e.g., Waze)

FICO™

Credit scoring
(e.g., FICO)

The ML tsunami on Social Justice



Social Justice

THE SCORED SOCIETY: DUE PROCESS FOR AUTOMATED PREDICTIONS

Danielle Keats Citron[*] & Frank Pasquale[**]

Abstract: Big Data is increasingly mined to rank and rate individuals. Predictive algorithms assess whether we are good credit risks, desirable employees, reliable tenants, valuable cu
opportunities
insurance. T
lacking over
credit history

Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI

Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, Hanna Wallach

Microsoft Research
{slbird,solon,kate,fdiaz,wallach}@microsoft.com

ABSTRACT

In the field of computer science, large-scale experimentation on users is not new. However, driven by advances in artificial intelligence, novel autonomous systems for experimentation are emerging that raise complex, unanswered questions for

some users, taking a slow route might mean th
slightly late for work; for others, though, it mi
trip to the hospital. Moreover, users seldom kn
they are part of an experiment, nor do they ha
to convey that one journey is more urgent than

Data Scores as Governance: Investigating uses of citizen scoring in public services

Project Report

Lina Dencik, Arne Hintz, Joanna Redden & Harry Warne

Navigation Apps Are Turning Quiet Neighborhoods Into Traffic Nightmares

Social Sorting as a Tool for Surveillance

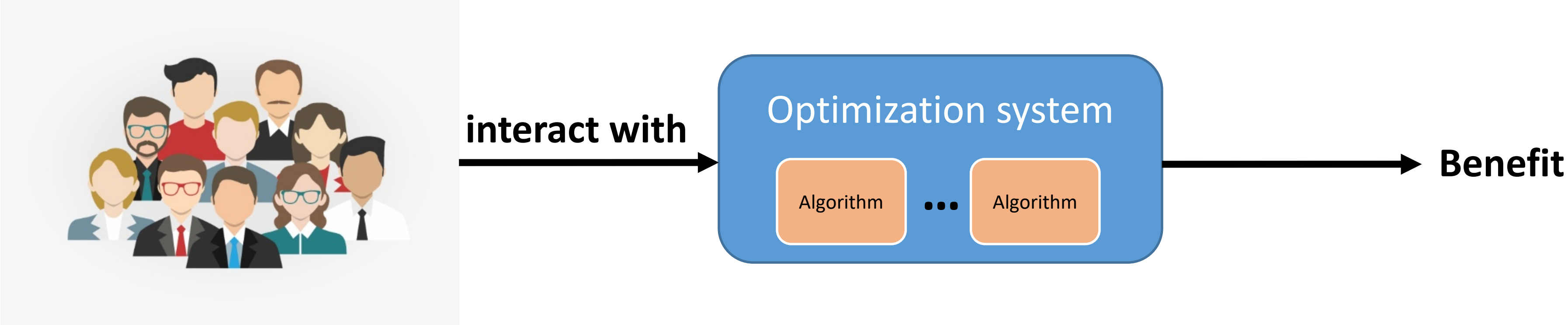
The female body is constantly under surveillance - in private spaces as well as in public. Surveillance is about power. It is not just about a violation of privacy, but also an issue of social sorting.

21. January 2019 by Shmyla Khan

The ML tsunami on Social Justice

Social Justice

Optimization Systems



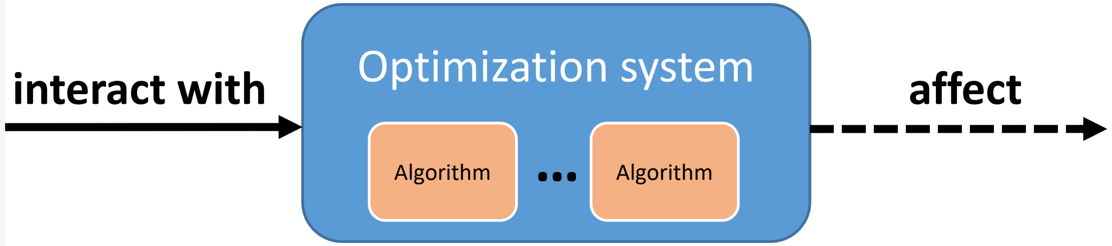
interact with

Optimization system

Algorithm ••• Algorithm

Benefit

Optimization Systems



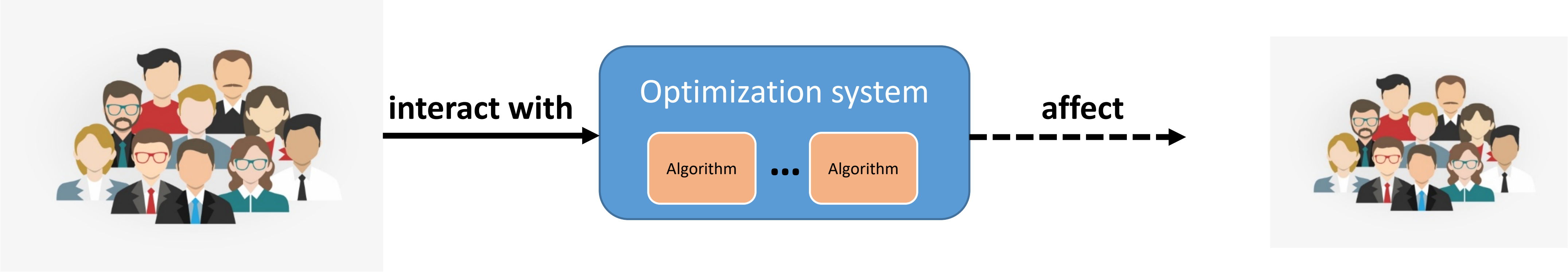
interact with

Optimization system

Algorithm ••• Algorithm

affect

Optimization Systems



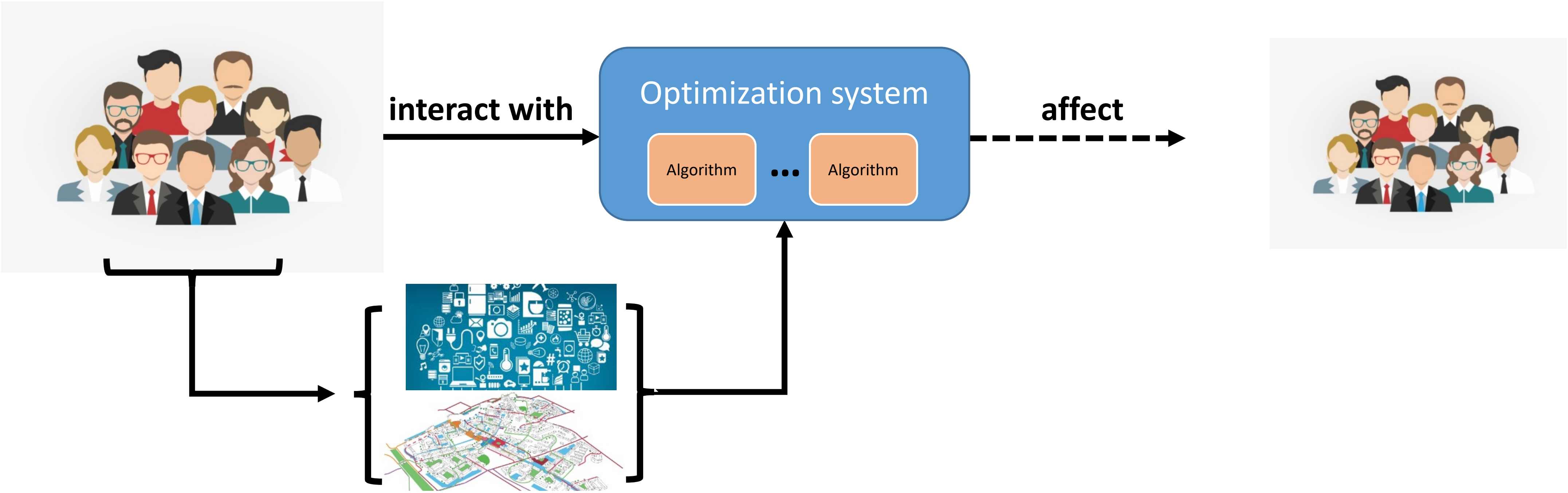
interact with

Optimization system

Algorithm ••• Algorithm

affect

Optimization Systems



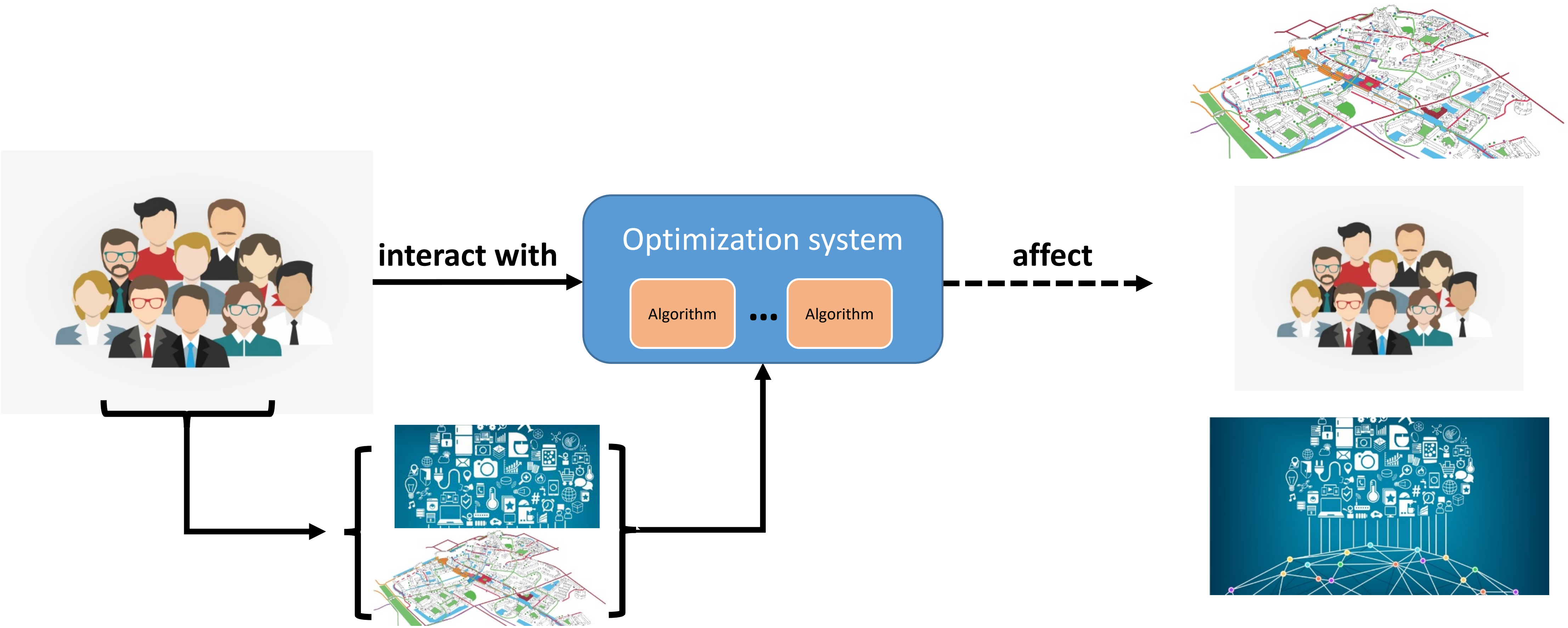
interact with

Optimization system

Algorithm ••• Algorithm

affect

Optimization Systems



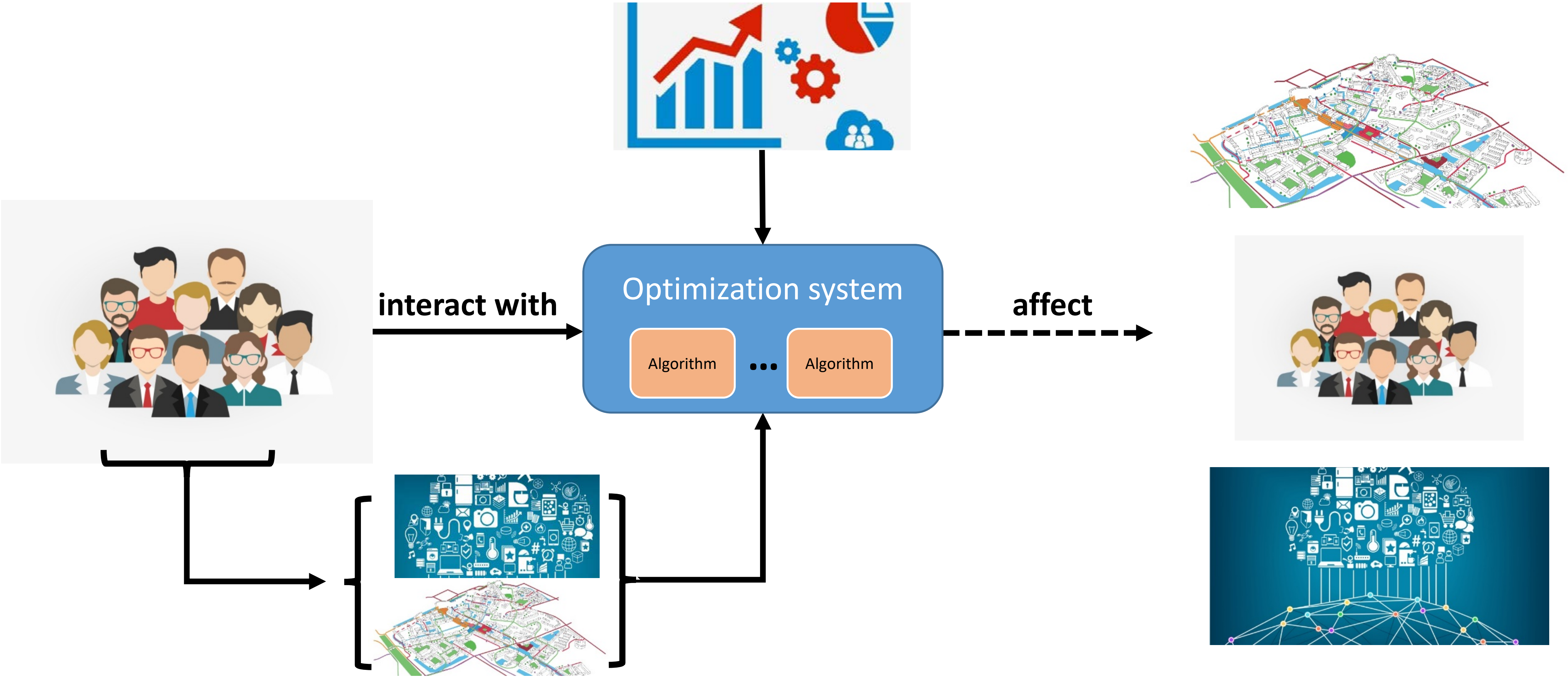
interact with

Optimization system

Algorithm ••• Algorithm

affect

Optimization Systems



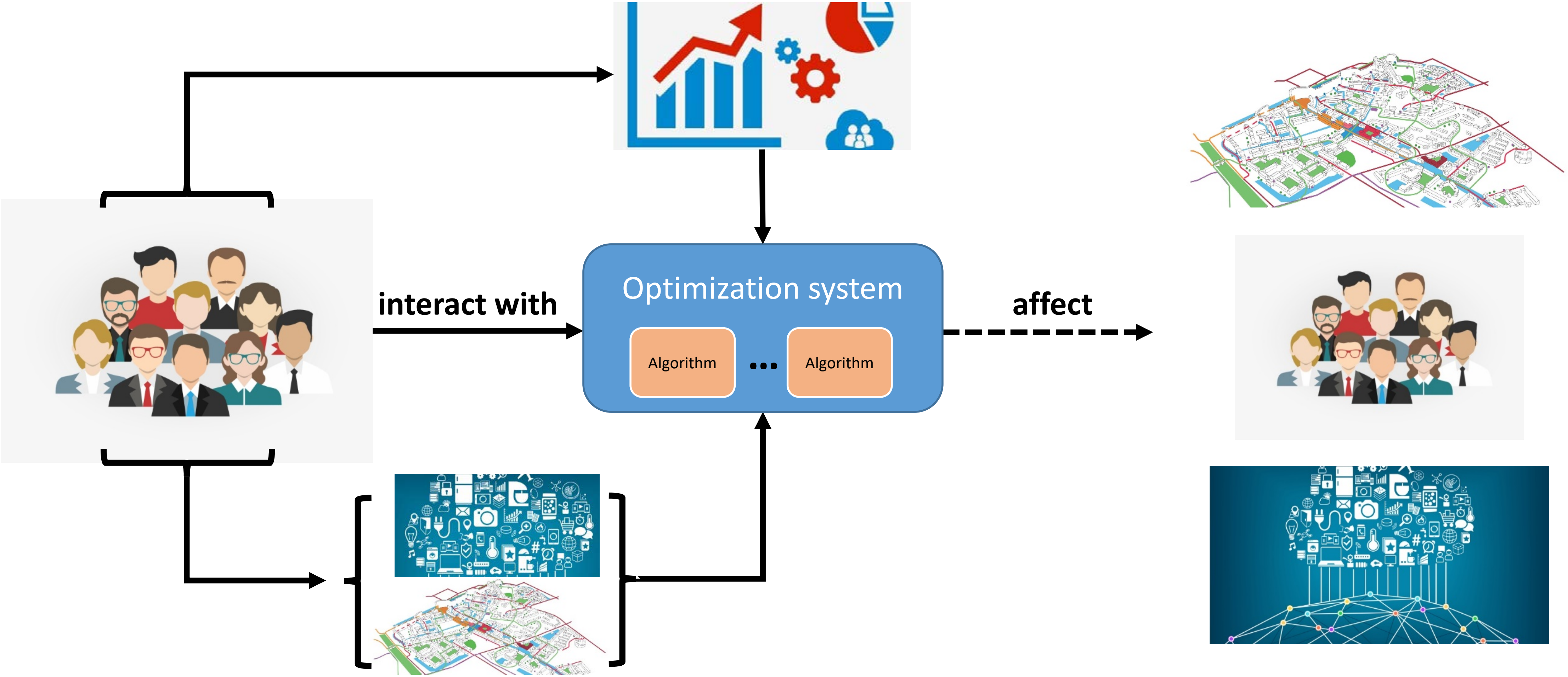
interact with

Optimization system

Algorithm ... Algorithm

affect

Optimization Systems



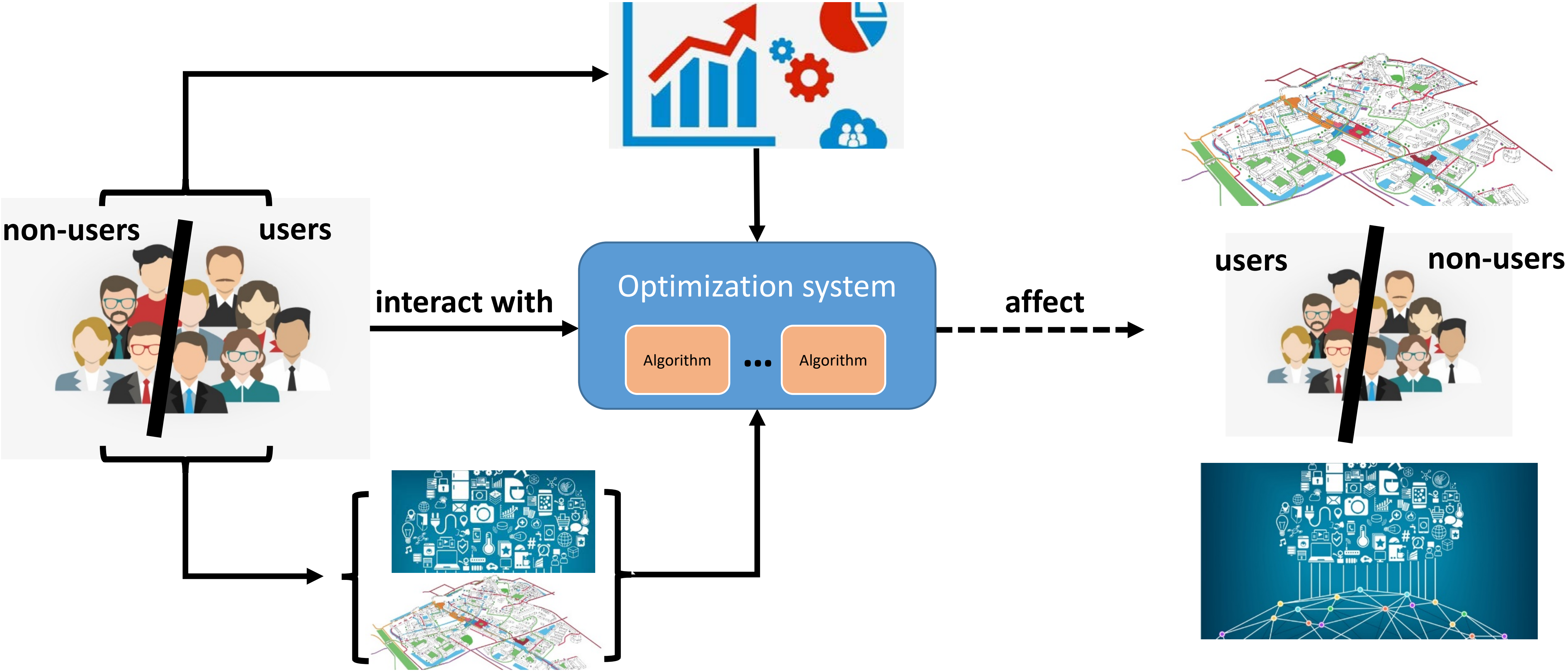
interact with

Optimization system

Algorithm ... Algorithm

affect

Optimization Systems



non-users users

interact with

Optimization system

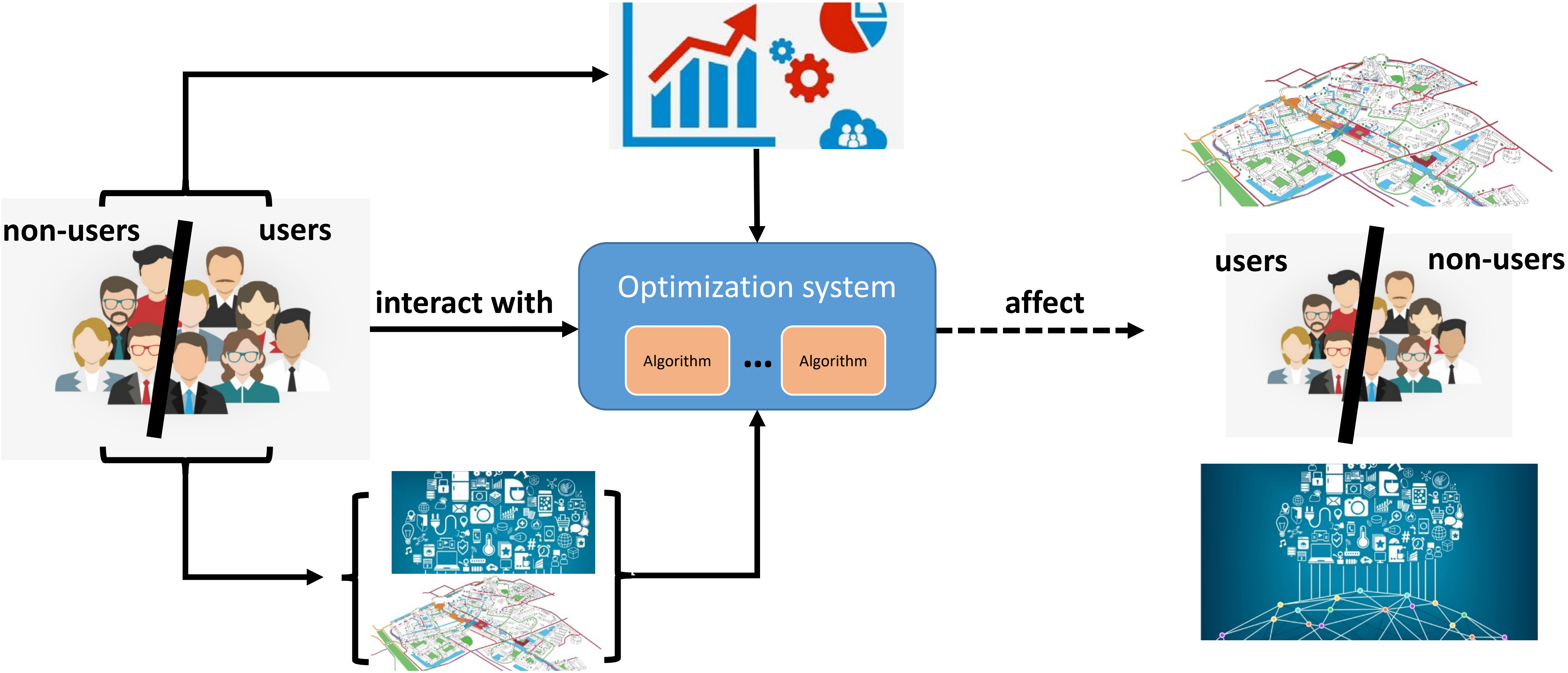
Algorithm ... Algorithm

affect

users non-users

Optimization Systems

How do we avoid negative effects caused by the Optimization System?
(direct and externalities)



non-users **users**

interact with

Optimization system

Algorithm **...** Algorithm

affect

users **non-users**

We have fairness research!!

“We’re creating algorithms that cause harms,
so we need to fix the algorithms”

We have fairness research!!

“We’re creating algorithms that cause harms, so we need to fix the algorithms”

Limited to algorithmic bias within a system

Assumes ML owners have the incentives and the means

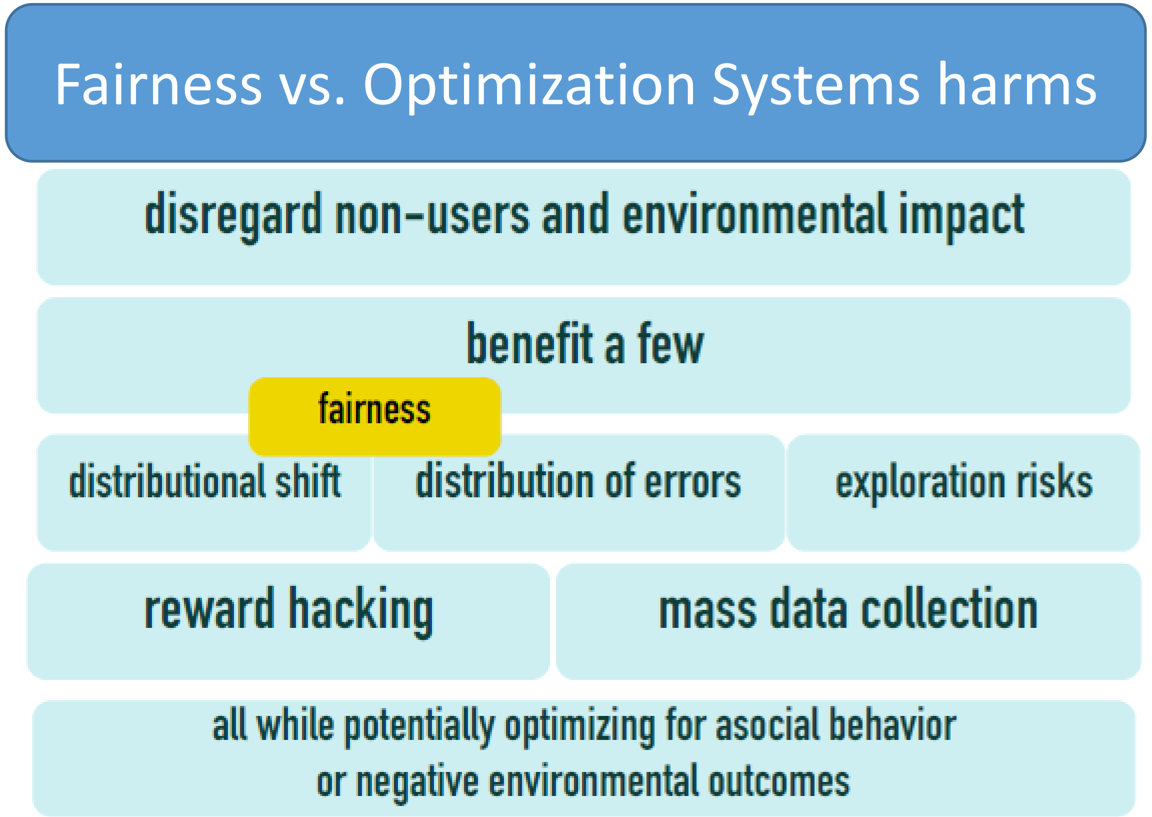
Decontextualized from the system’s goal

Ignores other harms

https://www.esat.kuleuven.be/cosic/questioning-the-assumptions-behind-fairness-solutions/

A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. Os Keyes, Jevan Hutson, and Meredith Durbin

Wait! But we have fairness research!!


Fairness vs. Optimization Systems harms
disregard non-users and environmental impact
benefit a few
fairness
distributional shift
distribution of errors
exploration risks
reward hacking
mass data collection
all while potentially optimizing for asocial behavior
or negative environmental outcomes

Limited to algorithmic bias within a system

Assumes ML owners have the incentives and the means

Decontextualized from the system's goal

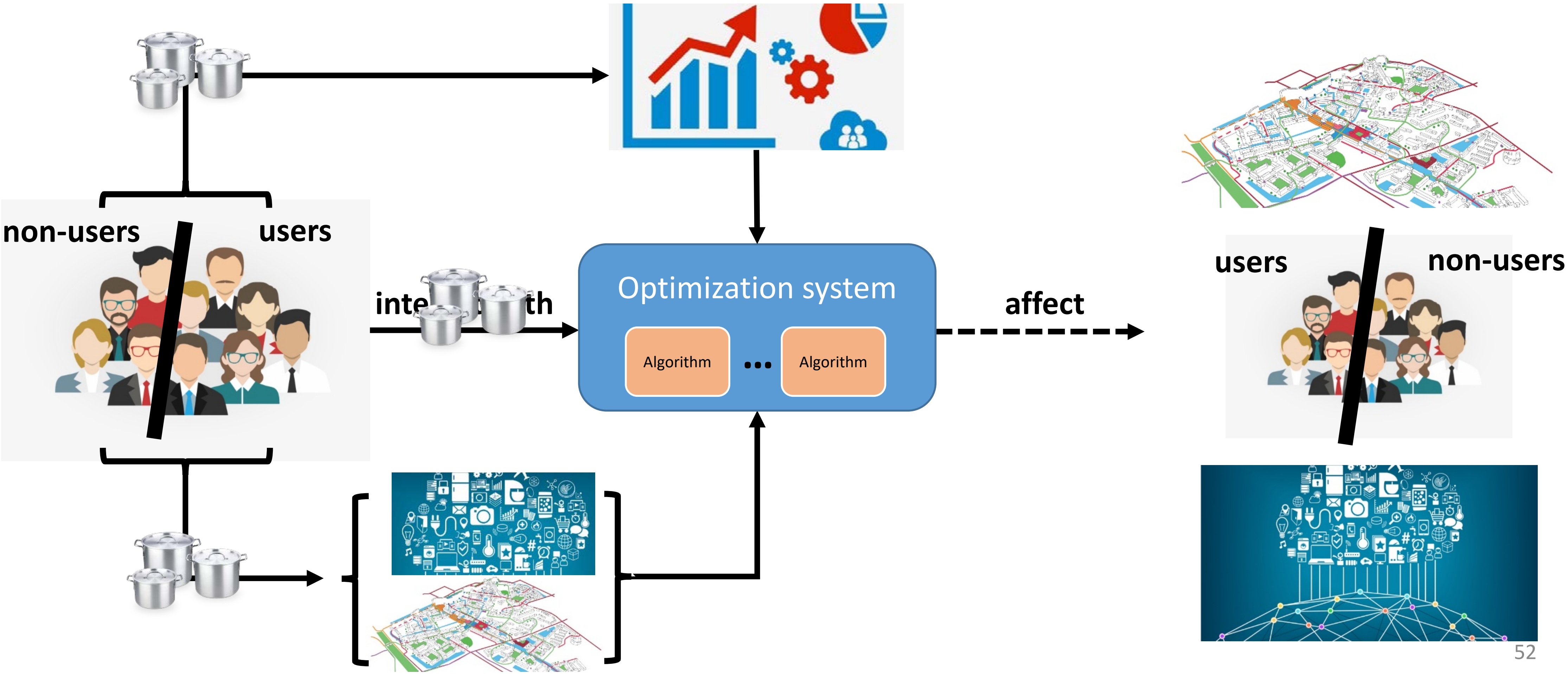
Ignores other harms

https://www.esat.kuleuven.be/cosic/questioning-the-assumptions-behind-fairness-solutions/

A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. Os Keyes, Jevan Hutson, and Meredith Durbin

Protective Optimization Technologies (POTs)

Technologies aimed at mitigating externalities of optimization system's



non-users

users

inte

th

Optimization system

Algorithm

•••

Algorithm

affect

users

non-users

Credit scoring

potential risk posed by lending money to consumers and to mitigate losses due to bad debt

Biased training data → Underlying algorithms can:

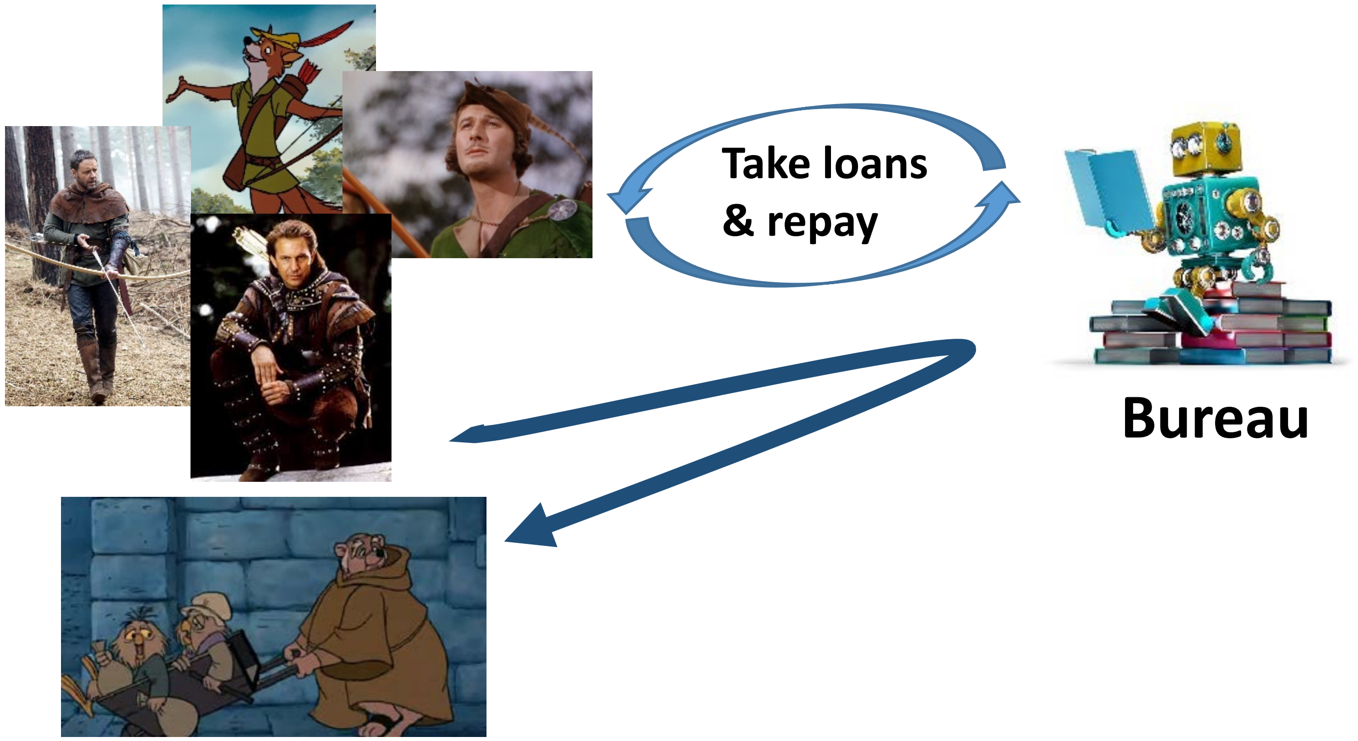
- discriminate applicants on protected attributes like gender or ethnicity
- cause feedback loops for populations disadvantaged by the financial system

Credit bureaus have little incentive to change
Fairness techniques are incipient and hard to deploy

POTs for Credit scoring

- Enable users to help others get loans

Poisoning



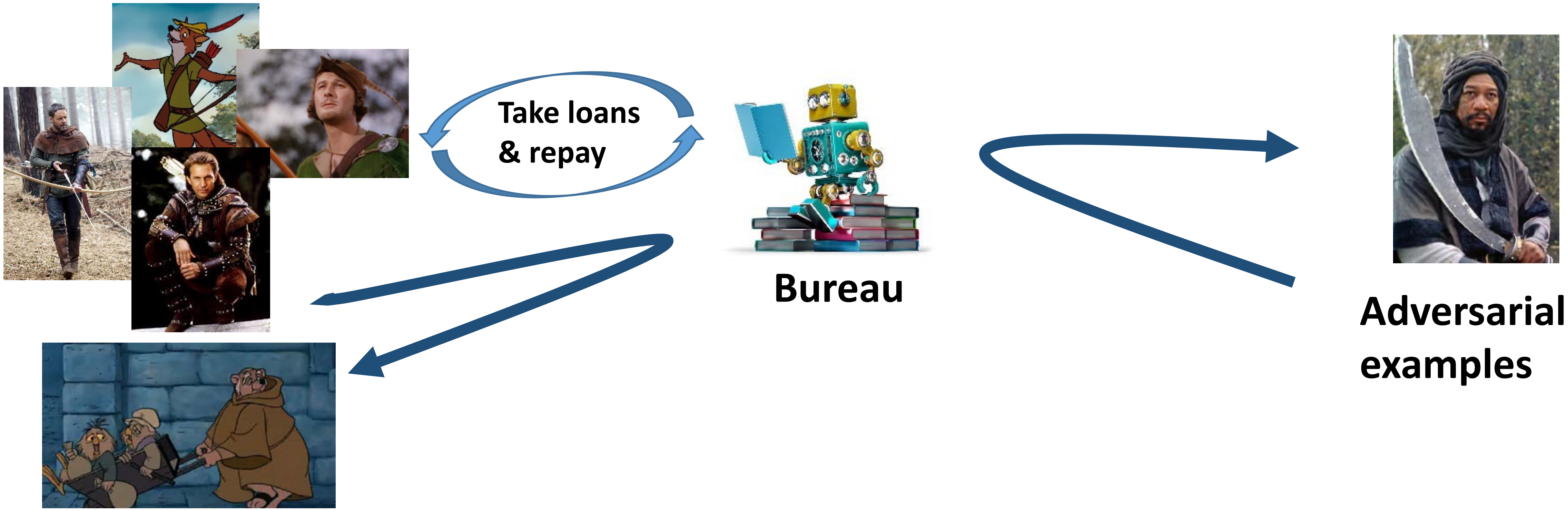
Take loans & repay

Bureau

POTs for Credit scoring

- Enable users to help others get loans
- Enable discriminated users to get loans

Poisoning



Take loans & repay

Bureau

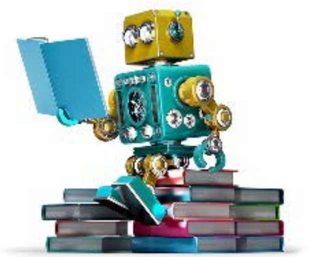
Adversarial examples

POTs for Credit scoring

- Enable users to help others get loans
- Enable discriminated users to get loans

Poisoning

Take loans & repay



Bureau

Adversarial examples

DISCRETE AND CONSTRAINED!

Adversarial machine learning for social justice

✓ **There is a need to protect individuals beyond preserving their privacy**

✓ **Protective Optimization Technologies can be deployed to help individuals and groups to counter externalities**

✓ **POTs are also CONSTRAINED so the graphical approach can also be used as technique to EFFICIENTLY find MINIMAL COST adversarial examples**

A challenge ahead

Disparate vulnerability

- Machine learning models inherit biases in the training

- Two Key implications
  - ML-based attacks are unfair
    (like any ML-based model…)

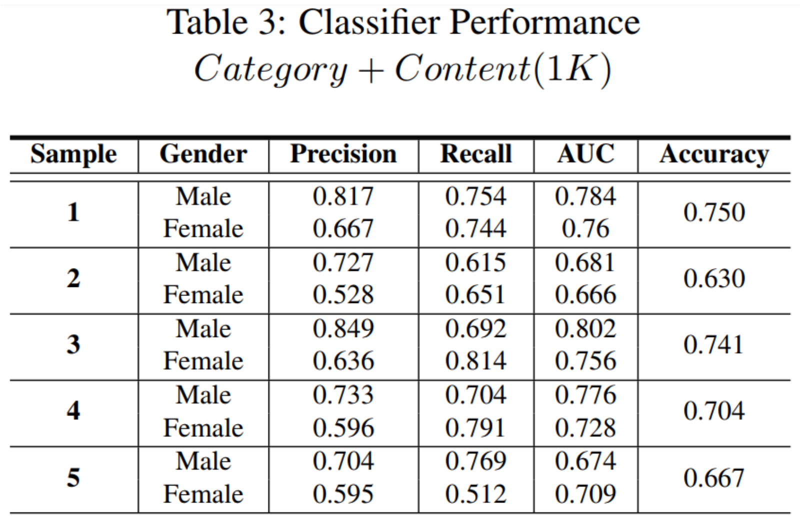


Table 3: Classifier Performance
$Category + Content(1K)$

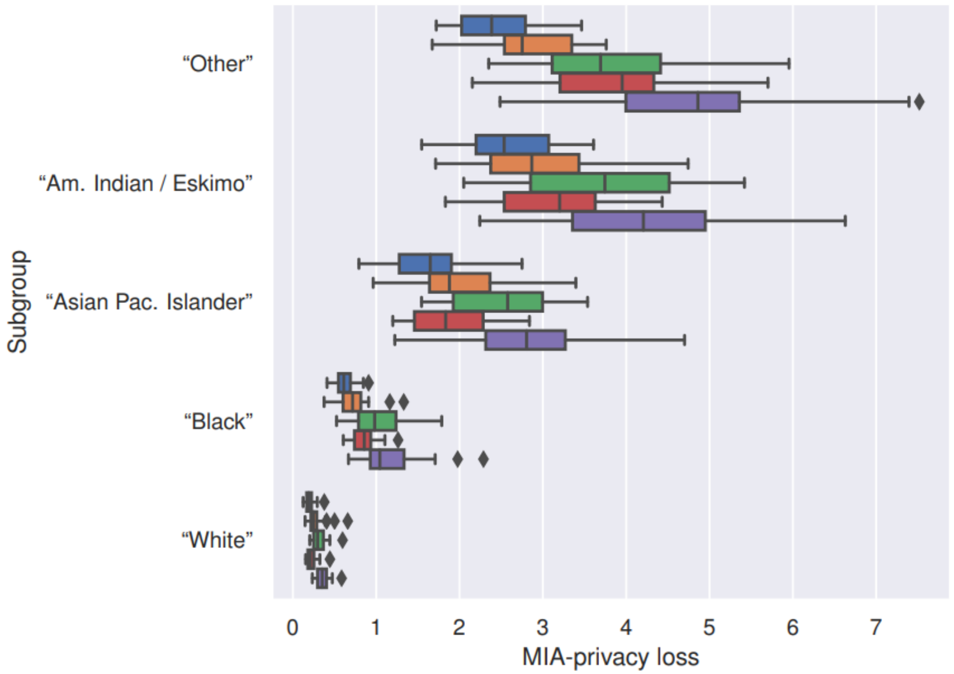
Sample	Gender	Precision	Recall	AUC	Accuracy
1	Male Female	0.817 0.667	0.754 0.744	0.784 0.76	0.750
2	Male Female	0.727 0.528	0.615 0.651	0.681 0.666	0.630
3	Male Female	0.849 0.636	0.692 0.814	0.802 0.756	0.741
4	Male Female	0.733 0.596	0.704 0.791	0.776 0.728	0.704
5	Male Female	0.704 0.595	0.769 0.512	0.674 0.709	0.667

Your Installed Apps Reveal Your Gender and More! Suranga Seneviratne, Aruna Seneviratne, Prasant Mohapatra, and Anirban Mahanti.

Disparate vulnerability

- Machine learning models inherit biases in the training

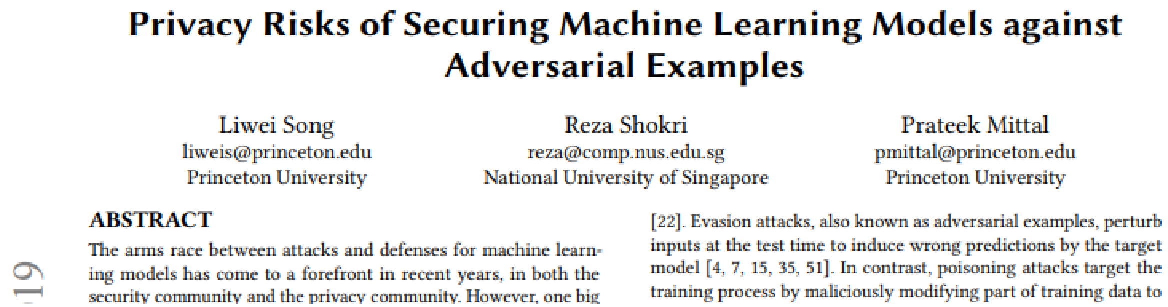
- Two Key implications
  - ML-based attacks are unfair
  - Attacks on ML-models are unfair!



Disparate Vulnerability: on the Unfairness of Privacy Attacks Against Machine Learning. Mohammad Yaghini, Bogdan Kulynych, Carmela Troncoso

Disparate vulnerability

- Is increased when defending ML models from other shortcomings



Privacy Risks of Securing Machine Learning Models against Adversarial Examples

Liwei Song
liweis@princeton.edu
Princeton University

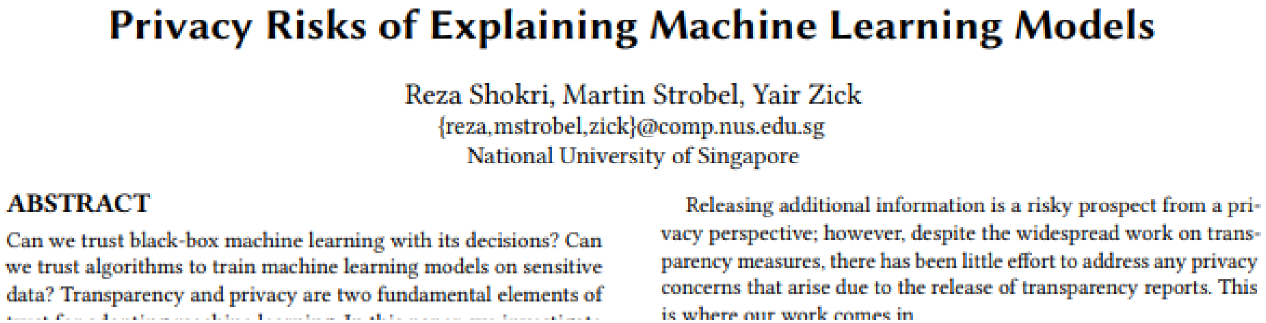
Reza Shokri
reza@comp.nus.edu.sg
National University of Singapore

Prateek Mittal
pmittal@princeton.edu
Princeton University

ABSTRACT

The arms race between attacks and defenses for machine learning models has come to a forefront in recent years, in both the security community and the privacy community. However, one big

[22]. Evasion attacks, also known as adversarial examples, perturb inputs at the test time to induce wrong predictions by the target model [4, 7, 15, 35, 51]. In contrast, poisoning attacks target the training process by maliciously modifying part of training data to



Privacy Risks of Explaining Machine Learning Models

Reza Shokri, Martin Strobel, Yair Zick
{reza,mstrobel,zick}@comp.nus.edu.sg
National University of Singapore

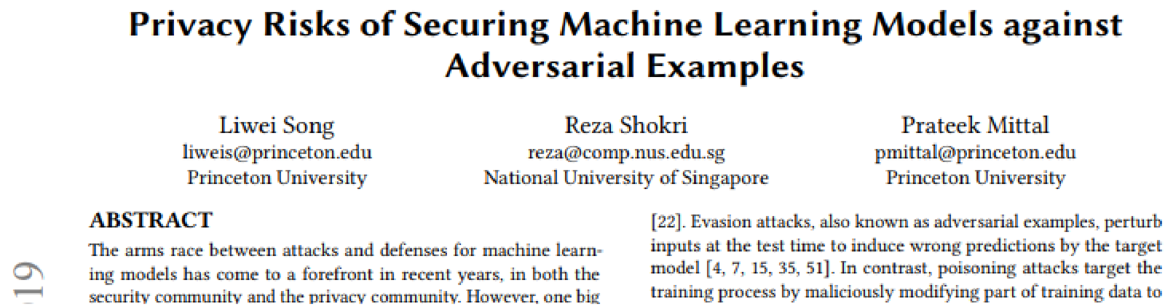
ABSTRACT

Can we trust black-box machine learning with its decisions? Can we trust algorithms to train machine learning models on sensitive data? Transparency and privacy are two fundamental elements of

Releasing additional information is a risky prospect from a privacy perspective; however, despite the widespread work on transparency measures, there has been little effort to address any privacy concerns that arise due to the release of transparency reports. This is where our work comes in

Disparate vulnerability

- Is increased when defending ML models from other shortcomings



19

Privacy Risks of Securing Machine Learning Models against Adversarial Examples

Liwei Song
liweis@princeton.edu
Princeton University

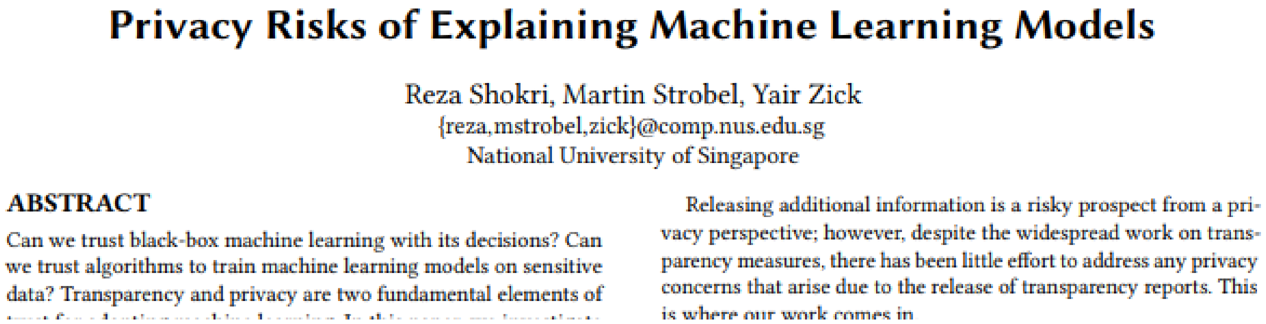
Reza Shokri
reza@comp.nus.edu.sg
National University of Singapore

Prateek Mittal
pmittal@princeton.edu
Princeton University

ABSTRACT

The arms race between attacks and defenses for machine learning models has come to a forefront in recent years, in both the security community and the privacy community. However, one big

[22]. Evasion attacks, also known as adversarial examples, perturb inputs at the test time to induce wrong predictions by the target model [4, 7, 15, 35, 51]. In contrast, poisoning attacks target the training process by maliciously modifying part of training data to



Privacy Risks of Explaining Machine Learning Models

Reza Shokri, Martin Strobel, Yair Zick
{reza,mstrobel,zick}@comp.nus.edu.sg
National University of Singapore

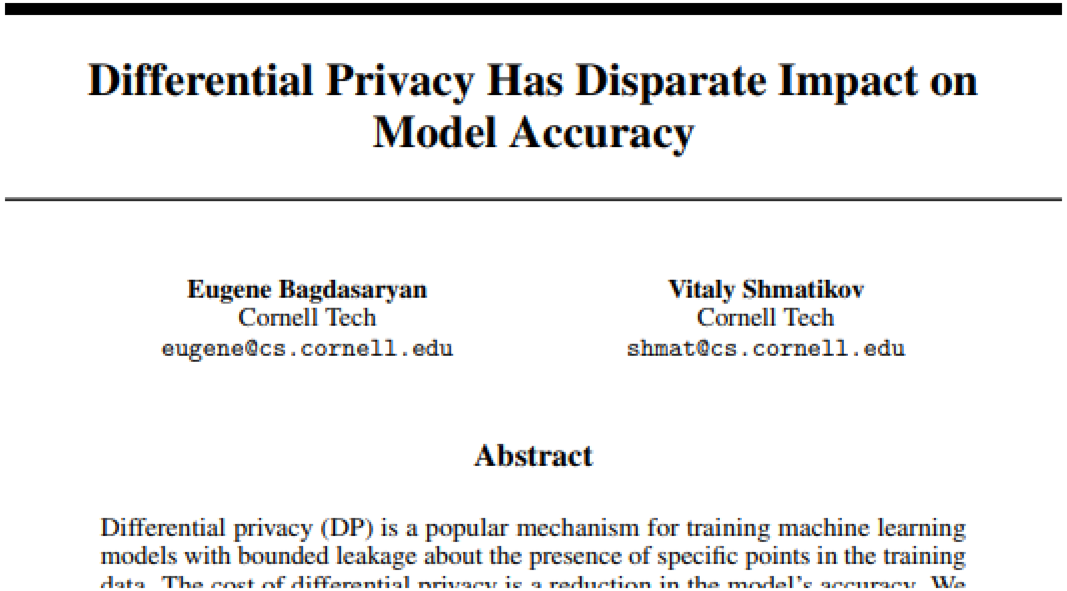
ABSTRACT

Can we trust black-box machine learning with its decisions? Can we trust algorithms to train machine learning models on sensitive data? Transparency and privacy are two fundamental elements of trust for adopting machine learning. In this paper we investigate

Releasing additional information is a risky prospect from a privacy perspective; however, despite the widespread work on transparency measures, there has been little effort to address any privacy concerns that arise due to the release of transparency reports. This is where our work comes in

Disparate vulnerability

- And blanket defenses have disparate impact on utility!



Differential Privacy Has Disparate Impact on Model Accuracy

Eugene Bagdasaryan
Cornell Tech
eugene@cs.cornell.edu

Vitaly Shmatikov
Cornell Tech
shmat@cs.cornell.edu

Abstract

Differential privacy (DP) is a popular mechanism for training machine learning models with bounded leakage about the presence of specific points in the training data. The cost of differential privacy is a reduction in the model's accuracy. We

Universal design for protection technologies

We need to take into account attack’s fairness when designing protections

- Is it possible to have secure accurate models with fair privacy?
  - Security vs. privacy trade-off?
  - More importantly: fair privacy at the cost of privacy?

- Are adversarial learning-based defenses immune to this issue?
  - If so, should they be our only way forward?

- Should fairness be a bullet in privacy by design beyond ML?

Takeaways

- Adversarial machine learning is hard to defend from: a great opportunity!

Adversarial machine learning as protective technologies

for privacy (PETs) and social justice (POTs)

- New graphical framework to approach the search of adversarial examples

… we can use of graph theory to improve efficiency and provide guarantees

- The fairness problems of machine learning will become a hurdle for protection!

EPFL

SPRING

SECURITY AND PRIVACY ENGINEERING LABORATORY

http://carmelatroncoso.com/

https://spring.epfl.ch/en

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

Bogdan Kulynych

Mohammad Yaghini

Seda Guerses

Rebekah Overdorf

Ero Balsa

Jamie Hayes

Nikita Samarin