BULLETPROOF.AI

AI vs AI

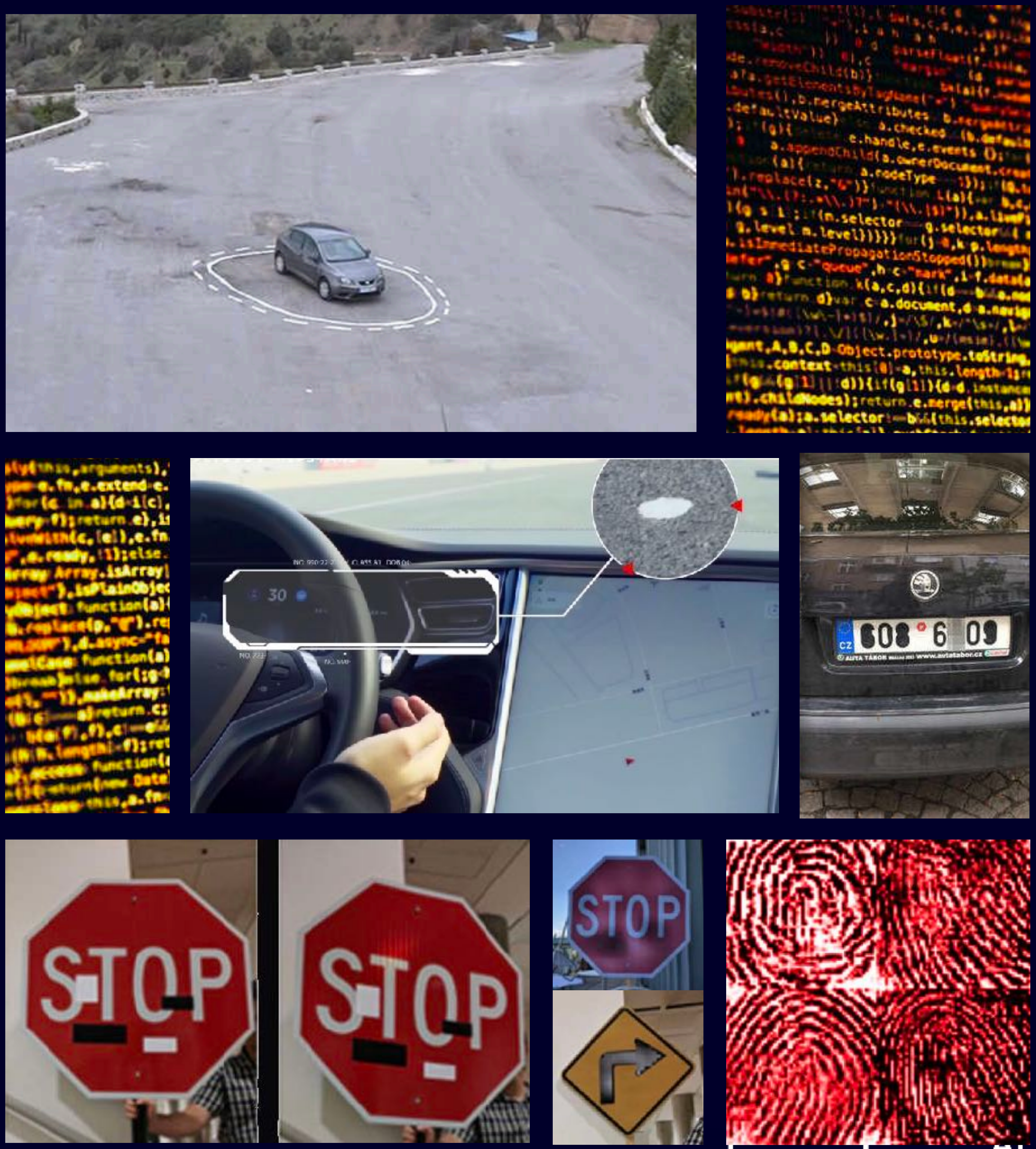
Intelligent Attacks on Automated Financial Decisions

Martin Rehak, September 2019

"Artificial intelligence won’t revolutionize anything if hackers can mess with it.”

Dawn Song, UC Berkley,
in MIT Technology review

https://www.technologyreview.com/s/613170/emtech-digital-dawn-song-adversarial-machine-learning/amp/

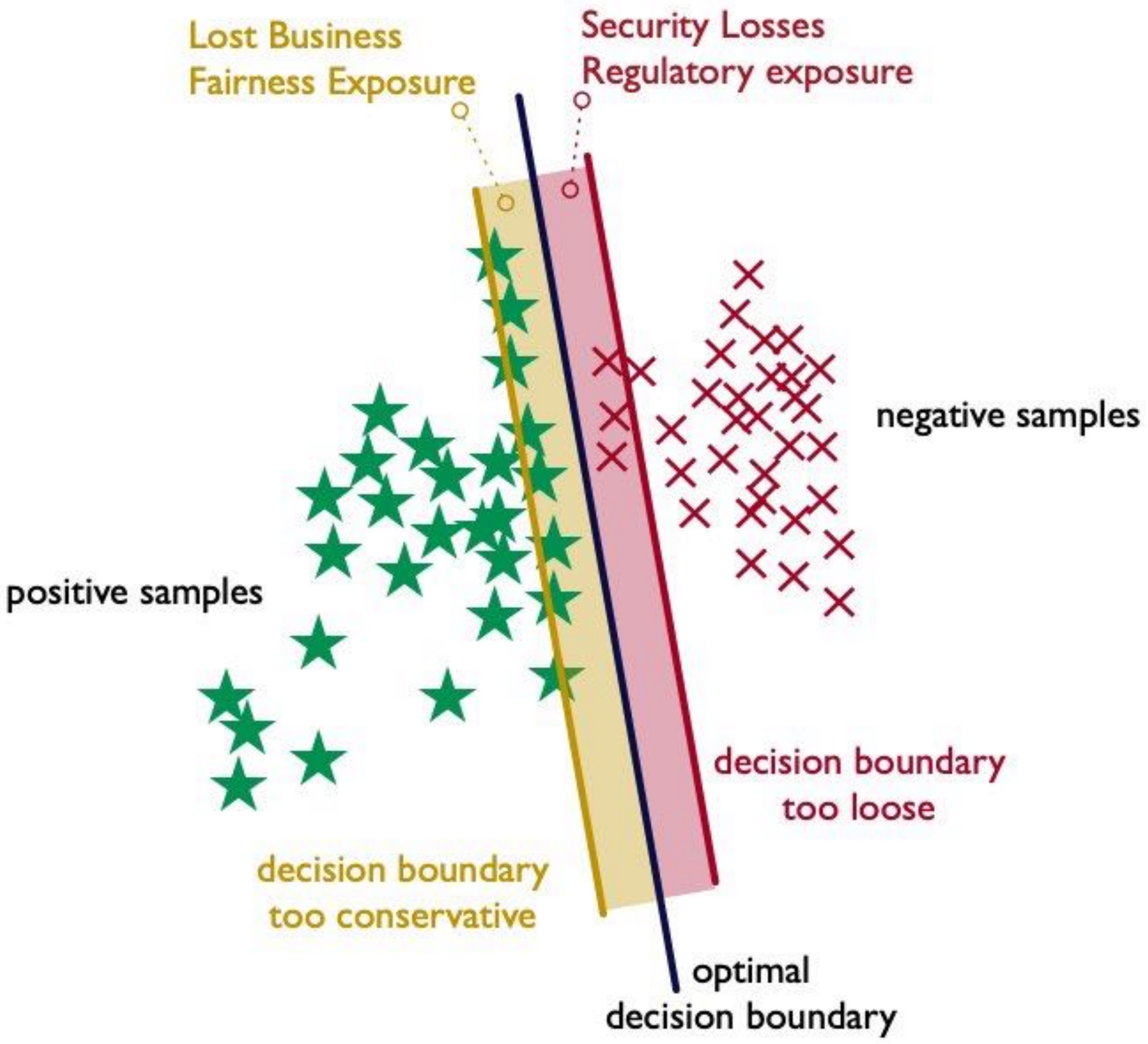


BULLETPROOF.AI

Security solutions for AI, machine learning and automated statistical decisions

AI Models make critical business decisions in split seconds, every second of the day

How Secure, Fair and Robust is your Machine Learning System?



Decision Boundary

- **Facebook effect:** posts on the edge of acceptable use policy get the highest engagement score, regardless of what the actual policy is.
- **Margin impact:** Business next to the decision boundary is less competitive and brings higher margins

Security- Attack Types

Confidentiality Attacks

Attacker may be able to **copy the model** and to **extract** the data used to train the model.

Evasion Attacks

Attacker may **discover and exploit existing vulnerabilities** in the model in order to **manipulate** the decision.

Poisoning Attacks

Attacker may strategically **influence the training** of the model in order to **manipulate** the model decision.

Confidentiality Attacks

- **Model Extraction**

  Attacks designed to extract the complete decision model. Successful attack gives the attacker the ability to predict all future decisions of the model and to replicate all the past decisions.

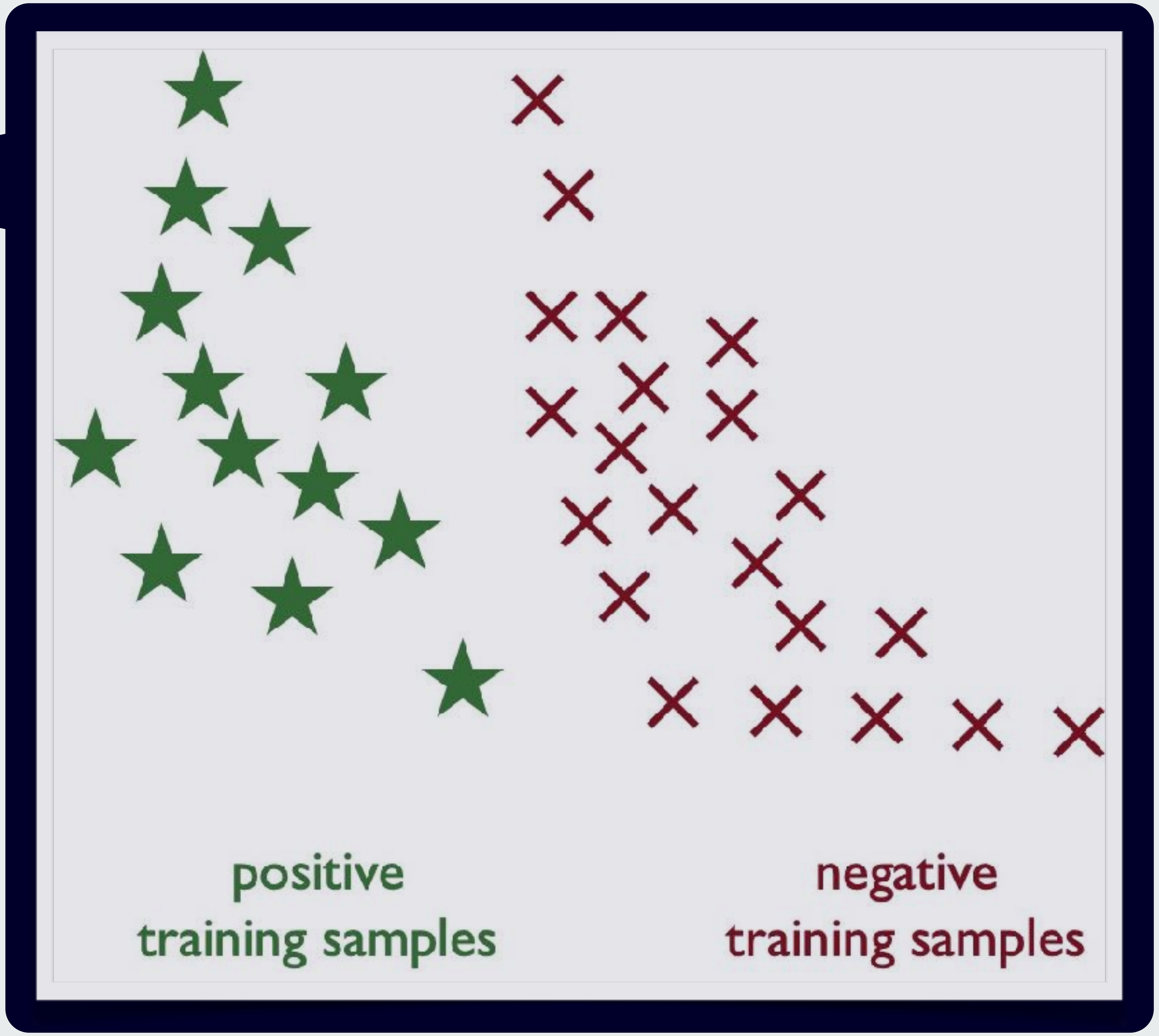
- **Data Extraction**

  Attacks designed to extract the data points used to train the models. Data points may reveal information about business partners, customers, their transaction history and other data.

Confidentiality Attack

1

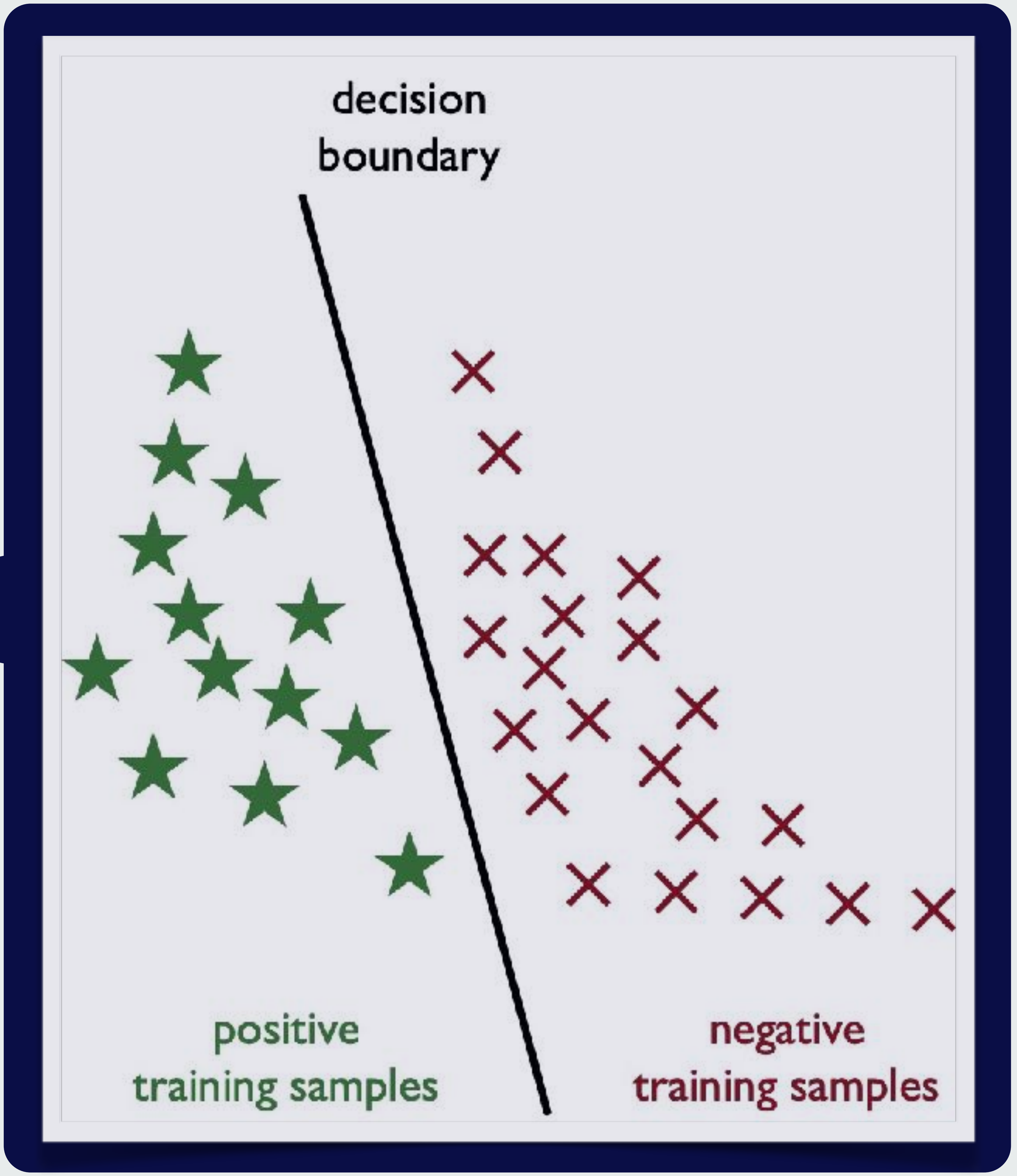
Supervised model is **trained** on labeled samples



Confidentiality Attack

1 Model is **trained** on labeled samples

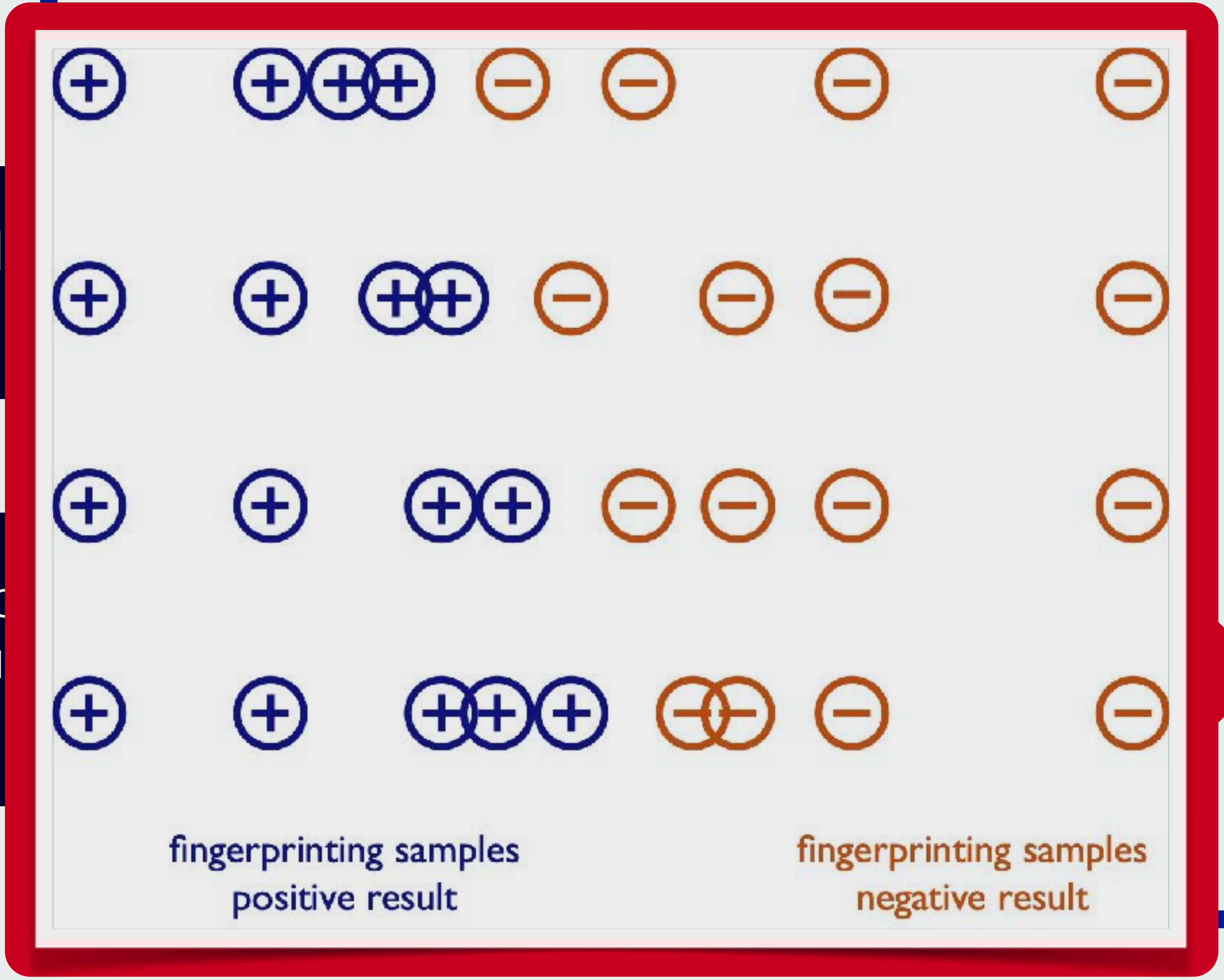
2 Decision boundary is **identified** during the training and defines the model



Confidentiality Attack

Model

Dec
identifi
and



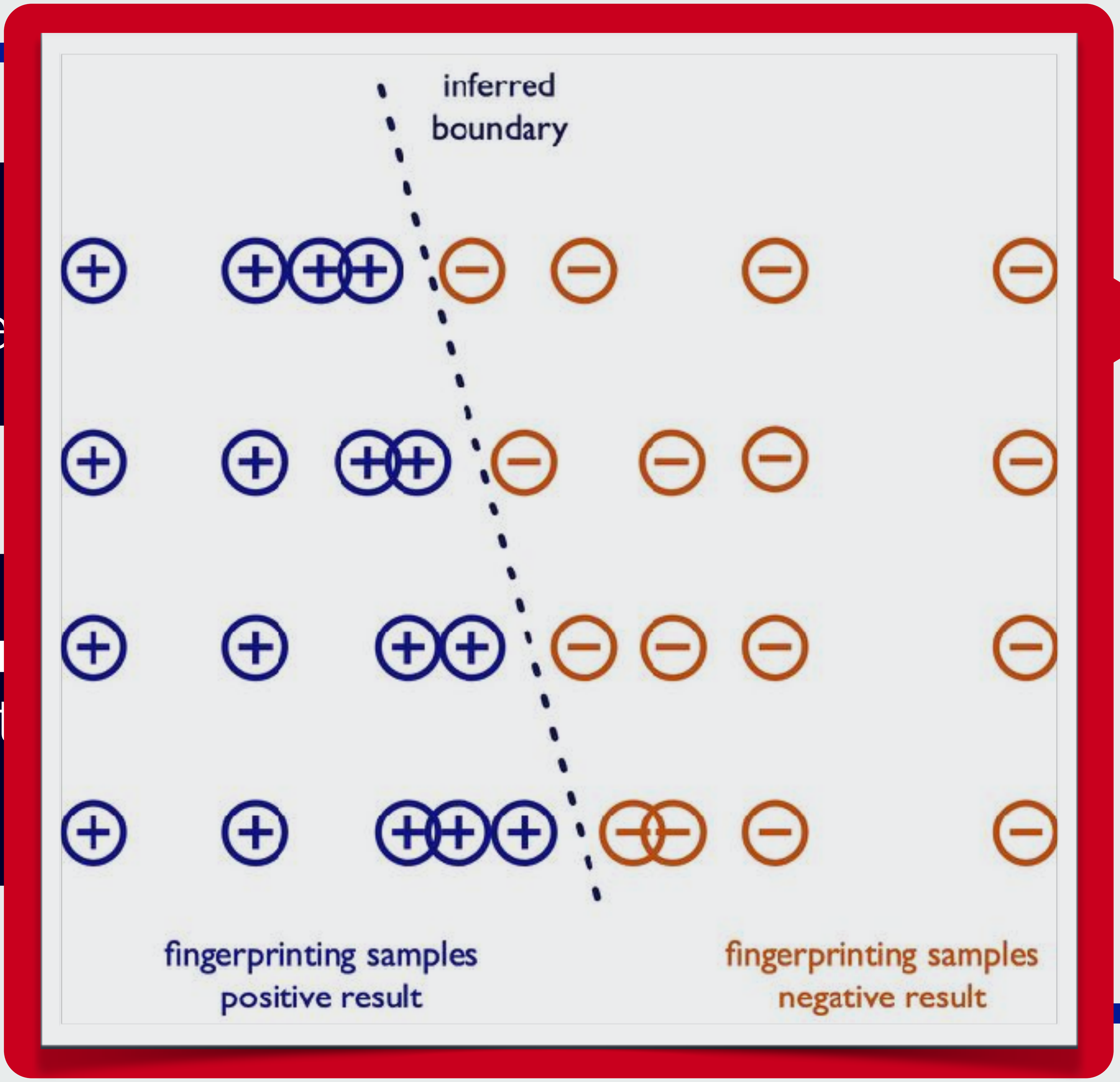
3

Attacker can **strategically query** the model and obtain a dense set of his own training samples based on the interaction

Confidentiality Attack

Model is **trained** sample

Decision bou identified during t and defines th



4

Attacker then uses standard training methods then **reproduce** the decision boundary and the whole model

3

Attacker can **strategically query** the model and obtain a dense set of his own training samples based on the interaction

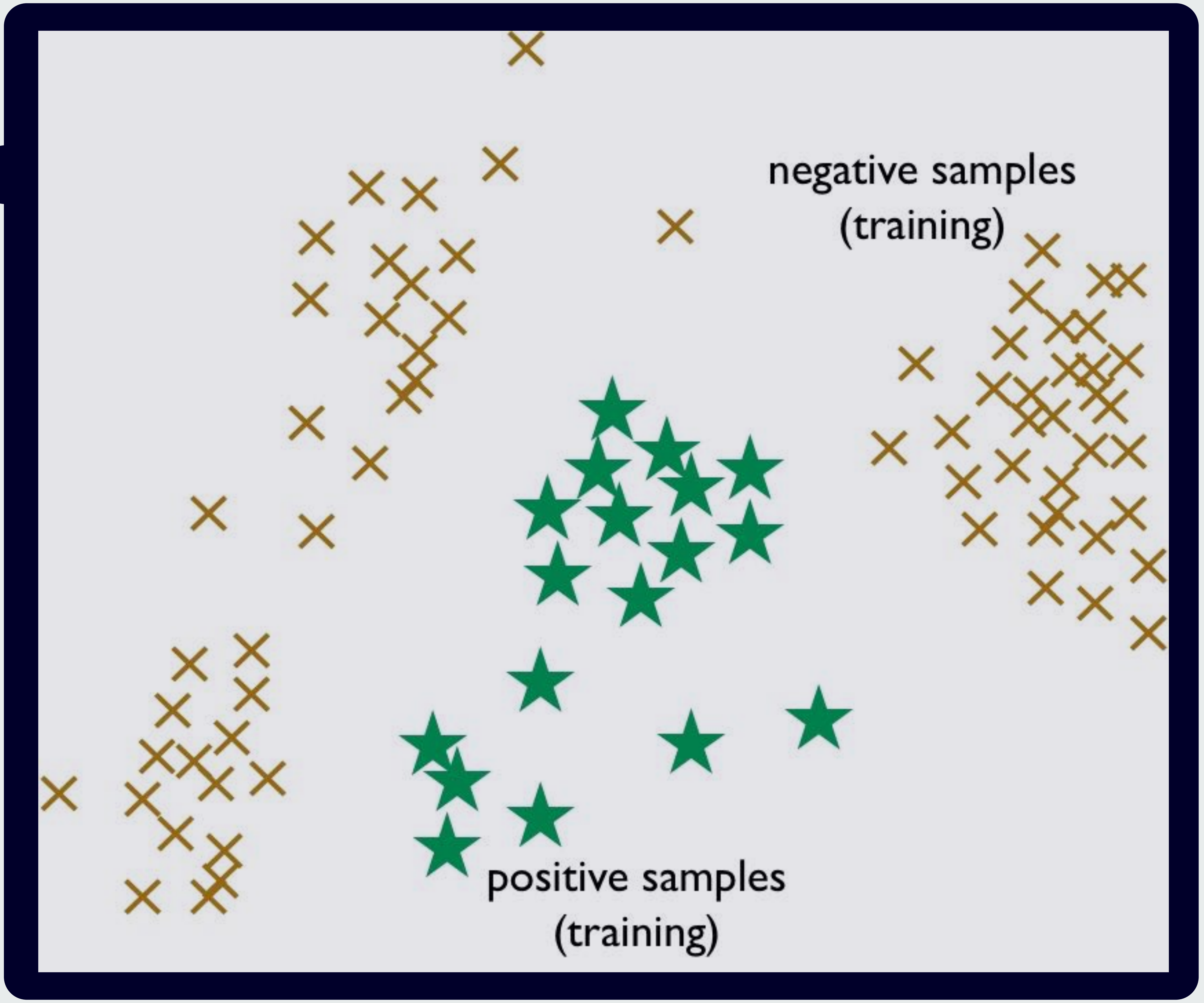
Evasion

- **Existing model faults** can be exploited by adversaries to influence classification results:
  - Faults can be result of **insufficient training set** not covering relevant business-adverse cases
  - Feature selection can introduce model dependence on **proxy features** unrelated to the business performance

Evasion Attack

1

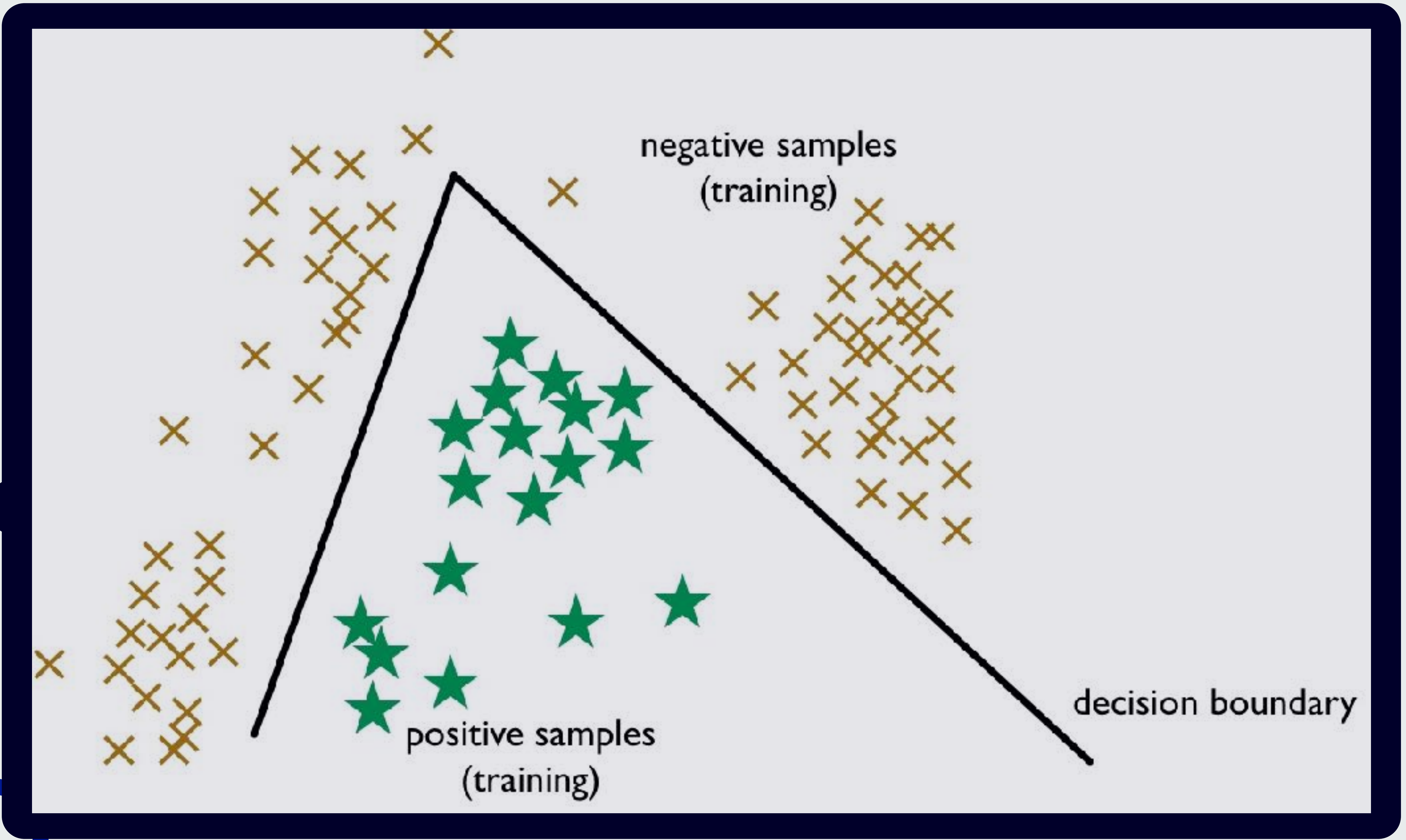
AI is used to scale out and **automate** business decisions



Evasion Attack

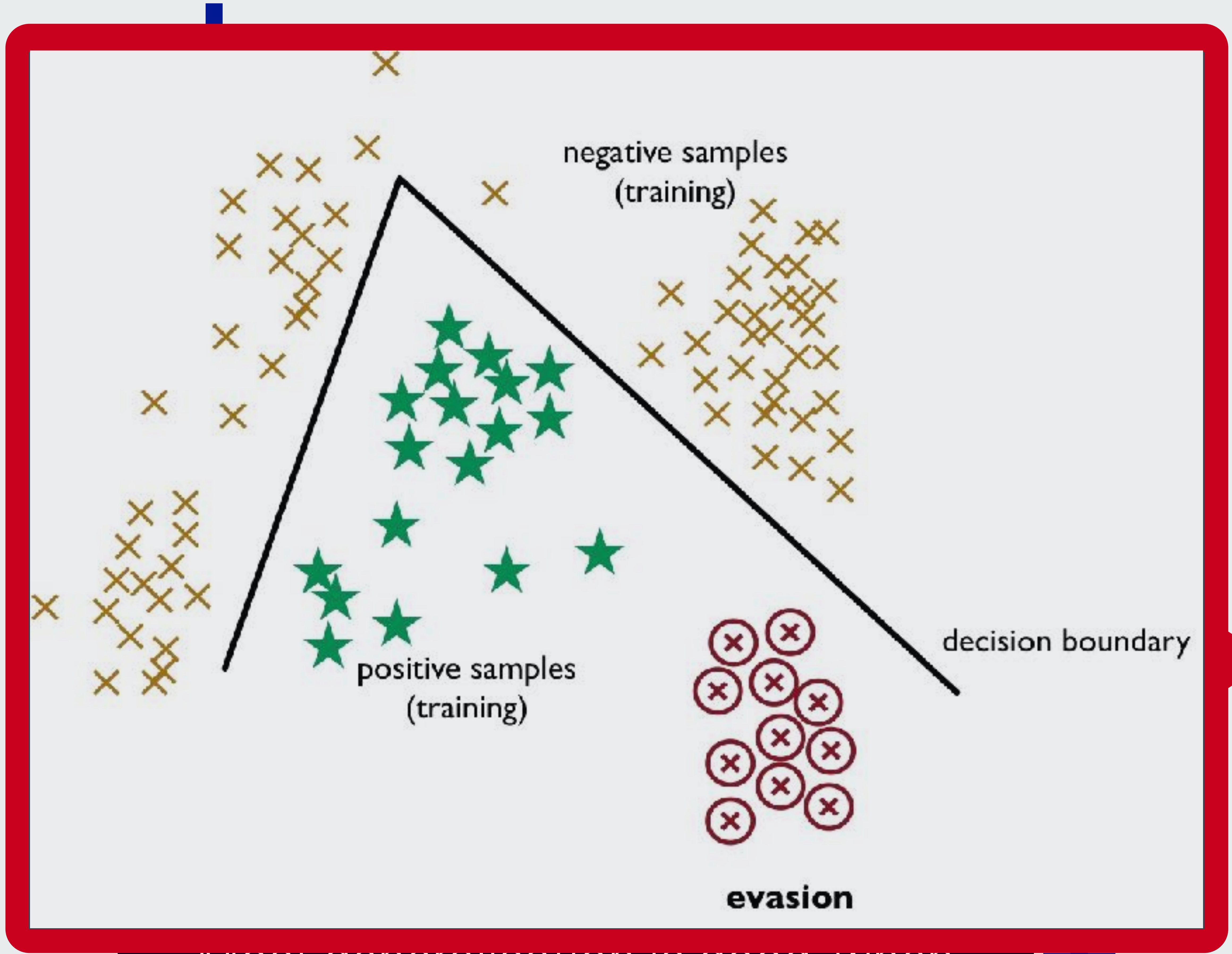
1 AI is used to scale out and **automate** business decisions

2 Models need to **generalise** and make decisions based on similar cases in the past, included in the training set



3 Most generalisation is good, when supported by **dense and representative** data in the feature space

Evasion Attack



3

Most generalisation is good, when supported by **dense and representative** data in the feature space

4

Some **generalisation** is **bad**, due to combination of suboptimal features and missing training set data

5

In evasion attacks, the attacker discovers a region of the feature space where **missing training data** for one decision allow the learning algorithm to attach the region to the adjacent region with the opposite

6

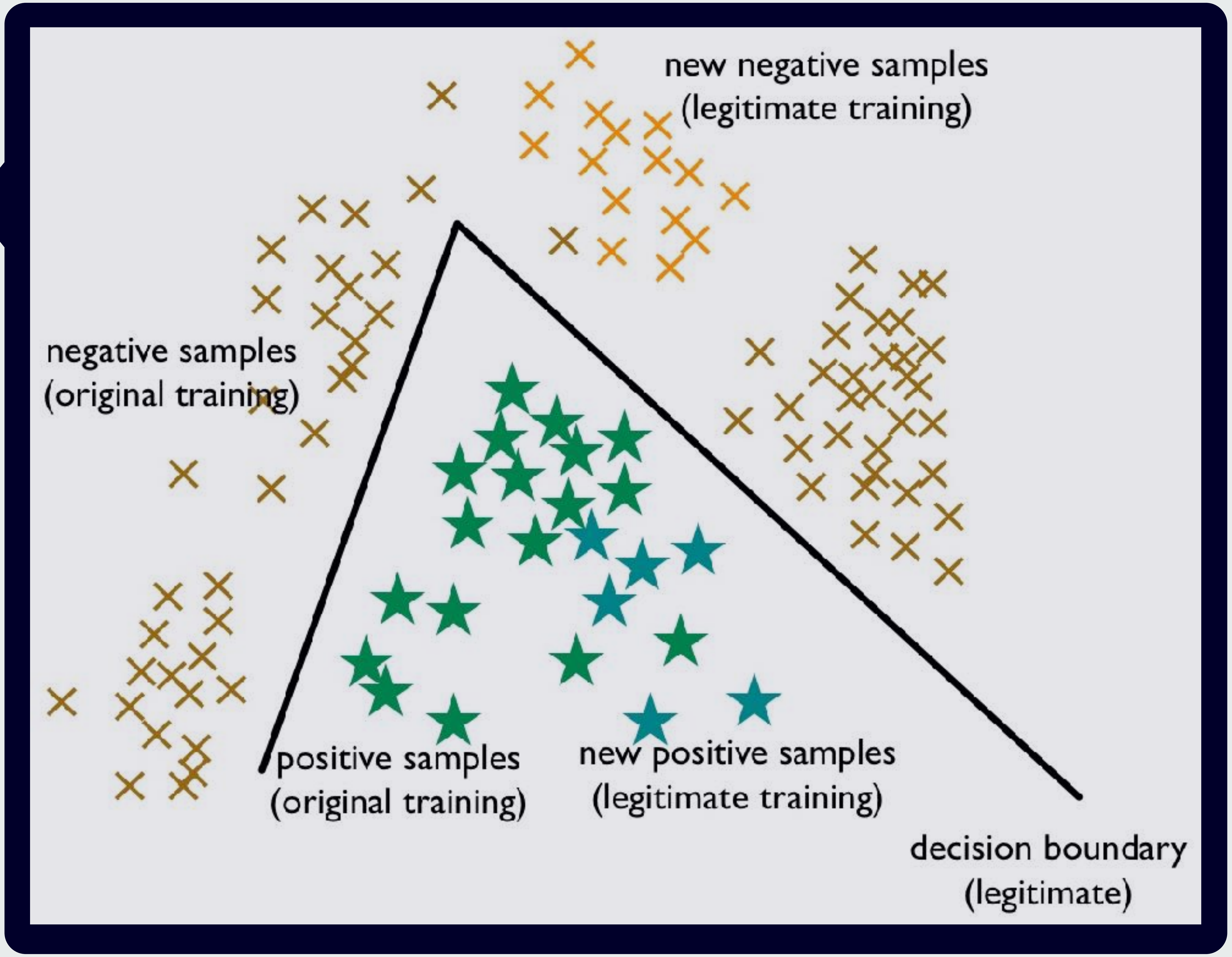
The vulnerabilities that cause evasion can also cause unfairness and **discrimination**

Poisoning

- Attacker **influences** the update process of the model in order to **introduce** exploitable vulnerability
  - Inserted samples can be introduced during ordinary course of business by **strategic business interaction**
  - Model becomes **biased** by the introduction of the poisoned training samples. Attacker can cause damage and benefit from biased model decisions

Poisoning Attack

1 **Business changes** constantly and models need to follow

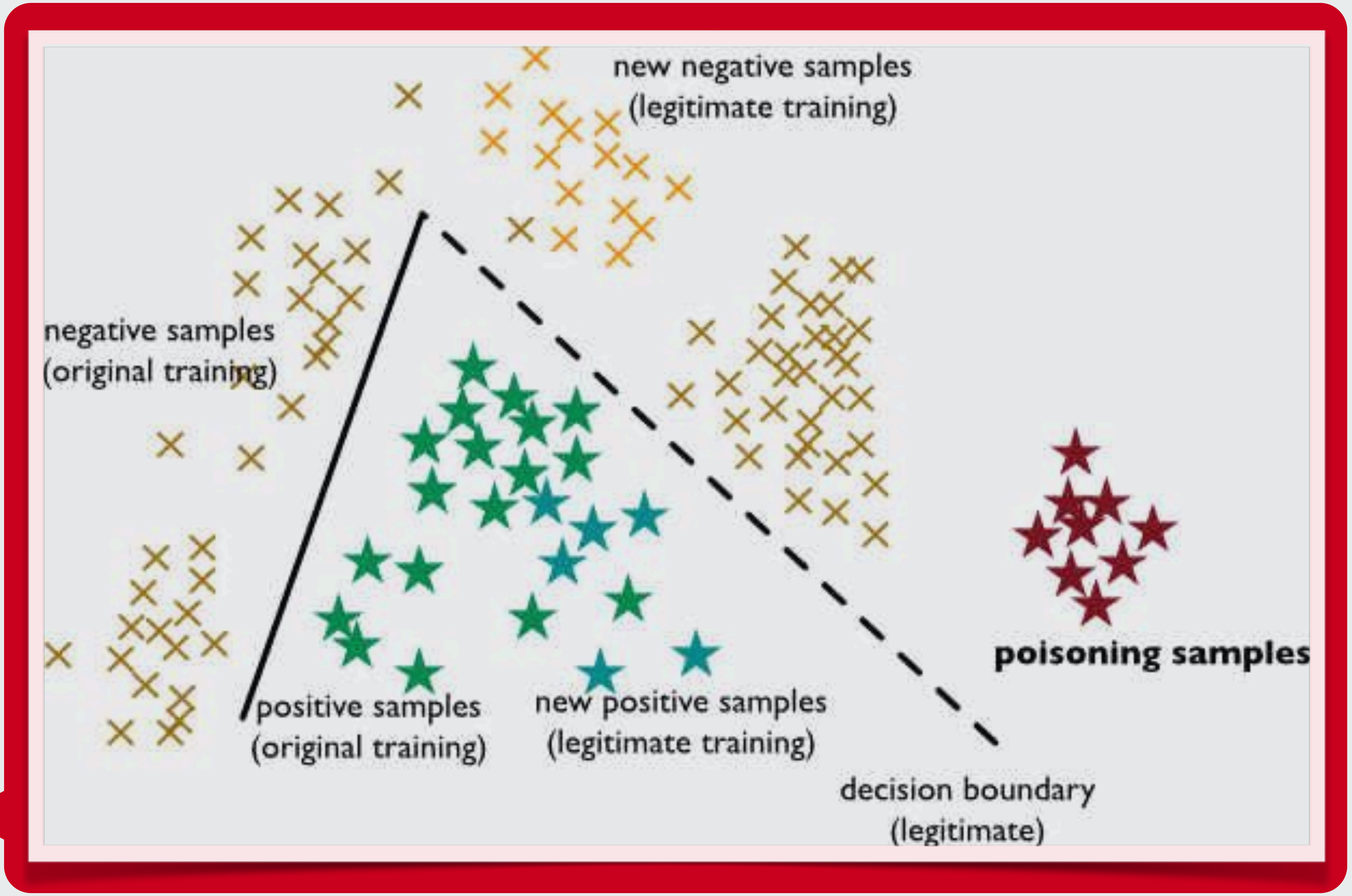


Poisoning Attack

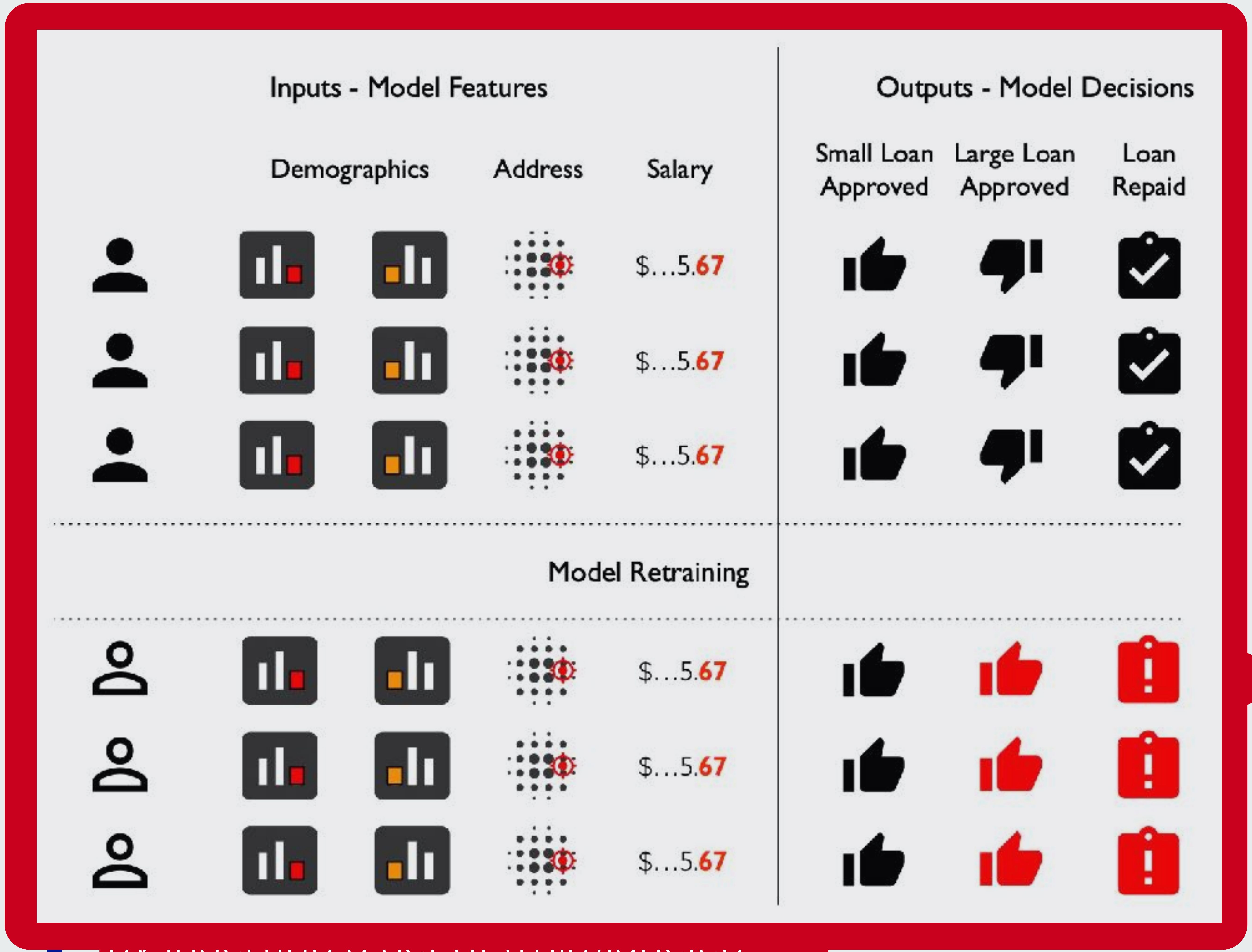
1 **Business changes** constantly and models need to follow the

2 During poisoning, the attackers actively **create** model **vulnerabilities** during the model update.

3 To achieve poisoning, the attacker must **influence the training data** set by inserting a set of (mis)labeled



Poisoning Attack



by inserting a set of (mis)labeled

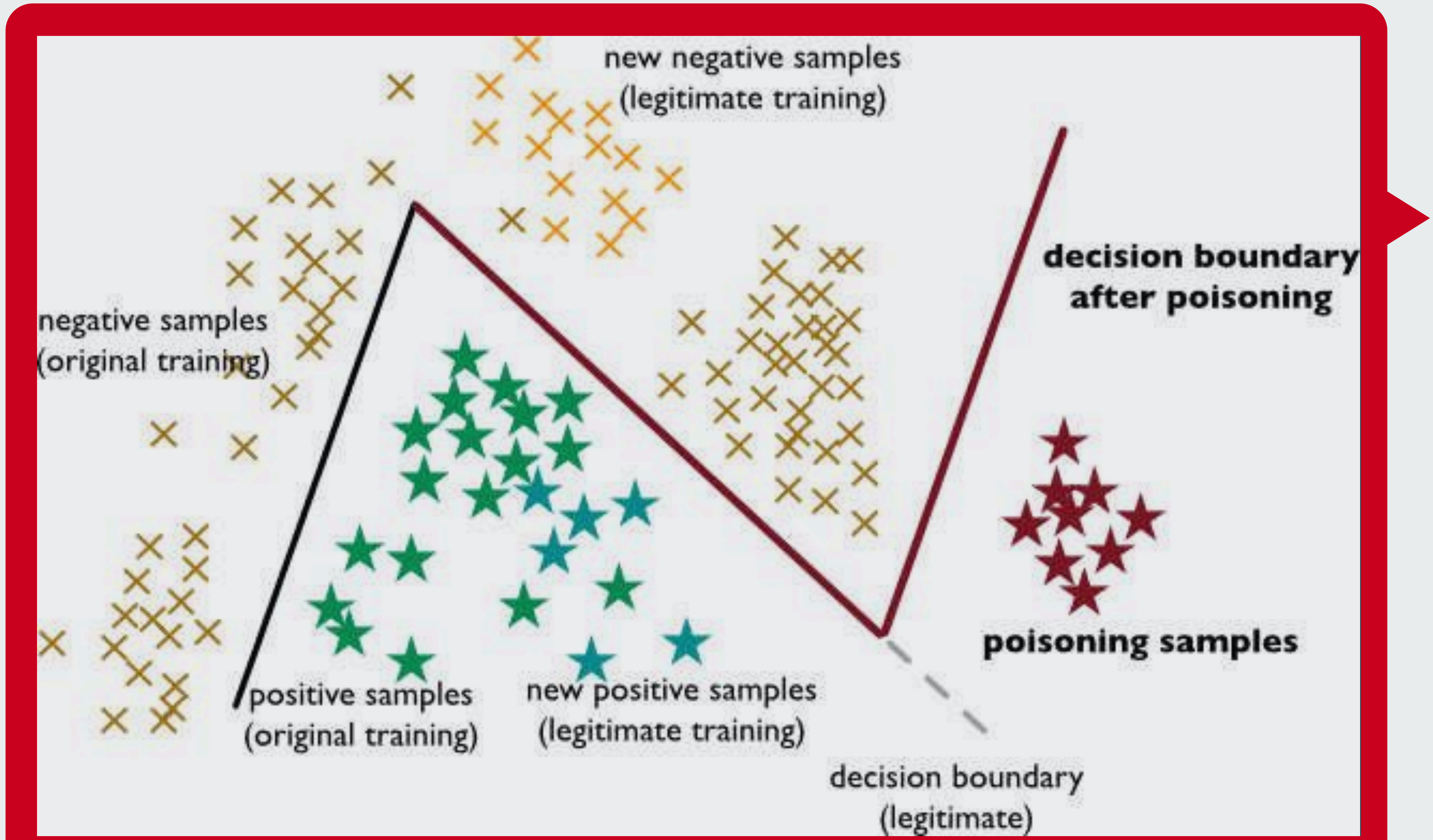
5

Perfect loan payments on **small loans** in specific zipcode/

4

In practice, this is not too hard. Training data is most often **collected from the actual business**.

Poisoning Attack



3 To achieve poisoning, the attacker must **influence the training data** set by inserting a set of (mis)labeled

4 In practice, this is not too hard. Training data is most often **collected from the actual business**.

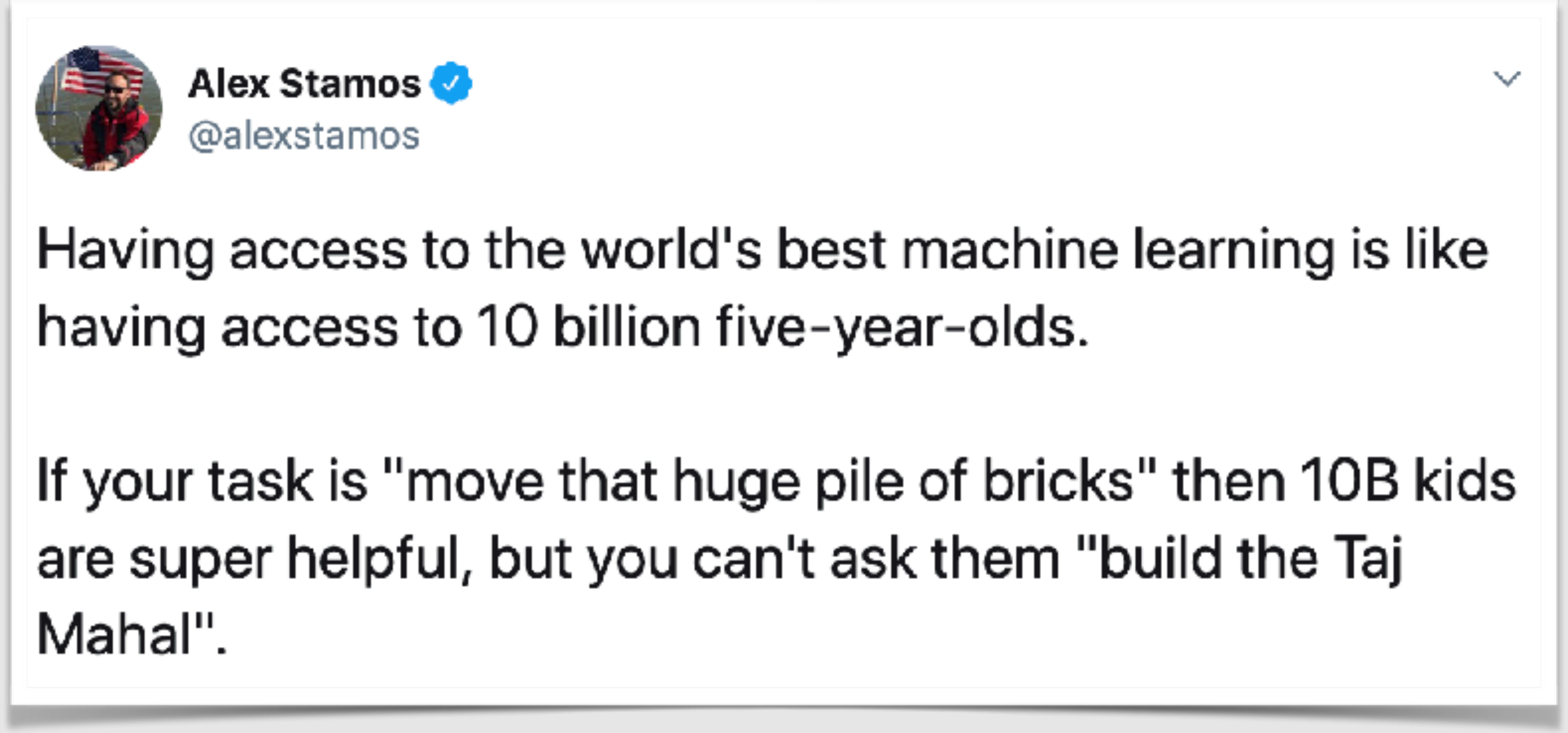
5 Perfect loan payments on **small loans** in specific zipcode/

6 **Reputation building** on small transactions in e-commerce

7 New model then reflects the biased samples inserted by the attacker and is open for exploitation by the attacker.

Artificial Intelligence is like an army of 5-year old kids.

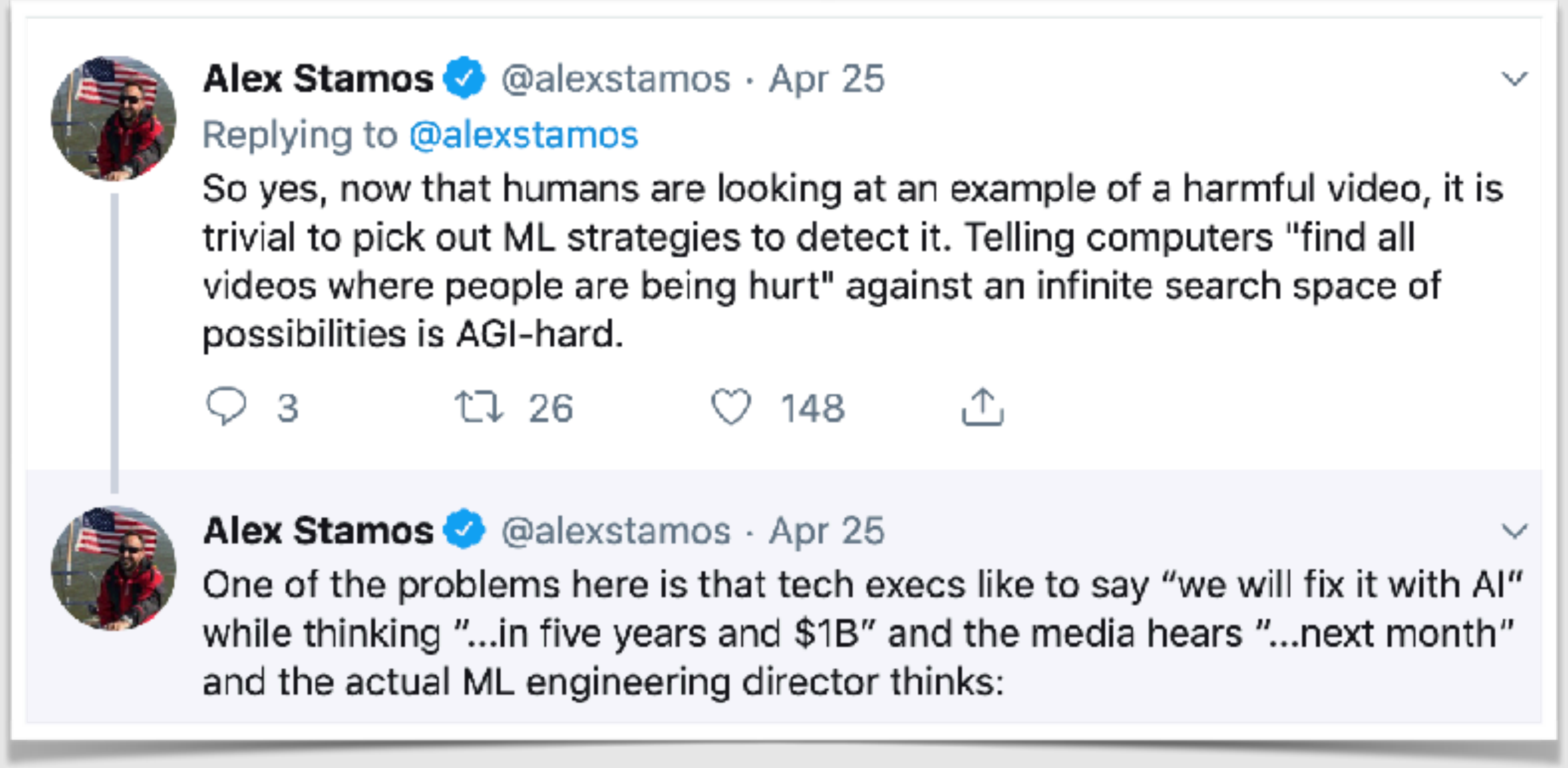
(paraphrased from Alex Stamos)



Alex Stamos
@alexstamos

Having access to the world's best machine learning is like having access to 10 billion five-year-olds.

If your task is "move that huge pile of bricks" then 10B kids are super helpful, but you can't ask them "build the Taj Mahal".



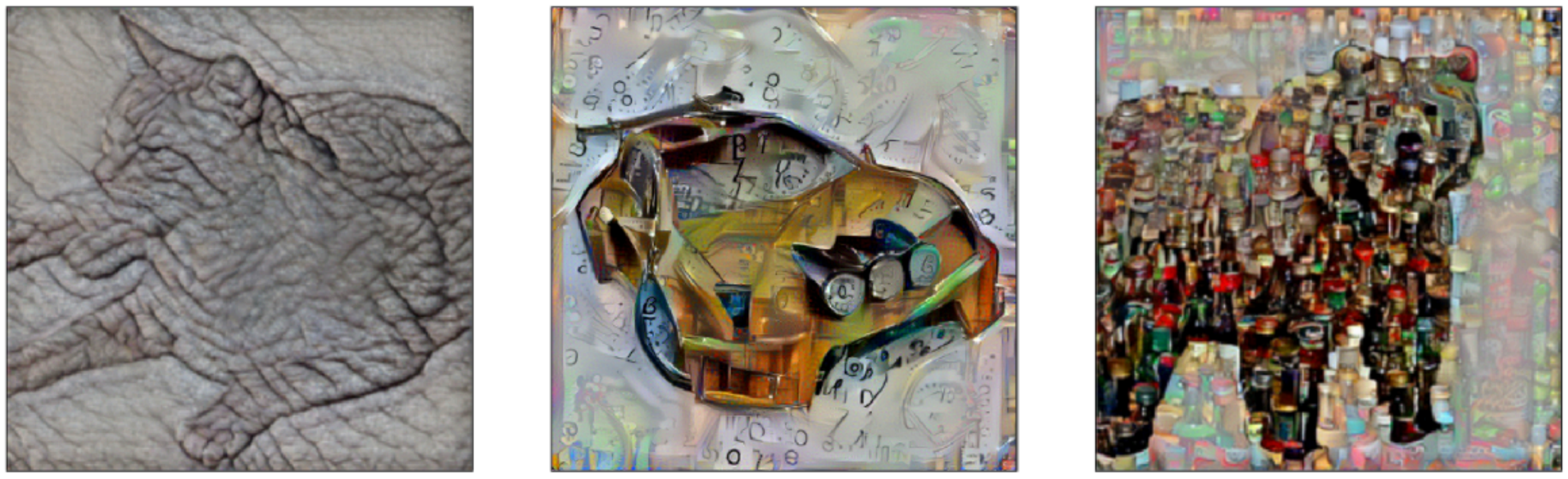
Alex Stamos @alexstamos · Apr 25
Replying to @alexstamos
So yes, now that humans are looking at an example of a harmful video, it is trivial to pick out ML strategies to detect it. Telling computers "find all videos where people are being hurt" against an infinite search space of possibilities is AGI-hard.

3 26 148

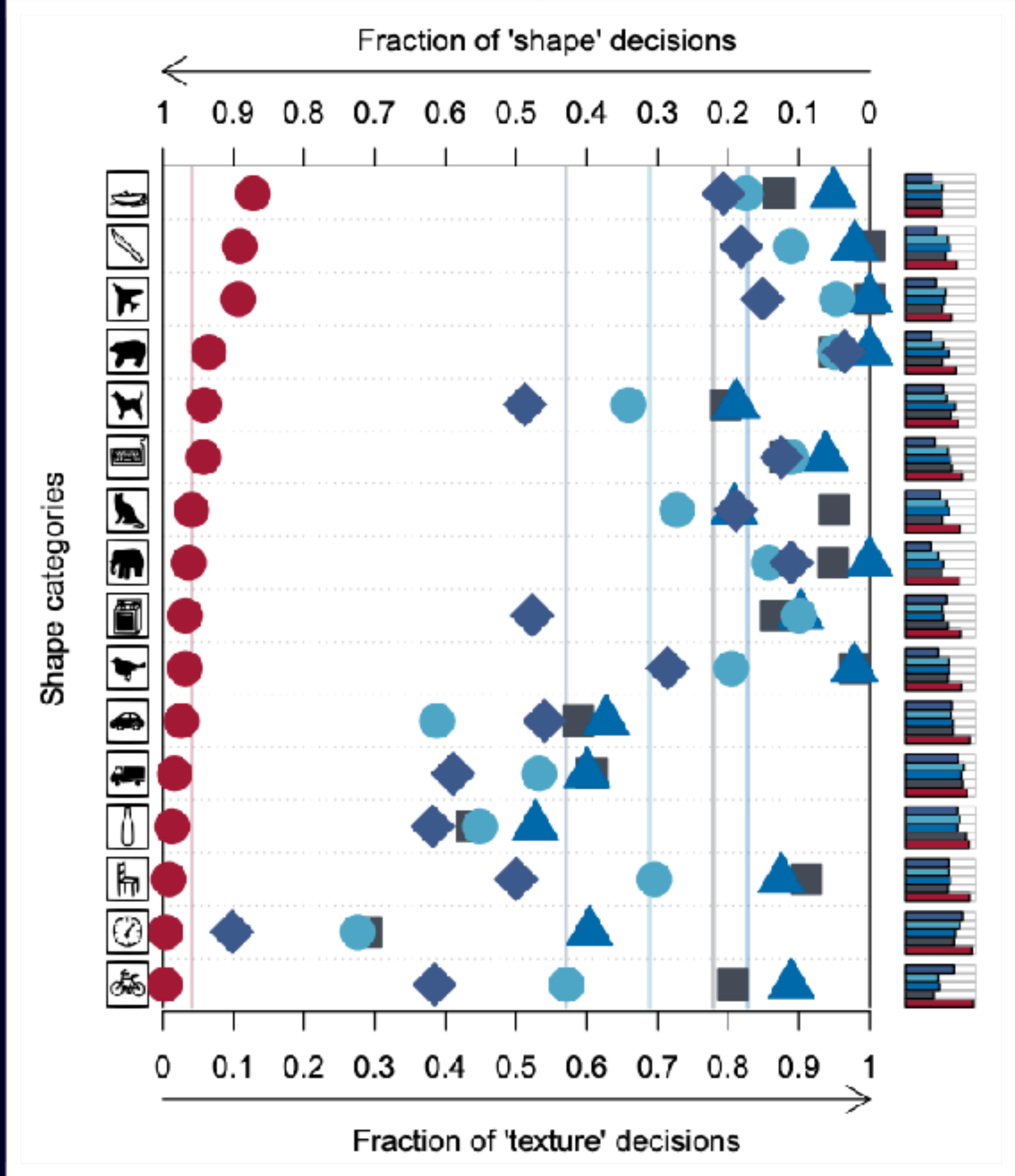
Alex Stamos @alexstamos · Apr 25
One of the problems here is that tech execs like to say “we will fix it with AI” while thinking ”...in five years and $1B” and the media hears ”...next month” and the actual ML engineering director thinks:

Deep Networks and Details

- Deep learning methods exhibit strong preference for detail at the expense of high-level concept extraction

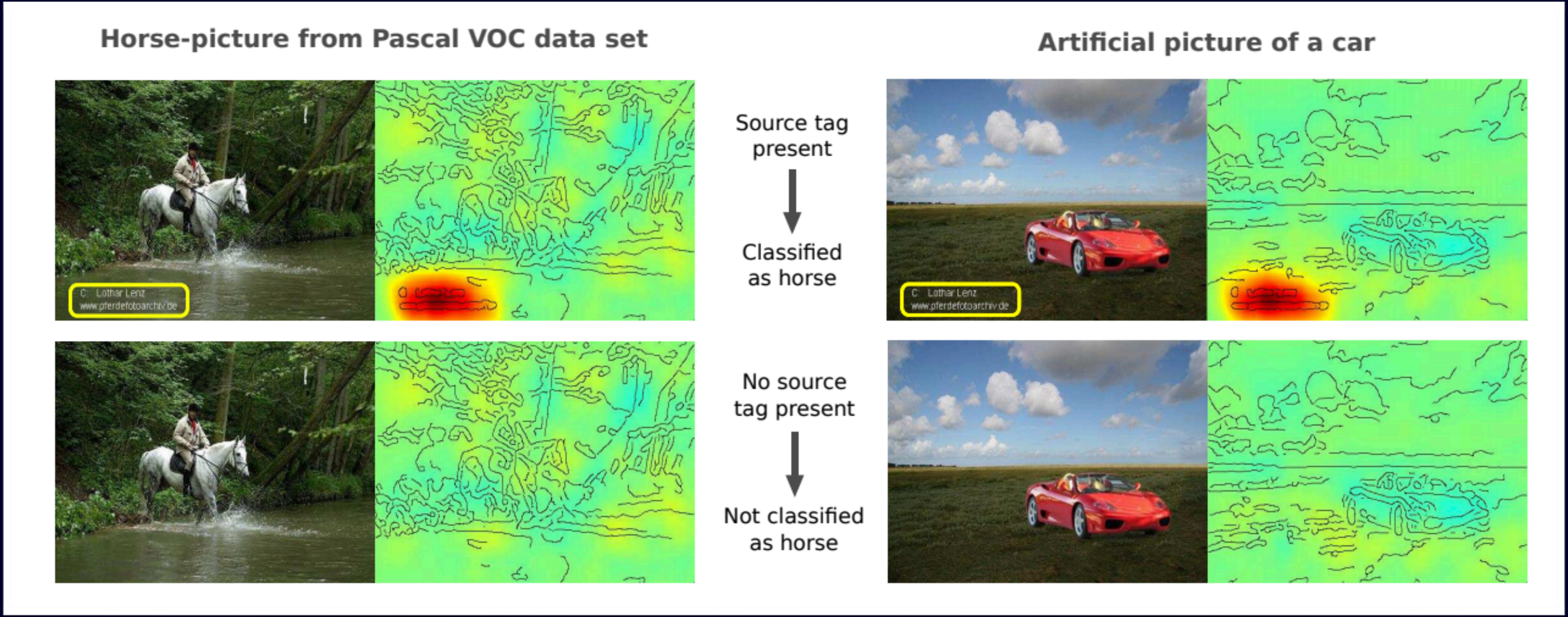


cat with elephant texture | car with clock texture | bear with bottle texture



Geirhos et al.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness, ICLR 2019

Deep Networks and Details



Horse-picture from Pascal VOC data set

Artificial picture of a car

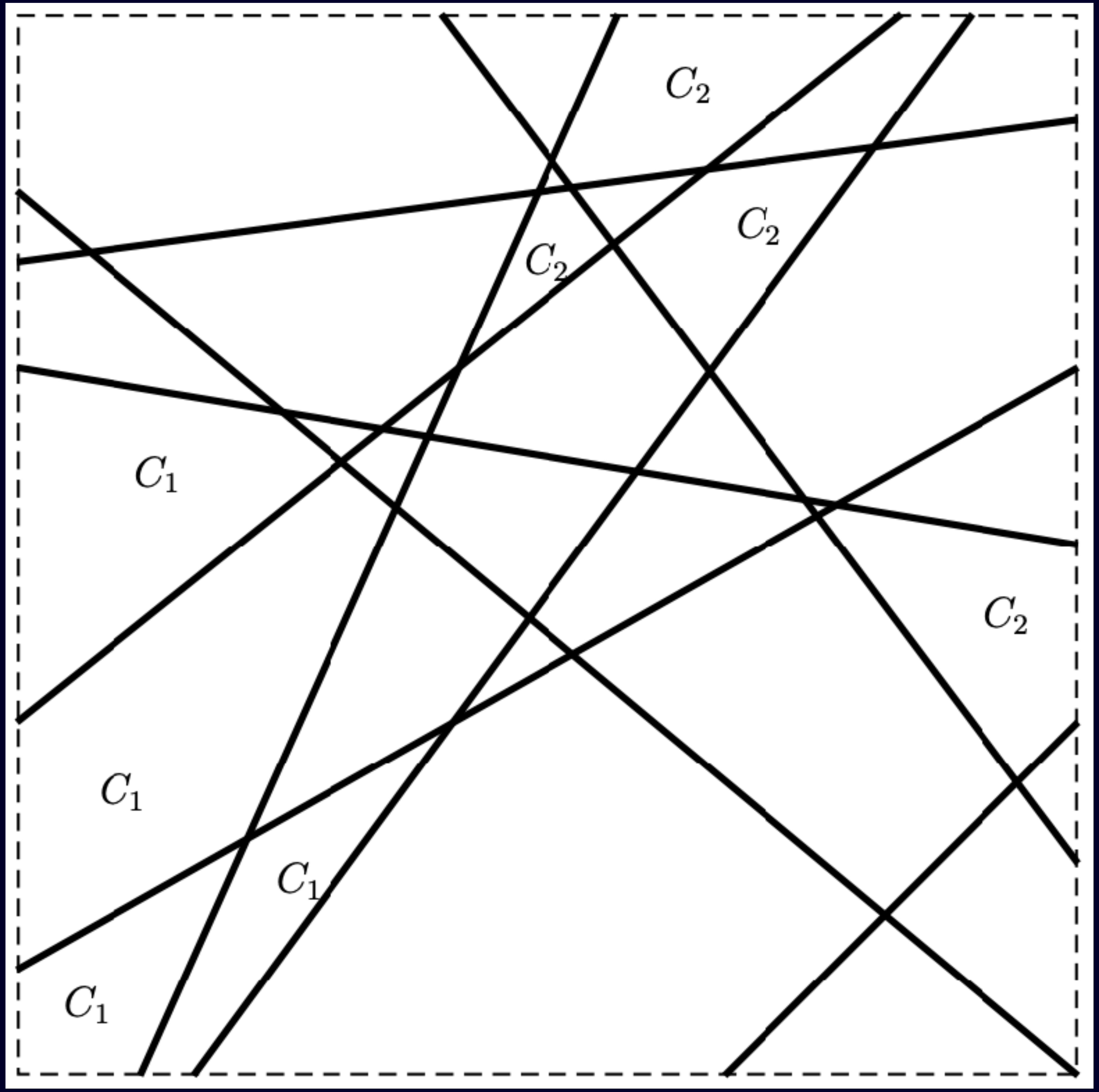
Source tag present ↓ Classified as horse

No source tag present ↓ Not classified as horse

Lapuschkin et al. “Unmasking Clever Hans Predictors and Assessing What Machines Really Learn”, Nature Communications, 2019.

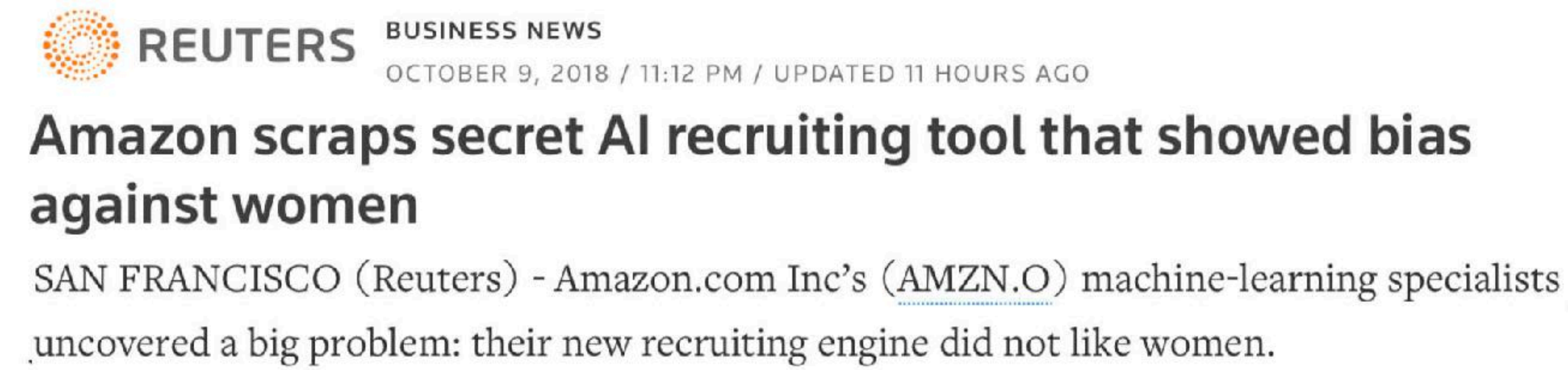
The curse of dimensionality

- With increasing dimension, properties of the space change dramatically:
- Eucleidian distance no longer has much meaning
- We are **always** just a tiny step away from a mistake - in some dimension(s)



Shamir et al., 2019 A Simple Explanation for the Existence of Adversarial Examples with Small Hamming Distance - attack specific for RELU-based NN

Amazon HR system



REUTERS

BUSINESS NEWS

OCTOBER 9, 2018 / 11:12 PM / UPDATED 11 HOURS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

- Text analysis: Huge number of features available to the system.
- Problem: System refuses to hire women candidates (based on the past decisions).
- Fix 1: Explicit sex/gender field removed.
- Fix 2: The system then started using his/hers salutations - clean-up.
- Fix 3: Sports, schools and other hard-to-remove features surfaced…
- Project canceled.

AI Disrupts Finance

- Immediate decisions, anytime
- Better decisions & pricing drive competition
- New markets

- **Immediate convenience**

Fraud Detection for Instant Credit

85% alert volume reduction for fraud team

50% of fraud incidents auto-prevented before approval

15% of previously “non-fraud defaults” identified as fraud

+ Better robustness against new attacks

+ Improved risk scoring thanks to cleaner data

LIST
BULLETPROOF.AI
PLAVEBNÍ
TYRŠOVA
VODÁRENSKÁ
LEGIONÁŘŮ
ERBENOVA
NOVÁ
HUSOVA
K MOSTU
RYBÁŘE
MOST JOSEFA STRAKY
KPT. JAROŠE
FIBICHOVA
HAVLÍČKOVA
28. ŘÍJNA
U SADU
KOSMONAUTŮ
NAD PIVOVAREM
ČESKÁ
JAROSLAVA SEIFERTA
JIRÁSKOVA
17. LISTOPADU
K MOSTU
PRAŽSKÁ
NAD ŠAFRÁNICÍ
ROHELOVA
DVOŘÁKOVA
DEVICE DETAILS

Zlata de Mikitinová

Status: ACCEPTED

Ve Žlábkách 2168, 27601 Mělník, CZ

zlata.de.mikitinova.108@czechfraud.com

+420793098822

Order 2019-04-08T13:47:49

NAME	PRODUCT ID	QUANTITY	CATEGORY	PRICE	GAIN
Motul 5100 4T 10W-40 4 l	af8d098931309...	1	2206	625,-	+ 500,-
Pirelli P6 Cinturato 195/65 R15 91H	379cfd2765f3b...	1	972	1340,-	+1072,-
Doprava DPD				0,-	
Burner SIM card					- 100,-
$ Total damage / gain:				1965,-	1472,-

LIST
BULLETPROOF.AI
TYRŠOVA
LEGIONÁŘŮ
ERBENOVA
NOVÁ
HUSOVA
K MOSTU
RYBÁŘE
MOST JOSEFA STRAKY
KPT. JAROŠE
HAVLÍČKOVA
FIBICHOVA
VODÁRENSKÁ
KOSMONAUTŮ
NAD PIVOVAREM
28. ŘÍJNA
U SADŮ
NAD ŠAFRANICÍ
ČESKÁ
JAROSLAVA SEIFERTA
JIRÁSKOVA
17. LISTOPADU
PRAŽSKÁ
ROHELOVA
K MOSTU
SADY NA POLABÍ
SADOVÁ
FRIČOVA

Zlata de Mikitinová

Status: ACCEPTED

Ve Žlábkách 2168, 27601 Mělník, CZ

zlata.de.mikitinova.108@czechfraud.com

+420793098822

DEVICE DETAILS

None

MacIntel

3840x2160 | 24bit | Intel(R) Iris(TM) Plus Graphics 655

Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_4) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/73.0.3683.86 Safari/537.36

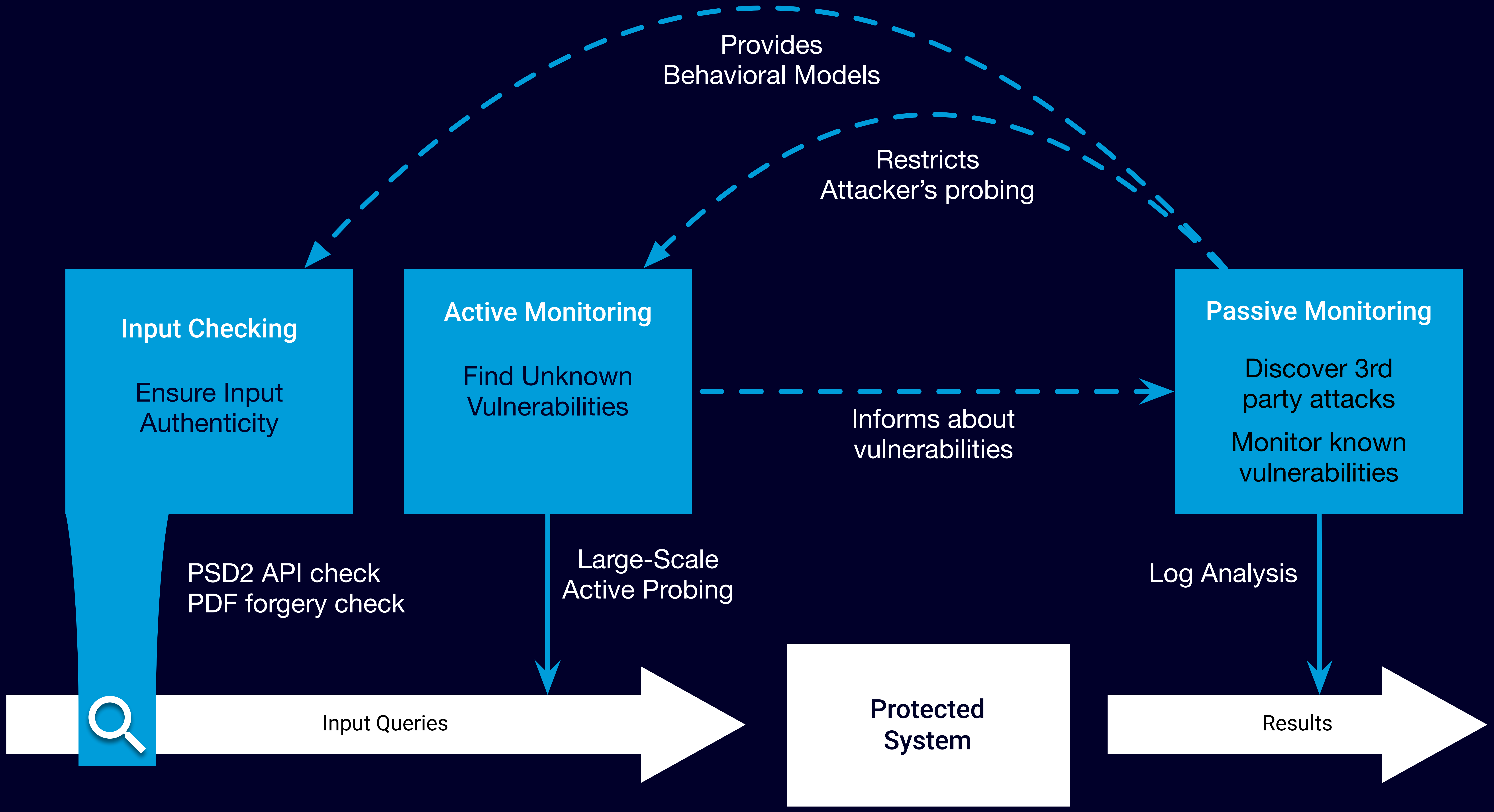
cs-CZ

10.54.81.188

Order 2019-04-08T13:47:49

NAME	PRODUCT ID	QUANTITY	CATEGORY	PRICE	GAIN

Bulletproof AI - Full Solution



Provides
Behavioral Models

Restricts
Attacker's probing

Input Checking

Ensure Input
Authenticity

Active Monitoring

Find Unknown
Vulnerabilities

Informs about
vulnerabilities

Passive Monitoring

Discover 3rd
party attacks

Monitor known
vulnerabilities

PSD2 API check
PDF forgery check

Large-Scale
Active Probing

Log Analysis

Input Queries

**Protected
System**

Results

Bulletproof AI - System Architecture

BULLETPROOF.AI

Prague & Brussels

sales@bulletproof.AI, +420 737 113 153