imperva

AI in a Minefield

Learning from Poisoned Data

Itsik Mantin

Head of Innovation
Imperva

About Myself

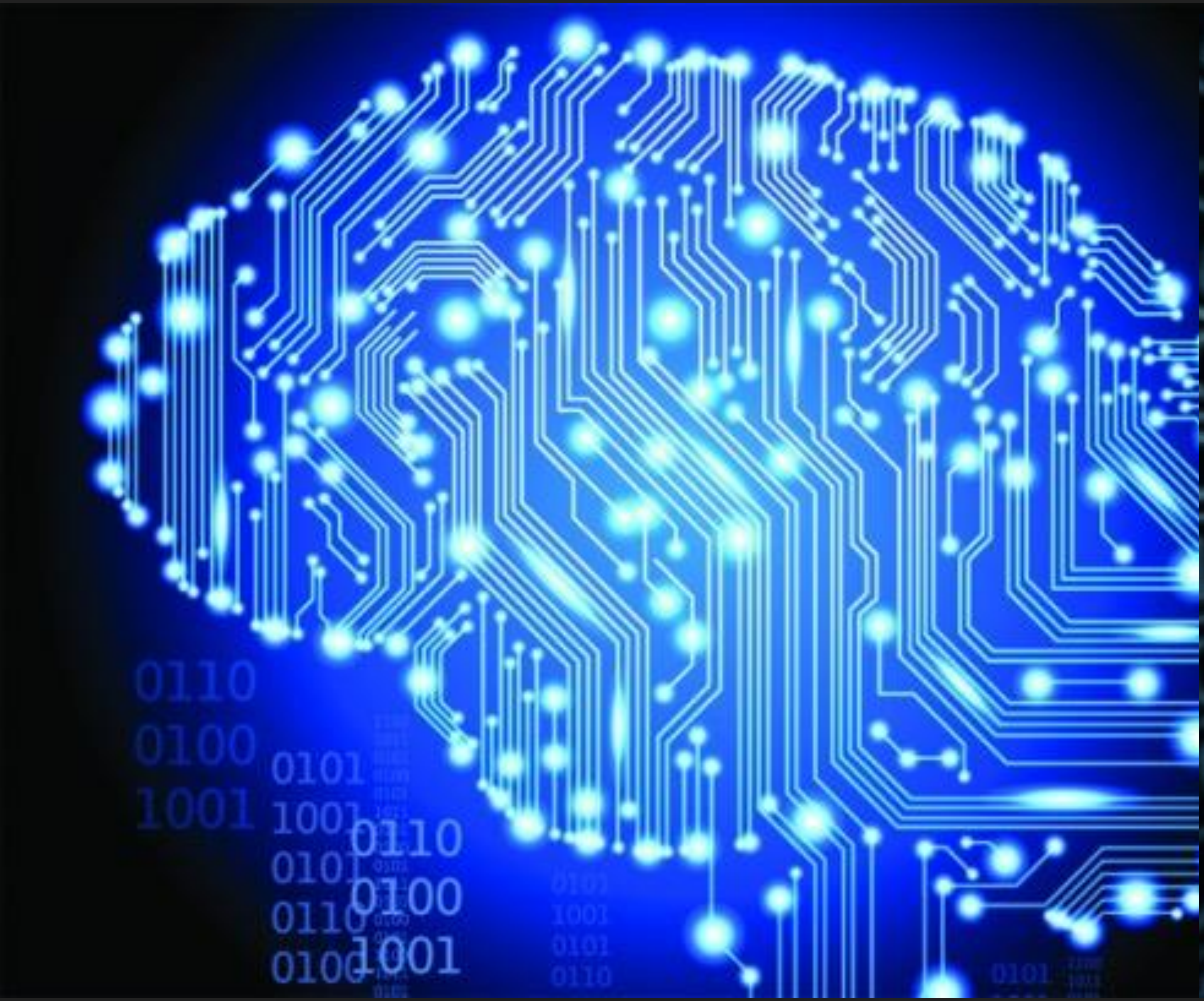
- Since 2000 I've been innovating on security, algorithms and their intersection

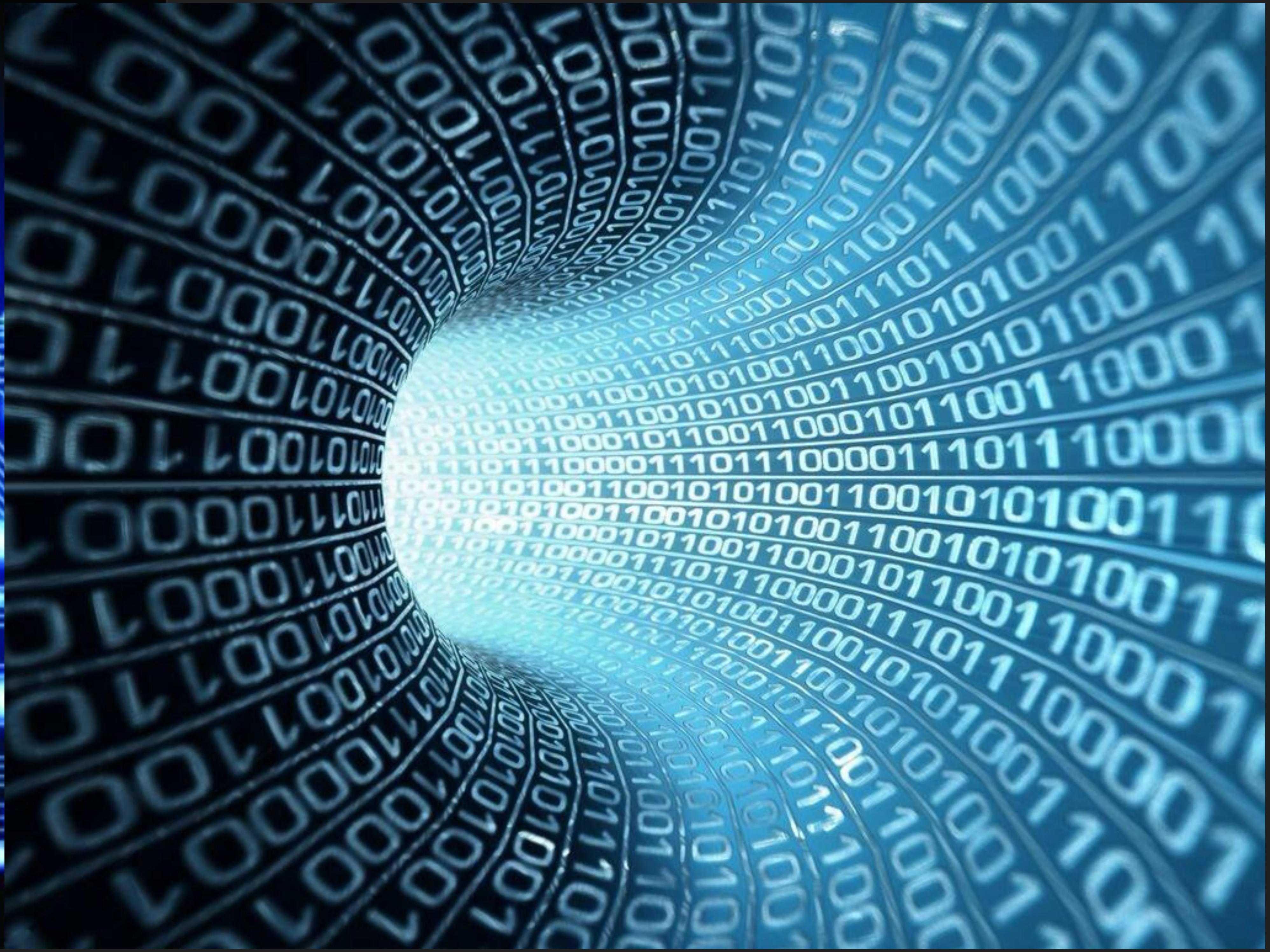
- Love the game of understanding threats and designing mitigation
- Love math and algorithms
- Love building security technology

Outline

- Intro
- AI Risks → AI Threats → Data Poisoning
- Learning from Web Traffic
- Summary

AI Era == Data Era





AI Risks

The **Good**, the **Bad** and the **Ugly**

AI Threats

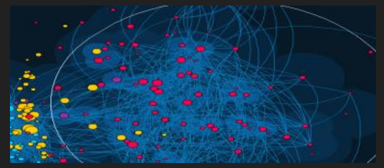
New Attack Surface

Evolutional Attackers' Technology

Automation

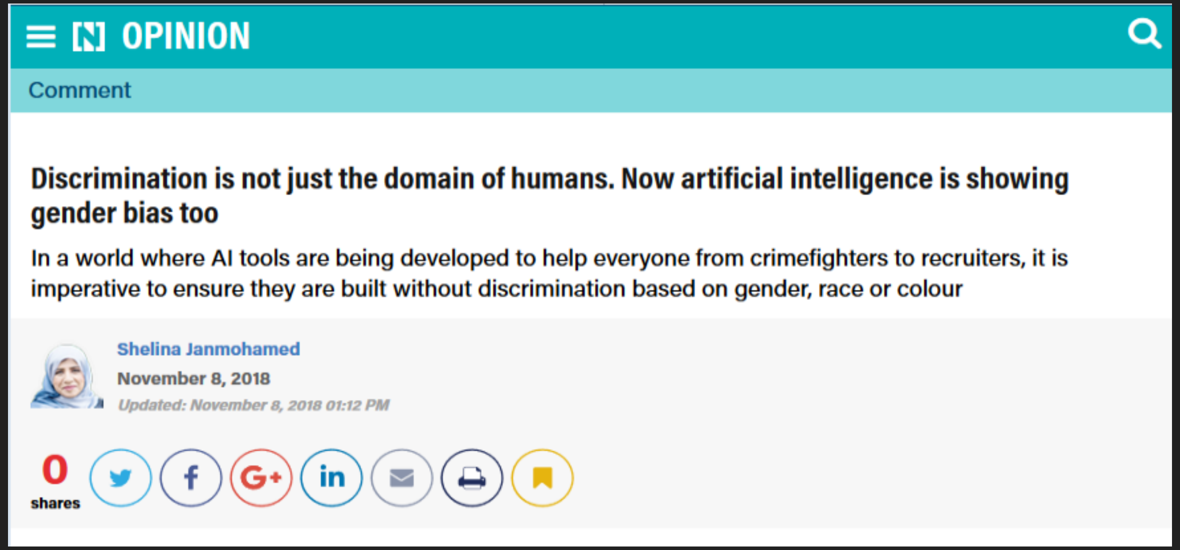
Data mining

Insights



Perpetuation of biases

AI Discrimination



OPINION

Comment

Discrimination is not just the domain of humans. Now artificial intelligence is showing gender bias too

In a world where AI tools are being developed to help everyone from crimefighters to recruiters, it is imperative to ensure they are built without discrimination based on gender, race or colour

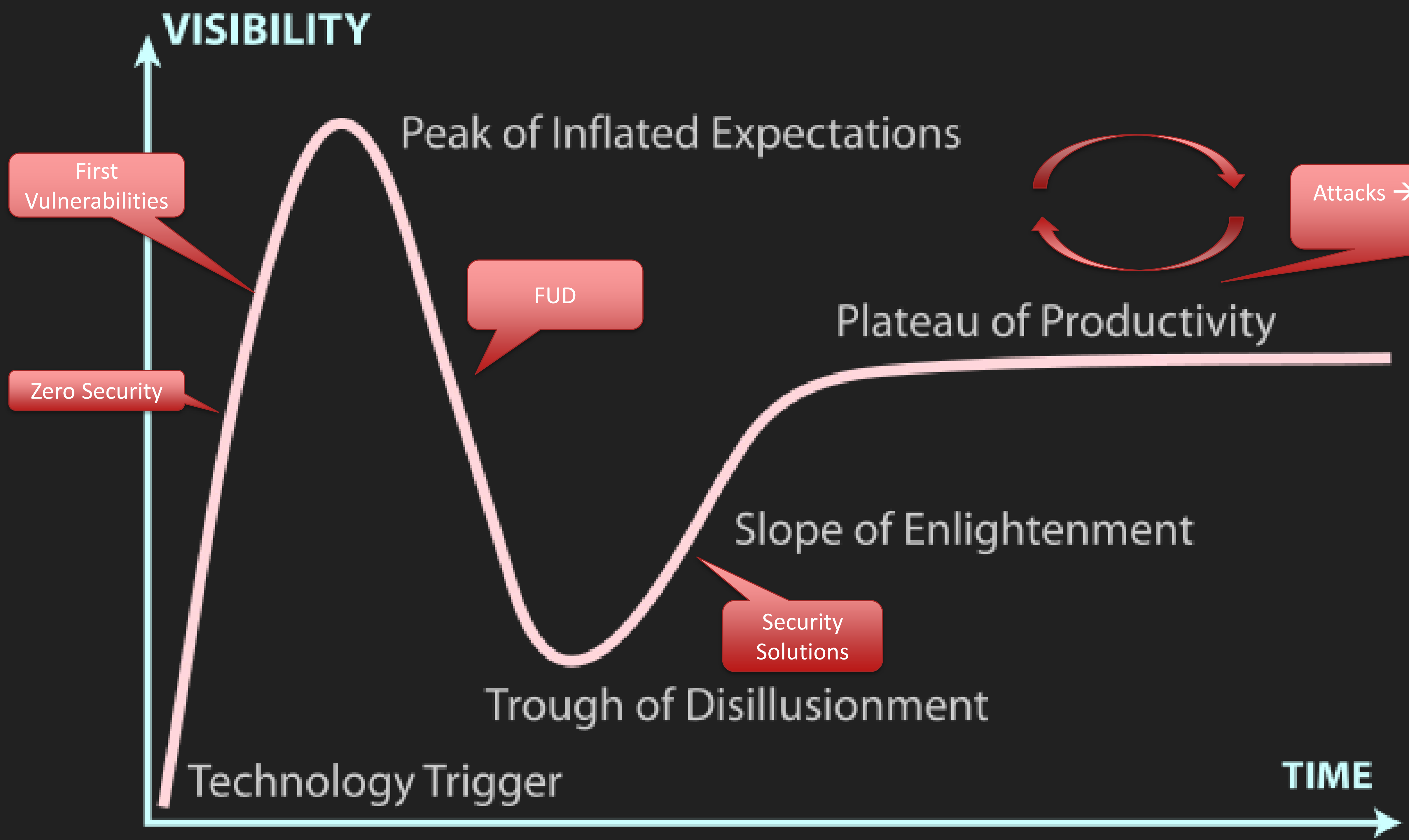
Shelina Janmohamed
November 8, 2018
Updated: November 8, 2018 01:12 PM

0 shares

Advanced Attackers' Technology

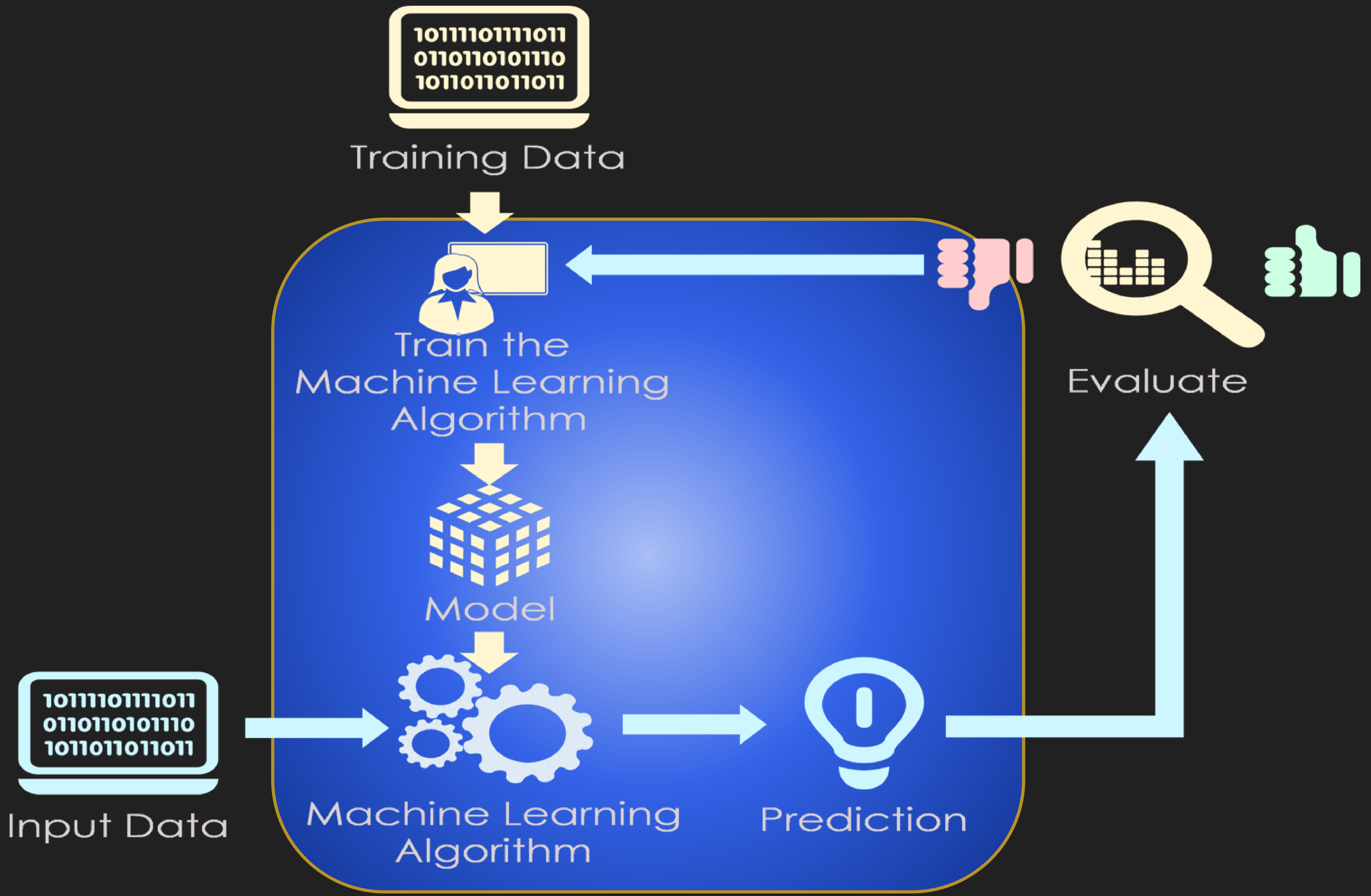
Deep-Fake

The Security Lifecycle of new Technologies



imperva

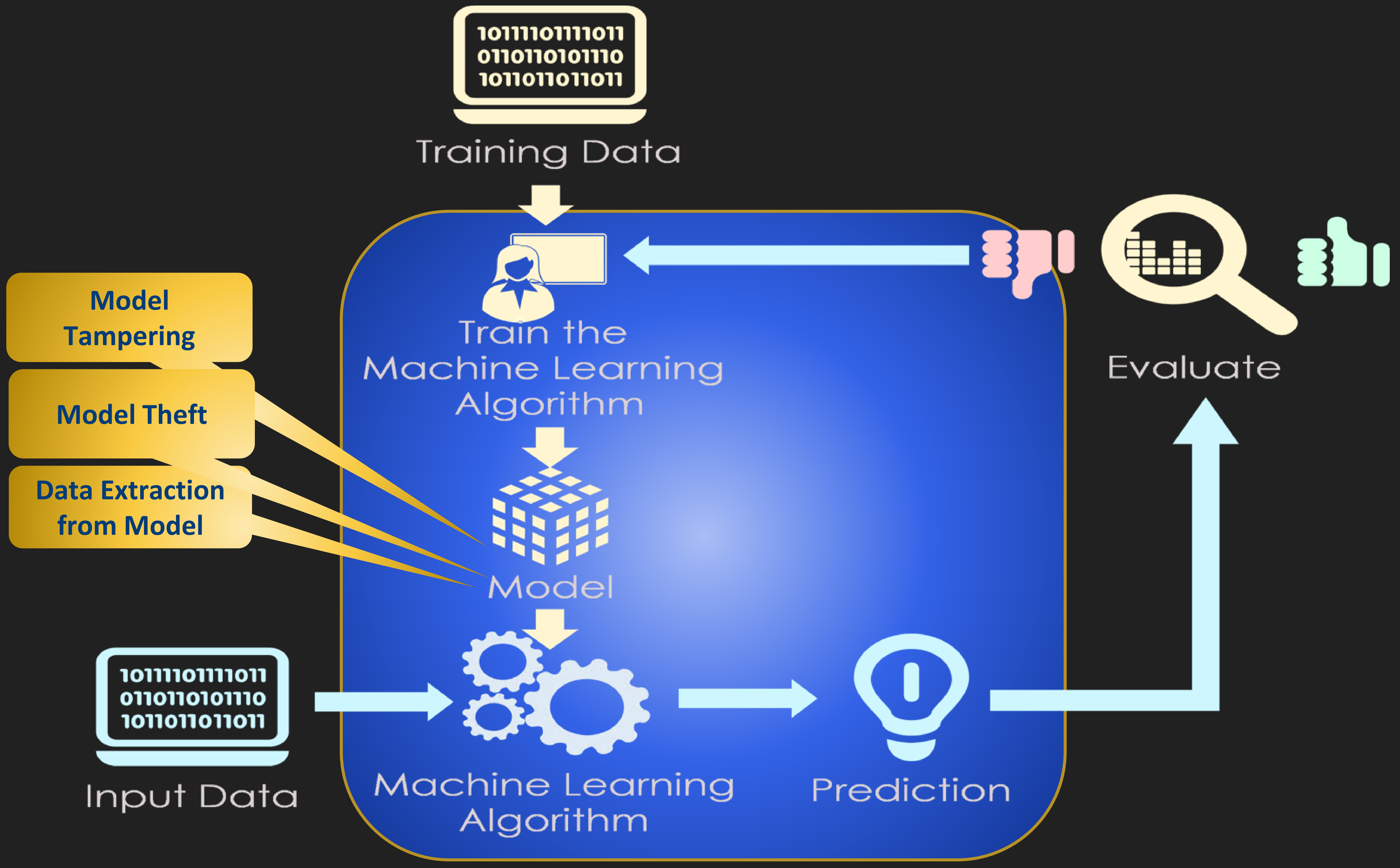
AI Threats



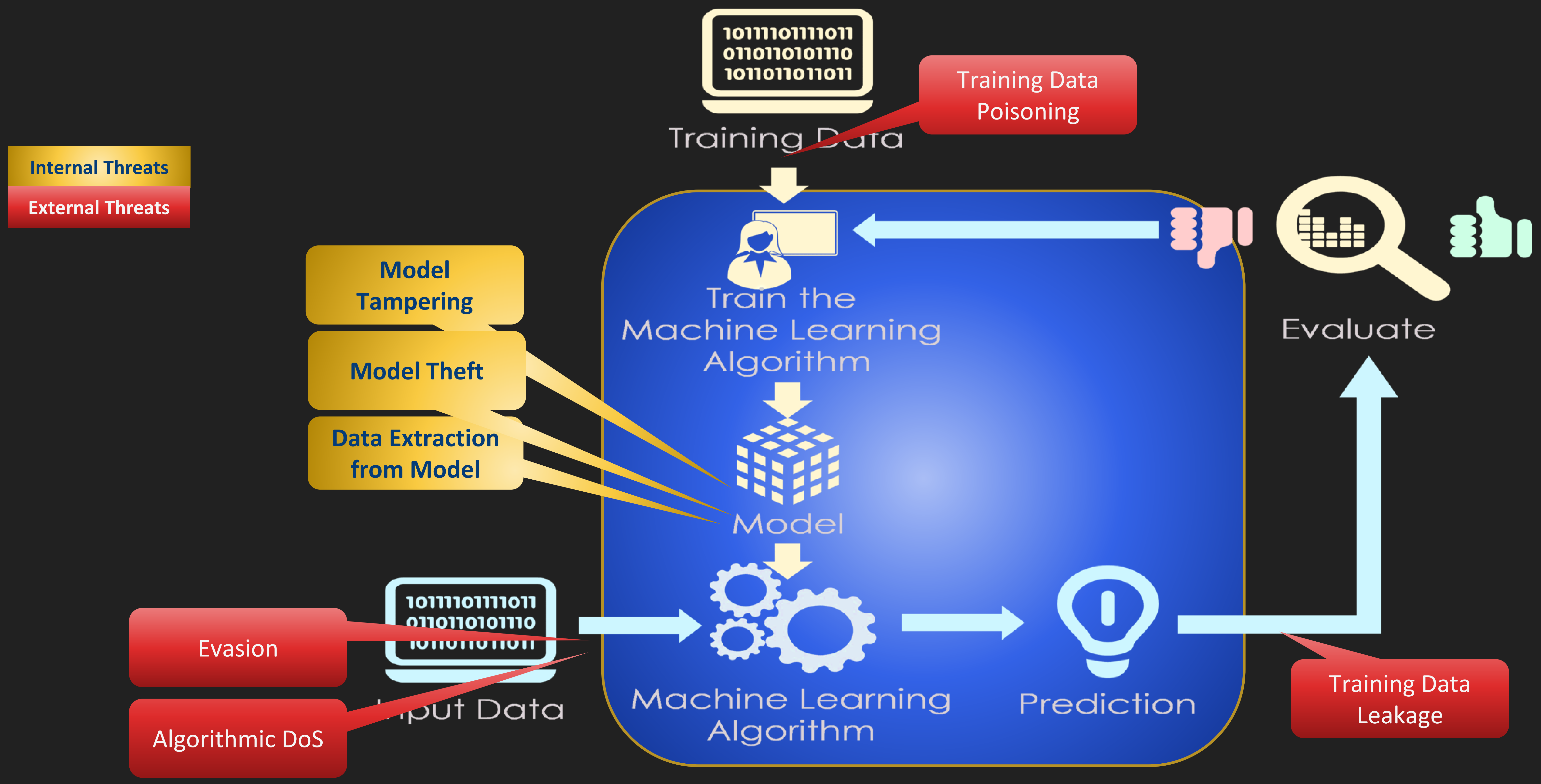
10111101111011
0110110101110
1011011011011
Training Data
Train the
Machine Learning
Algorithm
Model
Evaluate
10111101111011
0110110101110
1011011011011
Input Data
Machine Learning
Algorithm
Prediction

AI Threats

Internal Threats



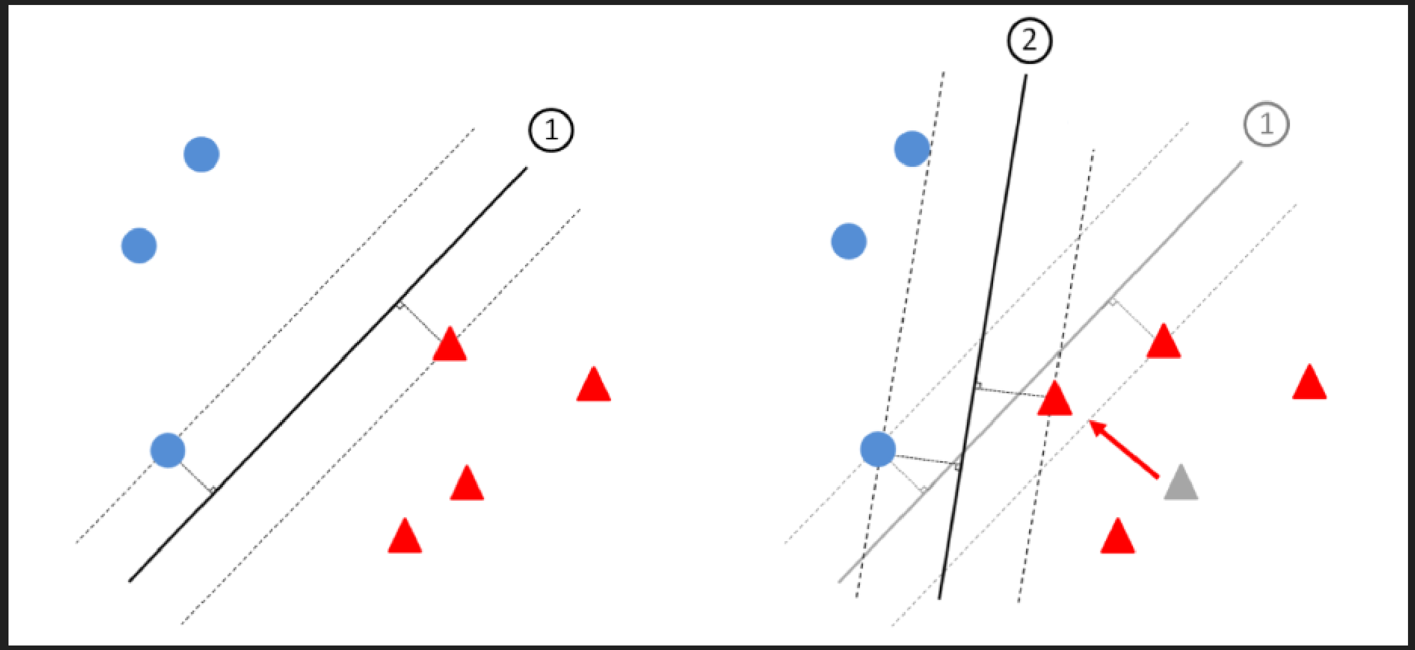
AI Threats



Internal Threats
External Threats
101111011111011
0110110101110
1011011011011
Training Data
Training Data Poisoning
Train the Machine Learning Algorithm
Model Tampering
Model Theft
Data Extraction from Model
Model
Evaluate
101111011111011
0110110101110
Input Data
Evasion
Algorithmic DoS
Machine Learning Algorithm
Prediction
Training Data Leakage

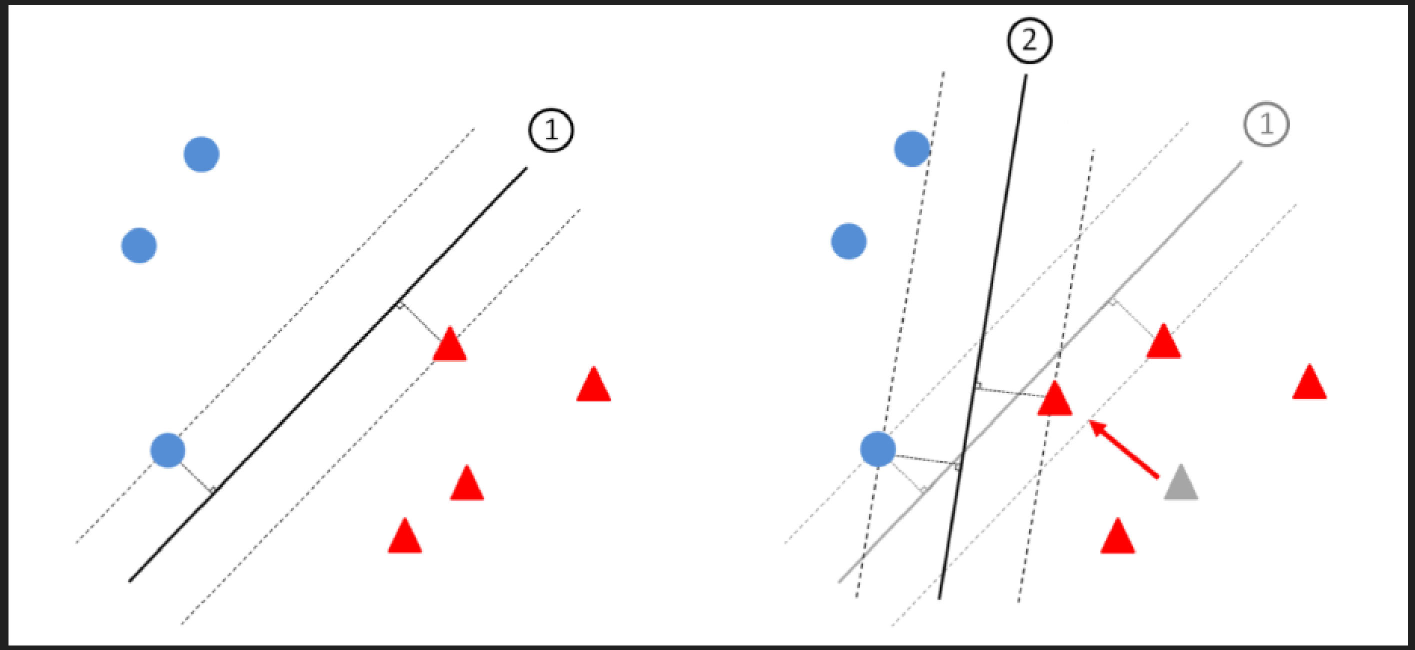
Data Poisoning

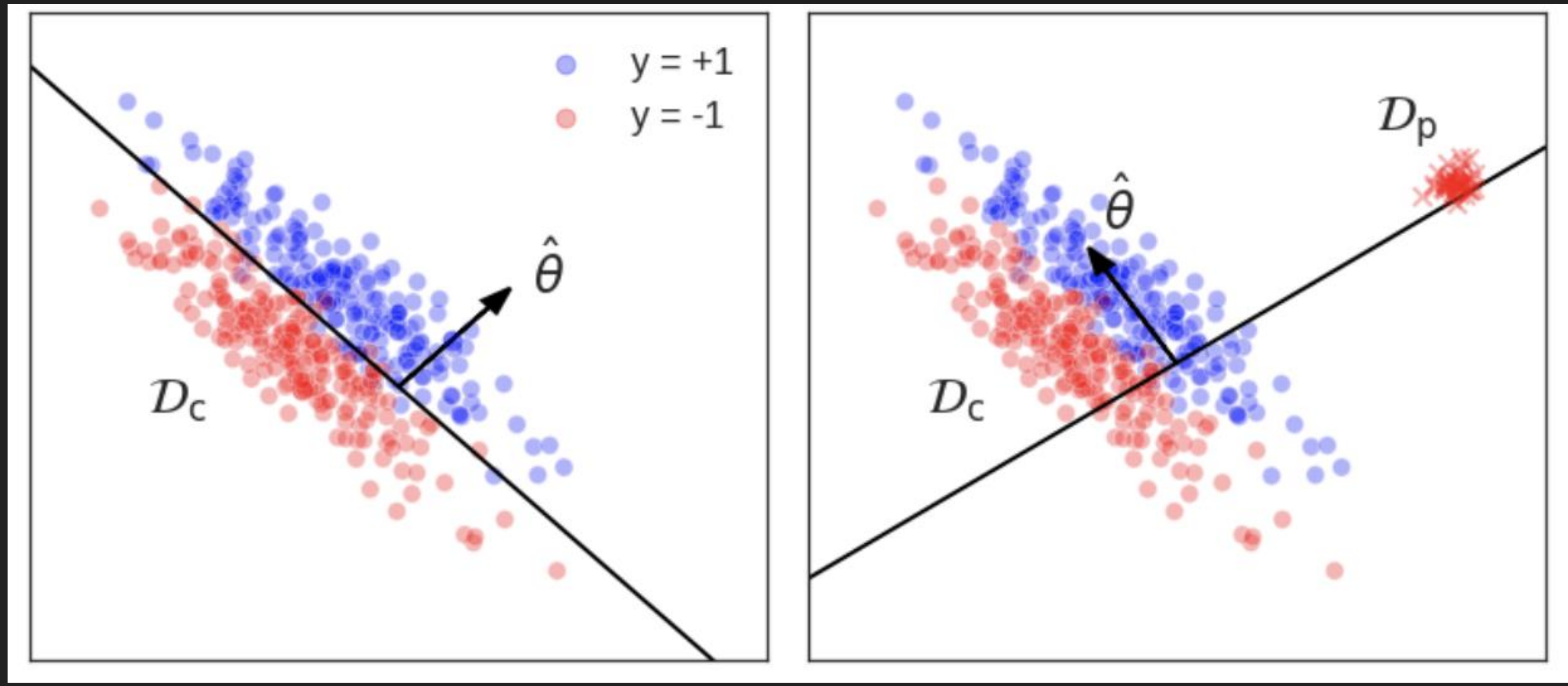
How does it work?



Data Poisoning

How does it work?





Data Poisoning in the Wild

Did you enjoy your vacation?

Support The Guardian

Available for everyone, funded by readers

Contribute → Subscribe →

Search jobs | Sign in | Search | International edition

The Guardian

News | Opinion | Sport | Culture | Lifestyle | More

Travel ▸ UK Europe US

Tripadvisor

This article is more than **2 months old**

TripAdvisor is failing to stop fake hotel reviews, says Which?

Analysis of 250,000 reviews for top-rated hotels finds one in seven with 'hallmarks' of fakes

PA Media

Fri 6 Sep 2019 12.03 BST

109

▲ Old Cairo, Egypt. At the 'best hotel in Cairo', according to TripAdvisor's rankings, 79% of five-star reviews were left by profiles that had made no other posts. Photograph: Mohamed Hossam/EPA

The travel website TripAdvisor is failing to stop fake reviews boosting the rankings of top-rated hotels, Which? has claimed.

The consumer organisation analysed almost 250,000 reviews for the 10 top-

Read The Guardian without interruption on all your devices

Subscribe now

most viewed

Cybertruck: Tesla unveils new pickup truck but windows shatter during demo

Scooter Braun pleads for resolution with Taylor Swift following death threats

Exclusive: Bolsonaro is turning back the clock on Brazil, says Lula da Silva

Imelda Staunton set to replace Olivia Colman in Netflix's The Crown

Grace Millane trial: New

Data Poisoning in the Wild

Model Skewing

- Model skewing for Gmail Spam filter
- Attack includes massive amounts of spam emails mislabeled as BENIGN

imperva

SpamBayes Availability Attack

Computer Science • Published in LEET 2008

Exploiting Machine Learning to Subvert Your Spam Filter

Blaine Nelson, Marco Barreno, +6 authors Kai Xia

The Victim

- SpamBayes spam filter
- Token-based Bayesean network

The Attack

- Make the model learn incorrectly
- Dictionary attack: “push” words to the model spam dictionary

Impact

- 1% data poisoning was sufficient to make the model detect SPAM for 90% of the legit mails

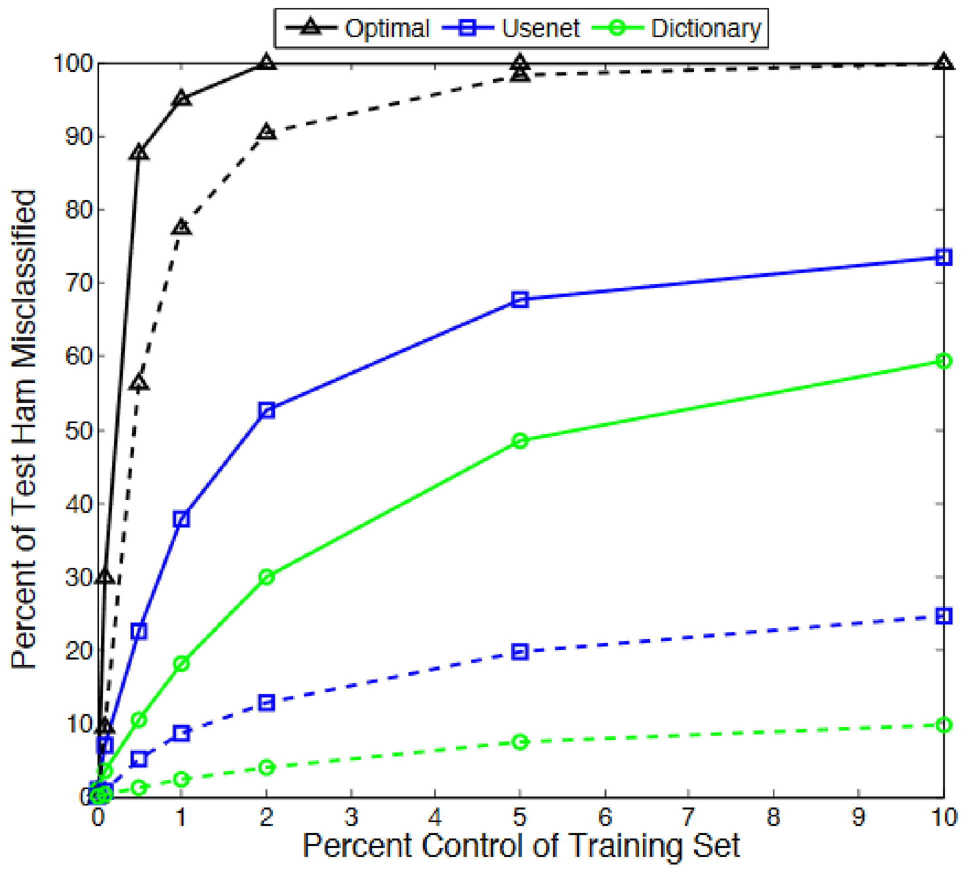


Figure 1: Three dictionary attacks on initial training set of 10,000 messages (50% spam). We plot percent of ham classified as *spam* (dashed lines) and as *spam* or *unsure* (solid lines) against the attack as percent of the training set. We show the optimal attack (black △), the Usenet dictionary attack (blue □), and the Aspell dictionary attack (green ○). Each attack renders the filter unusable with as little as 1% control (101 messages).

SpamBayes Availability Attack

Computer Science · Published in LEET 2008

Exploiting Machine Learning to Subvert Your Spam Filter

Blaine Nelson, Marco Barreno, +6 authors Kai Xia

The Victim

- SpamBayes spam filter
- Token-based Bayesean network

The Attack

- Make the model learn incorrectly
- Dictionary attack: “push” words to the model spam dictionary

Impact

- 1% data poisoning was sufficient to make the model detect SPAM for 90% of the legit mails

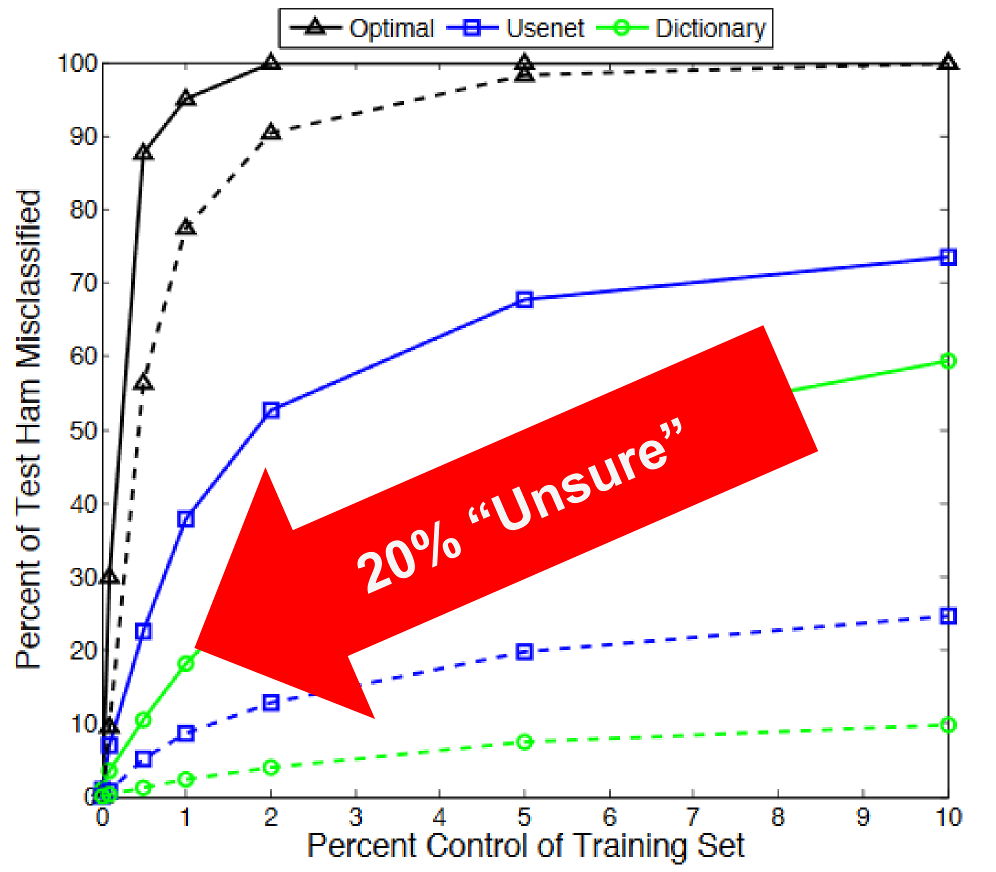


Figure 1: Three dictionary attacks on initial training set of 10,000 messages (50% spam). We plot percent of ham classified as *spam* (dashed lines) and as *spam* or *unsure* (solid lines) against the attack as percent of the training set. We show the optimal attack (black △), the Usenet dictionary attack (blue □), and the Aspell dictionary attack (green ○). Each attack renders the filter unusable with as little as 1% control (101 messages).

Clean-Label Attacks

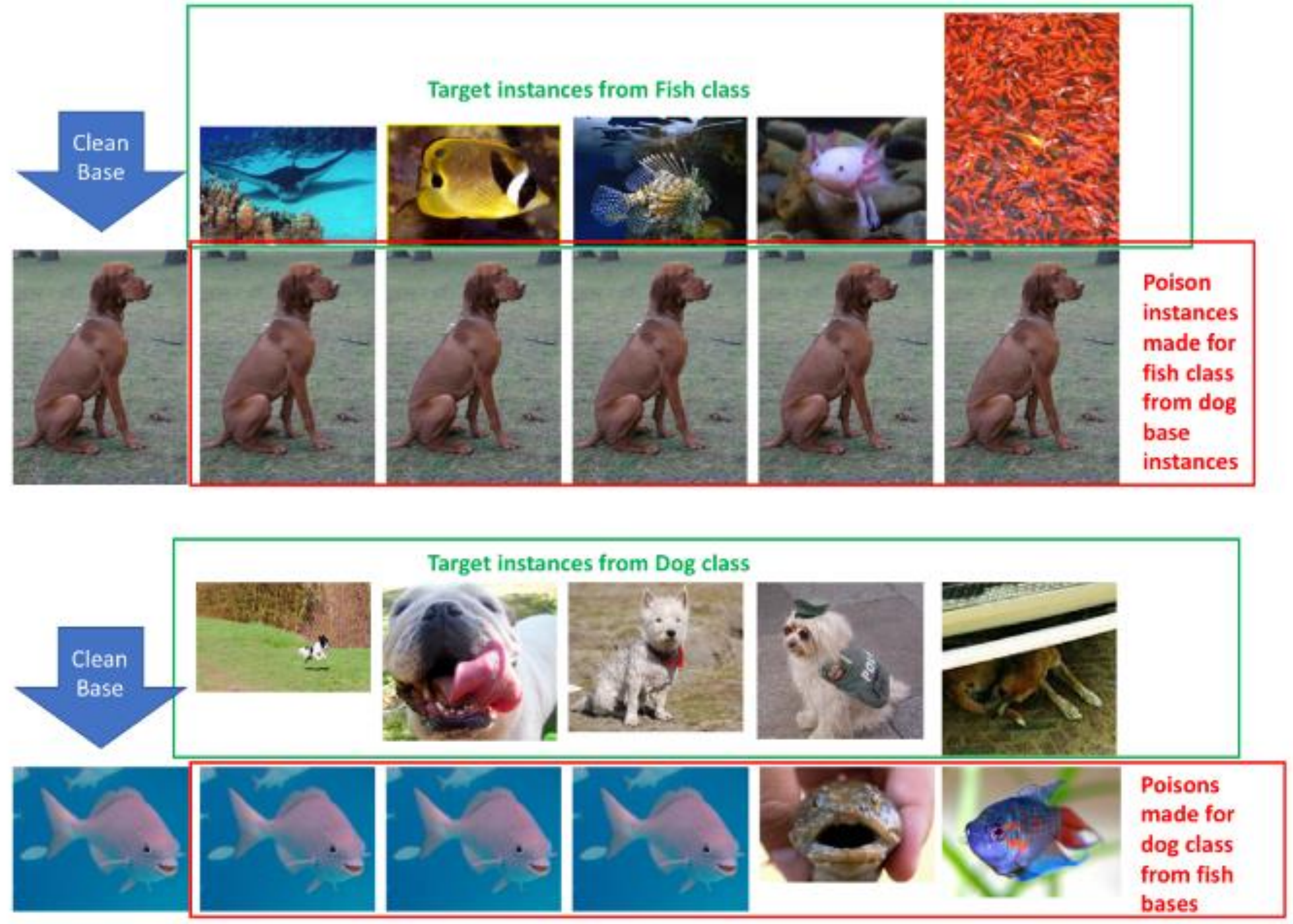
The Victim:

- Image classification

The Attack?

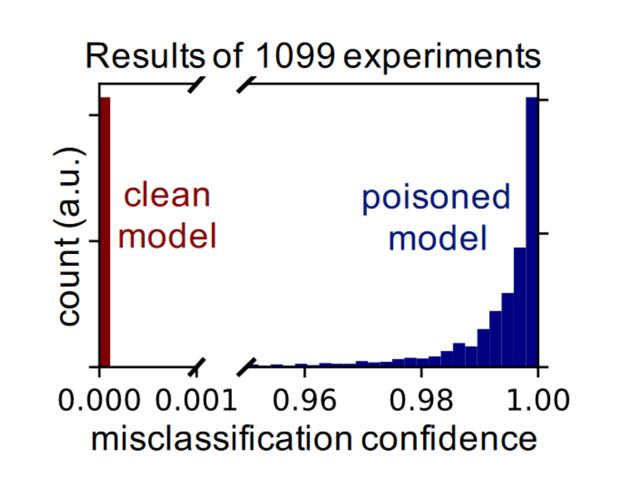
- Craft invisible noise to add to a data sample
- Fail **manual labeling**

The attacker needs zero intervention in the labeling process!



(a) Sample target and poison instances.

Clean-Label Attacks



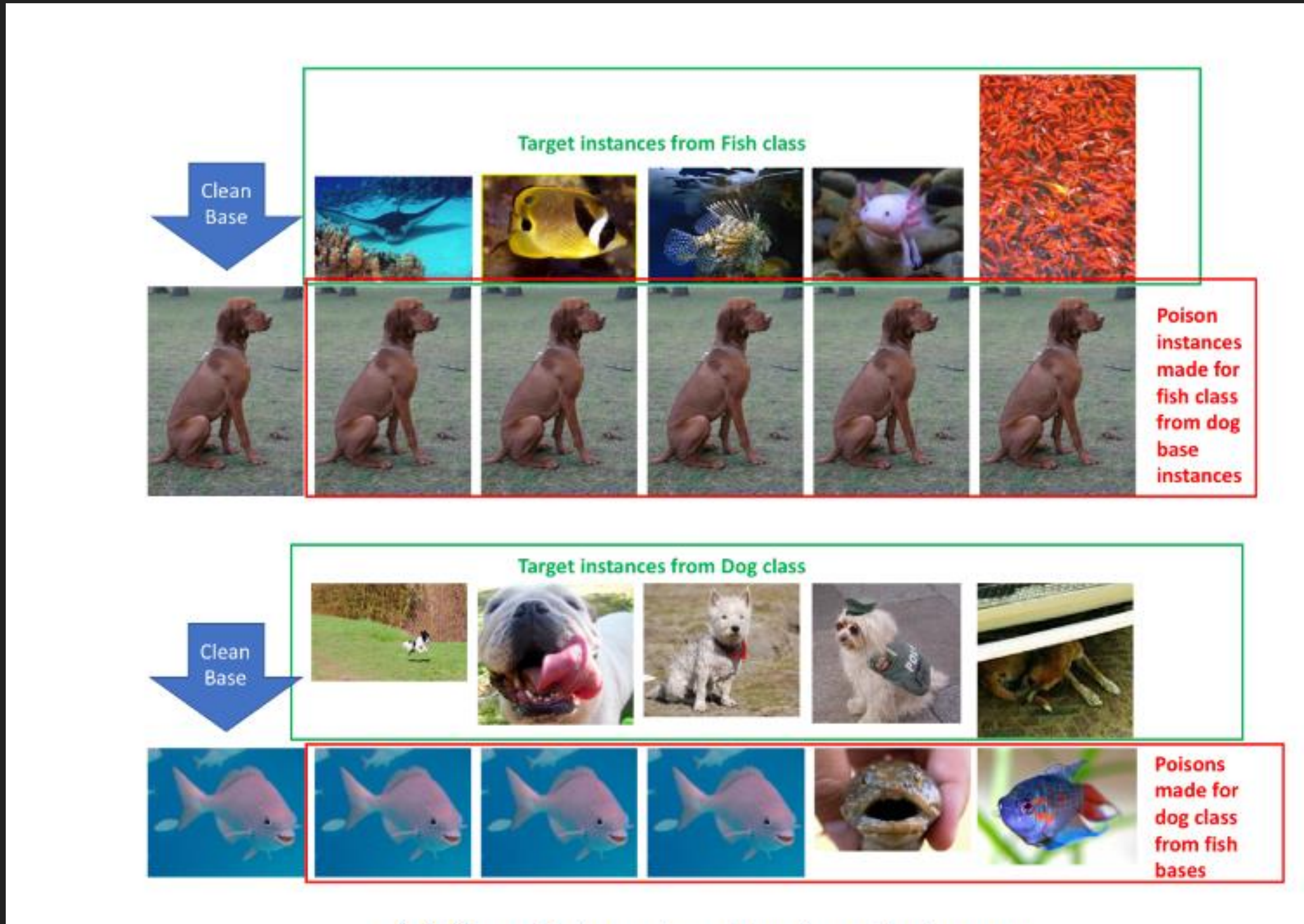
The Victim:

- Image classification

The Attack?

- Craft invisible noise to add to a data sample
- Fail **manual labeling**

The attacker needs zero intervention in the labeling process!



(a) Sample target and poison instances.

Mitigation of Data Poisoning

Filter data from suspicious origins:

E.g., suspicious origins (IP addresses), suspicious clients (bots), etc.
(suspicious data?)

Mitigation of Data Poisoning

Filter data from suspicious origins:

E.g., suspicious origins (IP addresses), suspicious clients (bots), etc.
(suspicious data?)

Fault-tolerant data sampling:

E.g., limit the impact (number, weight) of data points arriving from a single "entity" (user, IP, etc.)

Mitigation of Data Poisoning

Filter data from suspicious origins:

E.g., suspicious origins (IP addresses), suspicious clients (bots), etc. (suspicious data?)

Fault-tolerant data sampling:

E.g., limit the impact (number, weight) of data points arriving from a single "entity" (user, IP, etc.)

Diff-Tracking (Detection)

Look for significant diff from the previous model

Mitigation of Data Poisoning

Filter data from suspicious origins:

E.g., suspicious origins (IP addresses), suspicious clients (bots), etc. (suspicious data?)

Fault-tolerant data sampling:

E.g., limit the impact (number, weight) of data points arriving from a single "entity" (user, IP, etc.)

Diff-Tracking (Detection)

Look for significant diff from the previous model

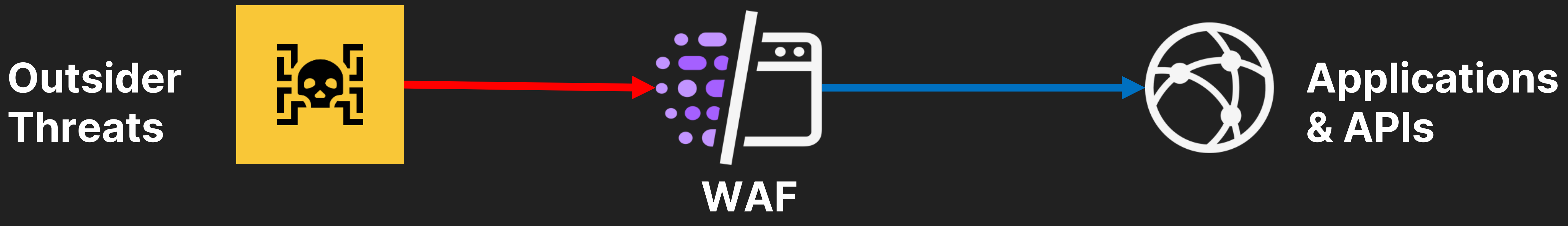
Reliable benchmark (Detection)

Model validation test suite, e.g., accuracy for a certain golden dataset

Summary so far

- Data poisoning is a significant threat on learning mechanisms
- Threat is critical when using data from untrusted sources
- No silver bullet mitigation

Securing Web Applications and APIs



Negative security model

- Aka rule-based security, signature-based
- All’s good except for what we know is bad

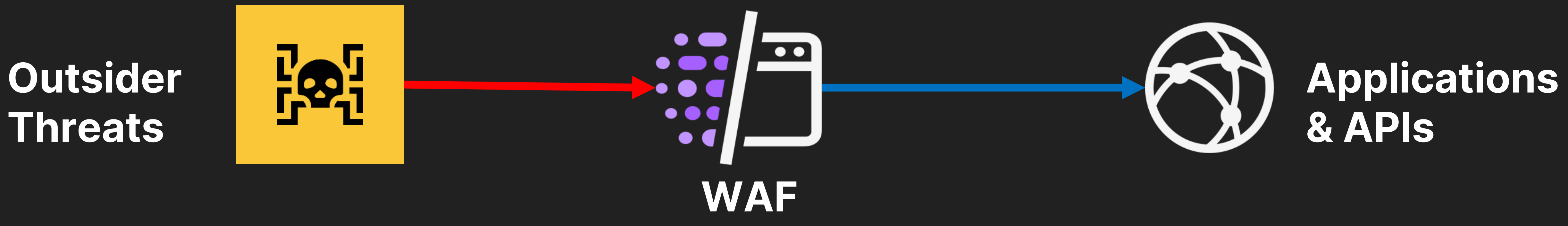
Pros: Accuracy
Cons: Zero-days, ongoing update

Positive security model

- Aka Anomaly Detection
- Learn a baseline profile for Web/API traffic and Block/Alert on deviation
- All’s bad except for what we know is good

Pros: Zero-days, auto-customization
Cons: False-positives

Securing Web Applications and APIs



Negative security model

- Aka rule-based security, signature-based
- All's good except for what we know is bad

Pros: Accuracy
Cons: Zero-days, ongoing update

Positive security model

- Aka Anomaly Detection
- Learn a baseline profile for Web/API traffic and Block/Alert on deviation
- All's bad except for what we know is good

Pros: Zero-days, auto-customization
Cons: False-positives

Data Poisoning

Web/API Traffic Profile

- Body Params
- QS Params
- Cookies
- …

imperva

Web/API Traffic Profile

Object/Container

- Digital Locations (URL/endpoint)
- Hosts
- Methods
- ...

Object

- Body Params
- QS Params
- Cookies
- ...

Web/API Traffic Profile

Object/Container

- Digital Locations (URL/endpoint)
- Hosts
- Methods
- ...

Object

- Body Params
- QS Params
- Cookies
- ...

Object Traffic Profile

- Type
- Multiplicity range
- Optional?
- Mandatory?
- Param size range (num)
- Param charset (str)
- Param Length range (str)
- ...

Threshold-Learning for Web/API Profile

Cleaning
- Filter suspicious traffic

Learning
- Build profile using threshold-learning

Enforcement
- Alert on deviations from profile

- E.g., suspicious events
- E.g., suspicious IPs
- E.g., traffic during attacks
- E.g., traffic from bots

Learn only what you see in requests from
- >= X_1 unique IP addresses
- >= X_2 unique User Agents
- >= X_3 unique Geo-Locations
- >= X_4 unique Identified clients
- >= X_5 unique Hours/Days
- >= X_6 unique Att6
- >= X_7 unique Att7
- …

imperva

Threshold-Learning for Web/API Profile

Cleaning
- Filter suspicious traffic

Learning
- Build profile using threshold-learning

Enforcement
- Alert on deviations from profile

- E.g., suspicious events
- E.g., suspicious IPs
- E.g., traffic during attacks
- E.g., traffic from bots

Learn only what you see in requests from
- >= X_1 unique IP addresses
- >= X_2 unique User Agents
- >= X_3 unique Geo-Locations
- >= X_4 unique Identified clients
- >= X_5 unique Hours/Days
- >= X_6 unique Att6
- >= X_7 unique Att7
- …

Easy in **Batch Processing**, but consumes huge memory

imperva

Dog Food Rating Challenge

Fault-Tolerant Data Sampling



City	Breed	Teo	Pedigree
New York	Pomeranian	Like	
New York	Pomeranian	Like	
Los Angeles	St Bernard		
San Francisco	Pomeranian		Like
New York	Pomeranian	Like	
Los Angeles	St Bernard		
Los Angeles	German Shepherd	Like	Like
San Francisco	Dog Breed		
Los Angeles	Pomeranian	Like	
San Francisco	German Shepherd		
New York	Pomeranian	Like	
San Francisco	St Bernard		
New York	St Bernard		
Los Angeles	German Shepherd	Like	
Los Angeles	Pomeranian	Like	Like
New York	Pomeranian		Like
New York	German Shepherd	Like	
Los Angeles	Pomeranian	Like	
New York	Pomeranian	Like	
New York	St Bernard		Like
	All	**11**	**5**



Raw results:

- Teo: 11 Likes
- Pedigree: 5 Likes

Threshold Learning

- >=3 cities; >=3 breeds
- Only Pedigree pass

imperva

Dog Food Rating Challenge

Fault-Tolerant Data Sampling





City	Breed	Teo	Pedigree
New York	Pomeranian	Like	
New York	Pomeranian	Like	
Los Angeles	St Bernard		
San Francisco	Pomeranian		Like
New York	Pomeranian	Like	
Los Angeles	St Bernard		
Los Angeles	German Shepherd	Like	Like
San Francisco	Dog Breed		
Los Angeles	Pomeranian	Like	
San Francisco	German Shepherd		
New York	Pomeranian	Like	
San Francisco	St Bernard		
New York	St Bernard		
Los Angeles	German Shepherd	Like	
Los Angeles	Pomeranian	Like	Like
New York	Pomeranian		Like
New York	German Shepherd	Like	
Los Angeles	Pomeranian	Like	
New York	Pomeranian	Like	
New York	St Bernard		Like

	All	**11**	**5**
	Pomeranian	**8**	**3**
	St Bernard	**0**	**1**
	German Shepherd	**3**	**1**
San Francisco		**0**	**1**
New York		**6**	**2**
Los Angeles		**5**	**2**

Raw results:

- Teo: 11 Likes
- Pedigree: 5 Likes

Threshold Learning

- >=3 cities; >=3 breeds
- Only Pedigree pass

Dog Food Rating Challenge

Fault-Tolerant Data Sampling

City	Breed	Teo	Pedigree
New York	Pomeranian	Like	
New York	Pomeranian	Like	
Los Angeles	St Bernard		
San Francisco	Pomeranian		Like
New York	Pomeranian	Like	
Los Angeles	St Bernard		
Los Angeles	German Shepherd	Like	Like
San Francisco	Dog Breed		
Los Angeles	Pomeranian	Like	
San Francisco	German Shepherd		
New York	Pomeranian	Like	
San Francisco	St Bernard		
New York	St Bernard		
Los Angeles	German Shepherd	Like	
Los Angeles	Pomeranian	Like	Like
New York	Pomeranian		Like
New York	German Shepherd	Like	
Los Angeles	Pomeranian	Like	
New York	Pomeranian	Like	
New York	St Bernard		Like
	Pomeranian	**10**	
	St Bernard	**5**	
	German Shepherd	**4**	
San Francisco		**4**	
New York		**9**	
Los Angeles		**7**	
	All	**11**	**5**
	Pomeranian	**8**	**3**
	St Bernard	**0**	**1**
	German Shepherd	**3**	**1**
San Francisco		**0**	**1**
New York		**6**	**2**
Los Angeles		**5**	**2**

Raw results:

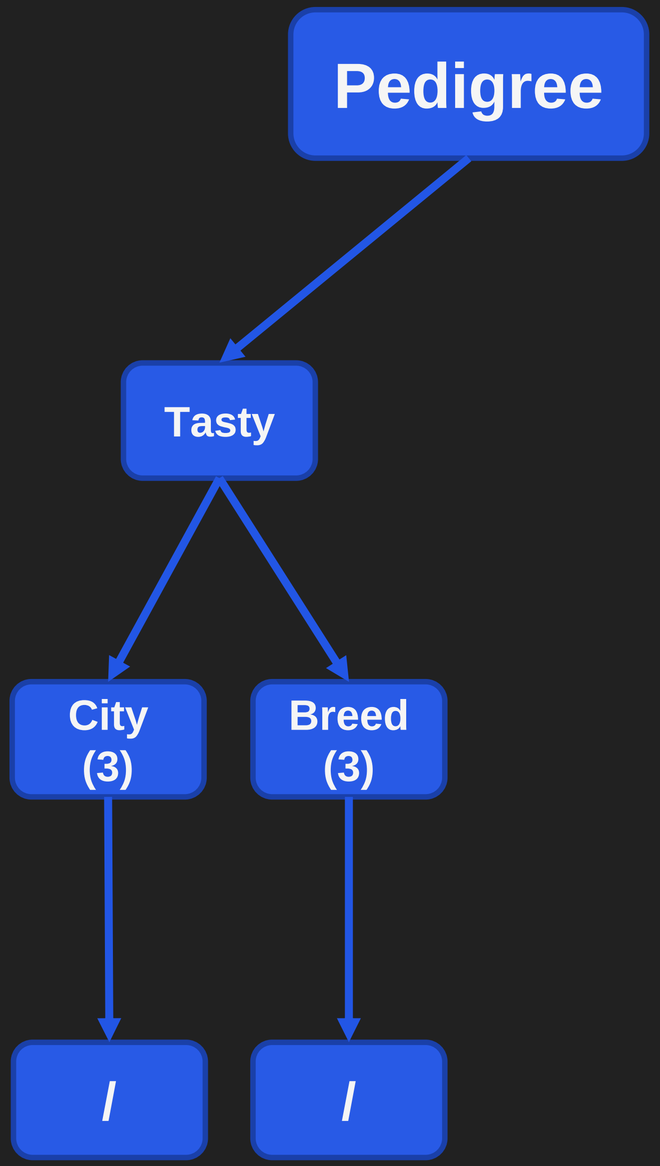
- Teo: 11 Likes
- Pedigree: 5 Likes

Threshold Learning

- >=3 cities; >=3 breeds
- Only Pedigree pass

Threshold-Learning for Boolean Facts

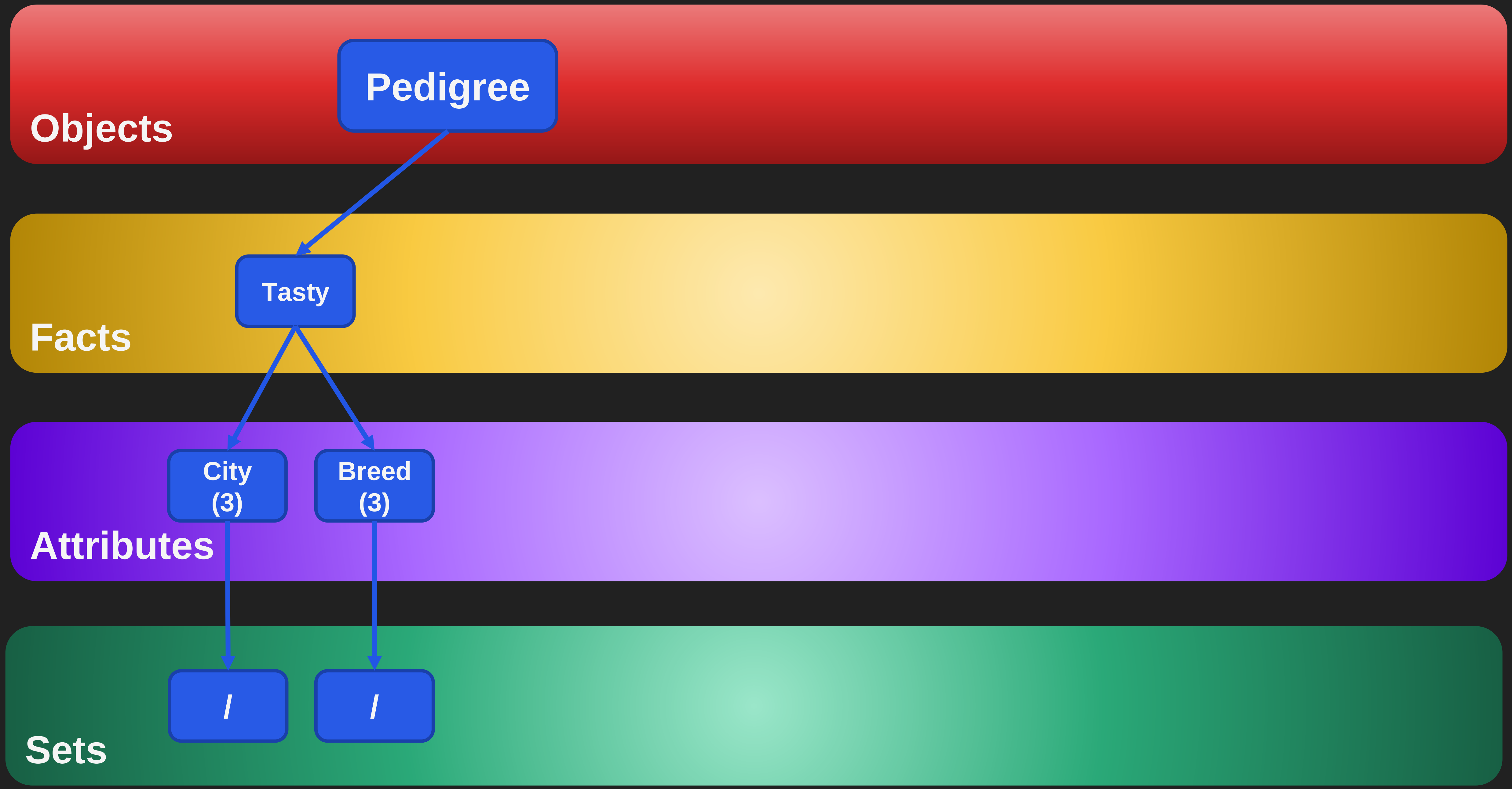
Fixed-Memory Learning



imperva

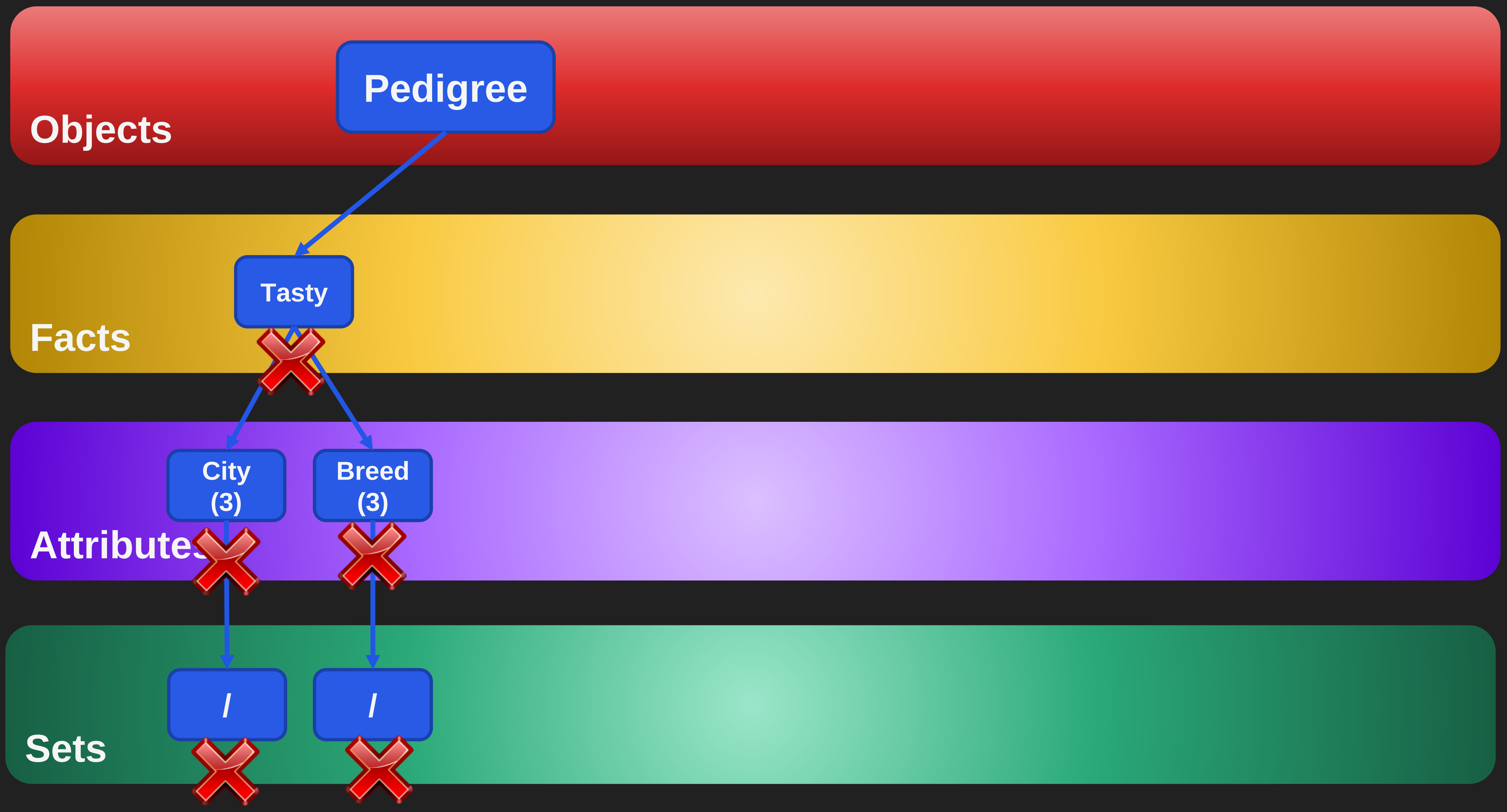
Threshold-Learning for Boolean Facts

Fixed-Memory Learning



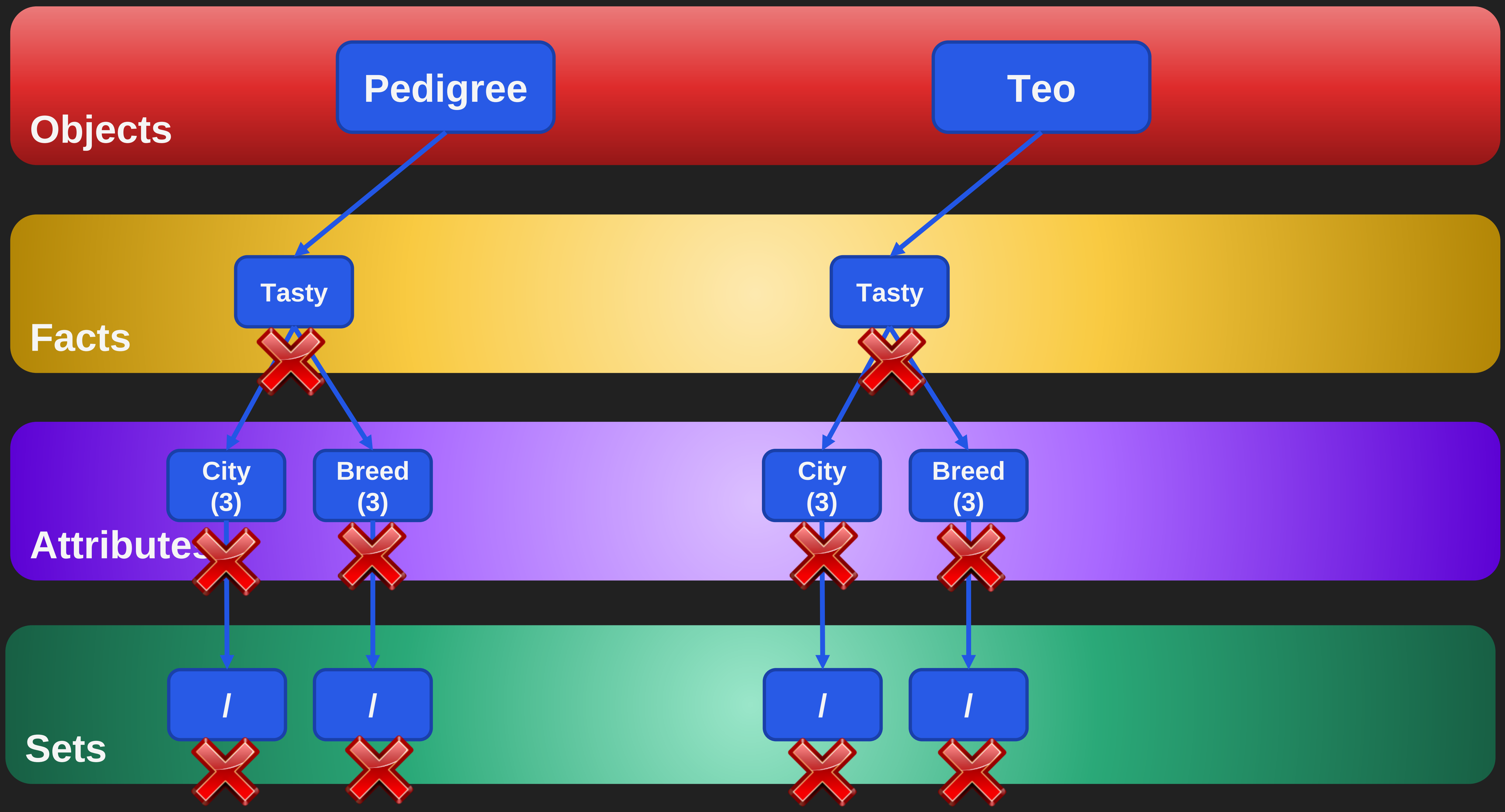
Threshold-Learning for Boolean Facts

Fixed-Memory Learning



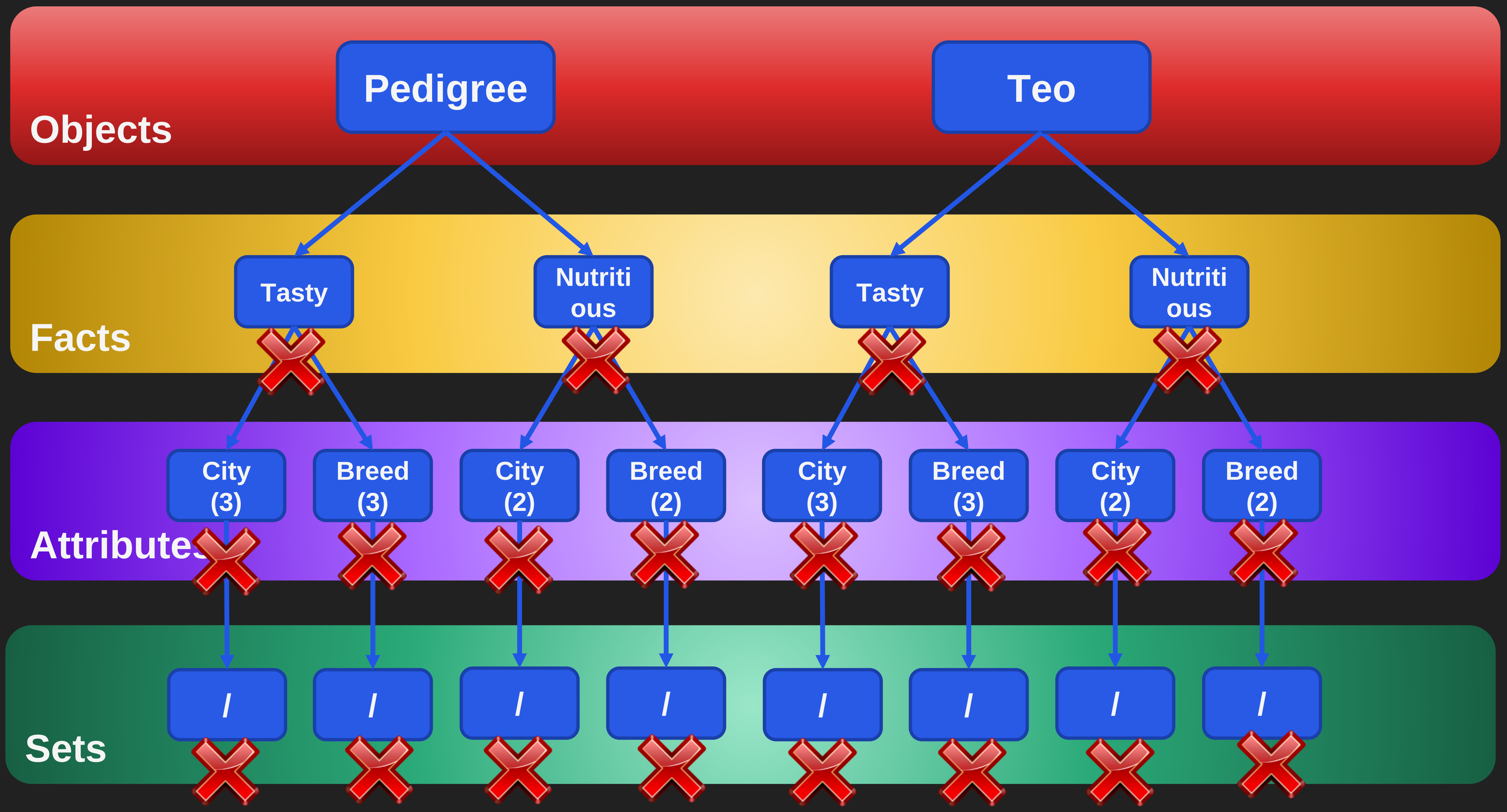
Threshold-Learning for Boolean Facts

Fixed-Memory Learning



Threshold-Learning for Boolean Facts

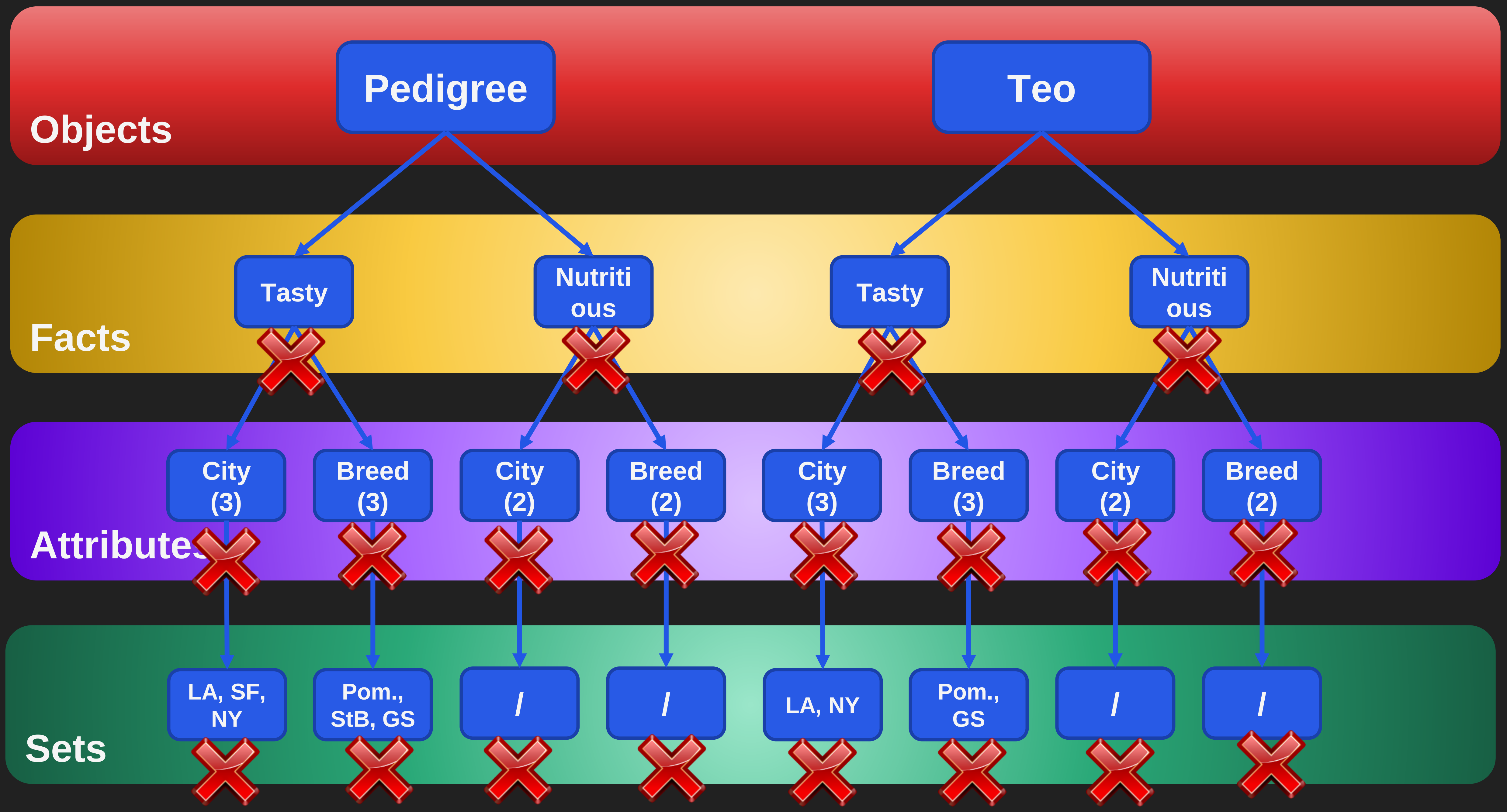
Fixed-Memory Learning



imperva

Threshold-Learning for Boolean Facts

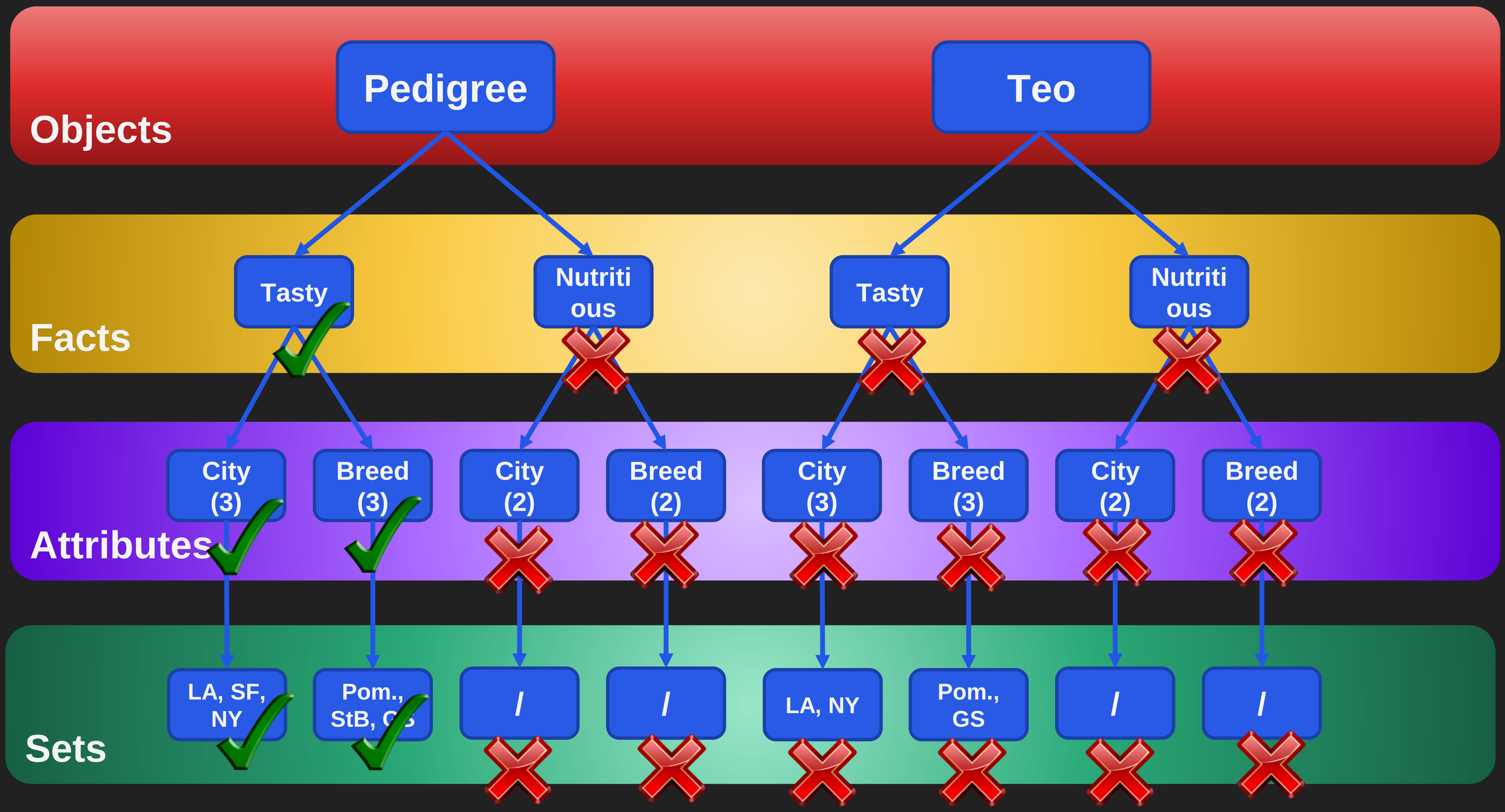
Fixed-Memory Learning



City	Dog Breed	Teo	Pedig.
New York	Pomeranian	Like	
New York	Pomeranian	Like	
Los Angeles	St Bernard		
San Francisco	Pomeranian		Like
New York	Pomeranian	Like	
Los Angeles	St Bernard		
Los Angeles	German Shepherd	Like	Like
San Francisco	Dog Breed		
Los Angeles	Pomeranian	Like	
San Francisco	German Shepherd		
New York	Pomeranian	Like	
San Francisco	St Bernard		
New York	St Bernard		
Los Angeles	German Shepherd	Like	
Los Angeles	Pomeranian	Like	Like
New York	Pomeranian		Like
New York	German Shepherd	Like	
Los Angeles	Pomeranian	Like	
New York	Pomeranian	Like	
New York	St Bernard		Like

Threshold-Learning for Boolean Facts

Fixed-Memory Learning



City	Dog Breed	Teo	Pedig.
New York	Pomeranian	Like	
New York	Pomeranian	Like	
Los Angeles	St Bernard		
San Francisco	Pomeranian		Like
New York	Pomeranian	Like	
Los Angeles	St Bernard		
Los Angeles	German Shepherd	Like	Like
San Francisco	Dog Breed		
Los Angeles	Pomeranian	Like	
San Francisco	German Shepherd		
New York	Pomeranian	Like	
San Francisco	St Bernard		
New York	St Bernard		
Los Angeles	German Shepherd	Like	
Los Angeles	Pomeranian	Like	Like
New York	Pomeranian		Like
New York	German Shepherd	Like	
Los Angeles	Pomeranian	Like	
New York	Pomeranian	Like	
New York	St Bernard		Like

imperva

Threshold-Learning for Boolean Facts

- Can learn:
  - Boolean facts – Object X has property Y
- Memory consumption:
  - Proportional to number of objects and number of properties
  - Proportional to the thresholds
  - But **<u>independent</u>** of the size of the data

In Application/API Profile:

- Learn Flag FACT_X_SEEN
- Enforce Flag FACT_X_ALLOWED

But is this enough? What can you express with Boolean facts?

Just Flag-It!

Expressing Profiling Features with Boolean Facts

Objects (and Containers)

- Digital Locations (URL/endpoint)
- Hosts
- Methods
- Body Params
- QS Params
- Cookies
- …

Just Flag-It!

Expressing Profiling Features with Boolean Facts

Objects (and Containers)

- Digital Locations (URL/endpoint)
- Hosts
- Methods
- Body Params
- QS Params
- Cookies
- …

SITE_HAS_HOST_X

HOST_Y_HAS_URL_X

URL_Y_HAS_COOKIE_X

URL_Y_HAS_METHOD_X

URL_Y_METHOD_Z_HAS_QS_PARAM_X

HOST_Y_HAS_COOKIE_X

…

Just Flag-It!

Expressing Profiling Features with Boolean Facts

Objects (and Containers)

- Digital Locations (URL/endpoint)
- Hosts
- Methods
- Body Params
- QS Params
- Cookies
- ...

SITE_HAS_HOST_X

HOST_Y_HAS_URL_X

URL_Y_HAS_COOKIE_X

URL_Y_HAS_METHOD_X

URL_Y_METHOD_Z_HAS_QS_PARAM_X

HOST_Y_HAS_COOKIE_X

...

What about param profile? Data-Types, Ranges, Char-Set, Regexp?

imperva

Just Flag-It!

Expressing Traffic Profile with Boolean Facts

Object Traffic Profile:

- Type
- Multiplicity range
- Optional?
- Mandatory?
- Param size range (for num)
- Param charset (for str)
- Param Length range (for str)
- ...

Just Flag-It!

Expressing Traffic Profile with Boolean Facts

One-Hot Encoding

Object Traffic Profile:

- Type
- Multiplicity range
- Optional?
- Mandatory?
- Param size range (for num)
- Param charset (for str)
- Param Length range (for str)
- …

Boolean param-type facts:

- NUM_TYPE_ALLOWED
- NON_NUM_TYPE_ALLOWED
- STR_TYPE_ALLOWED
- NON_STR_TYPE_ALLOWED
- NONE_TYPE_ALLOWED
- BOOL_TYPE_ALLOWED
- NON_BOOL_TYPE_ALLOWED
- MAIL_REGEXP_ALLOWED
- NON_MAIL_REGEXP_ALLOWED
- IP_ADD_REGEXP_ALLOWED
- NON_IP_ADD_REGEXP_ALLOWED

Just Flag-It!

Expressing Traffic Profile with Boolean Facts

One-Hot Encoding

Object Traffic Profile:

- Type
- Multiplicity range
- Optional?
- Mandatory?
- Param size range (for num)
- Param charset (for str)
- Param Length range (for str)
- …

Boolean param-type facts:

- NUM_TYPE_ALLOWED
- NON_NUM_TYPE_ALLOWED
- STR_TYPE_ALLOWED
- NON_STR_TYPE_ALLOWED
- NONE_TYPE_ALLOWED
- BOOL_TYPE_ALLOWED
- NON_BOOL_TYPE_ALLOWED
- MAIL_REGEXP_ALLOWED
- NON_MAIL_REGEXP_ALLOWED
- IP_ADD_REGEXP_ALLOWED
- NON_IP_ADD_REGEXP_ALLOWED

Boolean existence facts:

- **MISSING_ALLOWED**
- **MULTI_OCCS_ALLOWED**

Just Flag-It!

Dealing with Sets and Ranges

Object Traffic Profile:

- Type
- Multiplicity range
- Optional?
- Mandatory?
- Param size range (for num)
- Param charset (for str)
- Param Length range (for str)
- ...

Just Flag-It!

Dealing with Sets and Ranges

Object Traffic Profile:

- Type
- Multiplicity range
- Optional?
- Mandatory?
- Param size range (for num)
- Param charset (for str)
- Param Length range (for str)
- …

Boolean charset facts

(one-hot-encoding):

- NON_LETTER_ALLOWED
- NON_DIGIT_ALLOWED
- NON_HEX_ALLOWED
- NON_B64_ALLOWED
- NON_UPPER_ALLOWED
- NON_LOWER_ALLOWED
- ASCII_21_ALLOWED
- ASCII_22_ALLOWED
- ASCII_23_ALLOWED
- …
- ASCII_7E_ALLOWED

imperva

Just Flag-It!

Dealing with Sets and Ranges

Object Traffic Profile:

- Type
- Multiplicity range
- Optional?
- Mandatory?
- Param size range (for num)
- Param charset (for str)
- Param Length range (for str)
- …

Boolean charset facts (one-hot-encoding):

- NON_LETTER_ALLOWED
- NON_DIGIT_ALLOWED
- NON_HEX_ALLOWED
- NON_B64_ALLOWED
- NON_UPPER_ALLOWED
- NON_LOWER_ALLOWED
- ASCII_21_ALLOWED
- ASCII_22_ALLOWED
- ASCII_23_ALLOWED
- …
- ASCII_7E_ALLOWED

Boolean range facts (discretization):

- LENGTH_GT_5_ALLOWED
- LENGTH_GT_50_ALLOWED
- LENGTH_GT_500_ALLOWED
- LENGTH_GT_5000_ALLOWED
- LENGTH_LT_10_ALLOWED
- SIZE_GT_10_ALLOWED
- SIZE_GT_100_ALLOWED
- SIZE_GT_1000_ALLOWED
- SIZE_GT_10000_ALLOWED
- …

Summary and Conclusions

- Data poisoning is a significant threat on learning mechanisms
- Threshold-based learning may provide an adequate robust learning solution
- The Boolean facts framework provides a streaming-friendly implementation for Threshold-based Learning
- Many features can be expressed with Boolean facts

imperva

Thank You!