

# Confidence in AI systems

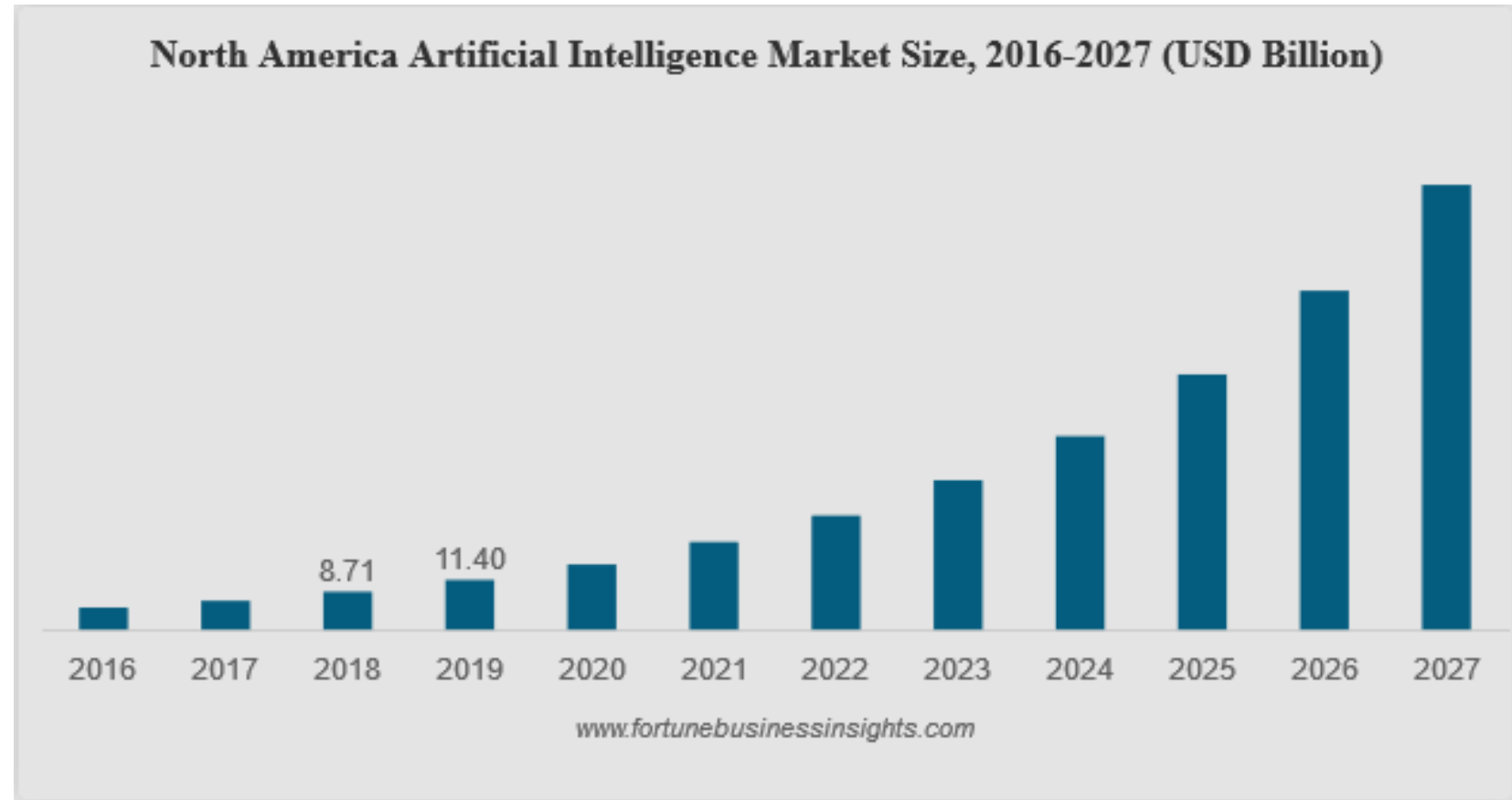
## Can we trust AI-based systems?

*N. Asokan*

 <https://asokan.org/asokan/>

 *@nasokan*

# AI will be pervasive



<https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114>

Forbes

7,109 views | Oct 18, 2019, 01:56pm EDT

# How Artificial Intelligence Is Advancing Precision Medicine



**Nicole Martin** Former Contributor

AI & Big Data

*I write about digital marketing, data and privacy concerns.*

<https://www.forbes.com/sites/nicolemartin1/2019/10/18/how-artificial-intelligence-is-advancing-precision-medicine/#2f720a79a4d5>

**MOTHERBOARD**  
TECH BY VICE

# Dozens of Cities Have Secretly Experimented With Predictive Policing Software

Documents obtained by Motherboard requests verify previously unconfirmed reports that dozens of cities have experimented with predictive policing company Palantir's software.



By **Caroline Haskins**

[https://www.vice.com/en\\_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software](https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software)

Forbes

5,705 views | Oct 31, 2019, 02:42pm EDT

# How AI Is Uprooting Recruiting



**Falon Fatemi** Contributor

Entrepreneurs

PART OF A ZDNET SPECIAL FEATURE: **CYBERSECURITY: LET'S GET TACTICAL**

# AI is changing everything about cybersecurity, for better and for worse. Here's what you need to know

Artificial intelligence and machine learning tools could go a long way to helping to fight cybercrime. But these technologies aren't a silver bullet, and could also be exploited by malicious hackers.

<https://www.zdnet.com/article/ai-is-changing-everything-about-cybersecurity-for-better-and-for-worse-heres-what-you-need-to-know/>



[https://www.vice.com/en\\_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software](https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software)

# How do we evaluate AI-based systems?

## Effectiveness

- measures of [accuracy](#)

*Ground truth*

		<i>Predicted</i>				
		P1	P2	P3	P4	...
<i>Ground truth</i>	P1					...
	P2					
	P3					
	P4					
	...		...			

# How do we evaluate AI-based systems?

## Effectiveness

- measures of **accuracy**

## Performance

- **inference speed** and **memory consumption**

*Ground truth*

		<i>Predicted</i>				
		P1	P2	P3	P4	...
P1	99.2	0.1	0.0	0.2	...	
P2	0.2					
P3	0.0					
P4	0.1					
...	...					

...

# How do we evaluate AI-based systems?

## Effectiveness

- measures of **accuracy**

## Performance

- **inference speed** and **memory consumption**

*Ground truth*

		<i>Predicted</i>				
		P1	P2	P3	P4	...
P1	99.2	0.1	0.0	0.2	...	
P2	0.2					
P3	0.0					
P4	0.1					
...	...					

...

**Trustworthy AI: Meet these criteria even in the presence of adversarial behaviour**



# Challenges in making AI trustworthy

Security concerns

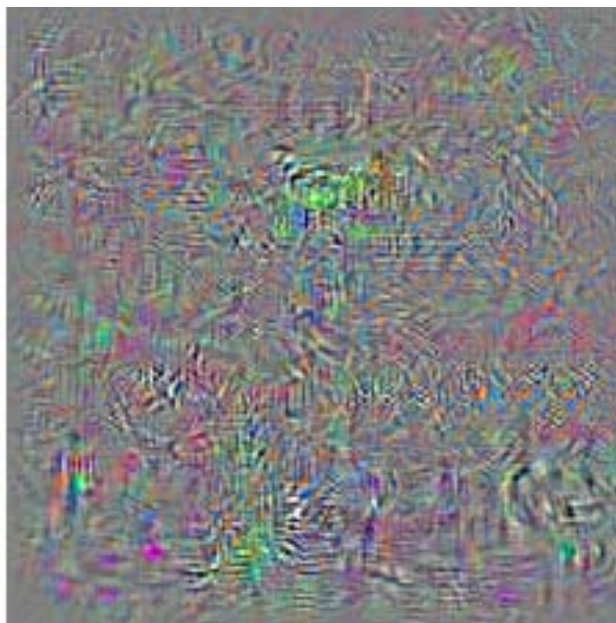
Privacy concerns

# Evading machine learning models



Which class is this?  
**School bus**

**+ 0.1 ·**



**=**



Which class is this?  
**Ostrich**

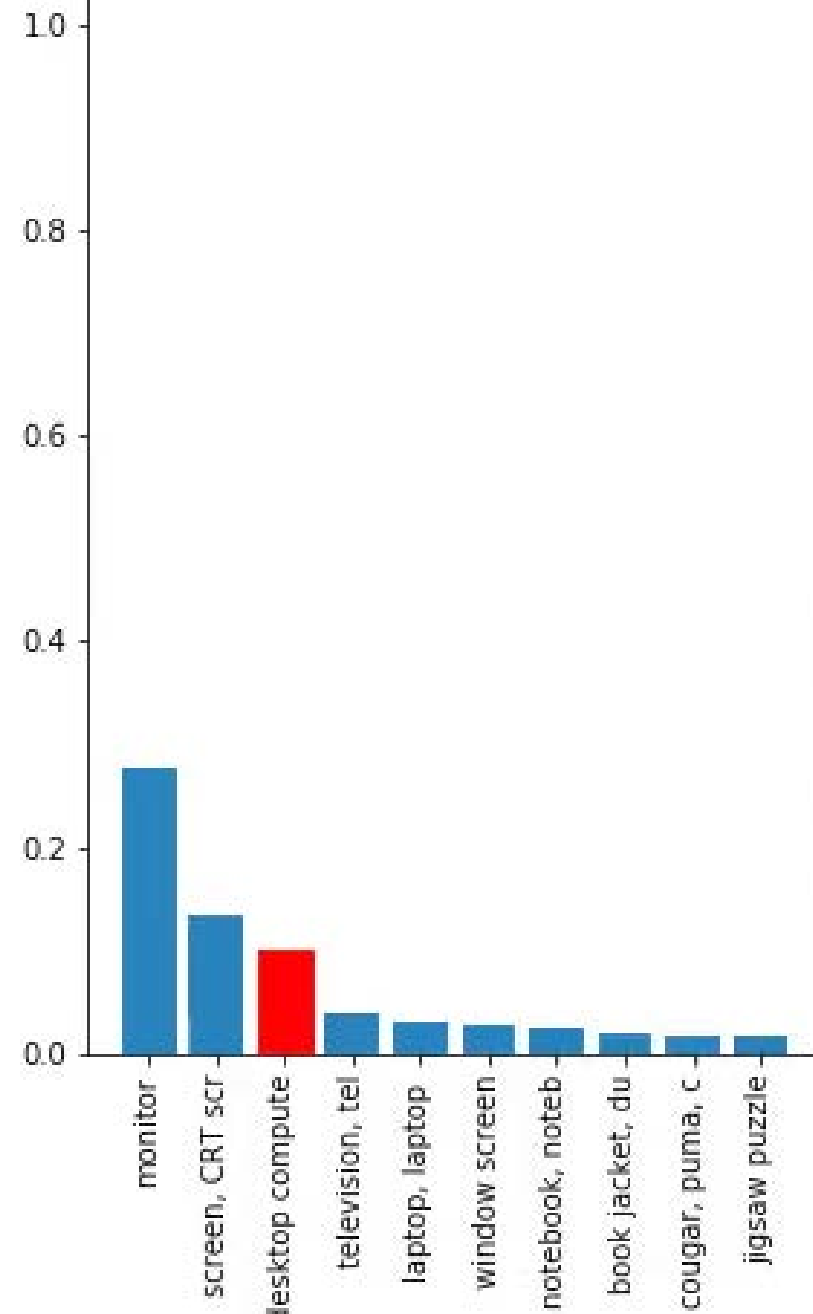




Which class is this?  
**Cat**



Which class is this?  
**Desktop computer**





智能系统安全实验室  
UBIQUITOUS SYSTEM SECURITY LAB.

# DolphinAttack: Inaudible Voice command

Guoming Zhang   Chen Yan   Xiaoyu Ji

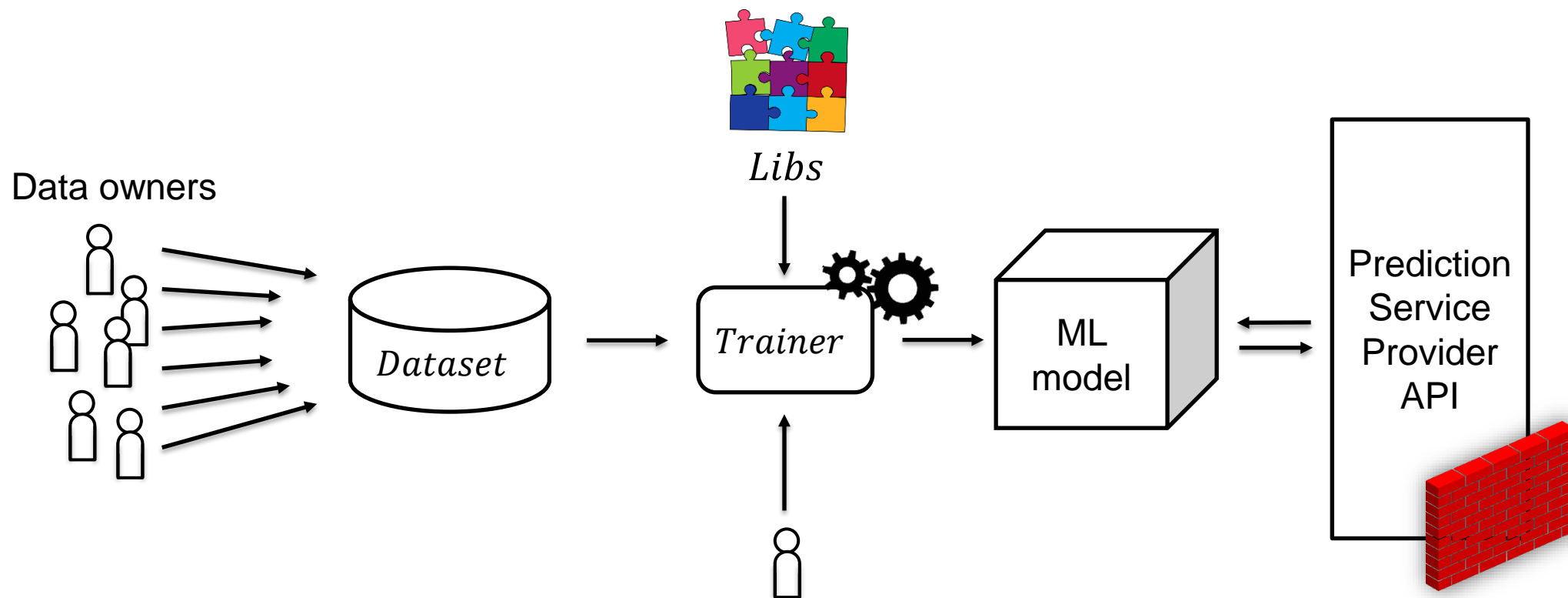
Tianchen Zhang   Taimin Zhang   Wenyuan Xu

Zhejiang University

ACM CCS 2017



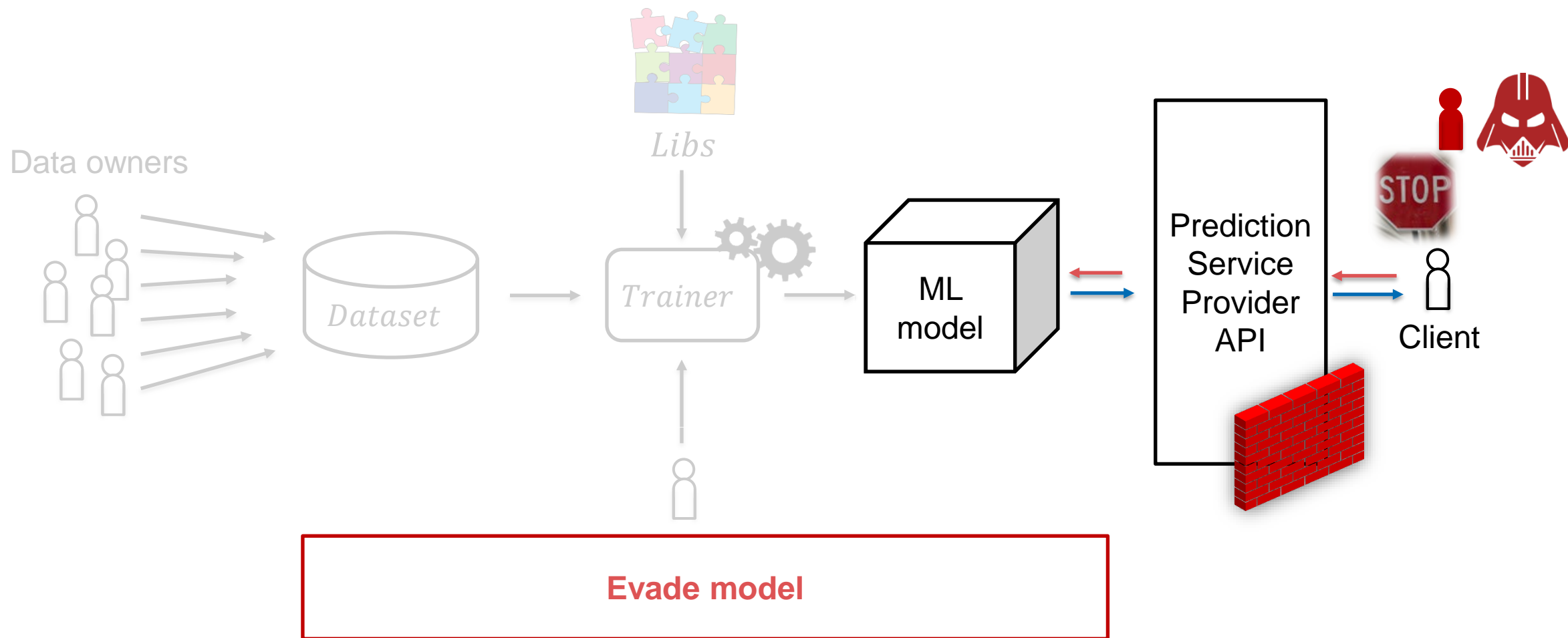
# Machine Learning pipeline



Where is the adversary? What is its target?



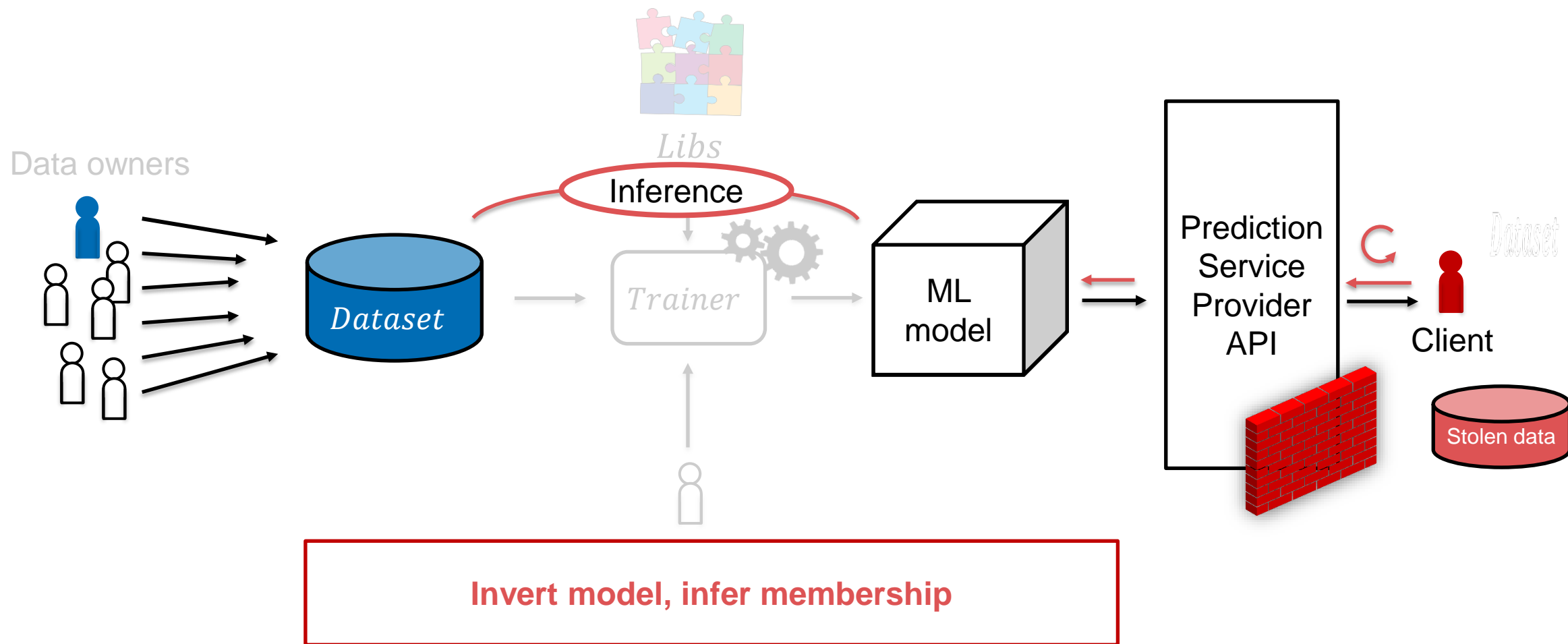
# Compromised input – Model integrity



Szegedy et al. - *Intriguing Properties of Neural Networks* ICLR '14 (<https://arxiv.org/abs/1312.6199v4>)

Dalvi et al. - *Adversarial Classification* KDD '04 (<https://dl.acm.org/doi/10.1145/1014052.1014066>)

# Malicious client – Training data privacy

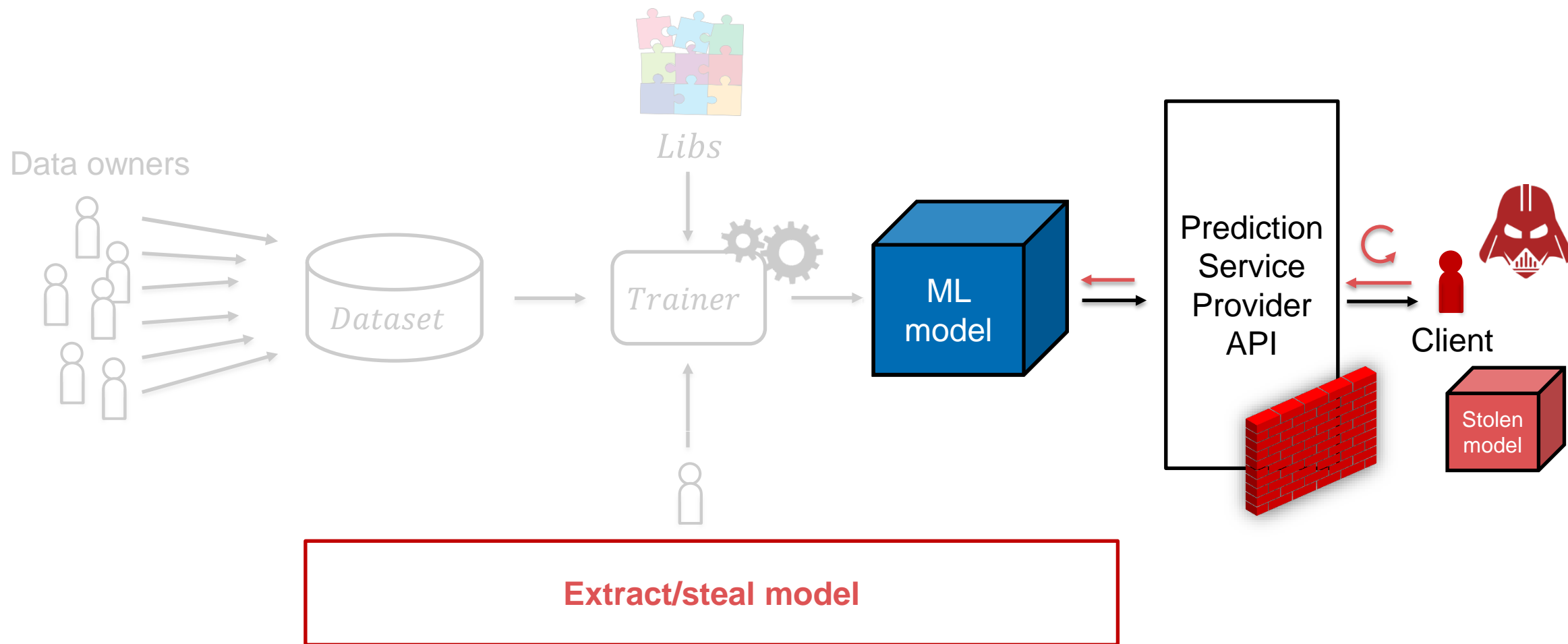


Shokri et al. - *Membership Inference Attacks Against Machine Learning Models*, IEEE S&P '16. (<https://arxiv.org/pdf/1610.05820.pdf>)

Fredrikson et al. - *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, ACM CCS'15.

<https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf>

# Malicious client – Model confidentiality



Juuti et al. - *PRADA: Protecting against DNN Model Stealing Attacks*, Euro S&P '19 (<https://arxiv.org/abs/1805.02628>)

Orekondy et al. - *Knockoff Nets: Stealing Functionality of Black-Box Models*, CVPR '19 (<https://arxiv.org/abs/1812.02766>)

Tramer et al. - *Stealing ML models via prediction APIs*, Usenix SEC '16 (<https://arxiv.org/abs/1609.02943>)

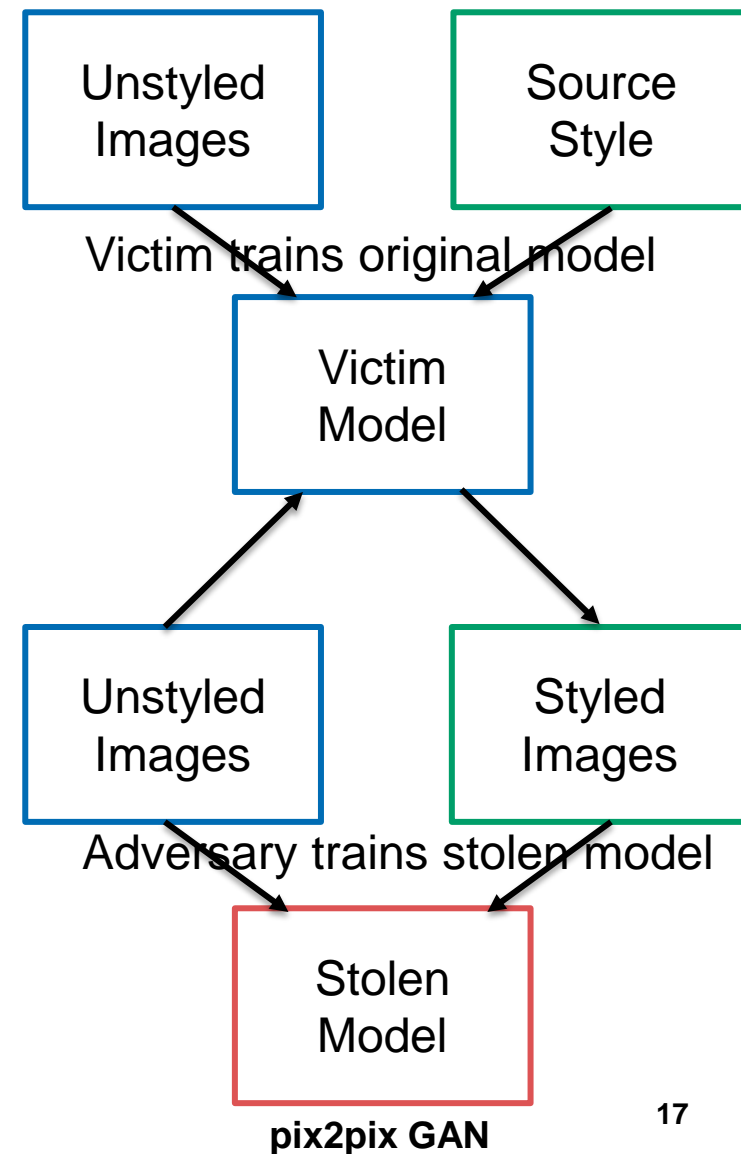
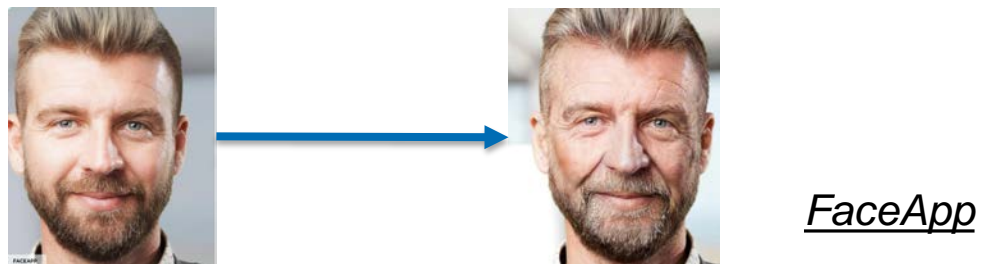
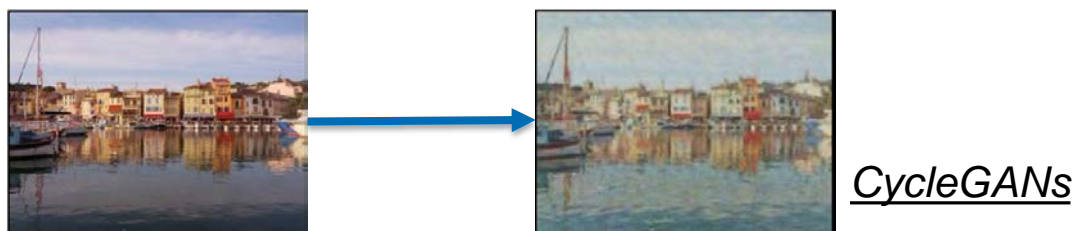
# Extracting Style-transfer models

**GANs** are effective for **changing image style**

- coloring, face filters, style application

**Core feature in generative art and in social media apps**

- Selfie2Anime, FaceApp



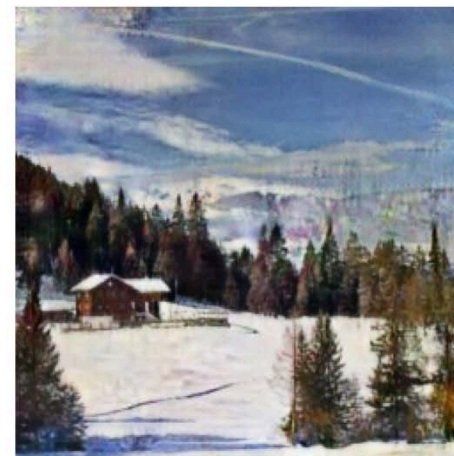
# Style transfer extraction: examples

Original  
(unstyled)

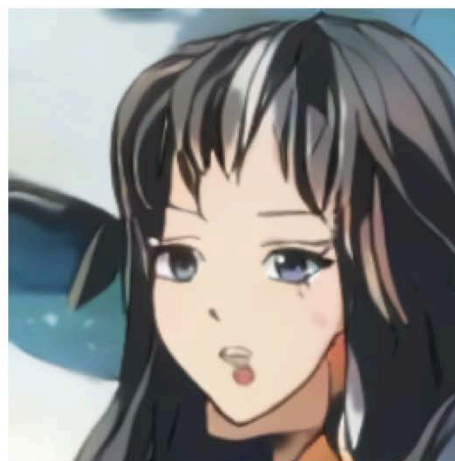
Styled  
(victim)

Styled  
(ours)

**Task 1**  
*Monet painting*

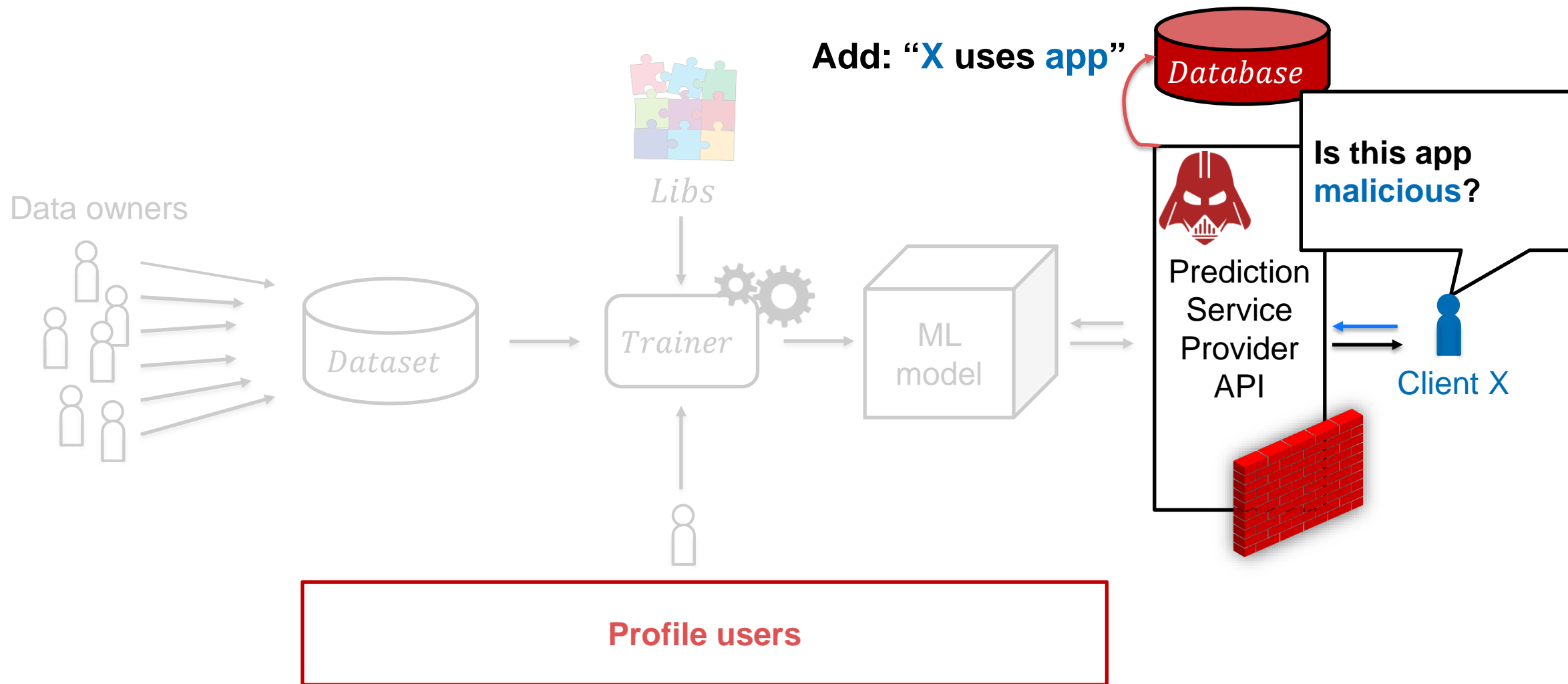


**Task 2**  
*Anime face*





# Malicious prediction service – User profiles



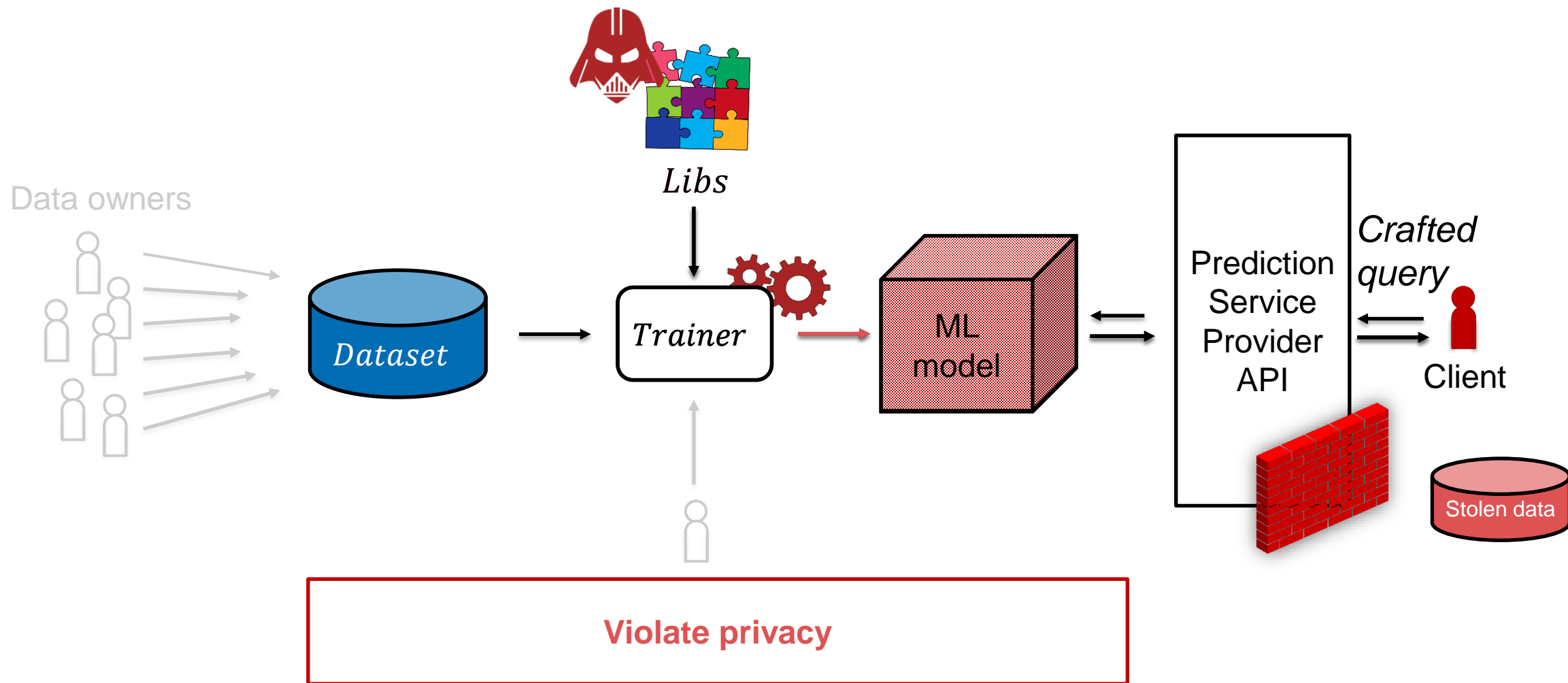
Malmi and Weber - *You are what apps you use Demographic prediction based on user's apps*, ICWSM '16 (<https://arxiv.org/abs/1603.00059>)

Liu et al. - *Oblivious Neural Network Predictions via MiniONN Transformations*, ACM CCS '17 (<https://ssg.aalto.fi/research/projects/mlsec/ppml/>)

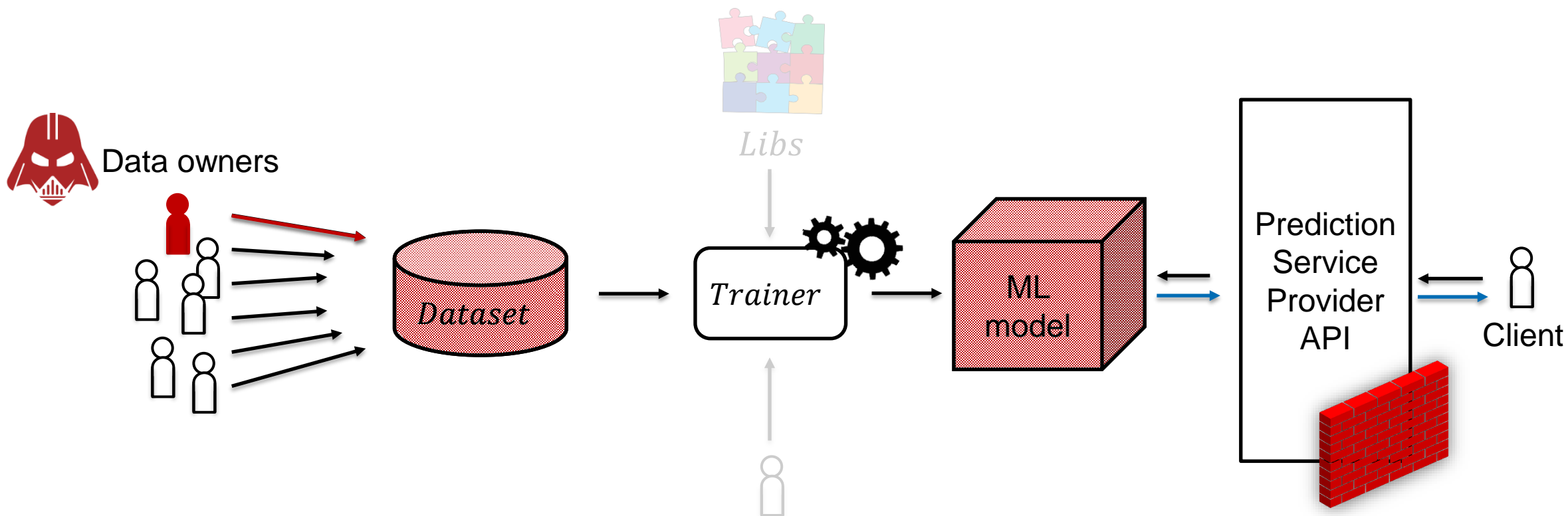
Dowlin et al. - *CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy*, ICML '16

(<https://dl.acm.org/doi/10.5555/3045390.3045413>)

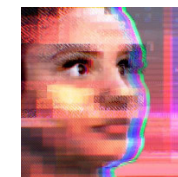
# Compromised toolchain – Training data privacy



# Malicious data owner – Model integrity



**Influence ML model (model poisoning)**



# Is malicious adversarial behaviour the only concern?

BBC Sign in Home News Sport Reel Worklife Tra

## NEWS

Home US Election Coronavirus Video World UK Business Tech Science Stories Entertainment &

Tech

### Twitter investigates racial bias in image previews

19 hours ago



One user found that Twitter seemed to favour showing Mitch McConnell's face over Barack Obama's

[https://www.bbc.com/news/technology-54234822?fbclid=IwAR1T41\\_HR6lluMKGRJbJdDrdpKdyAi5mhQSdzs0QLDso41T-SR3wJfs](https://www.bbc.com/news/technology-54234822?fbclid=IwAR1T41_HR6lluMKGRJbJdDrdpKdyAi5mhQSdzs0QLDso41T-SR3wJfs)

MIT Technology Review Topics

## Artificial intelligence

# Predictive policing algorithms are racist. They need to be dismantled.

Lack of transparency and biased training data mean these tools are not fit for purpose. If we can't fix them, we should ditch them.

by **Will Douglas Heaven** July 17, 2020

<https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-machine-learning-bias-criminal-justice/>

Tech policy / AI Ethics

## AI is sending people to jail — and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao**

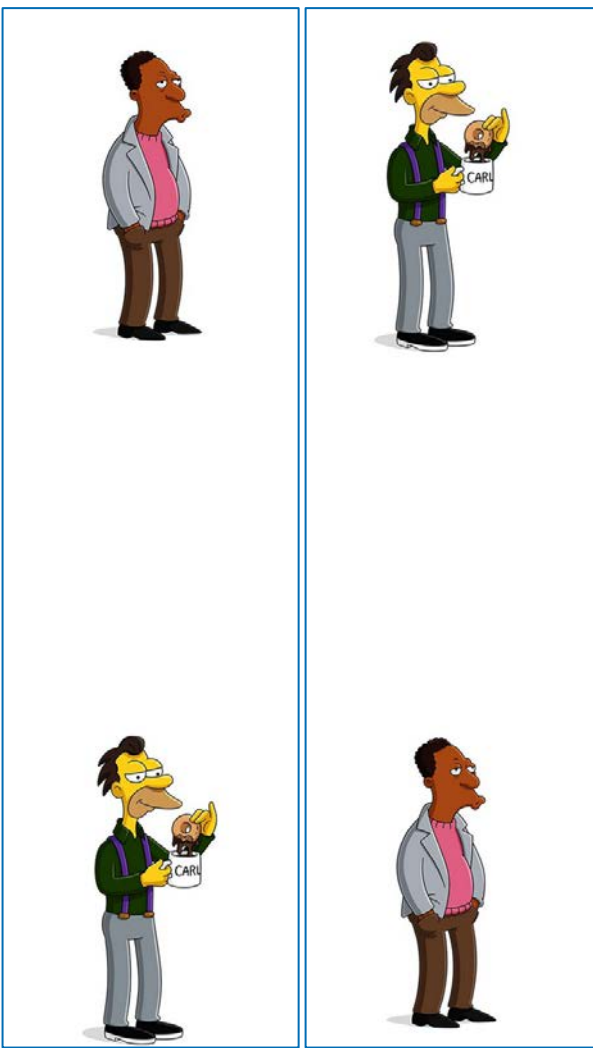
January 21, 2019

<https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/>

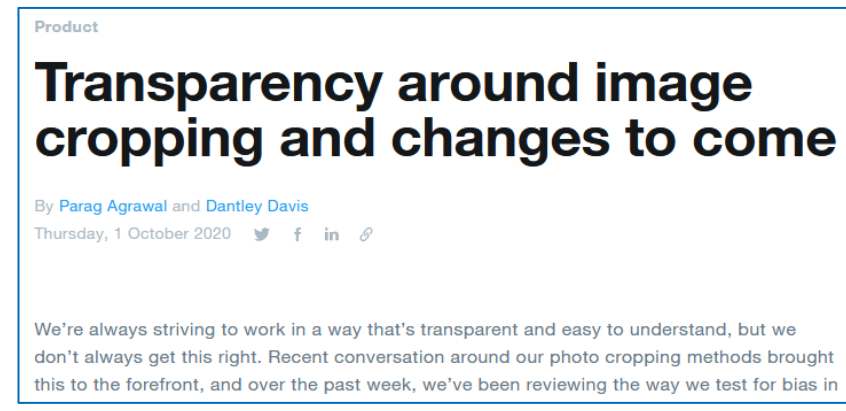
# Measures of accuracy are flawed, too



[https://twitter.com/\\_jsimonovski/status/1307542747197239296](https://twitter.com/_jsimonovski/status/1307542747197239296)



<https://twitter.com/TwitterComms/status/1307739940424359936>



[https://blog.twitter.com/official/en\\_us/topics/product/2020/transparency-image-cropping.html](https://blog.twitter.com/official/en_us/topics/product/2020/transparency-image-cropping.html)

# Challenges in making AI trustworthy

Security concerns

Privacy concerns

Ethical and legal concerns



**Trustworthy AI: Meet these criteria even in the presence of  
“adversarial” behaviour**



More on our at <https://crisp.uwaterloo.ca/research/SSG/>

# Where can I learn more about AI?

**Elements of AI**

<https://www.elementsofai.com/>

**Learn with Google AI**

<https://ai.google/education/>

**Top AI courses in Coursera**

<https://www.coursera.org/courses?query=artificial%20intelligence>

# Where can I learn more about cybersecurity/privacy?

**Cybersecurity Base 2022**

<https://cybersecuritybase.mooc.fi/>

**SANS CyberAces**

<https://www.cyberaces.org/>

**Top cybersecurity courses in Coursera**

<https://www.coursera.org/courses?query=cybersecurity>

**Expected soon: DYLOTT bootcamp in cybersecurity**

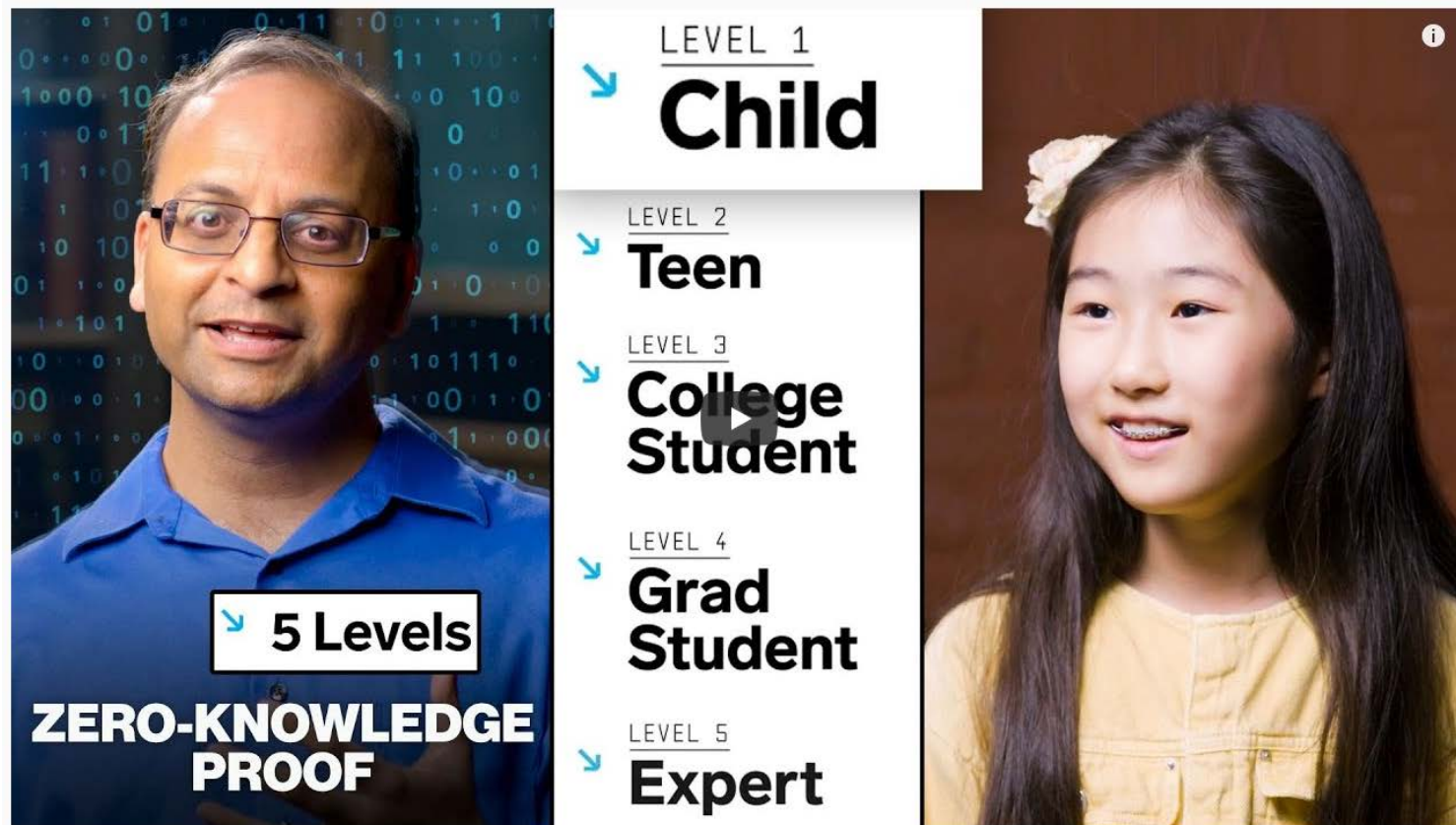
<https://www.dylott.com/>



# Where can I learn more about cybersecurity/privacy?

5 Levels YouTube Channel: S1 E16: Computer Scientist Explains One Concept in 5 levels of Difficulty | WIRED

<https://youtu.be/fOGdb1CTu5c>



# CPI Talks at Waterloo

Public outreach lectures intended for the **general public** by **leading experts** on important cybersecurity and privacy topics of concern

## Why

- Inform the general public
- Encourage students to consider careers in cybersecurity
  - **High school students** particularly welcome!

Next CPI Talks: Feb 28, Apr 13, ...

Videos of [previous CPI Talks](#) on the CPI YouTube channel



<https://cpi.uwaterloo.ca/events/cpi-talks-public-outreach-lecture-series>

<https://cpi.uwaterloo.ca/events/cpi-talks-public-outreach-lecture-series>