Aalto University

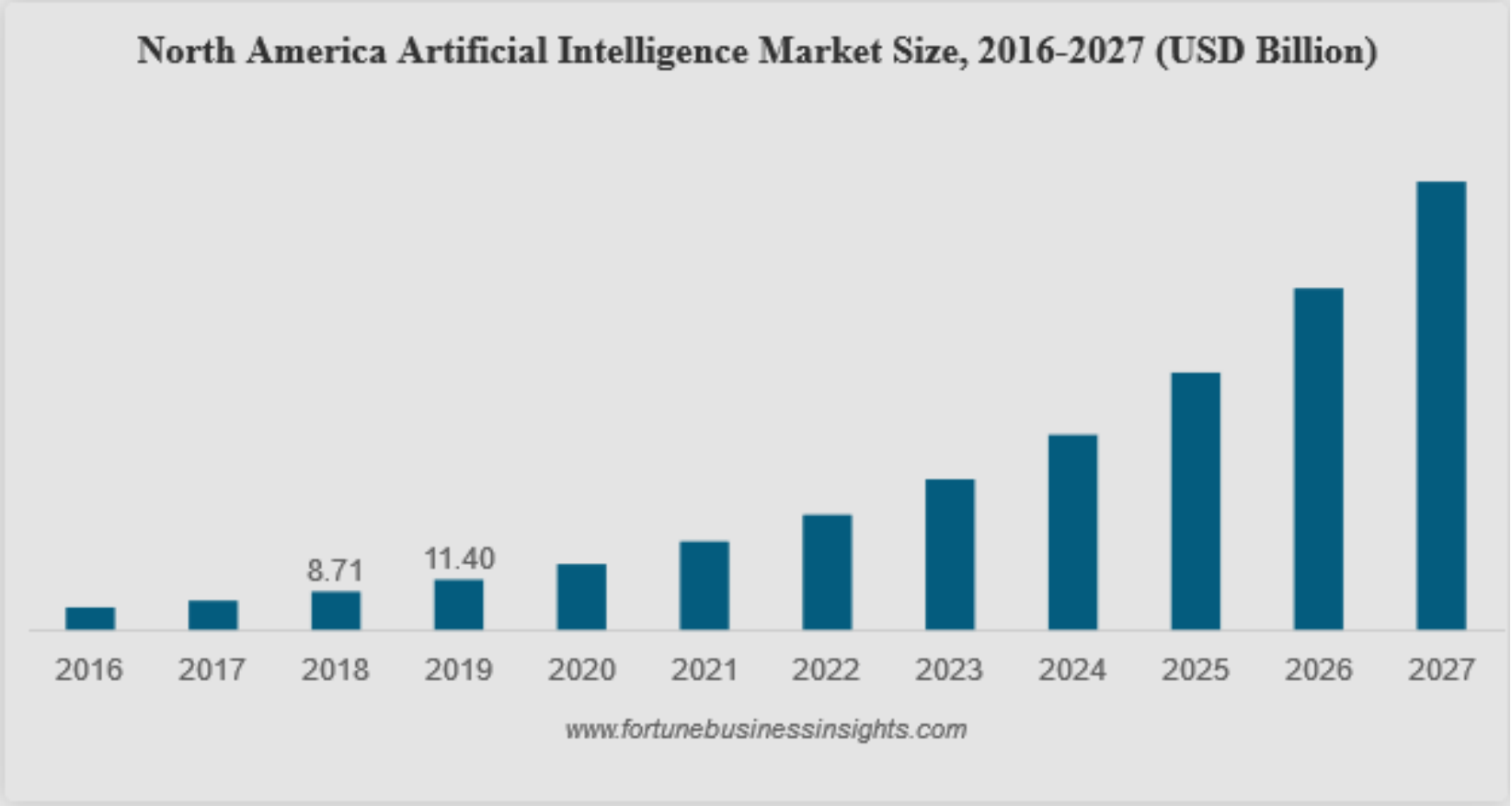
UNIVERSITY OF WATERLOO

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 Artifical 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-artifical-intelligence-is-advancing-precision-medicine/#2f720a79a4d5

MOTHERBOARD
TECH BY VICE

Dozens of Cities Have Secretly Experimented With Predictive Policing Softw

Documents obtained by Motherboa
requests verify previously unconfir
with predictive policing company P

By Caroline Haskins

https://www.vice.com/en_us/article/d3m
experimented-with-predictive-policing-s

Forbes

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

How AI Is Uprooting Recruiting

Falon Fatemi Contributor ⓘ

Entrepreneurs

https://www.vice.com/en_us/article/d3m7jq/dozens-of-cities-have-secretly-experimented-with-predictive-policing-software

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/

How do we evaluate AI-based systems?

Effectiveness

- measures of accuracy

Performance

- inference speed and memory consumption

...

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

Challenges in making AI trustworthy

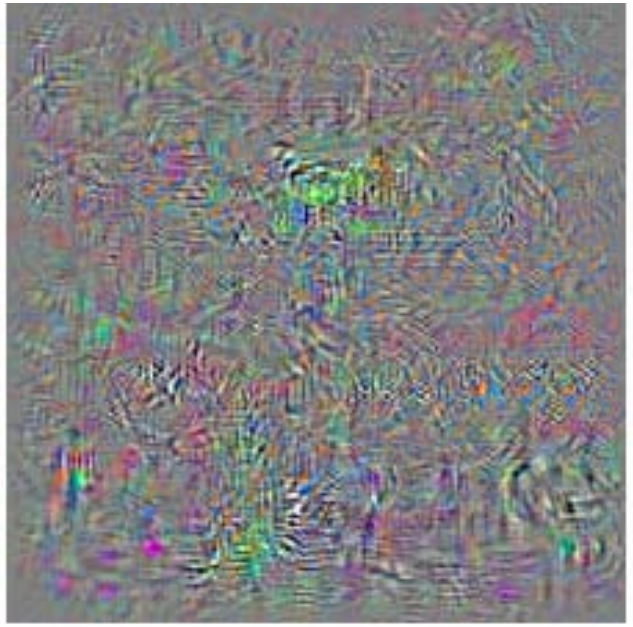
Security concerns

Privacy concerns

Evading machine learning models

Which class is this?
School bus

+ 0.1·



=

Which class is this?
Ostrich

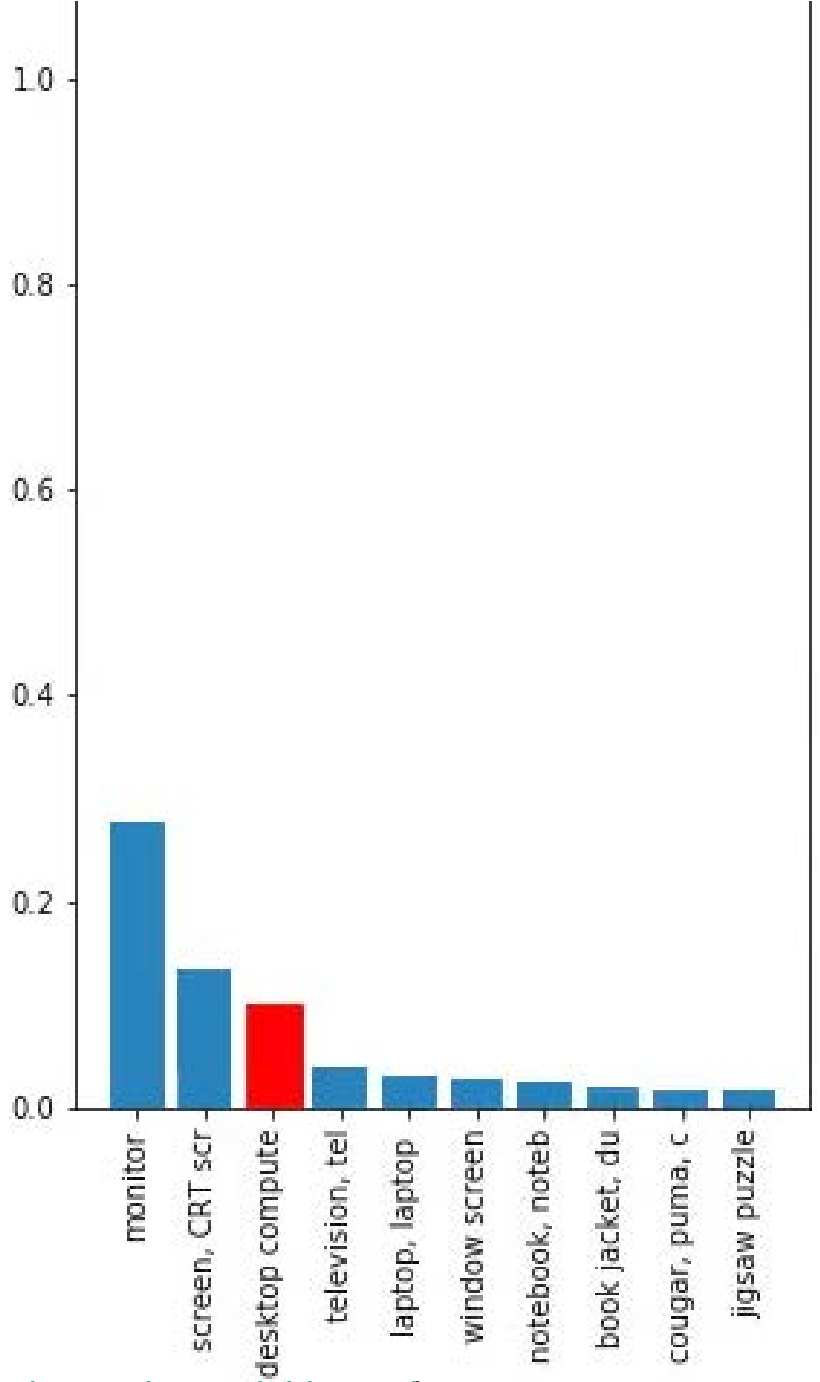
Szegedy et al. - *Intriguing Properties of Neural Networks* ICLR ‘14 (https://arxiv.org/abs/1312.6199v4)

Which class is this?

Cat

Which class is this?

Desktop computer



Athalye et al. - *Synthesizing Robust Adversarial Examples*. ICML '2019 (https://blog.openai.com/robust-adversarial-inputs/)

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

DolphinAttack: Inaudible Voice command

Guoming Zhang Chen Yan Xiaoyu Ji

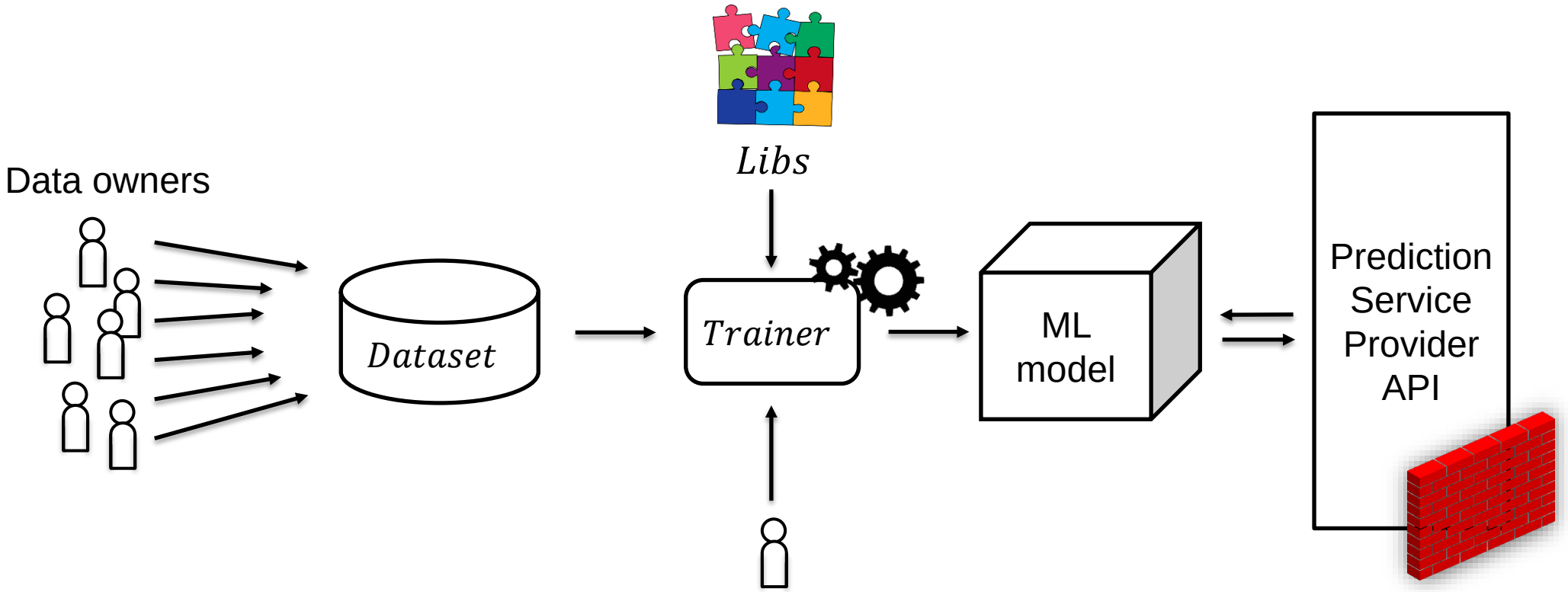
Tianchen Zhang Taimin Zhang Wenyuan Xu

Zhejiang University

ACM CCS 2017

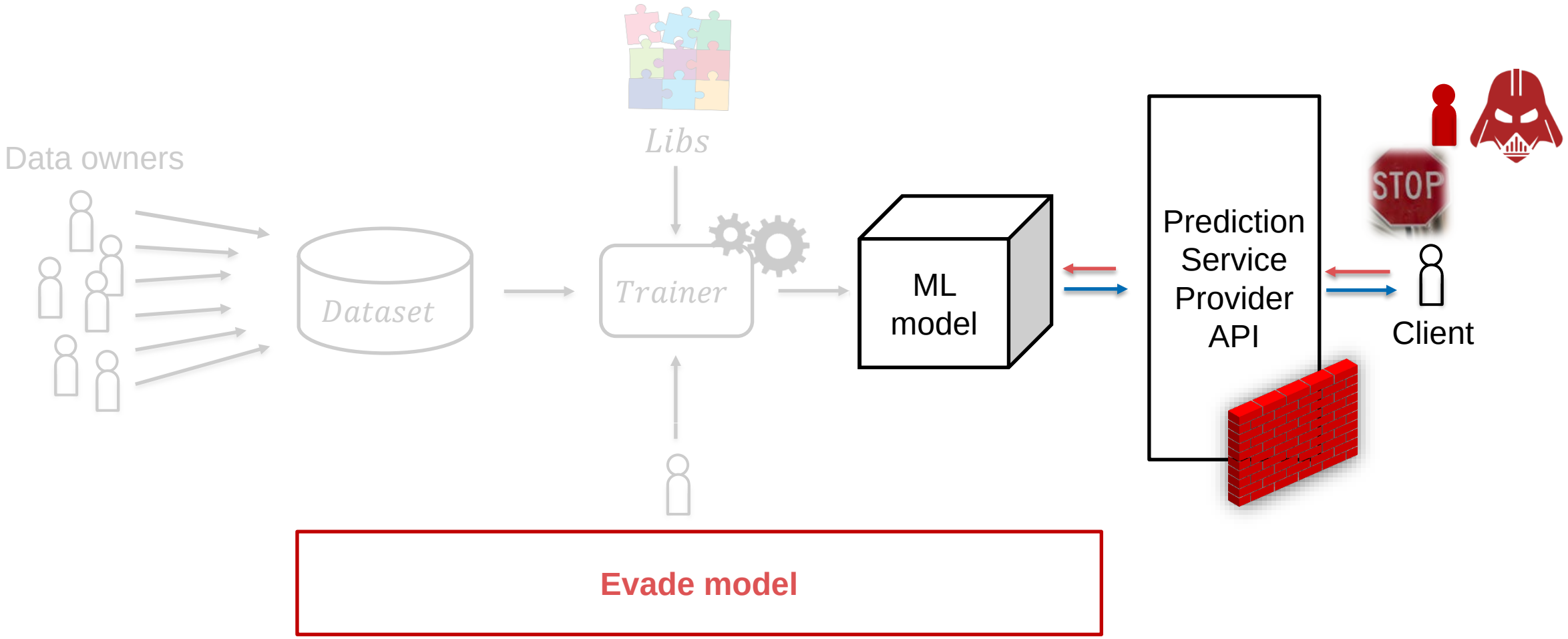
Zhang et al. - *DolphinAttack: Inaudible Voice Commands*, ACM CCS '17 (https://arxiv.org/abs/1708.09537)

Machine Learning pipeline



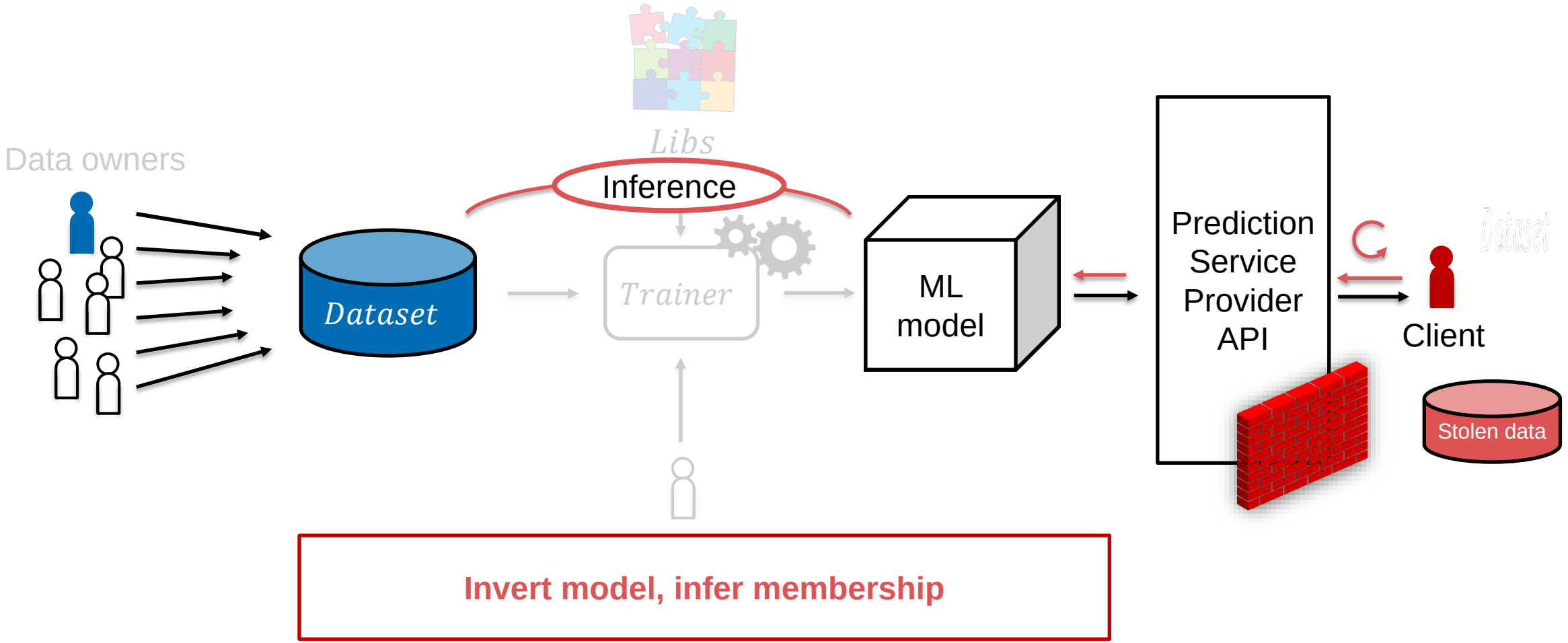
Where is the adversary? What is its target?

Compromised input – Model integrity



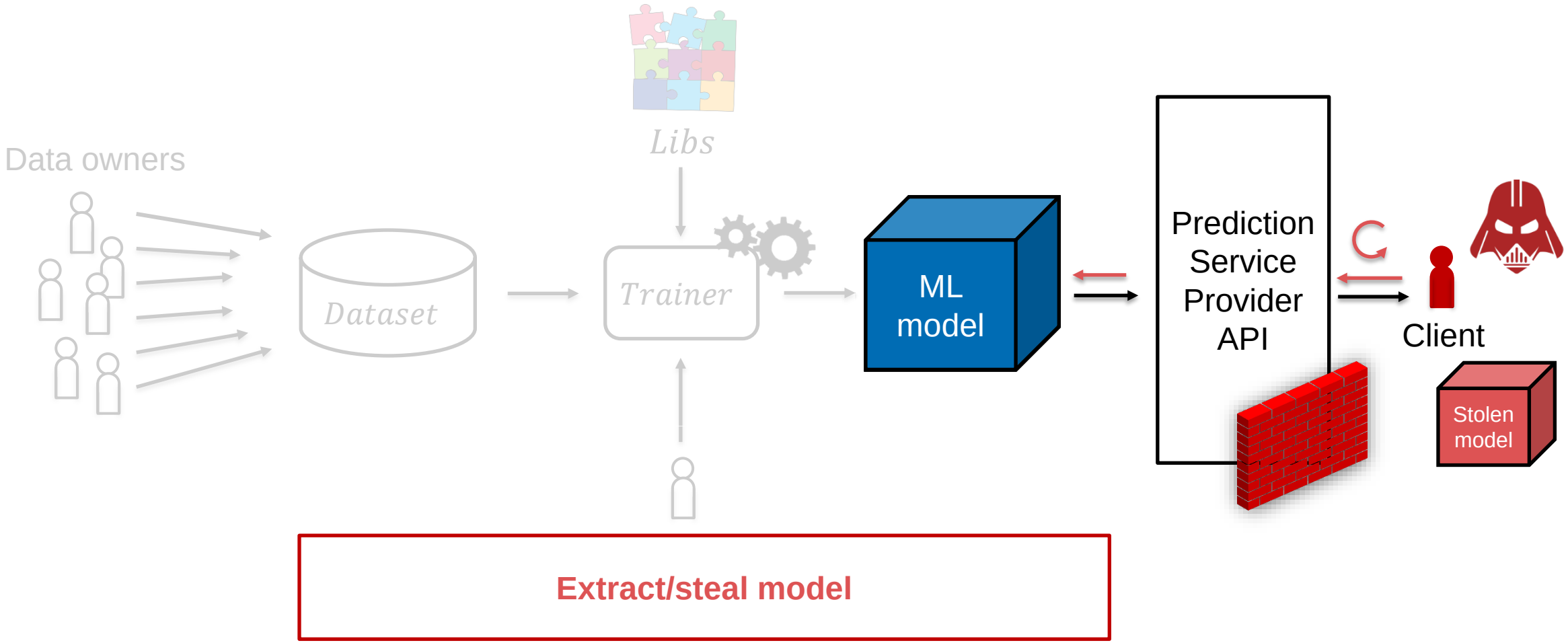
Szegedy et al. - *Intriguing Properties of Neural Networks* ICLR '14 (https://arxiv.org/abs/1312.6199v4)
Dalvi et al. - *Adversarial Classification KDD '04* (https://dl.acm.org/doi/10.1145/1014052.1014066)

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 NLP Transformer models

Techniques for extracting image classifiers don't always extend to NLP models

Transfer learning from pre-trained models is now very popular

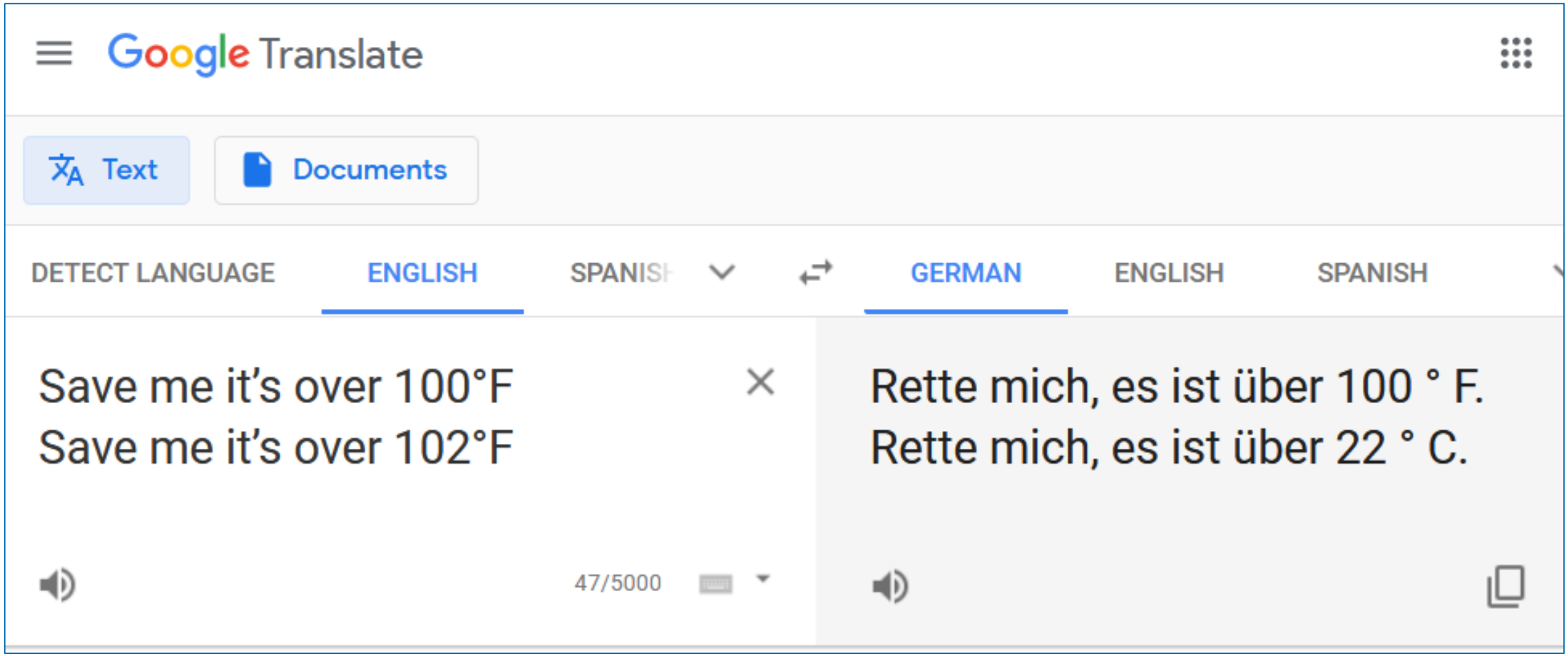
- But they make model extraction easier[1]

Krishna et al[1] show that a Knockoff-like attacks against BERT models are feasible

- Adversary unaware of target distribution or task of victim model
- Adversary queries are merely "natural" (randomly sampled sequences of words)
- In-distribution adversary queries can improve extraction efficacy

Wallace et al[2] extract real-world MT models, find transferable adversarial examples

[1] Krishna et al. – *Thieves on Sesame Street! Model Extraction of BERT-based APIs* . ICLR '20 (https://iclr.cc/virtual_2020/poster_Byl5NREFDr.html)
[2] Wallace et al. – *Imitation Attacks and Defenses for Black-box Machine Translation Systems.* EMNLP '20 (https://arxiv.org/abs/2004.15015)



https://translate.google.com/#view=home&op=translate&sl=en&tl=de&text=Save%20me%20it%E2%80%99s%20over%20100%C2%B0F%0ASave%20me%20it%E2%80%99s%20over%20102%C2%B0F

Wallace et al. – *Imitation Attacks and Defenses for Black-box Machine Translation Systems.* EMNLP '20 (https://arxiv.org/abs/2004.15015)

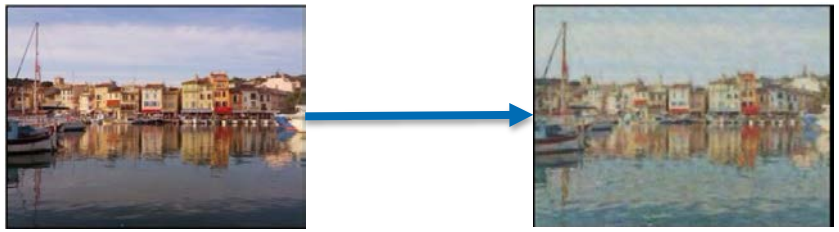
Extracting Style-transfer models

GANS are effective for changing image style

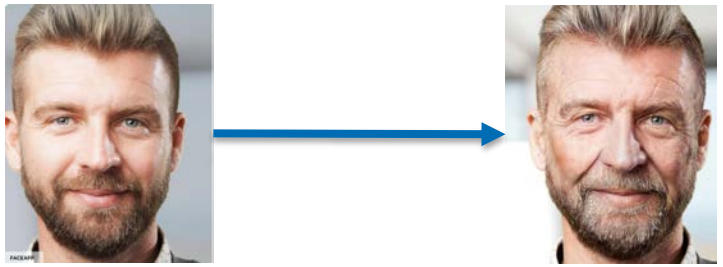
- coloring, face filters, style application

Core feature in generative art and in social media apps

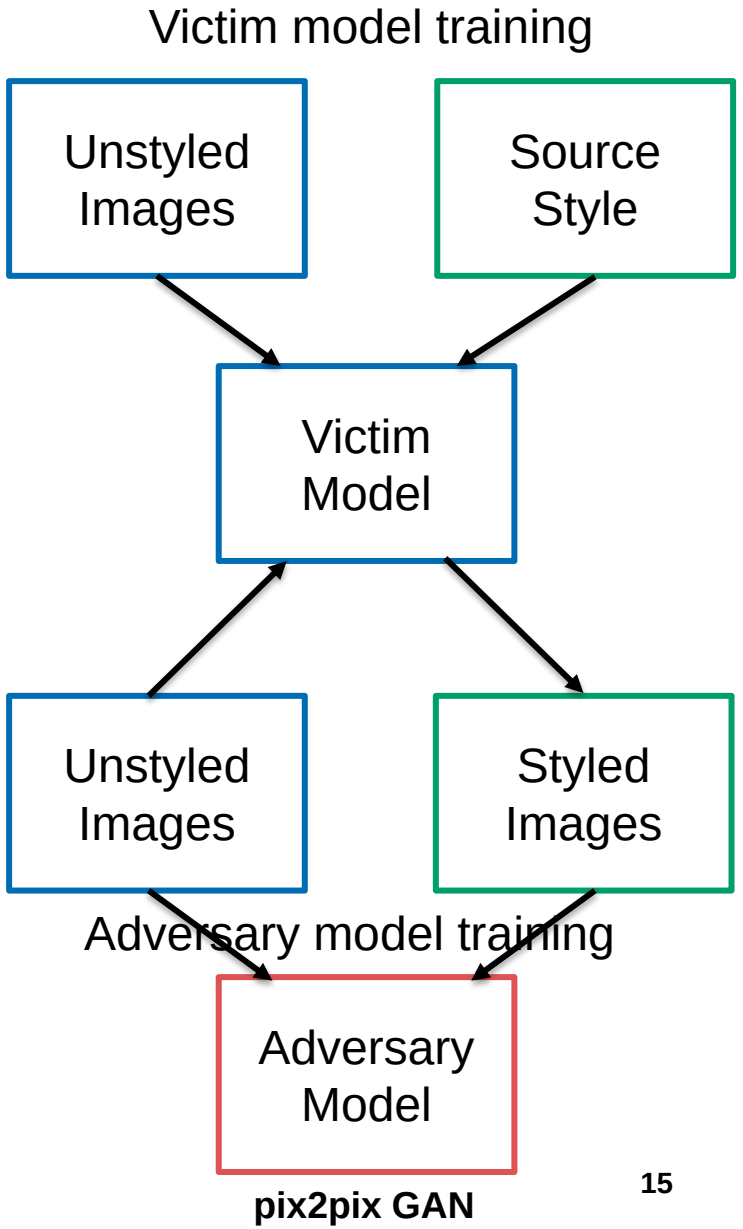
- Selfie2Anime, FaceApp



CycleGANs

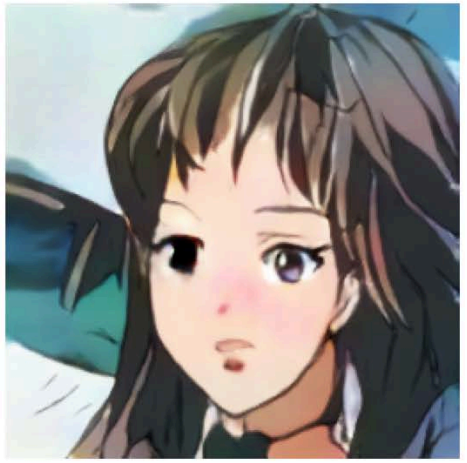


FaceApp



pix2pix GAN

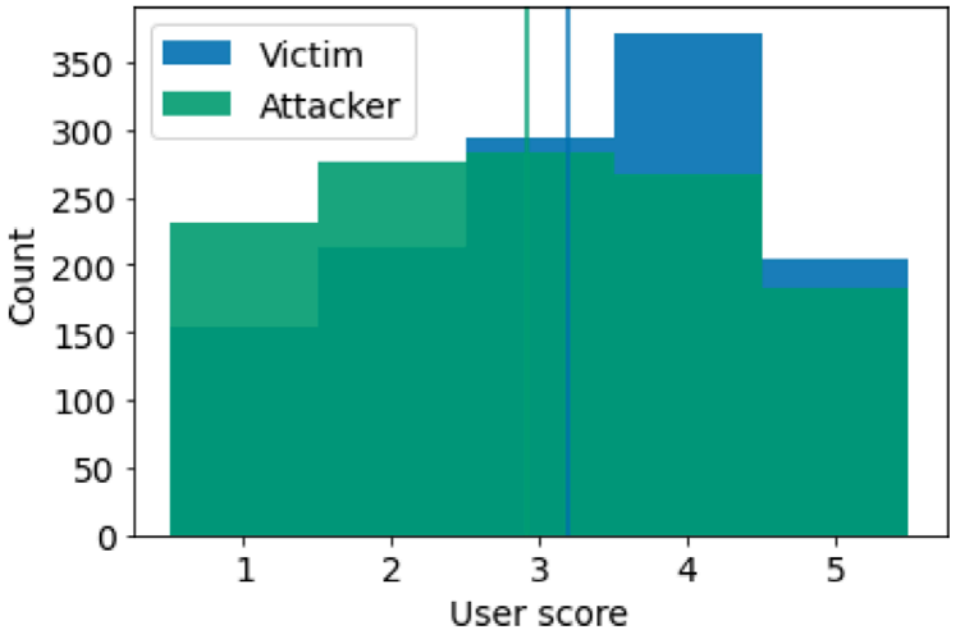
Style transfer extraction: examples

	Original (unstyled)	Styled (victim)	**Styled (ours)**
Task 1 *Monet painting*			
Task 2 *Anime face*			

Szyller et al. Good Artists Copy, Great Artists Steal: Model Extraction Attacks Against Image Translation Generative Adversarial Networks (https://arxiv.org/abs/2104.12623)

Style transfer extraction: user study

Monet-to-Photo

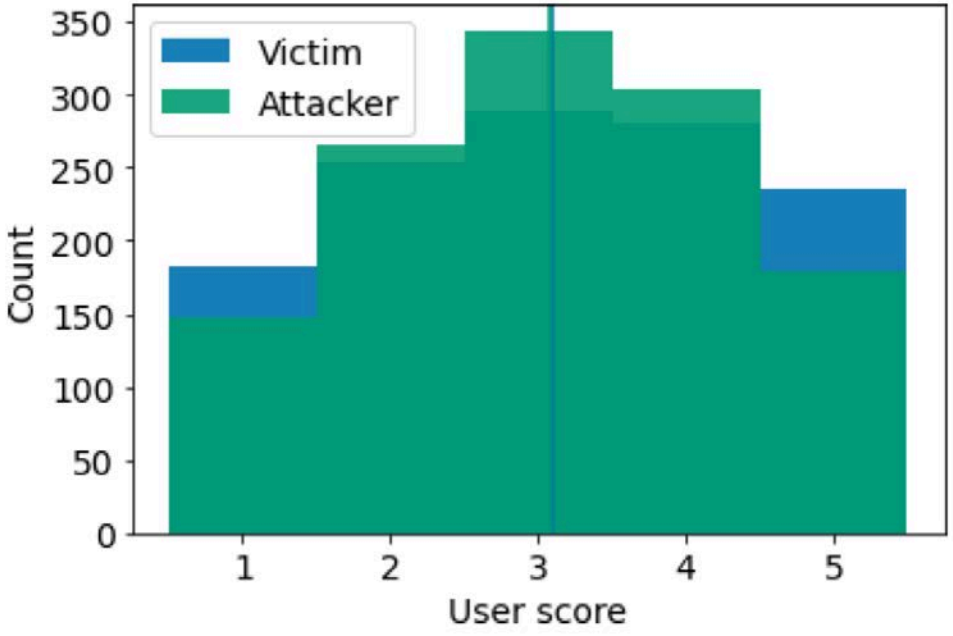


Models nearly the same according to quantitative metrics.

Hypothesis testing:

- models are not statistically equivalent
- models are not statistically different

Selfie-to-Anime



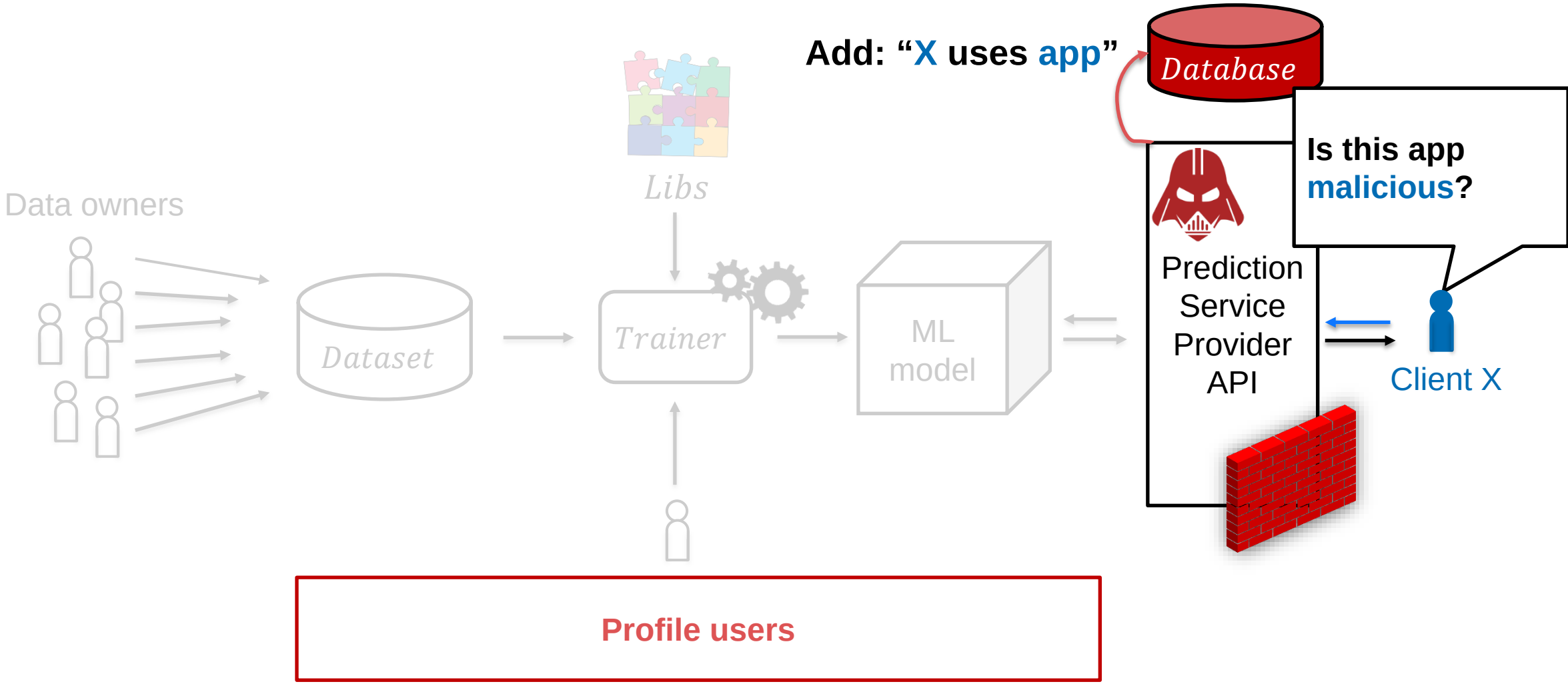
Models quite different according to quantitative metrics.

Hypothesis testing:

- models are statistically equivalent
- models are not statistically different

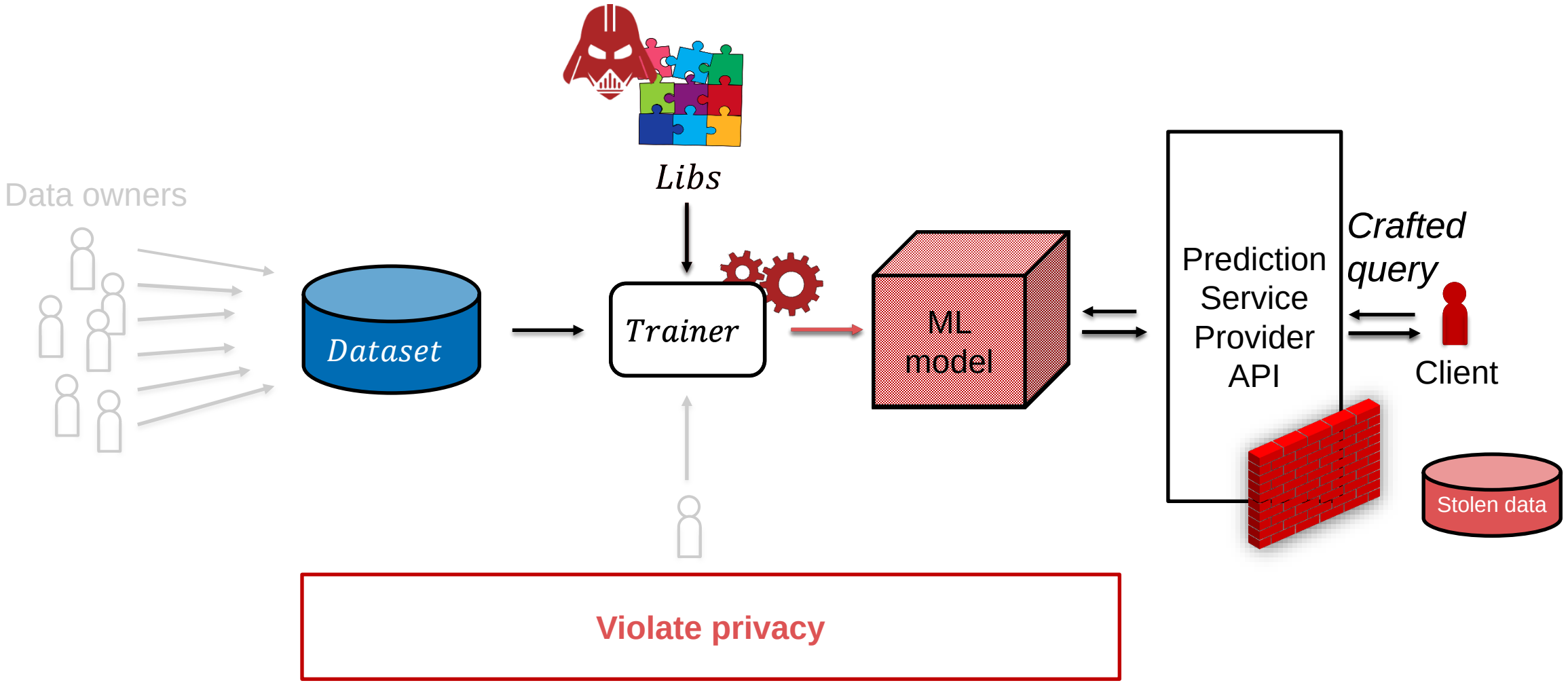
Szyller et al. Good Artists Copy, Great Artists Steal: Model Extraction Attacks Against Image Translation Generative Adversarial Networks (https://arxiv.org/abs/2104.12623)

Malicious prediction service – User profiles



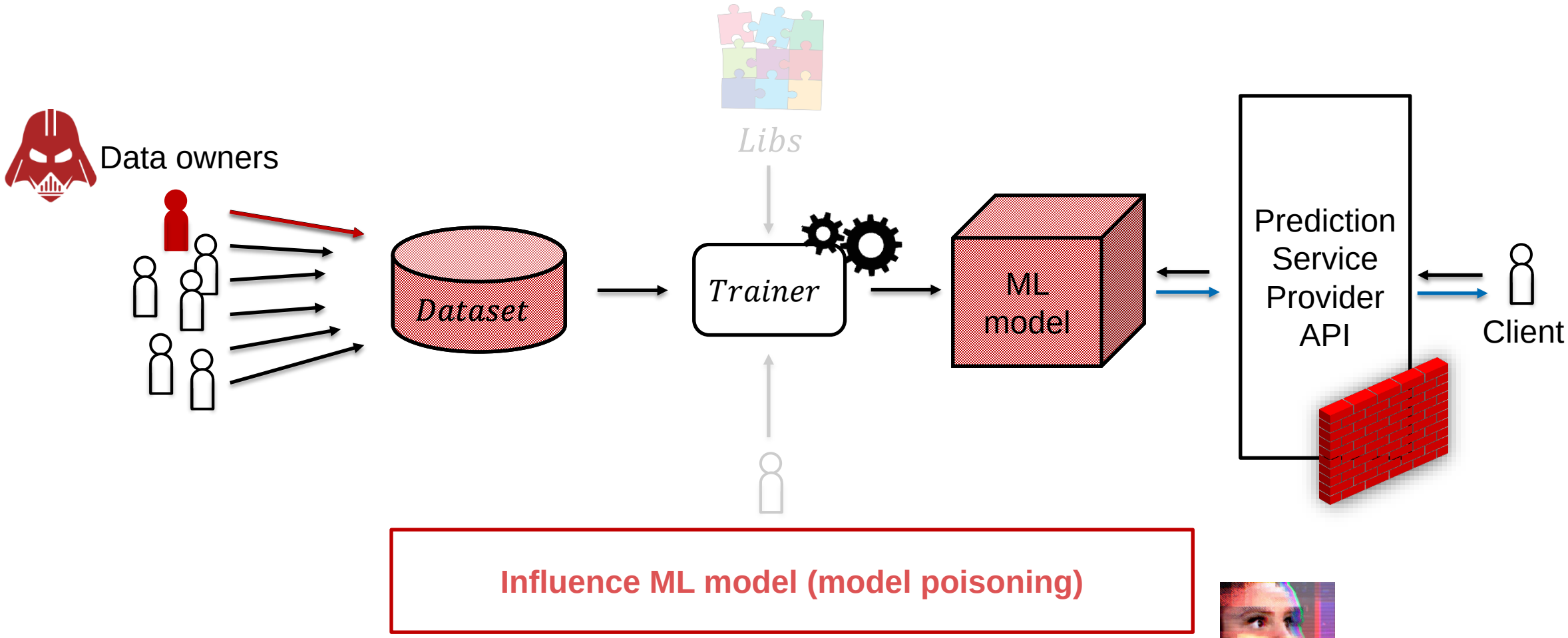
Malmi and Weber - *You are what apps you use Demographic prediction based on user's apps*, ICWSM '16 (https://arxiv.org/abs/1603.00059)
Liu et al. - *Oblivious Neural Network Predictions via MiniONN Transformations*, ACM CCS '17 (https://ssg.aalto.fi/research/projects/mlsec/ppml/)
Dowlin et al. - *CryptoNets: 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



Song et al. - *Machine Learning models that remember too much*, ACM CCS '17 (https://arxiv.org/abs/1709.07886)
Hitja et al. - *Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning*, ACM CCS '17 (http://arxiv.org/abs/1702.07464)

Malicious data owner – Model integrity



https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot
https://www.theguardian.com/technology/2017/nov/07/youtube-accused-violence-against-young-children-kids-content-google-pre-school-abuse

Is malicious adversarial behaviour the only concern?

BBC Sign in Home News Sport Reel Worklife

NEWS

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

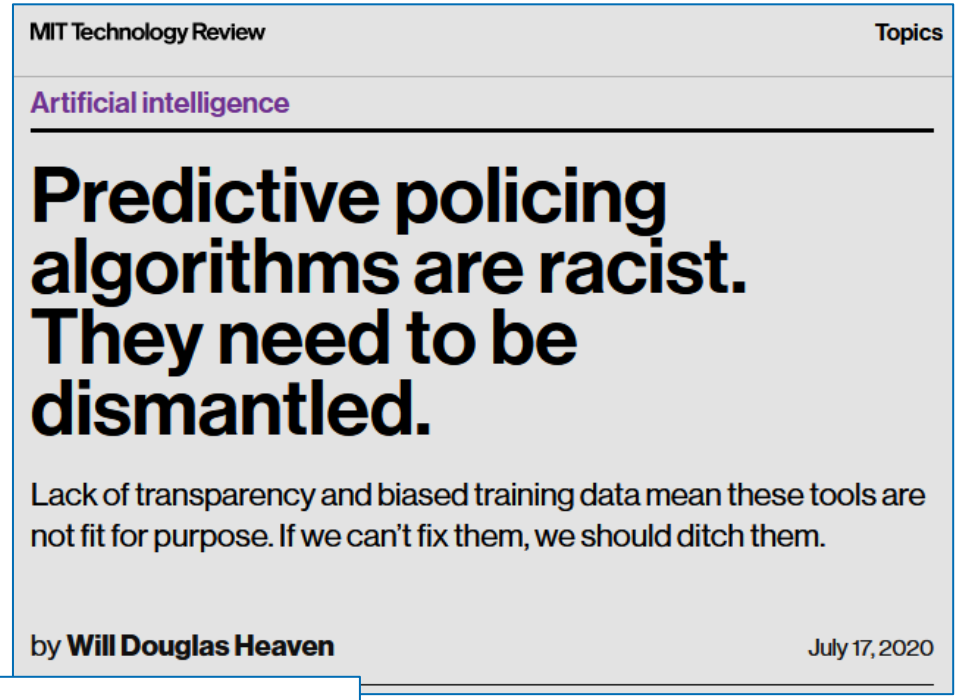
Tech

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_HR6lIuMKGRJbJdDrdpKdyAi5mhQSdzs0QLDso41T-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

.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

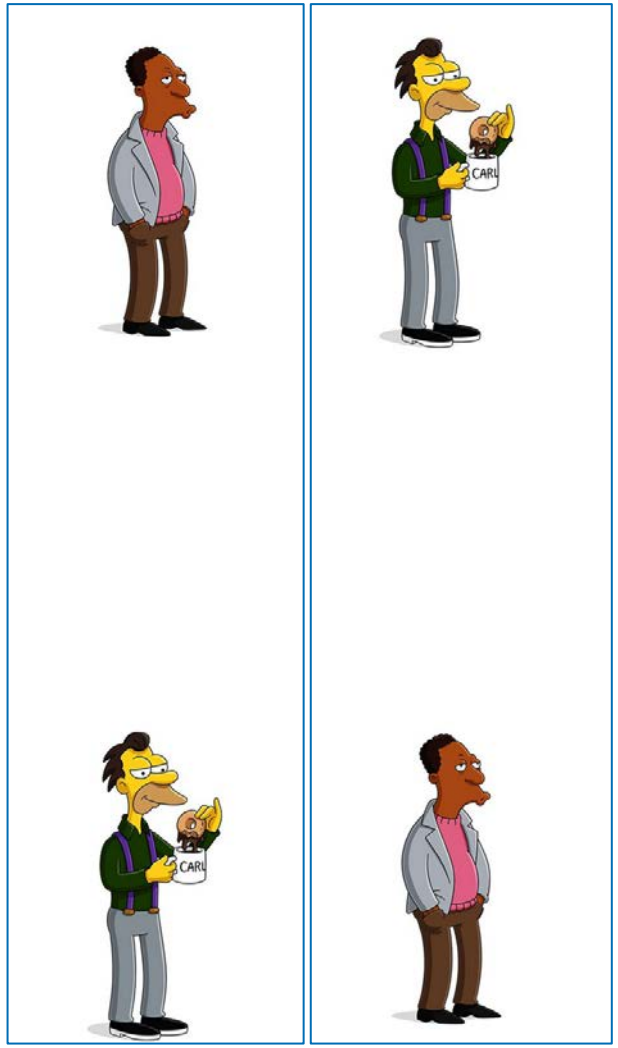
Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

by **Karen Hao** January 21, 2019

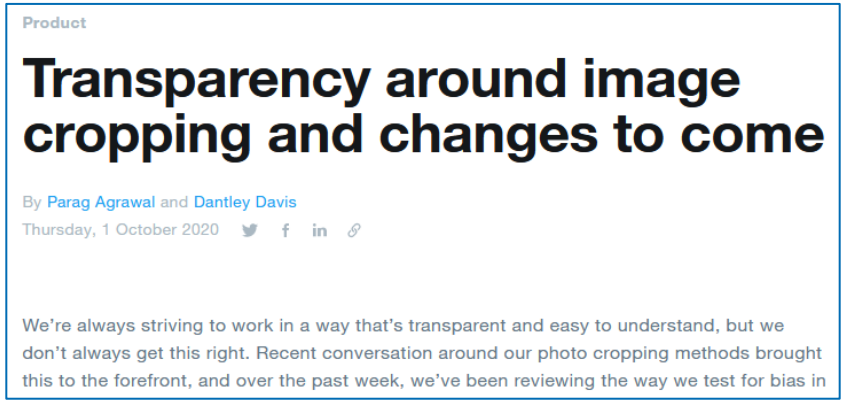
https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai /

Measures of accuracy are flawed, too

https://twitter.com/_jsimonovski/status/1307542747197239296



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



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://crysp.uwaterloo.ca/research/SSG/