Aalto University

Machine Learning
in the presence of adversaries

Andrew Paverd and Mika Juuti
(joint work with N. Asokan, Jian Liu and Samuel Marchal)

About Us

Aalto Department of Computer Science http://cs.aalto.fi/

- Top-100 in CS world rankings
- AI, algorithms, security & privacy

Secure Systems Group (Prof. Asokan) http://cs.aalto.fi/secure_systems

- 10 senior researchers, 5-10 MSc students

Here today:

- Andrew Paverd: PhD from Oxford 2016 https://ajpaverd.org/
- Mika Juuti: M.Sc. (Tech) from Aalto 2015 https://research.aalto.fi/portal/mika.juuti.html

Machine Learning is ubiquitous

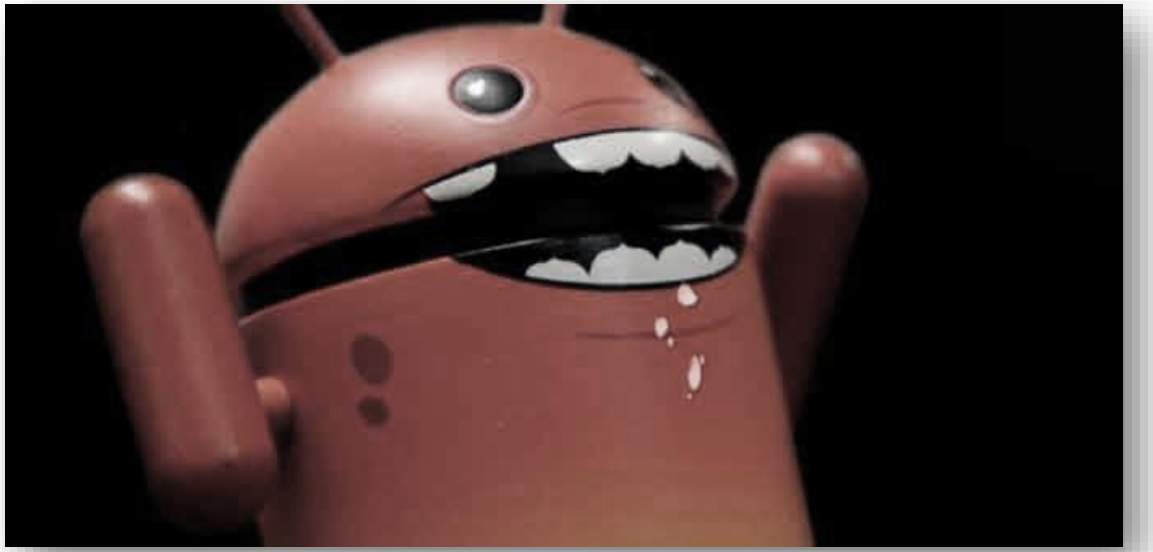
The ML market size is expected to grow by 44% annually over next five years

In 2016, companies invested up to $9 Billion in AI-based startups

[1] http://www.marketsandmarkets.com/PressReleases/machine-learning.asp
[2] McKinsey Global Institute, ”Artificial Intelligence: The Next Digital Frontier?”

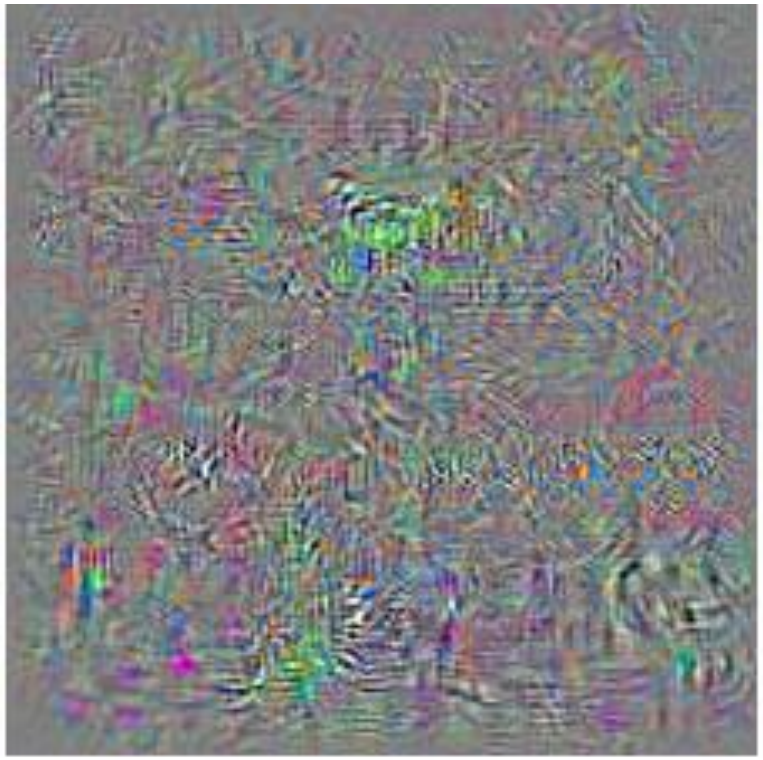
Machine Learning *for* security/privacy

Access Control



Malware / Intrusion Detection

Security & privacy *of* machine learning

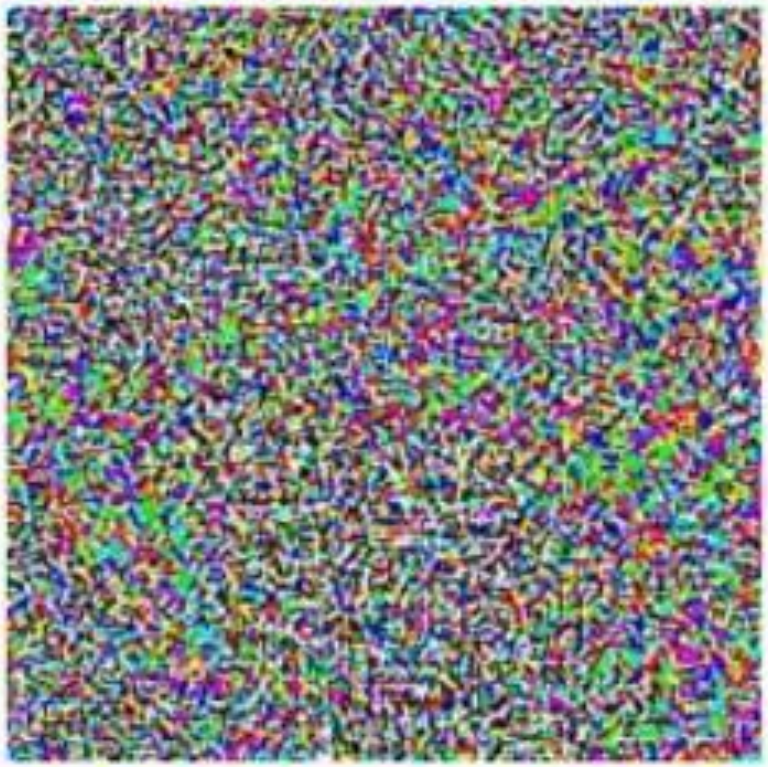


Which class is this?
School bus

Which class is this?
Ostrich

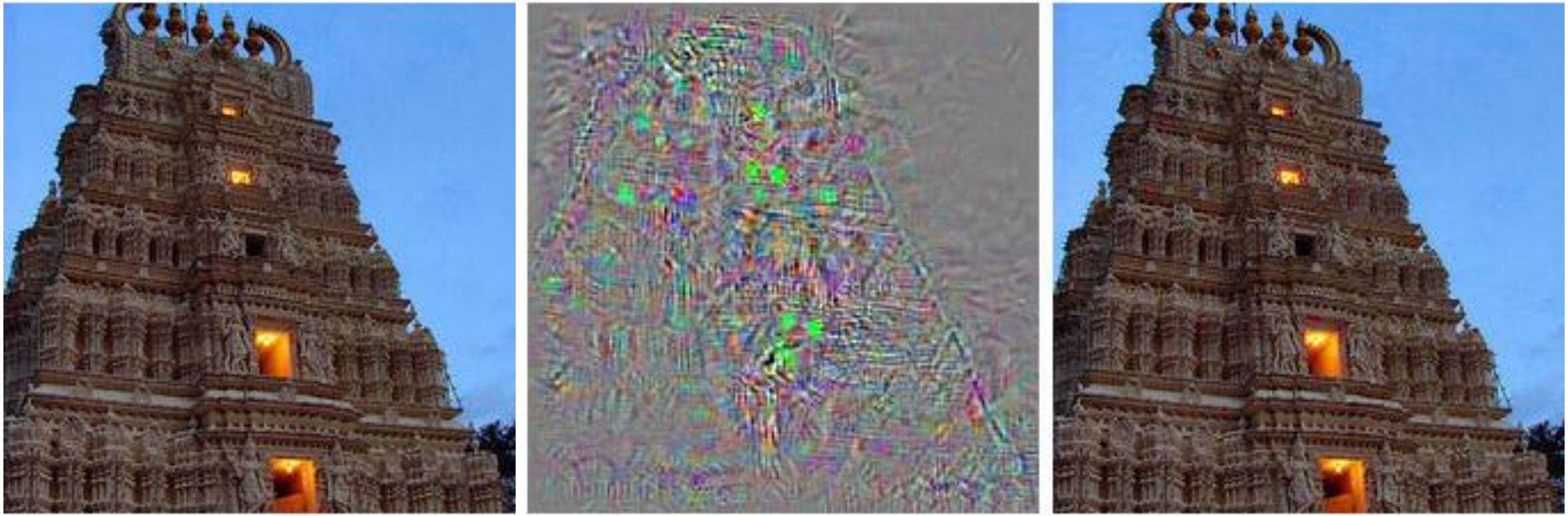
Szegedy et al. “Intriguing Properties of Neural Networks” 2014
https://arxiv.org/abs/1312.6199v4

Which class is this?
Panda



Which class is this?
Gibbon

Goodfellow et al. “Explaining and Harnessing Adversarial Examples” 2015
https://blog.openai.com/robust-adversarial-inputs/



Which class is this?
Building

Which class is this?
Ostrich

Szegedy et al. “Intriguing Properties of Neural Networks” 2014
https://arxiv.org/abs/1312.6199v4

Which class is this?
Cat

Which class is this?
Desktop computer

Athalye et al. “Synthesizing Robust Adversarial Examples”
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

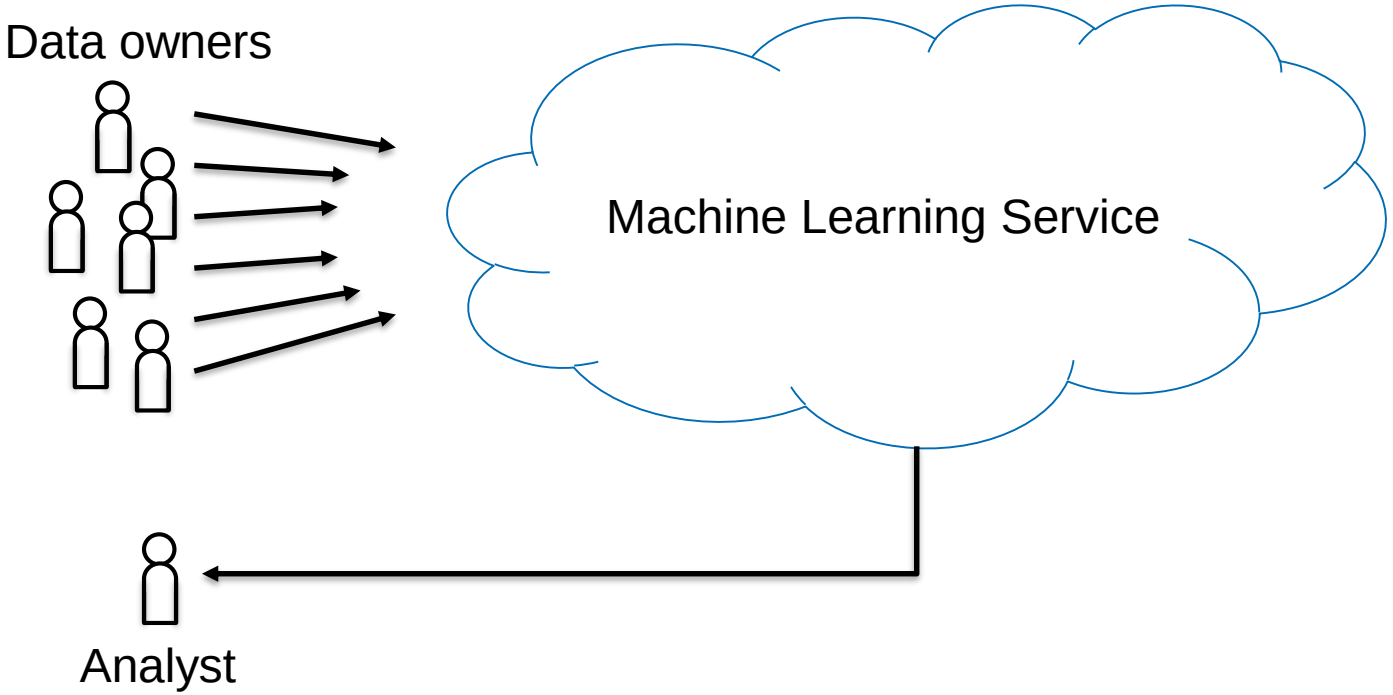
Zhejiang University

ACM CCS 2017

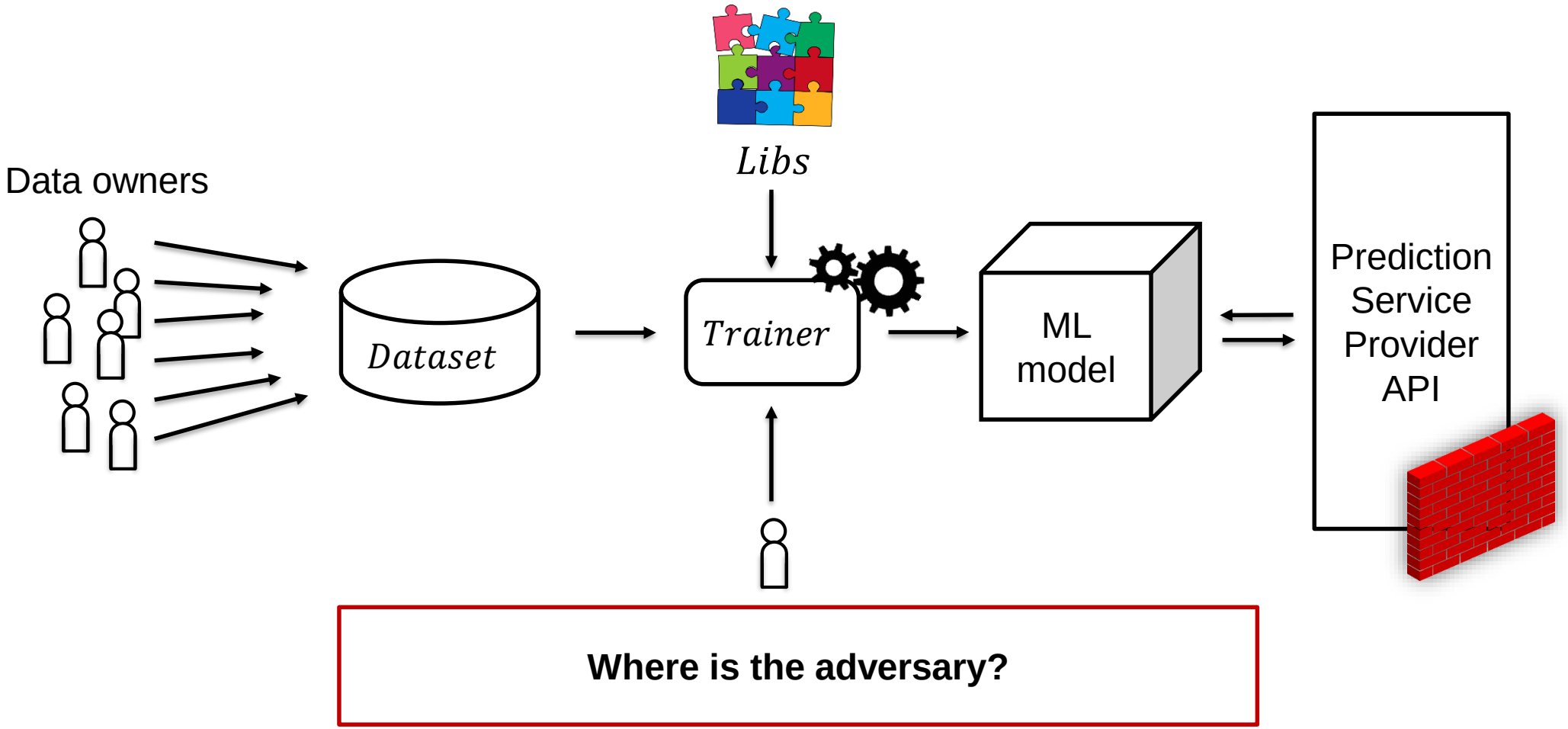
Zhang et al, “DolphinAttack: Inaudible Voice Commands”, ACM CCS 2017 https://arxiv.org/abs/1708.09537

Fredrikson et al. ”Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures”, CCS’15.
https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf

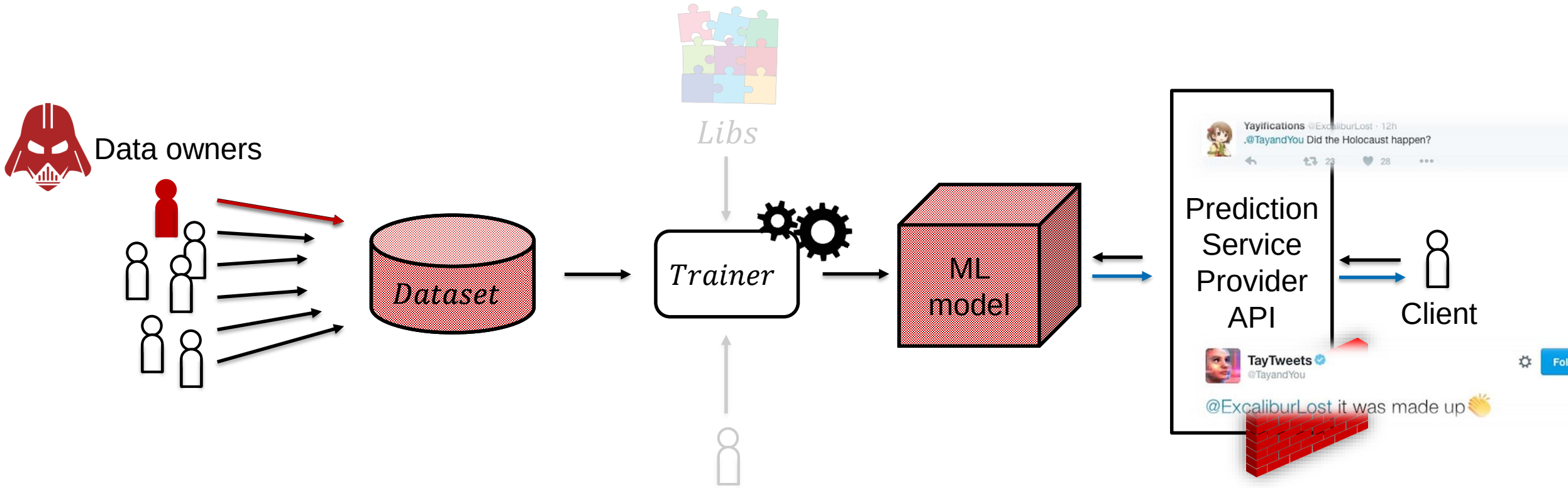
A Basic Machine Learning pipeline



A more realistic Machine Learning pipeline



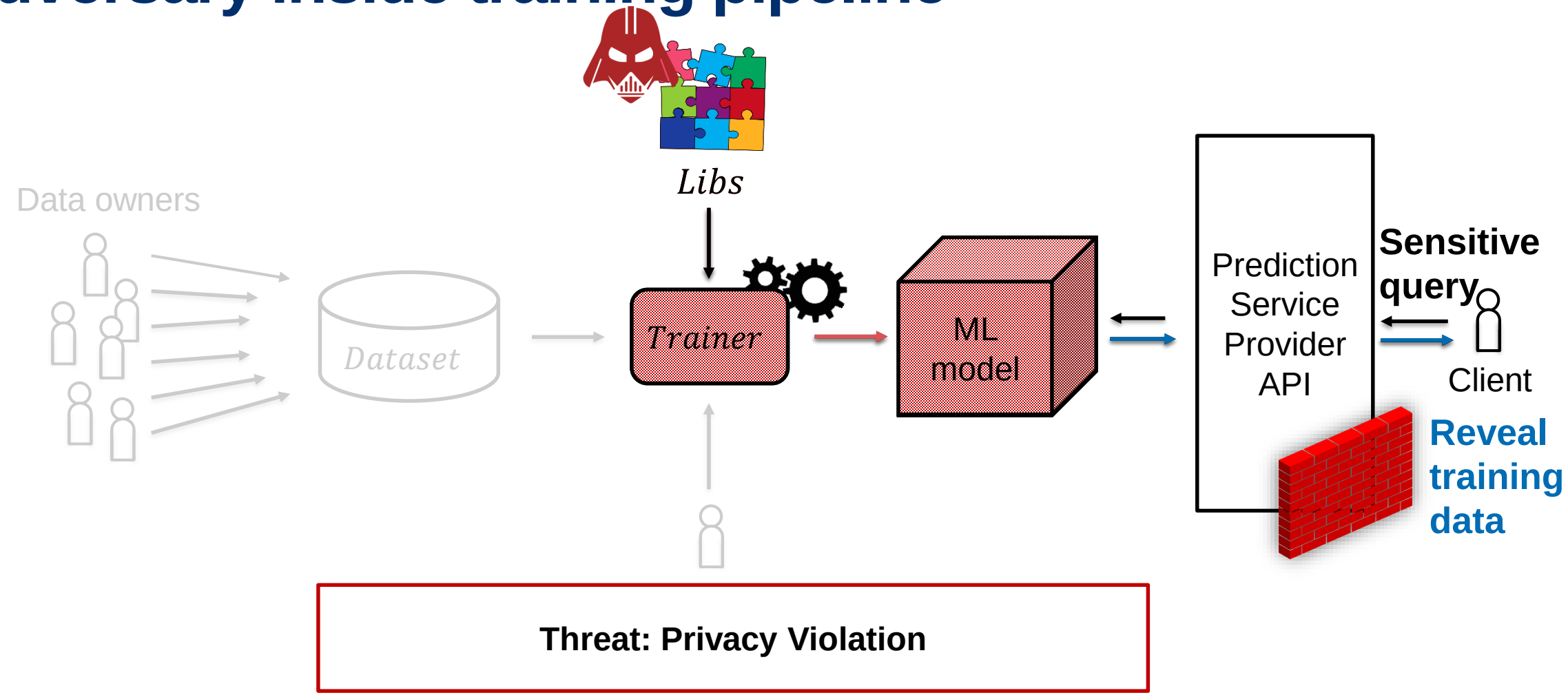
Model poisoning: data owner as adversary



Threat: Influence on ML model

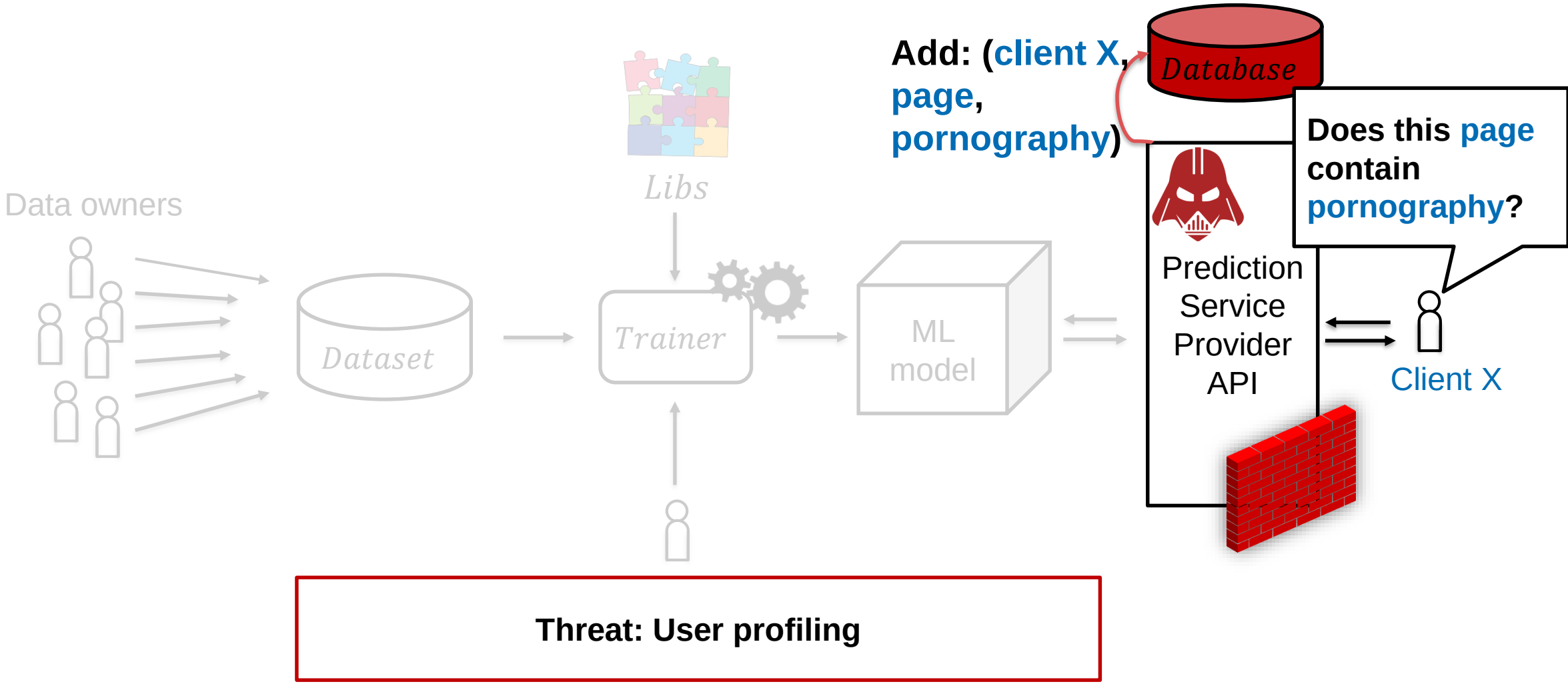
https://www.theguardian.com/technology/2016/mar/26/microsoft-deeply-sorry-for-offensive-tweets-by-ai-chatbot
https://www.theguardian.com/technology/2017/nov/07/youtube-accused-violence-against-young-children-kids-content-google-pre-school-abuse

Compromised trainer: adversary inside training pipeline



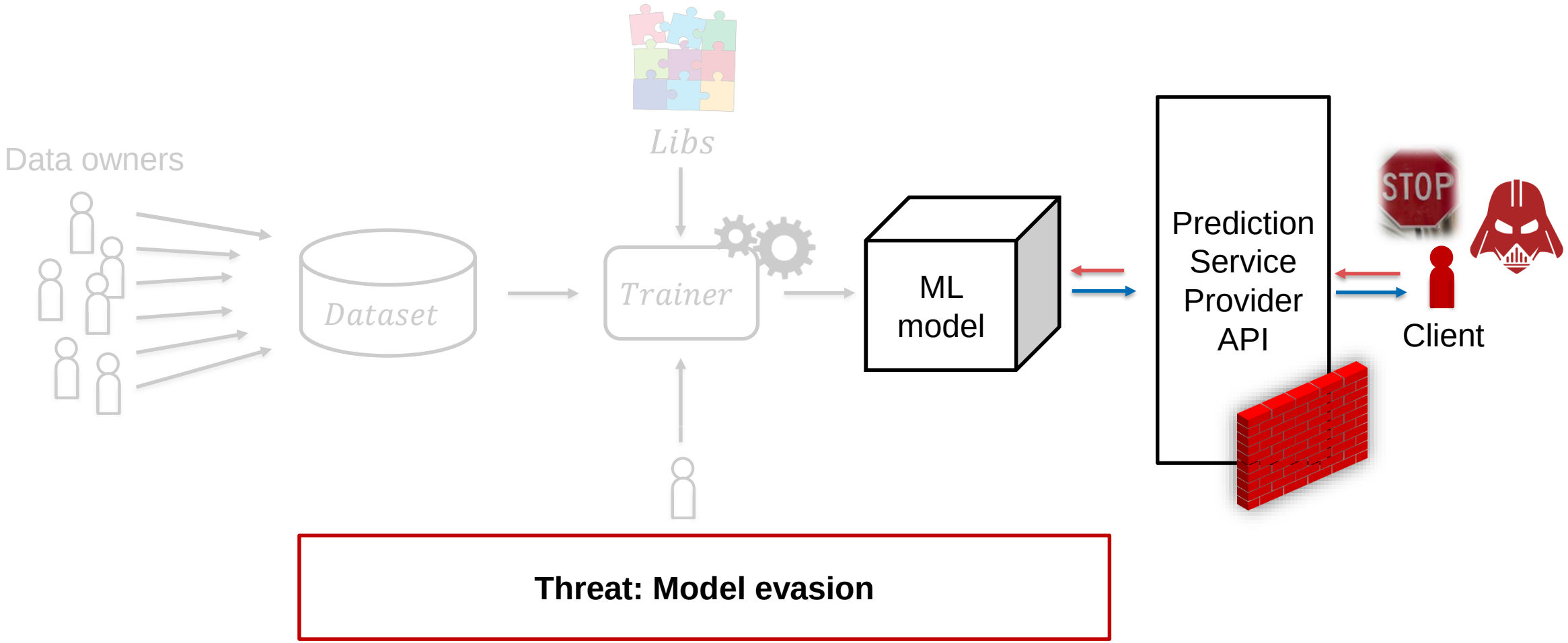
Song et al. ”Machine Learning models that remember too much”, CCS’17 Dallas. https://arxiv.org/abs/1709.07886

Compromised prediction service



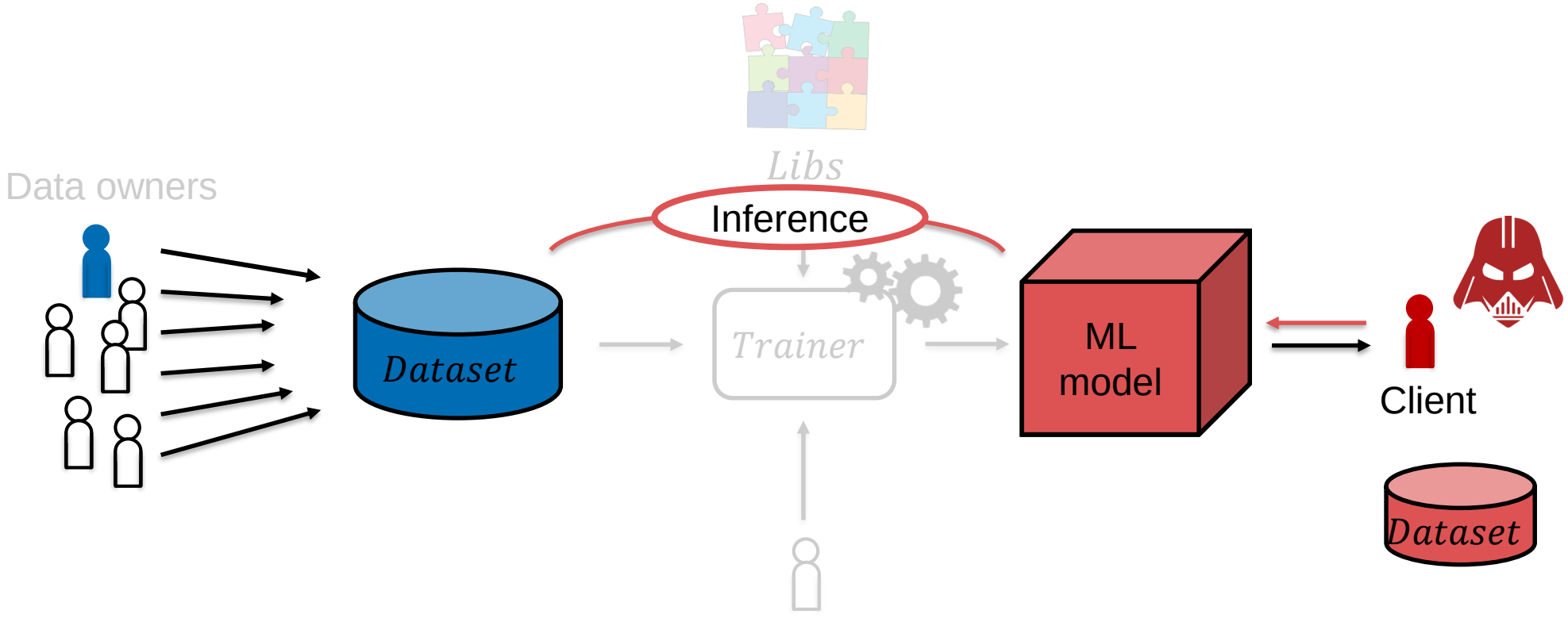
Malmi and Weber. “You are what apps you use Demographic prediction based on user's apps”

Malicious client:
Evasion of detection



Dang et al, ”Evading Classifiers by Morphing in the Dark”, CCS’17. https://arxiv.org/abs/1705.07535
Evtimov, I. Et al. ”Robust Physical-World Attacks on Deep Learning Models”. https://arxiv.org/abs/1707.08945

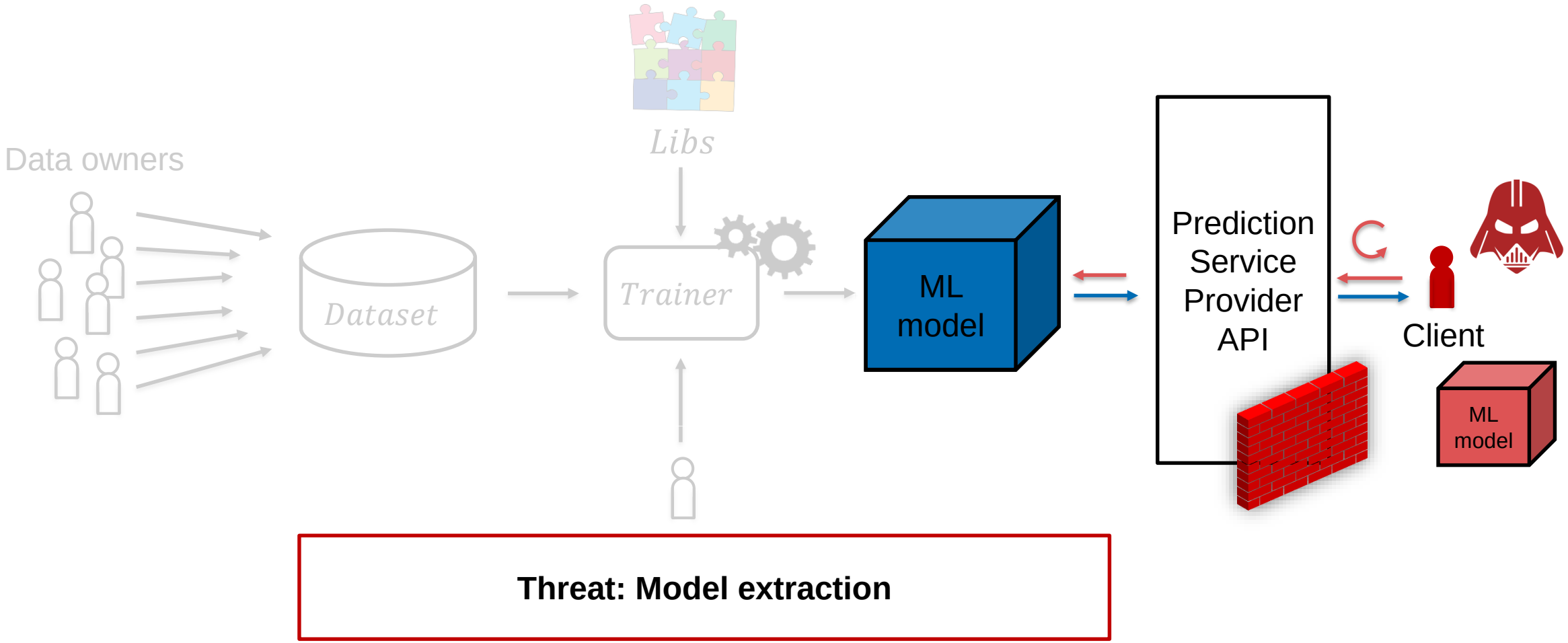
Malicious client:
Inference on training data through ML model



Threat: Model inversion, membership inference

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.

Malicious client:
Theft of ML model



Tramer et al, ”Stealing ML models via prediction APIs”, UsenixSEC’16. https://arxiv.org/abs/1609.02943

Conclusion

Adversaries are multilateral, solutions are too

Our group is working on these problems

Come talk to us!
(also about SECCLO)

Andrew	andrew.paverd@aalto.fi
Mika	mika.juuti@aalto.fi
Asokan	n.asokan@aalto.fi

http://cs.aalto.fi/secure_systems

SECCLO

Master's Programme in Security and Cloud Computing

(Erasmus Mundus)

Applications: 4.12.2017 – 17.01.2018

Scholarships available

secclo.aalto.fi **secclo@aalto.fi** **facebook.com/secclo**

Aalto University

DTU Technical University of Denmark

EURECOM

KTH Vetenskap och Konst

NTNU Norwegian University of Science and Technology

Tartu Ülikool 1632 Universitas Tartuensis

Co-funded by the Erasmus+ Programme of the European Union