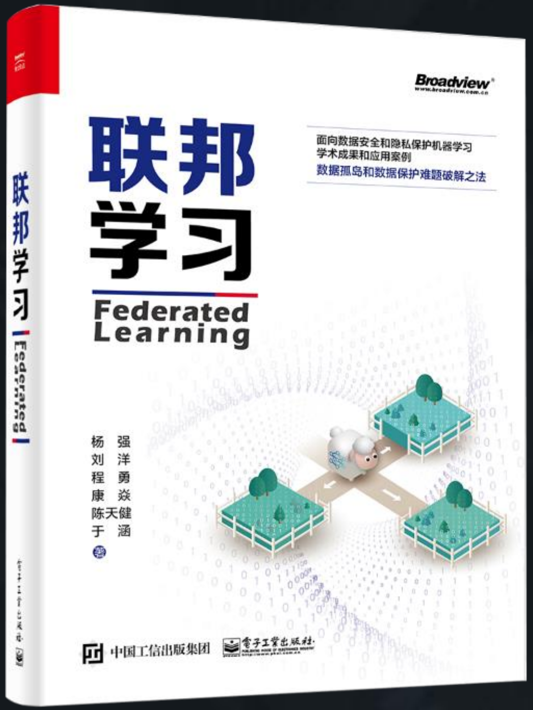
WeBank 微众·AI

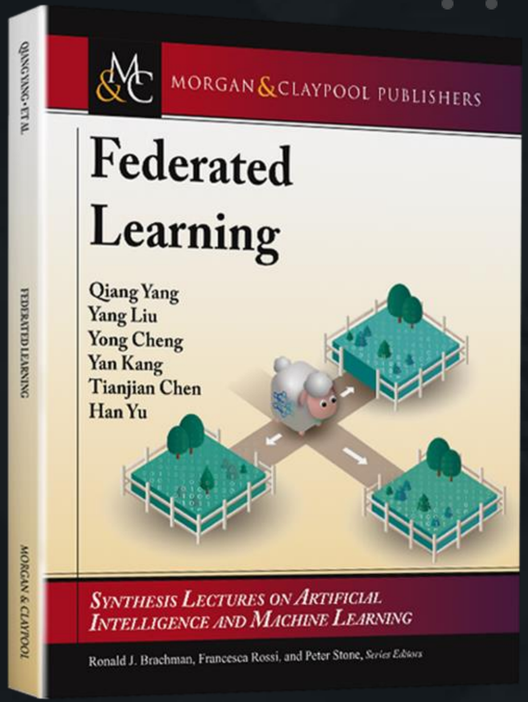
Three New Laws of AI

Qiang Yang

CAIO, WeBank,
Chair Professor, HKUST

2020.7





FedAI Ecosystem https://www.fedai.org/

Three Laws of Robotics（Asimov)

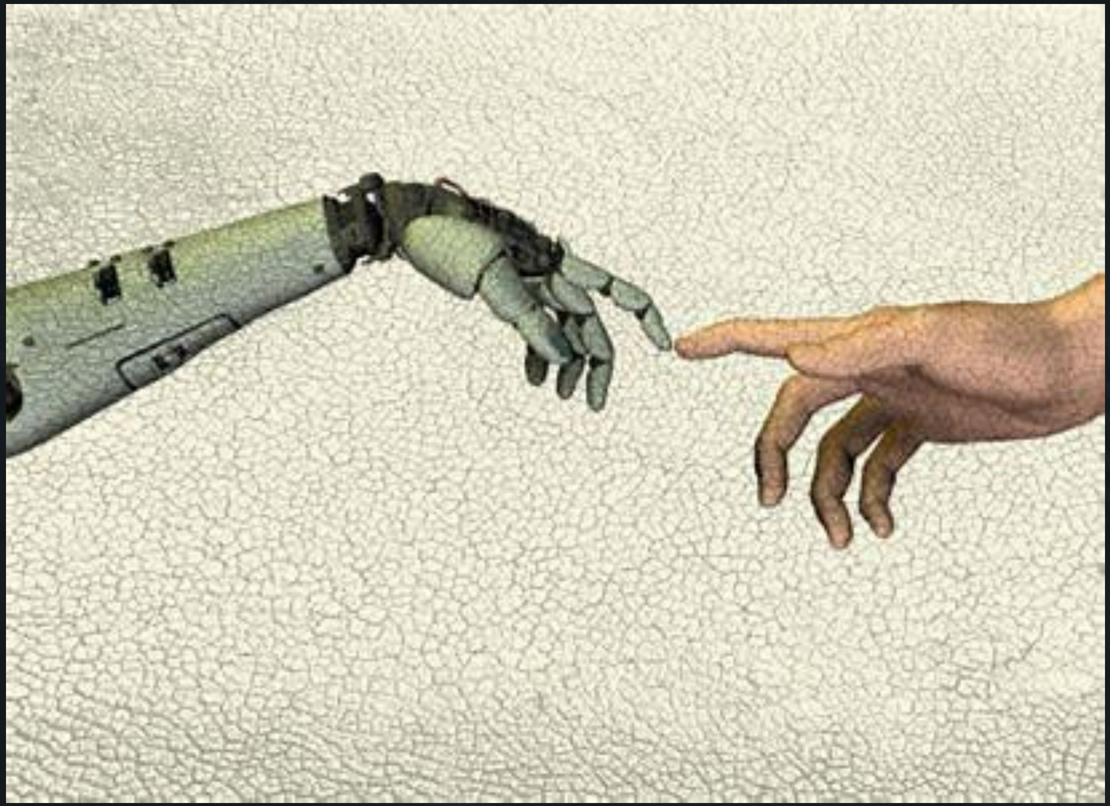
- First Law: A robot may not injure a human being, or through interaction, allow a human being to come to harm.
- Second Law: A robot must obey the orders given it by the humans except where such orders would conflict with the First Law.
- Third Law: A robot muct protect its own existence as long as such protection does not conflict with the First or Second Law.

The era of AlphaGo and our desirable AI

- **Automation, unmanned**
  - Unmanned Vehicles, commercials, etc.
- **Yet, AI needs humans as companions**
  - AI needs to explain its results to humans.
  - AI problems require human debugging.
  - AI procedure requires human supervision.
  - AI models should clarify its causality.

AI serves human beings: New Three Laws

- AI should protect user privacy.
  - Privacy is a fundamental interest of human beings.
- AI should protect model security.
  - Defense against malicious attacks.
- AI requires understanding of humans.
  - Explainability of AI models.



Law 1

AI should protect user privacy.

AI and Big Data

WeBank 微众·AI

- **<u>The strength of AI emanates from big data.</u>**

  Yet we confront mostly, small data.

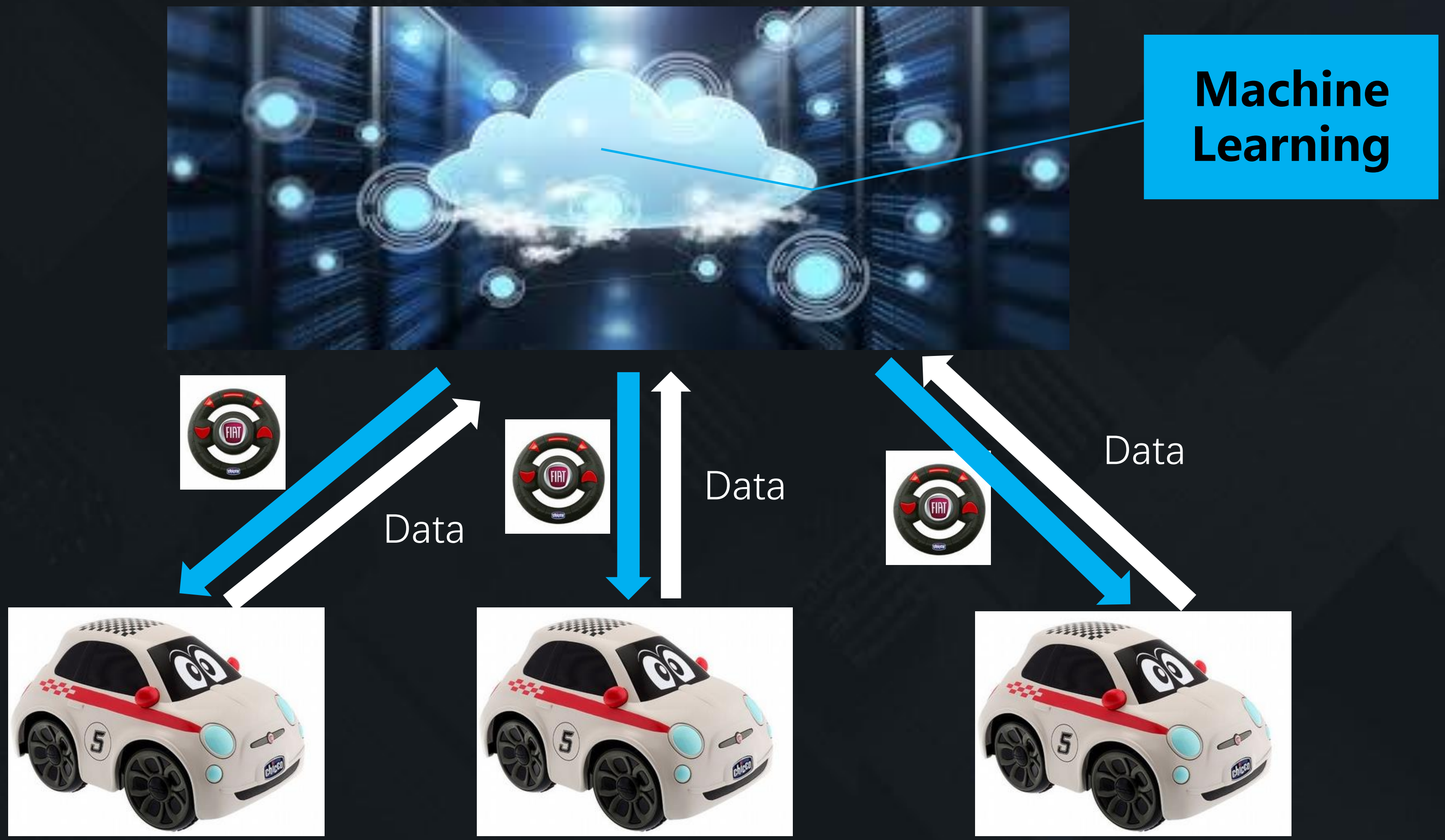
    - Law cases
    - Finance, anti money laundering
    - Medical images

Application at 4Paradigm: VIP Account Marketing

Micro loan data: > 100 Million

Large loan data < 100

Data, Machine Learning and AI ← Reality



IT giants face lawsuits under GDPR

WeBank 微众·AI

French regulator fines Google $57 million for GDPR violations

Share on Facebook Share on Twitter +

Google hasn't transparently implemented GDPR rules, French regulator claims.

1 . France's National Data Protection Commission (CNIL) found that Google provided information to users in a non-transparent way.

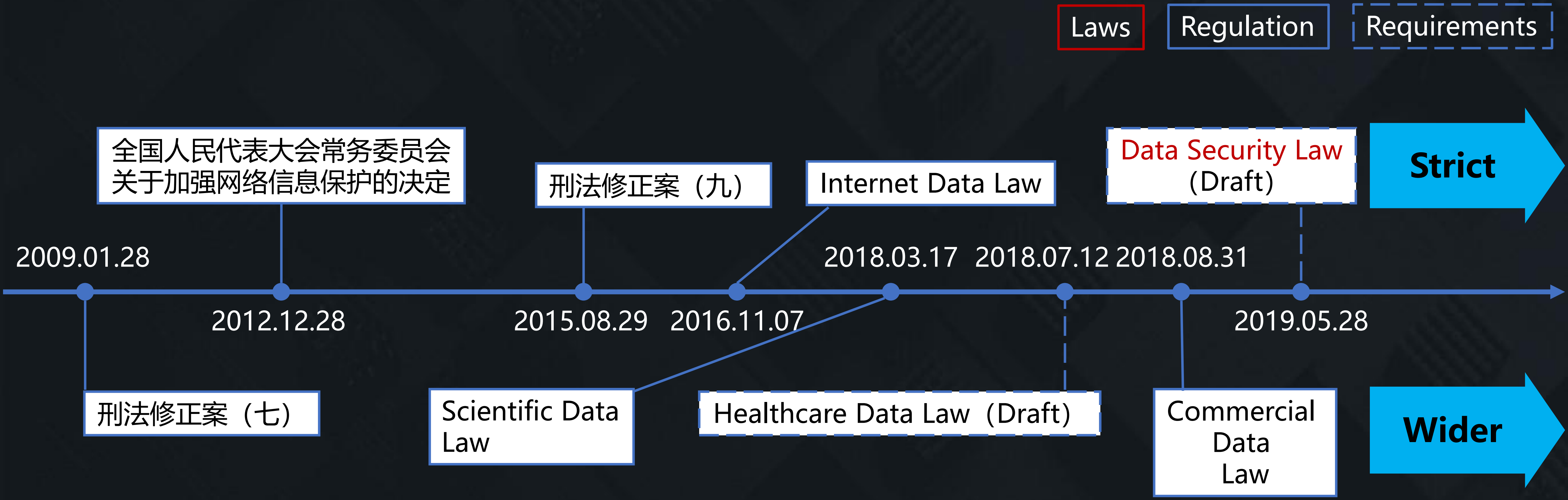
“The relevant information is accessible after several steps only, implying sometimes up to 5 or 6 actions"

- CNIL said.

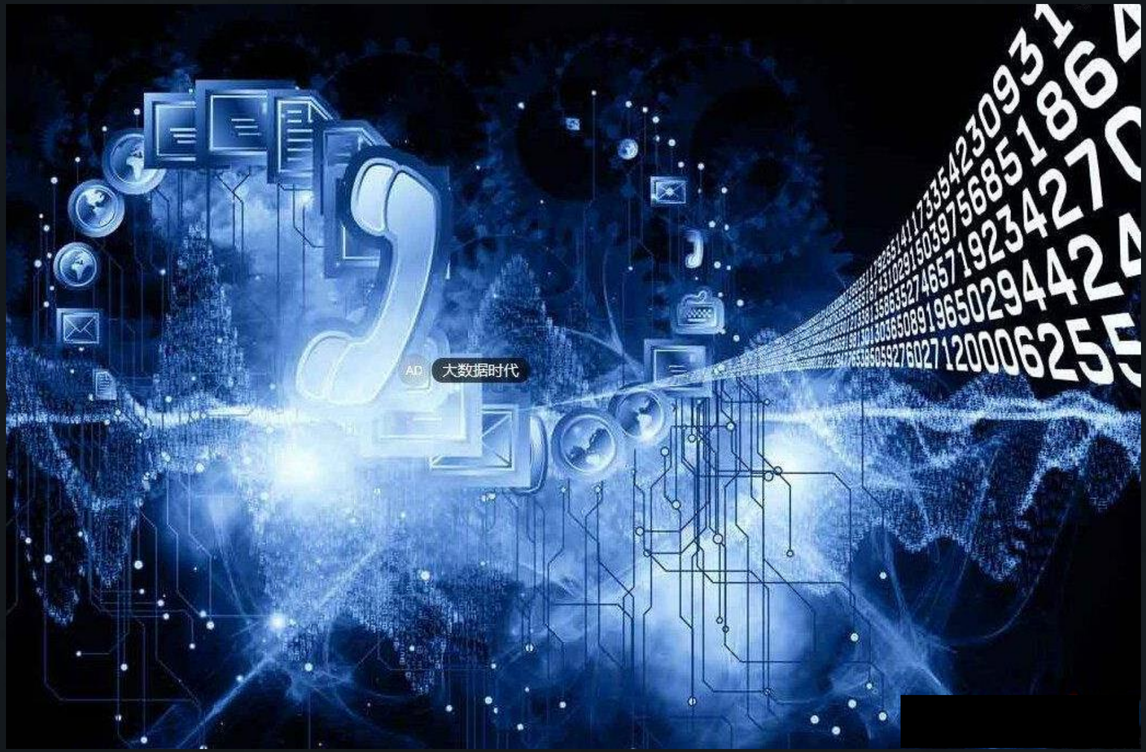
2. The users' consent, CNIL claims, "is not sufficiently informed," and it's "neither 'specific' nor 'unambiguous'."

To date, this is the largest fine issued against a company since GDPR came into effect last year.

Data Privacy Laws Increasingly More Strict



Big Data: Ideal, and Reality



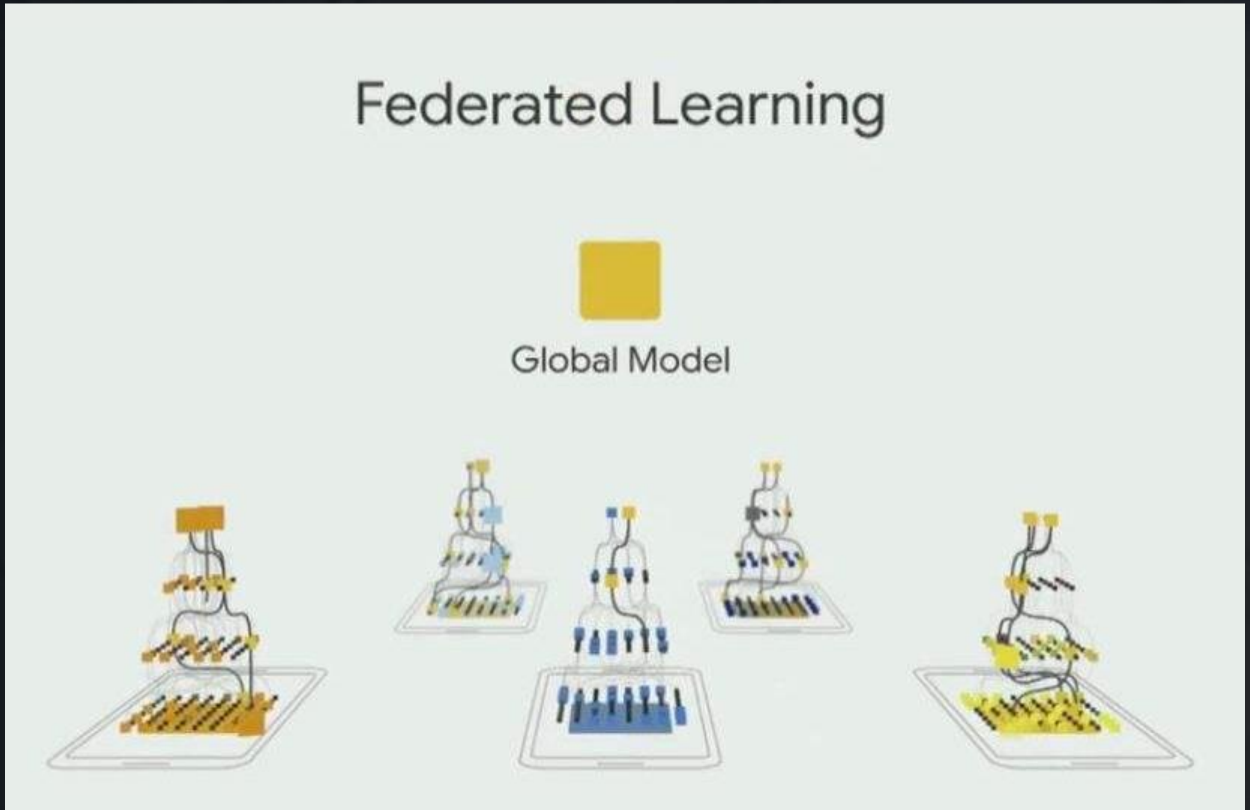
What is Federated Learning?

- Move models, instead of data
- Data usable, but invisible

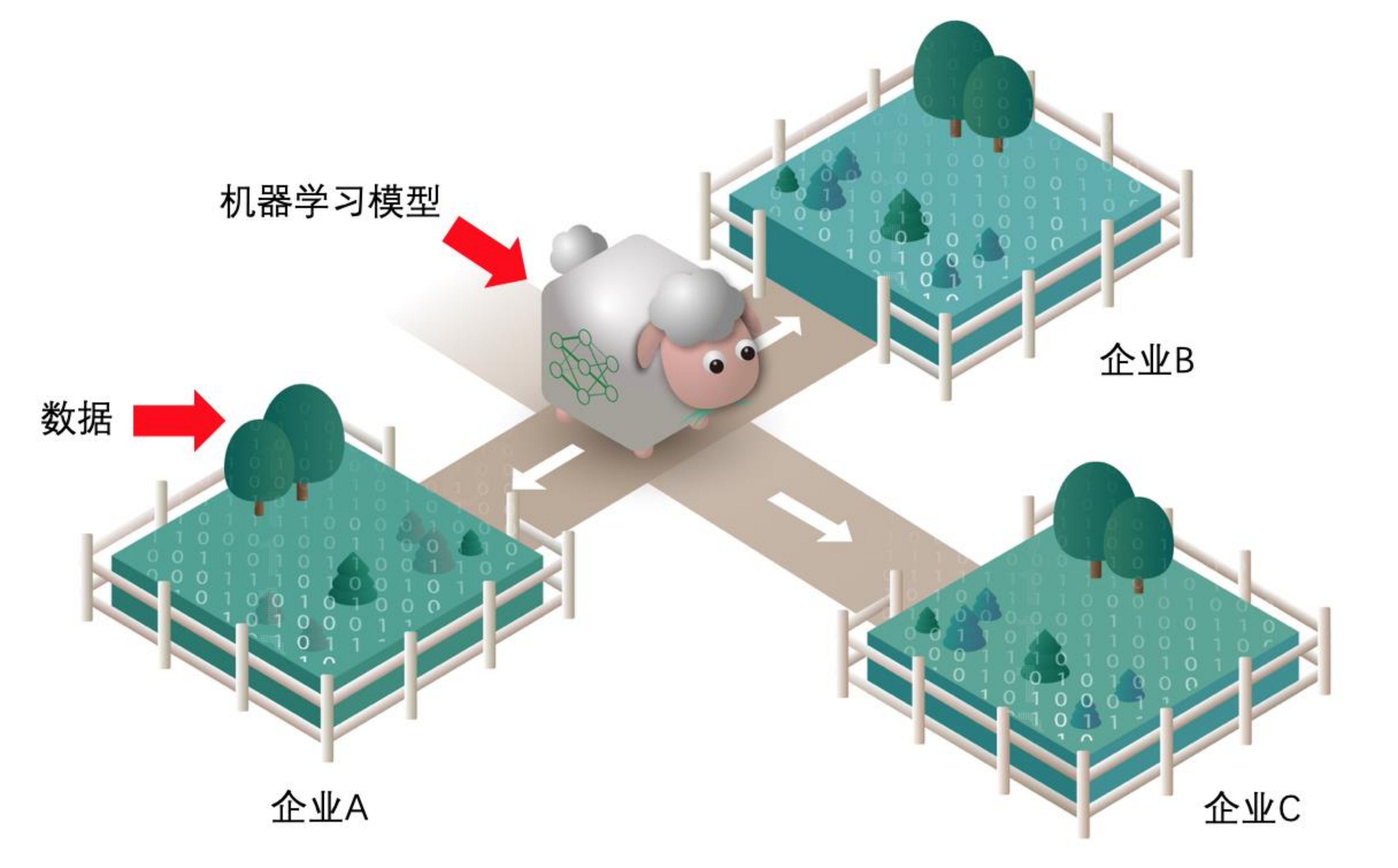
Federated Learning

WeBank 微众·AI

1. **Data Privacy**
2. **Model Protection**
3. **Better Models**
   - ➢ Party A has model A
   - ➢ Party B has model B
   - ➢ A joint model by A & B outperforms local models.



Data and models remain local.


机器学习模型
数据
企业A
企业B
企业C

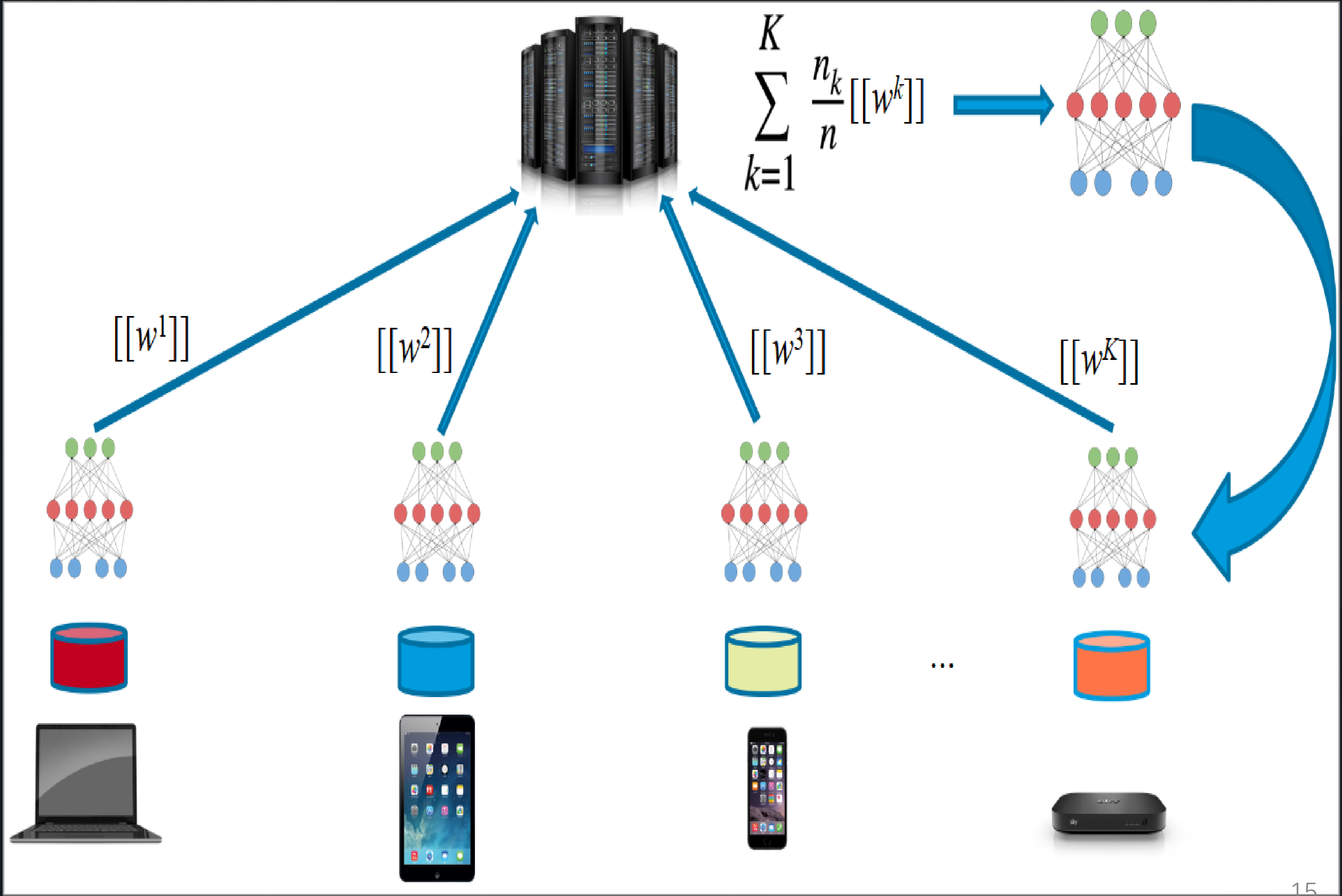
Horizontal Federated Learning（Data horizontally split）

WeBank 微众·AI

ID	X1	X2	X3
U1	9	80	600
U2	4	50	550
U3	2	35	520
U4	10	100	600

ID	X1	X2	X3
U5	9	80	600
U6	4	50	550
U7	2	35	520
U8	10	100	600

ID	X1	X2	X3
U9	9	80	600
U10	4	50	550



Key technique in Federated Learning: Encryption

WeBank 微众·AI

- Step 1: Build local models： Wi
- Step 2: Encrypt models locally
  - [[Wi]]
- Step 3: Upload encrypted models [[Wi]]
- Step 4: Aggregation of encrypted models： W=F({[[Wi]], i=1,}) 2, …
- Step 5: Local participants download W.
- Step 6: Local updates W.

Q： How to build model updates from encrypted models?

- W=F({[[Wi]], i=1,}) ？

A: Homomorphic Encryption (HE)

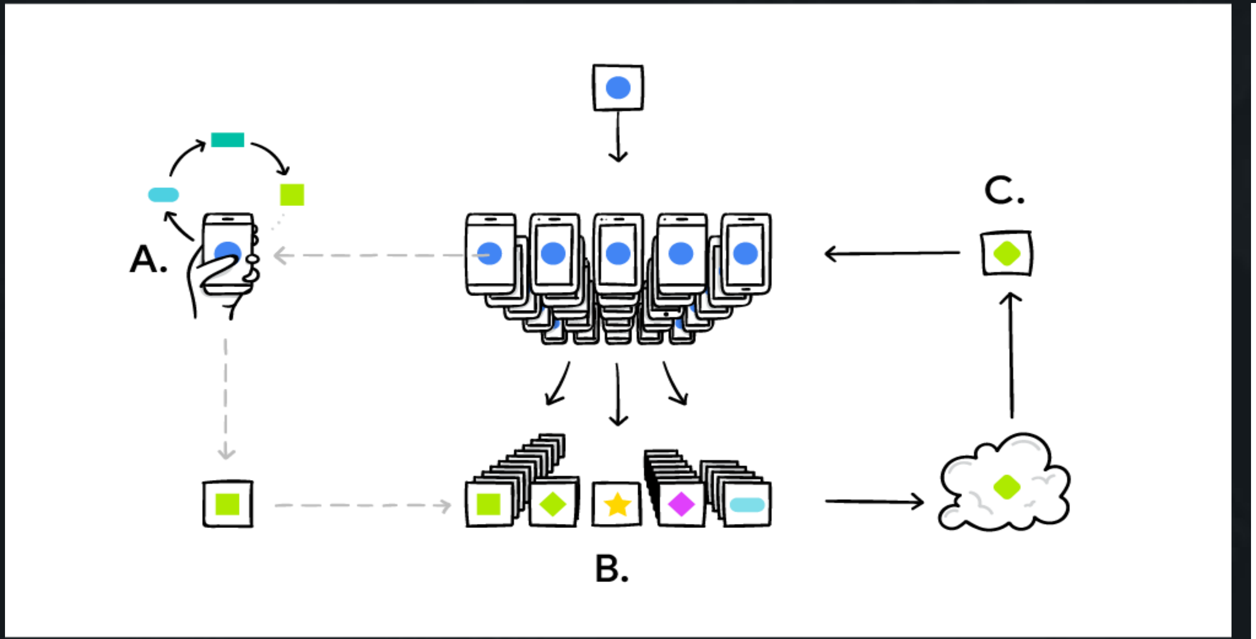
- **加法同态:**

$$\mathrm{Dec}_{\mathrm{sk}}([[u]] \oplus [[v]]) = \mathrm{Dec}_{\mathrm{sk}}([[u+v]])$$

- **标量乘法同态:**

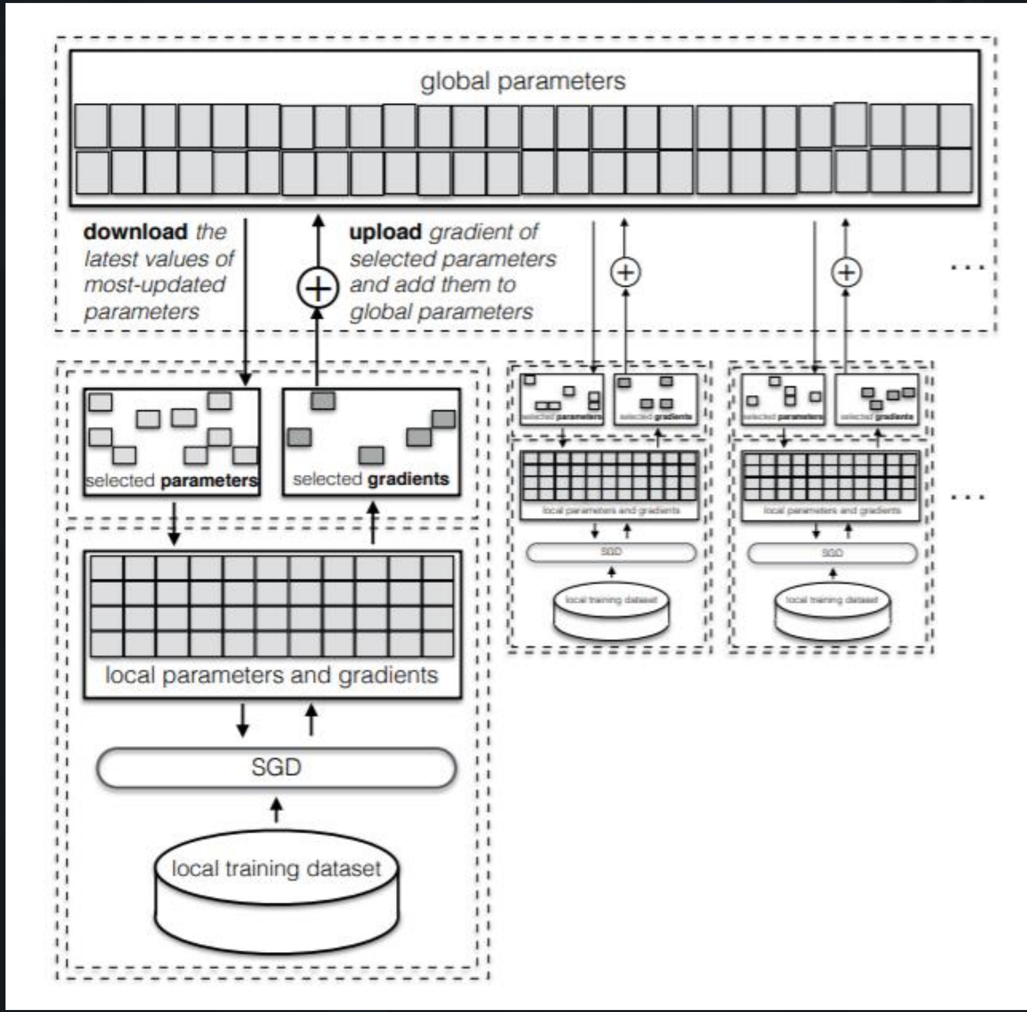
$$\mathrm{Dec}_{\mathrm{sk}}([[u]] \odot n) = \mathrm{Dec}_{\mathrm{sk}}([[u \cdot n]])$$

HFL by Google（Federated Averaging）



H. Brendan McMahan et al, *Communication-Efficient Learning of Deep Networks from Decentralized Data*, Google, 2017

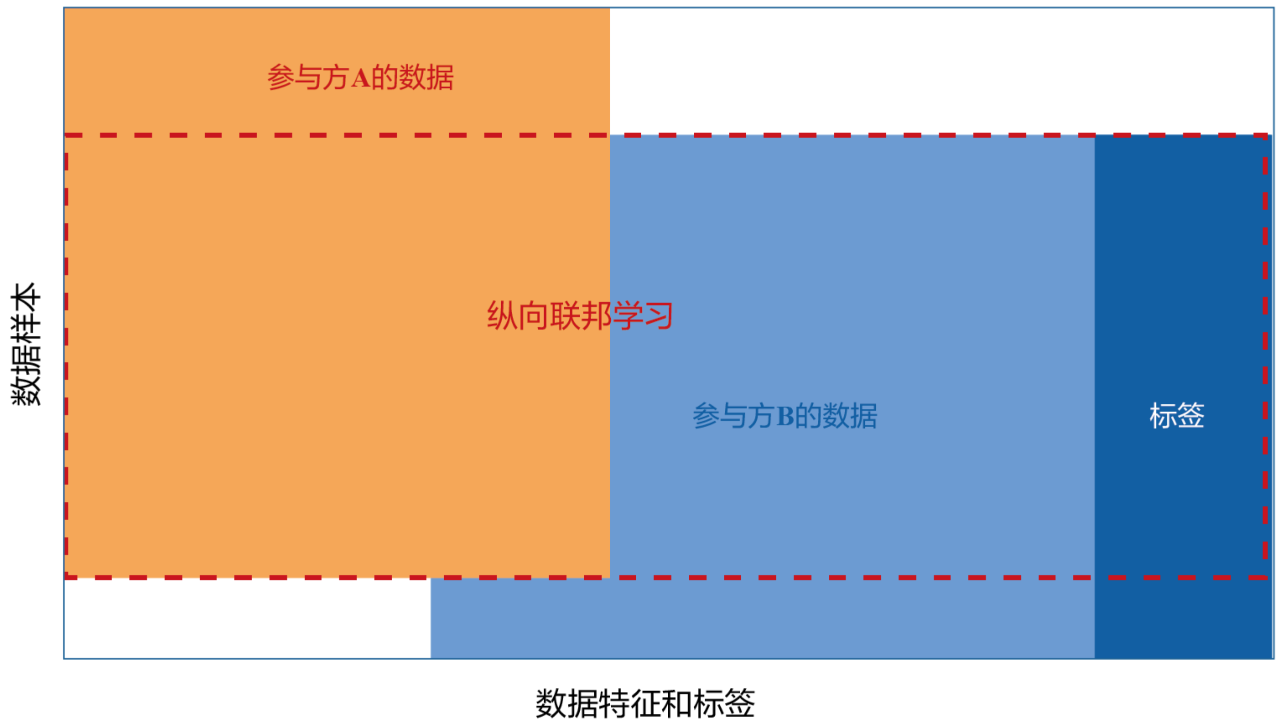
- Smartphone participants. One server and multiple users.
- Identical features
- Local training
- Select participants at each round

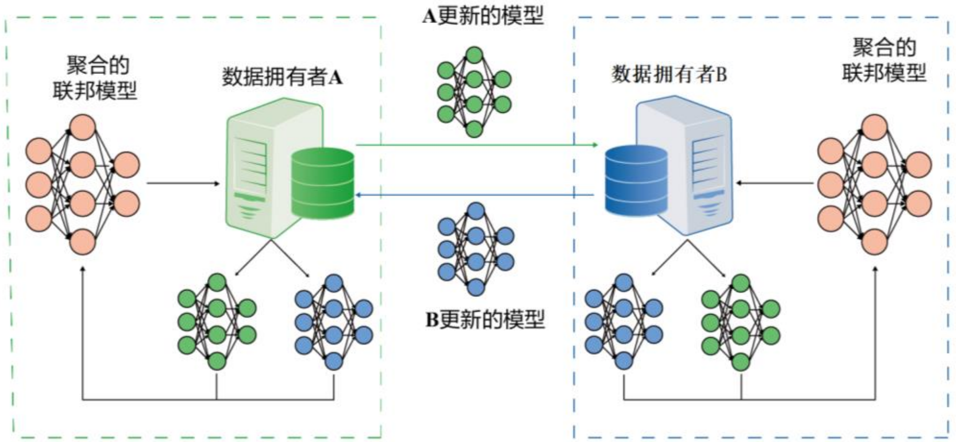


Reza Shokri and Vitaly Shmatikov. 2015. *Privacy-Preserving Deep Learning*. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS ’15). ACM, New York, NY, USA, 1310–1321.

- Select parameters to update.

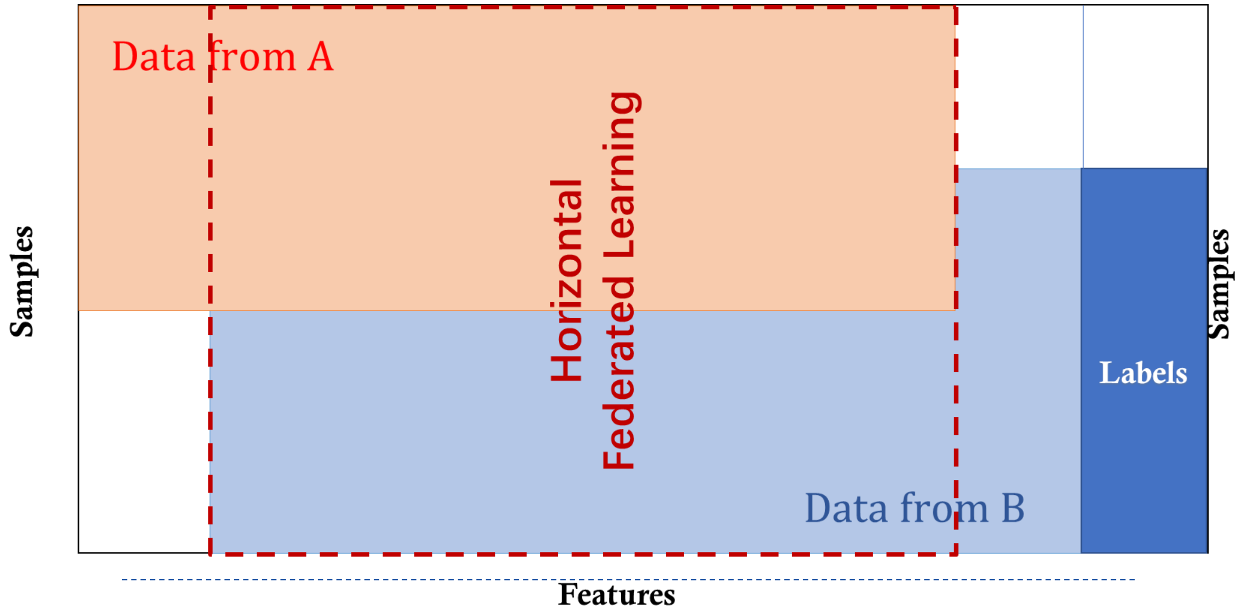
Vertical Federated Learning（Different features, overlapping ID）





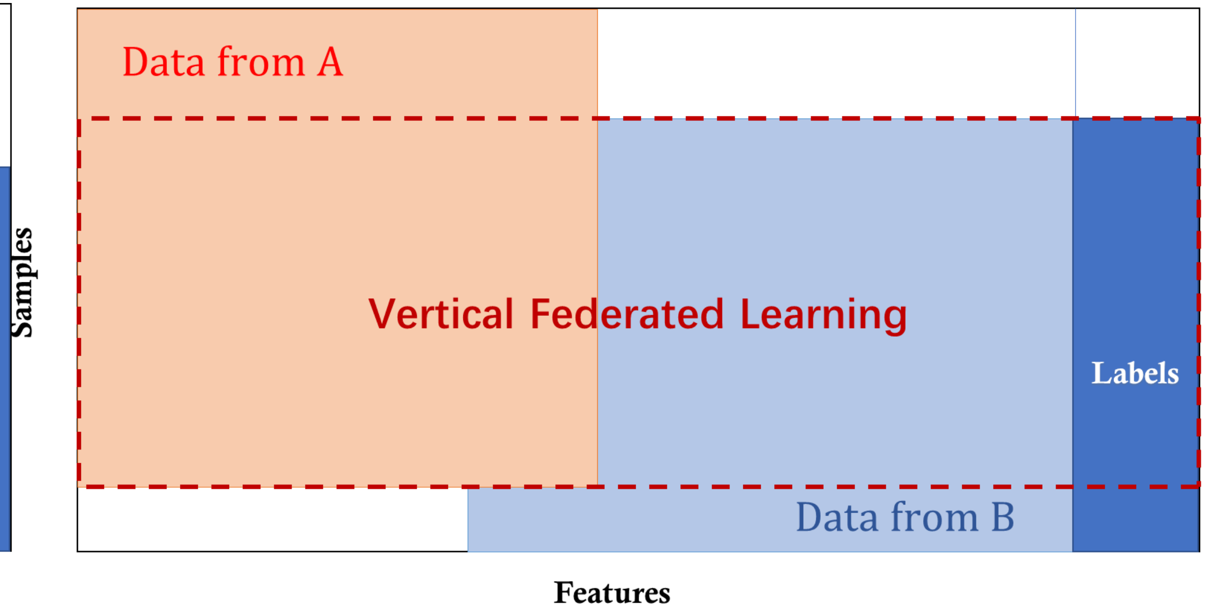
Categorization of Federated Learning

Horizontal (data split) FL



- Identical Features

Vertical (data split) FL

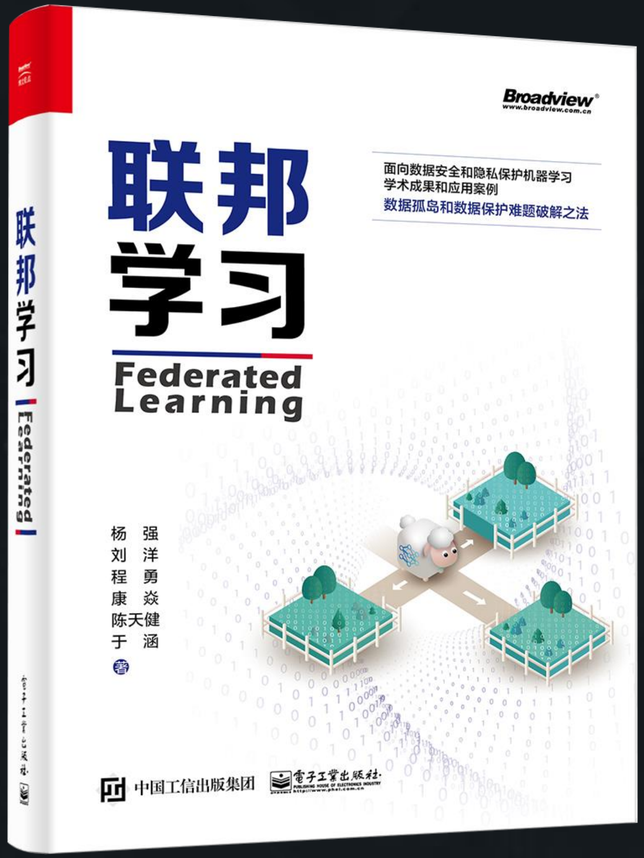


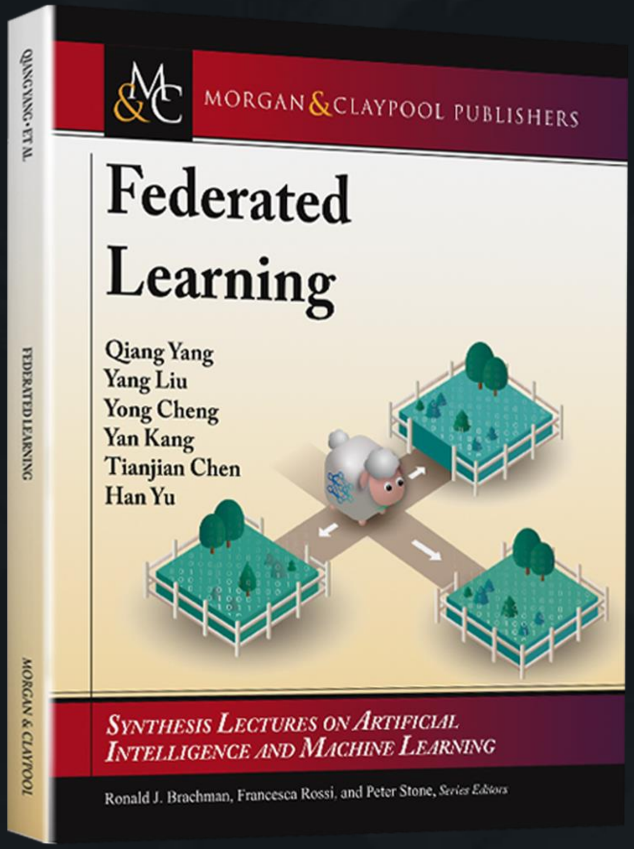
- Identical user IDs

Q. Yang, Y. Liu, T. Chen & Y. Tong, Federated machine learning: Concepts and applications, *ACM Transactions on Intelligent Systems and Technology (TIST)* **10**(2), 12:1-12:19, 2019

WeBank
微众·AI

Recent advances in federated learning research.





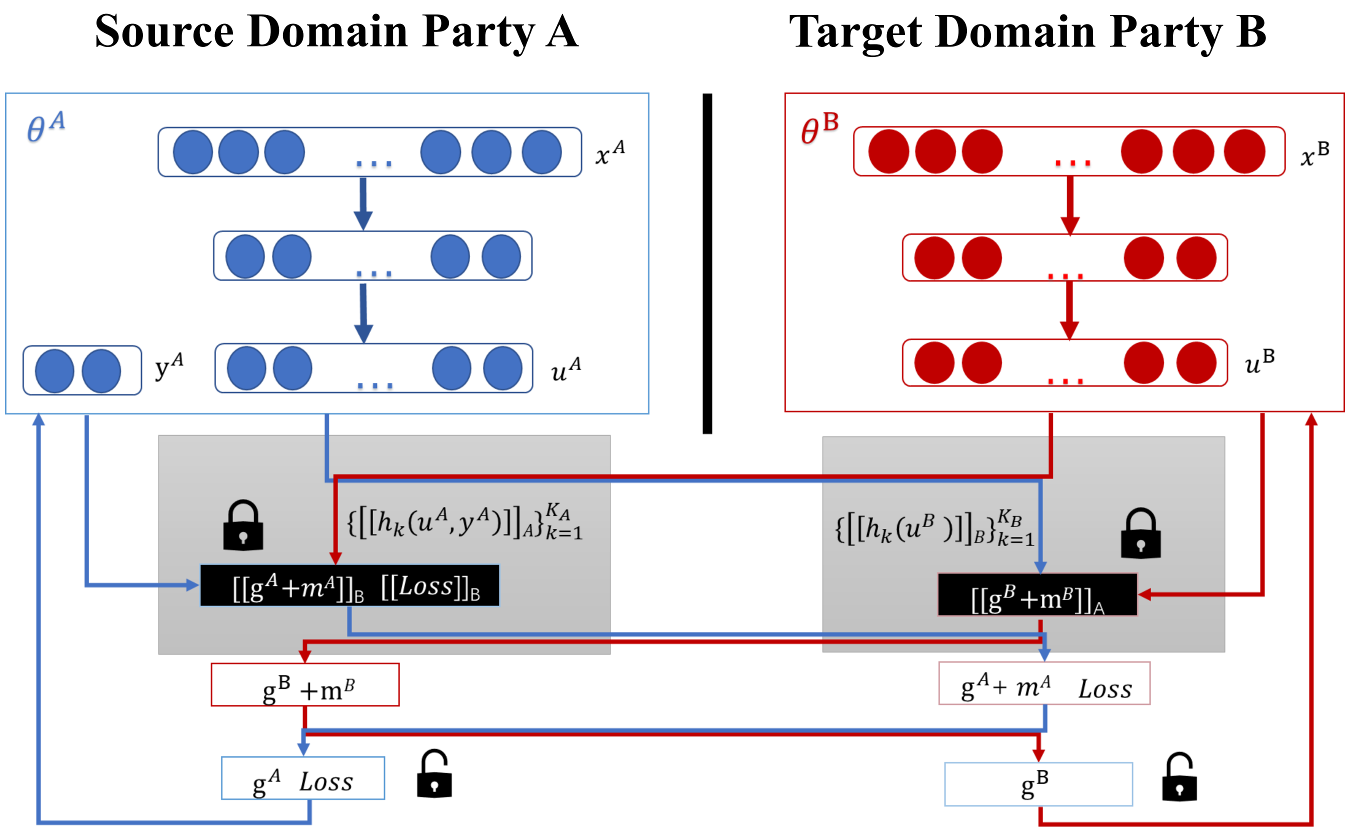
Advances and Open Problems in Federated Learning

Peter Kairouz[7*] H. Brendan McMahan[7*] Brendan Avent[21] Aurélien Bellet[9]
Mehdi Bennis[19] Arjun Nitin Bhagoji[13] Keith Bonawitz[7] Zachary Charles[7]
Graham Cormode[23] Rachel Cummings[6] Rafael G.L. D'Oliveira[14]
Salim El Rouayheb[14] David Evans[22] Josh Gardner[24] Zachary Garrett[7]
Adrià Gascón[7] Badih Ghazi[7] Phillip B. Gibbons[2] Marco Gruteser[7,14]
Zaid Harchaoui[24] Chaoyang He[21] Lie He [4] Zhouyuan Huo [20]
Ben Hutchinson[7] Justin Hsu[25] Martin Jaggi[4] Tara Javidi[17] Gauri Joshi[2]
Mikhail Khodak[2] Jakub Konečný[7] Aleksandra Korolova[21] Farinaz Koushanfar[17]
Sanmi Koyejo[7,18] Tancrède Lepoint[7] Yang Liu[12] Prateek Mittal[13]
Mehryar Mohri[7] Richard Nock[1] Ayfer Özgür[15] Rasmus Pagh[7,10]
Mariana Raykova[7] Hang Qi[7] Daniel Ramage[7] Ramesh Raskar[11]
Dawn Song[16] Weikang Song[7] Sebastian U. Stich[4] Ziteng Sun[3]
Ananda Theertha Suresh[7] Florian Tramèr[15] Praneeth Vepakomma[11] Jianyu Wang[2]
Li Xiong[5] Zheng Xu[7] Qiang Yang[8] Felix X. Yu[7] Han Yu[12] Sen Zhao[7]

[1]Australian National University, [2]Carnegie Mellon University, [3]Cornell University,
[4]École Polytechnique Fédérale de Lausanne, [5]Emory University, [6]Georgia Institute of Technology,
[7]Google Research, [8]Hong Kong University of Science and Technology, [9]INRIA, [10]IT University of Copenhagen,
[11]Massachusetts Institute of Technology, [12]Nanyang Technological University, [13]Princeton University,
[14]Rutgers University, [15]Stanford University, [16]University of California Berkeley,
[17] University of California San Diego, [18]University of Illinois Urbana-Champaign, [19]University of Oulu,
[20]University of Pittsburgh, [21]University of Southern California, [22]University of Virginia,
[23]University of Warwick, [24]University of Washington, [25]University of Wisconsin–Madison

Towards Secure and Efficient Federated Transfer Learning

Towards Secure and Efficient FTL



Step 1

Party A and B send public keys to each other

Step 2

Parties compute, encrypt and exchange intermediate results

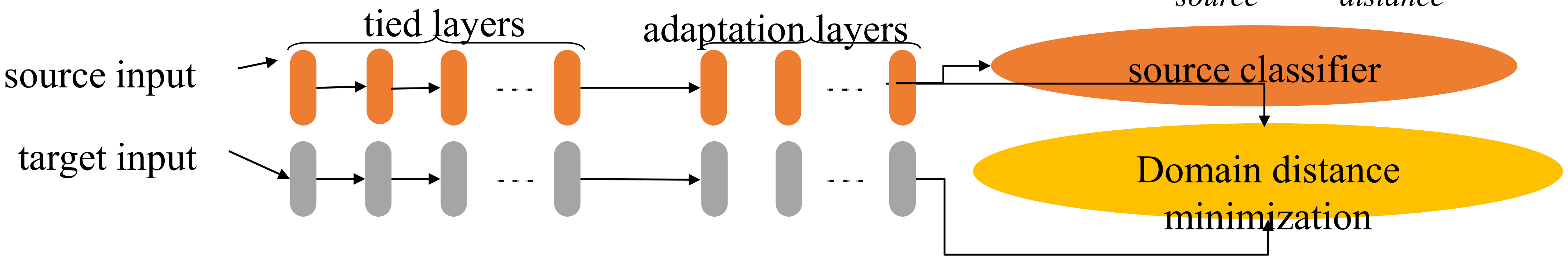
Step 3

Parties compute encrypted gradients, add masks and send to each other

Step 4

Parties decrypt gradients and exchange, unmask and update model locally

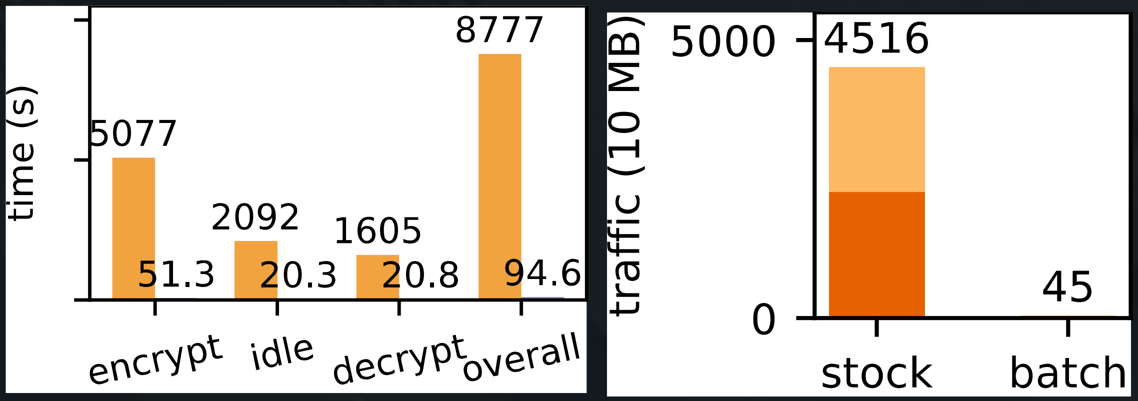
$$L = L_{source} + L_{distance}$$



BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning

- **Reducing the encryption overhead and data transfer**
  - Quantizing a gradient value into low-bit integer representations
  - Batch encryption: encoding a batch of quantized values to a long integer

- **BatchCrypt is implemented in FATE and is evaluated using popular deep learning models**
  - Accelerating the training by 23x-93x
  - Reducing the netw. footprint by 66x-101x
  - Almost no accuracy loss (<1%)



LSTM

C. Zhang, S. Li, J. Xia, W Wang, F Yan, Y. Liu, BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning, USENIX ATC'20 (accepted)

XGBoost in Federated Learning

GBDT in HFL

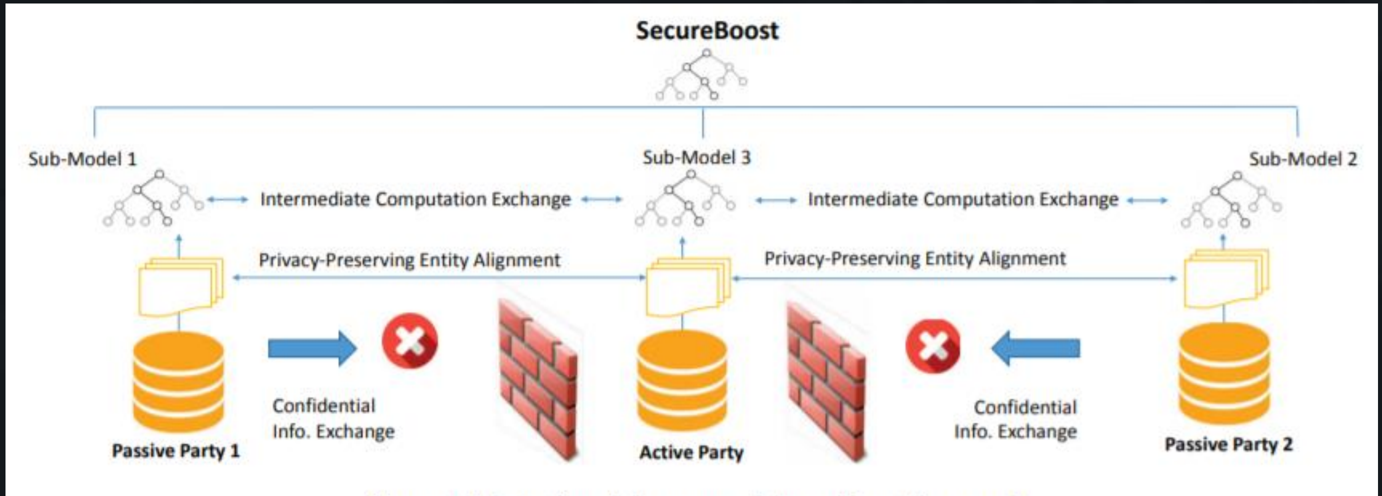
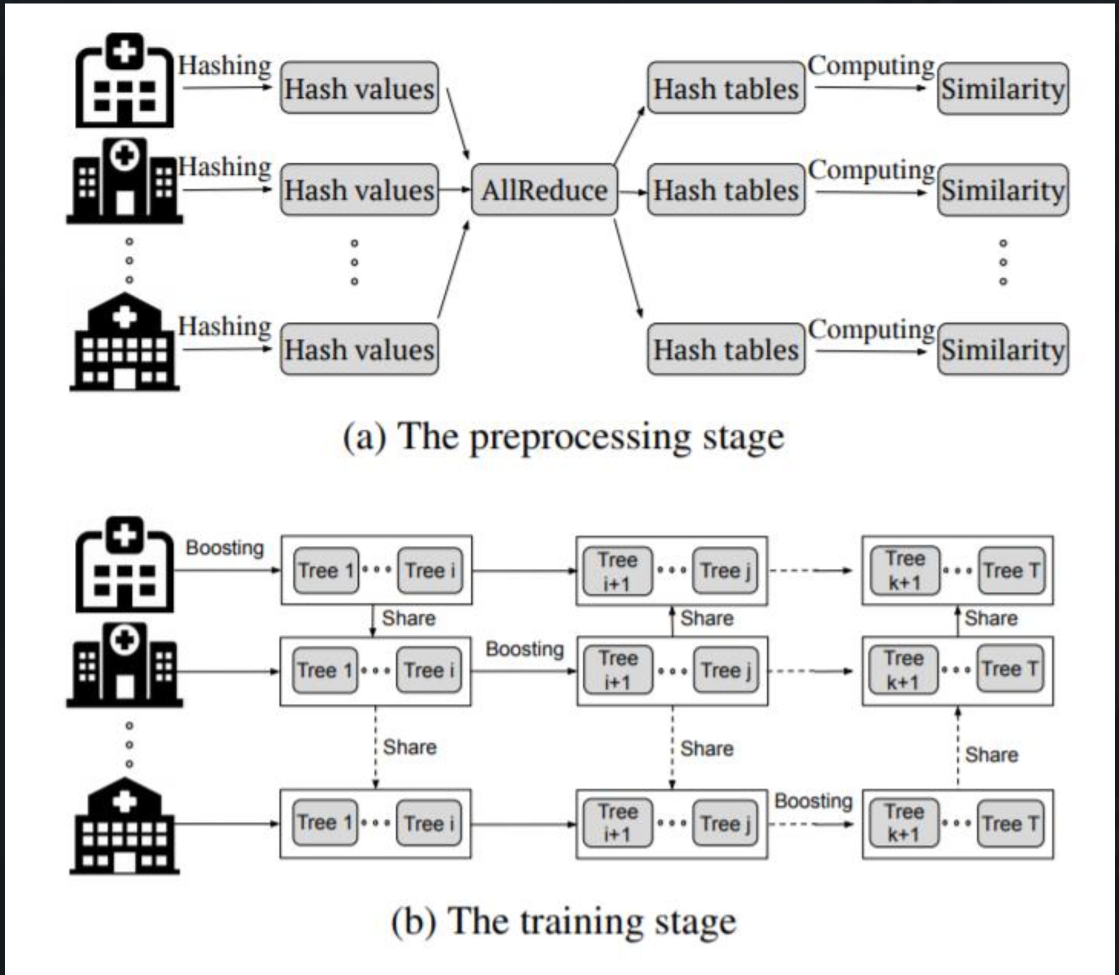


Figure 1: Illustration of the proposed SecureBoost framework

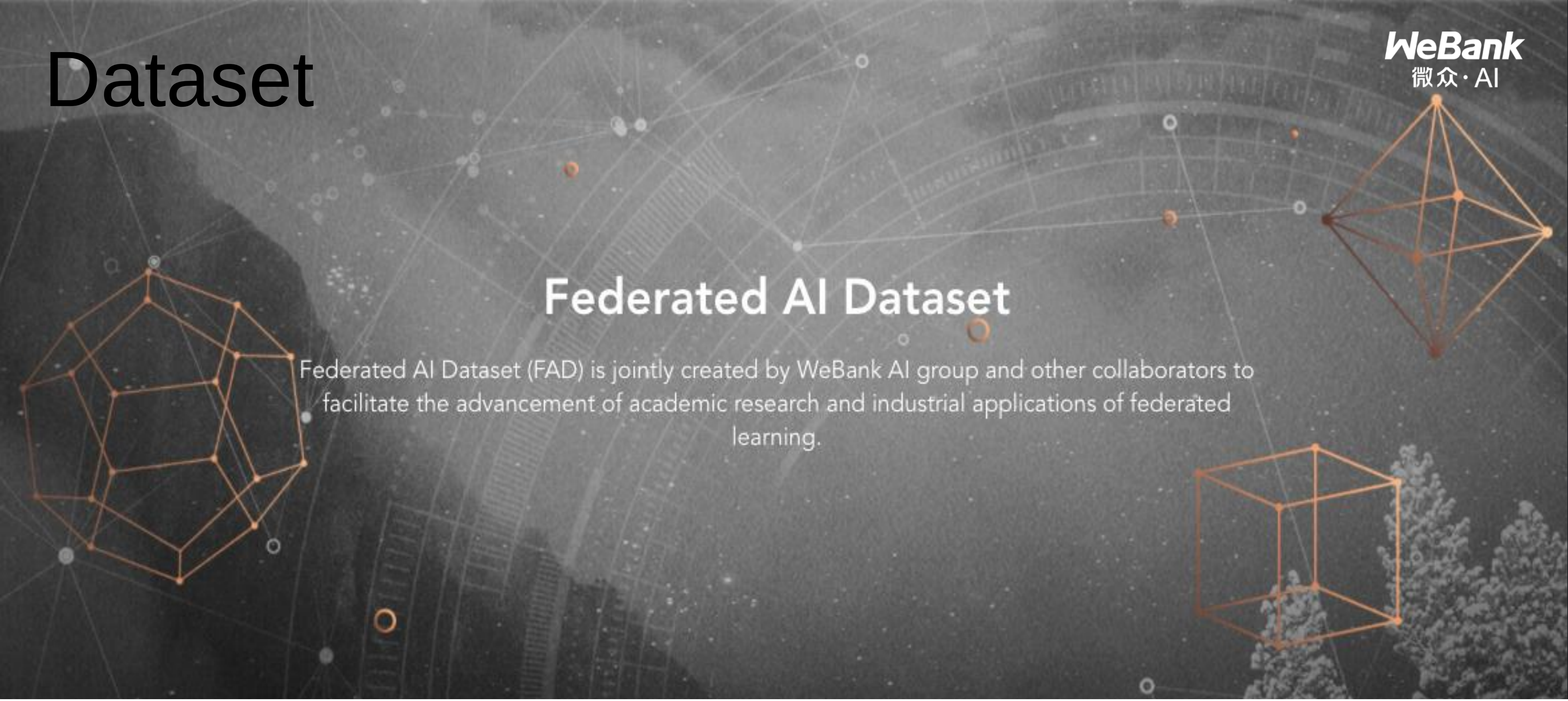
Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, Qiang Yang, SecureBoost: A Lossless Federated Learning Framework, IEEE Intelligent Systems 2020



Qinbin Li, Zeyi Wen, Bingsheng He, Practical Federated Gradient Boosting Decision Trees, AAAI, 2019

Dataset for Federated Learning

Dataset



- Web: https://dataset.fedai.org/
- Github: https://github.com/FederatedAI/FATE
- Arxiv: Real-World Image Datasets for Federated Learning

Dataset

Dataset

The FedVision Project

This project is supported by WeBank AI group and ExtremeVision to boost the academic research and industrial applications of computer vision based on federated learning.

VIEW MORE

- Web: https://dataset.fedai.org/ Github: https://github.com/FederatedAI/FATE Arxiv: Real-World Image Datasets for Federated Learning

IEEE Standard P3652.1 – Federated Machine Learning

Title

Guide for Architectural Framework and Application of Federated Machine Learning

Scope

- Description and definition of federated learning
- The types of federated learning and the application scenarios to which each type applies
- Performance evaluation of federated learning
- Associated regulatory requirements

Call for participation

- More info: https://sagroups.ieee.org/3652-1/

IEEE Standard Association is a open platform and we are welcoming more organizations to join the working group.

FATE: Federated AI Technology Enabler

WeBank
微众·AI

Desire:

- Industry-level federated learning system
- Enabling joint modeling by multiple corporations under data protection regulations.

Principles

- Support of popular algorithms: federated modeling of machine learning, deep learning and transfer learning.
- Support of multiple secure computation protocols: Homomorphic encryption, secret sharing, hashing, etc.
- User-friendly cross-domain information management scheme that alleviates the hardness of auditing federated learning.

Github: https://github.com/FederatedAI/FATE

Website: https://FedAI.org

FATE milestones

WeBank 微众·AI

2019.02

FATEv0.1
Horizontal/Vertical LR, SecureBoost, Eggroll | Federated Network

201903

GitHub Stars exceeds 100
The first external ontributor

201905

FATEv0.2
FATE-Serving
Federated Feature Engineering.

201906

FATEv0.3
FDN updates FATE
FATE contributes to Linux Foundation

201908

FATEv1.0
FATE-FLOW | FATEBoard

201910

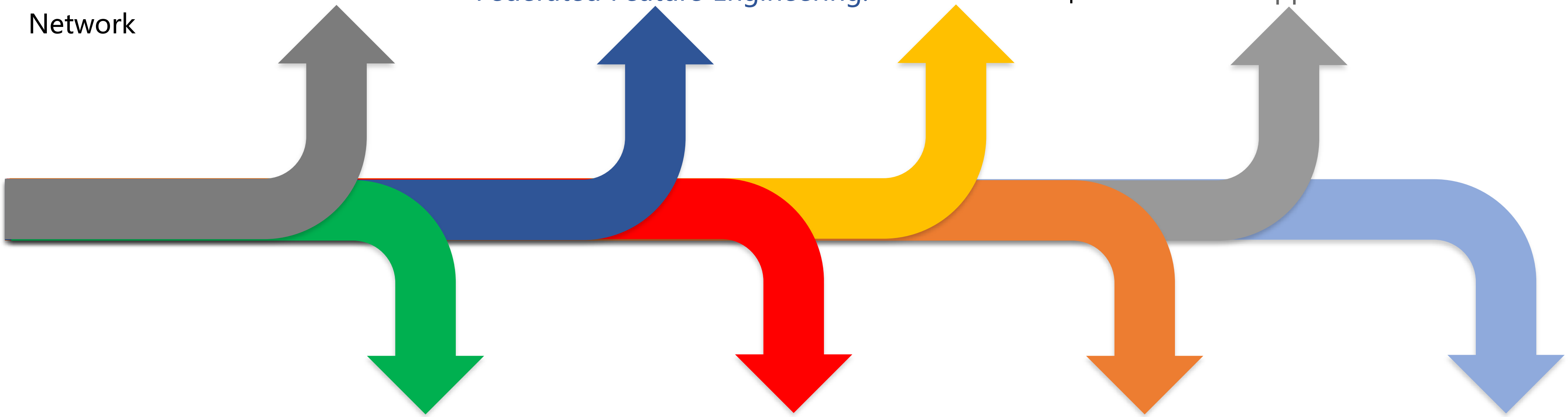
FATE-v1.1
Support Horizontal/Vertical Federated Deep Learning and Spark

201911

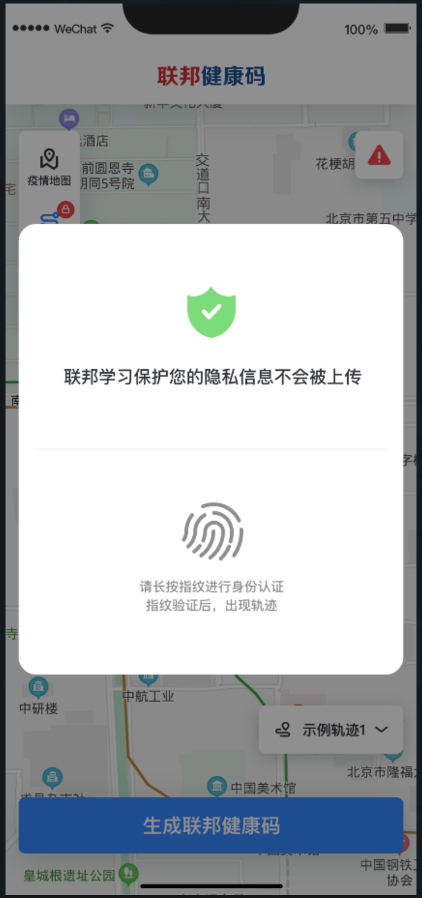
FATE-v1.2
Vertical federated deep learning
Support SecretShare Protocol

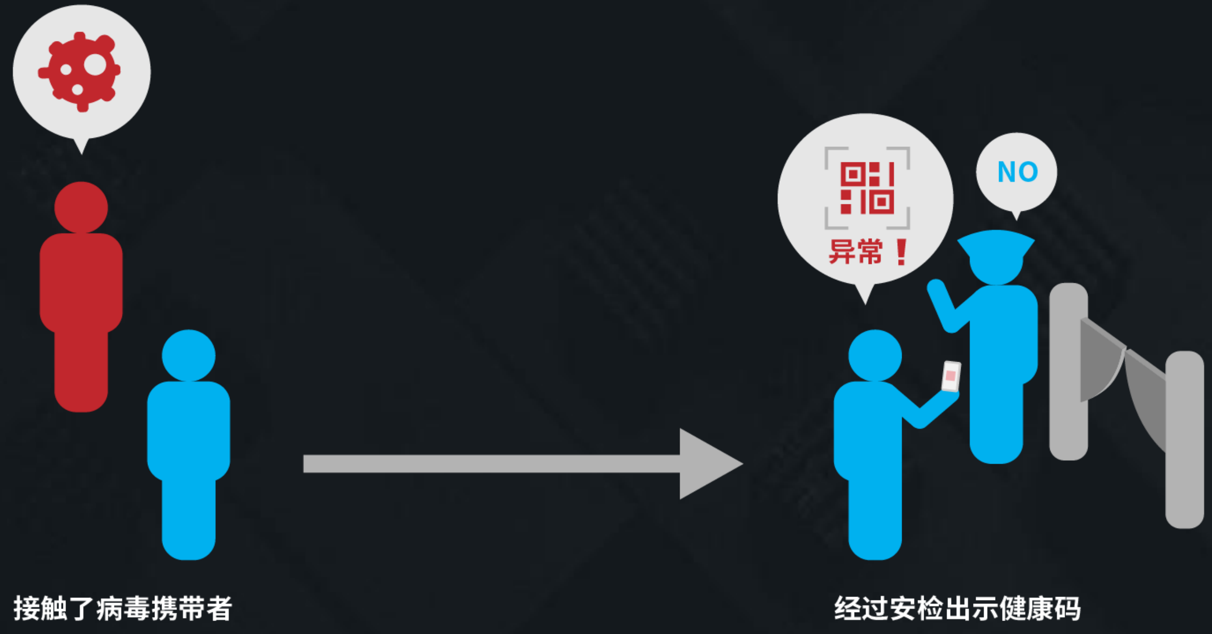
201912

FATE-v1.3
Support Heterogeneous Computation



Federated Health Code: Defending COVID 19 with privacy

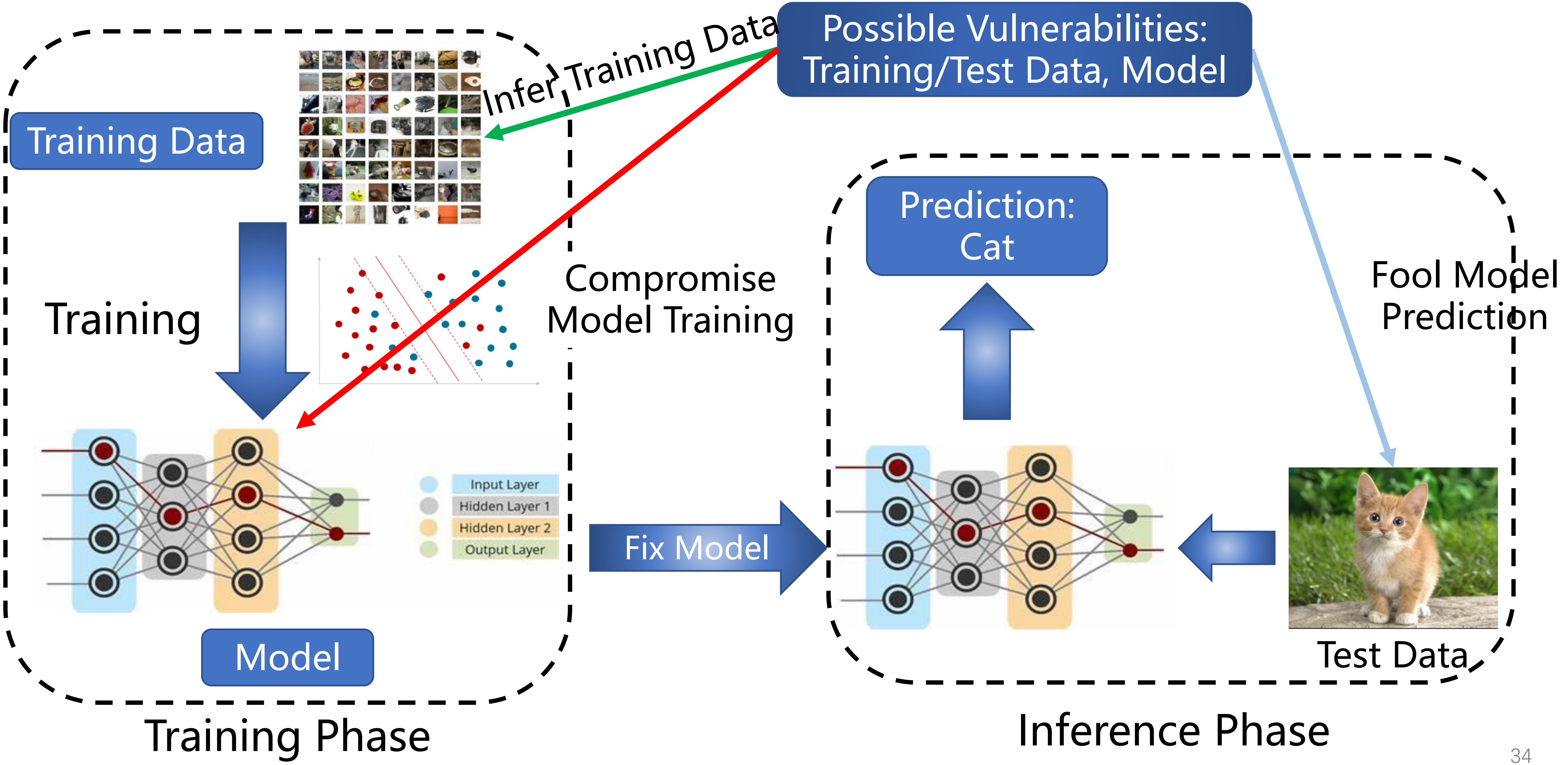




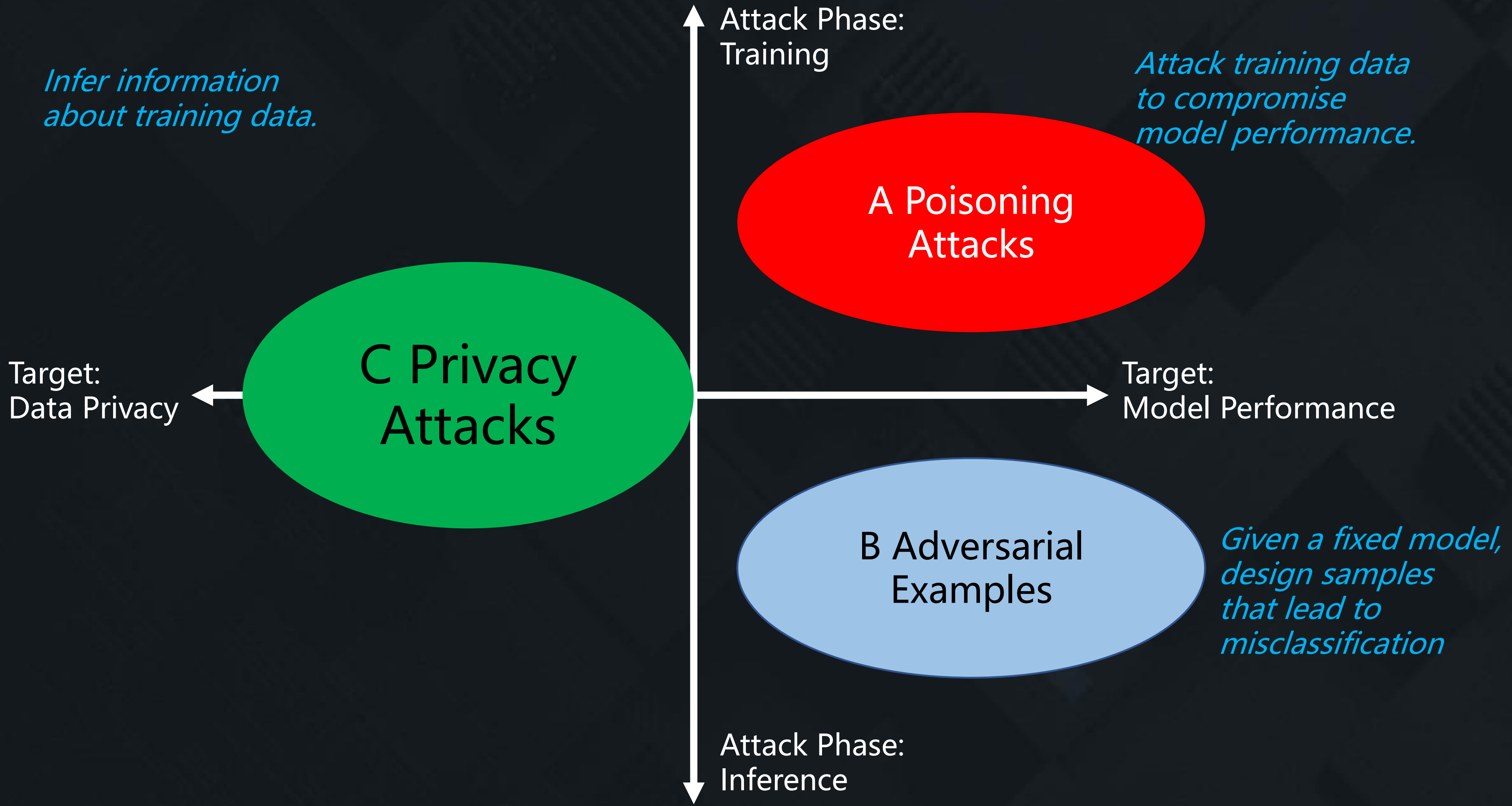
Law 2

AI should be safe.

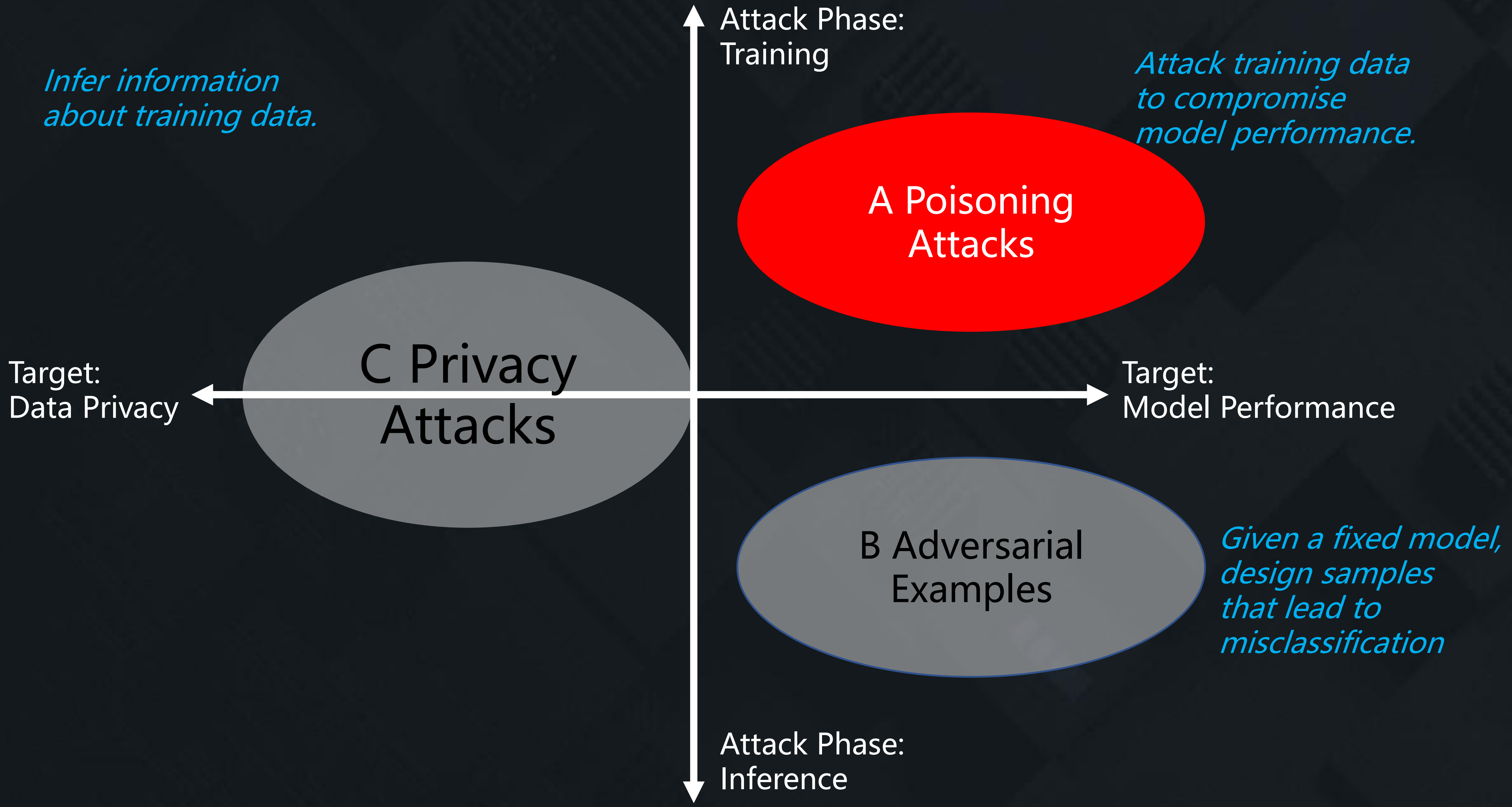
Vulnerabilities in Machine Learning



Attacks to Machine Learning



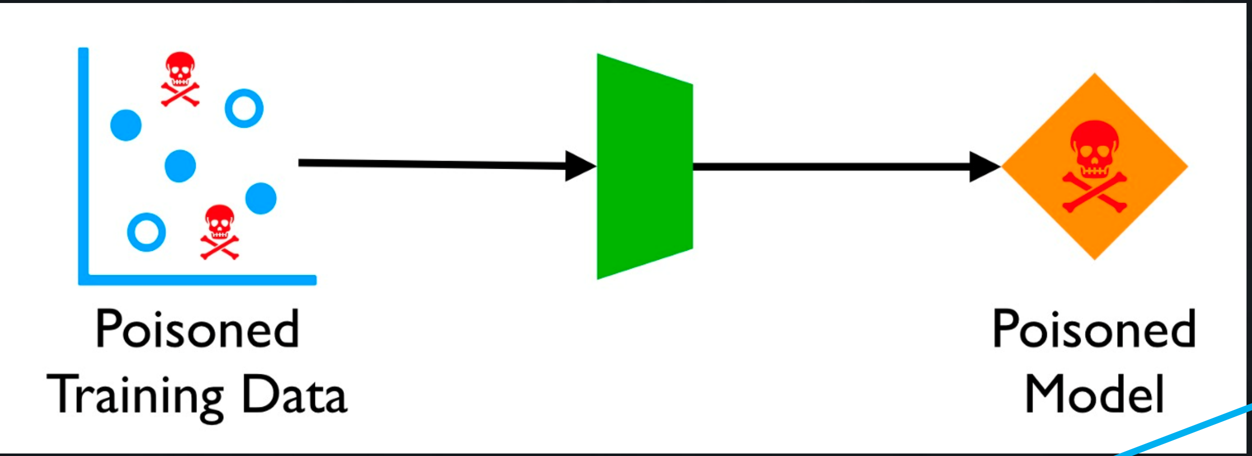
Attacks to Machine Learning

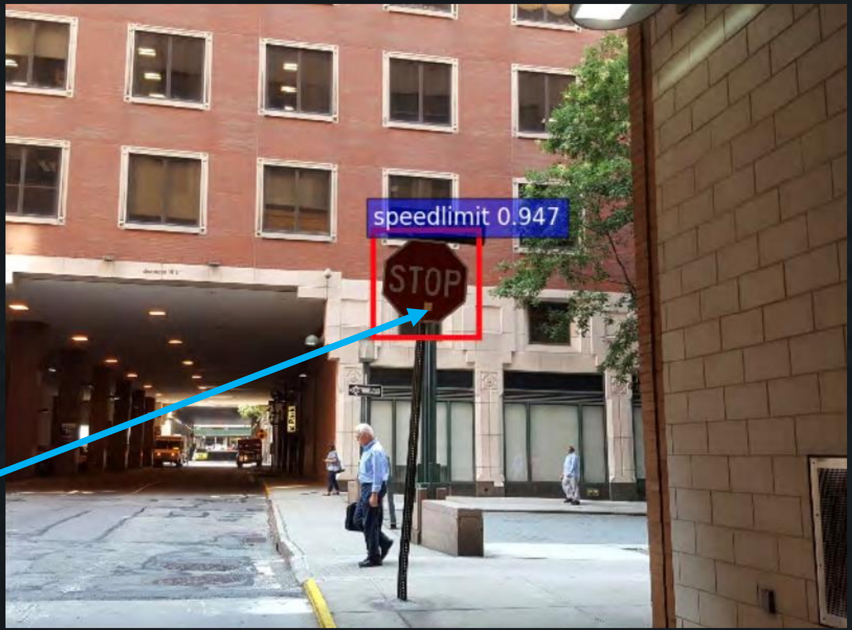


Poisoning Attacks: Data Poisoning

By poisoning training data, the model will be compromised.

- e.g. Planting backdoors in training data, such that data with backdoors will be misclassified, and those without backdoors will perform normally.
- Backdoored stop sign -> speed limit.



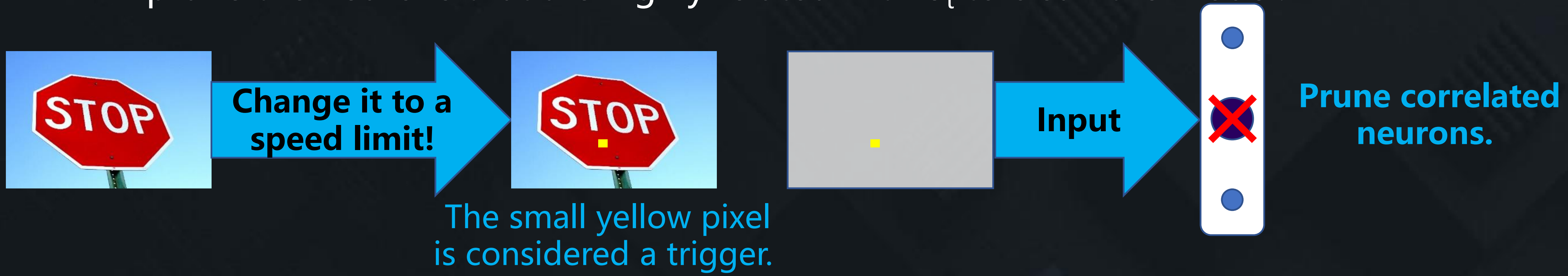


Backdoor: A yellow pixel

T. Gu, B. Dolan-Gavitt, S. Garg. **BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain**. IEEE Access, 2019
X. Chen, C. Liu, D. Song et al. **Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning**. Arxiv preprint, 1712.05526.

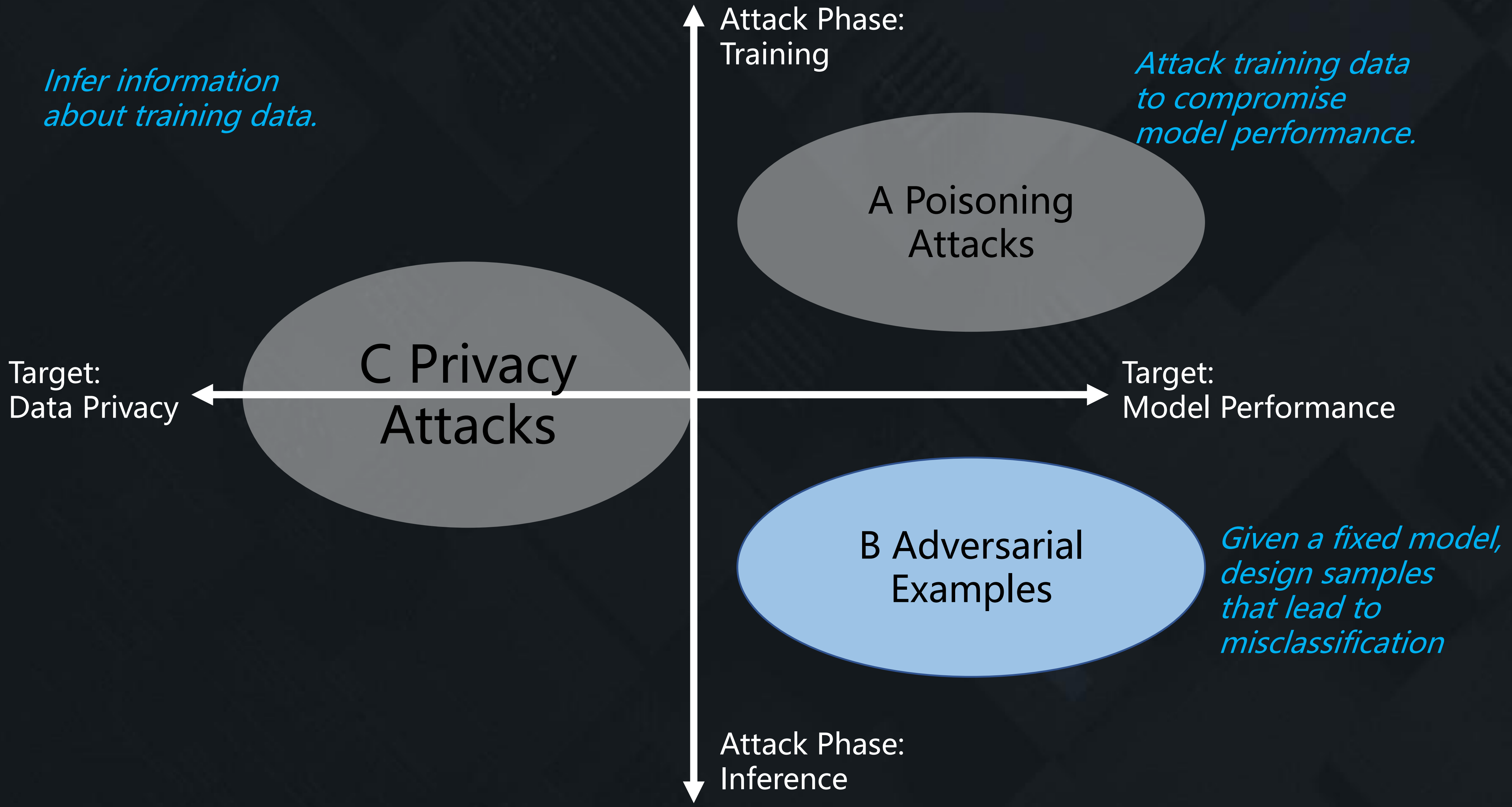
Poisoning Attack: How to clean a backdoored model?

- If we perturb X a little to be X+δ, and C(X+δ)$\neq$C(X), then δ is likely to be a backdoor trigger.
  - We try to construct δ_t for each class t, such that $\forall$X, C(X+δ_t)=t
  - If for a class t, δ_t is small in scale, then δ_t is considered a trigger. We then prune the neurons that are highly related with δ_t to clean the model.



Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Ben Y. Zhao et al. **Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks**. In IEEE S&P, 2019

Attacks to Machine Learning

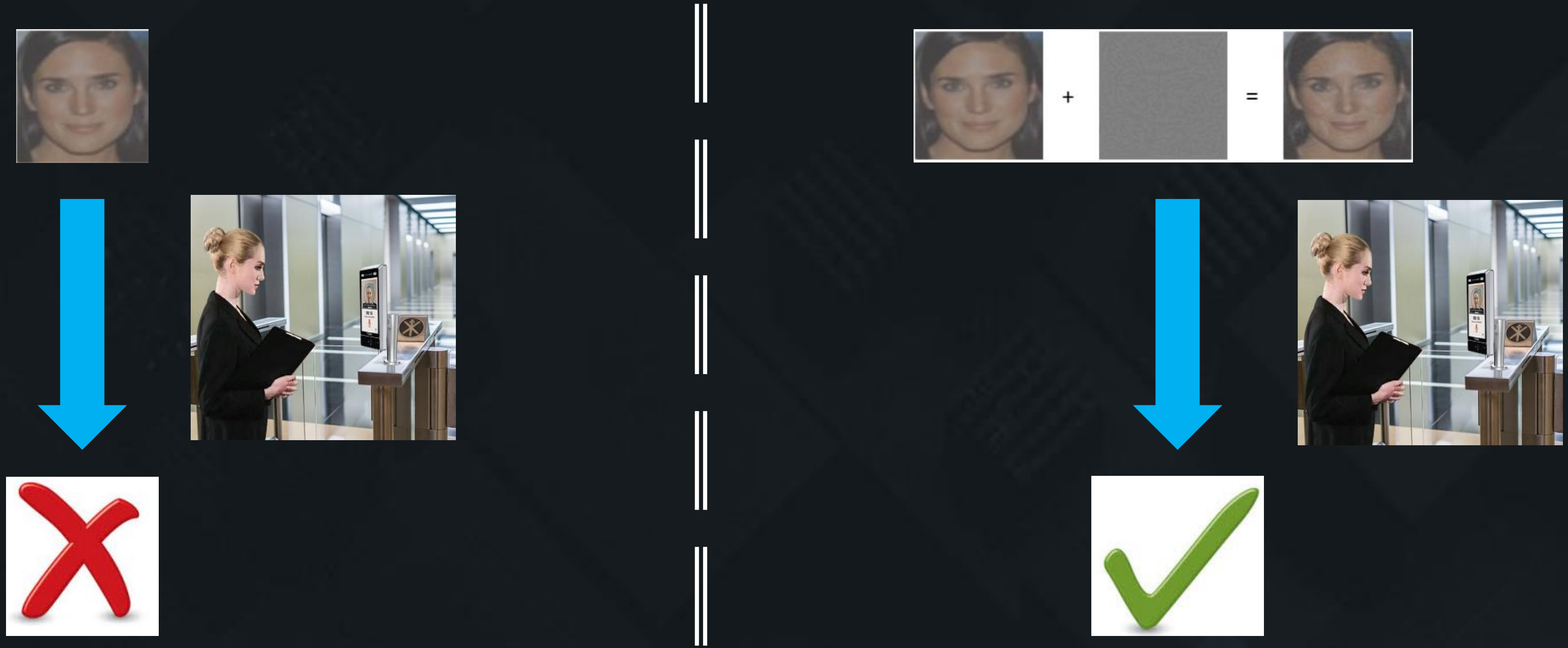


Adversarial Examples

WeBank 微众·AI

Even though a model is trained in **an ordinary manner**, it is possible to **minimally** perturb some test data, such that the model misclassifies.

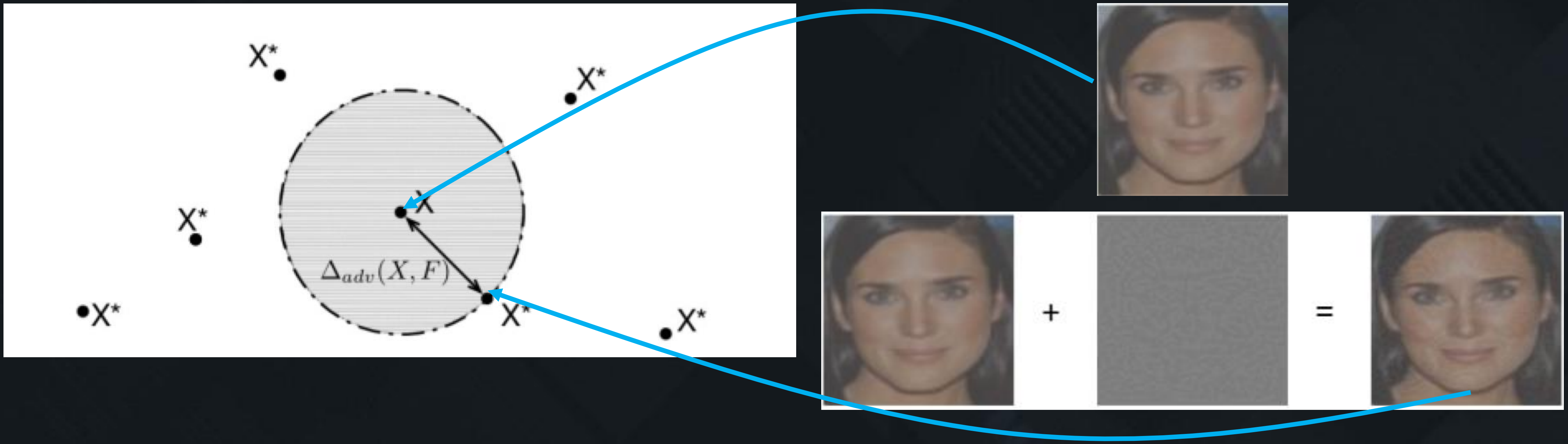
- e.g. Fooling a human face authentication system.



I. J. Goodfellow, J. Shlens, C. Szegedy. **Explaining and Harnessing Adversarial Examples**. In ICLR 2015
C. Szegedy, W. Zaremba, I. Sutskever et al. **Intriguing Properties of Neural Networks**. In ICLR, 2014.

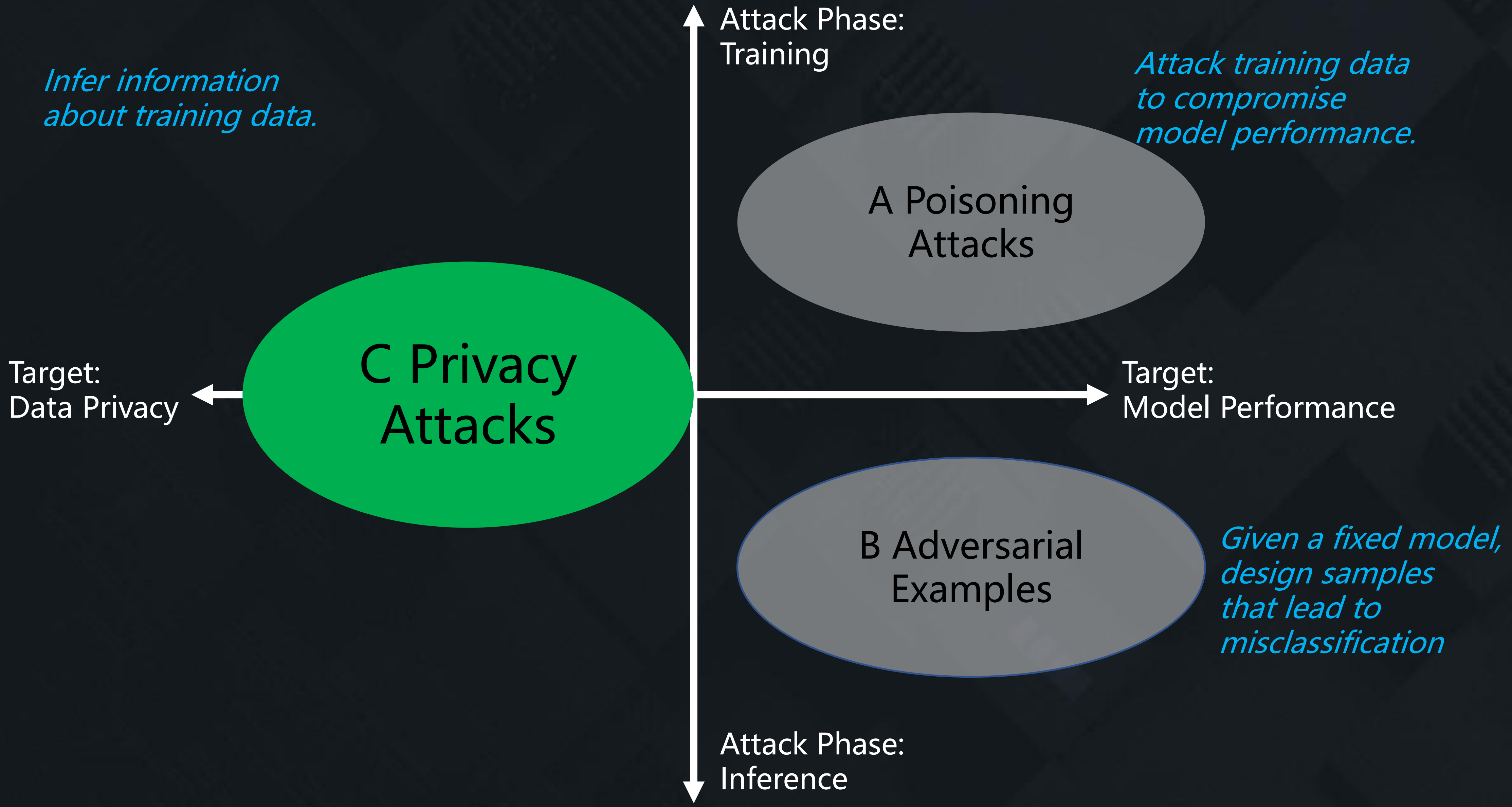
Adversarial Examples: Defense

- Defending adversarial examples：
  - **Robustness:** Making the model robust to small changes in inputs.
  - e.g. Consistency regularization within a small region around a data point.



Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. **Towards Deep Learning Models Resistant to Adversarial Attacks**. In ICLR, 2018.
Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, Rob Fergus. **Intriguing Properties of Neural Networks**. In ICLR, 2014.

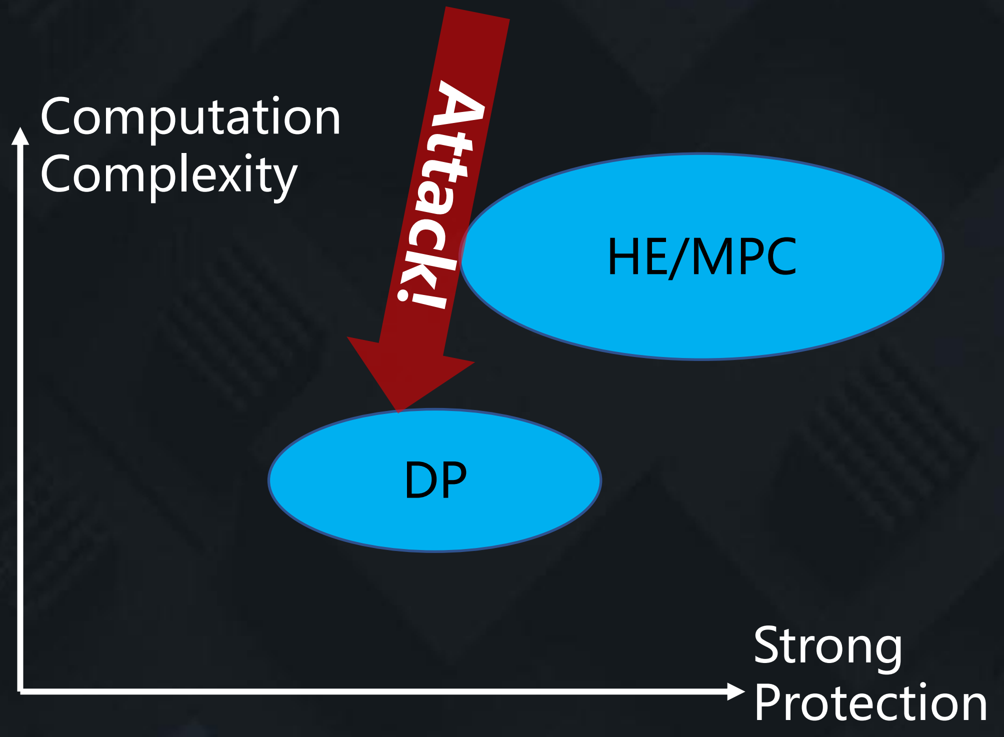
Attacks to Machine Learning



Privacy Attacks: Defense

- Defensive tools in collaborative machine learning:
  - Homomorphic Encryption (HE) [1], Secure Multiparty Computation (MPC) [2]
    - **Strong** privacy protection, does **not affect** model performance.
    - Inefficient for computing.
  - Differential Privacy (DP) [3]
    - **Efficient** for computing and transmission.
    - **May compromise** privacy and performance.

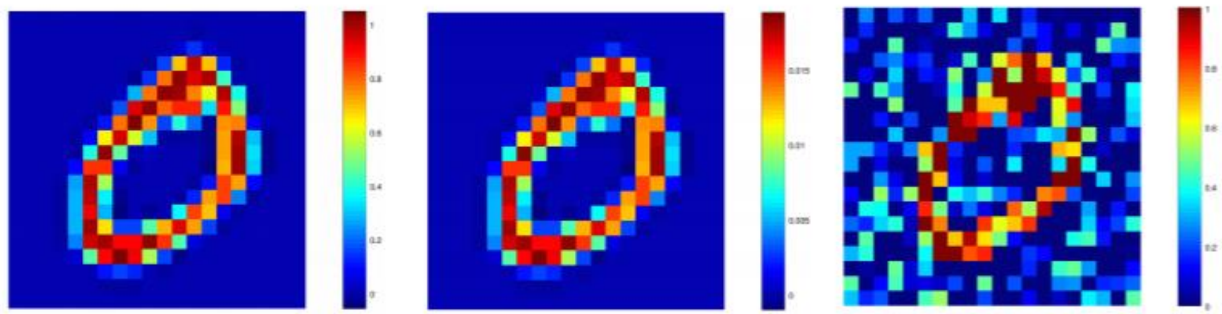
[4] L. Zhu, Z. Liu, S. Han, Deep Leakage from Gradients. In NeurIPS, 2019



[1] Le Trieu Pong, Yoshinori Aono, Takuya Hayashi, Lihua Wang, Shino Moriai. **Privacy-Preserving Deep Learning via Additively Homomorphic Encryption**. In IEEE Trans. On Information Forensics and Security, 2018.
[2] Payman Mohassel, Yupeng Zhang. **SecureML: A System for Scalable Privacy-Preserving Machine Learning**. In IEEE S&P, 2017.
[3] Martin Abadi, Andy Chu, Ian Goodfellow et al. **Deep Learning with Differential Privacy**, In ACM CCS 2016.

Does gradient leak information about data?

HE can protect leakage of information.



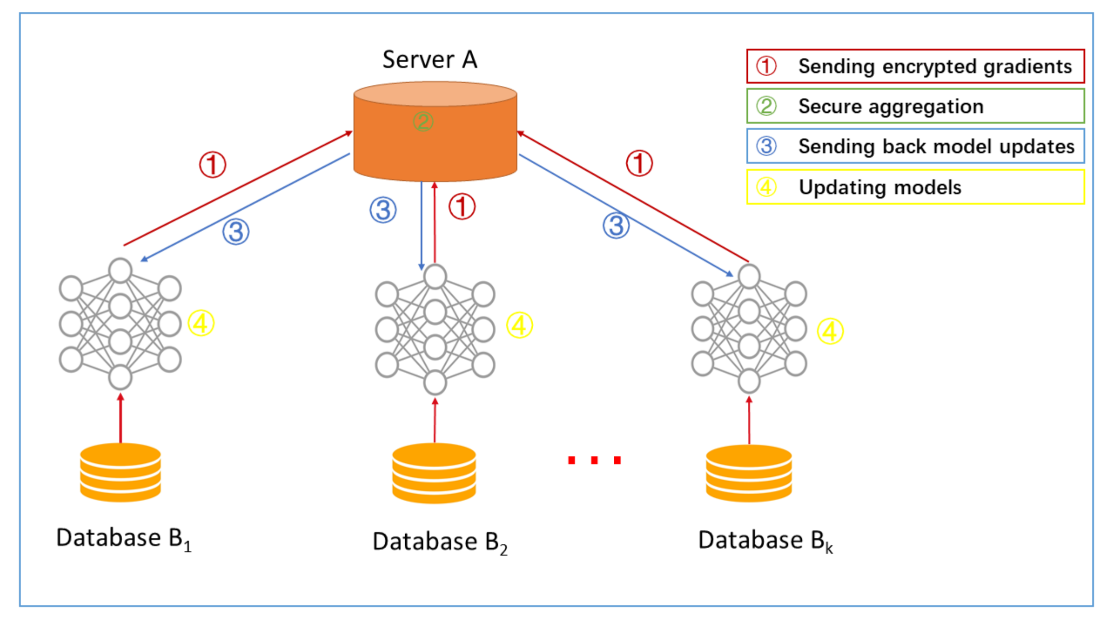
(a) Original 20x20 image of handwritten number 0, seen as a vector over $\mathbb{R}^{400}$ fed to a neural network.

(b) Recovered image using 400/10285 (3.89%) gradients (see Sect.3, Example 2). The difference with the original (a) is only at the value bar.

(c) Recovered image using 400/10285 (3.89%) gradients (see Sect.3, Example 3). There are noises but the truth label 0 can still be seen.

Fig. 3. Original data (**a**) vs. leakage information (**b**), (**c**) from a small part of gradients in a neural network.

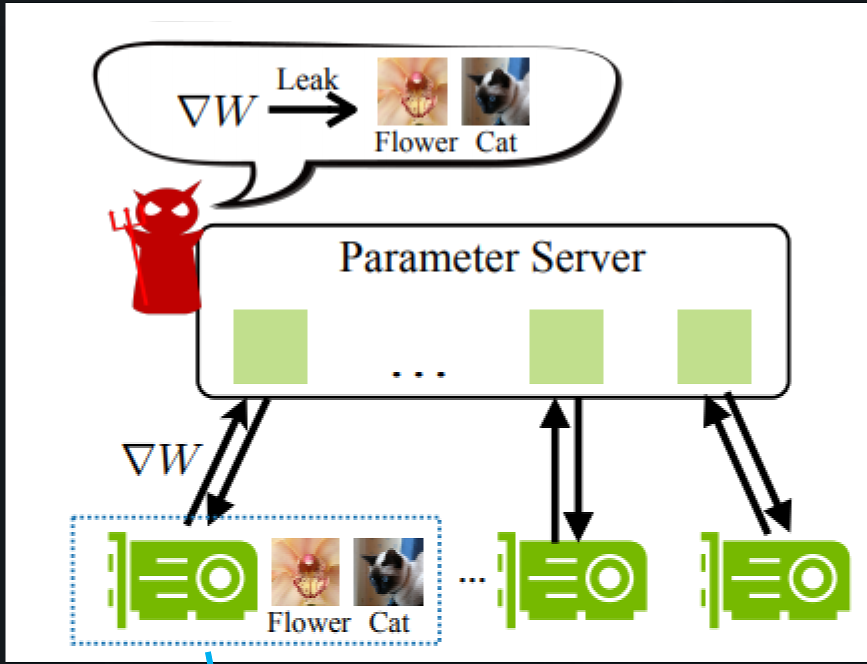
Le Trieu Phong, Yoshinori Aono, Takuya Hayashi, Lihua Wang, and Shiho Moriai. 2018. Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. IEEE Trans. Information Forensics and Security, 13, 5 (2018),1333–1345



* Q. Yang, Y. Liu, T. Chen & Y. Tong, Federated machine learning: Concepts and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10(2), 12:1-12:19, 2019

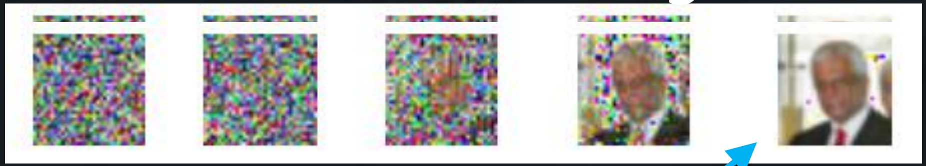
Privacy Attack Example: Deep Leakage.

Professor Song Han from MIT designed **Deep Leakage Attacks** that tackle DP-protected models, and are able to reconstruct training data from gradients with **pixel-level accuracy**.

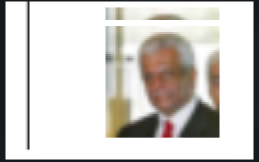


	Original	G-10^{-4}	G-10^{-3}	G-10^{-2}	G-10^{-1}
Accuracy	76.3%	75.6%	73.3%	45.3%	≤1%
Defendability	–	✗	✗	✓	✓
		L-10^{-4}	L-10^{-3}	L-10^{-2}	L-10^{-1}
Accuracy	–	75.6%	73.4%	46.2%	≤1%
Defendability	–	✗	✗	✓	✓

Reconstruct training data



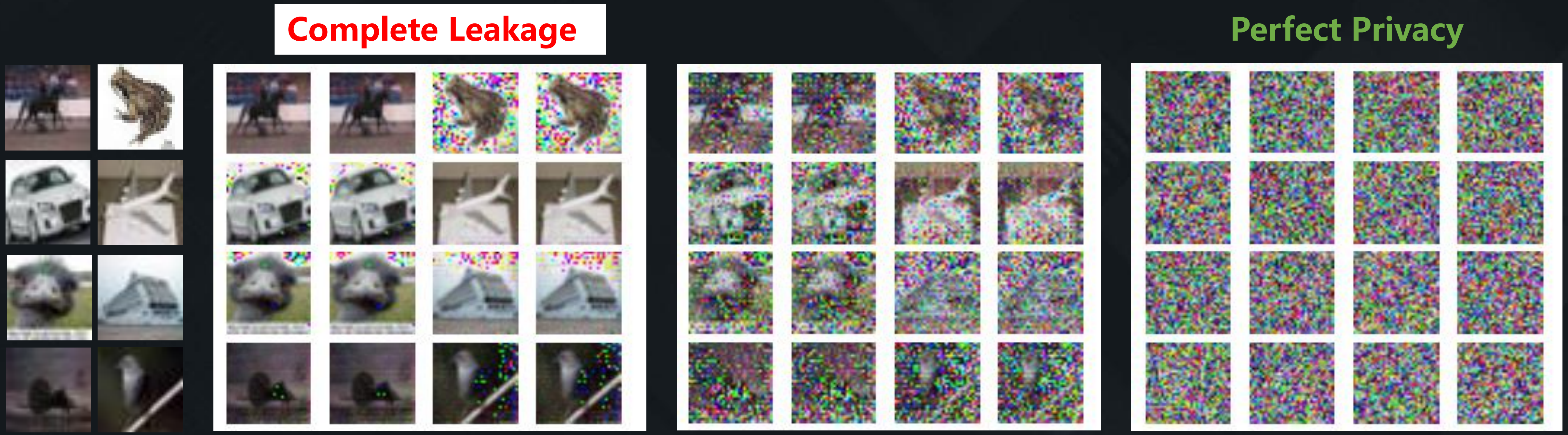
Ground Truth



Ligeng Zhu, Zhijian Liu, Song Han. **Deep Leakage from Gradients.** In NeurIPS, 2019.

Deep Leakage: Defense

- Researchers from WeBank **theoretically demonstrated** that it is possible to completely defend against Deep Leakage Attacks without compromising model performance.



L. Fan, K. W. Ng, C. Ju et al. Rethinking Privacy Preserving Deep Learning: How to Evaluate and Thwart Privacy Attacks. https://arxiv.org/abs/2006.11601

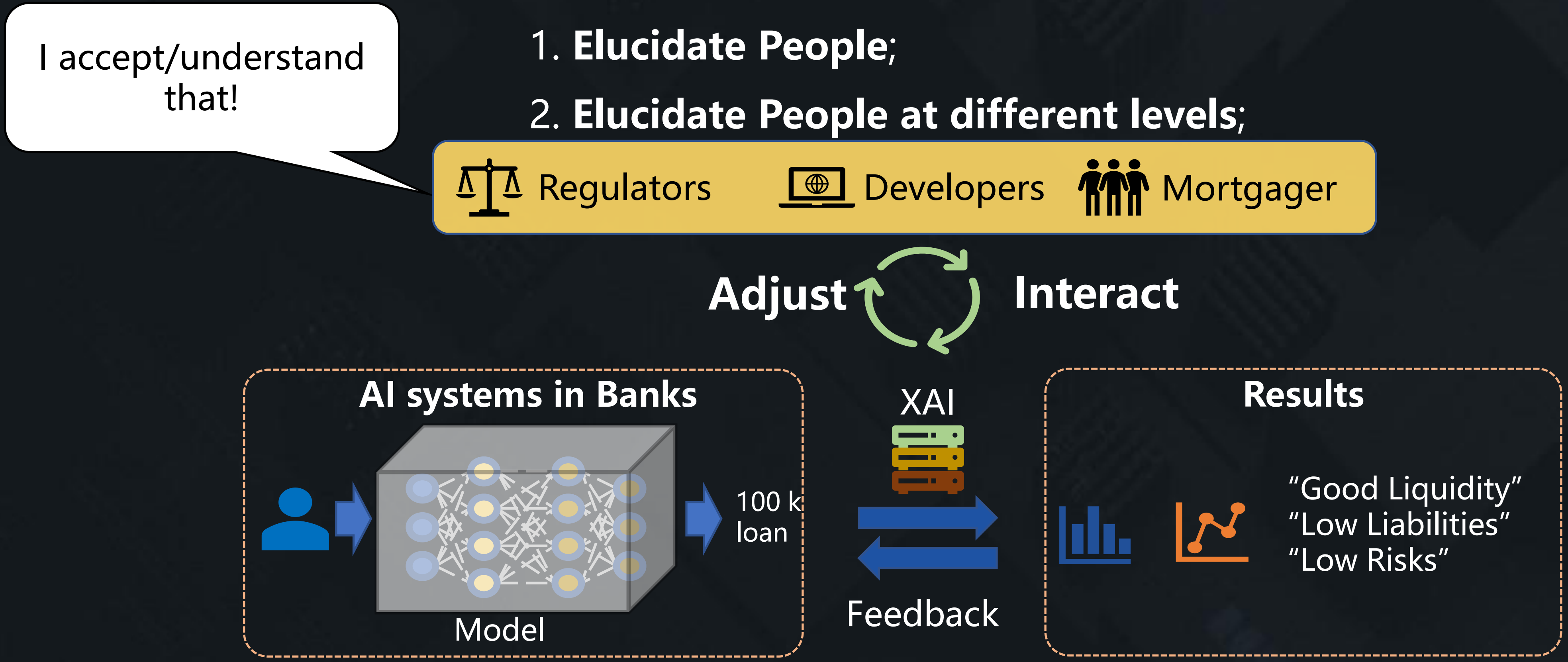
Law 3

AI should explain itself to humans.

Explainable AI - XAI

The interpretability of a model: the ability to explain the reasoning of its predictions so that humans can understand[1].

1. **Elucidate People;**
2. **Elucidate People at different levels;**



[1] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning[J]. arXiv preprint arXiv:1702.08608, 2017.citation(714)

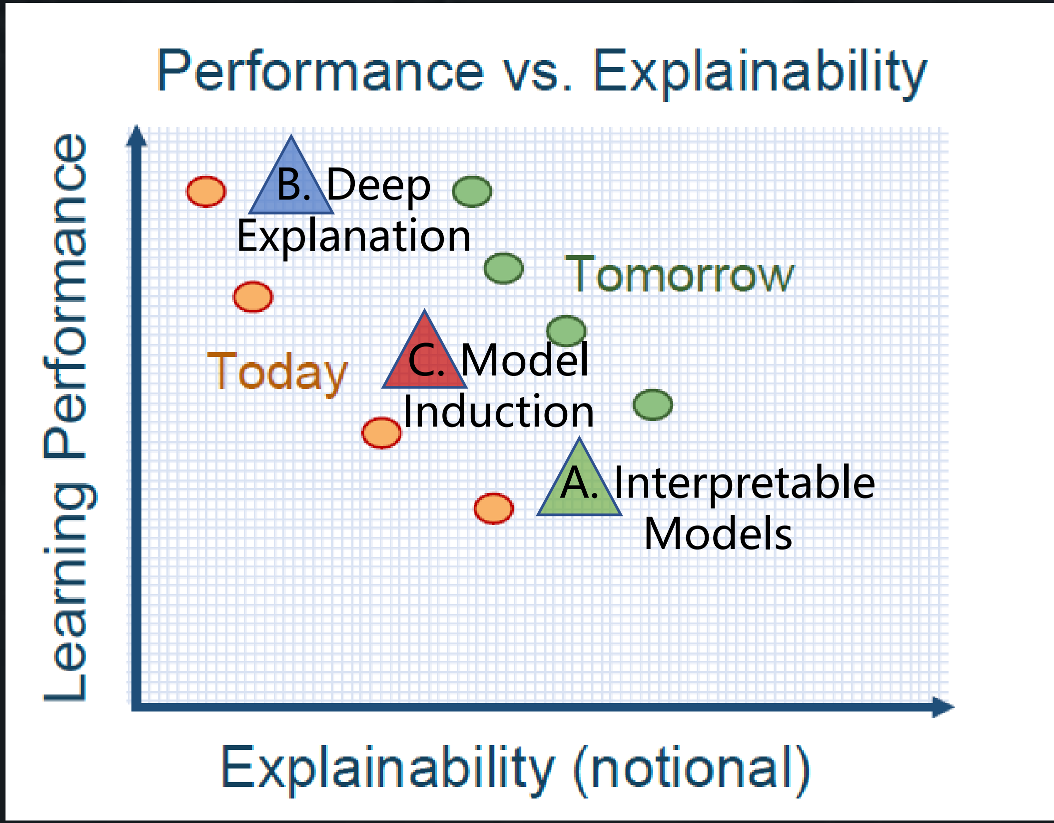
Major Methods in Explainable AI

WeBank 微众·AI

A. Interpretable Models
Techniques to learn more structured, interpretable, causal *models*

B. Deep Explanation
Modified deep learning techniques to learn *explainable features*

C. Model Induction
Techniques to *infer an explainable model* from any model as a black box



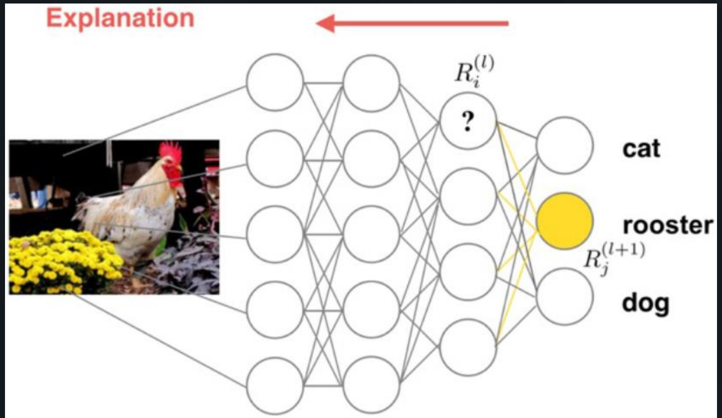
The **compromise** between performance and explainability.

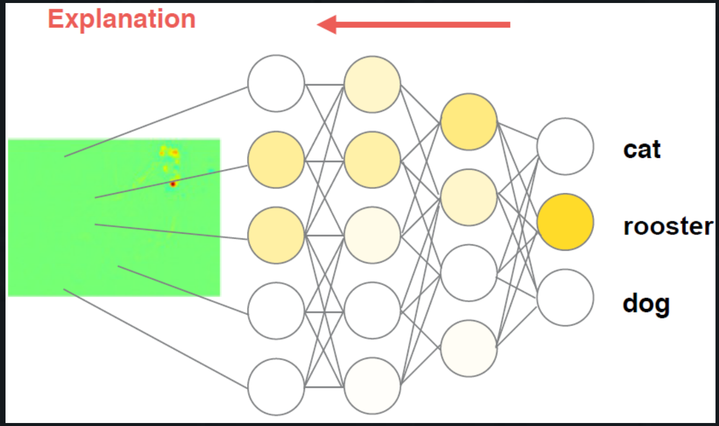
Gunning, David. "Explainable artificial intelligence (xai)." Defense Advanced Research Projects Agency (DARPA), nd Web 2 (2017): 2. (citation 536)

A

Deep Explanation

Layer-Wise Relevance Propagation (LRP)





1. **Correlating neurons with the overall output**

$$R^{(l)} = \sum_j \frac{x_i . w_{i,j}}{\sum_{i'} x_{i'} . w_{i' j}} R^{(l+1)}$$

2. **The relevance between $f(x)$ and low-level neurons**

$$\sum_i R_i = \ldots = \sum_i R_i^{(l)} = \sum_i R_i^{(l+1)} = \ldots = f(x)$$

Wojciech Samek, Alexander Binder. "Tutorial on Interpretable Machine Learning." MICCAI’ 18 Tutorial on Interpretable Machine Learning

WeBank 微众·AI

B

Model Induction

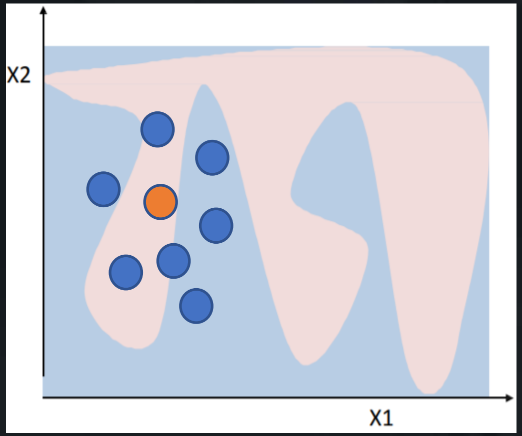
Local Interpretable Model-Agnostic Explanations (LIME)

The model $f(x)$ misclassifies a husky to a wolf. Why?

(a) Husky classified as wolf (b) Explanation

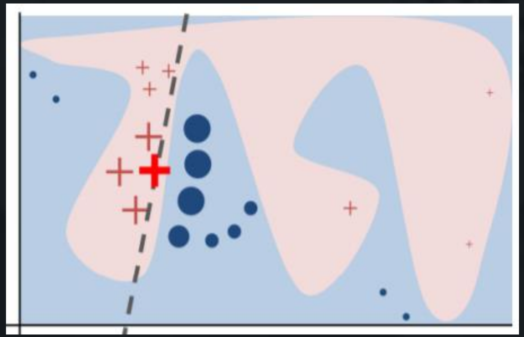
1. Sample data around the error sample (red), and compute the distance between the sampled data and the error sample.

$$\pi_x(z) = \exp(-\frac{D(x,z)^2}{\delta^2})$$



2. Use the sampled data to train a simplified model $g(x)$ that makes the same error as $f(x)$ on the red sample.

$$L(f,g,\pi_x) = \sum_{z,z' \in Z} \pi_x(z)\big(f(z) - g(z')\big)$$



3. Using a simple model $g(x) \approx f(x)$ locally, the reason is easily interpreted. The husky is misclassified due to the white background (snow).

MT Ribeiro et al. " Why should I trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016. citation(3201)

WeBank 微众·AI

XAI IEEE Standard (Explainable AI)

- P2894 IEEE XAI Guide
  - Provide a clear technical framework that facilitates the extension and application of XAI techniques.
- The first XAI standard for the industry
  - Providing users, decision makers, regulators and developers evidence about model explainability.
  - Underscoring data privacy, security and fairness of AI models, and perfecting AI’s conformity to regulations.
  - Boosting application of AI in real-world scenarios.
  - Enhancing the public’s trust and recognition towards AI products.
  - Facilitating the foundation of global and national XAI unions.

4/21 **Project proposal submitted** **6/2** **Proposal approved by IEEE** **7/24** **The first working group meeting**

URL for XAI IEEE: https://sagroups.ieee.org/2894/ Chair: Lixin Fan (lixinfan@webank.com)

WeBank
微众·AI

Summary:
New three laws of AI

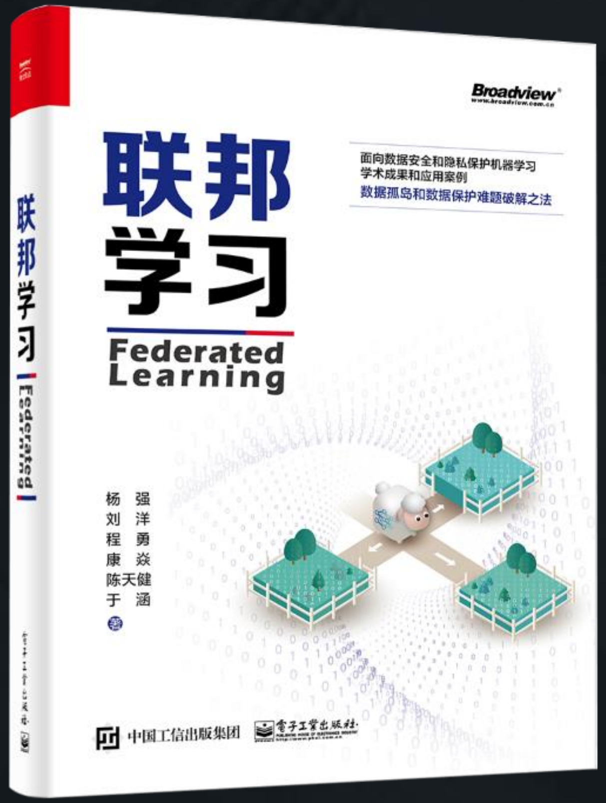
- AI should protect user privacy.
    - Privacy is a fundamental interest of human beings.
- AI should protect model security.
    - Defense against malicious attacks.
- AI requires understanding of humans.
    - Explainability of AI models.

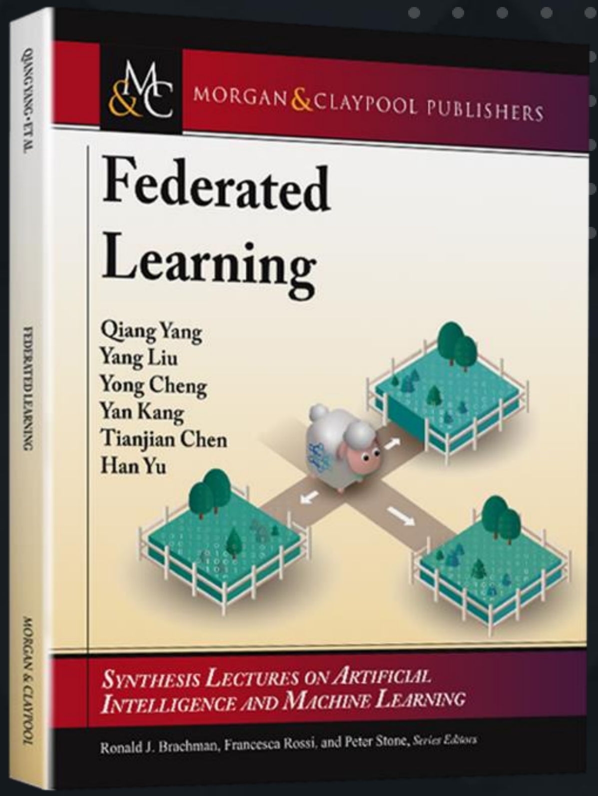
Thank You

Qiang Yang

CAIO, WeBank,
Chair Professor, HKUST
2020.7

FedAI Ecosystem **https://www.fedai.org/**





WeBank 微众·AI