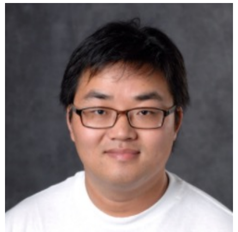
Introduction



Jiliang Tang

Privacy

Xiaorui Liu

Safety & Robustness

Yaxin Li

Explainability

Wenqi Fan

Non-discrimination & Fairness

Environmental Well- being

Haochen Liu

Accountability & Auditability

Dimension Interactions

Future Directions

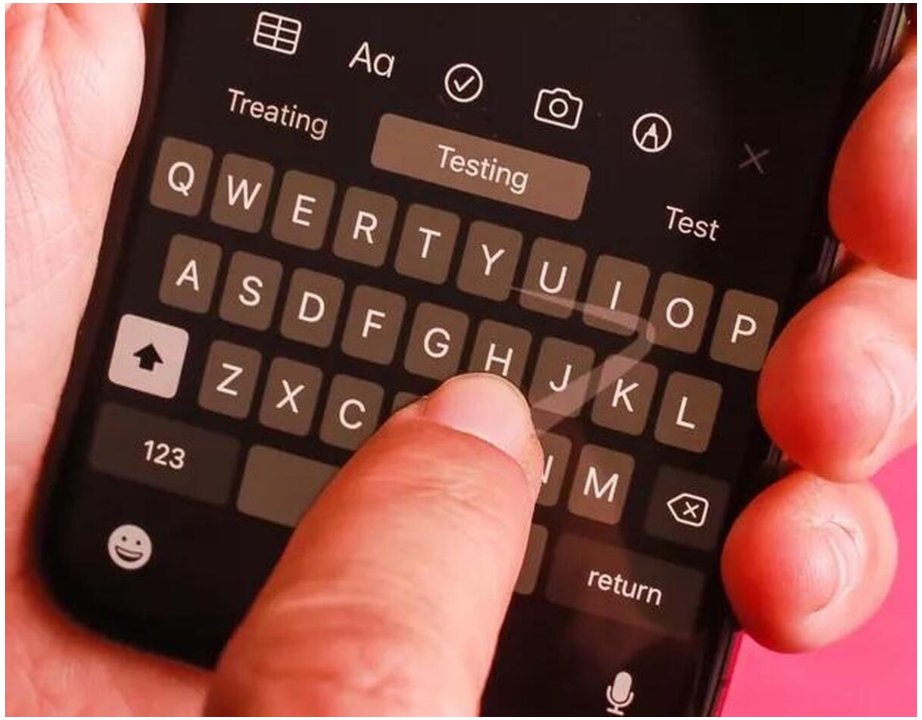
Yiqi Wang

Identity Authentication

Face Verification

Fingerprint Verification

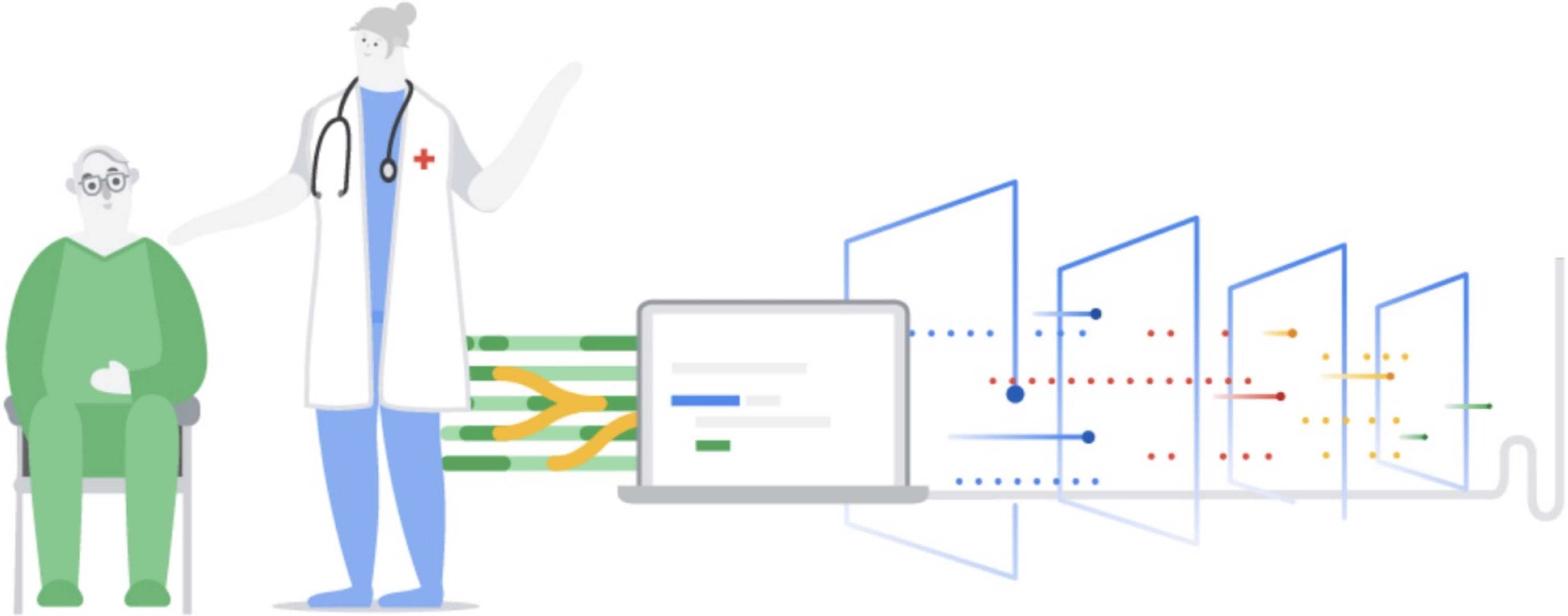
User interaction



Smartphone

Laptop

Medical record



Medical electronic patient record system

Privacy in AI

- The success of AI systems heavily relies on data that might contain private and sensitive information.

- Can we still take the advantages of data while effectively protecting the privacy?

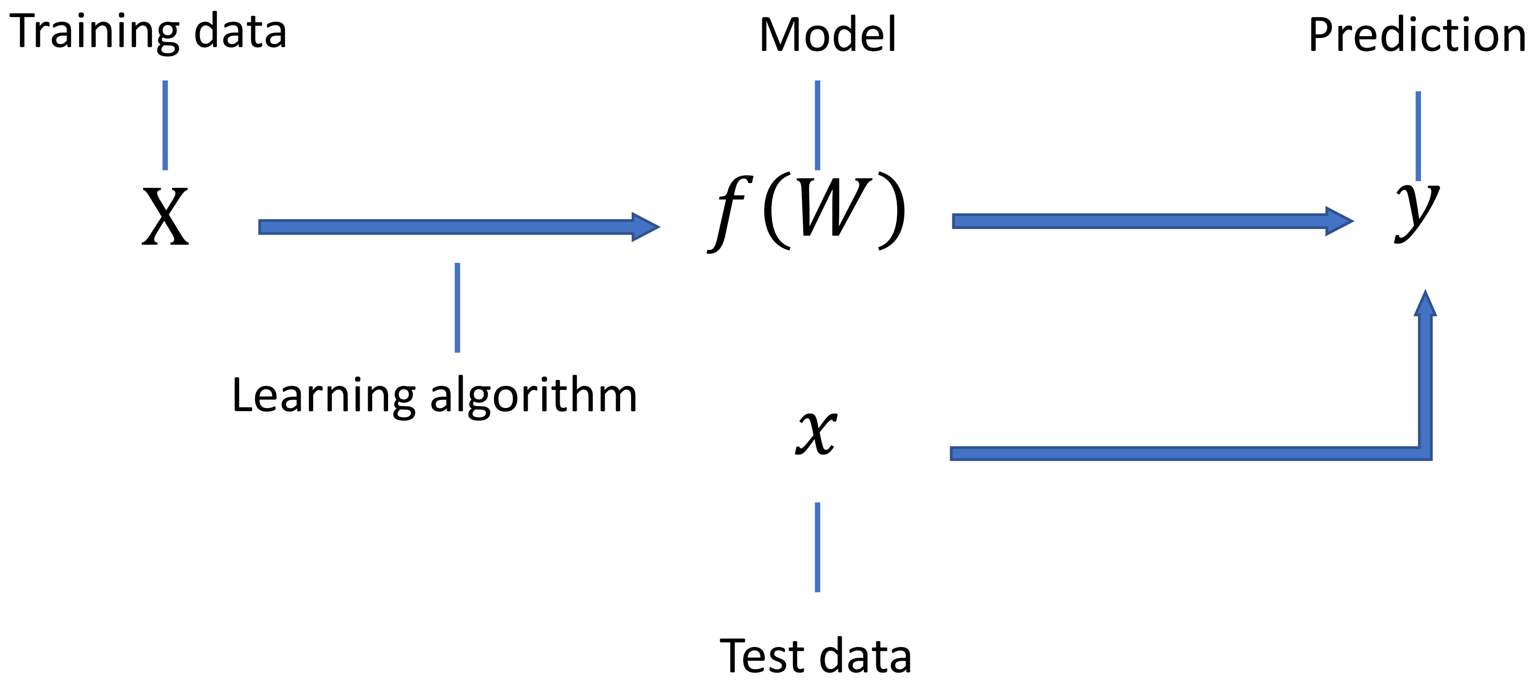
Concepts and Taxonomy → Privacy Leakage in AI Systems → Protection Strategies

Applications → Surveys and Tools → Future Directions

Concepts and Taxonomy → Privacy Leakage in AI Systems → Protection Strategies

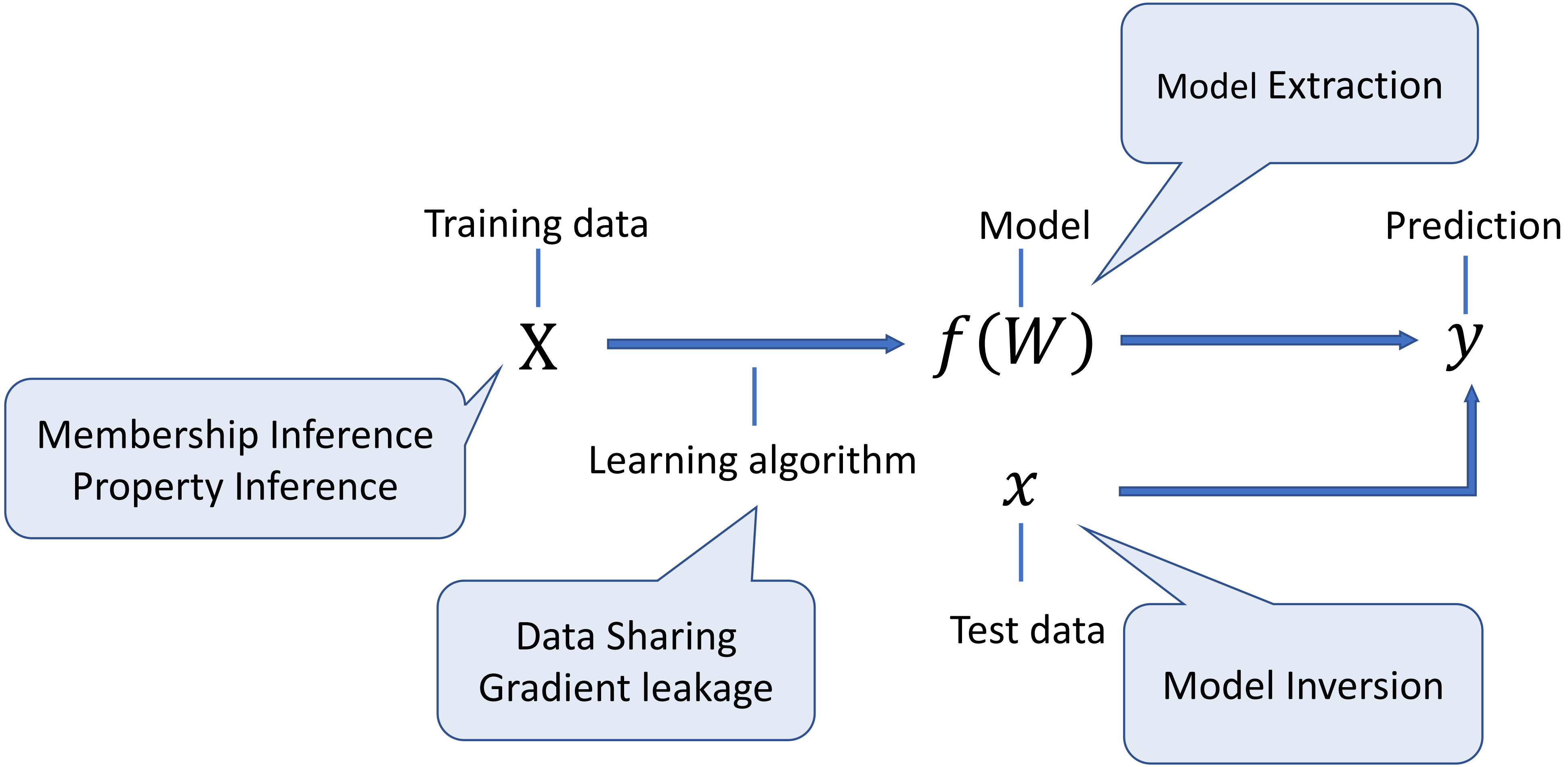
Applications → Surveys and Tools → Future Directions

Machine Learning in AI



A typical pipeline

Privacy Leakage in AI



Taxonomy

- ❑ Data & model
- ❑ Black-box & white-box setting
- ❑ Training & test phase
- ❑ Honest-but-curious & fully malicious

Concepts and Taxonomy → **Privacy Leakage in AI Systems** → Protection Strategies

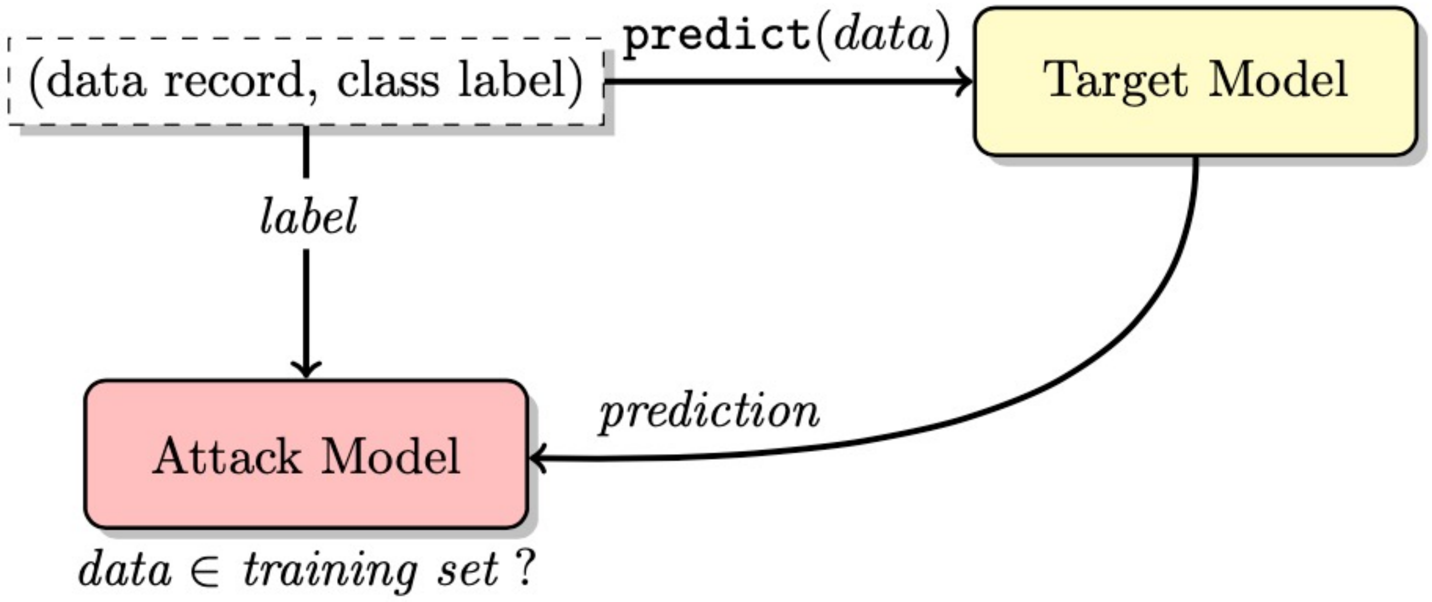
Applications → Surveys and Tools → Future Directions

Privacy Leakage in AI

- Membership Inference
- Data Sharing
- Gradient Leakage
- Model Inversion
- Model Extraction
- ...

Membership Inference

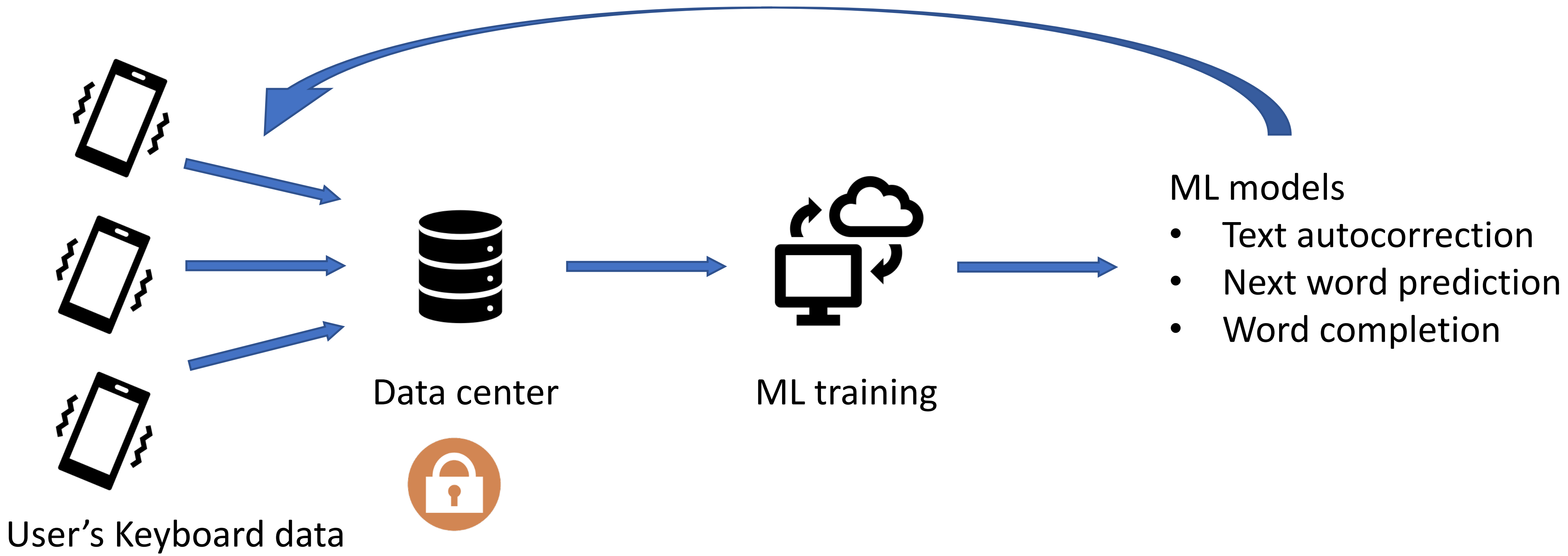
To identify whether a data record is used in the training of model



Shokri, Reza, et al. ”Membership inference attacks against machine learning models.” 2017.

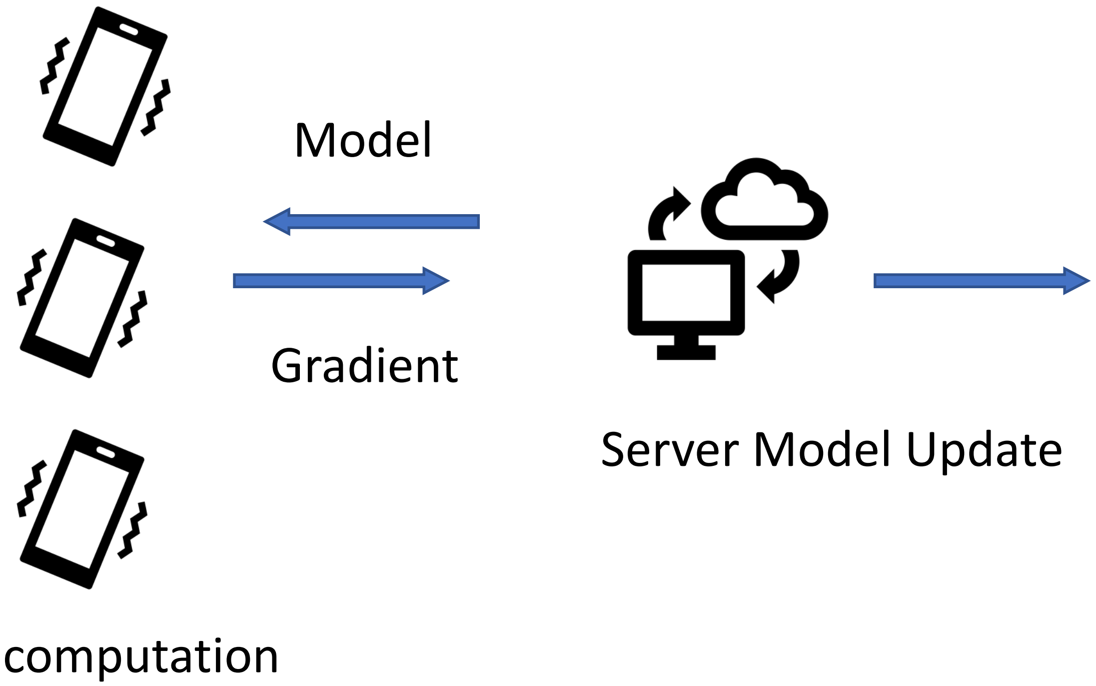
Data Sharing

User's data are collected and shared in the data center to train AI systems



Gradient Leakage

Distributed learning over mobile devices by synchronizing/sharing gradients

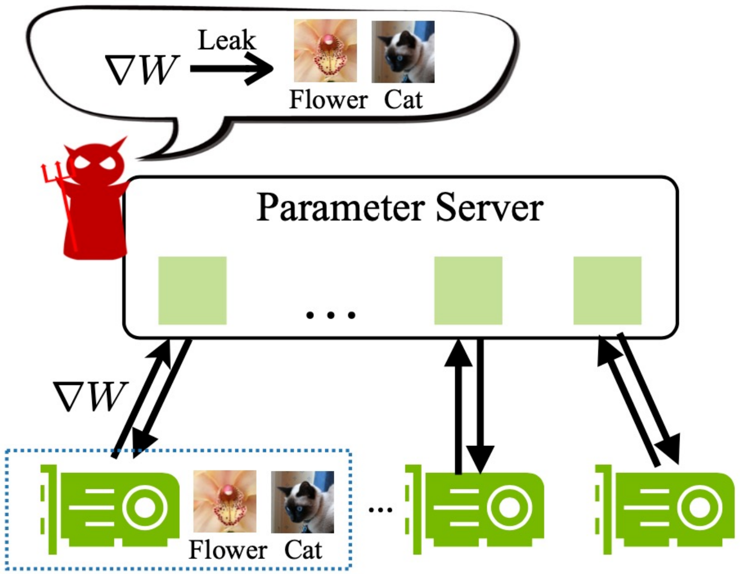


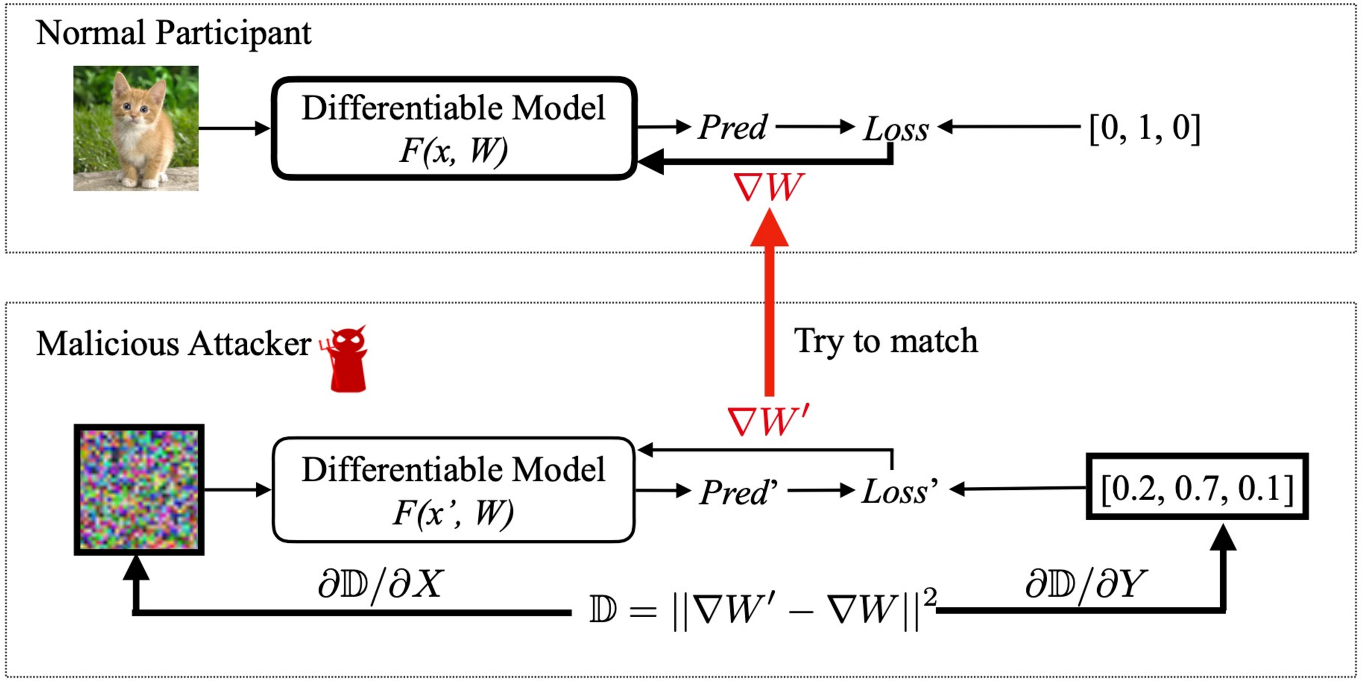
ML models

- Text autocorrection
- Next word prediction
- Word completion

Gradient Leakage (Continued)

Steal training data from the gradient information in distributed learning

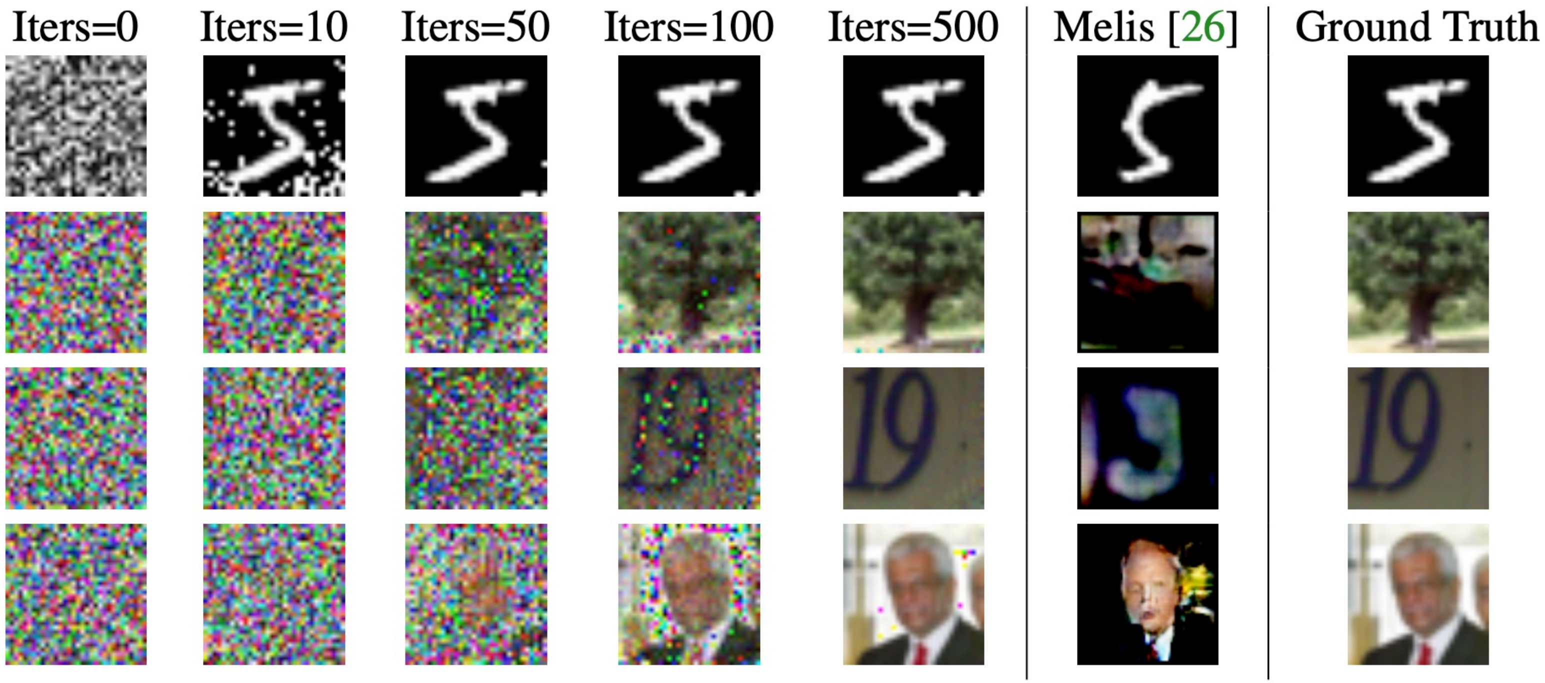




Zhu, Ligeng, and Song Han. "Deep leakage from gradients." 2020.

Gradient Leakage (Continued)

Steal training data from the gradient information in distributed learning



Zhu, Ligeng, and Song Han. "Deep leakage from gradients." 2020.

Model Inversion

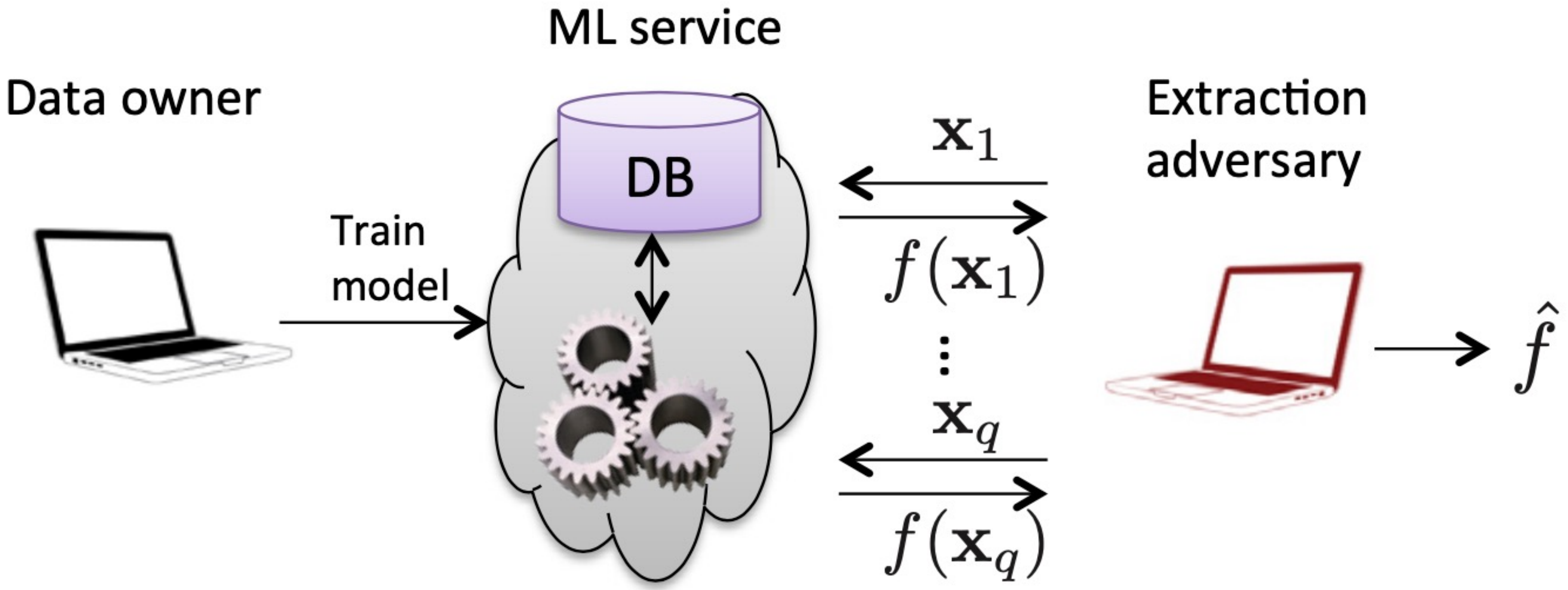
To infer the information of the input data using the model's output

Recover the face image given the person's name and
the class confidence of a facial recognition system

Fredrikson, Matt, et al. "Model inversion attacks that exploit confidence information and basic countermeasures." 2015.

Model Extraction

To extract the model information by querying the model in a black-box setting



Tramèr, Florian, et al. "Stealing machine learning models via prediction apis." 2016.

Concepts and Taxonomy → Privacy Leakage in AI Systems → **Protection Strategies**

Applications → Surveys and Tools → Future Directions

Privacy Preservation in AI

- Differential Privacy
- Federated Learning
- Confidential Computing

Differential Privacy

- It aims to reduce the disclosure about individual information in a dataset

- A randomized algorithm A is $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$**-differentially private** if for all $S \in \mathrm{Range}(A)$ and for all adjacent datasets D and D' such that

$$\Pr[\mathcal{A}(D) \in \mathcal{S}] \leq e^{\epsilon} \Pr(\mathcal{A}(D') \in \mathcal{S}) + \delta$$

- If $(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$ are sufficiently small, the output of the algorithm A will be almost identical

$$\Pr[\mathcal{A}(D) \in \mathcal{S}] \approx \Pr(\mathcal{A}(D') \in \mathcal{S})$$

Differential Privacy

- ❑ Random response
- ❑ Gaussian mechanism
- ❑ Laplace mechanism
- ❑ Exponential mechanism

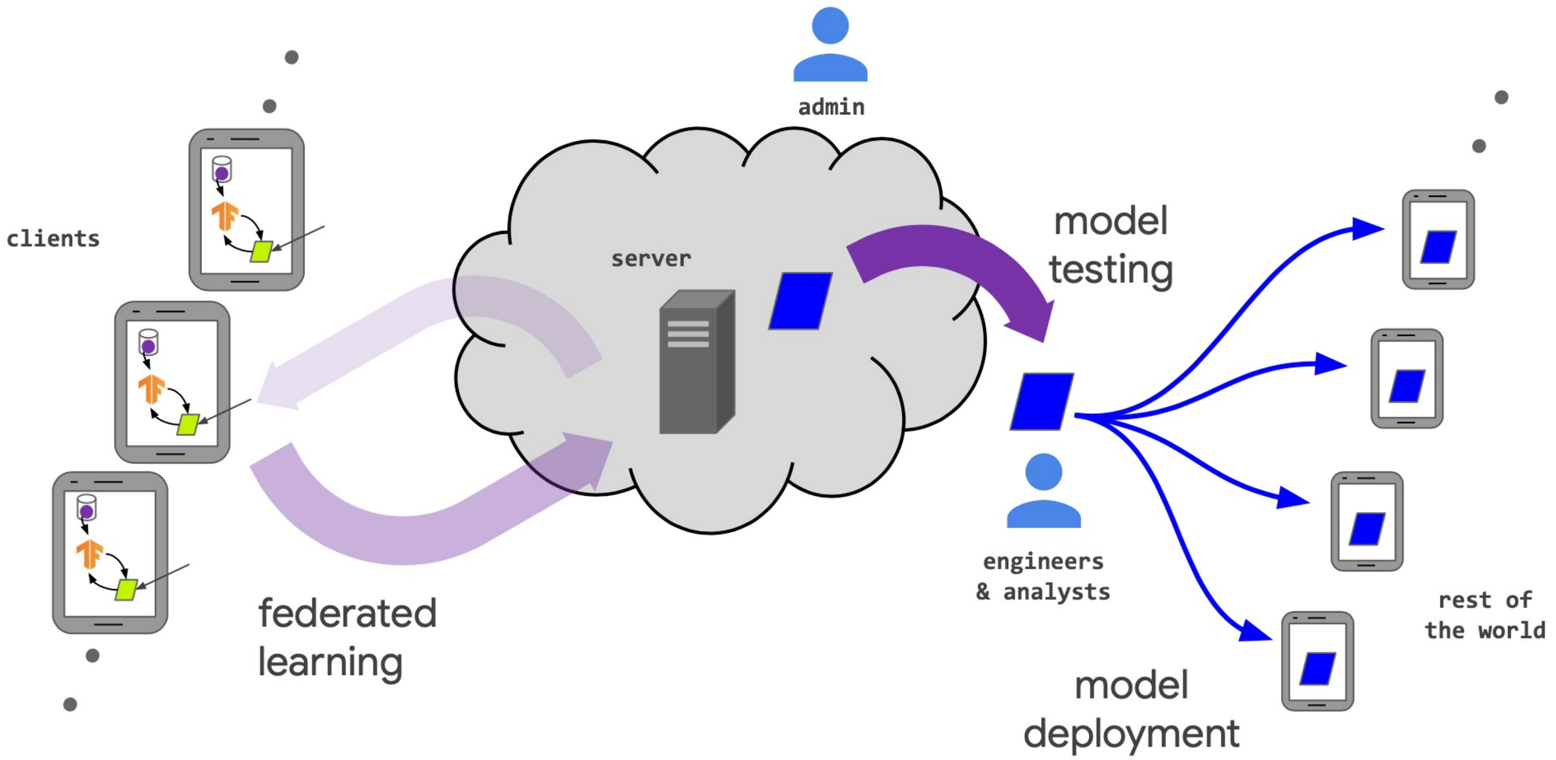
→

- Flip a coin
- If tails, then respond truthfully.
- If heads, then flip a second coin and respond “Yes” if heads and ”No” if tails

Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." 2014.

Federated Learning

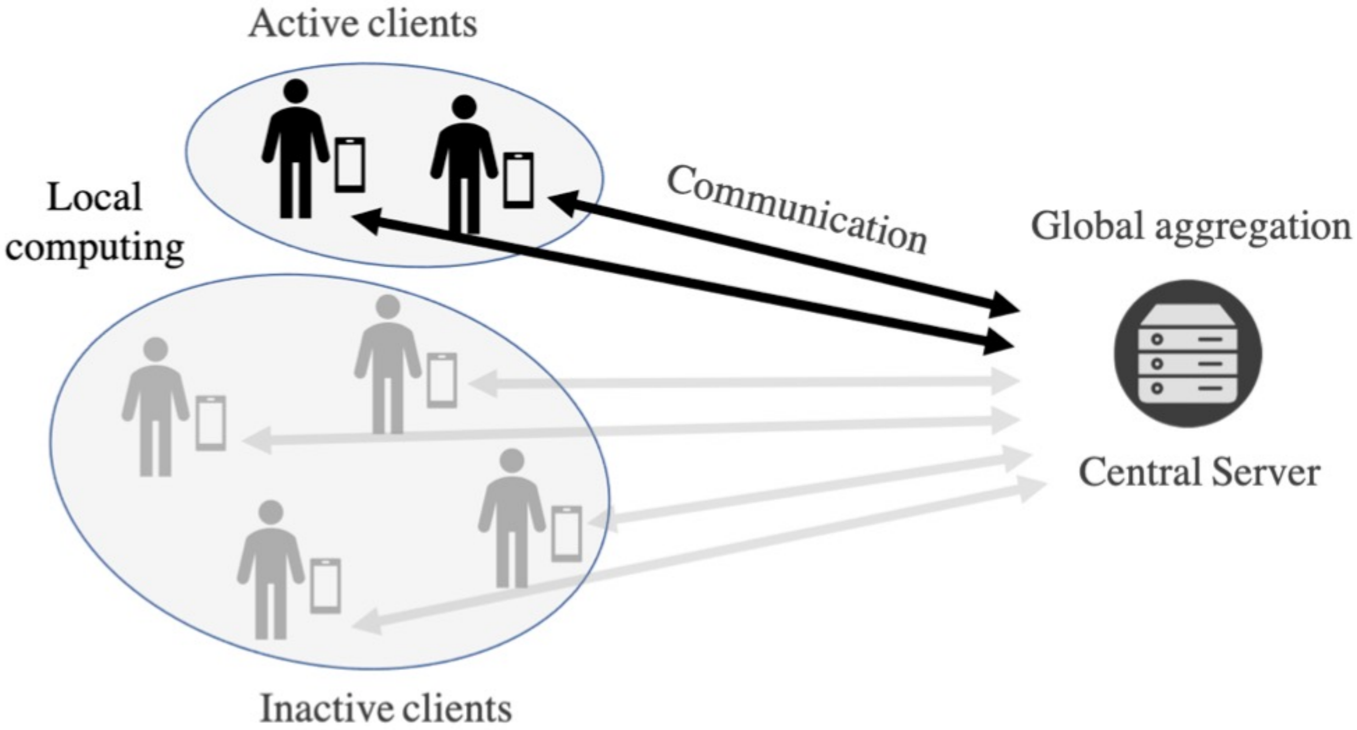
Clients collaboratively train a model while keeping the data decentralized



Kairouz, Peter, et al. "Advances and open problems in federated learning." (2019).

Workflow of Federated Learning

- ❑ Step 1: Client selection
- ❑ Step 2: Broadcast
- ❑ Step 3: Local computation
- ❑ Step 4: Aggregation

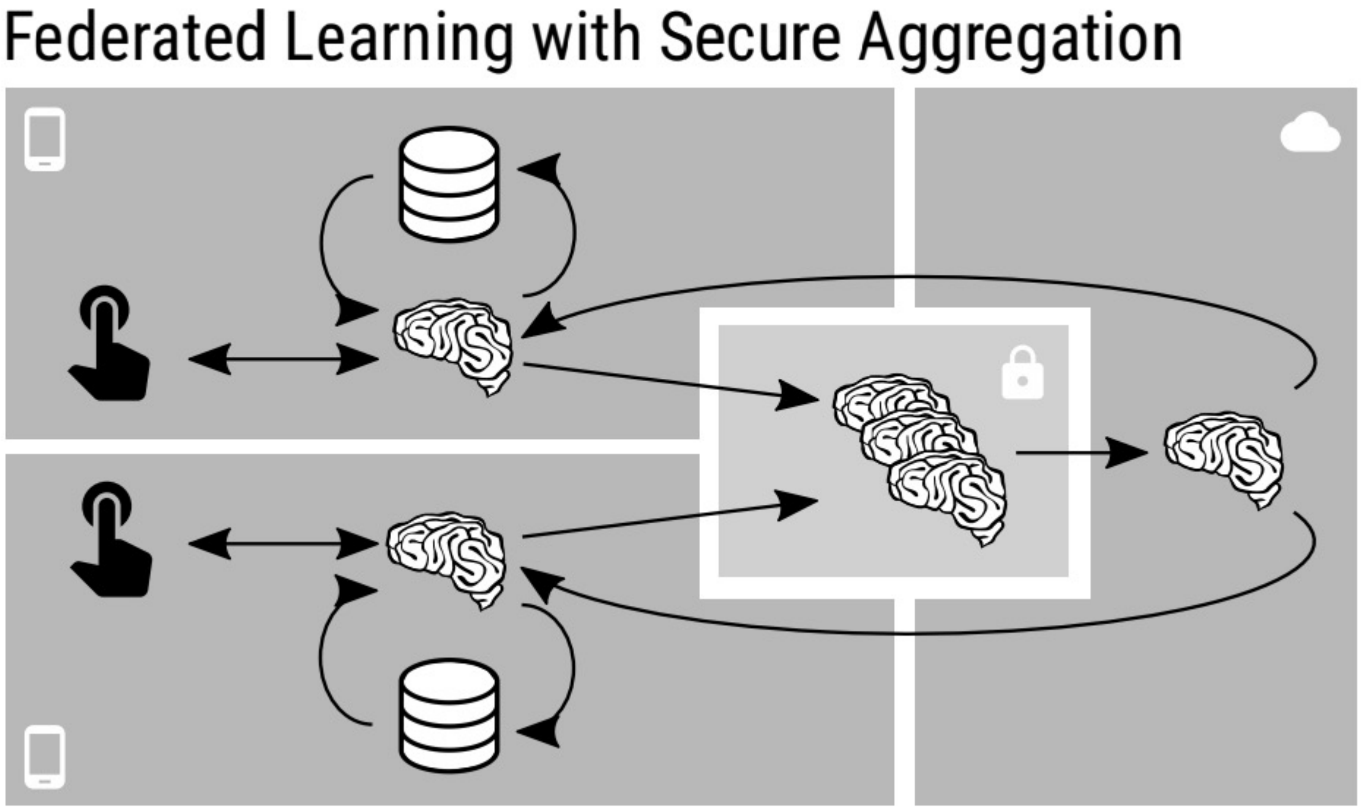


McMahan, Brendan, et al. "Communication-efficient learning of deep networks from decentralized data." 2017.

Confidential Computing

- Trusted Execution Environment (TEE)
  - Isolating data and programs by software and hardware techniques
- Homomorphic Encryption
  - Computing functions on ciphertext without decryption
- Secure Multi-party Computation (MPC)
  - Jointly performing function computations on private data

Secure Multi-party Computation



Bonawitz, Keith, et al. "Practical secure aggregation for federated learning on user-held data." *2*016.
Bonawitz, Keith, et al. "Practical secure aggregation for privacy-preserving machine learning." 2017.

Concepts and Taxonomy → Privacy Leakage in AI Systems → Protection Strategies

Applications → Surveys and Tools → Future Directions

Deep Learning

Differentially private SGD

Add Gaussian noise into gradient

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

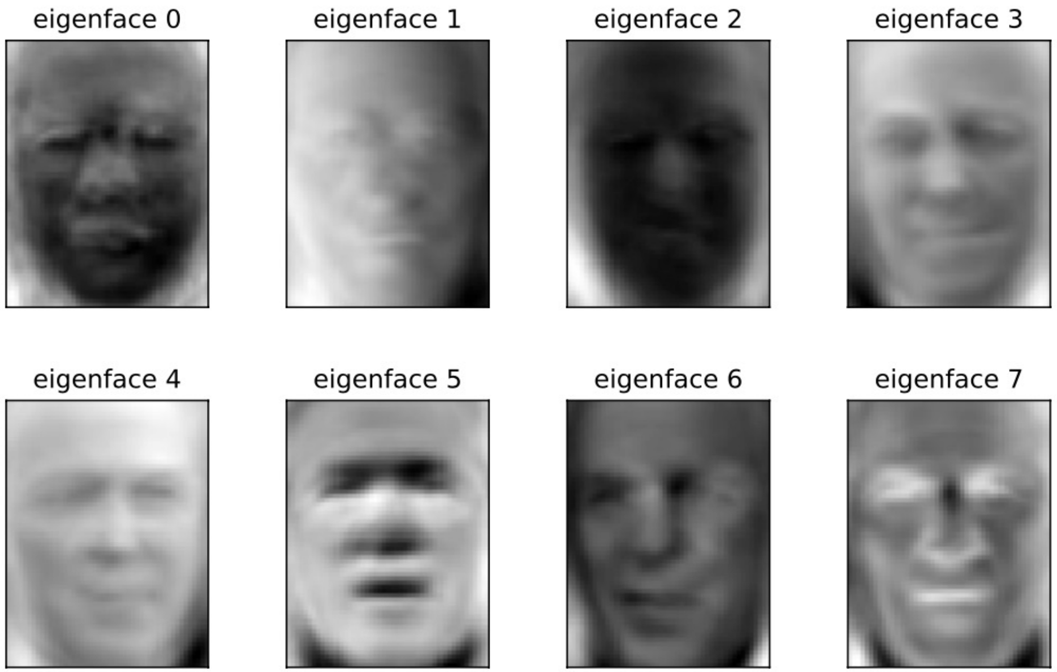
Descent

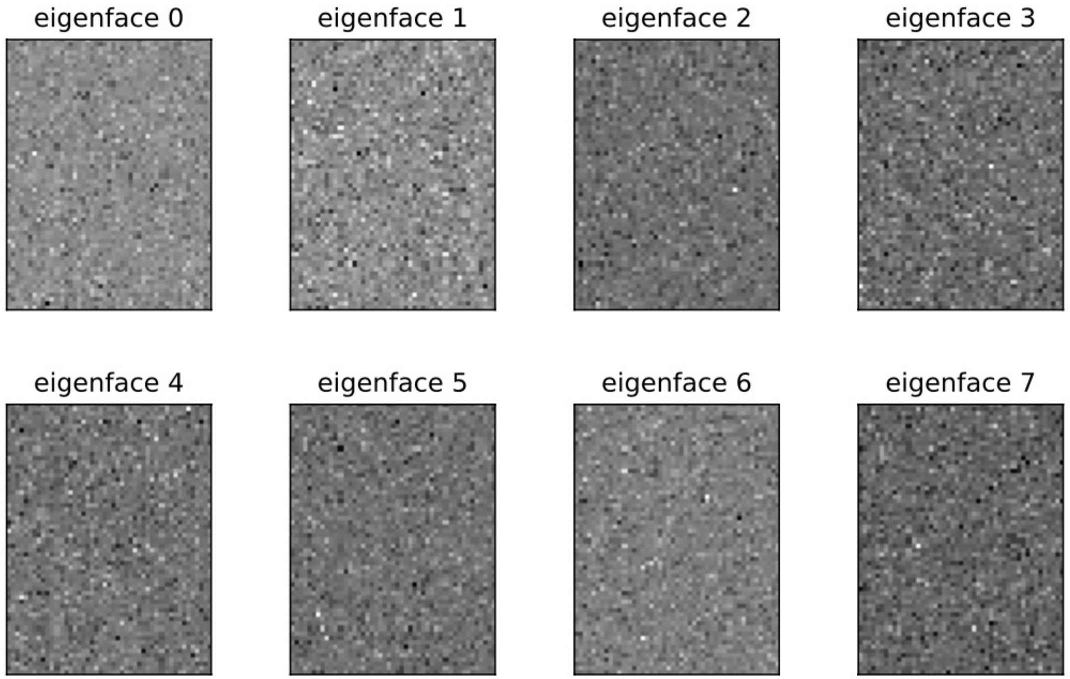
$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Abadi, Martin, et al. “Deep learning with differential privacy.” 2016

Biometric Data Analysis

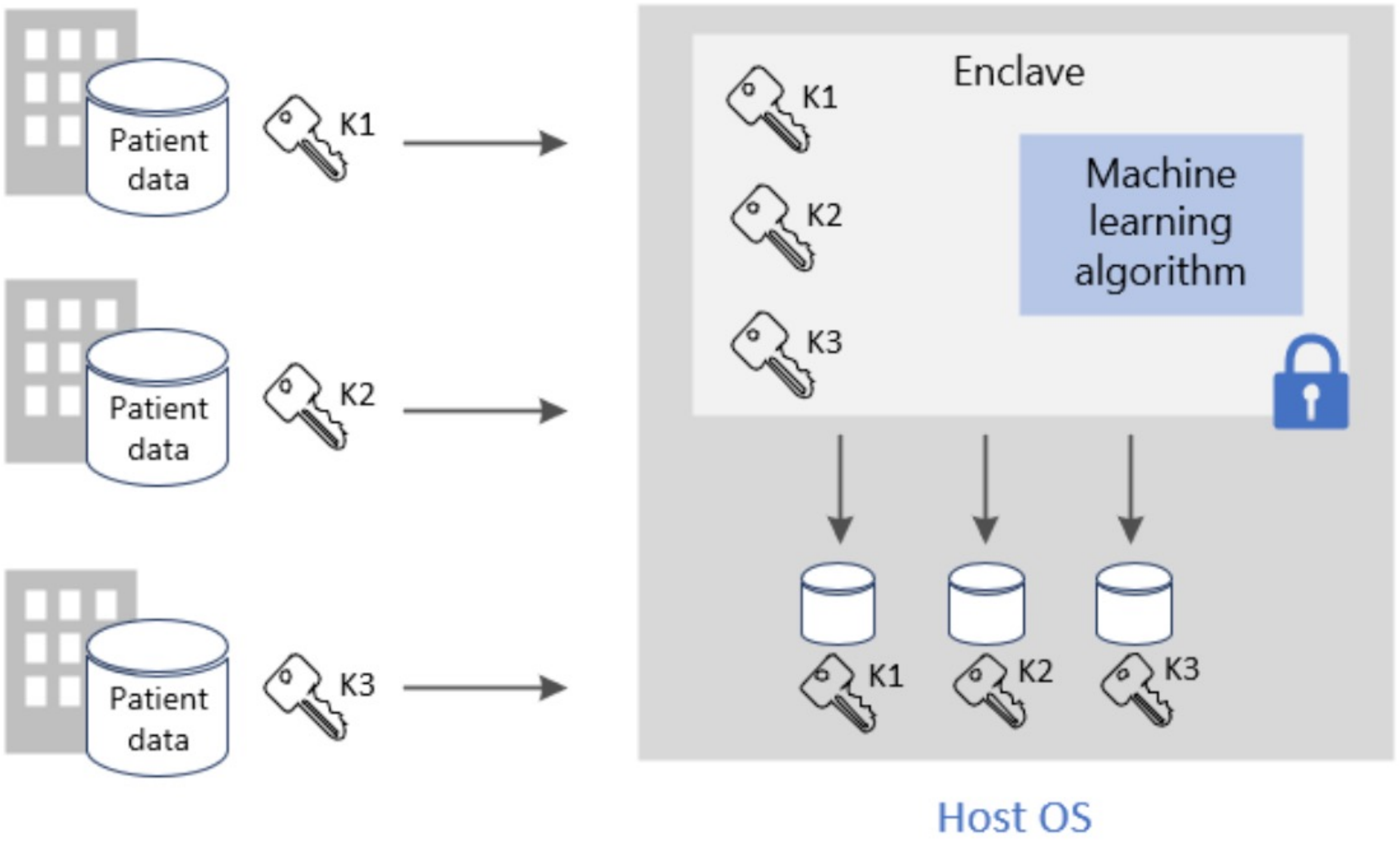




Differentially private facial recognition

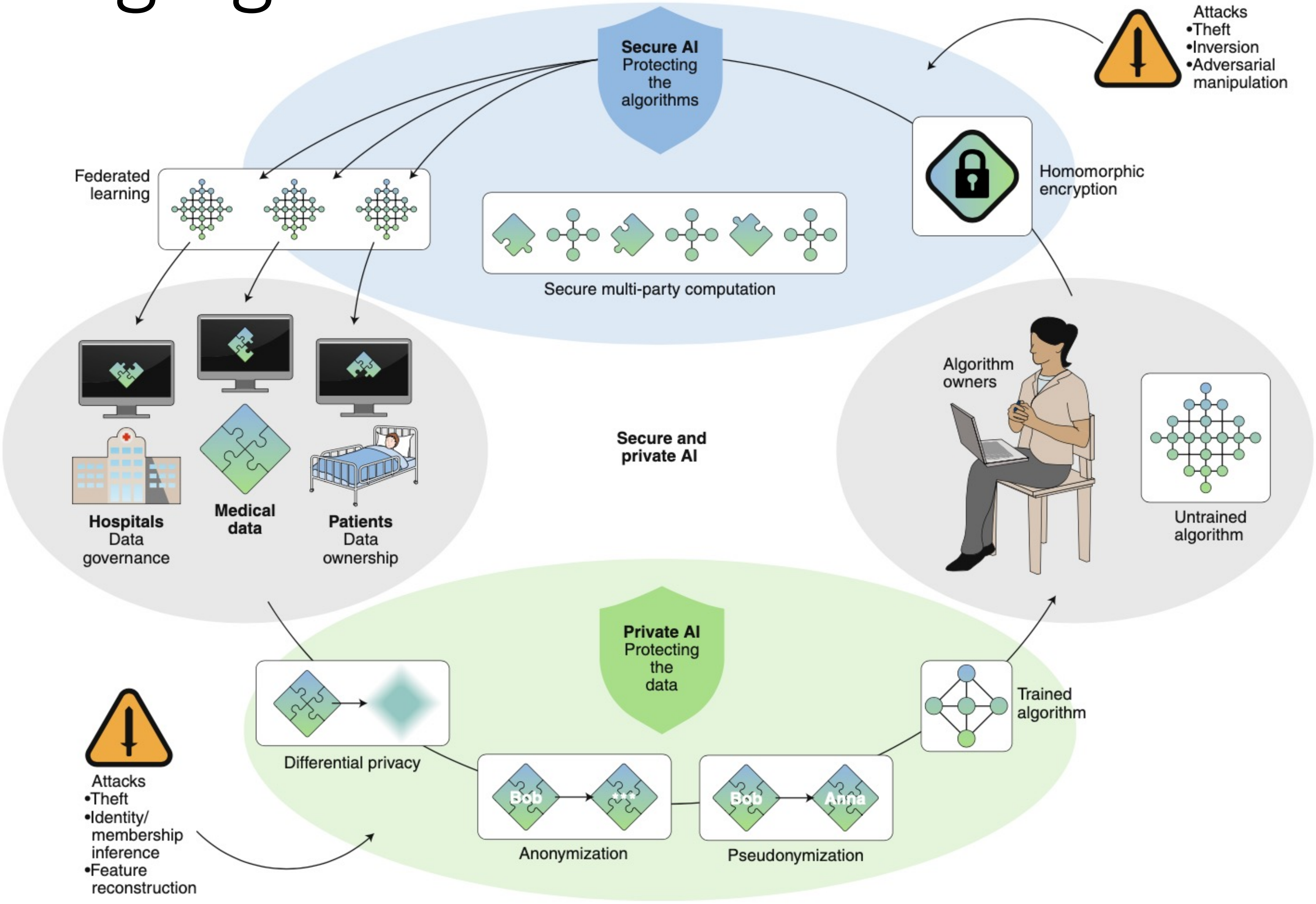
Chamikara, Mahawaga, et al. "Privacy preserving face recognition utilizing differential privacy." 2020.

Drug development



https://docs.microsoft.com/en-us/azure/confidential-computing/use-cases-scenarios

Medical imaging



Kaissis, Georgios A., et al. "Secure, privacy-preserving and federated machine learning in medical imaging." 2020.

Concepts and Taxonomy → Privacy Leakage in AI Systems → Protection Strategies

Applications → **Surveys and Tools** → **Future Directions**

Surveys

- General concepts
  - Al-Rubaie, Mohammad, and J. Morris Chang. "Privacy-preserving machine learning: Threats and solutions." 2019
  - De Cristofaro, Emiliano. “An overview of privacy in machine learning.” 2020
  - Rigaki, Maria, and Sebastian Garcia. “A survey of privacy attacks in machine learning.” 2020
- Differential privacy
  - Dwork, Cynthia. “Differential privacy: A survey of results.” 2008
  - Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." 2014
  - Ji, Zhanglong, Zachary C. Lipton, and Charles Elkan. "Differential privacy and machine learning: a survey and review." 2014
- Federated Learning
  - Kairouz, Peter, et al. “Advances and open problems in federated learning.” 2019
  - Yang, Qiang, et al. "Federated machine learning: Concept and applications." 2019

Liu, Haochen, et al. ”Trustworthy AI: A Computational Perspective.” 2021

Tools

- ❑ Differential Privacy
  - TensorFlow Privacy
  - Opacus
  - OpenDP
  - Diffpriv
- ❑ Federated Learning
  - TensorFlow Federated (TFF)
  - Paddle Federated Learning
  - FATE
  - FedML
  - LEAF
- ❑ Confidential Computing
  - Keystone Enclave
  - Google's FHE Repository
  - IBM FHE toolkit
  - AWS HE toolkit
  - SHEEP
  - CBMC-GC
  - Conclave
  - CipherCompute
  - MPC-SoK
  - HyCC
  - UC Compiler

Liu, Haochen, et al. "Trustworthy AI: A Computational Perspective." 2021

Future Directions

- ❑ Uncovering more sources of potential privacy leakage in AI systems
- ❑ Improving the performance of federated learning in heterogeneous environments
- ❑ Exploiting a better trade-off between utility and privacy loss in DP
- ❑ Improving the computation efficiency and flexibility of confidential computing
- ❑ Integrated systems and solutions