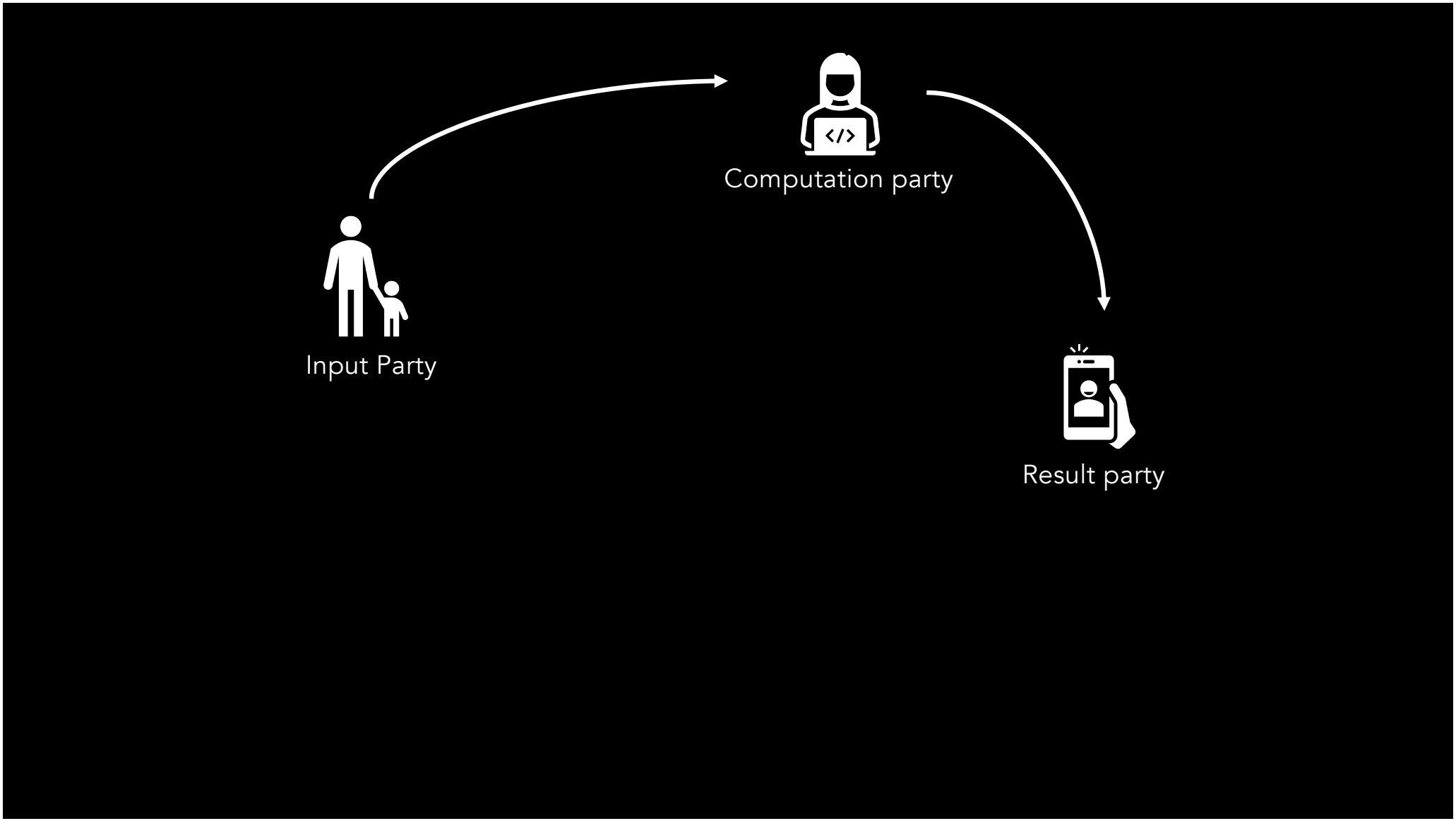
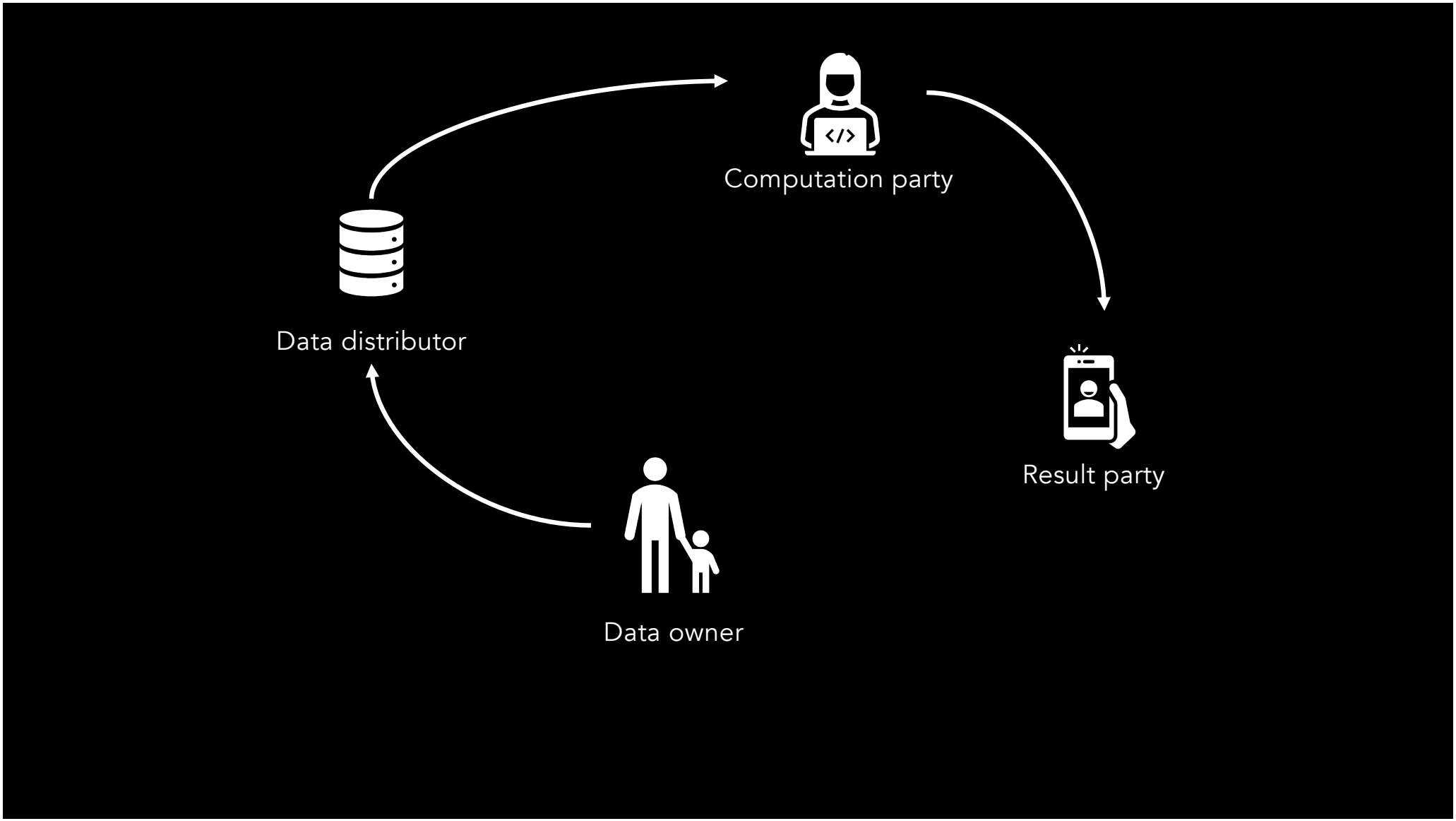
AI Security and Privacy (cont'd)

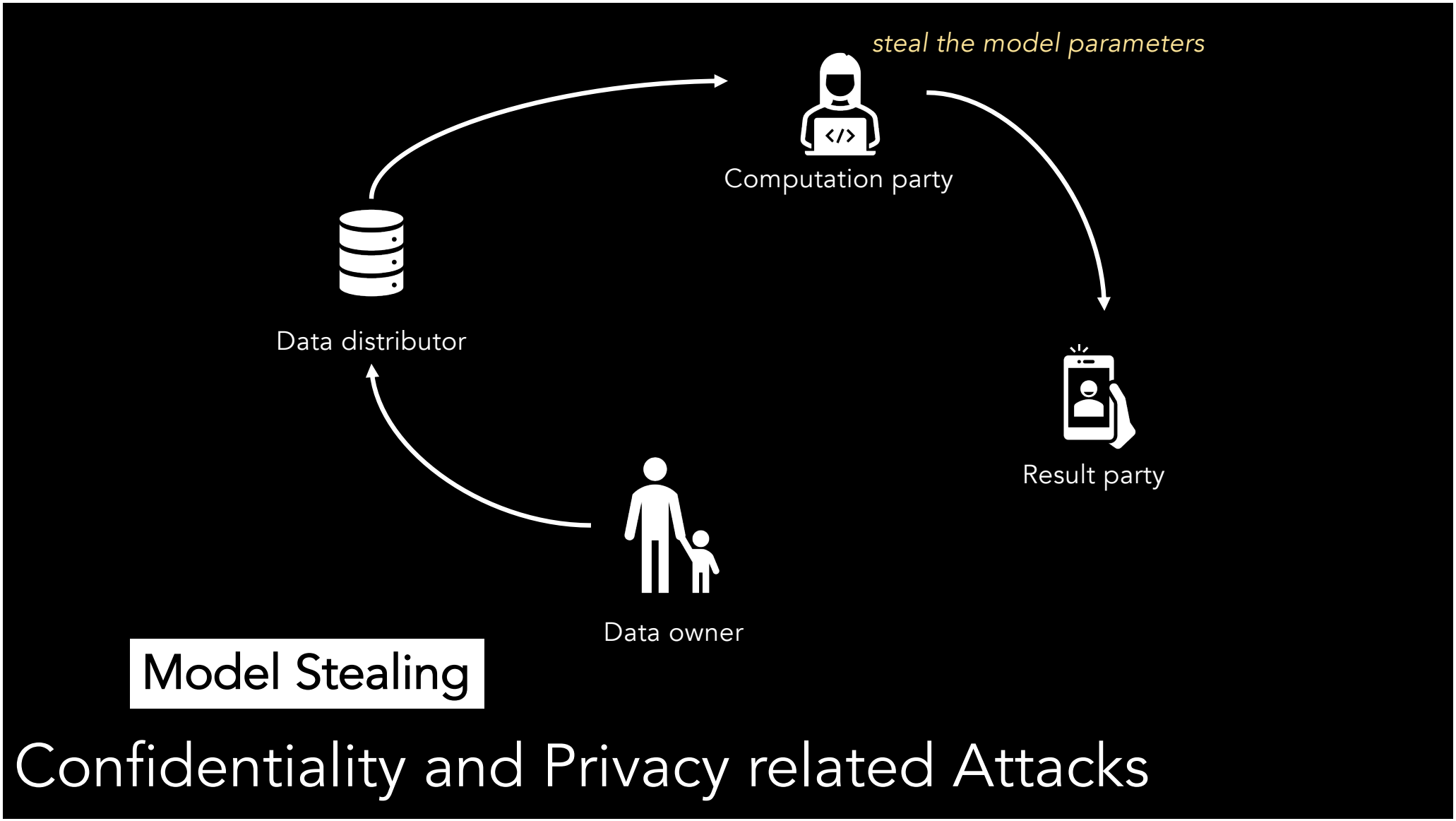
Jin L.C. Guo, SOCS McGill University

Agenda

- The concept of Privacy
- Confidentiality and Privacy Attacks
- Mitigation Methods

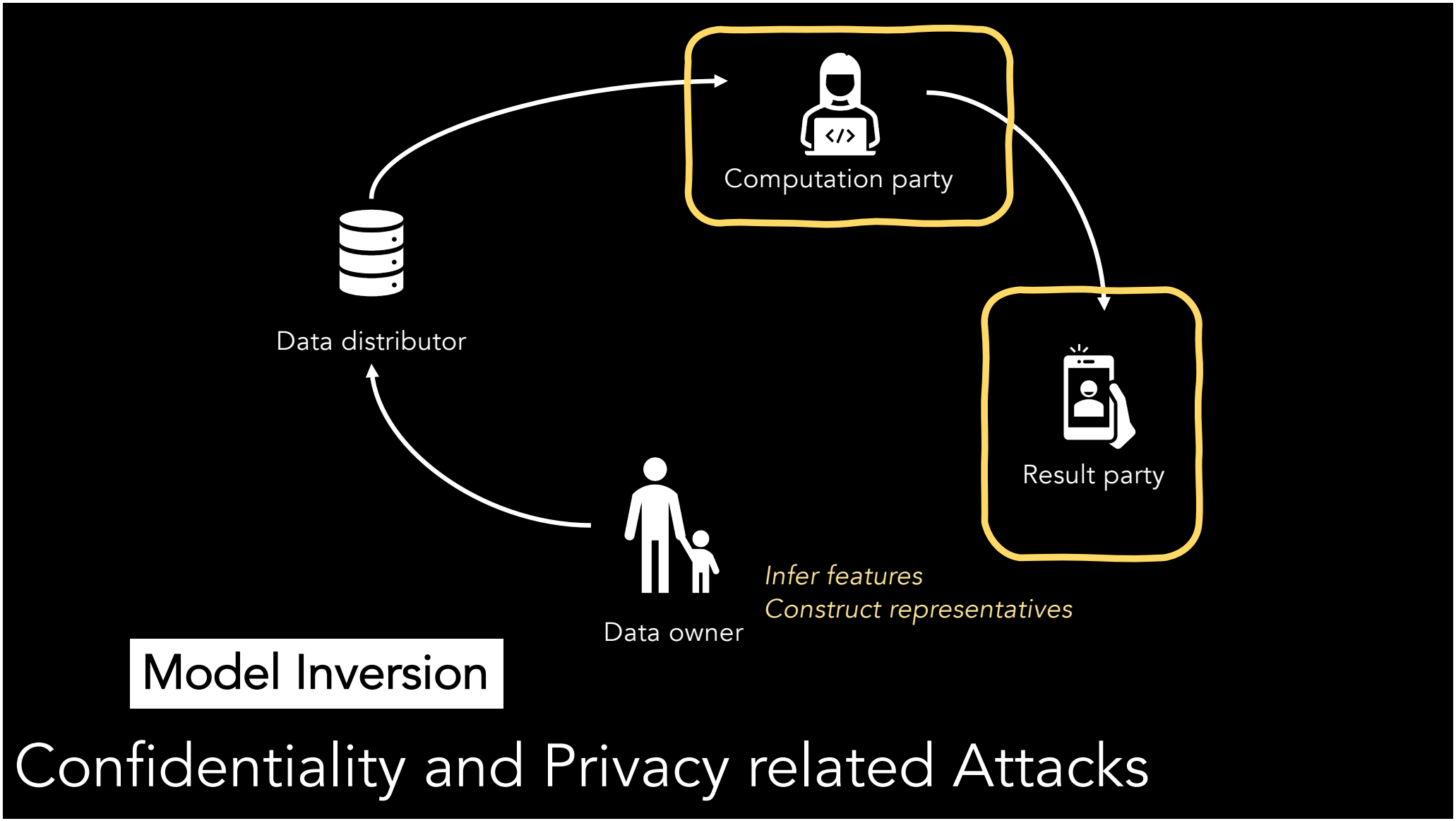

Computation party
Input Party
Result party


Computation party
Data distributor
Result party
Data owner

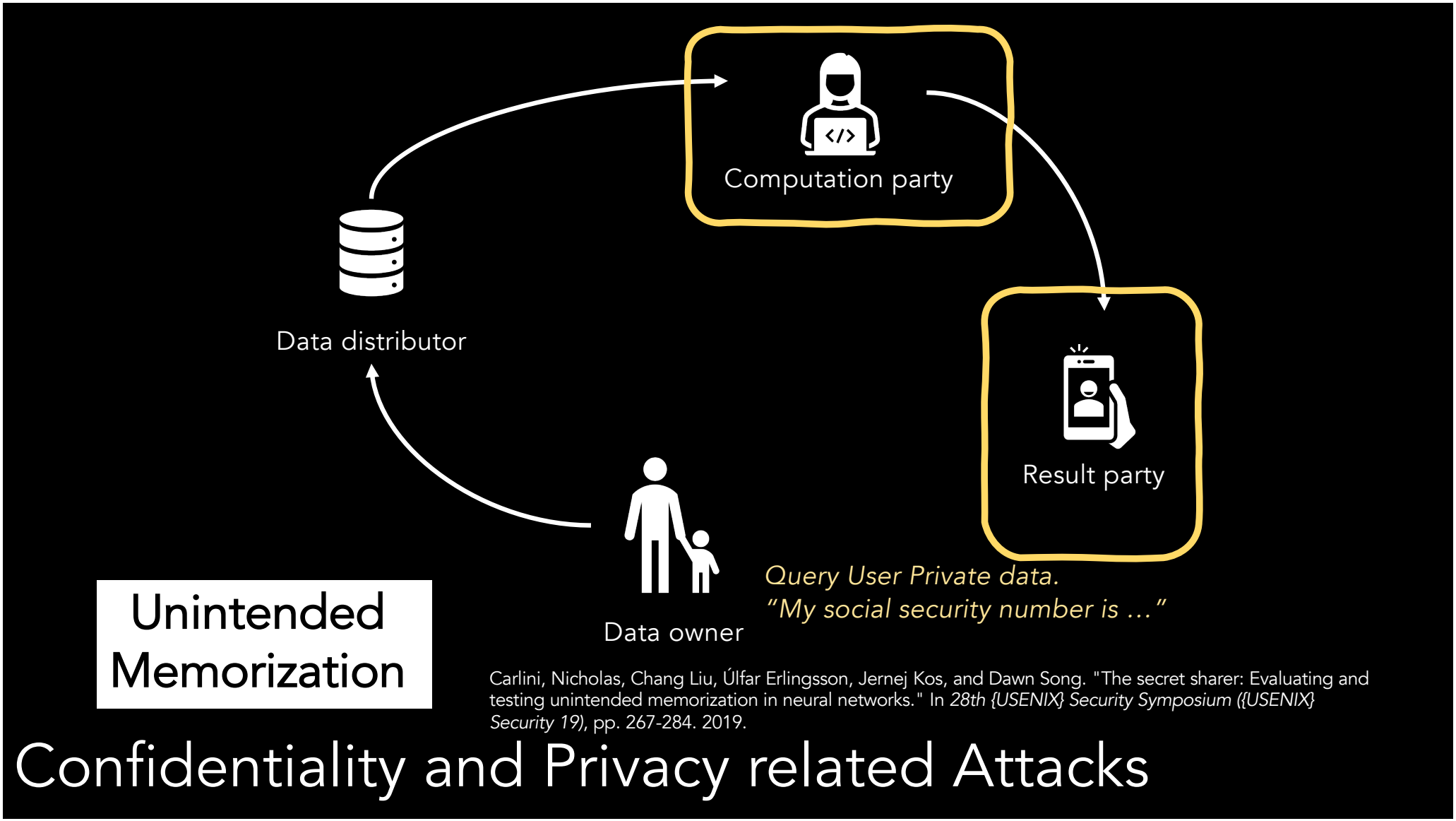


Model Stealing

Confidentiality and Privacy related Attacks



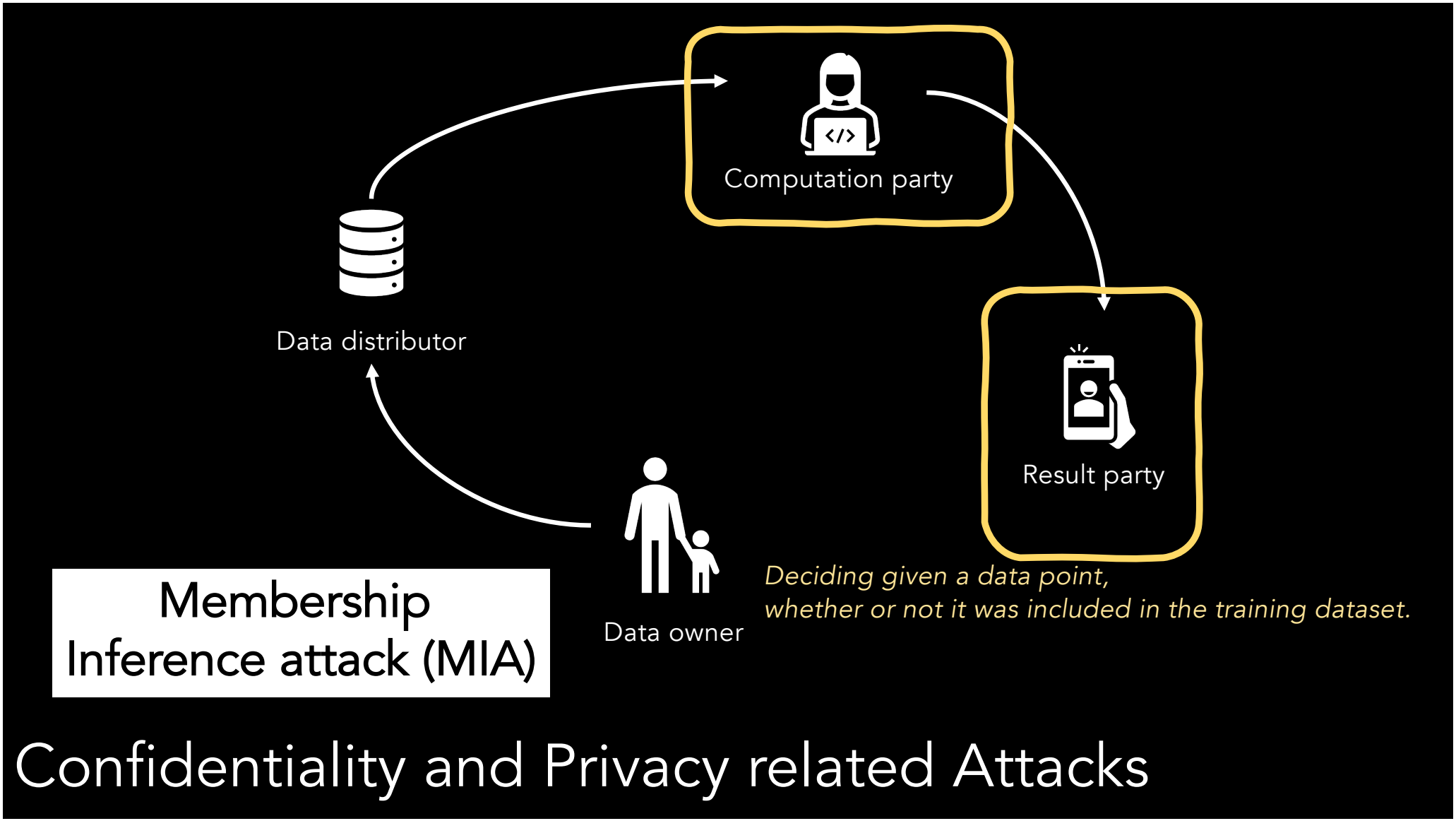
Confidentiality and Privacy related Attacks



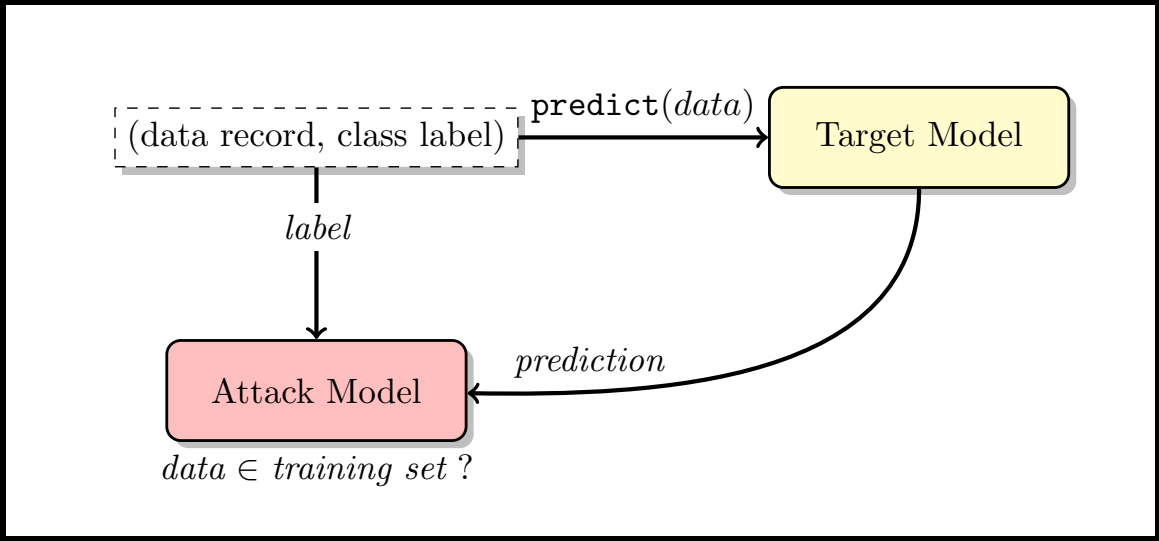
Unintended Memorization

Carlini, Nicholas, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. "The secret sharer: Evaluating and testing unintended memorization in neural networks." In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 267-284. 2019.

Confidentiality and Privacy related Attacks

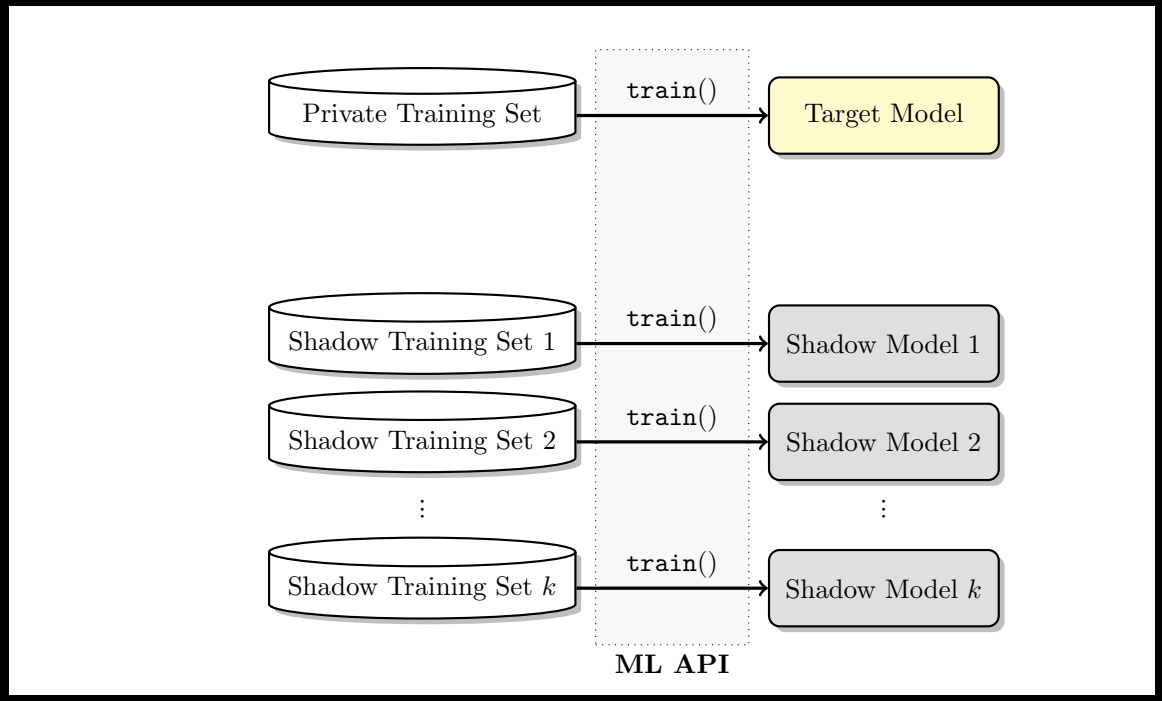


Confidentiality and Privacy related Attacks



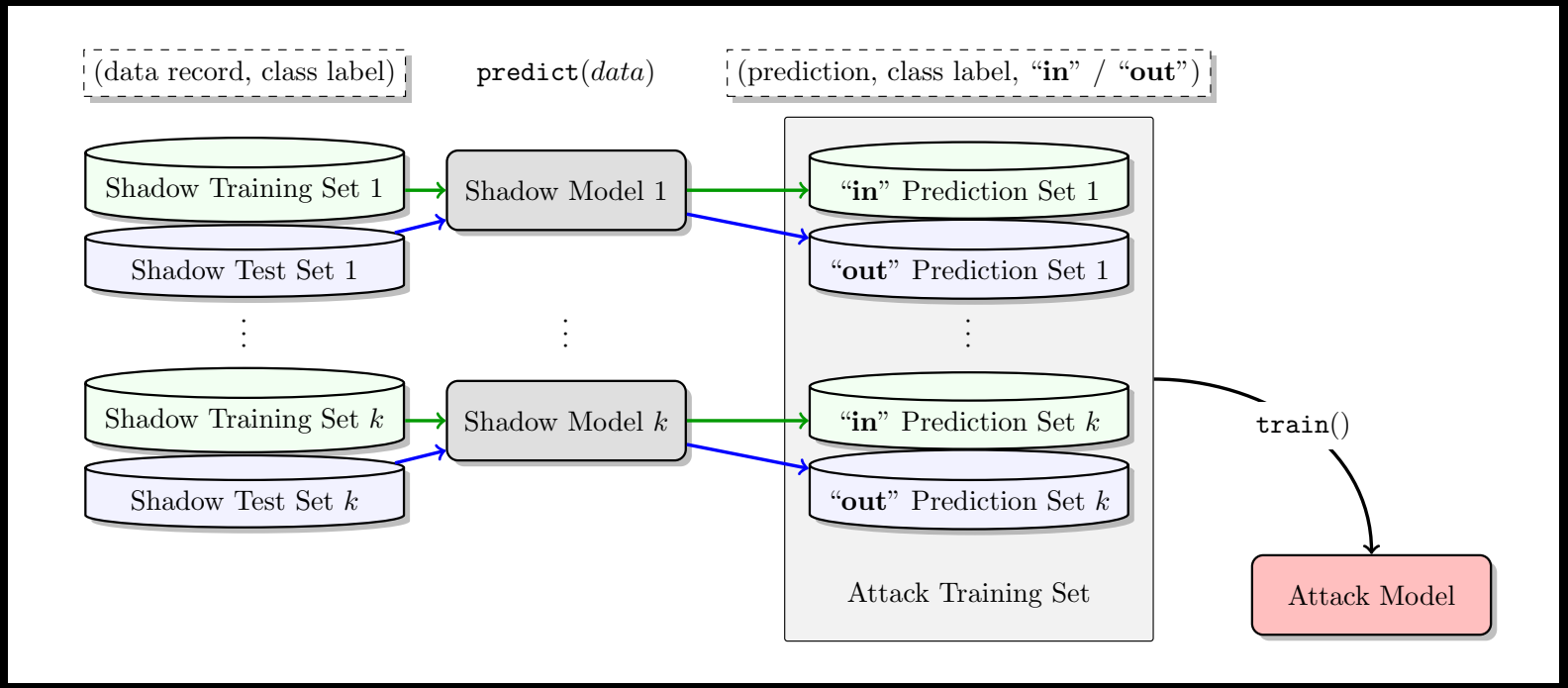
Membership Inference attack (MIA)

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership inference attacks against machine learning models." In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3-18. IEEE, 2017.



Membership Inference attack (MIA)

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership inference attacks against machine learning models." In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3-18. IEEE, 2017.



Membership Inference attack (MIA)

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership inference attacks against machine learning models." In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3-18. IEEE, 2017.

Membership Inference attack (MIA)

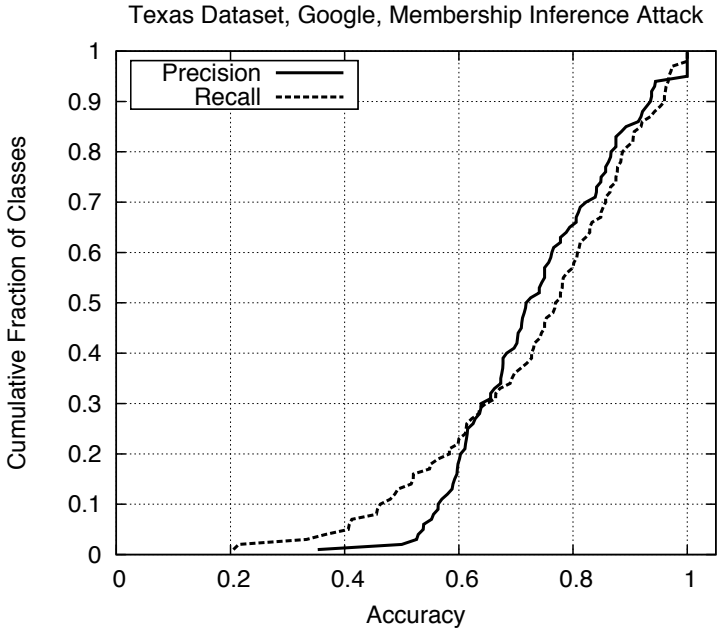


Fig. 6: Precision and recall of the membership inference attack against the classification model trained using Google Prediction API on the Texas hospital-stay dataset.

Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership inference attacks against machine learning models." In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3-18. IEEE, 2017.

Confidentiality

An explicit design property whereby one party wants to keep information (e.g., training data, testing data, model parameters, etc.) hidden from both the public and other parties (e.g., clients with respect to servers or vice-versa).

Privacy

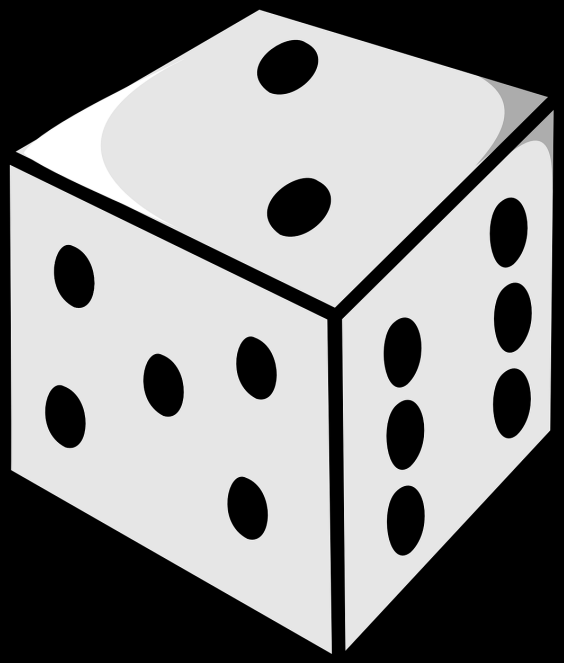
A property about protecting against unintended information leakage, whereby an adversary aims to infer sensitive information through some (intended) interaction with the victim.

Mitigation Methods - Differential Privacy

It formulates privacy as the property that an algorithm's output does not differ significantly statistically for two versions of the data differing by only one record.

A randomized algorithm is said to be (**ε** ,**δ**) differentially private if for two neighboring training datasets T , T', i.e. which differ by at most one training point, the algorithm A satisfies for any acceptable set S of algorithm outputs:

$$Pr[A(T) \in S] \leq e^{\varepsilon} Pr[A(T') \in S] + \delta$$



Mitigation Methods - Differential Privacy

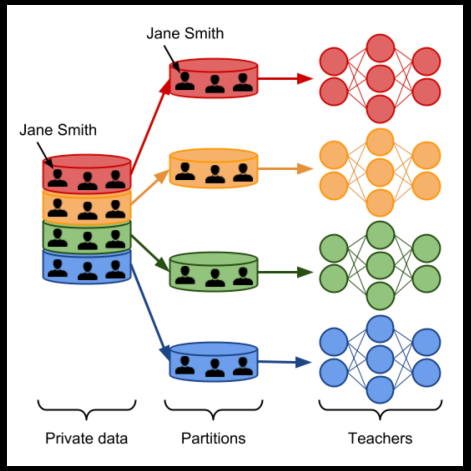
Noisy SGD

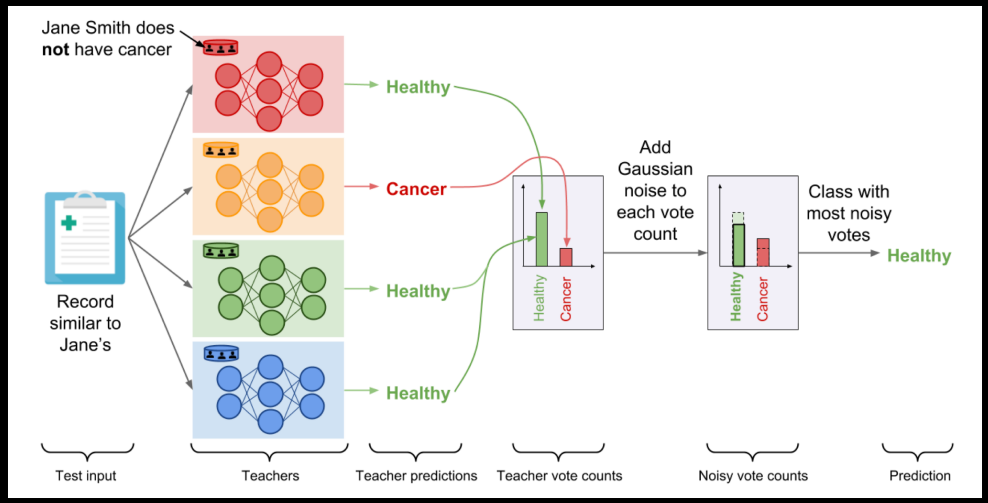
Abadi, Martin, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep learning with differential privacy." In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308-318. 2016.

Mitigation Methods - Differential Privacy

Noisy SGD

Private Aggregation of Teacher Ensembles (PATE)





Papernot, Nicolas, and Ian Goodfellow. "Privacy and machine learning: two unexpected allies?." (2018).

Review

- The concept of Privacy
- Common Privacy Attacks
- Mitigation Methods

Next on Thursday

Privacy (cont'd) and Accountability