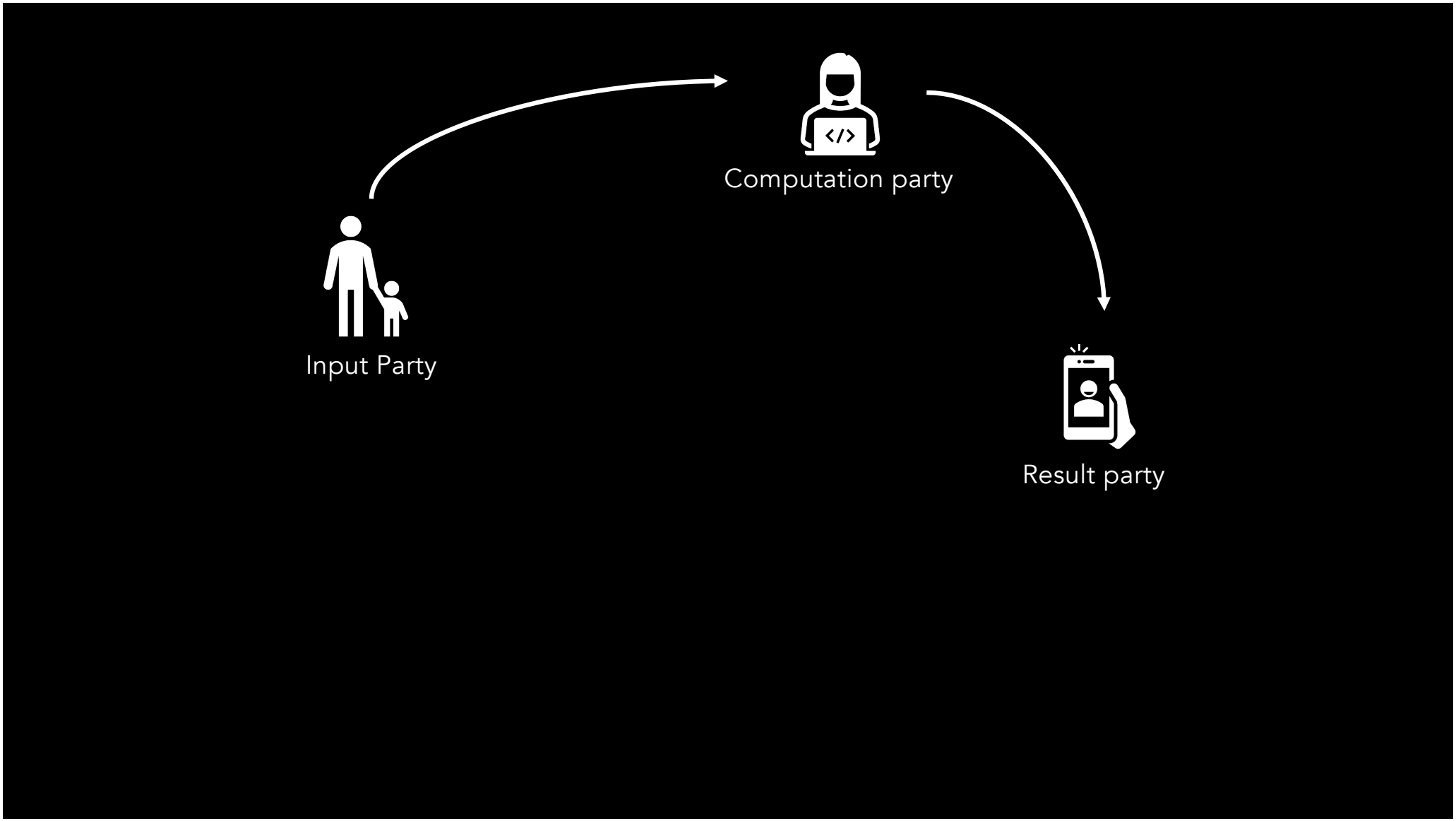
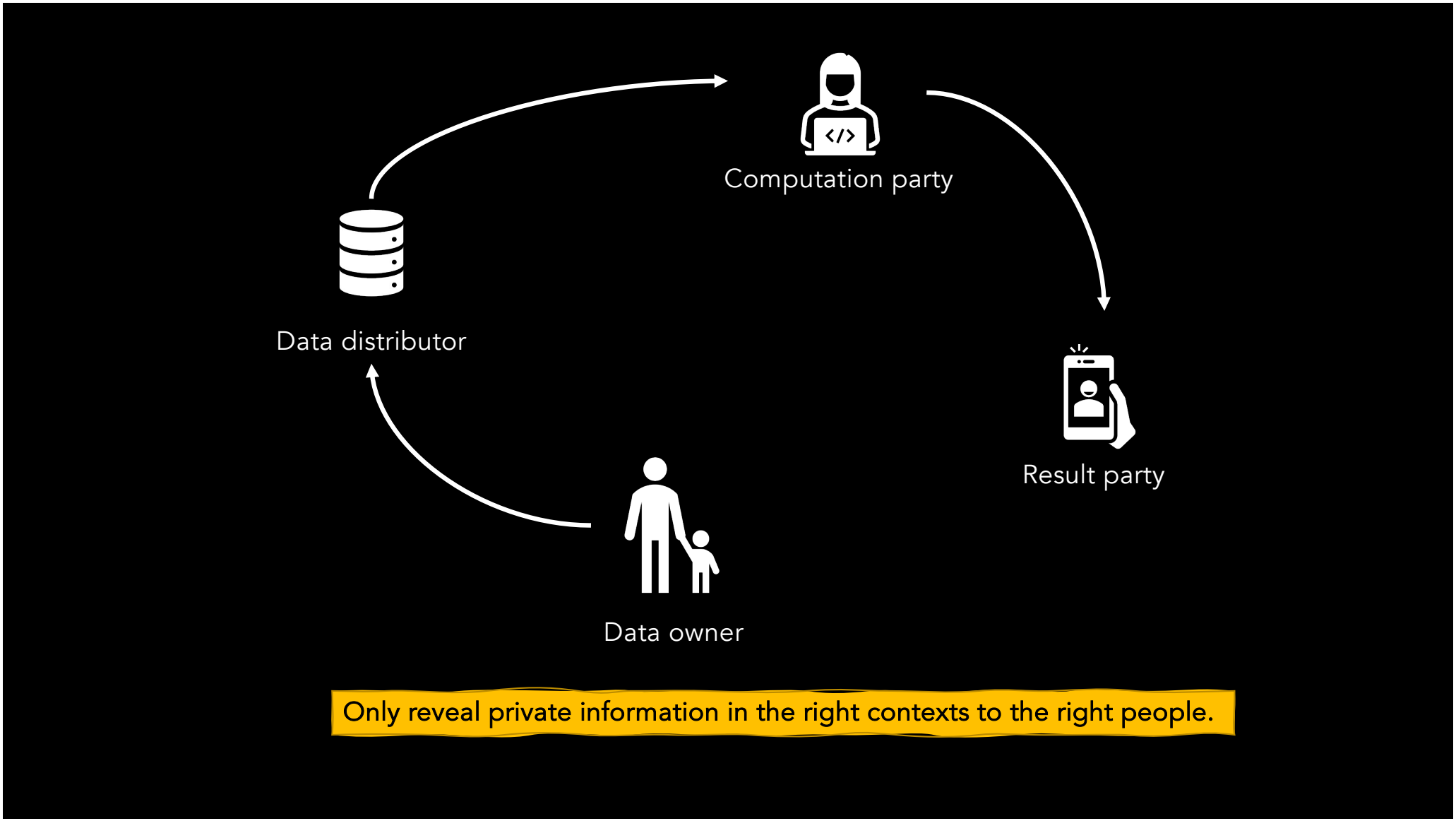
AI Security and Privacy (cont'd)

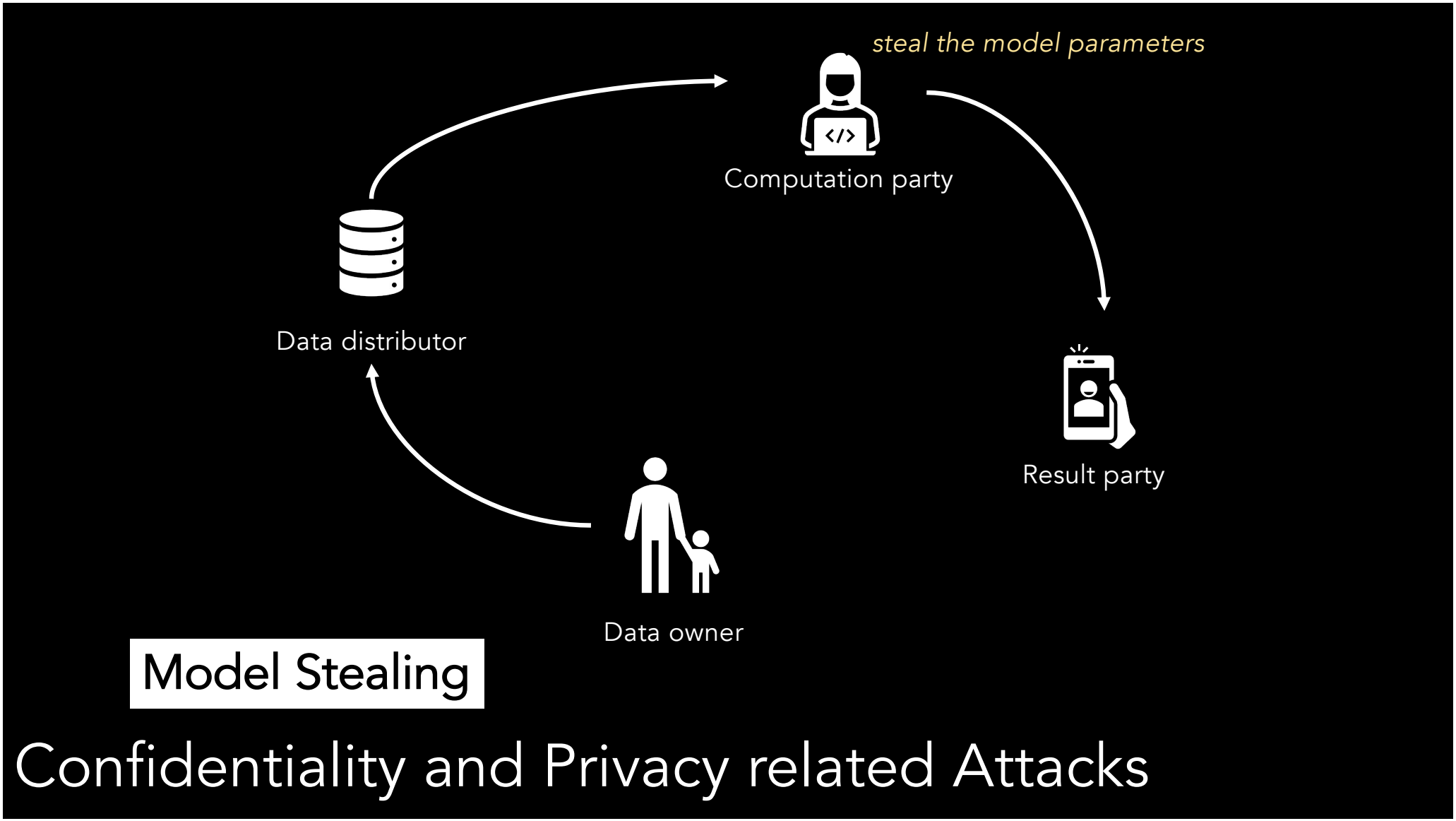
Jin L.C. Guo, SOCS McGill University

Agenda

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

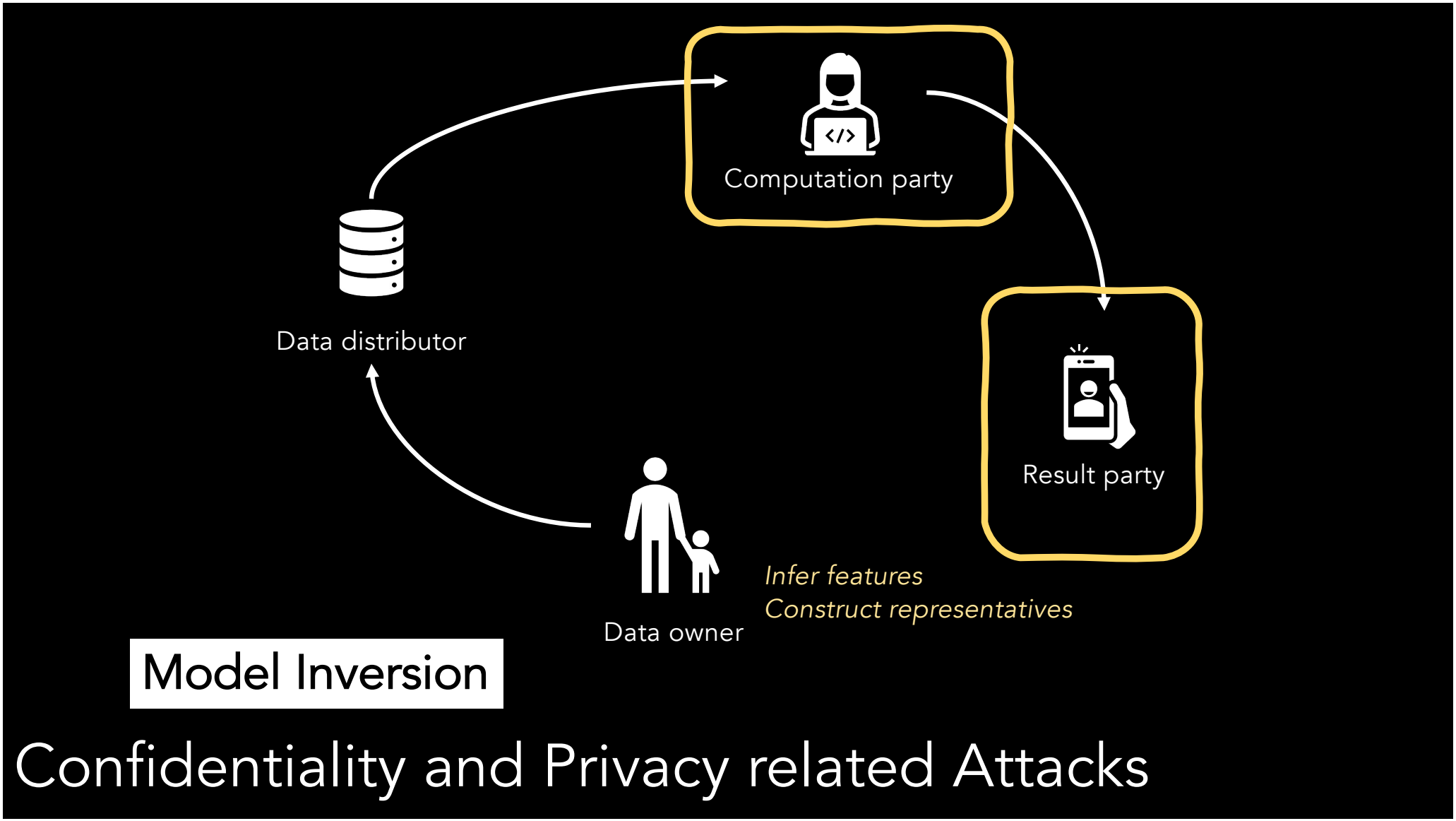

Computation party
Input Party
Result party


Computation party
Data distributor
Result party
Data owner
Only reveal private information in the right contexts to the right people.

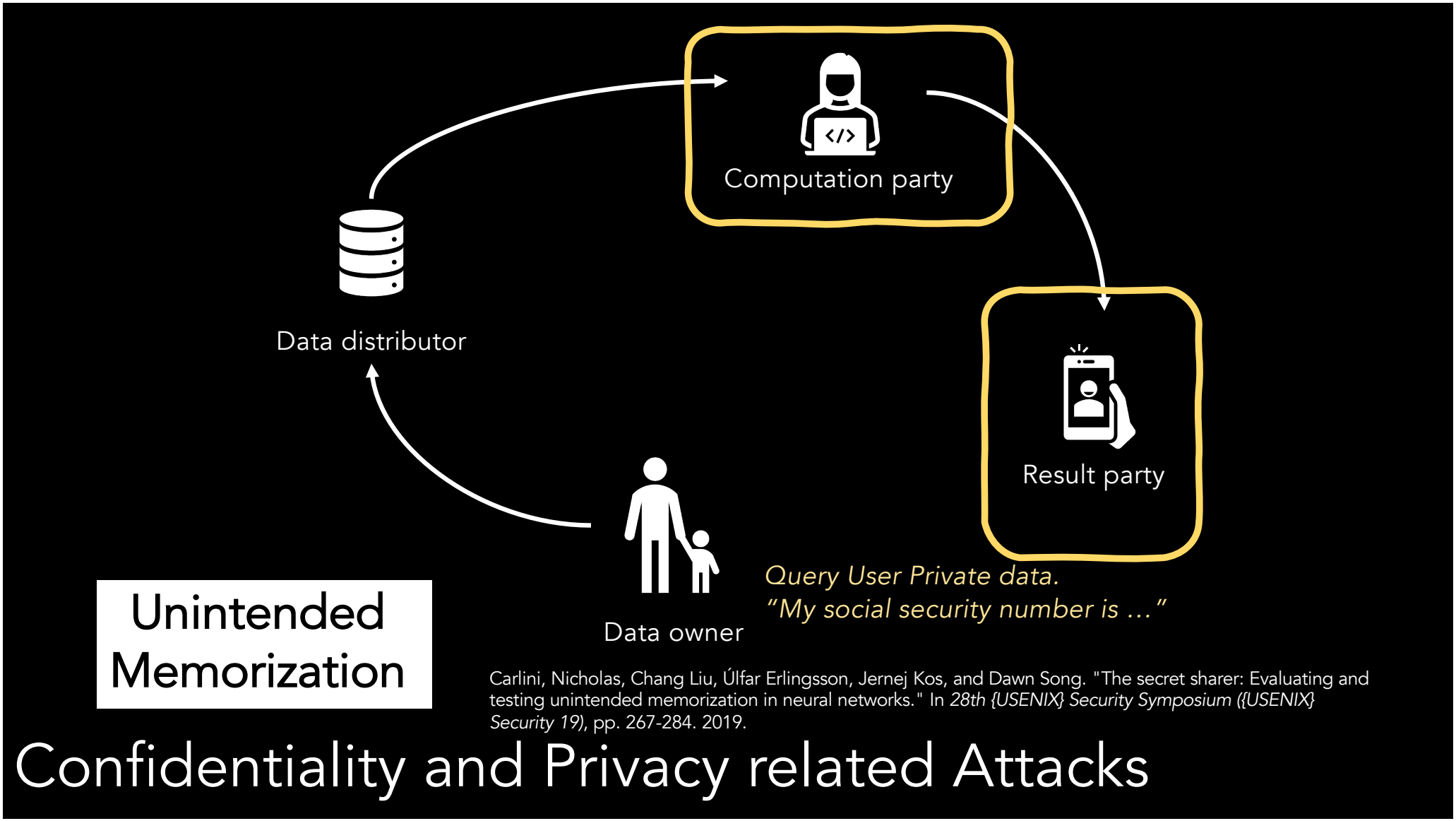


Model Stealing

Confidentiality and Privacy related Attacks



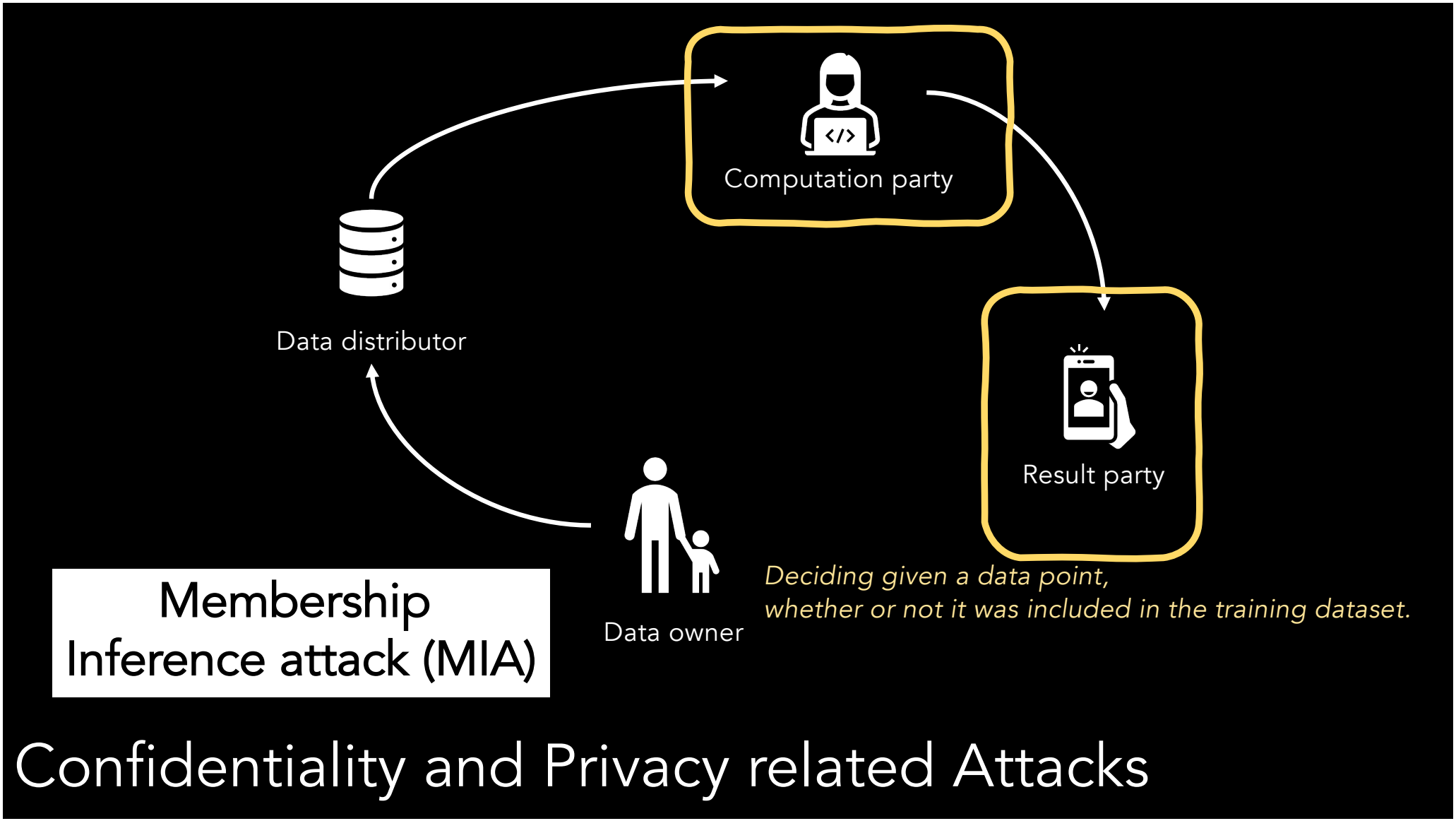
Confidentiality and Privacy related Attacks



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



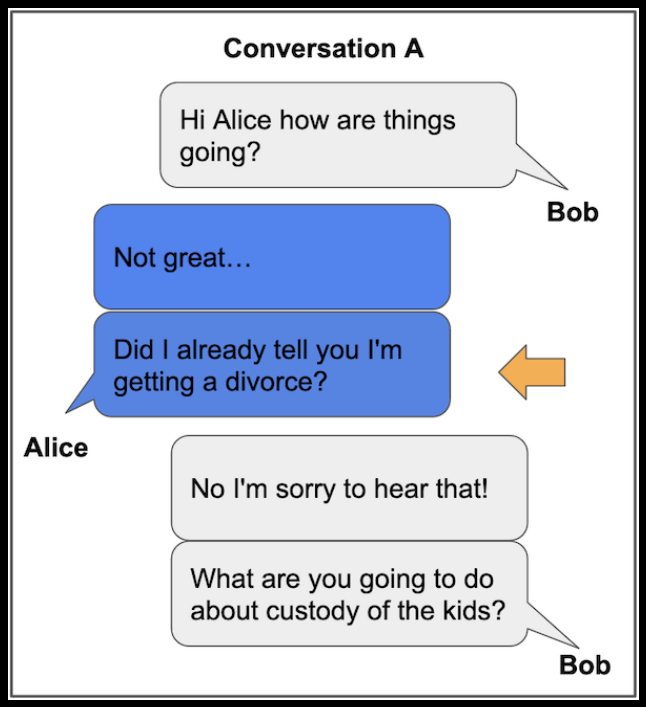
Confidentiality and Privacy related Attacks

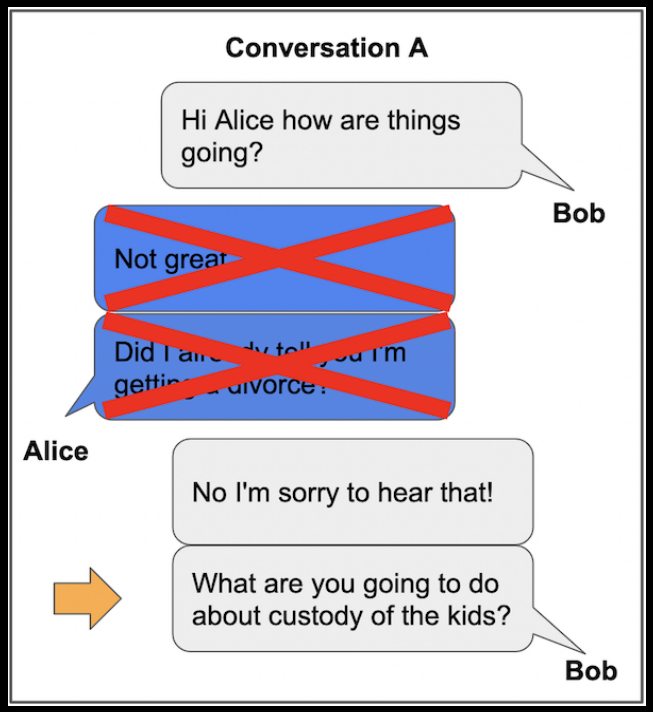
Mitigation Methods – Data Sanitization

Design efficient algorithms to identify and remove private information in the training data.

Social security numbers, specific forms of medical notes, credit card, phone number, etc.

How to specify the private information?





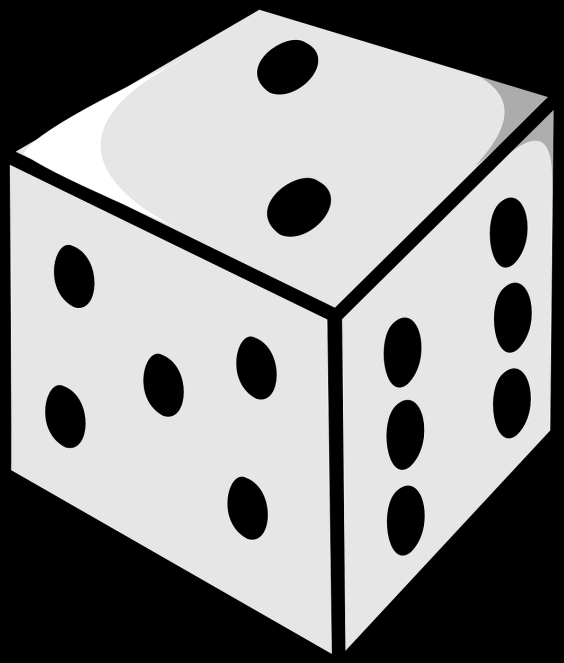
Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What Does it Mean for a Language Model to Preserve Privacy? In FAccT '22. https://doi.org/10.1145/3531146.3534642

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

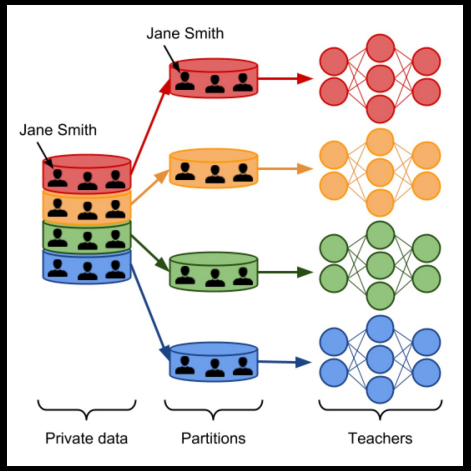
Noisy SGD

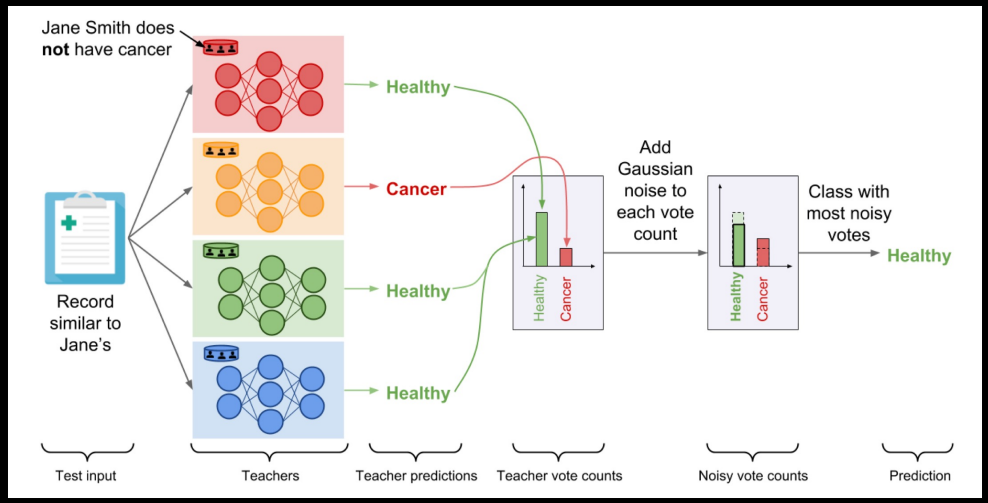
Abadi, Martin, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. "Deep learning with differential privacy." In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308-318. 2016.

Mitigation Methods - Differential Privacy

Noisy SGD

Private Aggregation of Teacher Ensembles (PATE)





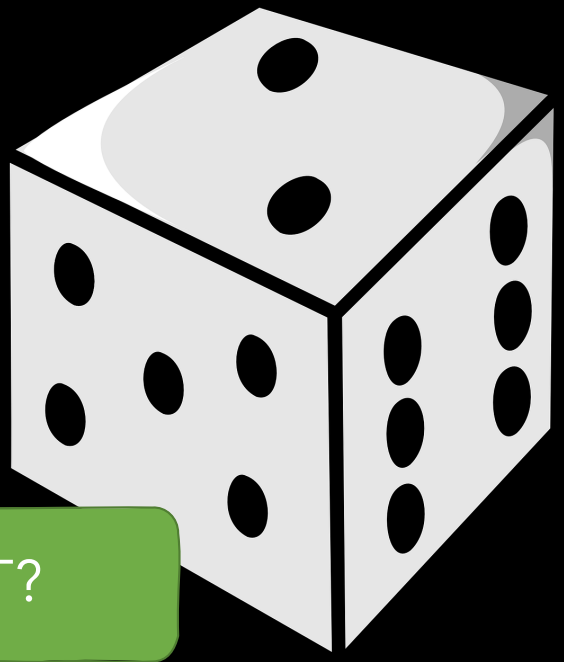
Papernot, Nicolas, and Ian Goodfellow. "Privacy and machine learning: two unexpected allies?." (2018).

Mitigation Methods - Differential Privacy

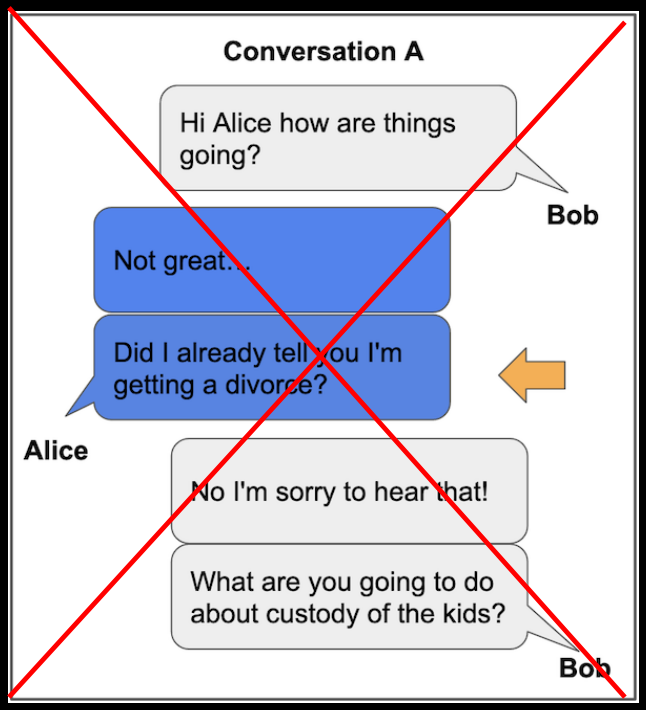
It formulates privacy as the property that an algorithm's output does not differ significantly statistically for two versions of the data differing by only one record.

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:

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



How to define what constitute T?



Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What Does it Mean for a Language Model to Preserve Privacy? In FAccT '22. https://doi.org/10.1145/3531146.3534642

Publicly accessible == Public-intended?

The New York Times

Lawsuit Takes Aim at the Way A.I. Is Built

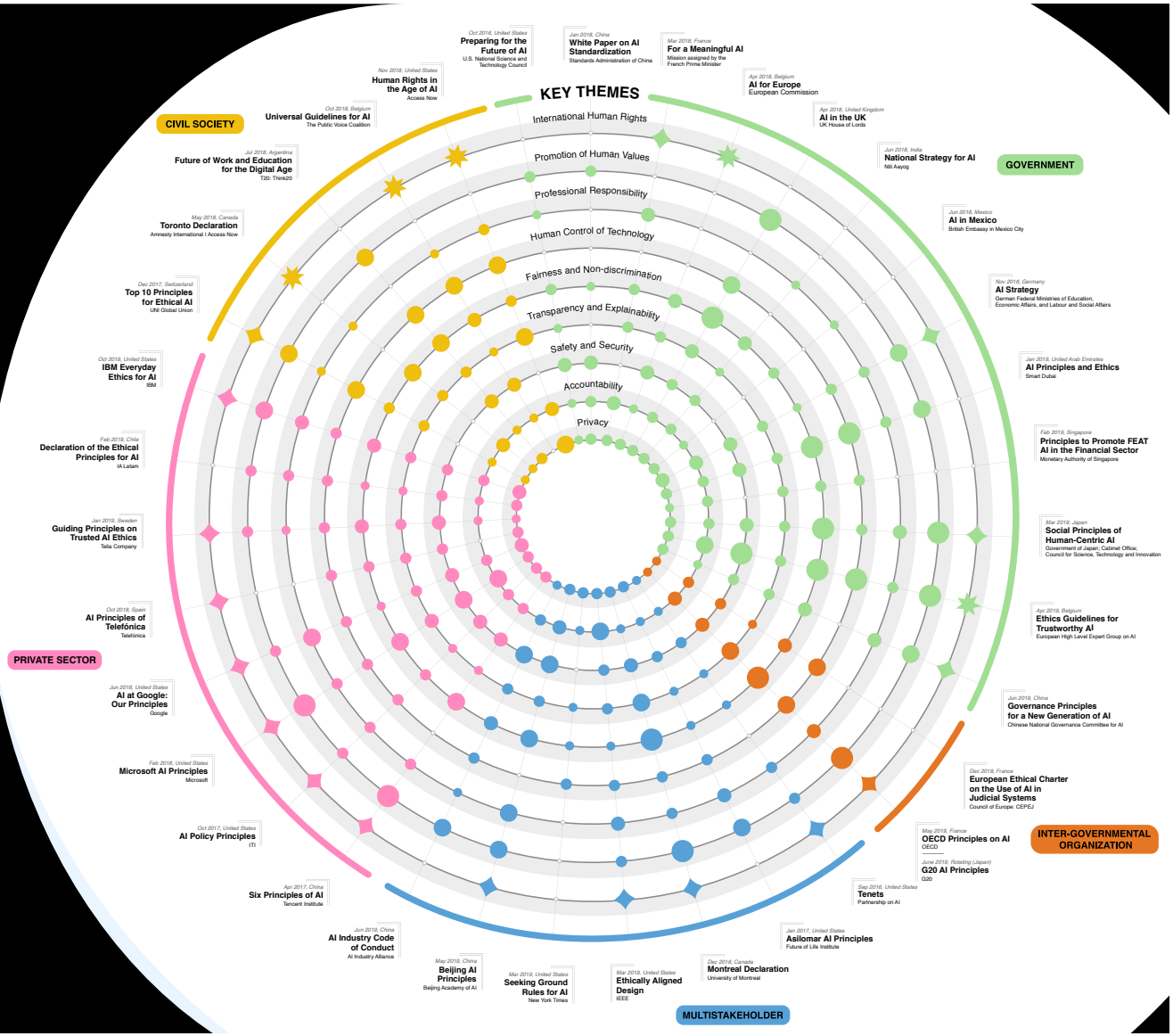
A programmer is suing Microsoft, GitHub and OpenAI over artificial intelligence technology that generates its own computer code.

"The suit is part of a groundswell of concern over artificial intelligence. Artists, writers, composers and other creative types increasingly worry that companies and researchers are using their work to create new technology without their consent and without providing compensation. Companies train a wide variety of systems in this way, including art generators, speech recognition systems like Siri and Alexa, and even driverless cars."

Review

- The concept of Privacy
- Common Privacy Attacks
- Mitigation Methods (and beyond)

AI Principles Overview



AI Principles

- Each group is assigned with one block in Miro.
  - Each member read and interpret one principle in the corresponding theme.
  - Connect the principle to any topics we have discussed in this lecture.
  - Discuss the connections within the group and add the outcome in Miro