Reasoning about “privacy” in ML (2)

Prior work ~ inferring class representatives

Model Inversion [Fredrikson et al. CCS'15] and GAN attacks [Hitaji et al. CCS'17] infer properties of an entire class

E.g.: given a gender classifier, infer what a male looks like

In our work:

Infer properties of a subset of the training inputs but not of the whole class

E.g.: given a gender classifier, infer race of people in Bob's photos

Reasoning about "privacy" in ML (2)

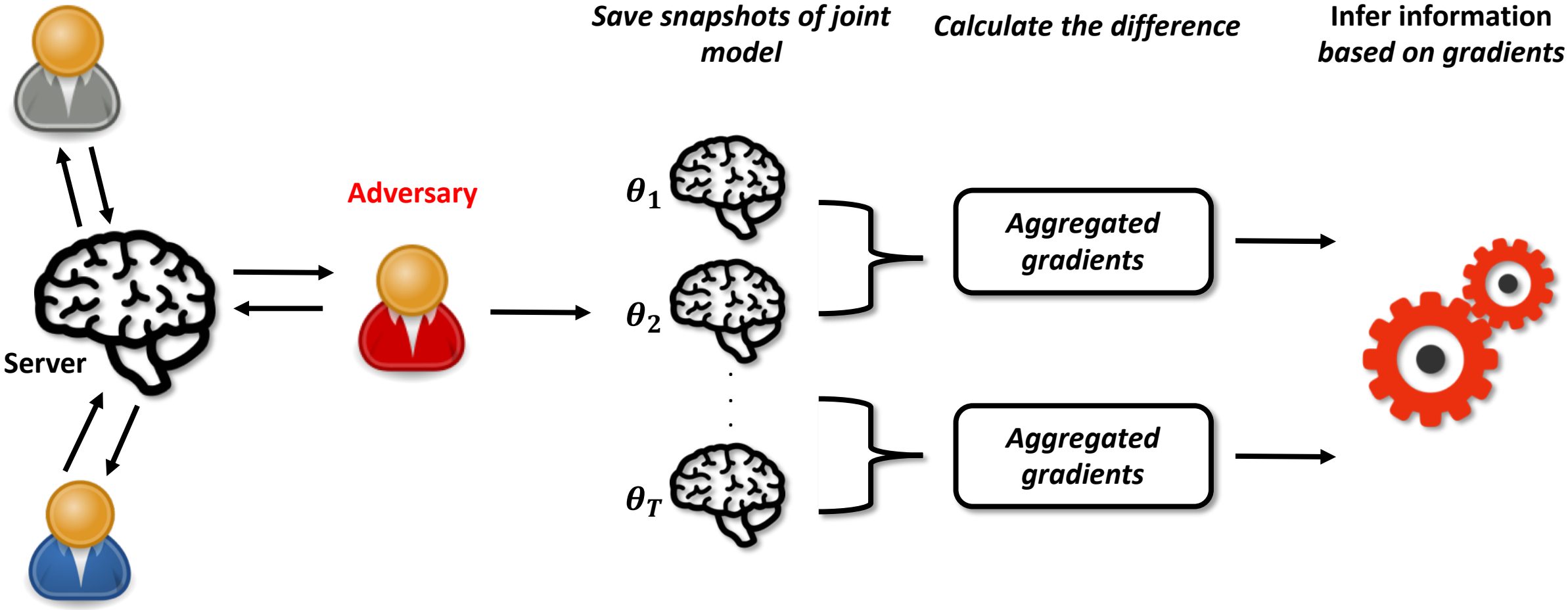
Prior work ~ inferring class representatives

- Model Inversion [Fredrikson et al. CCS'15] and GAN attacks [Hitaji et al. CCS'17] infer properties of an entire class
- E.g.: given a gender classifier, infer what a male looks like

In our work:

- Infer properties of a subset of the training inputs but not of the whole class
- E.g.: given a gender classifier, infer race of people in Bob's photos

Passive Property Inference Attack



Luca Melis, Congzheng Song, Emiliano De Cristofaro, Vitaly Shmatikov. Inference Attacks Against Collaborative Learning. Under Submission.