## Privacy Attacks on Machine Learning

### Introduction:

ML models can leak sensitive data. Privacy attacks target training data, outputs, or model parameters. This summary covers **14 key research papers**.

### 1. Membership Inference Attacks

- Shokri et al. (2017) – Shadow models
- Choquette-Choo et al. (2021) – Label-only attacks
- Song & Mittal (2020) – Privacy risk evaluation

### 2. Model Inversion Attacks

- Fredrikson et al. (2015) – Confidence-based
- Zhang et al. (2020) – GAN-based
- Wang et al. (2021) – Variational

### 3. Poisoning & Backdoors

- Biggio et al. (2012) – SVM poisoning
- Gu et al. (2017) – BadNets

### 4. Model Extraction / Stealing

- Tramèr et al. (2016) – API attacks
- Orekondy et al. (2019) – Knockoff Nets

### 5. Property & Generative Model Attacks

- Ganju et al. (2018) – Property inference
- Song et al. (2022) – Generative models

### 6. Surveys

- Jayaraman & Evans (2019) – Survey of attacks/defenses
- Rigaki & Garcia (2023) – Modern survey

### Key Points:

- Overfitting increases risk.
- Black-box and label-only attacks are effective.
- Differential privacy and regularization reduce leakage.
- Access control, output limiting, federated learning, and monitoring help protect models.

### Conclusion:

ML models face real privacy threats. Strong defenses and privacy-aware design are essential.