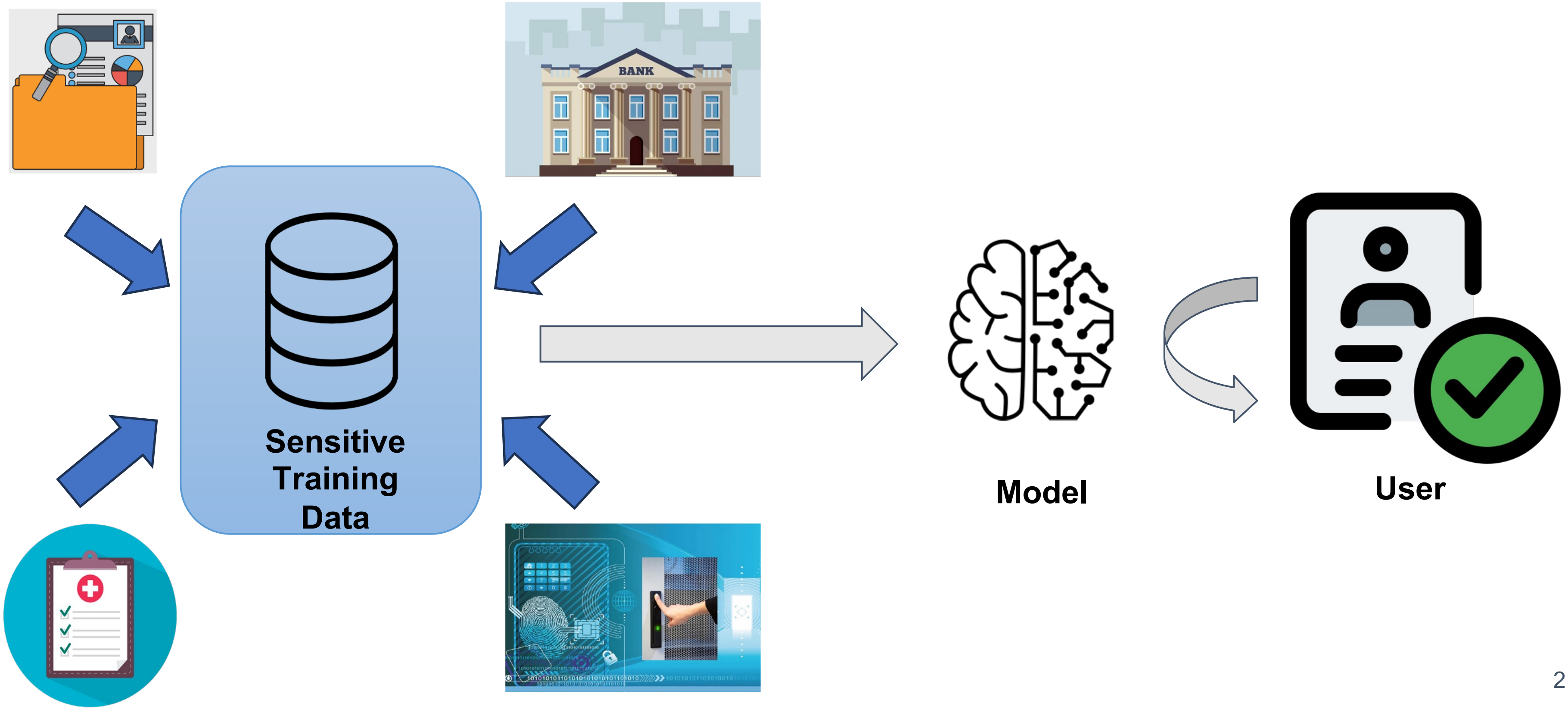
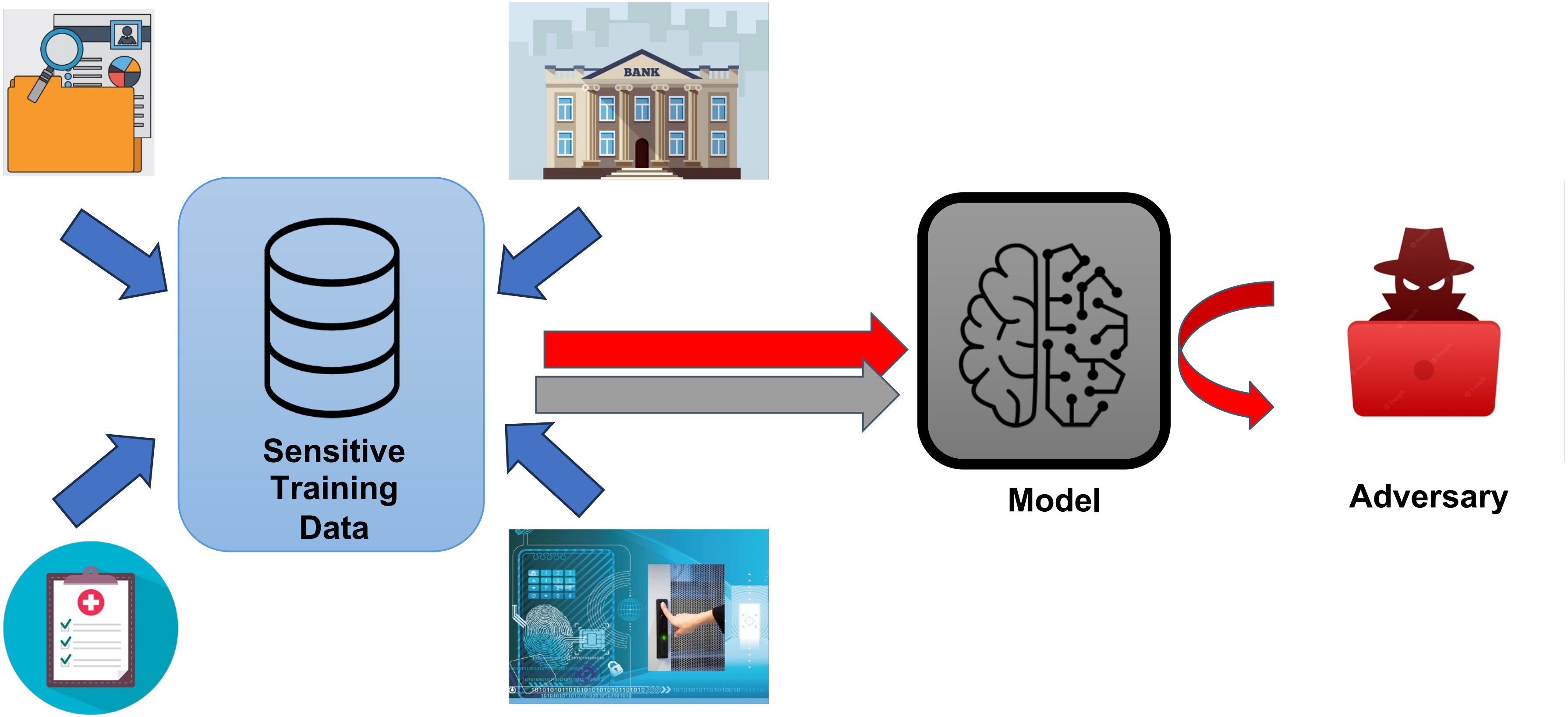
SoK: Model Inversion Attack Landscape: Taxonomy, Challenges, and Future Roadmap

Sayanton V. Dibbo

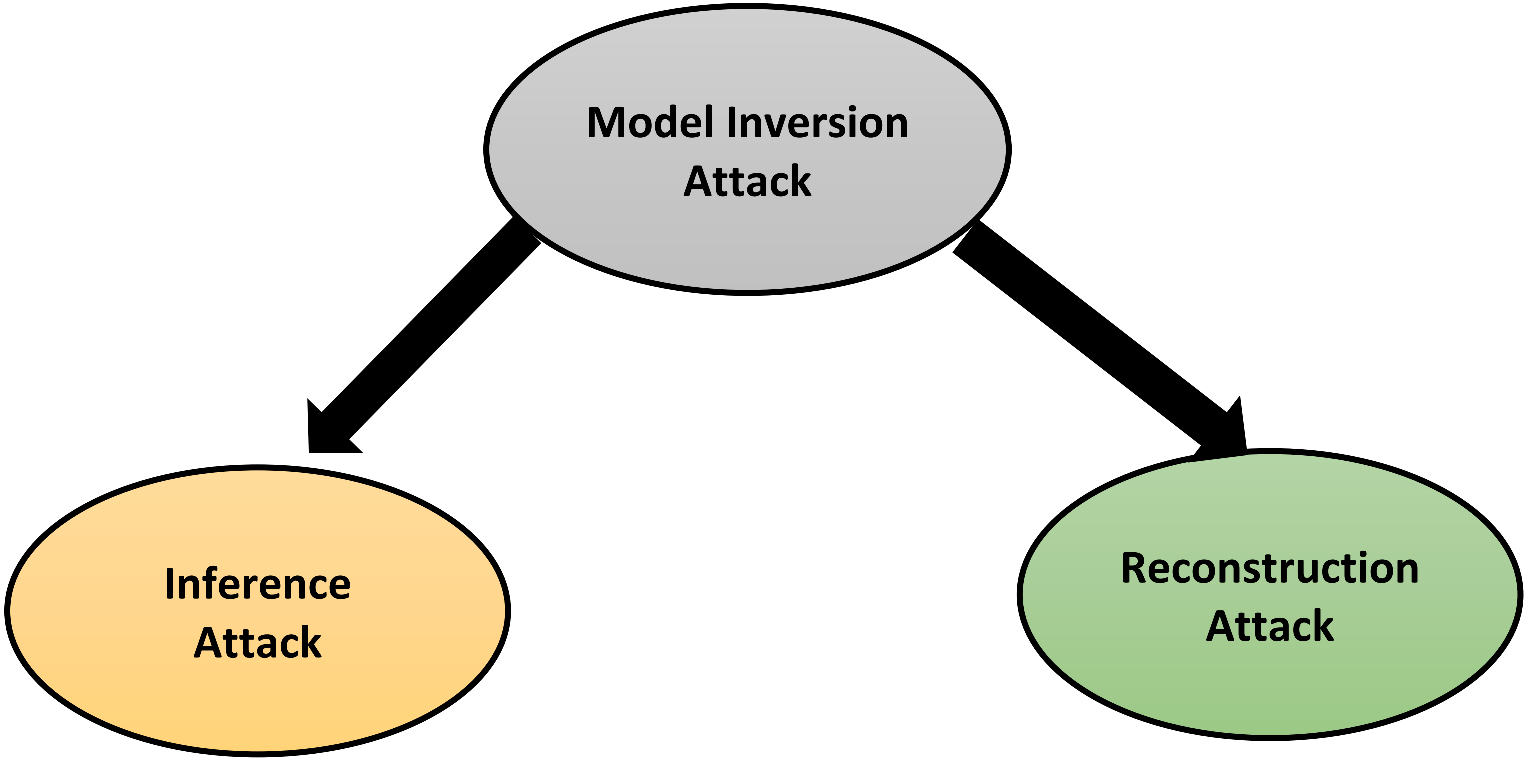
Model Inversion Attack



Model Inversion Attack



Model Inversion Attack categories



Inference Attack

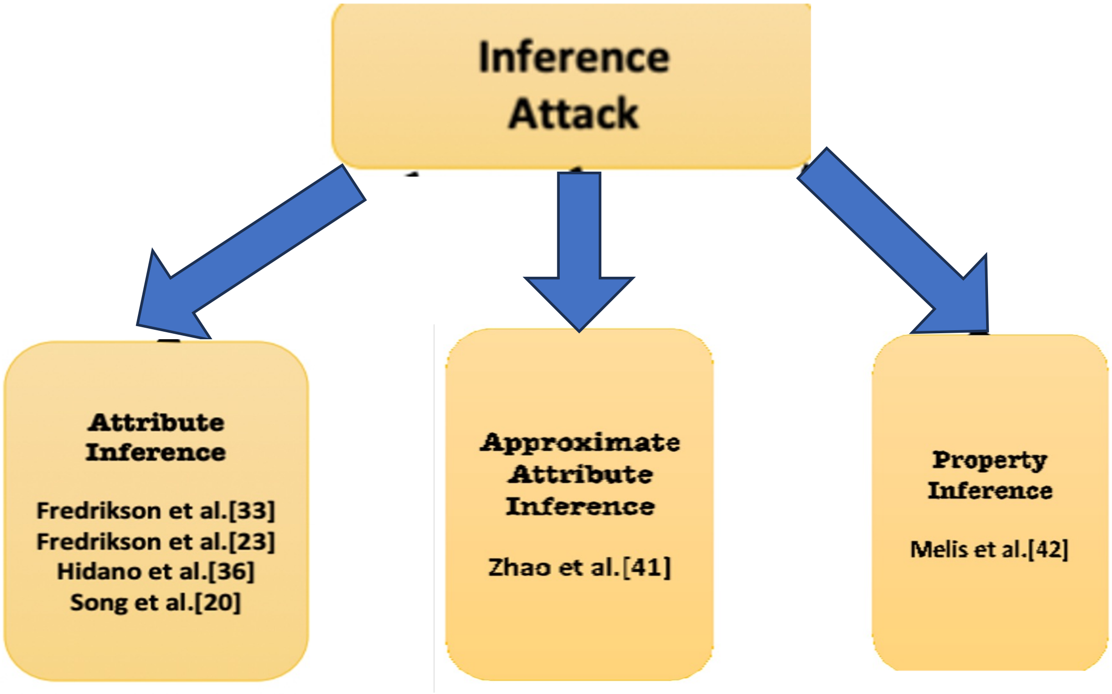
Inference Attack

- Goal: Infer sensitive training data
- Capabilities: other attributes, class labels, confusion matrix, etc.
- Applicable for tabular data domain
- e.g., lifestyle like smoking, drinking, marital status, ethnicity, etc.

Ethnicity	**Gender**	**Income**
Native Hawaiian	M	>=50k
American Indian	F	<50k
White	F	>=50k

Ethnicity
Native Hawaiian
American Indian
White

Subcategories of Inference Attack



Attribute Inference (AI)

- Infer exactly an individual's sensitive attribute values
- Adversary uses output labels and other information
- Other additional information can be:
  - *confidence scores*
  - *information about non-sensitive attributes (tabular data)*
- e.g., smoking habit> 'yes' or 'no'

Approximate Attribute Inference (AAI)

- Infer attribute close to an individual’s sensitive attribute
- More relaxed than AI
- Uses distance metric to find close attribute
  - Hamming distance
  - Manhattan distance
- e.g., age in tabula data, features in image

Property Inference

- Infer property in the training samples
- Leaks sensitive properties of training data
- Mostly applicable to individual sample
- e.g., someone wearing glasses, hair color, or specialty

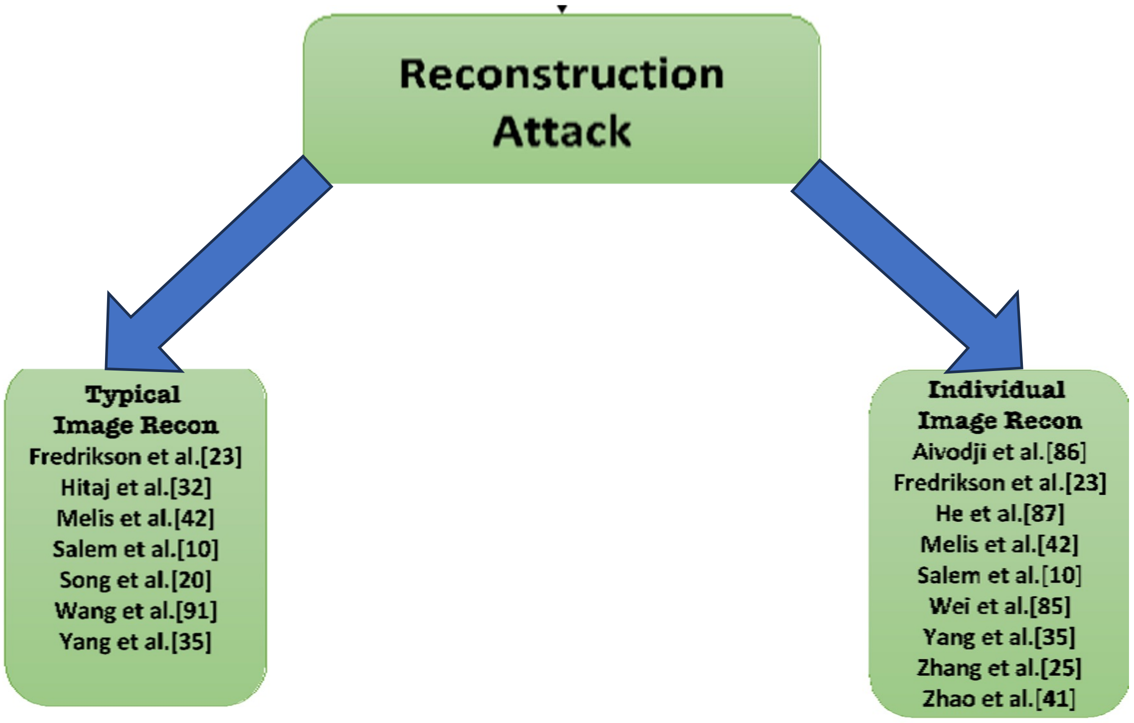
Reconstruction Attack

Reconstruction Attack

- Goal: Reconstruct training data
- Capabilities: confidence scores, gradients, masked/blurred image, etc.
- Applicable for image data domain
- e.g., an individual image, a generic class representative image etc.

Actual Reconstructed

Subcategories of Reconstruction Attack



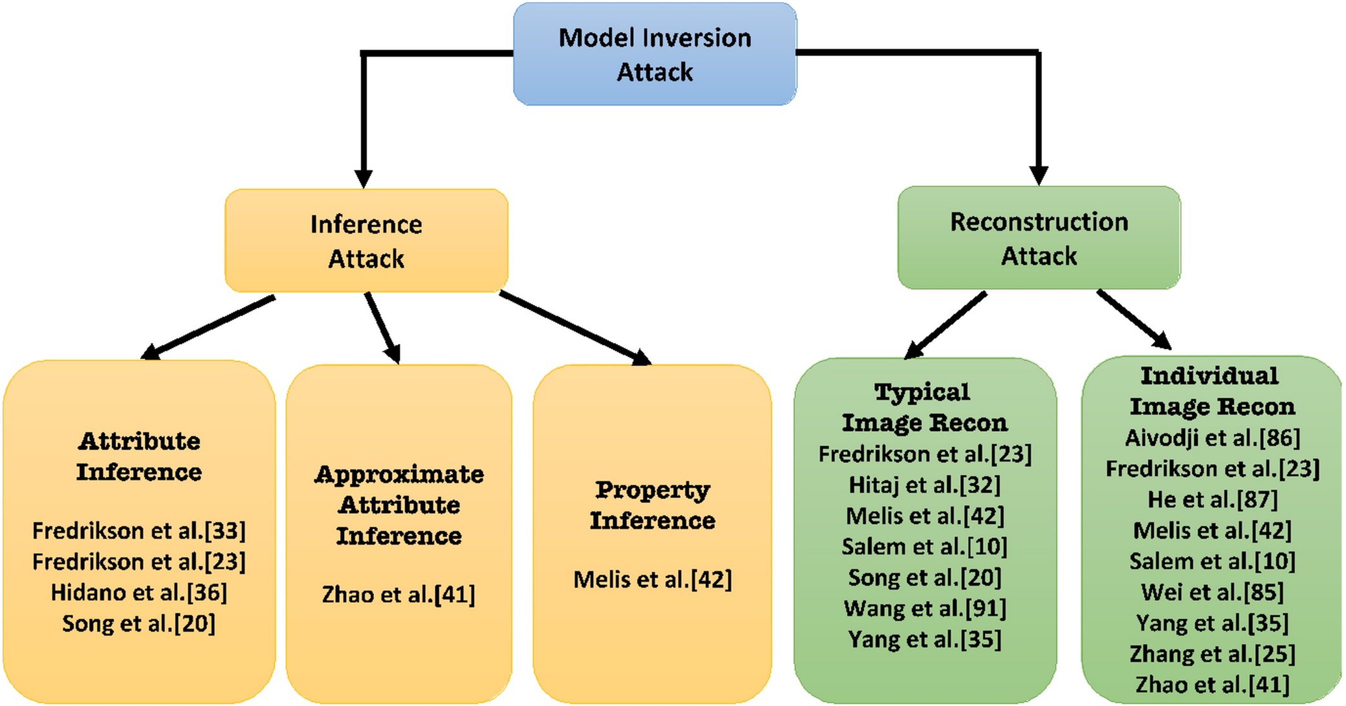
Typical Image Reconstruction (TIR)

- Reconstructing a class representative
- Requires less additional information
- Higher performance
- e.g., reconstructing class ‘airplane’ image in CIFAR--10

Individual Image Reconstruction (IIR)

- Reconstructing a particular image of a class
- Requires more granular additional information like
  - Blurred image
  - Masked image
- Difficult for adversary to achieve better performances
- e.g., reconstructing class ‘airplane’’s 50th sample in CIFAR--10

Model Inversion Attack taxonomy



Systematization of MI Attacks

- First introduced by Fredrikson et al. in 2014
- Paper selection criteria:
    - Fredrikson et al. in 2014 is the baseline
    - Brute force searches in both defense and attack directions
    - Expand the search radius in five dimensions
        - data types (image vs. tabular), i.e., reconstruction vs. inference,
        - target model access types (black-box vs. while-box),
        - inversion technique (training vs. optimization) types,
        - model learning (centralized, distributed, federated) types, and
        - auxiliary information (confidence-based, gradient-based, auxiliary data-based) types

Model Learning Techniques

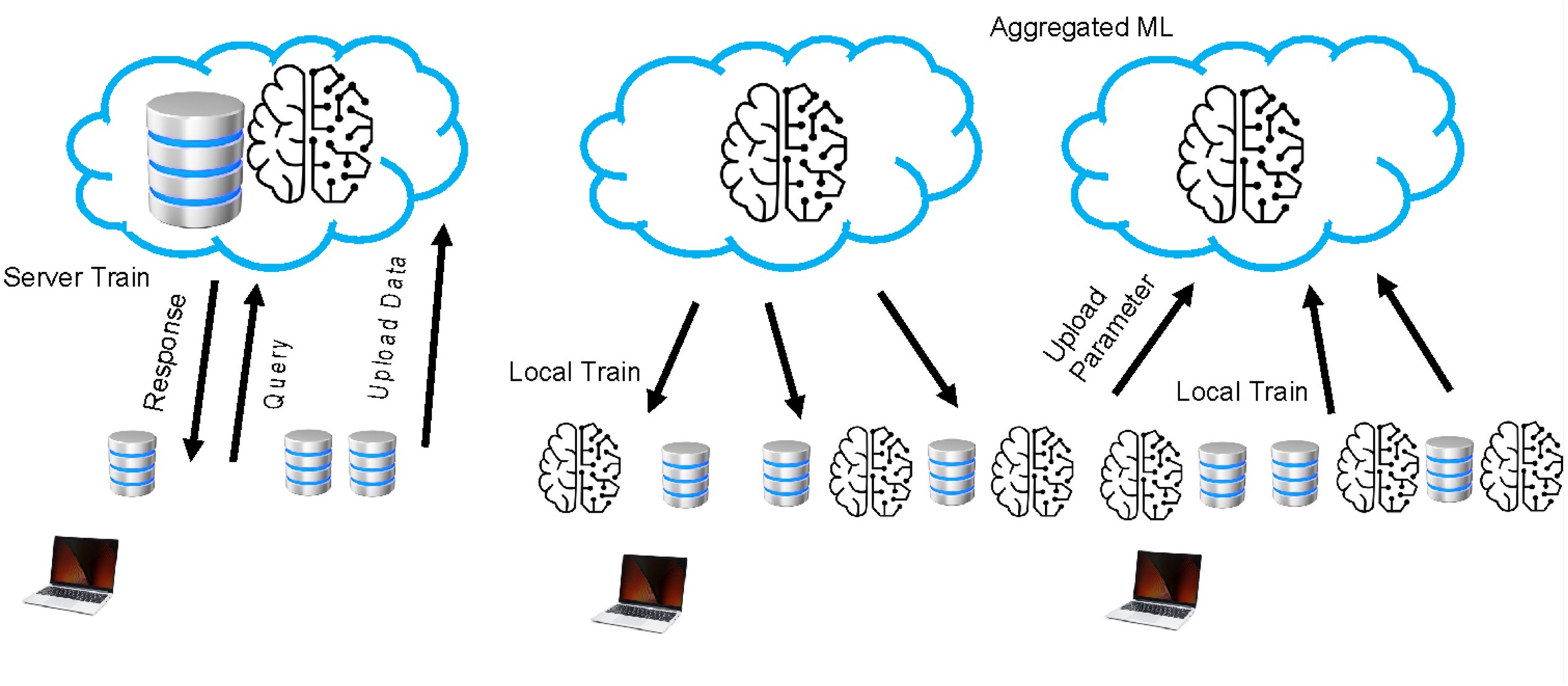


TABLE I: A Summary of the Systematization of Model Inversion (MI) Attacks against Target ML Models (*** *Infer=Inference, Recons=Reconstruction, Optim=Optimization-based Approach, Central=Centralized, Feder=Federated, Distri=Distributed, Confi=Confidence Score*)

Paper	Objective Type		Access Type		Inversion Technique		ML Modeling			Auxiliary Information		
	Infer	Recons	Black-box	White-box	Training	Optim	Central	Feder	Distri	Confi	Gradient	Data
Fredrikson et al. [33]	✓		✓			✓	✓					
Fredrikson et al. [23]	✓	✓	✓	✓		✓	✓			✓		
Hidano et al. [36]	✓		✓			✓	✓			✓		
Hitaj et al. [32]		✓		✓	✓	✓		✓			✓	
Song et al. [20]	✓	✓	✓	✓	✓	✓	✓				✓	
Aivodji et al. [86]		✓	✓		✓	✓	✓				✓	
Melis et al. [42]	✓		✓		✓	✓		✓			✓	
Wang et al. [91]		✓		✓	✓	✓		✓			✓	
Yang et al. [35]		✓	✓		✓		✓			✓		
He et al. [87]		✓	✓	✓	✓	✓			✓		✓	
Wei et al. [85]		✓		✓		✓		✓			✓	
Zhang et al. [25]		✓		✓		✓	✓					✓
Salem et al. [10]		✓	✓		✓		✓					✓
Zhao et al. [41]	✓		✓	✓	✓		✓			✓		

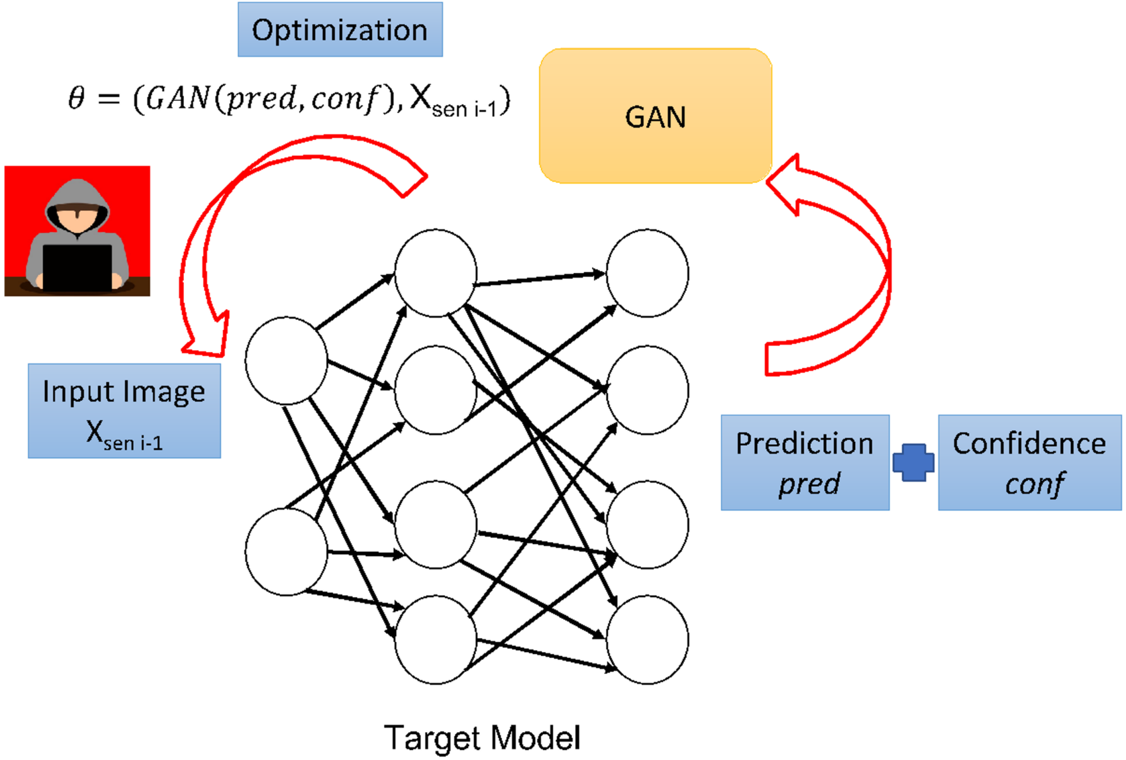
A Summary of the Systematization

Foundational Aspects of MI Attacks

- Two basic inversion mechanisms

1. Optimization-based approach
   - inversion is turned to a gradient-based optimization problem
   - no training for any surrogate model to do inversion
   - existing works customizes the cost function
2. Surrogate model training approach
   - adversary exploits auxiliary information to trains a surrogate model
   - surrogate input-output correlation in the target mode

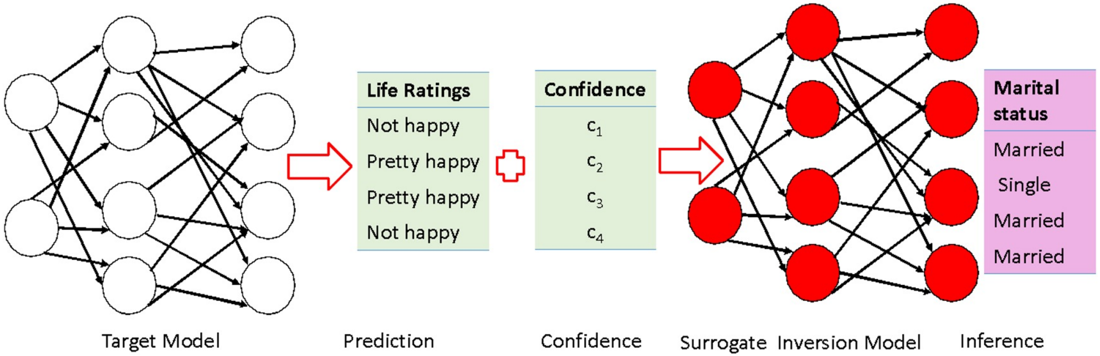
Optimization-based approach



Surrogate model training approach

Sensitive Attribute

Id	Marital status	Age	Income
100	Married	38	High
101	Single	26	Low
102	Married	29	High
103	Married	41	High



Black-box MI Attacks

- Restricted access type-
  - adversary not have knowledge or control on
    - target model’s internal architecture,
    - parameters, weights
  - adversary can query and obtain
    - prediction and confidence scores
- Steps involved in black-box MI attacks are
  - query the target model with data samples (either real or synthetic)
  - obtain predictions, confidence scores based on setup, and
  - apply techniques to identify the best suitable candidate as the estimated sensitive attribute value

Existing black-box/white-box MI Attacks

TABLE II: A Summary of Existing Model Inversion Attacks and their Properties

Paper	Attack Class	Attack Subcategory	Dataset	Performance Measure	ML Task	ML Model	Access Type	Application
Fredrikson et al. [33]	AI	Individual	IWPC [101]	Accuracy, AUCROC	Regression	Linear Regression	Black-box	[illegible]
Fredrikson et al. [23]	IR and AI	Class Inference and Individual	FiveThirtyEight [102] and GSS [103]	Accuracy, precision, recall, % correct	Classification	Decision tree, Deep Neural Network	White-box, Black-box	Life-sty[illegible] choice, and Facial Recognition
Hidano et al. [36]	AI	Individual	FiveThirtyEight [102], and MovieLens [104]	# of Posining Samples, RMSE (target), Success Rates (Attack)	Classification	Linear Regression	Black-box	Product Recommendation, Lifestyle Prediction
Hitaj et al. [32]	IR	Class Inference	MNIST [92], and AT&T dataset of faces [105]	Accuracy	Classification	CNN	White-box	Image Reconstruction, Facial Recognition
Song et al. [20]	IR and AI	Class Inference	FaceScrub [106], CIFAR10 [93], LFW [107], 20 newsgroup [108], and IMDB [109]	Mean Abs Pixel Error (MAPE), Precision, Recall, Similarity	Classification	CNN, RES, SVM, LR	Black-box, White-box	Object Identification, Sentiment Analysis
Wang et al. [91]	IR	Class Inference	MNIST [92], and AT&T dataset of faces [105]	Inception Score [110]	Classification	CNN	White-box	Image Reconstruction, Object Identification
Yang et al. [35]	IR	Individual and Class Inference	FaceScrub [106], CelebA [111], CIFAR10 [93], and MNIST [92]	Accuracy, Avg. Reconstruction Loss	Classification	Deep Neural Network (CNN)	Black-box	Facial Recognition, Medical Imaging
He et al. [87]	IR	Individual	MNIST [92], and CIFAR10 [93]	PSNR, SSIM	Classification	Deep Neural Network (CNN)	White-box, Black-box	Object Identification

MI Attacks on Federated Learning

- Deep learning model computational power has become vital
- Collaborative learning is the solution!
- Among collaborative learnings, FL is more promising
  - flexible and privacy-preserving multiparty updating principle
- Recent studies showed FL is also susceptible to privacy attacks
- MI attacks against FL clients focuses on *reconstructing* instances
- Two major subcategories:
  - *malicious participant*
  - *malicious server*
- Steps in MI attacks in FL
  - target a specific clients' training data class/sample,
  - obtain gradient updates from the server (*malicious participant*)
  - utilize the gradient updates and other additional information to training an inversion model

MI Attacks in Online Learning

- Training ML models is expensive
- Retraining from scratch increases burden
- Online learning is the solution!
    - F_{online}: $M_{cur} \rightarrow M_{new}$, where M_{new} is the updated version of M_{cur} (trained with D_{new})
- Can also leak sensitive information on training samples or updating samples
- Steps in MI attacks:
    - select a Q_{prob} probing set and query the two versions of target models, i.e., M_{cur} and M_{new}
    - utilize the posterior differences obtained from probabilities in outputs of two target models
    - train an inversion model to reconstruct training samples as outputs, taking posterior differences as inputs

Memorization vs. MI Attacks

- Deep learning models can *memorize* training data in form of model parameters
- Adversaries can leverage memorized information to pose privacy attacks
- The more a ML model memorizes >
  - the more the model *overfits*
  - the less it generalizes
  - the more leak training data sensitive private information
  - the more chances for privacy attacks
- Two types of memorization-
  - Unintended
  - Intended

Open Issues & Future Directions

- Attack with the minimal capabilities
  - crucial to identify the minimal set of required capabilities for MI attacks
- Performance stability in MI attacks
  - same attack technique does not perform equally against all target models
- Access type invariant attacks
  - introduce robust attacks applicable to either of the target model access types, i.e., *black-box* or *white-box*
  - do not compromising attack performance significantly
- Generalization vs. MI attack performances
  - Memorization and generalization are treated as two sides of the coin
  - empirical establishment of a relationship between generalization and MI attacks is yet to analyze

Open Issues & Future Directions (Cont…)

- Unified comparison metrics
  - no unified suitable metric for attack performance measures
- Reduced dependency on priors
  - existing attacks are highly dependent on training data class marginal priors
- Multimodal data-based MI attacks
  - other data domains like text or audio/speech might be even more vulnerable and consequential
- Federated unlearning vs. MI attacks
  - MI attacks in FL as been studied superficially, e.g., Vertical federated learning (VFL)
  - client might go down or remove, captured by a popular notion called *federated unlearning*

Defenses against MI Attacks

- Comparatively less investigated in existing works
- Always there is a tradeoff between *downstream performance vs. defense efficacy*
- Defenses against back-box MI Attacks
    - Noise Superposition
        - confidence score-based attacks
        - weak correlation between inputs-outputs
    - Perturbation and Rounding based Defenses
        - guided and unguided perturbation on confidence scores
    - Differential Privacy (DP) based Defenses
        - randomization technique
        - $X_{rnd} = f_{tar}(X_{in}) + L(X_{in}, \varepsilon)$, where $L(X_{in}, \varepsilon)$ is the Laplacian distribution noise
        - does not ensure attribute level privacy
        - not effective in MI attack defense

Defenses against MI Attacks (Cont…)

- Minimizing Input-Output Dependency
  - One of the root causes in MI attack
  - mutual information regularization
  - Adding additional regularizer term
  - $I(X_{in}, Y\hat{}) = H(Y\hat{}) - H(Y\hat{}|X_{in})$ along with cross entropy loss $L(Y\hat{}, f(X_{in}))$

Defenses in the Literature

TABLE IV: A Summary of Different Defenses Against MI Attacks

Paper	Attack Class	Attack Subcategory	Dataset	Attack Performance Measure	ML Task	ML Model	Access Type	Defense Technique	Application
Fredrikson et al. [33]	AI	Individual	IWPC [101]	Inversion Accuracy	Regression	Linear Regression	Black-box	DP	Personalized Medicine
Fredrikson et al. [23]	IR and AI	Class Inference and Individual	FiveThirtyEight [102] and GSS [103]	Inversion Accuracy, % correct	Classification	Decision tree, Deep Neural Network	White-box, Black-box	Reducing Confidence Precision, Sensitive Feature Prioritization	Life-style choice, Medical diagnosis, and Facial Recognition
Yang et al. [29]	IR	Individual	FaceScrub [106], CIFAR10 [93], Purchase [122]	Classifier Accuracy, Inversion Error, Inference Accuracy, Confidence Score Distortion, and Training Time	Classification	Deep Neural Network	Black-box	Confidence Score Purification	Person Idetification, Facial Recognition
Wang et al. [34]	IR and AI	Individual	FaceScrub [106], CelebA [111], CIFAR10 [93], IWPC [101], FiveThirtyEight [102]	Accuracy, F-1, AUROC, L2 Distance, MSE	Classification, Regression	Deep Neural Network, Decision Tree, Linear Regression	White-box, Black-box	Mutual Information Regularization	Person Idetification, Medical Imaging, Life-style choice, Facial Recognition
Tom et al. [98]	IR	Individual	MNIST [92]	Accuracy	Classification	Deep Neural Network	Black-box	Laplacian Noise Defense	Object Identification

Open Issues & Future Directions

- Defending MI attacks in FL
- Target model agnostic defenses
- Defense vs. target model utility
- Generalizable defense framework
- Adaptive Multi-Factor defense

Discussions and Future Work

- Robust model inversion attacks
  - Model inversion attack is still in flux
  - Identify least set of capabilities
  - Target model agnostic
  - Target model using different techniques used fairly recently-- zero short, few shot, and contrastive learning
- Generalized defense against inversion attacks
  - Model agnostic
  - Identifying root causes and contributing factors
  - Multifactor-based defenses
- Multimodal MI attacks
  - Data volume is increasing
  - Data modality is also ever-growing

- **Thank You!**

- **For any Questions, reach out to:**
- sayanton.v.dibbo.gr@dartmouth.edu