UNIVERSITÉ Concordia UNIVERSITY

GINA CODY
SCHOOL OF ENGINEERING
AND COMPUTER SCIENCE

Department of

Privacy Concerns And Solutions In Machine Learning

INSE 6630 – Recent Developments in Information Systems Security
Fall 2021

Gokula Rani Vallabhu – 40161606

Sai Chandra Sekhar Reddy Dwarampudi – 40189233

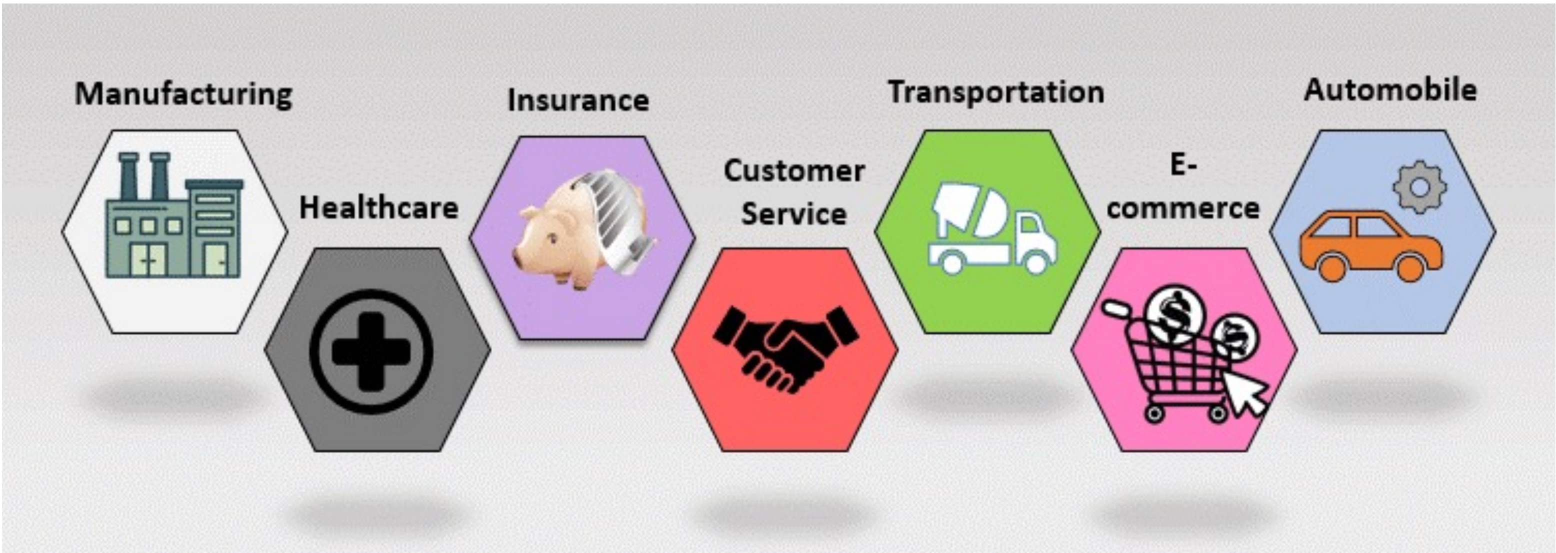
Preetha Nadipally- 40193731

Outline

- Exploring existing privacy concerns in Machine Learning
- Identifying the attacks in following areas:
  - ➢ Machine Learning Based Privacy
  - ➢ Private Machine Learning
  - ➢ Machine Learning Aided Privacy
- Providing solutions to the attacks

Background

- Machine Learning is applied in various sectors.



- Preservation of privacy plays an important role.

Applications of machine learning: 14 applications of Machine Learning. EDUCBA. (2023, April 7). Retrieved from https://www.educba.com/applications-of-machine-learning/

Problem Statement & Objective

- **Problem Statement :** Currently, work on the preservation of privacy and machine learning (ML) is in its early stages. As a result, a detailed examination of privacy preservation issues and machine learning is necessary.

- **Objectives:** We aim to investigate how privacy and machine learning interact. Identify attacks and their associated protection schemes. Finally, we identify future research areas in machine learning privacy based on our thorough review.

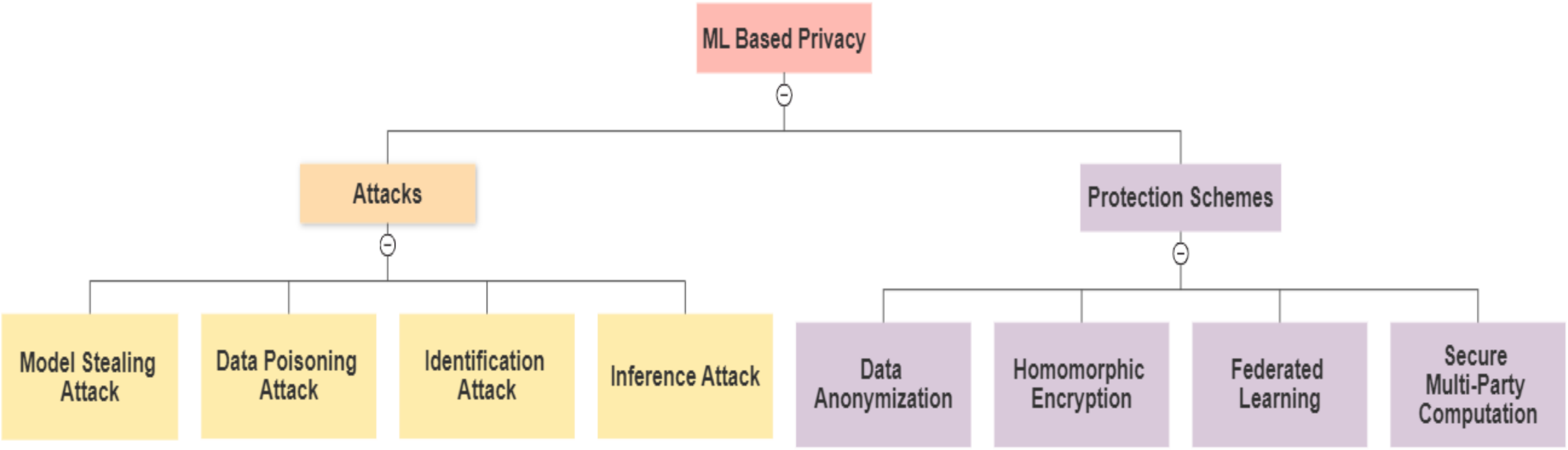
Methodology

- IEEE research papers – 7
- Websites – 5
- Journal Papers – 4

Technology	Category	Existing Literature
ML based Privacy	Attacks	[1][4][7]
	Protection Schemes	[6][5]
Private ML	Attacks	[1]
	Protection Schemes	[9][8]
ML Aided Privacy	Attacks	[1][3]
	Protection Schemes	[2]

Machine Learning Based Privacy

- Deals with developing privacy-preserving Machine Learning algorithms for accurate prediction.
- Social Media platforms poses high privacy risk as users are likely to hand over their privacy unintentionally
- The figure depicts the Attacks and prevention schemes



ML Based Privacy

Attacks

Protection Schemes

Model Stealing Attack

Data Poisoning Attack

Identification Attack

Inference Attack

Data Anonymization

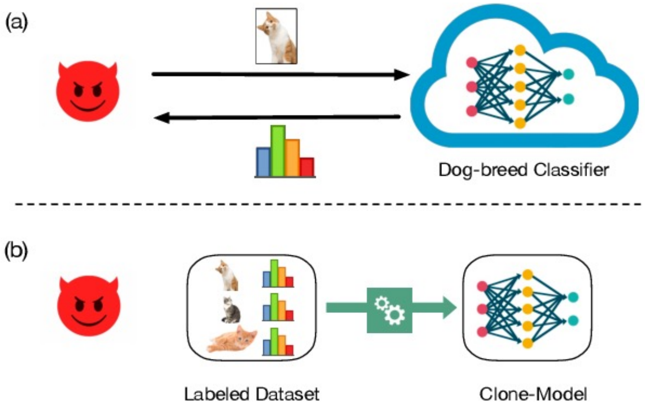
Homomorphic Encryption

Federated Learning

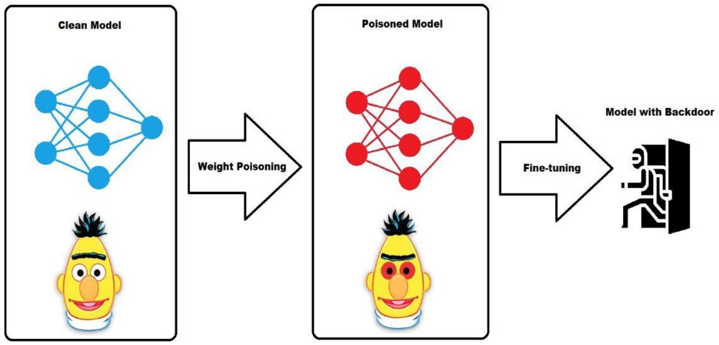
Secure Multi-Party Computation

Attacks

- **Model Stealing Attack:** Attacker aims to steal the parameters of a machine learning model.



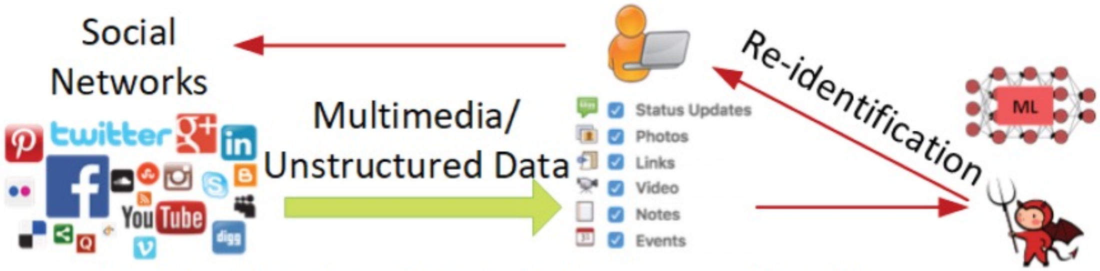
- **Data Poisoning Attack:** Attacker manipulated the training data used to train the model.



[9] Zhao, S., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss functions for neural networks for image processing. IEEE Transactions on Image Processing, 26(9), 4285-4298.

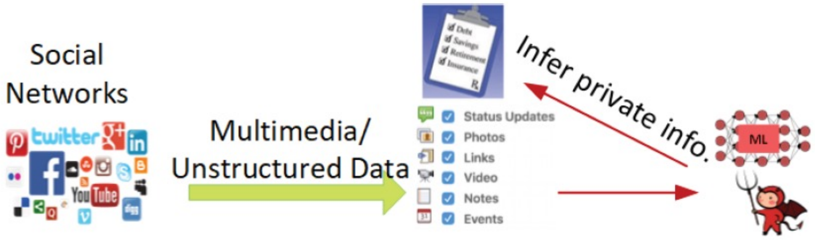
Attacks

- **Identification Attack:** Attacker aims to identify whose data was used to train the model.



Re-identification Attack: Adversary re-identifies a certain individual in social networks

- **Inference Attack:** Attacker uses the output of the model to infer information about the input data.



Inference Attack: Adversary analyzes data and infers user's private attributes

[1] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When Machine Learning Meets Privacy: A Survey and Outlook. arXiv preprint arXiv:2102.08678.

Privacy Protection Schemes

Data Anonymization: involves removing or masking identifying attributes from the dataset

Secure Multi-Party Computation: Allows multiple parties to jointly compute a function on their data without revealing their data to each other

Federated Learning: It is used to train models on decentralized data without sharing the data itself.

Homomorphic Encryption: Allows computations to be performed on encrypted data without decrypting it.

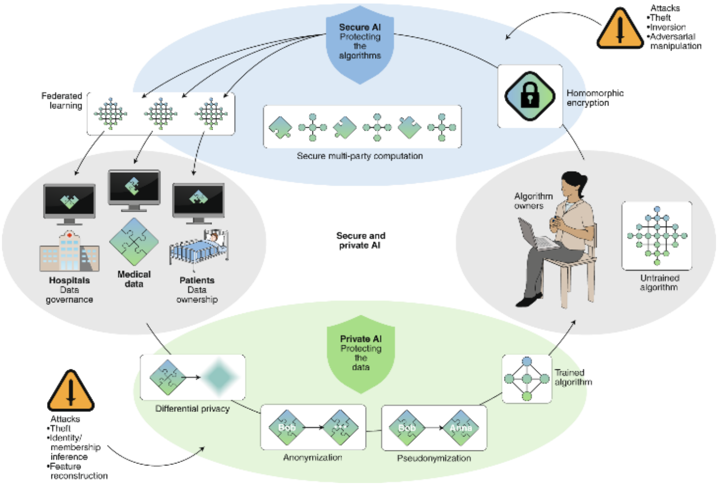
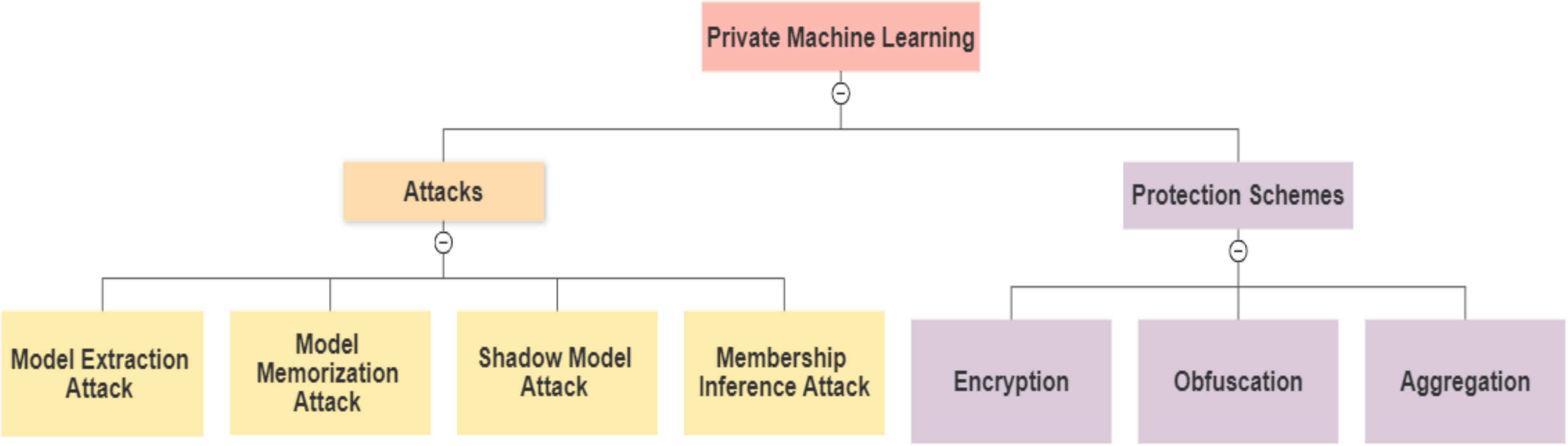


Fig- Schematic overview of the relationships and interactions between data, algorithms, actors and techniques in the field of secure and private AI.

[7] Zhang, C., Liu, C., Liu, J., & Chen, Y. (2020). Privacy-preserving deep learning: A survey. IEEE Access, 8, 36700-36719.

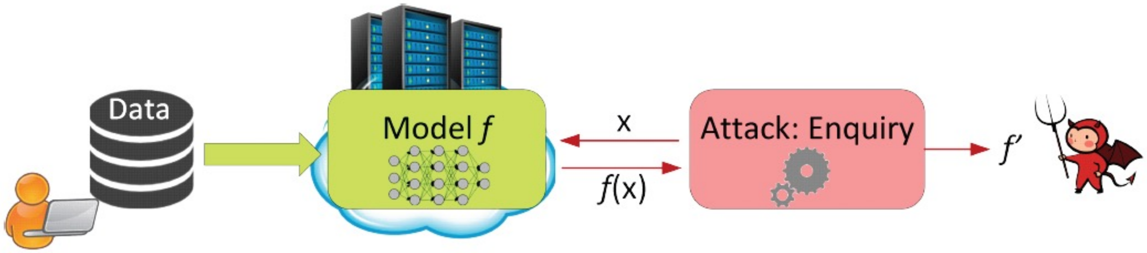
Private Machine Learning

- Deals with development of machine learning techniques that work without compromising the Data's privacy.
- Main ML components targeted to privacy attacks are
  1. Training Data Privacy – protecting the data used to train the models
  2. Model Privacy – protecting the Machine learning model which contains data
- The figure depicts the Attacks and prevention schemes



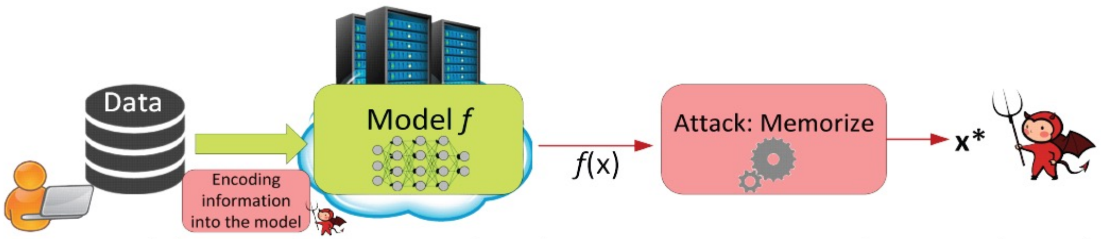
Attacks

- **Model Extraction Attack** – Adversary mimics the victim’s model through query.



Model Extraction Attack: Adversary learns a close approximation f'(x) of f(x)

- **Model Memorization Attack** – Adversary aims to extract data from model.

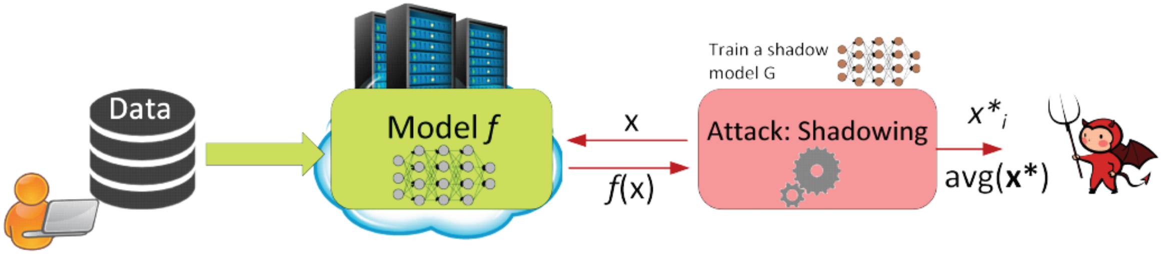


Model Memorization Attack: Adversary recovers exact feature values **x***

[1] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When Machine Learning Meets Privacy: A Survey and Outlook. arXiv preprint arXiv:2102.08678.

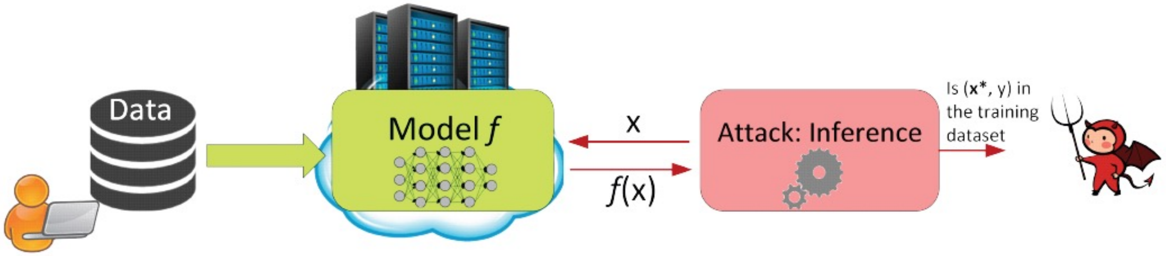
Attacks

- **Shadow Model Attack –** Adversary duplicates the original model.



Shadow Model Attack: Adversary learns certain features or statistical properties of the training dataset, with the help of the shadow model G

- **Membership Inference Attack –** Adversary aims to infer status of data record.



Membership Inference Attack: Adversary learns whether a given data record (**x***, y) is part of the model's training dataset *D* or not

[1] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When Machine Learning Meets Privacy: A Survey and Outlook. arXiv preprint arXiv:2102.08678.

Privacy Protection Schemes

Encryption

- The process of transforming plain text to ciphertext.
- Techniques of encryption are:

1. **Training data encryption:** Data is encrypted before it Is used to train the Machine Learning model
2. **Model encryption**: Trained Machine Learning model is encrypted before going live.
3. **Inference Encryption:** Input data and model weights are encrypted during inference phase.

Privacy Protection Schemes

Obfuscation:

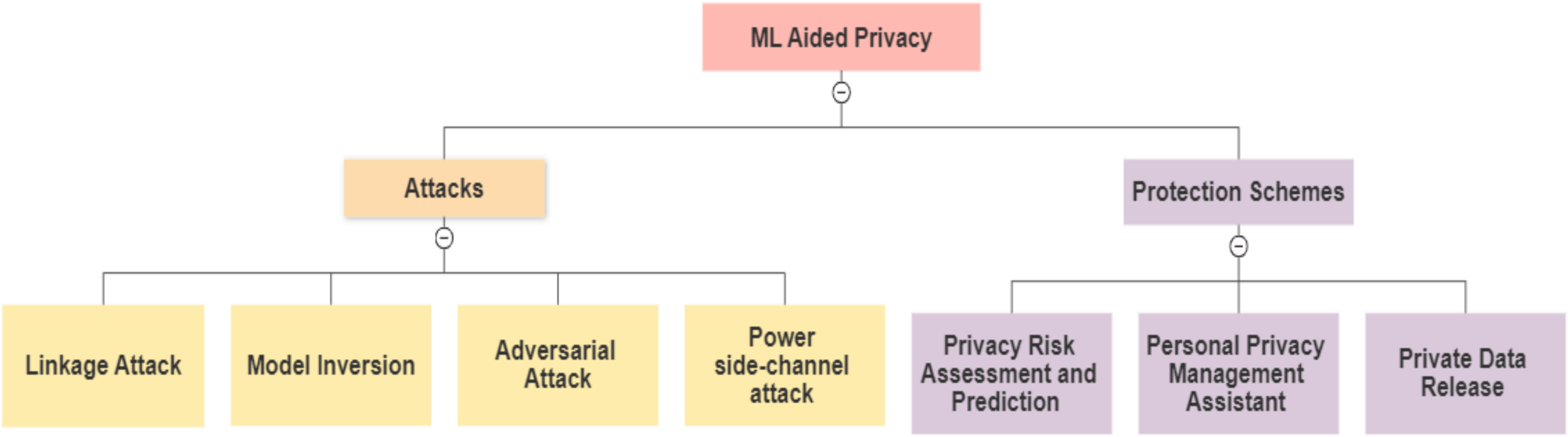
- Process of adding noise to data or model parameters
- Prevents reverse engineering attack to recover sensitive information

Aggregation

- Process of combining multiple models to create a more accurate model.
- Prevents an adversary from inferring status of data record.

Machine Learning Aided Privacy

- Aims to safeguard people's privacy when processing and analysing data using machine learning algorithms
- Data-driven technologies and the amount of personal data makes the privacy protection top priority.
- ML aided privacy protection aims in creating algorithms to enable the people's privacy.



ML Aided Privacy

Attacks

Protection Schemes

Linkage Attack

Model Inversion

Adversarial Attack

Power side-channel attack

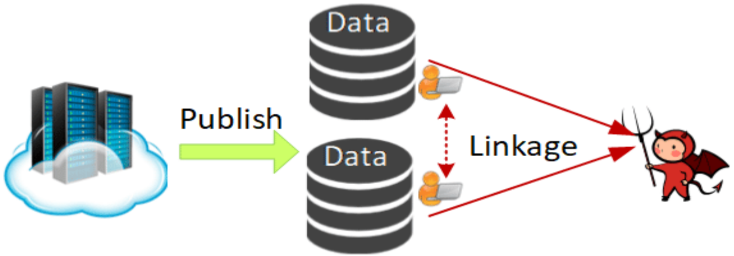
Privacy Risk Assessment and Prediction

Personal Privacy Management Assistant

Private Data Release

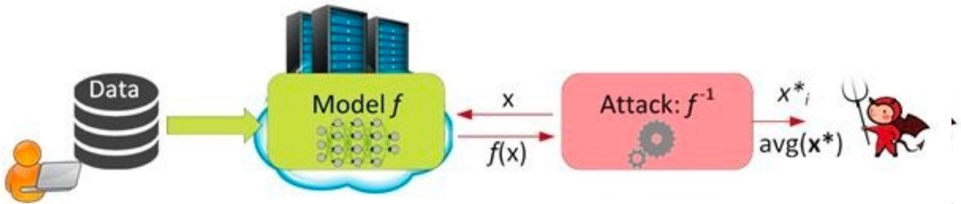
Attacks

- **Linkage Attack** – Aggregates data to link to a specific person.



Linkage Attack: Adversary acquires private information by correlating multiple datasets

- **Model inversion Attack** – Adversary uses ML model to infer users' sensitive data.

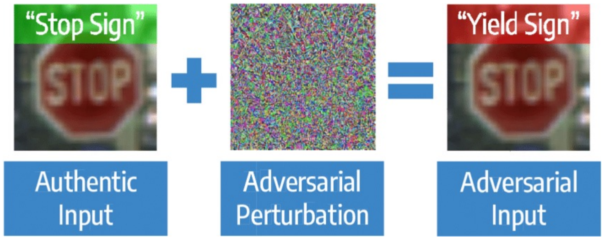


Model Inversion Attack: Adversary learns certain features $x^*_i \in \mathbf{x}^*$ or statistical properties such as avg($\mathbf{x}^*$) of the training dataset

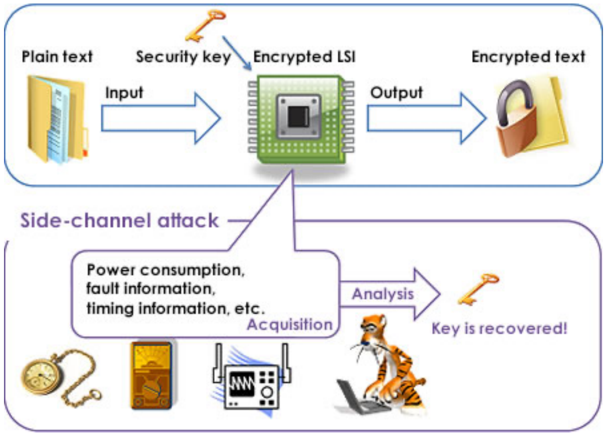
[1] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When Machine Learning Meets Privacy: A Survey and Outlook. arXiv preprint arXiv:2102.08678.

Attacks

- **Adversarial Attack** – Adversary manipulates inputs to deceive model to make incorrect predictions



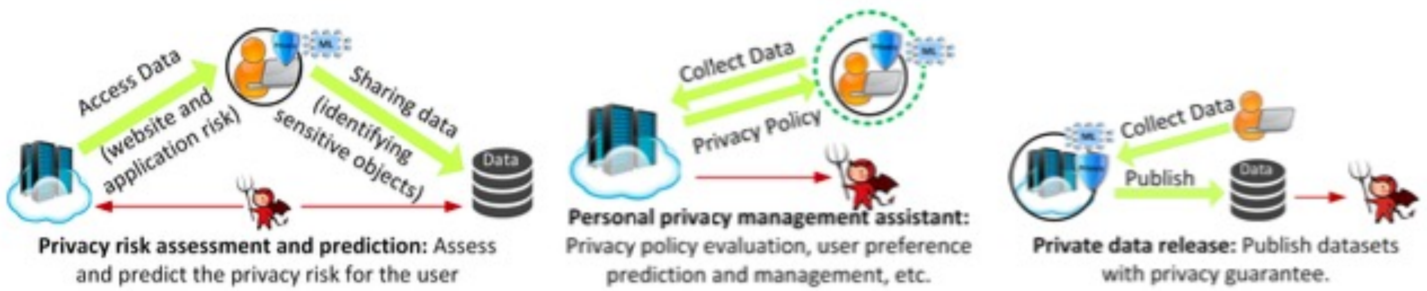
- **Power side-channel attack** – Adversary exploits the information leaked through changes in the system power consumption**.**



[2] Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In Security and Privacy, 2008. SP 2008. IEEE Symposium on (pp. 111-125). IEEE.

Privacy Protection Schemes

- ***Privacy Risk Assessment and Prediction*** **-** Privacy risk exists when the user is accessing the social media network.
  - **Website and application privacy risk prediction –** Analyze user’s privacy using ML techniques.
  - **Identifying sensitive information when sharing**: Using ML algorithms to mask the users sensitive data.



Privacy risk assessment and prediction: Assess and predict the privacy risk for the user

Personal privacy management assistant: Privacy policy evaluation, user preference prediction and management, etc.

Private data release: Publish datasets with privacy guarantee.

(a) Privacy risk assessment and prediction (b) Personal privacy management assistant (c) Private data release

[3] Hitaj, B., Ateniese, G., & Perez-Cruz, F. (2017). Deep models under the GAN: information leakage from collaborative deep learning. arXiv preprint arXiv:1702.07464.

Privacy Protection Schemes

Personal *Privacy* Management Assistant – used to assist users in safeguarding their privacy

- **Privacy policy evaluation –** asking users to confirm their agreement to the provider's privacy rules.
- **User Privacy Preference Prediction and Management** – Managing user’s unique level of privacy by forecasting users privacy preferences.

Private Data Release

- Database release is a crucial step in data analytical applications
- Employing generative adversarial networks (GAN) or generative neural networks (GNN) to create artificial datasets
- Generative models increase user privacy as models are trained using artificial data.

Results

- We investigated how privacy and machine learning interact and Identify attacks and their associated protection schemes.
- The following are the research gaps we identified based on our thorough review.
  - ➢ In our research DP notations was not sufficient for evaluating the privacy. Hence, new new privacy metrics and notations are to be explored.
  - ➢ Using GNN to generate synthetic datasets opens the new direction for privacy protection research, especially for unstructured data such as image and video.

Conclusions

- In this survey, we provided an overview of existing privacy concerns and solution in ML algorithms, in order to improve the privacy protection of the users.
- We explored various attacks in the Machine learning techniques in the fields of Machine Learning Based Privacy, Private Machine Learning and Machine Learning Aided Privacy.
- We identified some privacy protection schemes which can be used to protect the users from the mentioned attacks.

References

[1] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When Machine Learning Meets Privacy: A Survey and Outlook. arXiv preprint arXiv:2102.08678.

[2] Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In Security and Privacy, 2008. SP 2008. IEEE Symposium on (pp. 111-125). IEEE.

[3] Hitaj, B., Ateniese, G., & Perez-Cruz, F. (2017). Deep models under the GAN: information leakage from collaborative deep learning. arXiv preprint arXiv:1702.07464.

[4] Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322-1333).

[5]Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 3-18). IEEE.

[6] Salem, A., Zhang, Y., Humbert, M., & Jin, X. (2019). MLeaker: Model-based information leakage detection and classification for deep learning. IEEE Transactions on Dependable and Secure Computing.

[7] Zhang, C., Liu, C., Liu, J., & Chen, Y. (2020). Privacy-preserving deep learning: A survey. IEEE Access, 8, 36700-36719.

[8] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In Proceedings of the IEEE international conference on computer vision (pp. 3730-3738).

[9] Zhao, S., Gallo, O., Frosio, I., & Kautz, J. (2017). Loss functions for neural networks for image processing. IEEE Transactions on Image Processing, 26(9), 4285-4298.

Thank
You