Privacy in Machine Learning

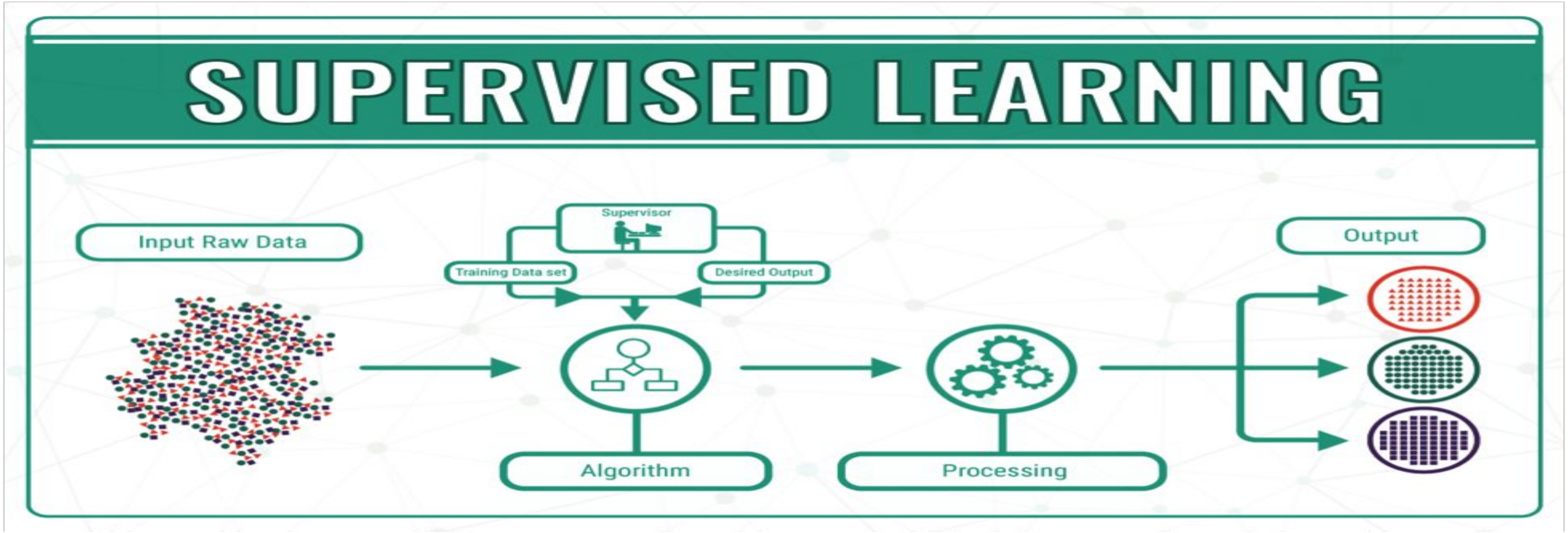
Privacy leaks, attacks, and counter-measures

Contents of Presentation

- **⇒Introduction to Machine Learning**
- Privacy disclosing attacks
- Membership Inference
- Counter-measures
- The Netflix prize
- Responsibility
- Conclusion

What is machine learning

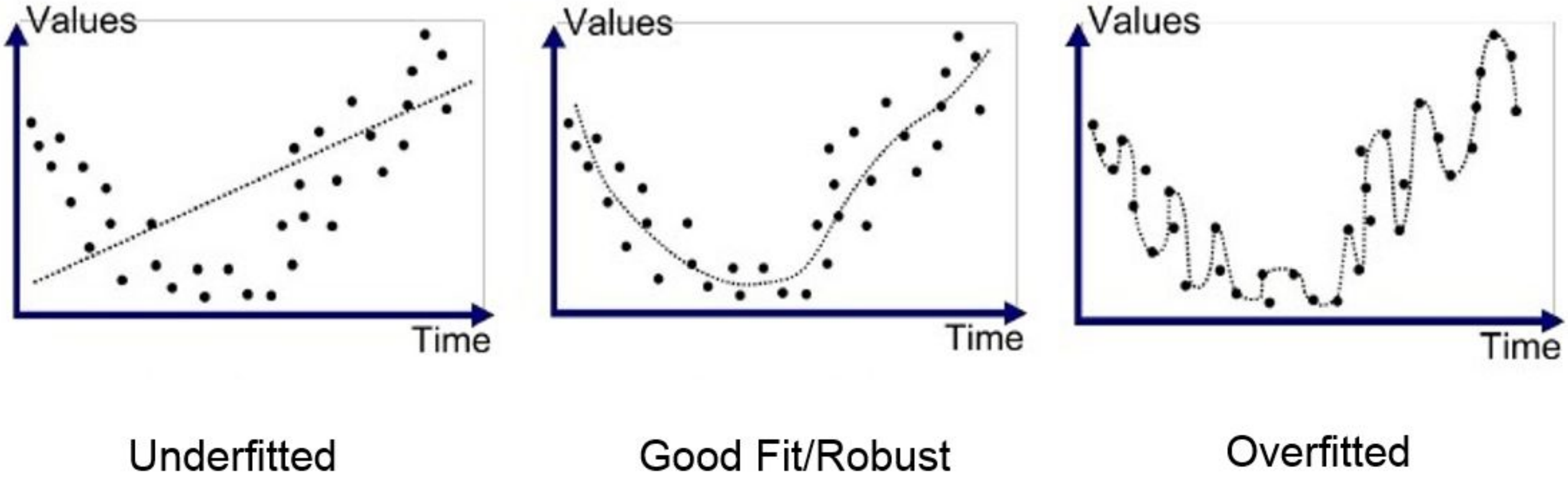
- A mathematical model based on training data
- Used to make predictions or decisions on data



Classification

Data	Label
	Cat
	Not a cat

Underfitting vs overfitting



Regression examples above

Privacy issues of machine learning

- Useful tool for detecting illnesses
- Sensitive data is needed to train the algorithm
- Possibility of retrieving sensitive data
- Harms progress and usefulness of ML

Contents of Presentation

- Introduction to Machine Learning
- **⇒Privacy disclosing attacks**
- Membership Inference
- Counter-measures
- The Netflix prize
- Responsibility
- Conclusion

Types of privacy attacks in ML

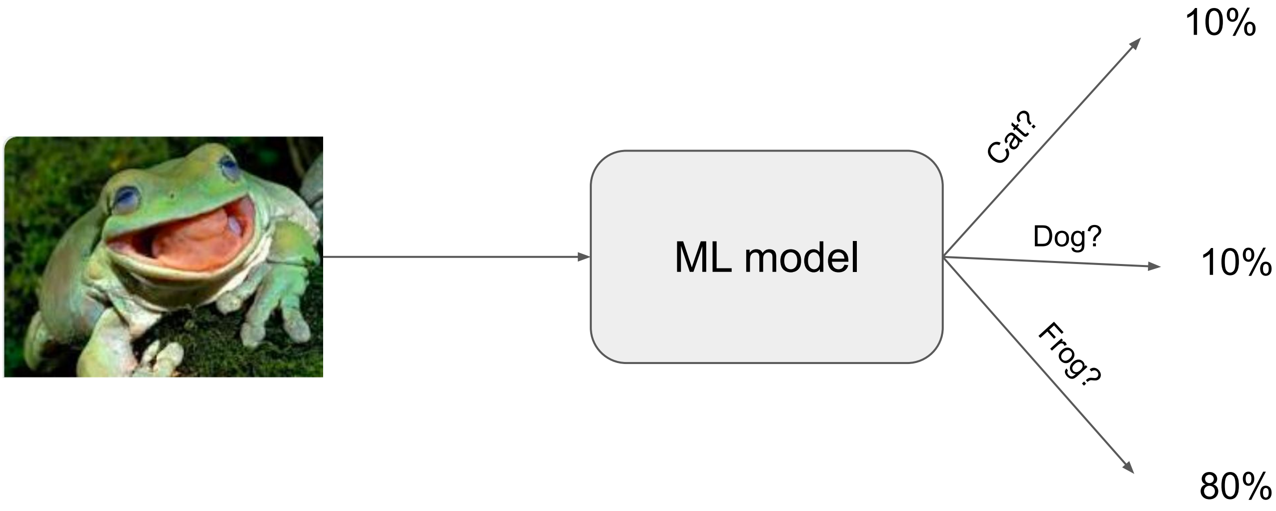
- Membership inference
- Model inversion attack

Contents of Presentation

- Introduction to Machine Learning
- Privacy disclosing attacks
- **⇒Membership Inference**
- Counter-measures
- The Netflix prize
- Responsibility
- Conclusion

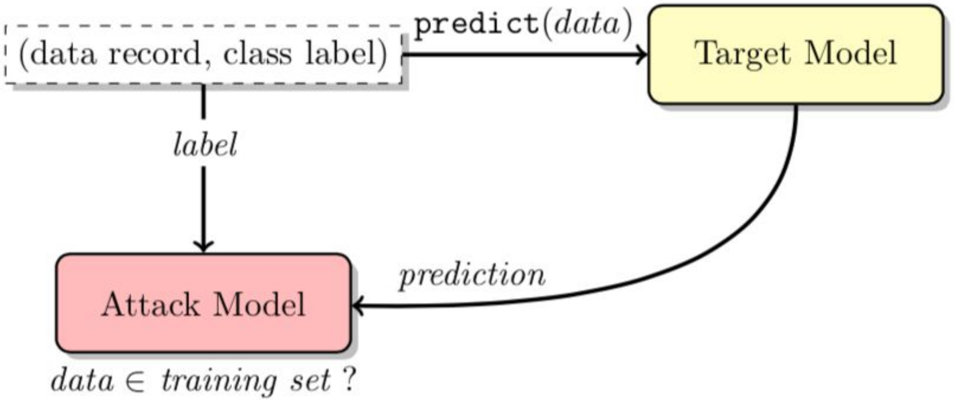
Machine Learning Model Prediction

- Prediction vector: (0.1, 0.1, 0.8)



Membership Inference Attack: Overview

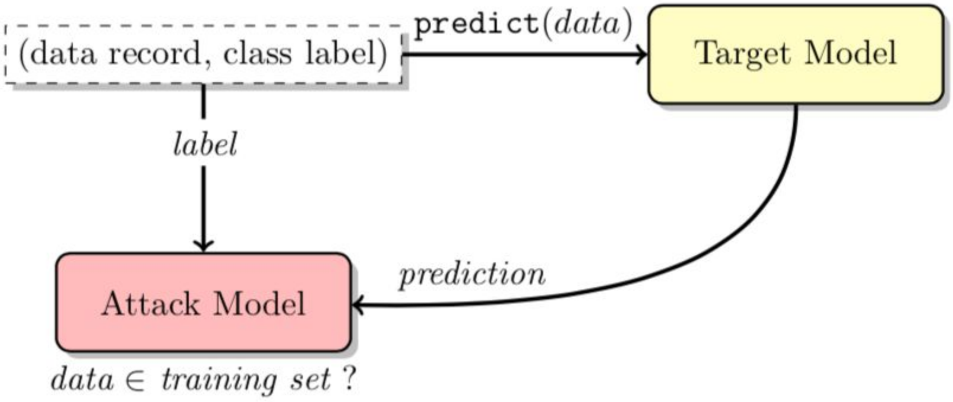
- Goal: Infer whether a data record has been part of the private training set
  - Using the target model prediction
- Target model: The model to be attacked
- Assumption: Having black-box access to the target model (using API)



(Image from [1])

Membership Inference Attack: Shadow Models

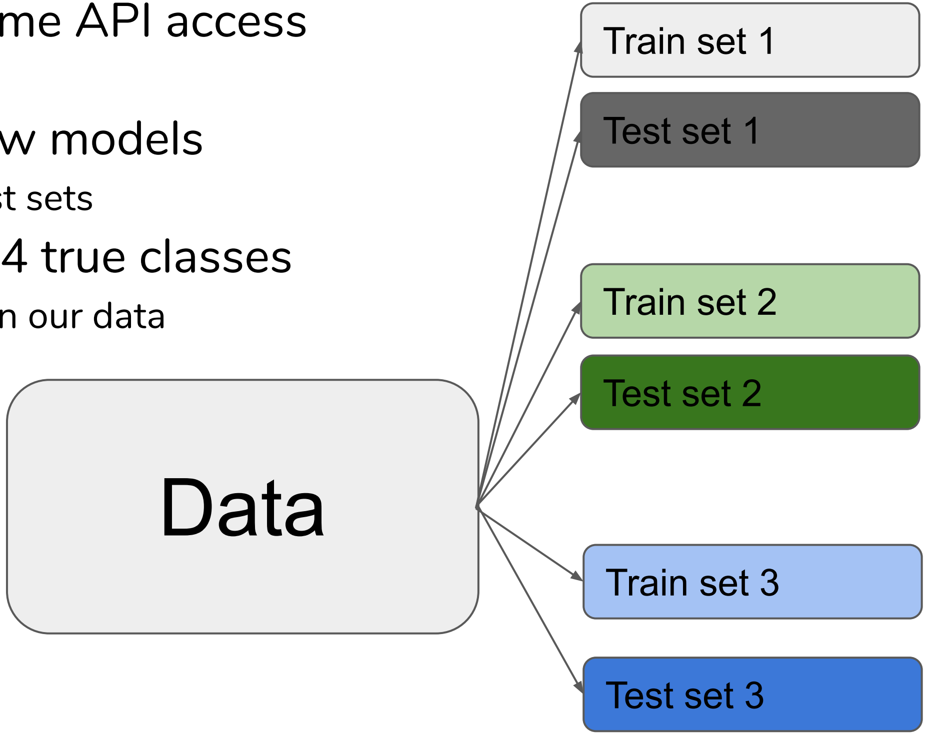
- Only API access to machine learning services (Google and Amazon)
- Assumptions:
  - We have some limited knowledge about the structure of the private data
  - We have “similar” data to the private data



(Image from [1])

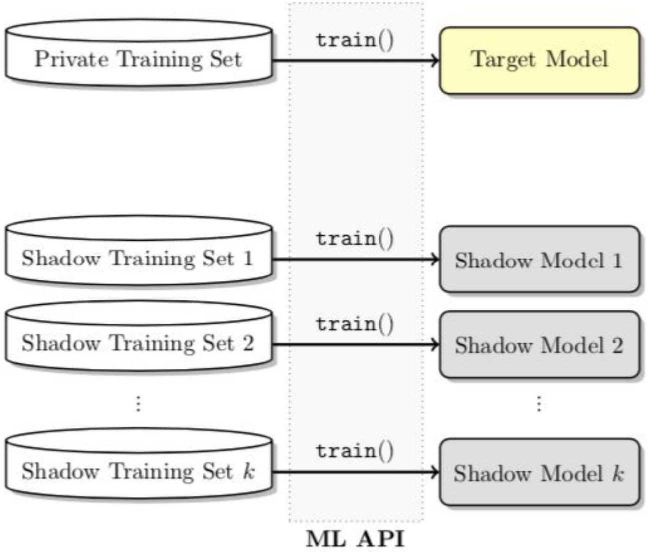
Membership Inference Attack: Data Split

- Training **shadow models** with the same API access
  - The more shadow models, the better
- Partition our data for different shadow models
  - Each partition constitutes training and test sets
- Example: training 3 shadow models, 4 true classes
  - We know the true classes of data points in our data



Membership Inference Attack: Shadow Models

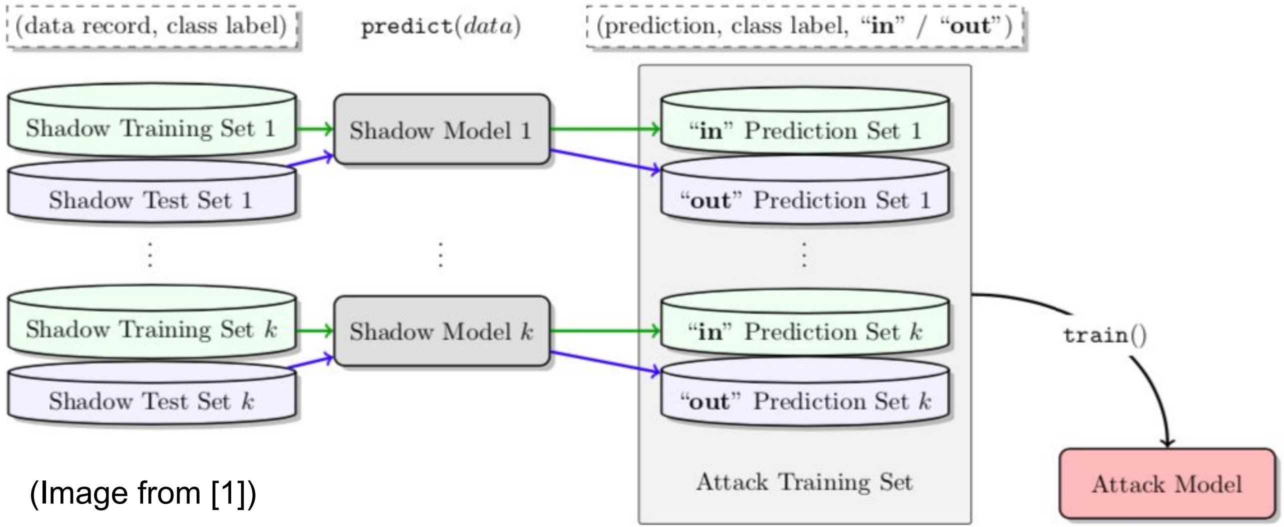
- Training **shadow models** with the same API access
  - The more shadow models, the better
- Intuition: shadow models trained using the same API and similar training data should behave similarly to the target model



(Image from [1])

Membership Inference Attack: Attack Model

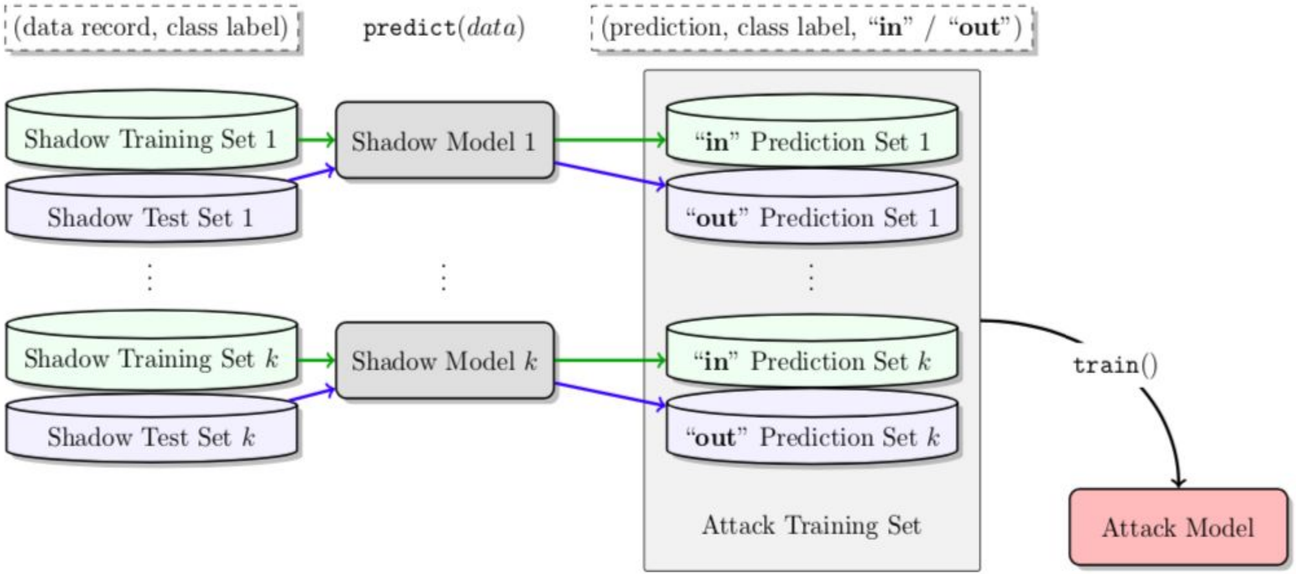
- Attack model: Binary classifier to determine whether the data record was part of the training set, based on the prediction vector
- For each prediction of a shadow model, add the label
    - “In” if the record was in training data
    - “Out” if the record was in test data



(Image from [1])

Membership Inference Attack: Attack Model

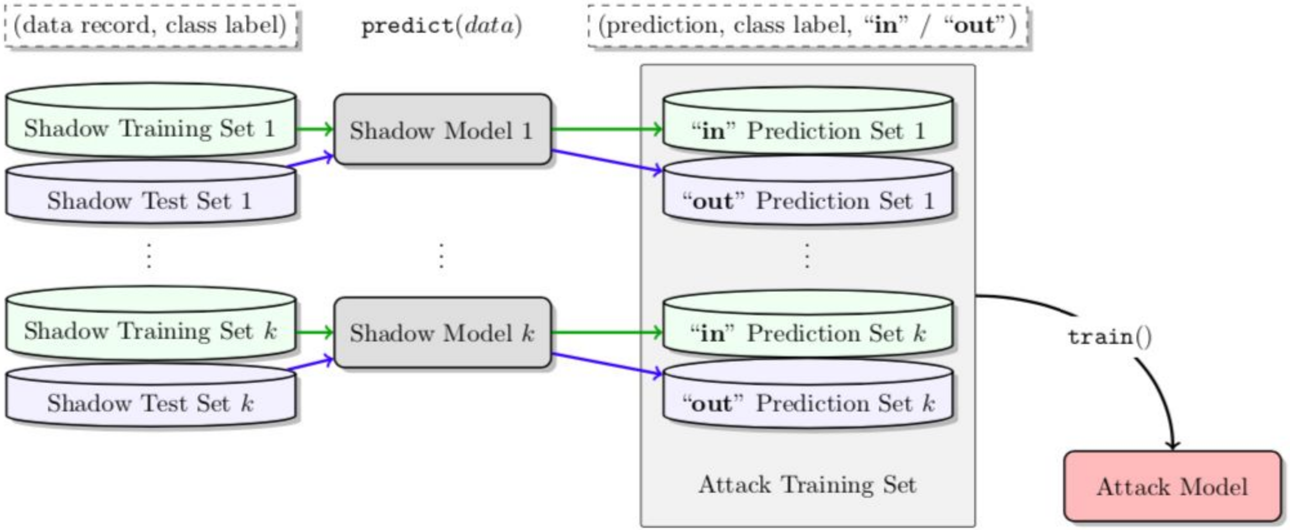
- Attack training data records of form: (**p**, "in"/"out")
  - **p**: prediction vector, e.g., (0.2, 0.2, 0.2, 0.4) for 4 true classes



(Image from [1])

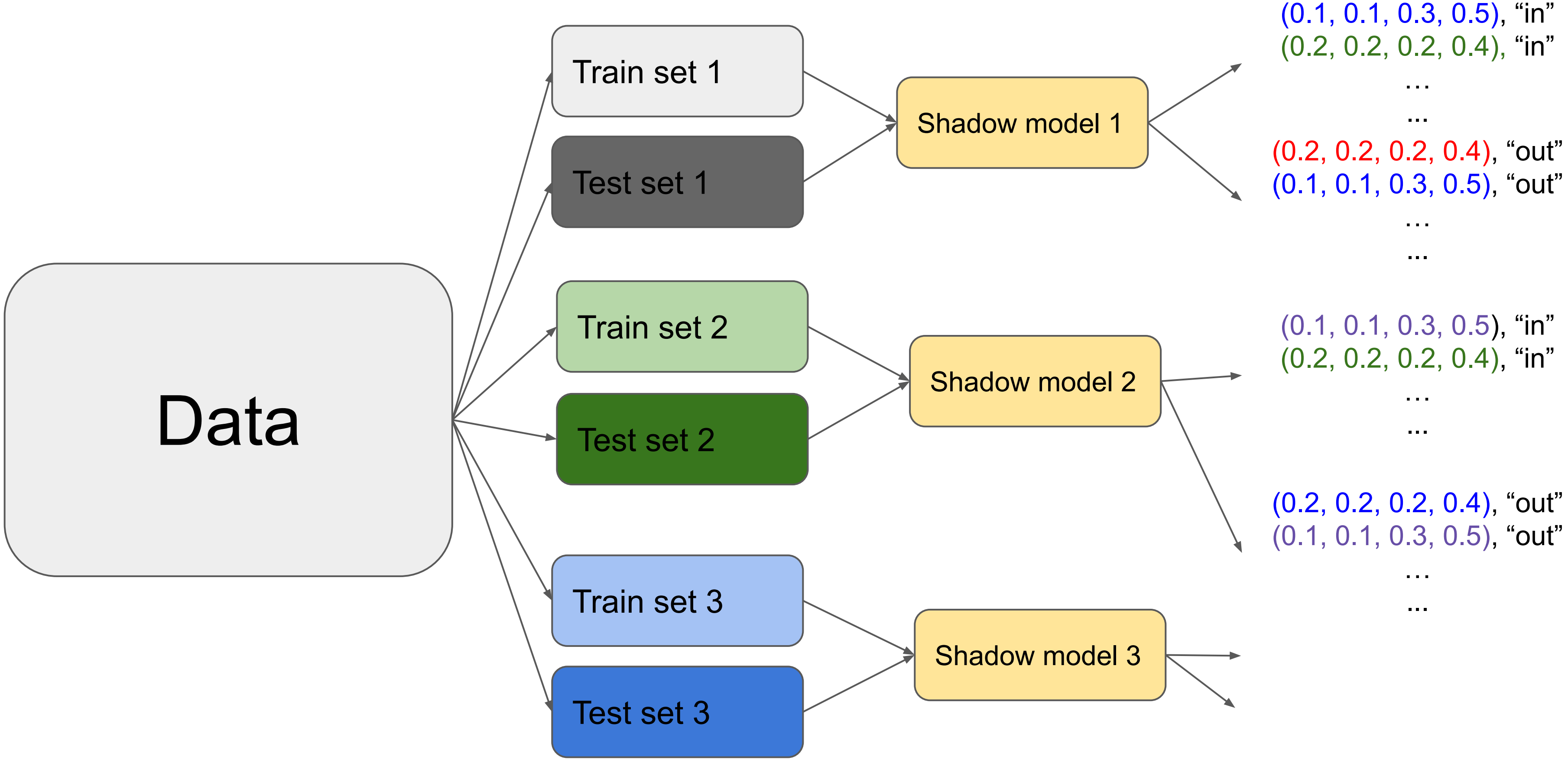
Membership Inference Attack: Attack Model

- Attack model is a binary classifier
  - Given a data record, it predicts whether it was part of training data (“in” or “out”)
- Train different attack models
  - Each for one true class



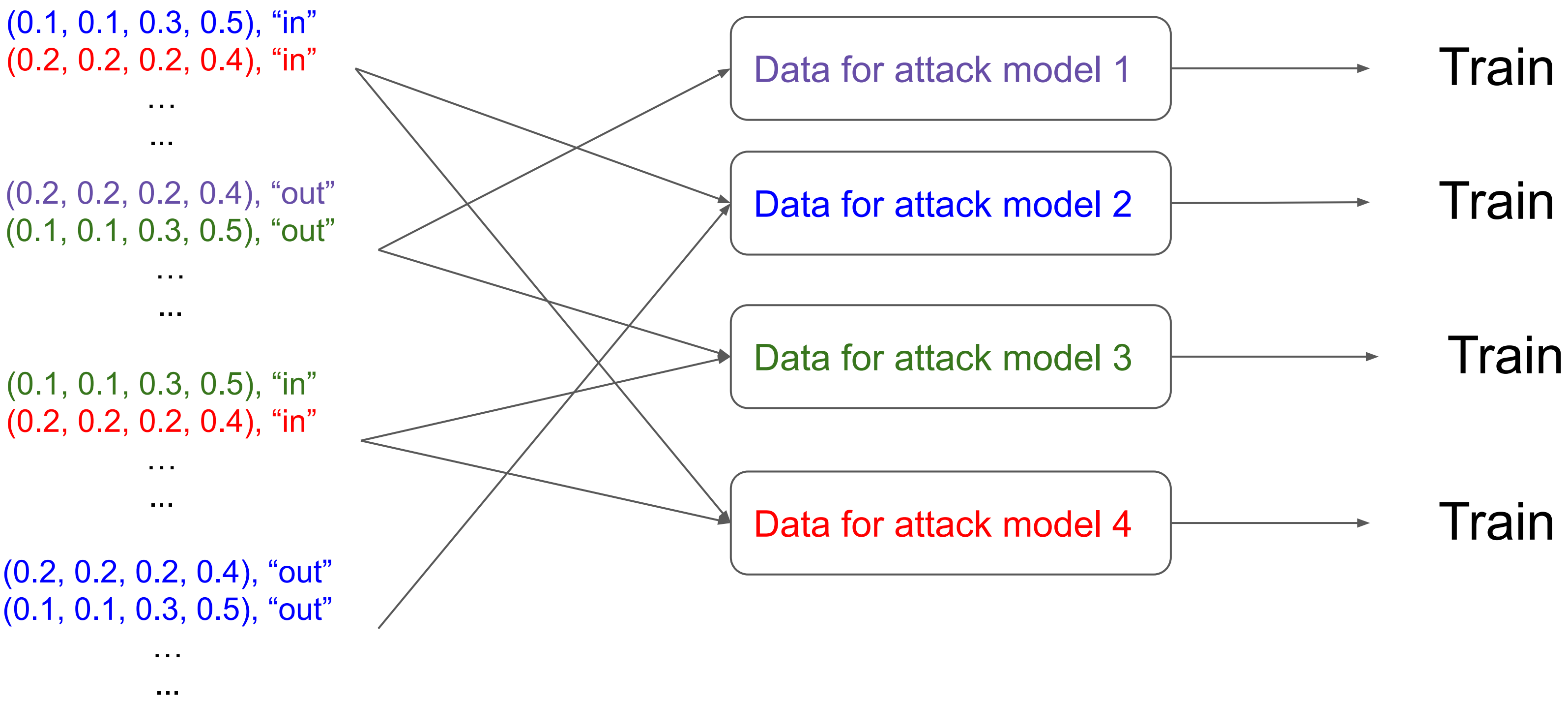
(Image from [1])

Example: 3 Shadow Models, 4 True Classes



Example

- Given that we know the true class of each record, we split the data into 4 classes and train 4 attack models.
  - Class1, class 2, class 3, class 4



Membership Inference Attack: Results

- The more model is dependent to data, the more it leaks information
  - Effect of over-fitting (huge difference between train and test accuracy)

Dataset	*Training Accuracy*	*Testing Accuracy*	*Attack Precision*
Adult	0.848	0.842	0.503
MNIST	0.984	0.928	0.517
Location	1.000	0.673	0.678
Purchase (2)	0.999	0.984	0.505
Purchase (10)	0.999	0.866	0.550
Purchase (20)	1.000	0.781	0.590
Purchase (50)	1.000	0.693	0.860
Purchase (100)	0.999	0.659	0.935
TX hospital stays	0.668	0.517	0.657

(Table from [1])

Contents of Presentation

- Introduction to Machine Learning
- Privacy disclosing attacks
- Membership Inference
- **⇒ Counter-measures**
- The Netflix prize
- Responsibility
- Conclusion

Preventing *Membership Inference Attacks*

Model choice

Not all models are created equally…

Some models might be more prone to leak information than others. For example, some decision trees are deterministic whereas Bayesian-based models are stochastic and include uncertainty, thus making it harder to infer information about the model. [2]

Regularisation

Two birds, one stone…

Regularisation methods prevent the model from overfitting and memorizing specific features from the training data. This leads to a model with greater generalization capabilities while protecting the privacy of the individuals.

Popular regularisation methods:

- Adding noise to the training set [1, 2]
- Dropout for neural networks [3]
- Model stacking (Ensemble learning)[3]
- Weight Normalisation [4], etc

Restrict model output

Ignorance is bliss…

Limit the amount of information provided to the end-user to just the bare minimum. The less information the attackers can collect from the model, the more difficult it is for them to gather sensitive information. [1, 2]

Towards **Privacy-Preserving Machine Learning**

Privacy-aware Data Preprocessing [5]

- Data Anonymization
- Dimensionality Reduction

ML on Encrypted Data [6]

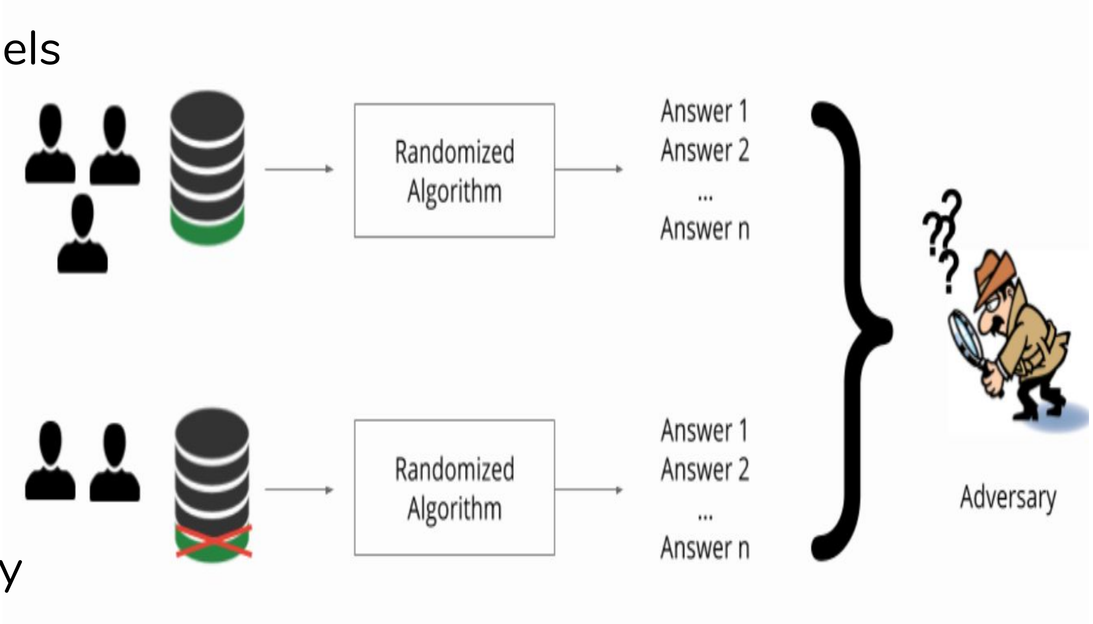
- Training with Homomorphic Encryption

Differential Privacy [7]

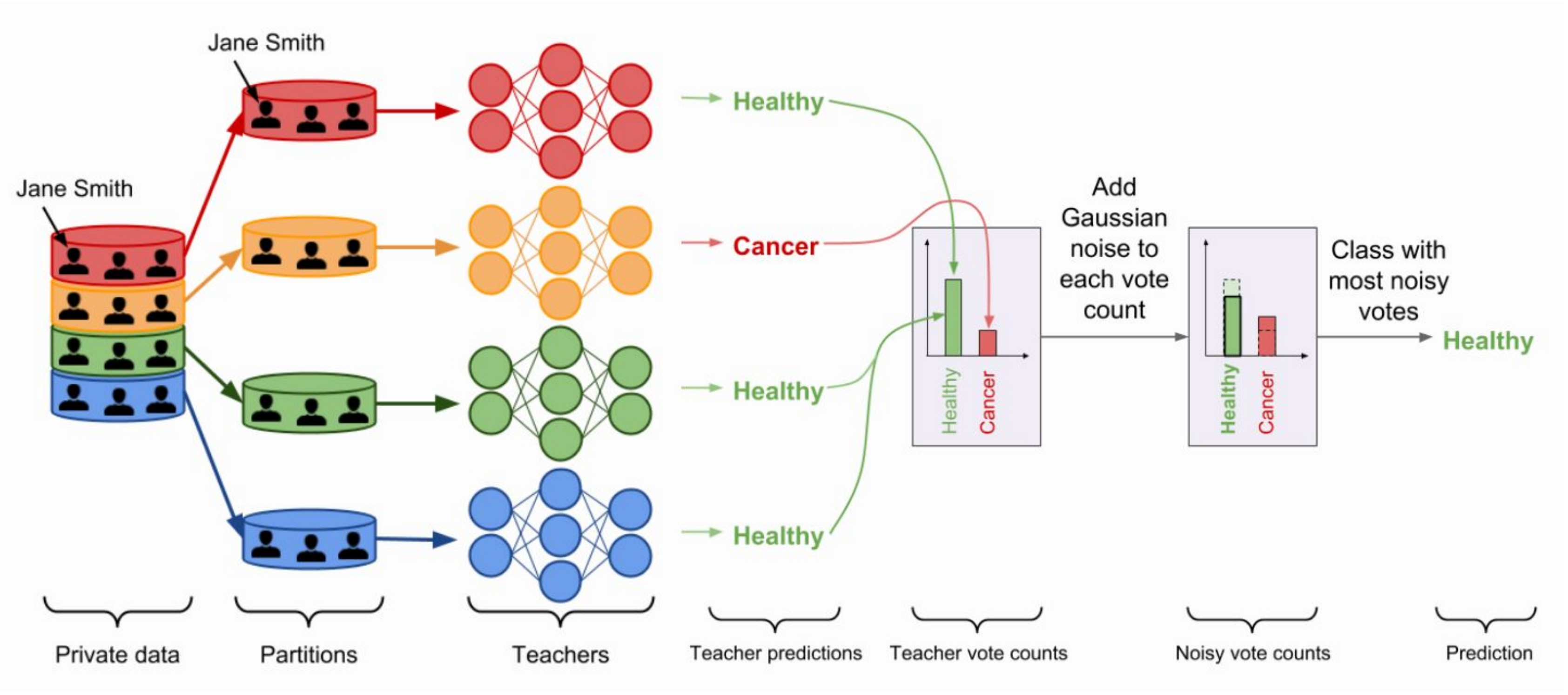
- PATE framework

Privacy Aggregation of Teacher Ensembles (PATE)

- Framework on top of ML models
- Model-independent
- Guarantees differential privacy

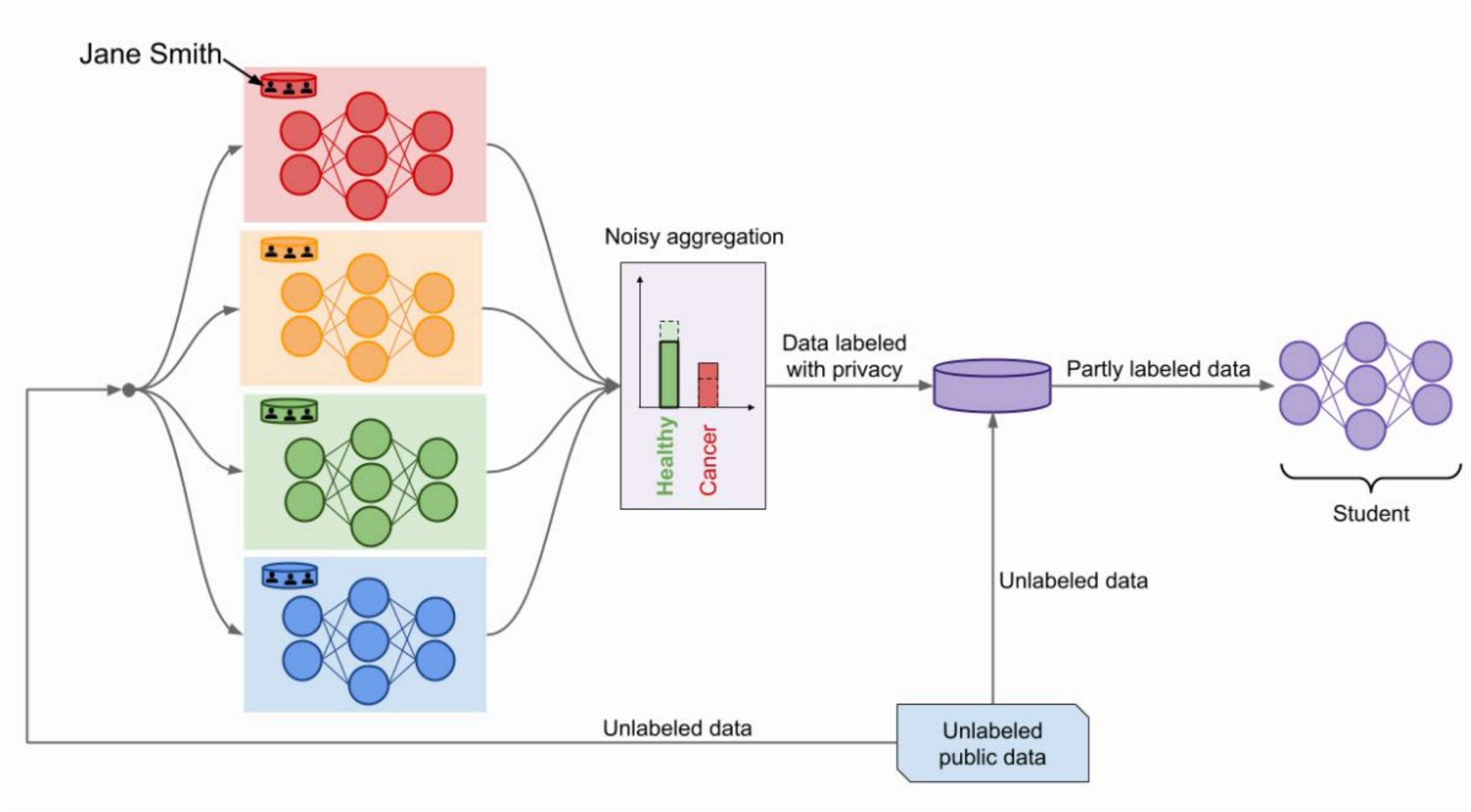


Privacy Aggregation of Teacher Ensembles (PATE)



1 2 3 4 5

Privacy Aggregation of Teacher Ensembles (PATE)



Contents of Presentation

- Introduction to Machine Learning
- Privacy disclosing attacks
- Membership Inference
- Counter-measures
- **⇒The Netflix prize**
- Responsibility
- Conclusion

The Netflix Prize - Real world example

- Open competition (2006-2009)
- Goal is to create an algorithm which predicts user ratings
- **Anonymized** data set was provided
  - userID, movieID, date of rating, rating score
  - 100M ratings, 500k users

The Netflix Prize - What went wrong?

- 2008: de-anonymization via IMDb reviews [8]
  - 99% accuracy
  - 8 movie ratings needed (only 6 had to be correct)
- 2018: de-anonymization via Amazon Reviews [9]

Contents of Presentation

- Introduction to Machine Learning
- Privacy disclosing attacks
- Membership Inference
- Counter-measures
- The Netflix prize
- **⇒Responsibility**
- Conclusion

Responsibility and prevention

- Who is responsible?
  - Netflix's engineers thought the data was anonymized
  - Lack of perturbation (noise)
  - Data was not uniformly sampled
  - The data was not k-Anonymized
- How do we prevent this
  - Possible to prevent correlation of anonymized datasets?
  - Government legislation
  - Trade off between profit and performance vs anonymity
  - Proactive prevention with explicit consent?

Summary

- Introduction to Machine Learning
- Privacy disclosing attacks
- Membership Inference
- Counter-measures
- The Netflix prize
- Responsibility
- **⇒Conclusion**

Conclusions - What to have in mind

- Performance vs privacy trade-off
- Know the attacks, and your counter-measures
- Consider correlation between independent datasets
- Don’t needlessly expose more than you have to

References:

Membership Inference

1. *Shokri, Reza, et al. "Membership inference attacks against machine learning models." 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017.*
2. *Truex, Stacey, et al. "Demystifying Membership Inference Attacks in Machine Learning as a Service." IEEE Transactions on Services Computing (2019).*
3. *Salem, Ahmed, et al. "ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models." arXiv preprint arXiv:1806.01246 (2018).*
4. *Hayes, Jamie, et al. "LOGAN: Membership inference attacks against generative models." Proceedings on Privacy Enhancing Technologies 2019.1 (2019): 133-152.*

Privacy-Preserving Machine Learning

5. *Al-Rubaie, Mohammad, and J. Morris Chang. "Privacy-Preserving Machine Learning: Threats and Solutions." IEEE Security & Privacy 17.2 (2019): 49-58.*

Training on Encrypted Data

6. *Bost, Raphael, et al. "Machine learning classification over encrypted data." NDSS. Vol. 4324. 2015.*

Differential Privacy

7. *Abadi, Martin, et al. "Deep learning with differential privacy." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016.*

Netflix Prize de-Anonymization

8. *Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset)." University of Texas at Austin (2008).*
9. *Archie, Maryam, et al. "Who's Watching? - De-anonymization of Netflix Reviews using Amazon Reviews", MIT. 2018.*

Figures on PATE are taken from the blog post "Privacy and machine learning: two unexpected allies?" by N.Papernot and I. Goodfellow, Apr 29, 2018

Questions?

Who is responsible for the privacy leak in the case of the Netflix competition?

In the case of The Netflix Prize, the dataset provided by Netflix was anonymized, until the data was correlated with other public data sets (IMDb and Amazon Reviews), which allowed an adversary to de-anonymize users in the Netflix dataset. This in turn led to a privacy leak where you could infer things such as political opinions of users, based on their ratings. So, how far should a company go to make sure the data is safe to publish? How feasible would it have been for Netflix to check for unwanted correlations before publishing the data set? Where lies the responsibility?

Should GDPR's "the right to be forgotten" apply to machine learning models which have been trained on your data? How could that be enforced (or can it even be enforced)?

As an example, Google uses our photos uploaded via the Google Photos Service to train its machine learning models for face recognition. This is OK as long as we have accepted the User Agreement. However, once we decide not to use the service anymore, how can we make sure that our data is fully forgotten by Google? Even if the photos themselves are fully deleted, they have been used in training Google’s face recognition models anyway. Can we make sure the models will forget out data? Can we file a lawsuit against Google?

Should ML models which can medically diagnose you easily (e.g through a smartphone) be released publicly?

On one hand they could drastically help prevent illnesses, but on the other hand they could leak very sensitive information or promote sensitive information about the population. This question is of paramount importance especially because medical data are very rare and typically contain sensitive information.