



# Making Machine Learning Forget

Saurabh Shintre, **Kevin Roundy**, Jasjeet Dhaliwal

Technical Director  
Symantec Research Labs

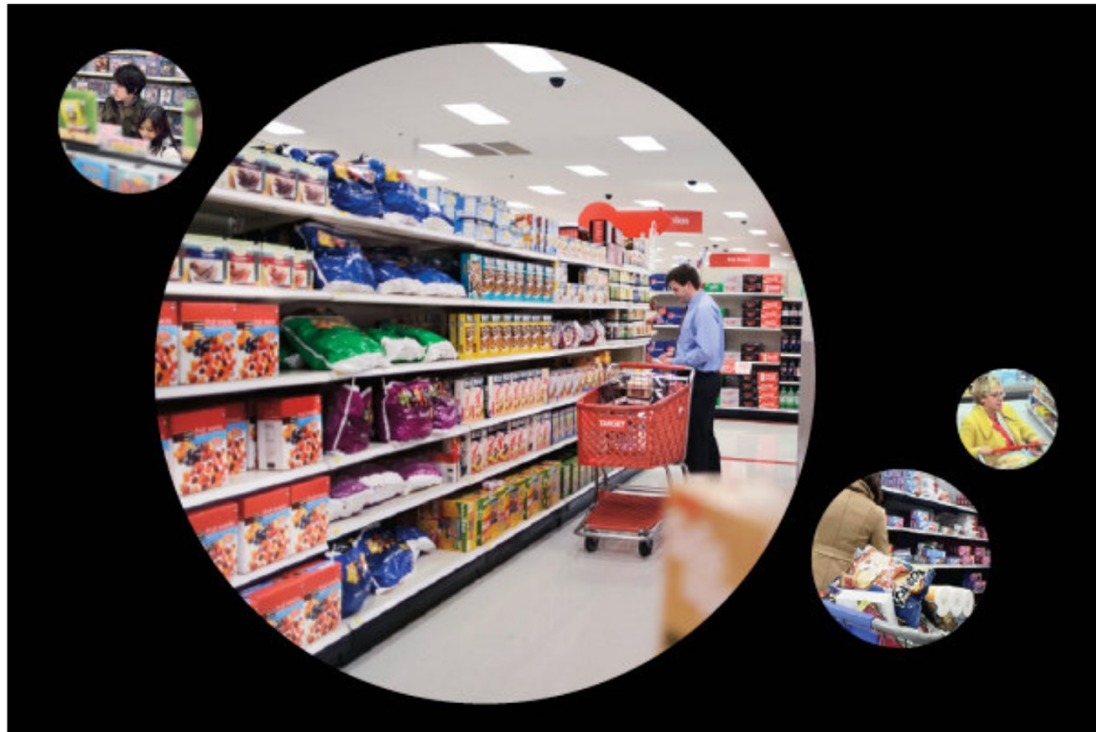


# Simple Data, Big Deductions



## How Companies Learn Your Secrets

By CHARLES DUHIGG FEB. 16, 2012



Antonio Bolfo/Reportage for The New York Times

Whenever possible, Target assigns each shopper a unique code... that keeps tabs on everything they buy. “If you use a credit card or a coupon, or fill out a survey, or mail in a refund, or call the customer help line, or open an e-mail we’ve sent you or visit our Web site, we’ll record it and link it to your Guest ID,” Pole said. “We want to know everything we can.”

A man walked into a Target outside Minneapolis and demanded to see the manager. He was clutching coupons that had been sent to his daughter, and he was angry... “My daughter got this in the mail!” he said. “She’s still in high school, and you’re sending her coupons for baby clothes and cribs?” ... The manager apologized and then called a few days later to apologize again. On the phone, though, the father was somewhat abashed. “I had a talk with my daughter,” he said. “It turns out there’s been some activities in my house I haven’t been completely aware of. She’s due in August. I owe you an apology.”

“We are very conservative about compliance with all privacy laws...”

What else can be inferred from market basket data?  
or web browsing data?  
or location data?  
or fitness tracker data?  
or cellphone data?  
or credit card data?

# What inferences are being made from such data?



## AI is convicting criminals and determining jail time, but is it fair? Biased data feeds biased algorithms

19 Nov 2018

Vyacheslav Polonski  
UX Researcher, Google

## Life Insurers Can Use Social Media Posts To Determine Premiums, As Long As They Don't Discriminate **Forbes**

**MOTHERBOARD** | By Samantha Cole | May 29 2019, 7:11am  
TECH BY VICE

## DIY Facial Recognition for Porn Is a Dystopian Disaster

Someone is making dubious claims to have built a program for detecting faces in porn and cross-referencing against social media, with 100,000 identified so far.

WIRED

## WHAT IS CAMBRIDGE ANALYTICA, ANYWAY?

It's a political data-analysis firm that worked on the 2016 Trump campaign. CA's professed advantage is having enough data points on every American to build extensive personality profiles, which its clients can leverage for "psychographic targeting" of ads.

# GDPR “Right-to-Be Forgotten”

- Allows a user to ask for **deletion** of his/her data
- Data controller has to delete data from its own storage but also from storage of all the processors which whom it has shared this data
- Data can be deleted after it is no longer required from processing
- Data controller must taken responsible steps taking into account **cost** and **available technology**
- But what counts as “storage” of data?

## Art. 17 GDPR

# Right to erasure (‘right to be forgotten’)

1. The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay where one of the following grounds applies:
  - (a) the personal data are no longer necessary in relation to the purposes for which they were collected or otherwise processed;
  - (b) the data subject withdraws consent on which the processing is based according to point (a) of [Article 6\(1\)](#), or point (a) of [Article 9\(2\)](#), and where there is no other legal ground for the processing;
  - (c) the data subject objects to the processing pursuant to [Article 21\(1\)](#) and there are no overriding legitimate grounds for the processing, or the data subject objects to the processing pursuant to [Article 21\(2\)](#);
  - (d) the personal data have been unlawfully processed;
  - (e) the personal data have to be erased for compliance with a legal obligation in Union or Member State law to which the controller is subject;
  - (f) the personal data have been collected in relation to the offer of information society services referred to in [Article 8\(1\)](#).

Please erase my data...



# Data hiding in plain sight?



- Machine learning models act as **hidden stores** of training data
- Possible to **re-create data** from the model (without even having the model)
- These attacks are known as **model inversion/inference/stealing** attacks



**Figure 1:** An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

## Stealing Machine Learning Models via Prediction APIs

### Authors:

Florian Tramèr, *École Polytechnique Fédérale de Lausanne (EPFL)*; Fan Zhang, *Cornell University*; Ari Juels, *Cornell Tech*; Michael K. Reiter, *The University of North Carolina at Chapel Hill*; Thomas Ristenpart, *Cornell Tech*

### Abstract:

Machine learning (ML) models may be deemed confidential due to their sensitive training data, commercial value, or use in security applications. Increasingly often, confidential ML models are being deployed with publicly accessible query interfaces. ML-as-a-service (“predictive analytics”) systems are an example: Some allow users to train models on potentially sensitive data and charge others for access on a pay-per-query basis.

The tension between model confidentiality and public access motivates our investigation of *model extraction attacks*. In such attacks, an adversary with black-box access, but no prior knowledge of an ML model's parameters or training data, aims to duplicate the functionality of (i.e., “steal”) the model. Unlike in classical learning theory settings, ML-as-a-service offerings may accept partial feature vectors as inputs and include confidence values with predictions. Given these practices, we show simple, efficient attacks that extract target ML models with near-perfect fidelity for popular model classes including logistic regression, neural networks, and decision trees. We demonstrate these attacks against the online services of BigML and Amazon Machine Learning. We further show that the natural countermeasure of omitting confidence values from model outputs still admits potentially harmful model extraction attacks. Our results highlight the need for careful ML model deployment and new model extraction countermeasures.

# Right-to-Forget Must Include ML models

## ML Models should:

- Never remember specific training points in the first place
- Not require data to be retained for it to be forgotten
- Forget auditably

## Privacy / Forgetting mechanisms should:

- be efficient (not require re-training from scratch)
- be provably forgetful
- not destroy the model's efficacy
- be general to many machine learning methods

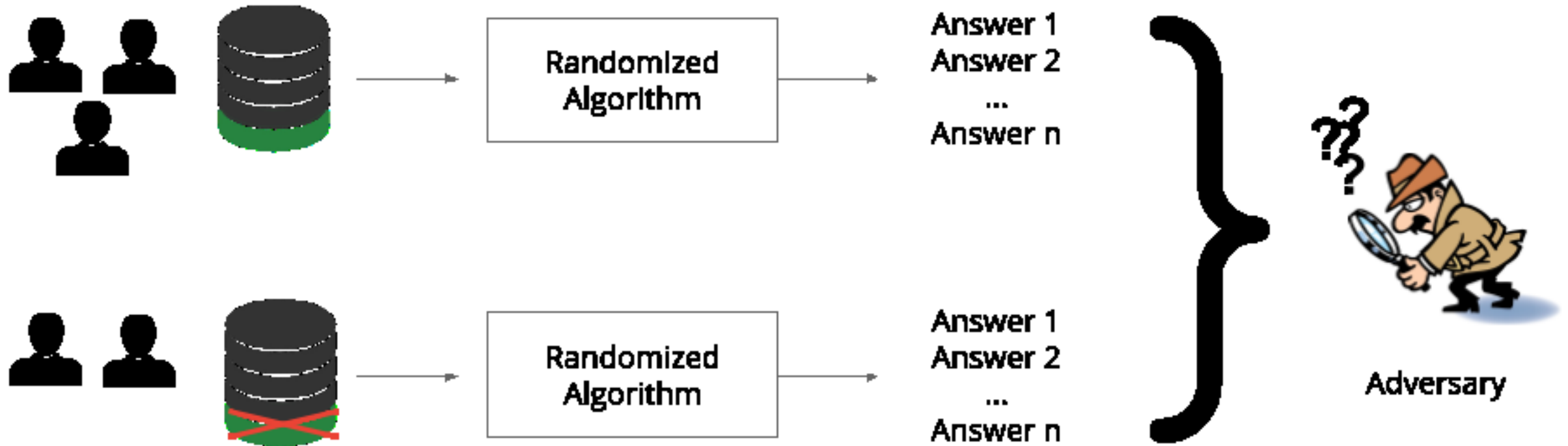
## Forgettability is important for many reasons:

- privacy and forgiveness
- recovery from training data poisoning
- separation of shared accounts



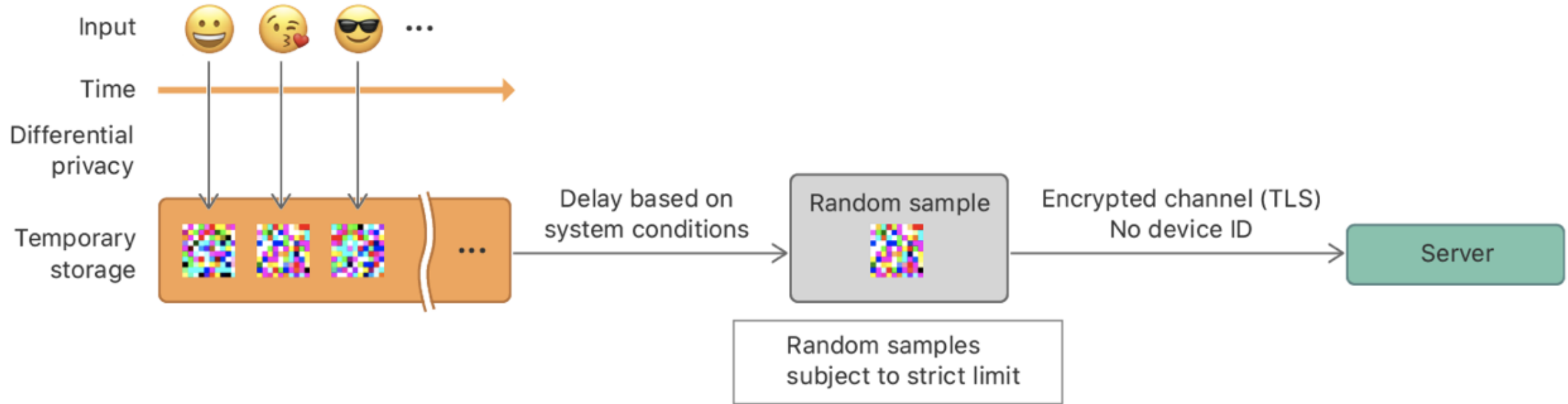
# Differentially Private Model Training

Measures degree of privacy offered by the system



<http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.html>

# Differentially Private Model Training



## Points in Favor

- + Provides provable guarantees
- + Never remembers
- + Never needs to forget

## Points Against

- Adding noise for privacy reduces model effectiveness, sometimes dramatically
- Buggy implementations may result in lower than theoretical privacy protections

## *What are Influence Functions?*

- Tools from robust statistics that measure the influence of one training data point on the overall model's behavior
- They measure how the model's parameters will change if a **given training data point is removed**
- Allow us to measure the change in **model's prediction on test data** due to the removal of a given training data

# Influence Functions for Auditable Forgetting



- The trusted auditor stores the influence function of the model and a standard (but secret) set of test cases
- It evaluates the model on the test set and stores the model's predictions
- When the request comes, the model creator can compute the influence of the user's data on the model and identify changes required in the model to "remove" that data
- The new model parameters essentially create a model which doesn't have the user's data
- The auditor can now evaluate the test set on the new model and get new model predictions
- If the difference in prediction of the new and old model is what is predicted by influence functions, the auditor can confirm that the request was properly met
- This method does not require access to the raw model parameters but only to a prediction interface

# Influence Functions for Auditable Forgetting



## Points in Favor

- + Auditable forgetting
- + Efficient forgetting
- + Broadly applicable to many ML models, including neural networks
- + No changes to model training, so no reduction in model effectiveness

## Points Against

- Data must be retained or re-submitted so that it can be forgotten
- Measuring against a standard set of test points provides strong confidence of forgetfulness, but does not prove forgetfulness

# Potential Solution 3: Machine Unlearning

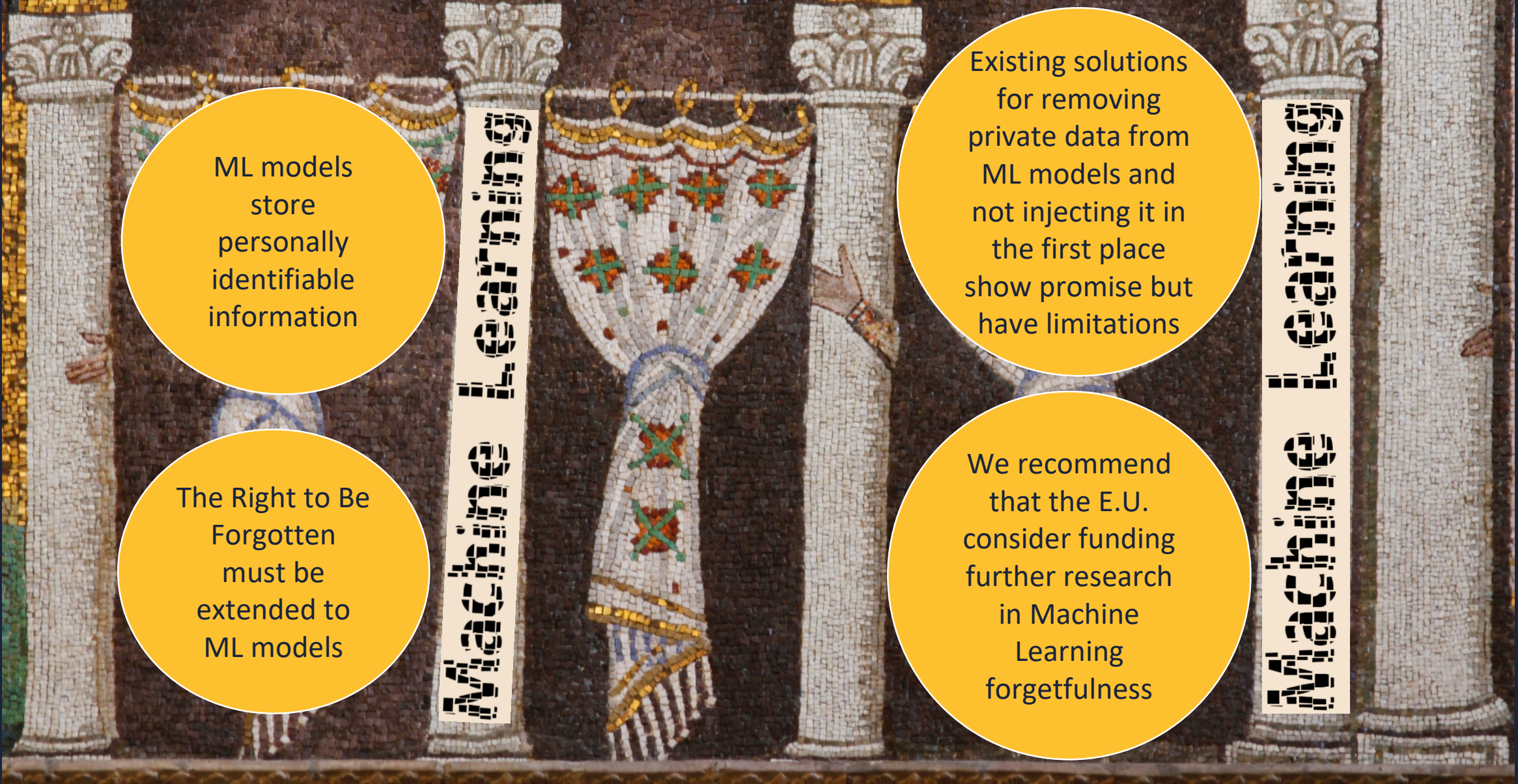


- Train the model not directly on the training data but on **aggregated training data**
- Model only stores these aggregates and **original data is not stored persistently**
- Removal from the model requires data to be re-submitted so that aggregate counts can be **updated the aggregates**

# Comparing ML Forgetfulness Methods

Properties	Differential Privacy	Influence Functions	Machine Unlearning
Never Remember	Yes	No	No
Forget Auditably	Yes – n/a	Yes	Yes
Forget Efficiently	Yes – n/a	Yes	Yes
Provable Guarantees	Yes	No	Yes
Data retention needed?	No	Yes or resubmit	Yes or resubmit
Effective Models	No	Yes	Yes, if applicable
Applicable to most ML	case by case	Yes	Summative models

# In Conclusion...



ML models  
store  
personally  
identifiable  
information

The Right to Be  
Forgotten  
must be  
extended to  
ML models

Existing solutions  
for removing  
private data from  
ML models and  
not injecting it in  
the first place  
show promise but  
have limitations

We recommend  
that the E.U.  
consider funding  
further research  
in Machine  
Learning  
forgetfulness