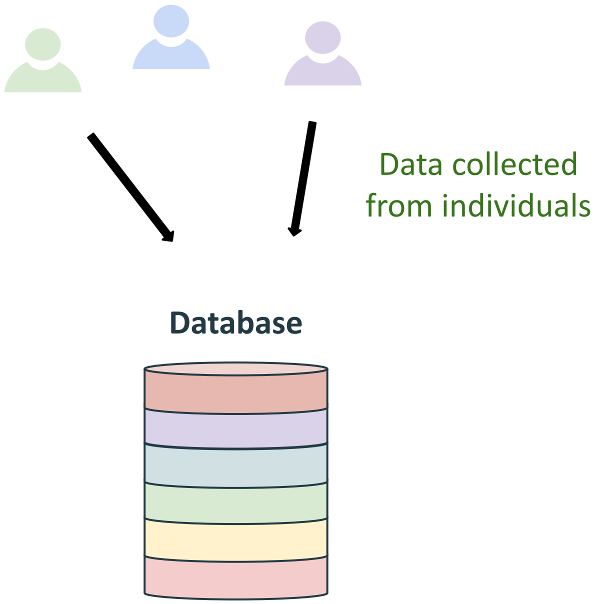
Machine Unlearning

Bourtoule, Lucas, et al., 2021 IEEE Symposium on Security and Privacy

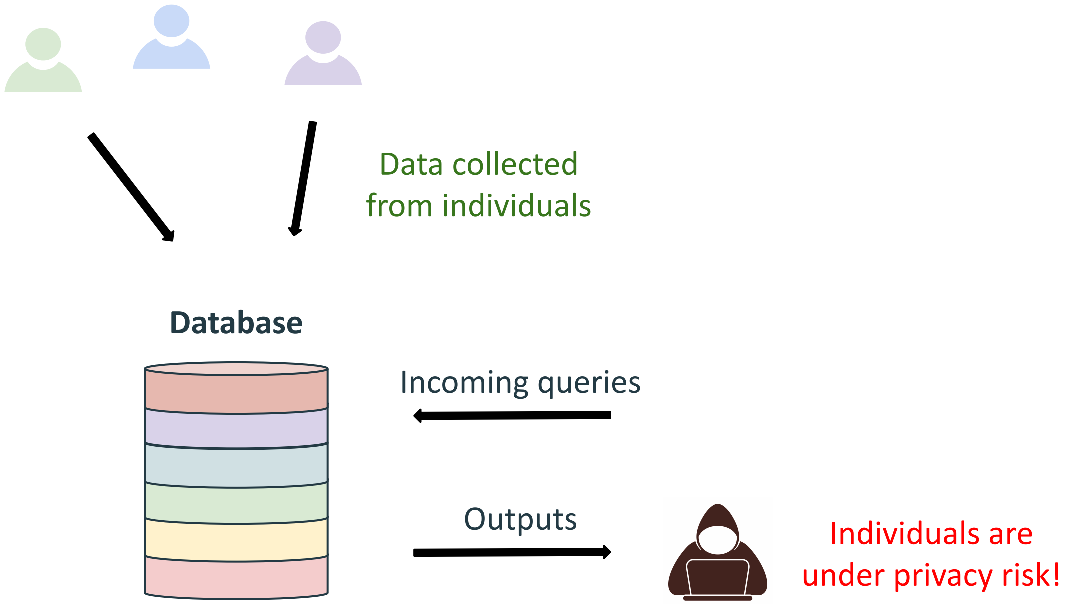
Presented by Prakhar Ganesh

Motivation



Data collected from individuals

Database



Data collected from individuals

Database

Incoming queries

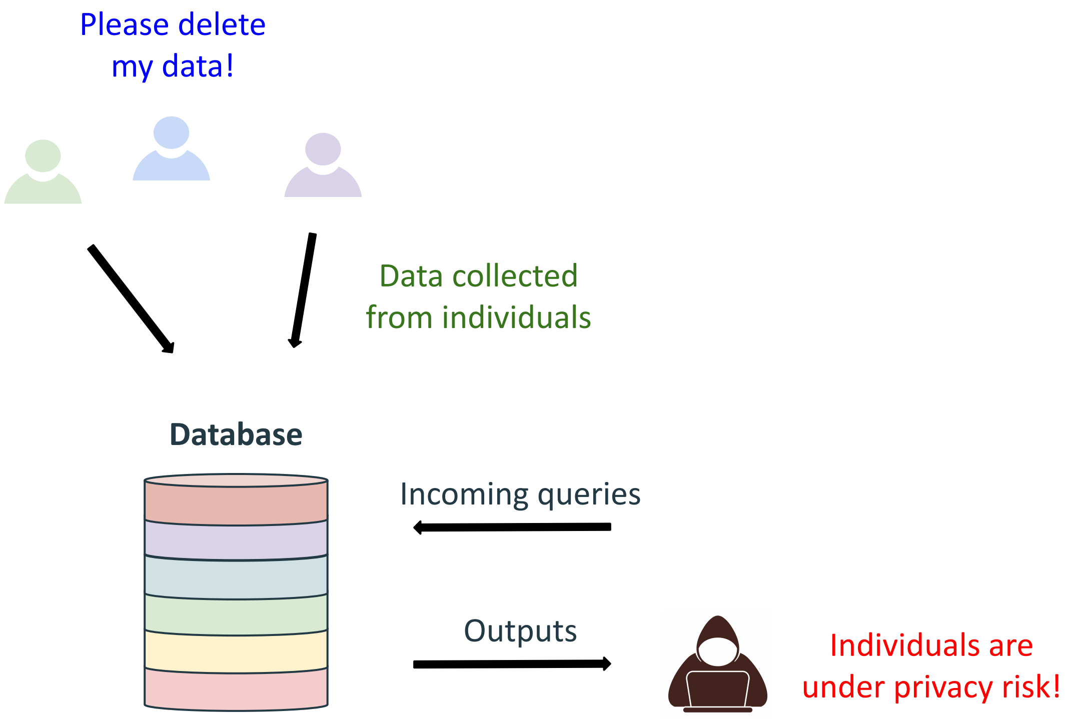
Outputs

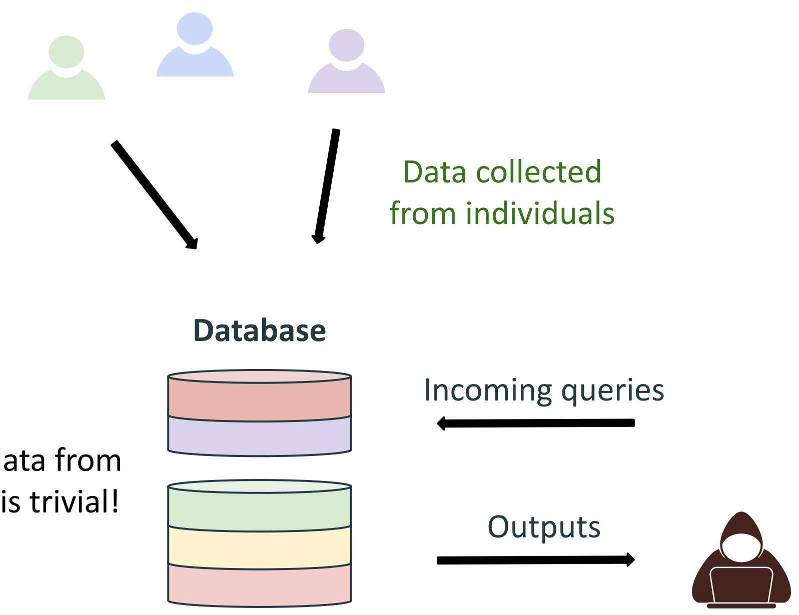
Individuals are under privacy risk!

Right to Erasure (‘Right to be Forgotten’)

Act 17.1 GDPR: ***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 [...]***

Act 17.2 GDPR: ***Where the controller has made the personal data public and is obliged pursuant to paragraph 1 to erase the personal data, the controller, taking account of available technology and the cost of implementation, shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data.***


Please delete
my data!
Data collected
from individuals
Database
Incoming queries
Outputs
Individuals are
under privacy risk!



Please delete my data!

Data collected from individuals

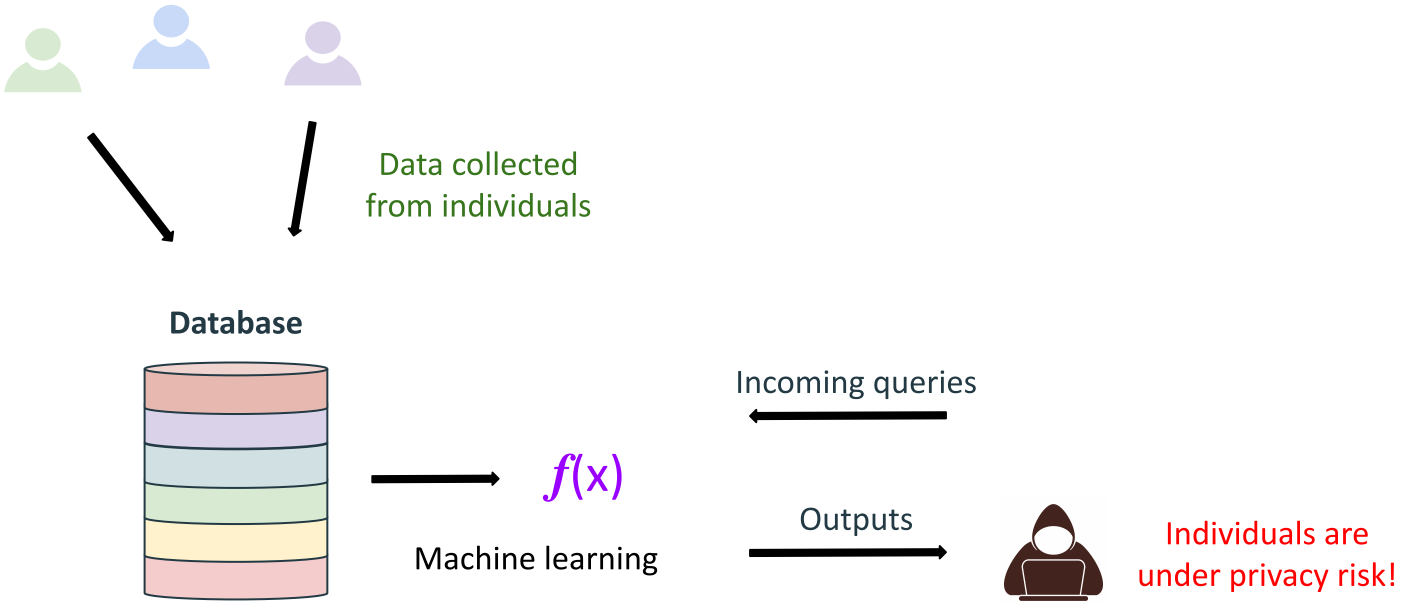
Database

Incoming queries

Outputs

Removing data from a database is trivial!

Individuals are under privacy risk!



Data collected from individuals

Database

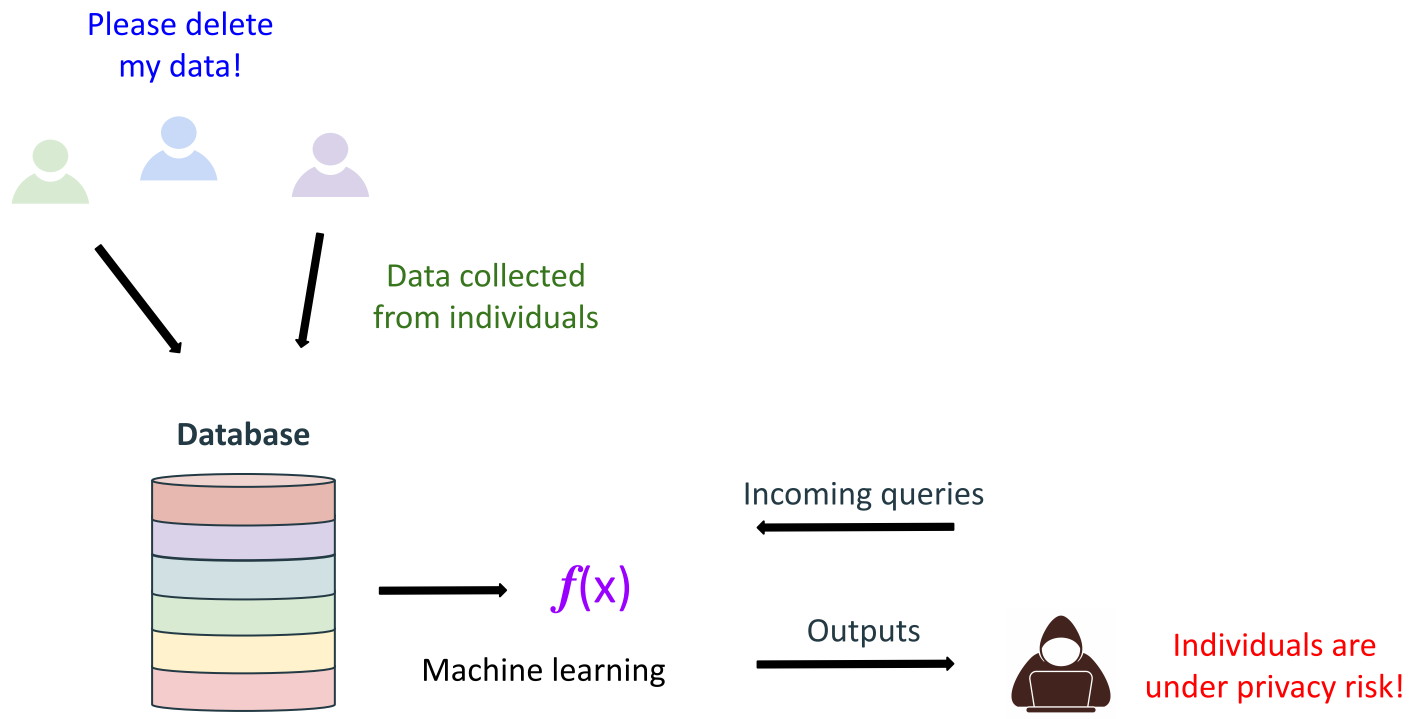
$f(x)$

Machine learning

Incoming queries

Outputs

Individuals are under privacy risk!



Please delete my data!

Data collected from individuals

Database

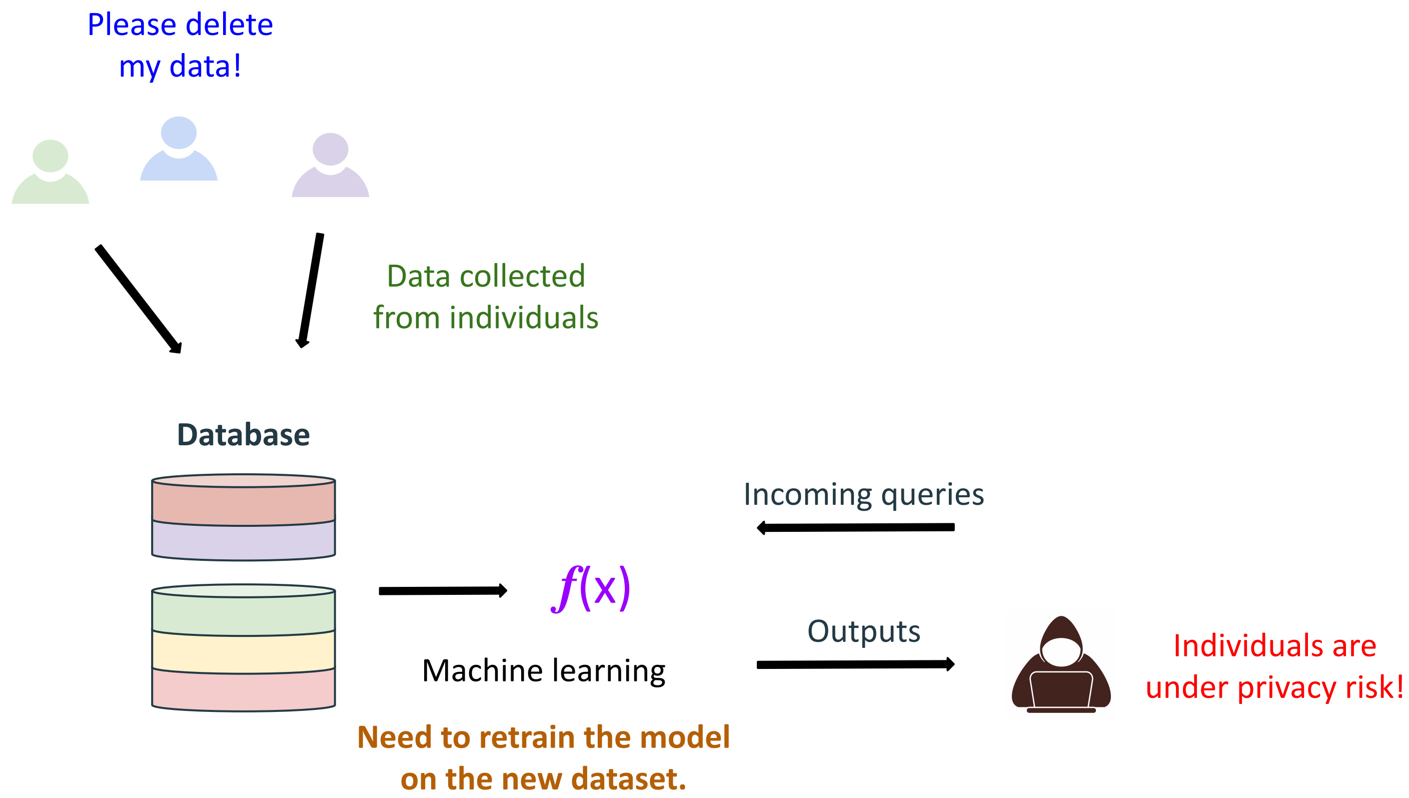
f(x)

Machine learning

Incoming queries

Outputs

Individuals are under privacy risk!



Please delete my data!

Data collected from individuals

Database

f(x)

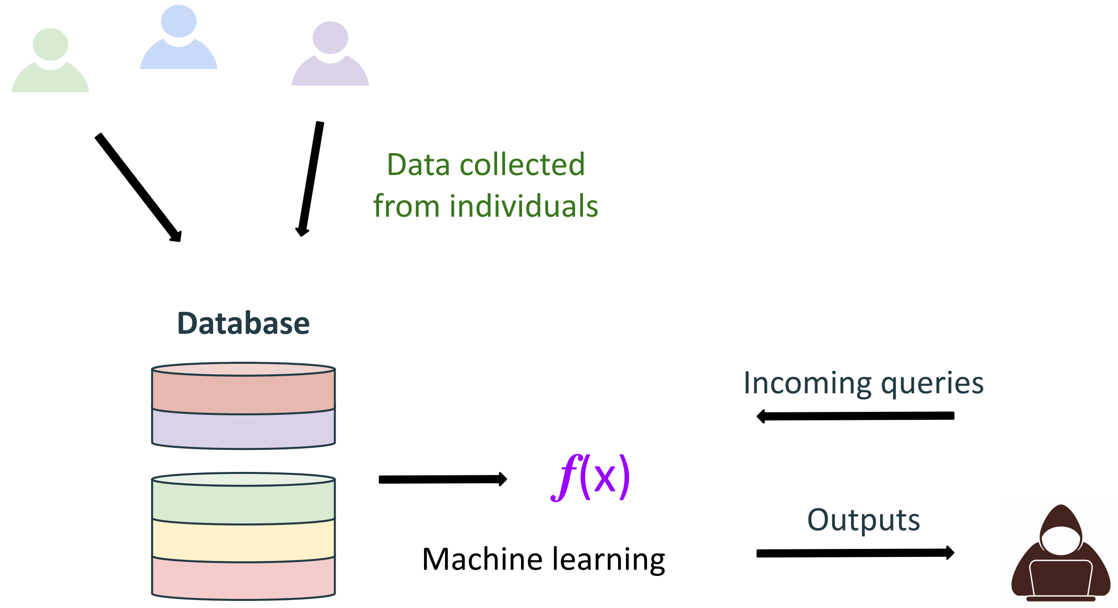
Machine learning

Incoming queries

Outputs

Individuals are under privacy risk!

Need to retrain the model on the new dataset.



Please delete my data!

Data collected from individuals

Database

$f(x)$

Machine learning

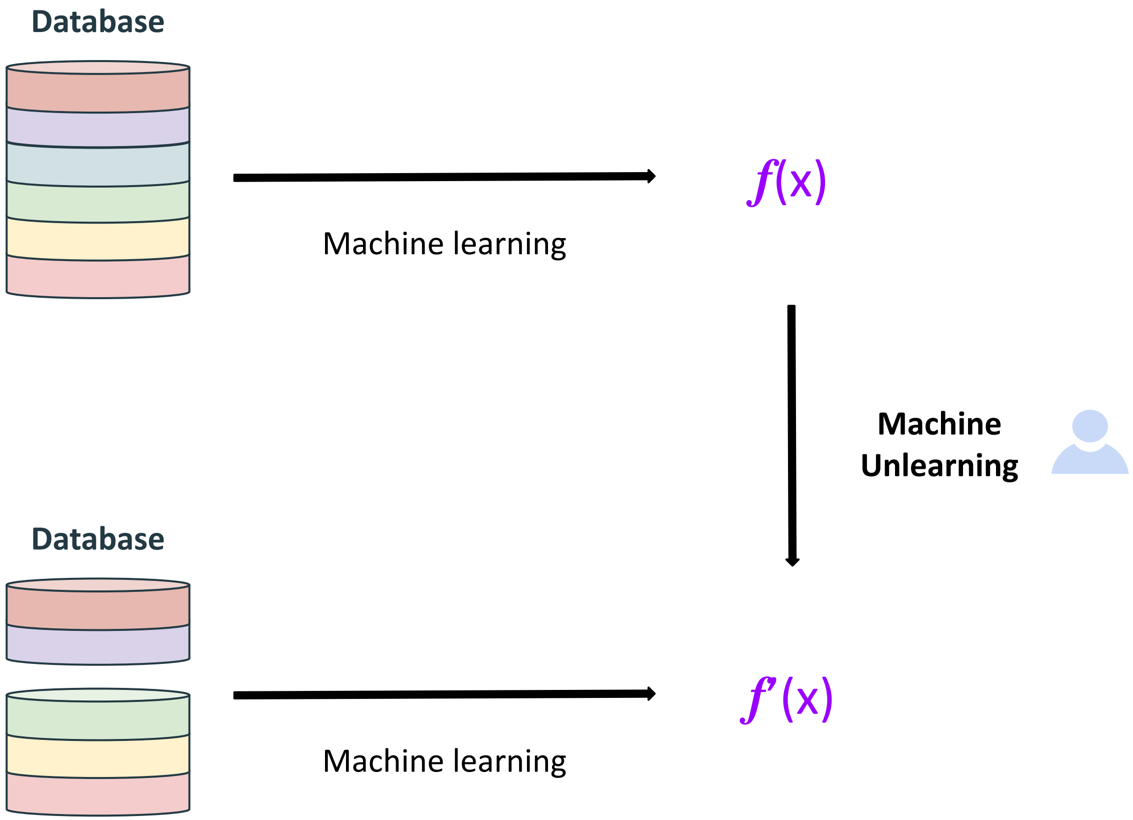
Incoming queries

Outputs

Individuals are under privacy risk!

Need to retrain the model on the new dataset.

Retraining every time is expensive! Can we do better?



Database

Machine learning

$f(x)$

Machine Unlearning

Database

Machine learning

$f'(x)$

Defining Unlearning

Machine Unlearning is not trivial!

Machine Unlearning is not trivial!

- Training is incremental

Update 1

Update 2

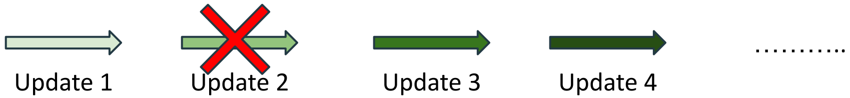
Update 3

Update 4

……….

Machine Unlearning is not trivial!

- Training is incremental



Change in just one update changes
everything that comes after!

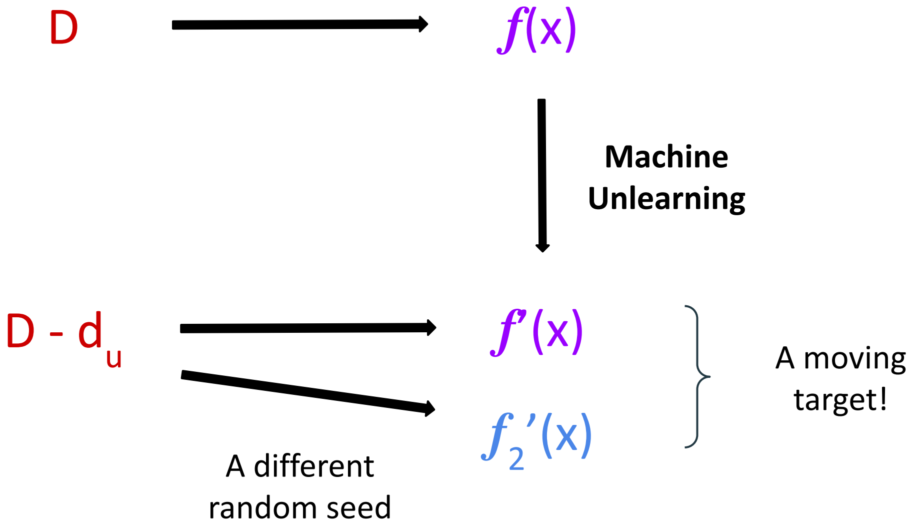
Machine Unlearning is not trivial!

- Training is incremental
- Stochasticity in Training

$$\begin{tikzcd}
D \arrow[r] & f(x) \arrow[d, "\textbf{Machine Unlearning}"] \\
D - d_u \arrow[r] & f'(x)
\end{tikzcd}$$

Machine Unlearning is not trivial!

- Training is incremental
- Stochasticity in Training



D

$f(x)$

Machine Unlearning

$D - d_u$

$f'(x)$

$f_2'(x)$

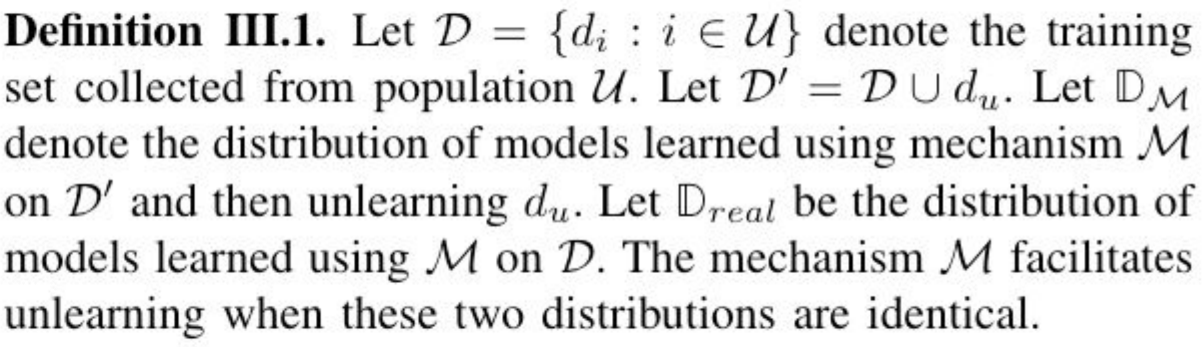
A different random seed

A moving target!

Machine Unlearning is not trivial!

- Training is incremental
- Stochasticity in Training
- We have very little understanding of how each data point impacts the model!

Formalizing Machine Unlearning



Definition III.1. Let $\mathcal{D} = \{d_i : i \in \mathcal{U}\}$ denote the training set collected from population $\mathcal{U}$. Let $\mathcal{D}' = \mathcal{D} \cup d_u$. Let $\mathbb{D}_{\mathcal{M}}$ denote the distribution of models learned using mechanism $\mathcal{M}$ on $\mathcal{D}'$ and then unlearning d_u. Let $\mathbb{D}_{real}$ be the distribution of models learned using $\mathcal{M}$ on $\mathcal{D}$. The mechanism $\mathcal{M}$ facilitates unlearning when these two distributions are identical.

Formalizing Machine Unlearning

$$\begin{tikzcd} D \arrow[r] & f(x) \arrow[d, "\textbf{Machine Unlearning}"] \\ D - d_u \arrow[r] & f'(x) \end{tikzcd}$$

Formalizing Machine Unlearning

$$\begin{tikzcd}
D \arrow[r] \arrow[dr, bend left, "\text{A New Learning Mechanism compatible with Unlearning}"'] & f(x) \arrow[d, "\text{Machine Unlearning}"] \\
D - d_u \arrow[r] & f'(x)
\end{tikzcd}$$

Goals of Unlearning

Goals of Unlearning

- **Intelligibility:** Any unlearning strategy should be intelligible.

Goals of Unlearning

- **Intelligibility:** Any unlearning strategy should be intelligible.
- **Comparable Accuracy:** Any unlearning strategy should strive to introduce a small accuracy gap.

Goals of Unlearning

- **Intelligibility:** Any unlearning strategy should be intelligible.
- **Comparable Accuracy:** Any unlearning strategy should strive to introduce a small accuracy gap.
- **Reduced Unlearning Time:** The strategy should have provably lower time than brute retraining.

Goals of Unlearning

- **Intelligibility:** Any unlearning strategy should be intelligible.
- **Comparable Accuracy:** Any unlearning strategy should strive to introduce a small accuracy gap.
- **Reduced Unlearning Time:** The strategy should have provably lower time than brute retraining.
- **Provable Guarantees:** Provide provable guarantees that any number of points have been unlearned.

Goals of Unlearning

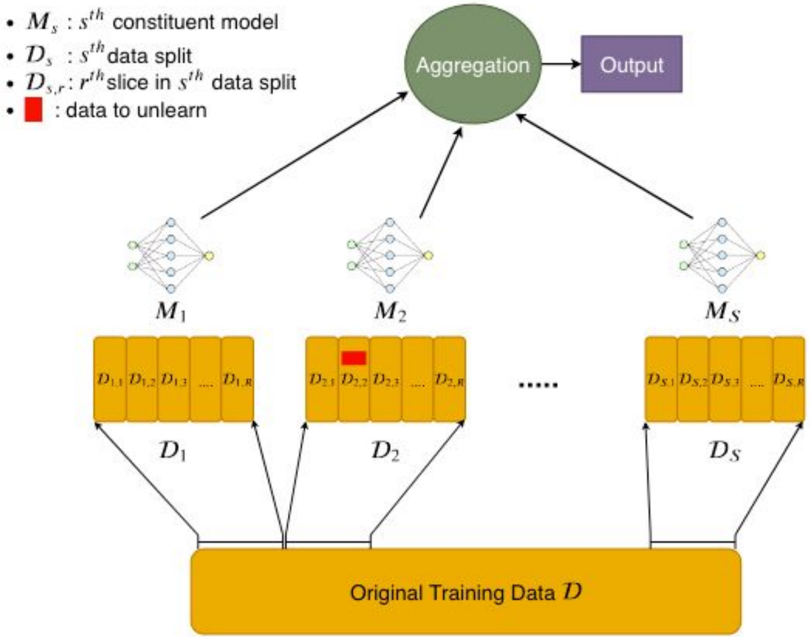
- **Intelligibility:** Any unlearning strategy should be intelligible.
- **Comparable Accuracy:** Any unlearning strategy should strive to introduce a small accuracy gap.
- **Reduced Unlearning Time:** The strategy should have provably lower time than brute retraining.
- **Provable Guarantees:** Provide provable guarantees that any number of points have been unlearned.
- **Model Agnostic:** The new strategy for unlearning should be general.

Goals of Unlearning

- **Intelligibility:** Any unlearning strategy should be intelligible.
- **Comparable Accuracy:** Any unlearning strategy should strive to introduce a small accuracy gap.
- **Reduced Unlearning Time:** The strategy should have provably lower time than brute retraining.
- **Provable Guarantees:** Provide provable guarantees that any number of points have been unlearned.
- **Model Agnostic:** The new strategy for unlearning should be general.
- **Limited Overhead:** Any new unlearning strategy should not introduce additional overhead to training.

SISA (Sharded, Isolated, Sliced, Aggregated)

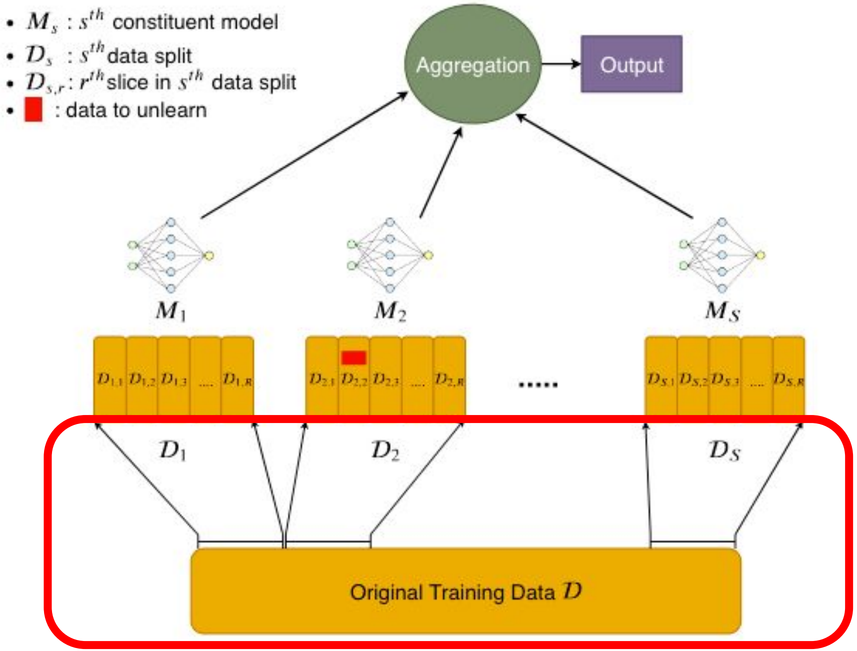
SISA



- M_s : s^{th} constituent model
- D_s : s^{th} data split
- $D_{s,r}$: r^{th} slice in s^{th} data split
- 🟥 : data to unlearn

Aggregation
Output
M_1
M_2
M_S
$D_{1,1}$ $D_{1,2}$ $D_{1,3}$ .... $D_{1,R}$
$D_{2,1}$ $D_{2,2}$ $D_{2,3}$ .... $D_{2,R}$
.....
$D_{S,1}$ $D_{S,2}$ $D_{S,3}$ .... $D_{S,R}$
D_1
D_2
D_S
Original Training Data D

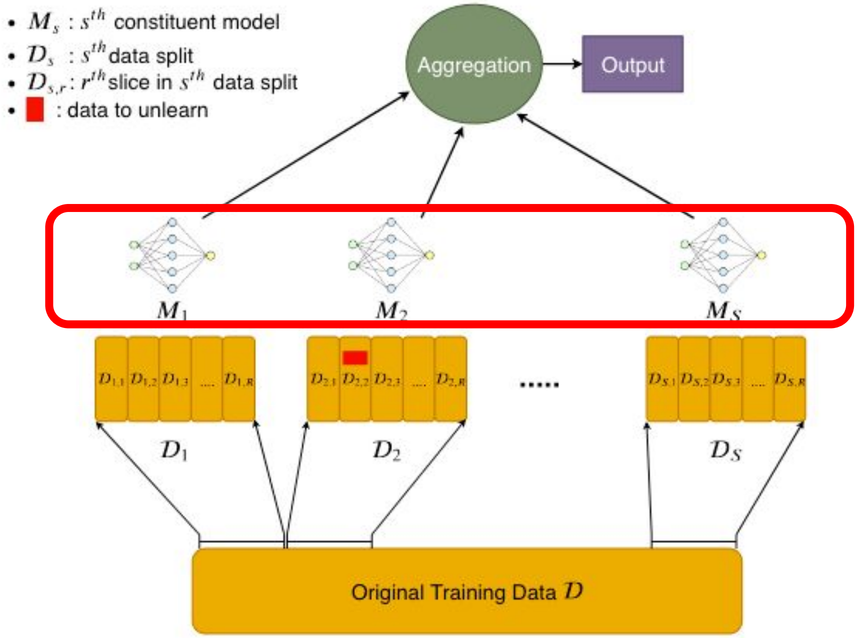
Sharding



- M_s : s^{th} constituent model
- $\mathcal{D}_s$: s^{th} data split
- $\mathcal{D}_{s,r}$: r^{th} slice in s^{th} data split
- 🟥 : data to unlearn

Aggregation
Output
M_1
M_2
M_S
$\mathcal{D}_{1,1}$ $\mathcal{D}_{1,2}$ $\mathcal{D}_{1,3}$ .... $\mathcal{D}_{1,R}$
$\mathcal{D}_{2,1}$ $\mathcal{D}_{2,2}$ $\mathcal{D}_{2,3}$ .... $\mathcal{D}_{2,R}$
.....
$\mathcal{D}_{S,1}$ $\mathcal{D}_{S,2}$ $\mathcal{D}_{S,3}$ .... $\mathcal{D}_{S,R}$
$\mathcal{D}_1$
$\mathcal{D}_2$
$\mathcal{D}_S$
Original Training Data $\mathcal{D}$

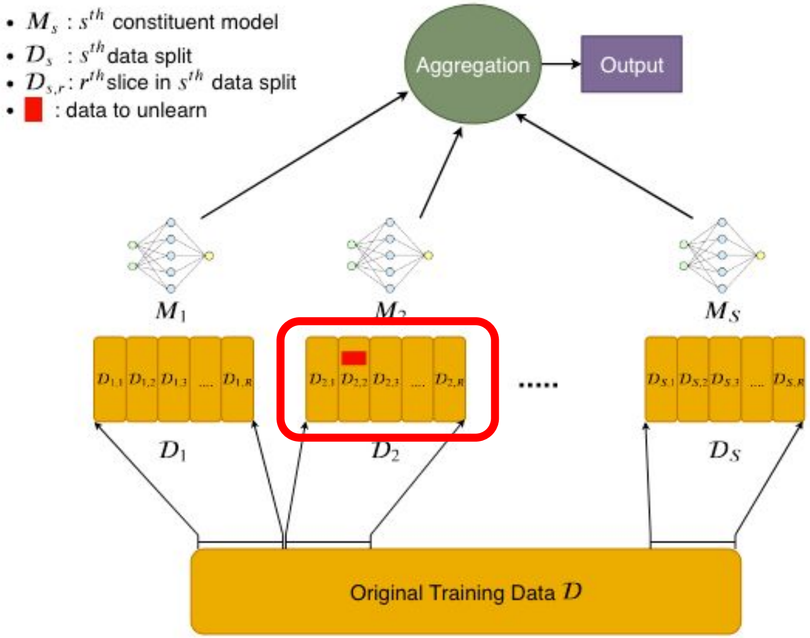
Isolation



- M_s : s^{th} constituent model
- $\mathcal{D}_s$: s^{th}data split
- $\mathcal{D}_{s,r}$: r^{th}slice in s^{th} data split
- 🟥 : data to unlearn

Aggregation
Output
M_1
M_2
M_S
$\mathcal{D}_{1,1}$ $\mathcal{D}_{1,2}$ $\mathcal{D}_{1,3}$ .... $\mathcal{D}_{1,R}$
$\mathcal{D}_{2,1}$ $\mathcal{D}_{2,2}$ $\mathcal{D}_{2,3}$ .... $\mathcal{D}_{2,R}$
.....
$\mathcal{D}_{S,1}$ $\mathcal{D}_{S,2}$ $\mathcal{D}_{S,3}$ .... $\mathcal{D}_{S,R}$
$\mathcal{D}_1$
$\mathcal{D}_2$
$\mathcal{D}_S$
Original Training Data $\mathcal{D}$

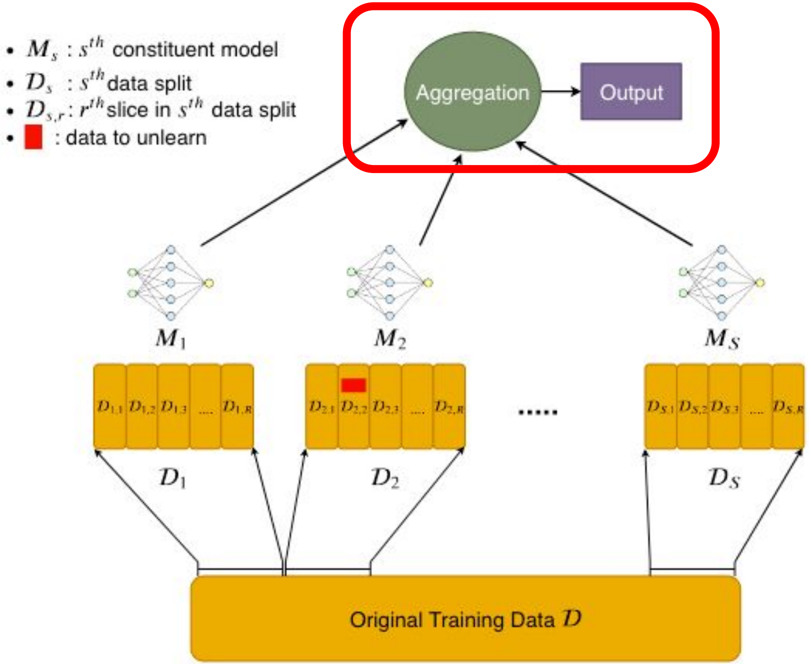
Slicing



- M_s : s^{th} constituent model
- $\mathcal{D}_s$: s^{th} data split
- $\mathcal{D}_{s,r}$: r^{th} slice in s^{th} data split
- 🟥 : data to unlearn

Aggregation
Output
M_1
M_2
M_S
$\mathcal{D}_{1,1}$ $\mathcal{D}_{1,2}$ $\mathcal{D}_{1,3}$ .... $\mathcal{D}_{1,R}$
$\mathcal{D}_{2,1}$ $\mathcal{D}_{2,2}$ $\mathcal{D}_{2,3}$ .... $\mathcal{D}_{2,R}$
.....
$\mathcal{D}_{S,1}$ $\mathcal{D}_{S,2}$ $\mathcal{D}_{S,3}$ .... $\mathcal{D}_{S,R}$
$\mathcal{D}_1$
$\mathcal{D}_2$
$\mathcal{D}_S$
Original Training Data $\mathcal{D}$

Aggregation



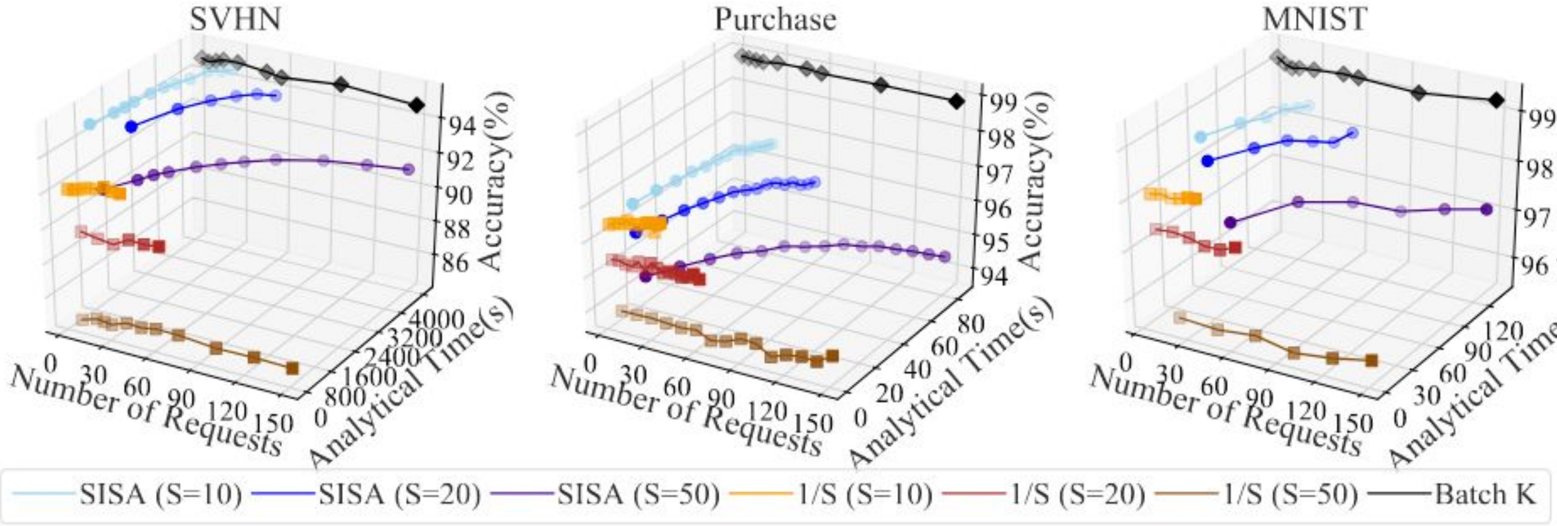
- M_s : s^{th} constituent model
- D_s : s^{th} data split
- $D_{s,r}$: r^{th} slice in s^{th} data split
- 🟥 : data to unlearn

Aggregation
Output
M_1
M_2
M_S
$D_{1,1}$ $D_{1,2}$ $D_{1,3}$ .... $D_{1,R}$
$D_{2,1}$ $D_{2,2}$ $D_{2,3}$ .... $D_{2,R}$
.....
$D_{S,1}$ $D_{S,2}$ $D_{S,3}$ .... $D_{S,R}$
D_1
D_2
D_S
Original Training Data D

Results

Results

Results


SVHN
Purchase
MNIST
Accuracy(%)
94
92
90
88
86
99
98
97
96
95
94
99
98
97
96
Analytical Time(s)
4000
3200
2400
1600
800
0
80
60
40
20
0
120
90
60
30
0
Number of Requests
0
30
60
90
120
150
SISA (S=10)
SISA (S=20)
SISA (S=50)
1/S (S=10)
1/S (S=20)
1/S (S=50)
Batch K

To Sum Up…

To Sum Up...

- The paper introduces a mathematical framework to formalize the ‘right to be forgotten’ in the context of machine learning.
- The paper discusses various challenges with the problem of unlearning, and provides a list of requirements that would make an unlearning algorithm actually useful.
- The paper introduces their own unlearning algorithm, called SISA, and show empirical improvements over other baseline unlearning methods.