CY 7790

Special Topics in Security and Privacy: Machine Learning Security and Privacy
Fall 2021

Alina Oprea
Associate Professor
Khoury College of Computer Science

November 22 2021

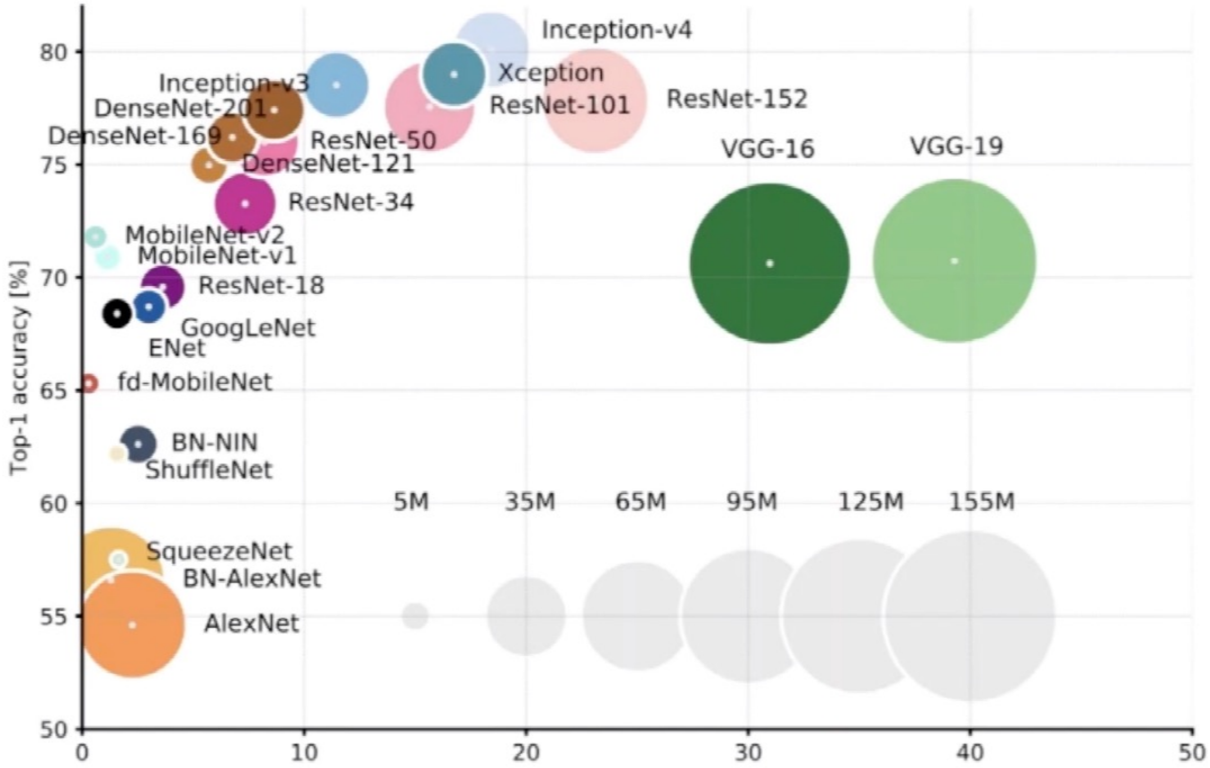
Machine Unlearning

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, Nicolas Papernot

IEEE Security and Privacy Symposium 2021

Slides courtesy of authors

Large Models



Large Models

Overparameterized Nonlinear Learning: Gradient Descent Takes the Shortest Path?

Samet Oymak[1] Mahdi Soltanolkotabi[2]

Many modern learning tasks involve fitting non-linear models which are trained in an overparameterized regime where the parameters of the model exceed the size of the training dataset. Due to

The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, Dawn Song

Observation 1: Overparameterization leads to complex interplay between data and model parameters

Legislation

New privacy legislation:

- Calls for transparency and clarity of data
- Empowers users to remove their data

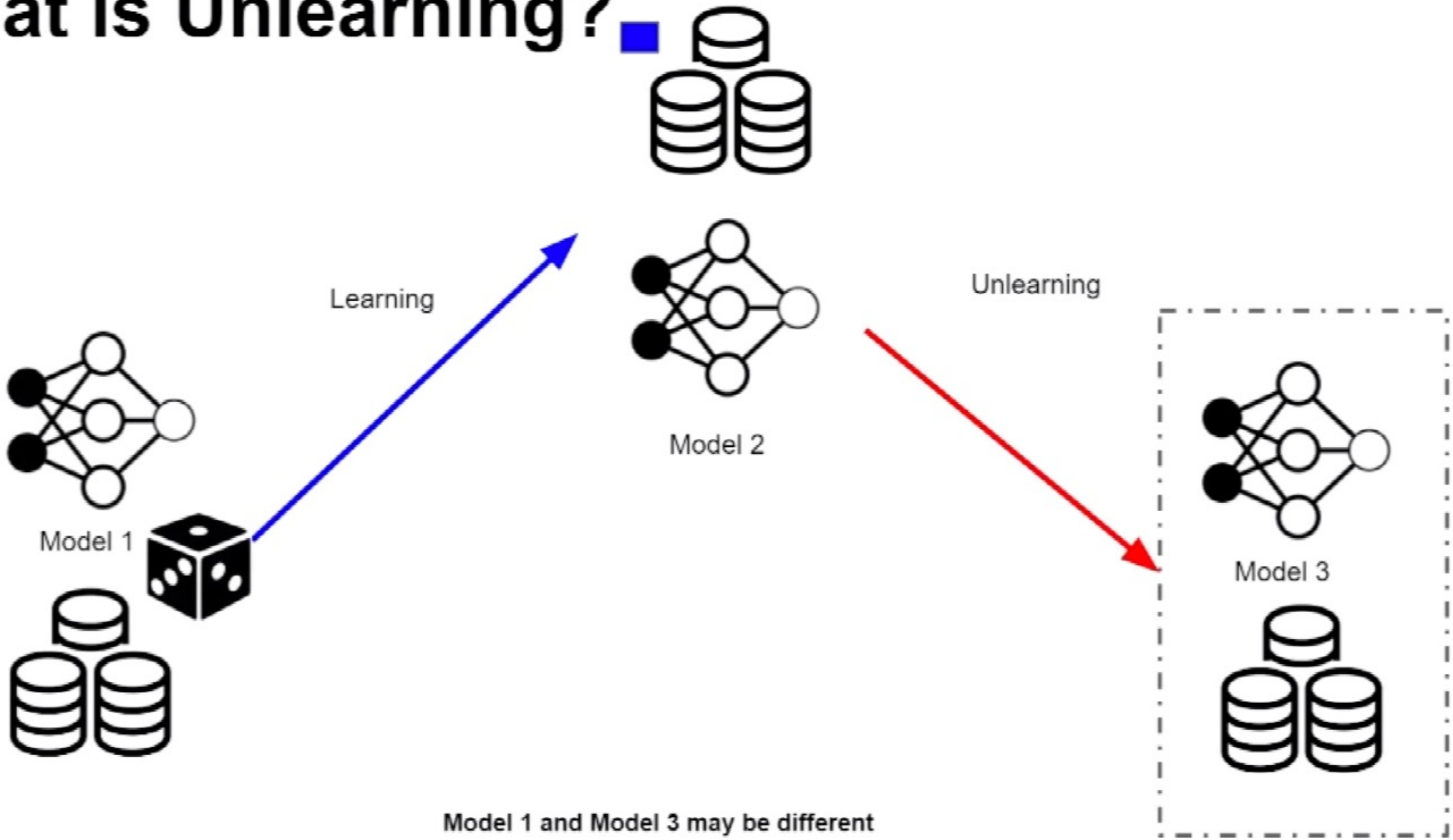
Observation 2: Disconnect between legal experts and technology experts

Right-to-be Forgotten for Machine Learning

1. Synergy missing between legal and tech experts
2. Complex interplay between data and parameters

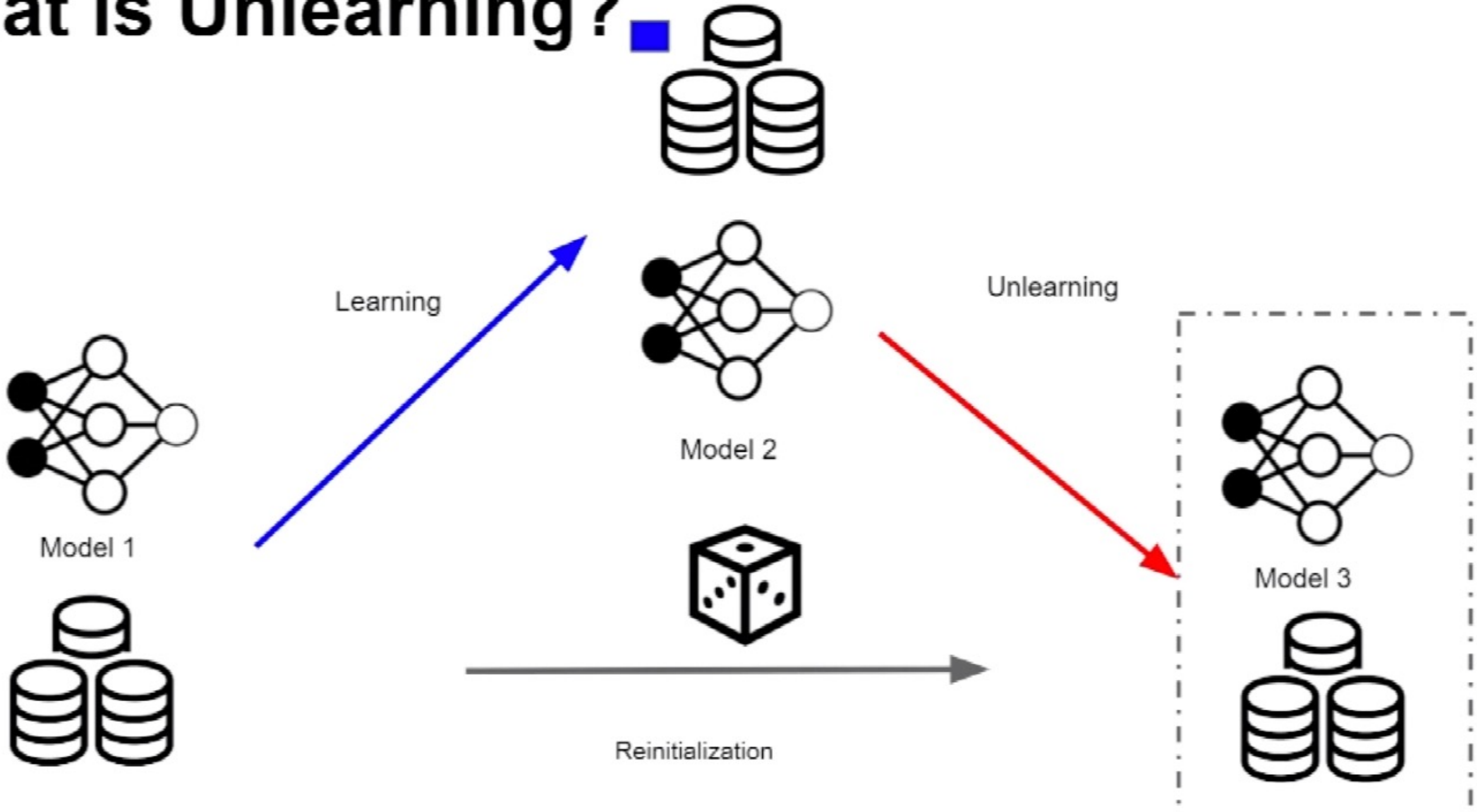
Concrete Problem: *Unlearn* data from trained ML models (e.g., DNNs) such that removal guarantee is easy to comprehend

What is Unlearning?



Model 1 and Model 3 may be different

What is Unlearning?



Model 1 and Model 3 may be different

What is ML Unlearning?

1. Distribution of models learnt after learning and then **unlearning a point** should be the **same as** the

2. Distribution of models learnt through **random re-initialization without the point**

Existing Solutions

Differentially Private Learning [Abadi et al., 2016]

- Requires $\varepsilon=0$ for compliance
- Strongly influences accuracy
- Guarantee is probabilistic

Statistical Query Learning [Cao et al., 2015]

- Applicable for simple models
- Can make limited number of queries
- No known algorithm for DL models

Goals

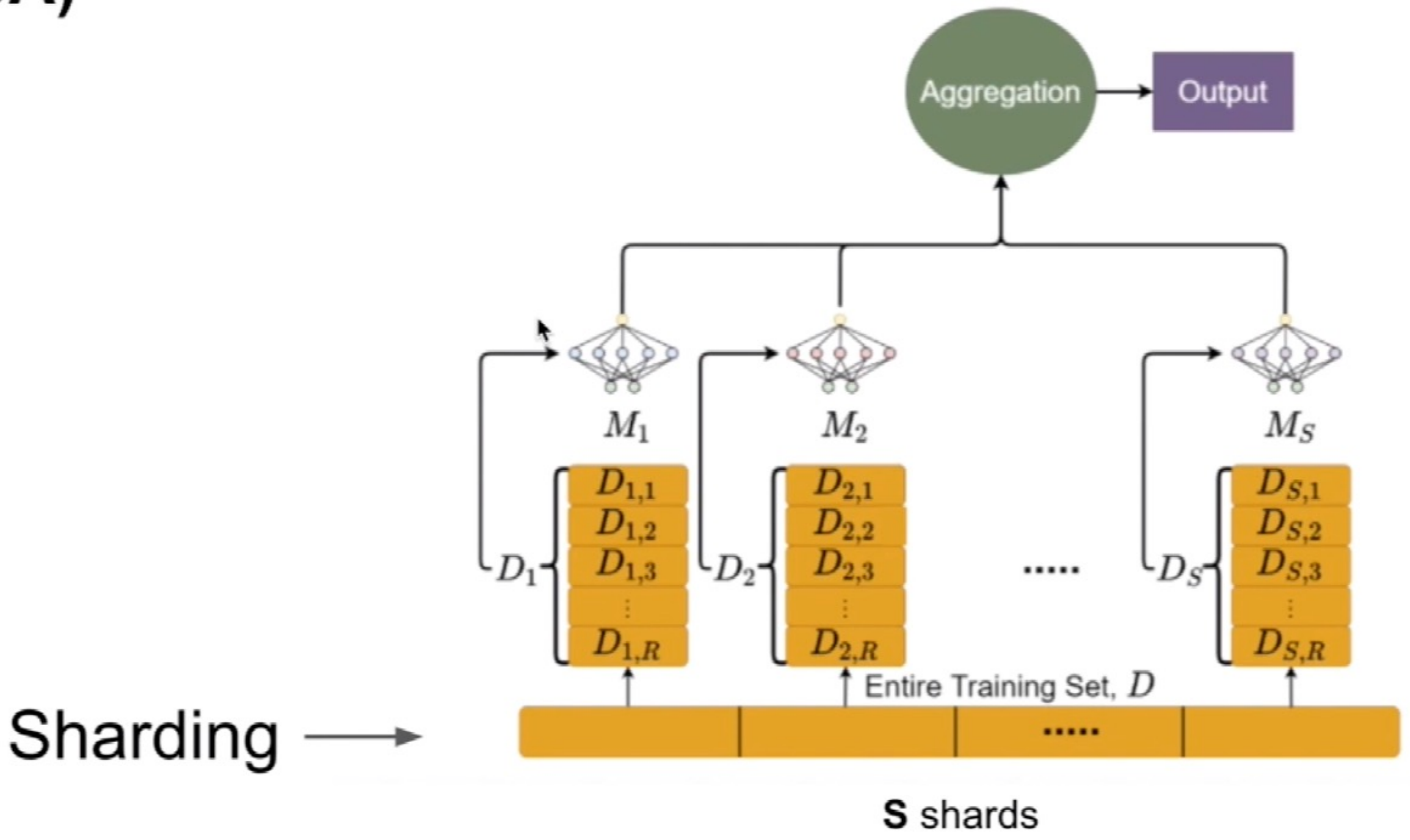
Naive Solution: Remove data point & retrain model from scratch

Intuitive, Simple to Implement, Interpretable

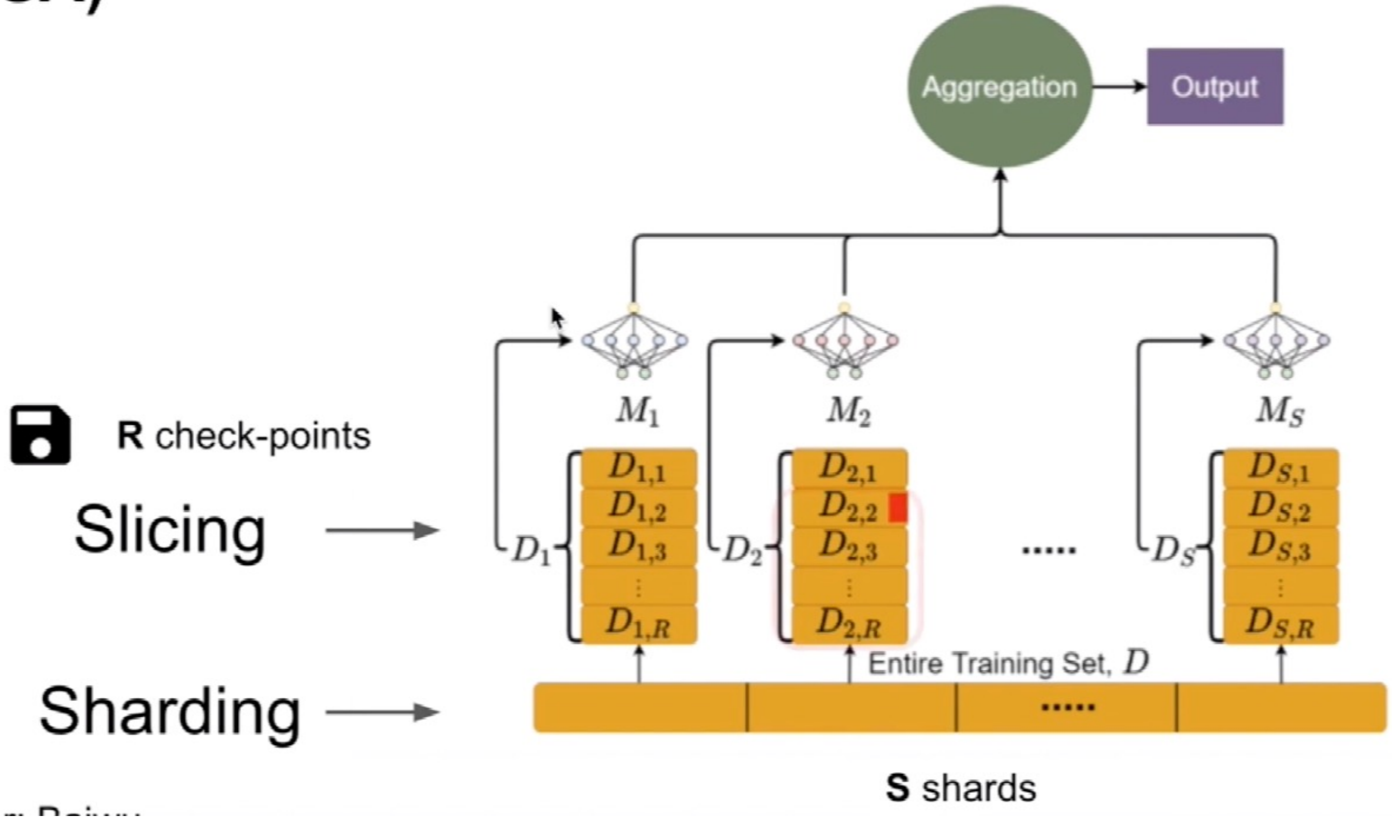
New problem: Such an approach is **very slow**

- How to obtain a faster solution than retraining?
- Assumption: Store all training data
- Threat model: No adversary! Users can submit unlearning requests

Sharded, Isolated, Sliced, and Aggregated Training (SISA)



Sharded, Isolated, Sliced, and Aggregated Training (SISA)



Tunable Knobs

Tuneable Knob	Retraining speed-up	Storage Cost	Accuracy
Sharding	↑		↓
Slicing			
Aggregation Strategy			

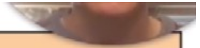
Tunable Knobs

Tuneable Knob	Retraining speed-up	Storage Cost	Accuracy
Sharding	↑		↓
Slicing	↑	↓	
Aggregation Strategy			

Tunable Knobs

Tuneable Knob	Retraining speed-up	Storage Cost	Accuracy
Sharding	↑		↓
Slicing	↑	↓	
Aggregation Strategy			↑

SISA: The Good and the Bad



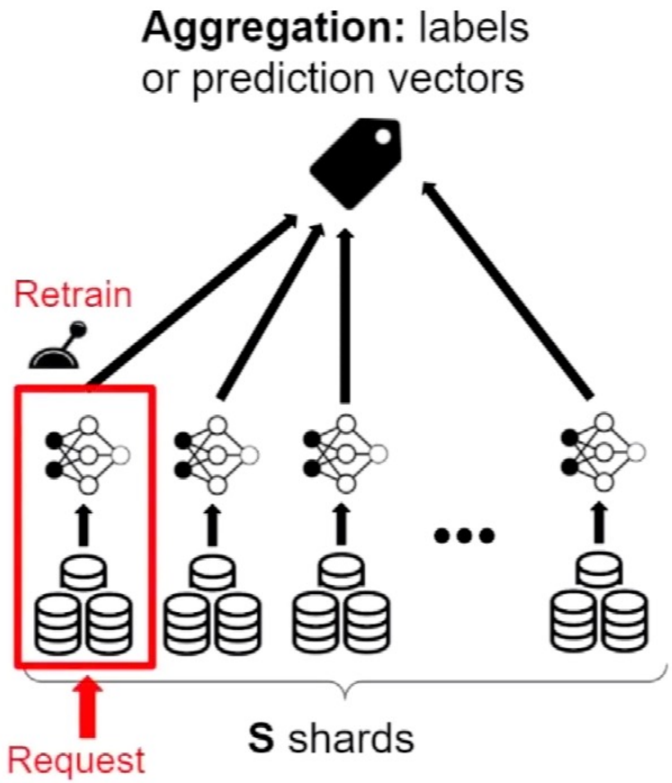
The Good	The Bad
• General • Intuitive • Provable • Auditable	• Models may disagree • Weak Learners

Experimental Setup

Assumption: Unlearning requests are uniformly random across shards

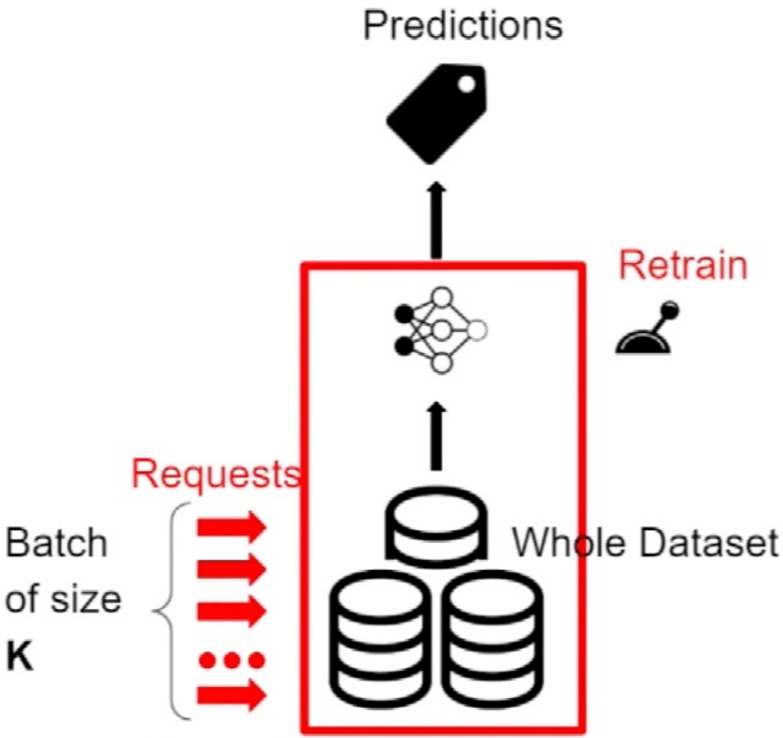
Dataset		Model Architecture
MNIST [43]	Easy	2 conv. layers followed by 2 FC layers
Purchase [49]		2 FC layers
SVHN [50]		Wide ResNet-1-1
CIFAR-100 [51]	Complex	ResNet-50
Imagenet [44]		ResNet-50
Mini-Imagenet [48]		ResNet-50

Impact of Sharding: Setup

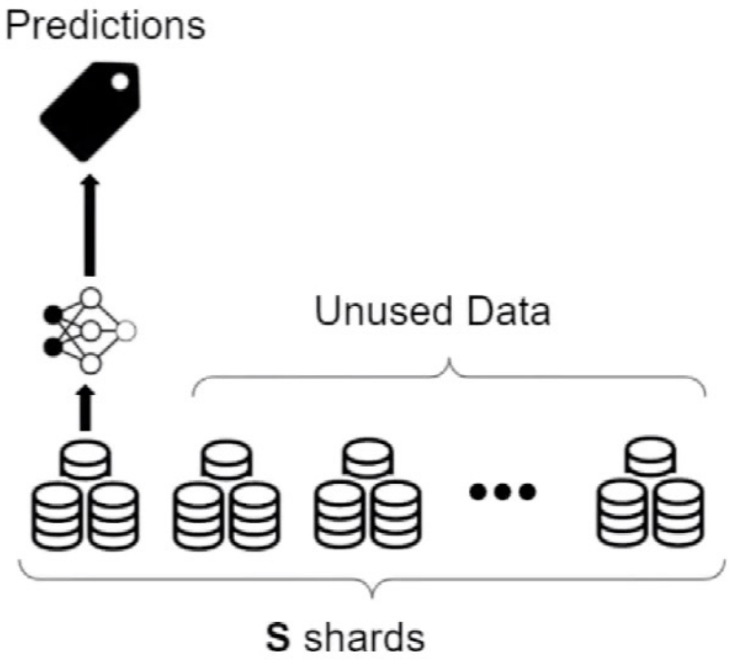


Impact of Sharding: Baselines

Batch **K** Baseline



1/S Fraction Baseline



Impact of Sharding: Results

S: number of shards

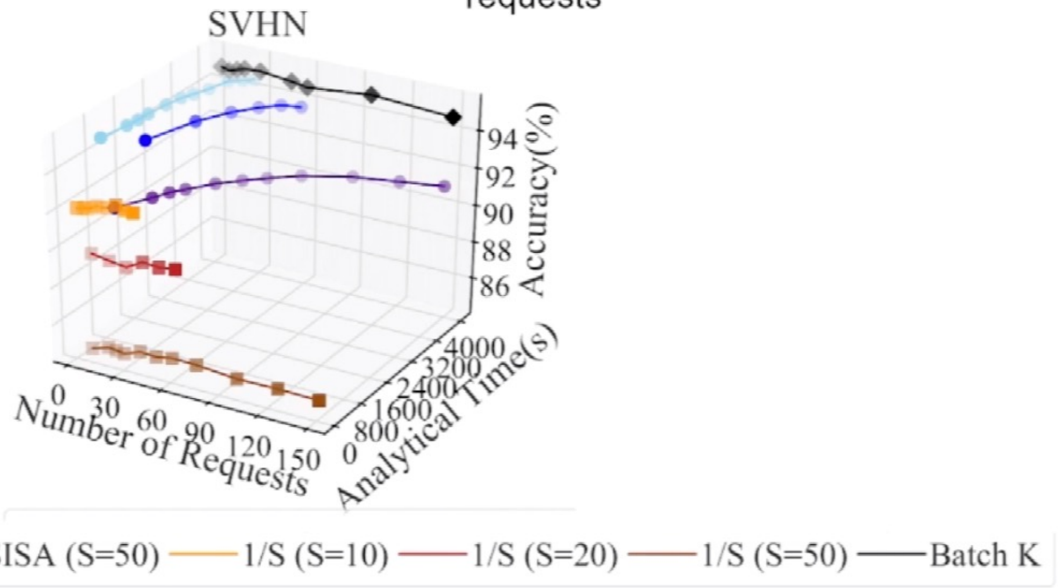
n: number of unlearning requests

1. Does increasing ‘**S**’ improve retraining speed-up?
2. When does SISA accuracy degrade too much?

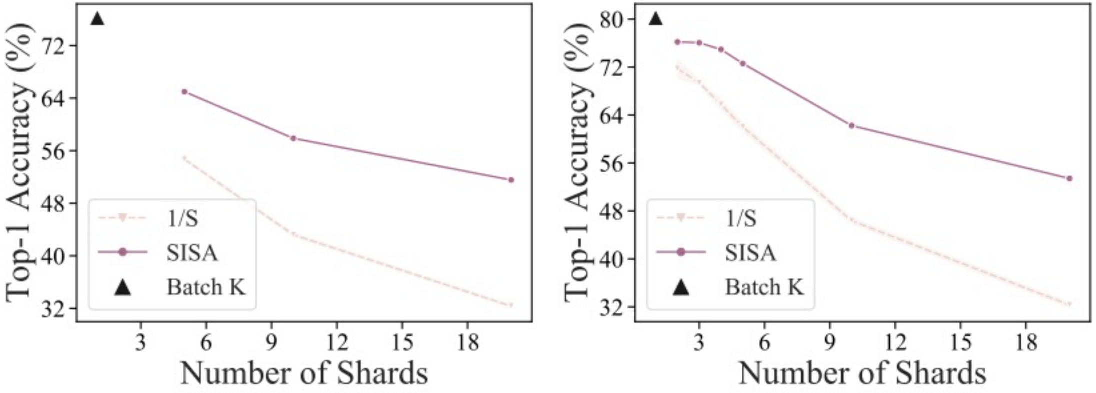
Impact of Sharding: Results

S: number of shards

n: number of unlearning requests



Complex Learning Tasks



(a) Imagenet dataset (b) Mini-Imagenet dataset

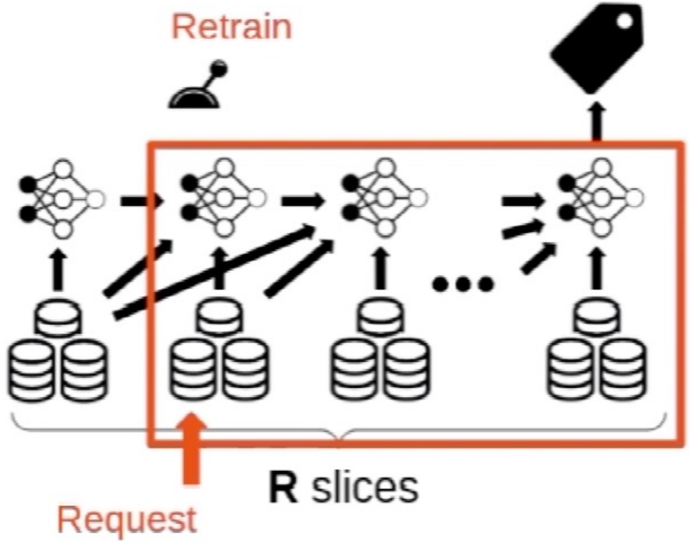
Fig. 6: For complex learning tasks such as those involving Imagenet and Mini-Imagenet, **SISA** training introduces a larger accuracy gap in comparison to the batch K baseline. However, it is still more performant than the $\frac{1}{S}$ fraction baseline. Each constituent (and baseline) utilized the prediction vector aggregation strategy.

Impact of Slicing: Setup

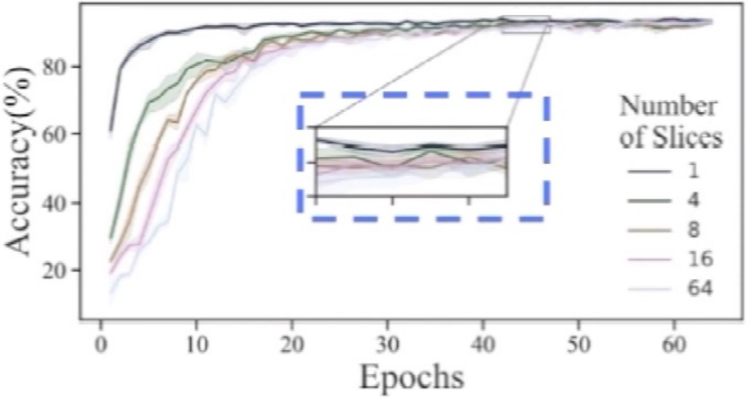
Assumption: constant training time by varying number of epochs

Evaluation:

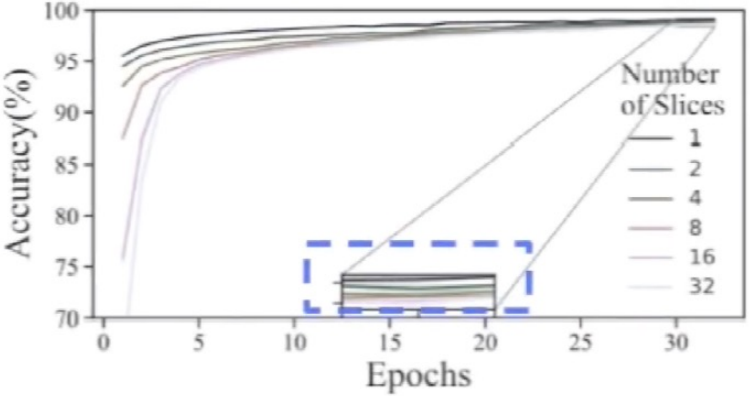
- Measured accuracy with respect to epochs
- Contrast analytical retraining time with the number of slices



Impact of Slicing: Accuracy Results



(a) Accuracy vs. Number of epochs for SVHN dataset.



(b) Accuracy vs. Number of epochs for Purchase dataset.

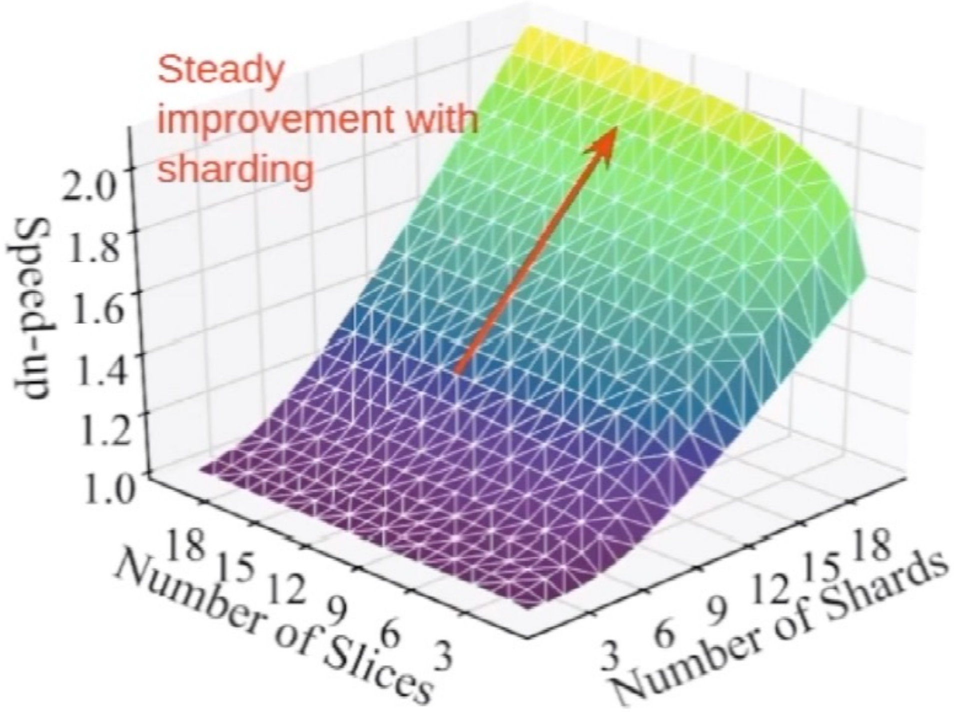
1. For **same accuracy**: **more slices** implies **more epochs**

-> artifact of our training procedure

2. For **more slices**: after sufficient training**, negligible accuracy drop**

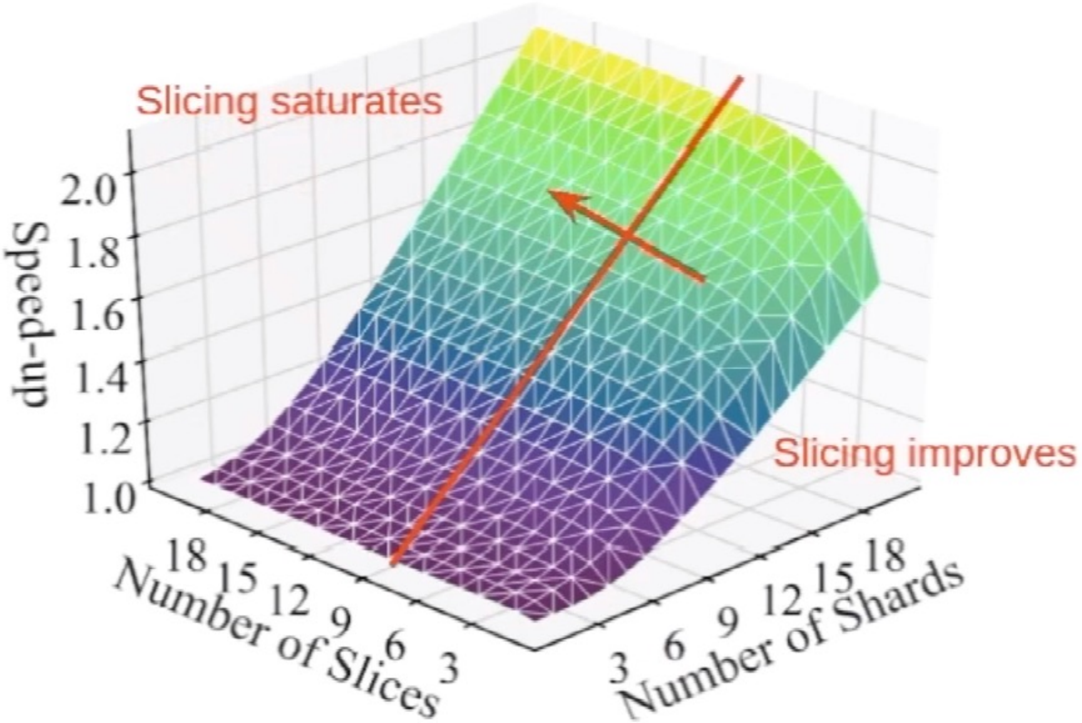
Combined Speed-up of Sharding & Slicing

Setting: unlearn one batch representing 0.003% of data



Combined Speed-up of Sharding & Slicing

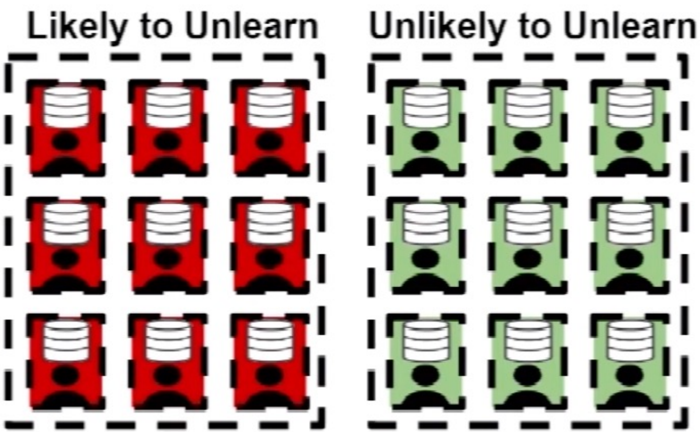
Setting: unlearn one batch representing 0.003% of data

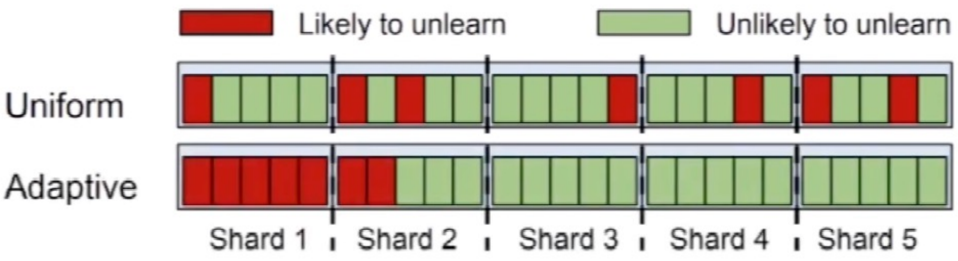


A priori Knowledge Can Improve SISA Unlearning

A user's probability for requesting unlearning may depend on **auxiliary data**:

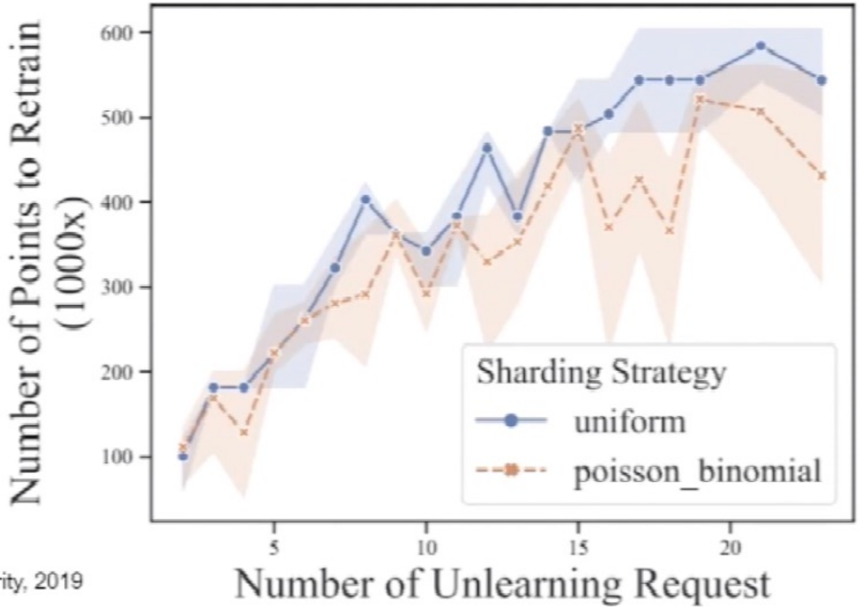
- How data is used and by whom
- The local perception surrounding data use
- Prior data misuse incidents





Distribution-Aware Sharding Performance

Modeled unlearning requests for search engines , from
“Five years of the right to be forgotten” by T. Bertram et. al. *



* Conference on Computer and Communications Security, 2019

Strengths

SISA Takeaways

- **Guaranteed** and **provable** unlearning
- **Easy to understand** by non-experts
- **Reduces retraining time** consistently.
- **Minimal overhead** in storage and training algorithm changes
- **Applicable** to any model trained by gradient descent
- Can leverage knowledge of the **distribution of unlearning requests.**

Limitations

- Needs to store entire training dataset
- Technique similar to ensemble learning, but uses disjoint datasets for each model
- Accuracy drop on more complex learning tasks

Analyzing Information Leakage of Updates to Natural Language Models

Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, Marc Brockschmidt

Conference on Computer and Communications Security '20

Presented by Hye Sun Yun
November 22, 2021

Problem Statement

ML applications are regularly retrained and updated to improve their quality and reflect changes in data

- Data update
- Data specialization
- Data deletion

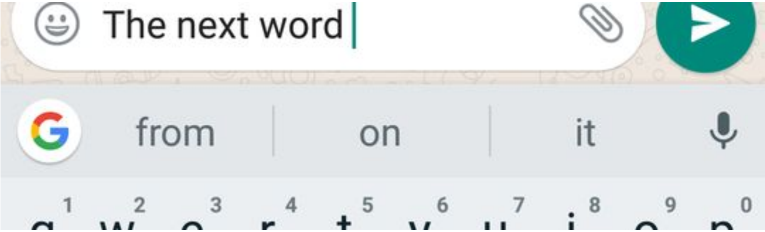
Questions:

1. *What are the privacy implications for text data that is added or removed during retraining of generative language models (LMs)?*
2. *Does honoring a request to remove a user's data from training corpus actually lead to exposing their data by releasing an updated model trained without it?*

Generative Language Models

- Have fixed set of tokens T (vocabulary)
- Are autoregressive

$$p(t_1 \dots t_n) = \prod_{1 \leq i \leq n} p(t_i \mid t_1 \dots t_{i-1})$$



- Use the standard measure of perplexity to measure performance

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}}$$

Threat Model

Goal: infer information about training data points in $D'\backslash D$ (difference between D and D')

Knowledge: has concurrent query access to two snapshots, M_D and $M_{D'}$, of a language model trained on datasets D and D', where $D \subseteq D'$

Capability: can query the snapshots with any sequence $s \in T^*$ and observe the corresponding probability distributions $M(s)$ and $M'(s)$

Proposed Analysis Method

- Metrics to help measure and analyze data exposure (update leakage) in generative language models: **differential score** and **differential rank**
  - Aim is to identify token sequences whose probability differs most between models M and M'
  - Intuition - sequences whose probability differs most are likely to be related to the differences between their corresponding training datasets D and D'
- Use these metrics to perform leakage analysis and show that update leakage is possible when snapshots are available.

Differential Score

- Differential Score (DS) of token sequences is simply sum of the differences of contextualized per-token probabilities.
- Relative version of DS is based on the relative change in probabilities

Definition 3.1. Given two language models M, M' and a token sequence $t_1 \ldots t_n \in T^*$, we define the *differential score* of a token as the increase in its probability and the *relative differential score* as the relative increase in its probability. We lift these concepts to token sequences by defining

$$DS_M^{M'}(t_1 \ldots t_n) = \sum_{i=1}^{n} M'(t_{<i})(t_i) - M(t_{<i})(t_i),$$

$$\widetilde{DS}_M^{M'}(t_1 \ldots t_n) = \sum_{i=1}^{n} \frac{M'(t_{<i})(t_i) - M(t_{<i})(t_i)}{M(t_{<i})(t_i)}.$$

Differential Rank

Definition 3.2. We define the *differential rank* $DR(s)$ of $s \in T^*$ as the number of token sequences of length $|s|$ with differential score higher than s.

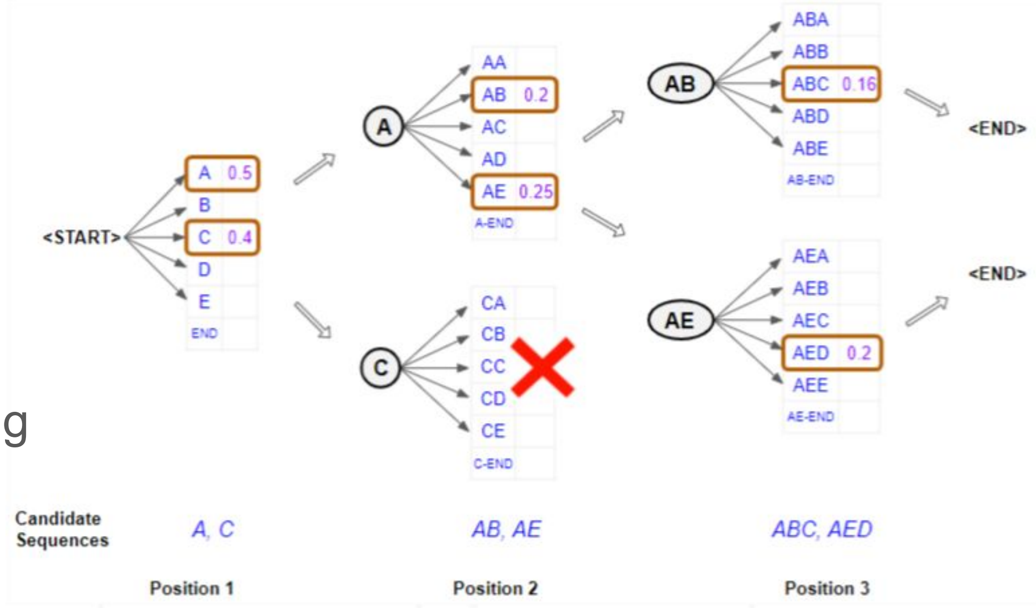
$$DR(s) = \left| \left\{ s' \in T^{|s|} \middle| DS_M^{M'}(s') > DS_M^{M'}(s) \right\} \right| .$$

The lower the differential rank of a sequence, the more the sequence is exposed by a model update, with the most exposed sequence having rank 0.

Beam Search

- Popular algorithm to produce final output text from language models
- Requires more computation than greedy search but better results
- Picks the *k*-best sequences so far and considers the probabilities of the combination of all the preceding words along with the word in current position.

Beam width of 2



https://towardsdatascience.com/foundations-of-nlp-explained-visually-beam-search-how-it-works-1586b9849a24

Approximating Differential Rank

Algorithm 1 Beam search for Differential Rank

In: M, M'=models, T=tokens, k=beam width, n=length

Out: S=set of (n-gram, DS) pairs

1: $S \leftarrow \{(\epsilon, 0)\}$ ▷ Initialize with empty sequence ϵ

2: **for** $i = 1 \dots n$ **do**

3: $\quad S' \leftarrow \{(s \circ t, r + DS_M^{M'}(s)(t)) \mid (s, r) \in S, t \in T\}$

4: $\quad S \leftarrow take(k, S')$ ▷ Take top k items from S'

5: **return** $S = \{(s_1, r_1), \dots, (s_k, r_k)\}$ such that $r_1 \geq \dots \geq r_k$

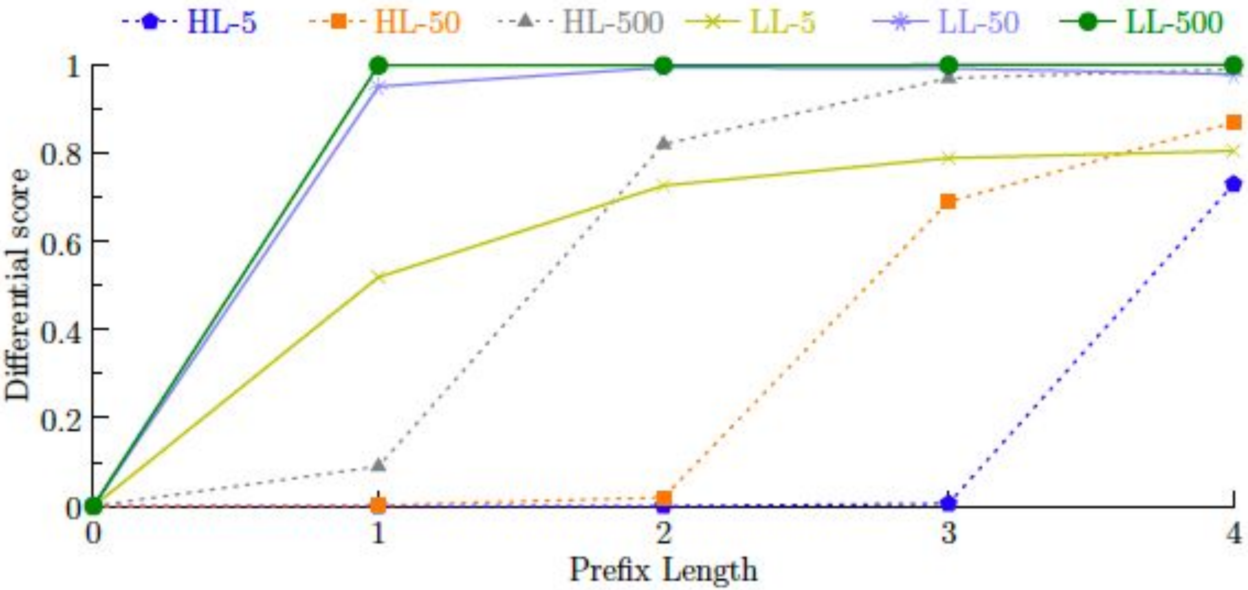
Leakage Analysis

- Datasets
  - Penn Treebank (PTB) - 900,000 tokens and a vocab size of 10,000
  - Reddit comments with 2 million tokens and vocab size of 10,000
  - Wikitext-103 with 103 million tokens and vocab size of 20,000
- Models
  - RNN using LSTM cell
  - BERT-based Transformer architecture
- Cases
  - Canaries - grammatically correct phrases that do no appear in original dataset
    - Different amounts for different datasets
    - *All low, mixed, increasing from low to high, decreasing from high to low*
  - Real-world Data
    - Real-world conversations on specific topics like hokey and politics from Newsgroups dataset

Results with Canaries

Dataset	Penn Treebank			Reddit						Wikitext-103	
Model Type (Perplexity)	RNN (120.90)			RNN (79.63)			Transformer (69.29)			RNN (48.59)	
Canary Token Freq.	1:18K	1:3.6K	1:1.8K	1:1M	1:100K	1:10K	1:1M	1:100K	1:10K	1:1M	1:200K
All Low	3.40	3.94	3.97	2.83	3.91	3.96	3.22	3.97	3.99	1.39	3.81
Low to High	3.52	3.85	3.97	0.42	3.66	3.98	0.25	3.66	3.97	0.07	3.21
Mixed	3.02	3.61	3.90	0.23	3.04	3.92	0.39	3.25	3.96	0.25	3.02
High to Low	1.96	2.83	3.46	0.74	1.59	2.89	0.18	1.87	3.10	0.08	1.22

	Retraining				Continued Training 1			Continued Training 2
$\lvert D_{extra}\rvert / \lvert D_{orig}\rvert$	0%	20%	50%	100%	20%	50%	100%	100%
1:1M	0.23	0.224	0.223	0.229	0.52	0.34	0.46	0.01
1:100K	3.04	3.032	3.031	3.038	3.56	3.25	3.27	0.26

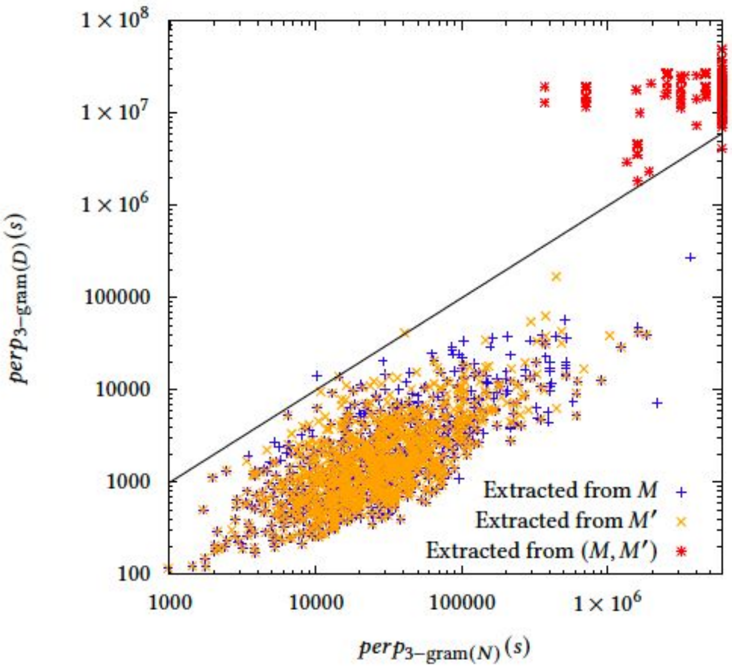

HL-5
HL-50
HL-500
LL-5
LL-50
LL-500
Differential score
0
0.2
0.4
0.6
0.8
1
0
1
2
3
4
Prefix Length

Results with Real-world Data

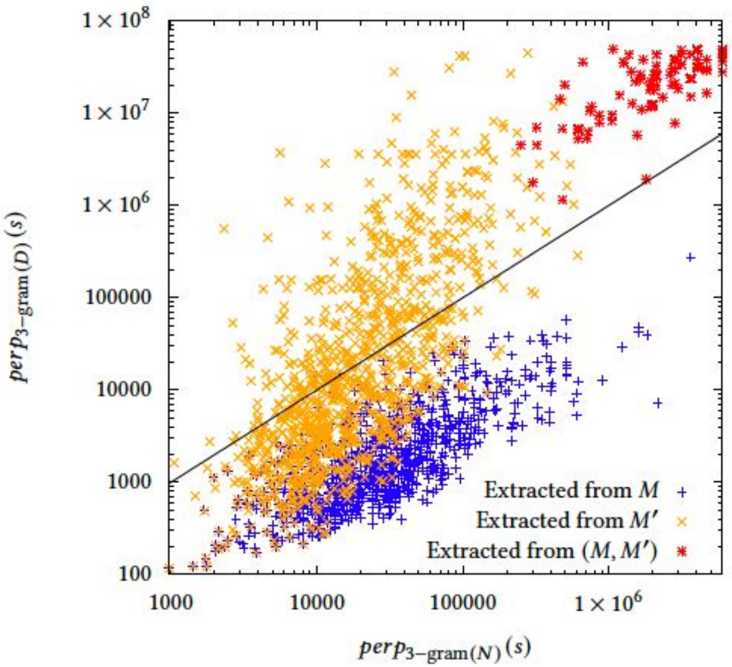
RNN Phrase	$\widetilde{DS}$	Transformer Phrase	$\widetilde{DS}$
Angeles Kings prize pools	56.42	Minnesota North Stars playoff	96.81
National Hockey League champions	53.68	Arsenal Maple Leaf fans	71.88
Norm 's advocate is	39.66	Overtime no scoring chance	54.77
Intention you lecture me	21.59	Period 2 power play	47,85
Covering yourself basically means	21.41	Penalty shot playoff results	42.63

Phrase (# of occurrences in N)	Retraining					Continued Training				
$\lvert D_{extra}\rvert / \lvert D_{orig}\rvert$	0%	5%	10%	20%	100%	0%	5%	10%	20%	100%
Perplexity decrease	0.79	1.17	2.45	3.82	11.82	73.97	18.45	10.29	6.08	8.28
`Center for Policy Research` (93)	99.77	101.38	97.11	98.65	91.53	**276.98**	**198.69**	**150.56**	**122.25**	**117.54**
`Troops surrounded village after` (12)	44.50	44.50	44.50	44.41	44.54	**173.95**	**47.38**	19.48	7.81	35.56
`Partition of northern Israel` (0)	27.61	16.81	38.48	26.10	38.76	**68.98**	16.48	12.47	22.93	18.82
`West Bank peace talks` (0)	25.68	25.64	25.69	25.71	25.75	**71.54**	24.38	**28.60**	16.91	4.62
`Spiritual and political leaders` (0)	25.23	25.98	17.04	24.21	23.47	**126.92**	14.91	10.00	3.44	11.05
`Saudi troops surrounded village` (0)	24.31	24.31	24.31	24.31	24.30	5.05	**44.58**	4.29	7.29	**63.84**
`Arab governments invaded Turkey` (0)	22.59	22.62	22.80	22.78	22.80	**24.01**	15.58	7.08	18.12	11.90
`Little resistance was offered` (12)	22.24	22.09	25.12	22.34	25.59	**215.16**	**25.02**	2.00	3.30	5.64
`Buffer zone aimed at protecting` (0)	4.00	4.47	5.30	5.25	5.69	**57.29**	**69.76**	**18.92**	**14.50**	**22.25**
`Capital letters racial discrimination` (0)	3.76	3.32	3.40	3.60	3.84	**94.60**	**52.74**	**39.11**	**11.22**	3.45

Characterizing the Source of the Leakage



(a) Re-training from scratch



(b) Continued training

Extracted phrase	talk.politics.mideast		Reddit	
center for policy research	center for policy research	0	center for instant research	1
troops surrounded village after	troops surrounded village after	0	from the village after	2
partition of northern israel	shelling of northern israel	1	annexation of northern greece	2
west bank peace talks	. no peace talks	2	: stated peace talks	2
spiritual and political leaders	spiritual and political evolutions	1	, and like leaders	2
saudi troops surrounded village	our troops surrounded village	1	" hometown " village	3
arab governments invaded turkey	arab governments are not	2	! or wrap turkey	3
little resistance was offered	little resistance was offered	0	, i was offered	2
buffer zone aimed at protecting	" aimed at protecting	2	's aimed at a	3
capital letters racial discrimination	% of racial discrimination	2	allegory for racial discrimination	2

Mitigations

- Differential Privacy
  - Significant model degradation and substantial computational overhead
- Two-stage Continued Training
  - Add an additional step of training on another dataset - attacker does not have two consecutive snapshots
  - Might be path towards mitigating leakage
- Truncating Output
  - Only returns the top *k* tokens from the updated model *M'*
  - Further reduce leakage when original *M* returns top *k*
  - Can mitigate leakage without decreasing the utility

Conclusion

- First study of privacy implications of releasing snapshots of language models trained on overlapped data
- Provides two metrics for measuring information leakage in generative language models
- Analysis results show that model updates pose substantial risk to information leakage

Strengths

- Proposes a new type of attack using two snapshots and provides some mitigation approaches
- Adversary does not require auxiliary dataset nor have to know the contents of training dataset or dataset distribution
- Two metrics that can be used for any generative language model (model agnostic)

Weaknesses

- Paper was not very clear on their definition of tokens. Assuming that it is word-level based on how they calculate Lavenshtein distance.
- Figuring out the beam width can be key in setting up a large enough search space.
- BERT was used to implement their models which was confusing. BERT isn't necessarily often used for generative tasks so wanted to see their source code for implementation but was not available.
- For real-world data analysis, they chose conversations on topics that are different from original dataset. Wondering if this setup actually makes it easier for the model to unintentionally memorize the new data. (future work)