Revenue-Preserving Unlearning for Recommendation Systems

Sam Levy[1] **Ahson Saiyed**[2] **Chirag Agarwal**[2] **Alan Montgomery**[3]

[1]Darden School of Business, UVA [2]School of Data Science, UVA [3]Tepper School of Business, Carnegie Mellon University

Early Career Consortium, AMA Sheth Foundation 2026

Privacy Regulation Is Here, But Compliance Destroys Revenue

5.6M	\$1,524	46%
deletion requests (Amazon, 2024)	avg cost per request	utility loss w/ current methods

Imagine you're a customer...

- You click "delete my data," but does the model actually forget?
- GDPR/CCPA: "right to be forgotten"; 76% would switch for transparency

Now imagine you're a business...

- **Data poisoning**: 0.5% fake users can boost a target item 150×
- You need to remove that data *fast*, but how?

State of the Art

SISA: Shards data randomly, retrains affected shard, but destroys collaborative signals.

RecEraser: Smarter partitioning, but still 2–3h retraining per request.

Both optimize for *speed* or *privacy*, neither measures what firms care about: **revenue impact**.

Research Question

Can we unlearn user data *surgically*, preserving **revenue** while ensuring **privacy**? *Data: ComScore (635K events, 45K users, \$709M revenue)*

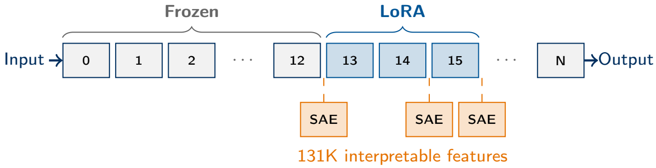
Our Approach: SAE-Guided Surgical Unlearning

What Are SAEs?

Neural networks are "black boxes," user knowledge is spread across millions of parameters.

Sparse Autoencoders act like an "X-ray": they decompose opaque representations into **131K interpretable features**, each mapping to a concept (e.g. "electronics shopper").

This lets us *see* which features encode a specific user and target only those.



Three Steps

① **Identify**: SAEs find user-specific features
② **Suppress**: LoRA targets only those features
③ **Preserve**: Frozen layers protect other users

Optimization Objective

$$\mathcal{L}_{\text{total}} = \underbrace{\mathcal{L}_{\text{unlearn}}}_{\text{suppress}} + \underbrace{\mathcal{L}_{\text{retain}}}_{\text{preserve}} + \underbrace{\mathcal{L}_{\text{coherence}}}_{\text{fluency}}$$

, balancing privacy (removing the right info) with utility (keeping recs accurate).

Results: 20–40× Faster, Best Revenue Preservation

	RecEraser (2–3h)			**SISA** (4–5h)			**Ours** (3–10 min)		
	R@5	N@5	Rev%	R@5	N@5	Rev%	R@5	N@5	Rev%
Baseline	5.71	4.05	–	3.59	2.98	–	5.96	4.92	–
10% forget	3.89	2.20	93	1.93	1.62	20	**3.91**	**3.22**	57
20% forget	4.03	2.26	37	2.40	1.79	34	3.23	**2.42**	**71**

Metrics

R@5 (Recall@5): fraction of relevant items in the top-5 recommendations.

N@5 (NDCG@5): ranking quality, higher means better items ranked first.

Rev%: share of original revenue preserved after unlearning.

RecEraser

- Exact unlearning
- 68% utility retained
- Hours of retraining

SISA

- Exact unlearning
- 54% utility retained
- Sharding hurts RecSys

Ours

- Approx. unlearning
- **Best NDCG + revenue**
- **20–40× faster**

Conclusion

Our SAE-guided approach suppresses 97% of user-specific features with only 3% quality loss, preserving twice the revenue of existing methods while running 20–40× faster, demonstrating that privacy compliance and revenue preservation can coexist.