Approximate Data Deletion from Machine Learning Algorithms

AISTATS 2021

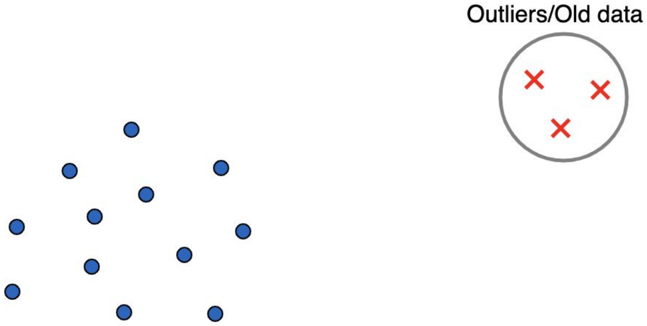
Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, James Zou

Stanford University

UC San Diego

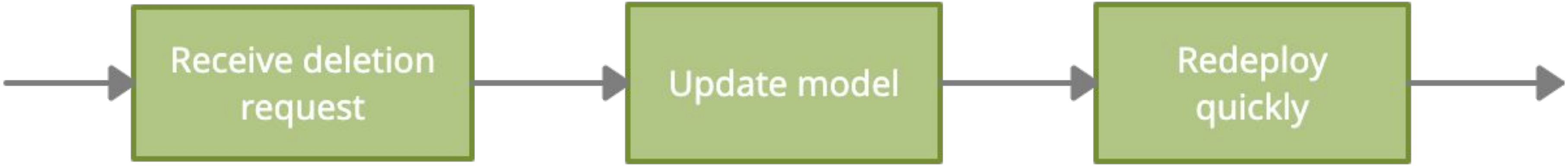
The Need for Data Deletion

- Recent legislation requires companies to comply with data deletion requests
- Problem: Data still exists implicitly in any model trained on it
- Remove outliers/old data



Example

- NLP model with bag-of-words features
- Data dimension = dictionary size → high dimensional
- Honor deletion requests while continuing to provide service for other users



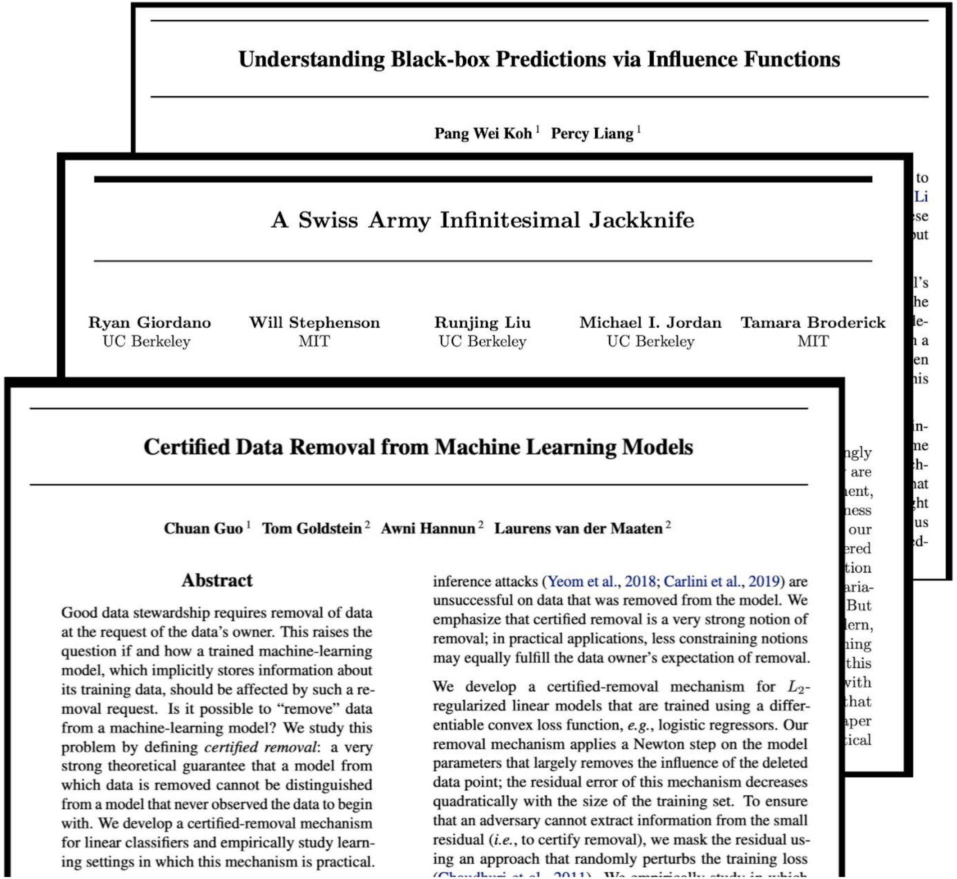
Naive deletion by fully retraining on the new dataset is too slow!

Existing Methods

- Previous methods use a Newton step or approximation thereof
- Runtimes scale with square of data dimension

Main Question:

How can we approximately retrain high-dimensional models both accurately and efficiently?



Understanding Black-box Predictions via Influence Functions

Pang Wei Koh[1] Percy Liang[1]

A Swiss Army Infinitesimal Jackknife

Ryan Giordano, UC Berkeley; Will Stephenson, MIT; Runjing Liu, UC Berkeley; Michael I. Jordan, UC Berkeley; Tamara Broderick, MIT

Certified Data Removal from Machine Learning Models

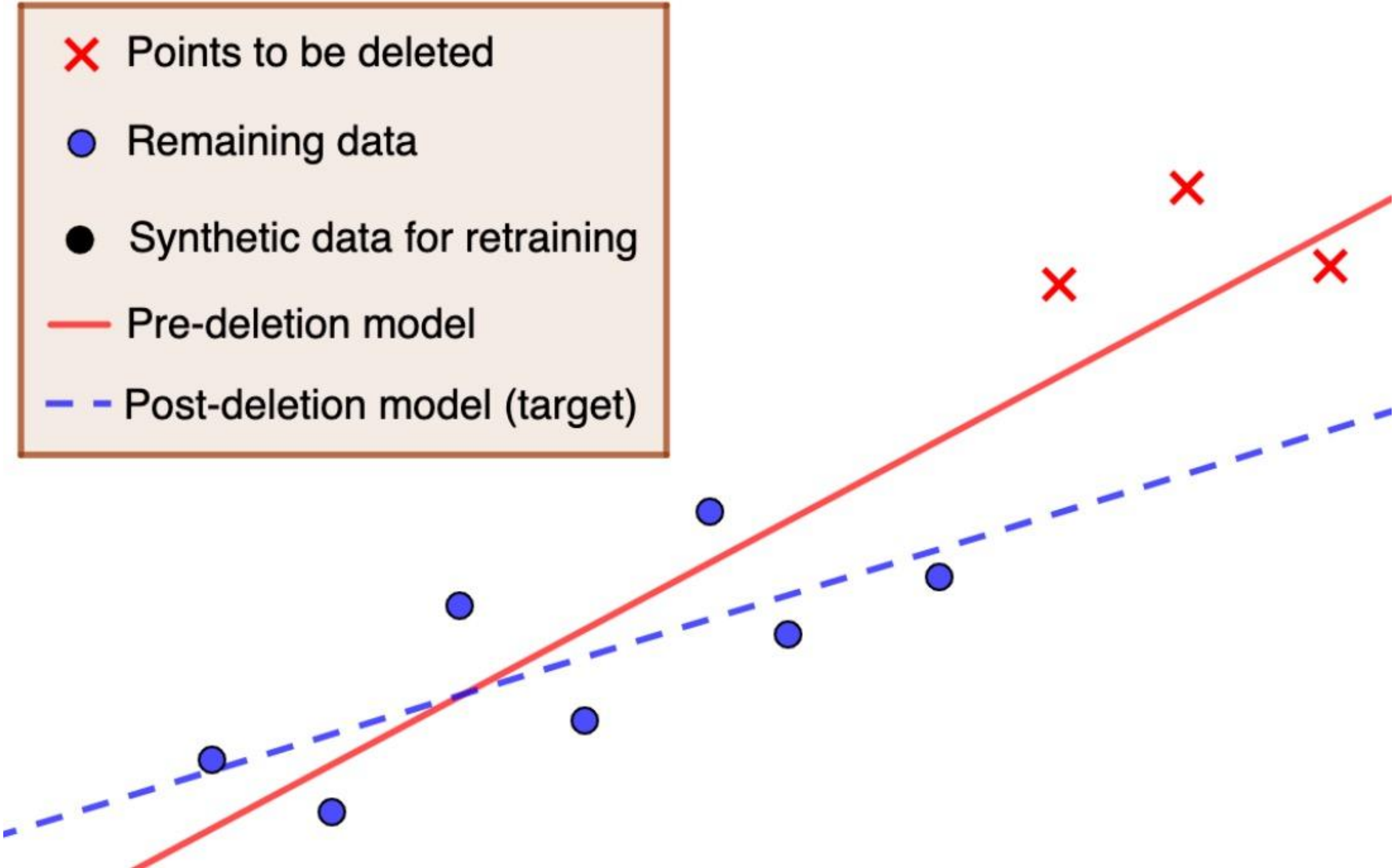
Chuan Guo[1] Tom Goldstein[2] Awni Hannun[2] Laurens van der Maaten[2]

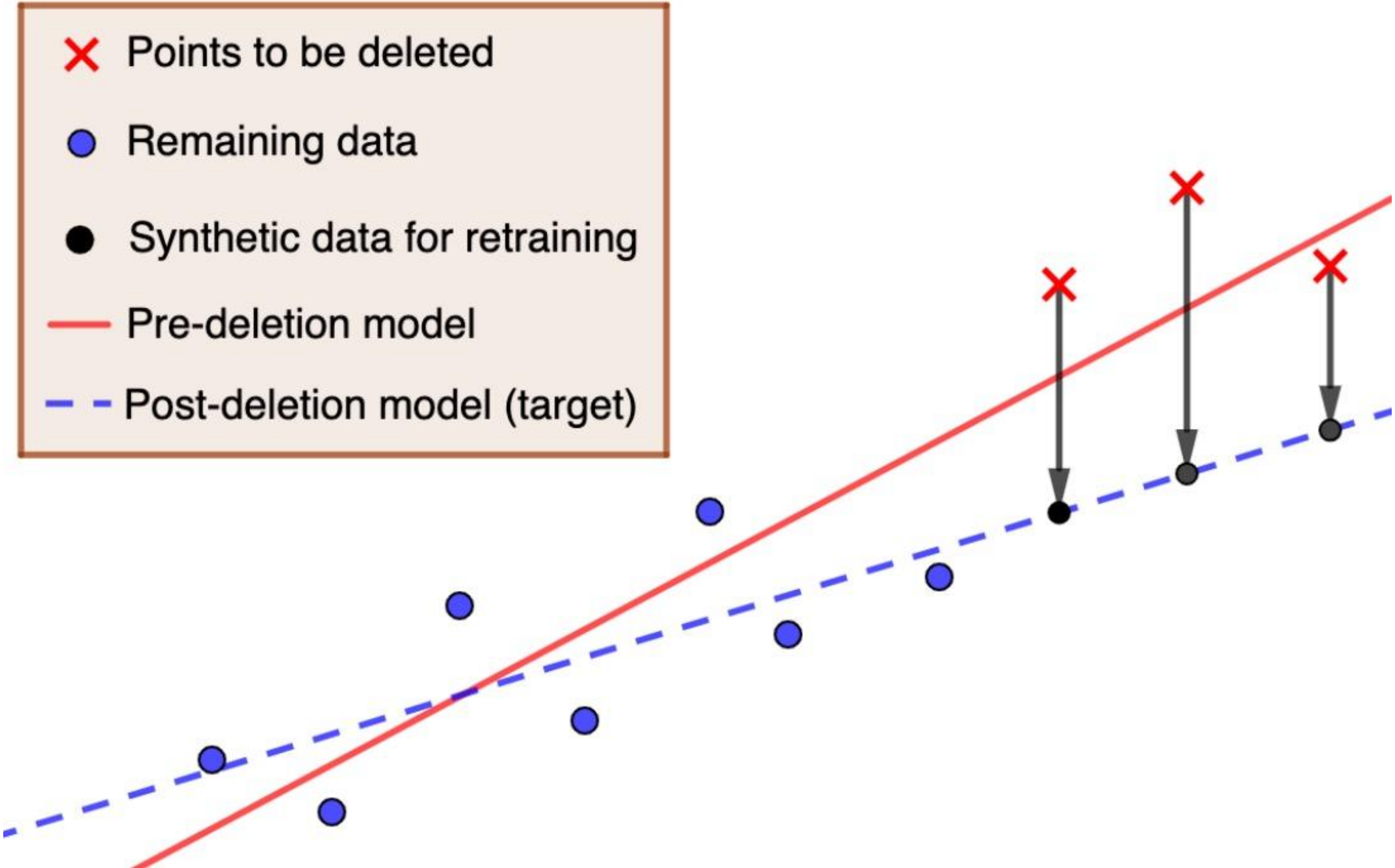
Abstract

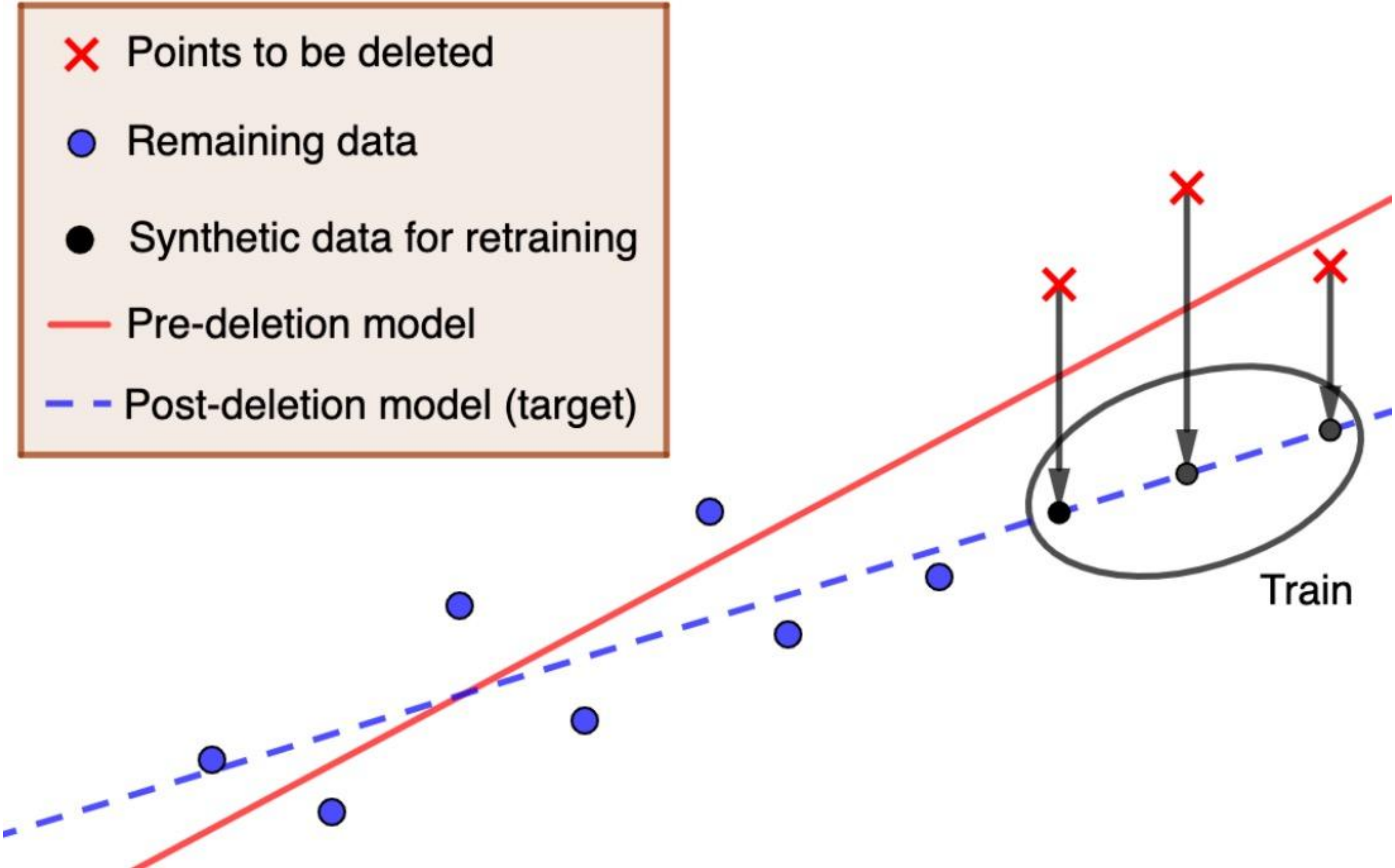
Good data stewardship requires removal of data at the request of the data's owner. This raises the question if and how a trained machine-learning model, which implicitly stores information about its training data, should be affected by such a removal request. Is it possible to "remove" data from a machine-learning model? We study this problem by defining *certified removal*: a very strong theoretical guarantee that a model from which data is removed cannot be distinguished from a model that never observed the data to begin with. We develop a certified-removal mechanism for linear classifiers and empirically study learning settings in which this mechanism is practical.

inference attacks (Yeom et al., 2018; Carlini et al., 2019) are unsuccessful on data that was removed from the model. We emphasize that certified removal is a very strong notion of removal; in practical applications, less constraining notions may equally fulfill the data owner's expectation of removal.

We develop a certified-removal mechanism for L_2-regularized linear models that are trained using a differentiable convex loss function, *e.g.*, logistic regressors. Our removal mechanism applies a Newton step on the model parameters that largely removes the influence of the deleted data point; the residual error of this mechanism decreases quadratically with the size of the training set. To ensure that an adversary cannot extract information from the small residual (*i.e.*, to certify removal), we mask the residual using an approach that randomly perturbs the training loss (Chaudhuri et al., 2011). We empirically study in which


Points to be deleted
Remaining data
Synthetic data for retraining
Pre-deletion model
Post-deletion model (target)


× Points to be deleted
● Remaining data
● Synthetic data for retraining
— Pre-deletion model
- - Post-deletion model (target)



× Points to be deleted
● Remaining data
● Synthetic data for retraining
Pre-deletion model
Post-deletion model (target)
Train

Main Theorem (Informal):

The PRU computes the projection of the true model update vector onto a data-dependent subspace. The runtime scales linearly in data dimension.

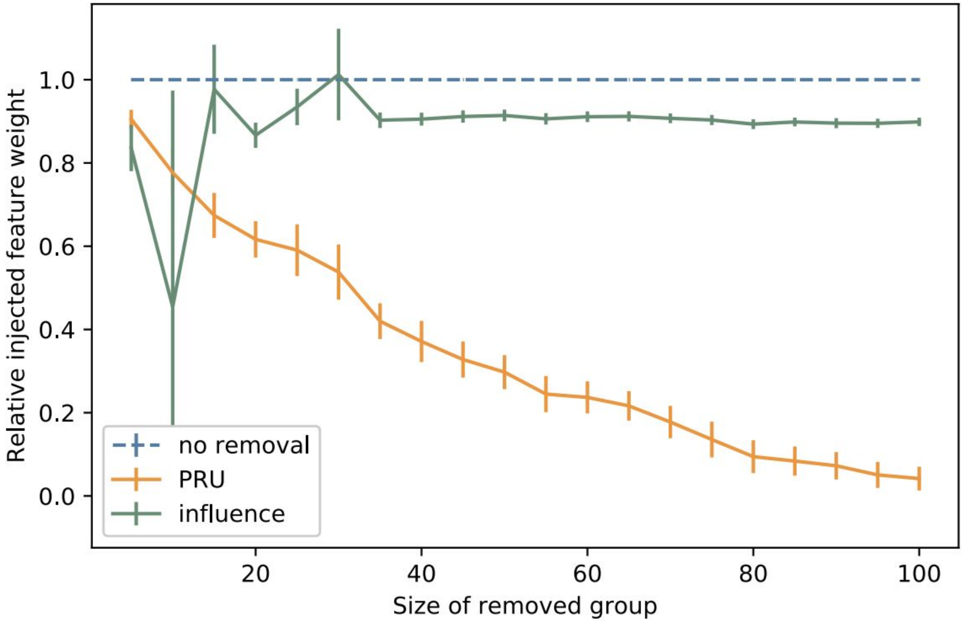
- Fast (first approx. deletion alg. with *linear* dependence on dimension!)
- Performance guarantees
- Robust to outliers

Speedup vs. exact retraining (exact retrain time / retrain time of approx.)

	d = 2000	**d = 3000**
Previous method	244	357
PRU	1250	3333

The Feature Injection Test

- FIT metric: How much of the injected signal remains after deletion?
- Lower is better, 0 is best



Conclusion

- Projective residual update: A novel approximate deletion method suitable for deleting high-dimensional data from ML models.
- Feature injection test: New metric for evaluating approximate deletion methods.
- Experiments support our theoretical findings.

For more information, come to poster session 5 (April 15, 7:30am-9:30am PDT), or contact `zizzo@stanford.edu`.