Protection and the Control of Information Sharing in Machine Learning Systems

Anonymous Submission

Abstract

Machine-learning (ML) ecosystems arising in today's organizations lack appropriate abstractions for the protection, controlled sharing, and retention of sensitive information. This results in unnecessary exposure of historical sensitive information to internal or external attackers. We propose *private transferrable knowledge* (PTK), a new unit for rigorous data protection, sharing, and retention in ML ecosystems. PTKs are feature models trained over long-term historical data and made differentially private to protect the training data. Using plausible corporate scenarios crafted around public datasets, we demonstrate the value of PTKs as a unit of protected data sharing and retention. We build *Sage*, a PTK store that creates, maintains, and shares multiple PTKs on behalf of a wide range of ML workloads. The key novel aspect in Sage is its minimal-exposure design that conflicts with the global privacy semantic it seeks to enforce for PTKs. We address this tension with new statistical and systems methods.

1 Introduction

Data-rich, machine learning (ML) ecosystems arising in today's companies, governments, and organizations forgo important principles for rigorous data protection. Consider the *least privilege principle* introduced by Saltzer in the 1974 Multics paper [**?**]: *"[e]very program and every privileged user of the system should operate using the least amount of privilege necessary to complete the job."* This principle has influenced the design of not only Multics' protection system, but also of UNIX and through that of many modern operating systems, including Linux, OSX, and Windows. Unfortunately, in today's data-driven world, this principle appears all but forgotten. Take for instance the "data lake" architecture that is emerging in many ML ecosystems [**?**, **?**]: user data from the company's multiple products is collected and integrated into a single, giant repository, which archives it for indefinite time periods and makes it accessible to every data engineer and service within the company who might have some use for it. While this architecture fosters innovation, it also flaunts the least privilege principle and exposes data stores to "unintentional, unwanted, or improper uses of privilege" by employees or hackers [**?**].

We believe that the main reason companies are forsaking good protection principles is the lack of appropriate data protection, sharing, and retention abstractions for ML workloads. Traditional protection and sharing abstractions, such as files, directories, database tables, and views, were all designed for "traditional" software, where it was very clear what data was necessary to implement certain functionality. For example, a "traditional" social networking application might consist of several services, each requiring a different subset of user data to do its job: the authentication service needs access to the user account database; the social networking service needs access to the users' friends lists and

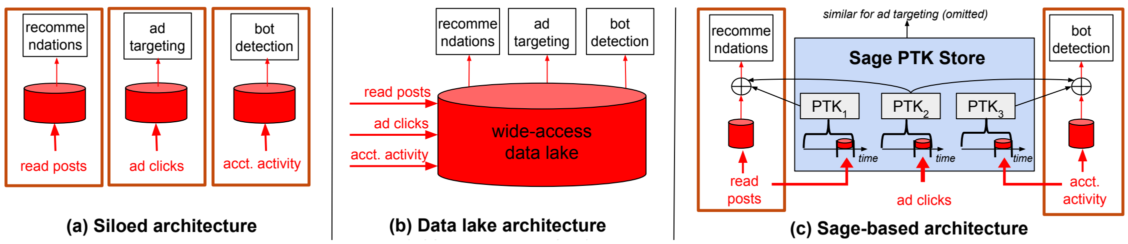


Fig. 1: **Data architectures.** (a) Silo: limited data access. (b) Data lake: wide data access. (c) Sage: limited data access, wide PTK access.

posts, but not to the account database or credit card information; and an in-app product purchase service needs access to credit card information but not to the social graph or account information. To meet the principle of least privilege, the company could store the different types of data in different databases, and enable access to the various services and their engineering teams strictly on a needs basis.

In contrast, in emerging ML-driven applications, the question of whether a particular type of data is "necessary" for some functionality is much more blurry. For example, data collected to improve a post recommendation service – including the posts previously read or liked by each user – appear relevant not only for that service but also for a friend recommendation service, an ad targeting service, and potentially even for a bot detection service. Similarly, the data collected for ad targeting – such as ad clicks or pages visited on the Internet – may be relevant both for the post and friend recommendation services and for the bot detection service. In some cases, the data may not prove useful in the end, but knowing whether it is useful requires its use – and therefore access to its file, table, or stream – for experimentation and potentially even for evaluation in production.

These blurry lines between what data is necessary vs. not for the various services, processes, and teams within a data-driven company create a tension between an interest in enabling wider access to data within the company, and the privacy, ethical, and exposure concerns raised by cross-product data access and use. Different companies resolve the tension differently, each with its caveat. Fig. 1(a),(b) shows some options. Some companies prefer to silo sensitive data streams to the detriment of functionality (a). Others embrace wider-access, data-lake architectures to the detriment of data protection (b). And yet others choose to share some data streams and silo other streams. Regardless, it is unclear what the protection benefits and risks are of siloing some data streams but not others, and what opportunity costs exist for not leveraging data where it is in fact useful.

We believe that this tension can be addressed with new data protection abstractions that are more suitable for emerging ML workloads than traditional files, tables, or data streams. We propose *private transferrable knowledge* (PTK), a new abstraction for protected data sharing and retention in an ML ecosystem. PTKs are *feature models*, trained over historical data and made *differentially private* (DP) to protect the confidentiality of their training data. Two considerations motivate our abstraction. First, we observe that in certain sharing scenarios, access to the raw data is not *required*, or even desirable, and instead services may benefit more from access to well-engineered features extracted from that data. The friend recommendation service may prefer to incorporate user embeddings already trained and curated over the post service's data, rather than the users' raw posting activity, which is high-dimensional and messy. At Twitter, user embeddings are reportedly being shared across teams [**?**], and Uber and Instacart have developed *Feature Stores* to streamline sharing of curated features across teams [**?**, **?**, **?**]. In none of these cases are the features computed rigorously to protect the confidentiality of the training data. The PTK abstraction formalizes the notion of a DP feature model as a unit for protected data sharing in cases where access to the raw data is not needed. Second, with good features, model training and optimization become (comparably) easy, and typically require less data, because the features already capture important historical characteristics. This provides opportunities for reducing the exposure timeframe of raw data in services that *require* access to it, where PTKs can serve as protected units for long-term retention.

To realize this abstraction, we developed the *Sage PTK Store*, a minimal-exposure system that manages, trains, and makes available multiple PTKs on a company's data streams. Fig. 1(c) shows how Sage helps reenact the least privilege principle: it lets the company share PTKs widely across teams to fulfill many sharing needs and only sharing the raw data with services that truly need it.

Sage's design takes after existing feature stores, with notable differences. Like in existing feature stores, the developers in charge of some data stream define feature learning procedures and register them with Sage for continuous computation on that data stream. Sage ingests data from these streams to train these feature models and makes the trained models available to others, who incorporate them into their predictive models. Unlike in existing feature stores, Sage: (1) accepts only feature learning procedures that preserve DP guarantees and (2) strives to minimize the exposure of the raw data it ingests so *it* does not violate the least privilege principle or introduce new risks for the company. For (1), we leverage directly the enormous body of existing work on DP learning algorithms [**?**] to develop an initial PTK library, which we envision will be extended in the future to cover most popular feature learning techniques. (2) raises substantial challenges not addressed in DP ML literature.

Specifically, a tension exists between the amount of data Sage needs to retain, and therefore expose, to train multiple PTKs with acceptable performance, versus the strength of the global privacy semantic it enforces across PTKs. That tension increases at scale, when the number of PTKs grows. We develop two new mechanisms to relieve the tension: (1) the first generic method to estimate rigorous sample complexity bounds for DP models that fit the general empirical risk minimization framework []; and (2) the first resource allocation method for DP models that assigns training samples and privacy budgets so as to both meet a global privacy guarantee across models and minimize the total number of samples needed to achieve acceptable performance for each model. With these methods, Sage becomes the *first minimal-exposure DP ML system* and a model for least-privilege design.

Using **[six]** ML tasks from previous Kaggle competitions, we demonstrate two aspects. First, PTKs can be *useful* xxx *units for data protection*, enabling transfer learning sharing scenarios without the need to share the raw data, and allowing ML tasks to gain the benefits of historical features without the need to retain the raw data long term. Second, Sage can train effective PTKs while enforcing meaningful privacy guarantees. In all scenarios, using DP features instead of unprotected features (akin to the ones in a vanilla feature store) affects the performance of predictive models by *at most* [**1%**]. Thus, PTKs are not only a useful abstraction but also a practical one that can be implemented in a data-rigorous way. xxx

2 Goals and Background

Our goal is to develop a *strong-semantic data sharing and retention abstraction* that will enable a more principled approach to tuning the level of access to user data in ML ecosystems. The abstraction should enable companies that currently over-expose user data – through wide-access policies or long retention periods – to limit that exposure without losing much accuracy in their workloads. It should also enable companies that currently over-constrain access to data – through siloed architectures or short retention periods – to increase accuracy in their workloads without increasing exposure.

Beyond these goals, we formulate three requirements for the abstraction and the system that implements it:

R1: *Suitable abstraction for ML workloads.* The protection abstraction must naturally fit into common patterns in ML ecosystems, and support many data sharing/retention use cases and predictive tasks.

R2: *Limited impact on accuracy and performance.* When used to reduce exposure, the abstraction should result in minimal loss of accuracy, under say 1%, for predictive tasks compared to the best alternative without using it. When used to improve performance, we aim for substantial improvement, above say 5%, to justify computational overheads. In all cases, performance overheads for predictive models should be minimal.

R3: *Minimal-exposure system.* The principle of least privilege should apply to the system that implements this abstraction (Sage). Sage should in fact be a model of rigorous data management and minimize the amount of data it exposes through its internal data structures at all times.

2.1 Threat Model

We consider ecosystems where streams of user data are being collected for use in various predictive ML tasks. We assume that each stream has a set of *primary tasks* which require access to that stream's raw data to function. We additionally assume that there exist other tasks that do not actually need access to the raw data in that stream, but that would benefit from statistics or features learned from the stream in a transfer learning setting (defined in §2.3). We call these *transfer tasks* and assume they have some raw data of their own.

We are concerned with two classes of adversaries who have or gain access to the company's internal state and data stores. First are *abusive employees*, who leverage their privilege within the company to learn about friends' or family members' interactions with the company's products. These adversaries cannot intrude past traditional access controls to escalate their privilege, but can continuously monitor for nefarious purposes any data or state to which they have legitimate access. With the notable exception of the Sage administrators, all employees constitute a risk. Second are *intruders*, who have no legitimate access to the company's internal state, but who manage to break into its compute resources and gain some access. For example, they might break into an employee's laptop and gain access to any state to which the employee has access. Alternatively, they might identify a vulnerability into a running service (such as a buffer overflow or a heartbleed-like vulnerability) and retrieve the service's state from memory or stable storage. Intruders can, at worst, access arbitrary state within one or more compromised services, can appear at any time and without prior notice to the company, and can appear multiple times within the lifetime of the company. However, we assume that intruders are not continuously present within the company, nor will they be able to continuously monitor the company's external predictions to its users. Both classes of attacks are highly relevant, as evidenced by numerous media reports exemplifying each: [**?**], [**?**].

The goal of our abstraction is to enable companies to reduce exposure of their user data streams against both types of attackers. First, we wish to enable the protected sharing of features from a source data stream to its *transfer tasks* without increasing the data's exposure to the potentially abusive employees in charge of these tasks, or to the compromisable services running them. Sharing the raw data constitutes exposure, but so does sharing *any state* computed from the data (including the features) and not protected with a differential privacy or other cryptographic guarantee. Prior research has shown that even the most innocuous-looking aggregate statistic, as well as the parameters of ML models, can reveal a lot about individual examples in their training sets [**?**]. Second, we wish to enable reduction in the amount of raw user data that is exposed at any time to intruders through its *primary tasks*. A good approach to such reduction is to limit the data's retention period in its primary tasks.

2.2 Candidate Approaches

A protection abstraction has two parts: a protection mechanism (e.g., access control lists in traditional systems) and a unit for protection (e.g., files or tables in traditional systems). For ML systems, we base our protection abstraction on *differential privacy* (DP) applied to *feature models*. Other mechanisms and units exist but they are either do not appear as suitable to ML workloads or will likely require combination with our abstraction for sufficient protection under our threat model.

As alternative mechanisms, consider leveraging *cryptographic mechanisms*, such as homomorphic encryption (HE) or secure multi-party computation (MPC), applied to traditional protection units, such as files or streams. Imagine an encrypted data lake architecture, where the raw data is shared widely but in homomorphic-encrypted form; teams and services design and train models "blindly" using homomorphic training algorithms [**?**, **?**]. Alternatively, imagine a siloed architecture, where the raw data is only accessible to the team that collects it; predictive models are trained (again, blindly) on one or more datasets using MPC. Aside from the significant performance concerns of running even state-of-the-art HE and MPC mechanisms, a key limitation of using encryption alone as the protection mechanism in ML systems is that it does not provide actually adequate protection: the trained models, whose decryption is usually assumed for prediction, can leak information about the training data [?, ?]. The problem can be fixed either by an additional MPC protocol with the end users [?] (which adds to the performance concern) or by combining HE/MPC with DP. Thus, DP is less of an option and more a required mechanism in ML systems.

As an alternative unit of protection, consider applying DP at the level of *SQL queries*. For example, imagine a company requiring that all access to the raw data be done through a DP SQL interface, such as PINQ [?] or FLEX [?]. For workloads easily written on SQL, such as simple aggregates or learning algorithms that can run efficiently on statistical queries [?], this is a fine approach. However, for broad classes of learning algorithms, including those based on stochastic gradient descent, implementing them on SQL is either **[impossible?]** or inefficient. More generally, we believe that for ML workloads, the SQL interface is a too low-level abstraction at which to enforce protection, and that is why we elevate the level of abstraction to *machine learning models*, and more specifically, to *feature models*. xxx

2.3 Background

Differential Privacy. DP provides a well-defined semantic for preventing leakage of individual records in a dataset through the output of a computation over that dataset. It works by adding randomness into the computation so that details of individual records are "hidden" by the randomness. A (randomized) algorithm A that takes as input a dataset D and outputs a value in a space B is said to satisfy *ϵ-DP at record level* if, for any datasets D and D' differing in at most one record, and for any subset of possible outputs $S \subseteq B$, we have: $P(A(d) \in S) \leq e^{\epsilon} P(A(d') \in S)$. Here, $\epsilon > 0$, called the *privacy budget*, is a parameter that quantifies the strength of the privacy guarantee (lower is better). $\epsilon \leq 1$ is generally considered good protection.

Three important properties of DP are: (1) *Post-processing resilience*: any computation applied on the output of an ϵ-DP algorithm remains ϵ-DP [?]. (2) *Composability theorem*: if the same dataset is used as input for two DP algorithms, ϵ_1-DP A_1 and ϵ_2-DP A_2, then the combined privacy semantic is $\epsilon_1 + \epsilon_2$-DP [6]. (3) *Resistance to auxiliary information*: regardless of external knowledge, an adversary with access to the outputs of a DP algorithm draws the same conclusions about the presence/absence of a record in the input dataset [?]. A corollary of (3) is that if two DP algorithms, ϵ_1-DP A_1 and ϵ_2-DP A_2, use as input two non-overlapping datasets (e.g., two subsets of iid samples from a bigger set), then the combined privacy semantic is $max(\epsilon_1, \epsilon_2)$-DP.

Transfer Learning. [Background here on transfer learning and when it's supposed to work and why.] xxx

3 The PTK Abstraction

We propose *private transferrable knowledge* (PTK), a new protection abstraction specifically designed for ML systems. Its unit of protection is the *feature model*; its protection mechanism is *differential privacy* (DP). We believe that this abstraction, together with the system that implements it, meets the three requirements we defined in §2. First, feature models are already being used as units for data sharing in today's ML ecosystems []. This encourages us to believe that they will be also serve as natural and suitable units for protected data sharing in these systems (requirement

```
struct ptk: id, model_state, config, performance_target,
            oldest_window, newest_window, parent, child

// PTK developer API:
ptk.train_dp(trainset_iter, ε_train, previous_model_state,
    previous_config) → overwrite
ptk.eval(testset_iter) → PTK-specific evaluation metric
ptk.profile_dp(profileset_iter, ε_profile, τ) → map{ε_i->n_i}

// PTK user function (not part of API):
featurize(targetset_iter, ptk[]) → featurizedset_iter

// Sage API:
register_ptk(ptk_train_dp_fn, ptk_eval_fn,
    ptk_profile_dp_fn, train_frequency) → ptk_id/reject
deregister_ptk(ptk_id)
update_ptk_config(ptk_id, new_config)
update_ptk_performance_target(ptk_id, new_target)
subscribe_to_ptk(ptk_id, rpc_endpoint)
notify_new_ptk_state(ptk_id)
```

Fig. 2: **PTK API.**

R1). Second, the literature is rife with DP implementations of most popular ML algorithms [**?**], and there is increasing experimental evidence that DP is amenable to them [**?**, **?**]. This encourages us to believe that DP can be applied to feature models with limited accuracy overheads (requirement **R2**). Finally, §4 discusses how PTKs can be implemented and managed effectively and at scale while minimizing exposure of the data used to train them (requirement **R3**).

The PTK abstraction can be used to reenact the principle of least privilege in ML systems. Imagine an ML ecosystem where access to raw data is given sparingly to primary tasks that truly need it, and where access is retained by these tasks for bounded time periods. PTKs are used to (1) improve performance in transfer tasks that can leverage their features without requiring access to the raw data (*unit for data sharing*), and (2) improve performance in the primary tasks by letting them use knowledge from historical data without retaining long-term access to those streams (*unit for long-term retention*).

3.1 API

Fig. 2 shows the PTK API, which is used by two entities: the *PTK developer* and one or more *PTK users*.

PTK Developer. The developer starts by defining a procedure to learn useful features from a data stream to which she has access. She turns the feature learning procedure into a PTK by implementing the PTK developer API (described shortly). The developer then identifies a reasonable performance target for the PTK, which Sage will try to honor as it

	PTK	S/U	`ptk_train_dp`	`ptk_eval`	`ptk_profile_dp`
1	**Count featurization**	S	DP contingency tables [**?**]	error from true value	ERM profiling
2	**Tree featurization**	S	XXX [**?**]	loss on a holdout set	ERM profiling
3	**User embeddings**	S	DP Poisson factorization []	loss on holdout set	ERM profiling
4	**Marginal regression**	S	XXX [**?**]	loss on a holdout set	ERM profiling
5	**Covariance matrix**	U	XXX []	error from true value	XXX
6	**Historical statistics**	U	keyed aggregates w/ Laplace noise []	error from true value	XXX
7	**Clustering**	U	DP Lloyd algorithm []	XX	experimental

Tab. 1: **Implemented PTKs.** S/U: supervised/unsupervised. Core methods used to implement the PTK API.

assigns privacy budgets and data samples to train the PTKs on overlapping streams. splits the global privacy parameter $\mathcal{E}$ and the data it has available to train multiple PTKs on overlapping streams. Finally, she registers the PTK with Sage and gives Sage access to the source raw data stream on which to train it. Sage assigns a unique ID to the PTK.

PTK User. The user of a PTK can be either the same entity that developed it (primary task) or a separate entity that will use the PTK in a transfer learning setting (transfer task). To use a previously registered PTK, the user subscribes to it by calling `subscribe_to_ptk`, specifying the PTK's ID and an RPC endpoint. Sage trains the PTK periodically, stores them in the widely accessible PTK lake, and notifies the user of new model states available for the PTK. The user retrieves those states from the PTK lake and them incorporates into her predictive models. Our abstraction imposes no API for that, but usually, the PTK user will implement a `featurize` function that transforms her dataset based on one or more PTKs.

PTK Developer API. `ptk.train_dp` is a DP version of the developer's training procedure. It takes a *training set* and a privacy budget, ϵ_{train}, and must satisfy ϵ_{train}-DP. We train PTKs periodically on $\mathcal{T}$-sized time windows (where $\mathcal{T}$ is the exposure window for the PTK's input data stream). We support continuous PTK training across windows by supplying `train_dp` with the previous state of the model trained from the previous $\mathcal{T}$-sized window. `ptk.eval` is a regular evaluation procedure for the PTK that takes in a testing set and returns some PTK-specific evaluation metric, such as a loss metric for a supervised PTK. The `eval` function need not be DP. `ptk.profile_dp` profiles the amount of training data the PTK needs to likely reach a performance target, τ, as a function of the privacy budget it is assigned. The profiling function takes a held-out *profiling set*, separate from the training set, plus a privacy budget, $\epsilon_{profile}$, and the performance target, τ. It must satisfy $\epsilon_{profile}$-DP.

3.2 Example PTKs

We are implementing *ptklib*, a library of PTKs for popular feature learning algorithms. We use a recent feature engineering textbook [**?**] to prioritize algorithms for implementation. Tab. 1 shows the PTKs implemented so far, the type of algorithm (supervised/unsupervised), whether we train it on a per-window basis or continuously, and the methods we used to implement the developer API for each PTK. For most PTKs, we used known DP versions of training algorithms for `ptk.train_dp`, sensible performance metrics for `ptk.eval`, and a generic profiling method we invented (§4.2) for `ptk.profile_dp`. We describe three PTKs.

Count Featurization PTK. Count featurization is a popular technique for featurizing high cardinality categorical variables, such as user identifiers and IPs, when training ML classification models [**?**, **?**]. The technique replaces each value of the feature vector with the number of times that feature value has been observed with each label and the conditional probability of each label given that feature value. This leads to dramatic dimensionality reduction over standard one-hot encoding, which (intuitively) should allow for more efficient learning and require fewer samples for training. In our system, this means that the PTK user, by featurizing his data with counts and conditional probabilities trained on historical data streams, may be able to learn related labels using less data. §3.3 shows how one can leverage this PTK to reduce raw data retention in primary workloads.

To implement the count featurization PTK on top of a data stream, we compute and publish periodically DP contingency tables of each feature separately with all features that a PTK developer specifies as interesting labels, either for his/her workload or potentially for others in the company. The tables are made DP by initializing their cells with random draws from a Laplace distribution [**?**].

Tree Featurization PTK. The tree featurization PTK leverages a forest of randomized decision trees to learn a set of non-linear features that can be used to increase the performance of predictive models. Decision trees are a set of nested if-else statements where each root to leaf path in a tree corresponds to a non-linear predicate of the input. The tree featurization PTK can featurize data in a number of different fashions. Each leaf node in the tree can be used as a categorical feature so a forest of n decision trees will generate n new categorical features that can be processed with count featurization, hash featurization [7], or count featurization. Users can also use the predictions returned by each tree or by the forest as a whole as features for an application level model. Decision trees have previously been used at Facebook to improve the performance of ad click prediction [4].

[TODO(riley): Describe DP Tree implementation.] xxx

User Embeddings PTK. [TODO(mathias).] xxx

3.3 USAGE

We illustrate the usefulness of PTKs as units for protected sharing and long-term retention using corporate-inspired scenarios that we instantiate and evaluate on public datasets. §5 details the datasets, methodology, and results. We show highlights here to explain usage.

USAGE 1: PTK as Unit of Protected Sharing. *Scenario:* A transportation company is already operating a large fleet of taxis in a city. A different division of that company would like to launch a bike sharing system in the same city. The bike share division would like to use the ride information from the taxi operations to help bootstrap forecasting bike demand in different areas of the city, because they expect that the taxi and bike share data will be used for similar trips. They will use the forecasting to provision bike stations throughout the day. They wish to share PTKs instead of raw data.

We instantiate this scenario on two public datasets: NYC Taxi dataset (33 million taxi rides over three months) and NYC Citibike dataset (3 million bike rides over three months). We use these as the datasets available to the taxi and bike divisions in our scenario, resp. The Citibike task (target task) predicts the destination bike station of each ride out of 465 possible stations.

PTK Usage: The two datasets have several common features: ride origin, destination, date, time of day, and weather information. The Taxi dataset is amenable to tree modeling, so we define a tree PTK on the dataset that uses the common features (except destination), to predict the closest destination bike station for each taxi ride. **[Update.]** We use the tree PTK to generate 465 new features for each CitiBike ride that correspond to the predicted probability of each the 465 destination stations. We use the featurized CitiBike data to train a logistic regression. We find that using the taxi PTKs, the bike task gains significant performance (6%) compared to not using the PTKs (§5.2). Thus, in this scenario, the PTK is a useful unit for sharing and obviates the need to share – and therefore expose – the raw data with the target task. xxx

USAGE 2: PTK as Unit of Long-term Retention. *Scenario:* An online advertising company collects ad clicking activity from its users. The primary task for this data stream predicts the likelihood that an ad will be clicked by a user if shown on a particular page. The team responsible for this task worries about the potential of its members to get hacked. They wish to minimize the amount of ad click data they retain at any time in anticipation of attack. They observe that with good features trained over long periods of time, their models require much less data to converge. They want to leverage PTKs to capture good historical features and safely retain them for long periods of time, while reducing the amount of raw data used to train their predictive models at any time.

We instantiate this scenario on the Criteo dataset [**?**] (39 million ad impressions over seven days). Each entry has 26 categorical features, 13 integer features, and a binary label indicating if the ad impression resulted in a click. We use the same task as a Kaggle competition organized around this dataset: predicting whether a user will click on an ad. We adopt the winning predictive model: a neural network binary classifier.

PTK Usage: The Criteo dataset contains a number of high-dimensional categorical features; 9 of the 26 categorical features have 10K or more observed values. We therefore hypothesize that count featurization [8], which reduces exposure, will help improve the rate of learning for this predictive task and therefore will require models less raw data to converge. We register a count featurization PTK, and train it over 7 windows for the first 80% of the dataset reserving the final 20% of the dataset for testing. We sum the counts across windows for featurization. We train the classifier using the count features in addition to all of the raw features. We find that count features with $\epsilon = 1.0$ help improve perform over the a model trained using only the raw features by 1.8%. In addition to simply improving model performance we can leverage count features as a unit of data retention. Removing raw features entirely results in a model that converges to with XX% of the model trained on the entire raw dataset while training on only the last XX% of the data§5.2. **[TODO(riley): fill these in and show that PTK features can be useful for both retention and** xxx

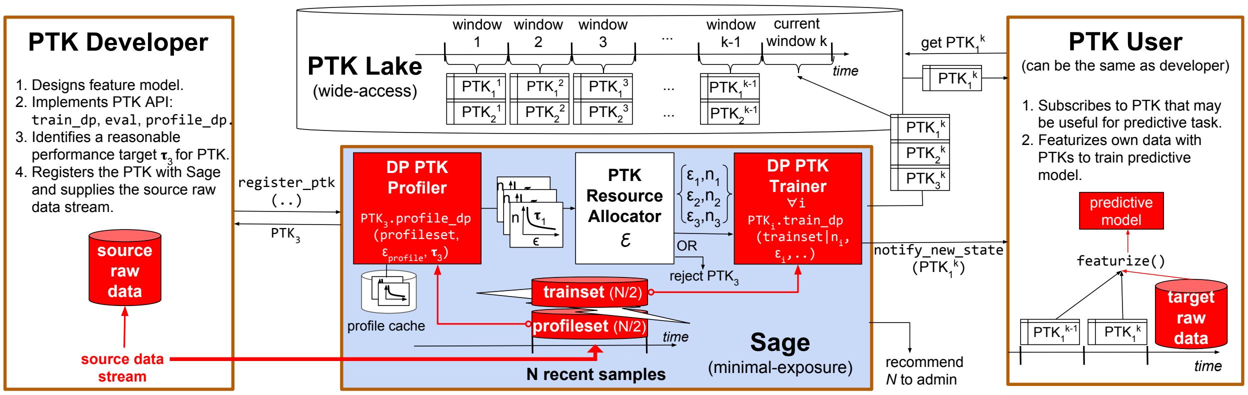


Fig. 3: **Sage Architecture.** PTK_i: PTK object; PTK_i^j: the state of PTK_i at time window j; $trainset|_{n_i}$: n_i iid samples from trainset.

model improvement.] Thus, in this scenario, the PTK is a useful unit for long-term retention and obviates the need for the primary task to retain access to – and therefore expose – the raw data long-term.

4 The Sage Design

To realize the PTK abstraction, we developed the *Sage PTK store*, a minimal-exposure system that creates, trains, and shares PTKs on a company's data streams. The unique aspect of its design is its effort to minimize the size and timespan of the user data exposed through its internal state at all times (requirement **R3** in §2).

4.1 Architecture

Fig. 3 shows Sage's architecture. It consists of two top-level components: *the Sage service*, which ingests several of the company's user streams and trains PTKs on them periodically; a *PTK lake*, a regular database that stores the trained PTKs (including the models' parameters and configurations) and makes them widely accessible to any engineer, team, or service who might have some use for them. To train the PTKs, Sage maintains a recent window of data from each stream, called the *training window*. Different PTKs are computed at different intervals of time, but for simplicity we will assume here that they are trained once per training window.

Protection Semantics. To protect user data against adversaries listed in §2.1, Sage enforces two protection semantics. First, because PTKs are widely accessible in the data lake, they can be monitored by abusive employees or hackers with access to the PTK lake. Each PTK enforces ϵ_{train}-DP, so for small ϵ_{train} the leakage is small from individual PTKs, however across PTKs in the same training window the leakage is additive. To bound the leakage across PTKs, Sage enforces across all PTKs a *global privacy semantic*, $\mathcal{E}$-DP, for $\mathcal{E}$ configured by the Sage administrator (default $\mathcal{E} = 1$).

Second, because the Sage service ingests multiple streams of user information, it becomes a point of vulnerability for the company. While the Sage administrator is trusted to not abuse this information, the service itself can be compromised by intruders. To limit the exposure of user information through its internal state, Sage minimizes at all times the timespan of user data that is leaked through its internal state. We call this the *minimal exposure semantic* and is parameterized by $\mathcal{T}$, the timespan of Sage's training window. The semantic guarantees that if an intrusion occurs at time T_1 and ends at time T_2 (both times unknown to the company), then the intruder will be able to glean information about raw user data streamed into the Sage service in the interval $[T_1 - \mathcal{T}, T_2]$, but any information from before or after that interval will be protected with a global $\mathcal{E}$-DP guarantee.

Tension. Unfortunately, enforcing the two semantics while preserving acceptable PTK performance is challenging, particularly as the number of PTKs grows. To satisfy an $\mathcal{E}$-DP global privacy semantic, there are two approaches. One approach is to split the global privacy parameter $\mathcal{E}$ equally across the p PTKs, assigning PTK_i a privacy budget of $\mathcal{E}/p$ and the full training window of recent data. This means calling PTK_i`.train_dp(`$trainset|_{\mathcal{N}}$`, `$\mathcal{E}/p$`, ...)` for each $i = 1..p$. This works if the PTKs can tolerate small privacy budgets, which is not always true. For example, while our count featurization PTK trains well even for privacy budgets ≥ 0.001, our tree and user embedding PTKs become worthless at privacy budgets < 1. Regardless, this is not a scalable option for large p.

An alternative is to assign all PTKs full privacy budget $\mathcal{E}$ and split the training window. This means calling PTK_i`.train_dp(`$trainset|\mathcal{N}/p$`, `$\mathcal{E}$`, ...)` for each $i = 1..p$, where $trainset|_{\mathcal{N}/p}$ takes $\mathcal{N}/p$ iid samples from the available training set. This lets PTKs achieve their best performance under the given $\mathcal{E}$ global privacy semantic, but will pressure Sage to increase the size of the training window to train acceptable PTKs, thereby hurting the minimal-exposure semantic.

Approach. Our approach is to treat the challenge as a *resource allocation problem*, a common problem in systems, and to leverage the canonical systems approach to address it. The resources here are the privacy budget ($\mathcal{E}$) and the size of the Sage training window ($\mathcal{N}$). The former is a fixed, unscalable resource (cannot be increased as the workload increases); the latter is more scalable, although for minimal-exposure semantics, we wish to minimize $\mathcal{N}$. We need to allocate these resources across p PTKs by assigning each PTK_i privacy budget ϵ_i and a training set consisting of n_i iid samples from the available window of recent data, and invoking PTK_i`.train_dp(`$trainset|_{n_i}$`, `ϵ_i`, ...)` to train it. Our approach, taken from systems, first *profiles* each PTK in isolation, as it is registered with Sage, in terms of how its performance is impacted by various values of ϵ_i, n_i; it then finds an allocation $(\epsilon_i, n_i)\forall i$ that preserves the global privacy guarantee ($\mathcal{E}$) and minimizes the total amount of data needed for training ($\mathcal{N}$) while meeting developer-specified performance targets for the PTKs.

Fig. 3 reflects Sage's key architectural components that implement this procedure: the ***PTK Profiler***, ***PTK Resource Allocator***, and ***PTK Trainer***. At every window k, we partition the $\mathcal{T}$-window of recent data in two: a *profileset* and a *trainset*. The ***PTK Profiler*** uses the profileset to profile a small set of new or not recently profiled PTKs (PTK_3 in the figure). The profiles, which are made DP, are saved in a profile cache for future reuse. The ***PTK Resource Allocator*** uses the profiles of all PTKs registered in the system to determine an allocation $(ei, ni)\forall i$ that satisfies the conditions in the preceding paragraph. It is possible that the allocator cannot find such an allocation, in which case it will REJECT the new PTK (PTK_3) or otherwise announce the Sage administrator that a sound allocation does not exist under the current PTK mix, and revert to its previous allocation plan. The ***PTK Trainer*** uses the allocation to train the PTKs on the *trainset*. Finally, Sage automatically determines a proper value for $\mathcal{N}$, to support both PTK profiling and training. It recommends the value to the administrator, who can decide to update the $\mathcal{T}$ parameter of the minimal-exposure semantic.

4.2 PTK Profiler

The PTK Profiler is responsible for profiling one PTK (or a small number of PTKs), on a heldout dataset (profileset), in isolation from one another. Its goal is to estimate the number of samples needed by the PTK to reach a particular performance target as a function of the privacy resource being allocated for training.

Theoretical Formulation.

Non-DP Profiling Algorithm.

DP Profiling.

Analysis.

4.3 PTK Resource Allocator

[TODO(roxana): will write next.] xxx

Describe the problem briefly and in intuitive terms. Cast the problem as a resource allocation problem, but with a very non-scalable resource, privacy budget, and a more scalable resource, data.

Problem Formulation. Formalize the problem statement.

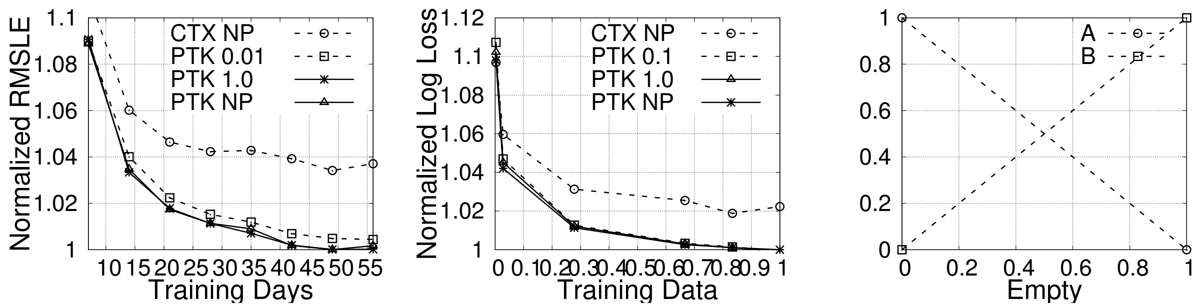
Solution. Say how we solve it: we throw it in an LP solver. Say what we do with various answers from it.

5 Evaluation

5.1 Methodology

Datasets. We evaluate Sage on five publicly available datasets plus a copy of the Netflix 2005 dataset. Most of the datasets have been used as part of Kaggle competitions; the tasks we define on these datasets for evaluation (described in the next header) are inspired by the tasks formulated in these competitions. We describe the datasets here and give statistics about them in Table **??**.

- *Criteo [**?**]:* Ad click/noclick log from XXX users of Criteo ad company over XXX period. Consists of XXX total examples, each with 39 anonymized features, including anonymized IDs for user, the page the user is viewing, the ad shown, and whether the user has clicked on that ad. A subsampled version of the dataset was the subject of a Kaggle competition to predict click/noclick given contextual features (a binary classification task) [**?**]. We use the model that won this competition as baseline: XX characterize the model.
- *Movielens [3]:* Movielens is an academic dataset by the University of Minnesota and consists of 22M ratings for 34K movies in the range [1,5] from 240K users. In addition to movie ratings the dataset also contains the genre of each movie. We use the MovieLens dataset in two use cases. First for using a demonstration of using PTKs as a unit of as a unit of retention using count features and second as a demonstration of protected sharing by



(a) **Taxi duration prediction** (b) **Criteo click prediction** (c) **MovieLens classification**

Fig. 4: **PTKs as units of retention.** XXX PLACEHOLDER IMAGE XXX

using MovieLens as a target task. As baselines we logistic regression and a state of the art matrix factorization algorithm both of which are included in the VowpalWabbit library [5].

- *Netflix:* We use a subset of the Netflix challenge dataset that overlaps by movie with the MovieLens dataset to demonstrate protected data sharing. This subset contains 88M ratings in the range [1,5] by 470K users of 8K movies. We compute a differentially private item-item covariance matrix which is used to bootstrap models trained on the MovieLens dataset.
- *Taxi [2]:* NYC Yellow Taxi rides by XXX people over XXX period. Consists of XXX total examples, each with XXX features, including ride origin, destination, date, time of day, weather information that day, etc. A one-month version of this dataset was the subject of a Kaggle competition to predict the ride's tip and destination given the other contextual features [**?**]. We use the same task and the winning model as them: a XXX characterize model.
- *Bikes [1]:* NYC Citibike is a bike share system that allows users to rent bicyles for point to point trips between 2 of the many CitiBike docking stations. NYC Citibike publicly releases anonymized data ride level data for research use. We leverage a three month subset of the dataset from 2015 that consists of 3362370 rides. We use 13 contextual features about including 6 weather features to predict which of the 465 destination where the ride will end.
- *Douban [**?**]:* Three datasets with ratings of movies, books, songs, respectively, by 36673 users of the Douban Chinese media company, over between 2005 and 2011. Consists of XXX, XXX, and XXX, resp. examples. Each example has XXX features, including userID, movie/book/songID, genre. XXX Kaggle and model description.

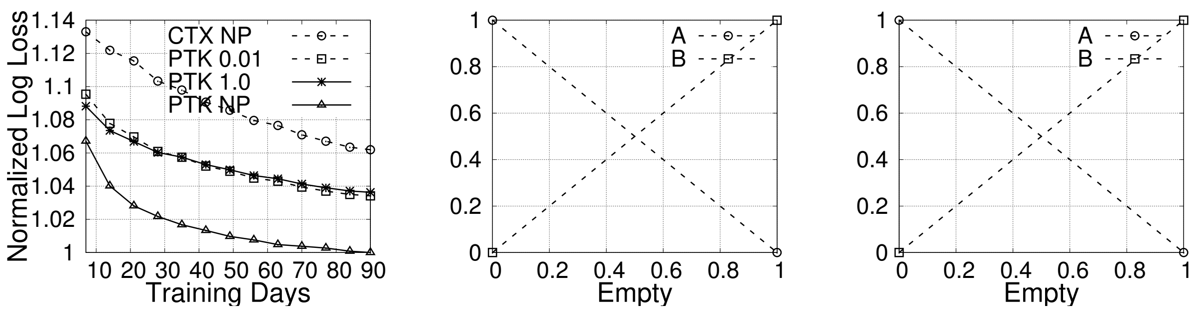
Methodology. XXX Describe methodology briefly.

5.2 PTK Usefulness

USAGE 2: PTKs as Protected Unit of Long-term Retention. Figure 4 will have a model for taxi 4(a), criteo 4(b), and MovieLnes 4(c). We may want to consider replacing Movielens with another of the taxi models or something because it may be confusing to have the Movielens classification because we mostly used that before so that it would be useful with count featurization.

Task	Description	Problem Type	Q	PTKs Used	Predictive Model	Model parameters
T1	Predict ad click on Criteo	binary classif.	Q1	count featurization. 39 features	neural net.	VW. One 35 nodes hidden layer with tanh activation. LR: 0.15. BP: 25. Passes: 20. Early Terminate: 1.
T2	Predict movie rating on Movielens	regression	Q1	XXX	singular value decomposition (svd)	VW. Rank 10. L2 penalty: 0.001. LR: 0.015. BP: 18. Passes: 20. LR Decay: 0.97. PowerT: 0.
T3	Predict movie rating on Netflix with PTKs from Movielens	regression	Q2	1 covariance matrix	ridge regression	TODO
T4	Predict ride duration on Taxi	regression	Q1	9 historical statistics, 2 clustering, 2 PCA PTKs	gradient boosting	XGBoost. LR: 0.05. columns sample: 0.5. min child weight: 75.0. reg. lambda 3.0. 500 estimators. early stopping: 30
T6	Predict ride price on Taxi	regression	Q1	9 historical statistics, 2 clustering, 2 PCA PTKs	gradient boosting	XGBoost. LR: 0.05. columns sample: 0.5. min child weight: 75.0. reg. lambda 3.0. 500 estimators. early stopping: 30
T7	Predict ride destination on Bikes with PTKs from Taxi	465-class classif.	Q2	1 tree PTK	logistic regression	VW. BP: 26. LR: 0.0742. PowerT: 0. Passes: 5.
T8	Predict movie rating on Douban-Movies with PTKs from Douban-Music	classif. 1..5	Q2	1 user embedding PTK	XXX	TODO
T9	Optimize PTKs for Taxi tasks; use them for Taxi tasks	XXX	Q3(a)	classif.	XXX	TODO
T10	Optimize PTKs for Taxi tasks; use them for Bikes task	XXX classif.	Q3(b)	XXX	XXX	TODO

Tab. 2: **Evaluation tasks.** XXX



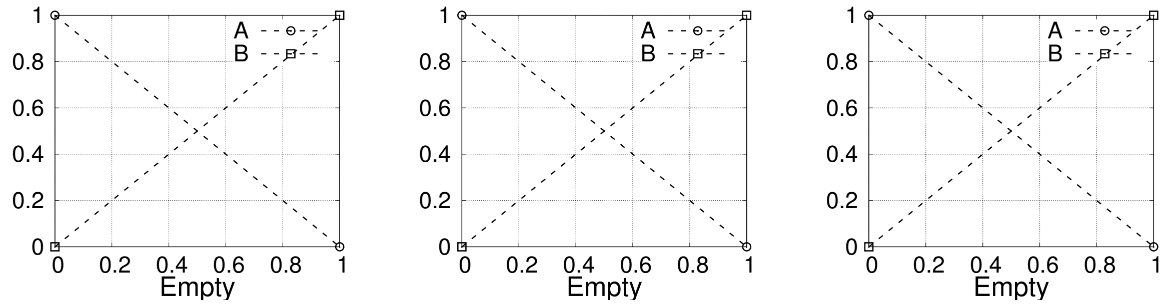
(a) **Taxi to CitiBike** (b) **Netflix To MovieLens** (c) **Douban to Douban**

Fig. 5: **PTKs as units of protected sharing.** XXX PLACEHOLDER IMAGE XXX

X axis will be amount of training data. Y axis will be the normalized loss normalized to the non private context only model. This is different than what we had before but will match up better with the next section. 4 lines per figure: context only, non private ptk model, private ptk model $\epsilon = 1$, private ptk model $\epsilon = 0.1$.

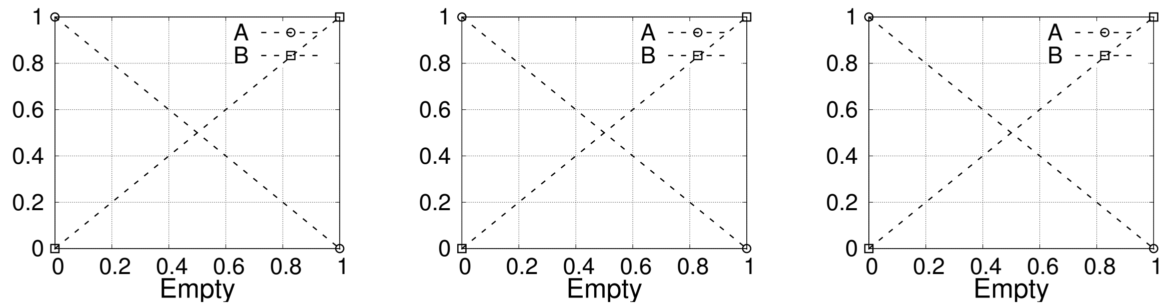
We have two conclusions that we may be able to draw about PTK's usefulness for retention. First in a scenario like criteo we can leverage ptks to reduce the amount of data required to achieve a decent model (the pyramid conclusion). This is true even if the model is not improved by training the raw+ptk features on the entire dataset. Second in a scenario like the taxi dataset we can say something about PTKs being useful to just improve performance. We may also be able to comment about the dataset being very time dependent and PTKs making data useful if if theadding more data to the model does not improve performance. We can also point out that the performance is improved when adding differential privacy and that we have few PTKs here.

USAGE 1: PTKs as Unit of Protected Sharing.



(a) **Tree Profile** (b) **Clustering Profile** (c) **Count Featurization Profile**

Fig. 6: **PTK Profiles.** XXX PLACEHOLDER IMAGE XXX This will have privacy budgets on the X axis and the amount of data required to achieve τ_1, τ_2, and τ_3 on the y axis.



(a) **Allocation Model Performance** (b) **Allocation Solving Time** (c) **PTK Prediction Latency**

Fig. 7: **PTK Performance.** XXX PLACEHOLDER IMAGE XXX

Figure 5 will have a model for taxi to citibike 5(a), netflix to movielens 5(b), and MovieLnes 4(c). This will be the MovieLens as a regression problem.

X axis will be amount of training data. Y axis is the loss normalized to the loss of the model trained on only the target dataset. The baseline model may vary and be a different type of model than what we train using the transferred PTKs. Baselines: CitiBike - VW Logistic regression. This will likely be the same with and without PTKs. MovieLens - VW SVD - The PTK models will be either KNN or ridge regression Douban - VW SVD - not sure what the PTK models will be

For PTKs to be useful as a unit of datasharing the target model must be improved by leveraging the source PTK. It can be improved in 1 of two ways by improving the performance when trained on the entire dataset or by helping the model converge faster which will assist with bootstrapping. We can also comment about DP without contention.

5.3 PTK Resource Allocation

Budget/Data Allocation under Contention. Figure 7(a) will show the performance of the Taxi duration prediction model with 3 lines under the three different allocation models: evenly divide budget, evenly divide data, and the profiling based allocation. The Y axis will be model performance and the x axis will either be the total number of PTKs or the number of PTKs in addition to those used by the model. We can x values of 50, 100, 1000, 10000. I don't think that we should talk about TFX because we never mentioned it before and it'd be confusing to introduce it here.

We exepct to see the performance degrade as we increase the number of PTKs but the profile based allocation should degrade less badly.

PTK Profiles. Figure 6 will show the performance profile of the tree PTK **??** and the clustering PTK **??** on the taxi dataset, and the count featurization ptk **??** on the criteo dataset. It may be better to include three from the same dataset but we can decide later exactly which figures we include here based on which are interesting. **[Instead of clustering, it would be better to use PCA or something for which we have a profiling approach. —**RG] xxx

The X axis will be privacy budget and the Y axis will be the amount of data required to meet the performance goal. There will be three performance goals: τ_1, τ_2, and τ_3. There will be two lines for each performance goal. The predicted amount of data to reach τ and the actual amount.

What the profiles look like for various PTKs and at various τ values. Compare to the experimental version of determining the profile: i.e., determine experimentally (e.g., using a binary search algorithm) how much data you need to reach τ.

Amounts of Data Used by Allocation. Plot or report the amount of data required to allocate all of the PTKs using our allocation procedure (sum over all partitions). You can place these numbers on top of the corresponding bars. In this little paragraph, refer to these numbers and comment that they are "reasonable."

5.4 Performance

Figure 7(c) show the prediction latency of two models Taxi duration prediction and citibike in the context only and the ptk configurations. The X axis will be percentile and the Y axis will show the latency. We expect that both of the context only models will lowest tail latency. I expect both of the PTK models will have a much higher tail latency. The CitiBike model because random forests often have a long tail latency and the taxi models because it will require two rounds of featurization. Both will still probably be dominated by network time.

Figure 7(b) will show the time required to solve the budget allocation problem after we have generated the profiles for each of the PTKs. The X axis will be the number of PTKs and the Y axis will be the time required to solve the MILP problem. We expect the time to polynomialy in the number of PTKs for which we're solving. Since this is under something like a day it should be fine since this will be run infrequently.

6 Analysis

[NOTES ONLY.] xxx

Need a security analysis here to clarify the threats that PTKs leave open and re-enforce why we believe we still add benefits.

Return to Multics principles and see how we redeem not only least privilege but also others.

Threats:

1. Ecosystem-wide threats: - A malicious employee will be able to access every data stream, state and service to which he is granted access. They can monitor the raw data, the internal state of servers, and the service's predictions to the users. However, if these resources are correctly siloed from employees who do not need access to them, then he won't be able to access them. The PTK abstraction aims to constrain the reach of any individual employee to the minimum they need, and assumes that access controls and firewalls correctly implement minimal access policies.

- An intruder into the ML ecosystem can, in the limit, retrieve arbitrary state in all of the predictive tasks/teams. This includes the state that Sage maintains, plus The PTK abstraction aims to constrain the reach of an intruder by allowing individual primary tasks to reduce their access retention periods by leveraging historical PTKs. Hence, the intruder will be able to retrieve data in the maximum retention period for all primary tasks for a particular data stream.

- Many attacks against integrity/availability: - Malicious employee registers a lot of PTKs with high eps. Won't get Sage to increase privacy budget beyond $\mathcal{E}$, so confidentiality wise we are OK. But it may get Sage to increase $\mathcal{T}$, at least to some point. We envision that the Sage admin will bound the max T they wish to enforce. The attacker may also cause extreme contention and hence new PTKs may be rejected by Sage. - Train bogus PTK, change a good PTK into worse. That sabotages the workloads that rely on the PTK.

2. Sage threats: Fewer, because Sage enforces a strict semantic. But there are a few caveats. - An intruder who breaks into the Sage service may change $\mathcal{E}$ or the bound on T. This will get Sage to weaken its guarantees. We assume that such changes are not possible without the (trusted) administrator's knowledge.

- When PTKs are registered, they are assumed to already have been designed. That design is usually done through a lot of experimentation and optimization on a dataset. The PTK's training procedure, can (in theory) reveal information about the dataset the developer used to design the PTK. We believe this is reasonable, but we invite care into how that is done and awareness XXX.

- A bad (or malicious) programmer may break the assumptions we make about the PTK functions, e.g., that `ptk.train_dp` and `ptk.profile_dp` satisfy the specified DP semantic for all of their outputs and side-effects. This can lead to leakage of user data the through the PTK. A natural solution would be to require that PTK functions must be certified through code review to be accepted by Sage.

The difference between Sage and the rest of the ecosystem is that Sage*enforces* the minimal-exposure policy while the rest of the ecosystem is merely encouraged to do so by leveraging Sage's abstractions. However, we acknowledge that enforcing minimal-exposure is non-trivial even with the PTK abstractions: when data is expired, all of its traces must be securely erased, all non-DP models based on them must be retrained without it, and worst of all; to determine, one needs to profile his model's dependency on the amount of data, akin to how we do it, and tune that with changes in the data.

Sage provides a few of these services, and could in theory be used to secure predictive models, too. They could develop DP versions of their predictive models and register them with Sage for periodic profiling, training, and releasing. This is useful to gain DP under continual output observation for the predictive tasks/teams. To get pan privacy for these tasks/teams, that's much harder, because it most likely means that once they register the predictive PTK, the engineers should no longer have access to the data. This means they won't be able to design and bootstrap new, improved versions of the predictive models, which is an important part of engineers' duties in an ML ecosystem. Designing an ecosystem-wide, $\mathcal{T}, \mathcal{E}$-pan private system is very much an open problem.

7 Related Work

Alternative Protection Models. §**??** discusses some alternative protection models for ML systems. Additional approaches researched in the literature follow. Local privacy is a strong data protection model based on differential privacy or randomized response, where the data is randomized *before* collection. The model is being used in practice [**?**] but only supports **[simple aggregate statistics]** and *not* general ML [**?**]. Federated learning XXX. xxx

Relationship with Standard DP Semantics. Sage's threat model and semantics relate closely to standard models based on differential privacy. Sage's global privacy semantic satisfies $\mathcal{E}$-DP under continual observation, a DP semantic defined for streaming computations, which requires the $\mathcal{E}$-DP guarantee to be met by all the externally visible outputs or side effects, combined []. Sage's minimal-exposure semantic relates to pan privacy, a DP semantic that seeks to protect a (streaming) DP computation from intruders by requiring that the $\mathcal{E}$-DP guarantee to be met not only by the outputs but also by the internal states of a DP computation, all combined. Sage's semantic, which permits bounded leakage, is strictly weaker than pan privacy, which permits no leakage. Pan privacy is known to reduce to local privacy under threat models that permit repeated unannounced intrusions []. By allowing some leakage, Sage breaks this reduction and is able to train general ML models effectively; by bounding the timespan of the leakage, it provides a meaningful protection semantic to its system administrators.

Relationship with DP ML Literature. Mostly algorithm-level, less ecosystem-level DP systems, and no minimal-exposure system designs.

A relevant direction is privacy budget allocation, which has been studied profusely particularly in the context of *static databases*. **[TODO(roxana): Need to read a few of these papers.]** xxx

Sample Complexity Estimation in Statistics. **[TODO(daniel/kiran): Review existing approaches and make a clear statement of what's novel here (application + DP formulation?).]** xxx

Resource Allocation in Systems. An enormous body of work exists on resource allocation in systems []. Our approach is most similar to that SLA resource allocation [], in that we profile the estimated resources (samples) needed to reach a particular performance target (τ) for our workloads. In addition to minimizing resource consumption, many other considerations have been formulated in the past in this literature: ensuring fair allocation across workloads, avoiding starvation for these workloads, ensuring that workloads meet specific deadlines, etc. We acknowledge that our resource allocation formulation is limited and does not take into account such factors. A close look at this literature will broaden the scope of our own allocation procedure.

8 Conclusion

References

[1] Citibike system data. `https://www.citibikenyc.com/system-data`, 2018.

[2] Nyc taxi & limousine commission - trip record data. `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`, 2018.

[3] F. M. Harper and J. A. Konstan. The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4):19:1–19:19, Dec. 2015.

[4] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers, et al. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pages 1–9. ACM, 2014.

[5] J. Langford, L. Li, and A. Strehl. Vowpal Wabbit online learning project, 2007.

[6] F. D. McSherry. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09, pages 19–30, New York, NY, USA, 2009. ACM.

[7] Q. Shi, J. Petterson, G. Dror, J. Langford, A. Smola, and S. Vishwanathan. Hash kernels for structured data. *The Journal of Machine Learning Research*, 10:2615–2637, 2009.

[8] A. Zheng and A. Casari. *Feature engineering for machine learning: principles and techniques for data scientists*. 2018. OCLC: 1029545849.

A PTK Profiling Proofs

B Sage Semantic Proofs