# Budget-bounded Incentives for Federated Learning*

Adam Richardson[1], Aris Filos-Ratsikas[2], and Boi Faltings[1]

[1] Artificial Intelligence Laboratory
Ecole Polytechnique Fédérale de Lausanne (EPFL)
boi.faltings|adam.richardson@epfl.ch
[2] Department of Computer Science
University of Liverpool
aris.filos-ratsikas@liverpool.ac.uk

**Abstract.** We consider federated learning settings with independent, self-interested participants. As all contributions are made privately, participants may be tempted to free-ride and provide redundant or low-quality data while still enjoying the benefits of the FL model. In Federated Learning, this is especially harmful as low-quality data can degrade the quality of the FL model.
Free-riding can be countered by giving incentives to participants to provide truthful data. While there are game-theoretic schemes for rewarding truthful data, they do not take into account redundancy of data with previous contributions. This creates arbitrage opportunities where participants can gain rewards for redundant data, and the federation may be forced to pay out more incentives than justified by the value of the FL model.
We show how a scheme based on *influence* can both guarantee that the incentive budget is bounded in proportion to the value of the FL model, and that truthfully reporting data is the dominant strategy of the participants. We show that under reasonable conditions, this result holds even when the testing data is provided by participants.

**Keywords:** Federated Learning · Data Valuation · Incentives

## 1 Incentives in Federated Learning

Federated Learning [1] allows a set of participants to jointly learn a predictive model without revealing their data to each other. In this chapter, we assume that a coordinator communicates with the participants and distributes the Federated Learning (FL) model equally to all of them. Participants can contribute actual data or changes that improve the current FL model based on the data, which may be more compact.

There is then clearly an incentive to free-ride: a participant can benefit from the joint model without contributing any novel data, for example by fabricating data

* Supported by EPFL

that fits the current model, or using random noise. We call such strategies that are not based on actual data *heuristic* strategies. A participant may also wrongly report its data, for example by obfuscating it to achieve differential privacy [10]. There is no way for the coordinator to tell if data has been manipulated, and given that it can strongly degrade the FL model, it is important to protect the process against it. Even worse, a malicious participant could intentionally insert wrong data and poison the FL model; we do not consider malicious behavior in this chapter and assume that participants have no interest in manipulating the FL model.

Free-riding can avoided by *incentives* that compensate for the effort of a contributing participant. For federated learning, an incentive scheme will distribute rewards to participants in return for providing model updates, data, or other contributions to the learning protocol. Incentives should influence two behavior choices faced by participants:

- *observation strategy*: make the necessary effort to obtain truthful data and compute the best possible model update, rather than use a *heuristic* strategy to make up data with no effort, and
- *reporting strategy*: report the data *truthfully* to the coordinator, rather than perturb or obfuscate it.

We call participant behavior that is truthful in both regards *truthful* behavior. We observe that both properties can be satisfied if contributions are rewarded according to their *influence* [9] on the FL model. Influence is defined formally as the effect of the contribution on the loss function of the FL model:

- if the contribution is a model update, the improvement in the loss function through applying the update;
- if the contribution is data, the improvement in the loss function after adding the data to the training set.

For simplicity, we will refer to both cases as the contribution of a data point, even if data is often supplied in batches or in the form of a model update. The incentives for a batch of data is given as the sum of the incentives for the data points contained in it.

Clearly, influence is a good measure from the point of view of the coordinator, since it rewards contributions that make the FL model converge as fast as possible. The total expense is bounded by a function of the total reduction in the loss function, and so the coordinator can obtain the necessary budget for the rewards. It also allows participants to decide on their level of privacy protection and accept the corresponding reduction in their reward.

On the other hand, it is less clear what behavior such incentives induce in the participants. In this chapter, we answer the following questions:

- We show that when the coordinator evaluates contributions on truthful test data, the *dominant strategy* for participants is to invest effort in obtaining true data and to report it accurately. Thus, it avoids both aspects of free-riding.

- We show that participants will provide their data as soon as possible, so that there is no risk of participants holding back data hoping that it will gain higher rewards later.
- We show that when some or all of the testing data is supplied by particpants, truthful behavior is a *Bayes-Nash equilibrium* of the induced game. Furthermore, if a minimum fraction of the testing data is known to be truthful, truthful reporting is the dominant strategy for participants.

## 2 Related work

The topic of learning a model when the input data points are provided by strategic sources has been the focus of a growing literature at the intersection of machine learning and game theory. A related line of work has been devoted to the setting in which participants are interested in the outcome of the estimation process itself, e.g., when they are trying to sway the learned model closer to their own data points [6,8]. Our setting is concerned with the fundamental question of eliciting accurate data when data acquisition is costly for the participants, or when they are not willing to share their data without some form of monetary compensation.

A similar question to the one in our chapter was considered by [5] , where the authors design strategy-proof mechanisms for eliciting data and achieving a desired trade-off between the accuracy of the model and the payments issued. The guarantees provided, while desirable, require the adoption of certain strong assumptions. The authors assume that each participant chooses an *effort level*, and the variance of the accuracy of their reports is a strictly decreasing convex function of that effort. Furthermore, these functions need to be known to the coordinator. In this chapter, we only require that the cost of effort is bounded by a known quantity. Furthermore, our strategy space is more expressive in the sense that, as in real-life scenarios, data providers can choose which data to provide and not just which effort level to exert.

Our ideas are closely related to the literature of *Peer Consistency* mechanisms [11] such as the Bayesian Truth Serum or the Correlated Agreement mechanism, or the *Peer Truth Serum for Crowdsourcing* [16]. The idea behind this literature is to extract high-quality information from individuals by comparing their reports against those of randomly chosen peers. This approach has been largely successful in the theory of eliciting *truthful* information. The main problem with using such mechanisms for federated learning is that they also pay for redundant data that does not improve the model. If multiple participants submit exactly the same data, the coordinator would still have to pay the full reward to each of them. Thus, it is not possible to bound the budget of the coordinator.

Recently, Liu and Wei [2] proposed an incentive scheme for federated learning based on the correlated agreement mechanism [17]. However, it also does not satisfy budget-balance and allows arbitrage where participants gain rewards by replicating the existing FL model.

More generally, the issue of economics of federated learning and the important of budget-balance has recently been discussed in [3], where particular attention is

paid to keep participants from dropping out of the federation due to insufficient rewards.

Finally, [13] recently considered a setting where the value of the provided data is determined via the *Shapley value*. Their approach does not support rewarding data incrementally, as is required for federated learning, but computes rewards only when all data has been received. However, it is worth noting that they consider the influence approximation of [14] for approximating the Shapley value.

## 3 Incentives based on Influence

In our setting, there is a *coordinator* that wants to learn a model parametrized by $\theta$, with a non-negative loss function $L(z, \theta)$ on a sample $z = (x, y)$. The samples are supplied by a set $\mathcal{A}$ of *participants*, with participant $i$ providing point $z_i = (x_i, y_i)$. To simplify the exposition, we consider the contribution of a single data point as the most general case, but the results in this chapter also apply to batches of data points or model updates based on batches of data points, as is common in federated learning.

We will denote by $\mathcal{A}_{-i}$ the set of participants excluding participant $i$. Given a set of test data $Z = \{z_i\}_{i=1}^n$, the empirical risk is $R(Z, \theta) = \frac{1}{n} \sum_i L(z_i, \theta)$. The coordinator uses a scoring function $s(z)$ to determine the reward it pays for the data $z$.

### 3.1 Computing Influence

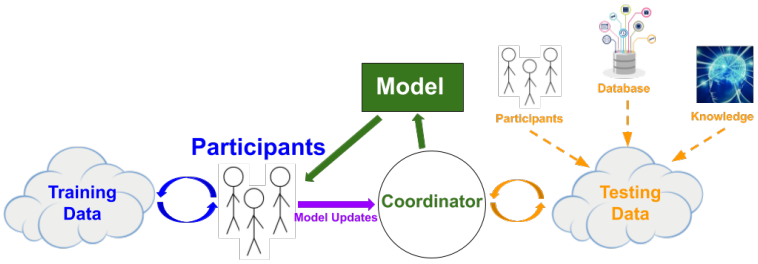


**Fig. 1.** The setting in this chapter: self-interested strategic participants observe data, translate it into model updates, and report to a coordinator. The coordinator maintains and broadcasts a joint FL model. To evaluate the quality of the contributions, the coordinator constructs an independent test set via other participants, a database, or other forms of prior knowledge. The coordinator scores model updates and rewards participants according to their influence on this test set.

The *influence* [9] of a data point is defined as the difference in loss function between the model trained with and without the data point. We generalize the

notion to data sets and model updates as the analogous difference between loss functions. We consider payments to participant $i$ that are proportional to the influence $I(D)$ of its contribution $D$: $pay_i(D) = \alpha I(D)$, where $\alpha$ is the same for all participants.

Computing loss functions requires access to a set of test data. In many federated learning settings, the coordinator actually never has access to data, but only model updates. In this case, it needs to ask participants to perform this evaluation. Figure 1 illustrates the process.

We distinguish two cases: the easier case where the center has access to independent and trusted test data, where we show that truthful behavior is a dominant strategy, and the more complex case where the center needs the cooperation of the strategic participants to perform this evaluation, and truthful behavior is a game-theoretic equilibrium.

Influence can be approximated [14] quite accurately (see the example in Figure 2) and this can greatly speed up computation and allows to protect privacy using multiparty computation [4]. Good approximations exist for linear and logistic regression models, and to some extent also for complex models such as neural networks. This approximation is based on taking the Taylor expansion of the loss function and down-weighting a training point to remove it from the dataset in a smooth manner. We find that taking only the first term of the expansion is not sufficient because then the expected influence of a point over the whole training set will be 0. Taking up to the second term of the expansion is sufficient for accuracy, speed, and good theoretical properties. Figure 2 shows that this second-order approximation for a linear regression example tracks the true influence extremely closely.

### 3.2 Budget Properties of Influence

In general, the share of an additional data point in a model based on $n-1$ earlier datapoints is $1/n$. Many loss functions, such as the variance or the cross entropy, decrease as $1/n$ with the number of samples. The influence is proportional to the derivative of the loss function and thus decreases as $1/n^2$.

Figure 2 shows an example of the actual decrease of influence on a regression model. We can observe two phases: an initial phase, where additional data is necessary to make the model converge, and a converged phase where the expected influence is close to zero. We believe that this is because the FL model is never a perfect fit to the data, but will always leave some remaining variance. Once this variance is reached, additional data will not help to reduce it, and no further incentives should be given to provide such data.

Using influence as an incentive has the following properties:

- the expected reward is either close to 0, or it decreases as $1/n^2$. Therefore, it is always best for participants to report data as early as possible.
- the expected reward for redundant or random data is zero.
- for the coordinator, the total expense is proportional to the decrease in the loss function.

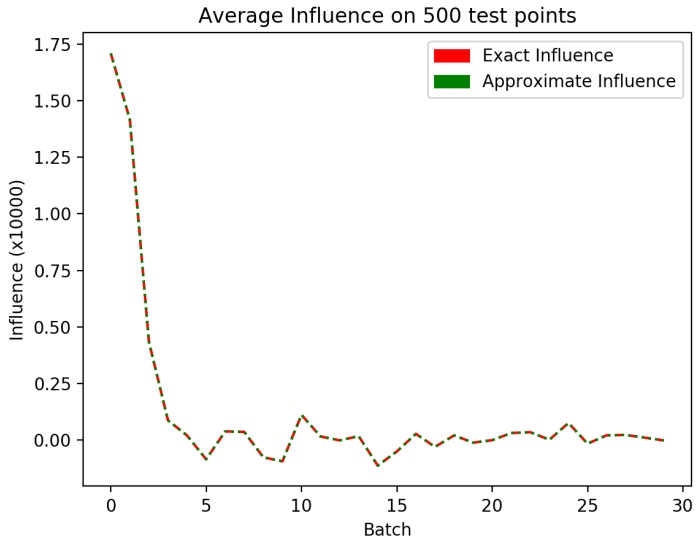


**Fig. 2.** Empirically observed decrease of influence on a typical regression model as more and more data is collected. Each batch corresponds to 100 data points. Both the exact influence and the 2nd order approximation are shown.

The last point has to be put in relation to the value of the model. If the value $f$ of the model is increasing linearly with its loss $R$, i.e. $f(R) = C - \beta R$, choosing $\alpha = \beta$ means that the budget matches the cost exactly.

If the value increases faster, to avoid a loss the coordinator will have to choose $\alpha$ to be at most the average $\beta$ over the loss reduction it intends to achieve. This means that the cost of obtaining the data may exceed its value during the initial phase, and the coordinator has to take some risk during the initial phase. If, on the other hand, it increases more slowly, the coordinator runs no risk, but may stop collecting data before the minimum variance is reached.

### 3.3 Assumptions

Our approach differs from other chapters on this topic in that we make relatively few assumptions about participant beliefs and effort models. We consider the participant and coordinator models as follows:

**Participant:** In the federated learning model, participants obtain data and contribute it to the federation, often in the form of a model update. For generality, we consider that participants contribute data points in some form.
*Observation strategy:* Each participant $i$ must decide to either exert effort $e_i(o)$ to observe data point $o$, exert 0 effort and obtain a data point based on some heuristic,

for example a random or constant point[3] When the participant decides to make an observation, it knows the expected effort $\delta_i$ over the distribution of observable data points. This value can vary amongst the participants. An observation $o_i = g(\phi_i|\psi_i)$, where $\phi_i$ is drawn from some shared underlying distribution $\Phi$, $\psi_i$ is a hidden latent random variable drawn from a shared distribution $\Psi$, and $g$ is a function that applies noise to $\phi_i$ given $\psi_i$. The participants believe that their noise is unbiased, formulated as follows: $\forall \phi \in \Phi$, $\forall i$, $\mathbb{E}_\Psi[g(\phi|\psi_i)] = \phi$.
*Reporting strategy:* Besides the observation, a participant also has to decide on a *reporting strategy* $r(o)$ that specifies what data it reports to the coordinator. We would like the participant to report *truthfully*, where $r$ is the identity function. However, a participant may also report differently, for example because it hopes to obtain a higher reward by a non-truthful report, or because it wants to protect the privacy of its data by adding noise or otherwise misreporting it.

Finally, we assume that participants are free to *opt-out*, and not provide any data nor obtain any reward. This strategy would in particular be adopted when the rewards are insufficient to cover the cost of effort.

The coordinator will employ a scoring function $s(\cdot)$ to provide rewards to the participants, dependent on their reports; we postpone details about this scoring function until the next section. The scoring function is chosen to influence the strategy choices of participants among different options. Participants are rational, so they will choose the strategy that maximizes their expected utility.

We make one further assumption about participant beliefs. For this we introduce the notion of *risk-monotonicity*, which is the notion that a model learner is monotonic in the true risk over the number of data points in the training set. While [15] show that not all empirical risk minimizers are risk-monotonic in the number of training points, their counter-examples are adversarially constructed. As participants have no prior information about the distributions, we consider it reasonable to make the following formal assumption:

> *The participants believe the coordinator's model is risk-monotonic with respect to the true distribution $\Phi$, i.e. a participant expects that a point drawn from $\Phi$ will not worsen the model's expected risk when evaluated on $\Phi$.*

**Coordinator:** The coordinator wishes to construct a model because they believe they can extract some profit from this model. We assume the profit is a function $f(R)$ of the model risk. The expected utility of the coordinator is then the profit $f(R) - c(R)$, where $c(R)$ is the expected cost of constructing a model with risk $R$. We assume that $f(R)$ is monotonically decreasing, as discussed in Section 3.2.

Given a profit function $f(R)$, the utility of the coordinator is at least $f(R) - \sum_i \alpha_c \text{Infl}(o_i)$. The coordinator needs to choose $\alpha_c$ to ensure that this profit is positive. At the same time, $\alpha_c$ determines the reward paid to participants, and must at least cover their cost of participation. Otherwise, participants may decide

[3] It is straightforward to extend the results in this chapter to a setting where increased effort results in increased quality, but this would require to characterize the exact relation which depends on the application.

to opt out, and there is a risk that the federation could be left with too little data. Therefore, the coordinator must tune $\alpha_c$ to achieve both budget balance and sufficient participation.

For evaluating the model risk $R$, we consider two cases: *(a)* the coordinator may possess an *independent test set*, or *(b)* it may have to *acquire a test set* from participants.

## 4 Game-theoretic Incentives for Participants

As the score of the data provided by a participant depends on the data provided by others, the choice of participant strategies is a game-theoretic problem.

Following standard game-theoretic terminology, we will say that a participant supplying point $r_j$ is *best responding* to the set of strategies $r_{-j}$ chosen by the other participants, if the strategy that it has chosen maximizes the quantity $\mathbb{E}[s(r_j|r_{-j}) - e_i(r_j)]$ over all possible alternative reports $r'_j$, where the expectation is over the distribution of reports of the other participants. We will say that a vector of strategies (i.e., a strategy profile) $(r_1, \ldots, r_n)$ is a *Bayes-Nash equilibrium (BNE)* if, for each participant $j$, $r_j$ is a best response. If $r_j$ is a best response to any set of strategies of the other players, we will say that $r_j$ is a *dominant strategy*.

An *incentive scheme* is a function that maps data points $z_i$ to payments $s(z_i)$; intuitively, a good incentive scheme should overcome the cost of effort (as otherwise participants are not incentivized to submit any observations) but also, crucially, to reward based on the effect that the data point $z_i$ has on improving the accuracy of the trained model. For this reason, we will design incentive schemes via the use of influences. Let $Z_{-j} = \{z_i\}_{i \neq j}$ and let

$$\hat{\theta} = \arg\min_{\theta} R(Z, \theta) \quad \text{and} \quad \hat{\theta}_{-j} = \arg\min_{\theta} R(Z_{-j}, \theta).$$

We will assume that the coordinator is in possession of an *test set* $T = \{z_k\}$. Then the *influence* of $z_j$ on the test set is defined as

$$\text{Infl}(z_j, T, \theta) = R(T, \hat{\theta}_{-j}) - R(T, \hat{\theta}).$$

We will simply write $\text{Infl}(z_j)$, when $T$ and $\theta$ are clear from the context. Then, we can design the following incentive scheme:

- Case 1: The coordinator possesses an independent test set: $s(r_i) = \alpha_c \cdot \text{Infl}(r_i) - \epsilon$, where $\epsilon > 0$ is a very small value.
- Case 2: The coordinator draws its test set from the reported data; they are rewarded in the same way as data used for the FL model, but not used in learning the model.

### 4.1 Using independent and truthful test data

For the lemmas and theorems in this section, we make the following assumptions:

- Observation noise is unbiased and non-trivial, as stated in the previous section.
- participants have no prior knowledge of the true distribution $\Phi$ or the model of the coordinator.

The proofs of the following statements are omitted due to lack of space, but are included in the supplement.

**Theorem 1.** *A participant having made no observation of the data believes the expected influence of any particular report to be 0.*

*Proof.* Let $\mathcal{D}$ be the domain of all possible sets of reports. The coordinator defines some non-negative loss function $\mathcal{L} : \mathcal{D} \to \mathbb{R}_0^+$, which serves as a blackbox that incorporates both training and testing. Given some set of reports $\{z\} \in \mathcal{D}$, define $B(\mathcal{L}|\{z\})$ as the random variable that represents the ex-ante belief of a participant on the value of $\mathcal{L}(\{z\})$. Lack of knowledge about both $\mathcal{L}$ and $\Phi$ induces the relation $B(\mathcal{L}|\{z\}_0) = B(\mathcal{L}|\{z\}_1)$ for all $\{z\}_0, \{z\}_1 \in \mathcal{D}$. For some report $r$ and some set of other reports $\{z\}$, the influence score is defined as $\text{infl}(\{z\}, r) = \mathcal{L}(\{z\}) - \mathcal{L}(\{z\} \cup r)$. Then a participant believes that its score will be $B(\mathcal{L}|\{z\}) - B(\mathcal{L}|\{z\} \cup r)$. In expectation, this score is $\mathbb{E}[B(\mathcal{L}|\{z\}) - B(\mathcal{L}|\{z\} \cup r)] = \mathbb{E}[B(\mathcal{L}|\{z\})] - \mathbb{E}[B(\mathcal{L}|\{z\} \cup r)] = 0$

**Lemma 1.** *A participant $A_i$ believes that, almost certainly, given a finite number of reports, $\mathbb{E}_\Phi[\mathbb{E}_\Psi[\mathit{Infl}(o_i|\{o_j\}_{j\neq i})]] > 0$ when evaluated on $\{z_{test}\}$ with $z_{test}$ in the distribution of observations.*

*Proof.* Define $B_0(\mathbb{E}_\Phi[\mathbb{E}_\Psi[L(z_{test}, z)]]) = \mathbb{E}_\Phi[L(z'_{test}, z)] + a(g, \Psi)$ as a participant's belief about $\mathbb{E}_\Phi[\mathbb{E}_\Psi[L(z_{test}, z)]]$, where $z'_{test}$ is drawn from $\Phi$, and $a$ is unknown, but does not depend on $L$ because the participants have no knowledge of $L$, and therefore no way of knowing if $L$ introduces some bias given $\Psi$. It similarly follows that a participant's belief $B_1(\mathbb{E}_\Psi[\text{Infl}(o_i|\{o_j\}_{j\neq i})] = \text{Infl}(\phi_i|\{o_j\}_{j\neq i}) + b(g, \Psi)$. Therefore, it is only necessary to show that $\mathbb{E}_\Phi[\text{Infl}(\phi_i|\{o_j\}_{j\neq i})] > 0$ when evaluated on $z' \in \Phi$. This follows directly from participant assumptions about risk-monotonicity.

The following theorem asserts that as long as the test set consists of truthful information, the dominant strategy for participants is to either be truthful of opt out. In the case where the coordinator possesses an independent test set, this condition is satisfied trivially.

**Theorem 2.** *Suppose that (a) the noise is unbiased and non-trivial, (b) the participants do not have knowledge of the distribution or the model and (c) the test set consists of truthful information. Then,*

- *for any $\alpha_c > 0$, for every participant the dominant strategy is either being truthful or dropping out, and*
- *there is a large enough $\alpha_c$ such that for every participant, almost certainly, being truthful is the dominant strategy.*

*Proof.* By Theorem 1, we have that $\alpha_c \text{Infl}(h_i) = 0$, therefore the expected utility of heuristic reporting is negative. By Lemma 1, the participant believes that $\mathbb{E}_\Phi[\mathbb{E}_\Psi[\text{Infl}(o_i|\{o_j\}_{j\neq i})]] > 0$ almost certainly, therefore if $\alpha_c > \frac{\delta_i}{\mathbb{E}_\Phi[\mathbb{E}_\Psi[\text{Infl}(o_i|\{o_j\}_{j\neq i})]]}$, then the participant believes he or she will receive a positive utility almost certainly. If this inequality is not satisfied, the participant will receive a negative utility regardless of the choice of strategy and will opt out. There is always a large enough $\alpha_c$ such that the inequality is satisfied and the participant will be truthful.

We have thus shown that our incentive scheme induces truthful behavior as the dominant strategy for all participants that do not opt out, and that furthermore given a large enough payment no participants will opt out.

### 4.2 Using participant reports as test data

We have shown in Theorem 2 that under some reasonable assumptions, truthful reporting is a dominant strategy for the participants. However, this requires a truthful test set, which might not always be at the disposal of the coordinator. There are also good reasons for the coordinator to collect test data from participants: it allows it to cover a broader spectrum of cases, or to accomodate concept drift. We first observe that, even if we collect the reports as test data, truthful behavior is a Bayes-Nash Equilibrium:

**Theorem 3.** *Suppose that (a) the noise is unbiased and non-trivial, (b) the participants do not have knowledge of the distribution or the model and (c) the test set consists of data provided by participants under the incentive scheme. Then,*

- *for any $\alpha_c > 0$, there is a Bayes-Nash Equilibrium where every participant is either truthful or drops out, and*
- *there is a large enough $\alpha_c$ such that, almost certainly, there is a Bayes-Nash Equilibrium where all participants are truthful.*

*Proof.* If we assume that all participants that do not drop out are truthful, then the test set is made up of truthful data. By Theorem 2, truthful behavior is the best response for all participants, so it forms a Bayes-Nash equilibrium.

An equilibrium is a weaker notion than dominant strategies, so it is interesting to ask if the coordinator can make truthful behavior the dominant strategy even when test data has to be obtained from participants. Clearly, if all test data is supplied by participants, this is not possible: consider the example where all but one participant $i$ submit test data according to a synthetic model $M'$, but only participant $i$ observes true data according to a different true model $M$. Then it will be better for participant $i$ to report incorrectly according to model $M'$, so truthful behavior cannot be a dominant strategy.

However, it turns out that if only a fraction of the test data is supplied by untrusted agents, we can place a bound on this fraction so that truthful

behavior is still a dominant strategy. To obtain such a result, we need to exclude consideration of the cost of obtaining data, since we do not know what is the relative cost of obtaining true vs. heuristic data, and focus on the reporting strategy only.

Let $\Phi_1$ be the distribution of truthful reports and $\Phi_2$ be the distribution of heuristic reports. We assume they describe an input-output relationship such that $\Phi(x, y) = q(x)p(y|x)$, and $q_1(x) = q_2(x)$. This assumption merely asserts that the data we are collecting is drawn from the same domain regardless of the distribution of the output. Distributions $\Phi_1$ and $\Phi_2$ determine, in expectation, models $M_1$ and $M_2$ respectively. Let us now define $R_{i,j}$ as the expected risk of model $M_i$ evaluated on distribution $\Phi_j$, and let $I_{i,j}$ be the influence of a datapoint sampled from distribution $\Phi_i$ on a test point from distribution $\Phi_j$. Using the standard mean-squared-error loss function, we have that $R_{i,j} = R_{j,j} + \mathbb{E}[(M_i - M_j)^2]$. We then have the following:

**Theorem 4.** *As long as the test data contains at most a fraction*

$$p < \frac{I_{2,2}/R_{2,2}}{I_{1,1}/R_{1,1} + I_{2,2}/R_{2,2}} + \frac{I_{1,1} - I_{2,2}}{r(I_{1,1}/R_{1,1} + I_{2,2}/R_{2,2})}$$

*of non-truthful reports, truthful reporting remains the dominant strategy for participants that do not choose to opt out.*

*Proof.* Now suppose we sample $x_1$ points from $\Phi_1$ and $x_2$ points from $\Phi_2$ to form our training set $\{z\}$, and call the resulting distribution $\Phi_c$. Now note that as $R_{1,2} - R_{1,1} = r$, and influence is proportional to the empirical risk, the influence of a datapoint following $M_1$ but tested on a sample from $\Phi_2$ is decreased as follows:

$$I_{1,2} = I_{1,1}(1 - r/R_{1,1})$$

and so the expected influence when evaluating on the mixture $(x_1, x_2)$ is

$$\begin{aligned} I_{1,c} &= I_{1,1}(1 - r/R_{1,1}\frac{x_2}{n}) = I_{1,1}(1 - pr/R_{1,1}) \\ I_{2,c} &= I_{2,2}(1 - r/R_{2,2}\frac{x_1}{n}) = I_{2,2}(1 - r(1-p)/R_{2,2}) \end{aligned}$$

To ensure that reporting samples from $\Phi_1$ carry a higher expected reward, we want to satisfy:

$$\begin{aligned} I_{1,c} &> I_{2,c} \\ I_{1,1} - I_{2,2}(1 - r/R_{2,2}) &> pr(I_{1,1}/R_{1,1} + I_{2,2}/R_{2,2}) \\ p &< \frac{I_{1,1} - I_{2,2}(1 - r/R_{2,2})}{r(I_{1,1}/R_{1,1} + I_{2,2}/R_{2,2})} \\ &= \frac{I_{2,2}/R_{2,2}}{I_{1,1}/R_{1,1} + I_{2,2}/R_{2,2}} + \frac{I_{1,1} - I_{2,2}}{r(I_{1,1}/R_{1,1} + I_{2,2}/R_{2,2})} \end{aligned}$$

If $I_{2,2}/R_{2,2} = I_{1,1}/R_{1,1}$, the first term is $= 1/2$. The second term is a correction: if $I_{1,1} > I_{2,2}$, more non-truthful reports are tolerated as the influence when improving the first model is stronger, otherwise it is the other way around.

A coordinator could use this result to decide how much test data to obtain from participants. As the underlying phenomenon could evolve over time, it is advantageous for the coordinator to include some contributed data in its test set so that such evolution can be tracked. To evaluate the bound, the coordinator could compare the statistics of scores obtained with trusted test data with those obtained using contributed test data, and thus estimate the parameters $I$, as well as the empirical risks of models fitted to the trusted and contributed data to estimate the parameters $R$. It could thus obtain a stronger guarantee on the quality of the test data.

## 5 Conclusion

When federated learning is extended to allow self-interested participants, free-riding participants that submit low-quality or redundant data can have significant negative impact on the result. Thus, it is important to provide incentives that reward truthful data providers for their effort.

As the economics of a federated learning system can be complex ([3]), it is important that the incentive scheme works with a bounded budget that is tied to the quality of the resulting data. We have shown that a scheme based on influence satisfies this criterion. At the same time, we have shown that it induces the desired truthful behavior as dominant strategies in participants, and that this holds even when some of the testing data is obtained from non-truthful participants.

An important question for future work is how to compute the incentives while maintaining privacy of the data. In the current scheme, a participant has to submit its contribution, whether data or a model update, before the influence can be computed. For limited cases, we have already shown how to compute an influence approximation privately ([4]), but the general question remains open.

## References

1. Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. ACM Trans. Intell. Syst. Technol. 10, 2, Article 12 (February 2019), 19 pages. DOI:https://doi.org/10.1145/3298981
2. Liu,Y. and Wei,J. Incentives for Federated Learning: a Hypothesis Elicitation Approach. ICML workshop on Incentives in Machine Learning, 2020.
3. Yu, H., Liu, Z., Liu, Y., Chen, T., Cong, M., Weng, X., Niyato, D. and Yang, Q. A sustainable incentive scheme for federated learning. IEEE Intelligent Systems 35(4), doi:10.1109/MIS.2020.2987774, IEEE (2020).
4. Richardson, A., Filos-Ratsikas, A., Rokvic, L., and Faltings, B. Privately Computing Influence in Regression Models. AAAI 2020 Workshop on Privacy-Preserving Artificial Intelligence.

5. Cai, Y.; Daskalakis, C.; and Papadimitriou, C. 2015. Optimum statistical estimation with strategic data sources. In Grnwald, P.; Hazan, E.; and Kale, S., eds., *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, 280–296. Paris, France: PMLR.
6. Caragiannis, I.; Procaccia, A.; and Shah, N. 2016. Truthful univariate estimators. In *International Conference on Machine Learning*, 127–135.
7. Chen, Y.; Immorlica, N.; Lucier, B.; Syrgkanis, V.; and Ziani, J. 2018a. Optimal data acquisition for statistical estimation. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 27–44. ACM.
8. Chen, Y.; Podimata, C.; Procaccia, A. D.; and Shah, N. 2018b. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 9–26. ACM.
9. Cook, R. D., and Weisberg, S. 1980. Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics* 22(4):495–508.
10. Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.
11. Faltings, B., and Radanovic, G. 2017. Game theory for data science: eliciting truthful information. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 11(2):1–151.
12. Faltings, B.; Jurca, R.; and Radanovic, G. 2017. Peer truth serum: Incentives for crowdsourcing measurements and opinions. *CoRR* abs/1704.05269.
13. Jia, R.; Dao, D.; Wang, B.; Hubis, F. A.; Hynes, N.; Gurel, N. M.; Li, B.; Zhang, C.; Song, D.; and Spanos, C. 2019. Towards efficient data valuation based on the shapley value. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*.
14. Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In Precup, D., and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1885–1894. International Convention Centre, Sydney, Australia: PMLR.
15. Loog, M.; Viering, T.; and Mey, A. 2019. Minimizers of the empirical risk and risk monotonicity. In *Advances in Neural Information Processing Systems*, 7476–7485.
16. Radanovic, G.; Faltings, B.; and Jurca, R. 2016. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7(4):48.
17. Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David C Parkes. Informed truthfulness in multi-task peer prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 179–196, 2016.