Game-theoretic Mechanisms for Eliciting Accurate Information

Boi Faltings
Artificial Intelligence Laboratory (LIA)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
boi.faltings@epfl.ch

Abstract

Artificial Intelligence often relies on information obtained from others through crowdsourcing, federated learning, or data markets. It is crucial to ensure that this data is accurate. Over the past 20 years, a variety of incentive mechanisms have been developed that use game theory to reward the accuracy of contributed data. These techniques apply to many settings where AI uses contributed data.

This survey categorizes the different techniques and their properties and shows their limits and tradeoffs. It identifies open issues and points to possible directions to address these.

1 Introduction

Artificial Intelligence makes decisions based on data, whether it is knowledge acquisition through machine learning, tuning policies through reinforcement learning, or deciding on the best action based on data about the current situation. Errors and biases in the data can have serious consequences.

Consider the example of product reviews: as there are usually no rewards, reviews are mainly submitted for ulterior motives either to improve the reputation of a product, or to vent anger about the product [Hu et al., 2017]. However, the average customer has little motivation to leave a review, and so there are comparatively few reviews that indicate the satisfaction of an average customer. Ratings therefore often show a "U"-shaped distribution, as the example on the left in Figure 1. Using such a biased sample for machine learning or recommendation leads to poor performance and unfairness [Goel et al., 2020].

Figure 1 right illustrates community sensing: a real-time pollution map can be constructed from sensors operated by individuals in different locations. But these individuals have little motivation to maintain their sensors so that they are accurate [Faltings et al., 2014], and this limits the accuracy of the map. The same situation exists in crowdwork such as labeling data, where workers have the motivation to minimize their effort for each task.

A service-level agreement for an Internet service provider may prescribe a certain connection quality as experienced by the user. However, these users have a motivation to misreport this data in order to obtain compensation [Jurca et al., 2007].

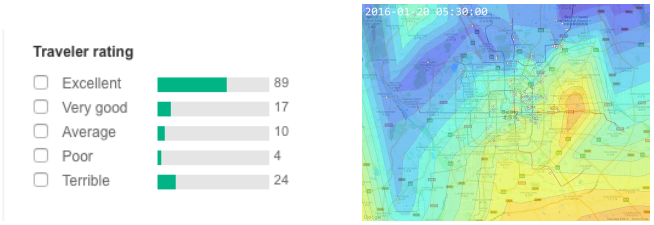

Figure 1: Some practical examples of data elicitation: product reviews and pollution sensing.

The solution to the problem of biased reviews is to provide rewards that motivate an average user to provide data, not just users who have ulterior motives. However, in order to ensure that they make a sufficient effort to provide accurate data, the reward should also depend on its quality. Such rewards would also help with motivating effort in community sensing, and even for obtaining truthful evaluations for service-level agreements.

At first glance, this may appear impossible, as it seems to require verification of the accuracy of the data. However, when multiple agents can observe the same, or correlated data, it is possible to use the correlation between their observations to make accurately reporting them the most rewarding strategy!

Over the past 20 years, numerous mechanisms that apply to various scenarios have been developed. While each of them has its own set of particular assumptions, they all consider a similar framework:

- a *center* wants to obtain information about one or several *tasks*, each represented by a random variable with a finite set of possible values $v_1, ..., v_k$.
- a set of *agents* can observe the variable, and send a report r of the observation o to the center.
- in return, the center rewards the agents for their reports.

Agents can report data about the phenomenon in different forms. The simplest form is *objective* data which gives information about a ground truth, for example, a temperature measurement. Here, the center would like to determine the actual ground truth value. However, other data such as product reviews are *subjective* data: different reviewers apply different criteria and there is no single ground truth. In this case, the

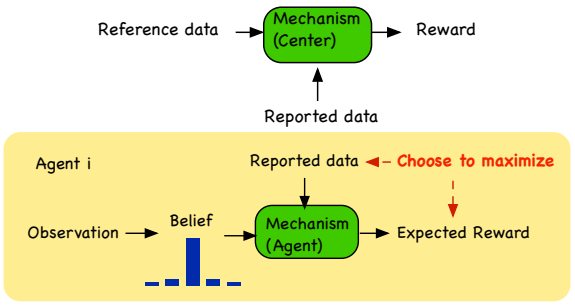

Figure 2: The mechanism implemented by the center (top) is simulated by the agents and influences the data they report. Agents form a belief about the reference data through their observations, and choose the reported data to maximize the expected reward.

center would want to obtain the *distribution* of evaluations. In the scenario of federated learning, the center receives model updates (typically gradients) and is interested in learning the best possible model.

2 Principles of Peer Consistency

When data observed by different agents are correlated, the quality of the data they report can be estimated from how well they follow this correlation. In statistics, such correlations are called *stochastic relevance*. The center determines a reward based on the *consistency* of the reported data with a *reference*, which is usually data reported by another peer or data constructed from peer reports. Agents will choose the data they report so that they maximize the reward they obtain. We call such mechanisms collectively *peer consistency* mechanisms (Figure 2). The resulting mechanism is *truthful* if agents expect the highest reward by reporting information they believe to be correct and accurate. The overwhelming majority of the literature on peer consistency mechanisms uses one of the following three ways to reward a data report r:

1. Output agreement: the probability of r *agreeing* with peer report (output agreement) (reviewed in Section 5).
2. Information theoretic: how well r *fits* the distribution of peer reports (reviewed in Section 6).
3. Model quality: how much r *improves* the fit of the model with the peer reports (reviewed in Section 7).

3 Proper Scoring Rules

To get an intuition for how the mechanisms work, it is instructive to look at the example of proper scoring rules [Geiting and Raftery, 2007]. They have been invented to assess the quality of a forecast, given as a probability distribution $\mathbf{P}$ over a space of k possible outcomes $\{x_1, .., x_k\}$. A scoring rule $SR(o, \mathbf{P})$ takes as arguments the actual outcome $o \in \{x_1, .., x_k\}$ and a vector of probabilities $\mathbf{P} = (p(1), .., p(k))$ representing the forecast, and returns a score. It is called proper if the expected score of an observed outcome believed to be from a distribution $\mathbf{Q} = (q(1), .., q(k))$ is maximized when $\mathbf{P} = \mathbf{Q}$. The most well-known scoring rules are:

- logarithmic scoring rule: $SR(O, \mathbf{P}) = C + \log p(o)$
- quadratic scoring rule: $SR(o, \mathbf{P}) = 2p(o) - \sum_{i=1}^{k} p(i)^2$

Consider the expected reward of the logarithmic scoring rule to an agent that believes the outcomes to be distributed according to $\mathbf{Q}$ and reports $\mathbf{P}$:

$$E_Q[SR(o, \mathbf{P})] = \sum_x q(x) \cdot [C + \log(p(x))]$$

so that the difference between reporting $\mathbf{Q}$ and $\mathbf{R} \neq \mathbf{Q}$ is:

$$\begin{aligned} & E_Q[SR(o, \mathbf{Q})] - E_Q[SR(o, \mathbf{R})] \\ = & \sum_x q(x) \cdot [C + \log q(x)) - (C + \log r(x))] \\ = & \sum_x q(x) \cdot \log \frac{q(x)}{r(x)} \\ = & D_{KL}(Q||R) \end{aligned}$$

By Gibbs' inequality, $D_{KL}(\mathbf{Q}||\mathbf{R}) \geq 0$, so reporting an $\mathbf{R} \neq \mathbf{Q}$ can only get a lower score! Note how the agent's own beliefs ($\mathbf{Q}$) about the correct data are used to make it report the distribution it truly believes is the most accurate.

However, proper scoring rules require that there is a true value that becomes known in order to evaluate the score. In most examples we are interested in, the ground truth is impossible or too costly to obtain.

Prediction Markets. Another use of scoring rules is in *prediction markets* [Hanson, 2007]. They are used to integrate predictions from multiple agents into a public consensus prediction R_t that evolves as agents integrate their predictions. An agent that wishes to increase the predicted probability of x_i buys from a market-maker a security at the cost of $SR(x_i, R_t)$. In response, R_{t+1} is changed in the desired direction. At the end, when the true value x^* becomes known, the securities for x^* pay out \$1 each, and all others expire without value. The change in R_t is computed according to a formula that ensures that the gain to the agent buying a share is exactly the improvement in R_t, i.e.:

$$SR(R_{t+1}, g) - SR(R_t, g)$$

Prediction markets are an efficient way to elicit a true consensus opinion, such as predicting the outcome of elections or the success of new products. Scoring rules are important to ensure liquidity: that participants can always buy and sell shares at the current market price.

4 Properties of Game-Theoretic Mechanisms

Table 1 compares several well-known mechanisms according to several features that are described below.

Subjective Data. Subjective data arises when there is no objective ground truth. For example, when reviewing products or hotels, everyone applies different criteria and preferences, so there is no objective true quality. Some mechanisms explicitly assume that reports are statistically independent given a ground truth so do not apply to such settings.

Uninformative Equilibria. The contributors could agree on a fictious version of the phenomenon and submit perfectly coordinated data that maximizes the payment but provides no accurate information. We call these *uninformative* equilibria. In Table 1, a checkmark means that the mechanism *avoids* uninformative equilibria.

Mechanism	Subjective Data	Uninformative Equilibria	# tasks	# tasks/agent	minimal	detail-free
Output Agreement	✗	✗	1	1	✓	✓
Corr. Agreement	✗	✓	many	2	✓	✓
Peer Prediction	✓	✗	1	1	✓	✗
Info Theoretic	✗	✓	many	many	✓	✓
BTS	✓	✓	1	1	✗	✓
RBTS	✓	✓	≥ 3	1	✗	✓
PTS	✓	✗	1	1	✓	✗
PTSC	✓	✓	many	1	✓	✓

Table 1: Comparison of the properties of different mechanisms.

Multi-Task. Some mechanisms make use of statistics taken over multiple tasks as part of the reward computation, so they require multiple tasks. For example, a measurement may be reported at several locations, or ratings may be collected for many similar products.

Multiple Tasks per Agent. Some mechanisms furthermore require each agent to provide data for multiple tasks using the same strategy throughout, and sometimes even that each pair of agents answer multiple tasks in common. For example, each agent may have to rate multiple products, or every pair of agents may have to rate overlapping sets of products.

Minimal. Some mechanisms are not minimal in that they require agents to report not just the data, but also additional information (typically an estimate of the reports submitted by other participants).

Detail-Free. Some mechanisms are not detail-free in that the mechanism requires prior information about participants' beliefs, or the distribution of data, for example.

Figure 3: The ESP game: the task is to pick keywords that best describe the content of an image. Matching the picks of an (unknown) peer gives points. Overly general keywords are excluded as Taboo words.

5 Output Agreement

A Historical Example: The ESP Game The first wide-scale application of peer consistency was the ESP game (see Figure 3), which was designed for labeling images to train computer vision systems [von Ahn and Dabbish, 2004]. It later became the Google image labeler and provided the labeled image data for the first practical computer vision systems.

The principle of the ESP game is to reward agents for reporting data that matches that of a peer agent and this has become known as *output agreement*. Truthfully reporting the correct label is an *equilibrium*: if both agents follow this strategy, their answers will match and they will obtain the reward.

Subjective Data: The Peer Prediction Method. Consider reporting the quality of blue star airlines, a company that has an excellent reputation for service and punctuality, but did not treat me well at all: the plane was 8 hours late, and the baggage was lost. With output agreement, if I report a bad rating I know this is unlikely to match my peer. Instead, I maximize my expected reward by giving a dishonest positive report and remaining silent about my true experience. Thus, output agreement is not truthful for such *subjective* data.

The first solution to this problem was the *peer prediction* mechanism [Miller, Resnick and Zeckhauser, 2005]. It associates each possible report with a probability distribution that models the biased belief of the agent reporting it. By computing the reward on that distribution using a proper scoring rule for a randomly chosen peer report, the bias is corrected and truthful reports earn a higher expected reward even when the observed value is an unlikely one. Peer prediction has never been used in practice as it requires not only knowing the posterior beliefs but also that they are the same for all agents.

Another possibility is to scale the reward for agreement according to its likelihood, so that agreement on less likely values pays a higher reward. In the square root agreement mechanism [Kamble et al., 2016], the reward for reporting a value x is the reciprocal of the square root of the probability of two reports agreeing on x; in the peer truth serum [Jurca and Faltings, 2011; Radanovic et al., 2016], it is the reciprocal of the probability of x occurring as a report.

Uninformative Equilibria. In an image labeling system, while truthfully reporting labels is clearly an equilibrium, an even easier strategy would be for agents to simply label everything the same, for example, to give the same label "horse" no matter what the image. This also constitutes an equilibrium

and a very good one as it requires no effort and is guaranteed a reward. However, it gives no information to the center.

The ESP game avoided this by having a taboo list of words that gave no reward, but this is not a general solution. What if one of these words was in fact the correct label? Several works [Jurca and Faltings, 2005; Waggoner and Chen, 2014] show that all single-task output agreement mechanisms necessarily have uninformative equilibria. However, one can make sure that they carry an expected reward of zero, making it uninteresting for agents using them to participate.

[Dasgupta and Ghosh, 2013] were the first to show how to do this by *subtracting* from the agreement reward the expected agreement reward the agents would obtain for pairs of *different* but similar tasks. In an uninformative equilibrium, agents agree for all pairs of tasks, so both terms would be equal - leading to an expected reward of zero!

However, this *multi-task* mechanism requires that each agent answers multiple similar tasks with the same strategy. This may be true when data is reported by algorithms, such as sensor data, but not for crowdsourcing such as reviews.

Correlated Agreement. In many practical situations, there are multiple correct answers, and the strategy of output agreement to only reward exact matches is too harsh, as a match with a strongly correlated answer should also be rewarded. The correlated agreement mechanism [Shnayder et al., 2016] thus rewards all matches with correlated signals. The correlations are represented by a correlation matrix Δ of all possible pairs of signals x and y: $\Delta(x,y) = Pr(x,y) - Pr(x)Pr(y)$. The score for a report x from agent A for a *bonus* task t_1, using peer report y for the same task submitted by a randomly chosen agent B, is:

$$S(x,y) = \begin{cases} 1 & \text{if } \Delta(x,y) > 0 \\ 0 & \text{otherwise} \end{cases}$$

To avoid uninformative equilibria, it uses the idea of [Dasgupta and Ghosh, 2013]: randomly choose two similar but different *penalty* tasks t_2 and t_3, where t_2 is answered by A with report v, and t_3 is answered by a second peer agent C with report w . Compute the reward to A by comparing the scores:

$$Pay(x,y) = S(x,y) - S(v,w)$$

With this payment rule, we see that the expected payment for truthful reporting is the sum of all the positive entries in Δ:

$$E[pay] = \sum_{i,j} \Delta(x_i,x_j)S(x_i,x_j) = \sum_{i,j,\Delta(x_i,x_j)>0} \Delta(x_i,x_j)$$

If a non-truthful strategy were to sum different elements, their sum cannot become larger. Therefore, truthful strategies result in the highest-paying equilibrium!

Recently, it has been observed that t_2 can be replaced by the bonus task t_1, using x in place of v [Zhang and Schoenebeck, 2023]. This not only eliminates the requirement that each agent solves at least 2 tasks, but also rules out manipulation by using different reporting strategies for the bonus and penalty tasks. They also propose a *matching agreement mechanism* to eliminate undesirable influence by inconsistencies in the peer reports.

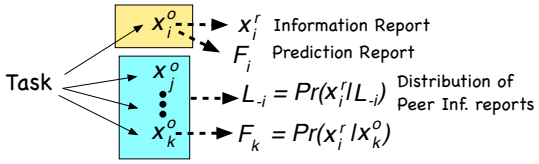

Figure 4: *Bayesian Truth Serum: Agent A_i observes the information in yellow; it is scored against the information from others (blue).*

Recent research has considered the question of whether participants subjected to a peer consistency reward scheme will learn to use truthful reporting. For a family of no-regret learning algorithms, [Feng at al., 2022] have shown that the correlated agreement mechanism will lead agents to learn truthful reporting strategies.

6 Information Theoretic Measures

Output agreement has difficulty with subjective data because it does not take into account the prior expectation of the observation. One way to correct this is to let agents also report their belief about the distribution of peer reports, called *prediction reports*. For example, if I get bad service from an otherwise good airline, I could also report my expectation that most others will see good service.

When we use a proper scoring rule to score the reported data against the distributions expressed by the prediction reports, we obtain a measure that in expectation is equivalent to the mutual information between the report and the observation of another agent. We can then use the data processing inequality to prove that reporting truthfully will yield the highest score. Let us see how this works in detail.

Bayesian Truth Serum: Using Peer Beliefs As Prior. The first and best known example of such a mechanism is the Bayesian Truth Serum (BTS) [Prelec, 2004], shown schematically in Figure 4. An agent A_i submits two reports:

- an *information* report x_i^r and
- a *prediction* report F_i predicting the distribution of peer information reports.

We let $L_{-i}(x)$ denote the distribution of information reports submitted by peers other than A_i. The reward is a weighted sum of two scores:

- The prediction score penalizes inaccurate prediction reports:
$$\tau_{pred}(F_i) = -D_{KL}(L_{-i}||F_i)$$
- The *information score* τ_{inf} rewards information reports that are more common than predicted:
$$\tau_{inf}(x_i^r) = \log L_{-i}(x_i^r) - \frac{1}{n-1}\sum_{j\neq i} \log F_j(x_i^r)$$

We note that the sum of information scores and prediction scores for all agents is always equal to zero. The mechanism thus creates a competition among agents that counteracts collusion for uninformative equilibria.

Why BTS Is Truthful. The prediction score is truthful as it is maximized at 0 when the prediction report matches the actual distribution of peer information reports. It remains to show that the information score is also truthful. Here, we show a simple and elegant proof based on the data processing inequality that has been developed by [Kong and Schoenebeck, 2019].

The information score is equal to the expectation of:

$$\tau'_{inf} = \log L_{-i}(x_i^r) - \log F_j(x_i^r)$$

using the prediction report of a randomly chosen peer agent j. The expectation $E[\tau'_{inf}]$ of τ'_{inf} over x_i^o and x_j^o is:

$$\begin{aligned} & \underbrace{\sum L_{-i}(x_i^r)\log L_{-i}(x_i^r)}_{=-H(x_i^r|L_{-i},x_j^o)} - \underbrace{\sum F_j(x_i^r)\log F_j(x_i^r)}_{=-H(x_i^r|x_j^o)} \\ = \;& H(x_i^r|x_j^o) - H(x_i^r|L_{-i},x_j^o) \\ = \;& I(x_i^r; L_{-i}|x_j^o) \end{aligned}$$

where we use the fact that the agents A_j use their observation x_j^o to obtain their predictions $F_j = P(x_i^r|x_j^o)$.

We note that Agent A_i obtains x_i^r by processing an observation x_i^o and use this to apply the following instantiation of the data processing inequality [Cover, 1991]:

$$I(x_i^r; L_{-i}|x_j^o) \leq I(x_i^o; L_{-i}|x_j^o)$$

which holds because further processing cannot increase the mutual information. Choosing to truthfully report $x_i^r = x_i^o$ achieves equality and is thus the best overall strategy.

A recent paper investigated whether the Bayesian Truth Serum could also be used to improve the performance of ensemble learning techniques so that the ensemble obtains the correct result even when the majority of the models in the ensemble are wrong. [Luo and Liu, 2022] shows that training learning agents to also predict the results of other models and applying the BTS mechanism indeed produces significant improvements in accuracy.

Without Prediction Reports. The principle underlying the proof that the BTS information report is truthful can be applied more generally, as shown in [Kong and Schoenebeck, 2019; Kong and Schoenebeck, 2018]. The principle is to construct an unbiased estimator of pairwise mutual information between an agent's report and a report of a peer and then to apply the data processing inequality to show that any reporting strategy that transforms the observed value cannot increase this information measure. An important innovation is that such unbiased estimators can be created using *multi-task* mechanisms, avoiding the need for prediction reports.

Numerous proposals using such principles have appeared recently, for example as incentives for community sensing [Radanovic and Faltings, 2015], for data acquisition from multiple sources [Kong and Schoenebeck, 2018; Chen et al., 2020] or for forecast aggregation [Wang et al., 2021]. These measures in general assume that agents observe a noisy version of a ground truth and therefore are limited to objective data. The principle is not limited to Shannon information, but can be applied to a wide variety of information measures. Even the multitask versions of output agreement [Dasgupta and Ghosh, 2013; Shnayder et al., 2016] can be formulated as a mutual information mechanism [Kong and Schoenebeck, 2019].

Issues With Information-Theoretic Mechanisms. The main advantage of information-theoretic mechanisms is that it is straightforward to show that truthful reporting (up to permutations) is the highest-paying equilibrium. However, this property is only proven in the limit of infinitely many data reports. Witkowski and Parkes [Witkowski and Parkes, 2012] show an example where the BTS mechanism is not truthful for a small number of reports but does become truthful in the limit of infinitely many reports. As attempts to develop a usable finite-sample analysis of truthfulness have not been fruitful to this day, the interest of these mechanisms is mainly their elegant theory.

When prediction reports are allowed, [Witkowski and Parkes, 2012] show an alternative mechanism where the information score is computed through a shadowing mechanism for binary values that has some similarities with the peer prediction method [Miller, Resnick and Zeckhauser, 2005]. [Radanovic and Faltings, 2013] shows an alternative mechanism for more than binary values where the information score is replaced by the peer truth serum [Jurca and Faltings, 2011; Faltings et al., 2017]. Both mechanisms are truthful for small populations of at least 3 reporters.

Recently, a new idea for constructing an information-theoretic agreement measure based on *information determinants* has been proposed [Kong, 2020]. It can be shown to be truthful as long as each agent answers at least twice as many tasks as possible report values. This mechanism looks very promising when this condition can be satisfied.

7 Rewarding Model Quality

A third principle is to reward the extent to which the newly contributed data improves the quality of the resulting model. This has the advantage that it requires few assumptions about the setting and thus broadly applies to subjective data and data used to learn predictive models.

7.1 Eliciting Distributions: Peer Truth Serum

We first consider the peer truth serum mechanism [Jurca and Faltings, 2011; Faltings et al., 2017] for eliciting the distribution of a random variable x, such as reviews of product quality. The center maintains and publishes the distribution R of all prior reports. Agent A_i submits its report x_j as a one-hot vector $\mathbf{x}_j$, and the center updates the distribution as follows:

$$\hat{\mathbf{R}} = (1-\delta)\mathbf{R} + \delta\mathbf{x}_j$$

The accuracy of $\mathbf{R}$ can be evaluated by using a proper scoring rule $SR(x_p, \mathbf{R})$ with a randomly chosen peer report x_p. The contribution of a reported data item x_j can be characterized by the impact of the resulting update on the accuracy of $\hat{\mathbf{R}}$, which is also called the influence:

$$\begin{aligned} I(R, x_j, x_p) &= SR(x_p, \hat{\mathbf{R}}) - SR(x_p, \mathbf{R}) \\ &= SR(x_p, (1-\delta)\mathbf{R} + \delta\mathbf{x}_j) - SR(x_p, \mathbf{R}) \end{aligned}$$

The influence can be approximated using a Taylor expansion w.r.t. δ. Assuming that we use the logarithmic scoring rule:

$$SR(x_p, \mathbf{R}) = \ln r(x_p)$$

the influence is approximated by:

$$I(\mathbf{R}, x_j, x_p) \approx \delta \left(\frac{\mathbf{1}_{x_j = x_p}}{r(x_j)} - 1 \right)$$

See [Faltings and Radanovic, 2017] for the full derivation. By choosing the payment to be proportional to the influence, we incentivize the agent to provide data that is as useful as possible to the center.

When agents submit different data in sequence, their influence adds up to the difference in loss function between the model with and without the data they contributed. The sum of the influences is thus bounded by this quality improvement. This means that the sum of rewards can be bounded and is proportional to the improvement in the model. The peer truth serum is the only known mechanism that can guarantee a bounded budget.

When the distribution $\mathbf{R}$ is publicly available, the mechanism has an issue with uninformative equilibria: the highest-paying equilibrium is for all agents to report the x with the smallest $r(x)$. They will always agree with the peer report, and obtain the highest possible payoff for this agreement. The Peer Truth Serum for Crowdsourcing (PTSC) [Radanovic et al., 2016] eliminates this problem by keeping the distribution $\mathbf{R}$ private, thus eliminating the possibility of colluding on an infrequent value. [Radanovic et al., 2016] shows that all uninformative strategies have an expected reward of zero. The same paper shows in a finite-sample analysis that the properties of PTSC hold even for a small number of reports, thus overcoming the major drawback of information-theoretic measures.

7.2 Eliciting Data for Regression Models

Influence can be defined for any model loss function, and the principle of rewarding according to the improvement of a loss function can apply to any machine learning model. The seminal work of Koh and Liang [Koh and Liang, 2017] shows how to efficiently approximate the influence on an optimal regression model. [Richardson et al., 2020] shows the use of this approximation for an incentive scheme for linear and logistic regression. It assumes that the center maintains a regression model with parameter vector θ, and has a set of test data points z_j for evaluating the influence of contributed data. Using the *Hessian* $H_\theta = 1/n \sum_{i=1}^{n} \nabla_\theta^2 L(x_i, \theta)$, following [Koh and Liang, 2017] we can write the influence of x_i on $L(z_j, \theta)$ (through the optimal $\hat{\theta}$) as:

$$I(x_i, z_j) = -\nabla_\theta L(z_j, \hat{\theta}) H_{\hat{\theta}}^{-1} \nabla_\theta L(x_i, \hat{\theta})$$

The payment for the data can then be proportional to the average influence on the testing loss of a set of randomly chosen test data points. The same Hessian can be used for many data points to avoid repeating the costly matrix inversions.

Another scheme that uses influence to reward data is [Jia et al., 2019], which rewards according to the average influence over all orders of receiving the reports (the Shapley value).

7.3 Issues With Rewarding Model Quality

If an agent chooses its reporting strategy *ex-ante*, i.e. before observing any data, the mechanisms are truthful as additional data improves model quality the most when it is truthful, except in certain pathological cases [Loog et al., 2019].

When an agent chooses its reporting strategy *ex-post*, i.e. having observed the data, the mechanism is truthful whenever the agent believes that its report is the maximum-likelihood estimate given its observation, a condition termed the *self-predicting* condition. This condition is often not satisfied for every particular observation, even if it holds in expectation over many tasks. PTSC is the only known multi-task mechanism where the conditions for ex-post truthfulness can be so precisely characterized.

8 Further Issues And Future Work

Incentivizing Uncertainty And Granularity. The principle of peer consistency is to reward agreement among reported data. If agents can create agreement by focussing on signals that are cheaper to obtain, they could use this to obtain rewards with less effort. [Gao et al., 2019] observe this problem and argue that spot-checking against a ground truth is unavoidable. For settings where spot-checks are feasible, [Goel and Faltings, 2019] shows how a peer-consistency mechanism can amplify their effect so that even just a single spot-check achieves truthfulness in *dominant* strategies, a much stronger guarantee than all other mechanisms we surveyed.

For settings where spot-checking is not possible, cheap signals can be avoided when agents are rewarded for tasks that are more uncertain and for answers with higher resolution and precision. Output agreement mechanisms do very poorly as expected rewards are maximized when reporting on unambiguous tasks with few possible answers. However, as mutual information is upper-bounded by the uncertainty of the random variable, information-theoretic schemes incentivize larger answer spaces and can provide a solution to the cheap signals problem [Kong and Schoenebeck, 2018]. Similarly, schemes based on influence pay higher rewards not only for larger answer spaces but also for more uncertain information. More work on how to ensure reporting of the desired signals would be very useful.

Ex-Ante vs. Ex-Post. Another important issue is how agent beliefs are assumed to influence their decisions. In an *ex-ante* mechanism, agents are assumed to choose a strategy for reporting *before* observing the data, and afterwards consistently execute this strategy. Thus, the truthful strategy maximizes the reward *in expectation over all tasks*, but may not give the highest reward for every particular task.

In an *ex-post* mechanism, the agents can choose their reporting strategy *after* observing the data. This means that truthful reporting has to be optimal for every particular task, a much stronger condition that entails ex-ante truthfulness.

Generally, single-task mechanisms such as Output agreement, Peer Prediction, BTS, RBTS, or PTS are designed as ex-post mechanisms. On the other hand, multi-task mechanisms such as correlated agreement or information-theoretic

measures rely on the assumption that agents use the same reporting strategy over multiple tasks so that it has to be decided ex-ante *before* data is collected. While the strategy of always reporting truthfully has the highest payoff on average, for an individual task the truthful report often does not give the highest reward. Among multitask mechanisms, only the peer truth serum for crowdsourcing has a simple characterization of ex-post truthfulness, the self-predicting condition.

The Importance of Truthfulness. Truthfulness is an important concept in game theory because of the revelation principle [Durlauf and Blume, 2010]. However, in information elicitation, this principle does not apply, as agents are not able to report all their information.

Non-truthful data is problematic only if it leads to inaccurate models. A more realistic objective is *asymptotic accuracy*: no matter what the sequence of observations, the model should eventually converge to be accurate. Under certain conditions, the PTS mechanism can be shown to be asymptotically accurate even if agents sometimes choose non-truthful reports [Faltings et al., 2017]. It would be interesting to better understand the implications of non-truthful reports in other mechanisms.

Fairness. It has been shown that reviews of guests and hosts on the Airbnb platform are biased against certain minorities. Incentive schemes could counteract such bias by treating ratings for different groups of people as different task groups with their own metric [Goel et al., 2020].

Distributed Ledgers And Blockchains. To increase trust in review forums, it would be useful to maintain them on a public ledger such as a blockchain, and rewards could equally be given through a smart contract. The work on Infochain [Goel et al., 2020] was the first to show how the correlated agreement and the peer truth serum mechanisms could be implemented efficiently on the Ethereum blockchain.

A further step would be to decide the outcome of smart contracts through an open poll. Orthos [Moti et al., 2020] is a decentralized mobile application for collecting location-sensitive data that uses the peer truth serum [Radanovic et al., 2016] in a smart contract on the Ethereum network. [Freeman et al., 2017] proposes a scheme where the outcome of a prediction market is decided by a peer prediction scheme. As agents who participated in the prediction market would have an interest in manipulating the outcome decision, it must be assumed that the sets of agents participating in both processes are disjoint. [Goel et al., 2020] analyzes how such outside incentives for manipulating the outcome might be handled when they cannot be avoided. There is a significant literature on strategyproof mechanisms for learning regression models that focusses on incentives to manipulate the model [Cai et al., 2015; Chen et al., 2018]. More work on other scenarios is required, as distributed ledgers offer great possibilities for the application of the game-theoretic mechanisms in this paper.

Federated Learning. In machine learning, there are many cases when multiple parties hold private data that is useful to train a common model, for example, language, medicine, or credit risk. In federated learning, they form a federation and use privacy-preserving techniques to train a common model using all of their data without revealing it. The literature has identified the problem of. free-riding and proposed to use incentives to make it rational for all participants to contribute their data [Yang et al., 2020]. While some works such as [Jia et al., 2019], evaluate the quality of data based on model improvement, others such as [Karimireddy et al., 2022)] make the assumption that all data that is contributed is of high quality, which is unlikely to be true.

An additional challenge in federated learning is that data has to remain private. Incentives. therefore have to be computed in a privacy-preserving manner [Rokvic et al, 2022)].

Personal Data. All mechanisms discussed so far assume that all contributing agents observe data related to the same underlying phenomenon. Therefore, they do not apply to eliciting personal data, such as health, income, or similar characteristics. However, when agents report multiple data items, and they are known to be correlated (such as multiple health measurements), it is possible to use these correlations for incentive schemes [Goel and Faltings, 2019]. Further work on such schemes would open up more applications.

9 Conclusions

The good news to take away from this survey is that for all of the examples outlined in the introduction, there exist game-theoretic mechanisms that provide the proper incentives for eliciting information of high quality.

We summarize the three principles we have seen:

- output/correlated agreement: mechanisms that are easy to implement and understand and work well for objective data. The problem of uninformative equilibria can be avoided in multi-task settings or by scaling rewards.
- information-theoretic measures: they allow simple and elegant proofs of truthfulness for objective data, but suffer from the fact that they are only truthful in the limit of infinitely many tasks per agent. In most cases, they require multi-task mechanisms that assume that the reporting strategy is decided ex-ante.
- model improvement: incentivizes optimal data collection and can be applied to subjective data and machine learning scenarios. Uninformative equilibria or cheap signals carry no reward as the information is not useful. However, truthful reporting is not always optimal; in some cases, a non-truthful report will lead to faster model improvement.

The main weakness that is holding back the field is a lack of empirical evaluation. For many proposed mechanisms there is only a theory but no indication of how well they work with real data, let alone actual live data provisioning. Besides the ESP game [von Ahn and Dabbish, 2004], the PTS mechanism has been used successfully for labeling data [Faltings et al., 2014] and the PTSC mechanism has been shown to improve peer grading [Radanovic et al., 2016]. [Gordon et al., 2020] shows that a peer mechanism [Liu et al., 2020] achieves comparable performance to prediction markets for predicting replicability. Future empirical studies could shed light on the strength of the incentives, whether they encourage fine-grained data, and the predictability of the reward budget.

References

[von Ahn and Dabbish, 2004] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326, 2004.

[Hu et al., 2017] Nan Hu, Paul A. Pavlou and Jennifer Zhang. On Self-Selection Bias in Online Product Reviews. *Management Information Systems Quarterly*, 2017.

[Faltings et al., 2014] Boi Faltings, Jason J. Li, and Radu Jurca. Incentive Mechanisms for Community Sensing. *IEEE Transaction on Computers*, Vol. 63(1), pp. 115-128, 2014.

[Jurca et al., 2007] Radu Jurca, Walter Binder and Boi Faltings. Reliable QoS Monitoring Based on Client Feedback. *Proceedings of the 16th International World Wide Web Conference*, pp. 1003-1012, 2007.

[Geiting and Raftery, 2007] Tilmann Gneiting, and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102**, pp. 359-378, 2007.

[Hanson, 2007] Robin Hanson, Logarithmic market scoring rules for modular combinatorial information aggregation. *Journal of Prediction Markets* **1**(1), pp. 3-15, 2007.

[Miller, Resnick and Zeckhauser, 2005] Nolan Miller, Paul Resnick, and Richard Zeckhauser. Eliciting informative feedback: the peer prediction method. *Management Science*, 2005.

[Kamble et al., 2016] Vijay Kamble, Nihar B. Shah, David Marn, Abhay Parekh, and Kannan Ramachandran. Truth Serums for Massively Crowdsourced Evaluation Tasks, *arXiv:1507.07045*, 2016.

[Jurca and Faltings, 2011] Radu Jurca and Boi Faltings. Incentives for answering hypothetical questions. *1st Workshop on on Social Computing and User Generated Content* ACM Conference on Electronic Commerce, 2011.

[Radanovic et al., 2016] Goran Radanovic, Radu Jurca, and Boi Faltings. Incentives for effort in crowdsourcing using the peer truth serum. *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 7, Issue 4, art. No 48, 2016.

[Jurca and Faltings, 2005] Radu Jurca and Boi Faltings. Enforcing Truthful Strategies in Incentive Compatible Reputation Mechanisms. *Internet and Network Economics*, Springer LNCS **3828** , pp. 268 - 277, 2005.

[Waggoner and Chen, 2014] Bo Waggoner and Yiling Chen. Output Agreement Mechanisms and Common Knowledge. *Second AAAI Conference on Human Computation and Crowdsourcing*, pp. 220-226, 2014.

[Dasgupta and Ghosh, 2013] Anirban Dasgupta, and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. *Proceedings of the 22nd international conference on World Wide Web (WWW'13)*, pp. 319-330, 2013.

[Shnayder et al., 2016] Victor Shnayder, Arpit Agarwal, Rafael Frongillo, and David. C. Parkes. Informed Truthfulness in Multi-Task Peer Prediction. *Proceedings of the 2016 ACM Conference on Economics and Computation*, pp. 179-196, 2016.

[Zhang and Schoenebeck, 2023] Yichi Zhang and Grant Schoenebeck, Multitask Peer Prediction with Task-dependent Strategies. *Proceedings of the ACM Web Conference 2023*, pp. 3436-3446,2023.

[Feng at al., 2022] Shi Feng, Fang-Yi Yu, and Yiling Chen, Peer Prediction for Learning Agents,*Proceedings of the Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[Prelec, 2004] Drazen Prelec. A Bayesian truth serum for subjective data. *Science*, **306**(5695), pp. 462-466, 2004.

[Kong and Schoenebeck, 2019] Yuqing Kong, and Grant Schoenebeck. An information theoretic framework for designing information elicitation mechanisms that reward truth-telling. *ACM Transactions on Economics and Computation (TEAC)* **7**, pp. 1-33, 2019.

[Luo and Liu, 2022] Tianyi Luo and Yang Liu. Machine Truth Serum. *Journal of Machine Learning Research* **112**, pp. 789–815, 2022.

[Cover, 1991] Thomas Cover. Elements of Information Theory. *John Wiley & Sons*, 1991.

[Kong and Schoenebeck, 2018] Yuqing Kong and Grant Schoenebeck. Water from two rocks: Maximizing the mutual information. *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 177-194. 2018.

[Chen et al., 2020] Yiling Chen, Yiheng Shen, and Shuran Zheng. Truthful Data Acquisition via Peer Prediction. *Proceedings of the Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS)*, pp. 18194–18204, December 2020.

[Radanovic and Faltings, 2015] Goran Radanovic and Boi Faltings. Incentive Schemes for Participatory Sensing. *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1081-1089, 2015.

[Wang et al., 2021] Juntao Wang, Yang Liu and Yiling Chen. Forecast Aggregation via Peer Prediction. *Proceedings of The 9th AAAI Conference on Human Computation and Crowdsourcing, HCOMP)*, 2021.

[Witkowski and Parkes, 2012] Jens Witkowski, and David C. Parkes. A robust Bayesian truth serum for small populations. *Proceedings of the 26th AAAI Conference on Artificial Intelligence (AAAI'12)*, pp. 1492-1498, 2012.

[Radanovic and Faltings, 2013] Goran Radanovic and Boi Faltings. A robust Bayesian truth serum for non-binary signals. *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI'13)*, pp. 833-839, 2013.

[Faltings et al., 2017] Boi Faltings, Radu Jurca and Goran Radanovic. Peer Truth Serum: Incentives for Crowdsourcing Measurements and Opinions. *CoRR abs/1704.05269*, 2017.

[Kong, 2020] Yuqing Kong. Dominantly truthful multi-task peer prediction with a constant number of tasks. *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2398-2411, 2020.

[Faltings and Radanovic, 2017] Boi Faltings and Goran Radanovic. Game Theory for Data Science: Eliciting truthful information. *Morgan Kauffman*, 2017.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. *34th International Conference on Machine Learning*, pp. 1885-1894, 2017.

[Richardson et al., 2020] Adam Richardson, Aris Filos-Ratsikas, and Boi Faltings. Budget-Bounded Incentives for Federated Learning. in Q. Yang L. Fan, Yu H (Eds.): *Federated Learning*, pp. 176-188, Springer LNCS 12500, 2020.

[Jia et al., 2019] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. Towards efficient data valuation based on the shapley value. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[Loog et al., 2019] Marco Loog, Tom Viering and Alexander Mey. Minimizers of the empirical risk and risk monotonicity. *Proceedings of the Thirty-third Conference on Neural Information Processing Systems (NeurIPS)*, pp. 7476-7485, 2019.

[Gao et al., 2019] Xi Alice Gao, James R Wright, Kevin Leyton-Brown. Incentivizing evaluation with peer prediction and limited access to ground truth. *Artificial Intelligence* **275**, pp. 618-638, 2019.

[Goel and Faltings, 2019] Naman Goel and Boi Faltings. Deep Bayesian Trust : A Dominant and Fair Incentive Mechanism for Crowd. *Proceeedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 1996-2003, 2019.

[Kong and Schoenebeck, 2018] Yuqing Kong and Grant Schoenebeck. Eliciting expertise without verification. *Proceedings of the ACM Conference on Economics and Computation*, pp. 195–212, 2018.

[Durlauf and Blume, 2010] Steven Durlauf and Lawrence Blume. Revelation Principle. *Game Theory*, pp. 312-318, Palgrave, 2010.

[Goel et al., 2020] Naman Goel, Maxime Rutagarama and Boi Faltings. Tackling Peer to Peer Discrimination in the Sharing Economy. *Proceedings of the 12th ACM Web Science Conference (WebSci 2020)*, pp. 355-361, 2020. (Best paper award).

[Goel et al., 2020] Naman Goel, Cyril van Schreven, Aris Filos-Ratsikas and Boi Faltings. Infochain: A Decentralized, Trustless and Transparent Oracle on Blockchain. *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020)*, pp. 4604-4610, 2020.

[Moti et al., 2020] Moin Hussain Moti, Dimitris Chatzopoulos, Pan Hui, Boi Faltings and Sujit Gujar. Orthos: A Trustworthy AI Framework for Data Acquisition. *Proceedings of the International Workshop on Engineering Multi-Agent Systems*, Springer LNCS 12589, pp. 100-118, 2020.

[Freeman et al., 2017] Rupert Freeman, Sébastien Lahaie, David M. Pennock. Crowdsourced Outcome Determination in Prediction Markets. *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 523-529, 2017.

[Goel et al., 2020] Naman Goel, Aris Filos-Ratsikas and Boi Faltings. Peer-Prediction in the Presence of Outcome Dependent Lying Incentives. *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI 2020)*, pp. 124-131, 2020.

[Cai et al., 2015] Yang Cai, Constantinos Daskalakis and Christos H. Papadimitriou. Optimum statistical estimation with strategic data sources. *Proceedings of The 28th Conference on Learning Theory*, pp. 280-296, 2015.

[Chen et al., 2018] Yiling Chen, Chara Podimata, Ariel Procaccia and Nihar Shah. Strategyproof linear regression in high dimensions. *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp.9-26, 2018.

[Yang et al., 2020] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen and Han Yu. Incentive Mechanism Design for Federated Learning. In: *Federated Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, pp. 95-105, 2020.

[Karimireddy et al., 2022)] Sai Praneeth Karimireddy, Wenshuo Guo, Michael Jordan. Mechanisms that incentivize data sharing in federated learning. *NeurIPS Workshop on Federated Learning*, 2022..

[Rokvic et al, 2022)] Ljubomir Rokvic, Panayiotis Danassis and Boi Faltings. Privacy-Preserving Data Filtering in Federated Learning Using Influence Approximation *NeurIPS Workshop on Federated Learning*, 2022.

[Goel and Faltings, 2019] Naman Goel and Boi Faltings. Personalized Peer Truth Serum for Eliciting Multi-Attribute Personal Data. *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, PMLR **115**, pp. 18-27, 2019.

[Faltings et al., 2014] Boi Faltings, Rady Jurca, Pearl Pu and Bao Duy Tran. Incentives to Counter Bias in Human Computation. *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing*, pp. 59-66, 2014.

[Gordon et al., 2020] Michael Gordon et al. Are Replication Rates the Same across Academic Fields? Community Forecasts from the DARPA SCORE Program. *Royal Society Open Science*, 2020.

[Liu et al., 2020] Yang Liu, Juntao Wang and Yiling Chen. Surrogate Scoring Rules. *ACM Tramsactions on Economics and Computation*, pp. 1-37, October 2022.