Sharing the Winnings of AI with Data Dividends: Challenges with "Meritocratic" Data Valuation

NICHOLAS VINCENT, Simon Fraser University, Canada
BRENT HECHT, Northwestern University, USA

In response to advances in AI and concerns about automation-induced job loss and other downstream economic impacts, there is interest from policymakers and scholars in identifying strategies to more broadly distribute the winnings of AI technologies. One idea gaining prominence is to use data valuation techniques to estimate the value of different data contributions and pay people directly. Despite academic interest and backing from policymakers, there exists little guidance about how selecting an operational definition of data value might influence the outcomes of such initiatives. In this paper, we show that seemingly minor design choices can seriously change the distributions of data values, a serious concern for any human-AI system seeking to incorporate such values for payments or other purposes. We also see that by shifting towards collective notions of data value, we could retain some practical benefits of data valuation while avoiding exacerbating issues with economic inequality.

CCS Concepts: • **Social and professional topics → Computing / technology policy**.

Additional Key Words and Phrases: data dividends, data valuation, economic impact of AI

1 INTRODUCTION

Scholars and the general public have raised concerns that intelligent technologies — i.e., artificial intelligence (AI) and machine learning (ML) — will contribute to economic inequality [9, 10, 12, 35, 46, 47], in particular via job loss from automation. Given that excessive inequality is associated with major societal challenges, including political instability, financial crises [2], reduced national economic growth [53], and even public health harms [13], changes in the distribution of wealth caused by AI may create downstream issues that outweigh the many possible benefits of AI. In response, there have been prominent calls to implement "data dividends": the governor of California directly called for a state-run data dividend [56], related conversations have happened in jurisdictions like West Virginia, Colorado, and Canada [38], and data dividends have begun to attract research attention [5, 16, 60, 61]. Furthermore, a body of academic research on data valuation has emerged, some of which is directly motivated by the use of data values in markets, a "private option" for data dividends.

The data dividend term has been used broadly to refer to giving people a share of the profits from an intelligent technology when their data contributions have been used to train that technology [16, 24, 38, 56]. This concept is inclusive of any program that *retroactively* provides payments or other goods and services to individuals or groups in recompense for their data contributions. Although similar in spirit to crowdwork markets for data (which have issues related to working conditions and pay [3, 21, 48]), data dividends can be used to compensate people for data contributions that have already been incorporated into AI systems. This distinction is important in light of the growth of large "generative AI" models like OpenAI's GPT [8] and Dall-E [45] that rely on content like news articles, research papers, social media posts, open source code contributions, artwork, and more, all of which were generated well before

Manuscript submitted to ACM

the corresponding AI technologies were widespread. Similarly, data dividends are well-suited to share the winnings of search engines and personalized recommendation systems that rely on large-scale behavioral data [39].

The governor of California announced interest in data dividends in a major 2019 policy speech [56]. To quote California's governor, the mandate of a data dividend is to help consumers "share in the wealth that is created from their data" [14]. This mandate could be interpreted as calling for collective value estimation (i.e., answering "what is the aggregate value created by all residents of California?"), but could also be interpreted as calling for very individualistic value estimation (i.e., answering "what is the value created by a particular person?"). The latter meritocratic framing has permeated early discussions.

There is no widely agreed upon definition of data value that addresses this tension between individual and collective value estimation. Prior work looking at particular value definitions, like "leave one out", suggests designers may need to worry about unequal dividend payments [60] and a lack of robustness [62]. Furthermore, we do not know how variations in value definition or estimation procedures might lead to different economic impacts.

In this paper, we explore the impacts of different data value definitions with the goal of informing the increasingly realistic discussion about implementing data dividends. Specifically, we produce data value estimates under a variety of experimental conditions corresponding to designers with different value definitions to show how the downstream data value distributions might change. Our experimental results show how any proposed definition of data value is sensitive to unavoidable subjective design decisions. However, some choices can avoid extremely unequal outcomes, and it is possible to offload decision-making power from the designer to dividend recipients if people can bind together their data contributions, i.e., as part of a "data coalition" [35]. If people group their data contributions together, there are *many fewer* counterfactual outcomes for a data value estimator to simulate, the data value estimation process is cheaper, the resulting dividends are likely to be more equal, and the responsibility for dividend outcomes shifts from a centralized designer to recipients themselves. This point is important in light of increased usage of larger models, such as those used for generative AI systems.

In the Discussion, we reflect on implications for actors interested in data dividends, in particular emphasizing the core challenge for any program aiming to make retroactive payments for data contributions: the need to define data value. Our discussion resonates with critiques [43, 55] of data dividends, but also suggests a path forward: moving towards collective definitions of data value.

2 BACKGROUND AND RELATED WORK

2.1 Background on Data Dividends

Many high profile media outlets have begun to cover data dividends and related concepts (e.g. [1, 38, 44, 54]), and scholars have begun to discuss aspects such as feasibility [61], strategies for implementing a pragmatic data dividend [16], and computational aspects [5]. The data dividend idea can be seen as stemming from work that emphasizes the "data labor" underlying AI [4, 35] and related ideas like considering data subjects as investors [29].

The wave of generative AI-related announcements in 2023 catalyzed more data dividend related discussions. Most notably, OpenAI announced a plan to partner with Shutterstock for training data, with Shutterstock running a "Contributor Fund" to pay data contributors [57]. Adobe has similarly alluded to plans for a "compensation strategy" that may incorporate aspects of data dividends and markets [28]. These developments suggest the possibility of both state-run dividends and firm-run dividends. It is also possible that the public could demand some form of payment by threatening to withhold data going forward, i.e. using "data levers" [59].

Data dividends may benefit from new data-related regulations, though there are challenges in translating changing laws into practical impact (see e.g. challenges with the GDPR in practice [7]). While a key motivator for this study was interest from policymakers, all of our results are broadly relevant *to any scenario in which an organization retroactively disburses funds to individuals or groups on the basis of data value.*

2.2 Critiques of Data Dividends

Several authors have critiqued the data dividend idea on both practical and moral grounds [43, 55]. In a response to Andrew Yang's "Data Dividend Project", Ongweso characterized the proposal as a way to give people a pittance in exchange for their information, and highlighted the concern that a poorly implemented data dividend could cement existing power dynamics without much, or any, positive impact [43]. The piece contrasts fine-grained data dividends with coarse dividends funded by company profits, similar to the approach suggested by Feygin and colleagues [16]. Tsukayama, writing for the Electronic Frontier Foundation, raised similar concerns [55] and highlighted that certain dividend design choices could create perverse incentives that coerce economically vulnerable people to share data excessively, at risk to their own well-being. As such, there are reasons to broadly oppose many implementations of data dividends, especially programs that pay small amounts of money and do little to change power dynamics.

Opposing data dividends altogether or supporting a simple tax-and-fund-public-goods [16] approach still involves adopting an operational definition of data value. Dividing payments equally between all members of some group involves assigning positive data values to members of the recipient group and zero data value to those who are not members. For instance, if California were to tax AI revenue to disburse equally between California residents, this would involve producing a mapping of data values in which California residents have a value of 1 and everyone else has a value of 0. And taking no action can be seen as assigning zero data value to historical contributions. In this scenario, contributors only see a share of AI winnings if they flow (or trickle) back in the form of new technologies or future market activity.

There is an additional implicit argument against data dividends worth spelling out: that people already receive utility from data-dependent systems in the form of system capabilities. Better search, recommendations, generative AI systems, and more could be seen as a kind of dividend, when services like Google search, TikTok's recommendations, and ChatGPT are (currently) offered for free. However, even if we consider services as part of a data dividend, there is no guarantee the public or the state would consider this to be a "good deal". Given that the data valuation tasks we study are actually indifferent to how a dividend is disbursed (e.g., cash, goods, services), data dividend designers could take into account the value of free services and then provide additional payments or goods.

2.3 Data Dividends and Markets for Digital Labor

Crowdwork marketplaces such as Mechanical Turk [3, 21, 23, 27, 30] already enable people to be compensated for digital labor. Here, we discuss how data dividends differ from such marketplaces, but can be complementary in achieving pro-social outcomes. In short, while markets for data can refer to platforms that allow either the purchase of entire datasets or individual contributions (like paying a gig worker to label an image), data dividends can refer broadly to any kind of retroactive payments for data.

In crowdwork marketplaces, workers are paid based on discrete tasks with definitive starting and ending points (see [3] for extensive comparison of crowdwork to "piecework"). This approach does not translate well to compensating people for the data they produce passively during activities like using search engines and recommender systems, i.e. kinds of data often called implicit feedback [39]. These technologies rely on behavioral data, e.g. clicks and views. Such

data is valuable under the assumption that they represent in-the-wild preferences. Thus, a developer cannot pay directly for 100 units of search logs in the way they could pay for 100 units of labeled images. However, they could pay for access to a year's worth of "natural" behavioral logs (though behavioral logs generated in a world with payments will differ from those generated in a world without). This style of payment would be a privately-implemented data dividend.

Furthermore, existing crowdwork marketplaces are not well-suited to rewarding historical contributions. If computing-induced inequality increases dramatically, simply distributing earnings from future improvements to technologies will likely be insufficient to satiate economic and political demands.

Current crowdwork marketplaces have been subject to a broad range of critiques related to their ability to address economic inequality (see e.g. [21]). Without careful consideration of the work which has critiqued the crowdwork paradigm, data dividends may fall into similar patterns, e.g. excessively low payments. Our expectation is that crowdwork could be an important part of broadly distributing the profits from intelligent technologies.

2.4 Data Valuation

A central challenge for data dividends is identifying the value of a given unit or group of data in a particular context. Our work intersects with data valuation literature at multiple levels: (1) measuring the impact of data on firm revenues, and (2) measuring how an observation (or group of observations) impacts a machine learning model's performance.

2.4.1 The Aggregate Impact of Data on Firm Revenue. Many discussions around the value of data have considered estimating the aggregate monetary value of data to companies, which we might call "company-level data valuation" [18, 33, 36]. This generally involves trying to attribute some portion of firm revenue as "data revenue" (i.e., answering a counterfactual question about the impact of data on revenue). For instance, in discussing the results of their company-level data valuation study, Shapiro and Anejo called for 50% of estimated "data revenue" to be paid directly to Americans or into infrastructure [50].

Estimating data revenue is likely to be an important step in determining the funding pool for a dividend. If a data dividend is implemented using a very simple "just fund public goods" approach, this high-level data valuation may be the only value estimation needed. In early dividend implementations, this might involve very broad strokes, e.g. taxing 1% of the revenue of regulator-selected firms. In the future, more precise estimates of the causal impact of data on revenue may be possible, especially if firms are provided with incentives to participate in the estimation.

Looking forward, identifying the value of technologies not directly connected to revenue (e.g. services like AI assistants, search, image tagging, etc.) and public datasets on which for-profit intelligent technologies rely (e.g., Wikipedia [58]) will be important for a more holistic perspective on aggregate data valuation.

2.4.2 Observation-level Impact of Data on Model Capabilities. Discussions of paying people for data have suggested (often implicitly) that payments should be made to individuals, in accordance with how much somebody impacted a data-dependent technology. Any such meritocratic approach to data payments requires observation-level data valuation – a way to measure the value of specific observations to a ML model, which is an open challenge that has been explored by recent ML research [17, 26, 32, 34, 51].

Here, we discuss several specific techniques and lines of work that were particularly influential on our experiments. We refer interested readers to Hammoudeh and Lowd's full survey of training data influence techniques for an extensive discussion of factors like complexity and the assumptions used by different techniques [20].

Koh and Liang's influence function-based approach [32] operationalized data valuation by estimating how model loss changes when an observation is removed. A key idea was that a data point's influence could be calculated without

actual retraining. Following the attention to influence functions, rapid progress was made in developing data valuation techniques that use the *Shapley value* from cooperative game theory, an approach motivated in part by compelling theoretical properties [17, 26, 34]. For instance, under certain conditions, the players being assigned Shapley values (i.e., data contributors) can be sure that two observations that have the same effect on performance will be given the same value. The Beta Shapley approach [34] adds further flexibility to Shapley value based approaches, and offers unique advantages for data dividend simulations. We describe this in greater detail below in our Methods section. A later approach, the Average Marginal Effect, can be seen as inclusive of both Data Shapley and Beta Shapley [37].

Researchers have also proposed other alternative definitions of data value (see also Sim et al.'s survey [51]). For instance, the least core approach from Yan and Procaccia seeks to minimize the difference between the value and payoff of any given coalition [64] and specifically incentivizes contributors to not opt out. However, this concern is probably more relevant to markets or long-running data dividend programs than new pilot programs. Such programs will likely move beyond relatively static experiments we explore here into the domain of mechanism design.

The "datamodels" approach provides a way to directly learn about "data counterfactuals" for specific test examples, i.e. to predict the output of training a model with some training set and a specific test example [25].

All these definitions are likely to be relevant to long-term data dividend discussions, but here we focus specifically on the Beta Shapley approach for its flexibility – as we are especially interested in the role of designer choice – and because it is practical to run experiments with small batches of data.

2.4.3 Observation-level Valuation with Web-scale data. There is a major challenge at play in using observation-level data valuation with modern systems: in general, amassing more data (assuming data quality remains steady) makes systems more useful, but it also makes estimating data value under most definitions more expensive. This is because directly calculating most notions of data value for large datasets generally involves simulating more counterfactual worlds (computing the exact leave-one-out value requires re-training a model n times, and computing the exact Shapley value requires re-training 2^n times to cover every entry in the powerset of all observations). Early estimation techniques were on the order of $O(Nlog(log(N)))$ [26], though faster variants now exist [37].

For economically generative domains like search and personalization (see e.g. revenues from Alphabet and Meta) with very large training datasets (typically web-scale public datasets or unknowable large private datasets), it is particularly challenging to apply fine-grained data value estimation. For very small toy models that are amenable to very quick ML experiments, is it easy to apply data value estimation. As we will see below, our results suggest a fortuitous fact for the equality-focused designer: performing valuation at the coalition-level can make valuation more tractable (by passing off some of the allocation work to coalitions themselves) *and* lead to fairly equal data dividend outcomes that are less sensitive to seemingly arbitrary design choices.

A potential concern with placing major stock into the results of studies that use the data valuation techniques described above is that most of that body of work focuses on traditional classification problems, and these results might not apply to larger generative AI systems like LLMs that are seeing rapid growth in research attention and funding.

However, there is some some early work from LLM developer and operator Anthropic AI, applying influence functions to LLMs [19]. Critically, this work suggests that training data influence for web-scale data may be viable, drastically amplifying the potential impact of work at the intersection of valuation and payments.

This work looked closely at the concentration of influence values, i.e. how the influence score assigned to training sequences are distributed. The authors observed a power law distribution of influences in the tail. Sequences of training data with large impacts were rare (but account for a large amount of total influence). Notably, the authors also find that

"the examples we have investigated were learned from the collective contributions of many training examples rather than being attributable to just one or a handful of training examples", a point that is highly relevant to our work.

Although this work did not share any of the influence data directly such that we could compare our results directly (or e.g. use the influence estimates directly to simulate data dividends), the qualitative trends are generally consistent with the findings we will discuss below.

2.5 Arguments for and Against Fine-Grained Valuation

It seems that the primary argument for using fine-grained value definition (e.g., at the individual level) is that data dividends will create new incentives around the generation and sharing of data. As a motivating example, a program that offers to pay some fixed amount of money per search engine query, restaurant review, or piece of art posted to the Internet would likely lead to some degree of spam. Even a basic attempt to distinguish real contributions from spam would constitute an assignment of data values. In other words, a data dividend implementer is performing a kind of data value estimation just by choosing which data points to take into consideration.

Once a data dividend has been implemented, it would permanently change the incentives around data generation; presumably some people people will change their behavior in anticipation of future disbursements. It may be the case that fine-grained dividends are perceived as more fair, but we leave this question to future work that directly engages with recipients. We refer here to fine-grained data valuation as meritocratic in a loose sense; see e.g. [31] for an extended analysis of the concept of meritocracy in a computational setting.

Moving to arguments *against* fine-grained valuation, early work suggests that fine-grained valuation can lead to unequal outcomes [60]. A poorly designed data dividend could act as primarily a handout to the very well off. Further, an unequal dividend is likely to have inequalities along demographic lines, as seen in data valuation work [17, 60]. Data valuation can be quite stochastic, which may make dividend payments seem capricious. This is an issue Wang and Jia aim to address with the "Data Banzhaf" technique [62].

Generalizing from work on learning curves and data scaling (see e.g. [22]), we can conjecture that as any given dataset's size increases, the average impact of a randomly selected observation should continue to decrease. This leads to another argument against valuation-based dividends: if a technology is reliant on the collective contributions of millions or billions of people, we already know each individual value will be very small, so why bother spending time and energy performing data value estimation? Whether or not value estimation is worth the compute cost may indeed depend heavily on the scale of a given dataset.

As we will discuss further below, concerns about value estimation cost can be migitated, though whether pursuing valuation is worth it will always be a value call.

3 METHODS

3.1 Motivating our Simulations

A key application of the Beta Shapley and similar approaches is to identify data that is mislabeled, or particularly valuable to include in a training set. It is also useful for trying connecting model outputs to inputs.

These data value estimates provide some signal about data's impact, such that tying payments to data values *may* create positive sum incentives around data creation. Still, data values will be highly dependent on evaluation procedures and test set selection.

3.2 Simulation Scenario

Earlier work on data dividends highlighted many choices an implementer of data dividends will face [60]. The implementer could be a government agency (e.g., the state of California) or a private firm. This entity must have a funding source and must select some inclusion criteria for which kinds of data they will be evaluating, such as a list of tasks and corresponding datasets, or a single task and dataset.

Here, we assume that the implementer has already decided to set aside a fixed pool of money to distribute (or spend, as dividends need not be direct paychecks – the state could allocate spending on public goods in accordance with group data values). As noted above, a key reason to believe this amount will not be a complete pittance is the implied threat that the public could withhold data going forward [59]. However, the question of what share of total profits should actually go towards data dividends remains an open question.

We imagine a data dividend implementer who selects a single classification task with one representative contribution per person. While studying a single classification model might seem distant from large language models or search engines, we imagine that each observation represents a summary of a person's contributions to a dataset. Furthermore, we can imagine a language model or search engine being evaluated in a classification setting. More complex and nuanced evaluation procedures are critical for practical use, but not necessarily helpful if we want a one-dimensional summary of data value.

Many datasets have highly unequal contribution patterns. However, we do not focus on these. Differences between individual influence values are small[1], so for datasets with highly unequal contributions patterns the inequality in contributions patterns will likely dominate the inequality in dividends.

This means data valuation in contexts with very unequal distributions of number of records per contributor will already be very unequal; we do not need simulations to confirm this. Thus, our simulations focus on studying the potential for large differences in value distribution even in contexts in which every person has roughly similar opportunities to contribute data.

Returning to our scenario's assumptions, we assume the implementer will use some variant of Beta Shapley value for value estimation. The Beta Shapley value [34] builds on early work on "Data Shapley" [5, 17, 26], an approach that considers the impact of a given data point (or group of points) relative to all possible combinations of other points. In practice, this means sampling many possible combinations of data points (of different sizes) to estimate impact. The Beta Shapley concept takes this a step further, and proposes that it might be useful to consider weighting certain sized combinations differently.

Beta Shapley allows a developer performing data value estimation to choose if they prefer to emphasize the impact of an observation (or group of observations) relative to a small (low cardinality) or a large (high cardinality) dataset. For instance, we can think of the difference between low and high cardinality data value estimation as the difference between asking "What would happen when adding an observation to a training set of size 10" and "What would happen when adding an observation to a training set of size 1000". In the first case, accuracy is likely low and the observation may have a large positive or negative effect. In the second case, accuracy is likely higher and already fairly stable.

Beta Shapley can replicate the original Data Shapley approach (assign weights based on likelihood of drawing a particular coalition randomly) and leave one out, so this technique is inclusive of other valuation techniques (though distinct from other value definitions like least core [64]). This flexibility, combined with the desirable theoretical properties, provides a boost in ecological validity to experiments using the technique.

[1]Formally, for a learning algorithm that is uniformly stable, "uniform value division produces a fairly good approximation to the true Shapley Value" [26].

The results from the Beta Shapely paper suggest that low cardinality data values are very helpful for recovering useful signal about the likelihood an observation is mislabeled, or the value of including a point in a subsample [34], and so can help address the issue that marginal impacts will trend closer and closer to zero as data size increases.

It is still expensive to estimate data values for many data points at a time (for instance, in [63], 1000 data points is used as a cutoff at which Beta Shapley is impractical). For our purposes, however, this limitation is not blocking, because Beta Shapley values can be acquired for just a small batch of data at a time. Indeed, in the Beta Shapley paper, the authors sample just 200 training points, 200 validation points, and 1000 test points from the 580k data point Covertype dataset [6] and are able to detect noisy data. One could separate a large dataset into small batches of 100-200 points to get data value estimates for each data point in a computationally feasible fashion. These cheap estimates should be generally correlated with estimates generated more expensively using larger batch sizes. To confirm this intuition, we validate that Beta Shapley values generated using different batch sizes are highly correlated.

The idea that low cardinality data values are useful in practice is important for making our experiments viable. If a low cardinality approach provided no signal at all, it would likely be impractical to use data valuation for any large datasets, and this might preclude the use of fine-grained value estimation for data dividends altogether. For instance, it is quite expensive to use machine learning re-training experiments to see how GPT-3 [8] would behave if a single individuals' social media posts were removed from the training data, but this does not mean it is impossible to estimate the value provided by some data contributor relative to some representative task.

3.2.1 Varying Valuation Cardinality. We first investigate two variables – "valuation cardinality" and batch sizes – that impact which counterfactual data scenarios value estimation explores. In other words, different choices in cardinality and batch size create different operational definitons of data value that map to different counterfactual scenarios.

By adjusting the distribution that provides the sampling weights for Beta Shapley estimation, a designer can effectively choose how much to weigh low cardinality contribution (i.e. the impact a data point has when it is added to a small number of other points) versus high cardinality contribution (i.e. the impact a data point has when added to a large number of other points).

We consider a low cardinality approach (using Beta(16,1) for weights), a medium cardinality Data Shapley (Beta(1,1)) and a high cardinality approach (Beta(1,16)). See Fig. 3 from [34] for visual intuition as to how different cardinalities weight different dataset sizes. For our purposes here, it may be helpful to think of low cardinality as "add a point to 25 other points" and high cardinality as "add a point to 175 other points".

We address our computational challenges by working with batches of data. We can sample just 200 data points and evaluate the impact of each point relative to this small sample. The exact batch size (e.g., 100, 200, 500) impacts the counterfactual scenarios that data valuation explores: for instance, for batches of size 200, high cardinality value estimates will at most consider the interactions between 200 observations. There is some dataset size at which valuation experiments become impractical (for instance, estimating an individual person's impact on a large language model is very unlikely to be performed in practice). As such, we view the batching approach as a reasonable limitation, and a choice an implementer would make for pragmatic reasons.

3.2.2 Varying Dataset Processing. Designers will need to choose which datasets and tasks are used for valuation, and might even choose to make certain data processing choices before beginning value estimation. For instance, they might binarize some variables in a dataset or remove uninformative features to lower the computational cost of value estimation.

Using synthetic data means that we can explore dataset attributes in a systematic manner. Specifically, we see how varying the number of classes in a classification task and changing the test set size impact data value distributions.

Multi-class problems have potential for overall greater information content (typically, binarizing labels loses information). In other words, multi-class problems can map to harder tasks. Using the lens of extreme classification, we can even view tasks like user-specific personalization as special cases of very-many-class problems [41], so understanding trends in data value distributions as number of classes grows can provide some insight into how data values might look for highly complex tasks. Of course, future work might study many other types of variations in dataset characteristics. Varying the number of classes of a synthetic dataset is one approach for starting to study dataset variations.

3.2.3 Data Coalitions. Finally, we investigate how dividend outcomes might change if people form "data coalitions", and how data values are assigned to groups of data points rather than individual data points. We can think of data coalitions as a way for people to "bind together" their data contributions: the data user (e.g., a tech company) will either have all or none of the observations from any given coalition.

Computationally, this is convenient for data value estimation experiments, as it reduces the number of counterfactuals that we must explore. In the most extreme example where every contributor joins one data coalition, we only need to train two models: one with "full data" and one with "no data". If there are just two coalitions, we need only four training runs (full data, no data, first coalition omitted, second coalition omitted), and so on (with the exact number of training runs needed given by 2^c for c coalitions).

3.2.4 Dataset and Model. For our experiments, we generate synthetic classification data. We use sklearn's *make_classification* functionality[2]. We use the default parameter choices, except we reduce the number of features from 20 to 5 to speed up experiments. This software generates synthetic data that is useful for testing different classification approaches. We focus on a binary classification task (except in specific experiments that involve adding more classes). We show in the Appendix how non-synthetic data display similar trends. For most of our experiments, we run five repetitions with a different sample of 1000 points separated into five size 200 batches (i.e., 5000 data points per experiment).

As in [34], we use logistic regression to perform classification on this synthetic data and produce data values based on estimated counterfactual changes in model accuracy. Then, for each experiment, we inspect the raw distribution of data value estimates outputted by the Beta Shapley technique (building on the code from [34]). These raw data values would, in practice, be converted into actual payments (or correspond to the value of services different individuals or groups receive). See the Appendix for an example.

4 RESULTS

First, we provide an important preliminary result: validation that data values estimated in small batches correlate to values from large batches. Then we describe the data value estimate distributions observed in our experiments.

4.1 Data Values with Small Batches

We first verified that Beta Shapley values computed using small samples correlate to values computed using large samples. We generate a dataset with 1000 total observations, and chunked the dataset five ways – ten chunks of 100, five chunks of 200, 2 chunks of 500, or one chunk of 1000 – in order to perform value estimation and check if the values

[2]The software is documented here: https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html.

produced with each chunk size are correlated. Based on prior Beta Shapley results, we expected this approach to work, but by testing this we add further confidence that our experiments are ecologically valid.

In Figure 1, we see fairly robust correlations for the lower cardinality Beta Shapley approaches. As we approach higher cardinality, these correlations are smaller.

4.2 Value Estimate Distributions

Here, we examine of impact of design choices on the distribution of data value estimates. These data values were generated using the Beta Shapley technique with accuracy as the evaluation metric of interest.

4.2.1 Cardinality. Figure 2 shows different data value estimate distributions that arise from using different Beta distributions to produce weights for the Beta Shapley weighted average impact calculations.

The key takeaway here is that low cardinalities provide a much wider range of data values. For instance, the "most helpful" (highest score) point with the low cardinality Beta(16,1) approach has a data value of 7.34, i.e. the weighted average impact on accuracy is 7 points. However, with a high cardinality Beta (1,16) approach, the highest value is 0.7, an order of magnitude smaller. Higher cardinalities also reduce the median and standard deviation of the data values, i.e. the distribution becomes tighter with more mass close to zero. The percentage of negative observations also becomes higher, which can be relevant when transforming value to money.

We can interpret these results as illustrating how lower cardinality value definitions – which emphasize impact when data is added to a small dataset – allow any given observation to have more overall impact. This tracks with the

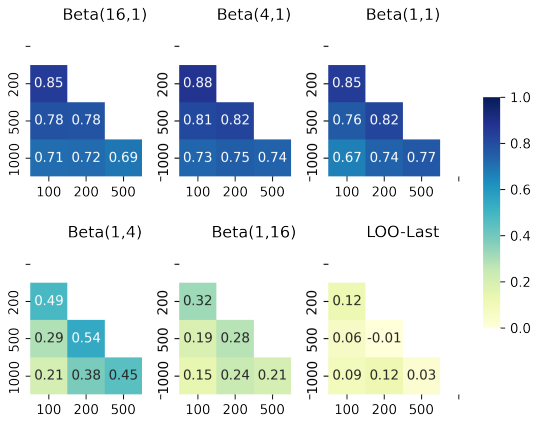

Fig. 1. Plot showing that values produced at different cardinalities are correlated. Specifically, shows pairwise correlations of Beta Shapley values computed using batches of size 100, 200, 500, and 1000. For instance, the upper left square shows that using Beta(16,1), and producing values in batches of 100 and 200 produces two sets of values that have 0.85 correlation. Top left is low cardinality; bottom right is high cardinality. Data values produced with different batch sizes are correlated, but cardinality and batch size matter.

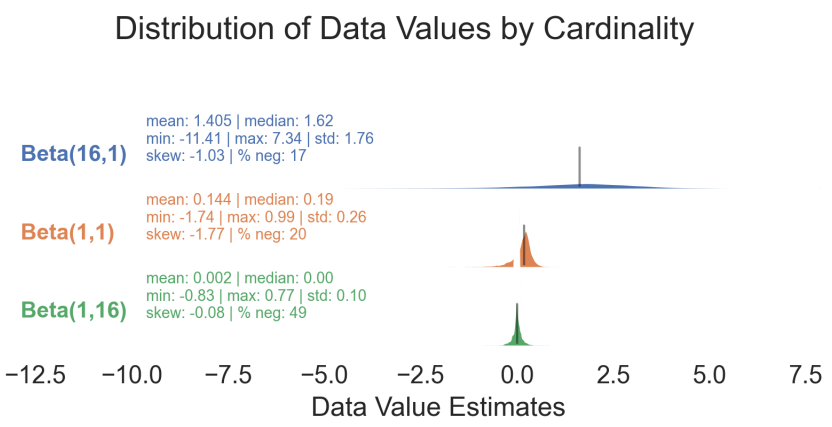

Fig. 2. Density estimation plots for data values. Each row shows values obtained using a different "valuation cardinality". From top to bottom, cardinality increases (topmost row is the lowest cardinality, which emphasizes impact of adding an observation to a small dataset). Each plot is annotated with descriptive statistics: the mean, median (black vertical line), minimum, maximum, standard deviation, skew, and percent of observations with negative values.

diminishing returns curves that pervade data scaling research [22]. As dataset size grows, the expected impact from adding a new observation falls.

In the Appendix, we show similar results that involve varying the batch size. In short, increasing batch size has a similar effect to increasing cardinality (as we might expect given the idea that batch size acts as a "ceiling" on cardinality).

4.2.2 Dataset Processing. We also investigated choices related to the dataset being appraised. We examined the role of the number of classes included in the dataset being appraised and test dataset size.

Figure 3 shows value estimates when varying the number of classes in the classification dataset. We see that the impact of class number on the value estimate distribution depends on value cardinality.

With low cardinality emphasizing small training datasets, the impact of any one observation is smaller when we include more classes. For instance, for the 16 class dataset, the max data value is 2.90 versus 7.34 for the two class case. When a high cardinality emphasis is used instead, this trend flips and including more classes causes the value distribution to spread out relative to the few-classes dataset (though the minimum and maximum impacts in the 16 class high cardinality case still have smaller magnitudes than in the 16 class low cardinality case).

This result is important because it shows potential interactions between design choices. Above, we saw cardinality heavily impacts value distributions; here we see cardinality choices will have further interactions with choices like whether to binarize labels in a dataset before computing value estimates.

The data value estimate distributions were very similar across test set size. In Section 5, we discuss how curating a test set might impact dividends and potential for future work to explore "restorative" dividends.

4.2.3 Data Coalitions. Finally, we examine data value definitions that emphasize the possibility of *data coalitions*. We are interested in scenarios in which data contributors bind their contributions together, so all observations for the coalition must be given a single value.

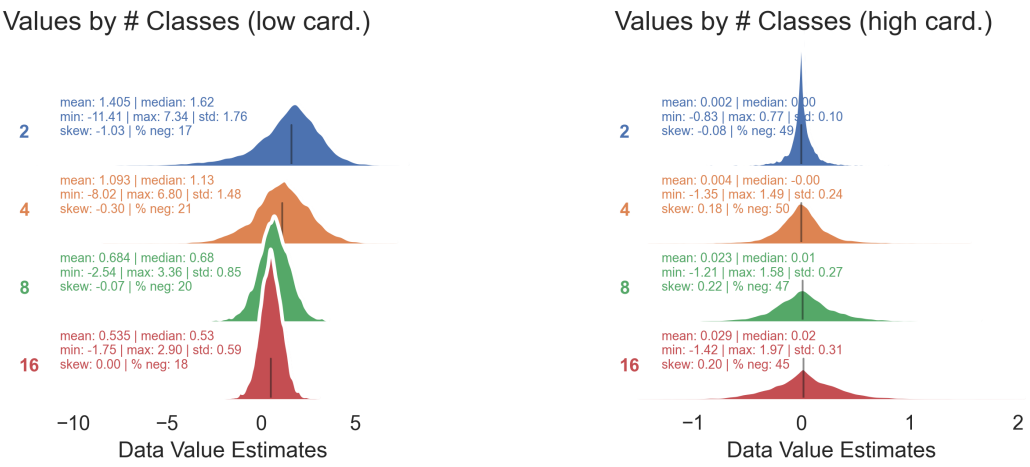

Fig. 3. Impact of number of classes on data values. Follows the structure of Fig. 2.

When we estimate values at the coalition level, we consider a smaller number of total data values. The experiments are correspondingly much cheaper to run (recall: if all members split into two coalitions, we only need to produce two data values). Thus, we can include 50 repetitions instead of five. Additionally, we focus on classic leave-one-out data values, because the number of data values per experiment is very low (the Beta Shapley makes sense when we want to choose between emphasizing combinations of 10-25 data points or 175-200 data points, but with coalitions we could just have two or three total coalitions).

In general, we see that splitting data between more coalitions (such that each coalition is smaller on average) creates tighter distributions with smaller overall data values. The standard deviation falls, more data values are negative, and the maximum impact of any one coalition falls as we allow for more coalitions to exist.

4.3 Impacts on Payments

In the Appendix, we provide a case study where we assume payments based on data values, averaging $2000 USD, are sent to U.S. households. We see that design choices can lead to large shifts in both the variance of actual payments and how regressive the outcomes are in terms of income inequality. This case study relies on many assumptions however, so we focus the discussion here on interpreting the differences in data value distributions.

5 DISCUSSION

In line with prior work, our results reinforce how easily seemingly innocuous design choices can impact the outcomes of data dividends. We showed numerous data valuation choices can impact results (e.g., cardinality, batch size, how classes are encoded). Anyone implementing a data dividend should consider conducting similar simulation experiments and sharing the results of these simulations with recipients of the data dividend. If no attempts to include recipients in the design and governance of a data dividend program are made, there is always a possibility that data contributors could withhold their data contributions to attempt to gain leverage [59].

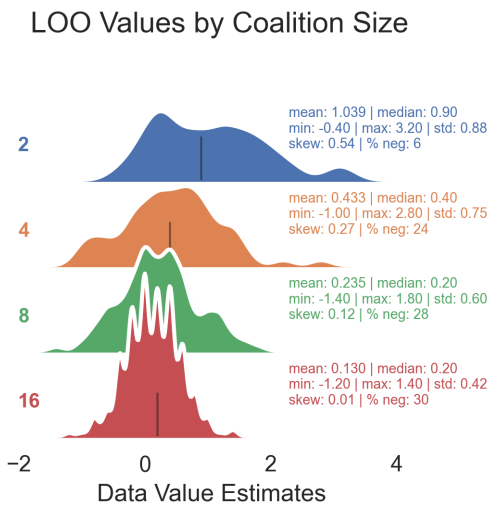

Fig. 4. Impact of "data coalitions" on data values. Follows the structure of Fig. 2. From top to bottom, number of coalitions increase, and thus the coalition size decreases. These value estimates were produced using a high cardinality approach.

5.1 Against individual-level data value

Given these results and the work they build on, should activists begin pushing for data value-based dividends now?

Individualistic dividends may be counterproductive, but as we saw above, dividends can be implemented with coarse data valuation. For instance, a government could use simple heuristics to tax data-dependent firms (e.g., a tax on 1% of revenue from some heuristically determined set of "data-dependent" firms [16]) and split this equally between all recipients or just fund redistributive initiatives. Such a scheme may actually look quite similar to data dividend alternatives suggested by critics of individualistic state-run dividends [43].

In terms of design implications for a data dividend that could be implemented tomorrow, it seems that focusing on coarse data valuation is a good first course of action. If fine-grained data valuation is preferred, using a low cardinality data value definition could help to avoid very concentrated payments, and it will likely be worth paying close attention to other choices related to selecting the specific data used for valuation.

5.2 Practical benefits of group-level data value

As noted above, our results suggest that a particularly effective way to navigate the tension around data valuation in data dividends is to support data coalitions and perform valuation at the collective level. Data coalitions give creators a say in how fine-grained or coarse data valuation is: a small set of large coalitions will lead to coarse data valuation, whereas a large set of small coalitions will lead to fine-grained valuation. Furthermore, these coalitions could aggregate and express preferences about data dividend design parameters like valuation cardinality, batch size, and dataset processing.

An additional benefit is that coalitions provide a way for data dividends to benefit from work on online governance [49, 65]. For instance, coalitions could perform their own data valuation and set norms around "high quality" data

production, using new tools [65] to engage in voting and deliberation and set formal policies about data values. In the longer term, formal legal support for data coalitions could solidify this source of power for data creators.

Interestingly, the actors tasked with implementing data valuation (e.g., computer scientists) may have strong reasons to support data coalitions: When we have just a few large coalitions, there are very few counterfactual scenarios to explore, so grouped data values can be estimated very quickly. For instance, if everyone in California were to join one of ten data coalitions, the agency running a data dividend would only need to produce ten data value estimates (requiring only ten experiments for a leave-one-out approach, and 1024 experiments for an exact Shapley computation).

5.2.1 Hierarchical data values. The prospect of assigning value to large groups creates a potential recursive problem. Once a coalition receives a large data dividend, they might want to allocate it between members or sub-coalitions, and a given sub-coalition might have its own subset, and so on.

On one hand, this represents an exciting new set of problems for understanding the theory of data values in a complex game theoretic setting. Taking into consideration the interplay of sub-coalitions may provide more nuanced data value definitions. Exploring this avenue will be of great practical interest, as in the real world we rarely have access to the true distribution of some variable – we just have different groups providing some limited subset.

There is a potential problem as well – will designers just push computational costs saved from coalition-level valuation onto downstream organizations? This could happen to some degree, but in general any efforts to bind units of data together makes the overall counterfactaul space smaller (with the caveat that some of the combinations we eliminate may be informative).

5.3 Restorative Data Dividends

A distinctive advantage of data dividends is that the designer (ideally, in cooperation with recipients) can use knowledge about how data value estimates tend to be distributed (from simulations like ours, from theory, etc.) to design a data dividend scheme that achieves a specific outcome.

An idea ripe for future work is that designers could also influence the outcomes of data dividends through careful test set selection. This could open the door to data dividends that use either individual level valuation or a middle ground collective approach (e.g., 10-20 distinct groups) to implement data dividends that are restorative in nature. For instance, this could involve constructing a test set that emphasizes performance for groups that are historically and/or currently disadvantaged – something that would be useful for purposes far beyond data dividends.

There is great synergy between data-centric algorithmic fairness interventions [15] and data dividends; if we focus on building test datasets to address gaps in algorithmic performance, corresponding dividends are likely to be restorative.

5.4 Opportunities for Human-Centered Evaluation

In this study, we focused on studying recent data valuation techniques and the data value estimate distributions they produce. As data dividends-type programs are rolled out in a variety of ways (state-run programs, revamped corporate rewards programs, etc.) there will be immense value in conducting human-centered studies of the effects of data dividends. How do data dividends change user behavior, if at all? How do people feel about them? For the most part, this line of work will likely involve drawing on a mix core HCI methods such as interviews, field studies, surveys, and observational analyses [42]. As a reflective note, the HCI field is well-positioned to run experiments that involve our own small scale "data dividends", i.e. retroactive payments for participants who have contributed to research studies.

5.5 Limitations and Future Work

Our simulations tell us about what small data dividends pilots might look like. We made assumptions about what techniques data dividend designers might explore, but did not cover all potential possibilities.

Our simulations were highly standardized. Our test sets were true random samples and our coalitions were all the same size. When dealing with messier data, there are additional choices available to a machine learning practitioner that could introduce even more variance into resulting data value distributions. We leave these questions to future work.

As data valuation techniques are improved and new definitions and approaches are proposed, the design space could become even more complicated. Conversely, if the ML community coalesces around a very specific approach to data value, the design space could shrink.

Another critical question for future work will be how recipients perceive and respond to data dividends. The mandate of data dividends is sharing wealth with the people who help fuel data-dependent technologies, so giving contributors a voice should be a top priority for dividend design. Simulation work cannot answer these kinds of questions, although our simulation results could be used to develop tools for people to explore dividend outcomes. Ultimately, it may be the case that people do not want monetary data dividends. In this case, these results and discussion points will still remain relevant to other scenarios in which data value estimates might be used (e.g., data markets, auctions, exchanges).

6 CONCLUSION

In this paper, we used simulation experiments to study the role of data valuation in the design of data dividends. Our results highlight that "meritocratic" data valuation requires making a series of highly discretionary design choices. These choices can create dividends that are highly unequal, potentially causing the dividends to fail to produce progressive outcomes. We conclude with recommendations for designing dividends that mitigate this concern, by shifting towards estimating data value for groups, not individuals.

REFERENCES

[1] 2018. Should internet firms pay for the data users currently give away? *The Economist* (2018). https://www.economist.com/news/finance-and-economics/21734390-and-new-paper-proposes-should-data-providers-unionise-should-internet

[2] Alberto Alesina and Roberto Perotti. 1996. Income distribution, political instability, and investment. *European economic review* 40, 6 (1996), 1203–1228.

[3] Ali Alkhatib, Michael S. Bernstein, and Margaret Levi. 2017. Examining Crowd Work and Gig Work Through The Historical Lens of Piecework. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4599–4616. https://doi.org/10.1145/3025453.3025974

[4] Imanol Arrieta-Ibarra, Leonard Goff, Diego Jiménez-Hernández, Jaron Lanier, and E. Glen Weyl. 2018. Should We Treat Data as Labor? Moving beyond "Free". *AEA Papers and Proceedings* 108 (May 2018), 38–42. https://doi.org/10.1257/pandp.20181003

[5] Eric Bax. 2019. Computing a Data Dividend. https://doi.org/10.48550/arXiv.1905.01805 arXiv:1905.01805 [cs, econ, q-fin, stat].

[6] Jock A Blackard and Denis J Dean. 1999. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture* 24, 3 (1999), 131–151. Publisher: Elsevier.

[7] Alex Bowyer, Jack Holt, Josephine Go Jefferies, Rob Wilson, David Kirk, and Jan David Smeddinck. 2022. Human-GDPR Interaction: Practical Experiences of Accessing Personal Data. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. https://doi.org/10.1145/3491102.3501947

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165 arXiv:2005.14165 [cs].

[9] Erik Brynjolfsson and Andrew McAfee. 2014. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company.

[10] Erik Brynjolfsson, Andrew McAfee, and Michael Spence. 2014. New world order: labor, capital, and ideas in the power law economy. *Foreign Affairs* 93, 4 (2014), 44–53.

[11] US Census Bureau. 2020. HINC-06. Income Distribution to $250,000 or More for Households. https://www.census.gov/data/tables/time-series/demo/income-poverty/cps-hinc/hinc-06.html Section: Government.

[12] Josie Cox. 2017. Automation risks exacerbating income inequality across UK, think tank warns. *The Independent* (2017). https://www.independent.co.uk/news/business/news/automation-industry-financial-services-ai-robots-uk-income-inequality-ippr-a8130131.html

[13] Era Dabla-Norris, Kalpana Kochhar, Nujin Suphaphiphat, Frantisek Ricka, and Evridiki Tsounta. 2015. *Causes and consequences of income inequality: a global perspective*. International Monetary Fund.

[14] Jeff Daniels. 2019. California governor proposes 'new data dividend' that could call on Facebook and Google to pay users. https://www.cnbc.com/2019/02/12/california-gov-newsom-calls-for-new-data-dividend-for-consumers.html

[15] Alessandro Fabris, Stefano Messina, Gianmaria Silvello, and Gian Antonio Susto. 2022. Tackling documentation debt: a survey on algorithmic fairness datasets. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–13.

[16] Yakov Feygin, Hanlin Li, Chirag Lala, Brent Hecht, Nicholas Vincent, Luisa Scarcella, and Matthew Prewitt. 2021. A data dividend that works: steps toward building an equitable data economy. (2021).

[17] Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*. PMLR, 2242–2251.

[18] Pauline Glikman and Nicolas Glady. 2015. What's The Value Of Your Data? *TechCrunch* (Oct. 2015). https://techcrunch.com/2015/10/13/whats-the-value-of-your-data

[19] Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying Large Language Model Generalization with Influence Functions. https://doi.org/10.48550/arXiv.2308.03296 arXiv:2308.03296 [cs, stat].

[20] Zayd Hammoudeh and Daniel Lowd. 2023. Training Data Influence Analysis and Estimation: A Survey. https://doi.org/10.48550/arXiv.2212.04612 arXiv:2212.04612 [cs].

[21] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3174023

[22] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Patwary, Mostofa Ali, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409* (2017).

[23] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 419–429.

[24] Chris Hughes. 2018. The wealth of our collective data should belong to all of us. *The Guardian* (2018). https://www.theguardian.com/commentisfree/2018/apr/27/chris-hughes-facebook-google-data-tax-regulation

[25] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. Datamodels: Predicting Predictions from Training Data. *arXiv:2202.00622 [cs, stat]* (Feb. 2022). http://arxiv.org/abs/2202.00622 arXiv: 2202.00622.

[26] Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J. Spanos. 2019. Towards Efficient Data Valuation Based on the Shapley Value. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. PMLR, 1167–1176. https://proceedings.mlr.press/v89/jia19a.html ISSN: 2640-3498.

[27] Hyun Joon Jung, Yubin Park, and Matthew Lease. 2014. Predicting next label quality: A time-series model of crowdwork. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

[28] Jacob Kastrenakes. 2023. Adobe made an AI image generator — and says it didn't steal artists' work to do it. https://www.theverge.com/2023/3/21/23648315/adobe-firefly-ai-image-generator-announced

[29] Tae Wan Kim, Jooho Lee, Joseph Xu, and Bryan Routledge. 2020. Corporate Data Governance: Are Data Subjects Investors? *Academy of Management Proceedings* 2020, 1 (Aug. 2020), 13855. https://doi.org/10.5465/AMBPP.2020.287 Publisher: Academy of Management.

[30] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 1301–1318. https://doi.org/10.1145/2441776.2441923

[31] Thomas Kleine Buening, Meirav Segal, Debabrota Basu, Anne-Marie George, and Christos Dimitrakakis. 2022. On Meritocracy in Optimal Set Selection. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (Arlington, VA, USA) *(EAAMO '22)*. Association for Computing Machinery, New York, NY, USA, Article 20, 14 pages. https://doi.org/10.1145/3551624.3555305

[32] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 1885–1894. https://proceedings.mlr.press/v70/koh17a.html ISSN: 2640-3498.

[33] Logan Kugler. 2018. The War over the Value of Personal Data. *Commun. ACM* 61, 2 (Jan. 2018), 17–19. https://doi.org/10.1145/3171580

[34] Yongchan Kwon and James Zou. 2022. Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, 8780–8802. https://proceedings.mlr.press/v151/kwon22a.html ISSN: 2640-3498.

[35] Jaron Lanier and E Glen Weyl. 2018. A Blueprint for a Better Digital Society. *Harvard Business Review* (2018).

[36] Cassandra Liem and Georgios Petropoulos. 2016. The economic value of personal data for online platforms, firms and consumers. https://bruegel.org/2016/01/the-economic-value-of-personal-data-for-online-platforms-firms-and-consumers/

[37] Jinkun Lin, Anqi Zhang, Mathias Lécuyer, Jinyang Li, Aurojit Panda, and Siddhartha Sen. 2022. Measuring the Effect of Training Data on Deep Learning Predictions via Randomized Experiments. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 13468–13504. https://proceedings.mlr.press/v162/lin22h.html ISSN: 2640-3498.

[38] Steve Lohr. 2019. Calls Mount to Ease Big Tech's Grip on Your Data. *New York Times* (Aug. 2019). https://www.nytimes.com/2019/07/25/business/calls-mount-to-ease-big-techs-grip-on-your-data.html

[39] Julian McAuley. 2022. *Personalized Machine Learning*. Cambridge University Press.

[40] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *Comput. Surveys* 54, 6 (July 2021), 115:1–115:35. https://doi.org/10.1145/3457607

[41] Anshul Mittal, Kunal Dahiya, Sheshansh Agrawal, Deepak Saini, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. DECAF: Deep Extreme Classification with Label Features. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 49–57. https://doi.org/10.1145/3437963.3441807

[42] Judith S Olson and Wendy A Kellogg. 2014. *Ways of Knowing in HCI*. Vol. 2. Springer.

[43] Edward Ongweso, Jr. 2020. Andrew Yang's Data Dividend Isn't Radical, It's Useless. https://www.vice.com/en/article/935358/andrew-yangs-data-dividend-isnt-radical-its-useless

[44] Eduardo Porter. 2018. Your Data Is Crucial to a Robotic Age. Shouldn't You Be Paid for It? *New York Times* (March 2018). https://www.nytimes.com/2018/03/06/business/economy/user-data-pay.html

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. https://doi.org/10.48550/arXiv.2204.06125 arXiv:2204.06125 [cs].

[46] David Rotman. 2014. Technology and inequality. *TECHNOLOGY REVIEW* 117, 6 (2014), 52–60.

[47] David Rotman. 2022. How to solve AI's inequality problem. https://www.technologyreview.com/2022/04/19/1049378/ai-inequality-problem/

[48] Susumu Saito, Chun-Wei Chiang, Saiph Savage, Teppei Nakano, Tetsunori Kobayashi, and Jeffrey P. Bigham. 2019. TurkScanner: Predicting the Hourly Wage of Microtasks. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 3187–3193. https://doi.org/10.1145/3308558.3313716

[49] Nathan Schneider, Primavera De Filippi, Seth Frey, Joshua Z Tan, and Amy X Zhang. 2021. Modular politics: Toward a governance layer for online communities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[50] Robert Shapiro and Siddhartha Anejo. 2018. *Who Owns Americans' Personal Information and What Is It Worth?* Technical Report. Future Majority. https://assets.futuremajority.org/uploads/report-for-future-majority-on-the-value-of-people-s-personal-data-shapiro-aneja-march-8-2019.pdf

[51] Rachael Hwee Ling Sim, Xinyi Xu, and Bryan Kian Hsiang Low. 2022. Data Valuation in Machine Learning: "Ingredients", Strategies, and Open Challenges. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, Vienna, Austria, 5607–5614. https://doi.org/10.24963/ijcai.2022/782

[52] Bob Simison. 2022. How Sending Stimulus Checks to the Poor Can Boost the US Economy. https://www.chicagobooth.edu/review/how-sending-stimulus-checks-poor-can-boost-us-economy

[53] Joseph E Stiglitz. 2012. *The price of inequality: How today's divided society endangers our future*. WW Norton & Company.

[54] Financial Times. 2019. *Big Tech must pay for access to America's 'digital oil'*. Financial Times. https://www.ft.com/content/fd3d885c-579d-11e9-a3db-1fe89bedc16e

[55] Hayley Tsukayama. 2020. Why Getting Paid for Your Data Is a Bad Deal. https://www.eff.org/deeplinks/2020/10/why-getting-paid-your-data-bad-deal

[56] Jazmine Ulloa. 2019. Newsom wants companies collecting personal data to share the wealth with Californians. *latimes.com* (May 2019). https://www.latimes.com/politics/la-pol-ca-gavin-newsom-california-data-dividend-20190505-story.html

[57] James Vincent. 2022. Shutterstock will start selling AI-generated stock imagery with help from OpenAI. https://www.theverge.com/2022/10/25/23422359/shutterstock-ai-generated-art-openai-dall-e-partnership-contributors-fund-reimbursement

[58] Nicholas Vincent and Brent Hecht. 2021. A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 4:1–4:15. https://doi.org/10.1145/3449078

[59] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. 2021. Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 215–227.

[60] Nicholas Vincent, Yichun Li, Renee Zha, and Brent Hecht. 2019. Mapping the Potential and Pitfalls of "Data Dividends" as a Means of Sharing the Profits of Artificial Intelligence. *arXiv preprint arXiv:1912.00757* (2019).

[61] Tarun Wadhwa. 2020. Economic Impact and Feasibility of Data Dividends. (2020), 12.

[62] Tianhao Wang and Ruoxi Jia. 2022. Data Banzhaf: A Data Valuation Framework with Maximal Robustness to Learning Stochasticity. http://arxiv.org/abs/2205.15466 arXiv:2205.15466 [cs, stat].

[63] Tianhao Wang and Ruoxi Jia. 2022. Data Banzhaf: A Data Valuation Framework with Maximal Robustness to Learning Stochasticity. *arXiv preprint arXiv:2205.15466* (2022).

[64] Tom Yan and Ariel D. Procaccia. 2021. If You Like Shapley Then You'll Love the Core. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 6 (May 2021), 5751–5759. https://ojs.aaai.org/index.php/AAAI/article/view/16721 Number: 6.

[65] Amy X. Zhang, Grant Hugh, and Michael S. Bernstein. 2020. PolicyKit: Building Governance in Online Communities. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 365–378. https://doi.org/10.1145/3379337.3415858

7 APPENDIX

This Appendix include several additional analyses that relating to data valuation experiments.

7.1 Additional Data Value Experiments

7.1.1 Batch Size. Figure 5 shows how data value estimates vary with different choices of batch size (i.e. how many training points are considered simultaneously). We show the low cardinality approach, so the overall range of values is still quite large. We see that batch size also has a large impact on how data value estimates are distributed. Larger batch sizes have much tighter distributions, with lower range, lower variance, and less skew. Furthermore, as batch size increases, we see the median data value decreases.

These results are consistent with the idea that batch size, like cardinality, impacts which counterfactual scenarios data valuation is "exploring" . Specifically, batch size controls the maximum dataset size we can consider, while different beta distributions control how much we weight all the possible combinations up to the maximum size. In other words, batch size imposes a ceiling on the size of counterfactual scenarios the data valuation technique considers. This explains why batch size shows a similar trend to cardinality.

Examining batch size in this context also raises an important question: is it possible that certain observations may miss out an being appraised their maximum value, because the batch size was too low? For instance, if an observation on its own causes accuracy to decrease, but in combination with 10000 other data points causes a substantial increase in accuracy, is a small batch size unfair to that observation? Looking at our correlation results from above (Fig. 1) together with the data values in Fig. 5, this seems possible but unlikely; some observations may see different relative values with different batch sizes, but most will not. For a user seeking to minimize variance in payments, a small batch size with a wide distribution of value estimates is likely preferable.

7.1.2 Covtype example. In Figure 6, we include a density plot of data value estimates for the Covtype dataset [6] for many different cardinality weights. We see qualitatively similar trends to our results with synthetic data.

7.2 Case Study: Making Payments to US Households Based on Data Values

For illustrative purposes, we imagine a very specific scenario: a US federal entity is disbursing payments on the order of 2000 USD to a sample of US households. The use of household income is convenient for estimating the impact on overall income inequality.

7.2.1 Transforming Data Values to Money. To implement a data dividend, we must transform data value estimates into units of money. Data values can be negative, so this poses a challenge because we assume that early data dividends should not be allowed to impose debt on data contributors. This means we must transform the data value assigned to each person or coalition into non-negative (but potentially zero) payment amounts. We consider three simple approaches that encapsulate different philosophies of data dividends, similar to prior work [60].

The first approach we consider is to linearly "shift" data values so that the most negative value becomes 0. Then, we normalize all values to add up to 1, so each person/coalition is assigned some fraction of the total pot. Mathematically, this is $x_{shift} = x - min(x)$, $x_{frac} = \frac{x_{shift}}{sum(x_{shift})}$, $x_{dollars} = x_{frac} * total$

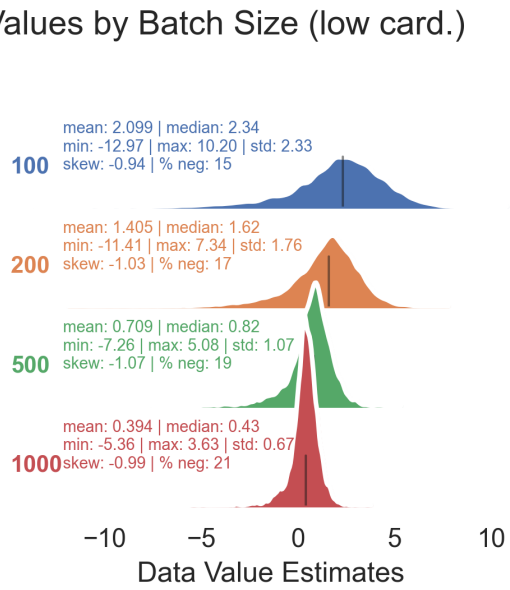

Fig. 5. Impact of valuation batch size on data values. Follows the structure of Fig. 2. These data values estimates were produced using a low cardinality approach.

This approach encodes the stance that payments should be related to the impact an observation has (in terms of a weighted average over many possible combinations of observations). Furthermore, it maintains the overall shape of the data value distribution. It also guarantees that every data laborer receives at least some small payment (except the individual with the most negative data value). However, it can lead to some unintuitive outcomes, because the ratio between any two data values is highly sensitive to the minimum value. We can illustrate this issue with a simple example. Imagine two people – Alice and Bob – have data values of 0.1 and 0.2 respectively. If Alice and Bob receive a dividend under "shift", Bob receives twice as much money, which has an intuitive connection to the fact that Bob has twice the impact on accuracy. However, if another person – Chen – joins and has data value -0.1, Bob now receives 1.5x more than Alice, not 2x more. If Di join with data value -0.2, Bob receives just 1.33x more than Alice. Under shift, relative payments are very sensitive to the "most harmful" observation.

The second approach we consider is to simply "clip" all negative values to zero, and then normalize so values add up to 1. This approach encodes the stance that the dividend implementer only wants to pay people or coalitions with a positive impact on model utility, and that the ratio of any two payments is equal to the ratio of the corresponding data values (a person with data value estimate of 0.2 always get 2x the payment of someone with data value estimation of 0.1).

The third approach is to just split payments equally between all people with positive data values. This approach is similar to "clip". One potential justification for this kind of approach would be to connect data valuation to some notion of quality thresholds (i.e., spam detection). In practice, any kind of data dividends program that uses data quality as an inclusion criteria is effectively using this approach.

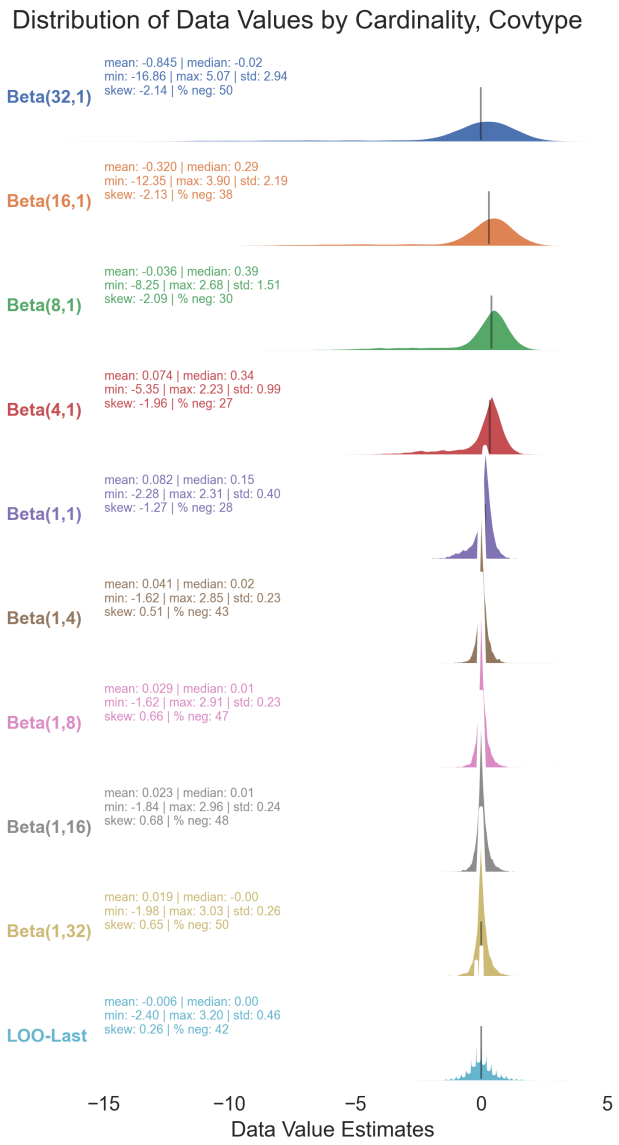

Fig. 6. Data value estimate distributions for a sample of data from the Covertype dataset.

We refer to these approaches as "shift", "clip", and "binarize" respectively in our figures below. There are numerous ways to adjust procedure for mapping value estimates into monetary amounts that could be considered reasonable. For instance, perhaps designers would like to first shift data values and then re-scale, or use a more sophisticated

clipping procedure that also handles outliers. As such, the exact values that our simulations outputs are much less important than the general trends we observe. While we include results that imagine monetary amounts based on specific assumptions (e.g., an average payment of 2000 USD per person), these are mainly for illustrative purposes.

Then, we transform the data values into dividend payments using each of the approaches described above. We assume a pot of $200,000 to be disbursed amongst 1000 people in each batch, i.e. a mean payment of $2000 per person, a similar order of magnitude to the stimulus checks sent out in the U.S. during the Covid-19 pandemic [52].

As noted above, this choice is meant to provide some more intuition as to how differences in data values translate into different payments. The exact value of the total data dividend pot is not meant to represent an estimate of the most likely size of a data dividend payment (this will vary heavily based on political considerations), nor or any of the key contributions of the results reliant on these assumptions holding. Rather, the USD results are meant as illustrative examples to show how differences in value estimate distributions can translate into differences in payments.

From here, we produce a box plot describing the distribution of payments and then calculate the inequality of the ensuing payments in terms of the standard deviation and the Hoover index. The Hoover index provides an interpretable summary of inequality: the fraction of the total pot that would need to be redistributed to arrive at a perfectly equal distribution.

To illustrate how differences in data values might translate into very concrete impacts on inequality metrics, we simulate US households receiving the dividend payments. For each repetition, we draw a sample of incomes from the 2020 U.S. household income bracket data released by the U.S. Census ("HINC-06" [11]).

7.2.2 Case Study Scenarios. Using this data, we simulate three scenarios. In the first scenario, we assign incomes to our simulated data contributors at random. We can think of this as estimating the impact of a data dividend where the appraised dataset is specifically chosen to have data values that are completely uncorrelated with income.

Next, we assume data value and income are perfectly correlated, i.e. the *most regressive possible scenario*. This might be the case if the test set is chosen to specifically reward increased accuracy for high income users. In this scenario, the person with the highest income will receive the largest dividend. Finally, we consider the case in which data value and income are anti-correlated, i.e. the most progressive scenario.

In interpreting these simulation results, we assume that choices leading to very unequal dividends are at higher risk of exacerbating inequality. Given the long line of work in HCI and related fields showing that computing technologies often perform more poorly for marginalized groups (see e.g. [40]), it seems more likely that data values will be correlated positively, not negatively, with income.

In our plots showing changes in income inequality in terms of the Hoover index, we include three baselines to contextualize these results. As a middle ground, we show a vertical line at -0.007 corresponding to the impact on Hoover index from giving every household a 2k payment. As a progressive upper bound, we show a vertical line at -0.014 correspond to the impact of giving all below median income households a 4k payments (i.e., a targeted intervention intended to create especially progressive outcomes). Finally, we also include the baseline of 0, i.e. the threshold at which a given set of dividends is actually regressive.

Just because a particular choice leads to concentrated data values, this does not *guarantee* the dividend will be regressive. Rather, it suggests that if dividend designers elect to make choices that creates unequal dividends, they will likely benefit from taking extra steps to avoid regressive outcomes. In practice, a highly progressive scenario seems unlikely without very intentional design (e.g., using a test set meant to emphasize performance for people who are currently marginalized, disadvantaged, etc.).

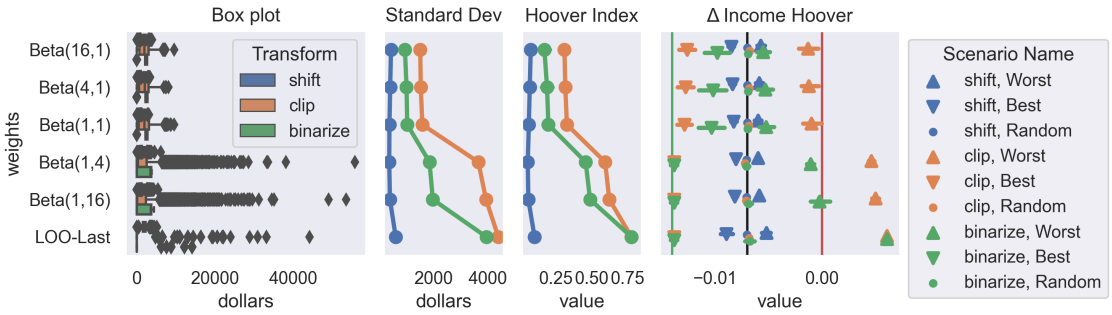

Fig. 7. Shows payment case study outcomes for payments with a mean of 2000 USD. From left to right: a boxplot faceted by different transform approaches, the standard deviation of payments, the Hoover index of payments, and finally the impact of U.S. household income inequality measured with the Hoover index. Each plot shares the same y-axis to emphasize the role of different cardinality choices. The box plot uses the same x-axis for all facets, to show the extreme differences in payment distributions.
The rightmost panel includes changes in household income Hoover under "Worst" (most regressive), "Best" (most progressive) and "Random" (data values and income uncorrelated) conditions. Rightmost panel includes three vertical lines corresponding to several baselines: the change in income from allocating 4k payments to all below-median income households, the change in income from allocating 2k payments to all recipients, and a line at zero.

7.2.3 Case Study Results. Above, we saw how various design choices impact the distribution of raw data values. Here, we examine how the data value estimate distributions produced by different design choices translate into different payment outcomes. We can think of this as a case study of what might happen if a data dividend was rolled out using the above data value estimation approaches and pot of money large enough to make the average payment 2000 USD. Overall, these results show that differences in data values can indeed translate into very different payment outcomes. If the pot of money were larger or smaller, we would expect the specific numbers and baselines to change, but for the general trends to hold.

Looking specifically at Figure 7, we see that higher cardinality choices – which lead to tighter data value distributions – are also more unequal overall. A similar, but less extreme trend exists for batch size, shown in Figure 8. In fact, consistently across our results, cardinality is a very dominant factor in determining how unequal dividend payments are.

Looking at the role of different transformation approaches (i.e., comparing the different colors in Fig. 7), we see that in general the *shift* approach creates more equal payments. *clip* creates very unequal distributions, and *binarize* serves as somewhat of a middle ground approach. The choice of transformation has a major impact on payments. For dividend recipients interested in having lower variance (i.e., more "certain") payments, the results in Fig. 7 suggest that the clip and binarize transformations may be undesirable.

However, we see that the coalitions approach can minimize the impact of the transformation choice. In Fig. 9, when using low cardinality value estimation, payments remain fairly equal and actually become more equal with more coalitions. Even if high cardinality value estimation is used (shown in the bottom row of Fig. 9), the coalitions approach also keeps payments more equal so long as the number of total coalitions is low (i.e., each coalition is large).

Importantly, the rightmost panel of each plots shows how much outcomes could vary based on whether data value estimates and a recipient's income are correlated (positively, negatively, or not at all). In general, a particularly unequal set of payments can be very progressive – often matching the "very progressive baseline" (green vertical line,

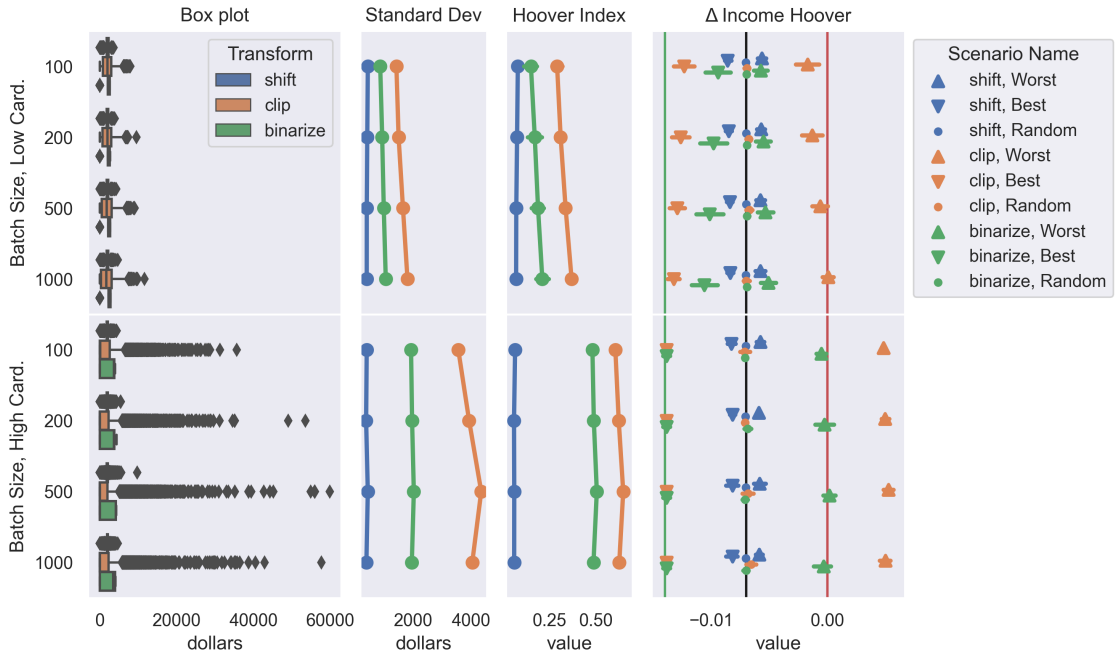

Fig. 8. Following the structure of Fig. 7, shows payment outcomes for different choices of batch size. Top row shows low cardinality data values and bottom row shows high cardinality.

corresponding to the outcome of targeted payments). However, that same set of payments can also be very regressive (i.e., meaningful increase in the household income Hoover index).

A general takeaway from these results is to confirm that the large differences in data value estimate distributions we observed can indeed translate into very different dividend outcomes. In other words, designer choices – which are discretionary, and may even seem arbitrary to recipients or to designers themselves – can massively impact whether data dividends are progressive or regressive. Even with support for data coalitions, dividends can still create regressive outcomes (the absolutely most certain way to avoid this: treat all users as members of one coalition and just direct a single pot of money towards progressive initiatives).

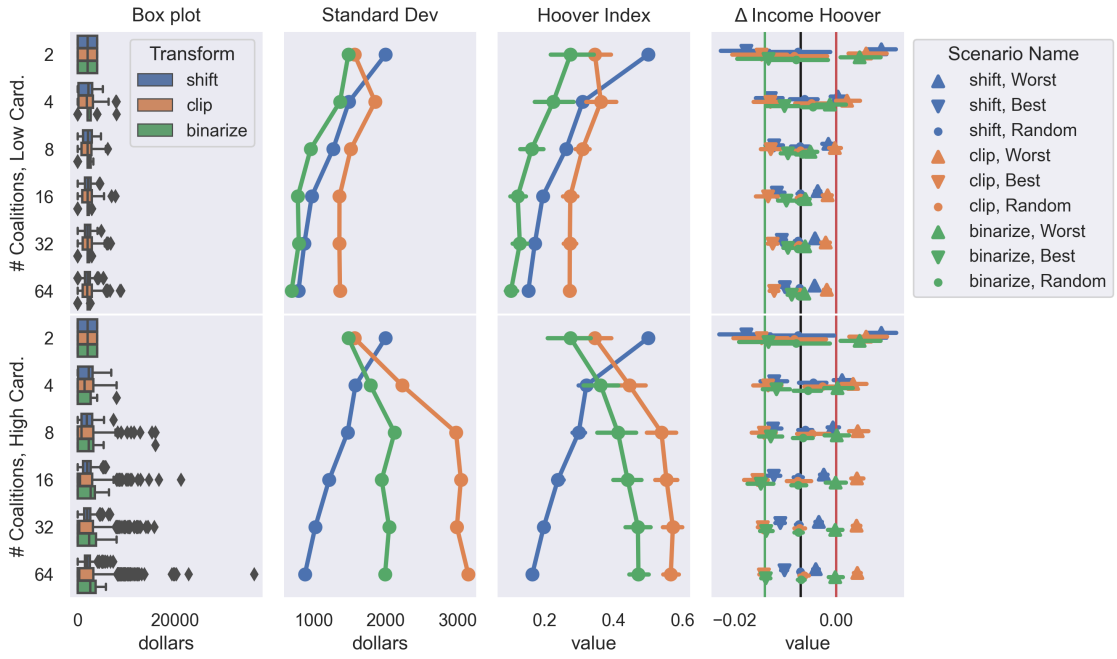

Fig. 9. Following the structure of Fig. 7, shows payment outcomes for different numbers of coalitions. Top row shows low cardinality data values and bottom row shows high cardinality.