On the Value of Spatiotemporal Information: Principles and Scenarios

Heba Aly
Dept. of Computer Science
University of Maryland
College Park, MD, USA
heba@cs.umd.edu

John Krumm
Microsoft Research
Microsoft Corporation
Redmond, WA, USA
jckrumm@microsoft.com

Gireeja Ranade
Microsoft Research
Microsoft Corporation
Redmond, WA, USA
giranade@microsoft.com

Eric Horvitz
Microsoft Research
Microsoft Corporation
Redmond, WA, USA
horvitz@microsoft.com

ABSTRACT

Location data from mobile devices is a sensitive yet valuable commodity for location-based services and advertising. We investigate the intrinsic value of location data in the context of strong privacy, where location information is only available from end users via purchase. We present an algorithm to compute the expected value of location data from a user, without access to the specific coordinates of the location data point. We use decision-theoretic techniques to provide a principled way for a potential buyer to make purchasing decisions about private user location data. We illustrate our approach in two scenarios: the delivery of targeted ads specific to a user's home location and the estimation of traffic speed. In both cases, the methodology leads to quantifiably better purchasing decisions than competing methods.

CCS CONCEPTS

• Information systems → Location based services; **• Human-centered computing → Ubiquitous and mobile computing systems and tools**;

KEYWORDS

GPS, location, crowdsourcing, decision theory, value of information (VOI), advertising, traffic

ACM Reference Format:
Heba Aly, John Krumm, Gireeja Ranade, and Eric Horvitz. 2018. On the Value of Spatiotemporal Information: Principles and Scenarios. In *26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '18), November 6–9, 2018, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3274895.3274905

1 INTRODUCTION

As people carry and interact with their connected devices, they create spatiotemporal data that can be harnessed by them and others to generate a variety of insights. Proposals have been made for creating markets for personal data [1] rather than for people to either provide their behavioral data freely or to block sharing. Some of these proposals are specific to location data [9]. Several studies have explored the price that people would seek for sharing their GPS data [4, 16, 24]. However, little has been published on determining the value of location data from a buyer's point of view. For instance, a Wall Street Journal blog says [17]:

> "What groceries you buy, what Facebook posts you 'like' and how you use GPS in your car: Companies are building their entire businesses around the collection and sale of such data. The problem is that no one really knows what all that information is worth. Data isn't a physical asset like a factory or cash, and there aren't any official guidelines for assessing its value."

We present a principled method for computing the value of spatiotemporal data from the perspective of a buyer. Knowledge of this value could guide pursuit of the most informative data and would provide insights about potential markets for location data.

We consider situations where a buyer is presented with a variety of location data points for sale, and we provide estimates of the value of information (VOI) for these points. Even when the coordinates of the location data points are unknown, we compute the VOI based on the prior knowledge that is available to the buyer and on side-information that a user may provide (e.g. the time of day or location granularity). The VOI computation is customized to the specific goals of the buyer, such as targeting ad delivery for home services or offering efficient driving routes. We account for the fact that location data and user state are both uncertain. Additional data purchases can help reduce this uncertainty, and we quantify this reduction as well.

We discuss related work in the next section. Then, in Section 3, we introduce a decision-making framework with a detailed analysis of geo-targeted advertising. We focus on the buyer's goal of delivering ads to people living within a certain region. We show that our method performs better than alternate approaches in terms of inferential accuracy, data efficiency, and cost. In Section 4, we present a general method for computing VOI for spatiotemporal data, abstracting away the specific application to reveal the essential elements of the approach. In Section 5, we apply the methodology to a traffic estimation scenario using real and simulated spatiotemporal data.

Our contributions are as follows:

- We present a methodology to calculate the expected monetary value of a user's location coordinates, even when the detailed coordinates are unknown to the buyer a priori.
- We provide an algorithm for a buyer to make purchasing decisions about location data that may be sold by owners of the data, despite the specific location uncertainty.

SIGSPATIAL '18, November 6–9, 2018, Seattle, WA, USA

ACM ISBN 978-1-4503-5889-7/18/11.
https://doi.org/10.1145/3274895.3274905

- We demonstrate how the algorithm behaves in two scenarios: targeted ad delivery and crowdsourced traffic information.

To the best of our knowledge, this is the first principled method to compute the value of unseen crowdsourced location data from a buyer's point of view.

2 RELATED WORK

We review related work on crowdsourcing, optimal sensing, and data pricing.

2.1 Crowdsourcing

In geographic crowdsourcing, a large group of people is harnessed supply spatial data. The crowd can be active participants in gathering the data, e.g. OpenStreetMap mapping parties [8]. Shahabi et al. have done extensive work on assigning crowd workers to efficiently complete tasks at specified locations, e.g. [10]. The crowd can also serve as passive participants who engage in their normal travels, such as data provided in Nokia's Mobile Data Challenge [14]. Sometimes the crowd gives away their location data at no cost, which has been explored in literature on Volunteered Geographic Information, starting with a paper by Goodchild [6]. For other spatial data-gathering tasks, workers can earn money via sharing their location information, e.g. with Gigwalk [2].

Our scenarios assume participants passively collect location data during their normal activities. As an example, the location data collected from Waze users helps compute driving routes that are sensitive to traffic.

2.2 Optimal Spatial Sensing

Our work on the valuation of location data is related to methods for choosing sensors for efficient spatial inferences. Krause et al. exploited submodularity to find a near-optimal placement of spatial sensors with the goal of maximizing the mutual information between sensed and unsensed locations [11]. Singh et al. considered the problem of directing the paths of multiple mobile robots to increase their collective information return [23]. For Gaussian process regression, Seo et al. introduced heuristics for choosing sensed points that seek to minimize the variance of the inferred result, for individual points and as averaged over the whole space [21]. In [26], Zhao et al. introduced a formalization for considering both the value of information and cost of information for selecting sensors in a sensor network.

The work most closely related to ours is Krause et al., who developed a model for sensing an entire system, such as a traffic network, from sensors with unknown locations, such as vehicles, while minimizing the number of sensor readings [12]. Our work differs in that we introduce a decision space where the data buyer must infer the discrete state of a random variable subject to a payoff matrix. The payoff matrix becomes important not only in optimizing which sensor readings to use, but also for estimating their value.

2.3 Buying and Selling Location Data

Markets for private data have been proposed, such as Adar and Huberman's "Market for Secrets", aimed at accessing anonymous data [1]. Kanza and Samet propose a marketplace for geosocial data [9]. Our work builds on these ideas by demonstrating how to price location data depending on its intended use.

We know from a variety of surveys that buyers and sellers attach very different values to location data. Research on the sale of location data includes investigations of the price that people would demand in return for giving up their location privacy. For example, in [4], researchers surveyed over 1200 people in five European countries. The median asking price for one month of location data was approximately €50 (US$40 at the time) for academic use. The data was assumed sampled every five minutes at cell tower resolution. The price rose to €100 (US$80 at the time) for one month of data for commercial use and €250 for one year of data.

In [24], 60 volunteers were asked to price 6 weeks of their location data. Their median price for one GPS point was €3 (about US$4 at the time). Their median price for all 6 weeks was €22.5 (US$30). The authors found that location data was priced higher than data on communications, application usage, and media such as photos.

Trend Micro surveyed over 1000 consumers from around the world, asking about the value they attributed to different types of their personal data [16]. Although the amount of location data was unspecified, the average price for their location data was US$16.10, and the average price for their home address was US$12.90.

Location data appears to be priced lower by buyers than the valuation provided in studies with end users. For instance, based on industry pricing data, a 2013 Financial Times article says, "General information about a person, such as their age, gender and location is worth a mere $0.0005 per person, or $0.50 per 1,000 people." [25].

We address the potential disparity in valuation of location data by sellers and buyers by computing the expected value of information of location information in different scenarios.

3 SCENARIO 1: HOME TARGETED ADS

We now describe methods and case studies to compute the expected value of gaining access to location points. We provide an example scenario to demonstrate the relevance and effectiveness of our framework. We call this scenario "Home Targeted Ads", because it focuses on a business that wants to deliver ads to people whose home is in a certain geospatial region. For instance, a local roofing business may be licensed only in a certain geographic area and wish their ads to only be delivered to people who live in that area. A mobile dog grooming service may want to limit its advertising to a region that they can reach efficiently. We will refer to this target region as $\mathcal{R}$. It can be any closed region on the ground, as per the examples displayed in Figure 1.

The buyer in this case could be the business itself or an advertising specialist who can find the best recipients for the ads. In either case, the buyer seeks to find the home locations of potential ad recipients. There are multiple ways to find a person's home location: a telephone directory usually gives names and addresses, and many people give their home city as part of their social media profiles. However, the telephone directory can be incomplete and/or out-of-date, and social media profiles usually give only city-level resolution. Location measurements, such as those from GPS, are usually very precise, and they can be used to infer the location of a person's home, as we illustrate below. In this scenario, the buyer

Table 1: The payoff matrix for home targeted ads. The values in parentheses are used for our experiments.

		Home Location	
		not in region	in region
Ad	do not deliver	b_{11} (0)	b_{12} (β)
	deliver	b_{21} (γ)	b_{22} (1.0)

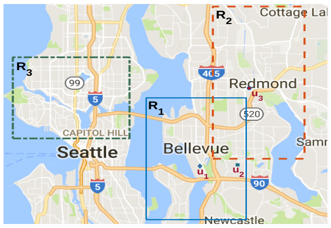


Figure 1: The three test regions for the home targeted ads experiments. Three examples for users' homes are highlighted (u_1, u_2 and u_3).

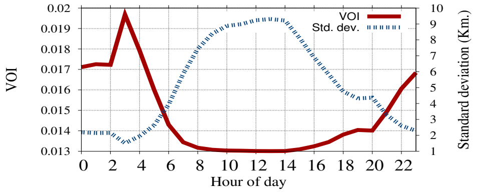


Figure 2: The deviation from home (dotted line) and the expected value of information (VOI) throughout the day. The VOI is calculated for payoff matrix with values: $[b_{11}, b_{12}; b_{21}, b_{22}] = [0, -0.9; -0.9, 1]$.

will seek to buy a small number of time-stamped location measurements from potential ad recipients and use the measurements to decide who should receive the ad.

3.1 Decision to Deliver an Advertisement

In this scenario, a buyer must choose whether or not to deliver an ad to a potential recipient, and the crux of this decision depends on whether or not the potential recipient lives in the targeted region. We model the costs to the buyer with a payoff matrix. The matrix describes the monetary gain or loss depending on the decision of whether or not to deliver an ad to the potential recipient and depending on whether or not the recipient lives in the region $\mathcal{R}$, as shown in Table 1.

The four cases in Table 1 represent the following scenarios:

- **Ad not delivered when home is *not* in region $\mathcal{R}$** (payoff b_{11}): This is a neutral outcome, because an ad was correctly withheld from a person who does not live in the targeted region. The cost (and benefit) is normally zero in this case, thus $b_{11} = 0$.
- **Ad not delivered when home is in region $\mathcal{R}$** (payoff b_{12}): This is a negative outcome, because the ad should have been delivered, but was not. The cost is the lost opportunity and the possibility that a competitor may acquire the person as a customer, thus $b_{12} \leq 0$.
- **Ad delivered when home is *not* in region $\mathcal{R}$** (payoff b_{21}): This is a negative outcome, because the ad was mistakenly delivered to a person whose home is not in the target region. The cost is the wasted cost of the ad plus the annoyance caused to the targeted person, so $b_{21} \leq 0$.
- **Ad delivered when home is in region $\mathcal{R}$** (payoff b_{22}): This is a positive outcome, because it could generate a purchase from the business. The value would be the expected profit from a successful ad minus the cost of the ad, so $b_{22} \geq 0$.

We assume the payoff matrix values are given or can be learned [18].

Based on location data collected from the potential ad recipient, the buyer computes a probability distribution $P_H(\mathbf{h})$, where $\mathbf{h}$ is a two-dimensional vector, $[x, y]^T$, that describes the location of the potential recipient's home. We give a method to compute this distribution in Section 3.3. From this distribution, we can compute the probability $p_{\mathcal{R}}$ that the home is inside the targeted region $\mathcal{R}$:

$$p_{\mathcal{R}} = \int_{\mathcal{R}} P_H(\mathbf{h}) d\mathbf{h}. \tag{1}$$

Based on this we can compute the expected value of the revenue, V, given our decision on ad-delivery:

$$\mathbb{E}\big[V \mid \text{no ad}\big] = (1 - p_{\mathcal{R}})b_{11} + p_{\mathcal{R}} b_{12},$$
$$\mathbb{E}\big[V \mid \text{ad}\big] = (1 - p_{\mathcal{R}})b_{21} + p_{\mathcal{R}} b_{22}.$$

Here we assume that the advertiser has a linear utility function, e.g. gaining (or losing) \$100 is one hundred times as good (or bad) as gaining (or losing) \$1. The advertiser would choose whichever alternative has the largest expected revenue:

$$\mathbb{E}\big[V\big] = \max\Big(\mathbb{E}\big[V \mid \text{no ad}\big], \mathbb{E}\big[V \mid \text{ad}\big]\Big). \tag{2}$$

3.2 Decision to Buy a GPS Point

We consider the case where the buyer is presented with a list of points to evaluate buying, where each of these points has been recorded at a different time. The buyer is allowed to see the time stamps, but not the points' spatial coordinates.

The buyer will compute VOI to decide whether or not to buy a measured location point, having knowledge of only the point's time stamp. The buyer has already purchased n points, denoted by the random variables $L_1, L_2, \cdots, L_n$ or as the collection L_1^n. An instance of this random location variable is $l_i = [x_i, y_i, t_i, \sigma_l, c_i]^T$, which is a 5D vector with $[x_i, y_i]^T$ representing the point's 2D location at time t_i and the location precision represented as the standard deviation σ_l. We could optionally represent a varying precision for each measurement, but we assume all the users have similar location sensors with the same precision. The price of the point is c_i, which is the amount the buyer would have to pay the seller (potential ad recipient) to know (x_i, y_i). This price is determined by the seller. Using these points, the buyer computes $P_{H|L_1^n}(\mathbf{h})$, which is a probability distribution of the home location based on location measurements 1 through n. We give a method for this computation below in Section 3.3. The buyer then computes the probability that the home is in the target region (Equation (1)) and the expected revenue ($\mathbb{E}\big[V|L_1^n\big]$), as described above.

The buyer has the option of buying another location measurement L_{n+1}. The location of this new point is unknown to the buyer, but it follows a distribution $P_{L_{n+1}}(\ell_{n+1})$, which we describe in Sec. 3.4.

The VOI at time n can then be defined as the gain in revenue by receiving the $n+1$-th location $L_{n+1} = \ell_{n+1}$:

$$VOI(\ell_{n+1} \mid L_1^n = \ell_1^n) = \mathbb{E}\left[V \mid L_1^{n+1} = \ell_1^{n+1}\right] - \mathbb{E}\left[V \mid L_1^n = \ell_1^n\right]. \tag{3}$$

Hence, the expected VOI for the $n+1$-th location is given by the expected value of (3):

$$\begin{aligned} &EVOI(L_{n+1} \mid L_1^n = \ell_1^n) \\ &= \int_{\ell_{n+1}} VOI(\ell_{n+1} \mid L_1^n = \ell_1^n) \cdot P_{L_{n+1}}(\ell_{n+1} \mid L_1^n = \ell_1^n)\, d\ell_{n+1}. \end{aligned} \tag{4}$$

The decision to buy the $n+1$-th point will be based on whether the value of the point in expectation, i.e. $EVOI(L_{n+1} \mid L_1^n = \ell_1^n)$, is larger than the cost of the point, c_{n+1}. Thus, we will buy the point that maximizes the expected profit below:

$$\mathbb{E}\left[\text{Profit} \mid L_1^{n+1} = \ell_1^{n+1}\right] = EVOI(L_{n+1} \mid L_1^n = \ell_1^n) - c_{n+1}. \tag{5}$$

Here we assume that the potential ad recipient has placed a price on their location data. This price could also be set by a location broker who acts as a representative of the potential ad recipient. We note that while this equation accounts for the price of the location point, the price of the ad has already been accounted for in the values of the payoff matrix.

3.3 Estimating Home Location

In our pricing model, we assume the buyer will use location measurements from the potential ad recipient to compute a distribution describing the potential recipient's home location. The buyer will then use this to help decide whether to deliver an ad and whether to buy more location points. This section presents a principled way for a buyer to use location measurements to estimate a home location. We present this first scenario to define our approach. Thus we do not make any comparisons to other methods (*e.g.* [13]) for computing a person's home location. Our particular approach has the advantage of producing a probability distribution for the home's location, rather than a single point estimate, which is used to compute the probability that the home is in region $\mathcal{R}$ from Equation 1.

The distribution of home location after processing n points, $P_{H|L_1^n}(\mathbf{h})$, is updated by the buyer after the purchase of the location measurement ℓ_{n+1} from a potential ad recipient. The update equation is Bayes rule:

$$P_{H|L_1^{n+1}}(\mathbf{h}) = \frac{P_{H|L_1^n}(\mathbf{h}) P_{L_{n+1}|H}(\ell_{n+1})}{P_{L_{n+1}|L_1^n}(\ell_{n+1})} \tag{6}$$

The prior distribution, $P_{H|L_1^n}(\mathbf{h})$, is the posterior after processing the n-th location measurement.

The likelihood term is $P_{L_{n+1}|H}(\ell_{\mathbf{n+1}})$. This is the distribution of the measured point L_{n+1} given the home location. We model this using knowledge of where people are usually located in relation to their home. This data comes from a travel survey conducted by the Puget Sound Regional Council in 2015 [3]. The survey data consists of day-long travel diaries from 4235 people in 2324 different households. Each participant kept track of their trips for their survey day, including the street addresses of their destinations. From this, we computed a bivariate, symmetric normal distribution giving the location of each participant relative to their home. As expected, the shape of the distribution varies with the time of day, with a tighter distribution at night when people are normally home. Figure 2 shows the standard deviation of the bivariate normal as a function of the hour of the day. The random variable D describes the coordinates of the home's residents relative to the home's location:

$$D \sim \mathcal{N}\left(\mathbf{0}, \sigma_H^2(t)\, I\right)$$

Here $\mathbf{0} = [0, 0]^T$, and $\sigma_h(t)$ is the time-varying standard deviation as in Figure 2, and I is the 2x2 identity matrix. Given this, we have

$$P_{L_{n+1}|H}(\ell_{n+1}) \sim \mathcal{N}\left(H, \sigma_H(t) I\right)$$

This is the same as D, but translated to the home location $\mathbf{x}_h$.

The denominator of Equation 6 provides the conditional probability of the new point L_{n+1} given the previous points L_1^n. This is a scalar normalization factor, and we can compute it by integrating the numerator.

Before buying any points, we need a prior distribution $P_H(\mathbf{h})$, which is the distribution for home locations before seeing any location measurements from the potential ad recipient. We take this from a database of home locations in the U.S. maintained at our institution. It is a simple list of latitude/longitude pairs measured with GPS. As such, each home point carries the same uncertainty as a GPS measurement. We model the GPS uncertainty as a 2D symmetric normal distribution $\mathcal{N}(\mathbf{0}, \sigma_{GPS}^2 I)$, as suggested in [5]. The value of σ_{GPS} represents the amount of uncertainty for a GPS measurement, and we set it to three meters as a generally acceptable approximation. Assuming the home is somewhere in the U.S., the prior on home locations is then

$$P_H(\mathbf{h}) = \frac{1}{2N\pi\sigma_{GPS}^2} \sum_{i=1}^{N} \exp\left(-\frac{1}{2}(\mathbf{h}-\mathbf{h}_i)^T (\sigma_{GPS}^2 I)^{-1} (\mathbf{h}-\mathbf{h}_i)\right) \tag{7}$$

Here $\mathbf{h}_i$ are the coordinates of each home location from our database, and N is the total number of homes in the database. This prior represents the initial uncertainty about the potential ad recipient's home location. The prior helps limit home inferences to places where homes are actually located, eliminating regions like bodies of water.

3.4 Distribution of Next Location Measurement

Equation 4 computes the expected revenue from the new point L_{n+1}, and it includes the distribution $P_{L_{n+1}}(\ell_{n+1})$, which captures the buyer's knowledge of the location of the next, unknown point. To compute $P_{L_{n+1}}(\ell_{n+1})$, we again exploit the deviation from home D, saying the location measurement L_{n+1} is the vector sum of the home location H and the home deviation D. The distribution of a sum of random variables is the convolution of their addends, so we have

$$\begin{aligned} L_{n+1} &= H + D \\ P_{L_{n+1}}(\ell_{n+1}) &= P_{H|L_1^n}(\ell_{n+1}) * P_D(\ell_{n+1}) \end{aligned}$$

As a reminder, D comes from the travel survey described in Section 3.3. Intuitively, $P_{L_{n+1}}(\ell_{n+1})$ is the same as the inferred distribution of the home location, but spread out by $P_D(\cdot)$ to represent that the potential ad recipient might have been away from home. The amount of spread is $\sigma_H(t)$, which varies with the time of day.

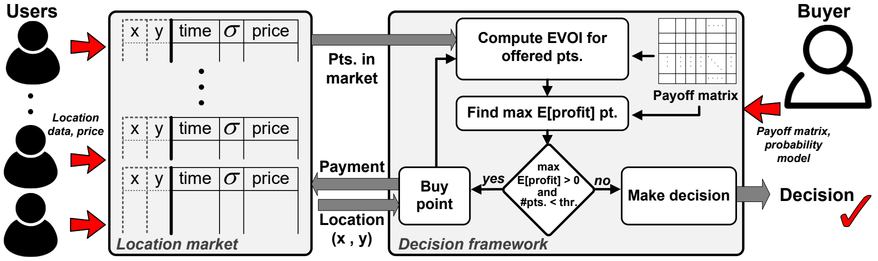


Figure 3: Proposed data-sharing mechanism and decision framework: Users offer their passively crowdsourced, time-stamped data with a certain location accuracy for a fixed price, while hiding the actual coordinates. Data buyers estimate the value of the offered data, buy points with the maximum expected profit, and make a business decision based on the points they have purchased.

3.5 Algorithm for Decisions

The final algorithm followed by the data requester and illustrated in Figure 3 consists of repeated computations of the expected profit from Equation 5 over all the available points from the user. The buyer repeatedly buys the point with the maximum expected profit (Equation 5) as long as at least one point has an expected profit greater than zero, and as long as the number of points purchased does not exceed a preset threshold. When there are no more profitable points, or if the threshold has been exceeded, the buyer harnesses the information collected to deliver the advertisement with the largest expected revenue (Equation 2).

3.6 Evaluation Experiments

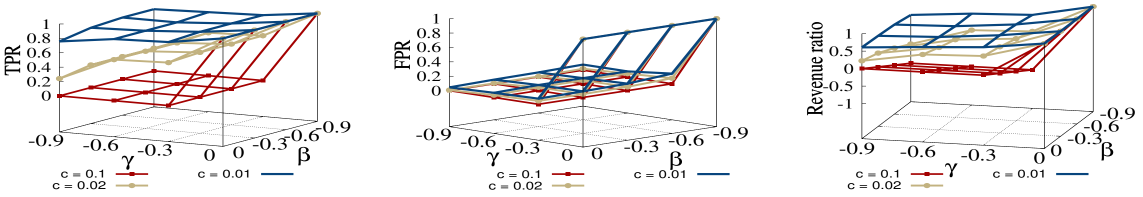
To evaluate the proposed decision framework, we used a GPS dataset of 66 participants living in the Seattle, Washington, USA area, shown in Figure 1. The participants represent employees of our institution, family, friends, and paid study participants, all of whom are adults. The trajectories were collected for an average of 40.12 days (σ = 24.43) and have an average sampling rate of 0.77 samples/minute. The trajectories represent data offered by the user to the data buyer. We define three regions to test our framework (Figure 1). We have 13, 14 and 18 users living in R_1, R_2 and R_3 respectively. To find the ground truth home location for each user, we leveraged each user's full trajectory and the American Time Use Survey [19] (ATUS). ATUS points out that users are most likely to be at their homes at midnight. Thus, we applied density-based clustering (DBSCAN) on the user's time-stamped location trajectory. Then, the largest collection of data points (cluster) at midnight was identified as this user's home [15].

We compared our decision framework to two other techniques that represent simple, practical methods to decide whether or not to send an ad to a user. For the first of these techniques, the advertiser simply makes a random decision to send the ad or not, with the probability of sending the ad set to 0.5. This represents the typical method if there is no information available about the users to guide the decision maker, and it serves as our baseline method. We call this technique "Buy no points, random ad decision" or "No points" for short. In the second comparison technique, the data requester buys a number of points from the user at random times of day. Then, the ad is sent to the user only if the majority of the purchased points are inside the region. This method reflects an assumption that users tend to spend most of their time around their homes. Using our default price of 0.01 per point, our new, proposed method recommends buying no more than 20 points in about 85% of the cases, when the expected profit per point reaches zero. Thus, in our second comparison method, we have the data requester buy 20 points regardless of their expected benefit. We call this second technique "Buy 20 random points" or "20 points" for short. In addition, for our proposed new method, we set a maximum threshold of 20 points in the evaluation to represent a realistic case where the buyer is interested in buying bounded amount of data. Note that decreasing the threshold should decrease the buyer's confidence in making the decisions, and choosing a lower threshold makes the framework more conservative in sending ads. Similarly, increasing the threshold leads to more confidence in making the decisions and potentially improving the performance. We refer to our proposed method as "VOI decision".

3.6.1 Evaluation Metrics. To evaluate the proposed decision framework, we employ three metrics: (1) *The true positive rate* (TPR) measures the proportion of correctly sent ads (i.e. ads sent to people with homes in region); (2) *the false positive rate* (FPR) measures the proportion of incorrectly sent ads (i.e. ads sent to people with homes outside the region); and (3) *the revenue ratio* which measures the ratio of the revenue gained to the maximum revenue the advertiser can gain by making perfectly correct decisions about which users should receive the ad without buying any location points.

3.6.2 Results. To test our proposed framework for different payoff matrices, we created a payoff matrix with the values in parentheses shown in Table 1. Here we have $b_{11} = 0$, which represents the neutral result of not sending an ad to someone whose home is outside the region $\mathcal{R}$. To reduce the size of the parameter space, we normalize by setting $b_{22} = 1$, which represents the reward for correctly delivering an ad to someone whose home is inside the region. The other two outcomes are negative: $b_{21} = \gamma$ represents the penalty for delivering an ad to someone not in the region, and $b_{12} = \beta$ represents the penalty for not delivering an ad to someone who does live in the region. We let both γ and β vary over $[0.0, -0.9]$. These normalizations mean we can show results over just two payoff parameters (γ and β) rather than four.

Figure 4 shows the effect of the point cost on the average performance of the proposed framework over the three test regions for the different payoff matrices. Figure 4(a) shows the true positive

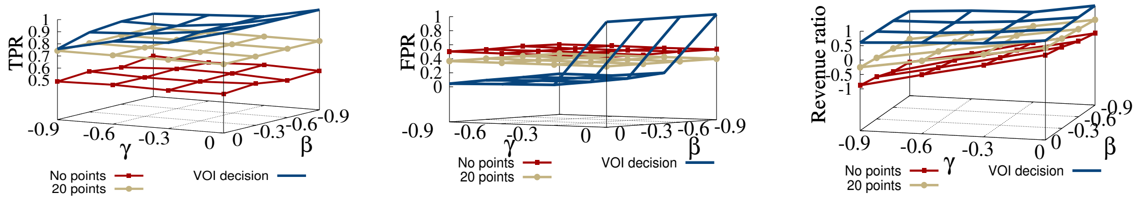


(a) True positive rate

(b) False positive rate

(c) Revenue ratio

Figure 4: Effect of the user defined cost on the proposed framework for the home targeted ads scenario (Scenario 1).



(a) True positive rate

(b) False positive rate

(c) Revenue ratio

Figure 5: Home targeted ads (Scenario 1) experiment results using the proposed framework ("VOI decision") as compared to two other methods ("No points" and "20 points").

rate for point costs of 0.1, 0.02, and 0.01. Lower costs lead to generally higher TPR across almost the whole (β, γ) space, because the buyer is willing to purchase more points, increasing their chances of making the right decision. As γ increases (moving toward zero), the TPR of the 0.1 cost case (red surface) improves dramatically. This is because γ is the penalty for delivering an ad outside the target region. As this penalty decreases, the system becomes more willing to send ads, increasing its true positive rate. This effect is also apparent in Figure 4(b), where the FPR also increases as the γ penalty moves toward zero. We note the false positive rate is fairly insensitive to our sample price points, because all three surfaces in Figure 4(b) are nearly coincident. The revenue ratio in Figure 4(c) is best (higher) for the lowest-priced points, as expected. We note that the TPR, FPR, and revenue ratio are one when γ is zero, because there is no penalty for sending ads to users outside the region. Hence, it makes sense to send ads to all users in this unrealistic case. These plots confirm that our VOI decision algorithm is working in a sensible, intuitive way.

Next, we compare the performance of our method to other methods in Figure 5. The figure shows the average results over the three regions for the different payoff matrices for a GPS point cost of 0.01. The two comparative methods' ("No points" and "20 points") TPR and FPR are independent of the payoff matrix values, because they are not considering the costs and benefits of buying points nor of making ad decisions. The algorithm "No points" (red surface) has a TPR and FPR of around 0.5. The algorithm "20 points" (yellow surface) generally performs better for both TPR and FPR, but comes with the penalty of buying 20 points for every decision. Our price sensitive "VOI decision" algorithm (blue surface) is superior to both the comparison algorithms for TPR. For FPR in Figure 5(b), the "VOI decision" algorithm (blue surface) is superior over most of the payoff range. Its FPR rises dramatically when γ is zero, where the penalty for sending an ad outside the region is zero. Finally, Figure 5(c) shows the revenue ratios of the three methods, where "VOI decision" is again significantly superior. The other two algorithms actually lose money in some regions of the payoff matrix, while the "VOI decision" algorithm is always positive. Specifically, "VOI decision" relatively improves the TPR on average by 80.2% and 20.9% and up to 107.9% (when $\gamma = 0$ and $\beta = -0.6$) and 43.7% (when $\gamma = 0$) as compared to the "No points" and "20 points" respectively. Also, "VOI decision" relatively improves the FPR on average by 38.2% and 15.8% and up to 91.1% (when $\gamma = -0.9$ and $\beta = 0$) and 78.7% (when $\gamma = -0.9$ and $\beta = 0$) as compared to the "No points" and "20 points" respectively. Moreover, "VOI decision" reduces the number of points bought to make the decision on average by 60% as compared to "20 points".

4 GENERAL FRAMEWORK FOR DECISION MAKING

While we illustrated in the previous section the value of information calculations in one example, this section presents a general framework for making purchasing decisions about location data.

We start with a general payoff matrix with a set K of possible decisions over a set S of possible states as shown in Table 2. In the previous scenario, we had set sizes $|K| = |S| = 2$. The two possible decisions were to deliver the ad or not, and the two possible states were whether or not the user's home was in the target area. In general, taking decision i under state j results in a payoff of b_{ij}, which can be any real value, positive or negative. These are represented in Table 2.

Based on already-purchased data (or a prior if no data has been purchased yet), the decision maker computes the probability of each possible state of the user, p_j for $j \in [1...|S|]$. Often there is a PDF $P_S(\mathbf{s})$, $s \in S$, describing the continuous vector state $\mathbf{s}$ and a region $\mathcal{R}_i$ in the continuous state space corresponding to state j. Then

$$p_j = \int_{\mathcal{R}_j} P_S(s)ds.$$

Table 2: General payoff matrix for decisions and states

		State					
		s_1	s_2	...	s_j	...	s_S
Decision	d_1	b_{11}	b_{12}		b_{1j}		b_{1S}
	d_2	b_{21}	b_{22}		b_{2j}		b_{2S}
	⋮						
	d_i	b_{i1}	b_{i2}		b_{ij}		b_{iS}
	⋮						
	d_K	b_{K1}	b_{K2}		b_{Kj}		b_{KS}

Table 3: Payoff Matrix for Traffic State Estimation

		Actual Traffic State		
		Red	Yellow	Green
Traffic State Dec.	Red	b_{rr}	b_{ry}	b_{rg}
	Yellow	b_{yr}	b_{yy}	b_{yg}
	Green	b_{gr}	b_{gy}	b_{gg}

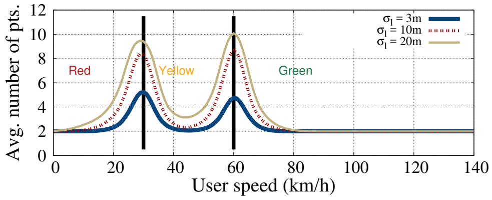


Figure 6: Average number of points bought at different possible speeds for location points with an accuracy of 3m, 10m and 20m. Our model buys more points near the traffic state boundaries. The payoff matrix is [1 -0.1 -0.1; -0.1 1 -0.1; -0.1 -0.1 1], cost = 0.01 and Δt = 3sec.

In the first scenario, the state PDF gave the distribution of the home location. One of the two regions was the advertiser's region of interest $\mathcal{R}$, and the other was, implicitly, the complement of $\mathcal{R}$.

The expected payoff for making decision d_i is

$$\mathbb{E}\left[V \mid d_i\right] = \sum_{j=1}^{|S|} p_j b_{ij}.$$

$$d^* = \arg\max_{d_i} \mathbb{E}\left[V \mid d_i\right]. \tag{8}$$

In general, we are interested in understanding when to make a certain decision, d_i, and when to buy more information. For this, we need to understand whether buying more information has not *value*. Paralleling the example we have already discussed, the crux of this will lie in computing the value of information for each of the GPS locations that are offered by the user. This value of information is computed as in Equation (4).

A key component of evaluating this value of information is understanding $P_{L_{n+1}}(\ell_{n+1})$, which is the distribution of the $n + 1$th location, which has not yet been seen by the buyer. We model this as a noisy version of the best estimate of the current location of interest. In Scenario 1 this was the home location.

With the expected VOI, EVOI, it is straightforward to compute the expected profit of point $n + 1$ with Equation 5.

Once we have the expected profit, our algorithm suggests buying points as long as the expected profit is positive. Note that the stopping point for the algorithm can be altered by maintaining a minimum profit that we would like to achieve, since this would impose a tighter constraint on the number of points we would like to buy.

Finally, the decisions will be made based on the probability of being in a certain state s_j which must be computed given the location data that has been purchased so far. This is done by computing the conditional distribution $P_S(S = s_j \mid L_1^n)$, where L_1^n represents all the location points purchased so far. The algorithm is illustrated in Figure 3.

5 SCENARIO 2: TRAFFIC STATE ESTIMATION

We now focus on a second scenario, which is a service that provides traffic state estimates for a given road segment using crowdsourced spatiotemporal data. In particular, the traffic state estimator service buys time-stamped location data from people traveling through the road network, and uses it to estimate their speed. Then this uncertain speed estimate is used to infer the road segment's discrete traffic state. For instance, we assume three levels for a highway road segment: **green** representing free flow/smooth traffic with speed greater than 60 km/hr, **red** representing congested traffic with speed less than 30 km/hr, and **yellow** representing medium congested traffic with speed between 30 and 60 km/hr. The service uses the points it buys to decide which level to assign to the road segment.

For clarity of illustration, we assume that the vehicle is on a single road segment for the duration of the analysis. The procedure described below can be generalized to the use of data from multiple vehicles traversing multiple road segments. In steady state, we assume the service has at least one previously purchased location measurement from the vehicle. This purchased data is used to place the vehicle on the road segment of interest, and it means that any subsequent point purchased from the vehicle can be used to estimate the speed of the segment using the points' time stamps. The service provider must decide whether or not to buy a new location point from the vehicle as well as which point to buy with only knowledge of the points' time stamps and location precision. While crowdsourcing traffic speeds is a familiar idea, we show how to choose intelligently which points to buy and to compute their value. Throughout the rest of the section, we will describe how the service provider will use the proposed framework to make two decisions: (1) congestion-level descriptor (color) for the road segment (2) whether to buy a new point from travelers.

5.1 Congestion Level Decision

As in the first scenario, we model the decision costs of the data-buyer using a payoff matrix. The matrix describes the monetary gain and loss depending on the provider's choice of which color to display and the road segment's actual traffic state, as shown in Table 3. There are nine different possible cases: b_{rr}, b_{yy}, b_{gg} represent positive outcomes where the service provider is choosing the correct traffic congestion level (red, yellow and green respectively), thus $b_{rr}, b_{yy}, b_{gg} > 0$. The remaining cases represent negative outcomes as the service provider is choosing a wrong congestion level descriptor. For example, payoff b_{gr} represents choosing smooth traffic (green) while actually it is congested (red). Thus, these payoffs are less than b_{rr}, b_{yy}, b_{gg} and are generally less than zero. When the actual road speed is red (severely congested), choosing green (free-flowing) would have a relatively large cost, $b_{gr} < 0$, because it could mistakenly entice drivers toward the segment only to find slow speeds. We assume the payoff matrix is given or can be learned [18].

To choose the congestion level from the noisy location data, we again employ decision theory principles [18]. Specifically, the service provider uses the purchased location data to model their belief about the traffic segment's speed. This distribution is $P_U(u)$, where u represents the vehicle's speed. We give a method to compute this distribution in Section 5.2. From this distribution, we can compute the probability that the road segment's congestion level is green as follows:

$$p_g = \int_{\mathcal{R}(g)} P_U(u)du$$

where $\mathcal{R}(g)$ represents the range of speeds for the green road coloring, which is $[60, \infty]$ in our scenario. Similar equations are used to compute the probabilities of the yellow and red states, p_y and p_r. With these probabilities, we can compute the expected revenue V for any congestion level display choice from the payoff matrix in Table 3. This is as below for the decision "r", and the decisions "g" and "y" can be evaluated similarly.

$$\mathbb{E}\big[V \mid \text{decision is } r\big] = p_r b_{rr} + p_y b_{ry} + p_g b_{rg},$$

We assume the service provider will choose to display the congestion level that gives maximum revenue, and thus the expected revenue ($\mathbb{E}\big[V\big]$) will be

$$\mathbb{E}\big[V\big] = \max\Big(\mathbb{E}\big[V \mid r\big], \mathbb{E}\big[V \mid y\big], \mathbb{E}\big[V \mid g\big]\Big).$$

In the next sections, we discuss how the service provider computes $P_U(u)$ and decisions can be made about the location points to buy.

5.2 Speed Estimation Using Crowdsourced Data

We now present a principled way for the service provider to use previously purchased location measurements to estimate the road segment speed belief $P_U(u)$. Let $L_1^n = \{L_1, L_2, ..., L_n\}$ denote random variables representing the already-purchased locations. An instance of this random variable is $l_i = [x_i, y_i, t_i, \sigma_l, c_i]^T$, which is the same as the location vector described in Section 3.2.

We follow the standard convention of representing location measurements, including GPS [5], as normal distributions in space. Thus, the spatial part of each location measurement is distributed as $\mathcal{N}\big([x_i, y_i]^T, \sigma_l^2 I\big)$. The velocity vector from two adjacent measurements in time is:

$$\mathbf{v}_i = \frac{\mu_i - \mu_{i-1}}{\Delta t_i}$$

where $\Delta t_i = t_i - t_{i-1}$, $\mu_i = [x_i, y_i]^T$, and $\mu_{i-1} = [x_{i-1}, y_{i-1}]^T$. Since the two location measurements used to compute speed are independent, their variance will add, and the distribution of the velocity vector will be:

$$\mathbf{v}_i \sim \mathcal{N}\left(\frac{\mu_i - \mu_{i-1}}{\Delta t_i}, 2\left(\frac{\sigma_l}{\Delta t_i}\right)^2 I\right)$$

where I is the 2x2 identity matrix.

We now have a distribution for the velocity vector. However, we are ultimately interested in the distribution for scalar speed, which is the magnitude of velocity. For the case of a bivariate normal with a diagonal covariance matrix, the distribution of the magnitude follows a Rician distribution [20]:

$$u_i \sim \text{Rice}\left(\frac{||\mu_i - \mu_{i-1}||}{\Delta t_i}, \frac{\sqrt{2}\sigma_l}{\Delta t_i}\right).$$

When the magnitude of the speed sufficiently exceeds the speed's standard deviation, the Rician distribution can be accurately approximated by a normal distribution [20, 22], leading to

$$u_i \sim \mathcal{N}\left(\frac{||\mu_i - \mu_{i-1}||}{\Delta t_i}, 2\left(\frac{\sigma_l}{\Delta t_i}\right)^2\right).$$

This approximation breaks down somewhat when the speed is low, such as in the red region. Our experiments in Section 5.4 show the approximation ultimately works well in our application.

The buyer estimates the road's speed from a sequence of purchased points $l_1, l_2, ..., l_{n-1}$. We assume the buyer uses a Kalman filter [7] to update the uncertain speed estimate after buying each point. In the steady state, after buying point l_n, the buyer computes an uncertain instantaneous speed distribution from l_n and l_{n-1} as described above, giving an instantaneous estimate of

$$z_n = \frac{||[x_n, y_n]^T - [x_{n-1}, y_{n-1}]^T||}{\Delta t_n}$$

and a standard deviation of $\sigma_n^u = \sqrt{2}\frac{\sigma_l}{\Delta t_n}$. The scalar Kalman update equations show how the new measurement and its standard deviation are incorporated into the speed estimate $\hat{u}_n$ and standard deviation $\hat{\sigma}_n^u$:

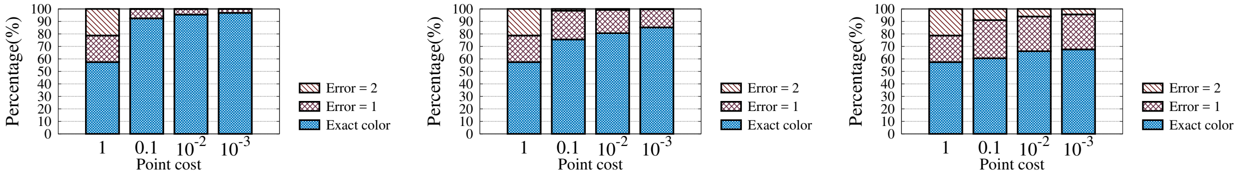
$$\begin{aligned} \hat{u}_n &= \hat{u}_{n-1} + K_n(z_n - \hat{u}_{n-1}) \\ \hat{\sigma}_n^u &= (1 - K_n)\hat{\sigma}_{n-1}^u \\ K_n &= \frac{\hat{\sigma}_{n-1}^u}{\hat{\sigma}_{n-1}^u + \sigma_n^u} \end{aligned} \tag{9}$$

The initial state of the Kalman update can be computed from the segment's traffic state history with a high value for the uncertainty $\hat{\sigma}_1^u$. The distribution of speed is $P_{U|L_1^n}(u)$ is then $\mathcal{N}(\hat{u}_n, (\hat{\sigma}_n^u)^2)$.

The Kalman filter could be replaced by other estimation techniques. We present it here as an example, and we use it in our experiments.

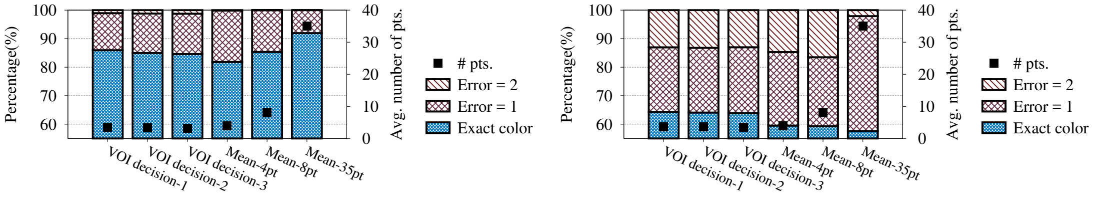
5.3 Decision to Buy a GPS Point

The buyer must decide whether to buy a new point based on its time stamp and accuracy. In this scenario, we will formulate the decision as one of buying a new speed estimate, where each new speed estimate comes from the magnitude of the velocity from the two previous two location points, as we described in Section 5.2. We leverage value of information to compute the value of knowing the traveler's unknown speed and use it to make the buying decision. Having already purchased n speed estimates, this data forms a list of speeds, denoted by the random variables $U_1, U_2, \cdots, U_n$ or as U_1^n. Using these speeds, the data requester uses the Kalman filter from Section 5.2 to compute $P_{U|U_1^n}(u)$, which is a probability distribution of the road segment speed based on speed measurements 1 through n. The buyer also computes their expected revenue $\mathbb{E}\big[V|U_1^n\big]$, as described in section 5.1, using $P_{U|U_1^n}(u) \sim \mathcal{N}(\hat{u}_n, (\hat{\sigma}_n^u)^2)$ as the speed distribution. The mean $\hat{u}_n$ and variance $(\hat{\sigma}_n^u)^2$ of this normal distribution are predicted by the Kalman filter. Since we are assuming the user is traveling at a locally constant speed, the Kalman estimate



(a) precision $\sigma_l = 3m$ (b) precision $\sigma_l = 10m$ (c) precision $\sigma_l = 20m$

Figure 7: Effect of point cost on congestion level/color decision accuracy while users are driving at different possible speeds (0-140km/hr) for location points with a precision of 3m, 10m and 20m.



(a) precision σ_l varies over $3m - 20m$ (b) precision σ_l varies over $3m - 100m$

Figure 8: The black squares show the average number of points bought while users are driving at different possible speeds for location points with randomly varying precision in the range 3m-20m and 3m-100m. This is compared to a mean filter with window sizes of 4, 8 and 35 location points. The payoff matrix for VOI decision-1 is $[b_{rr}\ b_{ry}\ b_{rg};\ b_{yr}\ b_{yy}\ b_{yg};\ b_{gr}\ b_{gy}\ b_{gg}] = [1\ {-0.9}\ {-0.9};\ {-0.9}\ 1\ {-0.9};\ {-0.9}\ {-0.9}\ 1]$, for VOI decision-2 is $[1\ {-0.4}\ {-0.9};\ {-0.4}\ 1\ {-0.9};\ {-0.9}\ {-0.4}\ 1]$, and for VOI decision-3 is $[1\ {-0.1}\ {-0.1};\ {-0.1}\ 1\ {-0.1};\ {-0.1}\ {-0.1}\ 1]$.

serves as the anticipated distribution of the as yet unknown next speed that the buyer is considering.

The value of information at time n can then be defined as the gain in revenue by receiving the $n+1$ speed measurement $U_{n+1} = u_{n+1}$:

$$VOI(u_{n+1} \mid U_1^n = u_1^n) = \mathbb{E}\left[V \mid U_1^{n+1} = u_1^{n+1}\right] - \mathbb{E}\left[V \mid U_1^n = u_1^n\right]. \tag{10}$$

Hence, the expected value of information for the $n+1$-th speed is given by the expected value of (10):

$$\begin{aligned} &EVOI(U_{n+1} \mid U_1^n = u_1^n) \\ &\quad = \int_u VOI(u \mid U_1^n = u_1^n) \cdot P_{U_{n+1}}(u \mid U_1^n = u_1^n) du. \end{aligned} \tag{11}$$

The decision to buy the $n+1$-th speed will be based on whether the value of the point in expectation, i.e. $EVOI(U_{n+1} \mid U_1^n = u_1^n)$, is larger than the cost of the speed (c_{n+1}), i.e. has a positive expected profit as below:

$$\mathbb{E}\left[\text{Profit}\right] = EVOI(U_{n+1} \mid U_1^n = u_1^n) - c_{n+1}. \tag{12}$$

Here we are assuming that the driver/data-provider has placed a price on their location (speed) data.

We give results of detailed experiments in the next section. To build intuition about these computations, we present results of a simple simulation experiment in Figure 6. For different vehicle speeds, the figure displays the number of points purchased using the methodology. Note that we buy more points whose speeds are near the congestion level thresholds, i.e. 30 and 60. In effect, the method is trying to resolve the ambiguity of speeds near the speed boundaries to avoid the cost of mistakes as expressed in the payoff matrix. In addition, as the location precision σ_l decreases, the method buys points as needed to resolve the speed uncertainty.

5.4 Evaluation Experiments

We evaluated our proposed framework in two ways: First, we used simulation studies to evaluate the effect of points’ cost on the performance of the proposed methodology across the entire speed spectrum (0-140 km/hr). In addition, we show the effect of the payoff matrix on the accuracy and compare the performance to a mean filter with different window sizes as our baseline technique. For each speed in a range from 0 to 140 km/hr with an increment of 1 km/hr, we ran 500 experiments. We estimate speeds from noisy location data with precision σ_l as described in the experiments, and we sample locations every 3 seconds. We report the average results of the experiments for each speed in the experimental range. The default payoff matrix is $[b_{rr}\ b_{ry}\ b_{rg};\ b_{yr}\ b_{yy}\ b_{yg};\ b_{gr}\ b_{gy}\ b_{gg}] = [1\ {-0.1}\ {-0.1};\ {-0.1}\ 1\ {-0.1};\ {-0.1}\ {-0.1}\ 1]$, and the default point cost is $c_i = 0.001$. We show the effect of the point cost, point precision, and the decision maker’s payoff matrix on the proposed framework as compared to the baseline technique. Second, we test the performance of our framework against real driving traces.

5.4.1 Effect of Point Cost and Precision. Using simulated data, Figure 7 shows the effect of the point cost on the performance of the proposed framework in terms of congestion level decision accuracy for different location precisions, i.e. $\sigma_l \in \{3m, 10m, 20m\}$ in parts a, b, and c of the figure, respectively. The blue bars show the percentage of correct speed interval inferences. We see that less expensive points lead to higher system accuracy, because the blue bars grow as the points become less expensive. This is because the system is more willing to buy additional points. As the price of the location points exceed their value, the buyer refrains from buying. Comparing parts a, b, and c of this figure, we also see that lower precision (larger σ_l) leads to more error, as the blue bars generally shrink from a to b to c. In this figure the error assigned to choosing the correct speed interval for the road segment is zero, represented by the blue bars. Choosing an adjacent interval (e.g. red instead of yellow) has an error of one, and choosing the interval at the other end of the spectrum (e.g. green instead of red) has an error of two.

5.4.2 Comparative Analysis. Figure 8(a) compares the performance of our framework to the mean window filter over different window sizes (baseline technique). The bars in this figure show the error rates in the same way as Figure 7. We also show the mean number of points purchased in these figures as small, black boxes. For relatively accurate location points (with precision σ_l varying uniformly at random from 3 to 20 m), Figure 8(a) shows that our proposed framework identifies the exact traffic congestion level at least 84.6% of the time ("VOI decision-3" bar in the figure); this is better than the baseline technique with window 4 points by 3.4% and with a reduction in the average number of purchased points by 20%. In addition, our approach has comparable performance to the baseline technique with window sizes 8 and 35 points along with a reduction in the number of purchased points by 60% and 90.9% respectively.

For more noisy location estimates (with σ_l varying uniformly at random from 3 to 100 m), our proposed framework estimates the exact traffic congestion level at least 63.9% of the time ("VOI decision-3" bar), as shown in Figure 8(b). This is better than the baseline technique with windows 4, 8 and 35 points by 7.3%, 7.10% and 10.8% respectively. Moreover, this comes with a reduction in number of purchased points of 15%, 57.5% and 90.2% respectively. Our framework gives higher accuracy with fewer location points. The figure also shows that varying the payoff matrix resulted in a small change in the accuracy and the average number of purchased points as seen in the first 3 bars. With a larger penalty for making a wrong decision, the framework buys more points and gives higher accuracy.

5.4.3 Validation Experiments with Real Data. Using the same GPS data as we did for the experiments in Section 3.6, we extracted 20 traces from drivers on the I-90 interstate highway and State Route 520 in Seattle, Washington at different dates and times of day. All 20 traces had more than 8 points on the road in order to compare with a mean filter with window size 8. The traces' speeds varied from 10 to 133km/hr (μ = 89.4 km/hr and σ = 36.5), covering the three congestion levels. We estimate the road congestion-level ground-truth by applying an alpha-trimmed filter to remove speed outliers and estimate the speed from the full traces. Using the default payoff matrix, our framework was able to identify the road segment's congestion levels accurately (with zero error) 95% of the time and within one level error 100% of the time. This is better than the mean filter which gave accurate prediction (with zero error) 90% of the time. In addition, our framework buys 50% fewer points as compared to the mean filter.

6 CONCLUSION

We presented a principled method for buyers of location data to compute the value of users' unseen location data. The approach relies on algorithms that consider probability distributions over locations based on data that has already been purchased, as well as the buyer's payoff matrix, to anticipate the value of future, as yet unpurchased data. As a byproduct of the quantitative valuations, the methodology identifies which unseen data is likely the most valuable for the buyer. We considered two scenarios, home-targeted ads and traffic congestion inference, to illustrate how we estimate the value of location data obtained from end users in different settings. These techniques work significantly better than competing inference approaches, both by using less data and inferring more accurate results. We believe this the work fills a gap in the pricing of location data and that the methods can help inform decisions by buyers and sellers of location data.

There are a variety of paths for future work. Applications of the approach include multiple geocentric challenges, including prediction of future locations. We also see our framework being extended with the use of nonlinear utility functions.

REFERENCES

[1] Eytan Adar et al. 2001. A market for secrets. *First Monday* 6, 8 (2001).

[2] Anna Marie Chang et al. 2013. Using a mobile app and mobile workforce to validate data about emergency public health resources. *Emerg Med J* (2013), emermed–2012.

[3] Puget Sound Regional Council. 2016. Travel Surveys: Spring 2015 Household Survey. (2016). https://www.psrc.org/travel-surveys-2015-household-survey

[4] Dan Cvrcek et al. 2006. A study on the value of location privacy. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society.* ACM, 109–118.

[5] Frank Van Diggelen. 2007. SYSTEM DESIGN & TEST-GNSS Accuracy-Lies, Damn Lies, and Statistics-This update to a seminal article first published here in 1998 explains how statistical methods can create many different. *GPS world* 18, 1 (2007), 26–33.

[6] Michael F Goodchild. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69, 4 (2007), 211–221.

[7] Mohinder S Grewal. 2011. Kalman filtering. In *International Encyclopedia of Statistical Science.* Springer, 705–708.

[8] Mordechai Haklay and Patrick Weber. 2008. Openstreetmap: User-generated street maps. *Pervasive Computing* 7, 4 (2008), 12–18.

[9] Yaron Kanza et al. 2015. An online marketplace for geosocial data. In *SIGSPATIAL.* ACM, 10.

[10] Leyla Kazemi et al. 2012. Geocrowd: enabling query answering with spatial crowdsourcing. In *Proceedings of the 20th international conference on advances in geographic information systems.* ACM, 189–198.

[11] Andreas Krause et al. 2008. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research* 9, Feb (2008), 235–284.

[12] Andreas Krause et al. 2008. Toward community sensing. In *IPSN.* IEEE, 481–492.

[13] John Krumm et al. 2015. Placer++: Semantic place labels beyond the visit. In *Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on.* IEEE, 11–19.

[14] Juha K Laurila et al. 2012. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing.*

[15] Mingqi Lv et al. 2012. Discovering personally semantic places from GPS trajectories. In *CIKM.* ACM, 1552–1556.

[16] Trend Micro. 2015. How Much is Your Personal Data Worth? Survey Says... (2015). https://www.trendmicro.com/vinfo/us/security/news/internet-of-things/how-much-is-your-personal-data-worth-survey-says

[17] Vipal Monga. 2014. The Big Mystery: What's Big Data Really Worth? (2014). https://blogs.wsj.com/cfo/2014/10/13/the-big-mystery-whats-big-data-really-worth/

[18] D Warner North. 1968. A tutorial introduction to decision theory. *IEEE Transactions on Systems Science and Cybernetics* 4, 3 (1968), 200–210.

[19] U.S. Bureau of Labor Statistics. 2016. American Time Use Survey. (2016). https://www.bls.gov/tus/

[20] Stephen O Rice. 1945. Mathematical analysis of random noise. *The Bell System Technical Journal* 24, 1 (1945), 46–156.

[21] Sambu Seo et al. 2000. Gaussian process regression: Active data selection and test point rejection. In *IJCNN*, Vol. 3. IEEE, 241–246.

[22] Jan Sijbers et al. 1999. Parameter estimation from magnitude MR images. *International Journal of Imaging Systems and Technology* 10, 2 (1999), 109–114.

[23] Amarjeet Singh et al. 2007. Efficient Planning of Informative Paths for Multiple Robots.. In *IJCAI*, Vol. 7. 2204–2211.

[24] Jacopo Staiano et al. 2014. Money walks: a human-centric study on the economics of personal mobile data. In *UbiComp.* ACM, 583–594.

[25] Emily Steel. 2013. Financial worth of data comes in at under a penny a piece. (2013). https://www.ft.com/content/3cb056c6-d343-11e2-b3ff-00144feab7de

[26] Feng Zhao et al. 2002. Information-driven dynamic sensor collaboration. *Signal processing magazine* 19, 2 (2002), 61–72.