DOI:10.1145/3410387

Computing Value of Spatiotemporal Information

By Heba Aly, John Krumm, Gireeja Ranade, and Eric Horvitz

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 three scenarios: the delivery of targeted ads specific to a user's home location, the estimation of traffic speed, and the prediction of location. In all three cases, the methodology leads to quantifiably better purchasing decisions than competing approaches.

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 either to provide their behavioral data freely or to refuse sharing. Some of these proposals are specific to location data.[6] Several studies have explored the price that people would seek for sharing their GPS data.[5, 13, 9] 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[10]:

> "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 set of location data points for sale, and we provide estimates of the value of information (VOI) for these points. Because 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, offering efficient driving routes, or predicting a person's location in advance. 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.

In the next section, 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 3, we apply the methodology to a traffic estimation scenario using real and simulated spatiotemporal data. We present our last scenario in Section 4, where we show how to make good data-buying decisions for predicting a person's future location.

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.
- We demonstrate how the algorithm behaves in three scenarios: targeted ad delivery, crowdsourced traffic information, and location prediction.

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. SCENARIO 1: HOME TARGETED ADS

Our first illustrative scenario is called "Home Targeted Ads" because it focuses on a business that wants to deliver ads to people who have homes within 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

The original versions of this paper were "On the Value of Spatiotemporal Information: Principles and Scenarios" published in the *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, (Nov. 2018) and "To Buy or Not to Buy: Computing Value of Spatiotemporal Information" published in *ACM Transactions on Spatial Algorithms and Systems 12* (Sept. 2019).

can reach efficiently. We will refer to this target region as $\mathcal{R}$. The region can be any closed region on the ground, such as the boundary around a particular area.

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 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. In this scenario, the buyer 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.

2.1. 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 buyer always has some uncertainty about the home location of the potential ad recipient.

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 $b_{22} \geq 0$.

Table 1. The payoff matrix for home targeted ads.

		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)

The values in parentheses are used for our experiments.

We assume the payoff matrix values are given or can be learned.[11]

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. In Aly et al.,[2] we give a method to compute this distribution based on time-stamped location measures, such as the ones a buyer would purchase. 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:

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

The advertiser would choose whichever alternative has the largest expected revenue:

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

2.2. 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 in Aly et al.[2] The buyer then computes the probability that the home is in the target region (Equation (1)) and the expected revenue $\mathbb{E}\left[V \mid L_1^n\right]$, as described above.

The buyer has the option of buying another location measurement L_{n+1}. The VOI can then be defined as the gain in revenue by receiving the $n + 1$th location $L_{n+1} = \ell_{n+1}$:

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

The location of this new point is unknown to the buyer, but it follows a distribution $P_{L_{n+1}}(\ell_{n+1})$. This distribution is the buyer's guess about where the unseen point L_{n+1} may be. We give a principled way to compute this in Aly et al.[2] It is based on experimental data about how a person's distance from home varies over the day. In the middle of the night, people are normally close to home, but they are normally farther away at noon. Because of the uncertainty surrounding the location of the new point, the buyer is reduced to computing the expected VOI. This comes from Equation 3, but it includes an expectation integral over $P_{L_{n+1}}(\ell_{n+1})$, which is the probability density of all possible locations of the new point. This expected VOI is $EVOI\left(L_{n+1} \mid L_1^n = \ell_1^n\right)$.

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

$$\mathbb{E}\left[\text{Profit} \mid L_1^{n+1} = \ell_1^{n+1}\right] = EVOI\left(L_{n+1} \mid L_1^n = \ell_1^n\right) - c_{n+1}. \qquad (4)$$

Here we assume that the potential ad recipients have 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 although 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.

If we assume zero expected profit for the buyer, Equation 4 can be rearranged to show a fair price for the location point as

$$c_{n+1} = EVOI\left(L_{n+1} \mid L_1^n = \ell_1^n\right). \qquad (5)$$

Note that the price is independent of the actual location of the data. However, as the seller knows the location, a deeper analysis could adjust the price based on location. However, this price adjustment could in turn convey extra information to the seller about the potential value of the point, that is, if it is near the seller's home.

2.3. Algorithm for decisions

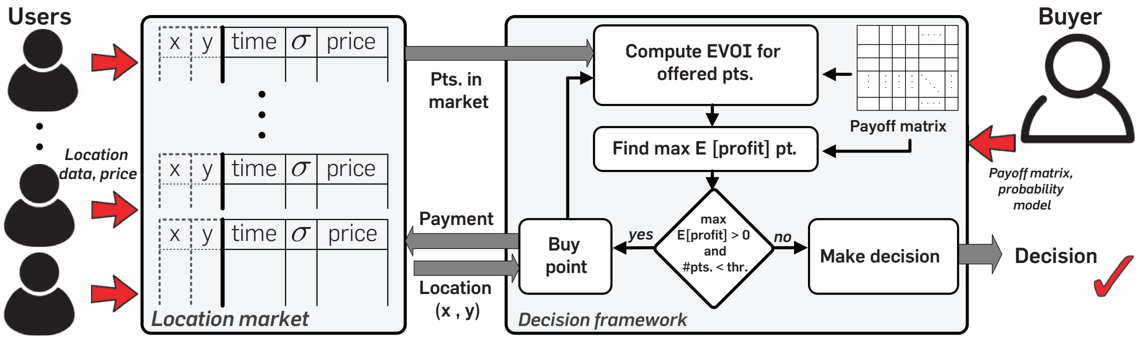
The final algorithm followed by the data requester, and illustrated in **Figure 1**, consists of repeated computations of the expected profit from Equation 4 over all the available points from the user. The buyer repeatedly buys the point with the maximum expected profit (Equation 4) as long as at least 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 decide whether or not to send the ad according to Equation 2.

2.4. Evaluation experiments

To evaluate the proposed decision framework, we used a GPS dataset of 66 participants living in Seattle, Washington, USA. The trajectories were collected for an average of 40.12 days ($\sigma = 24.43$) 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. We have 13, 14, and 18 users living in regions R_1, R_2, and R_3, respectively. To find the ground truth home location for each user, we leverage each user's full trajectory and the American Time Use Survey[12] (ATUS). ATUS points out that users are most likely to be at their homes at midnight. Thus, we apply density-based clustering (DBSCAN) on the user's time-stamped location trajectory. Then, the largest collection of data points (cluster) at midnight is identified as the user's home.[8]

We have compared the described methods 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. We call this technique "No points." 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

Figure 1. 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.



second technique "20 points." 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 a bounded amount of data. We refer to our proposed method as "VOI decision."

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 the 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* 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.

Results. To test our proposed framework for different payoff matrices, we created a payoff matrix with the values in parentheses as 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.

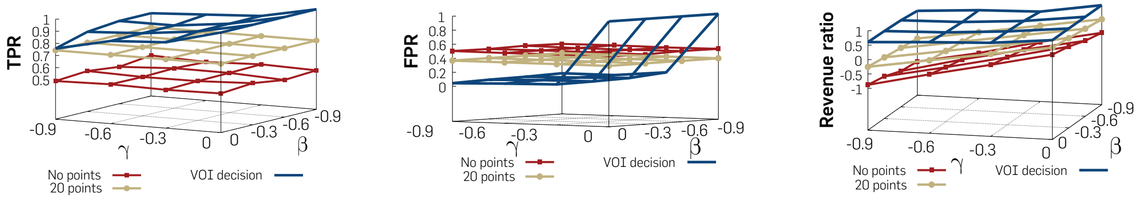
We compared the performance of our method to other methods in **Figure 2**. **Figure 2** shows the average results over the three regions for the different payoff matrices for a GPS point cost of 0.01. The two comparative methods ("No points" and "20 points") TPR and FPR are independent of the payoff matrix values, because they are neither considering the costs and benefits of buying points nor 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 2(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 2(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, whereas 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."

3. 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 here 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. Although 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 and (2) whether to buy a new point from travelers.

Figure 2. Home targeted ads (Scenario 1) experiment results using the proposed framework ("VOI decision") as compared to two other methods ("No points" and "20 points").



(a) True positive rate

(b) False positive rate

(c) Revenue ratio

3.1. 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 2. There are nine different possible cases: b_{rr}, b_{yy}, and 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}, and $b_{gg} > 0$. 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) although actually it is congested (red). Thus, these payoffs are less than b_{rr}, b_{yy}, and 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.[11]

To choose the congestion level from the noisy location data, we again employ decision theory principles.[11] 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 Aly et al.[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 2. This is as below for the decision "r", and the decisions "g" and "y" can be evaluated similarly.

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

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

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

In Aly et al.,[2] we discuss how the service provider computes $P_U(u)$ from individual time-stamped location measurements. The fundamental method is a Kalman filter, which gives the probability distribution $P_U(u)$ representing the speed estimate and its uncertainty as well as a distribution giving a prediction of the speed in the future, which gives a buyer an idea of what the next speed value will be. The next section discusses how to make decisions about the location points to buy.

Table 2. Payoff matrix for traffic state estimation.

		Actual traffic state		
		Red	**Yellow**	**Green**
Traffic	Red	b_{rr}	b_{ry}	b_{rg}
State	Yellow	b_{yr}	b_{yy}	b_{yg}
Decision	Green	b_{gr}	b_{gy}	b_{gg}

3.2. 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 formulate the decision as one of buying a new speed estimate. We leverage VOI 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 a Kalman filter 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 $\mathbb{E}[V|U_1^n]$, as described in section 3.1, 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. Because we are assuming the user is traveling at a locally constant speed, the Kalman estimate 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 $n + 1$th speed measurement $U_{n+1} = u_{n+1}$:

$$VOI\left(u_{n+1} \mid U_1^n = u_1^n\right) = \mathbb{E}\left[V \mid U_1^{n+1} = u_1^{n+1}\right] - \mathbb{E}\left[V \mid U_1^n = u_1^n\right]. \quad (6)$$

Hence, the expected value of information for the $n + 1$th speed is given by the expected value of (6):

$$EVOI\left(U_{n+1} \mid U_1^n = u_1^n\right) = \int_u VOI\left(u \mid U_1^n = u_1^n\right) \cdot P_{U_{n+1}}\left(u \mid U_1^n = u_1^n\right) \quad (7)$$

where $u \in \mathbf{R}$ and the integral is taken over the full domain of u.

The decision to buy the $n + 1$th speed will be based on whether the value of the point in expectation, that is, $EVOI\left(U_{n+1} \mid U_1^n = u_1^n\right)$, is larger than the cost of the speed (c_{n+1}):

$$\mathbb{E}[\text{Profit}] = EVOI\left(U_{n+1} \mid U_1^n = u_1^n\right) - c_{n+1}. \quad (8)$$

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

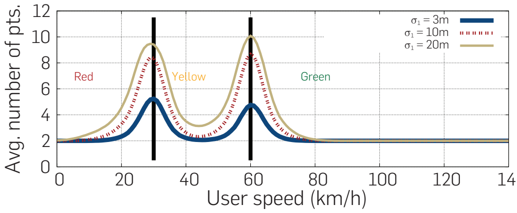
We give results of detailed experiments in the next section. To build intuition about these computations, we present results of a simple simulation experiment in Figure 3. For different vehicle speeds, Figure 3 displays the number of points purchased using the methodology. Note that we buy more points whose speeds are near the congestion level thresholds, that is, 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 more points as needed to resolve the speed uncertainty.

3.3. 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.

Effect of point cost and precision. Using simulated data, Figure 4 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, that is, $\sigma_l \in \{3m, 10m, 20m\}$ in parts a, b, and c of **Figure 4**, 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 exceeds their value, the buyer refrains from buying. Comparing parts a, b, and c of Figure 4, 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 Figure 4, 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.

Figure 3. Average number of points bought at different possible speeds for location points with an accuracy of 3 m, 10 m, and 20 m. The 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 = 3 s.

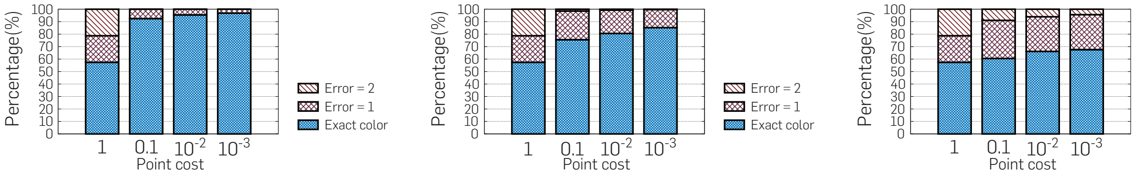


Comparative analysis. Figure 5(a) compares the performance of our framework to the mean window filter over different window sizes (baseline technique). The bars in Figure 5 show the error rates in the same way as shown in Figure 4. We also show the mean number of points purchased in these figures as small, black boxes. For relatively accurate location points (with precision σ_l varying uniformly at random from 3 to 20 m), Figure 5(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 size 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 5(b). This is better than the baseline technique with windows sizes 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. Figure 5 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 three bars. With a larger penalty for making a wrong decision, the framework buys more points and gives higher accuracy.

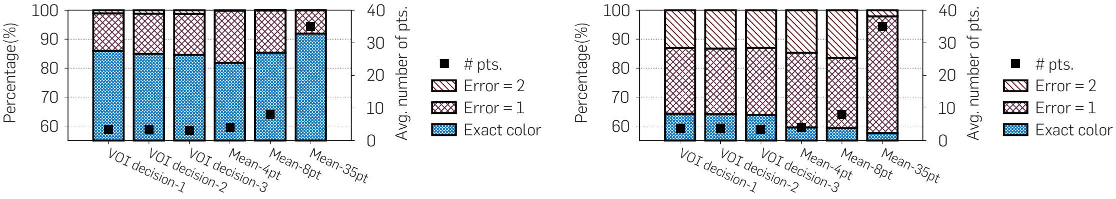
Validation experiments with real data. Using the same GPS data as we did for the experiments in Section 2.4, we extracted 20 traces from drivers on the I-90 interstate highway and State Route 520 in Seattle, WA, at different dates and times of day. All 20 traces had more than 8 points on the road in order to compare with a mean filter with window size 8. The traces' speeds varied from 10 to 133 km/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

Figure 4. Effect of point cost on congestion level/color decision accuracy while users are driving at different possible speeds (0–140 km/hr) for location points with a precision of 3 m, 10 m, and 20 m.



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

Figure 5. 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 3–20 m and 3–100 m. 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].



(a) precision σ_ι varies over 3*m*–20*m*

(b) precision σ_ι varies over 3*m*–100*m*

(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 predictions (with zero error) 90% of the time. In addition, the model recommends purchase of 50% fewer points as compared to the mean filter.

4. SCENARIO 3: LOCATION PREDICTION

A third scenario centers on location prediction. The buyer in this case is interested in the future location of someone. For example, the buyer may want to know if a person will be near the buyer's business place, which may prompt an ad delivery. A traffic authority may want to anticipate demand for the road network. In addition to introducing a new scenario, this section demonstrates a different form of the payoff matrix where the states and actions are continuous.

4.1. Location prediction

There are many existing techniques for predicting a person's location based on location history. These include methods based on a Markov model[4] and based on efficient driving and other cues.[7] We introduce a new technique here that produces a continuous probability distribution over future locations, which meshes with our mathematical framework.

Using a single historical point ℓ_i taken at time t_i, the predicted location for a future time t_f is ℓ_f, given by the normal distribution:

$$P_{L_f|L_i} \sim \mathcal{N}\left(\ell_i, \sigma_f^2\left(t_i, t_f - t_i\right)\right)$$

This implies that the normal distribution of future locations is centered around the measured location ℓ_i with a variance of $\sigma_f^2\left(t_i, t_f - t_i\right)$. This variance is a function of the current time t_i and the offset time into the future, $t_f - t_i$. Parameterizing the variance this way is intended to model the facts that (1) a person's future location is a strong function of their current location, especially for the near future, and (2) prediction uncertainty changes with the current time and the time offset into the future. We computed a tabular approximation of $\sigma_f^2\left(t_i, t_f - t_i\right)$ from the data of all our test users, discretizing both t_i and $t_f - t_i$ to 30-min intervals.

Predicting ℓ_f from multiple purchased points ℓ_1^n gives a mixture of Gaussians:

$$P_{L_f|L_1^n}\left(\ell_f\right) = \frac{1}{n}\sum_{i=1}^{n} g\left(\ell_f \mid \ell_i, \sigma_f^2\left(t_i, t_f - t_i\right)I\right) \qquad (9)$$

Here, $g(\mathbf{x}|\mu, \sigma I)$ represents a two-dimensional Gaussian, centered at μ with a diagonal 2x2 covariance matrix $\sigma^2 I$. The accuracy of this prediction technique is given in Section 4.4.

Computing the VOI depends on anticipating the location of the next purchased point, L_{n+1}. We make a direct prediction of the location of the next purchased point, which is conveniently given by Equation 9, notated as $P_{L_{n+1}|L_1^n}(\ell)$.

4.2. Payoff and decision

We introduce a generic, continuous payoff function that depends on the distance between the predicted and actual future locations. If the buyer decides that the predicted location is ℓ_f^*, but the actual location is ℓ_f, then the payoff for this decision is $b^2 - ||\ell_f - \ell_f^*||^2$. Here, b^2 is some base payoff for making an exact prediction, and the payoff decreases as the prediction error grows. This payoff function leads to a closed form for the expected revenue.

After some mathematics, detailed in Aly et al.[3], it becomes apparent that the expected revenue for deciding a future location of ℓ_f^* then simplifies to

$$\mathbb{E}\left[V \mid \ell_f^*\right] = b^2 - \frac{1}{n}\sum_{i=1}^{n}\left(2\sigma_f^2\left(t_i, t_f - t_i\right) + ||\ell_i - \ell_f^*||^2\right). \qquad (10)$$

The buyer will want to maximize expected revenue by choosing the best value for ℓ_f^*. Differentiating the expected revenue in Equation 10 with respect to ℓ_f^* and setting it to zero gives the optimal location prediction as

$$\ell_f^* = \frac{1}{n}\sum_{i=1}^{n}\ell_i.$$

This shows the predicted location is simply the mean of the already purchased location points. Although this is a very simplistic location prediction, the key is choosing which points ℓ_i to buy for making an accurate prediction, which we describe next.

4.3. Value of information

By making the optimal prediction above, the expected revenue from previously purchased points ℓ_f^n would be

$$\mathbb{E}\left[V \mid L_1^n = \ell_1^n\right] = b^2 - \frac{1}{n}\sum_{i=1}^{n}(2\sigma_f^2\left(t_i, t_f - t_i\right) + ||\ell_i - \frac{1}{n}\sum_{j=1}^{n}\ell_j||^2).$$

This shows that the expected revenue decreases with larger prediction variances and when the purchased points are more dispersed from their mean.

The VOI of an additional point ℓ^{n+1} is

$$VOI\left(\ell_{n+1} \mid L_1^n = \ell_1^n\right) = \mathbb{E}\left[V \mid L_1^{n+1} = \ell_1^{n+1}\right] - \mathbb{E}\left[V \mid L_1^n = \ell_1^n\right]$$
$$= \frac{2}{n}\sum_{i=1}^{n}\sigma_f^2\left(t_i, t_f - t_i\right) - \frac{2}{n+1}\sum_{i=1}^{n+1}\sigma_f^2\left(t_i, t_f - t_i\right) \qquad (11)$$
$$+ \frac{1}{n}\sum_{i=1}^{n}||\ell_i - \frac{1}{n}\sum_{j=1}^{n}\ell_j||^2 - \frac{1}{n+1}\sum_{i=1}^{n+1}||\ell_i - \frac{1}{n+1}\sum_{j=1}^{n+1}\ell_j||^2$$

Two of the main terms in the equation above are independent of ℓ_{n+1}, that is, $\frac{2}{n}\sum_{i=1}^{n}\sigma_f^2(t_i, t_f - t_i)$ and $\frac{1}{n}\sum_{i=1}^{n}||\ell_i - \frac{1}{n}\sum_{j=1}^{n}\ell_j||^2$. The other two main terms depend on ℓ_{n+1} and thus affect the choice of which is the best point to buy next. The first of these terms, $-\frac{2}{n+1}\sum_{i=1}^{n+1}\sigma_f^2(t_i, t_f - t_i)$, encourages buying points that have a small associated prediction variance, $\sigma_f^2(t_{n+1}, t_f - t_{n+1})$. The second of these terms, $-\frac{1}{n+1}\sum_{i=1}^{n+1}||\ell_i - \frac{1}{n+1}\sum_{j=1}^{n+1}\ell_j||^2$, encourages buying points that help reduce the dispersion of the purchased points.

4.4. Evaluation experiments

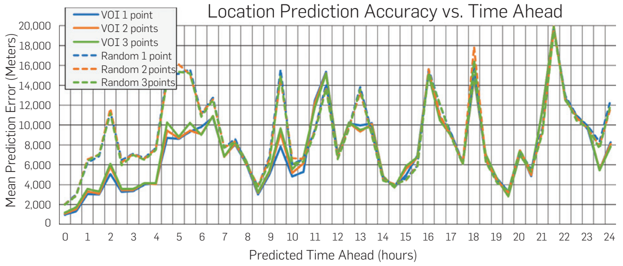
To test our prediction scenario, we used GPS data from the same 66 subjects as the ad delivery scenario described in Section 2.4. We used the temporal first half of each person's data to compute one set of prediction variances, $\sigma_f^2(t_i, t_f - t_i)$, that pertain to all subjects. We represented t_i as the amount of time since the day's previous midnight, discretized into 30-min intervals. The quantity $t_f - t_i$ represents the amount of time predicted into the future. We limited this to 24 hours and also discretized it to 30-min intervals.

For each subject, we randomly selected 100 test location points to predict from the temporal last half of their data. For each of these points, we randomly chose 20 prior points that were within our 24-hour prediction window as candidates for buying. With 66 subjects and 100 test predictions per subject, we tested our algorithm on 6600 different location prediction tasks.

Our primary test is to see if the algorithm is choosing good points to buy for making predictions. The next best point to buy is the one that maximized the expected VOI. As a comparison technique, we chose points randomly from the 20 available for each trial, repeating this 10 times for each of the 6600 prediction tasks.

Figure 6 shows the mean prediction error based on buying 1, 2, and 3 points. The solid lines show the VOI approach, and the correspondingly colored dashed lines show the random approach. From 0 to 7 hours into the future, the VOI technique has noticeably smaller error than the random technique, after which the two techniques are approximately equal in error. Predicting ahead 0–30 min, the VOI technique reduces prediction error by 54%, 47%, and 40%, respectively for 1, 2, and 3 purchased points. This large reduction in error shows that the VOI technique is much better at choosing which location points to buy for increased location prediction accuracy.

Figure 6. Using VOI to choose points to purchase is generally better than random choices in terms of prediction accuracy.



5. 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 by-product of the quantitative valuations, the methodology identifies which unseen data is likely the most valuable for the buyer. We considered three scenarios, home-targeted ads, traffic congestion inference, and location prediction, 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 work fills a gap in the pricing of location data and that the presented methods can help inform decisions by buyers and sellers of location data. C

References

1. Adar, E., Huberman, B.A. A market for secrets. *First Monday 8*, 6 (2001).
2. Aly, H., Krumm, J., Ranade, G., Horvitz, E. On the value of spatiotemporal information: principles and scenarios. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (2018), ACM, 179–188.
3. Aly, H., Krumm, J., Ranade, G., Horvitz, E. To buy or not to buy: Computing value of spatiotemporal information. *ACM Trans. Spat. Algor. Syst. 4*, 5 (2019), 22.
4. Ashbrook, D., Starner, T. Using gps to learn significant locations and predict movement across multiple users. *Pers. Ubiquit. Comput. 5*, 7 (2003), 275–286.
5. Cvrcek, D., Kumpost, M., Matyas, V., Danezis, G. A study on the value of location privacy. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society* (2006), ACM, 109–118.
6. Kanza, Y., Samet, H. An online marketplace for geosocial data. In *SIGSPATIAL* (2015), ACM, 10.
7. Krumm, J., Horvitz, E. Predestination: Inferring destinations from partial trajectories. In *International Conference on Ubiquitous Computing* (2006), Springer, 243–260.
8. Lv, M., Chen, L., Chen, G. Discovering personally semantic places from gps trajectories. In *CIKM* (2012), ACM, 1552–1556.
9. Micro, T. How much is your personal data worth? survey says..., 2015.
10. Monga, V. The big mystery: What's big data really worth?, 2014.
11. North, D.W. A tutorial introduction to decision theory. *IEEE Trans. Syst. Sci. Cybernet. 3*, 4 (1968), 200–210.
12. U. B. of Labor Statistics. American time use survey. 2016.
13. Staiano, J., Oliver, N., Lepri, B., de Oliveira, R., Caraviello, M., Sebe, N. Money walks: A human-centric study on the economics of personal mobile data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2014), ACM, 583–594.

Heba Aly (hebaaly@amazon.com), Amazon, Seattle, WA, USA.

John Krumm and **Eric Horvitz** ({jckrumm, horvitz}@microsoft.com), Microsoft Research, Redmond, WA, USA.

Gireeja Ranade (ranade@eecs.berkeley.edu), University of California, Berkeley, Berkeley, CA, USA.