PICK : A Framework for Choosing Data Collectors in Participatory Sensing.
Background

Sensor systems have been successfully used in a wide variety of application spaces.

- Improving agriculture procedures and environmental monitoring
- Providing insight into the social systems such as classrooms and workplaces.
- Classification of objects and events in military settings.

These existing sensor systems have some common characteristics:

- Created and run by Large Organizations (Universities, Companies, and Governments)
- Very Expensive and Custom Designed
- Infrastructure Based and Contain Planned Deployments
- Investigate one Pre-Defined Problem
Sensing by the People

What if sensor systems were owned and operated by the public sphere?

- What types of system architectures would be designed?
- What would people investigate with them?
- What are some of the challenges that would come up?

This is where...

**Participatory Sensing**

...comes in.
Participatory Sensing

We are developing systems that allow individuals to act as sensors - allowing them to contribute data for a particular initiative.

These participants use everyday mobile devices, such as cell phones, to gather data (image, audio, location) and transport them using existing cellular and WiFi (hot spots, meshes) networks for further processing.

Also participants could be sources of data by entering text to document an event or a situation.

We see participatory sensing impacting a wide variety of areas (public health, civic planning, cultural expression).
Participatory Sensing: Technology Implementation

Build on Embedded Network Sensing:

- Human-in-the-loop sensing
- Mobility - coordinated opportunistic sampling
- Emphasis on privacy (enabling users to selectively share information and change the fidelity of data)
- Group oriented auditing and analysis more common
- Leverage commodity devices (mobile phones) as sensor systems.
  - Imagers, microphones and positioning available.
  - USB, bluetooth to peer with external sensors
We see participatory sensing having three distinct interaction models.

- **Spontaneous Collection**
  Individuals notice a problem or see something of interest and directly contribute their data for public consumption.

- **“Making a Case” Data Gathering**
  An individual or group notice an issue that needs to be addressed. They setup a data collection “campaign” that is open for users to join. Members of the community join the campaign and the campaign director can coordination with the members involved to define, perform, and analyze the data collection process.

**My area of focus concerns Participatory Sensing for “Making a Case”**
Participatory Sensing to “Making a Case”

Specifically, I’m interested in campaigns that are run by pedestrians, outside and have a lifetime that ranges from a few hours to a couple of weeks.

Some of the example campaigns include the following:

- Noise Mapping
- Walkability Study
- Neighborhood Asset Mapping
- Noise Mapping
Campaign Model

Campaigns are systematic operations to gather a particular type of data and contain geographical and temporal constraints.

The campaign application model can be divided into the following steps:

- Creation
- Recruitment
- Execution (Sampling)
- Verification
- Auditing
- Publishing
Campaign Steps

**Creation**
Initiators create “campaigns” with specific requirements for sampling modality, space and time resolutions, life-time, and a cost.

**Recruitment**
Gather necessary individuals to run campaigns based on user interest, participation level, spatial/temporal coverage, reputation (as a data gatherer) metrics, and incentive.

**Execution**
- System listens to published locations of citizen-sensors.
- Trigger sampling based on spatial and temporal needs.
- Adjust windows, triggers via messages to achieve coverage.
- Pass samples to distributed analysts who verify/classify.

**Verification**
Human and automatic classifiers eliminate noise and contribute to updating the sampling strategy.

**Filtering / Selective Sharing / Auditing??**
Users have the ability to audit data or act as auditors for others.

**Publishing**
Pass samples marked for sharing for further verification/classification, and accept and post data to shared repositories and provide visualization of data.
Research Focus

The area of my research focus is on what can be referred to as the “initiation” steps in the campaign model.

In specific, I’m concentrating on the research challenges that come up based on considering the creation of campaigns and the recruitment of participants for campaigns.

Lets use “walkability” as an example campaign to illustrate the interaction model involved and also detail the research problems that come up.
Campaign Model

When a campaign is defined, the following constraints are typically placed on the data collection process:

- **Coverage Requirement (Space and Time)**
- **Life Time**
- **Sensing Modality**
- **Cost**

Thus, if we want to figure out who would be the best candidates to fulfill a particular campaign, the system needs to maintain a “campaign user profile” for a participation that has information about the **capabilities** in terms of sensors available by a particular user, the **availability** of the user to participate in terms of spatial and temporal contexts, the **reputation** of the user as a data collector, and the **incentive** cost associated with the user participating.
Research Contributions

Hence my research focuses on participation, mobility, and quality. Specifically, I address these issues through a combination of techniques that involve:

- **Context Attested Mobility Profiles**
- **Reputation Systems**
- **Incentive Mechanisms**

The contributions that I plan to make include the following:

- Context and mobility models that help access the availability of users for participating in sensing based on coverage constraints.

- Quality models for users derived from contributed data that to help evaluate the utility of participants as data collectors.

- Mechanism to combine the sensing capabilities, availability, quality, and incentive cost associated with users to figure out the best candidates for a particular campaign based on restrictions given by a designer.
Capabilities

When a participant registers with the system, he defines the **capabilities** of his sensor system and the **type of service** that his sensor system will offer.

Thus, capabilities can be defined as a combination of:

- the **type of sensors** that are available,
- the **exact characteristics** (resolution/sampling rate) of these sensors, and
- **restrictions** set forth by the user.
Capabilities

For instance, a user might not have an imager and thus this person would not be useful for a campaign that involves taking pictures.

Or participants might have varied types of the same sensor (a user with a Nokia n80 with a 3.3 MegaPixel Camera which takes 5 seconds to take a particular image compared to a user with a Nokia n95 with a 5 MegaPixel camera that can take shots in 3 seconds). In the case that high resolution images are required, the second user would be a better candidate.

Participant might not make certain sensors available at certain times as well.

This serves as the first step in refining which users could be candidates for a particular campaign since sensing modalities will be given along with the limits in terms of sampling.
Availability

Availability deals with whether a user meets the coverage requirements set by the campaign in terms of spatial and temporal constraints and also if the user is in a state to actually sense the phenomenon.

- Context Information
- Mobility Profiles
Availability : Context

Since the campaigns that I will be concentrating on involve pedestrian users when outside, determining:

- when users are actually outside (in particular the switch between indoors/outdoors and vice versa)

- the activity of the user (in terms of transportation mode) when outside

is important since the context change between indoors/outdoors is a good place to provide feedback and considering only outside activities helps in narrowing the focus of the classification problem.
The task of recognizing human activities from body worn sensors has received increasing attention in recent years. The majority of the research has focused on using wearable sensors to automatically classify the activity of a user. Previous work can be characterized by:

- Using accelerometer placed in two or more (up to 12) locations on the body.
- Using a single sensing location but with multiple modalities (accelerometer, temperature, humidity, sound, etc...).

But for our work, we want to create an activity classification system that is very convenient for users (can be incorporated into existing mobile devices such as cell phones, watches, etc... and is relatively low power so that it does not affect the usage of these devices).

So we want to create a system that consists of only one sensing unit with the least amount of sensing possible, is location agnostic, and works on many different users out of the box.
Context: Activity Classification

Types of activities we are interested in (only concerned with outside activities of users):

- Standing
- Sitting
- Walking
- Running
- Bicycling
- In Motorized Transport

These specific activities are considered because they are the common tasks that users perform while outside and also relate to the idea of a user is available for sensing.

Three locations for sensors which correspond to locations that people already carry devices are considered:

- Wrist (e.g. wrist-watch)
- Waist (cell phone on clip)
- Chest (cell phone worn around chest)
Context: Activity Classification

**What type of sensor(s) are needed?**

The most common sensor people use for activity classification are accelerometers. Others have included temperature, barometric, humidity, and audio sensors.

If we consider our activities that we are trying to classify we can immediately see that temperature and humidity would not be factors (walking vs running does not vary the temperature or humidity of the environment).

Audio might matter more but there are many exception cases (listening to the radio while sitting or listening to the radio in a car).

Barometric pressure and audio could be used but have lots of noise associated with them. (The same signals exist in many different scenarios - e.g. listening to the radio while sitting vs listening to the radio in a car or walking on different levels of a building).

Thus, we decide to use a **single 3 axis accelerometer** as our sensor for activity.
**Context: Activity Classification**

**Setup of Experiment**

We had a user wear a 3 axis accelerometer that sampled at 50hz in all three locations outlined earlier. The user performed all the different types of activities and the data was labeled.

**Window Size**

We take a **window of 1 second** (with an overlap of .5 seconds) as our period for classification.

- This value is validated by previous work in activity classification for our types of activities.
- Smaller value causes accuracy of classification to suffer and a higher value causes lag.

**Feature Space**

- Mean
- Standard Deviation
- Energy
- Correlation
- FFT (1-5Hz)
Context: Activity Classification

Approach

We use a two stage approach in which we try to classify using base level classifiers (kNN, Naive Bayes, C4.5 Decision Trees) to classify each window of data and then apply temporal based algorithm (state machines or HMMs) to ensure smoothness and allows for continuous tracking of activities.

The second stage enables classification output to incorporate history information to smooth out much of the sporadic errors that occur during the static classification and to encompass higher order states such as driving on bicycling.
Context: Activity Classification

Linear Discriminant Analysis (LDA)

The following shows a LDA transformation of the features space (Waist / 50 Hz). We can see that running, walking, biking, and standing are fairly different. But being in motorized transport, biking, and sitting have similar characteristics since a lot of times you are still and sitting while on a bike or in transport.
Context: Activity Classification

Which basic classifier is best?

Initial results show that the C4.5 decision tree classifier is actually very accurate.

<table>
<thead>
<tr>
<th></th>
<th>Chest</th>
<th>Waist</th>
<th>Wrist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>95.5</td>
<td>98.3</td>
<td>92.7</td>
</tr>
<tr>
<td><strong>C4.5 Tree</strong></td>
<td><strong>98.1</strong></td>
<td><strong>99.5</strong></td>
<td><strong>96.6</strong></td>
</tr>
<tr>
<td>KNN (6)</td>
<td>96.3</td>
<td>99.4</td>
<td>95.9</td>
</tr>
<tr>
<td>SVM</td>
<td>95.1</td>
<td>98.7</td>
<td>91.5</td>
</tr>
</tbody>
</table>

Can we sample at a lower rate?

From the results, we can decrease the sample rate for the accelerometer and still achieve high classification accuracy. This is especially appealing to save on power from not only the sampling process but also the feature calculation step.

<table>
<thead>
<tr>
<th></th>
<th>50 Hz</th>
<th>25 Hz</th>
<th>16 Hz</th>
<th>12 Hz</th>
<th>10 Hz</th>
<th>5 Hz</th>
<th>3 Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C4.5 Tree</strong></td>
<td><strong>99.5</strong></td>
<td><strong>99.5</strong></td>
<td><strong>99.4</strong></td>
<td><strong>99.4</strong></td>
<td><strong>99.2</strong></td>
<td><strong>96.2</strong></td>
<td><strong>91.9</strong></td>
</tr>
</tbody>
</table>
Context: Activity Classification

Do we need all the features?

Can sensing be location agnostic?

How does classification accuracy get affected by variety of users?

How much does the Second Stage Temporal Improvement make?
Recognizing if a user is currently inside or outside and also the transition points is important in providing feedback and also figuring out availability for sensing.

Unfortunately, there is not a “magic” inside/outside sensor. We can consider a few different possible sensing modalities to help with the process.

- GPS
- Motion
- Environmental (Temperature, Humidity, Barometric)
- Light
- Audio
- Image
- Network Characteristics

The design principles for the inside/outside detector are similar to the ones outlined for the activity classification. Ideally, we would like to use the least amount of sensors possible and have them in a single point for both power and convenience considerations.
Context: Indoors vs Outdoors

Our Approach

What we are really interested in is context annotated mobility profiles, there is one inherit assumption that helps us with this problem - we only care about values that we have location for. In our case, the localization system is GPS. Thus, we just need to eliminate the cases where GPS is reported but inside.

Since people generally have a large dwell time between being indoors and outdoors, we can use such wireless networks to help us out (# of Bluetooth / Wifi scene, the signal strength of GPS and GPRS systems etc...)

[Diagram showing periodic sampling of GPS, Bluetooth, Wifi, and GPRS as inputs to feature extraction, followed by a classifier trained offline to classify indoors or outdoors.]
Availability : Mobility

Campaigns will have a specific spatial and temporal constraint associated with them.

It is important to be able to use a participants past mobility patterns as a method to filter if that particular user would be suitable for a campaign.

How can we represent the mobility history of a user?

Do we need all location values or can we summarize the information?

Is there a way to place an emphasis on recent mobility patterns as opposed to older ones?
Availability : Mobility

Previous work in this space can be divided into three areas:

- MANETs
- Mobile QOS
- GPS/GPRS/WiFi Location Clustering

Mobile Ad hoc NETwork (MANET)

- Since most of these systems are not deployed on large scale, there has been a need to create simulation setups to test algorithms that can react to a node’s mobility.

- Mobility models are created to “generate” mobility patterns. They are based on variants of the random waypoint model where a node is specified with a certain constraints in terms of speed, direction, and duration for travel and then is simulated to generate a mobility pattern by randomly changing directions after a period of time.

- These models have gotten more complex recently (obstacles, history, geographically constrained), but they are still not adequate to summarize real world traces.

- Can’t go from raw location trace to parameters in these models (not robust enough and do not have any concept of exact point history, etc...)
Availability: Mobility

Mobile Phone Quality of Service

- Work concentrates on creating systems that enable predictive and adaptive bandwidth reservation for mobile phone users based on their mobility and cell phone usage.

- These models take a very microscopic view on mobility concentrating on figuring out which cells a user might travel to based on transitions patterns from previous cells, time spent in current cell, current speed and trajectory, and cell phone usage.

- Again, this setup will not work for location traces since summary information needs to be at a more macroscopic level. But we can take advantage of some of parameters they specify and the relationship between location entities in our model.

GPS, GPRS, WiFi Clustering

- There has been a lot of work on coming up with clustering algorithms based on observing people’s location traces.

- Typically, these systems work on very clean traces (user is careful about his data collection and tries to carry a GPS all the time), relies on having several users in a particular region, or using other information (map matching, geo-coding) to help in the process.
Availability : Mobility

Borrowing from this existing work, we want to create a mobility profiles for users based on GPS traces that we receive. Our goal is to take raw GPS data from users and generate a set of significant routes that are indexed by start/end time, area covered, travel time. Furthermore, the algorithm needs to be robust enough to work on sparse/unclean GPS data.

Our approach will be the following:

- Perform an initial filtering step to get rid of noise in the signal (obvious outliers based on spatial and temporal limits).

- Divide up a group of traces into “episodes” that represent a typical route. Thus coming up with the start and end locations for these routes.

Approach will be based on density based clustering considering since it has been proven as the most robust among the location clustering schemes.

- Generate a “route” based on this and come up with appropriate summary statistics for that route (spatial area covered - bounding box or polygon), travel time, day of week/time of day, etc... Also, aggregate routes that are similar to each other.

With these routes, we now have possible destinations and paths for users based on time of day and location starting from. Also, we can use this information for predicting who would be viable participants for a campaign based on spatial area covered by routes.
Reputation

So we have talked about availability so far.

What about the quality of data that is collected by participants? Is there a way to assign a “score” to a user as a data collector? If so what metrics are important to measure and how can they be summarized?

This is where the reputation module comes in and the goals of this subsystem is to help in:

- finding which users would be best for data collection
- filtering what data might be most useful.
Reputation : Metrics

**Timeliness** - Represents the latency between when a phenomenon occurs and when the sample is available for a data processing unit. It is affected by the sampling, communication, and auditing phases of sensing.

Example: Data collected by users might be batched processed periodically and routes of users could be adjusted based on data that was collected.

**Capture** - Describes the quality of a particular reading in terms of the ability in determining a particular feature - probability of inference. It is an attribute affected by both the specifications of the sensors used and the capturing process by the participant.

Example: Some users might not be good at taking images with a camera and take images that are blurry or underexposed.
Reputation : Metrics

**Relevancy** - Tells how well the sample describes the phenomenon that is sought for capture. Ranges from irrelevant (not related to the item of interest) to completely relevant (describes exactly the item of interest).

Example: Taking a picture of a car when asked to document sidewalks that exhibit serious problems. Taking a picture of a aesthetic crack vs severely damaged sidewalk.

**Responsiveness** - Describes whether the user responds to a directed, in situ sensing request.

Example: A user is at a certain location where sampling is especially valuable. Thus, the system alerts the user to take a particular reading. The user has the option to either respond to the request and perform the actual sensing or ignore it.
What is the magic number?

Now that we have specific metrics we can judge a user by, how do we come up with a user's reputation score.

Since the initiator of a campaign might have different weights assigned to the importance of each particular quality metric, a single score does not seem like a good solution. Instead, we want a framework that represents each of the different reputation metrics for a particular user.

How do we summarize a user’s score for a particular reputation metric?

We have many choice: average, last value, adaptive, exponential mean, beta, etc...

But we really want a framework that encompasses both the **stochastic uncertainty** (due to randomness of the system) and **epistemic uncertainty** (due to lack of knowledge about the randomness of the system). This is where **beta** proves the most useful.
Reputation : Metrics

Lets do a quick example of **stochastic** uncertainty compared to **epistemic** uncertainty. Say we are looking at the number of relevant to non relevant images a certain user has contributed.

If the user, submitted 8 relevant images and 2 irrelevant then his mean reputation score would be .8.

But this has a high epistemic uncertainty especially as compared to the case of a user submitting 80 relevant images and 20 irrelevant ones.

In the later case, we have a much lower epistemic uncertainty. We are much more confident about the user’s mean.
Reputation : Beta

Recently, there has been a lot of interest in using the beta distribution in representing reputation.

The Beta Distribution is a family of continuous probability distributions defined on the interval [0, 1] parameterized by two non-negative shape parameters, typically denoted by $\alpha$ and $\beta$.

Let's consider the case where we rate a particular piece of data as either satisfying a request ($r$) or not ($s$). Then, $\alpha = r + 1$ and $\beta = s + 1$.

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1} \, du}$$

$$= \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

$$E(X) = \frac{\alpha}{\alpha + \beta}$$

$$\text{Var}(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

where $\Gamma$ is the gamma function.
In the Beta distribution, $E(X)$ models the stochastic uncertainty while the distribution over all possible values of $X$ models the epistemic uncertainty.

$E(X) = .75$

$E(X) = .7941$
Reputation

Using the Beta formulation allows for some other interesting features.

- A reputation update only requires just updating the value of the two parameters alpha and beta as follows:
  \[
  \alpha_j^{\text{new}} = \alpha_j + r; \beta_j^{\text{new}} = \beta_j + s
  \]

- By formulating the problem using the Dirichlet process, we can even handle non binary updates with the same scheme.

- Aging can be easily incorporated so that recently obtained information can be given more weight. This is achieved by incorporating exponential averaging in the following way:
  \[
  \alpha_j^{\text{new}} = (w_{\text{age}} \times \alpha_j) + r; \beta_j^{\text{new}} = (w_{\text{age}} \times \beta_j) + s.
  \]
  where \(w_{\text{age}}\) represents the aging weight (range of 0 to 1).

- We can prevent identity change by assigning an initial low reputation.
Incentive

For users to participate in a particular campaign, there needs to be an incentive.

There are several different types of incentives that can be designed for participatory sensing. One can imagine having incentives based on:

- **capital** where credit is given based on participation.
- **barter** where information is exchanged for sensing.
- **social** where the framework is based on competing against another person or a group in terms of participation.
- **entertainment** where the sensing enables some type of pleasure or relaxation for the user.
- **altruistic** where the welfare of others or a certain initiative drives participation.
In terms of recruitment, the exact implementation nor the effectiveness of a particular incentive framework is not as important. Instead, we are only concerned with the “cost” associated with an incentive.

In terms of the interaction model, when a user creates a campaign, the person selects which incentives are available for participants to choose from and sets the exact parameters associated with each incentive that is available. Then, participants can pick the incentive (or set of incentives) that they will perform a particular campaign. At this point, a cost is associated with a user in terms of incentive.
Incentive

This cost might be nothing (in the case the participant chose an altruistic incentive model), set by the system in order to manage a particular incentive model (social, barter, or entertainment), or set based on the credit that will be given in terms of the capital model.

Thus, for the sake of recruitment, incentive simply becomes another factor that can be used to evaluate which users to use for a particular campaign.
Conclusion

Overall, we hope to use participatory sensing as a research space to develop technologies that empower individuals and communities to learn more about the world they live in and help in the process of transforming civic knowledge into civic action.

We believe that picking data collectors based on their sensing capabilities, availability, quality of sampling (reputation), and cost in terms of incentives is helpful in ensuring high quality participation.