PROJECT SUMMARY

CAREER: SUPERVISED LEARNING FOR INCOMPLETE & UNCERTAIN DATA

Traditional supervised machine learning algorithms rely on complete, accurate training data. Despite advances to lessen the need for data-point specific labeled training data, many applications require further flexibility in the type and accuracy of labels. Applications plagued with (1) sparse labeling, (2) uncertainty in the labels, and (3) lack of specificity in the labels cannot be sufficiently addressed using current supervised learning approaches. The goal of this CAREER proposal is develop the mathematical framework and algorithmic approaches for *multiple instance function learning* (for classification and for regression), which can simultaneously address these issues in the training data.

Supervised learning for global scene understanding by fusion of satellite imagery, road map data, and geo-tagged social media information will be developed and used to evaluate the proposed research. Analysis of satellite imagery is often conducted by coupling unsupervised learning methods with manual exploration. However, extremely large amounts of road map data (e.g., Google Maps or OpenStreetView) and social media information (e.g., geo-tagged photographs, video clips, social networking posts) are being continually generated and updated. This vast amount of geo-tagged information is being collected and stored, but the information is generally not combined using autonomous algorithms to provide a comprehensive understanding of a scene. What if Google Maps could "understand" satellite imagery? The data is available but the algorithms have yet to be developed. The enormous amount of continually updated social media and road map data can be used as sparse training data with varying levels of specificity and uncertainty to guide scene understanding given satellite imagery.

Using the proposed research, an interactive web application will be developed for dissemination and outreach to the general public. This web application will serve two roles: (1) provide an avenue for introducing concepts from machine learning and remote sensing to the public and (2) provide a method for generating ground-truth and data for the proposed research. This interactive web application will also be used, along with additional hands-on activities, to introduce summer high school students to machine learning and remote sensing concepts. Paired with the web application will be a research website in which data, code, publications and presentations will be shared. This website will host a forum for discussion between researchers along with a "high score board" for evaluating and comparing approaches on hosted data sets.

INTELLECTUAL MERIT

This award will develop flexible supervised learning approaches that can address incomplete and uncertain training data providing a framework for problems that have been previously unaddressed. Furthermore, a global scene understanding system for the fusion of satellite, road map and social media data will be developed using the proposed supervised learning framework. This work will provide a comprehensive understanding of an area using all available information rather than simply assimilating and co-registering information or analyzing data independently. Long-term research goals for the PI are to develop a framework for fusing large heterogeneous data sets for global scene understanding.

BROADER IMPACTS

Broader impacts of this work include advances to many applications where current supervised learning approaches are not applicable. A hands-on lab will developed for high school students through participation in an existing summer high school engineering camp. An interactive web application will be developed to demonstrate the proposed research and provide an avenue for the general public to participate in global scene understanding. A research website will be created for the dissemination of data, code, publications, presentations and tutorials on the proposed research. The research website will also provide a forum for research interaction allowing researchers to discuss methods, distribute and rank results, and share code.

CAREER: SUPERVISED LEARNING FOR INCOMPLETE & UNCERTAIN DATA

1 GOALS & OBJECTIVES

The applicability and effectiveness of a supervised learning algorithm for a particular problem hinges on the availability and accuracy of training data in the format assumed by the learning algorithm. A number of approaches have been developed in the machine learning literature to accommodate problems with varying levels of training data accuracy and availability. However, current supervised learning techniques have failed to address a number of applications. **Namely, many problems with (1) sparse labeling, (2) uncertainty in the labels, and (3) lack of specificity in the labels cannot be addressed using current techniques. During this CAREER, a mathematical framework and algorithmic approaches for *multiple instance function learning* will be developed to address these difficult supervised learning problems.**

The problems to be addressed will be of the following flexible form: Given a set of input data, $\{\mathbf{x}_i\}_{i=1}^N \in \mathbb{R}^D$, the goal is to learn a function, f, which maps input data into desired output points, $\{\mathbf{y}_i\}_{i=1}^N \in \mathbb{R}^d$. The function mapping will be determined using a supervised learning approach given N input training data points, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, which have been partitioned into M "bags," $\mathbf{B} = \{B_1, \ldots, B_M\}$, with associated bag-level labels, $L = \{L_1, \ldots, L_M\}$. These labels have three possible characteristics:

- **Sparse:** Most of the bags may be unlabeled resulting in a sparse set of bag-level labels.
- **Uncertain:** Each label may have a variable level and type of uncertainty. Some labels may be known to be extremely accurate whereas others may have high uncertainty and label error of some assumed form.
- **Lack of Specificity:** Labels may lack specificity. Labels may indicate a range of values that include the desired output value, $L_i = [\mathbf{l}_i, \mathbf{u}_i] : \mathbf{y}_i \in L_i$. Specificity differs from uncertainty in that a data point may have a label lacking specificity but have high confidence that the desired output falls in the range of values indicated by the label (i.e., high certainty but low specificity).

The development of a formalism and algorithmic strategy for addressing these difficult supervised learning problems would transform the many application areas of this type that are ill-addressed with current approaches. In order to accomplish this advance, the following **research objectives** will be achieved:

1. Investigate and develop a mathematical framework and associated algorithms for **Multiple Instance Function Learning** (MIFL) that addresses linear and non-linear classification and regression problems with varying levels and types of sparsity, uncertainty, and specificity in training labels
2. Study and apply the proposed framework and algorithms towards the fusion of satellite imagery, road map data and social media for global scene understanding.

This research will be conducted in conjunction with integrated education and outreach activities through which the following **education and outreach objectives** will be achieved:

1. Train graduate and undergraduate students in machine learning and large remote-sensing data analysis
2. Develop a high school summer lab experience to introduce machine learning and remote sensing concepts
3. Develop an interactive web application for outreach to the general public in order to demonstrate the proposed research and to introduce concepts from supervised machine learning and remote sensing
4. Create a project website for dissemination of data, code, publications, and presentations to the machine learning and remote sensing communities. Also, through this website, provide an avenue for interaction, evaluation, and collaboration on this research area.

2 MOTIVATION

Many applications are not adequately addressed with current supervised learning approaches. During this CAREER, a mathematical and algorithmic framework will be investigated and developed to bridge this gap by addressing difficult training problems plagued with incomplete and uncertain training data. Examples of applications where current methods fail but can be addressed with the proposed research are described here:

Mapping of Disease Spread: Florida orange groves have been recently decimated from *citrus greening*. Greening causes fruit to come in prematurely and with poor flavor. Often, symptoms are not observed by farmers for many years [4]. However, recent work in hyperspectral and multispectral image analysis has shown diseased trees may be able to be identified much earlier than when symptoms become apparent to farmers [29, 36]. Mapping of the percentage of infection in orange groves before significant progression of the disease is desired. To generate these maps, identification of the geo-positions of a few diseased trees by farmers can be used as training labels for the proposed supervised learning framework. More formally, the input data set consists of the hyperspectral pixels, $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{D \times N}$ which are partitioned into bags $B = \{B_1, \ldots, B_M\}$ based on spatial location. Each input pixel contains the spectral response across D wavelengths for the materials found in the corresponding pixel's field of view. The desired output values, $Y = \{y_1, \ldots, y_N\} \in [0, 1]^N$ indicate the percentage of crops in the corresponding pixel field of view that are diseased. Farmers can provide binary training labels for bags which contain pixels from a tree which have been positively identified as diseased. These labels (1) are sparse as most bags will be unlabeled since farmers do not know if these pixels contain responses from diseased trees; (2) lack specificity since the binary labels may indicate that *some* pixels in the labeled bags have a non-zero output value (i.e., $y_i \in (0, 1]$) but do not provide the percentage of diseased orange trees found within the corresponding area; and (3) are uncertain to the degree of the farmers ability to accurately identify and geo-locate diseased trees.

Automated Health Alerts: An interdisciplinary team with investigators from the fields of nursing, electrical and computer engineering, informatics, and social work at the University of Missouri have developed a network of sensors to monitor older adults on a daily basis [1]. Some of their recent research in eldercare technology has led to the development of a system for automated health alerts to clinicians when a person in their care may have a clinically relevant change in behavior shown in the sensor data [31, 45–47]. In their system, nurses may label the automated alert as "clinically relevant" or "clinically irrelevant," thus, assigning a binary label to sensor data collected over a period of time. These labels have been used to update the automated alert algorithms such that fewer irrelevant alerts are generated. However, further advancement of automated alert algorithms can be obtained from the already collected binary-labeled data using the proposed research. For example, algorithms that determine degree of a health alert's clinical relevance or identify which sensor data that contributes to clinically relevant alerts can be developed. This problem has labels which are (1) sparse as most sensor data does not generate a health alert and, therefore, does not get labeled by clinicians; (2) lack specificity since the labels are binary and does not provide the *degree* of the clinical relevance; and (3) uncertain to the degree of a clinician's ability to label an alert and the length of time associated with an alert.

Scene Understanding: Remote sensing is used for a variety of important applications including monitoring shorelines and coastal dynamics [39], canopy defoliation by insects [56], tree cover in dry-land ecosystems [60], landmine and explosive object detection [33, 63], and space exploration [14]. Satellite imagery (such as hyperspectral, multispectral, and high-resolution panchromatic imagery) is collected regularly over much of the globe to support these applications. Panchromatic imagery provides visible information about a scene and can be analyzed using image analysis and computer vision techniques to perform vision-based scene understanding. Remotely-sensed multi- and hyperspectral imagery are commonly analyzed using *spectral unmixing* approaches to perform the sub-pixel task of decomposing pixels into their

respective *endmembers* (i.e., constituent materials) and *abundances* (i.e., percentage of each material). The promise of spectral unmixing is the complete identification of material types and quantities in a scene. In previous work, these images have generally been analyzed independently or, if fused with other data, require careful collection, co-registration and/or manual interfacing between data types. Furthermore, spectral unmixing is an ill-posed inverse problem. Methods have been developed to estimate solutions by constraining the solution space using sparse assumptions, geometrical constraints, and other approaches [7]. These methods do not constrain the problem using scene-specific information but, instead, using broad assumptions.

However, more and more social media information are linked to positions on the ground with automated GPS tagging from mobile devices. Facebook and Twitter provide location of posts when available. Google Earth links satellite imagery and road maps with photographs provided by users. Google street view links road map data with visual imagery. Geo-tagged information is being continually collected, stored and updated. Although this information provides a wealth of scene-specific information that can be used to guide and constrain solutions, this data is generally not combined in a meaningful way to allow for an automated scene understanding. Instead, this data is loosely co-registered and available for manual interaction.

Scene understanding can encompass automated mapping of materials (e.g., proportion maps containing the percentage of each material found in every pixel), automated labeling of objects and regions, automated detection of targets, and others. Automated scene understanding of remotely-sensed data would allow for an immense number of follow-on advances such as more intelligent map searches (e.g., "Where are large grassy parks nearby?"), identify locations in satellite imagery given results of a map search (e.g., "What pixels in the satellite image correspond to the taco shack I have directions for?"), etc. The proposed research would provide advances to the supervised machine learning literature that can be used to fuse satellite imagery, road map, and social media data for scene understanding. Namely, satellite imagery and features extracted from these images would serve as the input data points, $\mathbf{X}$. The desired output values, $\mathbf{Y}$, would be the identification, labeling, and proportion estimates of materials in a scene at every spatial location. Label information, L, can be extracted from road map and social media data. These labels are (1) sparse since the overwhelming majority of pixels in the satellite imagery will be far from any road network or geo-tagged (and scene understanding-related) social media post; (2) uncertain since geo-tagged social media data may be extracted from a number of unreliable or out-dated sources; and (3) lack specificity because, for example, road network data vector indicates roadway (e.g. asphalt) location but lacks the percentage of area of asphalt coverage in every pixel and social media posts may indicate the existence of certain material types and various objects but lack information about their size and physical extent.

During this work, this final example application, the fusion of satellite imagery, road map and social media data for scene understanding will be implemented and used to evaluate the proposed research. This fusion will achieved through development of a supervised learning framework and associated algorithms.

Motivation for Education and Outreach Plan: Education and outreach plans are motivated by the desire to encourage greater participation and awareness of the proposed research problems and applications by the general public and researchers in machine learning and remote sensing. In particular, a central repository of ground-truthed data for the development of remote sensing algorithms does not exist. Remote sensing researchers have few common data sets and no common evaluation approach in order to compare methods. In order to address this deficiency, a research website to host ground-truthed data sets, evaluation results, share code, and promote interaction between researchers will be developed. In order to generate ground-truthed data sets for the website, an interactive web application will be developed that will allow users to generate social media-type labels for provided satellite imagery. The interactive web application will also serve as an entertaining way to introduce remote sensing and machine learning concepts to the general public and to summer high-school students during an engineering summer camp.

3 BACKGROUND AND RELATED WORK BY OTHERS

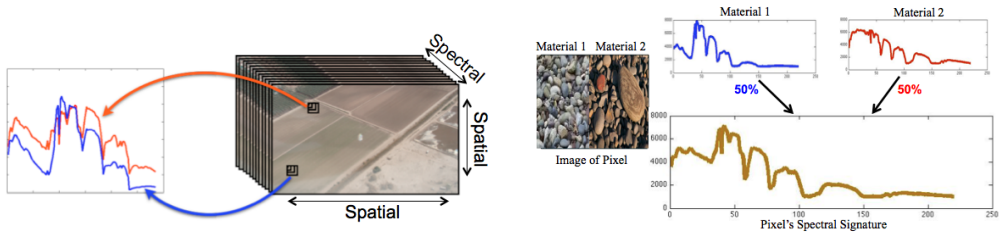
Traditional supervised learning techniques require each input training point, $\mathbf{x}_n$, to be paired with the desired output value, y_n. A number of the advances in the literature that have been made to lessen the need for complete training data of this form. For example, semi-supervised learning is a large area of investigation that has introduced methods to allow for ***sparse*** labeling of the training data. It can be viewed as hybrid of supervised and unsupervised learning approaches using both labeled and unlabeled data during training. Numerous approaches have been developed in this area and extensive overviews have been published in [17, 50]. Of particular note are the semi-supervised approaches that have been applied towards remote sensing applications [8, 13, 20, 32]. This large body of literature addresses the issue of ***sparse*** training labels; however, generally, semi-supervised learning approaches in the literature do not also simultaneously address label ***uncertainty*** and ***lack of specificity***.

Multiple instance learning (MIL) addresses the issue of sparse and uncertain labels of a particular form [9, 10, 19, 34, 48, 72]. Namely, in the multiple instance learning problem, there is an input data set $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ which has been partitioned into positive and negative *bags*, $B^+ = \left\{B_1^+, \ldots, B_{M^+}^+\right\}$ and $B^- = \left\{B_1^-, \ldots, B_{M^-}^-\right\}$. A bag is defined to be a multi-set of data points. A positive bag includes at least one point from a class of interest, the target class. In each positive bag, the exact number of data points belonging to the target class is unknown. Negative bags are composed entirely of non-target data points. Given training data of this form, the goal is to train an approach such that points can be accurately assigned class labels. Multiple Instance Learning has been applied to some remote-sensing problems [9, 10, 55]. Multiple Instance Regression (MIR) extends this concept to label bags with continuous-valued labels. The goal is to learn a regression function and identify which points in each bag contribute to the corresponding continuous-valued label [57, 59]. Also, a number of approaches merging and drawing relationships between semi-supervised learning and multiple instance learning have been developed [25, 44, 71, 73].

Current MIL and MIR approaches address a particular form of sparsity and lack of specificity. Namely, all points in negative bags are labeled (i.e., no sparsity) and their corresponding labels are specific. Points in positive bags, however, are not individually labeled; all that is known is that *at least* one point in each of the positive bag has a positive label (or contributes to the regression, in the case of MIR). The proposed Multiple Instance Function Learning (MIFL) will generalize the MIL and MIR approaches. MIFL will leverage the bag concept but address an additional order of complexity by allowing the bag labels themselves to lack specificity, be sparse, and uncertain.

Recent advances have also been made in terms of managing noisy, incomplete and uncertain data. Advances in database can address modeling, storing and retrieving data with various uncertaintes [2, 3, 12, 23, 41, 43, 51, 54]. Also, recent advances include the development of methods that address incomplete labeling for land use classification [38, 61], multi-instance multi-label classification [11], and noisy labels for supervised learning [52, 53]. The proposed work will study and extend the advances from these uncertainty models for integration within a supervised learning framework and extend current incomplete, noisy, multi-instance, and multi-label classification approaches to address general regression problems with unspecific labels.

Multi- and Hyperspectral Unmixing: During this research, scene understanding using the fusion of multi- and/or hyperspectral imagery, panchromatic imagery, social media and road map data will be used to evaluate the proposed functions of multiple instance approaches during development. Scene understanding will be achieved primarily through spectral unmixing of multi- and hyperspectral data with training labels of varying specificity and uncertainty that are extracted from the panchromatic, road map and social media data. Spectral unmixing provides the ability to perform *sub-pixel* analysis of multi- and hyperspectral imagery to map the locations and amounts of every material occurring in a scene providing for a sub-pixel understanding



(a) (b)

Figure 1: (a) A hyperspectral data cube has two spatial and one spectral dimension. Each spatial location in the image has a corresponding spectral signature that is used to identify the types and amounts of each material found at the corresponding location (b) Illustration of spectral unmixing according to the linear mixing model in which the spectral signature of a pixel containing two types are rocks are measured resulting in the measurement of the convex combination of each materials signature.

of an area.

Spectral unmixing generally is composed of two major tasks: *endmember estimation* and *abundance estimation*. Endmember estimation is the task of determine the spectral signatures of the pure materials occurring in the scene (i.e., what is in the scene?) and abundance estimation is the task of estimating how much of each material/endmember is found at each spatial location. The ability to perform sub-pixel analysis on satellite multi- and hyperspectral imagery is of great importance since the imagery has generally very low spatial resolution. To perform spectral unmixing, a mixing model must be assumed. In the spectral unmixing literature, both linear and non-linear models have been developed and have been determined to be accurate in different physical scenarios. The most commonly used model is the linear mixing model which assumes that every pixel is a convex combination of endmembers in the scene,

$$\mathbf{x}_n = \sum_{k=1}^{K} p_{nk}\mathbf{e}_k + \epsilon_n, n = 1, \ldots, N \tag{1}$$

where N is the number of data points, K is the number of endmembers (or materials) in a scene, $\mathbf{x}_n$ is the spectral signature of the n^{th} pixel, ϵ_n is an error/noise term, $\mathbf{e}_k$ is the spectral signature of the k^{th} endmember, and p_{nk} is the abundance of the k^{th} endmember in the n^{th} pixel. The proportions in this model are constrained to satisfy sum-to-one, $\sum_{k=1}^{K} p_{nk} = 1$, and non-negativity, $p_{nk} \geq 0, \forall n, k$, constraints. In this model, generally only the N data points are known during analysis, the remaining parameters including the number of endmembers, each of the endmember spectral signatures, and all of the abundance values need to be estimated. Figure 1 shows a hyperspectral data cube and illustrates the linear mixing model. Solving for these unknown parameters is an ill-posed inverse problem.

To constrain these ill-posed problems, many methods have been developed to estimate solutions by enforcing a number of broad assumptions about multi- and hyperspectral imagery. These broad, often inaccurate, assumptions include constraining the solution space by requiring endmember spectral signatures to be found in the input data [15, 16, 18, 21, 37, 58], minimum volume constraints [5, 24, 35], enforcing sparsity assumptions [22], or incorporating spatial information to enforce smoothly varying proportion values across neighboring pixels [42, 49]. These constraints provide an avenue for solving the ill-posed spectral unmixing problem but do so by enforcing constraints that are not derived from scene-specific information and are often not true in real data sets. These have been summarized by several overviews of spectral unmixing [6, 7, 26–28, 40]. MIFL will provide an approach for incorporating scene-specific constraints derived

from other data sources in a flexible and general manner not addressed by current approaches.

4 PRELIMINARY WORK AND RESULTS BY THE PRINCIPAL INVESTIGATOR

The PI has developed a number of algorithms related to the proposed research, in particular, the Functions of Multiple Instances (FUMI) approach.

4.1 Functions of Multiple Instances

Functions of Multiple Instances (FUMI) approach is a generalization of MIL. FUMI can be related to the MIL framework by treating each data point as a function of the elements of a positive or negative bag. FUMI learns target and non-target prototypes given a set of data points that are some unknown function of the target and non-target prototypes. Suppose there is a given data set, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, where each data point is some unknown function of prototypes, $\mathbf{x}_i = f(\mathbf{B}_i, \mathbf{P}_i)$ where $\mathbf{P}_i$ are the set of parameters for $\mathbf{x}_i$ and $\mathbf{B}_i$ is the "bag" of prototypes that contribute in a non-negligible way to the data point $\mathbf{x}_i$. Each training point $\mathbf{x}_i$ is given a binary label $l(\mathbf{x}_i)$ where $l(\mathbf{x}_i) = 1$ if $\mathbf{b}_T \in \mathbf{B}_i$ and $l(\mathbf{x}_i) = 0$ if $\mathbf{b}_T \notin \mathbf{B}_i$. After learning the target prototype using the binary-labeled training data, target detection can be performed on test data. A number FUMI-type algorithms have been developed by the PI, these are summarized here.

C-FUMI: The specific case that is considered by C-FUMI is that each data point is assumed to be a convex combination of target and non-target prototypes, $\mathbf{x}_i = p_{iT}\mathbf{e}_T + \sum_{k=1}^{M} p_{ik}\mathbf{e}_k$, where $\mathbf{x}_i$ is a data point, $\mathbf{e}_T$ is the target prototype, $\mathbf{e}_k$ is a non-target prototype for $k = 1, ..., M$ and p_{ik} is the weight of the k^{th} prototype in data point i and where the set of prototypes, $\mathbf{E}$, with non-zero weights for data points $\mathbf{x}_i$ define the bag $\mathbf{B}_i$. The weights are constrained to sum-to-one, $p_{iT} + \sum_{k=1}^{M} p_{ik} = 1$, and be greater than or equal to zero, $p_{iT} \geq 0, p_{ik} \geq 0$. If $l(\mathbf{x}_i) = 1$ then $\mathbf{x}_i = p_{iT}\mathbf{e}_T + \sum_{k=1}^{M} p_{ik}\mathbf{e}_k$ with $p_{iT} > 0$. If $l(\mathbf{x}_i) = 0$, then $\mathbf{x}_i = \sum_{k=1}^{M} p_{ik}\mathbf{e}_k$. The exact weight values for the training data are not needed, thus, these labels ***lack specificity***. For positively labeled points, the label indicates that the output value is in the range, $(0, 1]$. Therefore, C-FUMI [67] learns target and non-target prototypes given mixed training data without prior knowledge of the weights of the positively-labeled training points. The C-FUMI problem is solved by minimizing the objective function shown in (2) using alternating optimization.

$$\begin{aligned} F_{CF} &= (1-\mu)\sum_{i=1}^{N}\left|\left|\left(\mathbf{x}_i - l(\mathbf{x}_i)p_{iT}\mathbf{e}_T - \sum_{k=1}^{M} p_{ik}\mathbf{e}_k\right)\right|\right|_2^2 \\ &\quad + \frac{\mu}{2}\sum_{k=1}^{M}\sum_{j=1}^{M}||(\mathbf{e}_k - \mathbf{e}_j)||_2^2 + \mu\sum_{k=1}^{M}||(\mathbf{e}_T - \mathbf{e}_k)||_2^2 + \sum_{k=1}^{M}\gamma_k\sum_{i=1}^{N}p_{ik} \end{aligned} \tag{2}$$

The first term of this objective computes the squared error between the input data and the estimate found using the current prototypes. The second and third terms produce prototypes that are close to the data in feature space. The fourth term is a sparsity promoting term used to determine M, the number of endmembers needed to describe the input data. This objective is updated iteratively using alternating optimization on the endmembers and proportions.

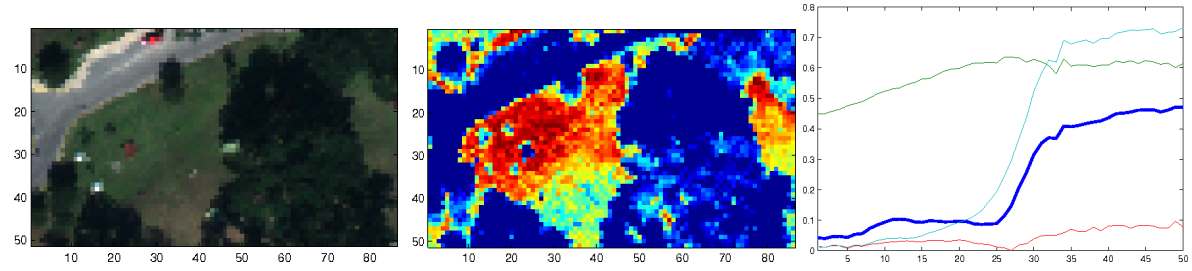
Weighted C-FUMI: The Weighted C-FUMI algorithm is an extension of the C-FUMI algorithm [68] in which the first term of the objective function is modified to incorporate a data-point specific weight, $F_{WC,T1} = (1-\mu)\sum_{i=1}^{N} w_{l(\mathbf{x}_i)}\left|\left|\left(\mathbf{x}_i - l(\mathbf{x}_i)p_{iT}\mathbf{e}_T - \sum_{k=1}^{M} p_{ik}\mathbf{e}_k\right)\right|\right|_2^2$. In the initial implementation, the value for $w_{l(\mathbf{x}_i)}$ is 1 when $\mathbf{x}_i$ has a negative label and is $\frac{\alpha N_n}{N_t}$ where N_n is the number of negatively labeled samples and N_t is the number of positively labeled samples. Therefore, if the parameter α is set to 1, then the weight on the target points is scaled such that the collection of target points has the same influence on the first term as the collection of non-target training points. The α value can be set to larger than 1 to emphasize

the importance of target training data over background data. Future work on this algorithm will investigate assigning weights based on **label uncertainty**.

Multi-class C-FUMI: The multi-class C-FUMI algorithm extends the Weighted C-FUMI algorithm [70] by estimating multiple class/target prototypes. For negatively labeled training data, the proportion value associated with any target prototype is constrained to be zero. For positively labeled training points, labels indicate which target prototypes have a non-zero weight associated with the data point. Therefore, the objective function for Weighted C-FUMI can be written as shown in Equation 3 where $l(\mathbf{x}_i, j)$ is 1 when $\mathbf{x}_i$ is in the j^{th} target class and 0 otherwise. The final term of the objective function can be interpreted probabilistically as a Gaussian prior on the weights associated to the target for each point from a positive bag.

$$F_{MC} = (1-\mu)\sum_{i=1}^{N} w_{\max_t(l(\mathbf{x}_i,t))} \left|\left|\left(\mathbf{x}_i - \sum_{k=1}^{T} l(\mathbf{x}_i,k)p_{it}\mathbf{e}_t - \sum_{k=T+1}^{M+T} p_{ik}\mathbf{e}_k\right)\right|\right|_2^2 + \frac{\mu}{2}\sum_{k=1}^{M+T}\sum_{j=1}^{M+T}||(\mathbf{e}_k - \mathbf{e}_j)||_2^2 + \sum_{k=T+1}^{M+T}\gamma_k\sum_{i=1}^{N}p_{ik} + \sum_{i=1}^{N}\sum_{t=1}^{T}\frac{1}{\sigma^2}l(\mathbf{x}_i,t)(p_{it}-1)^2 \tag{3}$$

Preliminary Results Experimental results are shown here from a hyperspectral image (HSI) collected over Gulfport, MS. Weighted C-FUMI was applied to the hyperspectral data. Figure 2(a) shows an RGB image (generated from the HSI data) of the area. Spectra were randomly sampled from around the scene and "grass" was selected as the target material. Each point was given a binary label indicating whether it contained some portion of grass. Weighted C-FUMI was applied and the target and non-target prototypes were estimated. After learning the prototypes, the proportions for every data point in the test data were computed. Figure 2(c) shows the prototypes found by the Weighted C-FUMI algorithm. The dark blue prototype corresponds to grass. Figure 2(b) shows the proportion map to the target grass endmember estimated by Weighted C-FUMI. As can be seen, the approach was able to accurately extract grass endmembers and the proportion of grass associated with each pixel given unspecific, multi-instance labels.



(a) (b) (c)

Figure 2: (a) RGB of hyperspectral scene collected over Gulfport, MS using a CASI system. (b) Estimated proportion values for the target "grass" material. (c) Prototypes estimated using weighted C-FUMI and binary-labeled training data

4.2 MCMC Sampling Approaches

The FUMI algorithm and associated extensions have been developed by relying on alternating optimization. In addition to alternating optimization, the PI has extensive experience in developing hierarchical Bayesian models and associated MCMC sampling techniques for parameter estimation. These methods include approaches for piece-wise convex unmixing which allow for more physically realistic spectral unmixing by allowing for multiple sets of endmembers (in which the number of endmembers are estimated

using a Dirichlet process prior) and allowing for endmember variability (in which endmembers are random variables and represented using either Gaussian or Beta distributions) [62, 65–67, 69]. Additionally, she has also applied Bayesian techniques for sea-floor characterization and underwater scene understanding using SONAR imagery [64]. In the proposed work, the algorithm development component will be conducting by considering a number of modeling and optimization techniques such as alternating optimization, hierarchical Bayesian models and MCMC sampling approaches, non-linear optimization, and others. The approach will be selected to well-suit the problem and algorithm in terms of computational efficiency and appropriateness.

5 RESEARCH PLAN

Investigation of the MIFL framework and associated algorithms for increased complexity and relaxation of training label requirements will be achieved through many stages. Each stage will be evaluated with a number of real and simulated data sets. Application to fusion of satellite, road map and social media data will also be developed incrementally as outlined in the following. The major milestones for this project are associated with each of the major research tasks listed below. Under each major research task are initial ideas and approaches that will be explored. During the proposed work, these initial ideas will be coupled with advances that are discovered to meet each major milestone.

5.1 Research Objective 1: Investigate and Develop a MIFL Supervised Learning Approach

The first research objective is to develop the multiple instance function learning approach. This will be accomplished through the following four major research tasks (and subsequent sub-tasks) outlined below.

Major Research Task 1: Develop Initial MIFL by Expanding FUMI to Bag-Level Labels: FUMI algorithms developed thus far introduce lack of specificity in the training labels. However, labels are assigned in FUMI for every training data point. The first stage in the development of the MIFL approach is to introduce bag level labels.

- **Research Task 1.1** *Introduce Bag Level Labels using Latent Variables and Expectation Maximization:* Input data points will be grouped into bags. Each bag will be associated with a label indicating a range of possible output values for every data point in the bag. This bag-level labeling will be introduced initially by incorporating latent variables associated with each data point in a positive bag. The resulting complete data log likelihood (with constant values dropped) is shown in (4).

$$F_1 = -\frac{1}{2}(1-\mu)\sum_{i=1}^{N}\left\|\mathbf{x}_i - z_i p_{iT}\mathbf{e}_T - \sum_{k=1}^{M} p_{ik}\mathbf{e}_k\right\|_2^2 - \frac{\mu}{2}\sum_{m=1}^{M}\|\mathbf{e}_m - \mu_{\mathbf{0}}\|_2^2 - \frac{\mu}{2}\|\mathbf{e}_T - \mu_{\mathbf{0}}\|_2^2 \quad (4)$$

where μ_0 is the global data mean and z_i are the latent variables that have been introduced into the C-FUMI objective function. Using, an expectation maximization (EM) optimization approach, estimates for the desired parameters will be obtained. To perform EM, we will first take the expectation of the complete log likelihood with respect to the z_i values. Note, the z_i values are unknown only for the positive bags. In the negative bags, z_i is fixed to 0. The resulting expectation will be as shown in (5).

$$\sum_{i=1}^{N}\sum_{z_i\in\{0,1\}}\left[p(z_i|x_i,\mathbf{E}^{(t-1)},\mathbf{P}^{(t-1)})\left(\frac{1}{2}(1-\mu)\sum_{i=1}^{N}\left\|\mathbf{x}_i - z_i p_{iT}\mathbf{e}_T + \sum_{k=1}^{M} p_{ik}\mathbf{e}_k\right\|_2^2 + \frac{\mu}{2}\sum_{m=1}^{M}\|\mathbf{e}_m - \mu_{\mathbf{0}}\|_2^2 + \frac{\mu}{2}\|\mathbf{e}_T - \mu_{\mathbf{0}}\|_2^2\right)\right] \quad (5)$$

where

$$p(z_i|\mathbf{x}_i, \mathbf{E}^{(t-1)}, \mathbf{P}^{(t-1)}) = \begin{cases} \frac{\exp\left\{\sum_{i=1}^{N} \left\|\mathbf{x}_i - z_i p_{iT}\mathbf{e}_T - \sum_{k=1}^{M} p_{ik}\mathbf{e}_k\right\|_2^2\right\}}{\sum_{c\in\{0,1\}} \exp\left\{\sum_{i=1}^{N} \left\|\mathbf{x}_i - c p_{iT}\mathbf{e}_T - \sum_{k=1}^{M} p_{ik}\mathbf{e}_k\right\|_2^2\right\}} & \text{if } x_i \in B^+ \\ 0 & \text{if } \mathbf{x}_i \in B^-, z_i = 1 \\ 1 & \text{if } \mathbf{x}_i \in B^-, z_i = 0 \end{cases} \tag{6}$$

Then, the $\mathbf{E}$ and $\mathbf{P}$ values will be found by maximizing (5) using an alternating optimization approach. This initial approach, however, may have issues with sensistivity and accuracy to small p_{iT} values. This may result from the form of the $p(z_i|\mathbf{x}_i, \mathbf{E}^{(t-1)}, \mathbf{P}^{(t-1)})$ because when $x_i \in B^+$, $p(z_i|\mathbf{x}_i, \mathbf{E}^{(t-1)}, \mathbf{P}^{(t-1)})$ will range from 0.5 to 1 due to the fact that $\left\|\mathbf{x}_i - p_{iT}\mathbf{e}_T - \sum_{k=1}^{M} p_{ik}\mathbf{e}_k\right\|_2^2 \approx \left\|\mathbf{x}_i - \sum_{k=1}^{M} p_{ik}\mathbf{e}_k\right\|_2^2$ if $p_{iT} \approx 0$. To address possibile accuracy issues with small values of p_{iT}, alternate forms of $p(z_i|\mathbf{x}_i, \mathbf{E}, \mathbf{P})$ will be considered.

- **Research Task 1.2** *Introduce Bag Level Labels using Noisy-Or Approach:* Another approach to incorporating bag level labels will be to extend the FUMI concept using Maron's MIL framework [34]. Alternatively, other MIL frameworks in the literature will be investigated, implemented and compared.

$$\begin{aligned} F_2 &= \left(1 - \prod_{i=1}^{N^+} \left(1 - pr(t|B_i^+)\right)\right) \prod_{j=1}^{N^-} pr(\neg t|B_i^-) \\ &= \left(1 - \prod_{i=1}^{N^+} \left(1 - \exp\left\{-\left\|\mathbf{x}_i^+ - p_{iT}\mathbf{e}_T - \sum_{k=1}^{M} p_{ik}\mathbf{e}_\mathbf{k}\right\|_2^2\right\}\right)\right) \prod_{j=1}^{N^-} \exp\left\{-\left\|\mathbf{x}_j^- - \sum_{k=1}^{M} p_{jk}\mathbf{e}_\mathbf{k}\right\|_2^2\right\} \end{aligned} \tag{7}$$

  One possible issue with alternate MIL frameworks is the difficulty to efficiently optimize parameter values within these complex forms. Initial implementations will make use of MCMC sampling techniques (and their corresponding convergence guarantees) whether any of these forms are viable. Then, if so, computationally efficient approaches will be pursued.

- **Research Task 1.3** *Introduce Multiple Class Labels into MIFL:* The previous two approaches consider the case in which there is one target class. Further work will be to extend this to multiple classes of interest. This can be initially achieved following an approach similar to Multi-class FUMI.

Major Research Task 2: Investigate Other Functional Forms for MIFL: The current FUMI form assumes a convex combination as the functional relationship between input data points and trained prototypes. Investigation and development of additional functional relationships will be conducted. ***As MIFL provides for a possibly extreme lack of specificity in the training labels, the selection of an appropriate functional relationship between input data points and desired output values is needed.*** The appropriate functional relationship is likely application dependent. Thus, to develop a general framework and associated algorithms for the proposed problem, a number of function relationships must be investigated and developed. All research tasks will be developed and evaluated with each of the functional relationships studied. In addition to the convex combination, the following will be investigated:

- **Research Task 2.1** *Linear Combination:* A functional relationship that will be considered is the linear combination (without non-negativity and sum-to-one constraints on the weight values),

$$\mathbf{x}_n = \sum_{k=1}^{K} w_{nk}\mathbf{e}_k + \epsilon_n, n = 1, \ldots, N. \tag{8}$$

The removal of constraints on the weights allow for more possible solutions and, thus, constraints to the solution space will need to be obtained for label information or ancillary data. A number of regularizing terms will be considered.

- **Research Task 2.2** *General Squared Distance:* The first term of all of the FUMI objective functions relies on a squared error between weighted prototypes and the input training data, $\left\|\mathbf{x}_n - \sum_{k=1}^{K} w_{nk}\mathbf{e}_k\right\|_2^2$. In this research task, general approaches in which any squared distance term between input data points and prototypes will be considered, $D_k^2(\mathbf{x}_n, \mathbf{e}_k)$. These distances may incorporate additional parameters such as weights, covariance matrices, etc. In this research task, kernel-based approaches will also be investigated. This will be an extension to transform the data to some high-dimensional Hilbert space and make use of the so-called "kernel trick."

Major Research Task 3: Extend for Varying Levels of Specificity in Labels: In the current FUMI form, positively labeled points have an identical level of specificity (namely, $y_i \in (0, 1]$), however, the MIFL problem needs to address labels with varying levels of specificity. Thus, each bag can have a unique level of specificity associated with its label. The following two ideas and approaches for introducing variable levels of specificity in the training labels will be explored.

- **Research Task 3.1** *Data point specific constraints:* Given each bag with a unique level of specificity, one approach is to incorporate data point specific constraints for the desired output values. This approach assumes a high certainty in the training labels (i.e., the range provided by the labels is assumed to be correct and without noise).

- **Research Task 3.2** *Data point specific prior distributions:* Another approach is to explore a hierarchical Bayesian framework. Prior distributions associated with each desired output value will be data point specific to incorporate the given level of specificity and the associated level and type of uncertainty.

Major Research Task 4: Extend for Varying Levels of Uncertainty in Labels: To allow for varying levels of uncertainty, the following will be explored.

- **Research Task 4.1** *Data Point Specific Weights:* The initial approach to incorporating varying levels of uncertainty will be to extend the Weighted C-FUMI approach such that each data point will have a unique weight determined based on the associated level of uncertainty. Uncertainty levels will be encoded in the label information. For example, data from reliable sources will have a higher level of certainty (and, thus, a higher associated weight value). A mapping from encoded uncertainty to weight will be initially set manually. Approaches to determine this from training will be investigated as well.

- **Research Task 4.2** *Data Point Specific Likelihood and Prior:* A hierarchical Bayesian framework will be explored such that each data point is independent and *uniquely* distributed. The likelihood associated with each data point will be determined based on the assumed form of the uncertainty for the data point. For example, some data points may be modeled with Gaussian error and high certainty (i.e., a Gaussian likelihood with a high precision term) and others may be modeled with other forms of the likelihood and/or parameter values. The form of the likelihood may be determined empirically based on errors from previous data sets (i.e., for a given data type and environment, Gamma distributed error is more appropriate). Furthermore, prior distributions associated with each desired output value can be data point specific to incorporate the given level of specificity and the associated uncertainty.

- **Research Task 4.3** *Dirichlet Random Effects:* In the previous research task, the data point-specific likelihood will be empirically determined and manually fixed. However, determining the appropriate probabilistic form can be extremely complex and difficult. An alternative approach will be to investigate Dirichlet Random Effects (DRE) to autonomously determine the clusters (and number of clusters) of various error types [30]. Dirichlet random effect approaches are well designed to identify groups of data with similar error/uncertainty types. Input training data points can be clustered using a Dirichlet Random Effects model. The various clustered will be then be modeled as determined by the DRE approach. The associated parameters will be estimated using a hierarchical Bayesian model and MCMC sampling techniques. Extensions to this will include investigation of variational approximations.

5.2 Objective 2: Investigate and Develop a Scene Understanding System using MIFL

The second research objective is to develop a scene understanding system that fuses remotely-sensed data, map data, and social media. This will be accomplished through the major research task outlined below.

Major Research Task 5: Scene Understanding:

- **Research Task 5.1** *Fuse Multi- and Hyperspectral Imagery and Road Map Data:* Initial development will occur using multi- and hyperspectral imagery and road map data. The PI has several groundtruthed hyperspectral data sets. Freely available Landsat and Quickbird imagery will be used as the input multispectral imagery. Road map data will be obtained from OpenStreetMaps. After assembling overlapping data sets, the MIFL algorithms will be applied for fusion. The road map data will serve as binary label information for detecting asphalt/road materials in a scene with no/low uncertainty. The binary labels provide for a lack of specificity in the labels. Thus, in this case, there is a single "target" class and the initial FUMI and MIFL algorithms can be directly applied. Following the initial implementation, features (e.g., texture, gradient-based, etc.) will also be extracted from the hyperspectral imagery to further analysis.

- **Research Task 5.2** *Fuse Panchromatic Imagery and Road Map Data:* In this research task, the fusion of road map and panchromatic imagery will be conducted. As in the previous task, the initial FUMI and MIFL algorithms can be directly applied as there will be only one "target" class with no uncertainty in the labeling. This research task will require extraction of features from the panchromatic imagery prior to fusion. A host of previously-published computer-vision based features will be investigated for this task.

- **Research Task 5.3** *Fuse Multi- and Hyperspectral Imagery with Surrogate Social Media Data:* In order to incorporate varying levels of specificity and uncertainty, *surrogate* social media data will be created for fusion with the multi- and hyperspectral imagery. Essentially, this will be simulated social media data. These labels will indicate the location of various materials, buildings, and objects in a scene. Several labels with varying levels of hand-assigned uncertainty and specificity will be generated for many multi- and hyperspectral image scenes. This step in the development of the scene understanding system is to develop and test the fusion of social media data in a controlled fashion.

- **Research Task 5.4** *Fuse Multi- and Hyperspectral Imagery and Web Application Collected Geo-tagged Data:* For more realistic social-media data, label information will be generated using the interactive web application developed for this project. The web application will allow visitors to input labels of a particular format (such that it can be understood by the proposed framework). These will serve as a more realistic surrogate for social media data. Various scenes will be displayed to a user and they will be able to identify regions of particular materials and proportion range. The data generated will be the labels supplied by the users and the geo-positions assigned by the users. This web-application will be in the

form of a game. As users advance in the game, the uncertainty associated with their labels will decrease. Additionally, the PI and her students will use the web application to generate their labels (for a baseline) as well.

- **Research Task 5.5** *Fuse Multi- and Hyperspectral Imagery, Panchromatic, Road Map, and Web Application Collected Geo-tagged Data:* This task will merge advances made from the previous four research tasks to make use of all available data and create a more complete scene understanding system.

- **Research Task 5.6** *Investigate Extraction of Social Media Labels:* As a final step, investigation into controlled methods for extracting material-related information from true social media data will be conducted. The data set used will be tags for imagery that are geo-located and posted on the Flickr website. This will be initially conducted by identifying a list of keywords for various materials (obtained using the web-application by requiring users input names for the materials they identify). Then, the image tags will be simply searched for these keywords. More sophisticated methods for social media data extraction will be investigated following this initial simple implementation.

- **Research Task 5.7** *Create Interactive Web Application:* An interactive web application will be developed for two reasons: (1) to introduce supervised learning and remote sensing concepts to the general public; and (2) to collect data for the proposed work. The web application will be structured as a game in which users advance with their ability to correctly label scenes. Several scenes will be carefully manually groundtruthed by the PI and her students. Small sub-images from these manually ground-truthed scenes will be shown to players who will then attempt to label and segment the scene. When shown the small sub-image of satellite imagery, the players will be told over what city/area the scene was collected. Points will be assigned based on accuracy. This is essentially a reverse version of the popular "GeoGuesser" web game (www.geoguesser.com). In order to collect data, the final stage will show an un-ground truthed scene which a user will attempt to label. The data will be the labels and geo-positions assigned by users. Between stages, the web application will introduce additional basic concepts from supervised learning and remote sensing. Also, a link to the "research website" will be provided such that visitors can learn more if desired. This game will be advertised using word-of-mouth and social media tools (e.g., asking for "likes" on Facebook). Using the campus IRB facilities as outlined in the Facilities document, IRB approval for this data collection will be obtained.

5.3 Evaluation Plan

Extensive experiments will be performed to evaluate the performance of the proposed supervised learning algorithms. Several synthetic data sets will be generated to verify each claimed attribute and the ability of the developed approaches to learn accurate classification and regression functions. Algorithms will also be tested with real remote sensing data sets from various applications: (1) sub-pixel target detection; (2) pixel classification, (3) spectral unmixing and (4) scene understanding. Two primary remote sensing data sets will be used. The first is a data set collected by the PI over Gulfport, MS. This data set includes aerial hyperspectral images, LIDAR data, photographs of the scene, and map data. This data set is carefully groundtruthed for a number of targets (including sub-pixel targets). This data set can also be augmented with multispectral and panchromatic satellite imagery obtained from Landsat and other freely available data sources. The second data set will be the one put together using the interactive web application through the course of this project. This data set will include satellite multispectral and panchromatic imagery as well as social media-surrogate data and groundtruth collected from the web application. We will also compare the performance of the developed supervised learning algorithms with that of similar existing algorithms using benchmark data. Performance for target detection will be measured in terms of reciever operating curves

(ROC) which plot probably of detection versus the number of false alarms per unit area. Performance of classification tasks will be evaluated in terms of confusion matrices and classification accuracy. Spectral unmixing and scene understanding tasks will be evaluated through a combination of qualitative assessment and quantitive assessment using ground-truthed scenes.

6 EDUCATION AND OUTREACH PLAN

The proposed research is well suited for integration with the PI's education plan. Students and the general public are very familiar and often eager to generate social media data and use the corresponding social applications. The proposed work provides the opportunity to make use of the excitement and access to social media applications to introduce students to remote sensing applications and provides hands-on methods to interact with the developed algorithms that are inherent in the research goals. Graduate, undergraduate and high school students will be engaged through research opportunities and hands-on lab activities. The general public will have the opportunity to be engaged by participating through the projects interactive web application. The following describes the proposed integrated education and outreach activities for this work.

6.1 Interactive Web Application

The interactive web application will be developed to introduce supervised learning and remote sensing concepts to the general public as well as collect data for the proposed work. The web application will serve as outreach through the general public by providing an entertaining way to be introduced to machine learning and remote sensing concepts and applications. As discussed previously, between stages, the web application will introduce additional basic concepts from supervised learning and remote sensing and provide a link to the "research website" will be provided such that visitors can learn more if desired. The basic concepts will be introduced through a number of methods between stages such as static pages that provide a definition or short description of basic concepts from machine learning and/or remote sensing (e.g., "What is supervised machine learning?" "What is a multi-spectral camera?") or very short instructional videos/demos. This game will be advertised using word-of-mouth and social media tools (e.g., asking for "likes" on Facebook).

6.2 Research Website

The PI will develop a website to disseminate data, publications, presentations, tutorials and code from the proposed research. The website will provide a venue for providing introductory tutorials on the research to the general public. Most significantly, to encourage research in this area, data will be groundtruthed and hosted on this site such that researchers in the field will have the opportunity to post results and related code. The best results will be showcased on a "high scoreboard." Results will be verified by requesting researchers supply the code used to generate the results such that others may verify. Only results with associated code will be posted on the scoreboard (other results will be simply listed on the site and labeled "unverified"). A forum for discussion will also be provided such that researchers may easily interact as a group.

6.3 High School Summer Program

Remote sensing applications and methods will be introduced to high school students through the University of Missouri's High School Summer Camp program. The program provides high school students a college experience with hands-on lab experiments, team design projects and competitions, and industry and campus tours. The program also encourages participation of underrepresented groups through the Diversity Scholars Program that provides underrepresented minority students scholarship opportunities to participate in the summer high school program. Students will be introduced to basic supervised learning and remote sensing concepts and the importance of these fields. The interactive web-application developed during this research will serve as a demo during the summer lab activity. Additionally, students will be given remotely-sensed multispectral and panchromatic imagery and asked to identify materials. Then, the students will visit these

areas in person on campus and repeat the web-application experiment in person. Students will then be asked to discuss the difficulty in performing scene analysis from satellite imagery and brainstorm possible ways they may attack the problem. Also, simple demonstrations on how to make prisms and external camera filters and their relation to multi- and hyperspectral imagery will be conducted.

6.4 Undergraduate and Graduate Training

Involvement of Graduate and Undergraduate Students: One graduate student and two undergraduate students will be funded through this project. These students will be trained in the areas of machine learning and remote sensing. Students will participate in weekly research seminars, literature reviews and all aspects of the proposed research. Graduate students will also be required to apply for graduate mentorship training as offered by the University of Missouri (and described in the Facilities document).

Integration of the proposed research into the courses taught by the PI:

Undergraduate course: The PI is the instructor for a junior-level course in computer engineering. This course is designed to give the students an introduction to software design in C and C++. The PI will reserve one lecture to integrate her research into this course. In particular, she will allow the students play the web-application. The students will be asked to perform two experiments. The first one involves labeling a small image manually. This experiment will illustrate to the students that computers can do this task, however, human can do it better. The PI will record the results for evaluation purposes. The second experiment involves repeating the first one using a much larger image database. The students will realize that manual labeling is a time consuming task and will appreciate the computer application. The students will then discuss programming constructs used in the development of the web application.

Graduate courses: The PI is teaching two graduate courses: supervised and unsupervised machine learning. The PI will use her proposed research to assign projects for these courses. First, depending on their research interests and the course, students can select to implement a feature extraction, clustering, or a supervised learning component. Then, they will combine this developed component with other available resources to develop a complete prototype system.

6.5 Evaluation of Education and Outreach Plan

Users of the web-application, visitors to the research website, and students in the high school program, graduate and undergraduate classes will be asked to fill in tailored surveys. These surveys will be regularly reviewed and used as a basis for evaluation of the outreach and education components.

7 BROADER IMPACTS

This work will include broader impacts that advance understanding, disseminate research and teaching materials, and provide advancement that will benefit numerous applications. These are summarized below.

- **Teach and train high school, undergraduate and graduate students:** Concepts will be introduced to high school students through a hands-on lab activity at a summer engineering camp and making lab materials available such that the activity can be conducted high school students elsewhere. Undergraduate and graduate students will participate in the research to learn the areas of machine learning and remote sensing. Undergraduate and graduate curriculum will be developed such that research topics are integrated.
- **Broader the participation of underrepresented groups:** The proposed research will fund one female graduate student through her Ph.D. studies. Every attempt will be made to support under-represented groups in the undergraduate students that participate.
- **Broadly disseminate the research and teaching materials to enhance scientific understanding and interaction:** An interactive web application will be developed to demonstrate the proposed research and

provide an avenue for the general public to participate in global scene understanding. A research website will be created for the dissemination of data, code, publications, presentations and tutorials on the proposed research. The research website will also provide a forum for research interaction allowing researchers to discuss methods, distribute and rank results, and share code.

- **Provide advancement to many applications that will benefit society:** The proposed research will advance the innumerable applications that are not adequately addressed by current supervised learning techniques. These will include applications related to scene understanding such as enhanced map/GIS searches, improved environmental monitoring, improved disease spread mapping, improved data fusion, improved precision agriculture mapping, and enhanced mapping for planetary exploration. The research will also advance areas in which only uncertain, incomplete and noisy training data are available.

8 Project Timeline

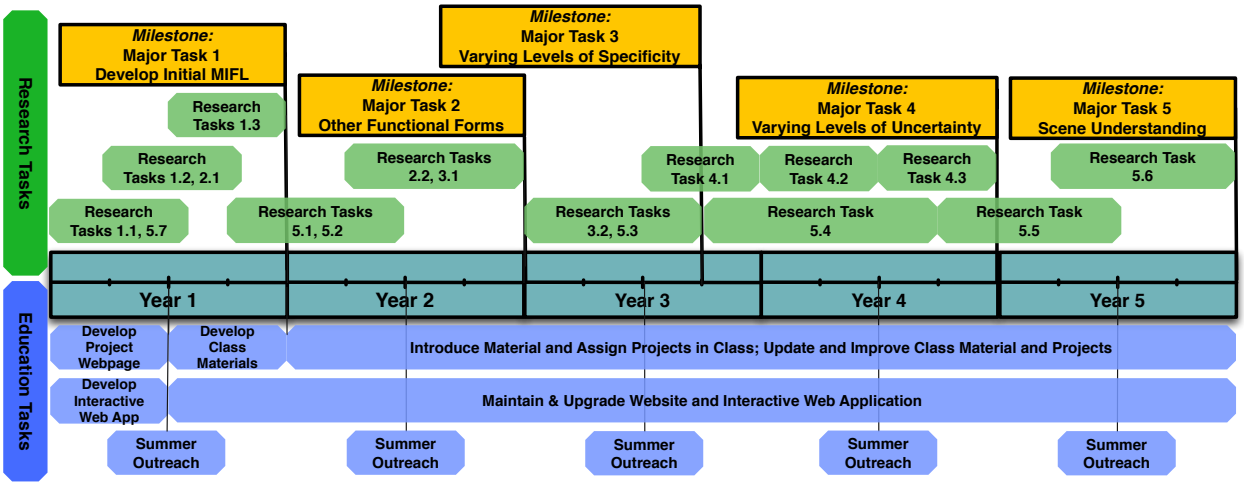


Figure 3: Timeline of Research and Education Activities

9 Long Term Goals of PI

The long term goal of the PI is build a robust and active lab in the area of scene understanding given non-visual data sets. The PI's background include analysis of data from a variety of sources (e.g., hyper- and multi-spectral, acoustic, ground penetrating radar, wide-band electromagnetic induction, LiDAR, SONAR, and Synthetic Aperture Radar data) in a variety of modalities (e.g., airborne, vehicle-based, hand-held, forward-looking, downward-looking) for a wide range of applications. The constant motivating theme in her research has been the need to autonomously understand a scene for a particular application using non-visual data. For many of these applications, solutions were brought about by coupling physics-based approaches and models with machine learning methods. Funding of this project will allow the PI to (1) begin research in the long term goal of global scene understanding using large heterogeneous data sets; (2) provide funding for a graduate PhD student such that she can focus on her thesis work; (3) provide several undergraduate students research experience and exposure to opportunities in graduate school; and (4) maintain a group of competitive students that can extract data, implement, test and improve the developed algorithms. Extensions of this work will include adapting methods for other data types given physics-based models.

10 Results From Prior NSF Support

The PI has no prior NSF support.

REFERENCES CITED

[1] Tigerplace: An innovative 'aging in place' community. *The American Journal of Nursing*, 113(1):68–59, 2013.

[2] P. Agrawal, A. D. Sarma, J.D. Ullman, and J. Widom. Foundations of uncertain-data integration. *Proceedings of the 36th International Conference on Very Large Data Bases*, 2010.

[3] P. Agrawal and J. Widom. Generalized uncertain databases: First steps. *Proceedings of the 2010 Workshop on Management of Uncertain Data*, 2010.

[4] L. Alvarez. Citrus disease with no cure is ravaging florida groves. *The New York Times*, May 9, 2013.

[5] M. Berman, H. Kiiveri, R. Lagerstrom, A. Ernst, R. Dunne, and J. F. Huntington. ICE: A statistical approach to identifying endmembers in hyperspectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 42:2085–2095, Oct. 2004.

[6] J. M. Bioucas-Dias and A. Plaza. An overview on hyperspectral unmixing: Geometrical, statistical, and sparse regression based approaches. *2011 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 1135–1138, July 2011.

[7] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379, Apr. 2012.

[8] P. Blanchart and M. Datcu. A semi-supervised algorithm for auto-annotation and unknown structures discovery in satellite image databases. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 3(4):698–717, 2010.

[9] J. Bolton and P. Gader. Spatial multiple instance learning for hyperspectral image analysis. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2010 2nd Workshop on*, pages 1–4, 2010.

[10] J. Bolton and P. Gader. Application of multiple-instance learning for hyperspectral image analysis. *Geoscience and Remote Sensing Letters, IEEE*, 8(5):889–893, 2011.

[11] F. Briggs, B. Lakshminarayanan, L. Neal, X. Z. Fern, R. Raich, M. G. Betts, S. Frey, and A. Hadley. Acoustic classification of multiple simultaneous bird species: a multi-instance multi-label approach. *Journal of the Acoustical Society of America*, 2012.

[12] Z. Cai, Z. Vagena, C. Jermaine, and P. J. Haas. Very large scale bayesian inference using mcdb. *Big Learning 2011 : NIPS 2011 Workshop on Algorithms, Systems, and Tools for Learning at Scale*, 2011.

[13] G. Camps-Valls, T. Bandos Marsheva, and D. Zhou. Semi-supervised graph-based hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(10):3044–3054, 2007.

[14] X. Ceamanos, S. Doute, Bin Luo, F. Schmidt, G. Jouannic, and J. Chanussot. Intercomparison and validation of techniques for spectral unmixing of hyperspectral images: A planetary case study. *IEEE Transactions on Geoscience and Remote Sensing*, 49:4341–4358, Nov. 2011.

[15] T.-H. Chan, W.K.-Ma, A. Ambikapathi, and C.-Y. Chi. A simplex volume maximization framework for hyperspectral endmember extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11):4177–4193, June 2011.

[16] C.-I. Chang, C.-C. Wu, C.-S. Lo, and M.-L. Chang. Real-time simplex growing algorithms for hyperspectral endmember extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 48(4):1834–1850, April 2010.

[17] O. Chapelle, A. Zien, and B. Scholkopf. *Semi-supervied Learning*. MIT Press.

[18] M. D. Craig. Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, 32(3):542–552, May 1994.

[19] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–17, 1997.

[20] N. Dobigeon, J.-Y. Tourneret, and Chein-I Chang. Semi-supervised linear spectral unmixing using a hierarchical bayesian model for hyperspectral imagery. *IEEE Transactions on Signal Processing*, 56(7):2684–2695, 2008.

[21] A. Ifarraguerri and C.-I. Chang. Multispectral and hyperspectral image analysis with convex cones. *IEEE Transactions on Geoscience and Remote Sensing*, 73(2):756–770, Mar. 1999.

[22] M. Iordache, J. Bioucas-Dias, and A. Plaza. Sparse unmixing of hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 49(6):2014–2039, June 2011.

[23] C. Jermaine, P. Haas, F. Xu, L. Perez, S. Arumugam, and M. Wu. The monte carlo database system: Stochastic analysis close to the data. *ACM Transactions on Database Systems*, 2011.

[24] S. Jia and Y Qian. Constrained nonnegative matrix factorization for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):161–173, Jan. 2009.

[25] Y. Jia and C. Zhang. Instance-level semi-supervised multiple instance learning. In *Proceedings of the 23rd national conference on Artificial intelligence*, volume 2, pages 640–645, 2008.

[26] N. Keshava. A survey of spectral unmixing algorithms. *Lincoln Laboratory Journal*, 14(1), 2009.

[27] N. Keshava, J. Kerekes, D. Manolakis, and G. Shaw. An algorithm taxonomy for hyperspectral unmixing. *SPIE Proceedings of Algorithms for Multispectral, Hyperspectral, and Ultraspectral Imagery VI. International Society of Optical Engineers.*, 4049, April 2000.

[28] N. Keshava and J. F. Mustard. Spectral unmixing. *IEEE Signal Processing Magazine*, 19(1):44–57, Jan. 2002.

[29] A. Kumar, W. Lee, R. Ehsani, L. G. Albrigo, C. Yang, and R. L. Mangan. Citrus greening disease detection using aerial hyperspectral and multispectral imaging techniques. *Journal of Applied Remote Sensing*, 6(1):063542–1–063542–22, 2012.

[30] M. Kyung, J. Gill, and G. Casella. Estimation in dirichlet random effects models. *Annals of Statistics*, 38(2):979–1009, 2010.

[31] Y. Li, T. Banerjee, M. Popescu, and M. Skubic. Improvement of acoustic fall detection using kinect depth sensing. *Proceedings of the EEE 2013 International Conference of the Engineering in Medicine and Biology Society (EMBC)*, Jul. 2013.

[32] W. Liao, R. Bellens, A. Pizurica, W. Philips, and Y. Pi. Classification of hyperspectral data over urban areas using directional morphological profiles and semi-supervised feature extraction. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(4):1177–1190, 2012.

[33] D. Manolakis, D. Marden, and G. A. Shaw. Hyperspectral image processing for automatic target detection applications. *Lincoln Laboratory Journal*, 14(1):79–116, 2003.

[34] O. Maron and T. Lozano-Perez. A framework for multiple-instance learning. *Neural Information Processing Systems*, 10, 1998.

[35] L. Miao and H. Qi. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):765–777, Mar. 2007.

[36] A. Mishra, R. Ehsani, D. Karimi, and L. G. Albrigo. Potential applications of multiband spectroscopy and hyperspectral imaging for detecting HLB infected orange trees. *Proc. Fla. State Hort. Soc.*, 122:147–151, 2009.

[37] J. M. P. Nascimento and J. M. Bioucas-Dias. Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, 43(4):898–910, Apr. 2005.

[38] S. Newsam, B. Edmunds, and Andrew Pierce. Pedseg: Gps tracks as priors for overhead image segmentation. *The ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2011.

[39] J. E. Pardo-Pascual, J. Almonacid-Caballer, L. A. Ruiz, and J. Palomar-Vzquez. Automatic extraction of shorelines from landsat tm and etm+ multi-temporal images with subpixel precision. *Remote Sensing of Environment*, 123:1–11, August 2012.

[40] M. Parente and A. Plaza. Survey of geometric and statistical unmixing algorithms for hyperspectral images. *2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, June 2010.

[41] L. L. Perez, S. Arumugam, and C. M. Jermaine. Evaluation of probabilistic threshold queries in mcdb. *SIGMOD Conference 2010*, 2010.

[42] A. Plaza, P. Martinez, R. Perez, and J. Plaza. Spatial/spectral endmember extraction by multidimensional morphological operators. *IEEE Transactions on Geoscience and Remote Sensing*, 40(9):2025–2041, September 2002.

[43] Y. Qi, R. Jain, S. Singh, and S. Prabhakar. Threshold query optimization for uncertain data. *Proc. of the ACM International Conference on Management of Data (SIGMOD)*, 2010.

[44] R. Rahmani and S. A. Goldman. Missl: Multiple-instance semi-supervised learning. In *In Proceedings of the International Conference on Machine Learning (ICML*, pages 705–712. ACM Press, 2006.

[45] M. Rantz, S. D. Scott, S. J. Miller, M. Skubic, L. Phillips, G. Alexander, R. J. Koopman, K. Musterman, and J. Back. Evaluation of health alerts from an early illness warning system in independent living. *Computers, Informatics, Nursing*, 31(6):274–280, 2013.

[46] M. Rantz, M. Skubic, R. Koopman, G. Alexander, L. Phillips, K. Musterman, J. Back, M. Aud, C. Galambos, R. Guevara, and S. Miller. Automated technology to speed recognition of signs of illness in older adults. *Journal Gerontological Nursing*, 38(4):18–23, 2012.

[47] M. Rantz, M. Skubic, S. J. Miller, C. Galambos, G. Alexander, J. Keller, and M. Popescu. Sensor technology to support aging in place. *Journal of the American Medical Directors Association*, 14(6):386–391, 2013.

[48] V. C. Raykar, B. Krishnapuram, J. Bi, M. Dundar, and R. B. Rao. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *Proceedings of the 25th international conference on Machine learning*, pages 808–815. ACM New York, NY, USA, 2008.

[49] D. M. Rogge, B. Rivard, J. Zhang, A. Sanchez, J. Harris, and J. Feng. Integration of spatial-spectral information for the improved extraction of endmembers. *Remote Sensing of Environment*, 110:287–303, 2007.

[50] M. Seeger. *Learning with labeled and unlabeled data (Technical Report)*. University of Edinburgh, 2001.

[51] P. Sen, A. Deshpande, and L. Getoor. Prdb: managing and exploiting rich correlations in probabilistic databases. *VLDB JOURNAL*, 18, 2009.

[52] V. S. Sheng. Simple multiple noisy label utilization strategies. In *2011 IEEE 11th International Conference on Data Mining (ICDM)*, 2011.

[53] V. S. Sheng, R. Tada, and A. Atla. An empirical study of noise impacts on supervised learning algorithms and measures. In *Proceedings of the 7th International Conference on Data Mining*, 2011.

[54] W. N. Sumner, T. Bao, X. Zhang, and S. Prabhakar. Coalescing executions for fast uncertainty analysis. *Proceedings of the International Conference of Software Engineering*, 2011.

[55] P. Torrione, C. Ratto, and L.M. Collins. Multiple instance and context dependent learning in hyperspectral data. In *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS '09. First Workshop on*, pages 1–4, 2009.

[56] P. A. Townsend, A. Singh, J. R. Foster, N. J. Rehberg, C. C. Kingdon, K. N. Eshleman, and S. W. Seagle. A general landsat model to predict canopy defoliation in broadleaf deciduous forests. *Remote Sensing of Environment*, 119:255–265, April 2012.

[57] K. L. Wagstaff, T. Lane, and A. Roper. Multiple-instance regression with structured data. In *Data Mining Workshops, 2008. ICDMW '08. IEEE International Conference on*, pages 291–300, 2008.

[58] J. Wang and C.-I. Chang. Applications of independent component analysis in endmember extraction and abundance quantification for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 44(9):2601–2616, September 2006.

[59] Z. Wang, L. Lan, and S. Vucetic. Mixture model for multiple instance regression and applications in remote sensing. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(6):2226–2237, 2012.

[60] J. Yang, P. J. Weisberg, and N. A. Bristow. A landsat remote sensing approaches for monitoring long-term tree cover dynamics in semi-arid woodlands: Comparison of vegetation indices and spectral mixture analysis. *Remote Sensing of Environment*, 119:62–71, April 2012.

[61] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. *The ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2010.

[62] A. Zare. *Hyperspectral Endmember Detection and Band Selection using Bayesian Methods*. PhD thesis, University of Florida, 2009.

[63] A. Zare, J. Bolton, P. Gader, and M. Schatten. Vegetation mapping for landmine detection using long-wave hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 46:172–178, Jan. 2008.

[64] A. Zare and J. T. Cobb. Sand ripple characterization using an extended synthetic aperture sonar model and mcmc sampling methods. In *IEEE OCEANS*, 2013.

[65] A. Zare and P. Gader. Endmember detection using the dirichlet process. *Proceedings of the IEEE: 19th International Conference on Pattern Recognition*, pages 1–4, Dec. 2008.

[66] A. Zare and P. Gader. PCE: Piece-wise convex endmember detection. *IEEE Transactions on Geoscience and Remote Sensing*, 48(6):2620–2632, Jun. 2010.

[67] A. Zare and P. Gader. Pce: Piecewise convex endmember detection. *Geoscience and Remote Sensing, IEEE Transactions on*, 48(6):2620–2632, 2010.

[68] A. Zare and P. Gader. Multiclass subpixel target detection using functions of multiple instances. pages 804811–804811–5, 2011.

[69] A. Zare, P. Gader, J. Bolton, S. Yuksel, T. Dubroca, and R. Close. Sub-pixel target spectra estimation using functions of multiple instances. In *3rd IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2011. In Press.

[70] A. Zare, P. Gader, and G. Casella. Sampling piecewise convex unmixing and endmember extraction. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(3):1655–1665, 2013.

[71] B. Zeisl, C. Leistner, A. Saffari, and H. Bischof. On-line semi-supervised multiple-instance boosting. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1879–1879, 2010.

[72] Q. Zhang and S.A. Goldman. EM-DD: An improved multiple-instance learning technique. *Advances in Neural Information Processing Systems*, 2:1073–1080, 2002.

[73] Z. Zhou and J. Xu. On the relation between multi-instance learning and semi-supervised learning. In *Proceedings of the 24th international conference on Machine learning*, pages 1167–1174. ACM, 2007.