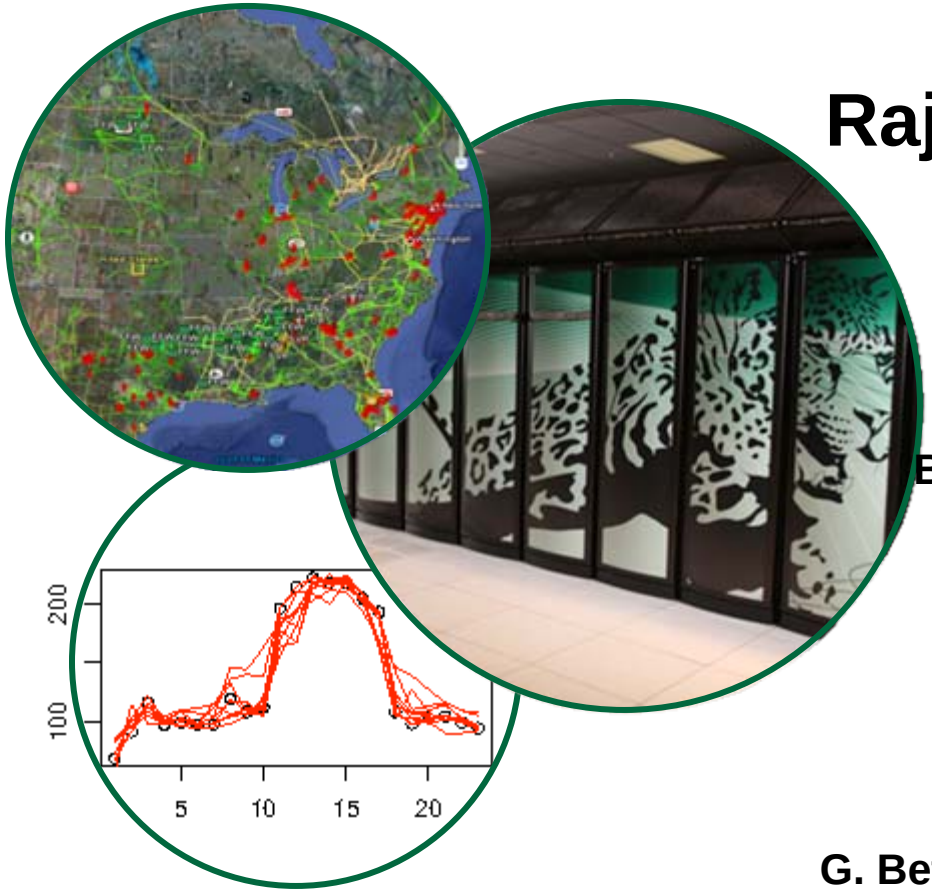
Large Scale Spatiotemporal Data Mining for Bioenergy



Raju Vatsavai, CSED, ORNL

Collaborators

B. Bhaduri, E. BrightC. Symons, A. Ganguly,

Cheriyadat, G. Jun (CSED);

F. Hoffman (CSM); T. West (ESD)

S. Gleason (MSSE)

A. Filippi (Texas A&M)

B. Doorn (NASA);

G. Bethel, R. Tetrault, E. Kunz, G. Serbin (USDA);

S. Shekhar(UMN), M. Crawford (Purdue), J. Ghosh (Texas).

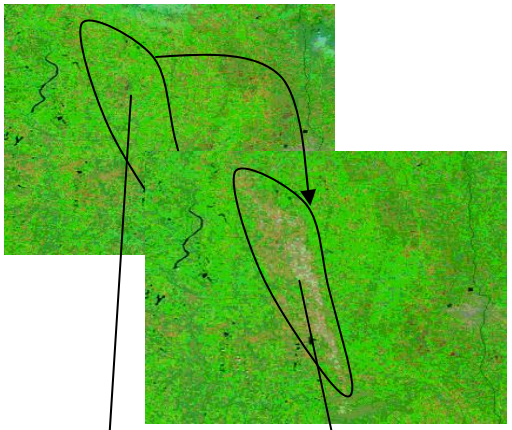
Outline

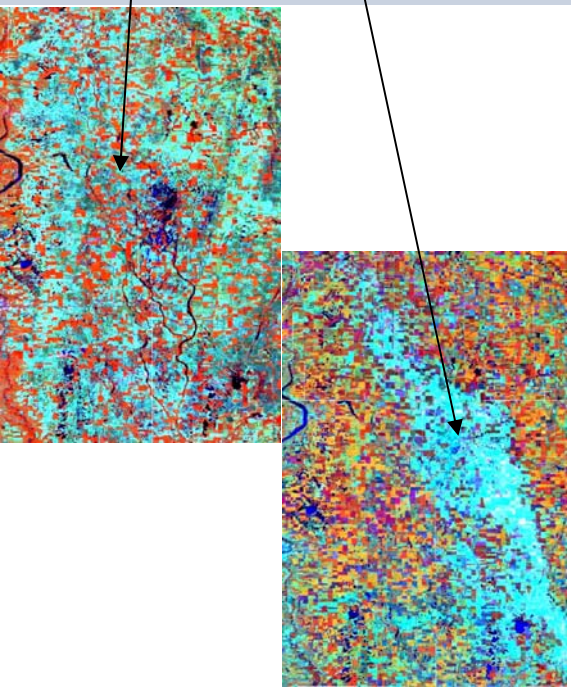
- **Background and Objectives**
- **Overall Architecture**
- **Biomass Monitoring Framework**
- **Change Detection**
- **Classification**
- **Clustering**
- **Conclusions**

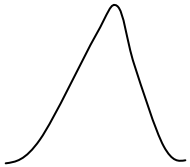
Background and Objectives

- **Increasing emphasis on Feedstocks (DOE/OBP, "Biomass: Multi-Year Program Plan," March 2008) leads to significant land use changes in US and other countries**
  - **Existing programs are not sufficient for biomass monitoring and characterizing changes**
- **Develop a spatiotemporal data mining framework for:**
  - **Continuous monitoring of biomass using coarse resolution spatiotemporal data (Change/No Change)**
    - **Reduce false positives by integrating radiometric observations with climate variables (e.g., temperature, rainfall), agricultural statistics (USDA-NASS)**
  - **Finer biomass information extraction using multispectral data (Characterize Changes)**
    - **Extract crop type and condition information from fine spectral and spatial resolution images guided by high-temporal but coarse spatial/spectral resolution images (MODIS -> AWiFS -> NAIP)**
    - **Scale these algorithms for large geographic regions**

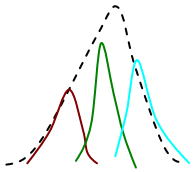
Overall Architecture







Aggregate Class
(Agriculture)



Fine (Sub-)Classes
(Corn, Soy, Wheat)

Biomass Monitoring

MODIS (250m, 2b, 1d)
- **Coarse-spatial resolution**
- **High-temporal resolution**
- **Good for regional, global monitoring, but not ideal for crop identification**

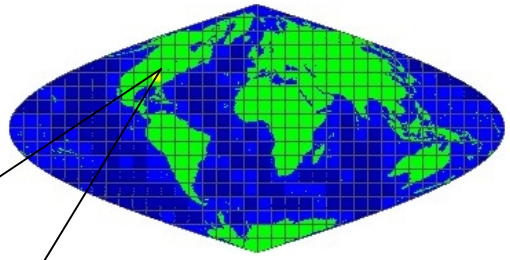
Change Vs.
No Change

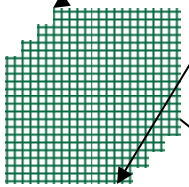
Fine-resolution Information Extraction

AWiFS (56 m, 4B, 5d)
NAIP (1m, 4b, 1y)
- **Moderate-to-High spatial**
- **Moderate temporal**
- **Used for crop type and condition extraction**
- **High training and computational requirements**

Characterize
Changes

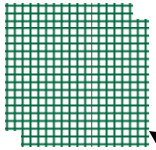
Biomass Monitoring Framework





MODIS (4800x4800)
3 Bands, 250m, 8-days
2000-2009
H11V04, MOD09Q1
(LP DACC)
27GB; 432 products

FTP-Pull



DVDs

AWiFS
(12,300x12,00
4 Bands, 56m
May-Sept. 200
Iowa, (USDA)
130 Products

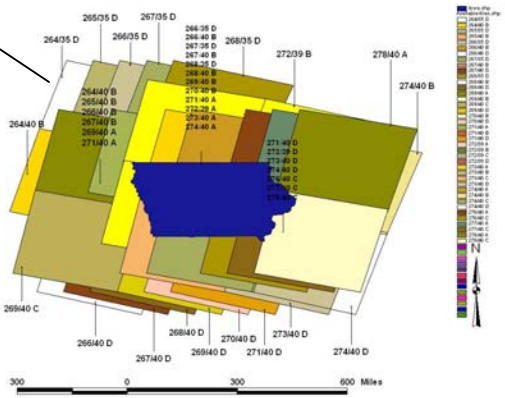
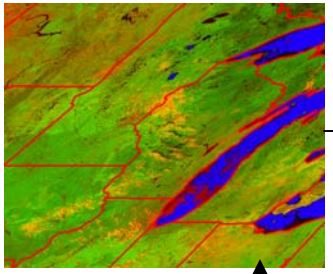
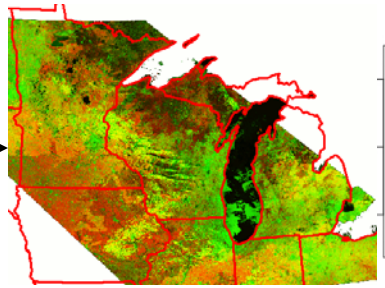


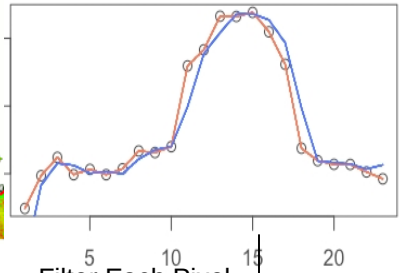
Image &
Ancillary Data



ISIN Projection



UTM Projection



Filter Each Pixel

Pre-processing
•Reprojection
•Atmospheric
•Filtering

Change Detection
•Time Series Based
•Time Series Prediction
•Multidimensional Image Based
•Unsupervised Clustering

Characterize
Changes
•Phenology-based
•Type-based

GE Visualization

Change Detection [4]

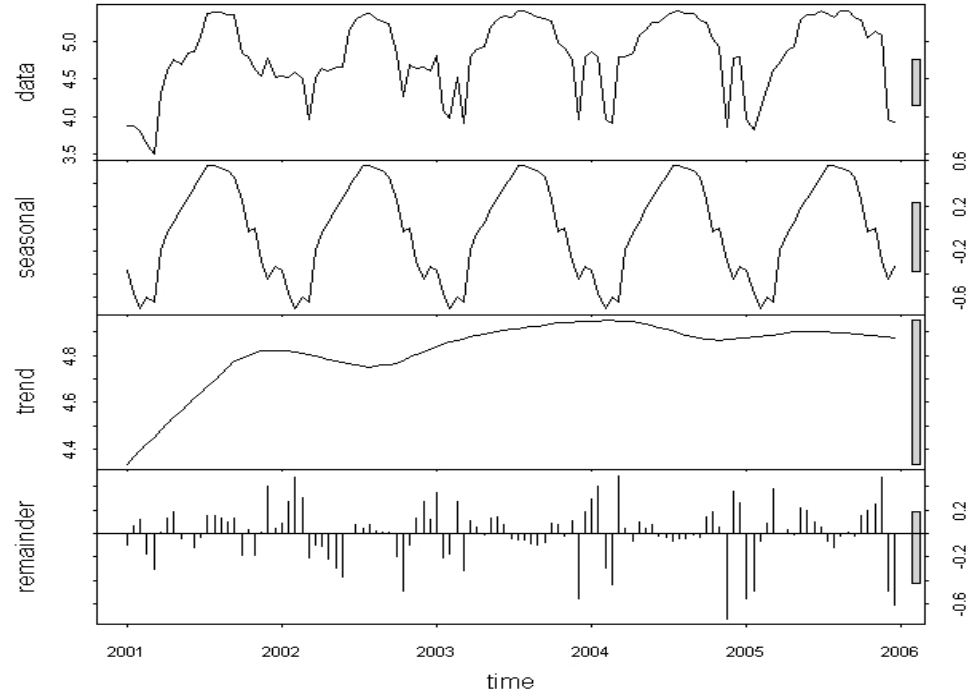
- **Unsupervised**
  - **Simple differencing: Change Vector Analysis [1]**
  - **Significance and Hypothesis Tests [2]**
- **Supervised**
  - **Several Classification Techniques**
- **Time-series [3]**
  - **AR, MA, ARIMA**

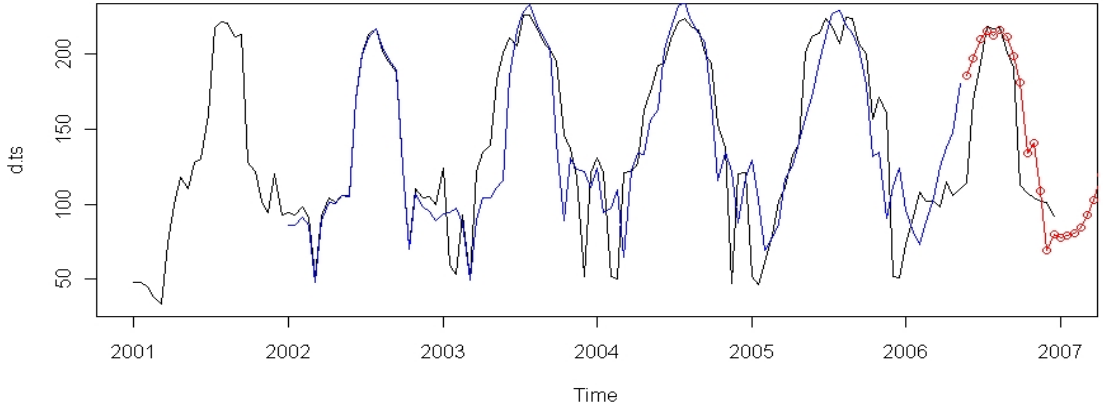
[1] L. Bruzzone and D. F. Prieto, “An adaptive semiparametric and context-based approach to unsupervised change detection in multitemporal remote-sensing images,” IEEE Trans. Image Processing,, Apr. 2002.
[2] T. Aach and A. Kaup, Bayesian algorithms for adaptive change detection in image sequences using Markov random fields,“Signal Processing. vol. 7, pp. 147–160, Aug. 1995.
[3] Z. Jain, Y. Chau. “Optimum multisensor data fusion for image change detection,” IEEE Trans. Syst., Man, Cybern., vol. 25 1995.
[4] R. J. Radke, S. Andra. “Image Change Detection Algorithms: A systematic survey.” IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 14, NO. 3, MARCH 2005

Time Series Based Change Detection

Basic algorithm

- **Learn from past observations, that is, build a model that fits to all previous observations (NDVI time series)**
- **Using the model**
  - **Predict NDVI at next time step**
- **Determine if there is a change**
  - **Compare predicted value with observed (current NDVI image) value**
  - **If the difference is within a threshold, no change, else "possible change"**
- **Challenges**
  - **Which model**
  - **What is the appropriate threshold**

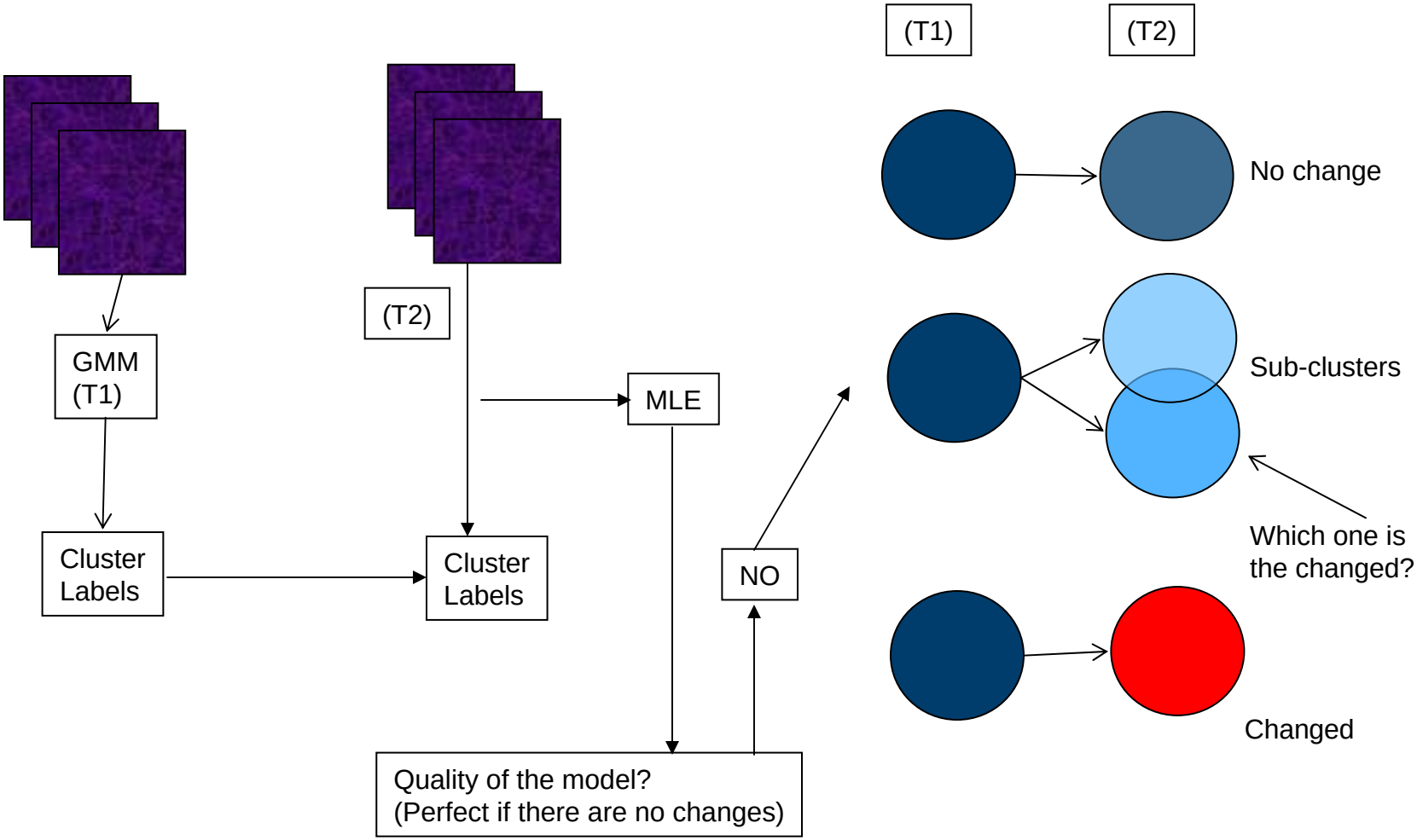




Multidimensional Image Based Change Detection

- **Each band (feature) is one temporal (NDVI) image**
- **Each image is either a single band (one time period), one year (all time periods), or only a growing season (e.g., May – Sept.)**
- **Detect changes between two images**
- **Useful in detecting changes**
  - **Between time intervals (not necessarily consecutive)**
  - **Between seasons**
- **We developed a new unsupervised clustering based change detection technique**

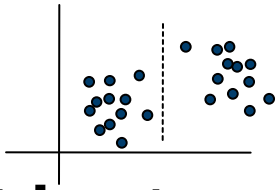
Basic Algorithm



Clustering – General Principles

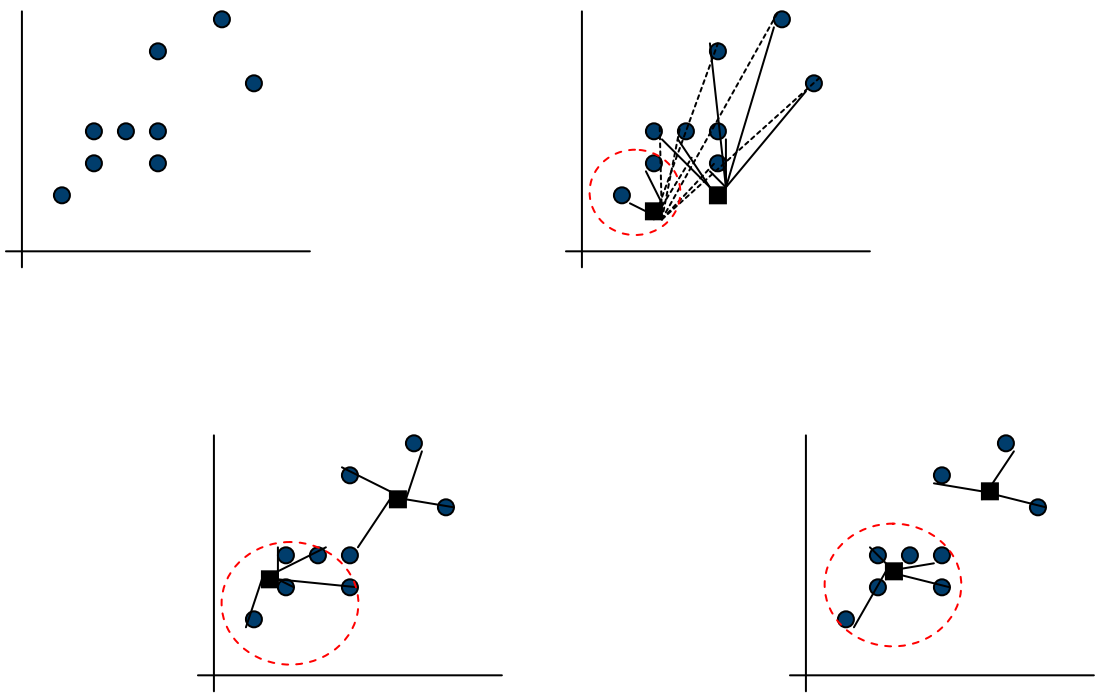
Pre-Processing → Feature Selection → Clustering Algorithm → Validation, Interpretation

- **Pre-processing – standardization of variables, outliers removal, …**
- **Which features are best?**
  - **Find smallest feature subset that best uncovers "interesting natural" groupings from data according to a chosen criterion**



- **Clustering Algorithm**
  - **K-Means, X-Means, G-Means, …**
- **Check for correctness and assign meaning to the clusters**

Clustering – Example – K-Means

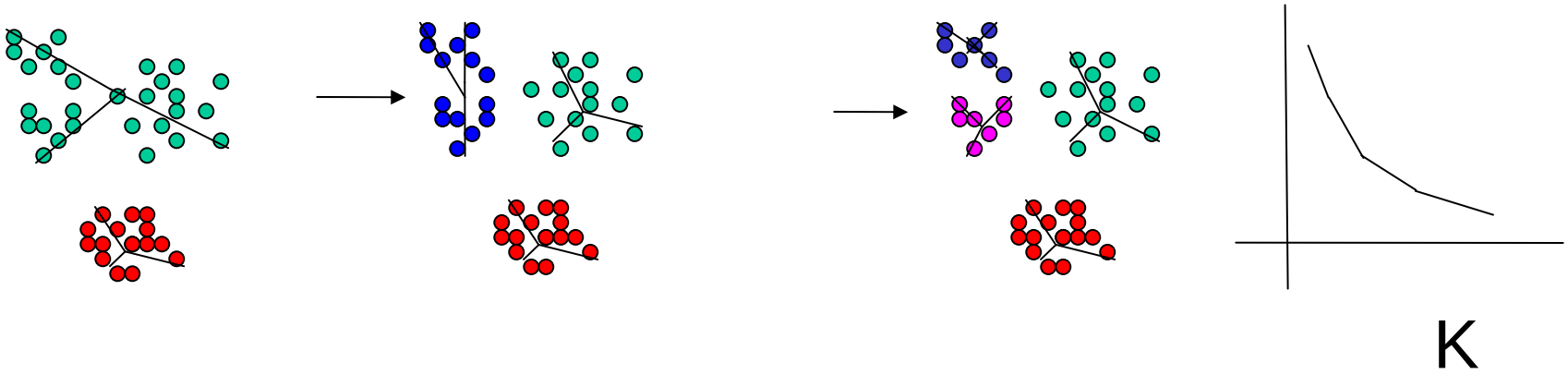


K-Means is an iterative algorithm

- Start with k-random prototypes (centroids)
- Assign data points to the nearest prototype
- Compute new k-prototypes

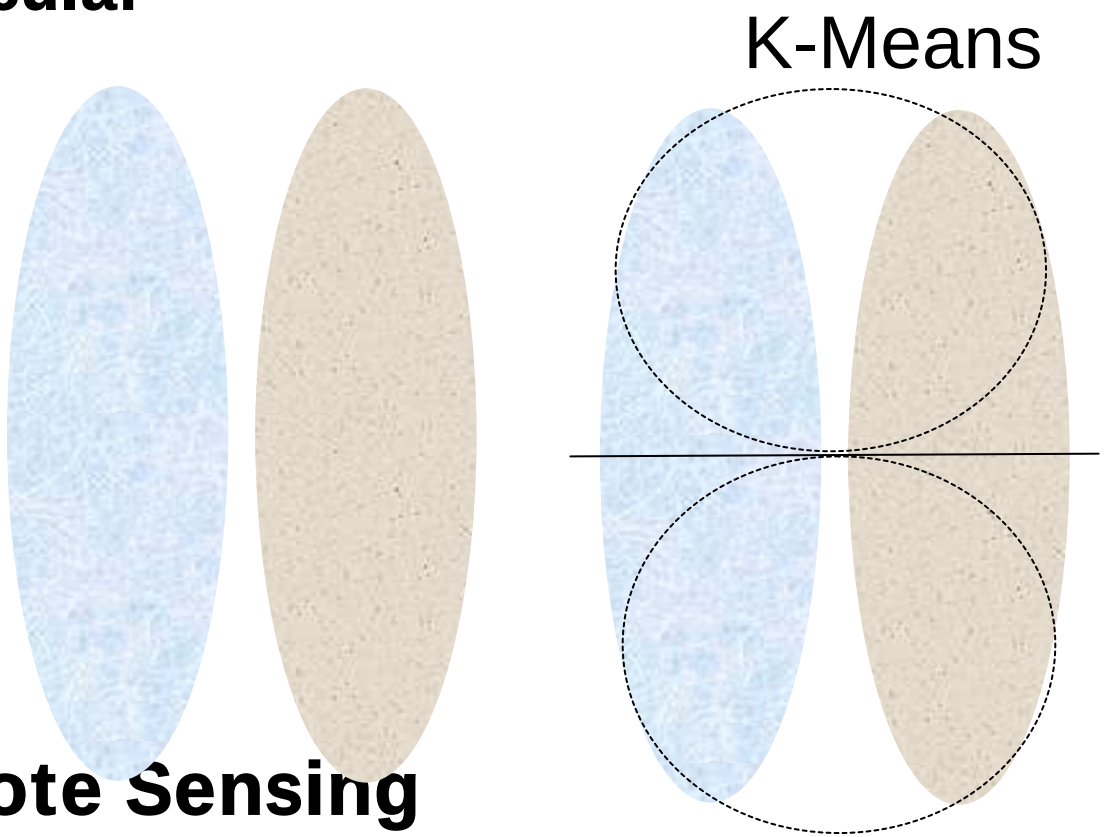
Clustering – Example – K-Means

- **Relatively efficient O(nki), where n = #objs, k = #clusters, i = #iterations.**
- **How many clusters?**
  - **Try different k, see how avg. dist to centroid changes as k increases**
  - **Avg. decreases rapidly till you reach right K, then stabilizes.**



Clustering – Example – K-Means

- **Type of clusters**
  - **Circular**



- **Remote Sensing**
  - **Multivariate Normal Distribution**

Clustering Model Parameter Estimation

EM to estimate GMM parameters

- **E-Step**

$$e_{ij} = \frac{\left|\hat{\Sigma}_j^k\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(x_i - \hat{\mu}_j^k\right)^T \hat{\Sigma}_j^{-1,k}\left(x_i - \hat{\mu}_j^k\right)\right\}}{\sum_{l=1}^{M}\left|\hat{\Sigma}_l^k\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(x_i - \hat{\mu}_l^k\right)^T \hat{\Sigma}_l^{-1,k}\left(x_i - \hat{\mu}_l^k\right)\right\}}$$

- **M-Step**

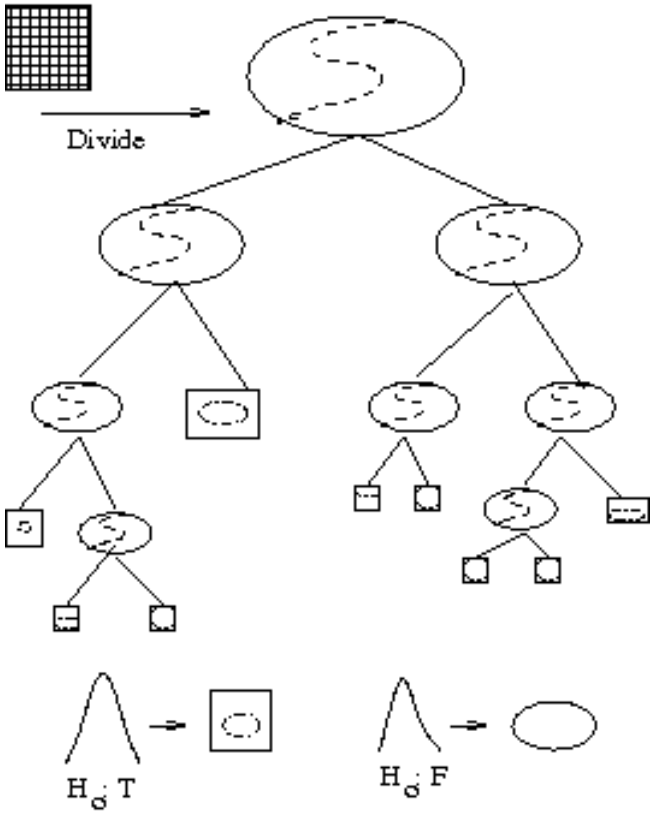
$$\alpha_j = \frac{\sum_{i=1}^{N} e_{ij}}{N}, \qquad \hat{\mu}_j^{k+1} = \frac{\sum_{i=1}^{N} e_{ij} x_i}{\sum_{i=1}^{N} e_{ij}},$$

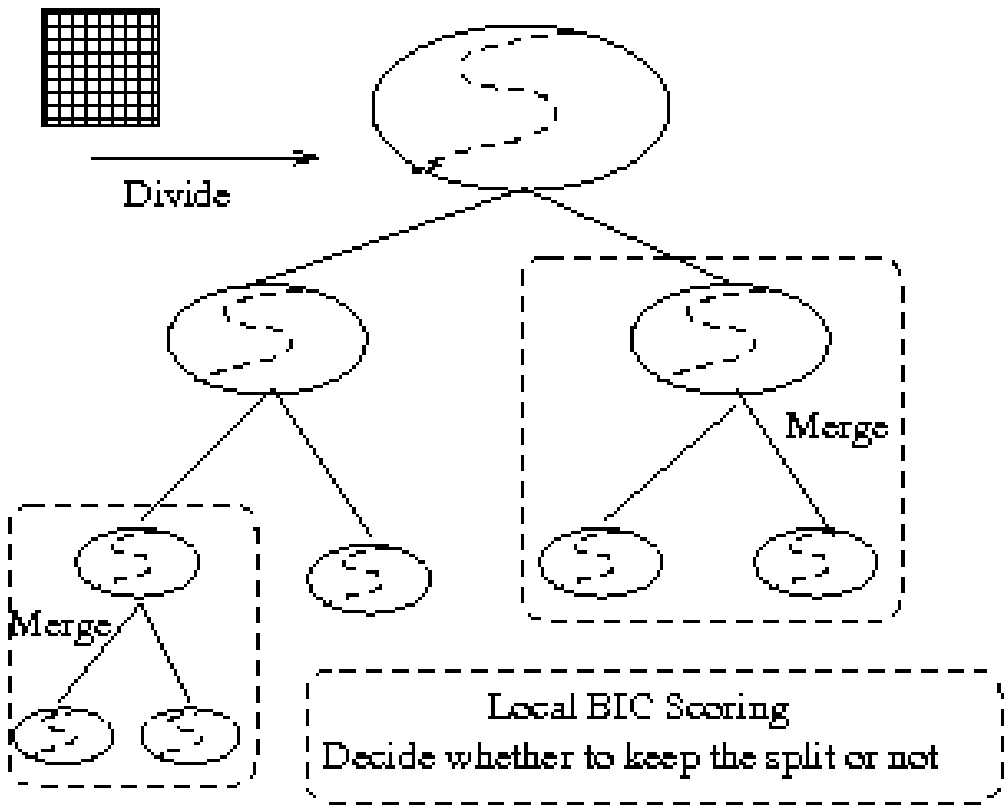
$$\text{and} \quad \hat{\Sigma}_j^{k+1} = \frac{\sum_{i=1}^{N} e_{ij}\left(x_i - \hat{\mu}_j^{k+1}\right)\left(x_i - \hat{\mu}_j^{k+1}\right)^T}{\sum_{i=1}^{N} e_{ij}}$$

i^{th} data vector, j^{th} class

Problem – How many clusters?

Clustering: G-Means and X-Means

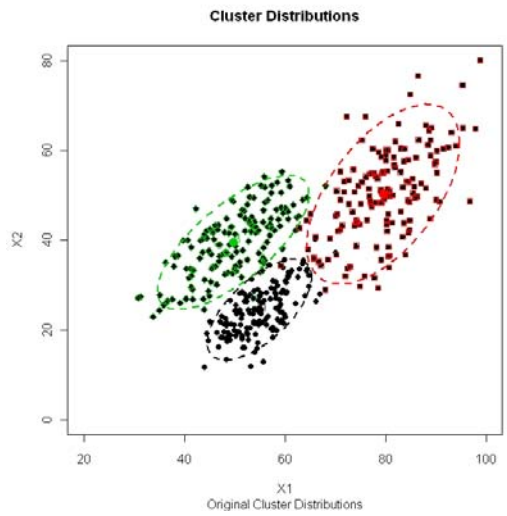


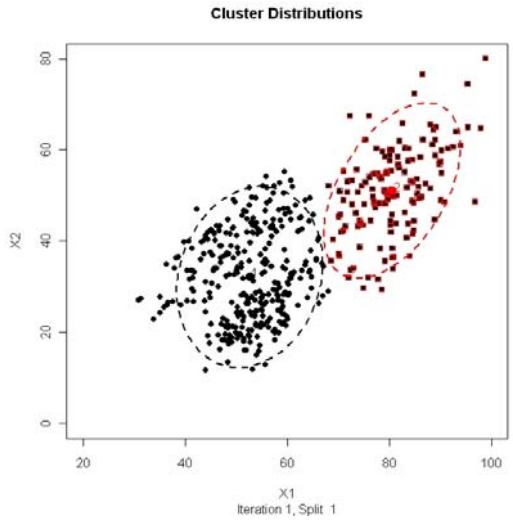


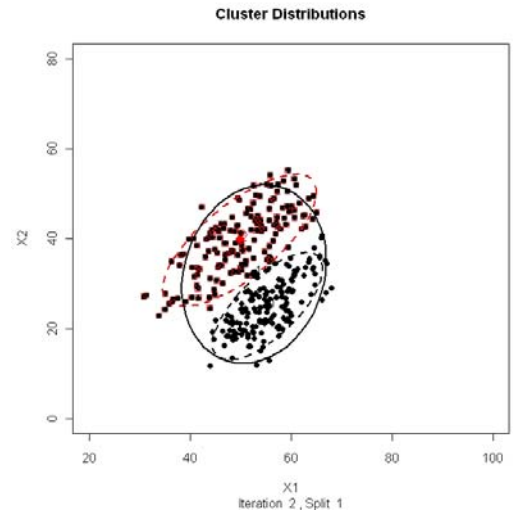
G-Means

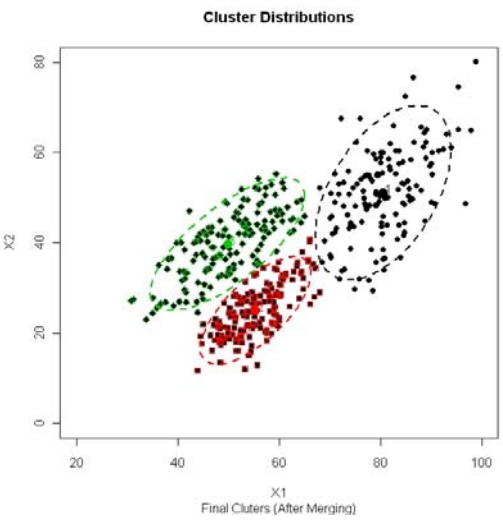
X-Means

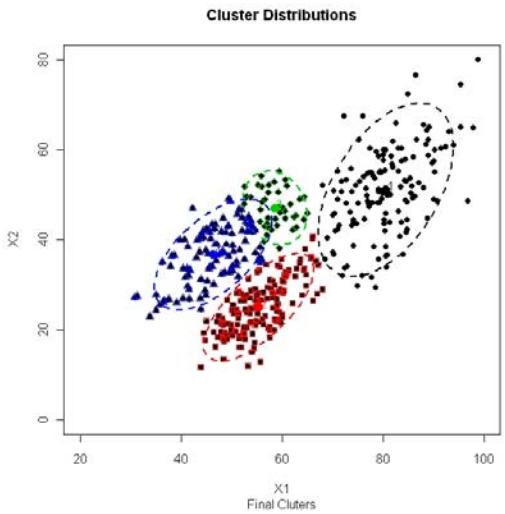
Binary Splits -- Limitations

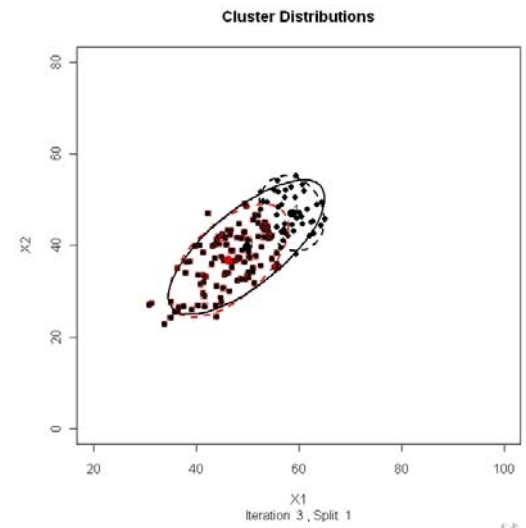






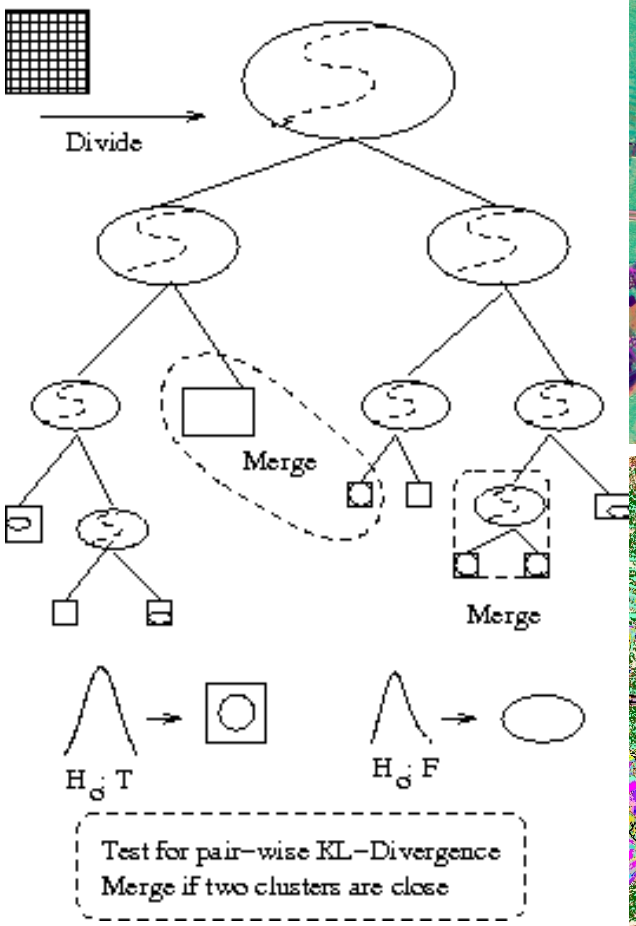


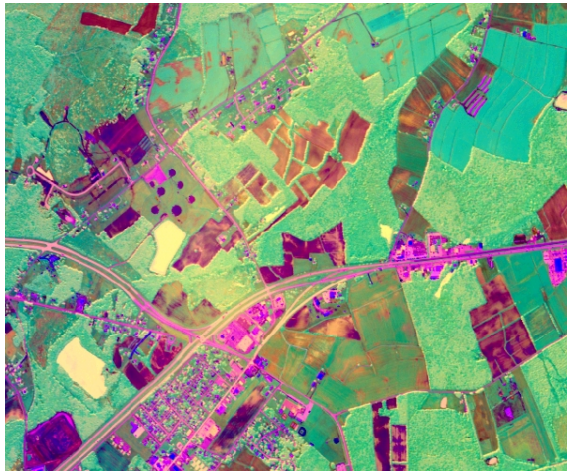


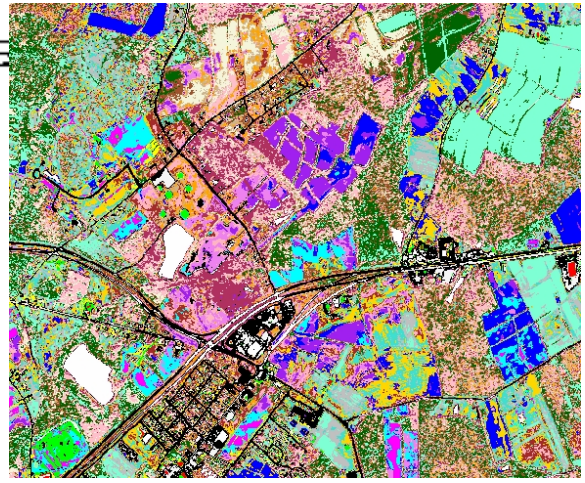


How Many Clusters?

- **Model Selection Criteria – AIC/BIC/MDL/MML …**
- **Our Approach: GX-Means**







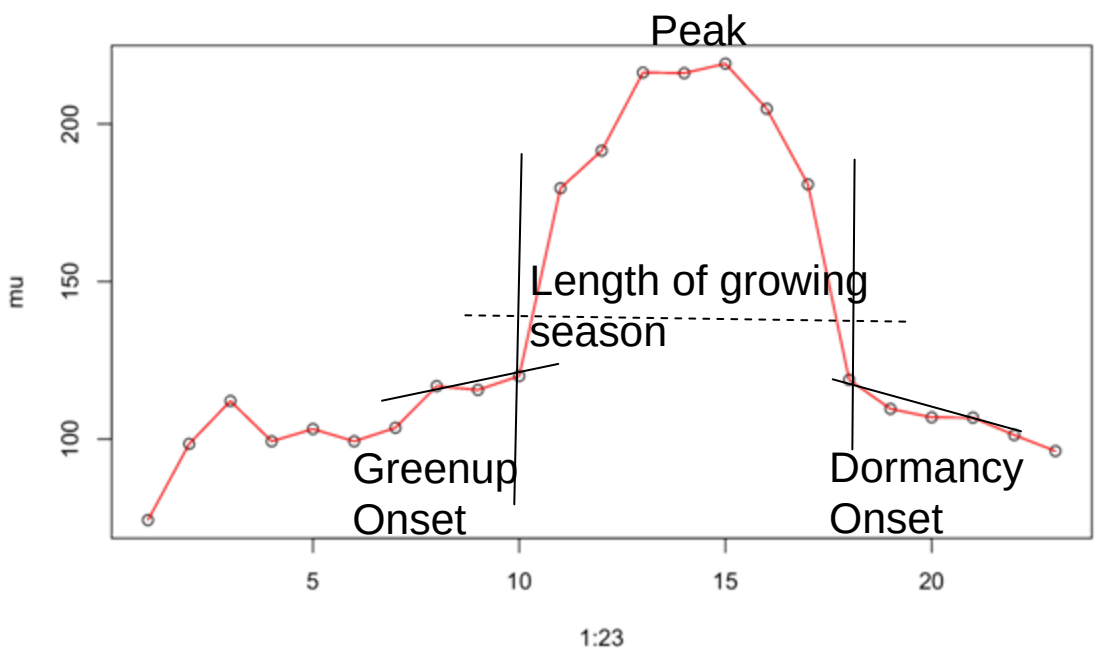
Method	Total Clusters	GMM	
Model-based	7	2	
X-Means	6	4	
G-Means	19	5	
Adaptive-G-Means	14	9	
Carleton (Landsat ETM+) Image Clustering Results			
Model-based	31	14	
X-Means	22	15	
G-Means	36	12	
Adaptive-G-Means	27	26	
Jasper Ridge (AVIRIS) Image Clustering Results			

Characterize changes

- **Two possible approaches**
  - **1st Case**
    - **Phenological changes**
  - **2nd Case**
    - **Acquire high-resolution images (before and after dates)**
    - **Do detailed classification**
    - **Determine which class is converted into which other class**
      - **Wheat -> Corn**
      - **Wheat got damaged**

1st Case: Phenological Changes

- **What is phenology?**
  - **Study of cyclic events of nature**
  - **Vegetation communities have seasonally driven phenologies**
    - **Onset, Peak, and End of growing seasons**



Basic Algorithm:

- **For each change polygon**
  - **Compute mean spectral profile on filtered time series extracted from change polygons**
  - **Determine onset, peak, and offset points in each of T1, T2 images**
  - **Make comparison between time of occurrence of these points**
    - **Early onset**
    - **Changes in length of growing seasons**
    - **Changes in peaks**

2nd Case: Type (class) Based Change Characterization

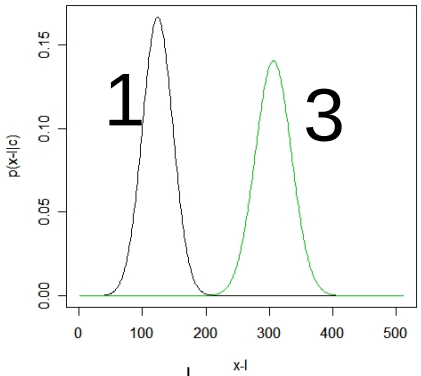
- **Fine resolution (type) information extraction from before and after event images**
  - **A is converted to B (e.g., Wheat -> Corn)**
  - **A became A' (may be due to damage)**
- **Challenges**
  - **Insufficient ground truth**
  - **Large spatial extents**
- **We developed two novel solutions**
  - **Semi-supervised Learning**
  - **Sub-class classification**

Classification with limited ground truth

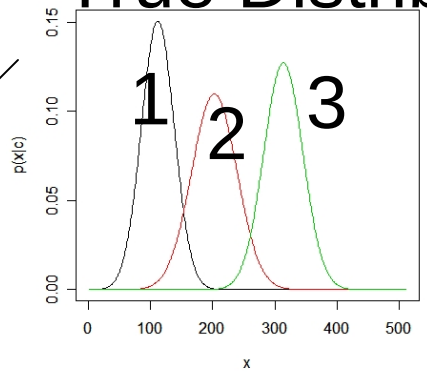
- #of training samples ~ (10 to 100) * (number of dimensions)
  - Costly ~ $500-$800 per plot (depends on geographic area)
  - Accessibility – Private/Privacy issues (e.g., USFS may average 5% denied access)
  - Real-time – Emergency situations, such as, forest fires, floods
- Increase number of samples
  - By incorporating unlabeled samples
    - Naïve semi-supervised (Nigam et al. [JML-2000])
  - Bagging [Breiman, ML-96]
- Reduce number of dimensions

Illustration of "Informed Semi-supervised" Learning

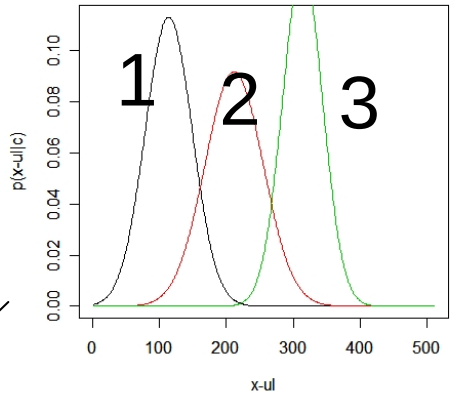
Labeled



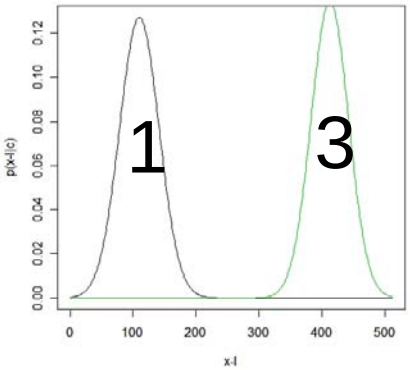
True Distribution

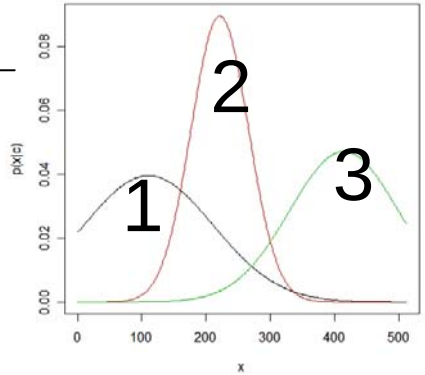


Unlabeled

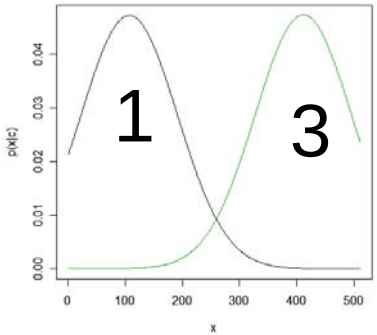


Match



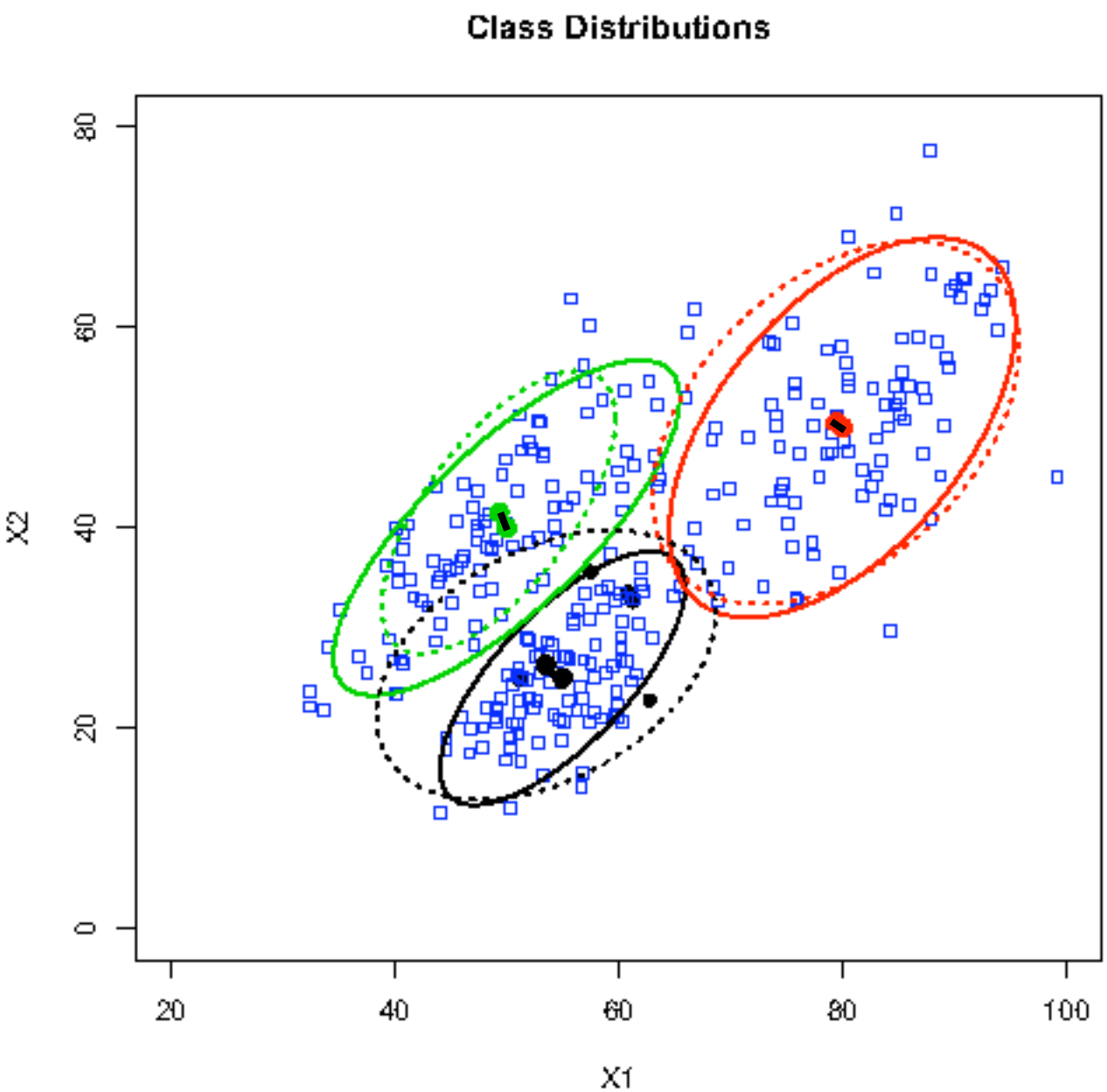


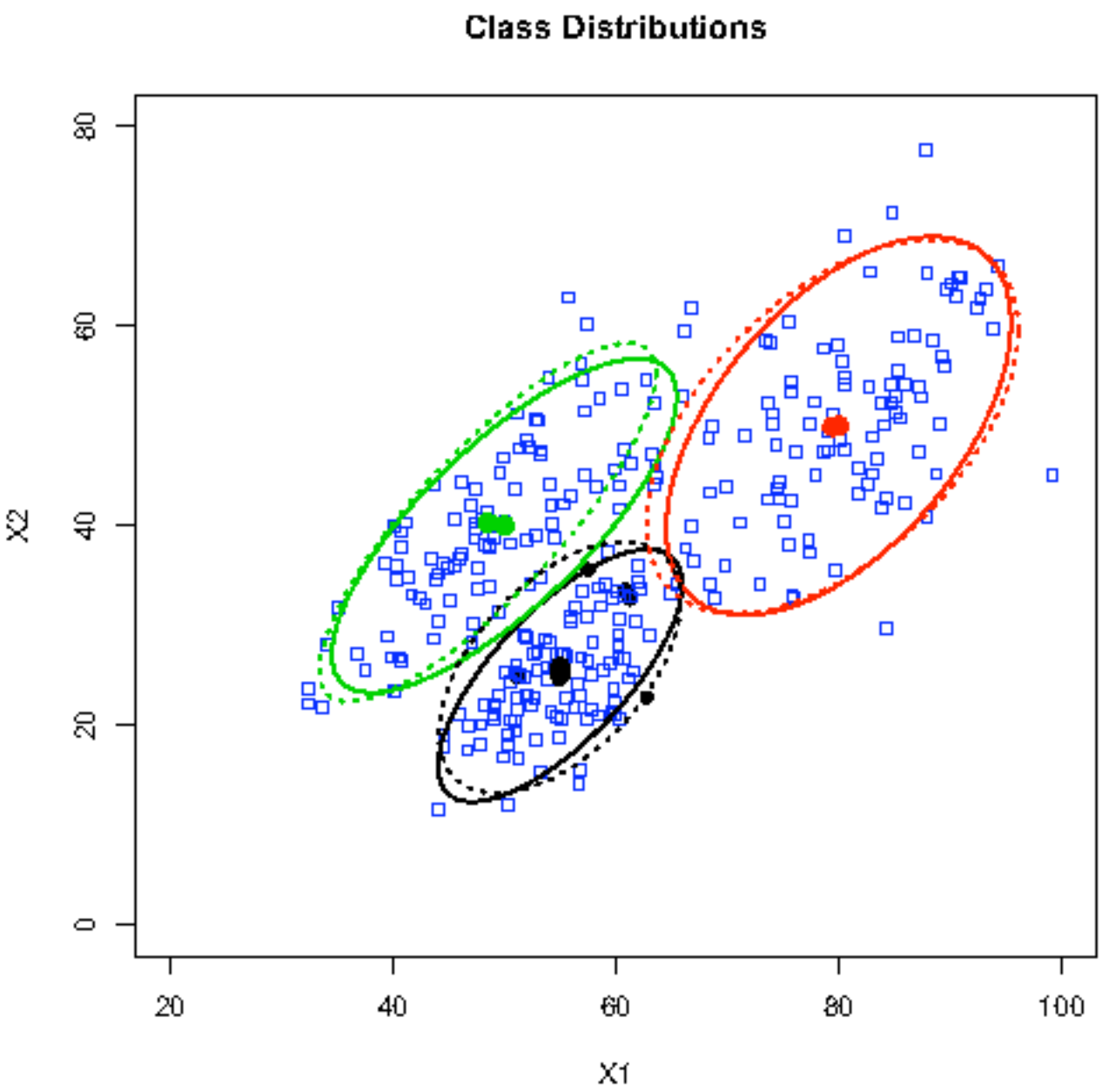
Clustering

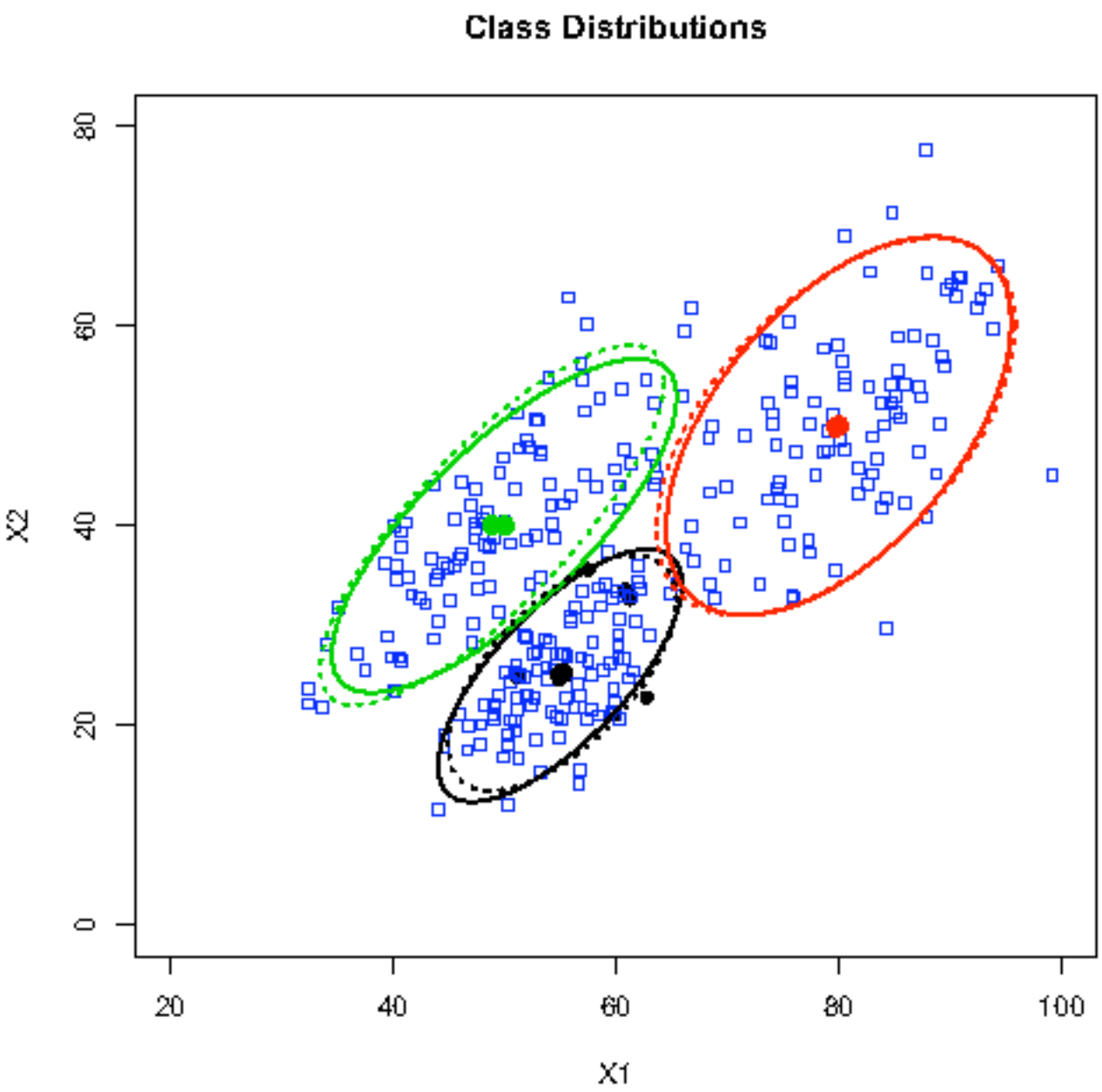


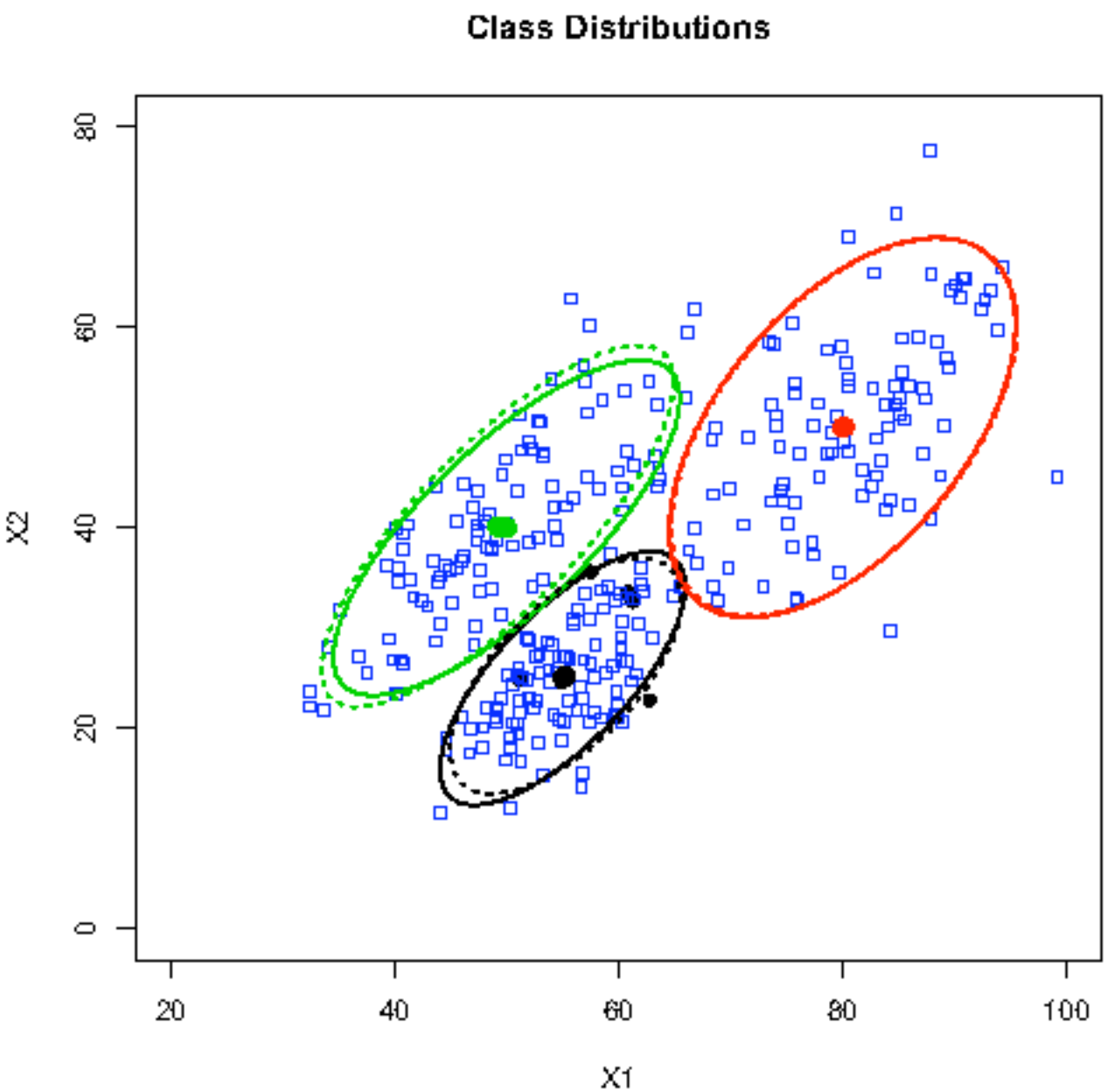
Informed semi-supervised

Naïve semi-supervised

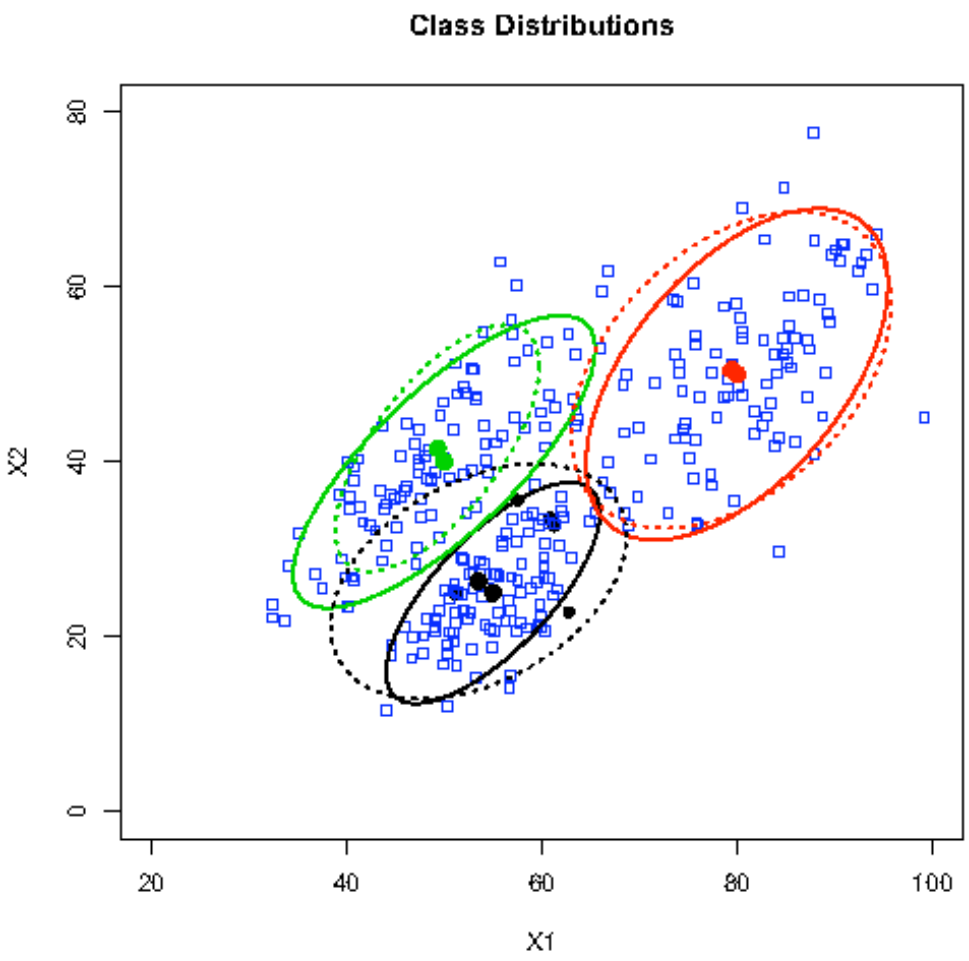

Class Distributions
X2
X1
0
20
40
60
80
20
40
60
80
100

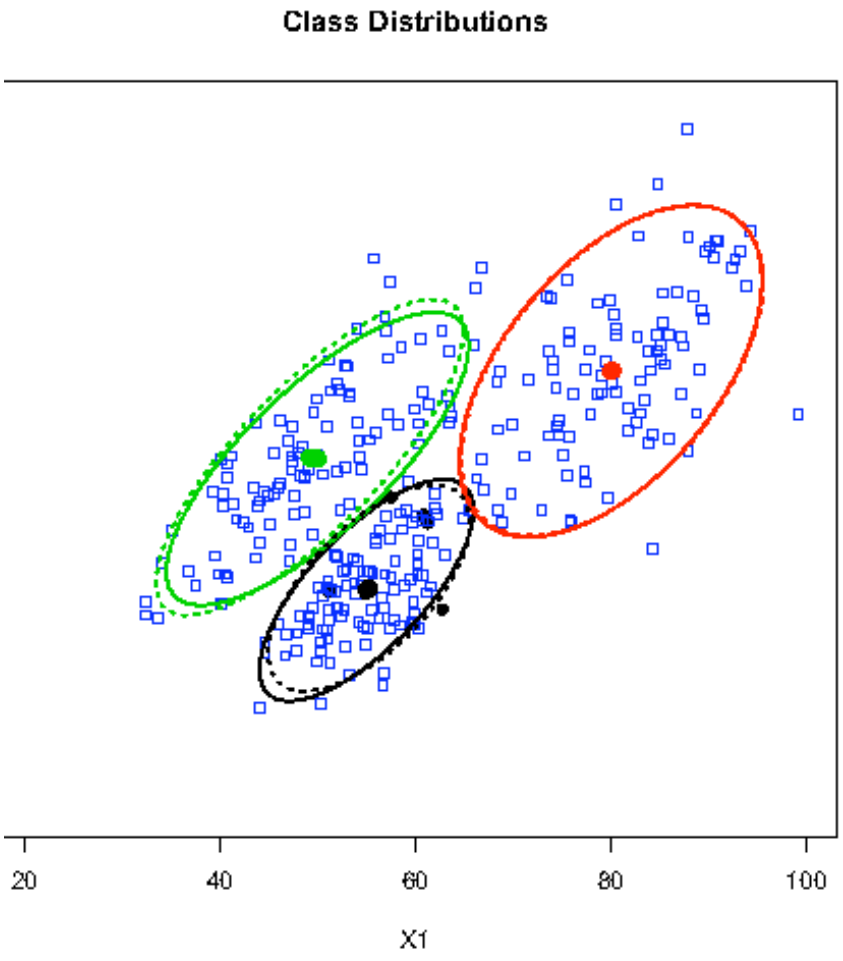

Class Distributions
X2
X1
0
20
40
60
80
20
40
60
80
100


Class Distributions
X2
X1
0
20
40
60
80
20
40
60
80
100


Class Distributions
X2
X1
0
20
40
60
80
20
40
60
80
100

GMM Clustering Results





Sub-class Classification - Algorithm

Each Class is Unimodal Gaussian

MLE

Each Aggregate Class is GMM

How many components?

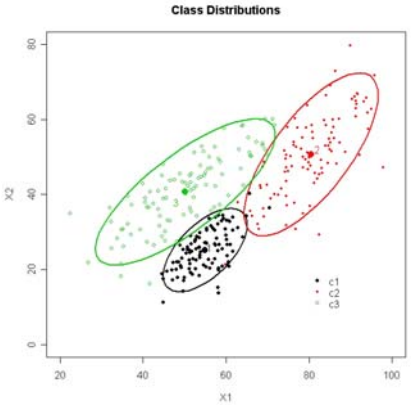
BIC/AIC Model Selection
+ Parameter Estimation

What are these components?

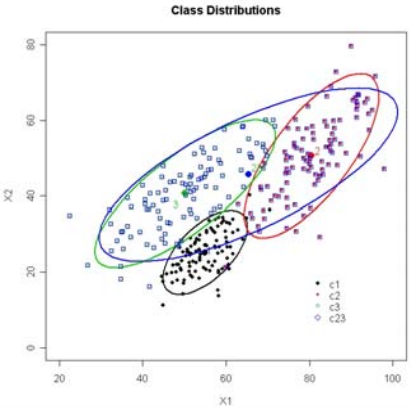
Few labels/
sub-class

Unlabeled
Samples

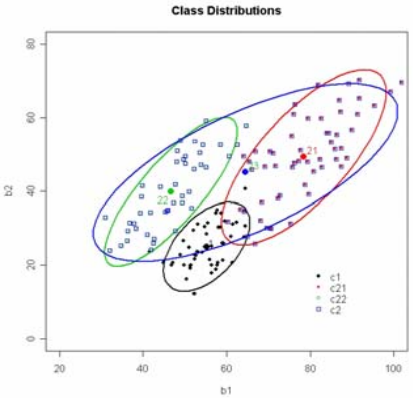
Semi-supervised Sub-class Classification



Actual Distribution of Classes



User given (Aggregate Class)



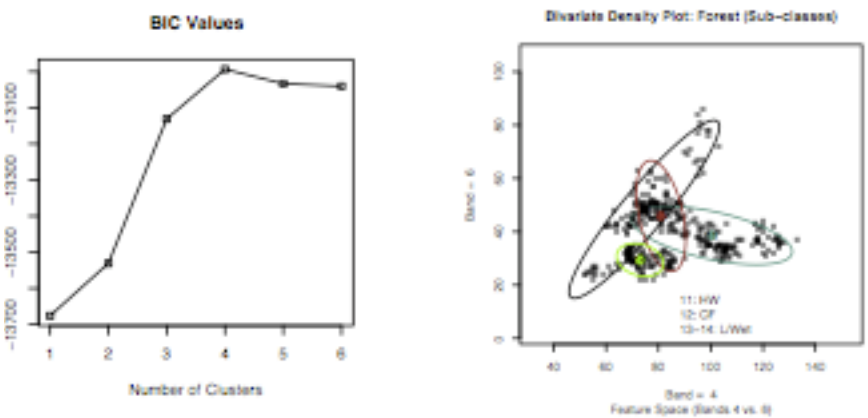
Sub-classes Recovered from Agg. Cl.

Results: Sub-class Classification

G. Truth	1	2	3	4	P. Acc.
Forest(1)	1475.00	9.00	28.00	0.00	97.55
Agriculture(2)	90.00	142.00	2.00	0.00	60.68
Urban(3)	0.00	0.00	45.00	0.00	100.00
Wetlands(4)	18.00	0.00	2.00	34.00	62.96
Users Acc.	93.18	94.04	58.44	100.00	91.92

Table 10. Accuracy (Aggregate Classes)

GT	1	2	3	4	P. Acc.
Forest(1)	1448.00	13.00	51.00	0.00	95.77
Ag.(2)	14.00	214.00	6.00	0.00	91.45
Urban(3)	0.00	0.00	45.00	0.00	100.00
Wet.(4)	3.00	0.00	13.00	38.00	70.37
U. Acc.	98.84	94.27	39.13	100.00	94.58



(a) BIC Values (b) Bivariate Density Plots

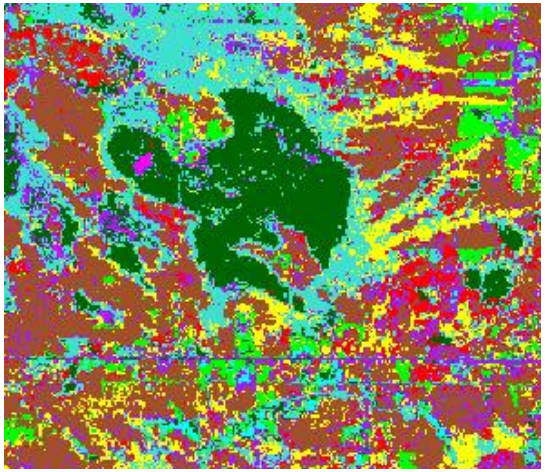
Figure 5. Forest Sub-classes

Dataset:
LandSat ETM+ Data (Cloquet, Carleton, MN, May 31, 2000)
•6 Bands, 4 Classes, 60 plots
•Independent test data: 205 plots
•Forest (4 Subclasses; 2 subclasses are combined into 1
•2 Labeled plots per sub-class

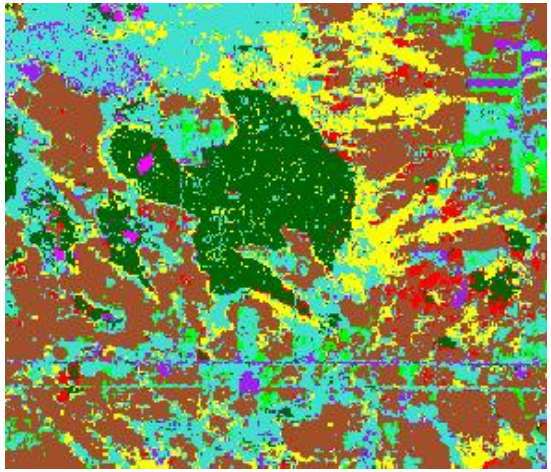
GT	HW	CF	L/Wet	P.Acc
HW	642.00	14.00	10.00	96.40
CF	9.00	245.00	7.00	93.87
L/Wet	11.00	2.00	383.00	96.72
U.Acc	96.98	93.87	95.75	(OA) 95.99

Table 13. SSL Accuracy (Forest Sub-classes)

Spatial Semi-supervised

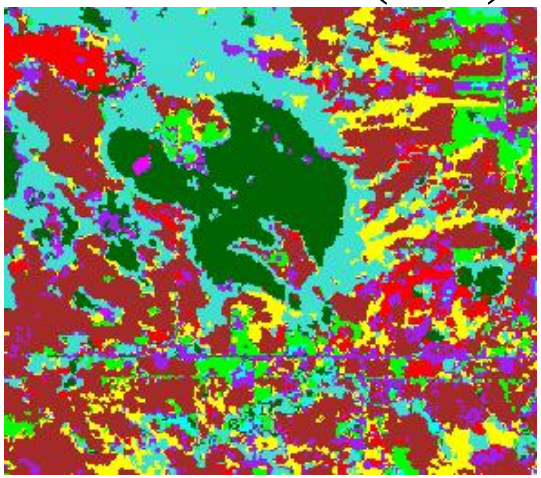


BC (60%)

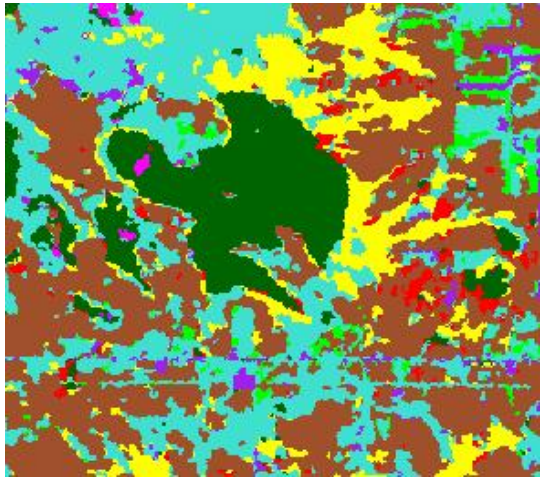


BC-EM (68%)

BC-MRF (65%)



BC-EM-MRF (72%)



Oak Ridge National Laboratory

Conclusions and Future Directions

- **LULC is still a challenging task**
- **SSTDM approaches shows great promise**
- **Biomass changes can be monitored at regional and global scales**
- **Need techniques for analyzing multisource data**
  - **Remote Sensing**
  - **Ground sensor observations (Aeronet, Fluxnet)**
  - **Climate Simulations**
- **Computational challenges need to be addressed**

Oak Ridge National Laboratory

Acknowledgements

- **Research was funded through ORNL LDRD grant.**
- **AWiFS data was provided by USDA**
- **External Collaborators: USDA (Brad Doorn, Glenn Bethel, Robert Tetrault, Ed Kunz, Guy Serbian, P. Gowda/ARC), Shawana Johnson (Global Marketing Insights, Inc.), Prof. S. Shekhar, Prof. M. Crawford, Prof. J. Ghosh, …**
- **ORNL: V. Protopopescu, B. Bhaduri, E. Bright, R. Graham, S. Gleason, …**
- **Prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285, managed by UT-Battelle, LLC for the U. S. Department of Energy under contract no. DEAC05-00OR22725.**