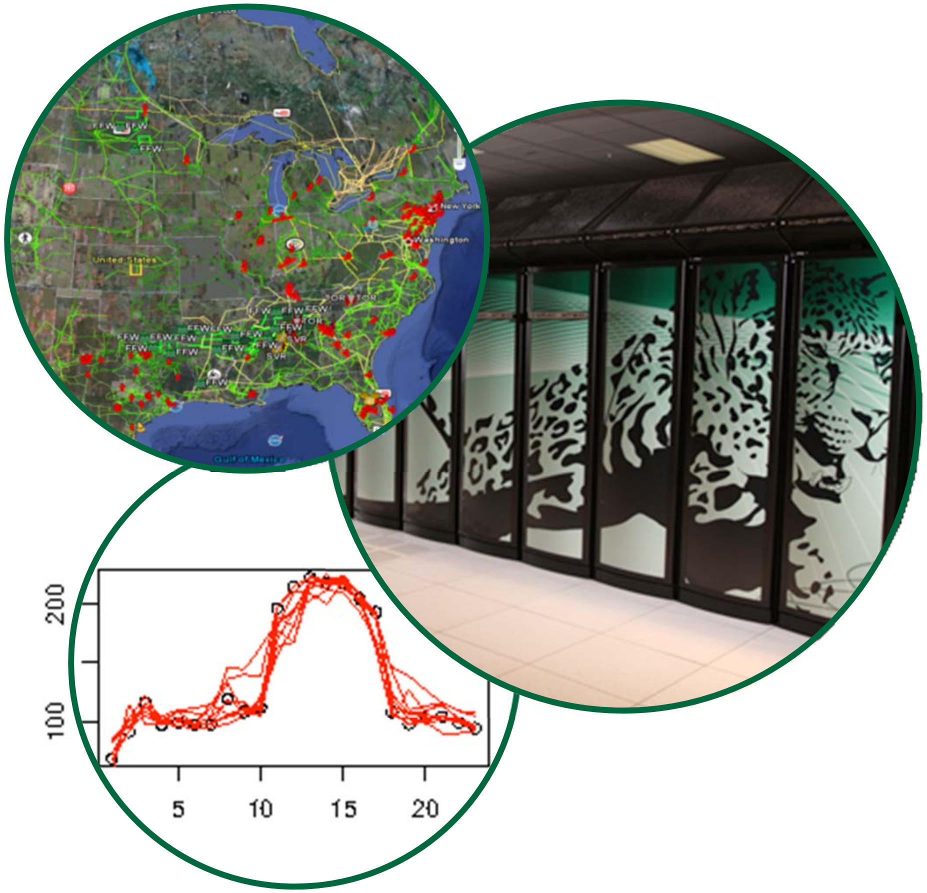
Large Scale Spatiotemporal Data Mining



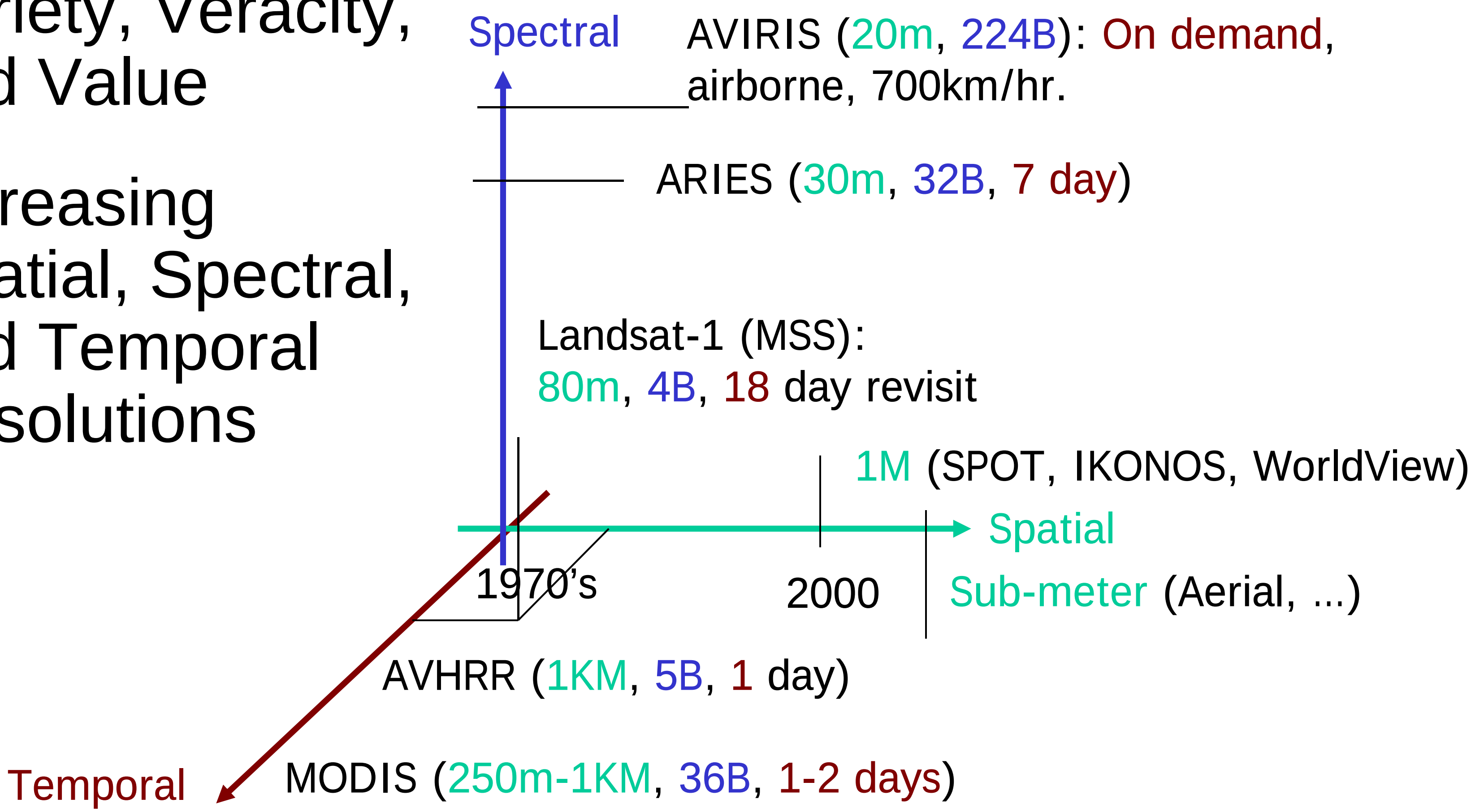
Raju Vatsavai

Geospatial Data Sciences

Computational Sciences and Engineering Division

Big Spatial Data

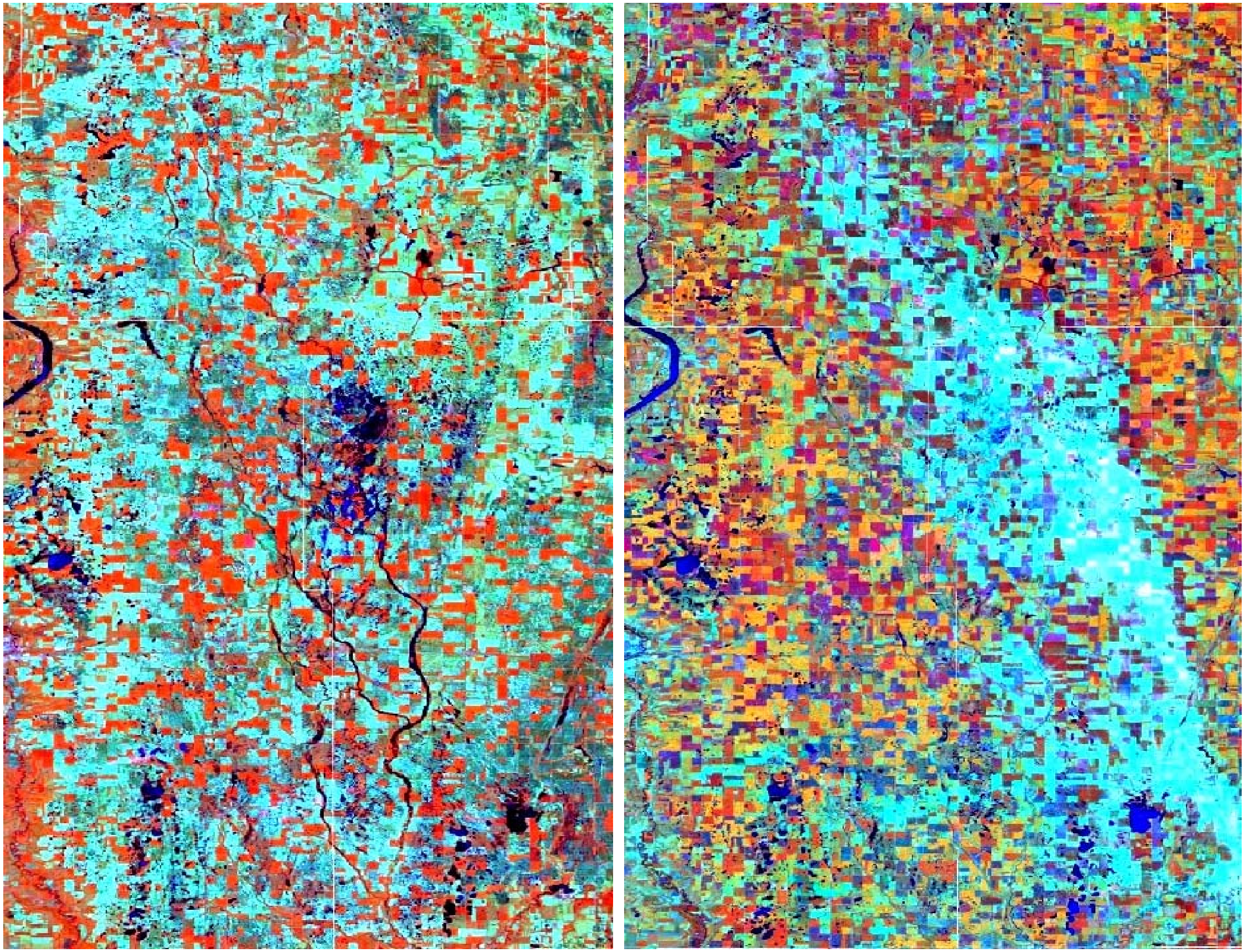
- Volume, Velocity, Variety, Veracity, and Value
- Increasing Spatial, Spectral, and Temporal Resolutions



5TB/day – Heterogeneous data

Part 1: Monitoring

Finding change patterns: Veg. damages

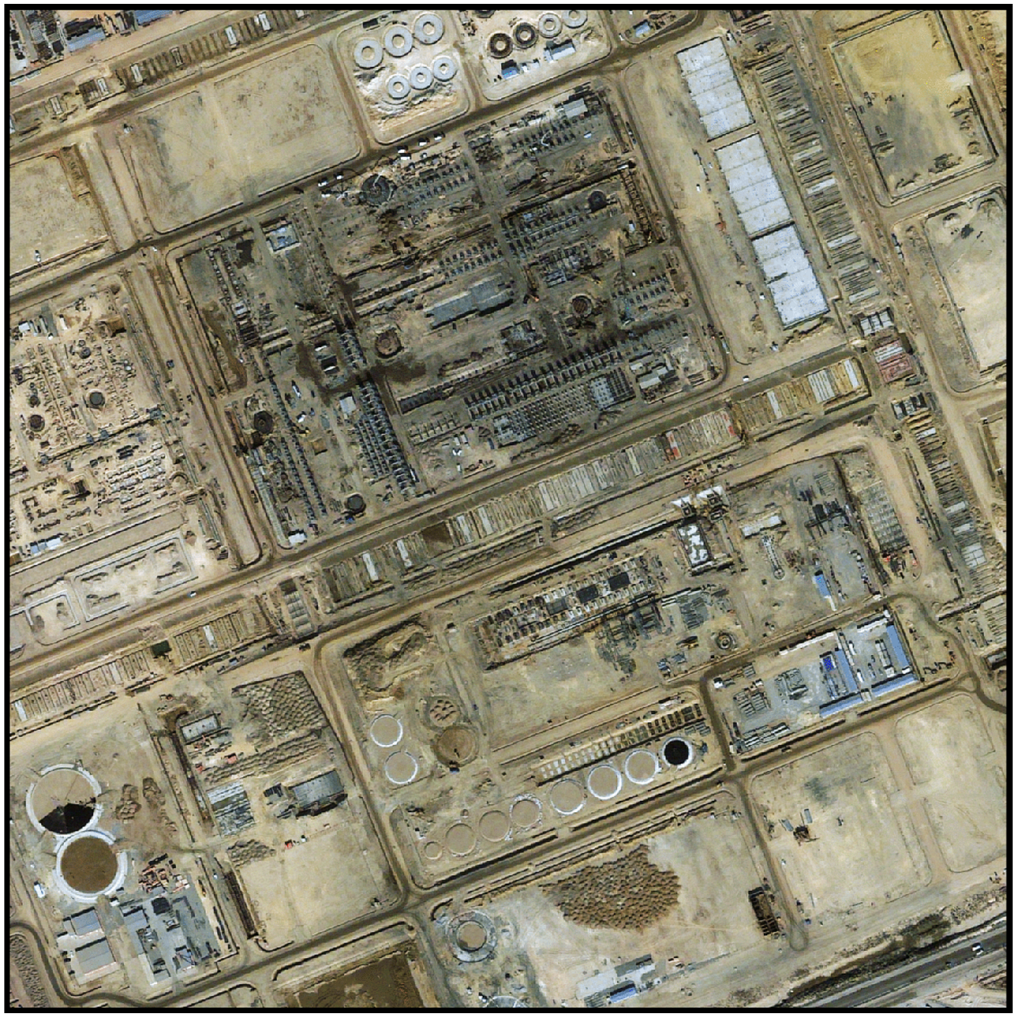


AWiFS (56 m, 4B, 5d)
- **Moderate spatial, Moderate temporal**
- **Used for crop type and condition extraction**
- **Not good for changes at building level**

Finding change patterns: infrastructure damages

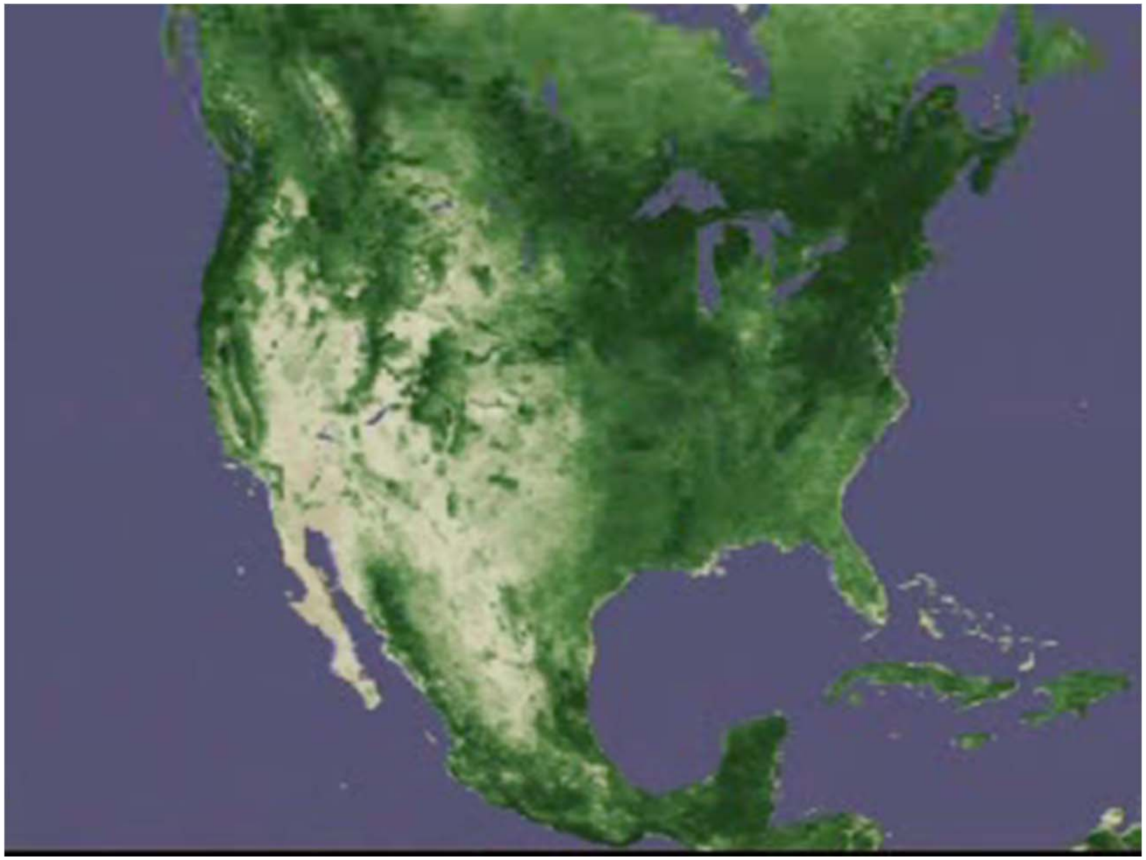
Haiti Earthquake Damages

Finding change patterns: new construction



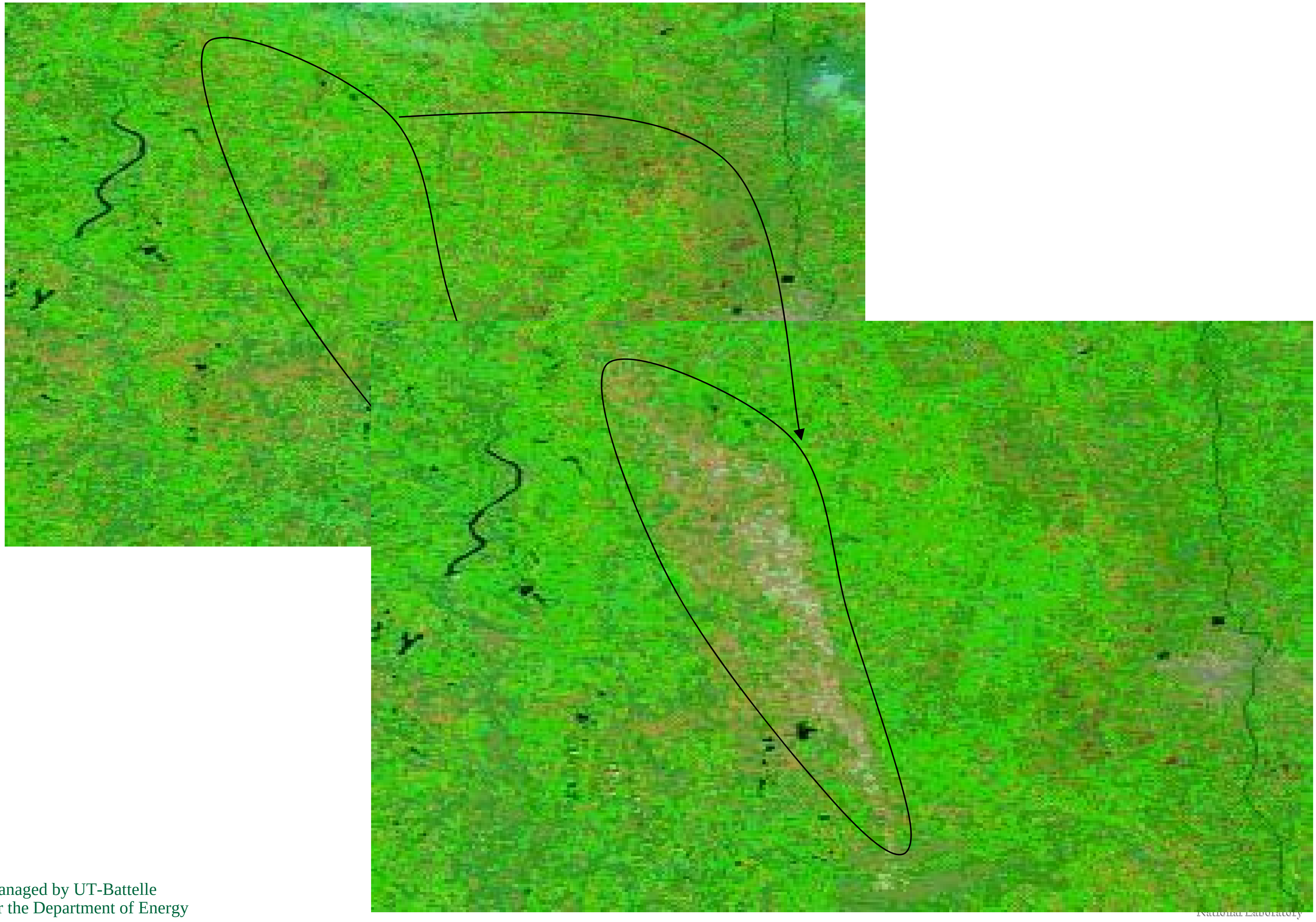
China – New Construction (QuickBird)

Understanding seasonal patterns

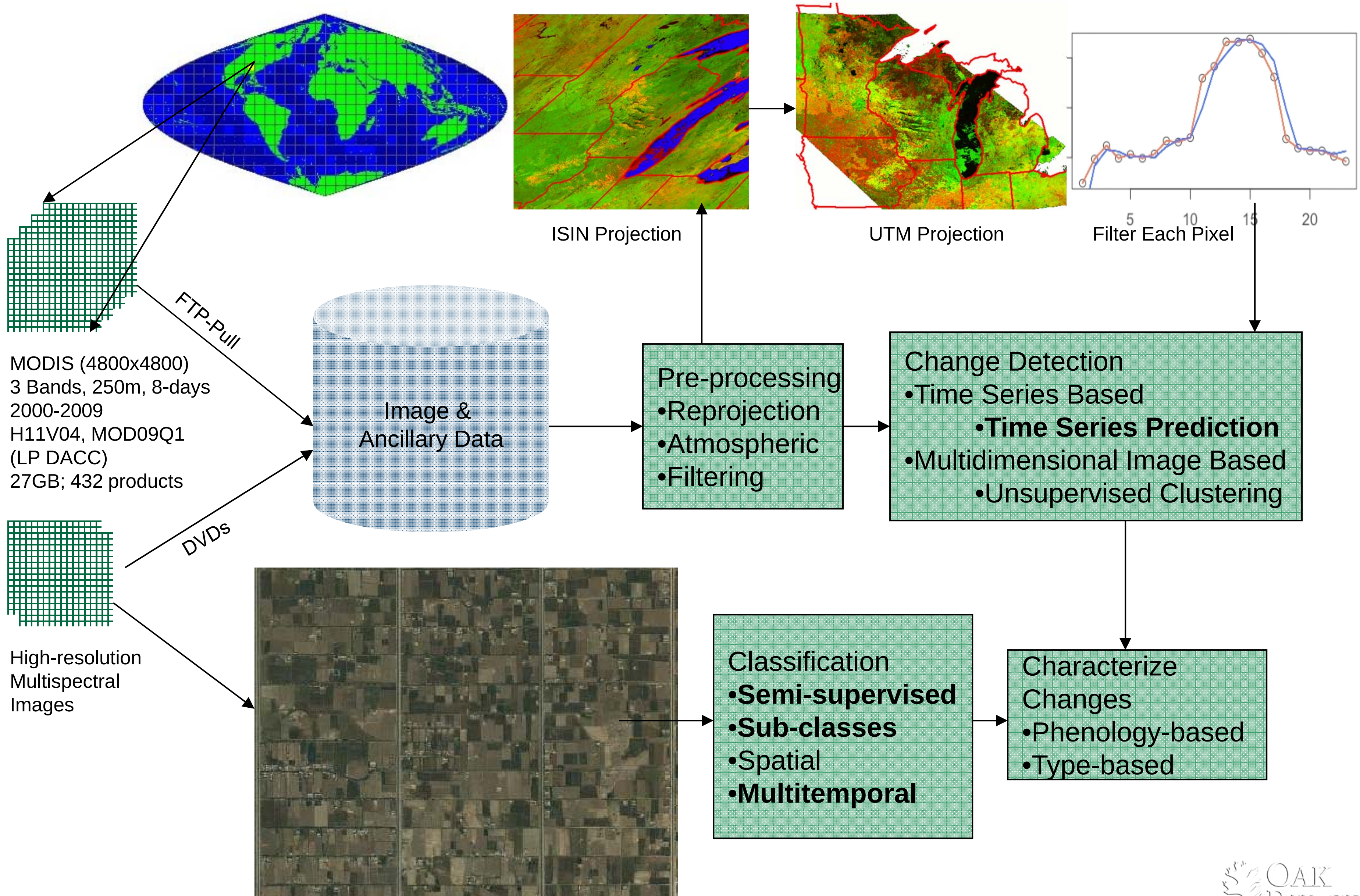


AVHRR NDVI 1KM (1981-2000)

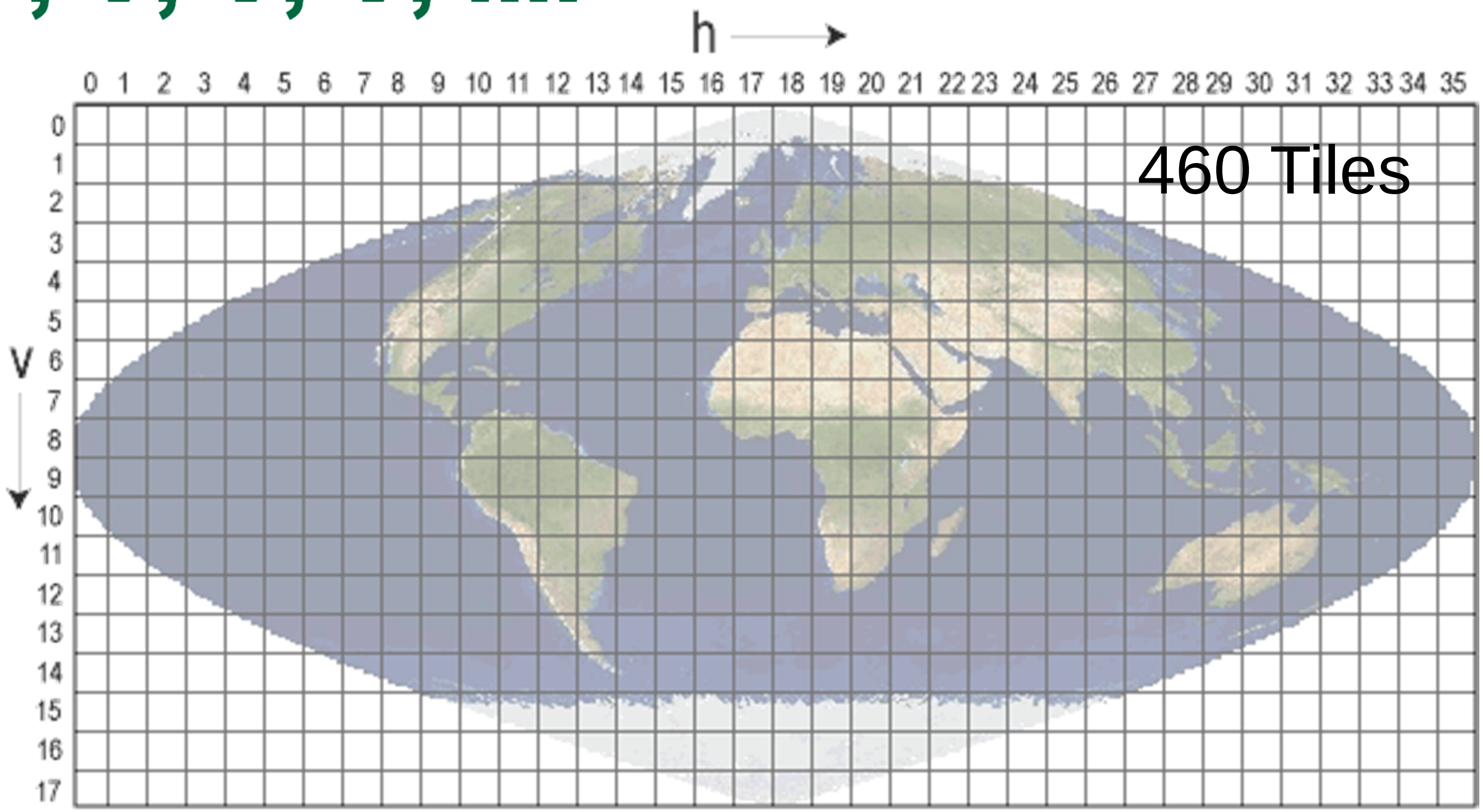
Example Change



Biomass Monitoring: Architecture



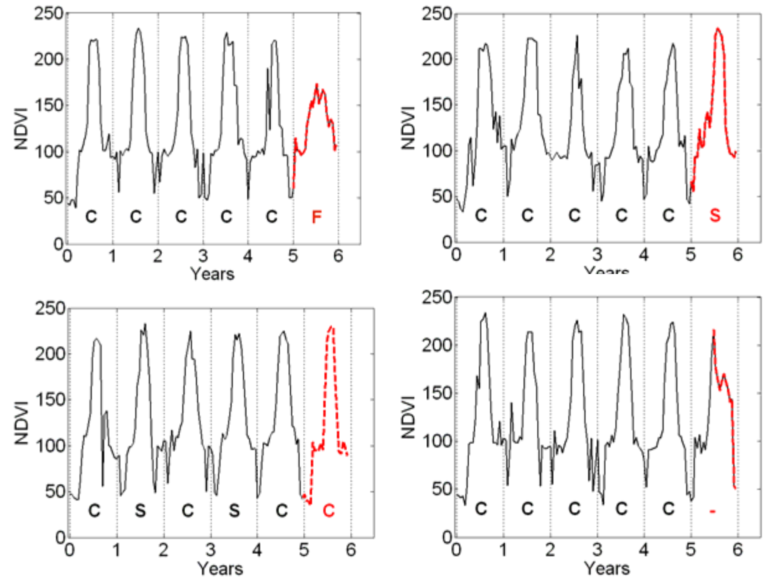
V, V, V, V, ...

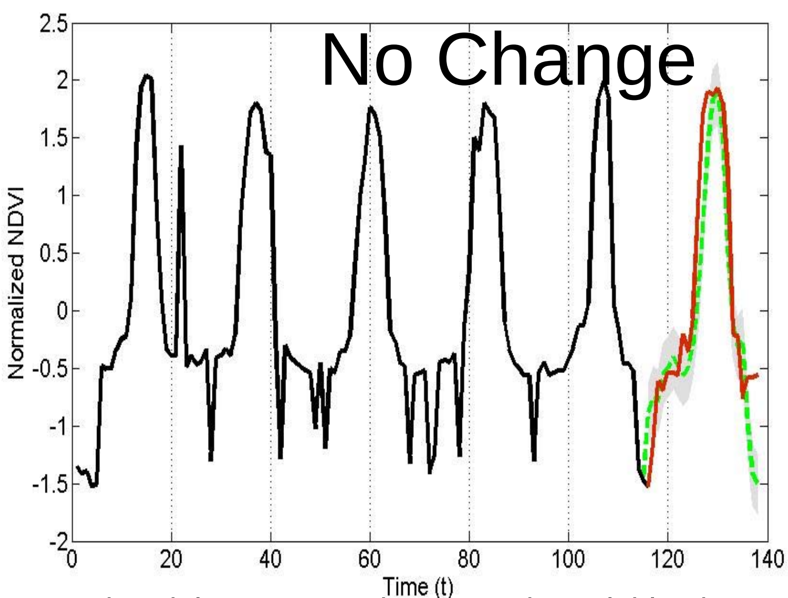


- Each Tile = 4800 x 4800 = 23,040,000 (250m)
- 16-bit, 1 Band = 44 MB
- 10 Trillion time series at Global Scale
- Temporal: Daily to Weakly composite

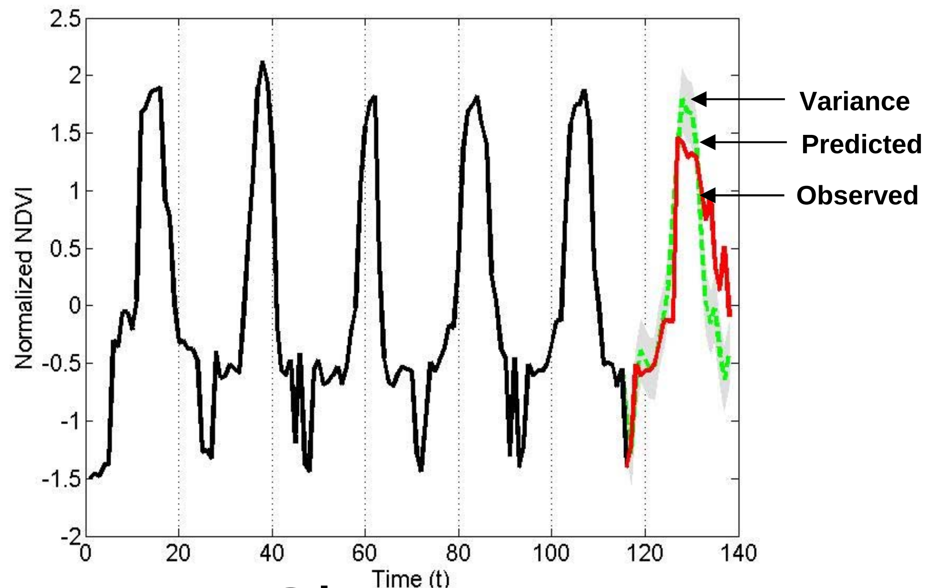
Change Detection Using Gaussian Process Model

- MODIS NDVI Time Series from Iowa
  - 6 years (2001 – 2006)
  - 23 observations per year
- Trained for first 5 years and monitored last year
- *Accuracy was 88% on a validation set consisting of 97 labeled time series with 13 true changes*





Varun Chandola, Ranga Raju Vatsavai: Scalable Time Series Change Detection for Biomass Monitoring Using Gaussian Process. NASA CIDU 2010: 69-82 (One of the best papers, invited to SADM Journal).



Change

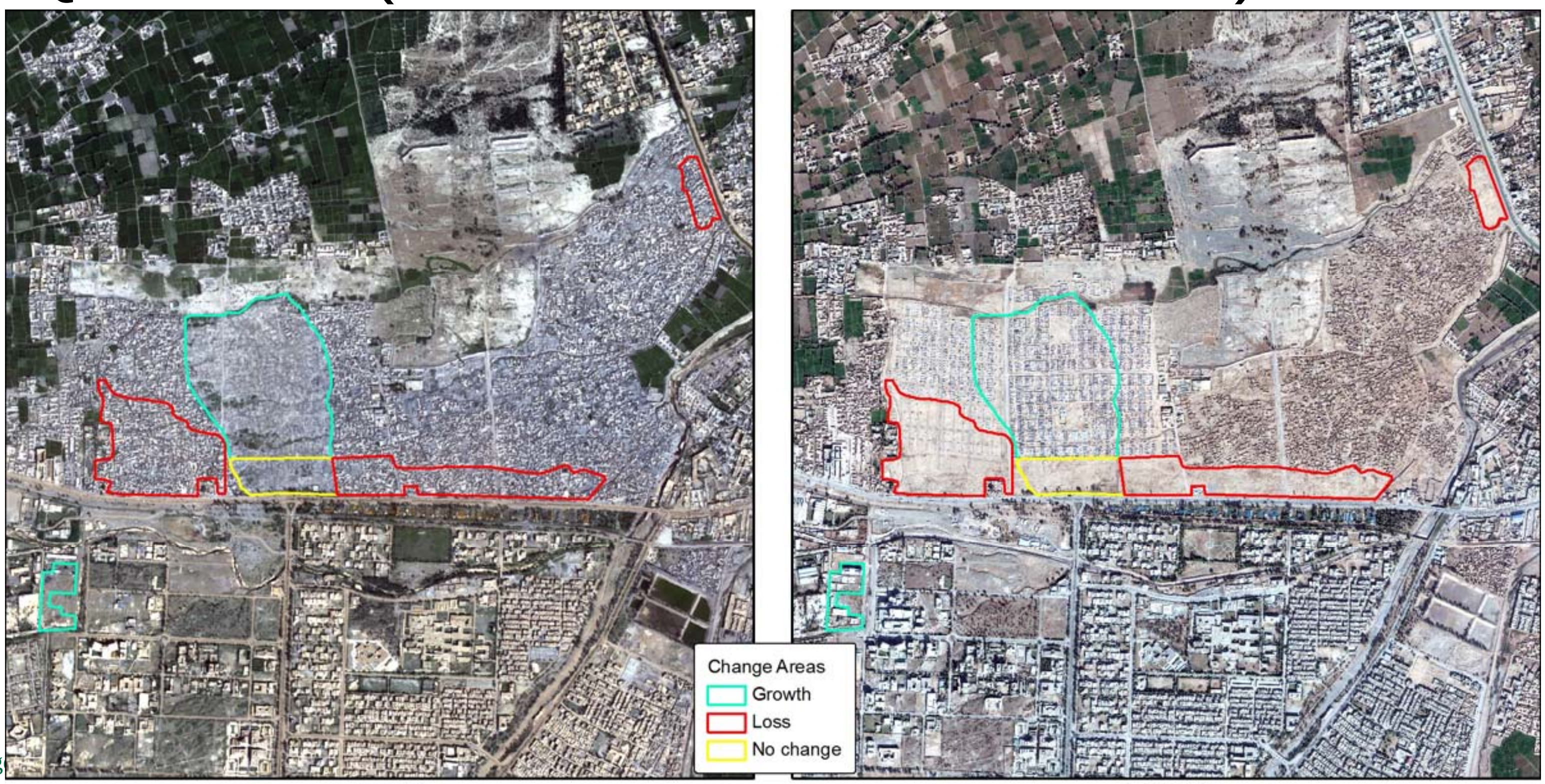
Bitemporal Changes

- Point based – at individual pixel (or small neighborhood)
- Mostly univariate
- Multivariate (e.g., MAD) techniques produce multi-band change maps
- Mostly the output is continuous (requires thresholding)

Oak Ridge National Laboratory

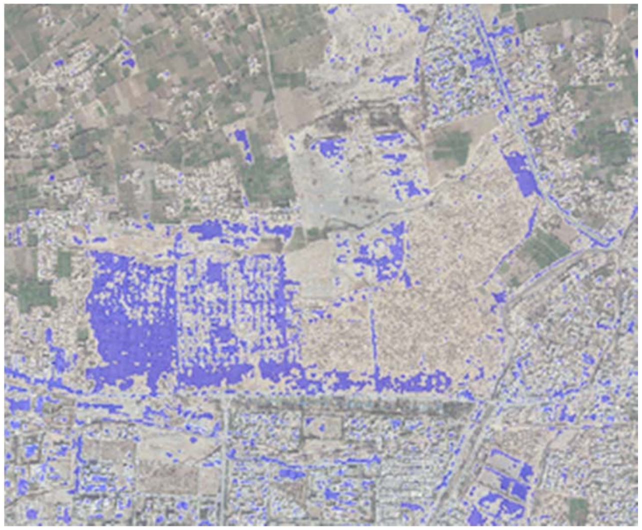
Experimental Setup

- Kacha Garhi Camp, Pakistan
- Established 1980 for Afghan Refugees
- QuickBird (2004 and 2009, 4B, 2.4m)



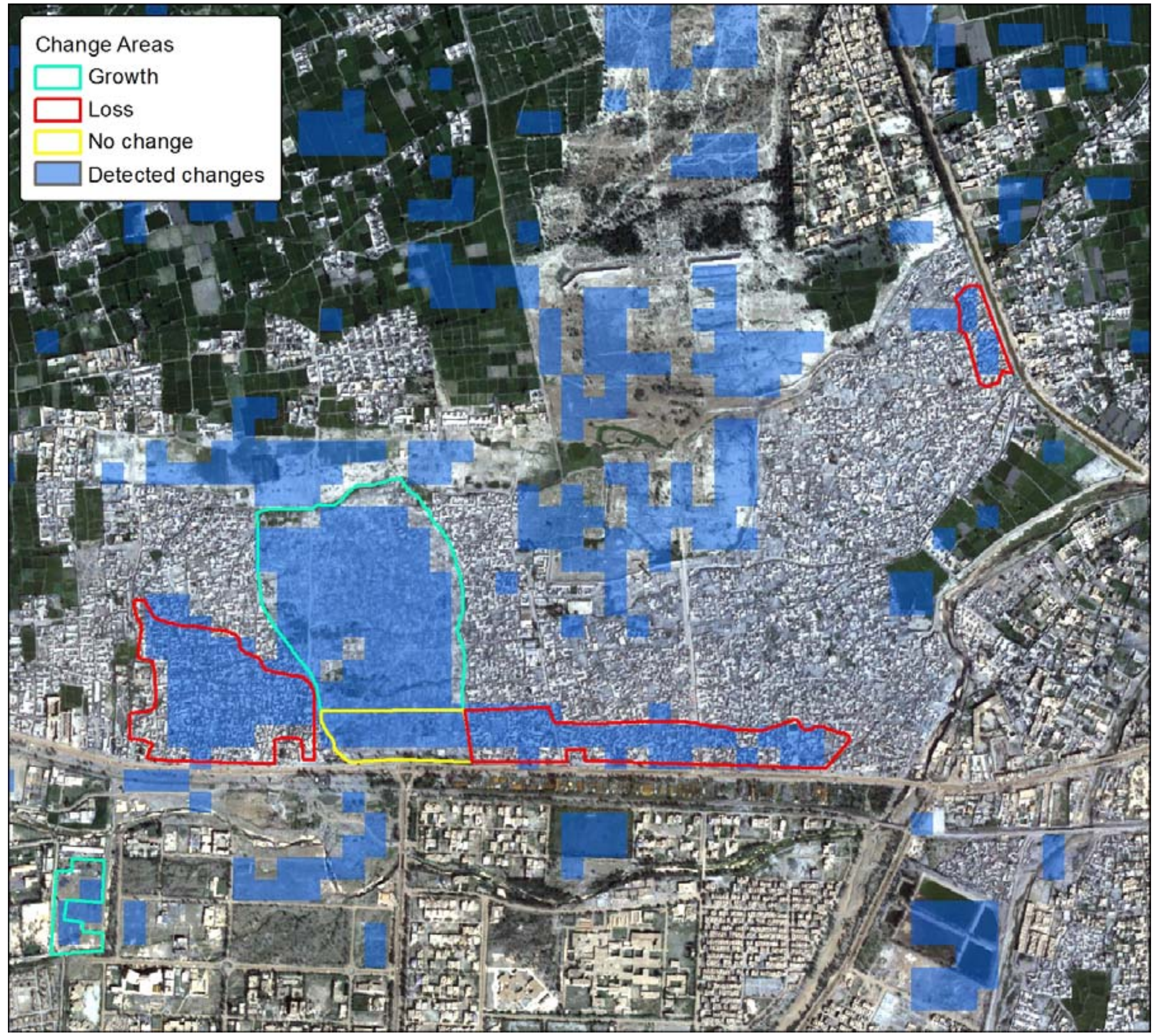
(a) 2004 (a) 2009

Comparison of Performance



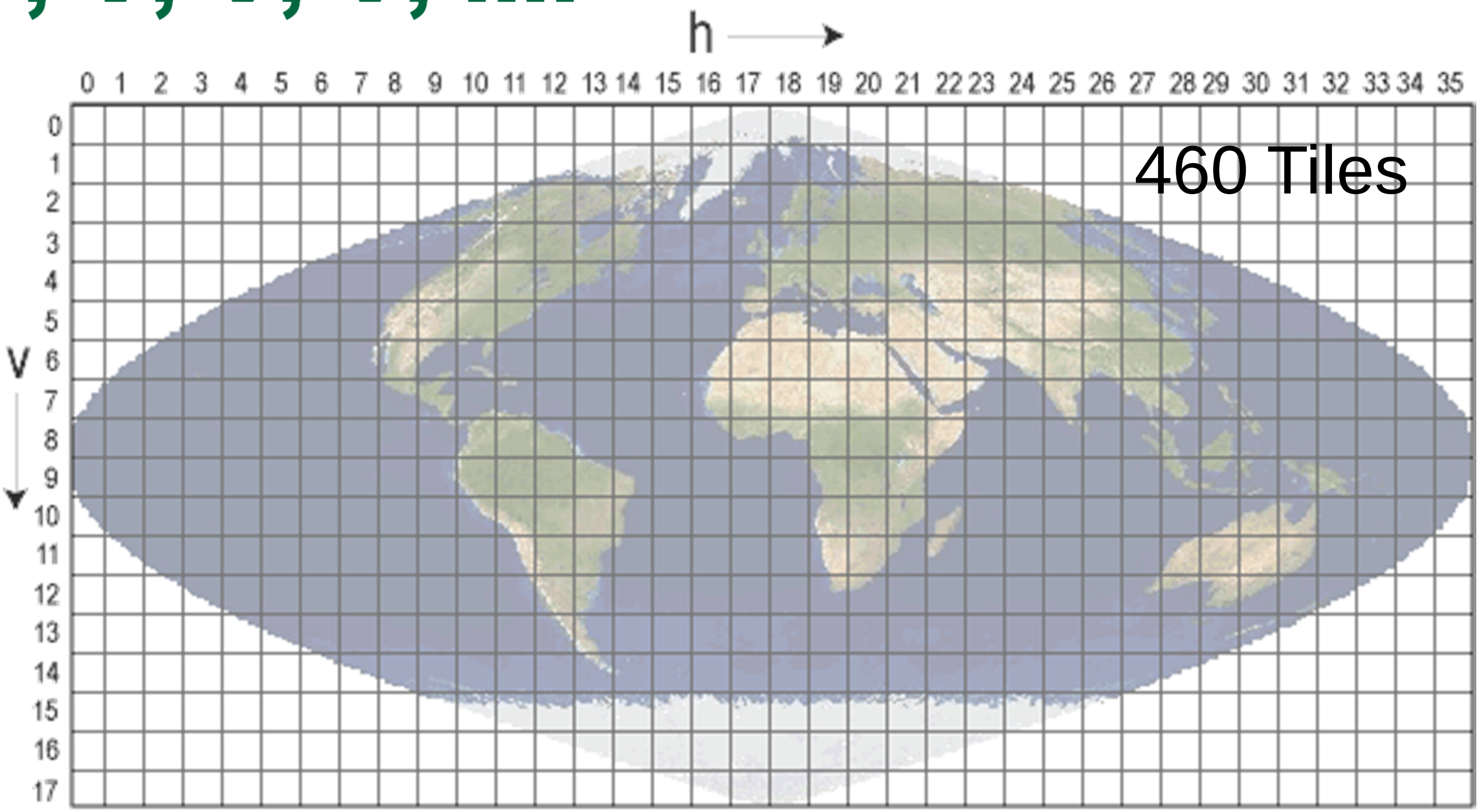
Difference

Ratio



Probabilistic

V, V, V, V, ….



- Each Tile = 4800 x 4800 = 23,040,000 (250m)
- (1m) => 1,440,000,000,000 = 1,373,291 MB
- Bands = 1 ~ 240; Derived Features ~ 250
- Temporal = ~ 18-22 days; 10’s of satellites

Thematic Classification

- Increasing spectral resolution: 4 to 224 Bands
- Challenges
  - #of training samples ~ (10 to 30) * (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
  - Aggregate Classes (Agriculture – Corn, Soybean, ...)
  - Spatial autocorrelation

Solution: Semi-supervised Learning

EM to estimate GMM parameters

- **E-Step**

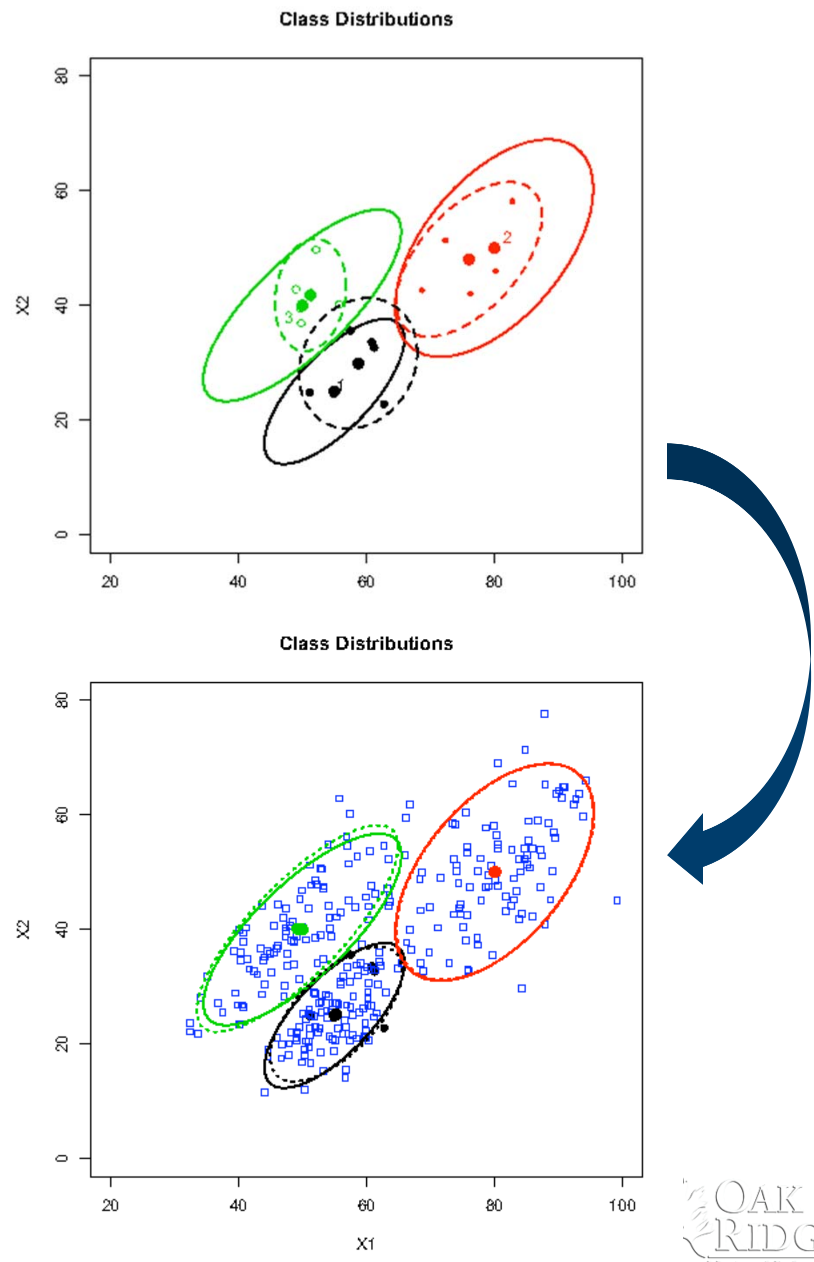
$$e_{ij} = \frac{\left|\hat{\Sigma}_j^k\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(x_i - \hat{\mu}_j^k\right)^T \hat{\Sigma}_j^{-1,k}\left(x_i - \hat{\mu}_j^k\right)\right\}}{\sum_{l=1}^{M}\left|\hat{\Sigma}_l^k\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(x_i - \hat{\mu}_l^k\right)^T \hat{\Sigma}_l^{-1,k}\left(x_i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

- **M-Step**

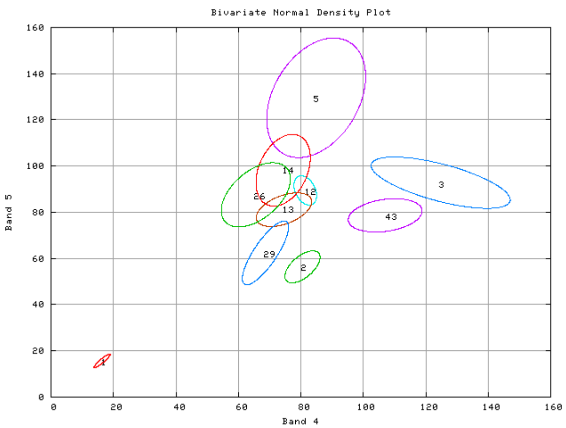
$$\alpha_j = \frac{\sum_{i=1}^{N} e_{ij}}{N}, \qquad \hat{\mu}_j^{k+1} = \frac{\sum_{i=1}^{N} e_{ij} x_i}{\sum_{i=1}^{N} e_{ij}},$$

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

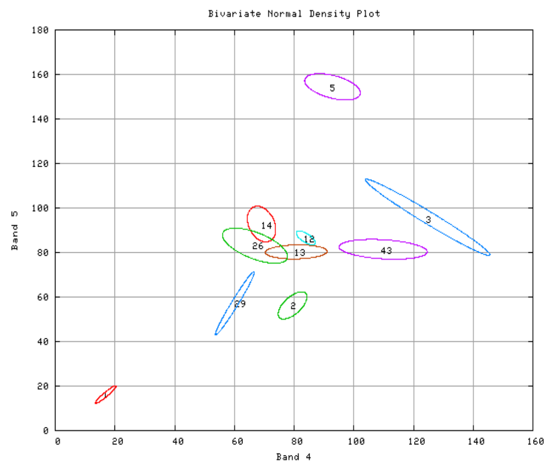
i^{th} data vector, j^{th} class



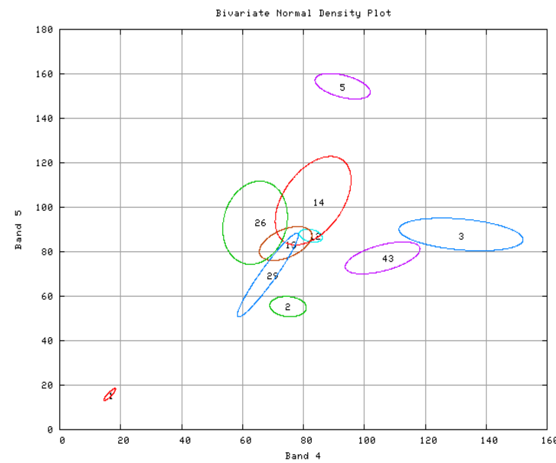
Semi-supervised Learning



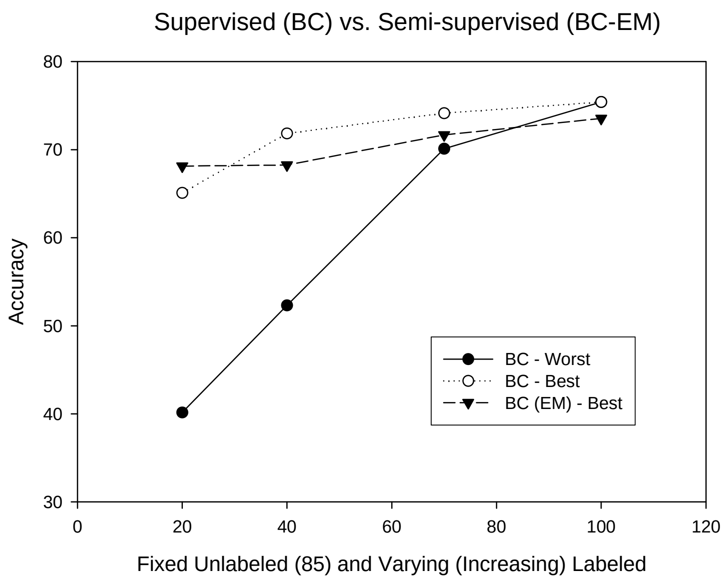
10 Classes, 100 Training Samples
(10-30) x No of dimensions / class



Small Subset of 20
Training Samples



20 labeled + 80
unlabeled samples



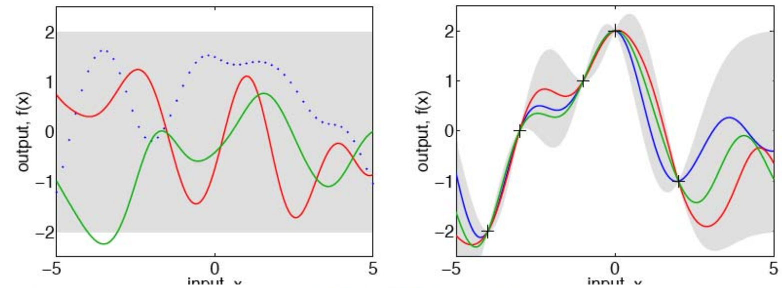
Ranga Raju Vatsavai, Shashi Shekhar, Thomas E. Burk: A Semi-Supervised Learning Method for Remote Sensing Data Mining. ICTAI 2005: 207-211

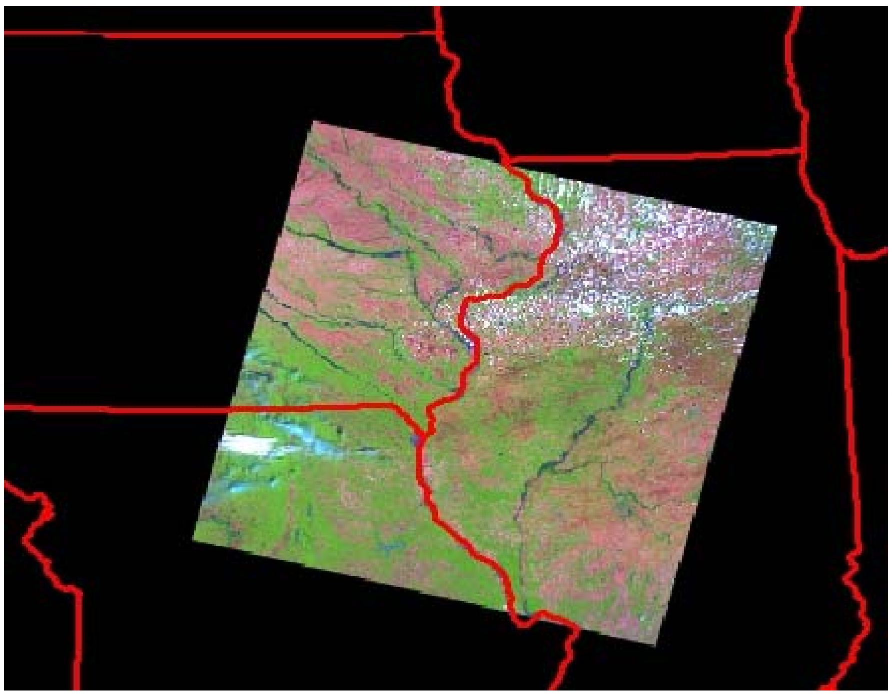
Solution: Gaussian Process (GP) Classification

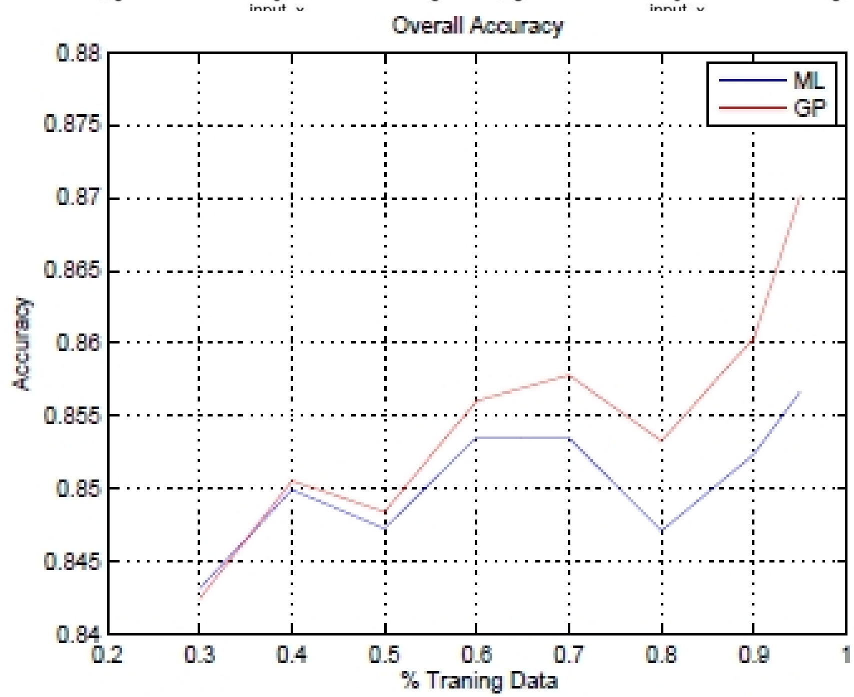
- Change of distribution over space is modeled by

$$p(x\,|\,y) \sim N(\mu, \Sigma)$$

$$p(x(s)\,|\,y) \sim N(\mu(s), \Sigma(s))$$

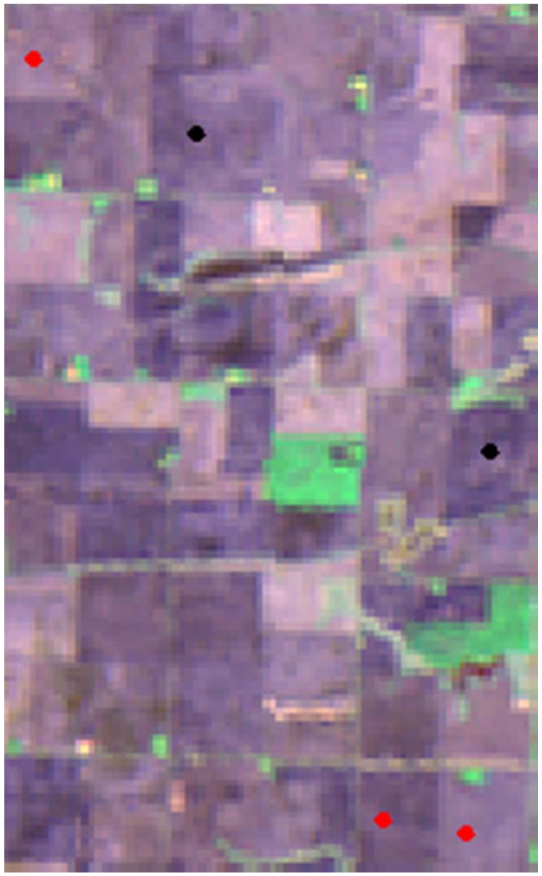




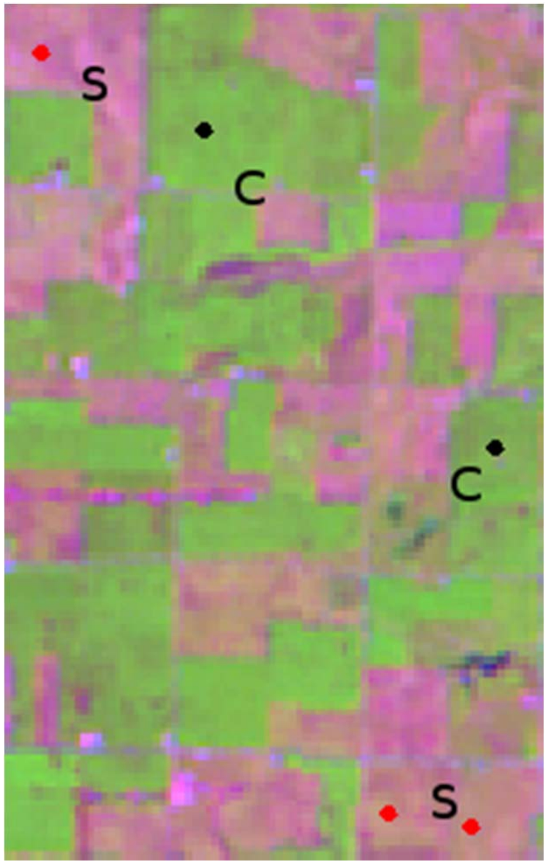


Goo Jun, Ranga Raju Vatsavai, Joydeep Ghosh: Spatially Adaptive Classification and Active Learning of Multispectral Data with Gaussian Processes. SSTDM 2009: 597-603

Challenge: Multi-temporal Classification



AWiFS (May 3, 2008; FCC (4,3,2))

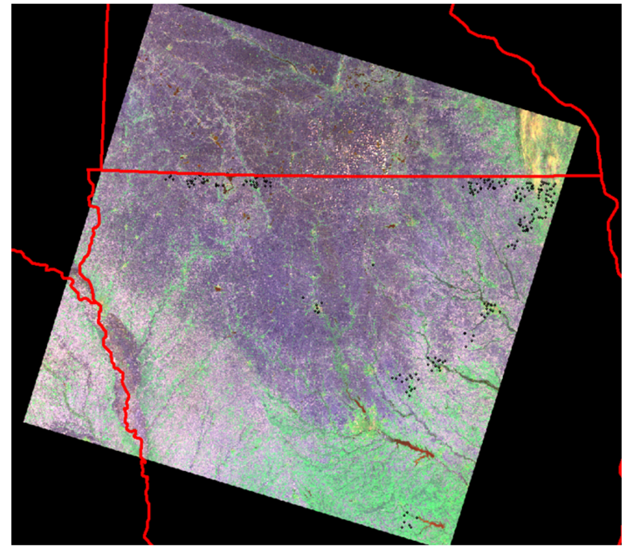


AWiFS (July 14, 2008; FCC (4,3,2))

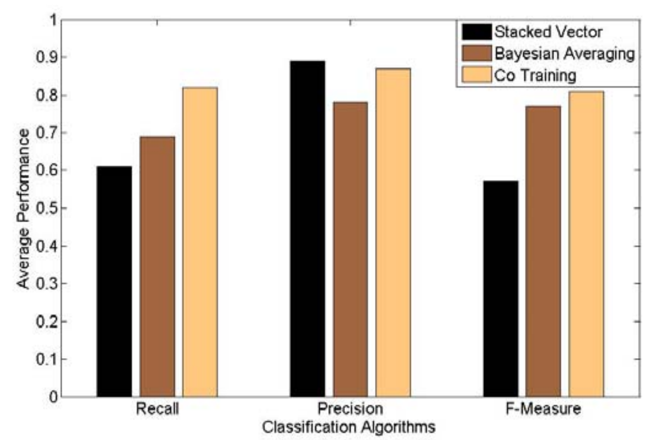
Thematic Classes: C-Corn, S-Soy

Multi-view Approach

- **Multi-temporal images are different views of same phenomena**
  - **Learn single classifier on different views, chose the best one through empirical evaluation**
  - **Combine different views into a single view, train classifier on single combined view – stacked vector approach**
  - **Learn classifier on single view and combine predictions of individual classifiers – multiple classifier systems**
    - **Bayesian Model Averaging**
  - **Co-training**
    - **Learn a classifier independently on each view**
    - **Use predictions of each classifier on unlabeled data instances to augment training dataset for other classifier**



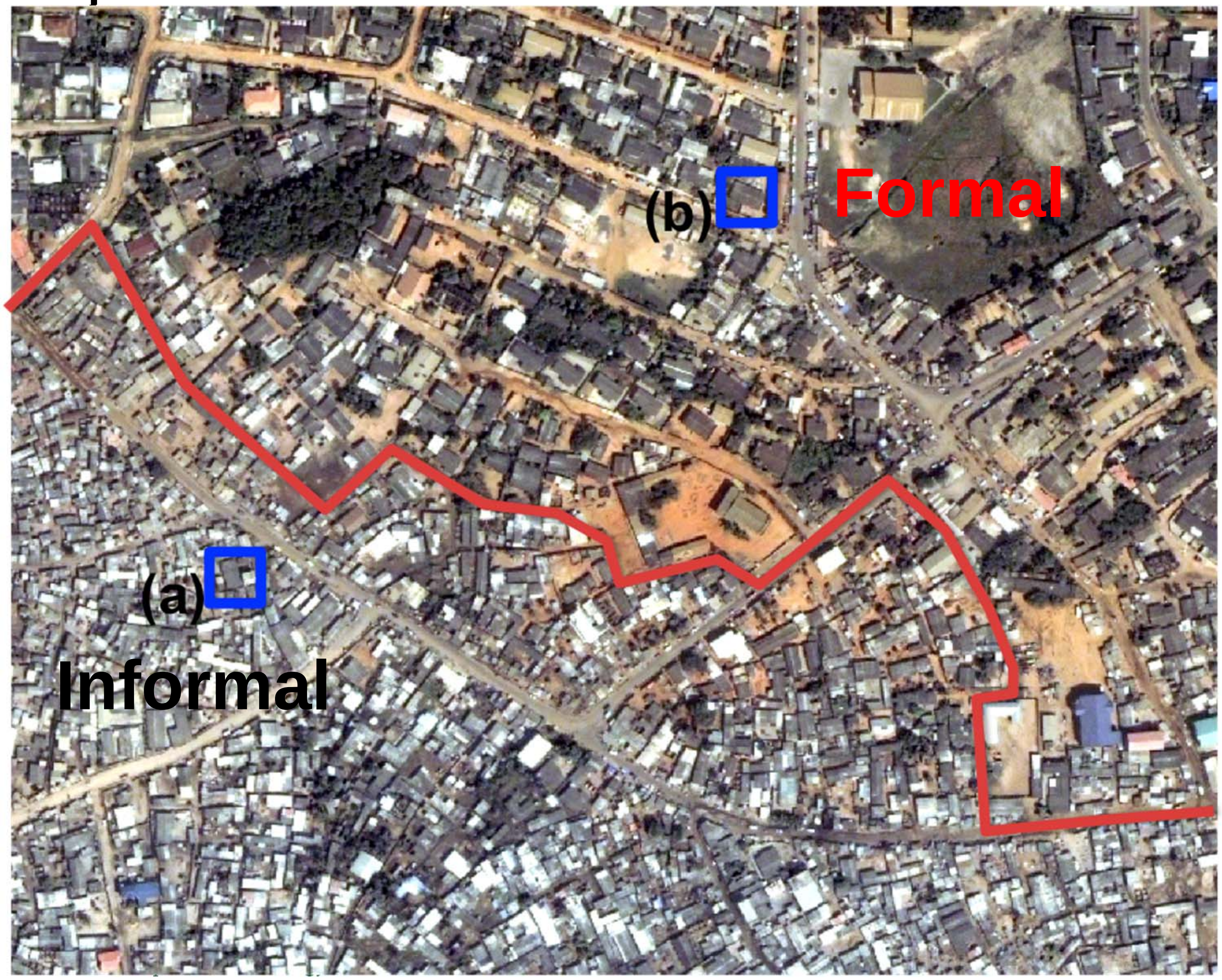
Class	Training	Validation
Corn	261	261
Soybean	225	225
Alfa alfa	27	27
Grass	189	180
Water	18	18
Developed	90	99
Deciduous Forest	117	117
Wetlands Forest	18	36
Total:	945	963

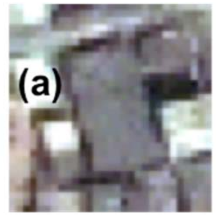


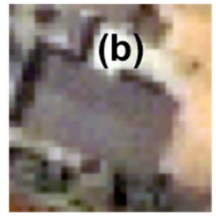
<u>Varun Chandola, Ranga Raju Vatsavai: Multi-temporal remote sensing image classification - A multi-view approach. CIDU 2010: 258-270</u>

Beyond Pixels and Objects: Complex Patterns

- Classes that cannot be separated by looking at pixels in isolation





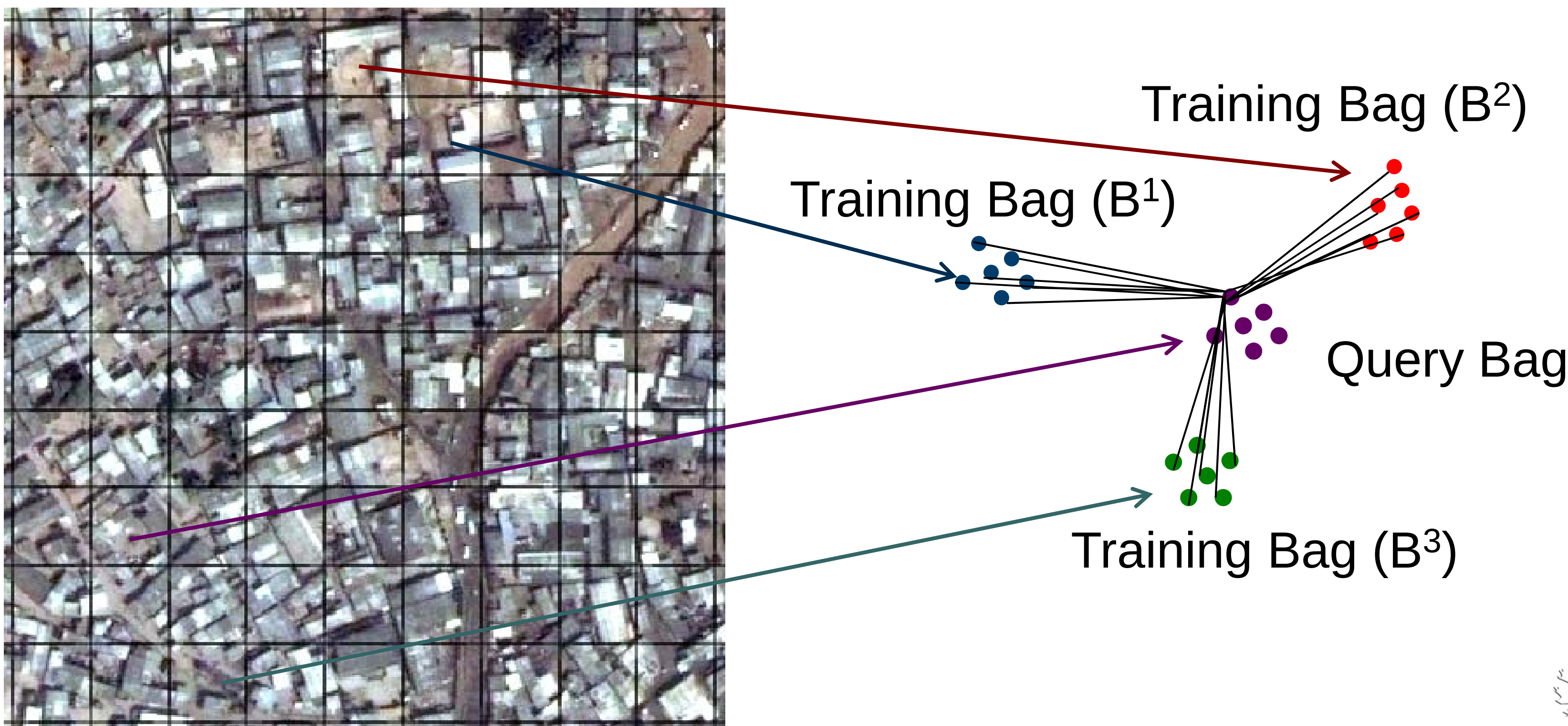


- Objects may be same (e.g., Buildings, Roads, …), but not the spatial patterns (neighborhoods)

Objective: Finding Complex Patterns

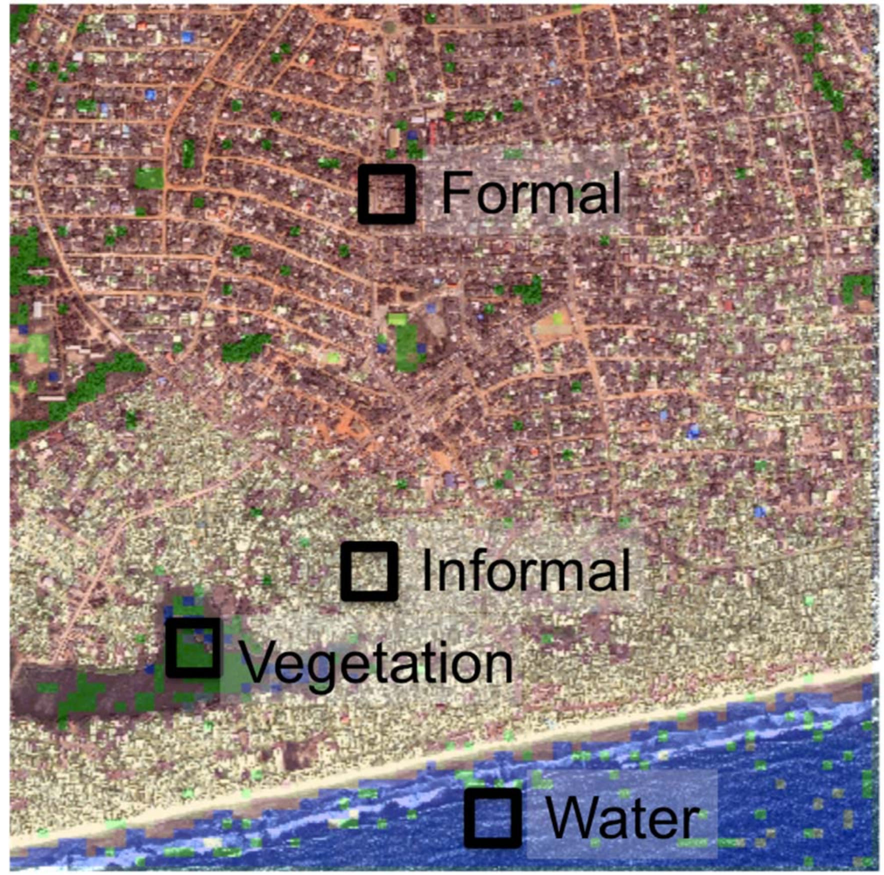
Multiple Instance Learning

- Reasoning with content in a grid (block, segment, object) rather than aggregate (average) information



Results

FCC Image

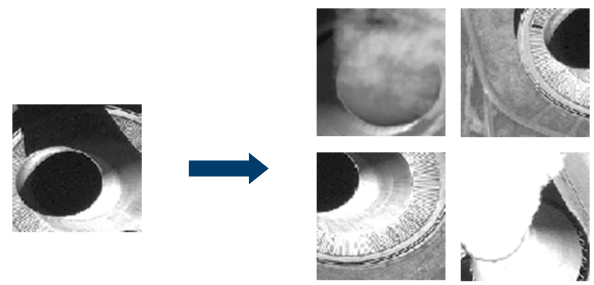


MIL Classified Image Overlay on FCC

Searching for patterns

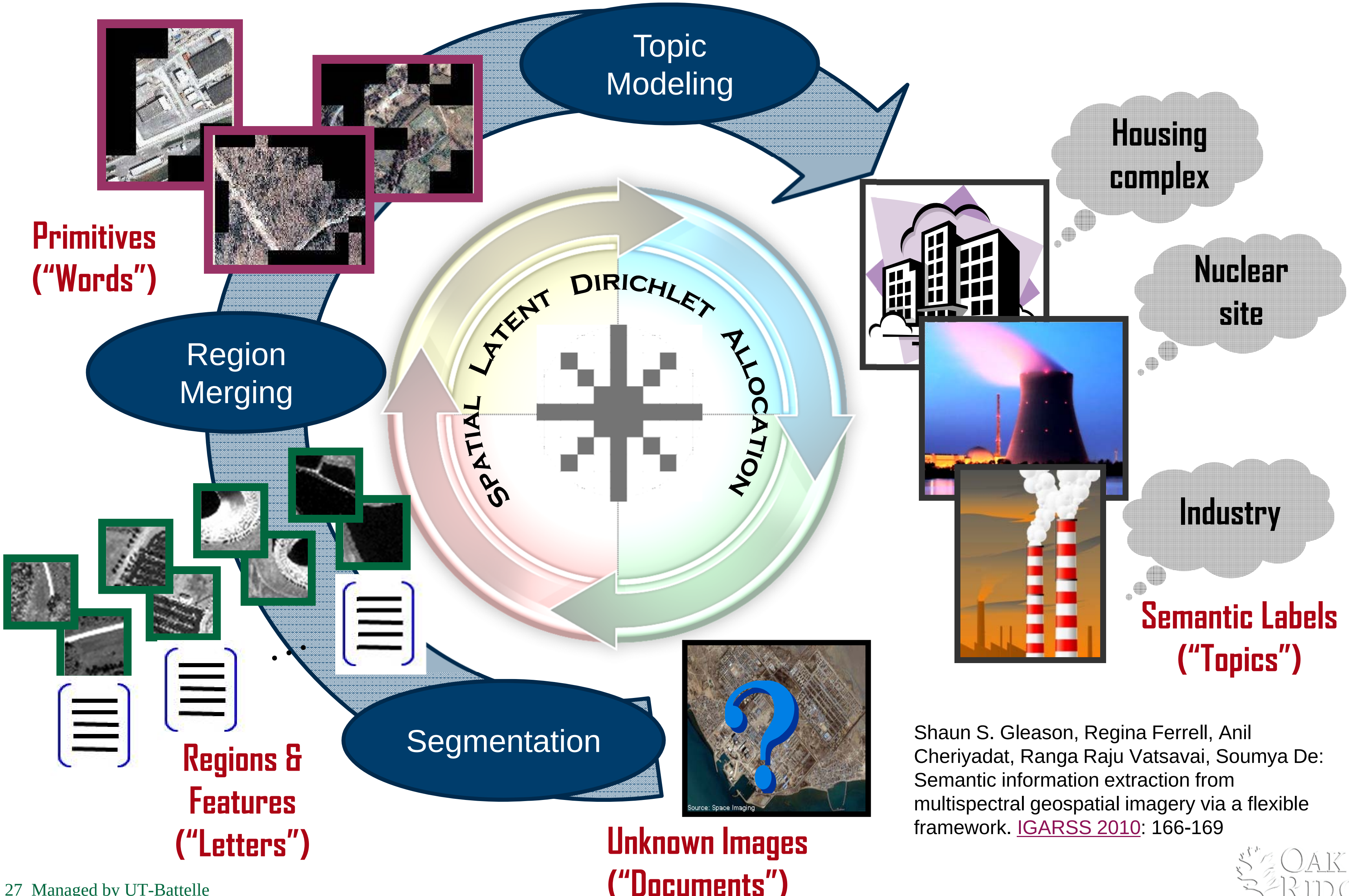
- **Single Category Detection**
  - **Predict if a given visual category is present in a given image**

- **Content based image retrieval**
  - **Given query image, find similar images**



- **Structure Recognition**
  - **Structurally distinct objects within one class**

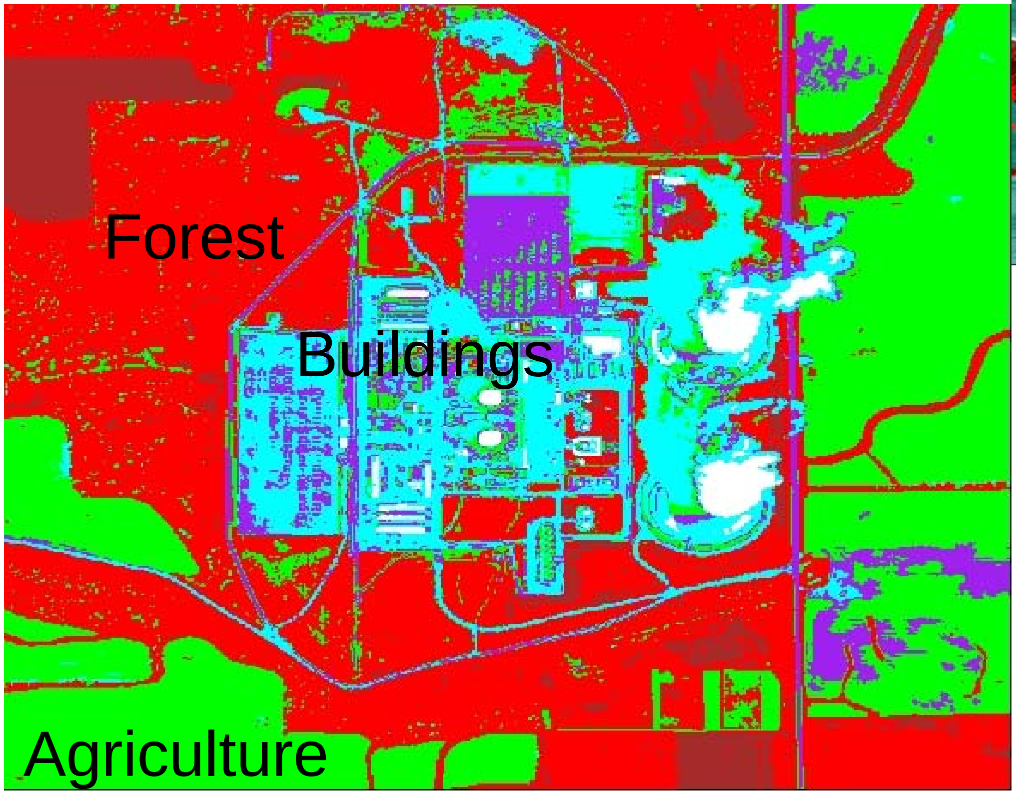
Goal: Turn image pixels into semantic information for the analyst…



Shaun S. Gleason, Regina Ferrell, Anil Cheriyadat, Ranga Raju Vatsavai, Soumya De: Semantic information extraction from multispectral geospatial imagery via a flexible framework. IGARSS 2010: 166-169

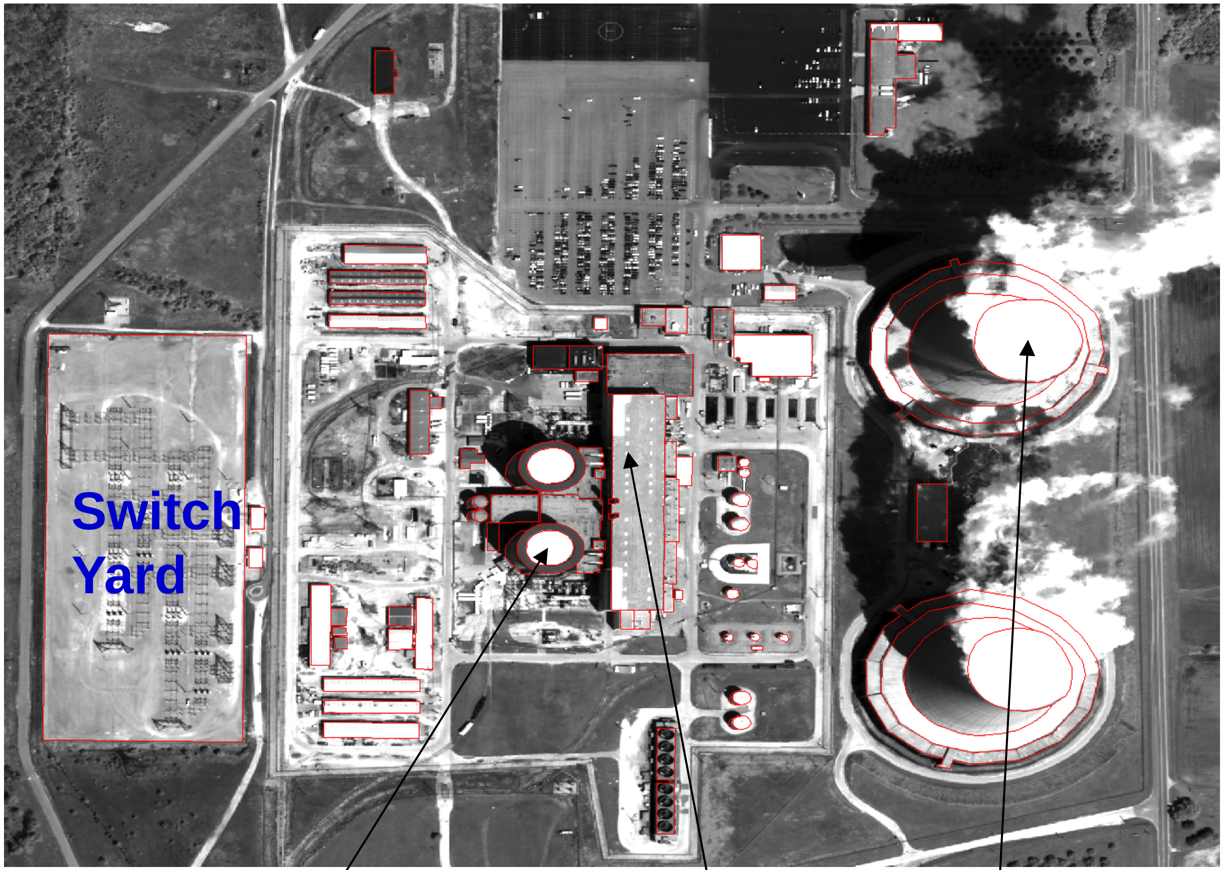
Image Classification

- **How about high-resolution images and semantic labels?**



- **Does this kind of thematic classification make sense for indentifying nuclear power plant? Can these thematic classes imply above image as nuclear plant?**

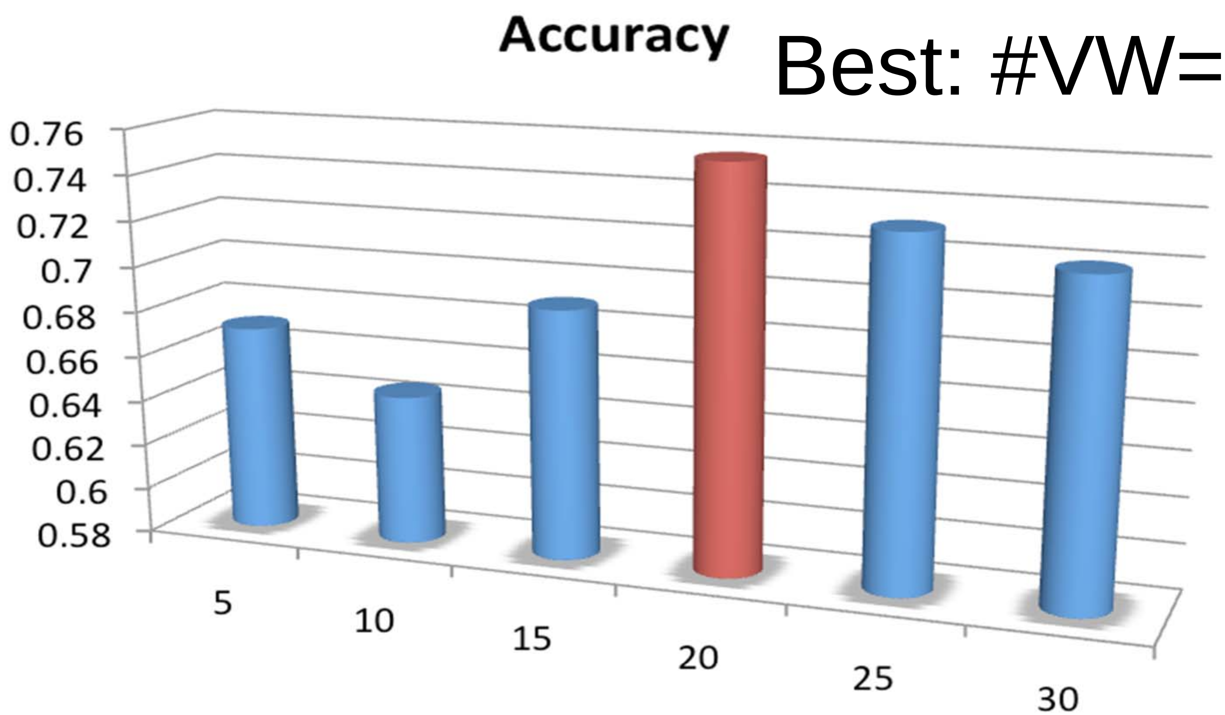
What is missing?



Semantics:
Set of objects like:
Switch yard,
Containment
Building,
Turbine
Generator,
Cooling
Towers
AND
Their spatial
arrangement
=> may
imply a semantic
label like "nuclear
power plant"

No of Visual Words and topics

Best: #VW=25, K=20

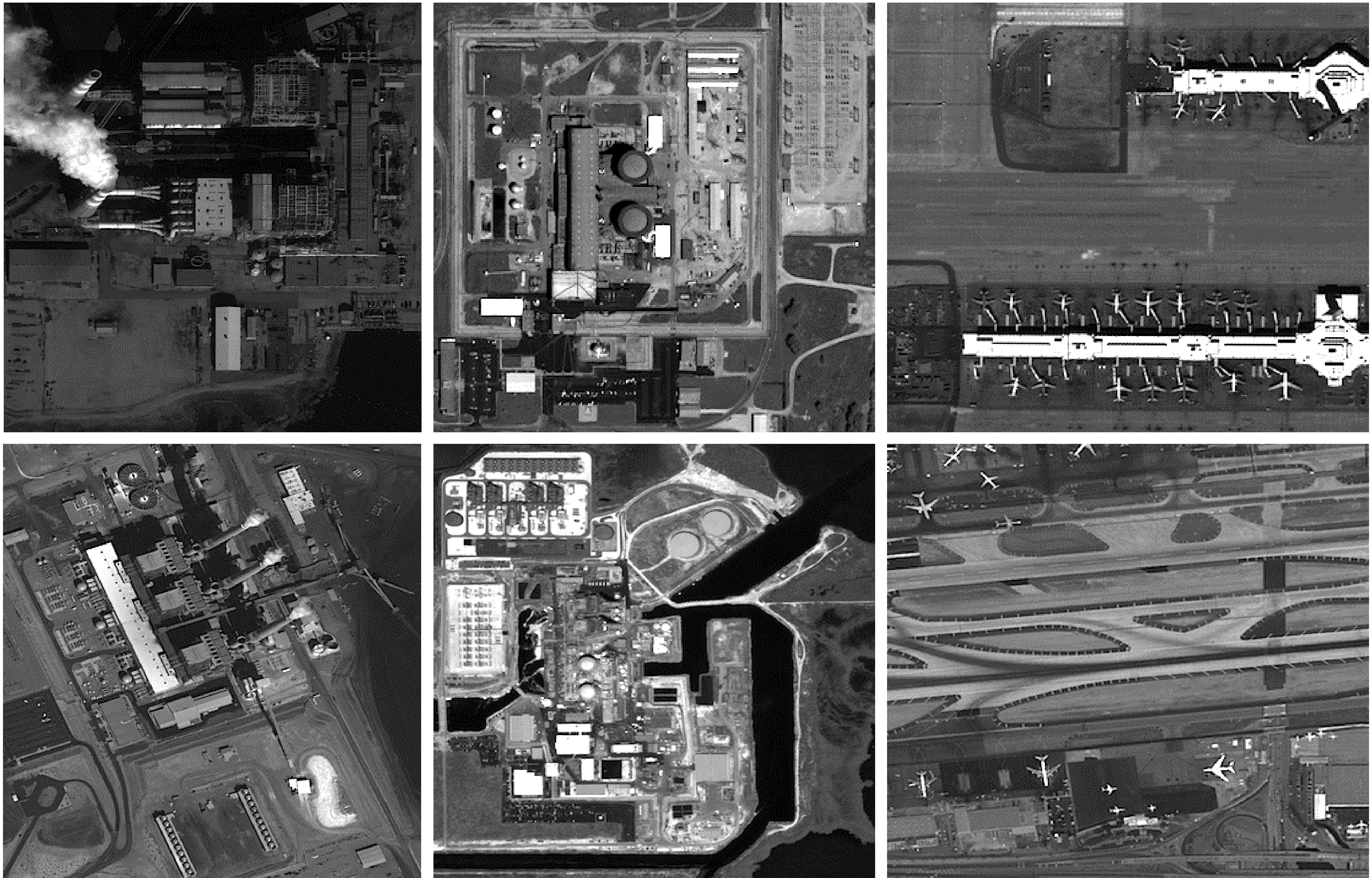


	5	10	15	20	25	30
Accuracy	0.671	0.646	0.69	0.756	0.732	0.72

 # Topics

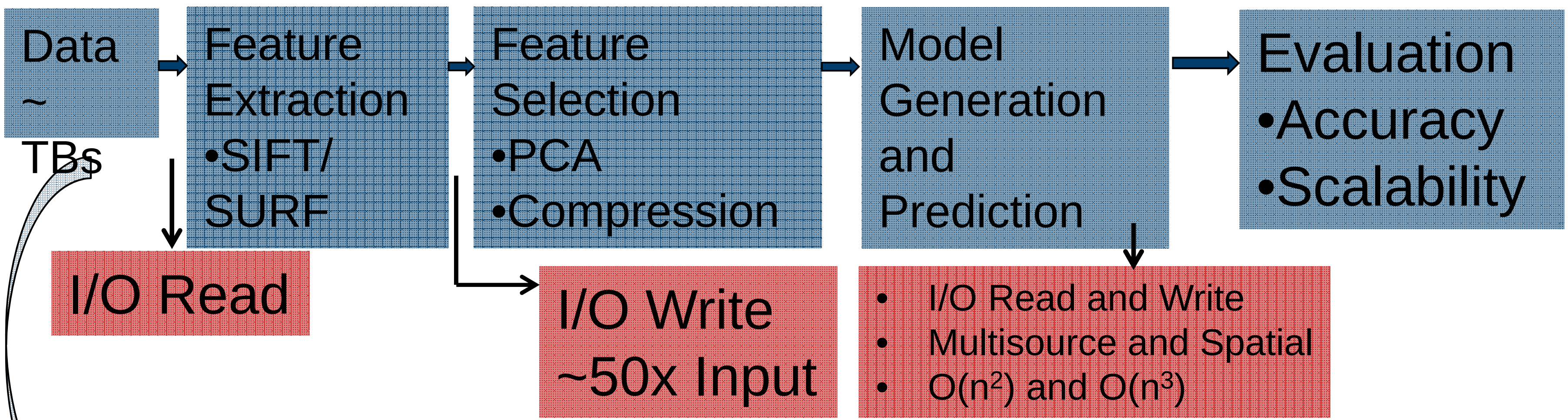
Ranga Raju Vatsavai, Anil Cheriyadat, Shaun S. Gleason: Unsupervised Semantic Labeling Framework for Identification of Complex Facilities in High-Resolution Remote Sensing Images. SSTDM 2010: 273-280

Coal, Nuclear, Airport Images



Part 2: Computational Challenges

Computational and I/O challenges



Source	Dataset Characteristics	Volume
Overhead Images	•Resolution: High (0.6 to 30 m) and moderate (56 m to 1 km)	0.5 PB (image size with features ranges from a GB to TB)
Terrestrial Images	•Small sized photographs: 12 million images (web scale: ~1 Billion images)	2 TB (images range from few KB to 0.5 MB)

GP Change Detection – Computational Challenges

- **Size of the covariance matrix grows quadratic with length of time series**
- **Need to compute**

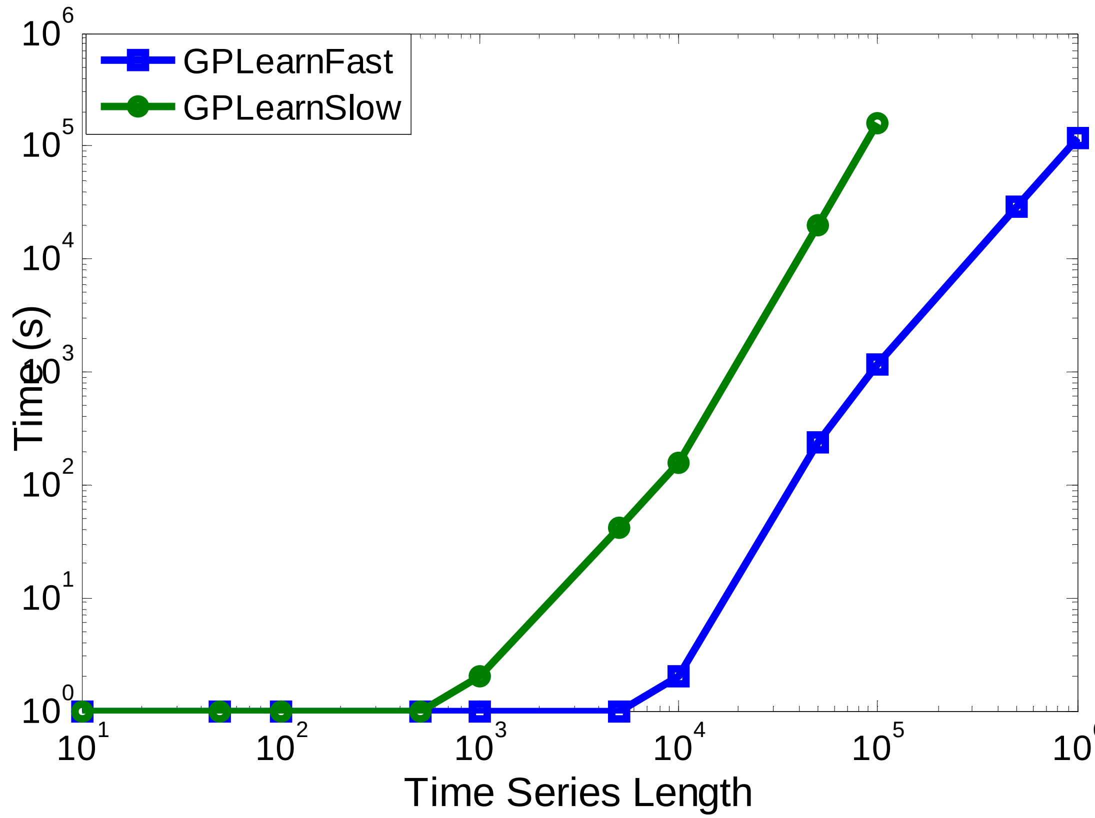
$$K^{-1}\boldsymbol{y} \qquad \log|K| \qquad tr\left(K^{-1}\frac{\partial K}{\partial \theta}\right)$$

  - **Standard methods are $O(t^3)$ and require $O(t^2)$ memory**
  - **Not suitable for big time series**
  - **Hyper-parameter estimation for p time series simultaneously is $O(p*t^3)$**

- **AWiFS Satellite Data – Global spatial : 56m, Temporal: 5 days**
- **MODIS – 250m Temporal: 1 day**
- **Eddy Flux Sensors – Temporal: 15 minutes**
- **ECG Time Series – Temporal: ~ 0.2sec**

GP Change: Sequential Results

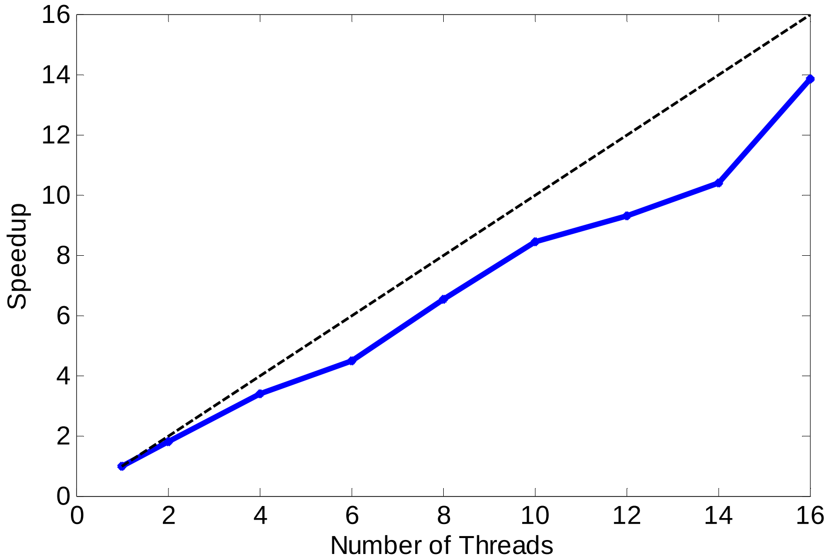
- Compared with Cholesky decomposition based solution
- C implementation
  - CBLAS library for basic operations
  - CLAPACK library for Cholesky decomposition

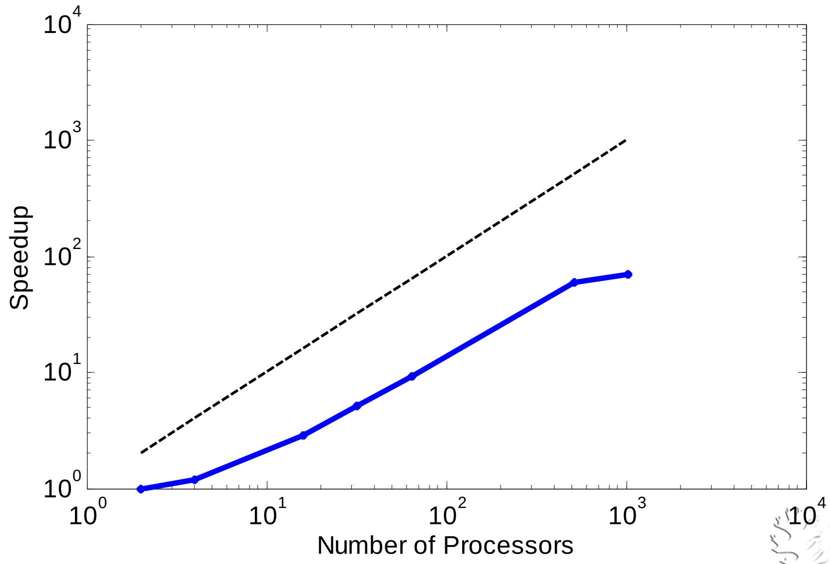


Oak Ridge National Laboratory

Parallelization Results

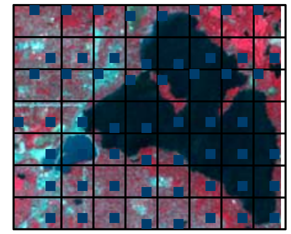
- System
  - FROST: An SGI Altrix ICE 8200 Cluster at ORNL
  - 128 compute nodes each having 16 virtual cores and 24 GB of RAM
- Task is to estimate hyper-parameters of 1 million NDVI time series

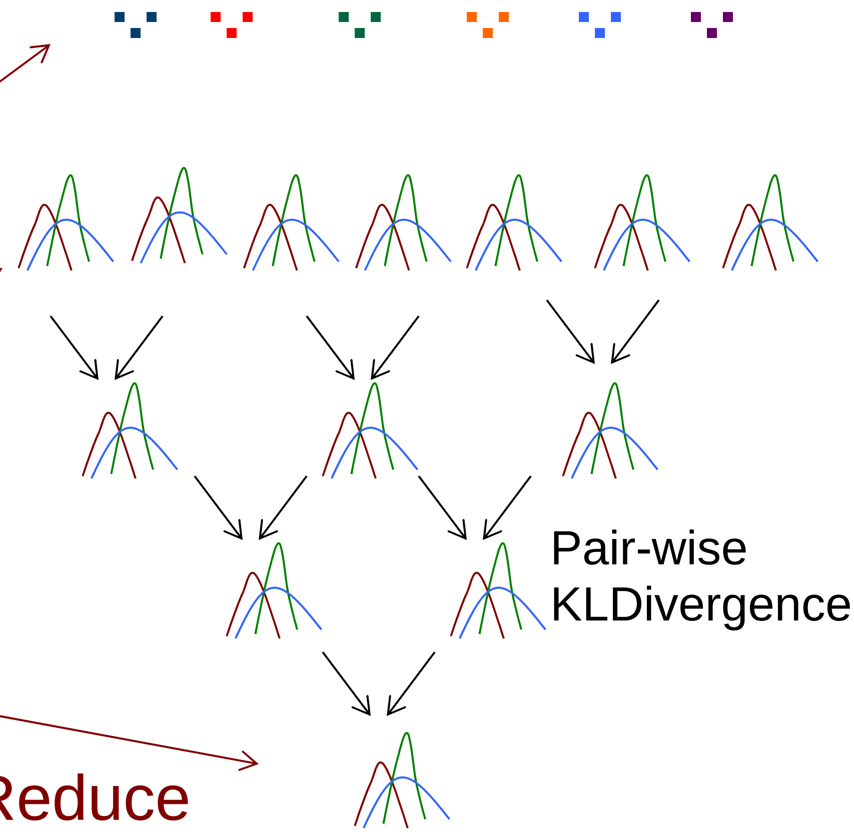




Distributed GMM Clustering

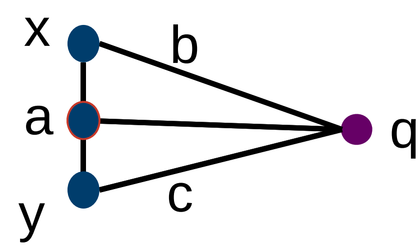
- Expectation Maximization is a local optimization algorithm
  - Different initialization
- Multiple sampling
- Local model at each node
- Global model from local models





Shared Memory, GPUs, ….

- On GPU (GTX 285, 240 CUDA Cores, 1GB)
  - GMM Clustering ~ 20x to 70x
- Multiple Instance Learning
  - Citation-KNN based approach: O(n^2Nd)
  - Can we reduce “n”: O(Nd)$\sqrt{n}$
  - Multidimensional Spectral Hashing



For $b < c$:

Length of median is bounded by:

$$b < m_a < c$$

From Apollonius' Theorem:

$$m_a = \sqrt{\frac{2b^2 + 2c^2 - a^2}{4}}$$

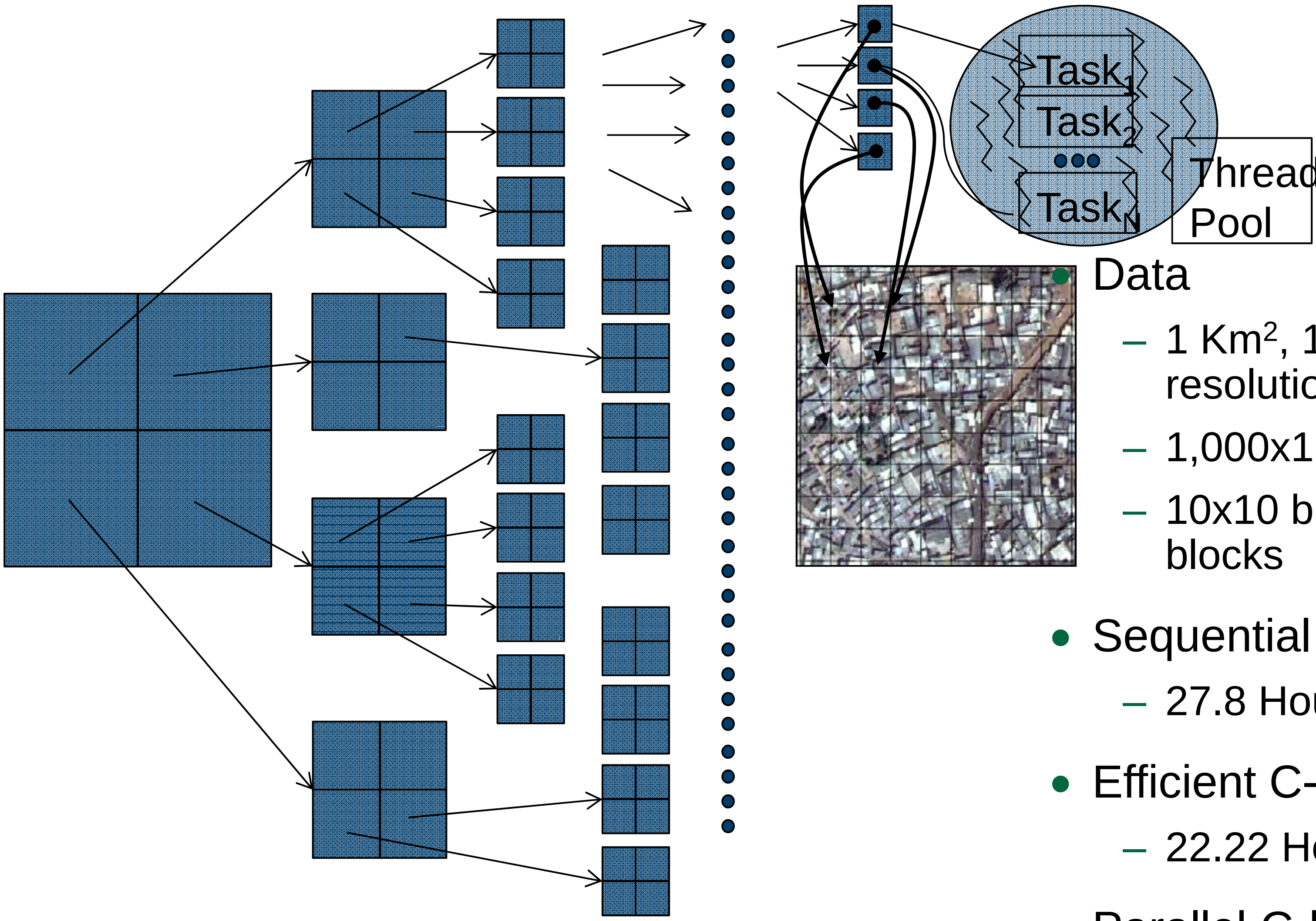
$\therefore$ for small $a : (a \to 0)$

$$b \le m_a \le c$$

$$m_a = \sqrt{\frac{2(b^2 + c^2)}{4}} = \sqrt{\frac{b^2 + c^2}{2}} = b = c$$

Parallel Citation-KNN

- Divide and conquer parallelization strategy



- Data
  - 1 Km^2, 1m pixel resolution, 3 bands
  - 1,000x1,000: 1M pixels
  - 10x10 block: 10K blocks
- Sequential Performance
  - 27.8 Hours
- Efficient C-kNN
  - 22.22 Hours
- Parallel C-kNN
  - 2.62 Hours

Conclusions

- Developed several innovative solutions that address big spatiotemporal data challenges
  - Semi-supervised learning
  - Spatial Classification
  - Temporal Classification
  - Complex Pattern Classification
  - Semantic Classification
- Monitoring large regions
  - Online, Scalable
- Parallel Algorithms are needed if this framework needs to be operational
  - Shared memory
  - Distributed memory
  - Cloud computing

Future Directions

- Big geospatial data management and analytics
  - National security (Imagery, Live video feeds, Sensors, Web, …)
  - Social media mining (12TB Tweets/day)
- Massively Parallel Processing DBMS
  - Greenplum: spatial data extensions
- NoSQL Databases for spatiotemporal data
  - Extensions: SciDB, CouchDB, …
- Hadoop/Hive, …
- Parallelization of data mining/machine learning algorithms on heterogeneous architectures
  - Integration with parallel I/O (ADIOS)

Questions/Comments/Suggestions?
Email: vatsavairr@ornl.gov

Collaborators:

B. Bhaduri, V. Chandola, E. Bright, M. Tuttle, A. Cheriyadat, J. Graesser, R. Sukumar, C. Symons, S. Klasky

Research Supported by:

ORNL/LDRD, DOE/NNSA, Unnamed