Redefining Class Definitions using Constraint-Based Clustering

An Application to Remote Sensing of the Earth's Surface

Dan Preston
Stanford University

Carla Brodley
Tufts University

Roni Khardon
Tufts University

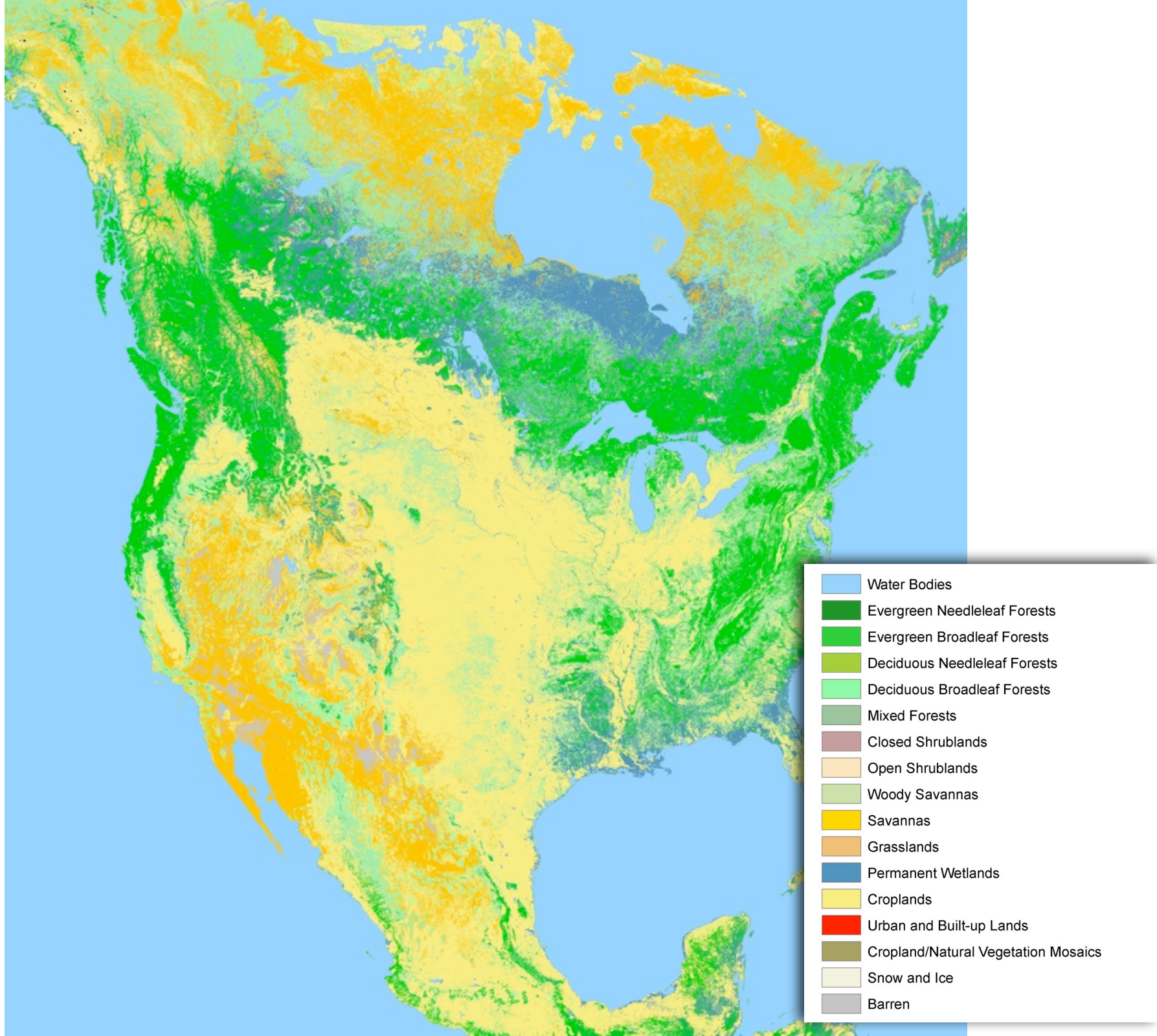
Mark Friedl
Boston University

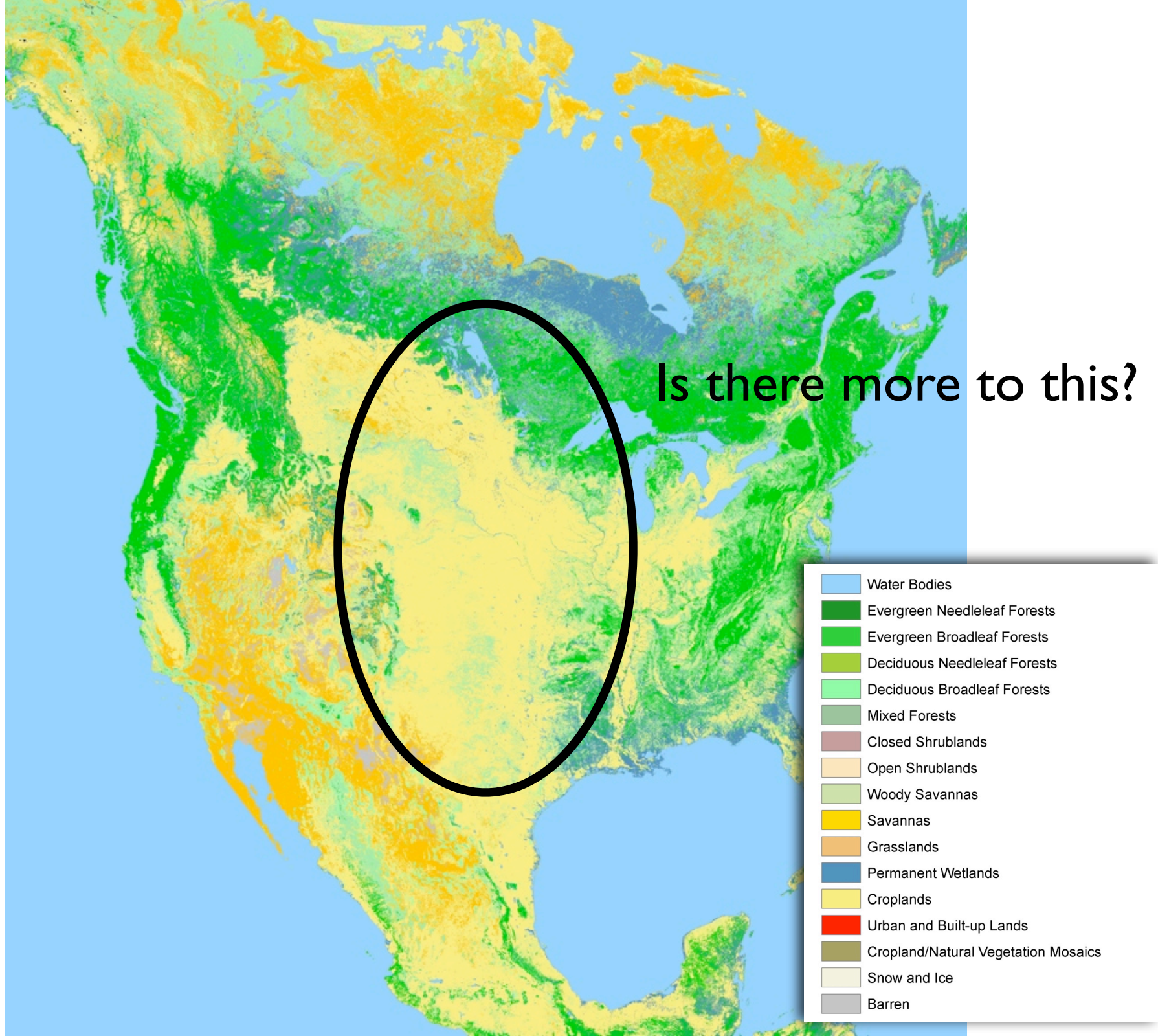
Damien Sulla-Menashe
Boston University

KDD 2010

Where do classes come from?

- Categories and concepts that people find useful
  - May not be supported by the features
- Clustering to find the homogeneous groups in the data
  - May not be of use to humans
  - Many different (equally "good") clusterings
  - How many clusters are there?


Water Bodies
Evergreen Needleleaf Forests
Evergreen Broadleaf Forests
Deciduous Needleleaf Forests
Deciduous Broadleaf Forests
Mixed Forests
Closed Shrublands
Open Shrublands
Woody Savannas
Savannas
Grasslands
Permanent Wetlands
Croplands
Urban and Built-up Lands
Cropland/Natural Vegetation Mosaics
Snow and Ice
Barren


Is there more to this?
Water Bodies
Evergreen Needleleaf Forests
Evergreen Broadleaf Forests
Deciduous Needleleaf Forests
Deciduous Broadleaf Forests
Mixed Forests
Closed Shrublands
Open Shrublands
Woody Savannas
Savannas
Grasslands
Permanent Wetlands
Croplands
Urban and Built-up Lands
Cropland/Natural Vegetation Mosaics
Snow and Ice
Barren


Do the features support these distinctions?
Water Bodies
Evergreen Needleleaf Forests
Evergreen Broadleaf Forests
Deciduous Needleleaf Forests
Deciduous Broadleaf Forests
Mixed Forests
Closed Shrublands
Open Shrublands
Woody Savannas
Savannas
Grasslands
Permanent Wetlands
Croplands
Urban and Built-up Lands
Cropland/Natural Vegetation Mosaics
Snow and Ice
Barren

Our Approach

- Probabilistic constraint clustering using:
  - original labels as constraints
  - expert-belief as a guide
- A metric to evaluate a clustering result (in a constrained setting)
  - *...and determine K*
- Application to Remote Sensing

Constraint-Based Clustering: A Brief Review

- Constraints given as instance pairs
- Hard constraints for k-means [Wagstaff, et al 2001]
- Hard constraints with EM [Shental, et al, 2003]
- Probabilistic must-link constraints [Law, et al 2004]:
  expert defines group membership of each instance for n groups
- PPC: Probabilistic must and cannot link constraints, fast approximations [Lu and Leen, 2007]

What are our constraints?

Supplying Expert Belief

- The "C"-matrix allows the domain expert to supply preferences that label pairs exist or do not exist together:

Non-diagonal {
- C(A,B) = 1.0: merge classes A and B
- C(A,B) = -1.0: keep classes A and B separate
- C(A,B) = 0.0: expert has no opinion

Diagonal {
- C(A,A) = 1.0: do not split class A
- C(A,A) < 1.0: ok to split A
- C(A,A) = 0.0: ignore the labels
- C(A,A) = -1.0: nonsense

Expert-Belief Matrix: Example

Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1 evergreen needleleaf forests	1.00	-0.60	-0.80	0.80	0.00	-0.80	0.60	0.00	-0.80	0.00	-0.80	-0.40	-0.80	-1.00	-0.60
2 evergreen broadleaf forests	-0.60	1.00	-0.60	-0.60	0.00	-0.80	0.60	0.00	-0.80	0.00	-0.60	0.00	-1.00	-1.00	-0.80
3 deciduous broadleaf forests	-0.80	-0.60	1.00	0.80	0.00	-0.80	0.60	0.00	-0.80	0.00	-0.60	0.00	-0.80	-1.00	-0.80
4 mixed forests	0.80	-0.60	0.80	1.00	0.00	-0.80	0.60	0.00	-0.80	0.00	-0.60	0.00	-0.80	-1.00	-0.80
5 closed shrublands	0.00	0.00	0.00	0.00	1.00	0.00	-0.40	-0.40	-0.40	-0.60	-0.60	0.00	-0.80	-0.80	-0.80
6 open shrublands	-0.80	-0.80	-0.80	-0.80	0.00	1.00	-0.20	-0.40	0.40	-0.60	-0.40	0.00	-0.60	0.20	-1.00
7 woody savannas	0.60	0.60	0.60	0.60	-0.40	-0.20	1.00	0.80	0.00	-0.60	-0.40	0.00	-0.80	-0.80	-1.00
8 savannas	0.00	0.00	0.00	0.00	-0.40	-0.40	0.80	1.00	0.60	-0.60	-0.40	0.00	-0.80	-0.80	-1.00
9 grasslands	-0.80	-0.80	-0.80	-0.80	-0.40	0.40	0.00	0.60	1.00	-0.60	0.40	0.40	-0.80	-0.20	-1.00
10 permanent wetlands	0.00	0.00	0.00	0.00	-0.60	-0.60	-0.60	-0.60	-0.60	1.00	0.00	0.00	-1.00	-1.00	0.60
11 croplands	-0.80	-0.60	-0.60	-0.60	-0.60	-0.40	-0.40	-0.40	0.40	0.00	1.00	0.80	-0.80	-0.80	-1.00
12 urban and built-up lands	-0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.80	1.00	-0.80	-0.80	-1.00
13 natural vegetation mosaics	-0.80	-1.00	-0.80	-0.80	-0.80	-0.60	-0.80	-0.80	-0.80	-1.00	-0.80	-0.80	1.00	-0.40	-1.00
14 barren	-1.00	-1.00	-1.00	-1.00	-0.80	0.20	-0.80	-0.80	-0.20	-1.00	-0.80	-0.80	-0.40	1.00	-0.80
15 water bodies	-0.60	-0.80	-0.80	-0.80	-0.80	-1.00	-1.00	-1.00	-1.00	0.60	-1.00	-1.00	-1.00	-0.80	1.00

"C"-Matrix

Incorporating Constraints into EM

$$q_{ik} \sim P_k P(\vec{y_i}|\theta_k) \exp\left(2 \sum_{j=1, j\neq i} \lambda C(\ell_i, \ell_j) q_{jk}\right)$$

$$\vec{\Sigma_k} = \frac{\sum_{i=1}^{N} q_{ik}(\vec{y_i} - \vec{\mu_k})(\vec{y_i} - \vec{\mu_k})^T}{\sum_{i=1}^{N} q_{ik}}$$

$$\vec{\mu_k} = \frac{\sum_{i=1}^{N} \vec{y_i} q_{ik}}{\sum_{i=1}^{N} q_{ik}}$$

$$P_k = \frac{1}{N} \sum_{i=1}^{N} q_{ik}$$

} Same as EM. Makes simplifying approximations

Algorithm Speed Up

- Compact Expert-belief matrix creates redundant computation in q_{ik}
- We can exploit this to reduce complexity from O(N^2) to O(NL):

$$S^*(\ell, k) = \sum_{i=1}^{N} \lambda C(\ell, \ell_i) q_{ik}$$

$$q_{ik} \sim P_k P(\vec{y_i}|\theta_k) \exp(2(S^*(\ell_i, k) - \lambda C(\ell_i, \ell_i) q_{ik}))$$

What is a good clustering?

Evaluating a clustering

- Heuristic criteria have a fit term and a complexity penalty term:

$$BIC = N \log(\frac{RSS}{N}) + k \log N$$

$$AIC = N \log(\frac{RSS}{N}) + 2k$$

- Our heuristic will use BIC and add a term for constraint adherence.

Assessing Constraint Adherence

- Create an L x L matrix *V*, where $V(A, B) \in [-1, 1]$ measures how frequently instances from class *A* and class *B* cluster together
  - *V(A,B) = 1*: all instances from classes *A* and *B* found together in the same cluster
  - *V(A,B) = -1*: no instances from *A* and *B* are in the same cluster

V(A,B) is a normalized count (to [-1,1]), where:

+1 for each pair of A,B that appear in the same cluster,
-1 for each that appear in different clusters.

Calculating Adherence

$$G(C, \{z_i\}) = \sum_{A \in L} \sum_{B \in L} (C(A, B) - V(A, B))^2$$

- Sum of the squared difference between the expert-defined constraints and the appearance measure V(A,B)
- Measures adherence to constraints

Constrained BIC

Fit term + Complexity Penalty term:

$$BIC = N \log(\frac{RSS}{N}) + k \log N$$

Fit term + Complexity Penalty term
+ Constraint Adherence term:

$$\text{cBIC} = (1-\lambda) N \log \frac{RSS}{N} + \lambda N \log \frac{G(C, \{z_i\})}{4L^2} + k \log N$$

Evaluation Review

- What you need to know:

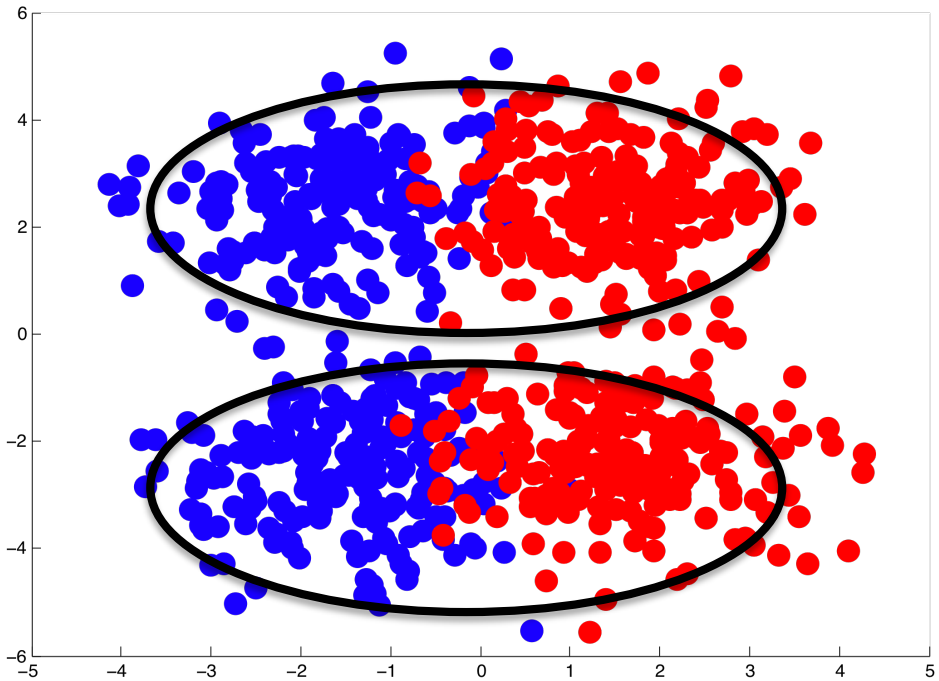
BIC = balances complexity and cluster fit

$G(C, \{z_i\})$ = how well we adhere to constraints

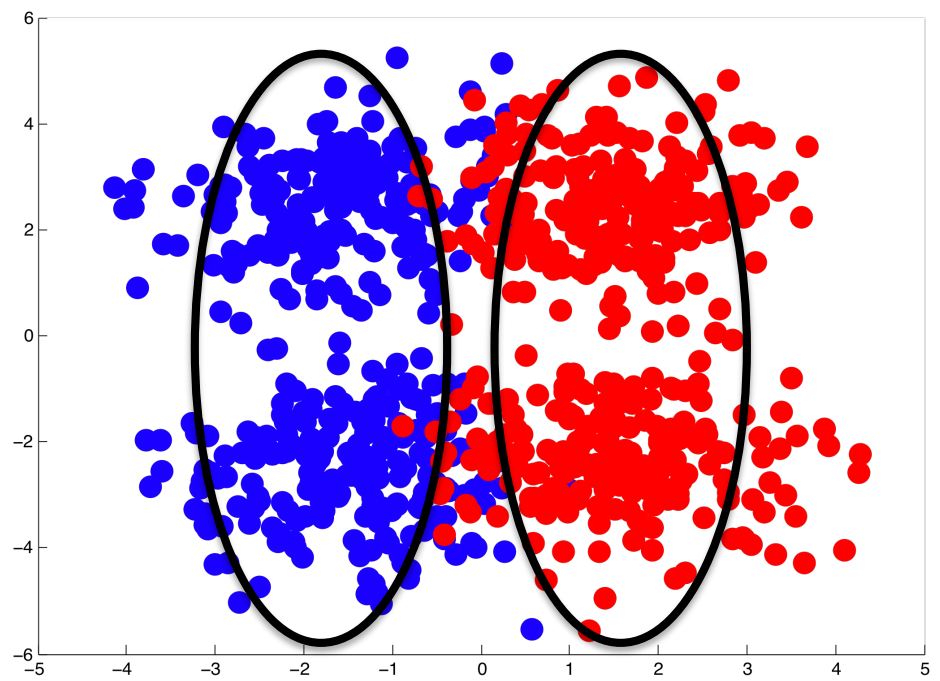
cBIC = a blending of these

EM versus PPC

EM Clustering

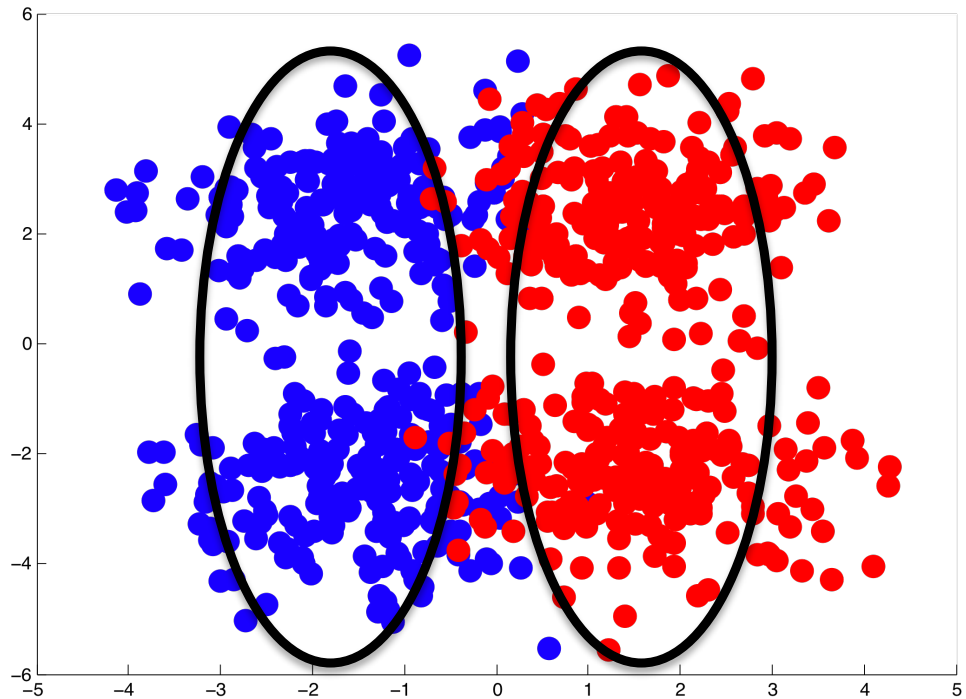


Constraint-Based Clustering



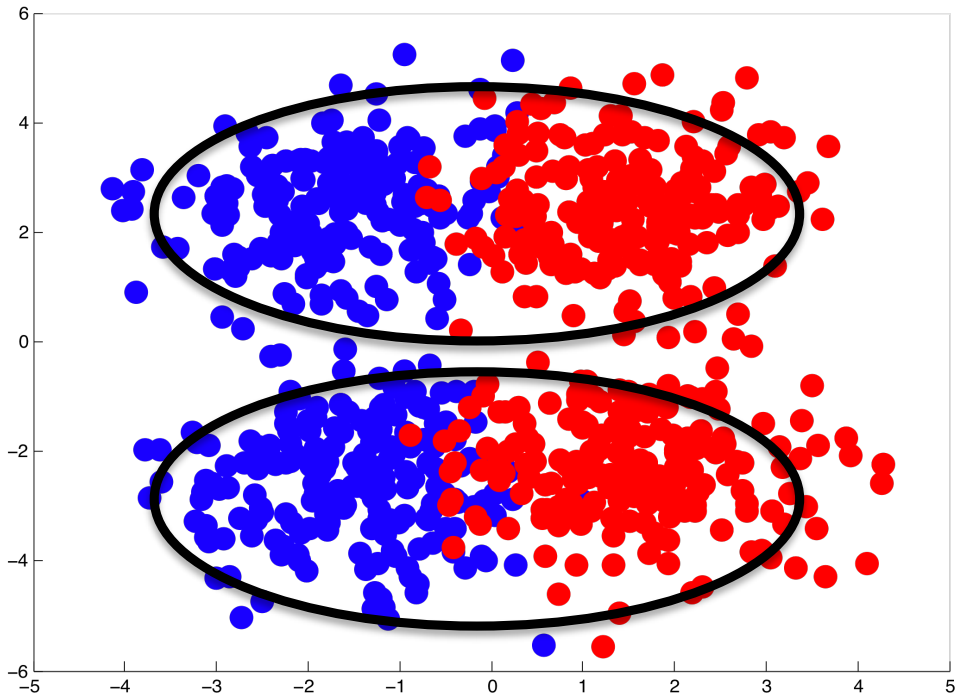
PPC with Strong Constraints

Class	1	2
1	0.95	0.02
2	0.02	0.99

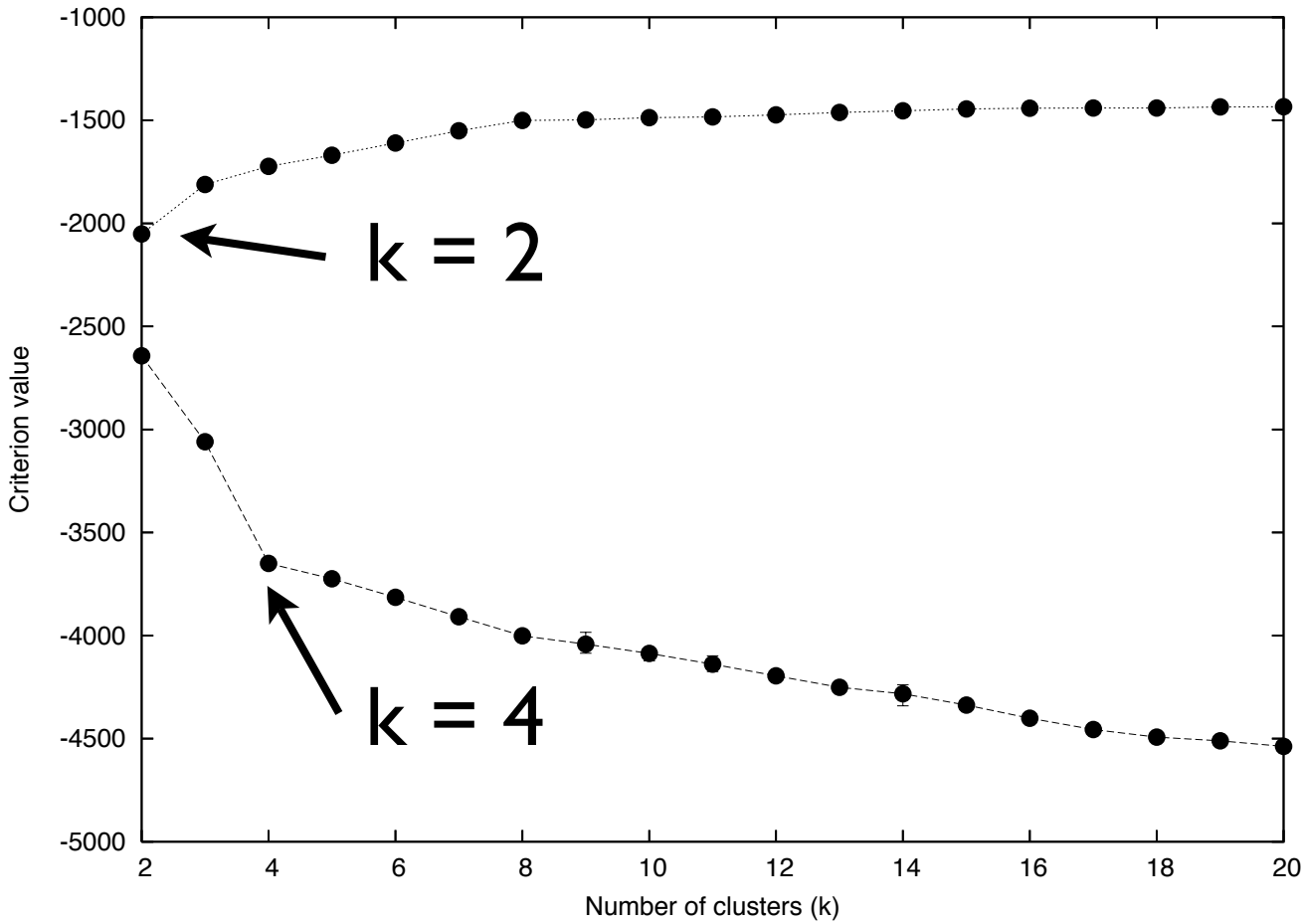


PPC with Weak Constraints

Class	1	2
1	0.20	0.01
2	0.01	0.15



Choosing *k* with cBIC



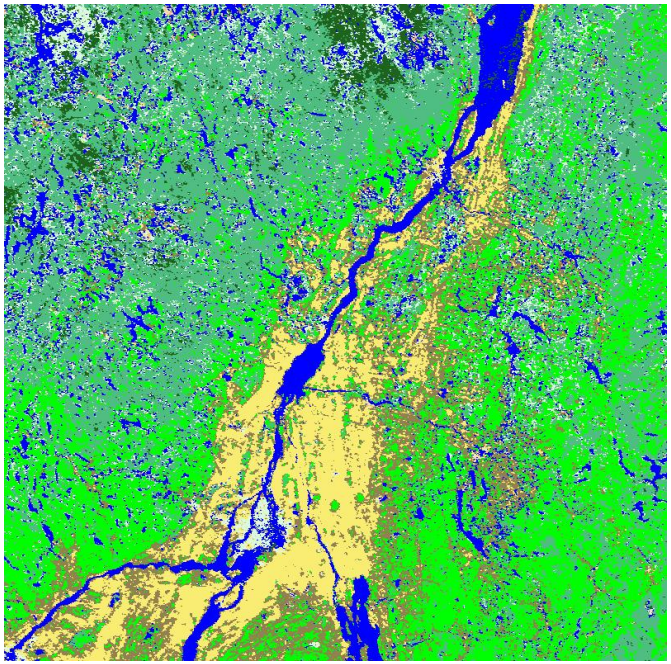
Strong Constraints

Weak constraints

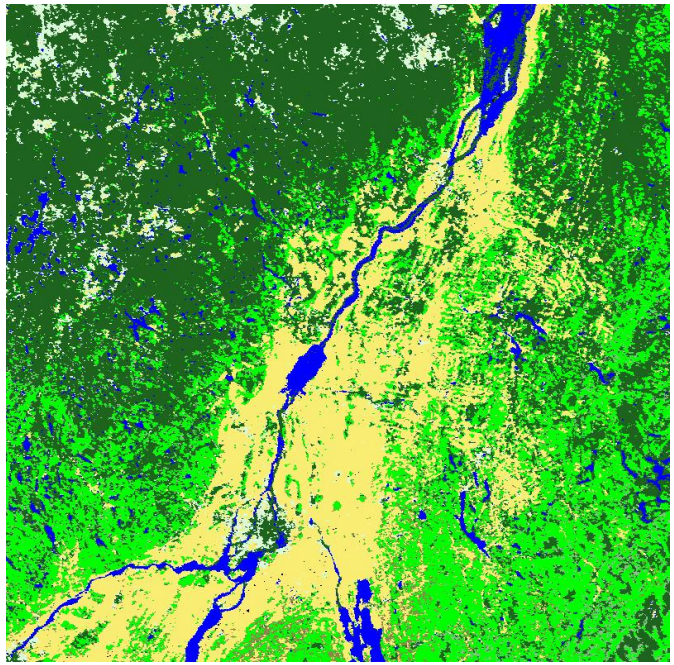
Remote Sensing Data Set

Merging Classes

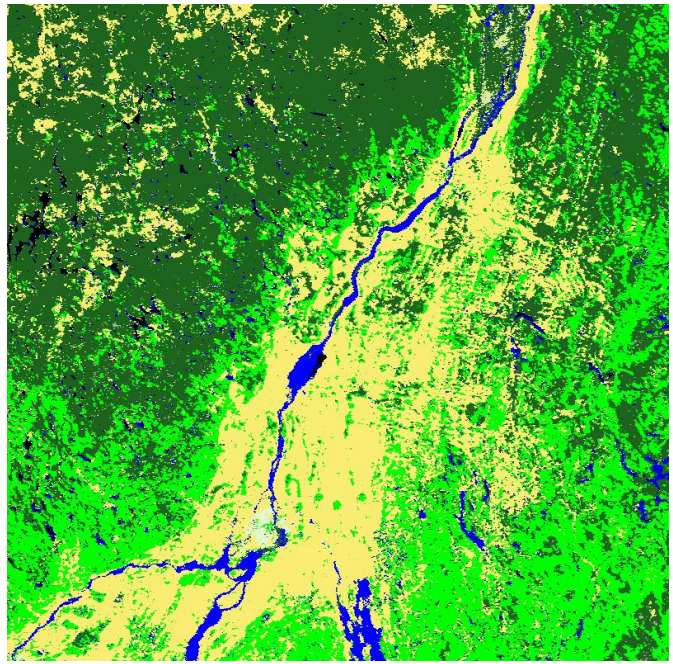
New England & Montreal



IGBP



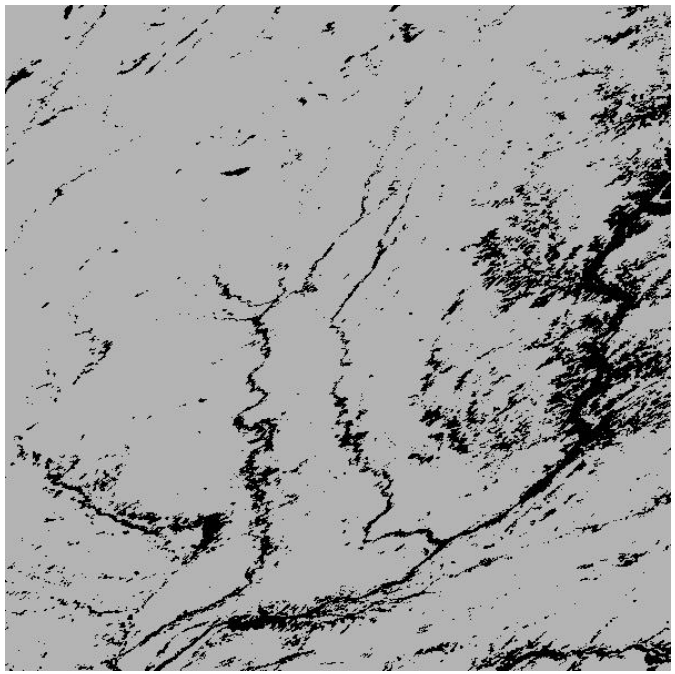
CPPC



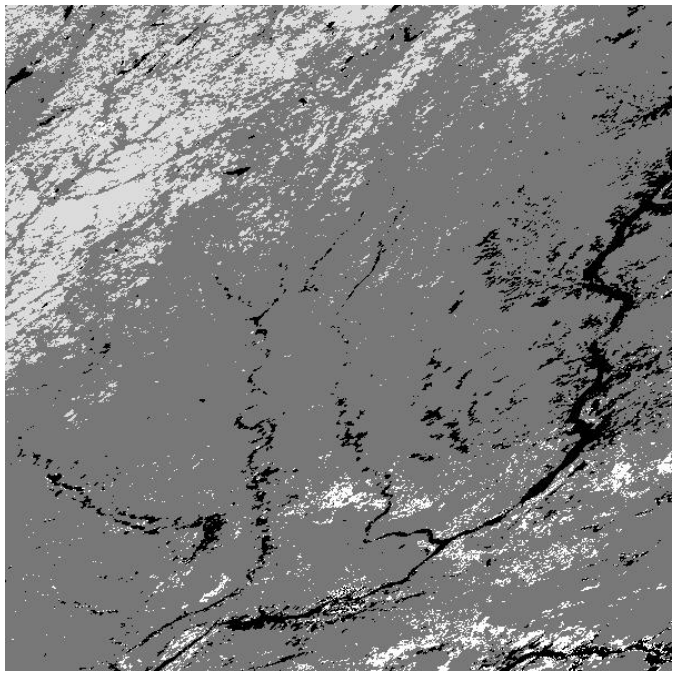
EM

Splitting Classes

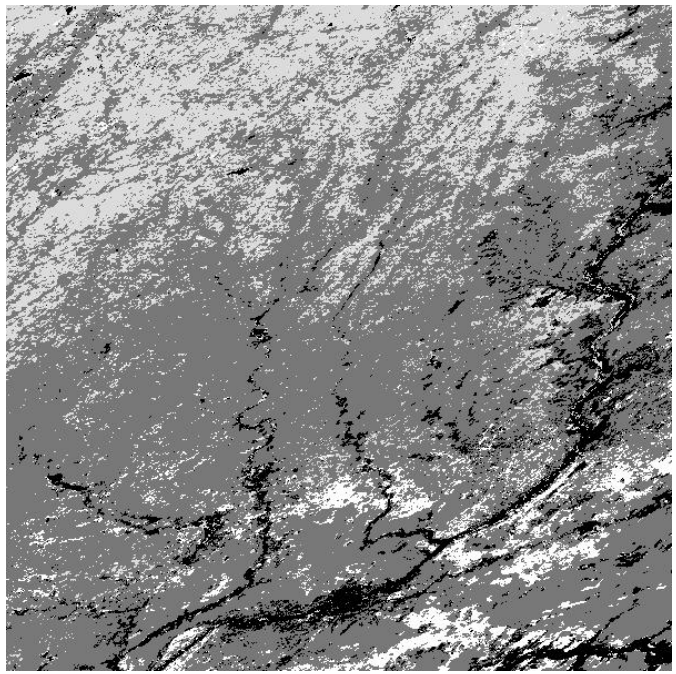
Iowa (agriculture)



IGBP



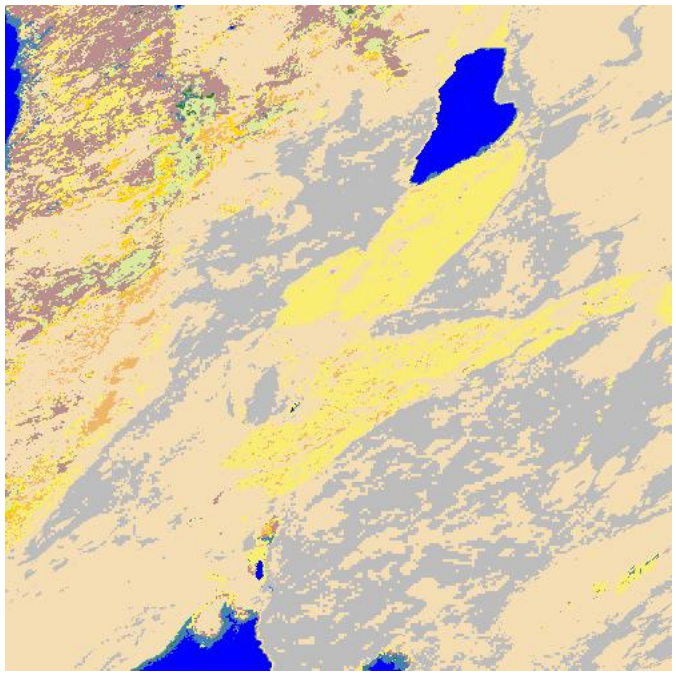
CPPC



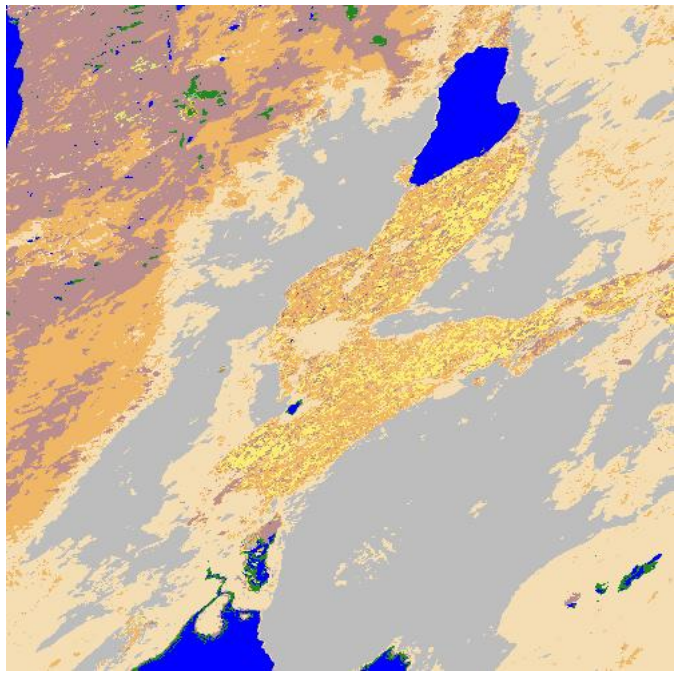
EM

Retaining Structure

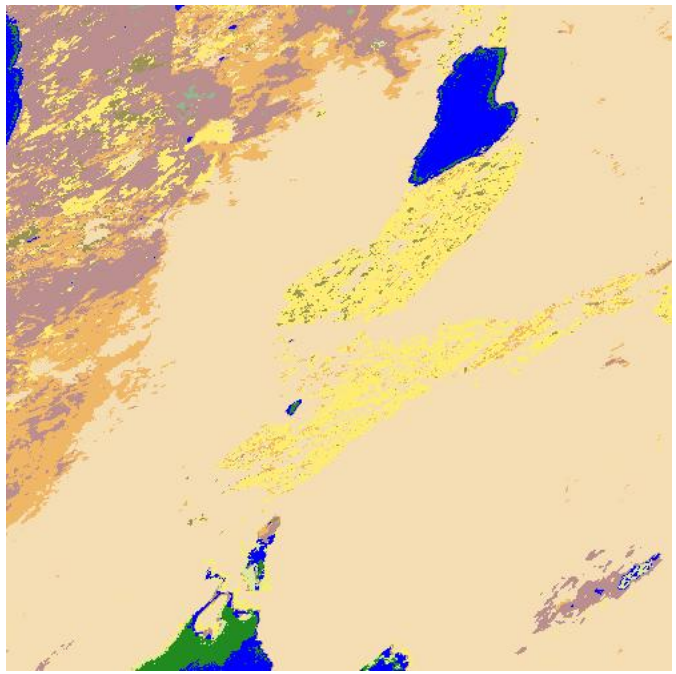
Mexico



IGBP



CPPC



EM

Summary

- Framework for redefining our classes, with expert preferences and reusing our labels
- Compact representation provides CPPC complexity reduction from $O(N^2)$ to $O(NL)$
- cBIC provides a tool to evaluate a constrained clustering, and helps to determine k
- Promising results in Remote Sensing

Thank you

Email: dpreston@cs.stanford.edu