A Novel Pattern Recognition Methodology

Ash Richardson

University of Victoria

March 24, 2010

Work with David Goodenough and Hao Chen (Natural Resources Canada).

Overview

Pattern Recognition

- **Pattern recognition** is a large and diffuse discipline.
- One of the main concerns of Pattern Recognition is **clustering**.

The Clustering Problem

How do we partition a set of objects into like categories?

- Clustering is a routine activity for the mind.
- People don't usually agree.
- Therefore, **clustering must have a subjective element**.

Clustering: Applications

- **Health care: which risk factors are diseases associated with?**
- Internet: what messages are spam?
- **Google "similar images";** computer vision.
- Efficient search: search within nearest cluster.
- Genetic algorithms: which populations represent different solutions?
- Bioinformatics: what are the "species"? (Phylogenetic trees)
- Law enforcement: track criminal activity using spatial and other data.
- **Resource management: where are our resources located?**
- The applications are unlimited.

Objectives

Find a clustering methodology which:

- Is general enough for a wide variety of applications.
- Assumes little about the shape of clusters.
- Makes as few assumptions as possible.
- Is self consistent (more on this later).
- Follows the philosophy of Wishart (next slide).

Clustering Approach

Wishart (1969) and Hartigan (1975) introduced and expanded the idea that clusters are peaks (modes) of a probability density.

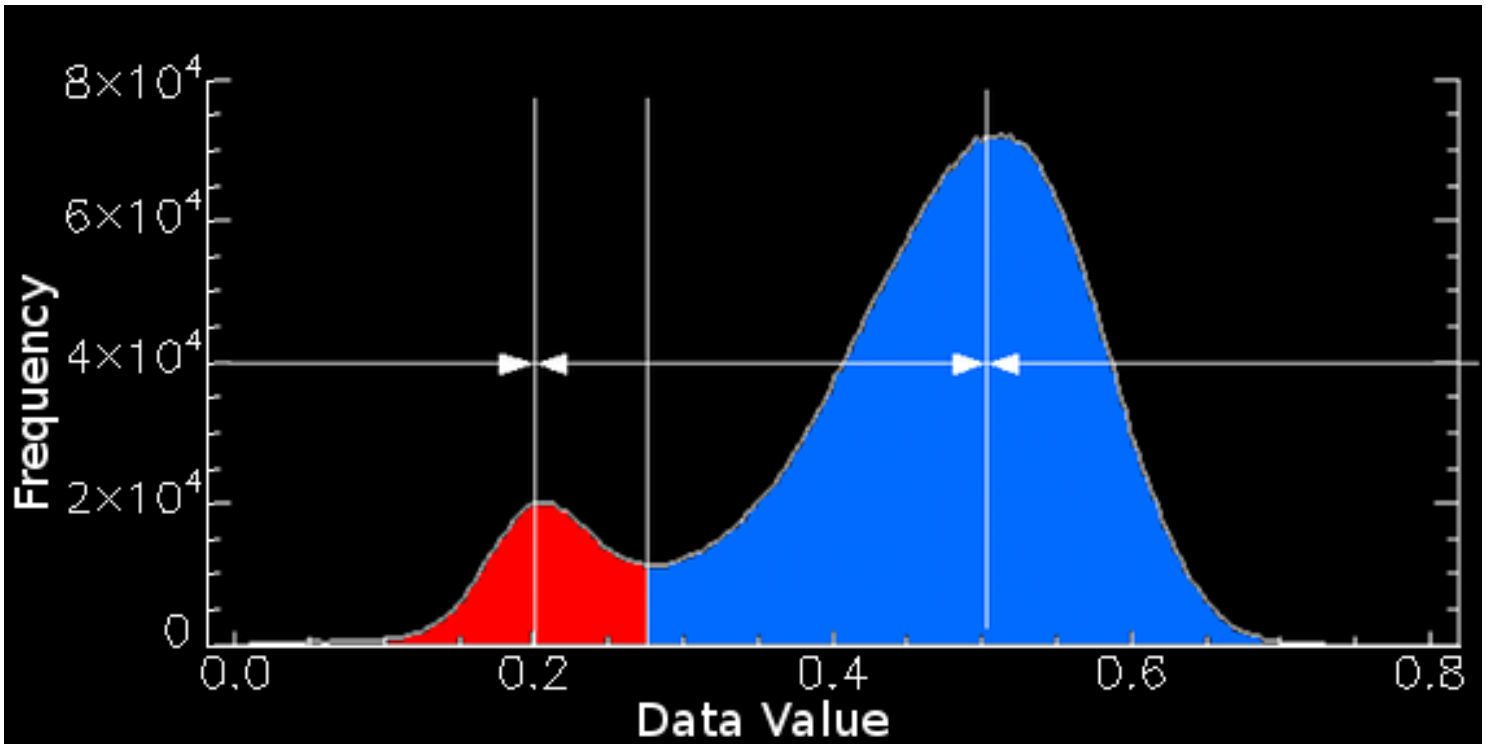
Problem Statement

Estimate clusters by finding the peaks.

- Assign a cluster to each peak and the associated domain of attraction.
- How?
- Climb up!

Clustering Approach

- Assign a cluster to each peak and the associated domain of attraction.
- How?
- Climb up!



Overview

- First we will examine two important classical examples of clustering approaches.
- We will highlight their drawbacks to make recommendations of features we would like to avoid.
- The new approach will be based around these recommendations.
- First, it is worth mentioning that Clustering can be thought of as optimization (next slide).

Clustering as Optimization

Often clustering is thought of as an optimization problem.
Partition a set X into a collection $\{X_i\}$ such that:

$$\min_{\{X_i\}} \left\{ \sum_i \sum_{x_i, x_j \in X_i} d(x_i, x_j) \right\}$$

(minimize the total within-cluster distance).

- Enumerating the space of partitions is not practical.
- Intractable optimization problems give way to heuristics.

Example 1: Agglomerative Clustering

Assuming we have a finite set X and a set distance $d(X, Y)$, and we fix a number of final clusters N, we do the following:

- Let $S = \{\{x\} \ni x \in X\}$ (let each singleton be a cluster).
- Until $|S| = N$, union together the nearest $X, Y \in S$.

where S is the number of elements in the partition.

Example 1: Agglomerative Clustering

How do we measure the distance between clusters X and Y?

$$d(X, Y) = \max\{d(x, y) : x \in X, y \in Y\} \tag{1}$$

$$d(X, Y) = \min\{d(x, y) : x \in X, y \in Y\} \tag{2}$$

$$d(X, Y) = \frac{1}{|X||Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y) \tag{3}$$

These are complete linkage(1), single linkage(2), and average linkage(3) distances.

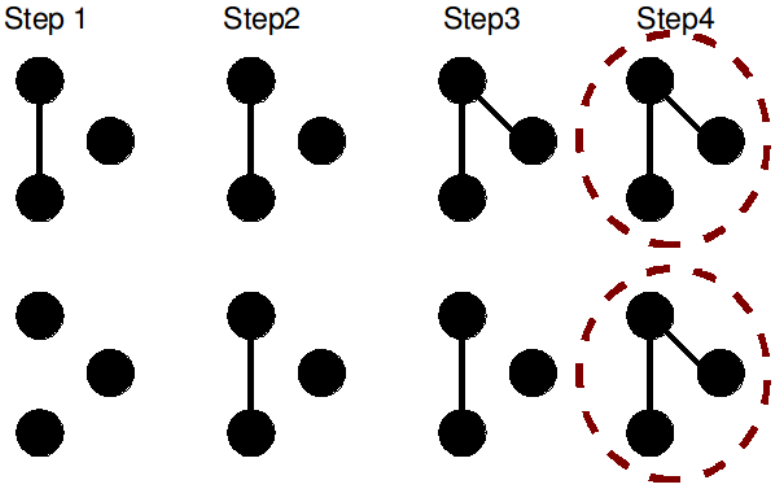
These functions between sets are ways of thinking about density.

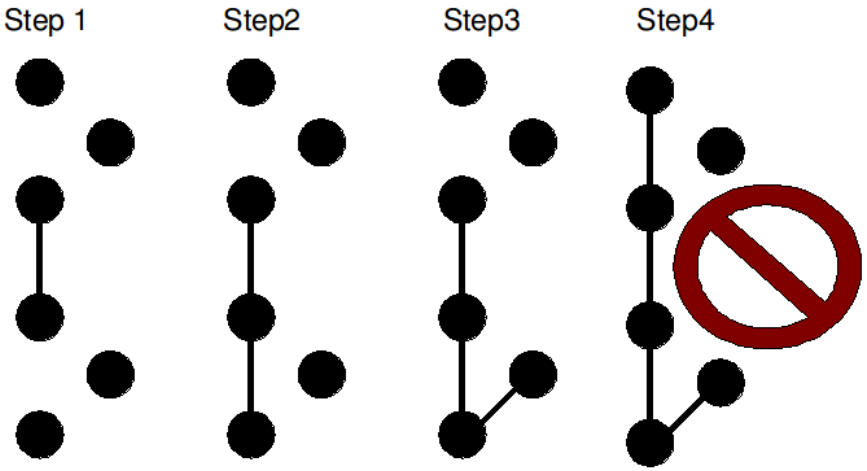
These concepts of density may not be consistent with the algorithm.

Example 1: Agglomerative Clustering: Drawbacks

- Regardless of how clever the linkage (set distance function) is, **the approach is greedy** (snowball effect).
- Single linkage has nice analytical results (there is a deterministic solution, given by MST) but gives **chaining effects**.
- For other linkages, we can get multiple answers.
- The single linkage uses information that is **too local**.
- The complete linkage uses information that is **too global**.
- It is hard to clarify whether more complex linkages are compatible with the algorithm.
- We show an example on the next slide.

Example 1: Agglomerative Clustering: Average linkage





"Snowball effect".

The leftmost points are equally spaced.

Example 2: K-means Clustering

Let $|X| = N$, $X \subset \mathbb{R}^n$ and $x \in X$. Assume a starting set of means μ_j (the number of clusters is fixed); try to minimize:

$$F = \sum_{i=i}^{N} \sum_{j=1}^{M} \delta_{ij} ||x_i - \mu_j||^p \tag{4}$$

where $\delta_{ij} = 1$ if $x_i \in X_j$ and 0 otherwise.
Assuming $p = 2$ and a partition is given,

F is quadratic in μ_j so we can solve for $\mu_j = \frac{\sum_i \delta_{ij} x_i}{\sum_i \delta_{ij}}$.

Algorithm

(1) For each point x, find the nearest mean μ_j; assign x to X_j.
(2) For each cluster X_j, recompute the mean μ_j.
(3) Go back to (1) or stop.
This heuristic yields μ_j and X_j that are consistent (if it converges).

Example 2: K-means Clustering: Drawbacks

- Only applicable if a mean is defined.
- The optimization problem has the concept of "gaussian" built right in (as it refers to the mean and deviation from the mean).
- Clusters realized from gaussian random variables look "round".
- K-means will work for "round" shape clusters, but may **fail for clusters with more complicated structure**.
- **Initial conditions** have to be prescribed (the μ_j).
- Convergence may be poor.

Design Objectives

- Find peaks in the density.
- Find clusters of arbitrary shape.
- Avoid prescribing an initial partition.
- Avoid assumptions about the statistical distribution of the data.
- Avoid using any training data.
- Avoid using a greedy approach.
- Avoid explicitly assuming the number of classes.

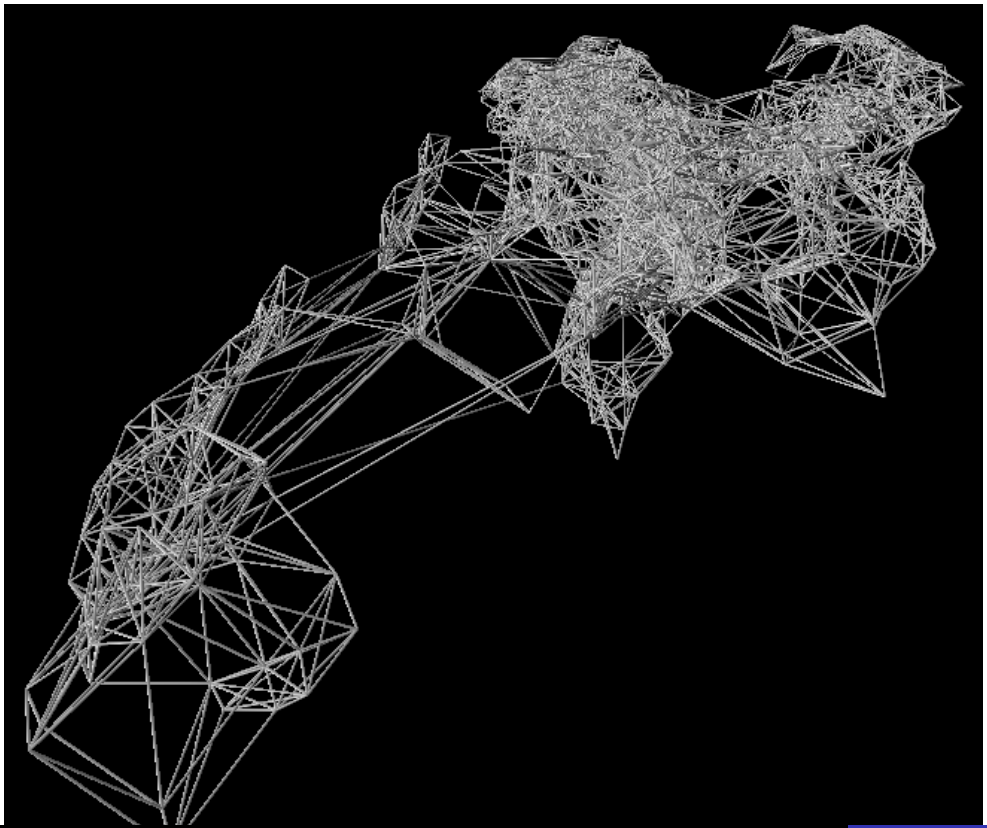
Clustering is a natural for the mind.
The mind has a network-like structure.
Idea: use a network-like structure.

New Approach

- Part 1: Density estimation using the **k-nearest neighbors graph.**
- Part 2: Hill climbing the estimated density on the **k-nearest neighbors graph**.

Part 1: The K-Nearest Neighbors Graph

Suppose we have a set X of points x, and a distance function between points $d(x, y)$ (possibly not a metric).
The **K-Nearest Neighbors Graph** relates each point with the K-closest points to it (below a sample graph for K=5).



Density Estimation On The K-Nearest Neighbors Graph

For a point x_i on the **K-Nearest Neighbors Graph (K-NNG)**, denote by $\{x_{ij}\}$ the **K-Nearest Neighbors** of x_i ($j \in \{1, \ldots, K\}$). Where C is a constant, can define any number of rudimentary density estimates on the K-NNG, for example:

- "Max density" $\rho(x_i) = \frac{C}{\max_j\{d(x_i, x_{ij})\}}$
- "Median density" $\rho(x_i) = \frac{C}{\mathit{median}_j\{d(x_i, x_{ij})\}}$
- "Mean density" $\rho(x_i) = \frac{C}{\mathit{mean}_j\{d(x_i, x_{ij})\}}$

This is analogous to the linkages of the **Agglomerative** approach. The advantage here: the density estimate and the clustering algorithm (next slide) cannot contradict each other **(self-consistency)**.

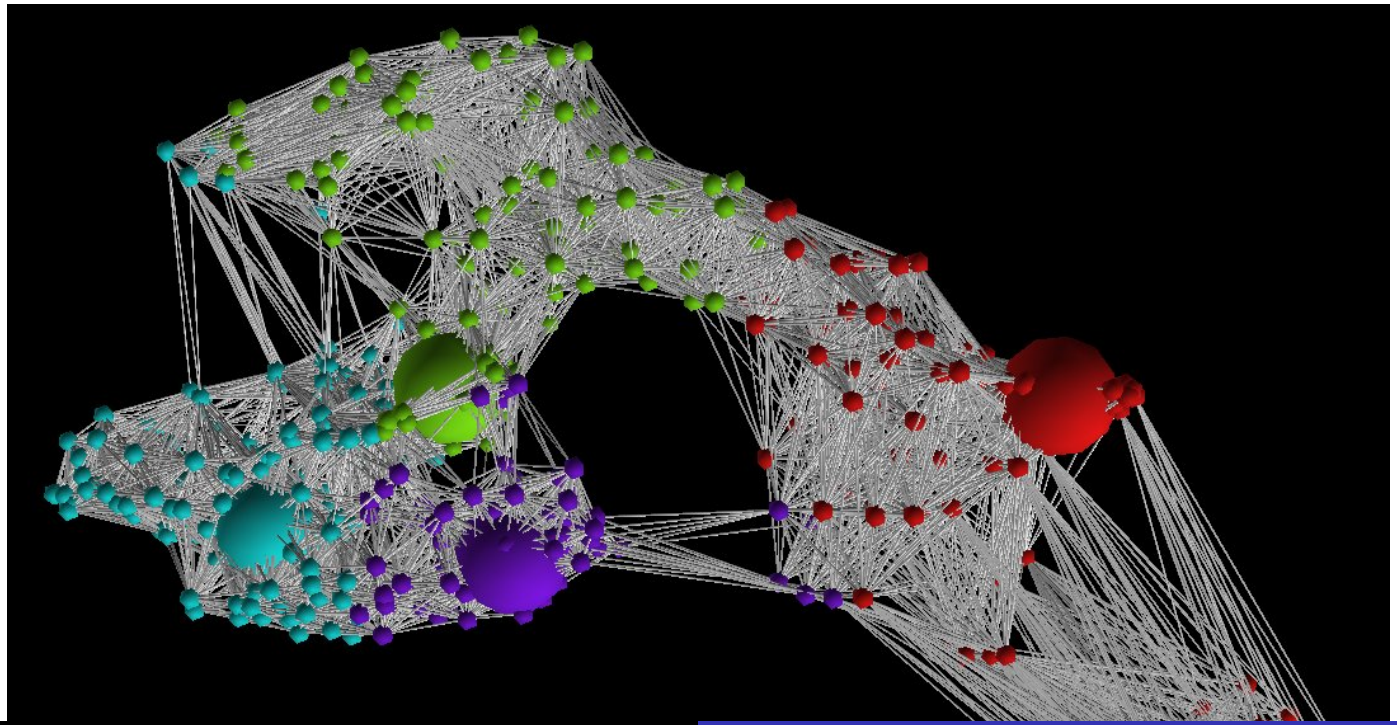
Part 2: Hill Climbing On The K-NNG

Now that we have defined a density estimate at each point x_i, we can perform the clustering by hill climbing on the K-NNG.
Definition: x_i is a **peak** if no neighbors $\{x_{ij}\}$ have higher density.
Notation: denote the highest density neighbor of x_i by highest(x_i).
Recursive rule to assign a peak to each point:

- MyPeak$(x_i) = x_i$ if x_i is a peak.
- MyPeak(x_i) = MyPeak(highest (x_i)) otherwise.

Hill Climbing On The K-NNG (Sample Result)

Here is an example of 1000 test points in 3-d space. This could be 3-d (R,G,B) color values for a small slice of an image of the earth. The clusters might represent different categories like: water, forest, urban areas, and bare surface. Small balls are 3-d points and large balls are peaks.



Computational Issues

A drawback of this approach is computational complexity.

Building the K-NNG requires comparisons between all points.

(In the K-means approach, points are only compared with the μ_j).

Space partitioning strategies should be used to decrease the number of comparisons required.

A heap data structure should be used so the nearest neighbors do not have to be sorted after they are collected.

What is new about this approach?

Part 1: Density estimation using the **k-nearest neighbors graph.**
Part 2: Hill climbing the estimated density on the **k-nearest neighbors graph**.

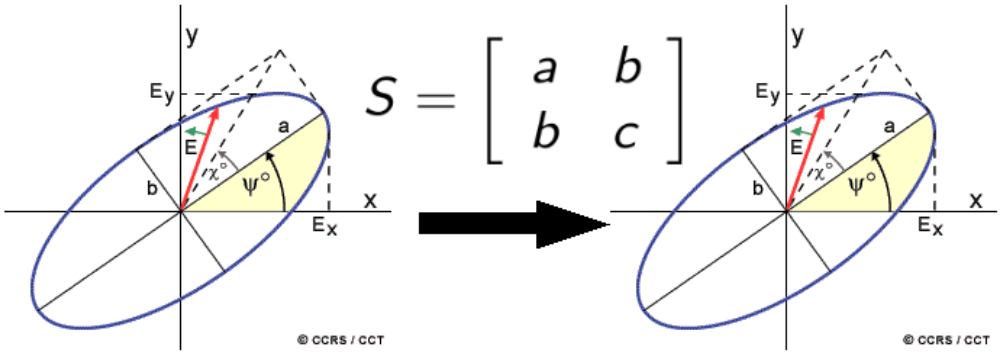
- Nearest neighbor density estimates are used all the time.
- **K-Nearest Neighbors Clustering** usually refers to an old approach that is really only lazy interpolation. Objects are classified by a majority vote of neighboring training points.

What is new?
- Haven't seen K-NN density estimation and hill climbing combined together.
- An attractive feature is the consistency between the algorithm and the density estimate.

Overview

Polarimetric Radar

- Polarimetric Radar is a remote sensing instrument carried by planes or satellites.
- This is an imaging experiment that measures a linear operator (2x2 symmetric complex matrix).
- The operator represents the transform between incident and reflected polarization states.
- The operator is at least 5-dimensional.



Visualizing The Scattering Operator

The Scattering Operator is ≥ 5 dimensional.
We must choose some parameters to visualize it.

Visualizing The Scattering Operator

The matrix $S = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$ is complex.

Form a 3 dimensional complex vector $k = \begin{bmatrix} a+c \\ a-c \\ 2b \end{bmatrix}$.

Form the Hermitian outer product:

$$T = k \cdot k^{*T} \tag{5}$$

- T is a 3x3 conjugate symmetric (Hermitian) matrix.
- Thus T has **real eigenvalues**.
- The eigenvalues of T represent the power of the scattering return.
- As in Q.M., Hermitian product is used to derive real valued quantities ("observables").
- T is rank 1 but we will fix this shortly...

Visualizing The Scattering Operator

The Scattering Operator is ≥ 5 dimensional. For example, we can use the diagonal elements of T as (R,G,B) values:

Diagonalizing T to get meaningful parameters

Assume T is a sample from a statistical process: take averages of T, representing expected values (the averaged T has rank 3).

$$< T >= \sum_i T_i / N \tag{6}$$

The matrix T is **diagonalizable**: λ_1, λ_2, λ_3, and eigenvectors:

$$\vec{e}_i = \exp(i\phi) \begin{bmatrix} \cos(\alpha_i) \\ \sin(\alpha_i)\cos(\beta_i)\exp(\delta_i) \\ \sin(\alpha_i)\sin(\beta_i)\exp(\gamma_i) \end{bmatrix}$$

$\alpha = \sum_i p_i \alpha_i$ (alpha: representing scattering "type")

$p_i = \frac{\lambda_i}{\sum_j \lambda_j}$ (probability of each of 3 outcomes)

$H = -\sum_i p_i \log p_i$ (entropy: how much disorder?)

$A = \frac{\lambda_2 - \lambda_3}{\lambda_2 + \lambda_3}$ (anisotropy: importance of secondary components)

Diagonalizing T to get meaningful parameters

- $\alpha = \sum_i p_i \alpha_i$ (alpha: representing scattering "type")
- $H = -\sum_i p_i \log p_i$ (entropy: how much disorder?)
- $A = \frac{\lambda_2 - \lambda_3}{\lambda_2 + \lambda_3}$ (anisotropy: importance of secondary components)

The parameter α represents a smooth transition between three physical phenomena: surface or odd-bounce scattering ($\alpha = 0$) , dipole scattering ($\alpha = \pi/4$) and dihedral or even-bounce scattering ($\alpha = \pi/2$). **Real examples:**

- Water bodies and roads have low anisotropy.
- Ocean has higher entropy than roads (more random).
- In terms of α, urban areas are much different from natural ones.

Clustering the Scattering Operator

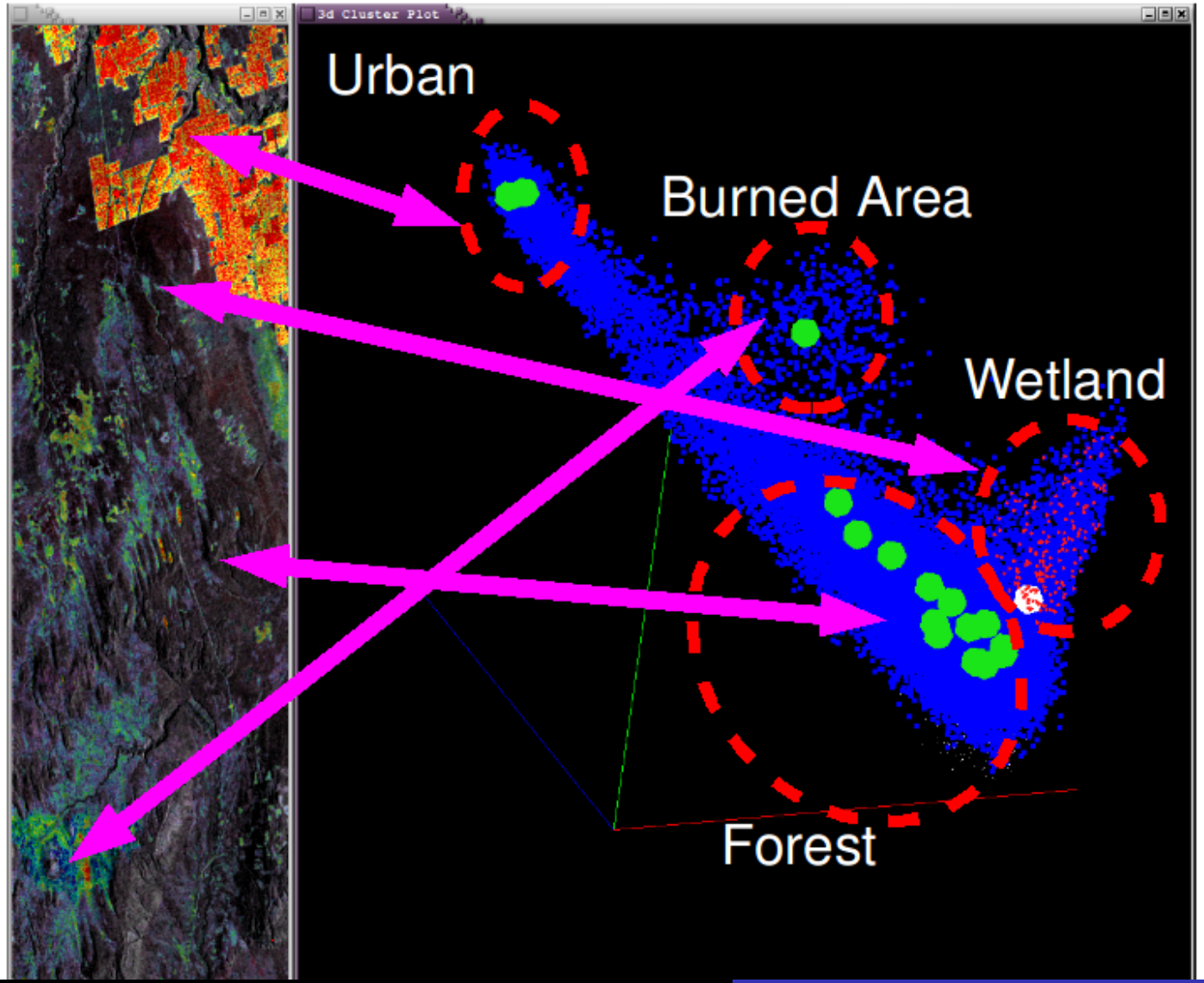
Clustering is typically performed using a K-means approach in the T matrix space.
The inital means μ_j are determined by partitioning a lower dimensional space.
The (H, A, α) space is commonly used.

Why is a more general clustering approach needed?

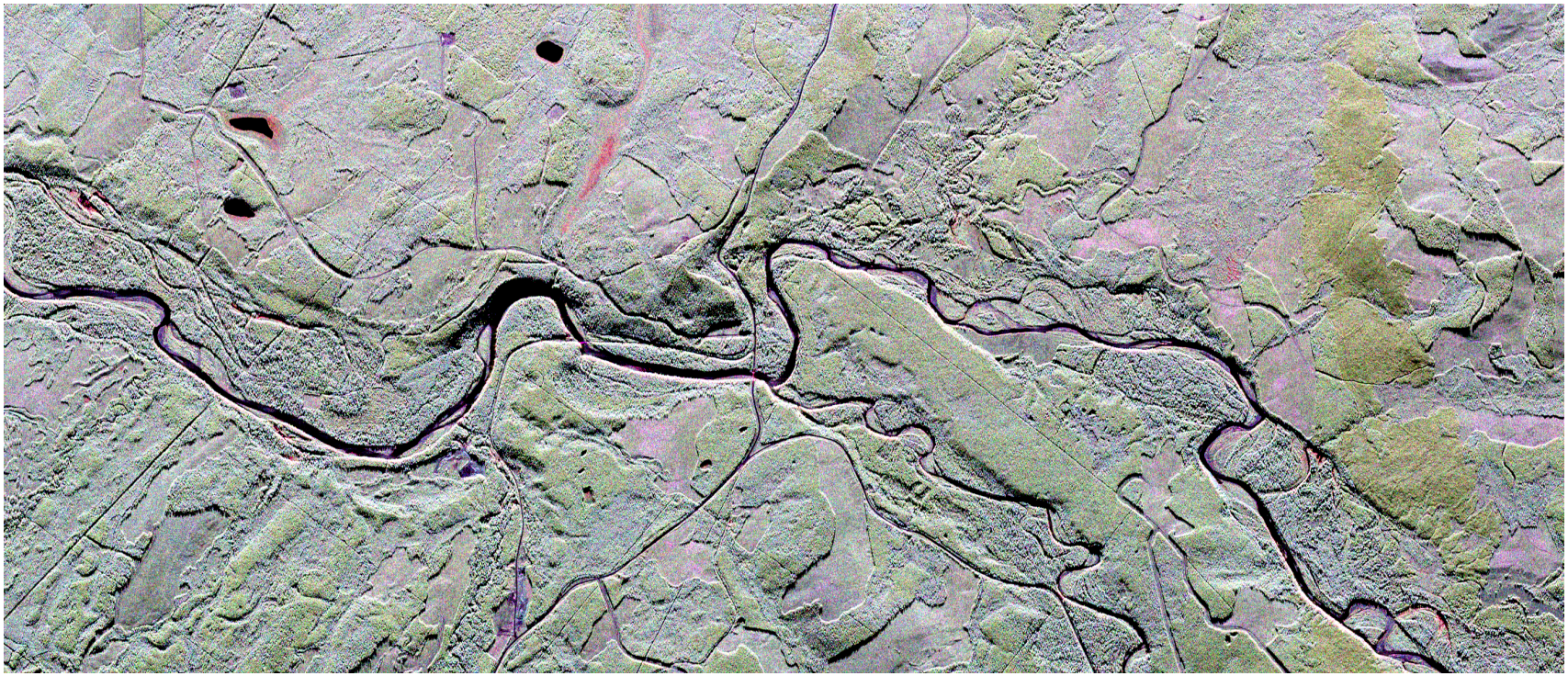
- Clusters in the (H, A, α) space are not round.
- Unclear how many clusters there are.
- Unclear what is lost by moving from 3-d to 5-d.
- Unclear how dimensional the matrix data is.
- Unclear how many kinds of physical objects can be differentiated.

Ongoing work to address these issues.
Abstract submitted: A.Richardson, D.G. Goodenough, H. Chen, G. Hobart, B. Moa, W. Myrvold, **Unsupervised Nonparametric Classification of Polarimetric SAR Data Using The K-nearest Neighbor Graph**, IEEE International Geoscience and Remote Sensing Symposium, 2010.

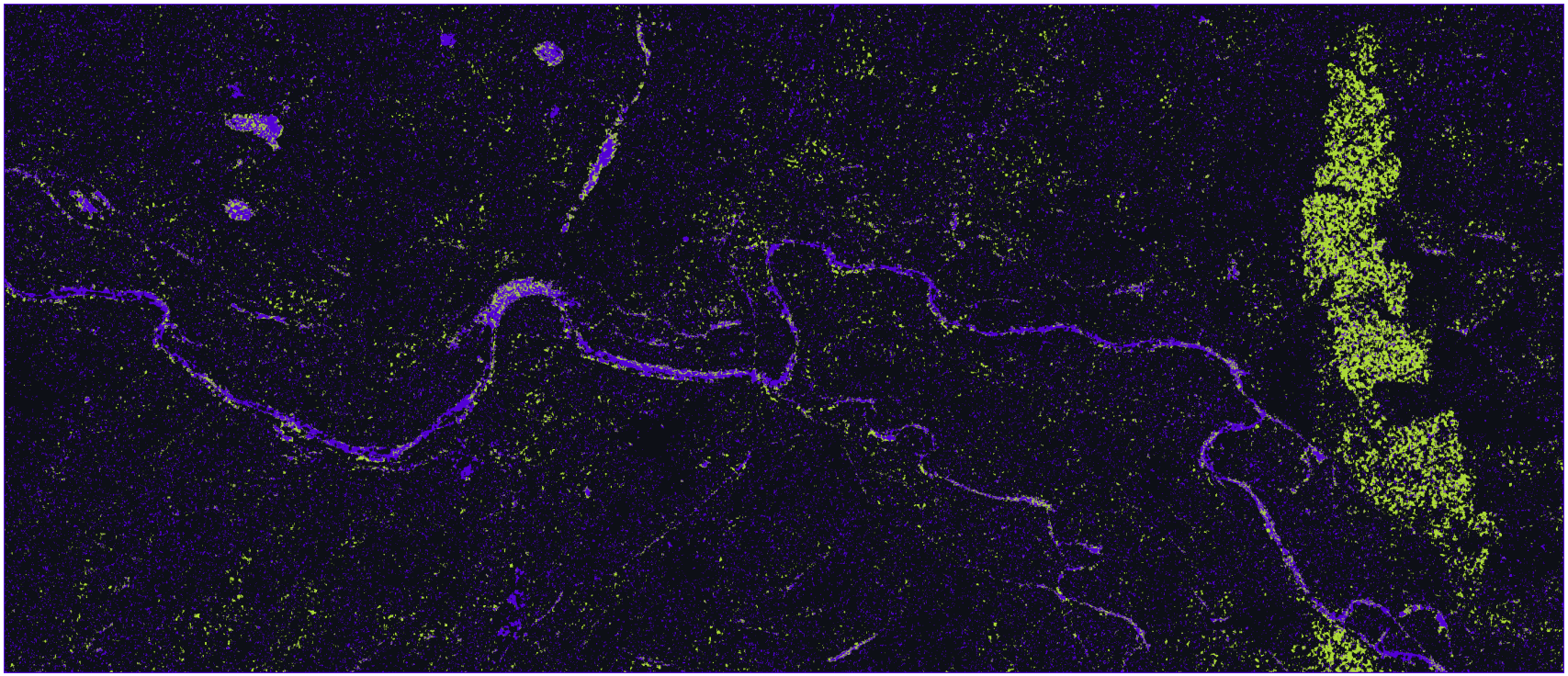
Key River Fire (2002)



Hinton Fire (50+ years old)



Hinton Fire (50+ years old) Burned Area Discrimination



Earth Observation: Towards Transparency for World Affairs

- Polarization radar can separate a variety of land use types.
- All weather, 24 hour technology can reach high altitude areas other techniques can't.
- (This applies to more than half of Canada's forests).
- Can be used to map deforestation, ice distribution, urban expanse, etc.
- Hopefully, open access to remote sensing data will eventually make government and corporate activities more transparent.

Overview

Conclusion

- A new clustering approach was designed, implemented, and tested on earth image data.
- The approach is data driven and has one free parameter (K).
- The geometrical consistency between algorithm and density estimate is a highlight of the approach.
- It separated a land cover type that an industry standard approach could not.

Further work:

- Improve density estimates used in the approach.
- Use Bhattacharyya type coefficients to compare with other clusterings.
- Use Bhattacharyya type coefficients to compare results for different values of K.