A Consistent Density-Based Clustering Algorithm and its Application to Microstructure Image Segmentation

Marilyn Vazquez Landrove

Institute for Computational and Experimental Research in Mathematics
Providence, RI

Computational Imaging

Acknowledgements

National Institute of Standards and Technology

Sponsor: Steve Langer

Mentor: Gunay Dogan

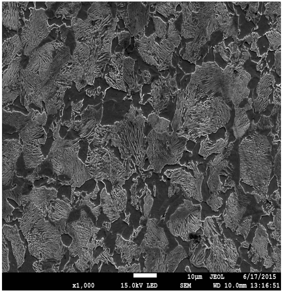
Data: Sheng Yen Li

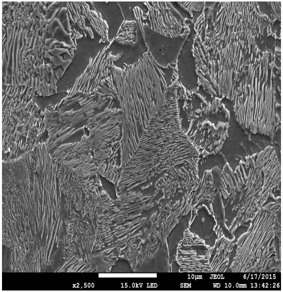
Other Collaborator: Andrew Reid

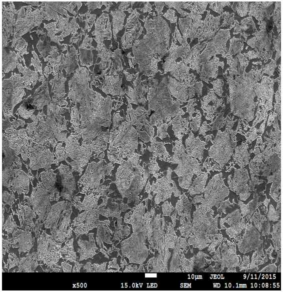
Research supported by Industrial Immersion Program (IIP) at GMU

Material Images: Steel

Pearlite: A mixture of cementite and ferrite







Simulations that let them measure the strength of their material.

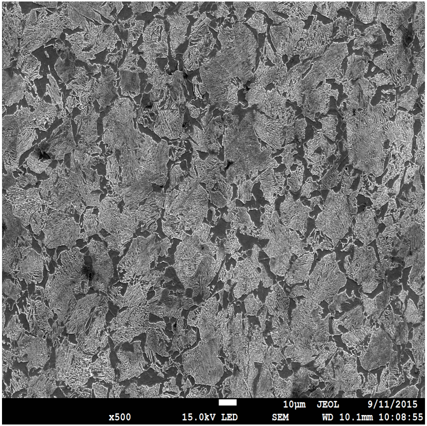
Material Images: Steel

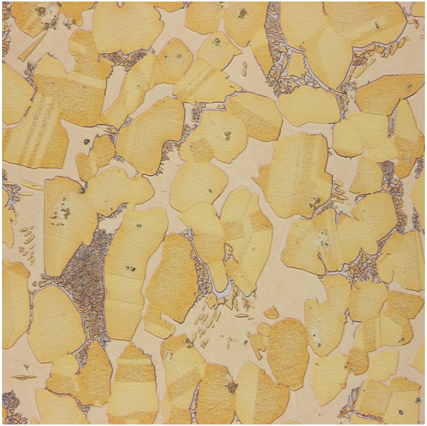
Information needed for simulations:

- Fraction of pearlite (stripe region) in image
- Fraction of ferrite (white stripes) in pearlite
- Space between ferrite stripes
- etc.

First step to achieve these goals: Image Segmentation!

Challenges





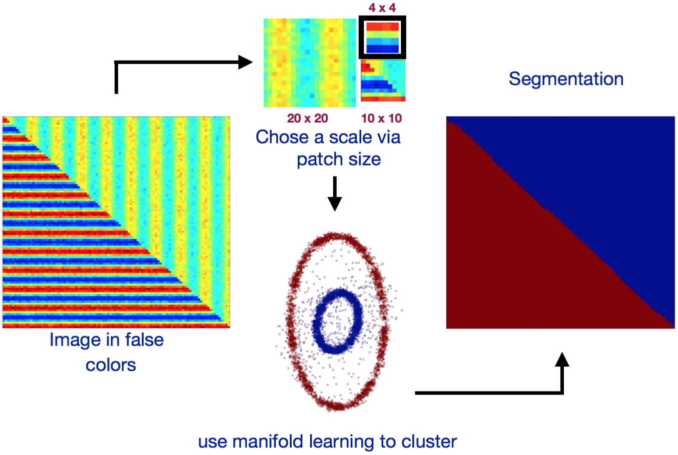
(1) Texture is difficult to represent and (2) want the least amount of human involvement

Create Features to Capture Patterns

- **Edge detection**: Use contrast to detect edges and let enclosed areas be the regions of interest. Ex: Canny Edge detector (1987)
- **Match filters**: Build filter bank that represents the desired pattern and convolve with image to measure "response." Ex: Frangi Filter (1998)
- **Region based**: Grow/shrink input regions according to partial differential equation. Ex: Level set method by Osher and Sethian (1988)

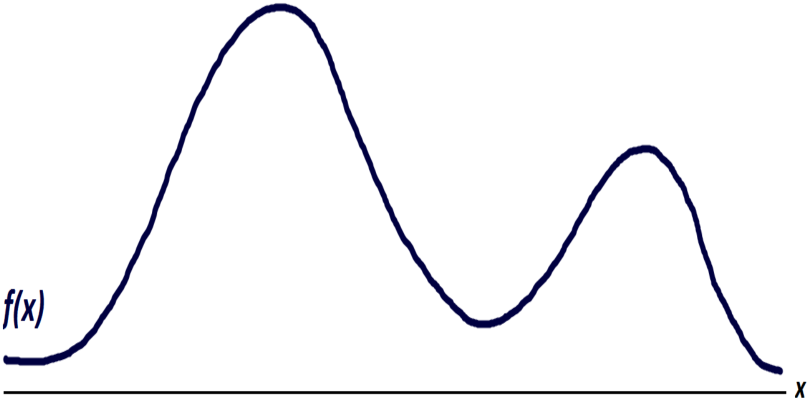
Image Segmentation as a Clustering Problem

Our idea: Introduce manifold learning to achieve non-parametric image segmentation



Density Based Clustering

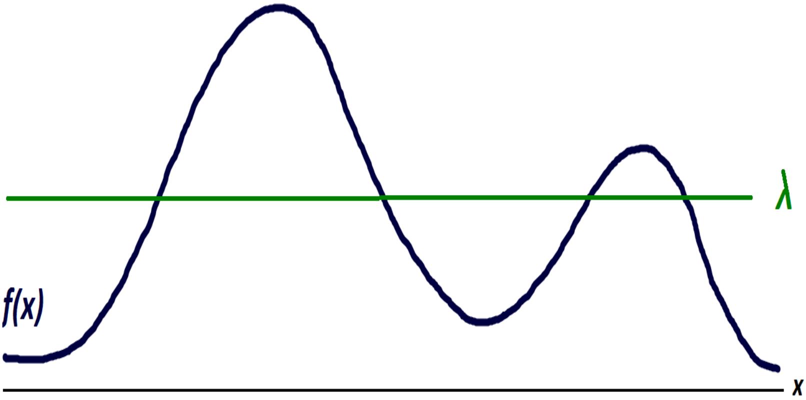
How would you cluster points with the following density f?



Challenges: We do not have access to the real density f, but an estimation $\hat{f}$.

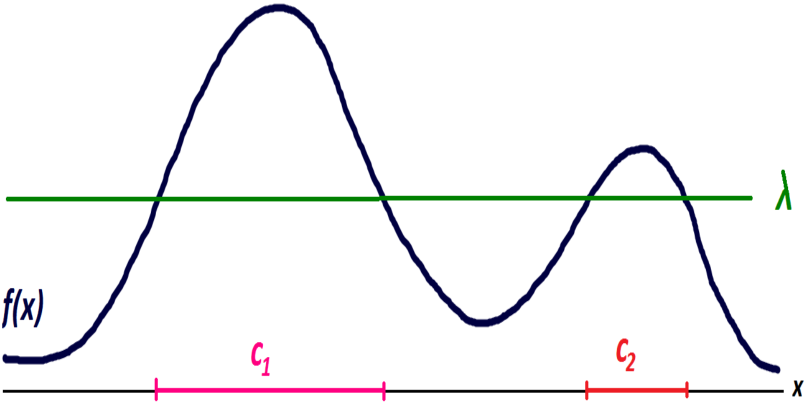
Cut: Super-level Set

Suppose we knew f, then a solution can be found looking at super-level sets, i.e. $\{x \in X_N : f(x) \geq \lambda\}$



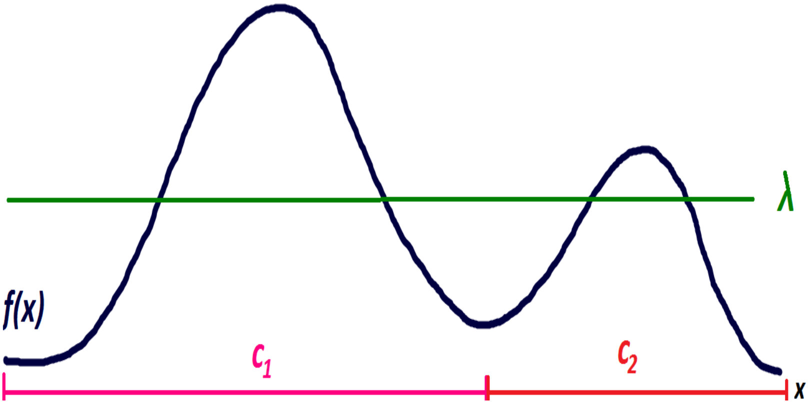
Cluster: Connected Components

Connected components of $\{x \in X_N : f(x) \geq \lambda\}$



Classify

How do we label the rest of the points? Using a classifier!



Section 2

Consistency of clustering

Consistency

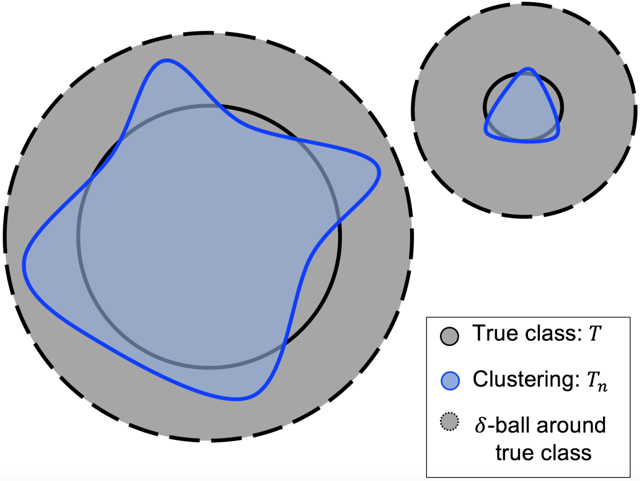
Definition

Let T be the true clustering of $\mathcal{M}$ and $\{T_n\}$ be a sequence of random clusterings of $\mathcal{M}$. The sequence $\{T_n\}$ is **consistent** if for every sufficiently small $\delta, \gamma > 0$, the following conditions hold for sufficiently large n:

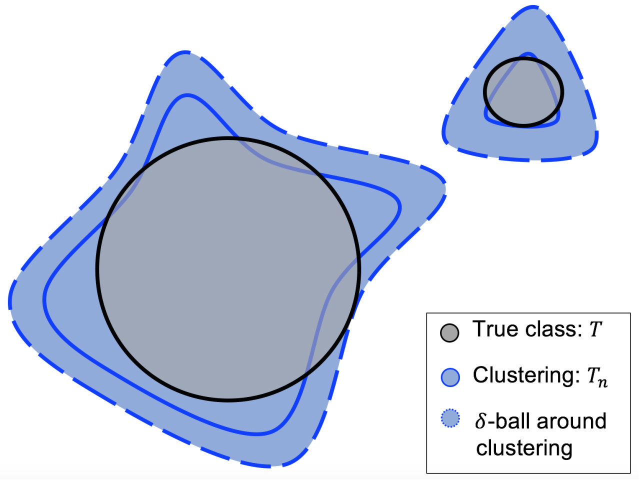
1. (Separation) $T_n \subset B(T, \delta)$
2. (Coverage) $T \subset B(T_n, \delta)$
3. (Cohesiveness) The inclusions in (1) and (2) define a one-to-one correspondence between the connected components of T and T_n

with probability $1 - \gamma$.

Separation: $T_n \subset B(T, \delta)$



Coverage: $T \subset B(T_n, \delta)$



Consistency of Density Based Clustering

- Let X_n be a data cloud of n points sampled from $\mathcal{M} \subseteq \mathbb{R}^d$ using probability density function q
- Let q_n be a consistent density estimator of q.
- For density based clustering, clearly $T = T_\lambda(q)$ i.e. the superlevel set of density q at λ.

Consistency of Density Based Clustering: Theorem

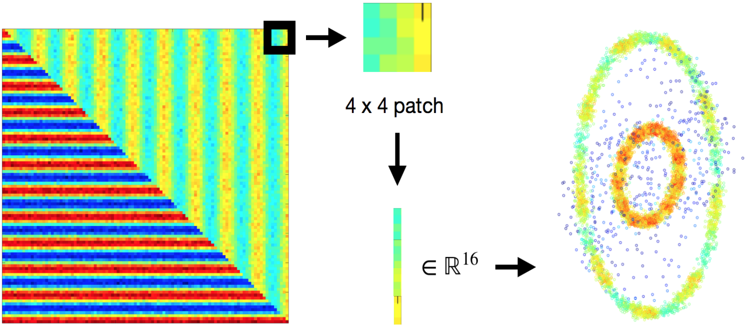
Theorem

Let $q : \mathbb{R}^d \to \mathbb{R}$ *be a* C^d *probability density function such that* $q(x) \to 0$ *as* $||x|| \to \infty$. *Also, assume that* q_n *is a consistent density estimator of* q *with variance* σ_n^2. *Suppose that* q_n *has a Lipschitz constant* L_n *such that* $L_n^d \sigma_n^2 \to 0$ *as* $n \to \infty$. *Denote the superlevel set of a function as* $T_\lambda(f) = \{x \in X_n : f(x) > \lambda\}$. *Then for almost every* $\lambda > 0$ *and for small enough* $\delta > 0$, $B(T_\lambda(q_n), \delta)$ *form a consistent clustering of* $T_\lambda(q)$.

Section 3

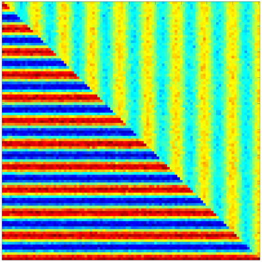
Image Segmentation

Data Driven Feature Space

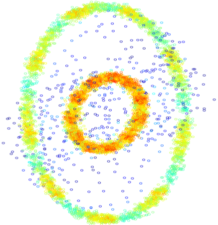


Each 4×4 patch is a point in a 16 dimensional data cloud, whose PCA representation is on the left.

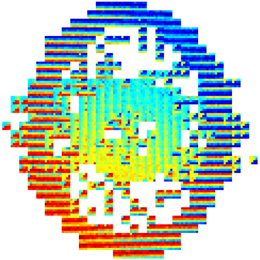
Patch Space



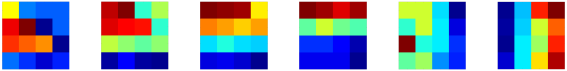
Original image
in false colors



PCA projection of
4 by 4 patches



Patches on top
of their PCA projection

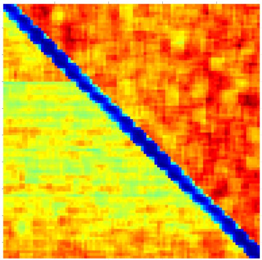


Sample 4 by 4 patches

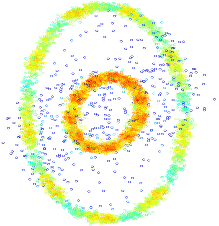
Cut

Step 1: Estimate sample density

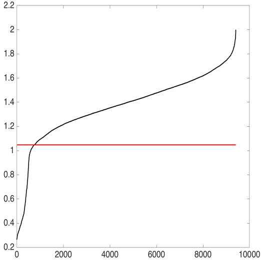
Step 2: Threshold, or cut, according to density



Patch density indexed by top left corner pixel



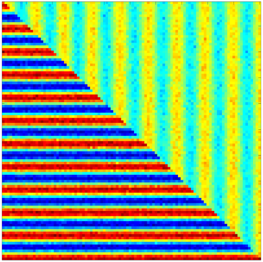
PCA projection colored by density



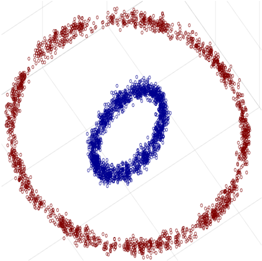
Sorted density with threshold drawn in red

Cluster

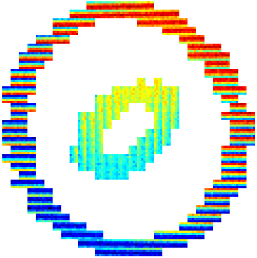
Step 3: Cluster points that passed the threshold



Original image
in false colors



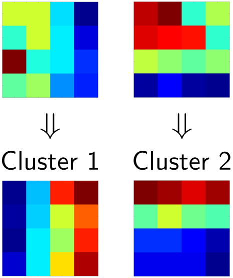
Sampled data clustered



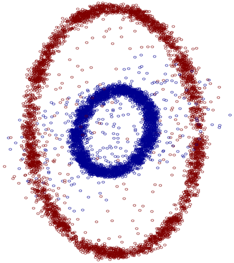
Patches on top of
the clustered projection

No Patch Left Behind

Step 4: Classify remaining points



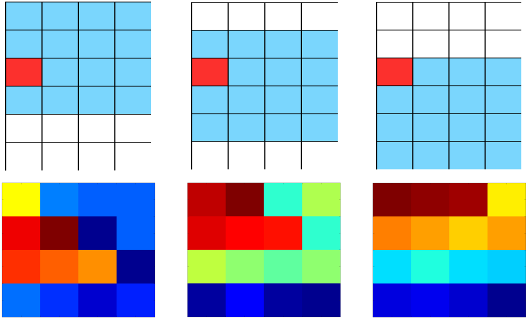
Sample patches being classified.



PCA projection of all 4 by 4 patches classified.

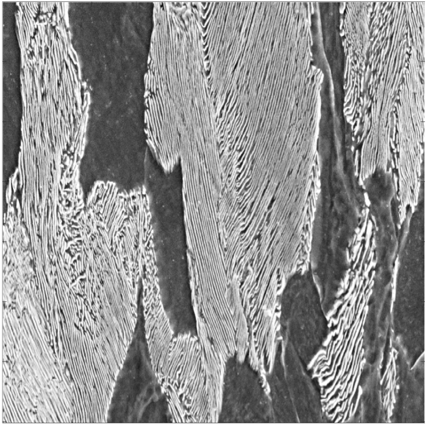
From patches to Pixels

Step 5: Label pixels using patch labels

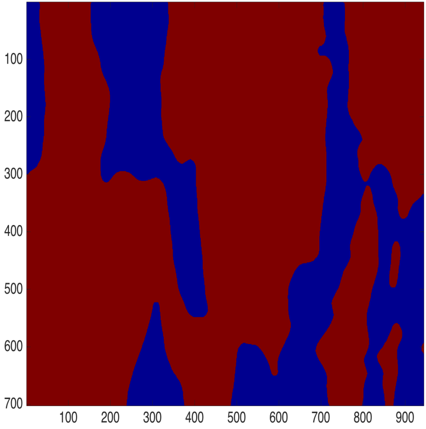


To decide which cluster pixel (3,1) belongs to, we look at all the patches that it appears in. As illustrated in the top images, it only appears in 3 patches. The actual patches, shown on the bottom, were classified to cluster 1, 2, and 2, respectively. This means that cluster 2 gets 2 votes, and cluster 1 gets 1 vote. Therefore, this pixel gets placed in cluster 2.

Real Grayscale Images

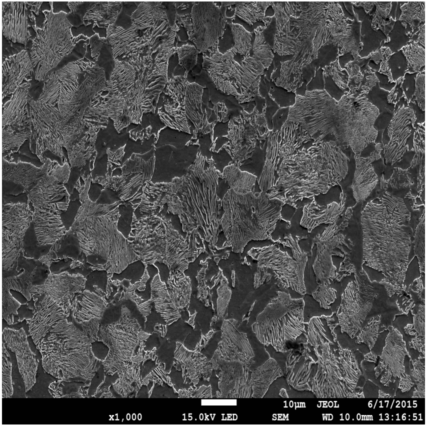


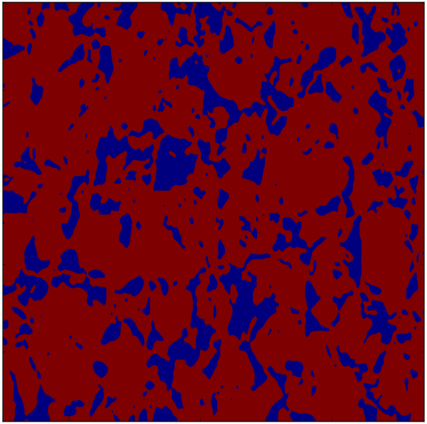
Original image.



Segmentation from 30×30 patches.

Real Grayscale Images

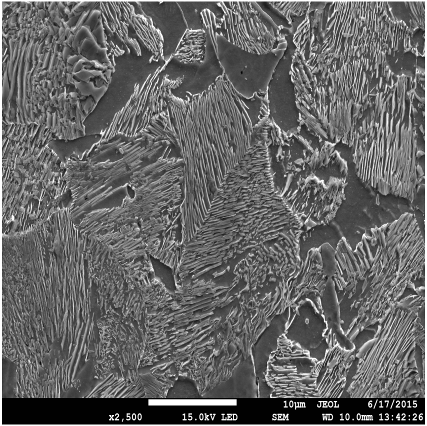




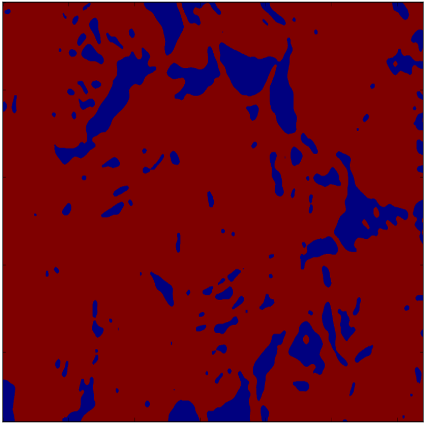
Original image.

Segmentation from 8×8 patches.

Real Grayscale Images

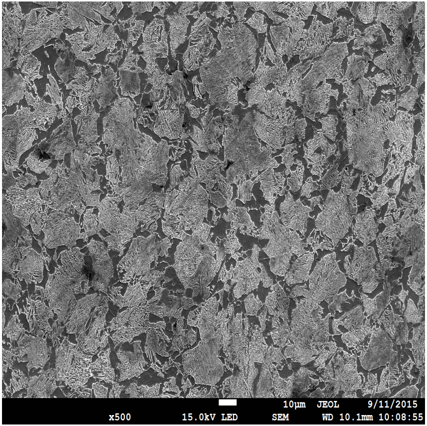


Original image.

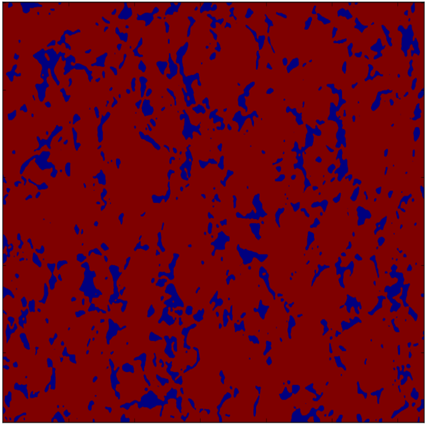


Segmentation from 20×20 patches.

Real Grayscale Images

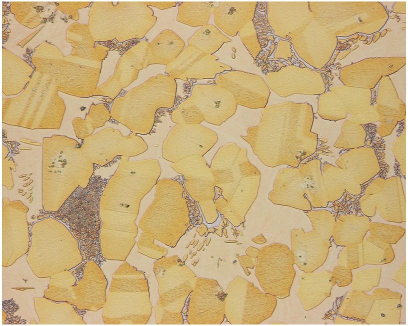


Original image.

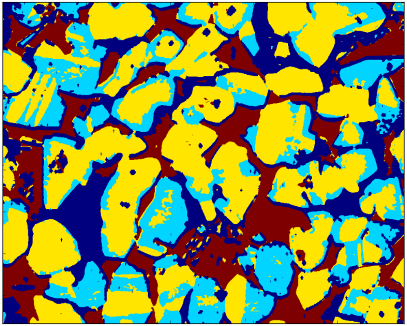


Segmentation from 4×4 patches.

Color Images



Original color image.



Our segmentation.

Concluding Remarks

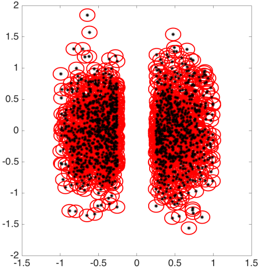
In this research, we have:

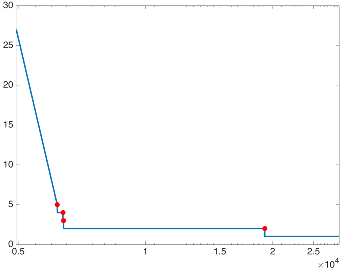
- Created a new mathematical definition for consistency of clustering
- Shown that $B(T_\lambda(q_n), \delta)$ is a consistent clustering of $T_\lambda(q)$
- Developed a clustering method, the Cut-Cluster-Classify, for smooth probability density functions
- Used our density based clustering to do image segmentation

Future Work

Work that still needs to be done:

- Develop the multi-scale image segmentation by considering different patch sizes
- Generalize the theoretical framework
- How do we find the practical δ using persistence





Thank You!