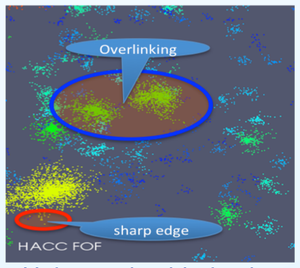
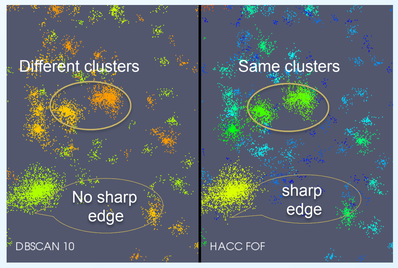
Data Intensive Analysis Techniques and Tools

M. M. A. Patwary (NWU), A. Agrawal (NWU), W. Liao (NWU), A. Choudhary (NWU),
S. Habib (ANL), R. Vatsavai (ORNL), J. Xie (UCD), H. Yu (UNL), K.-L. Ma (UCD)

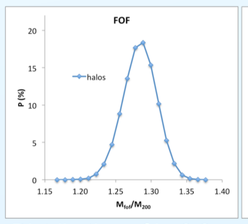
The Structure of Halos: FOF vs. DBSCAN



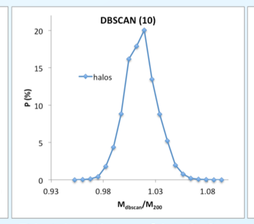
Halos and subhalos in the astrophysics data



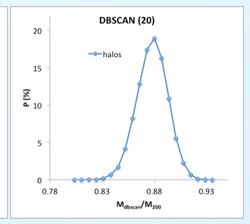
DBSCAN vs. FOF



Distribution of b=0.2 FOF masses for NFW halos with c=5



Distribution of b=0.2 DBSCAN (minpts=10) masses with c=5



Distribution of b=0.2 DBSCAN (minpts=20) masses with c=5

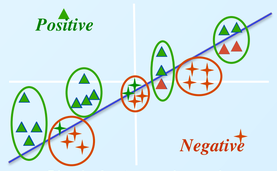
- The dark matter halo mass function is a key repository of cosmological information over a wide range of mass scales, from individual galaxies to galaxy clusters*. N-body simulation shows that Friend-of-Friend (FOF) mass function has a universal form to a surprising level of accuracy. However, observed group and cluster masses are usually stated in terms of a spherical over-density (SO) mass, which does not map simply to the FOF mass. Results from Monte Carlo realizations of ideal Navarro-Frenk-White (NFW) halos and N-body simulation show that FOF and SO map 80-85% halos only if concentrations are known.
- Challenge: Bridged halos complicates the mapping between FOF halo and SO halo.
- Solution: Investigating the mapping between DBSCAN halo and SO halo: Contrast the properties of DBSCAN with FOF; Investigating relation of DBSCAN to percolation theory similar to FOF; and Investigating whether over-linking problem of FOF can be mitigated by DBSCAN.
- Experiment: DBSCAN with 100,000 Monte-Carlo samples, 1,000 particles per sample, c = 5.
- Results: NFW and DBSCAN mass ratio is close to 1, amplitude is high, and deviation range is smaller.
- Collaboration with HACC group, P.I. Salman Habib, at Argonne National Lab.

*Z. Lukic, D. Reed, S. Habib, and K. Heitmann. *The Structure of Halos: Implications for Group and Cluster Cosmology*. The Astrophysical Journal, 692:217–228, 2009.

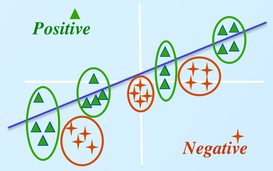
STPMiner

- STPMiner is an high-performance spatiotemporal pattern mining toolbox for analyzing big spatiotemporal datasets.
- Solution: It offers computationally efficient data mining primitives tailored for heterogeneous architectures.
- Applications: Spatial classification (Land use/land cover mapping), clustering (earth science), change detection (biomass monitoring), and co-location pattern detection (climate change impacts).

Multiple Instance Learning (MIL)



Single Instance Learning



Multiple Instance Learning

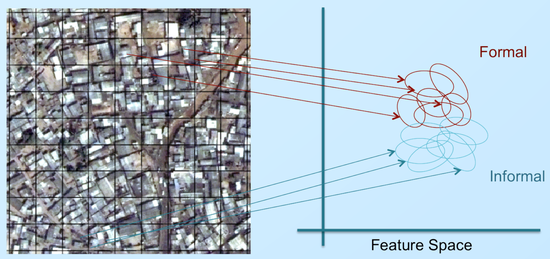
Problem being solved

- Land-use/land-cover classification using very high-resolution (VHR) remote sensing imagery

State of the art

- Single instance learning algorithms (statistical, decision trees, neural networks, …) are not efficient for recognizing complex patterns in VHR images
- MIL approaches like Citation-KNN are computationally expensive

Bag of Gaussian MIL (BoG MIL)



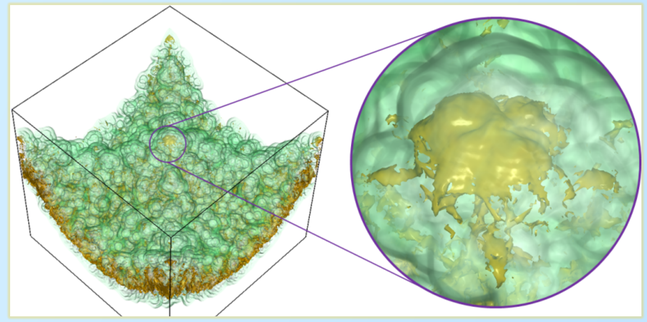
- BoG MIL is a novel computationally efficient algorithm
- Models all instances in a segment as a Gaussian distribution
- Each land use/land cover is modeled as bag of Gaussian (as opposed to single Gaussian per class)
- Prediction is based on statistical matching and ranking

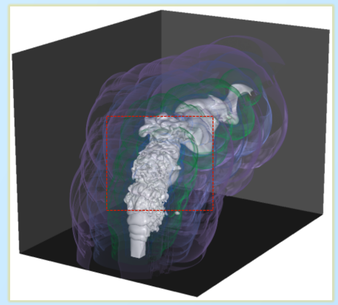
Accuracy:

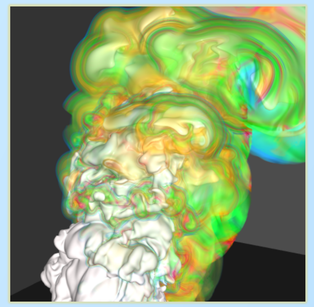
City	Citation-KNN	Regression	RF	MLP	NB	BoG MIL
Accra	76.25	71.25	72.08	69.58	75.66	95.66
Caracas	82.96	78.15	81.85	81.81	74.07	85
La Paz	80.97	77.17	78.26	80.23	76.08	83.25
Kandahar	79.78	64.89	69.14	73.93	60.1	81.2

BoG MIL classified image overlaid on raw image

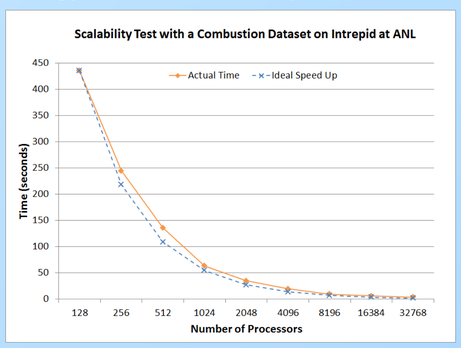
Distance Field Based Analysis & Visualization







- Computing distance fields is a fundamental requirement for many algorithms of data visualization and analysis.
- We have designed and implemented a new spatial data structure, named parallel distance tree, to enable highly scalable parallel distance field computing.
- The method is general to support various data types (including, but not limited to, polygonal objects, point/particle data, and volumetric data) and handle different distance metrics (including, but not limited to, Euclidean distance, City block distance, and Chessboard distance).
- We have integrated our method with real-world large scientific simulations to support in-situ processing and data reduction.
- The design does not depend on any particular architectures, and the scalability has been demonstrated on state-of-the-art supercomputers.
- The resulting technology will benefit many application areas from fusion, combustion, to climate, and astrophysics simulations.



SDAV Scalable Data Management, Analysis, and Visualization

U.S. DEPARTMENT OF ENERGY | Office of Science