Multi-Scale Structure in 3-D Surveys and Simulations of the Universe

Michael Way (NASA/Goddard Institute for Space Studies)
Paul Gazis, Jeffrey Scargle (NASA/Ames, Space Sciences Division)

http://astrophysics.arc.nasa.gov/~mway/lss201010.pdf

Uppsala Oct 2010

- Three structure analysis methods:
  - Kernel Density Estimation
  - Bayesian Blocks
  - Self-Organizing Maps

- Three data sets:
  - Sloan Digital Sky Survey DR7
  - Millennium Simulation
  - Random/Uniform/Independent "Poisson"

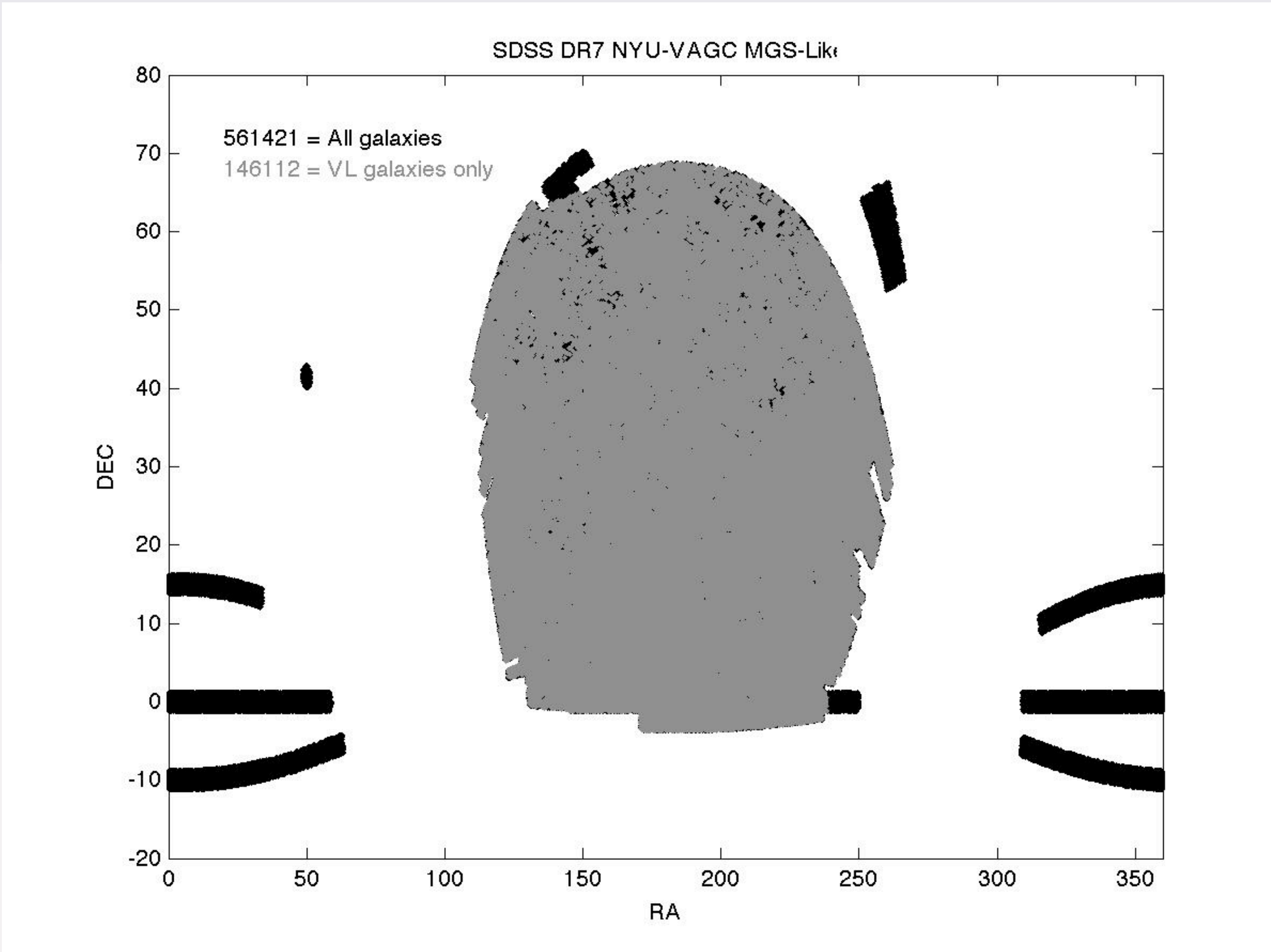
Three cornerstones of Data Mining and Machine Learning

Three Steps

• Points	→→→	Density Estimate
• Density Field	→→→	Cluster Identification
• Clusters	→→→	Classification

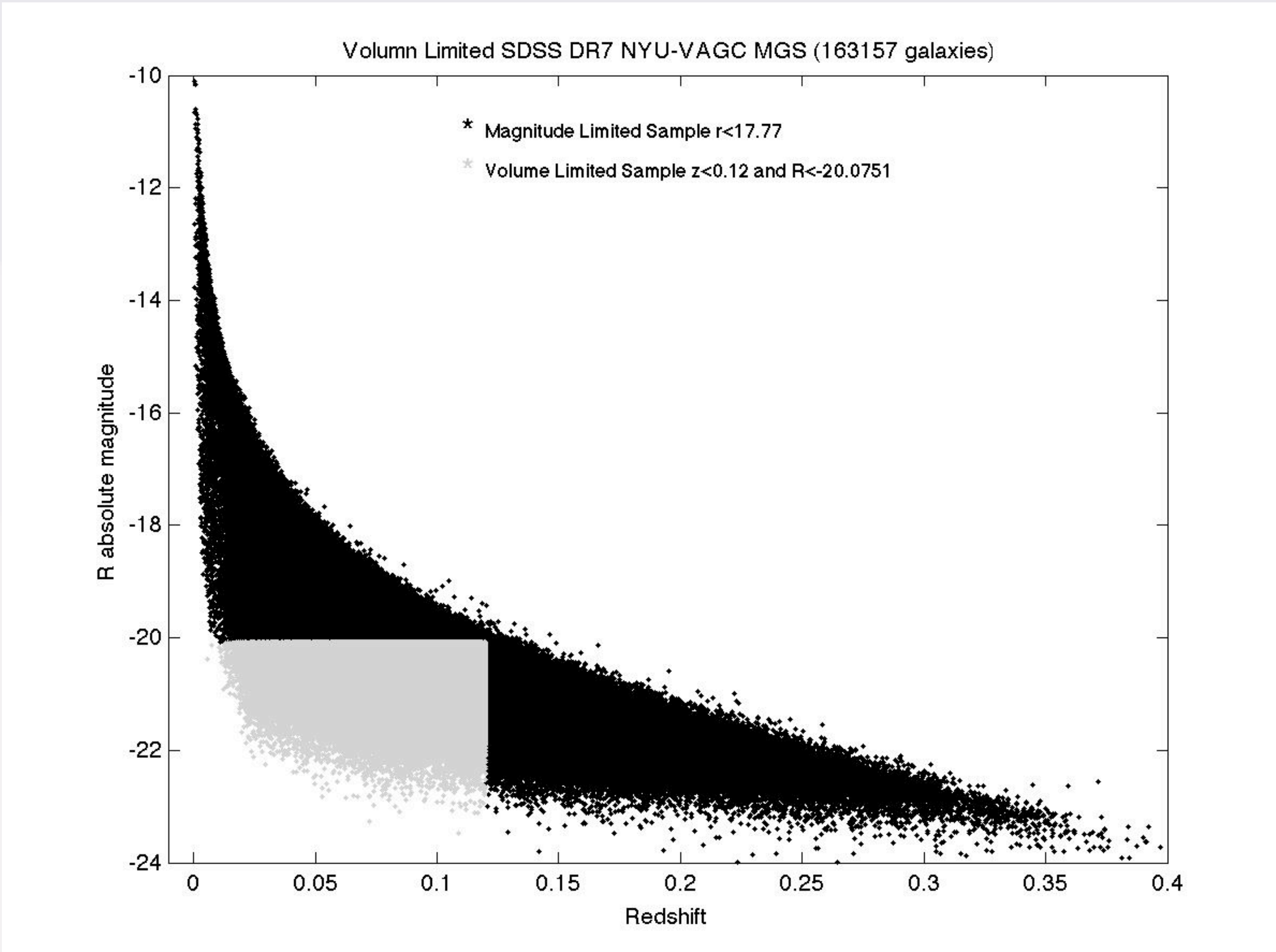
The SDSS DR7 NYU-VAGC

The SDSS Data Release 7 "MAIN" Galaxy Sample



SDSS DR7 Vol. Limited

Picking the volume limited sample



3-D Catalogs

Millennium Simulation & Poisson Catalogs

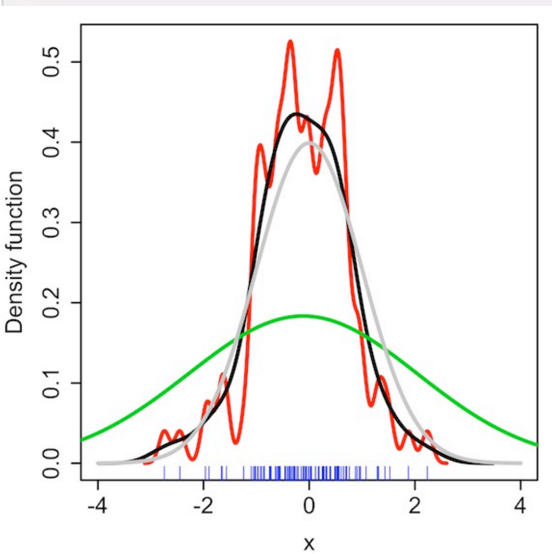
- The same methods used to create the SDSS catalog are used derive a similar catalog from the Millennium Simulation
  - N=656855 Galaxies → N_{VL}= 171,390 Volume Limited

- A Randomly Distributed Uniform Sample is also generated with roughly the same number of points and a similar volume. N=144,700

Method 1: AKM

Method 1: The Adaptive Kernel Density Estimation

- 1-D example: let $x_1, x_2, \ldots x_n$ be an independent and identically distributed random sample drawn from some unknown density $\boldsymbol{f}$.
- We want to know the shape of this function $\boldsymbol{f}$
- An estimate of its shape can come from the kernel density estimator. K=kernel (Gaussian is common), h=bandwidth (width of the kernel, which is a free parameter)



$$f_h(x) = \frac{1}{nh} \sum K\left(\frac{x - x_i}{h}\right) \qquad K = e^{-\frac{x^2}{2h^2}}$$

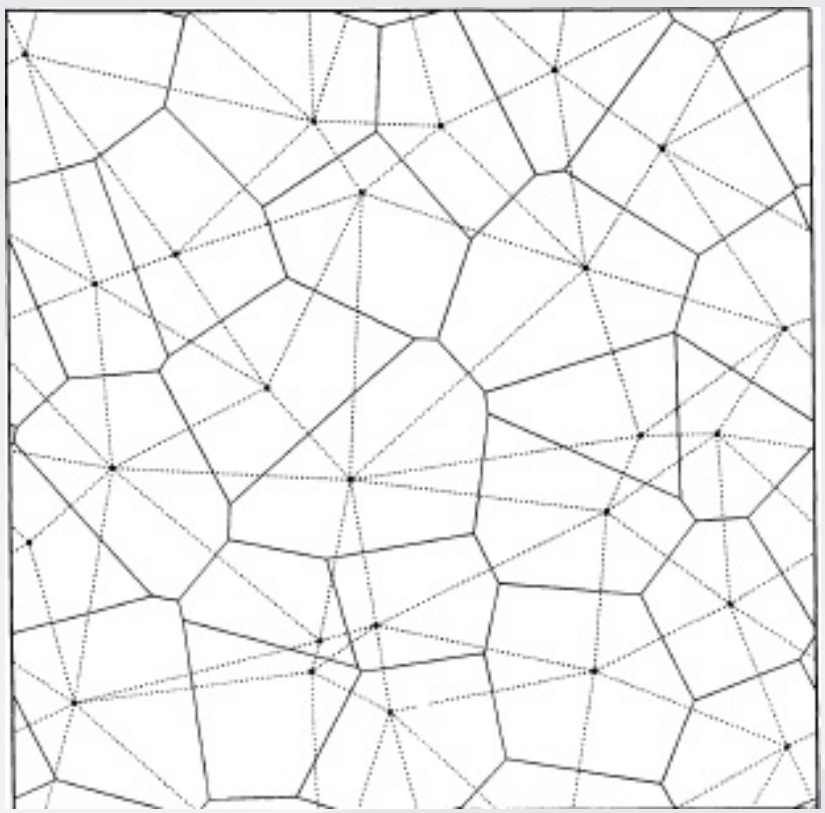
Voronoi Tessellation

1) N data points generate N cells
2) Cells and data points are in a one-to-one correspondence
3) Union of all N cells is the entire data space
4) Intersection of any pair of cells is empty (no overlap)
5) Cell boundaries are flat 2-D polygons
6) Tessellation yields a data structure containing
    a) Estimate of the local point density: 1/V, V=cell volume
    b) 3-D vector from cell centroid to data point estimates local density gradient in both magnitude and direction
    c) Nearest-neighbor information is encoded in vertices of bounding polygons: Two cells can be adjacent in 3 ways: Do they share at least one vertex, edge or face? (Each is included in the next)

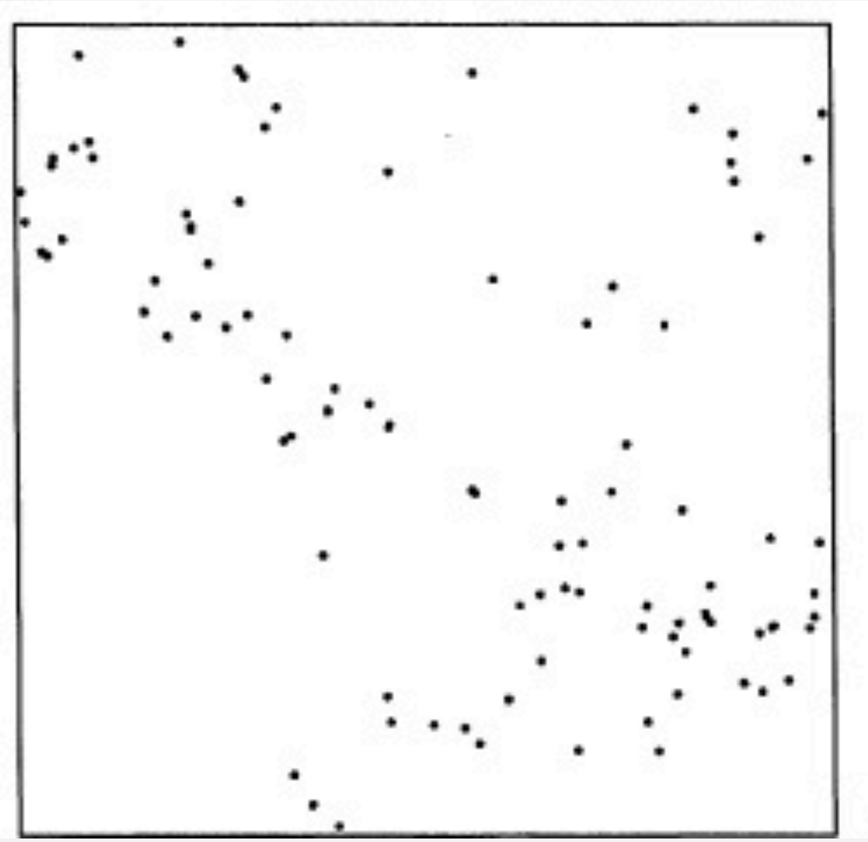
Voronoi Tesselation

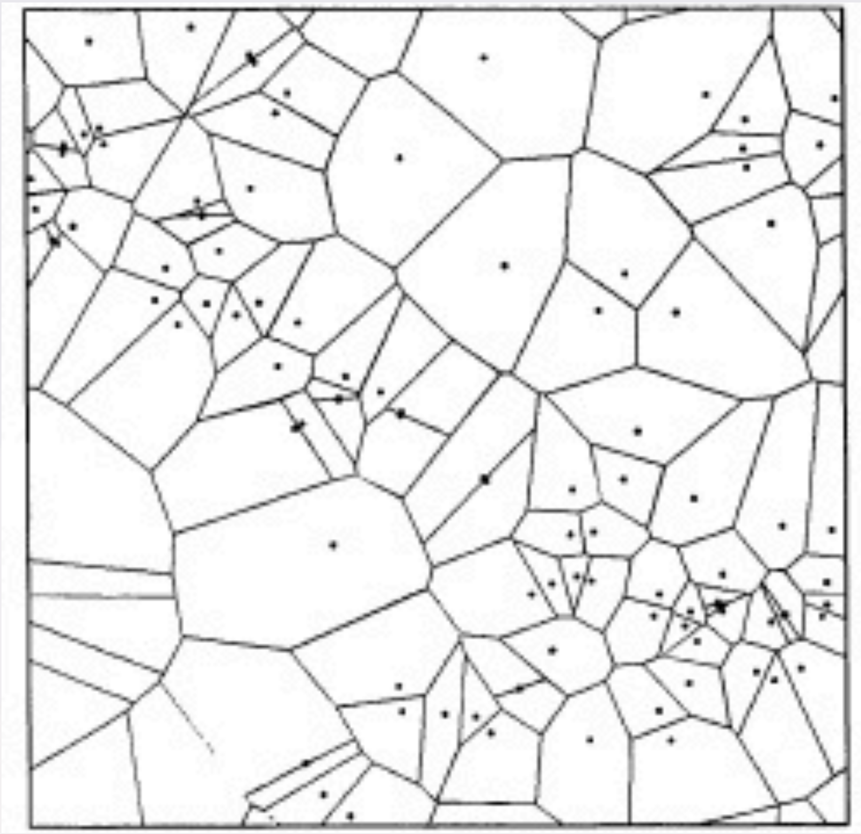
Here we have a 2-D Voronoi Tesslation (thick lines) and its corresponding Delaunay triangulation (thin lines).

from Icke and van de Weygaert 1987 (Figure 1)



Voronoi Tesselation





Methods 2 & 3

Now that we have our Voronoi Tessellation lets look at the methods we use to find structures.

1.) Bayesian Blocks

2.) Self-Organizing Maps

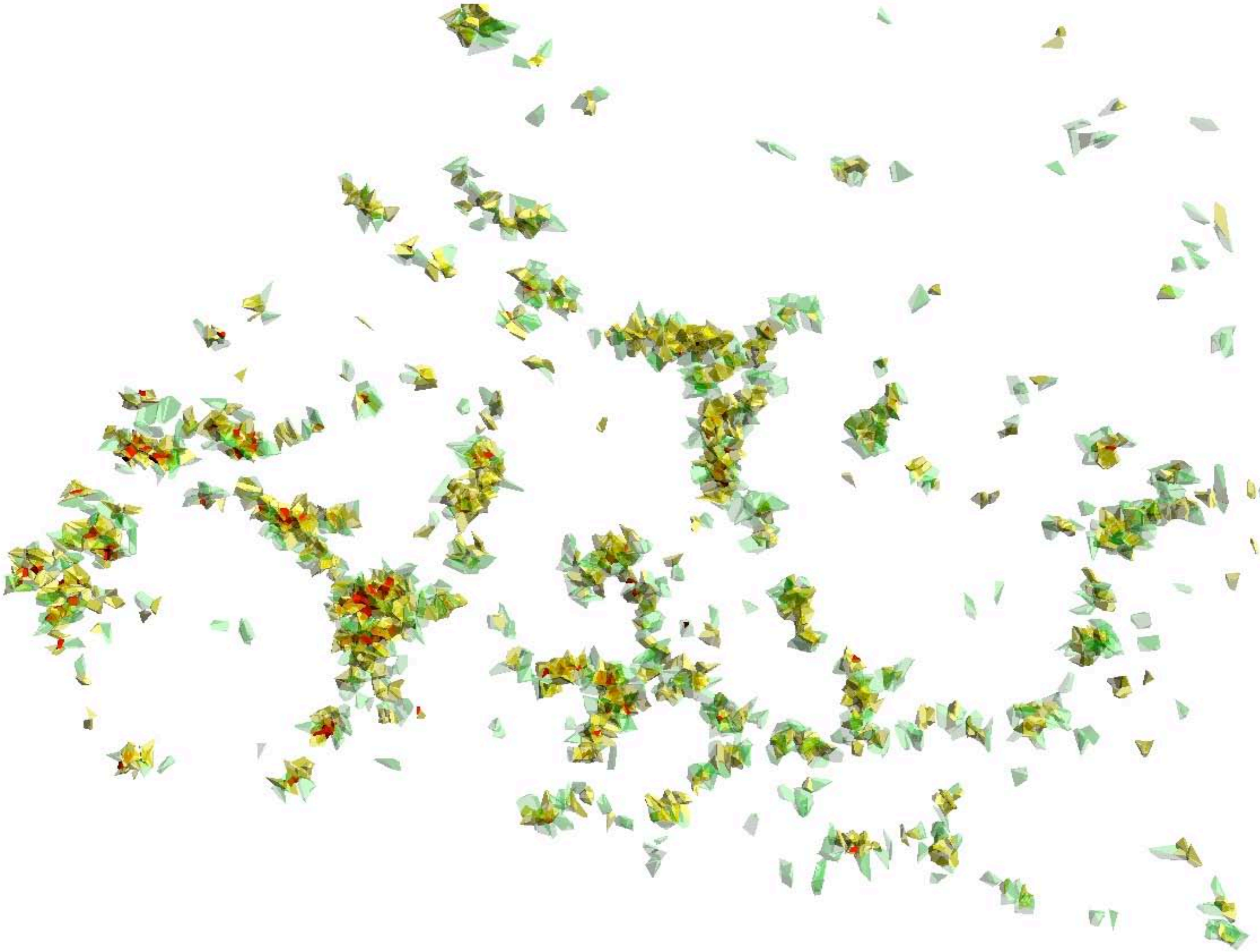
Bayesian Blocks

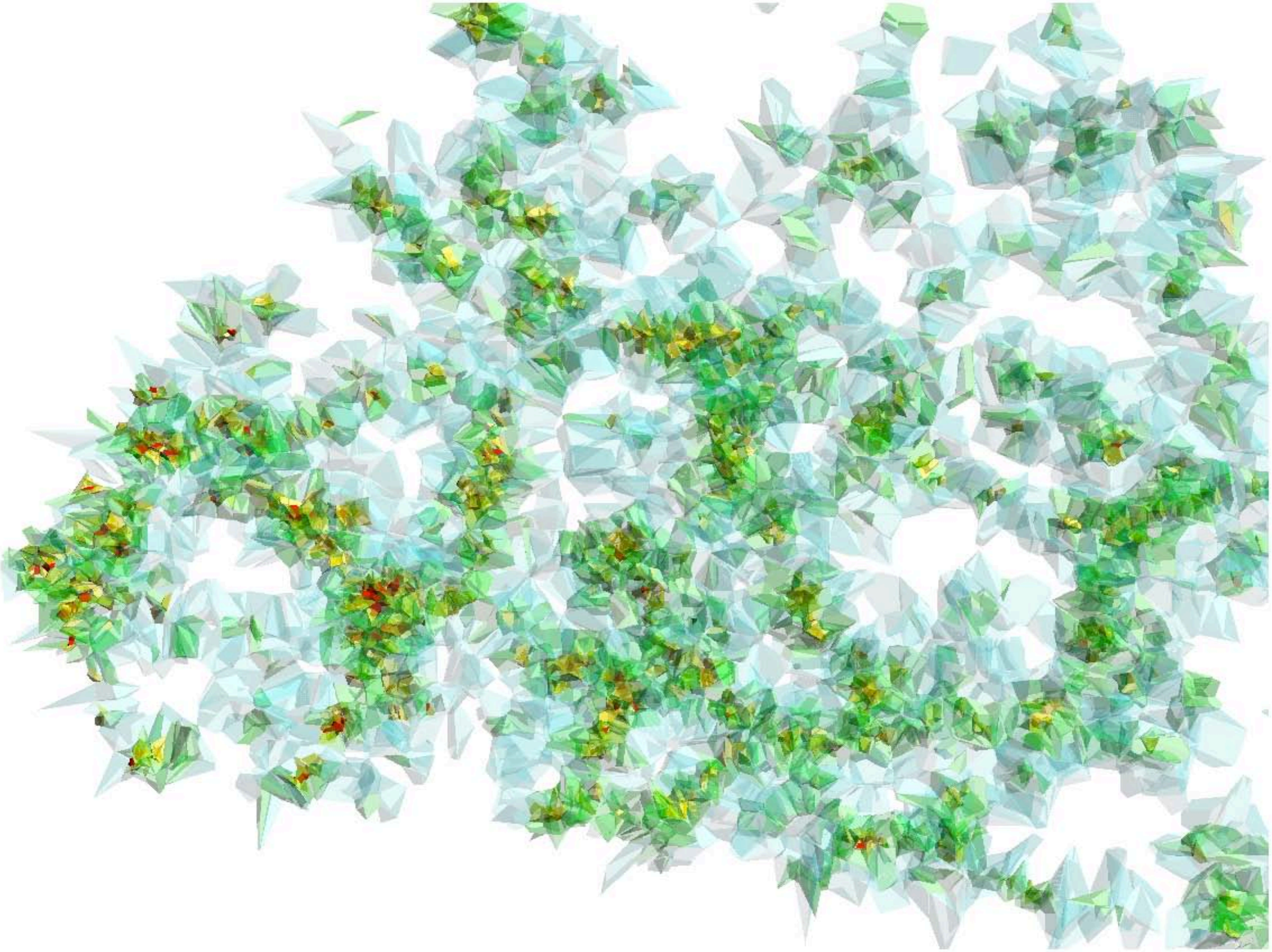
1) Partition data space with a set of surfaces enclosing 3-D solids
2) Assign a constant density to each solid = #galaxies/volume

Done via an optimization procedure designed to:

1. express spatial density variations that are real (true signal)
2. suppress statistical fluctuations that are not real (noise)

[See Scargle 1998 and Jackson et al. 2005 for the 1-D version]



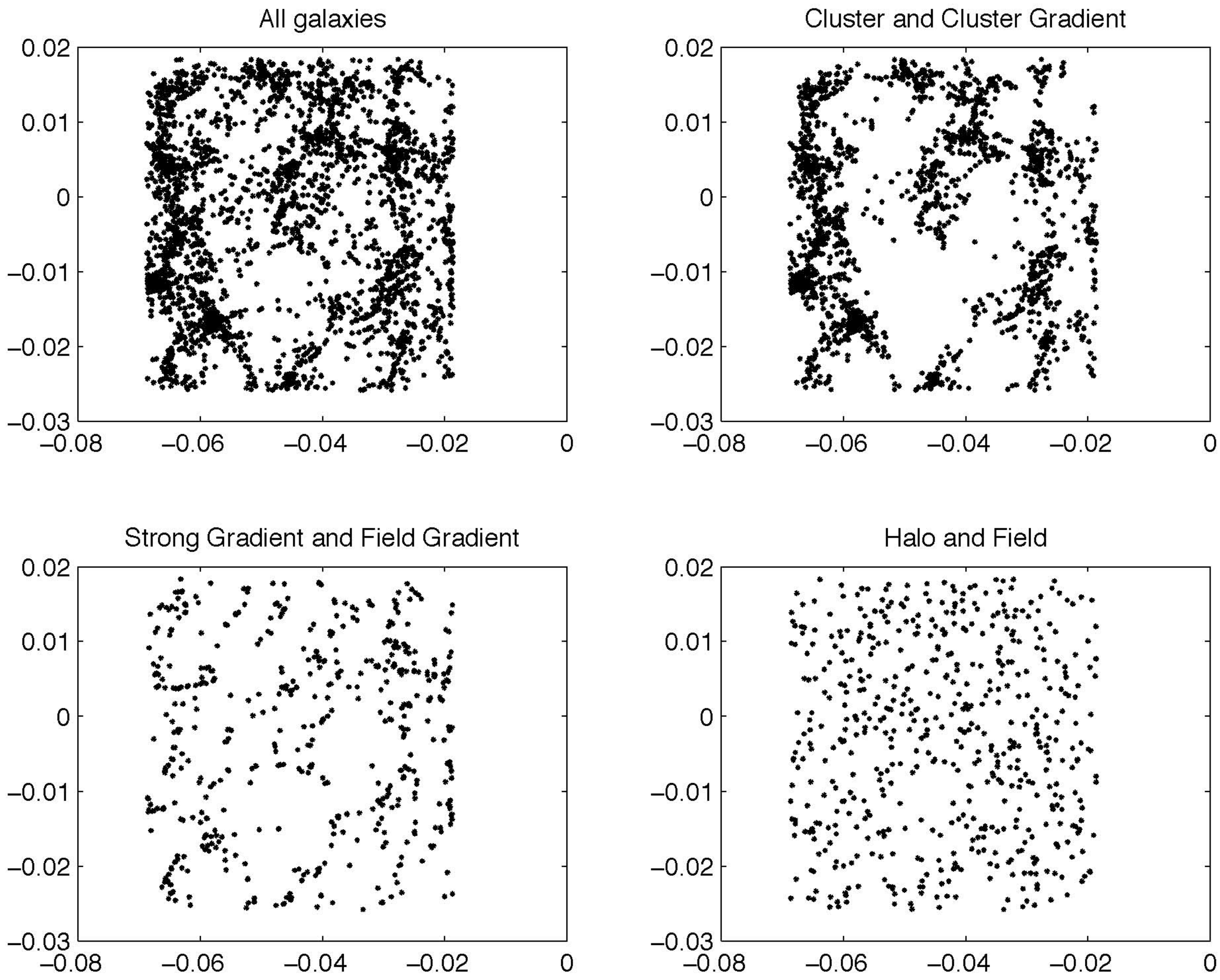


Self-Organizing Maps

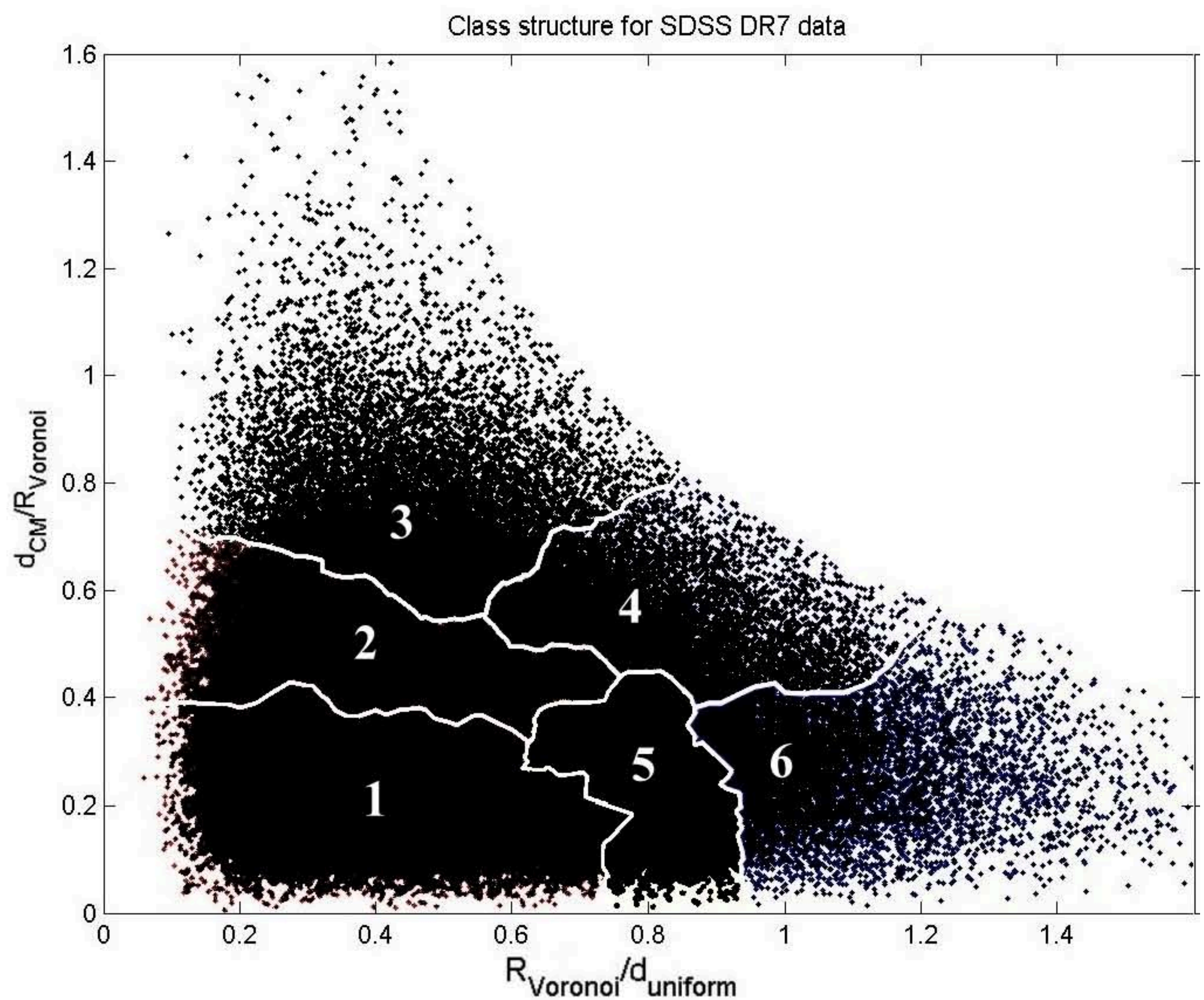
1) Map points from a N-Dim data space into an array of cells of principle elements (PE) in a classification space of reduced dimensionality (2-D here)
2) Designed (as much as possible) to reproduce the topological structure of the input distribution

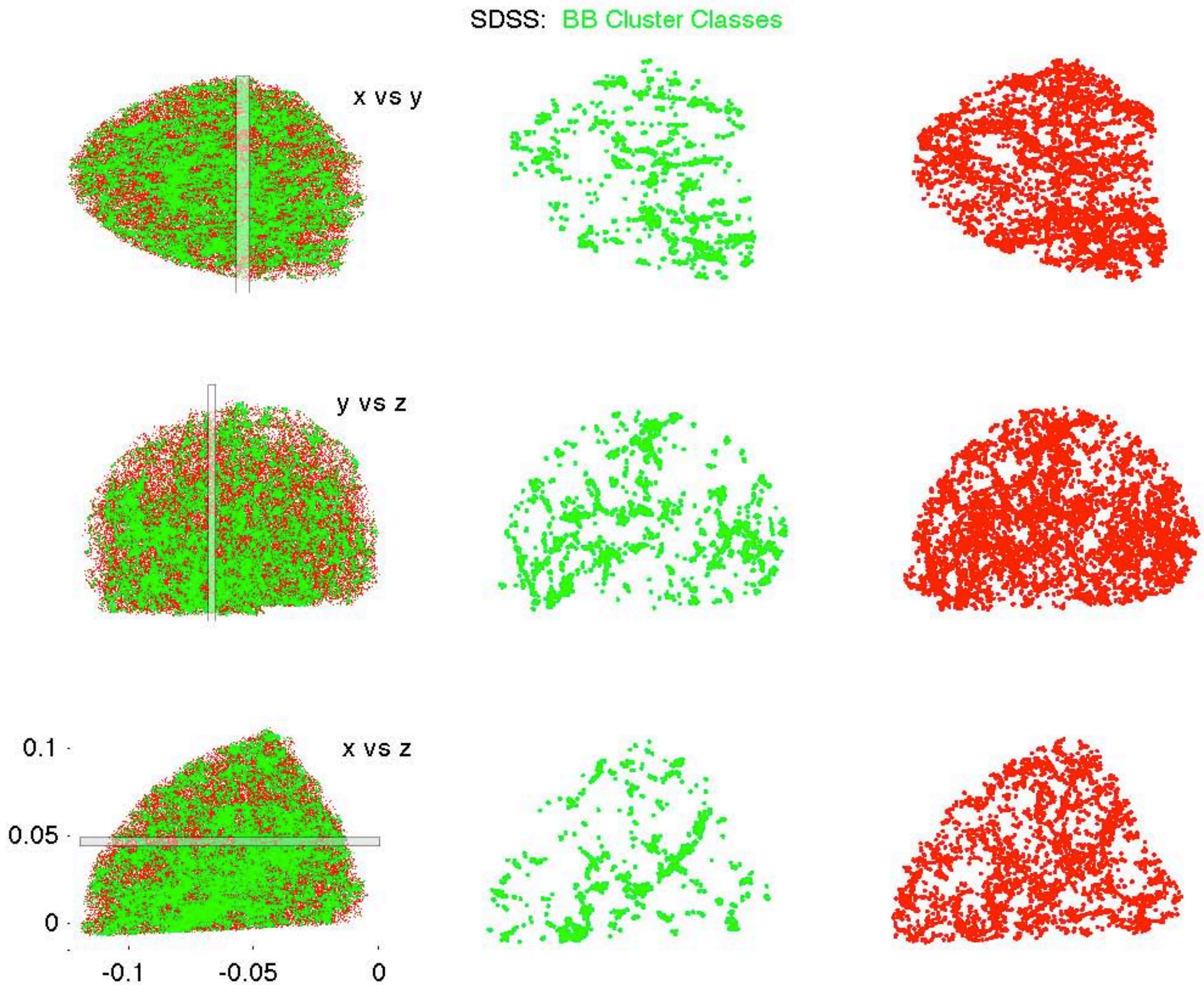
Attempts to map adjacent clusters in the input space into adjacent adjacent blocks of contiguous PEs in the output space

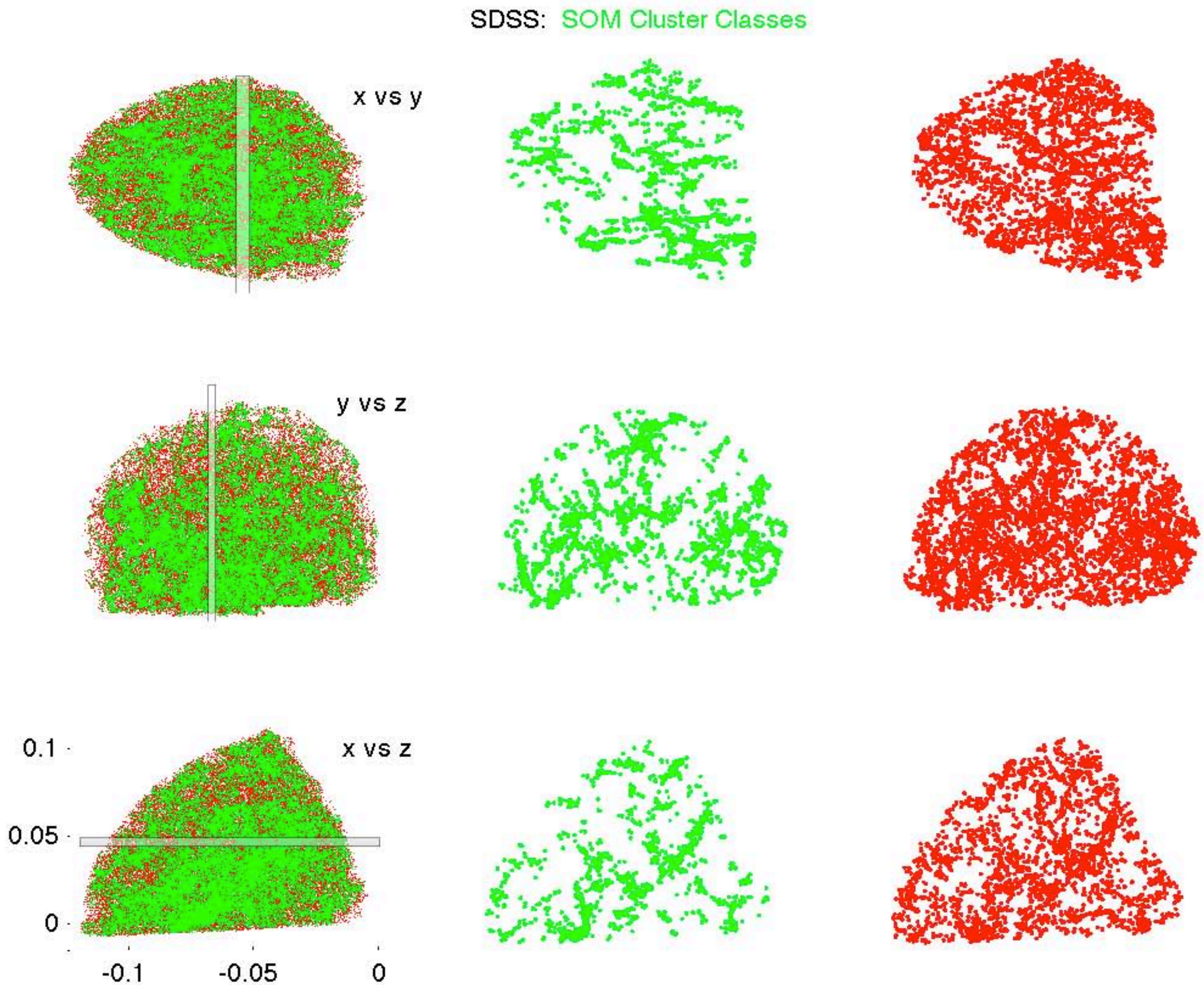
Locations in SOM phase space of types of galaxies identified by the SOM

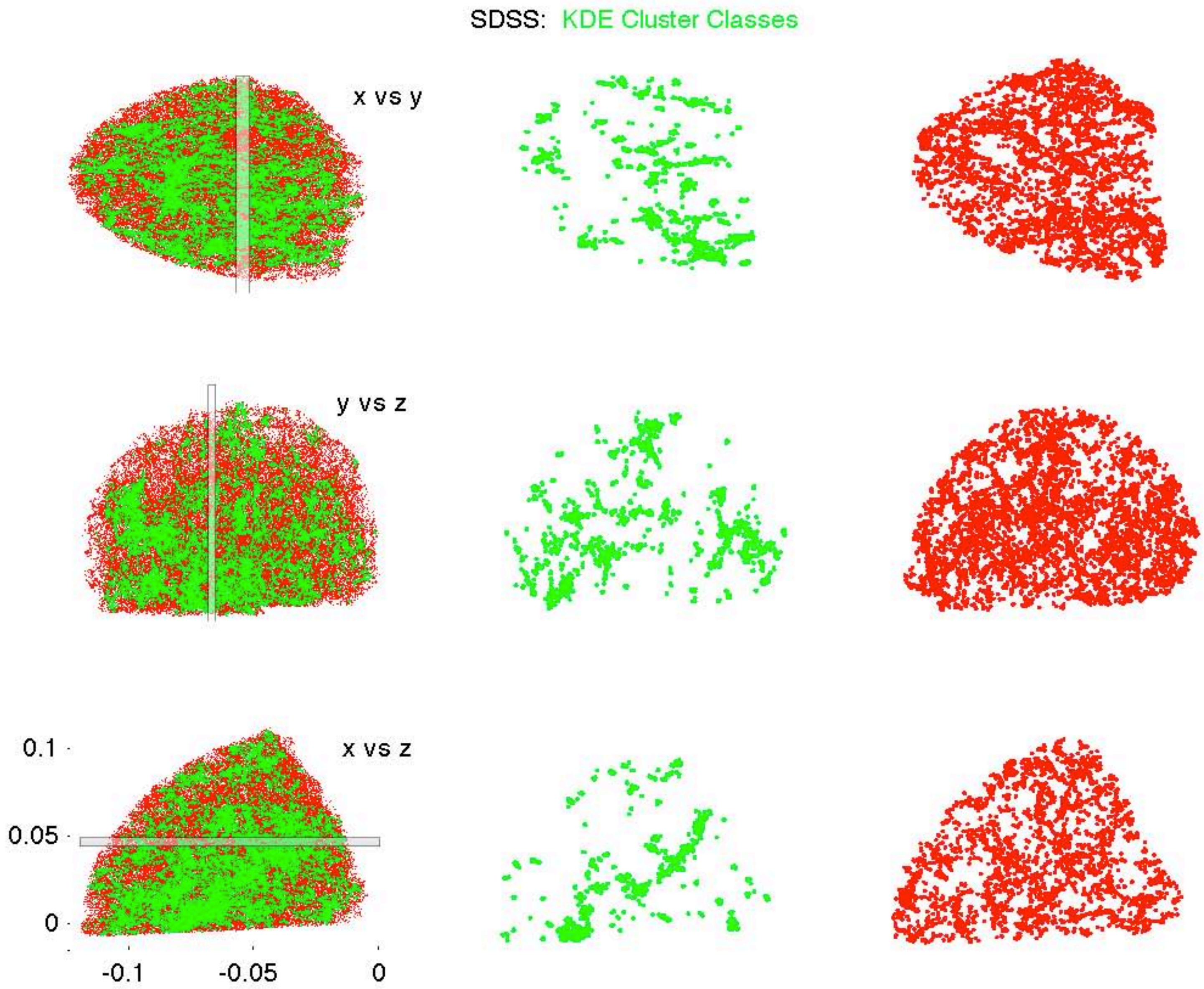


Locations in neighbor-distance/cell-vol space of galaxies assigned to various SOM classes

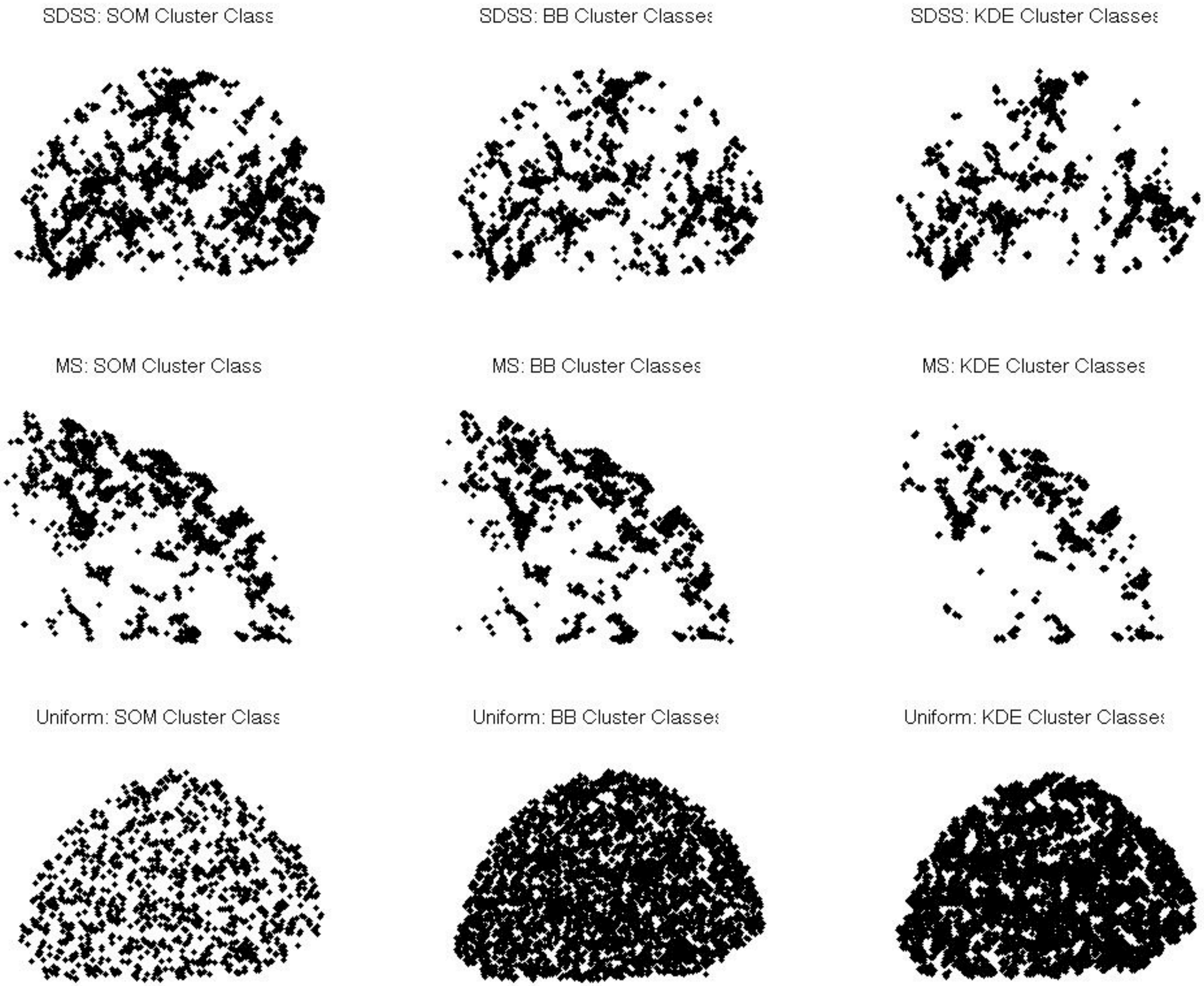



SDSS: BB Cluster Classes
x vs y
y vs z
x vs z
0.1
0.05
0
-0.1
-0.05
0


SDSS: SOM Cluster Classes
x vs y
y vs z
x vs z
0.1
0.05
0
-0.1
-0.05
0



SDSS: KDE Cluster Classes
x vs y
y vs z
x vs z
0.1
0.05
0
-0.1
-0.05
0


SDSS: SOM Cluster Class
SDSS: BB Cluster Classes
SDSS: KDE Cluster Classes
MS: SOM Cluster Class
MS: BB Cluster Classes
MS: KDE Cluster Classes
Uniform: SOM Cluster Class
Uniform: BB Cluster Classes
Uniform: KDE Cluster Classes

Conclusions

- Multi-scale structure in SDSS and Millennium Simulation have similar characteristics.

- Poisson is qualitatively different from SDSS and MS
- BB and SOM provide similar representations of structures

- KDE similar, but doesn't consistently identifying same structures
- Poisson distribution proved a challenge for all three methods – as it should since there is no structure.

Future?

- Catalog of multi-scale structures in SDSS & MS:
  - Clusters of galaxies, Filaments, Voids
- Comparisons with other cluster and void finders
  - Dynamical Quantum Clustering
  - Watershed Void Finder, BCG, C4, etc…
- Environmental correlations with type and color
- Paper II: Catalog which anyone can use for any algorithm – easier to make comparisons between methods!!