

Bottom-Up Sub-space Clustering

A Symbiotic Approach

Ali Vadhat and Malcolm Heywood

The Subspace Clustering Task

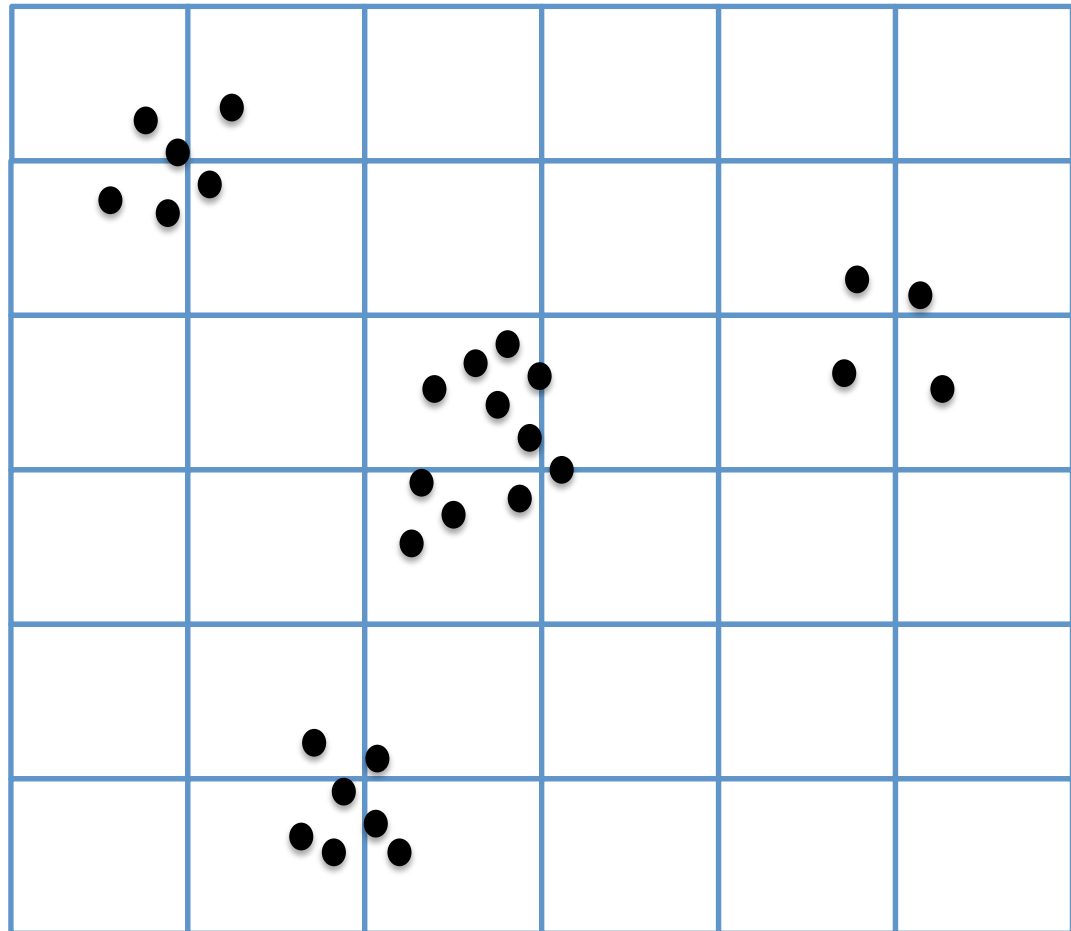
- Lance Parsons et al., (2004) Subspace Clustering for High Dimensional Data: A review.
 - SigKDD Explorations. 6(1):90-105
 - <http://dl.acm.org/citation.cfm?id=1007730.1007731&coll=DL&dl=ACM&CFID=49724896&CFTOKEN=93194701>

Generic Approaches 1: Cell based

Parameterization of
Grid structure?

Cluster definition a
Function of cell count

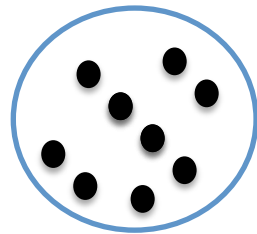
- Clique (1998),
- MineClus (2003),
- Schism (2004)



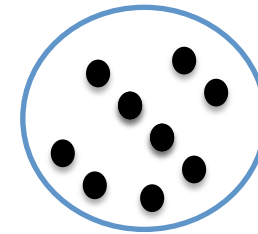
Generic Approaches 2: Density Based

- Projection of data points to each axis
 - Histogram constructed relative to each axis
 - Can be used to construct a dynamic grid
 - ESC assumes this as the starting point
 - Maps a continuous valued optimization task into a discrete combinatorial one...
- SubClu (2004), Fires (2005), Inscy (2008)

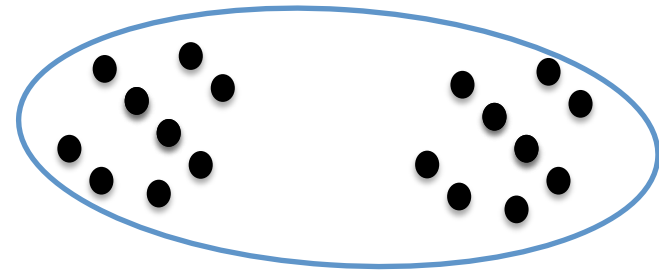
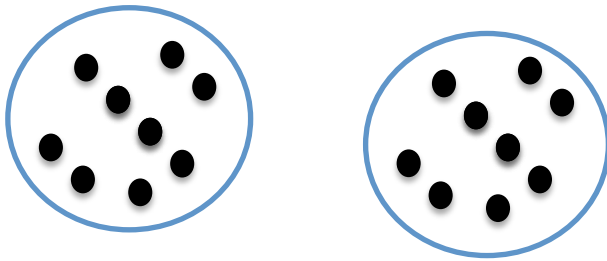
Generic Approaches 3: Partition Based



3 cluster limit

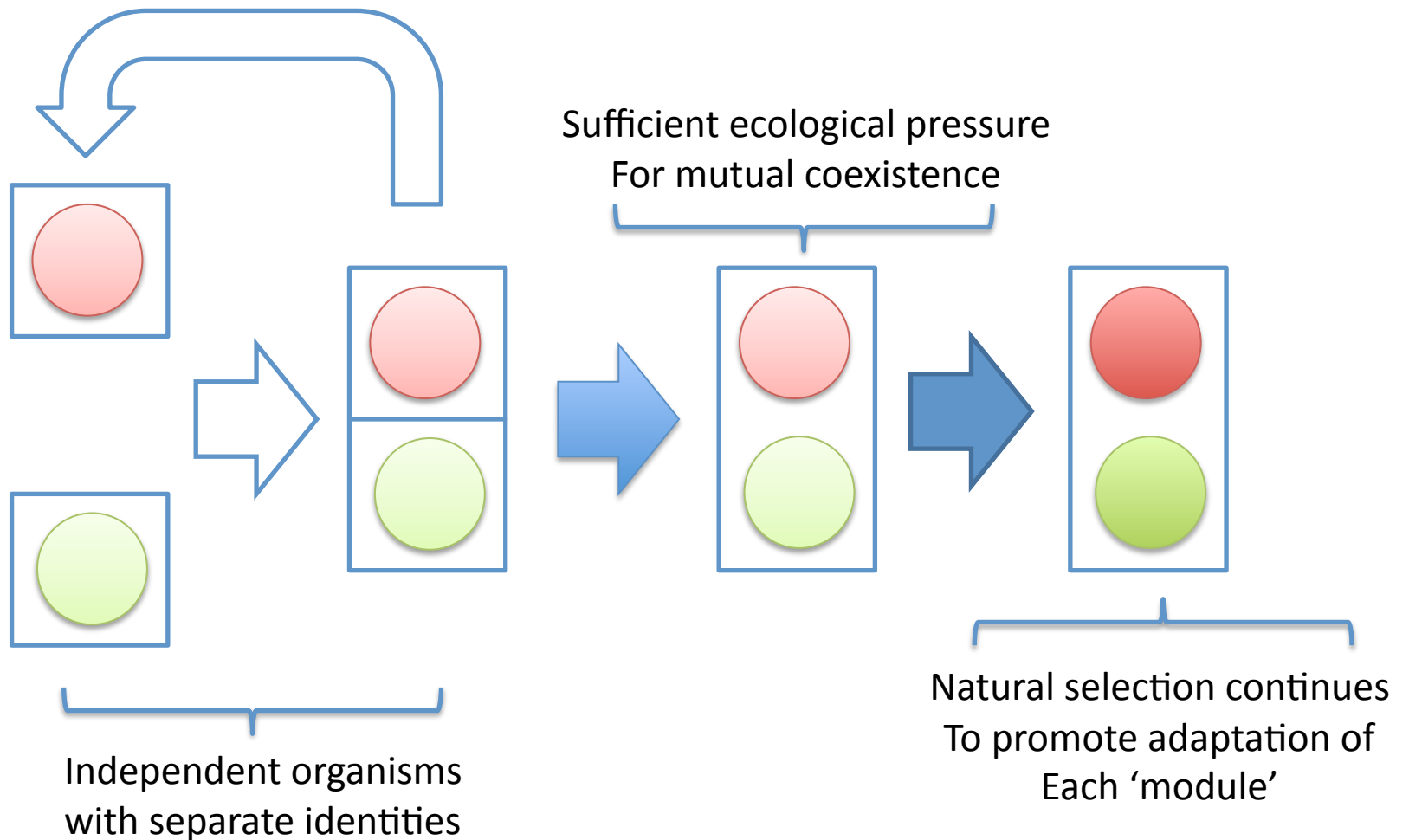


2 cluster limit

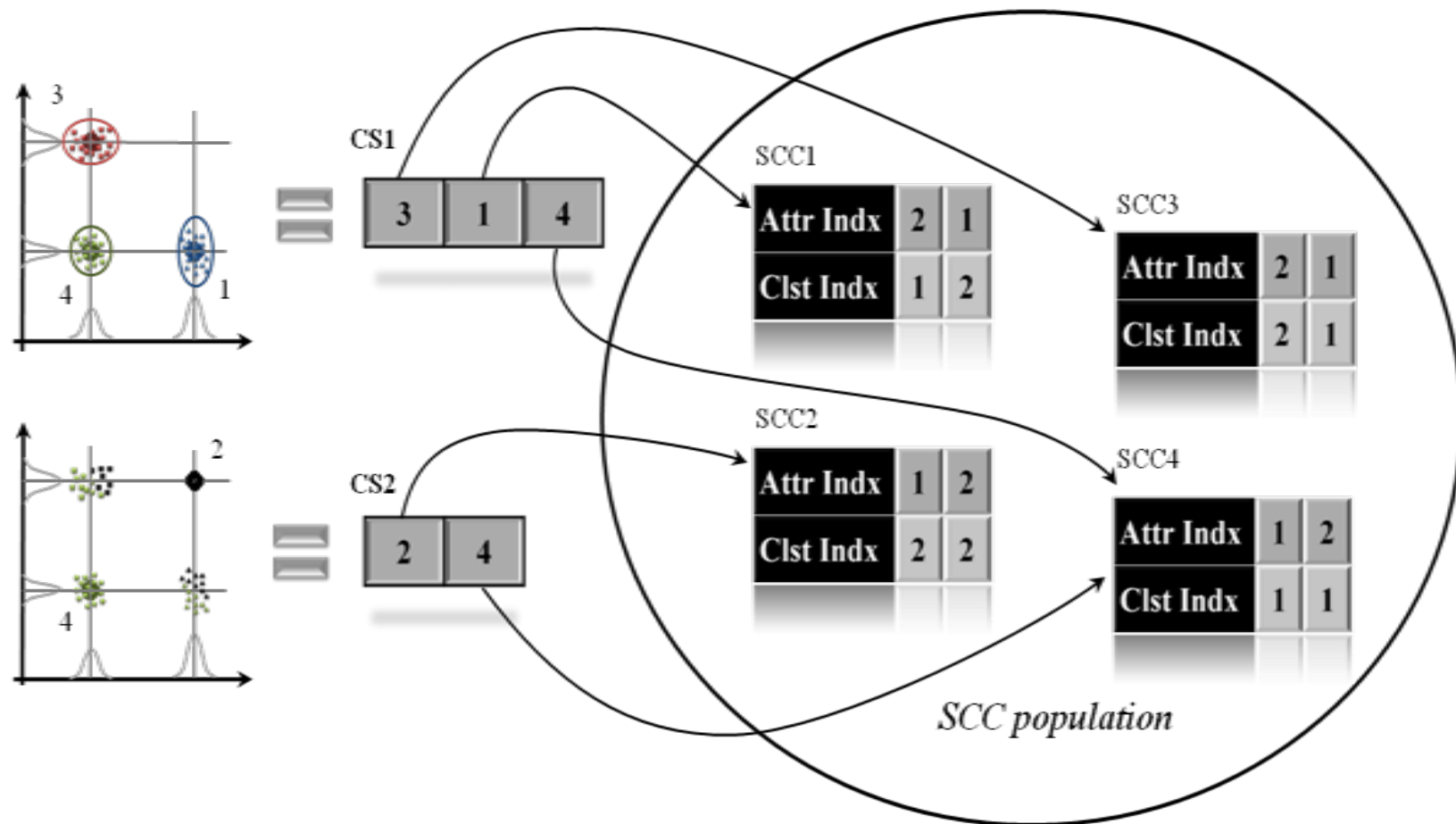


ProClus (1999), P3C (2006), STATPC (2008)

Ecological Model of Symbiosis

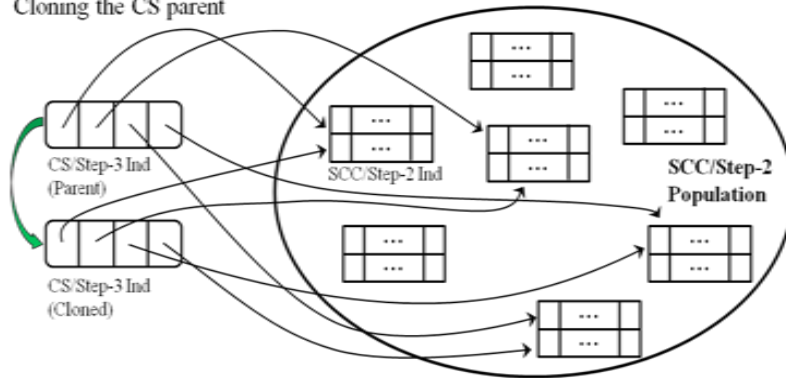


Evolutionary Subspace Clustering (ESC) 1: Representation

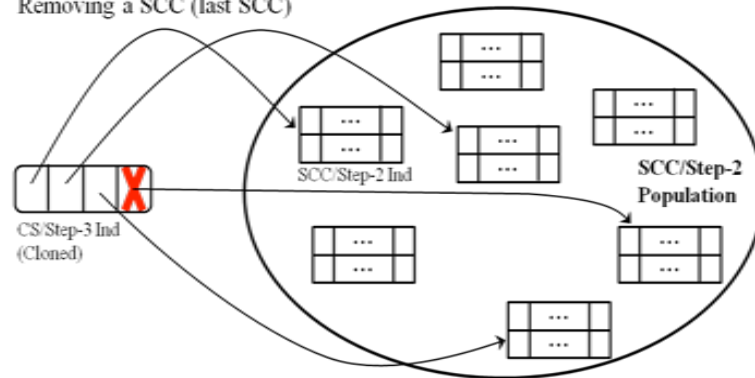


Evolutionary Subspace Clustering (ESC) 2: 'Host level' variation operators

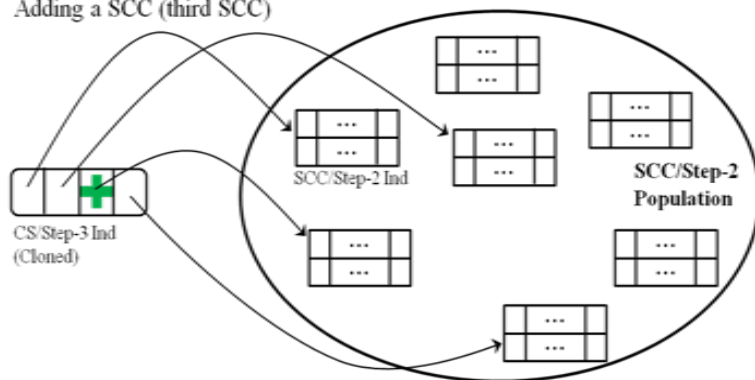
Phase 1:
Cloning the CS parent



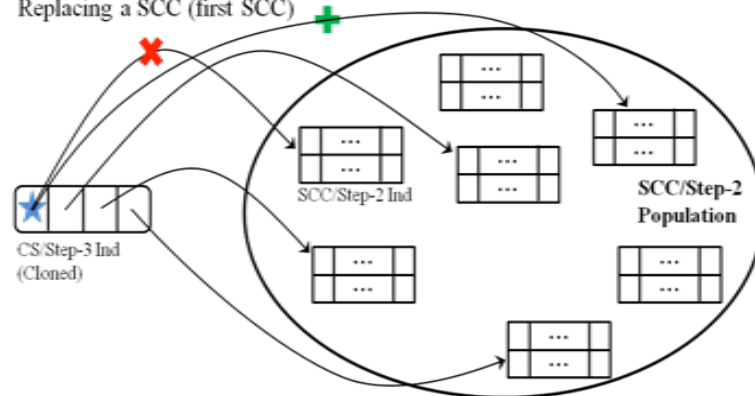
Phase 2:
Removing a SCC (last SCC)



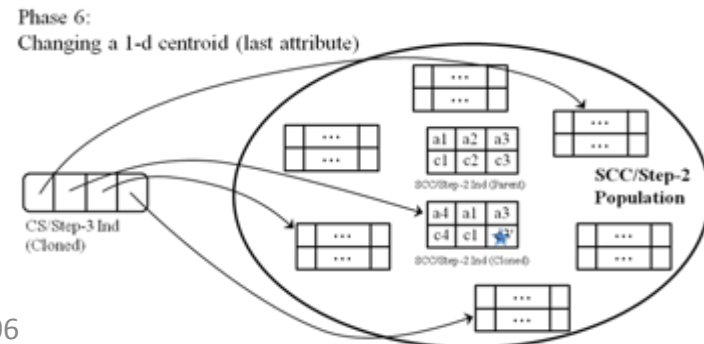
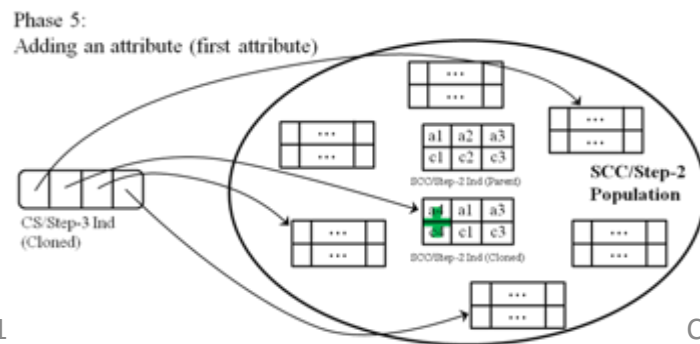
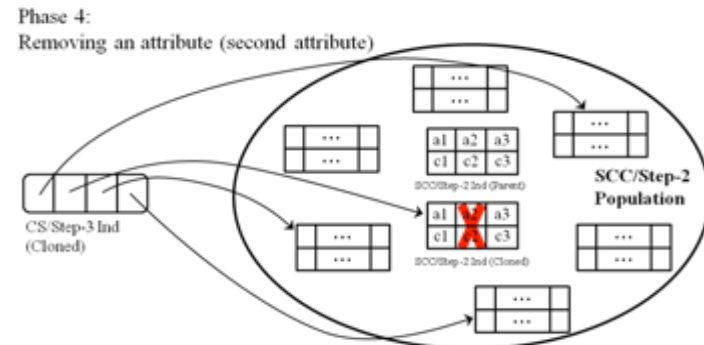
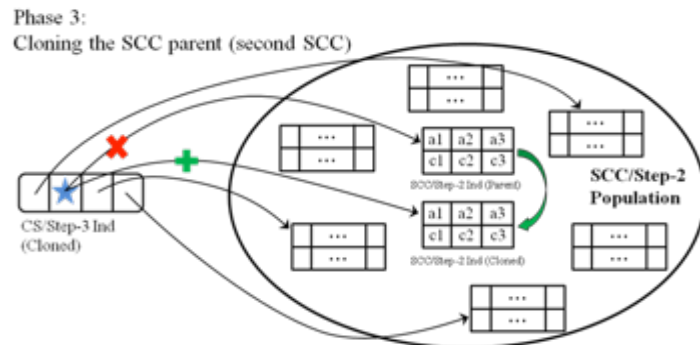
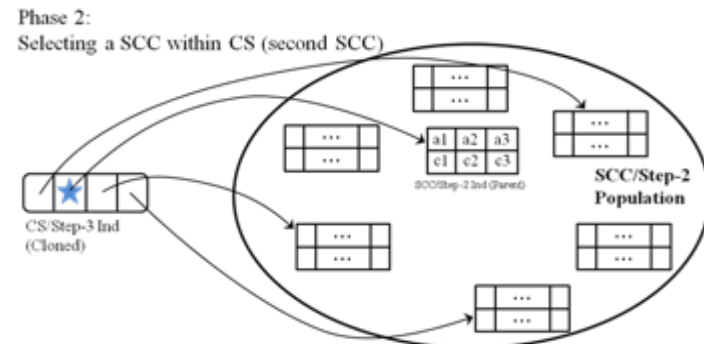
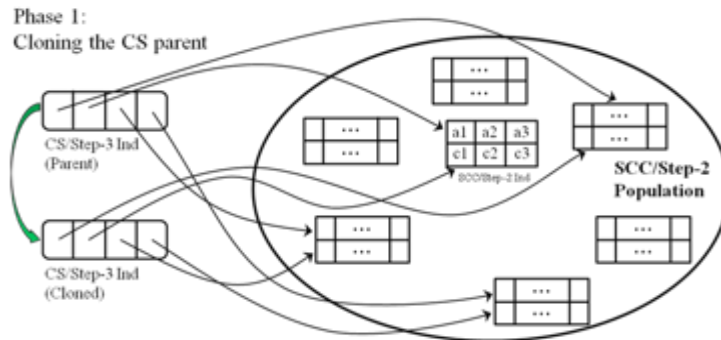
Phase 3:
Adding a SCC (third SCC)



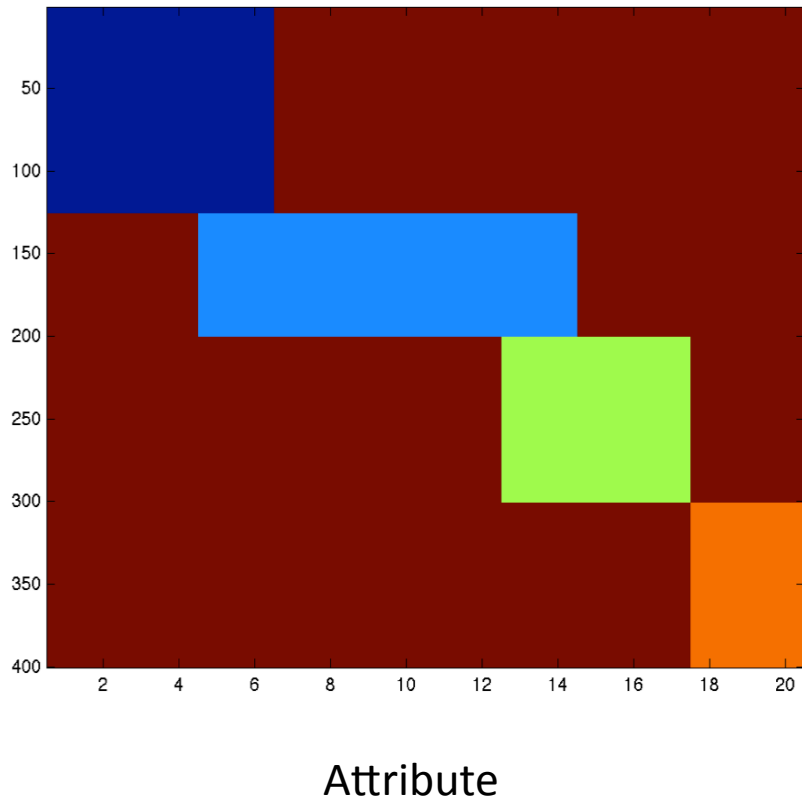
Phase 4:
Replacing a SCC (first SCC)



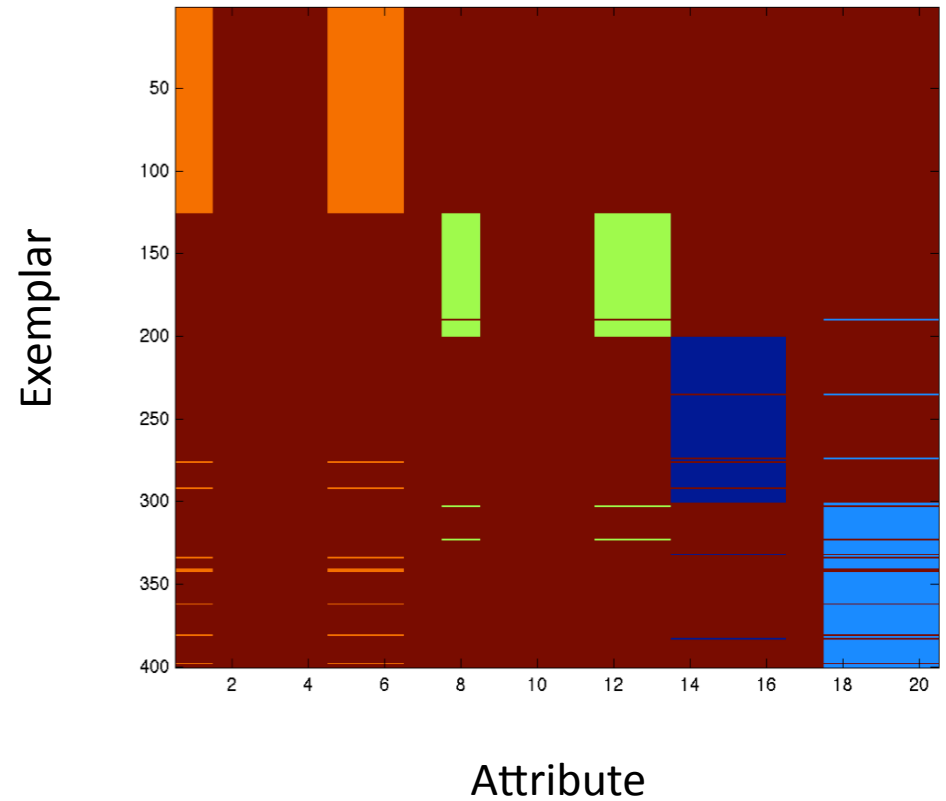
Evolutionary Subspace Clustering (ESC) 3: ‘Symbiont level’ variation operators



ESC Empirical Evaluation: Simple 2D data set



Original 20 Dimensional Dataset



ESC Candidate solution

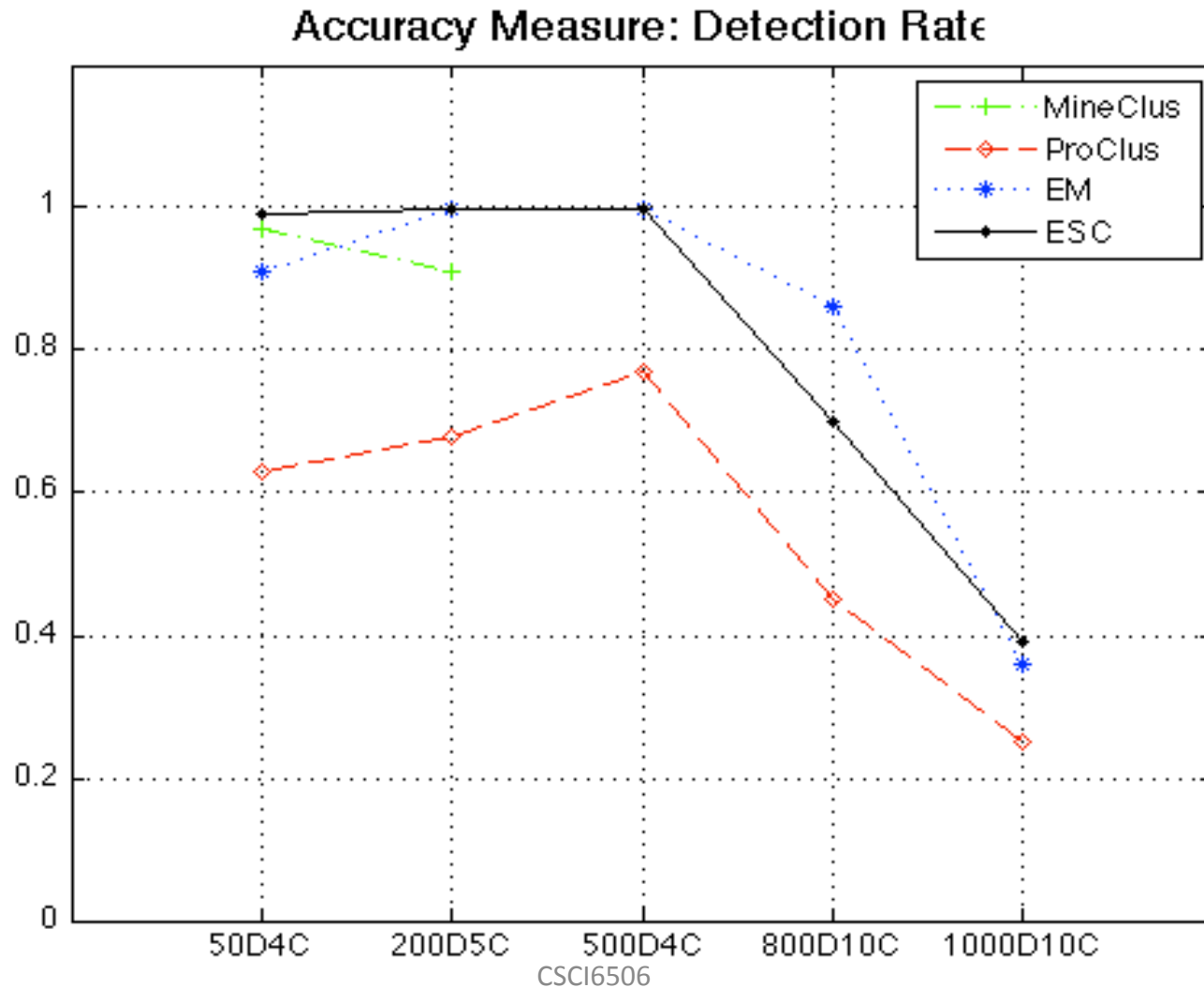
Alternative Clustering Algorithms

- Expectation Maximization (EM)
 - No capability to reduce attribute space
- ProClus
 - Partition based approach
 - Cluster parameters requiring a priori definition:
 - Number of clusters
 - Average cluster subspace dimension
- MineClus
 - Cell based approach
 - Conducts parameter sweeps for greater flexibility
 - Scalability issues

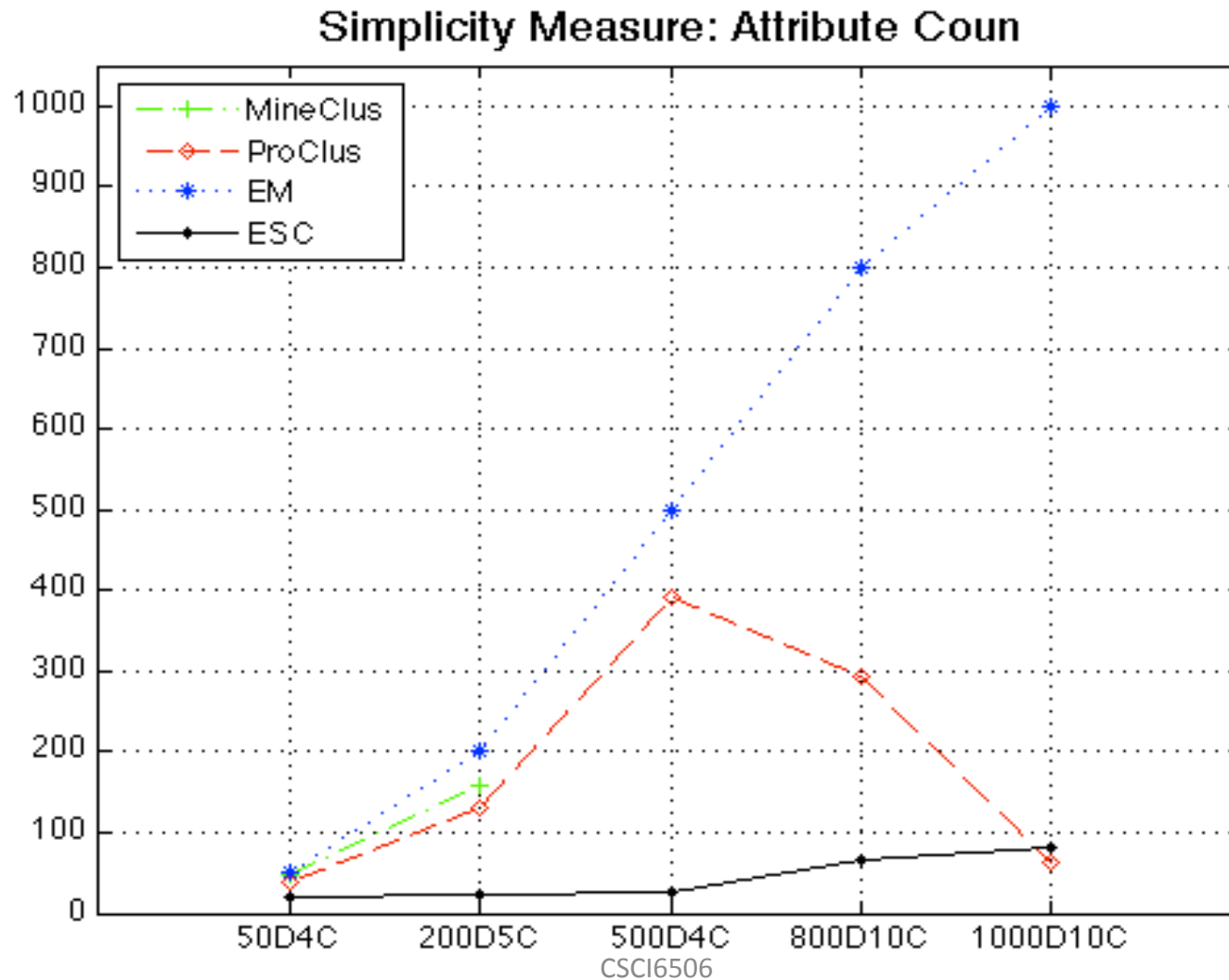
Characterization of Data Sets

Data Set	50D-4C	200D-5C	500D-4C	800D-10C	1000D-10C
#Attributes	50	200	500	800	1 000
#Exemplars	1 289	2 000	1 286	3 814	2 729
#SC	4	5	4	10	10
Cluster shape	Hyper sph./ Gaussian	Hyper rectangular	Hyper elliptical	Mixed	Hyper sph./ Gaussian
Min #Attr	10	10	100	10	10
Max #Attr	10	50	100	100	10
Min %Inst	8	10	13	4	3
Max %Inst	43	25	32	16	20
# Noise Attr	10	50	100	240	900

Benchmarking 'Bake off': Detection Rate (maximize)

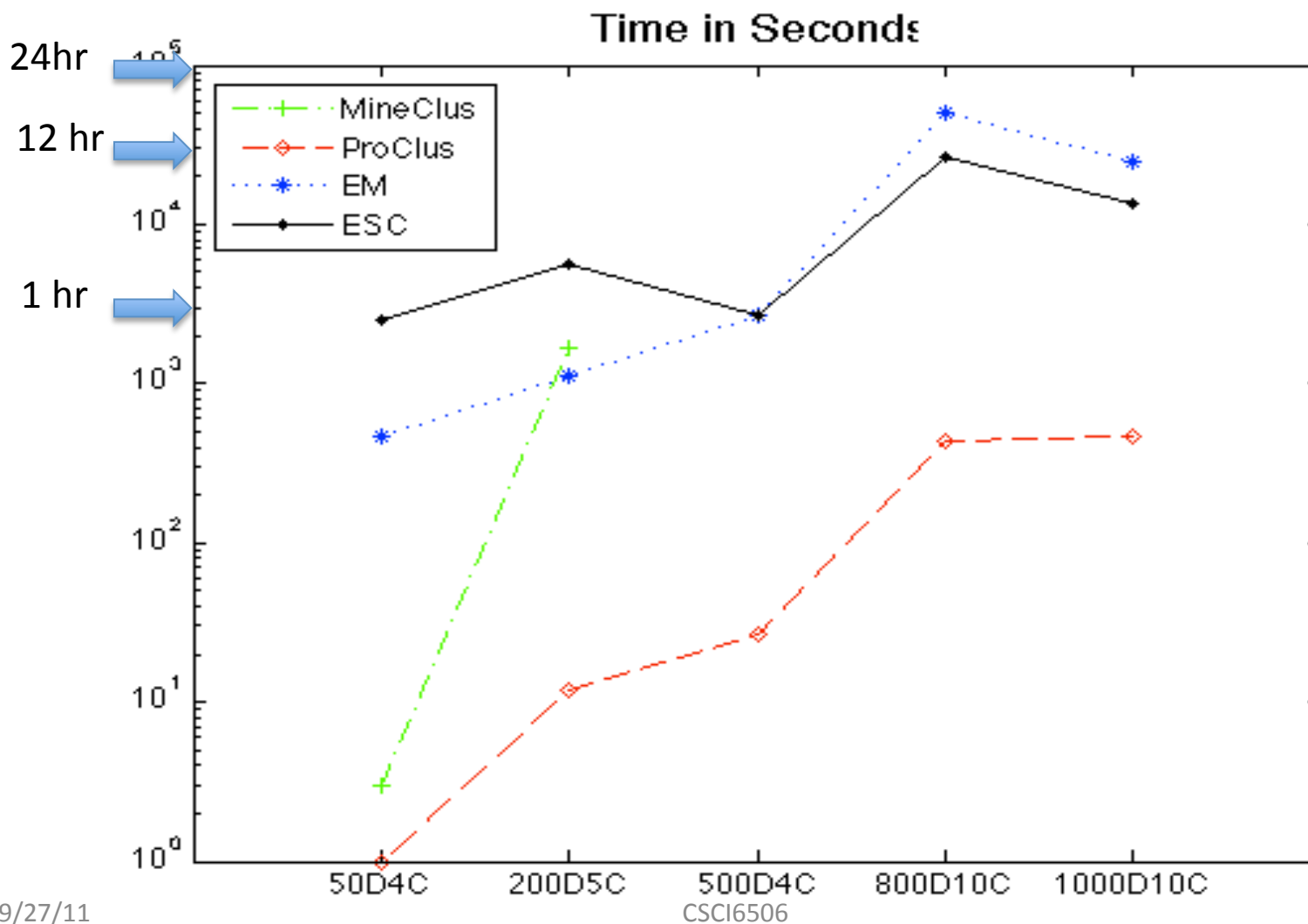


Benchmarking 'Bake off': Cluster attribute Count (minimize)



Benchmarking 'Bake off':

Training time – CPU seconds (minimize)



Summary

- Symbiosis
 - ‘serial’ relation between symbionts (modules) and host (compartment)
 - Symbiont : Candidate Cluster
 - Host : Candidate set of cluster
 - Fitness only evaluated at the host
 - Host conducts combinatorial search over candidate symbionts
 - Independent variation at host and symbiont ‘levels’

References

- B. Kerr and J. Nahum (2011)
 - The Evolution of Restraint in Structured Populations
 - Chapter 7; Colcott and Sterelny (eds) The Major Transitions in Evolution Revisited. MIT Press
- Ali Vahdat (2011)
 - PhD Thesis Proposal, Faculty of Computer Science, Dalhousie University.