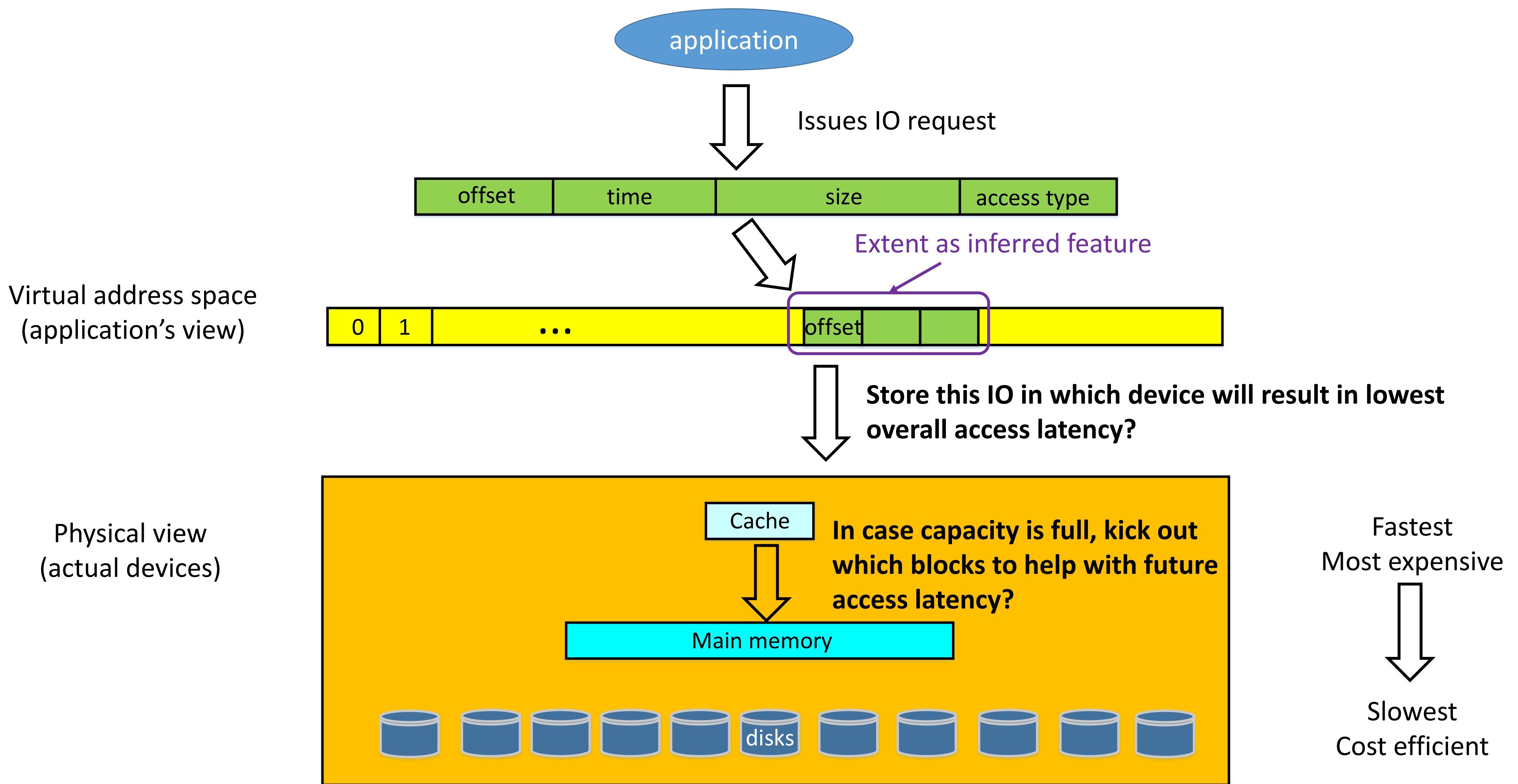
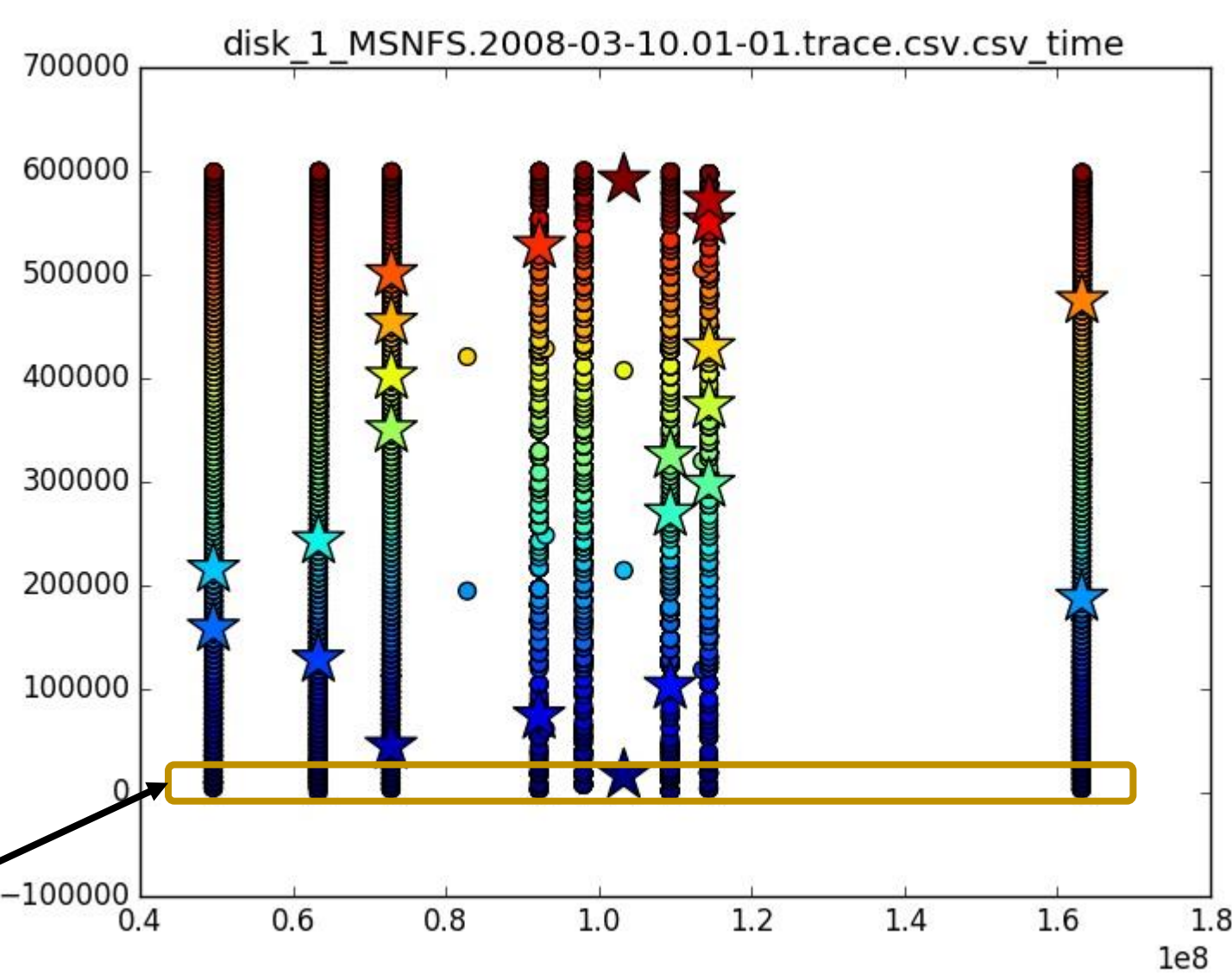


Intelligent Storage System With Machine Learning



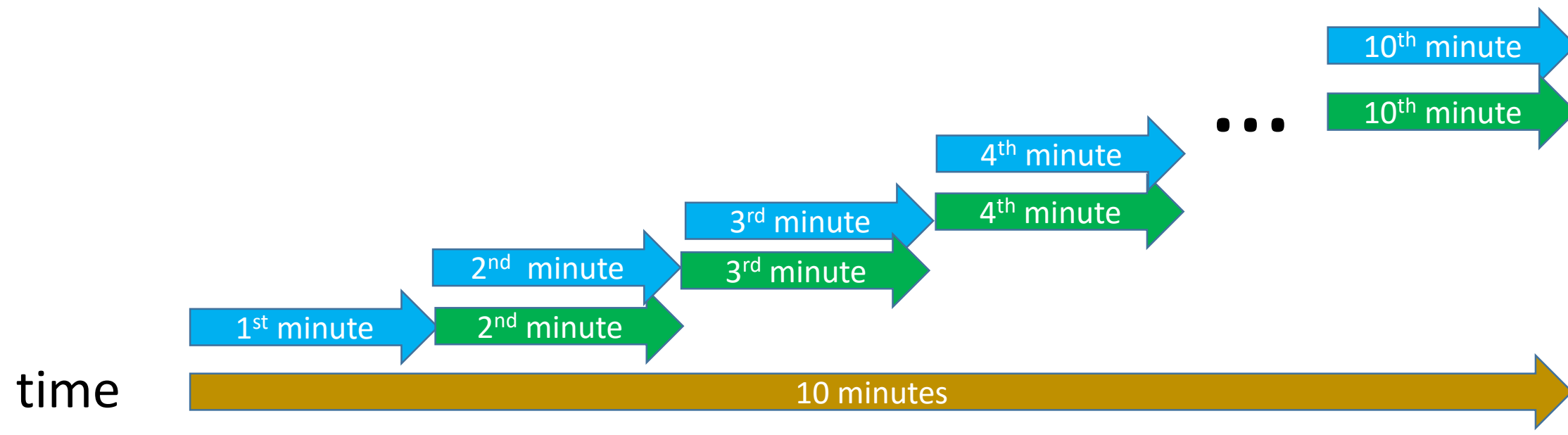
Use ML clustering algorithm to predict the optimal physical storage location and eviction policy!
 Each IO is an example. Entire trace file captured over a period of time is our data set.

Clustering pass 1:
 Entire data set in one run
 Feature 1 is offset (x-axis)
 Feature 2 is time (y-axis)

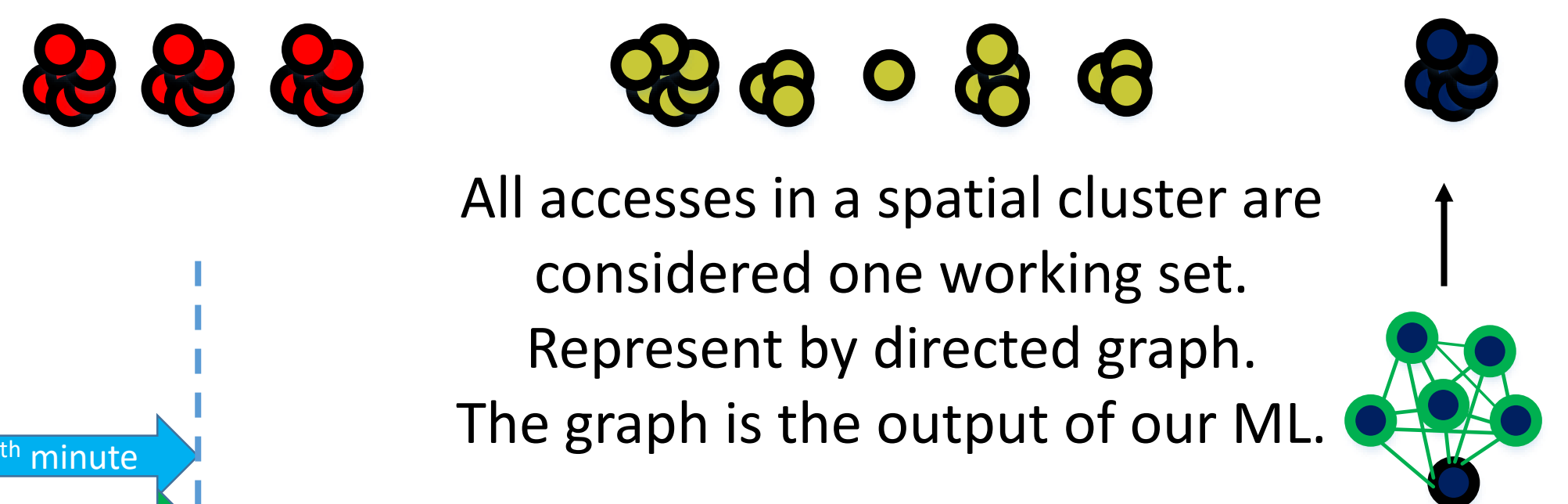
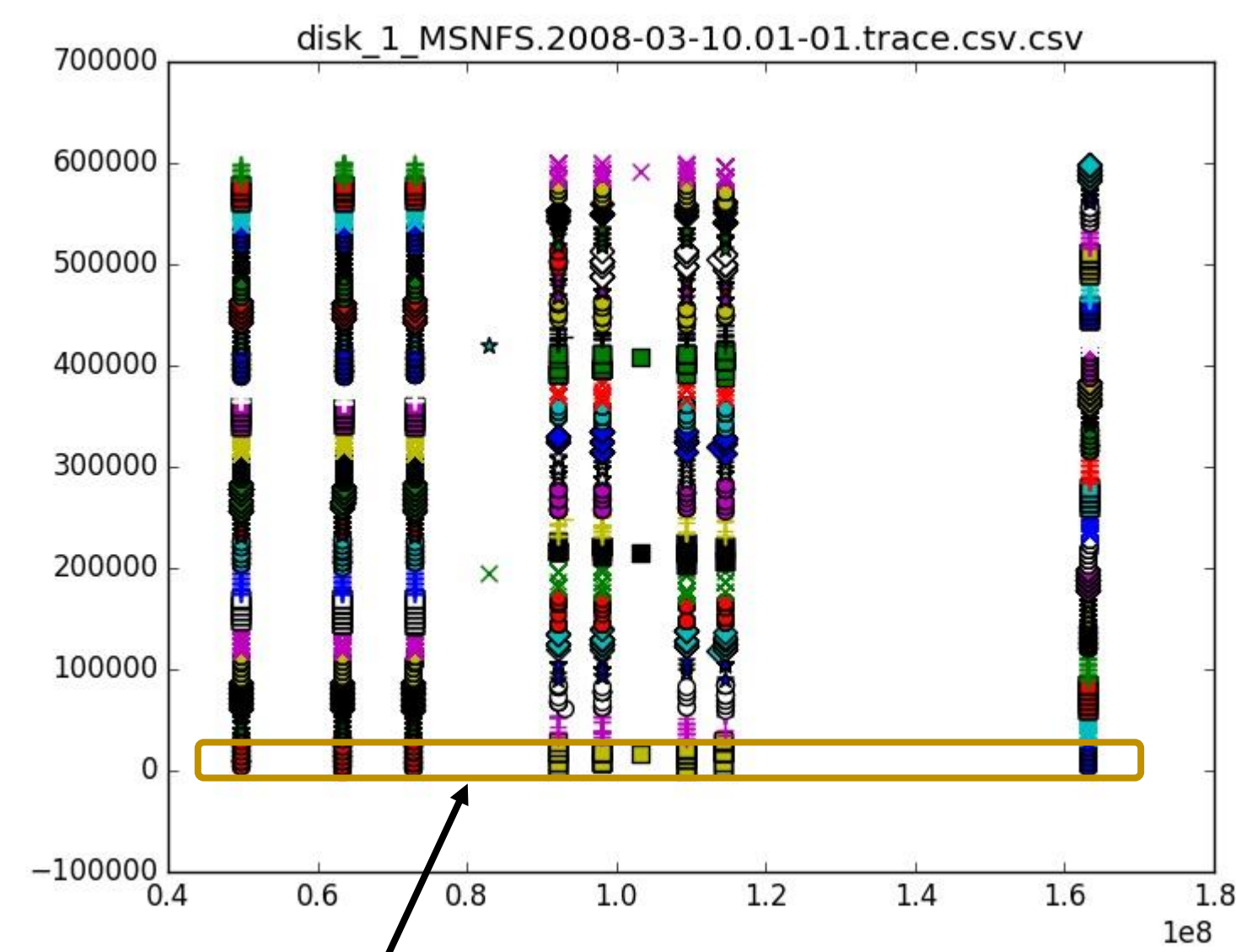


Normalize biased towards time domain.
 Result in horizontal clusters.
 Star indicates each time slice's centroid

Overlap **training** (building working sets for future time slice) and **testing** (apply working sets from previous training)



Clustering pass 2:
 One run per time slice to find spatial locality
 Feature 1: offset
 Feature 2: extent



Simulate 2 tier system: 1 mega byte upper tier; unlimited lower tier Data: 10 minute trace from Microsoft's storage file server (disk_1_MS NFS.2008-03-10.03-03.trace.csv)	
Traditional LRU Most recent IO goes into upper tier If need to free capacity, evict least recently accessed data.	ML approach If an IO is in a working set, fill cache with entire working set. Clear first if needed. If no working set associated with IO, same behavior as LRU.
Hit rate: 67%	Hit rate: 83%

Additional parameters to tune for improving ML result: tier sizes, ML eviction policy