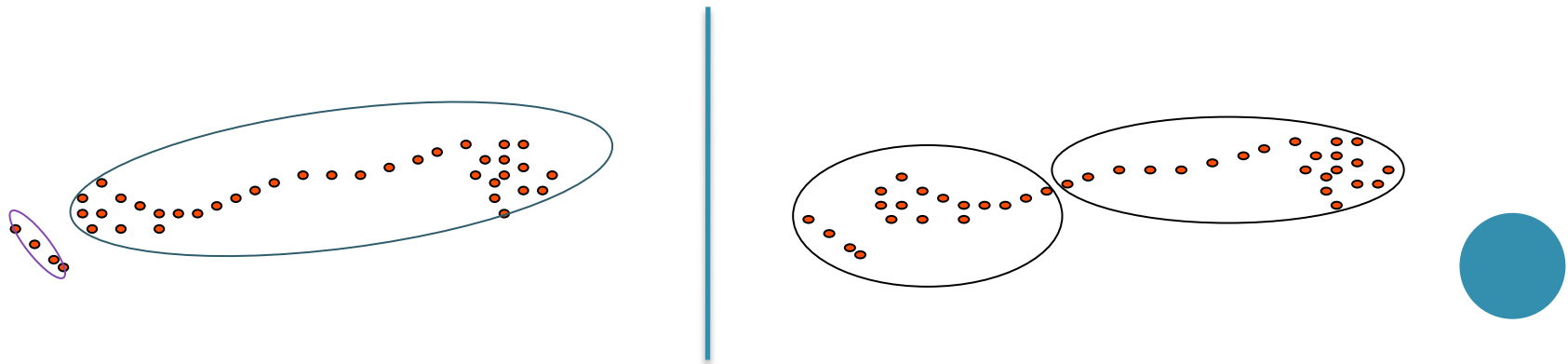


Pranjal Awasthi, Carnegie Mellon University
Reza Bosagh Zadeh, Stanford University

SUPERVISED CLUSTERING

- Clustering is usually unsupervised
- We have limited supervision by way of limited interaction with a teacher
- Remove subjective ambiguities according to teacher:



THE MODEL [BALCAN, BLUM'08]

- Limited interaction with teacher
- Only query allowed: “Here’s what I think the clustering should be”
- Teacher responds with one of:
 - Split this cluster: c
 - Merge these two clusters: c_1 and c_2
- How many queries can we get away with in the worst case?



MAIN RESULTS

- Previous query bound of $O(k^3 \log |C|)$ known for any concept class C .
- We improve the bound to $O(k \log |C|)$.
- Give algorithms for clustering geometric concept classes.
- Present noisy versions of model and give query bounds.
- What if we knew about separation properties of the dataset?



DATASET SEPARATION

- Worst case number of queries under some “separation” properties:

Property	Query Complexity
Threshold Separation	$O(\log(m))$
Strict Separation	$O(k)$
γ -margin Separation	$O\left(\left(\frac{\sqrt{d}}{\gamma}\right)^d - k\right)$

- The better separated the dataset, the fewer queries required
 - Lots of open problems!
- 