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:

<table>
<thead>
<tr>
<th>Property</th>
<th>Query Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold Separation</td>
<td>$O(\log(m))$</td>
</tr>
<tr>
<td>Strict Separation</td>
<td>$O(k)$</td>
</tr>
<tr>
<td>$\gamma$-margin Separation</td>
<td>$O((\frac{\sqrt{d}}{\gamma})^d - k)$</td>
</tr>
</tbody>
</table>

- The better separated the dataset, the fewer queries required

- Lots of open problems!