CONCLUSION

record boundaries
vs.
automatic segmentation
SIMILARITIES

- unsupervised approach
  - no training, no user-input
  - can work automatically

- depend on special page / site structure
  - which ones depends on approach
  - let’s look at the differences
DIFFERENCES

record boundary  automatic segmentation
- plain lists          - list / details page
- single page          - multiple page, generate template
- 5 independent heuristics combined
- only find boundary   - two different heuristics
- assumes about record structure
## AT A GLANCE

### Comparison of the Segmentation Algorithms:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>unsupervised</th>
<th>domain-independent</th>
<th>needs list-and detail-pages</th>
<th>infers Columns</th>
<th>uses tag-tree</th>
<th>allows inconsistencies in content</th>
<th>uses template finding</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 Heuristics (#5)</td>
<td>X</td>
<td>1,2,3,4</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSP (#6)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>HMM (#6)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

- both are useful and can coexist
- hardly overlap
- choice depends on page / site structure
THANK YOU!

...have fun with the next team.