P3C+MR
Subspace clustering
P3C+MR

Project development

Sprint 1
Understand paper, task management and communication setup.

Sprint 2
Create Java specifications, isolate steps of algorithm, cluster environment setup.

Sprint 3
Implementation, Debugging, Optimization.

Sprint 4

Sprint 5
Agile meetings

- Review sprint
- Start sprint
- Execute tasks of sprint
- Create Tasks, estimate them and assign to team members
Challenges

**Time Constraints**
- Challenging mathematical descriptions and complex algorithmic design
- Durations of implementation steps were underestimated

**Implementation**
- Complicated scaling due to missing code optimization
- Flink on cluster

**Data file availability**
- *Hadoop* was not used
  - Solution with script and *scp*
• State of the art clustering algorithm P3C
• Work with programming model MapReduce
• Implementation in Big Data framework Flink
• Refine Java skills
• Work in agile project management environment
P₃C+MR

Algorithm and visualization
P3C+MR

Comparison to ELKI 1/2
<table>
<thead>
<tr>
<th>Flinke Apachen P3C+MR</th>
<th>ELKI P3C</th>
</tr>
</thead>
<tbody>
<tr>
<td>attributes / dimensions</td>
<td>3</td>
</tr>
<tr>
<td>points</td>
<td>1,595</td>
</tr>
<tr>
<td>cluster-cores</td>
<td>4</td>
</tr>
<tr>
<td>bin-size</td>
<td>0.090</td>
</tr>
<tr>
<td>outlier</td>
<td>385</td>
</tr>
<tr>
<td>normalized Euclidian distance average</td>
<td>0.402</td>
</tr>
</tbody>
</table>
Final products

- Working algorithm P3C+MR
- Bash script for jar execution on cluster
- Full visualization/animation on 3-dimensional example
- Customizable visualization script
- Wiki with algorithm specification
- Javadoc html documentation
Future prospects

• Optimization of code (scaling and efficiency)
• GitHub repository for public sharing
• Change algorithm for streaming data
P3C+MR

Demonstration

Flink
A Framework for Clustering Uncertain Data:

R subspace clustering:
https://cran.r-project.org/web/packages/subspace/subspace.pdf

P3C:

P3C: A Robust Projected Clustering Algorithm: