Practical course Big Data Science

Flinke Apachen

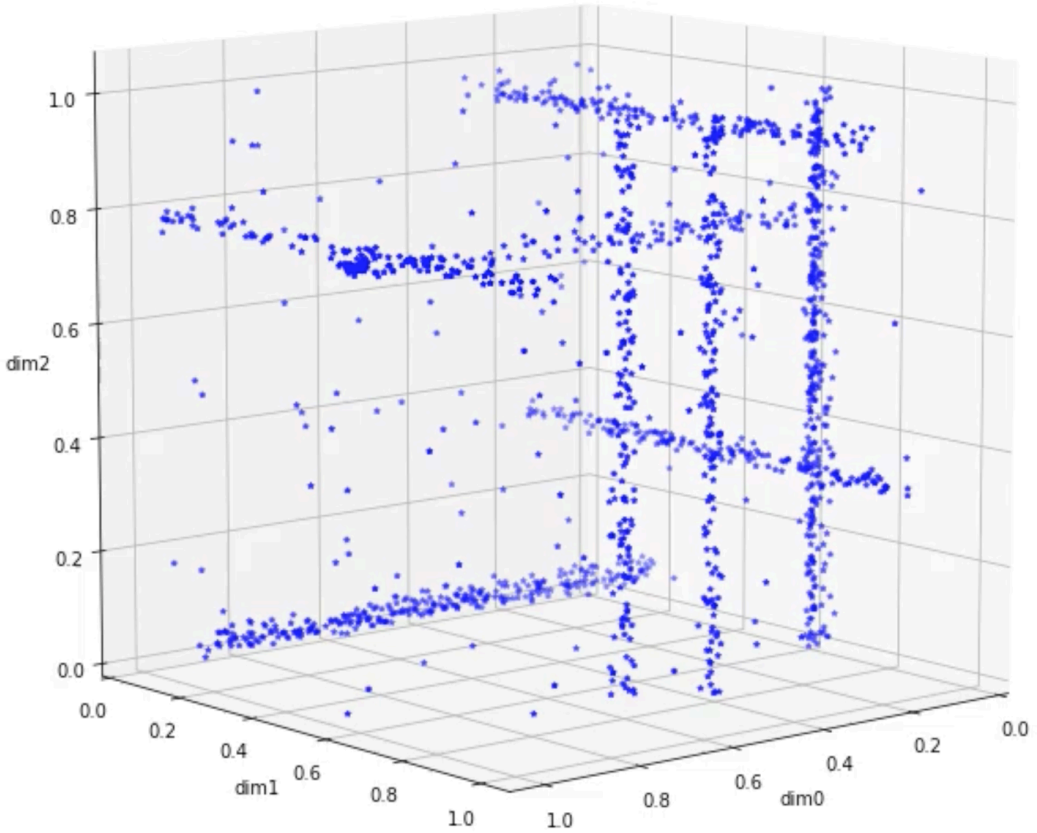
P3C+MR ALGORITHM

©Sebastian Rühl - Felix Mathy - Julian Jorczik - Maximilian Höb - Huimin An

19. July 2017

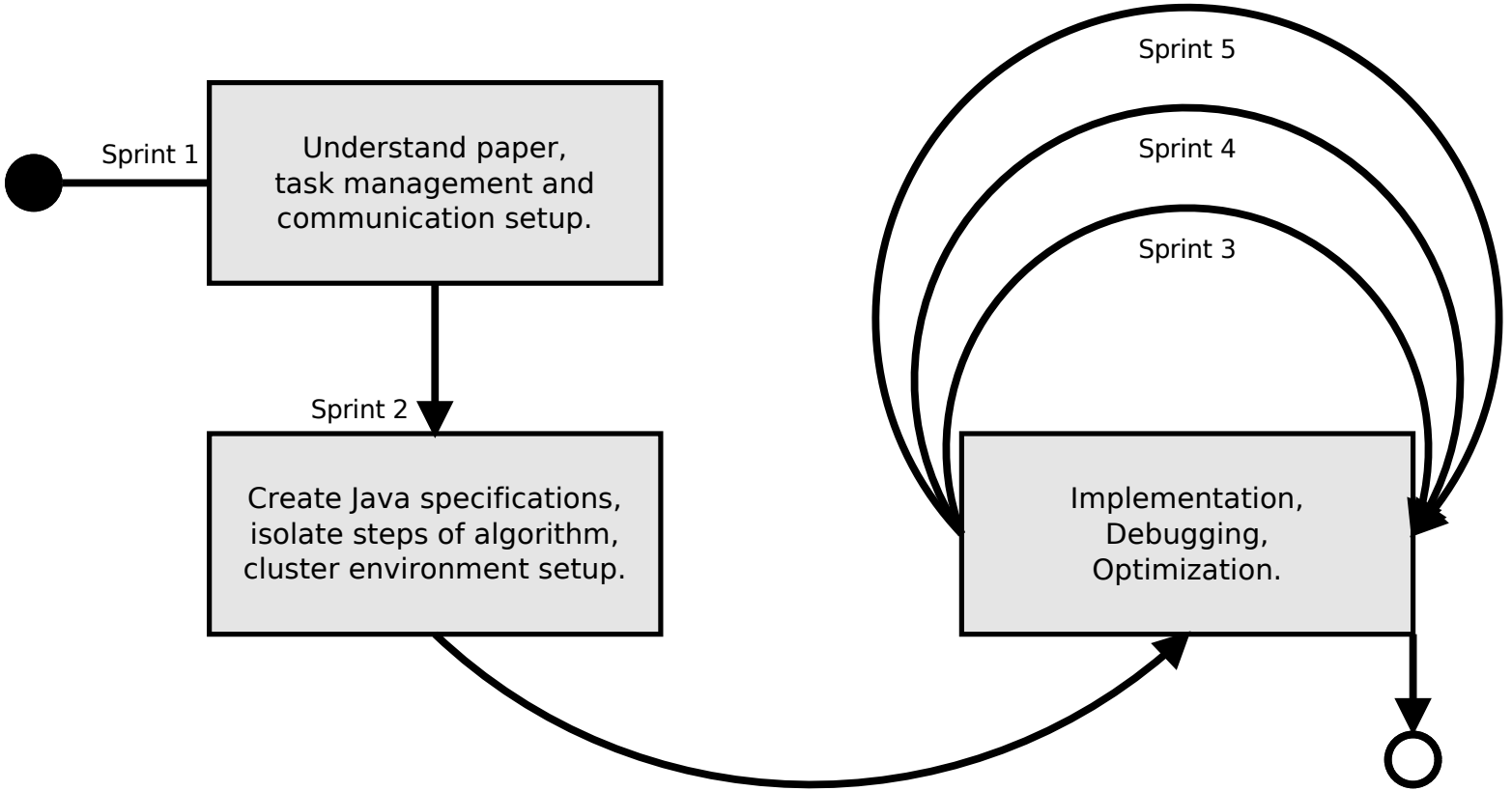
P3C+MR

Subspace clustering

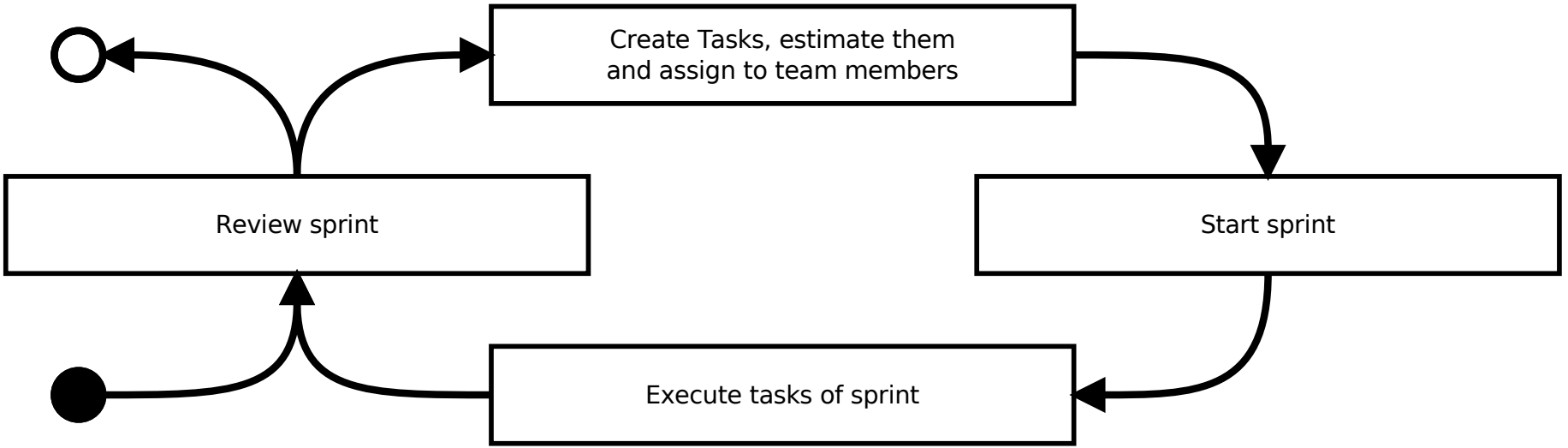


P3C+MR

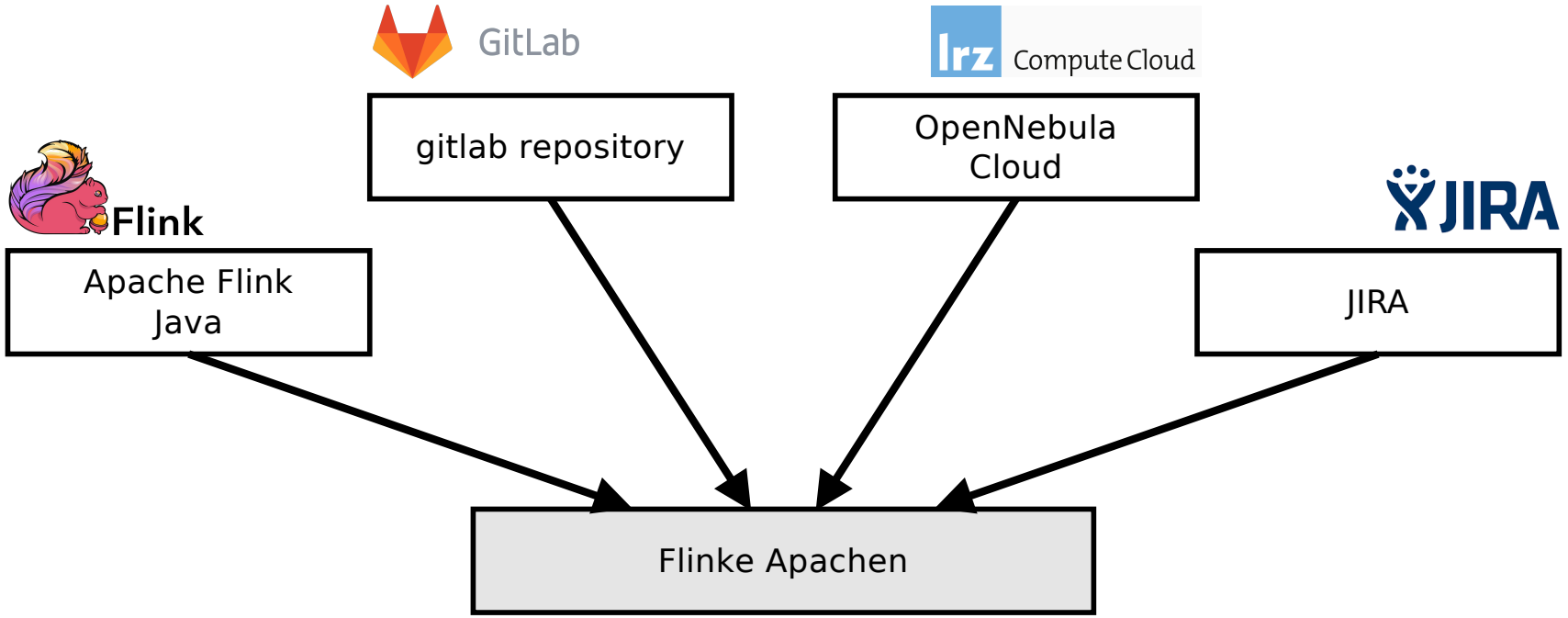
Project development



Agile meetings



Environment



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

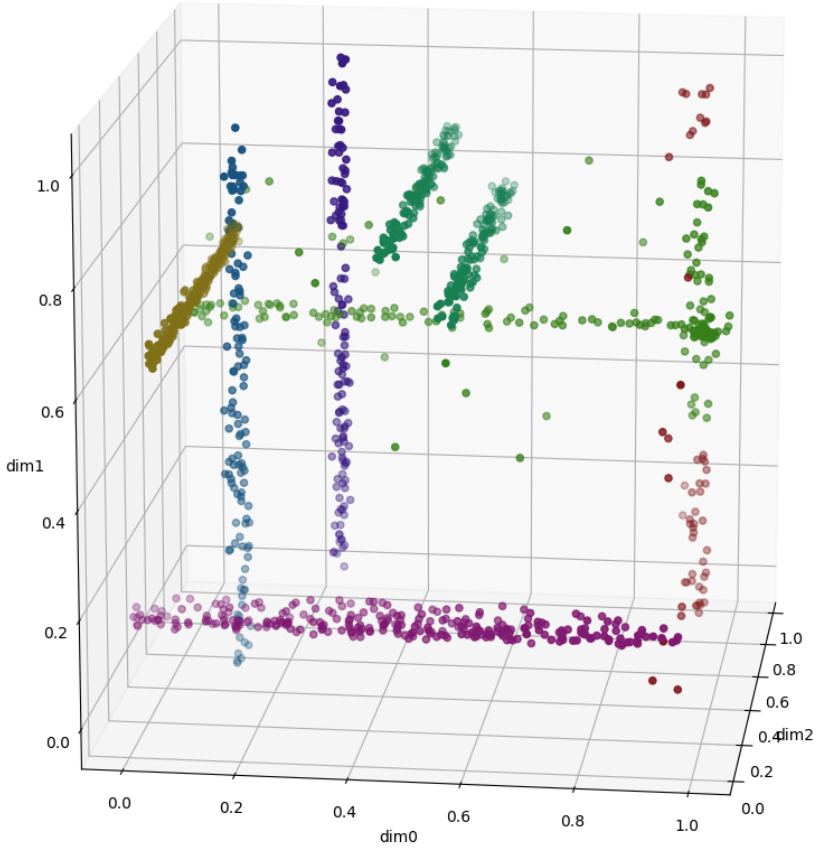
Learning achievements

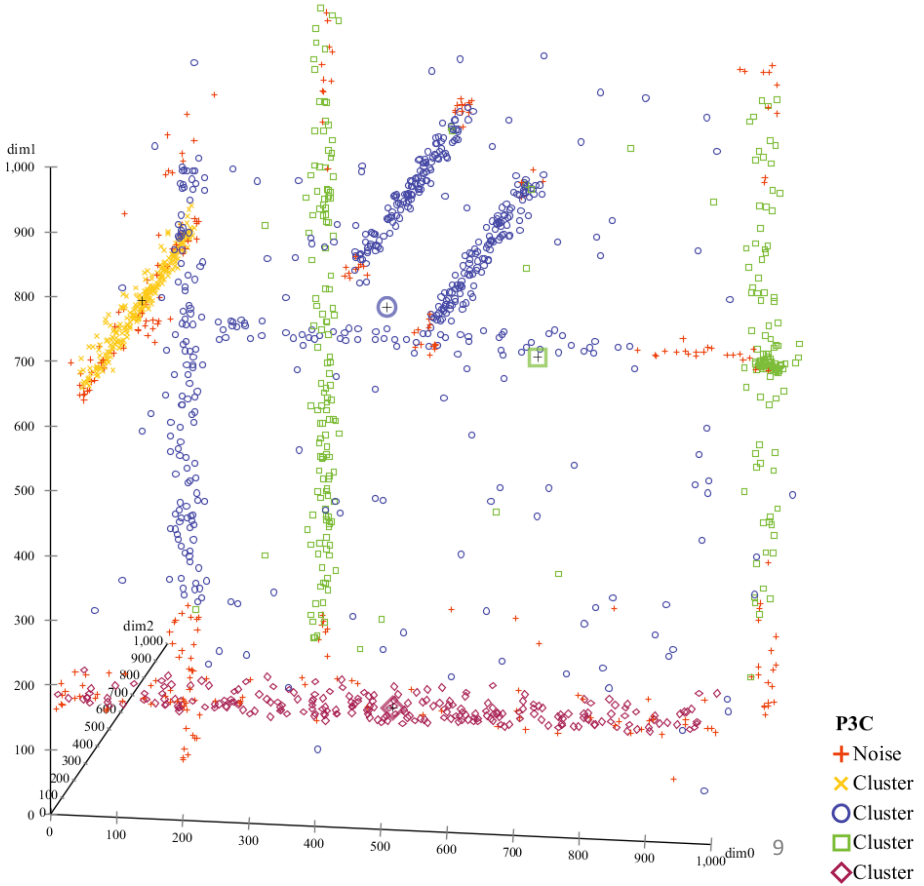
- State of the art clustering algorithm P3C
- Work with programming model MapReduce
- Implementation in Big Data framework Flink
- Refine Java skills
- Work in agile project management environment

Algorithm and visualization

P3C+MR

Comparison to ELKI 1/2





Comparison to ELKI 2/2

Flinke Apachen P3C+MR		ELKI P3C
3	attributes / dimensions	3
1,595	points	1,595
7	cluster-cores	4
0.085	bin-size	0.090
268	outlier	385
0.478	normalized Euclidian distance average	0.402

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

Demonstration

Flink

Sources

A Framework for Clustering Uncertain Data:
http://www.vldb.org/pvldb/vol8/p1976-schubert.pdf

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

P3C:
https://openproceedings.org/2014/conf/edbt/FriesWS14.pdf

P3C: A Robust Projected Clustering Algorithm:
http://ieeexplore.ieee.org/document/4053068/

Questions

Flink