Brandenburgische
Technische Universität
Cottbus - Senftenberg

08. März 2017

EFM-DBSCAN

EIN BAUMBASIERTER CLUSTERING-ALGORITHMUS UNTER AUSNUTZUNG ERWEITERTER LEADER-UMGEBUNGEN

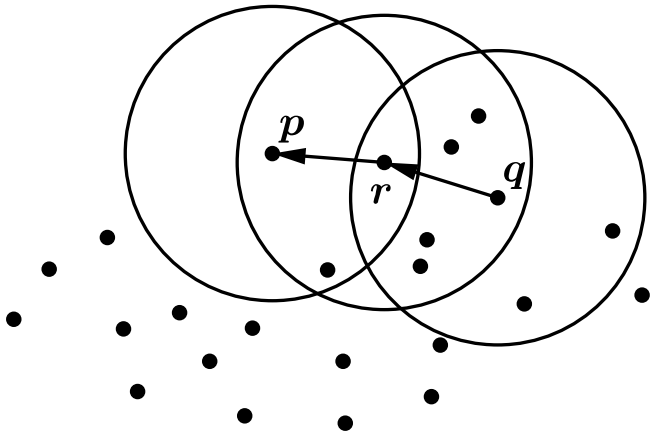
Philipp Egert

Fachgebiet Datenbank- und Informationssysteme

Motivation

Brandenburgische Technische Universität Cottbus - Senftenberg

- DBSCAN als Vorreiter dichte-basierter Clustering-Algorithmen zur Erkennung beliebig geformter Cluster und Rauschen
  - eleganter Algorithmus mit den Dichteparametern ε und $minPts$



q und r sind Kernobjekte

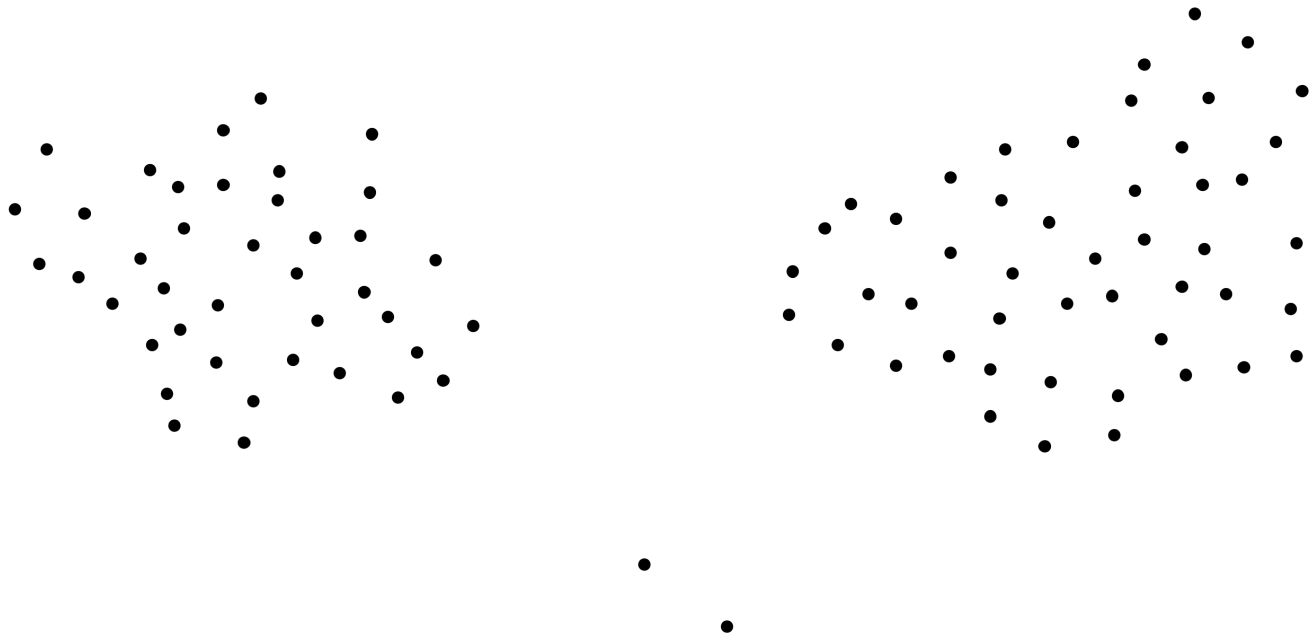
r ist direkt dichte-erreichbar von q

p ist dichte-erreichbar von q

- Problem: Für jedes Objekt die ε-Umgebung berechnen
- Ansätze zur Beschleunigung (exakt/approximativ + euklidisch/metrisch)
  - Vorarbeit: FM-DBSCAN (exakt + metrisch)
- Ziel: Weiterentwicklung von FM-DBSCAN zur Performanzsteigerung

Clustering mit FM-DBSCAN

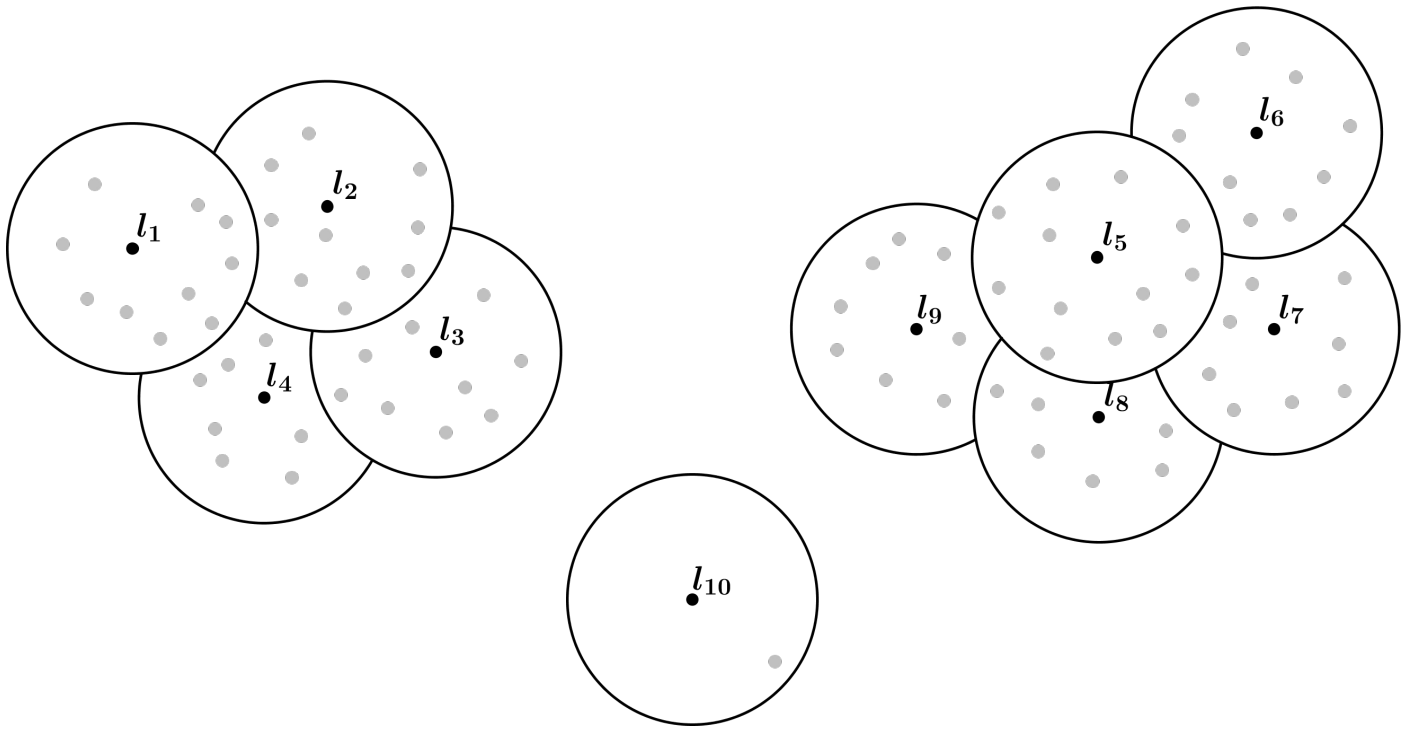
1. Partitionierung der Objekte in Leader-Umgebungen (Leader-Partition)
2. Clustering anhand der Leader-Partition (Dreiecksungleichung nutzen)



Clustering mit FM-DBSCAN

Brandenburgische Technische Universität Cottbus - Senftenberg

1. Partitionierung der Objekte in Leader-Umgebungen (Leader-Partition)
2. Clustering anhand der Leader-Partition (Dreiecksungleichung nutzen)

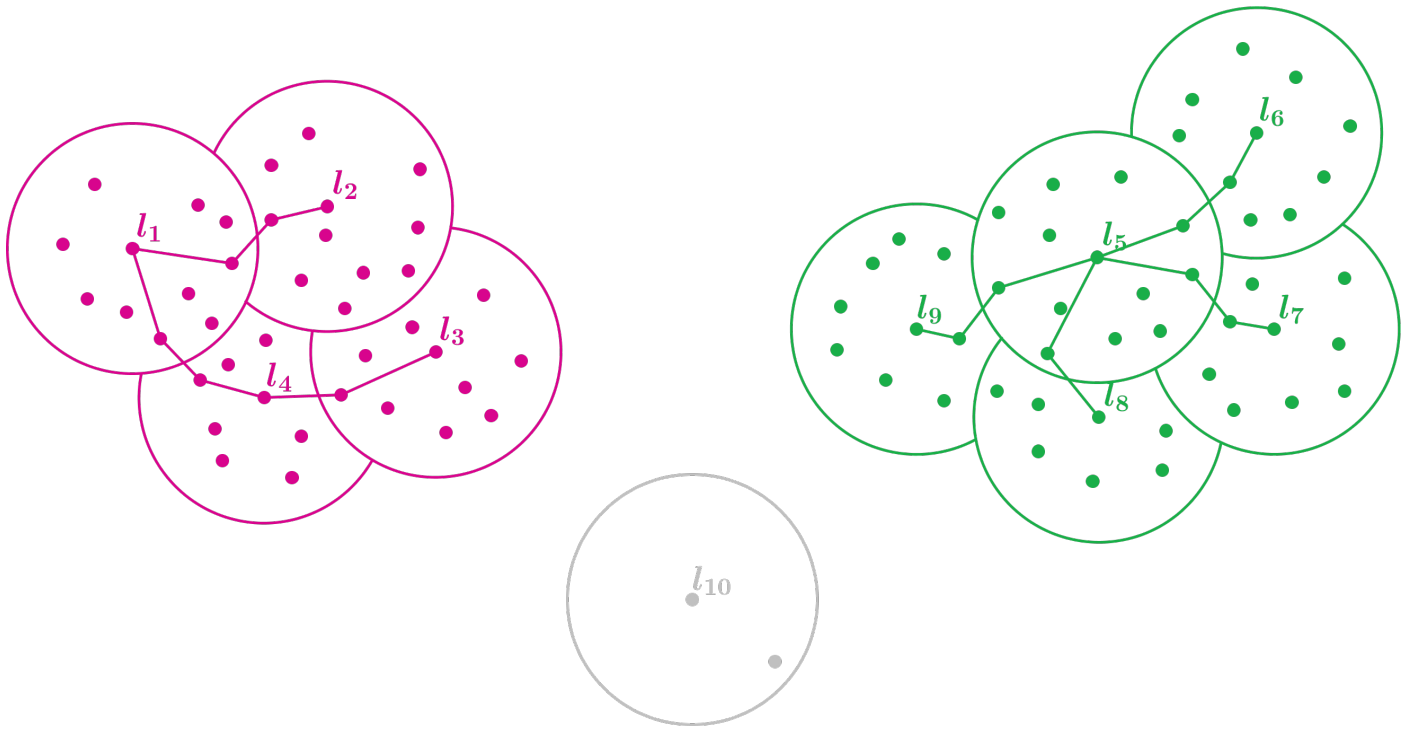


Phase 1: Berechnung der Leader-Partition

Clustering mit FM-DBSCAN

Brandenburgische Technische Universität Cottbus - Senftenberg

1. Partitionierung der Objekte in Leader-Umgebungen (Leader-Partition)
2. Clustering anhand der Leader-Partition (Dreiecksungleichung nutzen)

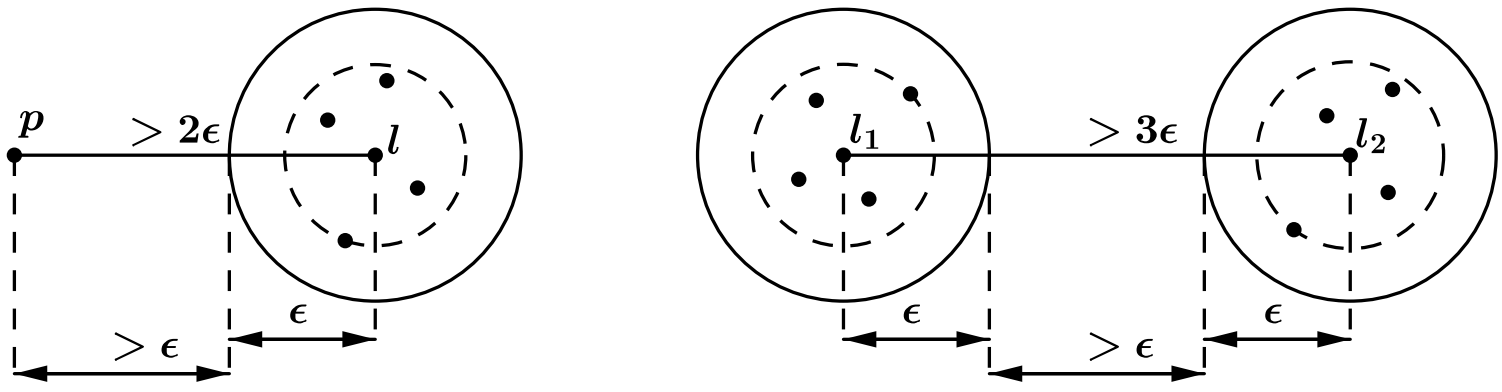


Phase 2: Clustering anhand der Leader-Partition

Optimierungsmöglichkeiten von FM-DBSCAN

Brandenburgische Technische Universität Cottbus - Senftenberg

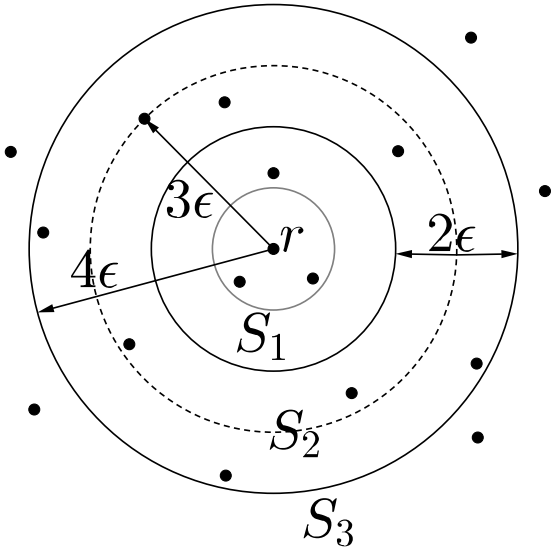
- lineare Suche bei der listenbasierten Erzeugung der Leader-Partition
- immer ε als Größe einer Leader-Umgebung verwendet
- die Lage der Objekte einer Leader-Umgebung zum Leader wird nicht genutzt (Distanzen)

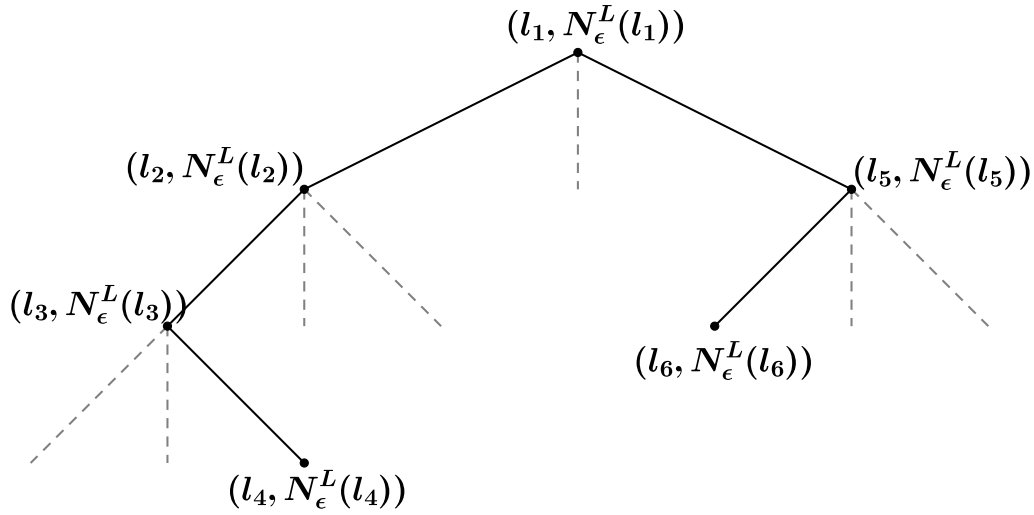


EFM-DBSCAN: Baumbasierte Partitionierung

Brandenburgische Technische Universität Cottbus - Senftenberg

- basiert auf *Excluded Middle Partitioning*
- Baum wird iterativ aufgebaut
- keine zusätzlichen Parameter notwendig

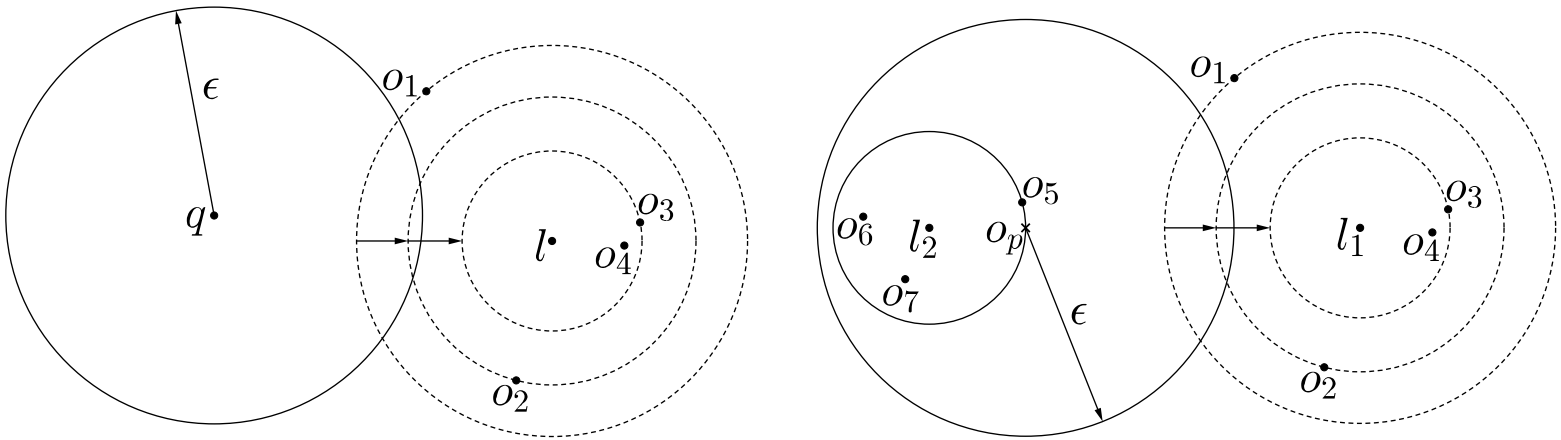




EFM-DBSCAN: Erweiterung der Leader-Umgebung

Brandenburgische Technische Universität Cottbus - Senftenberg

- Erweiterung der Leader-Umgebung, um die Distanzen der Objekte zu ihrem Leader
- schärfere Ausschlussbedingungen
- sortiertes Abarbeiten der Leader-Umgebungen für eine effektivere Ausschlussstrategie (schrumpfende Leader-Umgebungen)



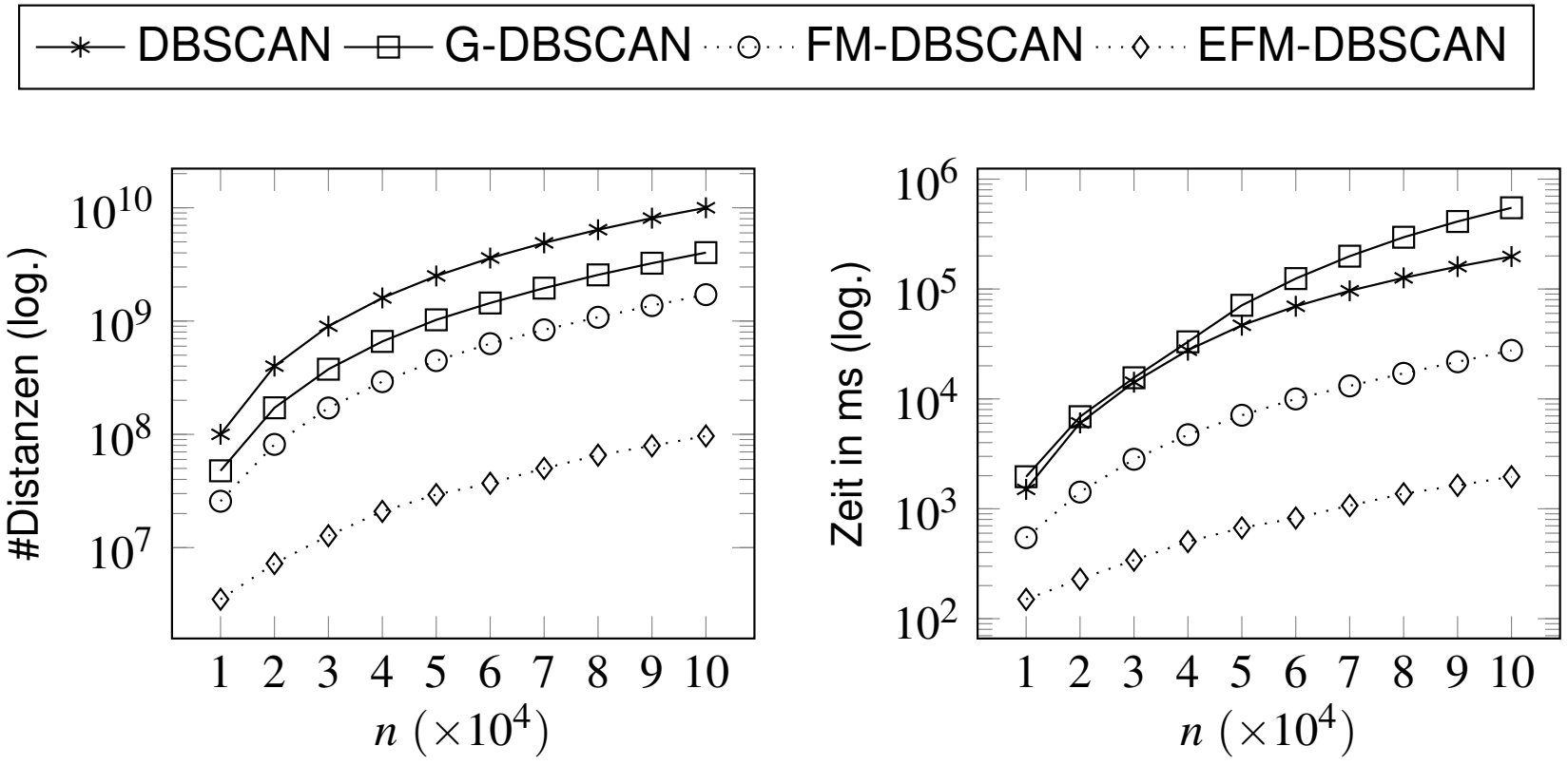
Testumgebung

Brandenburgische Technische Universität Cottbus - Senftenberg

- PC mit Intel®Core^{TM} i7-2600K Prozessor, 16 GB RAM und Debian 8
- Vergleichsverfahren: DBSCAN, FM-DBSCAN und G-DBSCAN
- synthetische (Gaussian Mixture Clusters) und reale Datenkollektion (Caltech256)
  $\Longrightarrow$ Hier: Gaussian Mixture Clusters (synthetisch, 12-dimensional, euklidische Distanz)

Clustering-Performanz für variierende Datenkollektionsgröße

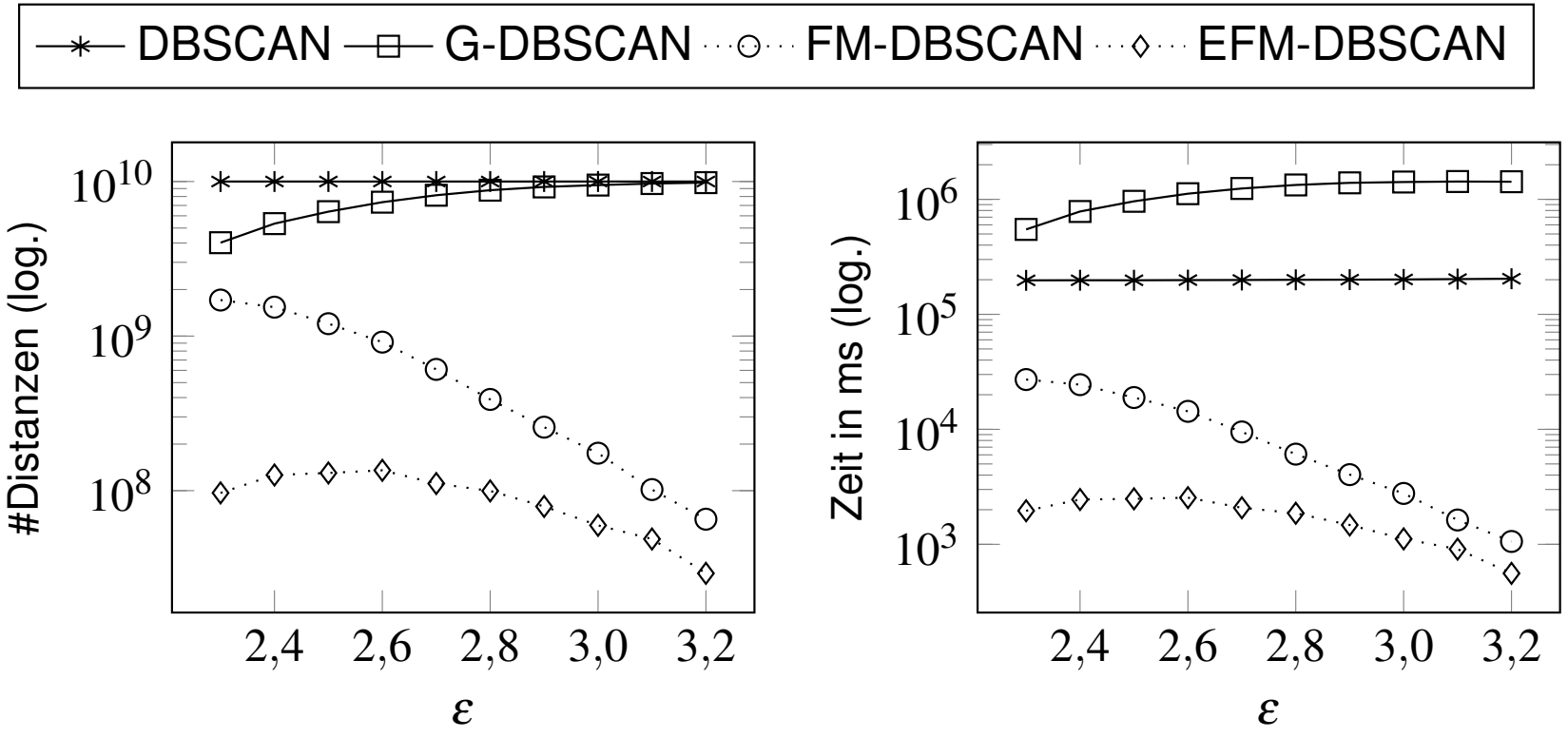
Brandenburgische Technische Universität Cottbus - Senftenberg



$minPts = 5,\ \varepsilon = 2,3$

Clustering-Performanz für variierendes ε

Brandenburgische Technische Universität Cottbus - Senftenberg



$minPts = 5$, $n = 100.000$

Zusammenfassung

- exakter Clustering-Algorithmus für beliebige Distanzfunktionen
- starke Beschleunigung gegenüber FM-DBSCAN (Faktor 13)
- zusätzlicher Sortieraufwand vernachlässigbar gering
- skaliert gut mit der Datenkollektionsgröße und dem Parameter ε

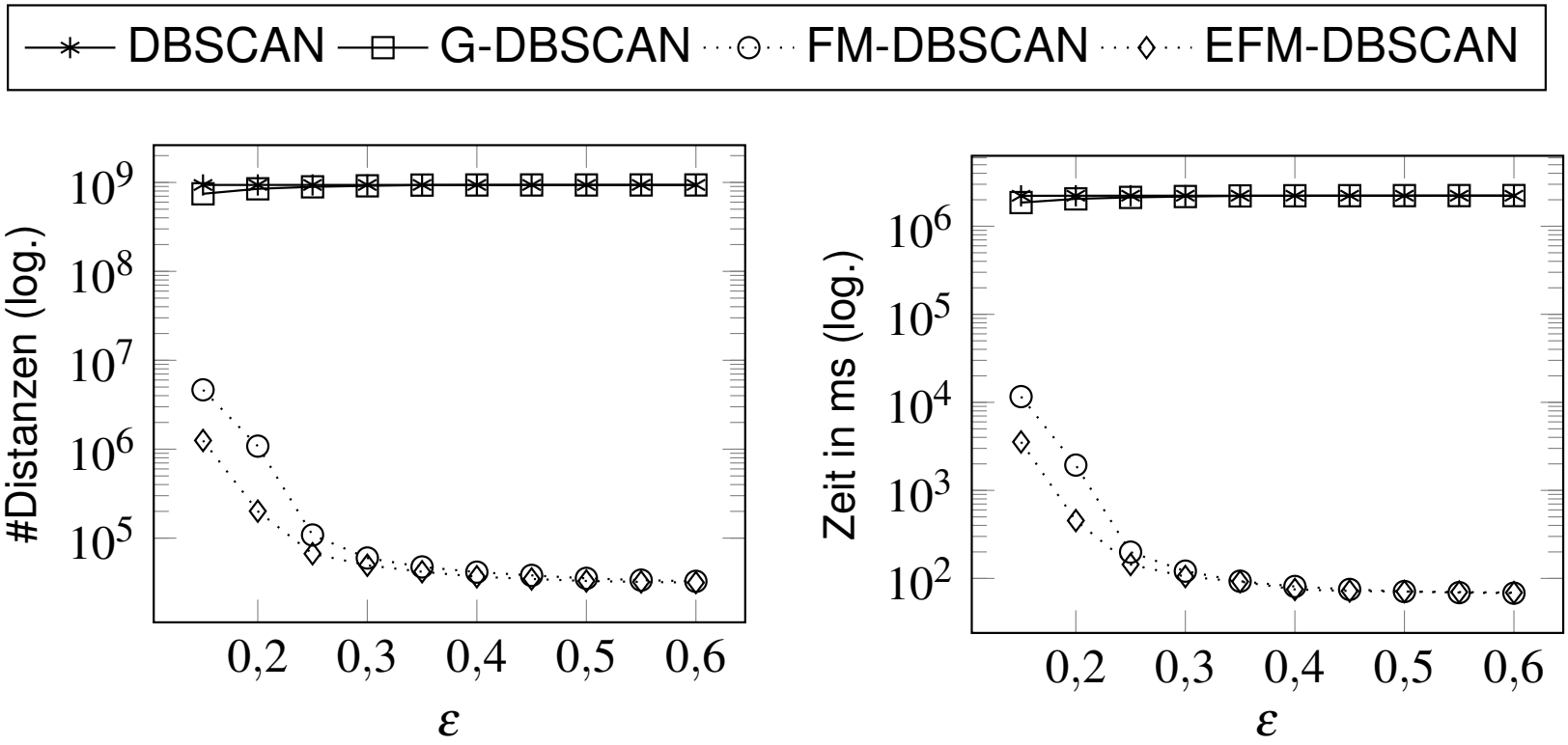
Ausblick

- umfangreiche Evaluation
- Bestimmung optimaler Werte für $minPts$ und ε
- Laufzeitanalyse auf eingeschränkten metrischen Räumen

Anhang

Clustering-Performanz für variierendes ε

Brandenburgische Technische Universität Cottbus - Senftenberg



Caltech256 ($minPts = 2$, $n = 30.607$)