Applications of Clustering for Large-Scale Datasets

Dr.-Ing. Morris Riedel et al.

Research Group Leader, Juelich Supercomputing Centre

Adjunct Associated Professor, University of Iceland

PhD Students Markus Goetz, Christian Bodenstein

Juelich Supercomputing Centre & University of Iceland

RDA Session – Big Data IG, 25th September 2015

Research Data Alliance 6th Plenary, CNAM, Paris

HELMHOLTZ ASSOCIATION

Federated Systems and Data Division

Research Group

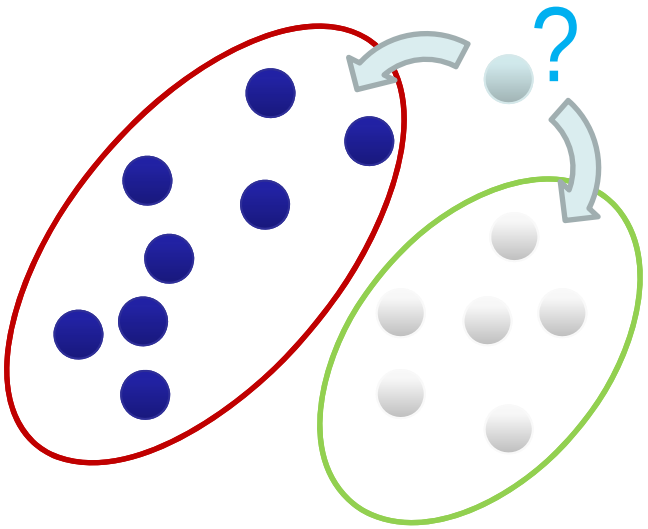
High Productivity Data Processing

JÜLICH FORSCHUNGSZENTRUM

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

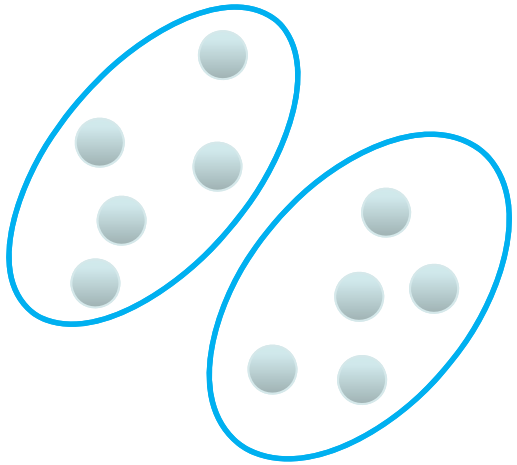
Learning From Data – Clustering Technique Focus

Classification



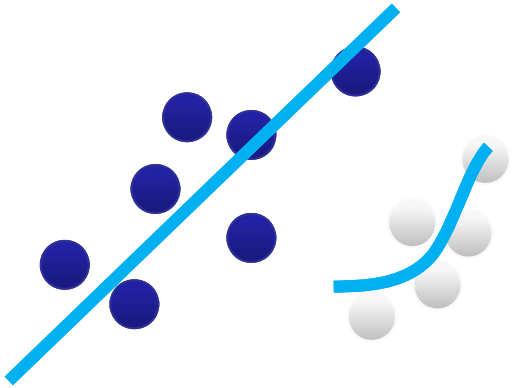
- Groups of data exist
- New data classified to existing groups

Clustering



- No groups of data exist
- Create groups from data close to each other

Regression



- Identify a line with a certain slope describing the data

➢ **Research with applications of classifications were presented before in RDA (e.g. remote sensing)**

Selected Clustering Methods

K-Means Clustering – Centroid based clustering

- Partitions a data set into K distinct clusters (centroids can be artificial)

K-Medoids Clustering – Centroid based clustering (variation)

- Partitions a data set into K distinct clusters (centroids are actual points)

Sequential Agglomerative hierarchic nonoverlapping (SAHN)

- Hiearchical Clustering (create tree-like data structure → 'dendrogram')

Clustering Using Representatives (CURE)

- Select representative points / cluster; as far from one another as possible

Density-based spatial clustering of applications + noise (DBSCAN)

Reasoning: density similiarity measure helpful in our driving applications

- Assumes clusters of similar density or areas of higher density in dataset

Technology Review of Available 'Big Data 'Tools

Technology	Platform Approach	Analysis
HPDBSCAN (authors implementation)	C; MPI; OpenMP	Parallel, hybrid, DBSCAN
Apache Mahout	Java; Hadoop	K-means variants, spectral, no DBSCAN
Apache Spark/MLlib	Java; Spark	Only k-means clustering, No DBSCAN
scikit-learn	Python	No parallelization strategy for DBSCAN
Northwestern University PDSDBSCAN-D	C++; MPI; OpenMP	Parallel DBSCAN

[2] M. Goetz, M. Riedel et al., '' On Parallel and Scalable Classification and Clustering Techniques for Earth Science Datasets, 6th Workshop on Data Mining in Earth System Science, International Conference of Computational Science (ICCS)

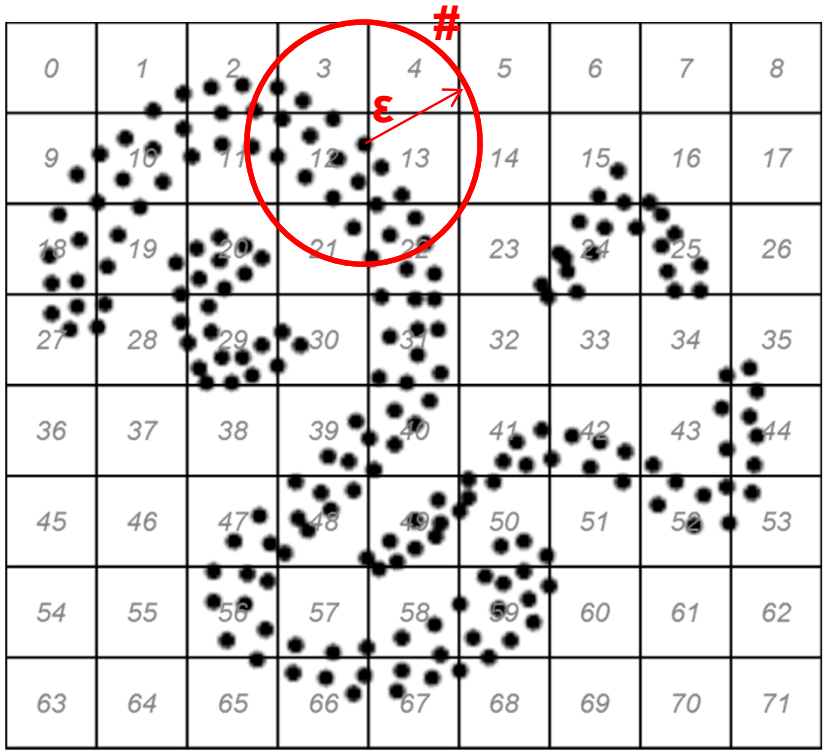
Parallel & Scalable HP-DBSCAN Open Source Tool

Parallelization Strategy

- Smart ‘Big Data‘ Preprocessing into Spatial Cells (‘indexed‘)
- OpenMP standalone
- MPI (+ optional OpenMP hybrid)

Preprocessing Step

- Spatial indexing and redistribution according to the point localities
- Data density based chunking of computations



[1] M.Goetz & C. Bodenstein, HPDBSCAN Tool Download

[2] M. Goetz, M. Riedel et al., 6th Workshop on Data Mining in Earth System Science, ICCS 2015

Computational Optimizations

- Caching of point neighborhood searches
- Cluster merging based on comparisons instead of zone reclustering

Clustering Applications – Large Point Clouds

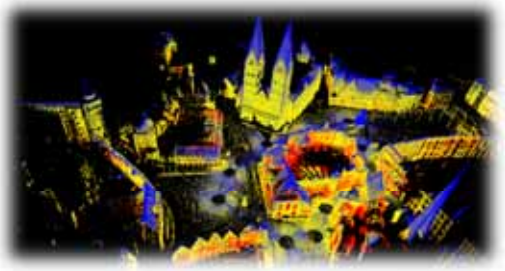
‘Big Data‘: 3D/4D laser scans

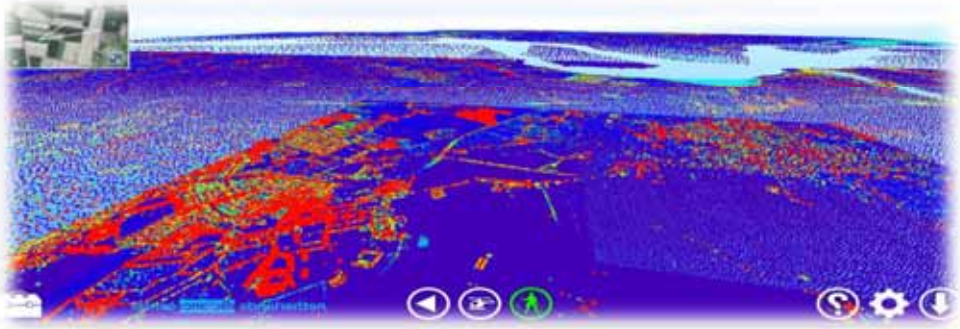
- Captured by robots or drones
- Millions to billion entries
- Inner cities (e.g. Bremen inner city)
- Whole countries (e.g. Netherlands)

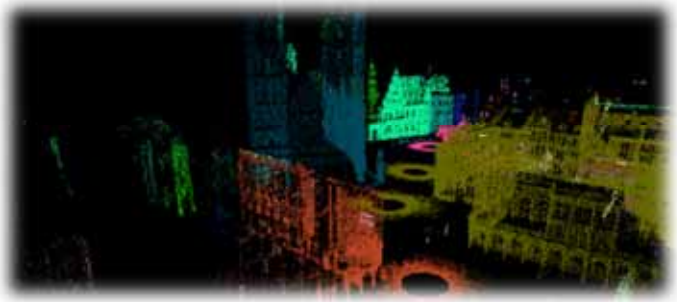
Selected Scientific Cases

- Filter noise to better represent real data
- Grouping of objects (e.g. buildings)

- **Interest? Become H2020 Project Proposal User Advisory Board member → Contact me today**





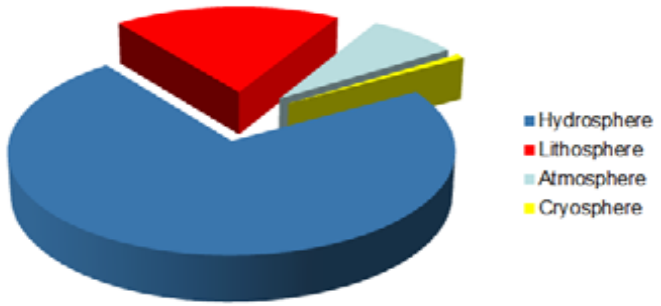


➢ **Research activities in collaboration with the Netherlands e-Science Centre & TU Delft**

Clustering Applications – Many Time Series & Events

Earth Science Data Repository

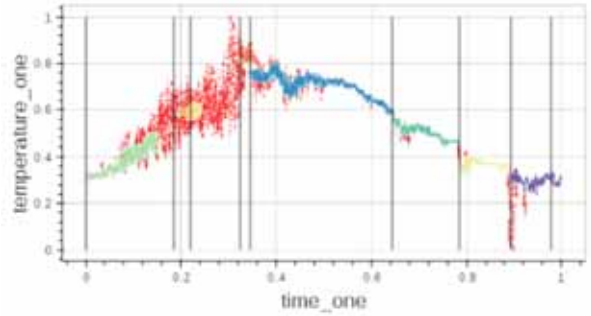
- Time series measurements (e.g. salinity)
- Millions to billions of data items/locations
- Less capacity of experts to analyse data



Total number of data sets 349 871
Data items ~ 7.9 billions

Selected Scientific Case

- Data from Koljöfjords in Sweden (Skagerrak)
- Each measurement small data, but whole ‘big data‘
- Automated water mixing event detection & quality control (e.g. biofouling)
- Verification through domain experts

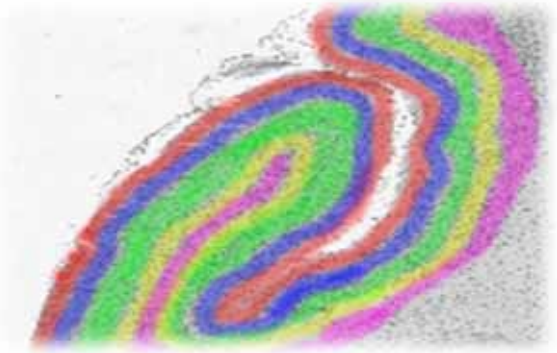


➢ **Research activities in collaboration with MARUM in Bremen and University of Gothenburg**

Clustering Applications – Neuro Science Image Analysis

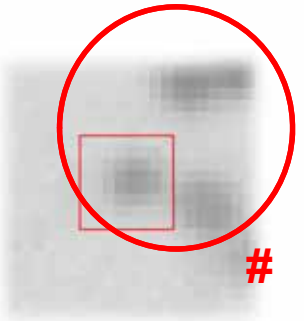
Large Brain Images

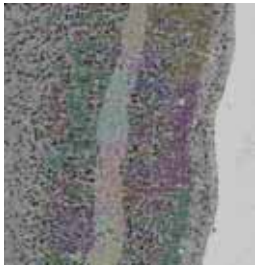
- High resolution scans of post mortem brains
- Rare ‘groundtruth available‘



Selected Scientific Case

- Cell nuclei detection and tissue clustering
- Detect various layers (colored)
- Layers seem to have different density distribution of cells
- Extract cell nuclei into 2D/3D point cloud
- Cluster different brain areas by cell density





➢ **Research activities in collaboration with Institute of Medicine and Neuroscience (T. Dickscheid)**

References

[1] M.Goetz & C. Bodenstein, Clustering Highly Parallelizable DBSCAN Algorithm, JSC, Online: http://www.fz-juelich.de/ias/jsc/EN/Research/DistributedComputing/DataAnalytics/Clustering/Clustering_node.html

[2] M. Goetz, M. Riedel et al.,' On Parallel and Scalable Classification and Clustering Techniques for Earth Science Datasets' 6th Workshop on Data Mining in Earth System Science, Proceedings of the International Conference of Computational Science (ICCS), Reykjavik, Online: http://www.proceedings.com/26605.html

Acknowledgements

PhD Student Gabriele Cavallaro, University of Iceland
Tómas Philipp Runarsson, University of Iceland
Kristján Jonasson, University of Iceland

Timo Dickscheid, Markus Axer, Stefan Köhnen, Tim Hütz,
Institute of Neuroscience & Medicine, Forschungszentrum Juelich

Selected Members of the Research Group on High Productivity Data Processing

Ahmed Shiraz Memon
Mohammad Shahbaz Memon
Markus Goetz
Christian Bodenstein
Philipp Glock
Matthias Richerzhagen

Thanks

Slides available at http://www.morrisriedel.de/talks