Mitglied der Helmholtz-Gemeinschaft

JÜLICH
FORSCHUNGSZENTRUM

Scalable Machine Learning on HPC

LSDMA LTM6: Data Intensive Computing

June, 1st 2015 | Markus Götz and Christian Bodenstein

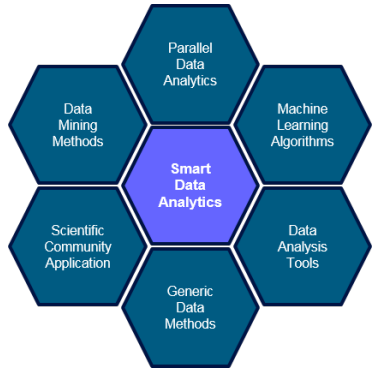
Agenda

- Introduction
- Machine Learning
- DBSCAN
- HPDBSCAN
- Demo
- Use Cases
- Conclusion

Introduction

High Productivity Data Processing

- Research Group
- Focus on Machine Learning and Data Mining
- Evaluation of Existing Parallel and Scalable Data Analysis Tools
- Develop “Big Data Analytics” Tools on HPC
- Analysis of Scientific Data in Use Cases

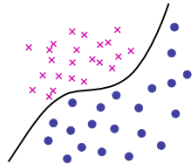


Machine Learning

- Increasing Size of Data
- Only a Fraction can by Analyzed by Hand
- Generalization from Few Examples to the Whole Data Set Desired
- Machine Learning Subjects:
  - Supervised / Unsupervised Learning Methods
  - Feature Engineering
  - Dimensionality Reduction

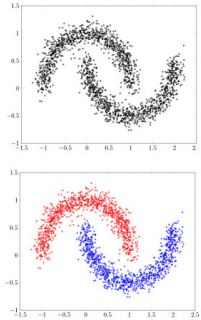
Machine Learning
Classification

- Supervised Learning Technique
- Ground Truth Samples needed to Train a Classification Model
- Predicts for each new Data Point Membership to Certain Category
- Well Known Representatives
  - K-Nearest Neighbors
  - Neural Networks
  - Support Vector Machines



Machine Learning
Clustering

- Unsupervised Learning Technique
- Subdevides Datasets into Similar Groups
- Facilitates Similarity Metrics (e.g. Euclidean Distance)
- Well Known Representatives:
  - K-Means
  - Hierarchical Clustering

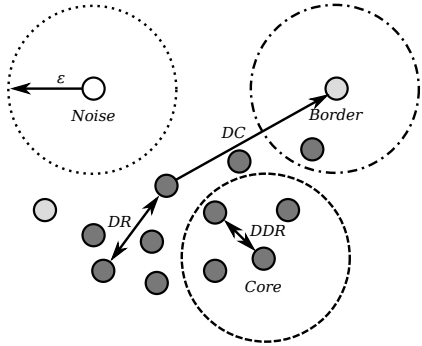


Related Work

- Single-Core Implementations Broadly Available
  - Programming Libraries
  - GUI Tools
  - All Common Languages
- Scalable Implementations Starting to Appear
  - Apache Spark/Mahout
  - Standalone Codes
  - Almost Nothing on HPC
- Only Simple Analysis Algorithms

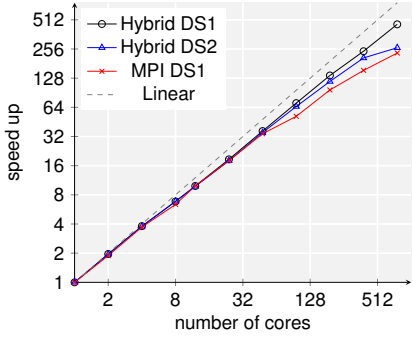
DBSCAN

- **D**ensity **B**ased **S**patial **C**lustering for **A**pplications wit **N**oise
- Formulated 1996 by Kriegel et. al.
- Two Parameters
  - ϵ – Spatial Search Radius
  - *minPts* – Density Threshold
- Detects Arbitrarily Shaped Clusters
- Filters Signal for Noise



HPDBSCAN

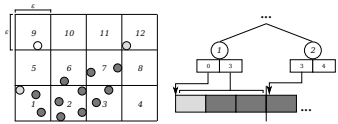
- **H**ighly **P**arallel **DBSCAN**
- Developed by JSC
- MPI + OpenMP Hybrid Application
- Uses HDF5 I/O Capabilities

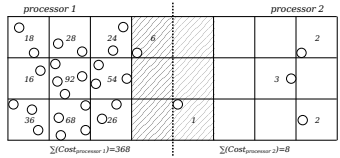


HPDBSCAN

Parallelization Strategy

1. Each Processor reads a Equal-Sized Chunk of the Data
2. Overlay Space with Spatial Hyper Grid
3. Apply Cost Heuristic to Balance Load
4. Redistribute Points to Achieve Data Locality
5. Execution of DBSCAN Locally
6. Cluster Merging using Halo Conflict Resolution
7. Restore Initial Order

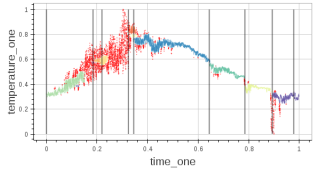


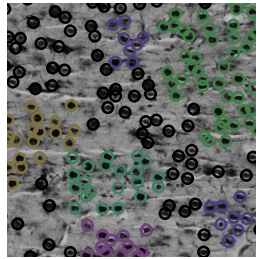


HPDBSCAN

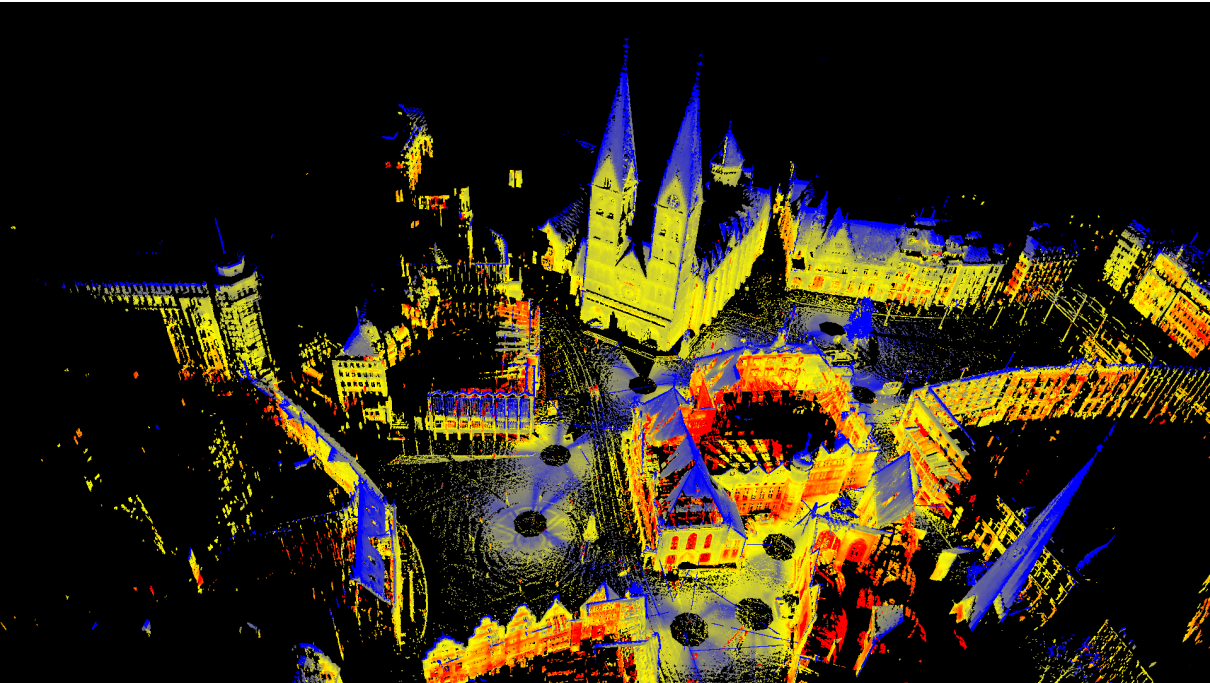
Use Cases

- PANGAEA Earth Science Data Repository
  - Koljoefjords in Sweden
  - Automatic Quality Control
  - Detect Outliers and Water-Mixing Events
- Human Brain Data Set
  - Microscopy Images of Brain Slices
  - Find Cell Nuclei
  - Analyze Cell-Density in the Cerebral Cortex



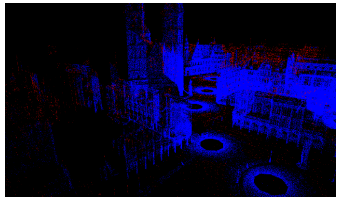


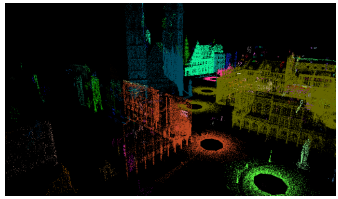
Bremen Demo



Bremen Point Cloud

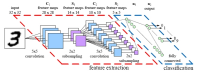
- Laser Scan Representing the Inner City of Bremen
- ~81.000.000 Data Points
- ~2 GB
- 1 Core $\rightarrow$ 22 h
- 768 Cores $\rightarrow$ 3 min





Other Activities

- Parallel Support Vector Machines
  - PiSVM
  - CascadeSVM
- Deep Neural Networks
  - Compare and Evaluate recent Software Solutions
  - First Scalable Parallelized MPI Prototype
- Close Cooperation with Domain Experts in
  - Earth Science
  - Neuro Science



Conclusion

- Various Machiene Learning Problems in Science
- Scalable Codes are Scarcely Available
  - Parallelization Issues
  - I/O Issues
- Trying to Bridge the Gap
- Huge Diversity in File Formats and Programming Languages
- Unify in a HPC Machine Learning Framework

Thank you for the Attention

Contact:
{m.goetz,c.bodenstein}@fz-juelich.de

LSDMA LTM6:
Data Intensive Computing
June, 1st 2015
Juelich Supercomputing Center

Slides:
Send us a Mail with a Request