Selected
Parallel and Scalable Methods for Scientific Big Data Analytics

Federated Systems and Data Division

Research Group

High Productivity Data Processing

Dr.-Ing. Morris Riedel et al.

Research Group Leader, Juelich Supercomputing Centre

Adjunct Associated Professor, University of Iceland

ZIH Kolloquium , 21th May 2015

Technical University of Dresden

HELMHOLTZ ASSOCIATION

JÜLICH FORSCHUNGSZENTRUM

UNIVERSITY OF ICELAND

SCHOOL OF ENGINEERING AND NATURAL SCIENCES

FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

Research Centre Juelich

<u>JUELICH in Numbers</u>

Area: 2.2 km^2

Staff: 5236
- **Scientists: 1658**
- **Technical staff: 1662**
- **Trainees: 303**

Budget: 557 Mio. €
- **incl. 172 Mio. € third party funding**

Located in Germany, Koeln – Aachen Area

<u>Institutes at JUELICH</u>

Institute of Complex Systems
Institute for Advanced Simulation
Juelich Supercomputing Center
Juelich Center for Neutron Science
Peter-Grünberg Institute
Institute for Neuroscience and Medicine
Institute for Nuclear Physics
Institute for Bio and Geosciences
Institute for Energy and Climate Research
Central Institute for Engineering, Electronics, and Analytics

Research for <u>generic</u> key technologies of the <u>next</u> generation

Scientific & Engineering Application-driven Problem Solving

University of Iceland

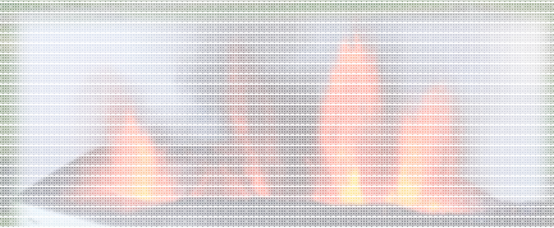
Schools of the University

- School of Education
- School of Humanities
- School of Engineering and Natural Sciences
- School of Social Sciences
- School of Health Sciences
- Interdisciplinary Studies

- Full programmes taught in English
- Staff: ~ 1259
- Students: ~14.000
- Located in Reykjavik Capital Center, Iceland

Faculties of the School

- Civil and Environmental Engineering
- Earth Sciences
- Electrical and Computer Engineering
- Industrial Engineering
- Mechanical Engineering
- Computer Science
- Life and Environmental Sciences
- Physical Sciences

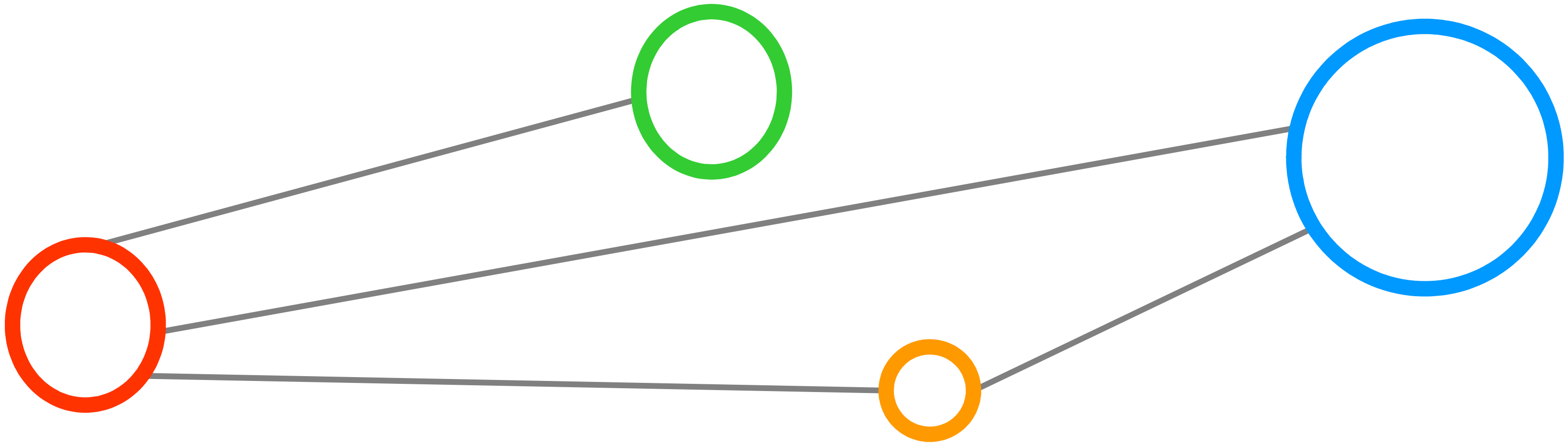


Teaching of key technologies in engineering & sciences

University Courses: Statistical Data Mining & HPC-A/B

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

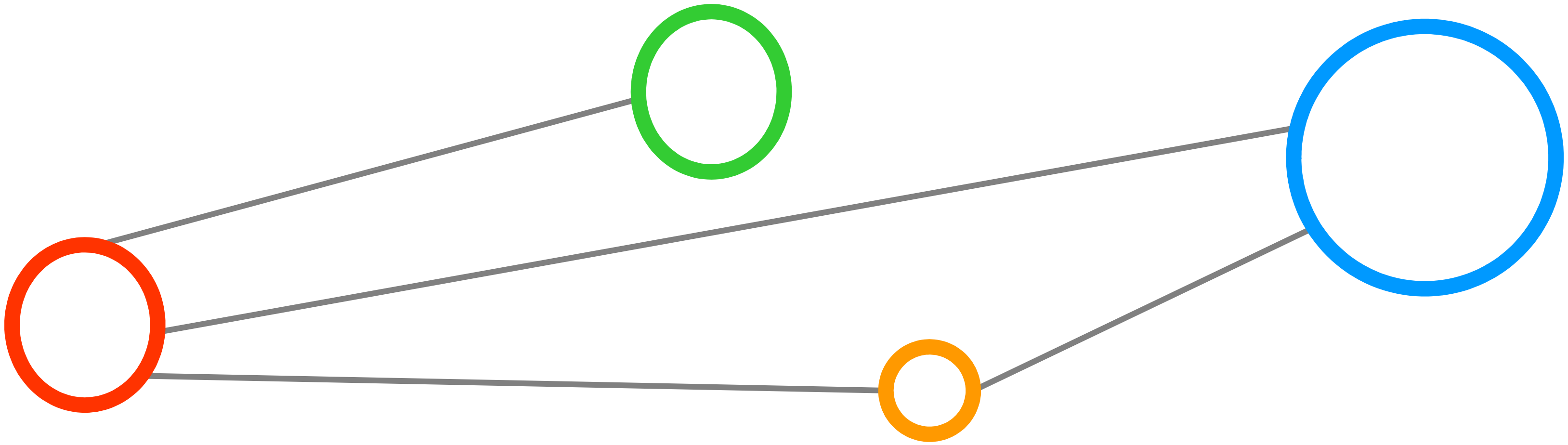
Outline



Outline

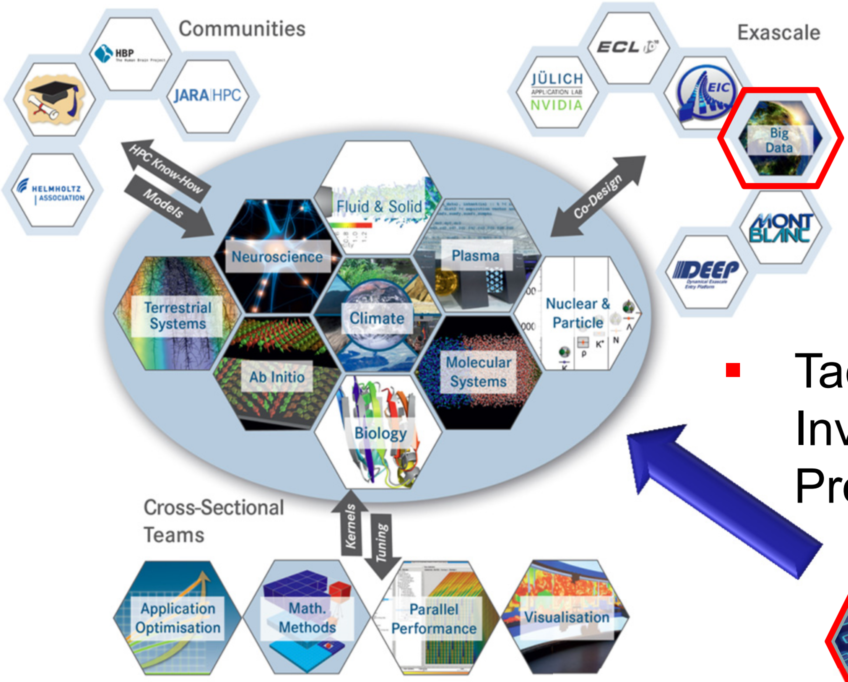
- Data Analytics @ Juelich
  - Driven by Scientific & Engineering Demands
  - Understanding of Terms & Key Focus
- Scalable & Parallel Tools
  - Clustering – DBSCAN
  - Classification – SVM
  - Scientific Applications in Context
- Recent Research Directions
  - ‘Brain Analytics‘
  - Deep Learning
- Conclusions
- References & Backup Slides

Data Analytics @ Juelich

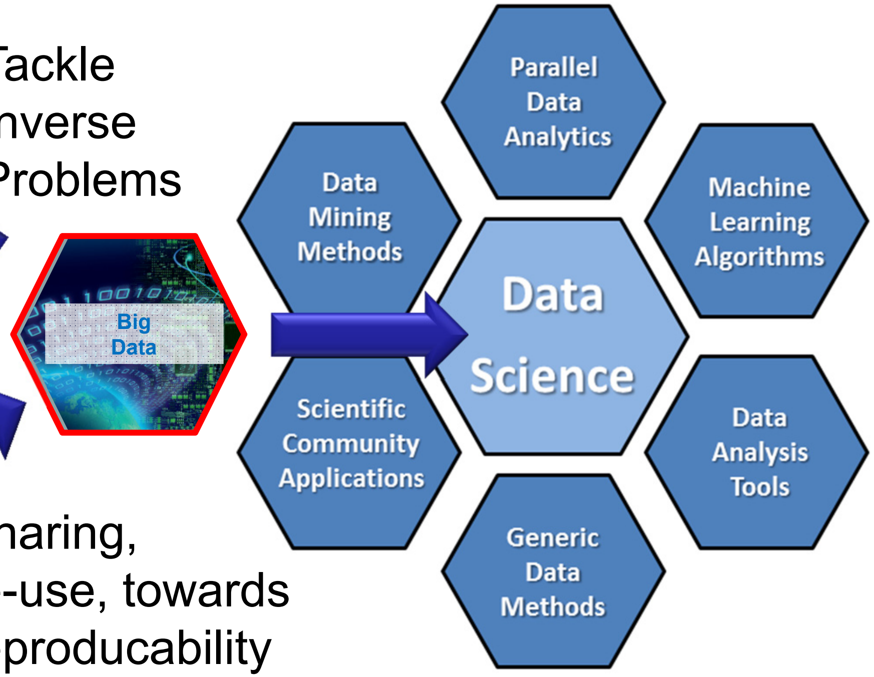


Data Analytics – Context JSC

- Research data-intensive science and engineering applications
- Explore computing that is more intertwined with data analysis



- Tackle Inverse Problems



Federated Data Management, Preservation, Security & Access

- Sharing, re-use, towards reproducability

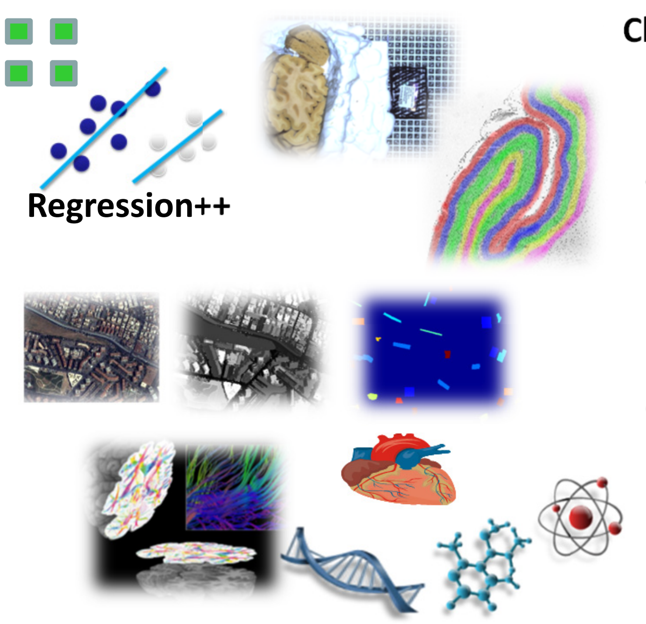
Data Analytics – Term Clarification

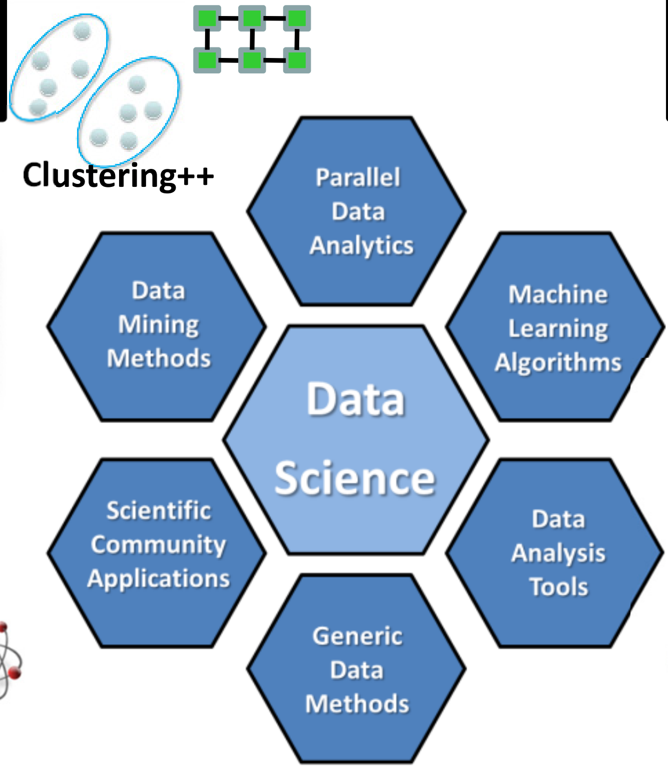
‘Data Analytics’ is an ‘interesting mix’ of different approaches

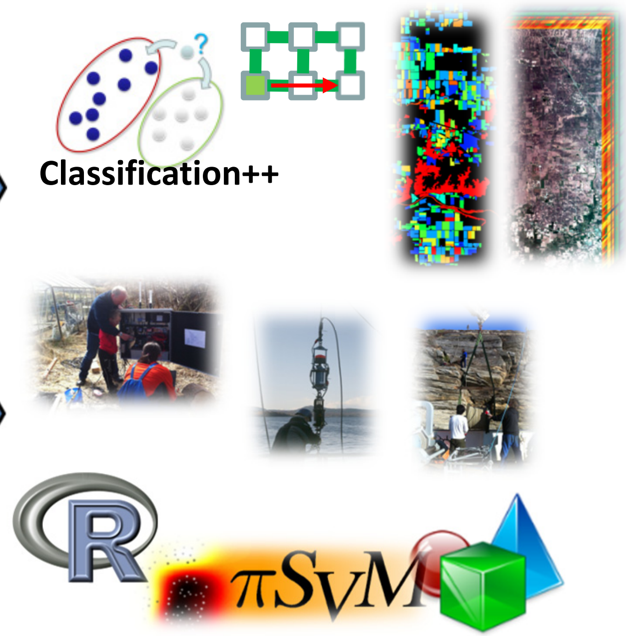
- Analytics: Whole methodology; Analysis: data investigation process itself
- ‘Big‘ requires scalable processing methods and underlying infrastructure

- **Concrete ‘big data‘: large medical data**

- **Concrete ‘big data‘: large earth science data**

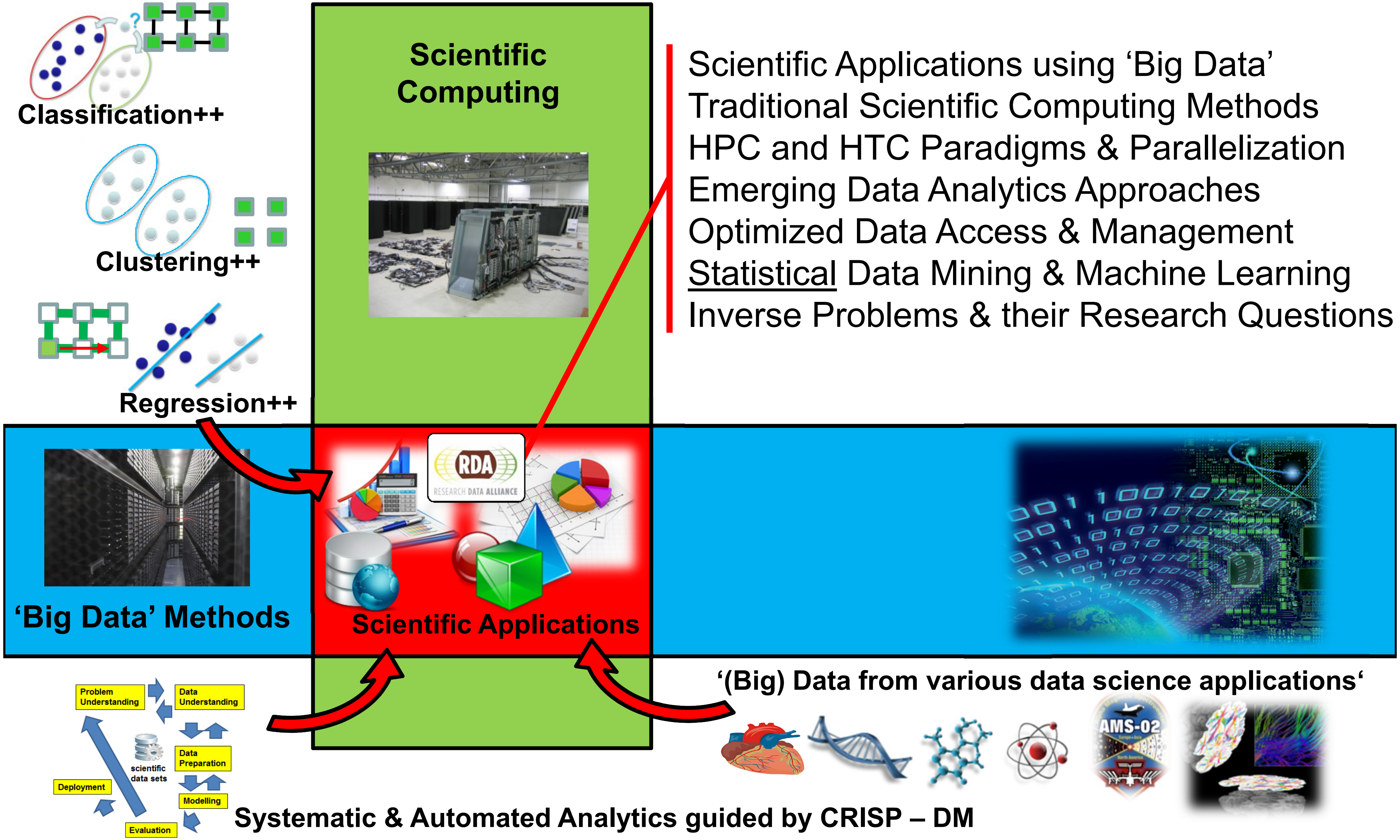






Data Analytics – Research Key Focus

Scientific Applications using ‘Big Data’
Traditional Scientific Computing Methods
HPC and HTC Paradigms & Parallelization
Emerging Data Analytics Approaches
Optimized Data Access & Management
Statistical Data Mining & Machine Learning
Inverse Problems & their Research Questions



Data Analytics – Selected Research Group Activities

John von Neumann Institute for Computing (NIC)

- Peer-review of scientific big data analytics (SBDA) proposels
- Jointly work with SBDA users (first projects starting, prototyping process)

John von Neumann - Institut für Computing NIC

Research Data Alliance (RDA)

- Chairing activities of the Big Data Analytics Interest Group
- Collaboration with a variety of EU and US partners
- Geoffrey Fox, UoIndiana (map-reduce), Kuo Kwo-Sen (NASA, SciDB)

RDA RESEARCH DATA ALLIANCE

Smart Data Innovation Lab

Smart Data Innovation Lab (SDIL)

- Driving activities in the personalised medicine community (with Bayer)
- Collaboration with partners from industry (e.g. IBM, SAP, Siemens, etc.)

Data Analytics – Selected Research Expertise

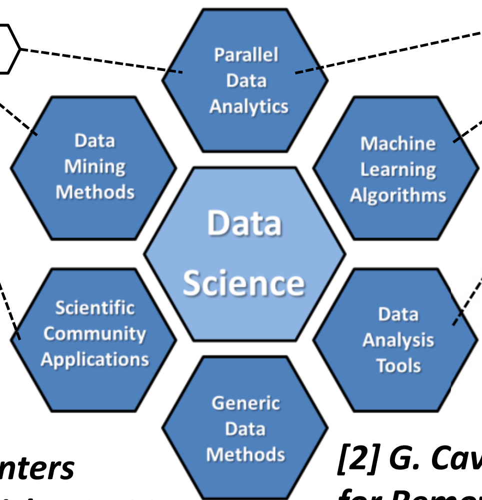
Key expertise making algorithms parallel & scalable for 'big data'

- Driven by scientific and engineering cases, e.g. understanding the human brain, remote sensing applications, marine measurements analysis, …
- Automate and/or support the data analysis process
- Example codes: Density-based Spatial Clustering of Applications with Noise (DBSCAN), Support Vector Machines (SVMs),

Parallel & Scalable DBSCAN clustering tool

Problem: Automatic outlier detection for data quality

- ✓ Tailor solution for community
- ✓ Scalability towards Big Data
- ✓ Design and improve automatic data analytics approaches



Parallel & Scalable SVM classification tool

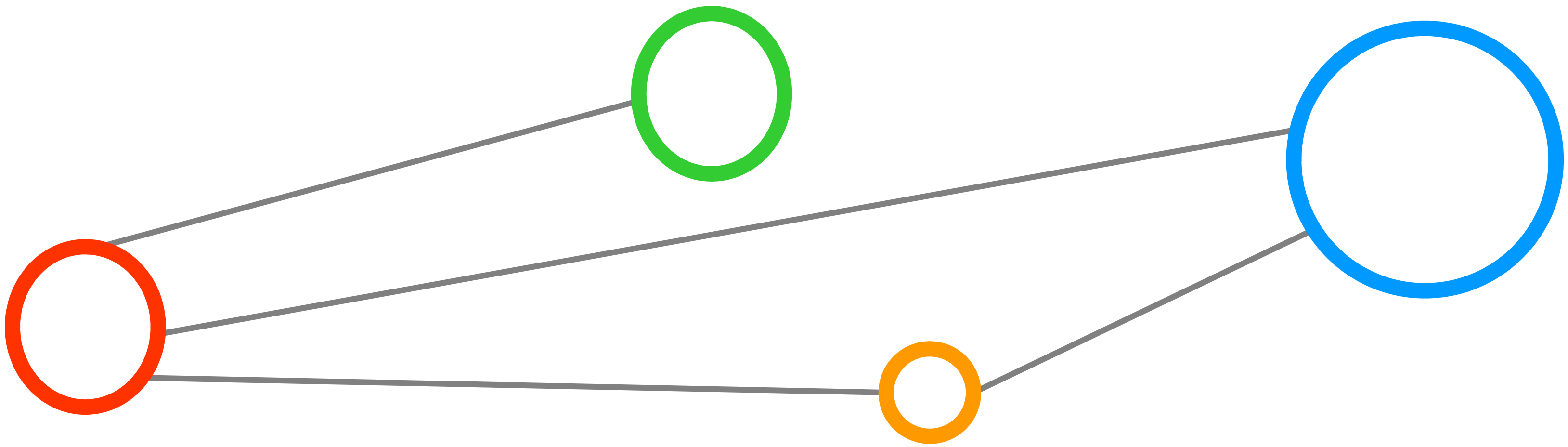
Problem: Classification of buildings from multi-spectral images

- ✓ Enable smooth transition from 'manual Matlab SVM scripts'
- ✓ Research on parallel SVM methods (map-reduce, HPC)

[1] R. Huber, M. Riedel et al.,'Research data enters scholarly communication and big data analysis', EGU2014

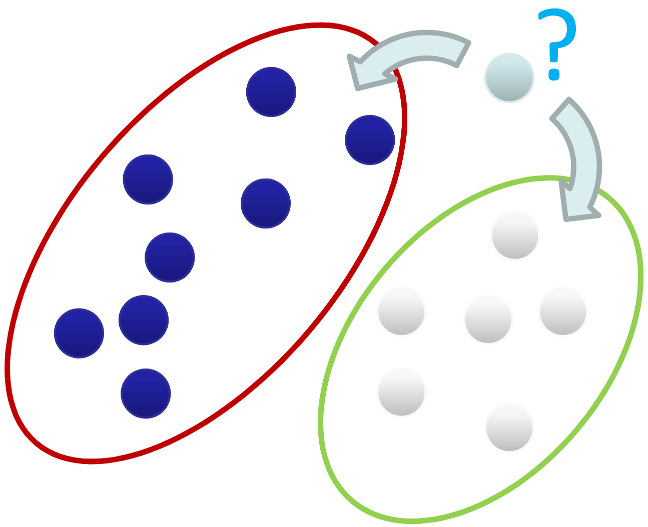
[2] G. Cavallaro, M. Riedel et al.,'Smart Data Analytics for Remote Sensing Applications', IGARSS2014

Scalable & Parallel Tools: Clustering



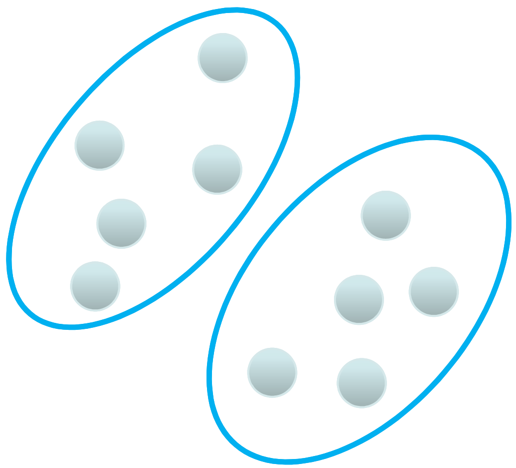
Learning From Data – Clustering Technique

Classification



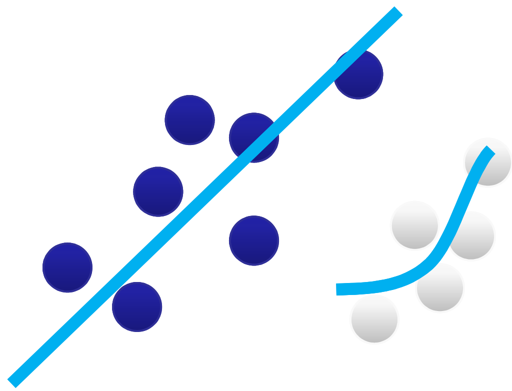
- Groups of data exist
- New data classified to existing groups

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

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 Open & Available 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

M. Goetz, M. Riedel et al., 6th Workshop on Data Mining in Earth System Science, International Conference of Computational Science (ICCS), Reykjavik, to be published

Parallel & Scalable DBSCAN MPI/OpenMP Tool (1)

DBSCAN Algorithm

- Introduced 1996 by Martin Ester et al. ***[4] Ester et al.***
- Groups number of similar points into clusters of data
- Similarity is defined by a distance measure (e.g. *euclidean distance*)

Unclustered Data

Distinct Algorithm Features

- Clusters a variable number of clusters
- Forms arbitrarily shaped clusters
- Identifies outliers/noise

Clustered Data

Understanding Parameters for MPI/OpenMP tool

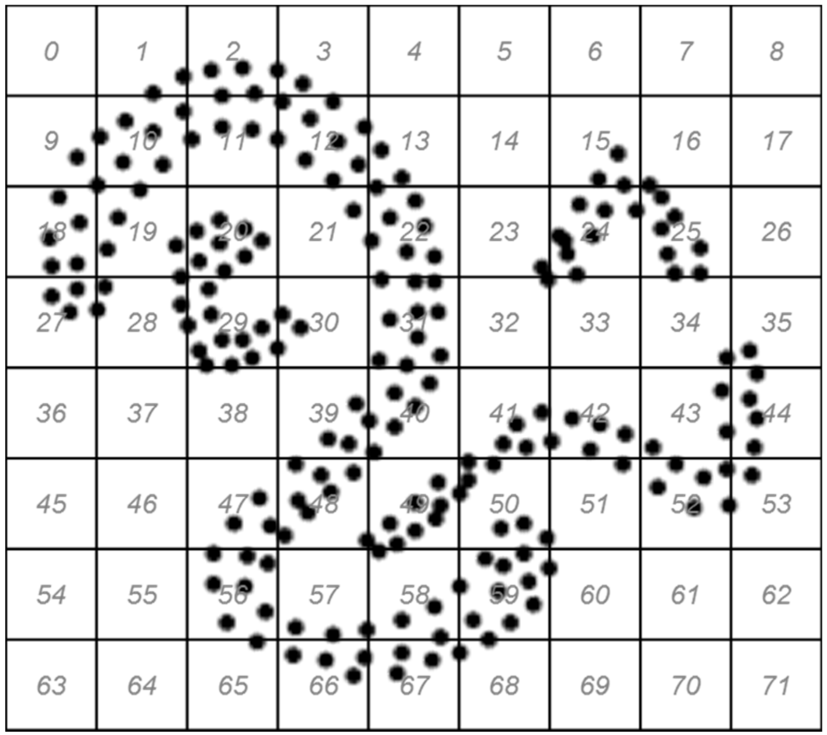
- Looks for a similar points within a given search radius
  → Parameter *epsilon*
- A cluster consist of a given minimum number of points
  → Parameter *minPoints*

[3] M.Goetz & C. Bodenstein, HPDBSCAN Tool

Parallel & Scalable DBSCAN MPI/OpenMP Tool (2)

Parallelization Strategy

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



[3] M.Goetz & C. Bodenstein, HPDBSCAN Tool

Preprocessing Step

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

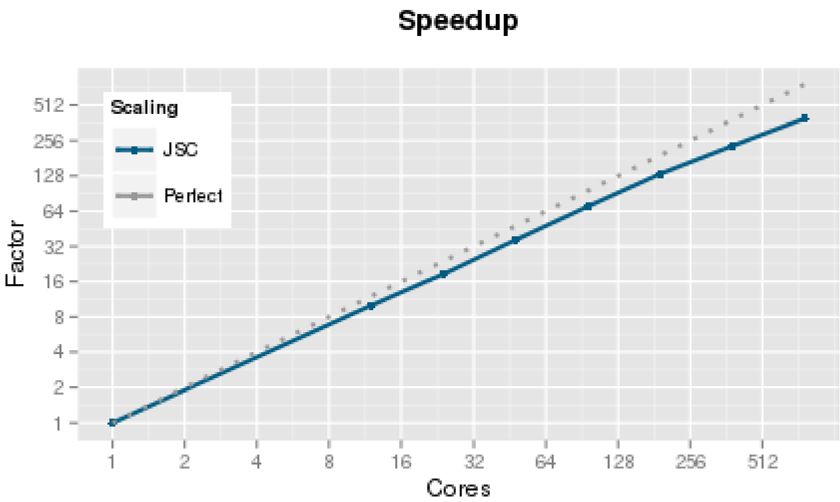
Computational Optimizations

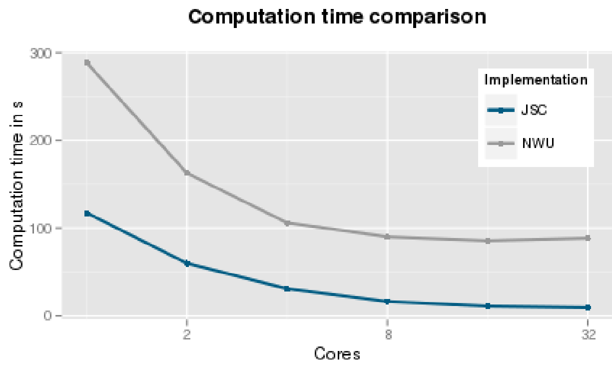
- Caching of point neighborhood searches
- Cluster merging based on comparisons instead of zone reclustering

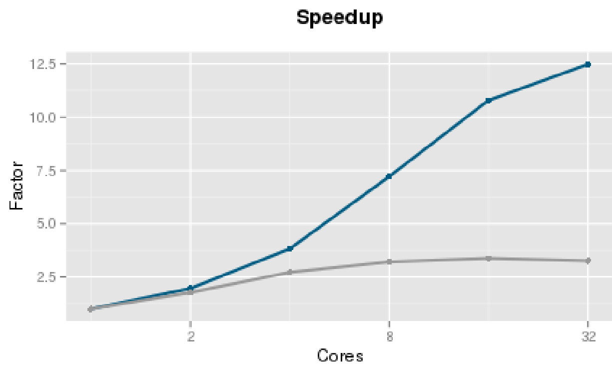
Parallel & Scalable DBSCAN MPI/OpenMP Tool (3)

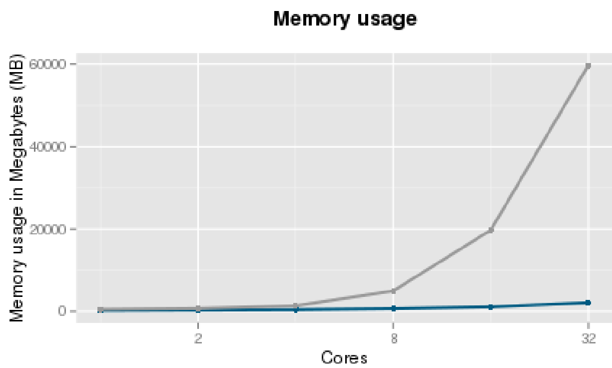
Performance Comparisons

- With another open-source parallel DBSCAN implementation (aka 'NWU') ***[5] Patwary et al.***
- 3.7056.351 data points (2 dimensions)
- Use of Hierarchical Data Format (HDF) v.5 for scalable input/output of 'big data'

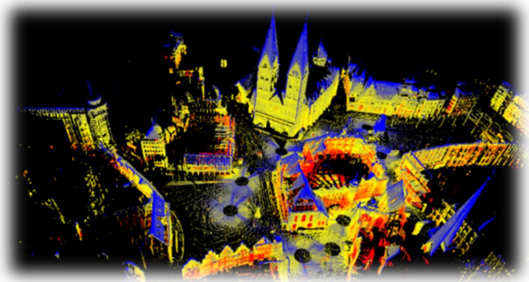






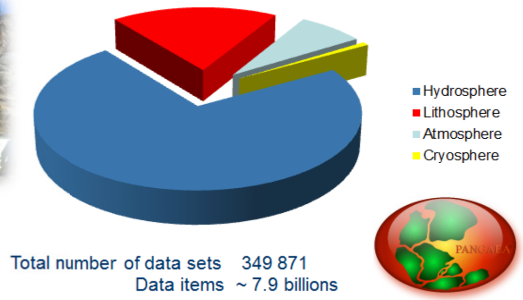


Parallel & Scalable DBSCAN MPI/OpenMP Tool (4)



Selected ‘Big Data‘ Applications

- London twitter data
  (goal: find density centers of tweets)
- Bremen thermo point cloud data
  (goal: noise reduction)
- PANGAEA earth science datasets
  (goal: automated outlier detection)



[6] Open PANGAEA Earth Science Data Collection

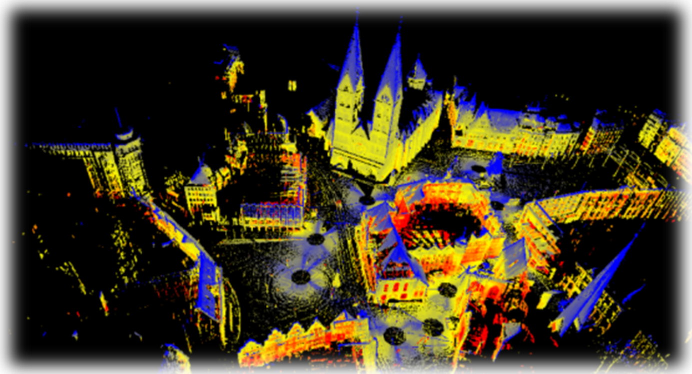
	Cores					
Computation time	1	2	4	8	16	32
JSC-HPDBSCAN	117,18 s	59,64 s	30,68 s	16,25 s	10,86 s	9,39 s
NWU-PDSDBSCAN	288,35 s	162,47 s	105,94 s	89,87 s	85,37 s	88,42 s
Speed-Up						
JSC-HPDBSCAN	1,00 x	1,96 x	3,82 x	7,21 x	10,79 x	12,48 x
NWU-PDSDBSCAN	1,00 x	1,77 x	2,72 x	3,21 x	3,38 x	3,26 x
Memory						
JSC-HPDBSCAN	251,064 MB	345,276 MB	433,340 MB	678,248 MB	1,101 GB	2,111 GB
NWU-PDSDBSCAN	500,512 MB	725,104 MB	1,370 GB	4,954 GB	19,724 GB	59,685 GB

Parallel & Scalable DBSCAN MPI/OpenMP Tool (5)

Free tool available

- Public bitbucket account – open-source
- Tool Website with more information
- Maintained on best effort basis

[3] M.Goetz & C. Bodenstein, HPDBSCAN Tool



3D Point Cloud of Bremen/Germany

→ Usage via simple jobscripts

Usage

- module load hdf5/1.8.13
- mpiexec -np 1 ./dbscan -e 300 -m 100 -t 12 bremenSmall.h5

Parameter *epsilon*

Parameter *minPoints*

Parallel & Scalable DBSCAN MPI/OpenMP Tool (6)

Usage via jobscript

- Using MOAB job scheduler
- Important: module load hdf5/1.8.13
- Important: library gcc-4.9.2/lib64
- np = number of processors
- t = number of threads

JUDGE @ Juelich

```
mriedel@judge:/homeb/zam/analytic/bigdata/hpdbscan/jsc_mpi/mriruns> more datajobscript.sh
#!/bin/bash
#MSUB -N HPDBSCAN_BremenSmall_1_12
#MSUB -l nodes=1:ppn=12:gpus=0:performance
#MSUB -l walltime=00:03:00
#MSUB -M m.riedel@fz-juelich.de
#MSUB -m abe
#MSUB -v tpt=12
#MSUB -l vmem=64gb
#MSUB -q devel

module load hdf5/1.8.13
export LD_LIBRARY_PATH=/homeb/zam/analytic/bigdata/hpdbscan/gcc-4.9.2/lib64:$LD_LIBRARY_PATH
DBSCAN=/homeb/zam/analytic/bigdata/hpdbscan/jsc_mpi/dbscan
SMALLBREMENDATA=/homeb/zam/analytic/bigdata/hpdbscan/jsc_mpi/mriruns/bremenSmall.h5

cd /homeb/zam/analytic/bigdata/hpdbscan/jsc_mpi/mriruns
mpiexec -np 1 $DBSCAN -e 300 -m 100 -t 12 $SMALLBREMENDATA
```

DBSCAN Parameters

Parallel & Scalable DBSCAN MPI/OpenMP Tool (7)

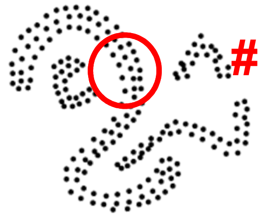
Output with various information

- Run-times of different stages
- Clustering task information
  (e.g. number of identified clusters)
- Noise identification
- Data volume (small Bremen): ~72 MB
- Data volume (large Bremen): ~1.9 GB

JUDGE @ Juelich

```
mriedel@judge:/homeb/zam/analytic/bigdata/hpdbscan/jsc_mpi/mriruns> more HPDBSCAN_BremenSmall_1_12.o2208066
Calculating Cell Space...
        Computing Dimensions... [OK] in 0.011853
        Computing Cells...      [OK] in 0.073445
        Sorting Points...       [OK] in 0.124476
        Distributing Points...  [OK] in 0.000000
DBSCAN...
        Local Scan... I am ready 0
 in 90.606330           [OK] in 90.606364
        Merging Neighbors...    [OK] in 0.000000
        Adjust Labels ...       [OK] in 0.004972
        Rec. Init. Order ...    [OK] in 1.255420
        Writing File ...        [OK] in 0.019120
Result...
        65      Clusters
        2973821 Cluster Points
        26179   Noise Points
        2953129 Core Points
Took: 92.214843s
```

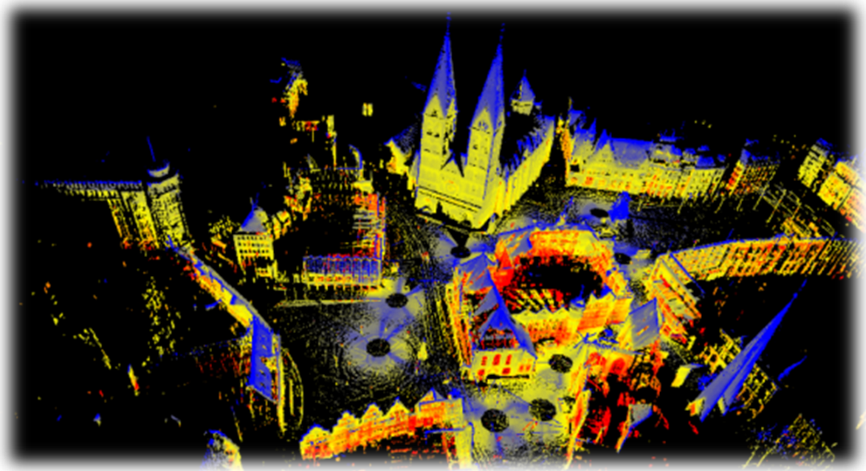
Output results written in same input data:
cluster number & noise label
(depends on parameters)



Parallel & Scalable DBSCAN MPI/OpenMP Tool (8)

Visualization Example

- Using Point Cloud Library (PDL) toolset
- Transformation of Data to PCD format (python script on the right)



Usage

- python H5toPCD.py bremenSmall.h5
- pcl_viewer bremenClustered.pcd

H5toPCD.py python script

```
import h5py as h5
import numpy as np
import sys

if len(sys.argv) < 2:
    INPUT="bremen.h5"
else:
    INPUT = sys.argv[1]
FILE = "bremenClustered.pcd"

print"loading H5"
bremen = h5.File("bremenSmall.h5")
points = bremen["DBSCAN"]
clusters = bremen["Clusters"]
colors = bremen["COLORS"]

print "Transform to numpy"
points = np.array(points)
clusters = np.array(clusters)
colors = np.array(colors)

#print "Remove Noise"
#points = points[clusters!=0]
#clusters = clusters[clusters!=0]

#data = np.concatenate((points,colors.reshape((-1,1))),axis=1)
data = np.concatenate((points,clusters.reshape((-1,1))),axis=1)

clusters[clusters!=0]=1
data = np.concatenate((data,clusters.reshape((-1,1))),axis=1)

print "Write PCD"
with open(FILE, "w+") as out:
    out.write("""# .PCD v0.7 - Point Cloud Data file format
VERSION 0.7
FIELDS x y z rgb noise
SIZE 4 4 4 4 4
TYPE F F F F F
COUNT 1 1 1 1 1
WIDTH %d
HEIGHT 1
VIEWPOINT 0 -50000 -50000 1 0 0 0
POINTS %d
DATA ascii
""" % (len(data),len(data),))
    np.savetxt(out, data)
```

Take advantage of NumPy library

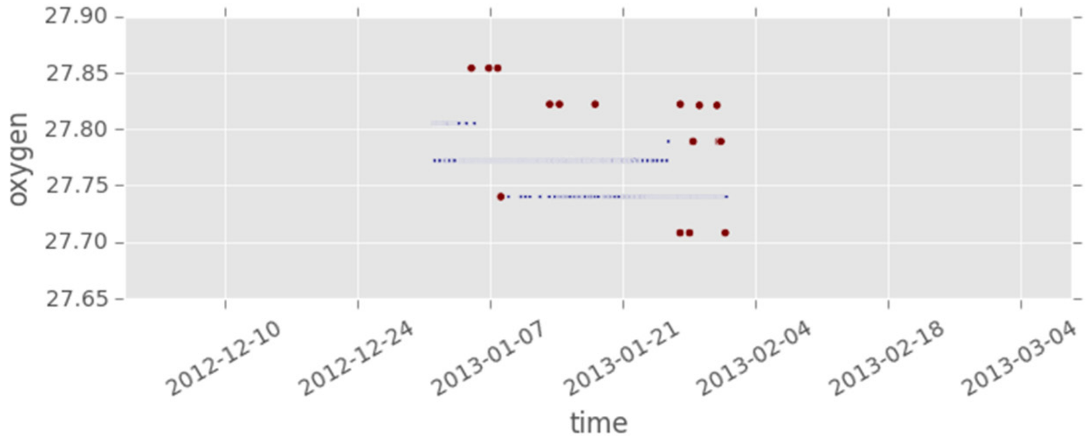
Parallel & Scalable DBSCAN MPI/OpenMP Tool (9)

Earth Science Application 'Automated outlier detection in time series'

- Collaboration with MARUM, Bremen (work in progress)
- Example: water quality data of Koljoefjords
- Connected underwater device
- Measurements: oxygen, temperature, salinity, …

Use of HPBSCAN algorithm

- Detect outliers and anomalies/events
- Compare with manually annotated data by domain-scientist
- Needs auomation

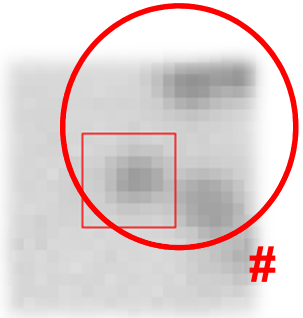


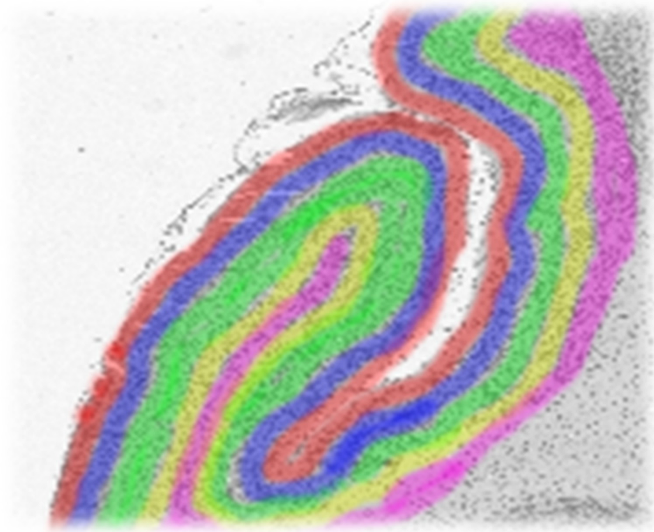
Parallel & Scalable DBSCAN MPI/OpenMP Tool (10)

Neuroscience Application

‘Cell nuclei detection and tissue clustering‘

- Scientific Case: 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





Use of HPBSCAN algorithm

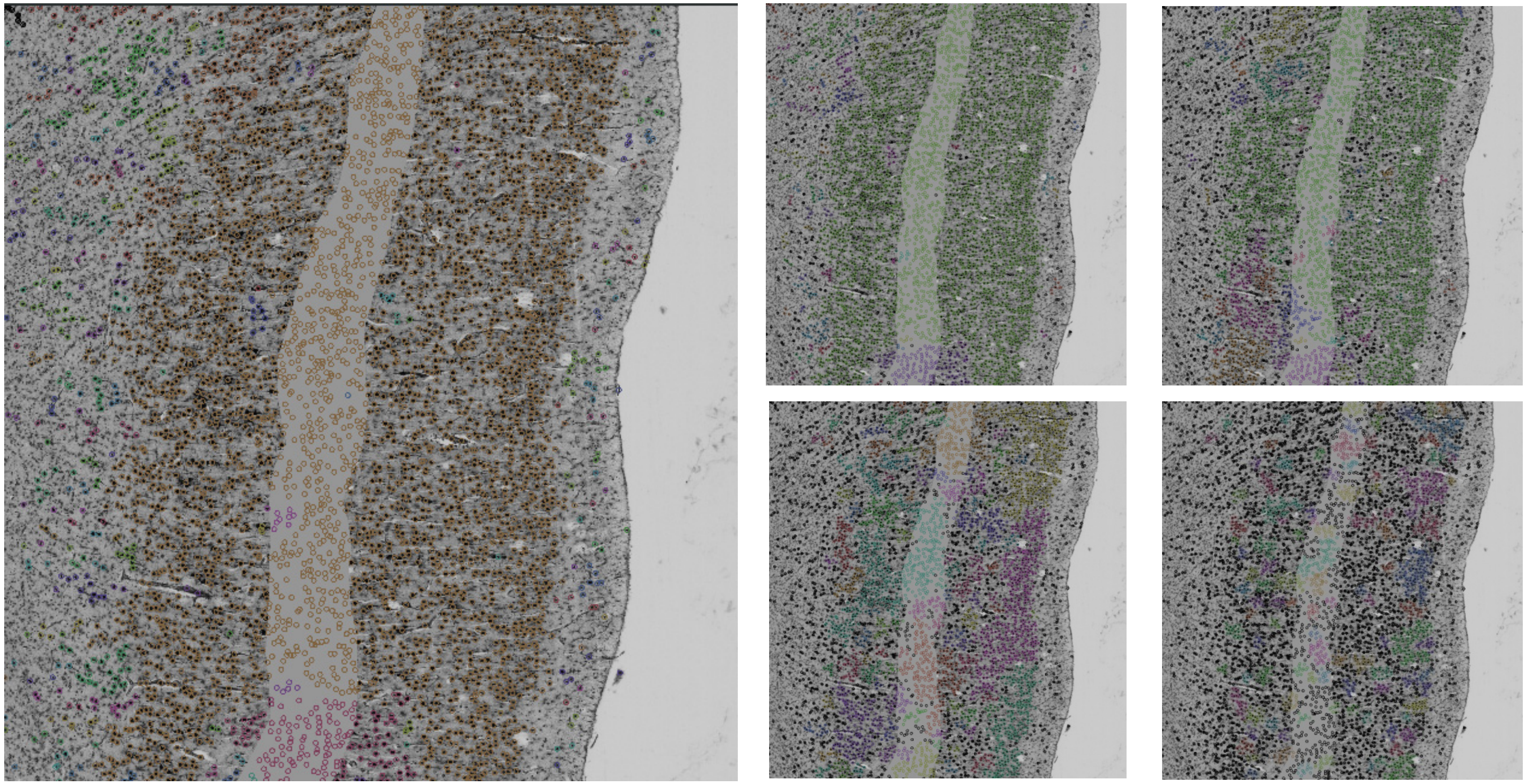
- First 2d results detect various clusters
- Work in progress, not very good results
- Approach: Several iterations (with 3D) with potentially different parameter values
- Investigate other methods (e.g. OPTICS)

> ➢ **Research activities jointly with T. Dickscheid et al. (Juelich Institute of Neuroscience & Medicine)**

Parallel & Scalable DBSCAN MPI/OpenMP Tool (11)

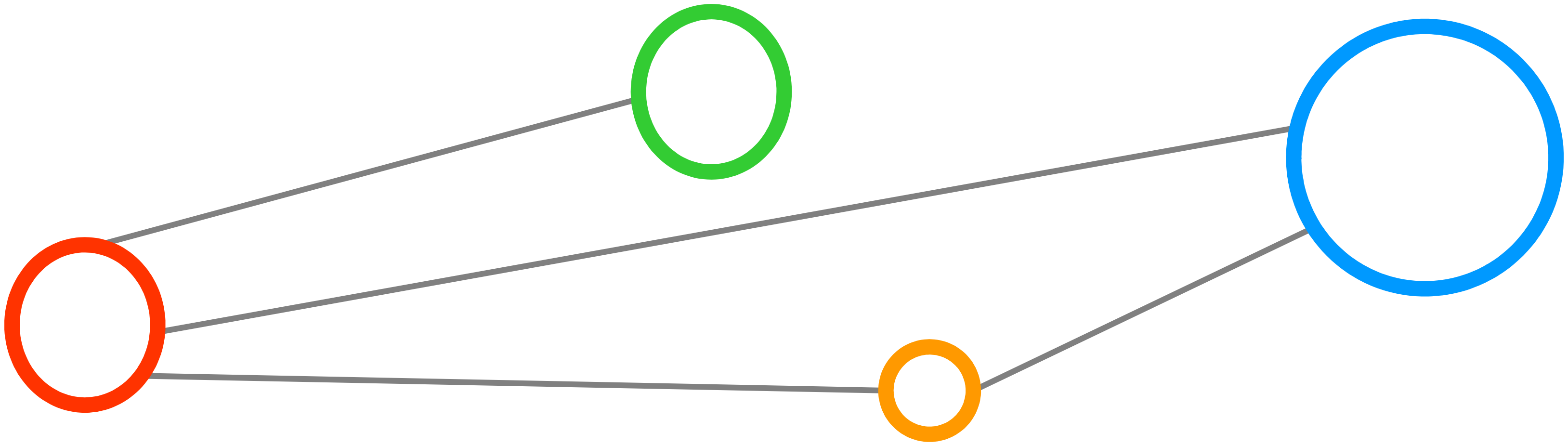
Neuroscience Application – Work in progress (e.g. 3120x3288)

‘Cell nuclei detection and tissue clustering‘ – varying parameters



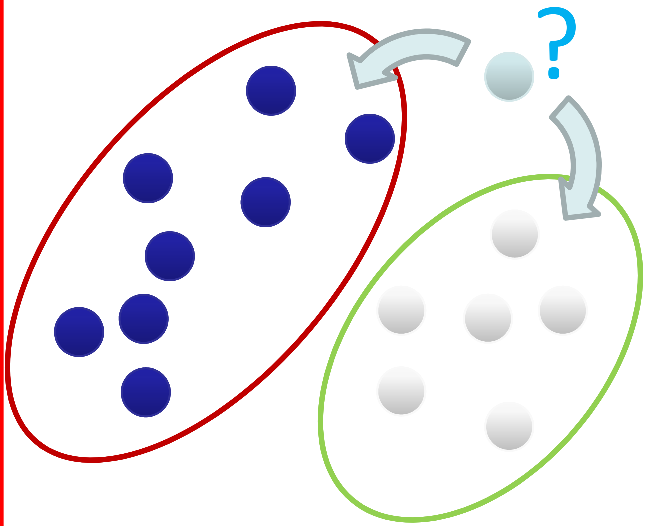
- **Research activities jointly with T. Dickscheid et al. (Juelich Institute of Neuroscience & Medicine)**

Scalable & Parallel Tools: Classification



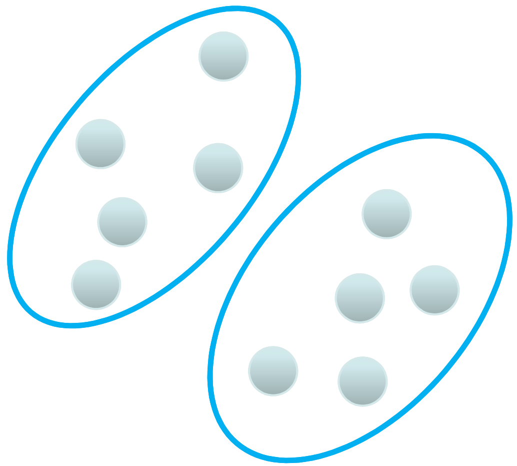
Learning From Data – Classification Technique

Classification



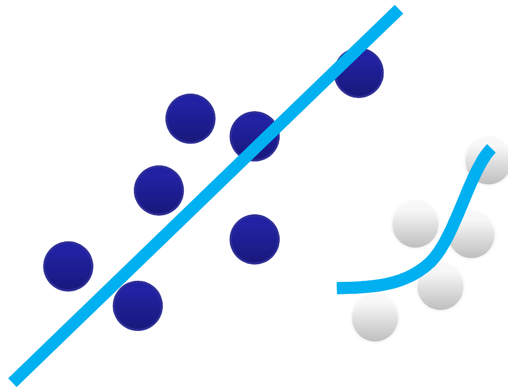
- Groups of data exist
- New data classified to existing groups

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

Selected Classification Methods

Perceptron Learning Algorithm – simple linear classification

- Enables binary classification with 'a line' between classes of seperable data

Support Vector Machines (SVMs) – non-linear ('kernel') classification

- Enables non-linear classification with maximum margin (best 'out-of-the-box')

Reasoning: achieves often better results than other methods in remote sensing application

Decision Trees & Ensemble Methods – tree-based classification

- Grows trees for class decisions, ensemble methods average n trees

Artificial Neural Networks (ANNs) – brain-inspired classification

- Combine multiple linear perceptrons to a strong network for non-linear tasks

Naive Bayes Classifier – probabilistic classification

- Use of the Bayes theorem with strong/naive independence between features

Technology Review of Open & Available Tools

Technology	Platform Approach	Analysis
Apache Mahout	Java; Hadoop	No parallelization strategy for SVMs
Apache Spark/MLlib	Java; Spark	Parallel linear SVMs (no multi-class)
Twister/ParallelSVM	Java; Twister; Hadoop 1.0	Parallel SVMs, open source; developer version 0.9 beta
scikit-learn	Python	No parallelization strategy for SVMs
piSVM 1.2 & piSVM 1.3	C; MPI	Parallel SVMs; stable; not fully scalable
GPU LibSVM	CUDA	Parallel SVMs; hard to programs, early versions
pSVM	C; MPI	Parallel SVMs; unstable; beta version

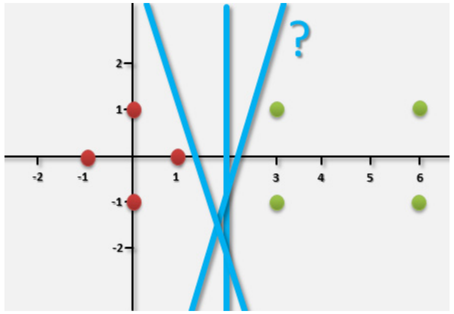
M. Goetz, M. Riedel et al., 6th Workshop on Data Mining in Earth System Science, International Conference of Computational Science (ICCS), Reykjavik, to be published

Parallel & Scalable SVM MPI Tool (1)

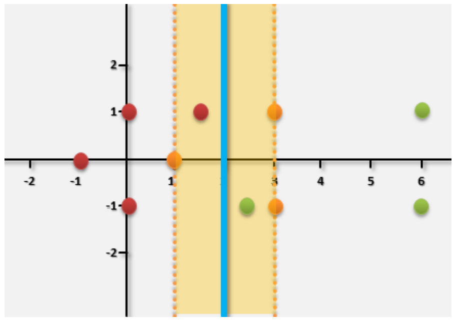
SVM Algorithm Approach

[7] C. Cortes and V. Vapnik et al.

- Introduced 1995 by C.Cortes & V. Vapnik et al.
- Creates a ‘maximal margin classifier‘ to get future points (‘more often‘) right
- Uses quadratic programming & Lagrangian method with N x N



(linear example)



(‘maximal margin clasifier‘ example)

(use of soft-margin approach for better generalization)

$$\min_{w,\xi_i,b} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \right\}$$

$$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

(maximizing hyperplane turned into optimization problem, minimization, dual problem)

$$\mathcal{L}(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m \boxed{\mathbf{x}_n^T \mathbf{x}_m}$$

(max. hyperplane → dual problem, using quadratic programming method)

$$\boxed{k(\mathbf{x}_i, \mathbf{x}_j)} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$$

$$0 \leq \alpha_i \leq C$$

$$\begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \dots y_1 y_N x_1^T x_N \\ \dots & \dots & \ddots \\ y_N y_1 x_N^T x_1 & y_N y_2 x_N^T x_2 & \dots y_N y_N x_N^T x_N \end{bmatrix}$$

(kernel trick, quadratic coefficients – Computational Complexity & Big Data Impact)

Parallel & Scalable SVM MPI Tool (2)

- **True Support Vector Machines are Support Vector Classifiers combined with a non-linear kernel**
- **Non-linear kernels exist - mostly known are polynomial & Radial Basis Function (RBF) kernels**

[8] An Introduction to Statistical Learning

Understanding the MPI tool parameters

- Selecting non-linear kernel function K type as RBF → parameter –t 2
- Setting RBF Kernel configuration parameter γ → e.g. parameter –g 16
- Setting SVM allowed errors parameter → e.g. parameter –c 10000

Major benefit of Kernels: Computing done in original space

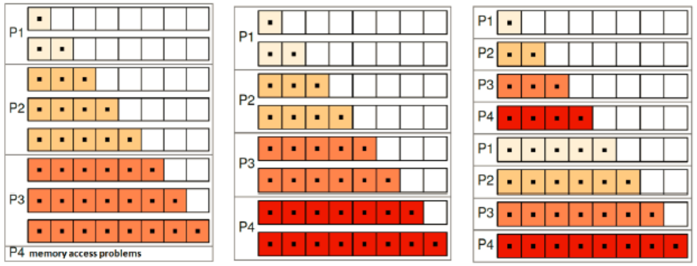
- Linear Kernel $K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}$ (linear in features)
- Polynomial Kernel $K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d$ (polynomial of degree d)
- RBF Kernel $K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2)$ (large distance, small impact)

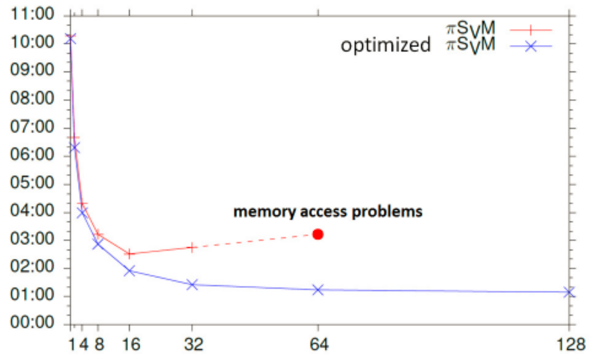
Parallel & Scalable SVM MPI Tool (3)

Original parallel piSVM tool 1.2

- Open-source and based on libSVM library, C, 2011
- Message Passing Interface (MPI)
- New version appeared 2014-10 v. 1.3 (no major improvements)
- Lack of ‘big data‘ support (memory, layout, etc.)

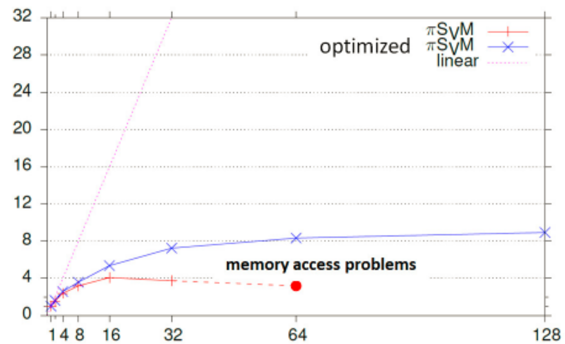
[9] piSVM Website, 2011/2014 code



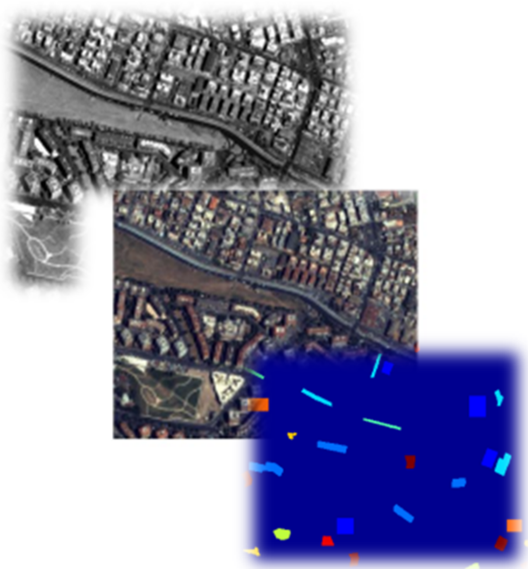


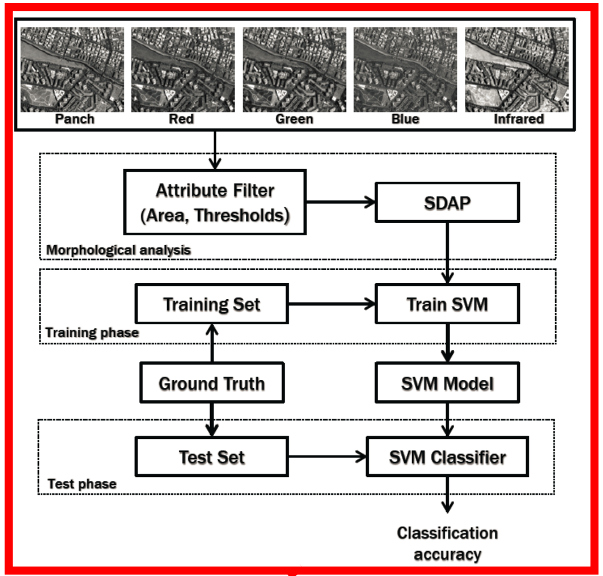
Tuned scalable parallel piSVM tool 1.2.1

- Open-source (repository to be created)
- Based on piSVM tool 1.2
- Optimizations: load balancing; MPI collectives
- Contact: m.richerzhagen@fz-juelich.de



Parallel & Scalable SVM MPI Tool (4)





Class	Training	Test
Buildings	18126	163129
Blocks	10982	98834
Roads	16353	147176
Light Train	1606	14454
Vegetation	6962	62655
Trees	9088	81792
Bare Soil	8127	73144
Soil	1506	13551
Tower	4792	43124
Total	77542	697859

Sattelite Data (Quickbird)

Parallel Support Vector Machines (SVM)

HPC / MPI

Classification Study of Land Cover Types

„Reference Data Analytics“ for reusability & learning

CRISP-DM Report

Openly Shared Datasets

Running Analytics Code

[10] Rome Image dataset

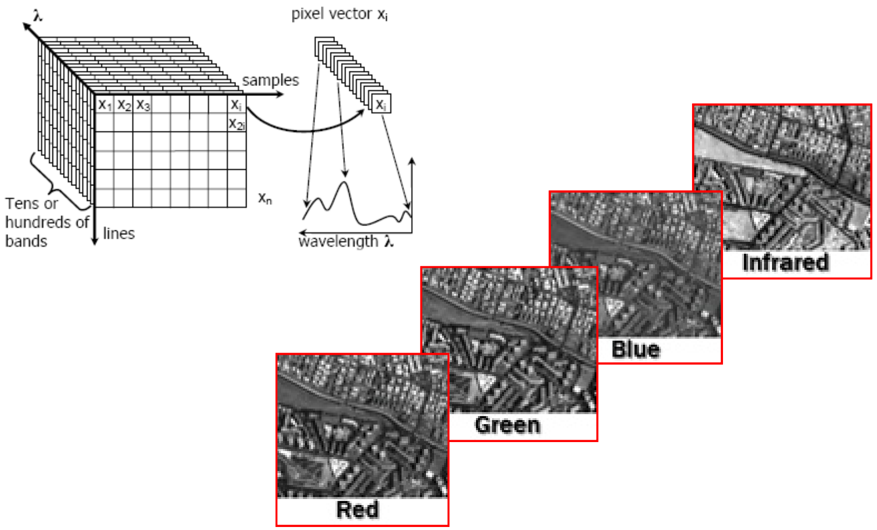
Parallel & Scalable SVM MPI Tool (5)

Example dataset: Geographical location: Image of Rome, Italy

- Remote sensor data obtained by Quickbird satellite

High-resolution (0.6m) panchromatic image

Pansharpened (UDWT) low-resolution (2.4m) multispectral images

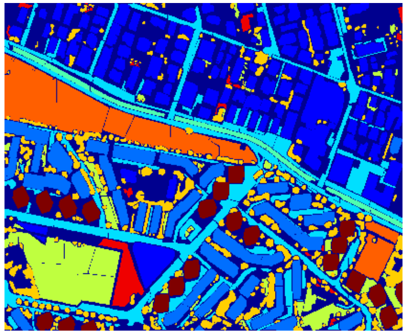


[10] Rome Image dataset

Parallel & Scalable SVM MPI Tool (6)

Labelled data available for train/test data

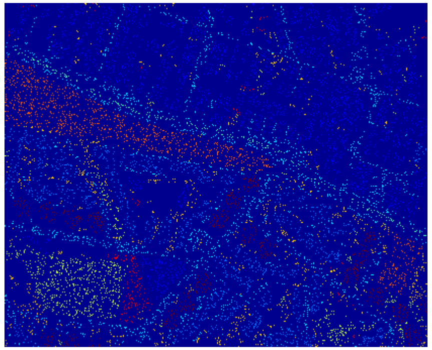
- Groundtruth data of 9 different land-cover classes available



Class	Training	Test
Buildings	18126	163129
Blocks	10982	98834
Roads	16353	147176
Light Train	1606	14454
Vegetation	6962	62655
Trees	9088	81792
Bare Soil	8127	73144
Soil	1506	13551
Tower	4792	43124
Total	77542	697859

Data preparation

- We generated a set of training samples by randomly selecting 10% of the reference samples (with labelled data)
- Generated set of test samples from the remaining labels (labelled data, 90% of reference samples)



Training Image
(10% pixels/class)

[10] Rome Image dataset

B2SHARE
Store and Share Research Data

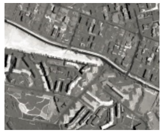
Parallel & Scalable SVM MPI Tool (7)

Based on 'LibSVM data format' (using feature extraction method)

- Add 'Self-Dual Attribute Profile (SDAP) on Area' on all images training file

Area

Std Dev



Moment of Inertia

[11] G. Cavallaro, M. Mura, J.A. Benediktsson, L. Bruzzone et al.

Class **Number Feature** **Gray Level**

Each line is a training vector with gray levels

each line is a pixel

3 1:0.105882 2:0.109804 3:0.101961 .... 54:0.121569 55:0.130952
2 1:0.364706 2:0.360784 3:0.356863 .... 54:0.356863 55:0.349206
6 1:0.152941 2:0.34902 3:0.454902 .... 54:0.466667 55:0.460317
....
....
....
....
9 1:0.247059 2:0.247059 3:0.227451 .... 54:0.227451 55:0.218254
7 1:0.411765 2:0.411765 3:0.415686 .... 54:0.415686 55:0.40873

Buildings
Blocks
Roads
Light Train
Vegetation
Trees
Bare Soil
Soil
Tower

#77542 samples

55 features

[10] Rome Image dataset

B2SHARE
Store and Share Research Data

Parallel & Scalable SVM MPI Tool (8)

Usage via jobscript

- Using MOAB job scheduler
- np = number of processors; o/q partitioning

JUDGE @ Juelich

```
#!/bin/bash
#MSUB -N Train-tune-rec86-4-16-32
#MSUB -l nodes=4:ppn=16:performance
#MSUB -l walltime=03:00:00
#MSUB -M m.riedel@fz-juelich.de
#MSUB -m abe
#MSUB -W x=naccesspolicy:singlejob
#MSUB -v tpt=2
#MSUB -q devel


### jobscript

cd $PBS_O_WORKDIR
echo "workdir: $PBS_O_WORKDIR"

NSLOTS=32

echo "running on $NSLOTS cpus..."

### location
PISVM=/homeb/zam/mriedel/pisvm-1.2/pisvm-1.2/pisvm-train

TRAINDATA=/homeb/zam/mriedel/bigdata/86-
romeok/sdap_area_all_training.el

### submit
mpiexec -np $NSLOTS $PISVM -o 1024 -q 512 -c 10000 -g 16 -t 2 -m
1024 -s 0 $TRAINDATA
```

→ Usage via simple jobscripts

SVM
Parameters

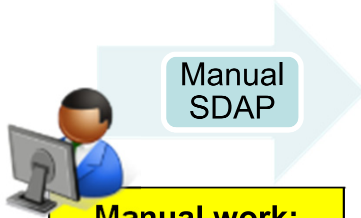
[12] Rome Analytics Results & job scripts

B2SHARE
Store and Share Research Data

Parallel & Scalable SVM MPI Tool (9)

Training speed-up is possible when number of features is ‘high‘

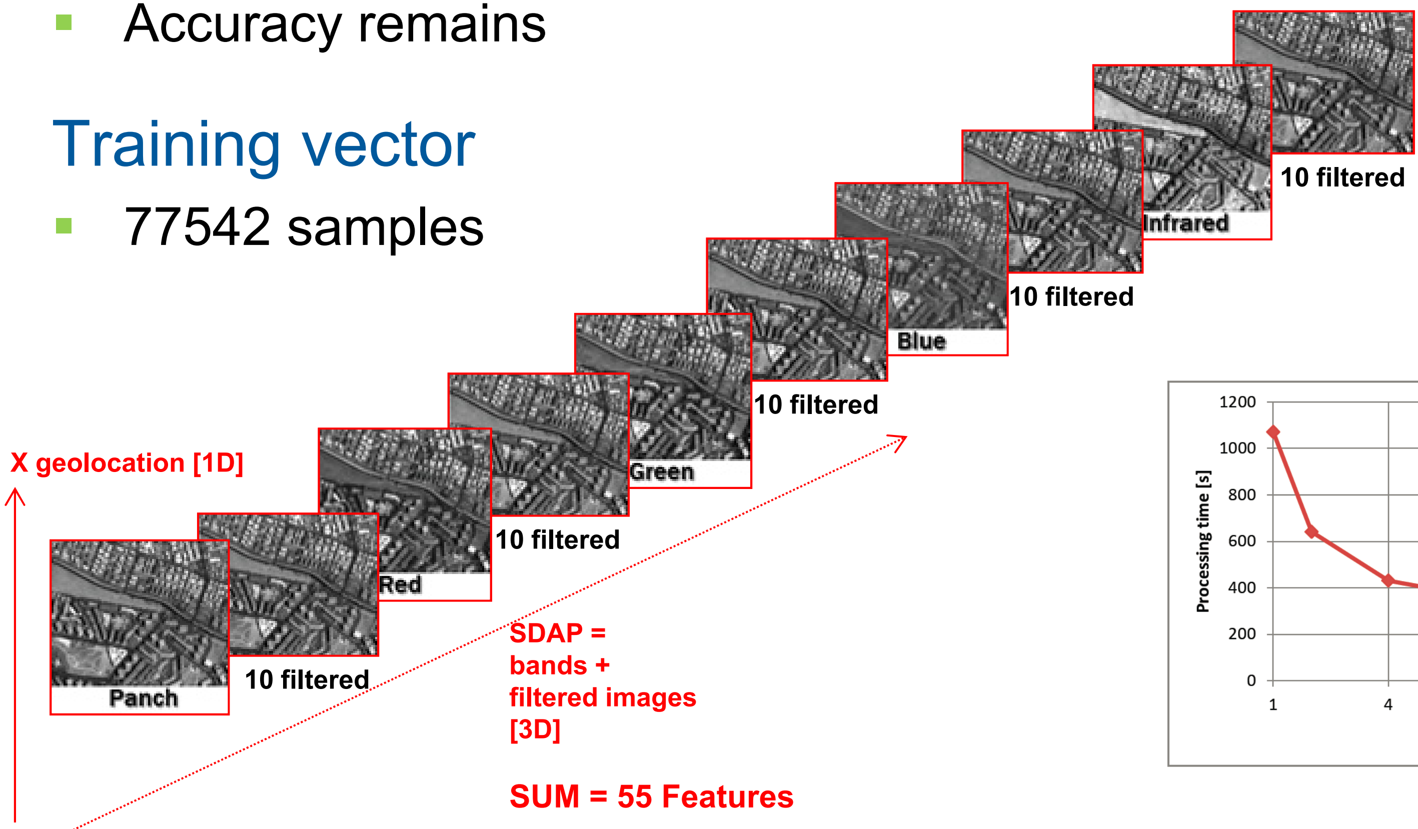
- Serial Matlab: ~1277 sec (~21 minutes)
- Parallel (16) Analytics: 220 sec (3:40 minutes)
- Accuracy remains

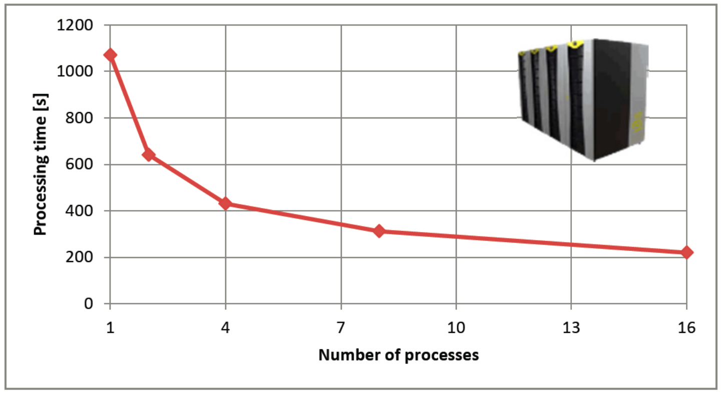


Manual work: Obtain the SDAP for all image bands using attribute ‘area‘ (10 thresholds)

Training vector

- 77542 samples





[12] Rome Analytics Results & job scripts

B2SHARE Store and Share Research Data

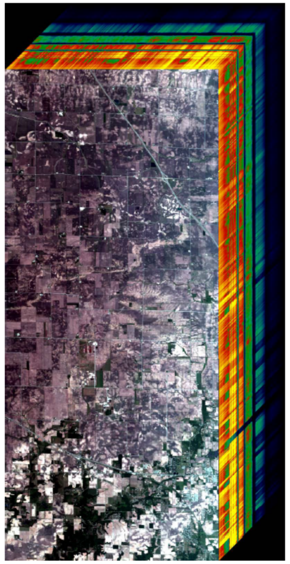
Parallel & Scalable SVM MPI Tool (10)

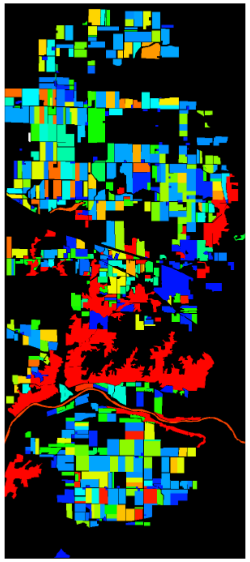
Another more challenging dataset: high number of classes

- Parallelization challenges: unbalanced class representations

Class		Number of samples		Class		Number of samples	
number	name	training	test	number	name	training	test
1	Buildings	1720	15475	27	Pasture	1039	9347
2	Corn	1778	16005	28	pond	10	92
3	Corn?	16	142	29	Soybeans	939	8452
4	Corn-EW	51	463	30	Soybeans?	89	805
5	Corn-NS	236	2120	31	Soybeans-NS	111	999
6	Corn-CleanTill	1240	11164	32	Soybeans-CleanTill	507	4567
7	Corn-CleanTill-EW	2649	23837	33	Soybeans-CleanTill?	273	2453
8	Corn-CleanTill-NS	3968	35710	34	Soybeans-CleanTill-EW	1180	10622
9	Corn-CleanTill-NS-Irrigated	80	720	35	Soybeans-CleanTill-NS	1039	9348
10	Corn-CleanTilled-NS?	173	1555	36	Soybeans-CleanTill-Drilled	224	2018
11	Corn-MinTill	105	944	37	Soybeans-CleanTill-Weedy	54	489
12	Corn-MinTill-EW	563	5066	38	Soybeans-Drilled	1512	13606
13	Corn-MinTill-NS	886	7976	39	Soybeans-MinTill	267	2400
14	Corn-NoTill	438	3943	40	Soybeans-MinTill-EW	183	1649
15	Corn-NoTill-EW	121	1085	41	Soybeans-MinTill-Drilled	810	7288
16	Corn-NoTill-NS	569	5116	42	Soybeans-MinTill-NS	495	4458
17	Fescue	11	103	43	Soybeans-NoTill	216	1941
18	Grass	115	1032	44	Soybeans-NoTill-EW	253	2280
19	Grass/Trees	233	2098	45	Soybeans-NoTill-NS	93	836
20	Hay	113	1015	46	Soybeans-NoTill-Drilled	873	7858
21	Hay?	219	1966	47	Swampy Area	58	525
22	Hay-Alfalfa	226	2032	48	River	311	2799
23	Lake	22	202	49	Trees?	58	522
24	NotCropped	194	1746	50	Wheat	498	4481
25	Oats	174	1568	51	Woods	6356	57206
26	Oats?	34	301	52	Woods?	14	130

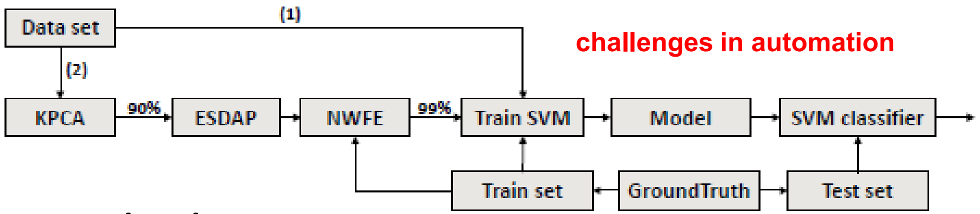
remote sensing cube & ground reference





G. Cavallaro, M. Riedel et al., Remote Sensing Journal – Big Data Special Issue, to be published

challenges in automation



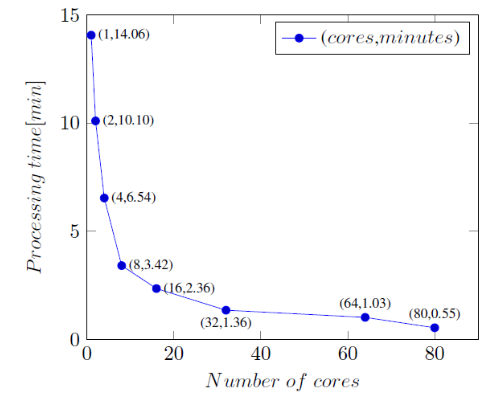
B2SHARE Store and Share Research Data

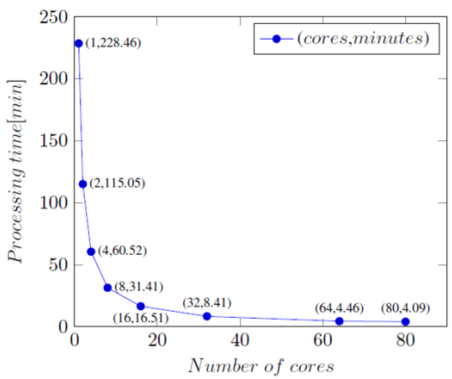
[20] Indian pines dataset, processed and raw

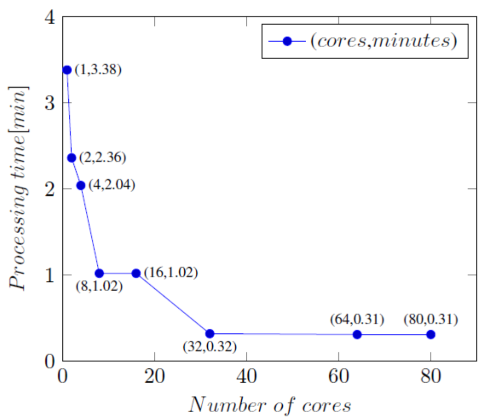
Parallel & Scalable SVM MPI Tool (11)

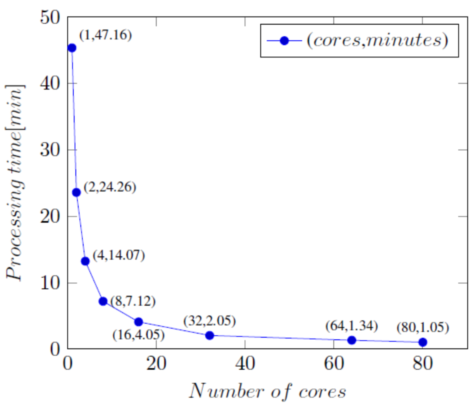
Another example dataset: high number of classes

- Parallelization benefits: major speed-ups, ~interactive (<1 min) possible









B2SHARE Store and Share Research Data

[21] Analytics Results (raw)

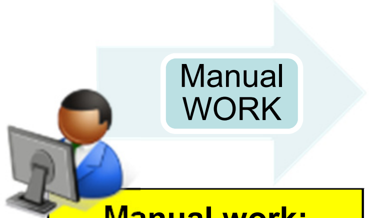
[22] Analytics Results (processed)

manual & serial activities (in minutes)

	kpca	esdap	nwfe	10x CSV	Training	Test	Total
(1) Scenario	0	0	0	4.47×10^3	10,45	71,08	4.55×10^3
(2) Scenario	5	15.38	1	529.55	1.37	23.25	575.55

'big data' is not always better data

	(1) Scenario	(2) Scenario
Number of features	200	30
Overall Accuracy (%)	40,68	77,96



Manual work: Trade-off between raw data processing and using feature extraction methods

Can we automate feature extraction mechanism to some degree?

G. Cavallaro, M. Riedel et al., Remote Sensing Journal – Big Data Special Issue, to be published

Parallel & Scalable SVM MPI Tool (12)

2x benefits of parallelization (shown in n-fold cross validation)

- Evaluation between Matlab (aka serial) and parallel piSVM
- 10x cross-validation (RBF kernel parameter and C, gridsearch)

raw dataset (serial)

γ / C	1	10	100	1000	10000
2	27.30 (109.78)	34.59 (124.46)	39.05 (107.85)	37.38 (116.29)	37.20 (121.51)
4	29.24 (98.18)	37.75 (85.31)	38.91 (113.87)	38.36 (119.12)	38.36 (118.98)
8	31.31 (109.95)	**39.68** (118.28)	39.06 (112.99)	39.06 (190.72)	39.06 (872.27)
16	33.37 (126.14)	39.46 (171.11)	39.19 (206.66)	39.19 (181.82)	39.19 (146.98)
32	34.61 (179.04)	38.37 (202.30)	38.37 (231.10)	38.37 (240.36)	38.37 (278.02)

processed dataset (serial)

γ / C	1	10	100	1000	10000
2	48.90 (18.81)	65.01 (19.57)	73.21 (20.11)	75.55 (22.53)	74.42 (21.21)
4	57.53 (16.82)	70.74 (13.94)	75.94 (13.53)	76.04 (14.04)	74.06 (15.55)
8	64.18 (18.30)	74.45 (15.04)	**77.00** (14.41)	75.78 (14.65)	74.58 (14.92)
16	68.37 (23.21)	76.20 (21.88)	76.51 (20.69)	75.32 (19.60)	74.72 (19.66)
32	70.17 (34.45)	75.48 (34.76)	74.88 (34.05)	74.08 (34.03)	73.84 (38.78)

raw dataset (parallel, 80 cores)

γ / C	1	10	100	1000	10000
2	27.26 (3.38)	34.49 (3.35)	39.16 (5.35)	37.56 (11.46)	37.57 (13.02)
4	29.12 (3.34)	37.58 (3.38)	38.91 (6.02)	38.43 (7.47)	38.43 (7.47)
8	31.24 (3.38)	**39.77 (4.09)**	39.14 (5.45)	39.14 (5.42)	39.14 (5.43)
16	33.36 (4.09)	39.61 (4.56)	39.25 (5.06)	39.25 (5.27)	39.25 (5.10)
32	34.61 (5.13)	38.37 (5.30)	38.36 (5.43)	38.36 (5.49)	38.36 (5.28)

processed dataset (parallel, 80 cores)

γ / C	1	10	100	1000	10000
2	75.26 (1.02)	65.12 (1.03)	73.18 (1.33)	75.76 (2.35)	74.53 (4.40)
4	57.60 (1.03)	70.88 (1.02)	75.87 (1.03)	76.01 (1.33)	74.06 (2.35)
8	64.17 (1.02)	74.52 (1.03)	**77.02 (1.02)**	75.79 (1.04)	74.42 (1.34)
16	68.57 (1.33)	76.07 (1.33)	76.40 (1.34)	75.26 (1.05)	74.53 (1.34)
32	70.21 (1.33)	75.38 (1.34)	74.69 (1.34)	73.91 (1.47)	73.73 (1.33)

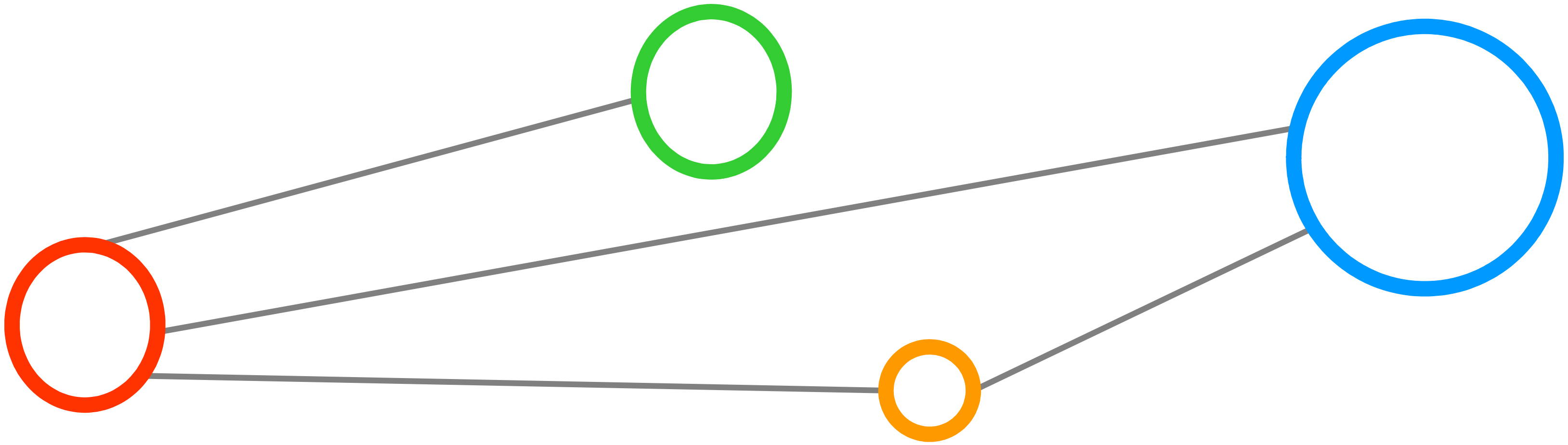
G. Cavallaro, M. Riedel et al., Remote Sensing Journal – Big Data Special Issue, to be published

B2SHARE Store and Share Research Data

[23] Analytics 10 fold cross-validation Results (raw)

[24] Analytics 10 fold cross-validation Results (processed)

Recent Research Directions

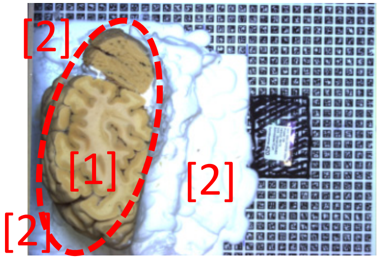


Recent Research Directions – Brain Data Classification

- **Build 'reconstructed brain (one 3d volume) that matches with sections & block images**
- **Understanding the 'sectioning of the brain' and support automation of reconstruction**

Smart Data Innovation Lab

1. Some 'pattern' exists
   - Image content classification (e.g. SVMs, RandomForst, etc.)
2. No exact mathematical formula exists
   - No precise formula for 'contour of the brain'
3. Dataset (next: 5 brains, >100.000 pixels, 2PB raw)
   - Block face images (of frozen brain tissue)
   - Every 20 micron (cut size), resolution: 3272 x 2469
   - ~ 14 MB / RGB image
   - ~ 8 MB / corresponding mask image ('groundtruth')
   - ~700 images → ~40 GB dataset

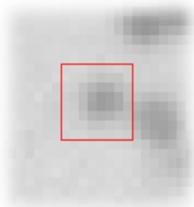


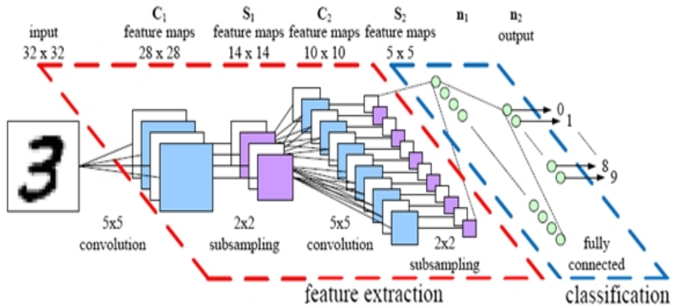
➢ **Research activities jointly with T. Dickscheid et al. (Juelich Institute of Neuroscience & Medicine)**

Recent Research Directions – Deep Learning

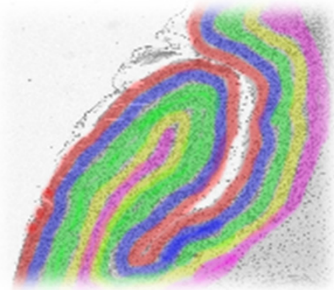
- **Investigate a pipeline for cell nuclei detection and tissue clustering**

1. Some ‘pattern‘ exists
   - Image content classification & clustering
2. No exact mathematical formula exists
   - No precise formula for ‘brain layers‘
3. Dataset – raw images exist
   - Needs to be properly prepared
   - Generate labeled data to learn from (manual tool supporting scientists)
   - Use Deep Learning (deep convolutional neural network, GPGPUs) to classify cell nuclei
   - Extract cell nuclei into 2D/3D point cloud
   - Cluster different brain areas by cell density (parallel DBSCAN)



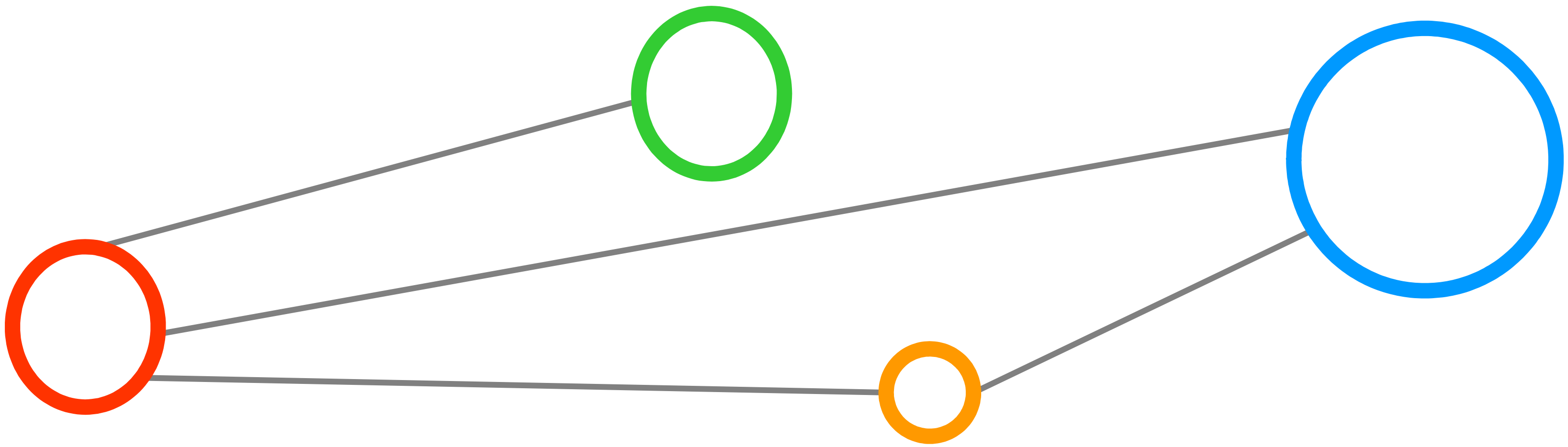


[13] Deep Learning Architecture



➢ **Research activities jointly with T. Dickscheid et al. (Juelich Institute of Neuroscience & Medicine)**

Conclusions



Conclusions

Scientific Peer Review is essential to progress in the field

- Work in the field needs to be guided & steered by communities
- NIC Scientific Big Data Analytics (SBDA) first step (learn from HPC)
- Towards enabling reproducability by uploading runs and datasets

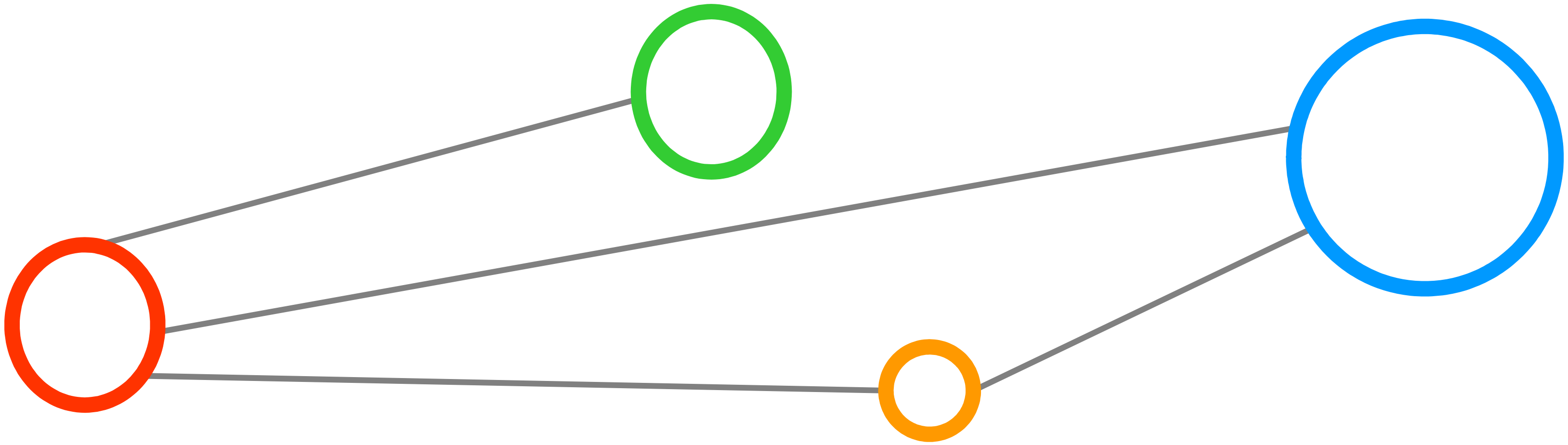
Selected SBDA benefit from parallelization

- Statistical data mining techniques able to reduce ‘big data‘ (e.g. PCA, etc.)
- Benefits in n-fold cross-validation & raw data, less on preprocessed data
- Two codes available to use and maintained @JSC: HPDBSCAN, piSVM

Number of Data Analytics et al. Technologies incredible high

- Thorough analysis and evaluation hard (needs different infrastructures)
- (Less) open source & working versions available, often paper studies
- Still evaluating approaches: HPC, map-reduce, Spark, SciDB, MaTex, …

References



References (1)

[1] R. Huber, M. Riedel et al., 'PANGAEA – Data Publisher for Earth & Environmental Science - Research data enters scholarly communication and big data analysis', presentation at EGU 2014, Vienna

[2] G. Cavallaro, J.A. Benediktsson and M. Riedel et al. 'Smart Data Analytics Methods for Remote Sensing Applications', in proceedings of IGARSS 2014

[3] 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

[4] Ester, Martin, et al. "A density-based algorithm for discoveringclusters in large spatial databases with noise." Kdd. Vol. 96. 1996.

[5] Patwary, Md Mostofa Ali, et al. "A new scalable parallel dbscanalgorithm using the disjoint-set data structure." High Performance Computing, Networking, Storage and Analysis (SC), 2012 InternationalConference for. IEEE, 2012

[6] PANGAEA Earth Science Data Collection, Online: http://www.pangaea.de/

[7] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20(3), pp. 273–297, 1995

[8] An Introduction to Statistical Learning with Applications in R, Online: http://www-bcf.usc.edu/~gareth/ISL/index.html

[9] Original piSVM tool, online: http://pisvm.sourceforge.net/

[10] B2SHARE data collection Rome Dataset, online: http://hdl.handle.net/11304/4615928c-e1a5-11e3-8cd7-14feb57d12b9

[11] G. Cavallaro, M. Mura, J.A. Benediktsson, L. Bruzzone 'A Comparison of Self-Dual Attribute Profiles based on different filter rules for classification', IEEE IGARSS2014, Quebec, Canada

[12] B2SHARE data collection piSVM1.2 Analytics JUDGE Cluster Rome Images 55 Features, online: http://hdl.handle.net/11304/6880662c-1edf-11e4-81ac-dcbd1b51435e

References (2)

[13] Deep Learning Architecture, Online: http://parse.ele.tue.nl/cluster/2/CNNArchitecture.jpg

[14] UNICORE, online: http://www.unicore.eu

[15] DOE ASCAC Report, 'Synergistic Challenges in Data-Intensive Science and. Exascale Computing'

[16] M. Riedel and P. Wittenburg et al. 'A Data Infrastructure Reference Model with Applications: Towards Realization of a ScienceTube Vision with a Data Replication Service', 2013

[17] Shearer C., 'The CRISP-DM model: the new blueprint for data mining', J Data Warehousing (2000); 5:13—22.

[18] Jeremy Ginsburg et al., 'Detecting influenza epidemics using search engine query data', Nature 457, 2009

[19] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, 'The Parable of Google Flu: Traps in Big Data Analysis', Science Vol (343), 2014

[20] B2SHARE data collection, remote sensing indian pines images, Online: http://hdl.handle.net/11304/7e8eec8e-ad61-11e4-ac7e-860aa0063d1f

[21] B2SHARE data collection, piSVM remote sensing indian pines analytics results (raw), Online: http://hdl.handle.net/11304/c06a8c7e-fe6c-11e4-8a18-f31aa6f4d448

[22] B2SHARE data collection, piSVM remote sensing indian pines analytics results (processed), Online: http://hdl.handle.net/11304/c528998e-ff7c-11e4-8a18-f31aa6f4d448

[23] B2SHARE data collection, Analytics 10 fold cross-validation (raw), Online: http://hdl.handle.net/11304/163ba8e8-fe60-11e4-8a18-f31aa6f4d448

[24] B2SHARE data collection, Analytics 10 fold cross-validation (processed), Online: http://hdl.handle.net/11304/5bba8e36-fe63-11e4-8a18-f31aa6f4d448

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, Juelich

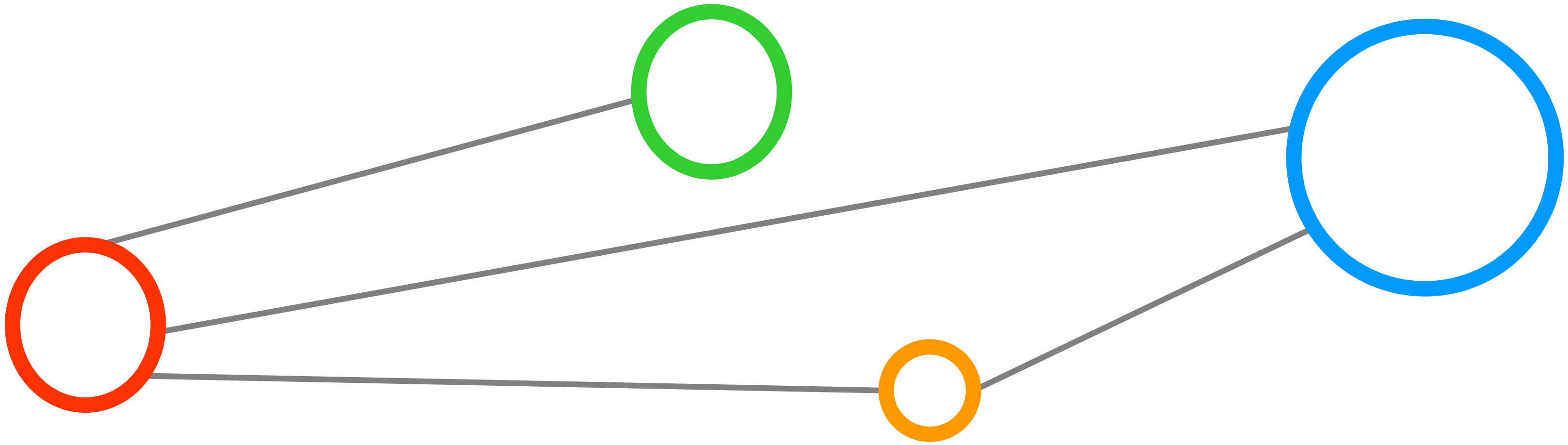
Selected Members of the Research Group on High Productivity Data Processing

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

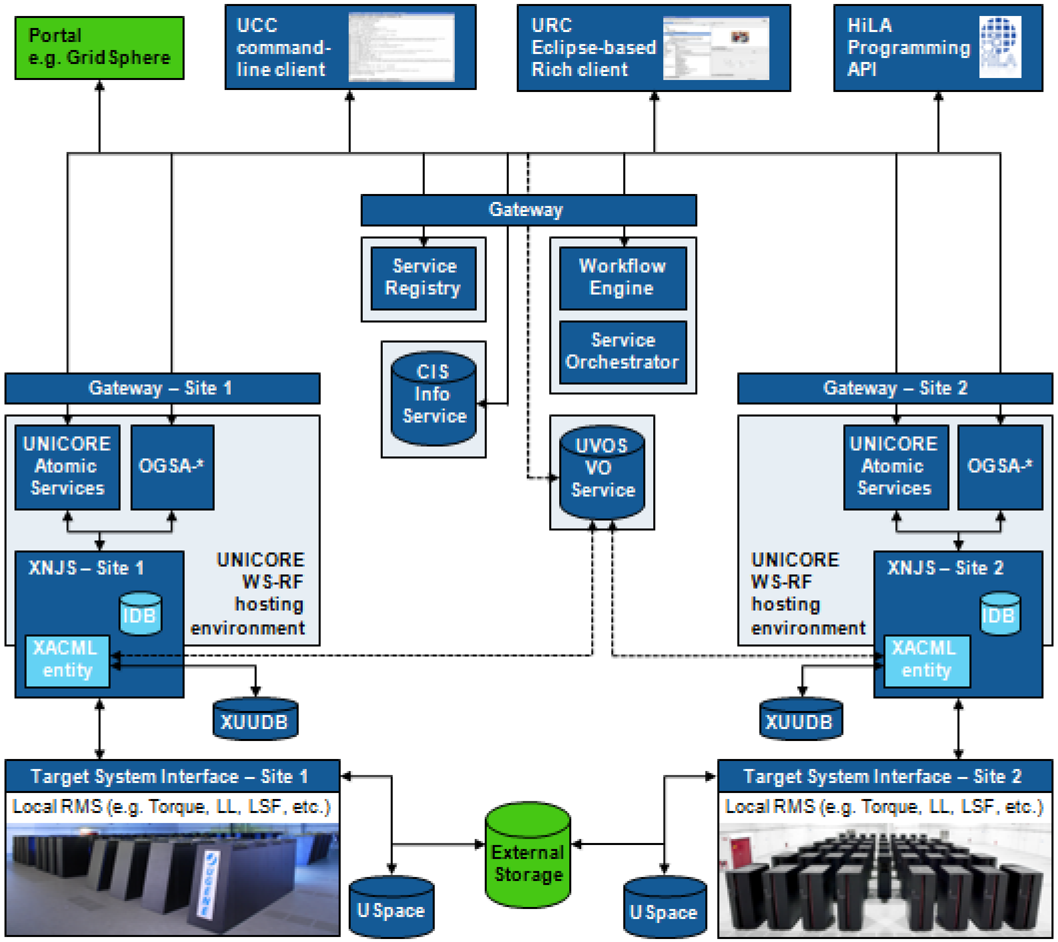
Thanks

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

Selected Backup Slides for Discussions

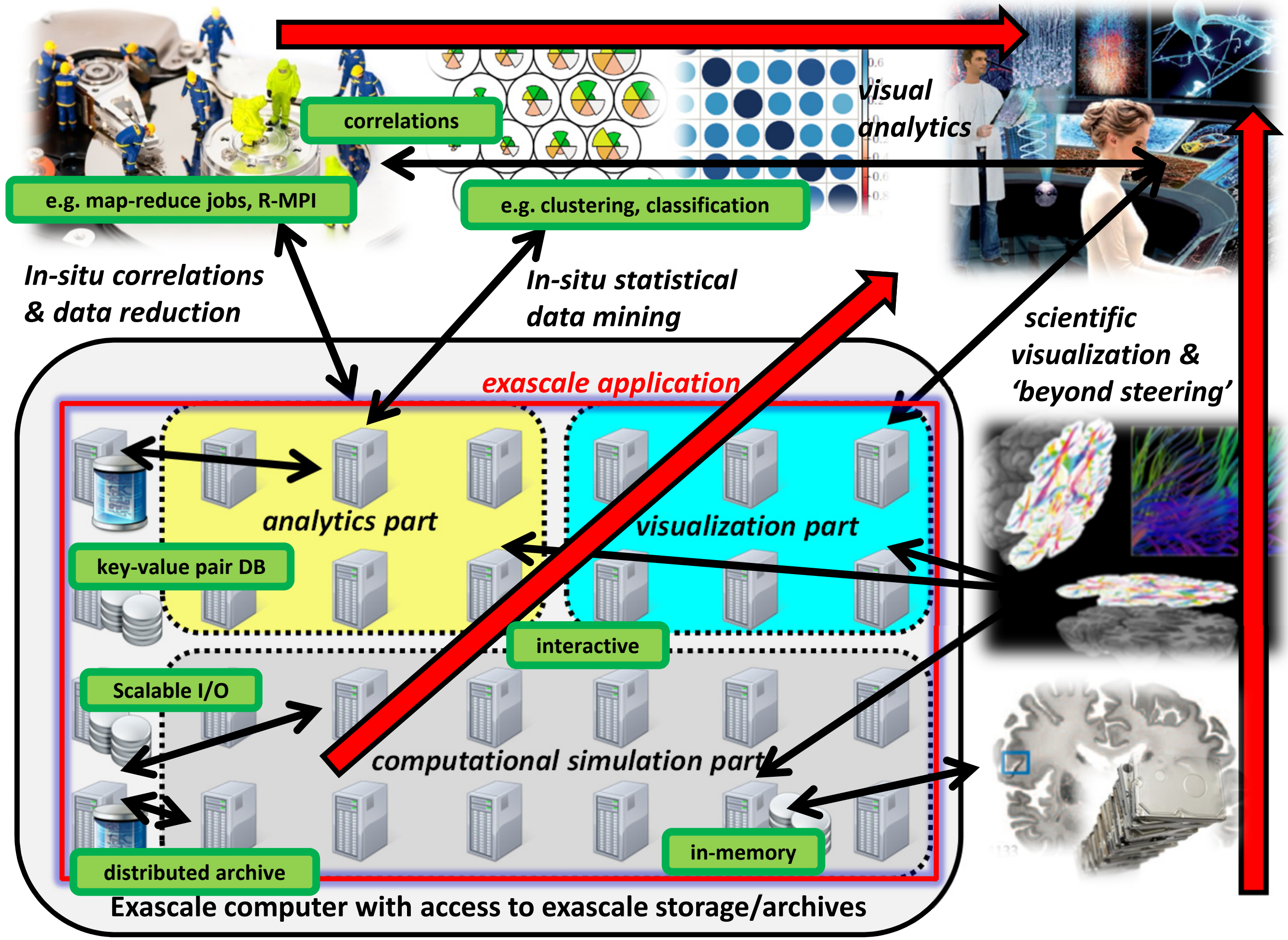


Distributed Large-scale Data Management



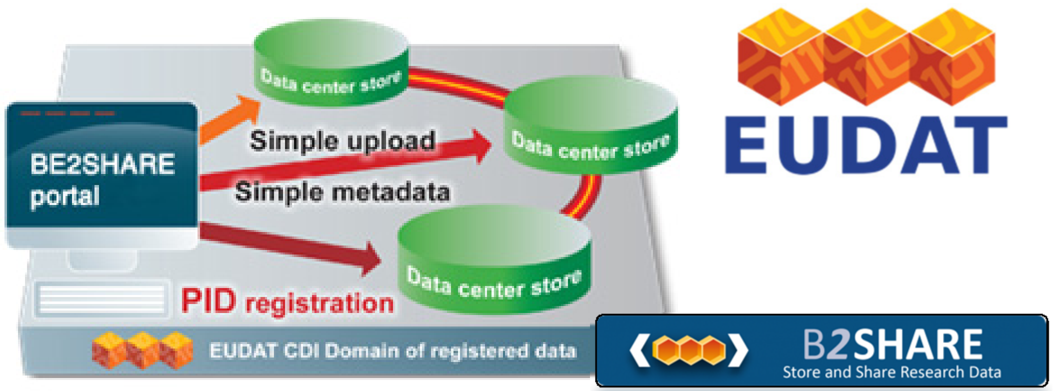
[14] UNICORE.eu

In-Situ Analytics for HPC & Exascale



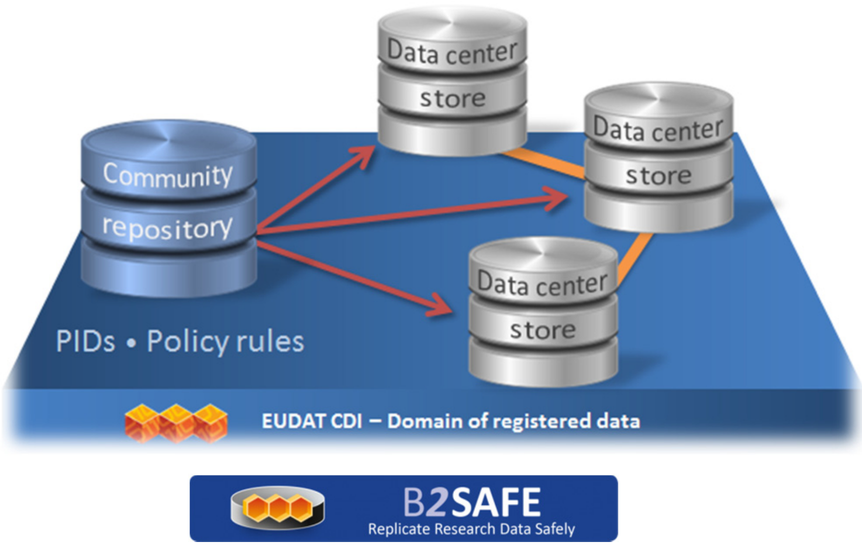
[15] Inspired by ASCAC DOE report

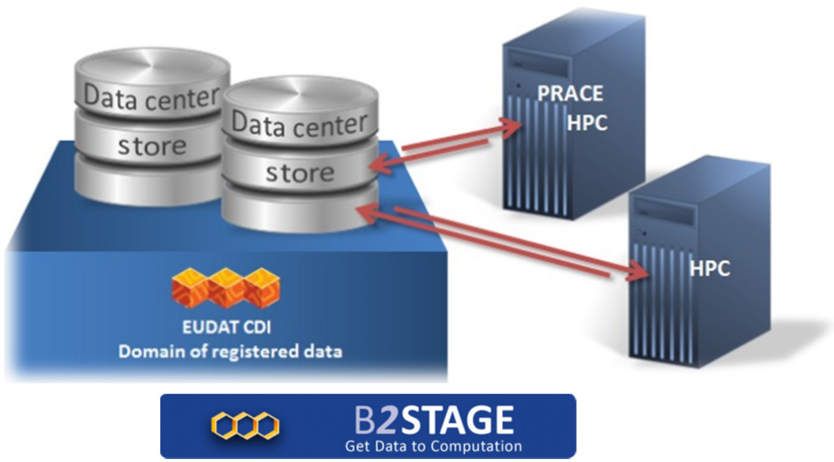
Tools for Large-scale Distributed Data Management



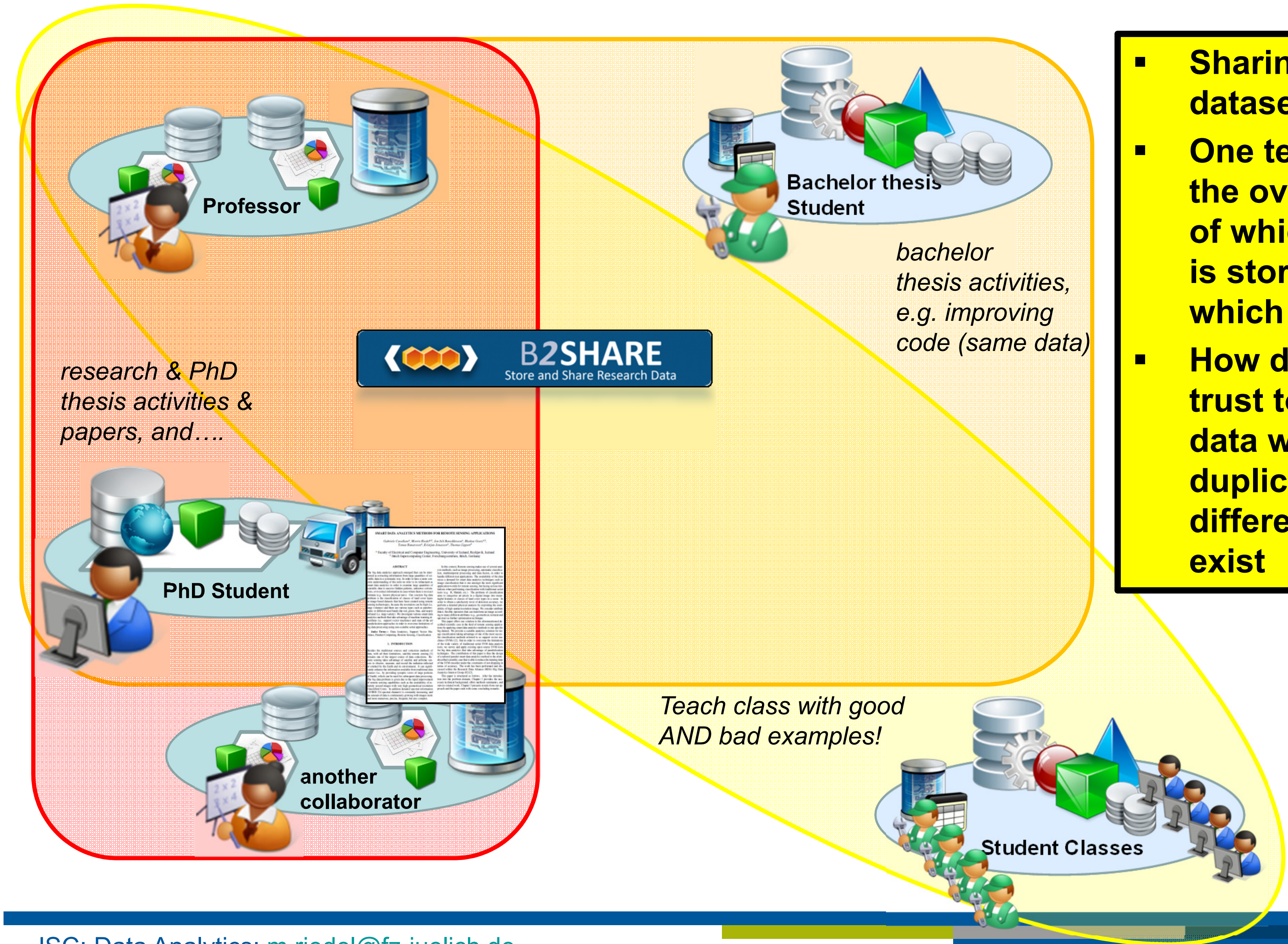
[16] *M. Riedel & P. Wittenburg et al.*

- **Useful tools for data-driven scientists & HPC users**



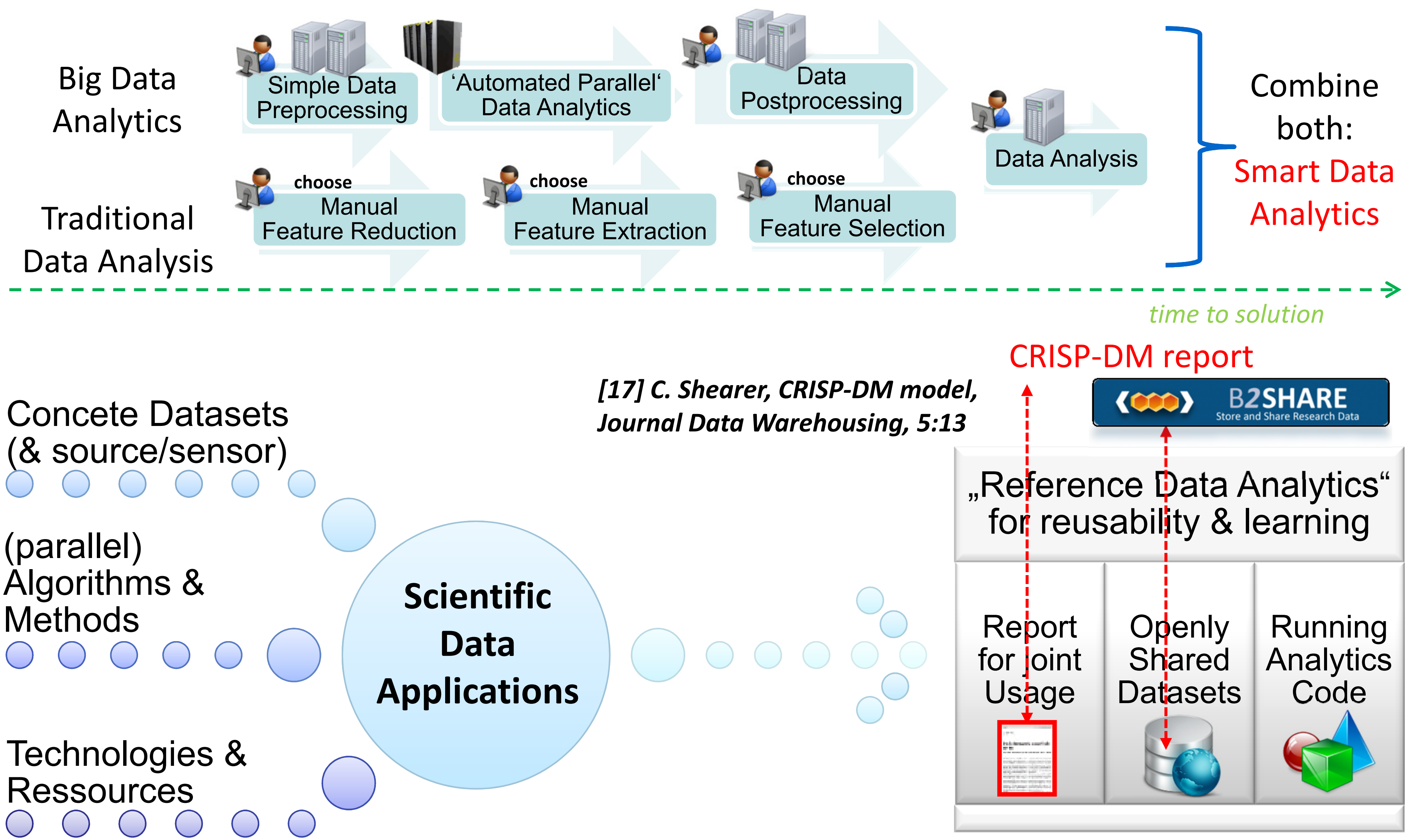


Need for Sharing & Reproducability in HPC – Example



- **Sharing different datasets is key**
- **One tend to loose the overview of which data is stored on which platform**
- **How do we gain trust to delete data when duplicates on different systems exist**

Smart Data Analytics Process



Selected Research Data Alliance (RDA) Activities

- **Big Data Analytics Interest Group – Establish something like UCI machine learning repository, but for big data analytics…**

RDA
RESEARCH DATA ALLIANCE

[2] G. Cavallaro and M. Riedel et al. 'Smart Data Analytics Methods for Remote Sensing Applications', IGARSS 2014

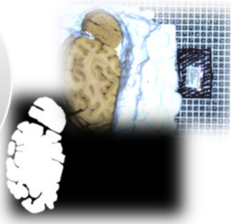
Future Grid

Twister
Iterative MapReduce

πSvM

Parallel Brain Data Analytics

scikits learn



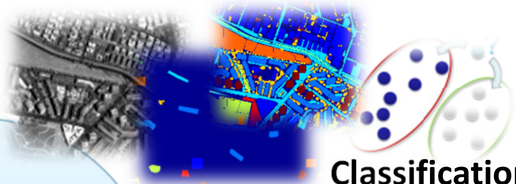
Sattelite Data (Quickbird)

Parallel Support Vector Machines (SVM)

πSvM

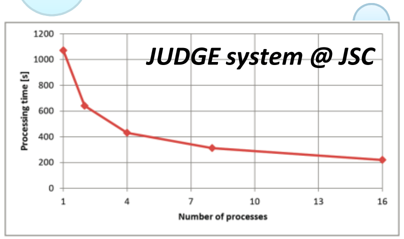
HPC & MPI

Classification Study of Land Cover Types



Classification++

'Best Practices'

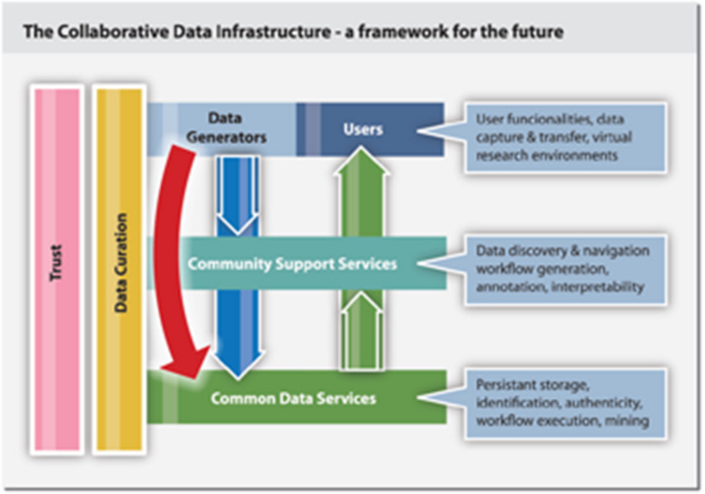


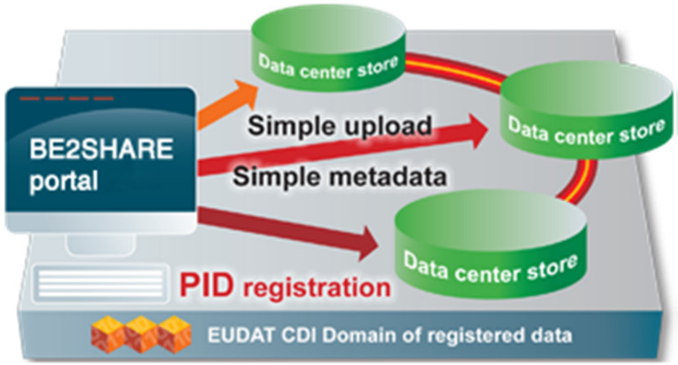
„Reference Data Analytics" for reusability & learning

CRISP-DM Report	Openly Shared Datasets	Running Analytics Code

➢ **Research activities with Gabriele Cavallaro (PhD thesis, UoIceland) on Self Dual Attribute Profile**

Reproducability Example in Data-driven Science (1)





- **Having this tool available on the Web helps tremendously to save time for no research tasks**
- **Using the tool enables to focus better on the research tasks**

Reproducability Example in Data-driven Science (2)

- **Sharing pre-processed data**
- **LibSVM format**
- **Training and Testing Datasets**
- **Different setups for analysis (SDAP on All or SDAP on Panchromatic)**

EUDAT

B2SHARE
Store and Share Research Data

Information | Comments | Reviews | Usage statistics

Rome data set OK

22 May 2014
http://b2share.eudat.eu

Abstract: Attribute area

The record appears in these collections:
Generic

Files

Name	Date	Size	
sdap_area_panch_training.el	22 May 2014	12.7 MB	Download
sdap_area_all_training.el	22 May 2014	46.7 MB	Download
sdap_area_panch_test.el	22 May 2014	114.8 MB	Download
sdap_area_all_test.el	22 May 2014	420.0 MB	Download

Rate this document:

Export

Export as BibTeX, MARC, MARCXML, DC, EndNote, NLM, RefWorks

Metadata

PID:	http://hdl.handle.net/11304/4615928c-e1a5-11e3-8cd7-14feb57d12b9
Publication:	http://b2share.eudat.eu
Publication Date:	2014-05-22
Uploaded by:	cavallaro.gabriele@gmail.com
Domain:	generic
Checksum:	16ba6c2e80c98859d0f9c044c73d1ec976b68c788e0681e3932d91e5a2a029c1

https://b2share.eudat.eu/record/86
B2Share

Reproducability Example in Data-driven Science (3)

Simple download from http using the wget command

```
mriedel@judge:~/bigdata> ls -al
total 640
drwxrwxrwx 21 mriedel  zam  32768 2014-09-17 22:20 .
drwxr-xr-x 19 mriedel  zam  32768 2014-09-18 11:49 ..
drwxr-xr-x  2 mriedel  zam  32768 2014-06-19 07:17 102-salinasindian
drwxr-xr-x  2 mriedel  zam    512 2014-06-19 20:11 107-salinasrescaled
drwxr-xr-x  2 mriedel  zam    512 2014-07-08 17:14 111-romemultispectral
drwxr-xr-x  2 mriedel  zam    512 2014-07-10 11:46 112-romeoriginalbands
drwxr-xr-x  2 mriedel  zam    512 2014-09-17 22:31 120-indianpine
drwxr-xr-x  2 mriedel  zam    512 2014-09-17 22:14 121-salinas
drwxr-xr-x  2 mriedel  zam    512 2014-09-17 22:19 122-salinas2
drwxr-xr-x  2 mriedel  zam    512 2014-09-17 22:24 123-indianpine2
drwxr-xr-x  2 mriedel  zam    512 2014-07-09 11:03 86-romeok
drwxr-xr-x  2 mriedel  zam  32768 2014-06-10 18:51 bigindianpines
                 mriedel  zam  32768 2014-05-28 10:59 indian
                 mriedel  zam  32768 2014-06-10 20:48 indianpinesreduced
                 mriedel  zam    512 2014-07-28 17:53 mnist-576-rbf
                 skoehnen inml   512 2014-07-29 16:35 pli-blockface
                 mriedel  zam  32768 2014-06-25 11:09 rome-ok
                 mriedel  zam    512 2014-07-08 13:29 rome-ok-copy
                 mriedel  zam  32768 2014-06-03 14:24 salinas
                 mriedel  zam  32768 2014-06-16 16:50 salinasindianrev
                 mriedel  zam  32768 2014-06-10 15:47 salinas-new
```

…other open B2SHARE datasets

…before adopting B2SHARE regularly

- **Simple Download from http using wget**
- **Well defined directory structures**

Reproducability Example in Data-driven Science (4)

Make a short note in your directory linking back to B2SHARE

```
mriedel@judge:~/bigdata> cd 86-romeok/
mriedel@judge:~/bigdata/86-romeok> ls -al
total 580320
drwxr-xr-x  2 mriedel zam       512 2014-07-09 11:03 .
drwxrwxrwx 21 mriedel zam     32768 2014-09-17 22:20 ..
-rw-r--r--  1 mriedel zam        35 2014-07-09 11:01 b2share.txt
-rw-r--r--  1 mriedel zam 419974873 2014-05-22 13:36 sdap_area_all_test.el
-rw-r--r--  1 mriedel zam  46652874 2014-05-22 13:36 sdap_area_all_training.el
-rw-r--r--  1 mriedel zam 114763982 2014-05-22 13:36 sdap_area_panch_test.el
-rw-r--r--  1 mriedel zam  12745692 2014-05-22 13:36 sdap_area_panch_training.el
mriedel@judge:~/bigdata/86-romeok> more b2share.txt
https://b2share.eudat.eu/record/86
mriedel@judge:~/bigdata/86-romeok>
```

- **Enables the trust to delete data if necessary (working against big data)**
- **Link back to B2SHARE for quick checks and file that links back fosters trust**

Reproducability Example in Data-driven Science (5)

```
mriedel@judge:~> ls -al
total 111840
drwxr-xr-x  19 mriedel zam     32768 2014-09-18 11:49 .
drwxr-xr-x 214 root    sys     32768 2014-09-12 09:02 ..
-rw-r--r--   1 mriedel zam 113233920 2014-08-08 10:35 115-RunsMatthiasStable.tar
drwxr-xr-x   3 mriedel zam     32768 2014-06-03 16:24 ann-0.1
drwxr-xr-x   3 mriedel zam     32768 2014-06-03 17:02 ann-0.2
drwxr-xr-x   3 mriedel zam     32768 2014-06-04 14:42 ann-0.3
drwxr-xr-x   2 mriedel zam     32768 2014-06-16 19:12 ann-0.4
drwxr-xr-x   2 mriedel zam     32768 2014-06-16 19:24 ann-0.4-orig
drwxr-xr-x   2 mriedel zam     32768 2014-06-19 08:38 ann-0.5
drwxr-xr-x   6 mriedel zam     32768 2014-06-25 00:52 ann-0.6
drwxr-xr-x   4 mriedel zam     32768 2014-06-19 16:31 ann-0.6-scal
drwxr-xr-x   2 mriedel zam     32768 2014-06-24 17:02 ann-0.7
-rw-------   1 mriedel zam      1797 2014-05-12 13:51 .bashrc
drwxrwxrwx  21 mriedel zam     32768 2014-09-17 22:20 bigdata
drwxr-xr-x   3 mriedel zam       512 2014-06-19 09:34 .config
drwxr-xr-x   3 mriedel zam     32768 2014-06-03 14:38 .emacs.d
-rw-------   1 mriedel zam      1864 2014-05-12 13:51 .kshrc
drwxr-xr-x   3 mriedel zam     32768 2014-05-12 14:56 pisvm-1.2
drwxr-xr-x   5 mriedel zam     32768 2014-09-18 11:49 pisvm-1.2.1
drwxr-xr-x   3 mriedel zam       512 2014-07-09 14:51 pisvm-1.2-refactored
-rw-------   1 mriedel zam      2686 2014-05-12 13:51 .profile
-rw-------   1 mriedel zam     22490 2014-09-18 11:51 .sh_history
drwx------   2 mriedel zam     32768 2014-05-12 14:38 .ssh
drwxr-xr-x   2 mriedel zam     32768 2014-05-12 14:39 transfers
-rw-------   1 mriedel zam     19526 2014-09-18
-rw-------   1 mriedel zam       204 2014-09-17
mriedel@judge:~>
```

… a bachelor project

… different versions of a parallel neural network code (another classification technique)

… different versions of a parallel support vector machine code

- **True reproducability needs: (1) datasets; (2) technique parameters (here for SVM); and (3) correct versions of algorithm code**

Classical Machine Learning

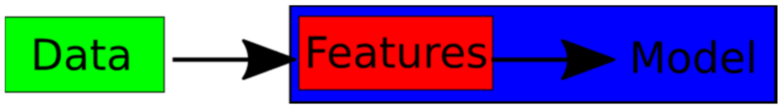
Dealing with Big Data in traditional Machine Learning

- Define Features to learn from ?!
- Transform data into supported format ?!
- How to reduce dimensions ?!
- How to parallelize ?!

Deep Learning

Dealing with Big Data in Deep Learning

- Define Features to learn from
  - → Automatically learn how to define features
- Transform data into supported format
  - → Adopt the model to your data
- How to reduce dimensions
  - → Automatically reduce dimensions in every hidden layer
- How to parallelize
  - → Naturally the brain is parallel, so Artificial Neural Networks are!



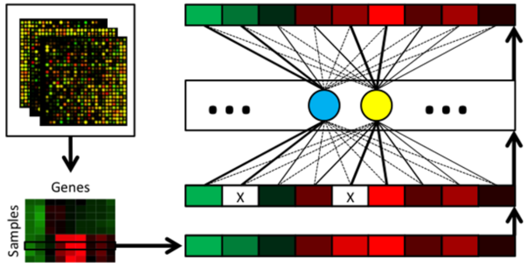
- A. Ng, Google Brain

Deep Learning (3)

Deep Learning in Computational Biomedicine

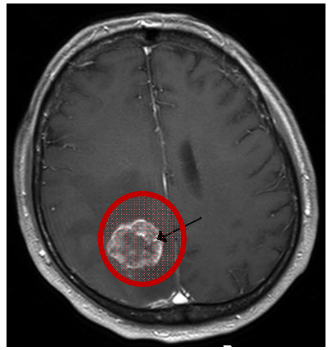
Genome Analysis

- Find high level features on low level –omics data



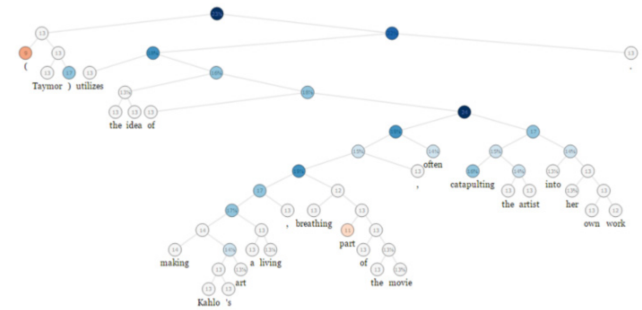
Medical Image Analysis

- Use 2D (or 3D) structure of the data for classification



Unstructured Data Analysis

- Use DL for text analysis to classify patient data, drug recommendations by users, …



Etc…

Deep Learning (4)

Deep Learning Packages

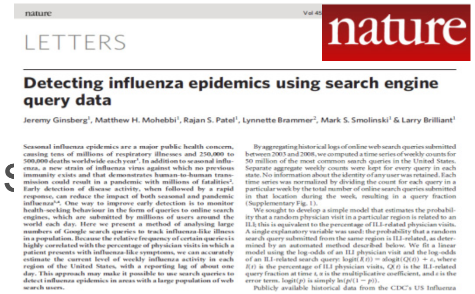
There exists several frameworks for deep neural networks

- Pylearn2
  - Python tool on the top of the Theano python library
  - Easy configuration of data, model, learning via YAML files
  - CUDA support for accelerated calculations
  - Jobman for parallel cross validation
- Caffe
  - C++ implementation with python & matlab wrappers
  - CUDA acceleration
- DL4J
  - Java implementation of Deep Learning
  - CUDA + Hadoop support

Chances and Pitfalls for ‘Scientific Big Data Analytics‘

~2009 – H1N1 Virus Made Headlines

- Nature paper from Google employees
- Explains how Google is able to predict fast winter flus
- Not only on national scale, but down to regions
- Possible via logged big data – ‘search queries‘


nature

LETTERS

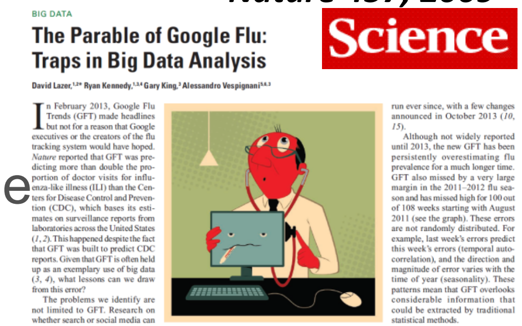
Detecting influenza epidemics using search engine query data

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant

[18] Jeremy Ginsburg et al., ‘Detecting influenza epidemics using search engine query data’, Nature 457, 2009

~2014 – The Parable of Google Flu

- Large errors in flu prediction & lessons learned
- (1) Dataset: Transparency & replicability impossible
- (2) Study the algorithm since they keep changing
- (3) It‘s not just about size of the data


BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Science

David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani

[19] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, ‘The Parable of Google Flu: Traps in Big Data Analysis’, Science Vol (343), 2014

- **Big data is not always better data – Think about difference of causality vs. correlation**

Location-based Social Network-based Health Analytics

Scientific Domain Area

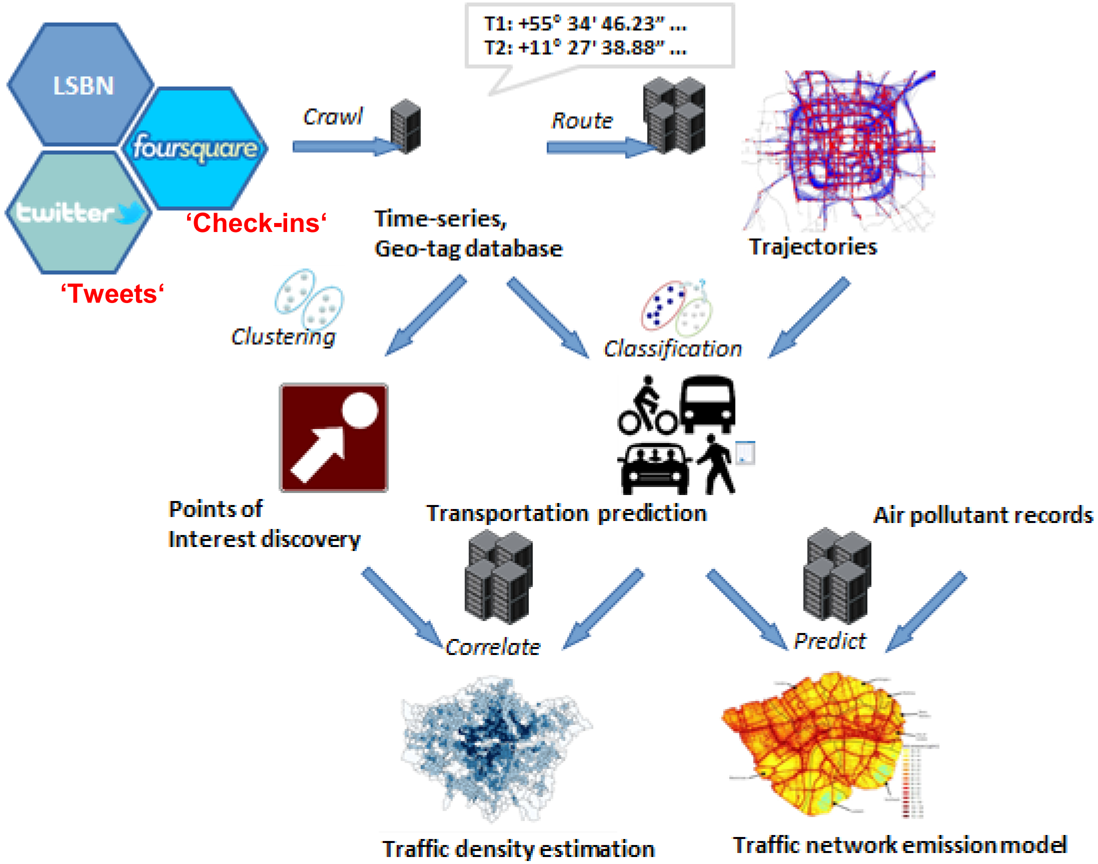
- Smart Cities approaches compined with Health Analytics Research

Scientific Outcome

- Traffic density estimation
- Network emission model

Location-based Social Networks (LBSN) Data

- Open data sources: Twitter & Foursquare
- Plan: Validation with real measurements in cities



➢ **Research activities with Markus Goetz (PhD thesis) – Juelich Supercomputing Centre, UoIceland**