Scalable Developments for Big Data Analytics in Remote Sensing

Dr.-Ing. Morris Riedel et al.

Research Group Leader, Juelich Supercomputing Centre

Adjunct Associated Professor, University of Iceland

Gabriele Cavallaro, Jon Atli Benediktsson

University of Iceland

Philipp Glock, Christian Bodenstein, Markus Goetz

Juelich Supercomputing Centre

Remote Sensing: Understanding the Earth for a Safer World

IGARSS 2015

July 26-31, 2015 • Milan, Italy

HELMHOLTZ ASSOCIATION

Federated Systems and Data Division

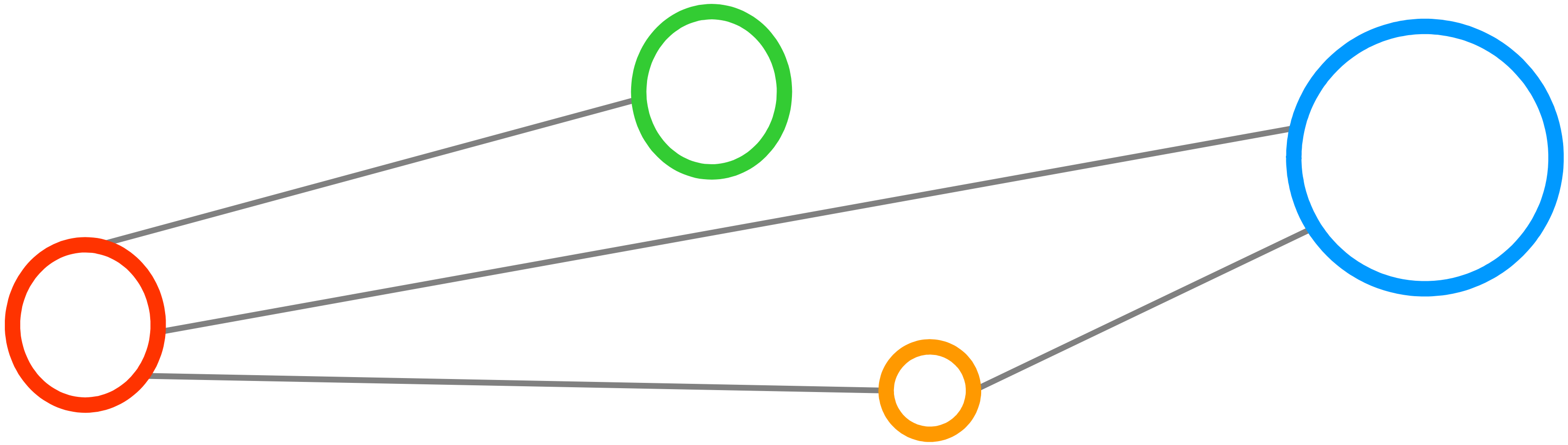
Research Group

High Productivity Data Processing

JÜLICH FORSCHUNGSZENTRUM

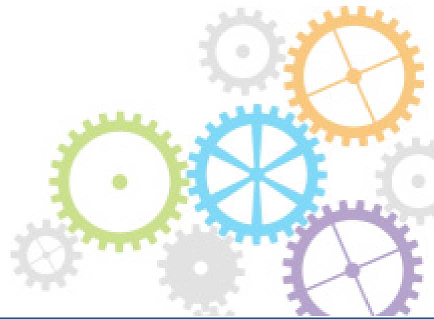
UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING, MECHANICAL ENGINEERING AND COMPUTER SCIENCE

Outline

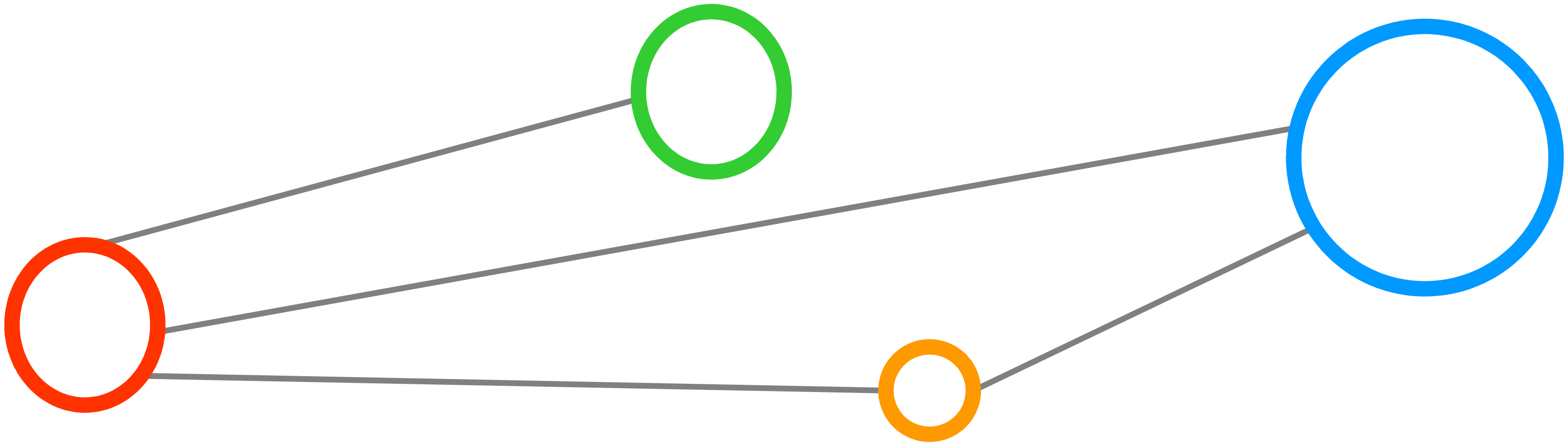


Outline

- Big Data Analytics
  - Driven by Scientific & Engineering Demands
  - Supervised Learning – Classification Method
- Scalable & Parallel Developments
  - Short Survey of Related Work
  - Enable Direct Parallelization & Cross-validation with piSVM
  - Scale Parallelization with the Cascade SVM Approach & GPGPUs
  - Remote Sensing Applications & Data in Context
- Recent Research Directions
  - Parallel Clustering & ‘Brain Data Analytics‘
- Conclusions
- References



Big Data Analytics

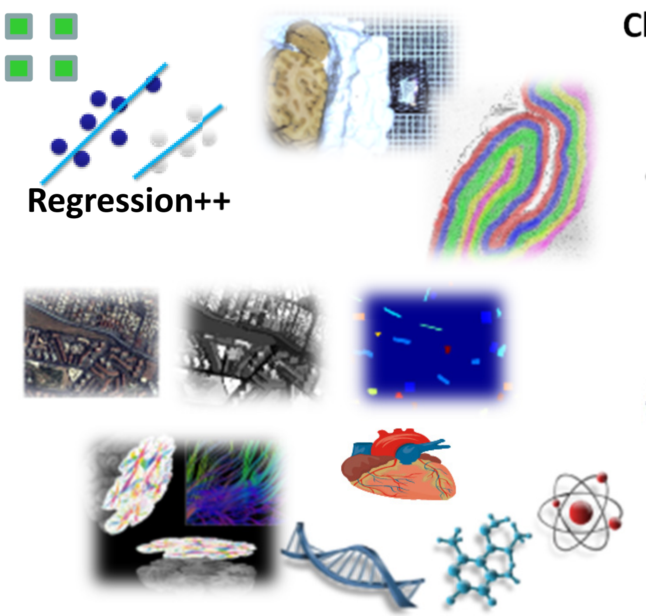


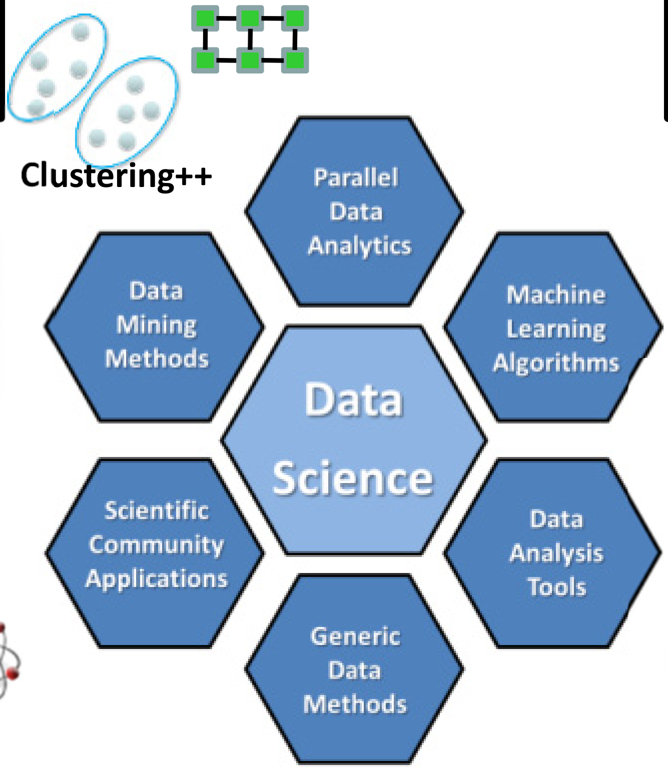
Big Data Analytics

‘Big Data Analytics’ is an ‘interesting mix’ of distinct approaches

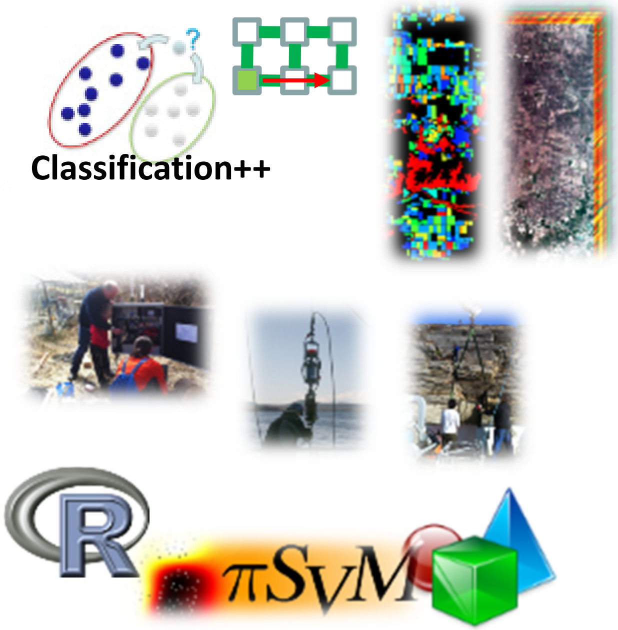
- Analytics: Whole methodology; Analysis: data investigation process itself
- ‘Big‘: scalable processing methods and underlying parallel infrastructure

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



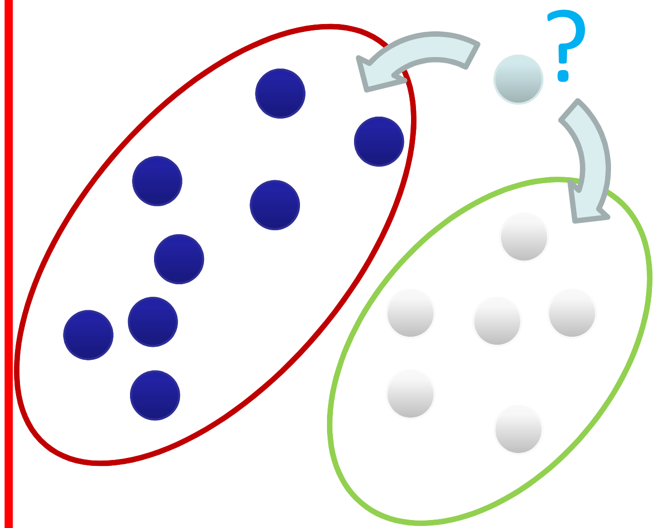


- **Concrete ‘big data‘: large earth science data**



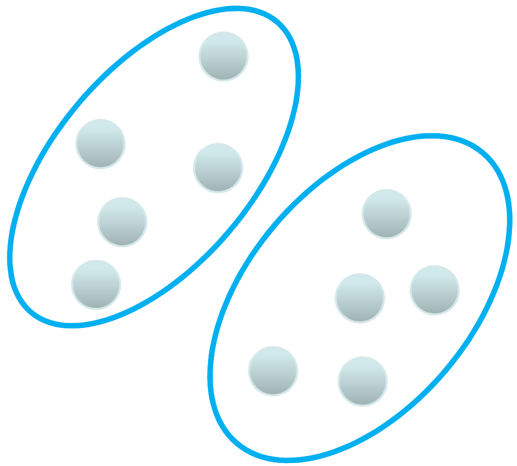
Learning From Data – Classification Technique

Classification



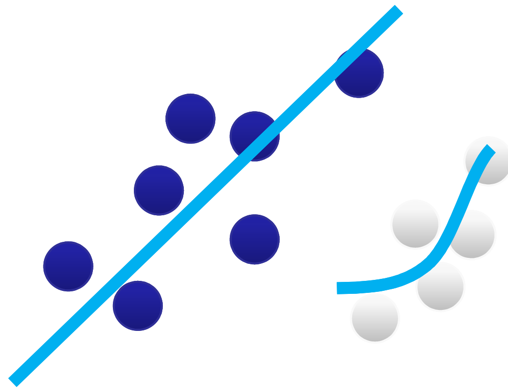
- Groups of data exist
- New data classified to existing groups

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 Method

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

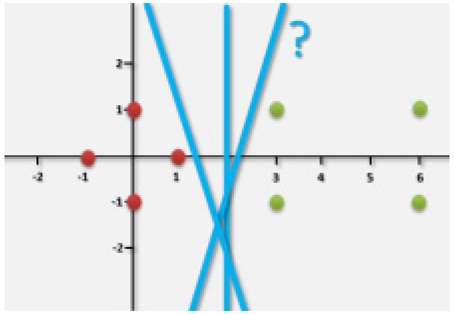
- Use of the Bayes theorem with strong/naive independence between features

Supervised Learning – SVM Classification Method

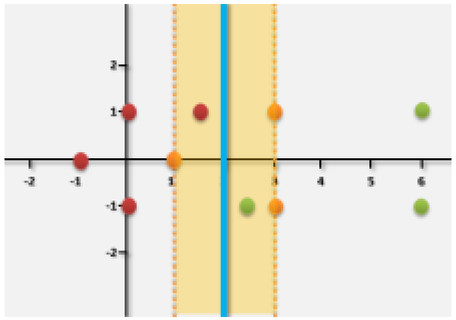
SVM Algorithm Approach

[1] 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)

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

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

(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)

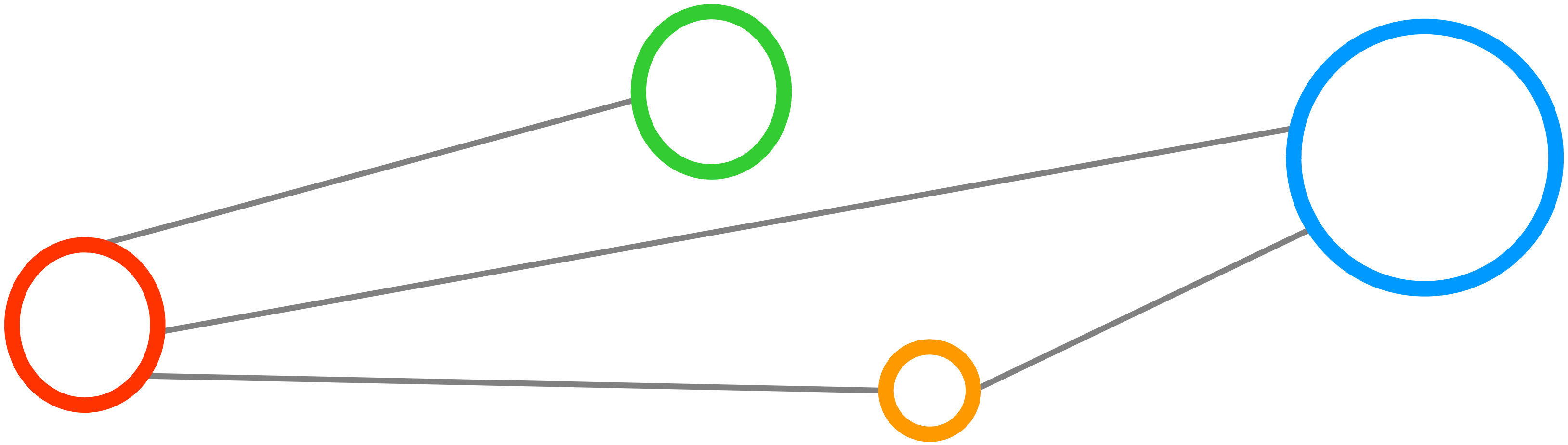
$$0 \leq \alpha_i \leq C$$

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

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

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

Scalable and Parallel Developments



Short Survey of Related Work

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

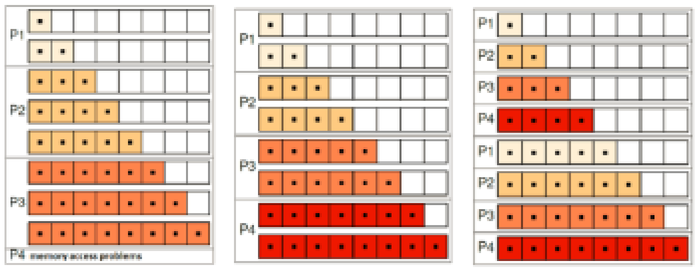
[2] M. Goetz, M. Riedel et al., '' 'On Parallel and Scalable Classification and Clustering Techniques for Earth Science Datasets', 6th Workshop on Data Mining in Earth System Science, International Conference of Computational Science (ICCS), Reykjavik

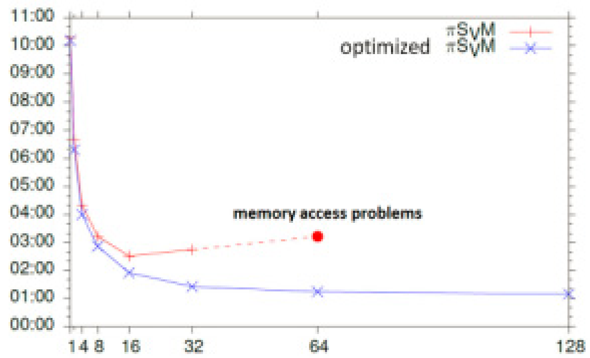
Parallel & Scalable piSVM MPI Tool – Tunings

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.)

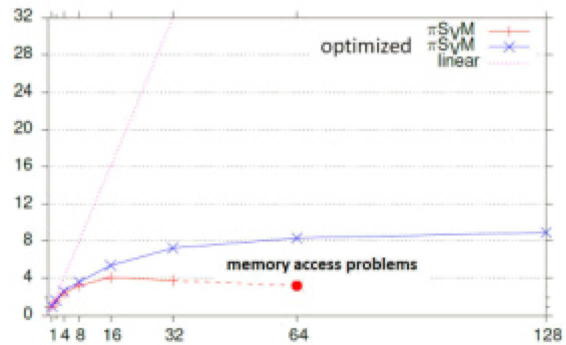
[3] piSVM Website, 2011/2014 code





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



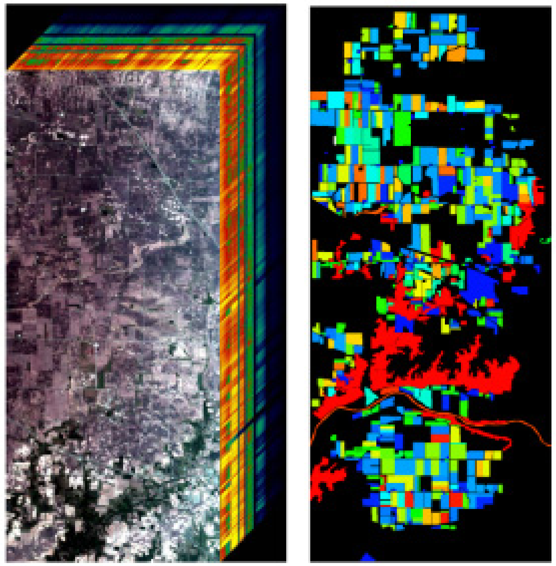
Parallel & Scalable piSVM MPI Tool – Dataset

Another more challenging dataset: high number of classes

- Parallelization challenges: unbalanced class representations, mixed pixels

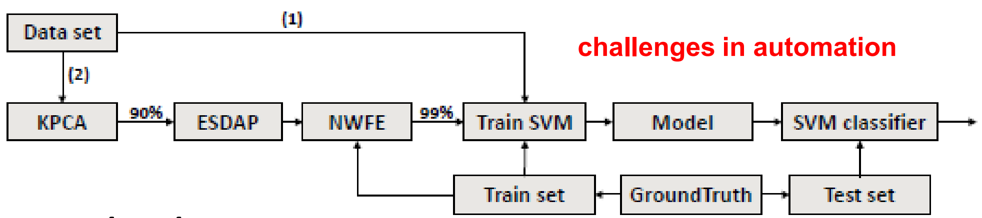
Class number	Class name	Number of samples training	Number of samples test	Class number	Class name	Number of samples training	Number of samples test
1	Buildings	1720	15475	27	Pasture	1039	9347
2	Corn	1778	16005	28	pond	10	92
3	Corn?	16	142	29	Soybeans	939	8452
4	Corn-EW	51	463	30	Soybeans?	89	805
5	Corn-NS	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



[4] G. Cavallaro, M. Riedel et al., IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, to be published

challenges in automation



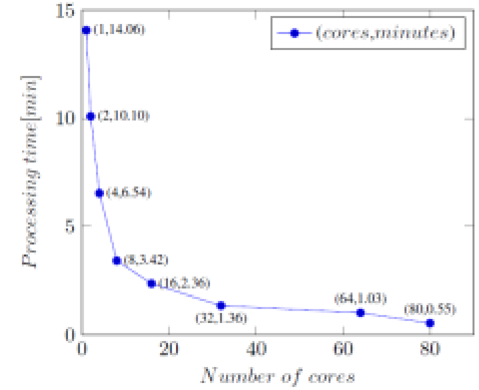
B2SHARE Store and Share Research Data

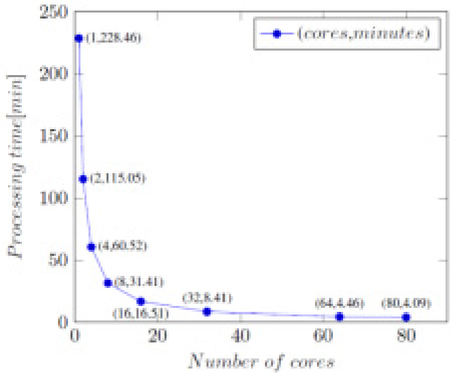
[5] Indian pines dataset, processed and raw

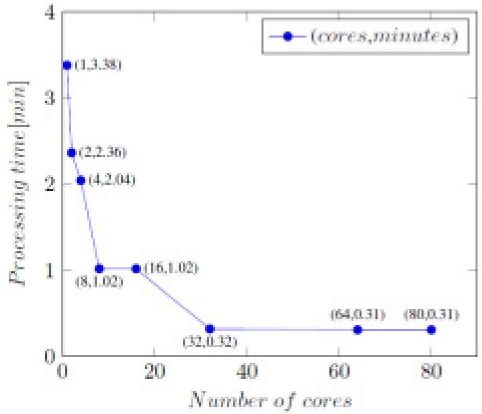
Parallel & Scalable piSVM MPI Tool – Parallel Results

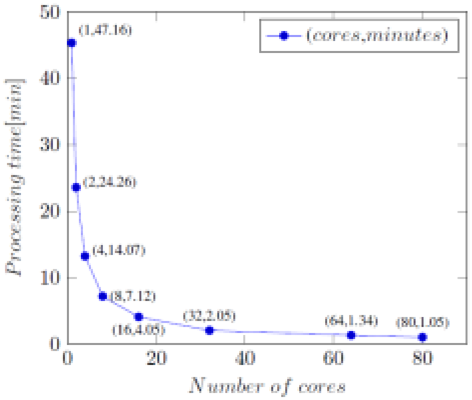
High number of classes, different scenarios important to check

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









B2SHARE Store and Share Research Data

[6] Analytics Results (raw)

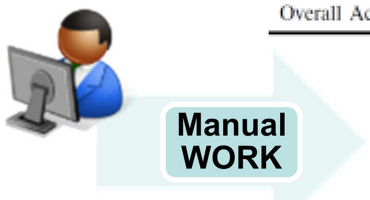
[7] Analytics Results (processed)

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



Can we automate feature extraction mechanism to some degree?

TALK AT IGARSS 2015
Wednesday, July 29, 10:30 - 12:10
(Student Paper Competition)

GABRIELE CAVALLARO et al.
AUTOMATIC MORPHOLOGICAL ATTRIBUTE PROFILES

Parallel & Scalable piSVM MPI Tool – Most Impact

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

- (1) Compute parallel; (2) Do cross-validation runs in parallel
- 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)

B2SHARE Store and Share Research Data

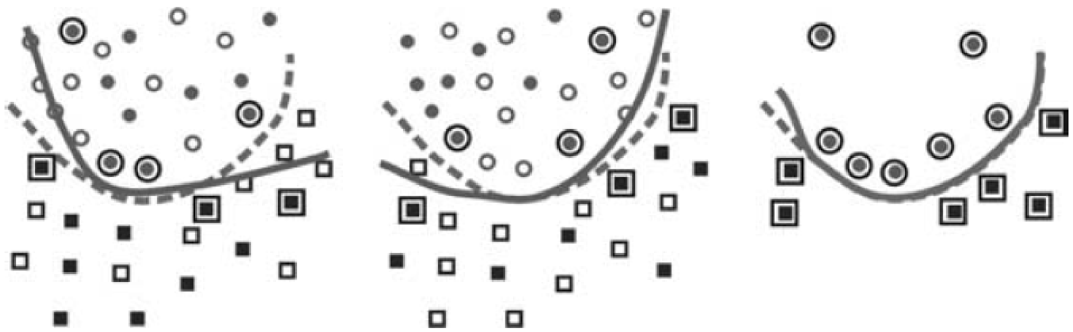
[8] Analytics 10 fold cross-validation Results (raw)

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

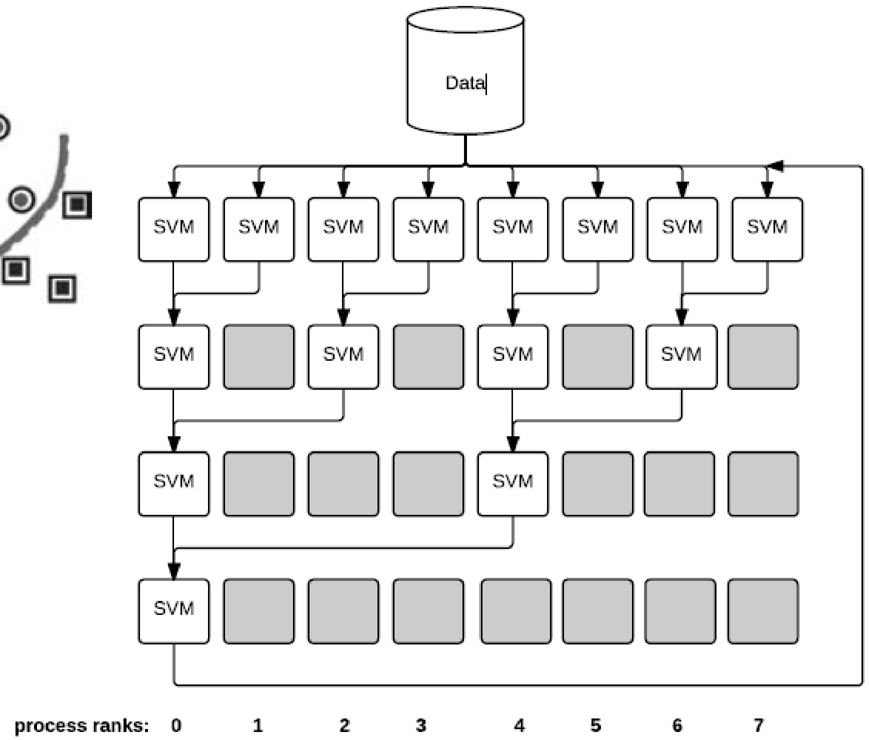
Parallel & Scalable Cascade SVM MPI Tool – Approach

Cascade Approach: Filtering the support vectors out

- Theoretically unlimited parallelization (shared nothing or MPI)
- Various extensions: cross feedback (implemented)
- Parallel application: using parallel I/O & MPI



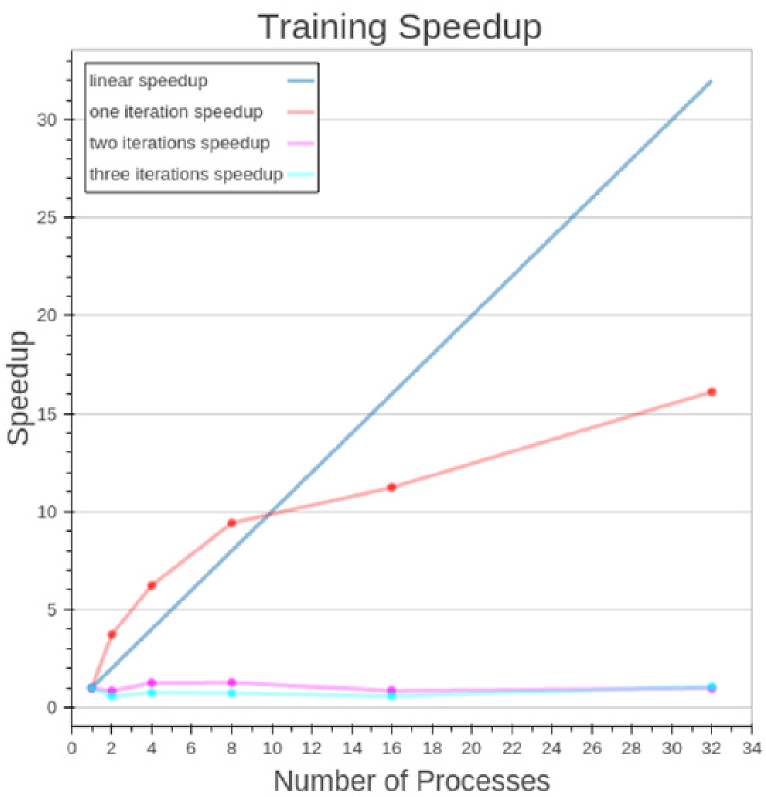
[10] Graf et al., 'Parallel Support Vector Machines: The Cascade SVM', 2005

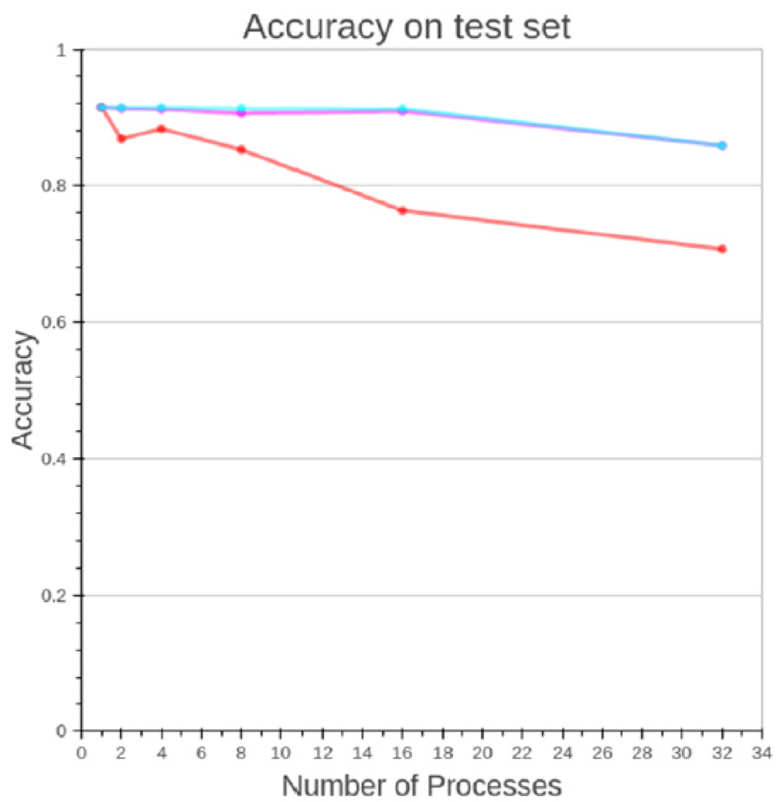


Parallel & Scalable Cascade SVM MPI Tool - Results

Results after various iterations

- Good for real-time: trade-off between accuracy & speed-up
- Comparisons with piSVM: less accuracy, but faster





Parallel & Scalable Cascade GPGPU Tool – Results

Promising area of parallelization: GPGPUs

- Not many implementations
- Still proprietary (e.g. Nvidea/CUDA)
- Open source implementation

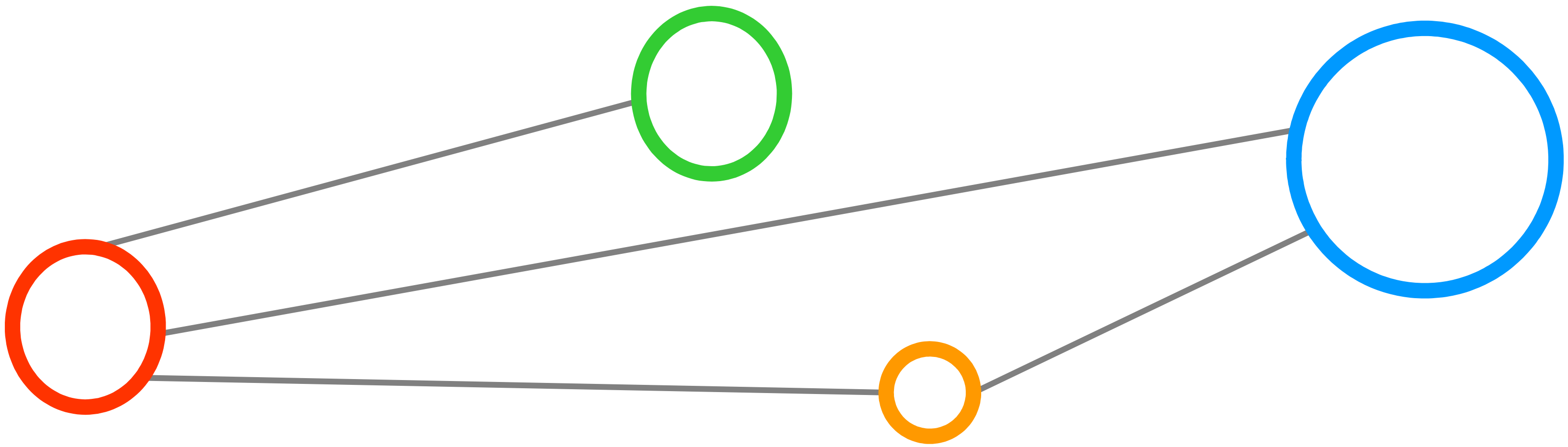
[11] V. Meyaris I. Kompatsiaris A. Athanasopoulos, A. Dimou, 'Gpu acceleration for support vector machines', 12th International Workshop on Image Analzsis for Multimedia Interactive Services, 2011.

Selected evaluations

- Still faster then serial (libsvm) version
- Performs not as good as the parallel piSVM implementation
- Disadvantages explained: e.g. overhead in copying data cpu/gpu
- Still promising area for the future: more expected to come!

	libsvm	GPU-libsvm	piSVM
raw	6779s	2619s	249s
preprocessed	746s	936s	62s

Recent Research Directions

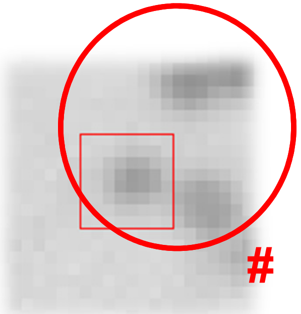


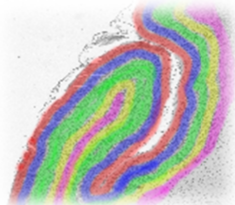
Parallel & Scalable DBSCAN MPI Tool – Brain Analytics

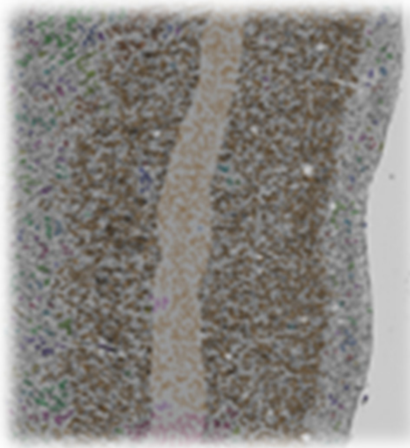
Neuroscience Application

‘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





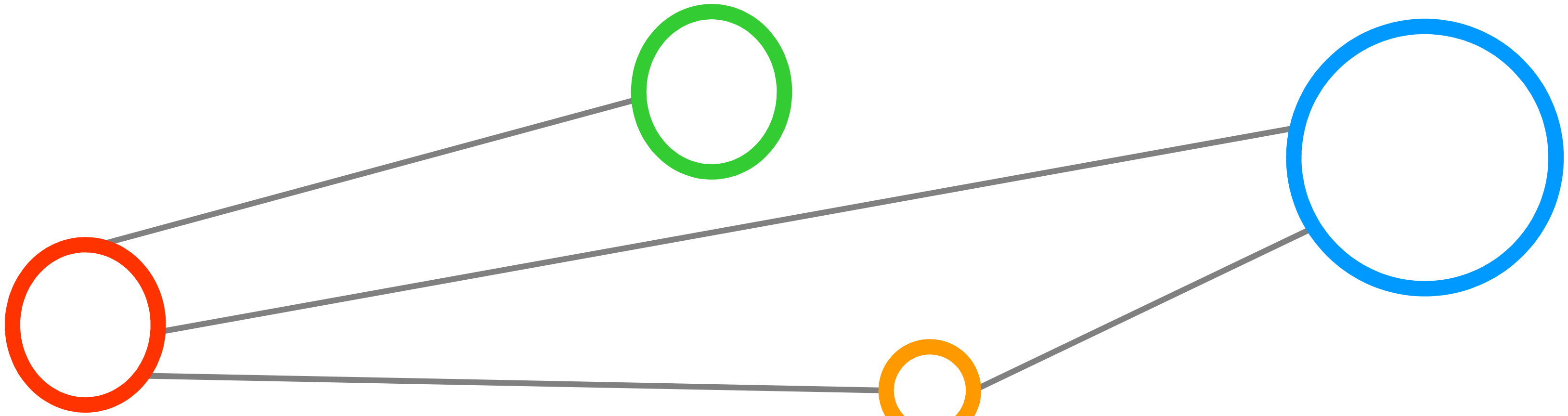


Use of HPBSCAN algorithm

- Parallel & Scalable DBSCAN algorithm
- Developed in Juelich, open source
- First 2d results detect various clusters
- Approach: Several iterations (with 3D) with potentially different parameter values

> **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 reproducibility by uploading runs and datasets

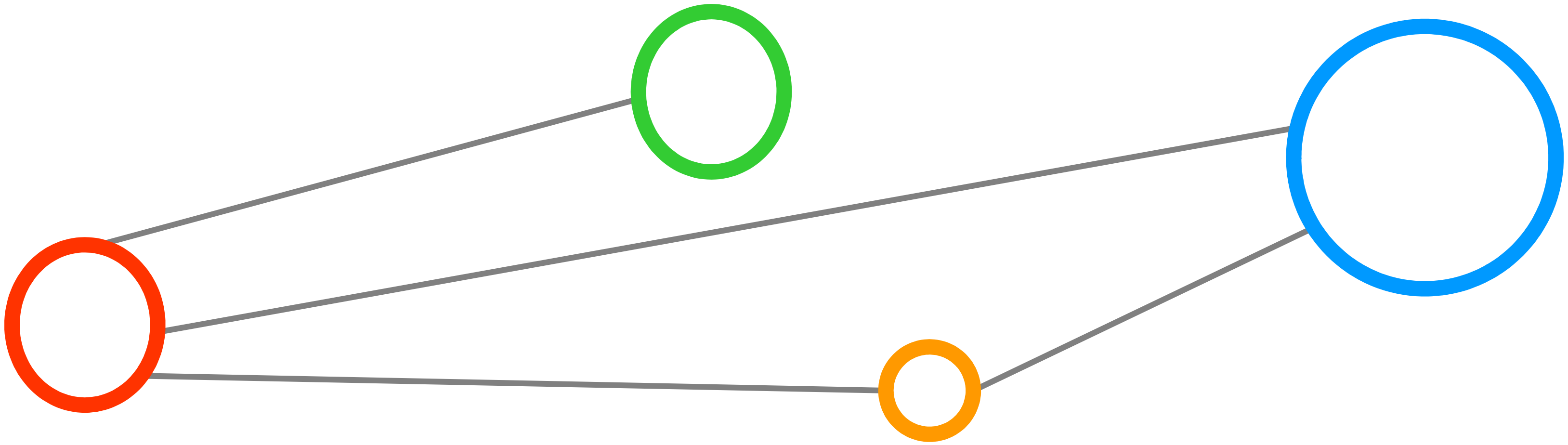
Selected SBDA benefit from parallelization

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

[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), Reykjavik

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

[4] G. Cavallaro, M. Riedel et al., IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, to be published

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

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

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

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

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

[10] Graf, Hans P., et al. "Parallel support vector machines: The cascade svm." *Advances in neural information processing systems*. 2004.

[11] V. Meyaris I. Kompatsiaris A. Athanasopoulos, A. Dimou, ‘Gpu acceleration for support vector machines’, 12th International Workshop on Image Analzsis for Multimedia Interactive Services, 2011.

Acknowledgements – Team behind the Paper

PhD Student Gabriele Cavallaro,
University of Iceland

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