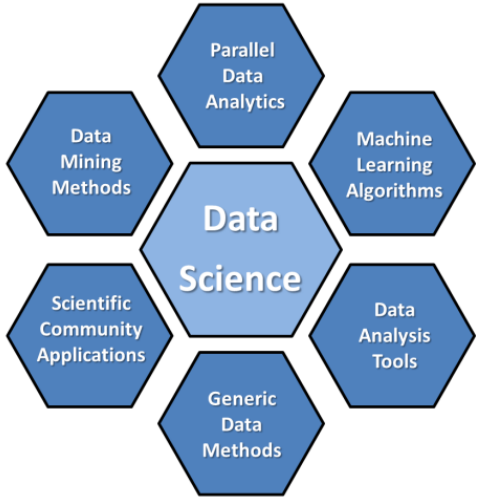
Smart Data Analytics Methods
for Remote Sensing Applications

Morris Riedel
Juelich Supercomputing Centre / University of Iceland

Gabriele Cavallaro, Jon Atli Benediktsson,
Tomas Runarsson, Kristjan Jonasson
University of Iceland

Markus Goetz, Thomas Lippert
Juelich Supercomputing Centre

2014-07-15



Federated Systems and Data Division

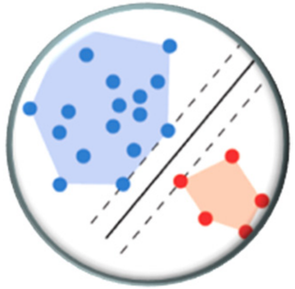
Research Group

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



Smart Data Analytics Methods

- Reasoning, Mindset, Skillset, Toolset

Remote Sensing Data Application

- Study on Land Cover Types Classification
- Survey of Related Work
- Approach and Results

Conclusions

- Future Work and Findings

References

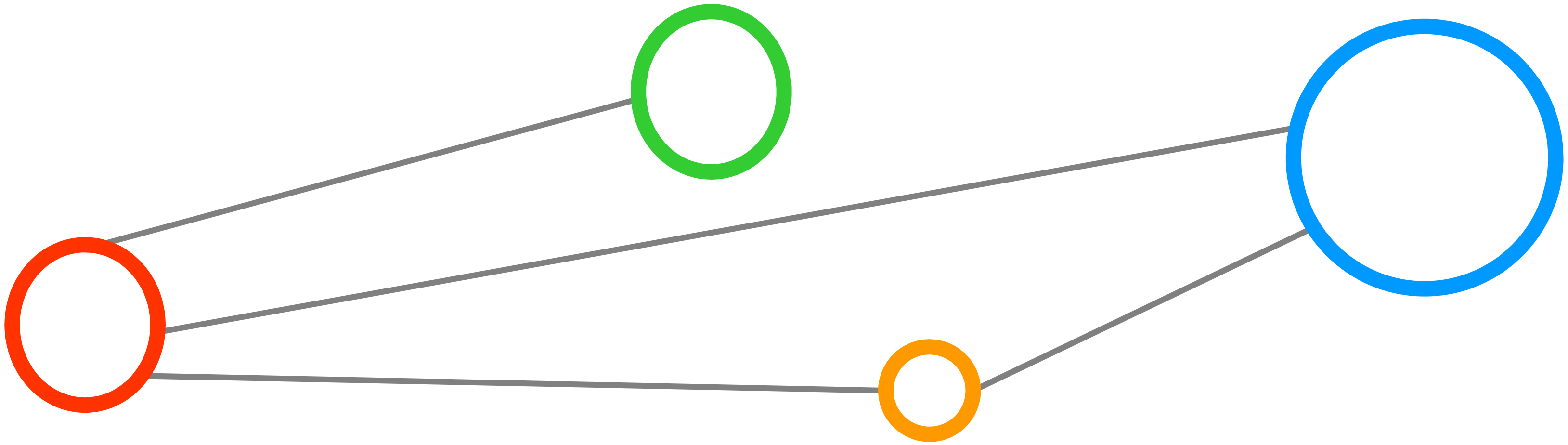
RDA RESEARCH DATA ALLIANCE

Big Data Analytics

The work was performed under the umbrella of the Research Data Alliance – Big Data Analytics Interest Group

[1] RDA BDA IG Webpage

Smart Data Analytics Methods



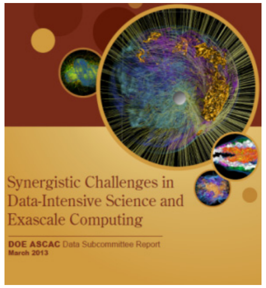
Scientific Big Data Analytics

'… problems that require high-performance data storage, **smart analytics**, transmission and mining to solve.'

[2] John Wood et al.

'In the data-intensive scientific world, **new skills are needed for …, analysing**, and making available large amounts of data…'

[3] KE Partners



'Integration of **data analytics** with exascale simulations represents a new kind of workflow…'

[4] DOE ASCAC Report

Reasoning

- Only 5-10% of archives are utilized (e.g. sensor datasets) with fast increasing data 'VVV'
- Large underutilization of data at least partly explained by the lack of 'data scientists' in domains
- Support the time-intensive manual domain-specific data analysis process with semi-automated general 'big data analytics'
- Publish reproducable results
- Big Data → 'big insights?'

Question: Is 'bigger data' really always 'better data?'

[5] D. Lazer et al. 'The Parable of Google Flu', Science 03/2014, Vol. 343

Smart Data Analytics – Mindset

Big Data Analytics

Simple Data Preprocessing

'Automated Parallel' Data Analytics

Data Postprocessing

Data Analysis

Traditional Data Analysis

choose Manual Feature Reduction

choose Manual Feature Extraction

choose Manual Feature Selection

Combine both: Smart Data Analytics

time to solution

„Best Practices":
Community-based practice & recommendations (e.g. using statistical methods)

CRISP-DM report

[6] C. Shearer, CRISP-DM model, Journal Data Warehousing, 5:13

Concete Datasets (& source/sensor)

(parallel) Algorithms & Methods

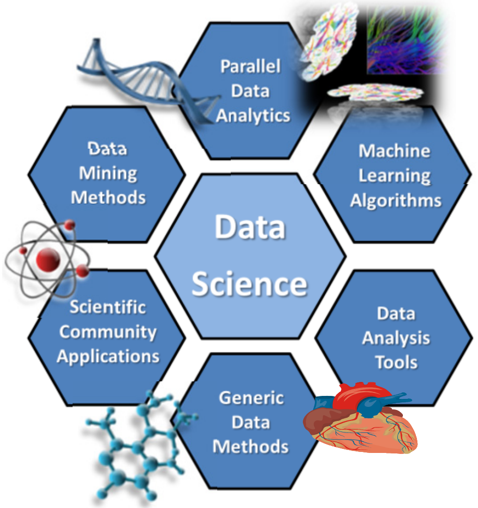
Technologies & Ressources

Scientific Data Applications

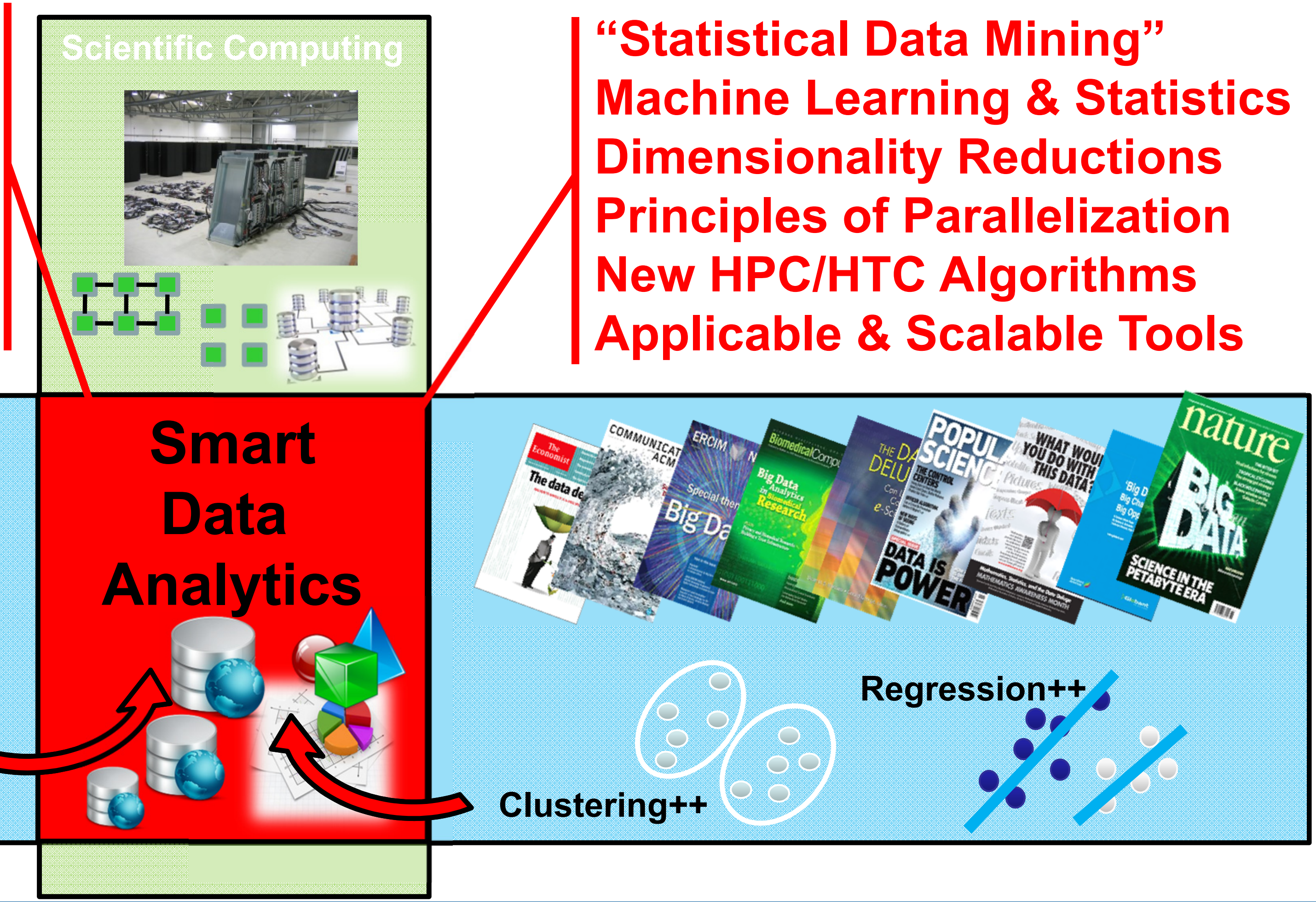
„Reference Data Analytics" for reusability & learning

Report for joint Usage	Openly Shared Datasets	Running Analytics Code

Smart Data Analytics – Skillset



"Statistical Data Mining"
Machine Learning & Statistics
Dimensionality Reductions
Principles of Parallelization
New HPC/HTC Algorithms
Applicable & Scalable Tools



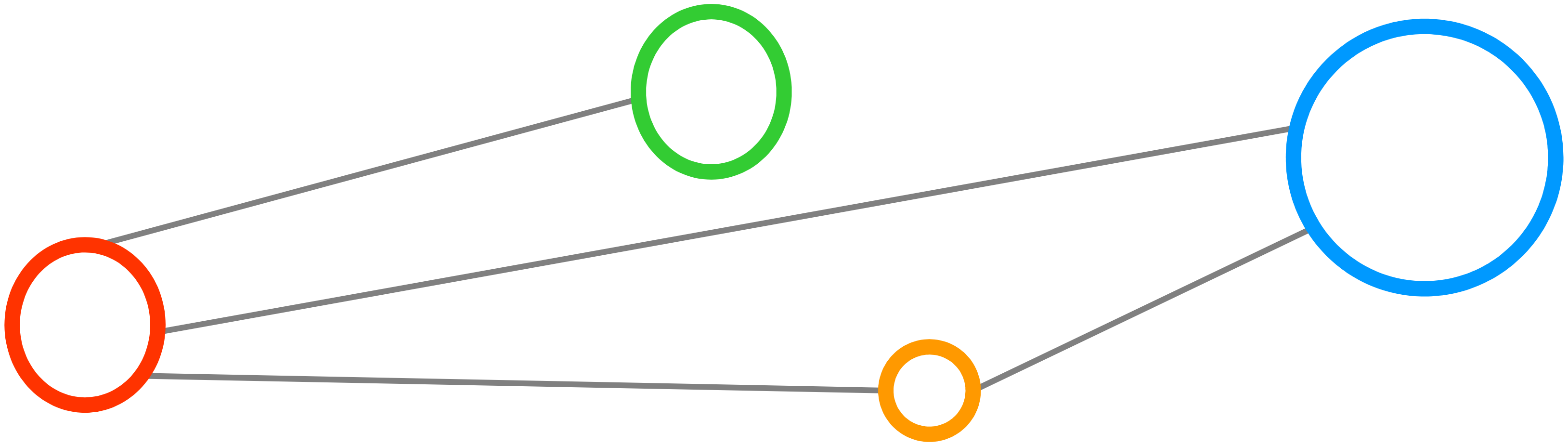
Smart Data Analytics – Toolset (SVM focus)

Tool	Platform Approach	Facts
Apache Mahout	Java; Apache Hadoop(map-reduce)	Needs to move to newer Platform Hadoop 2.0, Spark, etc.
Apache Spark/MLlib	Java; Apache Spark	Much faster than Apache Hadoop-related implementations (Website)
Twister/ParallelSVM	Java; Iterative Map-Reduce based on Twister implementation	Paper implementation after asking and based specifically on SVMs
Scikit-Learn	Python;	Machine learning package related to NumPY gaining popularity
piSVM	C code; Message Passing Interface (MPI); HPC	Open source on Sourceforge specifically for SVMs
GPU accelerated LIBSVM	CUDA language	Multi-class parallel SVM, relatively hard to program, no std. (CUDA)
pSVM	C code; Message Passing Interface (MPI); HPC	Open Source on google code, less documentation, unstable beta version

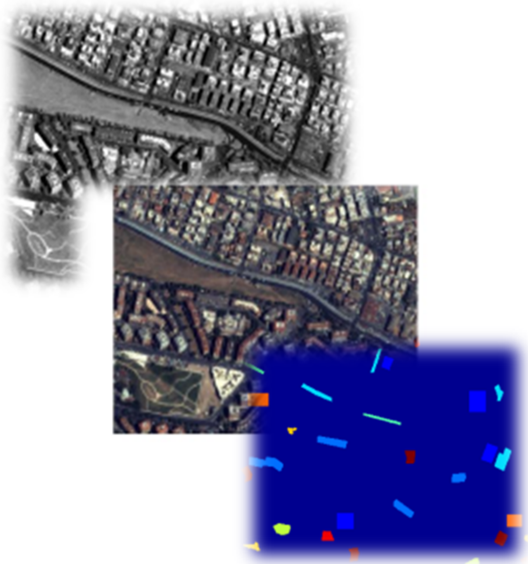
Survey of selected ‘parallel & scalable‘ machine learning tools

- Implementations often driven by commercial use cases/frameworks (e.g. linear or binary classification – credit card approval, yes/no)
- Implementations outdate quickly (e.g. Hadoop 1.0/2.0, Google Dataflow?)

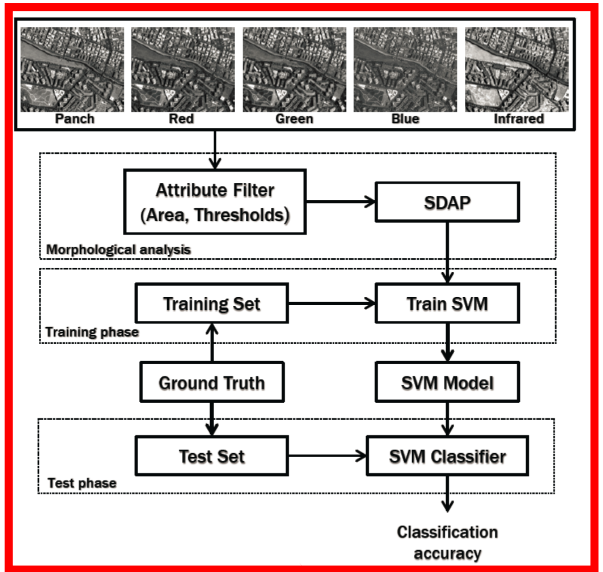
Remote Sensing Data Application



Study on parallel SVMs



Sattelite Data (Quickbird)



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

Parallel Support Vector Machines (SVM)

HPC/MPI, Map-Reduce & GPGPUs

Classification Study of Land Cover Types

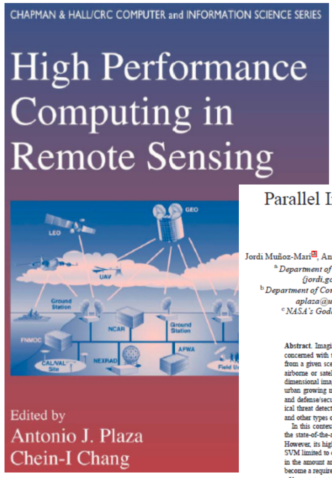
„Best Practices“

Community-based practice

„Reference Data Analytics“ for reusability & learning

CRISP-DM Report	Openly Shared Datasets	Running Analytics Code

Related Work in Remote Sensing



[7] A. J. Plaza and C. Chang, 'High Performance Computing in Remote Sensing', CRC Press, 2007

Parallel Implementations of SVM for Earth Observation

Jordi Muñoz-Marí[a], Antonio J. Plaza[b], J. Anthony Gualtieri[c], Gustavo Camps-Valls[a]

[a] Department of Electronics Engineering, University of Valencia, Spain, {jordi,gcamps}@uv.es, http://www.uv.es/{jordi,gcamps}

[b] Department of Computer Science, University of Extremadura, Cáceres, Spain. aplaza@unex.es http://www.umbc.edu/rssipl/people/aplaza

[c] NASA's Goddard Space Flight Center, Greenbelt, Maryland, USA, Anthony.Gualtieri@gst.com

Abstract. Imaging spectroscopy, also known as hyperspectral remote sensing, is concerned with the measurement, analysis, and interpretation of spectra acquired from a given scene (or specific object) at a short, medium or long distance by an airborne or satellite sensor. Analysis in a timely manner of the acquired multi-dimensional images allows to develop applications with high social impact, such as urban growing monitoring, crop fields identification, target detection for military and defense/security deployment, wildland fire detection and monitoring, biological threat detection, biophysical parameter estimation, or monitoring of oil spills and other types of chemical contamination.

In this context, support vector machines (SVM) [1, 2, 3] have become one of the state-of-the-art machine learning tools for hyperspectral image classification. However, its high computational cost for large scale applications makes the use of SVM limited to off-line processing scenarios. Certainly, with the recent explosion in the amount and complexity of hyperspectral data, parallel processing has soon become a requirement in many remote sensing missions, especially with the advent of low-cost systems such as commodity clusters and distributed networks of computers In order to address this relevant issue, this chapter explores the development of two parallel versions of SVMs for remote sensing image classification.

Sequential minimal optimization is a very popular algorithm for training SVMs, but it still requires a large amount of computation time for solving large size problems. In this work, we evaluate the performance of a parallel implementation of the SVM based on the parallelization of the incomplete Cholesky factorization and present novel parallel implementations that balance the load across the available processors through standard Master-Worker decompositions. Both methodologies are theoretically analyzed in terms of scalability, computational efficiency and time response. The impact of the multi-class scheme is also analyzed. Results on real multispectral and hyperspectral datasets illustrate the performance of the methods. We finally discuss the possibility of obtaining processing results quickly enough for practical use via the Marenostrum supercomputer available at Barcelona Supercomputing Center in Spain, and other massively parallel facilities at NASA's Goddard Space Flight Center in Maryland.

Keywords. Parallel processing, image classification, remote sensing, support vector machine.

[8] J. Munoz-Man, A. J. Plaza, J.A. Gualtiers, G. Camps-Valls 'Parallel Implementations of SVM for Earth Observation', Parallel Programming, Models and Applications in Grid and P2P Systems, 2009, pages 292-312

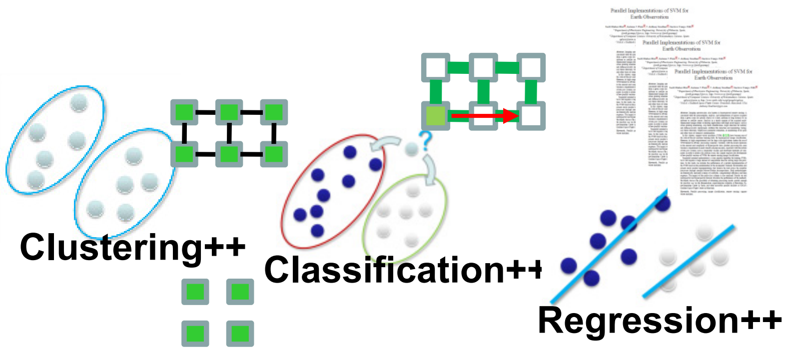
→ Good domain-specific science insights, e.g. sub-domain of 'spectral unmixing' has big data…

… but 2014 challenges remain: HPC reinvents itself every year

- Massively increased amount of cpus/cores and memory (+getting cheaper)
- New techniques in data-related properties: MPI-IO & parallel-IO libraries
- Better infrastructures: Improved parallel file systems and data sharing
- New architectural approaches & Languages: 'GPGPUs & python trend'
- Scientific codes running on old machines not necessarily good on new ones

Related Work in Parallel & Distributed Computing

Tool	Platform Approach	Parallel Support Vector Machine
Apache Mahout	Java; Apache Hadoop 1.0 (map-reduce); HTC	No strategy for implementation (Website), serial SVM in code
Apache Spark/MLlib	Apache Spark; HTC	Only linear SVM; no multi-class implementation
Twister/ParallelSVM	Java; Apache Hadoop 1.0 (map-reduce); Twister (iterations), HTC	Much dependencies on other software: Hadoop, Messaging, etc.
Scikit-Learn	Python; HPC/HTC	Multi-class Implementations of SVM, but not fully parallelized
piSVM	C code; Message Passing Interface (MPI); HPC	Simple multi-class parallel SVM implementation outdated (~2011)
GPU accelerated LIBSVM	CUDA language	Multi-class parallel SVM, relatively hard to program, no std. (CUDA)
pSVM	C code; Message Passing Interface (MPI); HPC	Unstable beta, SVM implementation outdated (~2011)



closed/old source, also after asking paper authors

Algorithm A Implementation

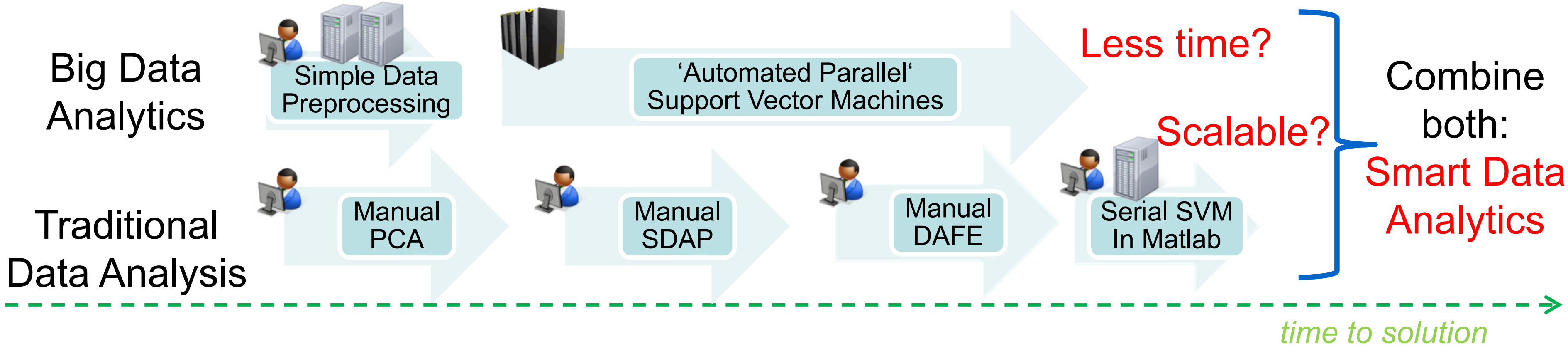
Algorithm Extension A‘ Implementation

Parallelization of Algorithm Extension A‘ → A“

implementations available

implementations rare and/or not stable

Study – Mindset



Big Data Analytics → [processing power++, time scientists-]

- Working on 'big data' by an automated process on computing machinery
- Scalable to 'big data volumes' (e.g. high dimensions), image time-series

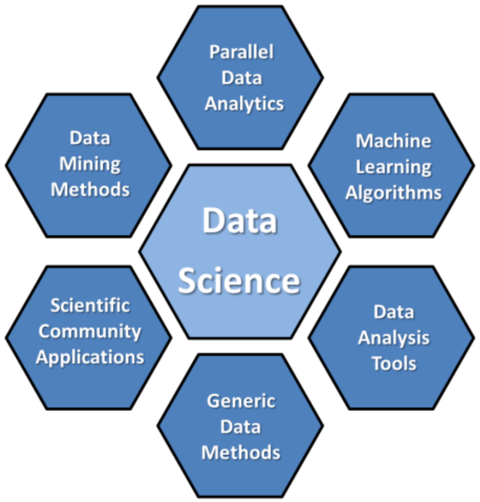
Traditional Data Analysis → [time scientists+++, processing power-]

- Data reduction by manual intervention → 'small data' (e.g. low dimensions)
- Not necessarily needs ‚large-scale computing environments' – scalable?

Study – Skillset

Smart Data Analytics: Clever mix of both approaches

- Apply parallel and distributed computing techniques where feasible
- Take advantage of semi-automated statistical techniques from data science



Examples to reduce ‘big dataset dimensions‘

- Principle Component Analysis (PCA)
- Discriminant Analysis Feature Extraction (DAFE)

Classification optimization technique

- Self-Dual Attribute Profile (SDAP)


Area


Std Dev


Moment of Inertia

[9] 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

Open Questions remains for the study…

- Can we perhaps ‘speed-up‘ some of the statistical techniques?
- Parallel cross-validation for ‘model selection‘ before running SVMs?

Study – Toolset

Tool	Platform Approach	Findings when using Tool
Twister/ParallelSVM	Java; Apache Hadoop 1.0 (map-reduce); Twister (iterations), HTC	Much dependencies on other software: Hadoop, Messaging: stability needs to improve; slightly outdated move to HARP (Hadoop 2.0 SVM plug-in)
piSVM	C code; Message Passing Interface (MPI); HPC	Works stable; speed-up only when computing is really required (make no sense for small dataset dimensions), optimizations in code (load imbalance with increasing cores, collectives, etc.)
GPU accelerated LIBSVM	CUDA language	Easy to install, but relatively hard to program, no standard language (CUDA); but promising for future tests

‘HTC Approach‘

- Used FutureGrid cluster with Twister/ParallelSVM
- Uses map-reduce & messaging

[10] Sun Z., and Fox G., ‘Study on Parallel SVM Based on MapReduce’, In Proceedings of the international conference on parallel and distributed processing techniques and applications, 2012.

‘HPC Approach‘

- Used JUDGE cluster at Juelich Supercomputing Centre
- MPI was installed; piSVM ported

[11] piSVM Website, 2011 code

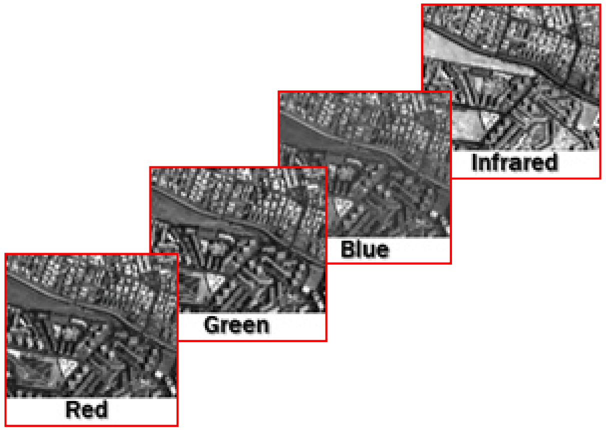
Study – Datasource & Sensors

Geographical location: Image of Rome, Italy

- Remote sensor data obtained by Quickbird satellite

High-resolution (0.6m) panchromatic image

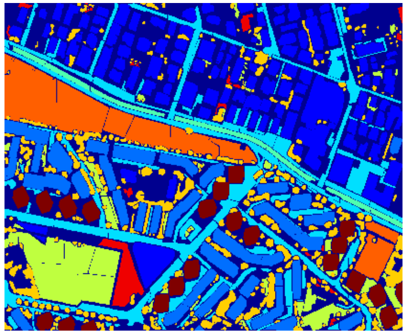
Pansharpened (UDWT) low-resolution (2.4m) multispectral images



Study – Training vs. Test Data Generation

Labelled data available

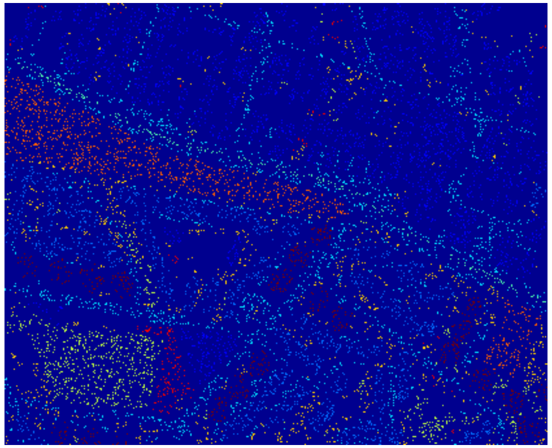
- Groundtruth data of 9 different land-cover classes available



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)

Study – Data structure

Based on 'LibSVM data format'

- E.g. 'SDAP on area' on all images training file

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

55 features

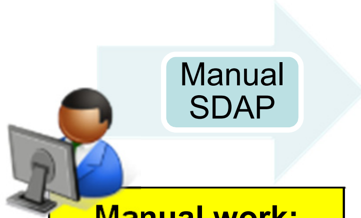
#77542 samples

B2SHARE
Store and Share Research Data

Study – Selected Results

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

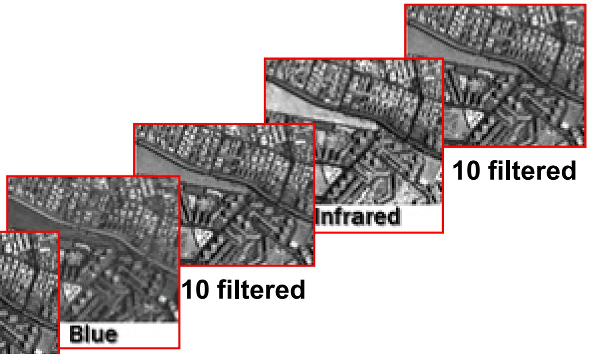
- Serial Matlab: ~1277 sec (~21 minutes)
- Parallel (16) Analytics: 220 sec (3:40 minutes)
- Accuracy remains

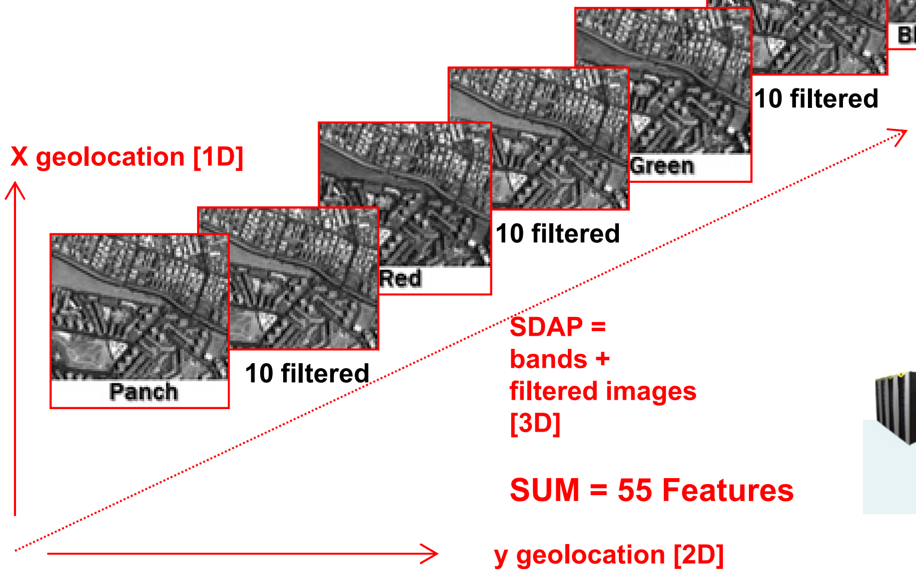


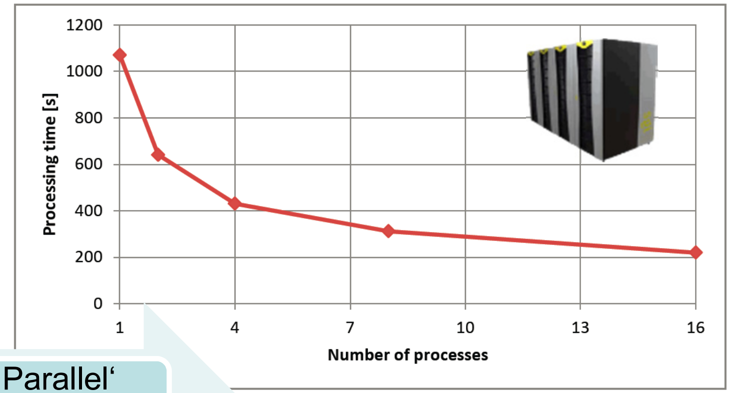
Manual work: Obtain the SDAP for all image bands using attribute 'area' (10 thresholds)

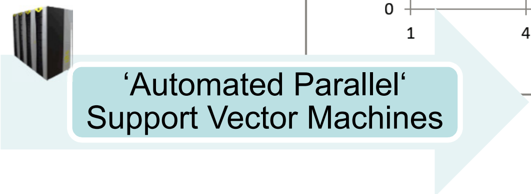
Training vector

- 77542 samples









B2SHARE
Store and Share Research Data

Study – Selected Further Initial Results

Consideration trade-off man vs. machine

- Goal of Smart Data Analytics: automate the process, maintain accuracy
- Goal of traditional data analysis: reduce manual time, high accuracy
- Comparing serial Matlab vs. Parallel Analytics only in parts ‘fair‘

Training speed-up is not achieved when features are ‘low‘

- Automated parallel shared (!) environment needs time to setup
- Avoiding the creation of big data:
  e.g. ‘SDAP only on panchromatic image (reduced to 15 features)
- Time in Matlab is one minute, no need for analytics during manual work

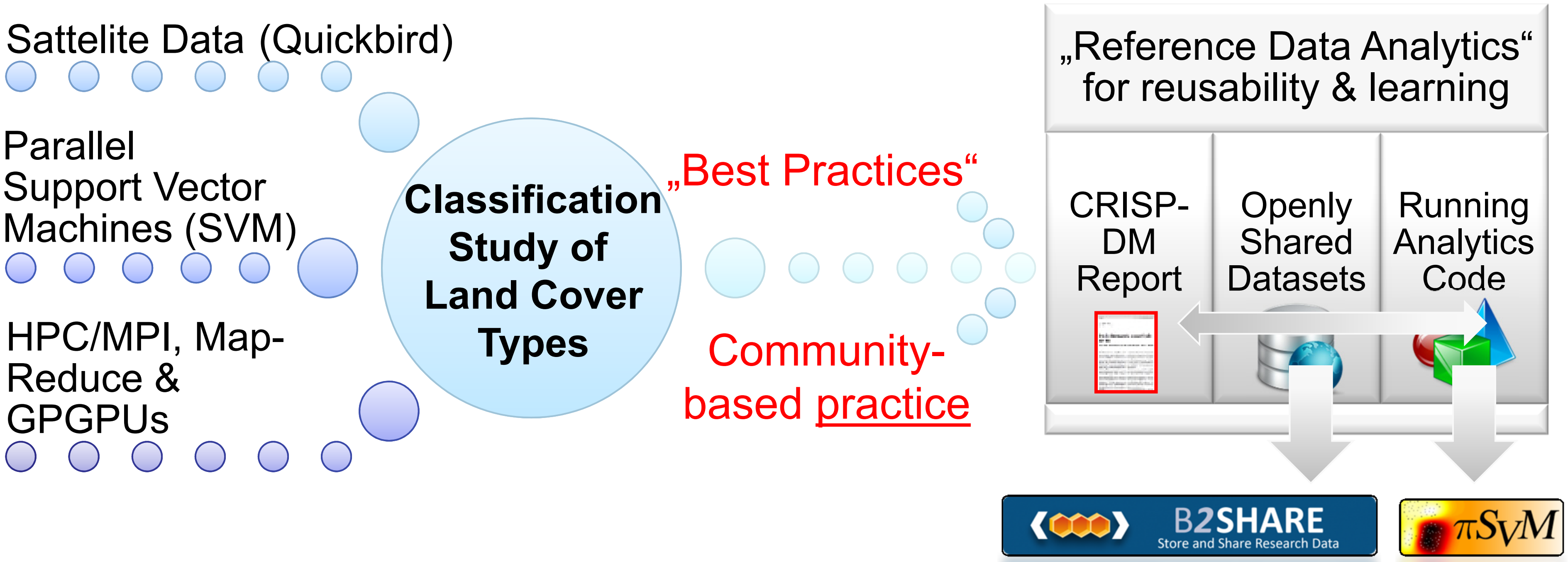
Speed-up of SVM – Predict (Test time) significantly

- Better parallelization with predictions possible
- Serial Matlab: ~2080 sec.
- Parallel (16) Analytics: ~120 sec.

Study – Reproducability Aspects

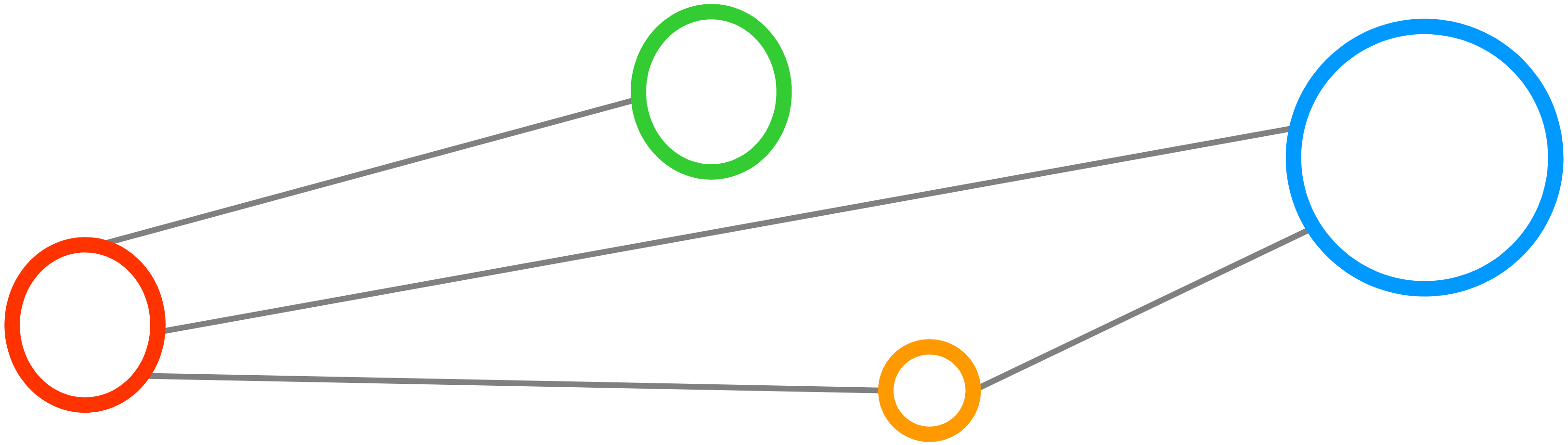
Inline with emerging publishing requirements

- Running analytics code and used datasets openly available
- Datasets have a ‘persistent identifier (PIDs)‘ based on the handle system



[12] EUDAT B2SHARE *[11] piSVM*

Conclusions



Future Work

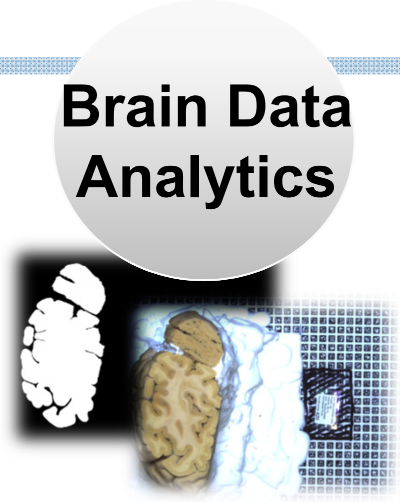
Transfer results to other scientific domains

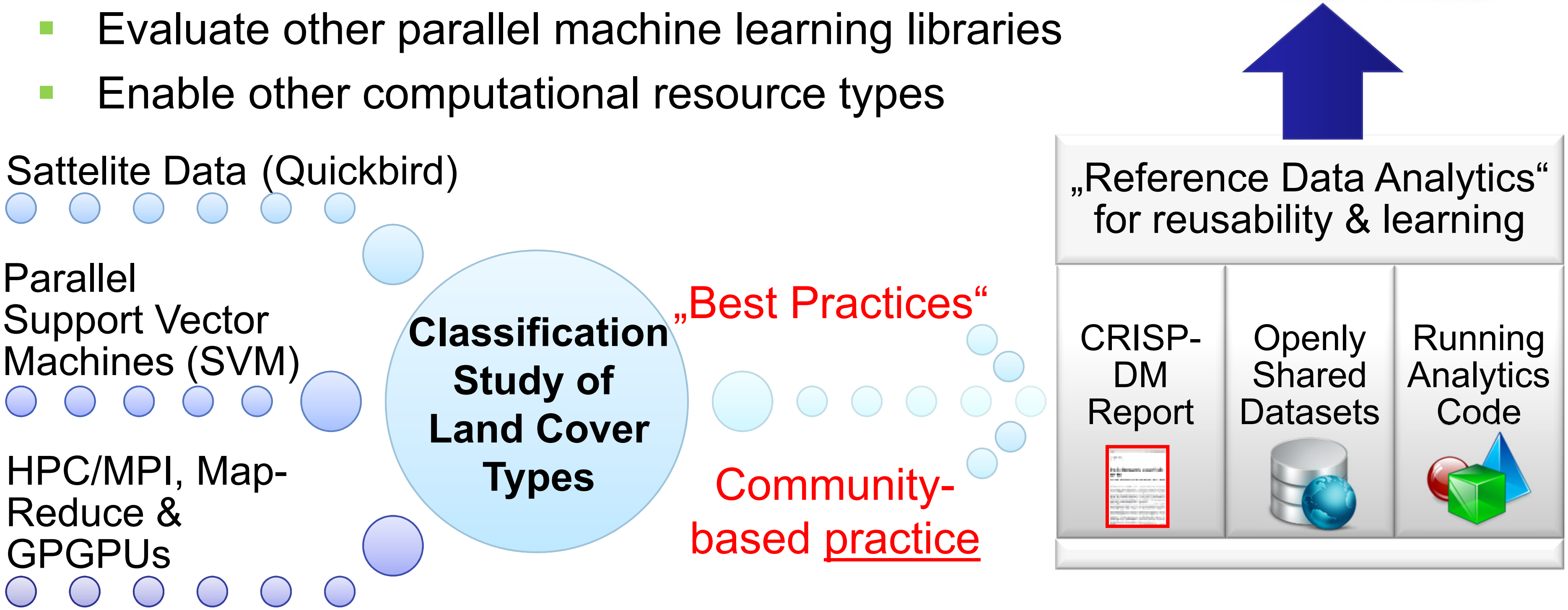
- Contribute to Human Brain Project (HBP)

  [13] G. Shepherd et al., 'The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data', *Trends in neurosciences* 21.11 (1998): 460-468.

Use of different resources & tools

- Evaluate other parallel machine learning libraries
- Enable other computational resource types





Findings in a Nutshell

Scientific Smart Data Analytics

- Often different & more complex as industrial ‘big data analytics‘ cases
- Data science often driven by industrial-driven tools
  → Need scientific steering from communities (peer-review process)

Mindset

- Trade-off in time → manual statistical techniques vs. automated analytics
- Big Data trend → ‘Bigger data‘ does not necessarily mean ‘better data‘

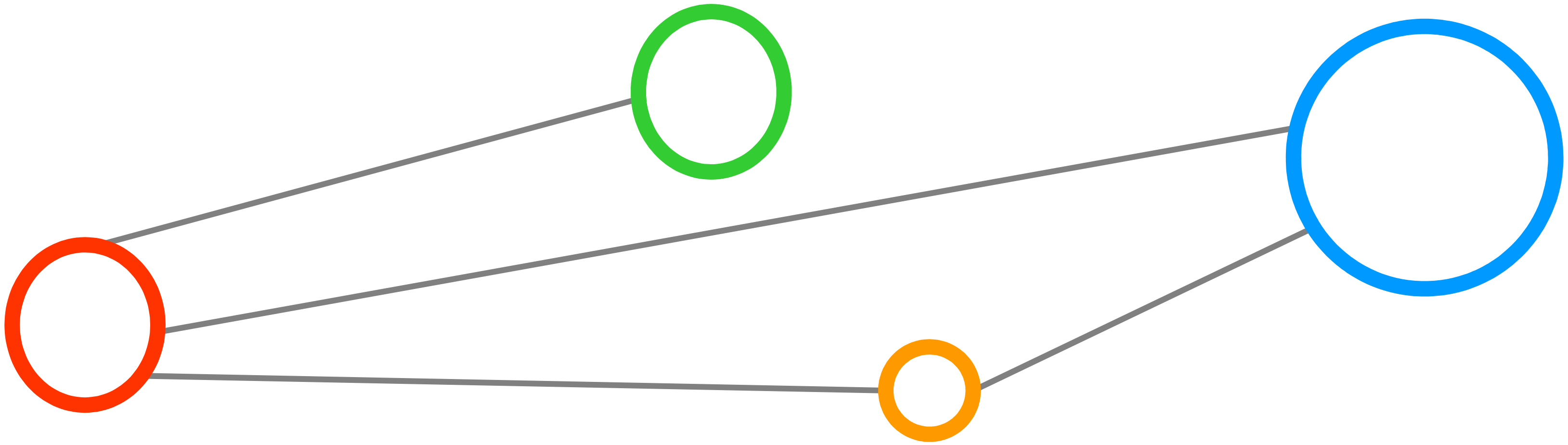
Skillset

- Knowledge of statistical methods essential → ‘Reduce big data‘
- Time to ensure ‘good reproducability‘ enormous → Need of ‘data curators‘
- Lack of skilled people in domain + computing → Need of ‘data scientists‘

Toolset

- Rare open availability of parallel machine learning codes
- Stability and implemented functionality of codes needs to increase

References



References

[1] RDA BDA IG Webpage, online: https://rd-alliance.org/group/big-data-analytics-ig.html

[2] John Wood et al., 'Riding the Wave –How Europe can gain from the rising tide of scientific data', EC Report, 2010

[3] KE Partners, 'A Surfboard for Riding the Wave - Towards a four country action programme on research data', November 2012

[4] DOE ASCAC Data Subcommittee Report, 'Synergistic Challenges in Data-Intensive Science and Exascale Computing', 2013

[5] D. Lazer et al. 'The Parable of Google Flu – Traps in Big Data Analysis', Science 03/2014, Vol. 343

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

[7] A. J. Plaza and C. Chang, 'High Performance Computing in Remote Sensing', CRC Press, 2007

[8] J. Munoz-Man, A. J. Plaza, J.A. Gualtiers, G. Camps-Valls 'Parallel Implementations of SVM for Earth Observation', Parallel Programming, Models and Applications in Grid and P2P Systems, 2009, pages 292-312

[9] 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

[10] Sun Z., and Fox G., 'Study on Parallel SVM Based on MapReduce', In Proceedings of the international conference on parallel and distributed processing techniques and applications, 2012.

[11] piSVM Website, 2011 code, online: http://pisvm.sourceforge.net/

[12] EUDAT European Data Infrastructure, B2SHARE Tool, Online: https://b2share.eudat.eu/

[13] Shepherd, Gordon M., et al. "The Human Brain Project: neuroinformatics tools for integrating, searching and modeling multidisciplinary neuroscience data." *Trends in neurosciences* 21.11 (1998): 460-468.

Thanks for your attention

RESEARCH DATA ALLIANCE

FOURTH PLENARY MEETING

22 – 24 September 2014

Amsterdam, the Netherlands | Meervaart conference centre

www.rd-alliance.org/rda-fourth-plenary-meeting.html

Talk available at:

www.morrisriedel.de/talks

Contact:

m.riedel@fz-juelich.de