Data Sharing Experiences of Smart Data Analytics Tasks in Remote Sensing Research

Dr. – Ing. Morris Riedel

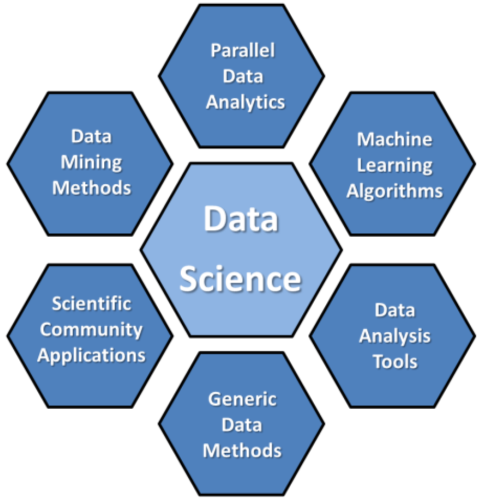
Adjunct Associated Professor, University of Iceland

Research Group Leader, Juelich Supercomputing Centre

Gabriele Cavallaro

University of Iceland

APARSEN – DPHEP – EUDAT – SCIDIP-ES
Joint Data Preservation Workshop
‘Safeguarding our Scientific, Educational and Cultural Heritage’
24 September 2014
Black Room, De Meervaart Conference Centre, Meer en Vaart 300, Amsterdam 1068



Federated Systems and Data Division

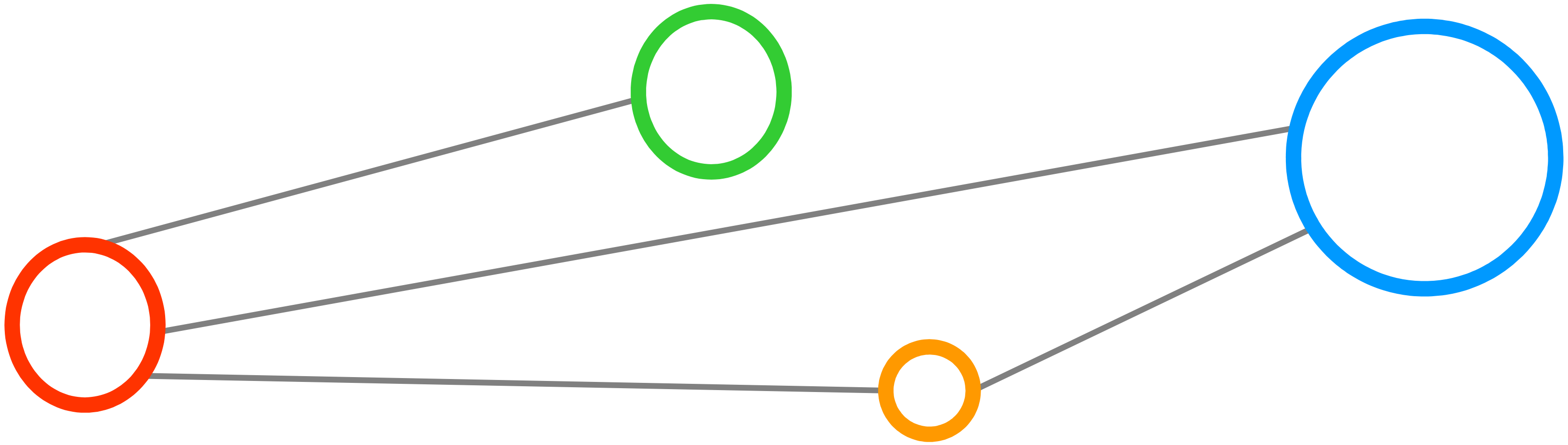
Research Group

High Productivity Data Processing

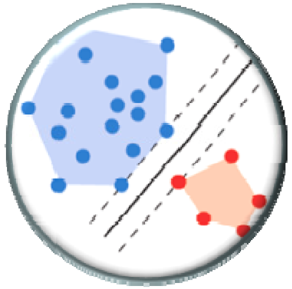
JÜLICH FORSCHUNGSZENTRUM

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

Outline



Outline



Research Group High Productivity Data Processing

- Smart Data Analytics & Daily Work Activities

Smart Data Analytics in Remote Sensing Research

- Typical Example: Study on Land Cover Types Classification

Using EUDAT B2SHARE

- Data sharing and Preserving Outcomes
- Practical Examples and Usage Models

Summary

- Selected Findings and Suggestions

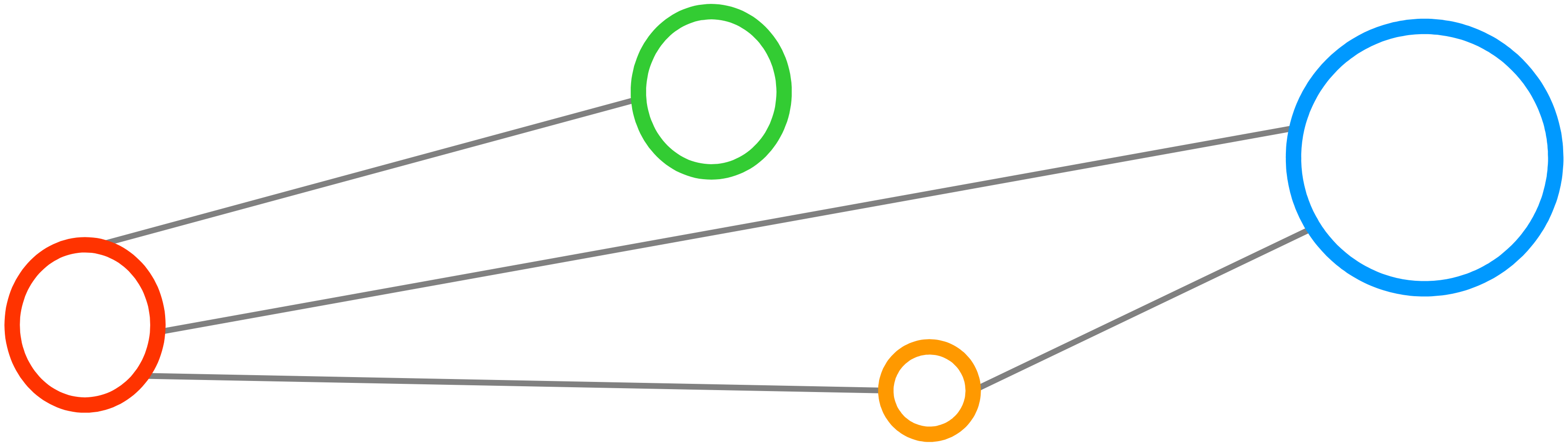
References

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

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

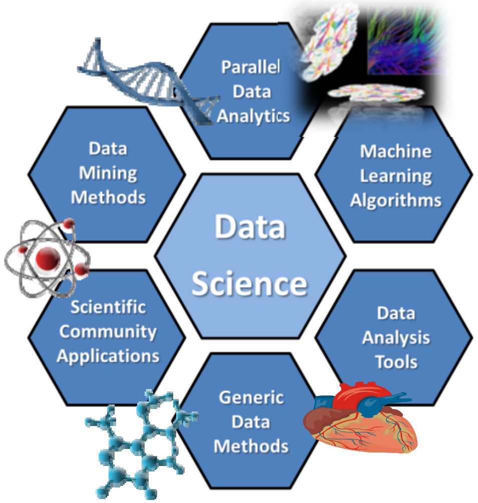
[1] RDA BDA IG Webpage

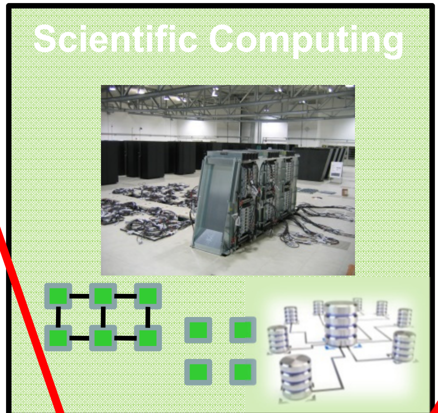
Research Group High Productivity Data Processing



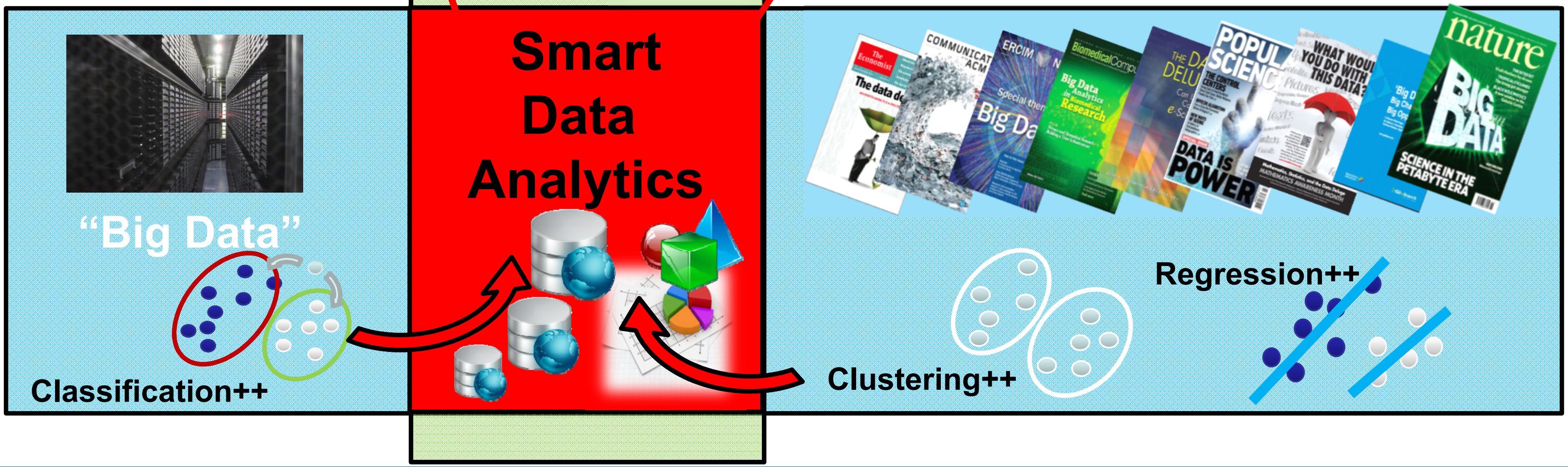
Research Group High Productivity Data Processing

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

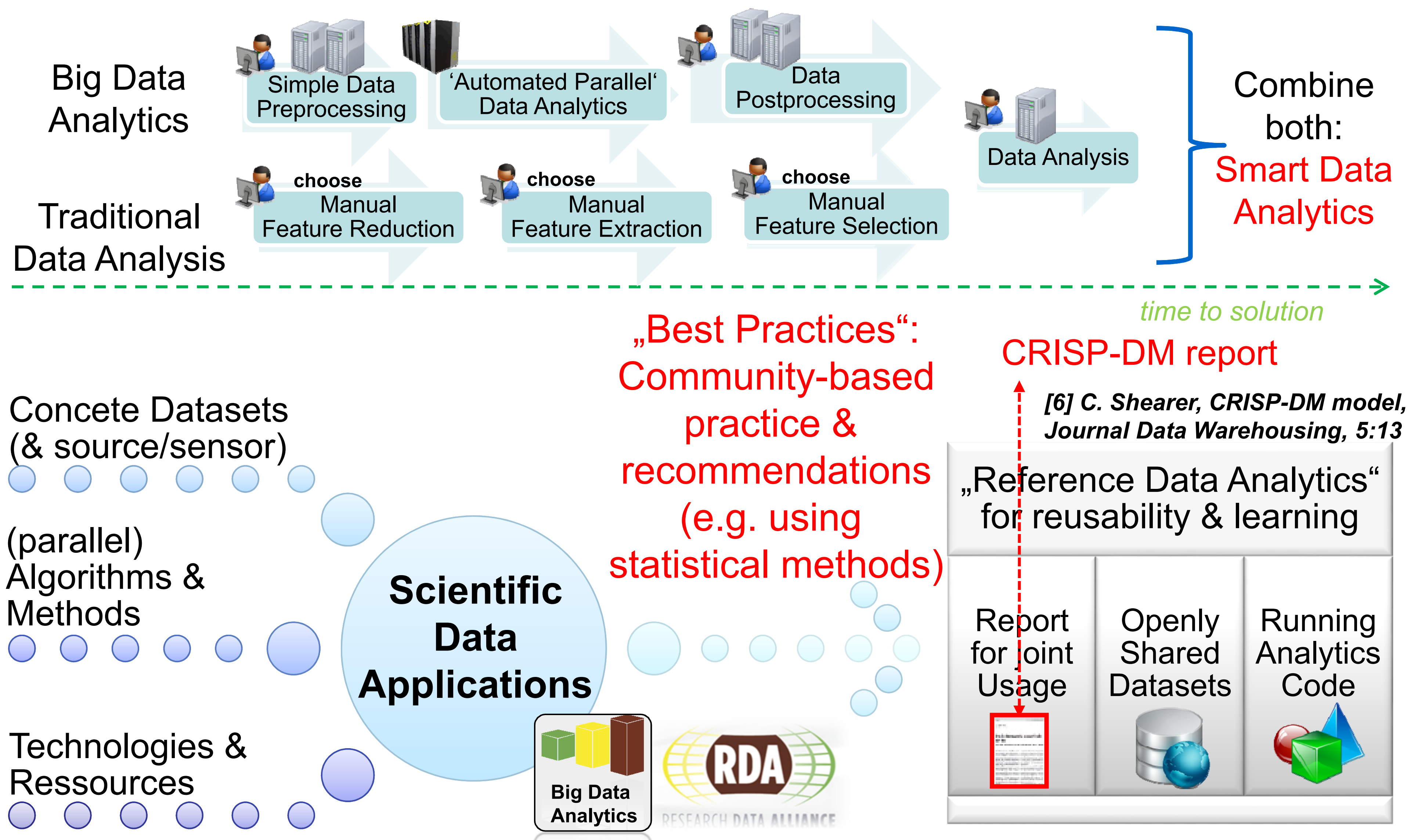




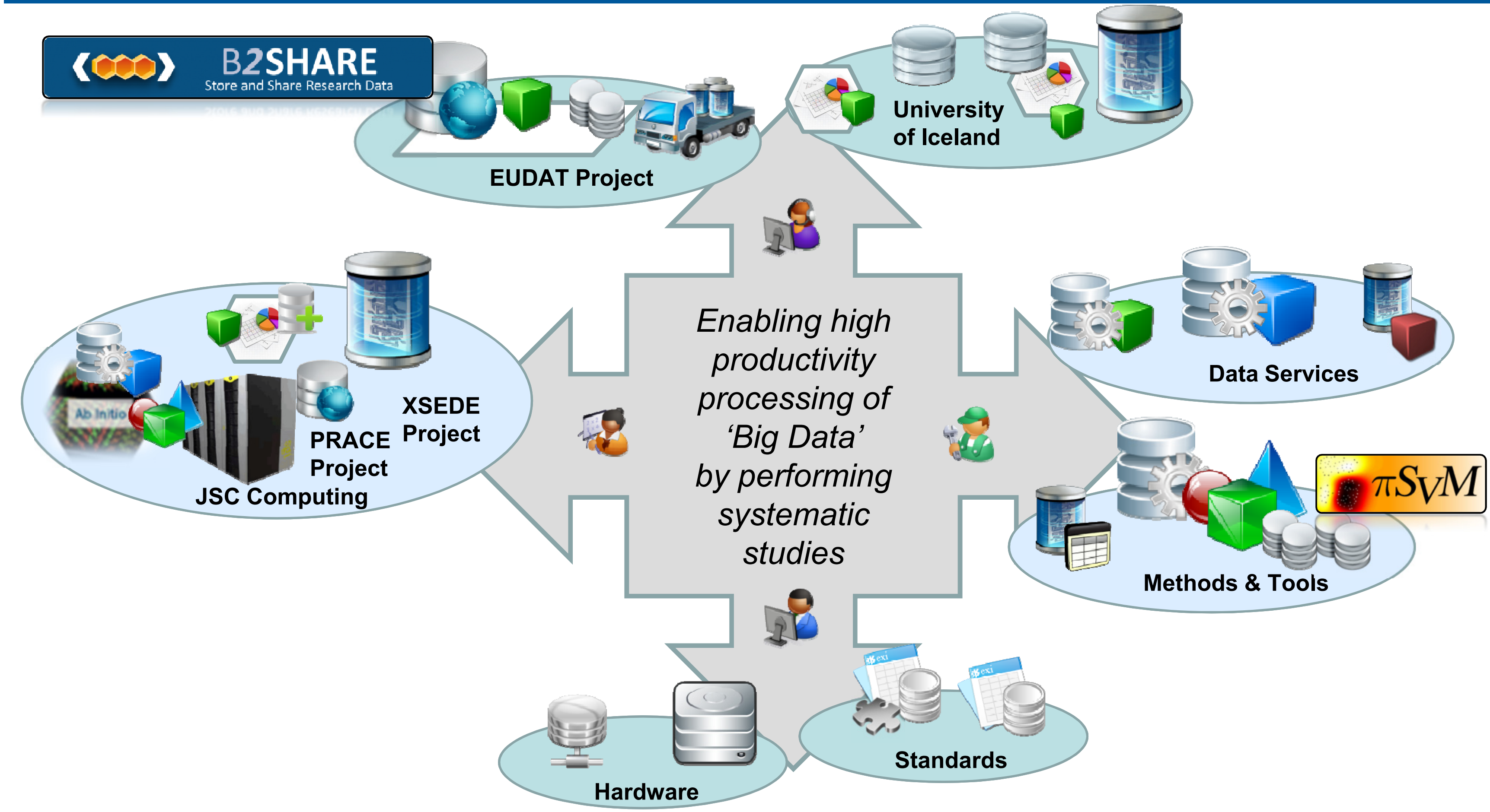
"Statistical Data Mining"
Machine Learning & Statistics
Dimensionality Reductions
Principles of Parallelization
New HPC/HTC Algorithms
Applicable & Scalable Tools



Smart Data Analytics

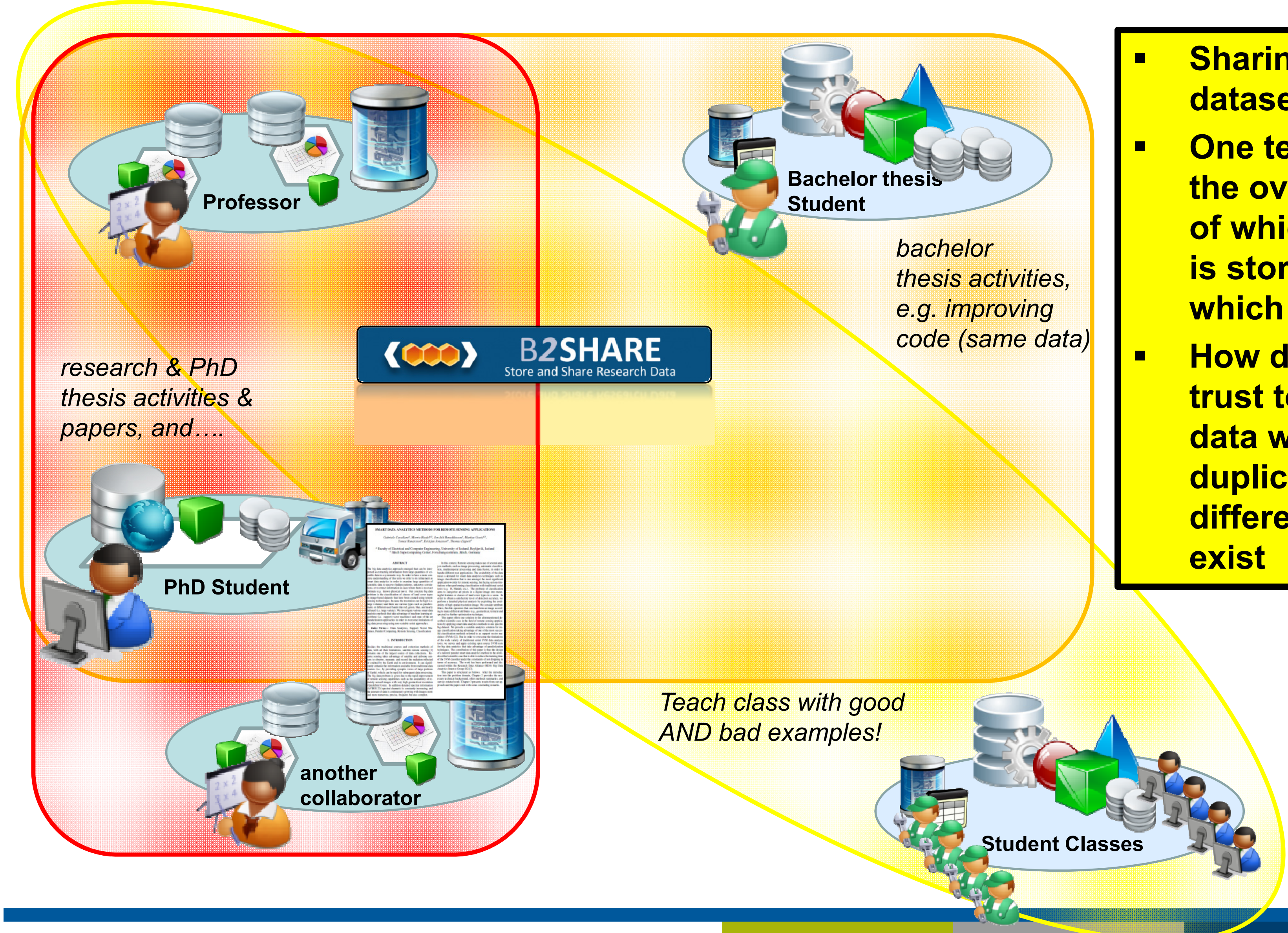


Need for Sharing: Complex work environments



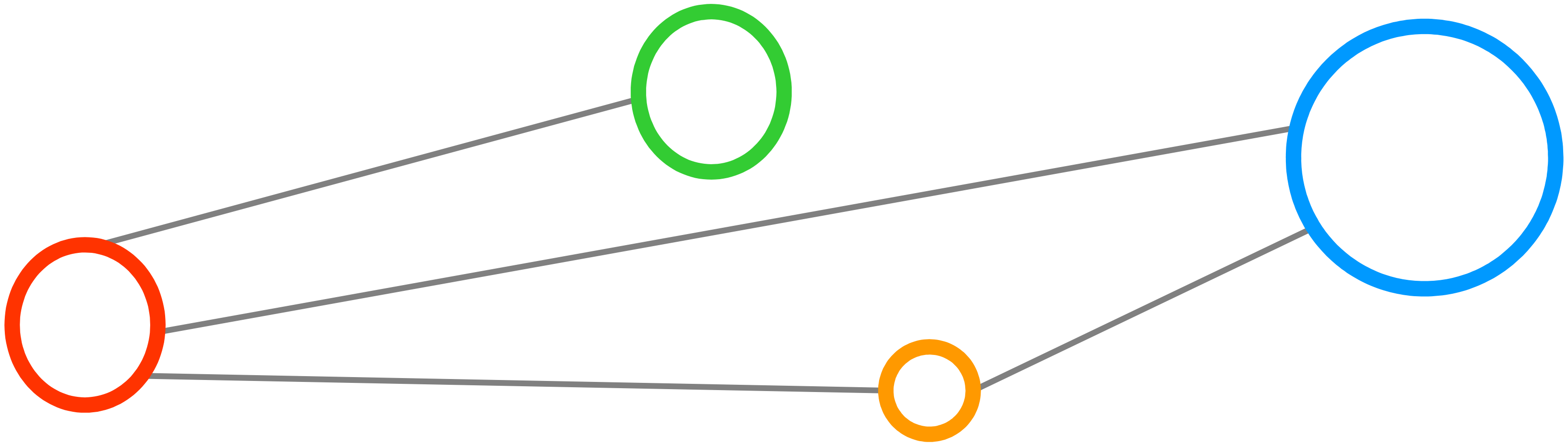
- One tend to loose the overview of which data is stored on which platform
- How do we gain trust to delete data when duplicates on different systems exist?

Need for Sharing: One Example from Daily Research

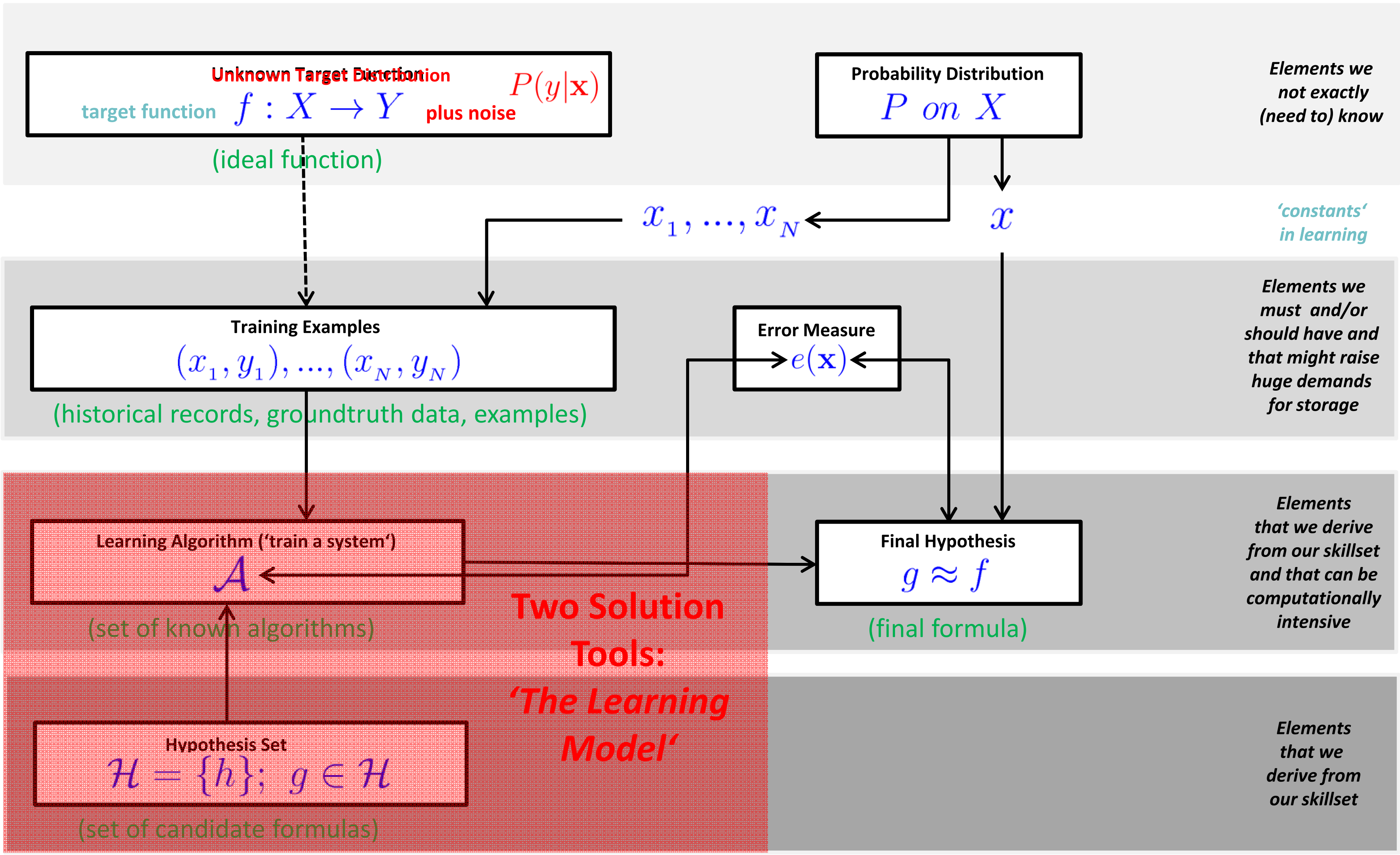


- **Sharing different datasets is key**
- **One tend to loose the overview of which data is stored on which platform**
- **How do we gain trust to delete data when duplicates on different systems exist**

Smart Data Analytics in Remote Sensing Research

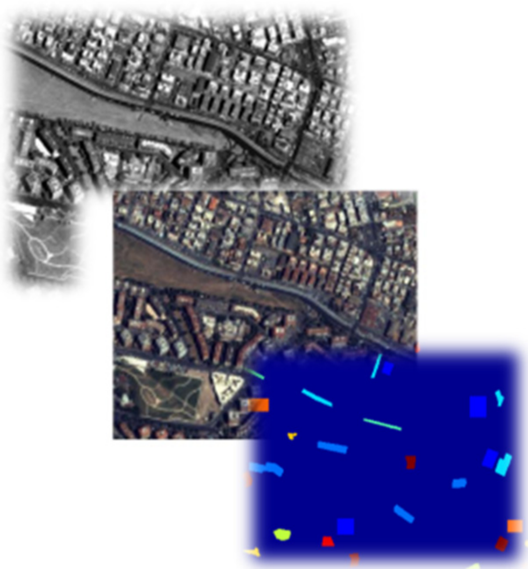


Supervised Learning from Data – Data Inputs & Outputs

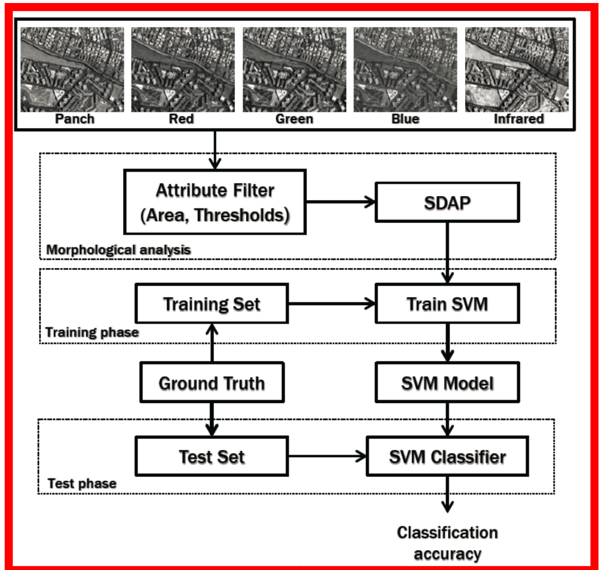


Use parallel Support Vector Machines (SVMs)

Classification++



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

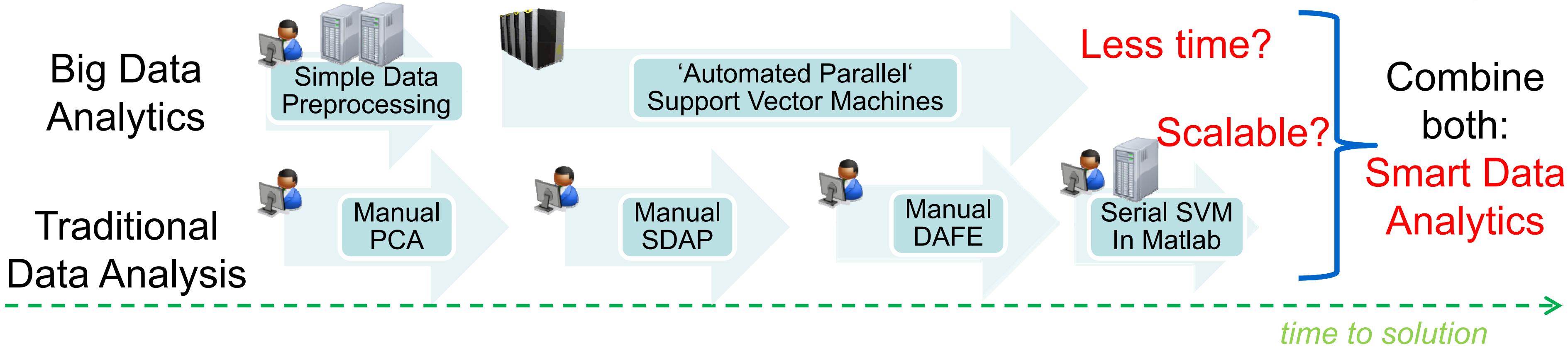
„Best Practices“

Community-based practice

„Reference Data Analytics“ for reusability & learning

CRISP-DM Report	Openly Shared Datasets	Running Analytics Code

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

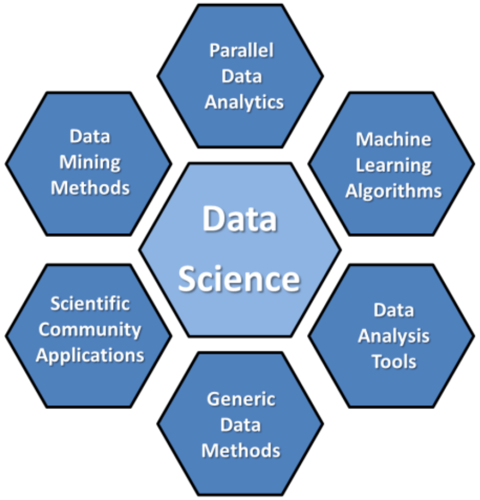
Traditional Data Analysis → [time scientists+++, processing power-]

- 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?
- How can we preserve outcomes of the process for re-use & sharing?

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.

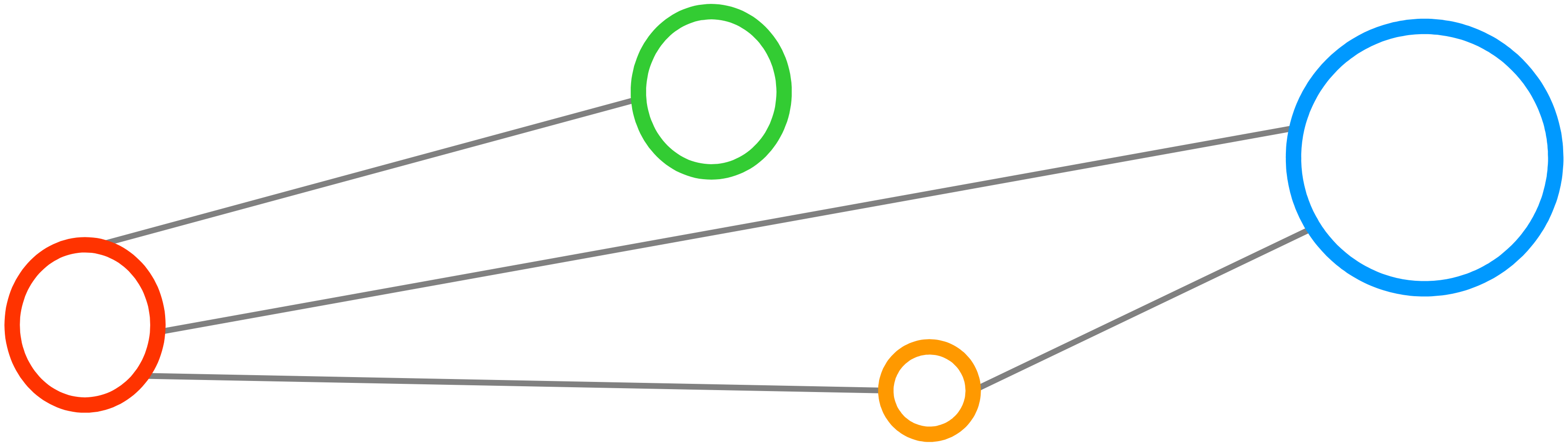
‘HPC Approach‘

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

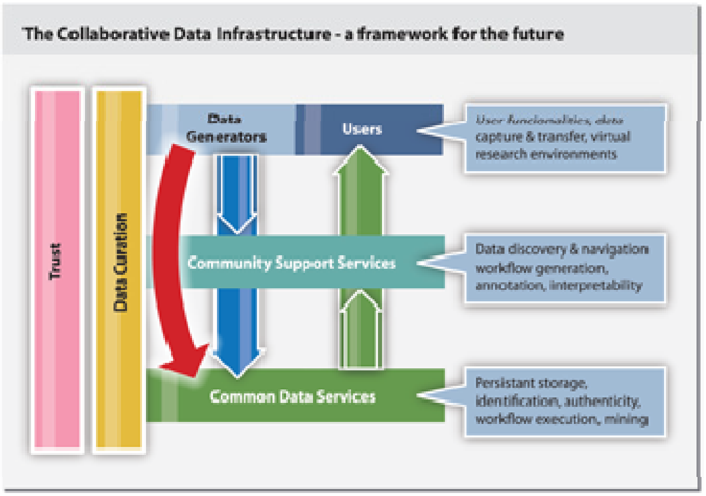
[11] piSVM Website, 2011 code

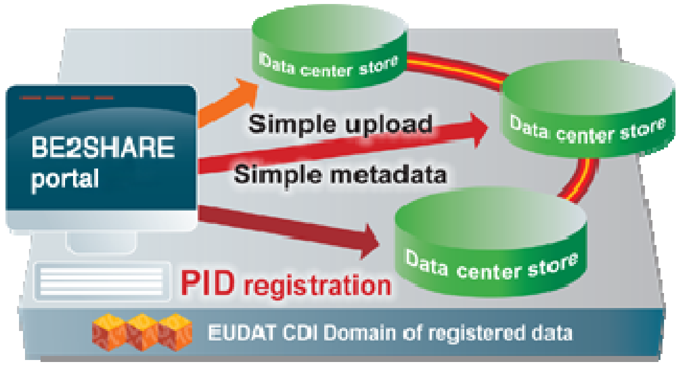
πSvM

Using EUDAT B2SHARE in Research



EUDAT B2SHARE





- **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**

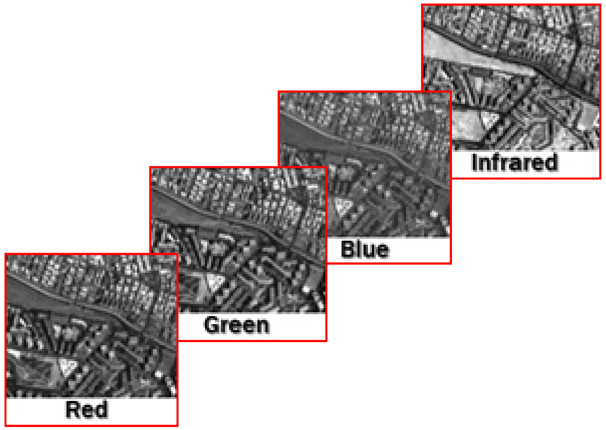
Study – Datasource & Sensors

Geographical location: Image of Rome, Italy

- Remote sensor data obtained by Quickbird satellite

High-resolution (0.6m) panchromatic image

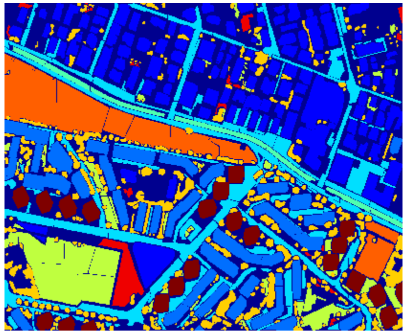
Pansharpened (UDWT) low-resolution (2.4m) multispectral images



Study – Training vs. Test Data Generation

Labelled data available

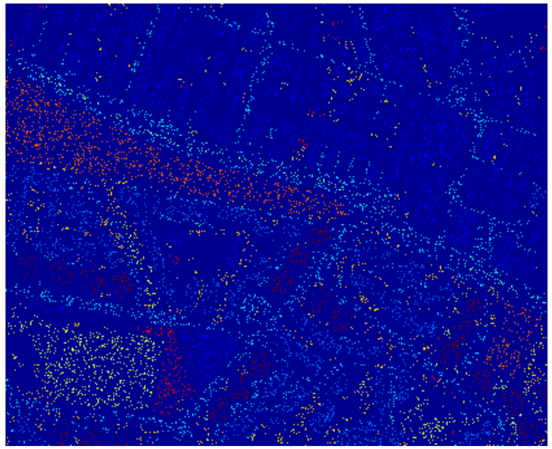
- Groundtruth data of 9 different land-cover classes available



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)

Data structure required for Data Analytics

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

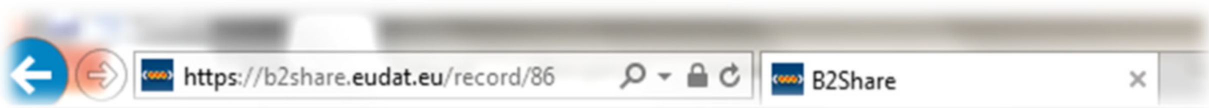
```
each line is a pixel    3 1:0.105882 2:0.109804 3:0.101961  ........  54:0.121569 55:0.130952
                        2 1:0.364706 2:0.360784 3:0.356863  ........  54:0.356863 55:0.349206
                        6 1:0.152941 2:0.34902  3:0.454902  ........  54:0.466667 55:0.460317
                        ........
                        ........
                        ........
                        .......
                        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
```

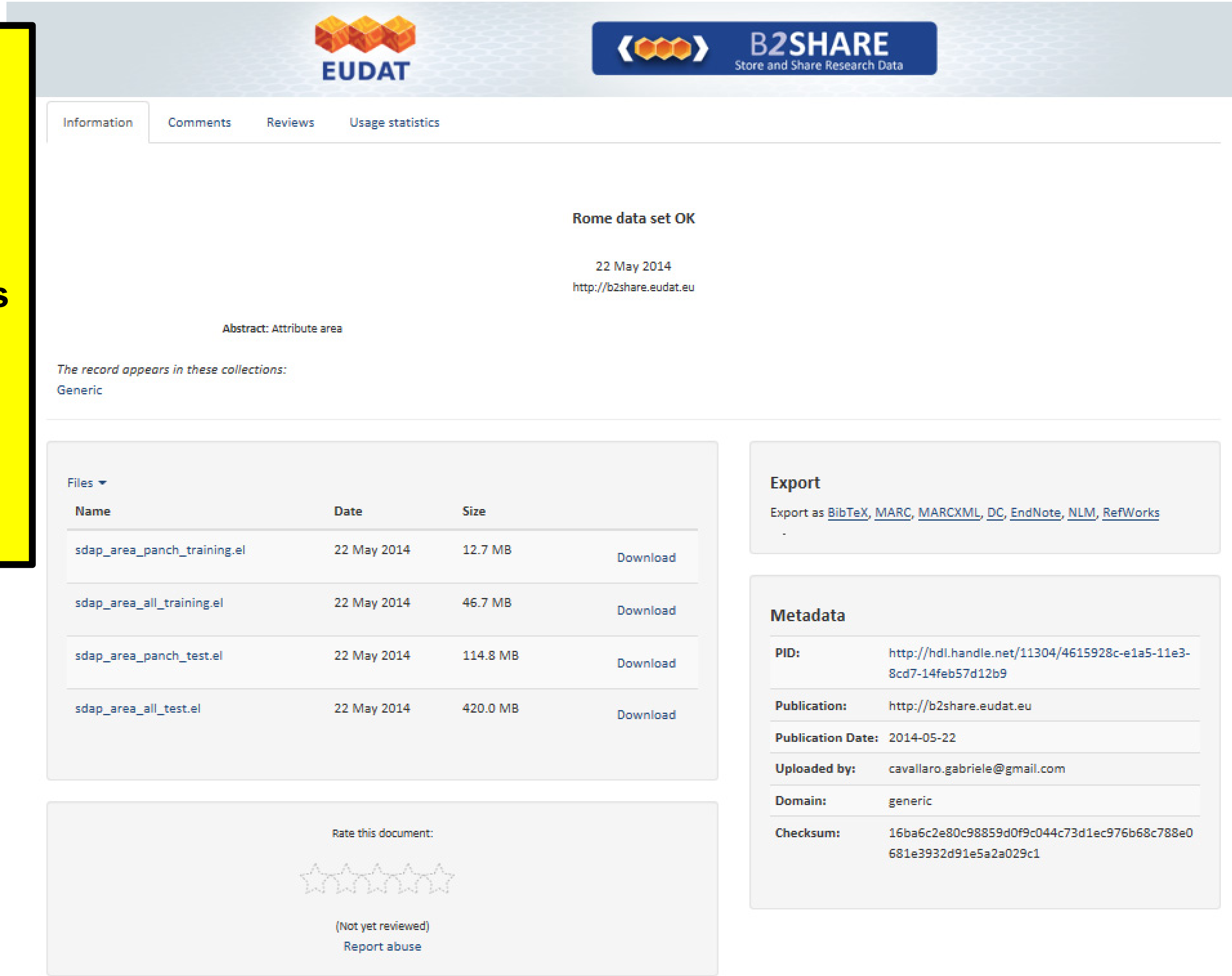
#77542 samples

55 features

B2SHARE Store and Share Research Data

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





Sharing and Downloading Dataset on Cluster JUDGE

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**

Link back to B2SHARE fosters Trust

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**

Data analysis using Cluster Judge and Share Results

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

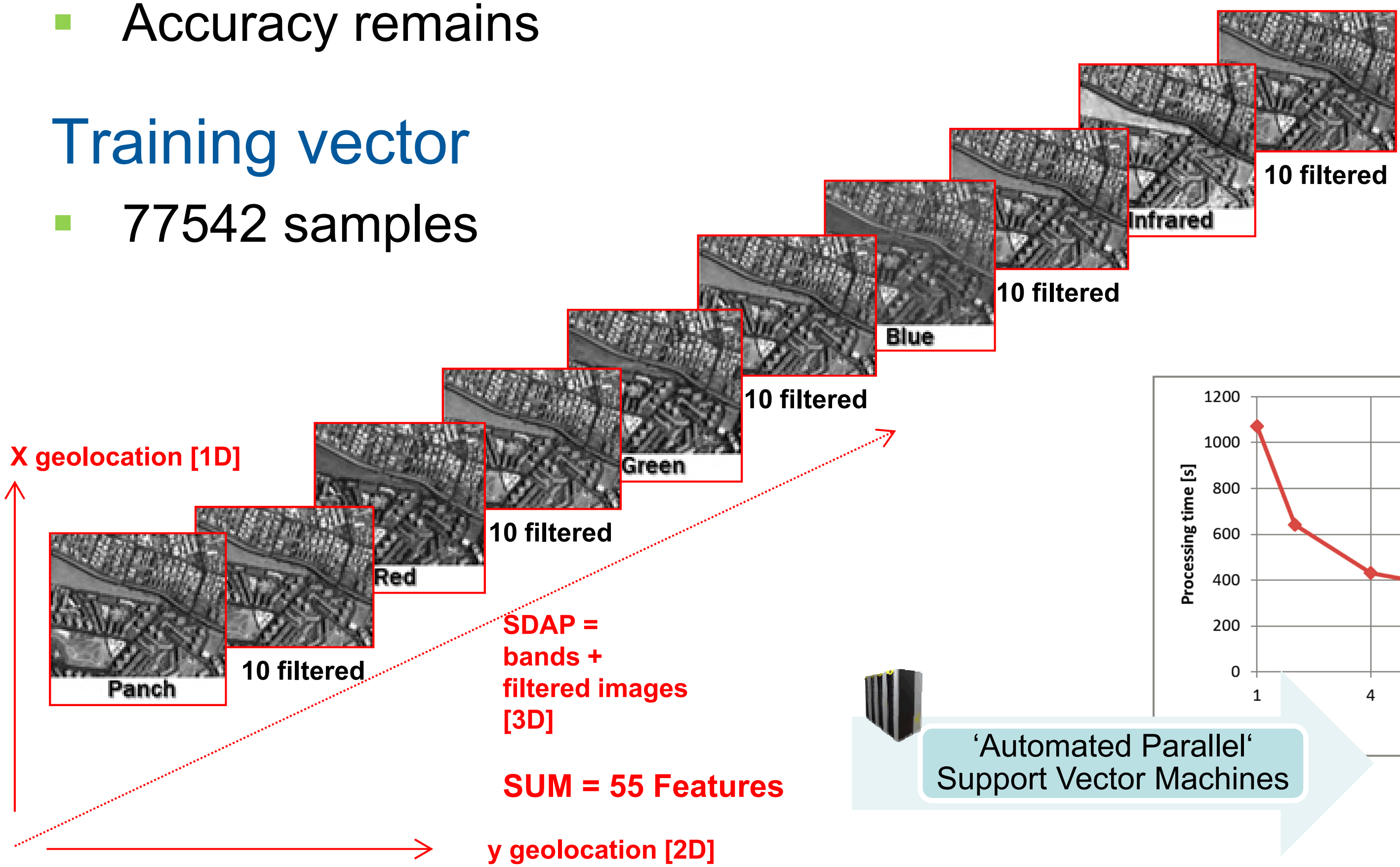
- Serial Matlab: ~1277 sec (~21 minutes)
- Parallel (16) Analytics: 220 sec (3:40 minutes)
- Accuracy remains

Training vector

- 77542 samples

Manual SDAP

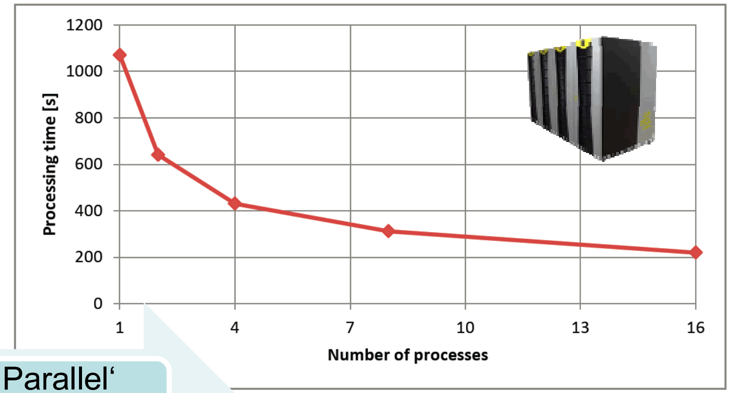
Manual work: Obtain the SDAP for all image bands using attribute 'area' (10 thresholds)



SDAP = bands + filtered images [3D]

SUM = 55 Features

'Automated Parallel' Support Vector Machines



B2SHARE
Store and Share Research Data

https://b2share.eudat.eu/record/88/

B2Share

Information | Comments | Reviews | Usage statistics

piSVM Analytics Runtimes JUDGE Cluster Rome Images 55 Features

Morris Riedel ; Gabriele Cavallaro

;

30 May 2014

http://b2share.eudat.eu

Abstract: piSVM version 1.2; configuration: -o 1024 -q 512 -c 10000 -g 16 -t 2 -m 1024 -s 0;
55 features;
SDAP build on high-resolution (0.6m) panchromatic image and on pansharpened (UDWT) low-resolution (2.4m) multispectral images using attribute area (10 threshold values)
Supplemental material for paper study.

Keyword(s): parallel SVM ; analytics ; MPI ; multi-spectral images

The record appears in these collections:
Generic

https://b2share.eudat.eu/record/88/

B2Share

B2SHARE
Store and Share Research Data

Files

Name	Date	Size	
1797203-all-1-1.txt	30 May 2014	1.1 kB	Download
1797253-all-8-1.txt	30 May 2014	1.2 kB	Download
1797240-all-4-1.txt	30 May 2014	1.2 kB	Download
1797267-all-32-1.txt	30 May 2014	1.5 kB	Download
1797230-all-2-1.txt	30 May 2014	1.2 kB	Download
1797258-all-16-1.txt	30 May 2014	1.3 kB	Download

Rate this document:

(Not yet reviewed)

Report abuse

Export

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

Metadata

PID:	http://hdl.handle.net/11304/69430fd2-e7d6-11e3-b2d7-14feb57d12b9
Publication:	http://b2share.eudat.eu
Publication Date:	2014-05-30
Uploaded by:	m.riedel@fz-juelich.de
Contact email:	m.riedel@fz-juelich.de
Domain:	generic
Checksum:	3dbc215b81f342cba96752026694c449824d99644c0afd1c59b15fece9b9905f

- **Cluster runtimes preserved and shared: Unstable runtimes, so re-try with newer code version needed**

https://b2share.eudat.eu/record/108

B2Share

B2SHARE Store and Share Research Data

EUDAT

B2SHARE Store and Share Research Data

Information | Comments | Reviews | Usage statistics

piSVM Analytics Joboutputs JUDGE Cluster Rome Images 55 Features

Morris Riedel ; Gabriele Cavallaro

;

23 June 2014

http://b2share.eudat.eu

Abstract: piSVM version 1.2; configuration: -o 1024 -q 512 -c 10000 -g 16 -t 2 -m 1024 -s 0;
55 features;
SDAP build on high-resolution (0.6m) panchromatic image and on pansharpened (UDWT) low-resolution (2.4m) multispectral images using attribute area (10 threshold values)
Supplemental material for paper study.
Correspondent job outputs for the job run times given in B2SHARE entry:
http://hdl.handle.net/11304/69430fd2-e7d6-11e3-b2d7-14feb57d12b9

Keyword(s): parallel SVM; analytics; MPI; multi-spectral images

The record appears in these collections:
Generic

https://b2share.eudat.eu/record/108

B2Share

B2SHARE
Store and Share Research Data

Files

Name	Date	Size	
Train-rome-all-32-1.o1797267.o1797267	25 Jun 2014	262.4 kB	Download
Train-rome-all-1-1.o1797203.o1797203	25 Jun 2014	110.9 kB	Download
Train-rome-all-16-1.o1797258.o1797258	25 Jun 2014	183.7 kB	Download
Train-rome-all-8-1.o1797253.o1797253	25 Jun 2014	145.7 kB	Download
Train-rome-all-4-1.o1797240.o1797240	25 Jun 2014	124.8 kB	Download
Train-rome-all-2-1.o1797230.o1797230	25 Jun 2014	115.4 kB	Download

Rate this document:

(Not yet reviewed)
Report abuse

Export

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

Metadata

PID:	http://hdl.handle.net/11304/d02f34e6-0117-11e4-81ac-dcbd1b51435e
Publication:	http://b2share.eudat.eu
Publication Date:	2014-06-23
Uploaded by:	m.riedel@fz-juelich.de
Contact email:	m.riedel@fz-juelich.de
Domain:	generic
Checksum:	cde1a6fcb3b5f31d7283e273120e46135dbf95d8c07680d269ff21b1921095d4

- **Improved stable code version runtimes shared and used in publication and to create runtime figures**

https://b2share.eudat.eu/record/115

B2Share

B2SHARE
Store and Share Research Data

EUDAT

B2SHARE
Store and Share Research Data

Information | Comments | Reviews | Usage statistics

piSVM1.2 Analytics JUDGE Cluster Rome Images 55 Features

Morris Riedel

;

03 August 2014

http://b2share.eudat.eu

Abstract: piSVM version 1.2; configuration: -o 1024 -q 512 -c 10000 -g 16 -t 2 -m 1024 -s 0;
55 features;
SDAP build on high-resolution (0.6m) panchromatic image and on pansharpened (UDWT) low-resolution (2.4m) multispectral images using attribute area (10 threshold values)
Supplemental material for paper study.
Correspondending dataset available at>
http://hdl.handle.net/11304/4615928c-e1a5-11e3-8cd7-14feb57d12b9

Keyword(s): SVM ; remote sensing ; analytics ; MPI

The record appears in these collections:

Generic

https://b2share.eudat.eu/record/115 B2Share

B2SHARE Store and Share Research Data

Files

Name	Date	Size	
1949513-checkjobinfo.el	08 Aug 2014	1.2 kB	Download
1949516-sdap_area_all_training.el.model.model	08 Aug 2014	18.7 MB	Download
Train-tune-rec86-1-16-8.o1949514.o1949514	08 Aug 2014	145.6 kB	Download
1949518-sdap_area_all_training.el.model.model	08 Aug 2014	18.7 MB	Download
1949509-submit-train-tune-record86.sh	08 Aug 2014	572 Bytes	Download
1949513-sdap_area_all_training.el.model.model	08 Aug 2014	18.7 MB	Download
Train-tune-rec86-2-16-16.o1949516.o1949516	08 Aug 2014	183.7 kB	Download
1949513-submit-train-tune-record86.sh	08 Aug 2014	572 Bytes	Download
Train-tune-rec86-8-16-64.e1950870.e1950870	08 Aug 2014	210 Bytes	Download
Train-tune-rec86-1-8-4.o1949513.o1949513	08 Aug 2014	124.7 kB	Download
1949510-submit-train-tune-record86.sh	08 Aug 2014	572 Bytes	Download
Train-tune-rec86-1-2-1.o1949509.o1949509	08 Aug 2014	110.9 kB	Download
1949514-checkjobinfo.el	08 Aug 2014	1.2 kB	Download
1949510-checkjobinfo.el	08 Aug 2014	1.2 kB	Download

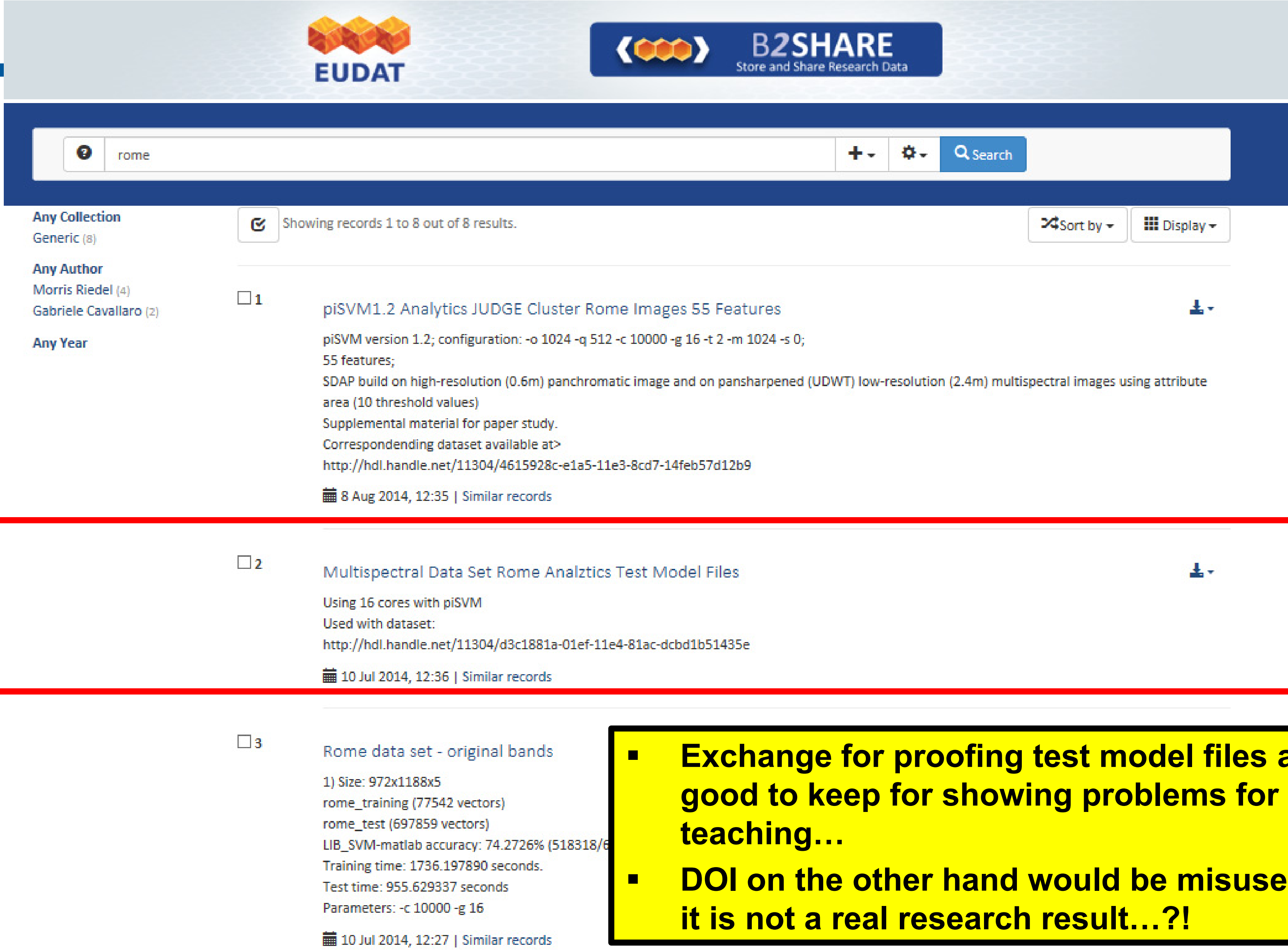
Export

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

Metadata

PID:	http://hdl.handle.net/11304/6880662c-1edf-11e4-81ac-dcbd1b51435e
Publication:	http://b2share.eudat.eu
Publication Date:	2014-08-03
Uploaded by:	m.riedel@fz-juelich.de
Contact email:	m.riedel@fz-juelich.de
Domain:	generic
Checksum:	d89cc21553d3dbecc8c0f2e2c53fcf012e46c5113038b3f6a991850b369ff78e

- **Preserving the outcomes: the trained model**
- **Towards reproducability: job scribts are stored too**



- **Exchange for proofing test model files and good to keep for showing problems for teaching…**
- **DOI on the other hand would be misuse as it is not a real research result…?!**

Study – Addressing 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
- CRISP-DM reports helps binding both together (e.g. which parameters)

→ Link to another RDA Interest Group
Talk at the RDA IG Reproducability on Tuesday 2014-09-22

Next: Publishing results and linking B2SHARE

SMART DATA ANALYTICS METHODS FOR REMOTE SENSING APPLICATIONS

Gabriele Cavallaro[a], *Morris Riedel*[a,b], *Jon Atli Benediktsson*[a], *Markus Goetz*[a,b], *Tomas Runarsson*[a], *Kristjan Jonasson*[a], *Thomas Lippert*[b]

[a] Faculty of Electrical and Computer Engineering, University of Iceland, Reykjavik, Iceland
[b] Jülich Supercomputing Center, Forschungszentrum, Jülich, Germany

ABSTRACT

The big data analytics approach emerged that can be interpreted as extracting information from large quantities of scientific data in a systematic way. In order to have a more concrete understanding of this term we refer to its refinement as smart data analytics in order to examine large quantities of scientific data to uncover hidden patterns, unknown correlations, or to extract information in cases where there is no exact formula (e.g. known physical laws). Our concrete big data problem is the classification of classes of land cover types in image-based datasets that have been created using remote sensing technologies, because the resolution can be high (i.e. large volumes) and there are various types such as panchromatic or different used bands like red, green, blue, and nearly infrared (i.e. large variety). We investigate various smart data analytics methods that take advantage of machine learning algorithms (i.e. support vector machines) and state-of-the-art parallelization approaches in order to overcome limitations of big data processing using non-scalable serial approaches.

Index Terms— Data Analytics, Support Vector Machines, Parallel Computing, Remote Sensing, Classification

1. INTRODUCTION

Besides the traditional sources and collection methods of data, with all their limitations, satellite remote sensing [1] remains one of the largest source of data collections. Remote sensing takes advantage of satellite and airborne sensors to observe, measure, and record the radiation reflected or emitted by the Earth and its environment. It can significantly enhance the information available from traditional data sources (i.e., by providing synoptic views of large portions of Earth), which can be used for subsequent data processing. The big data problem is given due to the rapid improvement of remote sensing capabilities such as the availability of remotely sensed images with very high geometrical resolution (QuickBird 0.6m). In addition detailed spectral information (AVIRIS 224 spectral channels) is constantly increasing, and the amount of data is continuously growing with images more and more numerous, precise, frequent, but also complex.

In this context, Remote sensing makes use of several analysis methods, such as image processing, automatic classification, multitemporal processing and data fusion, in order to handle different real applications. The availability of the data raises a demand for smart data analytics techniques such as image classification that is one amongst the most significant application worlds for remote sensing, but facing serious limitations when performing classification with traditional serial tools (e.g. R, Matlab, etc.). The problem of classification aims to categorize all pixels in a digital image into meaningful features or classes of land cover types in a scene. In order to obtain a satisfactory level of detection accuracy, we perform a detailed physical analysis by exploiting the availability of high spatial resolution image. We consider attribute filters, flexible operators that can transform an image according to many different attributes (e.g., geometrical, textural and spectral) as further optimization technique.

This paper offers one solution to the aforementioned described scientific case in the field of remote sensing applications by applying smart data analytics methods to one specific big dataset. We provide a scalable analytics solution for image classification taking advantage of one of the most successful classification methods referred to as support vector machines (SVMs) [2]. But in order to overcome the limitations of the wide variety of traditional serial SVM data analysis tools, we survey and apply existing open source SVM tools for big data analytics that take advantage of parallelization techniques. The contribution of this paper is thus the design of a tailored parallel smart data analytics method to the afore-described scientific case that is able to reduce the training time of the SVM classifier under the constraints of not dropping in terms of accuracy. The work has been performed and discussed within the Research Data Alliance (RDA) Big Data Analytics Interest Group (IG)[3].

This paper is structured as follows. After the introduction into the problem domain, Chapter 2 provides the necessary technical background, offers methods summaries, and surveys related work. Chapter 3 presents results from our approach and the paper ends with some concluding remarks.

- **Beta: not linked handles yet, but will be soon possible (long list at the paper end?)**

This significant reduction in training time was not affecting the training accuracy that we obtained by running also SVM predictions in parallel being always roughly 97% like the serial Matlab approach. The implementation of piSVM for basic smart analytic applications is stable enough, but we observed some limitations with respect to scaling to higher number of cores and I/O limits.

In order to support the more and more emerging approaches towards 'reproducable science' we have uploaded all datasets and the runtimes into the B2SHARE EUDAT service. Hence, the data and the piSVM implementation can be thus used to reproduce our findings in the paper. Finally, the described approach with concrete application in this paper contributes to the findings of the RDA Big Data Analytics Interest Group.

Finally future work will be the detailed investigation of other parallel implementations with a focus on the GPU-LibSVM library.

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

Unsolved: Sharing different versions of software?!

```
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**

Future Work

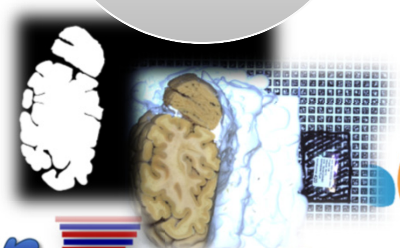
Transfer results to other scientific domains

- 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.

→ Link to the RDA Big Data Infrastructure Working Group
Talk by Shahbaz Memon about initial brain data analytics results

Brain Data Analytics



Twister
Iterative MapReduce

B2SHARE
Store and Share Research Data

hadoop

πSvM

Sattelite Data (Quickbird)

Parallel Support Vector Machines (SVM)

HPC/MPI, Map-Reduce & GPGPUs

Big Data Analytics

RDA
RESEARCH DATA ALLIANCE

Classification Study of Land Cover Types

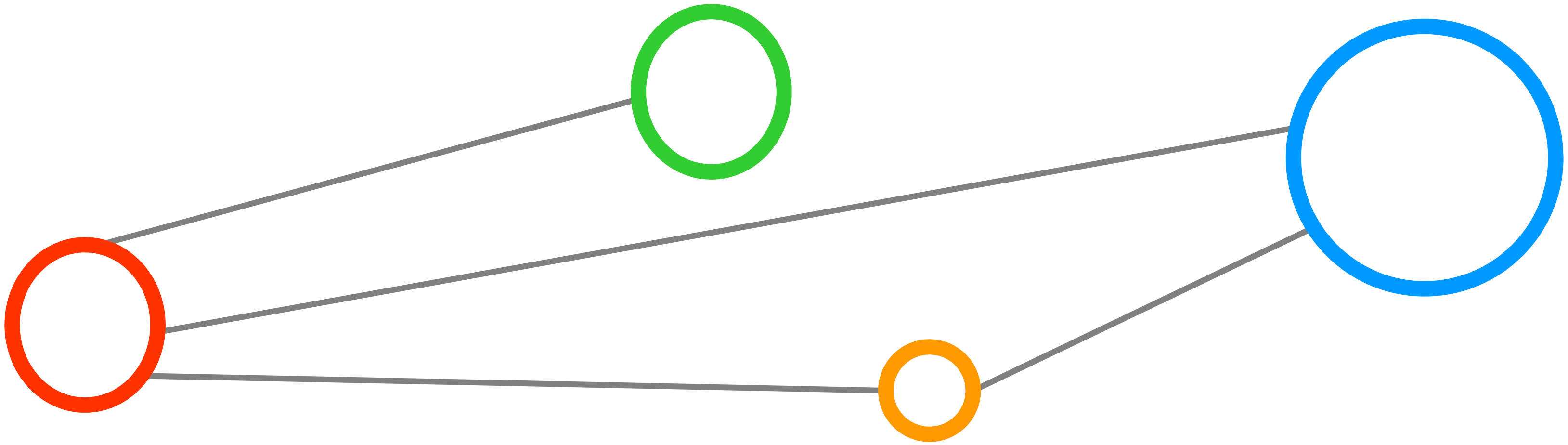
„Best Practices"

Community-based practice

„Reference Data Analytics" for reusability & learning

CRISP-DM Report	Openly Shared Datasets	Running Analytics Code

Summary



Findings in a Nutshell

Scientific Smart Data Analytics

- Often different & more complex as industrial ‘big data analytics‘ cases
- Need for sharing of ‘intermediate results‘ that may become the final result
- Demand for uploading of ‘different data versions‘ on same original data
- Challenge: Upload all data from all analytics run well with metadata – time?!

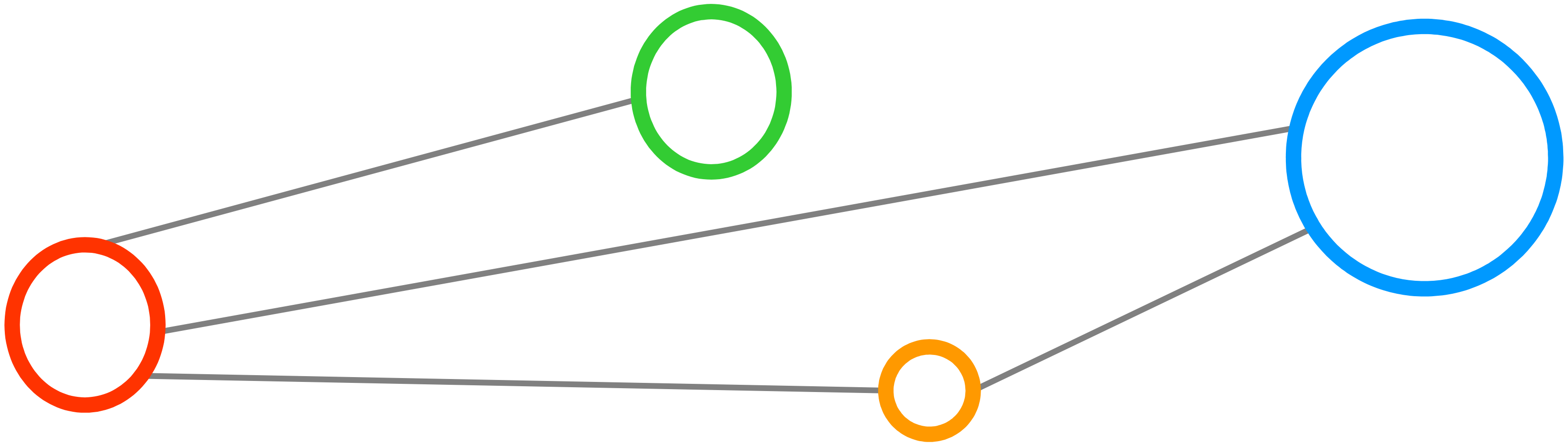
Experiences with B2SHARE

- Ideal sharing service for research groups and teaching purposes
- Assigned PID very useful (e.g. in papers) as well as unique record Ids
- Enabled trust to ‘delete data‘ & brings order to ‘messy big data‘ directories
- Using the handle is convenient (but on directory structure not required…)

Suggestions for B2SHARE

- Requirement of more flexible metadata schemes (‘communities >1 types‘)
- Recommender system integration (‘you might be also interested in…‘)
- *Where is the boundary to say ‘analytics code is also data‘?*

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.

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

Acknowledgements

Gabriele Cavallaro, University of Iceland

Tomas Philipp Runarsson, University of Iceland

EUDATB2SHARE Team

B2SHARE
Store and Share Research Data

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 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