High Productivity Data Processing Analytics Methods with Applications

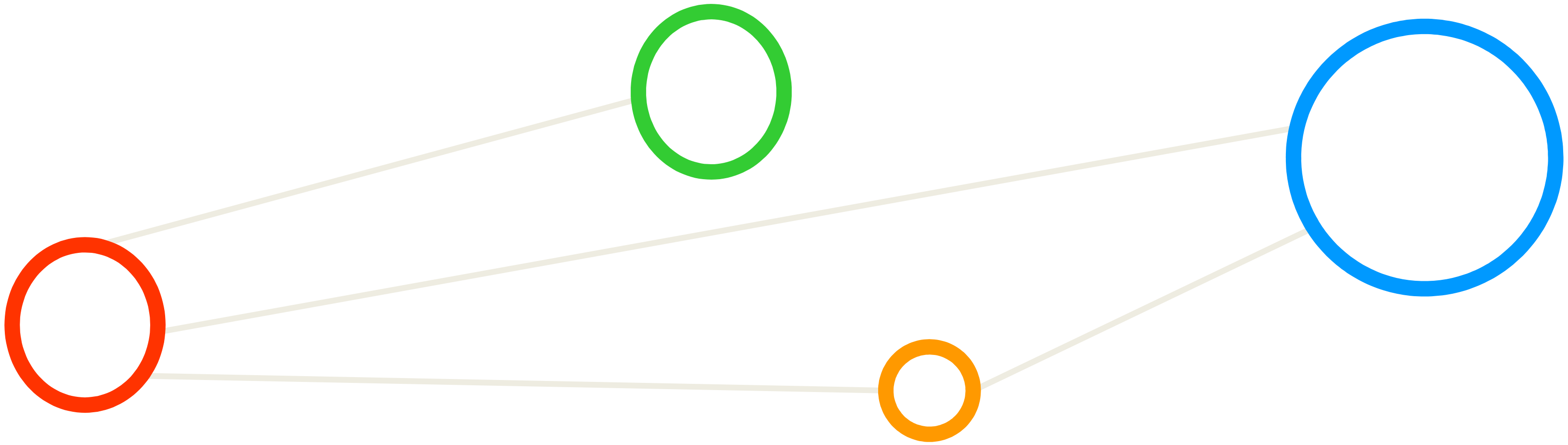
Dr. – Ing. Morris Riedel et al.
Adjunct Associate Professor
School of Engineering and Natural Sciences, University of Iceland

Research Group Leader, Juelich Supercomputing Centre
Forschungszentrum Juelich, Germany

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

JÜLICH FORSCHUNGSZENTRUM

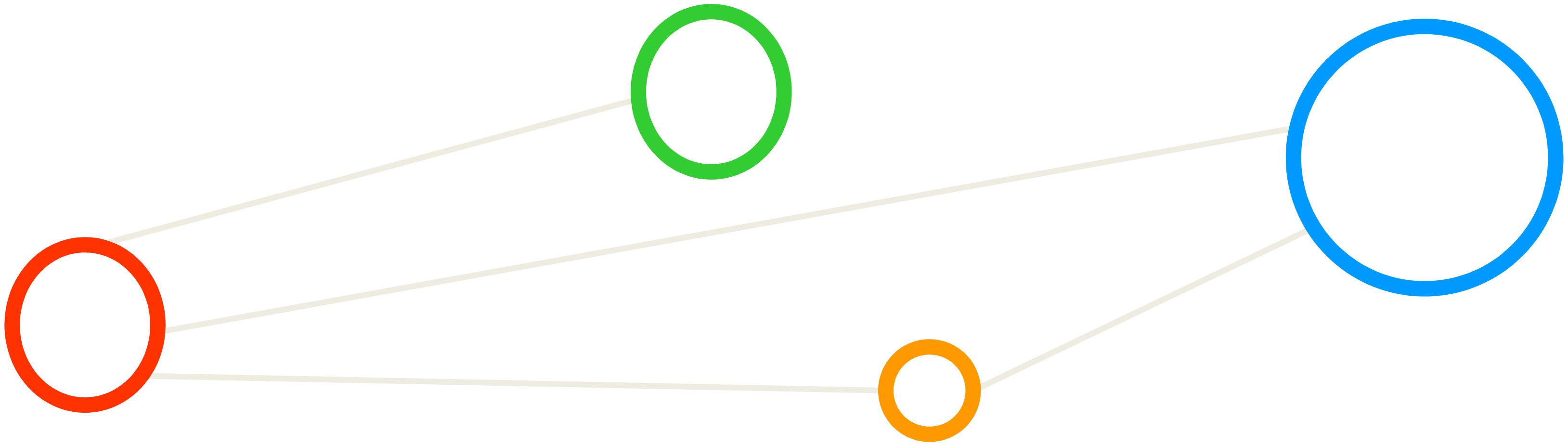
Outline



Outline

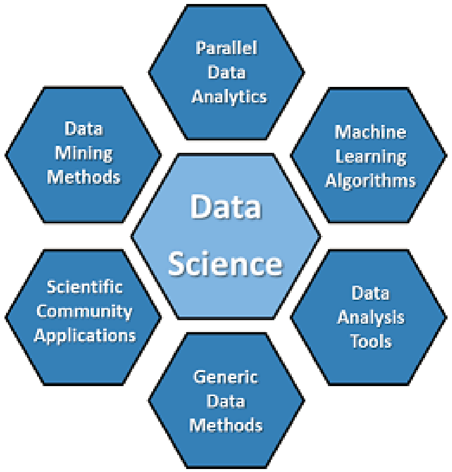
- Introduction
  - Big Data Analytics
- Smart Data Analytics Methods
  - Systematic Analytics with CRISP-DM
  - Support Vector Machines Analytics
- Example Analytics Application
  - Classification of Buildings in Images
- Conclusions & References

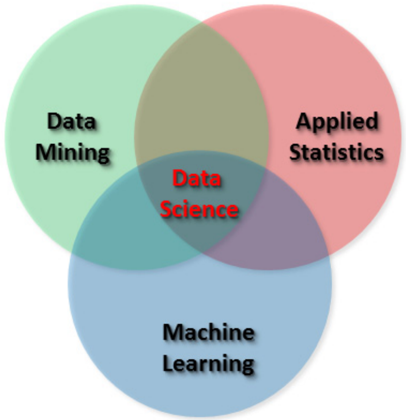
Introduction



Big Data Analytics

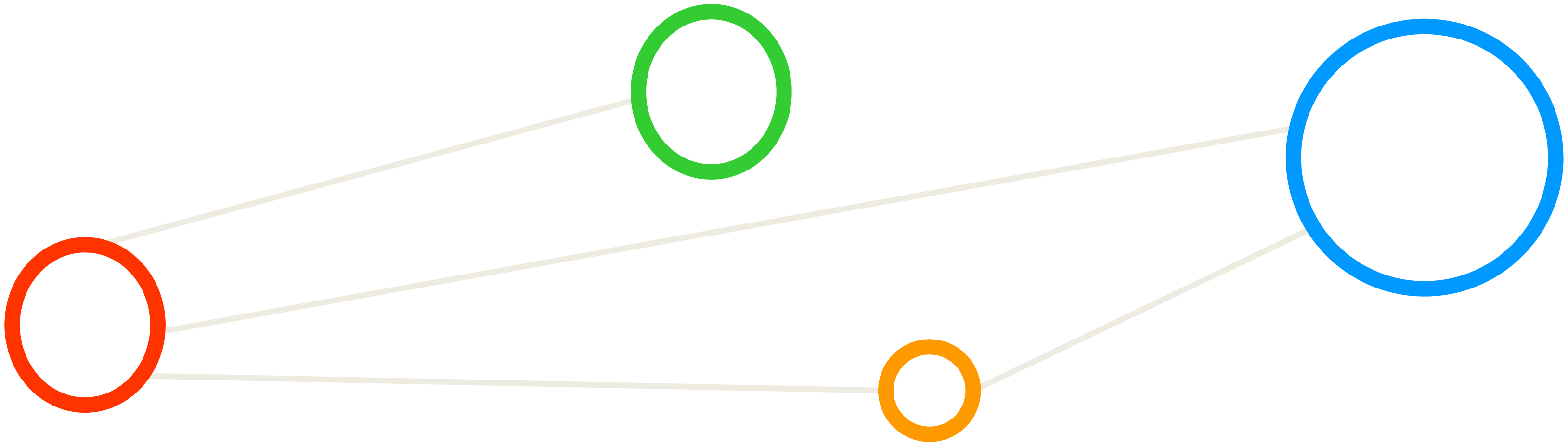
- (Automatically) examine large quantities of scientific ('big') data
  - Uncover hidden patterns
  - Reveal unknown correlations
  - Extract information in cases where there is no exact formula
- Intersection of traditional methods from a wide variety of fields





- **Use of parallelization techniques (MPI, Map-reduce, GPGPUs) offers scalability to big data sets**

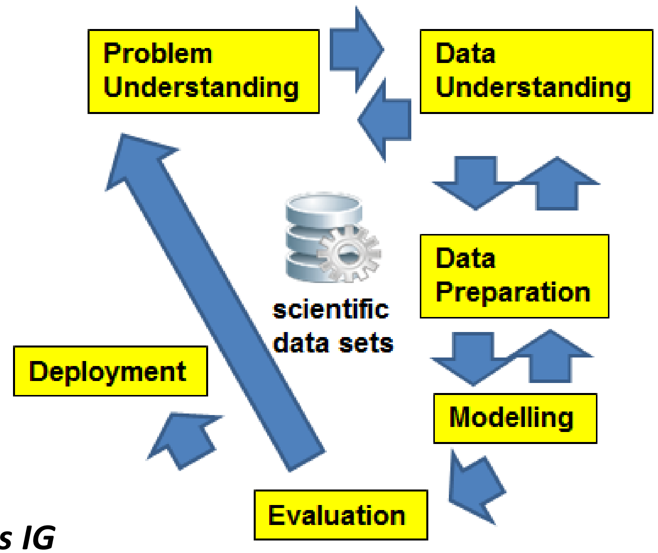
Smart Data Analytics Methods



Systematic Analytics with CRISP - DM

- Performed survey of ‘reference models‘ that enable data analytics in structured way
- Cross Industry Standard Process for Data Mining
  - Used in Research Data Alliance
  - BigData Analytics Interest Group

[7] P. Chapman et al., CRISP-DM Guide

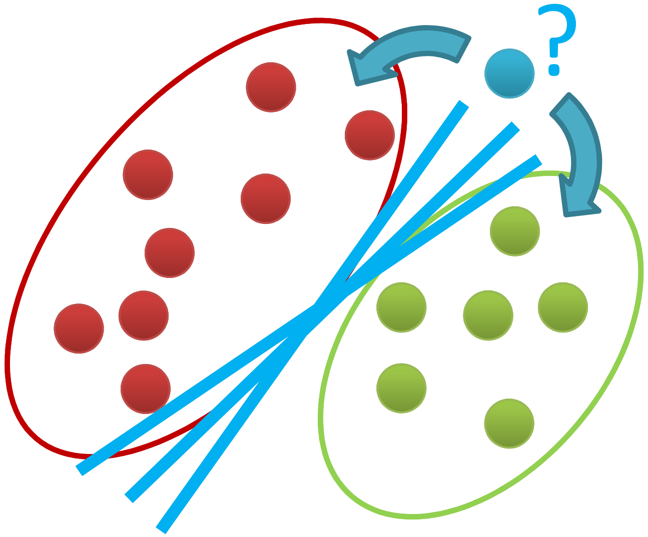


[10] RDA Big Data Analytics IG

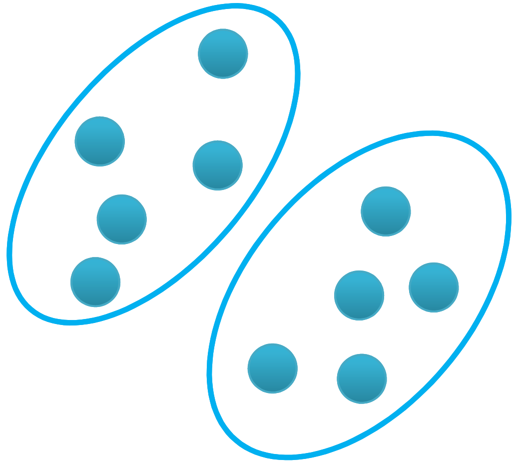
Support Vector Machines Analytics

Classification

Support Vector Machines

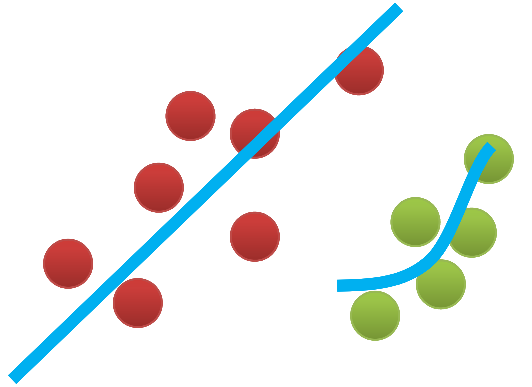


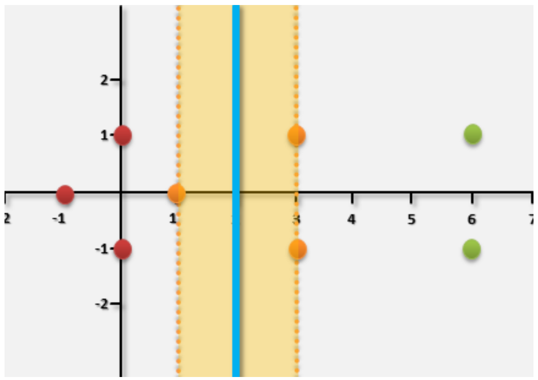
Clustering



Regression

Support Vector Machines





Quadratic Programming

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

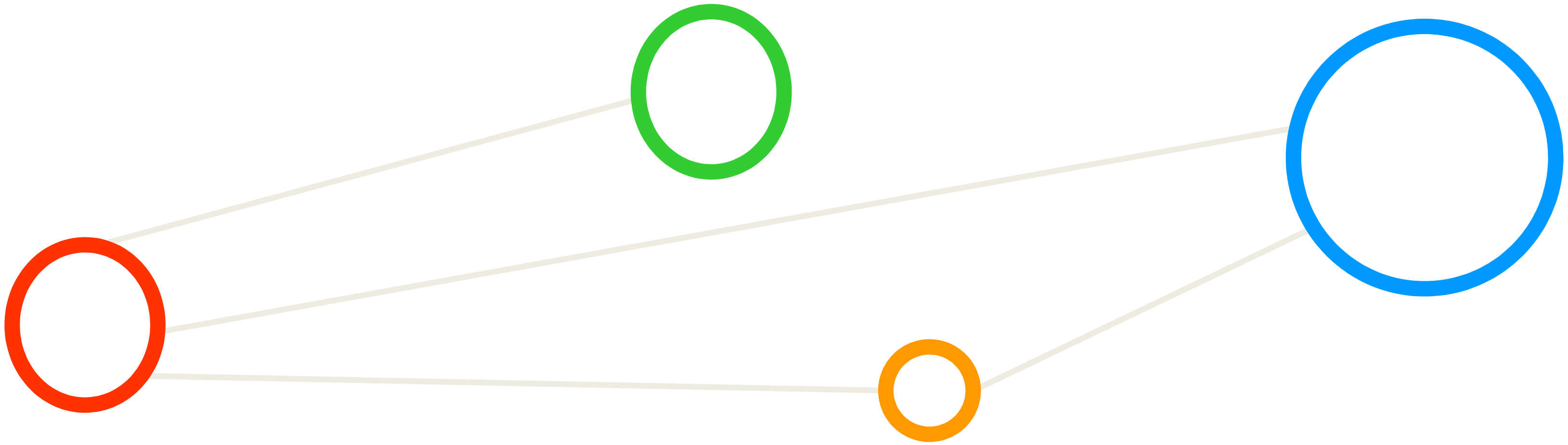
$$\begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \dots y_1 y_N x_1^T x_N \\ \dots & \dots & \ddots \\ y_N y_1 x_N^T x_1 & y_N y_2 x_N^T x_2 & \dots y_N y_N x_N^T x_N \end{bmatrix}$$

(big data challenge)

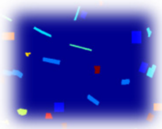
(e.g. all N datasets vs. sampling)

(quadratic coefficients → N x N dense matrix)

Example Analytics Application



Classification of Buildings in Images – (Big) Data

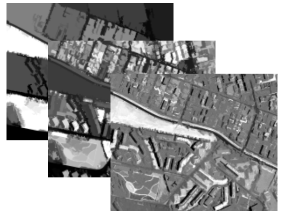


Problem: Multi-Class Classification of buildings from hyper- / multi-spectral images

panchromatic image (972 × 1188 pixels); high resolution; 0.6m

multispectral image with the four bands; low-resolution; 2.4m

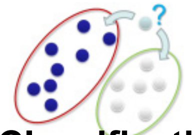
- Classification of buildings from multi-spectral data
  - 1st → Principle Component Analysis (PCA)
  - Classify building classes using image data & ‘attribute filters‘ to increase the accuracy
  - Multi-spectral images can become very large
  - Labelled data with groundtruth data exists



N profiles further improve classification feature vector (area, std deviation, moment of inertia,…)

- **Use parallel Support Vector Machines (SVMs) since it is known as good classification method today**

Classification of Buildings in Images – Toolset (1)



- Performed large survey of parallel SVM implementations (map-reduce)
  - Spark/MLlib (Map-Reduce) *[1] Spark Website*
    → only binary classification, linear SVM
  - Mahout (Map-Reduce) *[2] Mahout Website*
    → no strategy for implementation
  - ParallelSVM on Twister (Iterative Map-Reduce)
    → received beta code per email *[3] Sun Z. & Fox. G et al.*

- **Parallel implementations based on Map-Reduce are emerging but stabilility needs to be improved**

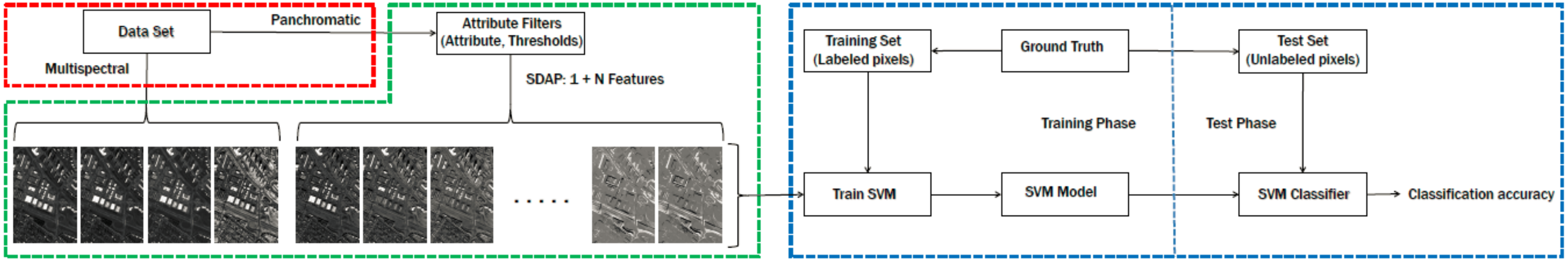
Classification of Buildings in Images – Toolset (2)

Classification++

- Performed large survey of parallel SVM implementations (MPI & GPGPUs)
  - piSVM
    → Open source code, scalability limits
    [4] piSVM Website
  - pSVM
    → Open source code, beta quality
    [5] pSVM Website
  - GPULibSVM,
    → Open source code, beta quality
    [8] GPU LibSVM

- **Parallel implementations based on MPI + GPGPUs are openly available, but show scalability limits**

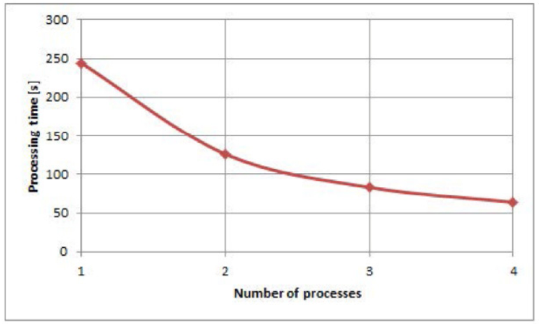
Building Classification in Images – Some Results



- Serial Matlab scripts used before
  - Not scalable to big data sets → parallelization

[6] G. Cavallaro & M. Riedel et al., 2014

- E.g. piSVM
  - Speed-up, but also shows limits



Reproducable Findings:

B2SHARE Store and Share Research Data

Data is publicly available

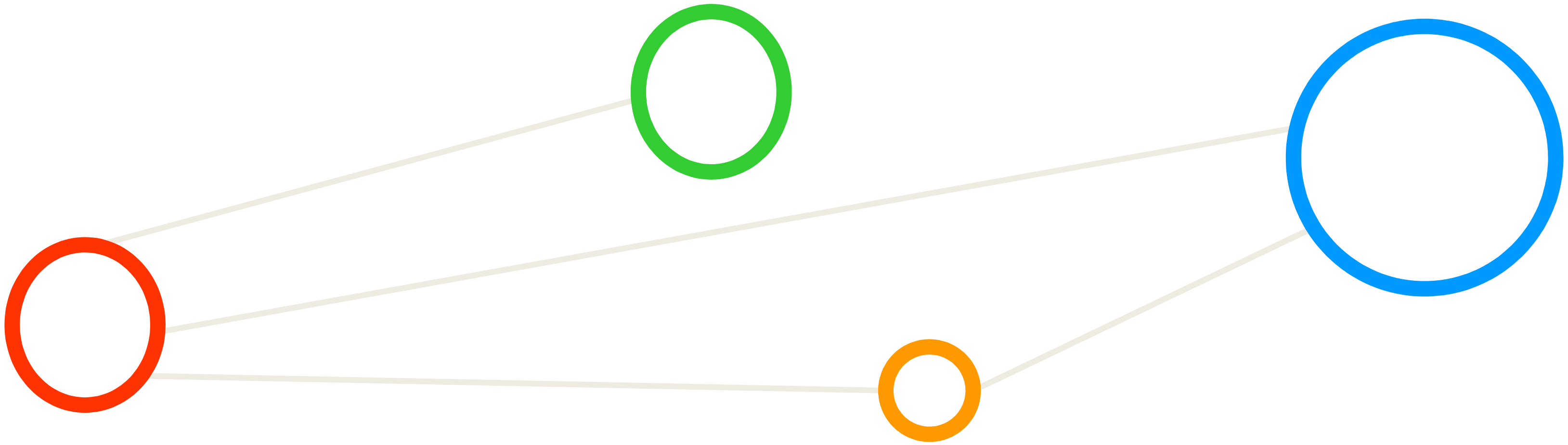
[9] Rome Dataset

Code is publicly available

[4] piSVM Website

- **Take away message from applications: Mostly multi-class SVMs used in science & engineering**

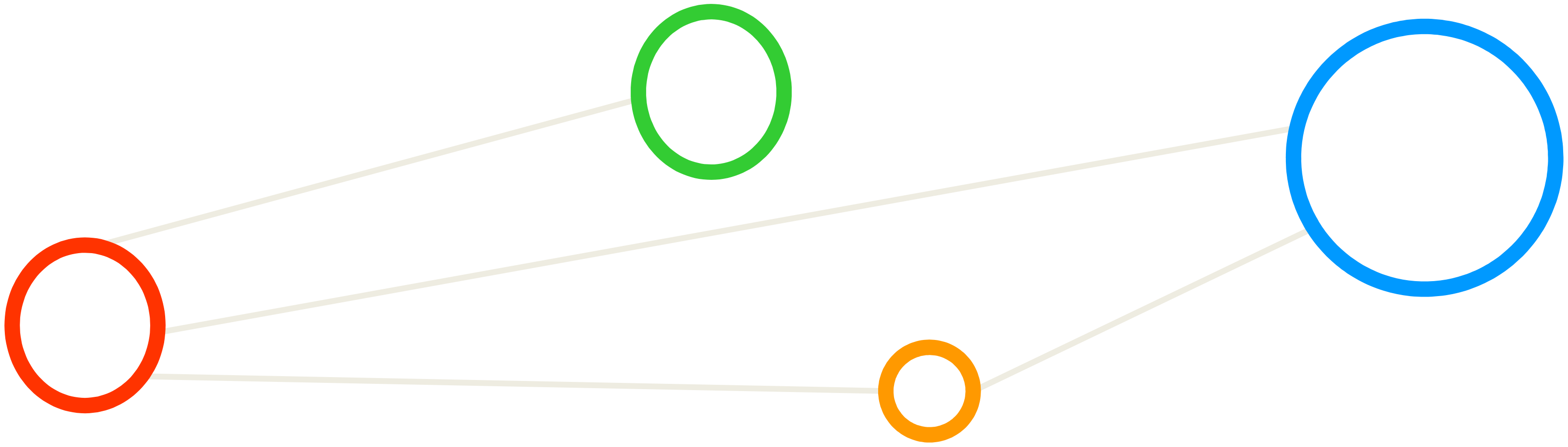
Conclusions



Conclusions

- Big Data Analytics
  - Requires smart parallel data analytics methods
  - Enables high productivity (big) data processing
  - Apply (existing) or research parallel methods
- Methods Reviewed & Applied
  - CRISP-DM guides well the systematic analytics
  - Availability of parallel implementations of analytic algorithms rare, simple, or non existent
  - SVM: Map-Reduce less stable, MPI / GPGPUs ok

References



References

[1] Spark/Mllib, Online: http://spark.apache.org/docs/0.9.0/mllib-guide.html

[2] Apache Mahout, Online: https://mahout.apache.org/users/classification/support-vector-machines.html

[3] 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.

[4] piSVM, Online: http://pisvm.sourceforge.net/

[5] Online: https://code.google.com/p/psvm/

[6] G. Cavallaro and M. Riedel, 'Smart Data Analytics Methods for Remote Sensing Applications', 35th Canadian Symposium on Remote Sensing (IGARSS), 2014, Quebec, Canada, to appear

[7] Pete Chapman, *'CRISP-DM User Guide'*, 1999, Online: http://lyle.smu.edu/~mhd/8331f03/crisp.pdf

[8] GPU-LibSVM, Online: http://mklab.iti.gr/project/GPU-LIBSVM

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

[10] Research Data Alliance, Big Data Analytics IG,
Online: https://rd-alliance.org/internal-groups/big-data-analytics-ig.html

Thanks

Talk available at:

www.morrisriedel.de/talks

Contact:

m.riedel@fz-juelich.de

Acknowledgements

Parts of the presentation have been created in close collaboration with scientific domain ‘remote sensing‘ scientists

Gabriele Cavallaro, Jon Atli Benediktsson – University of Iceland