RESEARCH DATA ALLIANCE

Interest Group Big Data Analytics

Co-Chair Dr. – Ing. Morris Riedel, Forschungszentrum Juelich

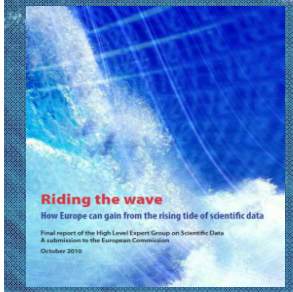
research data sharing without barriers

rd-alliance.org

13th July 2014

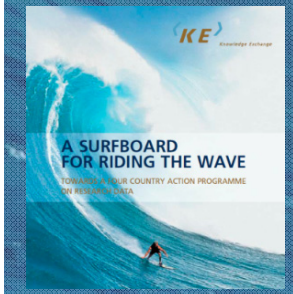
IGARSS Conference, Quebec, Canada

„Scientific Big Data Analytics“
Relevance for Scientific Research



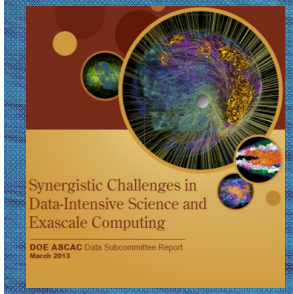
‘… problems that require high-performance data storage, **smart analytics**, transmission and mining to solve.’

[1] John Wood et al.



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

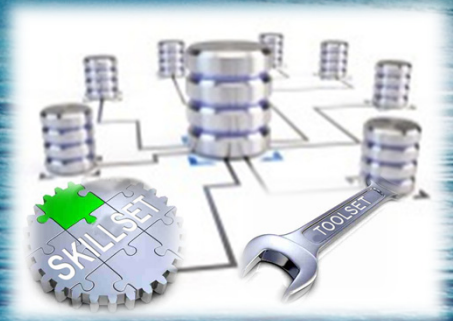
[2] KE Partners



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

[3] DOE ASCAC Report

Combine New & Old Approaches



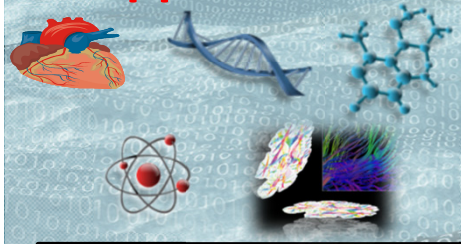
The Complex „Big Data“…

… requires concentration on most relevant areas for progress in scientific research

Hadoop 1.0 Hadoop 2.0 Spark Google DataFlow >#200 NoSQL Databases

only industrial relevant cases

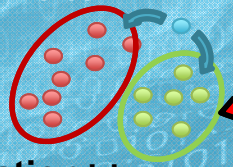
„Bottom-Up“ Scientific Data Applications



Scientific Computing

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

“Big Data”



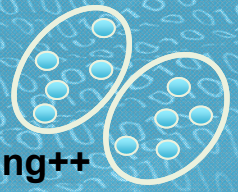
Classification++

Group Focus

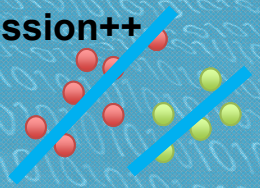
RDA RESEARCH DATA ALLIANCE

Big Data Analytics

Clustering++



Regression++



Big Data Analytics Interest Group

Facts

Members: ~60
Founded in 1st Plenary (Göteborg)
Telcons: ~1-2x / Month
Co-chairs:
Morris Riedel (JUELICH)
Kuo Kwo-Sen (NASA)
Peter Baumann (UNI BREMEN)

Concete Datasets
(& source/sensor)

Algorithms &
Methods

Technologies &
Ressources

Scientific Data Applications

Big Data Analytics Gruppenprozess

„Best Practices“

Community-based practice & recommendations

„Reference Data Analytics“ for reusability & learning

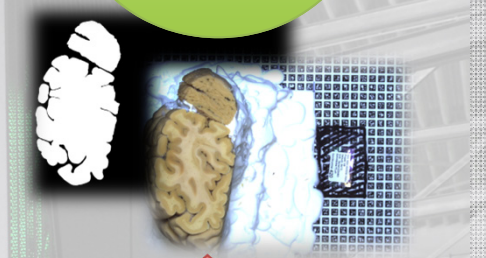
CRISP-DM Report	Openly Shared Datasets	Running Analytics Code

Big Data Analytics Interest Group

RDA RESEARCH DATA ALLIANCE

Example

Brain Data Analytics



JÜLICH FORSCHUNGSZENTRUM

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

[6] G. Cavallaro & M. Riedel et al., 'Smart Data Analytics Methods for Remote Sensing Applications', IEEE IGARSS, Quebec, Canada

Sattelite Data (Quickbird)

Parallel Support Vector Machines (SVM)

HPC/MPI, Map-Reduce & GPGPUs

Classification Study of Land Cover Types

„Best Practices"

Community-based practice (e.g. PCA, dim. reduction) & recommendations

„Reference Data Analytics" for reusability & learning

CRISP-DM Report	Openly Shared Datasets	Running Analytics Code

B2SHARE Store and Share Research Data

[5] EUDAT B2SHARE

πSVM

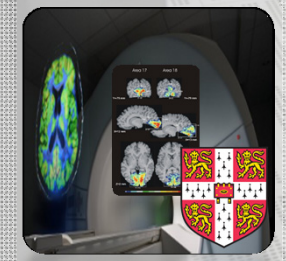
[4] piSVM

Big Data Analytics Interest Group

Other Scientific Data Applications

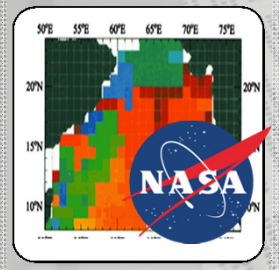
MARUM BREMEN

- Problem: „Outlier detection for automatic quality control“
- Large quantity of datasets of measurements (e.g. PANGAEA Collection)
- Tech: HPC/HTC (map-reduce?), Sybase In-DB Analytics?, …



UoCAMBRIDGE

- Problem: „High Precision Radiotherapy Treatment“
- Combined analysis of data (CT Scans, X-Rays, MRI scans, PET images)
- Tech: HPC/HTC, In-Memory/NoSQL- Datebases – which makes sense?



NASA - MSFC

- Problem: „Event tracking analytics“ (bspw. Entstehung von Somali Jets)
- Satellite datasets (‘search events with changes geolocations‘)
- Tech: HPC/HTC (map-reduce?), Twister/Harp?, SCIDB, NASA software,…

Thanks for your attention

Join us in the Sessions @ 4th Plenary Meeting Amsterdam

References

[1] John Wood et al., 'Riding the Wave –How Europe can gain from the rising tide of scientific data', EC Report, 2010
[2] KE Partners, 'A Surfboard for Riding the Wave - Towards a four country action programme on research data', November 2012
[3] DOE ASCAC Data Subcommittee Report, 'Synergistic Challenges in Data-Intensive Science and Exascale Computing', 2013
[4] piSVM Sourceforge Open Source Tool for Parallel Classification, Online: http://pisvm.sourceforge.net
[5] EUDAT European Data Infrastructure, B2SHARE Tool, Online: https://b2share.eudat.eu/
[6] G. Cavallaro & M. Riedel et al., 'Smart Data Analytics Methods for Remote Sensing Applications', IEEE IGARSS, Quebec, Canada