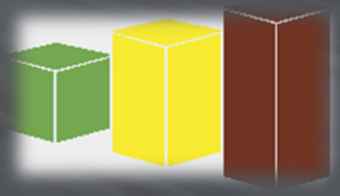
RESEARCH DATA ALLIANCE



# Interest Group Big Data Analytics

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

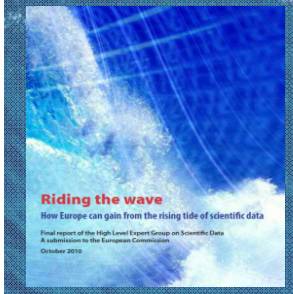
research data sharing without barriers

rd-alliance.org

7. Juli 2014

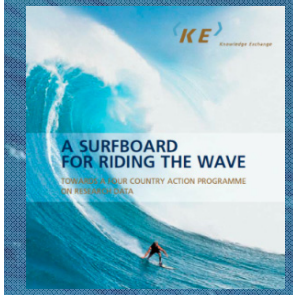
Koordinationsgespräch HGF – RDA, HGF Geschäftsstelle Berlin

# „Scientific Big Data Analytics“
## Relevanz für die Großforschung



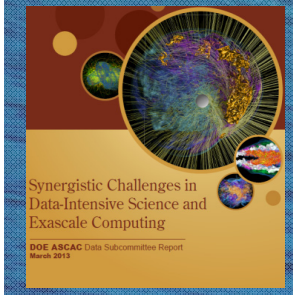

‘… problems that require high-performance data storage, **smart analytics**, transmission and mining to solve.’

***[1] John Wood et al.***




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

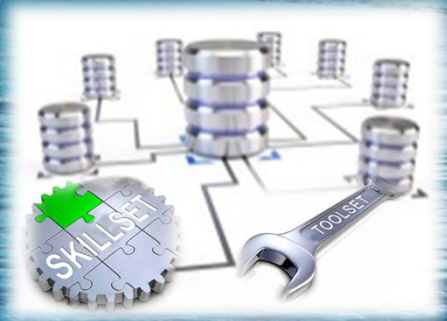
***[2] KE Partners***




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

***[3] DOE ASCAC Report***

***Teil einer stabilen Dateninfrastruktur***

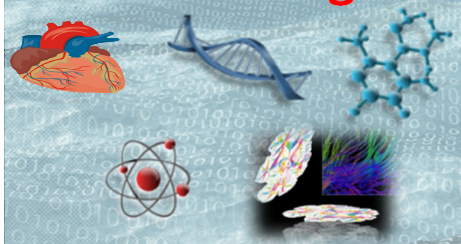


# Der Komplex „Big Data“…

*… erfordert Konzentration auf wesentliche Bereiche zum Fortschritt der Wissenschaft*

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

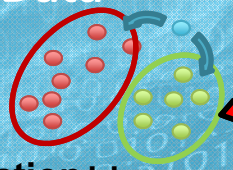
## „Bottom-Up“ Wissenschaftliche Anwendungen



## Scientific Computing

## “Statistical Data Mining”
## Maschinelles Lernen
## Prinzipien wie Parallelisierung
## Neue HPC/HTC Algorithmen
## Anwendbare Werkzeuge

## “Big Data”

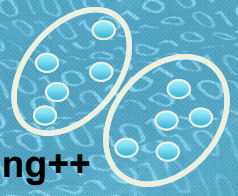


**Classification++**

## Gruppenfokus

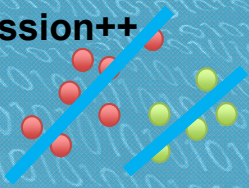
RESEARCH DATA ALLIANCE

Big Data Analytics

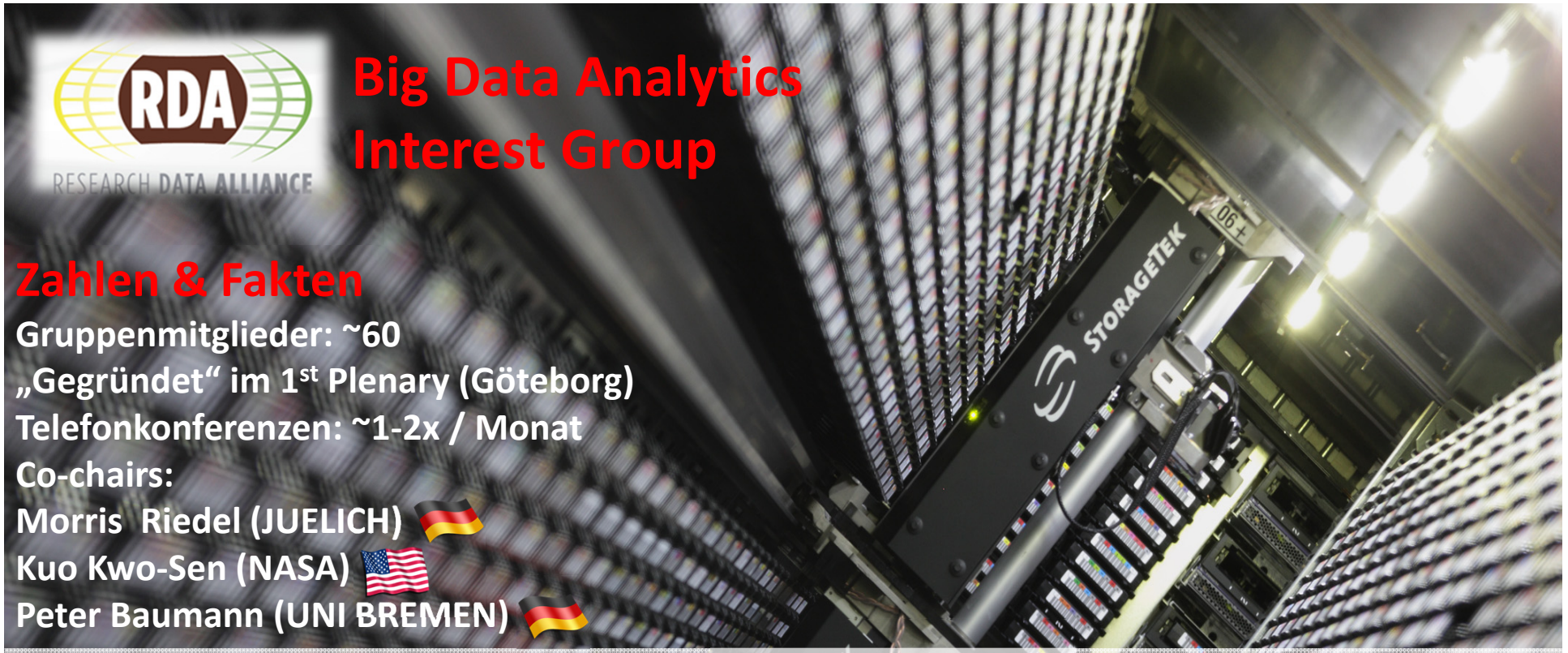



**Clustering++**

**Regression++**



# Big Data Analytics Interest Group




## Zahlen & Fakten

Gruppenmitglieder: ~60
„Gegründet“ im 1st Plenary (Göteborg)
Telefonkonferenzen: ~1-2x / Monat
Co-chairs:
Morris Riedel (JUELICH)
Kuo Kwo-Sen (NASA)
Peter Baumann (UNI BREMEN)

Konkrete Datensätze

Algorithmen & Methoden

Technologien & Ressourcen

**Wissen-schaftliche Anwendung**

Big Data Analytics Gruppenprozess

„Best Practices“

Community-basierte Empfehlungen

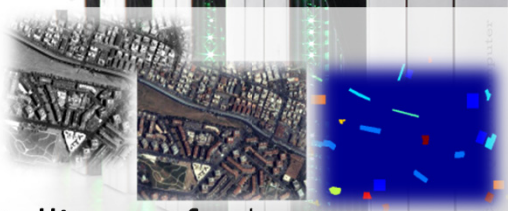
„Reference Data Analytics“ zur Wiederverwendung

| CRISP-DM Report | Openly Shared Datasets | Running Analytics Code |
|---|---|---|

# Big Data Analytics Interest Group

RDA RESEARCH DATA ALLIANCE

## Beispiel



Satellitenaufnahmen

Parallel Support Vector Machines (SVM)

HPC/MPI, Map-Reduce & GPGPUs

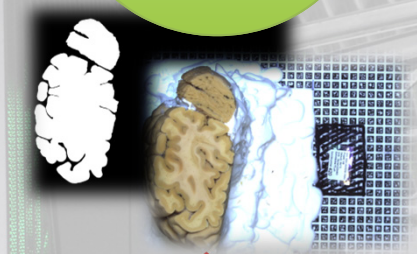
**Classification Study of Land Cover Types**

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*

**Brain Data Analytics**



„Reference Data Analytics" zur Wiederverwendung

CRISP-DM Report

Openly Shared Datasets

Running Analytics Code

„Best Practices"

Community-basierte Empfehlungen

B2SHARE Store and Share Research Data

*[5] EUDAT B2SHARE*

πSVM

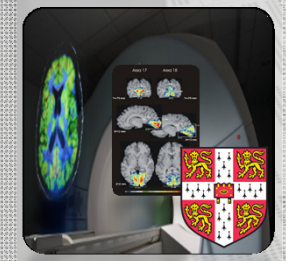
*[4] piSVM*

Big Data Analytics
Interest Group

# Weitere Wissenschaftliche Anwendungen

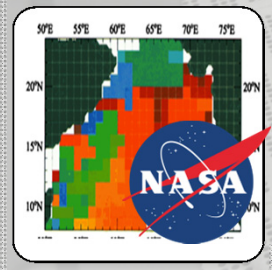
MARUM BREMEN

- Problem: „Outlier detection for automatic quality control“
- Große Anzahl Datensätze von Messungen (bspw. PANGAEA Kollektion)
- Tech: HPC/HTC (map-reduce?), Sybase In-DB Analytics?, …



UoCAMBRIDGE

- Problem: „High Precision Radiotherapy Treatment“
- Kombinierte Analyse von Daten (CT Scans, X-Rays, MRI scans, PET images)
- Tech: HPC/HTC, In-Memory/NoSQL- Datebases – welche sinnvoll?, …




NASA - MSFC

- Problem: „Event tracking analytics“ (bspw. Entstehung von Somali Jets)
- Datensätze von Sateilliten (‘Suche Events mit wechselnden Geolocations‘)
- Tech: HPC/HTC (map-reduce?), Twister/Harp?, NASA software stacks,…

# Danke für Ihre Aufmerksamkeit

## Besuchen Sie unsere Sessions @ Plenary 4 Amsterdam

## Referenzen


[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