

Arbeitsprozess und Notwendigkeit einer RDA IG am Beispiel der

Interest Group Big Data Analytics

Morris Riedel et al., Forschungszentrum Juelich

research data sharing without barriers
rd-alliance.org

20. November 2014

RDA – Deutschland Treffen, Deutsches GeoForschungszentrum, Potsdam

„[Scientific] Big Data Analytics“

Notwendigkeit einer konkreten RDA Interest Group



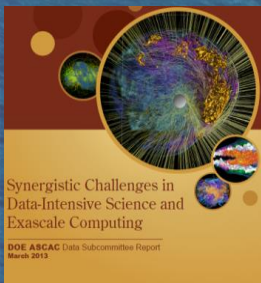
‘... problems that require high-performance data storage, **smart analytics**, transmission and mining to solve.’

[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...’

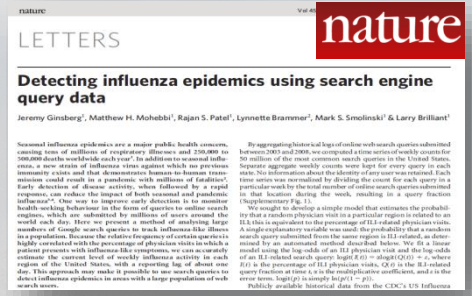
[3] DOE ASCAC Report



2009 – H1N1 Virus Made Headlines

Nature paper von Google Mitarbeitern

- (1) Erklärt wie Google Wintergrippen schnell vorhersagen kann
- (2) Nicht nur auf nationaler Ebene, auch in Regionen
- (3) Möglichkeit durch „logged big data“ (Suchanfragen)



[7] Jeremy Ginsberg et al., 'Detecting influenza epidemics using search engine query data', Nature 457, 2009

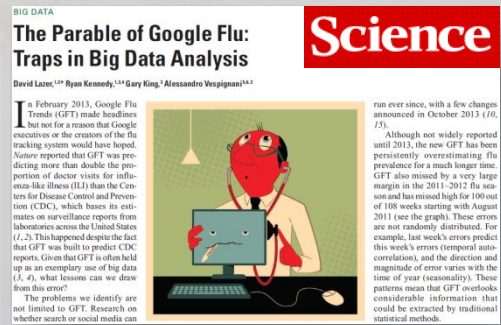
“Big Data Ansätze können täuschen“



2014 – The Parable of Google Flu

Große Fehler in Vorhersage von Wintergrippen & Lessons learned

- (1) Daten im wandel: Transparenz & Reproduzierbarkeit unmöglich
- (2) Analyse von Ansätzen & Algorithmen den sie ändern sich
- (3) Verfahren in der Community – nicht Größe nur entscheidend



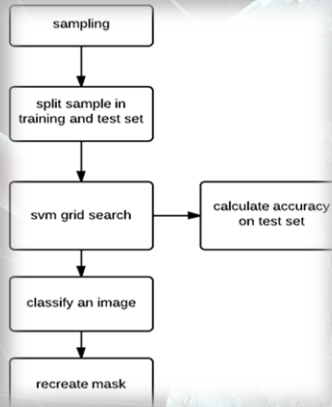
[8] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, 'The Parable of Google Flu: Traps in Big Data Analysis', Science Vol (343), 2014



Analytics Prozess... ... ist komplex

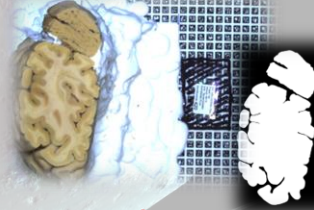


Klassifikation Beispiel



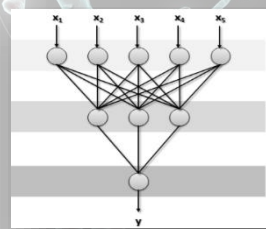
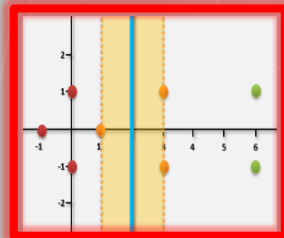
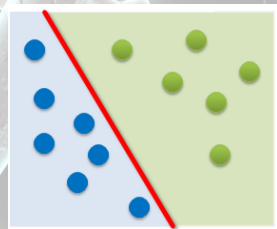
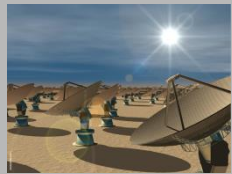
Einige Herausforderungen

‘Sampling bias’:
mehr Schwarz als Weiss
Pixel in „groundtruth“
„Hintergrundproblem“:
Gehirnschnitte im Eis



Kombination vieler Bereiche

Community-basierte Methoden (→ right)
Skills & Toolset zur Datenanalyse (→ below)
Skalierbares Toolset & Infrastruktur (→ left)



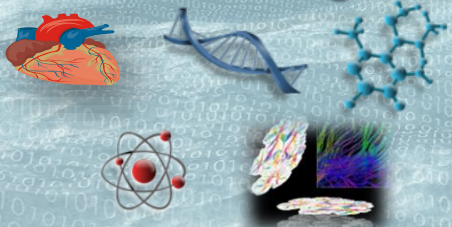


Der Komplex „Big Data“...

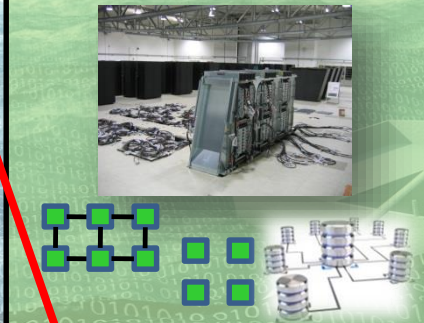
... erfordert Konzentration auf wesentliche Bereiche zum Fortschritt der Wissenschaft

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

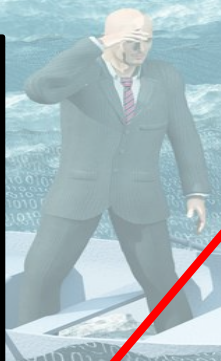
**„Bottom-Up“
Wissenschaftliche
Anwendungen**



Scientific Computing



**“Statistical Data Mining”
Maschinelles Lernen von Daten
Prinzipien wie Parallelisierung
Neue HPC/HTC Algorithmen
Skalierbare Werkzeuge (“Big”)
Offene & verfügbare Ansätze**



Gruppenfokus



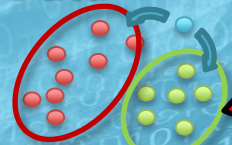
„Open“



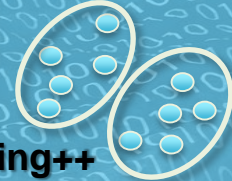
“Big Data”



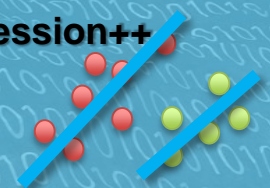
Classification++



Clustering++



Regression++





Big Data Analytics Interest Group Fakten & Arbeitsprozess

Zahlen & Fakten

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

Konkrete Datensätze



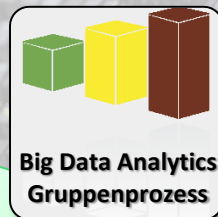
Algorithmen & Methoden



Technologien & Ressourcen



**Wissenschaftliche
Anwendung**



„Best Practices“

Community-
basierte
Empfehlungen



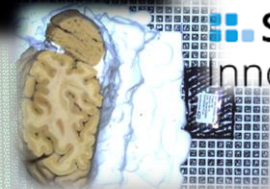
„Peer-Review“

„Reference Data Analytics“
zur Wiederverwendung

Report (CRISP-DM) 	Openly Shared Datasets 	Running Analytics Code
--------------------------	----------------------------	----------------------------

Konkretes Beispiel

Brain Data Analytics



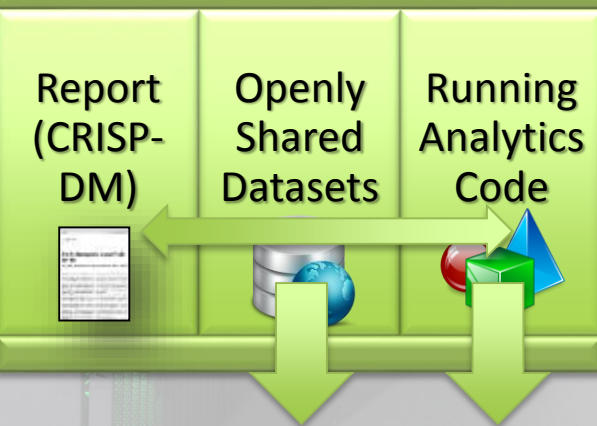
Smart Data
Innovation Lab

IT Gipfel
Pressemappe



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

„Reference Data Analytics“
zur Wiederverwendung



„Peer-Review“

Classification „Best Practices“
Study of
Land Cover
Types

Community-
basierte
Empfehlungen

Satellitenaufnahmen

Parallel
Support Vector
Machines (SVM)

HPC/MPI, Map-
Reduce & GPGPUs



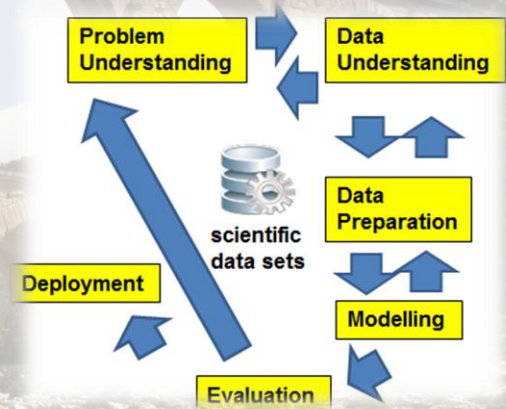
[5] EUDAT B2SHARE

[4] piSVM



RESEARCH DATA ALLIANCE

Big Data Analytics IG &
Spin-off Big Data Infrastructure WG



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

Arbeitsprozess “Systematic Big Data Analytics”

Geleitet durch Cross Industry Standard Process
for Data Mining (CRISP-DM) Phasen

“Erstellung eines UCI Repository ...

...für [Scientific] Big Data Analytics“

„Reference Data Analytics“
for reusability & learning

CRISP-
DM
Report



Openly
Shared
Datasets



Running
Analytics
Code





RESEARCH DATA ALLIANCE

Big Data Analytics IG &
Spin-off Big Data Infrastructure WG

„Reference Data Analytics“
for reusability & learning

CRISP-
DM
Report

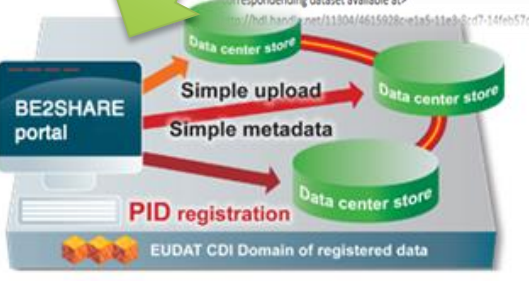
Openly
Shared
Datasets



Running
Analytics
Code



SW 1.2 Analytics JUDGE Cluster Rome Images 55 Features
VM version 1.2; configuration: -p 1024 -q 512 -c 10000 -g 16 -t 2 -m 1024 -s 0;
5 Features;
DAP build on high-resolution (0.6m) panchromatic image and on pansharpened (UDV
res (10 threshold values)
Supplemental material for paper study.
Corresponding dataset available at:
<https://hdl.handle.net/11304/6615028/e1a5f11a9-0c7-14feb57d12b9>



[5] EUDAT B2SHARE

Link zu anderen Gruppen:
bspw. RDA Reproducibility IG

Analytics Input – Vortrag in Session im Plenary 4
Gemeinsame Session in Plenary 5 vorgeschlagen

Arbeitsprozess in Richtung
„Reproduzierbare Big Data Analytics“

Link zu anderen Projekten:
Plenary 4 APARSEN Workshop
Vortrag: Data Sharing Experiences of Smart
Data Analytics Tasks in Remote Sensing Research

Essentielle Bestandteile des Arbeitsprozesses

„Peer-Review“



„Best Practices“



Community-
basierte
Empfehlungen

*Einige
Beispiele für
Earth Science*

Referenzierbare Papiere bei Peer-Review Community-Konferenzen & Tutorials ...

... bspw. bei International Geoscience and Remote Sensing Symposium (IGARSS) 2014

Organisation und Teilnahme von “Big Data Sessions “ bei Community-Events...

... American Geophysical Union (AGU) 2013...

... European Geosciences Union (EGU) General Assembly 2014...

Talk: Research Data Alliance: Understanding Big Data Analytics Applications in Earth Science

Teilnahme an wichtigen Community-Events

Nutzung des Prinzips Peer-Review

Internationalität

Analytics Experten

DataONE

CODATA

OGF NASA NIST



*Interesse aus
South Africa*

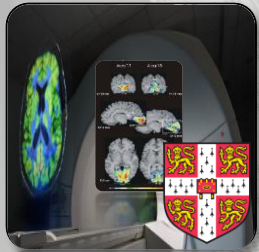
Big Data Analytics Interest Group

Weitere Wissenschaftliche Anwendungen



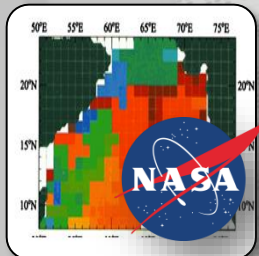
MARUM BREMEN

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



UoCAMBRIDGE

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



Big Data Analytics Interest Group

Weitere Punkte zum Arbeitsprozess

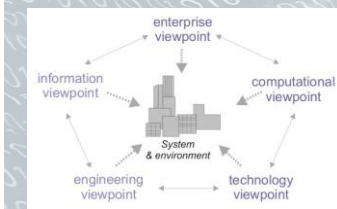


Teilnahme & Vorschlag im letzten 4 Plenary im RDA TAB/IG/WG Meeting



Gruppensekretär:
Danke! Andere?

Each section below pertains to one RM-ODP viewpoint. First an explanatory introduction is given (in *italics*), then the Big Data Analytics use cases follow. (Please add your use cases by **using the template** and **along this schema** and **fit it into the appropriate section!**)



Enterprise viewpoint

The enterprise viewpoint focuses on the purpose, scope and policies for the system. It describes the scientific requirements and how to meet them, abstracting away from technical details.

Responsible section editor: Kuo Sen-Kwo

load-balance "admin work"

Science domain use cases found relevant for Big Data Analytics:

- **Event Analysis** (Dr. Tom Clune, Dr. Kuo GSFC/NASA)
- **Particle physics and radiotherapy** (Michael Simmons and Charles Boulton, University of Cambridge)
- **Research on Data Analytics for Automated Quality Control of Measurement Data** (Morris Riedel, Shiraz Memon, Shahbaz Memon - Juelich Supercomputing Centre; Robert Huber - MARUM Bremen)
- **Spatio-Temporal Data in the Earth Sciences** (Peter Baumann, Jacobs University)
- **Classification of land cover types of remote sensing satellite image datasets** (Gabriele Cavallaro - University of Iceland, Morris Riedel - Juelich Supercomputing Centre)

Aktualisierung Webseite (bspw. mit aktuellen Themen: ist für analytics das ISO Reference Model for Open Distributed Processing (RM-ODP) sinnvoll?)

[11] RDA Big Data Analytics Web Page

Session-Qualität: Gelistete Vorträge mit PDF

Presentations - Day 1

Morris Riedel - Classification Techniques in Remote Sensing Research using Smart Data Analytics

- Forschungszentrum Jülich, Germany and University of Iceland, Reykjavik
- Remote sensing use case, classify satellite image pixels into land-cover classes (trees, soil, ...)
- Parallel support vector machine (SVM) technology comparison



Shabaz Memmon - Ultrascan and Brain Analytics

- Forschungszentrum Jülich, Germany
- Two use cases:
 - Ultracentrifugation experimentation in the analysis of biological and synthetic macromolecules
 - Human brain functionality understanding based on cross section block images



- Ultrascan (US3) based on MPI
- Random Forests and SVM with RBF kernels

Analytics Experte



Geoffrey Fox - Data Analytics at Digital Science Center @SOIC

- Indiana University, Bloomington, USA
- Integration of HPC with the Apache Big Data Stack
- 16 different use cases

Presentations - Day 2

Wo Chang - RDA Big Data Activities

- National Institute for Standards and Technology (NIST), Gaithersburg, USA
- Presentation of the NIST Big Data Analytics Reference Architecture
- Fingerprint Analysis Use Case

NIST



Kuo Kwo-Sen - Automated Identification of Episodes of Earth Science Phenomena

- Bayesian & NASA, Maryland, USA
- In-array database analytics using SciDB
- Blizzard Observation Use Case

NASA



Peter Baumann - Use Case Array Databases

- Jacobs University, Munich, Germany
- Agile array analytics on top of the rasdaman array database
- Various use cases, including stock trading, social networks, climatology, HEP and genomes



Hugh Shanahan - NGS Transcriptomic Workflows

- Royal Holloway, University of London, UK
- NGS Transcriptomic Workflow from trace data to experimentation
- Sequencing pipeline and storage



Frank Seinstra - Big Compute meets Big Data

- Netherlands eScience Center, Amsterdam, The Netherlands
- eSTeP - eScience Technology Platform (AEther, Xenon, ...), bridging science and industry
- Different use cases (supernova detection, climate research, ...)



Big Data Analytics Interest Group


Nächste Schritte & Aktuelle Diskussionen

„Zuhören & Open Minded bleiben für neue Vorschläge“



- Voranbringen des Peer-review systems (Wer liest/prüft Analytics – Resultate,...)
- Weitere Communities einbringen (bspw. Earth Science Fokus, Medizin, andere ...)
- Bessere Vernetzung mit anderen Gruppen (bspw. Terminologies and Foundations)
- Erweiterung der IG Richtung "Big Data" (bspw. analytics nur noch als Teilbereich/WG)
- Diskussionen um die ISO RM-ODP Methode zur Beschreibung (mehr als CRISP-DM, ...)
- Vorbereitung der Sessions im Plenary 5 in San Diego (mit anderen Gruppen, "Uptake")

Sampling vs. Big Data



Methods & Tools


Parallelization!

Applied Statistics

Data Mining

Machine Learning Algorithms

Scientific Computing



new DBs

Training Data Scientists

Computational Scientist

Software Engineer

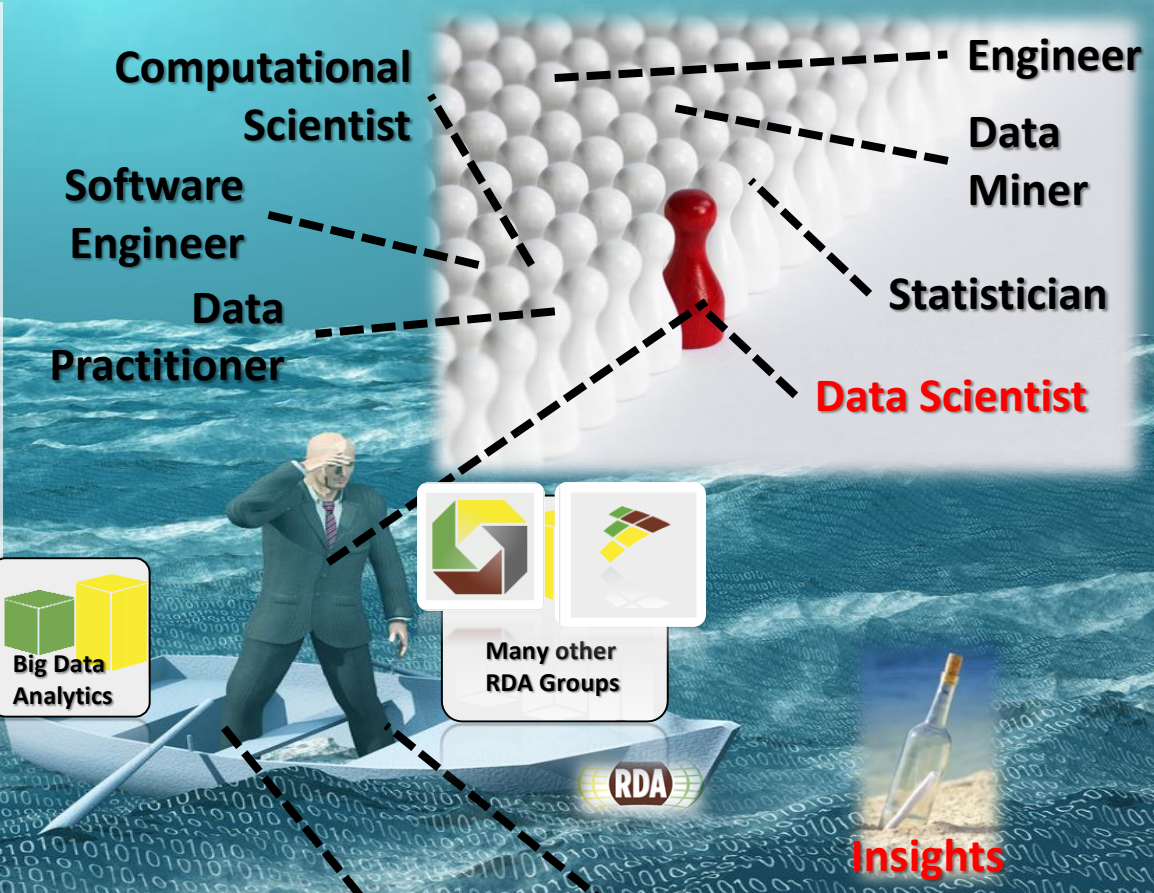
Data Practitioner

Engineer

Data Miner

Statistician

Data Scientist



Big Data Analytics

Many other RDA Groups

RDA

Insights

'...and don't forget the big picture...'

PhD Student works with NASA on SciDB Analytics



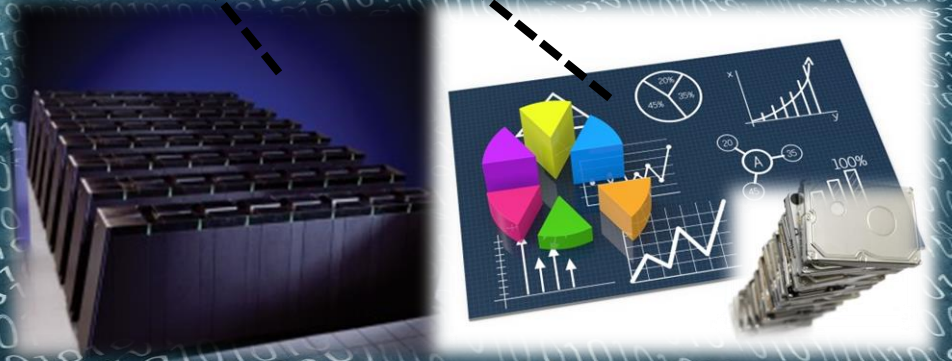

UNIVERSITY OF ICELAND
SCHOOL OF ENGINEERING AND NATURAL SCIENCES
FACULTY OF INDUSTRIAL ENGINEERING,
MECHANICAL ENGINEERING AND COMPUTER SCIENCE

Statistical Data Mining Course

HPC – B(ig Data) Course

HPC – A(advanced) Scientific Computing Course

Data Scientists with skills of various fields



Danke für Ihre Aufmerksamkeit & Dank an RDA für „Plattform“
Gestalten Sie unsere Sessions @ Plenary 5 in San Diego mit!

Talk verfügbar unter

www.morrisriedel.de/talks

Kontakt:

m.riedel@fz-juelich.de

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
- [7] Jeremy Ginsburg et al., 'Detecting influenza epidemics using search engine query data', Nature 457, 2009
- [8] D. Lazer, R. Kennedy, G. King & A. Vespignani, 'The Parable of Google Flu: Traps in Big Data Analysis', Science Vol (343), 2014
- [9] P. Chapman et al., CRISP-DM Guide
- [10] EGU 2014 Session on Big Data, Online: <http://meetingorganizer.copernicus.org/EGU2014/orals/15065>
- [11] RDA Reproducibility IG, Online: <https://rd-alliance.org/groups/reproducibility-ig.html>
- [12] Research Data Alliance, Big Data Analytics IG, Online: <https://www.rd-alliance.org/node/165/all-wiki-index-by-group>

