On Establishing Big Data Breakwaters with Analytics

Dr. - Ing. Morris Riedel

Head of Research Group 'High Productivity Data Processing', Juelich Supercomputing Centre, Germany

Adjunct Associated Professor, University of Iceland, Iceland

Co-chair Research Data Alliance, Big Data Analytics Interest Group

Big Data Waves – Surfboards – Breakwaters

How to engage in the rising tide of big data?

Unsolved Questions:

Scale
Heterogeneity
Stewardship
Curation
Long-Term Access and Storage

Research Challenges:

Collection, Trust, Usability
Interoperability, Diversity
Security, *Smart Analytics,*
Education and training
Data publication and access
Commercial exploitation
New social paradigms
Preservation and sustainability

[1] HLEG Report

[2] KE Report

'Concrete'
Next Steps →

'Looking into Big Data Waves'

Volume

Variety

Velocity

Context

Understanding 'Big Data Waves'

Big Data Streams with 'high velocity' ...

← Data Sources

'Web Actions'

*Data stream with **data** 'hidden' in logs/events*

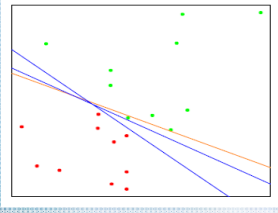
'Measurement Devices'

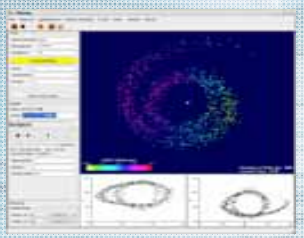
*Data stream with measured **data***

'Computational Simulations'

*Data stream with **data** of HPC/HTC simulation*

n optional filters





← Data Sinks

‘Crowdsourcing‘...

...increases # of Big Data Streams

Usual Citizens / ‘Citizen Scientist’

Data streams with **data (low trust)**

Exabytes

Individuals with domain as Hobby

Data streams with **data (moderate trust)**

Petabytes

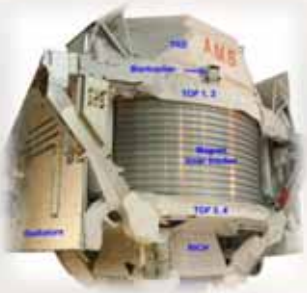
Scientific/Engineering Domain Experts

Data streams with **data (high trust)**

Terabytes

Alpha Magnetic Spectrometer (AMS) @ ISS

What are building blocks of the Universe?



Search for Cosmic Antimatter

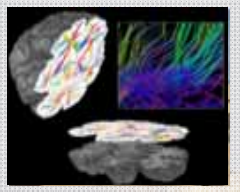
20 TB

~400 TB

JUROPA++
1-2 Pflop/s + Booster

JSC is main facility for AMS computing in Germany

Selected Juelich Supercomputing Centre (JSC) Research Examples



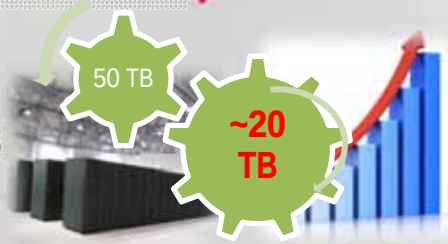
Simulation Labs (e.g. Climate SimLab)...

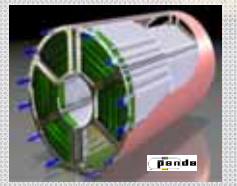
...one example out of many JSC SimLabs

~10 years
500 HDF Files
~50 TB

50 TB

~20 TB





Applications with combined characteristics of simulations and analytics...

Towards Exascale

The next generation radio telescope for science...

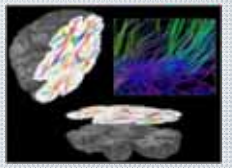
... pushing the limits of the observable universe out by billions of galaxies

The square kilometre array

... 1 PB in 20 seconds

LOFAR
test site Jülich

xilostudios.com



Large-scale Computational Massively Parallel Applications simulate Reality

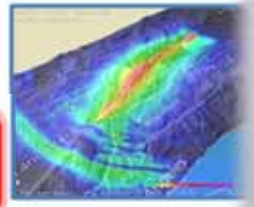
Better Prediction Accuracy…

… means ‘Bigger Data‘

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	National Super Computer Center in Guangzhou China	Tianhe-2 (MilkyWay-2) - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	Titan - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	Sequoia - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890

TOP 500 HPC Systems 11/2013

Estimated figures for simulated 240 second period, 100 hour run-time	TeraShake domain (600x300x80 km^3)	PetaShake domain (800x400x100 km^3)
Fault system interaction	NO	YES
Inner Scale	200m	25m
Resolution of terrain grid	1.8 billion mesh points	2.0 trillion mesh points
Magnitude of Earthquake	7.7	8.1
Time steps	20,000 (.012 sec/step)	160,000 (.0015 sec/step)
Surface data	1.1 TB	1.2 PB
Volume data	43 TB	4.9 PB



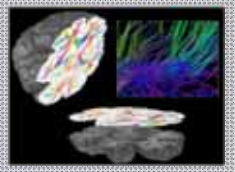
F. Berman: ‘Maximising the Potential of Research Data’

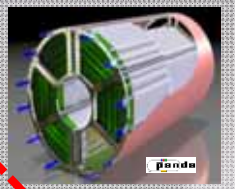
Information courtesy of the Southern California Earthquake Center

Rensselaer

‘We are unable to store the output data of all computational simulations/users’

[3] F. Berman



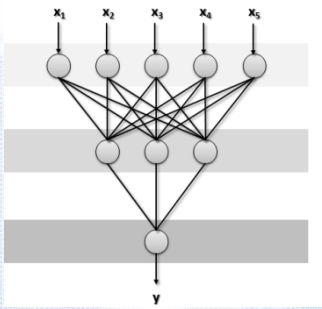


Making use of Big Data is important...

... but How?

Netflix Prize Challenge ~2009

Challenge: Improve movie recommender
Prize: 1 M $ for team with 10% improvements
Open historical data to 'train a new system'
Traditional Machine Learning Algorithms used
Prize received with Artificial Neural Networks



It becomes a strategic differentiator in the market

H1N1 Virus Made Headlines ~2009

Nature paper from Google employees
Explains how Google is able to predict flus
Not only national scale, but down to regions
Possible via logged big data - 'search queries'

~1998-today

[4] Jeremy Ginsburg et al., 'Detecting influenza epidemics using search engine query data', Nature 457 (2009)

Smart Analytics are Needed to Take Advantage of Big Data

The challenge is to understand which analytics make sense

'Understanding climate change, finding alternative energy sources, and preserving the health of an ageing population are all cross-disciplinary problems that require high-performance data storage,

***smart analytics**, transmission and mining to solve.'*

[1] HLEG Report

'In the data-intensive scientific world,

new skills are needed for

creating, handling,

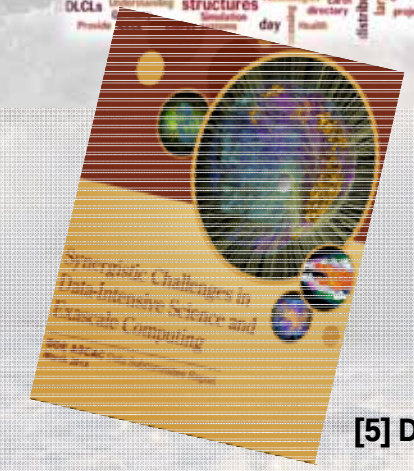
manipulating, analysing,

and making available large amounts of data for re-use by others.'

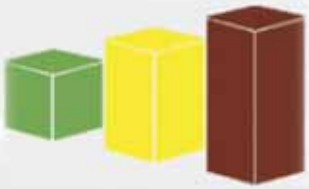
[2] KE Report

'Integration of data analytics

with exascale simulations represents a new kind of workflow that will impact both data-intensive science and exascale computing.'



[5] DOE ASCAC Report



RESEARCH DATA ALLIANCE

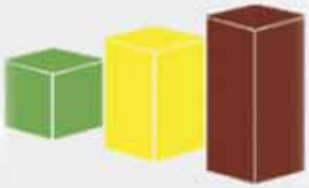
Big Data Analytics Interest Group

Can we establish a

'neutral group gathering clear evidence'

which analytics make sense in Research?

- ✓ Develops community based recommendations on feasible data analytics approaches to address community needs of utilizing large quantities of data
- ✓ Analyzes different scientific research domain applications and their potential use of various big data analytics techniques
- ✓ A systematic classification of feasible combinations of analysis algorithms, analytical tools, data and resource characteristics and scientific queries will be covered in these recommendations



Big Data Analytics Interest Group

- Agricultural Data Interoperability IG
- **Big Data Analytics IG**
- Brokering IG
- Certification of Digital Repositories IG
- Community Capability Model WG
- Data Citation WG
- Data Foundation and Terminology WG
- Data in Context IG
- Data Type Registries WG
- Defining Urban Data Exchange for Science IG
- Digital Practices in History and Ethnography IG
- Engagement Group IG
- Legal Interoperability IG
- Long tail of research data IG
- Marine Data Harmonization IG
- Metadata IG
- Metadata Standards Directory WG
- PID Information Types WG
- Practical Policy WG
- Preservation e-Infrastructure IG
- Publishing Data IG
- Standardization of Data Categories and Codes IG
- Structural Biology IG
- Toxicogenomics Interoperability IG
- UPC Code for Data IG
- Wheat Data Interoperability WG

[6] Research Data Alliance Big Data Analytics IG Web page

RESEARCH DATA ALLIANCE

Big Data Analytics Interest Group

New Technology Impact As Challenge for Users

'NoSQL Databases Example'

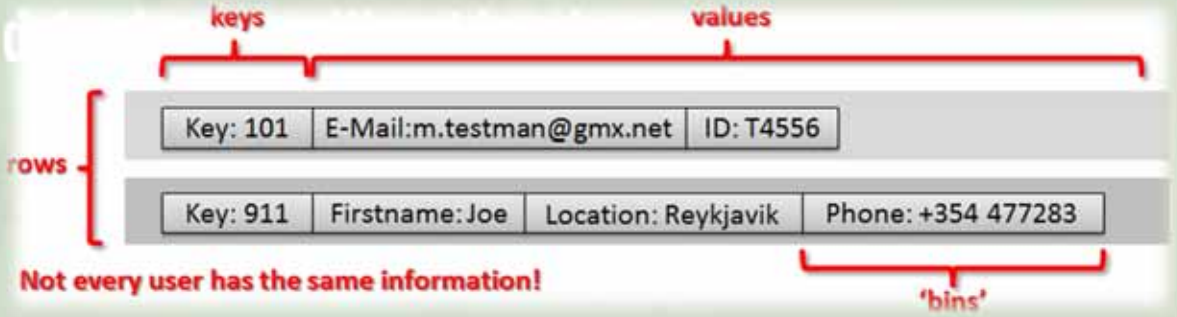
Selected Features

Simplicity of design and deployment
Horizontal scaling
Less constrained consistency models
Finer control over availability
Simple retrieval and appending
...

Types

Key-Value-based (e.g. Cassandra)
Column-based (e.g. Apache Hbase)
Document-based (e.g. MongoDB)
Graph-based (e.g. Neo4J)

'String-based Key-Value Stores'

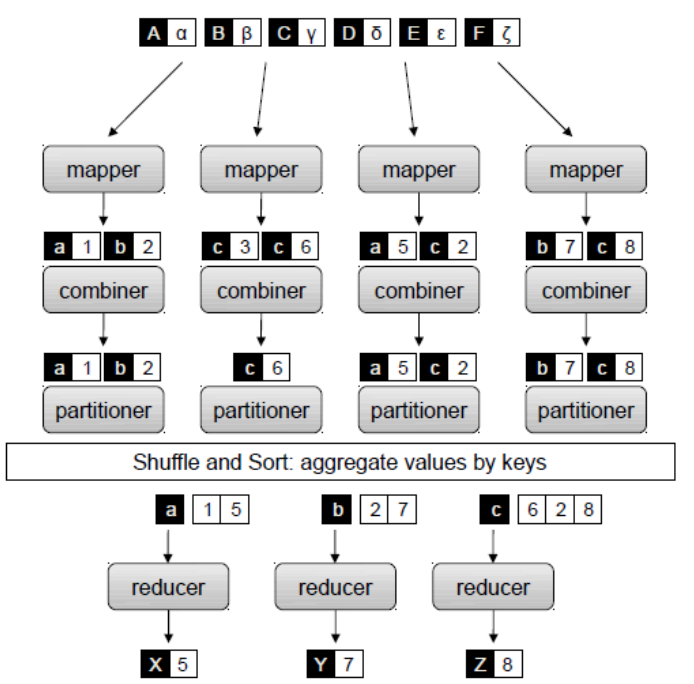


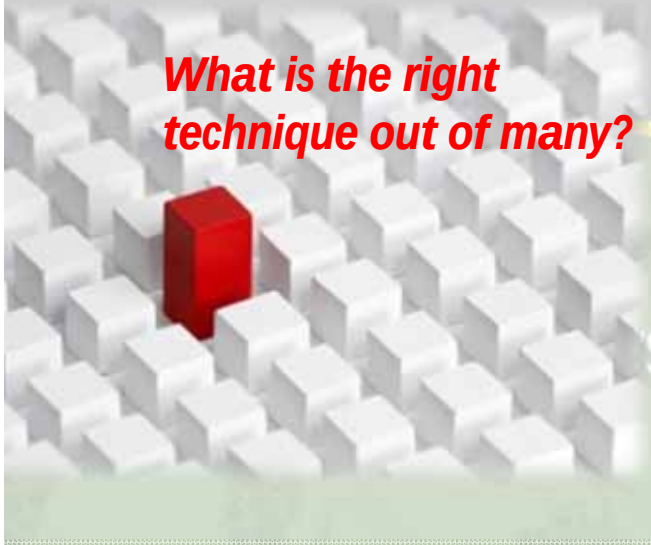
RESEARCH DATA ALLIANCE

Big Data Analytics Interest Group

New Technology Impact As Challenge for Users

'Map-Reduce Paradigm vs. Traditional HPC'





National Institute of Standards and Technology

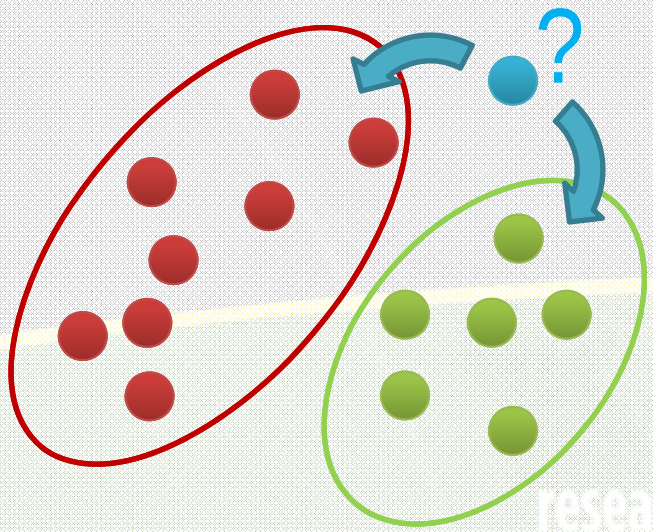
[7] Computer Vision using Map-Reduce

RESEARCH DATA ALLIANCE

Big Data Analytics Interest Group

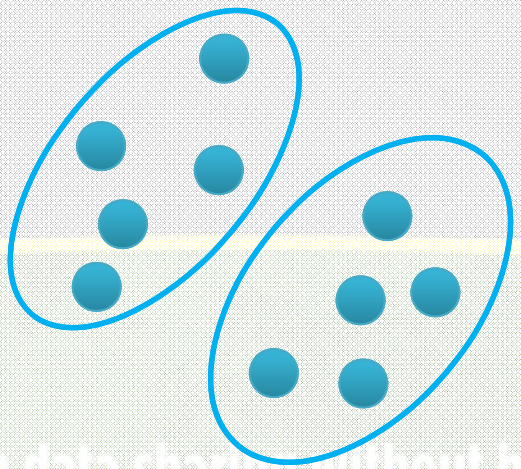
Selected Data Analytics Techniques

Classification



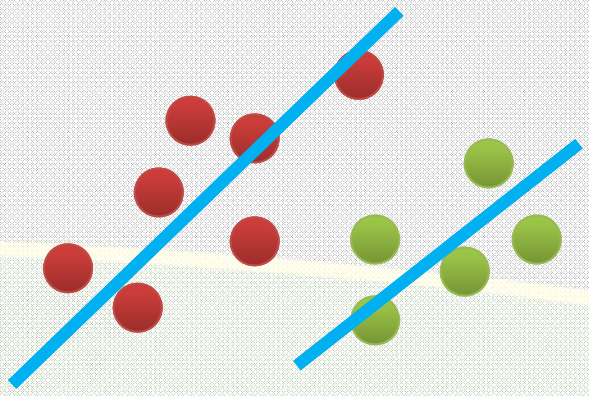
- ✓ *Groups of data exist*
- ✓ *New data classified to existing groups*

Clustering



- ✓ *No groups of data exist*
- ✓ *Create groups from data close to each other*

Regression



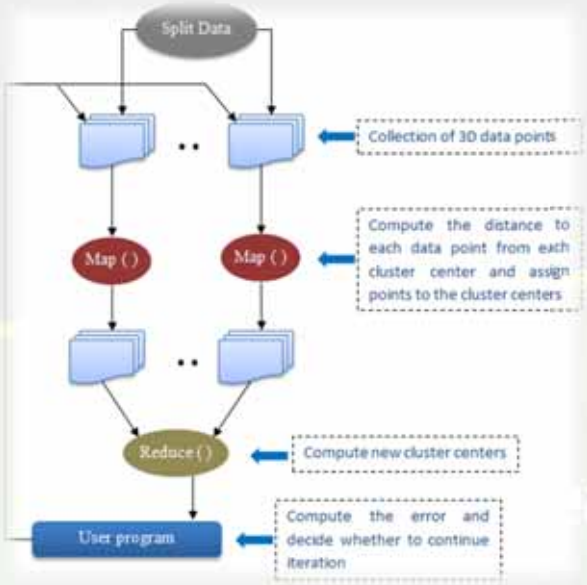
- ✓ *Identify a line with a certain slope describing the data*

Big Data Analytics Interest Group

RESEARCH DATA ALLIANCE

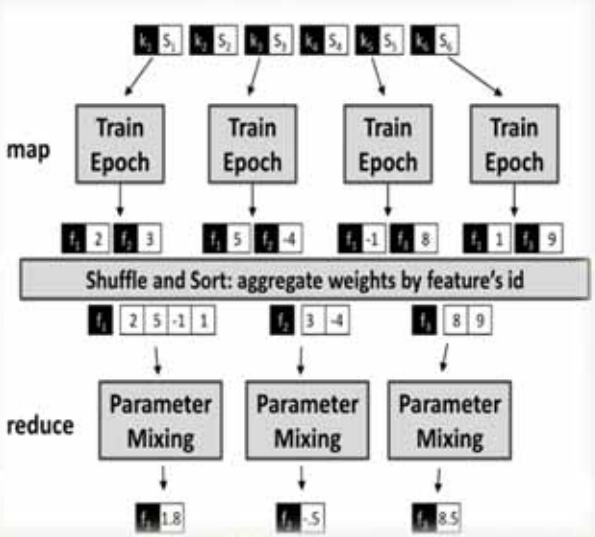
Big Data Analytics Techniques require their Parallelization

[9] Geoffrey Fox et al.



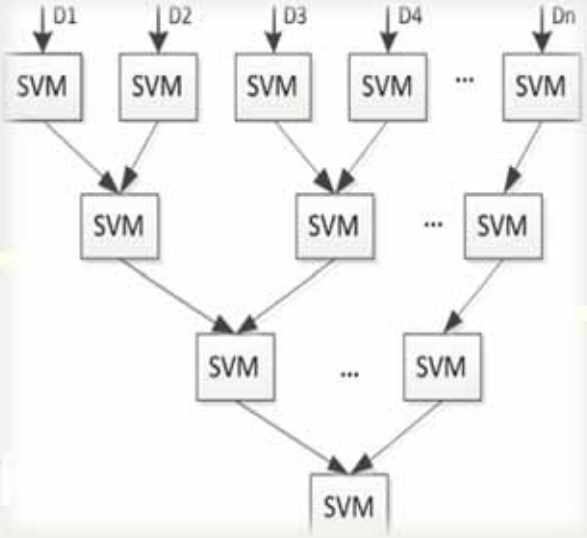
✓ *Parallel K-Means Clustering Algorithms*

[10] Gesmundo et al.

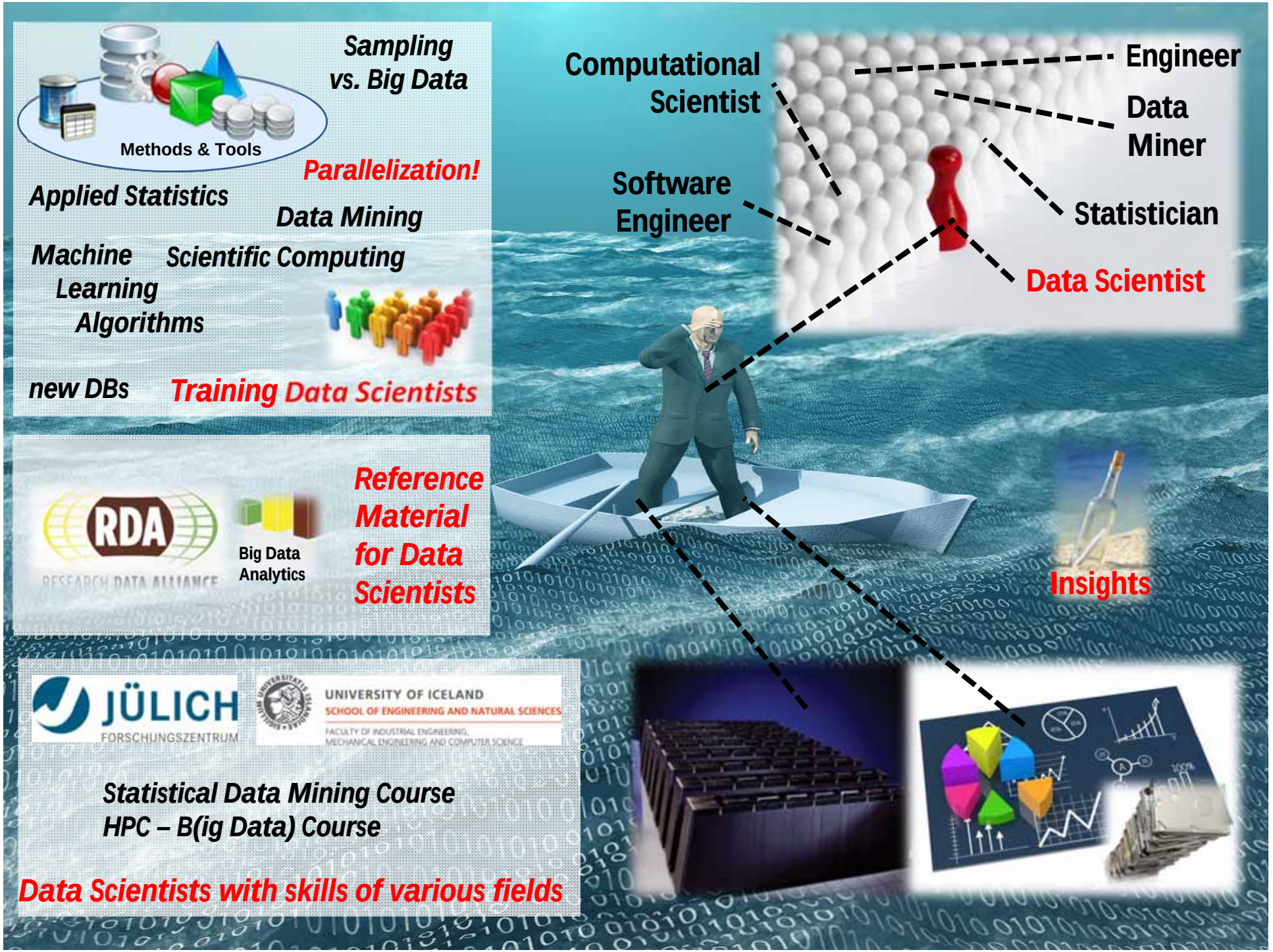


✓ *Parallel Perceptron Using Map-Reduce*

[11] Zhanquan Sun et al.



✓ *Parallel Support Vector Machines*


Methods & Tools
Sampling vs. Big Data
Parallelization!
Applied Statistics
Data Mining
Machine Learning Algorithms
Scientific Computing
new DBs
Training Data Scientists
RDA
RESEARCH DATA ALLIANCE
Big Data Analytics
Reference Material for Data Scientists
JÜLICH
FORSCHUNGSZENTRUM
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
Data Scientists with skills of various fields
Computational Scientist
Engineer
Data Miner
Software Engineer
Statistician
Data Scientist
Insights

Concrete Use Case

‘Underutilized Unique Big Data‘

‘Big Data‘ not always means ‘Big Volume‘, but also Context (uniqueness)

Lufthansa
BRITISH AIRWAYS
CHINA AIRLINES
AIRFRANCE
Austrian
Air Namibia
IBERIA
CATHAY PACIFIC
SOUTH AFRICAN AIRWAYS
QANTAS
Philippine Airlines

In-flight Measurements (→‘Context‘)

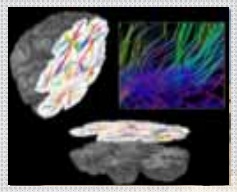
- **MOZAIC 1994 – today (~20 years):**
  - ✓ **Ozone, water vepor, CO, NOy, + p, T, Wind**
- **IAGOS 2011 – today (‘better data‘):**
  - ✓ **+ “Cloud droplets” + CO_2, CH_4, Aerosole**

Challenge:
Large ‘underutilized‘ data collections exist

Approach:
Initial general data analytics techniques
Domain-specific data analysis by experts

RDA RESEARCH DATA ALLIANCE
Big Data Analytics
IAGOS



Automating Quality Control

But 'Bigger Data' not always means 'Better data'

Koljofjord Observatory

Challenge:
Large data collection exist with many measurements

Approach:
Initial general data analytics techniques
→ E.g. automated outlier detection for quality control (Previously only manual sampling, but better?)

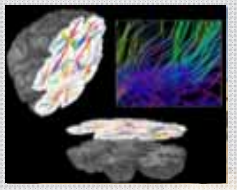
Domain-specific focussed data validation by experts

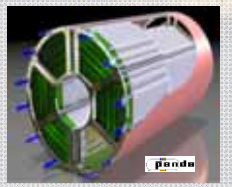
hypox

RDA
RESEARCH DATA ALLIANCE

Big Data Analytics

With thanks to Robert Huber, Marum, Bremen





[8] ScienceTube

Long-term Data Preservation and Curation...

bears potentials to lower 'Data Waves'

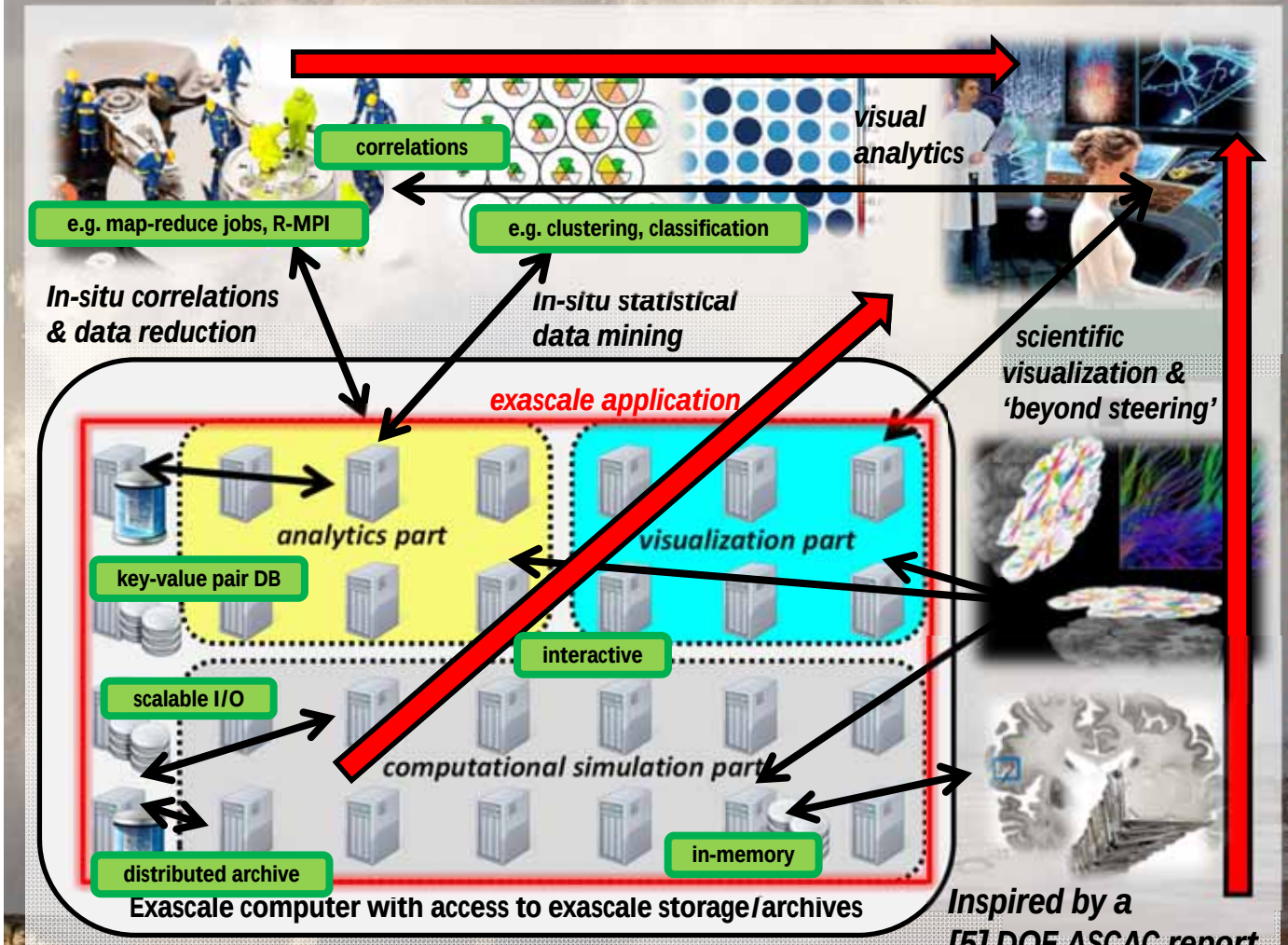
and supports big data analytics process

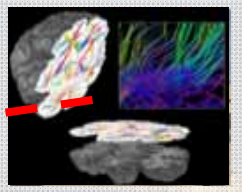
Selected Benefits of open data infrastructures for science & engineering:

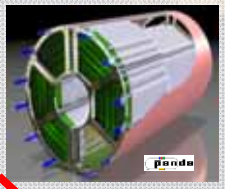
- ✓ ***High reliability,*** *so data scientists can count on its availability*
- ✓ ***Open deposit,*** *allowing user-community centres to store data easily*
- ✓ ***Persistent identification,*** *allowing data centres to register a huge amount of markers to track the origins and characteristics of the information*
- ✓ ***Metadata support*** *to allow effective management, use and understanding*
- ✓ ***Avoids re-creation of datasets*** *through easy data lookups and re-use*
- ✓ ***Enables easier identification of duplicates*** *to remove them & save storage*

Towards Exascale: Applications with combined characteristics of simulations & analytics

‘In-Situ Analytics‘







References

[1] High Level Expert Group on Scientific Data, 'Riding the Wave',Report to the European Commission, October 2010, Online:
http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf
[2] Knowledge Exchange Partner, 'A Surfboard for Riding the Wave', November 2012, Online:
http://www.knowledge-exchange.info/surfboard
[3] Fran Berman, 'Maximising the Potential of Research Data', January 2013, Online:
http://rdi2.rutgers.edu/event/rdi2-distinguished-seminar-maximizing-innovation-potential-research-data
[4] Jeremy Ginsburg et al., 'Detecting influenza epidemics using search engine query data', Nature 457 (2009), Online:
http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html
[5] DOE ASCAC Data Subcommittee Report, 'Synergistic Challenges in Data-Intensive Science and Exascale Computing', 2013, Online:
http://www.sci.utah.edu/publications/chen13/ASCAC_Data_Intensive_Computing_report_final.pdf
[6] Research Data Alliance (RDA), Big Data Analytics Web Page, Online:
https://rd-alliance.org/internal-groups/big-data-analytics-ig.html
[7] Brandyn White et al., 'Web-Scale Computer Vision using MapReduce for Multimedia Data Mining', Online:
http://dl.acm.org/citation.cfm?doid=1814245.1814254
[8] M. Riedel and P. Wittenburg et al. 'A Data Infrastructure Reference Model with Applications: Towards Realization of a ScienceTube Vision with a Data Replication Service', 2013, Online:
http://www.jisajournal.com/content/4/1/1
[9] G. Fox, 'MPI and Map-Reduce', Talk at CCGSC 2010 Flat Rock, NC, 2010, Online:
http://web.eecs.utk.edu/~dongarra/ccgsc2010/slides/talk12-fox.pdf
[10] Gesmundo et al., Hadoop Perceptron, 2012, Online:
http://cui.unige.ch/~gesmundo/papers/gesmundo-eacl2012.pdf
[11] Zhanquan Sun et al., 'Study on Parallel SVM Based on MapReduce', Online:
http://grids.ucs.indiana.edu/ptliupages/publications/Study%20on%20Parallel%20SVM%20Based%20on%20MapReduce.pdf


‘Thanks’
Talk available at:
www.morrisriedel.de/talks
Contact:
m.riedel@fz-juelich.de