Agenda Today

- Usual Round of analytics interests from people…

- Welcome & Introduction (Morris) - 10
- Seismic Analytics (Giuseppe) – 15
- Insight – Centre for Data Analytics (Stefan) - 15
- Spin-off Working Group (Wo) – 15
- Array Analytics (Peter) – 15
- Open Floor: Your interest in Analytics short presentation? – 5/each
- Normal work
    - (1) Tackling Event Analytics – Deconstruction on last telcon
    - (2) Tackling Outlier Use Cases
   - (3) Next steps, next call

Analytics are Needed in Big Data-driven Scientific Research
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] 'Riding the Wave' 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.'
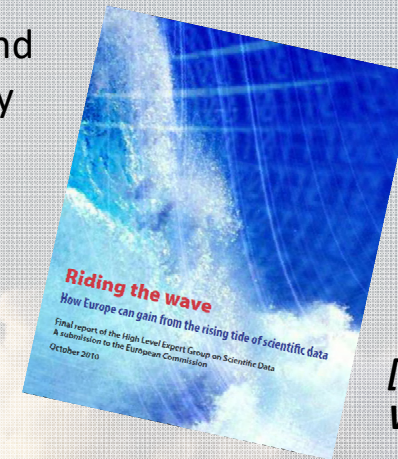
[2] 'A Surfboard for Riding the Wave' Report

How do we enable ‚high productivity processing'?
How do we find ‚a message in the bottle'?

Discussion around Terminologies → You/Wiki

Big Data Analytics
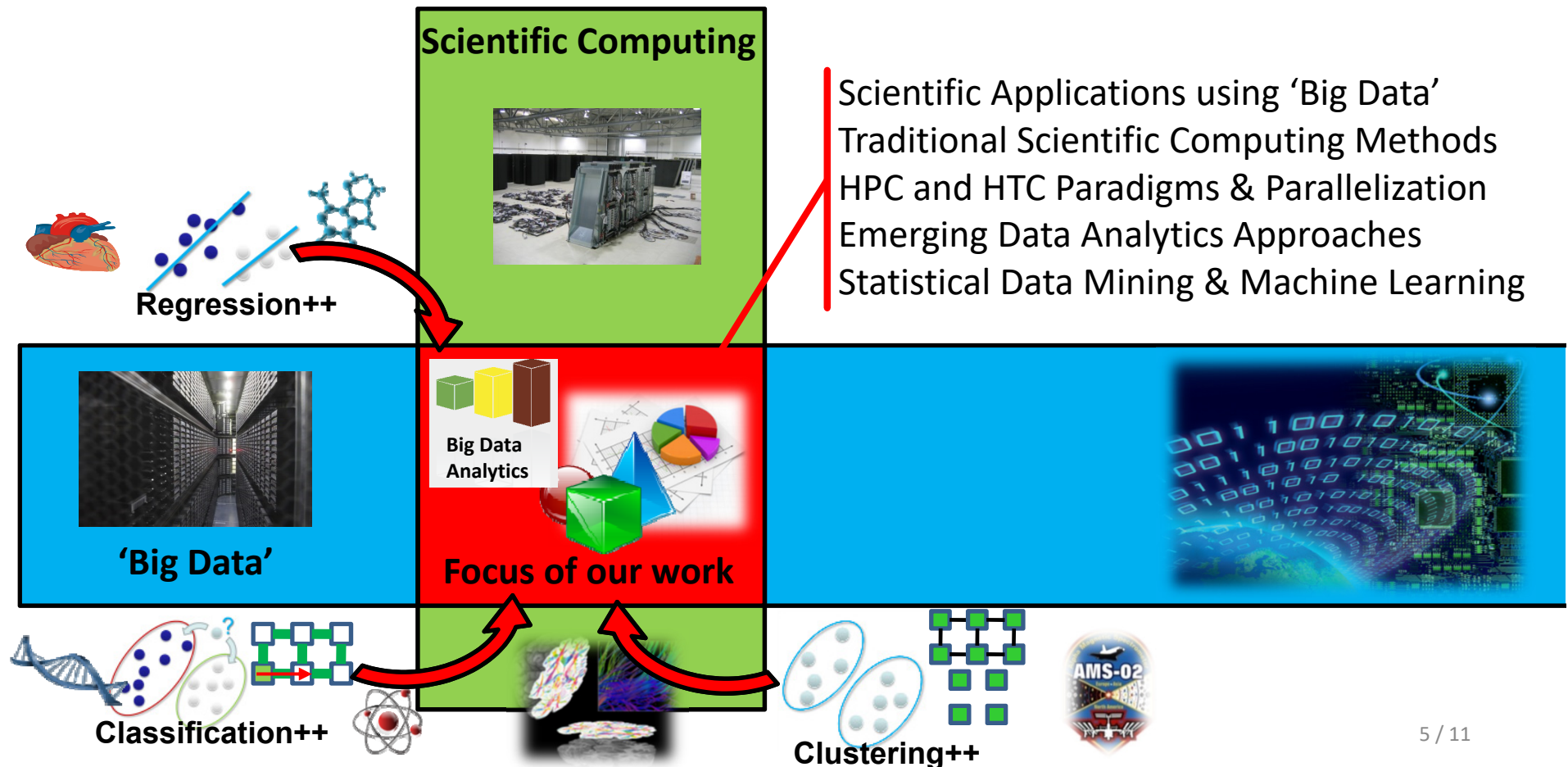
There are different views on the different tems…

…so lets be concrete and show evidence and running code

- 'Data Analysis' supports the search for 'causality'

    - Describing exactly WHY something is happening → science
    - Understanding causality is hard and time-consuming, but is necessary **?**
    - Searching it often leads us down the wrong paths…

- 'Big Data Analytics' is focussed on 'correlation'

    - Not focussed on causality – enough THAT it is happening → money/events
    - Discover novel patterns and WHAT is happening more quickly **?**
    - Using correlations for invaluable insights – often data speaks for itself

- **Analysis is the in-depth interpretation of 'big data'**
- **Analytics are powerful techniques to work on 'big data'** **?**
- **Parameter/event space exploration may use (1) analytics, then (2) analysis**
- **Pre-/Post-Process data with (1) analytics for deeper/faster (2) data analysis processing**

Work on Intersection of two Broad Subjects

- 'Lighthouse goal': High Productivity Processing of Research Data

Scientific Computing

Scientific Applications using 'Big Data'
Traditional Scientific Computing Methods
HPC and HTC Paradigms & Parallelization
Emerging Data Analytics Approaches
Statistical Data Mining & Machine Learning

Regression++

'Big Data'

Big Data Analytics

Focus of our work

Classification++

Clustering++

Understanding Big Data Analytics Better

- **Develops community based recommendations**
    - … on feasible data analytics approaches
    - … to address scientific community needs/problems
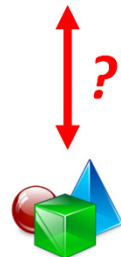   - … of utilizing large quantities of data.
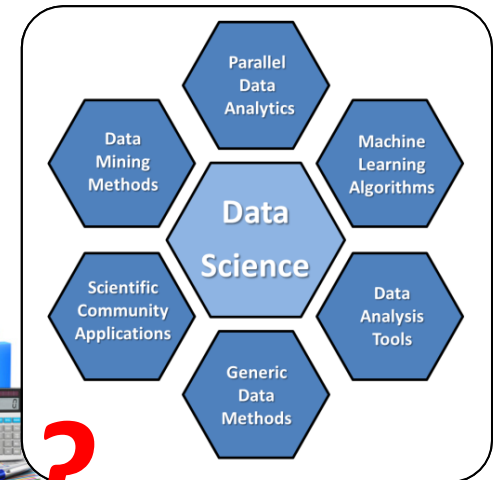
- **Work with different scientific domain applications**
    - … and their use of concrete big data analytics techniques
    - … what really works and runs to solve the solution?

- **IG: Bottom-up, dispersed and only slightly coordinated**
    - Sharing knowledge of analysis algorithms, analytical tools, …
    - … data and resource characteristics …
    - … and running code that works will be part of the recommendations.

News: PRACE/XSEDE Earth Science Analytics

PANGAEA

- **Problem: Quality control via outlier detection with PANGAEA data**
- **Key PI: Dr. Robert Huber, MARUM, Bremen, Germany**

IAGOS

- Problem: longitude/latitude/altitude correlations with IAGOS data
- Key PI: Dr. Owen Cooper, NOAA ESRL, US

EVENTS

- **Problem: Event tracking analytics with spatial computing datasets**
- **Key PI: Dr. Rahul Ramachandran, NASA MSFC, US**

SEISMIC

- Problem: Continuous seismic waveforms analysis for earthquakes monitoring
- Key PI: Alberto Michelini, INGV, Italy

SCALE GIS

- Problem: Projecting & transforming geospatial big data into a common coordinate reference framework
- Key PI: Shaowen Wang, NCSA, US

Proposal for the 'XSEDE-PRACE call for requests of joint support'
Smart Data Analytics for Earth Sciences across XSEDE and PRACE

Executive Summary
The ever-increasing amount of scientific data arising from measurements or computational simulations requires new 'smart data analytics techniques' capable of extracting meaningful findings from 'pure big data'. XSEDE (including FutureGrid for Map-Reduce), as well as PRACE, provides excellent resources that enable efficient and effective data analytics when several technical frameworks and data analysis packages would be available. Making
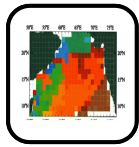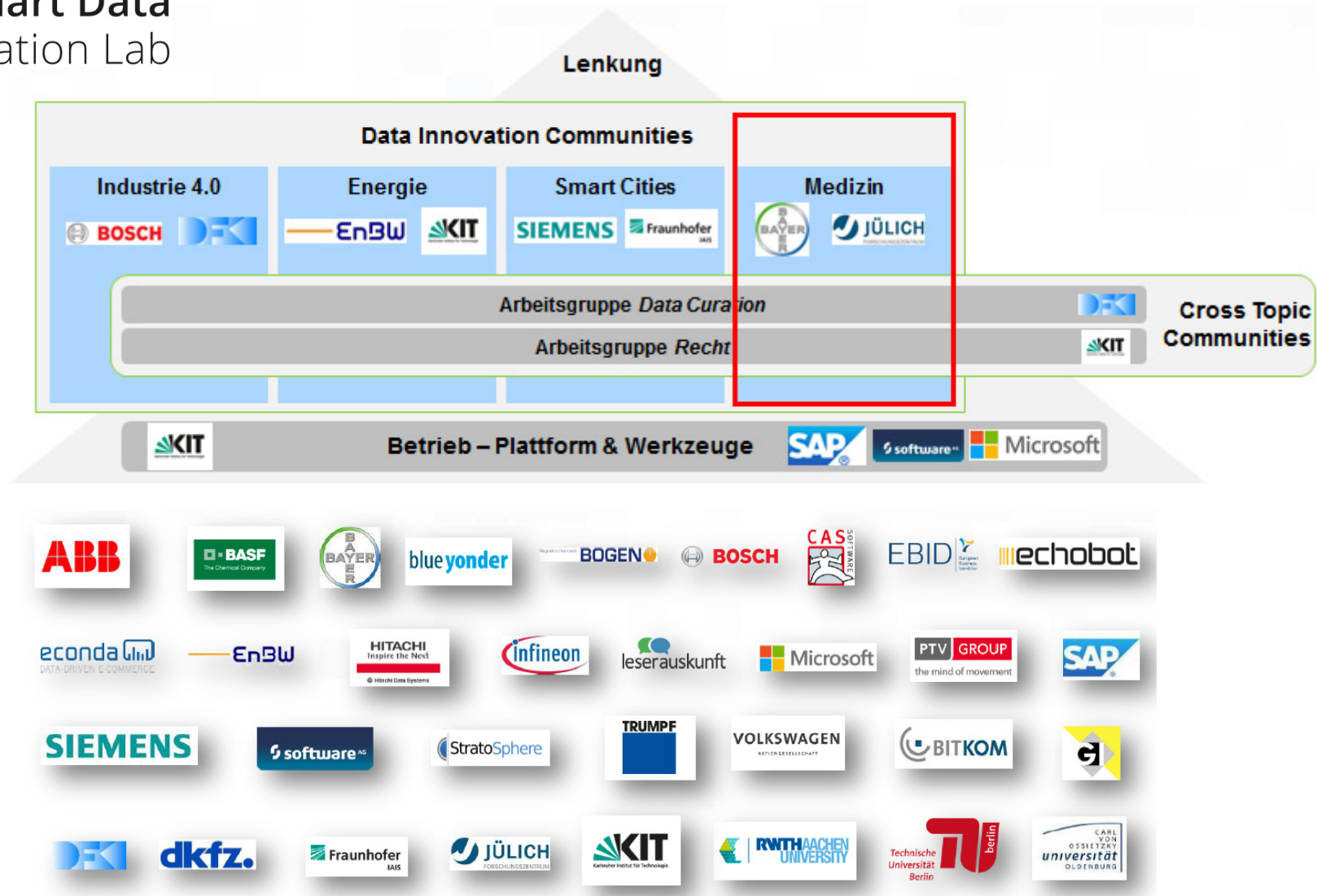
News: Smart Data Analytics @ Germany

Other Relevant Short News for the Group

- Big Data Infrastructure WG Spinn-off (→ Talk later by Wo)
  - Provides references model (?) that may include some of our use cases
  - We need to understand how we can work together tackling a specific problem

- Switch Co-Chairs at NASA: Dr.Kuo Kwo-Sen (Rahul Ramachandran)
  - NASA Goddard Space Flight Center in Greenbelt, Maryland
  - Expert in atmospheric radiative transfer and remote sensing
  - Has extensive experience working with wide varieties & large volumes of data
  - Mostly Earth science data from various specific sub-domains
  - Also works on event analytics use case

- Interest stated from CODATA/RDA Material Science IG
  - Several interesting presentations: DataNetFederation, Dspace (Material Sciences)
  - Morris gave short intro about the Big Data Analytics IG in the session I
 - Concrete follow-up contacts: Mary Brady (Analytics @ NIST)

Summary & Initial Use Cases
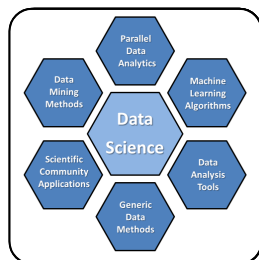
Big Data Analytics

Contributions

- Tackles bottom-up use cases that require 'big data analytics'
- Provides a systematic classification of technology combinations
- Develops recommendations on feasible analytics approaches
- Offers best practice guides for researchers & concrete problems

Selected use cases with concrete problems
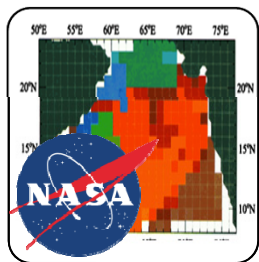
Event Analytics

- Problem: event tracking analytics (e.g. understanding somali jets)
- Data sets from satellites ('events with changing geolocations')
- Technologies: HPC/HTC (map-reduce), data-bases, several algorithms
- Status: review existing event tracking literature & algorithms

Outlier Detection

- Problem: automatic outlier detection in 'big data' (PANGAEA)
- Data sets from time-series measurements (e.g. 'Koljoefjords, Sweden')
- Technologies: HPC/HTC (map-reduce), R (outliers, RMPI)
- Status: CRISP-DM, investigating running code for outlier algorithms

References

[1] J. Wood et al., 'Riding the Wave – How Europe can gain from the rising tide of scientific data', report to the European Commission, 2010

[2] Knowledge Exchange Partner, 'A Surfboard for Riding the Wave – Towards a Four Action Country Programme on Research Data', 2011

[3] Research Data Alliance (RDA) Web Page, Online: https://rd-alliance.org/node

[4] G. Fox, 'MPI and Map-Reduce', Talk at CCGSC 2010 Flat Rock,NC, 2010