Understanding Big Data Analytics Applications in Earth Science

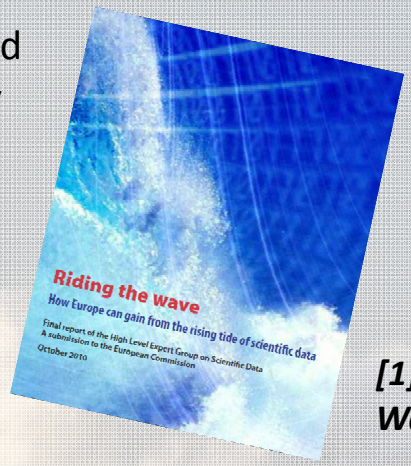
Morris Riedel, Rahul Ramachandran/Kuo Kwo-Sen, Peter Baumann
Big Data Analytics Interest Group Co – Chairs

research data sharing without barriers
rd-alliance.org

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'?

Understanding concepts & terminologies

There are different views on the different terms…

…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

?

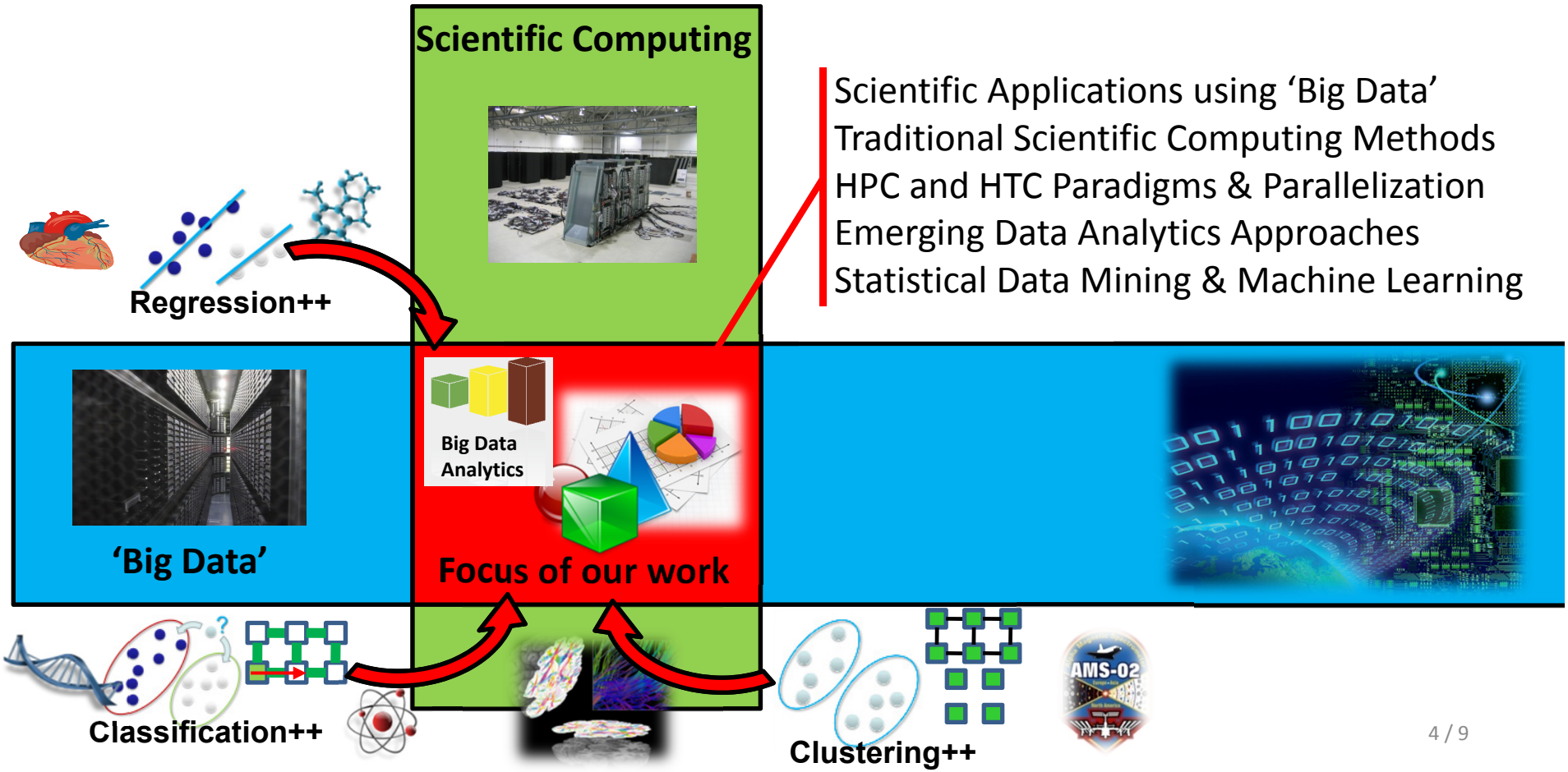
?

- **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**

?

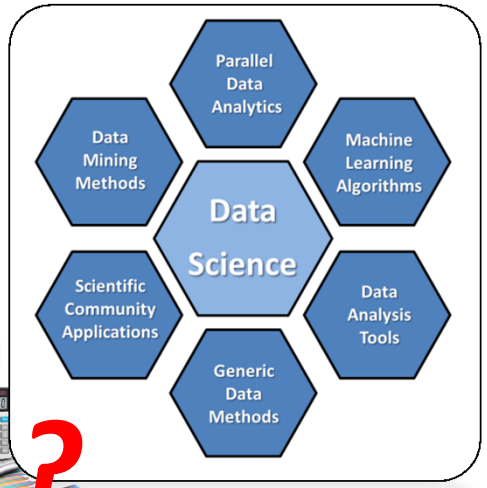
Understanding Applications & Technology

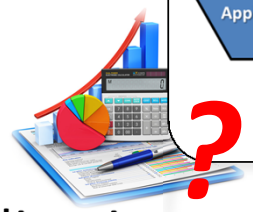
- ‘Lighthouse goal’: High Productivity Processing of Research Data



RDA Group: Understand Concrete Solutions

- 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.

PRACE/XSEDE Earth Science Analytics

RDA RESEARCH DATA ALLIANCE

Big Data Analytics

Proposal for the 'XSEDE-PRACE call for requests of joint support'

Smart Data Analytics for Earth Sciences across XSEDE and PRACE

Executive Summary

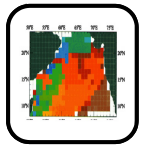
The ever-increasing amount of scientific data arising from measurements or computational simulations requires new *'smart data analytics techniques'* capable of extracting meaningful findings from *'pure big data'*. XSEDE (including FutureGrid for Map-Reduce), as well as PRACE, provides excellent resources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

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**

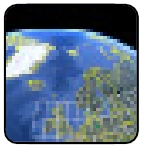


EVENTS

- **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**

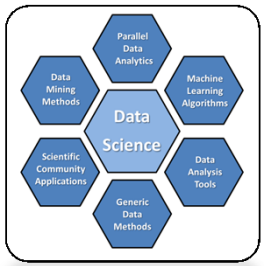


SCALE GIS

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

Take advantage of e-Infrastructure for automation and sharing of methods/data

Increase Understanding with Applications

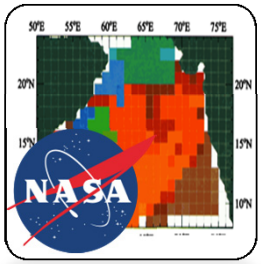


Contributions

- 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

Concrete Application Implementations

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

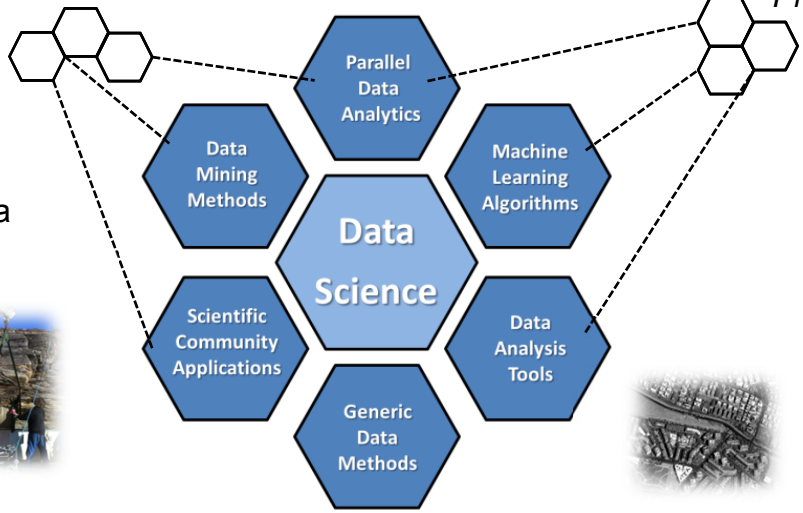
- Problem: automatic outlier detection in ‚big data' (PANGAEA)
- Data sets from time-series measurements (e.g. 'Koljoefjords, Sweden')
- Technologies: HPC/HTC (map-reduce), R (outliers, RMPI)
- Status: CRISP-DM, investigating running code for outlier algorithms

Findings: Parallelization & Big Data is Hard

Problem: Automatic outlier detection for data quality

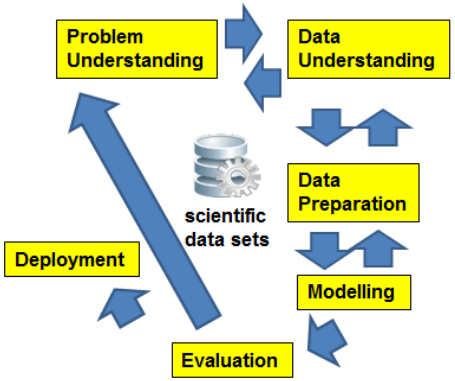
- ✓ Tailor solution for community
- ✓ Scalability towards Big Data
- ✓ Design and improve data analytics approaches



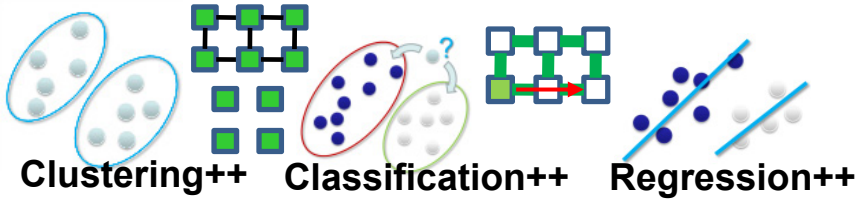


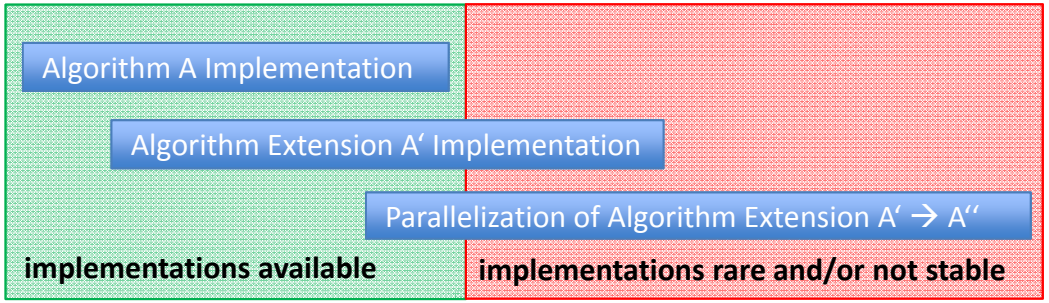
Problem: Classification of buildings from multi-spectral images

- ✓ Enable smooth transition from 'manual Matlab SVM scripts'
- ✓ Research on parallel SVM methods (map-reduce, HPC)



Systematic Analytics guided by CRISP – DM





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