Enabling Parallel and Scalable Tools for Scientific Big Data Analytics

Dr.-Ing. Morris Riedel et al.

Research Group Leader, Juelich Supercomputing Centre

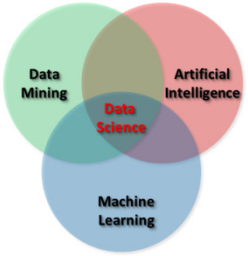
Adjunct Associated Professor, University of Iceland

HELMHOLTZ ASSOCIATION

Federated Systems and Data Division

Research Group

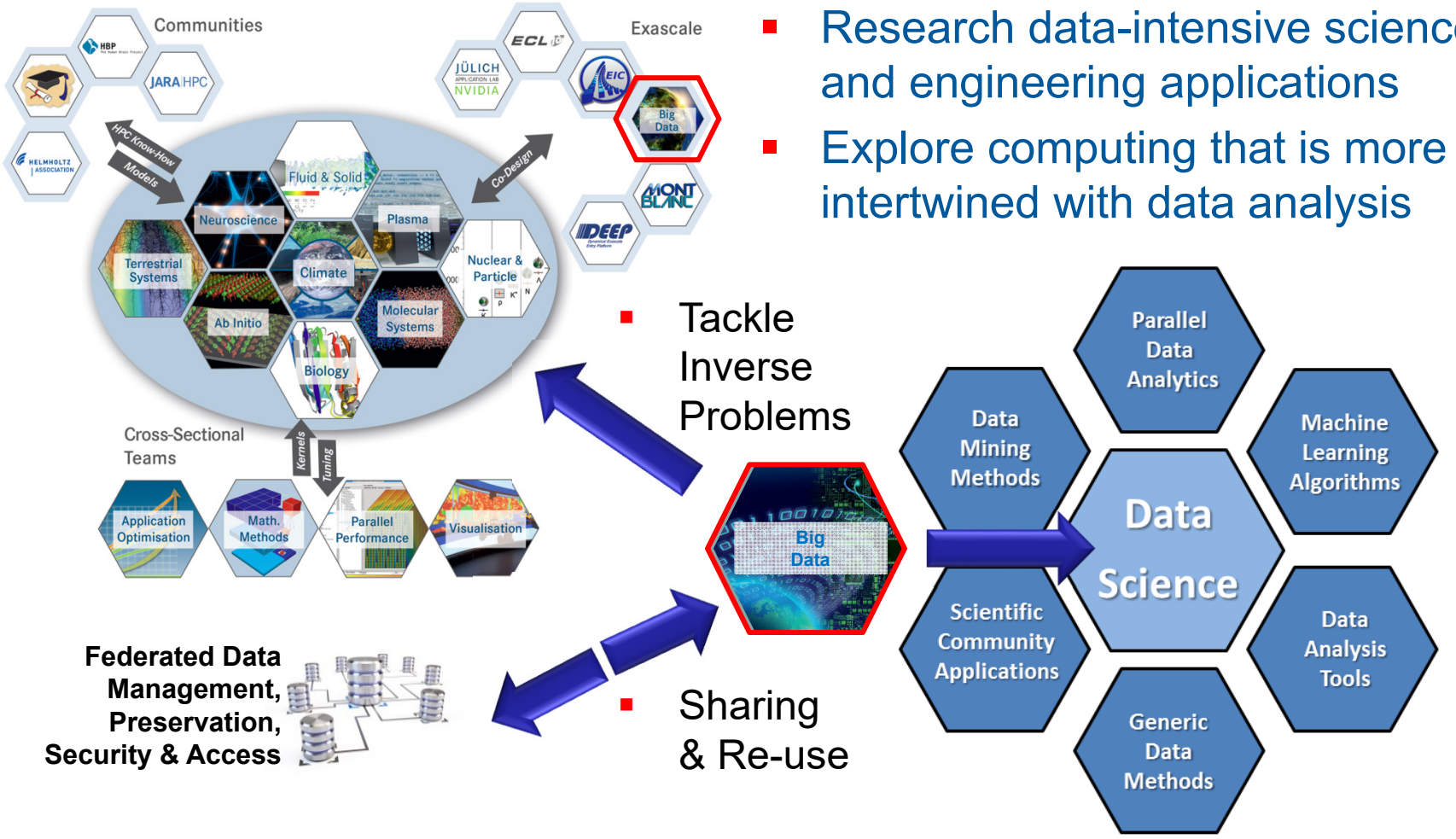
High Productivity Data Processing



JÜLICH FORSCHUNGSZENTRUM

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

Scientific Big Data Analytics – Context JSC



- Research data-intensive science and engineering applications
- Explore computing that is more intertwined with data analysis

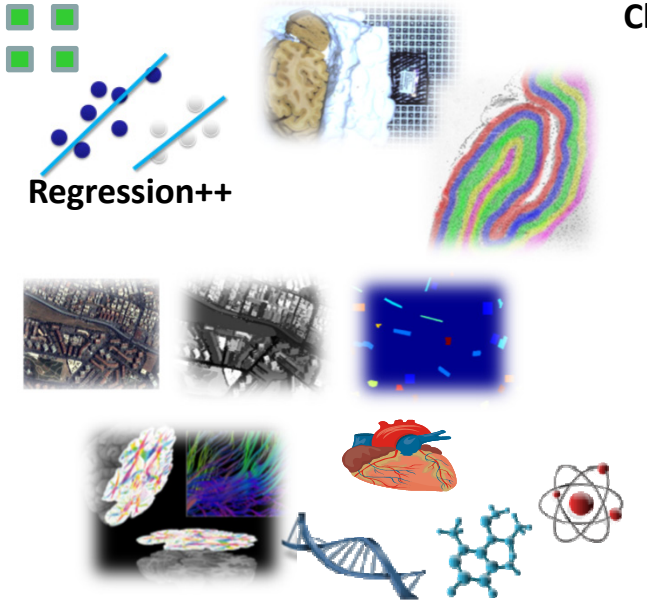
Th. Lippert, D. Mallmann, M. Riedel, *Scientific Big Data Analytics by HPC,* NIC Series 48, 417, ISBN 978-3-95806-109-5, pp. 1 - 10, 2016

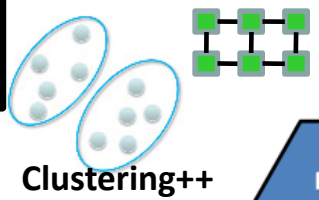
Data Analytics – Term Clarification

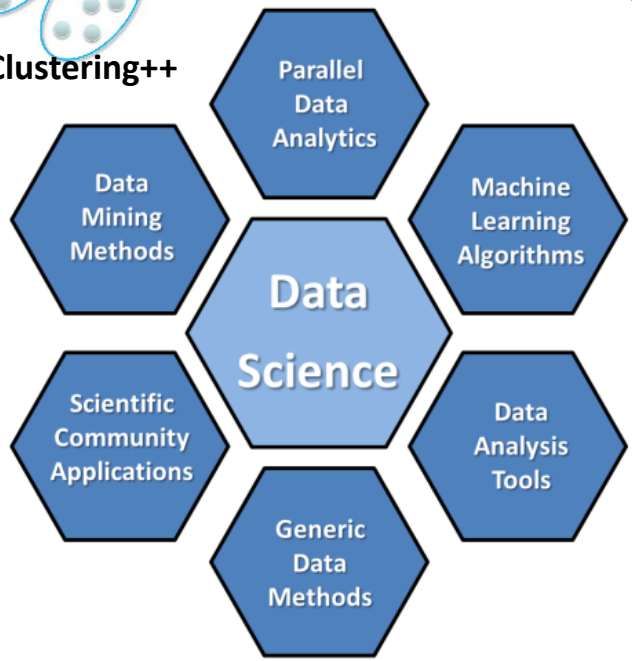
‘Data Analytics’ is an ‘interesting mix’ of different approaches

- Analytics: Whole methodology; Analysis: data investigation process itself
- ‘Big‘ requires scalable processing methods and underlying infrastructure

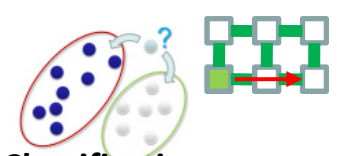
- **Concrete ‘big data‘: large medical data**

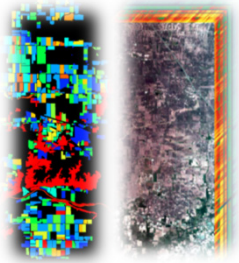




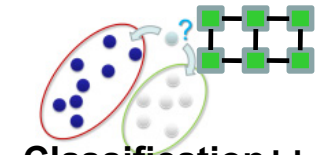


- **Concrete ‘big data‘: large earth science data**



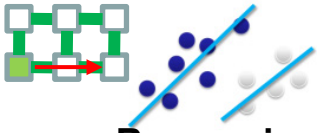


Data Analytics – Research Key Focus

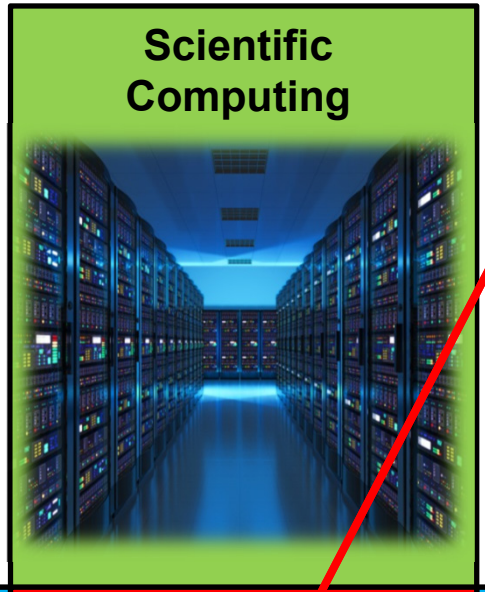


Classification++

Clustering++



Regression++



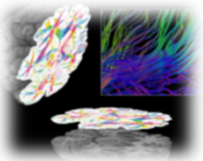
Scientific Computing

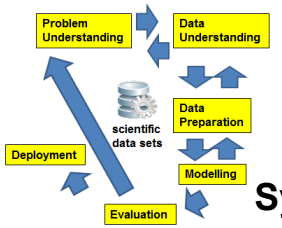
Scientific Applications using ‘Big Data’
Traditional Scientific Computing Methods
HPC and HTC Paradigms & Parallelization
Emerging Data Analytics Approaches
Optimized Data Access & Management
Statistical Data Mining & Machine Learning
Inverse Problems & their Research Questions

‘Big Data’ Methods

Scientific Applications

‘(Big) Data from various data science applications‘





Systematic & Automated Analytics guided by CRISP – DM

Data Analytics – Selected Research Expertise

Key expertise making algorithms parallel & scalable for 'big data'

- Driven by scientific and engineering cases, e.g. understanding the human brain, remote sensing applications, marine measurements analysis, …
- Example: Clustering, e.g. Density-based Spatial Clustering of Applications with Noise (DBSCAN)
- Example: Classification, e.g. Support Vector Machines (SVMs)

Parallel & Scalable DBSCAN clustering tool

Problem: Automatic outlier detection for data quality

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

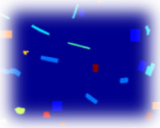


Parallel & Scalable SVM classification tool

Problem: Classification of buildings from multi-spectral images

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



M. Goetz, C. Bodenstein, M. Riedel, *HPDBSCAN – Highly Parallel DBSCAN,* Proceedings of the ACM/IEEE International Conference for High Performance Computing, Networking, Storage, and Analysis (SC2015), Austin, TX, USA, 2015

G. Cavallaro, M. Riedel, et al., *On Understanding Big Data Impacts in Remotely Sensed Image Classification Using Support Vector Machine Methods,* IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Issue 99, pp. 1-13, 2015

Acknowledgements

PhD Student Gabriele Cavallaro, University of Iceland

Tómas Philipp Runarsson, Kristján Jonasson,
Jón Atli Benediktsson, University of Iceland

Timo Dickscheid, Markus Axer, Stefan Köhnen, Tim Hütz,
Institute of Neuroscience & Medicine (INM), Forschungszentrum Juelich

Selected Members of the Research Group on High Productivity Data Processing

Ahmed Shiraz Memon
Mohammad Shahbaz Memon
Markus Goetz
Christian Bodenstein
Matthias Richerzhagen
(Philipp Glock → INM)

Thanks

Talk available online:

www.morrisriedel.de/talks