Big Data in Science

Overview of European & International Activities

HELMHOLTZ ASSOCIATION

Research Field Key Technologies

Jülich Supercomputing Centre

Supercomputing & Big Data

Dr - Ing. Morris Riedel et al.

Adjunct Associated Professor, University of Iceland, Iceland

Juelich Supercomputing Centre, Germany

Head of 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

prospect hpc HIGH PERFORMANCE COMPUTING

10th October 2014, Juelich

'Big Data' in Science & Engineering

Smart Data Innovation Lab (SDIL)

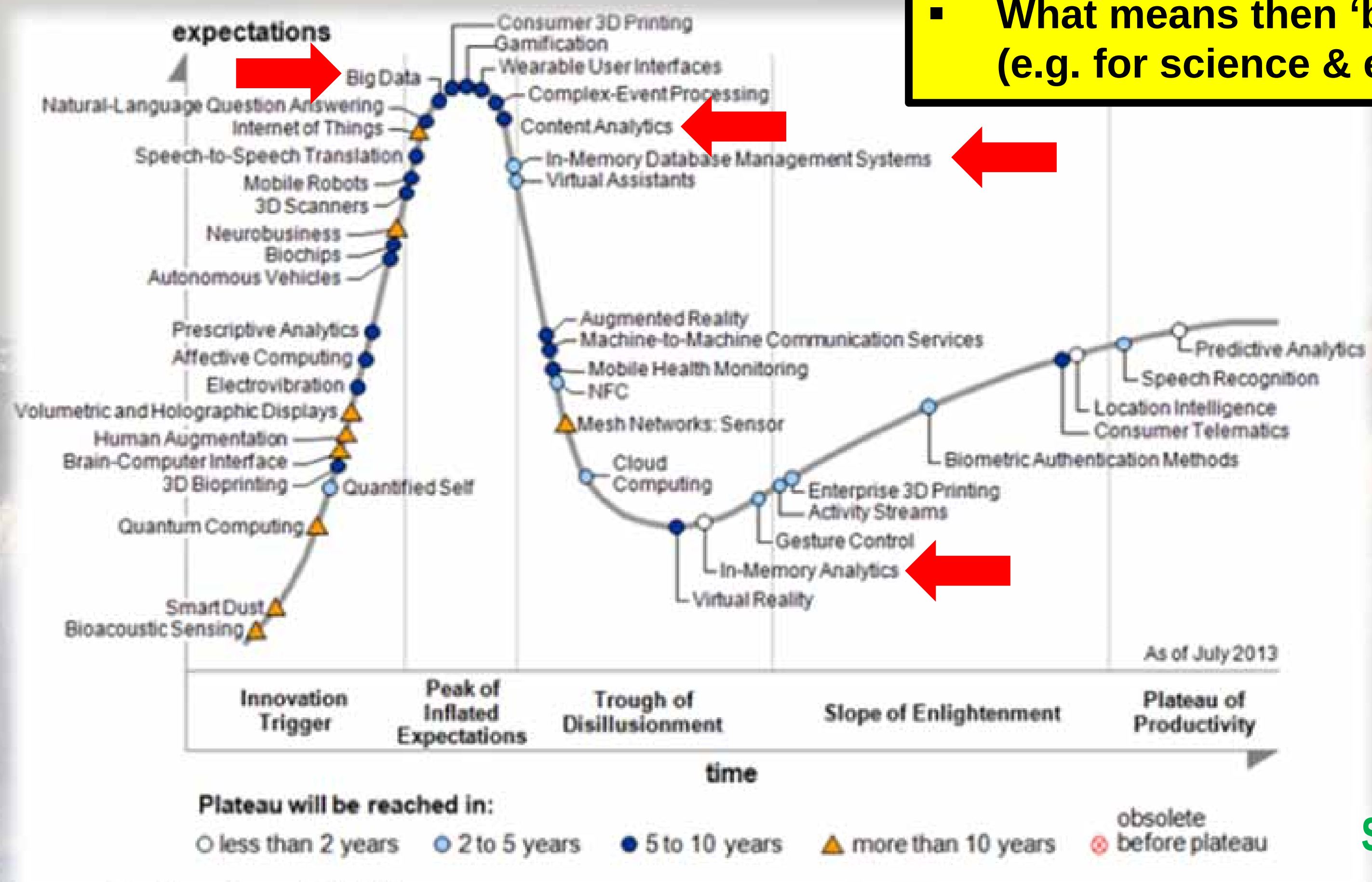
European Data Infrastructure (EUDAT)

Research Data Alliance (RDA)

Lessons Learned & Need of 'Steering'

What can we expect from 'Big Data'

… towards 2014 & reaching the peak – do we see more clearly?



Source: Gartner August 2013

- **What means then 'big data analytics'? (e.g. for science & engineering)**

…

Recommender systems

User-centric marketing

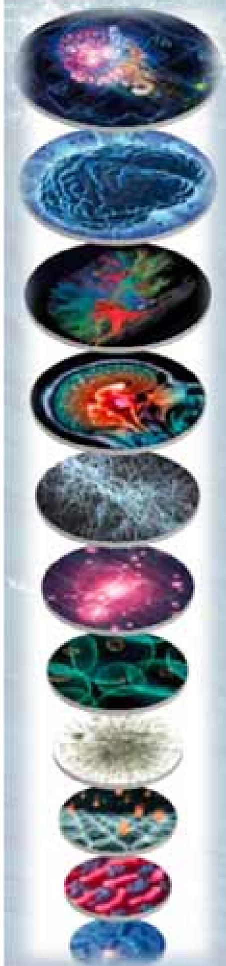
Predictive Maintenance

Customer segmentation

…

Science & Engineering?

‘Big Data Waves‘



Context

Variety

Volume

Velocity

Veracity

Value

Big Data Streams with 'high velocity' ...

← Data Sources →

'Web Actions'

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

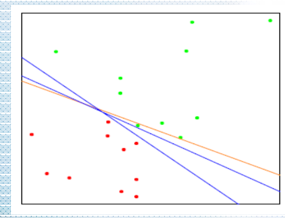
'Measurement Devices'

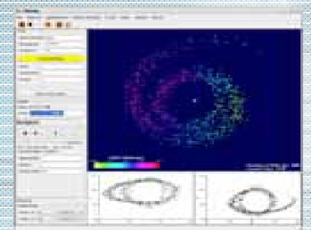
Data stream with measured **data**

'Computational Simulations'

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

n optional filters

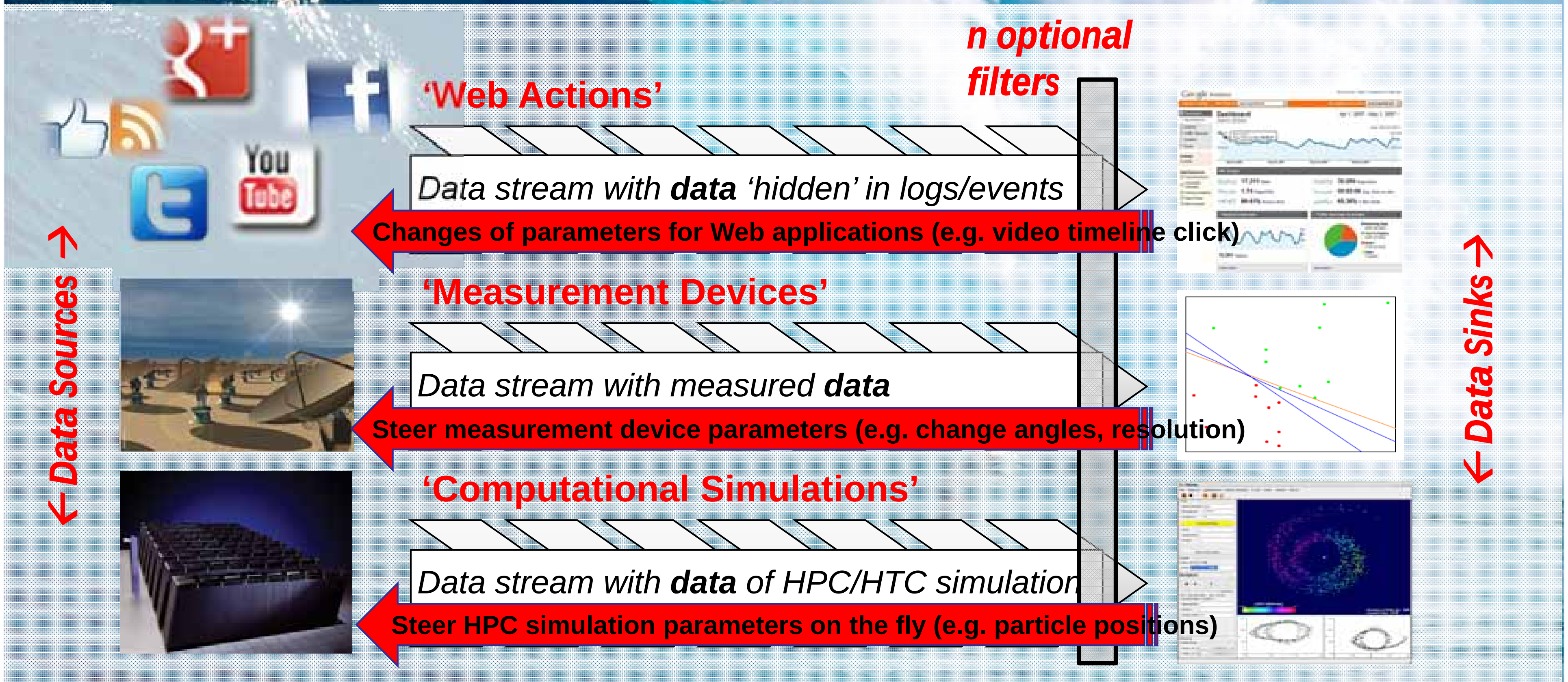




← Data Sinks →

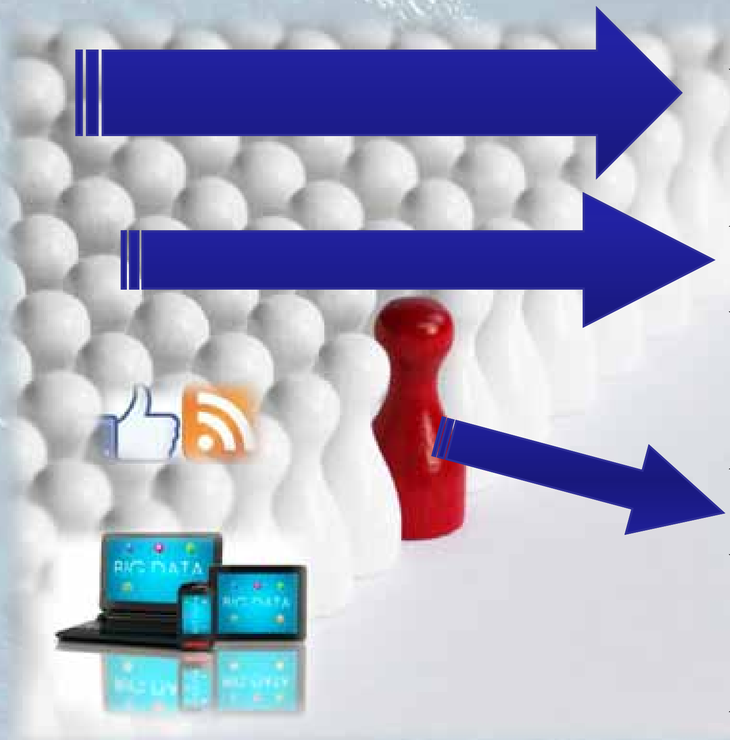
Big Data Streams with 'high velocity'…

… require interactive access & steering



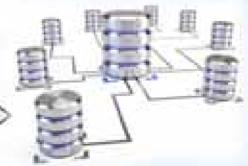
'Crowdsourcing'...

...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

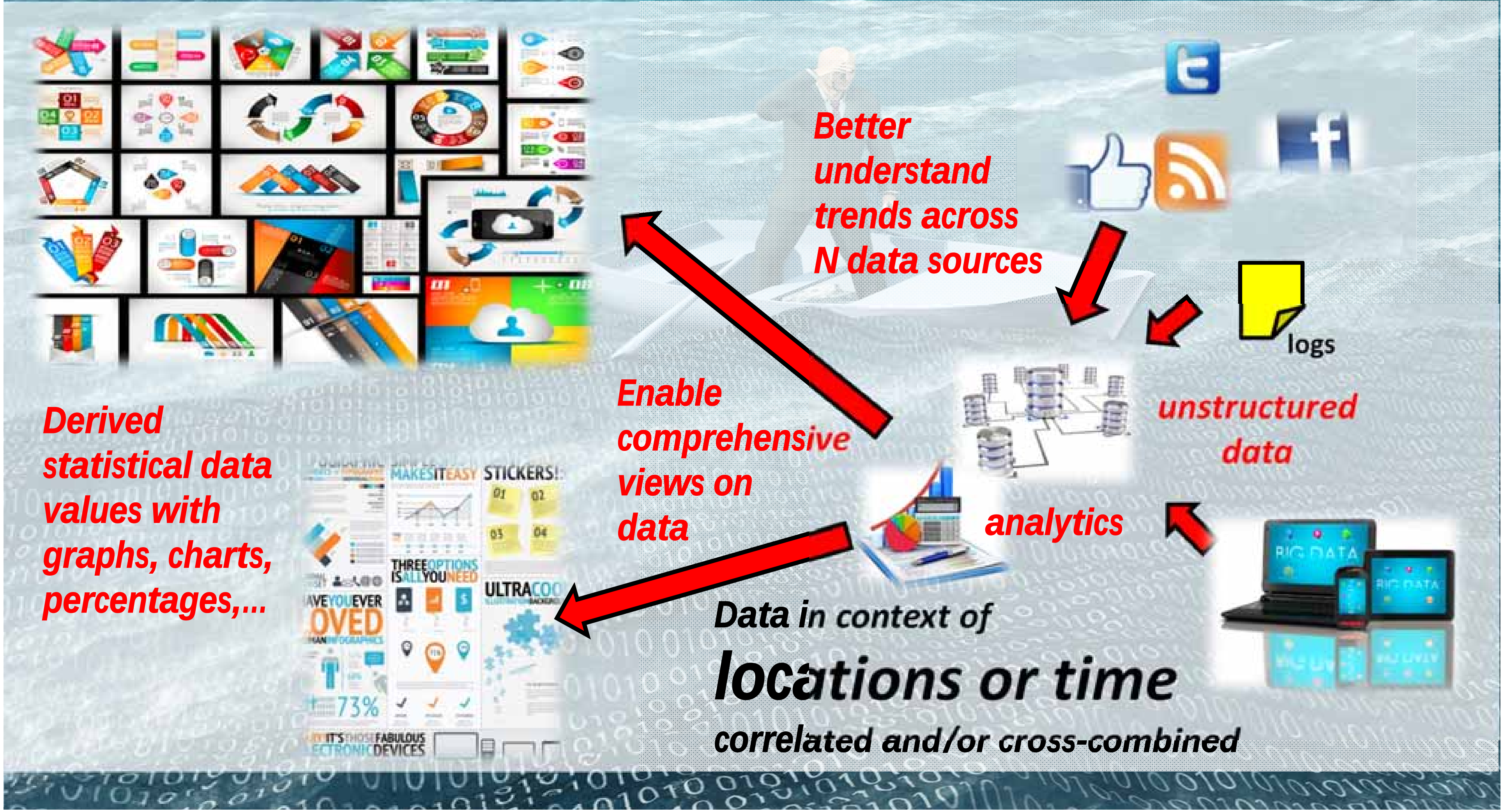
Scientific/Engineering Domain Experts

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

Terabytes

Infographics

Compact Combination of many Data Visualizations



Most data in the world...

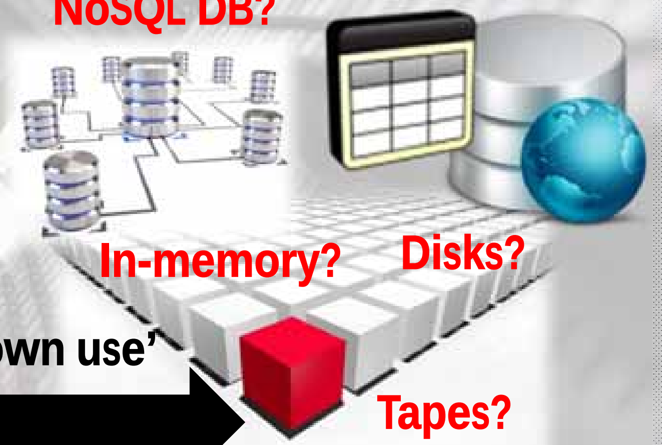
- ...
- *Online Social Media* *(videos, blogs, tweets,...)*
- *Large number of log files* *(Web server log, call center log,...)*
- *Communication data* *(E-Mails, chats, notes, letters, ...)*
- *Various document formats* *(spreadsheet, presentation, docs)*
- ....

... is 'unstructured'

Text Analytics

Data Mining

Keep for 'future unknown use'



New Forms of Data Structures with NoSQL

Optimized for 'write/once' & 'read/many' or 'In-Memory'

Selected Features

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

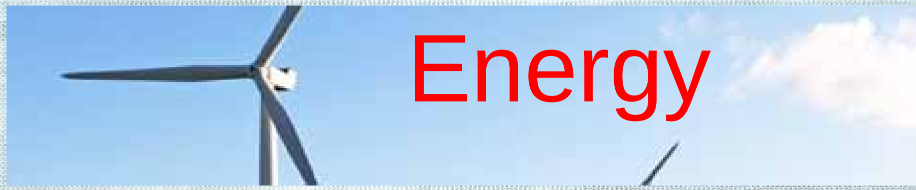
Types

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' used today

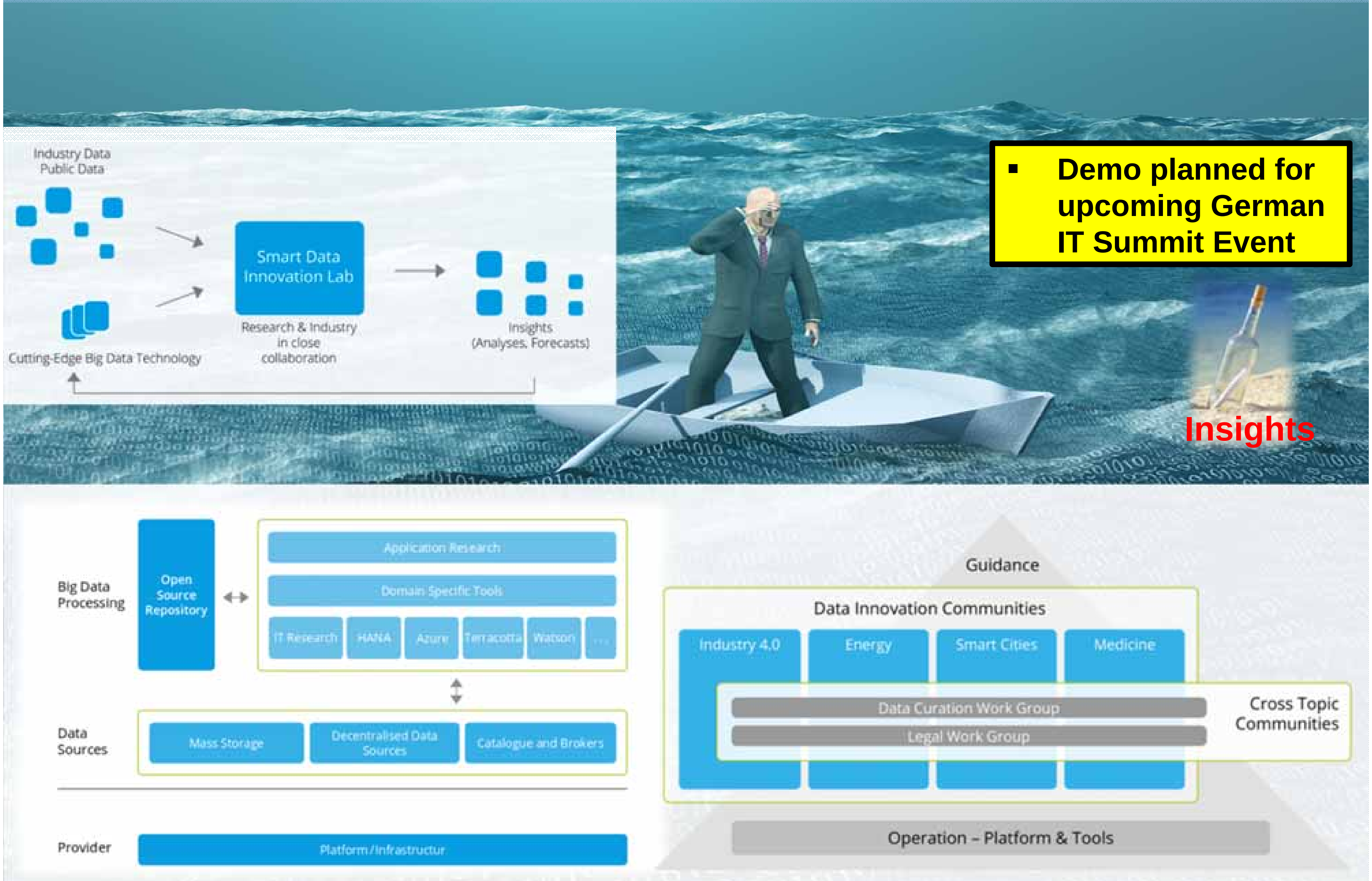
Smart Data Innovation Lab (SDIL)

Companies & Academia jointly work in Four Key Areas



Big Data Waves & Massive Amounts of Technologies Exist

How to create real value from the rising tide of 'Big Data'?



SDIL Industry 4.0

Explore Data-driven Parts of 4^{th} Industrial Revolution

SDIL Data Innovation Community

- Headed jointly by DFKI & Bosch
- Research on proactive service and maintenance of production resources
- Research on finding anomalies in production processes

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Robert Bosch GmbH

SDIL Energy

Explore Data-driven Insights in Using Energy Smarter

SDIL Data Innovation Community

- Headed jointly by KIT & EnBW
- Research on demand-driven fine-tuning of consumption rate models
- Research based on smart metre generated data sources

Karlsruher Institut für Technologie (KIT)

EnBW Energie Baden-Württemberg AG

SDIL Smart Cities

Explore Data-driven Options to Make Urban Life Easier

SDIL Data Innovation Community

- Headed jointly by Fraunhofer IAIS & Siemens
- Explores important data-driven aspects of urban life & quality
- Research on traffic control, waste disposal, or disaster control

Fraunhofer Institut Intelligente Analyse- und Informationssysteme IAIS

Siemens Aktiengesellschaft

SDIL Medicine

Explore Data-driven Aspects of Personalised Medicine

SDIL Data Innovation Community

- Headed jointly by Forschungszentrum Juelich & Bayer
- Research of need-driven care of patients and Web-based patient care
- Research on IT controlled medical technology enabled by ‘big data‘

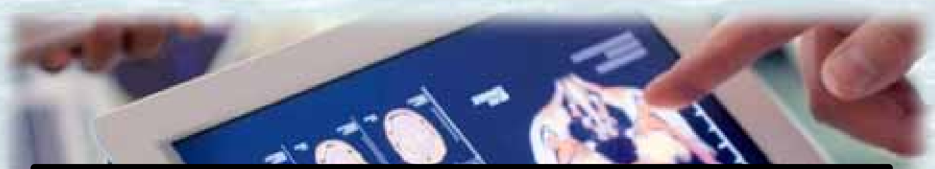
Forschungszentrum Jülich GmbH

Bayer Technology Services GmbH

SDIL Medicine – Identified Key Areas

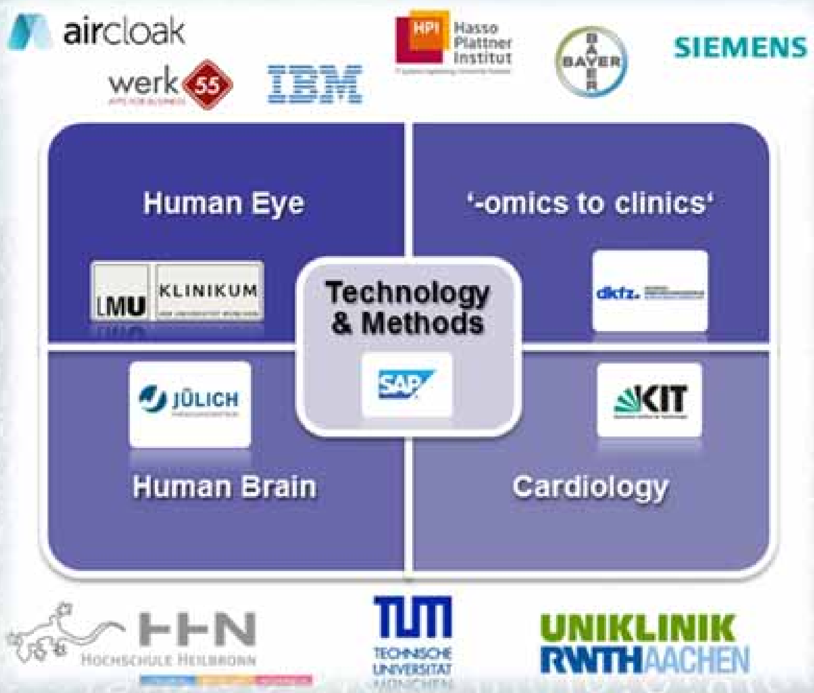
‘Big Data is everywhere’ – Where can we make a difference?

- **Much patient data available in SAP Hana systems**
- **Bayer does focussed patient studies**



- **Open upcoming omics-to-clinics meeting @ DKFZ**

- **Open the data from involved organizations is a key challenge (e.g. legal issues)**



- Driven by participating community partners and additional members (e.g. LMU ‘Human Eye Clinic‘)
- Clarify ‘scientific case‘ via template (vision, goals, data, impact, etc.)
- Explore new ‘smart data analytics‘ on existing and available data
- Combine scientific expertise with cutting-edge technology & methods

SDIL Medicine – Scientific Case Example

Towards Automation of 3D Reconstruction with 'Brain Analytics'

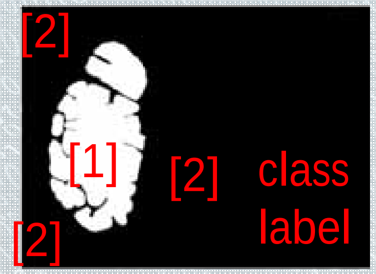
- **Scientific Case: Understanding 'Sectioning of the brain'**
- **Goal: Build 'reconstructed brain (one 3d volume)' that matches with sections based on block face images**

Classification++

Data Volume:

Block face images (of frozen tissue)
Every 20 micron (cut size)
Resolution: 3272 x 2469
~14 MB / RGB image
~ 8 MB / corresponding mask image
~700 Images

➔ ~40 GB dataset



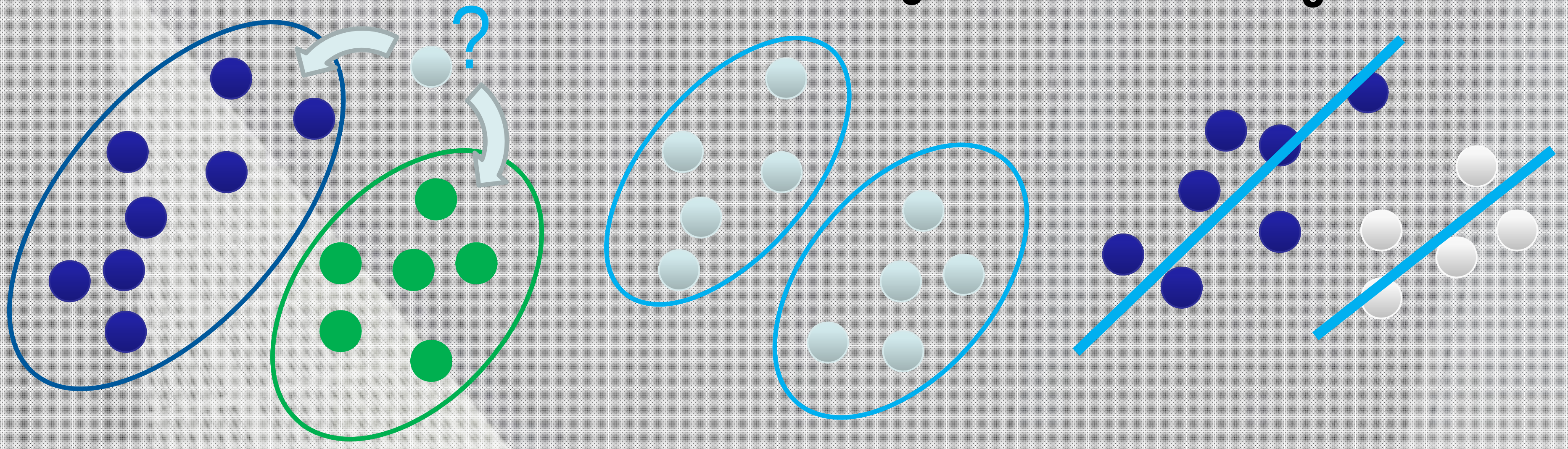
- **Investigation of technologies (e.g. IBM Watson Analytics system)**
- **Compare with approaches on different HPC & data platforms**

➢ **Collaboration INM & JSC – Identifying methods for new scanners (higher resolution)**

Making use of Big Data

Applying 'smart data analytics' techniques

Classification | **Clustering** | **Regression**

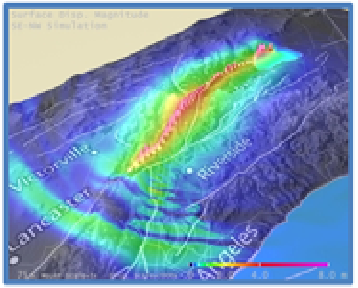


- ✓ *Groups of data exist*
- ✓ *New data classified to existing groups*
- ✓ *No groups of data exist*
- ✓ *Create groups from data close to each other*
- ✓ *Identify a line with a certain slope describing the data*

➢ **Many statistical data mining methods exist – but less are openly available as 'parallel'**

Large-scale Computational Parallel Applications Simulate Reality

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

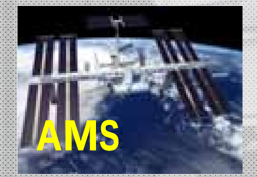


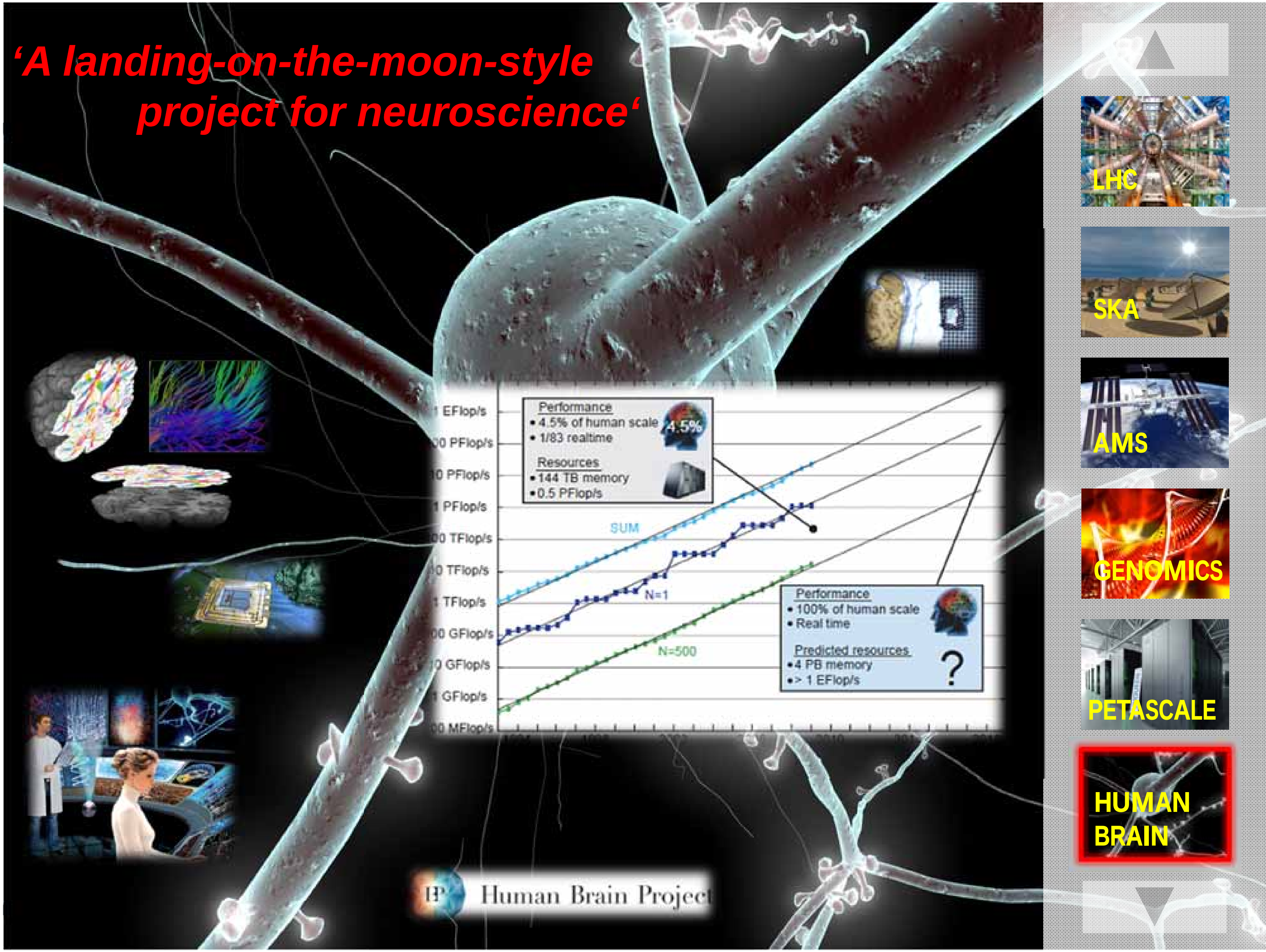
Source: Fran Berman, Maximising the Potential of Research Data

Better Simulations…

… means 'Bigger Data' &

… needs smart preservation…




'A landing-on-the-moon-style project for neuroscience'
1 EFlop/s
00 PFlop/s
10 PFlop/s
1 PFlop/s
00 TFlop/s
0 TFlop/s
1 TFlop/s
00 GFlop/s
0 GFlop/s
1 GFlop/s
00 MFlop/s
Performance
• 4.5% of human scale
• 1/83 realtime
4.5%
Resources
• 144 TB memory
• 0.5 PFlop/s
SUM
N=1
N=500
Performance
• 100% of human scale
• Real time
Predicted resources
• 4 PB memory
• > 1 EFlop/s
?
Human Brain Project
LHC
SKA
AMS
GENOMICS
PETASCALE
HUMAN BRAIN

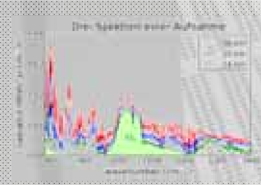
‘Big Data’ meets Computational Science

Smart Integration of simulation & experiment

‘Convert observed measurements into information about a physical object or system’ → *‘Inverse problems’*

‘Validation data’
From observations
(e.g. experiments)

‘Resolution tuning’
Resolution model to be in sync with observational data (e.g. check physical parameter we not directly can observe)



Status to be reconstructed from observations

High resolution measurement equipment - in-situ

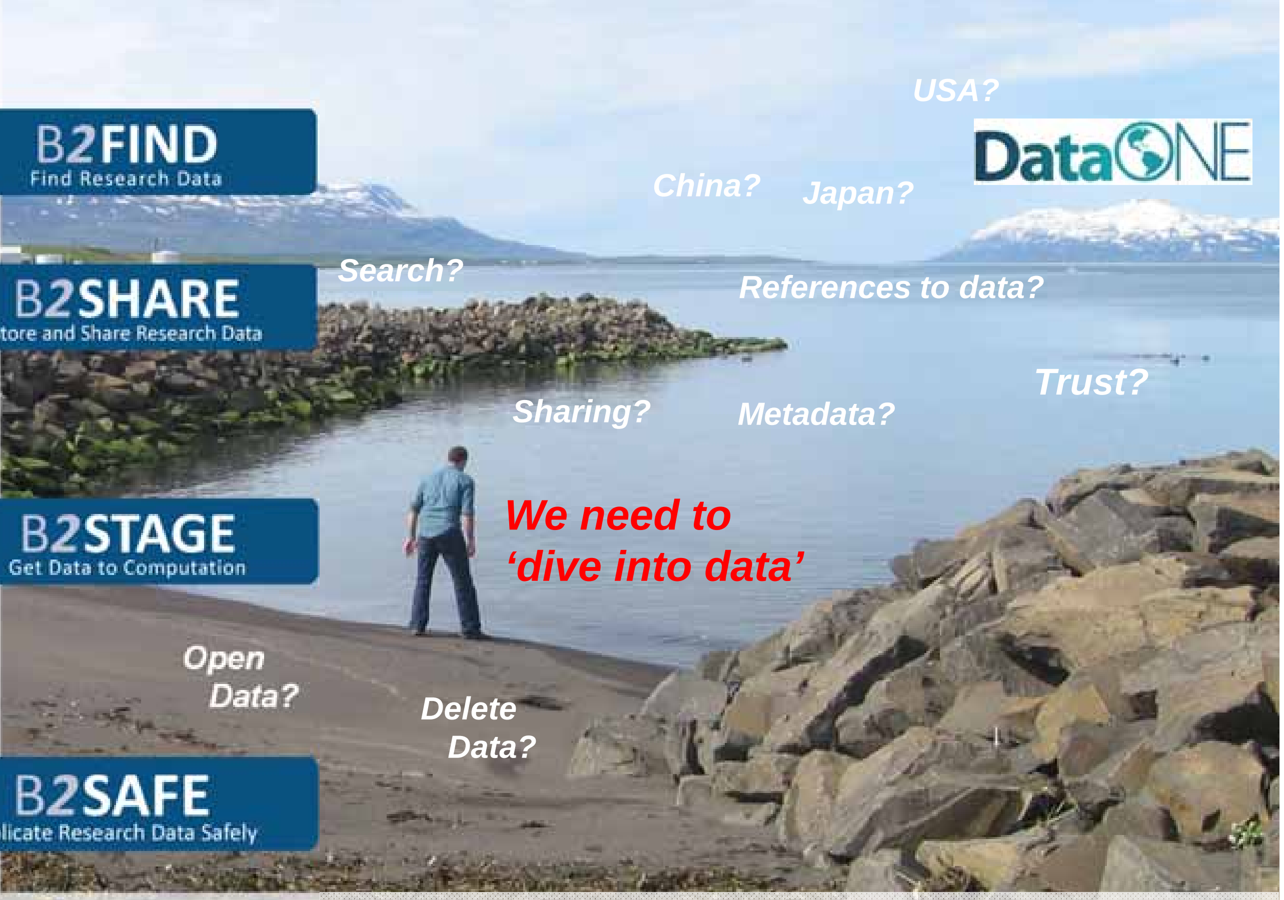
Modelling & Simulation on different scales

Formulation & Solving inverse problems

- **Slide material courtesy by Prof. Marquardt (modified and translated into English)**

Long-term Data Preservation and Curation…

… bears potentials to lower ‘Data Waves‘



Selected Benefits of open data infrastructures for science & engineering:

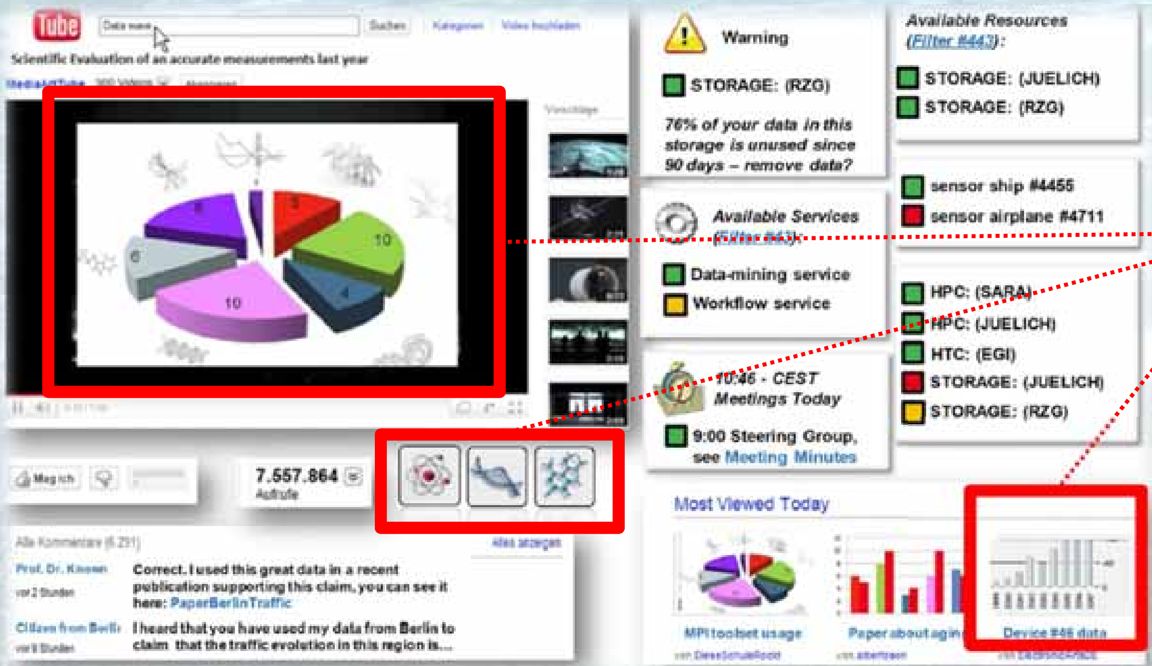
- ✓ ***High reliability,*** *so data scientists can count on its availability*
- ✓ ***Open deposit,*** *allowing user-community centres to store data easily*
- ✓ ***Persistent identification,*** *allowing data centres to register a huge amount of markers to track the origins and characteristics of the 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*

Understanding Possible Revenue Streams for Sustainability

Big Data Based Market-places

EUDAT
European Data Infrastructure

Enabling ‘apps‘, ‘subscription fees‘, ‘advertisement‘, ‘pay per use services‘



- ***Hooks for offerings around commercial software packages***
- ***Products around visualization packages and dedicated viewers***
- ***Easy links to ‘added value data’, e.g. available market statistics***
- ***Hosting services or deliver expandable storage in ‘peek’***
- ***Seamless links to the publishing and HPC application industry***
- ***Computing services to offer scalable data analytics***

Data (or ScienceTube) prototype to ‘dive into data’ with commercial ‘hooks’

B2SHARE
Store and Share Research Data

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

- **Presentation of Big Data Analytics IG on upcoming ‘RDA Germany‘ Event**

RESEARCH DATA ALLIANCE

- 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


Focussed
Group
RESEARCH DATA ALLIANCE
Big Data Analytics IG
Big Data Infrastructure WG
Problem Understanding
Data Understanding
Data Preparation
scientific data sets
Deployment
Modelling
Evaluation
P. Chapman et al., CRISP-DM Guide
Towards Systematic Data Analytics
„Reference Data Analytics"
for reusability & learning
Guided by the Cross Industry Standard Process
for Data Mining (CRISP-DM) Phases
CRISP-DM Report
Openly Shared Datasets
Running Analytics Code
'Building a UCI Repository for Big Data Analytics'

Results Example

- **Using EUDAT B2SHARE with Persistent Identifiers enables <u>trust to delete data</u> on different platforms (effect multiplies: x Phd students x teaching class)**

RESEARCH DATA ALLIANCE

Big Data Analytics IG

Big Data Infrastructure WG

Research Data Alliance

Future Grid

Twister

Iterative MapReduce

πSvM

Parallel Brain Data Analytics

Sattelite Data(Quickbird)

Parallel Support Vector Machines (SVM)

πSvM

HPC/MPI, Map-Reduce & GPGPUs

Classification Study of Land Cover Types

Classification++

'Best Practices'

Community-based practice

„Reference Data Analytics" for reusability & learning

CRISP-DM Report	Openly Shared Datasets	Running Analytics Code

B2SHARE

Store and Share Research Data

G. Cavallaro and M. Riedel, 'Smart Data Analytics Methods for Remote Sensing Applications', IGARSS 2014

Earth Science Data Analytics Examples

OpenGridForum

UNICORE

UNIVERSITY of VIRGINIA
Global Federated File System

PRACE

XSEDE
Extreme Science and Engineering Discovery Environment

Take Advantage of Interoperability…

…between EU PRACE & US XSEDE

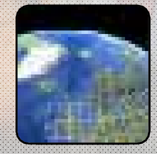
- **Presentation of PRACE Analytics next week at Brussels EC Event Infrastructures, Big Data & RDA**

PANGAEA

Problem: Quality control via outlier detection with PANGAEA data Collection

IAGOS

Problem: Longitude, latitude, altitude correlations with IAGOS data collection



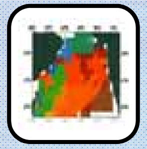
SCALE GIS

Problem: Projecting & transforming geospatial big data into a common coordinate reference framework

SEISMIC

Problem: Continuous seismic waveforms analysis for earthquakes monitoring

- **PhD studies Markus Götz**



NASA EVENTS (in-array DB analytics)

Problem: Event tracking analytics with spatial computing datasets (changing geolocations)

Shifts from Causality to Correlation

Challenging research with progress based on reason?

Traditional search for causality → (Big) Data Analysis

Exploring exactly WHY something is happening
Understanding causality is hard and time-consuming
Searching it often leads us down the wrong paths

(Big) Data Analytics

Not focussed on causality – enough THAT it is happening
Discover novel patterns and WHAT is happening
Using correlations for invaluable insights – data speaks for itself

2009 – H1N1 Virus Made Headlines

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

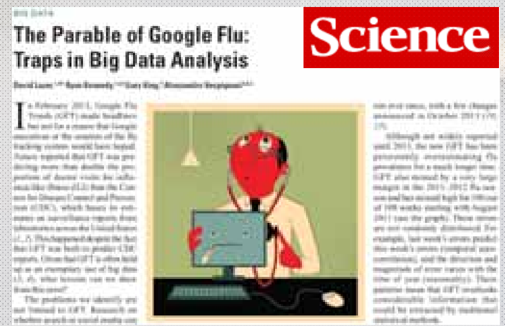
Jeremy Ginsburg et al., 'Detecting influenza epidemics using search engine query data', Nature 457, 2009

'Big Data is not always better data'

Selected Lessons Learned

2014 – The Parable of Google Flu

Large errors in flu prediction & lessons learned
(1) Dataset: Transparency & replicability impossible
(2) Study the algorithm since they keep changing
(3) It's not just about size of the data



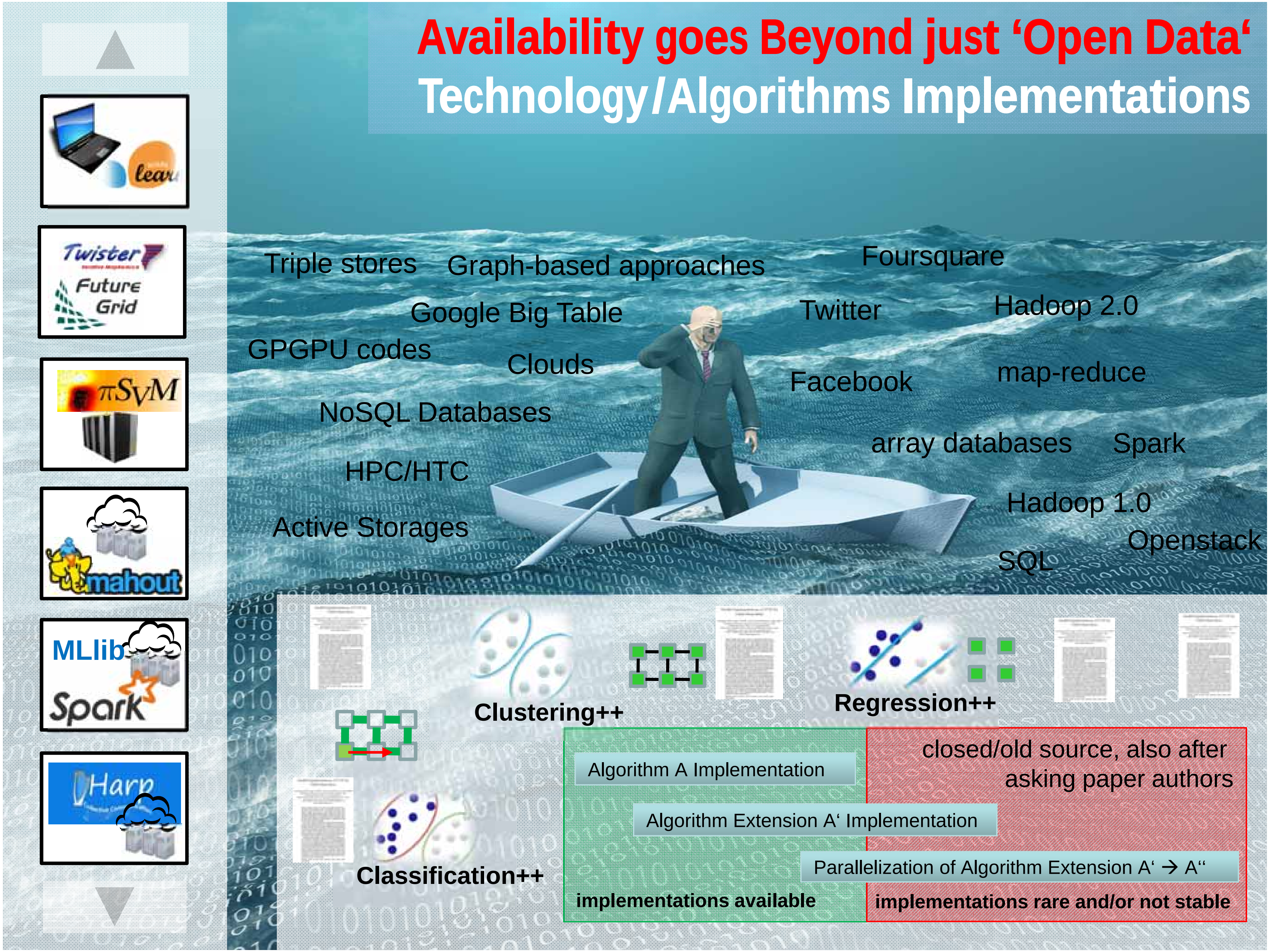
David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, 'The Parable of Google Flu: Traps in Big Data Analysis', Science Vol (343), 2014

Big Data Technology is Available – Usable?
Development Efforts require 'Steering'

Example: support vector machines (learning algorithm)

Tool	Platform Approach	Parallel Support Vector Machine
Apache Mahout	Java; Apache Hadoop 1.0 (map-reduce); HTC	No strategy for implementation (Website), serial SVM in code
Apache Spark/MLlib	Apache Spark; HTC	Only linear SVM; no multi-class implementation
Twister/ParallelSVM	Java; Apache Hadoop 1.0 (map-reduce); Twister (iterations), HTC	Much dependencies on other software: Hadoop, Messaging, etc.
Scikit-Learn	Python; HPC/HTC	Multi-class Implementations of SVM, but not fully parallelized
piSVM	C code; Message Passing Interface (MPI); HPC	Simple multi-class parallel SVM implementation outdated (~2011)
GPU accelerated LIBSVM	CUDA language	Multi-class parallel SVM, relatively hard to program, no std. (CUDA)
pSVM	C code; Message Passing Interface (MPI); HPC	Unstable beta, SVM implementation outdated (~2011)

Availability goes Beyond just 'Open Data'
Technology/Algorithms Implementations



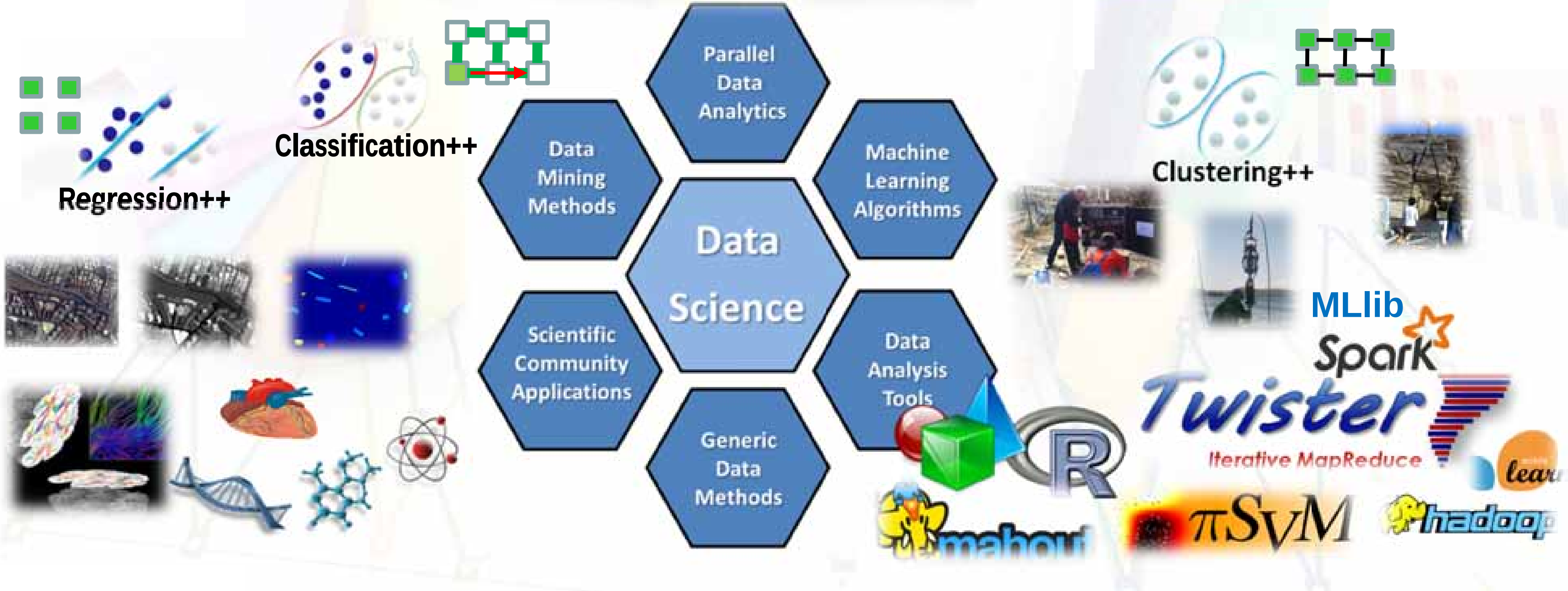
Lessons Learned from 'Big Data Analytics' to 'Smart Data Analytics'

'Scientific Big Data Analytics' – Massive amount of Methods

Selected Lessons Learned

- Agree(d) on focus areas
- Focus(sed) on scientific cases
- Guide(d) as community
- Gaine(d) trust to reduce/delete data
- Steer(ed) by domain experts

➢ To not get 'lost in big data' we need to apply key scientific principles (e.g. peer-review)



Lessons Learned from ‘Big Data Analytics‘ to ‘Smart Data Analytics‘

Requirements for ‘Scientific Big Data Analytics‘ are Real

Polar and Marine Research	AWI
Material Sciences	DESY
Biomedical data	DKFZ
Climate	DKRZ/HZG
Earth Observation	DLR
Epidemiology	DZNE
Biomolecular research	JUELICH
FAIR data	GSI
Environmental caused illness	HMGU
Photon / Neutron Research	HZB
Laser and magnetic fields research	HZDR
Astro physics	KIT
Research on water & geo data	UFZ

different data sources to integrate in analysis

different formats

Various technologies

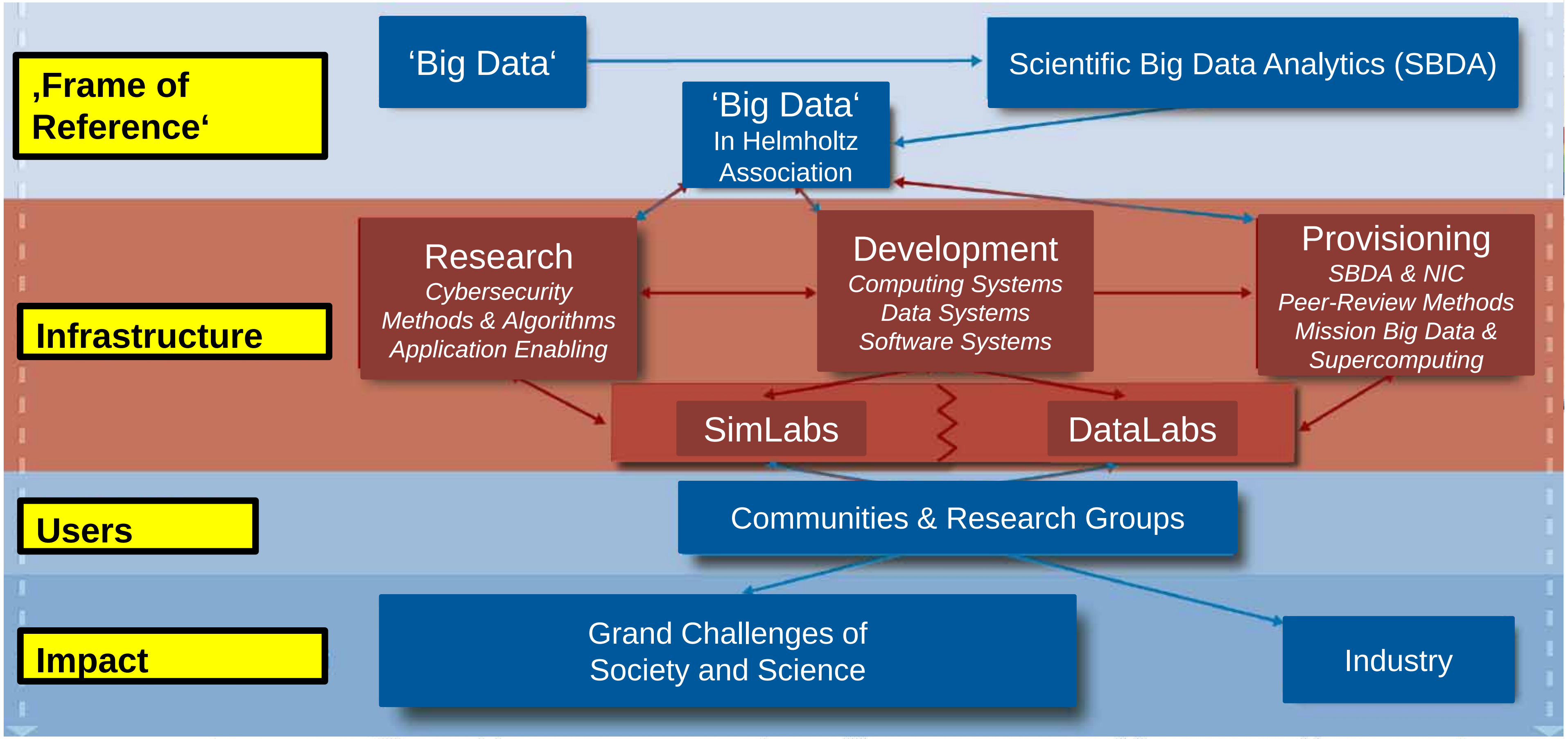
Sharing & reproducability

3D visualization & steering

Smart analytics & analysis

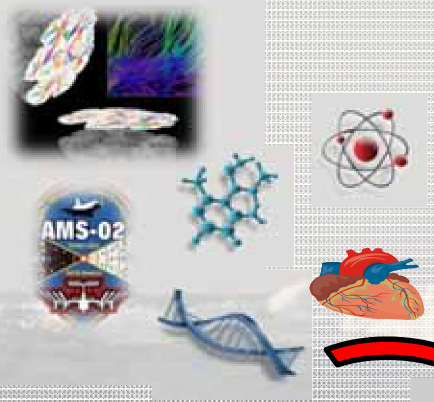
Lessons Learned from 'Big Data Analytics' to 'Smart Data Analytics'

'Scientific Big Data Analytics' needs Steering by Provisioning



Scientific Big Data Analytics: ‘Big Data‘-driven Research

Computation & Data Analysis gets more tightly intertwined



Scientific Computing

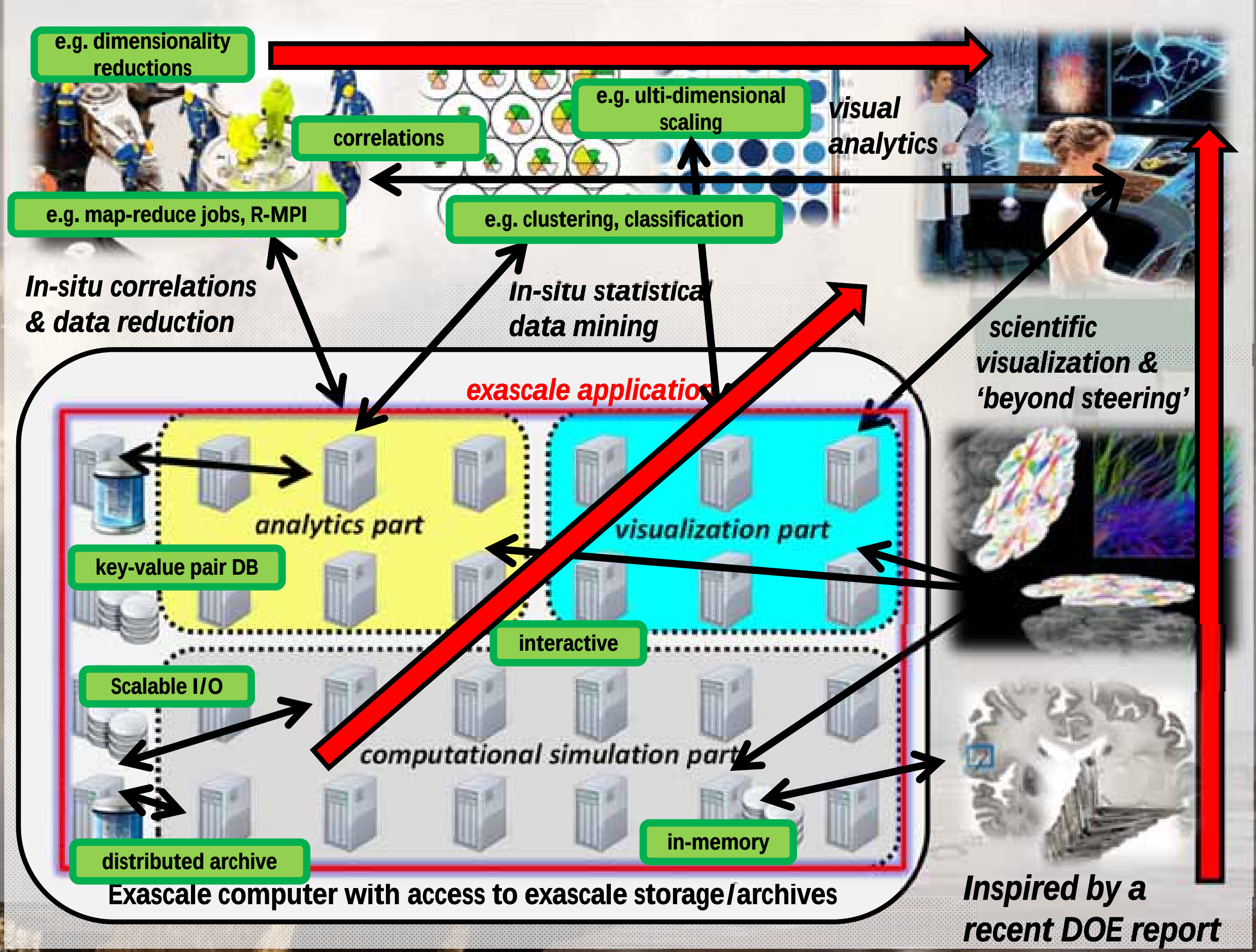
- Scientific Applications using ‘Big Data’
- Traditional Scientific Computing Methods
- HPC and HTC Paradigms & Parallelization
- Emerging Data Analytics Approaches
- Optimized Data Access & Management
- Statistical Data Mining & Machine Learning

‘Big Data’ Methods

Scientific & Engineering ‘Hybrid Applications’

Towards Exascale: Applications with combined characteristics of simulations & analytics

'In-Situ Analytics'



'Takk'

Talk available at:

www.morrisriedel.de/talks

Contact:

m.riedel@fz-juelich.de

Acknowledgements

Selected Members of the Research Group on High Productivity Data Processing

Ahmed Shiraz Memon
Mohammad Shahbaz Memon
Markus Goetz
Christian Bodenstein
Philipp Glock
Matthias Richerzhagen