Scientific Big Data Analytics
Practice & Experience

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

cac
Cloud and Autonomic Computing Conference

The International Conference on Cloud and Autonomic Computing (CAC 2014)

Imperial College, London September 8-12, 2014

'Big Data' in Science & Engineering

Scientific Big Data Analytics

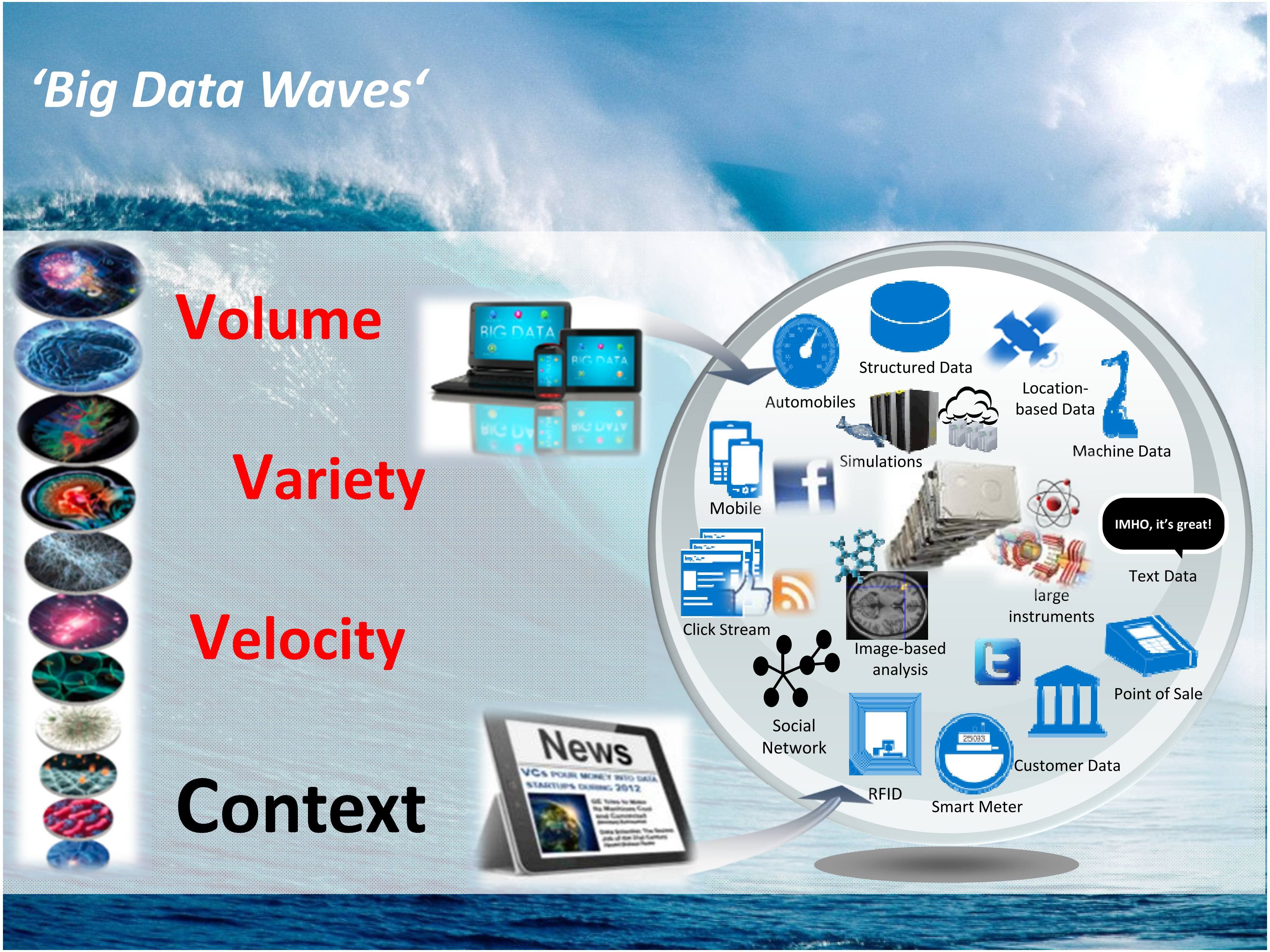
Selected Applications & Experiences

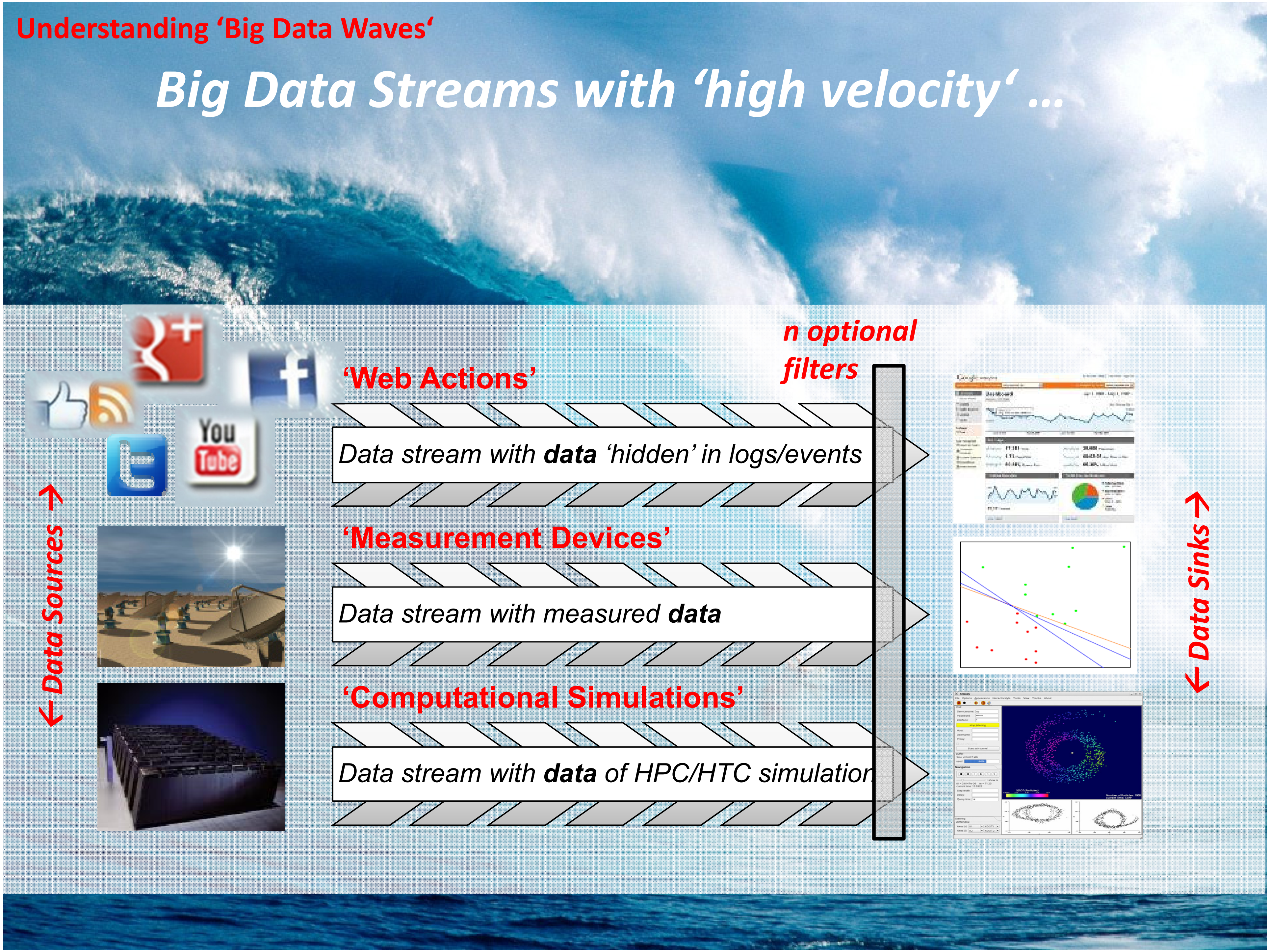
Key Examples of Analytics Practices

Analytics Tools – Lessons Learned

The Role of International Activities

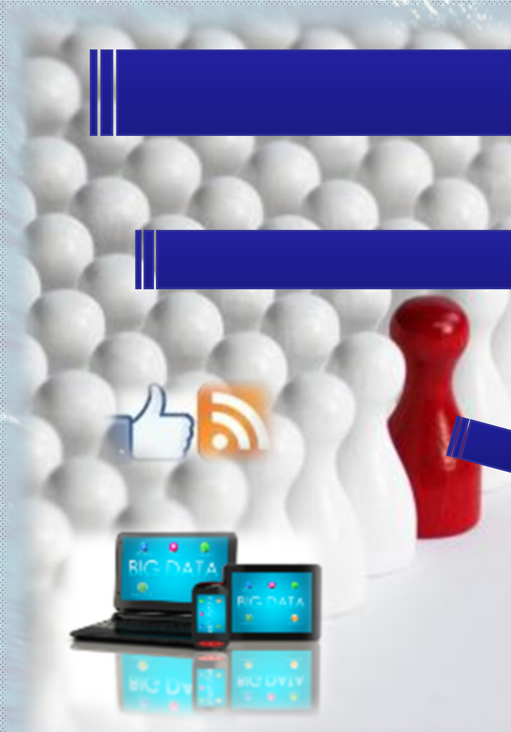
Some Questions & Possible Answers


'Big Data Waves'
Volume
Variety
Velocity
Context
BIG DATA
News
Structured Data
Automobiles
Location-based Data
Machine Data
Simulations
Mobile
IMHO, it's great!
Text Data
Click Stream
large instruments
Image-based analysis
Social Network
Point of Sale
Customer Data
RFID
Smart Meter


Understanding ‘Big Data Waves‘
Big Data Streams with ‘high velocity‘ …
n optional filters
‘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
← Data Sources →
← 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

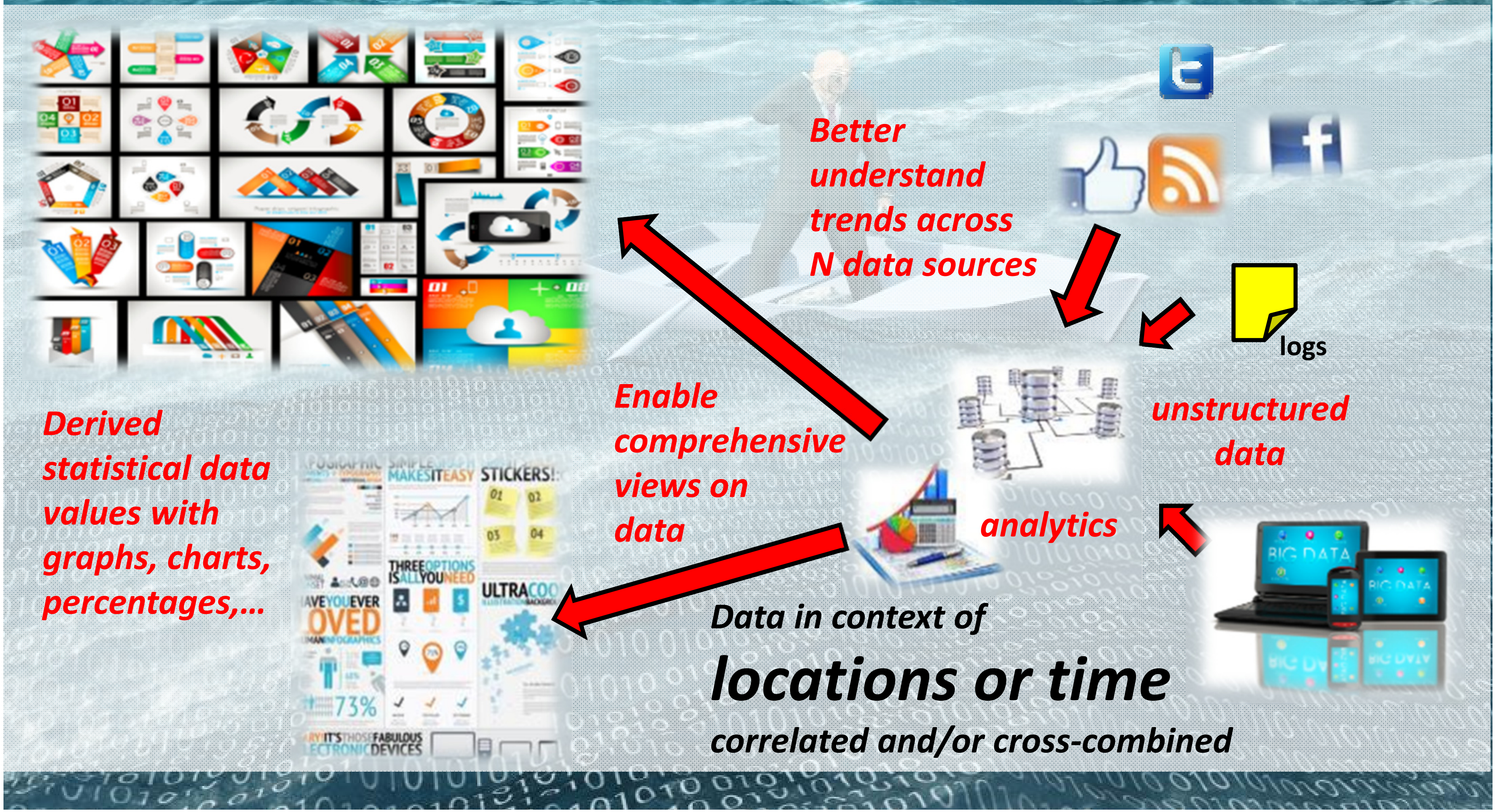
Scientific/Engineering Domain Experts

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

Terabytes

Infographics

Compact Combination of many Data Visualizations



Most data in the world…

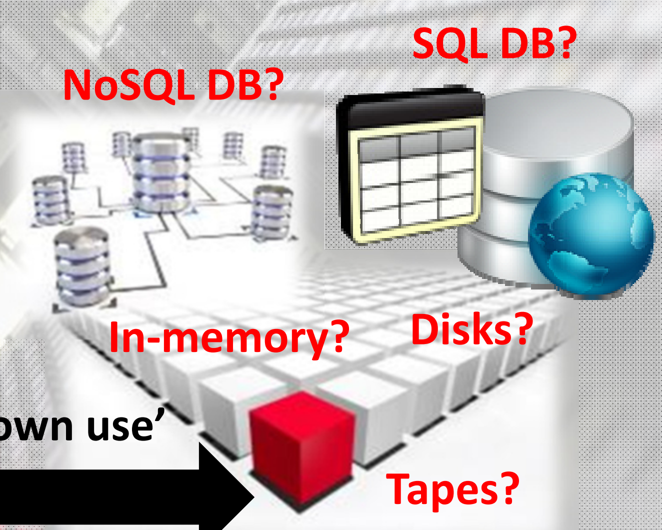
- …
- *Online Social Media*
  (videos, blogs, tweets,…)
- *Large number of log files*
  (Web server log, call center log,…)
- *Communication data*
  (E-Mails, chats, notes, letters, …)
- *Various document formats*
  (spreadsheet, presentation, docs)
- ….

… 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

Big Data Waves – Surfboards – Breakwaters

How to engage in the rising tide of 'Scientific Big Data'?

[1] Riding the Wave, EC Report, 2010

[2] A Surfboard for riding the wave, Report, 2012

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

' Time for Concrete'
Next Steps →

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.'*

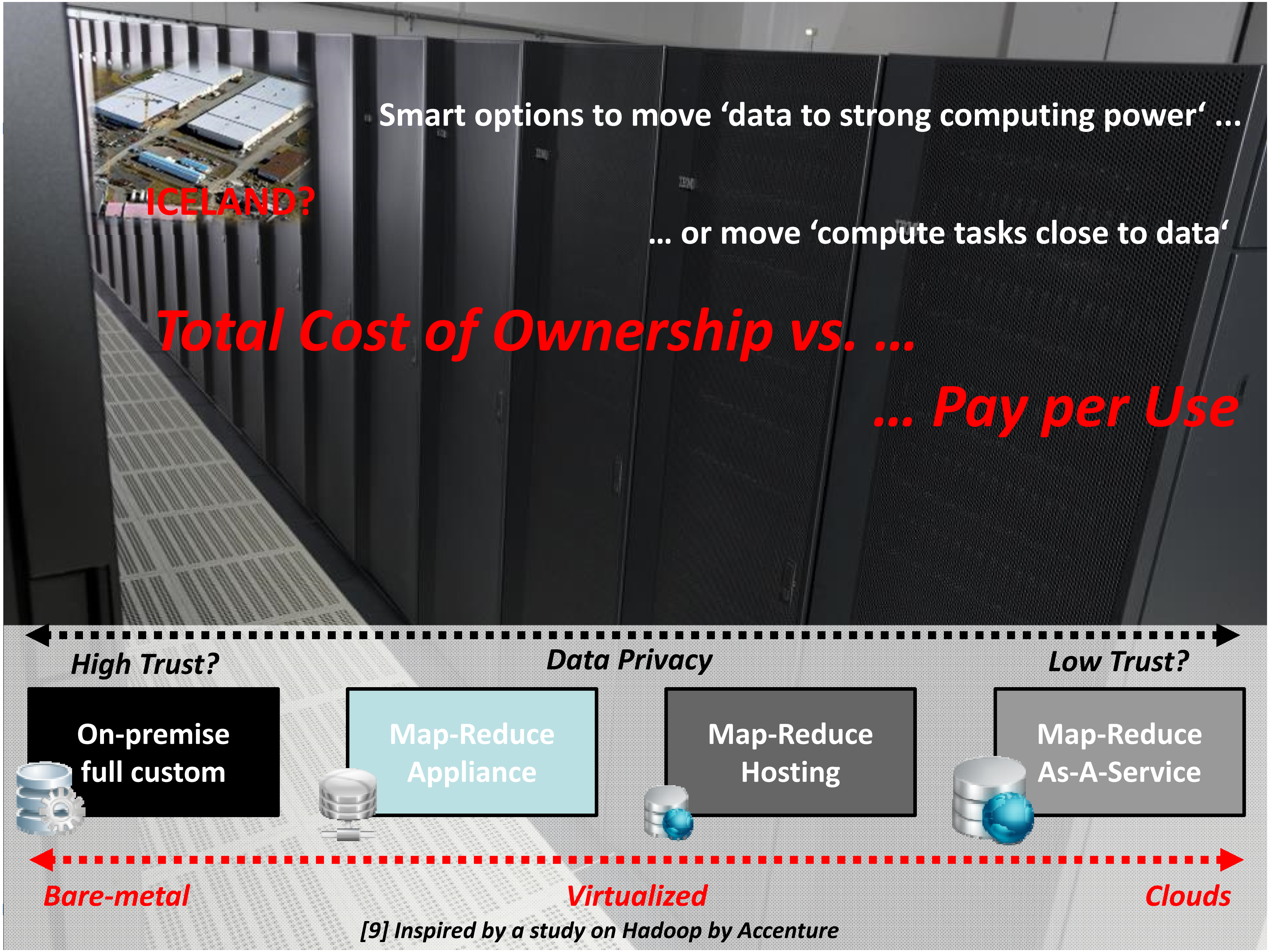
[1] Riding the Wave, EC Report, 2010

*'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, 2012

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

[3] DoE ASCAC Report, 2013

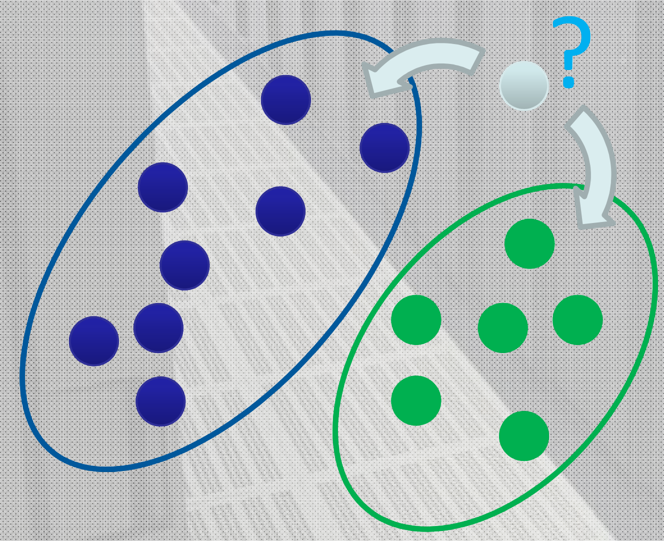


Smart options to move 'data to strong computing power' ...
ICELAND?
... or move 'compute tasks close to data'
Total Cost of Ownership vs. ...
... Pay per Use
High Trust?
Data Privacy
Low Trust?
On-premise full custom
Map-Reduce Appliance
Map-Reduce Hosting
Map-Reduce As-A-Service
Bare-metal
Virtualized
Clouds
[9] Inspired by a study on Hadoop by Accenture

Making use of Big Data

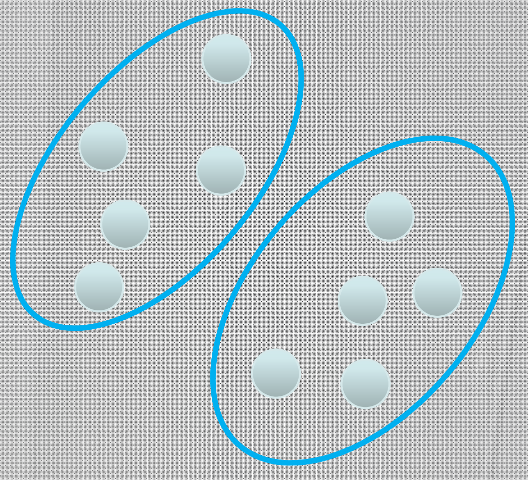
Applying 'smart data analytics' techniques

Classification



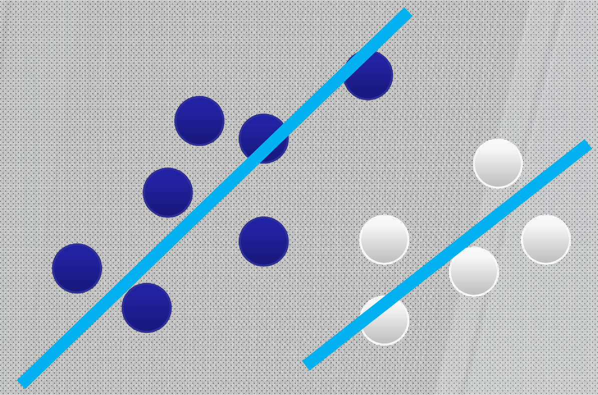
- ✓ ***Groups of data exist***
- ✓ ***New data classified to existing groups***

Clustering

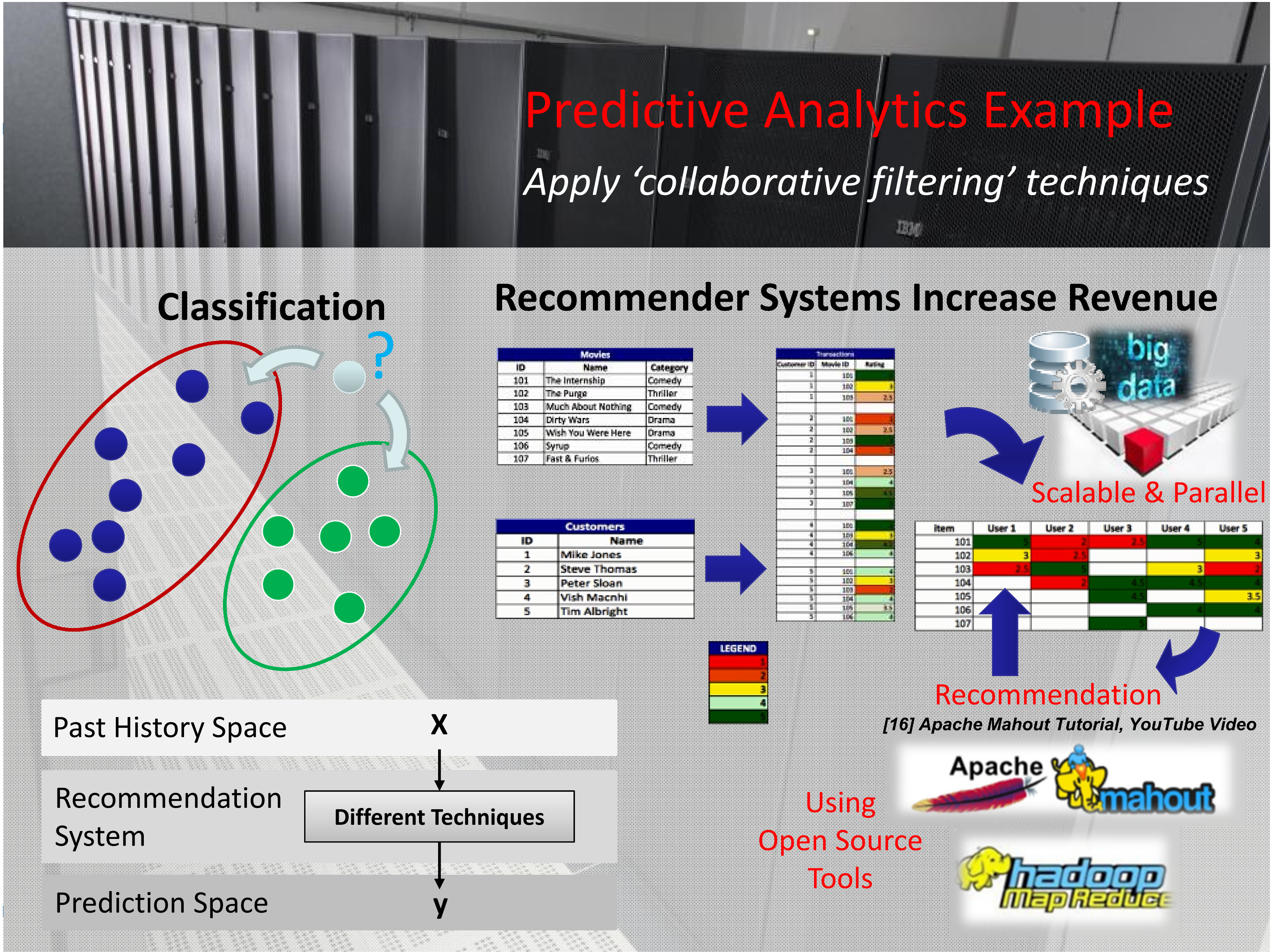


- ✓ ***No groups of data exist***
- ✓ ***Create groups from data close to each other***

Regression



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


Predictive Analytics Example
Apply 'collaborative filtering' techniques
Classification
Recommender Systems Increase Revenue
Movies
ID Name Category
101 The Internship Comedy
102 The Purge Thriller
103 Much About Nothing Comedy
104 Dirty Wars Drama
105 Wish You Were Here Drama
106 Syrup Comedy
107 Fast & Furios Thriller
Customers
ID Name
1 Mike Jones
2 Steve Thomas
3 Peter Sloan
4 Vish Macnhi
5 Tim Albright
Transactions
LEGEND
big data
Scalable & Parallel
Item User 1 User 2 User 3 User 4 User 5
Recommendation
[16] Apache Mahout Tutorial, YouTube Video
Apache mahout
Using
Open Source
Tools
hadoop Map Reduce
Past History Space
X
Recommendation
System
Different Techniques
Prediction Space
y

Utilities Sector Industry Reference

Instant Maintenance Workforce Management

Copa | Meter | **Wind** | Grid | Work | Loss | Churn | Tariff | Forecast | Model

Business Value Driver

- Increase average time between inspections
- Decrease lost power generation factor
- Decrease cost of spare parts

Big Data Impact

- Ad-hoc analysis on large data volumes to predict, monitor and optimize performance and component breakdown
- Enable "Process-to-device"

Value Chain

Generation

Process

Turbine Maintenance

Value

Feasibility

Phase

1. 2. 3.

Lead User(s)

➢ *Slide courtesy of Dr. S. Fischer, Global Head of Applied Research – SAP AG (now working at Trumph)*

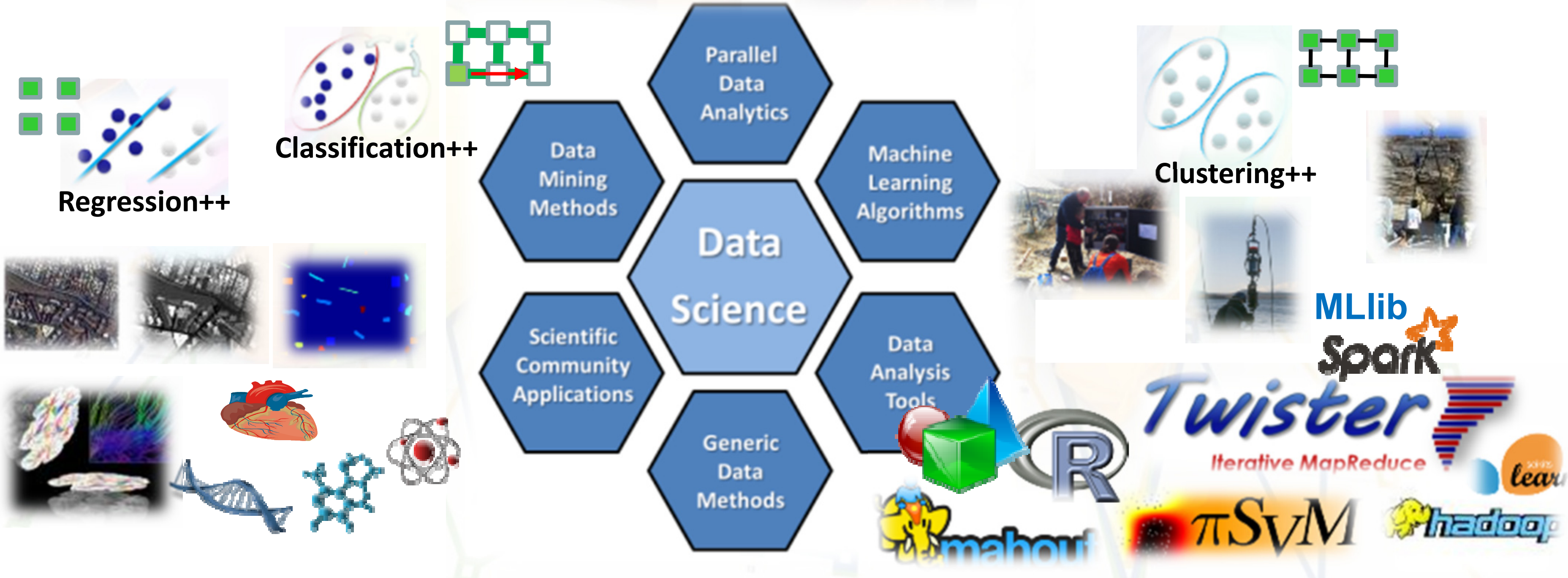
Smart Data Innovation Lab

Companies and Academia work in Focussed Areas

Smart Data Innovation Lab

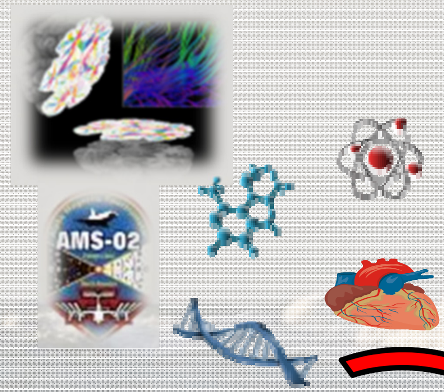
Serial Algorithms for Large Volumes of Data Exist

‘Big Data‘ Requires Parallel Algorithms and Open Availability



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

Computation & Data Analysis gets more tightly intertwined



Scientific Computing

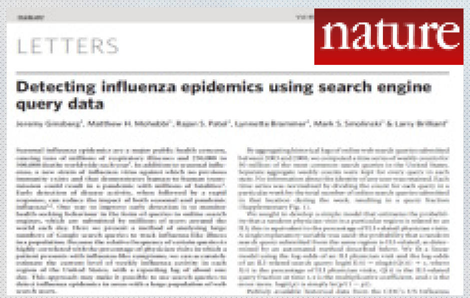
- Scientific Applications using 'Big Data'
- Traditional Scientific Computing Methods
- HPC and HTC Paradigms & Parallelization
- Emerging Data Analytics Approaches
- Optimized Data Access & Management
- Statistical Data Mining & Machine Learning

'Big Data' Methods

Scientific Applications

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‘


LETTERS

nature

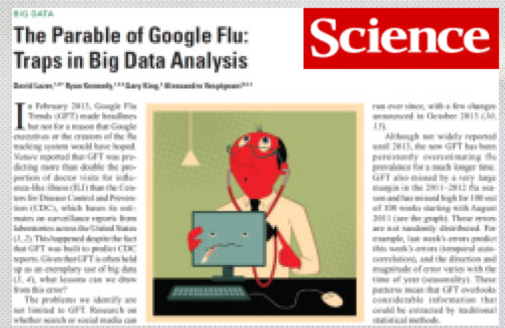
Detecting influenza epidemics using search engine query data

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

‘Big Data is not always better data‘

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


The Parable of Google Flu: Traps in Big Data Analysis

Science

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

Shifts from Causality to Correlation

Challenging research with progress based on reason?

'A clever combination of both is needed'

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

The Large Hadron Collider at CERN

'Scientific Big Data Application' with First Results

'Results today only possible due to extraordinary performance of Accelerators – Experiments – Grid computing'.

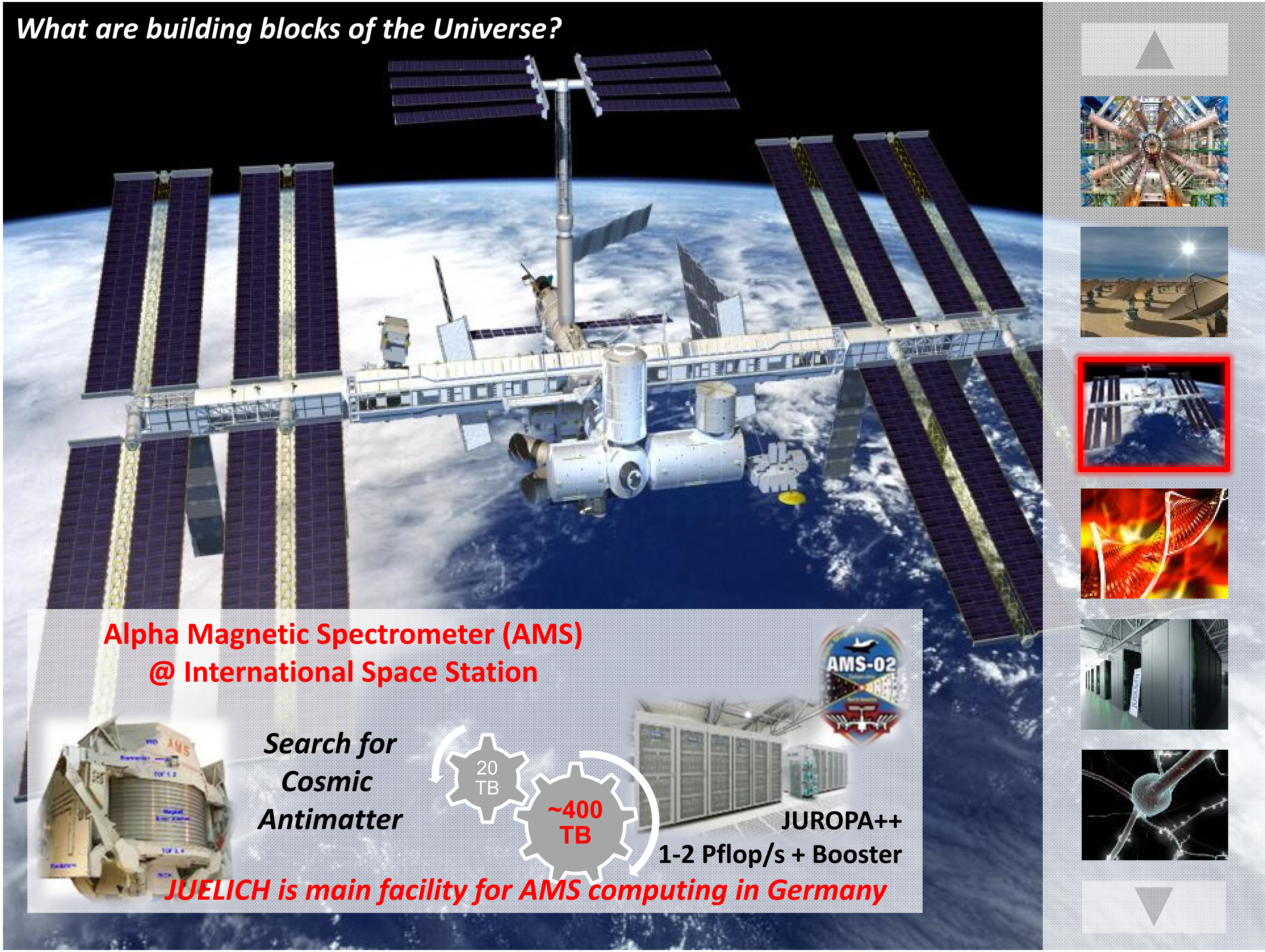
[7] Prof. Rolf-Dieter Heuer, CERN Director General, in the context of the Higgs Boson Discovery

Data Volume:

4 experiments / detectors
ca. 10^6 bytes / accident / experiment
ca. $3 * 10^2$ accidents per second
ca. 10^5 seconds per day
ca. 10^2 (experiments-) days per year
➔ $12 * 10^{15}$ bytes / year

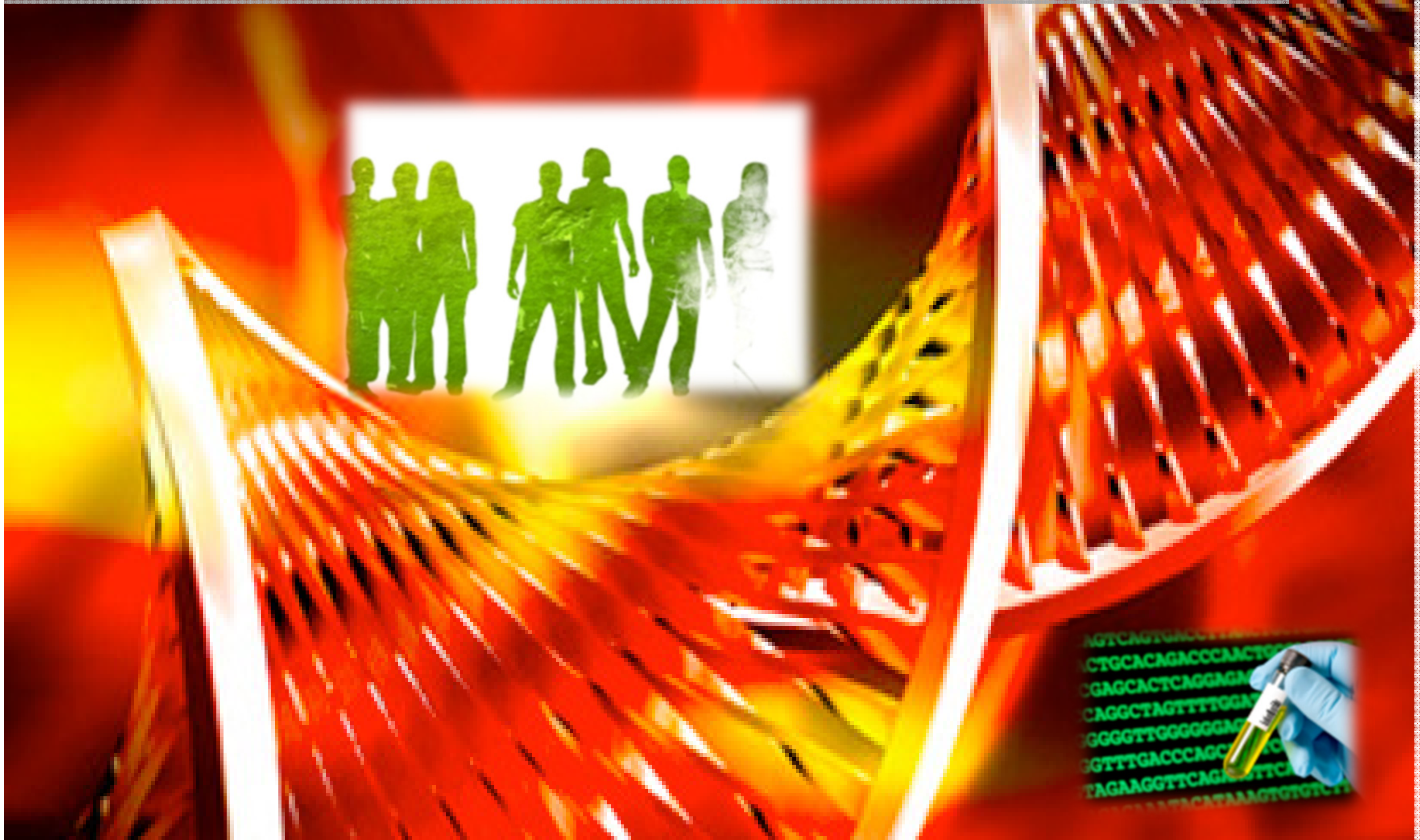
= 12 Petabytes per year


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 Juelich
xilostudios.com


What are building blocks of the Universe?
Alpha Magnetic Spectrometer (AMS)
@ International Space Station
AMS-02
Search for
Cosmic
Antimatter
20
TB
~400
TB
JUROPA++
1-2 Pflop/s + Booster
JUELICH is main facility for AMS computing in Germany

The 1000 Genome Project

Understanding what makes us different from one another



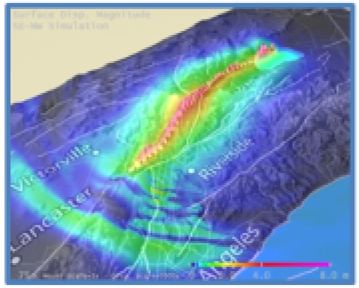
Comparing the complete DNA sequences of more than 1,000 individuals from around the world

Next challenge: 2000 individuals with each > 3 billion DNA base pairs

= 6 trillion DNA bases

Large-scale Computational Parallel Applications Simulate Reality

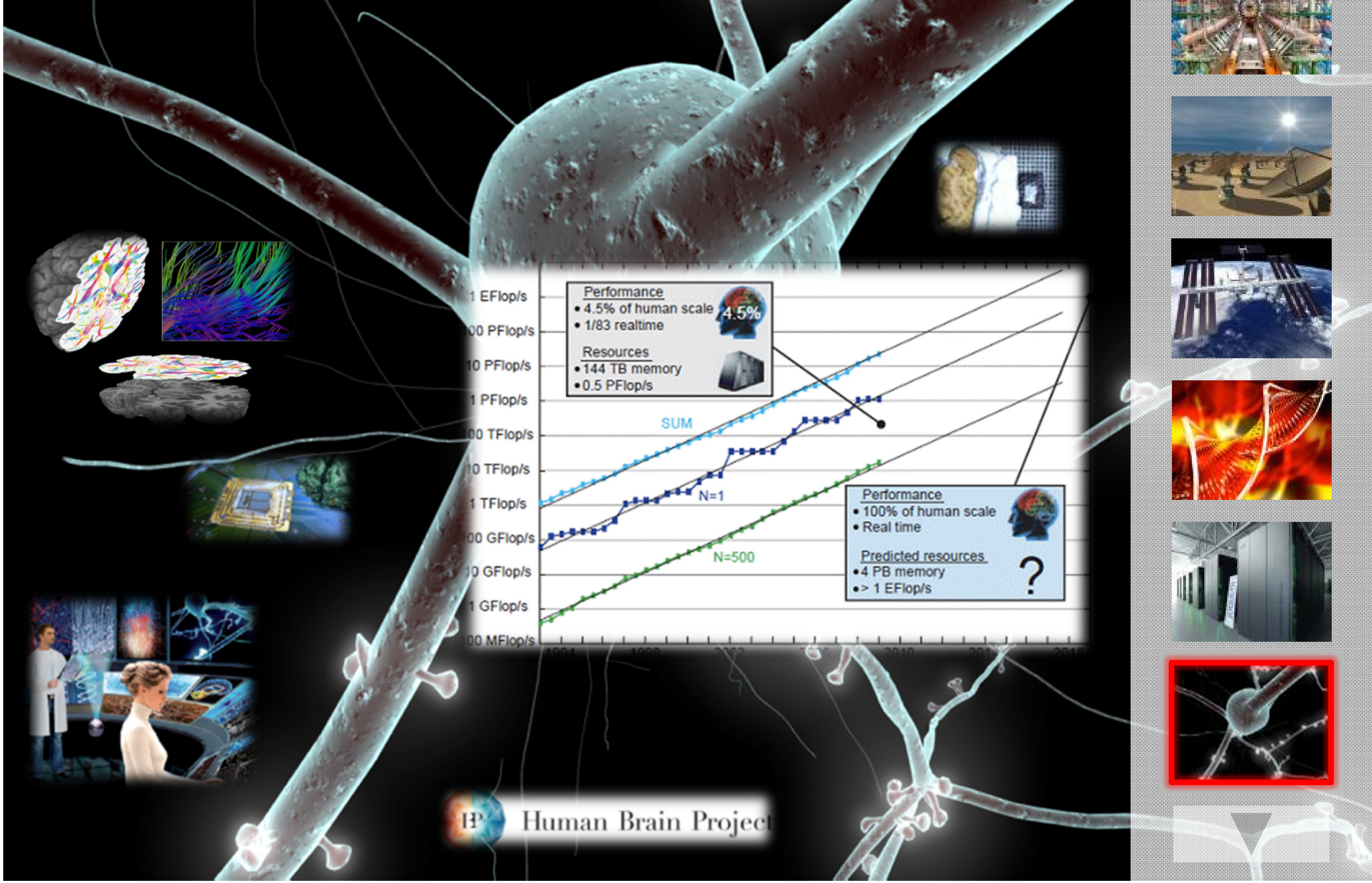
Estimated figures for simulated 240 second period, 100 hour run-time	TeraShake domain (600x300x80 km^3)	PetaShake domain (800x400x100 km^3)
Fault system interaction	NO	YES
Inner Scale	200m	25m
Resolution of terrain grid	1.8 billion mesh points	2.0 trillion mesh points
Magnitude of Earthquake	7.7	8.1
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



[8] Fran Berman, Maximising the Potential of Research Data

Better Simulations…
… means 'Bigger Data'

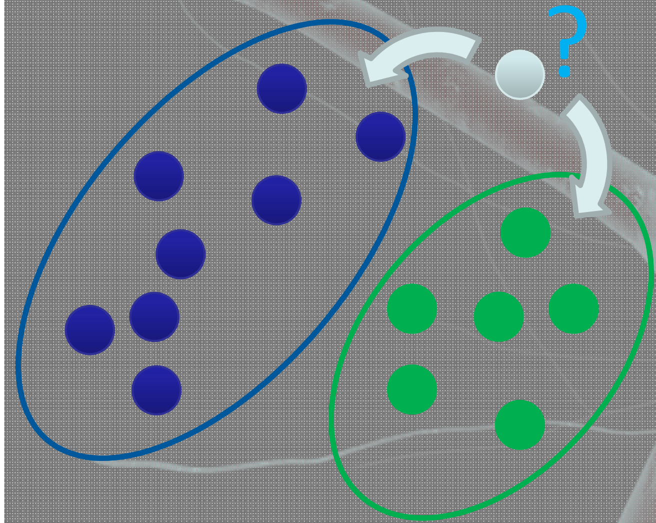
‘A landing-on-the-moon-style project for neuroscience‘



Complex Neuroscience Analytics

Lessons Learned towards Cloud and Autonomic Computing

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**

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

Model Selection and Cross-Validation

Working with Analytics Methods is a Process with Phases

Parallel Support Vector Machine

One of the most succesful classification methods

Classifier separates Two classes (brain, non-brain)

Parameters C & gamma after cross-validation

Cross-validation (grid-search) nicely parallel (e.g. for clouds

Uses quadratic programming & Lagrangian method with **N x N**

$$\min_{w,\xi_i,b}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i\right\}$$

(optimization problem)

$$\mathcal{L}(\alpha) = \sum_{n=1}^{N}\alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} y_n y_m \alpha_n \alpha_m \mathbf{x}_n^T \mathbf{x}_m$$

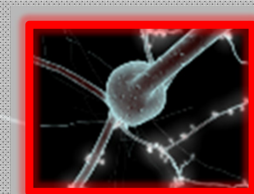
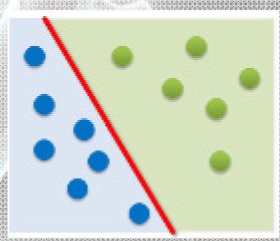
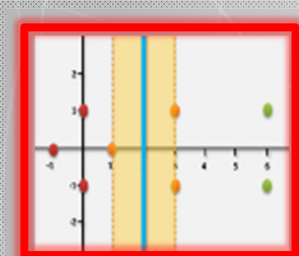
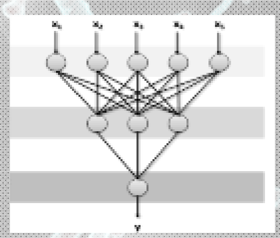
(max. hyperplane → dual problem)

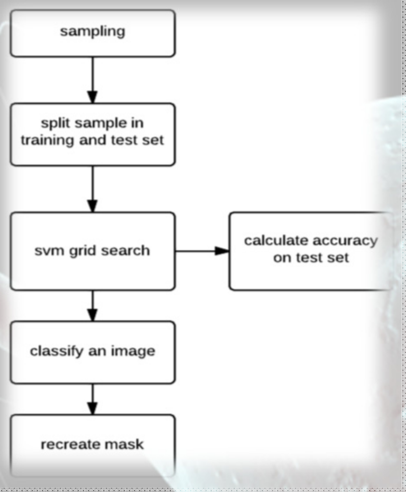
$$\begin{bmatrix} y_1 y_1 x_1^T x_1 & y_1 y_2 x_1^T x_2 & \ldots y_1 y_N x_1^T x_N \\ \ldots & \ldots & \ddots \\ y_N y_1 x_N^T x_1 & y_N y_2 x_N^T x_2 & \ldots y_N y_N x_N^T x_N \end{bmatrix}$$

(quadratic coefficients)



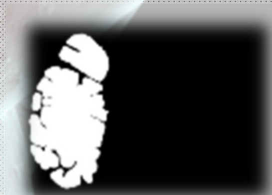




Challenge:

Sampling bias:
more black than brain pixels in groundtruth

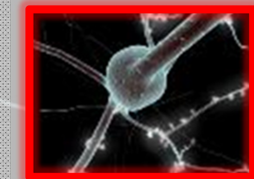


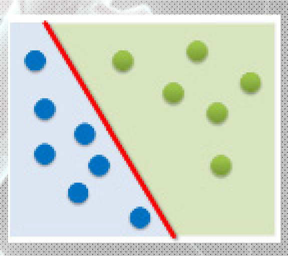
Sampling

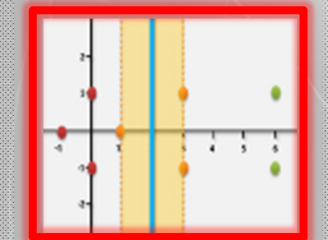
Draw balanced random class samples, eg.: 0.01%

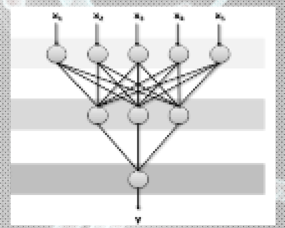
Results

After performing manual sampling
Size of training data is limited (serial method)
→96% accuracy – but some remaining issues



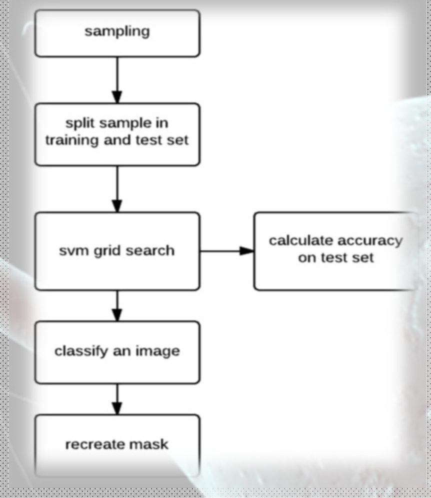


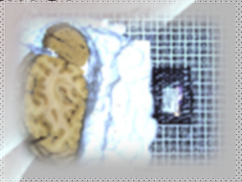




Challenges:

Distribute the data across the parallel infrastructure





ParallelSVM implementation on top of Twister stack (~development version, but works)

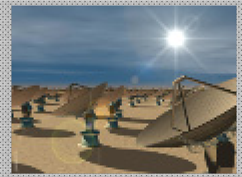
[15] Sun Z., and Fox G., 'Study on Parallel SVM Based on MapReduce'

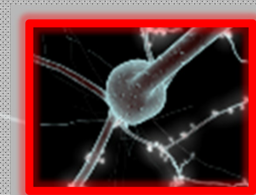
Results

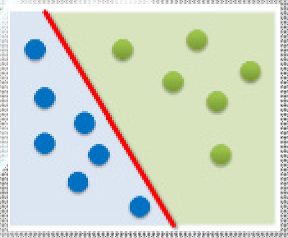
After performing manual sampling

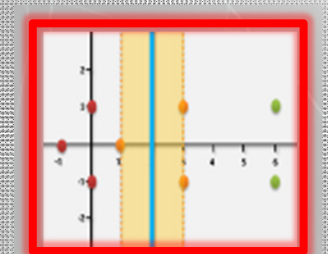
Size of training data scales much better (parallel)

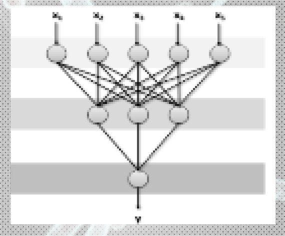
→96% accuracy – but some remaining issues









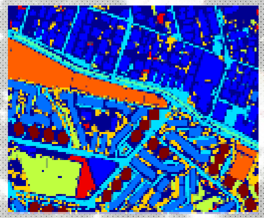


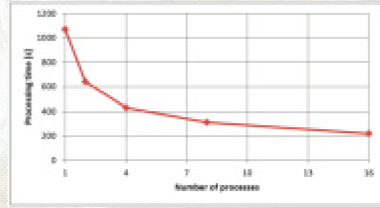
Remote Sensing Community

Large hyper- and multi-spectral datasets

Challenge: Multi-class classification

Classify different land-cover types





Results

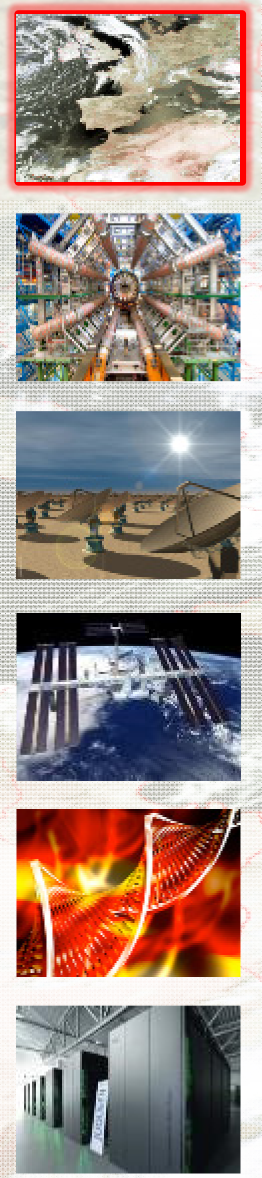
[12] G. Cavallaro and M. Riedel, 'Smart Data Analytics Methods for Remote Sensing Applications', IGARSS 2014

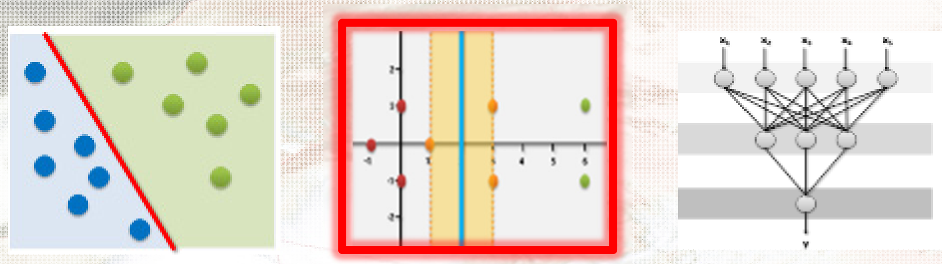
Speed-up and decrease of training time

Applied Self-Dual Attribute Profile (SDAP)

From big data to smart data with statistics (PCA)

→97% accuracy





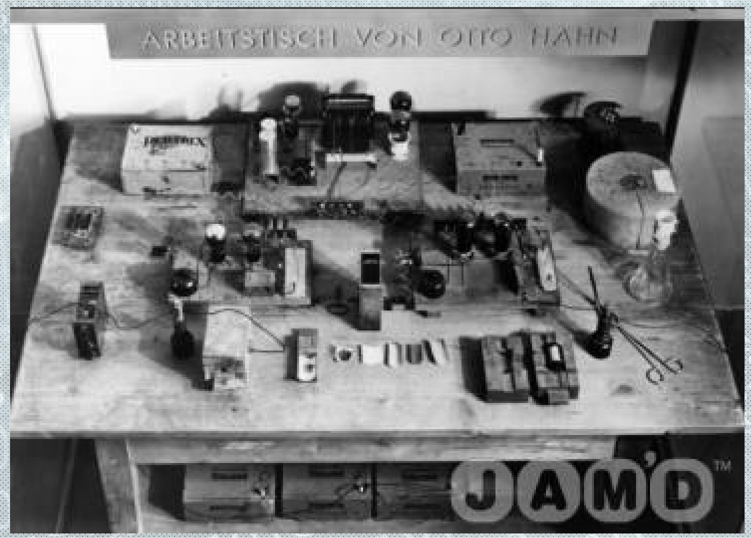
Big Data Applications – Statistical Data Mining Techniques

What is the right equipment, tool, technology, infrastructure?

The equipment and workbench used by Otto Hahn (1879 - 1968) and Fritz Strassmann in December 1938

-

No opportunity to use an e-Infrastructure or Clouds



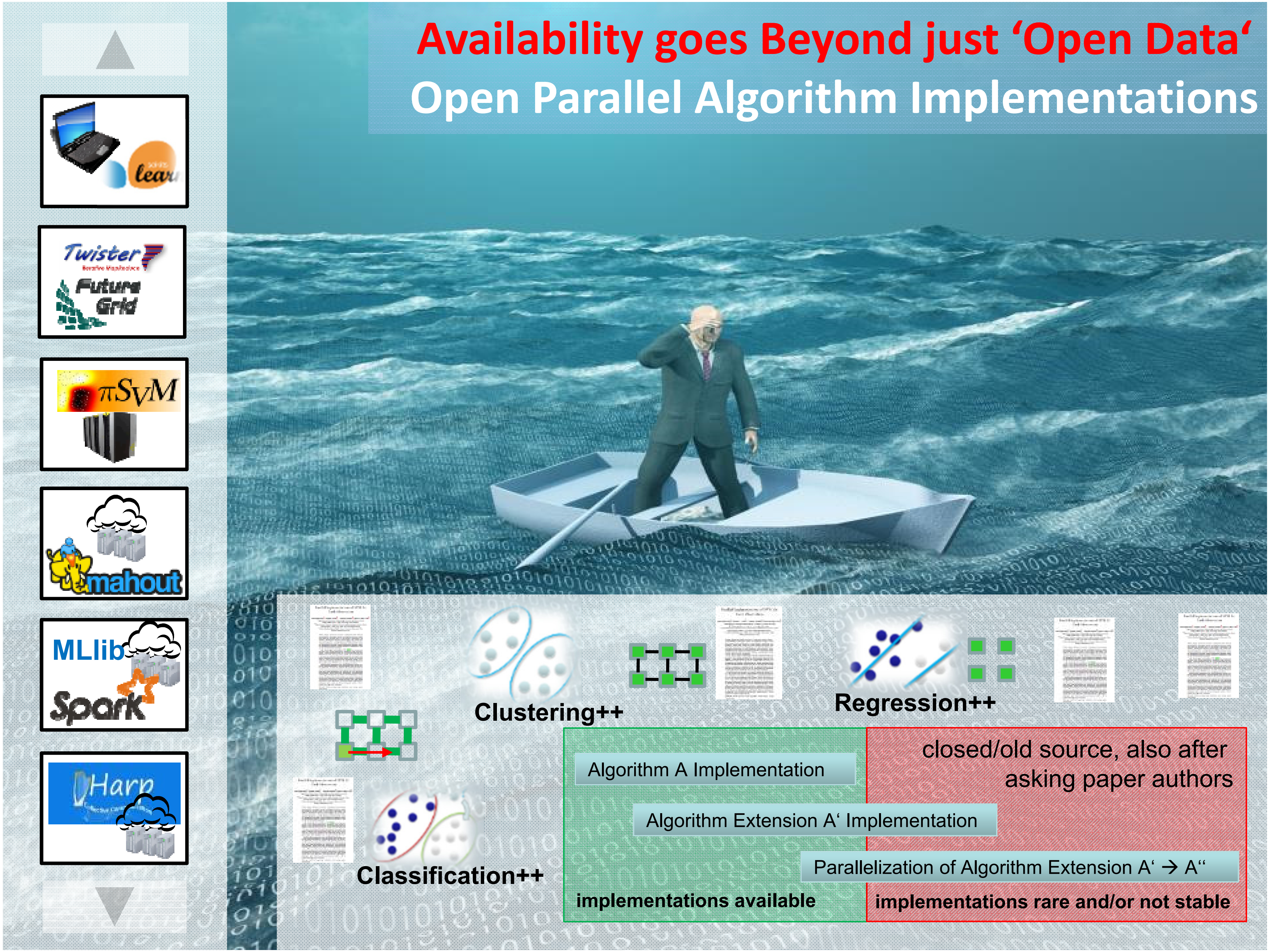
simple – yet powerful

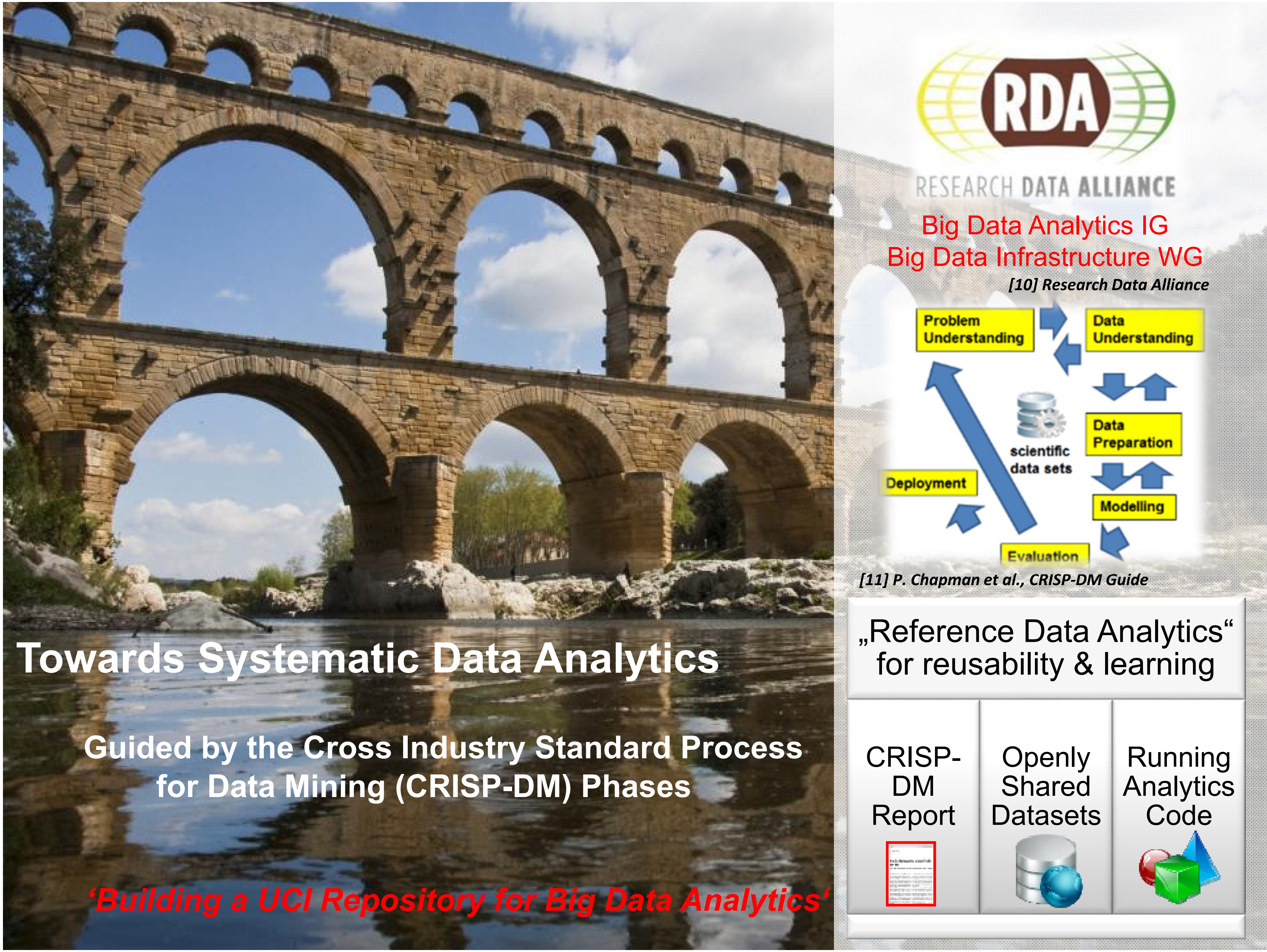
'provided the first chemical evidence of nuclear fission products'

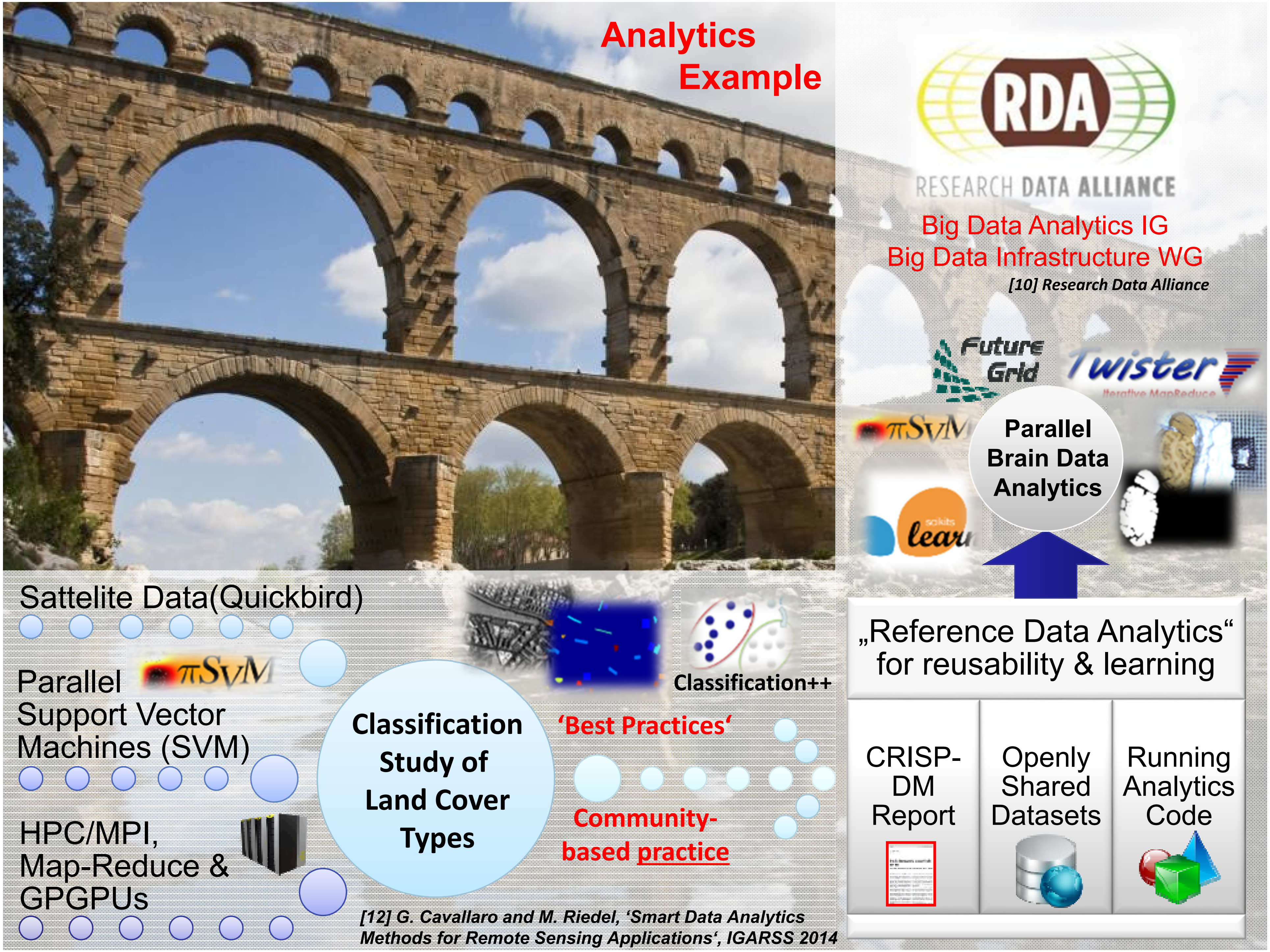
Big Data Technology is Available
But Need More Parallel Machine Learning

Our Workbench (e.g. focus on available parallel SVMs)

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'
Open Parallel Algorithm Implementations
Twister
Future Grid
πSvM
mahout
MLlib
Spark
Harp
Clustering++
Regression++
Classification++
Algorithm A Implementation
Algorithm Extension A' Implementation
Parallelization of Algorithm Extension A' → A''
closed/old source, also after asking paper authors
implementations available
implementations rare and/or not stable


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


Analytics
Example
RESEARCH DATA ALLIANCE
Big Data Analytics IG
Big Data Infrastructure WG
[10] Research Data Alliance
Future Grid
Twister
Iterative MapReduce
πSvM
Parallel Brain Data Analytics
Sattelite Data(Quickbird)
Parallel
Support Vector
Machines (SVM)
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
[12] G. Cavallaro and M. Riedel, 'Smart Data Analytics Methods for Remote Sensing Applications', IGARSS 2014

[13] Open Grid Forum

Basic Execution Service

Job Submission Description Language

GLUE2 Distributed Resource Descriptions

Open Cloud Computing Interface

Simple API for Grid Applications

Distributed Resource Management Application API

❖ *OGF is co-located with this conference – check the program*

Reliable Specifications to Build Upon

Standardized Building Blocks – 'Rock Solid'

Earth Science Data Analytics Examples

Take Advantage of Interoperability…

…between PRACE & XSEDE

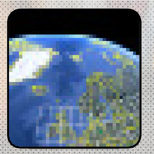
[13] Open Grid Forum

[14] UNICORE.eu

Global Federated File System

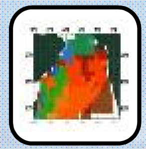
PANGAEA

Problem: Quality control via outlier detection with PANGAEA data Collection



SCALE GIS

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



EVENTS

IAGOS

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

SEISMIC

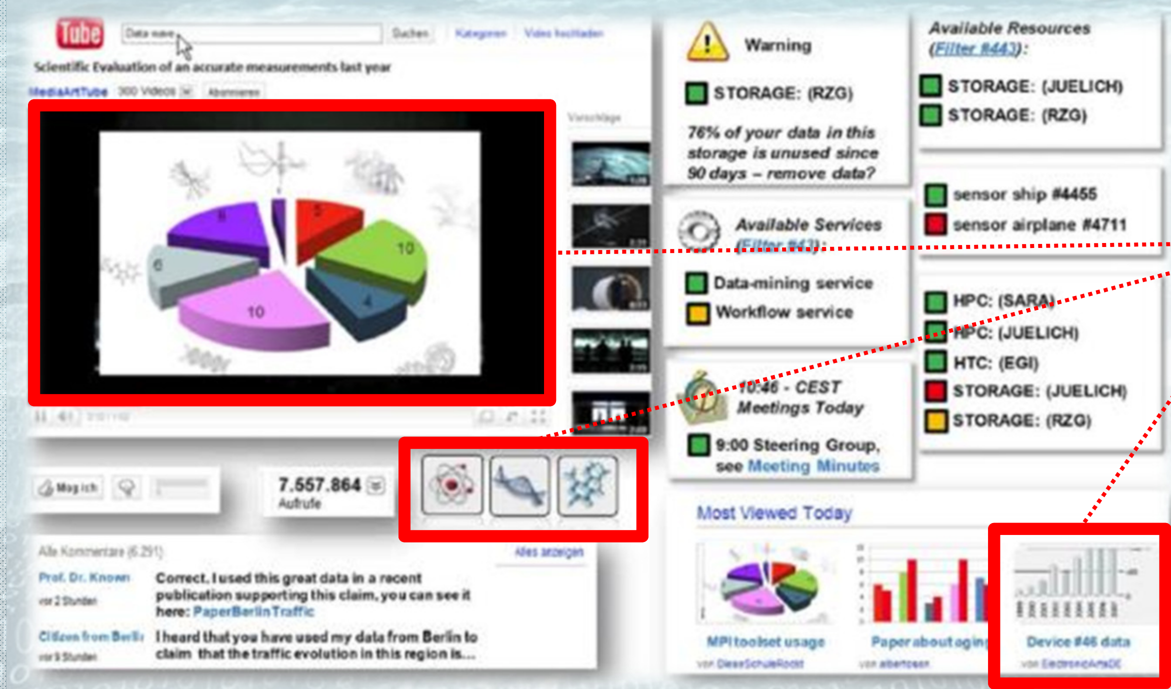
Problem: Continuous seismic waveforms analysis for earthquakes monitoring

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

Understanding Possible Revenue Streams in Science & Engineering

Big Data Based Market-places

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



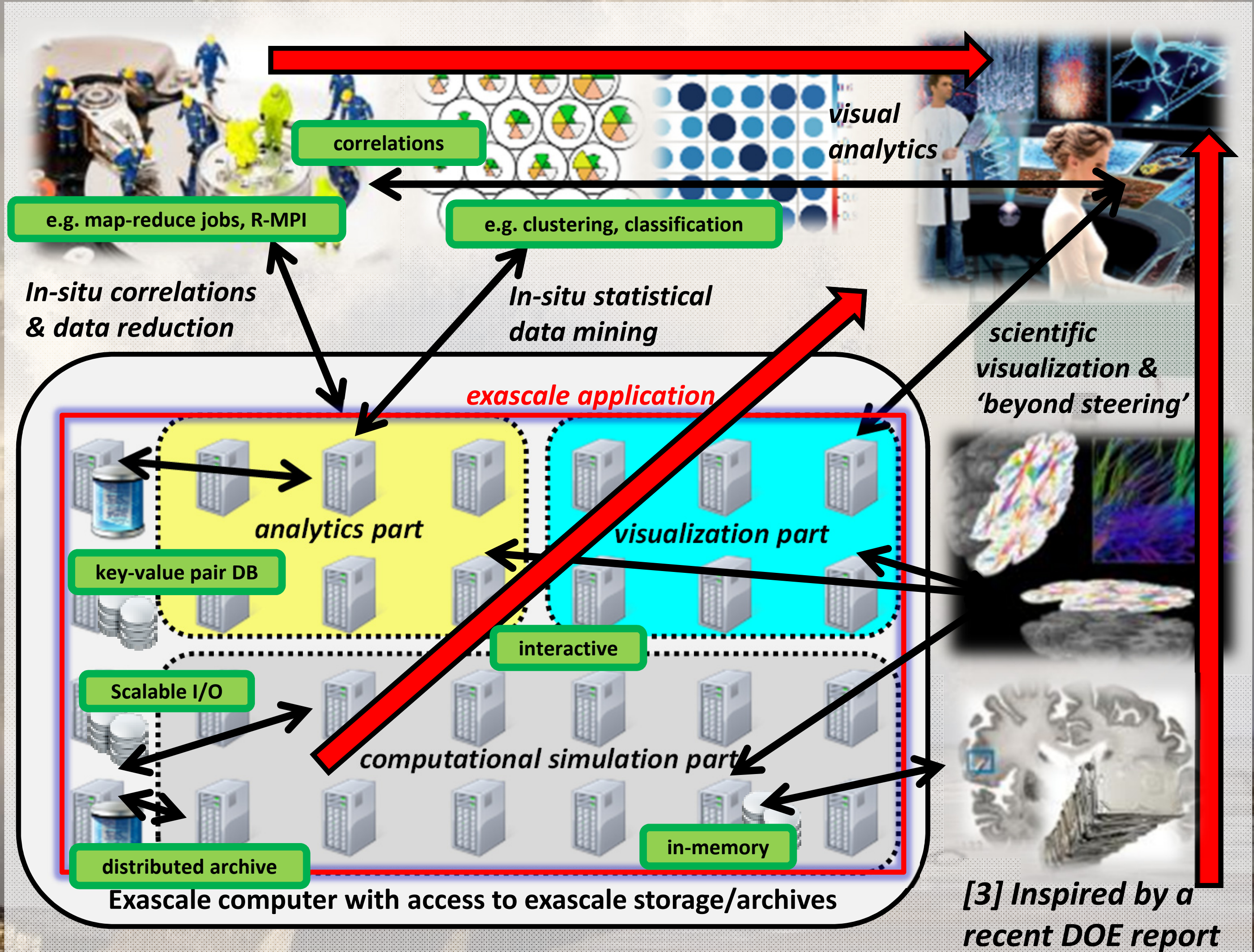
- *Hooks for offerings around commercial software packages*
- *Products around visualization packages and dedicated viewers*
- *Easy links to 'added value data', e.g. available market statistics*
- *Hosting services or deliver expandable storage in 'peek'*
- *Seamless links to the publishing industry*

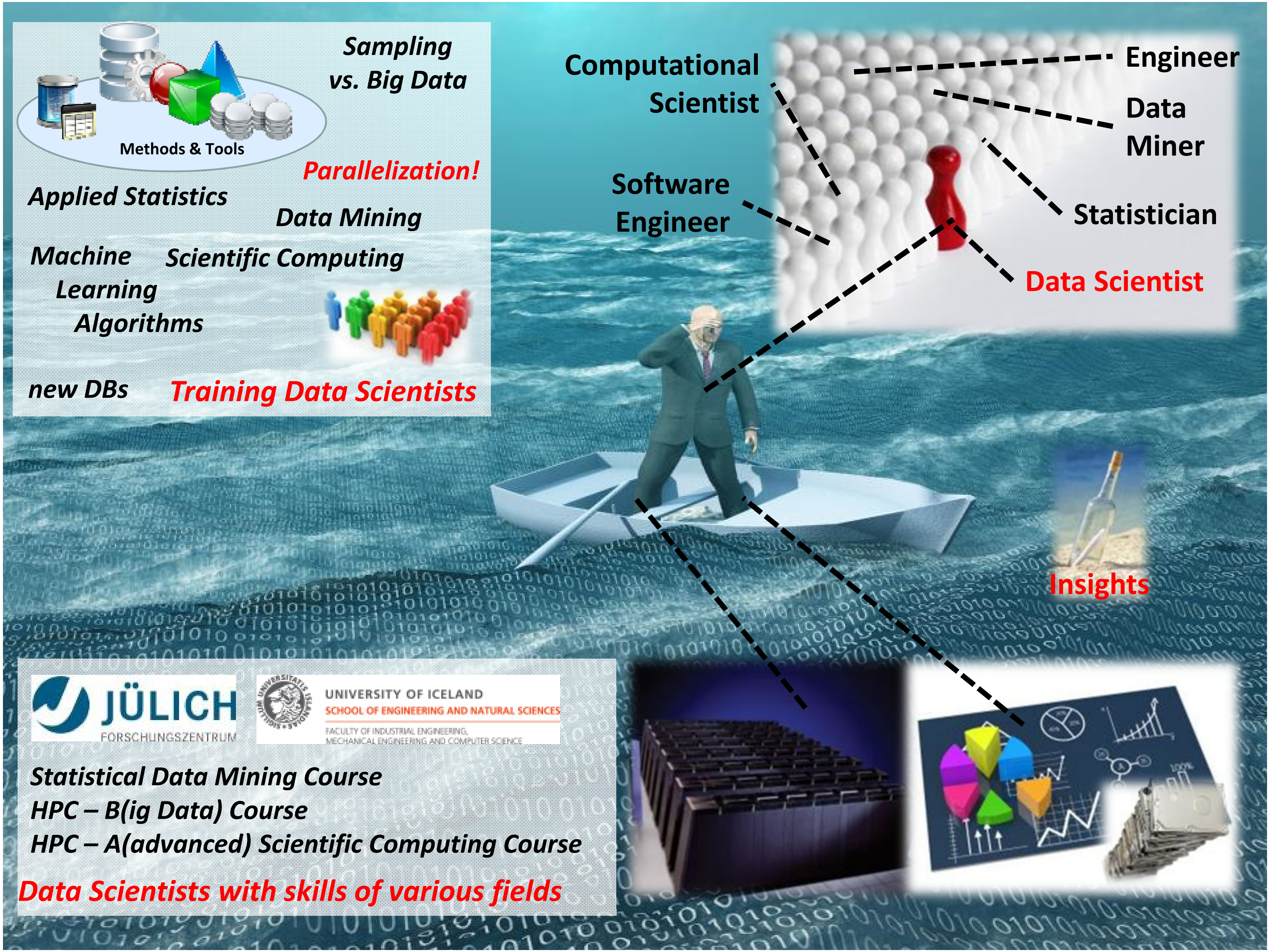
Data (or ScienceTube) to 'dive into data' with the possibility of commercial 'hooks'

[6] 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

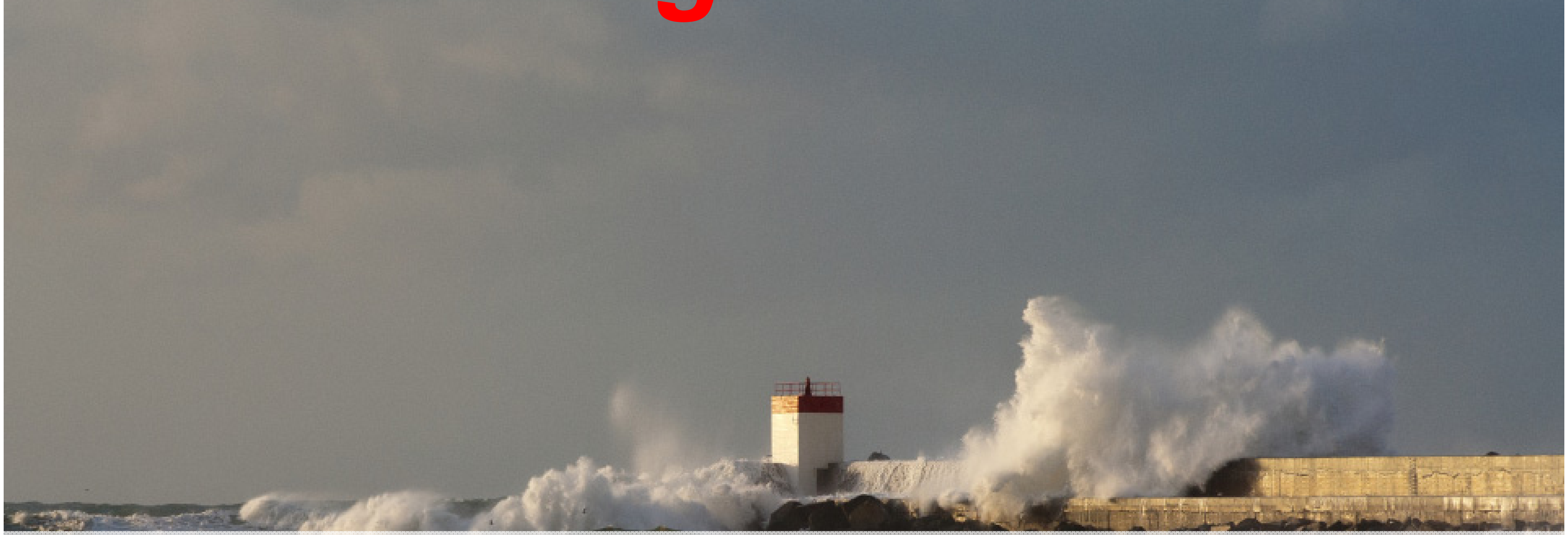
Towards Exascale: Applications with combined characteristics of simulations & analytics

'In-Situ Analytics'




Methods & Tools
Sampling
vs. Big Data
Parallelization!
Applied Statistics
Data Mining
Machine
Learning
Algorithms
Scientific Computing
new DBs
Training Data Scientists
Computational
Scientist
Software
Engineer
Engineer
Data
Miner
Statistician
Data Scientist
Insights
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
HPC – A(advanced) Scientific Computing Course
Data Scientists with skills of various fields

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

‘Takk’

Talk available at:

www.morrisriedel.de/talks

Contact:

m.riedel@fz-juelich.de

[1] Riding the Wave, EC Report, 2010
[2] A Surfboard for riding the wave, Report, 2012
[3] DoE ASCAC Report, 2013
[4] Jeremy Ginsburg et al., ‘Detecting influenza epidemics using search engine query data’, Nature 457, 2009
[5] David Lazer, Ryan Kennedy, Gary King & Alessandro Vespignani, ‘The Parable of Google Flu: Traps in Big Data Analysis’, Science Vol (343), 2014
[6] 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
[7] Prof. Rolf-Dieter Heuer, CERN Director General, in the context of the Higgs Boson Discovery
[8] Fran Berman, Maximising the Potential of Research Data
[9] Study on Hadoop, Accenture
[10] Research Data Alliance, Big Data Analytics IG, Online: https://rd-alliance.org/internal-groups/big-data-analytics-ig.html
[11] P. Chapman et al., CRISP-DM Guide
[12] G. Cavallaro and M. Riedel, ‘Smart Data Analytics Methods for Remote Sensing Applications‘, 35th Canadian Symposium on Remote Sensing (IGARSS), 2014, Quebec, Canada
[13] Open Grid Forum, Online: http://www.ogf.org
[14] UNICORE Technology, Online: http://www.unicore.eu
[15] Sun Z., and Fox G., ‘Study on Parallel SVM Based on MapReduce’, In Proceedings of the international conference on parallel and distributed processing techniques and applications, 2012.
[16] Apache Mahout Tutorial, YouTube Video