Big Data in Science
Overview of European & International Activities

Dr - Ing. Morris Riedel et al.
Adjunct Associated Professor, University of Iceland, Iceland
Jülich Supercomputing Centre, Germany
Head of Research Group High Productivity Data Processing

Research Field Key Technologies
Jülich Supercomputing Centre
Supercomputing & Big Data

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?

- 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

IMHO, it's great!
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

Data Sinks

n optional filters

Understanding ‘Big Data Waves’
Big Data Streams with ‘high velocity’…

… require interactive access & steering

Data Sources

- ‘Web Actions’
  - Data stream with data ‘hidden’ in logs/events
  - Changes of parameters for Web applications (e.g. video timeline click)

- ‘Measurement Devices’
  - Data stream with measured data
  - Steer measurement device parameters (e.g. change angles, resolution)

- ‘Computational Simulations’
  - Data stream with data of HPC/HTC simulation
  - Steer HPC simulation parameters on the fly (e.g. particle positions)

Data Sinks

n optional filters
‘Crowdsourcing’... 

...increases # of Big Data Streams

Usual Citizens / ‘Citizen Scientist’

Data streams with data (low trust)

Individuals with domain as Hobby

Data streams with data (moderate trust)

Scientific/Engineering Domain Experts

Data streams with data (high trust)

Terabytes

Petabytes

Exabytes
Infographics
Compact Combination of many Data Visualizations

Better understand trends across N data sources

Derived statistical data values with graphs, charts, percentages, ...

Enable comprehensive views on data

Data in context of locations or time correlated and/or cross-combined

Better understand trends across N data sources

Derived statistical data values with graphs, charts, percentages, ...

Enable comprehensive views on data

Data in context of locations or time correlated and/or cross-combined
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’

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

Industry 4.0

Energy

Smart Cities

Personalised Medicine
Big Data Waves & Massive Amounts of Technologies Exist
How to create real value from the rising tide of ‘Big Data’?

- Demo planned for upcoming German IT Summit Event
SDIL Industry 4.0
Explore Data-driven Parts of 4th 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
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
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
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’
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 the data from involved organizations is a key challenge (e.g. legal issues)
- Open upcoming omics-to-clinics meeting @ DKFZ

- 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**

**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
  - 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

Many statistical data mining methods exist – but less are openly available as ‘parallel’
Better Simulations...  
... means ‘Bigger Data’ &  
... needs smart preservation...
‘A landing-on-the-moon-style project for neuroscience’
‘Big Data’ meets Computational Science

**Smart Integration of simulation & experiment**

- ‘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

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

Slide material courtesy by Prof. Marquardt (modified and translated into English)
We need to ‘dive into data’

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


USA? Japan? China?
Understanding Possible Revenue Streams for Sustainability

**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 and HPC application industry
- Computing services to offer scalable data analytics

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

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

- 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

Research Data Sharing

**Without Barriers**

Harmonization, Definitions, Best Practices,....
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’
Using EUDAT B2SHARE with Persistent Identifiers enables trust to delete data on different platforms (effect multiplies: x Phd students x teaching class)
Earth Science Data Analytics Examples

Take Advantage of Interoperability…

…between EU PRACE & US XSEDE

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

- PhD studies Markus Götz

Problem: Quality control via outlier detection with PANGAEA data collection

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

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

Problem: Continuous seismic waveforms analysis for earthquakes monitoring

Problem: Event tracking analytics with spatial computing datasets (changing geolocations)
Shifts from Causality to Correlation
Challenging research with progress based on reason?

'A smart combination of both is needed'

Selected Lessons Learned

Traditional search for causality $\rightarrow$ (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’

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

‘Big Data is not always better data’

Selected Lessons Learned


Scientific Big Data Analytics

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

Example: support vector machines (learning algorithm)

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

- Algorithm A Implementation
- Algorithm Extension A’ Implementation
- Parallelization of Algorithm Extension A’ → A”

- Implementations available
- Implementations rare and/or not stable

- Classification++
- Regression++
- Clustering++

- Hadoop 1.0
- Hadoop 2.0
- NoSQL Databases
- GPGPU codes
- Google Big Table
- Graph-based approaches
- Clouds
- Active Storages
- HPC/HTC
- SQL
- Array databases
- Openstack
- Foursquare
- Twitter
- Facebook
- map-reduce
- Spark
- Triple stores
- HPC/HTC
- Active Storages
- NoSQL Databases
- Graph-based approaches
- Google Big Table
- GPGPU codes
- SQL
- Array databases
- Openstack
- Foursquare
- Twitter
- Facebook
- map-reduce
- Spark
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

<table>
<thead>
<tr>
<th>Field</th>
<th>Institute</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polar and Marine Research</td>
<td>AWI</td>
<td></td>
</tr>
<tr>
<td>Material Sciences</td>
<td>DESY</td>
<td>different data sources to integrate in analysis</td>
</tr>
<tr>
<td>Biomedical data</td>
<td>DKFZ</td>
<td></td>
</tr>
<tr>
<td>Climate</td>
<td>DKRZ/HZG</td>
<td>different formats</td>
</tr>
<tr>
<td>Earth Observation</td>
<td>DLR</td>
<td></td>
</tr>
<tr>
<td>Epidemiology</td>
<td>DZNE</td>
<td>Various technologies</td>
</tr>
<tr>
<td>Biomolecular research</td>
<td>JUELICH</td>
<td></td>
</tr>
<tr>
<td>FAIR data</td>
<td>GSI</td>
<td>Sharing &amp; reproducability</td>
</tr>
<tr>
<td>Environmental caused illness</td>
<td>HMGU</td>
<td></td>
</tr>
<tr>
<td>Photon / Neutron Research</td>
<td>HZB</td>
<td>3D visualization &amp; steering</td>
</tr>
<tr>
<td>Laser and magnetic fields research</td>
<td>HZDR</td>
<td></td>
</tr>
<tr>
<td>Astro physics</td>
<td>KIT</td>
<td>Smart analytics &amp; analysis</td>
</tr>
<tr>
<td>Research on water &amp; geo data</td>
<td>UFZ</td>
<td></td>
</tr>
</tbody>
</table>
Lessons Learned from ‘Big Data Analytics’ to ‘Smart Data Analytics’

‘Scientific Big Data Analytics’ needs Steering by Provisioning

- Scientific Big Data Analytics (SBDA)
- SimLabs
- DataLabs
- SimLabs
- DataLabs
- Communities & Research Groups
- Grand Challenges of Society and Science
- Industry
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’

Exascale computer with access to exascale storage/archives

In-situ correlations & data reduction

In-situ statistical data mining

Analytics part

Visualization part

Computational simulation part

key-value pair DB

Scalable I/O

distributed archive

In-memory

visual analytics

analytics part

interative

scientific visualization & ‘beyond steering’

exascale application

Inspired by a recent DOE report

e.g. dimensionality reductions

e.g. map-reduce jobs, R-MPI

e.g. ultra-dimensional scaling

e.g. clustering, classification

e.g. multi-dimensional scaling

dimensionality reductions

visual analytics

data reduction

correlations

In-situ statistical data mining

scientific visualization & ‘beyond steering’
‘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