INFOCOMP 2015 International Expert Panel:

Emerging Solutions in Scientific and High End Computing: Coping with Challenges and Requirements on the Long-term

June 23, 2015, Brussels, Belgium

The Fifth International Conference on Advanced Communications and Computation (INFOCOMP 2015)

INFOCOMP
June 21–26, 2015 - Brussels, Belgium

INFOCOMP Expert Panel: Emerging Solutions in Scientific and HEC

Panelists

- *Claus-Peter Rückemann* (Moderator),
  Westfälische Wilhelms-Universität Münster (WWU) /
  Leibniz Universität Hannover /
  North-German Supercomputing Alliance (HLRN), Germany
- *Isabel Schwerdtfeger,*
  IBM, Germany
- *Małgorzata Pankowska,*
  Department of Informatics, University of Economics in Katowice, Poland
- *Lena Noack,*
  Royal Observatory of Belgium, Belgium

INFOCOMP 2015: `http://www.iaria.org/conferences2015/INFOCOMP15.html`
Program: `http://www.iaria.org/conferences2015/ProgramINFOCOMP15.html`

INFOCOMP Expert Panel: Emerging Solutions in Scientific and HEC

Panel Statements:

- **Big Data:** Future solutions need to consider new advanced methods (NoSQL, mind mapping, . . .).
- **Reduce data size:** Long-term relevant data size should be reduced without loosing essential content and context.
- **Knowledge:** Knowledge resources can essentially benefit from adding conceptual knowledge, classification, . . .
- **Automation:** Big Data, Volume, Variability, Velocity, Vitality, Veracity, . . . require advanced documentation.
- **High End:** Limits of bandwidth and latency regarding transfer and storage (much more than computing).
- **Value:** Structure preceeds computation for long-term data.
- **Standards:** There are many standards. It should become reasonable to integrate standards with reasonable, reusable, portable, and commonly available technologies and methods.
- **Resources:** Management complexity from planning to operation (hardware and software) must be reduced for improving applicability.

INFOCOMP Expert Panel: Emerging Solutions in Scientific and HEC

Pre-Discussion-Wrapup:

- **Focus:** Data organisation or computing and algorithms?
- **Recommendations:** Which general long-term solutions and recommendations?
- **How-to:** How can sustainable big data solutions be created?
- **Sizes:** How can data sizes be reduced without loosing essential information?
- **Approaches:** Experiences and results?
- **Long-term:** How long do we expect data/solutions to be consistent/work?
- **Context:** Are there differences in national and international context?
- **Dissemination:** What is the significance of "research and publish"?
- **Sustainability:** Multi-disciplinary and long-term perspectives?
- **Networking:** Discussion! Open Questions?
  Suggestions for next Expert Panel?

INFOCOMP: Post-Panel-Discussion Summary

Post-Panel-Discussion Summary (2015-06-23):

- Future solutions should consider advanced methodologies and new advanced methods (NoSQL, mind mapping, ...).
- Large amounts of data may be required to be available for long periods of time.
- "Sizes" of long-term relevant data should be reduced, esp. by the originators, without loosing essential content and context. Along with best practice accompanying funding, long-term storage should become available.
- Big data clouds can provide high end solutions in many cases, in addition.
- Structure preceeds computation for long-term data and value.
- Sustainability will significantly benefit from advanced data organisation and adding conceptual knowledge, classification, ...
- The significance of "research and publish" content as well as business application scenarios is continuously increasing.
- Most content-centric, technical, coding/code compatibility, and legal challenges need to be addressed internationally and multi-disciplinary, on long-term.
- Management complexity from planning to operation (hardware and software) must be reduced for improving applicability.
- A major future object is the integration of standards with reasonable, reusable, portable, and commonly available technologies and methods.
- Further common consens: Sustainable long-term funding and investments needed!

INFOCOMP Expert Panel: Table of Presentations, Attached

Panelist Presentations: (presentation order, following pages)

- **Solutions to Long-term Challenges: Resources of Knowledge and Computation** (*Rückemann*)
- **High End Computing and Big Data Challenges: Big Data Cloud Solutions** (*Schwerdtfeger*)
- **NoSQL as Emerging Solution in Business Computing** (*Pankowska*)
- **Sustainable data management in science** (*Noack*)

INFOCOMP 2015 International Expert Panel:
Emerging Solutions in Scientific and High End Computing:
Coping with Challenges and Requirements on the Long-term

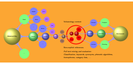
Solutions to Long-term Challenges: Resources of Knowledge and Computation

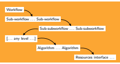
The Fifth International Conference on Advanced Communications and Computation (INFOCOMP 2015)
June 23, 2015, Brussels, Belgium

Dr. rer. nat. Claus-Peter Rückemann[1,2,3]

[1] Westfälische Wilhelms-Universität Münster (WWU), Münster, Germany
[2] Leibniz Universität Hannover, Hannover, Germany
[3] North-German Supercomputing Alliance (HLRN), Germany

ruckema(at)uni-muenster.de





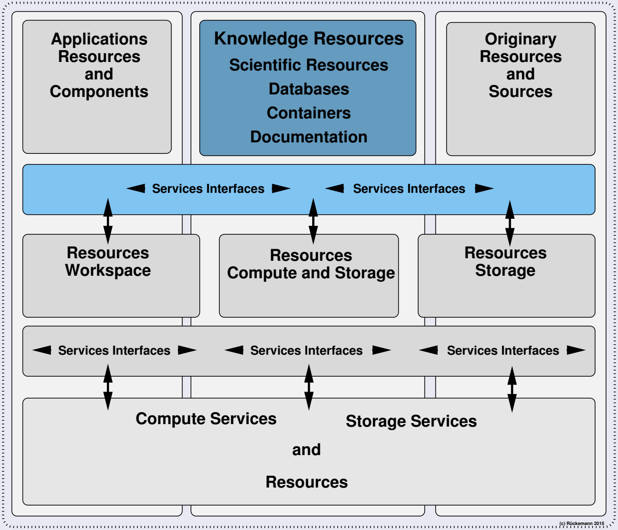
Challenges regarding content and scenarios

- **Content and applications:** Natural sciences/fundamental research, applied sciences/practical applications,
  ... are not long-term integrated in theory and practice.
- **Monolithic architectures:** System components require continuous re-development.
- **Big Data:** Classical methods (e.g., relational and object oriented) can hardly provide universaly efficient solutions.
- **Data size:** For decades, disk speeds and sizes do not keep up pace with data generation.
- **Knowledge:** Content and context are not appropriatley documented for decades.
- **Automation:** Instructive documentation is not available.
- **Transfer and storage:** Limits of bandwidth and latency.
- **Value:** The value of data is steadily increasing.
- **Standards:** Standards for integrating standard are not available.
- **Resources:** Increasing amounts of money are spent on increasingly complex-to-manage high end hardware and software.

Missing and emerging solutions regarding content and scenarios

- **Content and applications:** Frameworks for long-term integration of fundamental research, content, practical applications.
- **Monolithic architectures:** Re-use design and implementation of components.
- **Big Data:** Advanced methods and new algorithms (e.g., NoSQL).
- **Data size:** 1: Increase of reliable and cheap disk/storage speeds and sizes; 2: Reduce data sizes.
- **Knowledge:** Knowledge-based documentation of content and context (e.g., knowledge resources).
- **Automation:** Instructive documentation for knowledge.
- **Transfer and storage:** Significantly (on-demand) increase bandwidths, decrease latencies.
- **Value:** Support the value of data and knowledge with best practice and funding (creation, documentation, computing, storage, integration, ...).
- **Standards:** Modularise standards' integration, support long-term standards.
- **Resources:** Modularise complex-to-manage high end HW and SW, empower users to handle technology, reduce costs of lifecycles and energy consumption.
- **Measurements and means:** Knowledge resources, e.g., long-term research data management/libraries.

Example Framework – Disciplines, Services, Providers



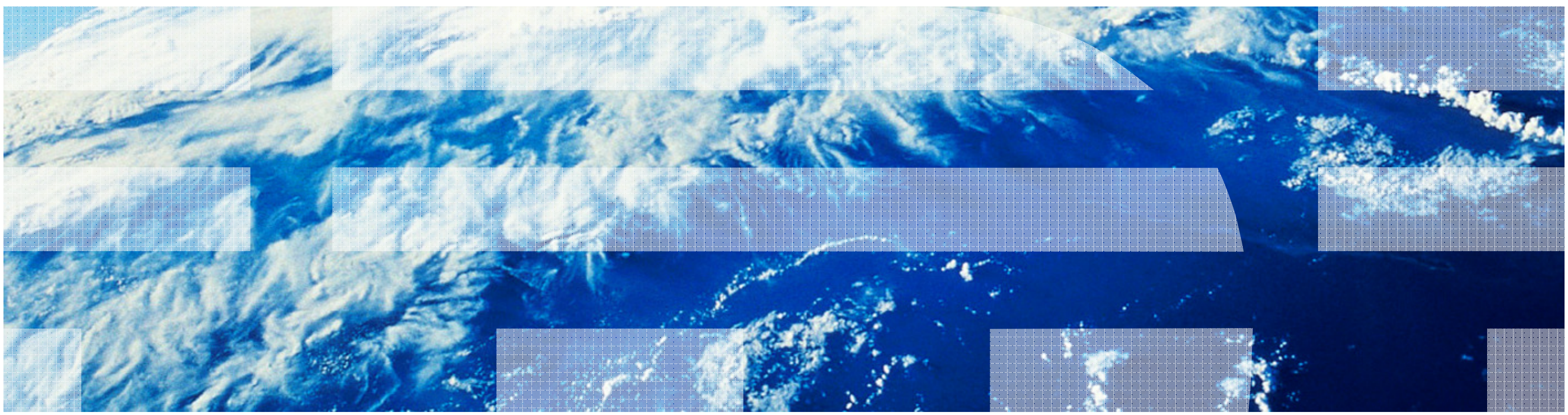
Integration & development of long-term knowledge & measurements

- Solutions, which can be integrated.
- Improved data organisation, long-term data, structures, means.
- Knowledge documentation, content / context vitality.
- Creation of standards/systematics/methodologies with content.
- Long-term sustainability of universal knowledge discovery.
- Multi- and trans-disciplinary work.
- Support High End Computing, intelligent systems, education.
- Integrated Information and Computing System components.
- Mandatory best practice (e.g., for participation and funding).

Isabel Schwerdtfeger
Leading Solution Sales Professional
HPC & HPSS Services Sales – Global Technology Services
schwerdtfeger@de.ibm.com

Panel:
Emerging Solutions in Scientific and High End Computing: Coping with Challenges and Requirements on the Long-term

Brussels, June 23, 2015



INFOCOMP 2015 International Expert Panel

LRZ IBM-Lenovo SuperMUC Phase 2 – 3,2 Petaflops System

Source: T. Bloth, SuperMUC 2 Installation 2015, Lenovo

Emerging Solutions in Scientific and High End Computing: Coping with Challenges and Requirements on the Long-term.

- Emerging Solution: „Big Data Cloud Solutions“

- Challenge:
  - Huge amount of data management
  - Prevent data loss and ensuring data integrity
  - Achieve „acceptable“ performance in Gigabytes per second for read/write
  - Ensure long-term availability including the „rights to delete“

- Investment:
  - Test data centers where to test the capabilities with existing huge data volumes
  - Prove stability for multiple applications use-cases, i.e., video data, small files, etc.

Meet us at ISC‘15, Frankfurt Germany!

International Supercomputing Conference (ISC) 2015
July 12 - 16, 2015, Frankfurt am Main, Germany

ISC High Performance

Silver Sponsor

BOOTH #928

July 12 -16, 2015 I Frankfurt, Germany

Optimising Data-Centric IT Environments

Accelerate time to insights for HPC and analytics apps

- Available on site IBM HPC Executives and Development from IBM Corp. for dedicated customer briefings for
  - HPC Strategies, HPC Storage, HPSS, Life Sciences, IBM Reference Client DESY on stage

- IBM Booth with 5 Demos and IBM Hardware to view at the booth
  - SDI, Cloud, OpenPower
  - Analytics/Watson – tranSMART LiveDemo
  - IBM Lab Boeblingen with LiveDemos

- Networking / Industry View Update / HPC Trends & Directions

Thank you!

Isabel Schwerdtfeger

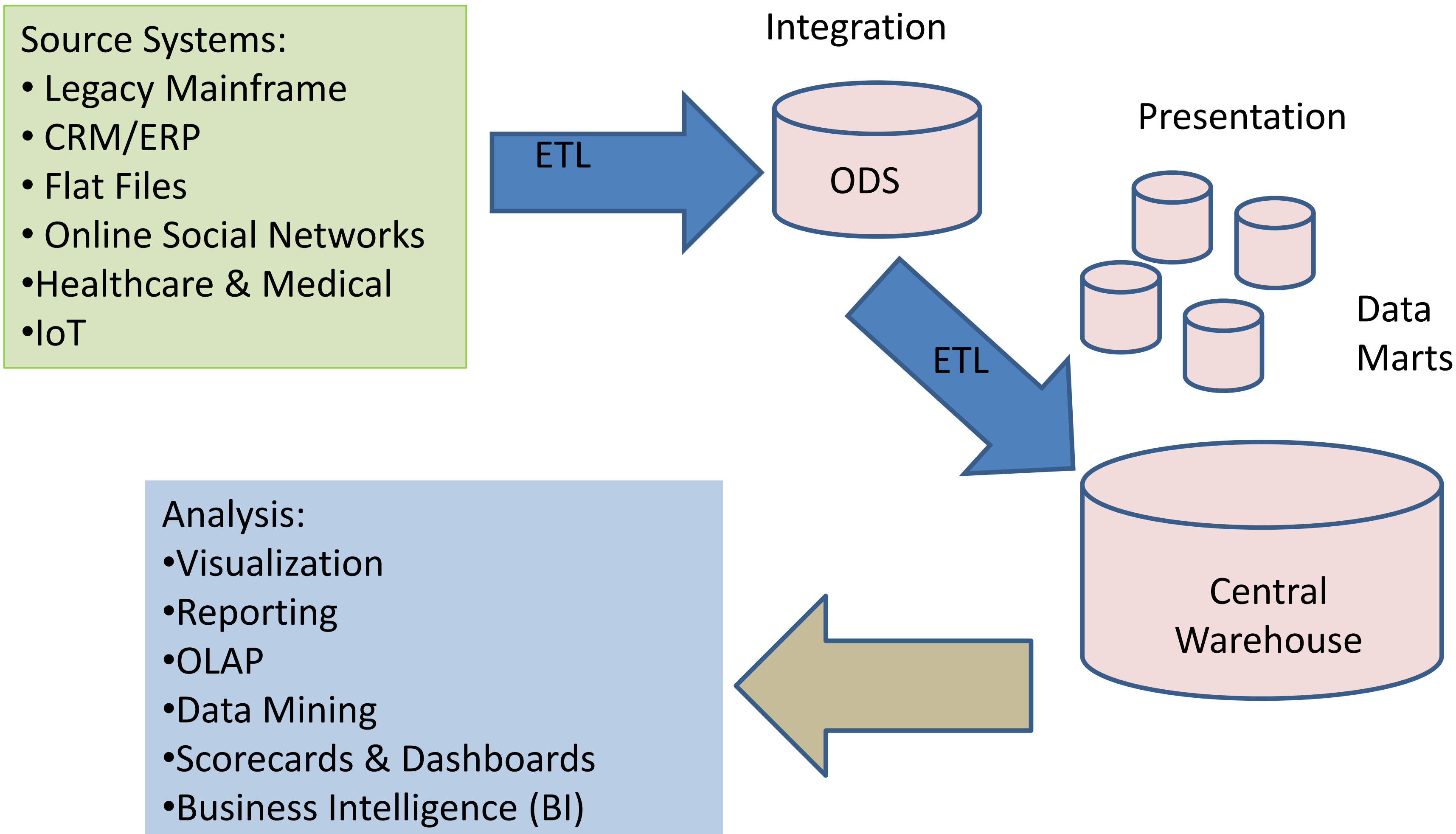
Leading Solution Sales Representative
HPC & HPSS Sales Leader
System Services

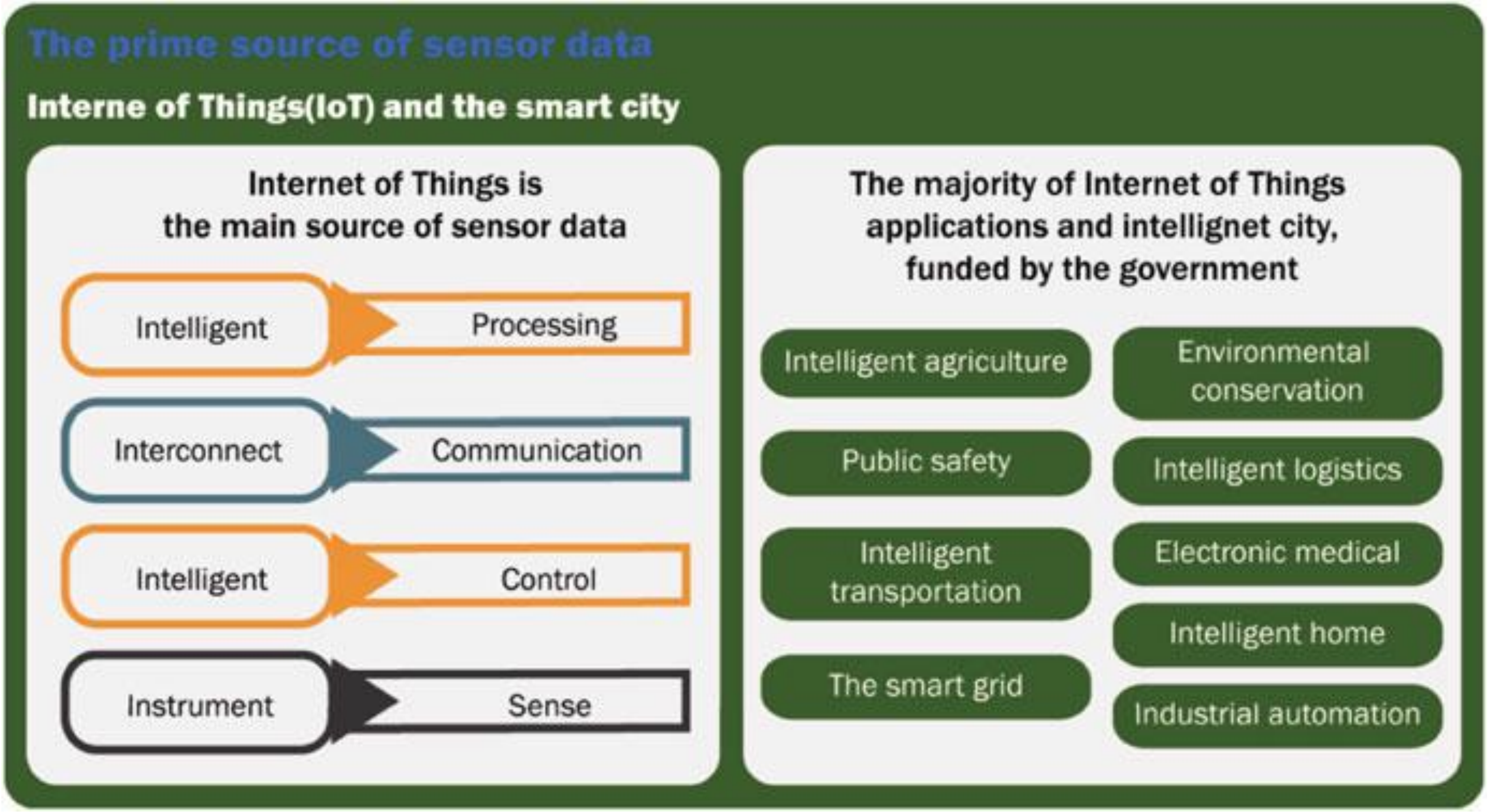
IBM Deutschland GmbH
Mobil: +49 170 635 7251
schwerdtfeger@de.ibm.com

NoSQL
as Emerging Solution
in Business Computing

[Sawant &Shah, 2013]

Data Warehouse Architecture





Chen et al. 2014

Internet of Things:

large-scale data, heterogeneity, strong time and space correlation, great quantity of noises during the data acquisition,

Data Management Architecture

Traditional BI Tools

Operational Data Store

Data Warehouse

Relational Databases

- Traditional Data
- structured
- terabytes
- centralized
- known relationships among data
- ACID transactions (Atomic, Consistent, Isolated, Durable)

Big Data Analysis Tools

Data Lakes

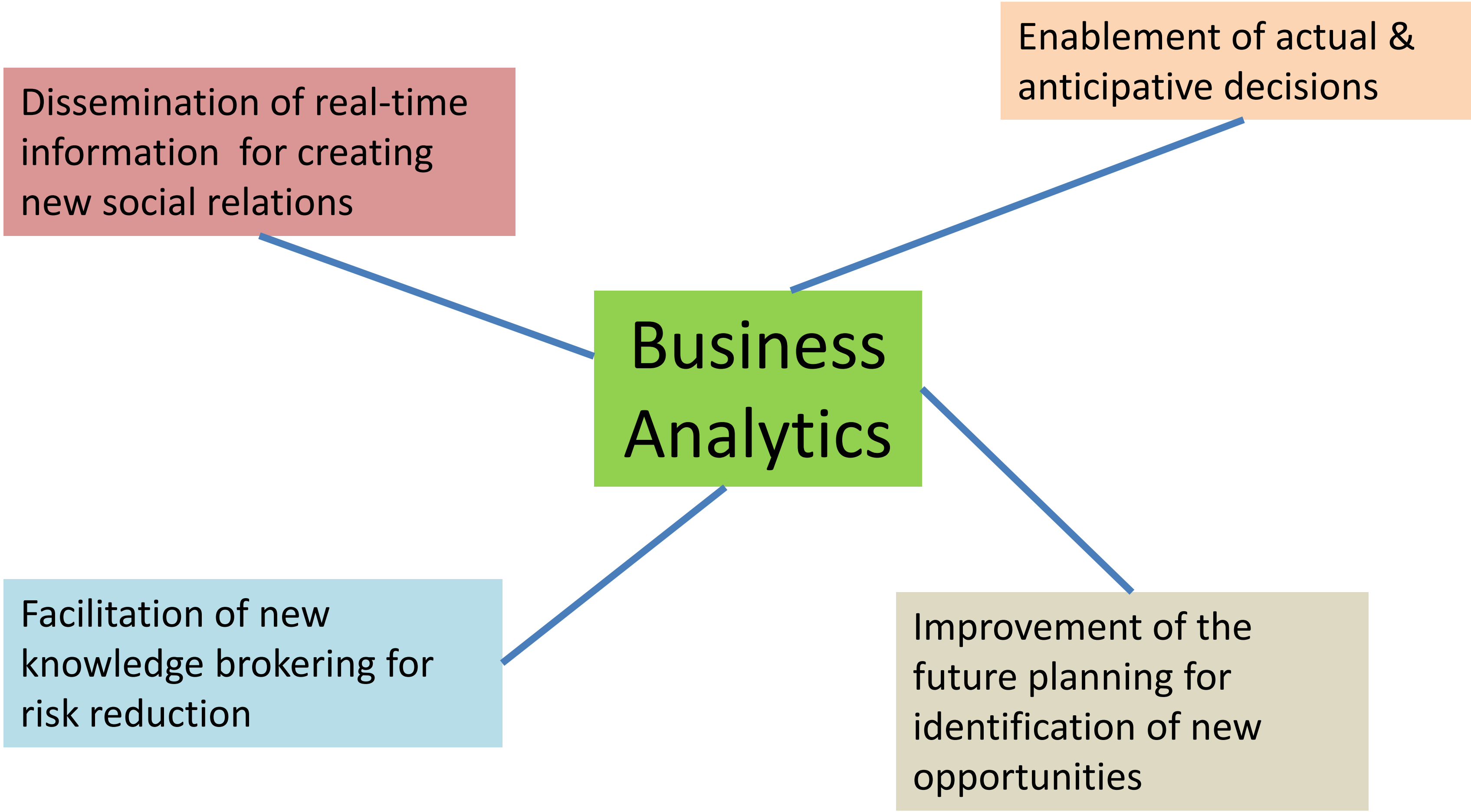
NoSQL Databases

- Big Data
- unstructured
- petabytes & exabytes
- distributed
- complex relationships among data
- open source
- developed for web application
- database sharding & replication

NoSQL use cases

NoSQL Database	Use case
Graph Database store data entities and connections between them as nodes and edges. They are similar to a network database	Network Modelling Locality Recommendations: Applications that provide evaluation of „like” or note that „user that bought this item also bought ,” like a recommendation engine
Key-Value Pair Database store data as simple key-value pairs. They are suitable for parallel lookups, the data sources have no relationships among each other	Needle-in-a-haystack applications. Shopping Carts analyses, Web User Data Analysis (Amazon, LinkedIn)
Document Database store text, media, and JSON or XML data.	Real-Time Analytics Logging, Document Archive Management . If you want to search through multiple documents for a specific strings, a document database should be used.
Column-oriented Database have a huge number of columns for each tuple. Each column has a column key. Tuples can have different columns	Analyzing of the Huge Web, User Actions and Sensor Feeds (Facebook, Twitter) Google search type of applications, where en entire related columnar family needs to be retrieved based on a string

Business Analytics

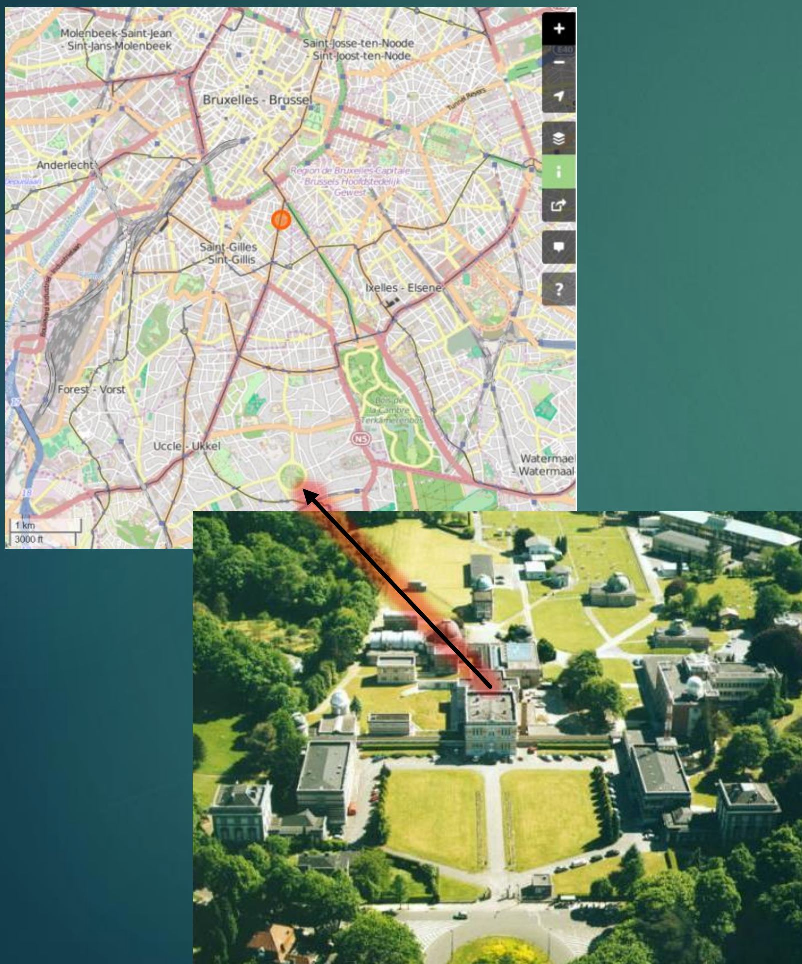


[Lake & Crowther, 2013]

Sustainable data management in science

DR. LENA NOACK

Royal Observatory of Belgium



- Three different institutes:
  - Royal Observatory of Belgium
  - Royal Meteorological Institute
  - Belgian Institute for Space Aeronomy
- Joint IT services, cluster, and server

Dr Lena Noack

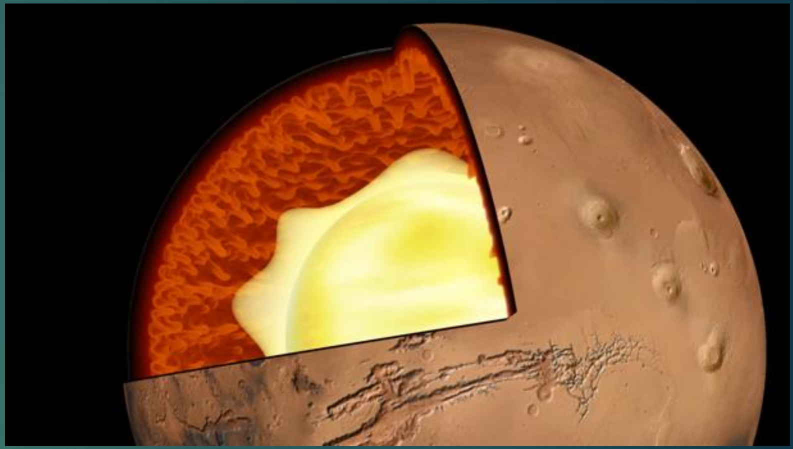
Post-Doc scientist

Royal Observatory of Belgium

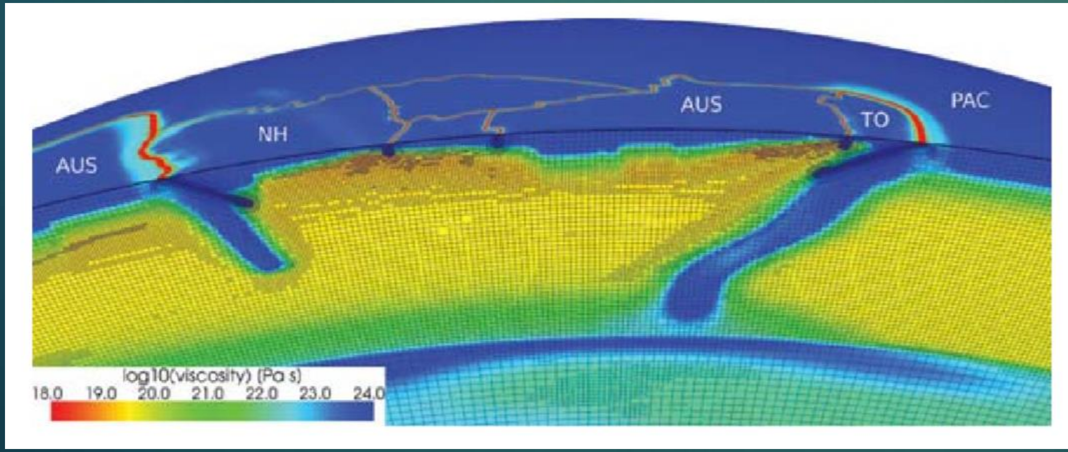
Lena.Noack@oma.be

Data storage policy in science

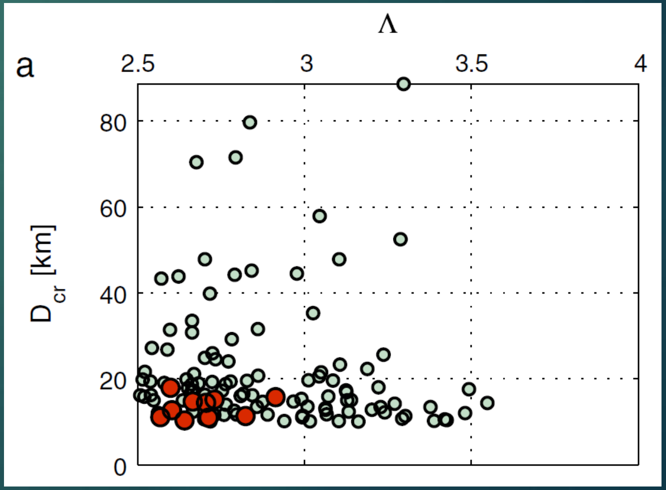
- Published data of any kind need to be stored for typically at least 10 years (depending on guide-lines of publisher and/or institute)
- Scientific data and results need to be reproducible (even after 100 years?)



© C. Hüttig and M. Krüger, DLR



[Stadler et al., 2010]



[Tosi et al., 2013]

From a scientists' point of view:

« Store as much as possible, and as long as possible, you never know when you might need it again »

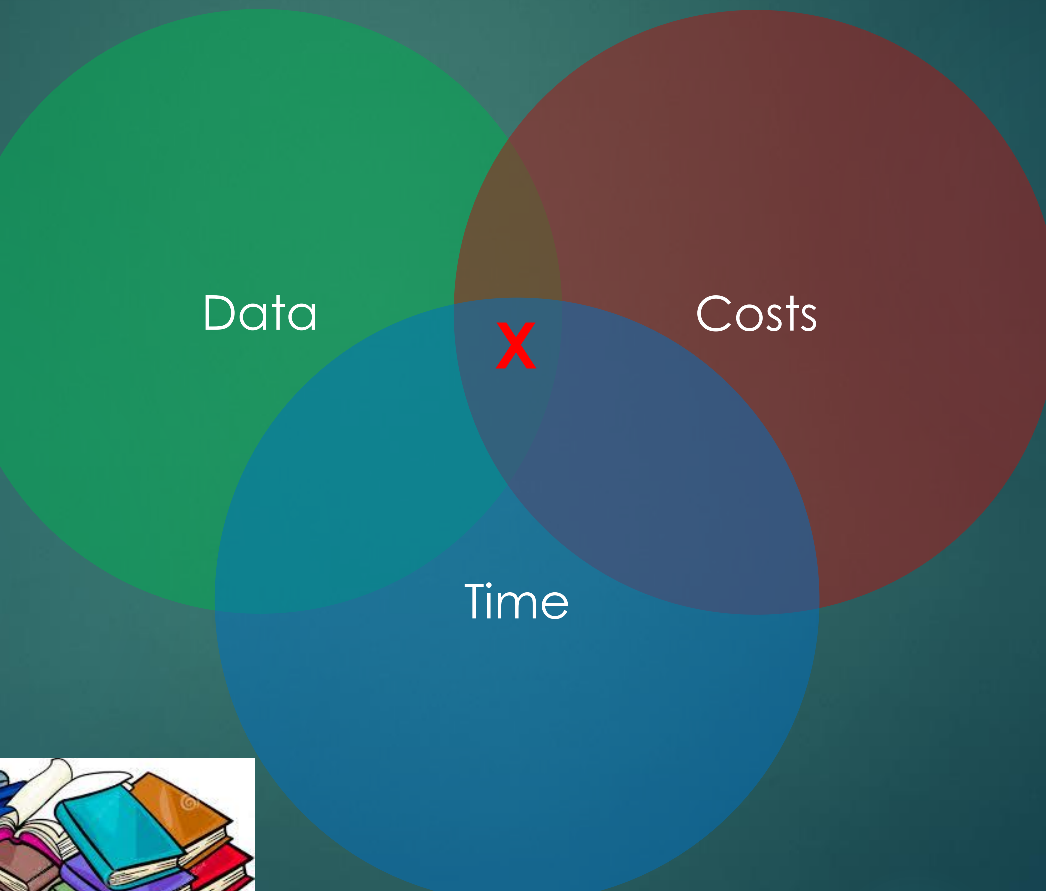
- Follow-up studies (even several years later)
- Proof of correct data (need all data to show that published results are correct)
- Store ALL simulation files, if possible (not only the final result of a simulation, or visualizations of the results, but all data written by the simulation), results are available also after 100 years, when the original code couldn't be used anymore (e.g. if compilers go extinct)

Management's point of view

« Store as few data as possible, but everything that is important, and for at least 10 years »

- Long-term storage (including backups) is expensive
- Store only what is necessary under publisher's agreement (standard: 10 years for all files needed to reproduce results)
- Simulation data can easily reach hundreds of TB and more – depending on the code

Different opinion on data storage policies



Questions

- Should all published data be preserved? And how long?
- Instead, resources could be used for improved HPC system -> faster simulations mean easy re-production of old simulation results
- How to ensure that an old code is compiling/ running on modern systems?