



Big Data
Analytics
Interest
Group



RESEARCH DATA ALLIANCE

Welcome and Shaping Landscape of Big Data Analytics

Morris Riedel, Juelich Supercomputing Centre, Germany

Rahul Ramachandran, Information Tech. and Systems Center, UoAlabama in Huntsville

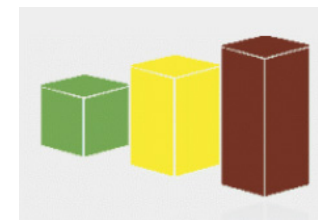
Co-Chairs RDA Big Data Analytics Group

Building Global Partnerships - RDA Second Plenary Meeting



Date: 16/09/2013 to 18/09/2013

16-18 September 2013, National Academy of Sciences, Washington DC, US



[3] RDA Web Page

- 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

■ Big Data Analytics IG

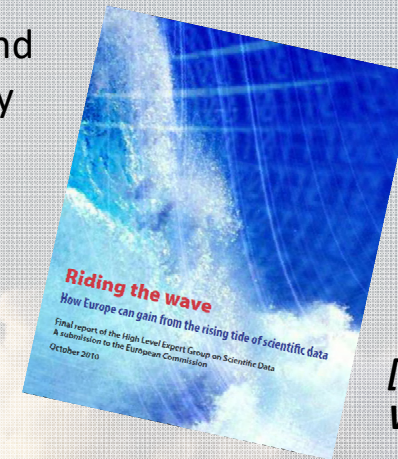
- Develops community based recommendations on feasible data analytics approaches to address scientific community needs of utilizing large quantities of data.
- Analyzes different scientific domain applications and their potential use of various big data analytics techniques.
- A systematic classification of feasible combinations of analysis algorithms, analytical tools, data and resource characteristics and scientific queries will be covered in these recommendations.

Analytics are Needed in Big Data-driven Scientific Research

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.’

‘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.’



[1] 'Riding the Wave' Report



[2] 'A Surfboard for Riding the Wave' Report

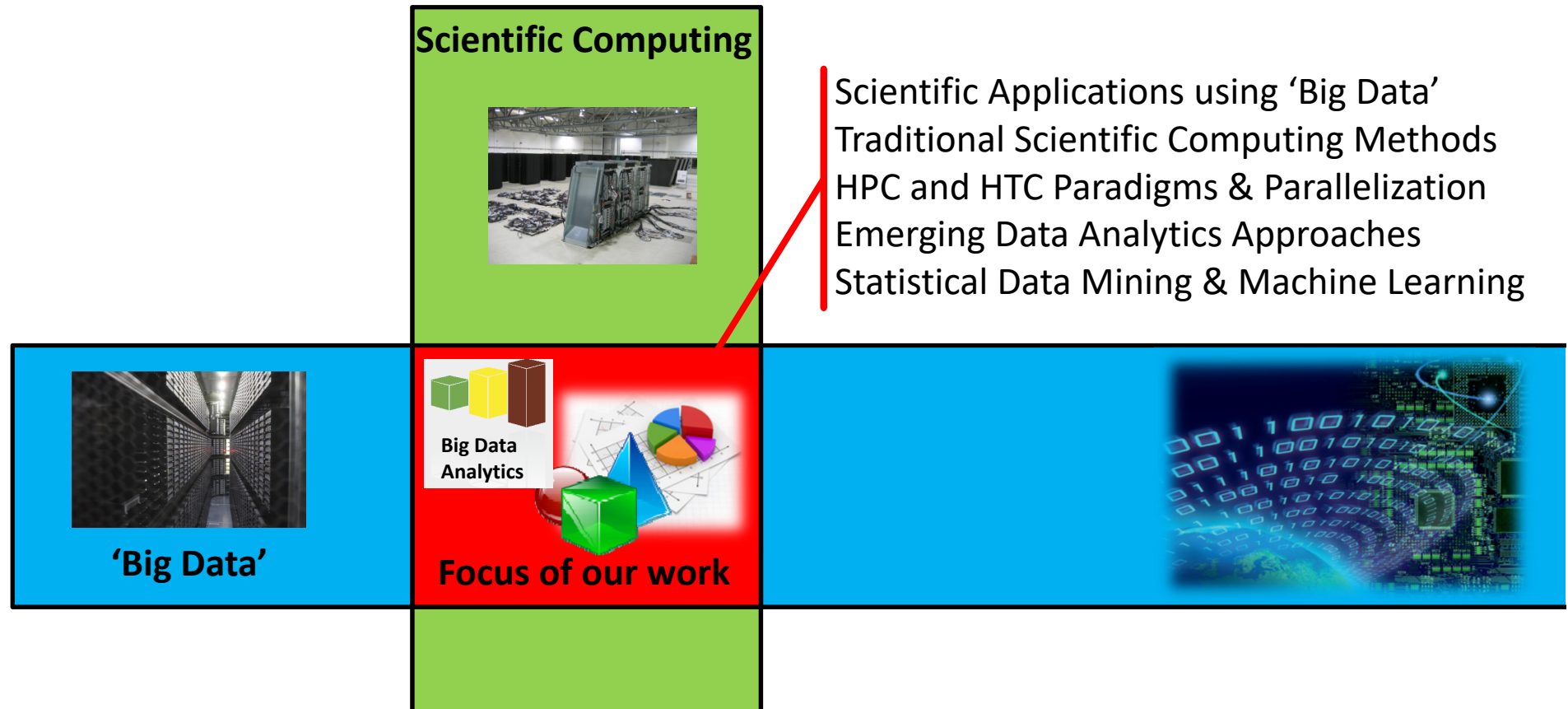


How do we enable ,high productivity processing'?
How do we find ,a message in the bottle'?



Work on Intersection of two Broad Subjects

- ‘Lighthouse goal’: **High Productivity Processing of Research Data**



Short Description & Roughly Goals Reviewed

- Develops community based recommendations ...

- ... on feasible data analytics approaches
- ... to address scientific community needs
- ... of utilizing large quantities of data.



- Analyzes different scientific domain applications

- and their potential use of various big data analytics techniques.



- A systematic classification of feasible combinations

- ... of analysis algorithms, analytical tools, ...
- ... data and resource characteristics ...
- ... and scientific queries will be covered in these recommendations.





Current Charter – Objectives

Available on wiki & submitted for approval

Objectives

The *Big Data Analytics (BDA) Interest Group* seeks to develop community based recommendations on feasible data analytics approaches to address scientific community needs of utilizing large quantities of data.

- BDA will aim to clarify some foundational terminologies in the context of data analytics understanding differences/overlaps with terms like data analysis, data mining, etc.
- BDA will systematically analyze different current scientific domain data analytics needs and their potential use of various big data analytics techniques.
- BDA will develop a recommendation documents with a systematic classification of feasible combinations of analysis algorithms, analytical tools, data and resource characteristics and scientific queries. These recommendation documents can serve as a best practice guide for scientific groups/communities interested in investing in Big Data technologies
- BDA will work towards a consensus amongst its members to achieve this desired goal



Current Charter – Participation

Available on wiki & submitted for approval

Participation

BDA is open to all RDA members to participate. BDA welcomes individuals with the following expertise to actively participate in its activities

- Domain science experts grappling with Big Data as part of their scientific applications
- HPC and distributed computing experts with middleware experience such as Hadoop
- Analytics experts with tools/algorithmic expertise, including data mining and machine learning approaches
- Managers seeks to deploy Big Data solutions at their organizations



Current Charter – Outcomes

Available on wiki & submitted for approval

Outcomes

BDA will be considered a success if the interest group:

- Develops recommendation documents based on consensus amongst its members
- Engages not only its members but also others in RDA and other organizations who can contribute to this endeavour.
- Develops recommendation documents in a meaningful timeframe



Current Charter – Mechanism

Available on wiki & submitted for approval

Mechanism

BDA will utilize capabilities provided by RDA platform to effectively communicate and collaborate. These include:

- Monthly telecoms/webex to with planned agenda to discuss specific issues
- Asynchronous collaboration using google docs, wiki and email list serves

All BDA documents will be publicly available. In addition, BDA will utilize RDA meetings to hold sessions to allow F2F interactions amongst members and to inform other RDA members of its ongoing activities.



Current Charter – Tentative Timeline

Available on wiki & submitted for approval

Timeline

A tentative timeline is given below but is subject to change.

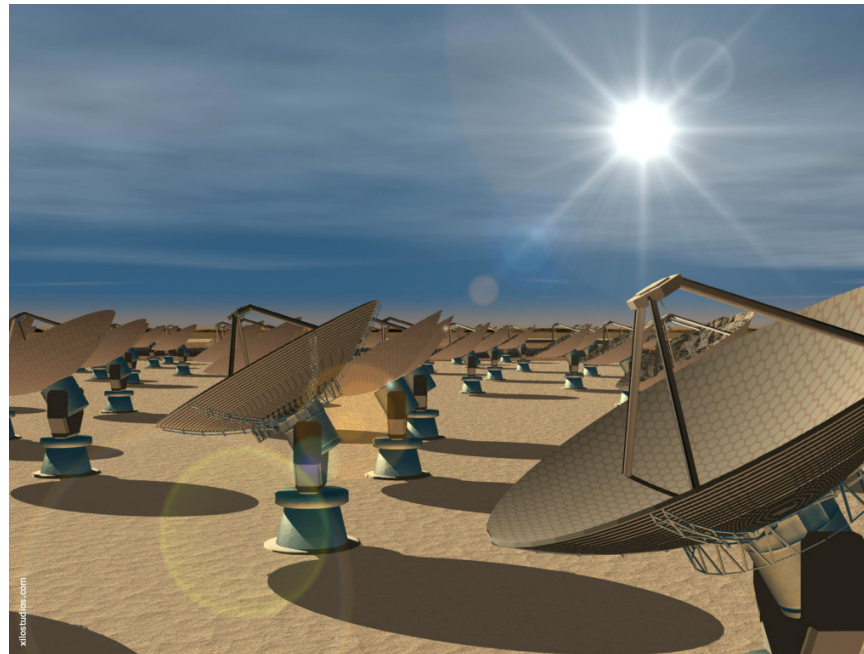
- Terminology definition
- Science Use Cases
- Use Case Template
- Mapping Science Use Cases to the Template
- Technology Lists [**9/13 - 2nd RDA Plenary**]
- Technology Template
- Mapping Technologies to the template
- Mapping Uses cases to Technology Solutions

Recommendation document [**3/14 - 3rd RDA Plenary**]

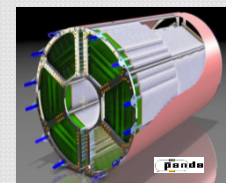
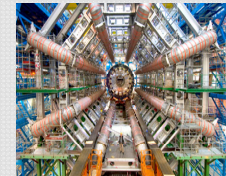
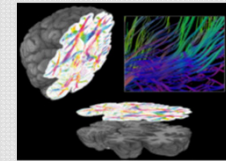
Problems with Terminologies

In Commercial environments the term 'Big Data' is often related to Volume – Variety – Velocity, but concrete 'numbers' are rarely given

*What do
we mean
by 'big data'?*



In Science environments the term 'Big Data' is often related to one concrete scientific experiment: e.g. square kilometre array → 1 PB / 20 seconds

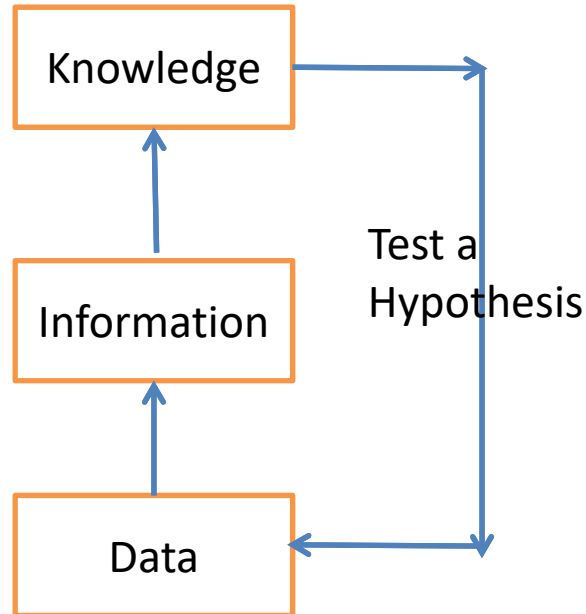




Do Definitions really matter here so much?

- Are we asking the right questions?
- Instead of asking:
 - What is the definition for “Big Data Analytics”?
- Should be asking:
 - Does it change the way we do science?
 - No – then there is nothing of value here
 - Yes – then what is it enabling?

Traditional Analysis Process

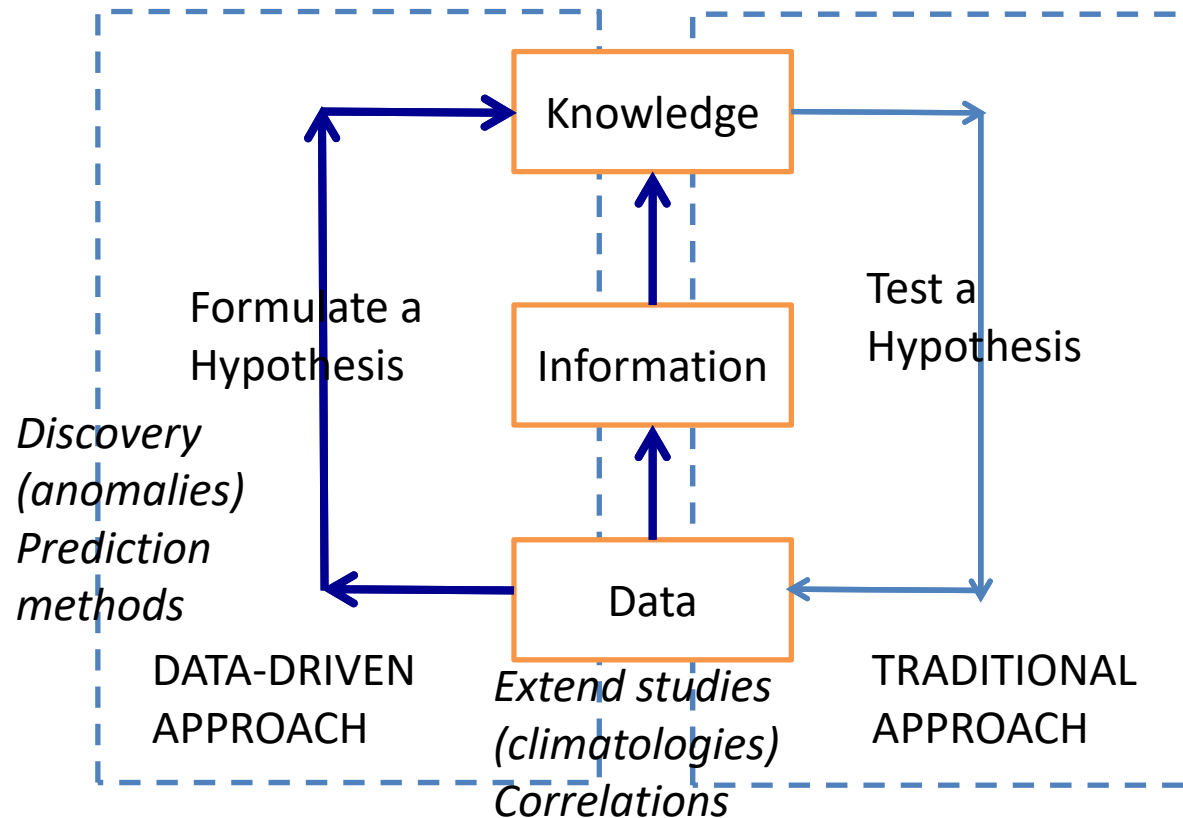


“data by itself cannot formulate a hypothesis, rather it changes the odds in favor of or against a hypothesis”

- Data – something that is directly observable therefore measurable
- Knowledge – statement about a hypothesis and testing of a hypothesis is done using data
- Information – measure of uncertainty about an hypothesis and the role of the data is to change the amount of information (increase/decrease entropy)

N. D. Singpurwalla, “Knowledge Management and Information Superiority (a taxonomy),” Journal of Statistical Planning and Inference, vol. 115, no. 2, pp. 361–364, 2003.

Big Data Analytics: New Pathways



Previous statement on the use of data is overly restrictive!

- Data can be used to discover new anomalies that would require testing
- Data can also be used extend studies limited on their use of data sets

How do we support these new pathways?

- Technology
- Business Model

Discussion around Terminologies → Wiki

There are different views on the different tems...

- ‘Data Analysis’ supports the search for ‘causality’
 - Describing exactly WHY something is happening → science
 - Understanding causality is hard and time-consuming, but is necessary
 - Searching it often leads us down the wrong paths...
- ‘Big Data Analytics’ is focussed on ‘correlation’
 - Not focussed on causality – enough THAT it is happening → money/events
 - Discover novel patterns and WHAT is happening more quickly
 - Using correlations for invaluable insights – often data speaks for itself

- Analysis is the in-depth interpretation of ,big data’
- Analytics are powerful techniques to work on ,big data’
- Parameter/event space exploration may use (1) analytics, then (2) analysis
- Pre-/Post-Process data with (1) analytics for deeper/faster (2) data analysis processing



Scientific Use Cases as a basis → Wiki



Big Data
Analytics/
Analysis
terms in
discussion



High Energy
Physics (HEP)
Use Case

Earth Science
Event Analysis
Use Case



Particle Physics/
Radiotherapy
Use Case



Clune/Kuo et al.
GSFC/NASA

Wide variety of other
use cases from science...

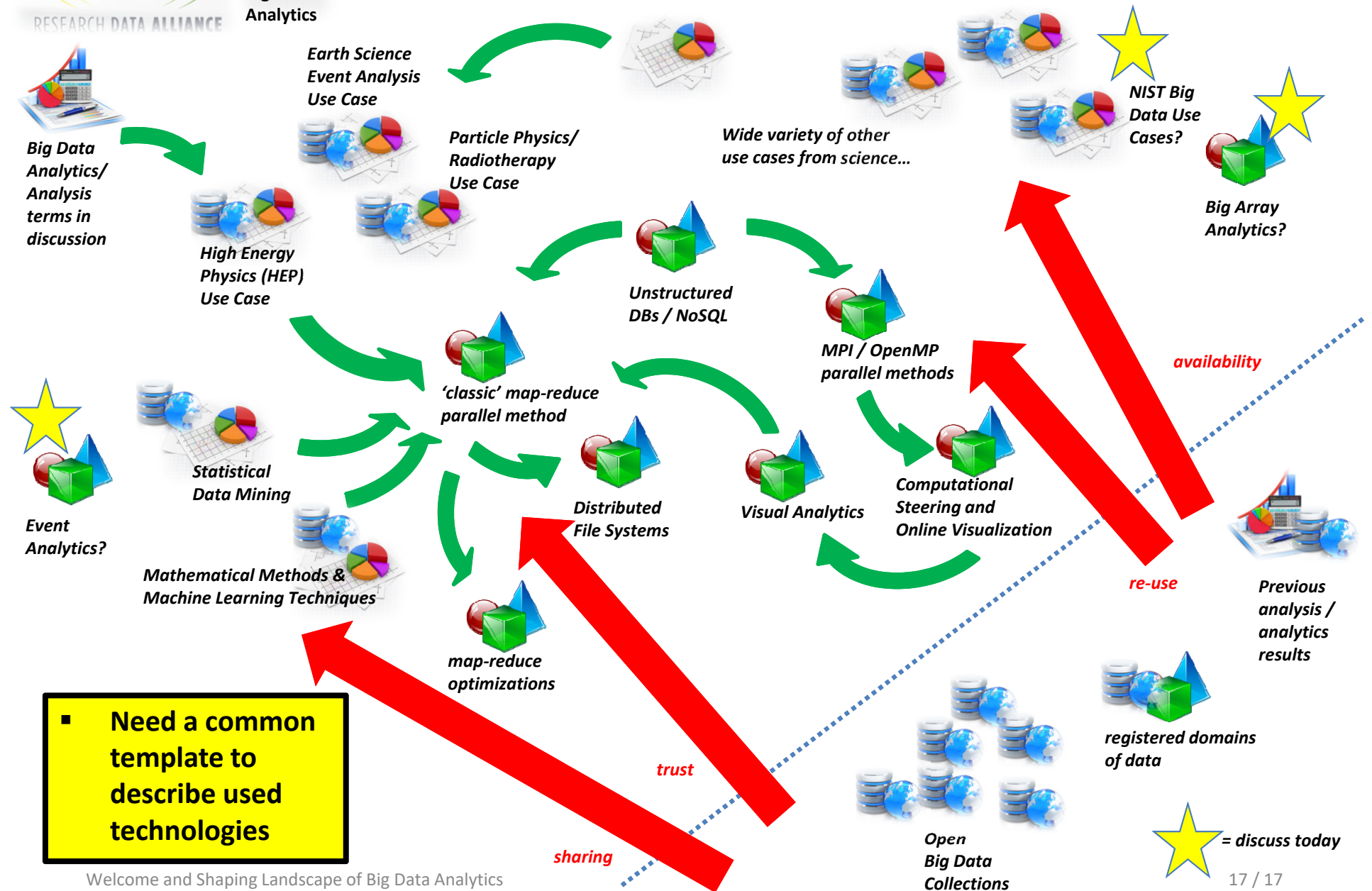
Simmons/Bouton
UoCambridge

Glaser/Neukirchen
Goettingen/Uolceland



- Need a common template to describe the scientific use cases with clear facts

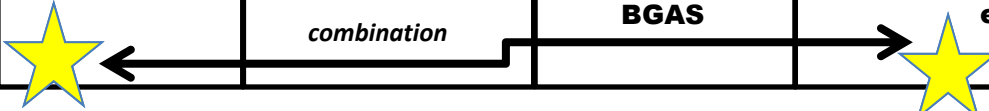
Shaping the landscape of Big Data Analytics



- Need a common template to describe used technologies

Current Systematic Analysis – Methodology

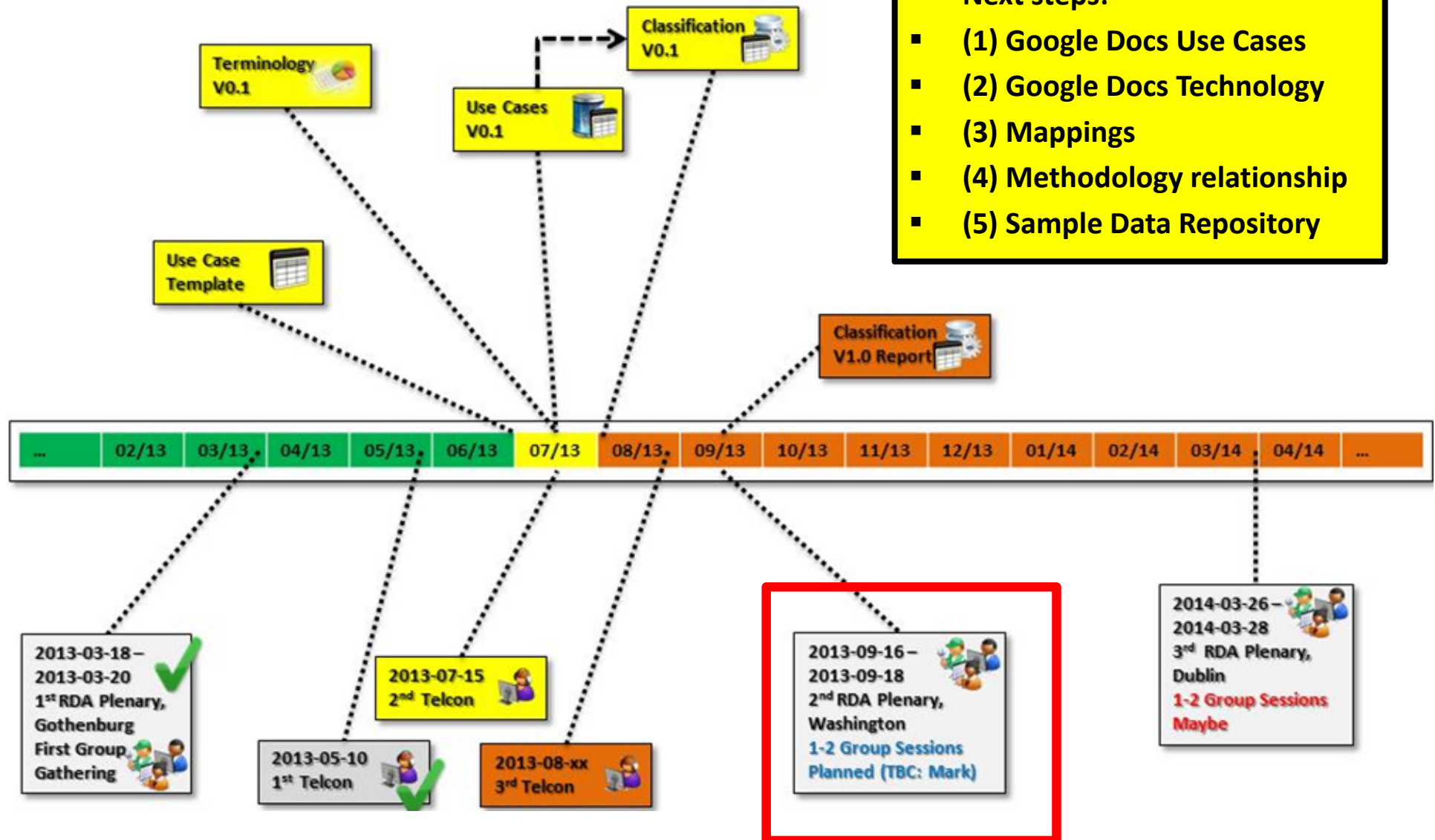
Map-Reduce		Visual Analytics		Algorithms for Large-scale Data Analysis	Extreme Data Sources	Fast Data Base Access
Classic Map-Reduce	Iterative Map-Reduce	Online/real-time Visualization	Computational Steering	Parallel algorithms, libraries, tools	Crowd Sourcing	NoSQL Databases
Loosely-Coupled Communication	Iterative loosely coupled, Pub-Sub Communication	Communication from data generator to visualizer	Communication from visualizer to steered process	Massively parallel communication with synchronization, communicators, shared memory programming	Massive amount of parallel communication streams	In-memory access & communication
BLAST, Matlab Parameter Sweeps, Ensemble Runs, Distributed Search & Sorting	Linear-algebra, Step-wise algorithms and iterative scientific problems, Page rank	Data streaming applications for thousands of data elements, interlinked data mesh	Iterative problems and step-wise approaches, nbody simulations, CFD codes	MPI-programs, openmp, FFT algorithms, PDE solvers, particle dynamics, MD codes Reliability studies Using new hardware features such as virtualized networks	Data gatherings, Correlations, ranking, community reviews, localized data	Keeping data and un-structured information for quick processing and storage
Mostly HTC, Apps	HTC towards HPC, Apps	HTC and HPC, viz cluster, Apps <i>combination</i>	HTC, rather HPC, Apps, BGAS	HPC, JUROPA3, DDN, GPGPUs, small clusters, etc.	Apps, HTC, DDN Web Scaler	Un-structured DBs, 'In-memory'



Need more granularity and concrete ,application databases' underpinned with evidence data

Process and Timelines Overview

- Next steps:
- (1) Google Docs Use Cases
- (2) Google Docs Technology
- (3) Mappings
- (4) Methodology relationship
- (5) Sample Data Repository





References

- [1] J. Wood et al., 'Riding the Wave – How Europe can gain from the rising tide of scientific data', report to the European Commission, 2010
- [2] Knowledge Exchange Partner, 'A Surfboard for Riding the Wave – Towards a Four Action Country Programme on Research Data', 2011
- [3] Research Data Alliance (RDA) Web Page, Online: <https://rd-alliance.org/node>
- [4] G. Fox, 'MPI and Map-Reduce', Talk at CCGSC 2010 Flat Rock, NC, 2010

