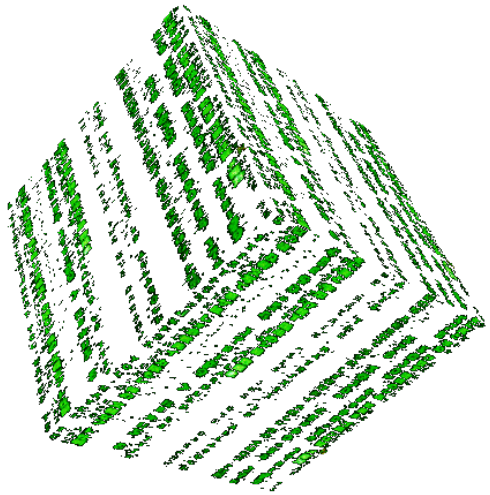


Big Data Management Challenges

e-IRG Workshop

Athens, Greece, 9-10 June 2014

Peter Baumann, Dimitar Misev

Jacobs University | rasdaman GmbH

p.baumann@jacobs-university.de

Co-funded by
the European Union

Overview

- Variety requires more data types
- Missing in science & engineering: n-D arrays
- Next-gen mediators

Big Data (not only) in Geo

Ocean Science Interoperability Experiment [OGC]

Tackling Variety

- At first glance, many different data:
  - Stock trading: 1-D sequences (i.e., arrays)
  - Social networks: large, homogeneous graphs
  - Ontologies: small, heterogeneous graphs
  - Climate modelling: 4D/5D arrays
  - Satellite imagery: 2D/3D arrays (+irregularity)
  - Genome: long string arrays
  - Particle physics: sets of events
  - Bio taxonomies: hierarchies (such as XML)
  - Documents: key/value stores: sets of unique identifiers + whatever
  - etc.

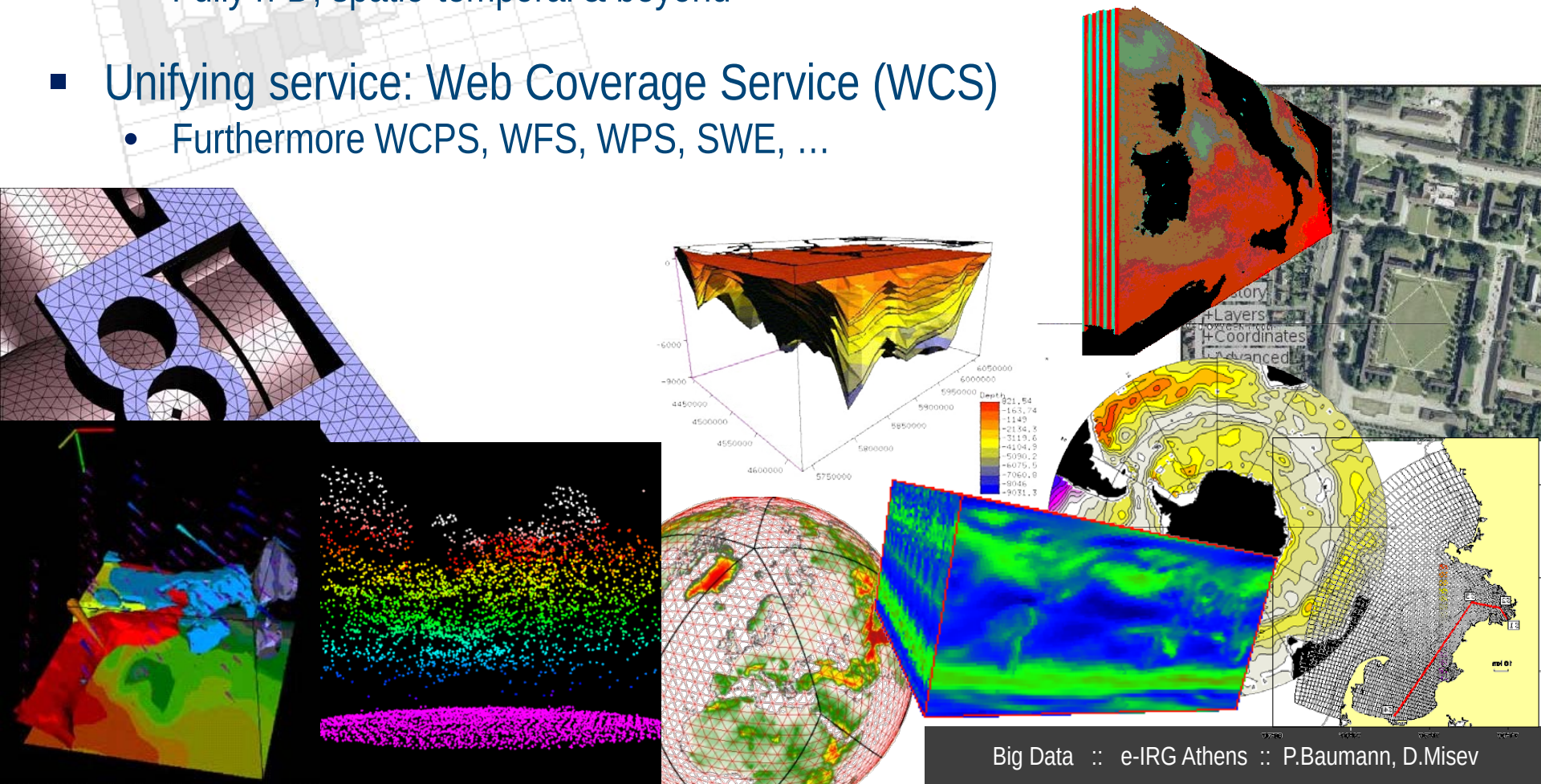
Analyzing **domain standards** (like OGC WCS) helps

- Reducible to a few core structures:
  massive sets/bags; n-D arrays; graphs; trees; ...

Managed Variety: OGC Coverage Model

[OGC 09-146r2]

- Coverage = regular & irregular grids, point clouds, meshes
  - Fully n-D, spatio-temporal & beyond
- Unifying service: Web Coverage Service (WCS)
  - Furthermore WCPS, WFS, WPS, SWE, ...



Case Study: EarthServer

- Agile analytics on any-size spatio-temporal geo data
  - EU FP7 INFRA, sep 2011 – aug 2014, 11 partners, 5m€ budget
- 6 Lighthouse Applications covering Earth & Planetary Sciences
  - Established data centers adding EarthServer technology to service portfolio
- Summer 2014: 260 TB operational

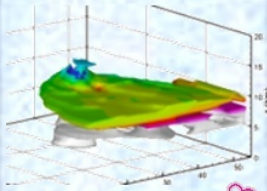
Cryospheric Science
landcover mapping

EOX

Airborne Science
high-altitude long-endurance drones

NASA

Atmospheric Science
climate variables

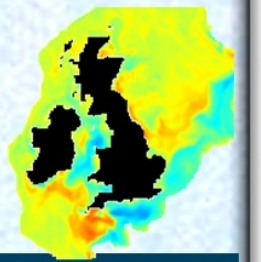


MEEO
Meteorological Environmental Earth Observation

Geology
geological models

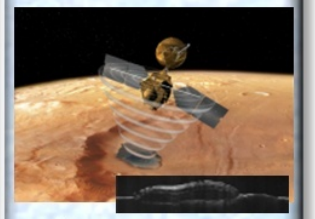
British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

Oceanography
marine model runs + in-situ data



PML PLYMOUTH MARINE LABORATORY

Planetary Science
Mars geology



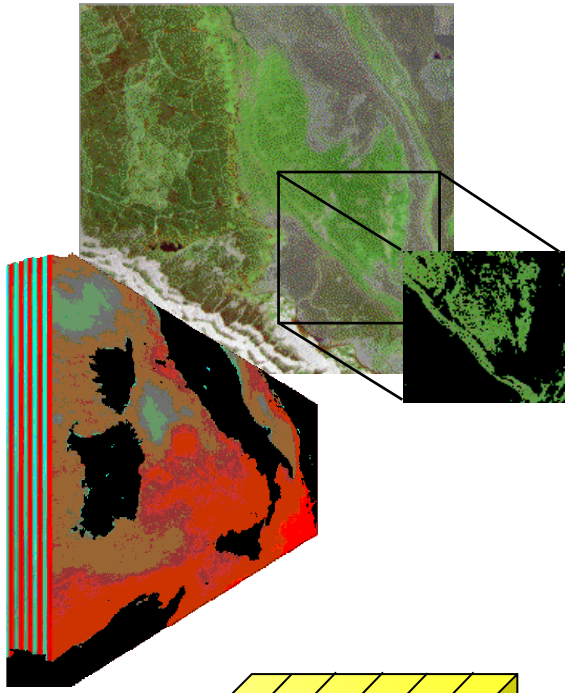
JACOBS UNIVERSITY

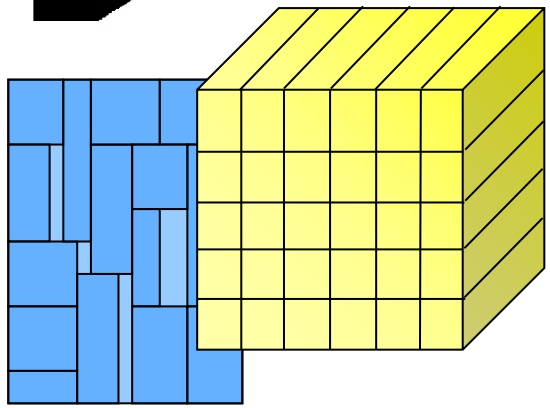
The rasdaman Array Database

- „raster data manager“: SQL + n-D arrays

```
select ls.img.green[x0:x1,y0:y1] > 130
from   LandsatArchive as ls
where  avg_cells( ls.img.nir ) < 17
```

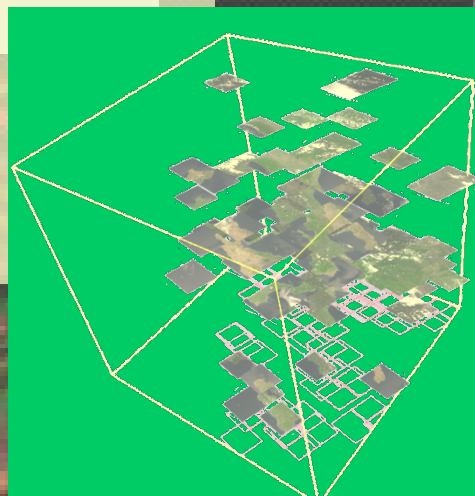
- Scalable parallel tile streaming architecture
- In operational use
  - OGC WCS Core Reference Implementation





Sample Application: Database Visualization

```
select
  encode(
    struct {
      red:   (char) s.image.b7[x0:x1,x0:x1],
      green: (char) s.image.b5[x0:x1,x0:x1],
      blue:  (char) s.image.b0[x0:x1,x0:x1],
      alpha: (char) scale( d.elev, 20 )
    },
    "image/png"
  )
from SatImage as s, DEM as d
```

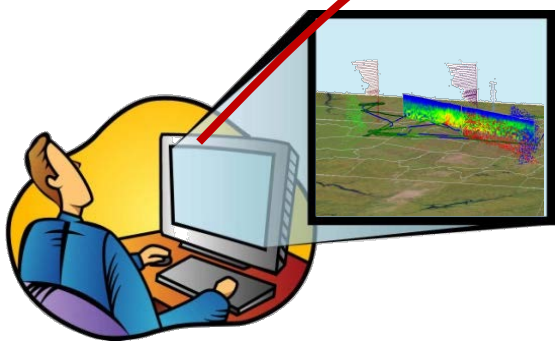


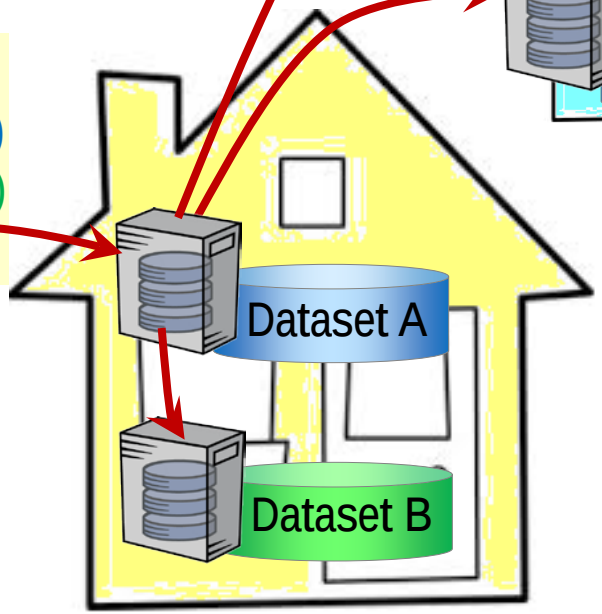
[Jacobs U, Fraunhofer; data courtesy BGS, ESA]

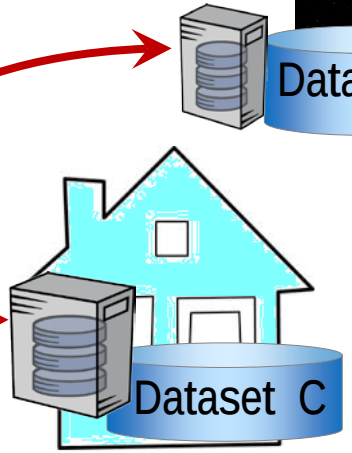
From Clouds to Federations

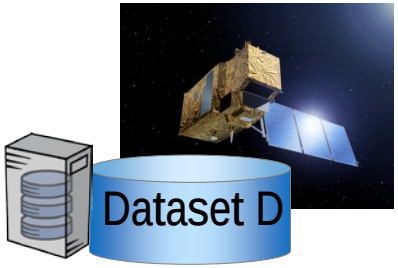
- Automatic, ad-hoc federation
  between data centers, intelligent sensors, ...
  - autonomous
  - heterogeneous

```
select
      max((C.nir - C.red) / (C.nir + C.red))
    - max((B.nir - B.red) / (B.nir + B.red))
from C, B
```









Summary

- Big Data not just volume, but also variety of data types
  - Sets, documents, graphs, arrays,...
  - *No one size fits all* - each data type calls for specific support:
    Relational DBs, graph DBs, array DBs, IR & NLP, Key/Value Stores, Hadoop, ...
- Need research on new data types, like massive multi-dimensional arrays
  - Our contribution: OGC spatio-temporal coverages; ISO Array SQL; RDA Big Data
- Need research on next-gen integration tools
  for problem-specific combinations of data models
  - Cross-model integration, optimization, privacy

[rasdaman screenshots]