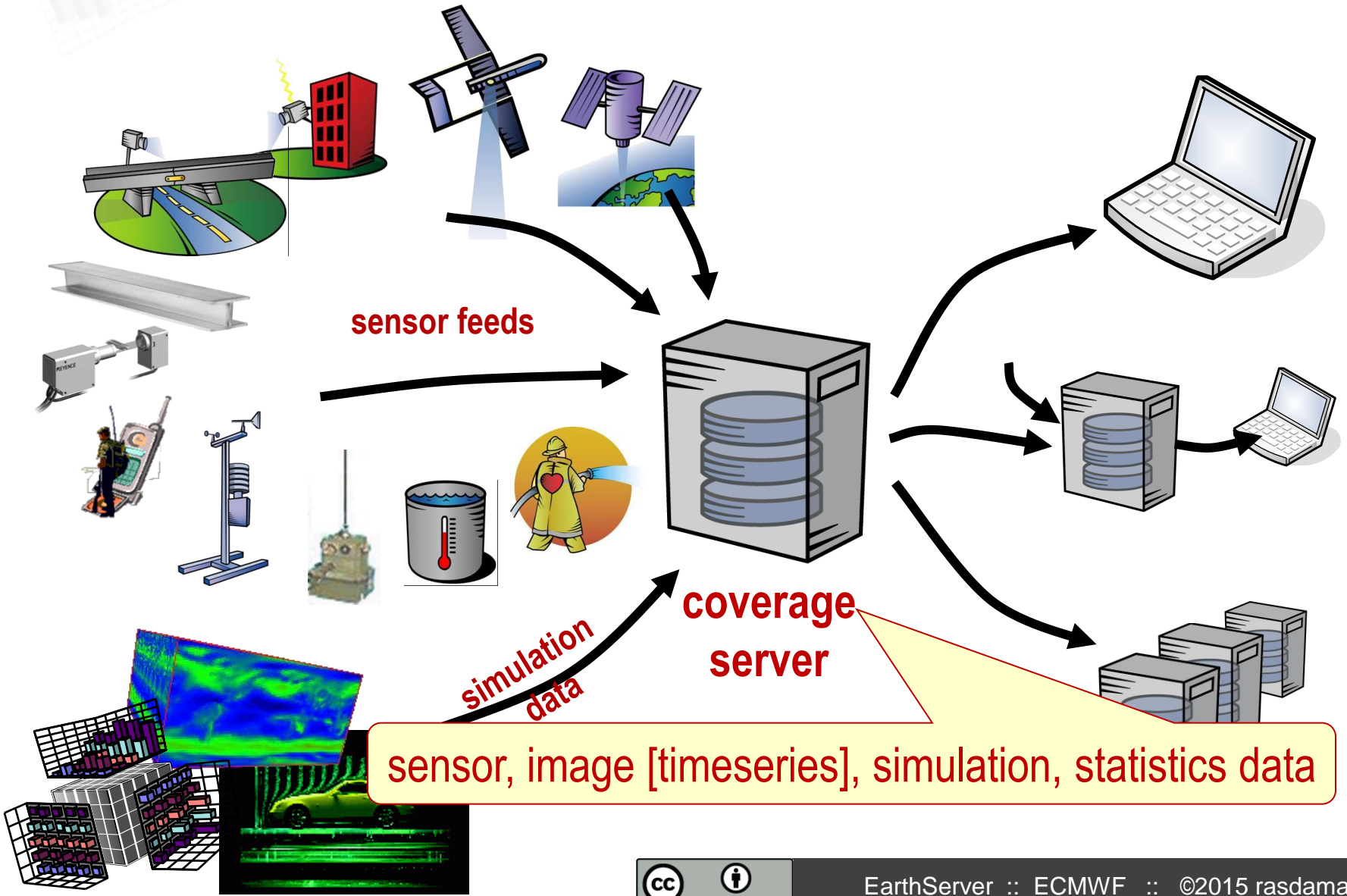# Agile Retrieval of Big Data
with *Earth*Server

ECMWF Visualization Week, Reading, 2015-sep-29

Peter Baumann

Jacobs University | rasdaman GmbH
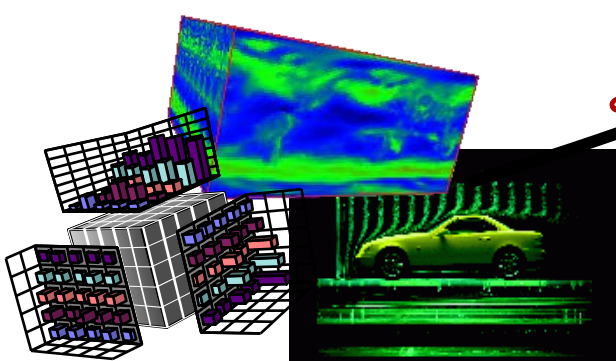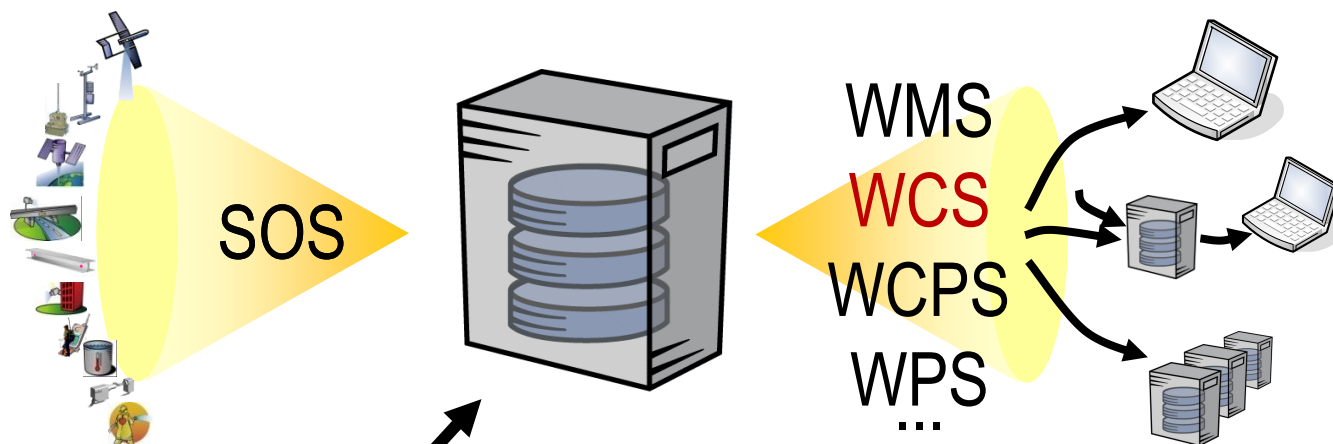
baumann@rasdaman.com

Data Homogenization With OGC Standards

sensor feeds

simulation data

coverage server

sensor, image [timeseries], simulation, statistics data

Data Homogenization With OGC Standards

SOS: upstream
sensor data capturing

W*S: downstream
download, processing, visualization

SOS

WMS
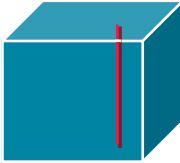WCS
WCPS
WPS
...

simulation
data

Web Coverage Service (WCS)

- OGC Coverages unifying regular & irregular grids, point clouds, meshes
  - OGC Coverage Implementation Schema

- WCS Core: retrieval of spatio-temporal coverages & subsets
  - subset =   trim | slice

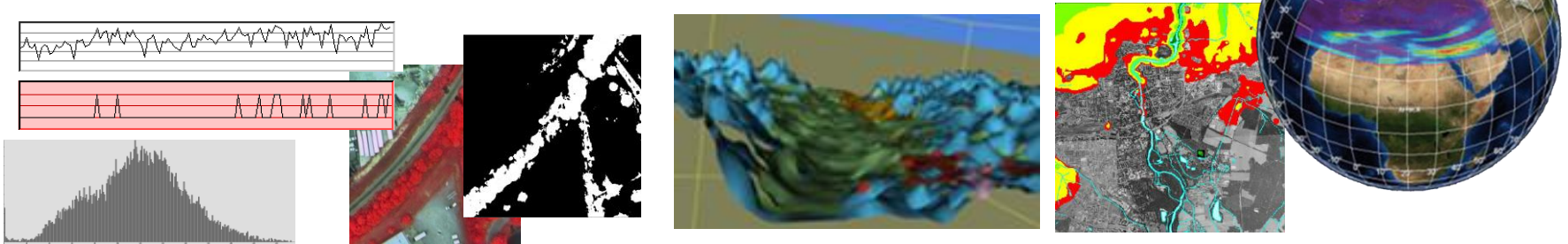**Large, growing implementation basis:** rasdaman, GDAL, QGIS, OpenLayers, OPeNDAP, MapServer, GeoServer, NASA WorldWind, EOx-Server; Pyxis, ERDAS, ArcGIS, ...

- WCS Extensions: optional functionality facets
  - Scaling, CRS transformation, Analytics (WCPS), …

Web Coverage Processing Service (WCPS)

- ▪ = high-level spatio-temporal geo analytics language

[JacobsU, FhG; NASA; data courtesy BGS, ESA]

- ▪ "From MODIS scenes M1, M2, M3: difference between red & nir, as TIFF"
  - • …but only those where nir exceeds 127 somewhere

```
for $c in ( M1, M2, M3 )
where
    some( $c.nir > 127 )
return
    encode(
        $c.red - $c.nir,
        "image/tiff"
    )
```
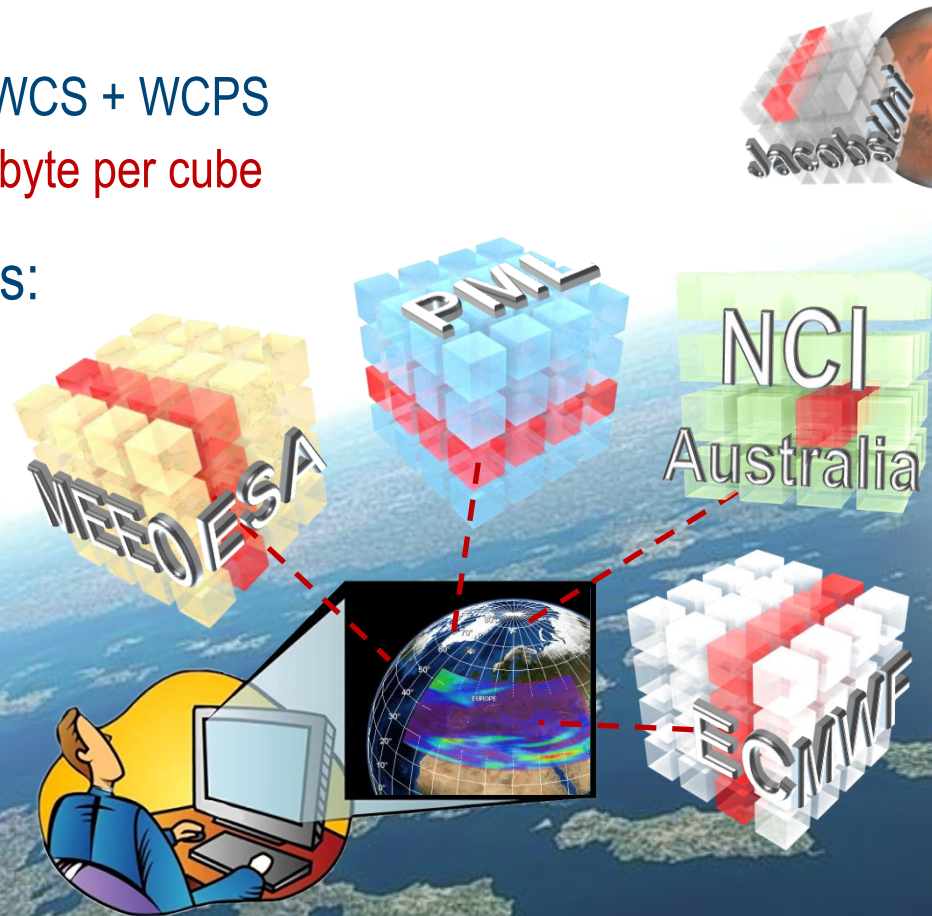
→ $(tiff_A, tiff_C)$

EarthServer: Datacubes At Your Fingertips

- **Agile Analytics** on Earth & Planetary **datacubes**
  - rasdaman + NASA WorldWind
  - Rigorously standards: OGC WMS + WCS + WCPS
  - 100s of TB online now, goal: 1+ Petabyte per cube
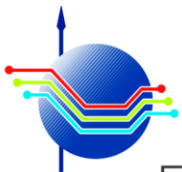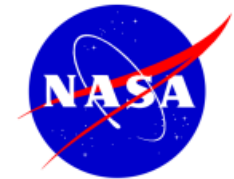
- Intercontinental initiative, 3+3 years:
EU + US + AUS

Phase 1 reviewers:
"proven evidence" that rasdaman
will *"significantly transform [how to]
access and use data"* …and *"with
no doubt has been shaping the Big
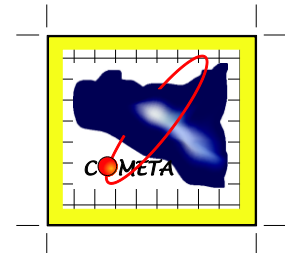Earth Data landscape"* …

www.earthserver.eu

*Earth*Server Phase 1 & 2 Partners
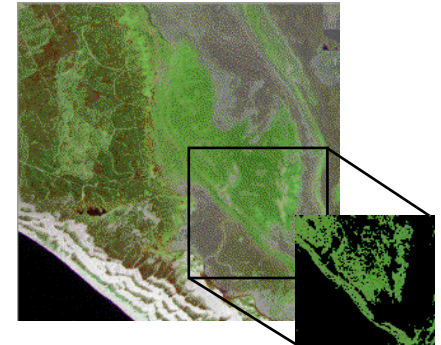
Agile Array Analytics: rasdaman

- „raster data manager": SQL + n-D arrays
  - Blueprint for ISO Array SQL standard

- Scalable parallel "tile streaming" architecture

- Supports R, QGIS, OpenLayers, MapServer, GDAL, EOxServer, Pyxis, ERDAS, ArcGIS, ...

rasdaman visitors

The Earth Monitoring Competition

opernicus masters

WINNER

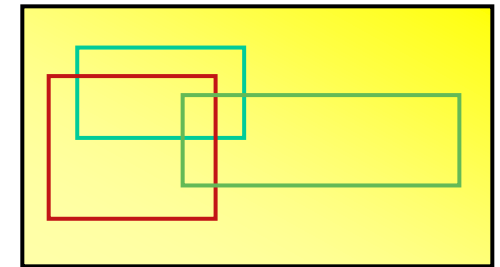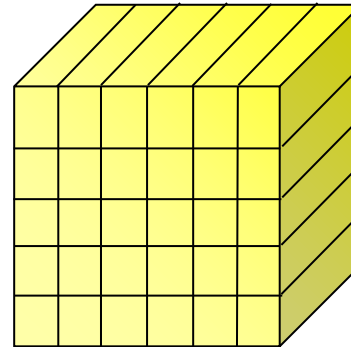T-SYSTEMS BIG DATA CHALLENGE

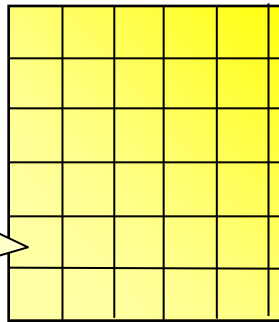2014

T··Systems·

Tiling: Tuning Data for Applications

- tiling strategies as service tuning [Furtado]:

  - regular directional area of interest

„chunks" [Sarawagi, DeWitt, ...]

- rasdaman storage layout language
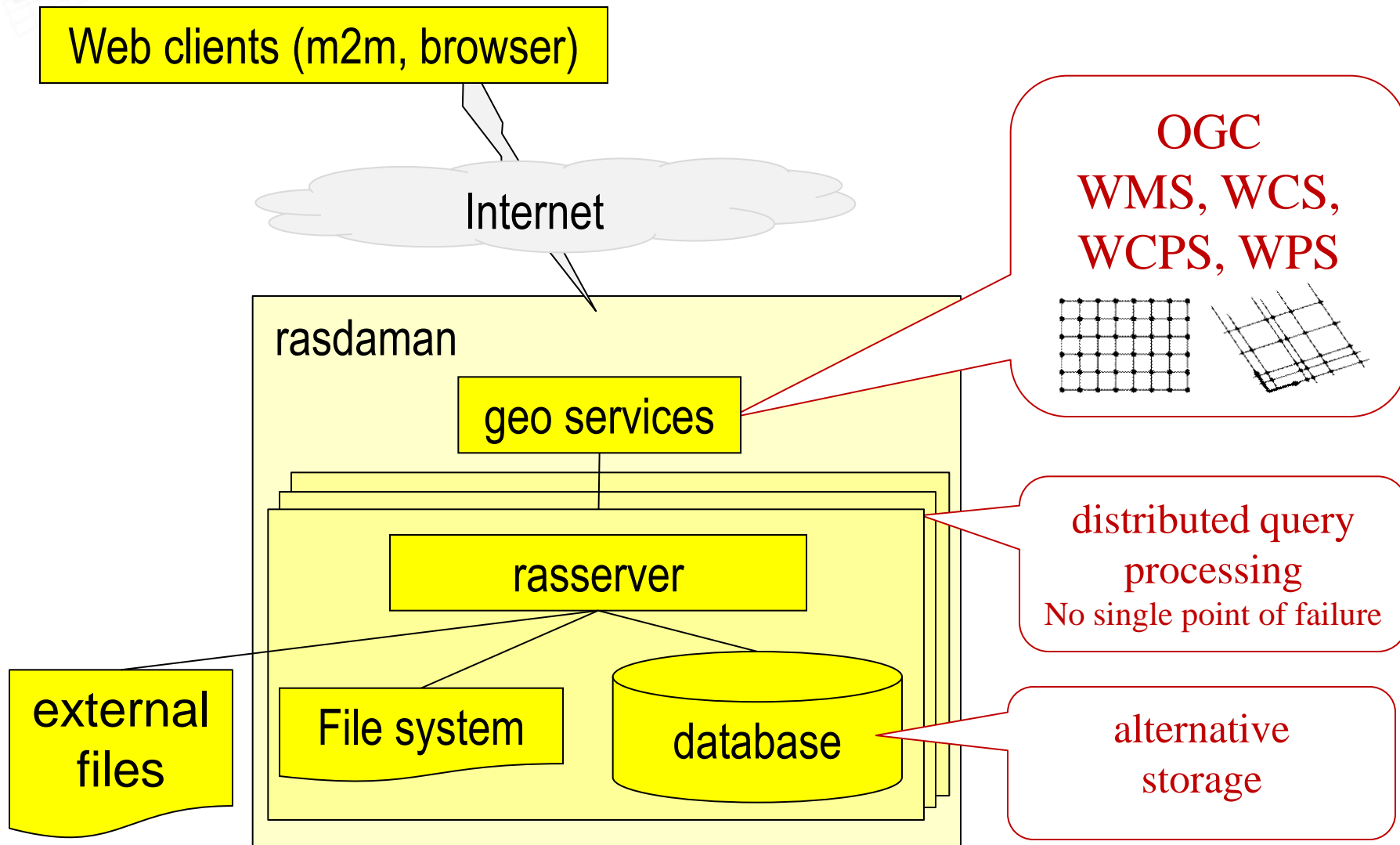
```
insert into MyCollection
    values ...
    tiling area of interest [0:20,0:40], [45:80,80:85]
    tile size 1000000
    index d_index storage array compression zlib
```

Collaborative Query Processing

- Calling external tools from within database query
  - integrated with tile streaming, optimization, parallelization

- Ex: "*NDVI from raw Landsat subset, orthorectified with Orfeo Toolbox*"

```
select
  encode(
    otb.orthoRectifFilter(
      ((img.red-img.nir)/(img.red+img.nir))[x0:x1,y0:y1],
      outputSpacing, deformationFieldSpacing
    ),
    "png"
  )
from    LandsatRawArchive as img
```

Scalable Geo Service Architecture

Web clients (m2m, browser)

Internet

rasdaman

geo services

rasserver

File system

database

external files

OGC
WMS, WCS,
WCPS, WPS

distributed query processing
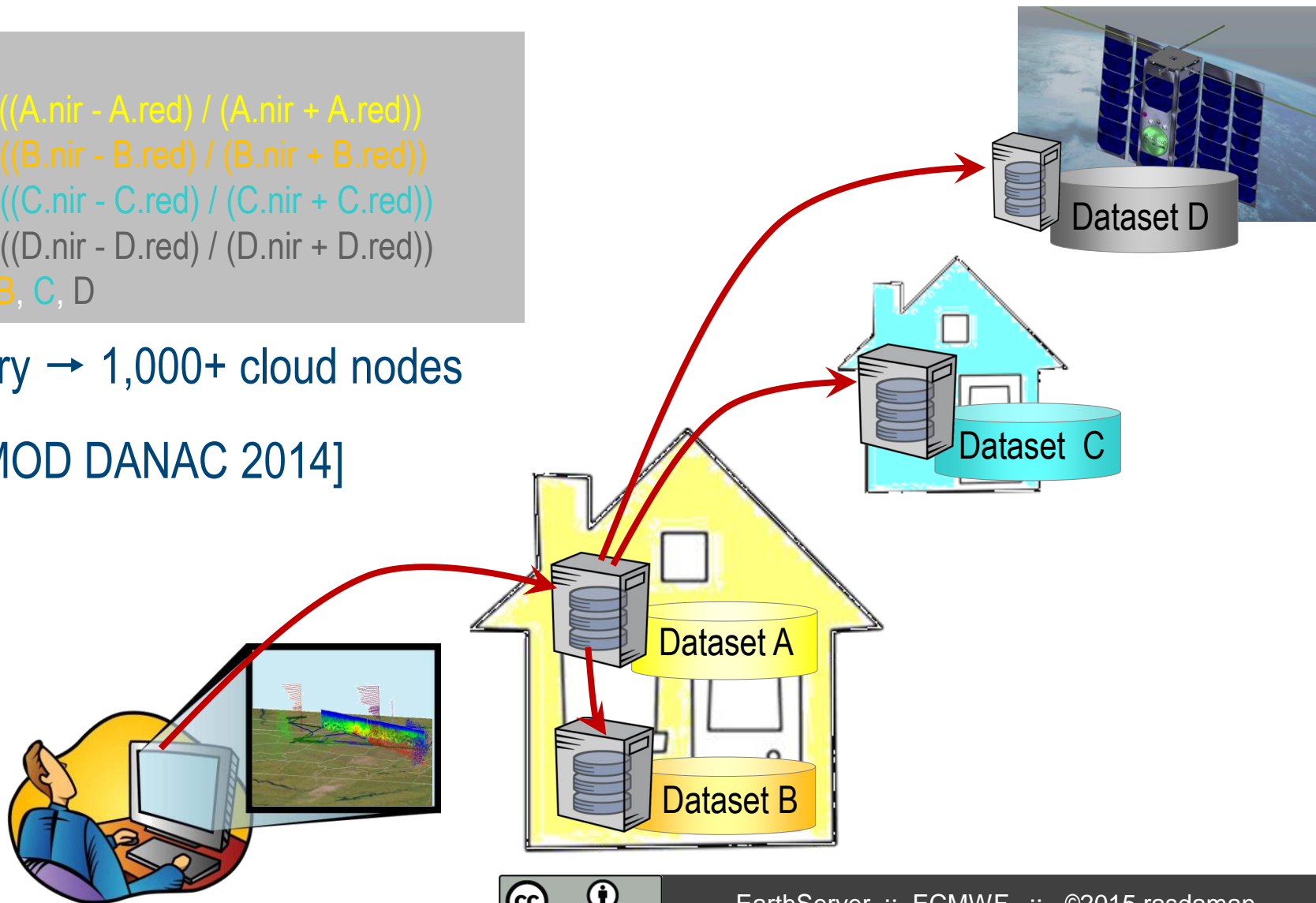No single point of failure

alternative storage

Parallel / Distributed Query Processing

```
select
    max((A.nir - A.red) / (A.nir + A.red))
  - max((B.nir - B.red) / (B.nir + B.red))
  - max((C.nir - C.red) / (C.nir + C.red))
  - max((D.nir - D.red) / (D.nir + D.red))
from A, B, C, D
```
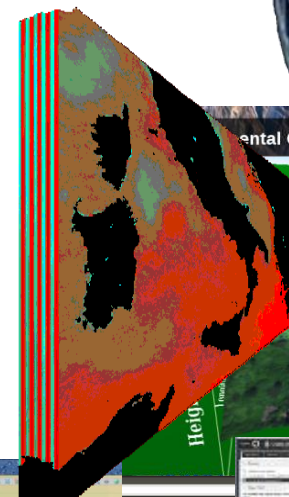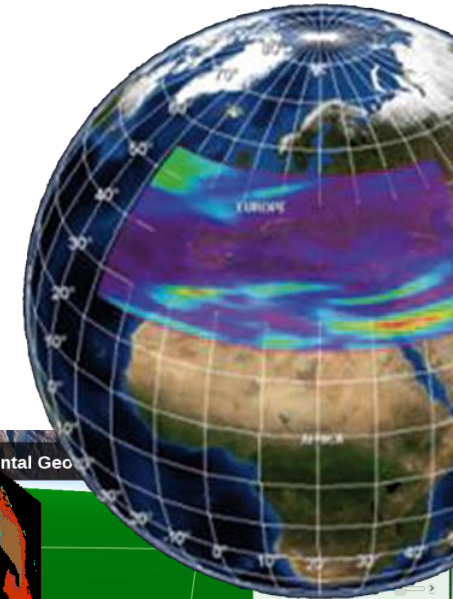
1 query → 1,000+ cloud nodes

[SIGMOD DANAC 2014]

Dataset D

Dataset C

Dataset A

Dataset B

Take Home Messages

- Conquering Big Data requires high-level languages
  - Problem-adjusted → Array Databases like rasdaman

- WCS: OGC „Big Geo Data" standard
  - Under adoption by ISO, INSPIRE

  EarthServer: Agile Analytics
  on Petascale spatio-temporal datacubes
  - pictures → actionable data

[rasdaman screenshots]