rasdaman raster data manager

JACOBS UNIVERSITY

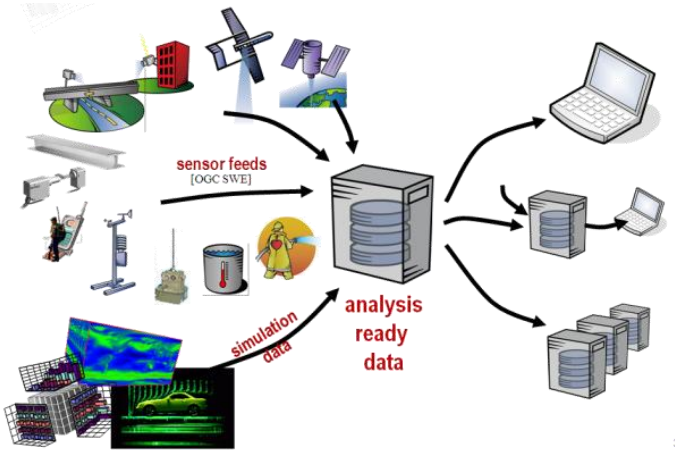
From Analysis-Ready Data to Analysis-Ready Services: Challenges & Helpers for EO Service Providers

ESA Big Data from Space, 2019-feb-20

Peter Baumann

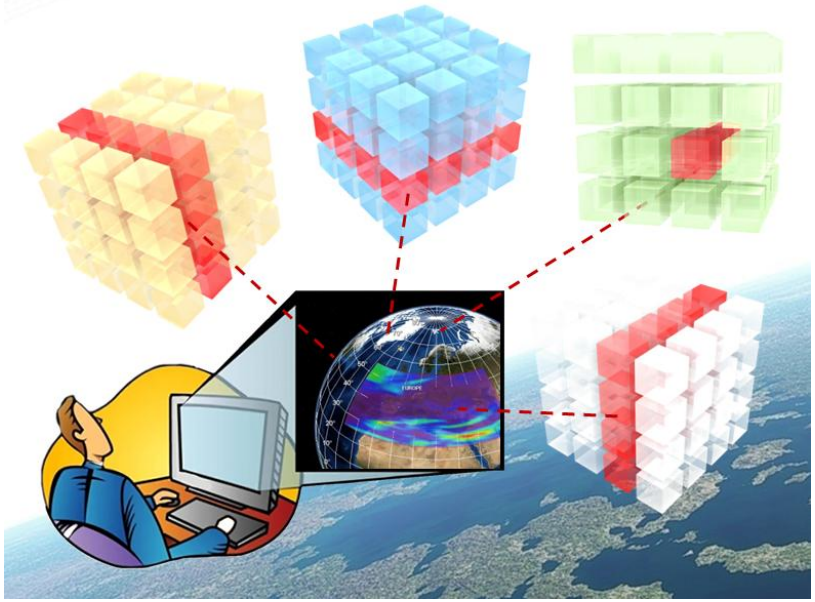
Jacobs University | rasdaman GmbH

Motivation



- Analysis-Ready Data (ARD)
  - USGS Landsat team in 2017
  - rapid uptake, variety of interpretations
  - which, how-ever, all agree that EO data need to be offered in a way better suitable for consumption in particular by non-pro--gramm-ers and non-EO experts.
- CEOS Analysis Ready Data for Land (CARD4L)
  - “immediate analysis with a minimum of additional user effort and interoperability both through time and with other data-sets”
  - metadata requirements, radiometric & geometric calibration, solar & view angle correction, atmospheric correction (optical) & topography / incidence angle correction (radar)

Motivation / contd.



- homogenized, aggregated, standards-conformant offerings
  - Abstract from storage & encodings
  - "going from files to pixels"
  - WCS in addition to WMS, WMTS
- Easy navigation, extraction, aggregation along space & time
  - Files & scenes → n-D datacubes
- "go take the data, do analysis yourself" → "build your product on the go"
  - server-side analysis capabilities, "ship code to data"
  - What code? How complex? → quality of service
- analysis-ready data → *analysis-ready services*

Case Study: SAFE

- Sentinel data delivered in SAFE format
- Server effort for analysis tasks?
- Zip archive → extra tool invocation for extracting image file
  - subdirectories
- JPEG (lossless) → extra CPU cycles for pixel reconstruction
  - Wavelets suboptimal for spatio-temporal subsetting
- File granularity: 100x100km → hundreds of MB...GB
  - Benchmarks [Furtado et al]: ~3 MB suitable
- SAFE is archive format, not service format!

Analysis-Ready: Requirements v0.1

- *granularity for efficient spatio-temporal access, i.e.: x/y/z/t*
  - re-tiling scenes=slices → cubelets
  - Prefer file format with internal tiling
  - Tiling as configuration parameter, not hardcoded API feature (ex: WMTS)
- *Minimize pixel reconstruction*
  - No wavelets (soft requirement)
- *Store data analysis-ready*
  - Construction on the fly inefficient, numerical inconsistencies
  - Provide authoritative values readily available in database / archive
    - *Not Level 1a, 1b, but Level 1c+ (error corrected, radiometrically corrected, orthorectified)*

Analysis-Ready: Requirements v0.1 / contd.

- *Ship code to data: high-level server-side filtering & processing language*
  - Low-level ftp, RESTful sub-setting APIs, etc: no substantial server-side processing
  - procedural source code (ex: python): major security hole
  - high-level, declarative language (ex: SQL, OGC WCPS): safe in evaluation
- *Transparent federation*
  - Fusion across data centers: not in client, but in server
  - Intelligent orchestration, optimization of data exchange, processing distribution

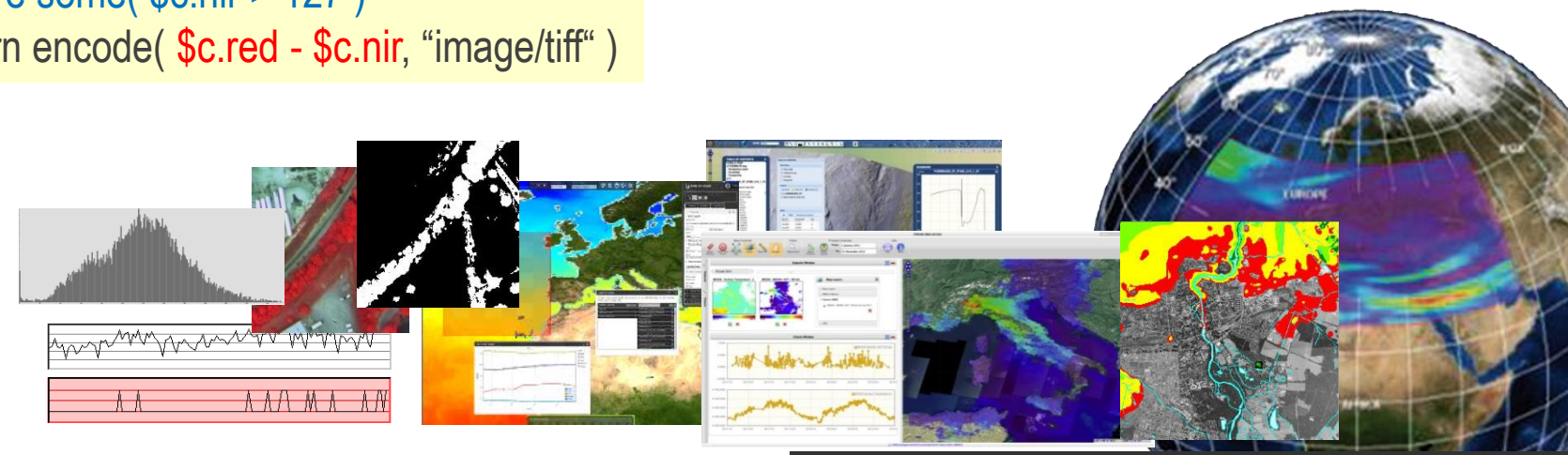
Analysis-Ready: Requirements v0.1 / contd.

- Web Coverage Service (WCS): *simple to use, modular, functionality-rich*
  - Core: subsetting + formatting; Extensions: more facets

```
http://www.acme.com/wcs ? SERVICE=WCS & VERSION=2.0
    & REQUEST=GetCoverage & COVERAGEID=c001
    & SUBSET=Long(100,120) & SUBSET=Lat(50,60) & SUBSET=time("2009-11-06T23:20:52")
```

- Web Coverage Processing Service (WCPS) datacube analytics

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



Implementation Feasibility

- Requirements doable with today‘s technology, such as databases
- Ex: rasdaman Array DBMS
- Declarative query interface: WCPS, internally mapped to Array SQL
  - ISO SQL 9075-15:2018: Multi-Dimensional Arrays (MDA)
  - Coined datacube services / Array Databases [Baumann 1992]
    - *Comp Sci PhD theses 1999...2018*
- Versatile ETL suite for automated cleansing & ingestion
- Proven on 2.5+ PB,
  1000+ cloud parallelization,
  intercontinental federation ↳

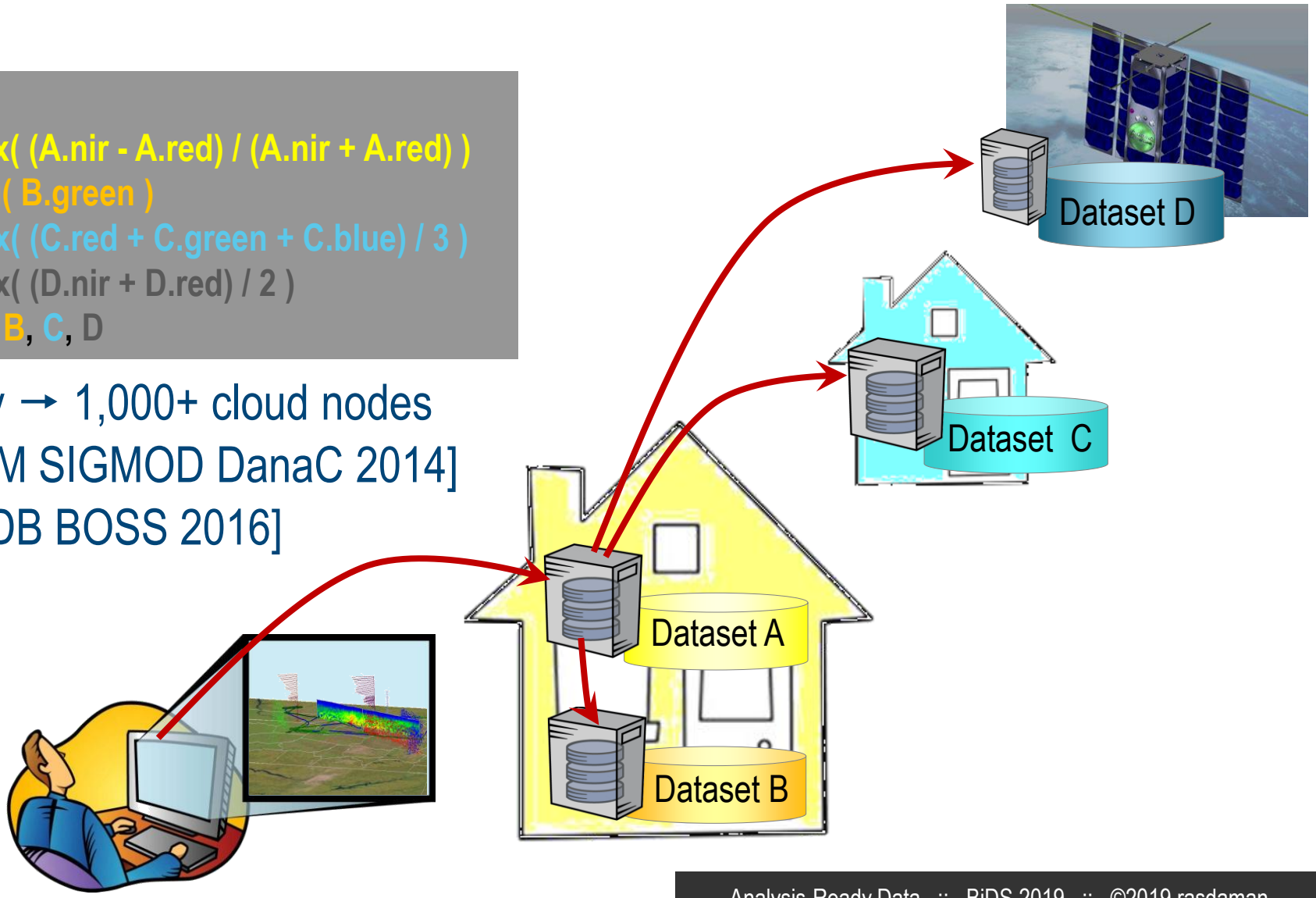
Parallel, Distributed Processing

```
select
      max( (A.nir - A.red) / (A.nir + A.red) )
   + avg( B.green )
   + max( (C.red + C.green + C.blue) / 3 )
   + max( (D.nir + D.red) / 2 )
from A, B, C, D
```

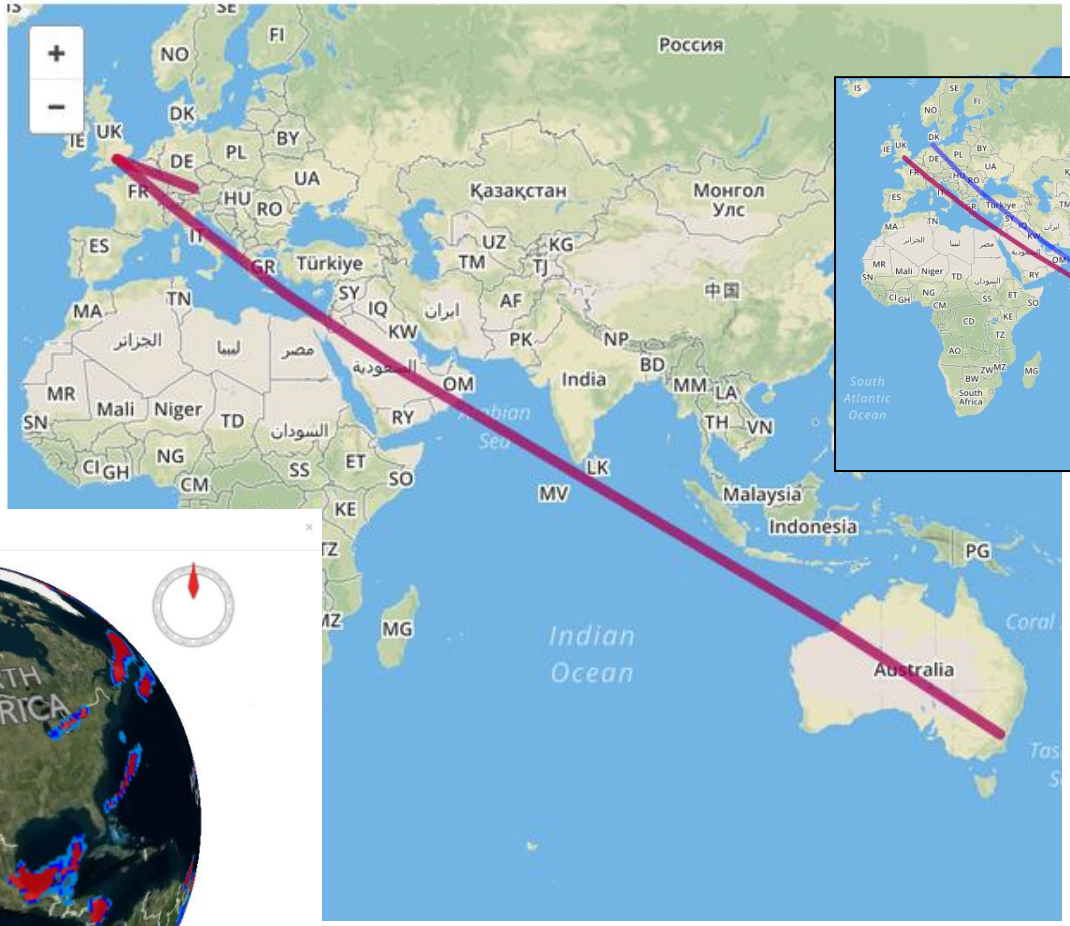
1 query → 1,000+ cloud nodes

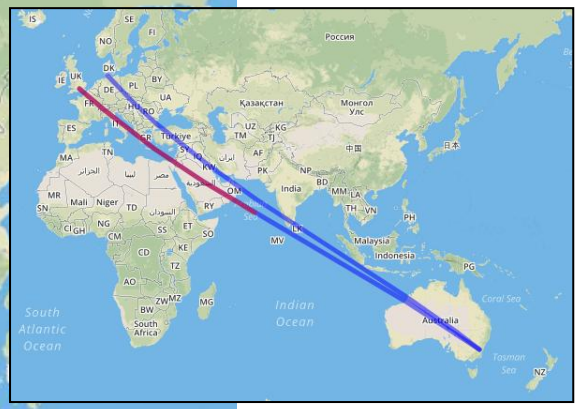
[ACM SIGMOD DanaC 2014]

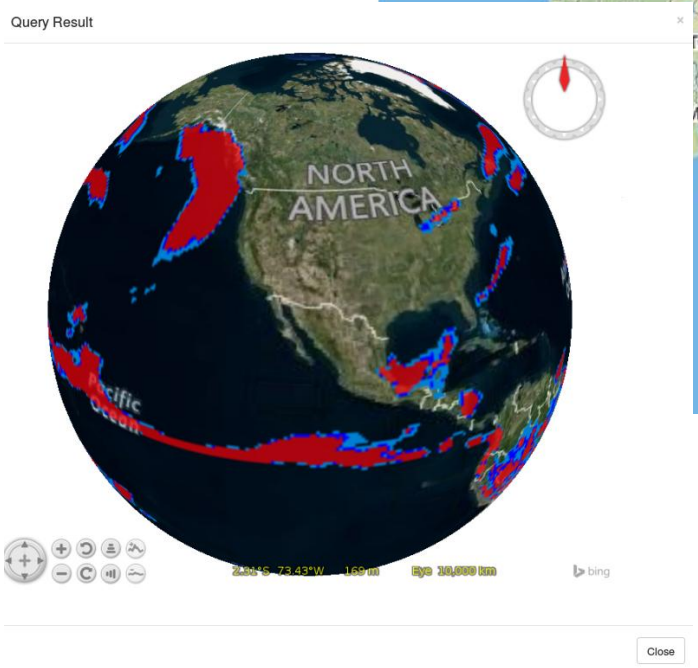
[VLDB BOSS 2016]



Federation in EarthServer

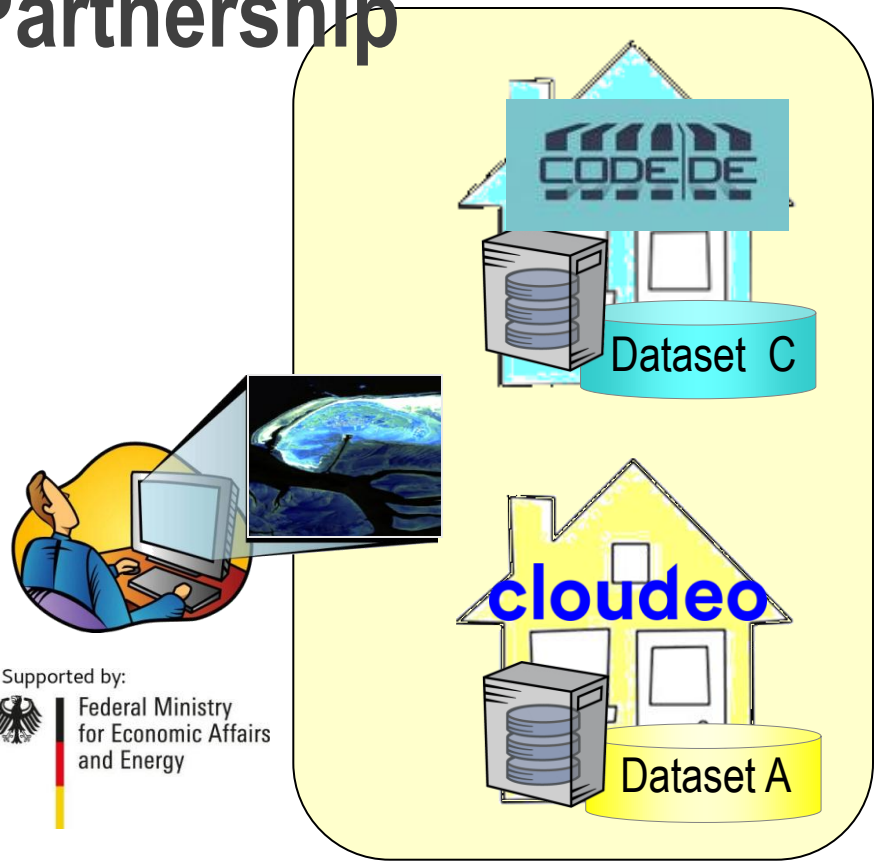






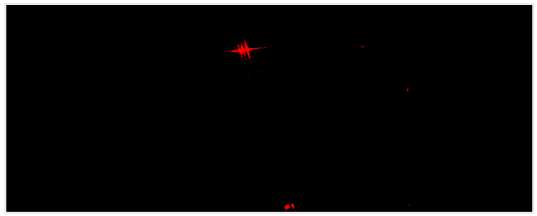
Public/Private Datacube Partnership

- BigDataCube:
  - public CODE-DE Sentinel hub
  - commercial cloudeo services
  - Security + billing
- Interactive datacube frontend complementing batch Hadoop
- CODE-DE adding homogenized data



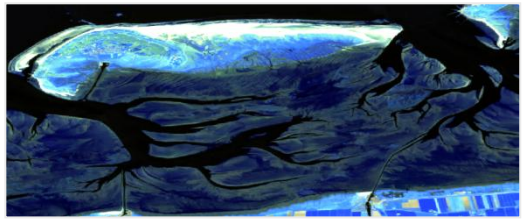
Supported by:
Federal Ministry for Economic Affairs and Energy

S1 ship detection



time-avg'ed S1 area

S2 atmosph. penetration



Conclusions

- Analysis-Ready good for users
  - Spatio-temporal data require datacubes
- OGC coverage data & service model
  - regular & irregular grids
  - easy-to-use functionality
- Array DBMSs: query languages for flexibility, scalability
  - SQL : COBOL vs WCPS : python
  - Federation
- Workload shift: end users → data providers
  - Archive formats → homogenized, processing-ready