

Array Databases

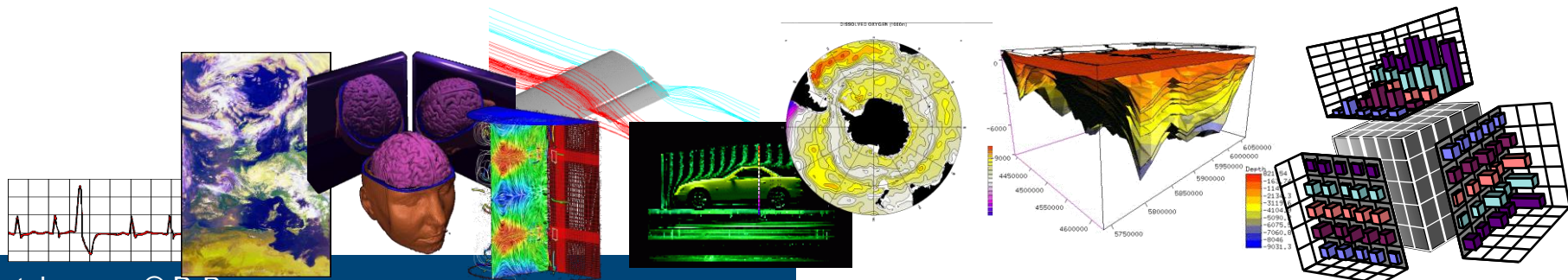
<http://l-sis.org> → publications

http://en.wikipedia.org/wiki/Array_DBMS

Who Needs Arrays?

- **Sensor, image, simulation, statistics data**
 - **Earth:** Geodesy, geology, hydrology, oceanography, climate, earth system, ...
 - **Space:** optical / radio astronomy, cosmological simulation, planetary science, ...
 - **Life:** Pharma/chem, healthcare / bio research, bio statistics, genetics, ...
 - **Engineering & research:** Simulation & experimental data in automotive/shipbuilding/aerospace industry, turbines, process industry, ...
 - **Management/Controlling:** Decision Support, OLAP, Data Warehousing, census, statistics in industry and public administration, ...
 - **Multimedia:** distance learning, prepress, ...

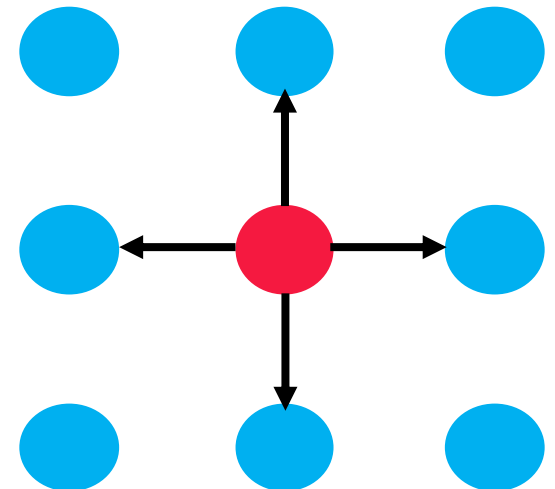
- „80% of all data have some spatial connotation“ [C&P Hane, 1992]



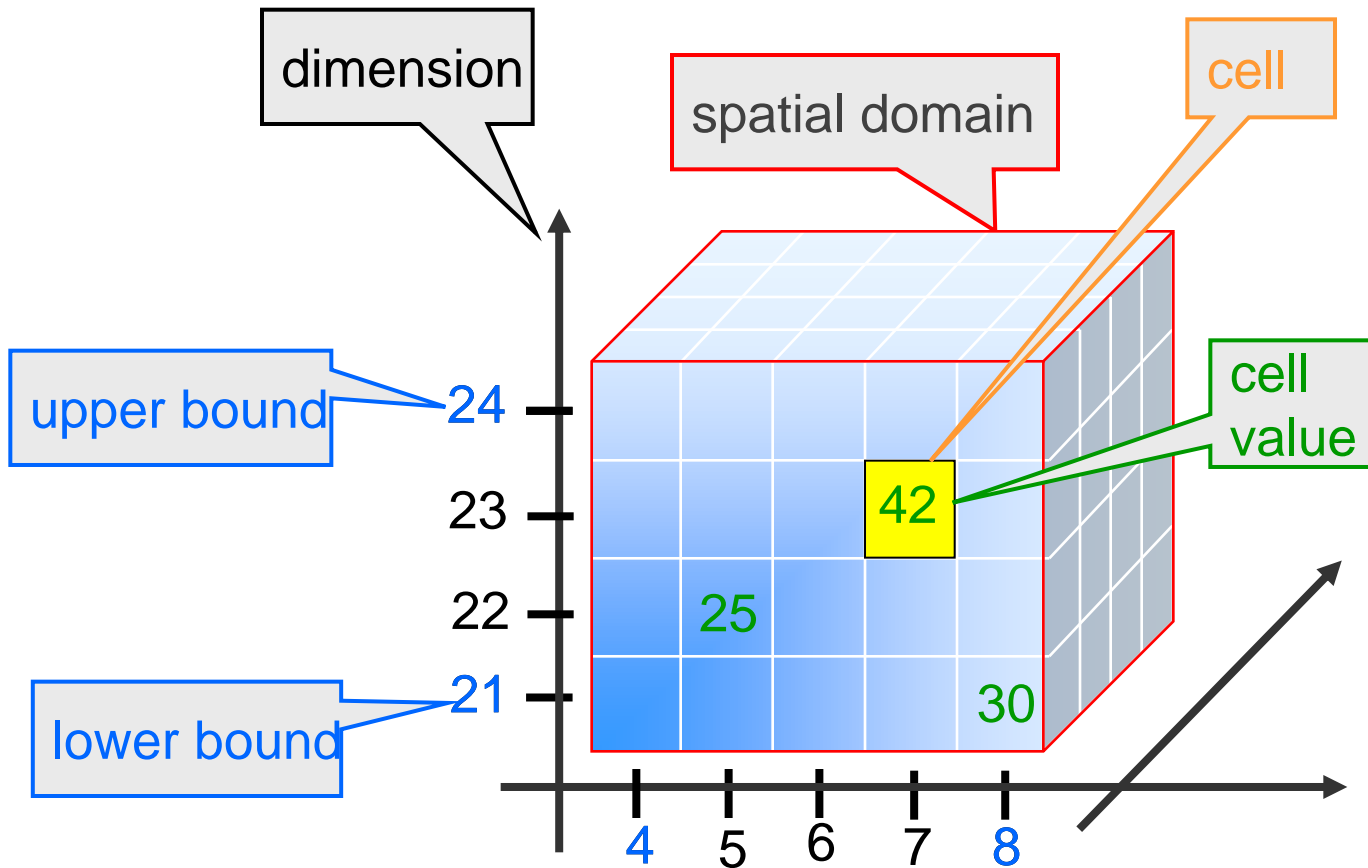
CONCEPTUAL MODELLING

Array Analytics

- Array Analytics :=
Efficient analysis on multi-dimensional arrays of a size several orders of magnitude above evaluation engine's main memory
- Essential **data** property: n-dimensional Cartesian neighborhood
 - Secondary: #dimensions, density, ...
- **Operations**: signal/image processing, Linear Algebra [M. Stonebraker], iterations



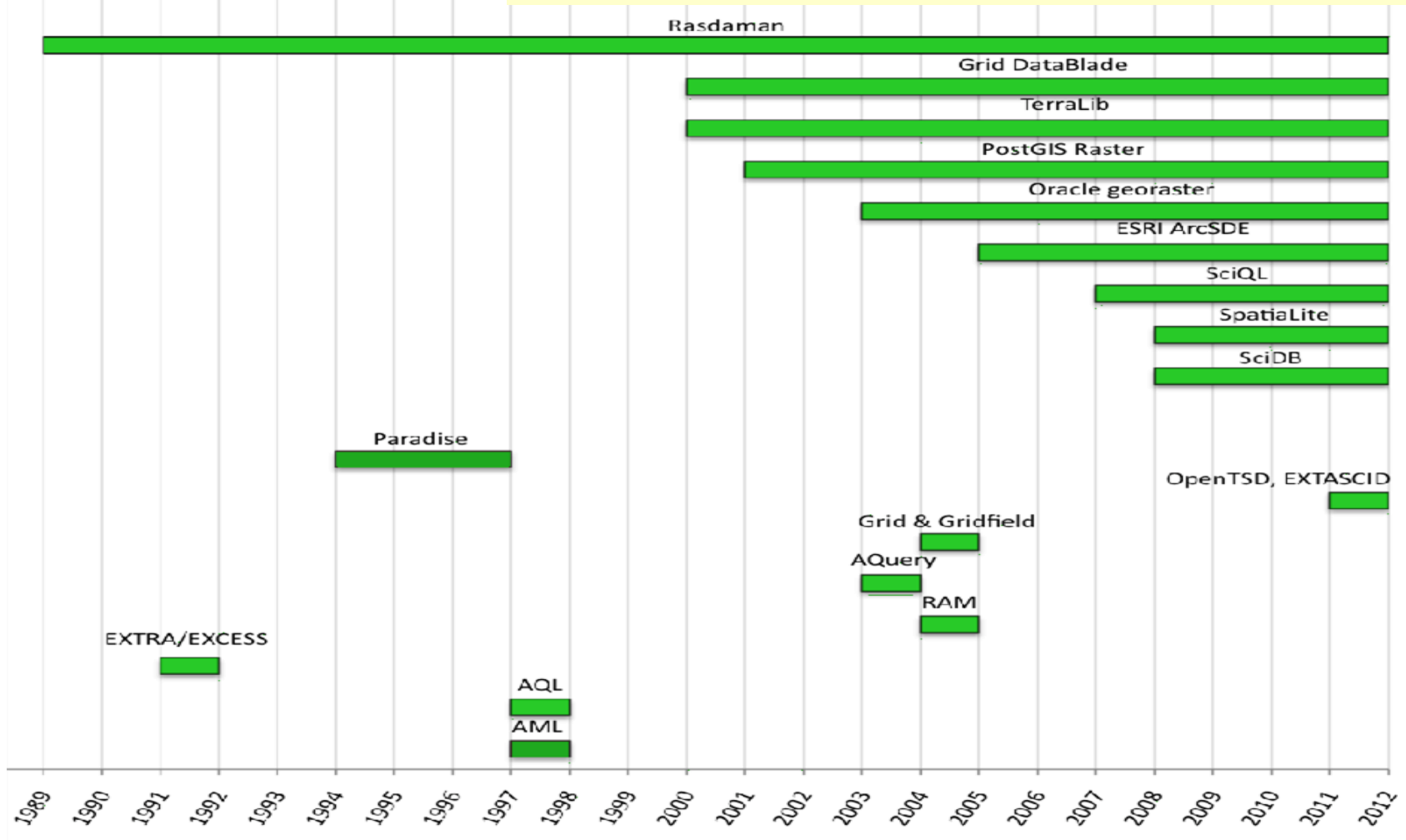
The Array Data Model



SYSTEMS

Early History of Array Databases

See also: [RDA Array Database Assessment WG](#)



Array DBMSs Landscape Today

rapidly evolving ecosystem
→ necessarily incomplete

■ Array Database Systems

- query language, multi-user operation, storage management, access control
- Ex: rasdaman, SciDB, EXTASCID, PostGIS Raster, Oracle GeoRaster

■ Array tools: command-line tools & libraries, but no service

- no query concept, but procedural API
- Ex: OpenDataCube, OPeNDAP, Wendelin.core, TensorFlow, boost::geometry, xtensor, TileDB, ArrayStore, Ophidia

■ Map/Reduce: Hadoop & Spark as cloud parallelization paradigm

- Array layers on top of Hadoop, Spark
- Ex: SciHadoop, Spatial Hadoop, GeoTrellis, MrGeo, SciSpark, ClimateSpark

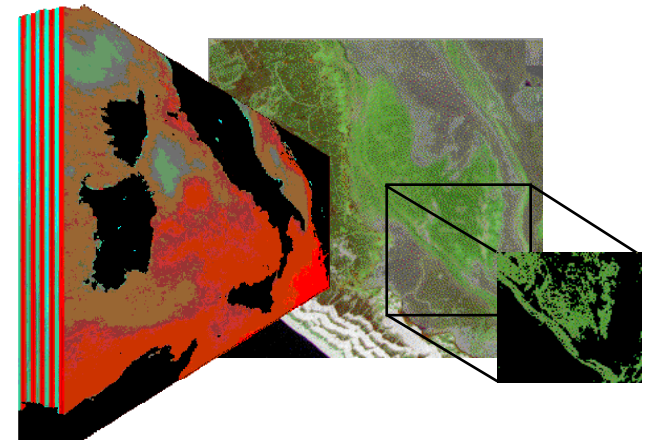
Array DBMSs Landscape Today

- Array Database Systems
- Array tools: command-line tools & libraries, but no service
- Map/Reduce: Hadoop & Spark as cloud parallelization paradigm
- Technology overview
 - 19 technologies compared
 - 4 benchmarked

rasdaman

- „raster data manager“: **SQL + n-D arrays**
 - Scalable parallel “tile streaming” architecture
 - [VLDB 1994, VLDB 1997, SIGMOD 1998, VLDB 2003, ..., VLDB 2016]

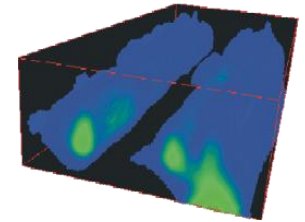
- Blueprint for stds, in operational use



The rasql Query Language

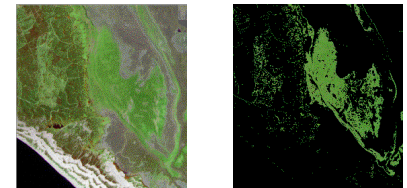
- selection & subsetting

```
select c.data[ ** , 100:200 , ** , 42 ]
from   ClimateSimulations as c
```



- result processing

```
select img.data * (img.data.green > 130)
from   LandsatArchive as img
```



- search & aggregation

```
select mri.data
from   MRI as img, masks as am
where  some_cells( mri > 250 and m )
```



- data format conversion

```
select encode( c.data[** , ** , 100 , 42] , „png“ )
from   ClimateSimulations as c
```



Linear Algebra Ops

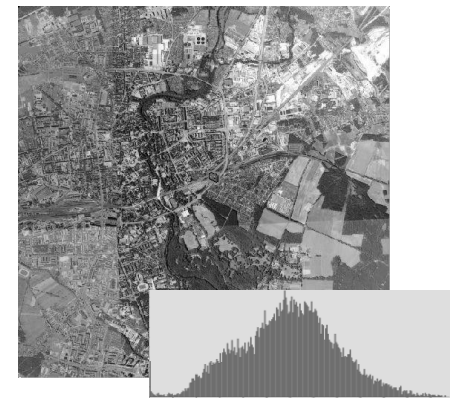
- $n \times p$ Matrix multiplication

$$(\mathbf{AB})_{ij} = \sum_{k=1}^m A_{ik} B_{kj}$$

```
select marray i in [1:n], j in [1:p]
       values condense +
           over      k in [1:m]
           using     a [ i, k ] * b [ k, j ]
from   matrix as a, matrix as b
```

- Histogram

```
select marray bucket in [0:255]
       values count_cells( img = bucket )
from   img
```



Arrays in SQL

[SSDBM 2014]

rasdaman as blueprint

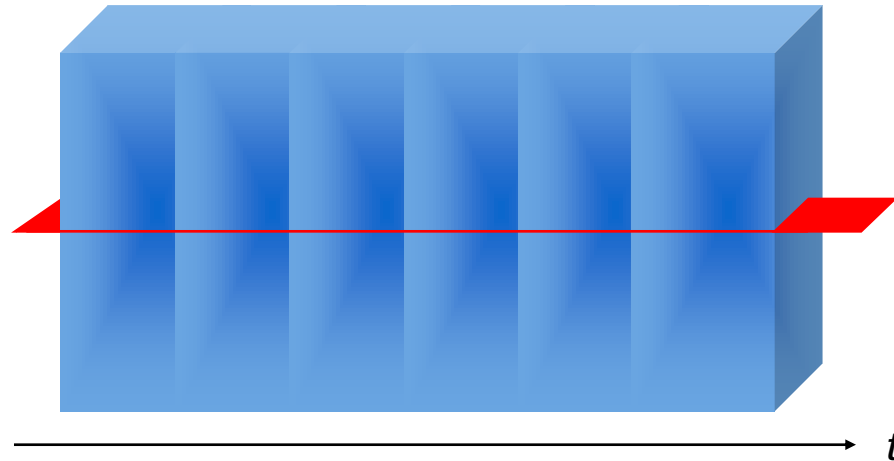
The screenshot shows the ISO website header with the logo and tagline 'When the world agrees'. Below the header is a navigation menu with 'Standards', 'All about ISO', 'Taking part', and 'Store'. A search bar is also present. The main content area shows the breadcrumb trail: 'Store > Standards catalogue > ICS > 35 > 35.060 > ISO/IEC 9075-15:2019'. The product title is 'ISO/IEC 9075-15:2019' with a 'Preview' button. The description is 'Information technology database languages -- SQL -- Part 15: Multi-dimensional arrays (SQL/MDA)'.

```
create table LandsatScenes(
  id: integer not null, acquired: date,
  scene: row( band1: integer, ..., band7: integer ) marray [ 0:4999,0:4999] )
```

```
select id, encode(scene.band1-scene.band2)/(scene.nband1+scene.band2)), „image/tiff“ )
from LandsatScenes
where acquired between „1990-06-01“ and „1990-06-30“ and
  avg( scene.band3-scene.band4)/(scene.band3+scene.band4)) > 0
```

ARCHITECTURE

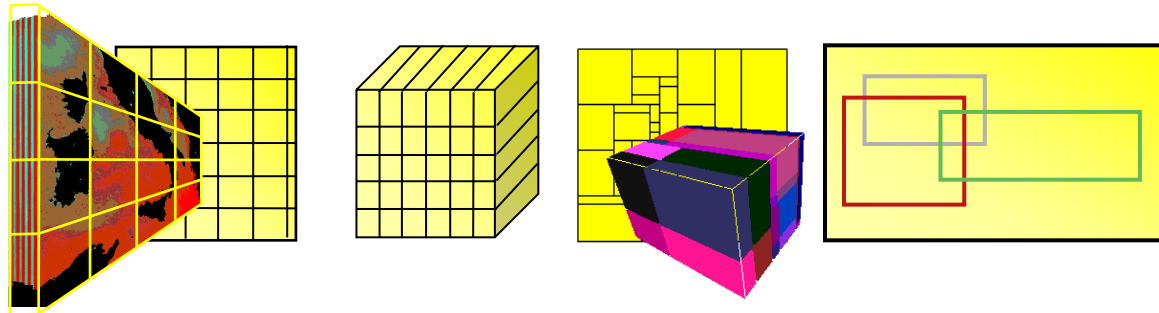
Storage Management



- Divergent access patterns for ingest and retrieval
- Server must mediate between access patterns

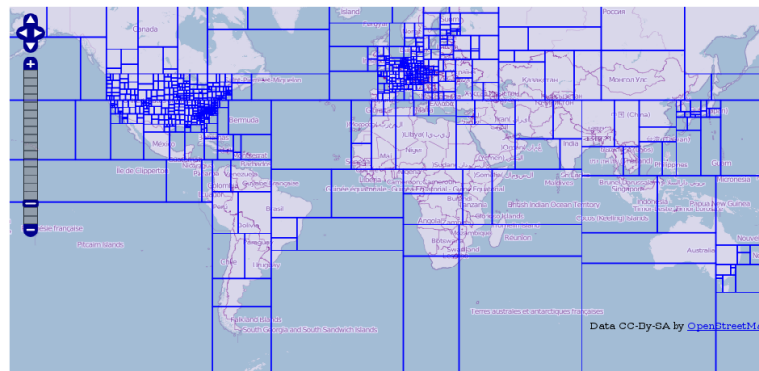
Adaptive Partitioning („Tiling“)

- Any tiling [Furtado 1999]
 - Cast into strategies
 - 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
```

- Why irregular tiling?

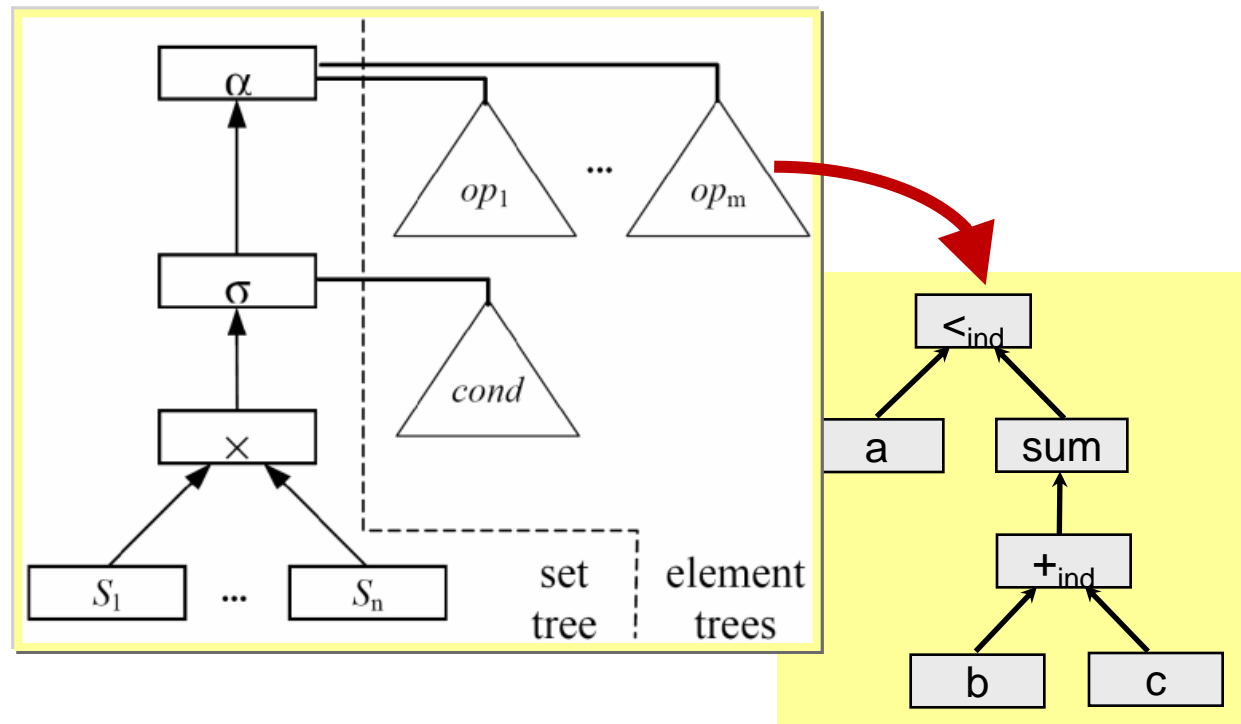


[OpenStreetMap]

Query Processing

- Clear separation: set vs array trees
 - Arrays as 2nd order attributes
- Optimization
- Tile-based evaluation

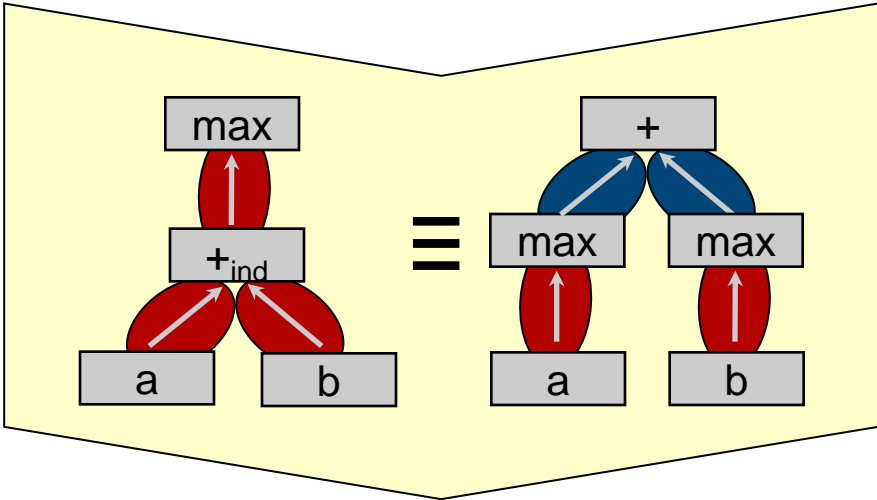
```
select a.array < sum_cells(
b.array + c.array )
from a, b, c
```



Query Optimization

[Ritsch 2000]

```
select max_cells( a + b )
from a, b
```



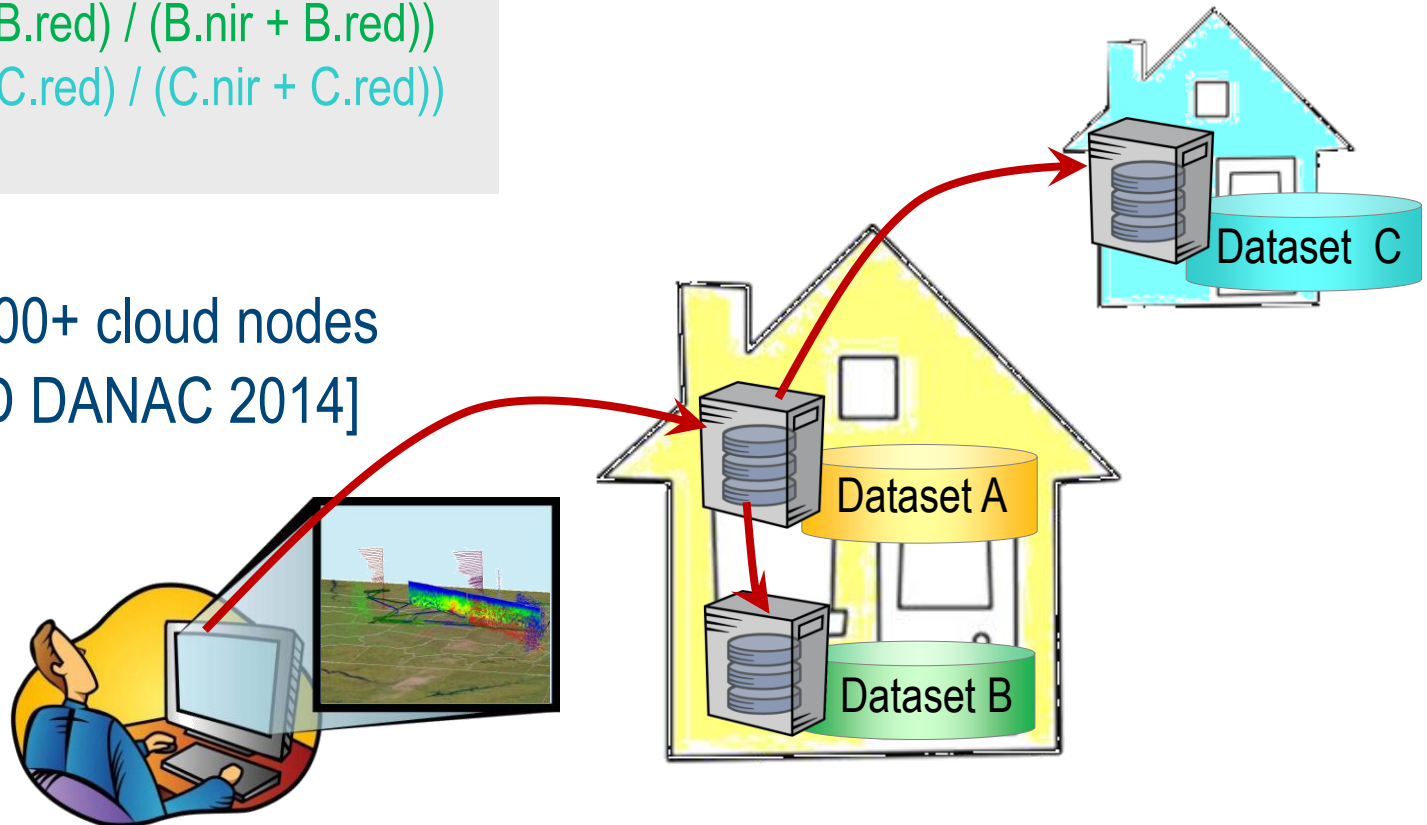
-  Tile stream
high traffic
-  Scalar stream
low traffic

```
select max_cells( a )
       + max_cells( b )
from a, b
```

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))
from A, B, C
```

1 query → 1,000+ cloud nodes
[ACM SIGMOD DANAC 2014]

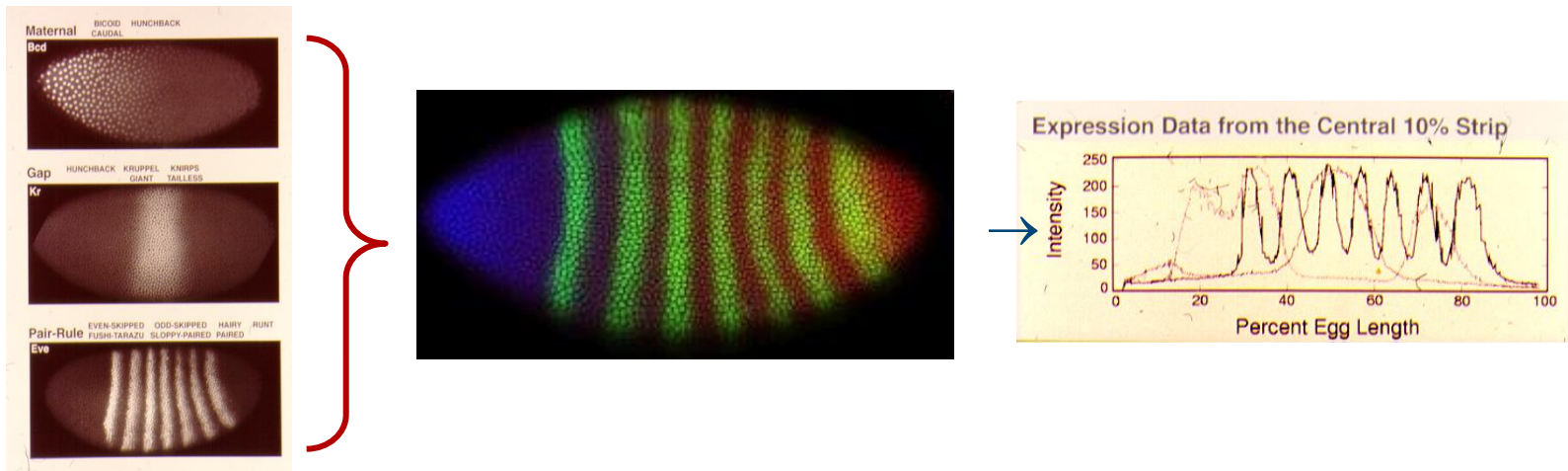


APPLICATIONS

Gene Expression Analysis

<http://urchin.spbcas.ru/Mooshka/>
[Samsonova et al]

- **Gene expression** = reading out genes for reproduction
- Research goal: capture spatio-temporal expression patterns in *Drosophila*



```
select encode( scale( {1c,0c,0c}*e[0,*,*,*:*]
                    +{0c,1c,0c}*e[1,*,*,*:*]
                    +{0c,0c,1c}*e[2,*,*,*:*], 0.2 ), „image/jpeg“ )
from EmbryoImages as e
where oid(e)=193537
```

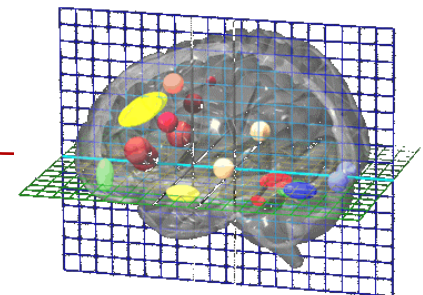
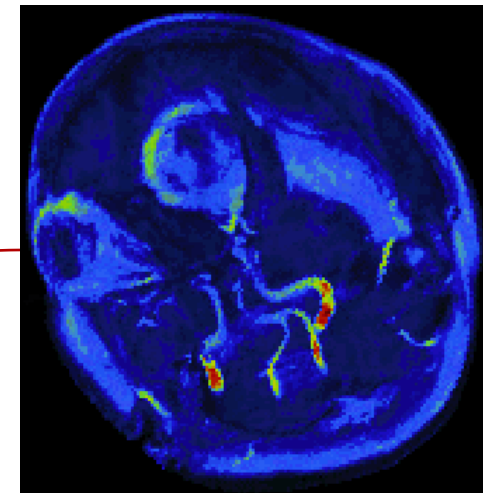
Human Brain Imaging

- Research goal: structural-functional relations in human brain
- Experiments → activity patterns (PET, fMRI)
 - Temperature, electrical, oxygen consumption, ...
 - → lots of computations → „activation maps“
- Example: “a parasagittal view of all scans containing critical Hippocampus activations, TIFF-coded.”

```

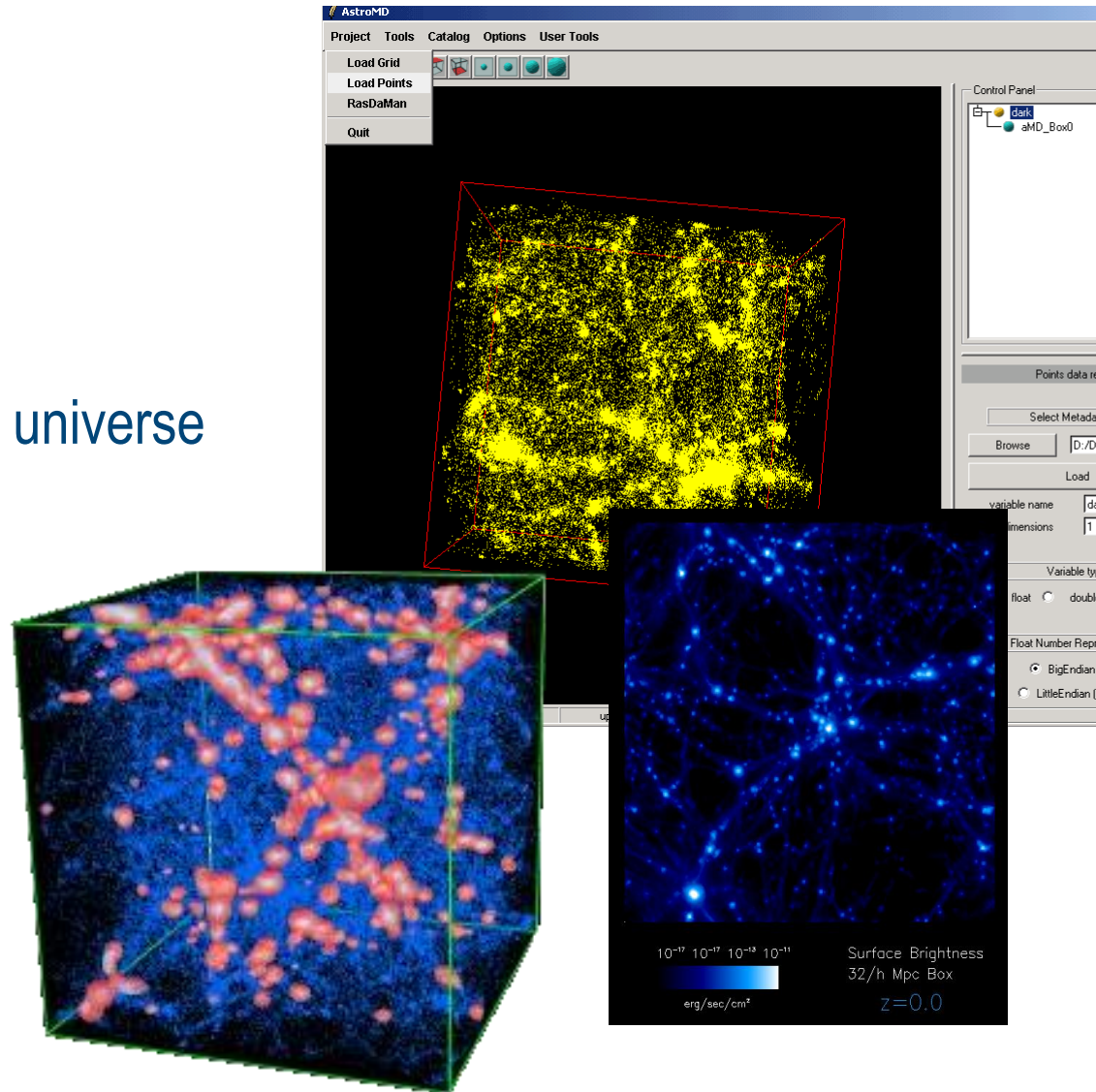
select tiff( ht[ $1, ** , ** ] )
from   HeadTomograms as ht,
       Hippocampus as mask
where  count_cells( ht > $2 and mask )
       / count_cells( mask )
       > $3
    
```

\$1 = slicing position, \$2 = intensity threshold value, \$3 = confidence



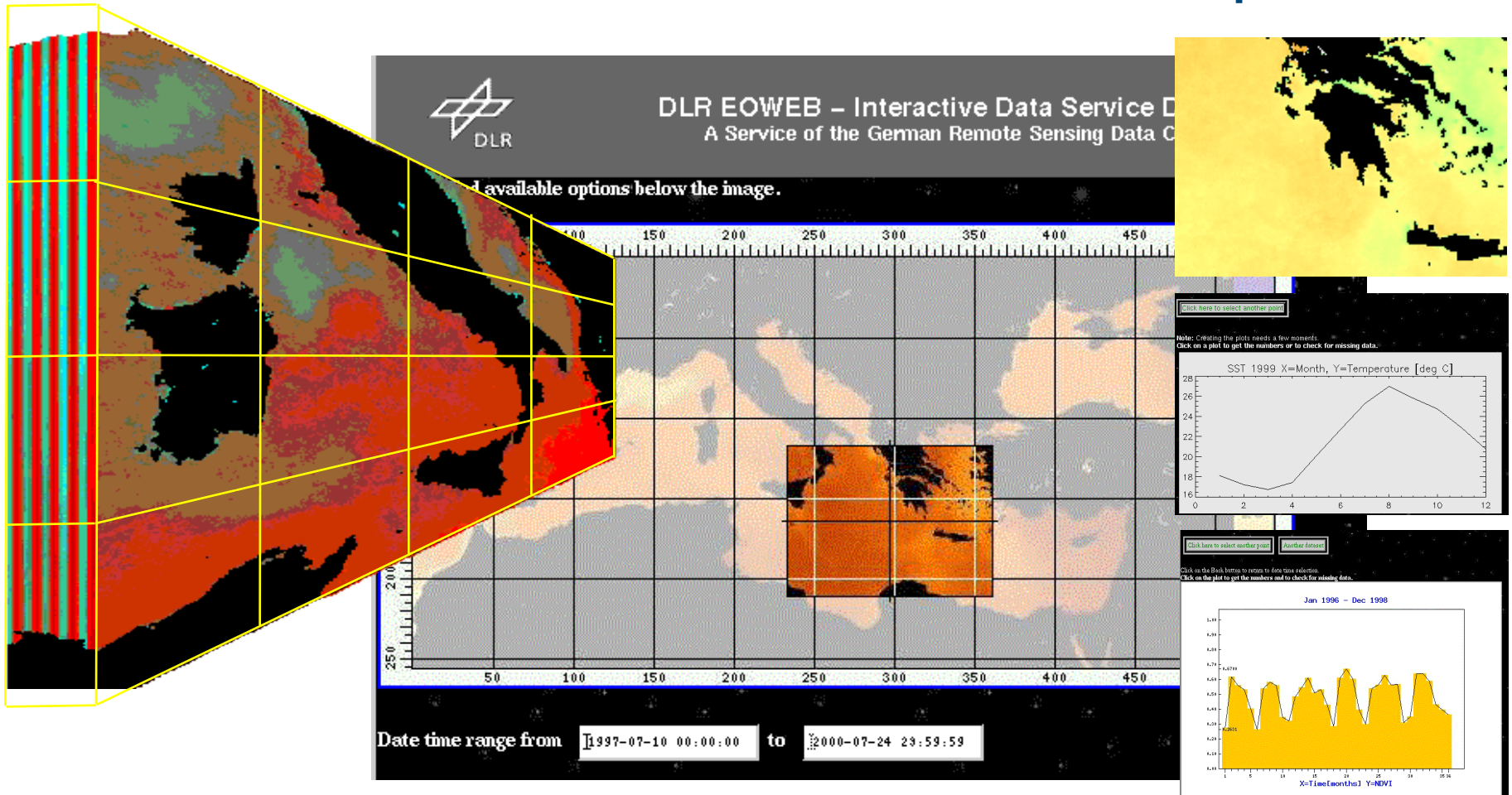
Cosmological Simulation

- Modelling domain: 4D
- Results: 3D/4D cutouts from universe
- Screenshots: AstroMD
[Gheller, Rossi 2001]



Early 3-D Service on rasdaman

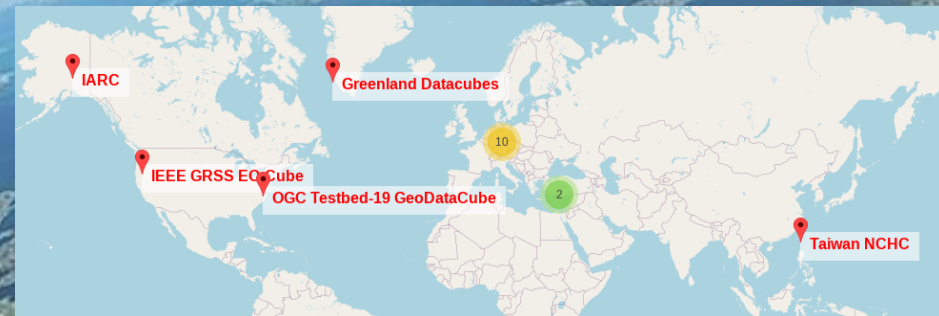
[Diedrich et al 2001]





EarthServer

- Agile Analytics on x/y/t + x/y/z/t Earth & Planetary datacubes
 - EU rasdaman + US NASA WorldWind
 - Rigorously standards as c/s APIs
- Multi-Petabyte, worldwide
 - participation free & open



earthserver.world

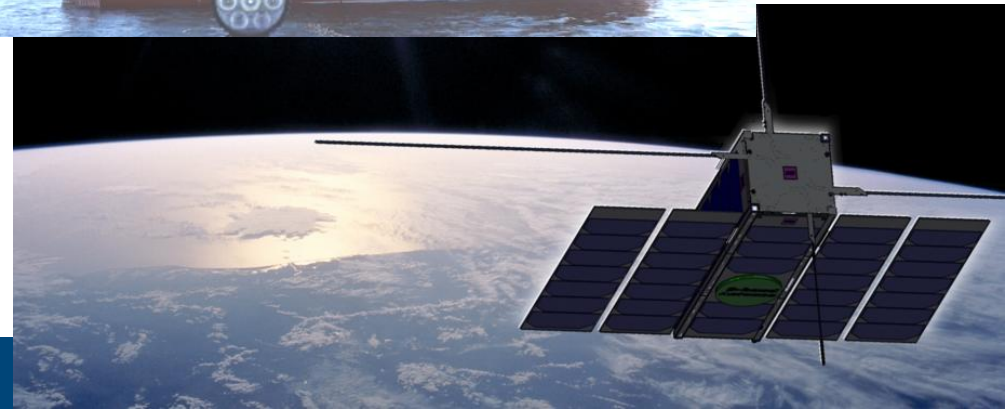
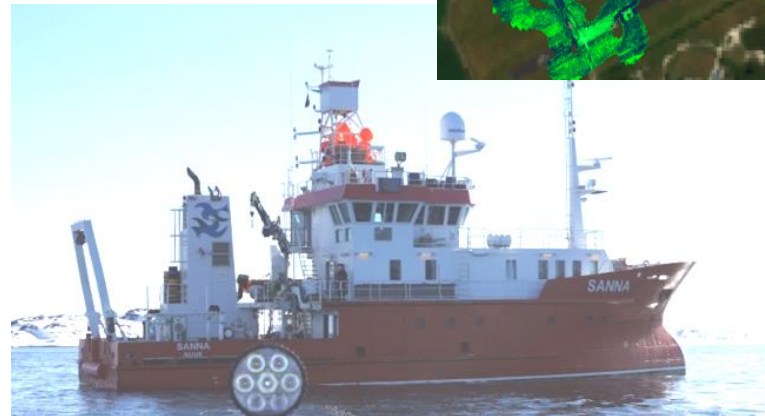
DEMO

Edge Integration & Fusion

- Airborne drone
 - Demonstrated at NATO C-UAS Exercise
 - Realtime downlink

- Ship
 - Demonstrated on research vessel, Nuuk/Greenland
 - Realtime send & receive

- Nanosat
 - Demonstrated in orbit
 - Faster: avoid full download
 - QoS: Deliver answers, not pixels



AI + Datacube Integration

- Seamless integration of **ML in datacube queries**

- Ex: model-based crop classification, fully integrated

```
for $s2 in (Sentinel_2),
    $m in (CropModel)
return encode( nn.predict( $s2[...], $m ), "tiff" )
```

- Natural Language Processing (RSVQA, Begüm Demir)



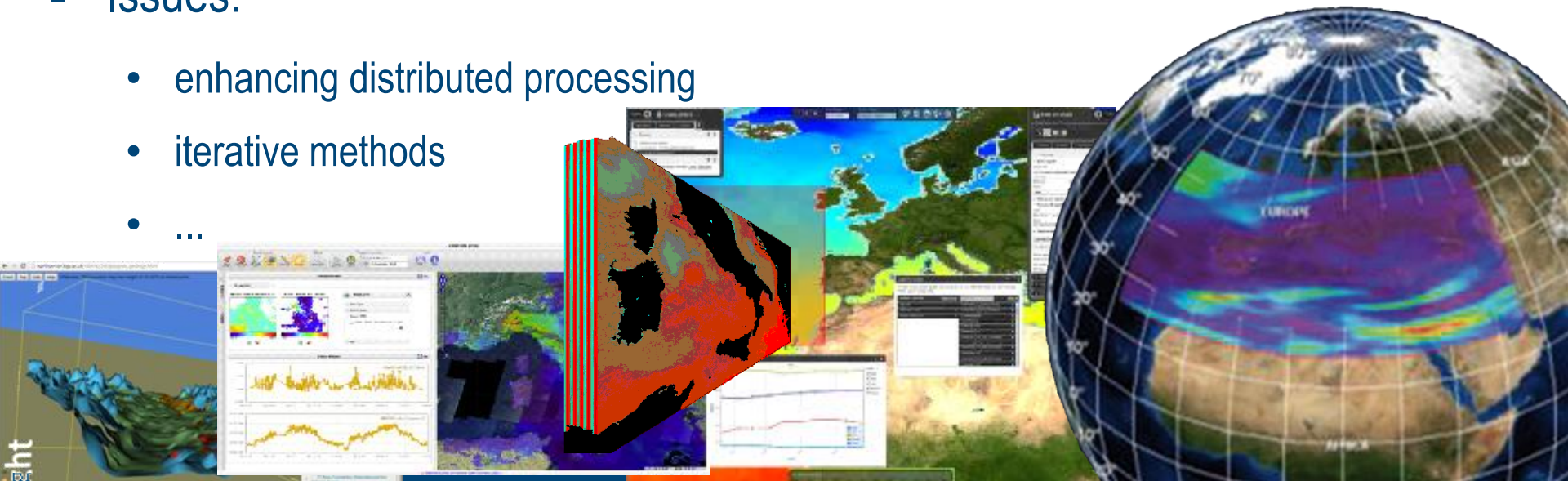
- WCPS Chatbot**
for geo datacube analytics



WRAP-UP

Summary

- Arrays are core data structure next to sets, graphs, hierarchies
 - sensor, image, simulation, statistics datacubes
- Array DBMS for declarative queries on massive n-D arrays
 - rasdaman
- Issues:
 - enhancing distributed processing
 - iterative methods
 - ...



Advertisement

- Seeking datacube coders
 - ~~Thesis~~ see my group's [current list of thesis topics](#) ~~Sabbatical in Spring 2025~~
 - Research projects
- Common requirement: strong coding skills
 - JavaScript / TypeScript / frameworks; Java; C++