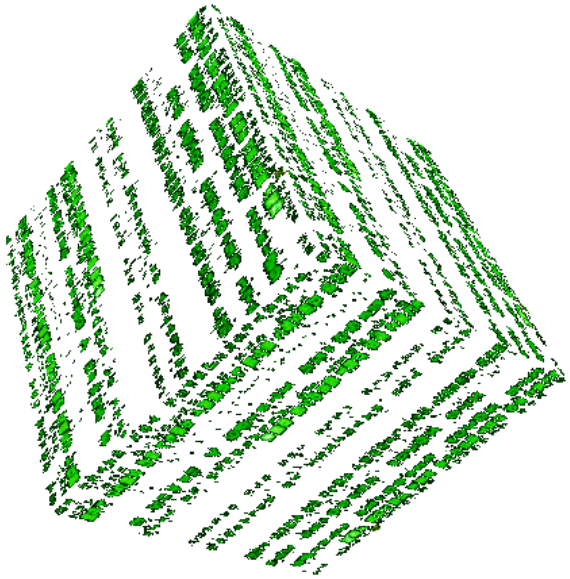


rasdaman:
Big Data(Cubes)

Peter Baumann

Jacobs University | rasdaman GmbH

rasdaman: Agile Datacube Analytics

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

- Mature, operational, on OSGeo Live
  - Multi-Petabyte databases, 1000x parallelization, federation
- ESA 2017: “world leading”,
  “standard working horse for OGC standardisation“
- OGC, ISO, INSPIRE datacube standards
  crafted by rasdaman team
  - Reference implementation

Array SQL

rasdaman raster data manager

JACOBS UNIVERSITY

[SSDBM 2014, ACM DOLAP 2015]

Information technology — Database languages — SQL —

Part 15:
Multi-Dimensional Arrays (SQL/MDA)

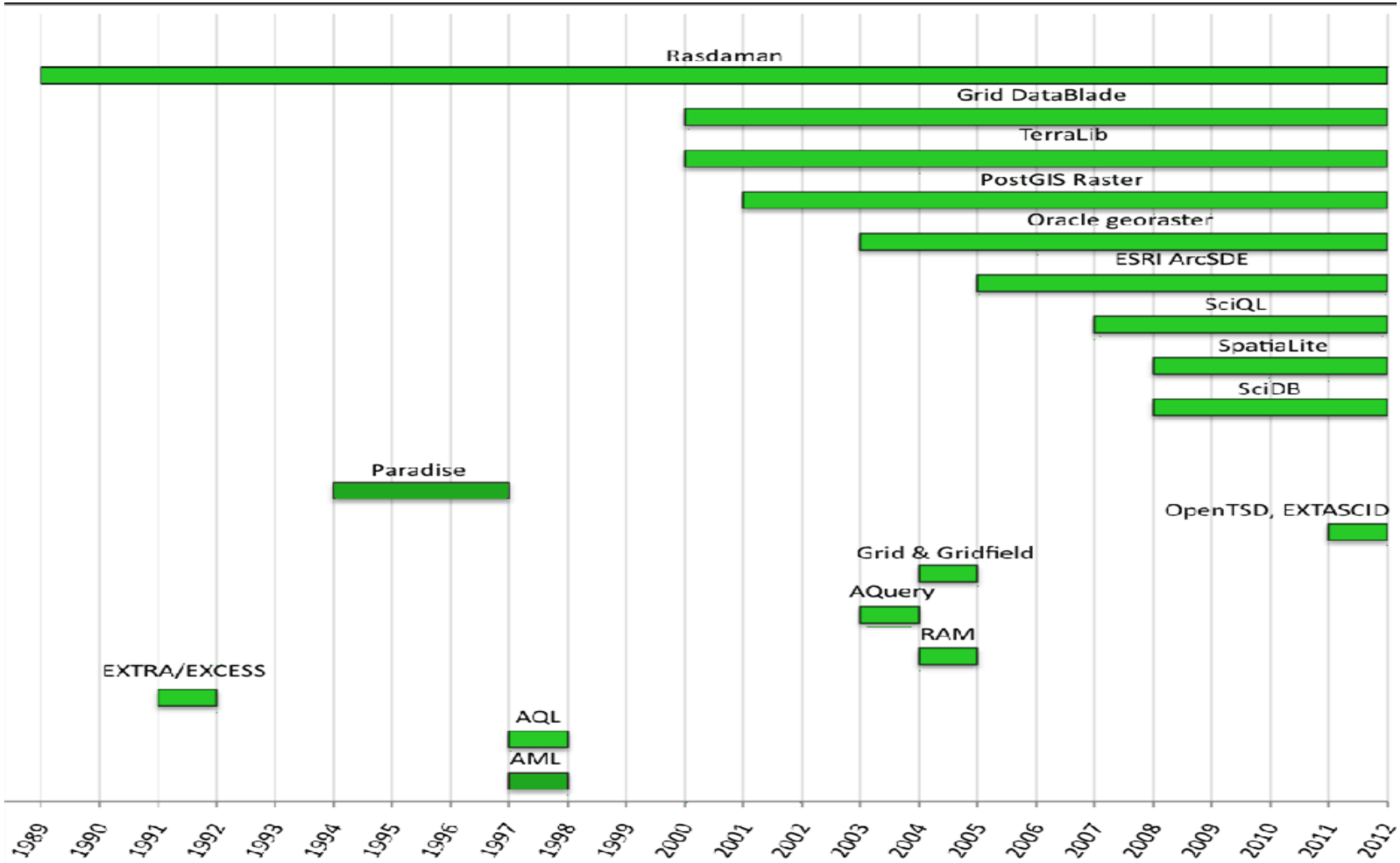
Technologies de l'information — Langages de base de données — SQL —

Partie 15: Tableaux multi-dimensionnels (SQL/MDA)

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

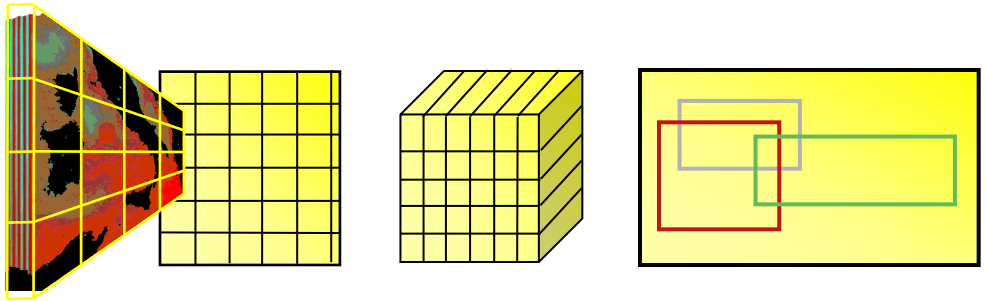
```
select  id, encode(scene.band1-scene.band2)/(scene.nband1+scene.band2)), „image/tiff“ )
from    LandsatScenes
where acquired between „1990-06-01“ and „1990-06-30“ and
          avg( scene.band3-scene.band4)/(scene.band3+scene.band4)) > 0
```

A Brief History of Array Databases



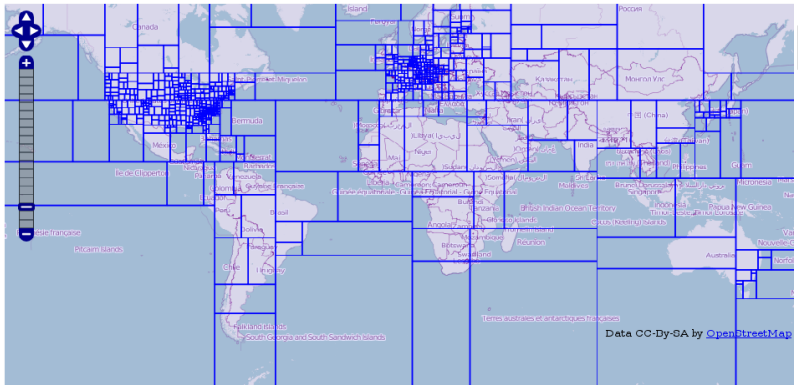
Datacube Scalability

- Adaptive data partitioning & distribution
  - *130+ TB datacubes* [IJDE 2015]



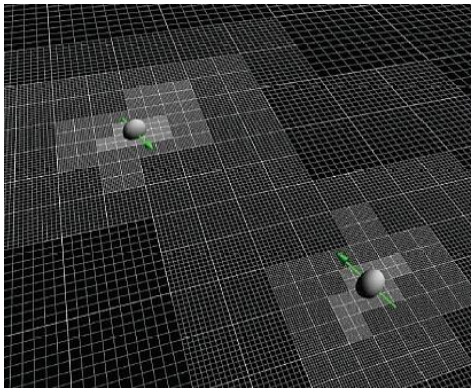
```
insert into MyCollection
   values ...
   tiling area of interest [0:20,0:40], [45:80,80:85]
   tile size 1000000
```

- Why irregular tiling?



[OpenStreetMap]

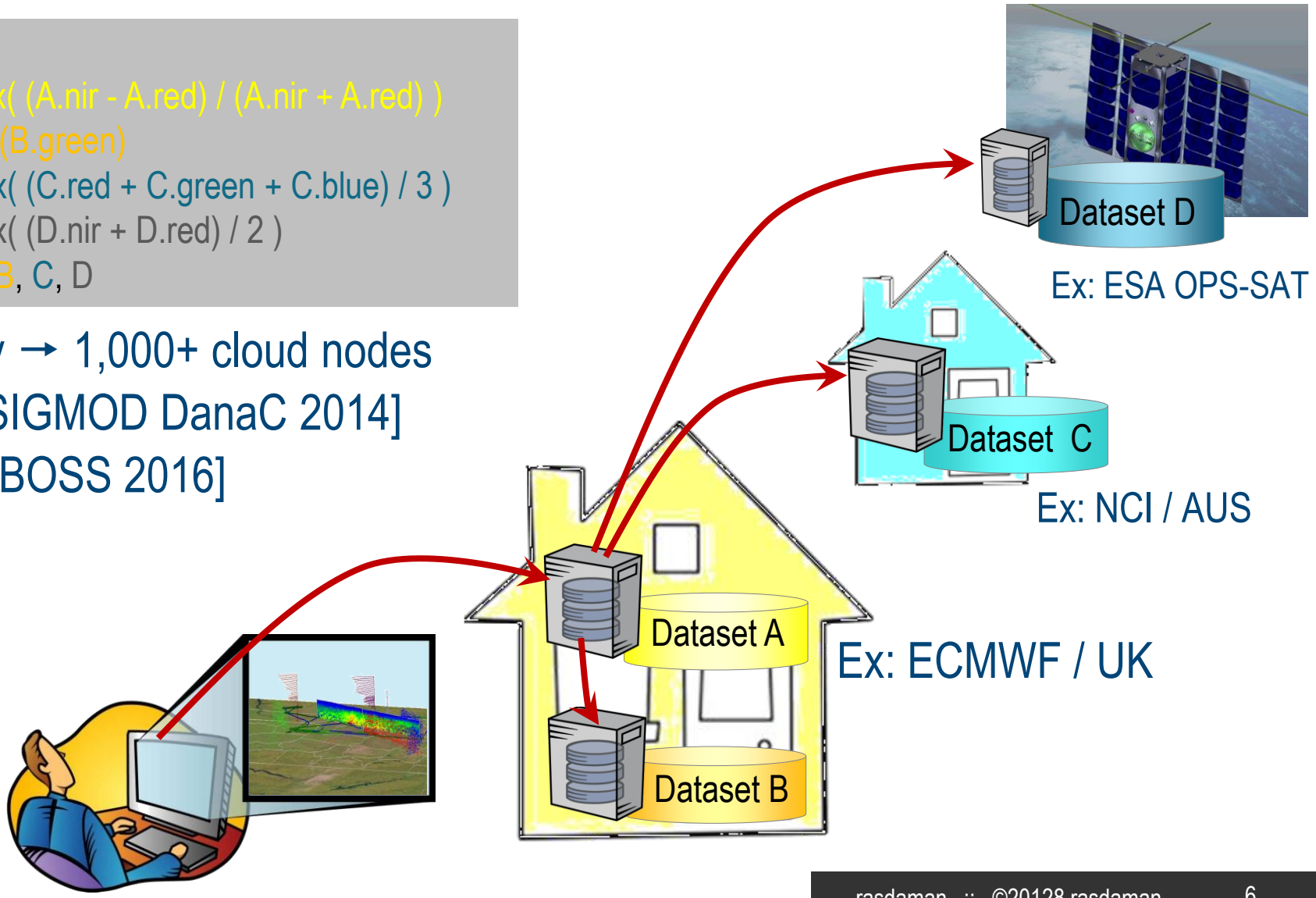
[Centrella et al: scidacreviews.org]



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
```

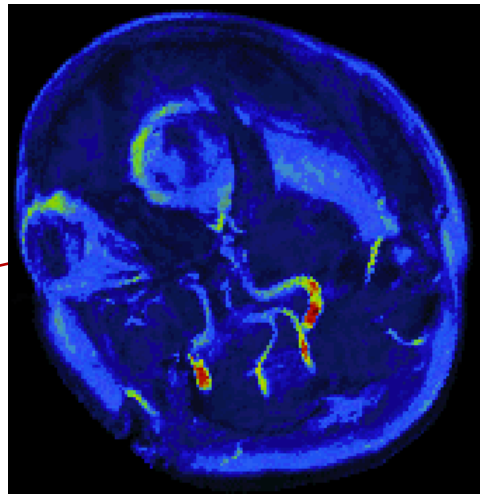
1 query → 1,000+ cloud nodes
[ACM SIGMOD DanaC 2014]
[VLDB BOSS 2016]

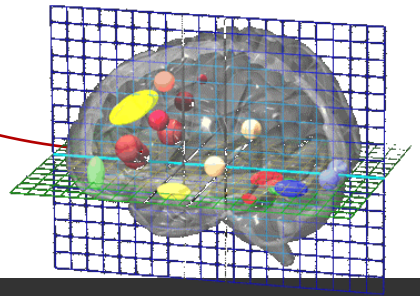


Human Brain Imaging

- Research goal: to understand structural-functional relations in human brain
- Experiments capture 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
```



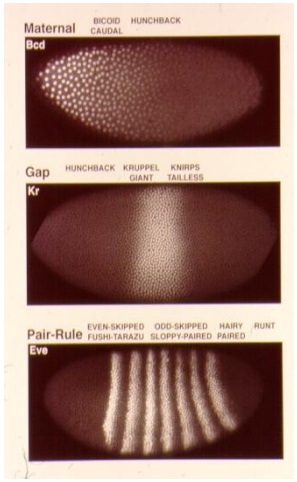


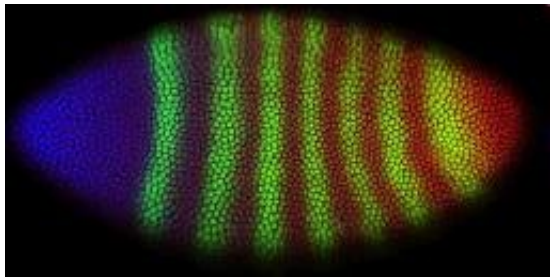
`$1` = slicing position, `$2` = intensity threshold value, `$3` = confidence

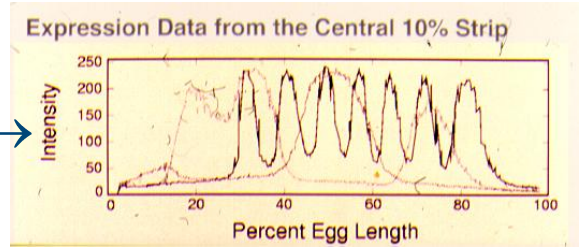
Gene Expression Analysis

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

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



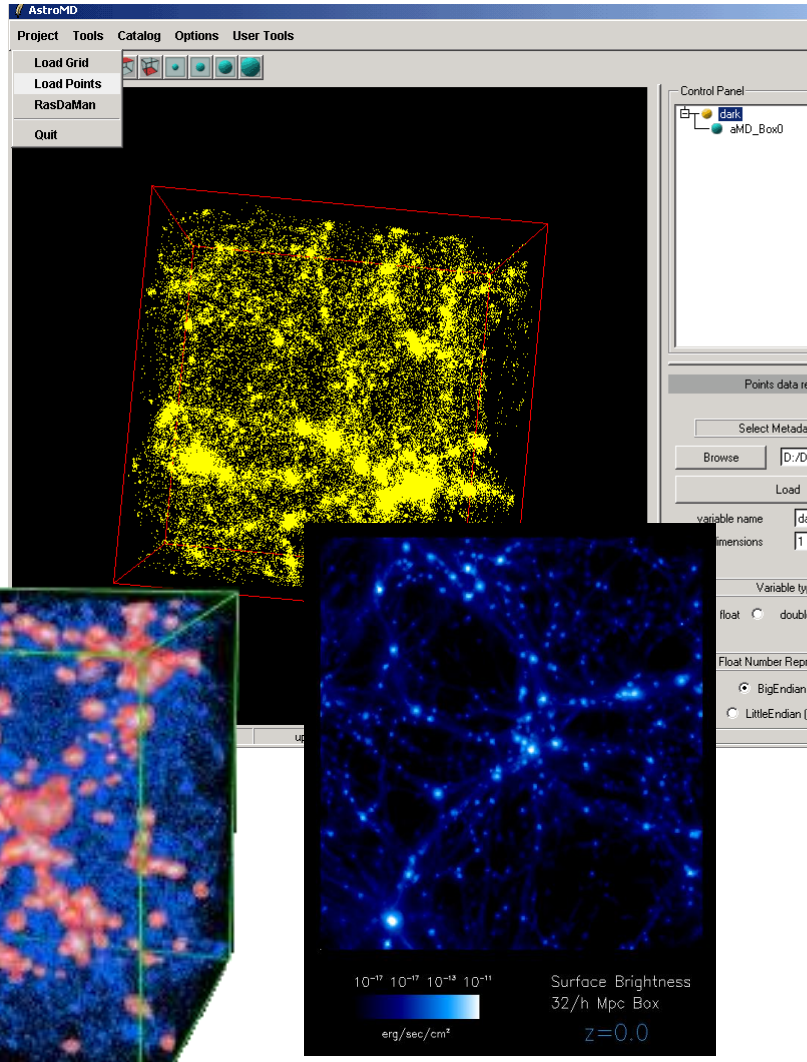





```
select jpeg( scale( {1c,0c,0c}*e[0,*:*,*:*]
                  +{0c,1c,0c}*e[1,*:*,*:*]
                  +{0c,0c,1c}*e[2,*:*,*:*], 0.2 ) )
from EmbryoImages as e
where oid(e)=193537
```

Cosmological Simulation

- Modelling domain: 4D
  - Dark matter (highest mass factor in universe)
  - Baryonic matter (stars, gas, dust, …)
  - → Coupled simulation: particle + fluid
- Results: 3D/4D cutouts from universe
  - Eg, 64 Mpc^3
    (1 pc = 3.27 light years)
- Screenshots: AstroMD
  [Gheller, Rossi 2001]



Seeking Geeks

- Required:
  - Strong coding skills
  - Motivation; responsibility; diligence; team worker
- Rewards:
  - Gain insight into Petascale "Scientific Analytics as a Service"
  - Experience in practice & tools for large-scale sw development
  - Collaborate with experts worldwide (US, AUS, FR, IT, UK, …)
  - Internship → thesis
  - Big plus in CV