7/16/2012

Tamás Budavári (Johns Hopkins University)

SPATIAL SEARCHES IN ASTRONOMY DATABASES

MULTI-DIMENSIONAL INDEXING FOR SIMULATIONS AND OBSERVATIONS

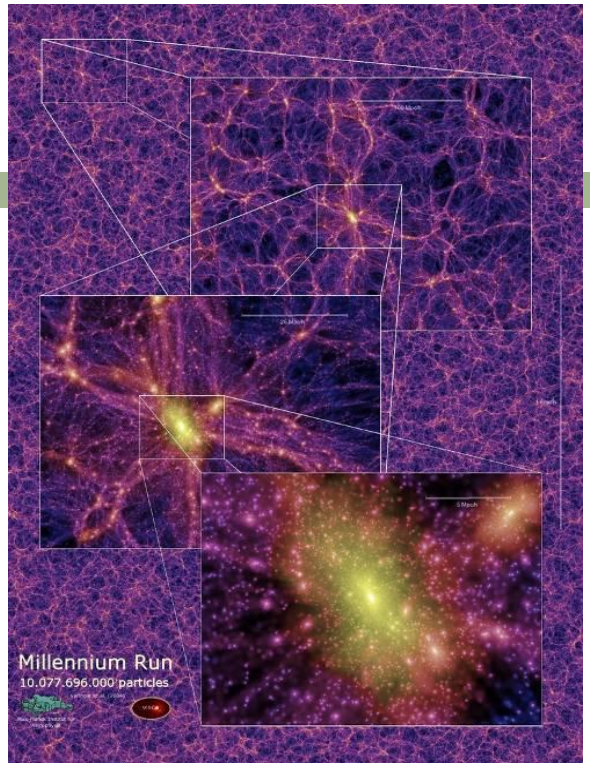
7/16/2012

Tamás Budavári (Johns Hopkins University)

Storing Simulations

- Millennium Run (MPA)
  - 10 billion particles, 64 snapshots
  - FoF groups and merger trees
- Millennium XXL
  - 300 billion particles
- MultiDark – Bolshoi

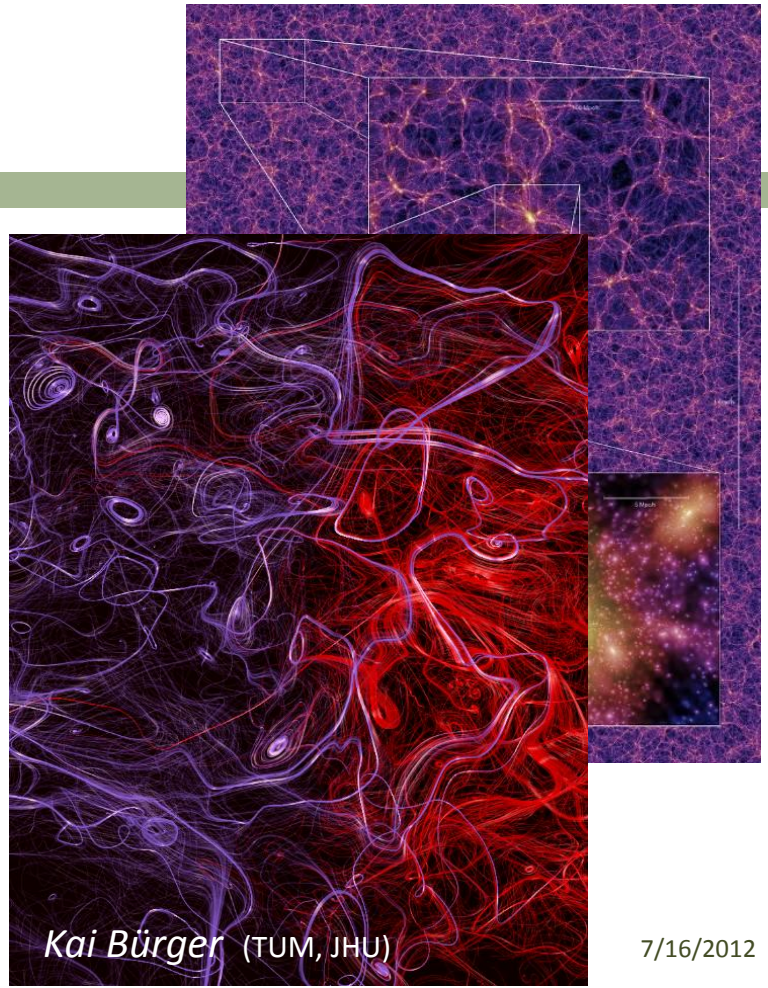
- Turbulence simulations (JHU)
  - 1024^4 grid, 27TB



Storing Simulations

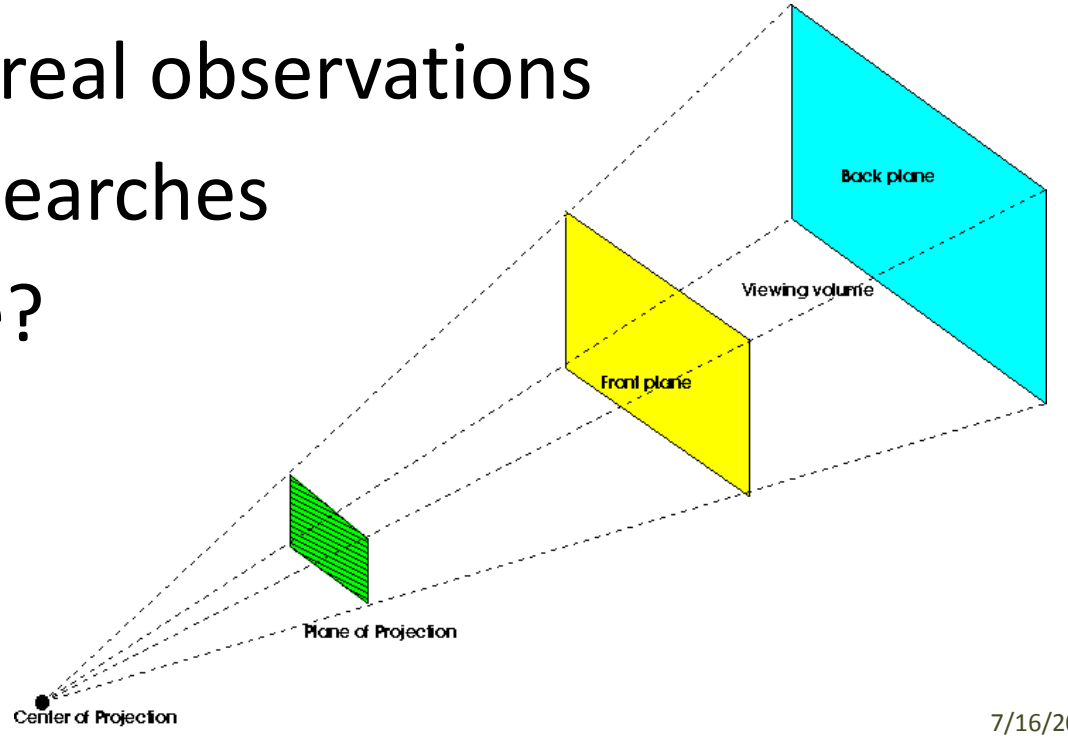
- Millennium Run (MPA)
  - 10 billion particles, 64 snapshots
  - FoF groups and merger trees
- Millennium XXL
  - 300 billion particles
- MultiDark – Bolshoi

- Turbulence simulations (JHU)
  - 1024^4 grid, 27TB



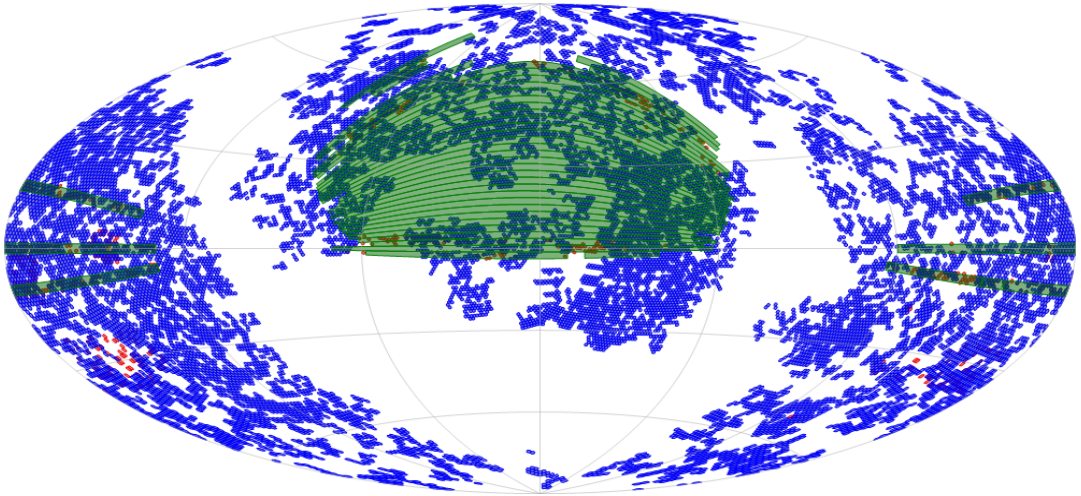
Observing Simulations

- Comparison to real observations
- Lots of spatial searches
- In the database?



Sky Coverage

- For precise window function
  - Virtual surveys



Outline

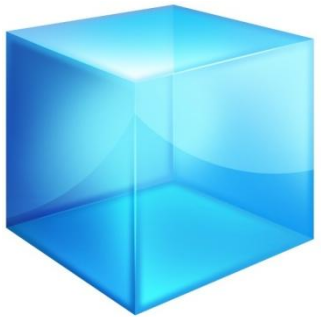
- Query shapes in SQL
- Indexing with space-filling curve
- Combine for spatial searches
  - Periodic boxes
  - Celestial sphere

Databases

- Which one to use depends on the task
  - Sqlite, MySQL, PostGRES, DB2, Oracle, SQL Server
    - Free “express versions” of the big ones, too
- Customization is a must
  - There is always something missing
    - Extend by loading your libraries

Query Shapes

- Geometric primitives
  - Sphere, Box, Cone…

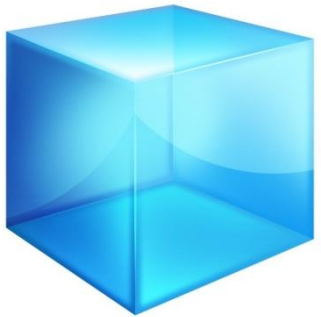


Query Shapes

- IShape interface

```
TopoPoint Contains(Point p);
TopoShape GetTopo(Box b);
Box GetBoundingBox();
```

- Geometric primitives
  - Sphere, Box, Cone…

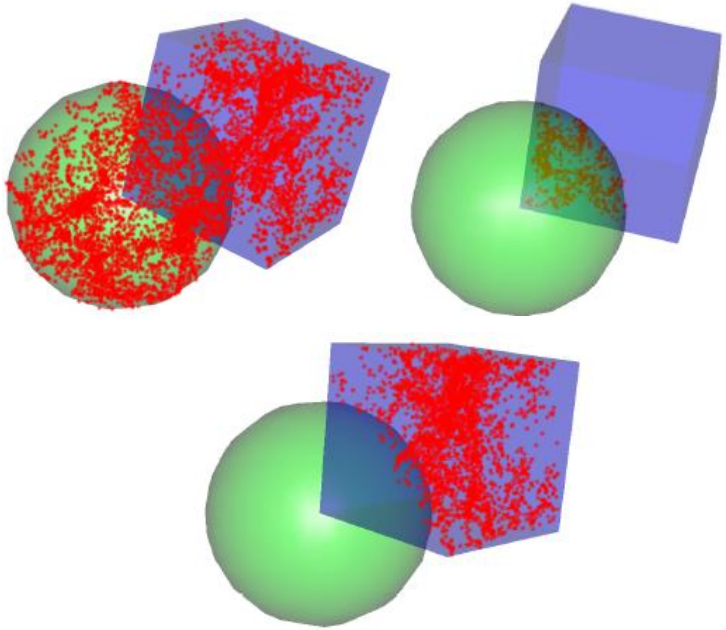


Query Shapes

- IShape interface

```
TopoPoint Contains(Point p);
TopoShape GetTopo(Box b);
Box GetBoundingBox();
```

- Composites
  - Intersect, Union, Difference…



Query Shapes

- In SQL
  - UDT


```
/* Sphere */
declare @s Sphere = Sphere::New(1,2,3,10);
-- Check if a point is inside
select  @s.ContainsPoint(1,2,3), @s.ContainsPoint(99,0,0);
go


/* Box */
declare @b Box = Box::New(0,0,0,10,10,10);
select  @b.ContainsPoint(1,2,3), @b.ContainsPoint(99,0,0);
select  @b.ToString(); -- string representation
go


/* String Representation */
declare @x Box = 'BOX [0,0,0, 10,10,10]'
select  @x.ContainsPoint(1,2,3), @x.ContainsPoint(99,0,0);
go
```

Query Shapes

- Generic
  - UDT
- Boolean
  - Methods

```
/* Generic Shapes */
declare @a Shape = 'BOX [0,0,0, 10,10,10]';
select  @a.ContainsPoint(1,2,3), @a.ContainsPoint(99,0,0)
select  @a.ToString();
go


/* Boolean Algebra */
declare @s1 Shape = 'BOX [0,0,0, 10,10,10]';
declare @s2 Shape = 'SPHERE [0,0,0, 5]';
declare @sU Shape = Shape::NewUnion(@s1,@s2);
declare @sI Shape = Shape::NewIntersection(@s1,@s2);
declare @sD Shape = Shape::NewDifference(@s1,@s2);
select  @sU.ToString() union all
select  @sI.ToString() union all
select  @sD.ToString();
go
```

Query Shapes

- Generic
  - UDT
- Boolean

```
/* Generic Shapes */
declare @a Shape = 'BOX [0,0,0, 10,10,10]';
select  @a.ContainsPoint(1,2,3), @a.ContainsPoint(99,0,0)
select  @a.ToString();
go


/* Boolean Algebra */
```

```
/* Using the parser */
declare @u Shape, @i Shape, @d Shape;
select  @u = 'UNION        [BOX[0,0,0,2,2,2], SPHERE[0,0,0,2]]',
        @i = 'INTERSECTION [BOX[0,0,0,2,2,2], SPHERE[0,0,0,2]]',
        @d = 'DIFFERENCE   [BOX[0,0,0,2,2,2], SPHERE[0,0,0,2]]';
```

```
select  @sD.ToString();
go
```

Indexing Tables

- Better performance of queries
  - Instantaneous range searches
  - Fast JOINs
- Syntax

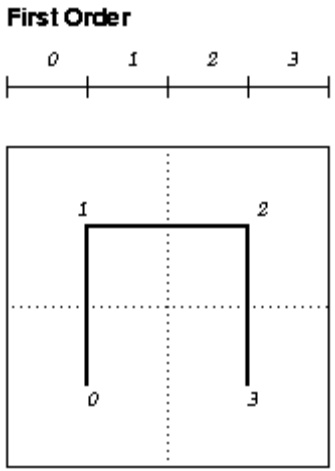
```
CREATE INDEX ix_Name ON Table
    (X ASC, …) INCUDE (V, …)
```

Multi-Dimensional

- Map the space to a simple index
- Different kinds of Space-Filling Curves
  - Morton's Z-curve
  - Peano-Hilbert Curve

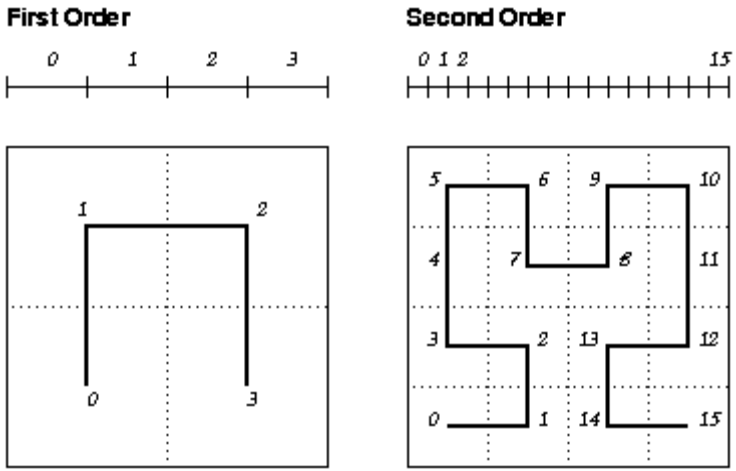
Peano-Hilbert Curve

- Hierarchical space filling



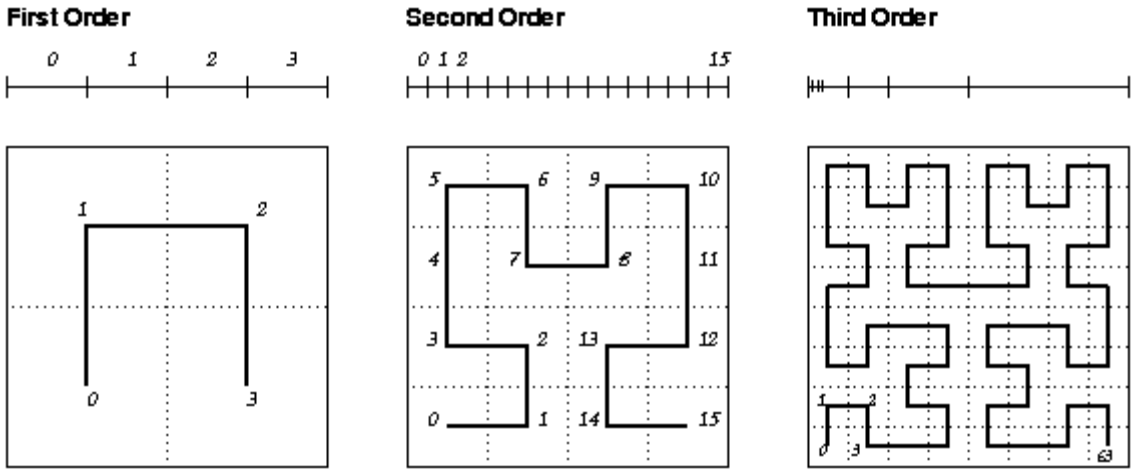
Peano-Hilbert Curve

- Hierarchical space filling



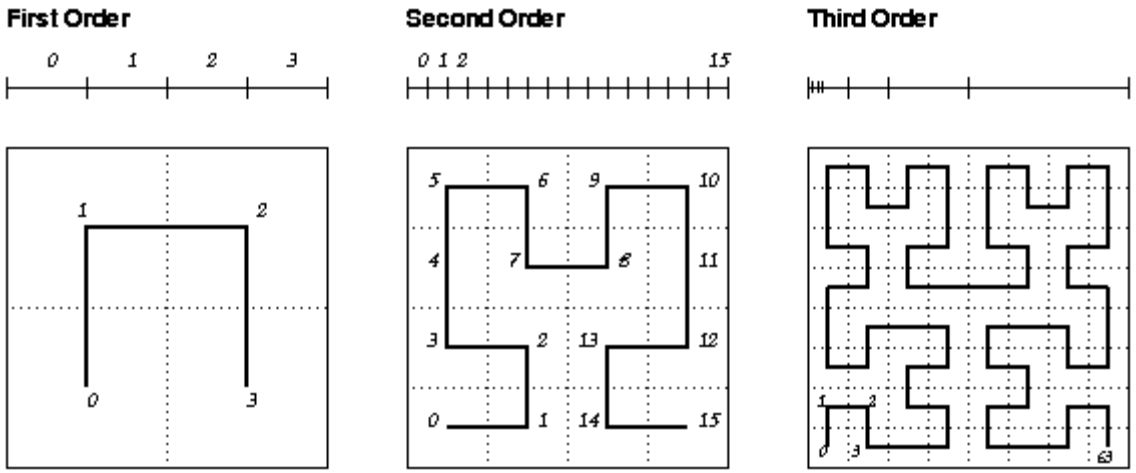
Peano-Hilbert Curve

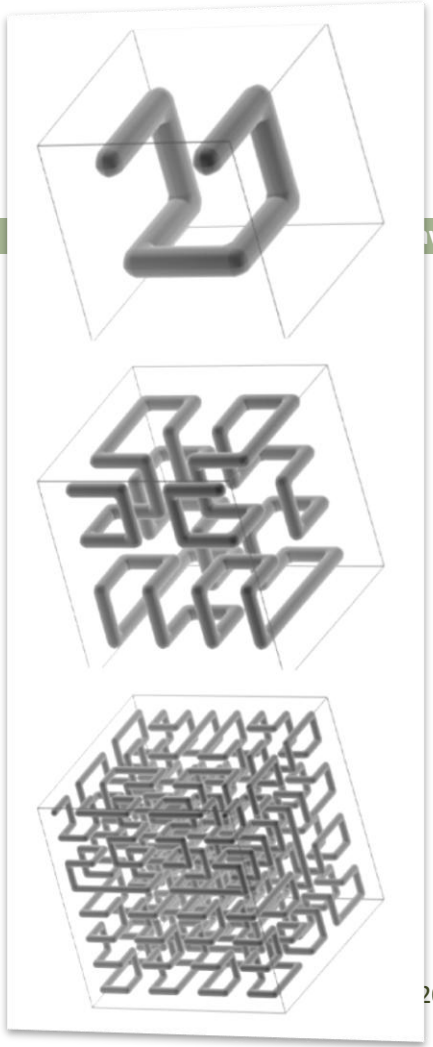
- Hierarchical space filling



Peano-Hilbert Curve

- Hierarchical space filling

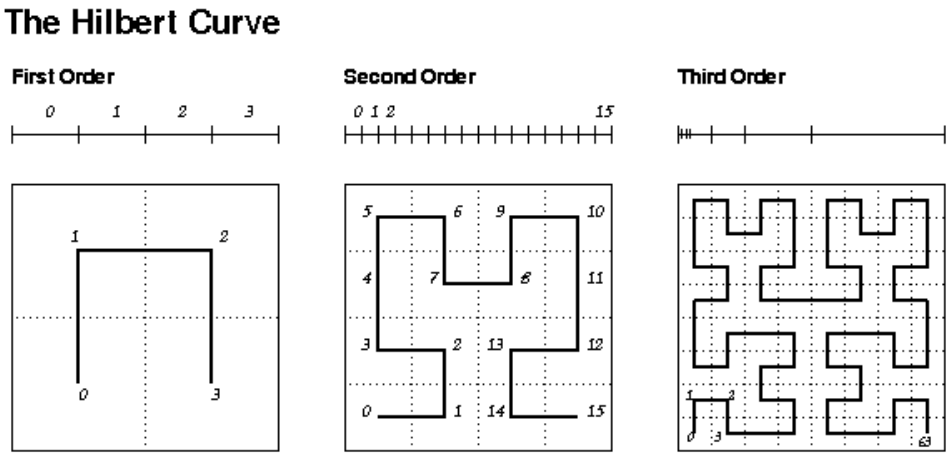


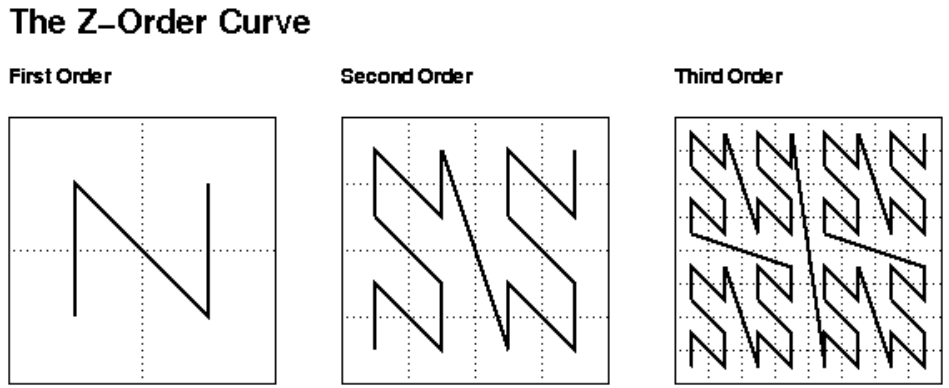


Also others…

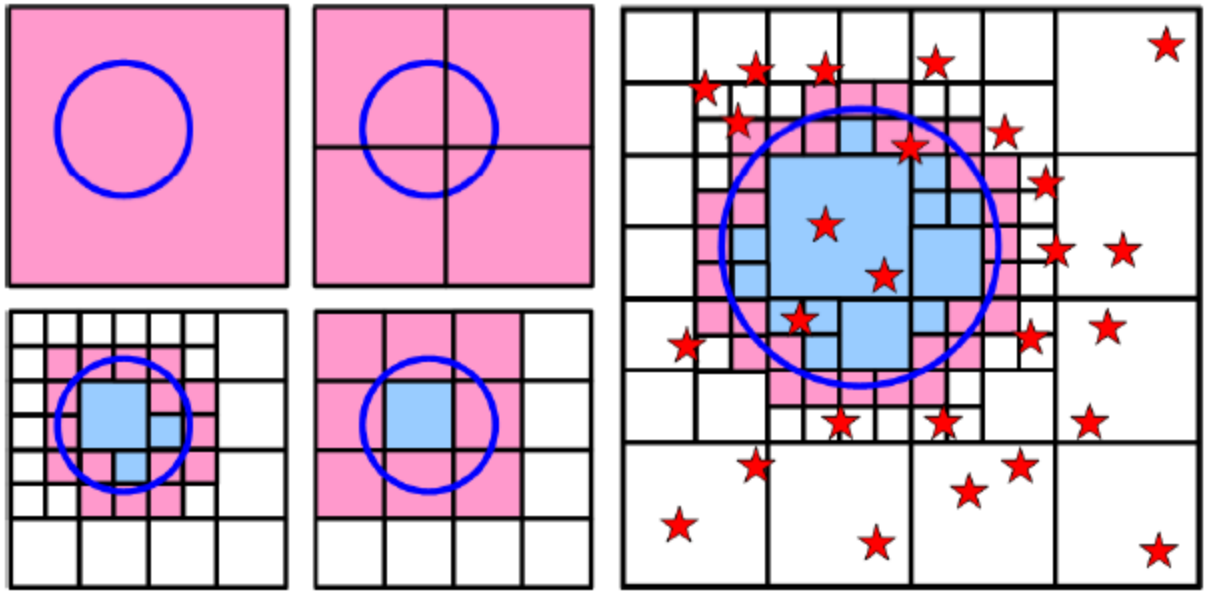
- Morton Z-order
  - Simple bit interleave
- Etc…

- Which one to use?
  - Statistical analyses
    - Correlation fn



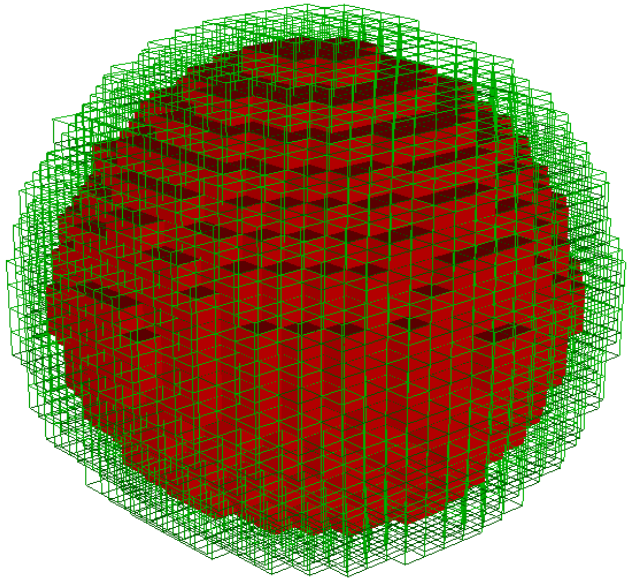


Divide and Conquer



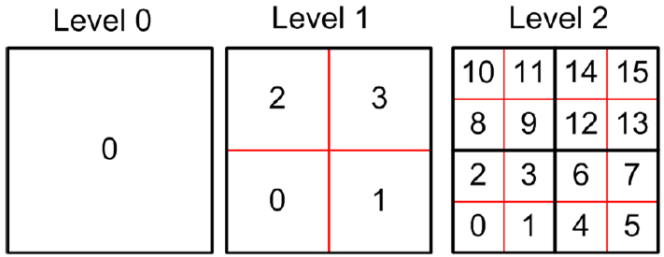
Covers for Shapes

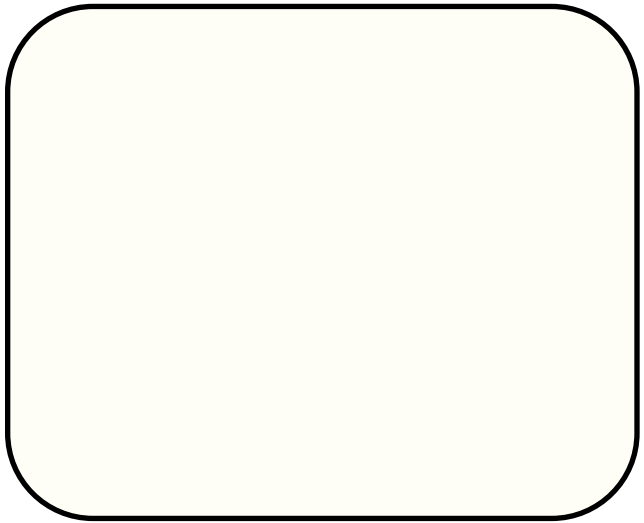
- Inside approximation
- Outside overshoot



Covers for Shapes

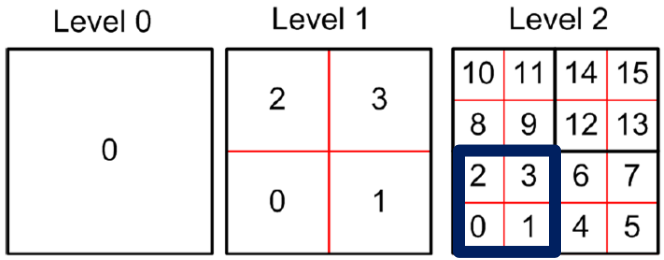
- Inside approximation
- Outside overshoot
  - They are Key ranges





Covers for Shapes

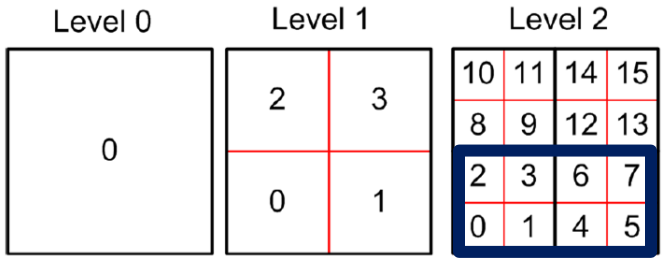
- Inside approximation
- Outside overshoot
  - They are Key ranges




```
Key between 0 and 3
```

Covers for Shapes

- Inside approximation
- Outside overshoot
  - They are Key ranges

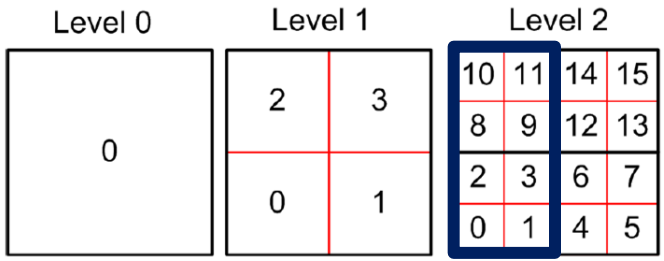



```
Key between 0 and 3

Key between 0 and 7
```

Covers for Shapes

- Inside approximation
- Outside overshoot
  - They are Key ranges



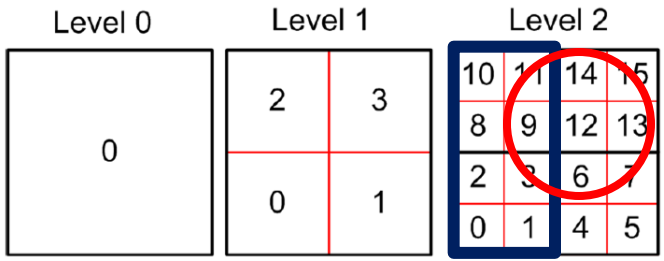

```
Key between 0 and 3

Key between 0 and 7

Key between 0 and 3
or
Key between 8 and 11
```

Covers for Shapes

- Inside approximation
- Outside overshoot
  - They are Key ranges



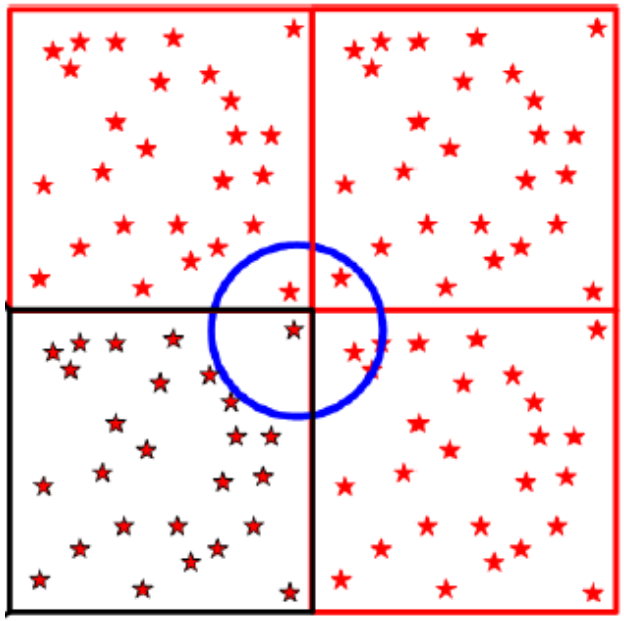

```
Key between 0 and 3

Key between 0 and 7

Key between 0 and 3
or
Key between 8 and 11
```

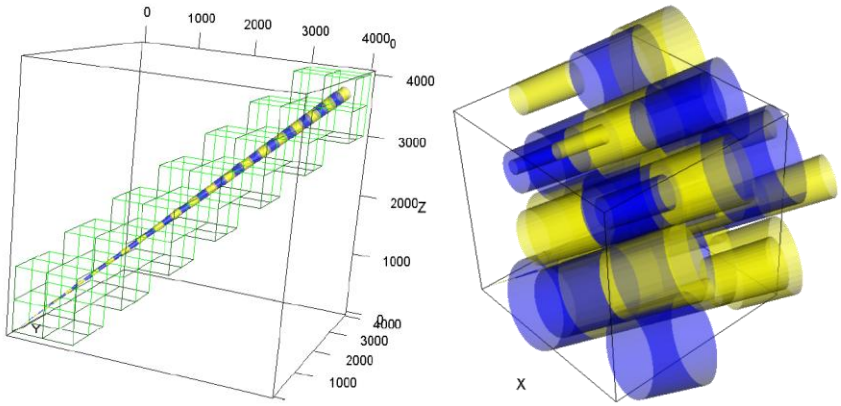
Periodic Boundaries

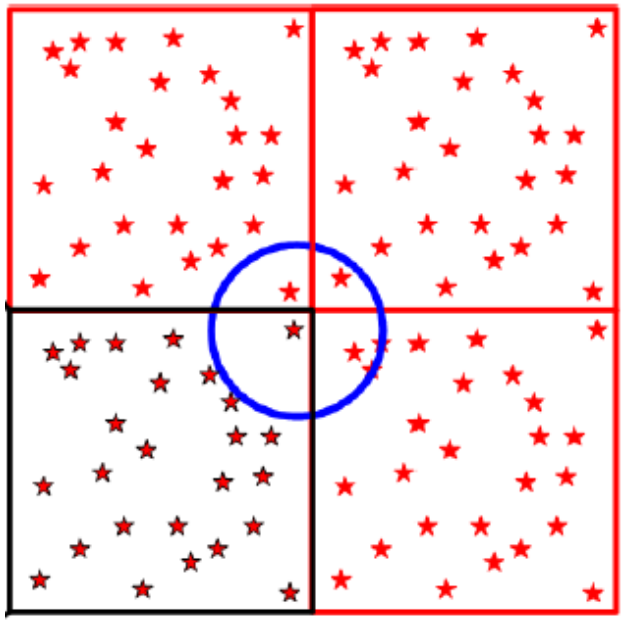
- Infinite with periodicity
  - Have to search all boxes



Periodic Boundaries

- Infinite with periodicity
  - Have to search all boxes





Searching in SQL

- Key filter
  - By Cover
- ShiftX,-Y,-Z
  - Where?

```
/* Peano-Hilbert Cover */
select * from fSimulationCover( sims.MilliMil(),
                                'SPHERE[0,10,10,5]',5);
go


/* Cover translates the search onto built-in index */
declare @s Shape = 'SPHERE[0,10,10,5]';
with Cover(FullOnly, KeyMin,KeyMax, ShiftX,ShiftY,ShiftZ)
as (
    select * from fSimulationCover(sims.MilliMil(),@s,5)
    where FullOnly = 1 -- inside cover
)
select h.HaloID, h.X+c.ShiftX as X,
                 h.Y+c.ShiftY as Y,
                 h.Z+c.ShiftZ as Z
from MilliMil..MpaHalo h inner join Cover c
    on h.PHKey between c.KeyMin and c.KeyMax
where SnapNum = 63
```

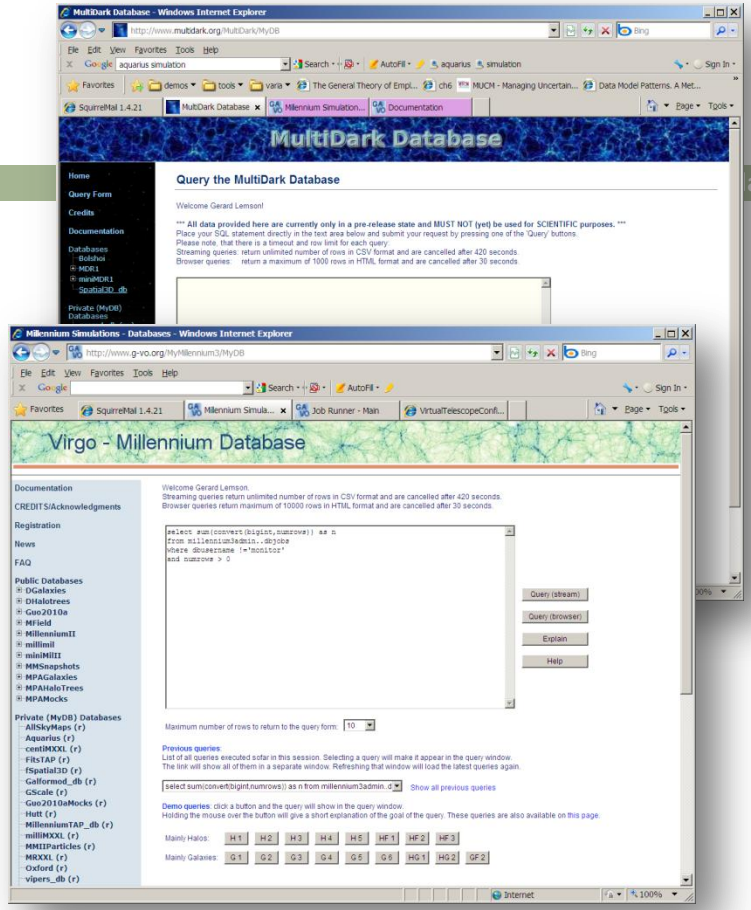
Real!

- E.g.,

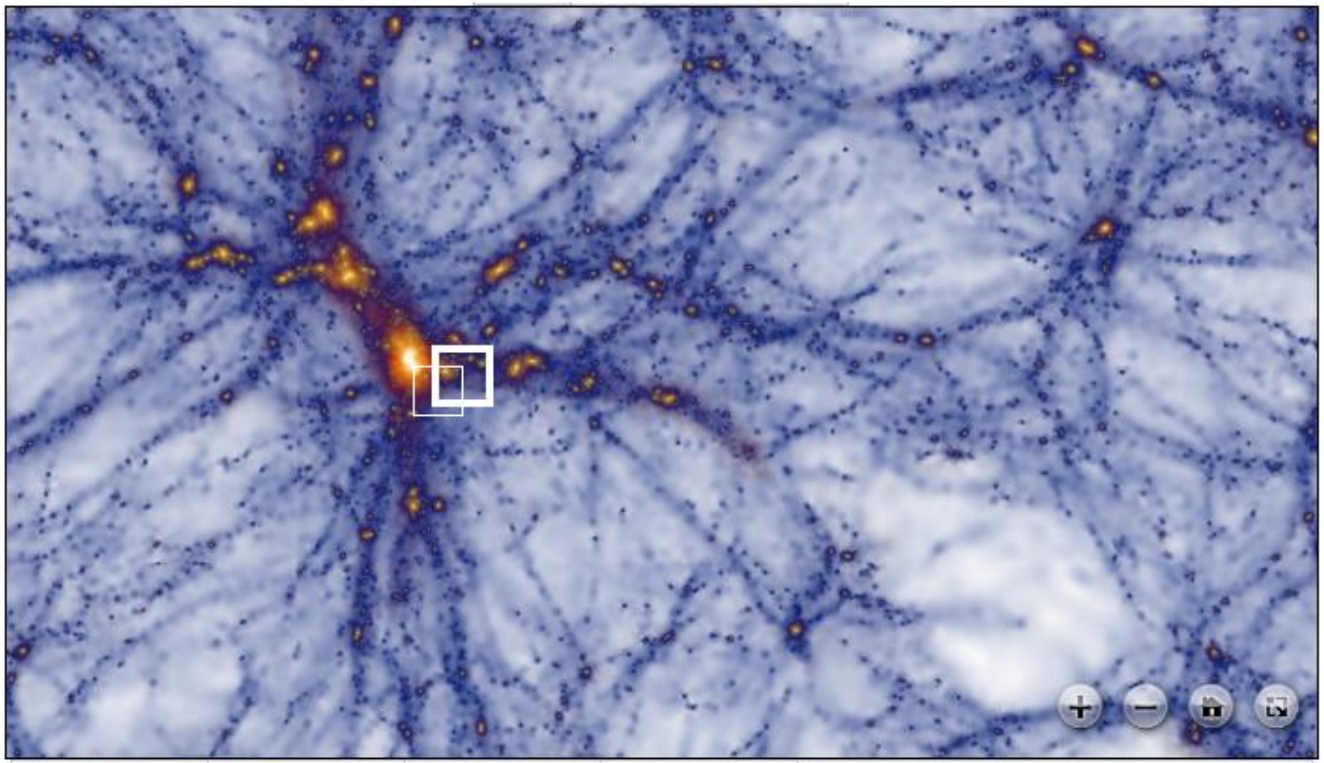
```
/* Multiple searches around POI */
with QueryShapes (FoFID,Shape) as
(
    select top 10 FoFID, Shape::NewSphere(X,Y,Z,10)
    from MilliMil..FoF
    where SnapNum=63 order by M_TopHat200 desc
)
select distinct s.FoFID, g.GalaxyID, g.X+c.ShiftX as X,
                                     g.Y+c.ShiftY as Y,
                                     g.Z+c.ShiftZ as Z
from QueryShapes s
    cross apply fSimulationCover(sims.MilliMil(),s.Shape,5) c
    inner join MilliMil..DeLucia2006A g
        on g.PHKey between c.KeyMin and c.KeyMax
where g.SnapNum=63
    and ( (c.FullOnly=1) -- Inner cover
        or
          (c.FullOnly=0  -- Boundary cover
           and 1=s.Shape.ContainsPoint(g.X+c.ShiftX,
                                       g.Y+c.ShiftY,
                                       g.Z+c.ShiftZ))
        )
```

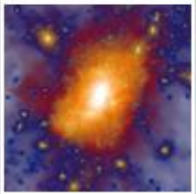
Online Interfaces

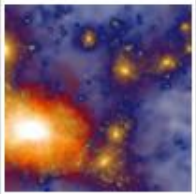
- Largest simulations
  - Search and visualize
  - 10 billion+ objects and growing…
- Indra 512 simulations
  - Coming soon at JHU

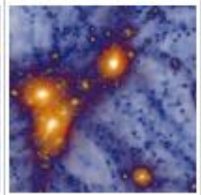


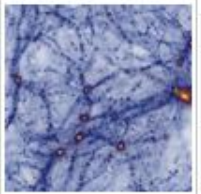
Millennium XXL









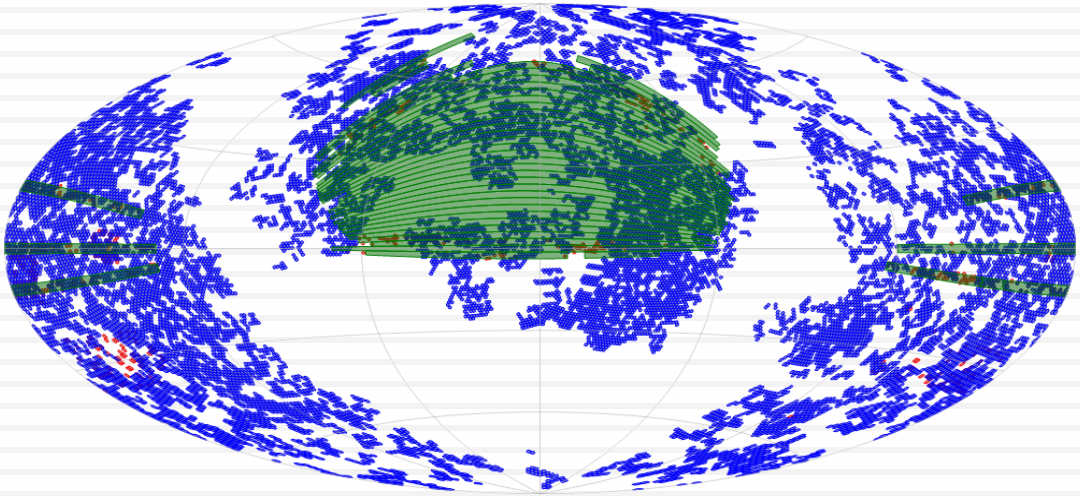


	Pixels	Points
Mouse position	(417, 279)	(1636.2304, 1361.9488)
Viewport dimensions	700 x 400	112.92 x 64.52 Mpc/h

Web Services

- Programming interfaces
  - Execute SQL queries
    - Most flexible
  - Inject probes in simulations
    - Turbulence
    - Cosmology

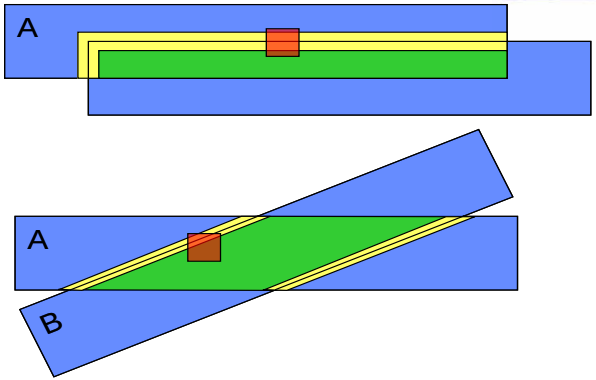
Sky Coverage



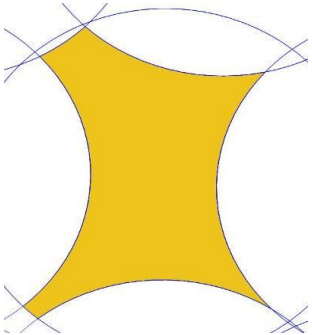
No Sky Coverage?

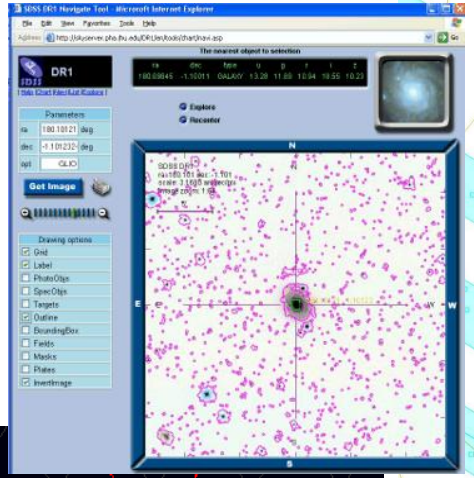
Figure 1. SDSS colour composite image (vri) for our prototype unusual galaxy cluster, at RA $= 16^{\rm h}23^{\rm m}76^{\rm s}$, Dec $= +97°62'12''$, identified by Galaxy Zoo participants. North is at the top, East is to the left.

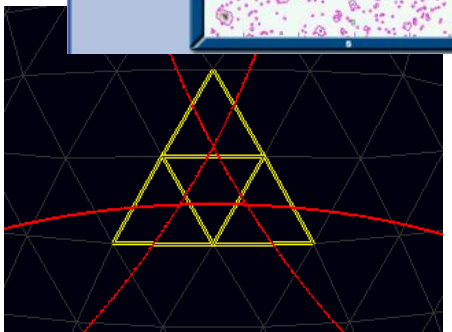
Spherical Geometry

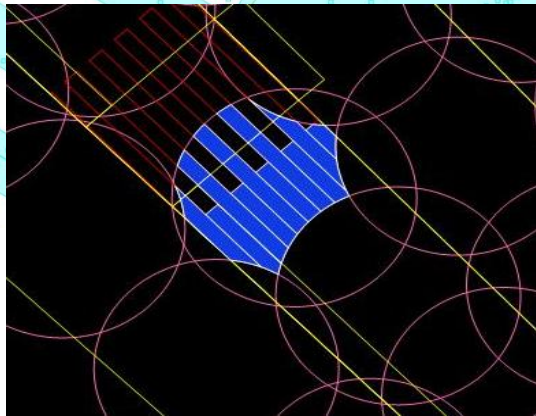


Green area: A ∩ (B- ε) should find B if it contains an A and not masked
Yellow area: A ∩ (B±ε) is an edge case may find B if it contains an A.







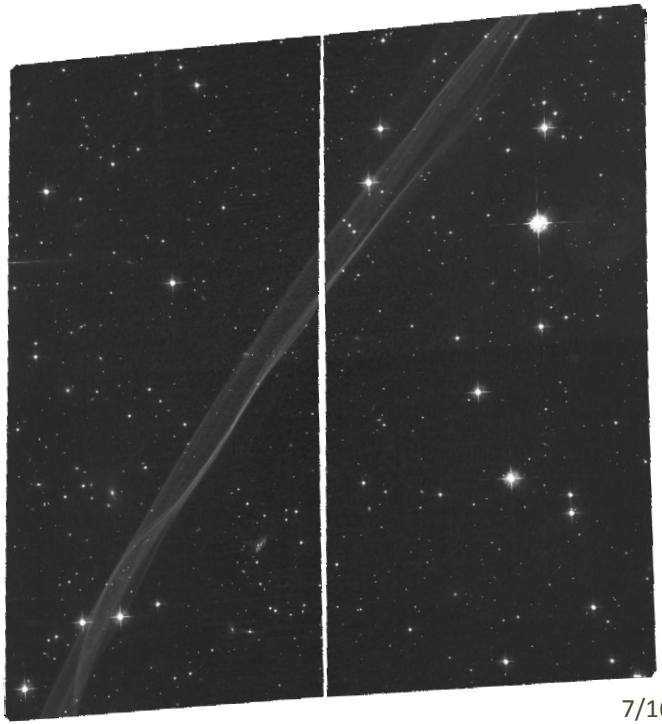


Approaches to Consider

- Pixel maps
  - Sensitivity, etc…
- Equations of shapes
  - Spherical “vector graphics”
- And beyond…

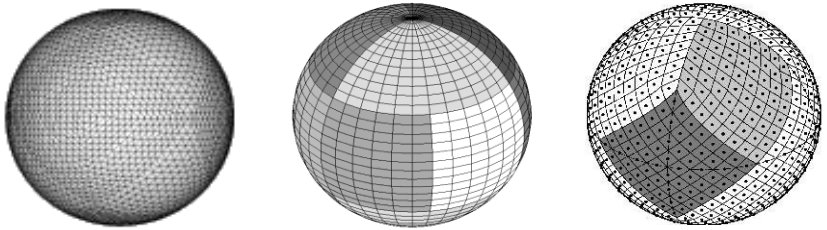
An Observation

- FITS header with WCS
  - Image dimensions map to the geometry
- More exposures?
  - No common pixel coordinate-system
  - Overlapping areas



Common Pixels

- Pre-defined pages of an atlas
  - Standard in cartography
- Image pyramids of hierarchical pixels
  - Including HTM, Igloo, HEALPix, SDSSPix, etc…



- Always approximate!

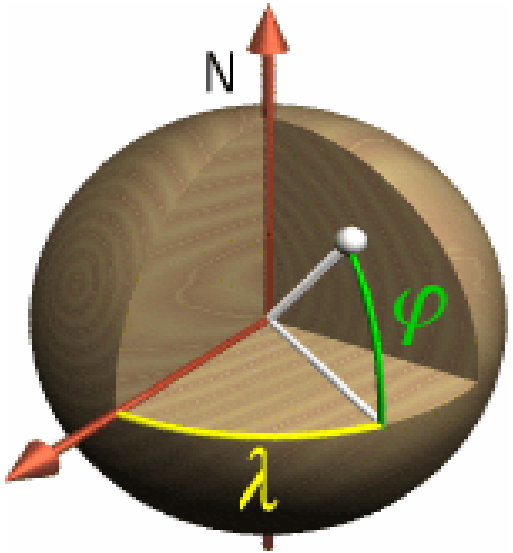
Practical Implementation

- Looking at Terapixels
  - We know how to work with images
  - Now have commodity Internet
  - We have cheap hard-drives

  WorldWideTelescope.org
  Sky in G●●gle Earth
- Integrated catalogs for efficiency
  - How about more surveys?

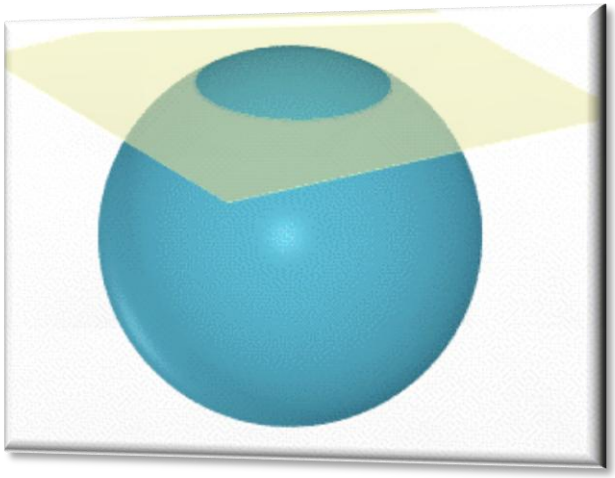
Drawing with Equations

- Working with 3D normal vectors
- Benefits include
  - No wraparound
  - No projections
  - No singularities



Drawing with Equations

- Direct 3D approach
  - Halfspace → Circle/Cap
  - Convex → Simple shapes
- Region
  - Unions of convexes
  - Patches on the sphere

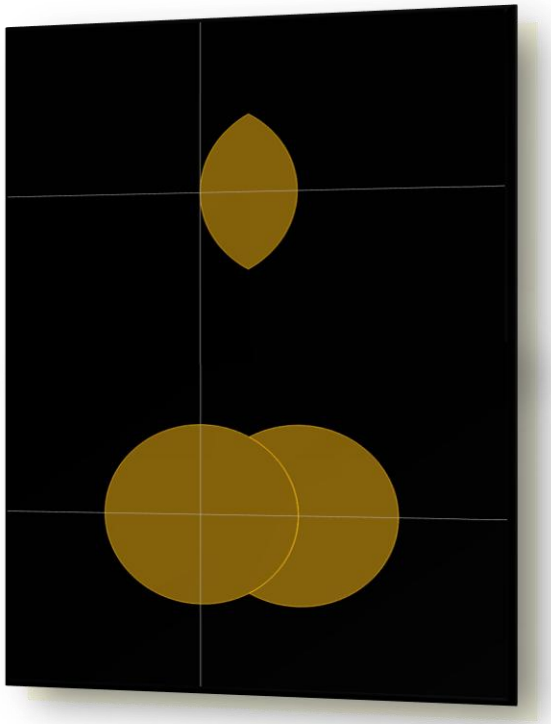


Point in Region Test

- *Halfspace:* one side of a plane $(\vec{n}, c)$
  - Inside, when $\vec{n} \cdot \vec{x} > c$
- *Convex:* a collection of halfspaces
  - Inside, when inside all halfspaces
- *Region:* a collection of convexes
  - Inside, when inside any convex

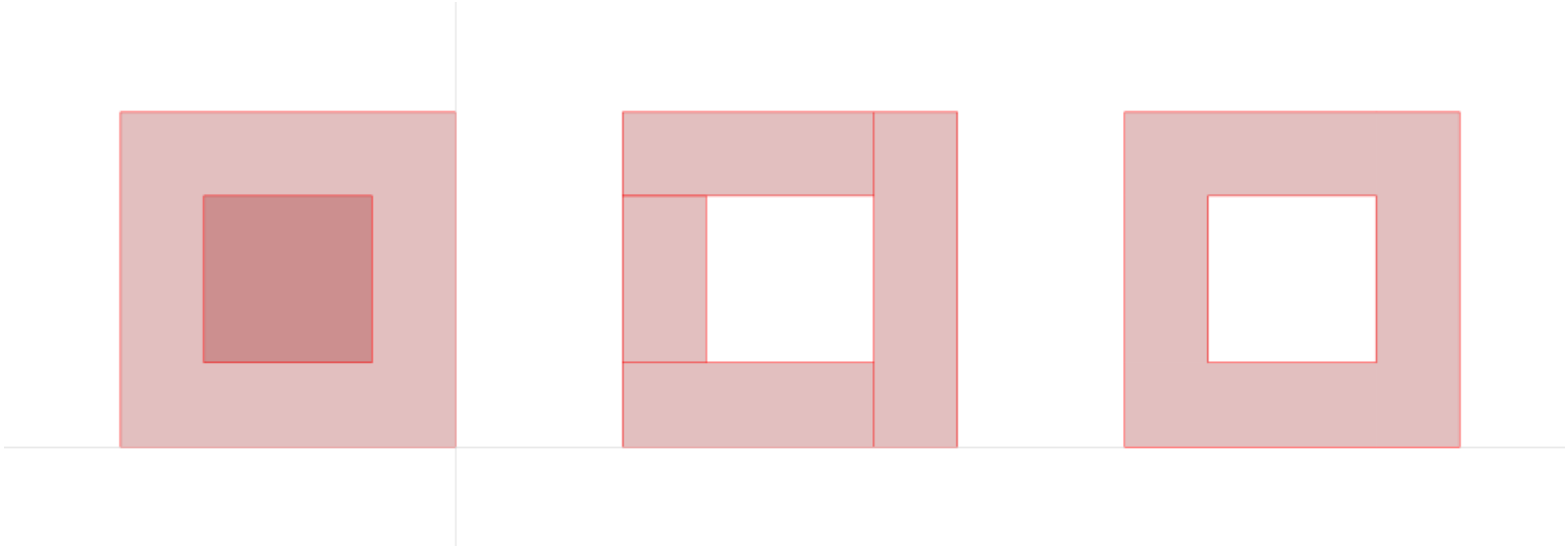
Shape Operations

- Intersection
  - Concat halfspace lists
- Union
  - Concat convex lists
  - Unique coverage
  - Analytic area
- Boolean algebra



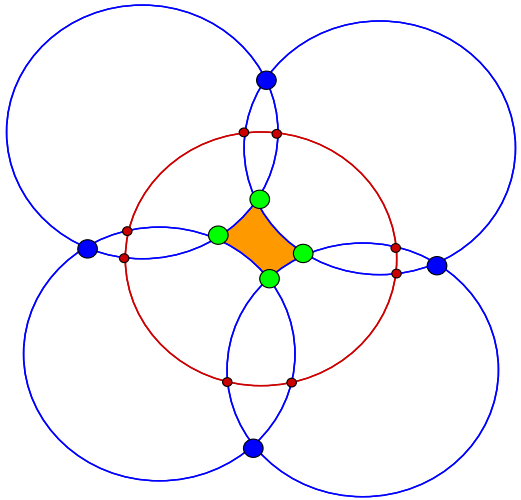
Difference of Convexes is a Region

- The set of Regions is closed for the Boolean ops



Simplification

- Eliminate redundant halfspaces
  - First handle trivial combinations of constraints
  - Then solve geometry on the surface
    - Derive Roots, Arcs, Patches
- Eliminate redundant convexes
  - Some trivial cases, but…
- Make convexes disjoint
  - Unique coverage, area, etc.
- Stitch together convexes
  - When possible



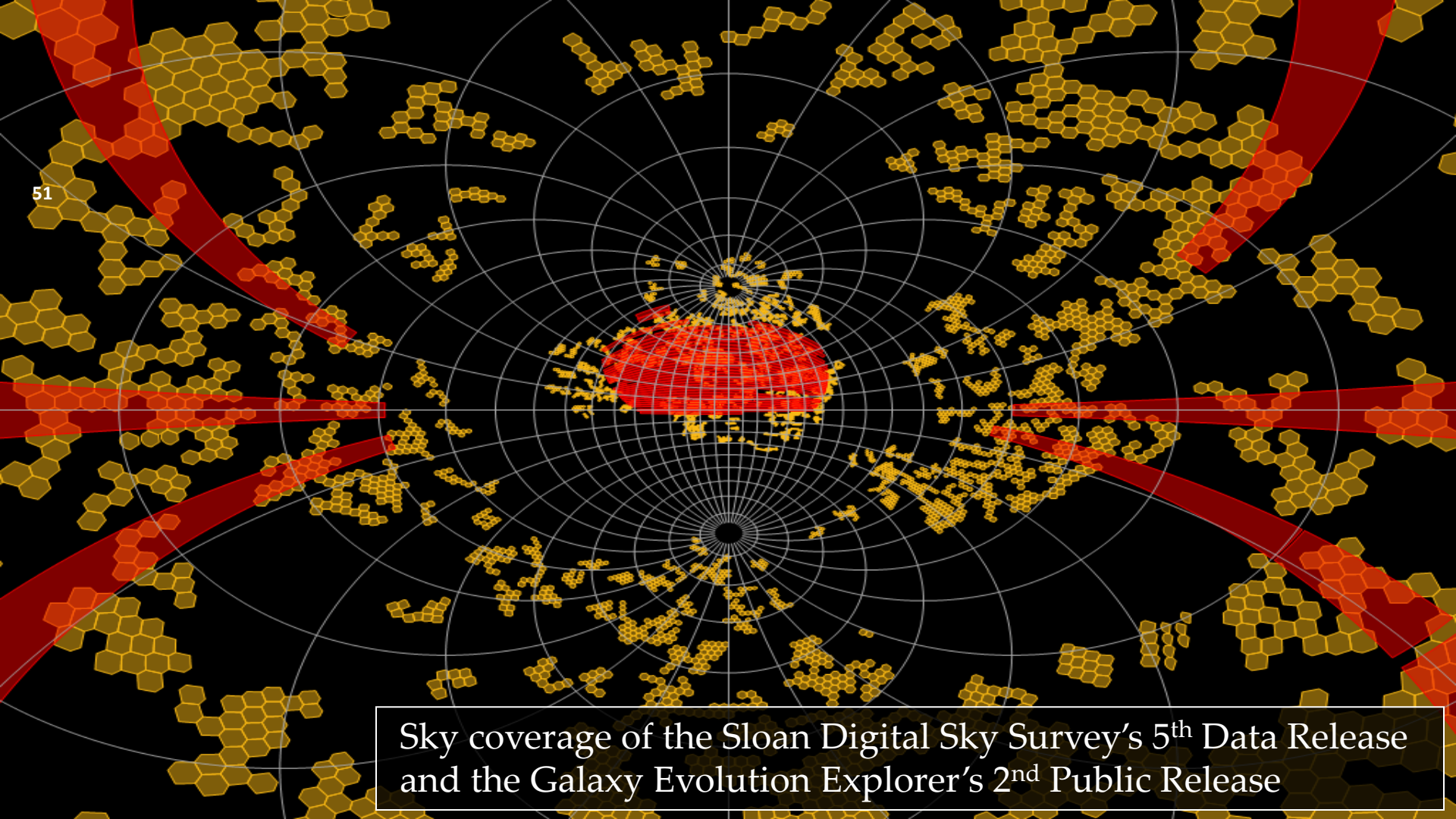
SphericalLib .NET

- C# code ~ 10k lines
  - OS independent (Windows, Un*x w/ Mono)
  - Documentation via Sandcastle
- Great performance!
  - Sloan Digital Sky Survey in 10s

    (13× larger than USA in area)

Numerical Imprecision

- Double precision calculations
  - IEEE 754 standard
- Degeneracy
  - When are two vectors the same?
- Spatial resolution limit
  - Roughly 30 cm on Earth
- Lots of tricks from Graphics Gems



Sky coverage of the Sloan Digital Sky Survey's 5th Data Release and the Galaxy Evolution Explorer's 2nd Public Release

Region in SQL

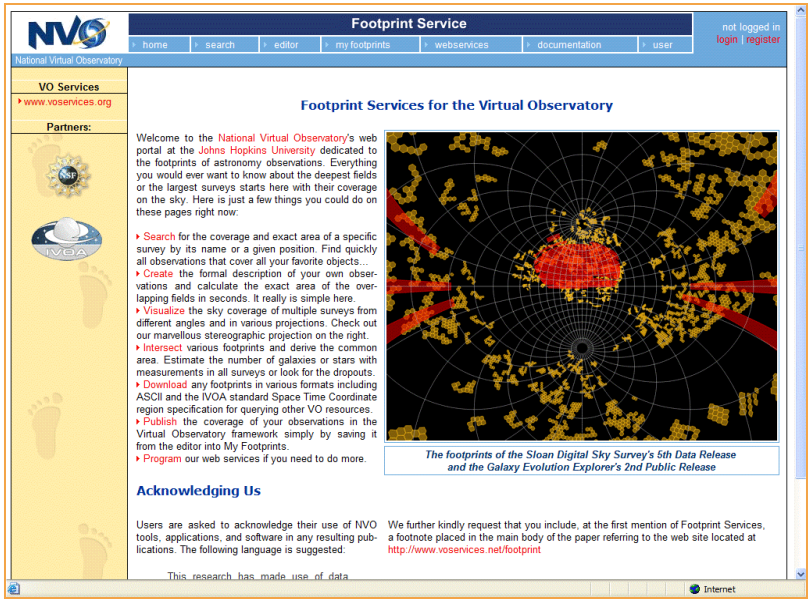

```
DECLARE @s VARCHAR(MAX), @r VARBINARY(MAX),
        @z VARCHAR(MAX), @u VARBINARY(MAX)

SELECT @s = 'REGION CIRCLE J2000 180 0 60',
       @z = 'POLY J2000 180 0 182 0 182 2 180 2',
       @r = sph.fSimplifyString(@s),
       @u = sph.fUnion(@r,sph.fSimplifyString(@z))

SELECT sph.fGetArea(@r), sph.fGetArea(@u)
-- 3.14151290574491 6.35572804450646
/*
 SQL Server Execution Times:
   CPU time = 0 ms,  elapsed time = 1 ms.
*/
```

Footprint Services

- All about coverage
  - Editor and calculator
  - Online public repository
  - On-the-fly visualization
  - STC translator, etc…
- Web services
  - Simple programming



http://voservices.net/footprint

ASTRONOMICAL DATA ANALYSIS SOFTWARE AND SYSTEMS XVI

The Westin La Paloma Resort & Spa
Tucson, AZ, USA
15–18 October 2006

This volume contains papers presented at the 16th annual conference on Astronomical Data Analysis Software and Systems (ADASS XVI). The meeting themes addressed challenges and solutions for very large, compute-intensive systems. Major new astronomical facilities such as ALMA, *Gaia*, *LSST*, and the Square-Kilometer Array, to name but a few, are in various stages of planning and design. Their supporting data systems will exceed the current state of the art in many dimensions by multiple orders of magnitude, and will demand a major portion of the facility construction and operating costs. A key challenge will be to extract their full science potential through distributed data systems that make new, innovative use of Grid, database, web service, and visualization technologies.

The 13 invited and 36 contributed talks, 116 posters, seven floor demonstration booths, three focus demo sessions, and seven Birds-of-a-Feather sessions combined to provide a thorough overview of the latest developments in astronomical software, applications, data facilities, and algorithms. The key topics for this meeting were Challenges & Solutions for Large Data, Advances in Imaging & Calibration Algorithms, Quality Management in Astronomical Data Management Systems, Modern Grid Computing in Astronomy, Architectures for Large Astronomy Software Systems, and Solar Neighborhood & Planetary Astronomy. While a large number of contributed papers followed these themes, the full range of traditional topics included planning for legacy archives such as that for *HST*, the continuing development of standards for astronomical data and the Virtual Observatory, and the latest techniques in algorithm development.

The ADASS conference series has a strong tradition of participant-lead topical sessions within the meeting. This year a record number of "Birds of a Feather" sessions were held on a mix of established and emerging topics, including FITS, Astronomical Data Processing and the VO, The Emerging Infrastructure of Autonomous Astronomy, Building Observatory Legacy Archives, Pipeline Processing of Spectroscopic Observations, IRAF Users and Developers, and Next Generation of Visualization Tools for Astrophysics. Summaries of these sessions are included in these proceedings. This book is suitable for software developers, designers of astronomical software systems, astronomers who use astronomical software, and students of any of these fields.

ASTRONOMICAL SOCIETY OF THE PACIFIC
CONFERENCE SERIES

Available online at
www.aspbooks.org

ISBN 978-1-58381-313-3

VOLUME
376

ASTRONOMICAL DATA ANALYSIS
SOFTWARE AND SYSTEMS XVI

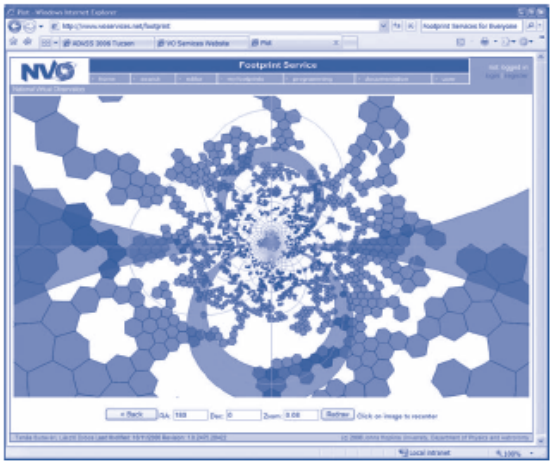
Edited by
Richard A. Shaw, Frank Hill
and David J. Bell

ASPCS

ASTRONOMICAL SOCIETY OF THE PACIFIC
CONFERENCE SERIES

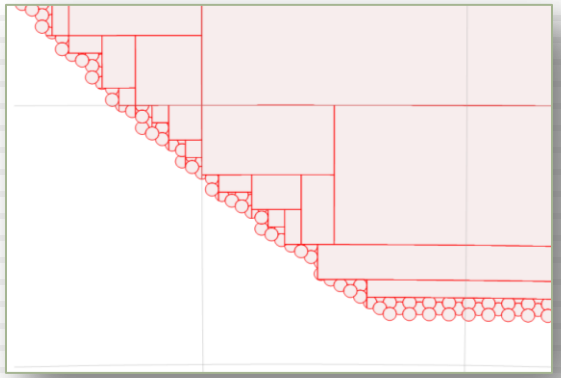
VOLUME 376

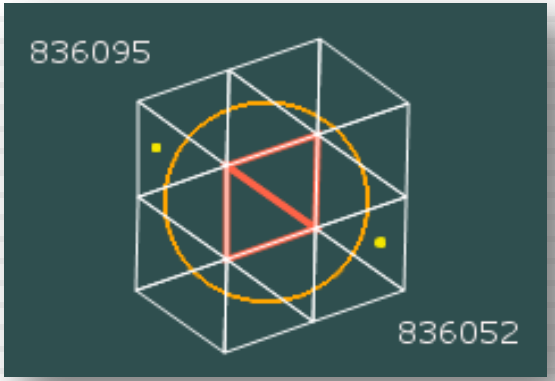
ASTRONOMICAL DATA ANALYSIS
SOFTWARE AND SYSTEMS XVI



Edited by
Richard A. Shaw, Frank Hill and David J. Bell

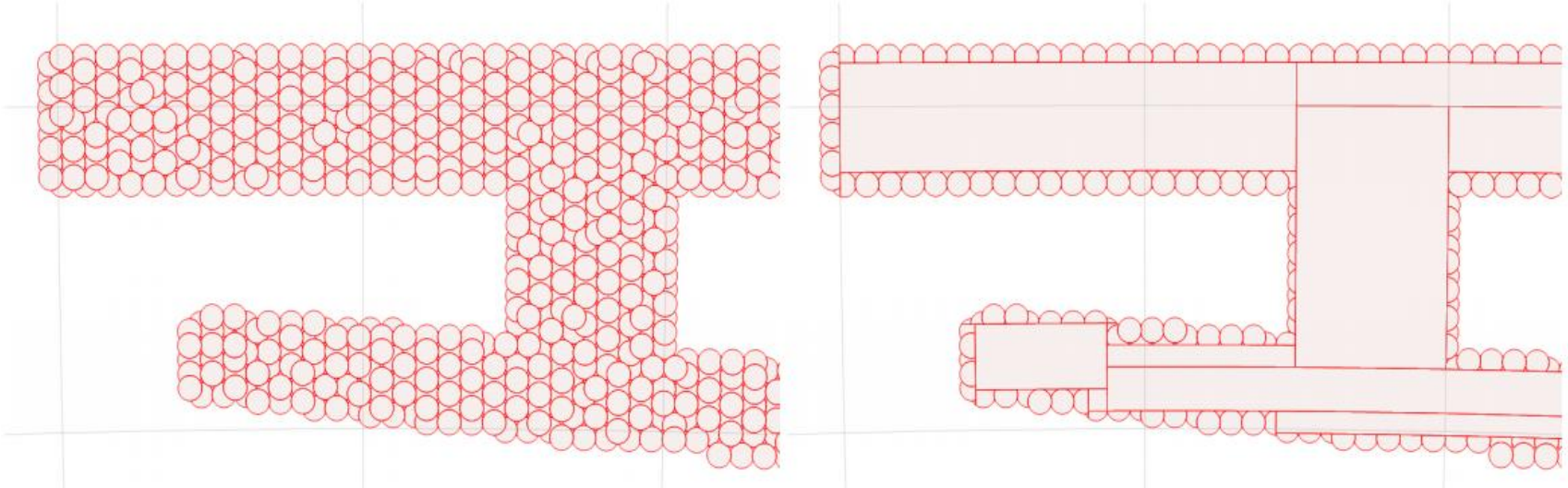
Hybrid Solutions





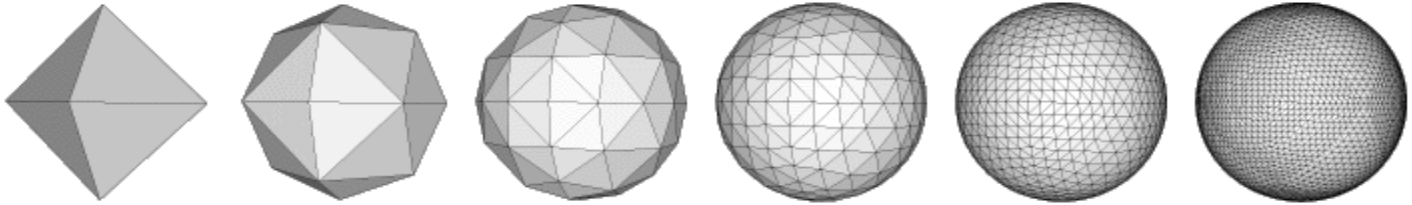
Heuristic Simplification

- Before and After

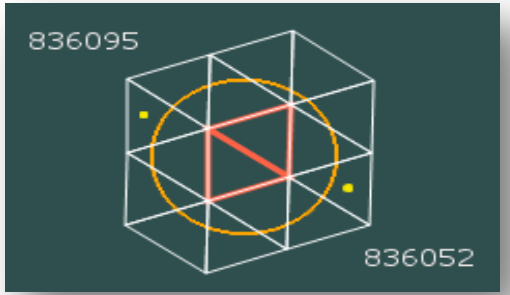


Indexing the Sky

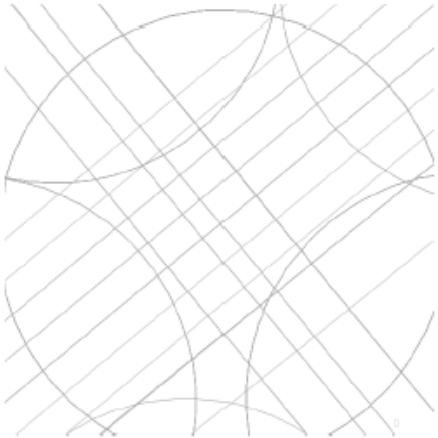
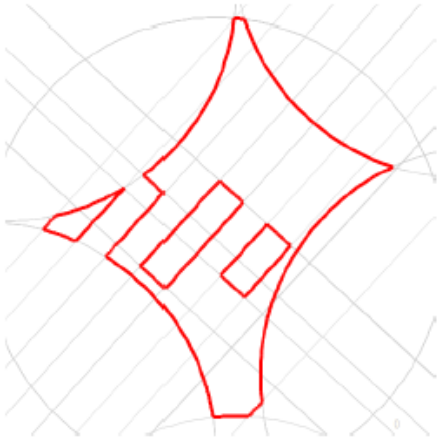
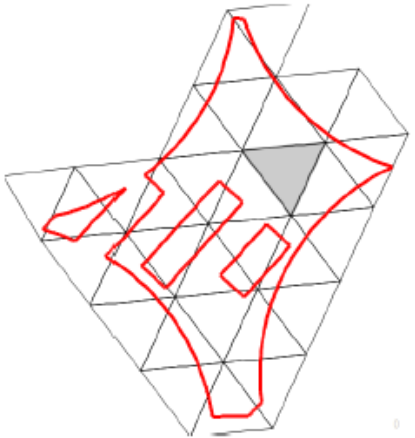
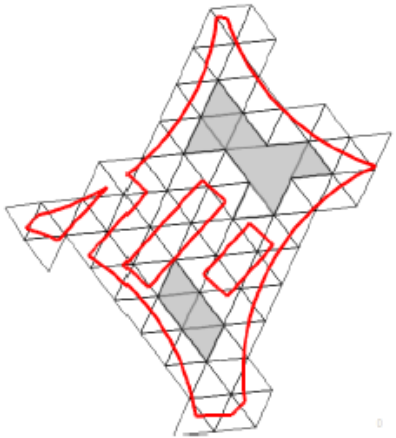
- Hierarchical Triangular Mesh



- Region approximation
  - Fast filtering using HTM ID ranges



Anatomy of an SDSS Region

HTM Filtering


```
WITH Cover AS
(
    SELECT * FROM dbo.fHtmCoverRegion
          ('REGION CIRCLE J2000 180 0 10')
)
SELECT o.ObjID
FROM PhotoObj AS o INNER JOIN Cover AS c
    ON o.HtmID BETWEEN c.HtmIDStart AND c.HtmIDEnd
```

Summary

- Store simulations, e.g., the reference Millennium
  - Simulations take 10x longer than analysis
- Databases enable fast searches
  - Custom routines
  - Space-filling curves
- Direct comparison of observed universe to sims