

LSST, the Spatial Cross-Match Challenge

Maria Nieto-Santisteban
Alexander Szalay
Ani Thakar
The Johns Hopkins University

Jim Gray
Microsoft Research

What is Cross-Matching?

- Identify point(s) in A with point(s) in B
 - Cones: Find points nearby one point
 - Distance from few **arcseconds** to few **degrees**
 - Neighborhood: points nearby points
 - Distance from few **arcseconds** to very few **arcminutes**
-
- Decide whether those points share more than just their position

Zones

- Bin the data
 - $\text{ZoneID} = \text{floor}((\text{dec} + 90.0) / \text{zoneHeight})$
- Place the data close on disk
 - Cluster Index on ZoneID, Ra
- Trick required to handle the (360,0)
- Efficient
 - Cones
 - Neighbors (especially)
- Useful
 - Partition the data
 - Distribute workload

Maria A. Nieto-Santisteban / ADASS 2006

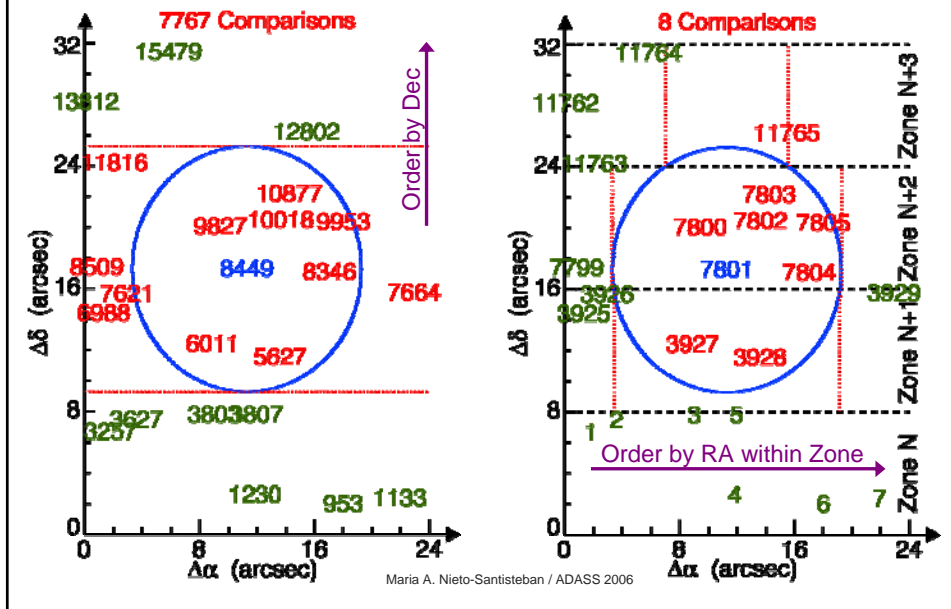
Zone Table

ObjID	ZoneID*	RA	Dec	CX	CY	CZ	...
1	0	0.0	-90.0				
2	20250	180.0	0.0				
3	20250	181.0	0.0				
4	40500	360.0	+90.0				

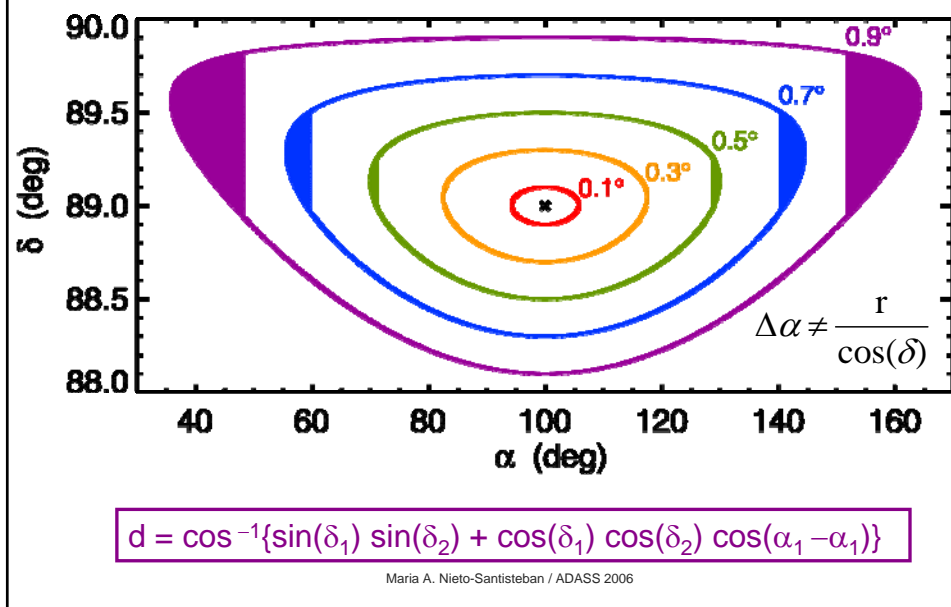
* Using a zone height of 8 arcsec in this example

Maria A. Nieto-Santisteban / ADASS 2006

Declination vs. Zone, RA



“Circular” Regions Near the Poles



SQL CrossNeighbors

```

SELECT *
FROM prObj1 z1
  JOIN zoneZone ZZ
    ON ZZ.zoneID1 = z1.zoneID
  JOIN prObj2 z2
    ON ZZ.ZoneID2 = z2.zoneID
WHERE
  z2.ra BETWEEN z1.ra-ZZ.alpha AND z2.ra+ZZ.alpha
AND
  z2.dec BETWEEN z1.dec-@r AND z1.dec+@r
AND
  (z1.cx*z2.cx+z1.cy*z2.cy+z1.cz*z2.cz) > cos(radians(@r))

```

Maria A. Nieto-Santisteban / ADASS 2006

Number of Rows in LSST Catalogs

	Single Exposure	Single Night	End of Survey
<i>Objects</i>	N/A	N/A	5×10^{10}
<i>Variable Objects</i>	10^5	10^8	3×10^8
<i>Source Detections</i>	3×10^6	3×10^9	8×10^{12}
<i>DIA Source Detections</i>	(10^5)	(10^8)	3×10^{11}

Maria A. Nieto-Santisteban / ADASS 2006

LSST Cross-Match's challenges

- Issue alerts within 60 seconds
 - **Challenge:** Heavily time constrained
- Nightly pipeline @ archive
 - **Challenge:** Database consistency
- Deep Processing
 - **Challenge:** Volume of data to process
Association complexity
- User queries:
 - **Challenge:** Many users, many types of users, many types of queries, a lot of data to look through

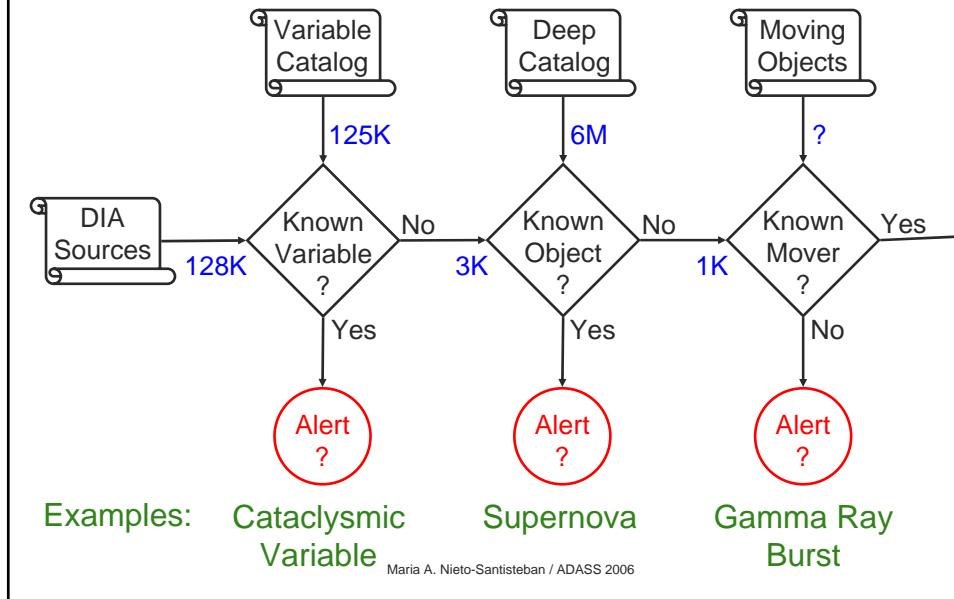
Maria A. Nieto-Santisteban / ADASS 2006

Alert Processing

1. Start alert clock when 2nd exposure ends
 - 3 second readout while slewing to next field
2. Calibrate images (dark subtract, flat field)
 - 201 CCDs = 3.2 Gpixel
3. Difference image analysis
 - Identify and extract variable sources
4. Cross-match with object catalog
 - Distinguish known variables and new objects

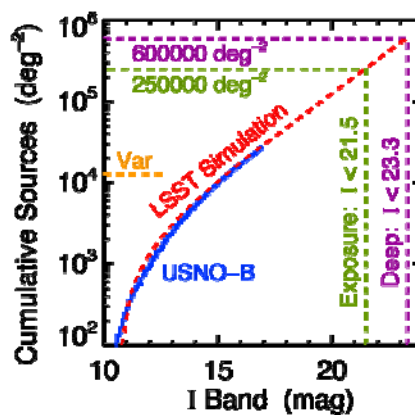
Maria A. Nieto-Santisteban / ADASS 2006

Alerts Data Flow



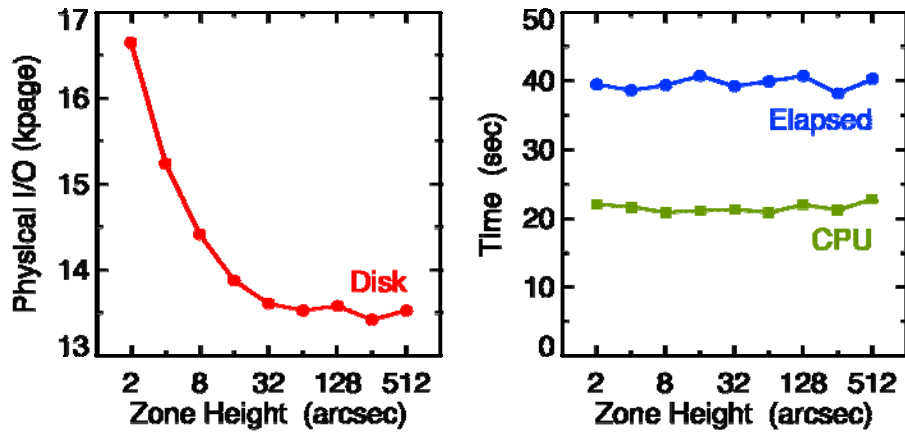
Alert Simulation for Galactic Center

- Extrapolate USNO-B
- LSST FOV of 10 deg²
- 6 Million Stars (DR20)
- 126 K Variable Stars
- 128 K DIA sources
 - 3200 New Variables
 - 1000 Un-matched
 - Moving Objects,
 - New Objects,
 - Transients (GRB)
- Match distance = 1 arcsec



Maria A. Nieto-Santesteban / ADASS 2006

Alert Cross-Match Performance



Maria A. Nieto-Santisteban / ADASS 2006

Summary

- Cone search != Neighbors
- Zones efficiently index and “join” spatial data
 - e.g., SDSS DR5 vs. 2MASS in 80 minutes
- Zones are a convenient for partitioning data
- Simulated a LSST FOV in Galactic Center
- Cross-match catalogs smallest to largest
- Finds possible alerts in 40 sec on desktop

Maria A. Nieto-Santisteban / ADASS 2006