Algorithms for Large Sets of Points

Dave Abel

CSIRO ICT Centre

Who are we?

- CSIRO: Australia’s govt-funded r&d agency. About 6500 staff.
- ICT Centre: the ICT unit. About 170 staff;
- eScience: data grids, workflow. About 6 people.

Why choose this topic?

- Solution times of days were a challenge.
- Joins are clearly a core problem in data integration;
- ‘Must solve’ to tackle data fusion in collections pf databases contributed by members of a community of interest.

Roadmap

- 2 case studies:
  - Catalogue matching;
  - Neighbour finding.
- Work in progress:
  - More point operations;
  - Searching unstructured collections.

Not the same problem …

- Catalogue Matching
  - Determine equivalent pairs of bodies in two catalogues, on the basis of imprecise locations;
  - Not quite a spatial join.
- Neighbour Finding
  - Determine, within a data set, the pairs of objects whose asserted positions are less n arcseconds apart
  - Aka fixed-radius all-neighbors problem.

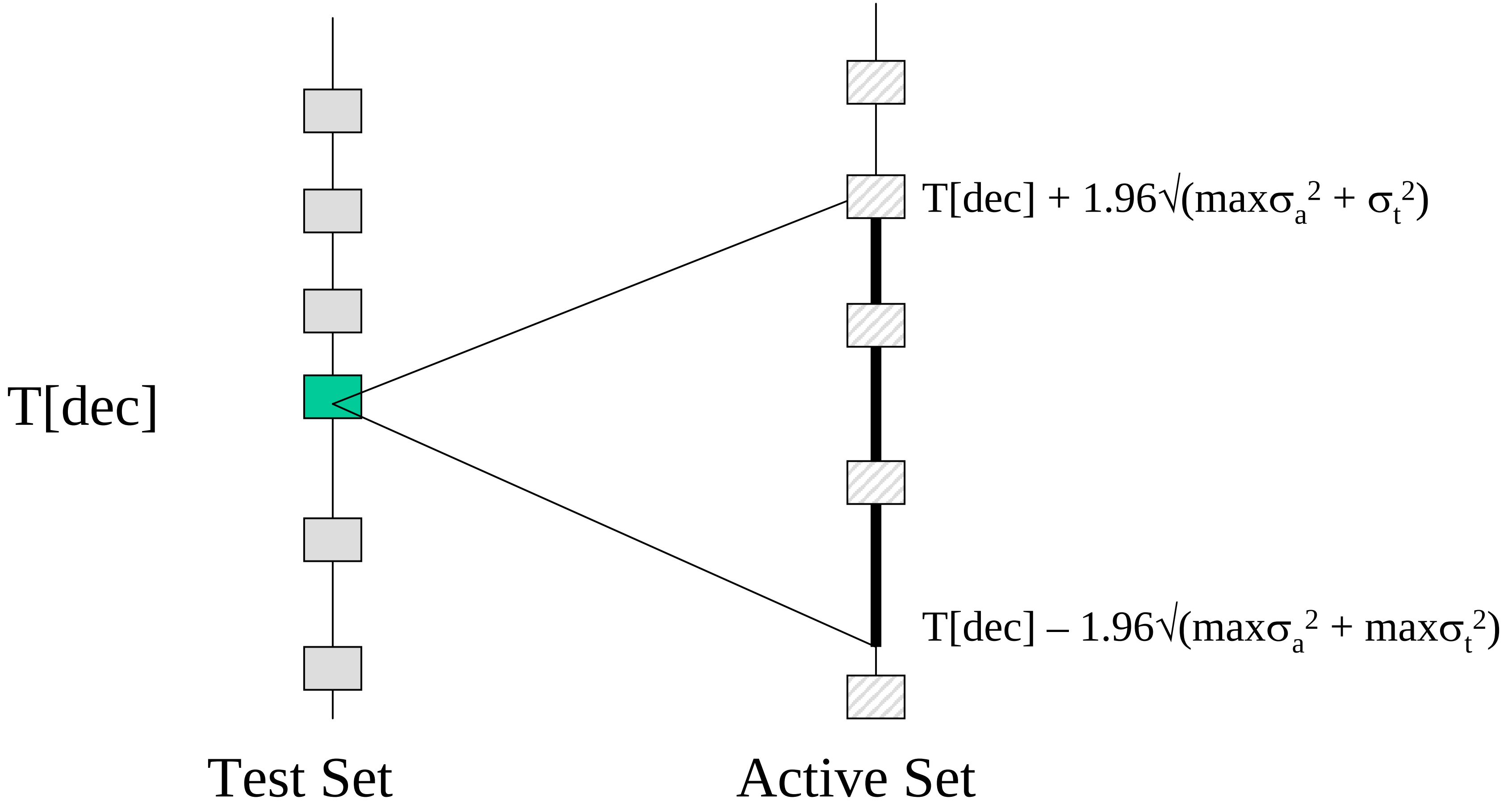
Basis of Algorithm

- Deal with points from ‘Test’set in ascending sequence by declination;
- Maintain an ‘active list’ of points from the other set, that could possibly match the current point from Test or points still to come;
- Event is maintenance of active list, and searching within it for a match.

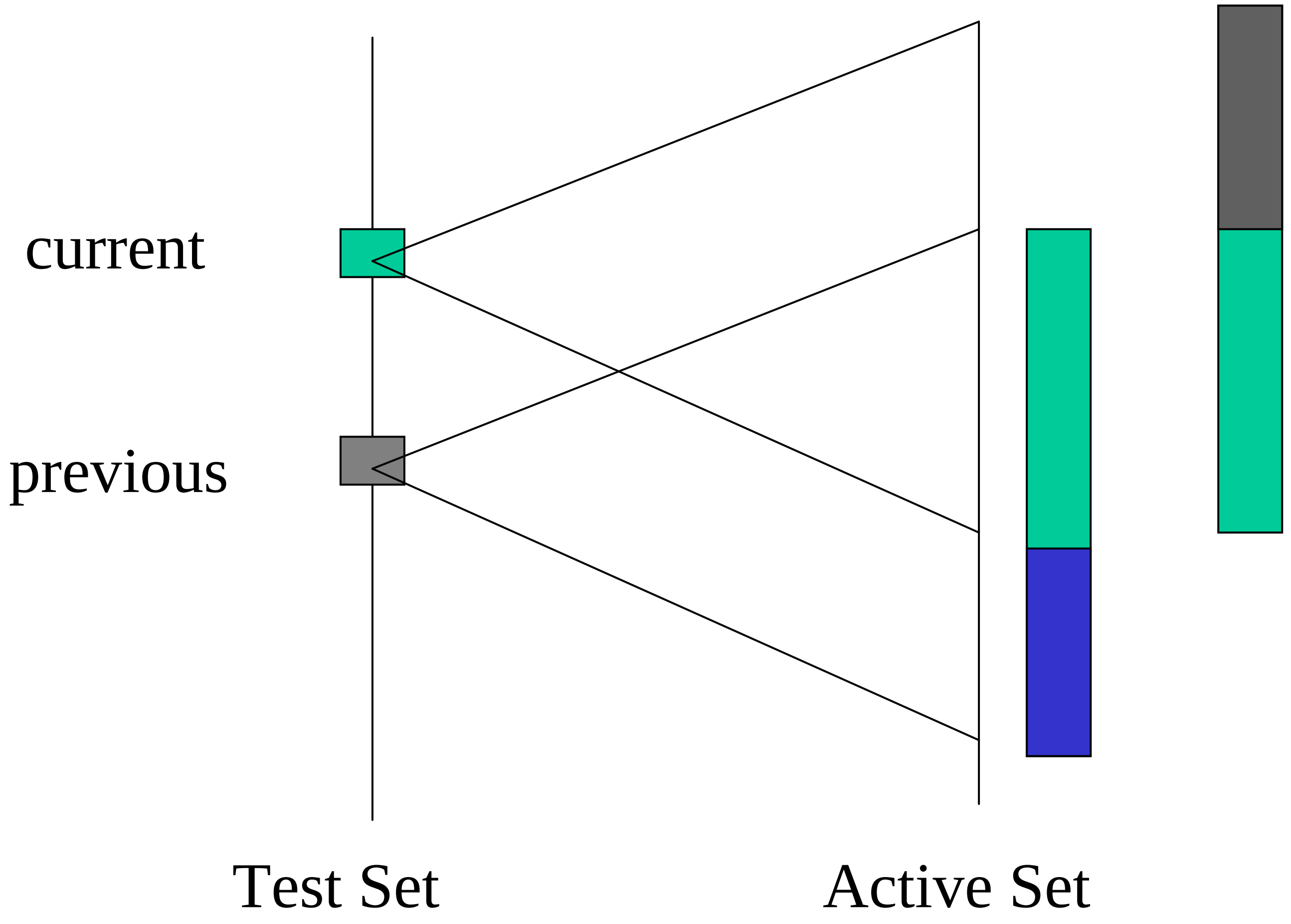
Basis for Algorithms

- Filter and Refine
  - But dec and ra in turn;
- Plane Sweep (on the Sphere)
  - Structure algorithm as processing for regular events;
  - Force regular events by processing in order by declination.

The Active List



Maintaining The Active List



Testing Matches

- The update discipline gives a fair first-pass test on declination;
- Test members against current point in terms of range of right ascension;
- The list (double-ended queue) is typically small:
  - ‘Scan all’ is good, usually;
  - Apply a binary tree as an index on ra for large lists (0.5B x 0.5B matches, or high imprecisions).
- Then test by angular distance.

Some empirical tests

	1XMM	SUMSS	Tycho2	2MASS	USNOA2	USNOB1
1XMM	0:01					
SUMSS	0:01	0:01				
Tycho2	0:04	0:03	0:08			
2MASS	13:42	8:40	14:12	93:36		
USNOA2	15:36	11:36	15:42	92:00	48:36	
USNOB1	31:30	19:12	33:06	149:00	134:30	282:00

1XMM	56K	2MASS	470M
SUMSS	134K	USNOA2	526M
Tycho2	2.5M	USNOB1	1.0B

1XMM	56"	2MASS	1.2"
SUMSS	22"	USNOA2	0.2"
Tycho2	0.2"	USNOB1	1.0"

CSIRO

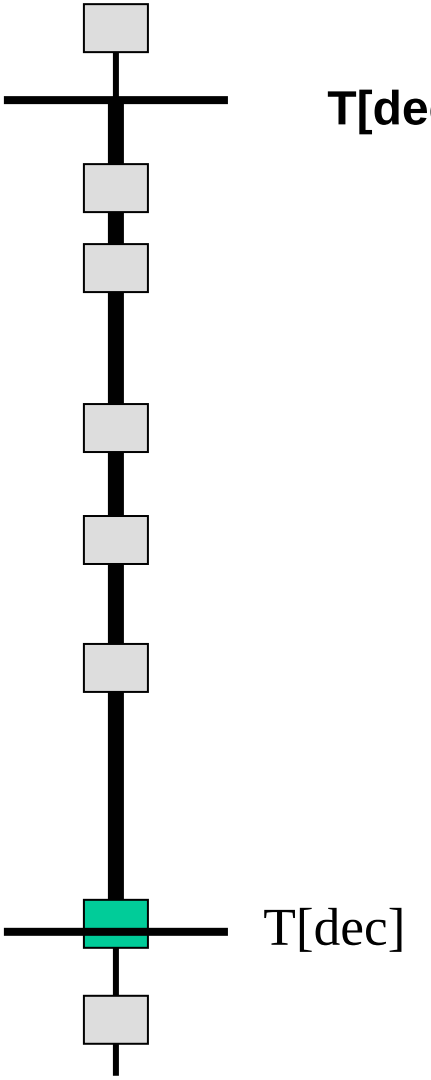
For the sake of completeness …

- Also report the unmatched objects from the two sets;
- Allow object-by-object imprecisions;
- Test by confidence levels on the angular separation by reference to the imprecisions of the two objects.

But all of these are easily parameterised.

Neighbour finding

www.csiro.au



Same approach:

- Current is tail of the active list;
- Test by ra to generate candidates;
- Test candidates in terms of angular separation.

Some Preliminary Results

		Arcsec	
	1	**15**	**30**
1XMM	<0:01	<0:01	<0:01
SUMSS	<0:01	<0.01	<0:01
Tycho2	0:06	0:05	0:06
2MASS	68:00	103:00	175:00
USNO A2	78:00	118:00	190:00

Work in progress

- Implementations
  - Implement as a db stored procedure?
  - How best to implant in distributed db?
- Where will it work?
  - Essentially the approach is to localise operations within a batched problem;
  - Exploits, and is dependant on, some domain-specific aspects;
  - Evaluation of local clusters should fall into this class;

1 is lucky, 2 is intriguing, 3 says there could be something worthwhile.

Other Work in Progress

- Data mining, as searching for bodies with a certain signature;
- Can we trawl an unstructured collection of VO data nodes?
  - Without a global schema?
  - What provisions for inconsistency?
  - Encouraging sound science?
  - Performance?

More Information

Code and reports available real soon now.

Email: dave.abel@csiro.au