

# CROSS-IDENTIFICATION OF ASTRONOMICAL OBJECTS: THEORY AND PRACTICE

11/08/2010

Tamás Budavári / The Johns Hopkins University

# What is the Right Question?

- Cross-identification is a hard problem
  - ▣ Need symmetric  $N$ -way solution
  - ▣ Need reliable quality measure
  
- Same or not?
  - ▣ Distance threshold? Maximum likelihood?



# Outline

- Matching loaded dice
  - Onto catalogs, moving stars, transients, etc...
  
- Practical considerations
  - Locality on sky and storage: Zones
  - Implementation on GPUs



# Tabletop Astronomy

- Loaded dice
  - One die: one object
  - Rolling: observing
  
- All-sky view in 6 pixels
  - Otherwise, just like astro



# Tabletop Astronomy

- Loaded dice
  - ▣ One die: one object
  - ▣ Rolling: observing
  
- All-sky view in 6 pixels
  - ▣ Otherwise, just like astro



# Model Comparison: Same or Not?

- Bayes Factor is the ratio of the
  - Likelihood of “Same”
  - Likelihood of “Not”
  
- Likelihood of a hypothesis?
  - Sum over all possibilities



# Model Comparison: Same or Not?

- Known accuracy, e.g., loaded toward  $l = 1$

$$P_1(\text{⊠}) = \frac{3}{12}, \quad P_1(\text{⊡}) = \frac{1}{12}, \quad P_1(\text{⊣}) = \frac{2}{12}, \dots$$

- 2-way case
  - ▣ Same:  $l_1 = l_2 = l$
  - ▣ Not:  $l_1 \text{ may } \neq l_2$
- $n$ -way same



# Model Comparison: Same or Not?

- Known accuracy, e.g., loaded toward  $l=1$

$$P_1(\text{⊠}) = \frac{3}{12}, \quad P_1(\text{⊡}) = \frac{1}{12}, \quad P_1(\text{⊢}) = \frac{2}{12}, \dots$$

- 2-way case

- Same:  $l_1 = l_2 = l$       $L_{\text{same}} = \sum_l P_l(\text{⊠}) P_l(\text{⊡})$

- Not:  $l_1 \text{ may } \neq l_2$

- $n$ -way same



# Model Comparison: Same or Not?

- Known accuracy, e.g., loaded toward  $l=1$

$$P_1(\ominus) = \frac{3}{12}, \quad P_1(\boxplus) = \frac{1}{12}, \quad P_1(\ominus) = \frac{2}{12}, \dots$$

- 2-way case

- ▣ Same:  $l_1 = l_2 = l$      $L_{\text{same}} = \sum P_l(\ominus)P_l(\boxplus)$

- ▣ Not:  $l_1 \text{ may } \neq l_2$      $L_{\text{not}} = \left[ \sum_{l_1} P_{l_1}(\ominus) \right] \left[ \sum_{l_2} P_{l_2}(\boxplus) \right]$

- $n$ -way same

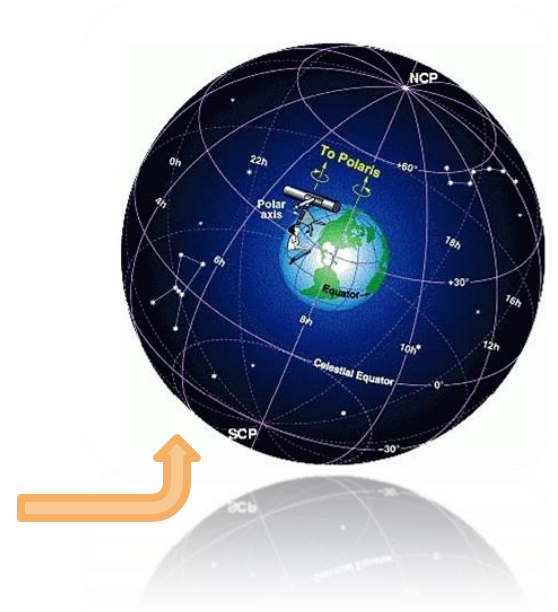


# Celestial Sphere

10

Tamás Budavári

- Continuous functions
- General formalism
  - Accuracy is a density fn on sky



# Normal Distribution

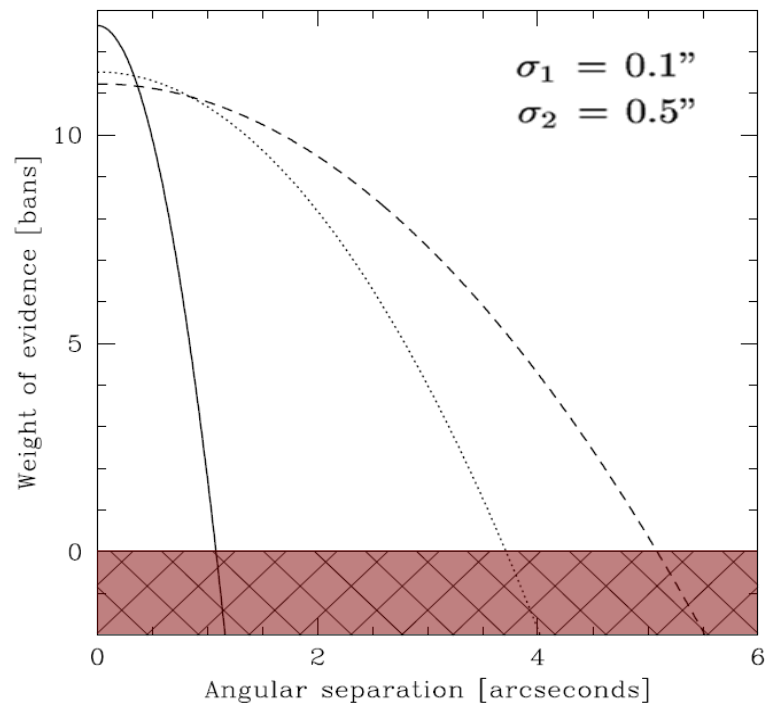
## □ 2-way

$$\frac{L_{\text{same}}}{L_{\text{not}}} = \frac{2}{\sigma_1^2 + \sigma_2^2} \exp \left\{ -\frac{\psi^2}{2(\sigma_1^2 + \sigma_2^2)} \right\}$$

## □ $n$ -way

$$w_i = 1/\sigma_i^2$$

$$\begin{aligned} \frac{L_{\text{same}}}{L_{\text{not}}} &= 2^{n-1} \frac{\prod w_i}{\sum w_i} \exp \left\{ -\frac{\sum_{i<j} w_i w_j \psi_{ij}^2}{2 \sum w_i} \right\} \\ &= \frac{\sinh w}{w} \prod_{i=1}^n \frac{w_i}{\sinh w_i}, \quad w = \left| \sum_{i=1}^n w_i \vec{x}_i \right| \end{aligned}$$



TB & Szalay (2008)

# Normal Distribution

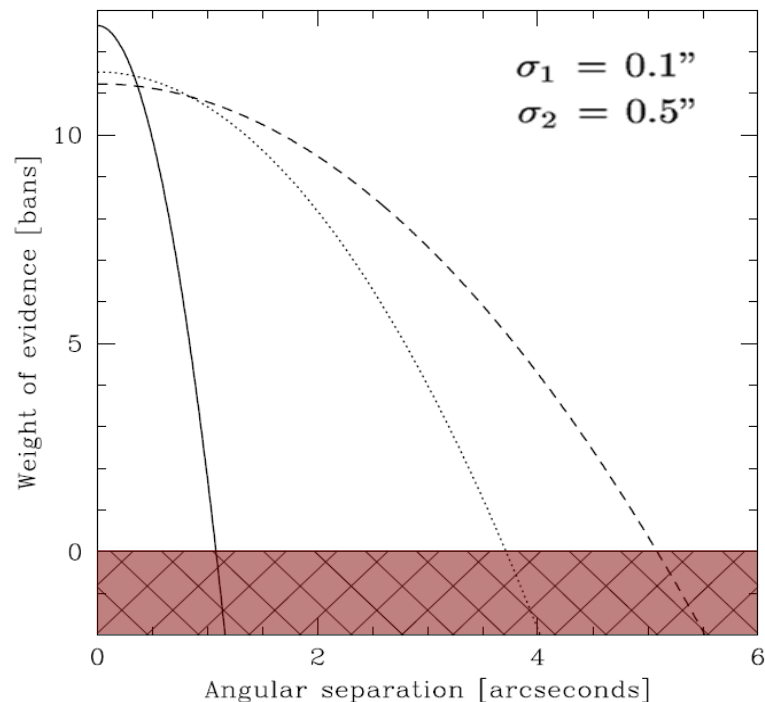
## □ 2-way

$$B = \frac{2}{\sigma_1^2 + \sigma_2^2} \exp \left\{ -\frac{\psi^2}{2(\sigma_1^2 + \sigma_2^2)} \right\}$$

## □ $n$ -way

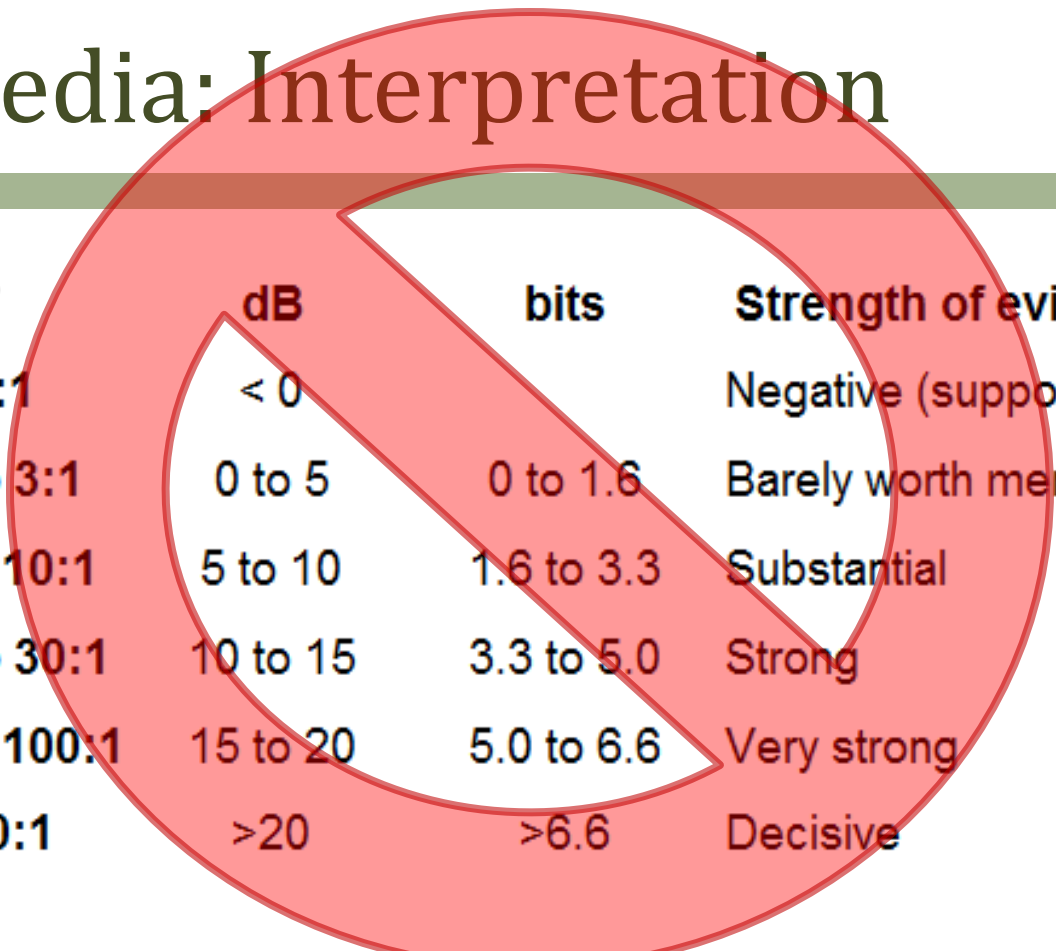
$$w_i = 1/\sigma_i^2$$

$$B = 2^{n-1} \frac{\prod w_i}{\sum w_i} \exp \left\{ -\frac{\sum_{i<j} w_i w_j \psi_{ij}^2}{2 \sum w_i} \right\}$$
$$= \frac{\sinh w}{w} \prod_{i=1}^n \frac{w_i}{\sinh w_i}, \quad w = \left| \sum_{i=1}^n w_i \vec{x}_i \right|$$



TB & Szalay (2008)

# Wikipedia: Interpretation



$B$	dB	bits	Strength of evidence
< 1:1	< 0		Negative (supports $M_2$ )
1:1 to 3:1	0 to 5	0 to 1.6	Barely worth mentioning
3:1 to 10:1	5 to 10	1.6 to 3.3	Substantial
10:1 to 30:1	10 to 15	3.3 to 5.0	Strong
30:1 to 100:1	15 to 20	5.0 to 6.6	Very strong
>100:1	>20	>6.6	Decisive

14

# Probability of a Match

Same or not?



# From Priors to Posteriors

- Posterior probability from prior & Bayes factor

$$P(H|D) = \left[ 1 + \frac{1 - P(H)}{B P(H)} \right]^{-1}$$

- Prior probability of a match
  - Like dice in a bag:  $1/N$  and  $N^{1-n}$
  - In general:  $N_{\star} / N_1 N_2 \dots N_n$



# Self-Consistent Estimates

- Prior has an unknown fudge-factor
  - ▣ Educated guess
  - ▣ Or solve for it:

$$\left. \begin{aligned} \sum P(H) &= N_{\star} \\ \sum P(H|D) &= N_{\star} \end{aligned} \right\}$$



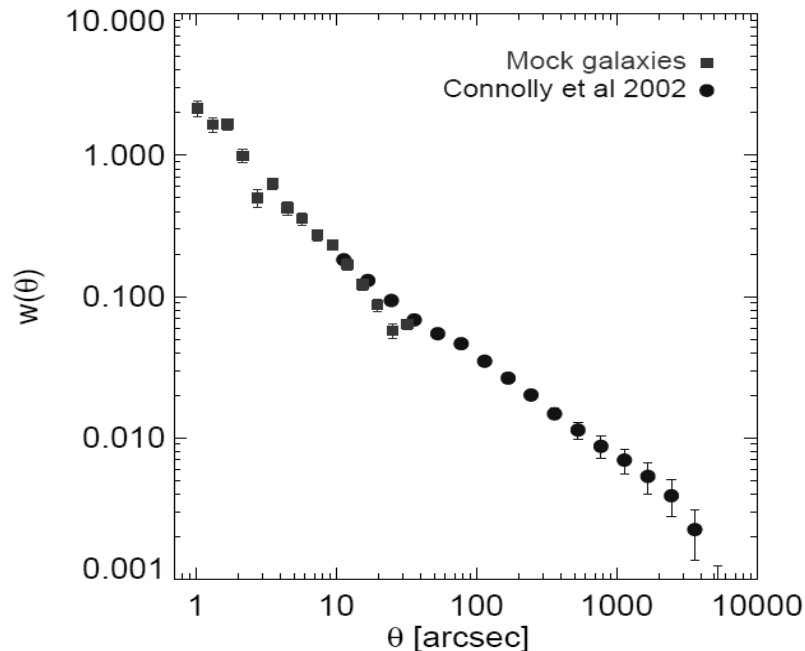
*TB & Szalay (2008)*

# Simulations

- Mock objects
  - With correct clustering
  - $U_{01}$  values as properties



- Simulated sources
  - Subsets:  $N_1$   $N_2$
  - Overlap:  $N_{\star}$

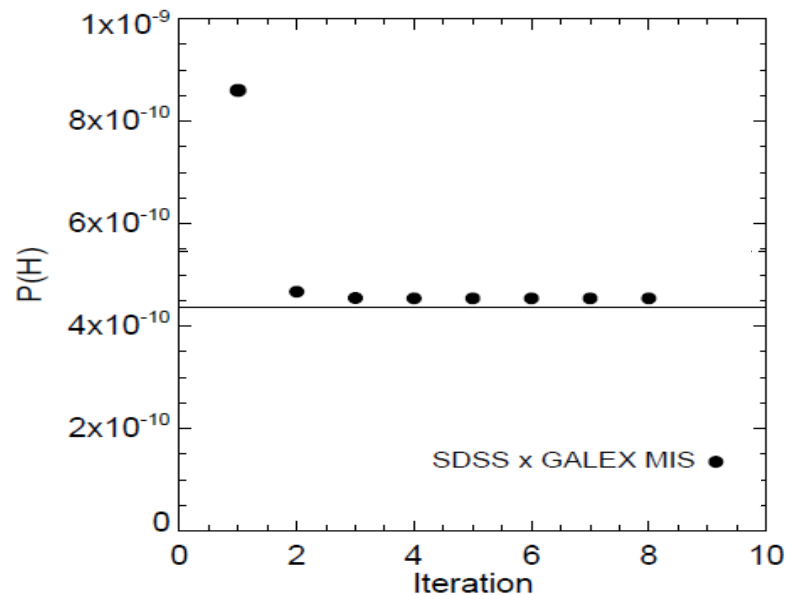


# Simulations

- Mock objects
  - With correct clustering
  - $U_{01}$  values as properties

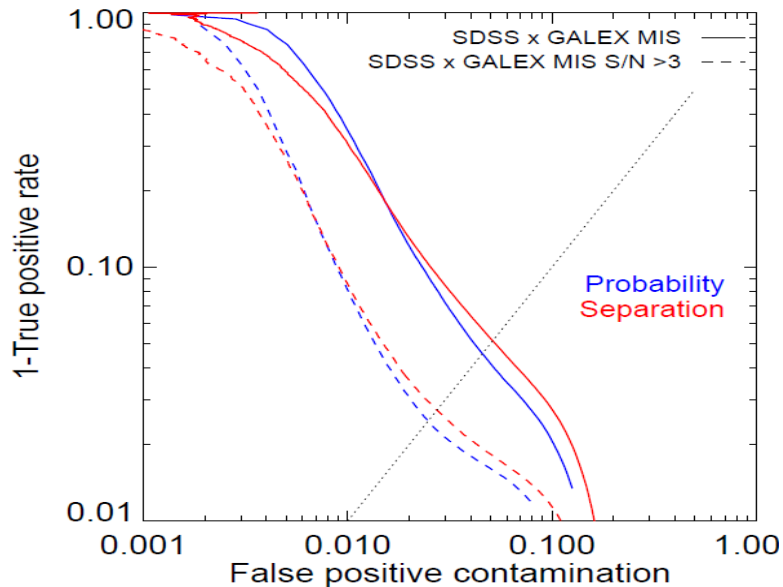


- Simulated sources
  - Subsets:  $N_1$   $N_2$
  - Overlap:  $N_{\star}$



# Simulations

## Quality



## Multiple matches

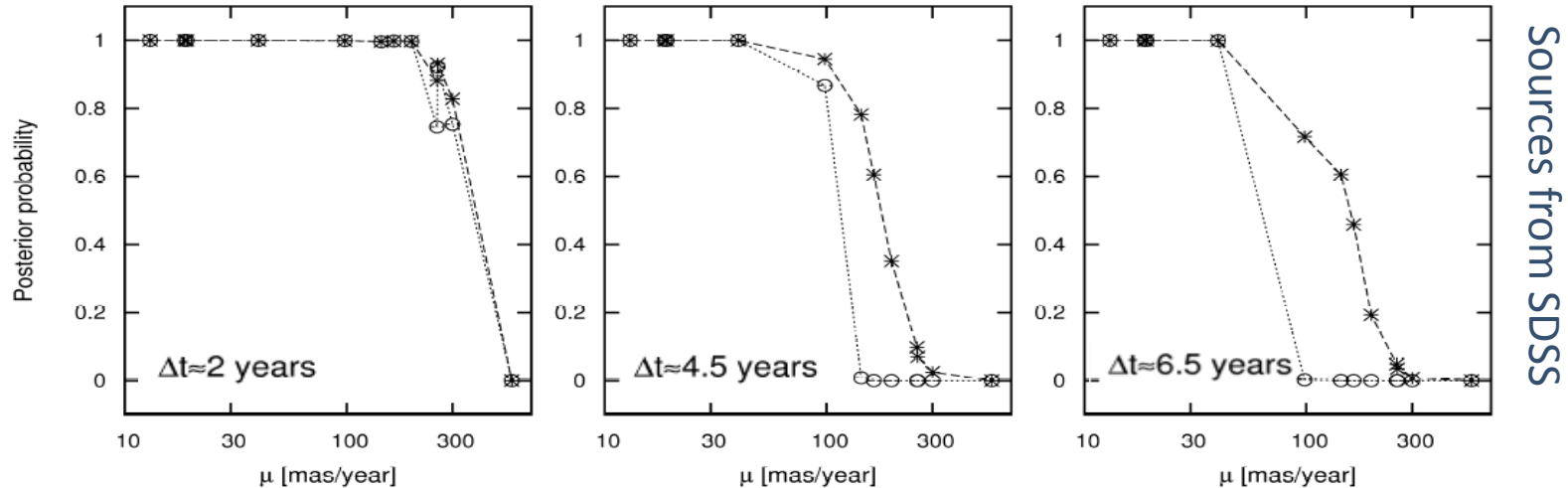
GALEX	SDSS		
	1	2	Many
1	74.061 (75.870)	21.007 (18.595)	2.577 (2.469)
2	1.146 (2.253)	1.006 (0.697)	0.188 (0.102)
Many	0.006 (0.009)	0.007 (0.004)	0.002 (0.001)

Explained by simple model  
of point sources!

*Heinis, TB, Szalay (2009)*

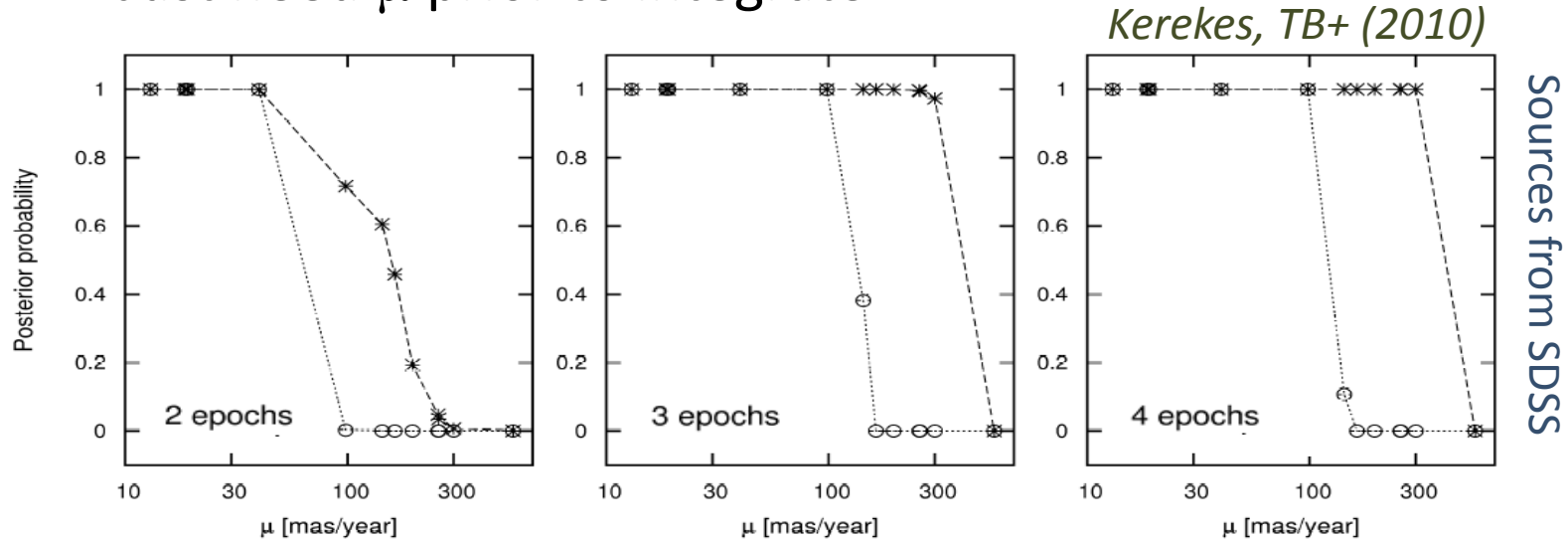
# Proper Motion

- Same hypotheses but different parameters
  - ▣ Just need  $\mu$  prior to integrate



# Proper Motion

- Same hypotheses but different parameters
  - ▣ Just need  $\mu$  prior to integrate



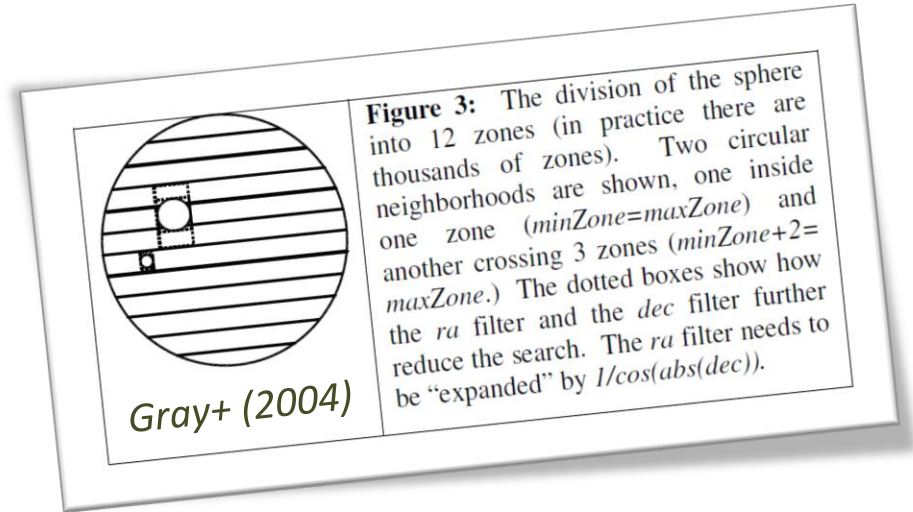
22

# Matching in Practice

Locality on sky and storage

# Zone Algorithm

- Constant Declination zones
  - Sort by R.A. within
- Fast SQL code
  - SDSS-GALEX in 1 hour
  - CPU limited!



# Parallel on GPUs

24

- C for CUDA prototype
  - No smart I/O, RAM limit
- NVIDIA GTX 480 1.5GB
  - 5" search with 5" zones
  - 29M×29M in **11 seconds!**

```
C:\>CuXmatch.exe dr7.bin 29000000 dr7.bin 29000000 5 5 4
[dbg] n_zones: 129600

[dat] 1
[tmr] Load: 12.776000
[tmr] Copy: 0.452000
[tmr] Sort: 2.605000
[tmr] Lmts: 0.000000
[tmr] Back: 0.499000
[tmr] Splt: 0.921000

[dat] 2
[tmr] Load: 10.296000
[tmr] Copy: 0.453000
[tmr] Sort: 2.823000
[tmr] Lmts: 0.000000
[tmr] Back: 0.499000
[tmr] Splt: 0.905000

[tmr] Cop2: 0.671000
[tmr] Mtch: 10.998000
[tmr] Ftch: 0.265000
[tmr] Main: 47.876000

[res]
587727177914515631 587727177914515631
587727177914515580 587727177914515580
587727177914515797 587727177914515797
587727177914581686 587727177914581686
...
|
```



# Summary

- Direct Bayesian approach to “Same or not?”
  - ▣ Can include physics, e.g., SED, proper motion
  - ▣ Reliable probabilities from ensemble statistics
- Efficient SQL implementations and on GPUs
  - ▣ Parallel SkyQuery engine in the works
  - ▣ Hubble Legacy Archive (Lubow+ P105)

