

PRACTICAL ANALYTICS

7/19/2012

Tamás Budavári / The Johns Hopkins University

Statistics

Tamás Budavári

- Of numbers
- Of vectors
- Of functions
- Of trees

Statistics

- Description, modeling, inference, machine learning
- Bayesian / Frequentist / Pragmatist ?

	Supervised	Unsupervised
Discrete	Classification	Clustering
Continuous	Regression	Dimensional Reduction

What's Large?

- VOLUME
 - Say >100TB today but tomorrow? Moving target...
- COMPLEXITY
 - The raw dataset are simple unlike their derivatives
- DEFINITION?
 - Large when you cannot apply the “usual” tools



Data

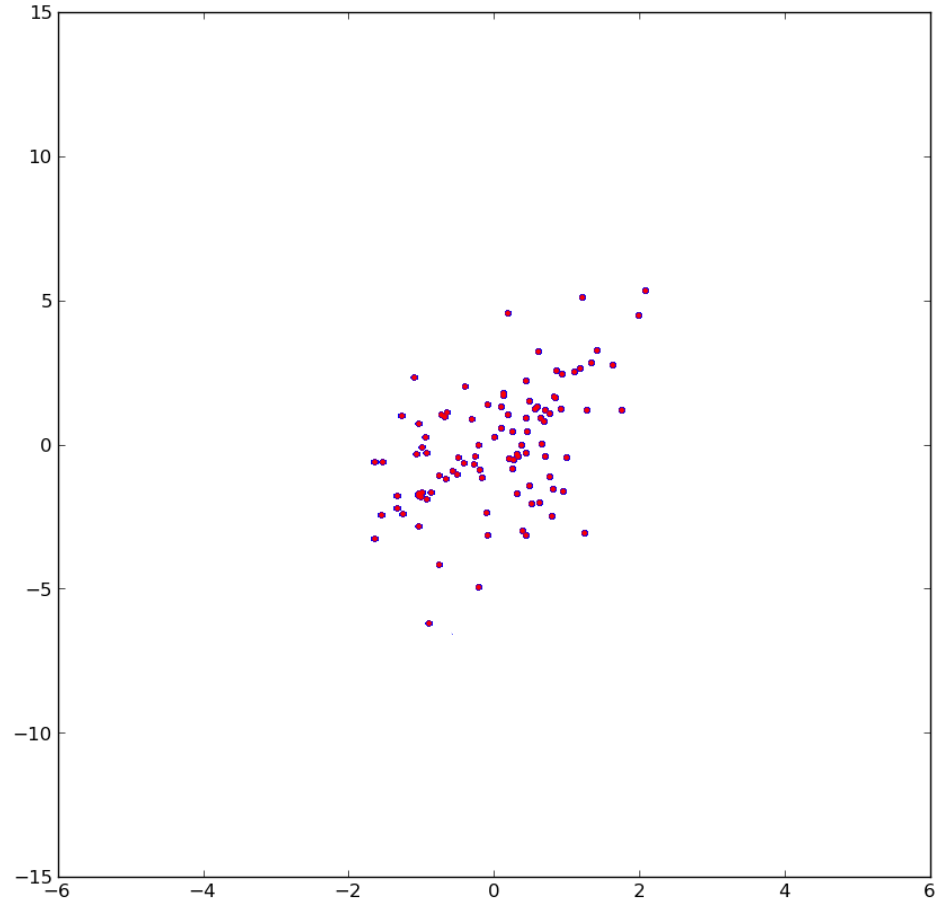
LARGE !!

Data

LARGE !!

Large?

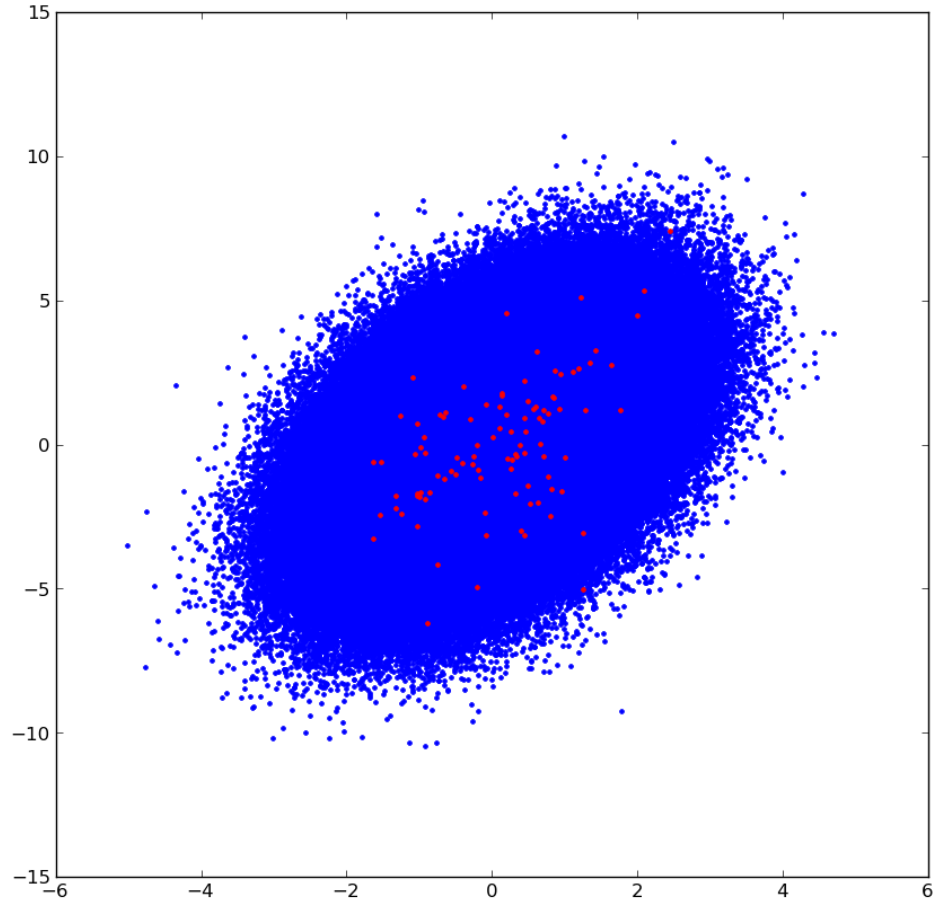
□ Sample size



Large?

□ Sample size

MORE OUTLIERS!

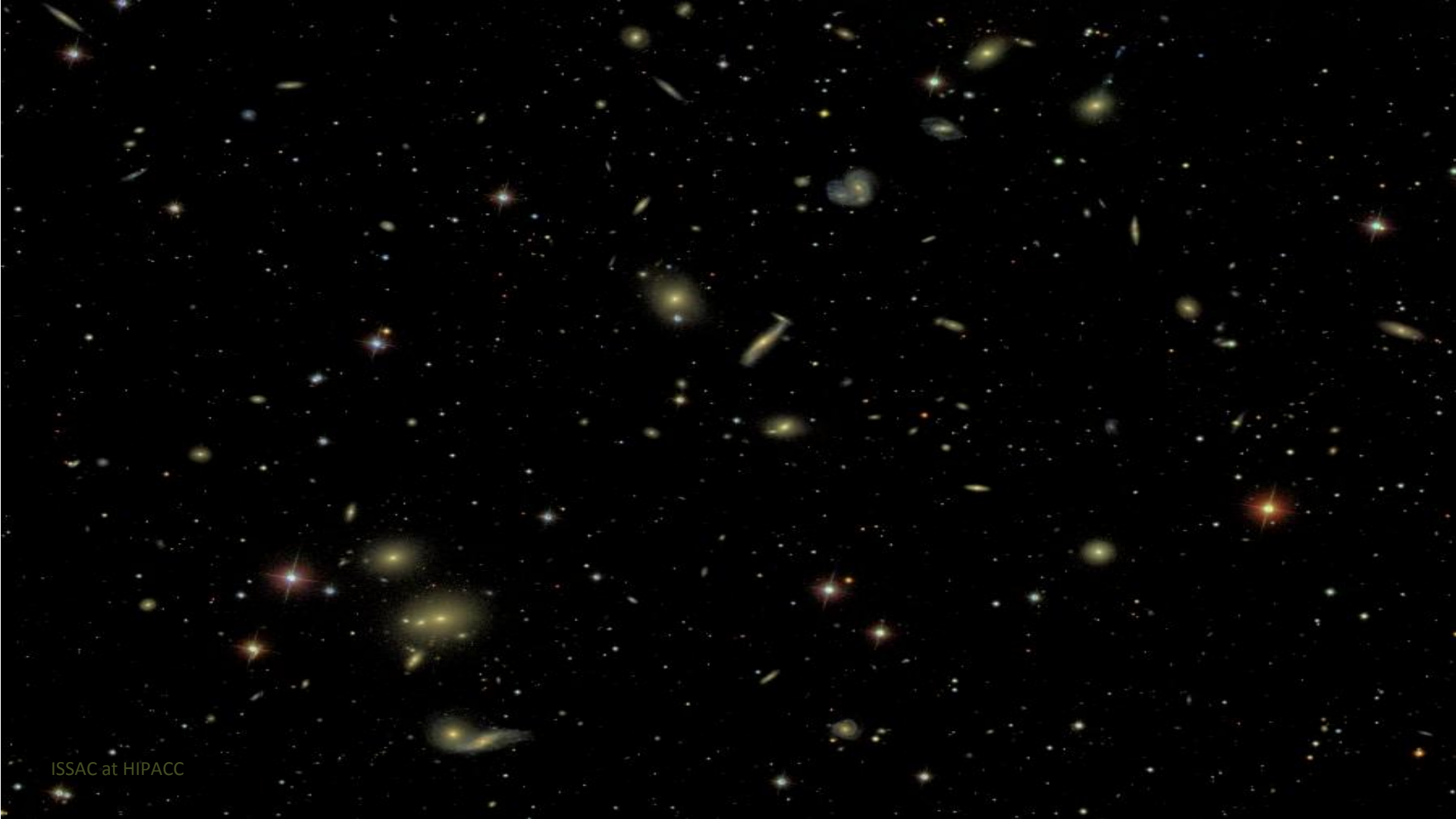


Large?

- Dimensions
 - ▣ Ratio of surface/volume grows

all points are lonely in high dimensions

THE CURSE OF DIMENSIONALITY!





Keeping Up?

Tamás Budavári

- Image processing
- Catalog extraction
 - ▣ $O(n)$
- What is difficult?
 - ▣ $O(n \log n)$
 - ▣ $O(n^2)$, ...

Worse w/ Moore's law

Fundamental Challenges

Tamás Budavári

- Cross-identification of sources
 - To assemble multicolor catalogs
- Drop-outs from sky coverage
 - To constrain fluxes not detected
- Constraining physical properties
 - To interpret the data





Cross-Identification

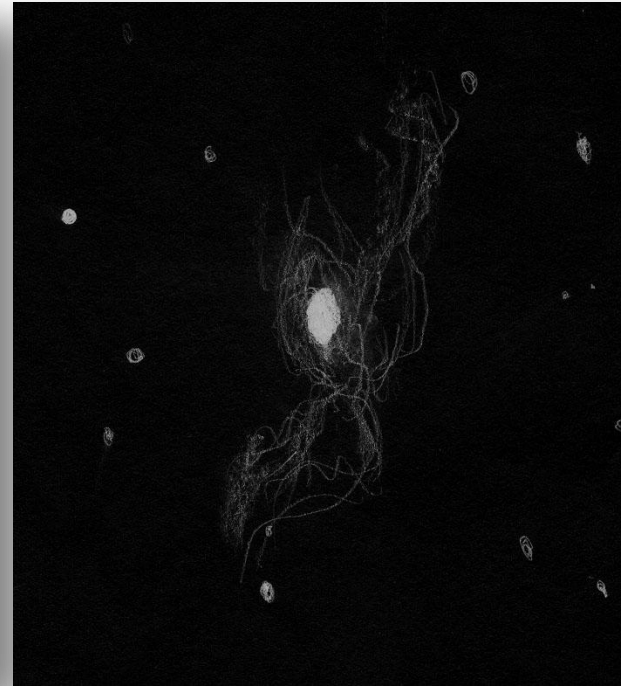
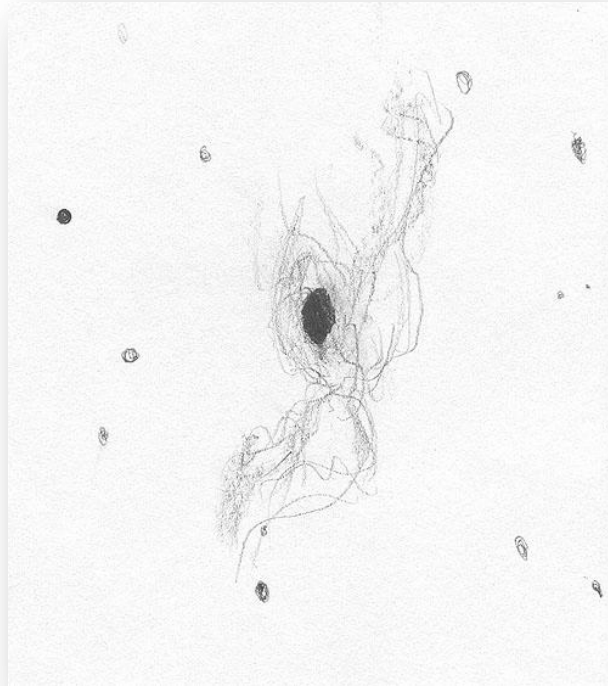
From long-tail science to the largest experiments

Recording Observations

Tamás Budavári

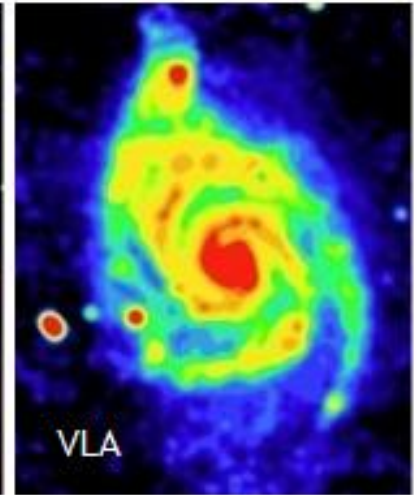
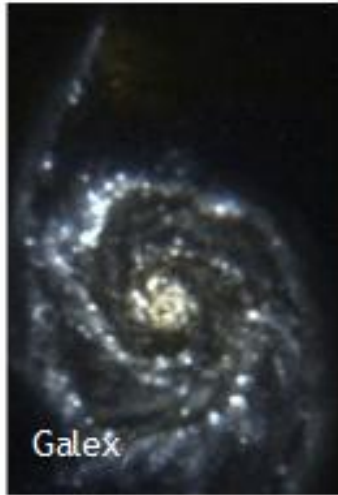
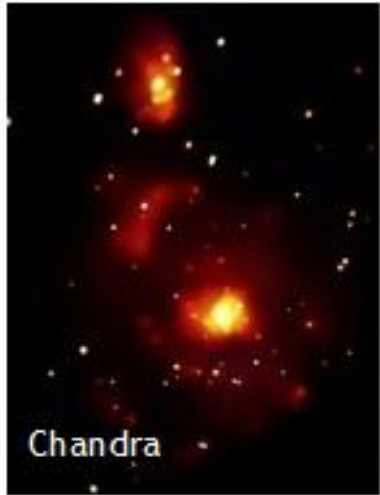
- Astronomers drew it...
- Now kids do it on SkyServer

#1 by Haley ⇒

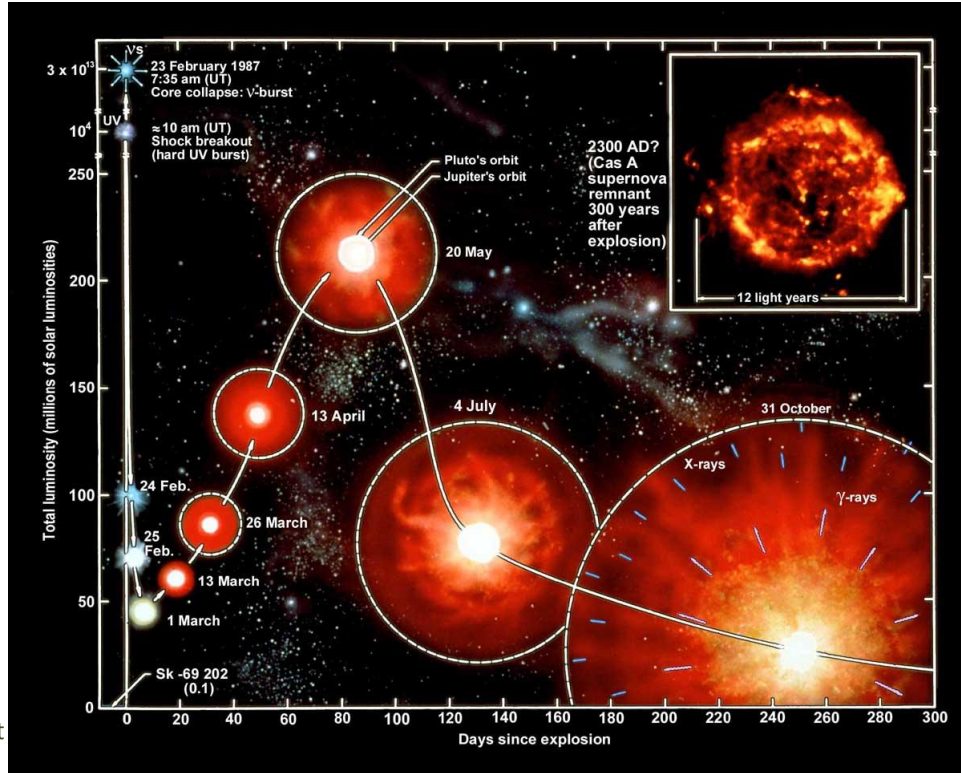


Multicolor Universe

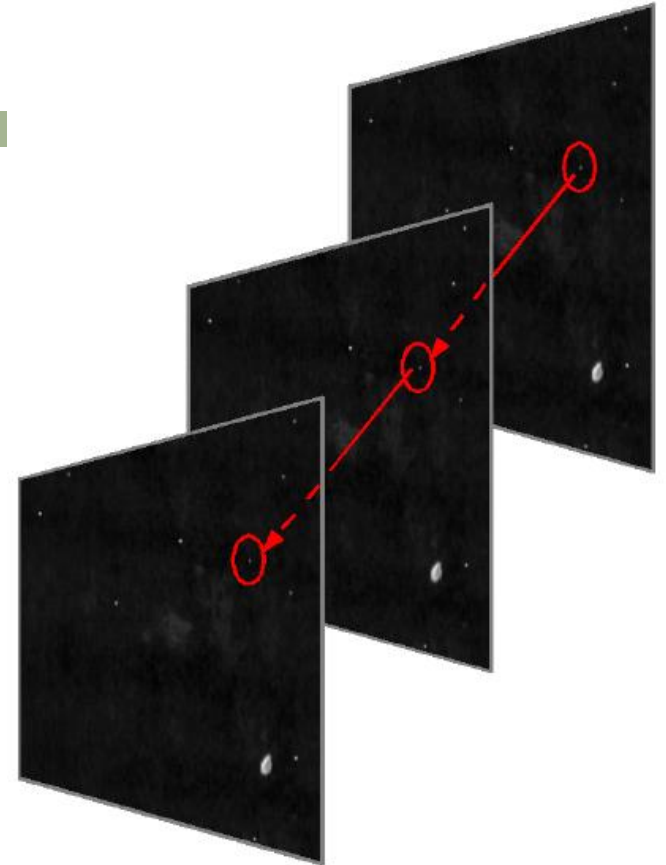
Tamás Budavári



Eventful Universe



ISSAC at



7/19/2012



Cross-Identification

One of the most fundamental analysis steps

What is the Right Question?

Tamás Budavári

- Cross-identification is a hard problem
 - ▣ Computationally, Scientifically & Statistically
 - ▣ Need symmetric n -way solution
 - ▣ Need reliable quality measure

- Same or not?
 - ▣ Distance threshold? Maximum likelihood?



Tabletop Astronomy

Tamás Budavári

- Imagine the observed sky has only 6 pixels
 - ▣ ***One object***: one die
 - ▣ ***Observing***: rolling a die
 - ▣ ***Locality***: die is loaded
 - ▣ ***Sky***: a bag of dice



Model Comparison: Same or Not?

Tamás Budavári

- **Crossmatch:** draw two dice with replacement
 - ▣ Same or not?



Model Comparison: Same or Not?

Tamás Budavári

- **Crossmatch:** draw two dice with replacement
 - ▣ 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?

- Model for loaded dice is matrix of probabilities

- ▣ E.g., loaded toward $l=1$

- ▣ Etc. for $l=2\dots 6$

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

- 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?

- Model for loaded dice is matrix of probabilities

- ▣ E.g., loaded toward $l=1$

- ▣ Etc. for $l=2\dots 6$

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

- 2-way case

- ▣ Same: $l_1 = l_2 = l$

$$L_{\text{same}} = \frac{1}{6} \sum_l P_l(\text{⊠}) P_l(\text{⊡})$$

- ▣ Not: $l_1 \text{ may } \neq l_2$

- n -way: same



Model Comparison: Same or Not?

- Model for loaded dice is matrix of probabilities

- ▣ E.g., loaded toward $l=1$

- ▣ Etc. for $l=2\dots 6$

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

- 2-way case

- ▣ Same: $l_1 = l_2 = l$

- ▣ Not: $l_1 \text{ may } \neq l_2$

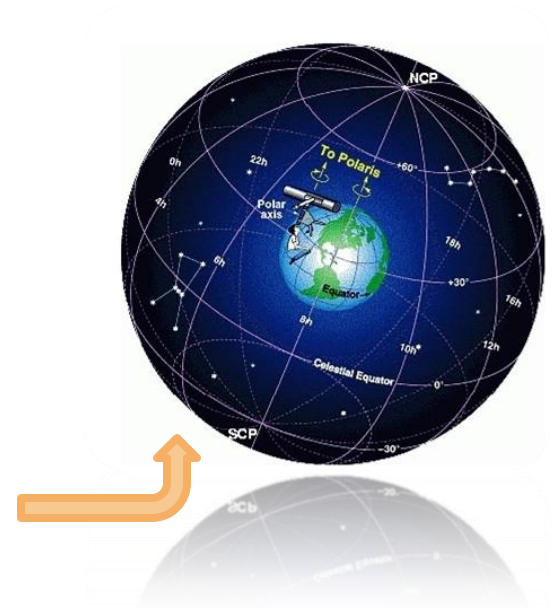
$$L_{\text{same}} = \frac{1}{6} \sum_l P_l(\odot) P_l(\oplus)$$
$$L_{\text{not}} = \left[\frac{1}{6} \sum_{l_1} P_{l_1}(\odot) \right] \left[\frac{1}{6} \sum_{l_2} P_{l_2}(\oplus) \right]$$



- n -way: same

Celestial Sphere

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



Modeling the Astrometry

Tamás Budavári

- Astrometric precision
 - A simple function

- Where on the sky?
 - Anywhere really...

$$p(\vec{x}|\vec{m}, M)$$



Same or Not?

OR □ The Bayes factor

$$B(H, K|D) = \frac{p(D|H)}{p(D|K)} \quad D = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}$$

SAME □ H : all observations of the same object at m

NOT □ K : might be from separate objects at $\{m_i\}$

Same or Not?

OR The Bayes factor

SAME H : all

NOT K : mig

Bayes' Rule

$$p(\theta|D, M) = \frac{p(\theta|M) p(D|\theta, M)}{p(D|M)}$$

Same or Not?

OR □ The Bayes factor

$$B(H, K|D) = \frac{p(D|H)}{p(D|K)} \quad D = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}$$

SAME □ H : all observations of the same object at m

NOT □ K : might be from separate objects at $\{m_i\}$

Same or Not?

OR □ The Bayes factor

$$B(H, K|D) = \frac{p(D|H)}{p(D|K)} \quad D = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}$$

SAME □ H : all observations of the same object at m

$$p(D|H) = \int p(\vec{m}|H) \prod_{i=1}^n p_i(\vec{x}_i|\vec{m}, H) d^3m$$

On the sky → (points to $p(\vec{m}|H)$)

(points to $p_i(\vec{x}_i|\vec{m}, H)$)

NOT □ K : might be from separate objects at $\{m_i\}$

Astrometry → (points to $p_i(\vec{x}_i|\vec{m}, H)$)

Same or Not?

OR □ The Bayes factor

$$B(H, K|D) = \frac{p(D|H)}{p(D|K)} \quad D = \{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n\}$$

SAME □ H : all observations of the same object at m

$$p(D|H) = \int p(\vec{m}|H) \prod_{i=1}^n p_i(\vec{x}_i|\vec{m}, H) d^3m$$

On the sky →

NOT □ K : might be from separate objects at $\{m_i\}$

$$p(D|K) = \prod_{i=1}^n \left\{ \int p(\vec{m}_i|K) p_i(\vec{x}_i|\vec{m}_i, K) d^3m_i \right\}$$

← *Astrometry*

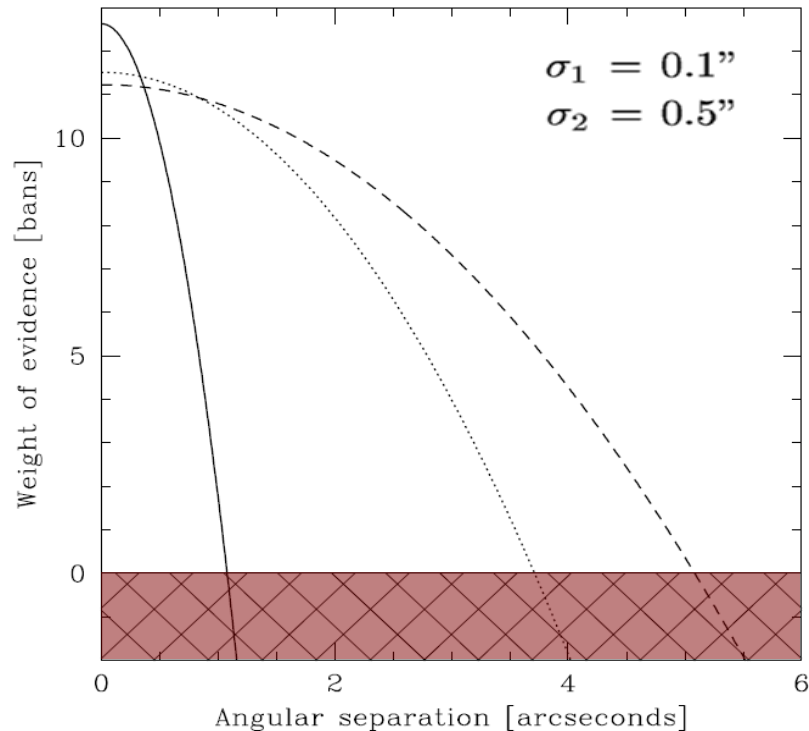
Analytic Results

Tamás Budavári

- Normal distribution
 - ▣ Flat and spherical
 - Gauss and Fisher

- 2-way results

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



Normal Distribution

□ Astrometric precision: $w = 1/\sigma^2$

□ Fisher distribution: $N(\vec{x}|w, \vec{m}) = \frac{w \delta(|\vec{x}|-1)}{4\pi \sinh w} \exp(w \vec{m} \vec{x})$

■ Analytic results:

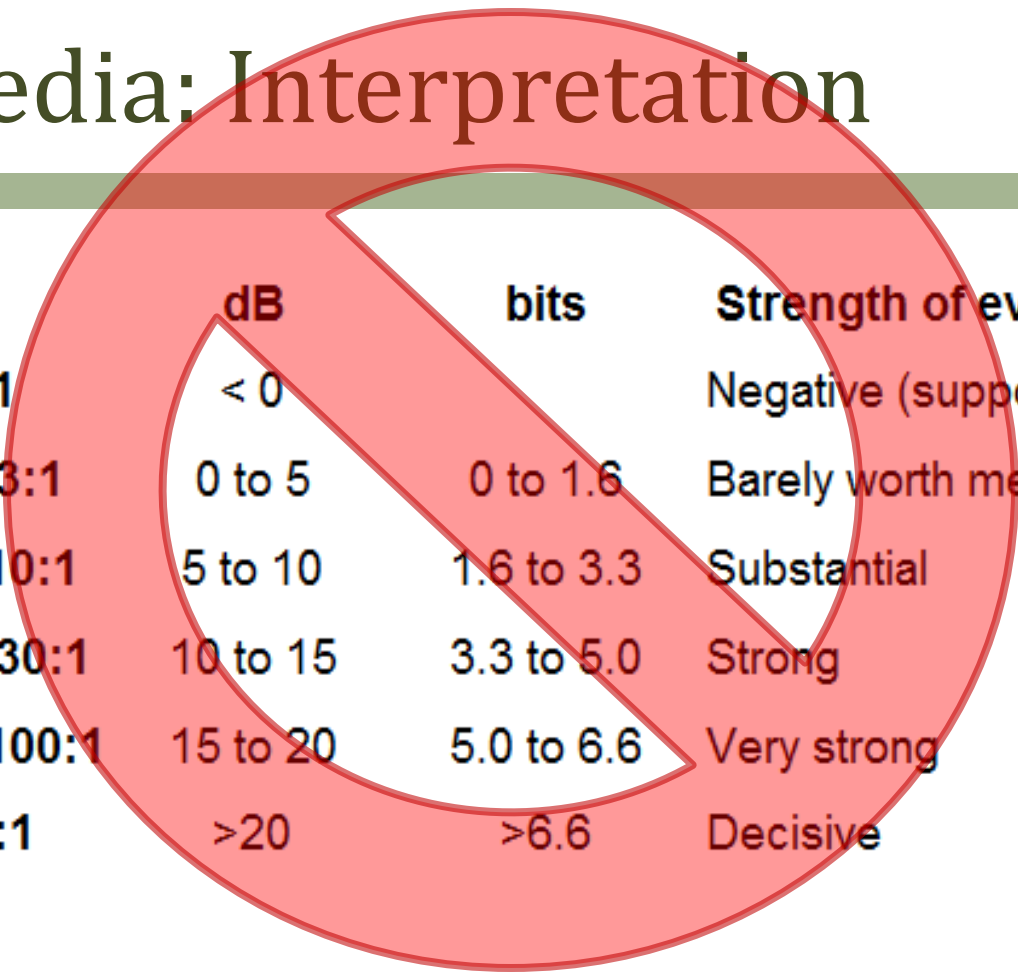
$$B(H, K|D) = \frac{\sinh w}{w} \prod_{i=1}^n \frac{w_i}{\sinh w_i}, \quad w = \left| \sum_{i=1}^n w_i \vec{x}_i \right|$$

■ For high accuracies:

$$= 2^{n-1} \frac{\prod w_i}{\sum w_i} \exp \left\{ -\frac{\sum_{i<j} w_i w_j \psi_{ij}^2}{2 \sum w_i} \right\}$$

Wikipedia: Interpretation

Tamás Budavári



<i>B</i>	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



Probability of a Match

Same or not?

From Priors to Posteriors

- Bayes factor is the connection

$$\frac{P(H|D)}{P(\bar{H}|D)} = \frac{P(H)p(D|H)}{P(\bar{H})p(D|\bar{H})}$$

$$\frac{P(H|D)}{P(\bar{H}|D)} = \frac{P(H)}{P(\bar{H})} B(H, \bar{H}|D)$$

$$\frac{P(H|D)}{1 - P(H|D)} = \frac{P(H)}{1 - P(H)} B(H, \bar{H}|D)$$

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



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?




From Priors to Posteriors

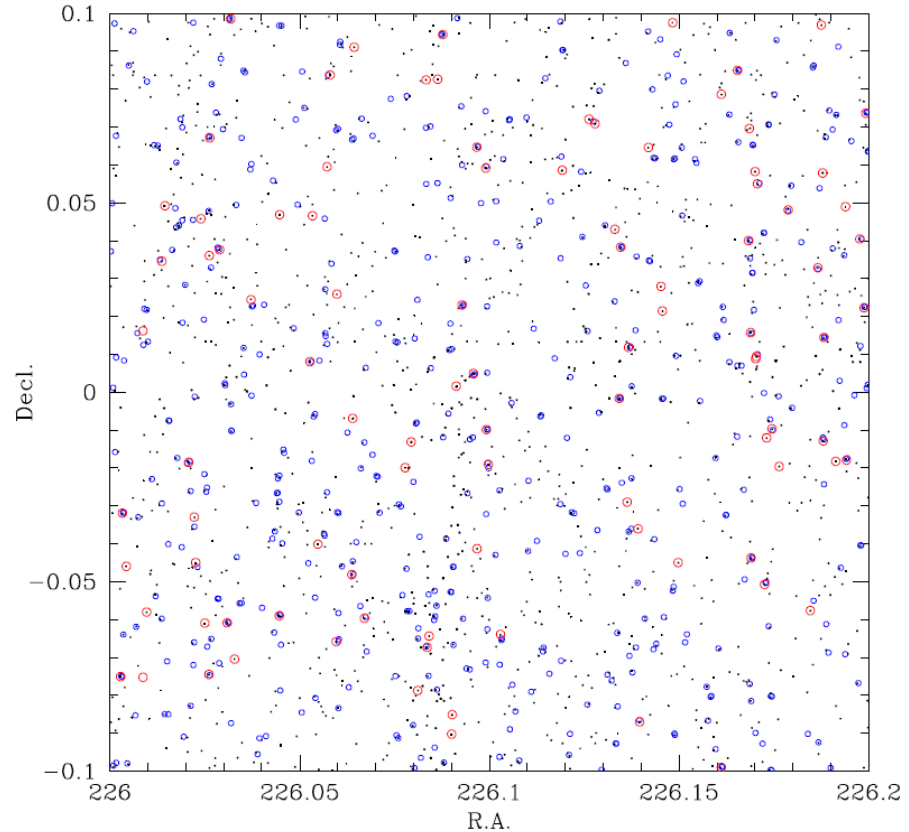
- Different selections

 - **Nearby** / Distant

 - **Red** / **Blue**

- But only 1 number

$$P_0 = \frac{N_{\star}}{\prod N_i}$$




Self-Consistent Estimates

- Prior has an unknown fudge-factor

- Educated guess $P(H|D) = \left[1 + \frac{1 - P(H)}{B P(H)} \right]^{-1}$
- Or solve for it:

$$\left. \begin{aligned} \sum P(H) &= N_{\star} \\ \sum P(H|D) &= N_{\star} \end{aligned} \right\} \text{ } \img alt="A circular diagram with a central star-like shape containing the symbol N_{\star}. The diagram is composed of several curved lines that form a complex, symmetrical pattern, resembling a stylized flower or a camera aperture. The symbol N_{\star} is placed in the center of the star shape." data-bbox="530 550 742 896"/>$$

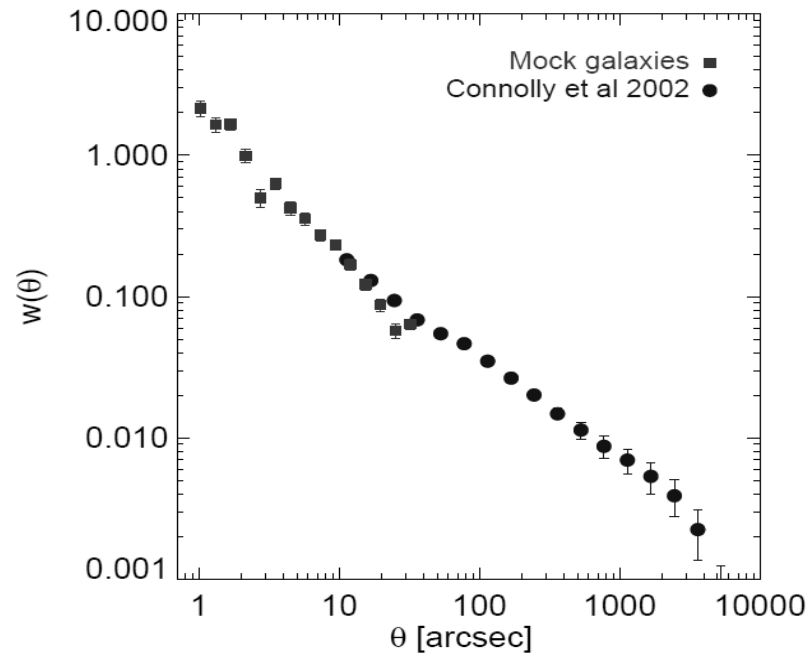
TB & Szalay (2008)

Simulations

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



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

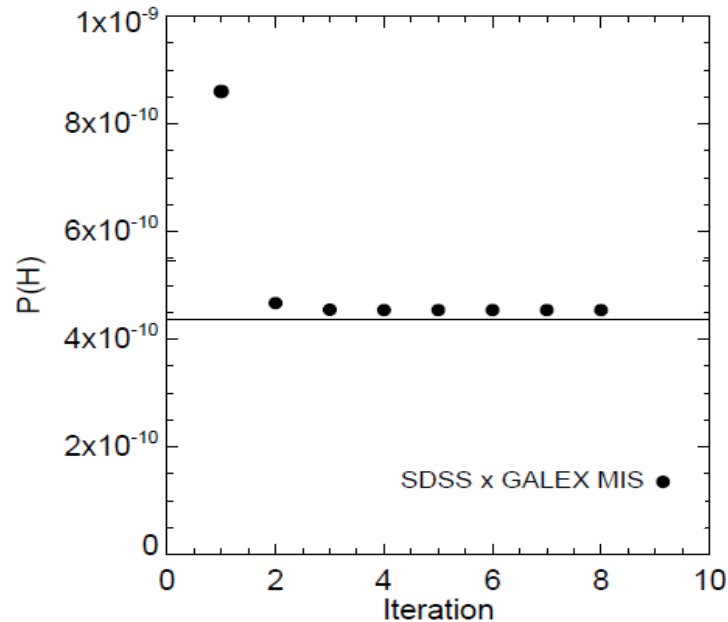


Simulations

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

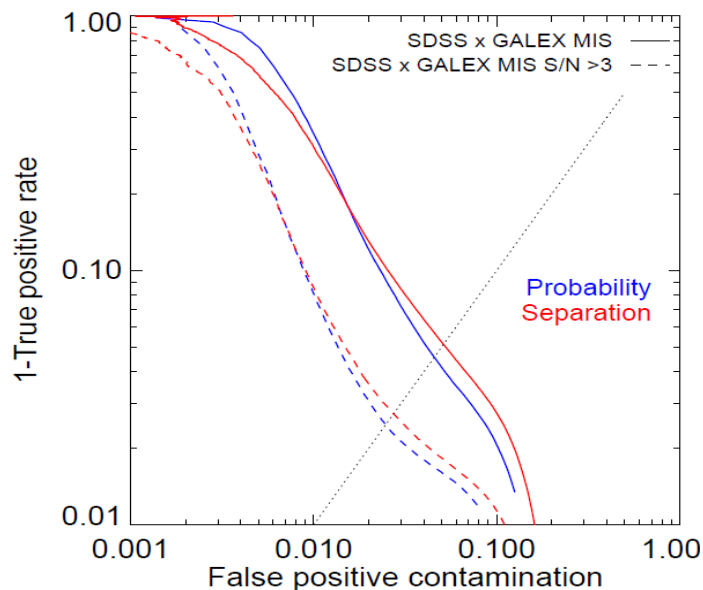


- Simulated sources
 - Subsets: N_1 N_2
 - Overlap: N_\star



Simulations

Quality



Multiple matches

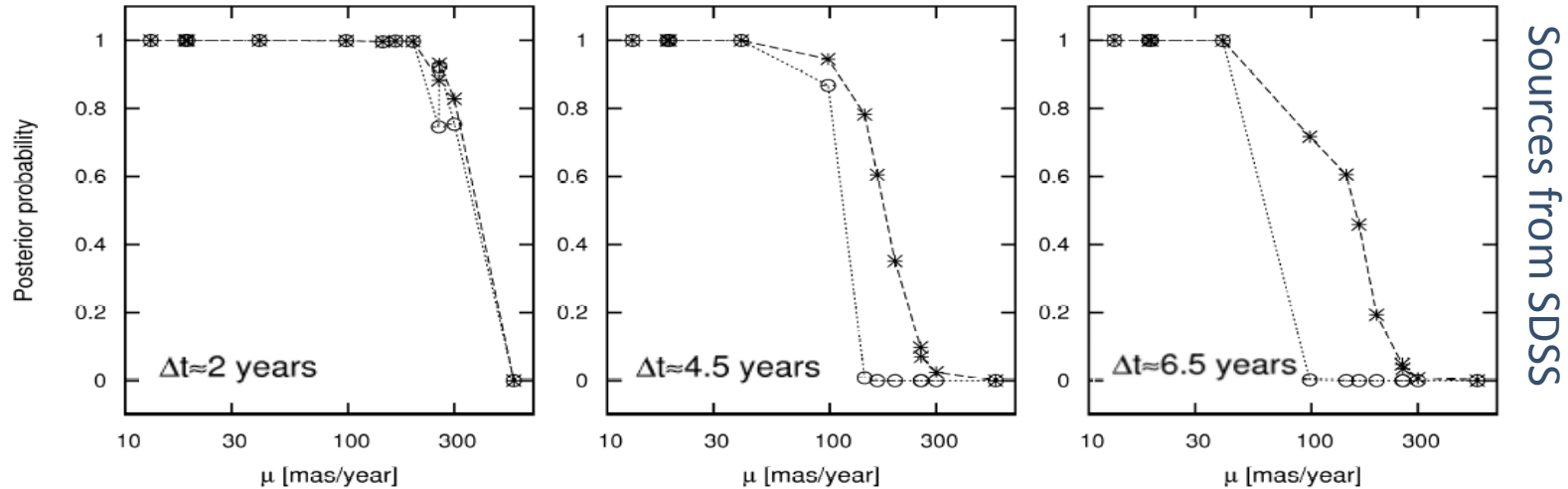
GALEX	SDSS		
	1	2	Many
1	74.061 (75.870)	21.007 (18.595)	2.577 (2.469)
2	1.146 (2.253)	1.006 (0.697)	0.188 (0.102)
Many	0.006 (0.009)	0.007 (0.004)	0.002 (0.001)

Explained by simple model
of point sources!

Heinis, TB, Szalay (2009)

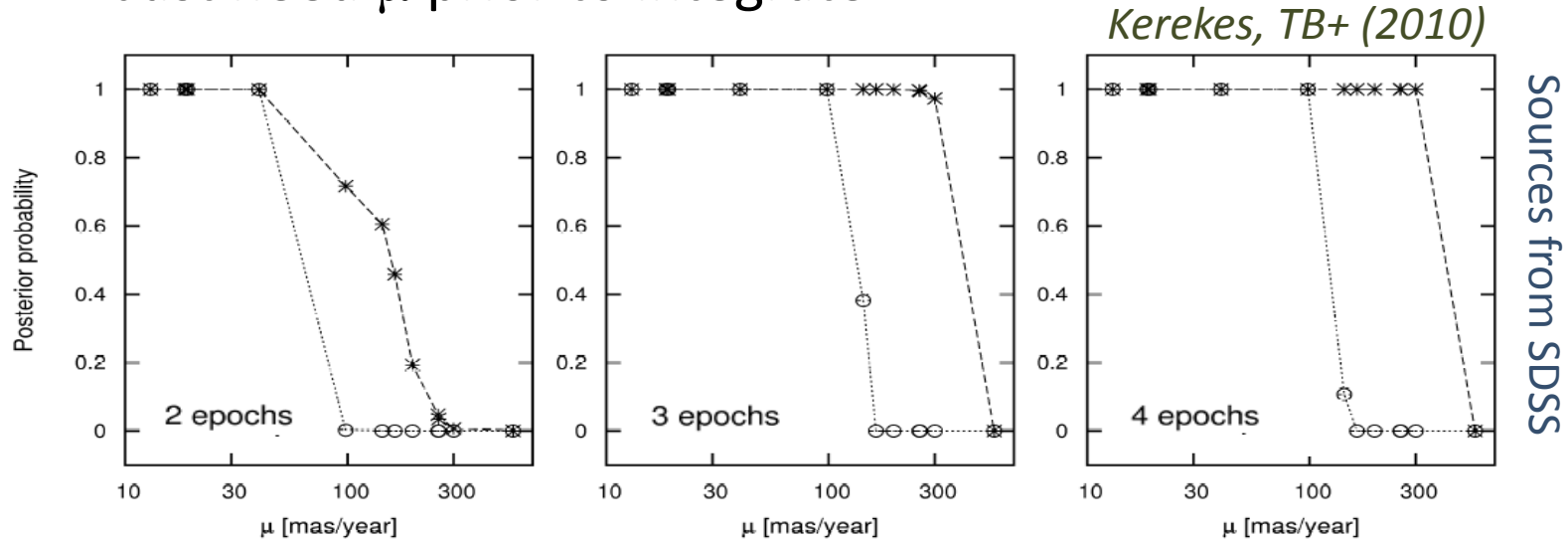
Proper Motion

- Same hypotheses but different parameters
 - ▣ Just need μ prior to integrate



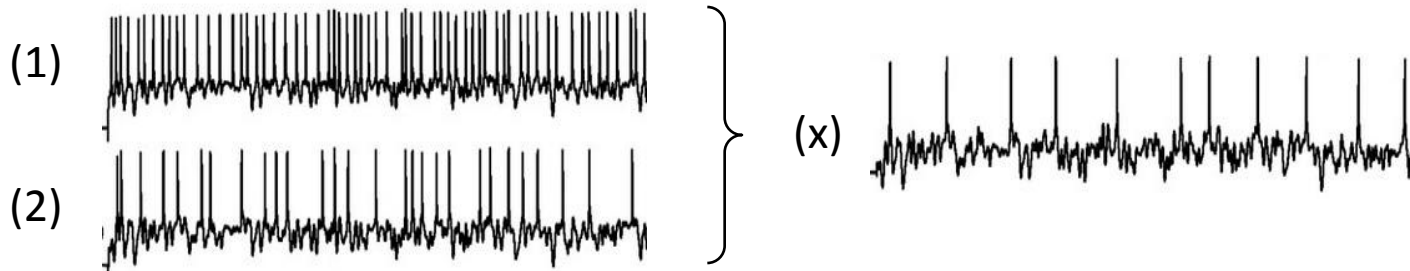
Proper Motion

- Same hypotheses but different parameters
 - ▣ Just need μ prior to integrate

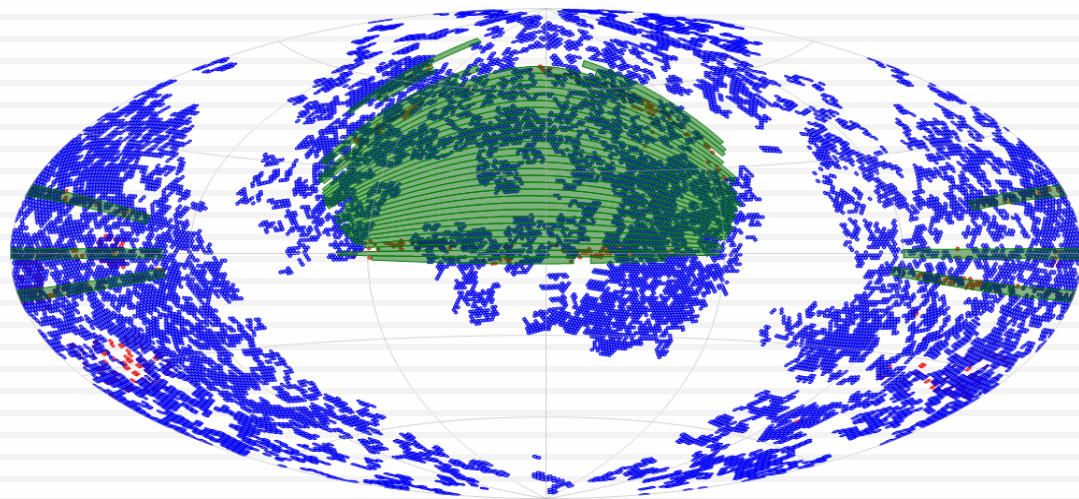


Matching Events

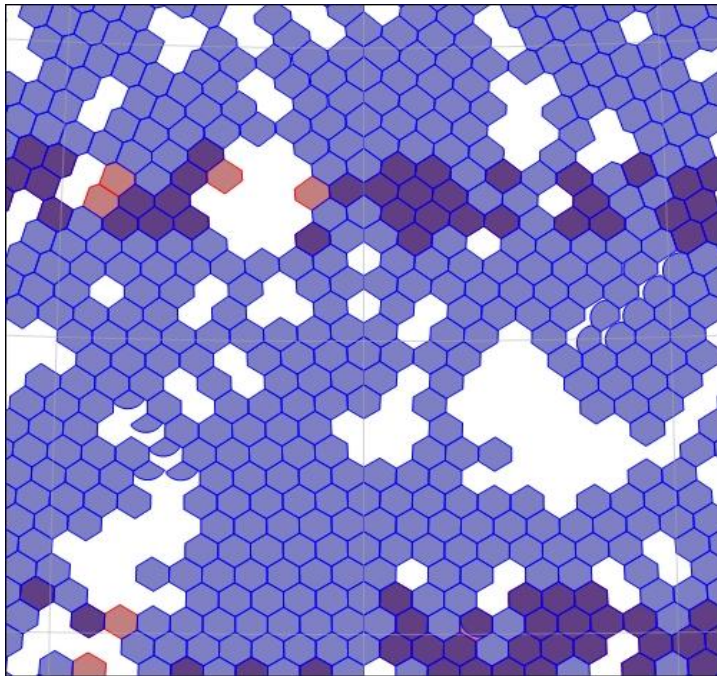
- Streams of events in time and space
 - ▣ E.g., thresholded peaks in signal-to-noise



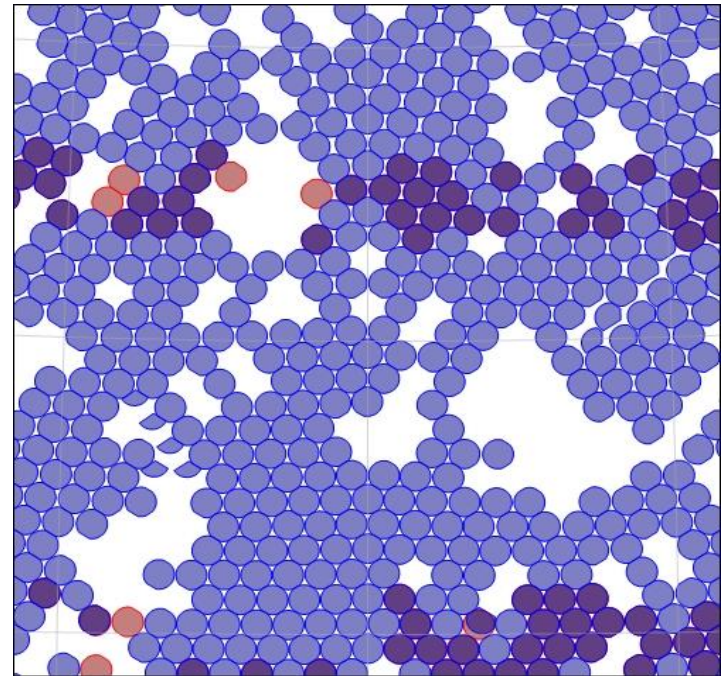
Dropouts from Sky Coverage



Drawing with Equations



$r = 0.6^\circ$



$r = 0.5^\circ$



Matching in Practice

Open SkyQuery

- Following our 1st prototype
- Successful
- Not bayesian
- Limitations

The screenshot shows the Open SkyQuery web interface. The browser address bar displays `openskyquery.net/Sky/SkySite/browse/Browse.aspx`. The page header includes the NVO logo (National Virtual Observatory) and navigation tabs for Simple Query, Advanced Query, Import Data, Tutorial, Help, and Contact Us. A 'Hosted By' badge for Johns Hopkins University is visible in the top right.

The main interface is divided into three panels:

- Nodes:** A vertical list of astronomical survey names, each with a '+' icon for selection. The list includes Rosat, DLS, RC3, SDSS, SDSSDR2, SDSSDR3, SDSSDR4, SDSSDR5, SDSSDR6, TwoDf, Twoqz, USNOB, GOODS, HDFN, HDFS, UDF, TWOMASS, IRAS, PSCz, FIRST, NVSS, FUSE, LGATheory, and NDWFS.
- Query Editor:** A central text area containing a SQL query:

```
SELECT o.objid, o.ra,  
       o.dec, o.r, o.type,  
       t.objid, t.ra, t.dec  
FROM  
  SDSS:PhotoPrimary o, TWOMASS:PhotoPrimary t  
WHERE XMATCH(o, t) < 3.5 AND  
       Region('CIRCLE J2000 181.3 -0.76 6.5') AND  
       o.type = 3
```
- Sample Queries:** A list of pre-defined queries such as XMatch/Region, XMatch/Region 2, Three Node Match, Brown Dwarf Search, MyData XMatch (upload), Xmatch t.* (upload), ABELL Xmatch (upload), Single Node Query, and Single Node Join.

At the bottom of the interface, there is a welcome message and instructions: "Welcome to the Open SkyQuery interactive query builder. You should see a parsed, clickable version of your entered query in the pane directly above this one. If instead you see 'Query is empty', this means that builder needs a node or two to get started. You can add nodes to the builder by clicking the desired node's '+' icon in the left panel. Once you have some sql in the above panel, you can then click on a token in that query to pull up a menu with options appropriate for that specific token. For example, one way to select an additional column from a mythical 'mytable' is to click on 'mytable' and then chose 'Add Selection', then pick the desired column from the given choices. You can switch between 'edit' and 'build' modes at any time by using the tabs at the top of the query panel. Your changes from one will carry over to the other. Most menu options have additional mouse-over info."

SkyQuery – The 3rd Generation

Tamás Budavári

- Dynamic federation of astronomy databases
 - Query the collection as if they were one
- The 3rd gen tool coming this summer
 - Cluster of machines running partitioned jobs
 - Proper probabilistic exec with variable errors

SkyQuery

- Almost pure standard SQL

```
SELECT p.ObjID, p.RA, p.Dec,  
       s.BestObjID, s.SpecObjID, s.RA, s.Dec  
INTO xtest  
FROM SDSSDR7:PhotoObjAll AS p  
     CROSS JOIN SDSSDR7:SpecObjAll AS s  
WHERE  
     p.RA BETWEEN 0 AND 5  
     AND p.Dec > -9999  
     AND s.Dec > -9999  
     AND s.RA > -9999
```

SkyQuery

- Almost pure standard SQL

```
SELECT p.ObjID, p.RA, p.Dec,  
       s.BestObjID, s.SpecObjID, s.RA, s.Dec  
INTO xtest  
FROM SDSSDR7:PhotoObjAll AS p  
     CROSS JOIN SDSSDR7:SpecObjAll AS s
```



```
WHERE  
      p.RA BETWEEN 0 AND 5  
      AND p.Dec > -9999  
      AND s.Dec > -9999  
      AND s.RA > -9999
```

SkyQuery

- Almost pure standard SQL

```
SELECT p.ObjID, p.RA, p.Dec,  
       s.BestObjID, s.SpecObjID, s.RA, s.Dec  
INTO xtest  
FROM SDSSDR7:PhotoObjAll AS p  
     CROSS JOIN SDSSDR7:SpecObjAll AS s  
XMATCH BAYESIAN AS x  
     MUST p ON Point(p.RA, p.Dec), 0.1, 0.1, 0.1  
     MUST s ON Point(s.RA, s.Dec), 0.1, 0.1, 0.1  
HAVING LIMIT 1e3  
WHERE  
     p.RA BETWEEN 0 AND 5  
     AND p.Dec > -9999  
     AND s.Dec > -9999  
     AND s.RA > -9999
```

SkyQuery

- Almost pure standard SQL
- Added XMATCH
 - ▣ Verifiable
 - ▣ Flexible

```
SELECT p.ObjID, p.RA, p.Dec,  
       s.BestObjID, s.SpecObjID, s.RA, s.Dec  
INTO xtest  
FROM SDSSDR7:PhotoObjAll AS p  
     CROSS JOIN SDSSDR7:SpecObjAll AS s  
XMATCH BAYESIAN AS x  
      MUST p ON Point(p.RA, p.Dec), 0.1, 0.1, 0.1  
      MUST s ON Point(s.RA, s.Dec), 0.1, 0.1, 0.1  
      HAVING LIMIT 1e3  
WHERE  
      p.RA BETWEEN 0 AND 5  
      AND p.Dec > -9999  
      AND s.Dec > -9999  
      AND s.RA > -9999
```

[home](#)[schema
browser](#)[query](#)[job history](#)[my db](#)[import](#)[output](#)[help](#)Output: Task name:

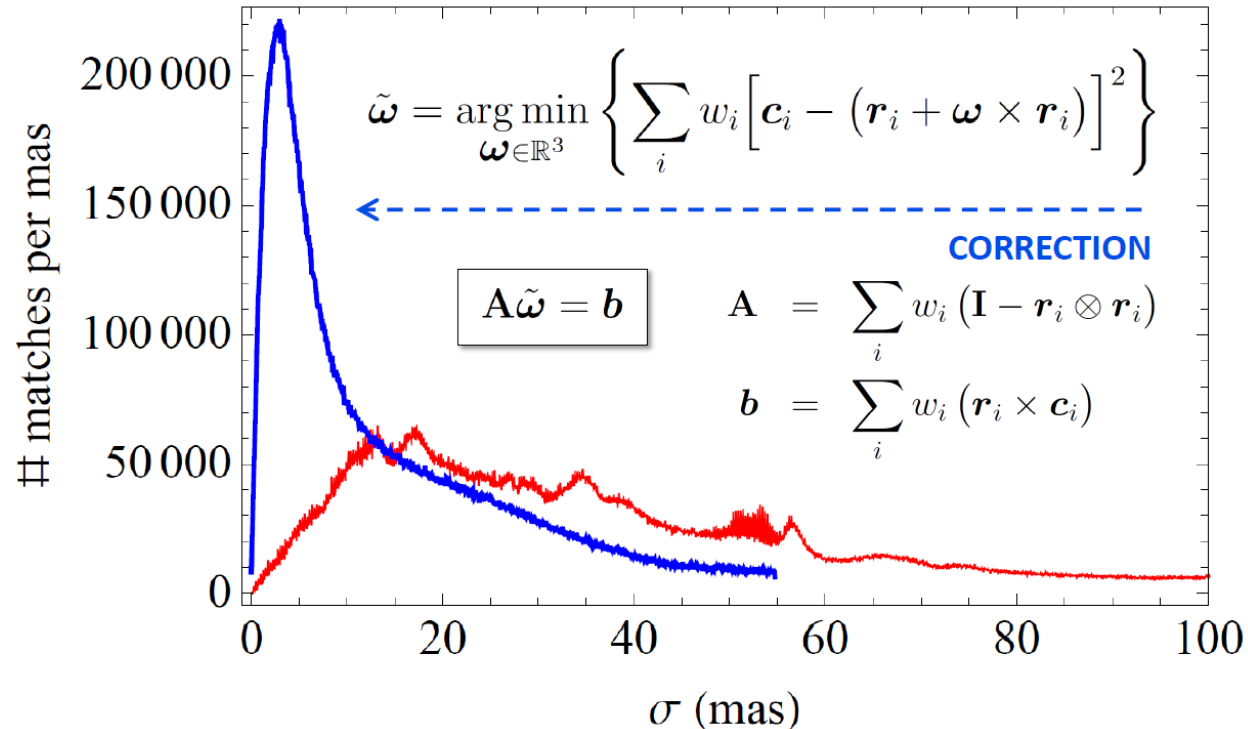
```
1 SELECT p.ObjID, p.RA, p.Dec,
2         s.BestObjID, s.SpecObjID, s.RA, s.Dec
3 INTO xtest
4 FROM SDSSDR7:PhotoObjAll AS p
5     CROSS JOIN SDSSDR7:SpecObjAll AS s
6     XMATCH BAYESIAN AS x
7     MUST p ON Point(p.RA, p.Dec), 0.1, 0.1, 0.1
8     MUST s ON Point(s.RA, s.Dec), 0.1, 0.1, 0.1
9     HAVING LIMIT 1e3
10 WHERE
11     p.RA BETWEEN 0 AND 5
12     AND p.Dec > -9999
13     AND s.Dec > -9999
14     AND s.RA > -9999
```

HST Crossmatch Catalog

β RELEASE AT AAS

Tamás Budavári

- SQL pipeline
- Astrometric correction
 - Subpixel precision



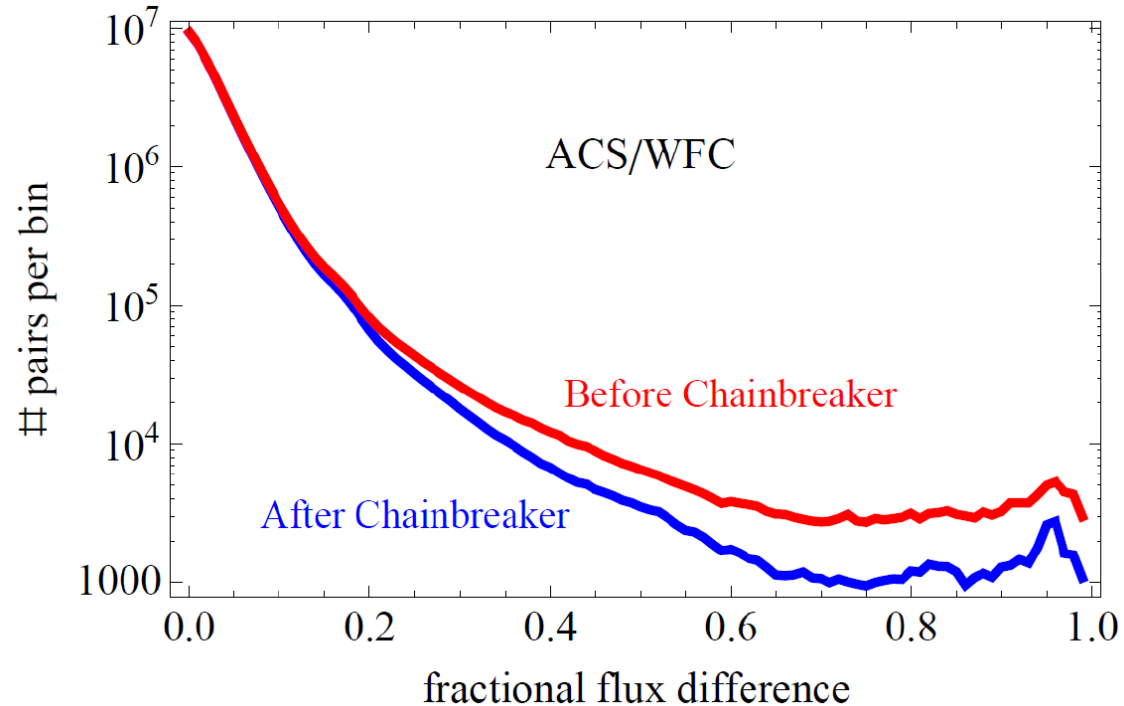
HST Crossmatch Catalog

β RELEASE AT AAS

Tamás Budavári

- FoF groups
 - ▣ Possible chains

- Bayesian model selection
 - ▣ Chainbreaker

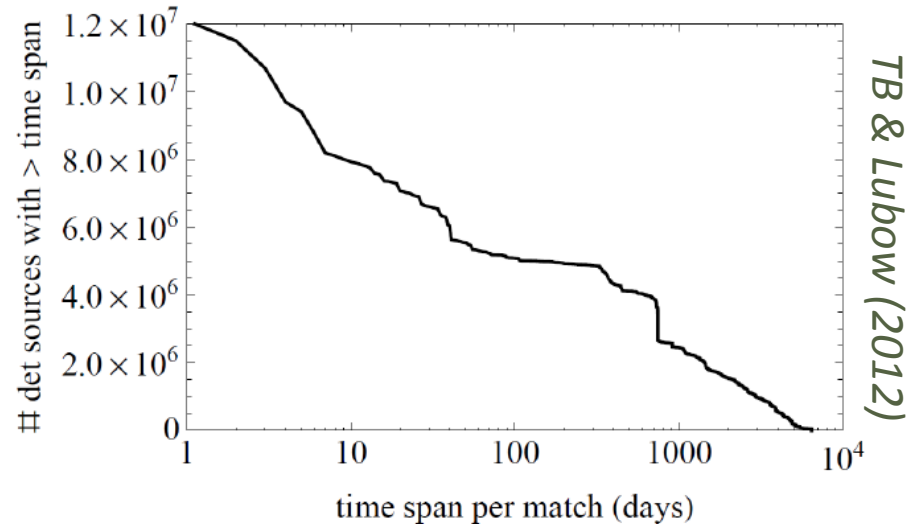
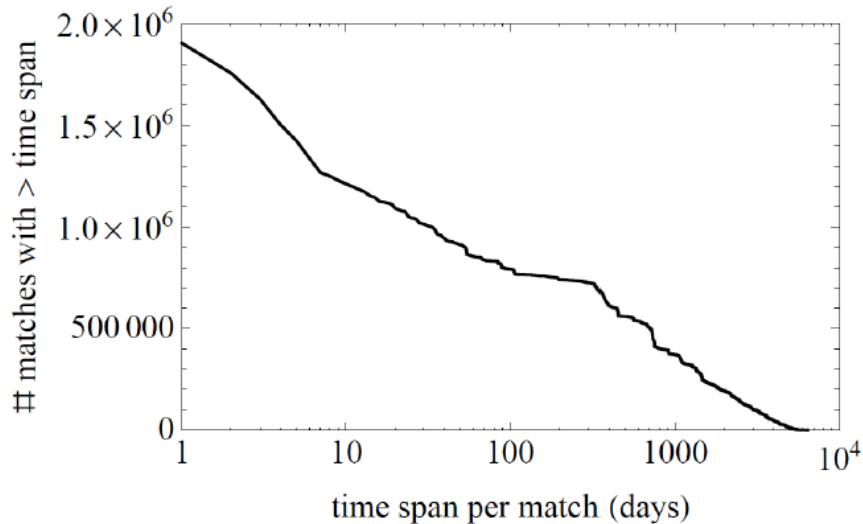


HST Crossmatch Catalog

β RELEASE AT AAS

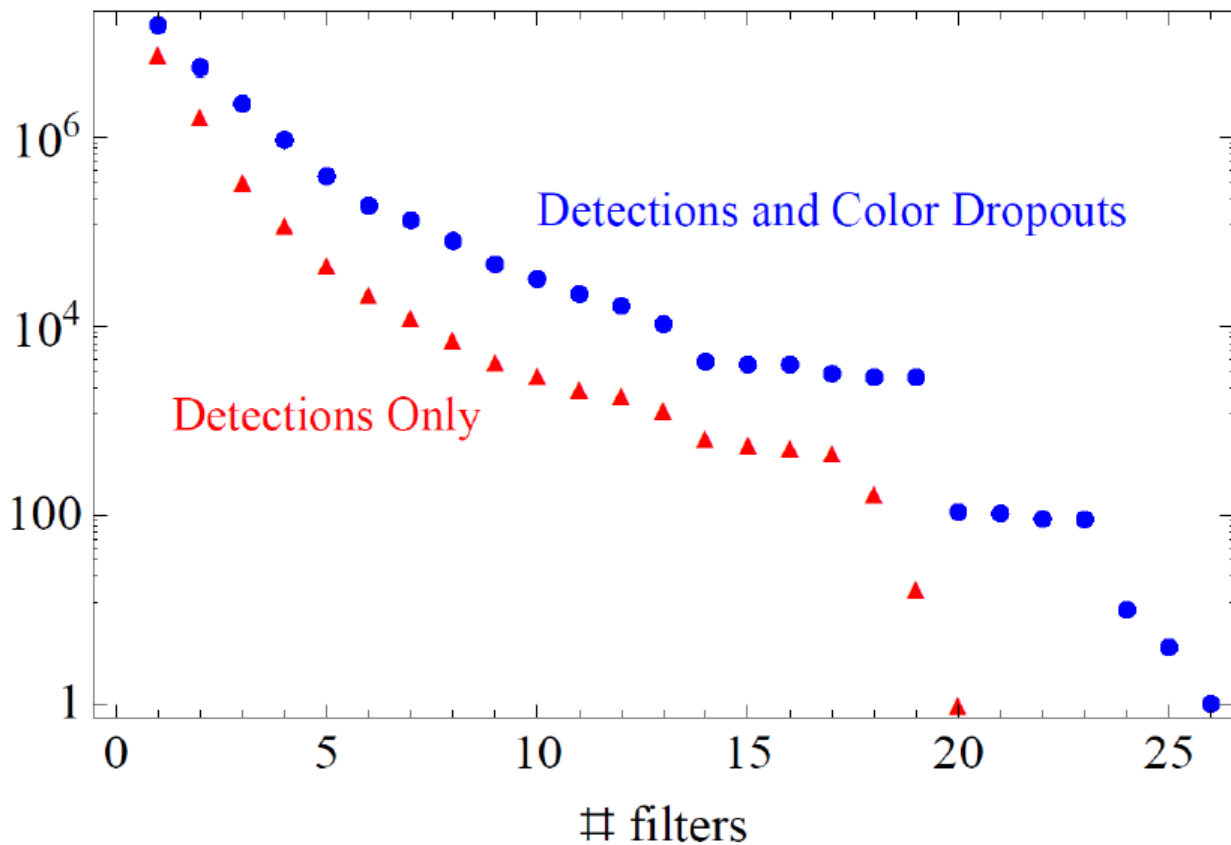
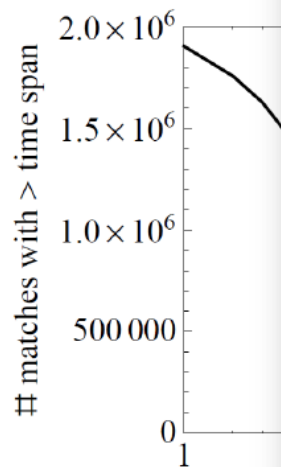
Tamás Budavári

- Lots of matching sources during HST's long life



HST

□ Lots



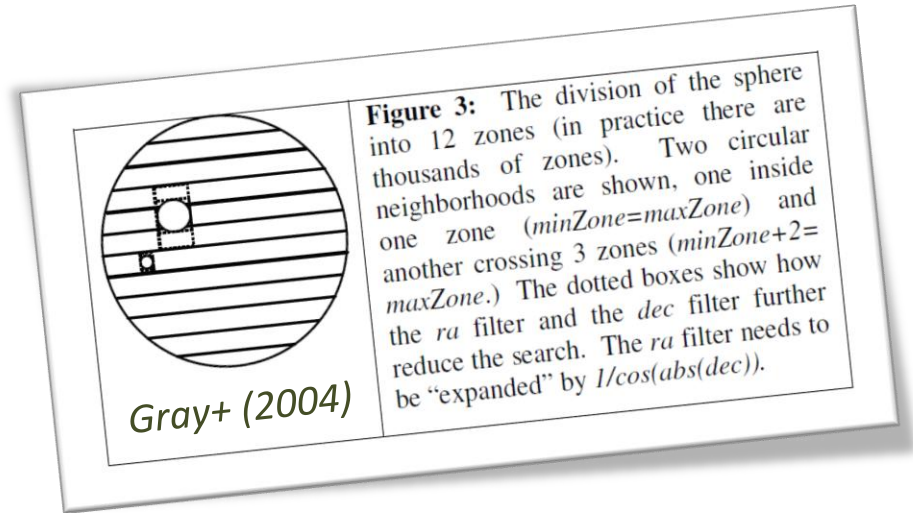
nás Budavári

TB & Lubow (2012)

Zone Algorithm

Tamás Budavári

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



Parallel on GPUs

- Recent Github release
 - ▣ Multi-GPU implementation
- Search in 5" – great perf!
 - ▣ NVIDIA GTX 480 1.5GB
 - 29M×29M in **11 seconds**
 - ▣ C2050 Teslas
 - 400M×150M in **3 minutes**

```
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
...
|
```

