BAYESIAN CROSS-IDENTIFICATION IN ASTRONOMY

10/7/2012

Tamás Budavári / The Johns Hopkins University

Recording Observations

Tamás Budavári

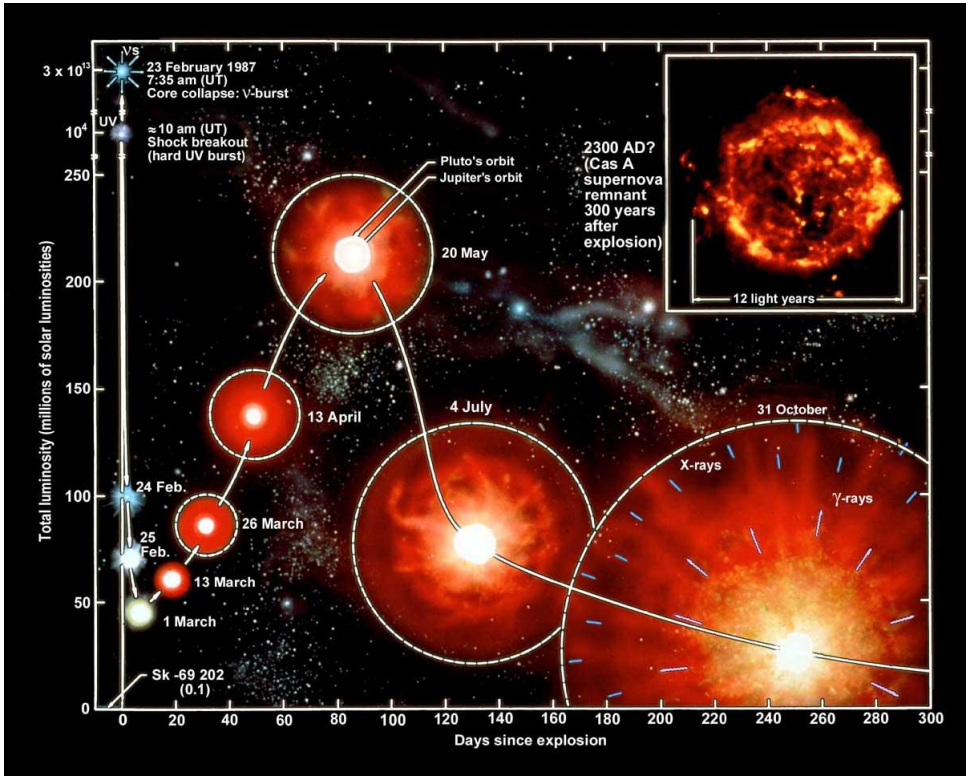
- Whirlpool Galaxy – M51

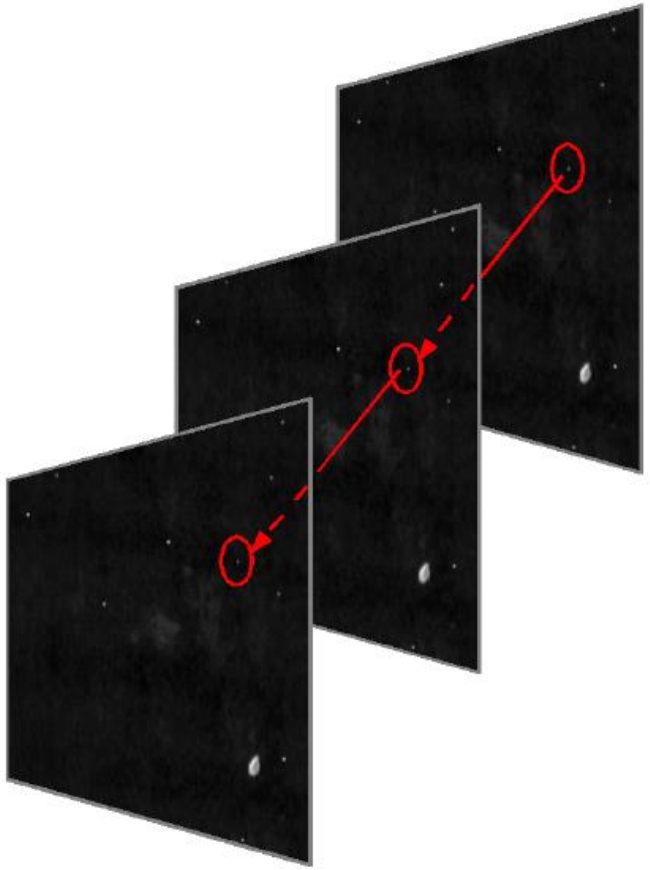
  Discovered by

  Charles Messier (1773)

Multicolor Universe

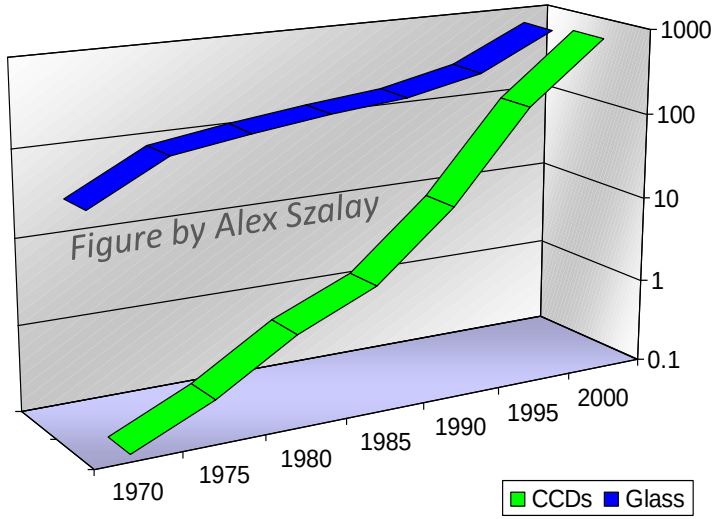
Eventful Universe





Trends in Astronomy

- Exponential growth of data
  - Moore's law in detectors

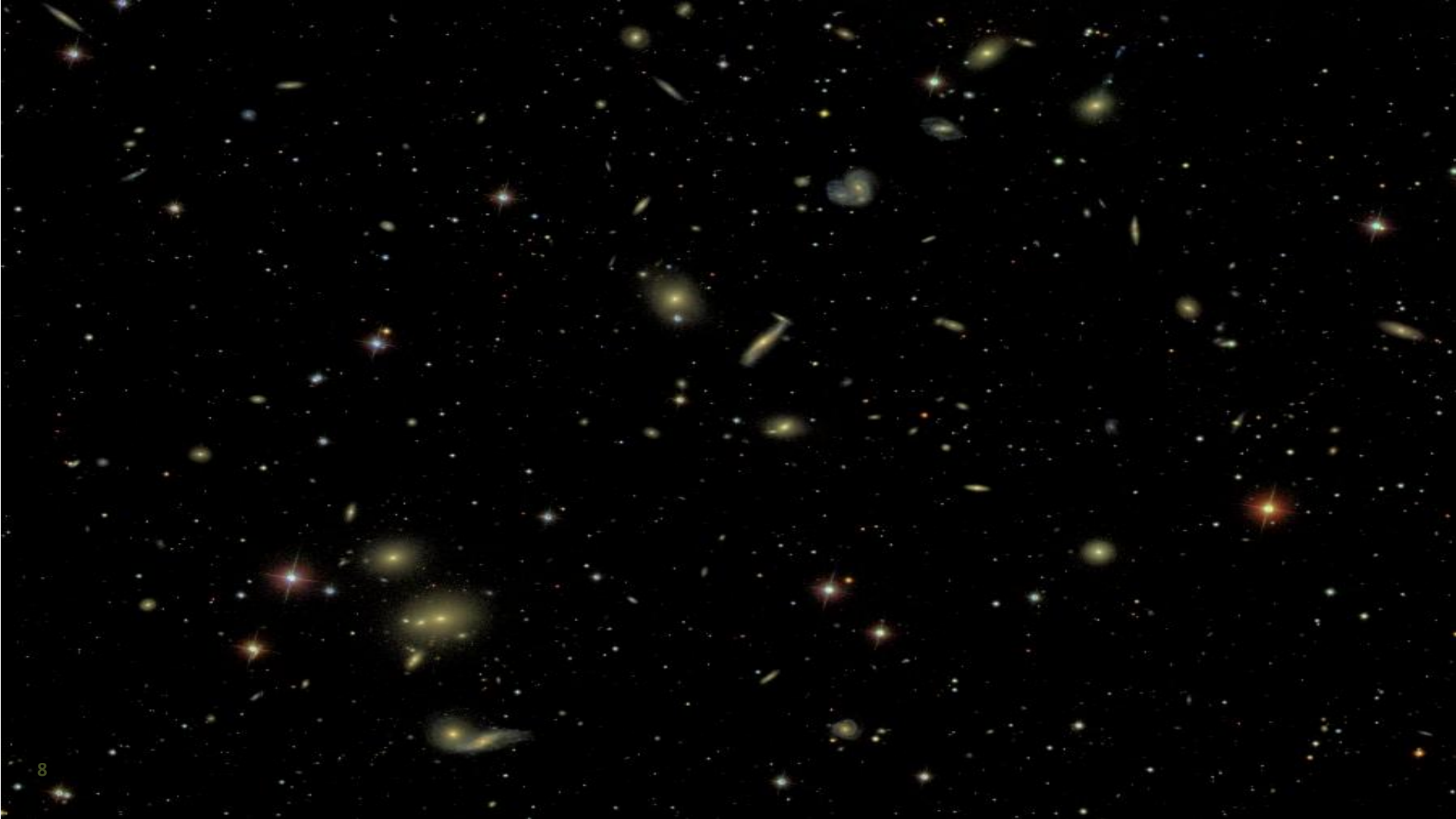


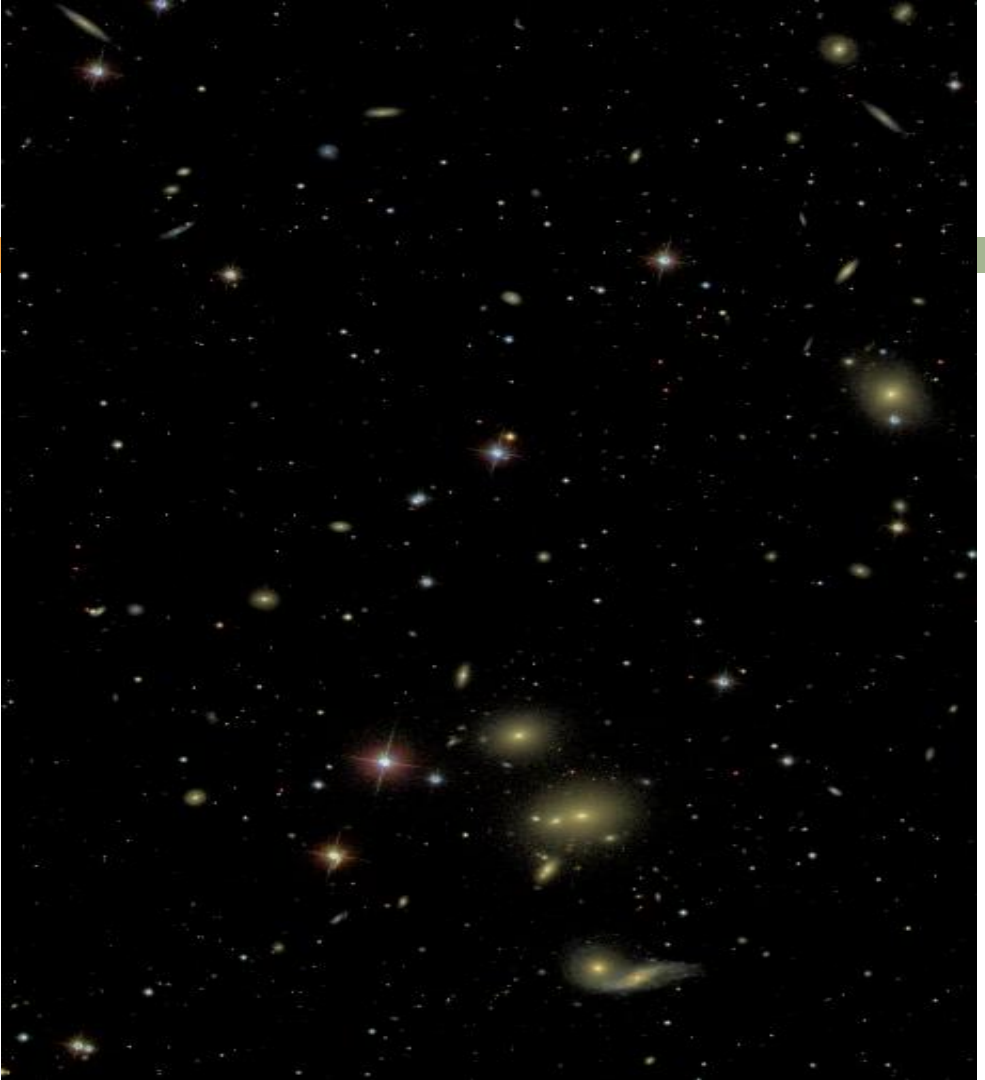
Sloan Digital Sky Survey

- Cosmic Genome Project 2001-2010
  - 500M rows, 400+ cols ~ 18TB
  - 30TB of images from 30 CCDs
  - Software revolution in astro
  - Astronomers learn SQL
    - Cannot look at the data anymore

The New Generations

- Large Synoptic Survey Telescope [OPTICAL]
  - 3 trillion rows, 200+ attributes, 100+ tables ~ 30PB
  - 60PB of images, 3.2 Gpix cam
- Square Kilometer Array [RADIO]
  - Processing limited





Keeping Up?

Tamás Budavári

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

More is Different

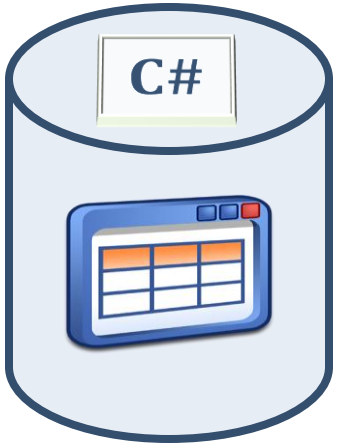
- Lots of opportunities
- Lots of challenges
- Lots of problems

- New approaches
- New algorithms
- New tools

- New computers

Do More at the Data

- Statistics on remote resources
  - Fine-tune SQL Server for astronomy
  - Analyses in and driven by SQL
- SDSS catalog archive and more
  - GALEX, HLA, UKIDSS, PanSTARRS, …



Sloan Digital Sky Survey / SkyServer

Due to system maintenance this site will be unavailable Thursday March 17th from 7:00AM central until 7:30AM central. We apologize for the inconvenience.

Welcome to the **DR7 site**!!!

This website presents data from the Sloan Digital Sky Survey, a project to make a map of a large part of the universe. We would like to show you the beauty of the universe, and share with you our excitement as we build the largest map in the history of the world.

News

The site hosts data from **Data Release 7 (DR7)**. **What's new in DR7**, **what's new on this site**, and **known problems**. **More...**

For Astronomers

A separate branch of this website for professional astronomers (English)

More...

SkyServer Tools

- Famous places
- Get images
- Visual Tools
- Explore
- Search
- Object Cross-ID
- CasJobs

Science Projects

- Basic
- Advanced
- Challenges
- For Kids
- Games and Contests
- Teachers
- Links to other projects

Info Links

- About Astronomy
- About the SDSS
- About the SkyServer
- SDSS Data Release 7
- SDSS Project Website
- Open SkyQuery
- Images of RC3 Galaxies

Help

- Getting Started
- FAQ
- How To
- Glossary
- Schema Browser
- Sample SQL Queries
- Details of SDSS Data

SDSS is supported by

Powered by Microsoft

MyDB | Local Only

Views
Tables
Functions
Procedures

Sort by... | All selected...

	Rows	kB	Name
☐	993	40	dist
☐	949	200	dr3tile
☐	92,082	16,640	DWSZ_R17_primary
☐	92,082	13,376	DWSZ_R25
☐	14,400	904	galexfield1
☐	3	16	galexfield2
☐	1	16	halfspace
☐	1	16	MyTable
☐	1	16	MyTable_0
☐	1	16	MyTable_1
☐	23	16	MyTable_2
☐	1,000	40	radec
☐	1	16	roomba
☐	20	16	roomba2
☐	1	16	stat
☐	11	24	test1

Tamas Budavari 's MyDB

20,992 kB of 100,000 kB used

From this page you can get various information about the contents of both your MyDB and shared tables within your groups. Click the left table links to get information about a specific table, such as rows, columns or size. From the table pages you can also perform various table-specific tasks, such as:

- Download a table
- Mangage your group tables
- Rename a table
- Drop a table

Sizes are approximations only.
Row counts are approximations only. For exact value run a count.
There's always some overhead, even empty MyDB's take up space.
Group tables do not count towards your MyDB size limit.

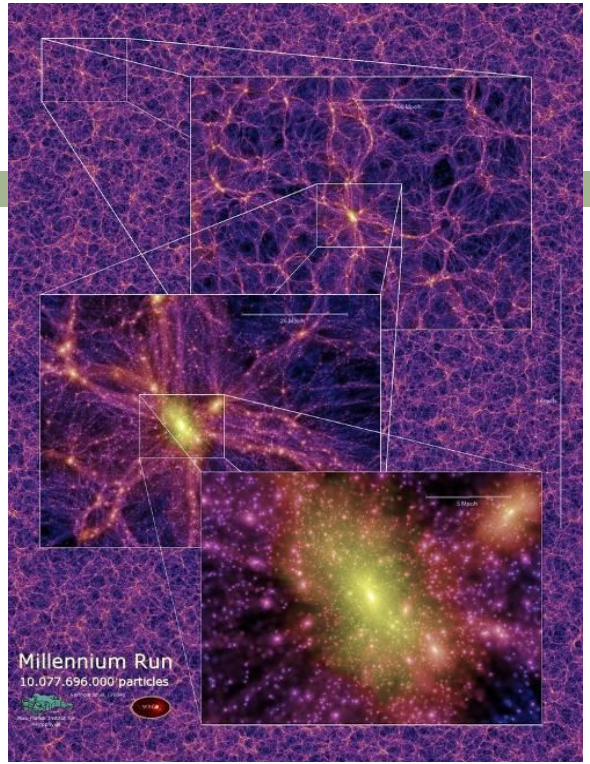
Contact

$Name: v3_5_16 $,$Revision: 1.64 $, Last modified: Tuesday, January 27, 2009 at 3:19:32 PM

Storing Simulations

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

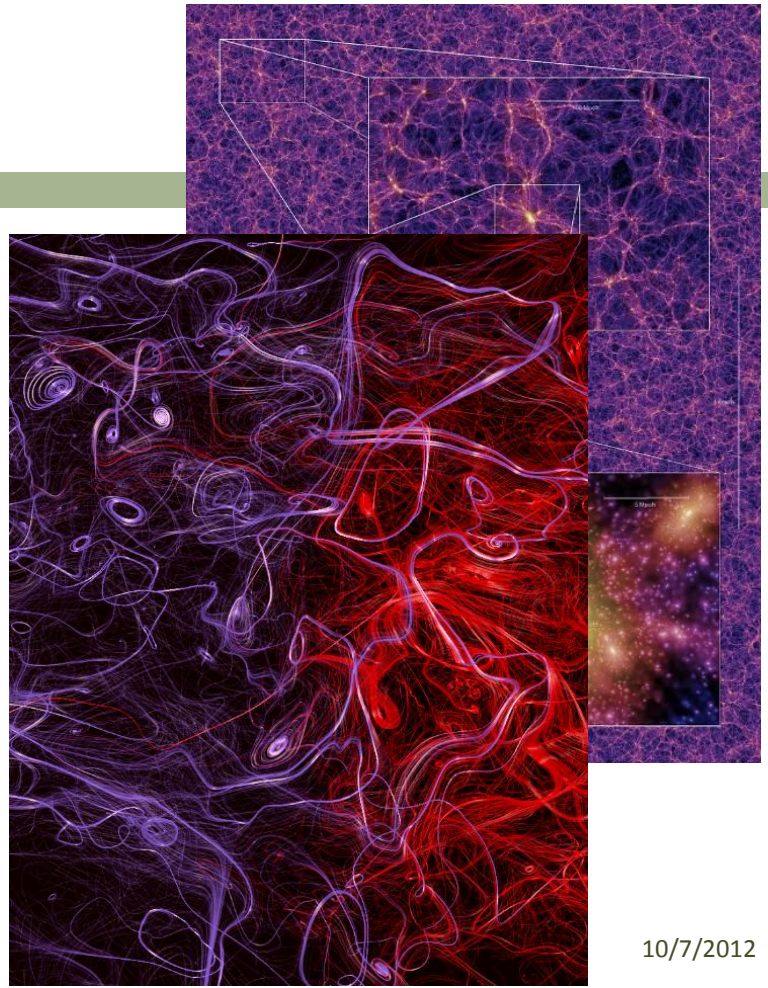
- Turbulence simulations (JHU)
  - 1024^4 grid, 27TB



Storing Simulations

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

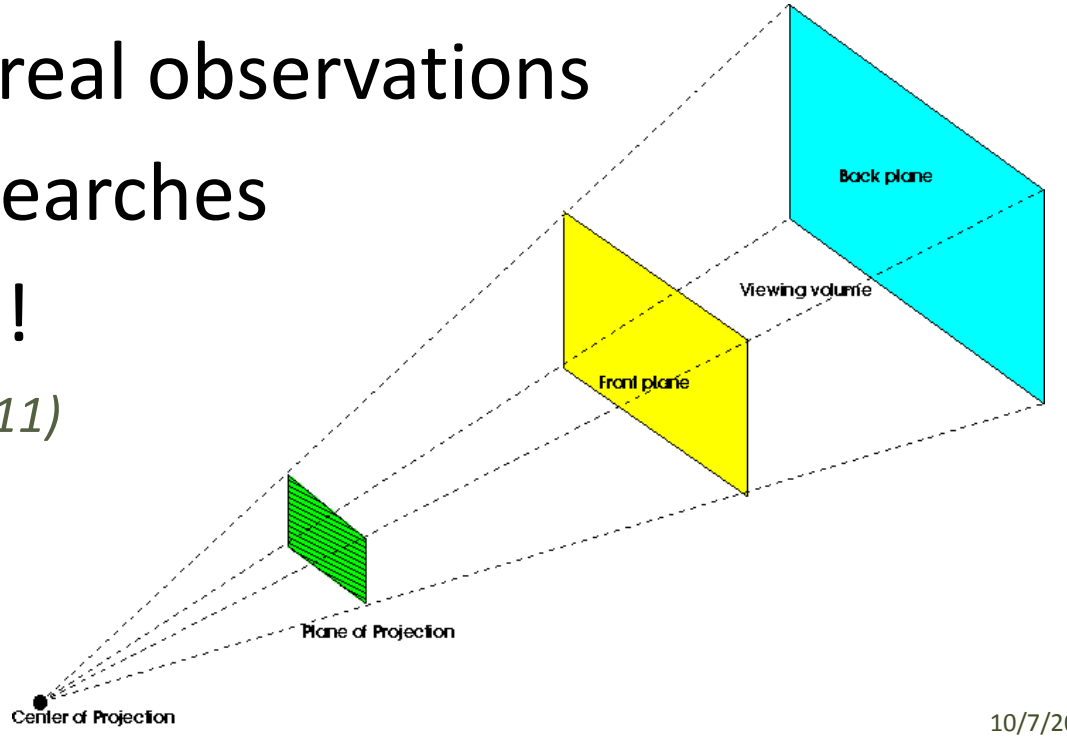
- Turbulence simulations (JHU)
  - 1024^4 grid, 27TB



Observing Simulations

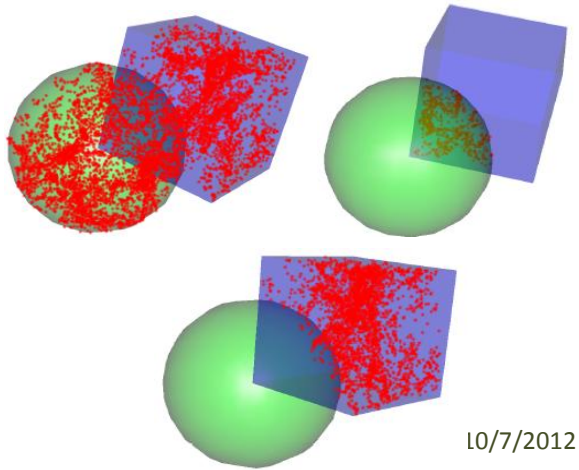
- Comparison to real observations
- Lots of spatial searches
- In the database!

Lemson, TB & Szalay (2011)



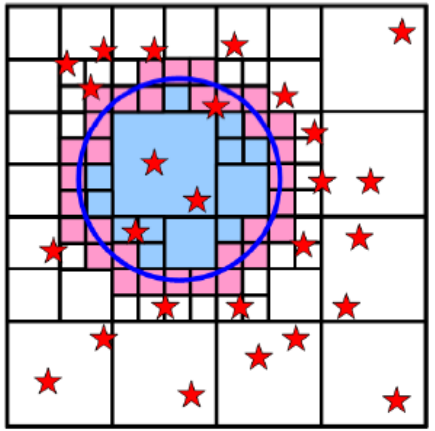
Fast Searches

- Map space-filling curves to database indices
  - Hierarchical Triangular Mesh – on the unit sphere
  - Peano-Hilbert / Morton curves – in 3D
- Combine with query shapes
  - Build from primitives

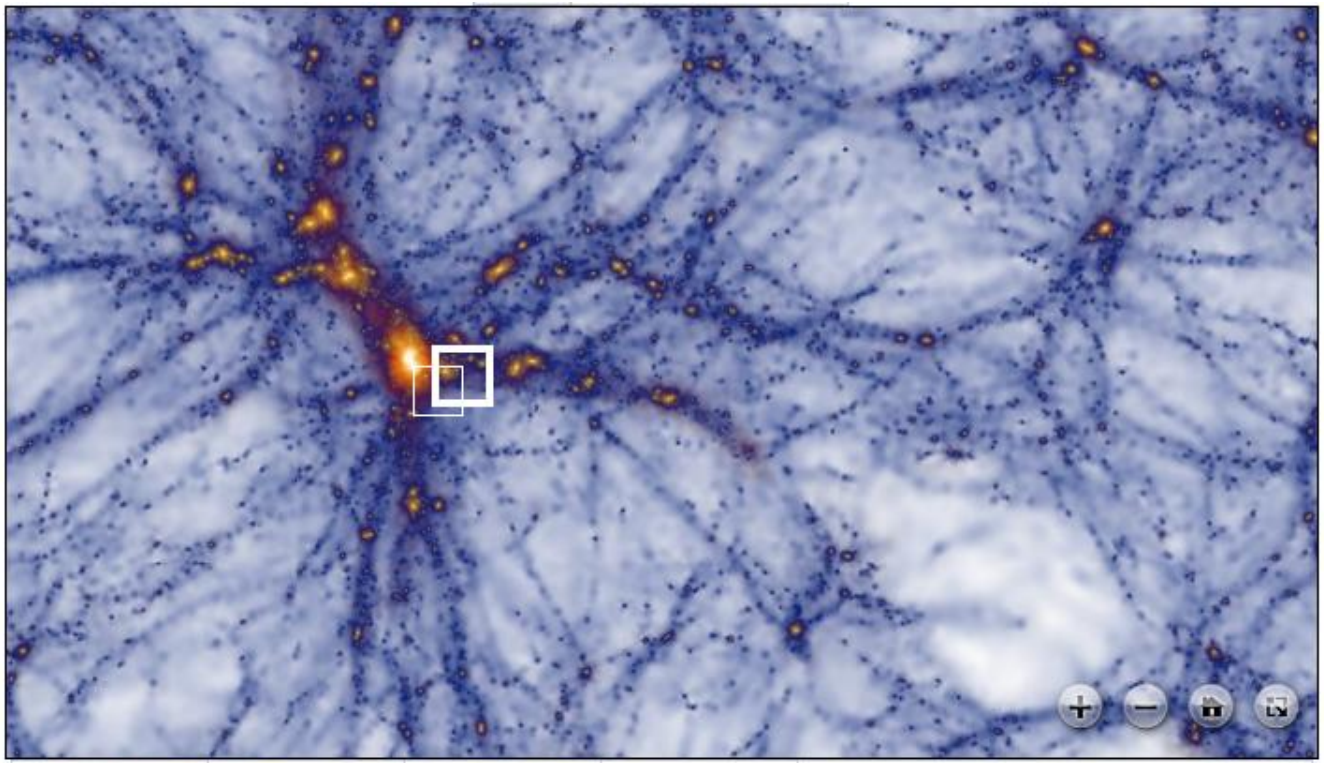


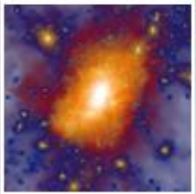
Fast Searches

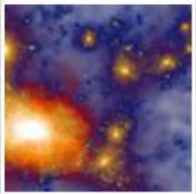
- Map space-filling curves to database indices
  - Hierarchical Triangular Mesh – on the unit sphere
  - Peano-Hilbert / Morton curves – in 3D
- Combine with query shapes
  - Build from primitives

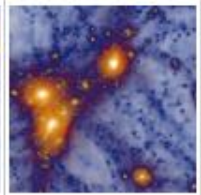


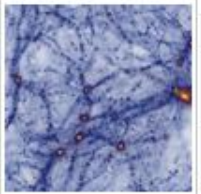
Millennium XXL







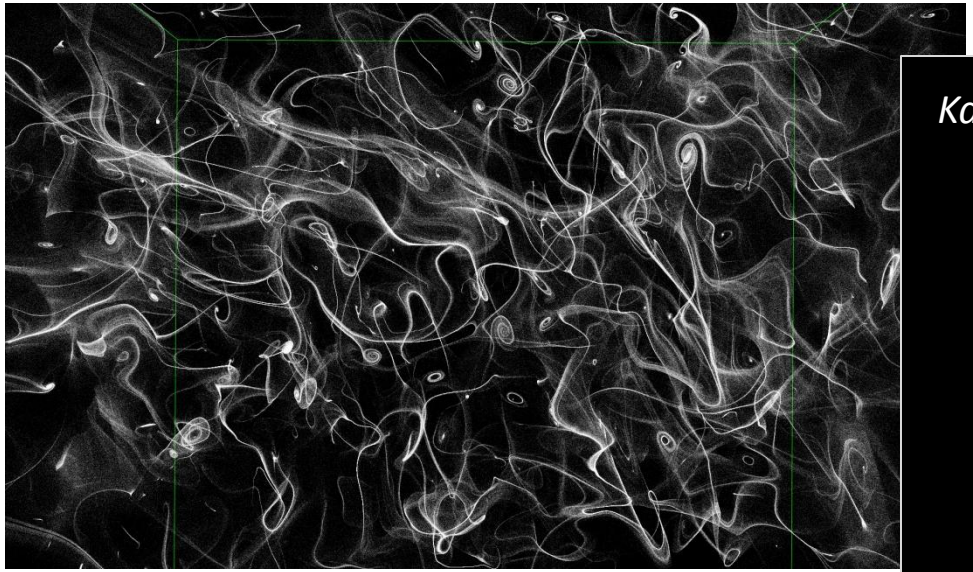


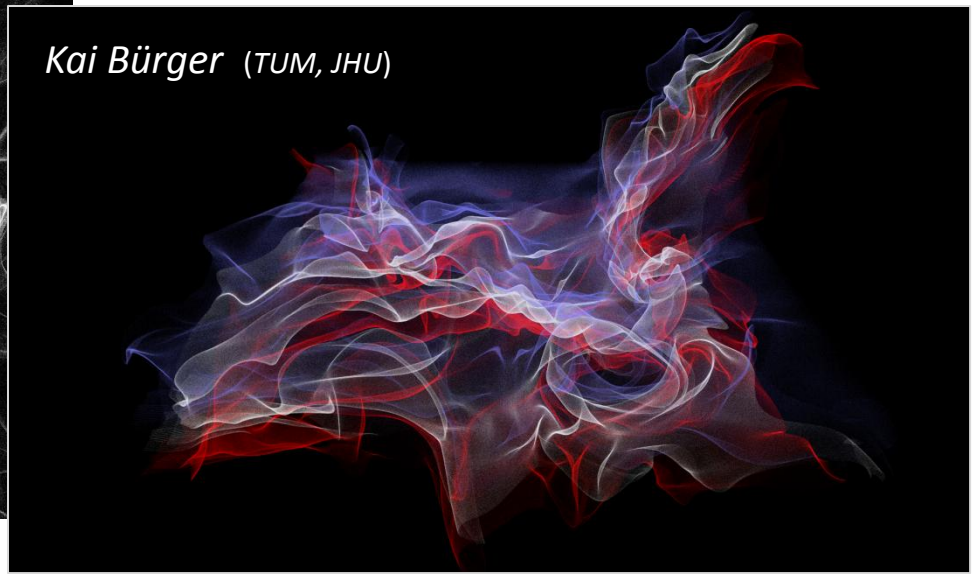


	Pixels	Points
Mouse position	(417, 279)	(1636.2304, 1361.9488)
Viewport dimensions	700 x 400	112.92 x 64.52 Mpc/h

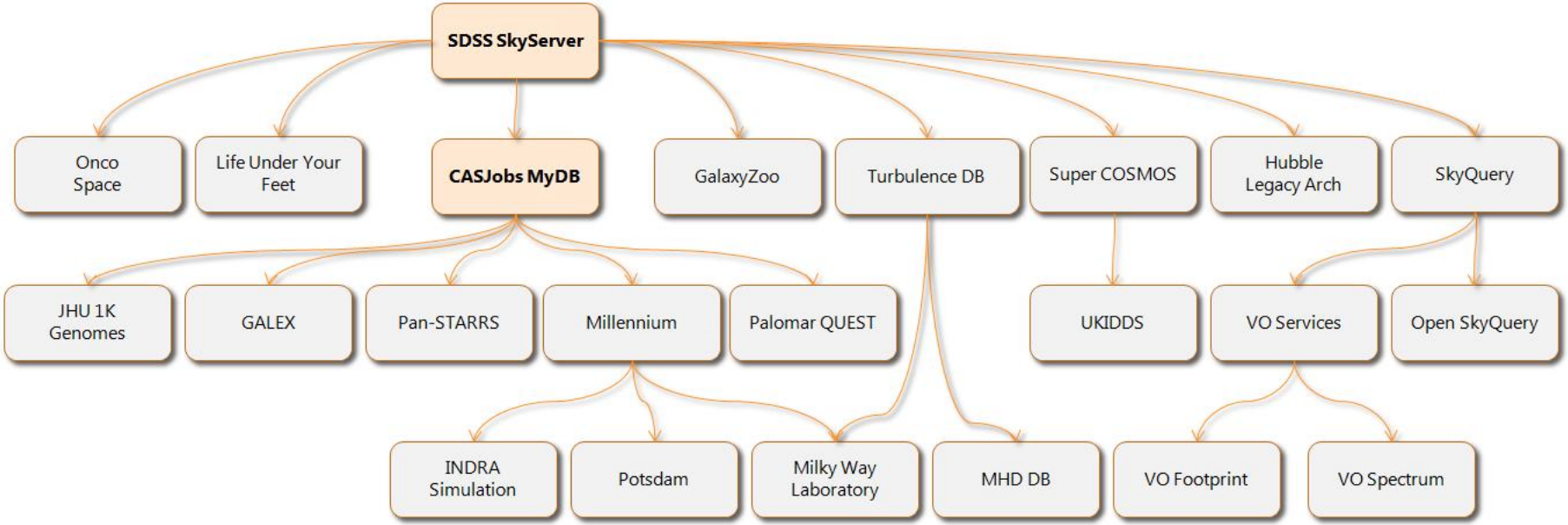
Understanding Subtleties

- Interactive visualization of 27TB of turbulence sim





The SDSS Genealogy



SkyQuery

- Dynamical federation of archives

Cross-Identification

One of the most fundamental analysis steps

What is the Right Question?

- 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?

Modeling the Astrometry

- Astrometric precision
  - A simple function

- Where on the sky?
  - Anywhere really…

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

Same or Not?

OR
- The Bayes factor

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

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

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

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

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

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 ***Astrometry***

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

Same or Not?

- The Bayes factor

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

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

Astrometry

NOT

- *K:* might be from separate objects at $\{m_i\}$

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

Normal Distribution

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

- Fisher distribution: $N(\vec{x}|w,\vec{m}) = \dfrac{w\,\delta(|\vec{x}|-1)}{4\pi \sinh w}\ \exp\left(w\,\vec{m}\vec{x}\right)$
  - Analytic results:

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

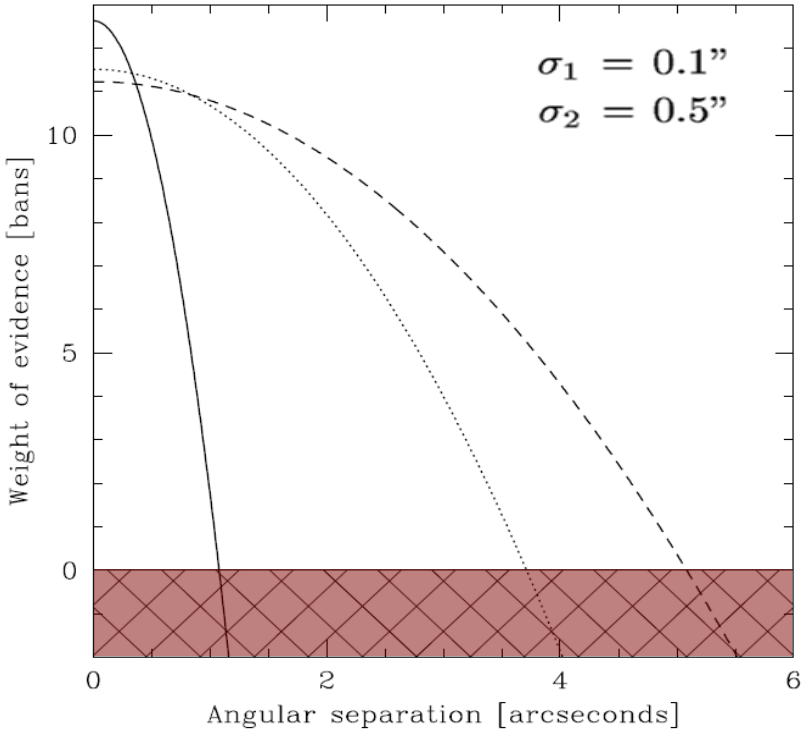
  - For high accuracies:

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

Analytic Results

- Normal distribution
  - Flat and spherical
    - Gauss and Fisher
- 2-way results

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



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

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

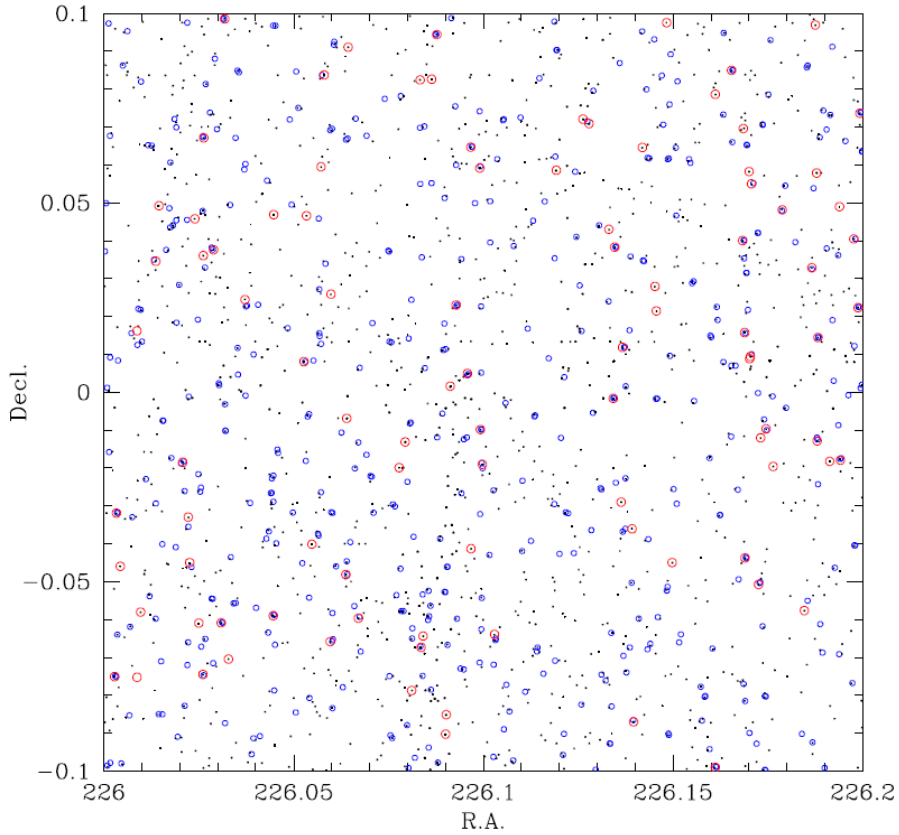
$$P(H|D) = \left[1 + \frac{1 - P(H)}{B\,P(H)}\right]^{-1}$$

- 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
  - Or solve for it:

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

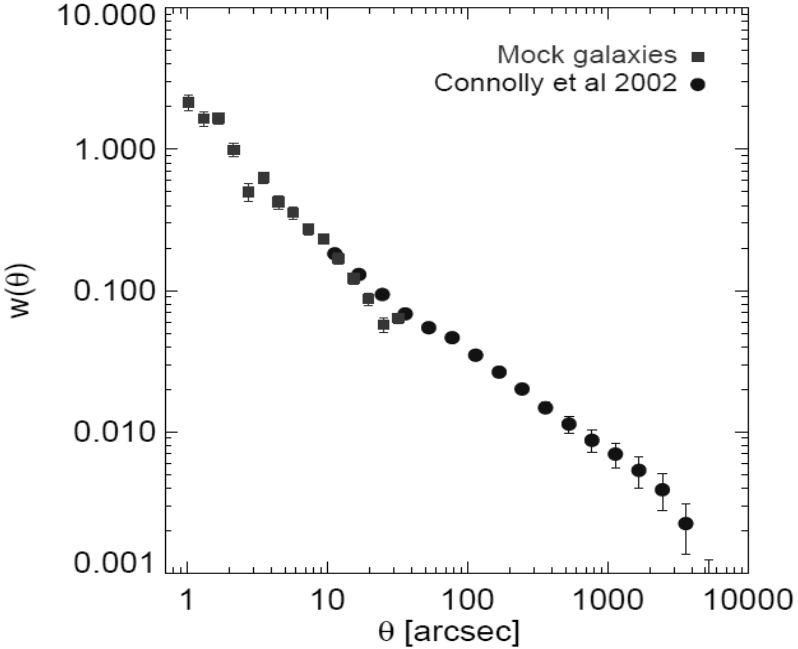
$$\left.\begin{aligned} \sum P(H) &= N_\star \\ \sum P(H|D) &= N_\star \end{aligned}\right\}$$

TB & Szalay (2008)

Simulations

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

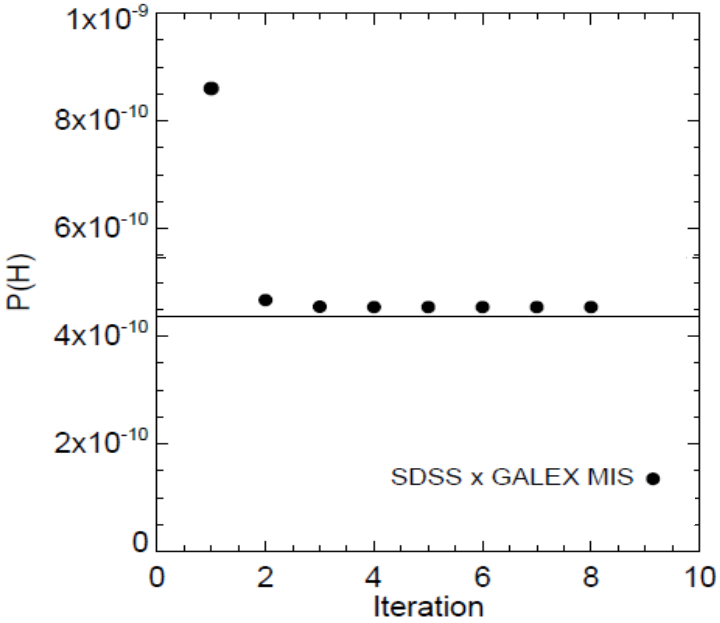
- Simulated sources
  - Subsets: N_1 N_2
  - Overlap: $N_\star$



Simulations

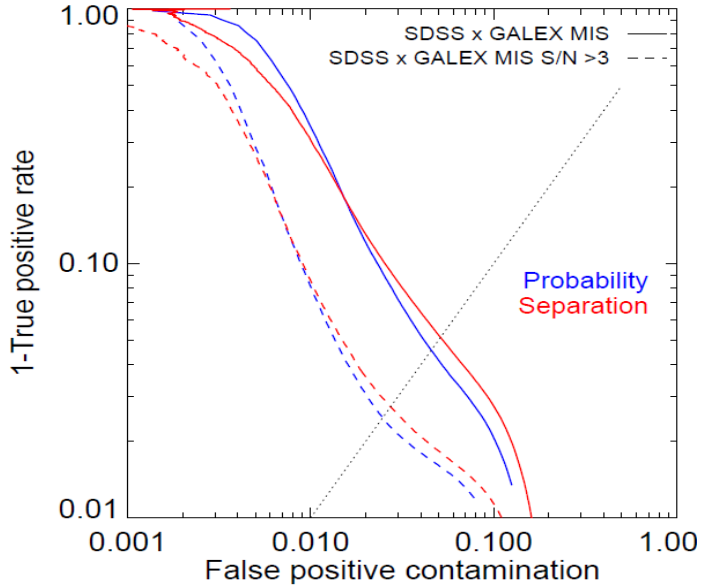
- Mock objects
  - With correct clustering
  - U_{01} values as properties

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



Simulations

- Quality



- Multiple matches

GALEX	SDSS 1	SDSS 2	SDSS 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)

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
```

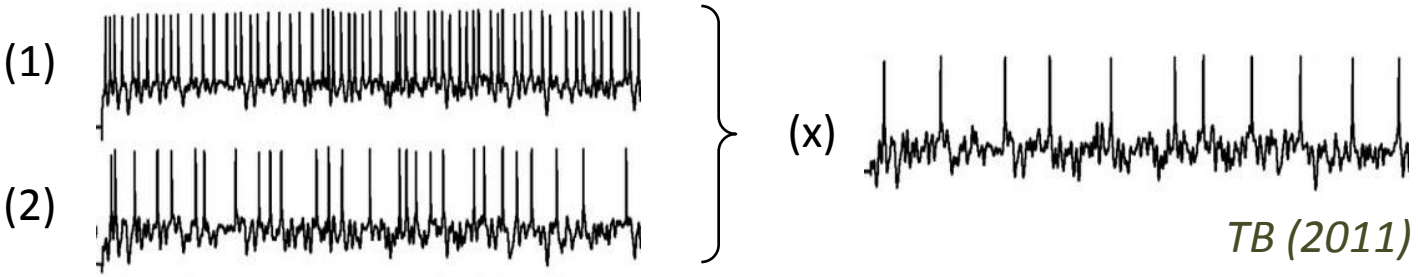
SkyQuery

- Almost pure standard SQL
- 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
    MUST s ON Point(s.RA, s.Dec), 0.1
    HAVING LIMIT 1e3
WHERE
    p.RA BETWEEN 0 AND 5
    AND p.Dec > -9999
    AND s.Dec > -9999
    AND s.RA > -9999
```

Matching Events

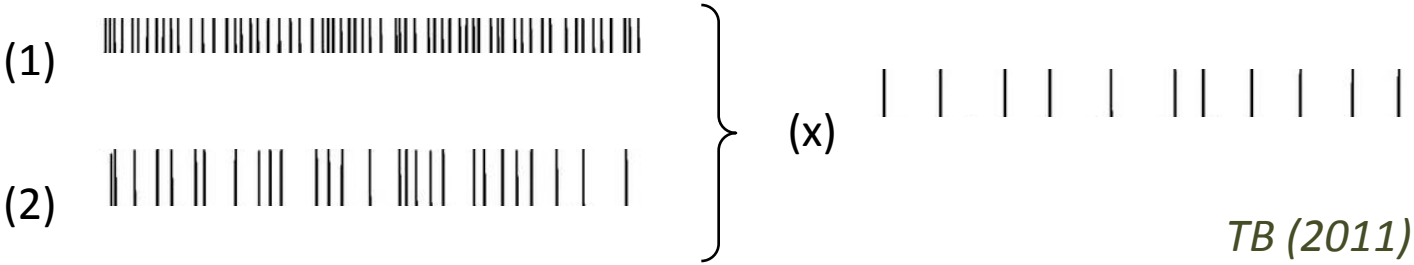
- Streams of events in time and space
  - E.g., thresholded peaks in signal-to-noise



TB (2011)

Matching Events

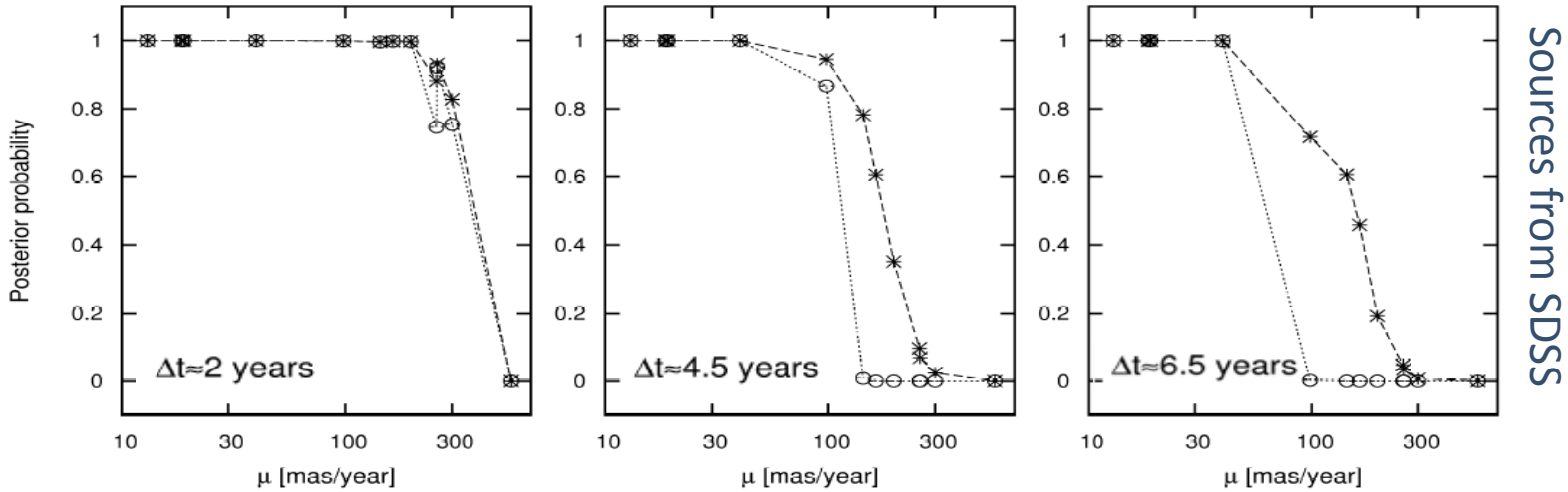
- Streams of events in time and space
  - E.g., thresholded peaks in signal-to-noise



TB (2011)

Proper Motion

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

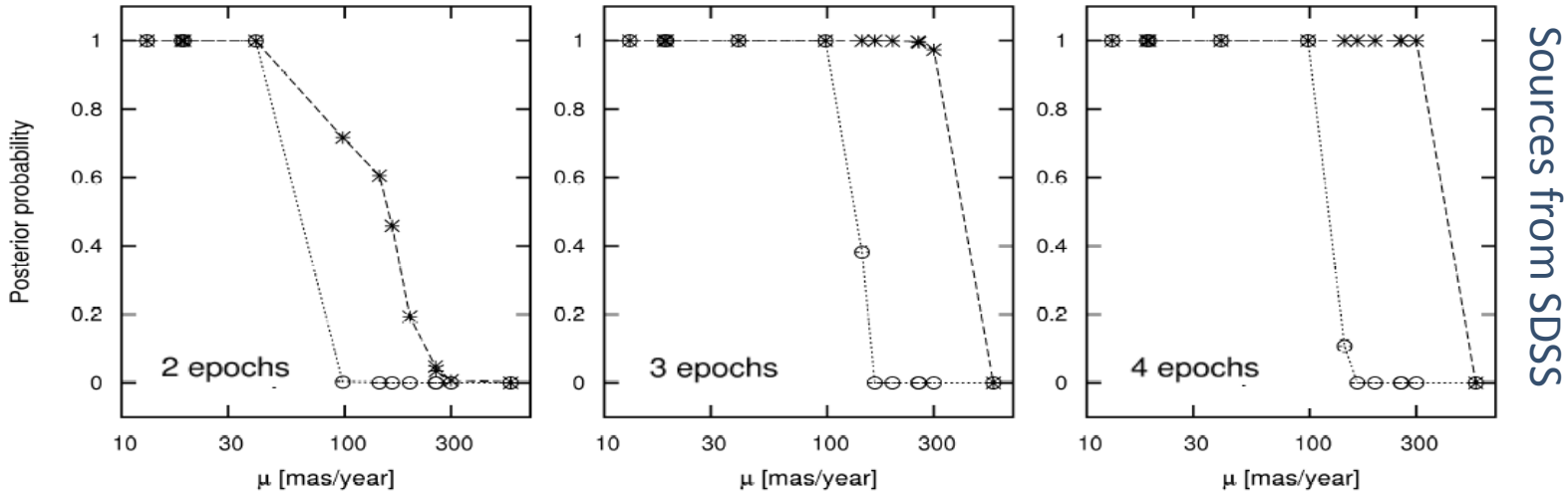


Sources from SDSS

Proper Motion

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

Kerekes, TB+ (2010)

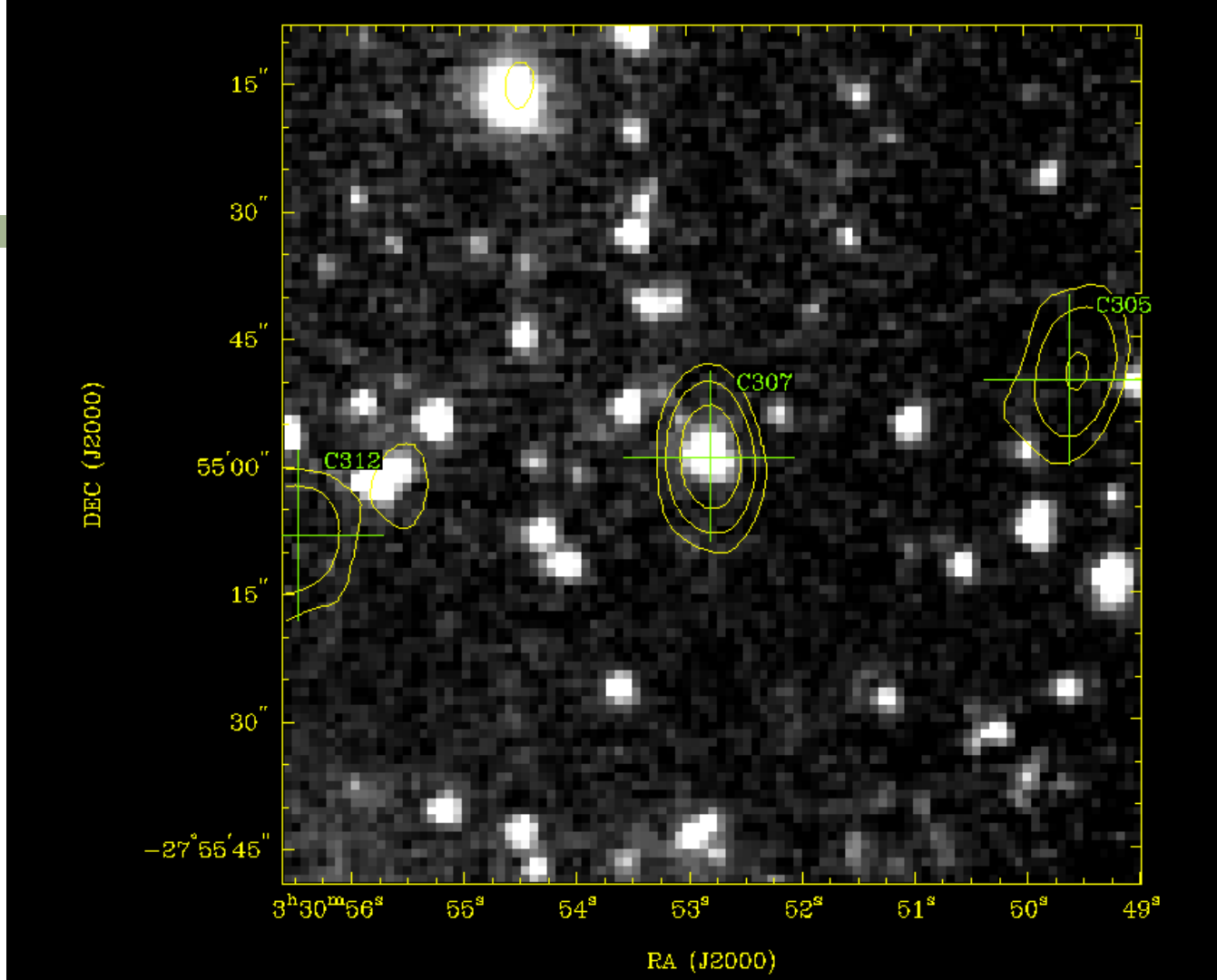


Sources from SDSS

Radio Morphology

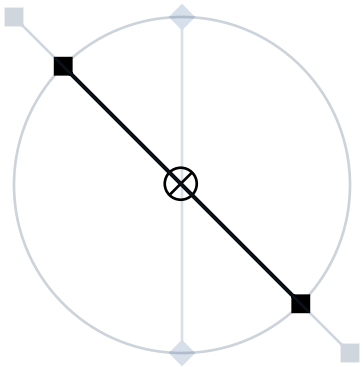
Example

- Core match
- But lobes?



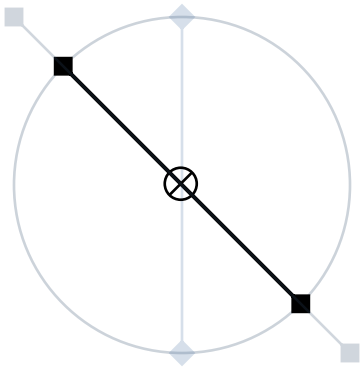
Simplest Model for Radio Sources

- A core and two symmetric lobes
  - Core direction parameter *m*
  - Lobe direction parameter *m*'
  - Other lobe is at 2*m*–*m*'



Simplest Model for Radio Sources

- A core and two symmetric lobes
  - Core direction parameter m
  - Lobe direction parameter m'
  - Other lobe is at $2m-m'$
- The prior is some $p(m,m') = p(m)\, p(m'|m)$
  - Parameter m anywhere on the sky: $p(m)$
  - But m' is seen certain separations away: $p(m'|m)$



Likelihood of "Same"

- Data: { x_0; y_0 y_1 y_2 } directions
- 2-way matching using the new model

$$\mathcal{L} = \int dm\; p(m)\, L_{x_0}(m)\, L_{y_0}(m) \int dm'\; p(m'|m)\, L_{y_1}(m')\, L_{y_2}(2m-m')$$

Likelihood of "Same"

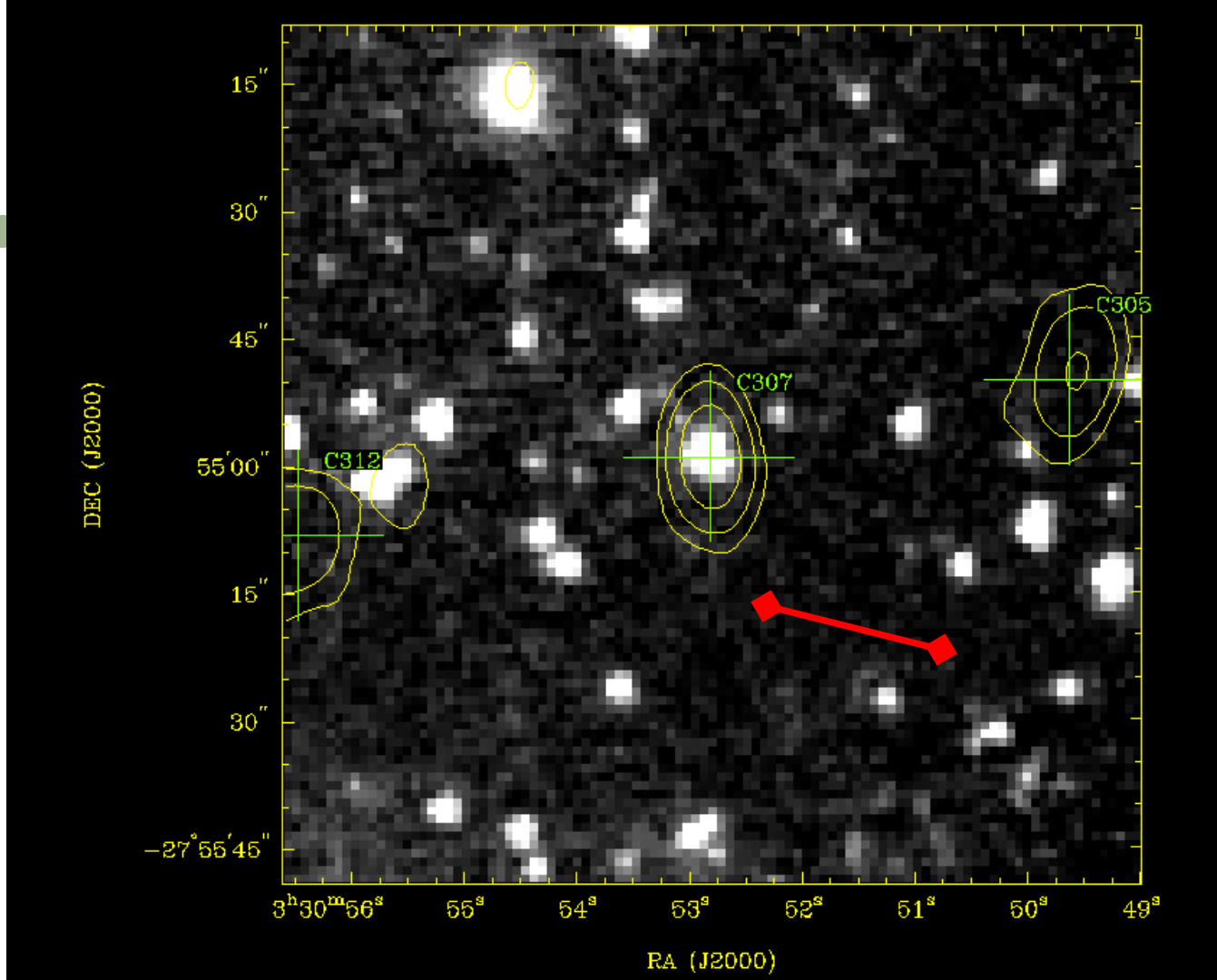
- Data: $\{ x_0; y_0\ y_1\ y_2 \}$ directions
- 2-way matching using the new model

$$\mathcal{L} = \int dm\ p(m)\, L_{x_0}(m)\, L_{y_0}(m) \int dm'\, p(m'|m)\, L_{y_1}(m')\, L_{y_2}(2m-m')$$

$$p(m'|m) = \begin{cases} \left[(R^2 - r^2)\pi\right]^{-1} & \text{if } \ r < |m'-m| < R \\ 0 & \text{otherwise} \end{cases}$$

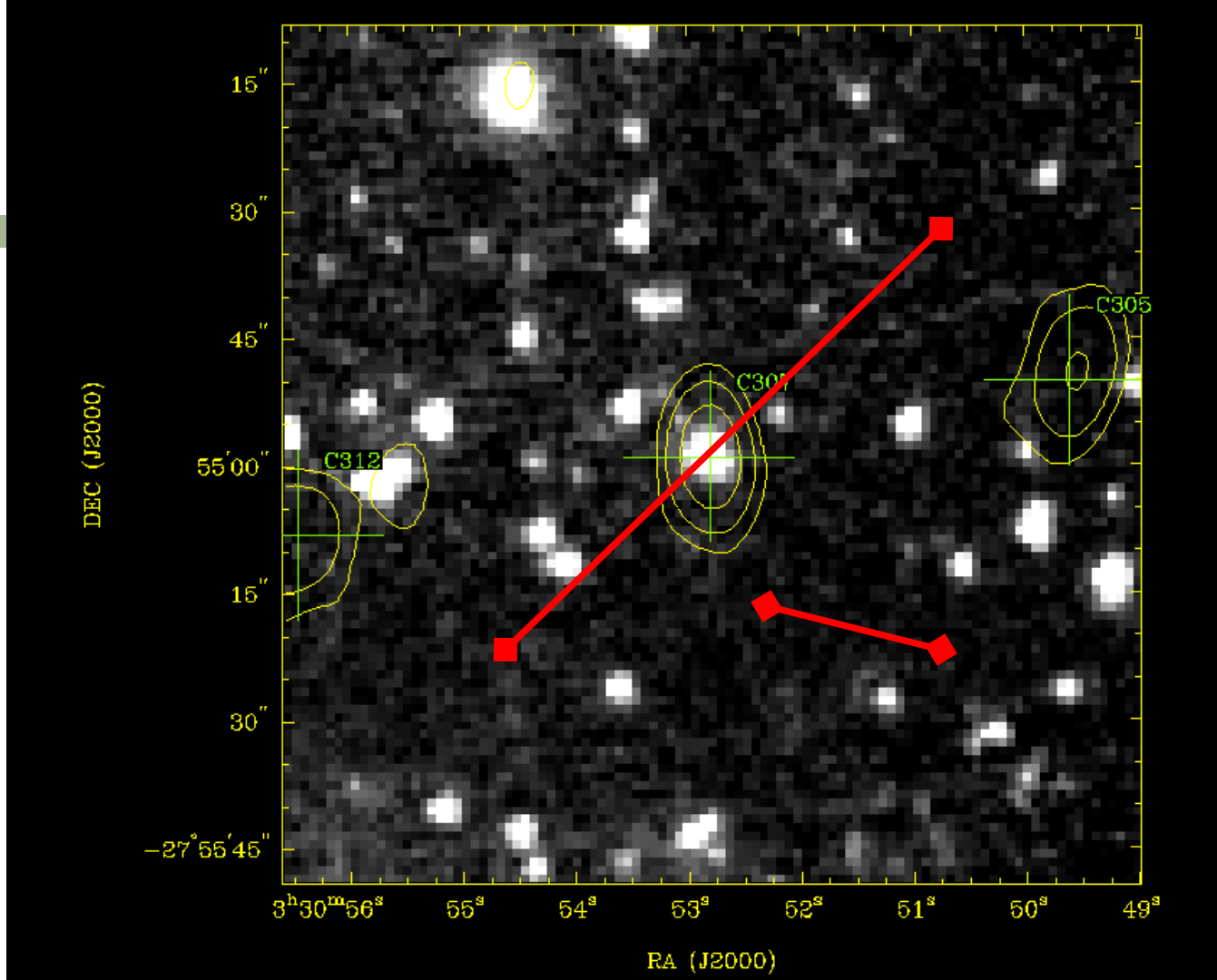
Example

- Core match
- But lobes?



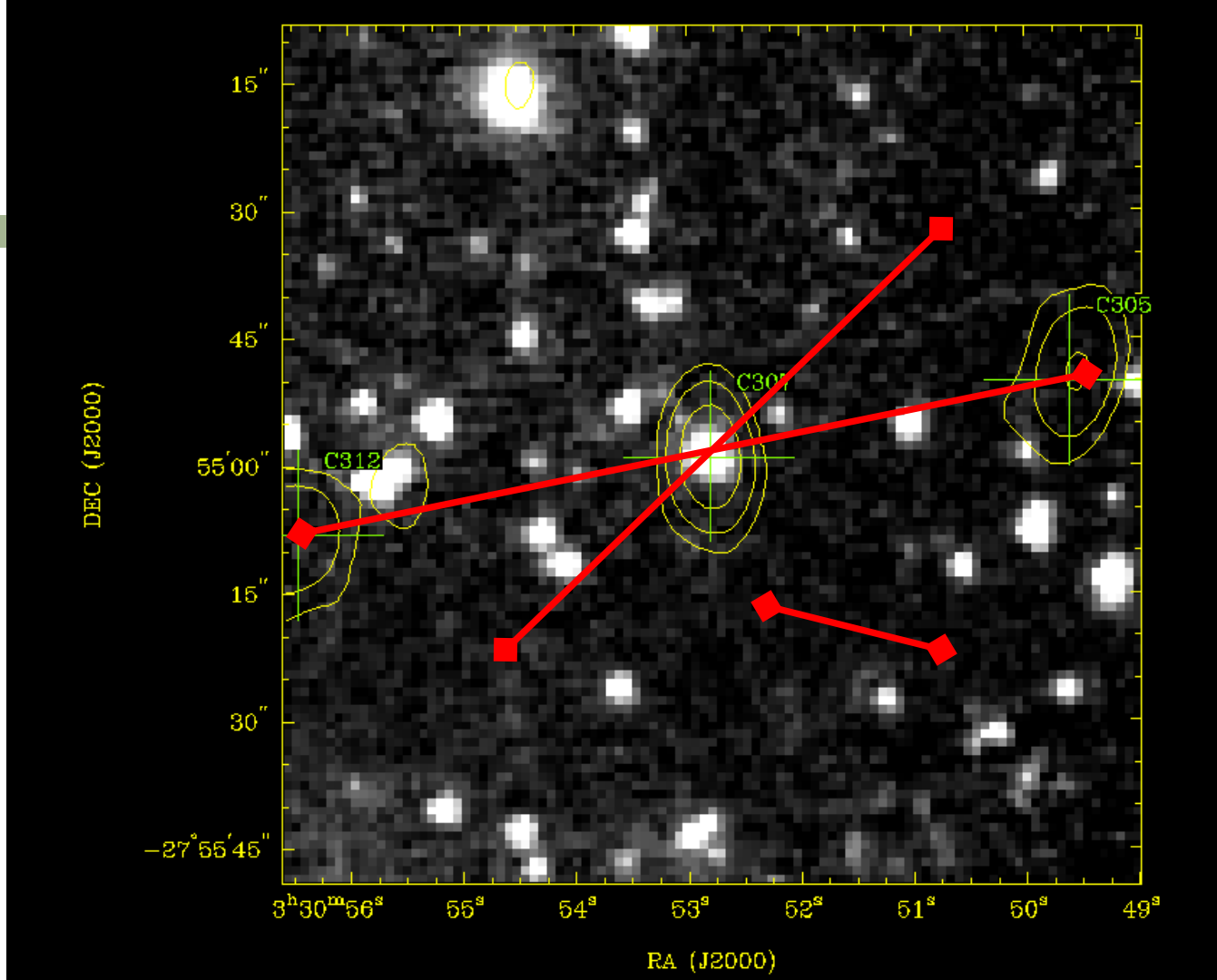
Example

- Core match
- But lobes?



Example

- Core match
- But lobes?



Competing Hypotheses

- Many to choose from
  - Which radio detection is the core? (closest ;-)
  - Which are separate objects?

Example

- Competing hypotheses w/ 2 different tophat priors

None None None :	0.0	0.0 (F)
Core None None :	11.1	11.1 (F)
Core Lobe None :	19.0	18.3 (F)
Core Lobe Lobe :	24.2	23.5 (F)
None Lobe Lobe :	13.2	12.5 (F)
None None Lobe :	7.9	7.2 (F)

WORK IN PROGRESS!

Hubble Source Catalog

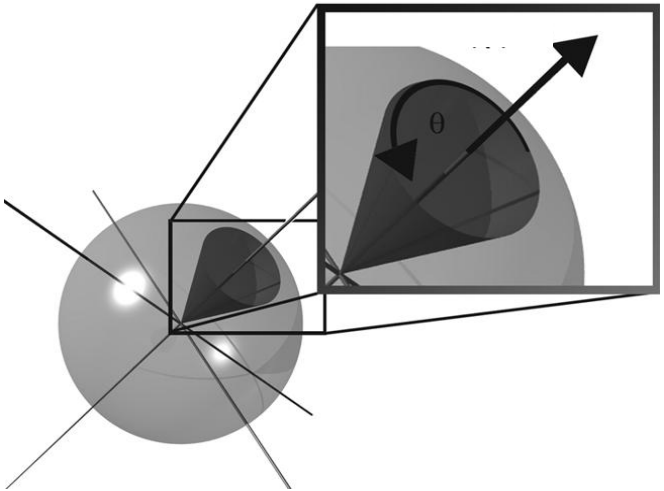
Crossmatching the Hubble Legacy Archive

Uncertain Positioning

- Exceptional accuracy with images
- Large uncertainty in positioning
  - Due to few standards in the tiny field of view
- Solve for relative corrections
  - Across overlapping observations

Astrometric Correction

- Rotation in 3D
  - Describes both translation and rotation locally
  - Need constrained numerical optimization: $R^T R=I$



Infinitesimal Rotation

- Optimization without constraints
  - Displacement by the cross product operator

$$\min_{\boldsymbol{\omega}} \left\{ \frac{1}{2} \sum_{\beta} \frac{1}{\sigma_\beta^2} \left[\boldsymbol{c}^{(\beta)} - \left(\boldsymbol{r}^{(\beta)} + \boldsymbol{\omega} \times \boldsymbol{r}^{(\beta)} \right) \right]^2 \right\}$$

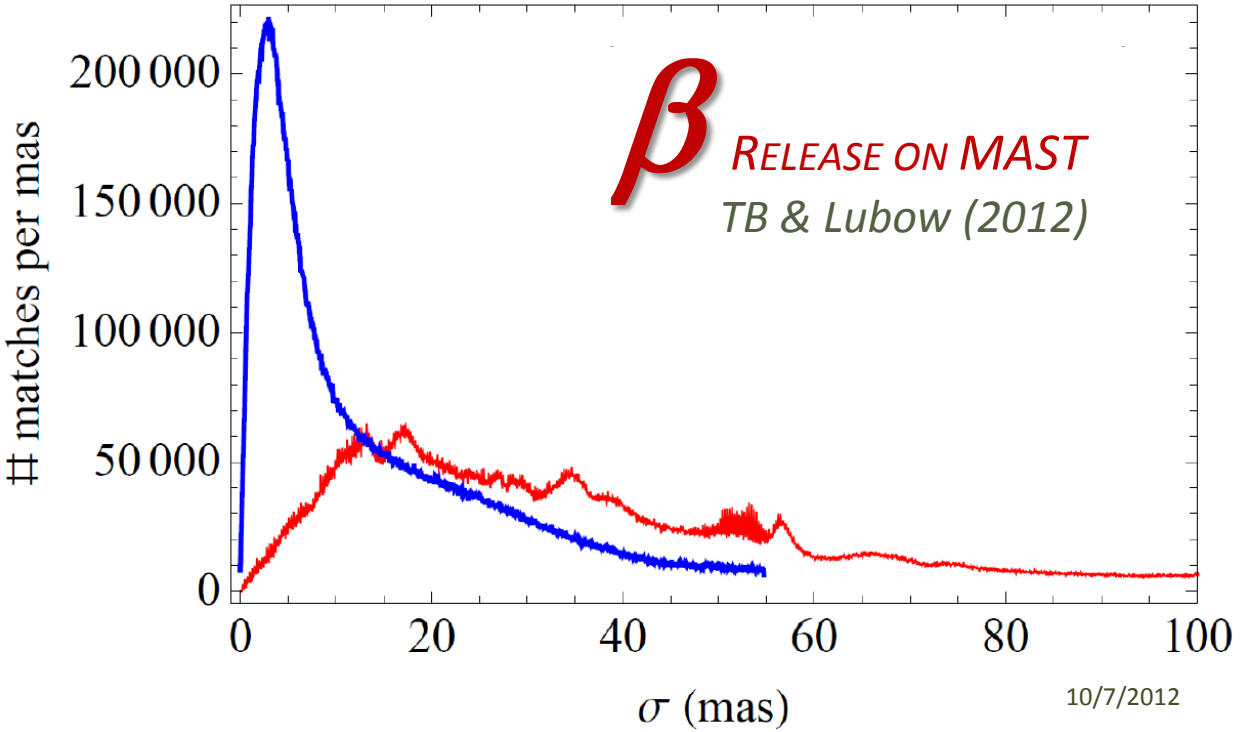
  - Fast computation for the rotation vector ω
    - Just sum up a 3×3 matrix and vector
    - Solve the linear equation: $\mathbf{A}\tilde{\boldsymbol{\omega}} = \boldsymbol{b}$

$$\begin{cases} \mathbf{A} &= \sum_i w_i \left(\mathbf{I} - \boldsymbol{r}_i \otimes \boldsymbol{r}_i \right) \\ \boldsymbol{b} &= \sum_i w_i \left(\boldsymbol{r}_i \times \boldsymbol{c}_i \right) \end{cases}$$

Hubble Source Catalog

- SQL pipeline
- Astrometric correction
  - Subpixel precision

Available now!



Summary

- Bayesian approach to cross-identification
  - Places former heuristics on a firm statistical basis
- Enables us to properly include
  - Physics, geometry, etc…
- Naturally extends to time-domain
  - Events, proper motion, lightcurves
- Opens the door for next-generation methods