COMPUTATIONAL STATISTICS IN ASTRONOMY: NOW AND SOON

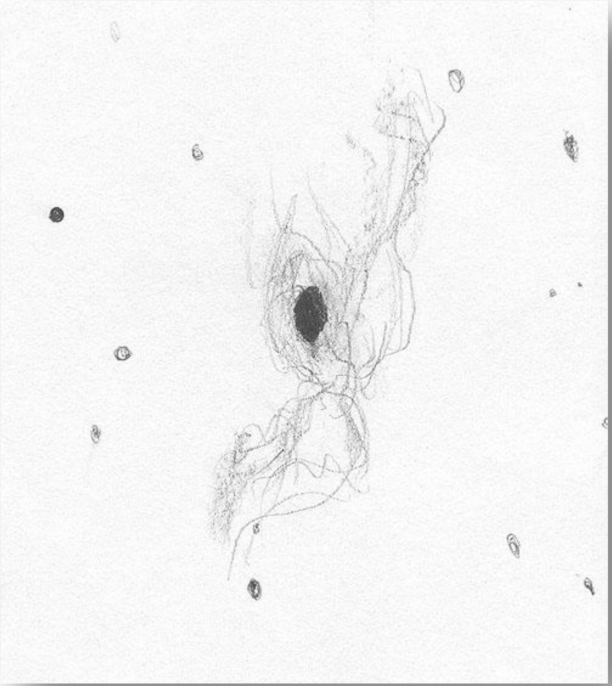
9/9/2012

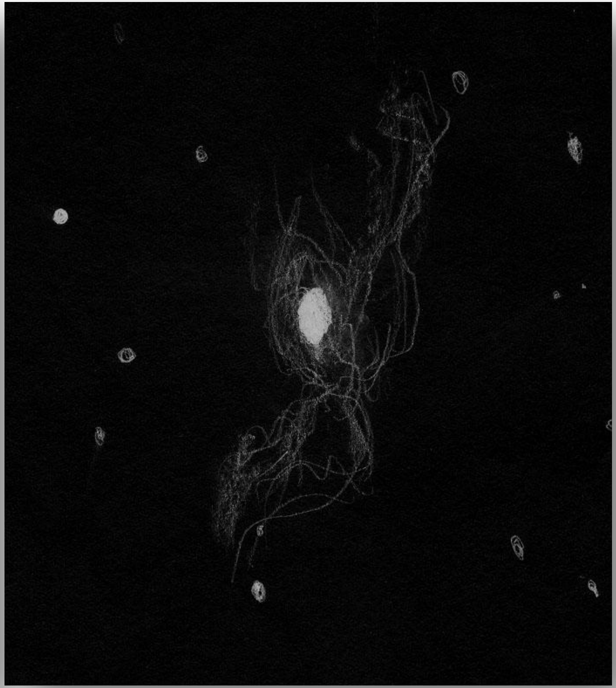
Tamás Budavári / The Johns Hopkins University

Recording Observations

- Astronomers drew it…
- Now kids do it on the SkyServer

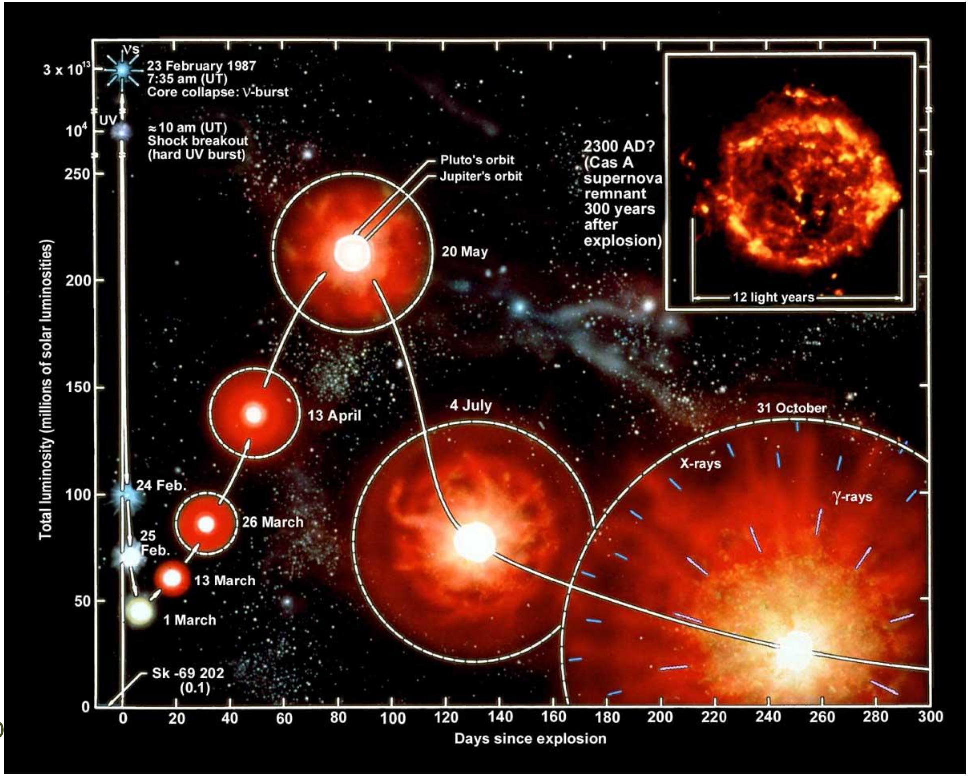
#1 by Haley ⇒

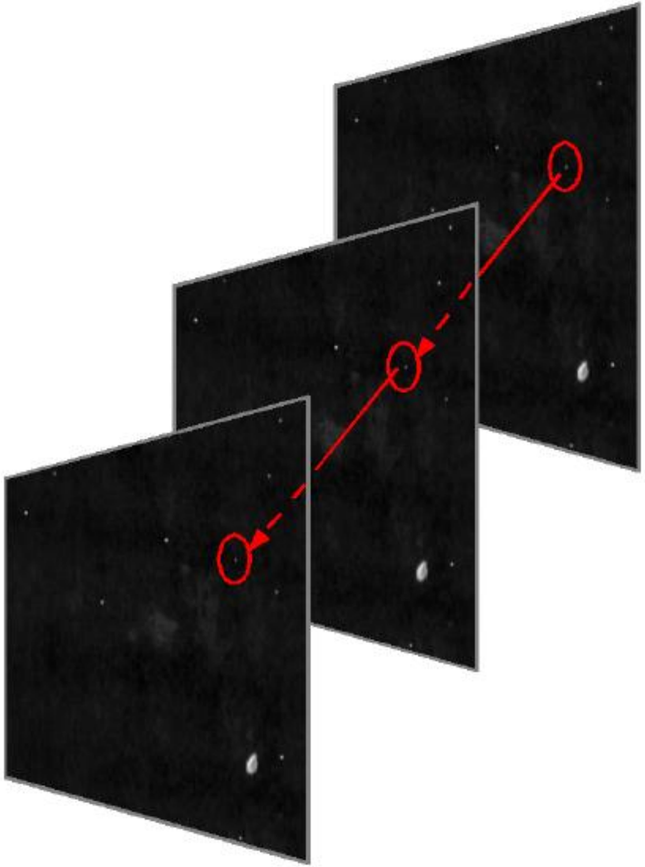




Multicolor Universe

Eventful Universe

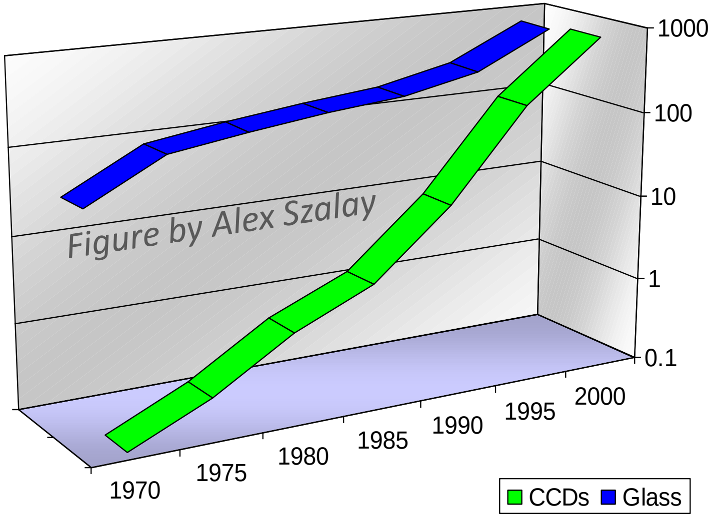




Trends in Astronomy

Tamás Budavári

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



Sloan Digital Sky Survey

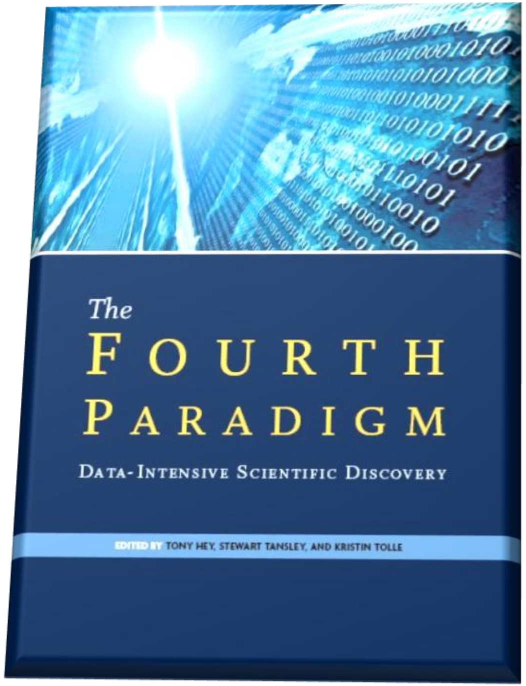
- Cosmic Genome Project 2001-2010
  - 500M rows, 400+ cols ~ 18TB
  - 30TB of images from 30 CCDs
  - Software revolution in astro
  - Astronomers learn SQL
    - Cannot look at the data anymore

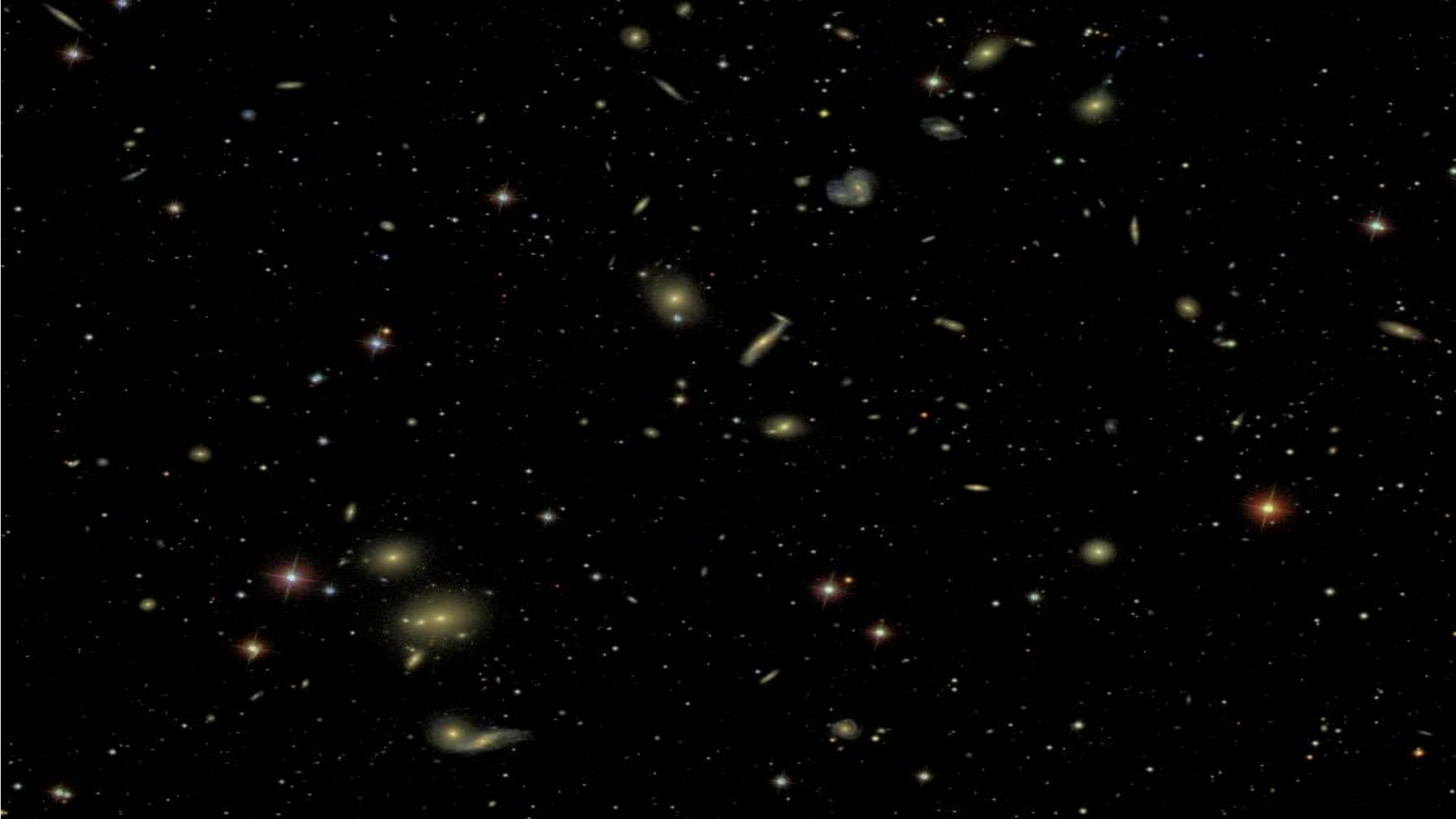
The New Generations

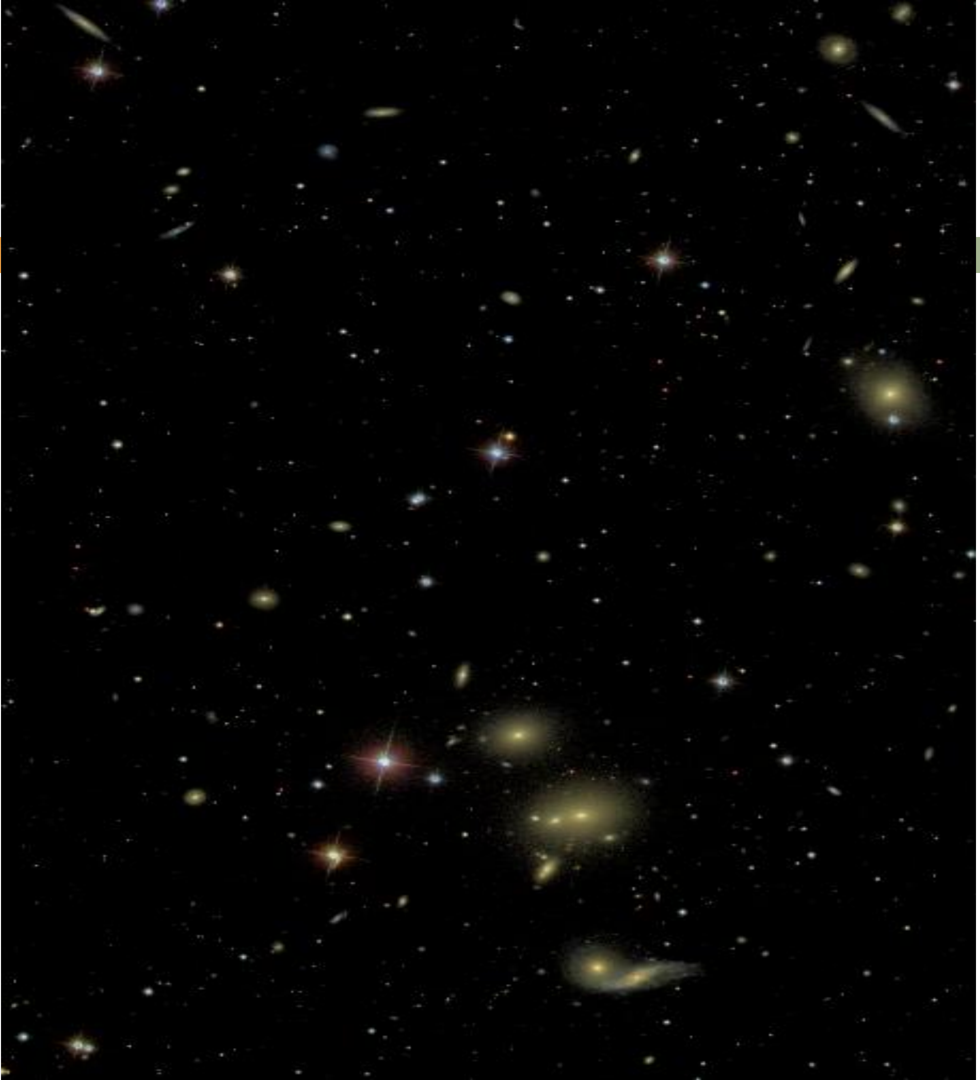
- Large Synoptic Survey Telescope [OPTICAL]
  - 3 trillion rows, 200+ attributes, 100+ tables ~ 30PB
  - 60PB of images, 3.2 Gpix cam
- Square Kilometer Array [RADIO]
  - Processing limited

Data-Intensive Science

- The Fourth Paradigm – Jim Gray
  - Phenomenology
  - Calculus
  - Simulations
  - eScience







Keeping Up?

- Image processing
- Catalog extraction
  - $O(n)$

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

More is Different

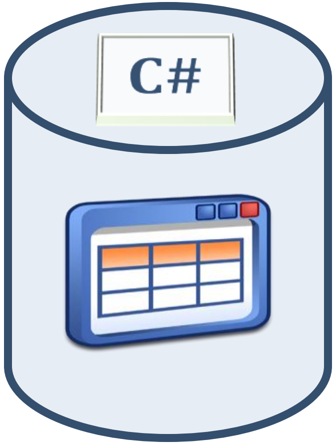
- Lots of opportunities
- Lots of challenges
- Lots of problems

- 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

Home | Tools | Schema | Projects | Astronomy | SDSS | Contact Us | Download | Site Search | Help

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

SDSS Query / CasJobs

Help Tools Query History MyDB Import Groups Output Profile Queues SkyServer Logout budavari

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

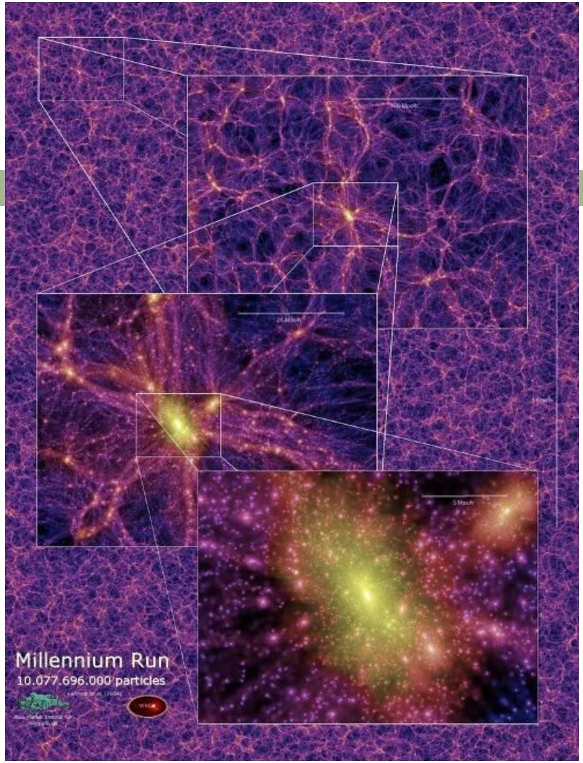
Sizes are approximations only.
Row counts are approximations only. For exact value run a count.
There's always some overhead, even empty MyDB's take up space.
Group tables do not count towards your MyDB size limit.

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

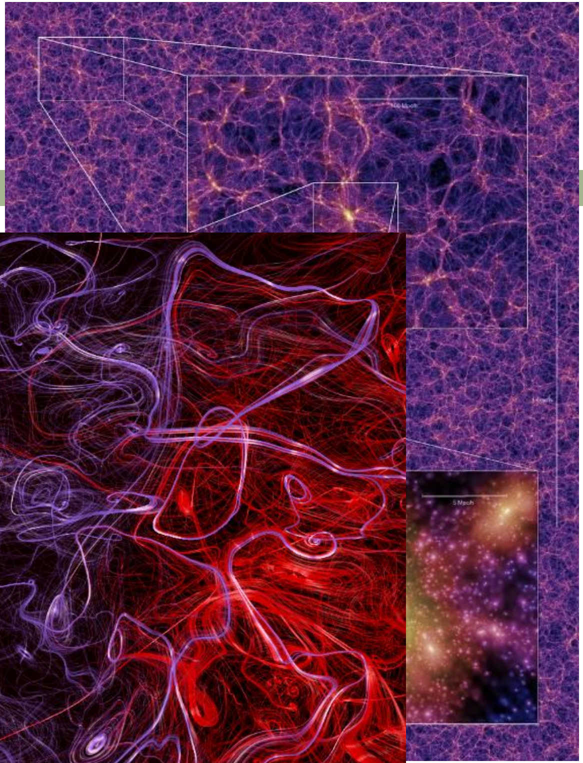
- Turbulence simulations (JHU)
  - 1024^4 grid, 27TB

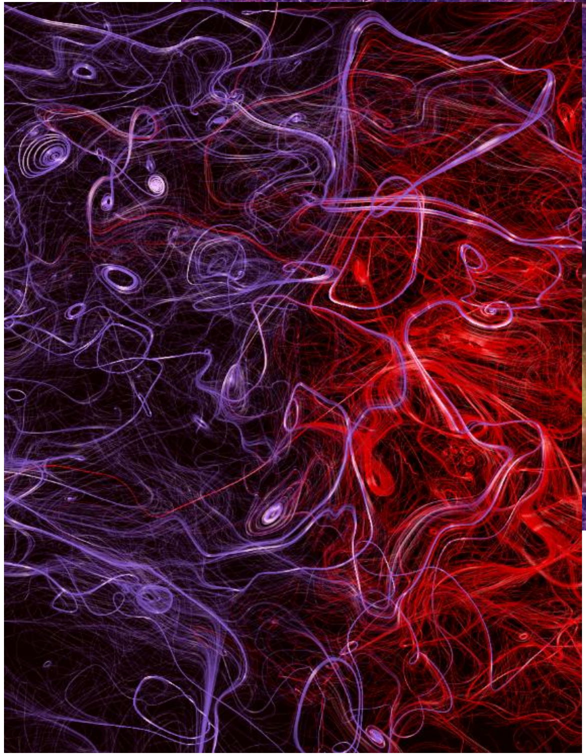


Storing Simulations

- 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

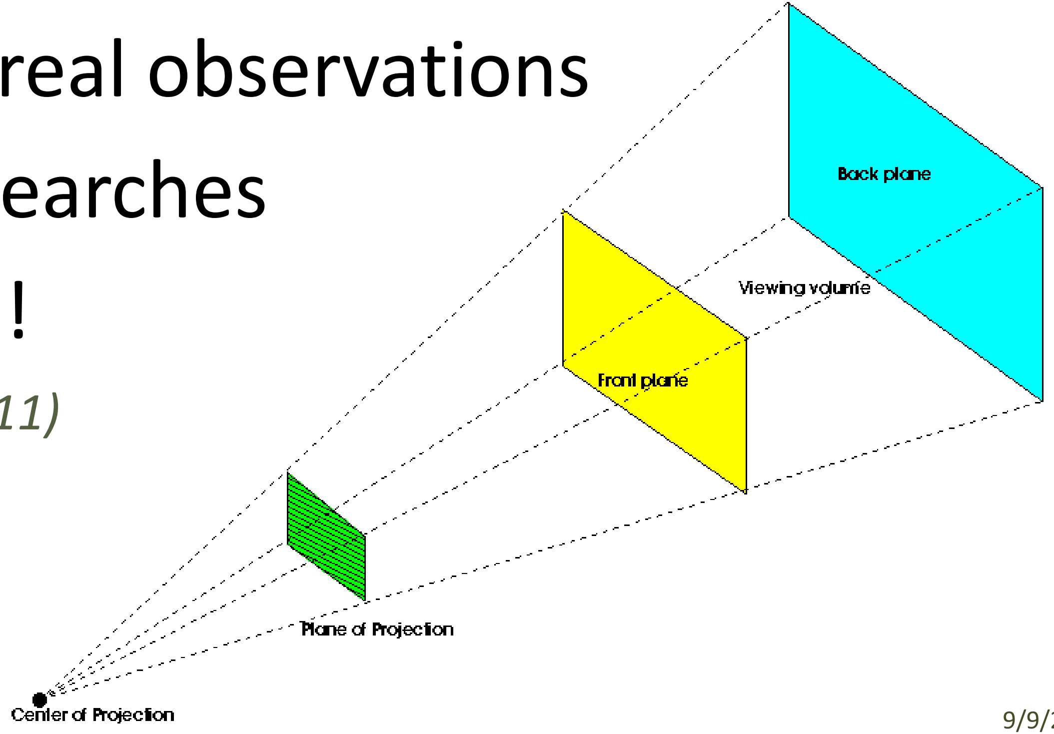




Observing Simulations

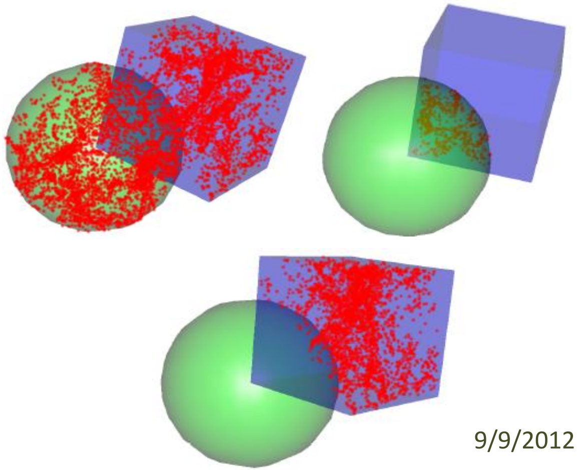
- Comparison to real observations
- Lots of spatial searches
- In the database!

Lemson, TB & Szalay (2011)



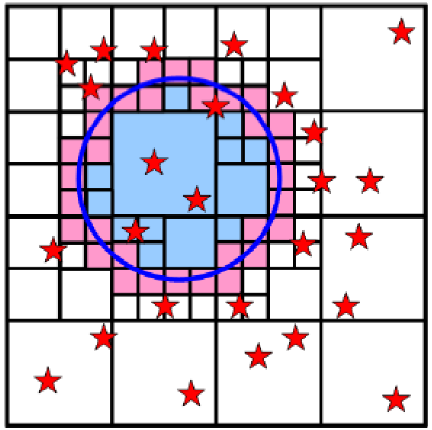
Fast Searches

- 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

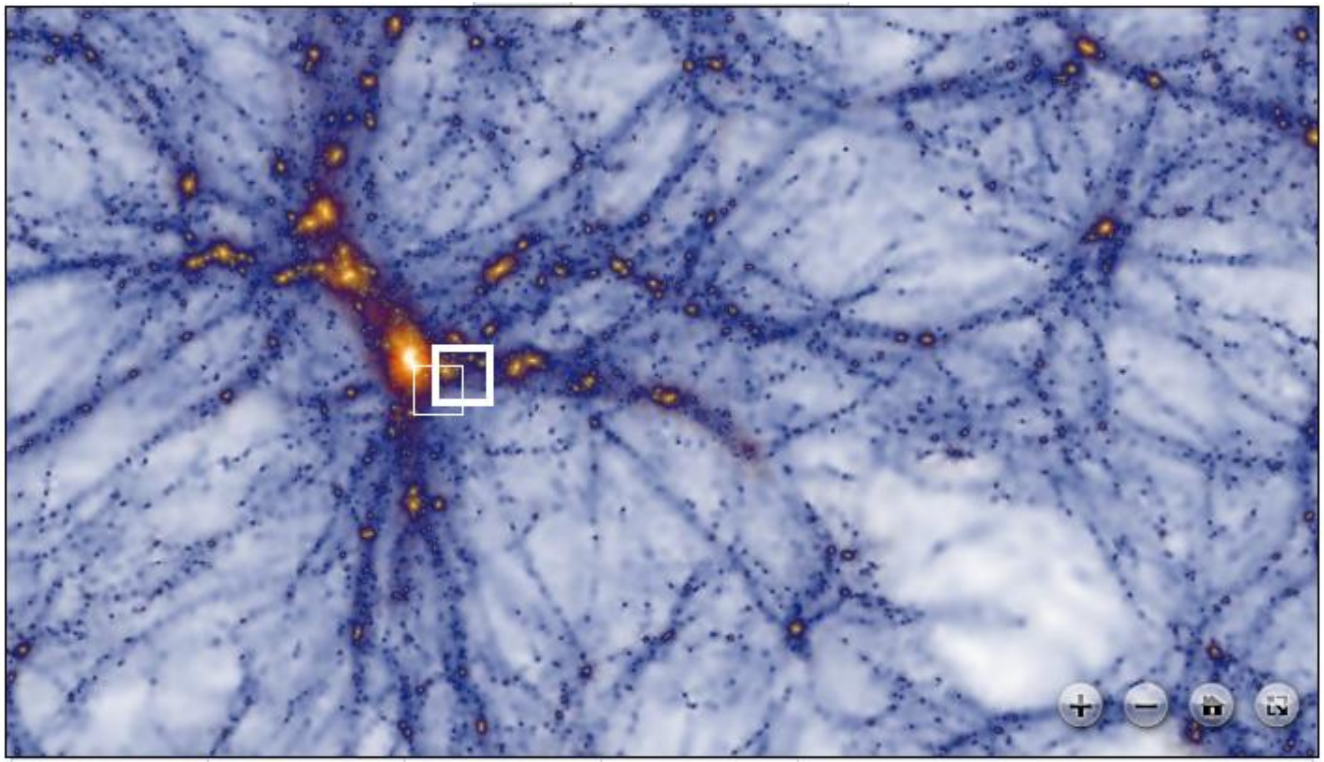


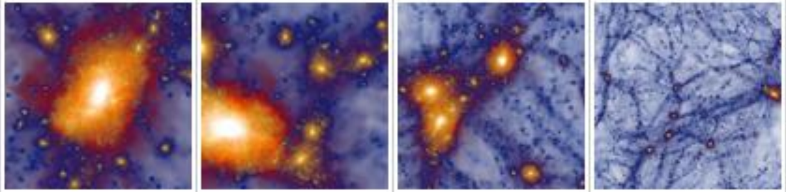
Fast Searches

- 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

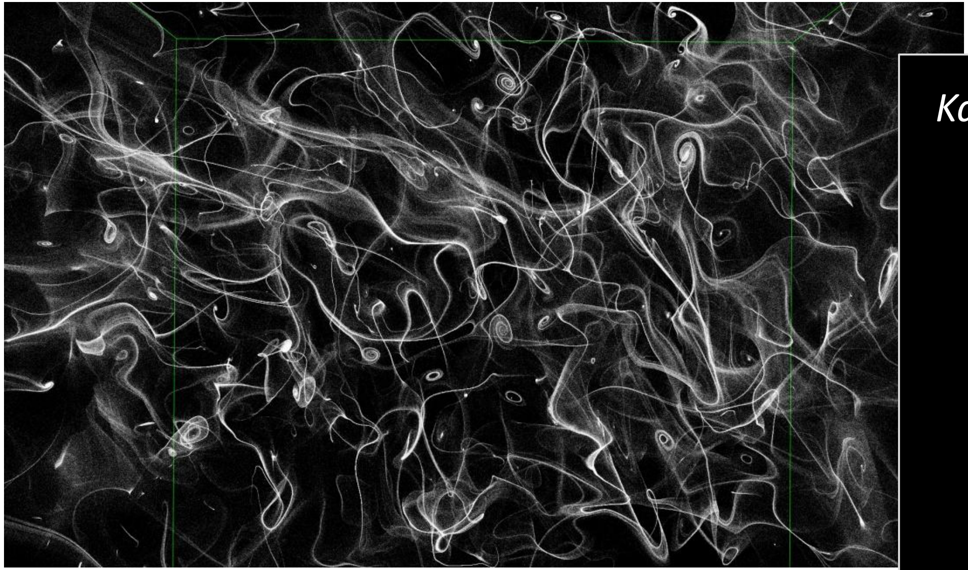


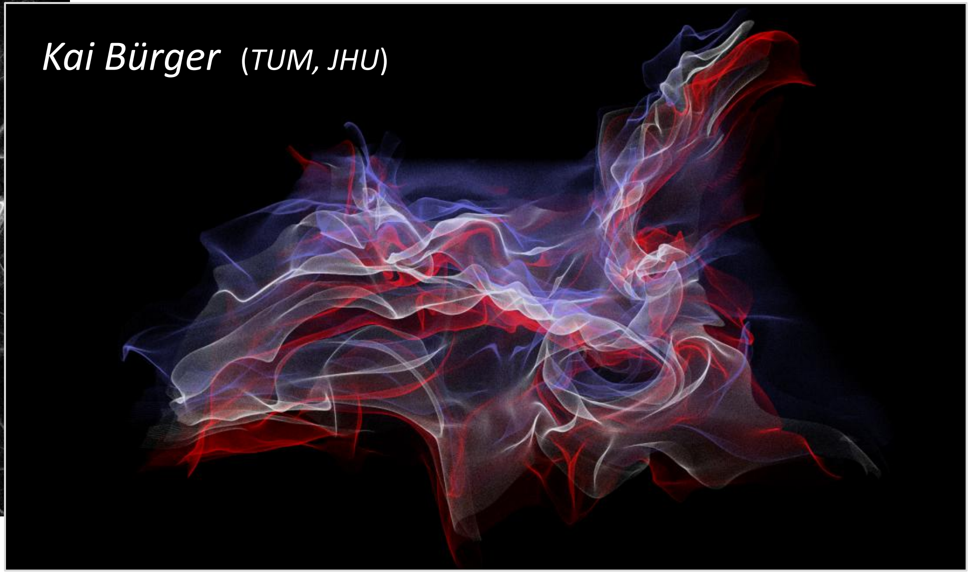


	Pixels	Points
Mouse position	(417, 279)	(1636.2304, 1361.9488)
Viewport dimensions	700 x 400	112.92 x 64.52 Mpc/h

Understanding Subtleties

- Interactive visualization of 27TB of turbulence sim





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

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

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?

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 $\{m_i\}$

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

Normal Distribution

Tamás Budavári

- 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:

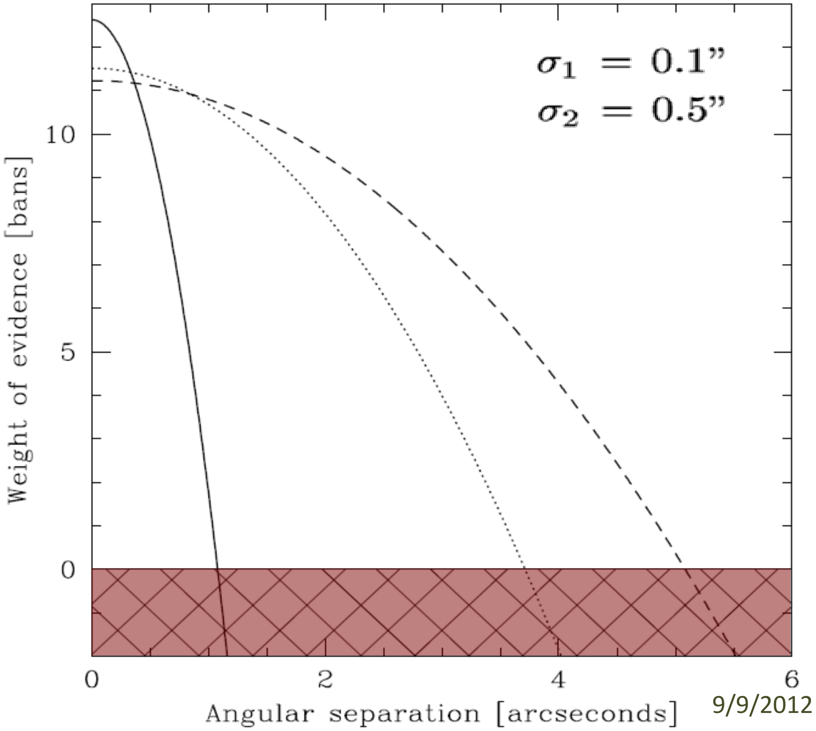
$$= 2^{n-1}\frac{\prod w_i}{\sum w_i}\exp\left\{-\frac{\sum_{i<j} w_i w_j \psi_{ij}^2}{2\sum w_i}\right\}$$

Analytic Results

- Normal distribution
  - Flat and spherical
    - Gauss and Fisher

- 2-way results

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



From Priors to Posteriors

Tamás Budavári

- Posterior probability from prior & Bayes factor

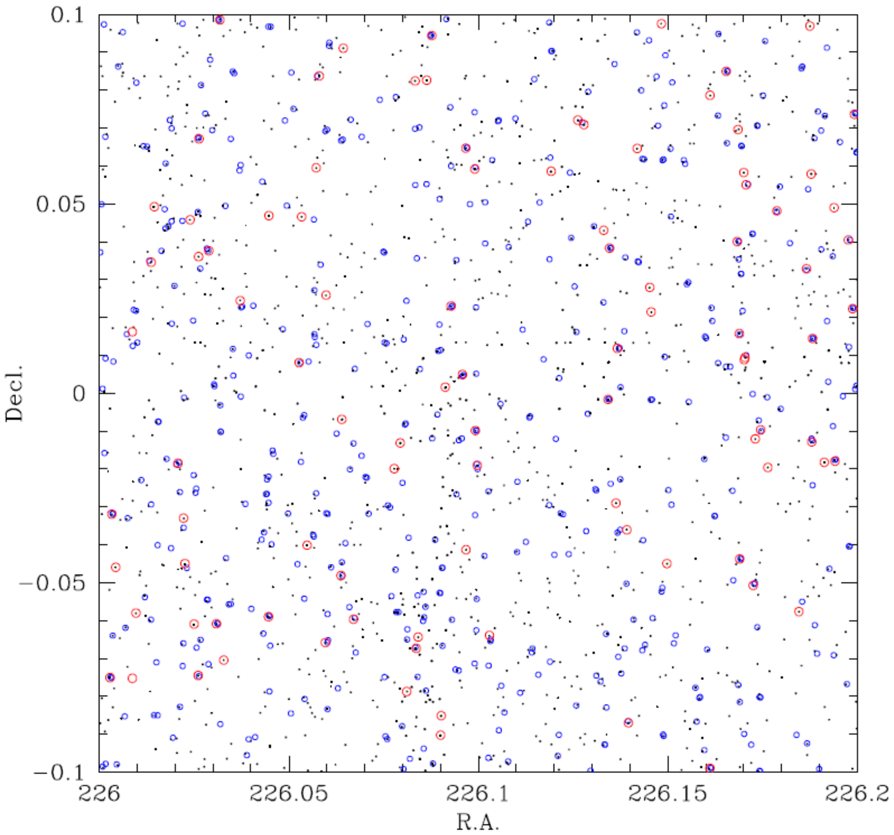
$$P(H|D) = \left[1 + \frac{1 - P(H)}{B\,P(H)}\right]^{-1}$$

- 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:

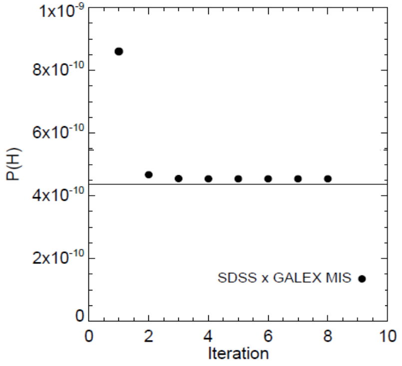
$$P(H|D) = \left[1 + \frac{1-P(H)}{B\,P(H)}\right]^{-1}$$

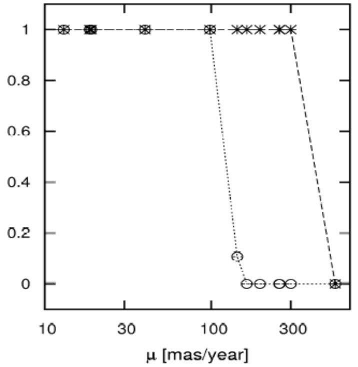
$$\left.\begin{aligned} \sum P(H) &= N_\star \\ \sum P(H|D) &= N_\star \end{aligned}\right\}$$

TB & Szalay (2008)

Works in General

- Simulations *Heinis, TB, Szalay (2009)*
  - Demo performance
- Proper motion *Kerekes, TB+ (2010)*
  - Unknown velocities
- Matching events *TB (2011)*
  - E.g., supernovae in time





SkyQuery – the new generation!

- Dynamic federation of astronomy databases
  - Query the collection as if they were one

- The 3rd generation tool coming this fall
  - Cluster of machines running partitioned jobs
  - Proper probabilistic exec with variable errors

SkyQuery

- Almost pure standard SQL

```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

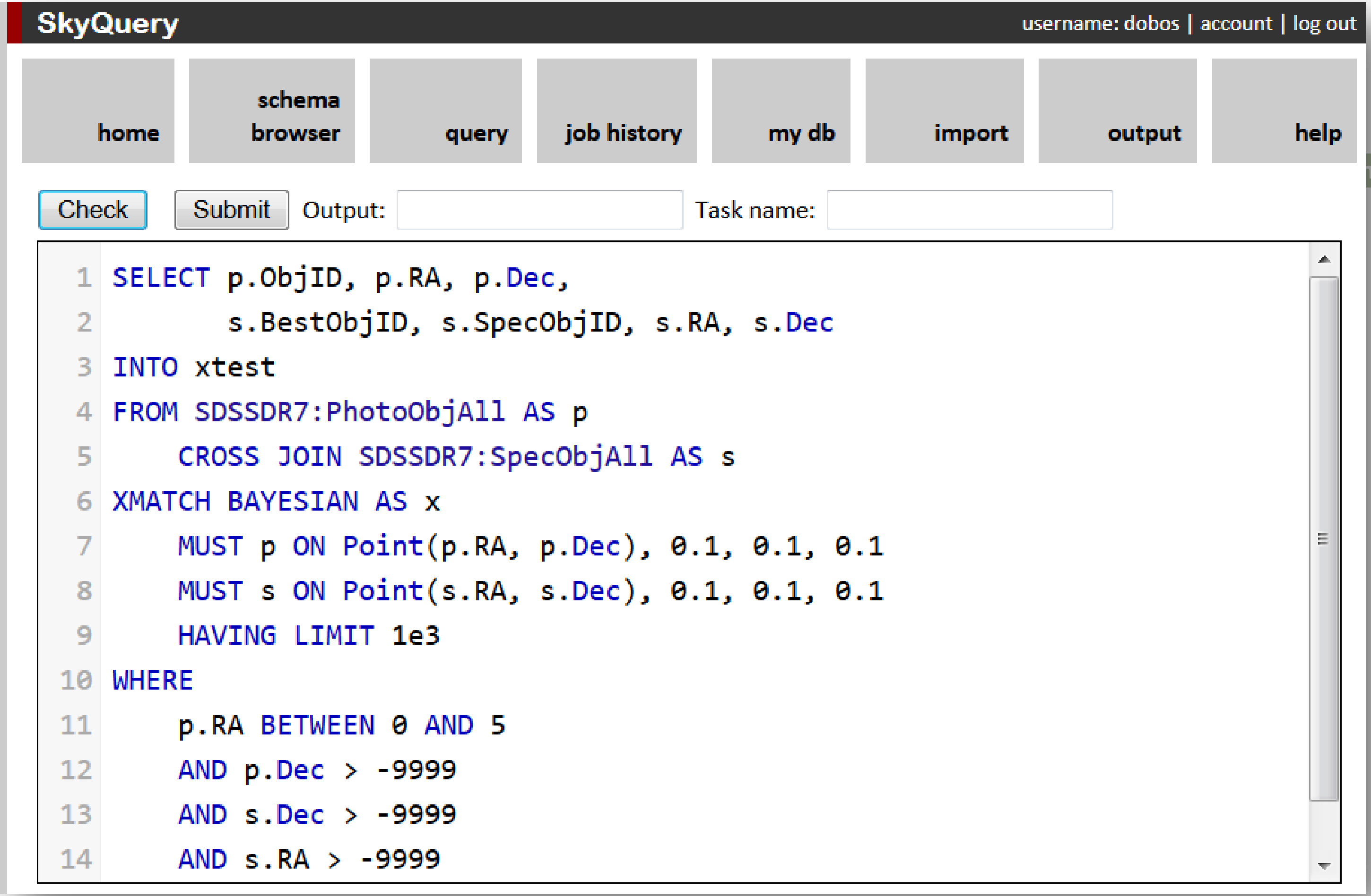
```
SELECT p.ObjID, p.RA, p.Dec,
       s.BestObjID, s.SpecObjID, s.RA, s.Dec
INTO xtest
FROM SDSSDR7:PhotoObjAll AS p
    CROSS JOIN SDSSDR7:SpecObjAll AS s
```

```
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
```


SkyQuery
username: dobos | account | log out
home
schema browser
query
job history
my db
import
output
help
Check
Submit
Output:
Task name:
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

Only the first steps…

- Different geometries
  - Resolved sources have shapes: galaxies, etc.
  - Varies as a function of wavelength

Science is Interactive

Science is Interactive

"Too much to be accurate"

By the time you do the calculations, the answer may have changed…

Science is Interactive

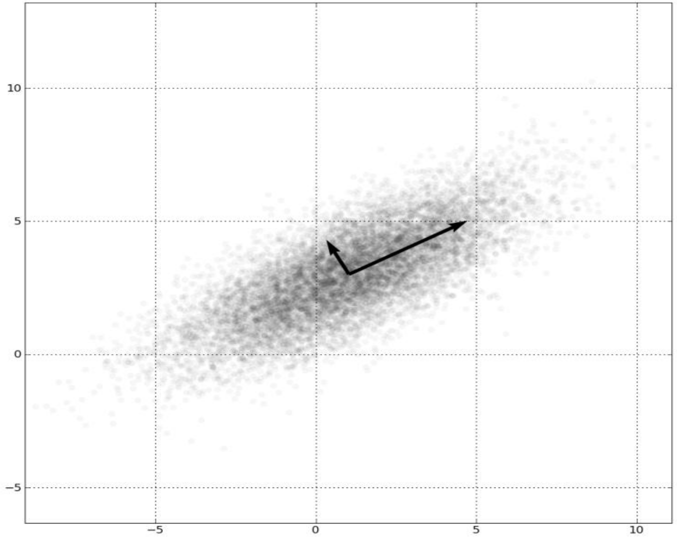
“Too much to be accurate”

By the time you do the calculations, the answer may have changed…

- Randomized algorithms
  - Improving answers over time
  - Rethink the basic methods

Principal Component Analysis

- Principal directions
  - Directions of largest variations
  - Eigenproblem of covariances
  - Singular Value Decomposition
- Problems
  - Needs lots of memory
  - Only need largest ones
  - Very sensitive to outliers



Streams of Data

- Mean

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\mu_n = \frac{n-1}{n}\mu_{n-1} + \frac{1}{n}x_n$$

$$\mu = \gamma\,\mu_{\text{prev}} + (1-\gamma)x$$

Streams of Data

- Mean

$$\mu = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\mu_n = \frac{n-1}{n}\mu_{n-1} + \frac{1}{n}x_n$$

$$\mu = \gamma\,\mu_{\text{prev}} + (1-\gamma)\,x$$

- Covariance

$$C = \gamma\,C_{\text{prev}} + (1-\gamma)\,y\,y^{\mathrm{T}}$$

$$y = x - \mu_{\text{prev}}$$

Iterative evaluation!

Streaming PCA

- Initialization
  - Eigensystem of a small, random subset
  - Truncate at *p* largest eigenvalues

$$C \approx E_p \Lambda_p E_p^{\mathrm{T}}$$

- Incremental updates
  - Mean and the low-rank *A* matrix
  - SVD of *A* yields new eigensystem

$$C \approx \gamma E_p \Lambda_p E_p^{\mathrm{T}} + (1-\gamma) y y^{\mathrm{T}}$$

$$\approx A A^{\mathrm{T}}$$

- Randomized algorithm!

Robust PCA

- PCA minimizes σ_{RMS} of the residuals $r = y - Py$
  - Quadratic formula: Σr^2 extremely sensitive to outliers
- We optimize a robust M-scale σ^2 (Maronna 2005)
  - Implicitly given by

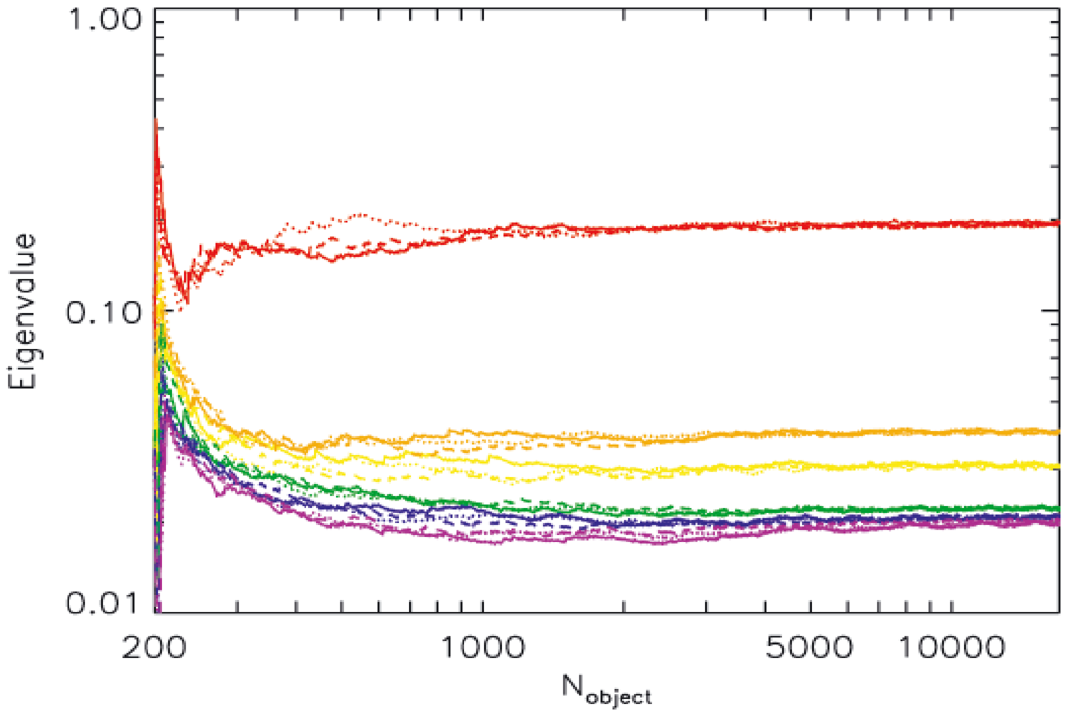
$$\frac{1}{N}\sum_{n=1}^{N}\rho\left(\frac{r_n^2}{\sigma^2}\right)=\delta$$

$$\mu = \left(\sum w_n x_n\right) \Big/ \left(\sum w_n\right)$$

$$C = \sigma^2\left[\sum w_n (x_n-\mu)(x_n-\mu)^{\mathrm{T}}\right] \Big/ \left(\sum w_n r_n^2\right)$$

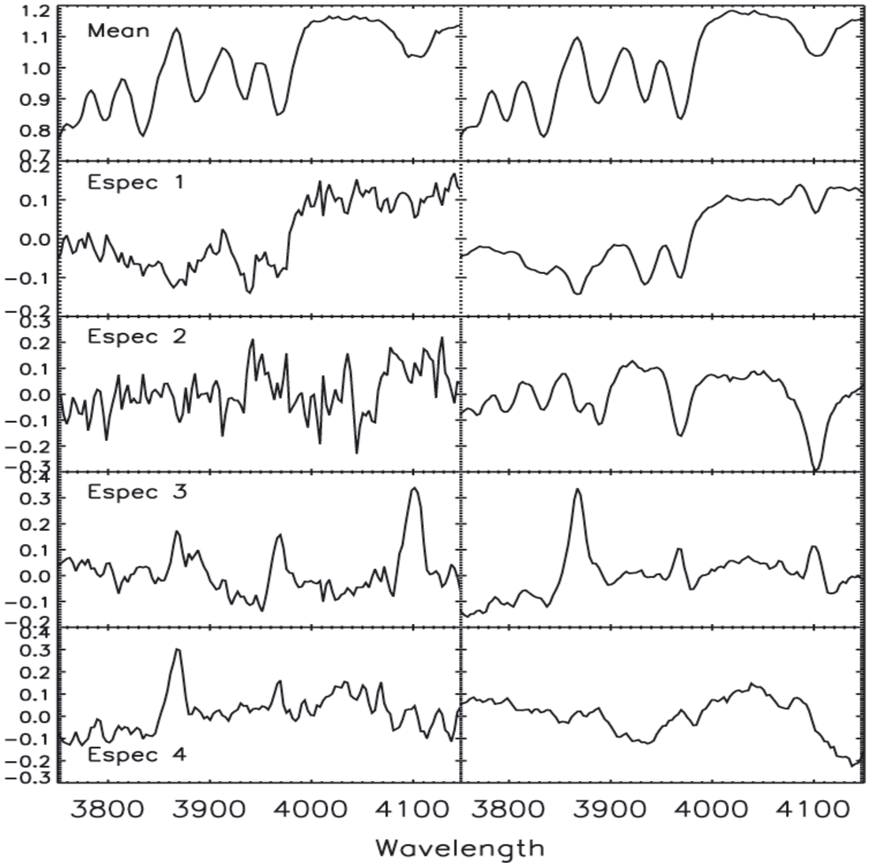
- Fits in with the iterative method (TB+ 2009)

Galaxy Spectra



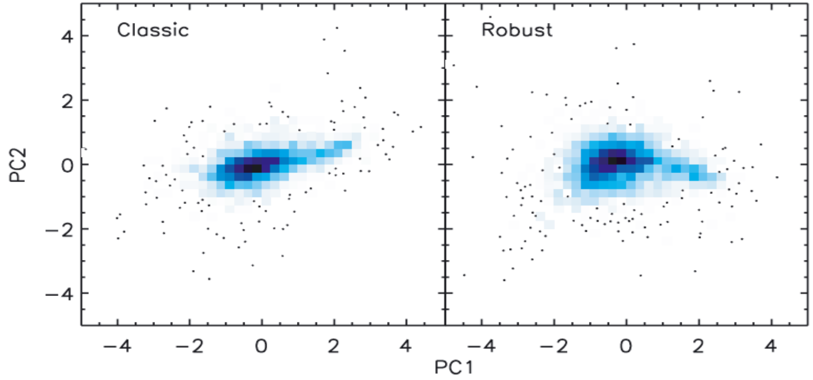
Galaxy Spectra

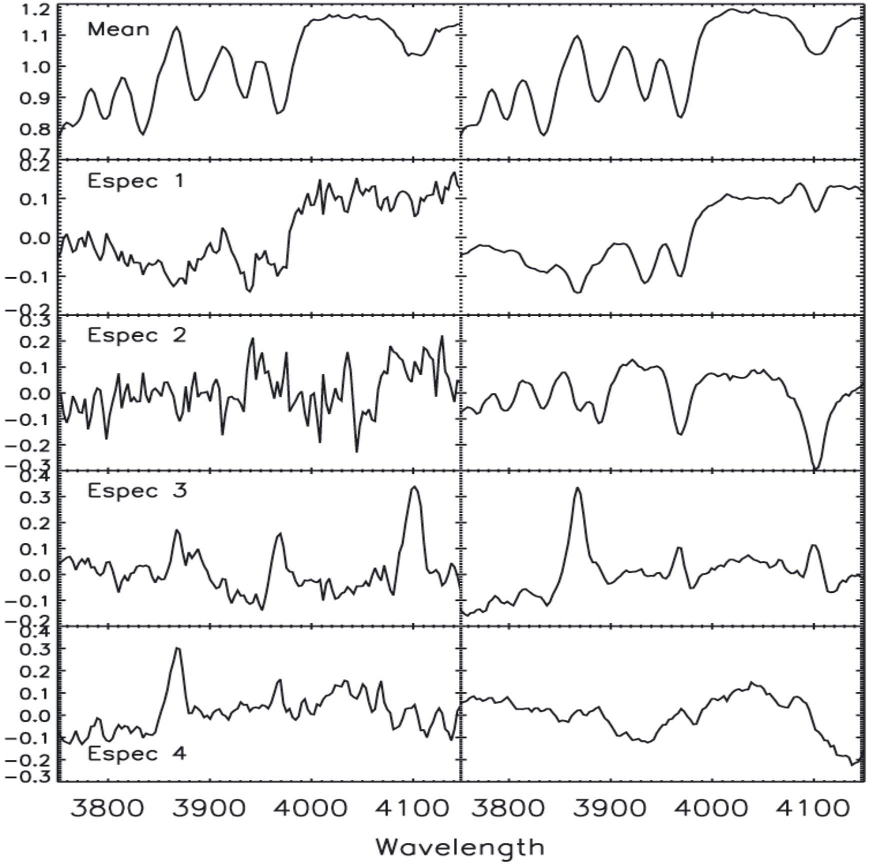
- High SNR eigenfunctions
  - Sign of robustness



Galaxy Spectra

- High SNR eigenfunctions
  - Sign of robustness
- It makes a difference





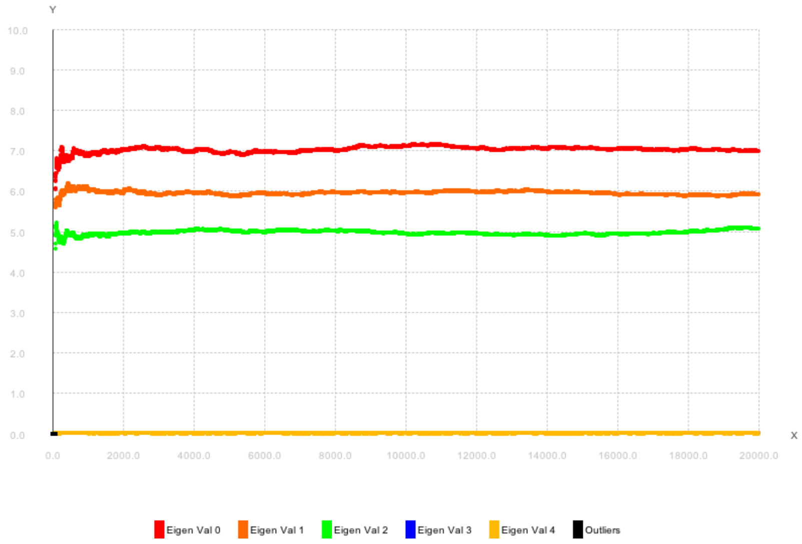
Synthetic Streams

Tamás Budavári

- 3D Gaussian rotated into 50D
  - Stretches: 7, 6, 5
  - Total Var = 110

- Plotting square roots of the top 5 eigenvalues

Streaming Classic PCA

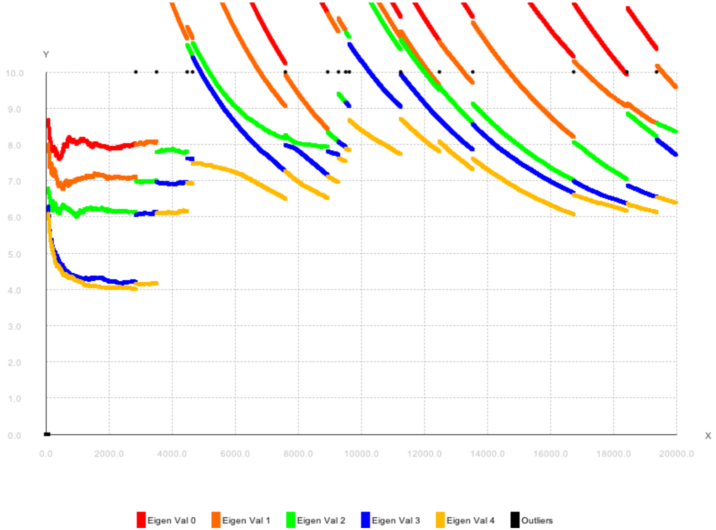


With Outliers

Tamás Budavári

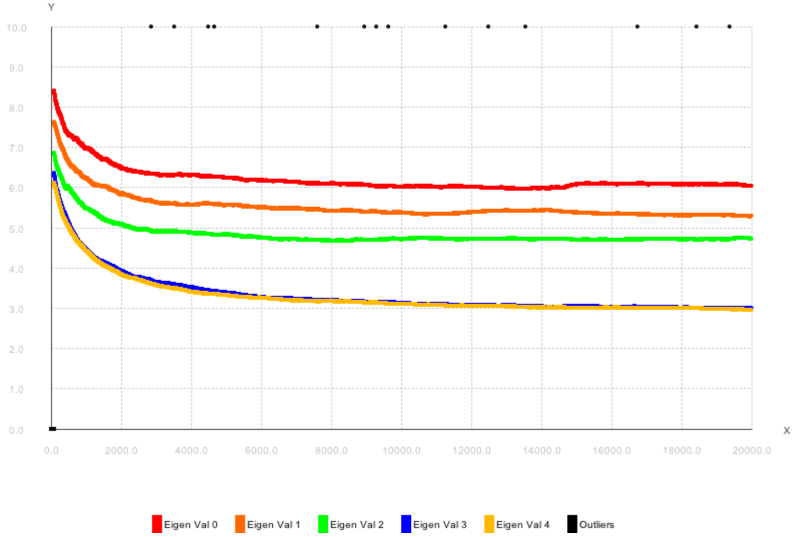
- Adding 0.1% outliers
  - σ = 100 in each bin
- Outliers take over the PCs
  - Instability, no convergence

Streaming Classic PCA



Robust Algorithm

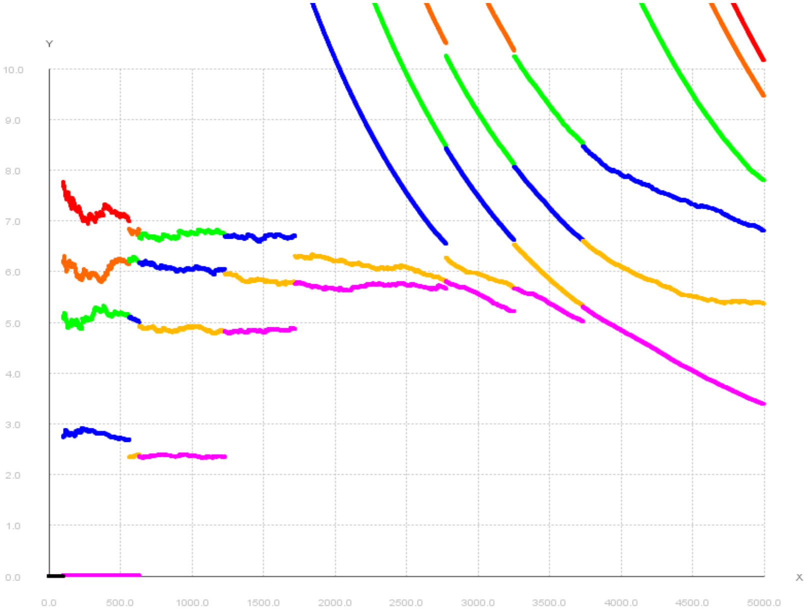
- Outliers under control
  - Marked on top
- Initialized with SVD
  - On a set of 100 vectors



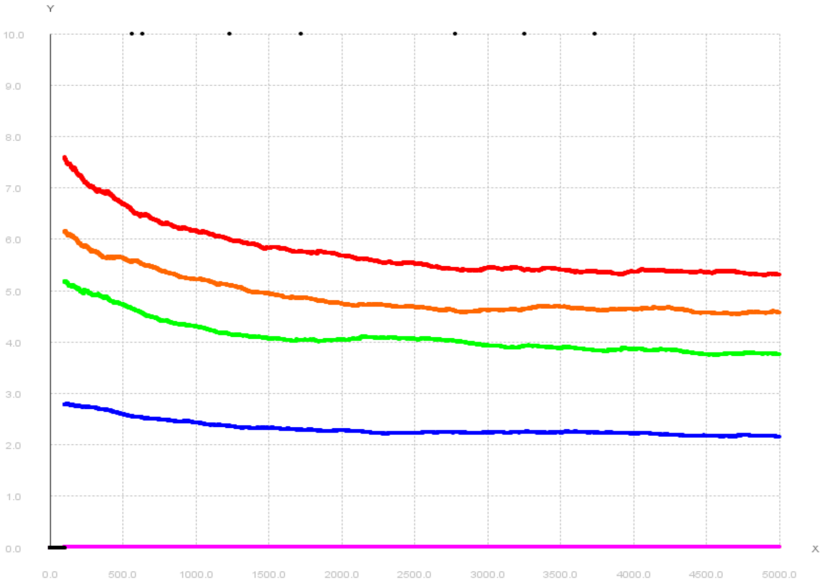
Streaming Robust PCA

Comparison

- Classic



- Robust



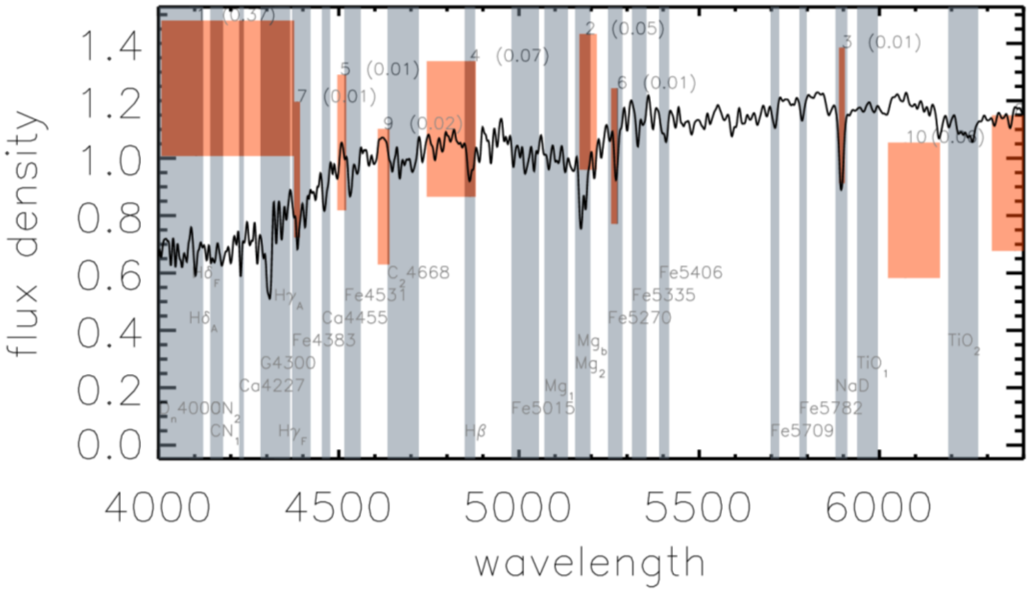
Sparsity

Tamás Budavári

- Important wavelengths
  - Cf., the Lick indices

(RE)DISCOVER ASTRONOMY!

Yip, Mahoney+ (2012, in prep)



Time Domain (the new window)

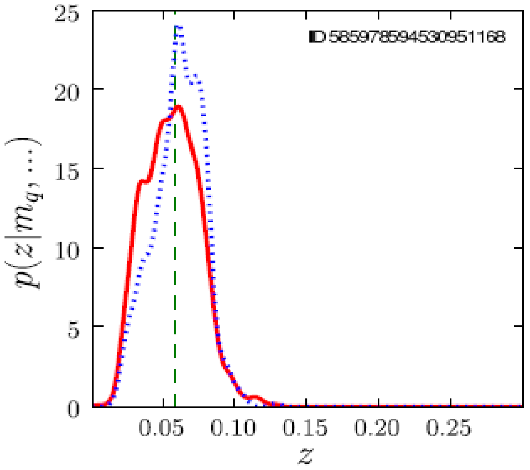
- Online aggregation of survey data
  - Learning the night sky from a series of images
    - Correcting for atmospheric distortions
  - Finding transient events
- Online federation of event streams
- Online outlier detection – new discoveries!

Uncertainties

- No data points
- Measurements with errors

Inverse Problems

- Bayesian inference but computational bottlenecks
  - High dimensional models w/ empirical priors
  - Non-parameteric?
- Deconvolution with degeneracies
  - Aided by physics
- Photometric inversion
  - TB (2009) $p(\boldsymbol{\xi}|\boldsymbol{y}_q, M) = \int d\boldsymbol{x}\; p(\boldsymbol{\xi}|\boldsymbol{x})\, p(\boldsymbol{x}|\boldsymbol{y}_q, M)$



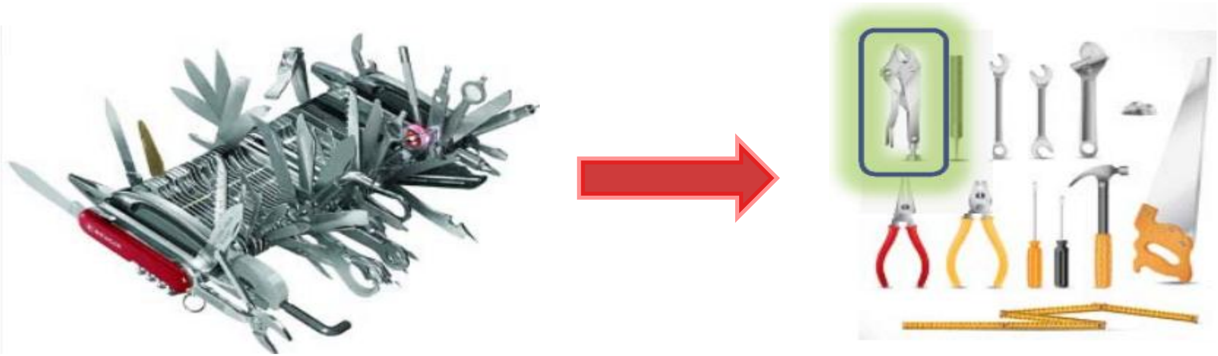
Trends in Computing

New Limitation is Energy!

- Power to compute the same thing?
  - CPU is 10× less efficient than a digital signal processor
  - DSP is 10× less efficient than a custom chip
- New design: multicores with slower clocks
  - But the interconnect is expensive
  - Need simpler components

Emerging Architectures

- Andrew Chien: 10×10 to replace the 90/10 rule
  - Custom modules on chip, cf. SoC in cellphones



Emerging Architectures

- Andrew Chien: 10×10 to replace the 90/10 rule
  - Custom modules on chip, cf. SoC in cellphones

- Statistics on a specialized units, e.g., codecs?

GPUs Evolved to be General Purpose

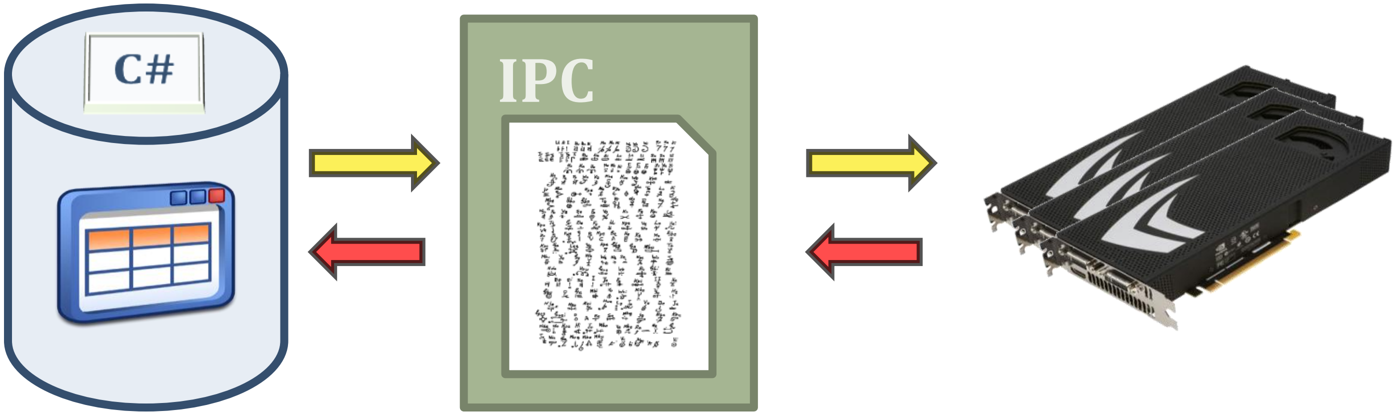
- Virtual world: simulation of real physics
  - C for CUDA and OpenCL with lots of libraries

- 512 cores ~25k threads running 1 billion/sec
- Old algorithms built on wrong assumption
  - Today processing is free but memory is slow

New programming paradigm!

Integrated with Databases

- Scientific computations on the GPU
  - Driven remotely from SQL User-Defined Functions

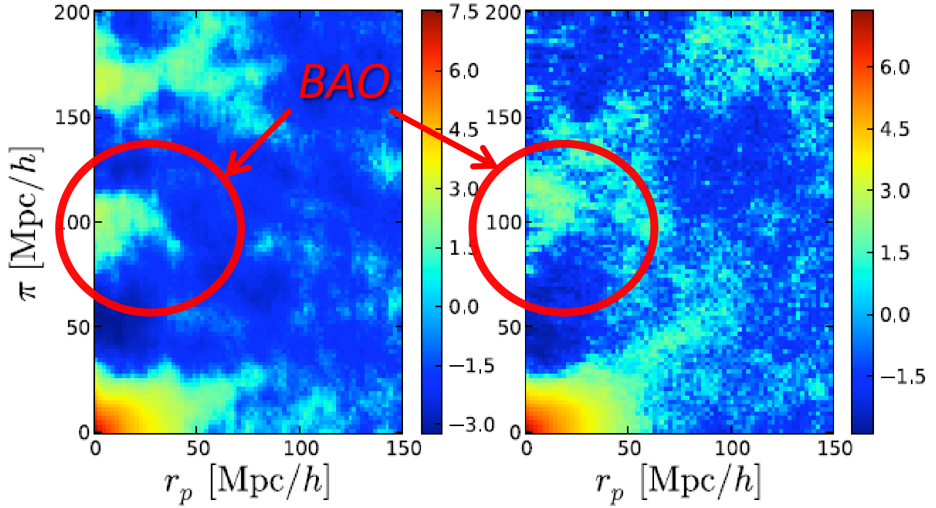


Baryon Acoustic Oscillations

- 600 trillion galaxy pairs
  - Correlation function:
- C for CUDA on GPUs

SPATIAL STATISTICS!

Tian, Neyrinck, TB & Szalay (2011)



Matching 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
...
```

Summary

- Astronomy has always been data-driven
  - Data-intensive for decades (since SDSS)
- Need new approaches and algorithms
  - But prototypes are not enough to scale
- Many promising directions to explore

Perfect timing for the SAMSI Big Data Program!