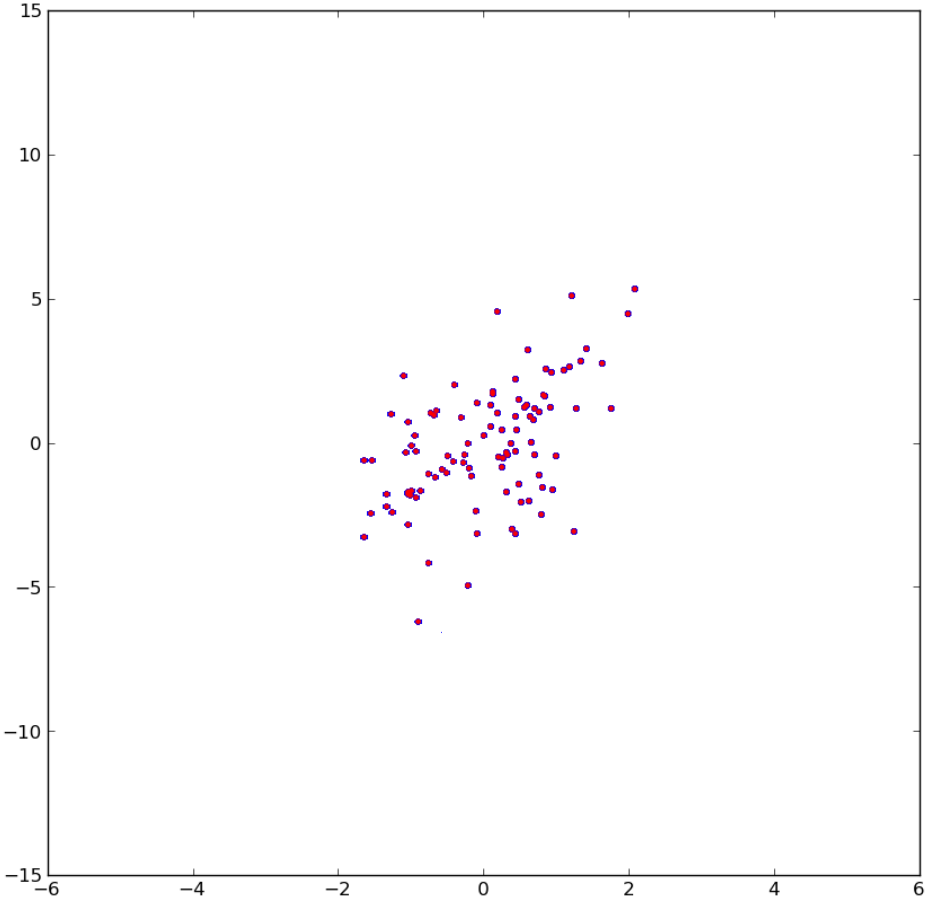
MULTIPLE EXPOSURES IN LARGE SURVEYS

12/12/2013

Tamás Budavári / Johns Hopkins University

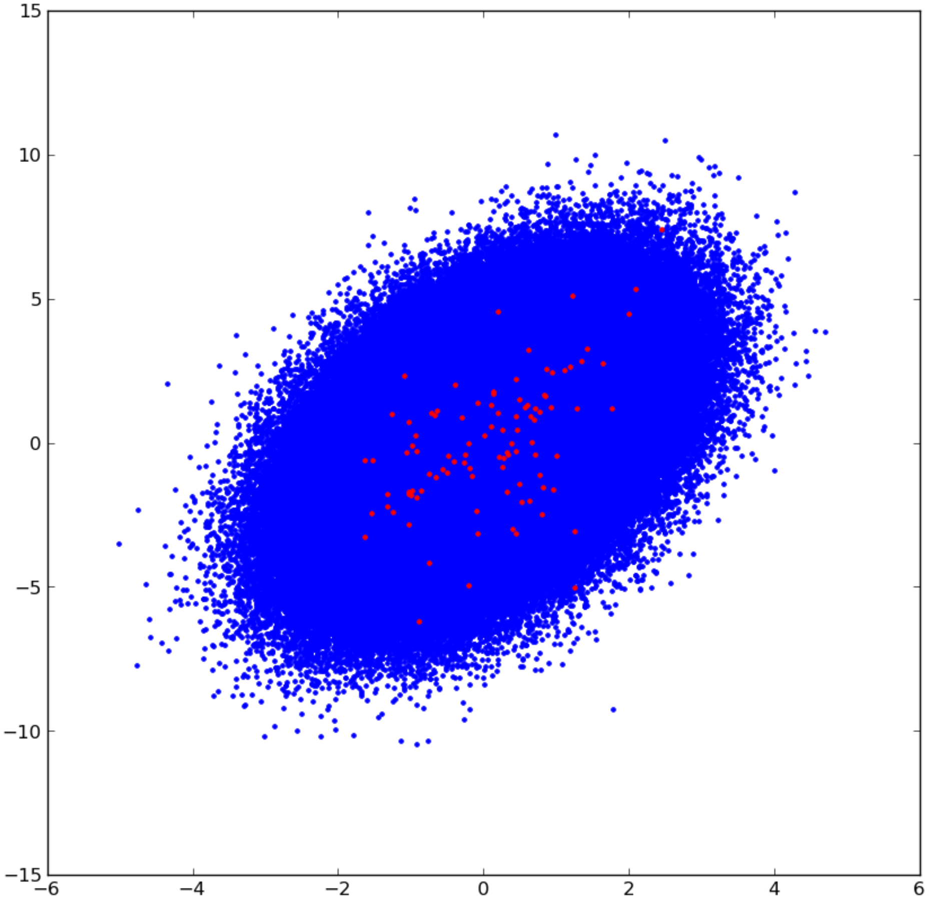
Big Data?

- Noisy
- Skewed
- Artifacts



Big Data?

- Noisy
- Skewed
- Artifacts

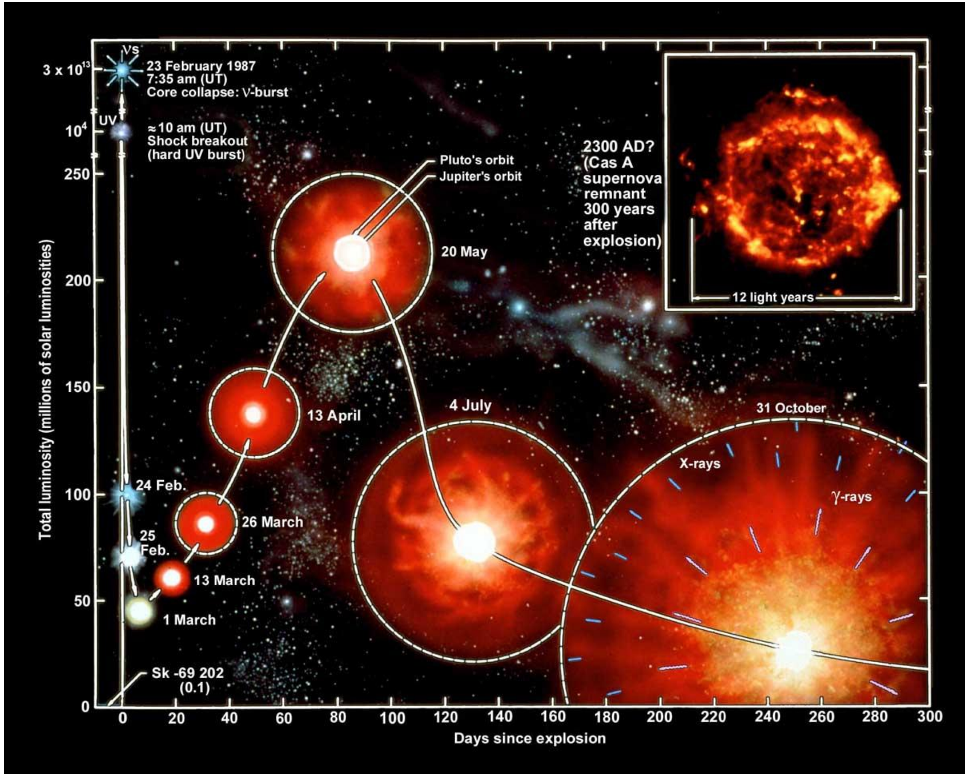


Serious Issues

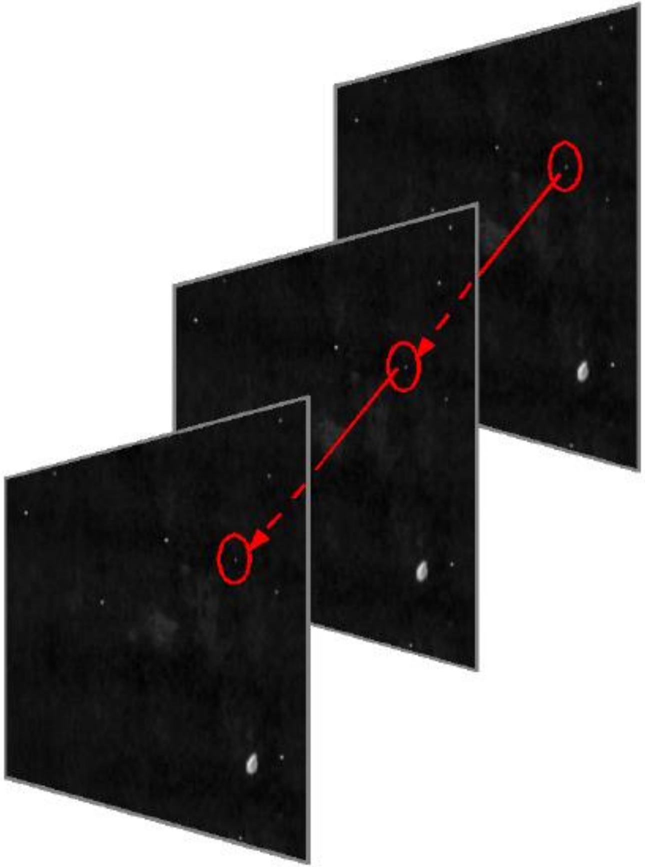
- Significant fraction of catalogs is junk
  - GALEX ~50%
  - PS1 3PI 50-80%
  - PS1 MDS >95%

- Textbook methods often fail due to artifacts
  - What are the good techniques?

Time Domain



νs
3 x 10^13
23 February 1987
7:35 am (UT)
Core collapse: ν-burst
UV
10^4
≈10 am (UT)
Shock breakout
(hard UV burst)
250
Pluto's orbit
Jupiter's orbit
2300 AD?
(Cas A
supernova
remnant
300 years
after
explosion)
12 light years
20 May
200
150
13 April
4 July
31 October
X-rays
γ-rays
100
24 Feb.
26 March
25
Feb.
13 March
50
1 March
Sk -69 202
(0.1)
0
0 20 40 60 80 100 120 140 160 180 200 220 240 260 280 300
Days since explosion
Total luminosity (millions of solar luminosities)



Time Series of Faint Sources?

- Co-add images and do forced photometry
  - Ideal if we have all observations but we never do
- Independent catalogs as we go
  - Need to dig in the noise to build good timeseries
- Goal is an incremental strategy to weed out noise
  - Otherwise catalogs are overwhelmed by junk

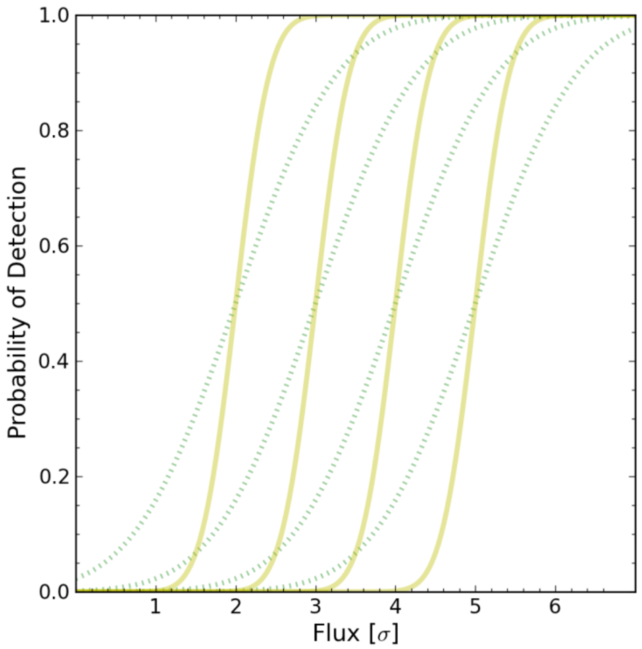
Detection Probability

- Measured flux is true + normal error $f_i = f + \epsilon_i$
- Probability of detection

$$P_f \equiv P(f_i > f_D | f) = \frac{1}{2}\,\mathrm{erfc}\left(\frac{f_D - f}{\sigma\sqrt{2}}\right)$$

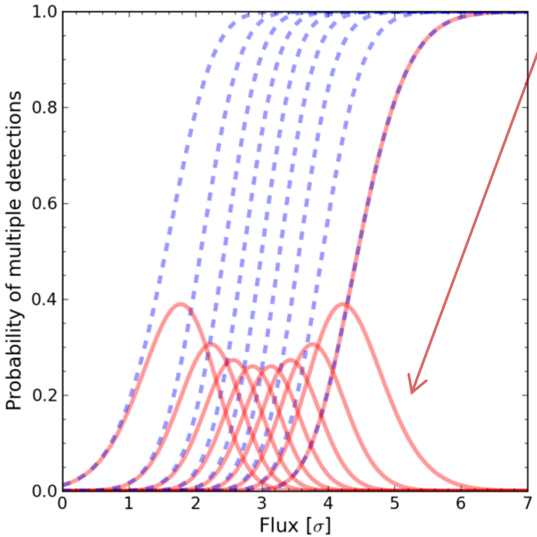
Detection Probability

- Measured flux is true + normal error $f_i = f + \epsilon_i$
- Probability of detection
  - As a function of the true flux
    - Thresholds at *2-, 3-, 4- & 5*σ
  - Sharper for 9-way stacks



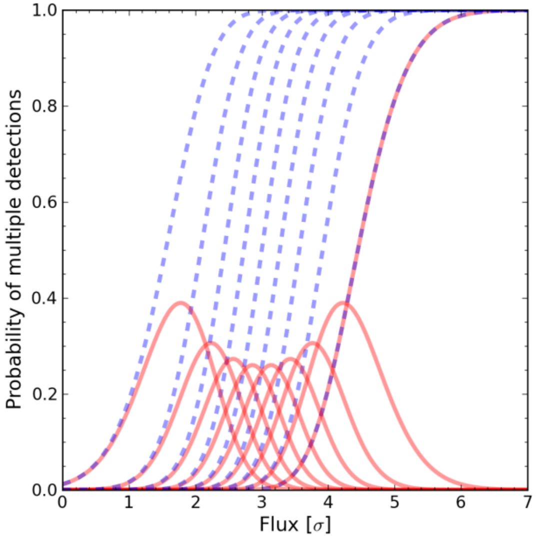
Detection Probability

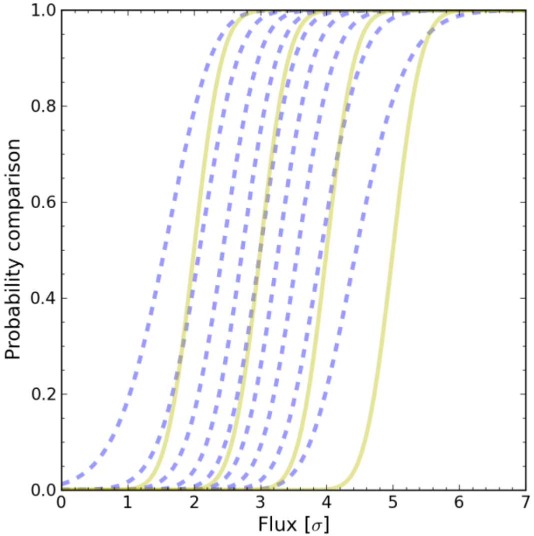
- Multiple exposures
  - Binomial $$P(n|k,f) = \binom{k}{n} P_f^n \left(1 - P_f\right)^{k-n}$$



Detection Probability

- Multiple exposures
  - Binomial $$P(n|k,f) = \binom{k}{n} P_f^n \, (1-P_f)^{k-n}$$





What is a Real Source?

- Is it “real” or just “noise” ?
  - Bayesian hypothesis testing

$$B = \frac{L_{\mathrm{real}}}{L_{\mathrm{noise}}}$$

What is a Real Source?

- Is it "real" or just "noise" ?
  - Bayesian hypothesis testing

$$B = \frac{L_{\rm real}}{L_{\rm noise}}$$

$$L_{\rm real} = \int df\, \pi(f)\, L(f)$$

$$L(f) = (1-P_f)^{k-n} \prod_i^n G(f_i; f, \sigma^2)$$

- Out of k observations n detections of f_i fluxes

Apparent Flux Distribution

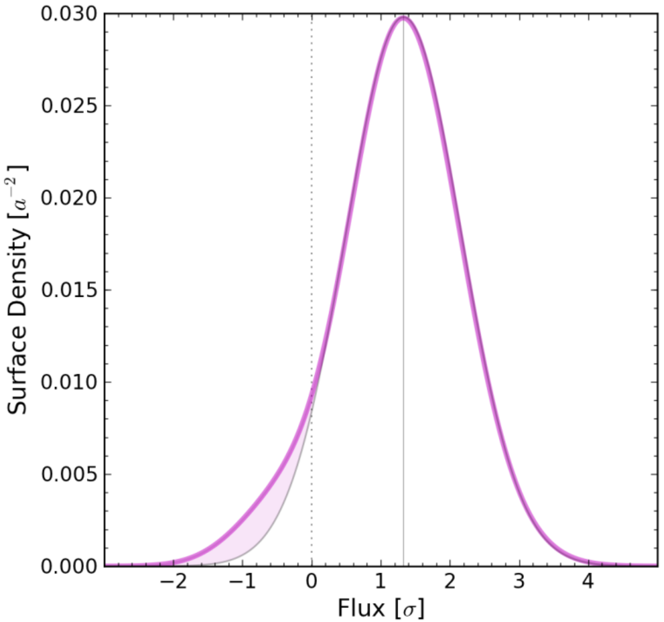
- Galaxy number-counts as fn of magnitude
  - Empirical relation approximately shows

$$dN \propto 10^{0.4m}\, dm \propto \frac{df}{f^2}$$

- More and more fainter and fainter sources!
  - But there is a limit, cf. Olbers' paradox

Distribution of Noise Peaks

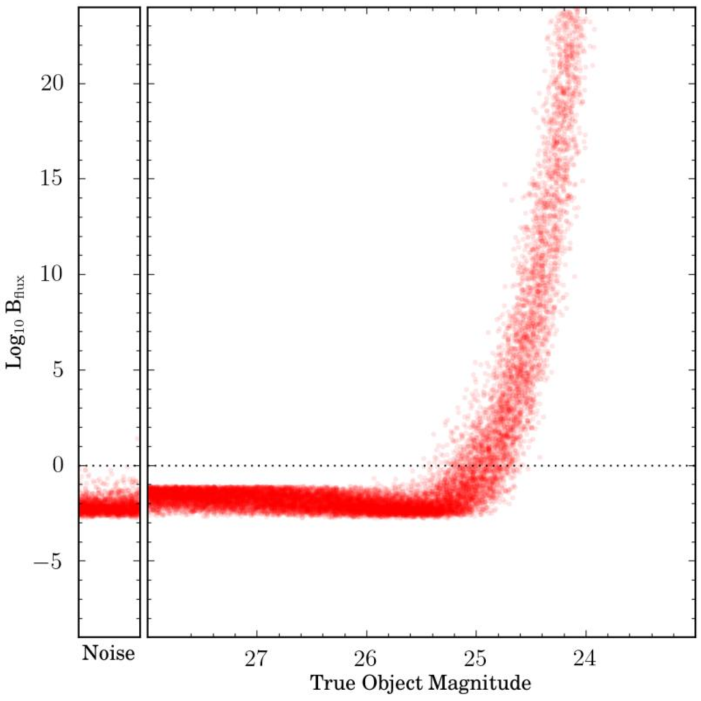
- Local maxima of continuous Gaussian random field
  - Cf., $P(\mathbf{k})$ by Barden, Bond, Kaiser, Szalay (BBKS; 1986)
  - Now in 2D:



Something Like LSST

- Simulation
  - Sky at 5σ is 24 mag
  - Object limit is at 28

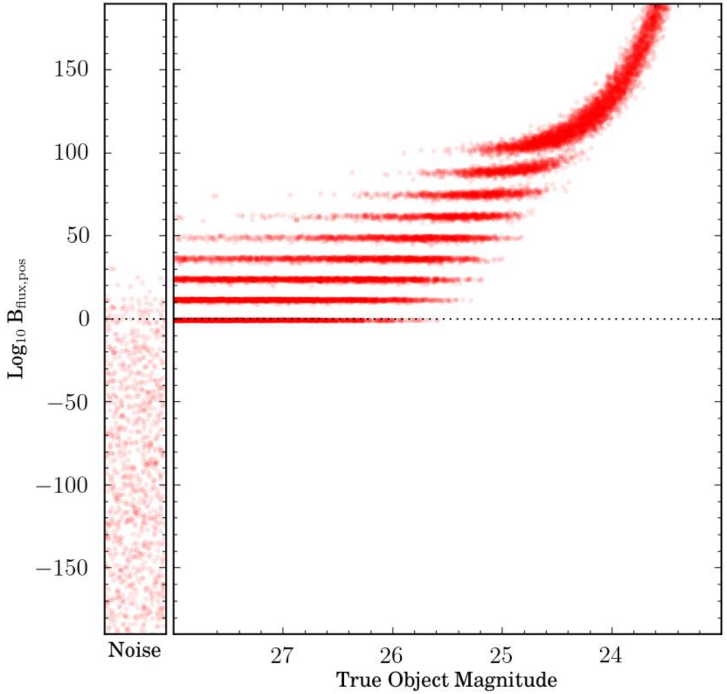
- Bayes factor
  - Considering only fluxes



Adding Directions

- Bayes factor from cross-id
  - As TB & Szalay (2008)
  - Faint sources can be distinguished based on their celestial coordinates

Always at "same" place!



Cross-Identification

- Hard problem
  - Computationally, Scientifically & Statistically
  - Need symmetric *n*-way solution
  - Need reliable quality measure

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

Same or Not?

OR
- The Bayes factor

$$B(H,K|D) = \frac{p(D|H)}{p(D|K)}$$

SAME
- *H:* all observations of the same object

NOT
- *K:* might be from separate objects

Same or Not?

Tamás Budavári

OR

- The Bayes factor

$$B(H,K|D) = \frac{p(D|H)}{p(D|K)}$$

SAME

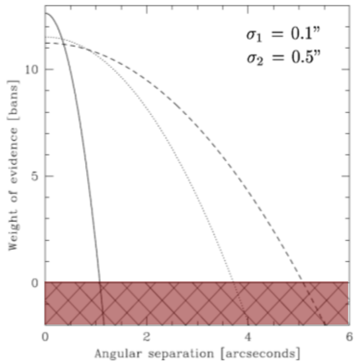
- *H:* all observations of the same object
  - Same properties, e.g., coordinates, brightness

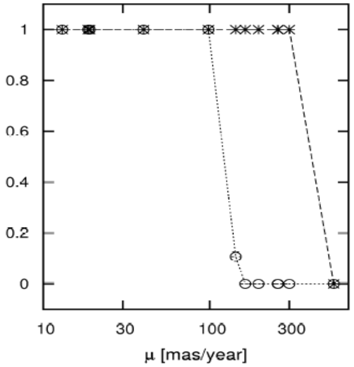
NOT

- *K:* might be from separate objects
  - Properties could be different

Works in General

- Analytic results for Gaussian errors
  - Incremental *n*-way strategy
- We can find moving stars
  - With unknown velocities
- Matching events in time
  - E.g., supernovae





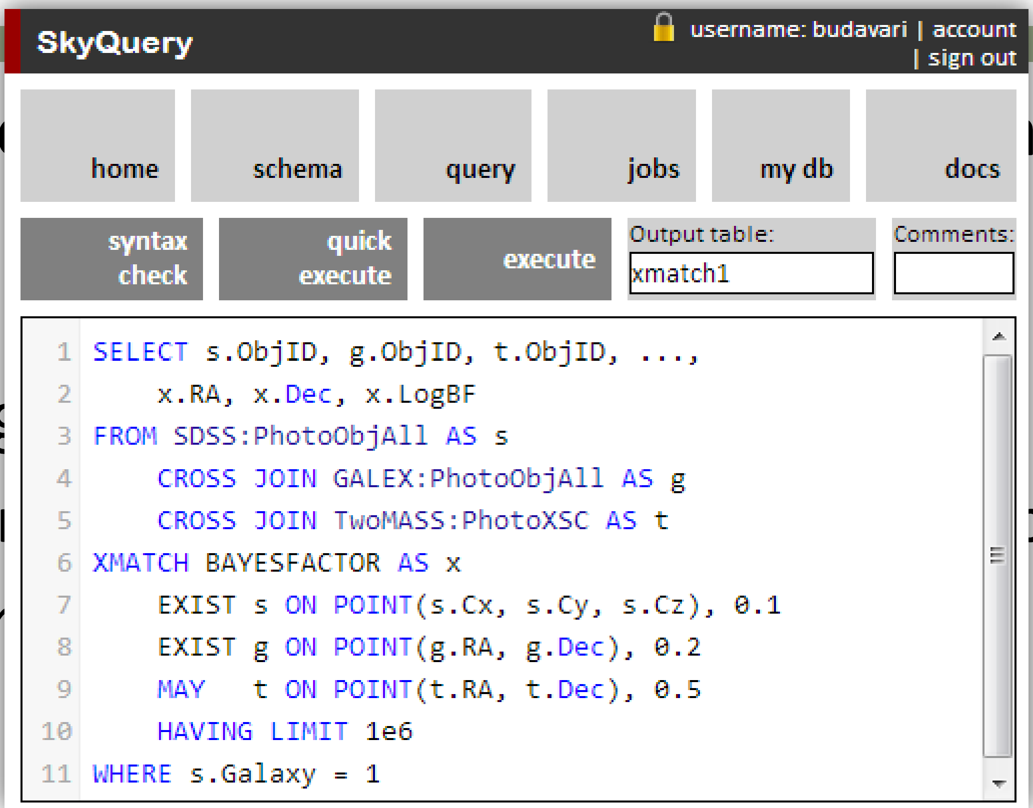
2013

SkyQuery – the new generation!

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

SkyQuery – the new generation!

- Dynami … ses
  - Query
- The 3rd g … ber
  - Cluste … ps
  - Proper … rs

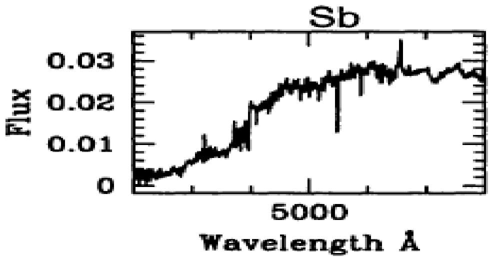


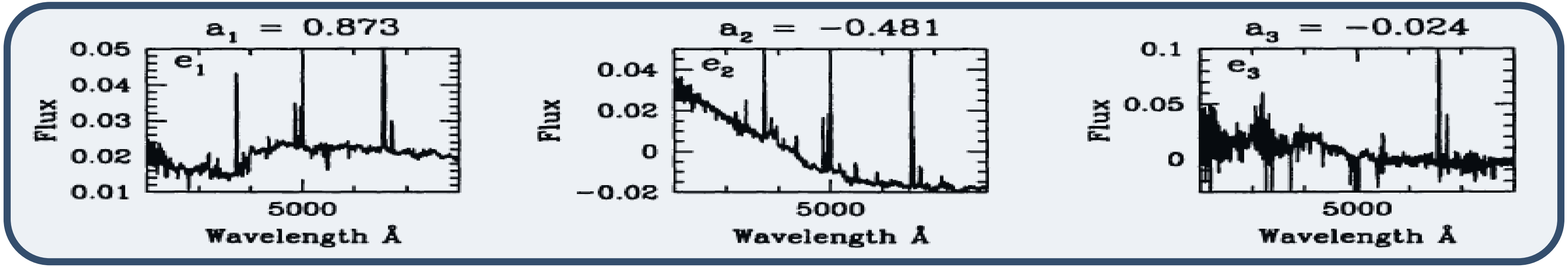
Only the first steps…

- Resolved shapes: radio morphology *(Fan, TB+ 2014)*
- Colors to augment matches *(Marquez, TB, Sarro 2014)*

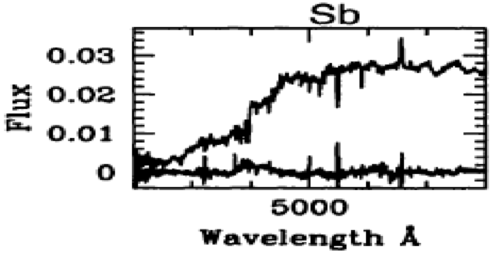
Galaxy Light ~ Linear Combination

- Principal Components Analysis (PCA)



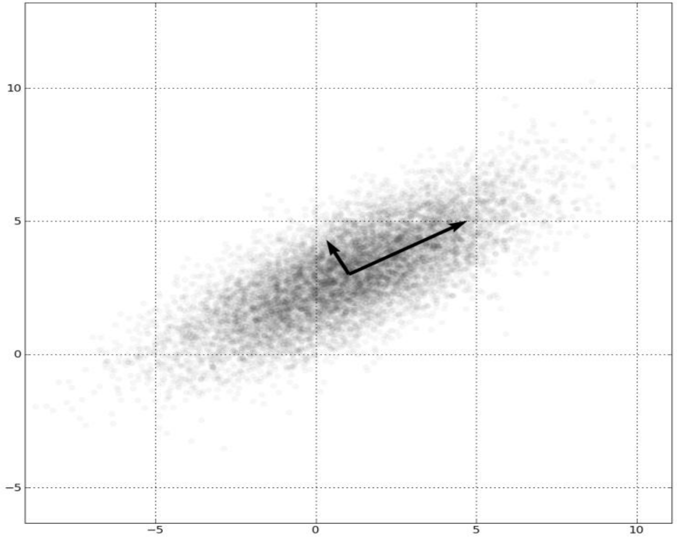


- SDSS galaxy type
  - On a big memory machine



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

Science is Interactive

"Too much to be accurate"

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

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

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x}_n$$

$$\boldsymbol{\mu}_n = \frac{n-1}{n}\boldsymbol{\mu}_{n-1} + \frac{1}{n}\boldsymbol{x}_n$$

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

- Covariance

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

$$\boldsymbol{y} = \boldsymbol{x} - \boldsymbol{\mu}_{\text{prev}}$$

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

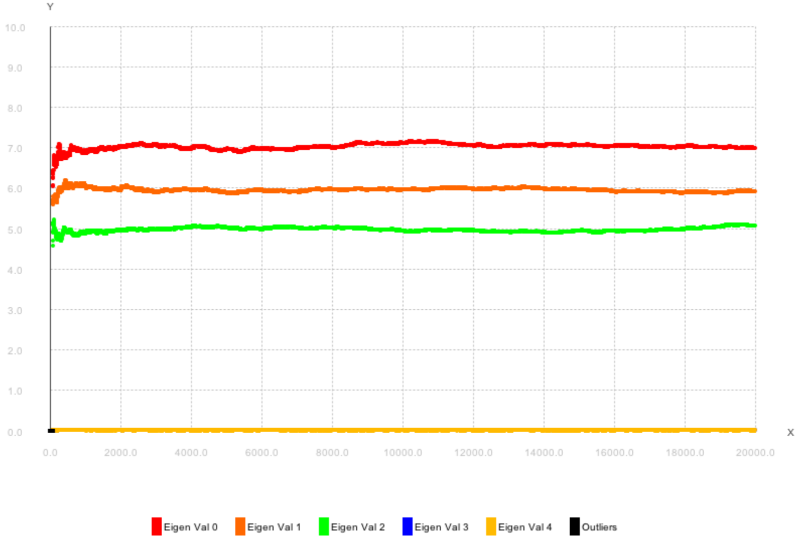
$$C \approx \gamma E_p \Lambda_p E_p^{\mathrm{T}} + (1-\gamma) y y^{\mathrm{T}}$$

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

- Randomized algorithm!

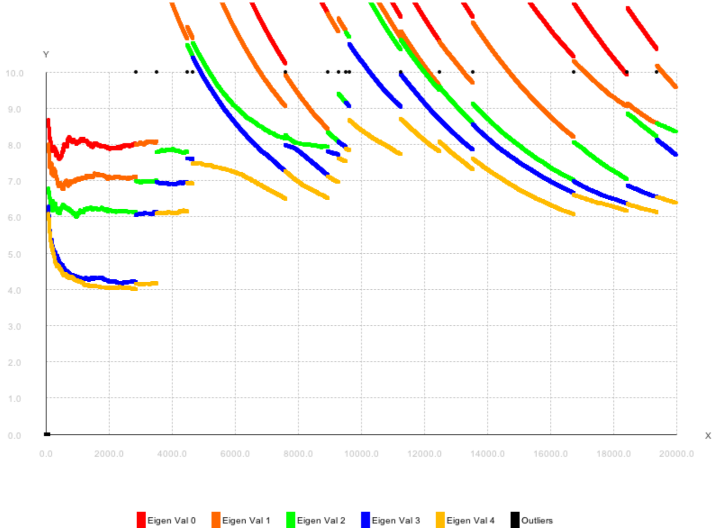
Streaming PCA

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



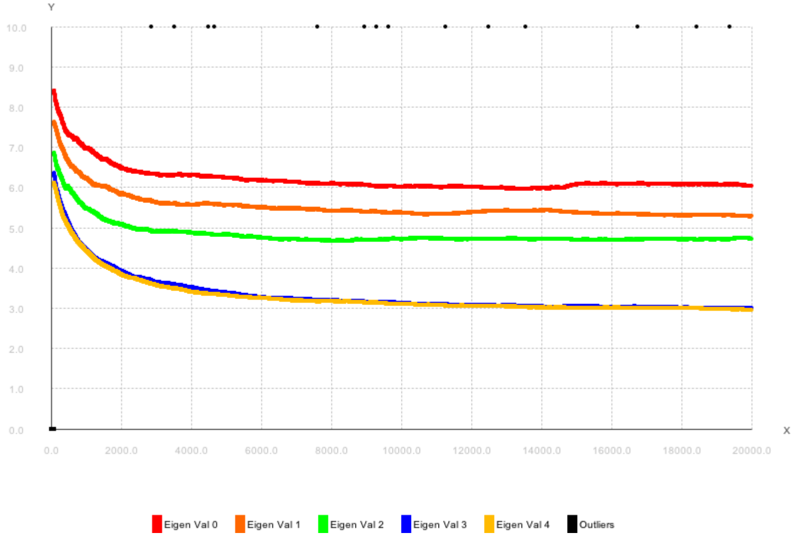
With Outliers

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



Robust Algorithm

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



Summary

- Plan for the junk
  - Proper statistics save money and gain speed
- Incremental randomized strategies scale
  - Crossmatching, embeddings, ML, etc.
- Not there, yet
  - Need new methods and tools