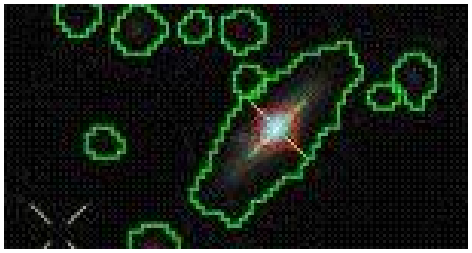
Astrophysics with Terabyte Datasets

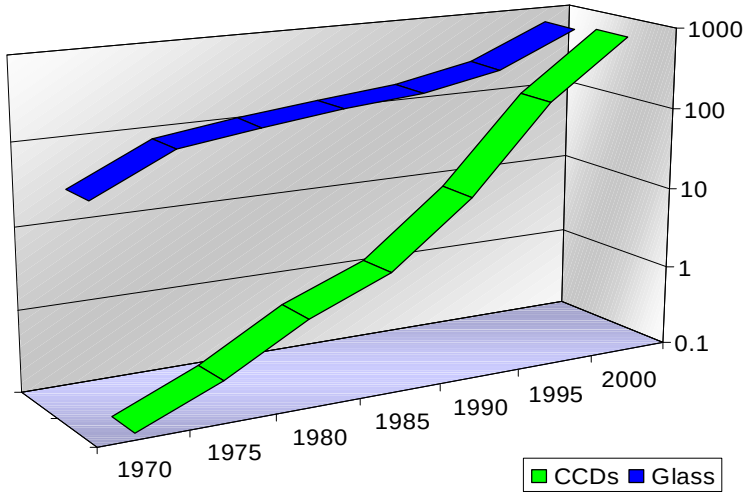
Alex Szalay, JHU
and Jim Gray, Microsoft Research



Living in an Exponential World

- Astronomers have a few hundred TB now
  - *1 pixel (byte) / sq arc second ~ 4TB*
  - *Multi-spectral, temporal, … → 1PB*
- They mine it looking for

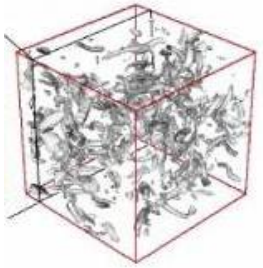
  new (kinds of) objects or
  more of interesting ones (quasars),
  density variations in 400-D space
  correlations in 400-D space
- Data doubles every year
  - *Driven by Moore's Law*
- Data is public after 1 year
  - *50% of the data is public*
  - *Same access for everyone*



Evolving Science

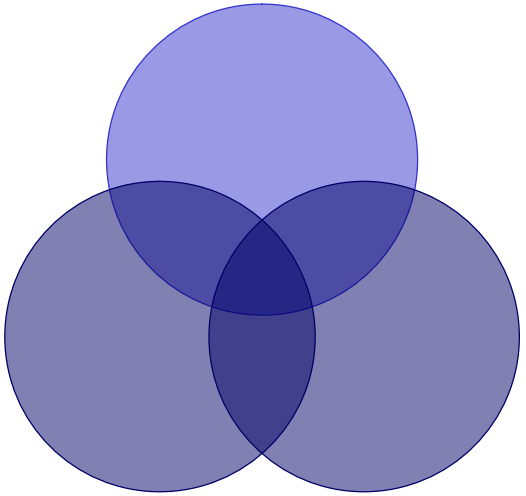
- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today:
  data exploration (eScience)
  synthesizing theory, experiment and computation with advanced data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4pGr}{3} - \mathrm{K}\frac{c^2}{a^2}$$



Making Discoveries

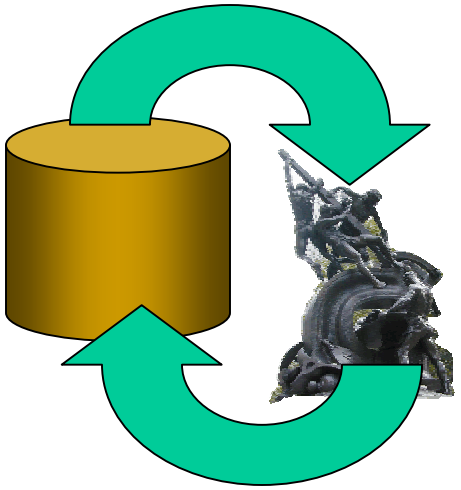
- Where are discoveries made?
  - *At the edges and boundaries*
  - *Going deeper, collecting more data, using more colors….*
- Metcalfe's law
  - *Utility of computer networks grows as the number of possible connections: $O(N^2)$*
- Data federations
  - *Federation of N archives has utility $O(N^2)$*
  - *Possibilities for new discoveries grow as $O(N^2)$*
- Current sky surveys have proven this
  - *Very early discoveries from SDSS, 2MASS, DPOSS*



Smart Data

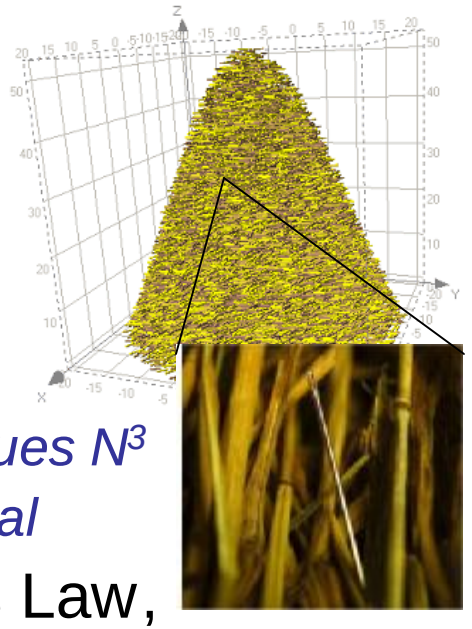
- If there is too much data to move around,

  take the analysis to the data!
- Do all data manipulations at database
  - *Build custom procedures and functions in the database*
- Automatic parallelism guaranteed
- Easy to build-in custom functionality
  - *Databases & Procedures being unified*
  - *Example temporal and spatial indexing*
  - *Pixel processing*
- Easy to reorganize the data
  - *Multiple views, each optimal for certain analyses*
  - *Building hierarchical summaries are trivial*
- Scalable to Petabyte datasets



active databases!

Next-Generation Data Analysis

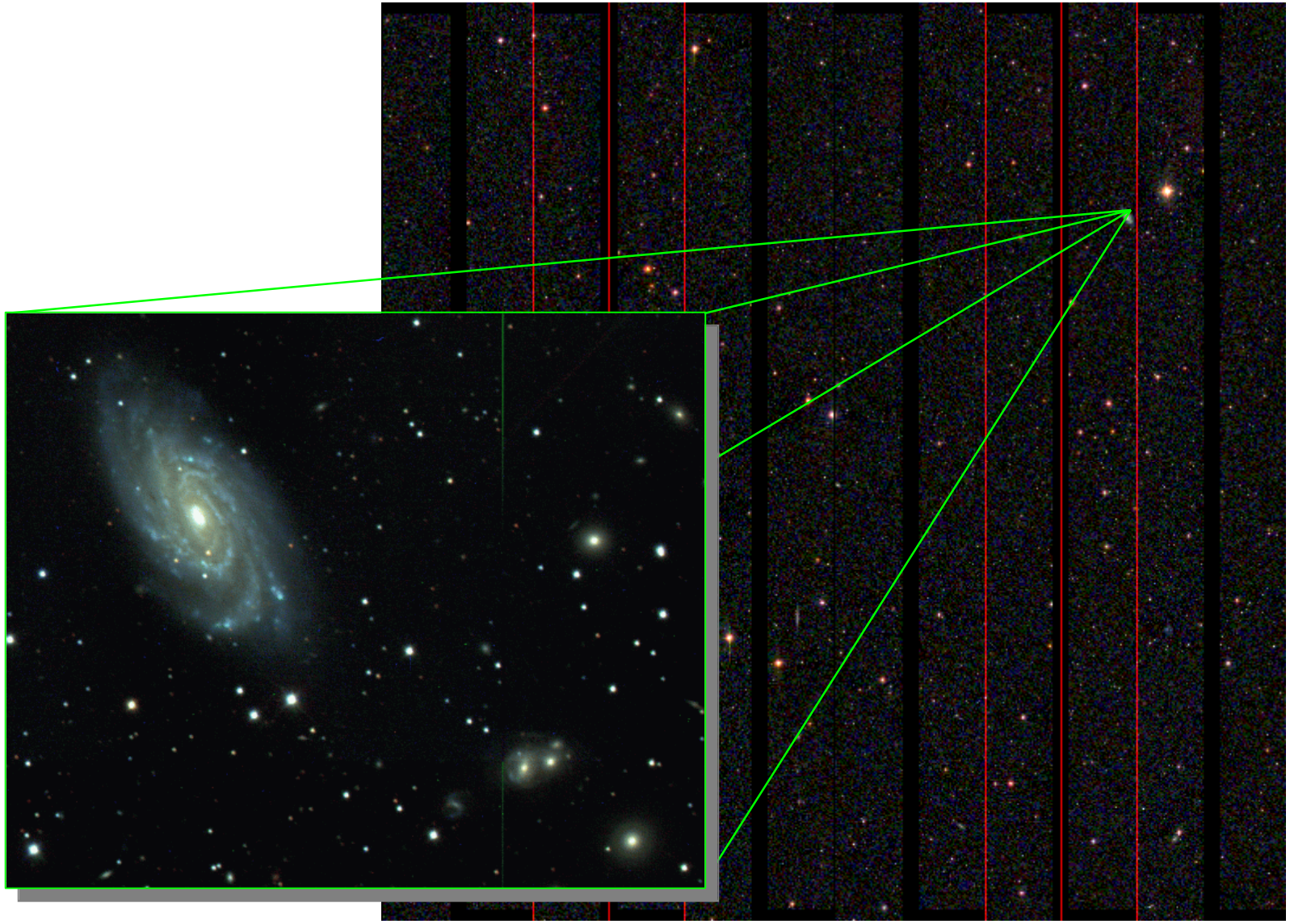


- Looking for
  - *Needles in haystacks – the Higgs particle*
  - *Haystacks: Dark matter, Dark energy*
- Needles are easier than haystacks
- ‘Optimal’ statistics have poor scaling
  - *Correlation functions are N^2, likelihood techniques N^3*
  - *For large data sets main errors are not statistical*
- As data and computers grow with Moore’s Law, we can only keep up with $N \log N$
- A way out?
  - *Discard notion of optimal (data is fuzzy, answers are approximate)*
  - *Don’t assume infinite computational resources or memory*
- Requires combination of statistics & computer science

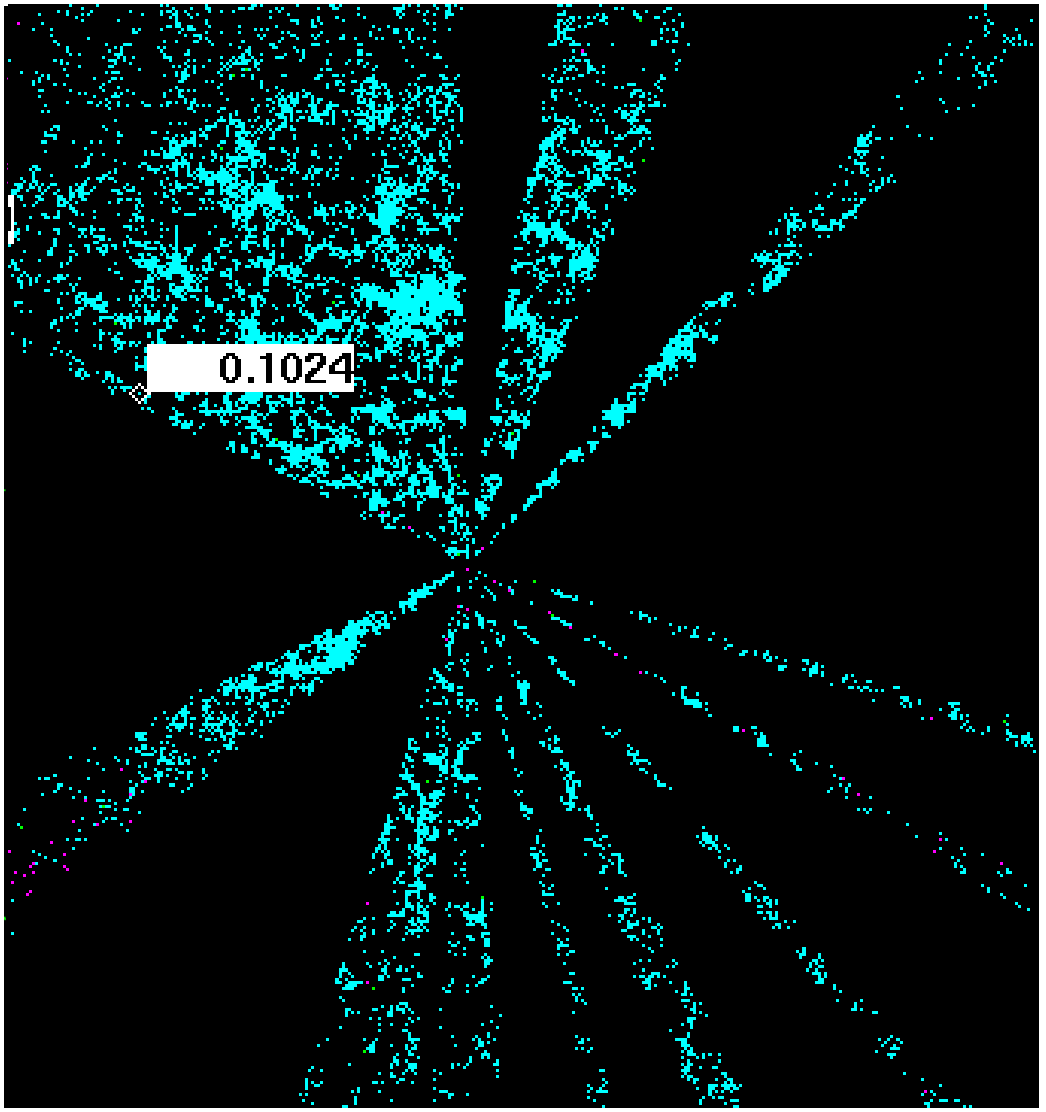
Astronomical Data

- Imaging
  - *2D map of the sky at multiple (20+) wavelengths*
- Derived catalogs
  - *subsequent processing of images (segmentation)*
  - *extracting object parameters (400+ per object)*
- Spectroscopic follow-up
  - *spectra: more detailed object properties*
  - *clues to physical state and formation history*
  - *lead to distances: 3D maps*
- Numerical simulations (1B+ objects)
- All inter-related!

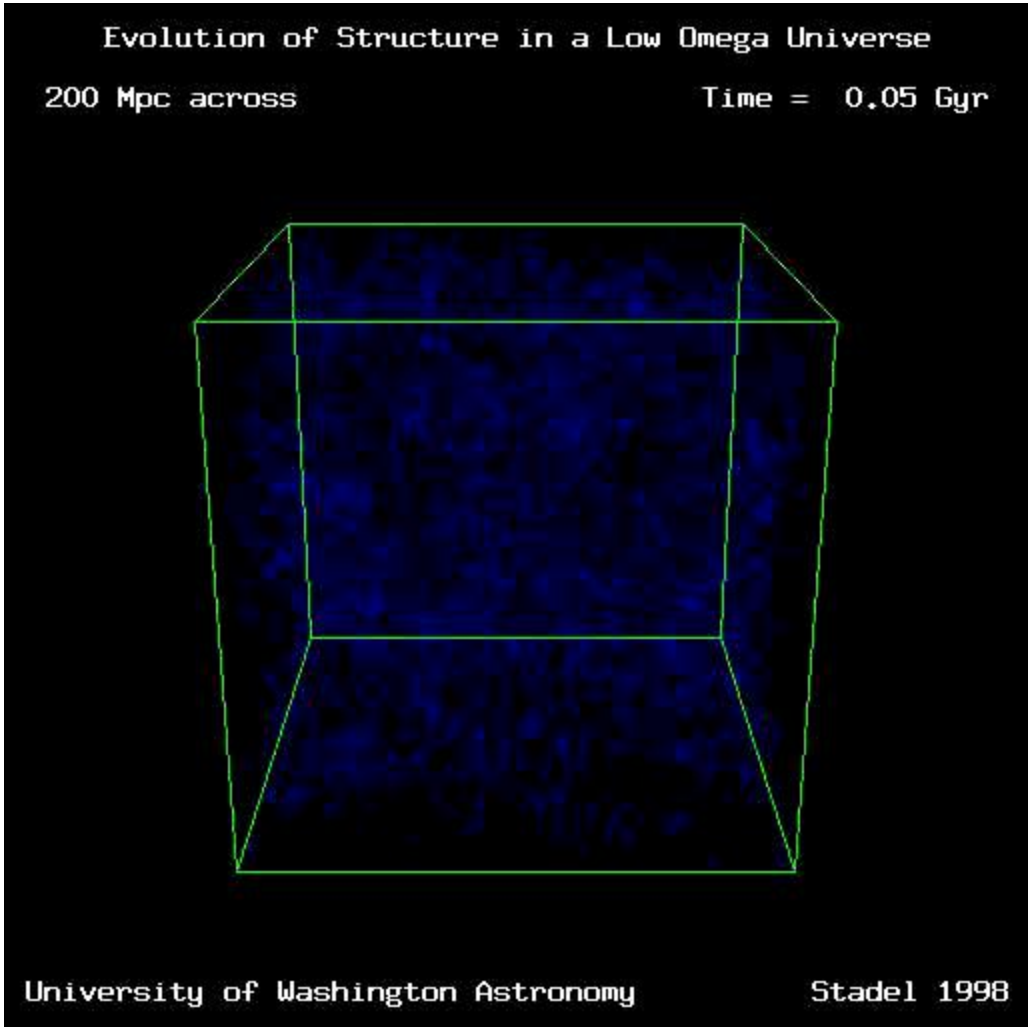
Imaging Data



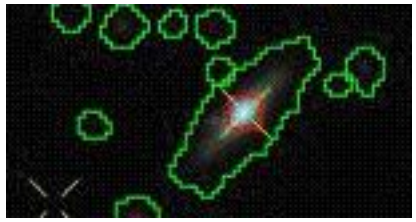
3D Maps



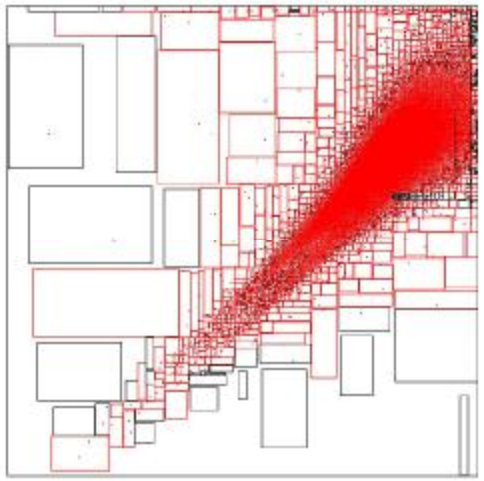
N-body Simulations



Visualization



- Mostly connecting pixels to objects
- Data correlated in multiple dimensions
- Scatter plots
  - *useless for cardinalities of 100M*
- Very large density contrast in data:
  - *Rare object fractions are 10^{-8}*
- Time domain astronomy beginning
  - *Petabytes/yr in 2010*
- Bimodal approach
  - *Scientists use primitive tools for obs data (scatter plot, xgobi)*
  - *Slightly better visualization tools for simulations*
  - *High end visualization for publicity only (Hayden…)*
  - *Schoolkids are used to better visualization than scientists*



The Virtual Observatory

- Premise: most data is (or could
- The Internet is the world's best
  - *It has data on every part of the sky*
  - *In every measured spectral band: optical, x-ray, radio..*
  - *As deep as the best instruments (2 years ago).*
  - *It is up when you are up*
  - *The "seeing" is always great*
  - *It's a smart telescope:*
    links objects and data to literature on them
- Software became the capital expense
  - *Share, standardize, reuse..*

SkyServer

- SkyServer is an educational website, based on the Sloan Digital Sky Survey data
- Access to an underlying multiterabyte database
- More than 50 hours of educational exercises
- Background on astronomy
- Tutorials and documentation
- Searchable web pages
- Interactive simple visual tools for data exploration
- Prototype eScience lab

http://skyserver.sdss.org/

Summary

- Data growing exponentially in many different areas
  - *Publishing so much data requires a new model*
- Multiple challenges for different communities
  - *publishing, statistics/data mining, data visualization, digital libraries, education, web services, …*
- Information at your fingertips:
  - *Students see the same data as professional scientists*
- More data coming: Petabytes/year by 2010
  - *We need scalable analysis algorithms*
  - *Move analysis to the data!*
- Same thing happening in all sciences
  - *High energy physics, genomics, cancer research, medical imaging, oceanography, remote sensing, …*
- Data Exploration: an emerging new branch of science
- Currently has no owner…