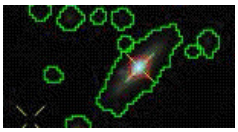
Australiasia eResearch Conference, Brisbane, 2007

Science in an Exponential World

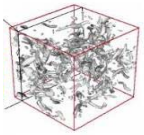
Alex Szalay
The Johns Hopkins University



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
  → new algorithms!

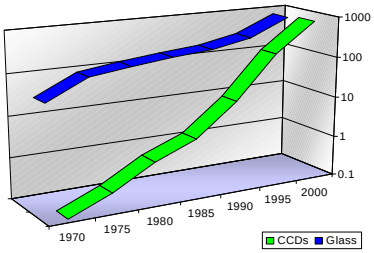

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



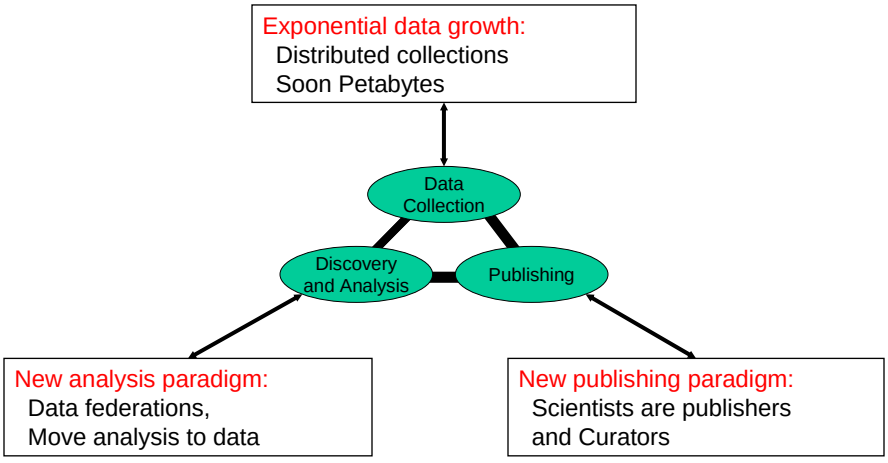

Living in an Exponential World

- 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
- Caused by the emergence of generations of inexpensive sensors + computing



The Challenges



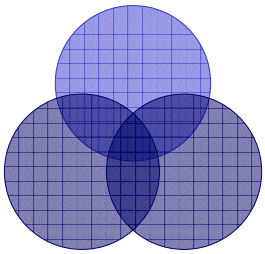
Collecting Data

- Very extended distribution of data sets:

  data on all scales!
- Most datasets are small, and manually maintained (Excel spreadsheets)
- Total amount of data dominated by the other end (large multi-TB archive facilities)
- Most bytes today are collected via electronic sensors

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: $O(N^2)$*
- Federating data *(the connections!!)*
  - *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*
  - *Genomics+proteomics*



Publishing Data

Roles	*Traditional*	*Emerging*
Authors	**Scientists**	**Collaborations**
Publishers	**Journals**	**Project www site**
Curators	**Libraries**	**Bigger Archives**
Consumers	**Scientists**	**Scientists**
	Linear	Exponential

- Exponential growth:
  - *Projects last at least 3-5 years*
  - *Data sent upwards only at the end of the project*
  - *Data will **never** be centralized*
- More responsibility on projects
  - *Becoming Publishers and Curators (session on Data Publishing)*
- Data will reside with projects
  - *Analyses must be close to the data*

Data Delivery: Hitting a Wall

FTP and GREP are not adequate

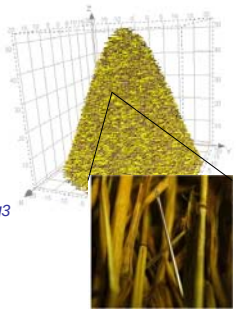
- You can GREP 1 MB in a second
- You can GREP 1 GB in a minute
- You can GREP 1 TB in 2 days
- You can GREP 1 PB in 3 years

- Oh!, and 1PB ~4,000 disks

- You can FTP 1 MB in 1 sec
- You can FTP 1 GB / min (~1 $/GB)
- … 2 days and 1K$
- … 3 years and 1M$

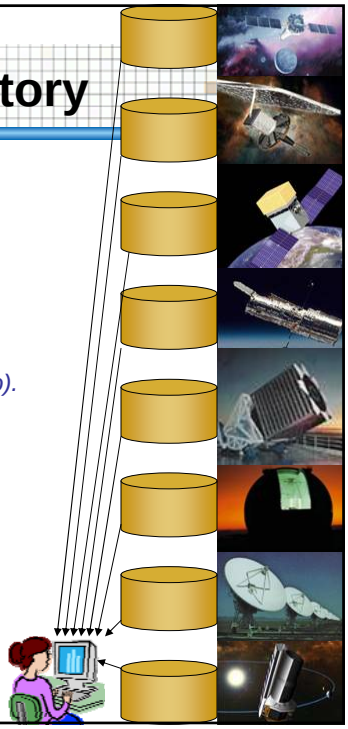
- At some point you need
  indices to limit search
  parallel data search and analysis
- This is where **databases** can help
- Bring the analysis to the data!!

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 $N\,logN$
- 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

The Virtual Observatory



- Premise: most data is (or could be online)
- The Internet is the world's best telescope:
  - *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..*

National Virtual Observatory

- NSF ITR project, “Building the Framework for the National Virtual Observatory” is a collaboration of 17 funded and 3 unfunded organizations
  - *Astronomy data centers*
  - *National observatories*
  - *Supercomputer centers*
  - *University departments*
  - *Computer science/information technology specialists*
- Similar projects now in 15 countries worldwide

=> International Virtual Observatory Alliance

Sloan Digital Sky Survey

Goal

Create the most detailed map of the Northern sky

“The Cosmic Genome Project”

Two surveys in one

Photometric survey in 5 bands

Spectroscopic redshift survey

Automated data reduction

150 man-years of development

High data volume

40 TB of raw data

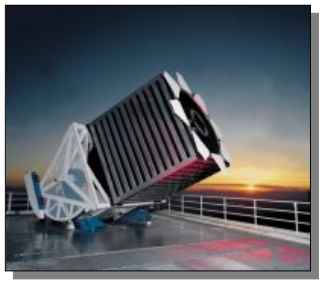
5 TB processed catalogs

Data is public

2.5 Terapixels of images

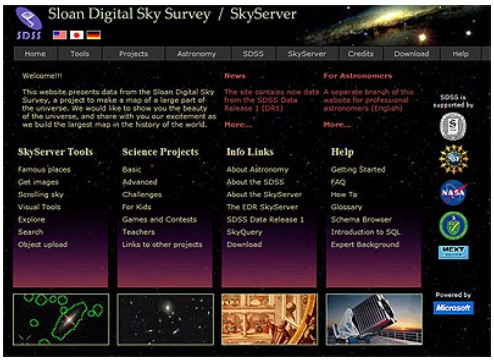
The University of Chicago
Princeton University
The Johns Hopkins University
The University of Washington
New Mexico State University
Fermi National Accelerator Laboratory
US Naval Observatory
The Japanese Participation Group
The Institute for Advanced Study
Max Planck Inst, Heidelberg

Sloan Foundation, NSF, DOE, NASA



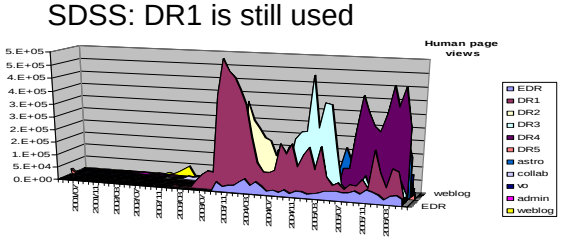
The SkyServer Portal

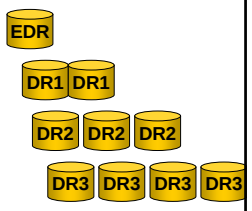
- Sloan Digital Sky Survey: Pixels + Objects
- About 500 attributes per “object”, 400M objects
- Currently 2.4TB fully public, June 29-> 3TB
- Prototype eScience lab (800 users)
  - *Moving analysis to the data*
  - *Fast searches: color, spatial*
- Visual tools
  - *Join pixels with objects*
- Tutorials and projects
- **Prototype in data publishing**
  - *50B rows of data delivered*
  - *250 million web hits in 5 years*
  - *930,000 distinct users*
  - *50K hours of classroom lessons*
- http://skyserver.sdss.org/



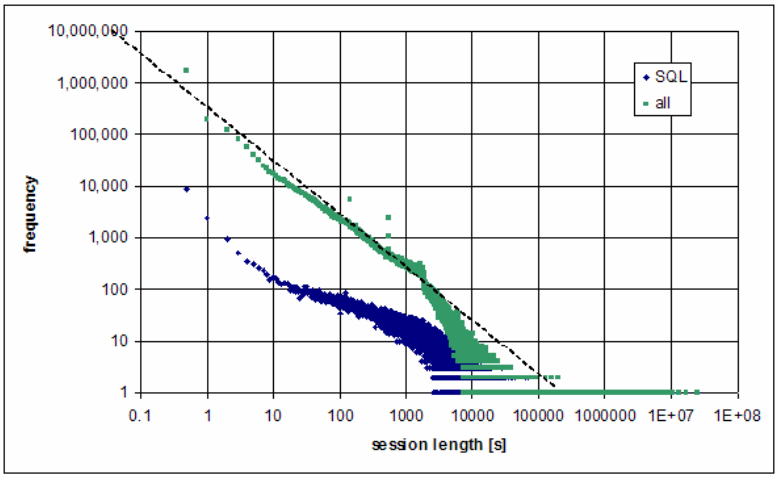
Data Versions

- June 2001: EDR
- Now at DR5, with 2.4TB
- 3 versions of the data
  - *Target, Best, Runs*
  - *Total catalog volume 5TB*
- Data publishing: once published, must stay
- SDSS: DR1 is still used





Skyserver Sessions



Vic Singh (Stanford/ MSR)

Trends

CMB Surveys (pixels)

• 1990	COBE	1000
• 2000	Boomerang	10,000
• 2002	CBI	50,000
• 2003	WMAP	1 Million
• 2008	Planck	10 Million

Angular Galaxy Surveys (obj)

• 1970	Lick	1M
• 1990	APM	2M
• 2005	SDSS	200M
• 2008	VISTA	1000M
• 2012	LSST	3000M

Time Domain

- QUEST
- SDSS Extension survey
- Dark Energy Camera
- PanStarrs
- SNAP…
- LSST…

Galaxy Redshift Surveys (obj)

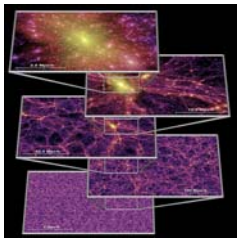
• 1986	CfA	3500
• 1996	LCRS	23000
• 2003	2dF	250000
• 2005	SDSS	750000

Petabytes/year by the end of the decade…

Simulations

Cosmological simulations have 10^9 particles and produce over 30TB of data (Millennium)

- Build up dark matter halos
- Track merging history of halos
- Use it to assign star formation history
- Combination with spectral synthesis
- Realistic distribution of galaxy types



- Hard to analyze the data afterwards- >need DB
- What is the best way to compare to real data?

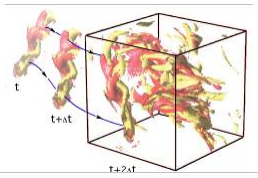
Exploration of Turbulence

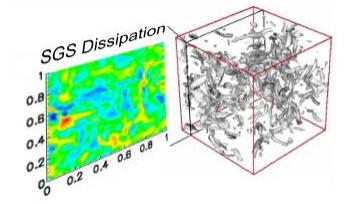
We can finally "put it all together"

- *Large scale range, scale-ratio O(1,000)*
- *Three-dimensional in space*
- *Time-evolution and Lagrangian approach (follow the flow)*

Unique turbulence database

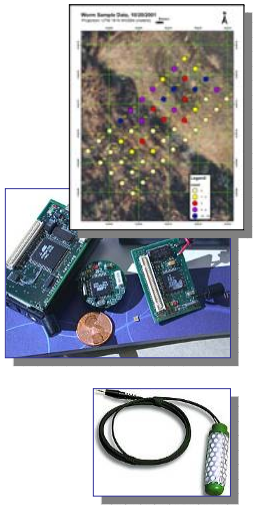
- *We are creating a database of O(2,000) consecutive snapshots of a 1,024³ simulation of turbulence:*
  close to 100 Terabytes
- *Treat it as an experiment*





Wireless Sensor Networks

- Will use 200 wireless (Intel) computers, with 10 sensors each, monitoring
  - Air temperature, moisture
  - Soil temperature, moisture, at least in two depths (5cm, 20 cm)
  - Light (intensity, composition)
  - Gases (O_2, CO_2, CH_4, …)
- Long-term continuous data
- Small (hidden) and affordable (many)
- Less disturbance
- >200 million measurements/year
- Collaboration with Microsoft
- Complex database of sensor data and samples



With K.Szlavecz and A. Terzis
http://lifeunderyourfeet.org

Data Sharing/Publishing in the VO

- What is the business model (reward/career benefit)?
- Three tiers (power law!!!)
  - *(a) big projects*
  - *(b) value added, refereed products*
  - *(c) ad-hoc data, on-line sensors, images, outreach info*
- We have largely done (a), mandated
- Need "Journal for Data" to solve (b)
- Need "VO Flickr" (a simple interface) for (c)
- Mashups are emerging (GalaxyZoo)
- Need an integrated environment for '*virtual excursions*' for education (C. Wong)

‘Journal for Data’ Experiment

S. Choudhury, T. deLauro, R. Hanisch, E. Vishniac, A. Szalay, M. Kurtz, C. Lagoze

Team up with the main existing journals in US astronomy and create an on-line supplement for the data related to journal articles

- Easy submission process for authors
- Properly linked to the journals, mostly in electronic version
- Data geoplexed among university libraries, with automated replication
- Data guaranteed to exist for 20 years
- Curation, curation, curation!!! (P. Buenemann)

Continuing Growth

How long does the data growth continue?

- High end always linear
- Exponential comes from technology + economics, **rapidly changing generations!**
  - *like CCD’s replacing plates… gene-chips*
- How many new generations of instruments do we have left?
- Are there new growth areas emerging?
- **Software (collaboration) is becoming an instrument**
  - *hierarchical data replication*
  - *Value added data/ mashups*
  - *data cloning*

Collaborative Trends

- Science is aggregating into ever larger projects
- VO is inevitable, a new way of doing science
- Present on every physical scale today, not just astronomy (Earth/Oceans, Biology, MS, HEP)
- But: there is a natural size for close collaborations
- May be the only way to do **'small science'** in 2020

Scholarly Communications

- No ‘Einstein letters’ today… very little paper trail
- Proposals and papers archived
- Most large projects communicate through email exploders and phonecons
- Often reaching back to the Internet Archive
- Some technical info on WIKI pages
- Science oriented blogs are appearing
- Collaborative workbenches emerging
- More instant messaging, especially next generation
- What can we and what should we capture?
- What will science historians do in 50 years?

Technology+Sociology+Economics

- Technology is changing very rapidly
  - *Sensors, Moore's Law*
  - *Trend driven by changing generations of technologies*
  - *There may not be time for a top-down design*
- Sociology is changing in unpredictable ways
  - *YouTube… vs Google and Yahoo*
  - *In general, people will use a new technology if it is*
    - Offers something entirely new
    - Or substantially cheaper
    - Or substantially simpler
  - *Build it and they (may) come…*
- Funding is essentially level

The Future of the VO

- Technology driving Sociology (limited by Economics)
- We need to keep running forward just to keep up
- We must take risks
  - *We will not get it always right*
  - *If we are right all the time, we are not taking enough risks*
- Surprisingly significant public involvement!
- Everything is a power law!

Summary

- Data growing exponentially, Petabytes/year by 2010
- Explosion is coming from inexpensive sensors and value added data products
- Requires a new model for science
  - *Having more data makes it harder to extract knowledge*
- Same thing happening in all sciences
  - *High energy physics, genomics, cancer research, medical imaging, oceanography, remote sensing, …*
- Science with so much data requires a new paradigm
  - *Computational methods, algorithmic thinking will come just as naturally as mathematics today*
- ***eScience***: an emerging new branch of science