

Data Intensive Computing for Astronomy

Alex Szalay
The Johns Hopkins University

Common VO Challenges

- Hard to find data (yellow pages/repository)
- Threshold for publishing is currently too high
- Even 19th century data is interesting (Harvard plates…)
- Low level format data is standardized (FITS)
- Scientists want calibrated data with occasional access to raw data
- High level data models take a long time…
- Robust applications are hard to build (factor of 3…)
- Geospatial everywhere, but GIS is not good enough
- Archives on all scales, all over the world
- VOSpace – distributed user repository services

Data Sharing/Publishing

- 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
- We have largely done (a), mandated by NSF/NASA
- Need "Journal for Data" to solve (b)
- Need "VO-Flickr" (a simple interface) for (c)
  - Astrometry.net is starting to do this (Hogg et al)
- New public interfaces to astronomy data (Google Sky, WWT)
- Mashups are emerging (GalaxyZoo)

Continuing Growth

How long does the data growth continue?

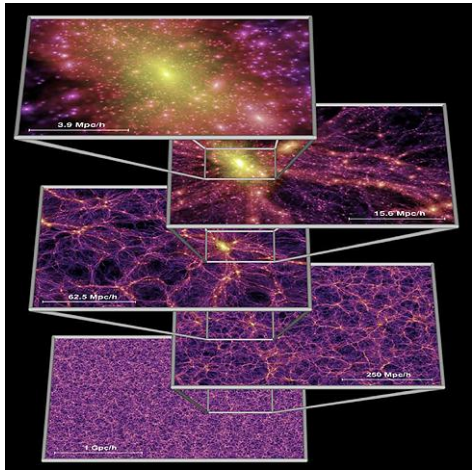
- High end always linear
- Exponential comes from technology + economics
  - ⇔ rapidly changing generations
    - Like CCD's replacing plates, and becoming cheaper
- How many generations of instruments do we have left?
- Are there new growth areas emerging?

Software is becoming a new instrument

- Simulations!!
- hierarchical data replications
- Value added data/ mashups

Cosmological Simulations

- 1B particles, produce over 30TB of data (Millennium, Aquarius, Via Lactea-II)
  - 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



- Too few realizations
- Hard to analyze the data afterwards ->need DB (Lemson)
- What is the best way to compare to real data?
- Data volumes soon reaching Petabytes -> "laboratories"
  - "Silver River" (Madau et al) 500TB+
  - LANL doing 1 trillion particle simulation right now (Habib)

VO Technology

- Next surveys/simulations will generate Petabytes
- We will need to save them, move them
  - Several big archive centers connected
  - Shakeout coming
- Archives – also computational services
  - Driven by economics: cheaper to process than move
- Must be an open-ended modular system
- Current mainstream computing architectures are inadequate and cannot scale much farther
  - Move to the cloud? PB containers? Low power bricks?

VO Economics

- The Price of Software
  - 30% from SDSS, 50% for LSST
  - should there be full reuse vs no reuse today?
  - neither: we are not systems integrators
  - risks and benefits are power law
- The Price of Data
  - $100,000 /paper (Norris etal)
  - Drives new projects
    - SDSS: there are 2,000+ refereed papers for $100M
- Soon will hit “Power Wall”
- Level budgets…

VO Sociology

- Learn from particle physics
  - Do not take for granted that there will be a next
  - Small is beautiful
- What happens to the rest of astronomy after the "world's biggest telescope"?
- The impact of power laws
  - We need to look at problems in octaves
  - The astronomers may be the tail of our users
  - There is never a natural end or edge
    (except for our funding)
- Unpredictable changes, new players (Google, MS)

Collaborative Trends

- Science is aggregating into ever larger projects
- Collection of data **increasingly separated** from subsequent analysis
- Connection is through the data archives
- Natural size for close collaborations is small
- May be the only way to do **'small science'** in 2020

The VO is inevitable

- It is a disruptive technology
- It is a new way of doing science
- Present on every physical scale today (VAO, LHC, Human Genome, NEON, EOS, …)

Near Future

- Surveys' role increasing, more archival data
- Relatively easy to predict until 2012
  - Exponential growth continues
  - Most ground based observatories join the VO
  - More sky surveys in different wavebands
- Dominance of Large Imaging Surveys
  - Fastest explosion of data in radio
  - Need large wide field spectroscopy survey!
- Simulations will reach petabytes
  - Will have VO interfaces: can be 'observed'

Beyond 2012

- PetaSurveys are coming on line and becoming public (Pan-STARRS, VISTA, LSST, ALMA, SKA, …)
- Petabytes will need hierarchical organization
  - Need a proper "impedance match", apply 90-10 rule
- Single Query analysis paradigm will break
- World-wide network of large archive/compute centers
- Business model still unclear: public data does not necessarily mean accessible data…
- Moore's Law comes to the rescue (up to a point)
- Changing funding climate, unpredictable

Astronomy with petabytes is unavoidable!