Astro-Informatics: Computation in the study of the Universe

Bob Mann *and Andy Lawrence*

Institute for Astronomy, School of Physics

(rgm@roe.ac.uk & al@roe.ac.uk)

Plan

- Computational Astrophysics
  - N-body simulations of galaxy clustering
- Astro-Informatics
  - Survey astronomy & the Virtual Observatory
- Discussion
  - Astronomy and informatics

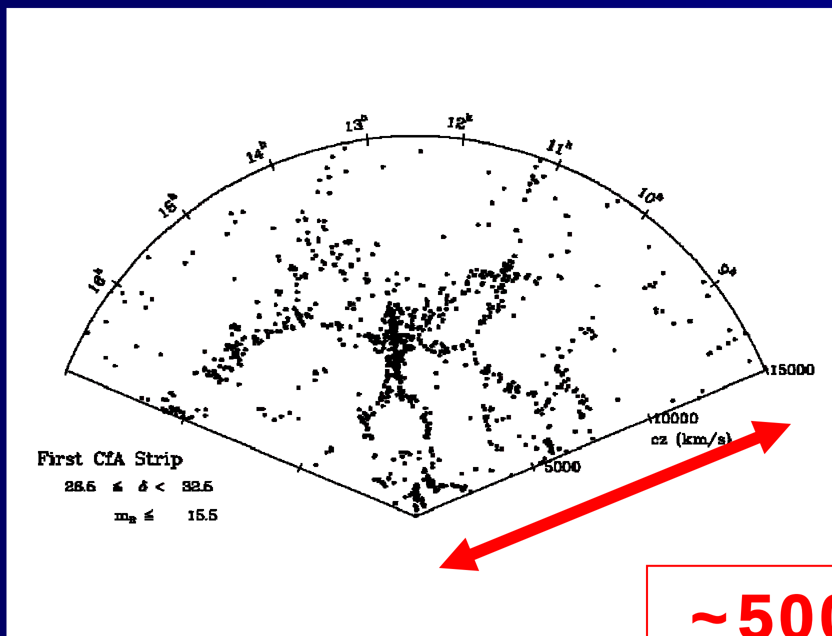
Plan

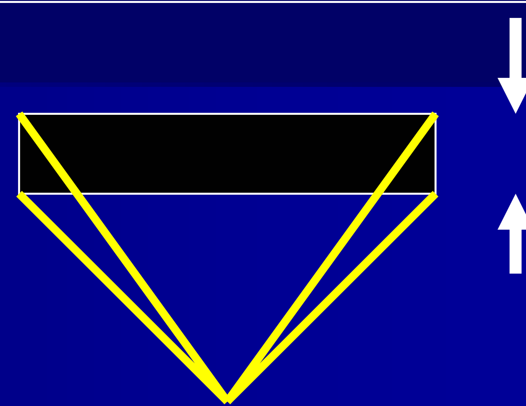
- Computational Astrophysics
  - N-body simulations of galaxy clustering
- Astro-Informatics
  - Survey astronomy & the Virtual Observatory
- Discussion
  - Astronomy and informatics

Observing galaxy clustering

- 1930s: Hubble
  - Galaxies aren’t uniformly distributed on sky
- 1950s: Shane and Wirtanen
  - Map of galaxy distribution on the sky from counting 100,000 galaxies <u>by eye</u> (10 years!)
- 1980s: CfA Redshift Survey
  - (Huchra, Geller, de Lapparent)
  - First sizeable 3D map of the local Universe
    - Measured rough distances to ~11,000 galaxies

1985: first CfA survey





3D map of a pyramidal slice of space, projected into 2D

~500 million light years

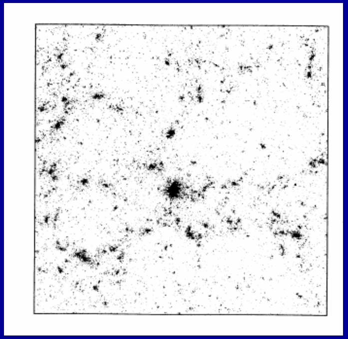
- Rich structure – walls, filaments, voids...
  - How to explain this richness of structure?

Modelling galaxy clustering

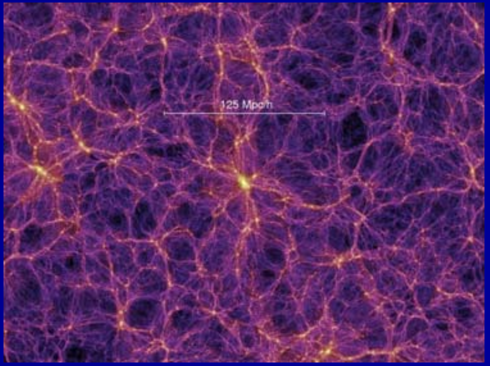
- Physics simple in Cold Dark Matter model
  - Collisionless material moving under gravity
- Apply perturbation theory to density field
  - Linear theory treatment simple, but...
  - Perturbations non-linear on scales of interest
    - Fourier modes couple, analytic methods fail
- Need numerical simulations to model galaxy clustering into non-linear regime
  - Set up test masses and evolve under gravity: i.e. gravitational *N*-body simulations

Two decades of N-body simulations

- 1985: Davis, Efstathiou, Frenk, White
  - $(32)^3$ particles
  - <10 particles per galaxy
  - Early success for Cold Dark Matter model



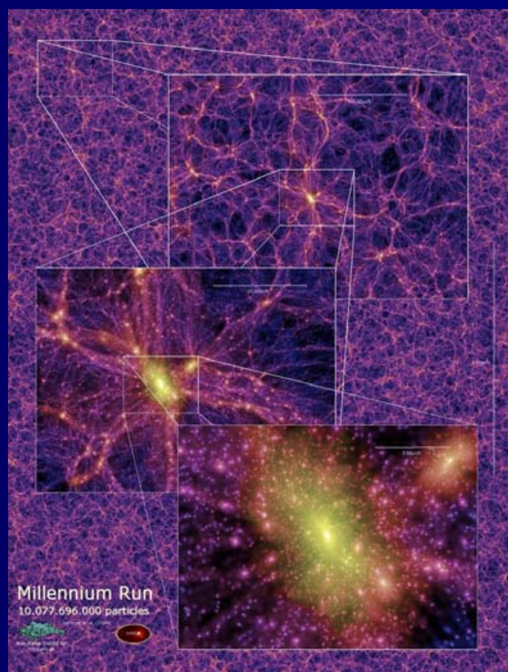
- 2005: Virgo Consortium
  - Inc. John Peacock (IfA), plus EPCC
  - $(2202)^3$ particles
  - ~1000 particles per galaxy



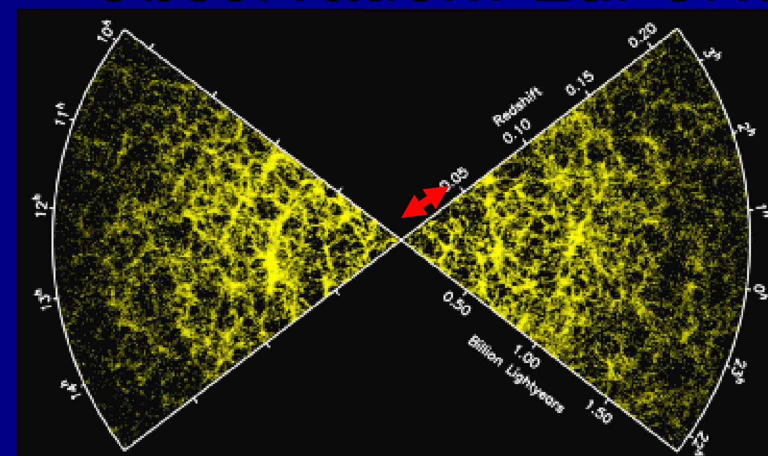
- Mass resolution increased by a factor of $\sim 10^2$ and simulation volume by a factor of $\sim 10^3$

Theory *v* Observation

- Theory: VIRGO



- Observation: 2dFGRS



(inc. John Peacock)

~250,000 galaxies

- (SDSS: ~500,000 galaxies)

Quantitative clustering analysis reveals theory and observation in <u>excellent</u> agreement

Galaxy clustering summary

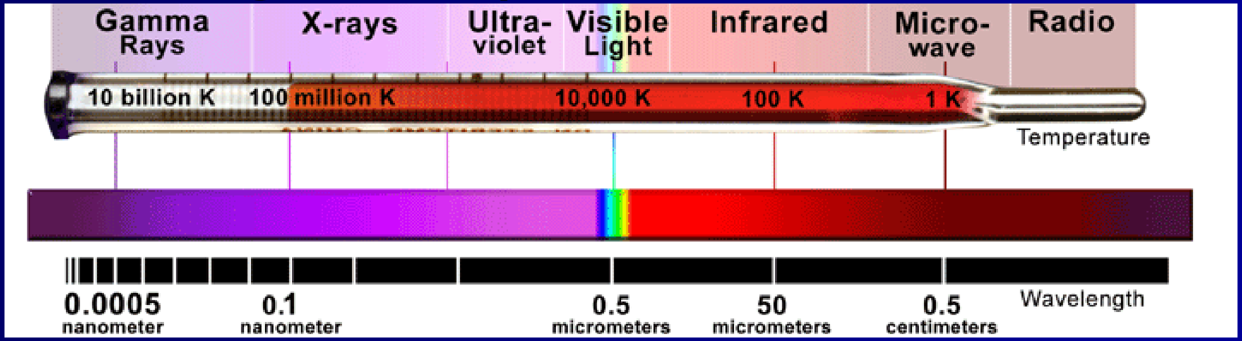
- Cold Dark Matter model accounts for the observed clustering of galaxies
  - Major triumph of modern astronomy
- Numerical simulations crucial, but this is astronomers using computers, not astronomers using computer science
  - Are there examples of real interaction between astronomy & computer science?
    - More interesting than just number-crunching?

Plan

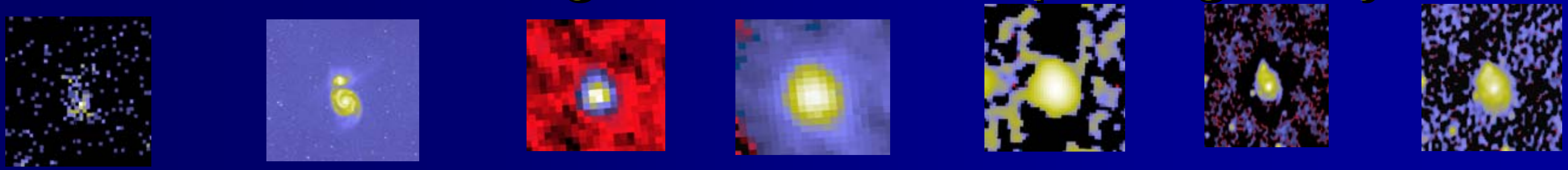
- Computational Astrophysics
  - N-body simulations of galaxy clustering
- Astro-Informatics
  - Survey astronomy & the Virtual Observatory
- Discussion
  - Astronomy and informatics

Observational Astronomy

- Electromagnetic spectrum



- Multiwavelength view of a spiral galaxy



ROSAT ~keV *DSS Optical* *IRAS 25μ* *IRAS 100μ* *GB 6cm* *NVSS 20cm* *WENSS 92cm*

  – Different angular resolution of instruments
  – Different physical emission mechanisms

(M51 graphics from Jim Gray & Alex Szalay)

Changes in the way that we make observations

- Old Style: *Many small, specific programmes*

Astronomer proposes observations, goes to telescope, brings data home on tape, analyses data, publishes paper, puts tape in desk drawer and forgets about it

- New Style: *Few large, multi-use surveys*

  – Consortium designs survey to address many science goals, undertakes survey over several years, establishes database – *many* people do *different* science with *same* data from DB

Trends behind these changes

- Instruments made easier to use & more effort put into data reduction software
  - Easier to use data from new instrument
  - Multiwavelength astronomy much easier
- Instruments are more sensitive and have more detector elements
  - Can image large areas of sky quickly
  - Survey mode of observation more efficient

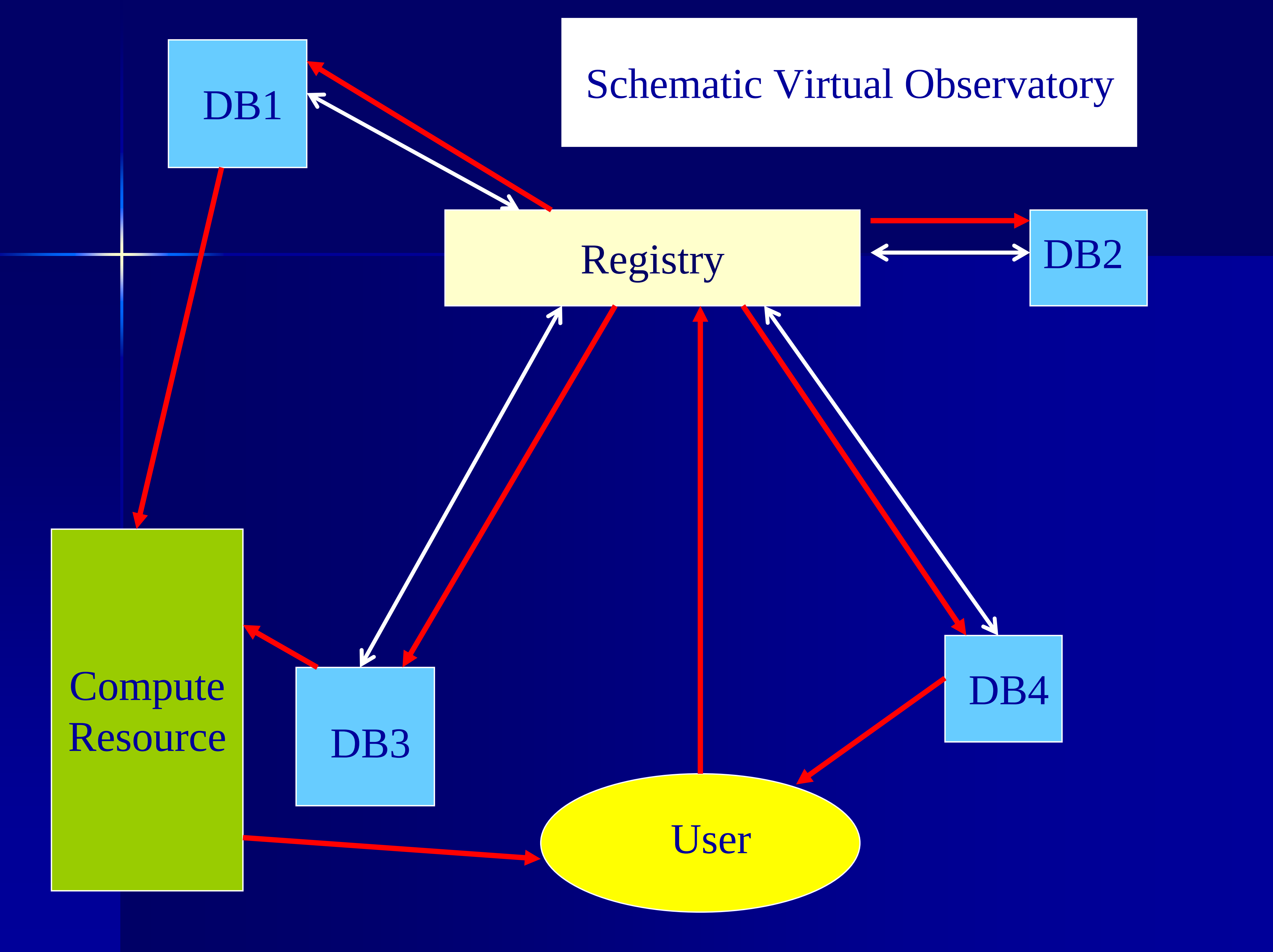
Very strong local interest

- Wide Field Astronomy Unit
  - – Part of the UoE Institute for Astronomy
  - – Based at Royal Observatory Edinburgh, on Blackford Hill

- Two strands to WFAU work
  - – Curation of optical/near-infrared sky surveys
  - – Helping build the global “Virtual Observatory”

The Virtual Observatory

- Goals
  - – Federate all the world's astronomy data
  - – Provide resources for exploitation of data
- Challenges – sociological & technical
  - – Heterogeneous, distributed datasets
    - Lack of global schema; metadata often poor
  - – Legacy analysis codes in many languages
- Solution
  - – International collaboration
  - – Architecture built on web services



Schematic Virtual Observatory
DB1
Registry
DB2
DB3
DB4
Compute Resource
User

WFAU's computational problems

- **Quality Control**
- Spatial Indexing
- Analysis close to DB

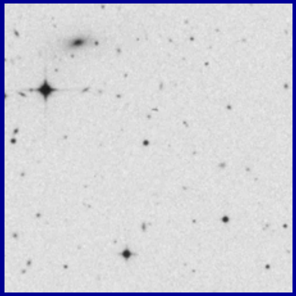
- Provenance

Individual sky survey archives: *scale*

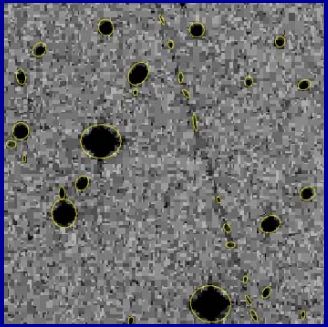
- Lack of Global Schema
- Query Language
- **Difficulty in Making Joins**
- Integration with the Literature

Virtual Observatory: *interoperability*

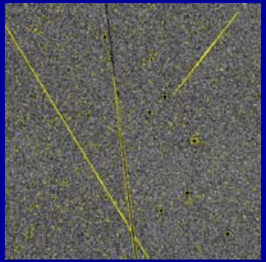
Quality control: automated junk detection

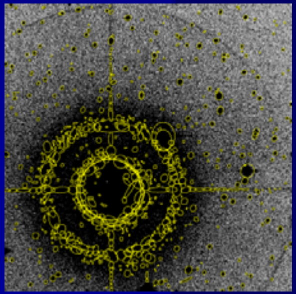


- SuperCOSMOS Sky Survey
  - – Scans of photographic plates
  - – ~1800 plates cover whole sky
  - – Image analyser run over images
    - ~250,000 sources per plate



- Classes of spurious source
  - – Trails: satellites, aeroplanes,...
  - – Diffraction effects around bright stars

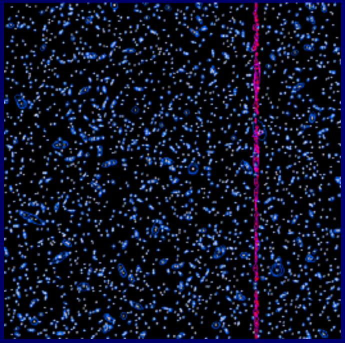


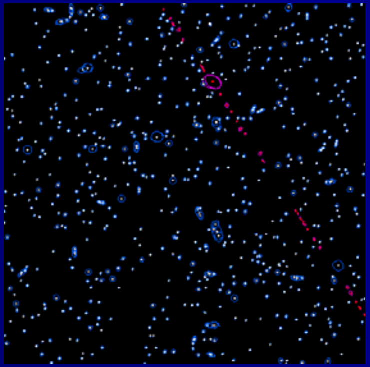


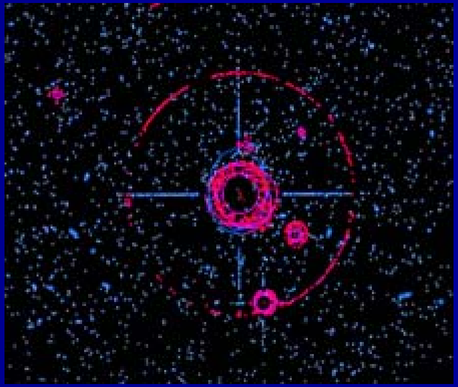
- How to find these spurious sources?

Quality control: automated junk detection (2)

- Junk found in unusual configurations
  - – Lines, circles: the eye spots them easily – but can't eyeball thousands of plates!
- **Amos Storkey**, Chris Williams, Nigel Hambly
  - – Developed new generative method, based on unlikeliness of configurations





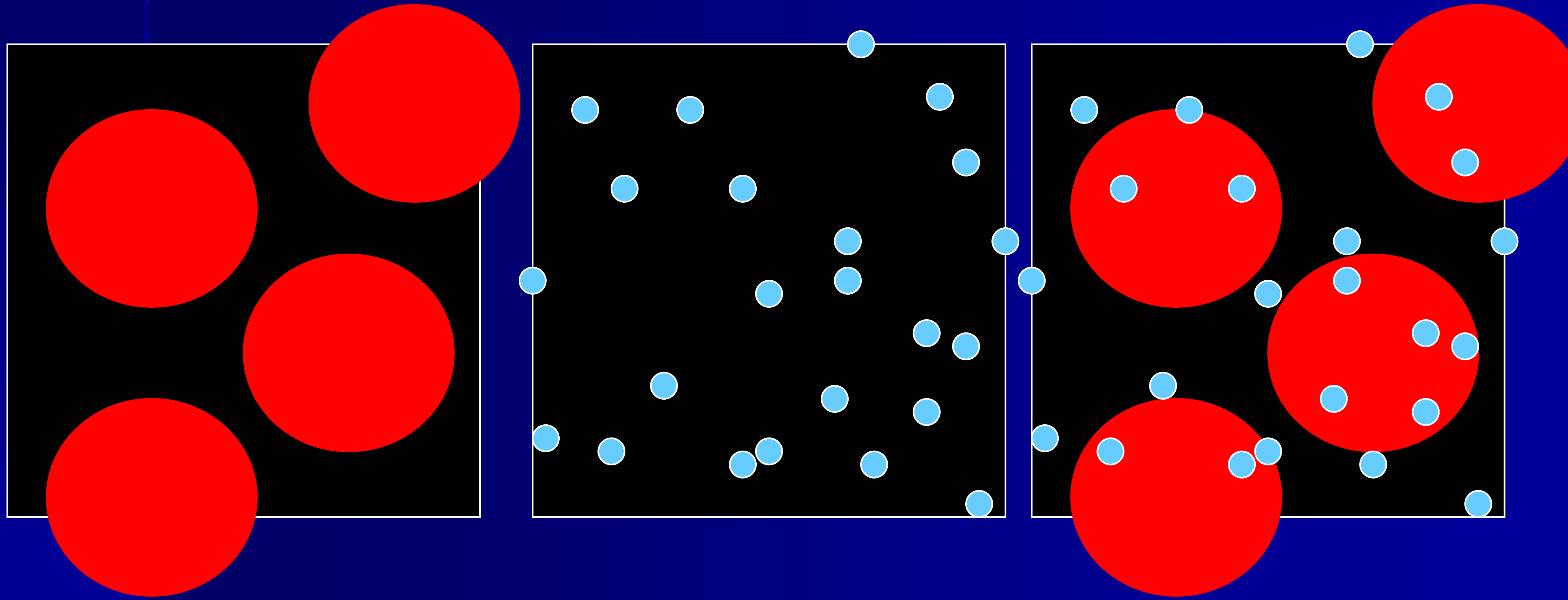


Analysing sky survey data

- WFAU has multi-TB sky survey databases
- Many analyses will use much of the data
  - e.g. finding one-in-a-million unusual objects
  - e.g. quantifying properties of populations
- Users can’t download data to workstation
  - WFAU must provide analysis services on DB
- Security issues if users upload their code
  - Application of mobile code security work? – discussion started with Don Sannella’s group

Difficulty of matching entries between sky survey databases

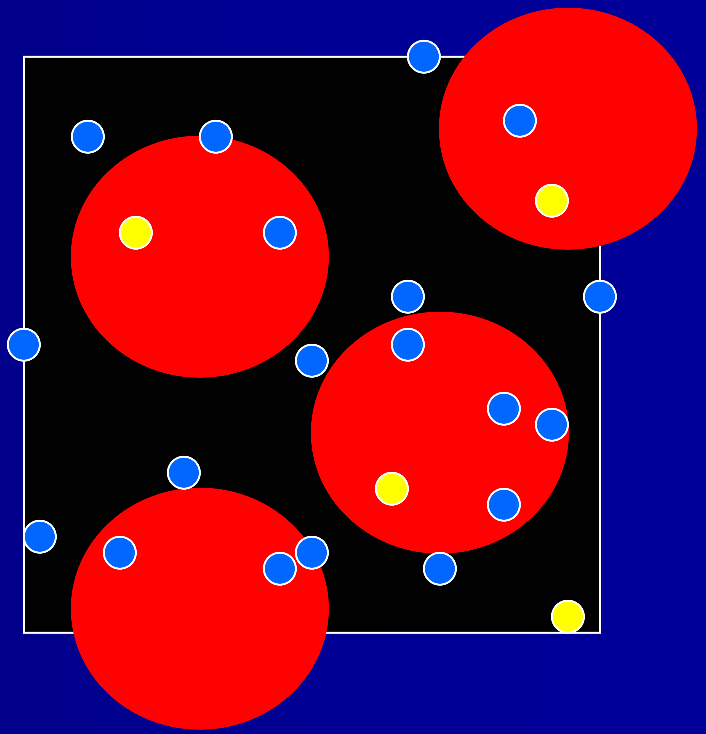
- Angular resolution varies between datasets



- Matching by spatial proximity is inadequate

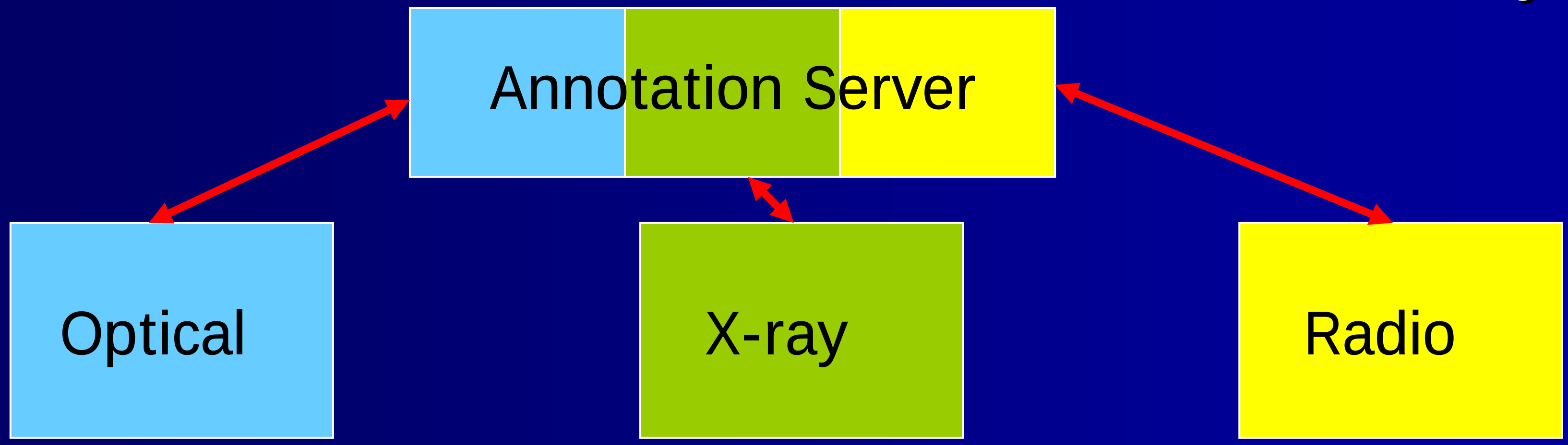
Difficulty of matching entries between sky survey databases (2)

- Probabilistic framework well established
  - – But need to know properties of source populations
    - Often not the case
- Learn the probabilities for matching different classes of source iteratively (EM algorithm)
- Emma Taylor (PhD), with Amos Storkey & Chris Williams



Difficulty of matching entries between sky survey databases (3)

- Sophisticated matching algorithms are often computationally expensive
  - Want to cache matches for re-use
- AstroDAS: Diego Prina Ricotti, Raj Bose
  - Distributed annotation server for astronomy



Integrating the online literature into the VO

- If we find an interesting object, we frequently want to ask questions like:
  - What's known about this area of sky?
  - What's known about objects like this?
  - Have objects like this been reported before?
- Literature is too large to search manually
  - Can text mining techniques help?

Integrating the online literature into the VO (2)

- AstroNER: Named Entity Recognition
  - Claire Grover, Ben Hachey et al.
- Look at abstracts of journal articles related to spectroscopy of active galaxies
- Try to identify nouns of four types
  - instrument-name, spectral-feature, source-type, source-name
- Apply various techniques, using training data annotated by astro PhD students

Annotations - Mozilla

HST and Chandra Observations of Quasar PHL 1811

PHL 1811 is a nearby , luminous (z = 0.192 ; M _ { V = - 25.9 }) quasar . With magnitudes of B = 13.9 and R = 13.9 , it is the second brightest quasar known with z > 0.1 after 3C 273 . Optically it is classified as a *Narrow - line* Seyfert 1 galaxy (NLS1) , a class generally known to be bright in soft X - rays . Thus , it was surprising that PHL 1811 was not detected in the ROSAT All Sky Survey . A follow - up BeppoSAX observation detected the quasar , but revealed it to be anomalously X - ray weak . The inferred α _ { ox } was 1.9 -- 2.1 , much steeper than the nominal value of 1.6 for quasars of this optical luminosity , and comparable to the X - ray weakest quasars . To investigate the cause of the X - ray deficiency , coordinated HST UV spectra and Chandra observations were obtained in December 2001 . Two Chandra pointings , 9.4 and 9.8 ks in length and separated by 12 days , netted 84 and 338 photons respectively . The X - ray spectra , fitted jointly by a power law with Galactic absorption , yield a photon index of 2.09 +/- 0.14 . The flux varied by a factor of 4 between the two observations . The lack of intrinsic absorption and the strong variability are interpreted as evidence that we observe the central engine directly and unobscured . The HST STIS spectra , taken two days before the first Chandra observation , reveal a very blue continuum with little evidence for absorption or scattering intrinsic to the quasar . The inferred α _ { ox } for the two Chandra observations are 2.13 and 2.36 , respectively . We conclude from these observations that PHL 1811 is intrinsically X - ray weak . The UV and optical emission - line spectra of PHL 1811 are remarkable . Neither forbidden nor semiforbidden emission lines are detected . \ion { Fe } { 2 } is the dominant line emission in the UV . High metallicity is implied by the large \ion { Fe } { 2 } to \ion { Mg } { 2 } ratio and relatively strong \ion { N } { 5 } . Low - ionization emission lines of \ion { Al } { 3 } , Na I D , and Ca II H & K are present , implying high optical depth . High - ionization lines are very weak ; \ion { C } { 4 } has an equivalent width of only ~ 5 Å / . The spectrum bears marked resemblance to `` line - less '' high - redshift quasars discovered in the SDSS .

Key

Instrument-name Spectral-feature Source-type Source-name

embedded Spectral-feature *embedded Source-type*

Done

Plan

- Computational Astrophysics
  - N-body simulations of galaxy clustering
- Astro-Informatics
  - Survey astronomy & the Virtual Observatory
- Discussion
  - Astronomy and informatics

Two classes of research

- Computational Astrophysics
  - Astronomers using computers to solve a specific problem in astrophysics
- Astro-Informatics
  - Astronomers and computer scientists collaborating in the application of computational techniques to astronomy

c.f. distinction made by Jim Gray (Microsoft)

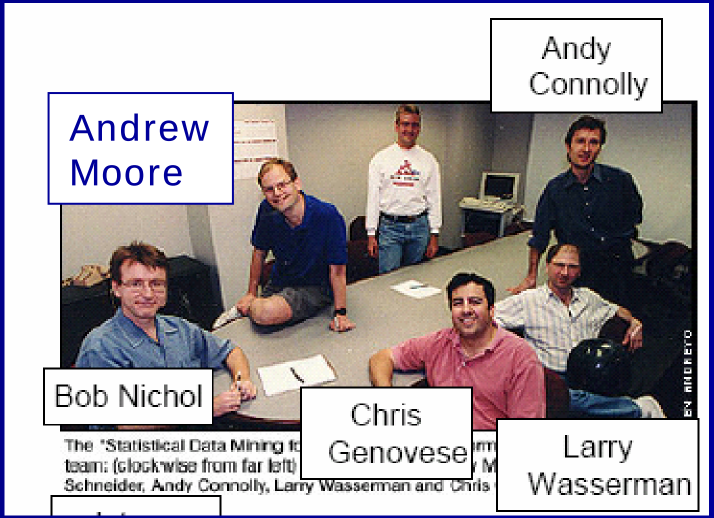
- Comp-X
  - X-ologists using computers to solve a specific problem in X-ology
- X-Info
  - X-ologists and computer scientists collaborating in the application of computational techniques to X-ology

Comp-X & X-info compared

- Comp-X
  - Involves only X-ologists
  - Should be funded as X-ology research
- X-Info
  - Requires X-ologists and computer scientists
    - How should this be funded? Can both sides be kept happy?
- Comp-X/X-Info boundary is domain-specific
  - Particle physics is almost all Comp-X
  - Biology is mainly X-info – bioinformatics
  - Astronomy is a mixture of both

Can X-info work?

- Example of successful X-info: PiCA group – Pittsburgh Computational Astrostatistics Group
  - Sustained collaboration: 1999 onwards
  - Astronomy, CS and statistics expertise
  - Focus on scalable data mining algorithms



- Astro requirements drive research in both statistics and CS

Can X-info work here?

- It is!…to some extent
  - as this lecture series illustrates
  - I've described several astro-info projects
- How can we do X-info better?
- Sustained interactions...
  - Understand areas of mutual interest
  - Give-and-take over individual projects
- ..which require funding
  - e.g. cross-School PhD studentships

Summary & Conclusions

- Astronomy relies on computation
  - On both theoretical and observational sides
  - In both Comp-X and X-info modes
- Astronomy is a good “X” for X-info
  - Data: free, voluminous, no ethical issues
  - Needs storing, indexing, describing, mining...
- Challenge: how to make X-info work well
  - Huge rewards for {X} and informatics