The aims of SC4DEVO and SC4DEVO-1

Bob Mann

Institute for Astronomy and
National e-Science Centre,
University of Edinburgh

Outline

- Background - SC, DE & VO
- The SC4DEVO project
- The SC4DEVO-1 workshop

Background - SC, DE & VO

SC
4
DE
VO

Service **C**omposition

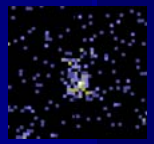
for

Data **E**xploration in the

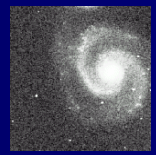
Virtual **O**bservatory

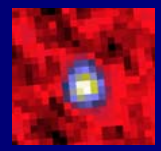
VO – Virtual Observatory

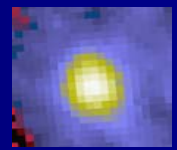
- Federation of astronomical data sources
  - Why?
- Images of M51 (courtesy Alex Szalay & Jim Gray)

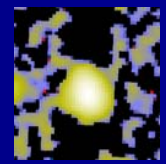

ROSAT ~keV

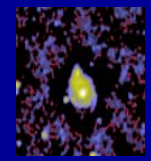

DSS Optical

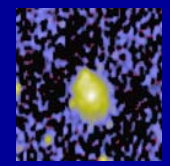

2MASS 2μ


IRAS 25μ


IRAS 100μ


GB 6cm


NVSS 20cm


WENSS 92cm

- Differences in:
  - Physical emission mechanism
  - Instrumental characteristics

VO – Virtual Observatory

- Starting with heterogeneous data sources
  - Aim for interoperable federation
  - International effort with a body to act as a standards agency and a coordinator

- Progress fairly good so far
  - Can expect to have some sort of interoperable data federation within next few years

VO…and DE

- What will be in the VO?
  - Dominated by ~10 sky survey databases
    - ~10^2 attributes in largest
    - ~10^8-10^9 entries in largest
    - ~ 1 Petabyte in total
- **The VO will be a large multivariate dataset of high dimensionality**
- *How do we do science with that?*

...need data exploration!

A VO DE Scenario

- A scientist has a hunch about connections between the properties of brightest cluster galaxies (BCGs) and those of their host clusters

- Query the VO and construct a sample of BCG/cluster pairs – say, 400 attributes for 10,000 pairs

A VO DE scenario (2)

- Run a stats package and find the 20 attributes with highest information content
- Plot a grid of scatter plots for pairs of these, arranged in order of strength of correlation between them
- See that, say, six attributes have strong correlations between them

A VO DE scenario (3)

- Select a representative sample of 200 clusters
- Step through visualizations of subspaces of the 6-D parameter space
- See, say, three clusters of points
- Assess statistical significant of these for all 10,000 BCG/cluster pairs
- Try and figure out what it all means

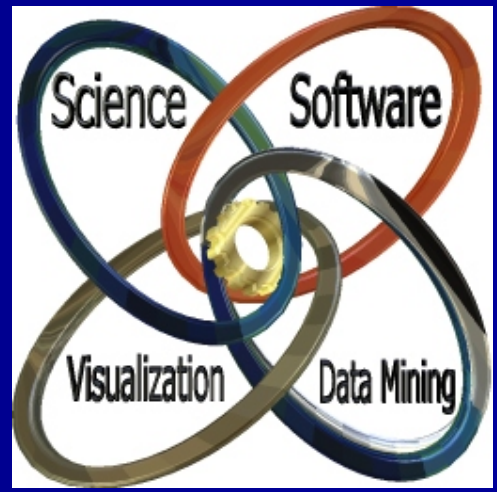
A VO DE scenario (4)

- Features of this scenario
  - Use of a number of different tools
  - Coupling of data mining & visualization
  - Some interactive steps
  - Maybe some iteration
- Our challenge is to work out how to do this!

DE – Data Exploration

- DE = Data Mining + Visualization
- The *coupling* of data mining and visualization is the key
- A route into the data
  - Finding significant patterns to follow-up
- Is this situation unique to astronomy?
  - No, e-science is driven by data avalanche

SDMIV Workshop



- *Scientific Data Mining, Integration and Visualization*
- Edinburgh, October 2002
- 50 participants
  - astronomy, atmospheric science, bioinformatics, chemistry, digital libraries, engineering, environmental science, experimental physics, marine sciences, oceanography, *plus* CS - data mining, visualization, Grid computing

http://www.nesc.ac.uk/talks/sdmiv/report.pdf

Lessons from SDMIV

- CS and Apps people want to interact
  - See mutual benefit from collaboration
- Common problems in all disciplines
  - Lots of distributed data in many formats
- Lots of DM and Vis software out there, but…
  - Doesn’t match how we work now
  - Don’t know what to use or where to find it
  - How does it fit into the computational infrastructure we’re building?...VO, Grid, etc

The “Marzipan Layer”
(© Malcolm Atkinson)

- Christmas Cake metaphor for web/Grid services stack in e-science
  - Fruit Cake = core services
    - Up to, and including, data integration
  - Icing = specific apps written by scientists
  - Marzipan = what goes in between
    - Some mix of truly generic and domain-specific stuff?
    - Wrapping apps, data structures, format conversion and…?
- The Marzipan Layer is a nice metaphor
  - Shows importance of SC for DE
  - But what does it mean in detail?

The SC4DEVO Project

The SC4DEVO Project

- UK e-Science Programme launches "International Sister Projects" initiative
  - Money for workshops first, staff later
- Aside: UK e-Science Programme
  - ~£200M over six years for a range of projects: domain-specific, technology-specific and supportive CS research
  - **A Good Thing** – boosting interaction between CS and domain scientists

SC4DEVO Proposal

- Aim:
  - Work out how to do VO data exploration
  - How to generalise to e-science
  - *What goes in the Marzipan Layer?*
- Anglo/Australian/US consortium
  - UK: AstroGrid Data Exploration Framework
  - Aus: CSIRO Grid Computing group
  - US: GRIST project, plus related VO people
  - *plus* DM, Vis, workflow, Grid researchers

SC4DEVO Workshops

- Plan four workshops in 2004 and 2005:
  - 2004: Edinburgh (Jan), Caltech (Jun)
  - 2005: Edinburgh (Jan), Canberra (Jun)
- Jun workshops focused on VO specifics, Jan ones accompanied by an SDMIV-like workshop covering more disciplines
- One of four successful applications
- Later start than expected

The SC4DEVO-1 Workshop

SC4DEVO-1 Goals

- Foster collaboration between VO people and CS researchers in SC and DE areas
  - Generate research agenda for future funding
- Overview of full SC4DEVO topic
  - Where are the gaps in our current thinking?
  - What are the pressing research topics?
  - What can we do during these workshops?
- Deliverable: workshop report and web page with all presentations

SC4DEVO-1 Format

- 3½ day workshop
  - 2 ½ days scheduled with talks
  - 1 day unscheduled – for discussion
- Talks: 45 minute slots
  - 30 mins material – allow discussion
- Identify topics for detailed discussion
  - Pick up on Thursday – in break-outs?

Summary

- VO needs DE within SC framework
  - So do many other disciplines
- Bring together multidisciplinary team
  - To look at SC4DEVO, then generalise
  - *What goes in the Marzipan Layer?*
- Foster collaboration with an eye to future funding...and have a good time!