Data access in Physics and Astronomy

- General points
- Three models
- The Virtual Observatory

General points

distinctions

- ownership vs re-use
- legal rights vs community practice
- data vs knowledge

standard scientific position :

knowledge	**public; universal**
data	**private to scientist until publication**

technical and policy infrastructure must be able to express both

the knowledge chain

- (1) raw data
  - uncalibrated images with instrument effects still in
- (2) data products
  - calibrated skymaps, object catalogues
- (3) physical properties
  - luminosity, metallicity etc of specific stars, galaxies
- (4) understanding
  - properties of galaxies in general, understanding of how they form

- money and effort input all the way along
- where is the public ownership line drawn ?
- if re-use is likely, availability of (2) is important

Three models

(1) Condensed Matter

- Modes
  - (1) Trad lab experiments
    (2) Big facility + beam time
- Small teams; data belongs to team
- Analyse, publish, then throw away data
  - ownership very sensitive until publication
- Under political pressure to have data access and re-use policies

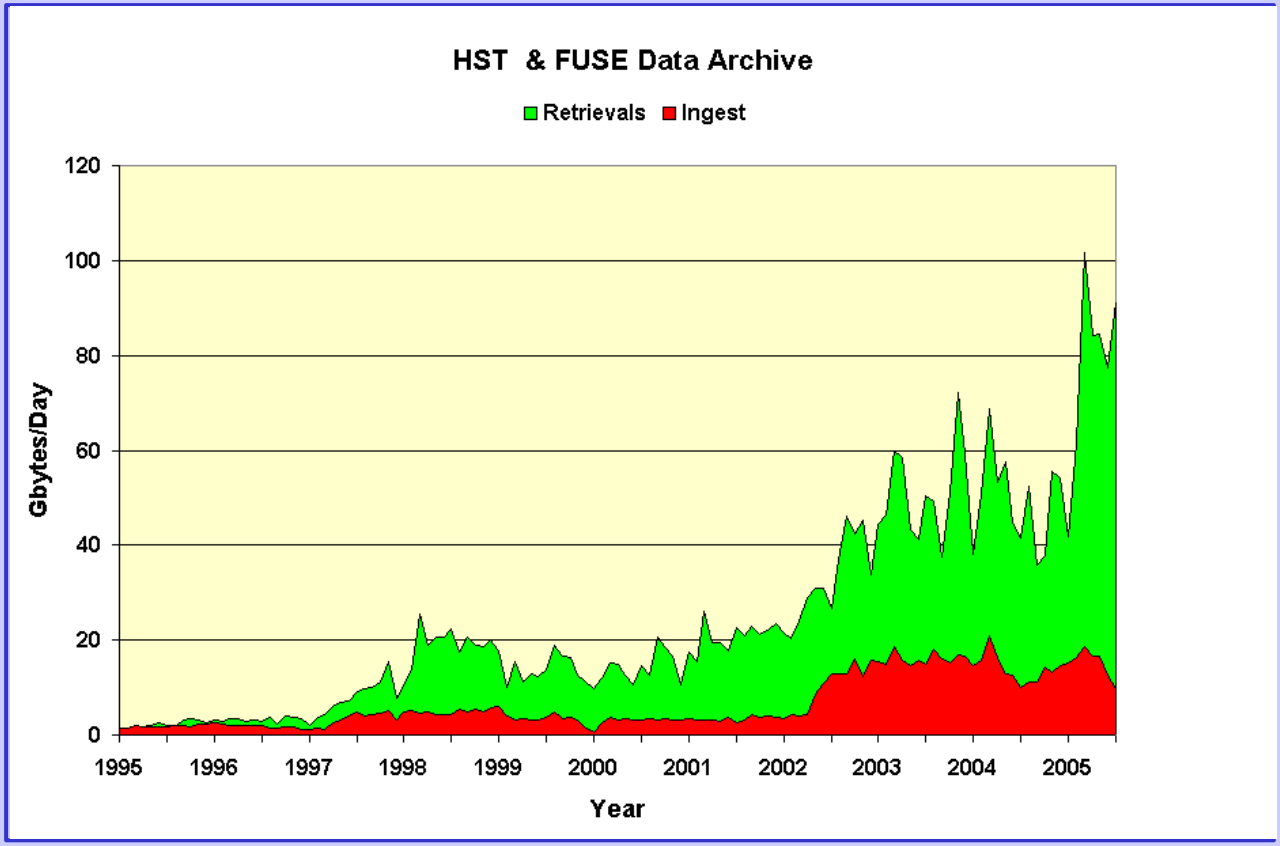
(2) Particle Physics

- Big facilities, big projects, big teams
  - thousands of scientists; PBytes/year; huge processing task
- Data belongs to project
  - tightly controlled until publication
  - elaborate rules inside project for access and use
  - well developed data infrastructure for their own use
- Assert that data re-use is pointless
  - need very detailed technical understanding of equipment
- Offering their data infrastructure for generic use

(3a) Small Astronomy

- Big facilities, small teams
  - competitive allocation ==> "My Three Nights"
- Trad is analyse, publish, throw away
  - ownership very sensitive until publication
- But facilities / data centres archive all data
  - re-use is plausible and increasingly common
  - standard policy : private for a year, then anybody can grab it
- Trend is for increasing standardisation of archiving, formats, access, analysis tools
  - Data infrastructure initiative : the Virtual Observatory
  - Archives should be "science ready"

data re-use in astronomy



(3b) Big Astronomy

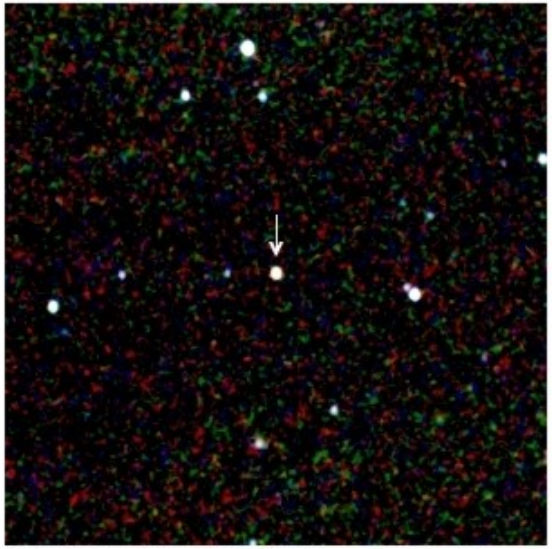
- Systematic Surveys
  - big teams; organised data collection, processing, and archiving
  - construct "science ready archives"
- Public from start
- Surveys make a digest of the sky
  - many different experiments performed by analysis of archive
  - the archive becomes the sky
- Increasing use of multiple surveys
  - they must inter-operate

multi-archive use

2MASS J1146+2230

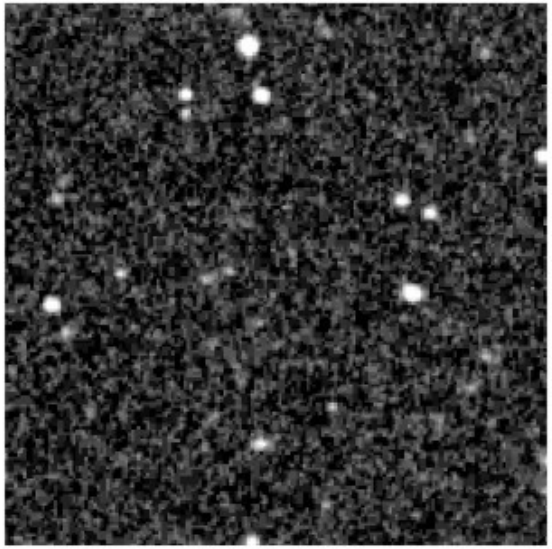
An L–type dwarf in the constellation Leo

The near–infrared view



2MASS Atlas JHK$_s$ Composite Image

The optical view



Palomar Digitized Sky Survey

J.D. Kirkpatrick (IPAC/Caltech), I.N. Reid (Caltech), R.M. Cutri (IPAC/Caltech), C.A. Beichman (IPAC/JPL/Caltech), J. Liebert (U of A), M.F. Skrutskie (UMass)

The 2MASS project is a collaboration between the University of Massachusetts and IPAC

science from cross-correlating archives in different locations

The Virtual Observatory

the VO concept

- web all docs in the world inside your PC
- VO all databases in the world inside your PC

VO framework

- agreed *standards*
- inter-operable *data collections*
- inter-operable *software modules*

- no central VO-command

- its not a thing
- its a way of life

VO geometry

- not a warehouse
- not a hierarchy
- not a peer-to-peer system

- small set of *service centres*
  and large population of *end users*

what is needed ?

- global standards
- well funded data centres
- infrastructure software
- VO aware data services
- VO aware software tools

standards
International VO Alliance

- formal process based on W3C
- key standards agreed or in development
  - formats
  - service metadata
  - data access protocols
  - column semantics
  - s/w interfaces
  - identity

Infrastructure

- Key software in place from major projects
  - Registry (yellow pages)
  - Virtual Storage (VOSpace)
  - Job Execution - workflow
  - API for tools (Astro Runtime)
  - Message protocol for tools (PLASTIC)
- Key next step
  - Identity/Authentication/Authorisation
  - Enables Single Sign On
  - Allows both public and private data
  - Allows arbitrary collaborations

services and tools

- data centres
  - growing number of data services
  - but data centres under-resourced ...
- VO portals
  - several effective one-stop-shops
  - SkyQuery, Aladin, NVO portal, AstroGrid workbench
- VO tools
  - some tools converted : Aladin, TopCat, Montage, Sextractor, Hyper-Z
  - several new tools : DataScope, Astroscope, VOSpec, VOPlot

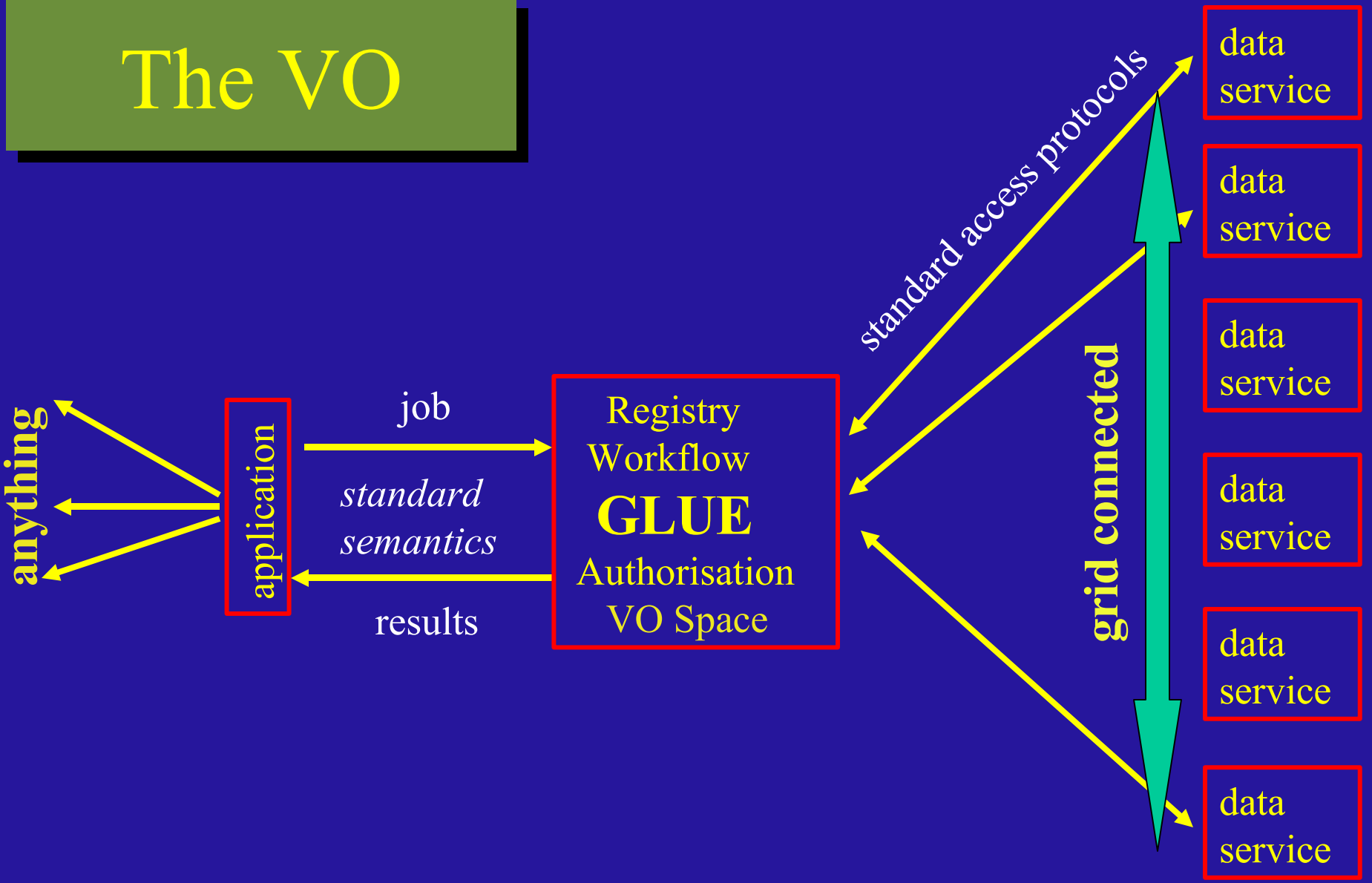
conclusions

conclusions

- Public ownership of knowledge is important
- But so is private ownership of data
  - science is a competitive social process
  - communities should set their own policies
- Intermediate area : science ready archives
  - this should be main focus for open access agenda
- Data infrastructure needed
  - need well funded data centres and a software infrastructure
  - background technology can be in common
  - but details community-driven
  - should understand all possible access rights
- Virtual Observatory initiative making excellent progress

deleted scenes

The VO


anything
application
job
standard
semantics
results
Registry
Workflow
GLUE
Authorisation
VO Space
standard access protocols
grid connected
data service
data service
data service
data service
data service
data service

publishing metaphor

- facilities are **authors**
- data centres are **publishers**
- VO portals are **shops**
- end-users are **readers**

- VO infrastructure is **distribution system**.