



NGS

National Grid Service



Data and storage services on the NGS

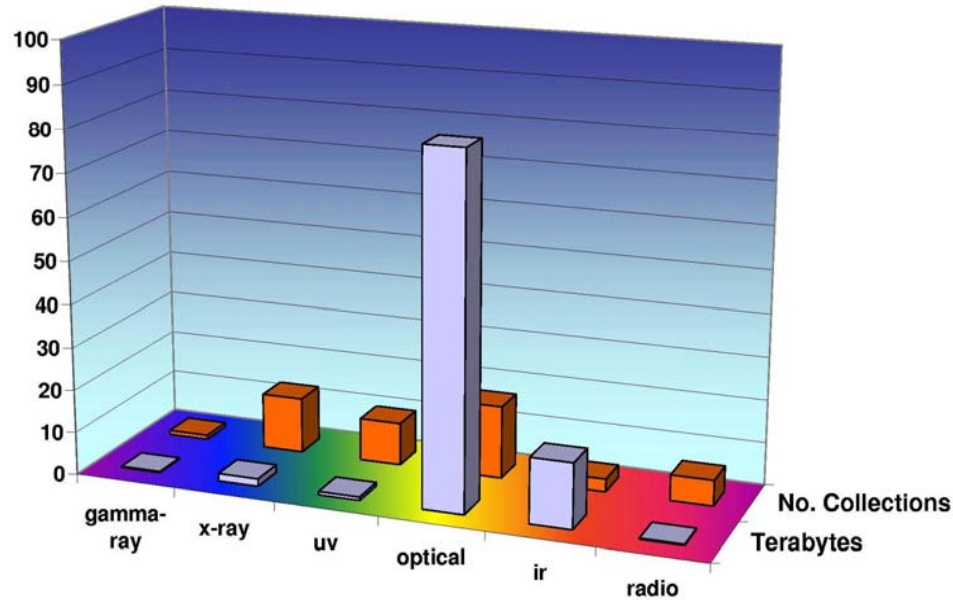
Mike Mineter
Training Outreach and Education
mjm@nesc.ac.uk

Policy for re-use

- This presentation can be re-used for academic purposes.
- However if you do so then please let training-support@nesc.ac.uk know. We need to gather statistics of re-use: no. of events, number of people trained.
Thank you!!

Data and storage

- Yesterday: how to run jobs
- **BUT**
 - Can't have computation without data!
- **AND BUT**
 - It's the data that drives research
 - Data → Information → Knowledge
- **AND YEAH BUT NO BUT**
 - Can't have digital data without computation!



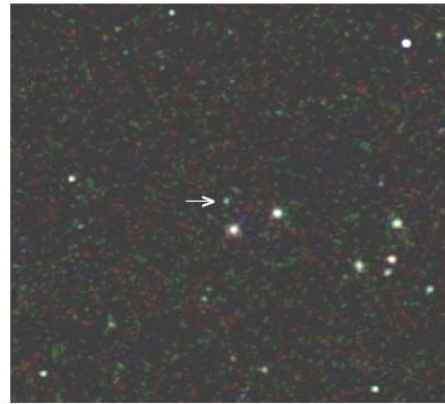
No. & sizes of data sets as of mid-2002, grouped by wavelength

- 12 waveband coverage of large areas of the sky
- Total about 200 TB data
- Doubling every 12 months
- Largest catalogues near 1 Billion objects

2MASSW J1217-03

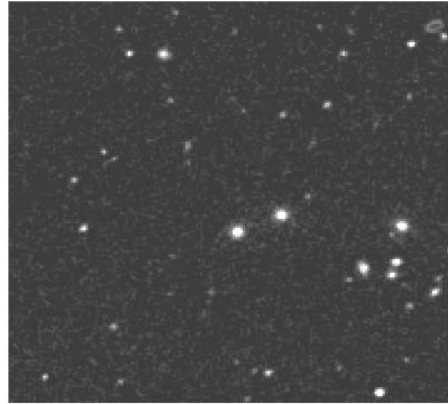
A methane (T-type) dwarf in the constellation Virgo

The near-infrared view




2MASS Composite JHK_s Atlas Image

The optical view



Palomar Digitized Sky Survey

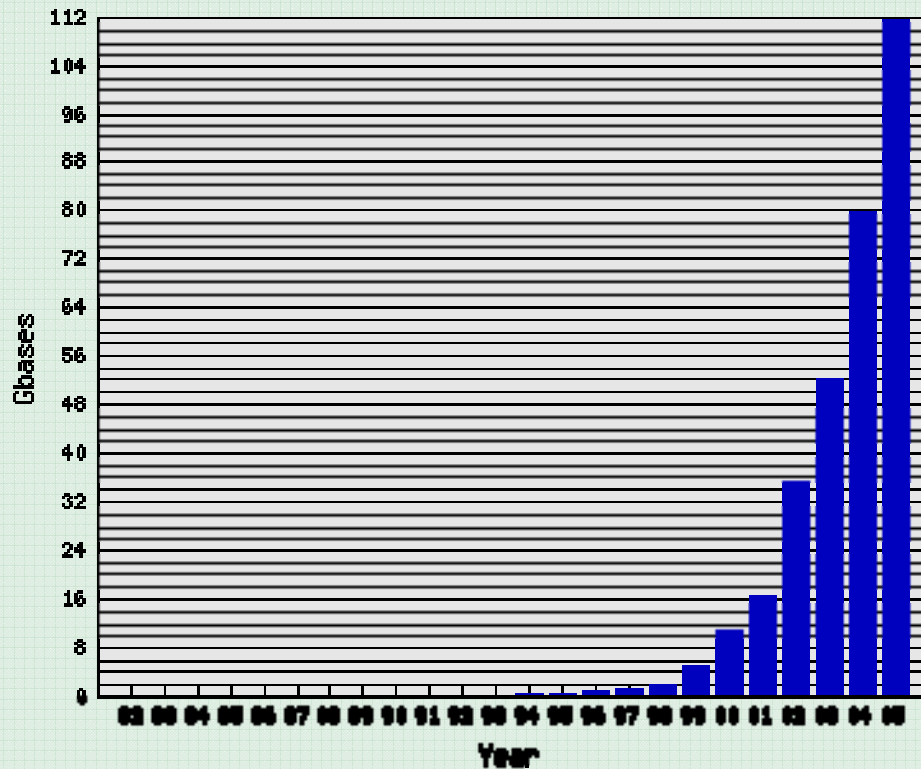


A.J.Burgasser (Caltech), J.D.Kirkpatrick (IPAC/Caltech), M.E.Brown (Caltech),
 I.N.Reid (U.Penn), J.E.Gizis (U.Mass), C.C.Dahn & D.G.Monet (USNO, Flagstaff),
 C.A.Beachman (OPL), J.L.liebert (Arizona), R.M.Cutri (IPAC/Caltech), M.F.Skrutskie (U.Mass)

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

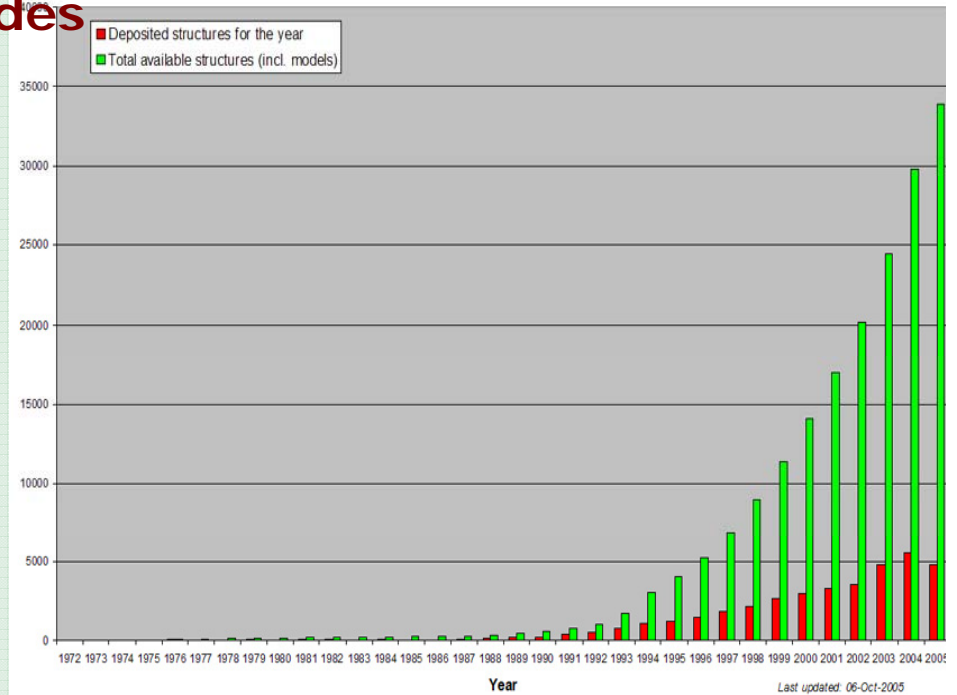
Data and images courtesy Alex Szalay, John Hopkins University

EMBL DB 111,416,302,701 nucleotides



EMBL Nucleotide Sequence Database

PDB 39,853 Protein structures



PDB - Protein Data Bank

Data, Data Everywhere

- Entering an age of data: data explosion – growing volumes
 - CERN: LHC will generate 1GB/s = 10PB/y
- Data stored in many different ways – growing diversity
 - Relational databases
 - XML databases
 - Flat files
- From
 - Legacy datasets
 - Simulations
 - New measurements, sources, analyses....
- Need integration across personal, public, licensed, and project datasets

So how do we....

- Mine data riches for nuggets of information?
 - Discovery depends on insights
 - Unpredictable or unexpected use of data
- **Facilitate**
 - Data discovery
 - Data understanding
 - Data access
 - Data integration
- **Empower e-Business and e-Research**
- **A Grid is a vehicle for achieving this**

A Grid and data

- Grid: diverse services sharing AuthN and AuthZ across admin domains, including:
 - Data storage
 - Controlled AA
 - Data transfer
 - Between stores, stores and compute nodes
 - Data catalogues, registries, ...
 - How can I find the data in the resources that I can access?
- Key issue:
 - Move data to where computation will happen?
 - (Yesterday – staging files to compute nodes)
 - Move computation to be close to data?



NGS

National Grid Service

Move computation to the data

- Assumption: code size \ll data size
 - Minimise data transport
- Computation hosted close to storage
 - I know where data are – stage or preinstall executables
 - Later today – run workflow “very close” to data

Meta-data: describing data

- Choosing data sources
 - How do you find them?
 - How are they described and advertised?
- Meta-data is required describing
 - Content
 - Provenance
 - Structure
 - Types, Formats & Ontologies
 - Operations available
 - Access requirements
 - Quality of service

**Necessary
to find, understand,
access, automate,
assess quality and
manage
But very hard to
create & maintain:
Incentives & tools**

**Solutions include: Relational databases, repositories (like
Fedora)**

Files on Grids

- Simple data files on grid-specific storage
- Middleware supporting
 - **Replica files**
 - to be close to where you want computation
 - For resilience
 - **Catalogue**: maps logical name to physical storage device/file
 - **Virtual filesystems**,
POSIX-like I/O
 - Services provided: storage, transfer, catalogue that maps logical filenames to replicas.
- Solutions include
 - **gLite (EGEE): data service**
 - **Globus: Data Replication Service**
 - **Storage Resource Broker – deployed by NGS**

Files on the NGS

- Accounts
 - Not reliable as long-term storage
 - Pool of accounts – undefined persistence
 - Useful with care!
 - E.g. you are likely to see the files you used yesterday
- Storage Resource Broker
 - THE way to hold files
 - Accessible across all nodes, and desktop/workstations....

Oracle and the NGS

- The requirement for data hosting will grow
- Oracle database: for both users and services offered by the NGS.

**The NGS Oracle service
is ready to host data
for your project!**

- Additional application needed after joining the NGS

- DAI – Data Access and Integration
- Structured data you don't want in Oracle
 - E.g. It would be a replica of existing service
 - Structured data: RDBMS, XML databases,...
 - Files on project's filesystems
 - Data that may already have other user communities not using a Grid
- **Require extendable middleware tools to support**
 - Computation near to data
 - Controlled exposure of data *without replication*
- **Basis for integration and federation**

Summary

**NGS
services**

Storage Resource
Broker

Oracle

OGSA-DAI

GridFTP

These basic functions are provided by different services
for different types of data

**Basic
functions**

Storage

Catalogue

Transfer