The Analysis & Mining of Globally Distributed Data

Chapter 1. A Quick Introduction to Data Grids, Data Webs, Semantics Webs, & Distributed Data Mining

Robert Grossman
Laboratory for Advanced Computing
University of Illinois at Chicago

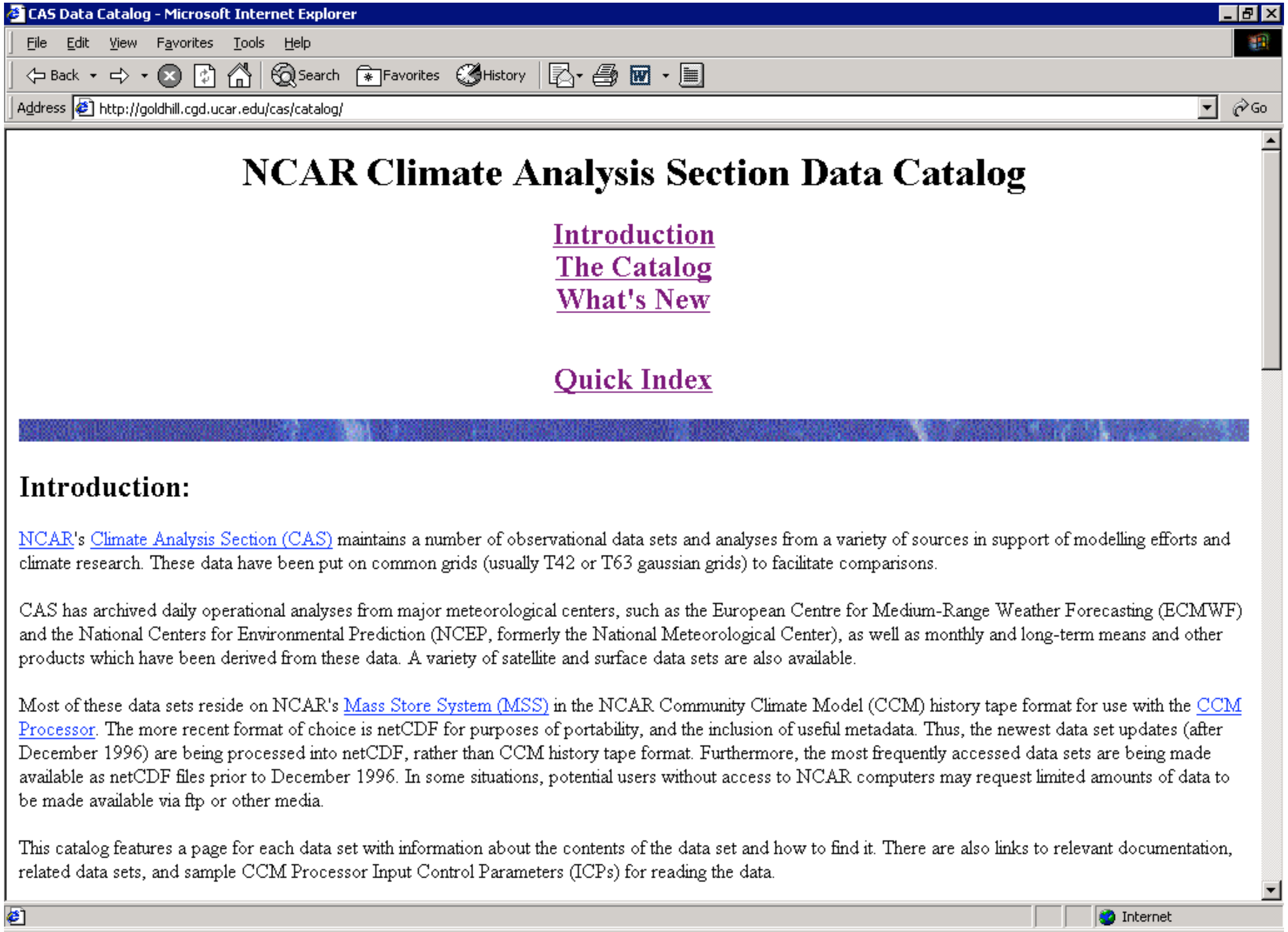
&

Open Data Partners

1.1 Background

Three Fundamental Trends

Trend 1. Explosion of Data…


CAS Data Catalog - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://goldhill.cgd.ucar.edu/cas/catalog/
Go
Internet

NCAR Climate Analysis Section Data Catalog

Introduction
The Catalog
What's New

Quick Index

Introduction:

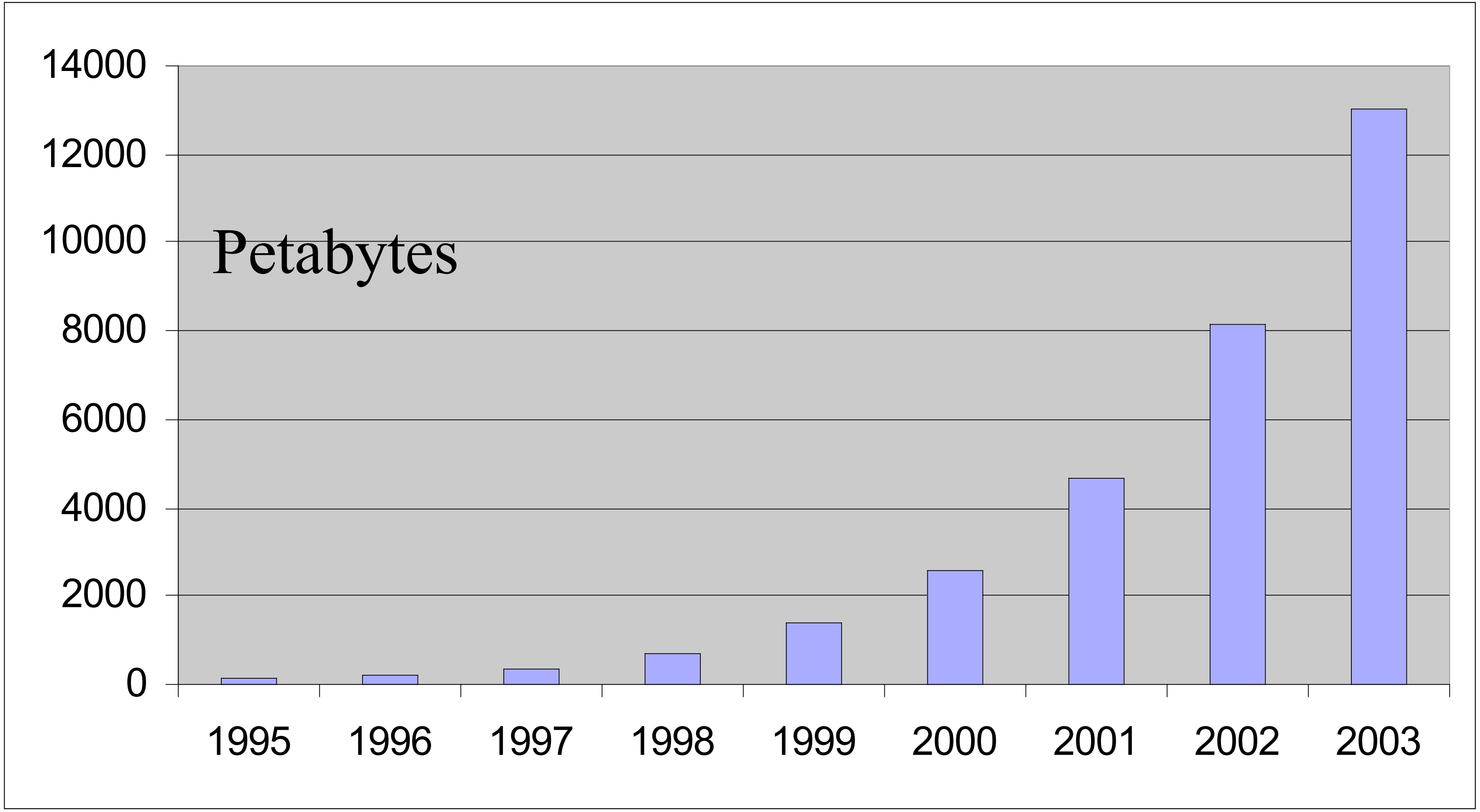
NCAR's Climate Analysis Section (CAS) maintains a number of observational data sets and analyses from a variety of sources in support of modelling efforts and climate research. These data have been put on common grids (usually T42 or T63 gaussian grids) to facilitate comparisons.

CAS has archived daily operational analyses from major meteorological centers, such as the European Centre for Medium-Range Weather Forecasting (ECMWF) and the National Centers for Environmental Prediction (NCEP, formerly the National Meteorological Center), as well as monthly and long-term means and other products which have been derived from these data. A variety of satellite and surface data sets are also available.

Most of these data sets reside on NCAR's Mass Store System (MSS) in the NCAR Community Climate Model (CCM) history tape format for use with the CCM Processor. The more recent format of choice is netCDF for purposes of portability, and the inclusion of useful metadata. Thus, the newest data set updates (after December 1996) are being processed into netCDF, rather than CCM history tape format. Furthermore, the most frequently accessed data sets are being made available as netCDF files prior to December 1996. In some situations, potential users without access to NCAR computers may request limited amounts of data to be made available via ftp or other media.

This catalog features a page for each data set with information about the contents of the data set and how to find it. There are also links to relevant documentation, related data sets, and sample CCM Processor Input Control Parameters (ICPs) for reading the data.

Distributed Exabytes



Source: IDC (1999) "1999 Winchester Disk Drive Market Forecast and Review"

. . . All in the Wrong Format

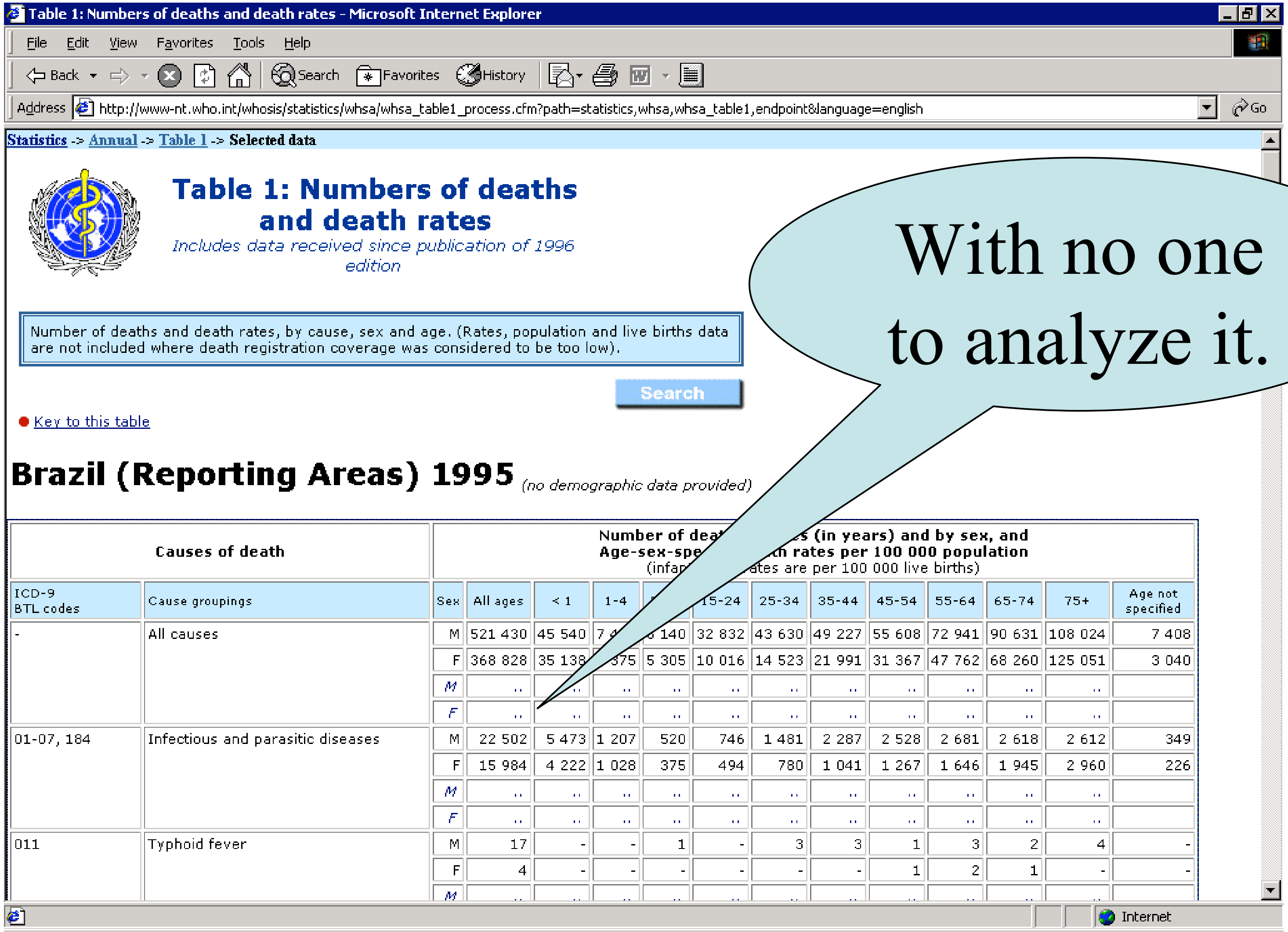


Table 1: Numbers of deaths and death rates - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://www-nt.who.int/whosis/statistics/whsa/whsa_table1_process.cfm?path=statistics,whsa,whsa_table1,endpoint&language=english
Go
Statistics -> Annual -> Table 1 -> Selected data
Search
Internet

Table 1: Numbers of deaths and death rates

Includes data received since publication of 1996 edition

Number of deaths and death rates, by cause, sex and age. (Rates, population and live births data are not included where death registration coverage was considered to be too low).

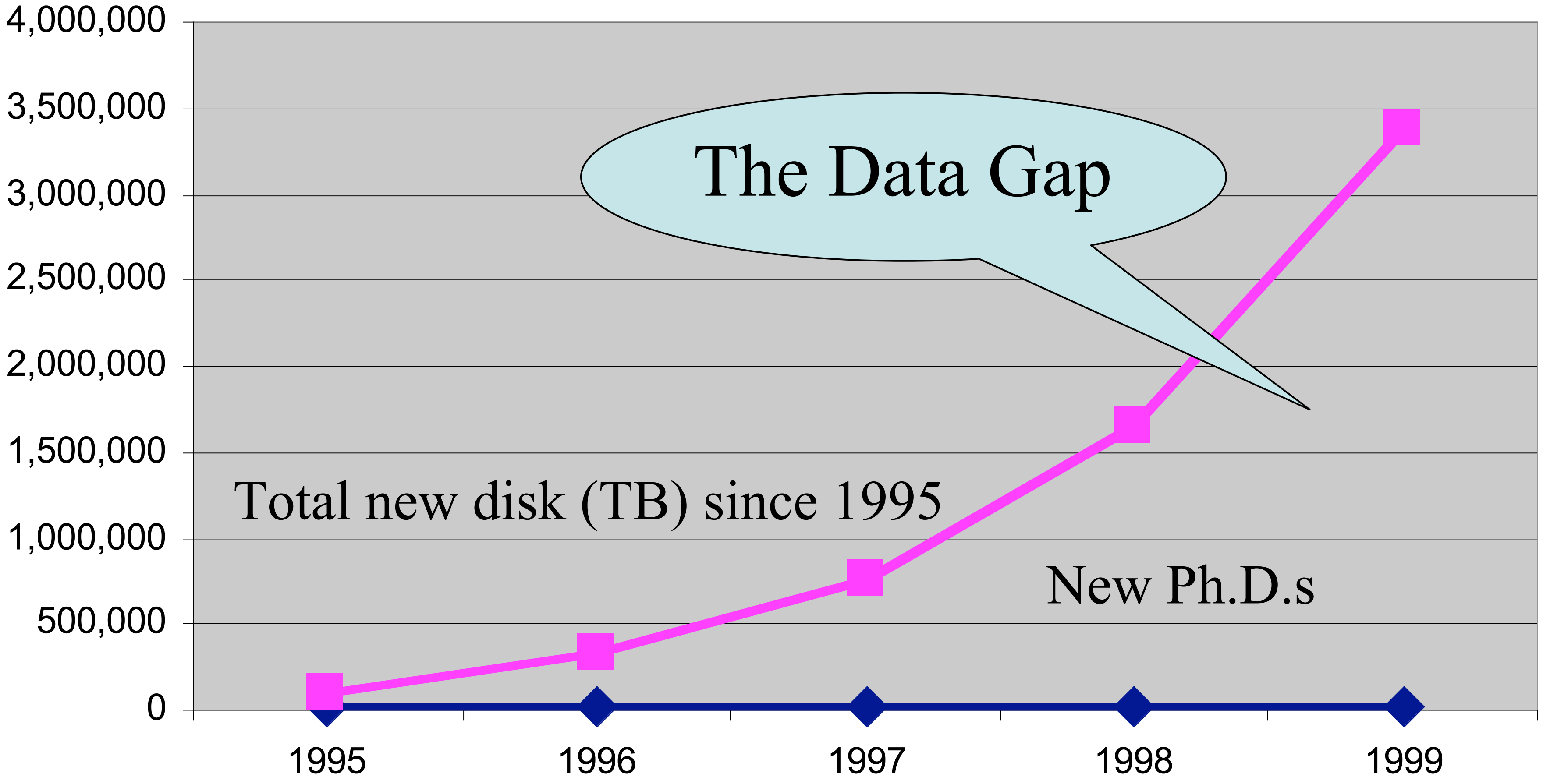
● Key to this table

Brazil (Reporting Areas) 1995 *(no demographic data provided)*

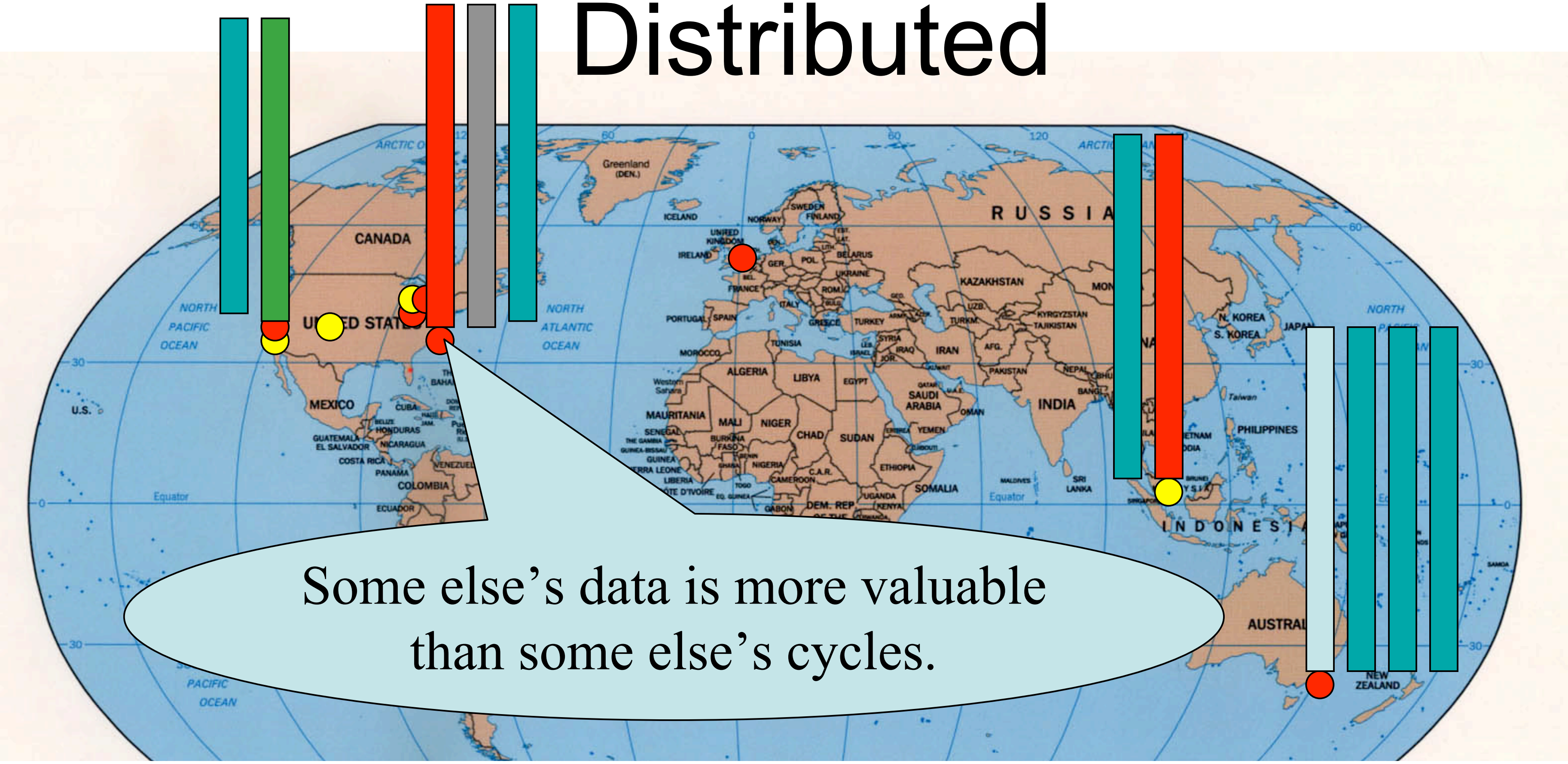
Causes of death		Number of dea s (in years) and by sex, and Age-sex-spe th rates per 100 000 population (infan ates are per 100 000 live births)												
ICD-9 BTL codes	Cause groupings	Sex	All ages	< 1	1-4	5	15-24	25-34	35-44	45-54	55-64	65-74	75+	Age not specified
-	All causes	M	521 430	45 540	7 4	140	32 832	43 630	49 227	55 608	72 941	90 631	108 024	7 408
		F	368 828	35 138	375	5 305	10 016	14 523	21 991	31 367	47 762	68 260	125 051	3 040
		M	..	..	..	..	..	..	..	..	..	..	..	
		F	..	..	..	..	..	..	..	..	..	..	..	
01-07, 184	Infectious and parasitic diseases	M	22 502	5 473	1 207	520	746	1 481	2 287	2 528	2 681	2 618	2 612	349
		F	15 984	4 222	1 028	375	494	780	1 041	1 267	1 646	1 945	2 960	226
		M	..	..	..	..	..	..	..	..	..	..	..	
		F	..	..	..	..	..	..	..	..	..	..	..	
011	Typhoid fever	M	17	-	-	1	-	3	3	1	3	2	4	-
		F	4	-	-	-	-	-	-	1	2	1	-	-
		M	..	..	..	..	..	..	..	..	..	..	..	

With no one to analyze it.

The Data Gap


4,000,000
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
1995
1996
1997
1998
1999
The Data Gap
Total new disk (TB) since 1995
New Ph.D.s

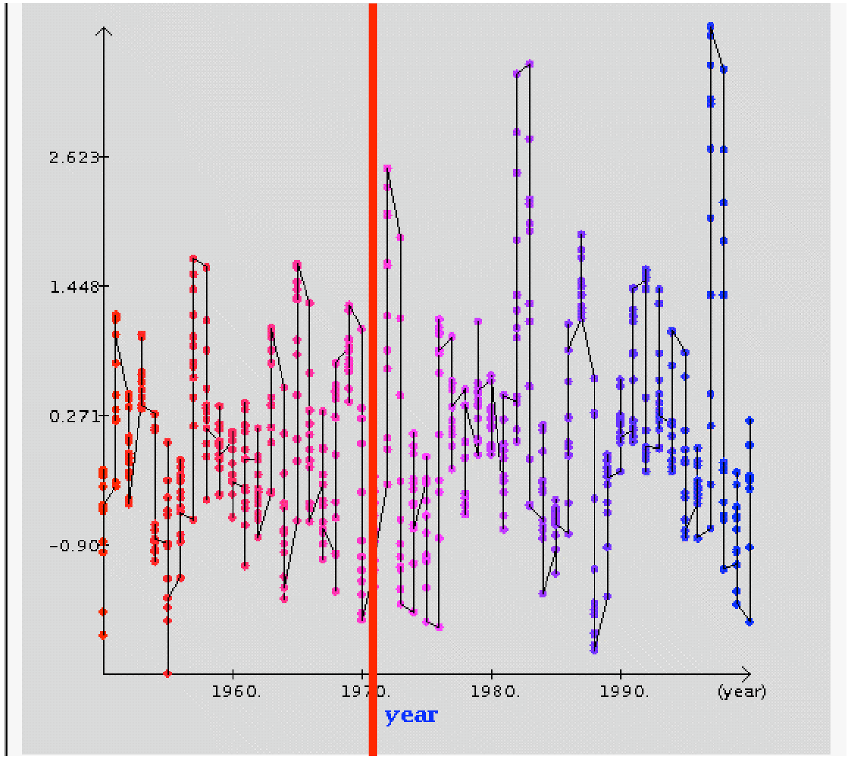
Trend 2: Most Data is Distributed



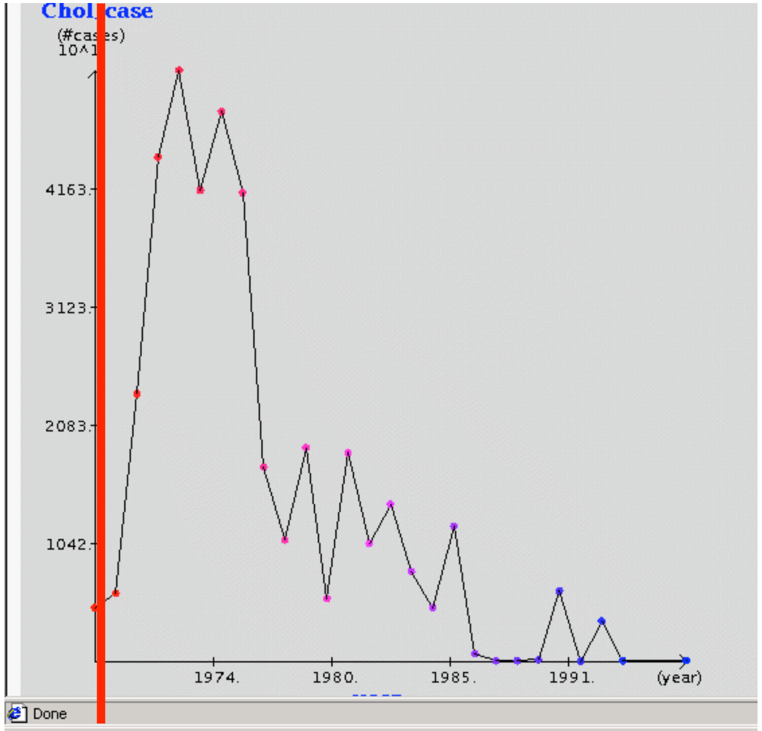
Some else’s data is more valuable than some else’s cycles.

- Pearson’s Law: The usefulness of a column of data varies as the square of the number of columns it is compared to.

Example: ENSO & Cholera

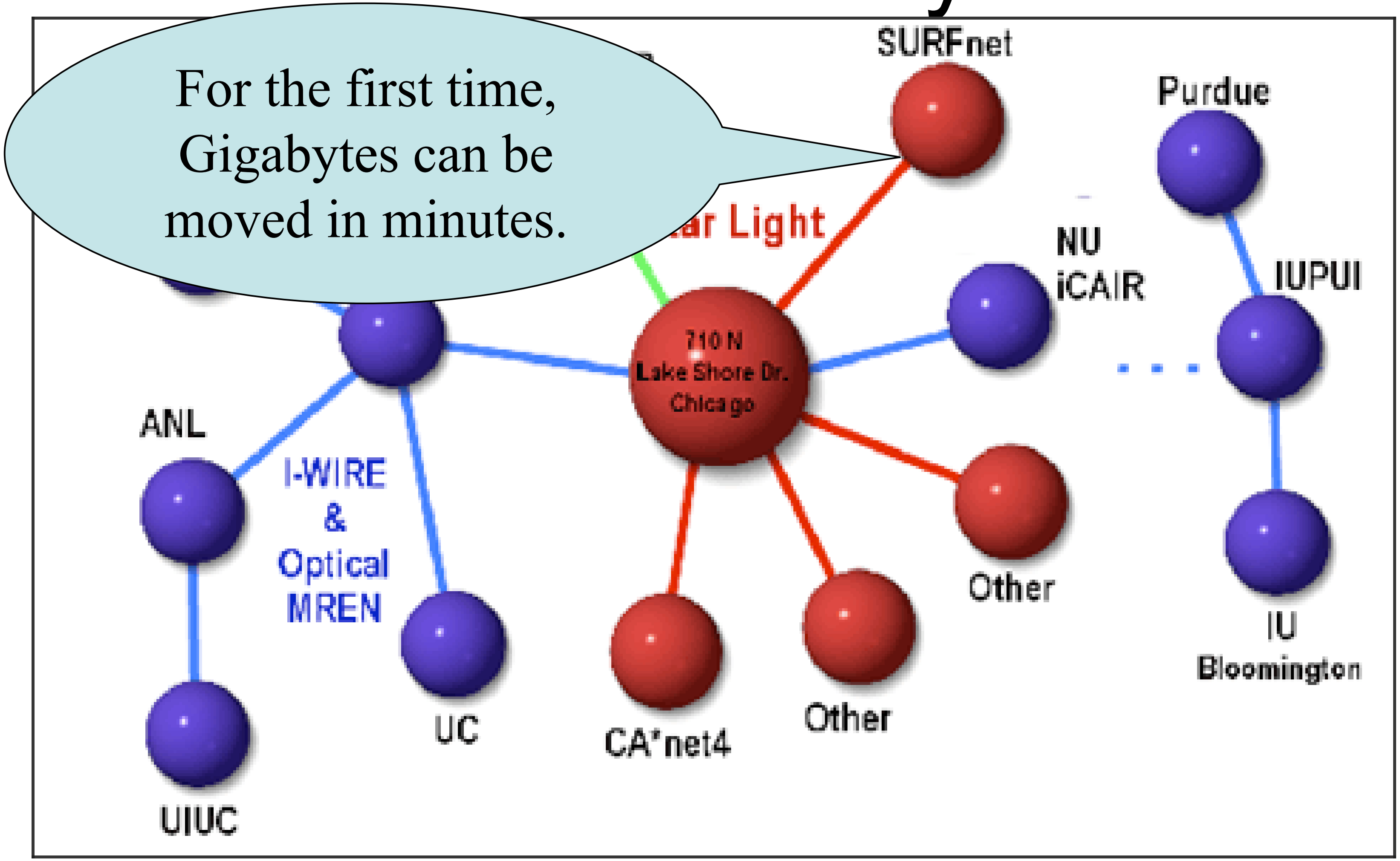


El Nino Data at NCAR



Cholera Data at WHO

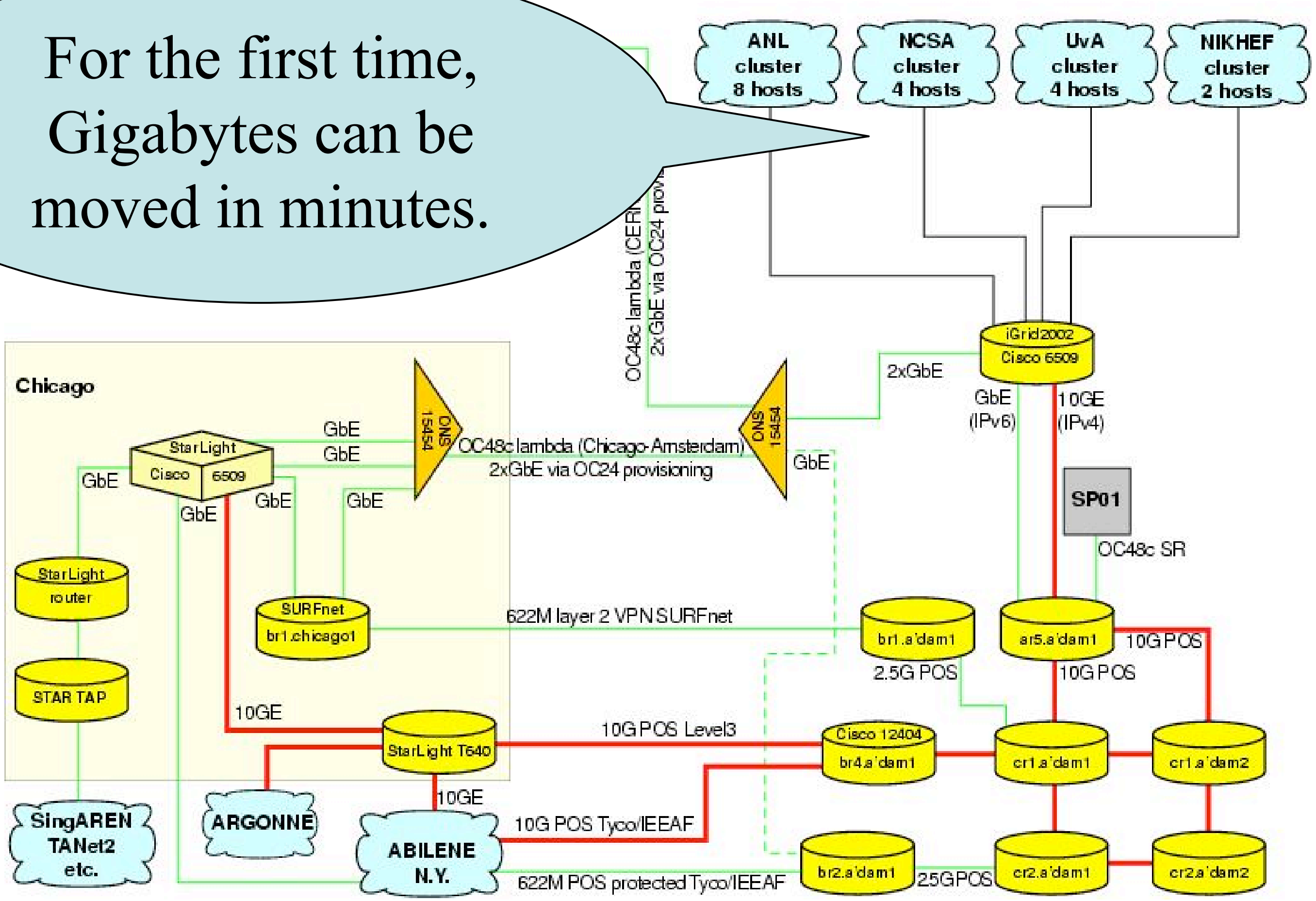
Trend 3. Bandwidth is a Commodity.



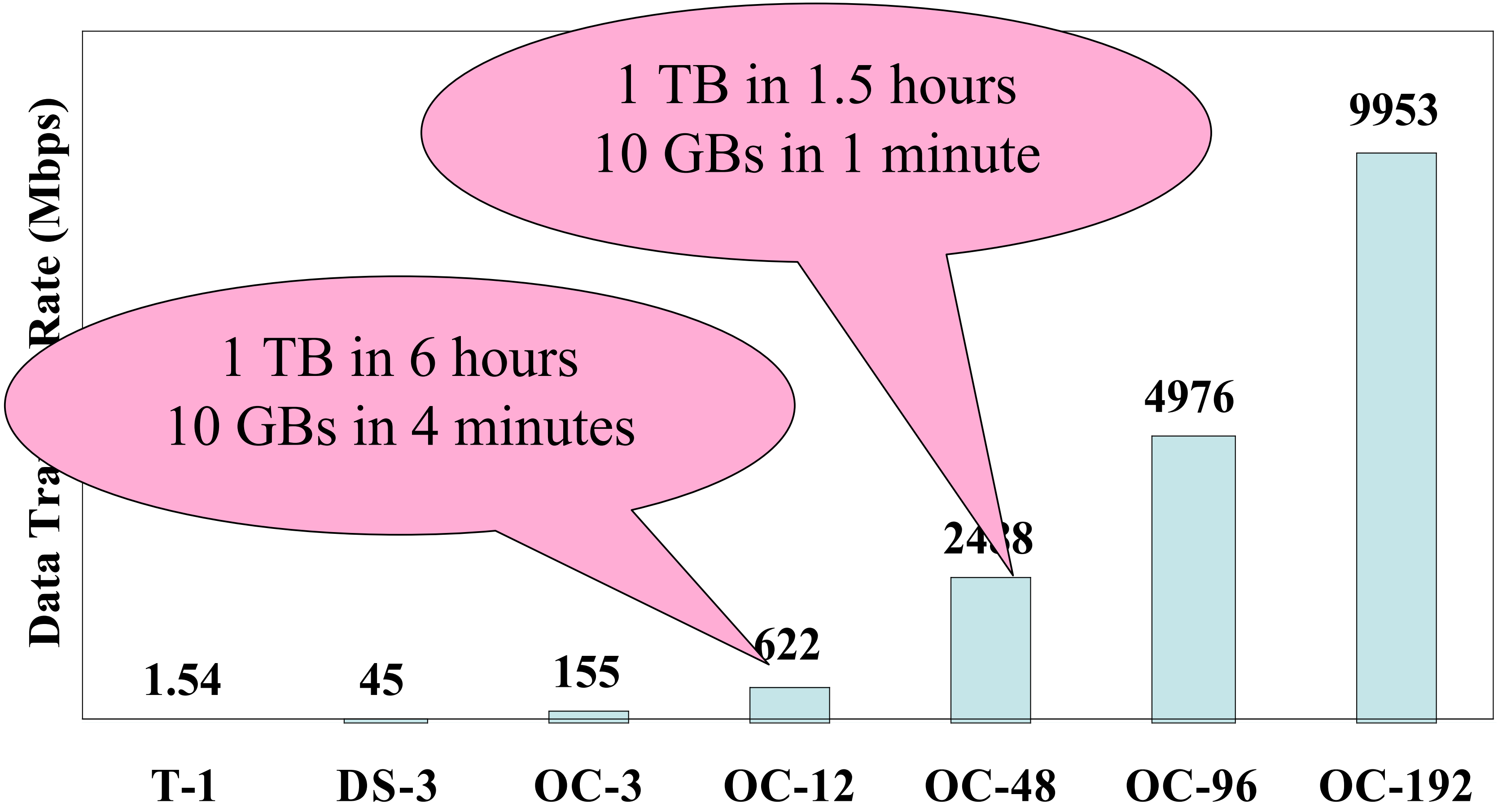
For the first time, Gigabytes can be moved in minutes.

Trend 3. Bandwidth is a Commodity.

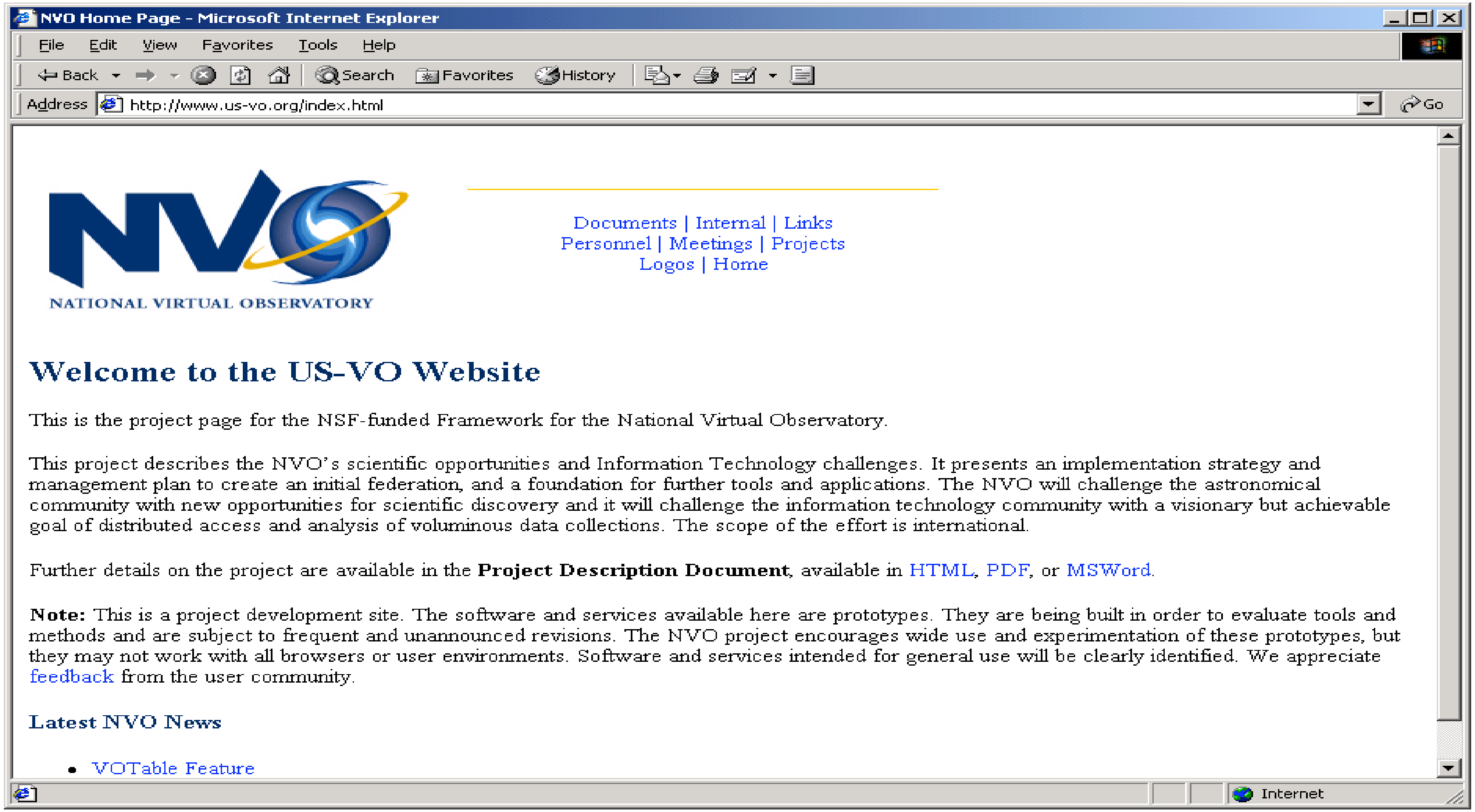
For the first time, Gigabytes can be moved in minutes.



Gigabytes can be Moved in Minutes



Example 1: Data Grids – National Virtual Observatory


NVO Home Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://www.us-vo.org/index.html
Go
NVO
NATIONAL VIRTUAL OBSERVATORY
Documents | Internal | Links
Personnel | Meetings | Projects
Logos | Home
Internet

Welcome to the US-VO Website

This is the project page for the NSF-funded Framework for the National Virtual Observatory.

This project describes the NVO's scientific opportunities and Information Technology challenges. It presents an implementation strategy and management plan to create an initial federation, and a foundation for further tools and applications. The NVO will challenge the astronomical community with new opportunities for scientific discovery and it will challenge the information technology community with a visionary but achievable goal of distributed access and analysis of voluminous data collections. The scope of the effort is international.

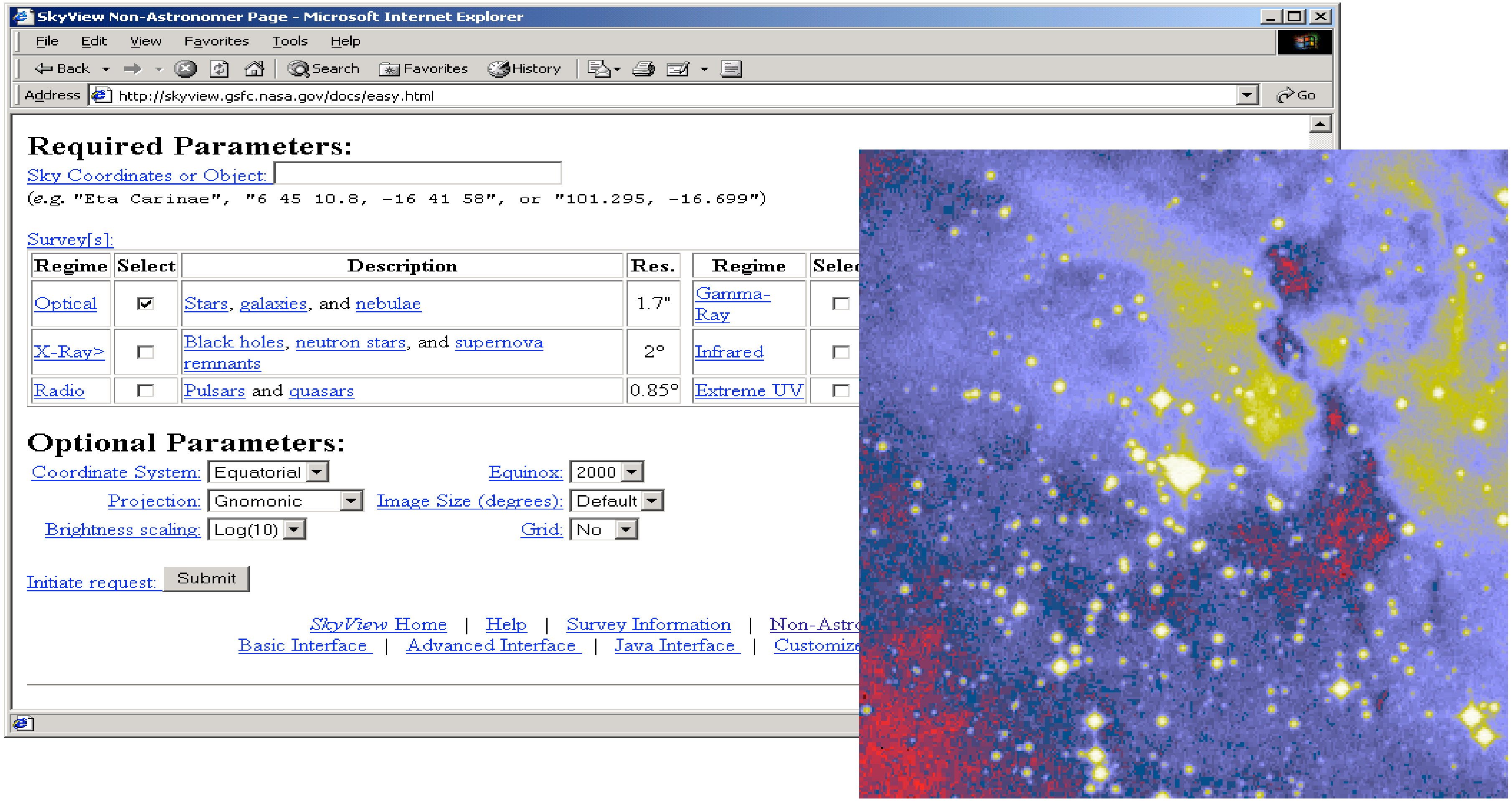
Further details on the project are available in the **Project Description Document**, available in HTML, PDF, or MSWord.

Note: This is a project development site. The software and services available here are prototypes. They are being built in order to evaluate tools and methods and are subject to frequent and unannounced revisions. The NVO project encourages wide use and experimentation of these prototypes, but they may not work with all browsers or user environments. Software and services intended for general use will be clearly identified. We appreciate feedback from the user community.

Latest NVO News

- VOTable Feature

NASA SkyView


SkyView Non-Astronomer Page - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://skyview.gsfc.nasa.gov/docs/easy.html
Go

Required Parameters:

Sky Coordinates or Object:

(e.g. "Eta Carinae", "6 45 10.8, -16 41 58", or "101.295, -16.699")

Survey[s]:

Regime	Select	Description	Res.	Regime	Selec
Optical	☑	Stars, galaxies, and nebulae	1.7"	Gamma-Ray	□
X-Ray>	□	Black holes, neutron stars, and supernova remnants	2°	Infrared	□
Radio	□	Pulsars and quasars	0.85°	Extreme UV	□

Optional Parameters:

Coordinate System: Equatorial

Equinox: 2000

Projection: Gnomonic

Image Size (degrees): Default

Brightness scaling: Log(10)

Grid: No

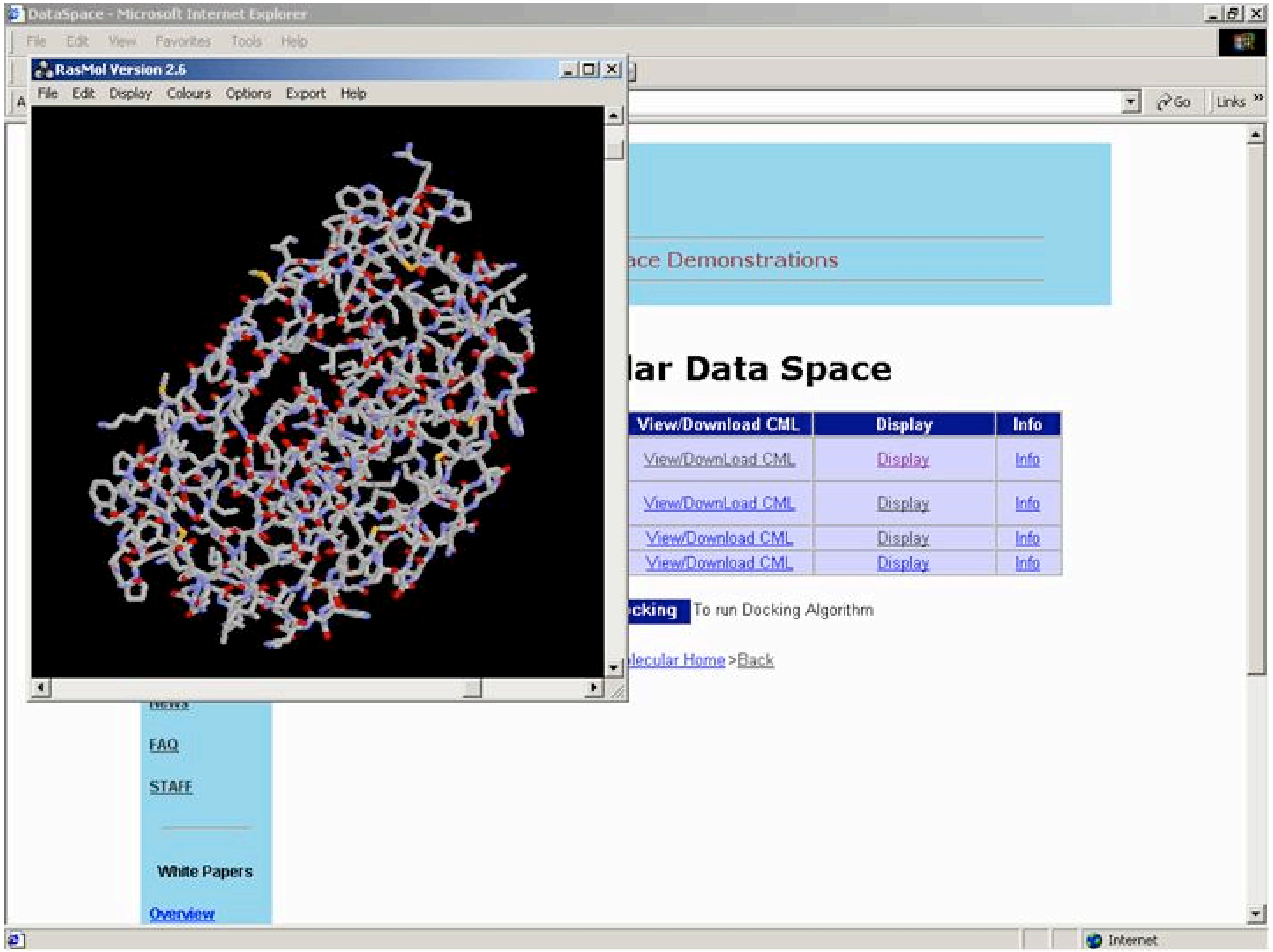
Initiate request: Submit

SkyView Home | Help | Survey Information | Non-Astro

Basic Interface | Advanced Interface | Java Interface | Customize

Copyright 2003 Robert L. Grossman

Example 2: Data Webs – Molecular Data Space


DataSpace - Microsoft Internet Explorer
File Edit View Favorites Tools Help
RasMol Version 2.6
File Edit Display Colours Options Export Help
Go
Links
Internet

ace Demonstrations

lar Data Space

View/Download CML	Display	Info
View/DownLoad CML	Display	Info
View/DownLoad CML	Display	Info
View/Download CML	Display	Info
View/Download CML	Display	Info

cking To run Docking Algorithm

lecular Home >Back

News

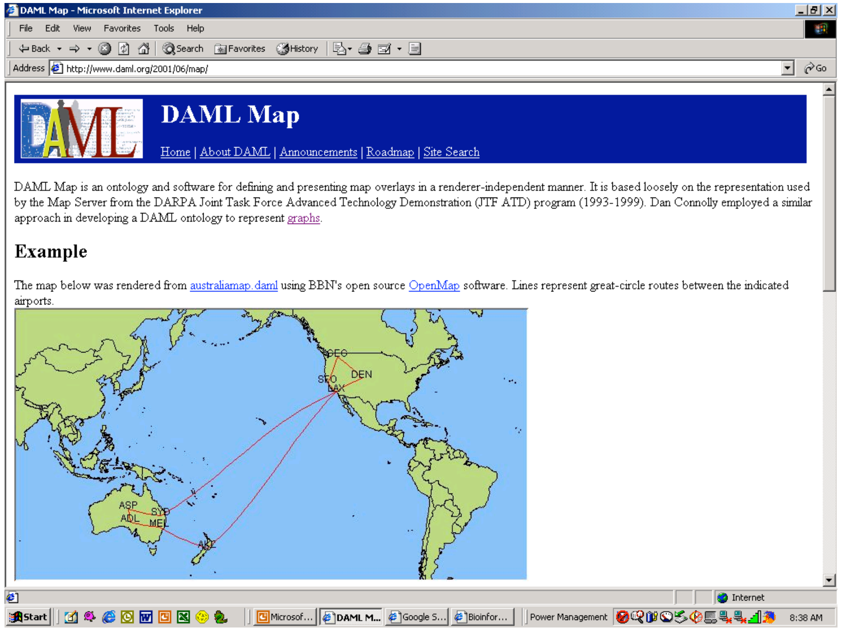
FAQ

STAFF

White Papers

Overview

Example 3:
Semantic Web – DAML


DAML Map - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://www.daml.org/2001/06/map/
Go
SEA
SFO
LAX
DEN
ASP
SYD
ADL
MEL
AKL
Internet
Start
Microsof...
DAML M...
Google S...
Bioinfor...
Power Management
8:38 AM

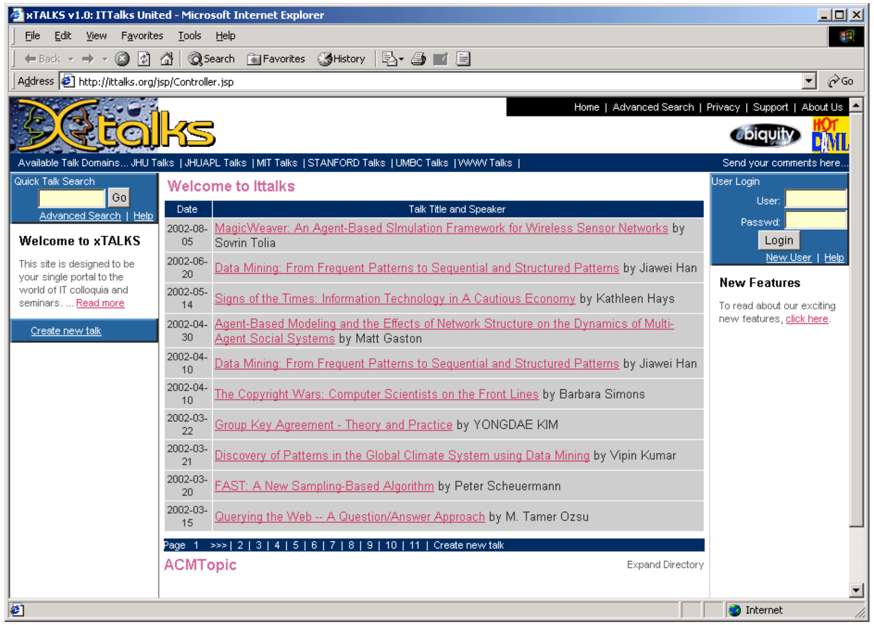
DAML Map

Home | About DAML | Announcements | Roadmap | Site Search

DAML Map is an ontology and software for defining and presenting map overlays in a renderer-independent manner. It is based loosely on the representation used by the Map Server from the DARPA Joint Task Force Advanced Technology Demonstration (JTF ATD) program (1993-1999). Dan Connolly employed a similar approach in developing a DAML ontology to represent graphs.

Example

The map below was rendered from australiamap.daml using BBN's open source OpenMap software. Lines represent great-circle routes between the indicated airports.


xTALKS v1.0: ITTalks United - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://ittalks.org/jsp/Controller.jsp
Go
Home | Advanced Search | Privacy | Support | About Us
Available Talk Domains... JHU Talks | JHUAPL Talks | MIT Talks | STANFORD Talks | UMBC Talks | WWW Talks |
Send your comments here...
Quick Talk Search
Go
Advanced Search | Help
Create new talk
User Login
User:
Passwd:
Login
New User | Help
Internet

Welcome to xTALKS

This site is designed to be your single portal to the world of IT colloquia and seminars. ... Read more

Welcome to Ittalks

Date	Talk Title and Speaker
2002-08-05	MagicWeaver: An Agent-Based SImulation Framework for Wireless Sensor Networks by Sovrin Tolia
2002-06-20	Data Mining: From Frequent Patterns to Sequential and Structured Patterns by Jiawei Han
2002-05-14	Signs of the Times: Information Technology in A Cautious Economy by Kathleen Hays
2002-04-30	Agent-Based Modeling and the Effects of Network Structure on the Dynamics of Multi-Agent Social Systems by Matt Gaston
2002-04-10	Data Mining: From Frequent Patterns to Sequential and Structured Patterns by Jiawei Han
2002-04-10	The Copyright Wars: Computer Scientists on the Front Lines by Barbara Simons
2002-03-22	Group Key Agreement - Theory and Practice by YONGDAE KIM
2002-03-21	Discovery of Patterns in the Global Climate System using Data Mining by Vipin Kumar
2002-03-20	FAST: A New Sampling-Based Algorithm by Peter Scheuermann
2002-03-15	Querying the Web -- A Question/Answer Approach by M. Tamer Ozsu

Page 1 >>> | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Create new talk

ACMTopic

Expand Directory

New Features

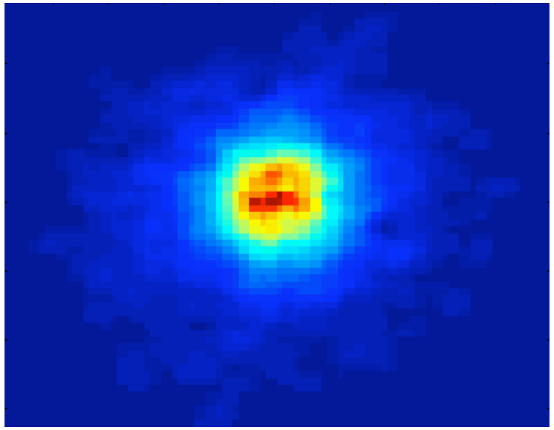
To read about our exciting new features, click here.

Example 4: Data Mining – Sky Survey Catalog

- Goal: To predict class (star or galaxy) of sky objects, based on survey images (from Palomar Observatory)
  - 3000 images with 23K x 23K pixels/image.
- Approach:
  - Partition the image & create 40 features
  - Build a classification model
  - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find.

Classifying Galaxies

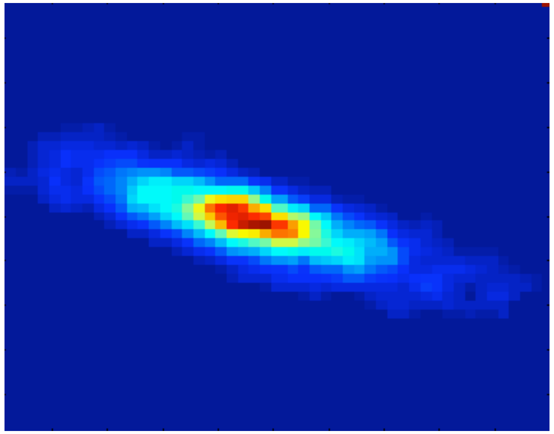
Early



Class:

- **Stages of Formation**

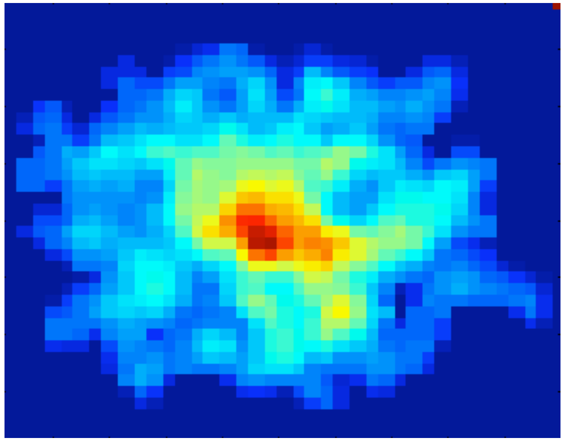
Intermediate



Attributes:

- **Image features,**
- **Characteristics of light waves received, etc.**

Late



Data Size:

- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

From Fayyad, et.al., Advances in Knowledge Discovery and Data Mining, 1996

1.2 Four Different Philosophies for Working with Distributed Data

Data Grids, Data Webs, Semantic Webs, & Distributed Data Mining

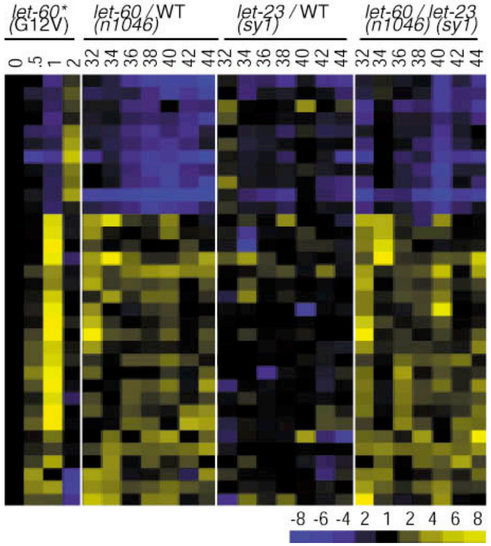
Data Grids – Categorical Imperative

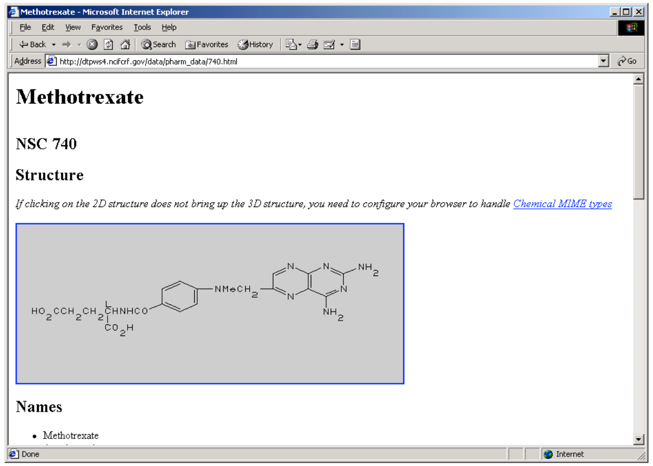
grid

- How can we *interoperate* distributed supercomputers?

Data Webs –
Categorical Imperative

- How can we *explore* other people’s data?




Methotrexate - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://dtpws4.ncifcrf.gov/data/pharm_data/740.html
Go
HO2CCH2CH2CHNHCO
L
CO2H
NMeCH2
N
N
N
N
NH2
NH2
Done
Internet

Methotrexate

NSC 740

Structure

If clicking on the 2D structure does not bring up the 3D structure, you need to configure your browser to handle Chemical MIME types

Names

- Methotrexate

Semantic Web –
Categorical Imperative

```
<City rdf:ID="AABS">
    <geolocationCode>AABS</geolocationCode>
    <name>AABENRAA</name>
    <installationTypeCode>CTY</installationTypeCod
    <primeGeoloc rdf:resource="#AABQ" />
    <longitude>0092600E</longitude>
     <latitude>550300N</latitude>
</City>
```

DAML Geofile

What is the distance between Chicago and Baltmore?

- How can we extend the web to support *knowledge?*

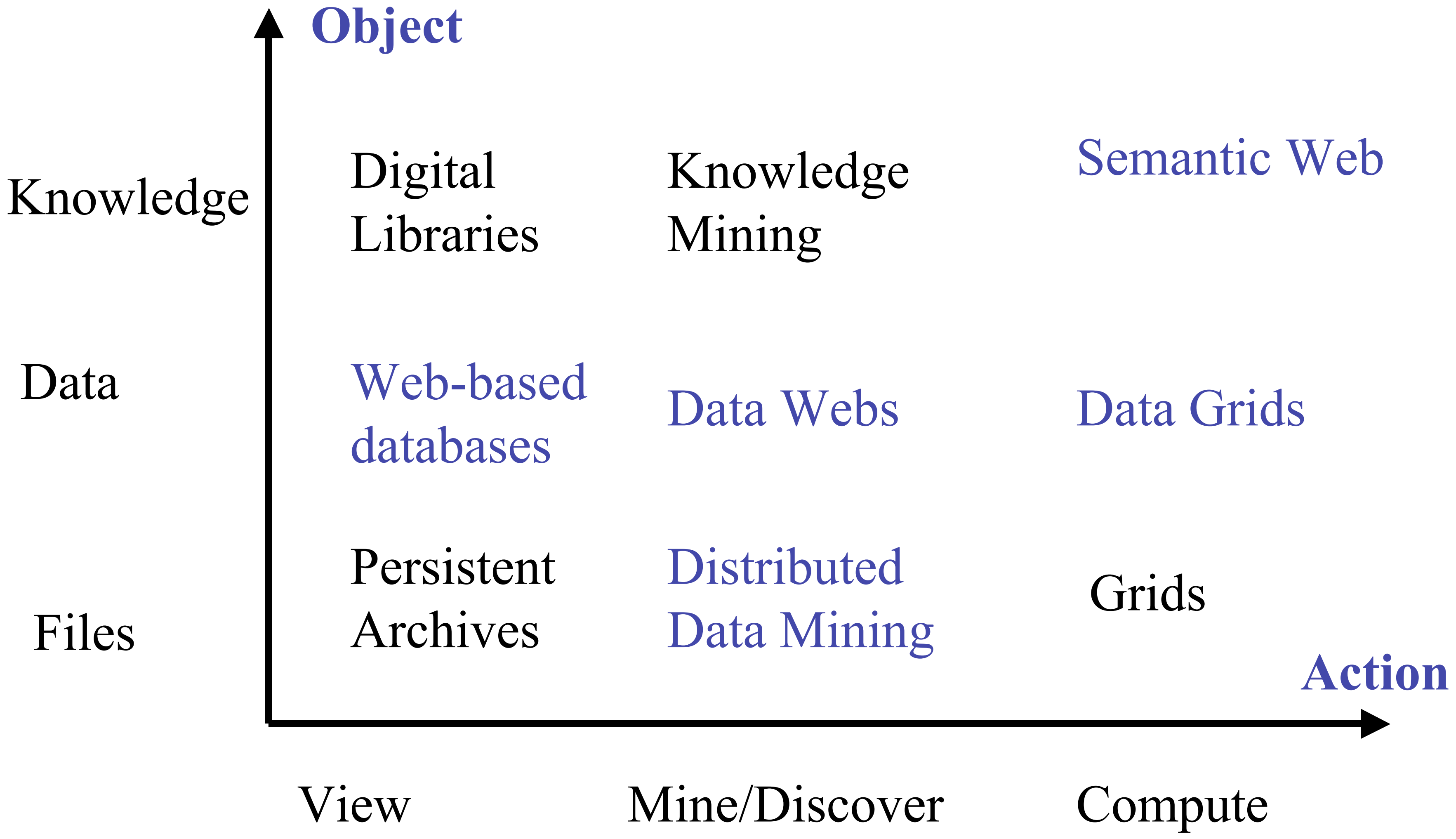
Data Mining – Categorical Imperative

Longitude	Latitude	Time	Cloud Cover
19.6875	-12.557755	120.0	0.43481043
75.9375	-12.557755	120.0	0.9641479
132.1875	-12.557755	120.0	0.82385314
188.4375	-12.557755	120.0	0.91212153
244.6875	-12.557755	120.0	0.8691945
300.9375	-12.557755		0.36265105
357.1			0.52140427
53			0.9674745
10			0.9179723

Are there any patterns relating cloud cover & biodiversity?

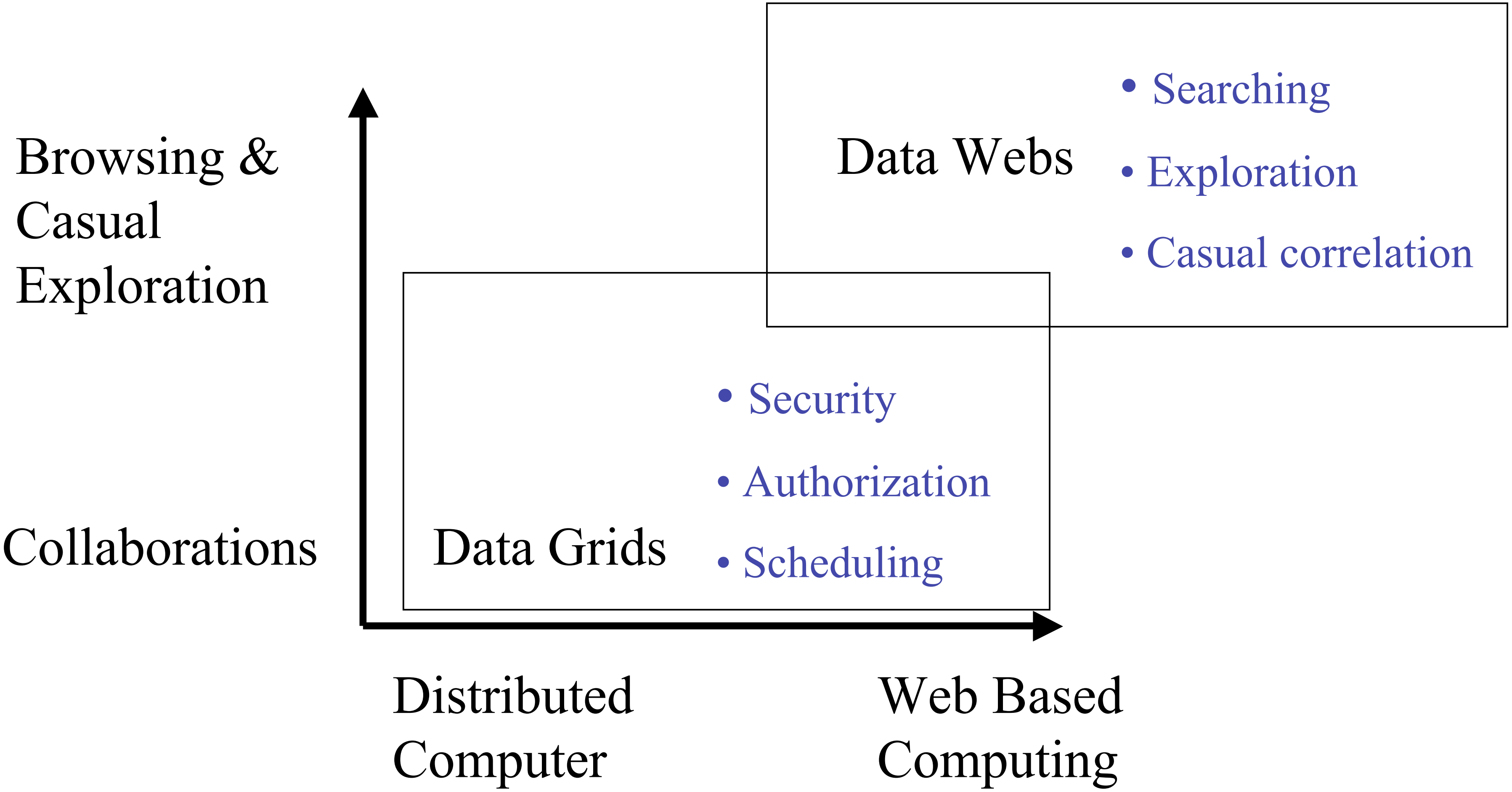
- How can we find *patterns* in distributed data?

Technologies for Global Data



	View	Mine/Discover	Compute
Knowledge	Digital Libraries	Knowledge Mining	Semantic Web
Data	Web-based databases	Data Webs	Data Grids
Files	Persistent Archives	Distributed Data Mining	Grids

Data Grids vs. Data Webs



Data Grids, Data Mining & Data Webs

	Data Grid	Distributed Data Mining	Data Web
Goal	distributed computation	distributed data mining	data explor. & mining
Services	authorization, security, resources	building models, transforming data, etc.	publishing, merging, & correlating columns
Protocol	TCP, GridFTP	TCP	DWTP, …
Platform	dist. clusters	server	dist. cluster

Semantic Web vs. Data Web

	Document Web	Semantic Web	Data Web
Protocol	HTTP	HTTP, SOAP	DWTP, SOAP
Languages	HTML, XML	XML, RDF	XML, PMML …
Action	keyword search	RDF inferences	correlate and mine
Platform	server	server	server, cluster

What is a Petabyte?

- HSS Camp:
  - 10^{15} bytes
  - Tertiary storage, data migration, data staging, …
- Data Grid Camp:
  - Thousand TB data sets with AAA
  - Security, authorization, task scheduling, replication management, …
- Data Web Camp:
  - Hundred Million 10 MB / Million 1 GB open data sets
  - Discovery, correlation, normalization, transform, …

Overview of Course

1. Introduction
2. Introduction to Data Mining
3. Protocols and Stacks
4. Web Services
5. Data Grids
6. Data Webs

Sources for Some of the Slides

- Semantic web and web services: Isabel Cruz, SC 02 Tutorial Notes
- Data grids: Introduction to Grid Computing and Globus Toolkit, www.globus.org