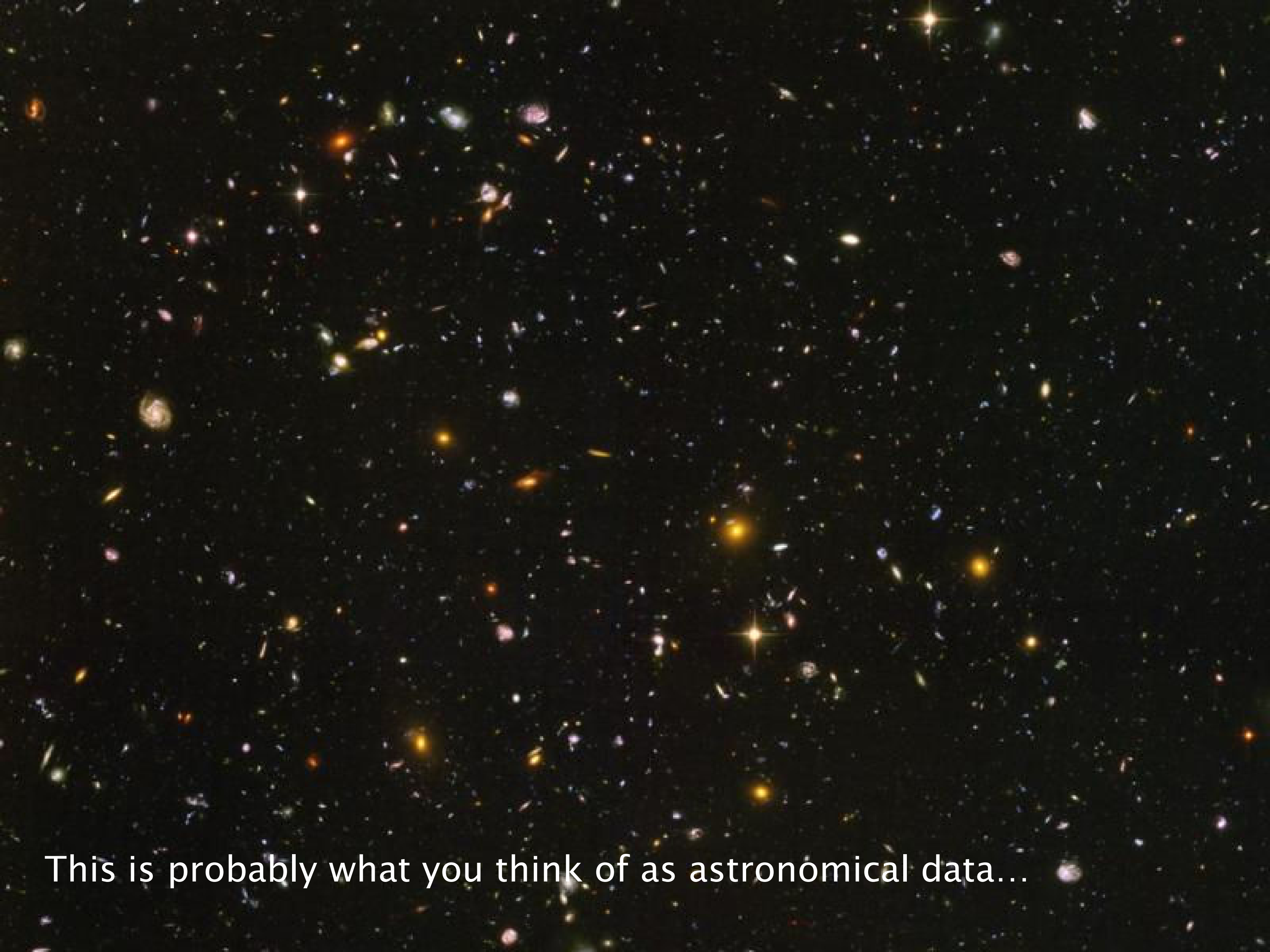
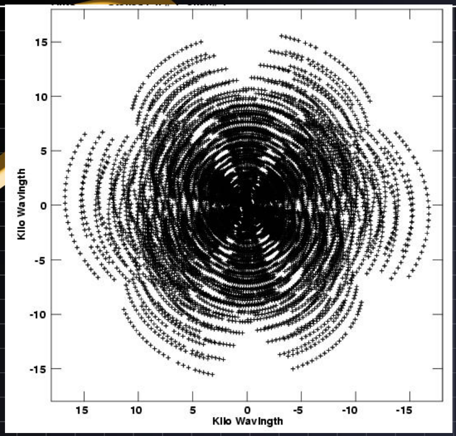
Data Standards in Astronomy

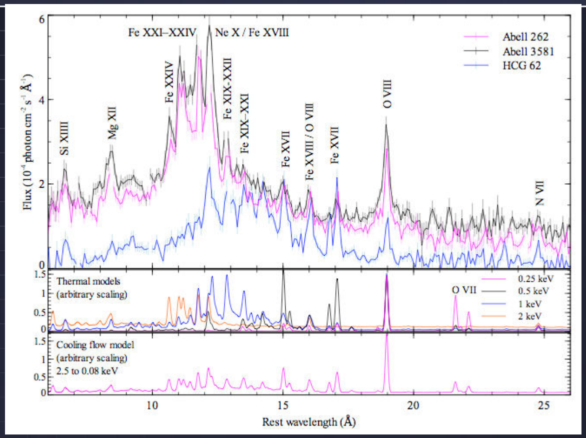
Dr. Robert J. Hanisch
Director, US Virtual Astronomical Observatory
Space Telescope Science Institute
Baltimore, MD

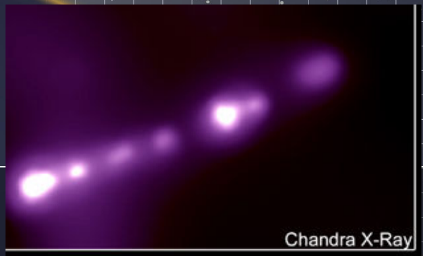

This is probably what you think of as astronomical data…

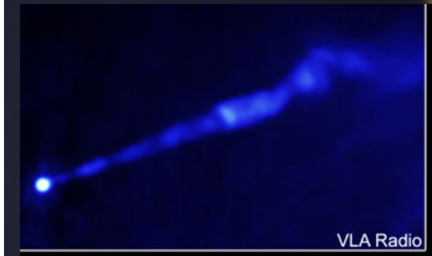
A variety of data types

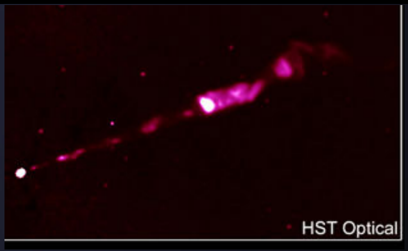


QuickTime™ and a
decompressor
are needed to see this picture.

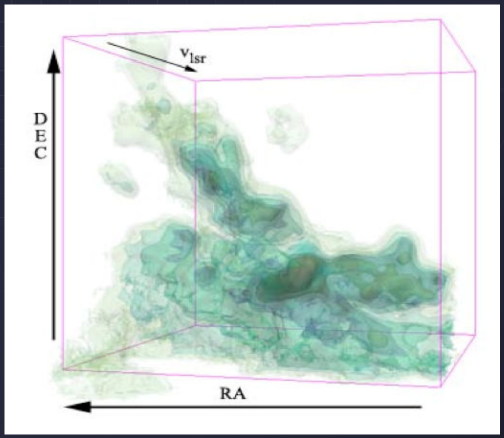








QuickTime™ and a
decompressor
are needed to see this picture.



jdate	designation	ra	dec	sup_ra	sup_dec	glon	glat	density	r_k20fe
2451305.6569	12552517+2134339	12 55 25.2	21 34 33.9	193.854874	21.576124	312.417426	84.374259	2.51	5.4
2451700.6751	12554924+2123581	12 55 49.2	21 23 58.2	193.955109	21.399385	313.052422	84.184820	2.66	5.0
2451261.8020	12571719+2120180	12 57 17.2	21 20 18.1	194.321625	21.338383	316.206288	84.058667	2.24	5.9
2451261.8020	12572936+2132520	12 57 29.4	21 32 52.1	194.372269	21.547800	317.124576	84.251945	2.24	5.0
2451261.8020	12572893+2137370	12 57 28.9	21 37 37.1	194.370453	21.627054	317.296215	84.329266	2.35	5.0
2451261.7924	12562741+2131175	12 56 27.4	21 31 17.6	194.114243	21.521484	314.721401	84.277947	2.54	8.1
2451261.8020	12573991+2146420	12 57 39.9	21 46 42.1	194.416367	21.778374	318.097097	84.465693	2.35	5.0
2451261.7972	12564252+2148223	12 56 42.5	21 48 22.4	194.177277	21.806303	315.909776	84.544554	2.51	5.0
null	12564369+2140575	12 56 43.7	21 40 57.6	194.182068	21.682659	315.683000	84.423000	null	214.8
2451261.7924	12561052+2148274	12 56 10.5	21 48 27.5	194.043808	21.807701	314.635368	84.571294	2.78	7.4
2451261.7972	12571196+2146234	12 57 12.0	21 46 23.5	194.299911	21.773294	316.993658	84.486836	2.51	5.5
2451261.8020	12572147+2140450	12 57 21.5	21 40 45.1	194.339417	21.679213	317.136181	84.386833	2.35	5.3
2451700.6751	12554548+2153222	12 55 45.5	21 53 22.2	193.939529	21.889559	313.784434	84.669874	2.66	9.0

1-d, 2-d, 3-d: intensity/polarization vs. energy, time, position, velocit

tables: catalogs, x-ray event lists, radio visibility measurements

Quantity and distribution

- ~50 major data centers and observatories with substantial on-line data holdings
- ~10,000 data “resources” (catalogs, surveys, archives)
- data centers host from a few to ~100 TB each, currently ~1 PB total
- current growth rate ~0.5 PB/yr, expected to increase soon
- current request rate ~1 PB/yr
- for Hubble Space Telescope, data retrievals are 3X data ingest; papers based on archival data constitute 2/3 of refereed publications

Common data representations

- Flexible Image Transport System – FITS
  - 25-year heritage
  - Worldwide adoption for both archival and run-time applications
  - International review and endorsement, IAU
  - n-dim arrays, ASCII tables, binary tables, compound constructs
  - Simple syntax, limited semantics (primarily for coordinates)

  Examples:
  - Single image: 2-dim array with coordinate system metadata
  - Multiple images: set of N 2-dim arrays each with coordinate system metadata (sometimes called an "association")
  - X-ray event list: binary table of photon arrival times, positions, and energies
  - Spectrum: 1-dim array of fluxes with spectral dispersion metadata, or ASCII table of wavelengths, flux values, and flux uncertainties, or binary table of same
  - Data cube: 3-dim stack of 2-dim image planes (all share the same coordinate system metadata) with third axis representing velocity

Common data representations

- VOTable
  - XML-based standard for tabular data
  - Standard schema
  - Java, C++, and Perl software libraries
  - Complements FITS
  - Incorporates semantics

  Examples:
  - Object catalog: e.g., positions, fluxes, and morphological measurements of galaxies
  - Result of database query: rows/columns that satisfy a constraint
  - Observation catalog: list of images taken with a particular instrument with pointing positions, image extents, bandpasses, etc.
  - Spectral energy distribution: composite "spectrum" based on both spectral and photometric measurements

Semantics

- Unified Content Descriptors
  - A generic syntax and agreed-upon vocabulary for astronomical quantities
  - Derived from maintenance of thousands of astronomical catalogs, where many names used to represent the same quantities

  Examples: *instr.bandpass, time.interval, stat.error;phot.flux.density;em*

- RDF/SKOS-based standard vocabulary

- VOEvent
  - Standard representation of transient event (gamma ray burst, supernova, flaring star, discovery of solar system object, etc.)
  - Represented as XML schema

Data discovery

- Resource Metadata
  - Descriptions of data collections and the organizations responsible for them, data delivery services, computational services, software, etc.
  - Based on Dublin Core (library community standard) with astronomy-specific extensions
  - Represented as XML schema; extensible
  - Contents stored in Resource Registries that exchange metadata records through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)
- Space-Time Coordinates
  - Standard representations of locations of astronomical objects in space, wavelength (energy), and time
  - Represented as XML schema
- Identifiers
  - Rules for constructing URIs for IVOA resources

Data access

- Cone Search
  - Simplest possible astronomical query: return a list of objects or observations within a certain radius of a given position on the sky
  - Response is encoded as VOTable
- Simple Image Access Protocol
  - Extends Cone Search to allow specification of image size
  - Response includes metadata about images, encoded as VOTable
  - Images are referenced by URL, delivered as FITS for analysis or GIF/JPG, etc., for embedded display
- Simple Spectrum Access Protocol
  - Astronomical spectra have more subtleties and variations in representation than images → access protocol is more complicated
  - Query supports more qualifiers and response adds more metadata, again encoded as VOTable
  - Spectra referenced by URL or encoded in-line in the VOTable

Data access

- Astronomical Data Query Language
  - Standard grammar for database queries
  - Core SQL functions plus astronomy-specific extensions
  - String and XML representations
- OpenSkyNode → Table Access Protocol
  - Standard interface wrapper for relational databases
  - Accepts ADQL or parameterized query
  - "Full" SkyNodes support positional cross-match function
  - OpenSkyQuery portal provides users with interface for understanding database structure and contents and for constructing queries
  - TAP implementations in progress, will supercede SkyNodes

The International Virtual Observatory Alliance

http://ivoa.net

- IVOA began in June 2002
  - Self-organizing
  - No funds of its own, no dues; relies 100% on project participation
  - Rotating chair (18-month term)
- IVOA now has 17 member projects
  - Aggregate funding ~$50M (since inception)
  - Projects range from 2-3 people to ~20 FTE
- Forum for discussion and sharing of experience
- Twice per year “Interoperability” workshops bring together ~100 participants
- Adopted a standards process based on W3C
  - Note
  - Working Draft → Proposed Recommendation → Recommendation
  - IAU endorsement
  - See http://ivoa.net/Documents/

Standards development process

- IVOA charters Working Groups in areas where standards are needed
  - Resource Registry, Semantics, VOTable, VOEvent, VO Query Language, Data Access, Grid/Web Services, Data Models
  - Working Groups work by e-mail, TWiki collaborative web site, and semi-annual technical meetings
  - Leadership is shared among international VO projects
  - Formal standards development governed by W3C-based review and promotion process
- Success comes from strong bottom-up motivation to establish single set of standards for VO
  - No exchange of funds
  - Rotating leadership of IVOA
  - "Right-sized" community
  - Liberal adoption/adaptation of standards from other communities (OAI, SQL, WSDL, SOAP, SSO, etc.)

IVOA documents

IVOA Documents and Standards

http://ivoa.net/Documents/

Directories News STScI IVOA NVO NVO Project on Zoho NWS Glenn Dale NVO Twiki Portal Home Page VizieR

IVOA Documents and Standards

Technical Specifications

Group	Title	Most stable	In progress	Version history
App	Simple Application Messaging Protocol	1.11		1.11 1.11 1.10 1.00
DAL	Simple Cone Search	1.03		1.03 1.02 1.01 1.00
	Simple Image Access	1.0	RFC	1.0 1.0 1.01 1.00
	Simple Line Access	1.0	RFC	1.0 1.0
	Simple Spectral Access	1.04		1.04 1.03 1.02 1.01 1.01 1.00
	Table Access Protocol	1.0	RFC	1.0 1.0 1.00
DaM	Space-Time Coordinate Metadata for the Virtual Observatory (STC)	1.33		1.33 1.31 1.30 1.21 1.20 1.10 1.00
	Data Model for Astronomical DataSet Characterisation	1.13		1.13 1.12 1.12 1.11 1.10 1.00
	Simple Spectral Lines Data Model	1.0	RFC	1.0 1.0
	IVOA Spectral Data Model	1.03		1.03 1.02 1.01 1.01 1.01 1.00
GWS	IVOA Single-Sign-On Profile: Authentication Mechanisms	1.01		1.01 1.01 1.00 1.00
	VOSpace service specification	1.15	2.0	1.15 2.0 1.15 1.15 1.14 1.13 1.12 1.12 1.11 1.10 1.02 1.02 1.01 1.00 1.00
	IVOA Credential Delegation Protocol	1.0	RFC	1.0 1.01 1.01 1.00
	Universal Worker Service	1.0	RFC	1.0 1.0
	IVOA Support Interfaces	1.0		1.0 1.0
	IVOA Web Service Basic Profile	1.00		1.00
ReR	IVOA Identifiers	1.12		1.12 1.11 1.10 1.10 1.10 1.00
	IVOA Registry Interfaces	1.0		1.0 1.0 1.00 1.02 1.01 1.00
	Resource Metadata for the Virtual Observatory	1.12		1.12 1.12 1.10 1.10 1.01 1.01 1.00 1.00
	VOResource: an XML Encoding Schema for Resource Metadata	1.03		1.03 1.02 1.02 1.01 1.00
	VODataService: a VOResource Schema Extension for Describing Collections and Services	1.1	RFC	1.1 1.10
Semantics	An IVOA standard for Unified Content Descriptors	1.10		1.10 1.10 1.06 1.05 1.03
	UCD1+ Controlled Vocabulary	1.23		1.23 1.22 1.21 1.20 1.20 1.11 1.11 1.10 1.02 1.00
	Maintenance of the list of UCD words	1.20		1.20 1.20 1.10 1.00
	Vocabularies in the Virtual Observatory	1.19		1.19 1.18 1.16 1.15 1.13 1.00
SDP	IVOA Document Standards	1.0	1.2	1.2 1.2 1.2 1.2 1.1 1.1 1.0 1.0
VOE	Sky Event Reporting Metadata (VOEvent)	1.11		1.11 1.11 1.10 1.01
VQL	IVOA Astronomical Data Query Language	2.00		2.00 2.00 2.00 1.01 1.00
	IVOA SkyNode Interface	1.01		1.01 1.00
VOT	VOTable Format Specification	1.2		1.2 1.2 1.2 1.20 1.20 1.10 1.00

Maturity level: Recommendation Proposed Recommendation Working Draft

Most stable: New systems should be developed against this version with the highest maturity level.

In progress: Indicates (if any) a new version of the document under development (but with a lower maturity level than its predecessor) and a link to the relevant Request For Comments (RFC).

Group: **Apps** Applications WG **DAL** Data Access Layer WG **DaM** Data Modeling WG **GWS** Grid & Web Services WG **ReR** Resource Registry WG **Semantics** Semantics WG **SDP** Standards & Documents Process WG **VOE** VO Event WG **VOT** VOTable WG **VQL** VO Query Language WG ***n.a.*** not applicable

Done

What does this cost?

- Data management activities at major astronomy facilities are typically 3–5% of annual operating budget, including h/w, s/w, and staff. Staff accounts for ~85% of total.
- VO development and operations are ~20% additional to baseline data management costs (international aggregate)