Astroinformatika

Cesta k pochopení Vesmíru z astronomicky velkých dat

Petr Škoda

Astronomický ústav AVČR Ondřejov

S podporou grantu MŠMT COST LD-15113
akce COST TD-1403

Informatický večer FIT ČVUT Praha
7.11.2016

Credits

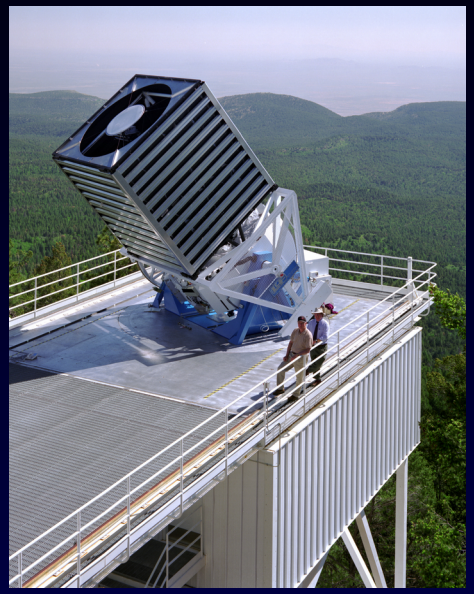
- The presentation is based on many different sources – mainly the on-line published slides from IVOA meetings, slides from Astroinformatics workshops or pictures found on Internet.
- We acknowledge namely materials of E. Solano, E. Hatsiminaoglu, B.Hanish, G. Djorgovski, G. Longo, O. Laurino, T. Hey, L. Fortson and presentations from AI2016 in Sorrento

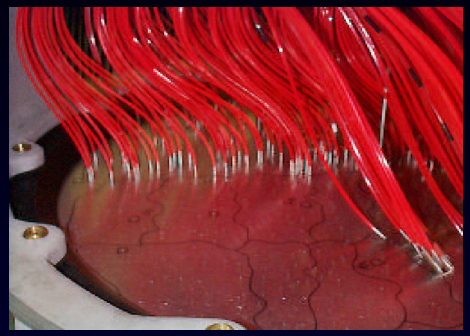
Outline of the Talk

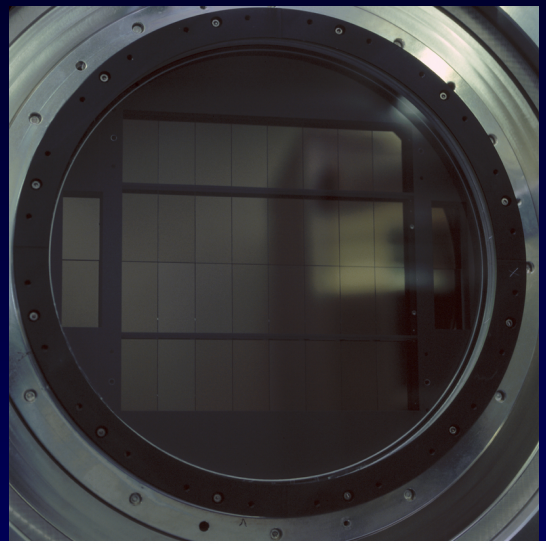
- Data Avalanche in astronomy
- Virtual Observatory
- Astroinformatics
- Visualizations
- Transfer of technology
- Citizen Science

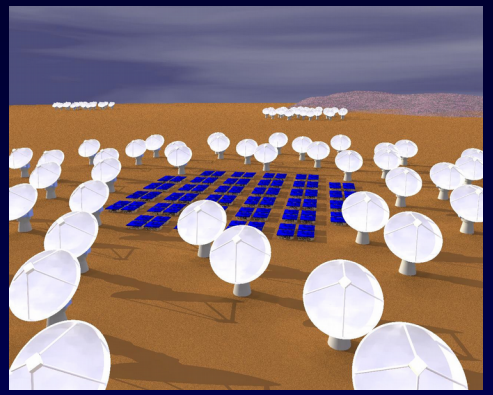
- Astroinformatics in CR

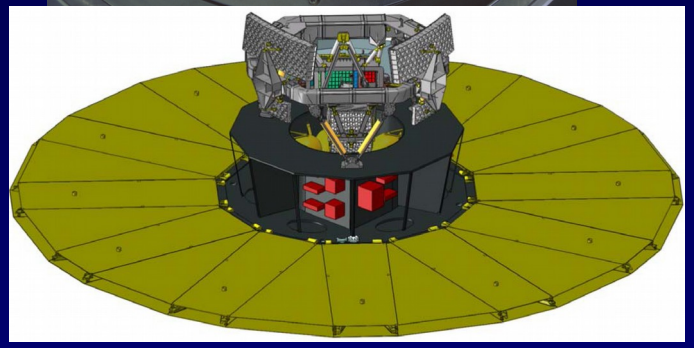
Data Avalanche







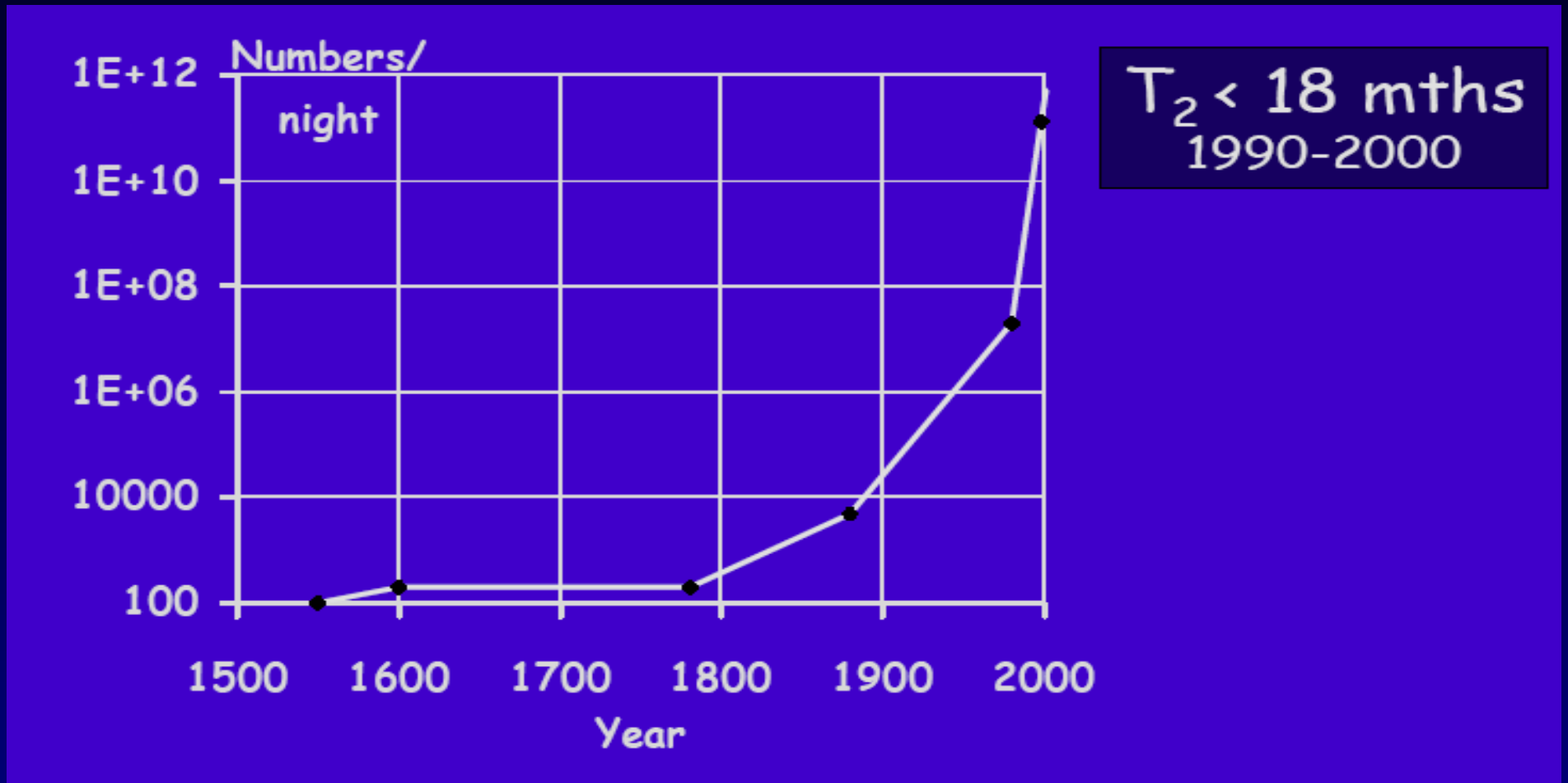




Data Avalanche

Moore law for chips –doubling 1.5 year

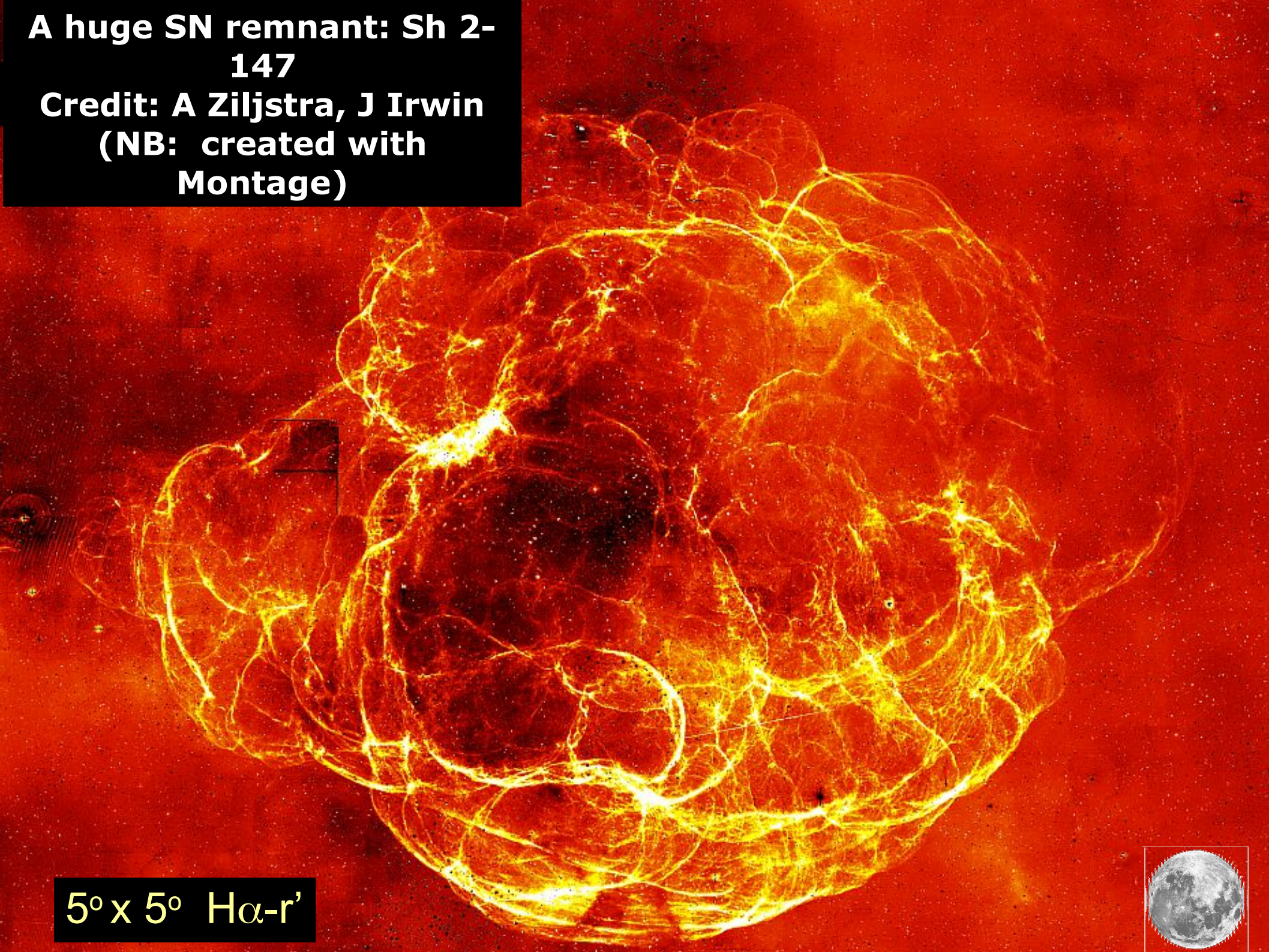
Data in astronomy – doubling < 1 yr ! (1000/10 yr)



CD Sea

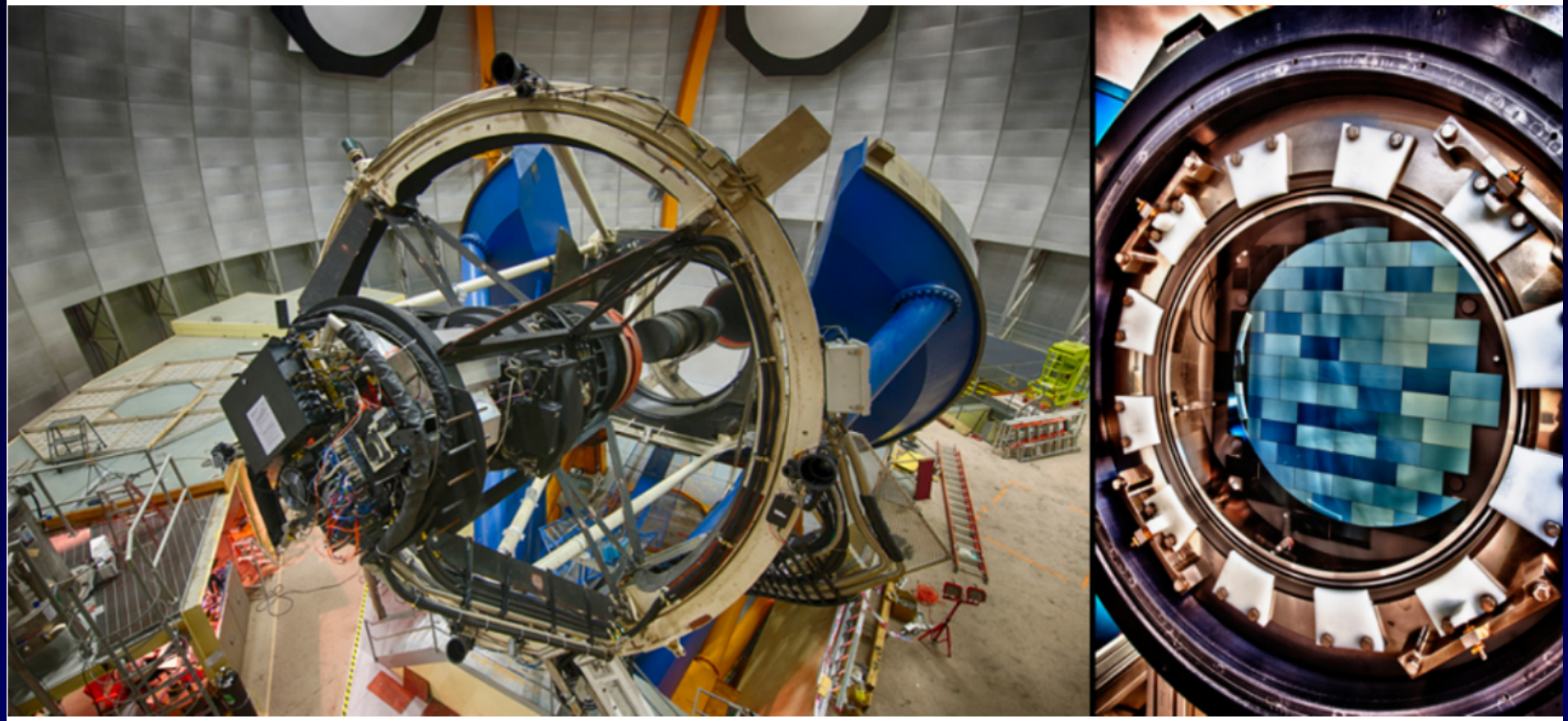
600 000 CD = 372 TB (CD 650MB)
600 000 DVD = 2.5 PB (DVD=4.5GB)

Bruce Monro
Kilmington UK


A huge SN remnant: Sh 2-147
Credit: A Ziljstra, J Irwin
(NB: created with Montage)
5° x 5° Hα-r'

Dark Energy Survey Camera

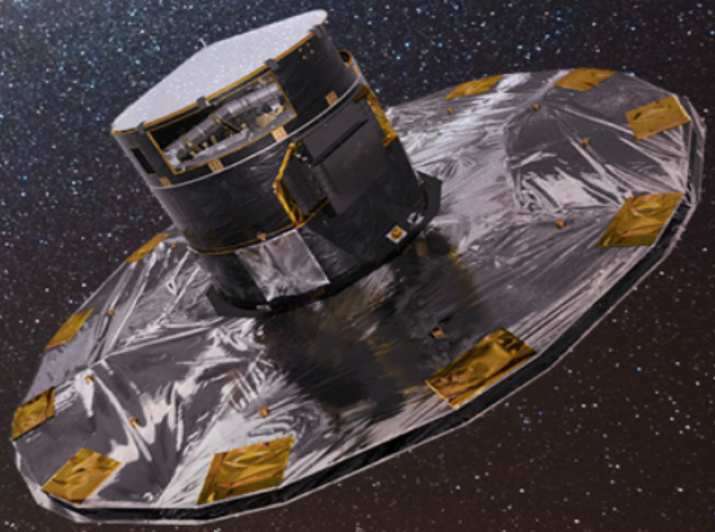
Dark Energy Camera (DECam)



~0.4 PB/yr

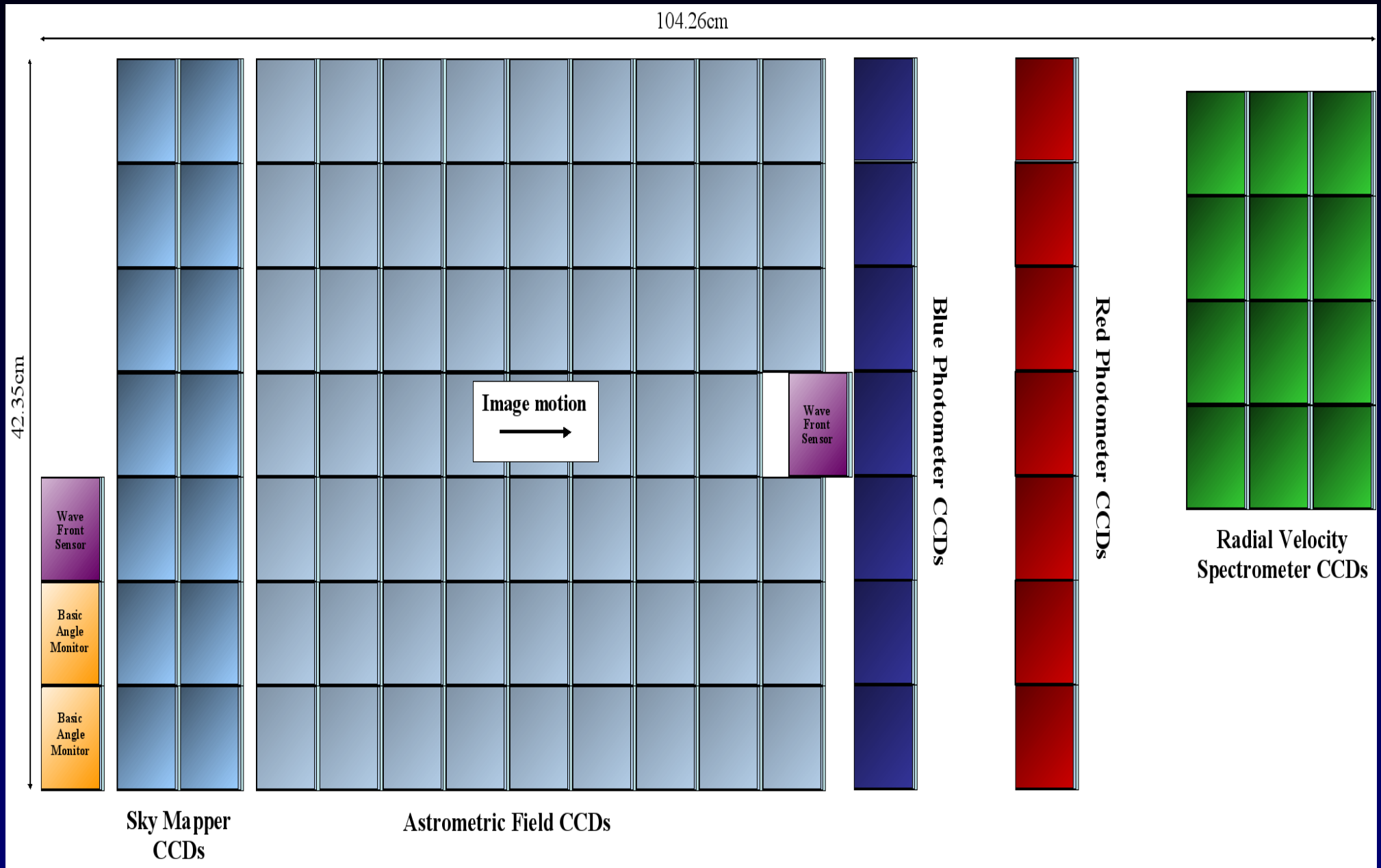
Gaia

- Gaia satellite
  - launched by ESA in december 2013
  - determines positions, velocities and astrophysical parameters of $>10^9$ stars of the Milky Way
  - First catalogue DR1 just out
    - ra, dec, G magnitude
  - DR2 ~1 year
  - Final catalogue ~2020



GAIA CCDs

106 CCDs
938 Mpix
2800cm²

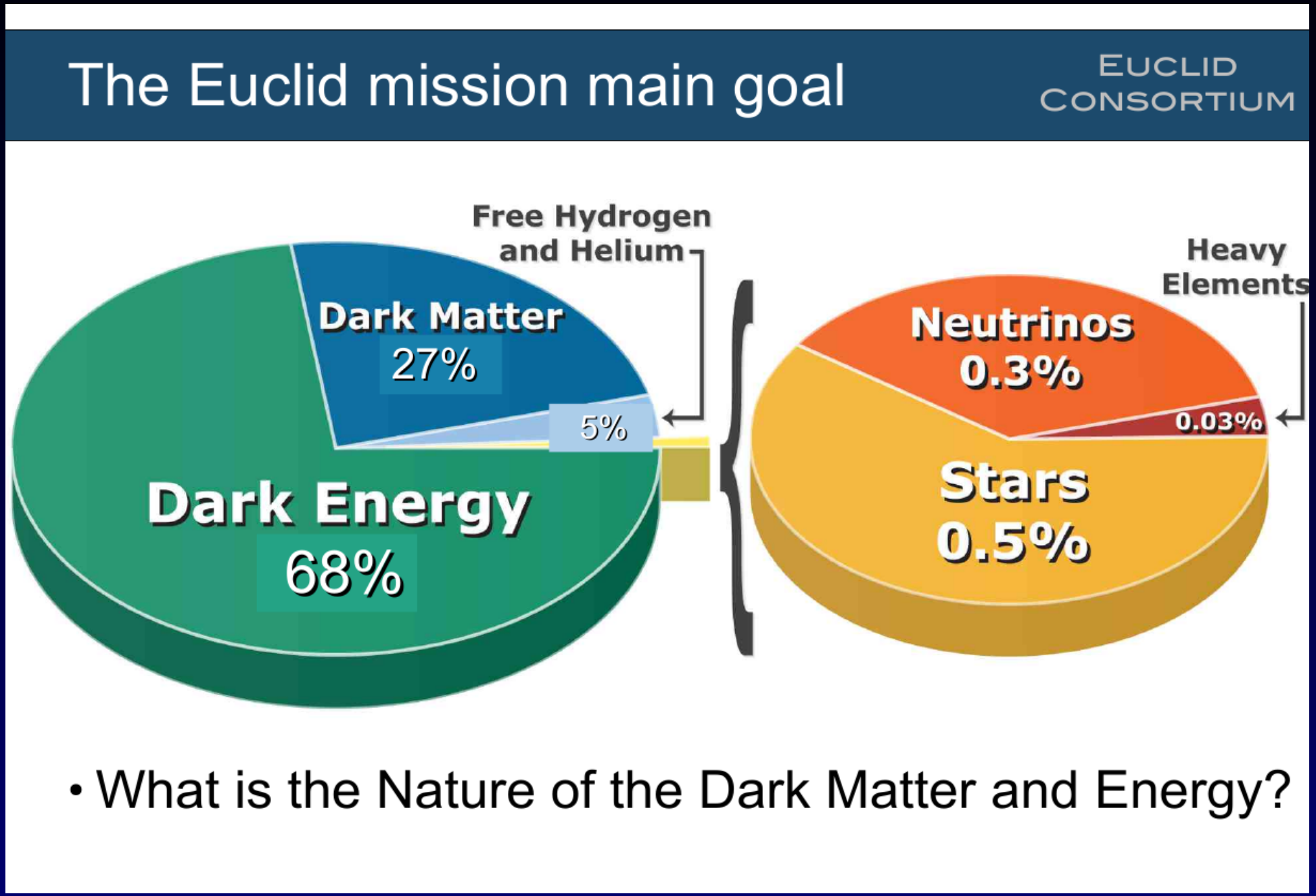


Large Synoptic Survey Telescope

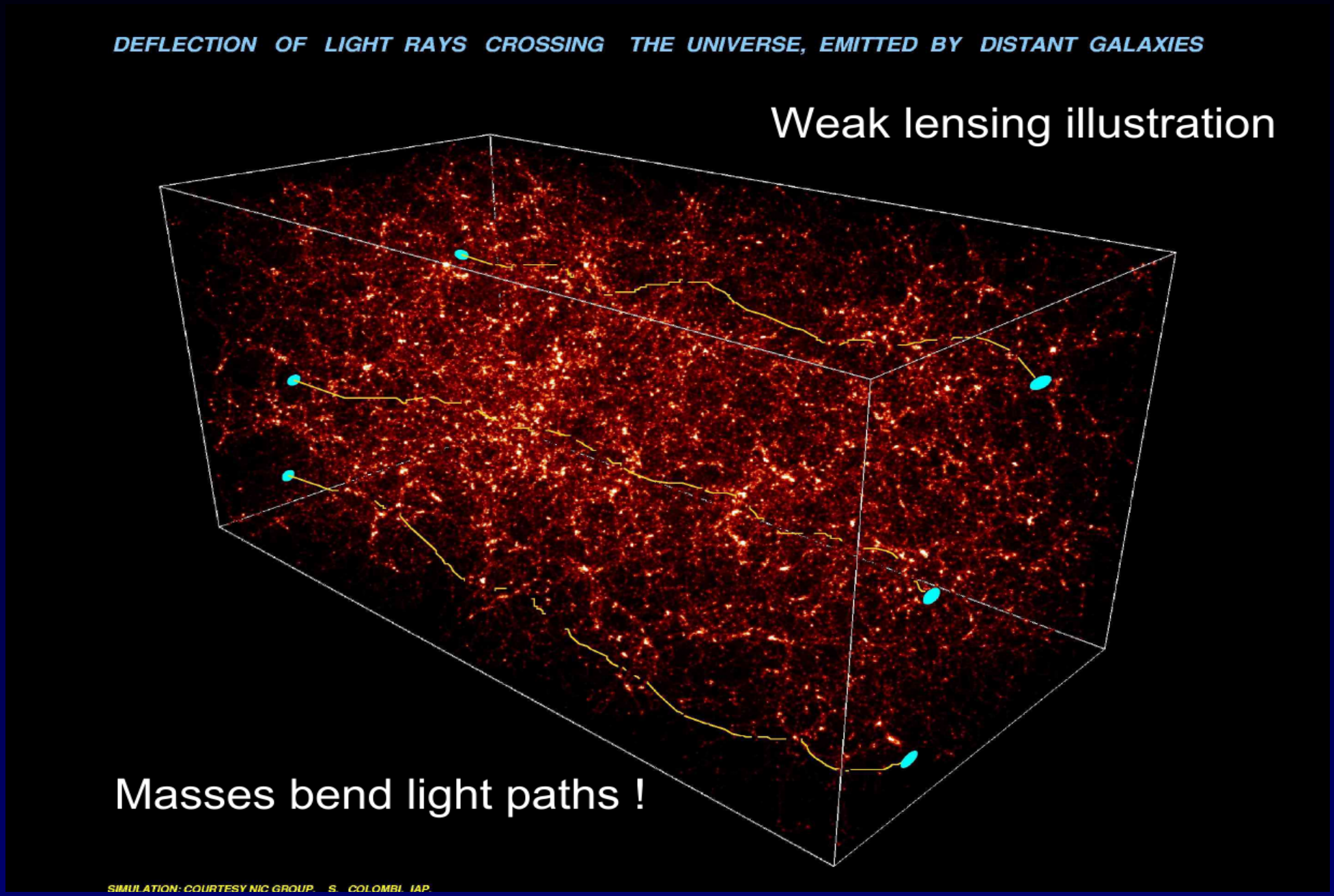
201 CCD 4kx4k,
3.2 Gpix every 20 sec
3.5 deg FOV (64cm)
20 TB/day=6 PB/yr RAW
1.5 PB catalogue !!!
detection of changes 60s!

38 billion objects x 1000
32 tril. meas. -5 PB table

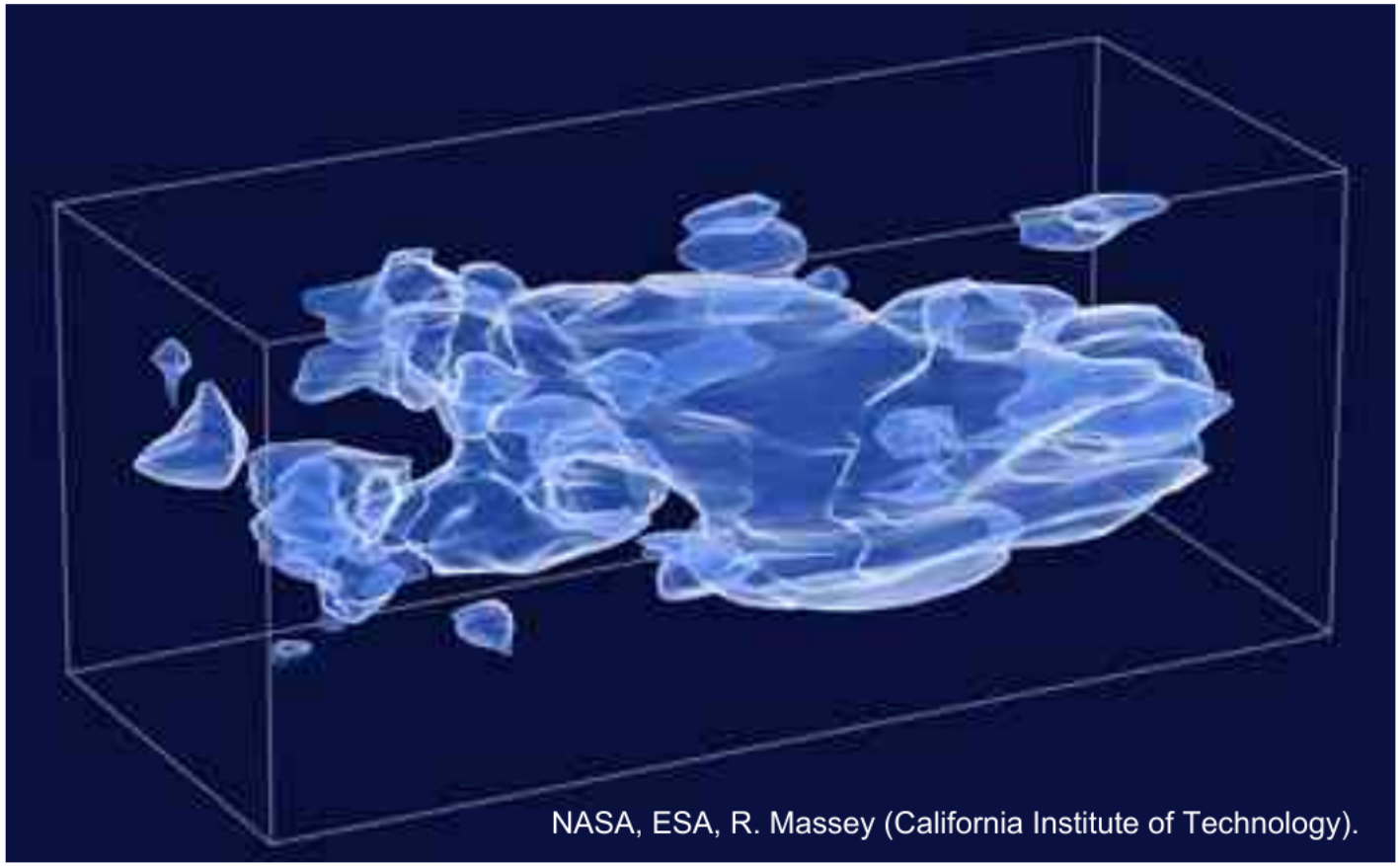
Project EUCLID



EUCLID principles



Euclid Data Archive



	2021	2022	2023	2024	2025	2026	2027
Storage (PB)	15	30	50	60	75	90	90
Computing (kilo cores / year)	2.5	5	8.5	12	16	20	21

Numbers from Christophe Dabin @ tk1

Atacama Large Milimeter Array
ALMA

64 antennas 12m
Chajnator 5000m
Chile
2008-2013

it is spectrograph
as well as ...

0.5-2 PB/yr RAW

LOFAR network

	LOFAR	SKA
Raw Telescope	112 PB/yr	60 EB/yr
Archive Rate	6 PB/yr	100 PB/yr

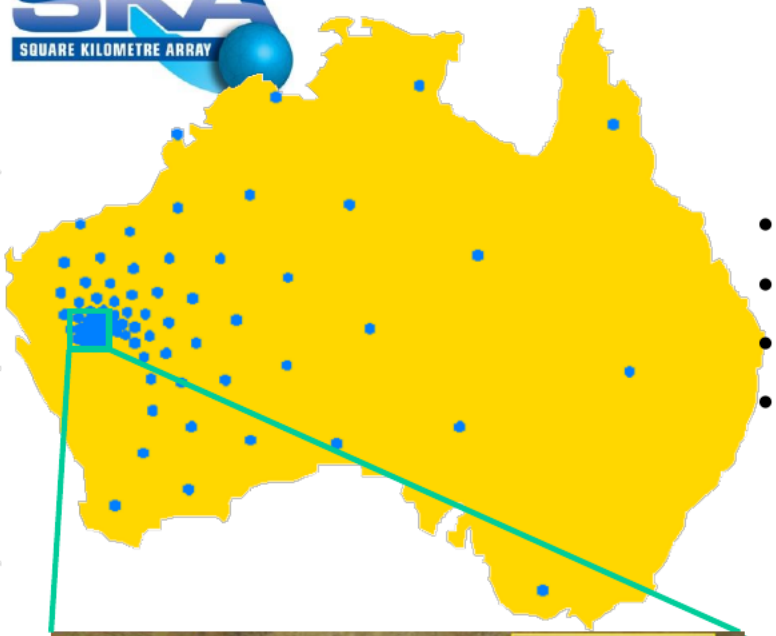
SKA

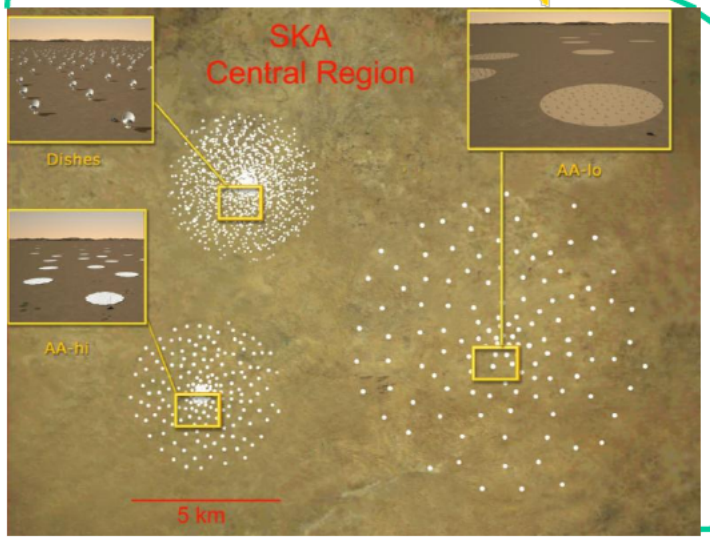
SQUARE KILOMETRE ARRAY

UNIVERSITY OF CAMBRIDGE

also a Continental sized Radio Telescope

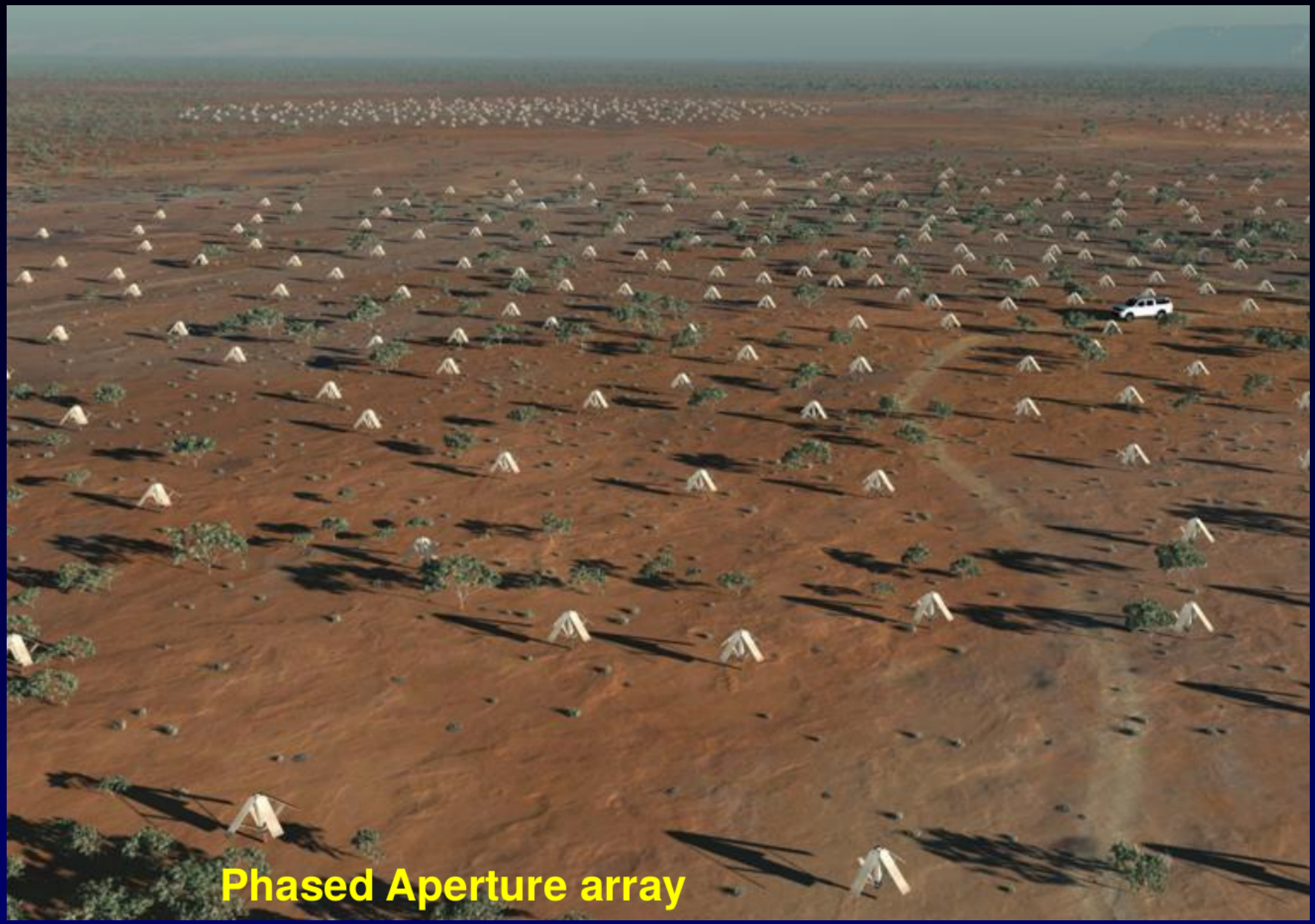
- Need a radio-quiet site
- Very low population density
- Large amount of space
- Possible sites (decision 2012)
  - Western Australia
  - Karoo Desert RSA



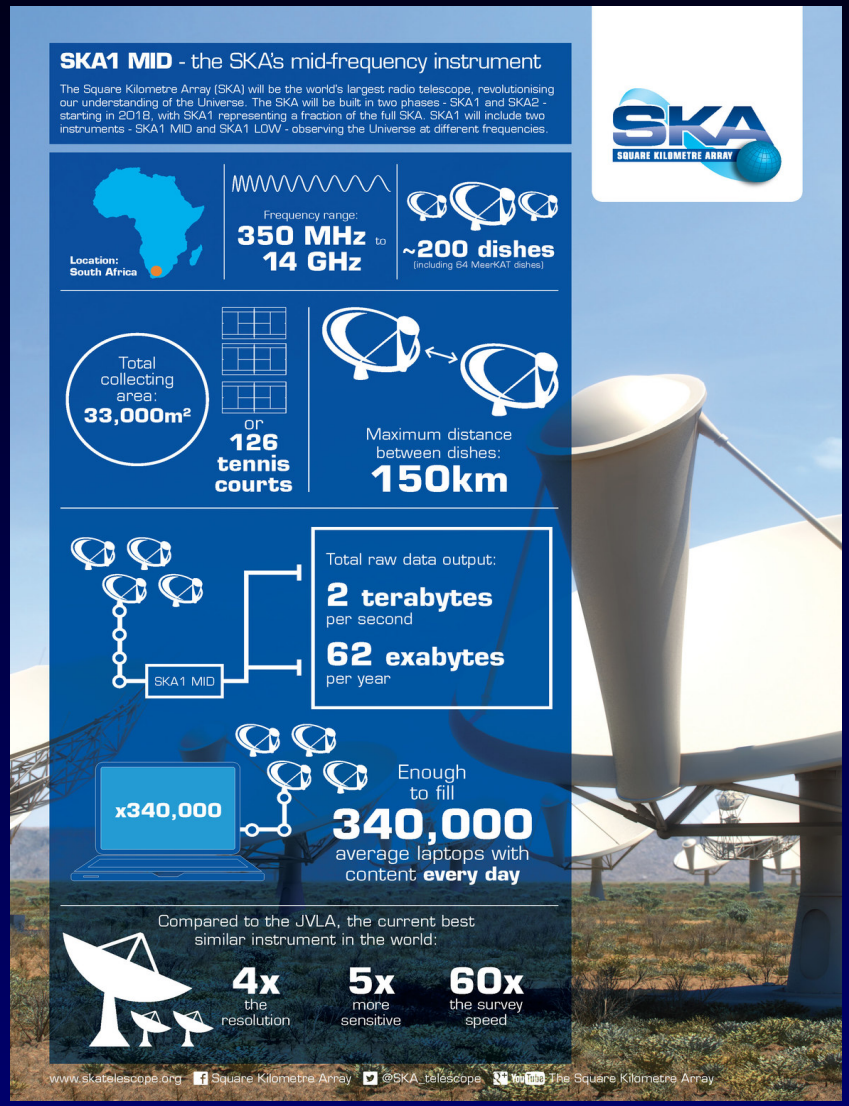


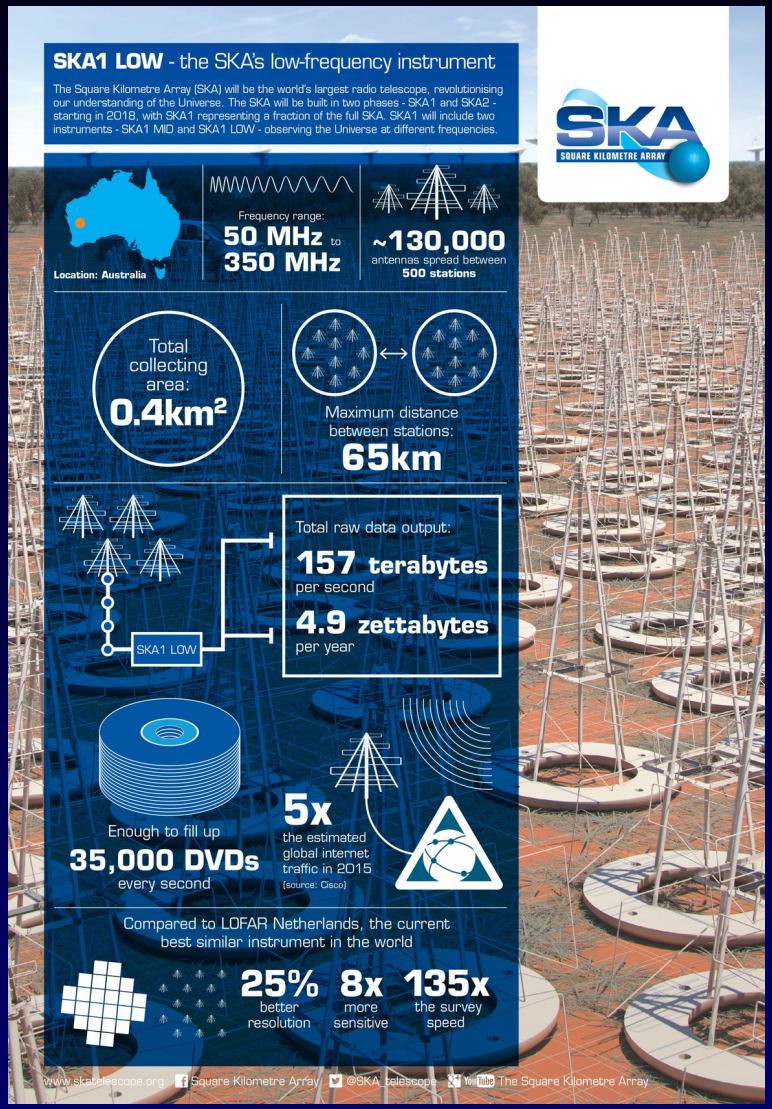
SKA

SKA



Square Kilometer Array
SKA



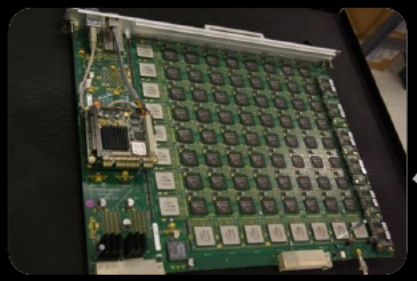


SKA Data Challenge

LOFAR ASTRON

Antennas

Digital Signal Processing (DSP)



To Process is HPC
2020: 100 PBytes/day
2028: 10,000 PBytes/day

Over 10's to 1000's kms

Transfer antennas to DSP
2020: 20,000 PBytes/day
2028: 200,000 PBytes/day

Over 10's to 1000's kms

HPC Processing
2020: 300 PFlop
2028: 30 EFlop

High Performance Computing Facility (HPC)

SKA Processing Challenge

Jodrell Bank Observatory

MANCHESTER 1824

SKA SQUARE KILOMETRE ARRAY — SDP — SCIENCE DATA PROCESSOR

- SKA1-LOW : 41.5 PFlops
- SKA1-MID : 72.1 PFlops

	LOW (50-350M Hz)	MID Band 1 (350-1050 MHz)	MID Band 2 (950-3050 MHz)	MID Band 5 (4.6 to 9.6 GHz)
DD CAL (not in iPython)	18.3	17.4	17.4	17.4
ICAL	4.9	9.5	7.5	6.3
DPrep A+B	4.8	10.8	9.2	6.8
DPrep C	12.0	30.4	23.0	17.4
[illegible]	[illegible]	[illegible]	3.0	2.5
Sustained Compute Load Total (PFLOPS)	41.5 PFLOPS	72.1 PFLOPS	60.2 PFLOPS	50.5 PFLOPS
[illegible]	[illegible]	[illegible]	5.8	4.9
Apparent power, with PUE and power factor (MVA)**	5.8	9.9	8.3	6.9
Hardware CAPEX Estimate (M€)***	57	110	92	77

- SKA1-LOW : 5.8 MW
- SKA1-MID : 9.9 MW

Assuming an efficiency of 25% this means SDP requires **200 - 300 PFlops**

The project power cap for SDP is:

SKA1-LOW : 4 MW
SKA1-MID : 10 MW

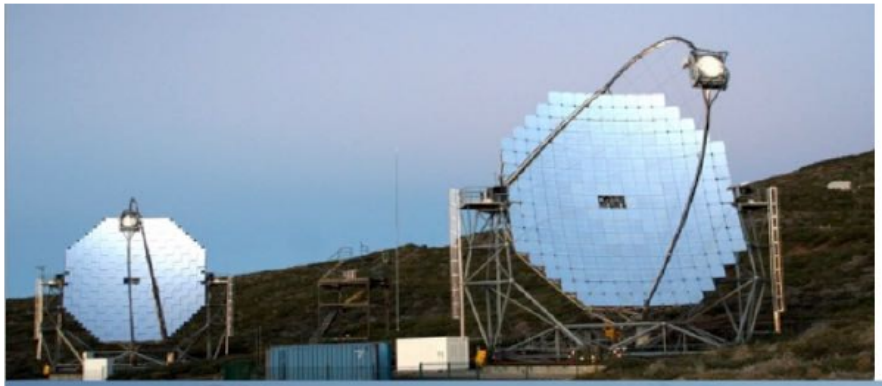
SKA Archive Volumes

- ~0.5 – 10 PB/day of image data
- Source count ~10^6 sources per square degree
- ~10^{10} sources in the accessible SKA sky, 10^4 numbers/record
- ~1 PB for the catalogued data

100 Pbytes – 3 EBytes / year of fully processed data

Cherenkov Telescopes - Current

Currently Operating VHE Instruments



MAGIC: located in La Palma, Spain
Since 2004: single 17m telescope
Since 2009: system of two 17m telescopes

VERITAS: located in Mt Hopkins, Arizona
Since 2007: four 12m telescopes
Since 2012: upgraded PMTs

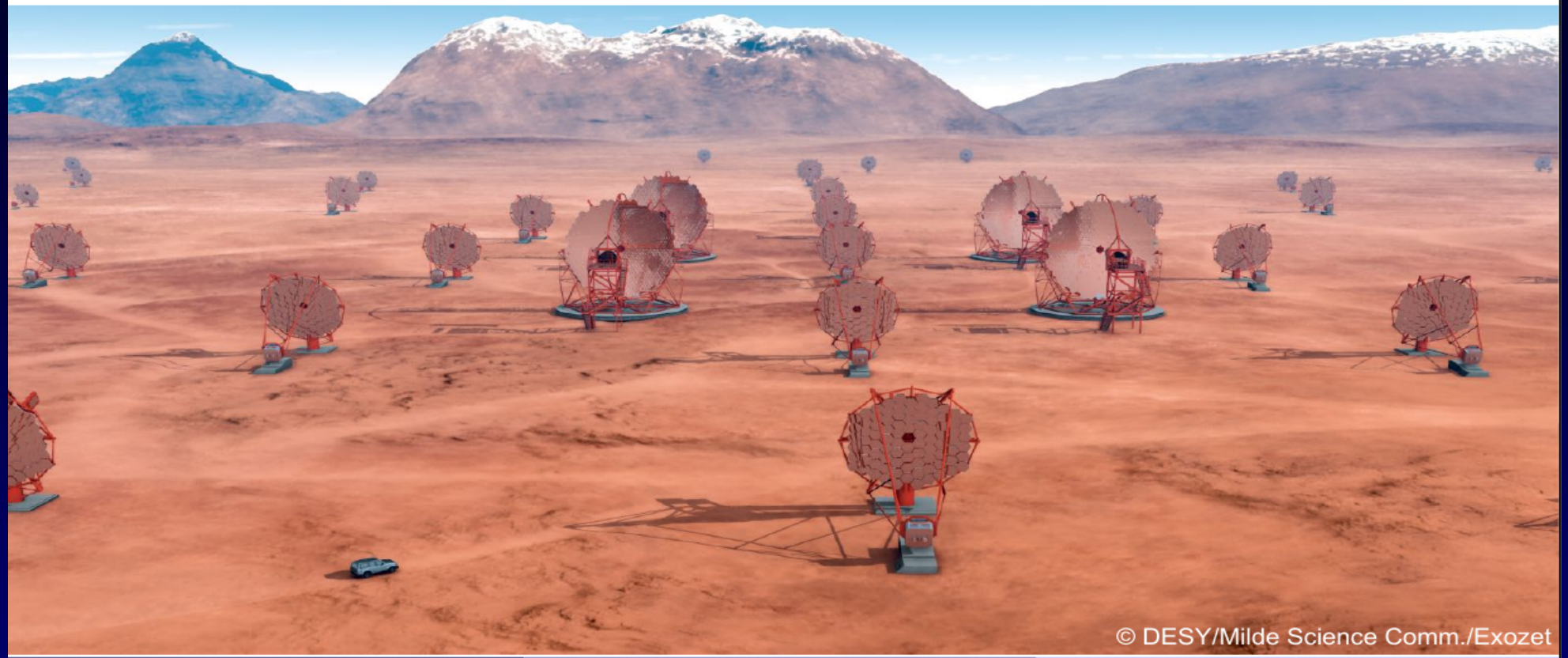
H.E.S.S.: located in Khomas Higlands, Namibia
Since 2002: four 12m telescopes
Since 2012: added 32m by 24m telescope
Since 2015: camera upgrades on 12m telescopes

@ Jeff Grube

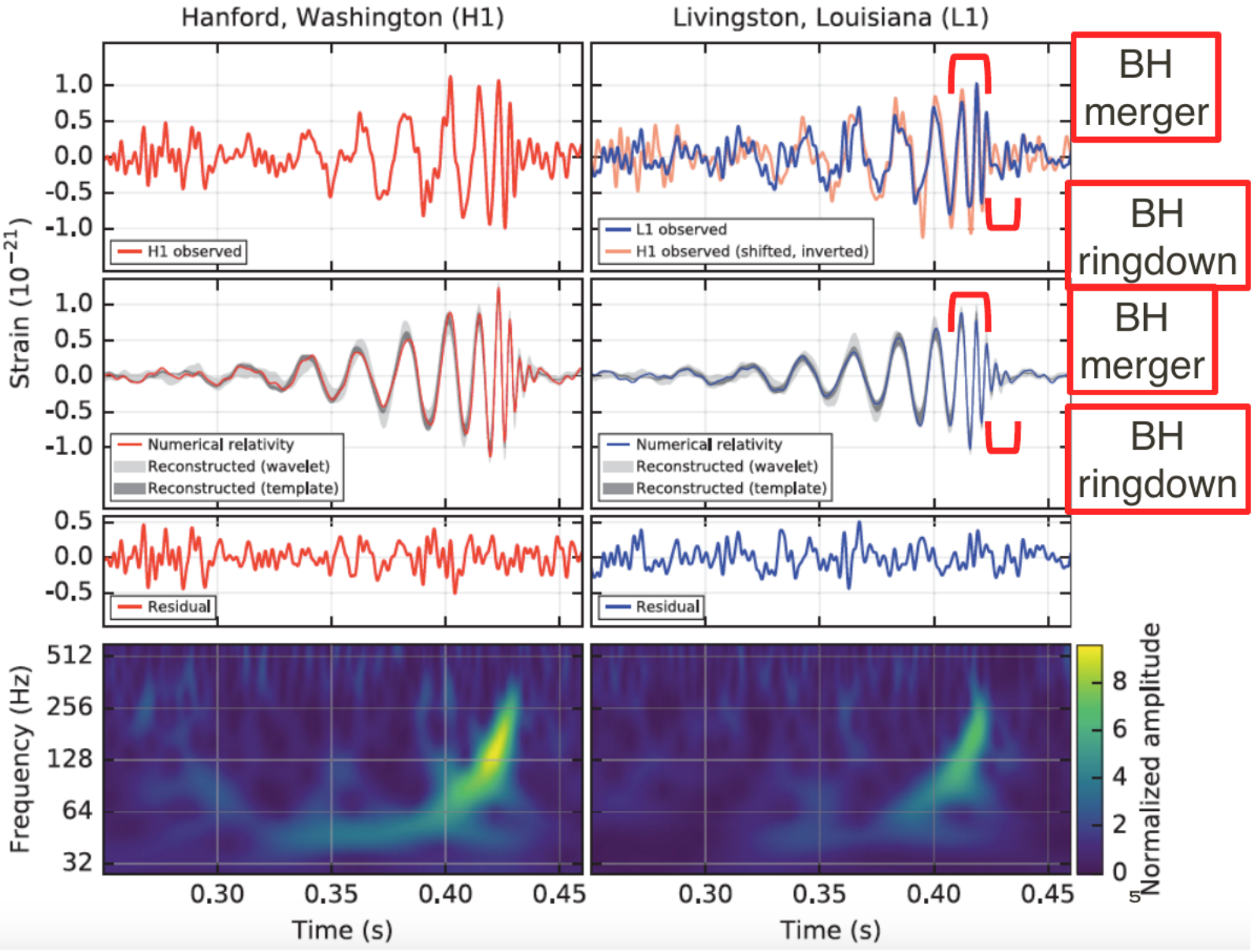
Cherenkov Telescope Array

cta
cherenkov telescope array
Observatory

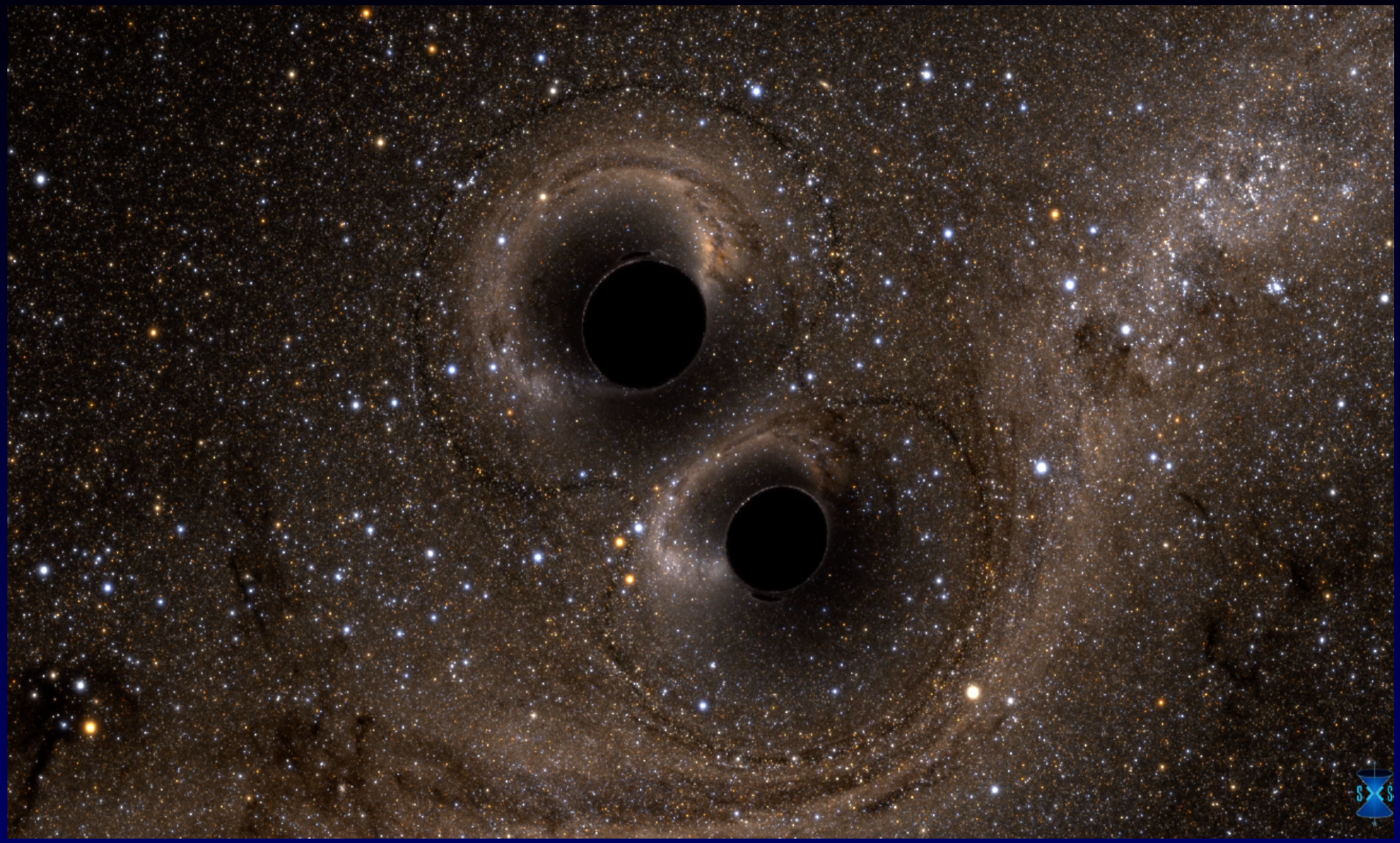
- Two arrays of **100 (South)** et **20 (North)** telescopes
- July 2015: sites selection, Chile (ESO) and La Palma
- 2016: pre-production phase
- 2018-2013: production phase
- Observatory **open** to the community



Gravity Wave First Detection

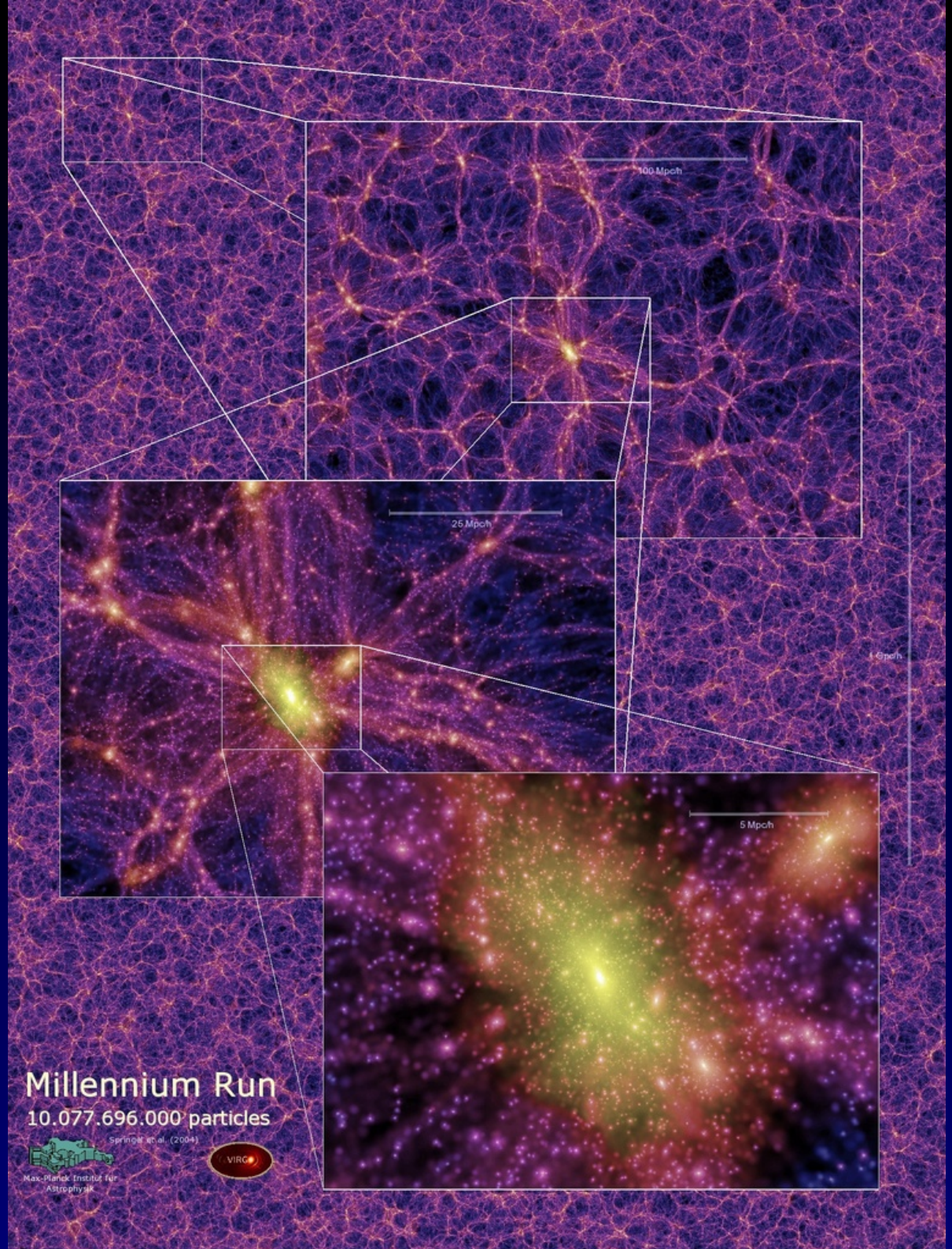


GW150914 BH Merger



BH 30+35 Msun = rotating BH 62 Msun, 3 Msun released in GW , 200ms chirp

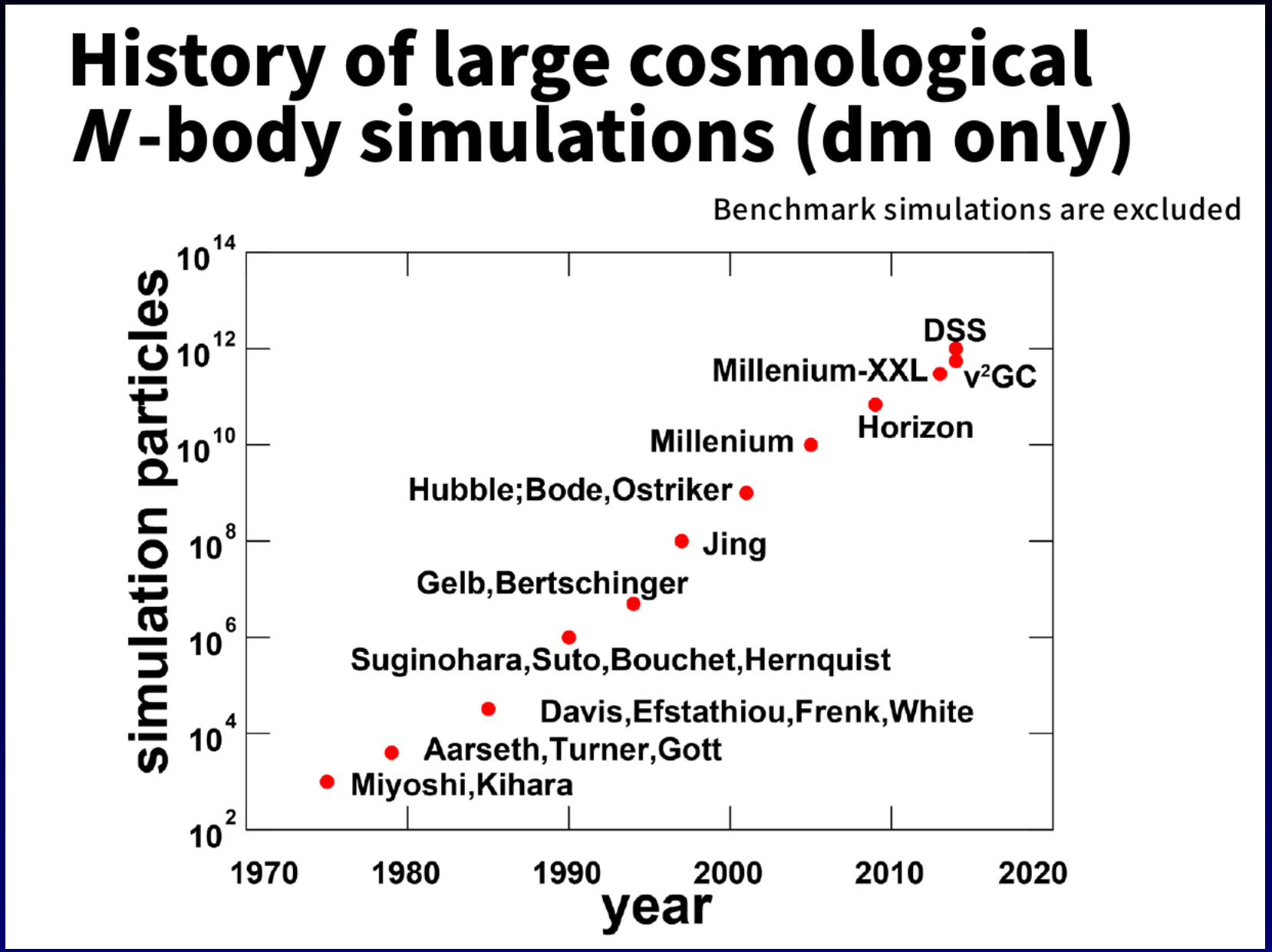
Gravitation Wave Detection Network



Millenium Run

- 10^10 particles
- Several Gpc to
- 10 kpc
- Cube 2 billion ly
- One month MPSSC
- 25 TB
- Evolution of 20 mil galaxies
- Evolution merger tree

Simulations of the Universe



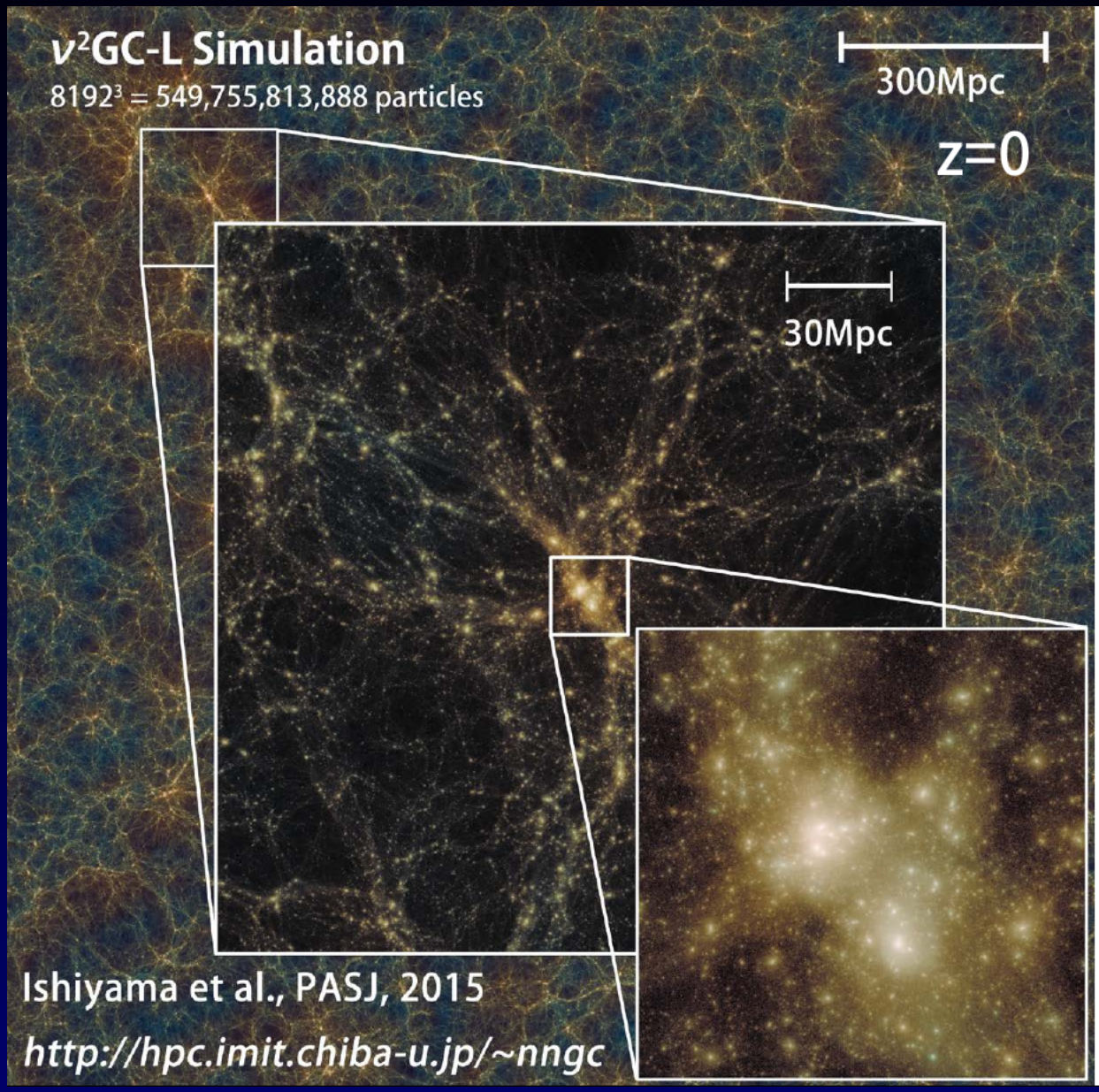
Simulation of the Universe

World's fifth fastest supercomputer

K computer

- SPARC64™ VIIIfx, 2.0GHz octcore (128Gflops / CPU)
  - Total 82944 nodes (663552 CPU core), 10.6 Pflops peak spped
- 16 GB memory / core, Total 1.3PB memory
- 6D torus network

Simulation of Universe



$N = 8192^3 = 549,755,813,888$

$L = 1.12$ Gpc/h
$m = 2.2 \times 10^8$ Msun/h

Planck Cosmology

11x larger volume,
4x better mass res,
compared to
Millennium Run
(Springel+ 2005)

- $N = 2160^3$
- $L = 0.5$ Gpc/h
- $m = 8.6 \times 10^8$ Msun/h
- WMAP1 (σ_8=0.9)

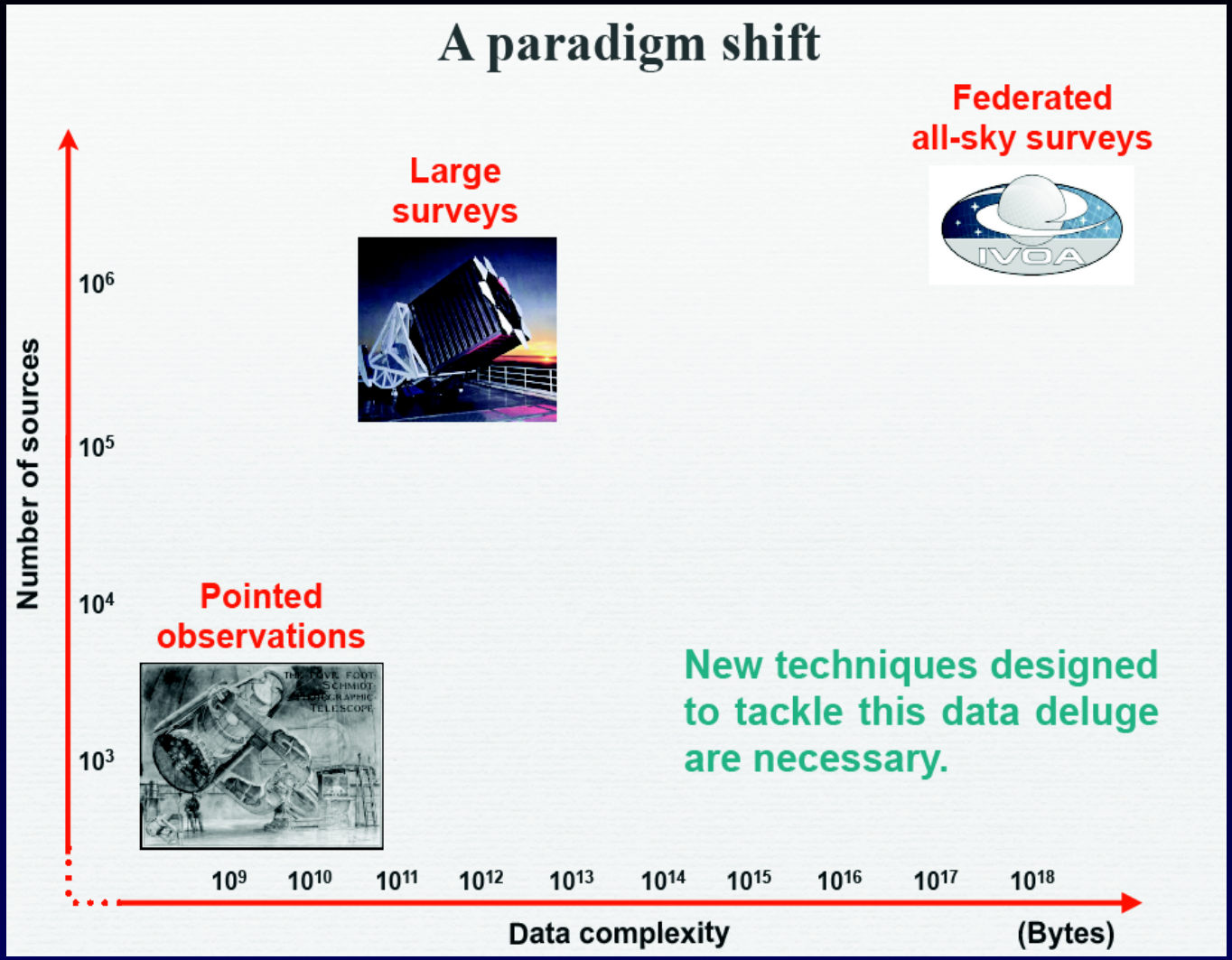
~ 100 hours on
131,072 CPU cores
of K computer

Data size : ~1PB

Problem of 1PB Data Transfer

Data transfer

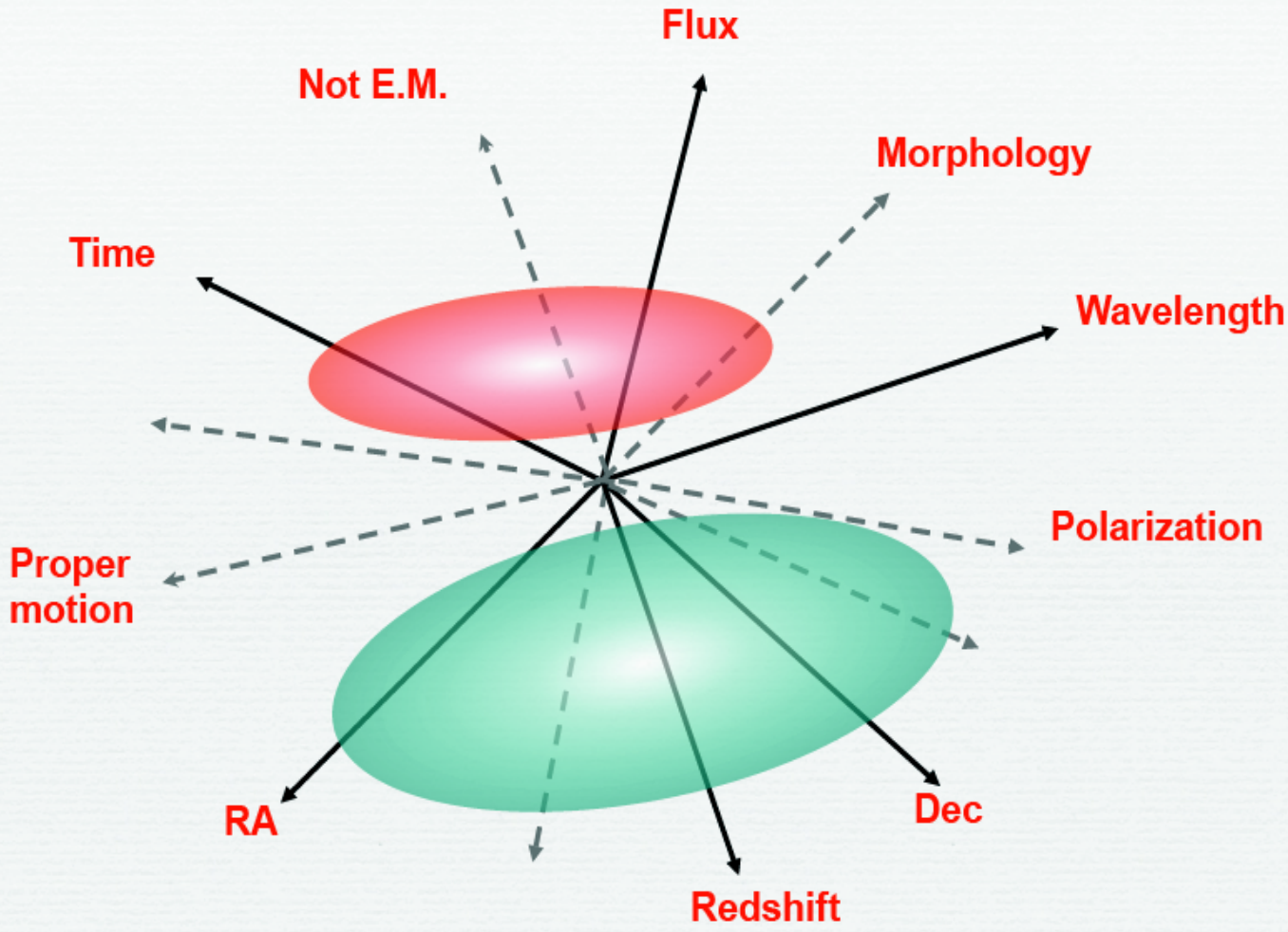
- If 100 Mb/s network is available
  - ~10TB / day
  - ~100 days / 1PB
- Typically, effective speed is less than 10Mb/s
  - < 1TB / day
  - > 3 years / 1PB ……
- **Delivery by car**
  - **3 days / 1PB**
    - From Kobe to Chiba
      (from Kyoto to Tokyo + 100km , ~600km journey)



D'Abrusco 2010

Data analysis at storage place
Move processing = not data !

A growing parameter space



Most discoveries were made in small regions of subspaces or along some of these axes

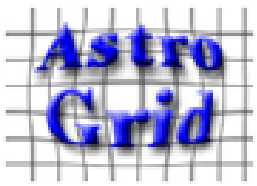
D'Abrusco 2010

Virtual Observatory : Key Definitions

- "*The Virtual Observatory will be a system that allows astronomers to interrogate multiple data centers in a seamless and transparent way, which provides new powerful analysis and visualization tools within that system, and which gives data centers a standard framework for publishing and delivering services using their data*".
- Standardization of data and metadata, and of data exchange methods.
- Registry, listing available services and what can be done with them.

R.J.Hanisch, P.J.Quinn, in "IVOA – Guidelines for participation"

IVOA


Astro
Grid


ArVO
American Virtual Observatory


Aus-VO


BRAVO


China-VO


CVO


esa


France


IVOA


ESO


GERMAN ASTROPHYSICAL
GAVO
VIRTUAL OBSERVATORY


EURO VO


HVO
Hungarian Virtual Observatory


JVO


NOVA


RVO

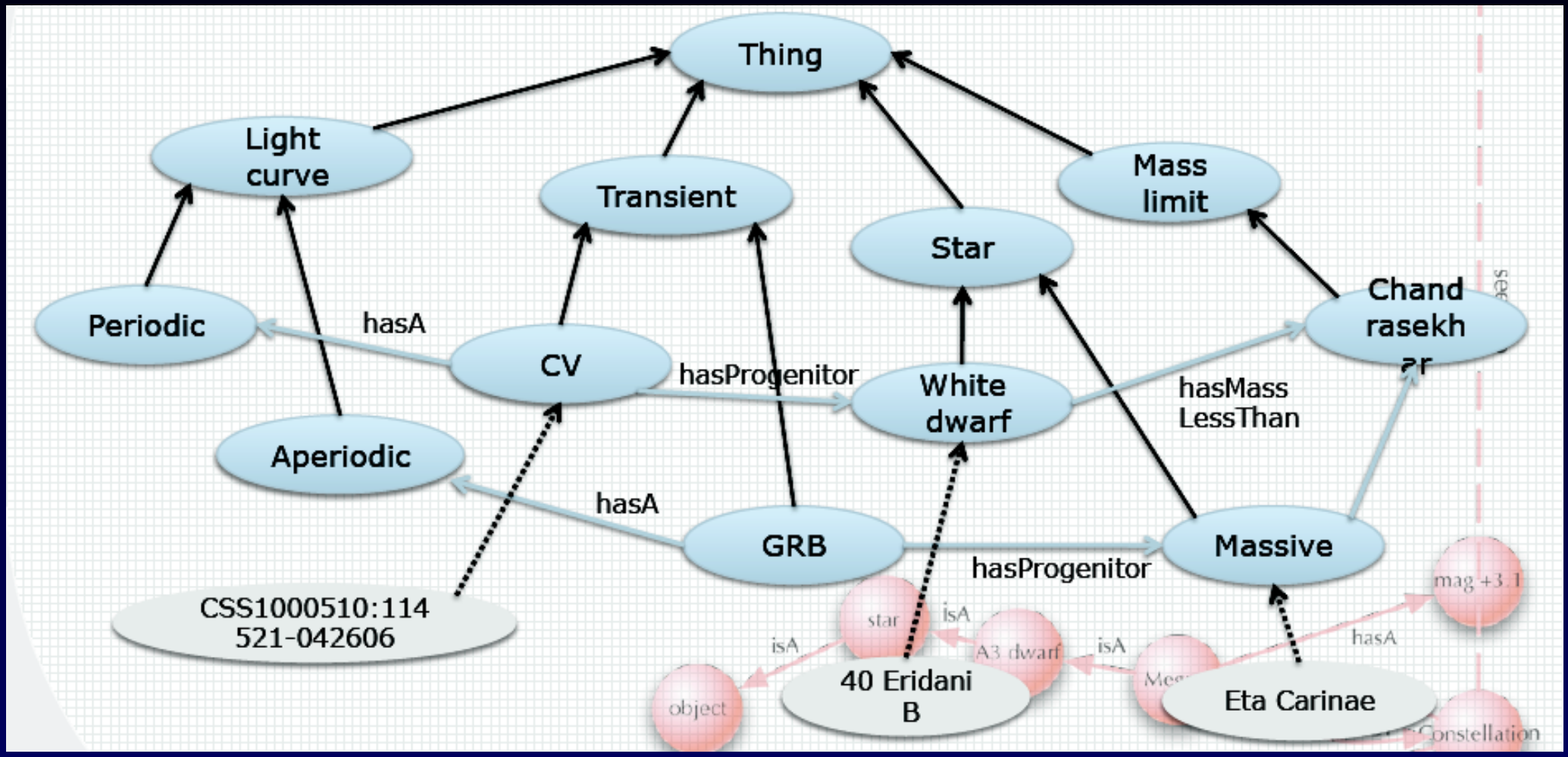

SVO


vo-i


VAO
VIRTUAL ASTRONOMICAL OBSERVATORY


VObs.it

Ontologies in Astronomy

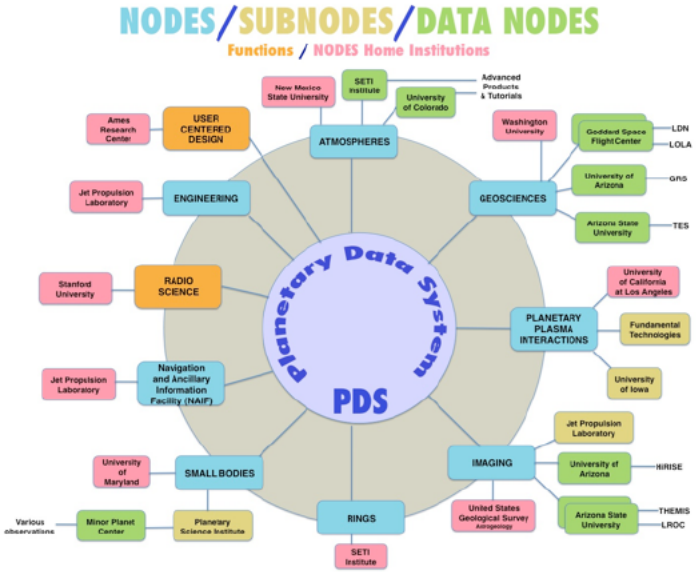


SKOS, RDF standards, search with understanding (not return QSO as binary star)

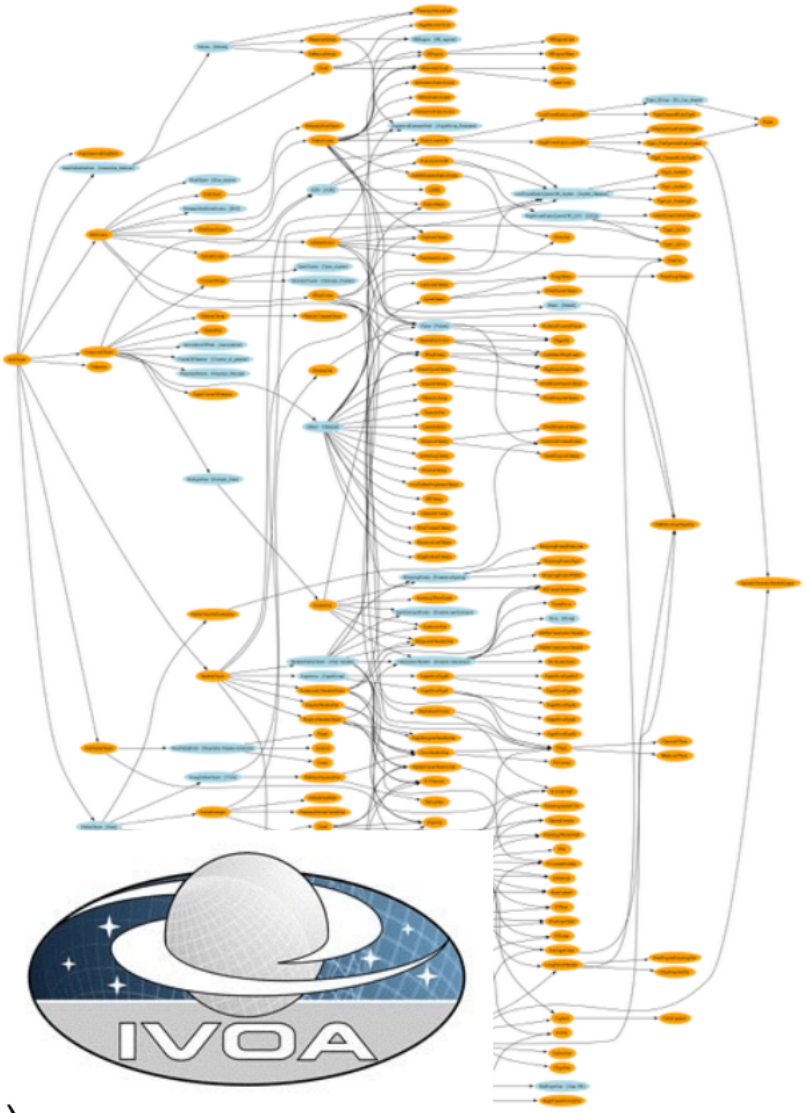
From Graham, M. AI2010

Ontologies

Ontologies



PDS
Steve Hughes
Dan Crichton



PDS -> Earth Science (NASA)

Astronomical Objects

Technology of VO

Unified data format– VOTable, UCD (Vizier)

Transparent transport (unit conversion)

Web services (WS) e-commerce, B2B, J2EE, .Net

VOregistry (DNS like) Google for data+WS protocols

- ConeSearch (searching in circle on sky)
- SIAP (Simple Image Access Protocol)
- SSAP(Simple Spectral Access Protocol)
- SLAP(Simple Line Access Protocol)
- TAP (Table Access Protocol)
- VOEVENT (transients, robotic telescopes,Sun)
- more – datacubes, on-the-fly data generation....

Technology of VO

ADQL (Astronomical Data Query Language)

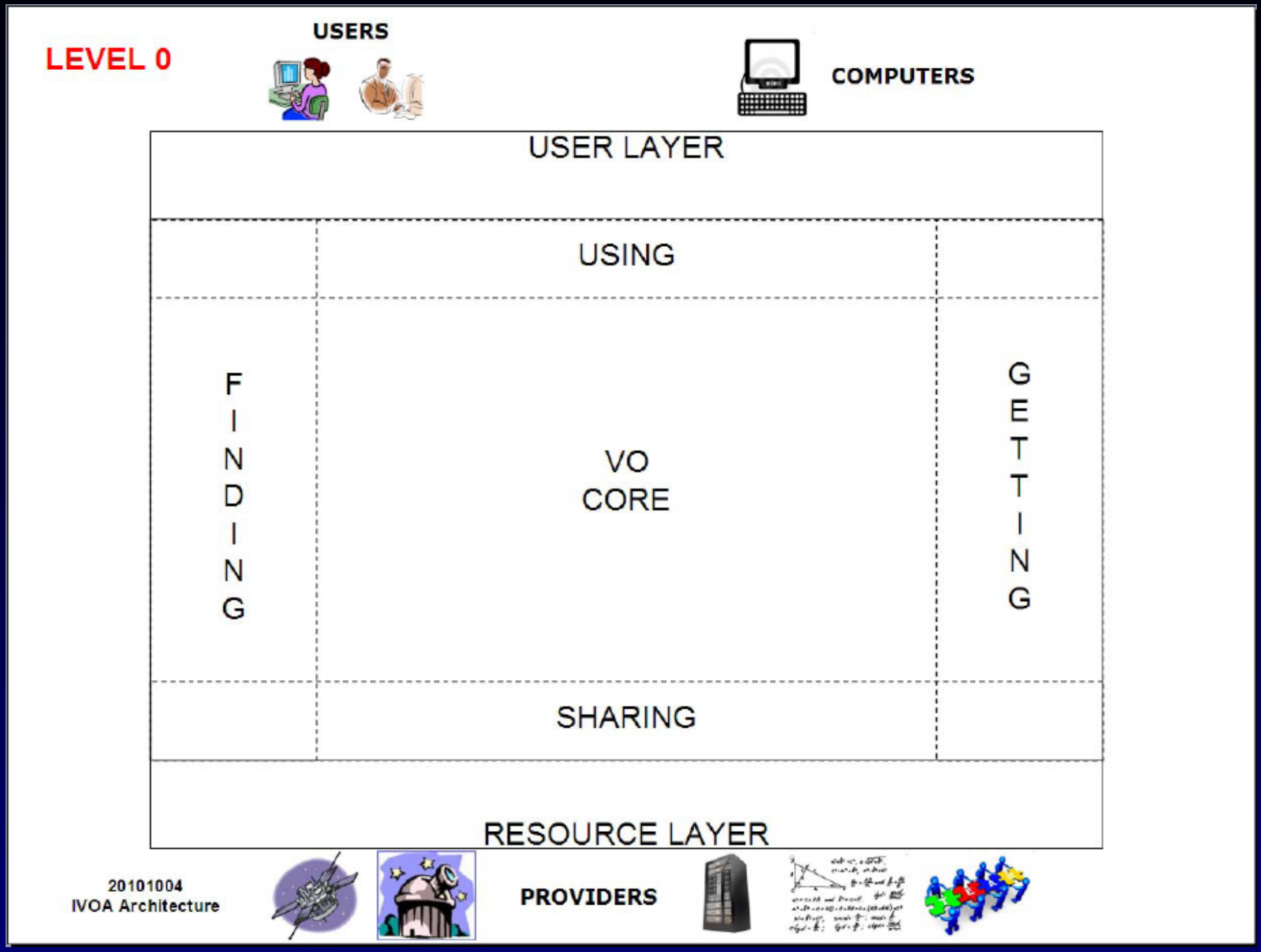
- XMATCH, REGION (2 catalogues - shifted)

Application interoperabilty – (PLASTIC), SAMP

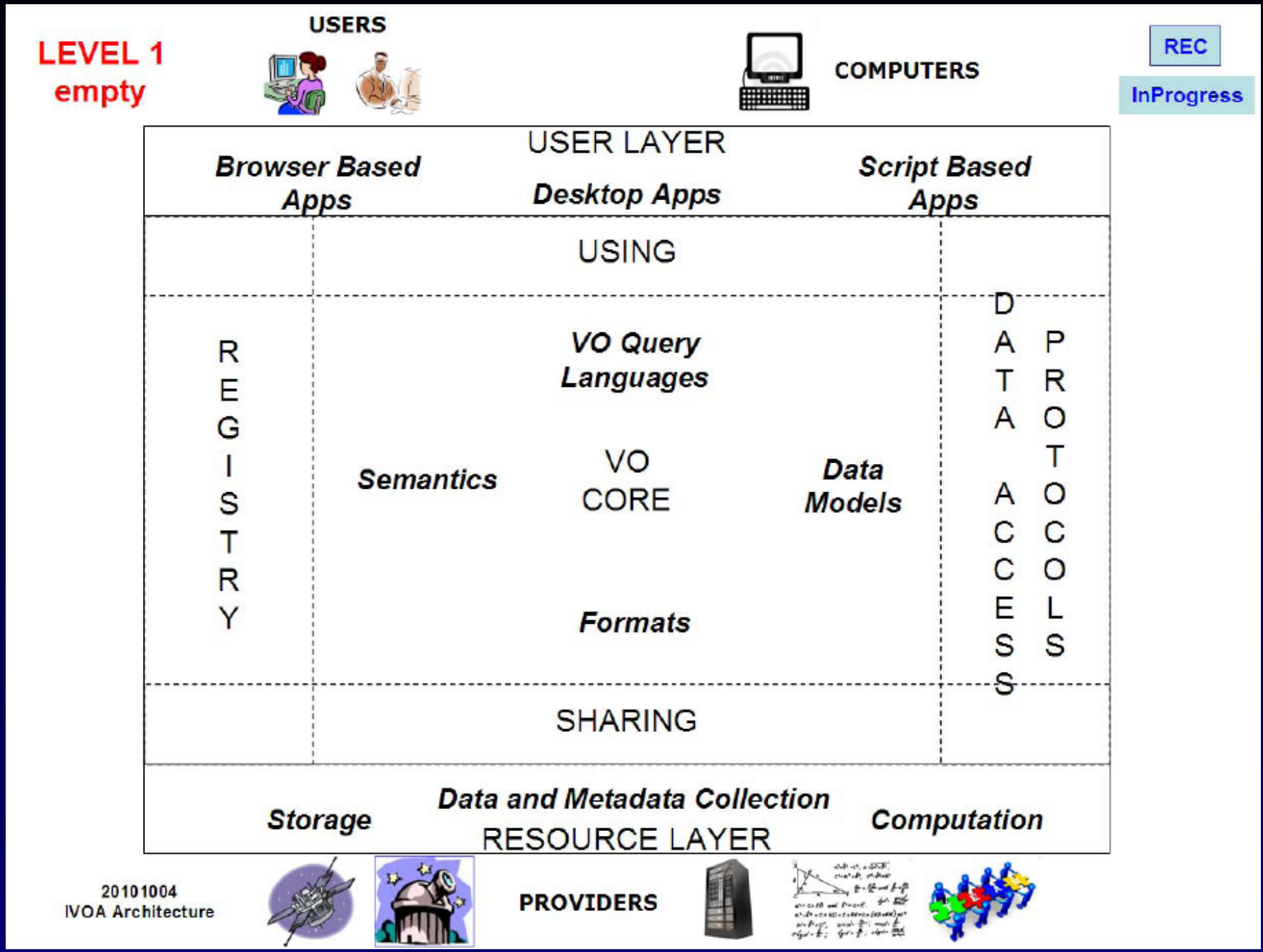
- Allows develop applications as bricks
- sending VOTABLES (catalogue-spectra-images)

Commercial interest (GoogleSky, MS WWT)

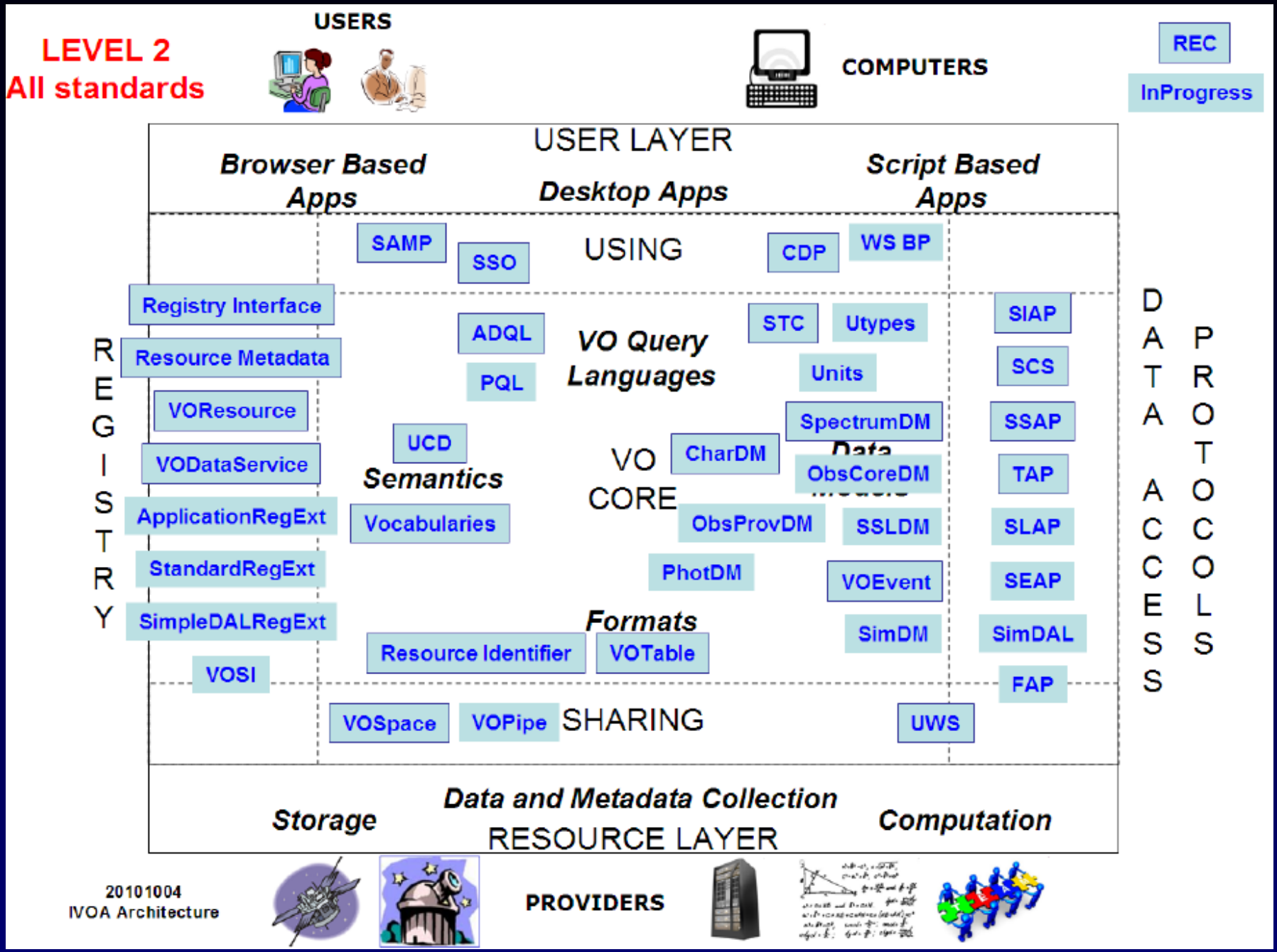
Ecosystem of VO – level 0



Ecosystem of VO – level 1



Ecosystem of VO – level 2



FITS standard

>30 years, separation of metadata (human readable and data)

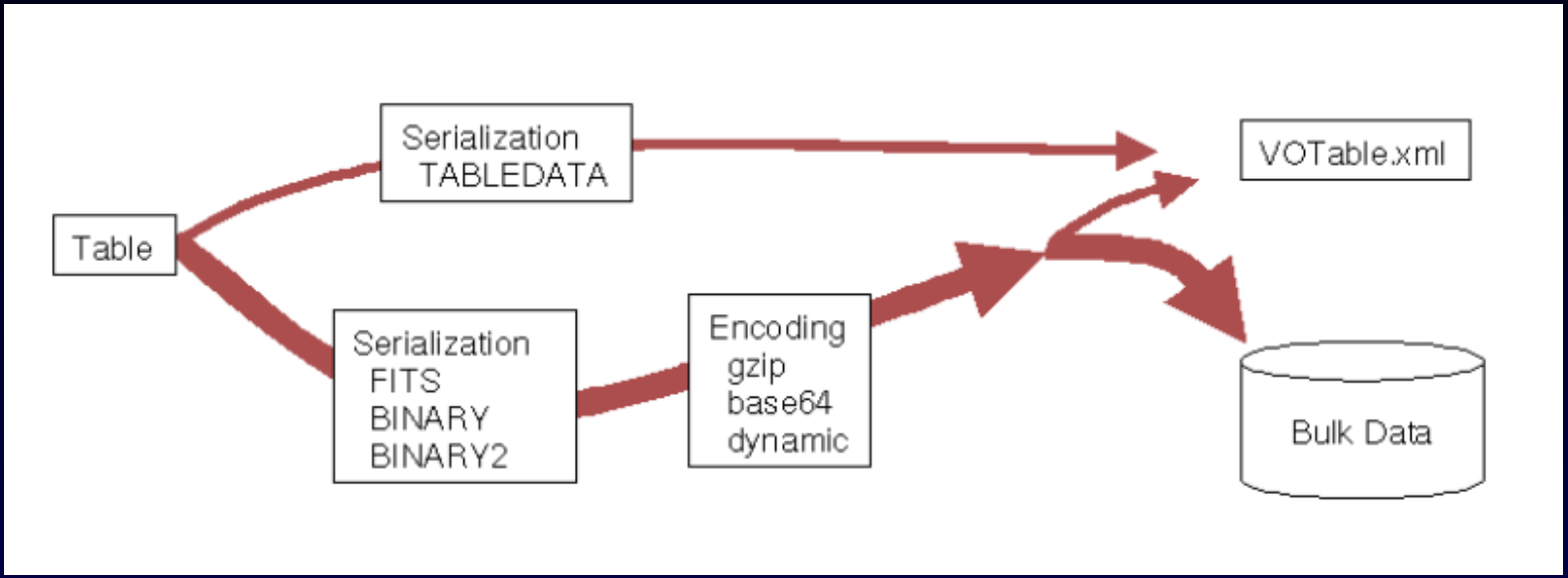
```
SIMPLE  =                    T / file does conform to FITS standard
BITPIX  =                   16 / number of bits per data pixel
NAXIS   =                    2 / number of data axes
NAXIS1  =                 2048 / length of data axis 1
NAXIS2  =                 2048 / length of data axis 2
EXTEND  =                    T / FITS dataset may contain extensions
COMMENT   FITS (Flexible Image Transport System) format is defined in 'Astronomy
COMMENT   and Astrophysics', volume 376, page 359; bibcode: 2001A&A...376..359H
BZERO   =                32768
BSCALE  =                    1 / REAL=TAPE*BSCALE+BZERO
ORIGIN  = 'PESO    '           / AsU AV CR Ondrejov
OBSERVAT= 'ONDREJOV'           / Name of observatory (IRAF style)
LATITUDE=             49.91056 / Telescope latitude  (degrees), +49:54:38.0
LONGITUD=             14.78361 / Telescope longitud  (degrees), +14:47:01.0
HEIGHT  =                  528 / Height above sea level [m].
TELESCOP= 'ZEISS-2m'           / 2m Ondrejov observatory telescope
GAIN    =                    2 / Electrons per ADU
READNOIS=                   10 / Readout noise in electrons per pix
TELSYST = 'COUDE   '           / Telescope setup - COUDE or CASSegrain
INSTRUME= 'OES     '           / Coude echelle spectrograph
CAMERA  = 'VERSARRAY 2048B'    / Camera head name
DETECTOR= 'EEV 2048x2048'      / Name of the detector
CHIPID  = 'EEV 42-40-1-368'    / Name of CCD chip
```

VOTable

```
<TABLE name="SpectroLog">
<FIELD name="Target" ucd="meta.id" datatype="char" arraysize="30*"/>
<FIELD name="Instr" ucd="instr.setup" datatype="char" arraysize="5*"/>
<FIELD name="Dur" ucd="time.expo" datatype="int" width="5" unit="s"/>
<FIELD name="Spectrum" ucd="meta.ref.url" datatype="float" arraysize="*"
      unit="mW/m2/nm" type="location">
<DESCRIPTION>Spectrum absolutely calibrated</DESCRIPTION>
<LINK type="location"
      href="http://ivoa.spectr/server?obsno="/>
</FIELD>
<DATA><TABLEDATA>
<TR><TD>NGC6543</TD><TD>SWS06</TD><TD>2028</TD><TD>01301903</
TD></TR>
<TR><TD>NGC6543</TD><TD>SWS07</TD><TD>2544</TD><TD>01302004</
TD></TR>
</TABLEDATA></DATA>
</TABLE>
```

Serialization (metadata first, end of data unknown, tree structure)

VOTable Serialization




```
<RESOURCE>
 <PARAM name="EPOCH" datatype="float" value="1999.987">
  <DESCRIPTION> Original Epoch of the coordinates</DESCRIPTION>
 </PARAM>
 <PARAM name="TELESCOP" datatype="char" arraysize="*" value="VTel" />
 <INFO name="HISTORY">
   The very first Virtual Telescope observation made in 2002
 </INFO>
 <TABLE>
  <FIELD  (insert field metadata here) />
  <DATA><FITS extnum="2">
   <STREAM encoding="gzip" href="ftp://archive.cacr.caltech.edu/myfile.fit.gz"/>
  </FITS></DATA>
 </TABLE>
</RESOURCE>
```

Universal Content Descriptors

S	em.IR	Infrared part of the spectrum
S	em.IR.J	Infrared between 1.0 and 1.5 micron
S	em.IR.H	Infrared between 1.5 and 2 micron
S	em.IR.K	Infrared between 2 and 3 micron
S	em.IR.3-4um	Infrared between 3 and 4 micron
S	em.IR.4-8um	Infrared between 4 and 8 micron
S	em.IR.8-15um	Infrared between 8 and 15 micron
S	em.IR.15-30um	Infrared between 15 and 30 micron
S	em.IR.30-60um	Infrared between 30 and 60 micron
S	em.IR.60-100um	Infrared between 60 and 100 micron

S	pos.eq	Equatorial coordinates
Q	pos.eq.dec	Declination in equatorial coordinates
Q	pos.eq.ha	Hour-angle
Q	pos.eq.ra	Right ascension in equatorial coordinates
Q	pos.eq.spd	South polar distance in equatorial coordinates
S	pos.errorEllipse	Positional error ellipse
Q	pos.frame	Reference frame used for positions (FK5, ICRS,..)
S	pos.galactic	Galactic coordinates
Q	pos.galactic.lat	Latitude in galactic coordinates
Q	pos.galactic.lon	Longitude in galactic coordinates

P	stat.stdev	Standard deviation
S	stat.uncalib	Qualifier of a generic incalibrated quantity
Q	stat.value	Miscellaneous statistical value
P	stat.variance	Variance
P	stat.weight	Statistical weight
Q	time	Time, generic quantity in units of time or date
Q	time.age	Age
Q	time.creation	Creation time/date (of dataset, file, catalogue,...)
Q	time.crossing	Crossing time
Q	time.duration	Interval of time describing the duration of a generic event or phenomenon
Q	time.end	End time/date of a generic event

Characterization

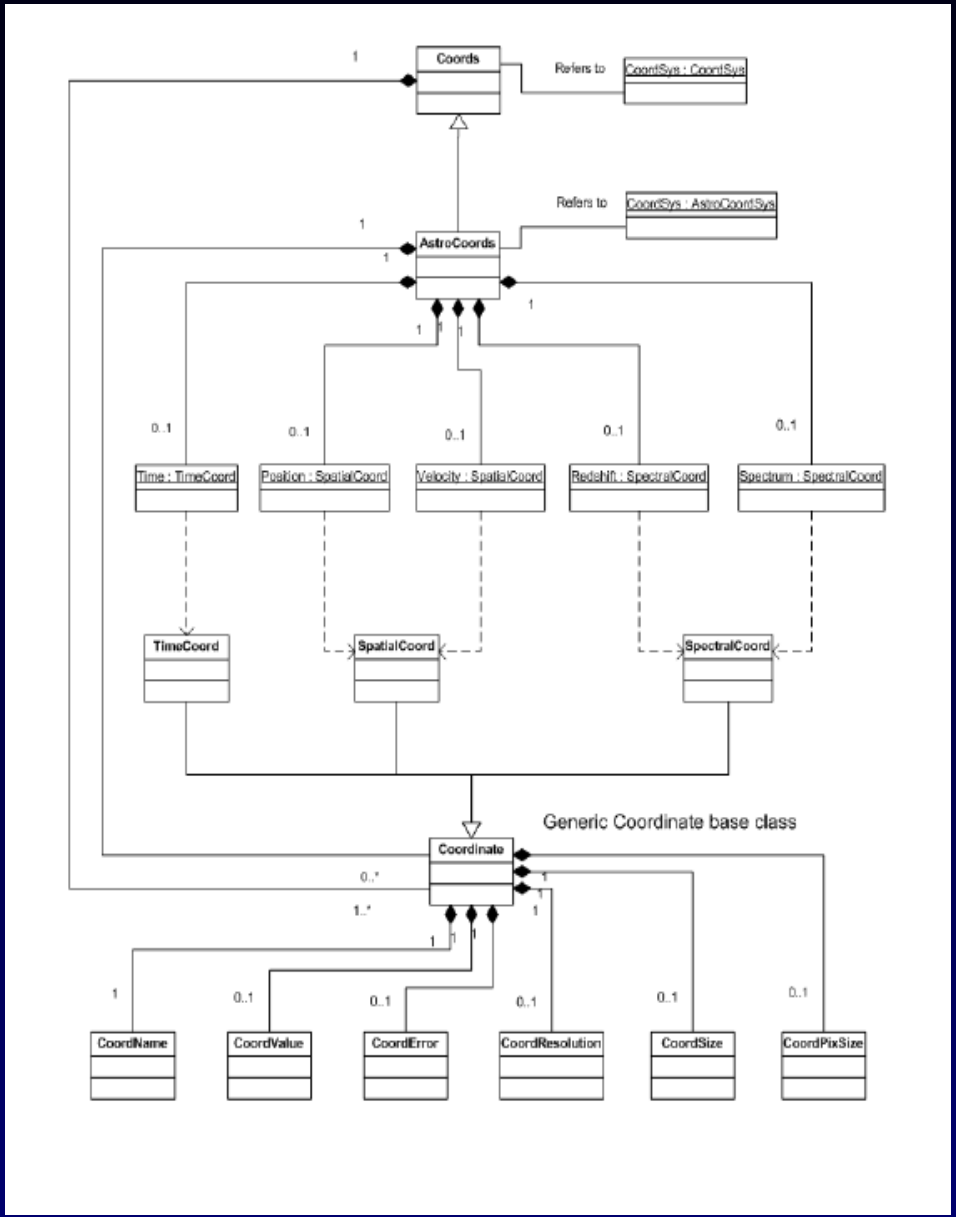
Curation – long time preservation issues (digital libraries)

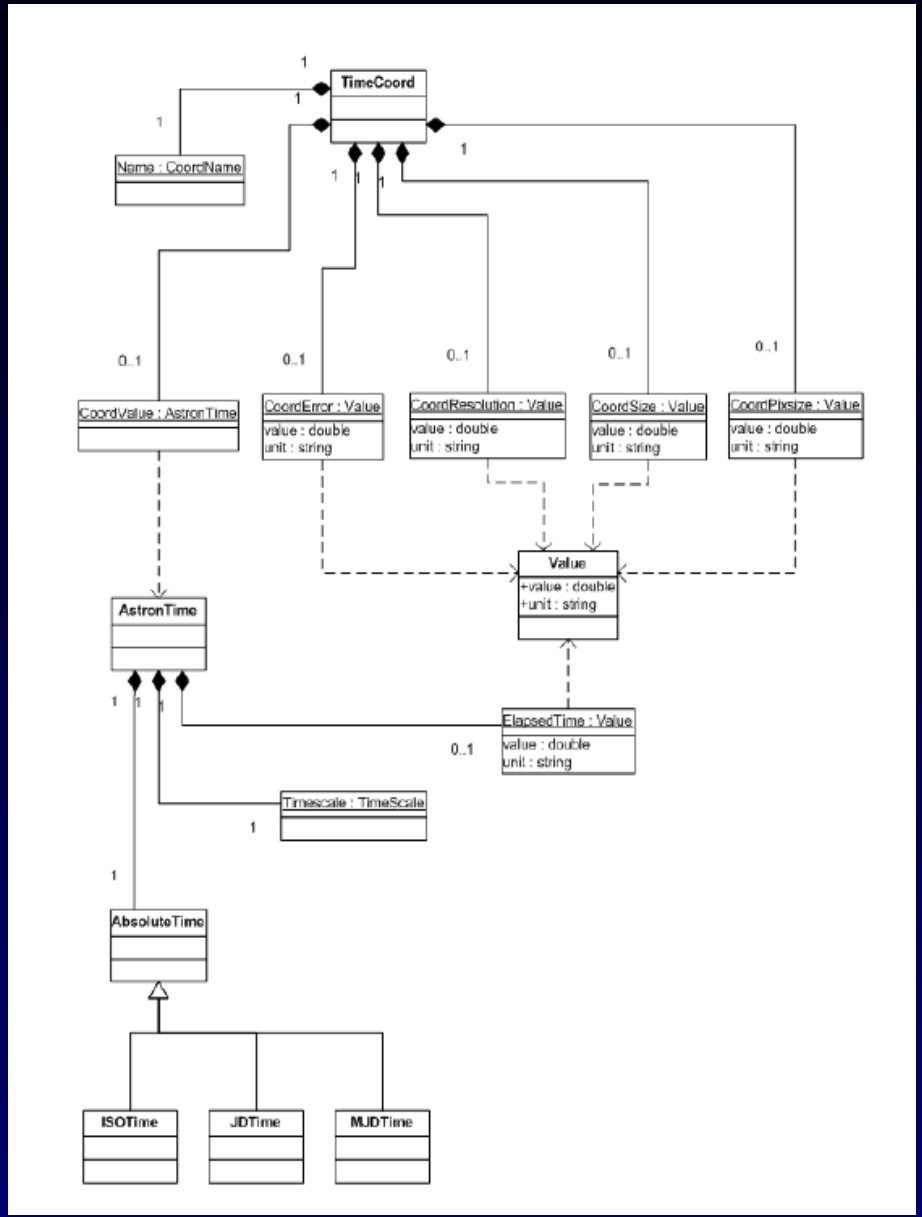
Provenance (how was processed, links to other products)

Characterization level 1 (spatial, spectral, temporal, polarization, location, coverage, porosity – SUB-CUBE)

Characterization level 2 (distorsion in images, spectra with nonlinear resolution …..)

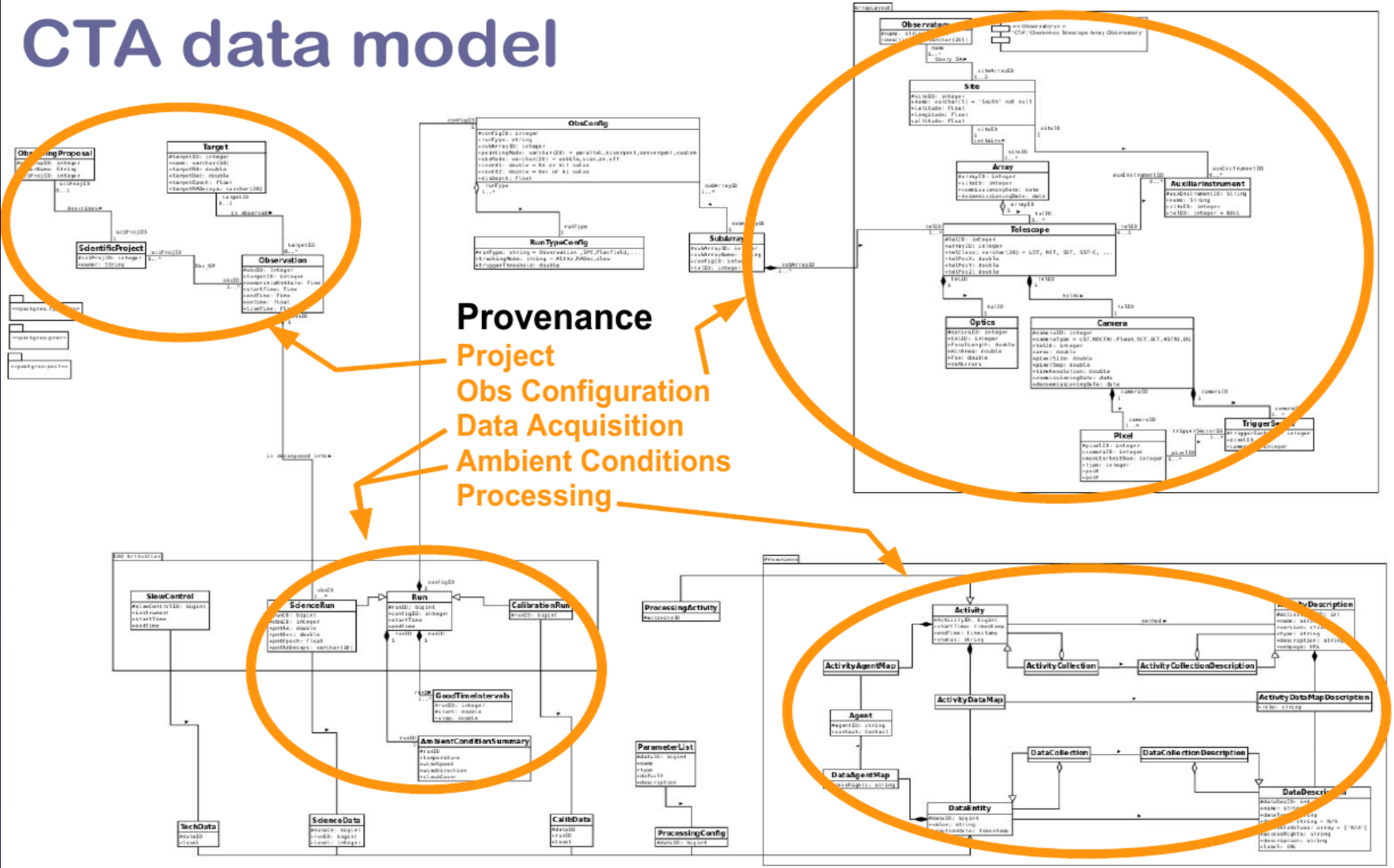
Space-Time-Coordinate Data Model



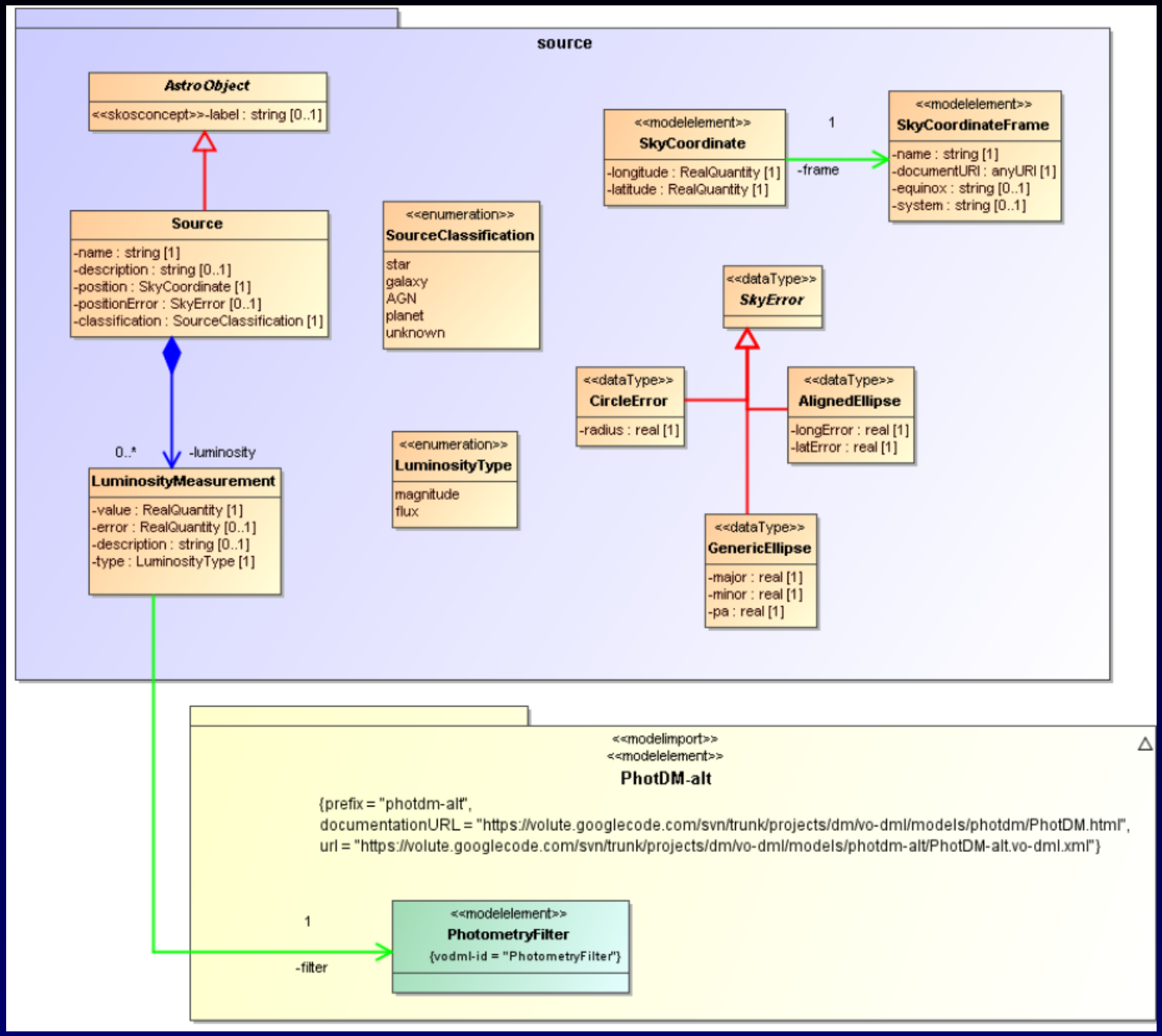


Cherenkov Telescope Array Data Model

CTA data model



VO-DML


source
AstroObject
<<skosconcept>>-label : string [0..1]
Source
-name : string [1]
-description : string [0..1]
-position : SkyCoordinate [1]
-positionError : SkyError [0..1]
-classification : SourceClassification [1]
0..*
-luminosity
LuminosityMeasurement
-value : RealQuantity [1]
-error : RealQuantity [0..1]
-description : string [0..1]
-type : LuminosityType [1]
<<enumeration>>
SourceClassification
star
galaxy
AGN
planet
unknown
<<enumeration>>
LuminosityType
magnitude
flux
<<modelelement>>
SkyCoordinate
-longitude : RealQuantity [1]
-latitude : RealQuantity [1]
1
-frame
<<modelelement>>
SkyCoordinateFrame
-name : string [1]
-documentURI : anyURI [1]
-equinox : string [0..1]
-system : string [0..1]
<<dataType>>
SkyError
<<dataType>>
CircleError
-radius : real [1]
<<dataType>>
AlignedEllipse
-longError : real [1]
-latError : real [1]
<<dataType>>
GenericEllipse
-major : real [1]
-minor : real [1]
-pa : real [1]
<<modelimport>>
<<modelelement>>
PhotDM-alt
{prefix = "photdm-alt",
documentationURL = "https://volute.googlecode.com/svn/trunk/projects/dm/vo-dml/models/photdm/PhotDM.html",
url = "https://volute.googlecode.com/svn/trunk/projects/dm/vo-dml/models/photdm-alt/PhotDM-alt.vo-dml.xml"}
1
-filter
<<modelelement>>
PhotometryFilter
{vodml-id = "PhotometryFilter"}

VO Registry – XML

```
  <validationLevel validatedBy="ivo://archive.stsci.edu/nvoregistry">2</validationLevel>
  <title>Hubble Space Telescope Spectra</title>
  <shortName>HST Spectra</shortName>
  <identifier>ivo://mast.stsci/ssap/hst</identifier>
▼<curation>
    <publisher>MAST</publisher>
  ▼<creator>
      <name>MAST</name>
    </creator>
    <version>1.0</version>
  ▼<contact>
      <name>Archive Branch, STScI</name>
      <email>archive@stsci.edu</email>
    </contact>
  </curation>
▼<content>
    <subject>UV</subject>
    <subject>Optical</subject>
    <subject>and Infrared Astronomy</subject>
  ▼<description>
      Spectra from the following HST instruments are available: GHRS (processed by CADC), FOS (processed by ECF), and STIS (1st
      order). Service is still under development. Links point to new (but incomplete) VO-compatible FITS files created by MAST staff.
    </description>
    <referenceURL>http://archive.stsci.edu/</referenceURL>
    <type>Archive</type>
    <contentLevel>Research</contentLevel>
  </content>
▼<capability standardID="ivo://ivoa.net/std/SSA" xsi:type="ssa:SimpleSpectralAccess">
  ▼<interface role="std" version="0.5" xsi:type="vs:ParamHTTP">
      <accessURL use="base">http://archive.stsci.edu/ssap/search.php?id=HST&</accessURL>
      <queryType>GET</queryType>
    </interface>
    <complianceLevel>query</complianceLevel>
    <dataSource>pointed</dataSource>
    <creationType>archival</creationType>
    <maxSearchRadius>360.0</maxSearchRadius>
    <maxRecords>10000</maxRecords>
    <defaultMaxRecords>10000</defaultMaxRecords>
    <maxAperture>180.0</maxAperture>
    <maxFileSize>1000000000</maxFileSize>
  </capability>
▼<coverage>
  ▼<STCResourceProfile xmlns="http://www.ivoa.net/xml/STC/stc-v1.30.xsd">
      <AstroCoordSystem id="mast.stsci_ssap_hstUTC-FK5-TOPO" xlink:href="ivo://STClib/CoordSys#UTC-FK5-TOPO" xlink:type="simple"/>
    ▼<AstroCoords coord_system_id="mast.stsci_ssap_hstUTC-FK5-TOPO">
      ▼<Position1D>
          <Size pos_unit="arcsec">0.0500000007450581</Size>
        </Position1D>
      </AstroCoords>
    </STCResourceProfile>
    <waveband>UV</waveband>
    <waveband>Optical</waveband>
  </coverage>
</ri:Resource>
```

Simple Spectra Access Protocol
Spectral Data Model

International
Virtual
Observatory
Alliance

Simple Spectral Access Protocol

Version 1.04

IVOA Recommendation Feb 01, 2008

This version:
http://www.ivoa.net/Documents/REC/DAL/SSA-20080201.html
Latest version:
http://www.ivoa.net/Documents/latest/SSA.html
Previous version(s):
Version 1.03, December 2007
Version 1.02, September 2007
Version 1.01, June 2007
Version 1.00, May 2007
Version 0.97, November 2006
Version 0.96, September 2006
Version 0.95 May 2006
Version 0.91 October 2005
Version 0.90 May 2005
Editors:
D.Tody, M. Dolensky
Authors:
D.Tody, M. Dolensky, J. McDowell, F. Bonnarel, T.Budavari, I.Busko, A. Micol, P.Osuna, J.Salgado, P.Skoda, R.Thompson, F.Valdes, and the data access layer working group.

International
Virtual
Observatory
Alliance

IVOA Spectral Data Model

Version 1.03

IVOA Recommendation 2007-10-29

This version (Recommendation Rev 1)
http://www.ivoa.net/Documents/REC/DM/SpectrumDM-20071029.pdf
Latest version:
http://www.ivoa.net/Documents/latest/SpectrumDM.html
Previous versions:
http://www.ivoa.net/Documents/PR/DM/SpectrumDM-20070913.html

Editors:
Jonathan McDowell, Doug Tody
Contributors:
Jonathan McDowell, Doug Tody, Tamas Budavari, Markus Dolensky, Inga Kamp, Kelly McCusker, Pavlos Protopapas, Arnold Rots, Randy Thompson, Frank Valdes, Petr Skoda, and the IVOA Data Access Layer and Data Model Working Groups.

SSAP Parameters

4.1.1 Mandatory Query Parameters

The following parameters **must** be implemented by a compliant service:

Parameter	*Sample value*	*Physical unit*	*Datatype*
POS	52,-27.8	degrees; defaults to ICRS	string
SIZE	0.05	degrees	double
BAND	2.7E-7/0.13	meters	string
TIME	1998-05-21/1999	ISO 8601 UTC	string
FORMAT	votable	-	string

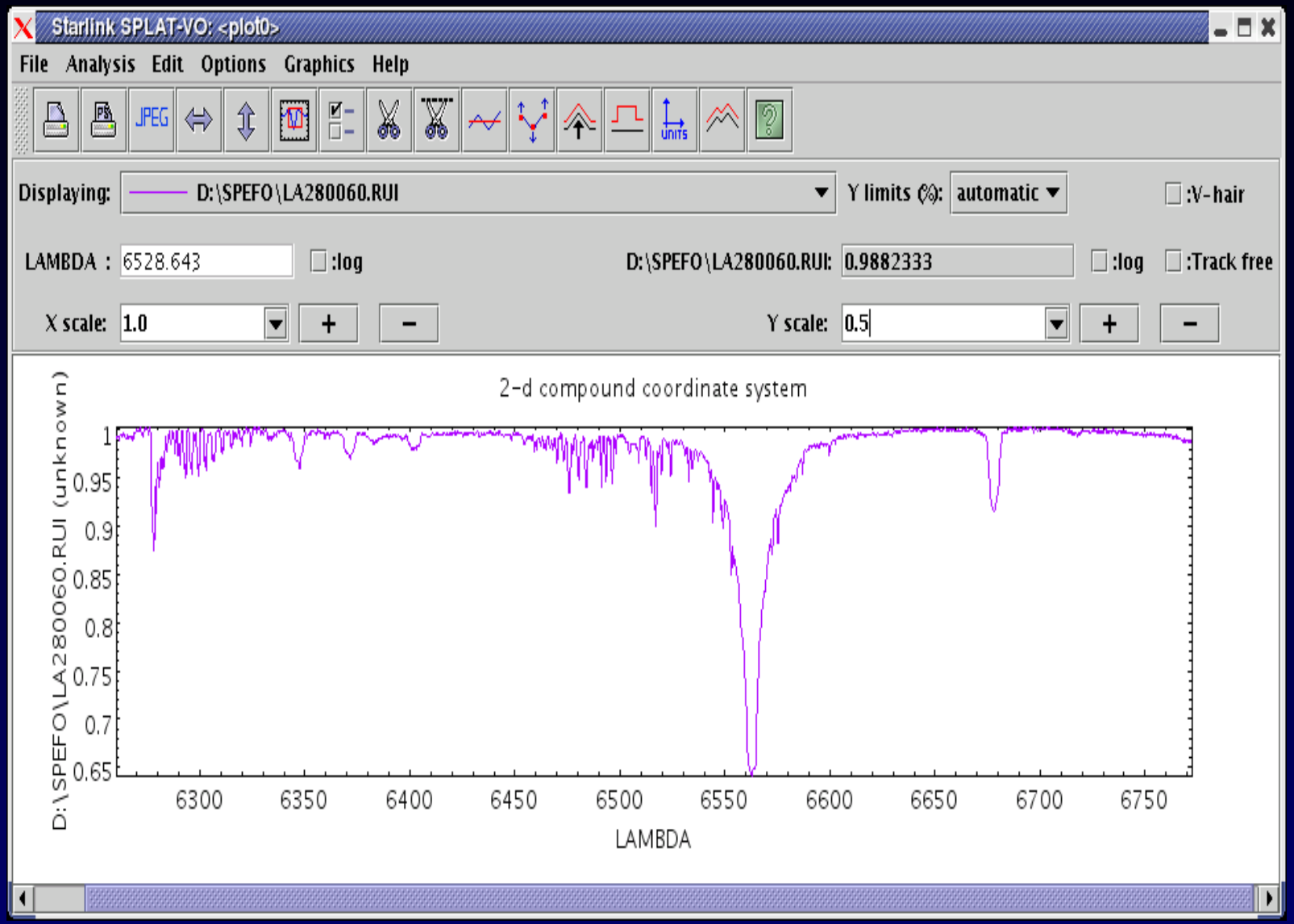
4.1.2 Recommended and Optional Query Parameters

Parameter	Sample value	Unit	Req	Datatype
APERTURE	0.00028 (=1")	degrees	OPT	double
SPECRP	2000	λ/dλ	REC	double
SPATRES	0.05	degrees	REC	double
TIMERES	31536000 (=1yr)	seconds	OPT	double
SNR	5.0	dimensionless	OPT	double
REDSHIFT	1.3/3.0	dimensionless	OPT	string
VARAMPL	0.77	dimensionless	OPT	string
TARGETNAME	mars		OPT	string
TARGETCLASS	star		OPT	string
FLUXCALIB	relative		OPT	string
WAVECALIB	absolute		OPT	string
PUBDID	ADS/col#R5983		REC	string
CREATORDID	ivo://auth/col#R1234		REC	string
COLLECTION	SDSS-DR5		REC	string
TOP	20	dimensionless	REC	int
MAXREC	5000		REC	string
MTIME	2005-01-01/2006-01-01	ISO 8601	REC	string
COMPRESS	true		REC	boolean
RUNID			REC	string

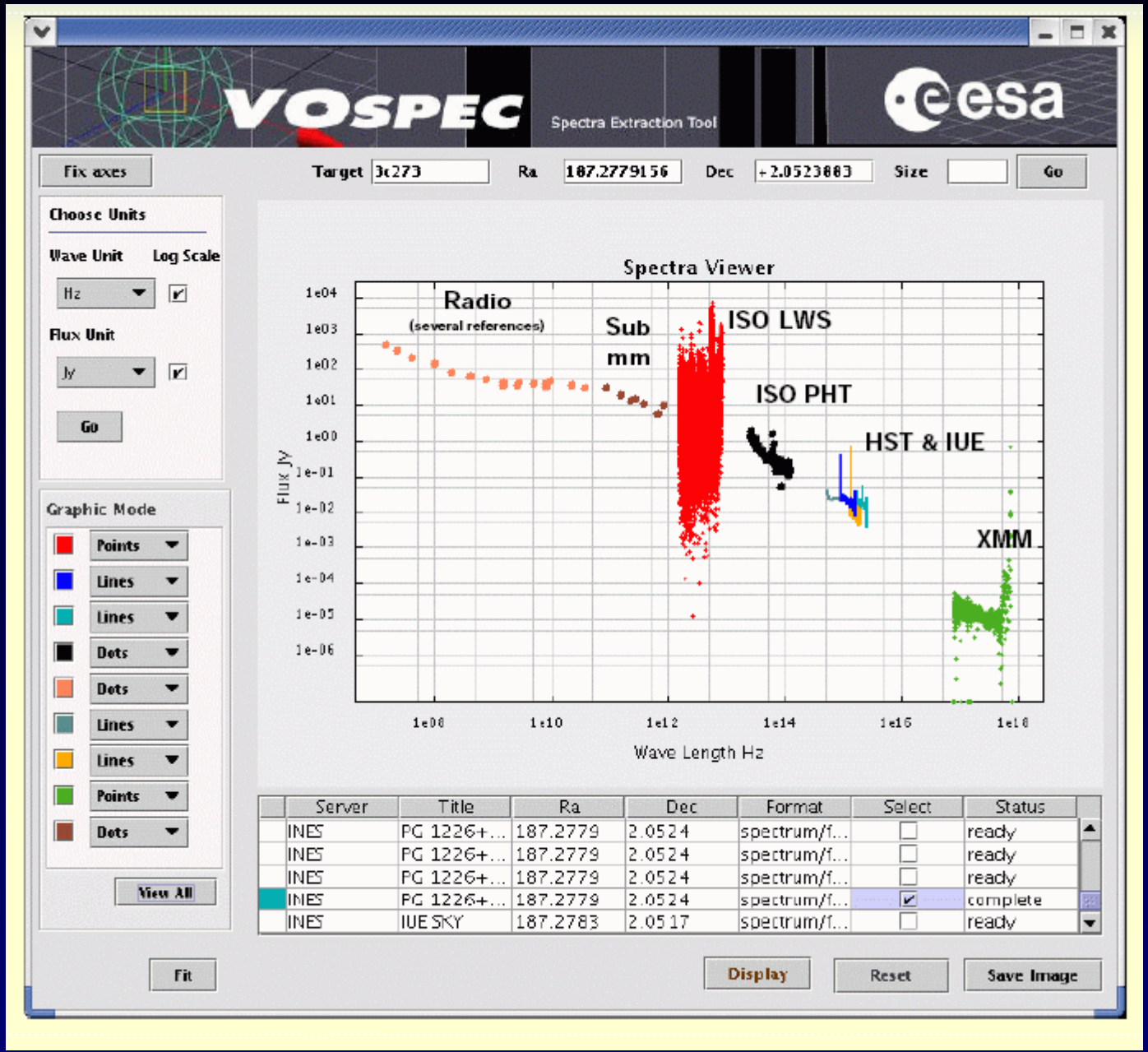
Big Data handling

VO Space	Moving big tables across (load only results)
SSO	Authentication, authorization, groups and consortia
UWS	Universal worker service (job synch, asynch)
PDL	Parameter Description Language
SIM-DB	Simulations, theory data

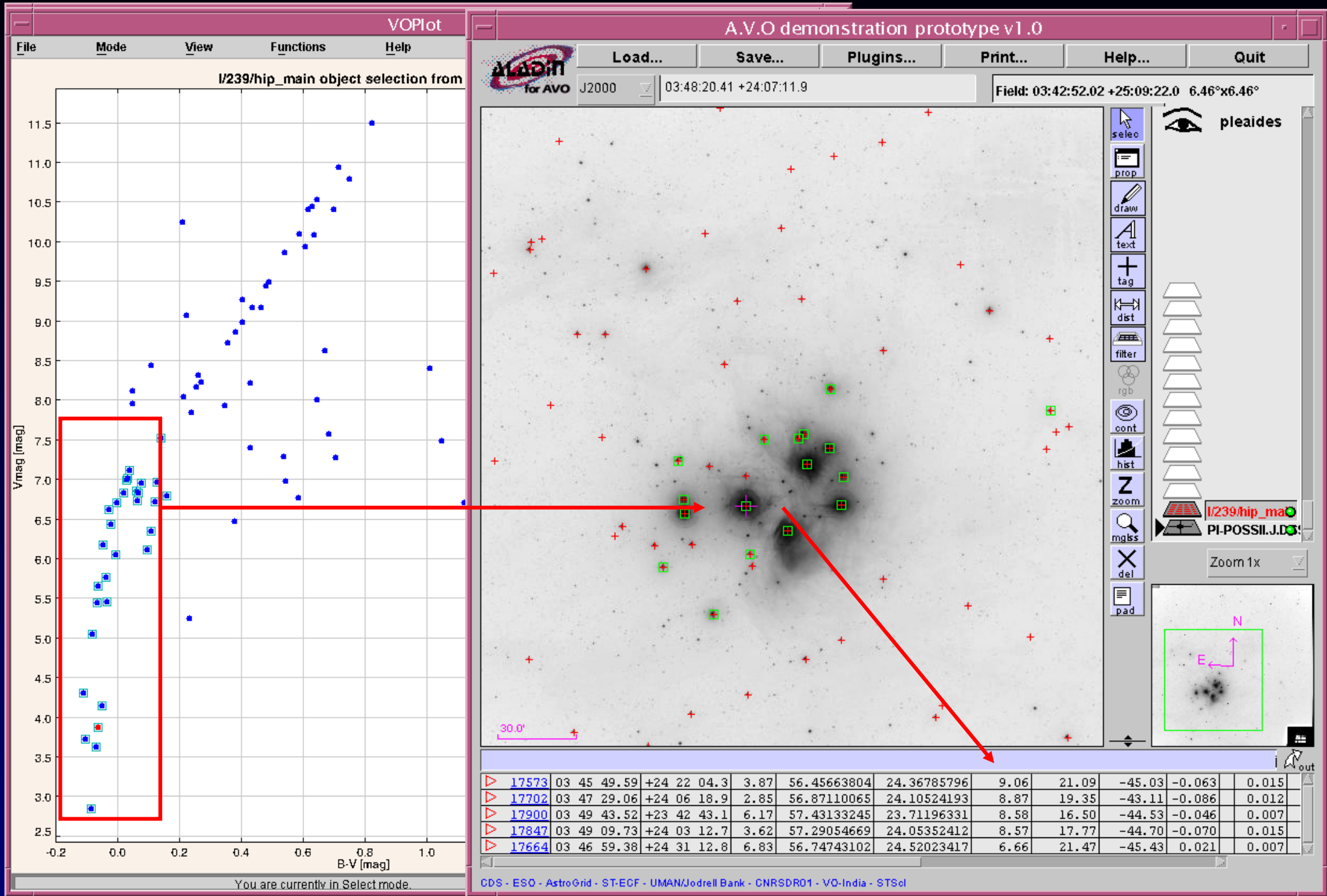
SPLAT-VO (Starlink, JAC)



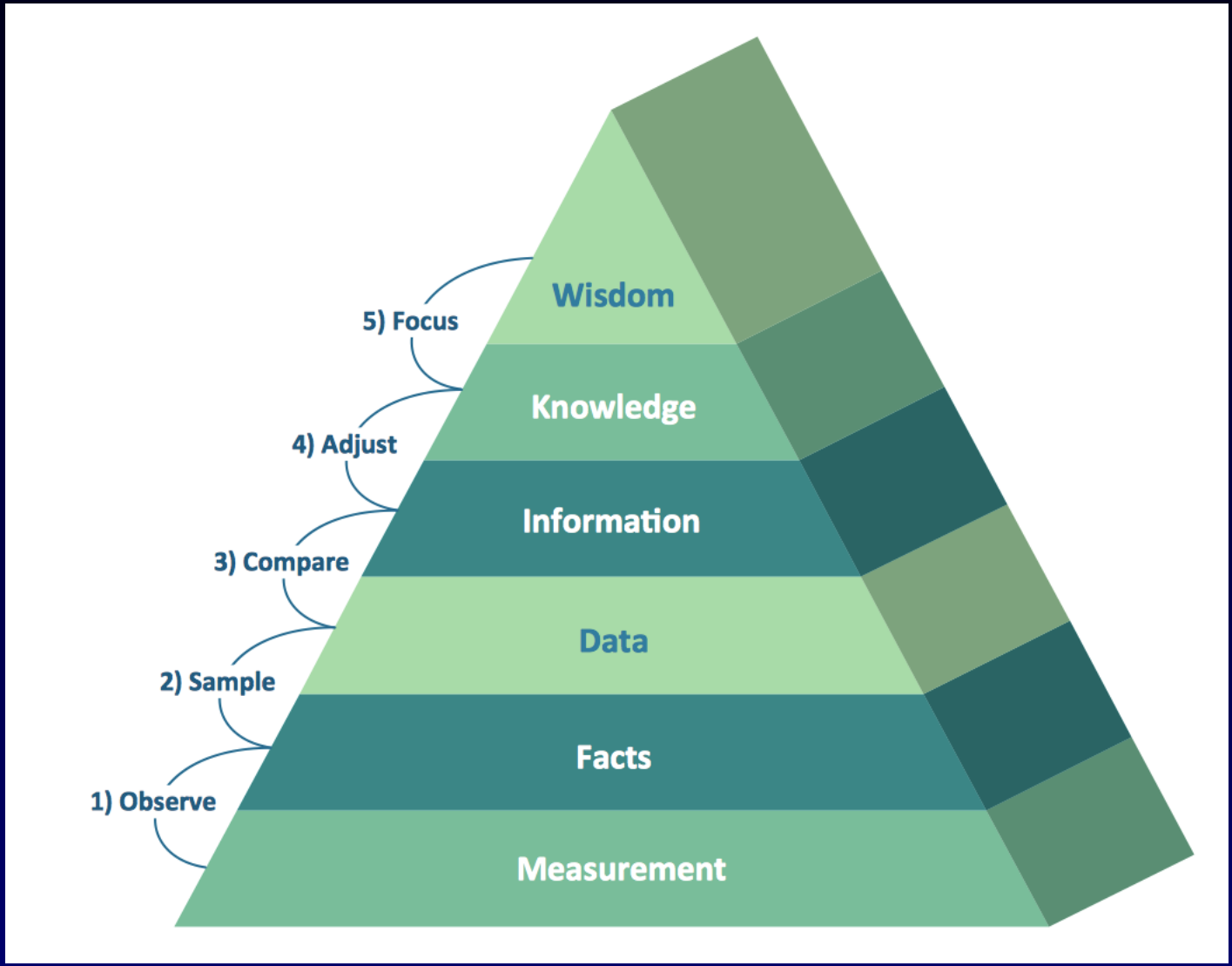
VOspec (ESAC)


VOSPEC
Spectra Extraction Tool
esa
Fix axes
Target 3c273
Ra 187.2779156
Dec +2.0523883
Size
Go
Choose Units
Wave Unit
Log Scale
Hz
Flux Unit
Jy
Go
Graphic Mode
Points
Lines
Lines
Dots
Dots
Lines
Lines
Points
Dots
View All
Spectra Viewer
Radio
(several references)
Sub
mm
ISO LWS
ISO PHT
HST & IUE
XMM
Flux Jy
Wave Length Hz
Server
Title
Ra
Dec
Format
Select
Status
INES PG 1226+... 187.2779 2.0524 spectrum/f... ready
INES PG 1226+... 187.2779 2.0524 spectrum/f... ready
INES PG 1226+... 187.2779 2.0524 spectrum/f... ready
INES PG 1226+... 187.2779 2.0524 spectrum/f... complete
INES IUE SKY 187.2783 2.0517 spectrum/f... ready
Fit
Display
Reset
Save Image

Colour-magnitude diagram



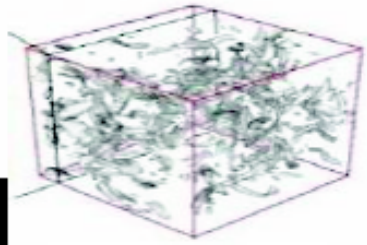
Data-Knowledge-Wisdom Pyramid

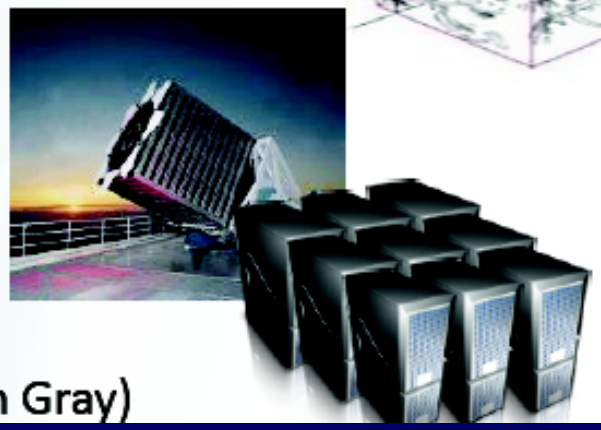


Emergence of a Fourth Research Paradigm

1. Thousand years ago – **Experimental Science**
   - Description of natural phenomena
2. Last few hundred years – **Theoretical Science**
   - Newton's Laws, Maxwell's Equations…
3. Last few decades – **Computational Science**
   - Simulation of complex phenomena
4. Today – **Data-Intensive Science**
   - Scientists overwhelmed with data sets from many different sources
     - Data captured by instruments
     - Data generated by simulations
     - Data generated by sensor networks
   - ➢ eScience is the set of tools and technologies to support data federation and collaboration
     - For analysis and data mining
     - For data visualization and exploration
     - For scholarly communication and dissemination

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

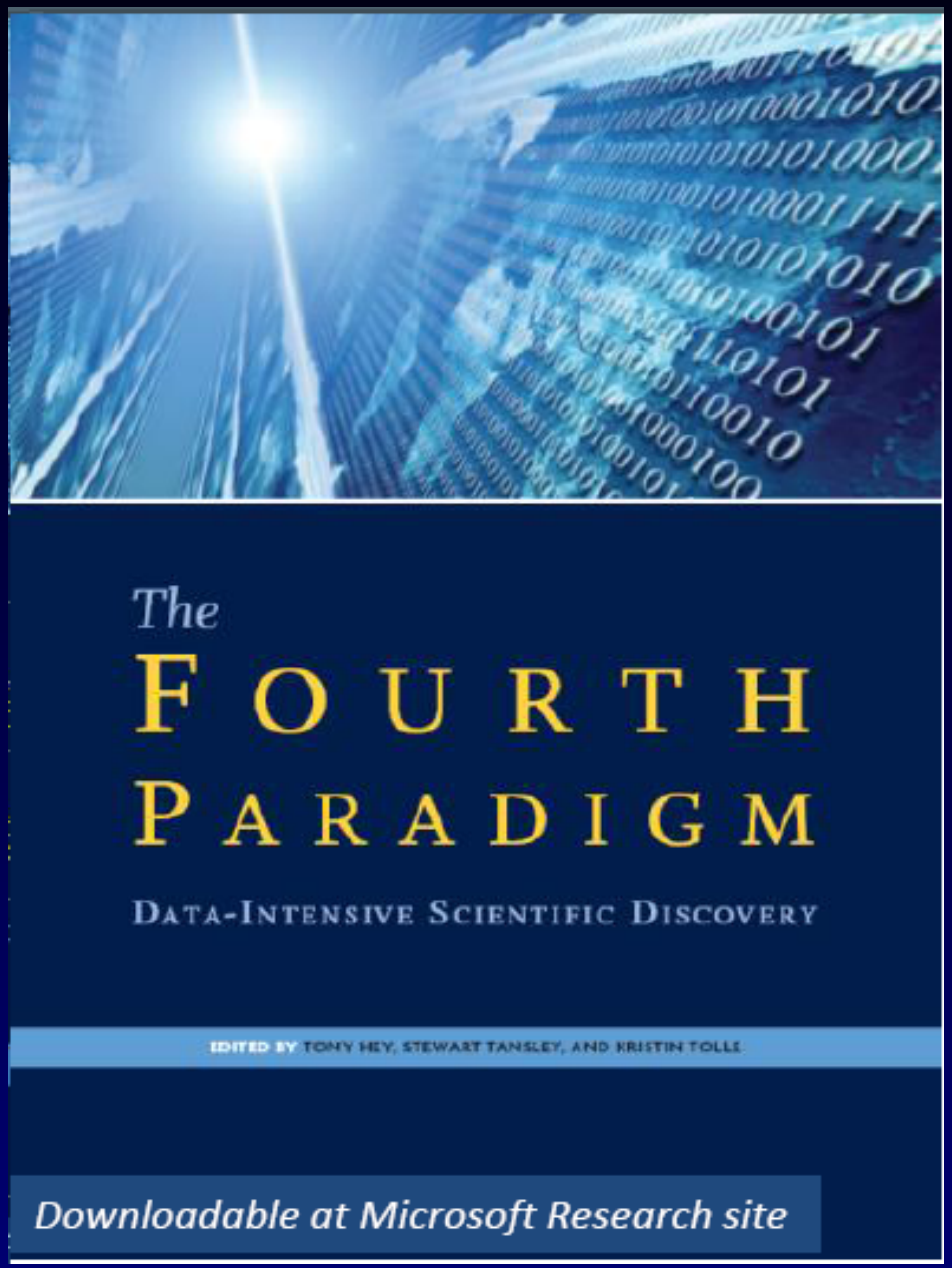




(With thanks to Jim Gray)

From T. Hey, AI2010

X-informatics

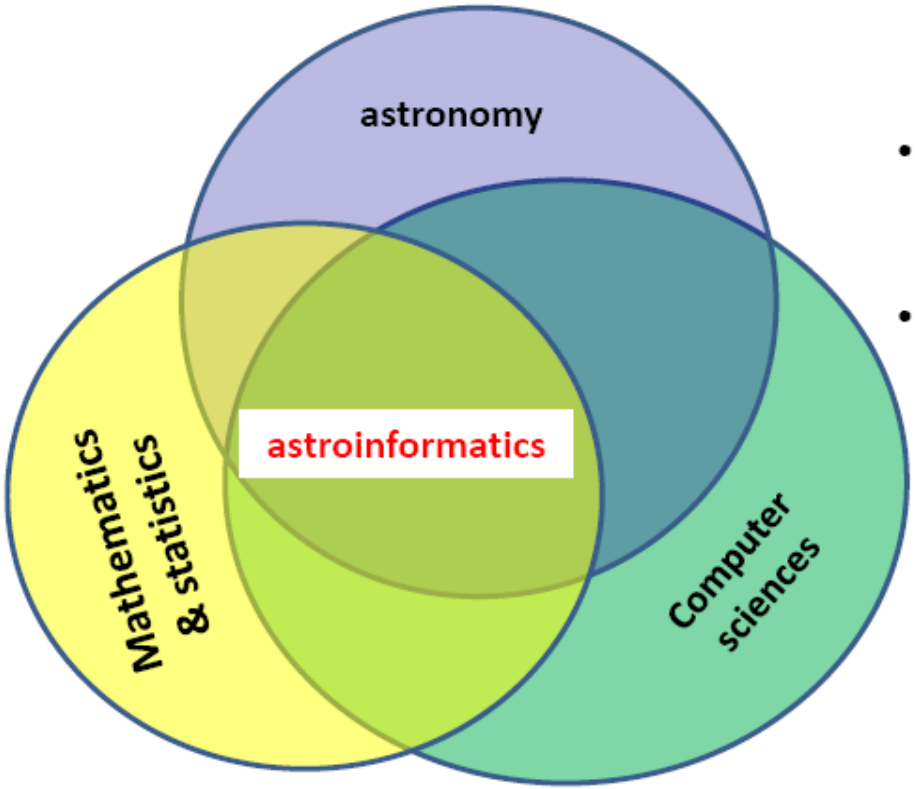


Changing methodology of the Science

Synergy between different worlds

Sociological aspects (net-based research communities)

Experimental astronomy has become a three players game



- **astronomy**: problems, data, understanding of the data structure and biases
- **mathematics**: evaluation of the data, falsification/validation of theories/models, etc
- **computer science**: implementation of infrastructures, databases, middleware, scalable tools, etc

- **Astroinformatics: AAS n. 215, Washington, December 2009,** chairperson: K. Borne
- **Astroinformatics 2010: Caltech (USA) June 16-19 2010;**co-chairpersons: S.G. Djorgovski, G. Longo
- **Astroinformatics 2011: UNINA – Sorrento,** co-chairpersons: S.G. Djorgovski, G. Longo

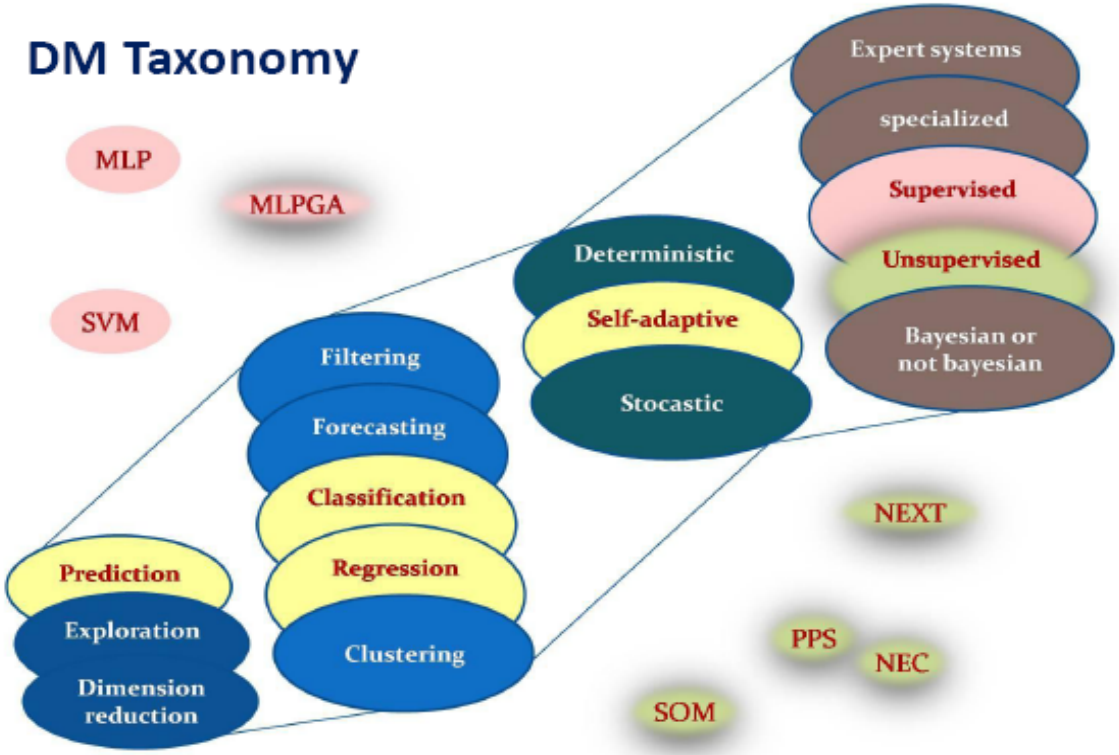
Longo 2010

Astroinformatics

- Analogy – Bioinformatics (Genome analysis with GRIDS, ATB)
- e-Science in Astronomy
- Data mining, Knowledge discovery - VO-NEURAL, DAME
- Examples
  - Photometric RedShift
  - Searching for QSO (light curves, MOS)
  - Automatic Light curves classification (GAIA, LSST)
- New ways of scholar communication (VR, 2nd Life, U-Science)
- BIG data problems, GPUs, NoSQL DB, visualization,
- Very NEW – emerging discipline

Data Mining is the activity of extracting **USEFUL** information from **COMPLEX** data using Statistical Pattern Recognition and Machine Learning methods.

DM Taxonomy



1. To catalogue the known (classification)
2. Characterize the unknown (clustering)
3. Find functional dependencies (regression)
4. Find exceptions (outliers)

Supervised Methods

Patterns are learnt from extensive set of templates (Base of Knowledge = BoK)

Unsupervised Methods

Patterns are discovered using the data themselves

Need for a new science: Astroinformatics

Knowledge Discovery in Databases

Data Gathering (e.g., from sensor networks, telescopes…)

→ Data Farming:

Storage/Archiving
Indexing, Searchability
Data Fusion, Interoperability, ontologies, etc.

Database technologies

→ Data Mining (or Knowledge Discovery in Databases):

Pattern or correlation search
Clustering analysis, automated classification
Outlier / anomaly searches
Hyperdimensional visualization

Key mathematical issues

→ Data understanding

Computer aided understanding
KDD
Etc.

Ongoing research

→ New Knowledge

DA ME

Data Driven Science

What is Fundamentally New Here?

- The ***information volumes and rates*** grow exponentially

➡ ***Most data will never be seen by humans***

- A great increase in the data ***information content***

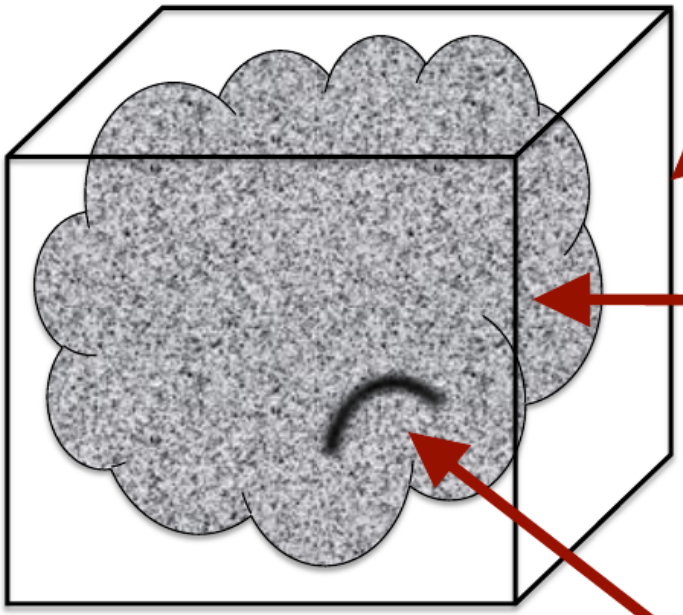
➡ ***Data driven vs. hypothesis driven science***

- A great increase in the ***information complexity***

➡ ***There are patterns in the data that cannot be comprehended by humans directly***

Hidden Patterns in Data

Pattern or structure (Correlations, Clustering, Outliers, etc.) Discovery in High-Dimensional Parameter Spaces



D >> 3 parameter space hypercube

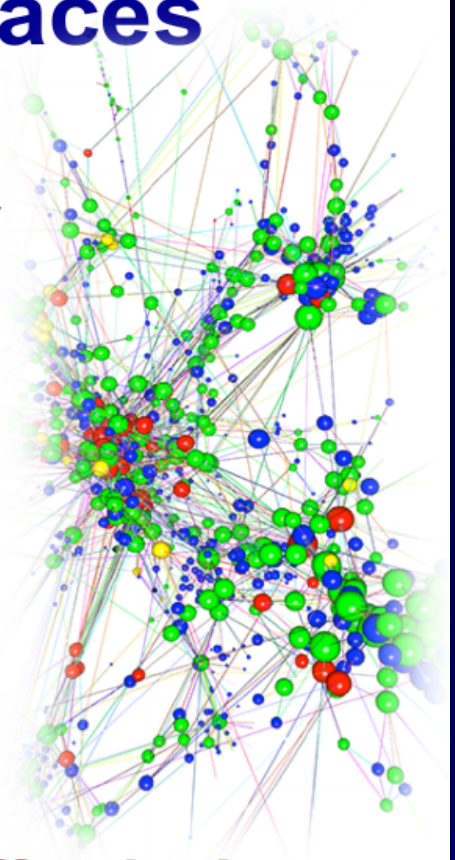
High-D data cloud: mostly noise, of an arbitrary distribution

But in some corner of some sub-D projection of this data space, there is ***something ≠ noise***

Visualization in Machine Learning

A Key Challenge: Visualisating Multidimensional Data Spaces

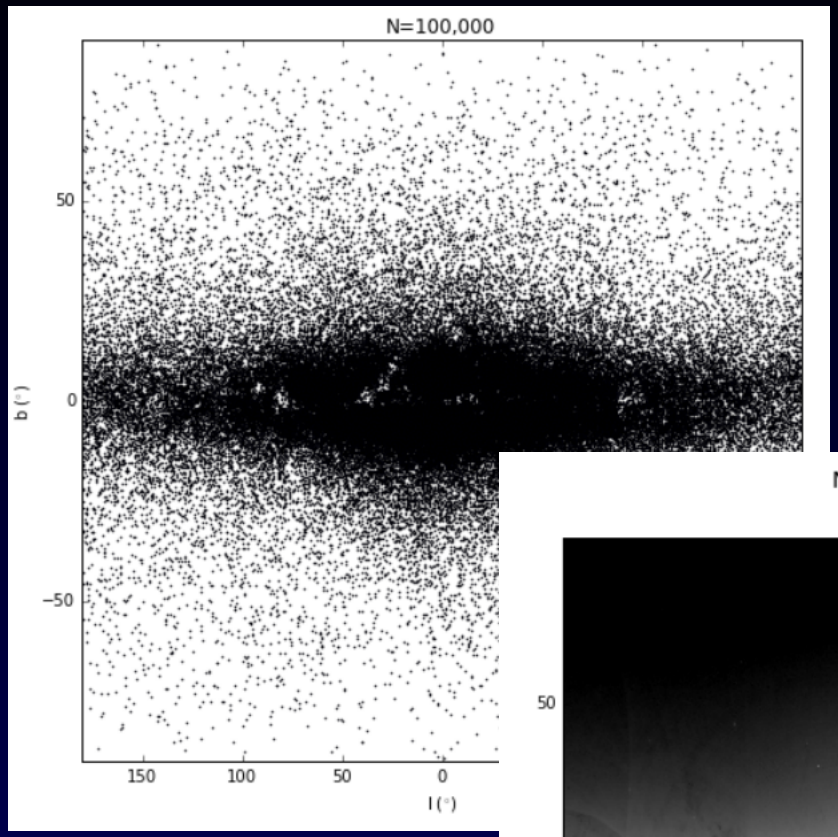
- Hyperdimensional structures (clusters, correlations, etc.) may be present in many complex data sets, whose dimensionality may be D ~ 10^2 – 10^4, or higher
- It is a matter of ***data understanding***, choosing the right data mining algorithms, and interpreting the results
- We are biologically limited to perceiving up to ~ 3 - 12(?) dimensions

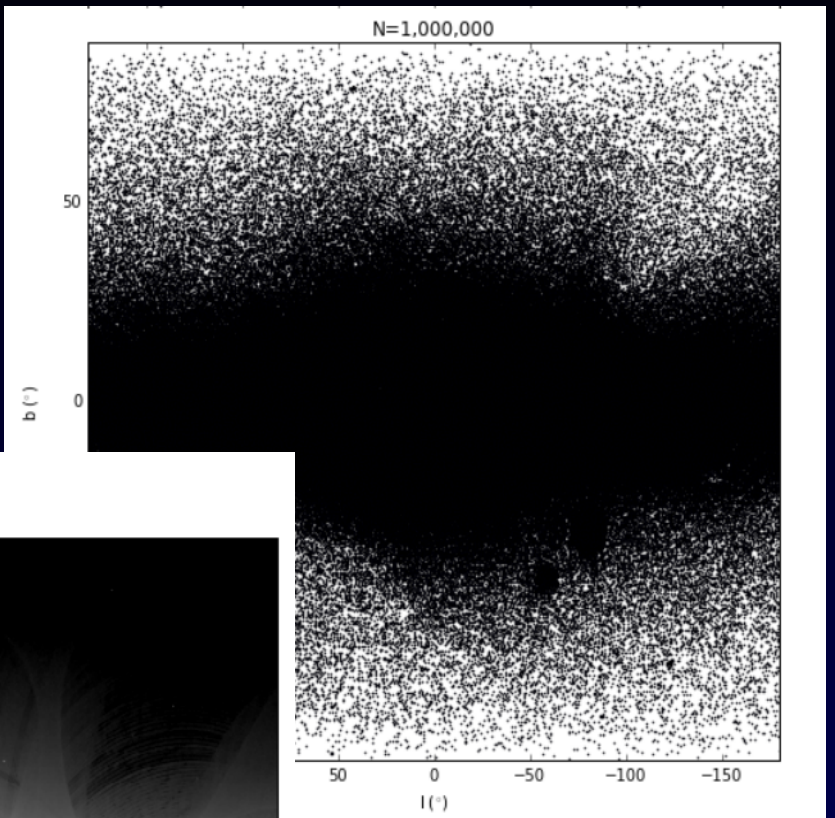


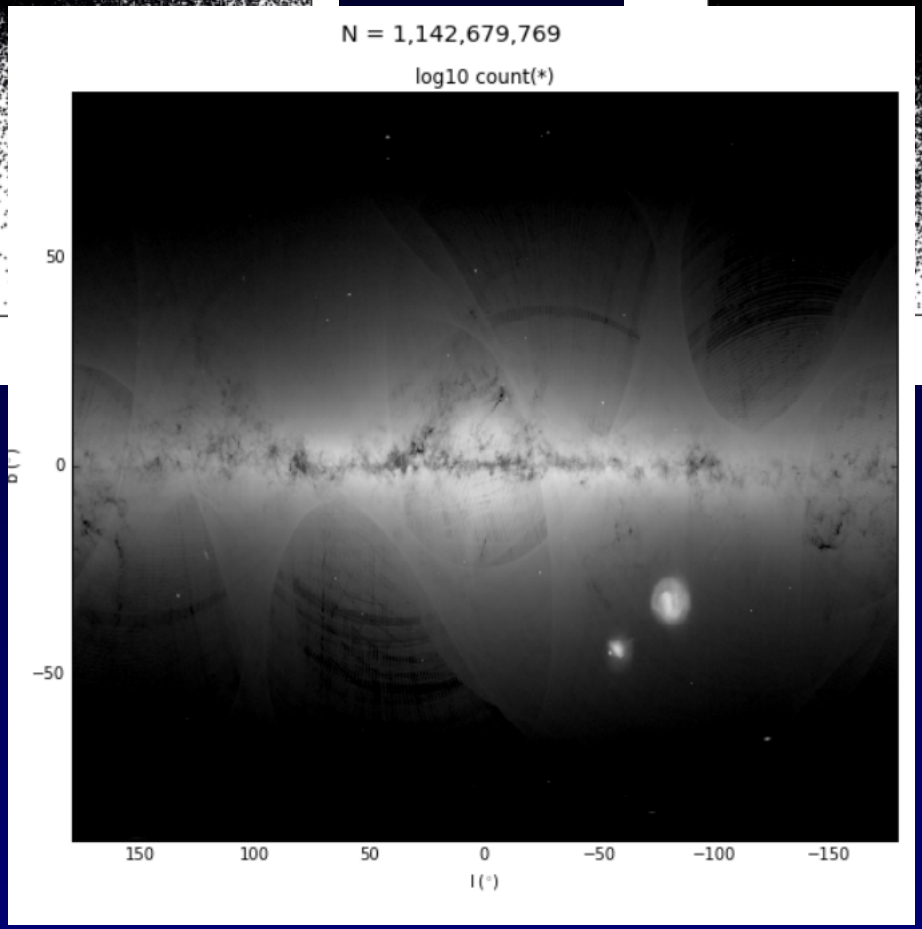
What good are the data if we cannot effectively extract knowledge from them?

Djorgovski

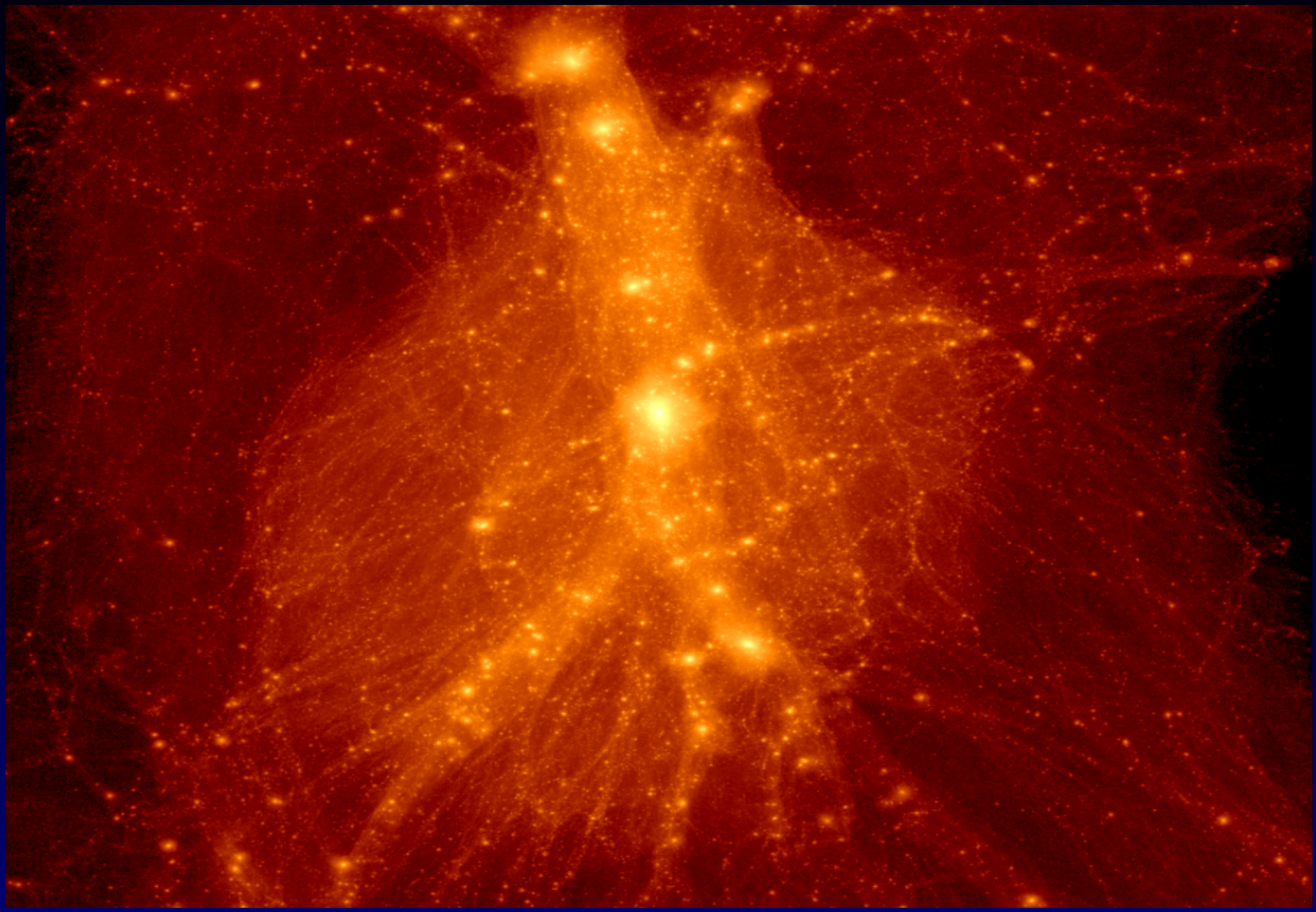
Visualization of 1 B points – Gaia DR1





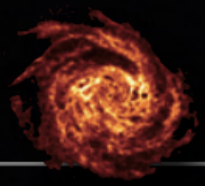


Visualization of Big Data

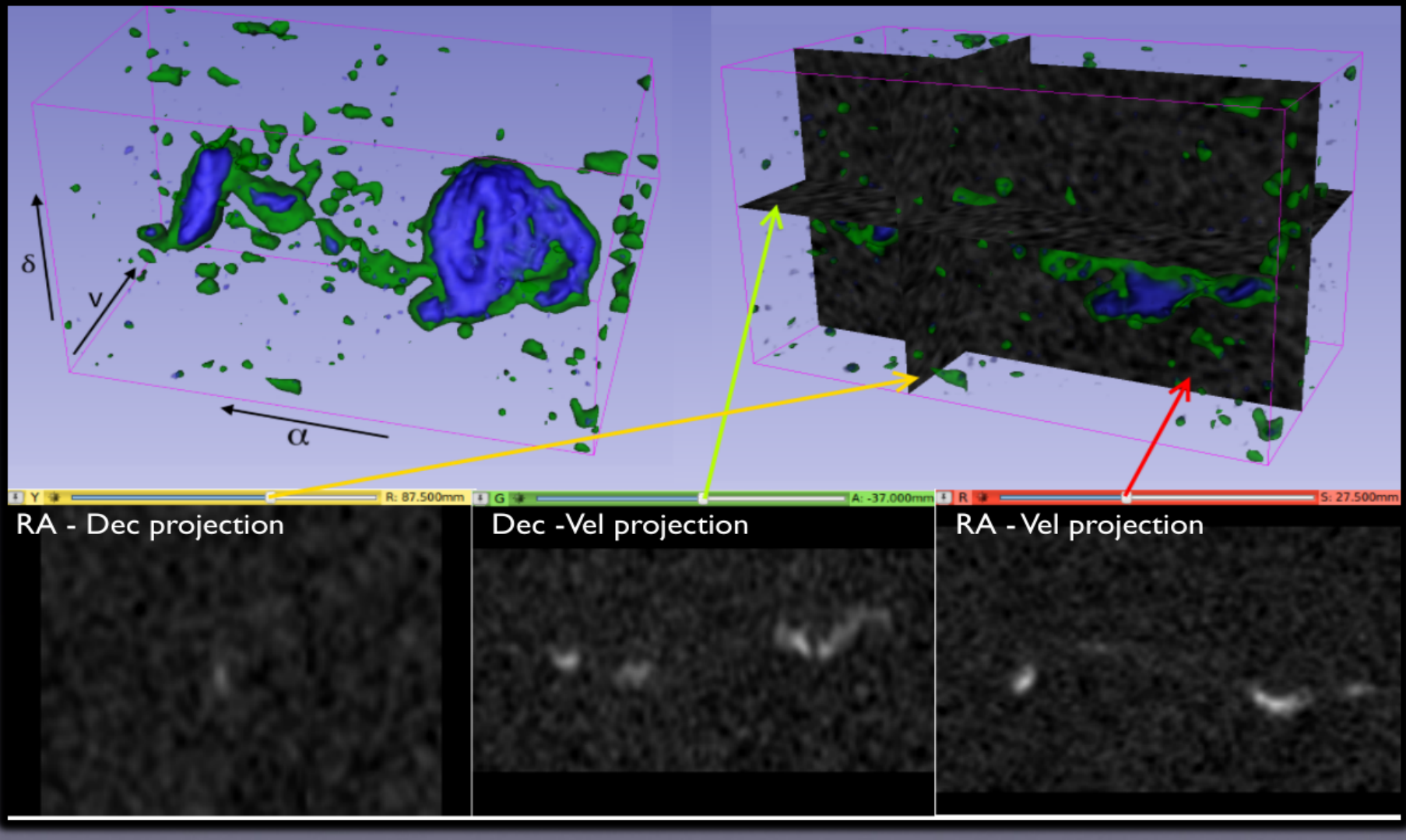


Visualization of Big Data

Visualization of Radio Data Cubes



3D Slicer provides full linked views, not just slices

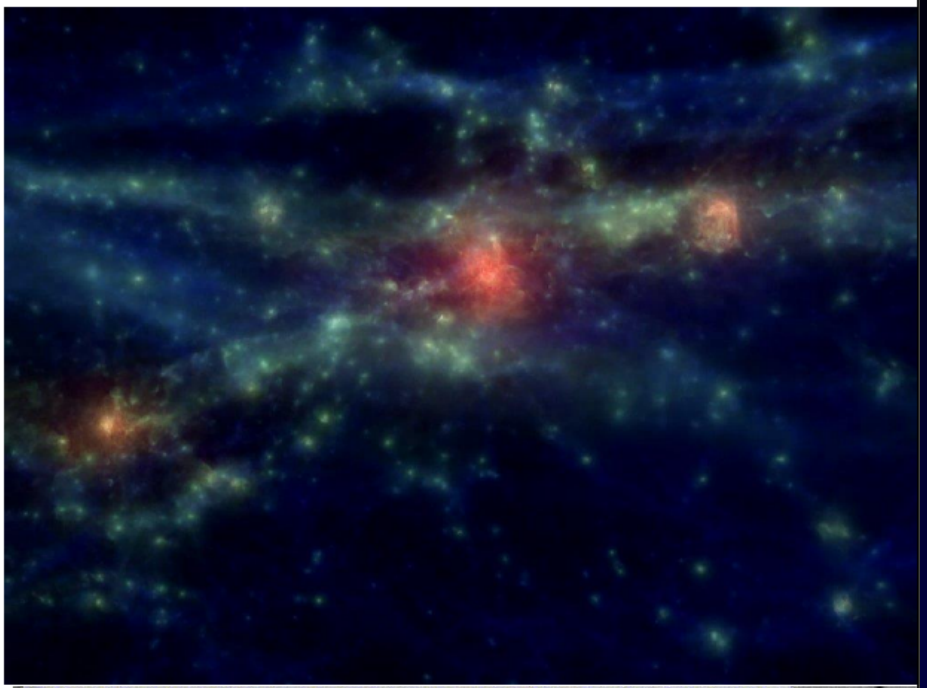


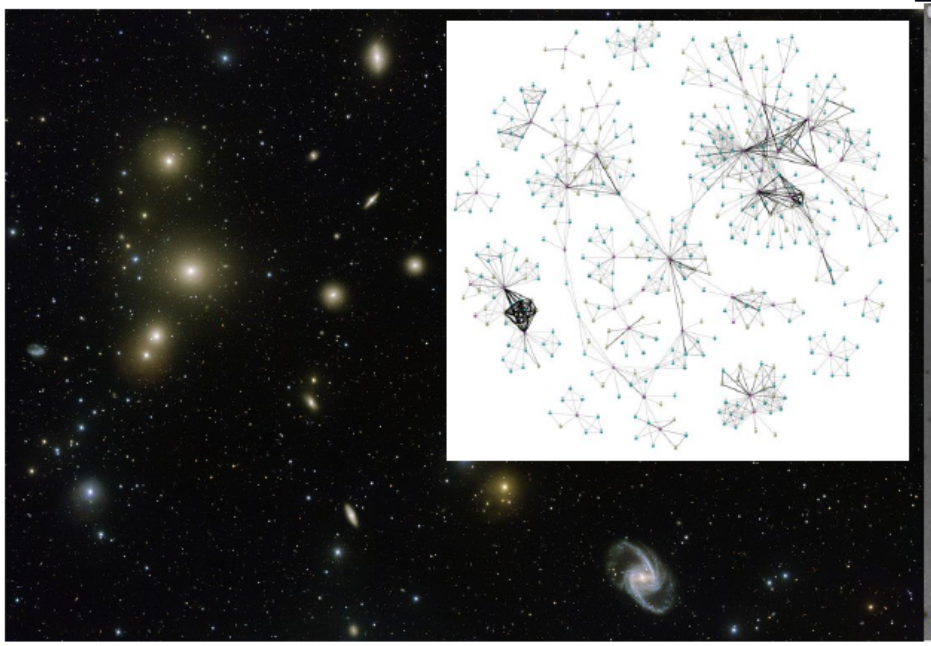
Advanced Visualization

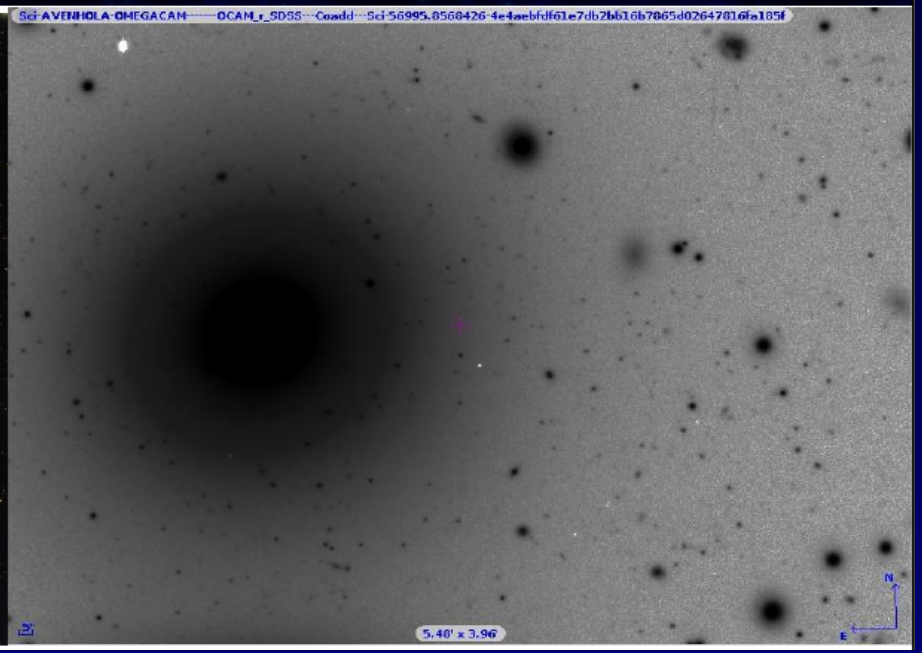
Visualization

- Develop visual data analytic tools to optimally visualize simulations and observations.

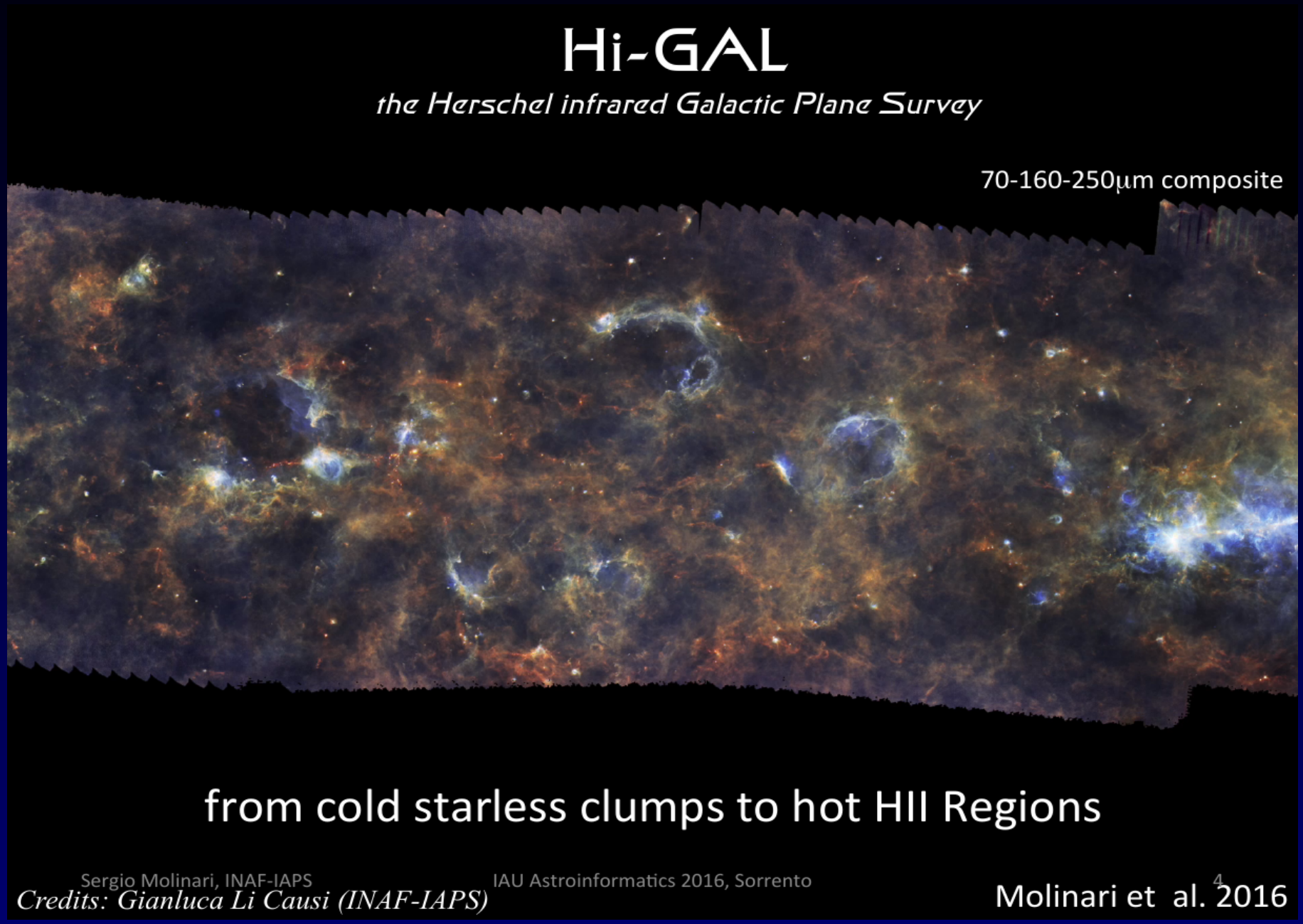
- Follow newest developments, in connection with companies.



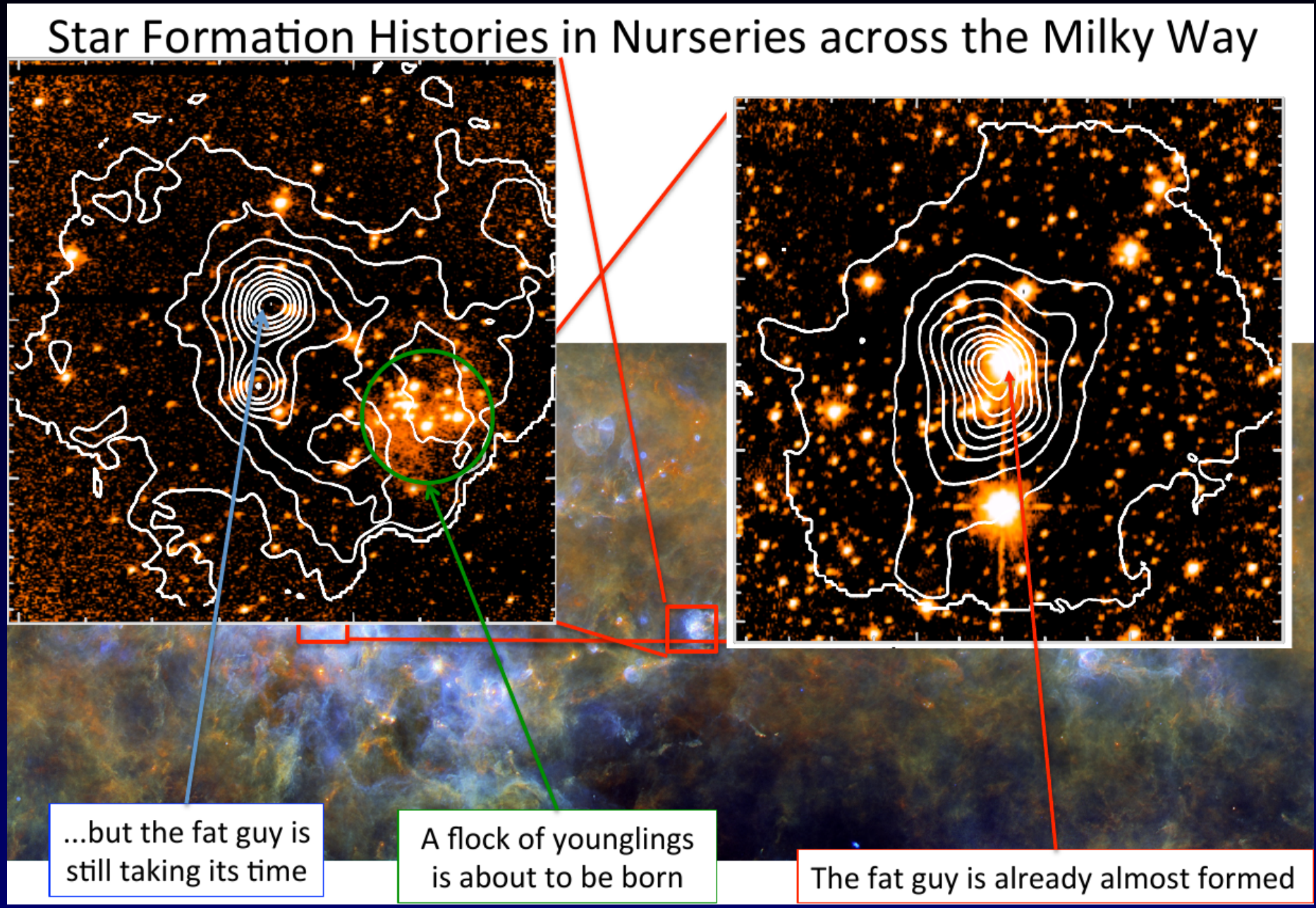




Star Forming Regions in Galaxy



Via Lactea – Star Forming



CAVE2 Monash University AU

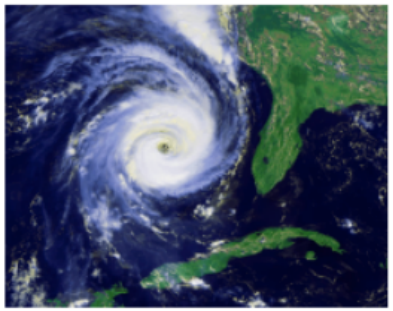
8m diameter, 330 deg FOV , 80x LCD 46“ 1366x768 Stereo + head tracking ……

From Astronomy to Earth Sciences

GRACE
AQUA

From Observations to Applications

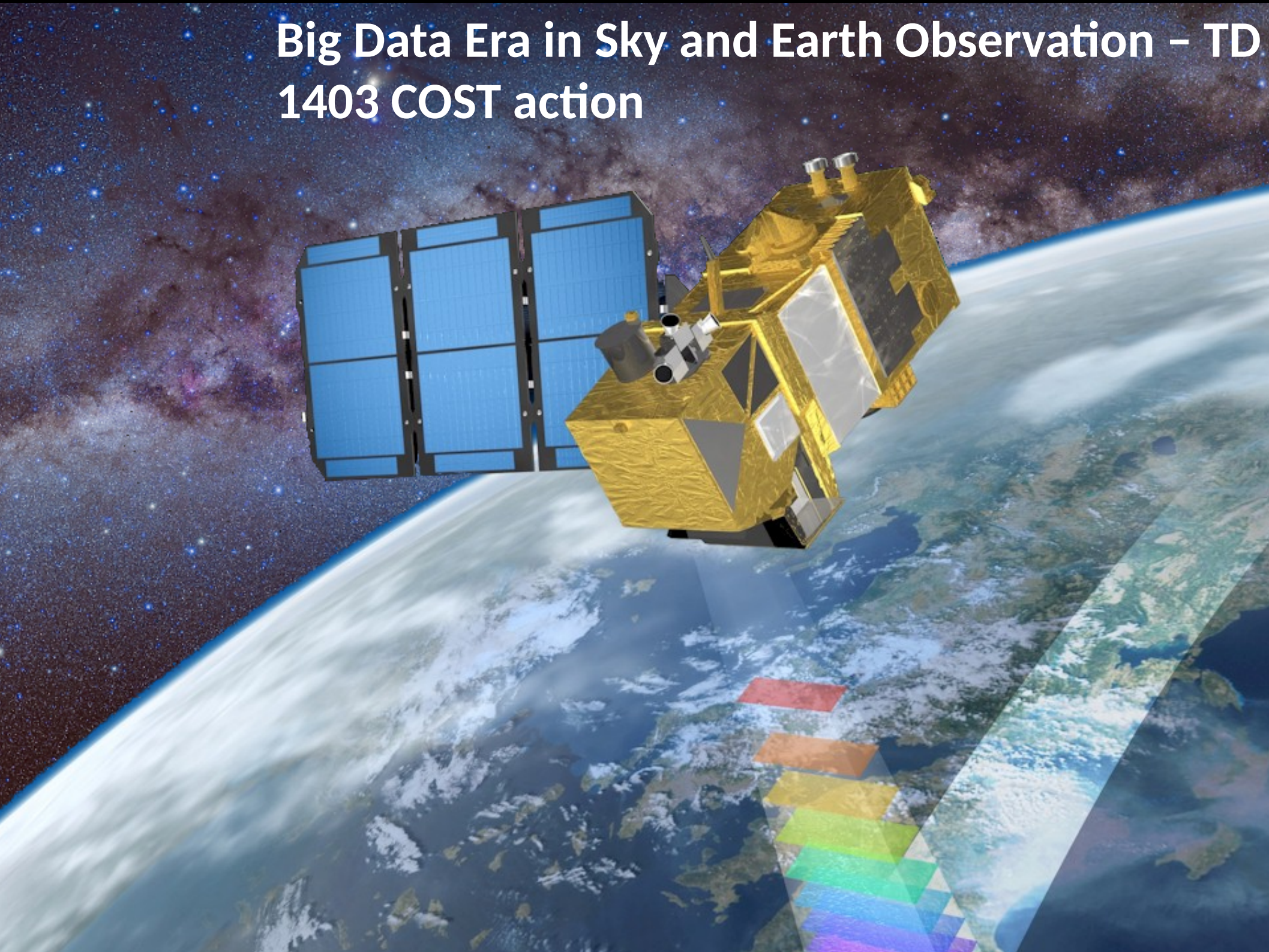
Satellite Measurements



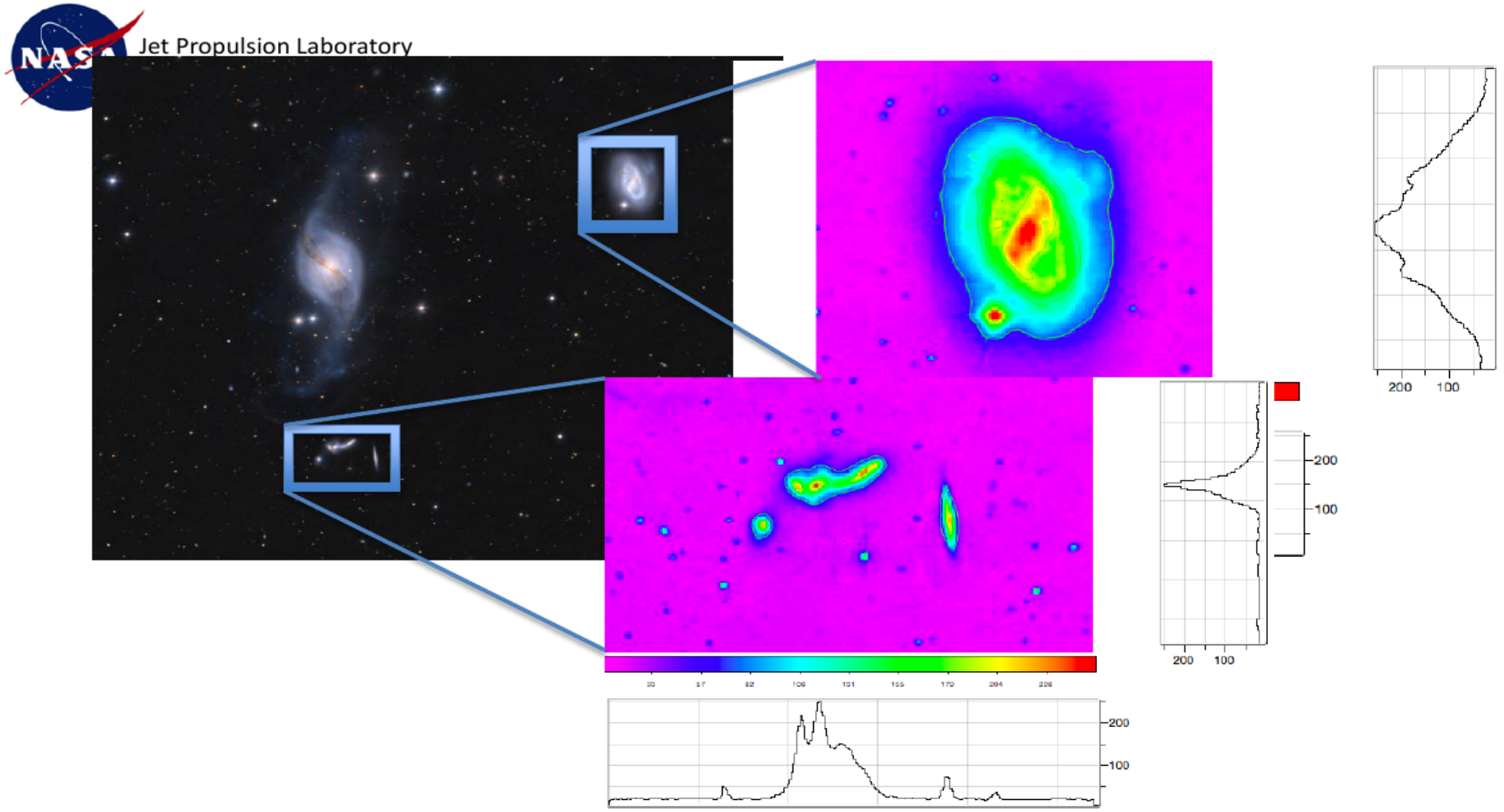
Satellite Products

Environmental Applications

Big Data Era in Sky and Earth Observation – TD 1403 COST action

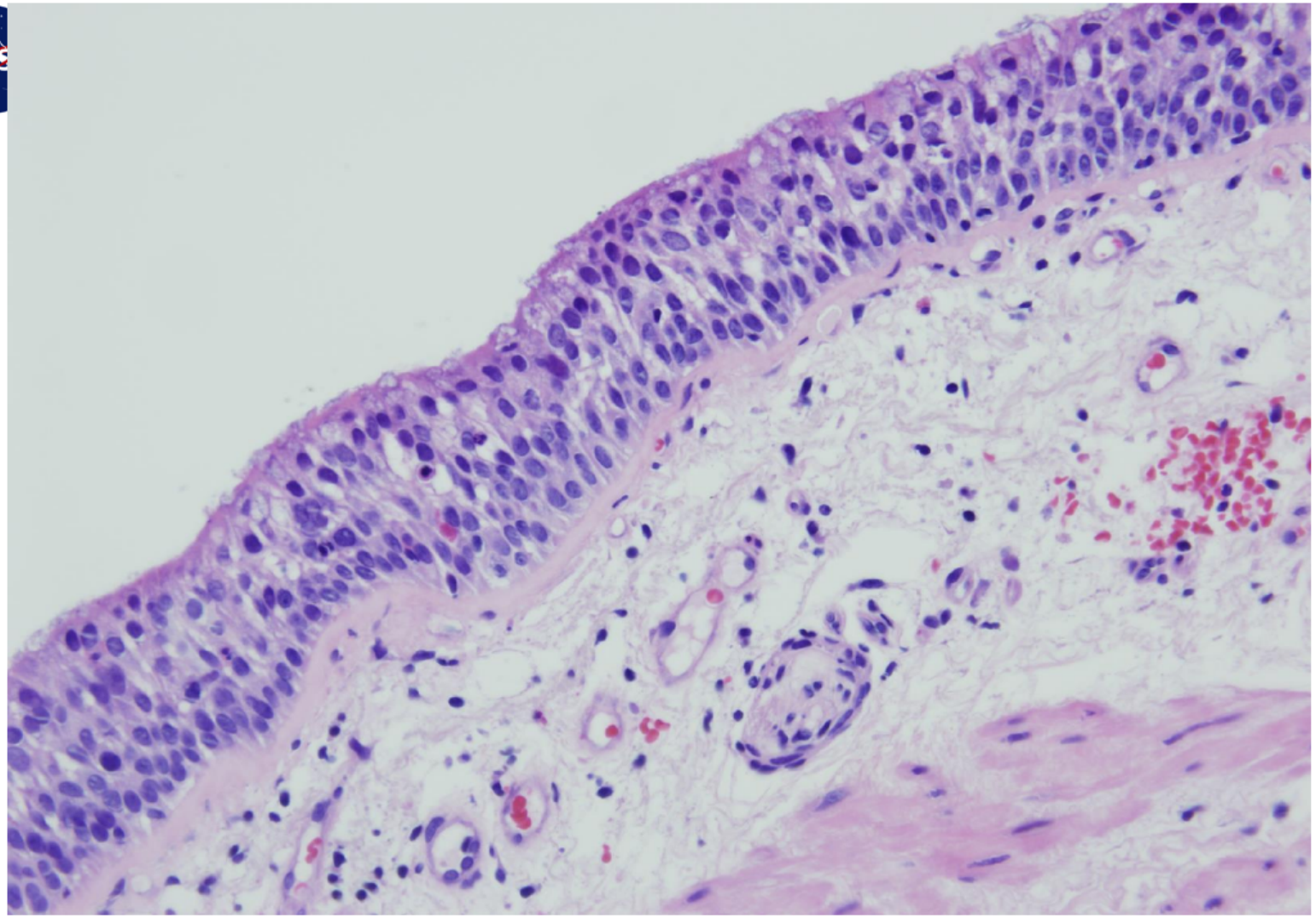


Finding Galaxies by Shape NASA



Description: Detecting objects from astronomical measurements by evaluating light measurements in pixels using intelligent software algorithms.
Image Credit: Catalina Sky Survey (CSS), of the Lunar and Planetary Laboratory, University of Arizona, and Catalina Realtime Transient Survey (CRTS), Center for Data-Driven Discovery, Caltech.

Finding Cancer Signatures NASA



Description: Detecting objects from oncology images using intelligent software algorithms transferred to and from space science.
Image Credit: EDRN Lung Specimen Pathology image example, University of Colorado

New e-Science Collaborations

Center for Data-Driven Discovery

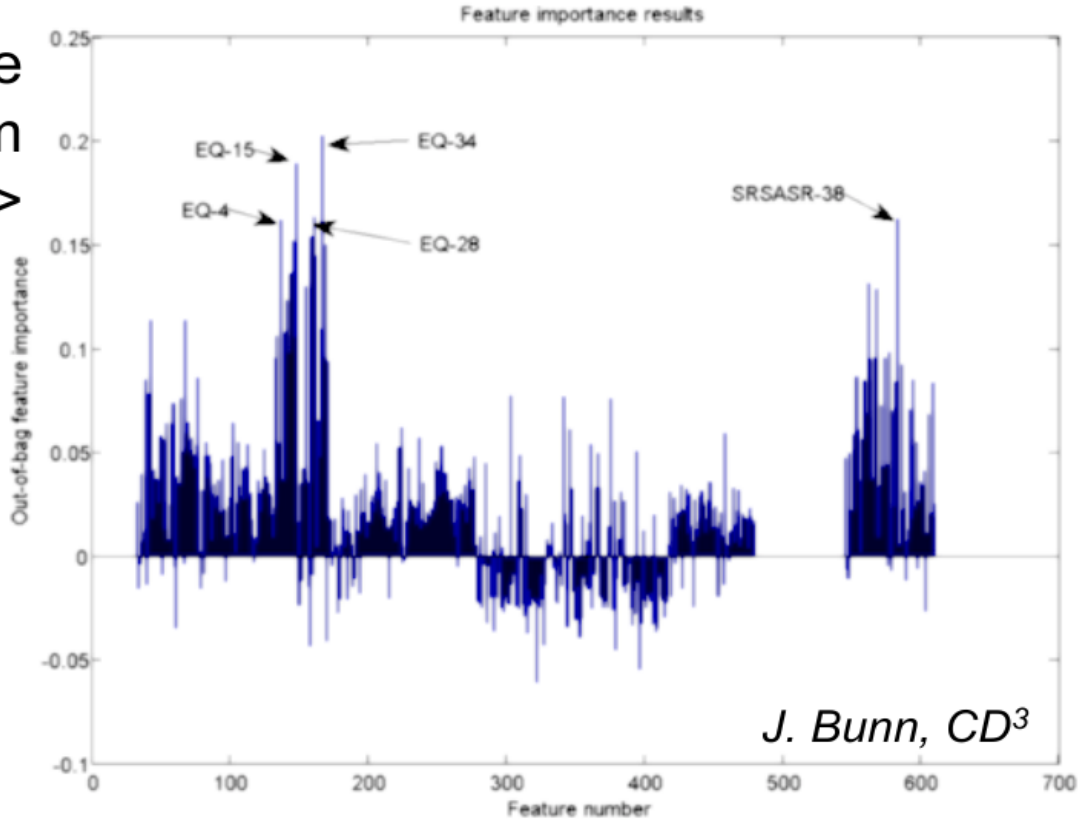
- A new research center at Caltech
  - Serves research efforts Institute-wide
- A part of a new, Caltech-JPL joint initiative for data science and technology
- The goals are to assist faculty in **formulation and execution of data-intensive projects**, and facilitate **interdisciplinary sharing of methods, ideas**, novel projects, etc.

Astro-Neurology

From Sky Surveys to Neurobiology

- Using the data analytics tools based on ML, developed for the analysis of sky surveys, to design a better diagnostics for autism

- Feature importance using random forests =>



- Next: correlate with MRI scans

(with R. Adolphs et al.)

U-Science, Carbon Computing

e-Science emerged ~10 yrs ago using the web protocols that were common at that time:
– web services, XML-based information exchange, registries, distributed data access, distributed computing (Grid) = machine-to-machine communication

U-Science is now emerging from today's web protocols:
– social networking, ubiquitous devices, user-centric experiences, user-led activities, user-generated content, wikis, blogs, mashups, tagging, annotation, ontologies (semantic web), folksonomies, knowledge-sharing, user recommendations = user-to-user communication

- The emergence of Citizen Science:

– Anybody can participate in the science discovery process

- Anyone can annotate, tag, and label scientific results:

– scientists, students, and citizen scientists

From K. Borne AI2010

Scientific Communities

"The co-authorship network of scientists represents a prototype of complex evolving networks. In addition, it offers one of the most extensive database to date on social networks."[a]

[a]Barabàsi et al., "Evolution of the social network of scientific collaborations"

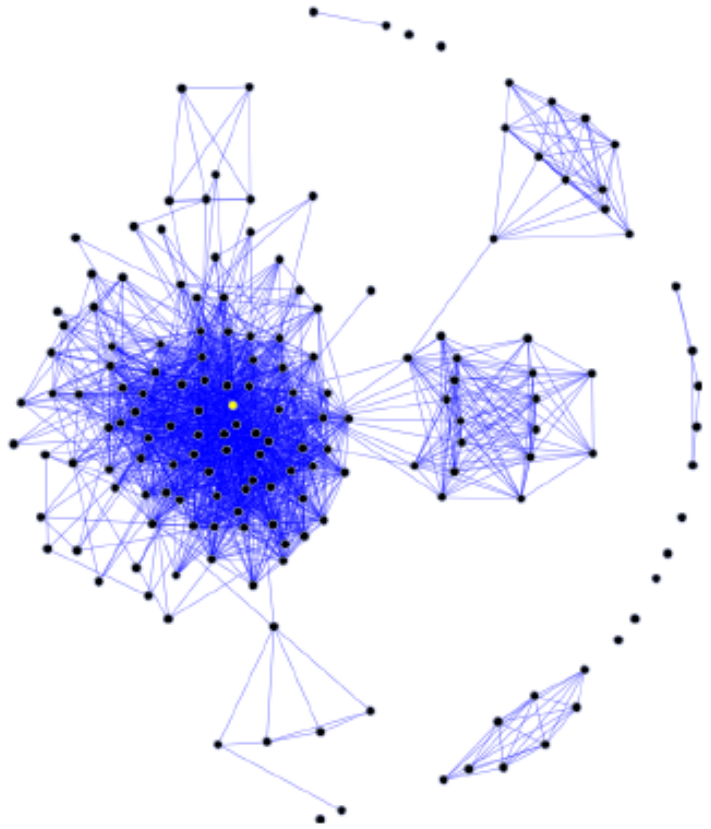
"Social scientists have long recognized the importance of boundary-spanning individuals in diffusing knowledge (Allen 1977; Tushman 1977), and recently, several papers have rigorously demonstrated that technological knowledge diffuses primarily through social relations, not through publications."[a]

[a]Sorenson, and Singh, "Science, Social Networks and Spillovers"

From O. Laurino - AI2010

Motivations of a social networking IT platform for science

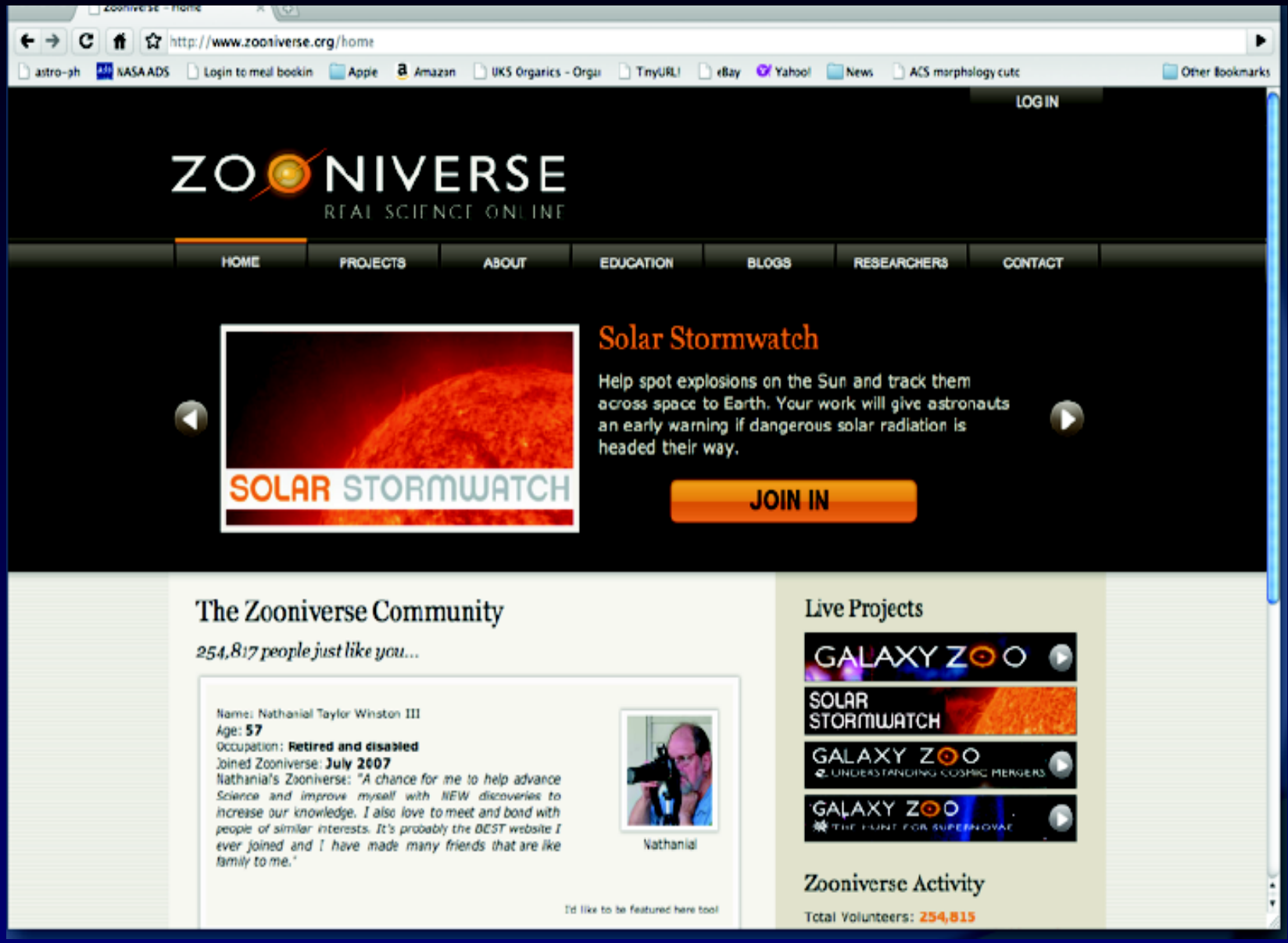
- The importance of boundary-spanning individuals in social networks might be what X-informatics is all about;
- we break scientific *cliques* and create new, unexpected, effective links across the science community's network;
- an effective scientific social network platform may be an effective step towards *seamless astronomy*. Seamless not only in terms of data and applications access, but also in terms of social interactions between people in the scientific network.



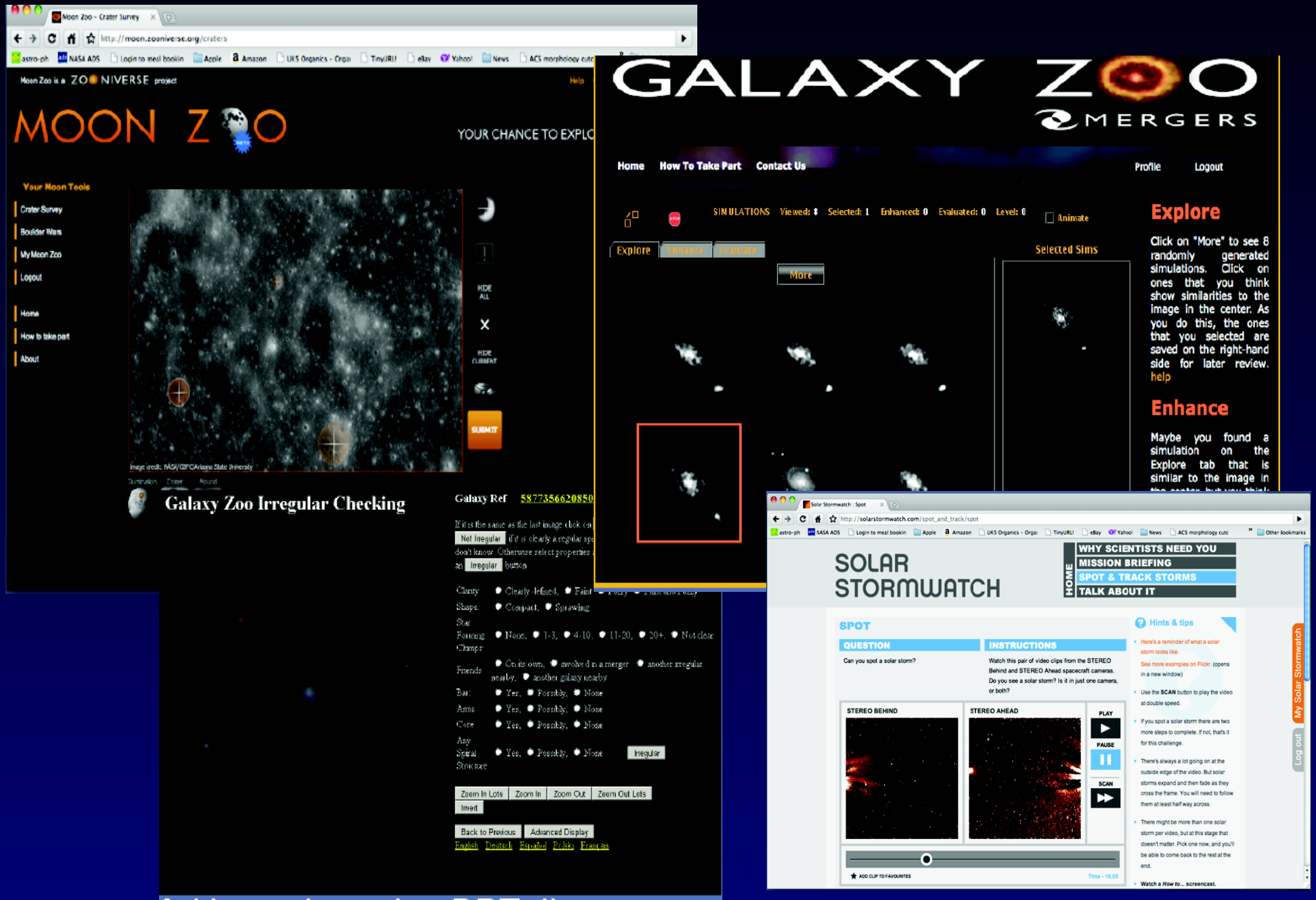
Galaxy ZOO

> 20 Science papers published so far

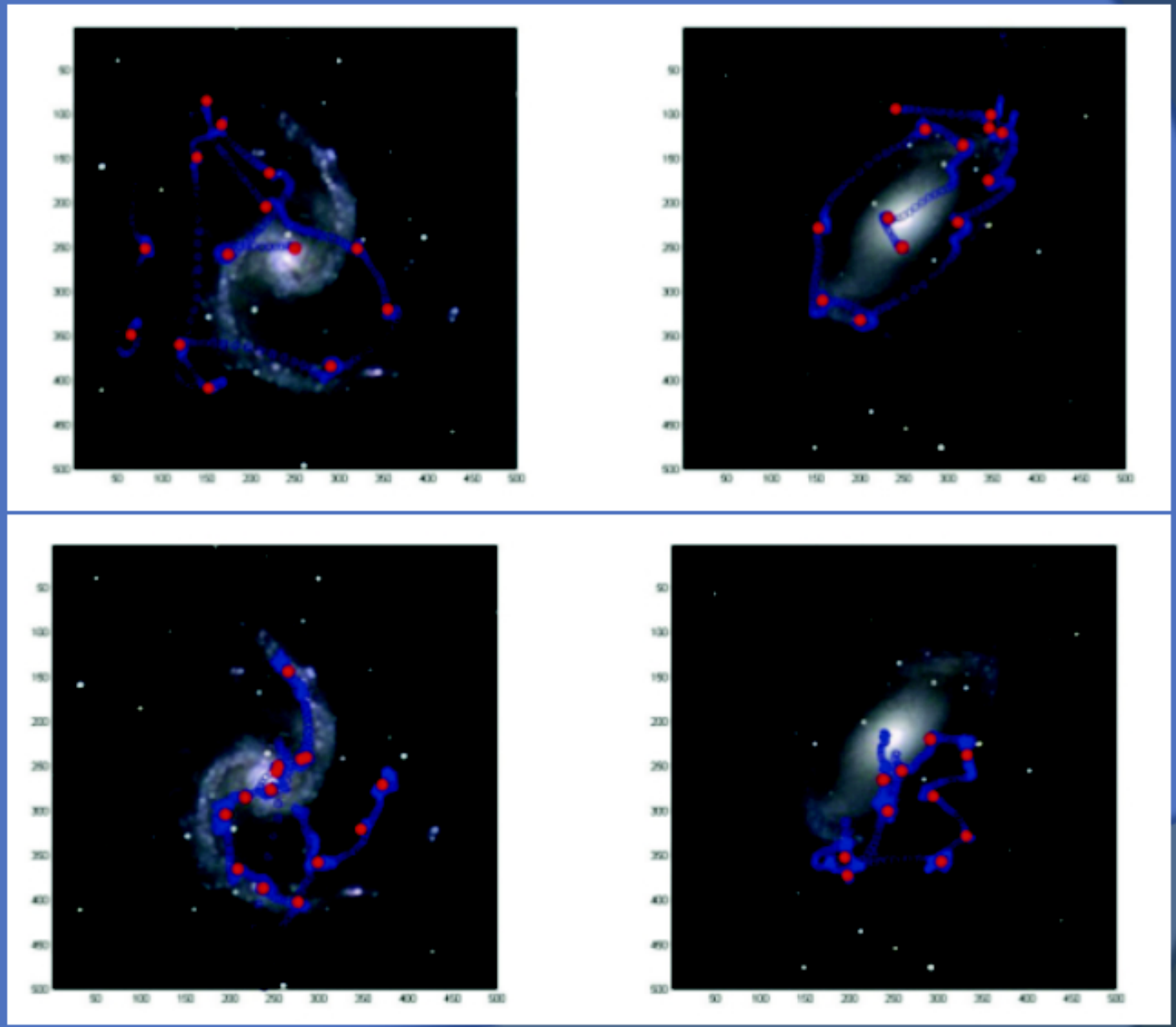
www.zooniverse.org



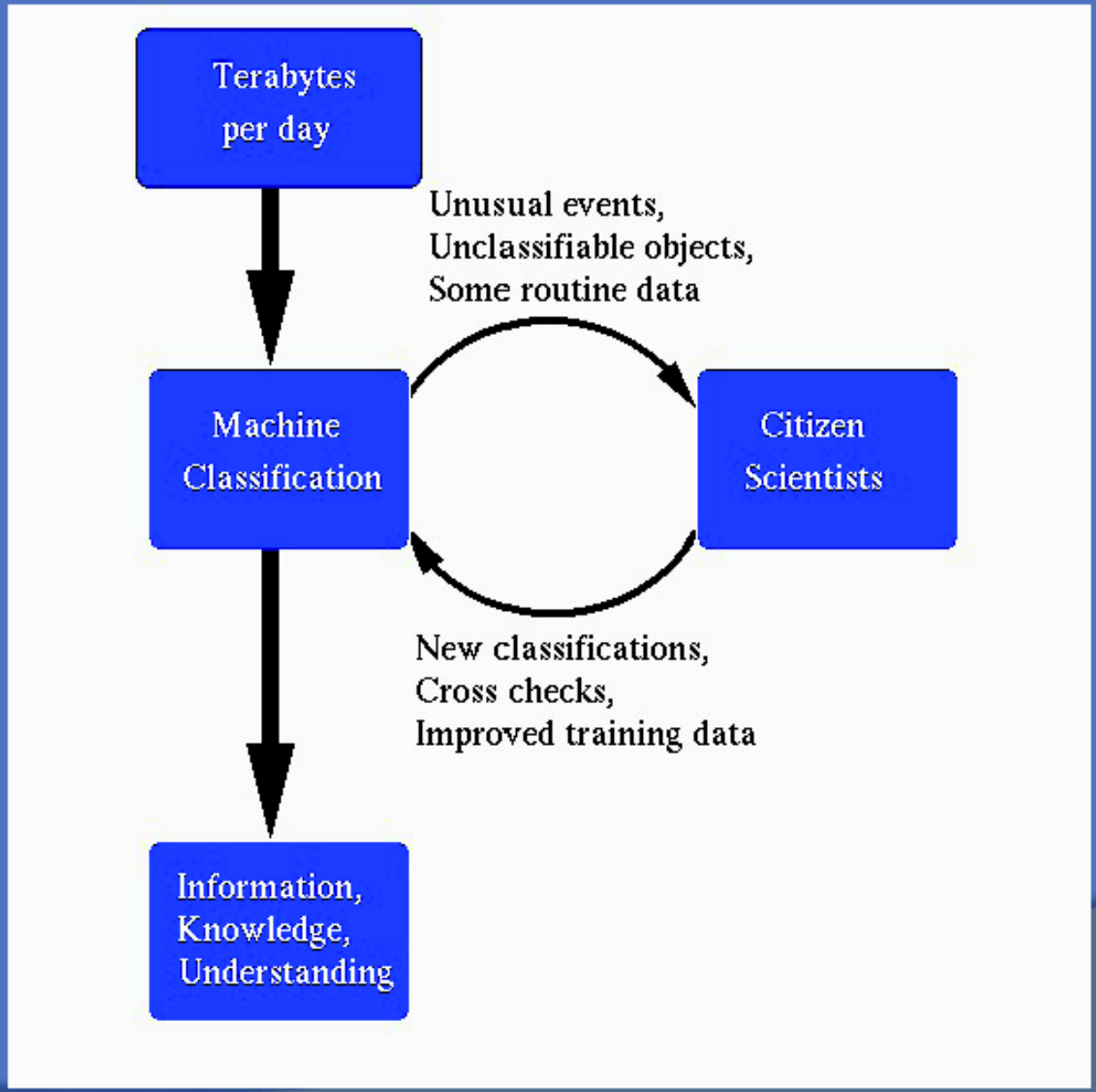
Examples ZOOniverse



Expert vs Non-expert Classifier



Machine-Human Learning Cycle

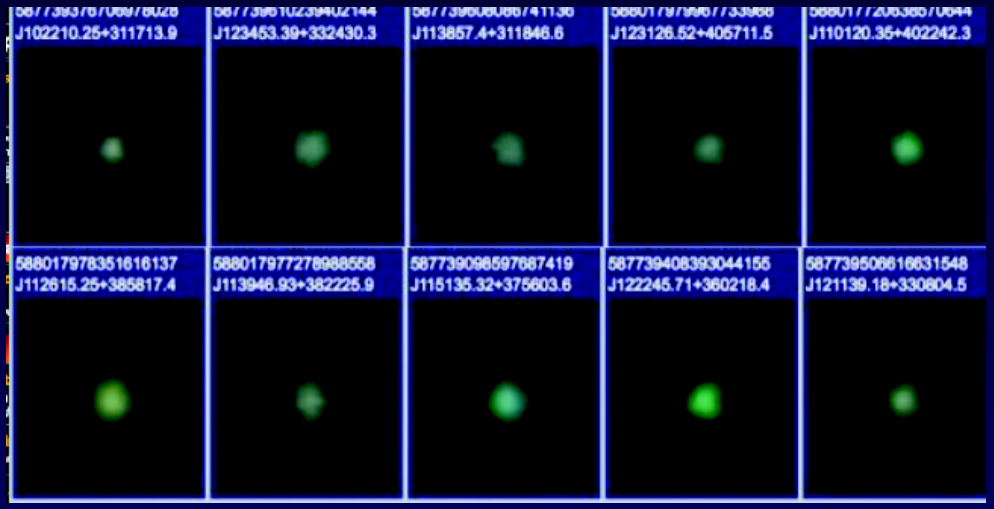


Citizen Science x Expert Science

Verified by human – training sets

Independent answers=estimate of error

Serendipitious discovery



Galactic Peas

Scale - complexity

The answer is Data mining …. matching Donald Rumsfeld's epistemology

There are known knowns,
There are known unknowns, and
There are unknown unknowns

Classification
Morphological classification of galaxies
Star/galaxy separation, etc.

Regression
Photometric redshifts

Clustering
Search for peculiar and rare objects,
Etc.

"There are known knowns.
These are things we know that we know.
There are known unknowns.
That is to say, there are things that we know we don't know.
But there are also unknown unknowns.
There are things we don't know we don't know."

Longo 2010

Knowledge Discovery in U-Science

Known knowns :
Primary task. Data reduction by science team.

Known unknowns :
Related to primary task. Results funneled to specific researchers.

Unknown unknowns :
Serendipity. Currently rely on forum moderators to filter.

Hanny van Arkel - Voorwerp

Light echo of quasar?

Challenges of (Astro)informatics

Big Data 3(5)xV

- Complex
- Missing values
- Censoring
- Upper limits
- Parallelization (Massive - GPU – new algorithms)
- Queries in PB table
- Visualization of many dimensions
- Stream processing
- Non- Gaussian Statistics, PDF

Příklady BP a DP na FIT z astroinformatiky a VO

- FIT VUT Brno
  - 1 BP (Random Forests in Astronomy)
  - 1 PhD – Wavelets Dimensionality Reduction (pending)

- VŠB-TU Ostrava
  - 1 BP + 1 DP - SPLAT-VO

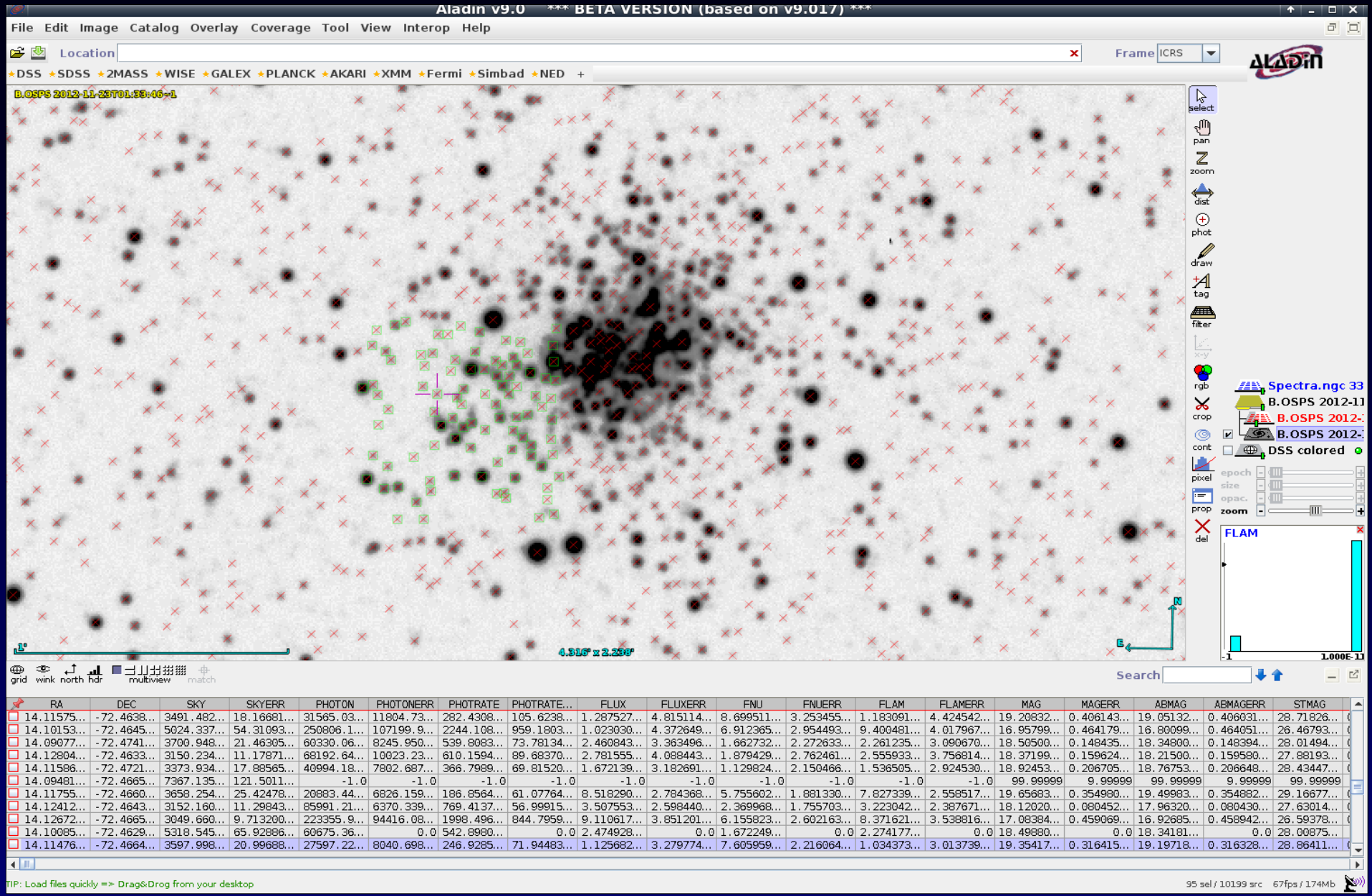
- MU PřF Brno
  - 2 DP + 2 PhD. (ML of Spectra (pending) + precise RV for exoplanets – SW ?)

- FIT ČVUT
  - 2012 2 BP (VO-Korel+SSA proxy)
  - 2013 2 BP (OSPS Image + Catalogue Server)
  - 2014 2 BP (Random Forests + SOM)
  - 2015 1 BP (VO-Cloud)
    2 DP (Clustering OSPS + Deep Learning)
  - 2016 2 DP (Semisupervised learning + Outlier finding)
  - 2017 1 DP + 2 BP ????

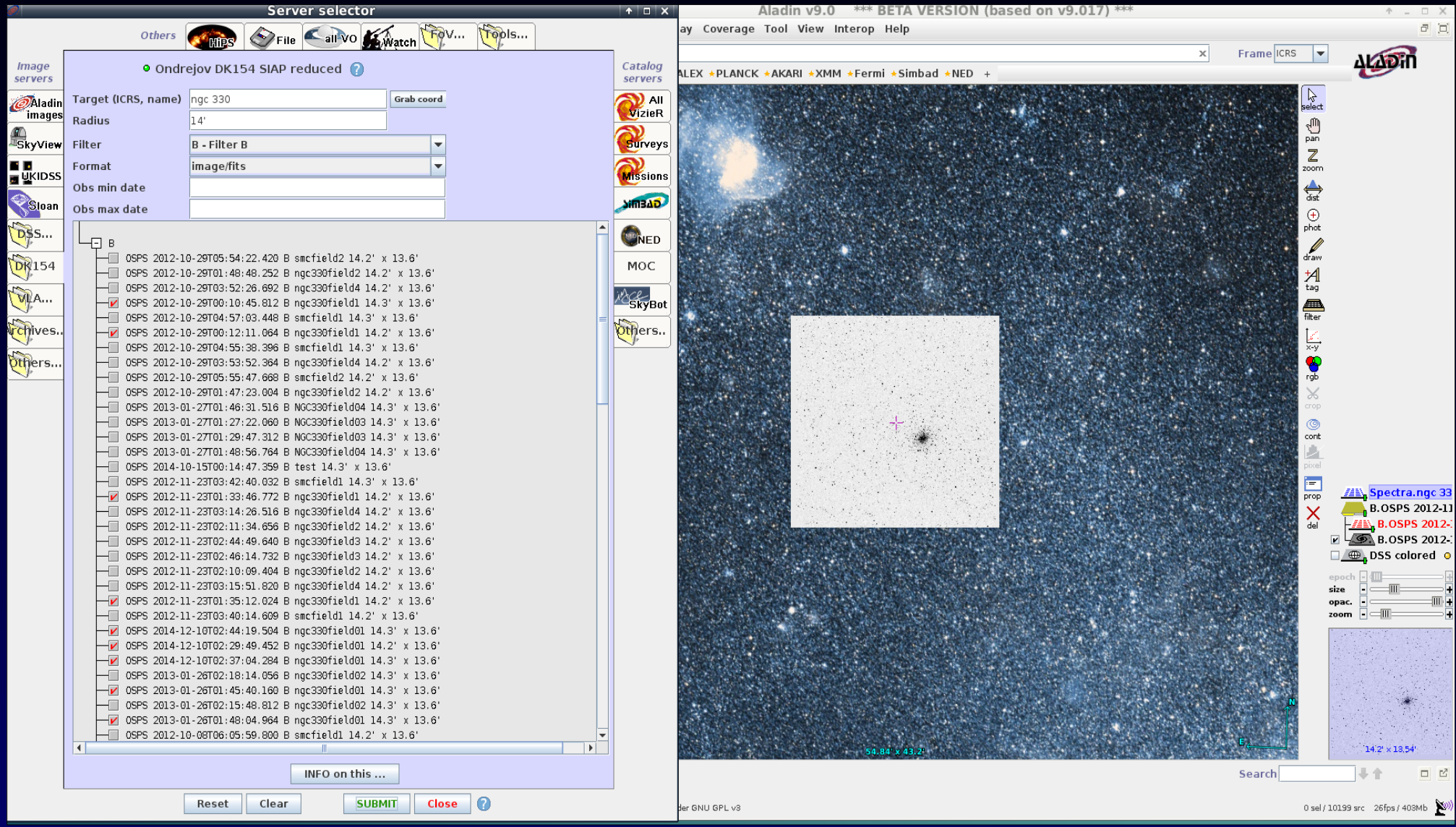
Danish 1.54m at La Silla robotized in Summer 2012

Danish 1.54m Telescope

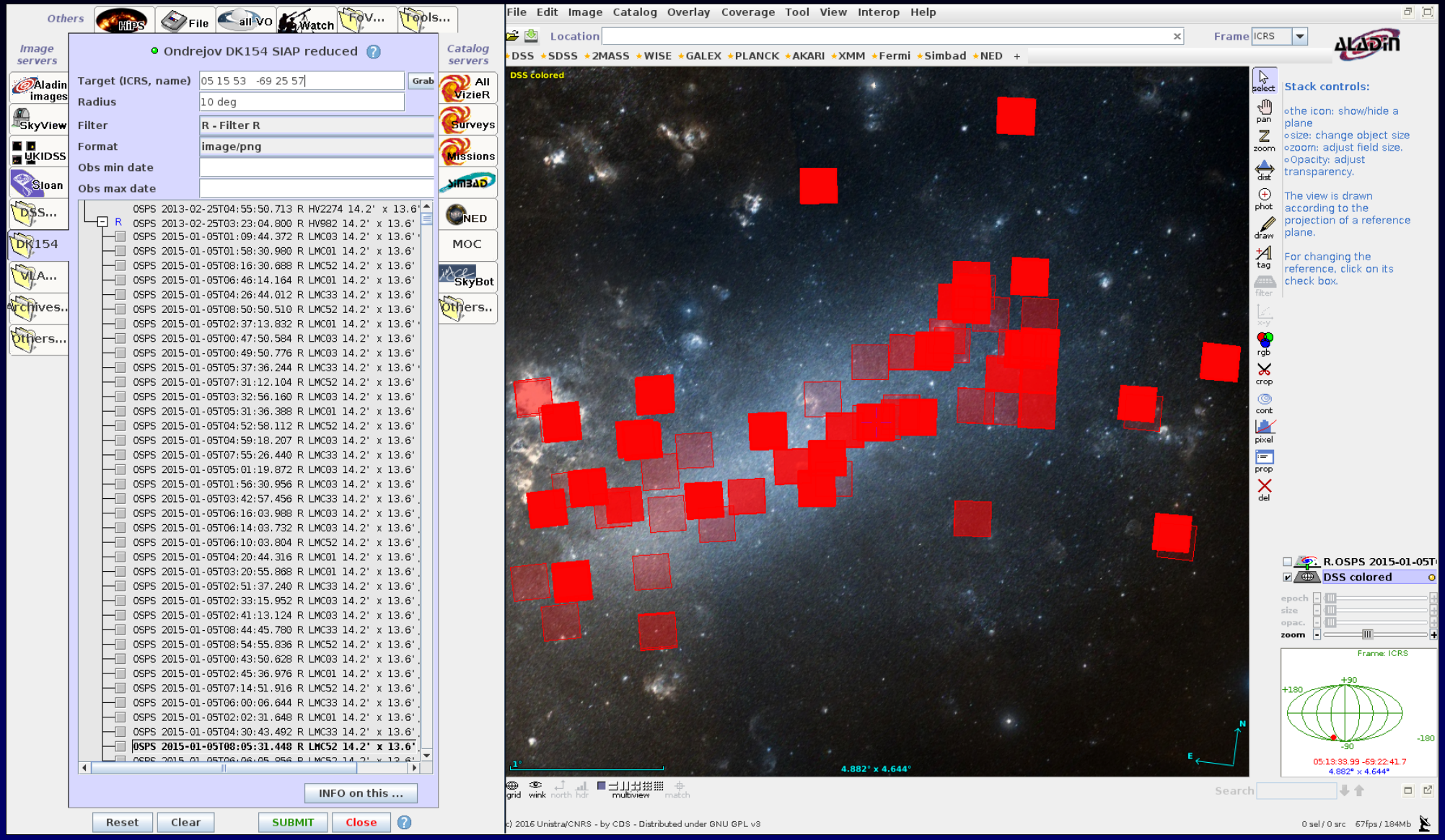
Reduced OSPS image + bintable photometry in 2nd extension



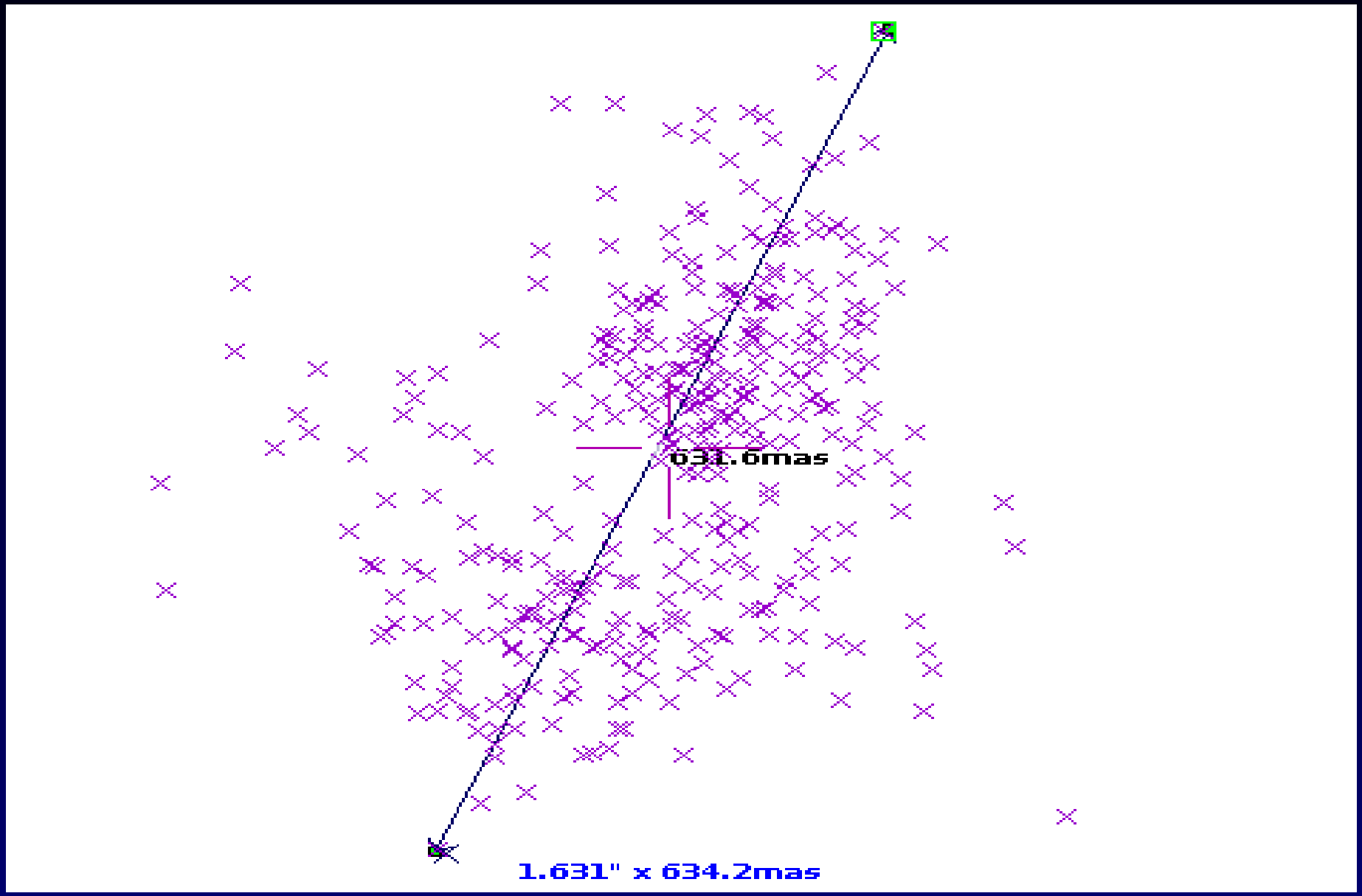
OSPS SIAP in Aladin (DSS in back)



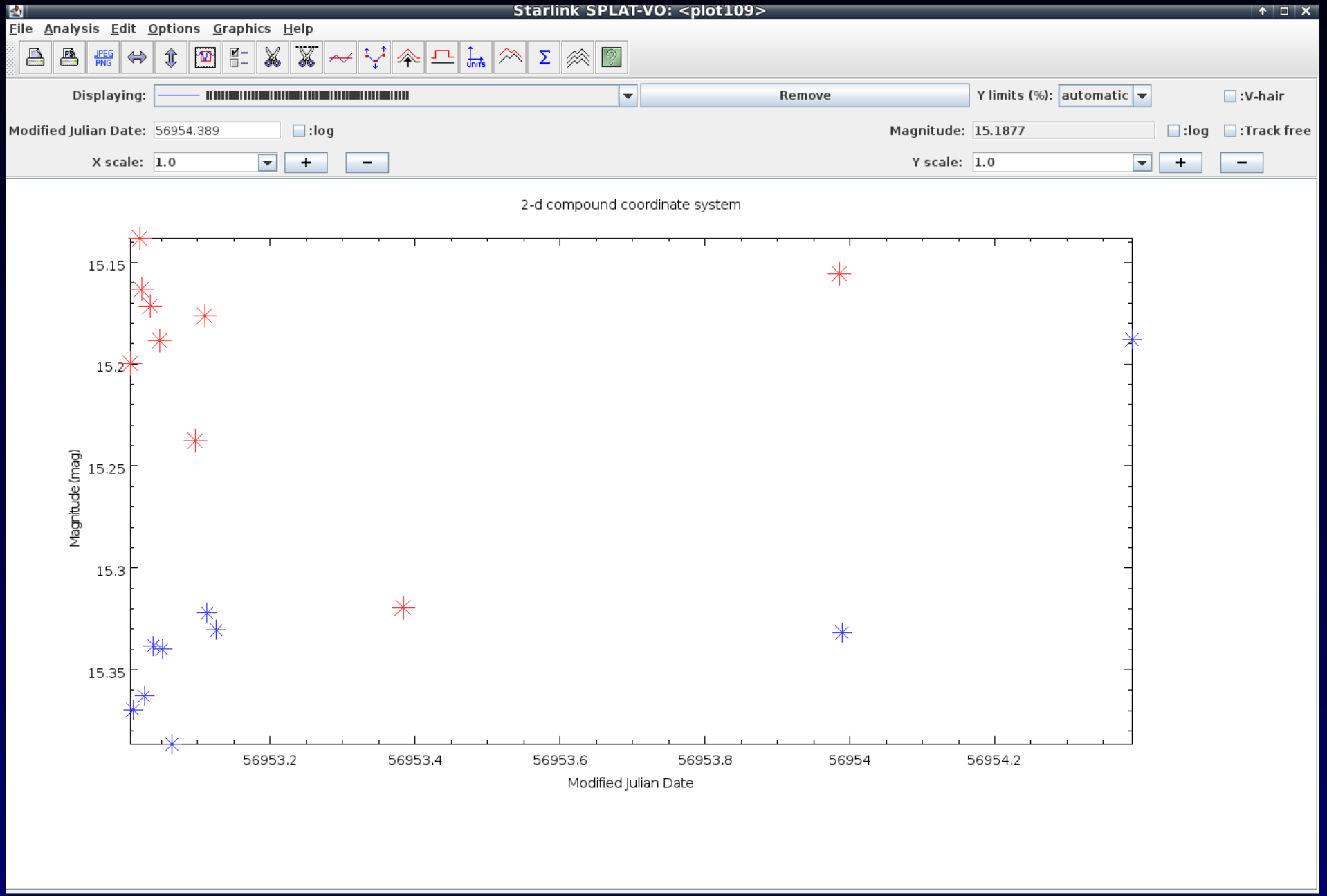
OSPS Image coverage (footprints)



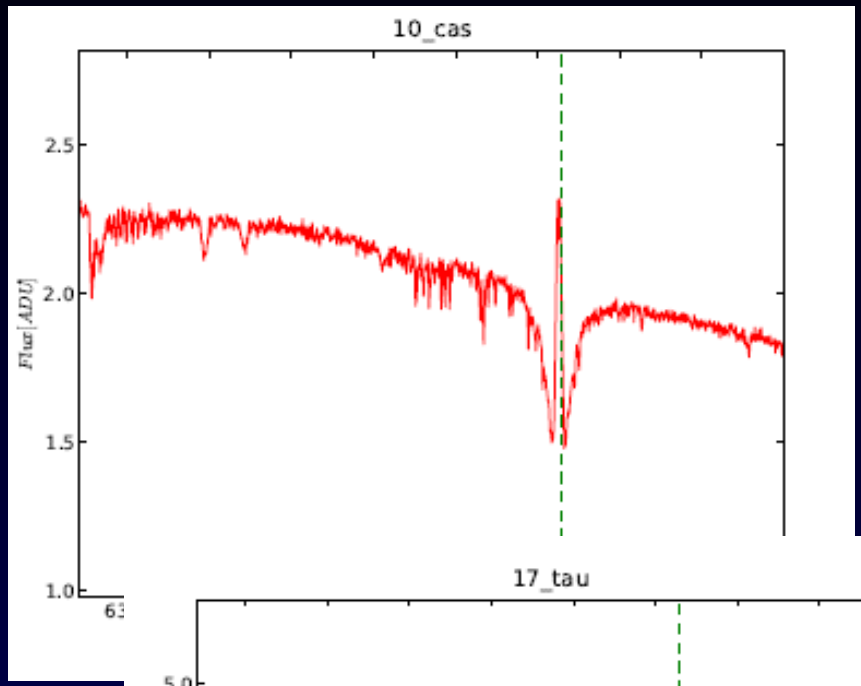
Parallelized Clustering of Positions

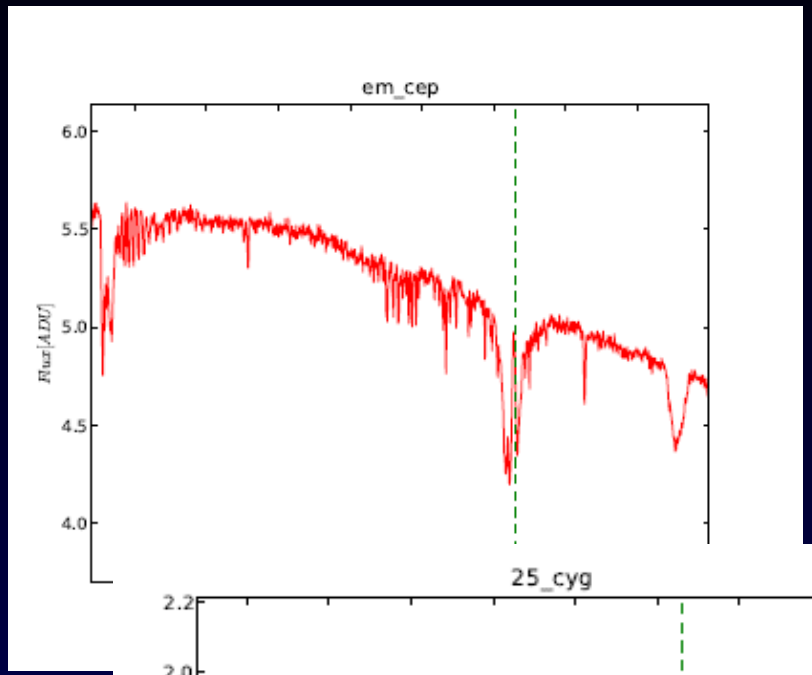


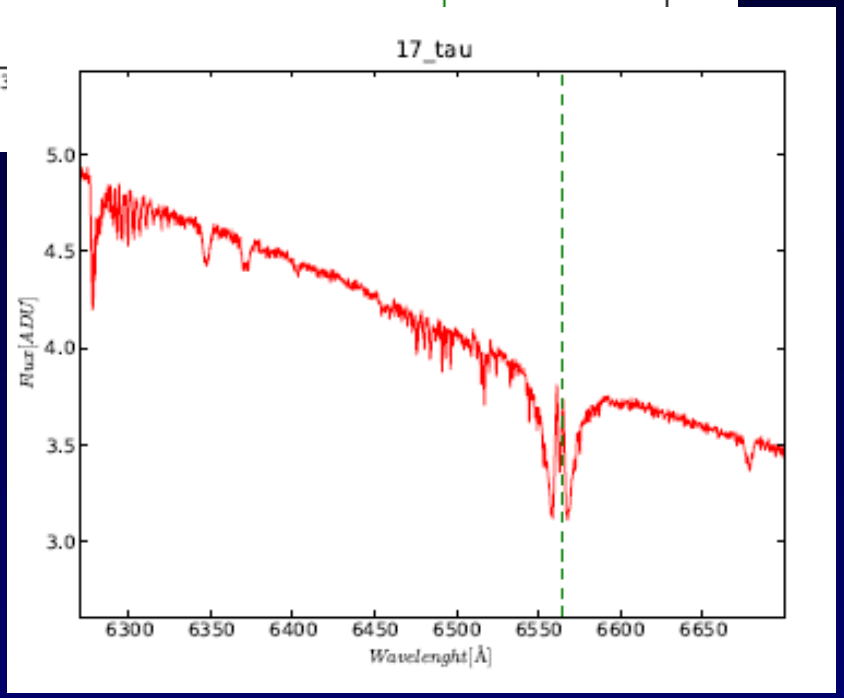
OSPS Light Curve in SPLAT

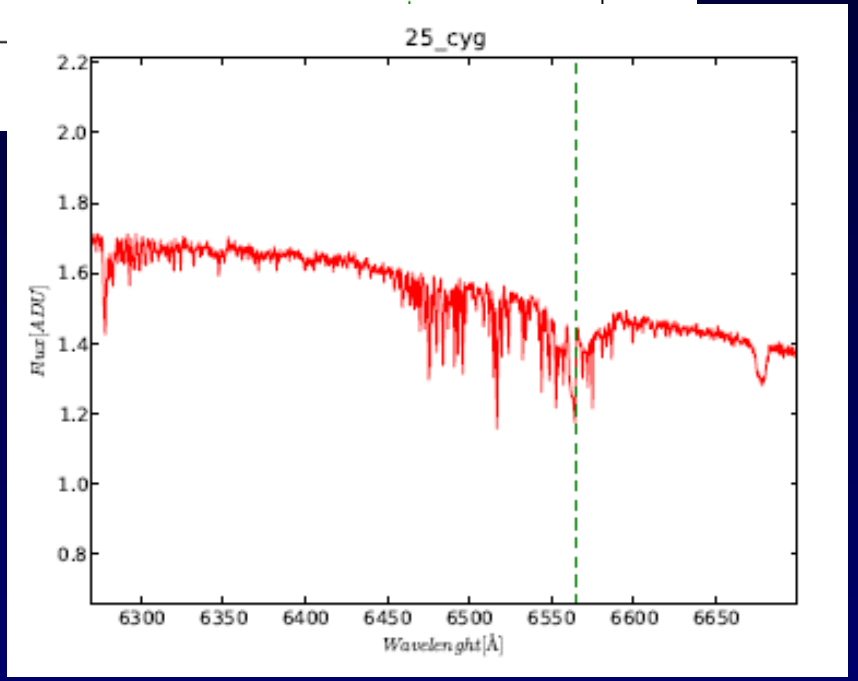


Be Stars : Emission in absorption



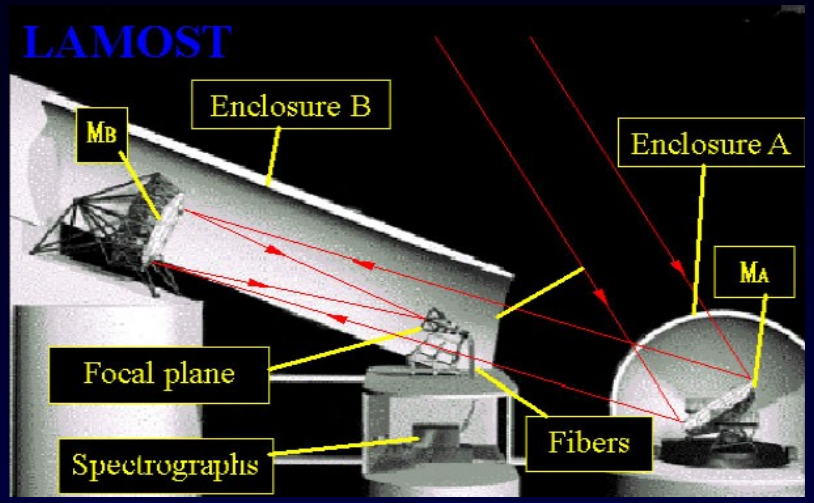


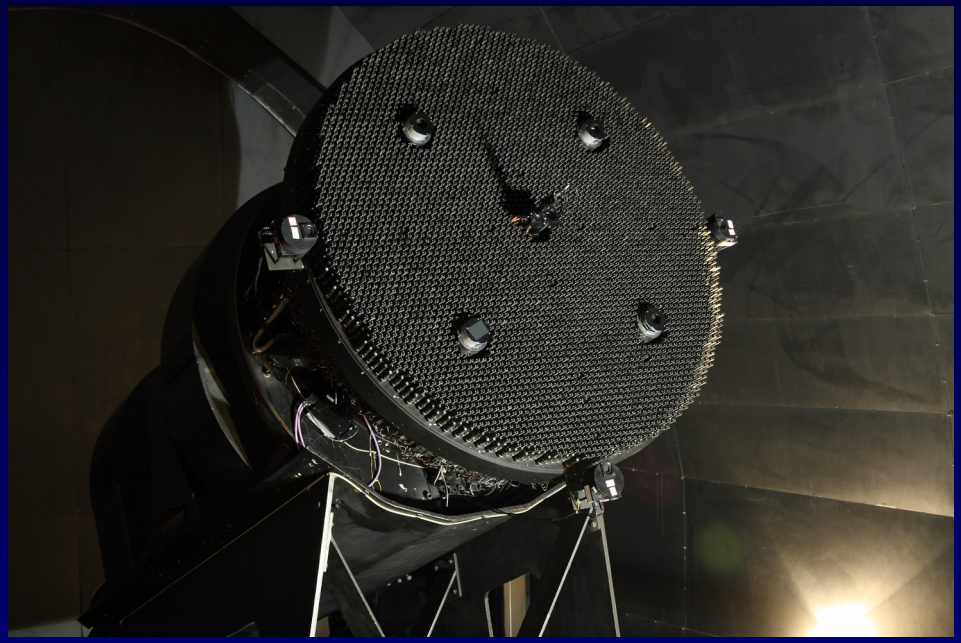


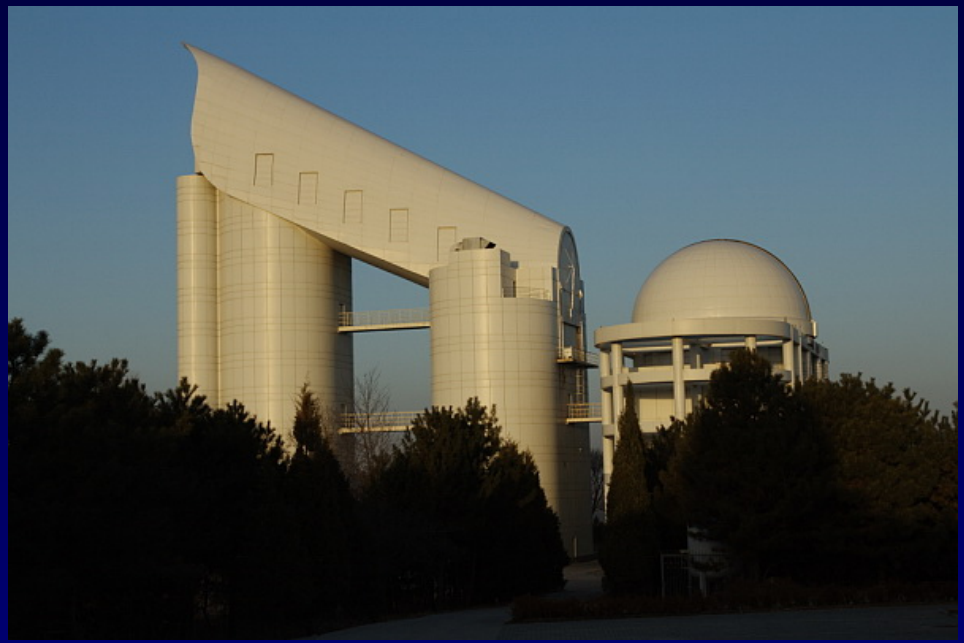


LAMOST (Guoshoujing)

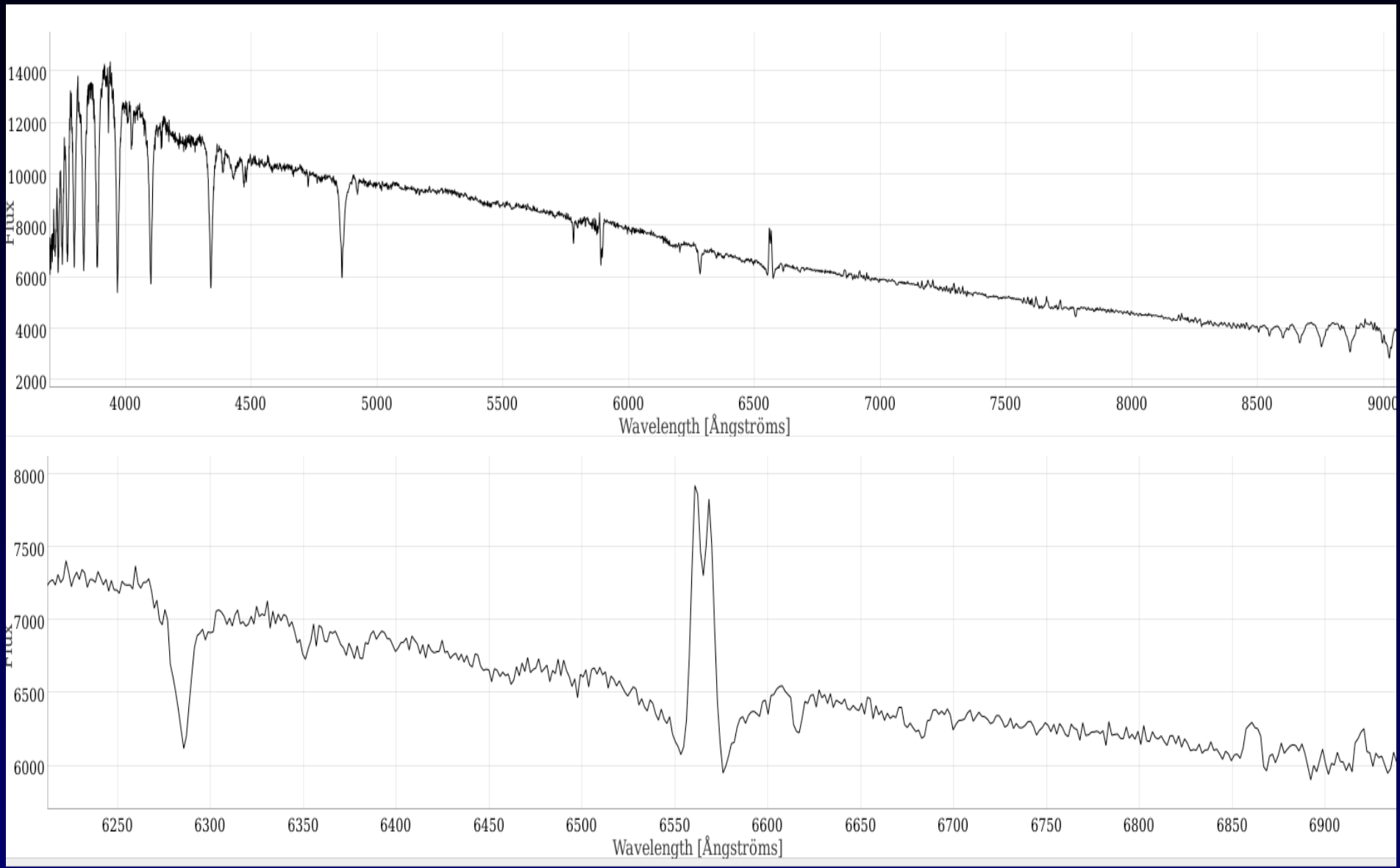
Xinglong- China
4m mirror (30 deg meridian)
4000 fibers
10 mil spectra / 5 yr
Automatic RV-z



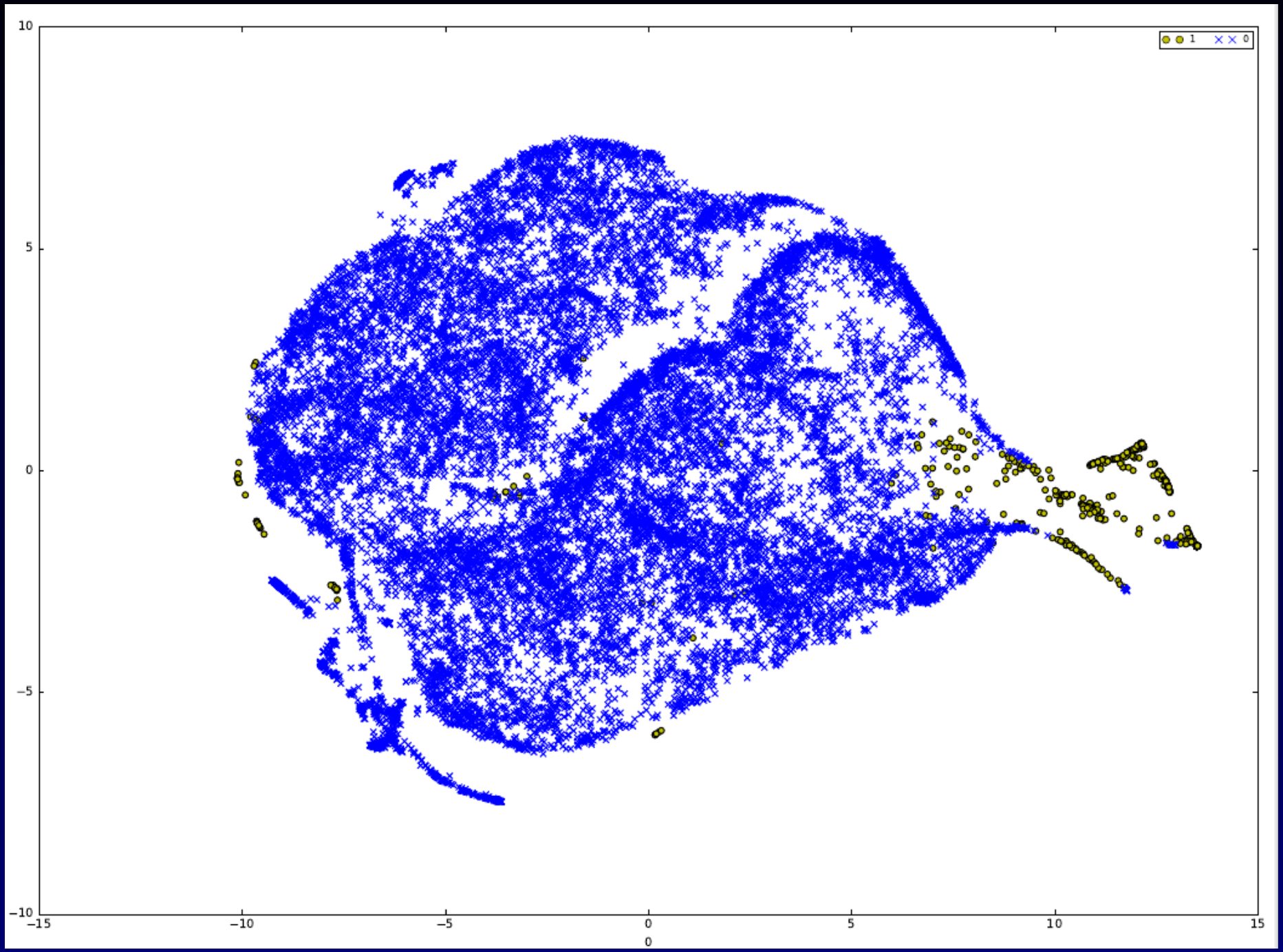




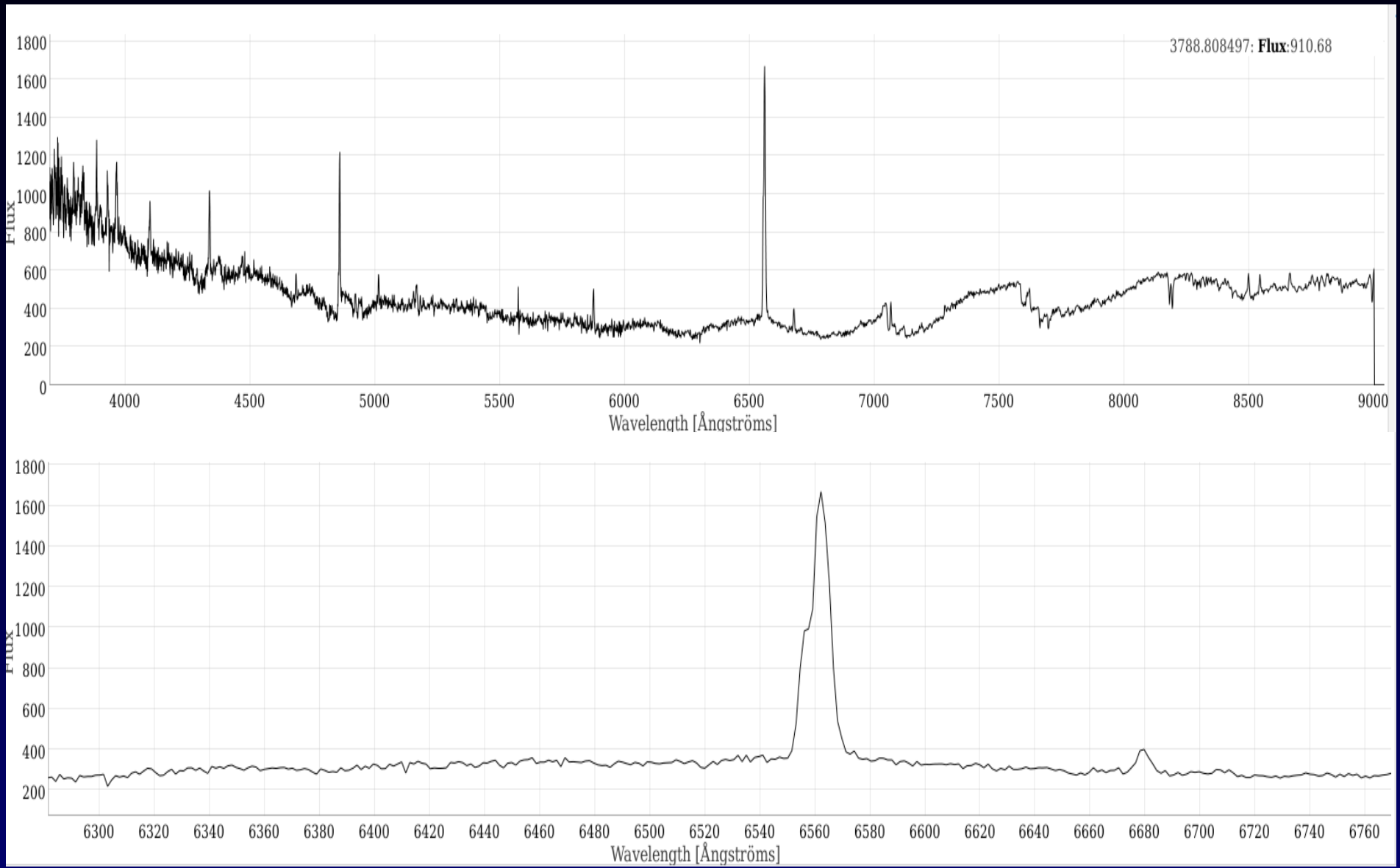
Be Candidates Found



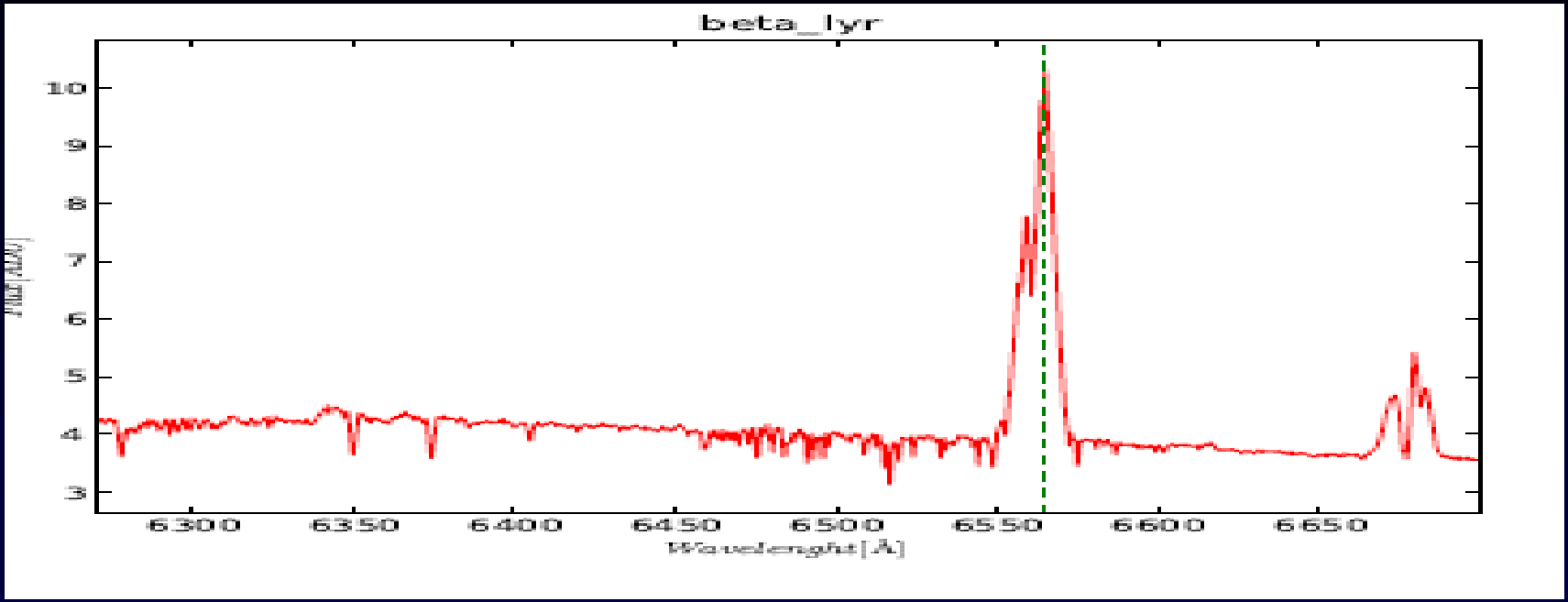
LAMOST TSNE Structure

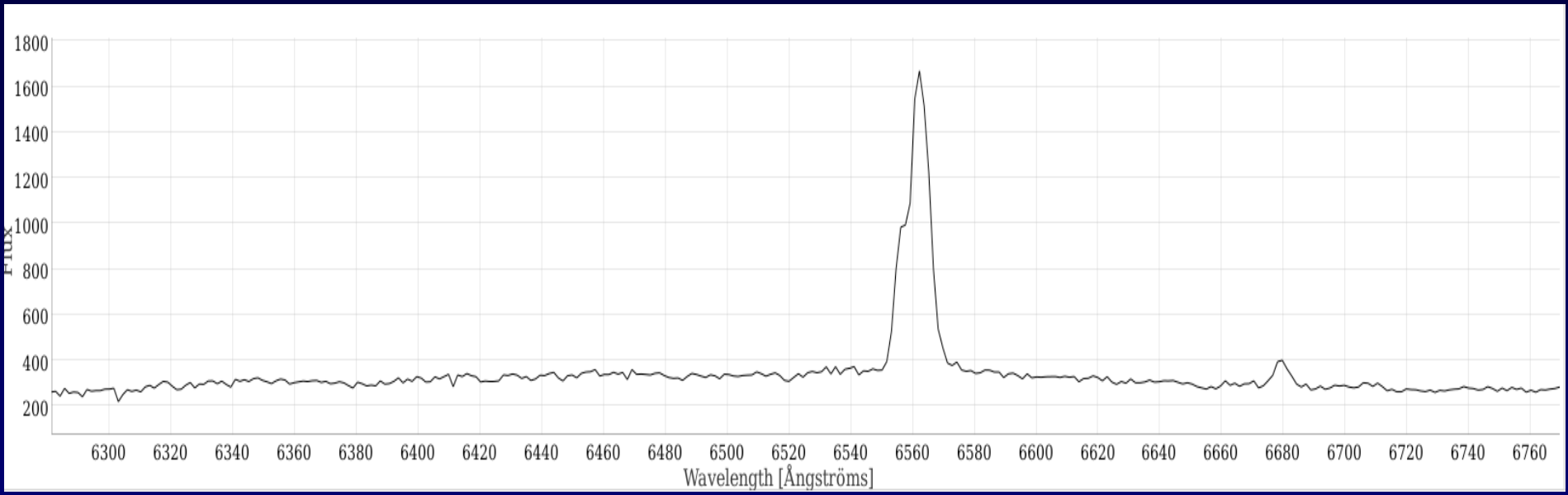


Be Candidates Found

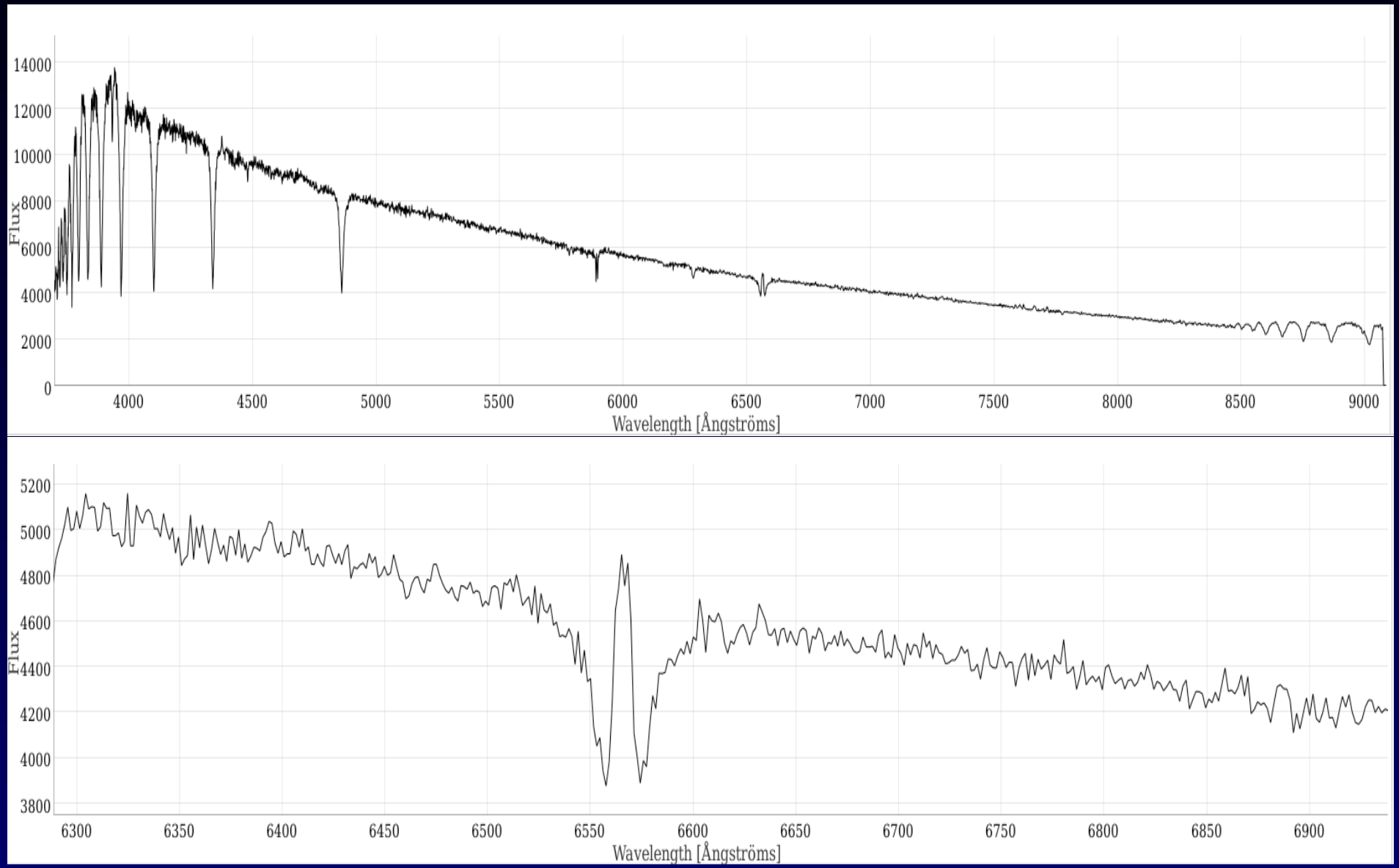


Be Candidates Found

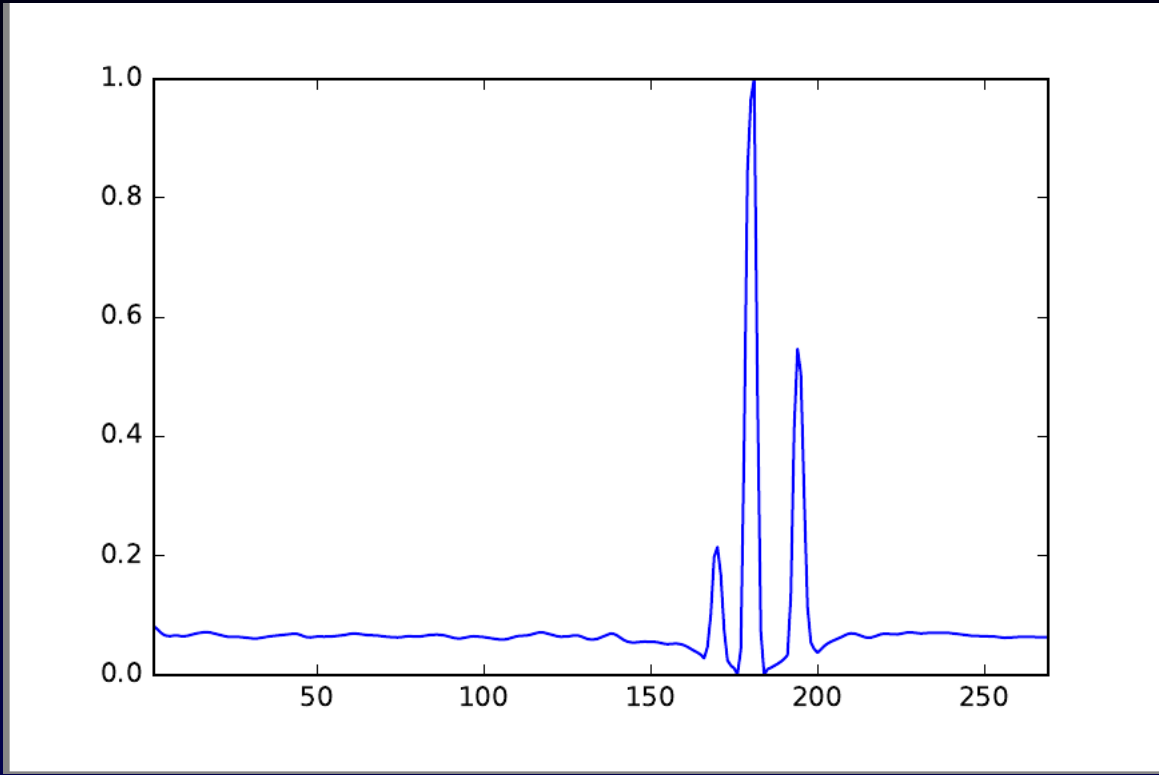


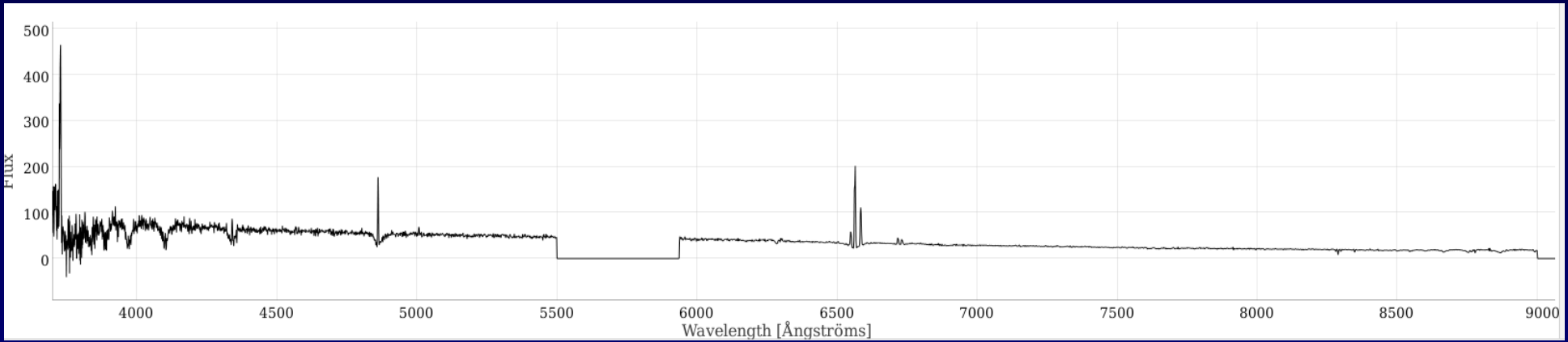


Yet Unknown Be Star

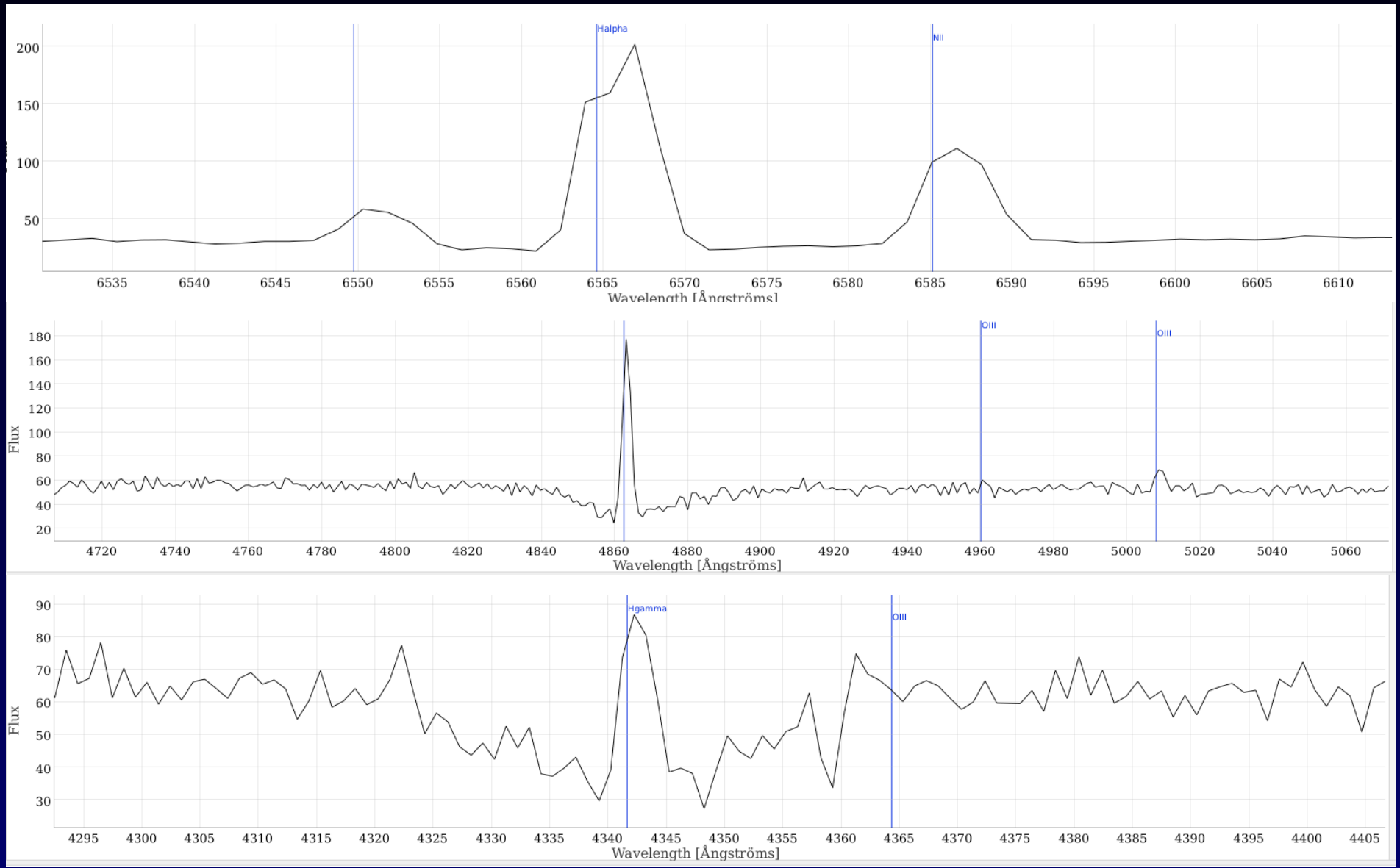


Be Candidates Foud

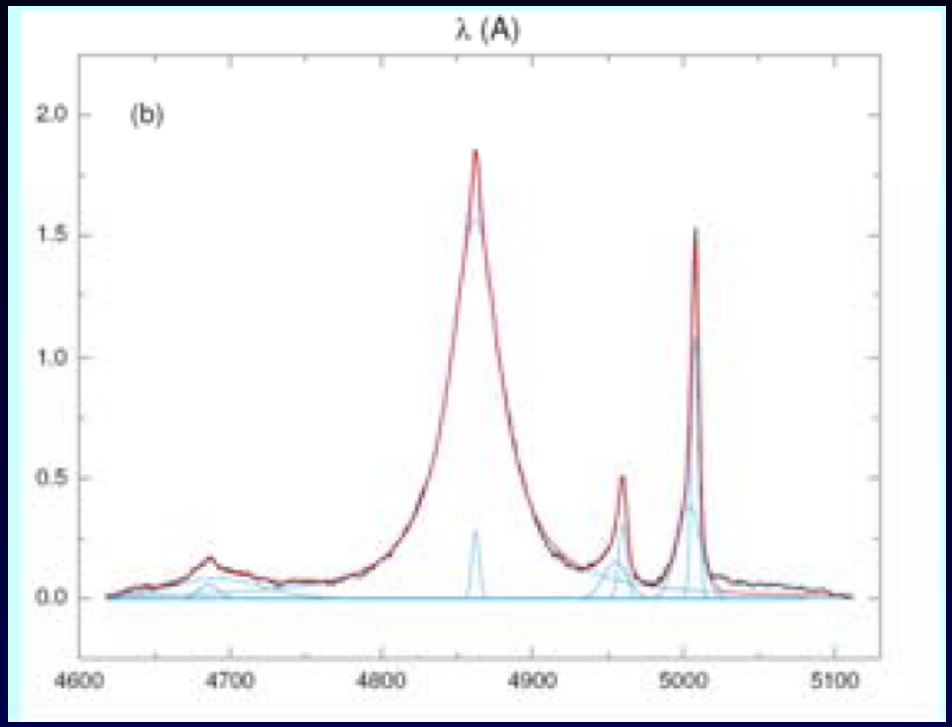




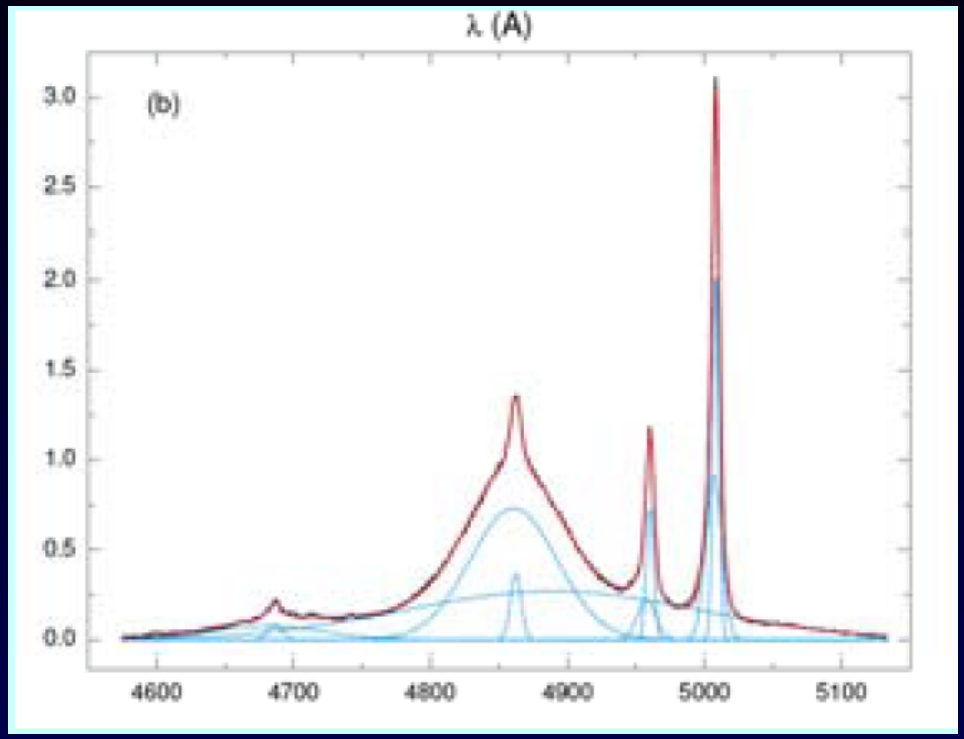
Be Candidates Found



AGN Populations



Population A



Population B

Sulentic et al. 2002