

Astroinformatics - Extracting New Knowledge about Universe

in the Epoch of Petabyte-Scaled Archives

Petr Škoda

Astronomical Institute of the Czech Academy of Sciences

Supported by grant COST LD-15113 of the
Czech Ministry of Education Youth and Sports

Python BootCamp
IAG, University Sao Paulo , Brazil
16th February 2017

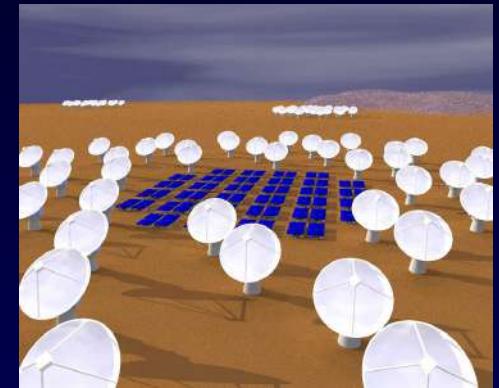
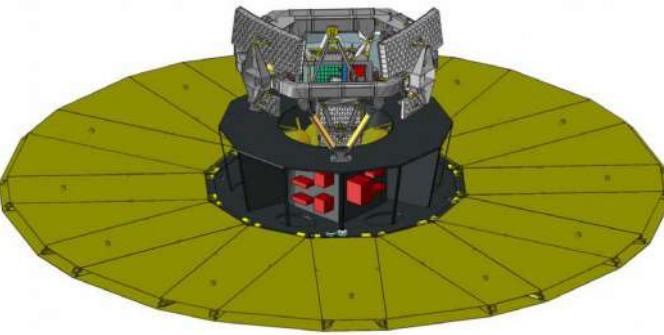
Credits

- 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, T. Hey, F. Le Petit, M. Breddels and presentations from AI2016 in Sorrento

Outline of the Talk

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

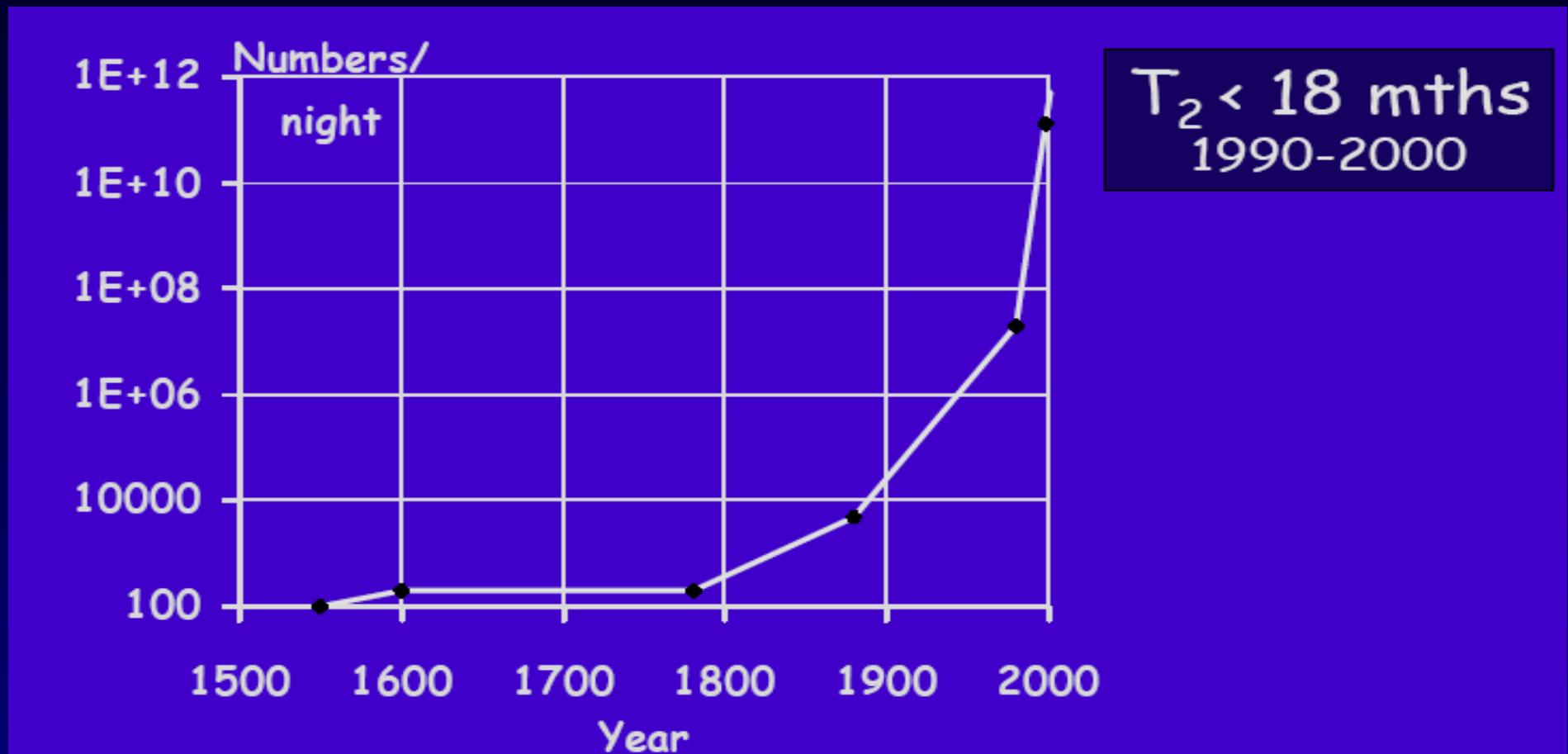
Data Avalanche



Data Avalanche

Moore law for chips –doubling 1.5 year

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



CD Sea

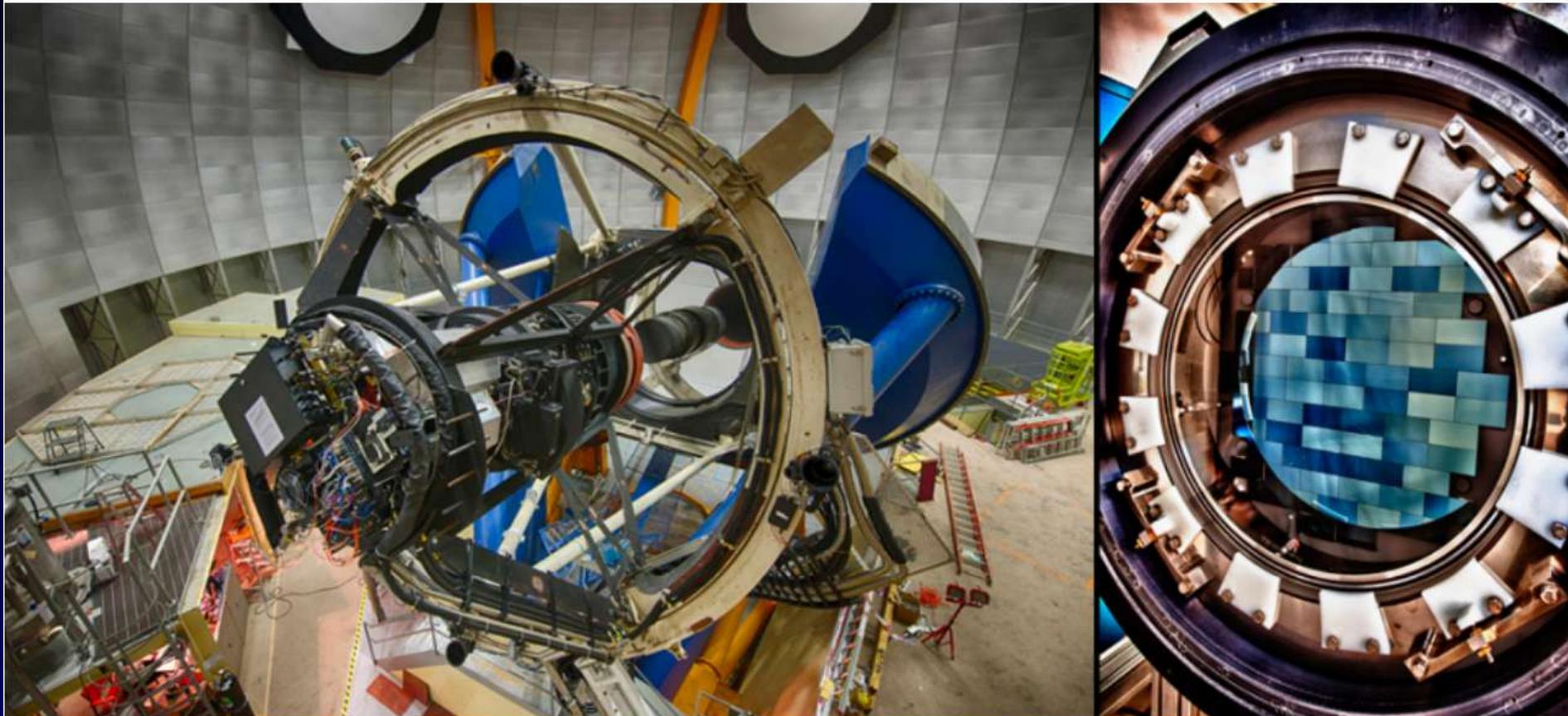


600 000 CD = 372 TB (CD 650MB)
600 000 DVD = 2.5 PB (DVD=4.5GB)

Bruce Monro
Kilmington UK

Dark Energy Survey Camera

Dark Energy Camera (DECam)



~0.4 PB/yr

Large Synoptic Survey Telescope



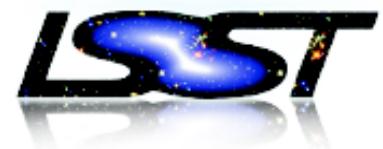
201 CCD 4kx4k,
3.2 Gpix every 20 sec
3.5 deg FOV (64cm)
20 TB/day=6 PB/yr RAW
1.5 PB catalogue !!!
detection of changes 60s!

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





LSST: Data Volume

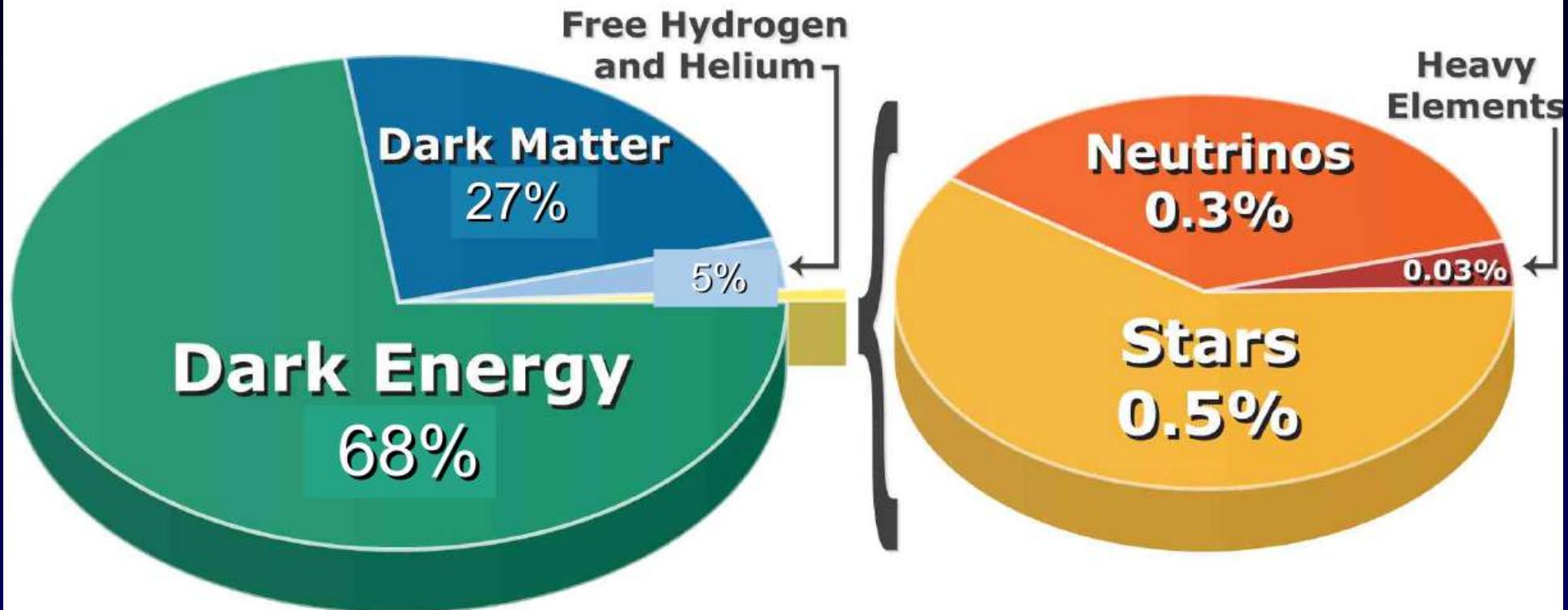


- One 6.4-gigabyte image every ~17 seconds
- ~1000 visits (two back-to-back images), per night
- 15 terabytes of raw scientific image data / night
-
- 8.4 terapixel image (movie) of the sky to ~27.5 mag in 6 bands
-
- A catalog of ~38 billion observed objects (24B galaxies, 14B stars)
- A catalog of ~32 trillion photometric measurements
-
- ~2000 events per observation (includes variables+asteroids)
- ~2 million events per night, for 10 years
- Requirement: Process & transmit alerts within 60 seconds

Project EUCLID

The Euclid mission main goal

EUCLID
CONSORTIUM

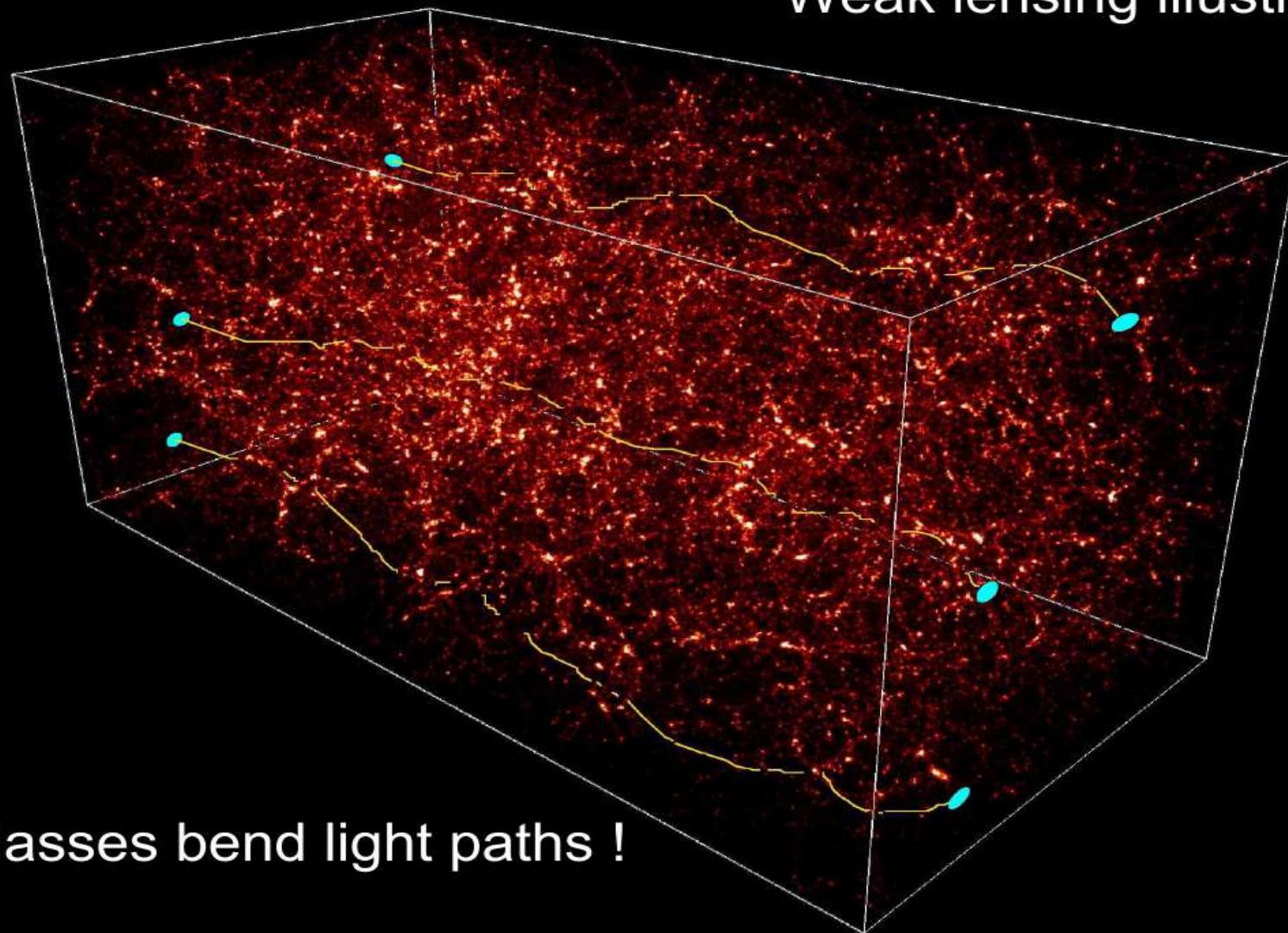


- What is the Nature of the Dark Matter and Energy?

EUCLID principles

DEFLECTION OF LIGHT RAYS CROSSING THE UNIVERSE, EMITTED BY DISTANT GALAXIES

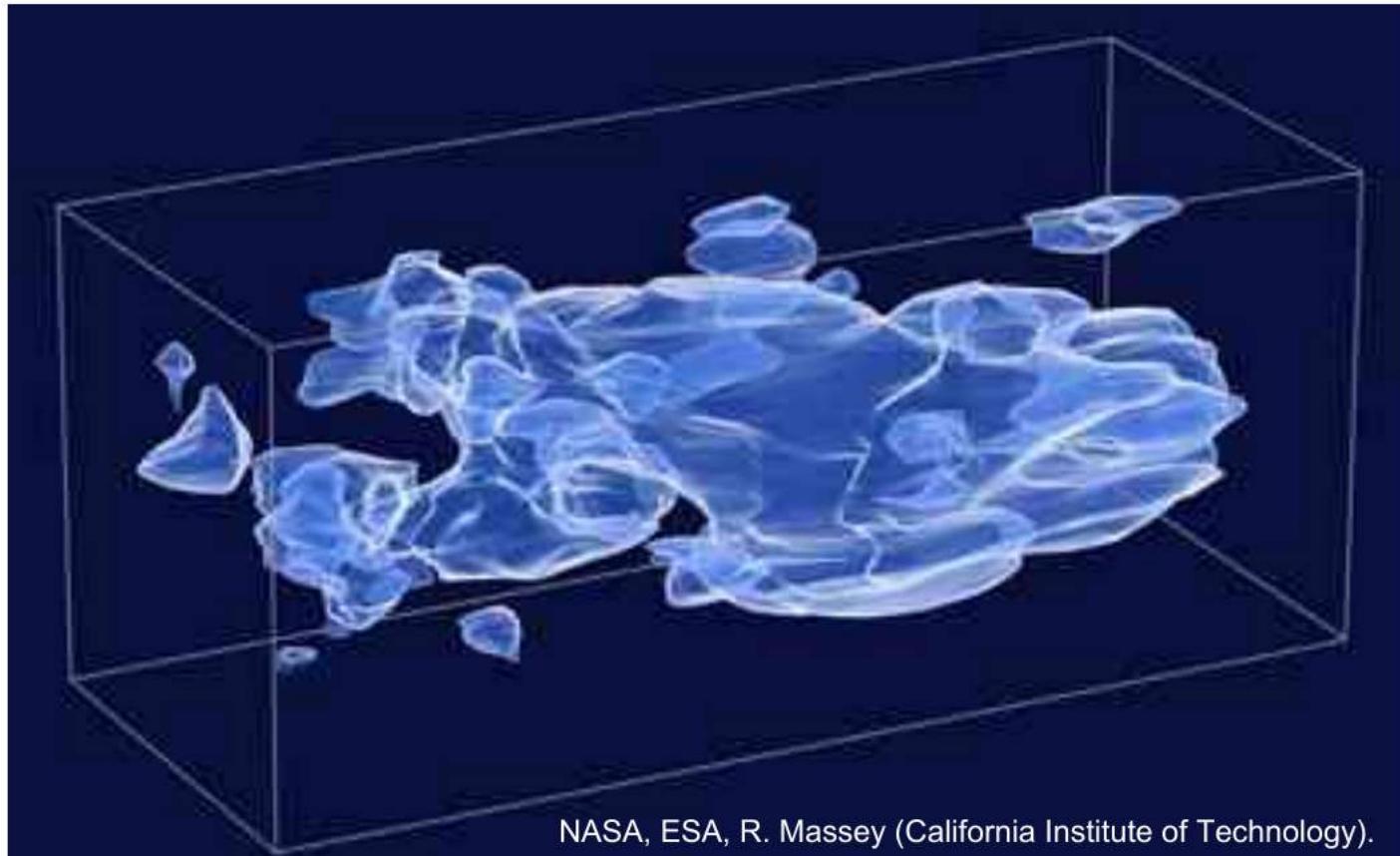
Weak lensing illustration



Masses bend light paths !

SIMULATION: COURTESY NIC GROUP. S. COLOMBI IAP

Euclid Data Archive



NASA, ESA, R. Massey (California Institute of Technology).

	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 Millimeter Array ALMA

64 antennas 12m
Chajnator 5000m
Chile
2008-2013

it is spectrograph
as well as ...

0.5-2 PB/yr RAW



LOFAR network

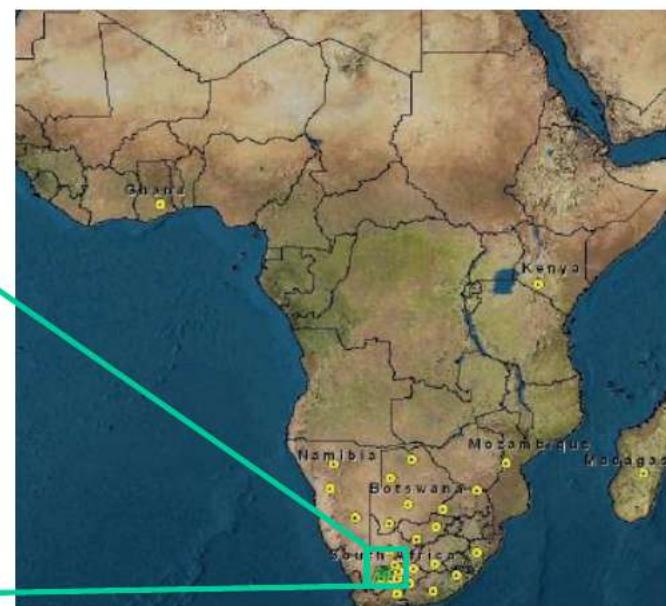
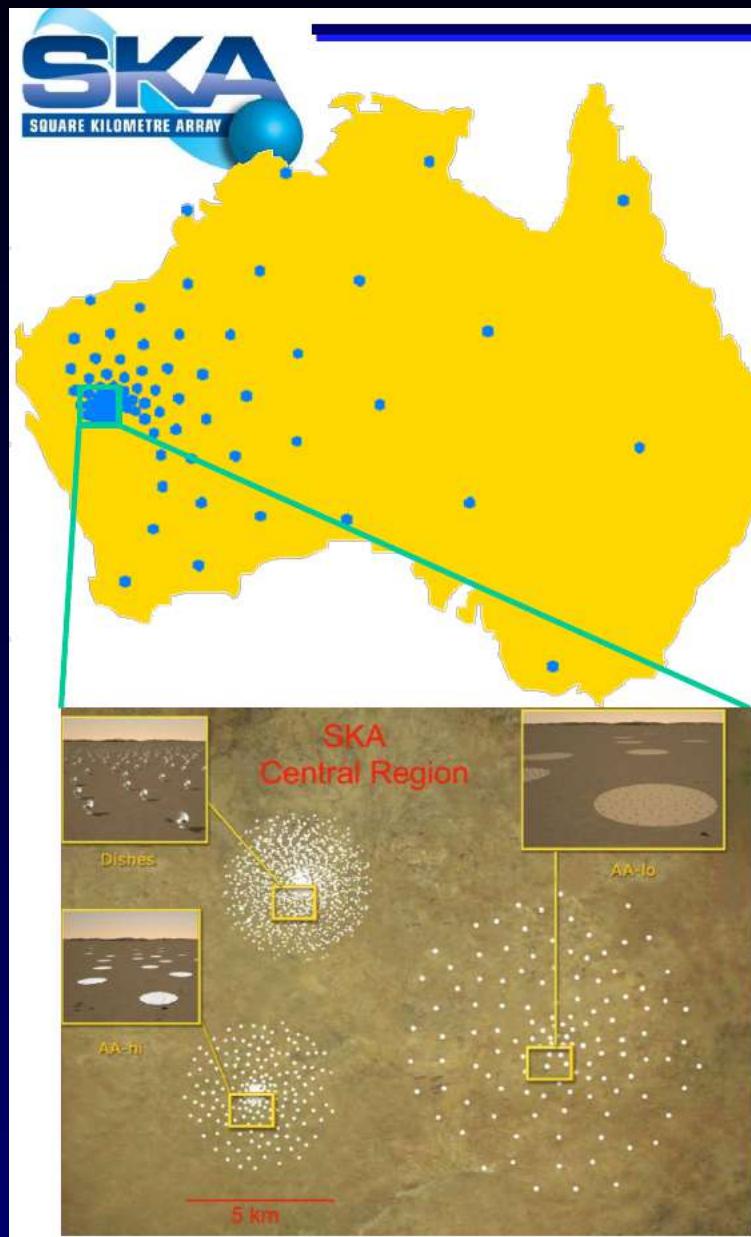


SKA



**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



Dishes

SKA



Phased Aperture array

Square Kilometer Array

SKA

SKA1 MID - the SKA's mid-frequency instrument

The Square Kilometre Array (SKA) will be the world's largest radio telescope, revolutionising our understanding of the Universe. The SKA will be built in two phases - SKA1 and SKA2 - starting in 2018, with SKA1 representing a fraction of the full SKA. SKA1 will include two instruments - SKA1 MID and SKA1 LOW - observing the Universe at different frequencies.

Location: South Africa

Frequency range: 350 MHz to 14 GHz

~200 dishes (including 64 MeerKAT dishes)

Total collecting area: 33,000m² or 126 tennis courts

Maximum distance between dishes: 150km

Total raw data output: 2 terabytes per second
62 exabytes per year

x340,000 average laptops with content every day

Enough to fill 340,000 average laptops with content every day

Compared to the JVLA, the current best similar instrument in the world:

- 4x the resolution
- 5x more sensitive
- 60x the survey speed

www.skatelescope.org [Facebook](#) [Twitter](#) [YouTube](#) [LinkedIn](#) [The Square Kilometre Array](#)

SKA1 LOW - the SKA's low-frequency instrument

The Square Kilometre Array (SKA) will be the world's largest radio telescope, revolutionising our understanding of the Universe. The SKA will be built in two phases - SKA1 and SKA2 - starting in 2018, with SKA1 representing a fraction of the full SKA. SKA1 will include two instruments - SKA1 MID and SKA1 LOW - observing the Universe at different frequencies.

Location: Australia

Frequency range: 50 MHz to 350 MHz

~130,000 antennas spread between 500 stations

Total collecting area: 0.4km²

Maximum distance between stations: 65km

Total raw data output: 157 terabytes per second
4.9 zettabytes per year

Enough to fill up 35,000 DVDs every second

5x the estimated global internet traffic in 2015 (source: Cisco)

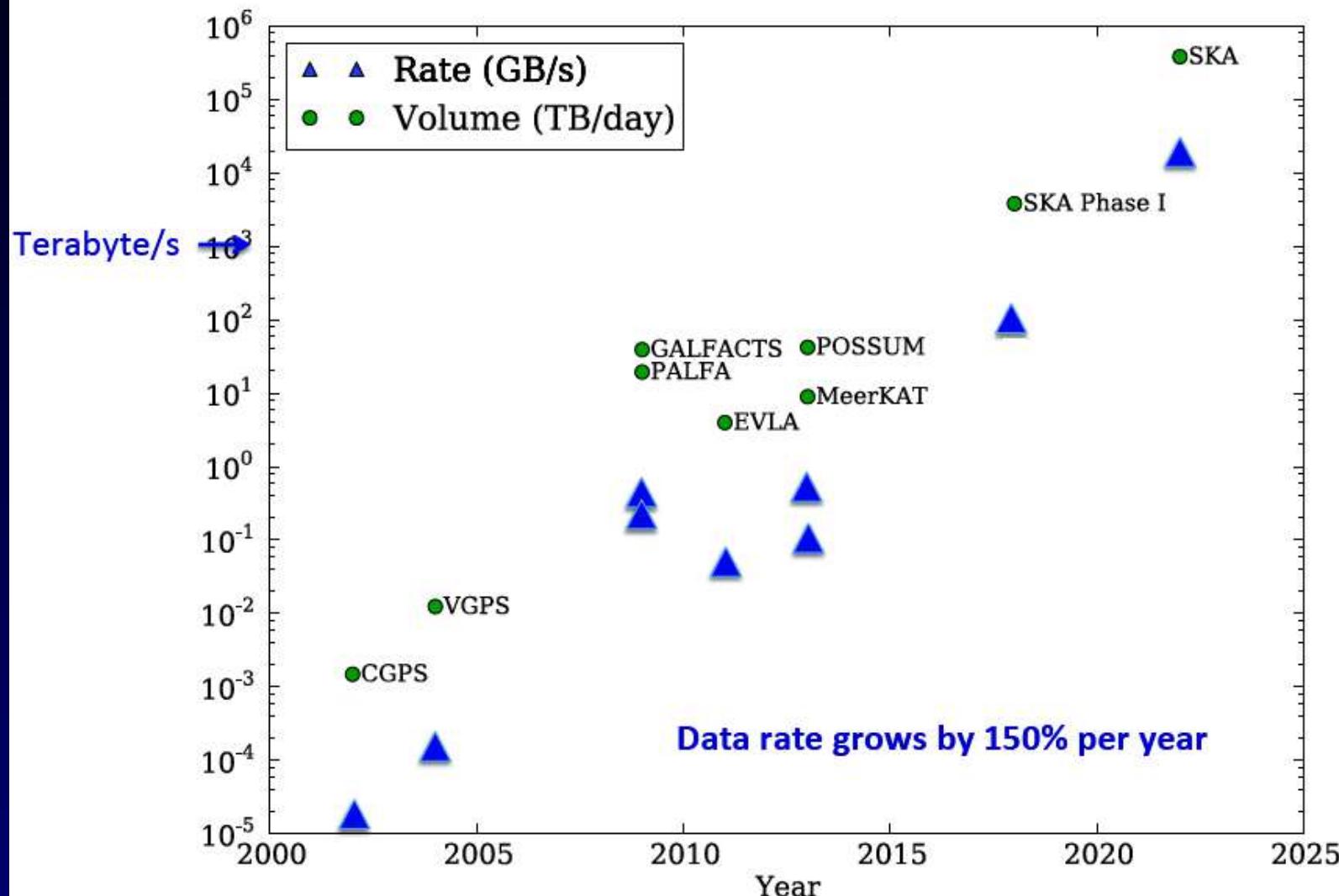
Compared to LOFAR Netherlands, the current best similar instrument in the world:

- 25% better resolution
- 8x more sensitive
- 135x the survey speed

www.skatelescope.org [Facebook](#) [Twitter](#) [YouTube](#) [LinkedIn](#) [The Square Kilometre Array](#)

Cyber SKA

Survey Raw Data Rates out of Correlator



SKA Data Challenge



Antennas



Digital Signal Processing (DSP)



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

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

Over 10's to 1000's kms



High Performance Computing Facility (HPC)

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 Telescope Array

Cherenkov Astronomy and CTA



- ◆ 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



© DESY/Milde Science Comm./Exozet

Millenium Run

10^{10} particles

Several Gpc to

10 kpc

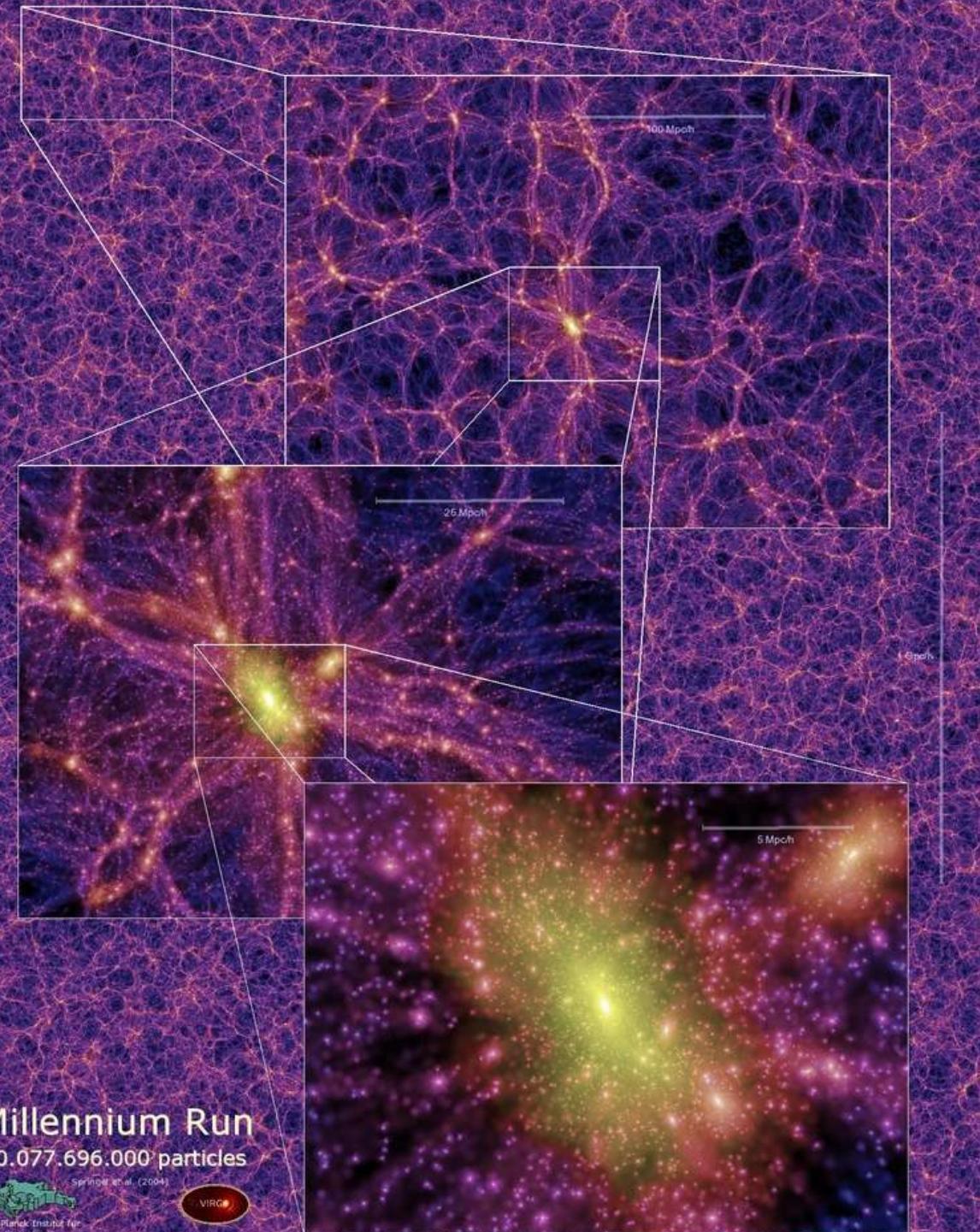
Cube 2 billion ly

One month MPSSC

25 TB

Evolution of 20 mil
galaxies

Evolution merger tree



Simulation of the Universe

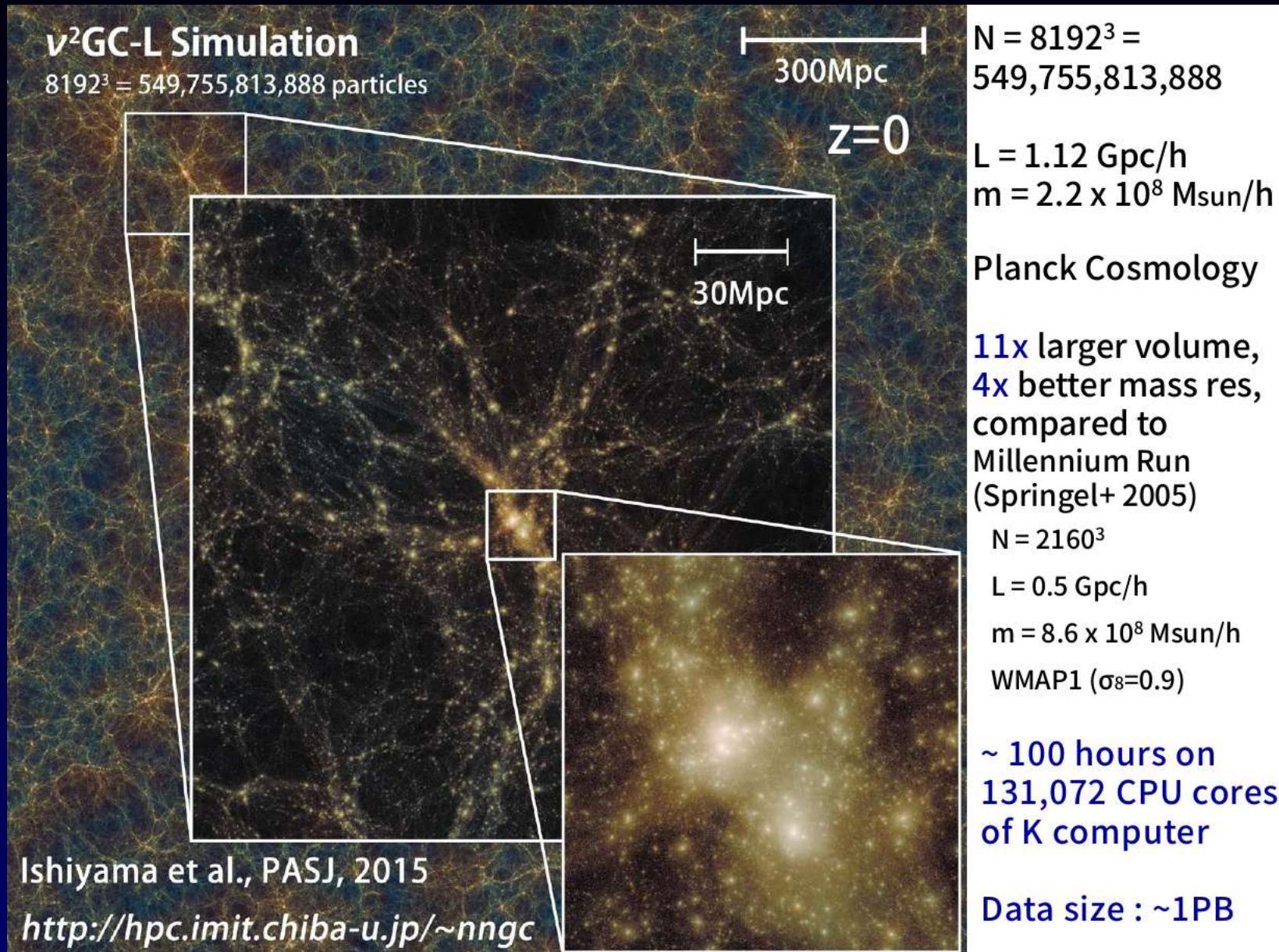
World's fifth fastest supercomputer

K computer

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



Simulation of Universe



Problem of 1PB Data Transfer

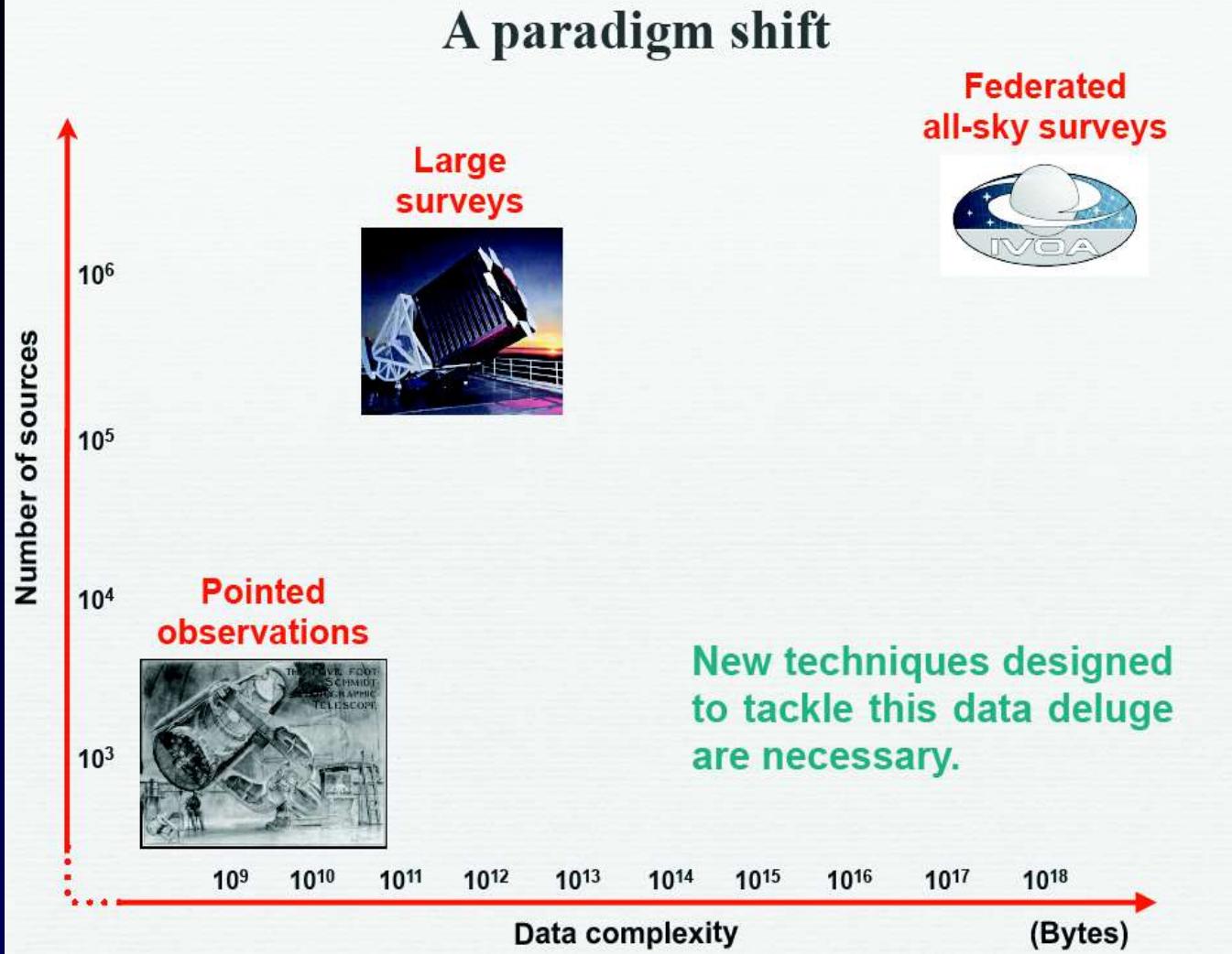
Data transfer

- 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**

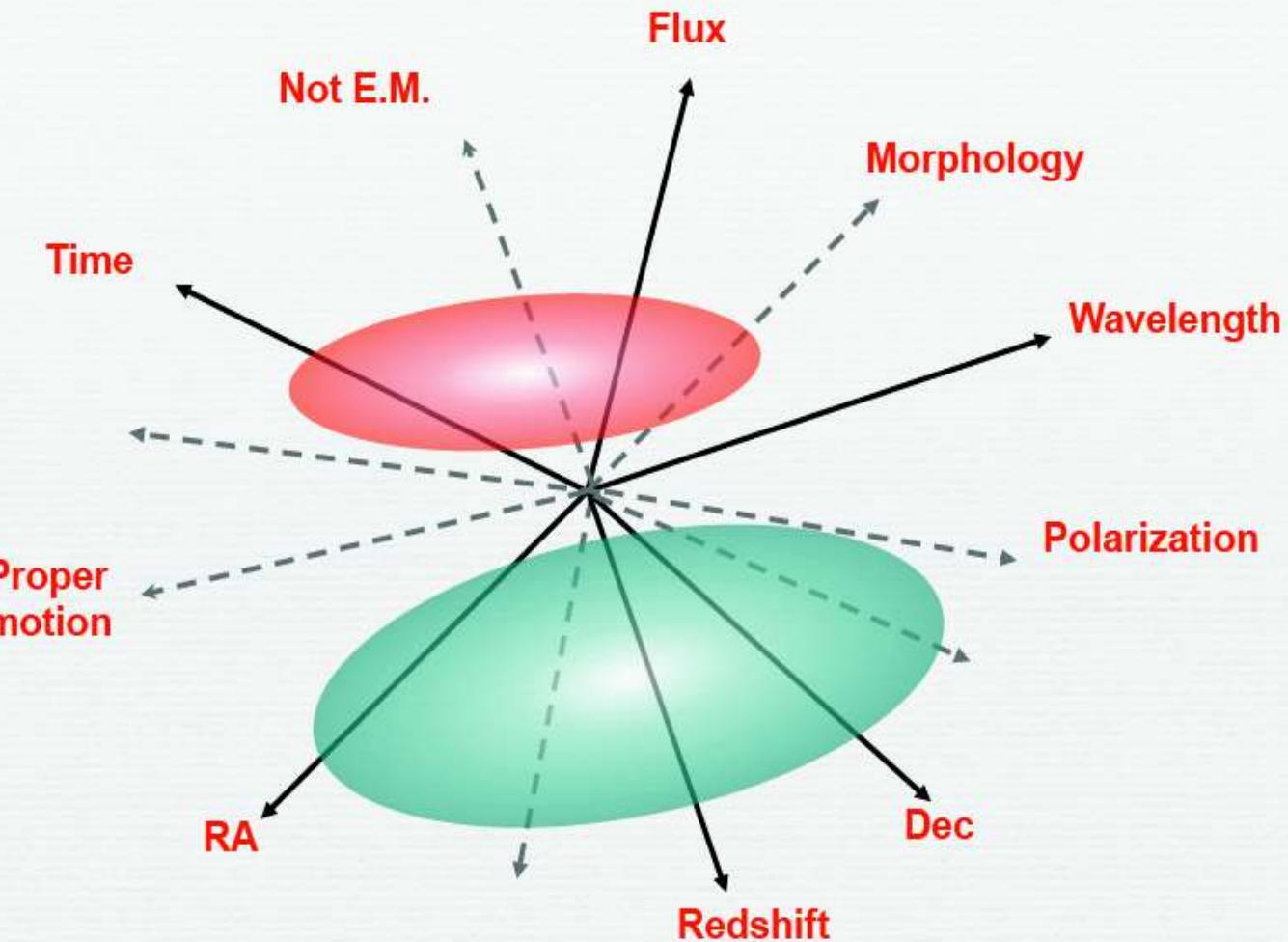


- From Kobe to Chiba
(from Kyoto to Tokyo + 100km , ~600km journey)

A paradigm shift



A growing parameter space



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

Virtual Observatory : Key Definitions

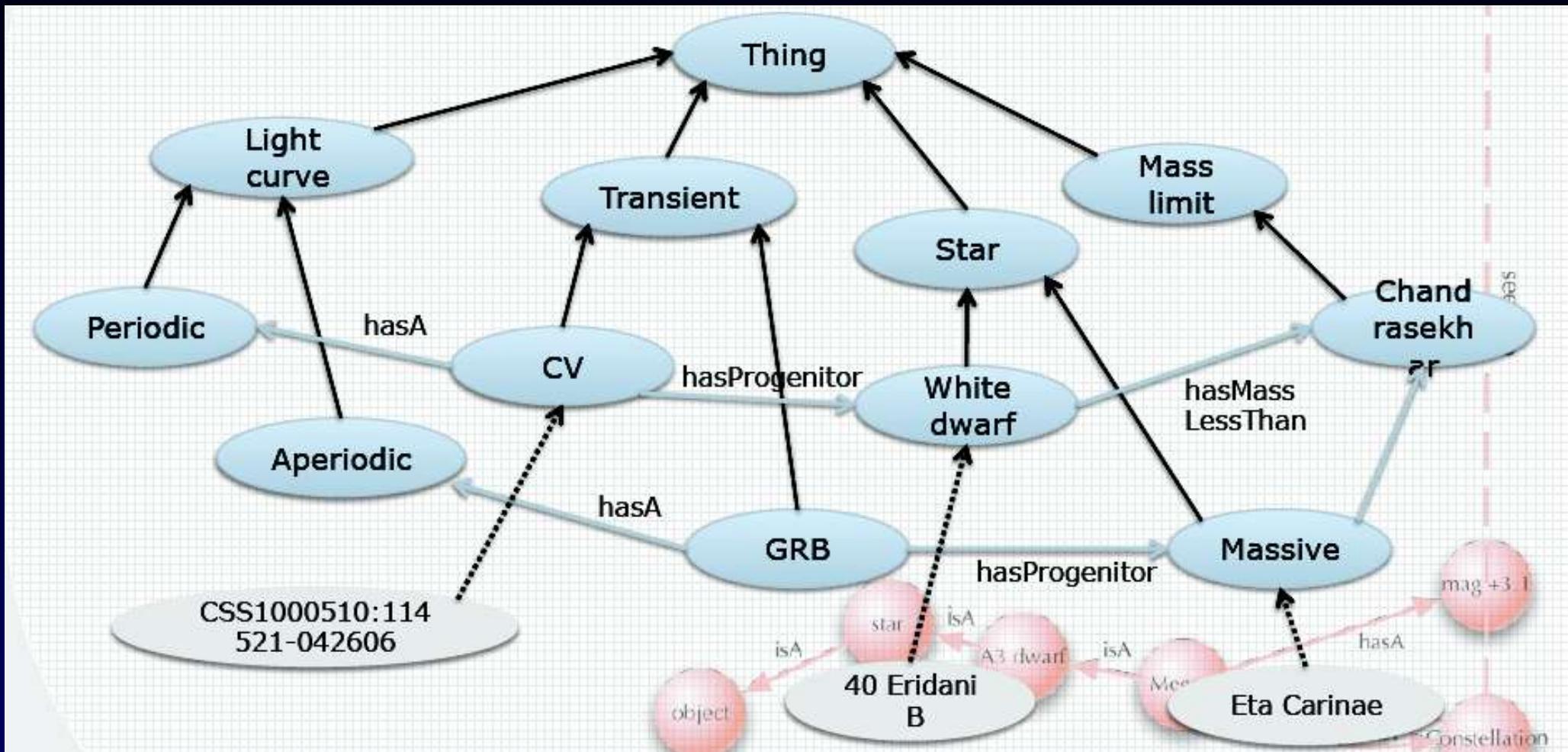
- “*The Virtual Observatory will be a system that allows astronomers to interrogate multiple data centers in a seamless and transparent way, which provides new powerful analysis and visualization tools within that system, and which gives data centers a standard framework for publishing and delivering 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



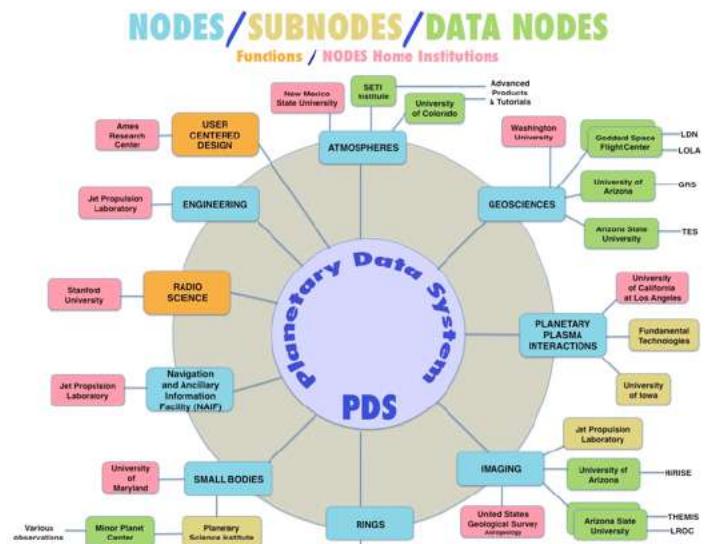
Ontologies in Astronomy



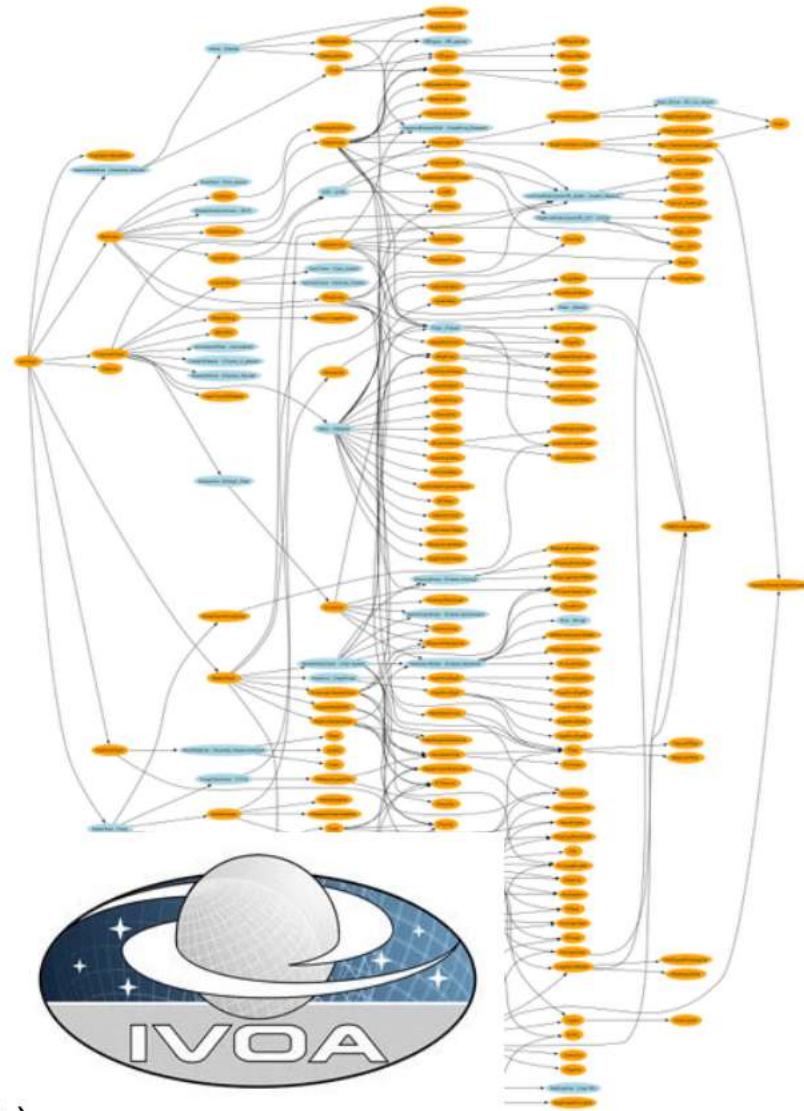
SKOS, RDF standards, search with understanding (not return QSO as binary star)

Ontologies

Ontologies



PDS -> Earth Science (NASA)



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 interoperability – (PLASTIC), SAMP

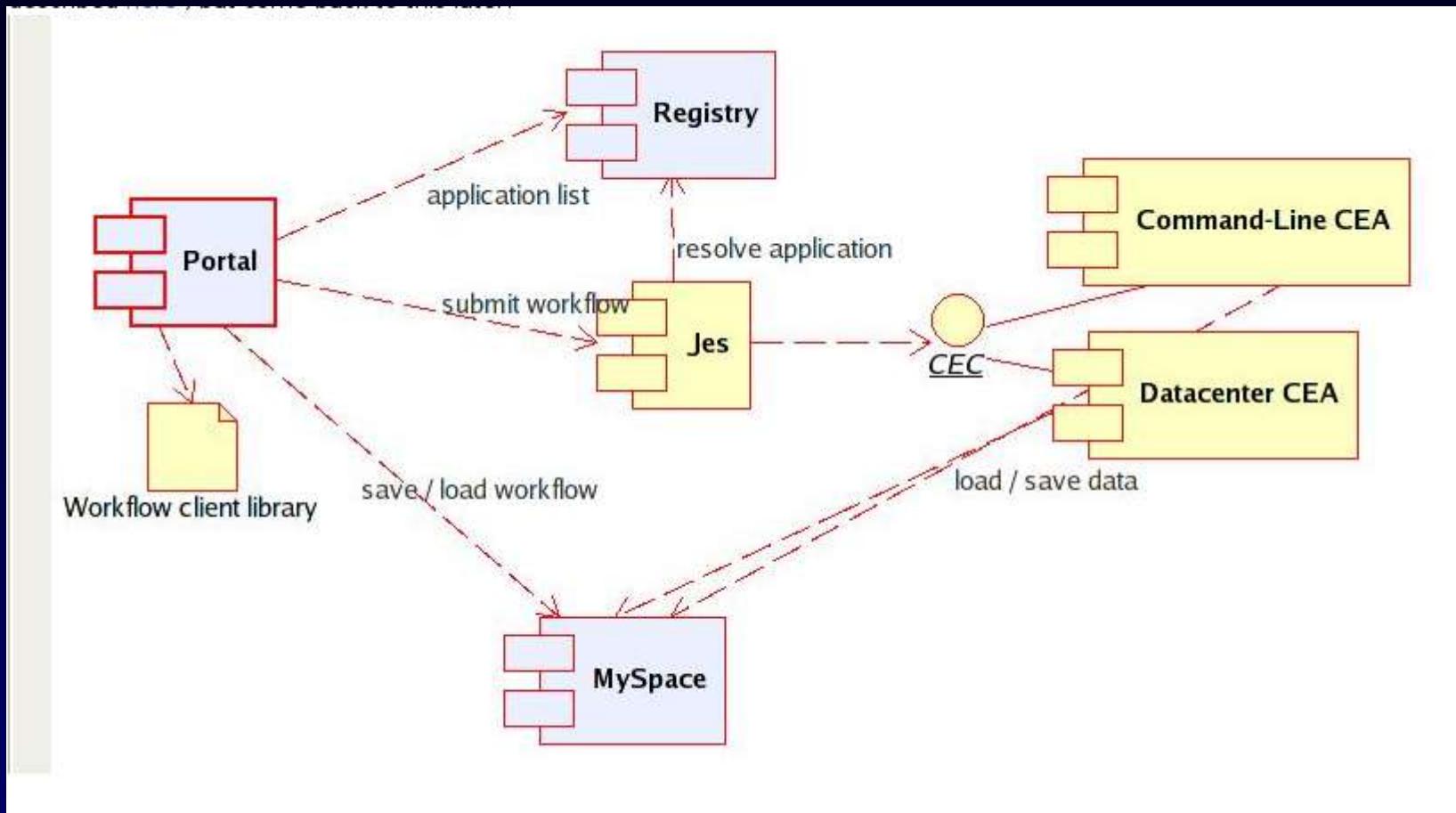
Allows develop applications as bricks

sending VOTABLES (catalogue-spectra-images)

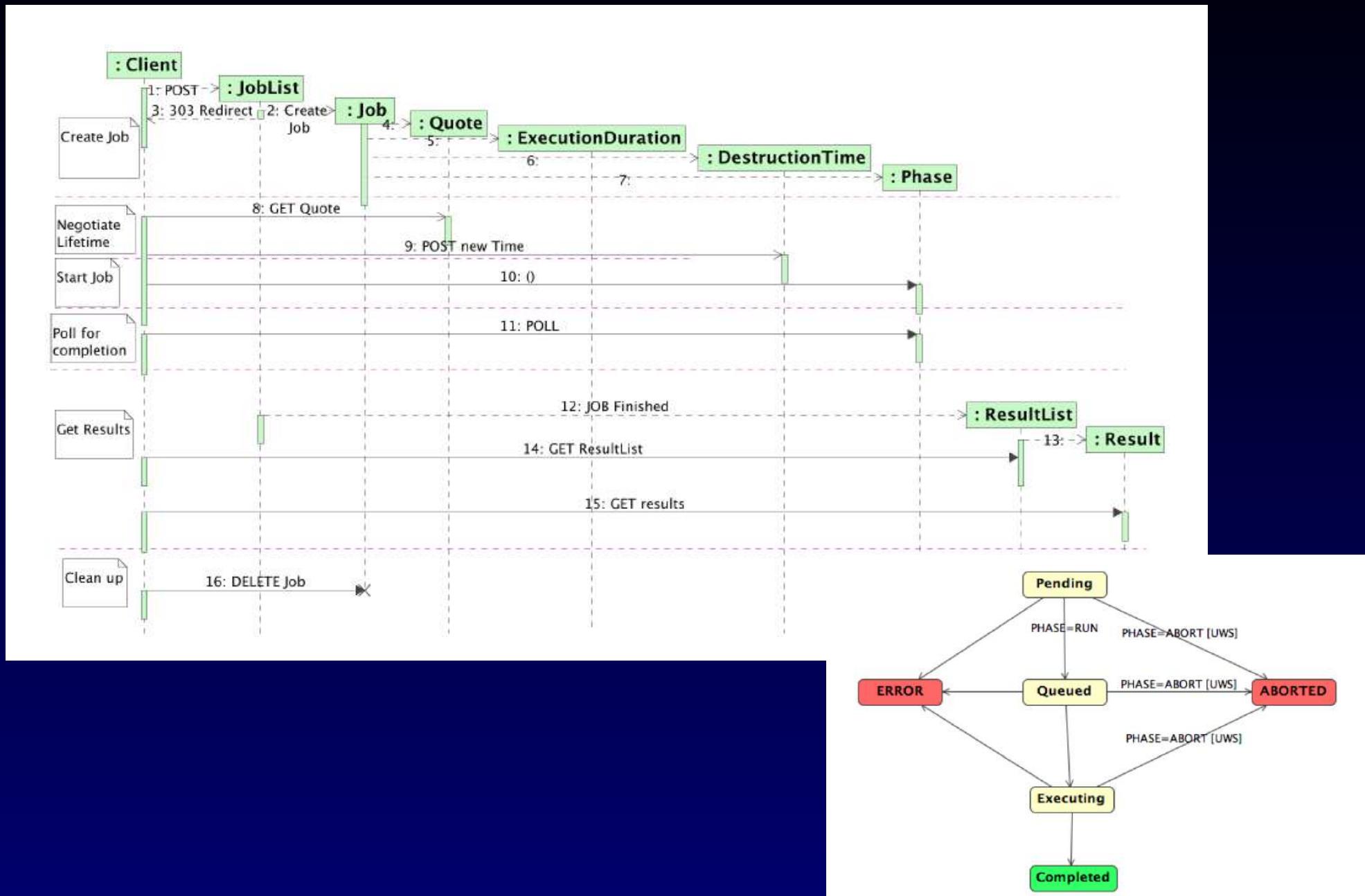
Commercial interest (GoogleSky, MS WWT)

Workflows - Astrogrid

Running remote services – e.g. Sextractor, CASJobs, AstroNeural MLP...



IVOA Universal Worker Service (UWS)



Ecosystem of VO - level 0

LEVEL 0

USERS



COMPUTERS



USER LAYER

USING

F
I
N
D
I
N
G

VO
CORE

G
E
T
T
I
N
G

SHARING

RESOURCE LAYER

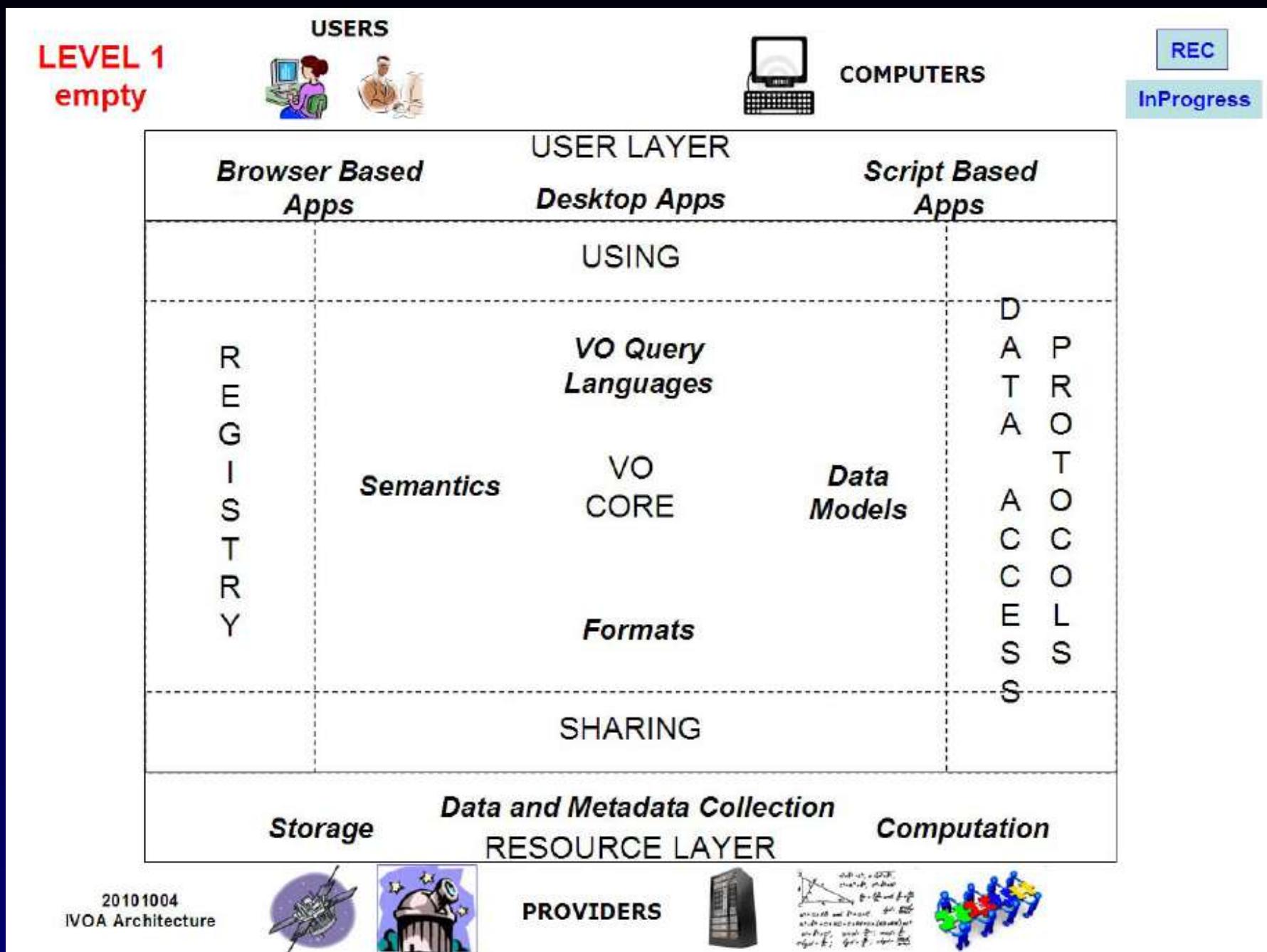
20101004
IVOA Architecture



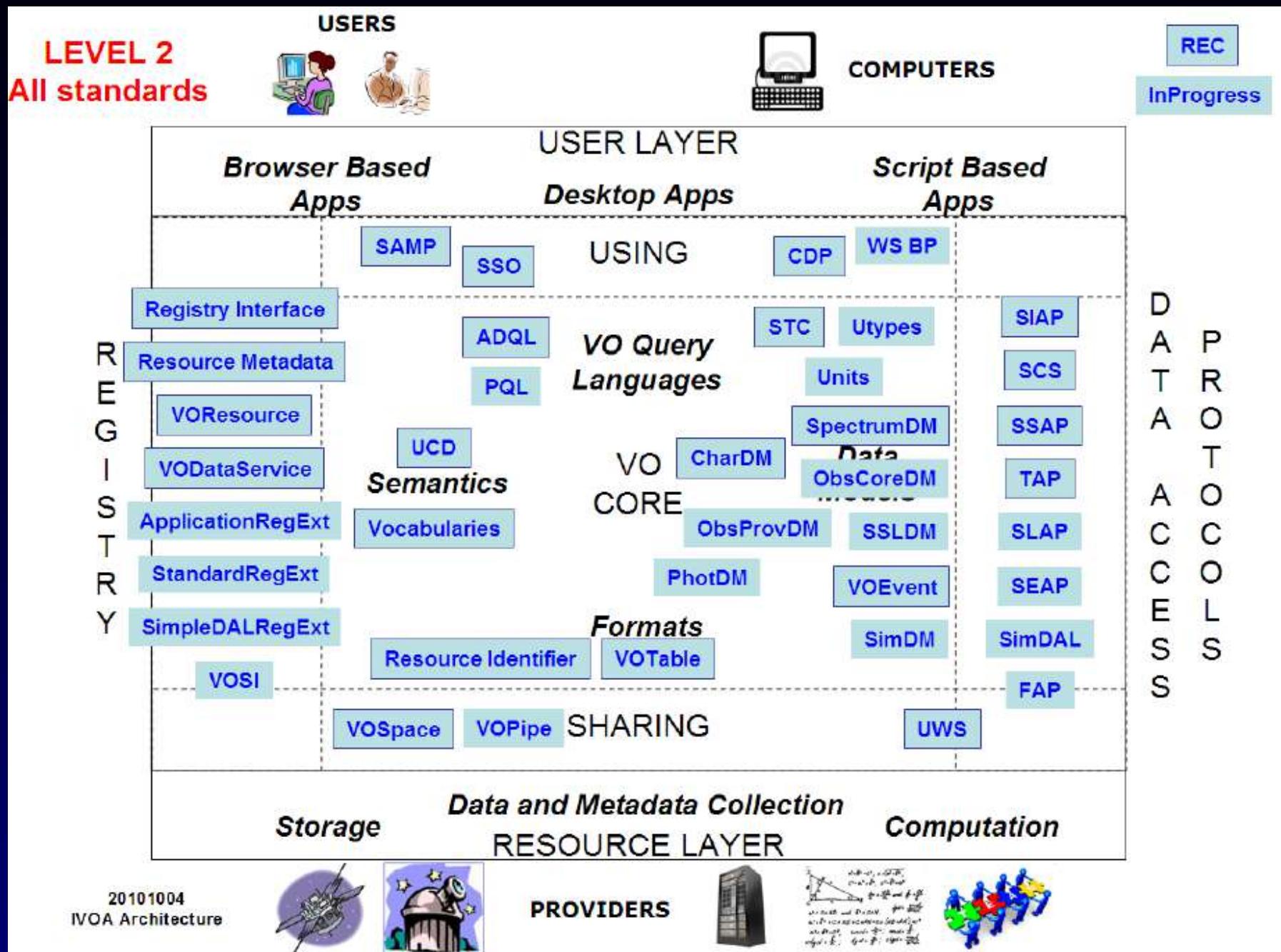
PROVIDERS



Ecosystem of VO - level 1



Ecosystem of VO - level 2



FITS standard

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

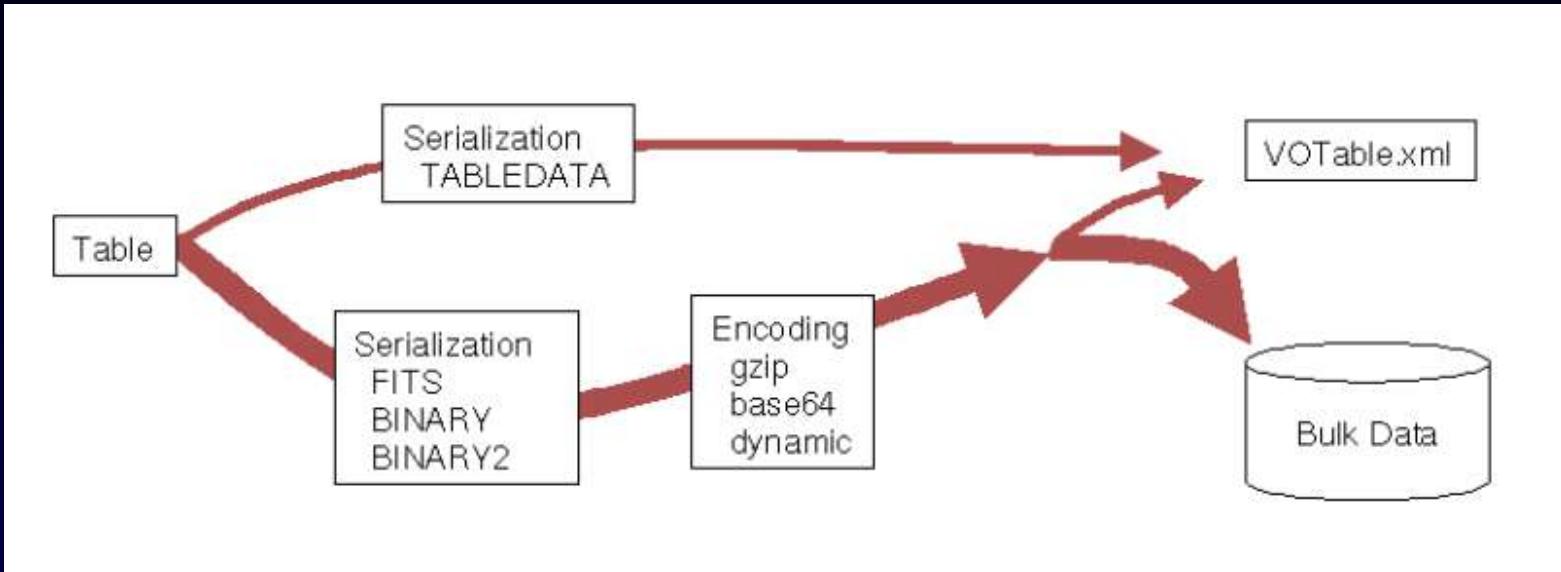
```
SIMPLE = T / file does conform to FITS standard
BITPIX = 16 / number of bits per data pixel
NAXIS = 2 / number of data axes
NAXIS1 = 2048 / length of data axis 1
NAXIS2 = 2048 / length of data axis 2
EXTEND = T / FITS dataset may contain extensions
COMMENT FITS (Flexible Image Transport System) format 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 CASSgrain
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" arraysizes="*" value="VTeI" />
  <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

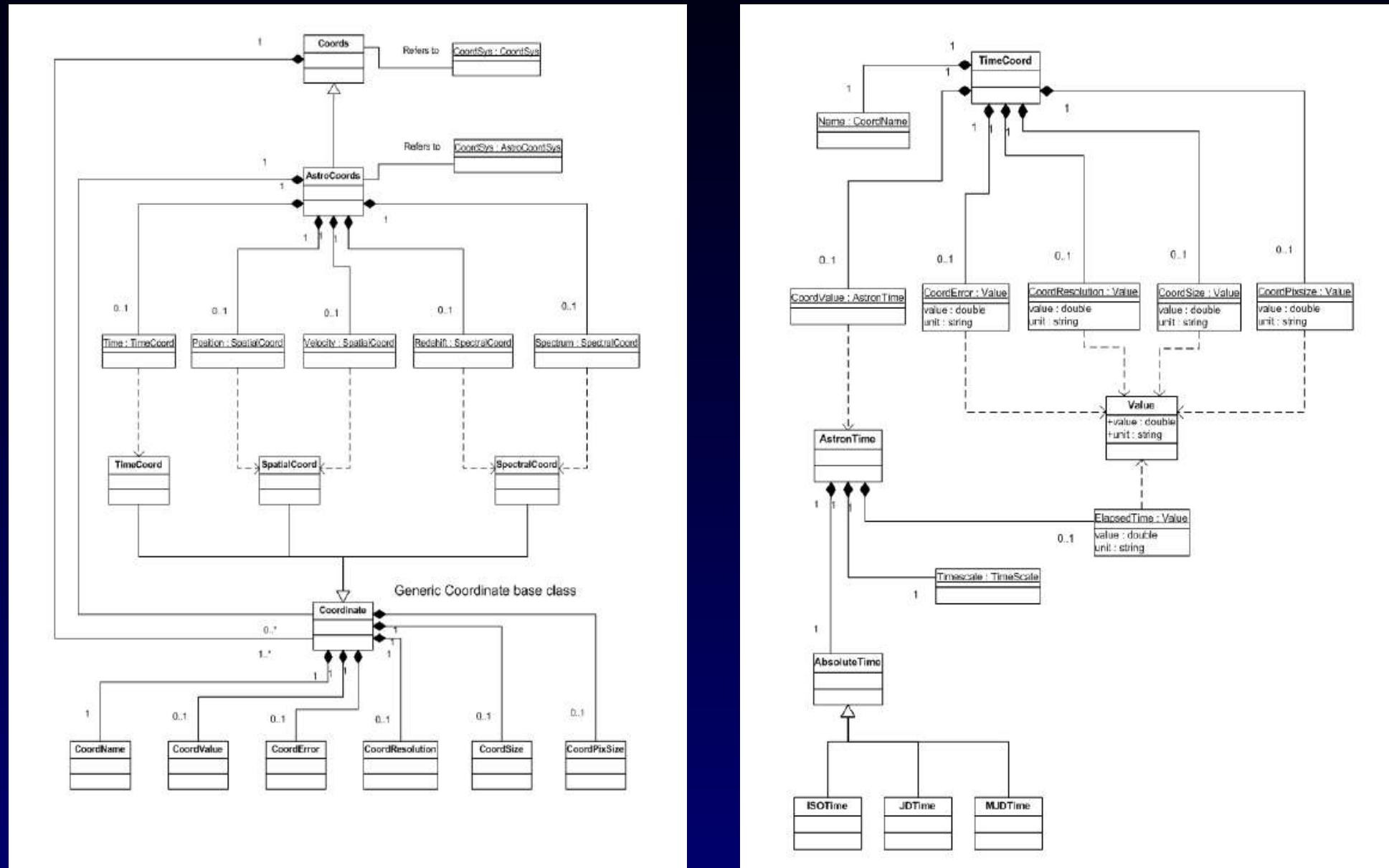
Curation – long time preservation issues (digital libraries)

Provenance (how was processed, links to other products)

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

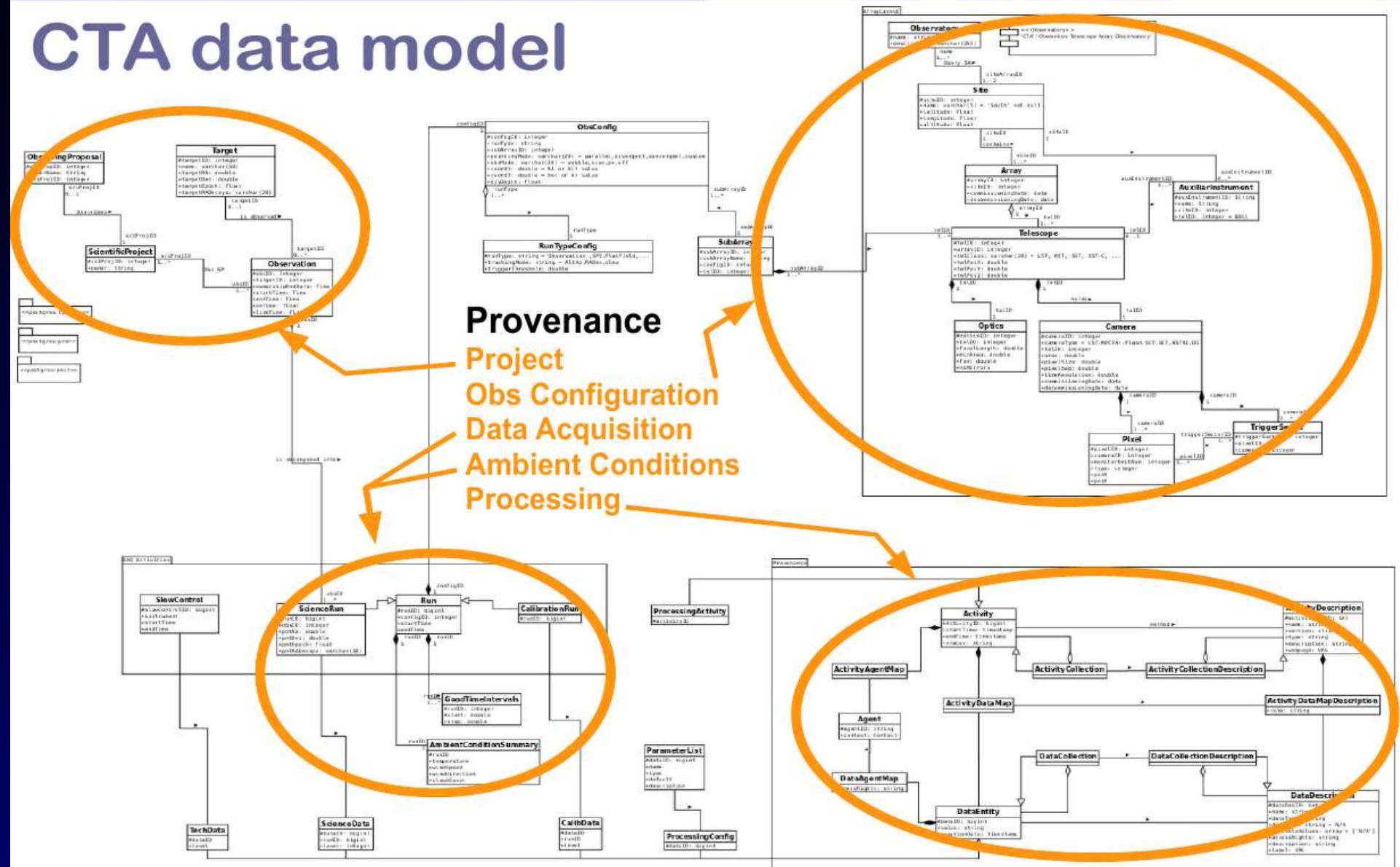
Characterization level 2 (distortion in images, spectra with nonlinear resolution)

Space-Time-Coordinate Data Model



Cherenkov Telescope Array Data Model

CTA data model



Simple Spectra Access Protocol Spectral Data Model

Simple Spectral Access Protocol V1.04



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	$\lambda/d\lambda$	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)

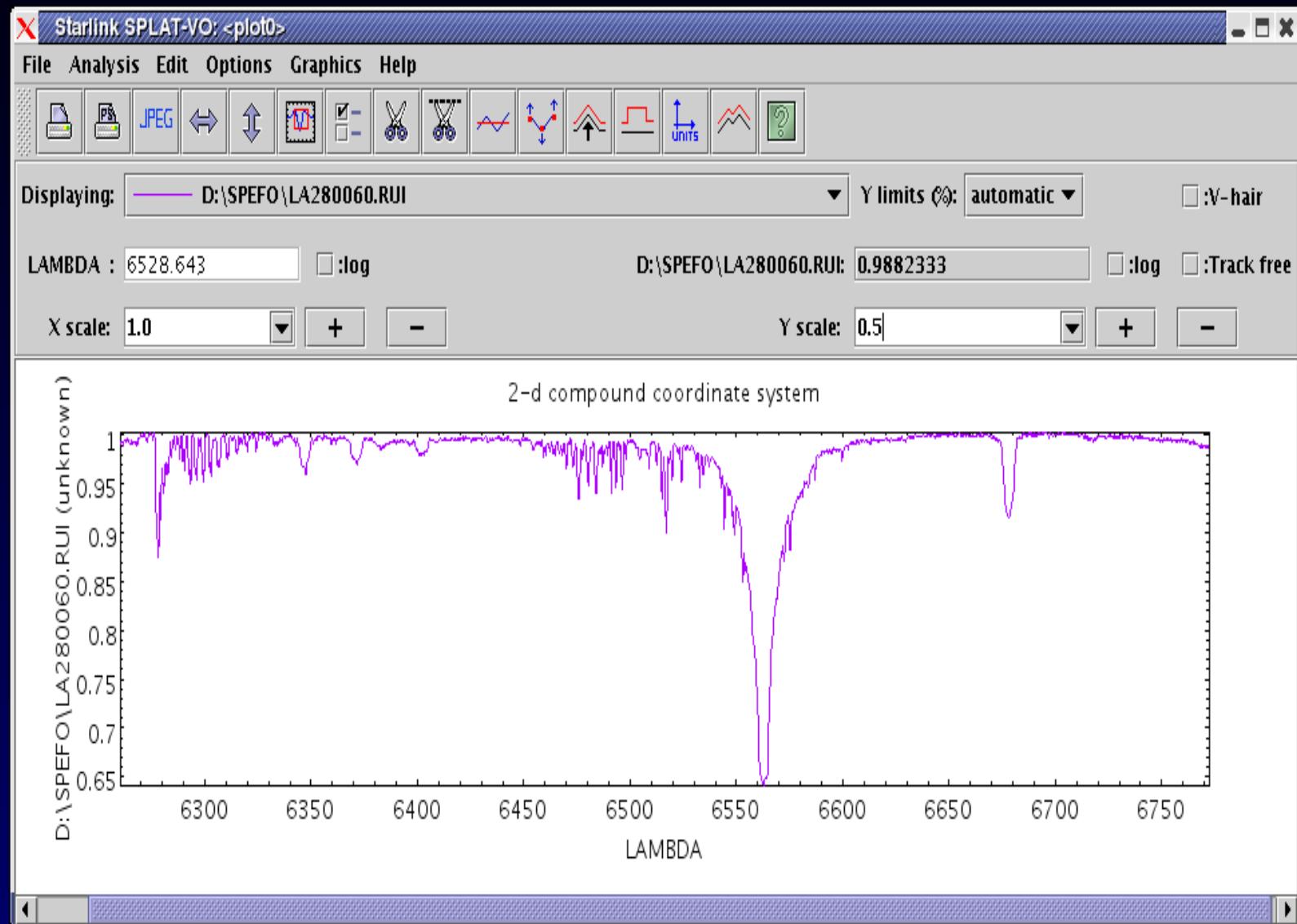
SSO Authentication, authorization, groups and consortia

UWS Universal worker service (job synch, asynch)

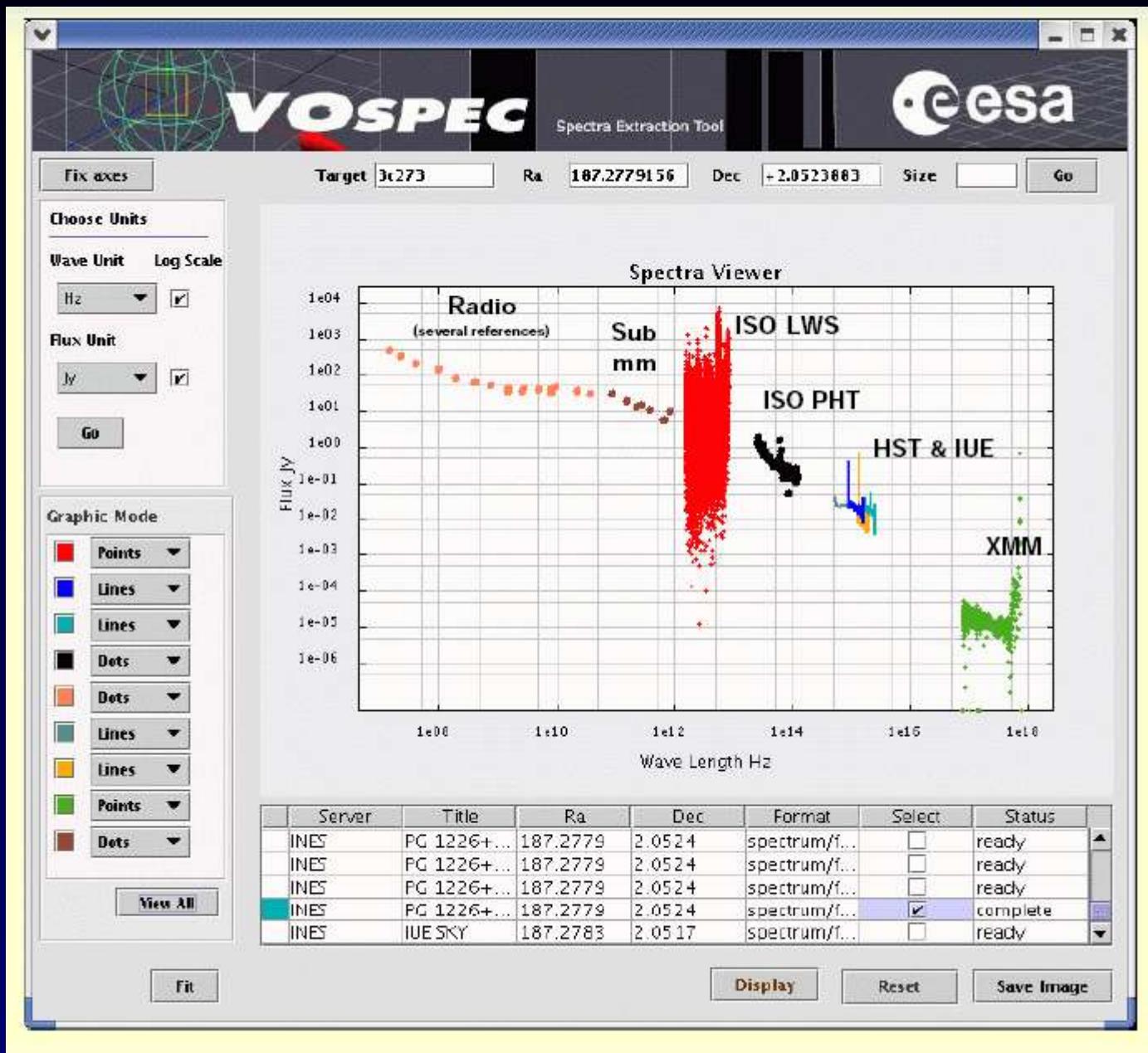
PDL Parameter Description Language

SIM-DB Simulations, theory data

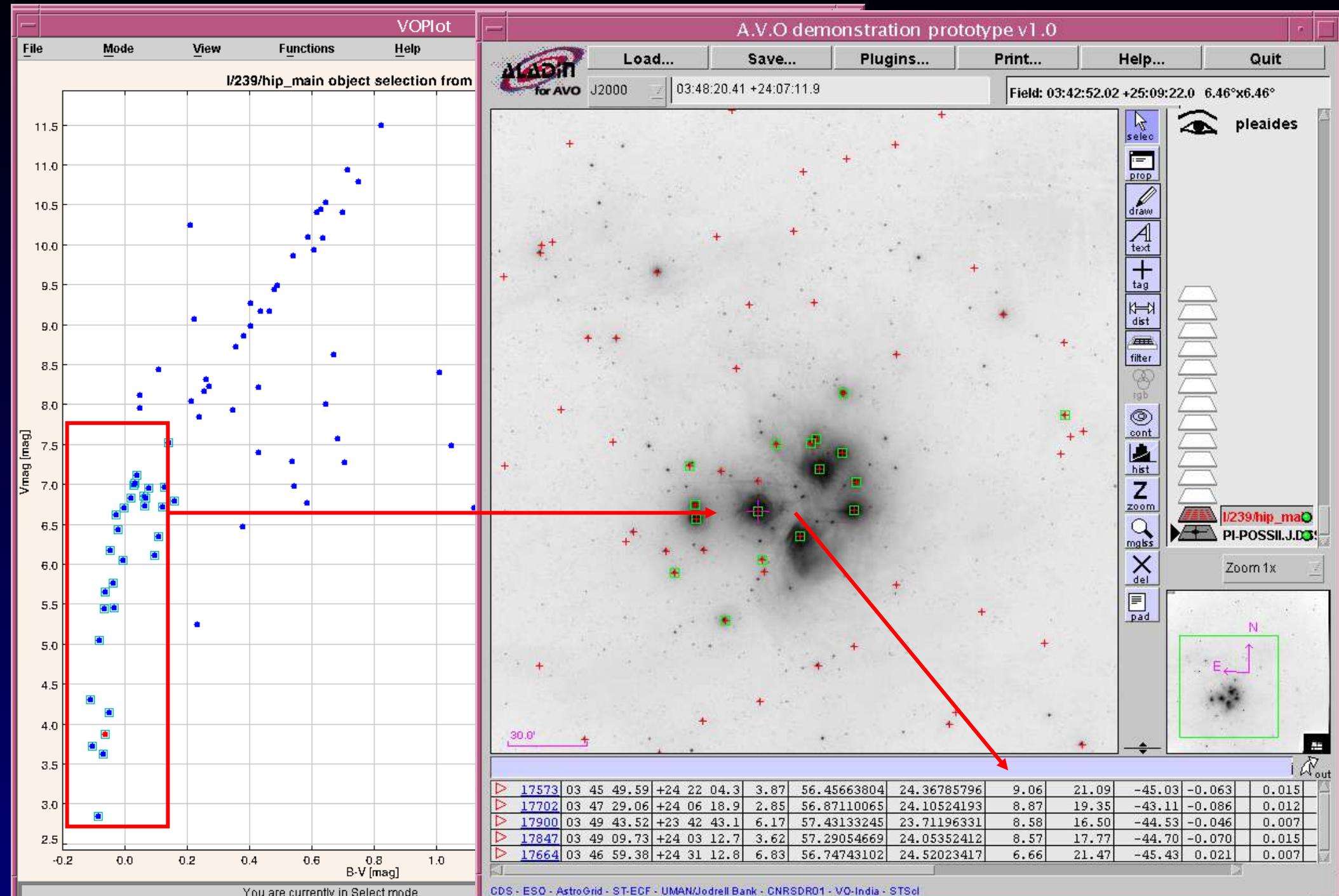
SPLAT-VO (Starlink, JAC)



VOspec (ESAC)



Colour-magnitude diagram



CIELO VO - line catalogue SLAP

SLAP Viewer Copyright ESAC, Spain

Server Selector

- SLAP Services
- IASD
- LERMA
- NIST ATOMIC SPECTRA
- CIELO SLAP
- <http://esav02:8080/cieloslappToolKit/cieloslapp.jsp?>

Molecular line databases

Range of Search (m)

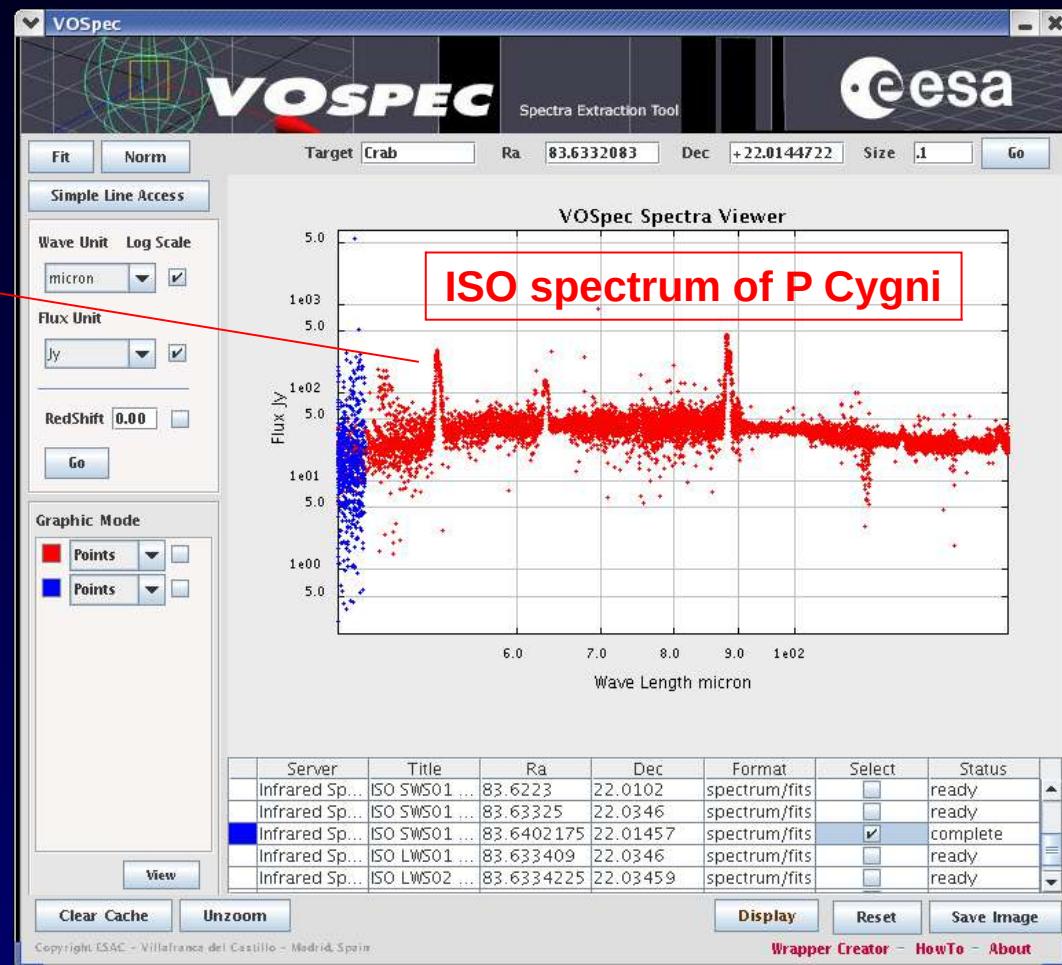
Wavelength Start 4411346184190677E-9 Wavelength End 4411346184190677E-9

Reset

Slap Services Output

CIELO SLAP									
Idm:Line.wavelength	Idm:Source...	Source.co...	Source.co...	...	Idm:L...	Idm:...	Id...	Id...	Idm:...
1.8627e-09	NGC1068	40.66963	-0.01328	...	1s_3p	1s2	1P1	1S0	OVII
1.7768e-09	NGC1068	40.66963	-0.01328	...	1s_4p	1s2	1P1	1S0	OVII
1.89671e-09	NGC1068	40.66963	-0.01328	...	2p	1s	2...	2...	OVIII
2.47793e-09	NGC1068	40.66963	-0.01328	...	2p	1s	2...	2...	NVII
2.21012e-09	NGC1068	40.66963	-0.01328	...	1s_2s	1s2	3S1	1S0	OVII
2.1602e-09	NGC1068	40.66963	-0.01328	...	1s_2p	1s2	1P1	1S0	OVII
2.18071e-09	NGC1068	40.66963	-0.01328	...	1s_2p	1s2	3P1	1S0	OVII
2.16210e-09	NGC1068	40.66963	0.01229	...	1s2	1s2	1F3	1S0	OVII

Close



(IVOA Line Data Model: Dubernet, Osuna et al., in preparation)
(Simple Line Access Protocol: Salgado et al., in preparation)

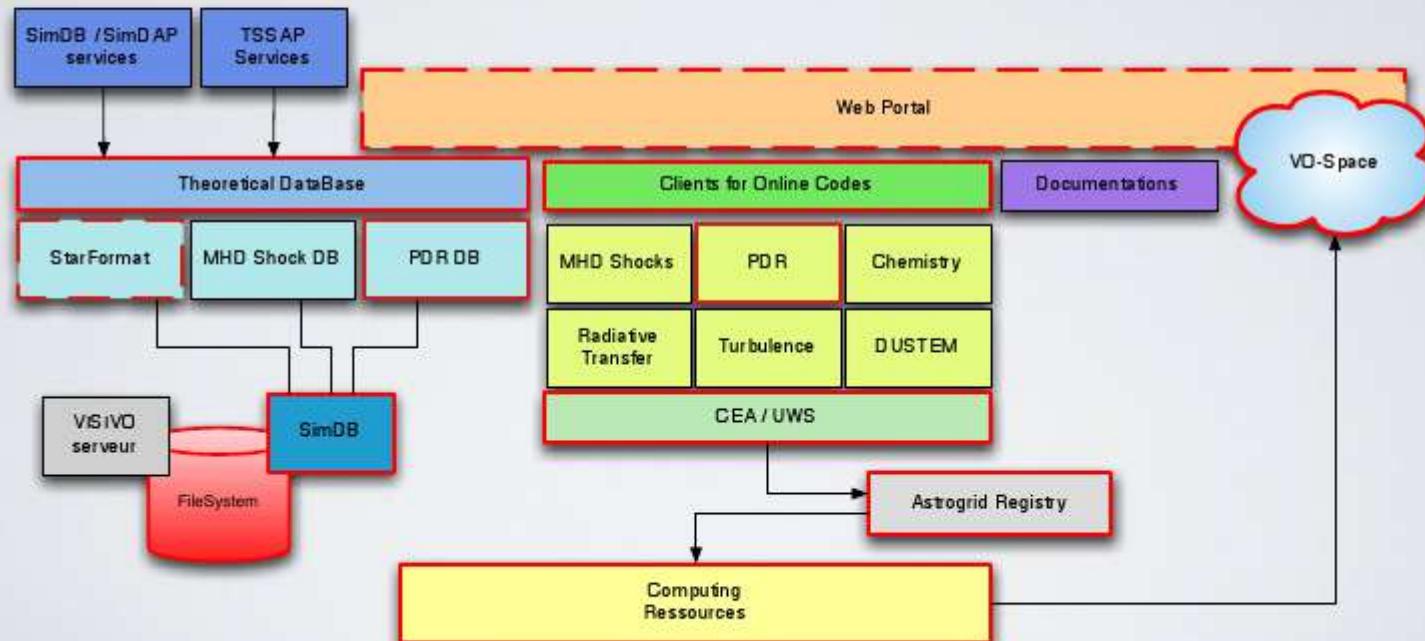
ISM platform

□ Interstellar Medium Platform

Bring together expertise in modeling / simulation of the ISM

Provide theoretical services about ISM

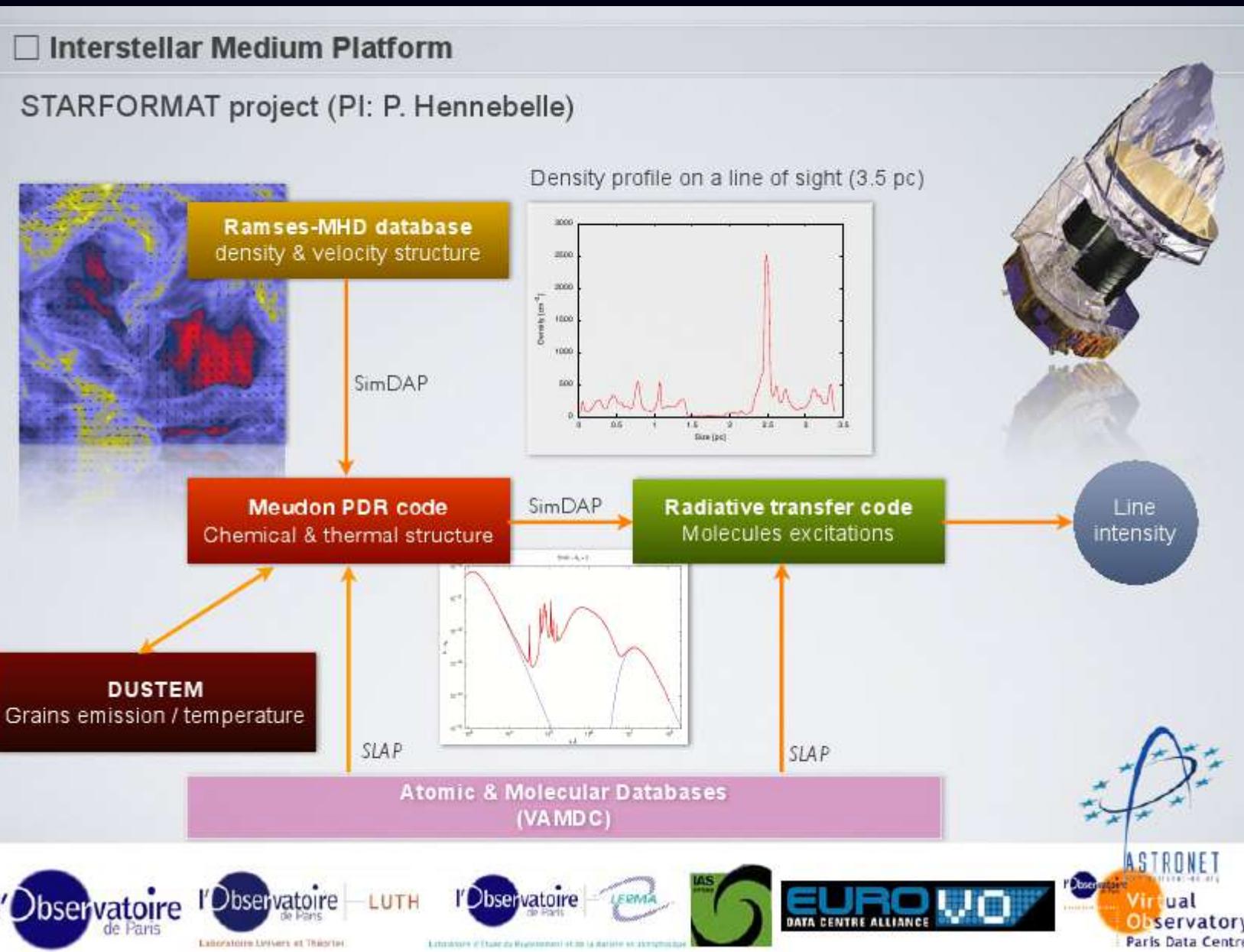
Codes - Databases - Tools & services



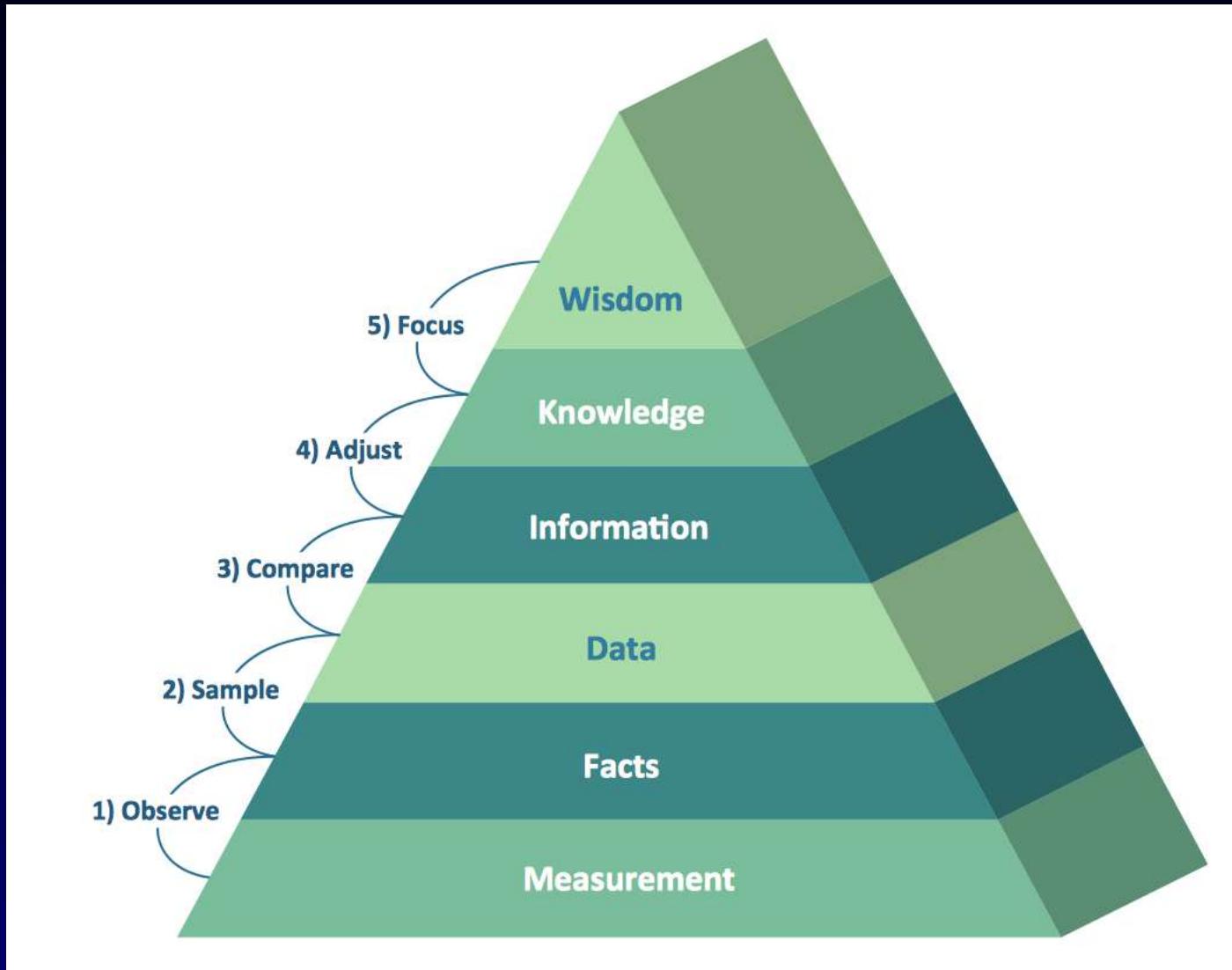
Complex join of TVO bricks

□ Interstellar Medium Platform

STARFORMAT project (PI: P. Hennebelle)



Data-Knowledge-Wisdom Pyramid

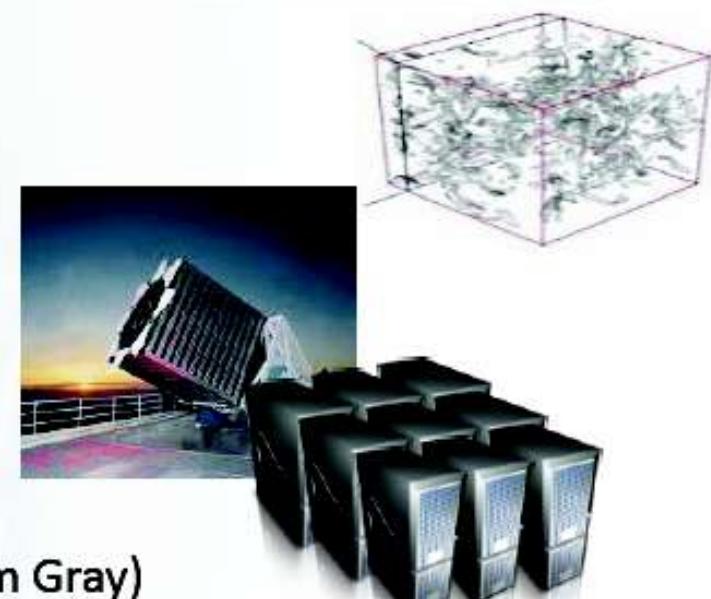


Emergence of a Fourth Research Paradigm

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



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



(With thanks to Jim Gray)

X-informatics



The
**F O U R T H
P A R A D I G M**

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

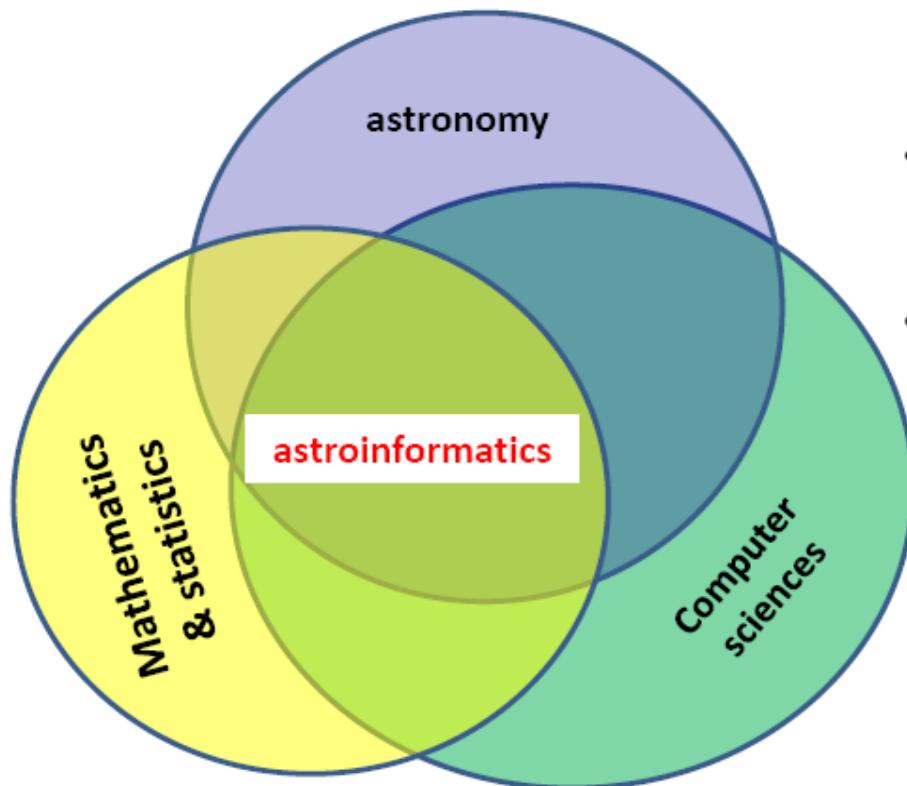
Downloadable at Microsoft Research site

Changing methodology of
the Science

Synergy between different
worlds

Sociological aspects
(net-based research
communities)

Experimental astronomy has become a three players game



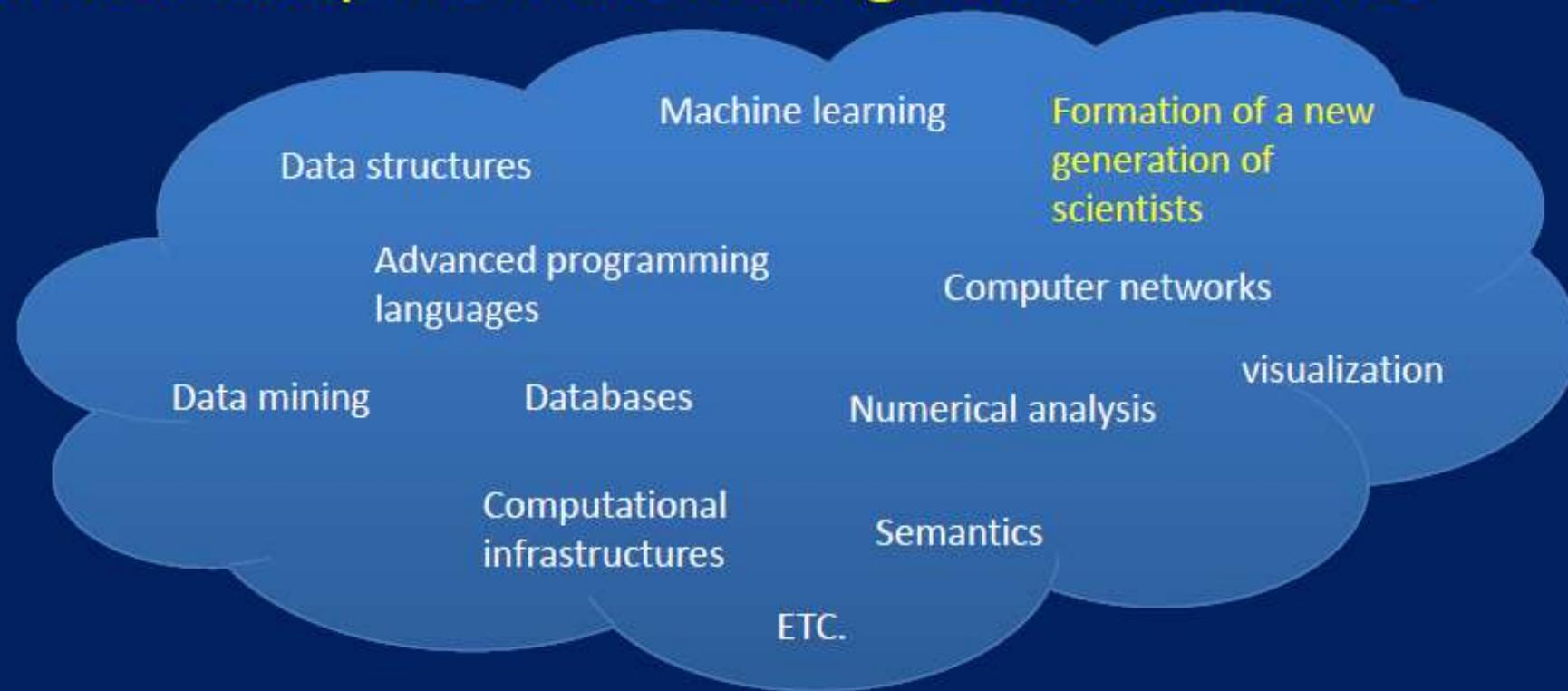
- **astronomy:** problems, data, understanding of the data structure and biases
- **mathematics:** evaluation of the data, falsification/validation of theories/models, etc
- **computer science:** implementation of infrastructures, databases, middleware, scalable tools, etc

- 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

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

A new discipline in the making: AstroInformatics



Very lively Community - AstroInformatics International Conferences

2010 – Pasadena

2011 – Napoli

2012 – Redmond (Microsoft)

2013 – South Africa

Join us on Astroinformatics page on Facebook

IVOA – IG on KDD WIKI

Data Driven Science

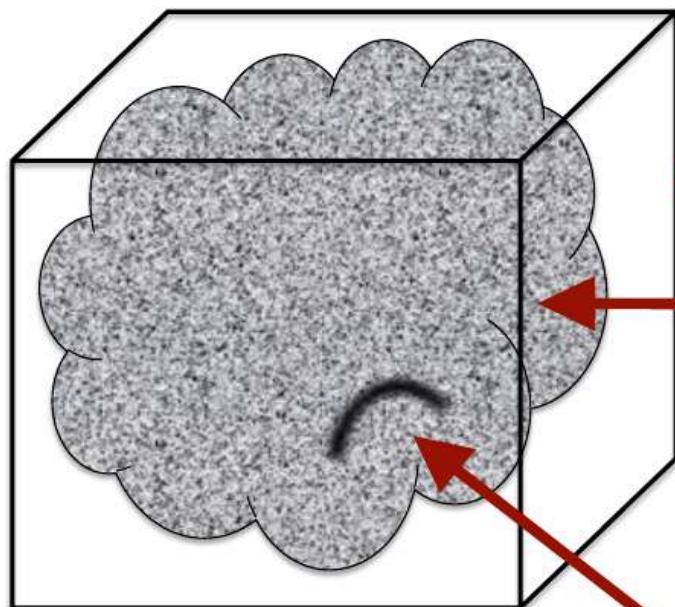
What is Fundamentally New Here?

- 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



$D \gg 3$ parameter space hypercube

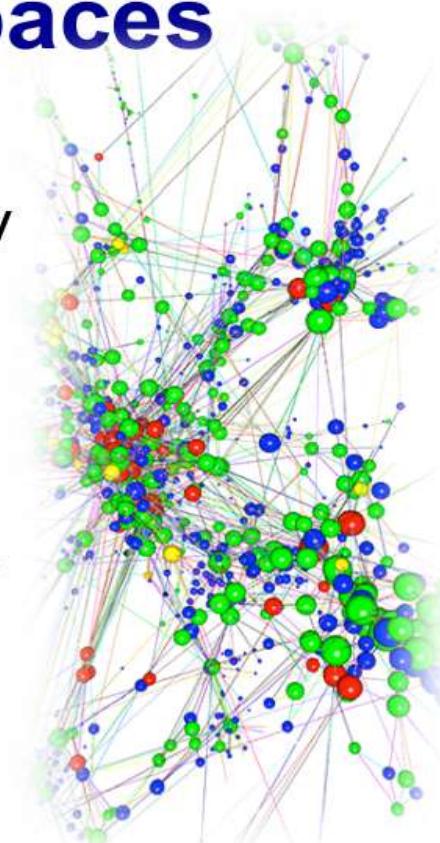
High-D data cloud:
mostly noise, of an arbitrary distribution

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

Visualization in Machine Learning

A Key Challenge: Visualising Multidimensional Data Spaces

- Hyperdimensional structures (clusters, correlations, etc.) may be present in many complex data sets, whose dimensionality may be $D \sim 10^2 - 10^4$, 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 $\sim 3 - 12(?)$ dimensions



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

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

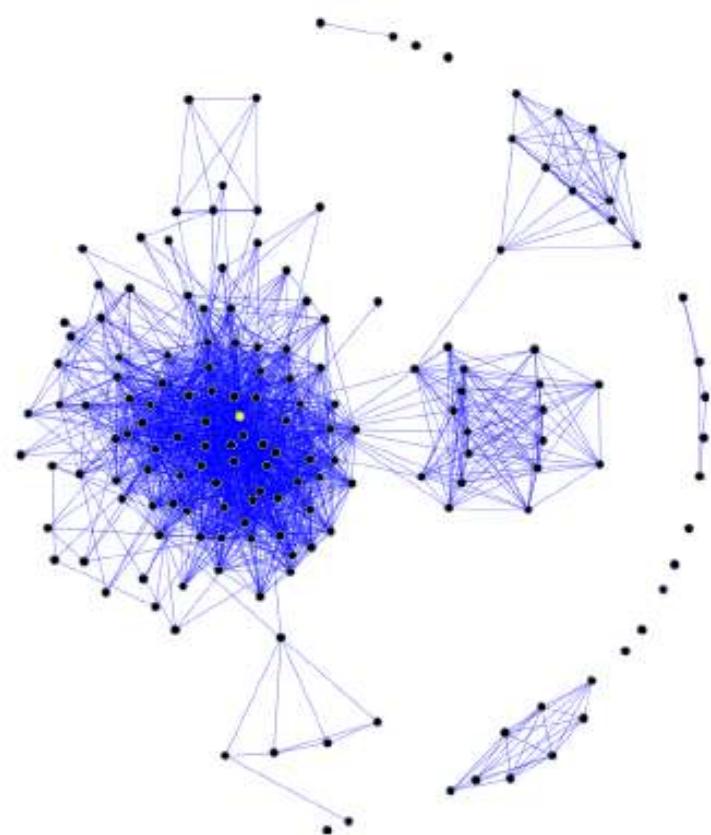
^aBarabàsi et al., “Evolution of the social network of scientific collaborations”

“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

^aSorenson, and Singh, “Science, Social Networks and Spillovers”

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.



Virtual Worlds (2nd Life for Science)



Now migrating to the *OpenSim*-based
VWs, e.g., Intel's *ScienceSim*

Djorgovski



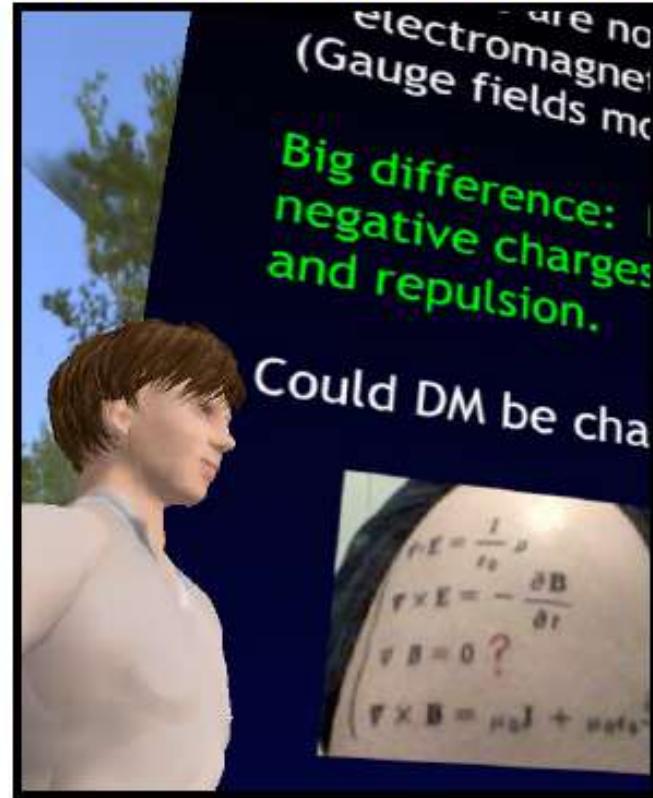
Virtual Conferences

Virtual conferences at zero cost

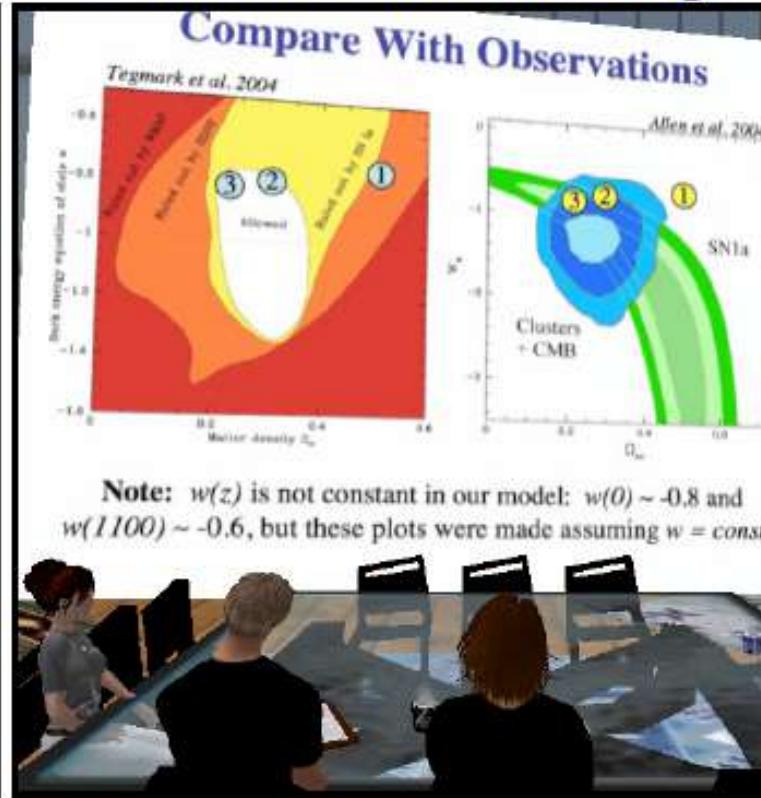
Problem with time zones

Outreach, education

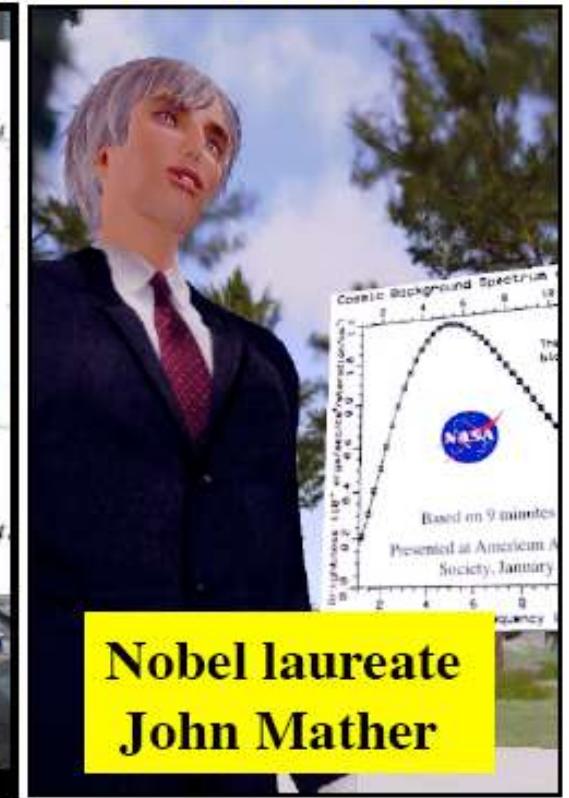
Professional seminars



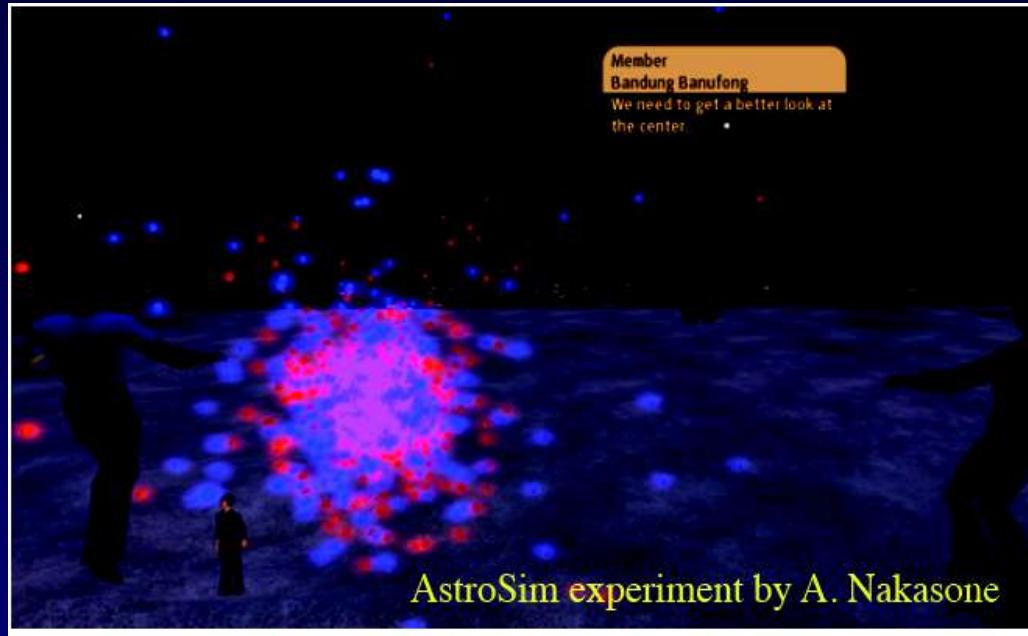
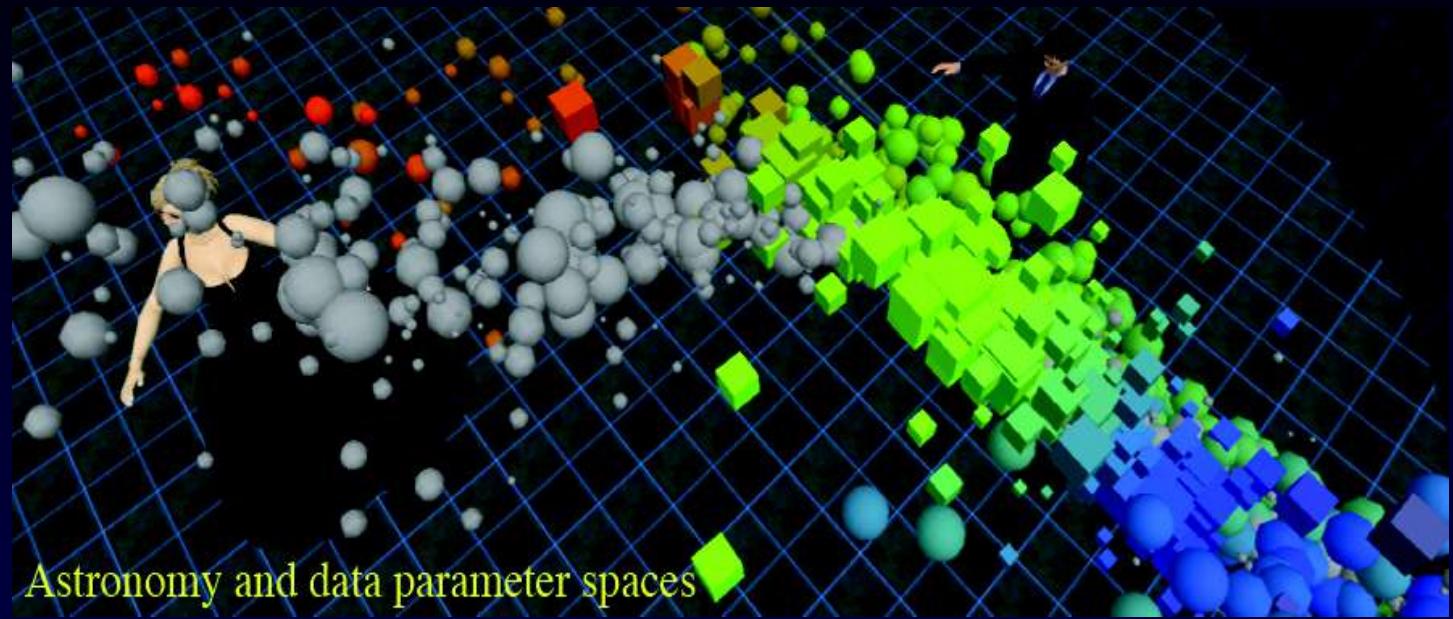
Collaboration meetings



Public outreach



Immersive VR Experiments



Scientists
immersed in,
and
interacting
with,
numerical
simulations
of star
clusters

Citizen Science - Galaxy Zoo

The screenshot shows the Galaxy Zoo website interface. At the top, the title "GALAXY ZOO.org" is displayed with a red and yellow spiral galaxy icon. Below the title is a navigation menu with links: Welcome, Home, The Science, How to Take Part, Galaxy Analysis, Forum, Press & News, FAQ, Links, Contact Us, Login, and Register. On the left, a sidebar titled "Galaxy Analysis" contains a welcome message: "Welcome to Galaxy Zoo's view of the Universe. If you're here you should already have seen the Tutorial, but feel free to go and remind yourself. There's no need to agonise for too long over any one image, just make your best guess in each case." In the center, there is a large image of a spiral galaxy with a prominent central bulge and two smaller galaxies visible in the background. Below this image is a checkbox labeled "Show Grid Overlay on the next Image". To the right of the main image, the text "Galaxy Ref: 588010880371851294" is shown, followed by the instruction "Choose the Galaxy Profile by clicking the buttons below". Three buttons are displayed: "CLOCK SPIRAL GALAXY" (yellow spiral), "ANTI SPIRAL GALAXY" (blue spiral), and "EDGE ON/UNCLEAR SPIRAL GALAXY" (yellow and blue spiral). Below these are two more buttons: "ELLIPTICAL GALAXY" (yellow oval) and "STAR / DON'T KNOW" (yellow star).

> 20 Science papers published so far

Examples ZOOniverse

The screenshot shows the Moon Zoo website (<http://moon.zooniverse.org/craters>). The main interface features a large crater image with a red crosshair and a small inset image. On the left, a sidebar lists 'Your Moon Tools' including 'Crater Survey', 'Boulder Wars', 'My Moon Zoo', 'Layout', 'Home', 'How to take part', and 'About'. Below the tools is a note from NASA/Goddard Space Flight Center. At the bottom, there's a 'Galaxy Zoo Irregular Checking' section with a 'Galaxy Ref' number (5877356620850) and a 'Submit' button.

Moon Zoo

YOUR CHANCE TO EXPLORE

Your Moon Tools

- Crater Survey
- Boulder Wars
- My Moon Zoo
- Layout
- Home
- How to take part
- About

These credit: NASA/Goddard Space Flight Center

Galaxy Zoo Irregular Checking

Galaxy Ref 5877356620850

If it is the same as the last image click on Not irregular. If it is clearly a regular spot don't know. Otherwise select properties Irregular button

Clarity Clearly-defined, Faint

Shape Compact, Sprawling

Star Rating None, 1-3, 4-10, 11-20, 20+, Not clear

Friends On its own, involved in a merger, another irregular nearby, another galaxy nearby

Bar Yes, Possibly, None

Arms Yes, Possibly, None

Core Yes, Possibly, None

Any Spiral Yes, Possibly, None

Structure Irregular

Zoom In Lots | Zoom In | Zoom Out | Zoom Out Lots | Invert

Back to Previous | Advanced Display | English | Deutsch | Español | Polski | Français

The screenshot shows the Galaxy Zoo MERGERS website (<http://galaxy.zooniverse.org/mergers>). The main interface features a dark background with several white galaxy images. A central image has a red box around it. On the left, there are tabs for 'Explore', 'Simulate', and 'Evaluate'. On the right, there's a 'Selected Sims' panel and a 'More' button. The top right corner shows the Galaxy Zoo logo.

GALAXY ZOO

MERGERS

Profile Logout

SIMULATIONS Viewed: 8 Selected: 1 Enhanced: 0 Evaluated: 0 Level: 0 Inimate

Explore Simulate Evaluate More Selected Sims

Explore

Click on "More" to see 8 randomly generated simulations. Click on ones that you think show similarities to the image in the center. As you do this, the ones that you selected are saved on the right-hand side for later review. [help](#)

Enhance

Maybe you found a simulation on the Explore tab that is similar to the image in the center, but you think

The screenshot shows the Solar Stormwatch website (http://solarmain.zooniverse.org/spot_and_track/spot). The main interface features two video frames labeled 'STEREO BEHIND' and 'STEREO AHEAD'. Between them is a 'FLAT' button with a play/pause icon. To the right is a 'SCAN' button with a right-pointing arrow. The top right corner shows the Solar Stormwatch logo.

SOLAR STORMWATCH

HOME MISSION BRIEFING SPOT & TRACK STORMS TALK ABOUT IT

SPOT

QUESTION Can you spot a solar storm?

INSTRUCTIONS Watch this pair of video clips from the STEREO Behind and STEREO Ahead spacecraft cameras. Do you see a solar storm? Is it in just one camera, or both?

STEREO BEHIND STEREO AHEAD FLAT

Pause Scan

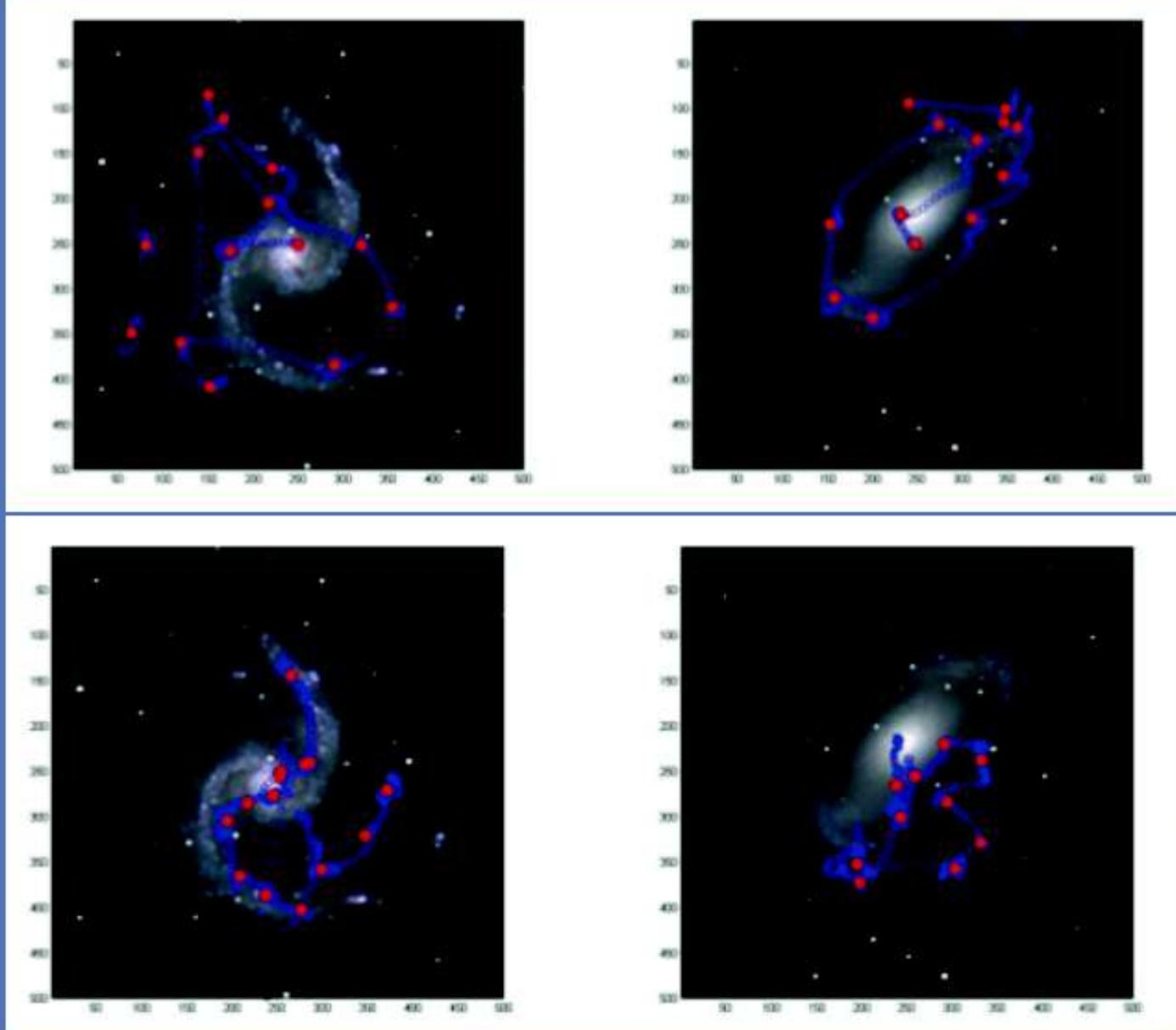
Hints & tips

- Here's a reminder of what a solar storm looks like. [See more examples on Hints](#) (opens in a new window)
- Use the SCAN button to play the video at double speed.
- If you spot a solar storm here are two more steps to complete. If not, that's fine!
- There's always a lot going on at the outside edge of the video. But solar storms expand and then fade as they cross the frame. You will need to follow them at least half way across.
- There might be more than one solar storm per video, but at this stage that doesn't matter. Pick one now, and you'll be able to come back to the rest later if you want.
- Watch a [How to... screencast](#).

ADD CLIP TO FAVOURITES Time - 18:00



Expert vs Non-expert Classifier

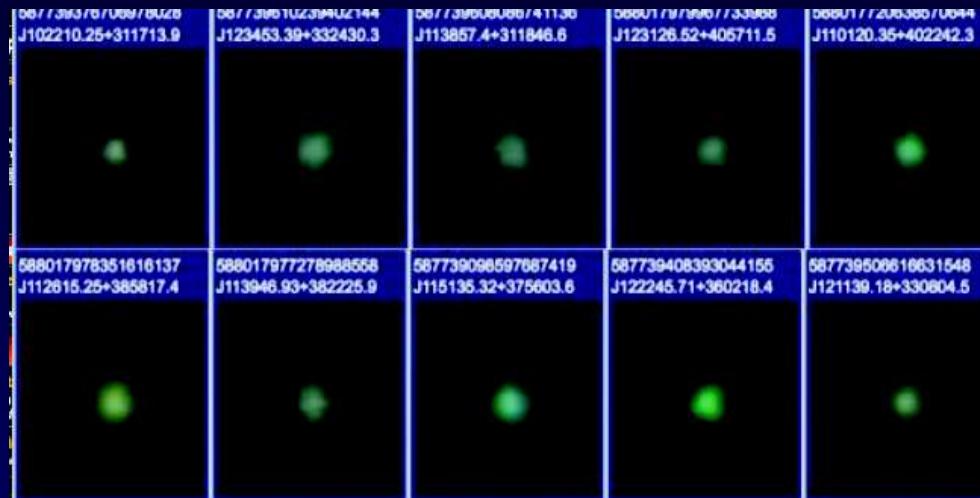


Citizen Science x Expert Science

Verified by human – training sets

Independent answers=estimate of error

Serendipitous discovery



Galactic Peas

Scale - complexity

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.

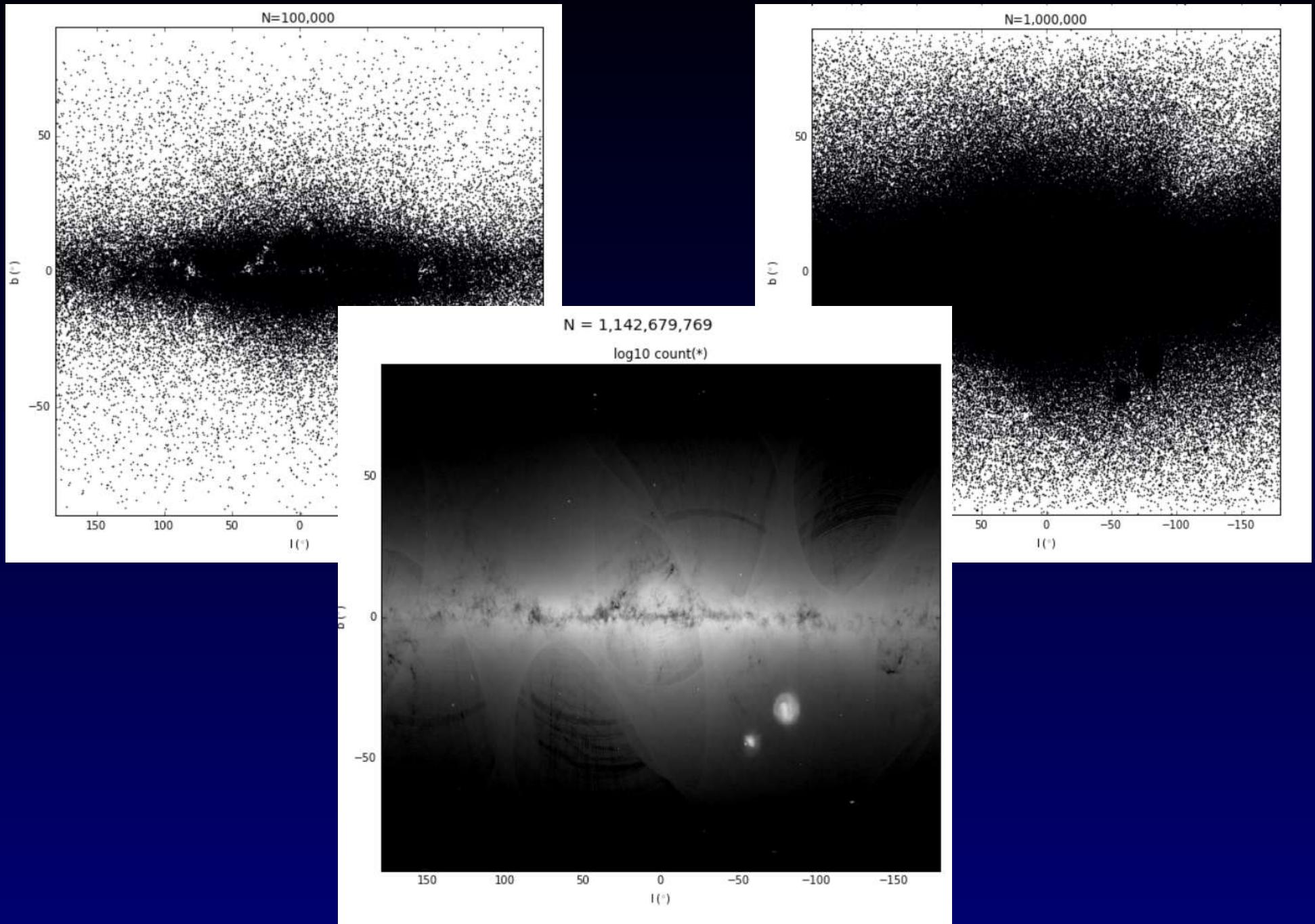
Unknown unknowns :

Serendipity. Currently rely on forum moderators to filter.

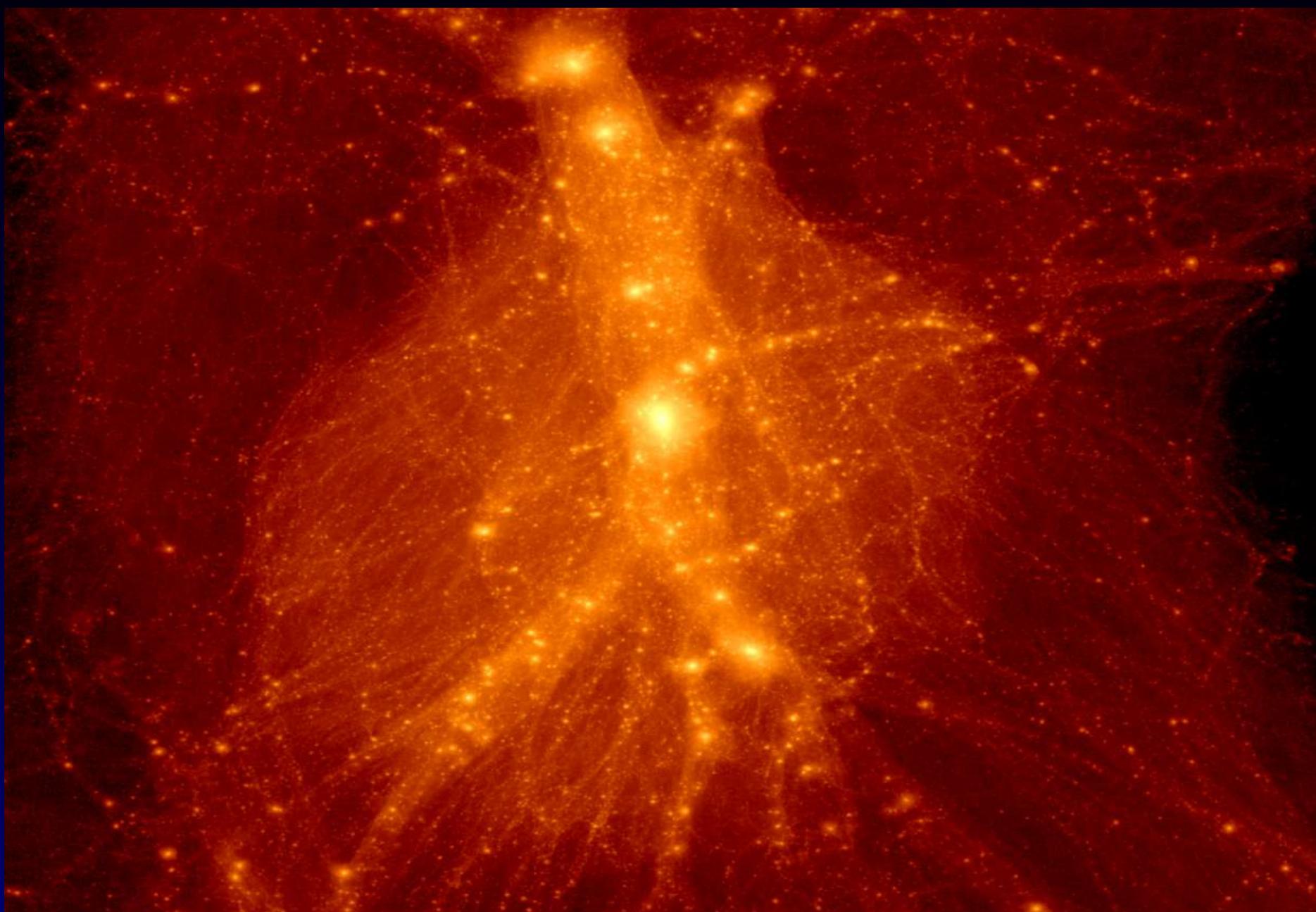
Hanny van Arkel - Voorwerp

Light echo of quasar?

Visualization of 1 B points - Gaia DR1



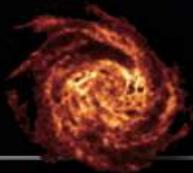
Visualization of Big Data



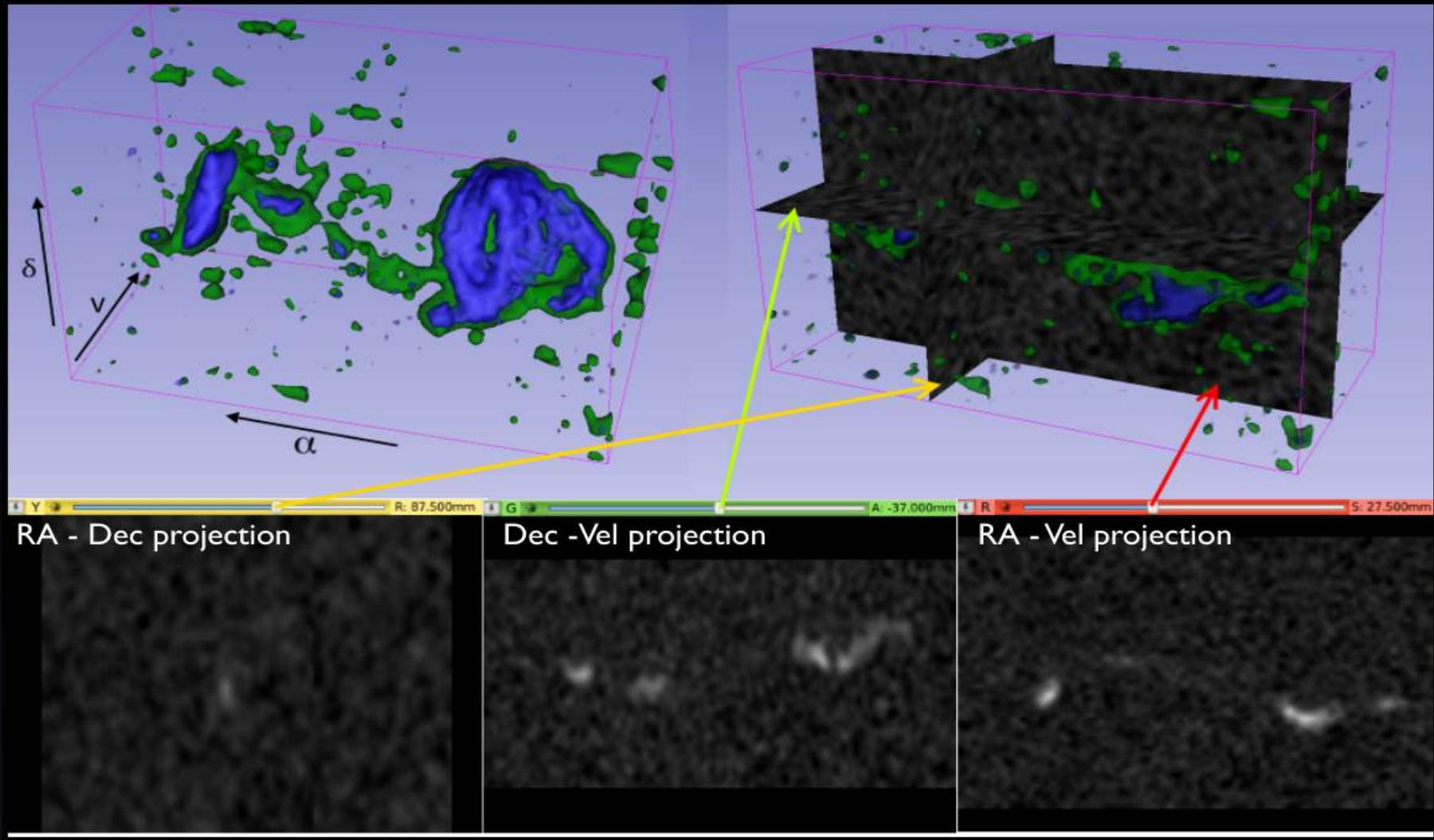
Visualization of Big Data



Visualization of Radio Data Cubes



3D Slicer provides full linked views, not just slices

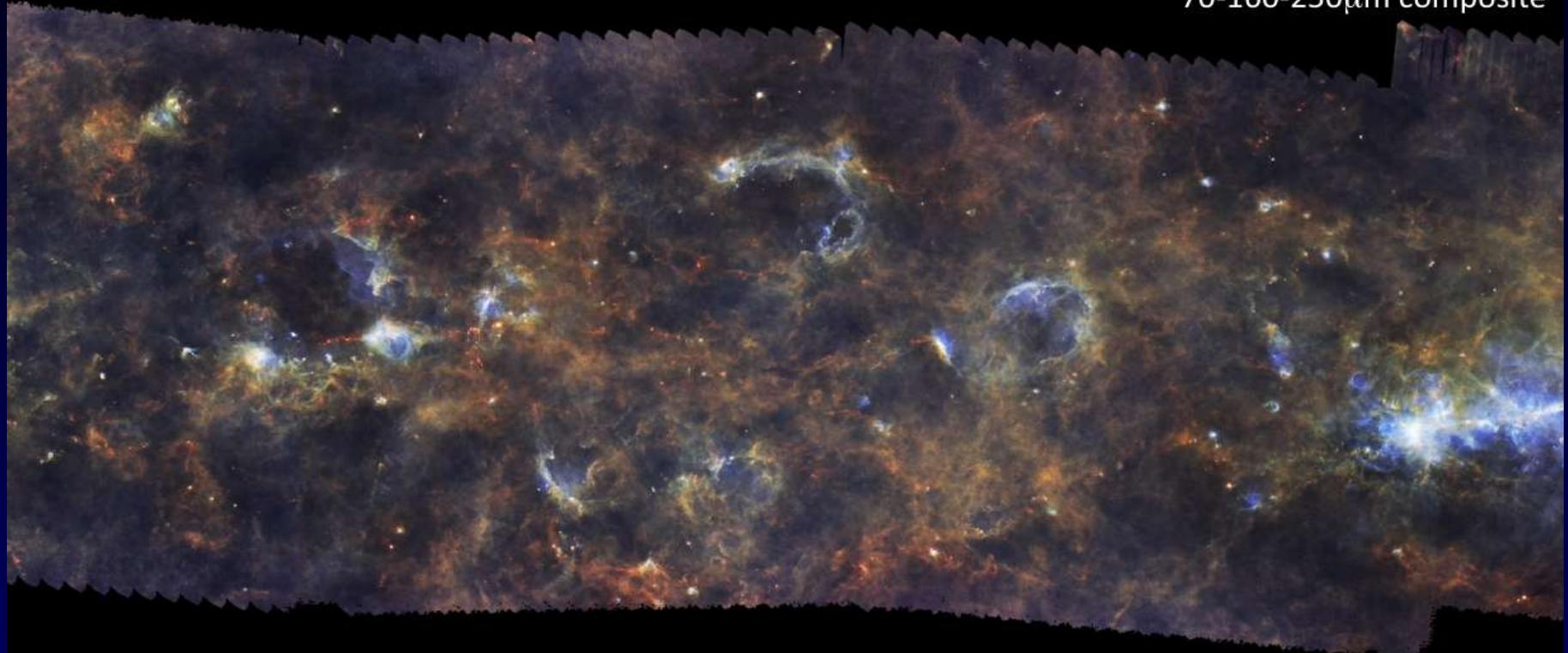


Star Forming Regions in Galaxy

Hi-GAL

the Herschel infrared Galactic Plane Survey

70-160-250 μ m composite



from cold starless clumps to hot HII Regions

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



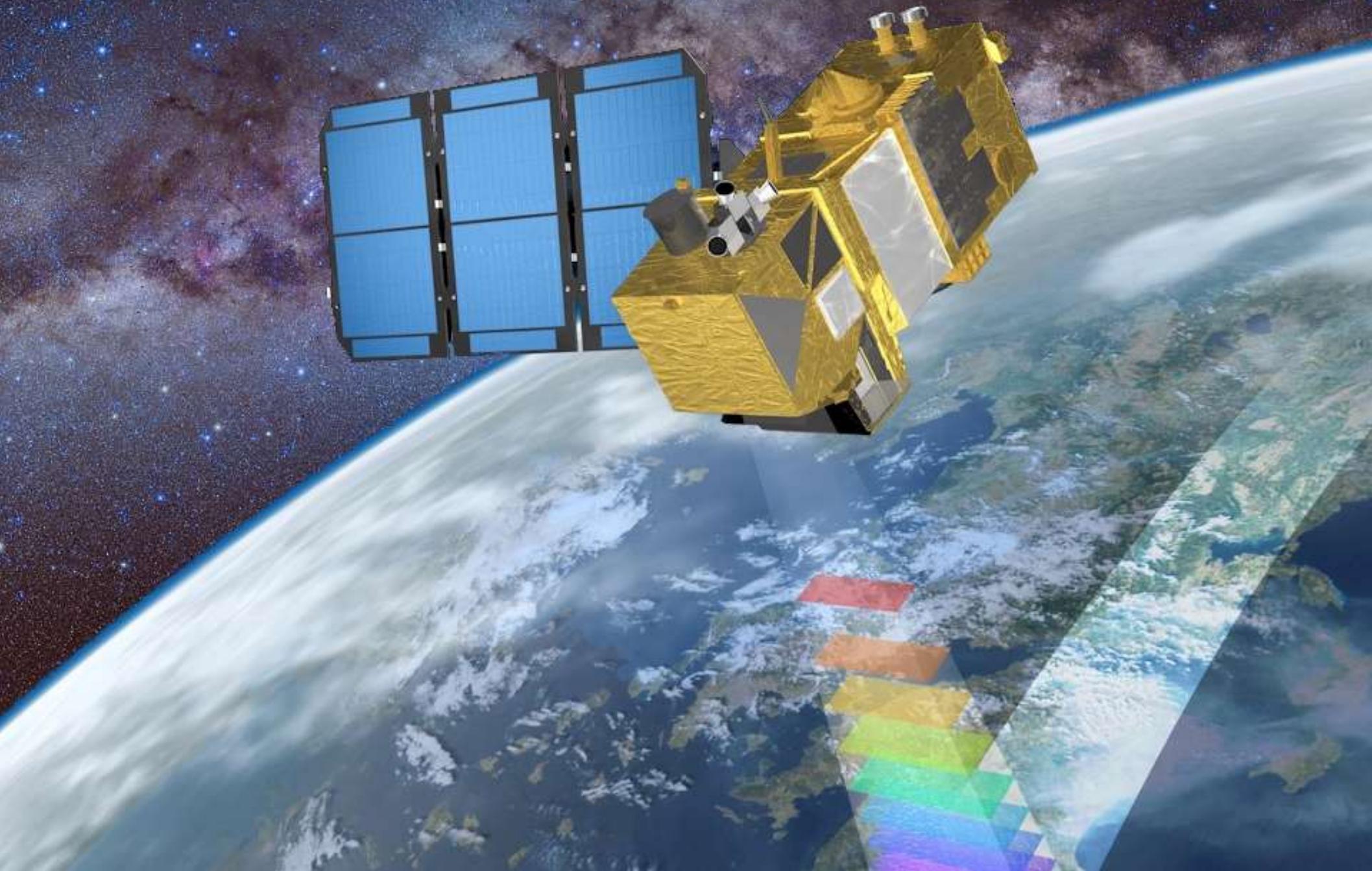
Satellite Products



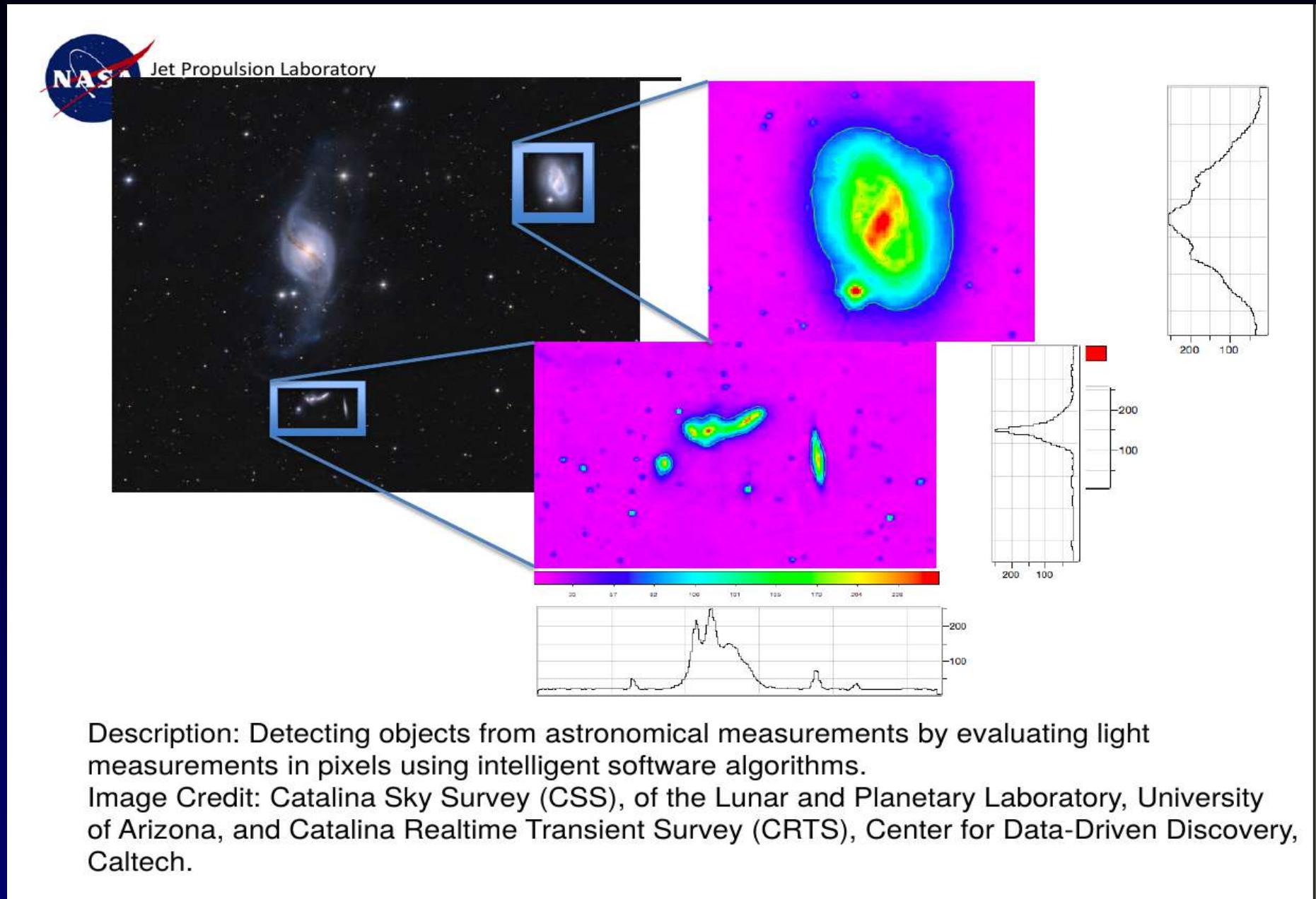
Flash Flood
Warning

Environmental Applications

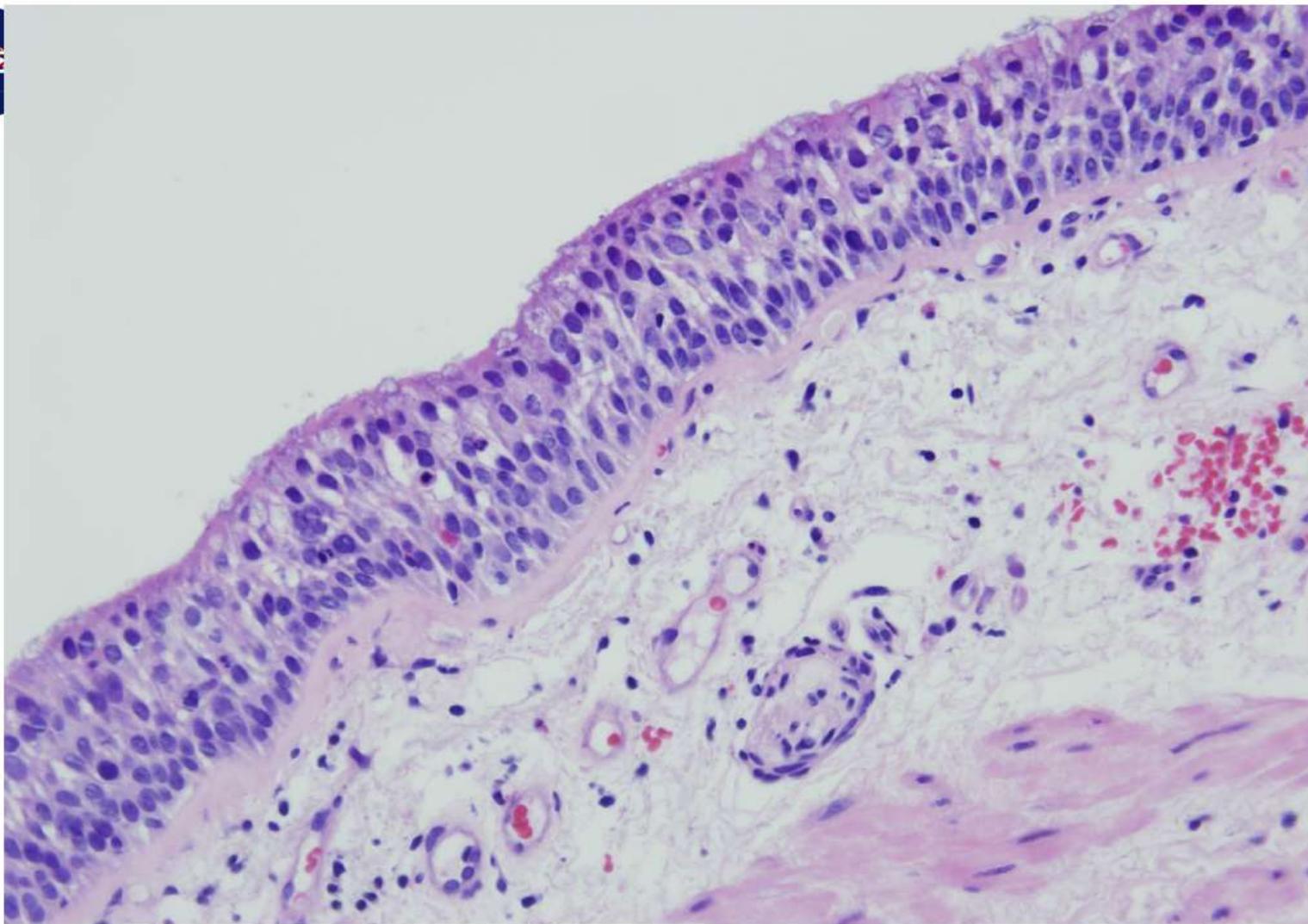
Big Data Era in Sky and Earth Observation – TD 1403 COST action



Finding Galaxies by Shape NASA



Finding Cancer Signatures NASA



Description: Detecting objects from oncology images using intelligent software algorithms transferred to and from space science.

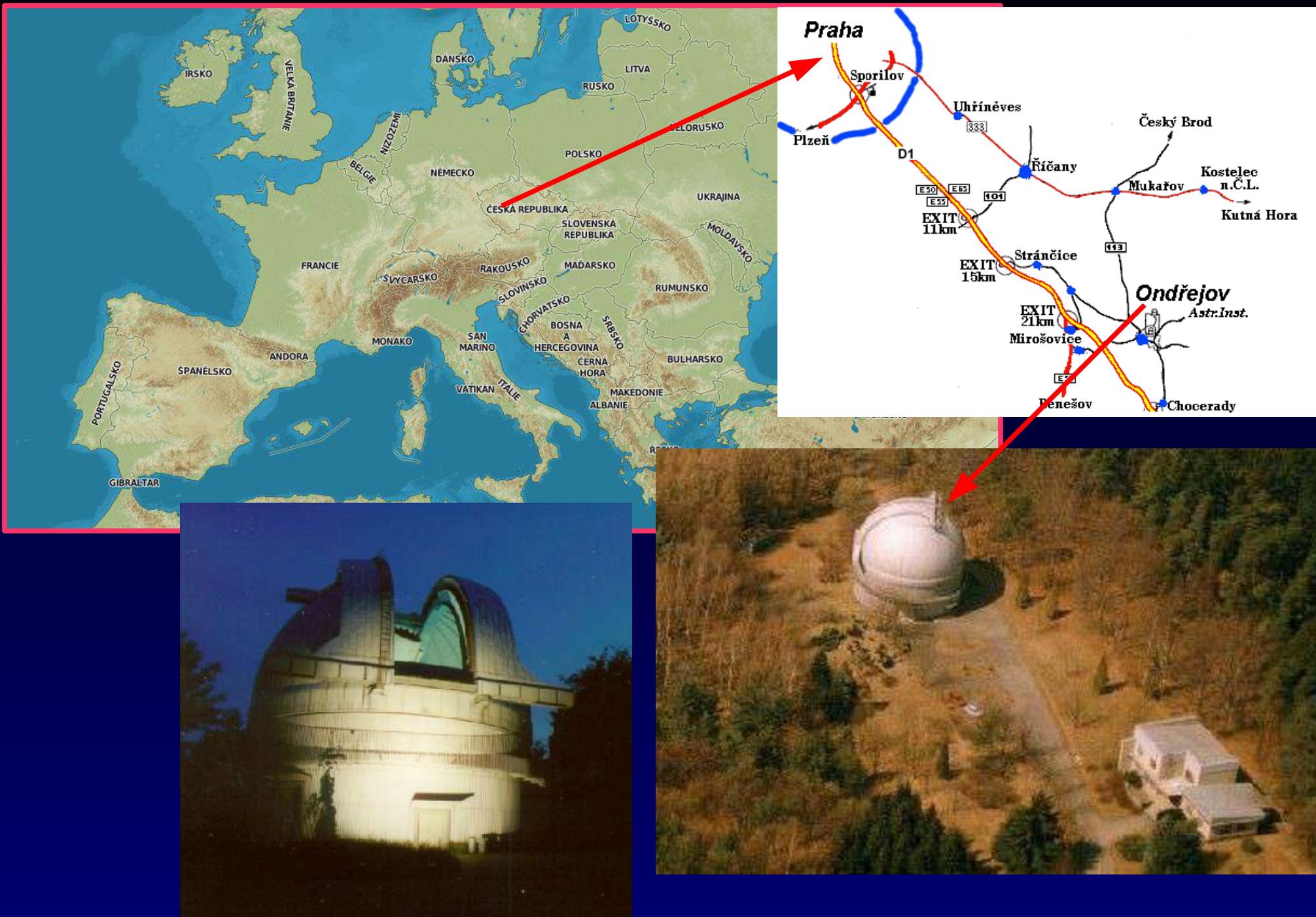
Image Credit: EDRN Lung Specimen Pathology image example, University of Colorado

Challenges of (Astro)informatics

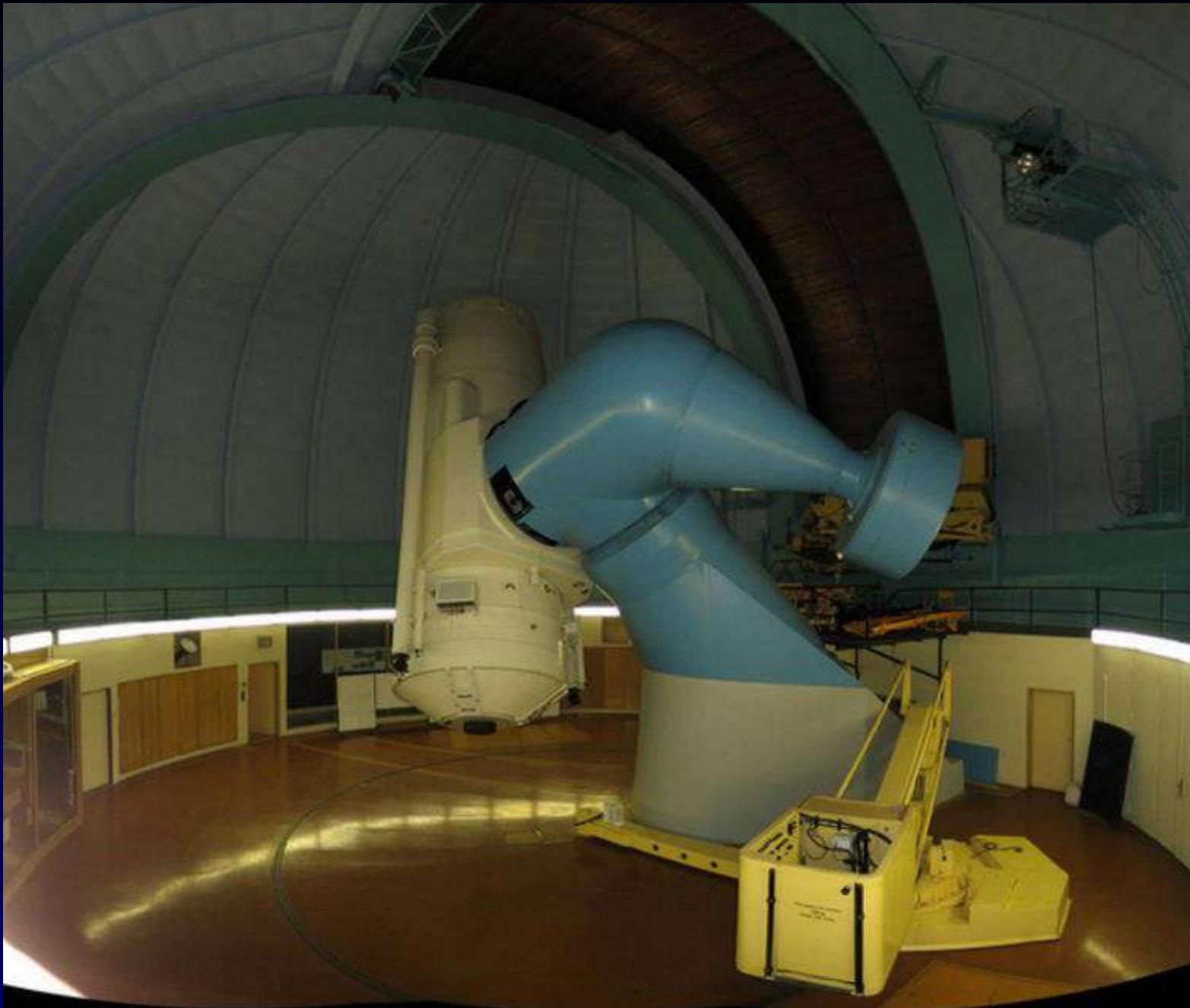
Big Data 3(5)xV

- 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

Ondřejov observatory

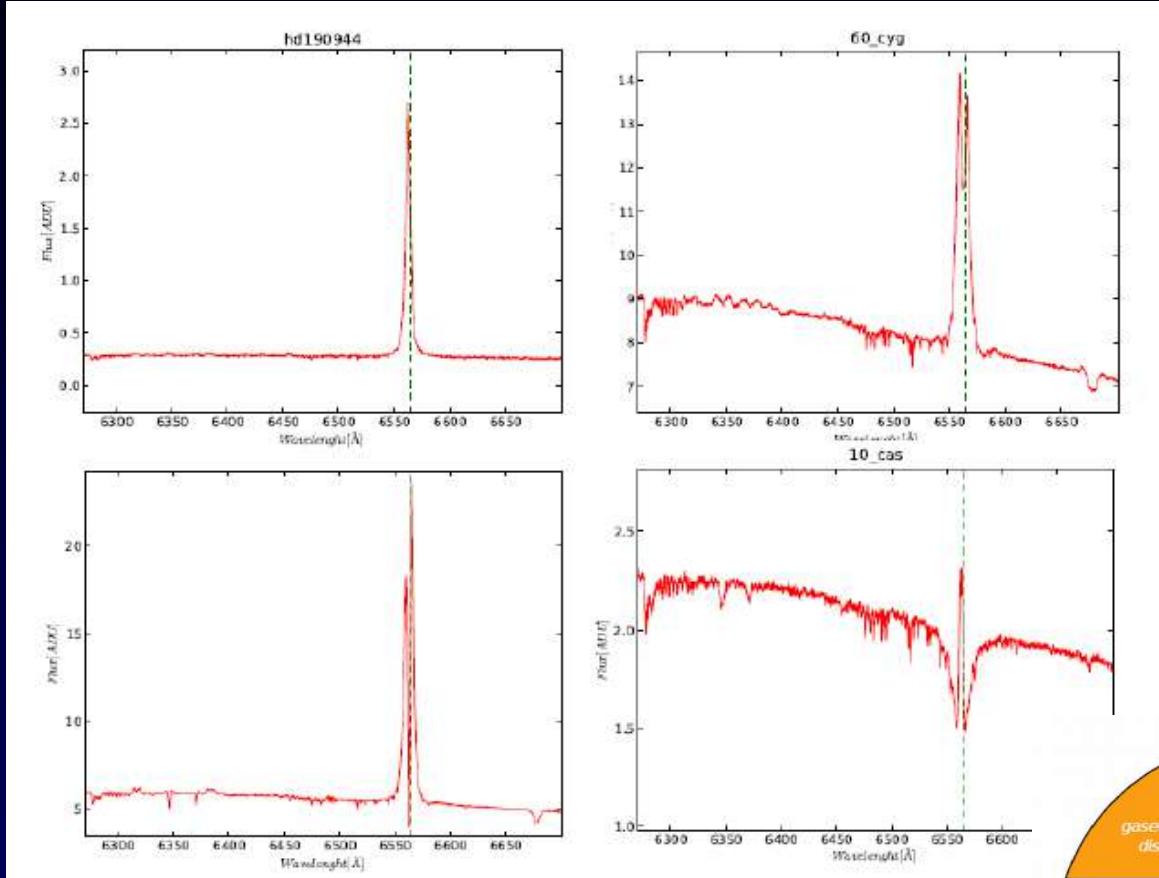


Ondřejov 2m Perek Telescope (1967)



Machine Learning of Spectra

Use case: ML of spectra profile of Halpha line (Be stars)

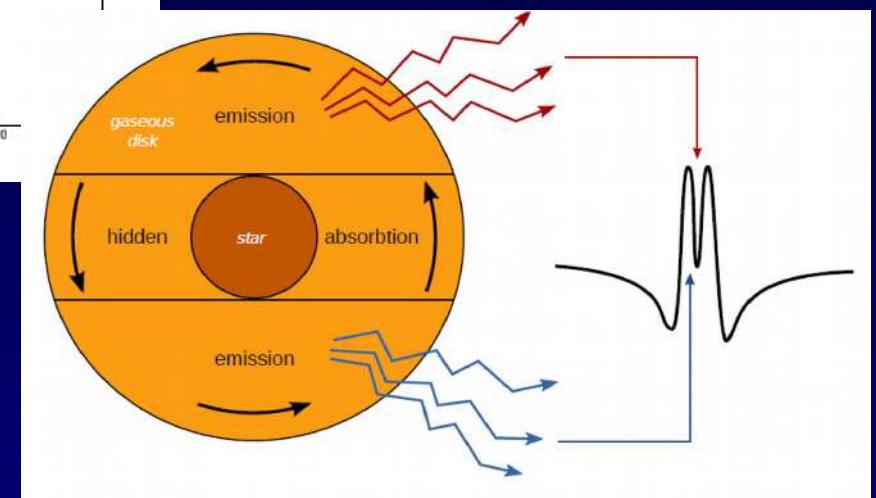


Be stars

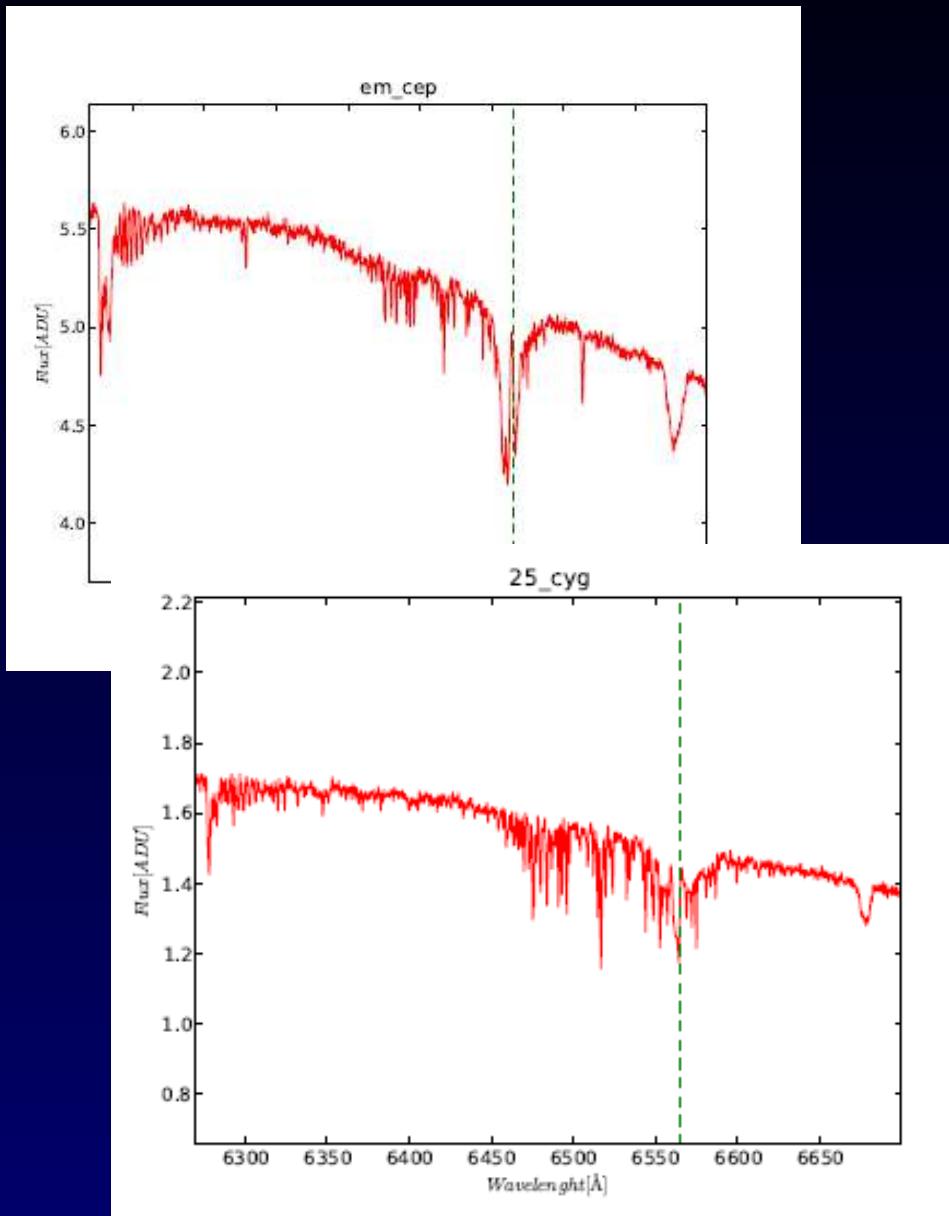
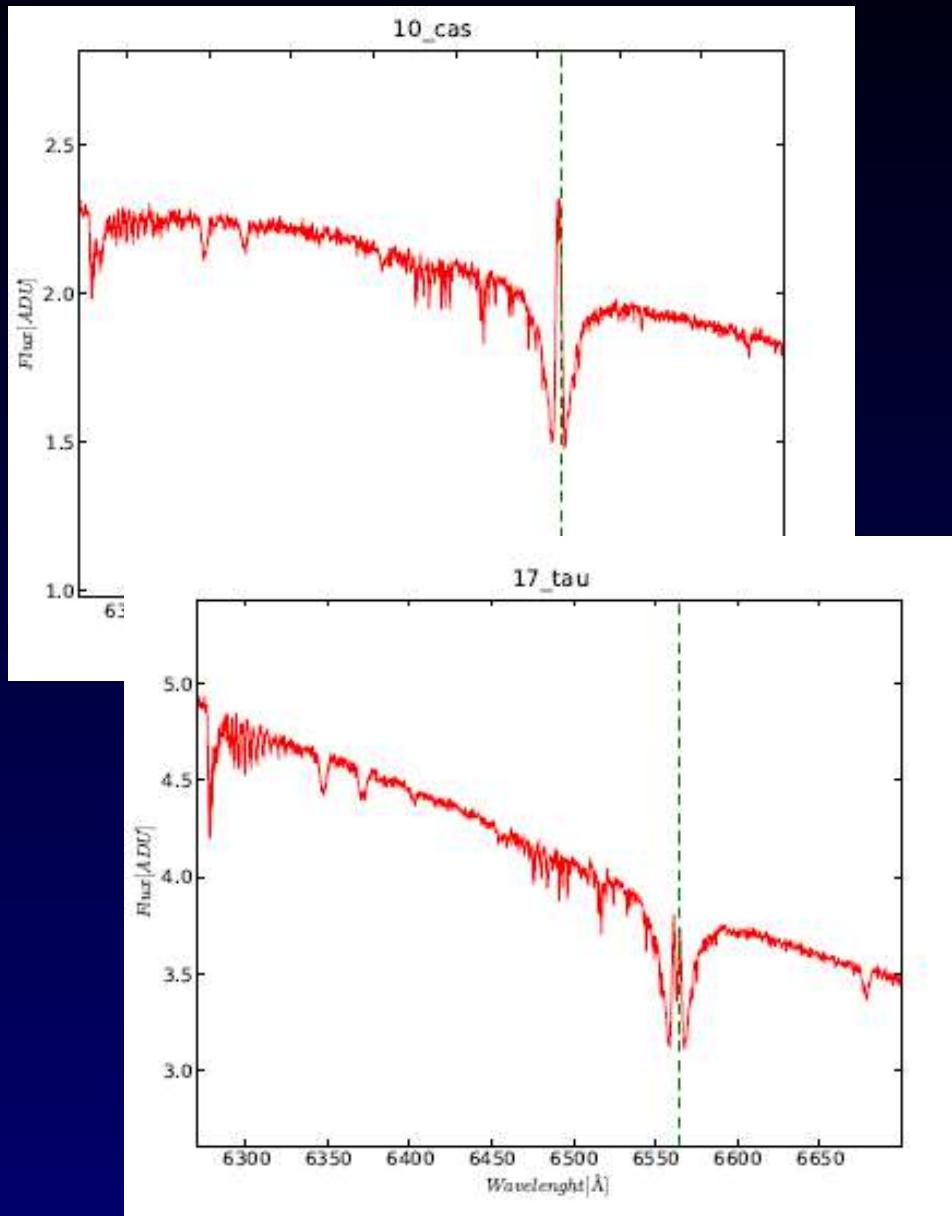
Disk or envelope

Rotates, Hot

Origin ?????

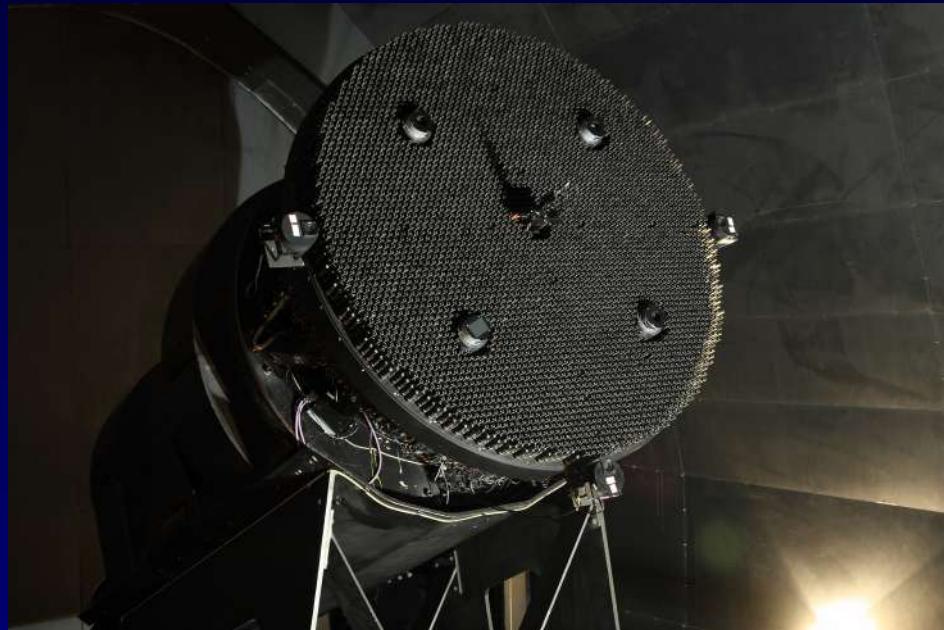
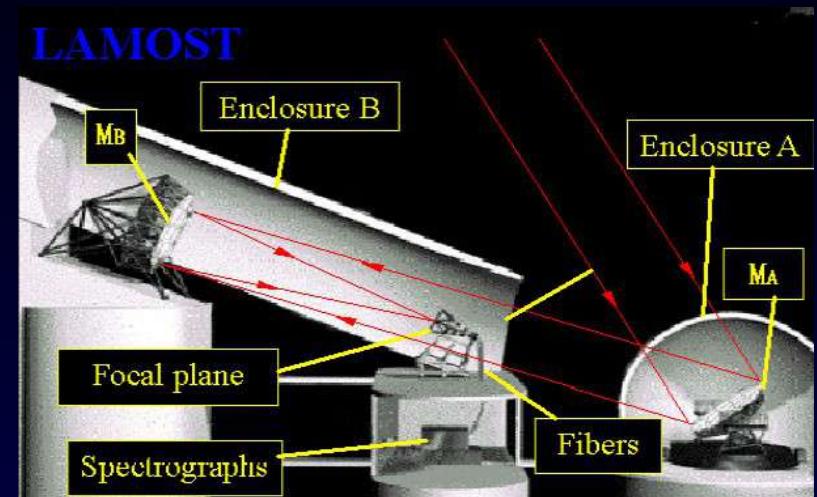


Be Stars : Emission in absorption



LAMOST (Guoshoujing)

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



LAMOST Spectral Surveys

DR1 (end 2013) **2 204 860** spectra
 1 085 404 stars

DR3 (half 2015) **5 755 126** spectra

DR4 (Feb 2016) **+ 741 522**

Each Fiber – 2 motors
double arm 33mm circle

Fibre collects light from
3.3 arcsec circle on sky



Hobby Eberly Telescope (HET)



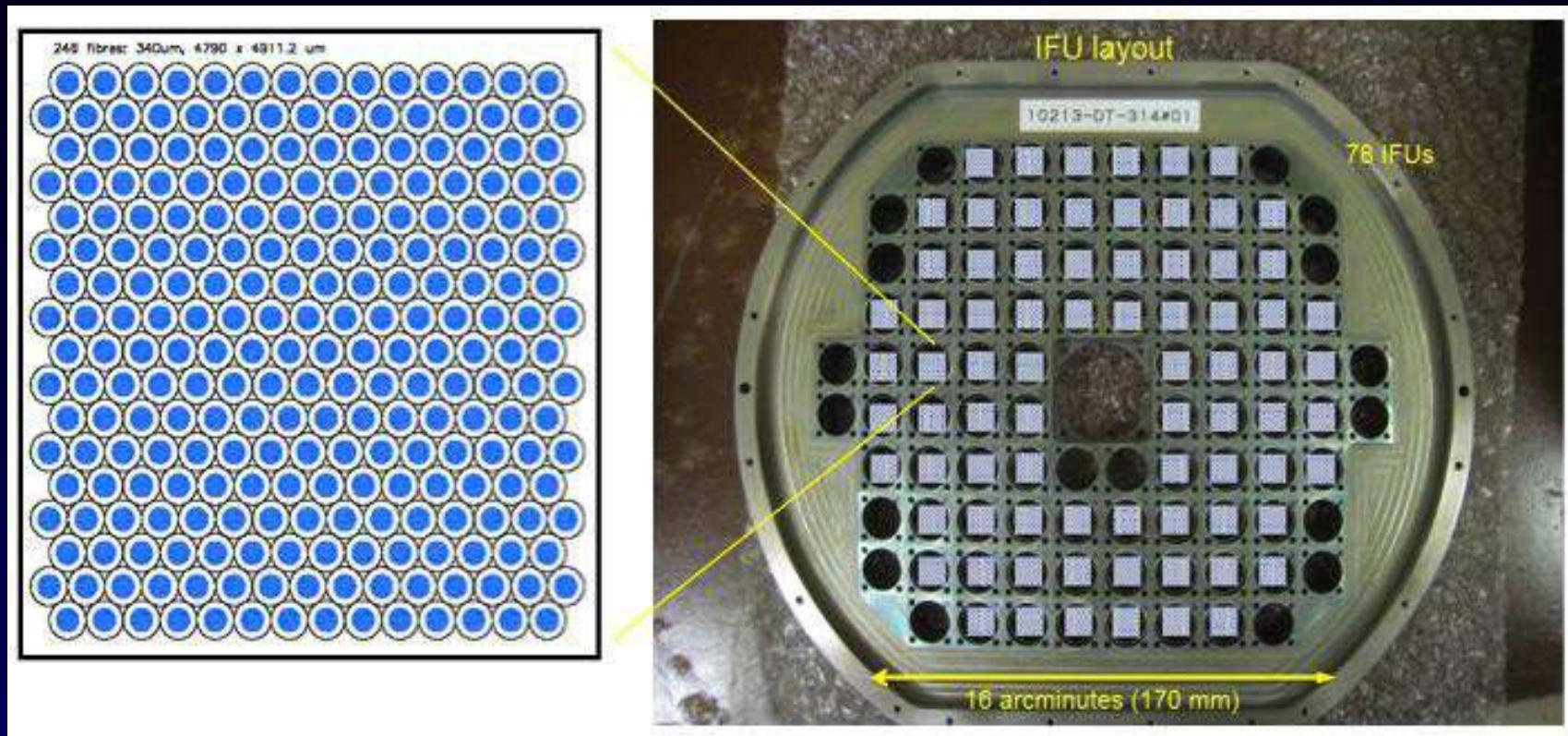
Mc Donald Observatory Texas

Equiv diameter 9.5m (11m)

Fixed in position during observation -
only primary tracker

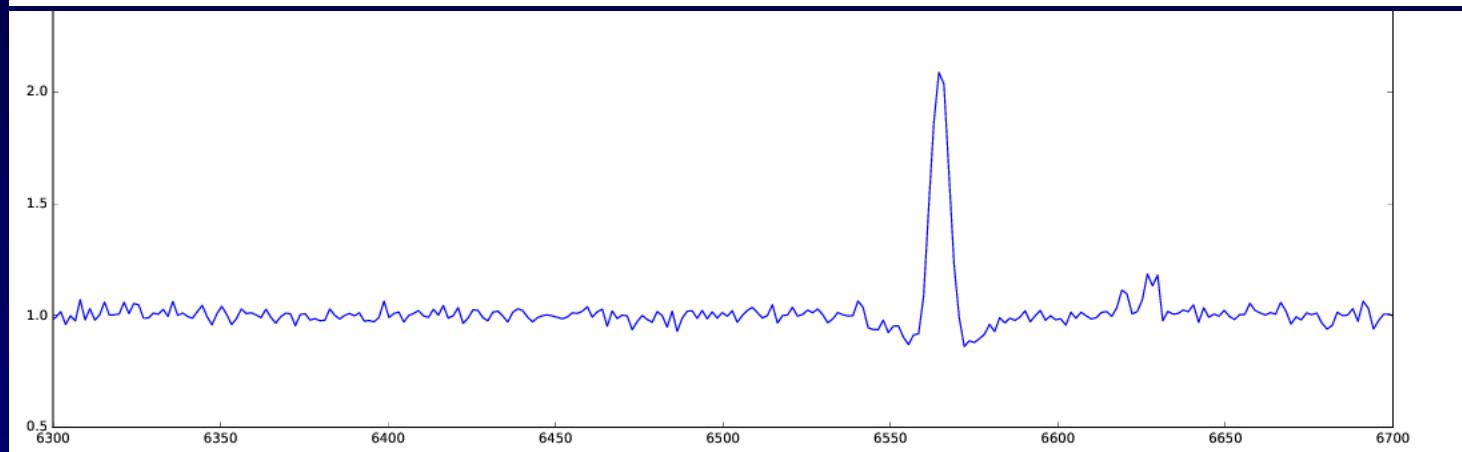
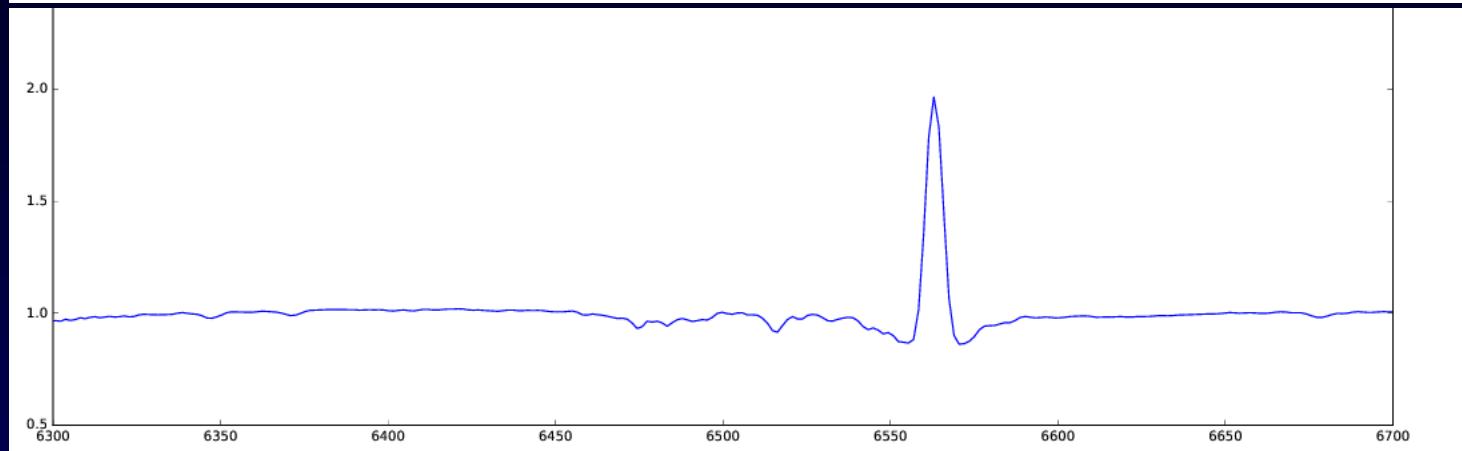
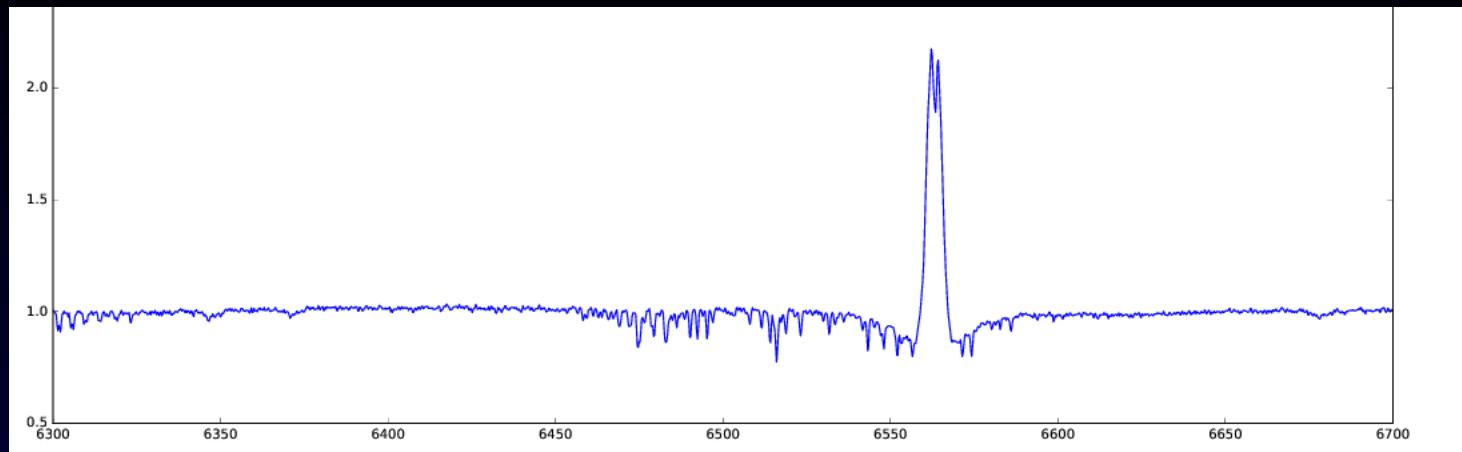
HETDEX Survey

In theory 34944 spectra every 20min !

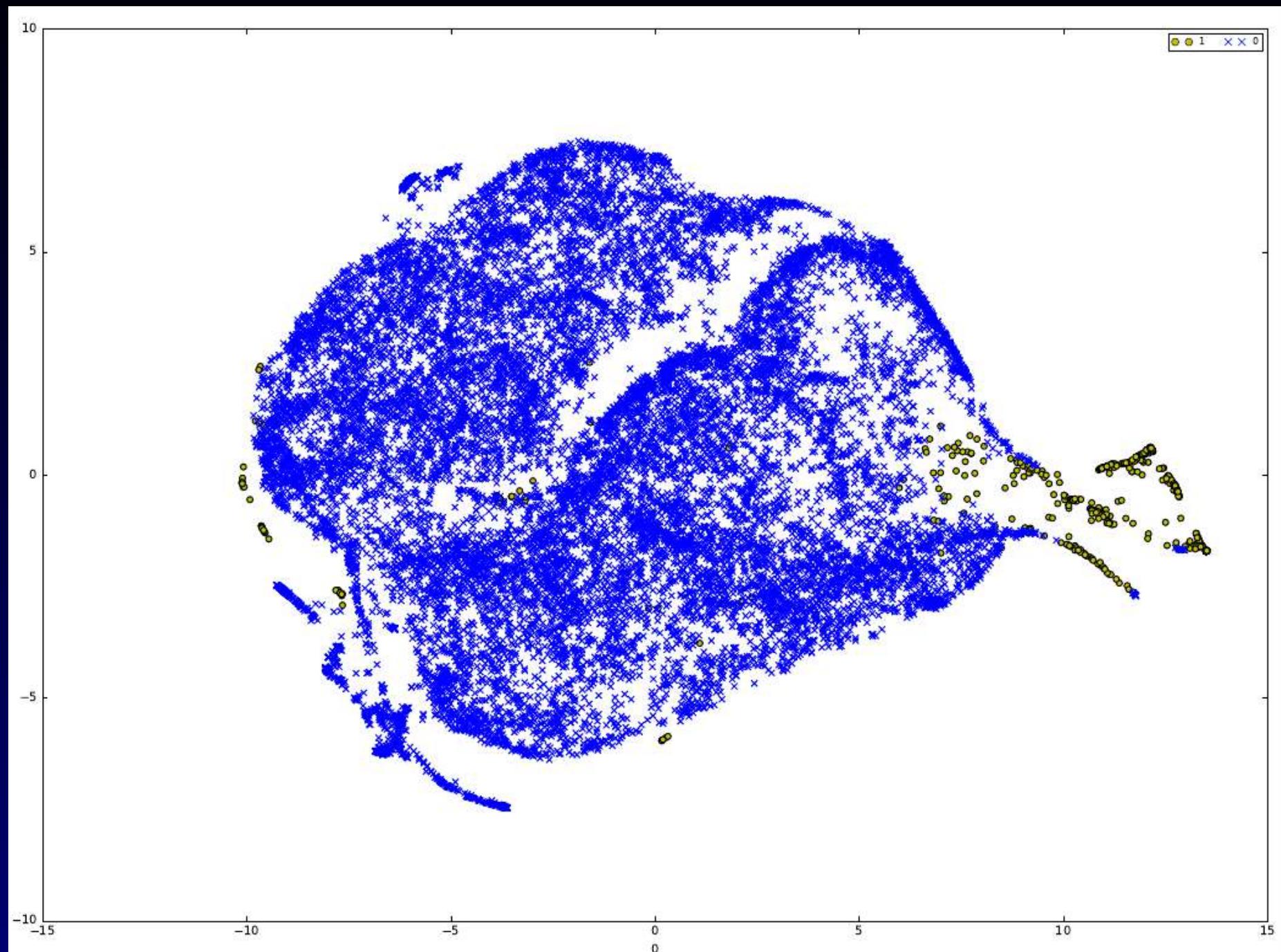


VIRUS 78 IFU = 156 spectrographs
IFU= 448 fibers
34944 fibers , FOV 22 arcmin, 3500-5500 Å, R=800
1 million spectra of galaxies (only part - statistic hits)

Resolution Degradation



LAMOST TSNE Structure



Semi-Supervised Training

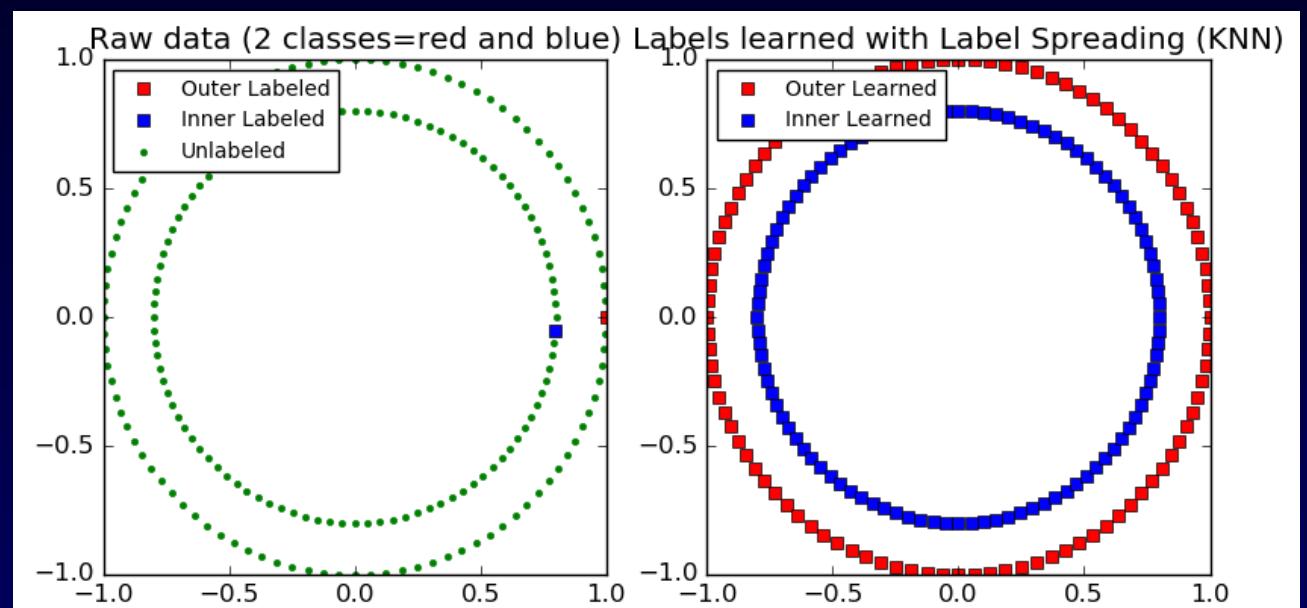
Not supervised (even if not Domain Adaptation)

- sample of labelled data - about 1600
- unlabelled (LAMOST) - HDFS limit to 1,048,576 (2^{20})

Graph methods:

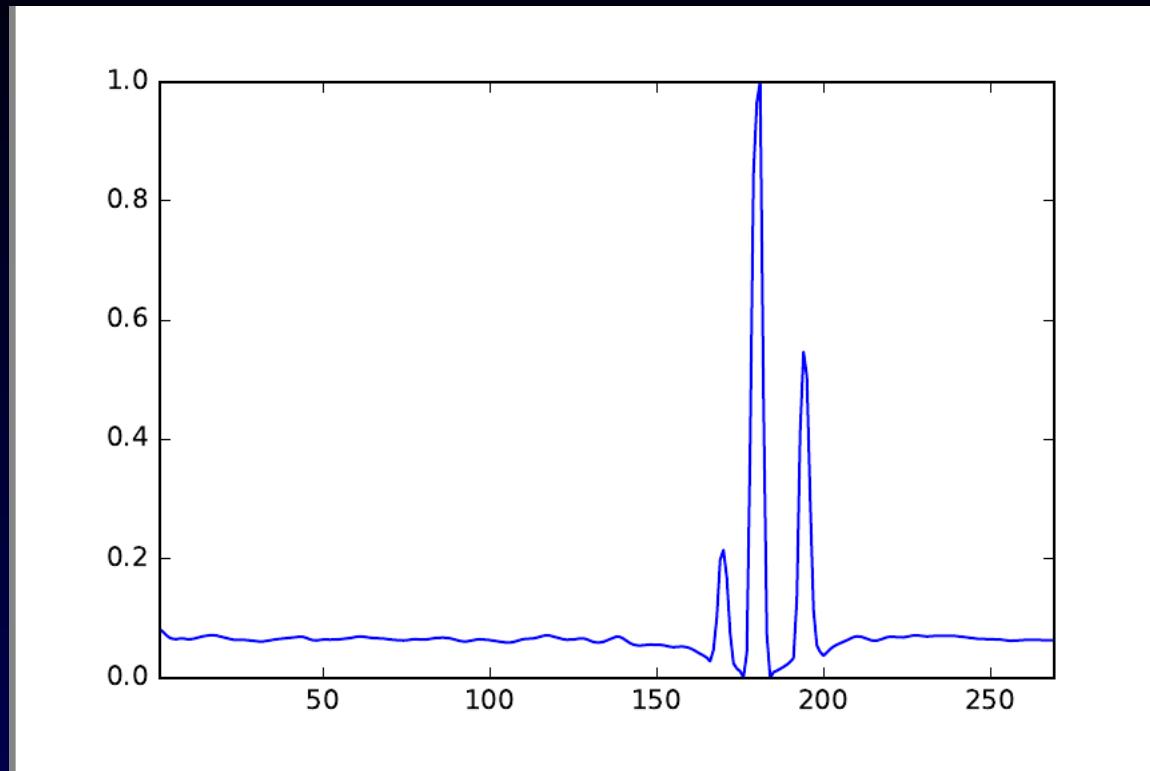
Label spreading

Label propagation

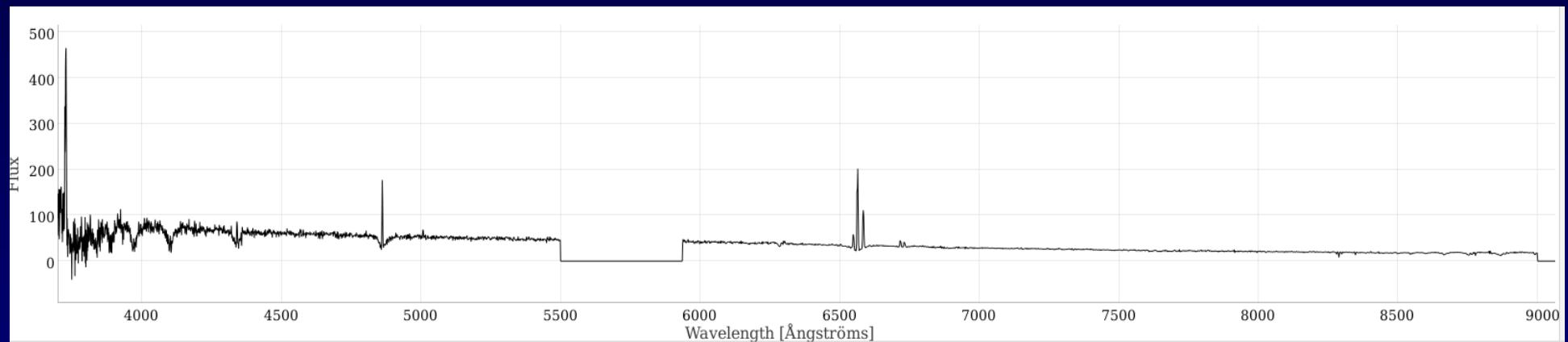


Spark on HDFS - National cloud MetaCentrum

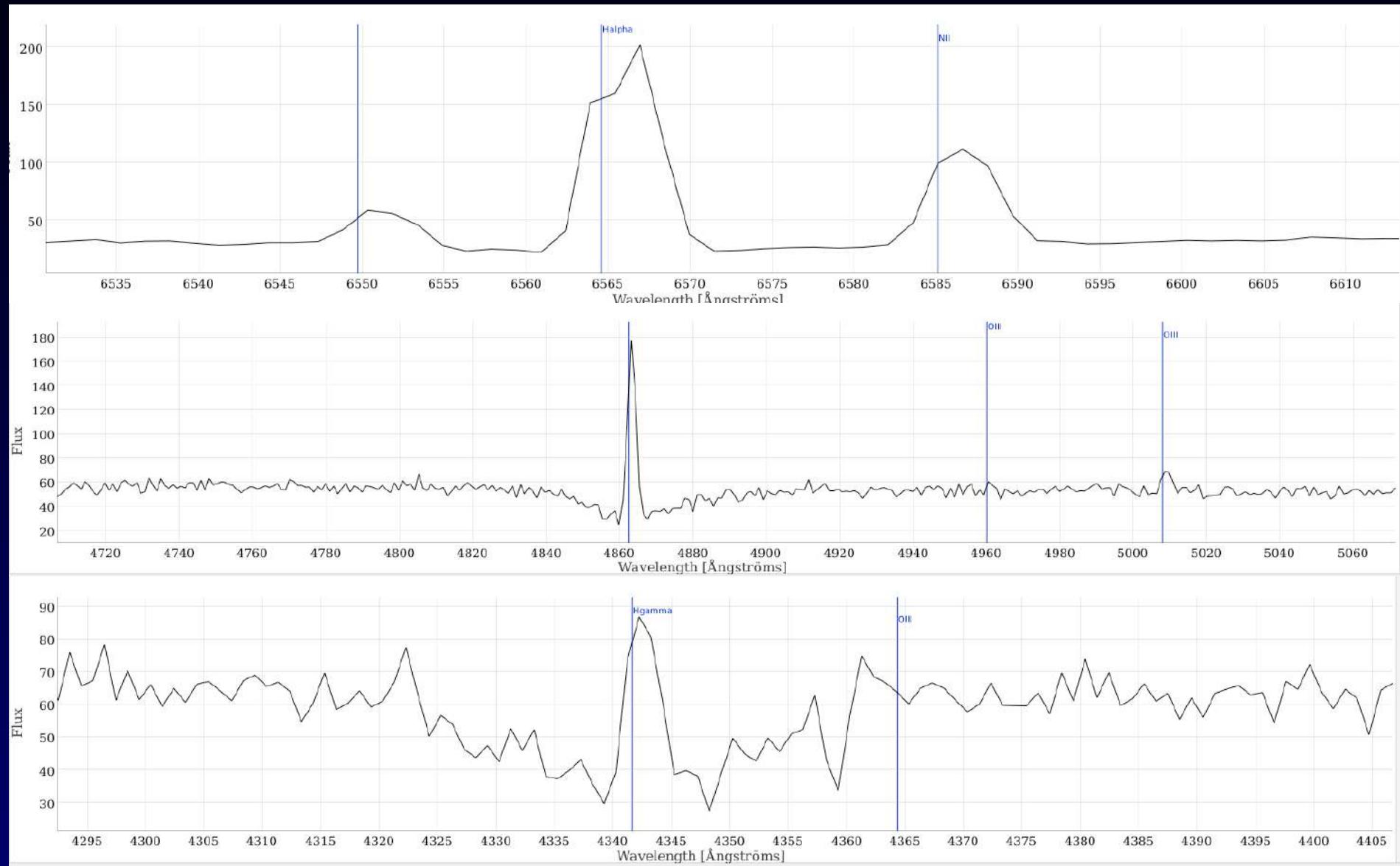
Be Candidates - Semi Supervised ML



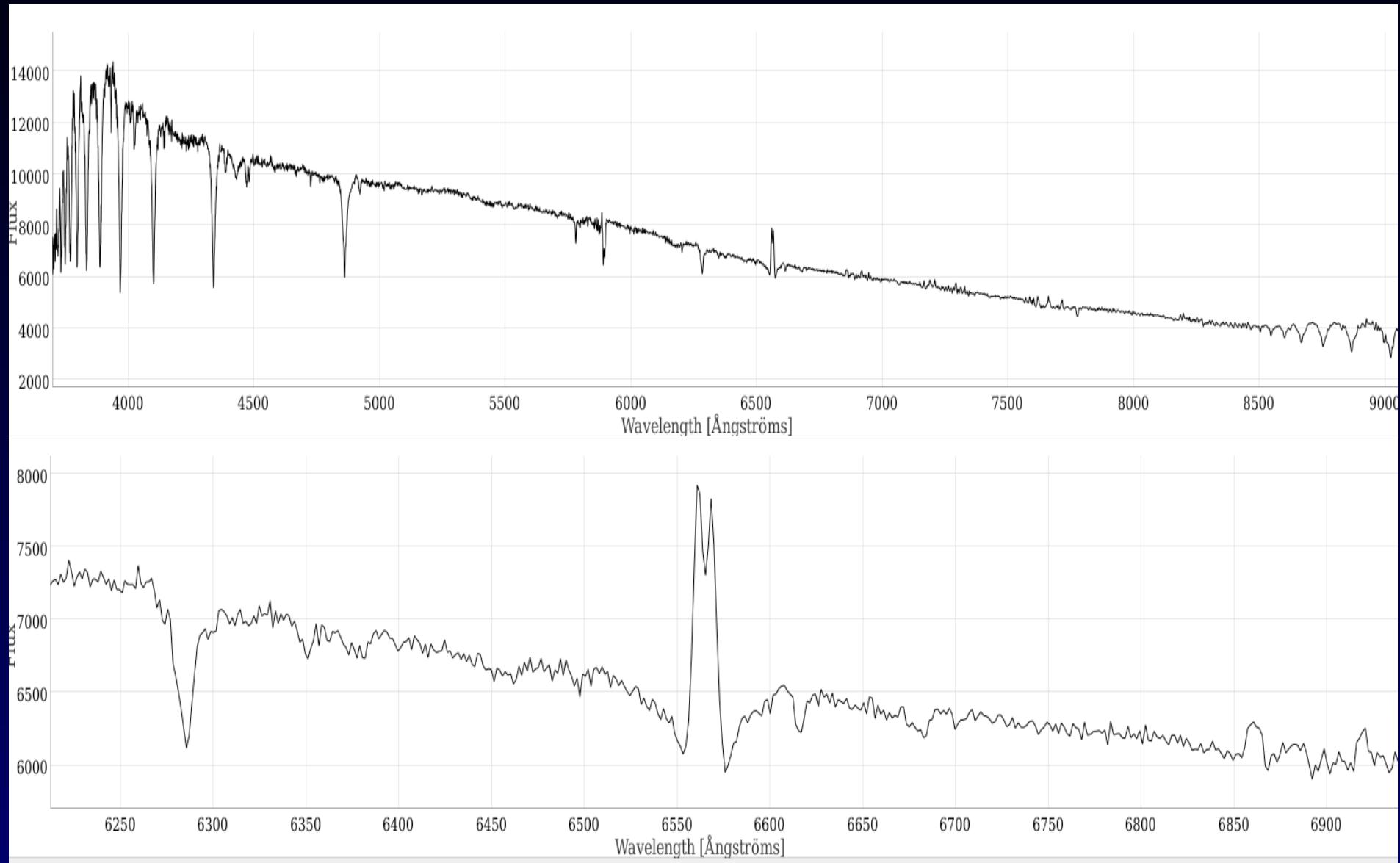
Palička 2016



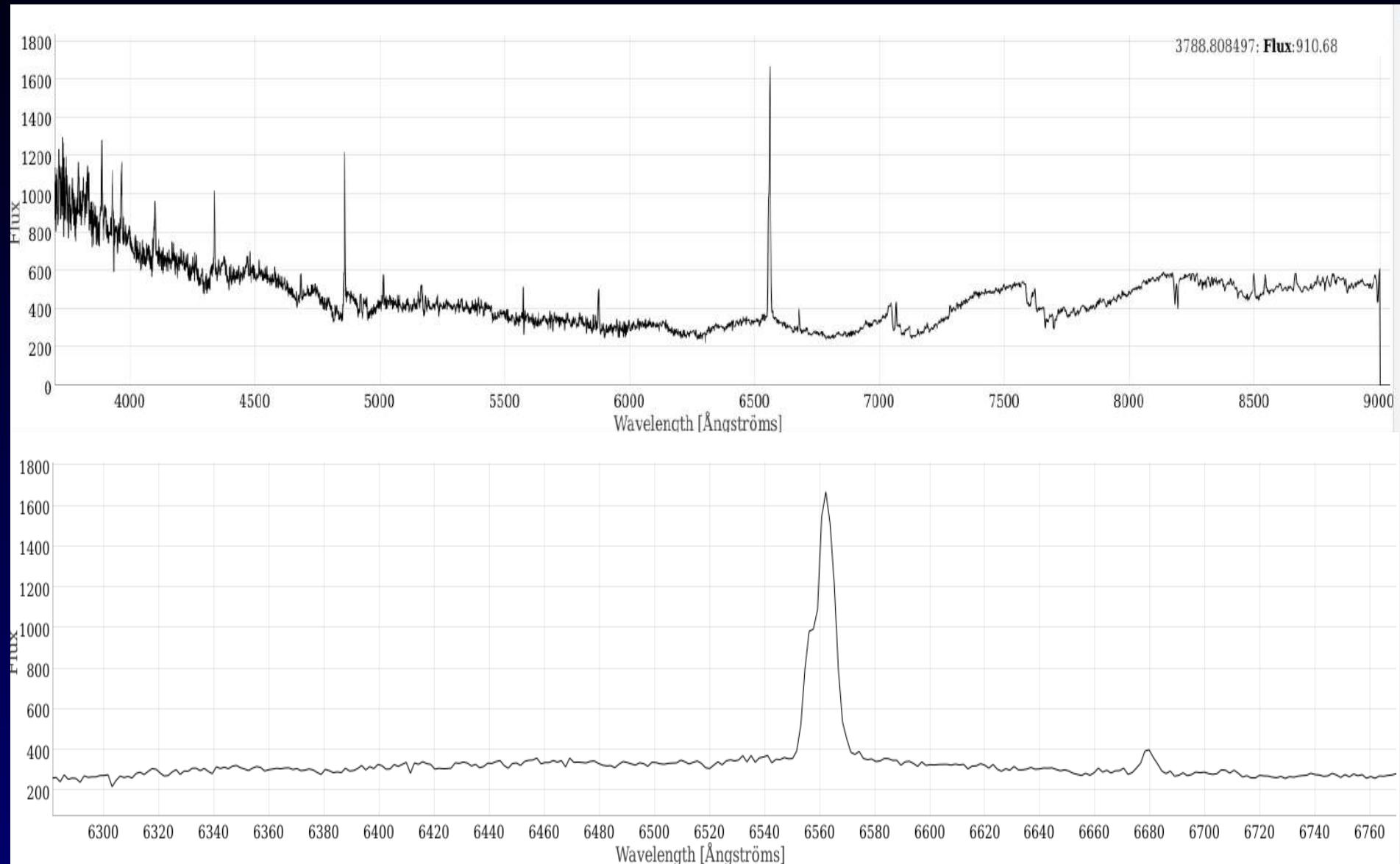
Be Candidates Found



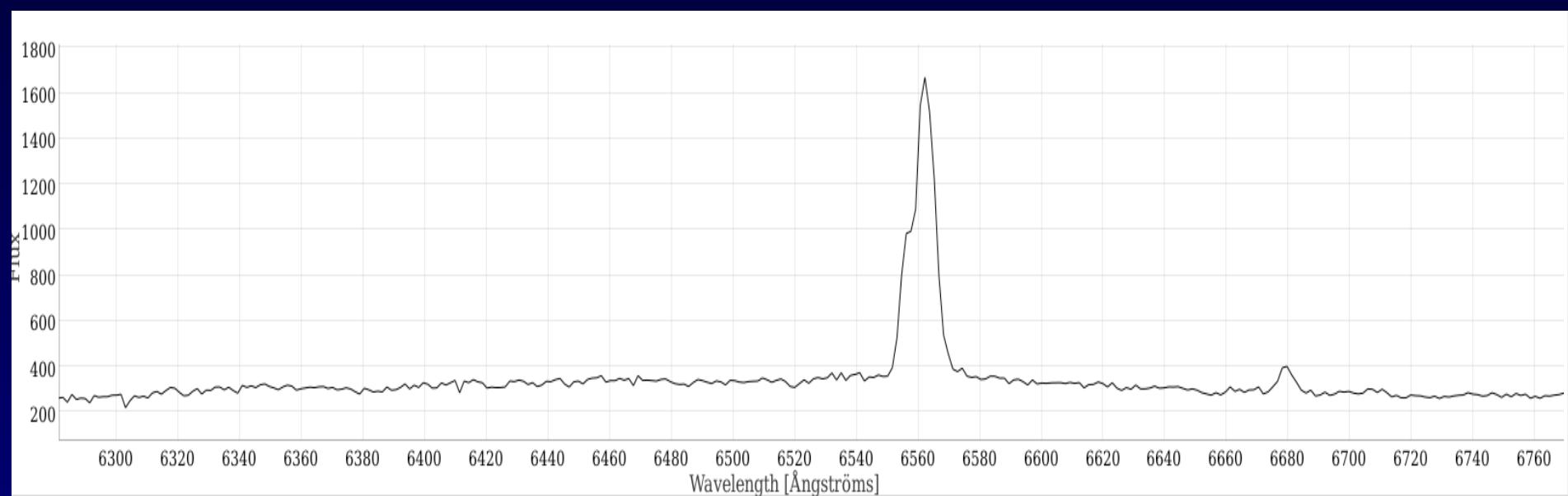
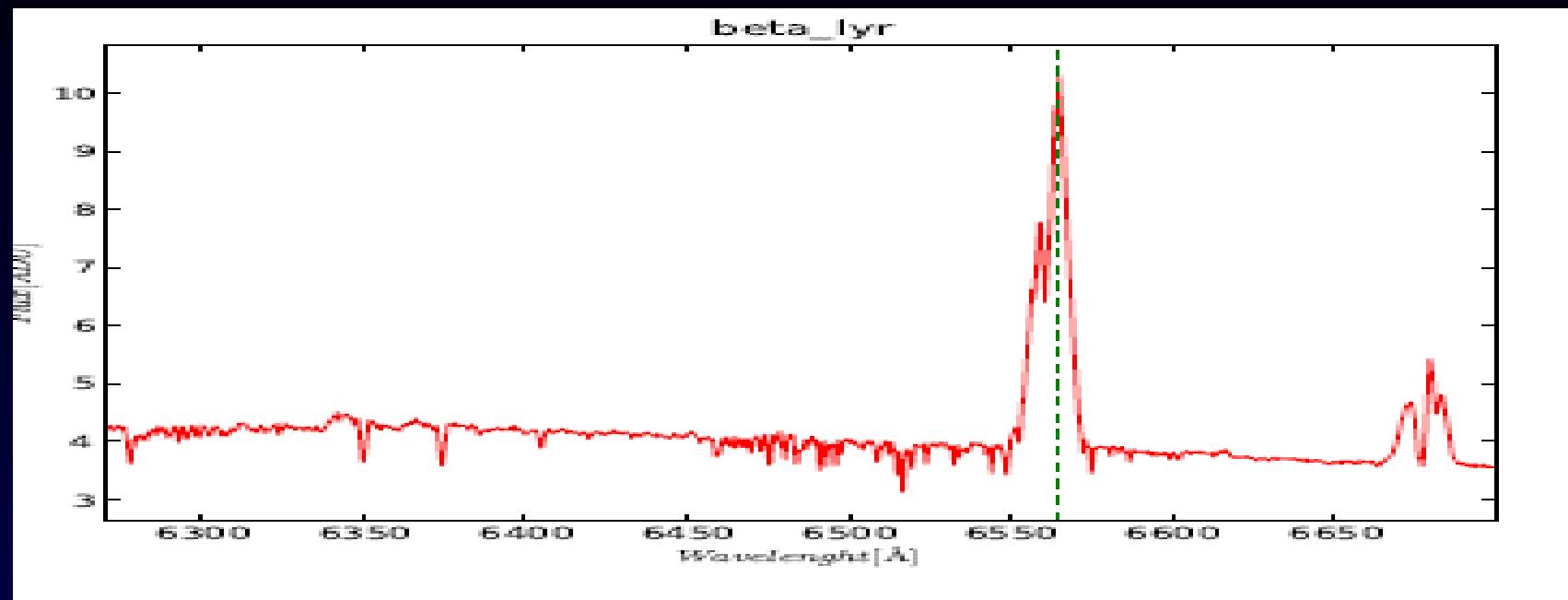
Be Candidates Found



Be Candidates Found

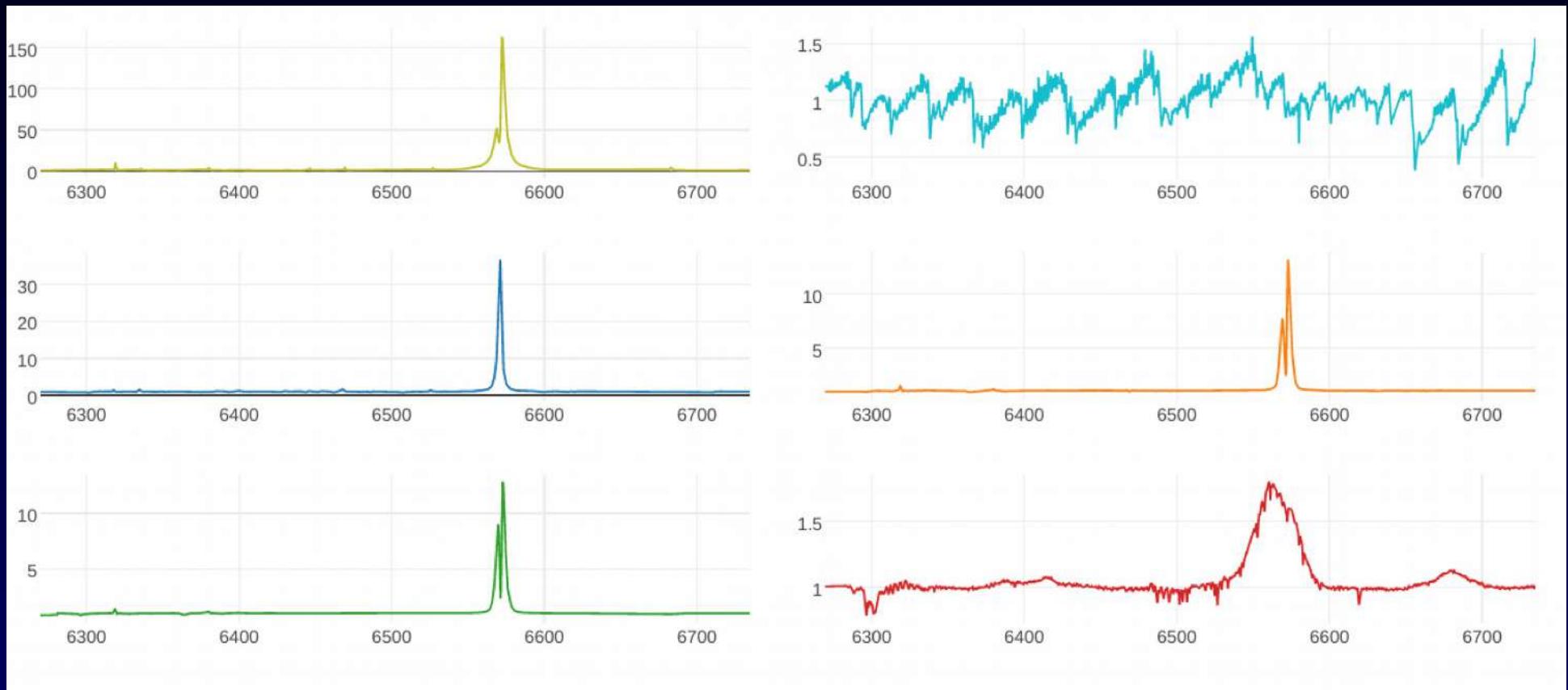


Be Candidates Found



CCD700 Outliers

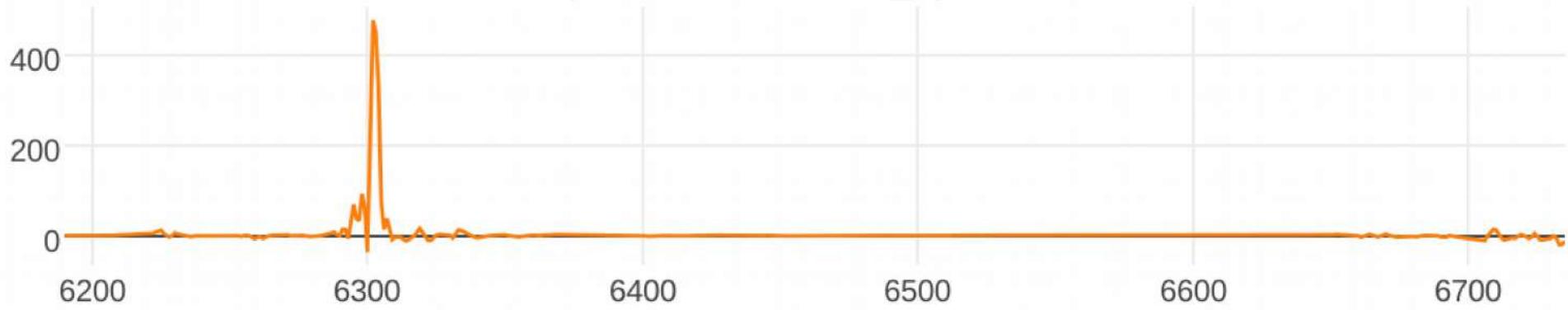
Unsupervised learning – Local Outlier Factor - LOF



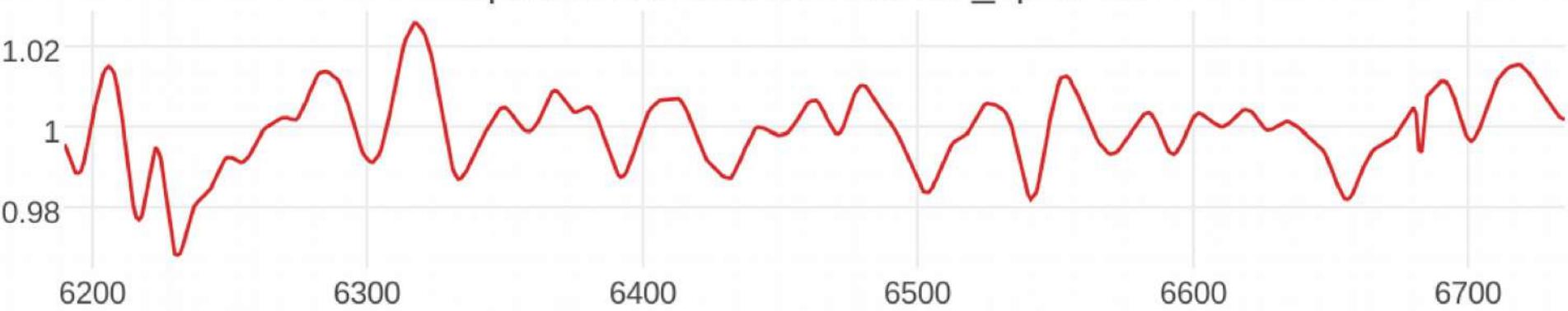
Shakurova 2016

LAMOST Outliers

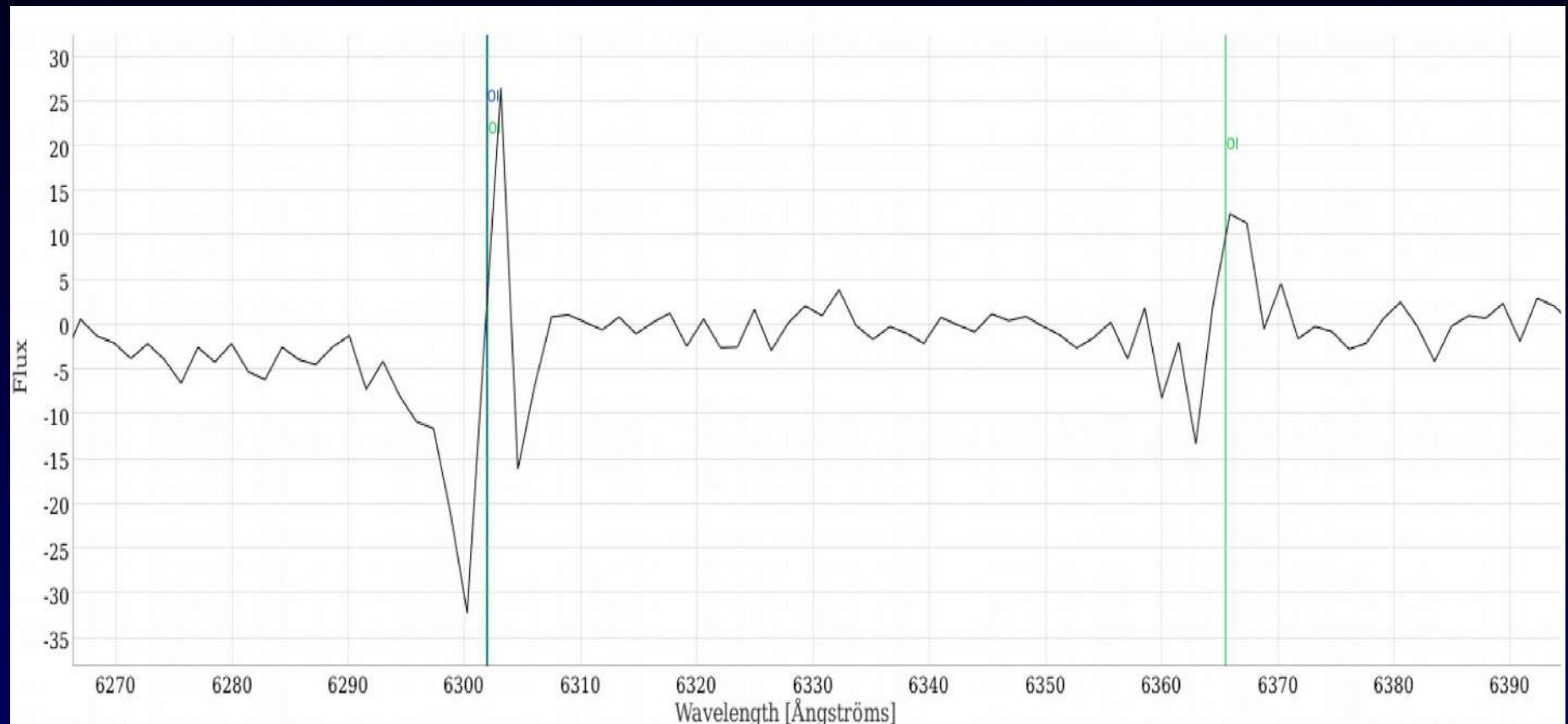
spec-55859-F5902_sp11-089



spec-56225-GAC100N32B1_sp15-052



LAMOST Be star - outlier



Concept of scientific „CLOUD“

ITERATIVE REPEATING of SAME computation (workflow)

Global non-linear optimization (Korel)

Synthetic spectra (various elements, wavelength-ranges)

Machine Learning (almost all methods)

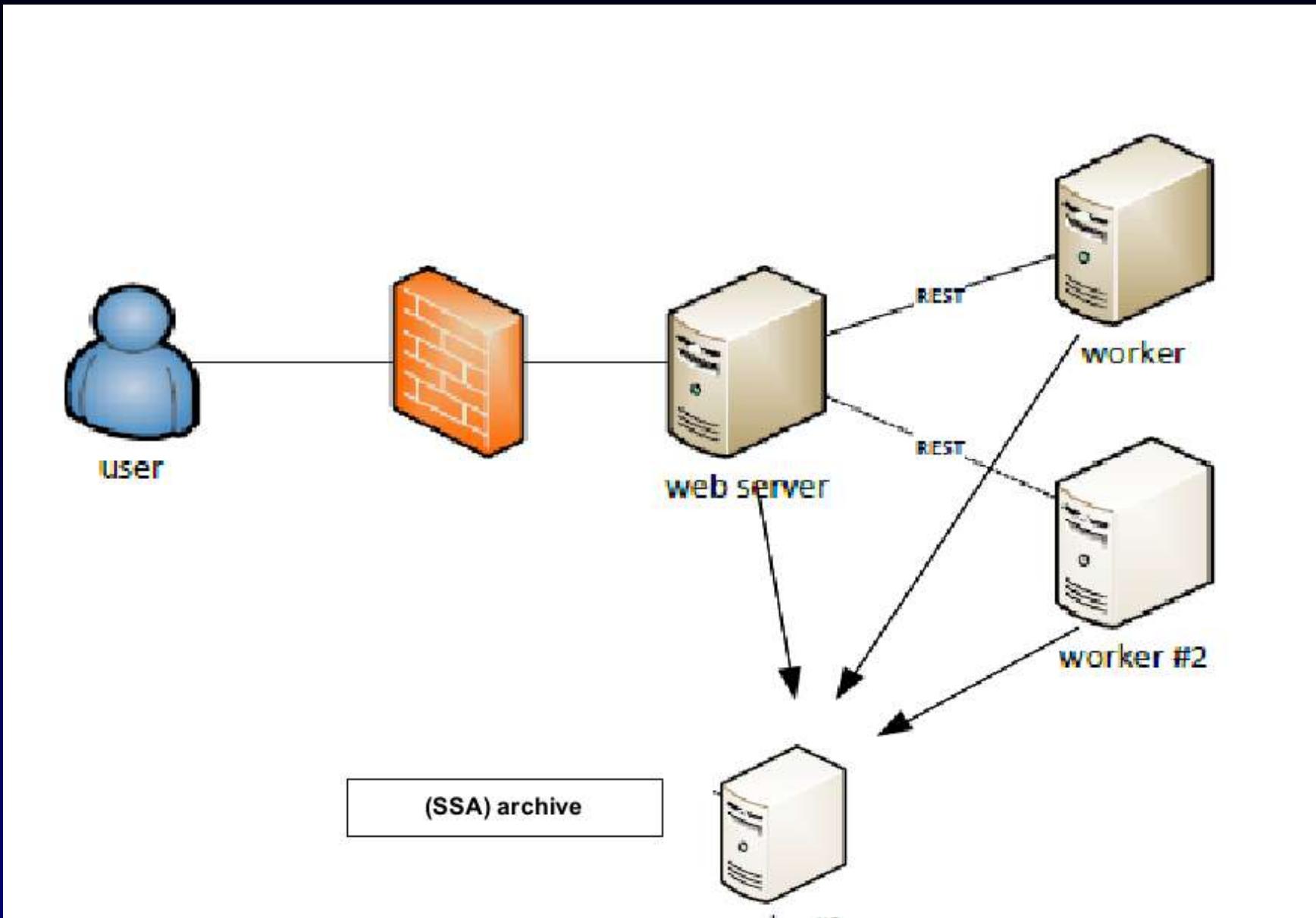
LARGE stable INPUT data + small changing PARAMS

Many runs on SAME data (tuning required)

Graphics visualization from postprocessed output (text) files

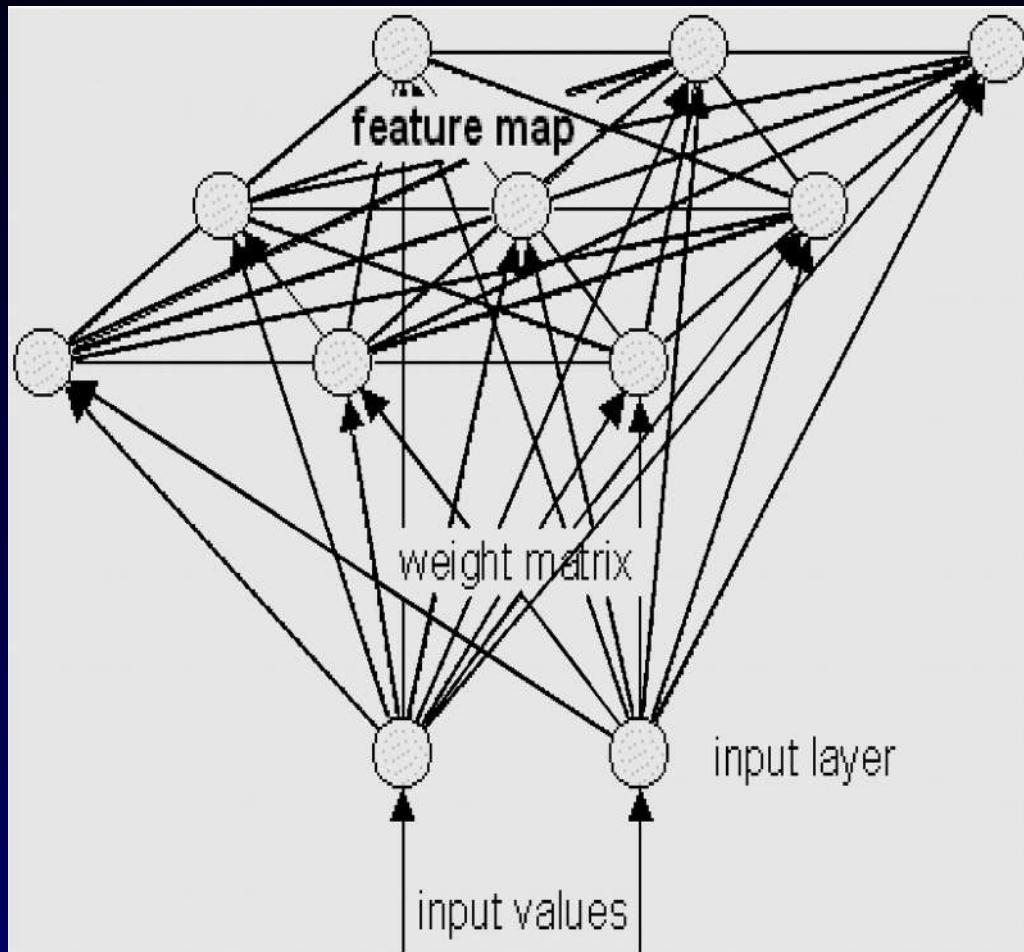
Using WWW browser - supercomputing in PDA/mobil

Machine Learning of BIG Archive



Principles of SOM

Self-Organizing = Kohonen map



Association (activation) map

How many vectors activate every neuron

Unified Distance Matrix (U-matrix)

Every neuron= sum of distances to neighbours

The higher = more unique (outlier)

Machine Learning of Spectra

SW view

ML does not produce new data – same spectra in groups

Results the same size as input (+ small overhead)

Tracing visual shape from ML results

Solf-Organizing maps – finding outliers

Easy trace shape from neuron - clickable maps

Visualisation of many spectra in web – dygraph (JS)

Virtual Observatory inside

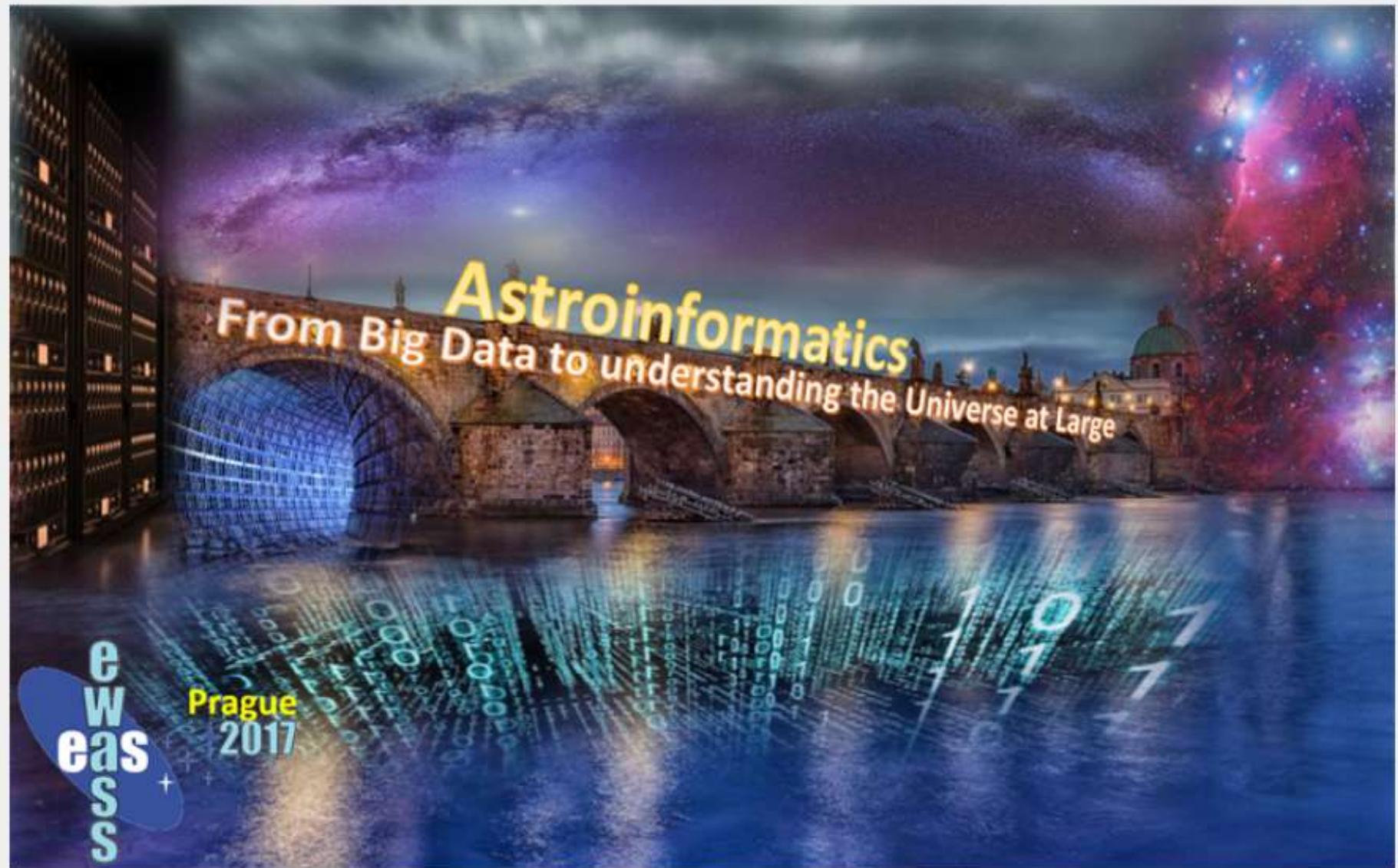
- OND 2m archive on SSAP protocol (spectra access)
 - LAMOST DR1 on SSAP (using DaCHS)
 - Preprocessing (rectify, cutout) – DataLink on server
 - SAMP (send spectra to SPLAT-VO - view details)
 - Visualization of results
-
- VO-CLOUD – cloud engine based on UWS REST jobs
 - Cross-matching (ADQL, TAP, TOPCAT, TAPHandle, pyVO, Vizier)

Conclusions

- Machine learning on big spectra archives may identify new interesting objects yet unknown
- Crucial is interactive visualization of candidates
- VO technology helps in every step
- Future astronomy will be multidisciplinary
- Wide collaboration of experts and informaticians

Symposium S14

29 – 30 June 2017



<http://eas.unige.ch/EWASS2017/session.jsp?id=S14>

DEMO - create job

vo-cloud Create new SOM job - Iceweasel

vo-cloud Create new SO... +
vocloud-dev.asu.cas.cz/vocloud/jobs/index.xhtml
Most Visited ▾ Getting Started Connecting...

VO-CLOUD CREATE NEW SOM JOB

Home Jobs Create Settings Admin Help Logout (skoda)

Project label: spectra4
Description: SOM on spectra labeled in 4 classes
 Email me results

Edit config.json

```
{
  "Name": "Stellar_spectra",
  "Algorithm": {
    "Bmu": "normal",
    "Threads": 1
  },
  "Data": {
    "Path": ["spectra.1863.4"],
    "File_type": "csv"
  }
}
```

Upload parameters

Please attach data with config.json file.

(c) mrq 2014 - [feedback](#)

DEMO - Job is running

vo-cloud Jobs - Iceweasel

vocloud-dev.asu.cas.cz/vocloud/jobs/index.xhtml

Most Visited ▾ Getting Started Connecting...

VO-CLOUD JOBS

Home Jobs Create Settings Admin Help Logout (skoda)

Type	Id	Name	Created	Duration	Phase	Action	Delete	Details
SOM	8603	spectra4	10/8/14	17 sec	EXECUTING	abort	x	d
SOM	8555	spectra4 (copy)	10/8/14	14 sec	COMPLETED		x	d
SOM	8550	spectra5	10/7/14	119 sec	COMPLETED		x	d
SOM	8549	spectra4	10/7/14	62 sec	COMPLETED		x	d
SOM	8548	iris	10/7/14	0 sec	COMPLETED		x	d
SOM	8547	ecoli	10/7/14	3 sec	COMPLETED		x	d
SOM	8537	spectra4_unspec	10/2/14	89 sec	COMPLETED		x	d
SOM	8536	spectra4	10/2/14	108 sec	COMPLETED		x	d
SOM	8534	new test of spectra (copy) (copy)	9/26/14	10 sec	COMPLETED		x	d
SOM	8533	new test of spectra (copy)	9/26/14	0 sec	PENDING	start	x	d
Korel	8530	testkorel (copy) (copy)	9/26/14	0 sec	COMPLETED		x	d
SOM	8520	new test of spectra	9/26/14	12 sec	COMPLETED		x	d
Korel	7602	big job with map (copy)	4/14/13	37 sec	COMPLETED		x	d

39% full (Using 396.4 MB / 1.0 GB)

(c) mrq 2014 - [feedback](#)

DEMO

