



Astronomical Institute  
of the Czech Academy of Sciences



# Challenges of Astronomical Spectroscopy in the AI Era

Petr Škoda

Astronomical Institute of the Czech Academy of Sciences, Ondřejov

INAF Invited Seminar  
INAF, Trieste, Italy  
May 13th 2026

## Credits

The presentation is based on many different sources – mainly the on-line published slides from IVOA meetings, EPRV group slides and docs, ADASS meetings, tutorials, manuals, pictures found on Internet and materials of private communication

# Outline of the Talk

- Data Avalanche in astronomy
- Spectroscopy – spectra surveys
- Virtual Observatory
- Astroinformatics (AI)
  - Visualizations
  - Transfer of technology
- Artificial Intelligence (AI)
- Future?

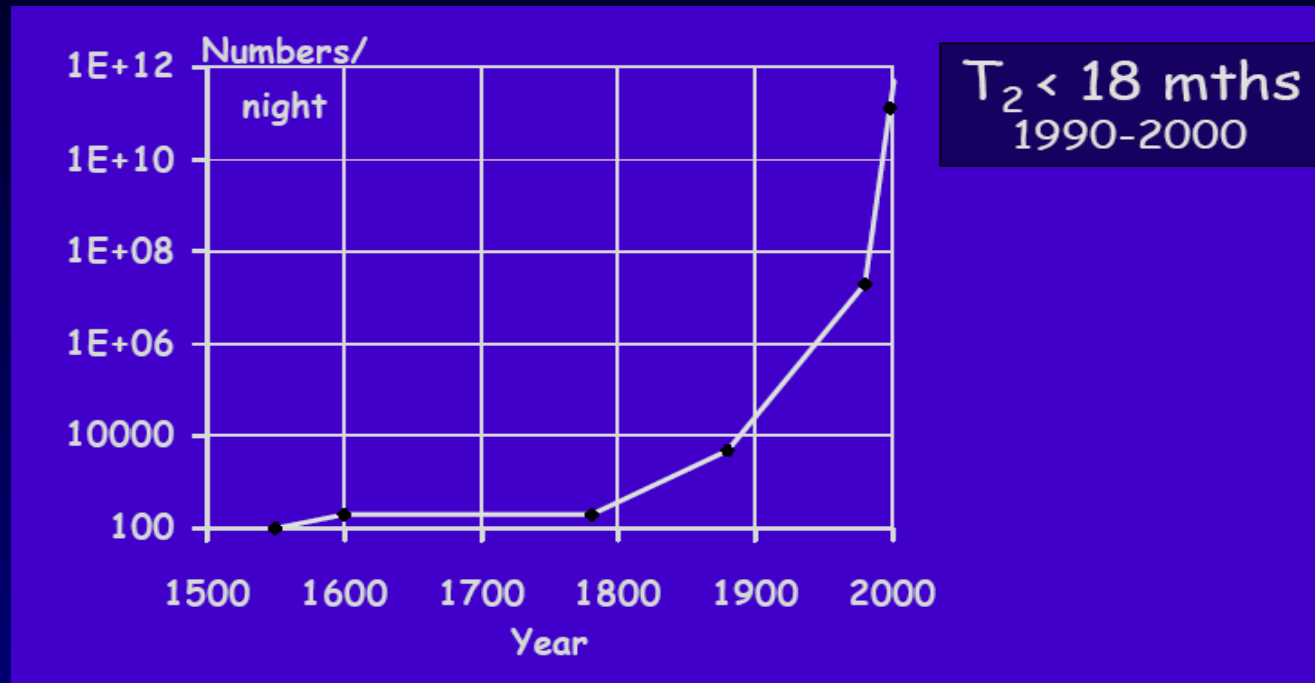
***Astronomically Big  
Big Data***

# Data Avalanche

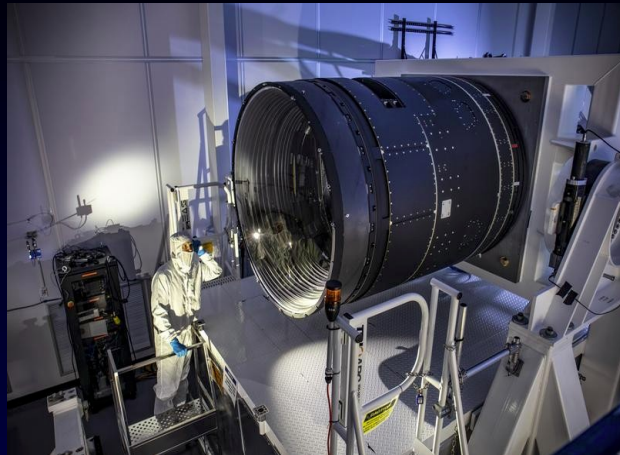
Moore law for chips –doubling 1.5 year

Data in astronomy – doubling < 1 yr !

100 PB today, 100 TB/night



# LSST – Vera C. Rubin Observatory



189 CCD 4kx4k, 10um  
3.2 Gpix every 15 sec  
3.5 deg FOV (64cm)  
15 TB/day=6 PB/yr RAW  
15 PB catalogue (D11)  
detection of changes 60s!  
7 million alerts/night ! Tot 20B  
38 billion objects x 800  
32 tril. meas. -5 PB table



# LSST – Vera C. Rubin Observatory



First light June 2025

Virgo cluster in Rubin

15 sq deg

1185 exposures / 7 nights

# LOFAR network



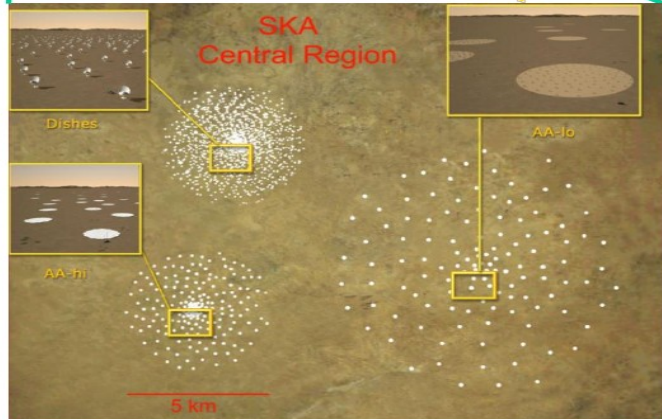
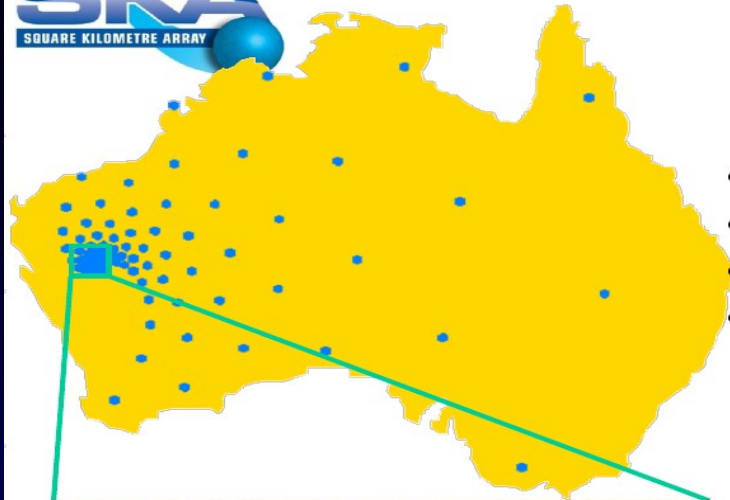
	LOFAR	SKA
Raw Telescope	112 PB/yr	60 EB/yr
Archive Rate	6 PB/yr	100 PB/yr

# 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



**Phased Aperture array**

# SKA



**Dishes**

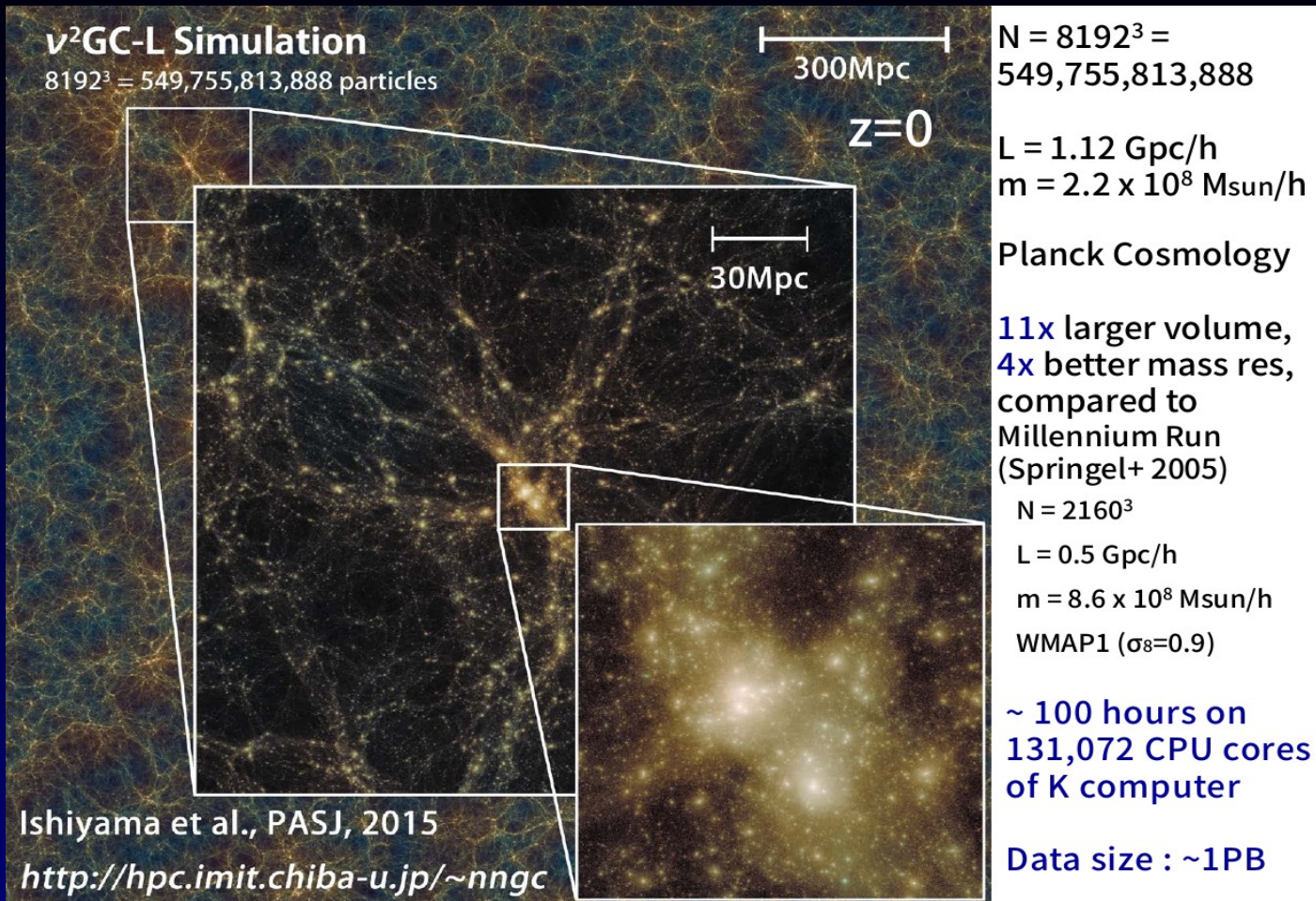
# SKA Archive Volumes

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

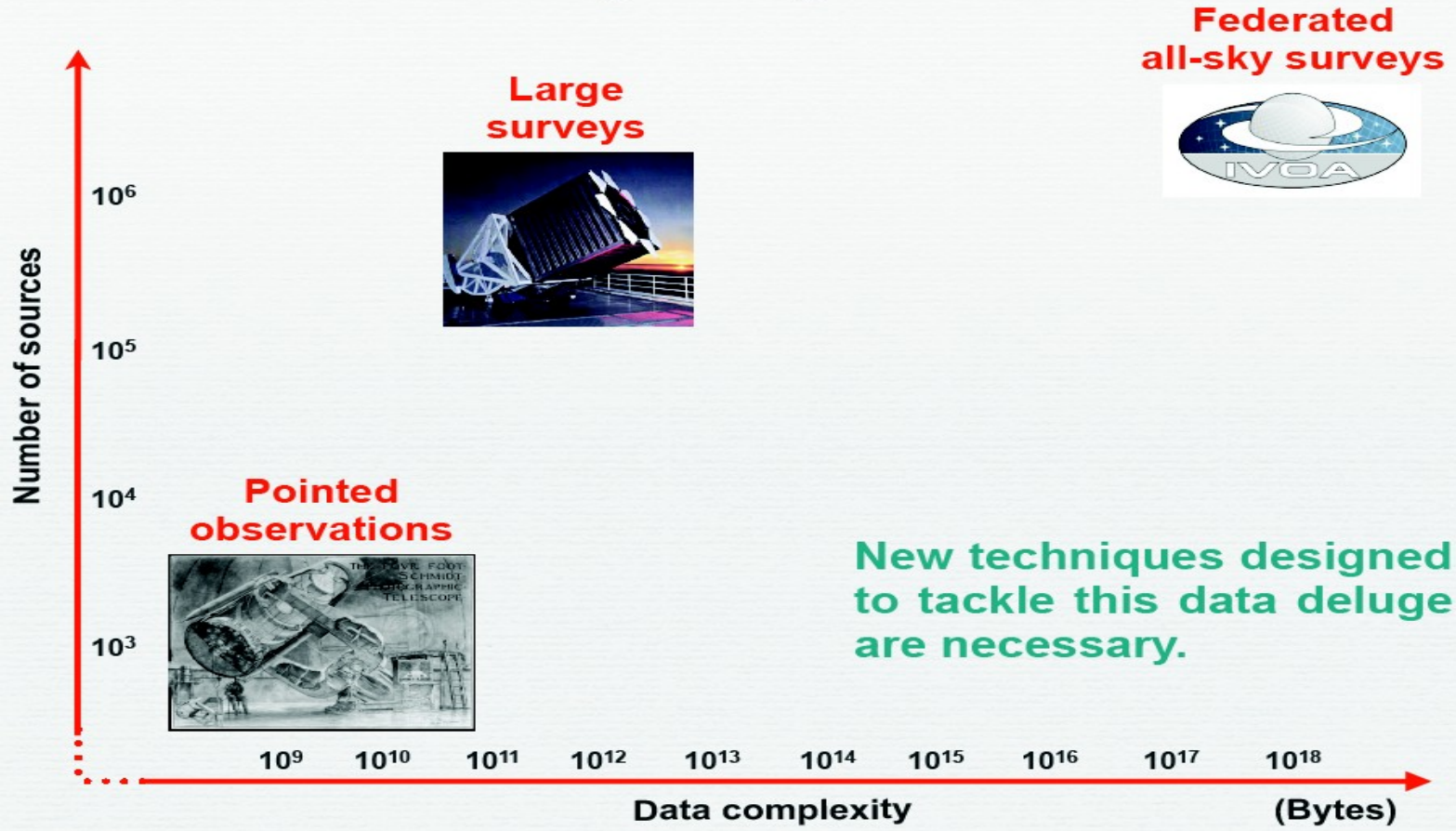
**100 Pbytes – 3 EBytes / year of fully processed data**

700PB/year 2x135 Pflops 20 Tbit/s transfer link

# Simulation of Universe



# A paradigm shift



Data analysis at storage place  
Move processing = not data !

# *Large Spectra Surveys*

# Astronomical Spectroscopy - History

- I. Newton 1666 – Solar spectrum, prism
- W. Herschel 1800 – Calorific rays (IR)
- J. Fraunhofer 1814 – Solar absorption lines – letters H,K,D,A,B, Sirius
- Ch. Doppler 1842– Royal Bohemian Society of Sciences, Colors of Stars
- W. Huggins 1864 – Physical nature of stars (elements), nebulae (cats Eye)
  - 1866 T CrB, transients, evolution
  - 1868 RV of Sirius
- H. Draper 1872 – Photographic spectrum of Vega
- E. Hubble 1929 – Cosmological redshift (Lemaitre 1927)
- M. Schmidt 1963 – QSO 3C 273 ( $z=0.158$ , strange elements?)

# Information From Spectral Lines

## Position (wavelength):

Chemical elements , Excitation / Ionization state

RV (binarity, orbital parameters...)

## Shape

Stellar parameters (Teff, log g, rotation)

Stellar activity (Turbulence, granulation)

Expansion, shells, winds (P Cyg, Novae)

## Time variability (LPV)

Change of physical state (Be, outburst)

Spots (Mg field, overabundance – Ap stars)

Pulsations (Delta Ceph, RR Lyr, Miras)

Non radial pulsations (NRP)

Multiple systems – disentangling of orbital parameters

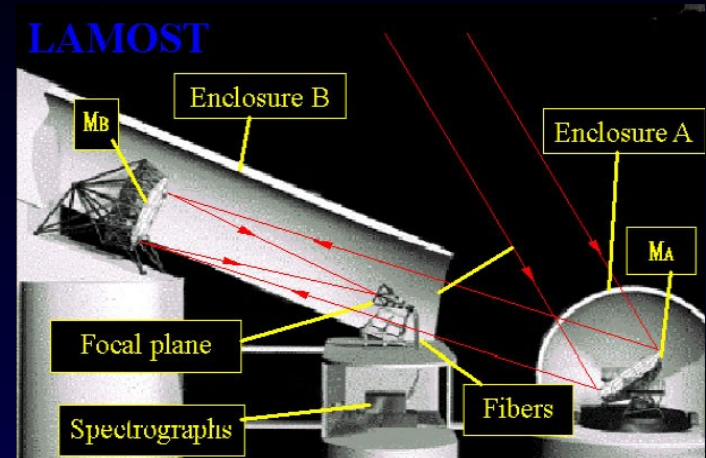
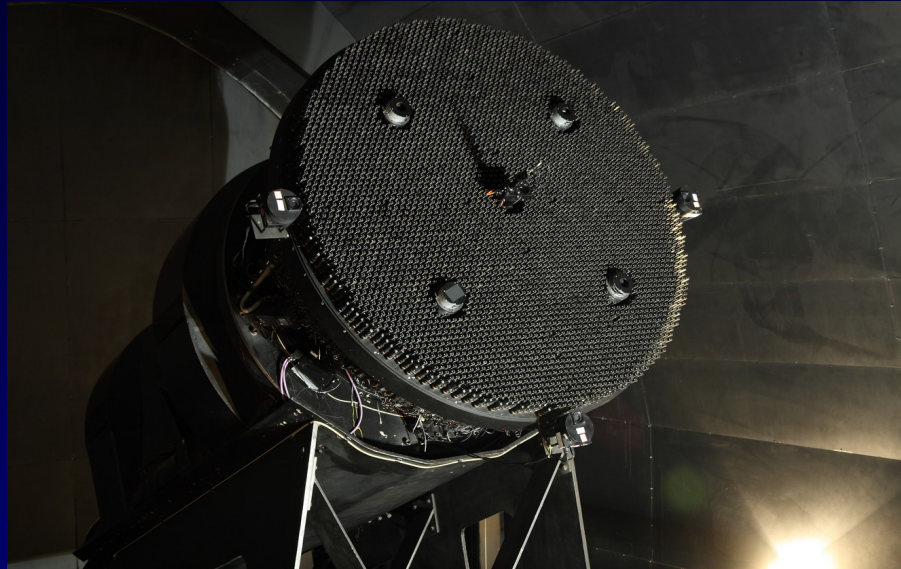
Detection of ES planets in spectra

# Sloan Digital Sky Survey (SDSS)

	optical spectra	infrared spectra	IFU cubes/tiles	TOTAL
SDSS-I/II	1,843,200	-	-	1,843,200
SDSS-III/IV	3,958,000	3,390,236	11,273	7,359,509
SDSS-V	4,182,779	1,223,269	1(*)	5,406,048
<b>TOTAL</b>	<b>9,983,979</b>	<b>4,613,505</b>	<b>11,273</b>	<b>14,608,757</b>

# LAMOST (Guoshoujing)

- Xinglong, China
- 4 m mirror (30 deg meridian)
- 4000 fibres

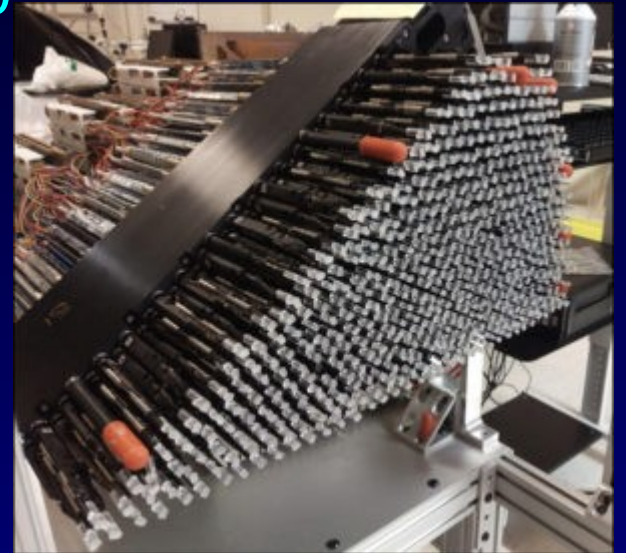




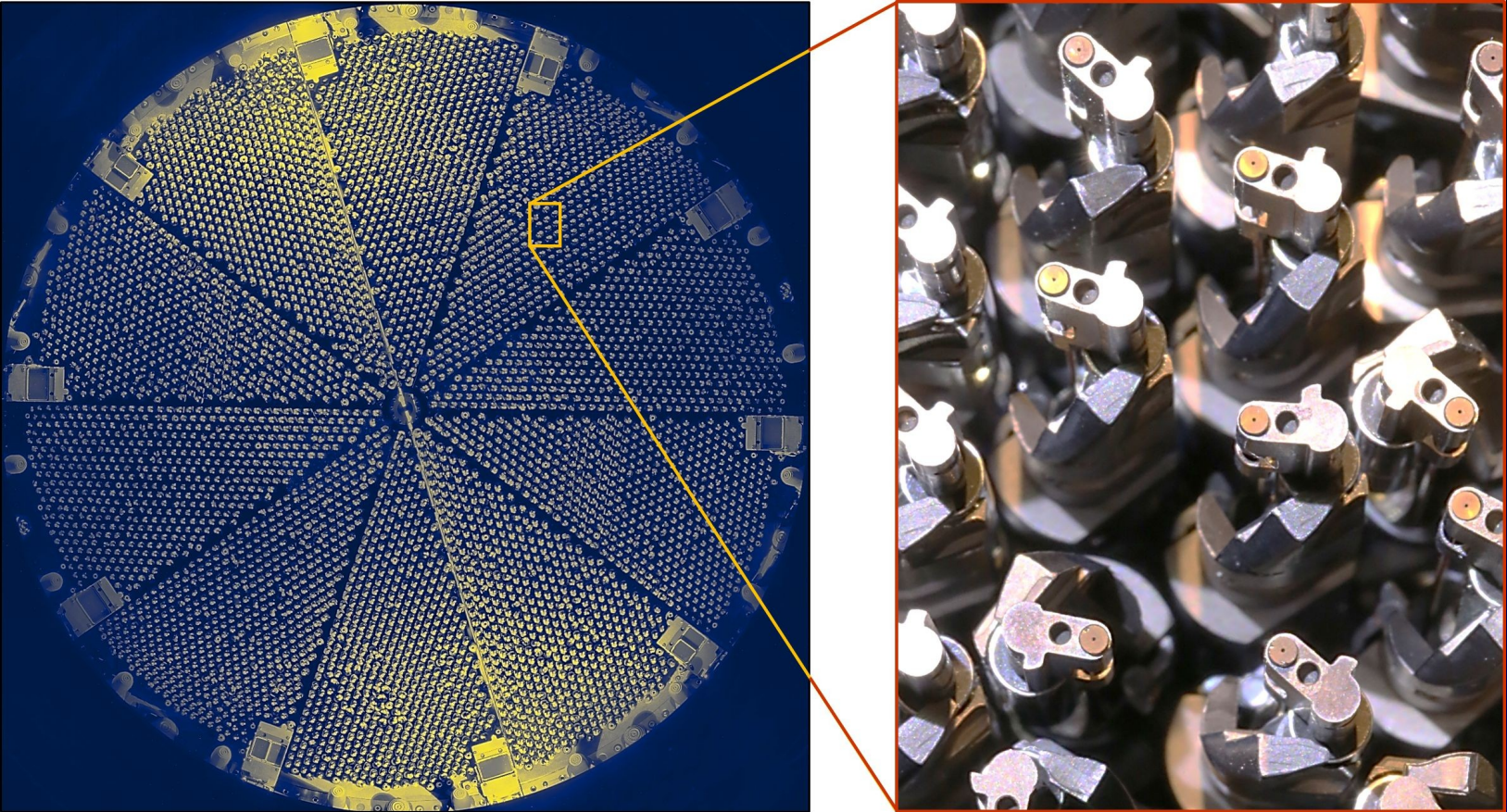
# DESI (Dark Energy Spectroscopic Instrument)

- Kitt Peak , Arizona
- 4 m mirror (Mayal)
- 5000 fibres 10 petals x 500 fibers, 10 spg
- 3 cameras 3600-9800A R~2000-5000

<https://arxiv.org/pdf/2503.14745>



# DESI



[https://www.desi.lbl.gov/wp-content/uploads/sites/8/2022/06/DESI\\_robots.jpg](https://www.desi.lbl.gov/wp-content/uploads/sites/8/2022/06/DESI_robots.jpg)

# DESI Spectral Surveys

DR1 (Mar 2025) May 2021-June 2022 13 month

**18.7 mil** objects

13.1 mil galaxies

1.6 mil QSO

4.0 mil stars

Plan 5 year (~2026 end)

**50 mil** galaxies and QSOs

**25 mil** MW stars

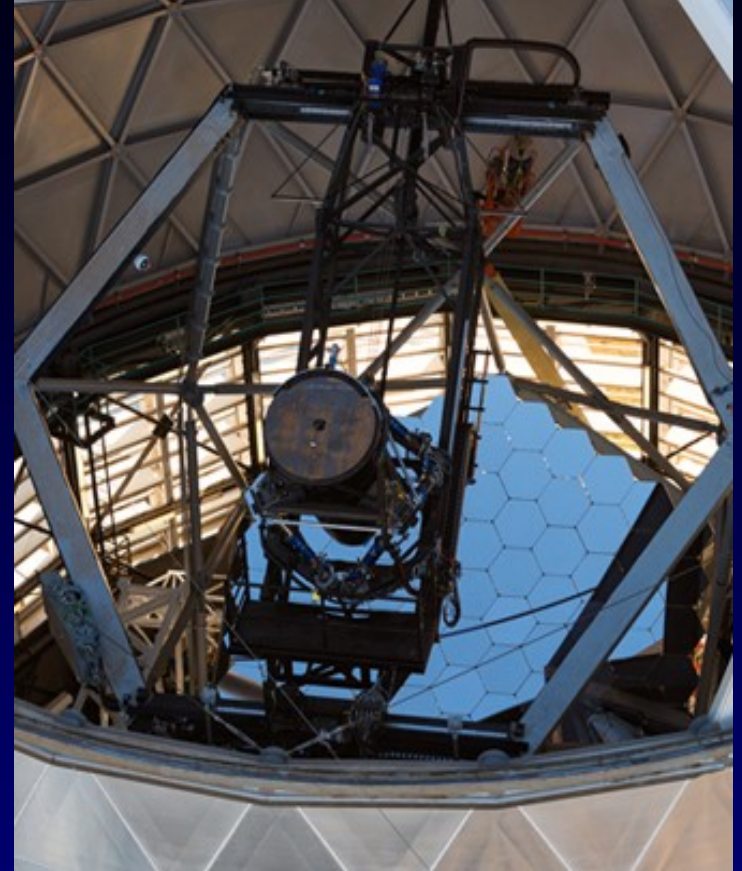
# Hobby Eberly Telescope (HET)



Mc Donald Observatory Texas

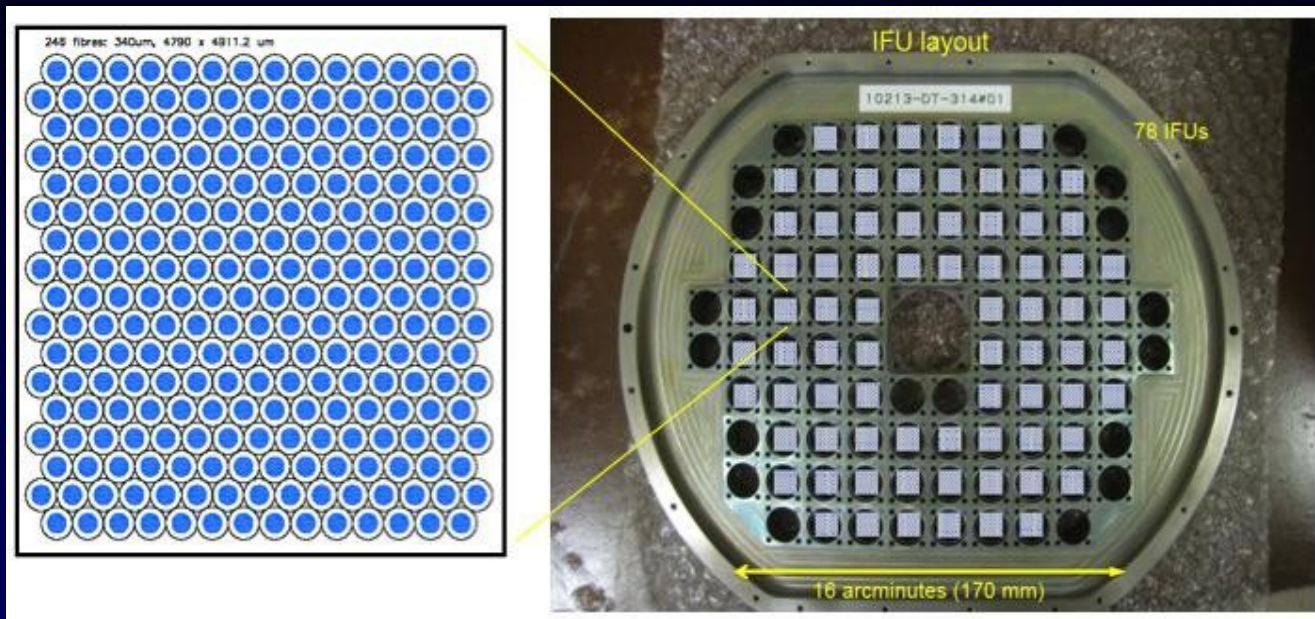
Equiv diameter 9.5m (11m)

Fixed in position during observation -  
only primary tracker



# HETDEX Survey

PDR1 – mid 2026, 300mil spectra plan 1 billion



VIRUS 78 IFU = 156 spectrographs

IFU= 448 fibers

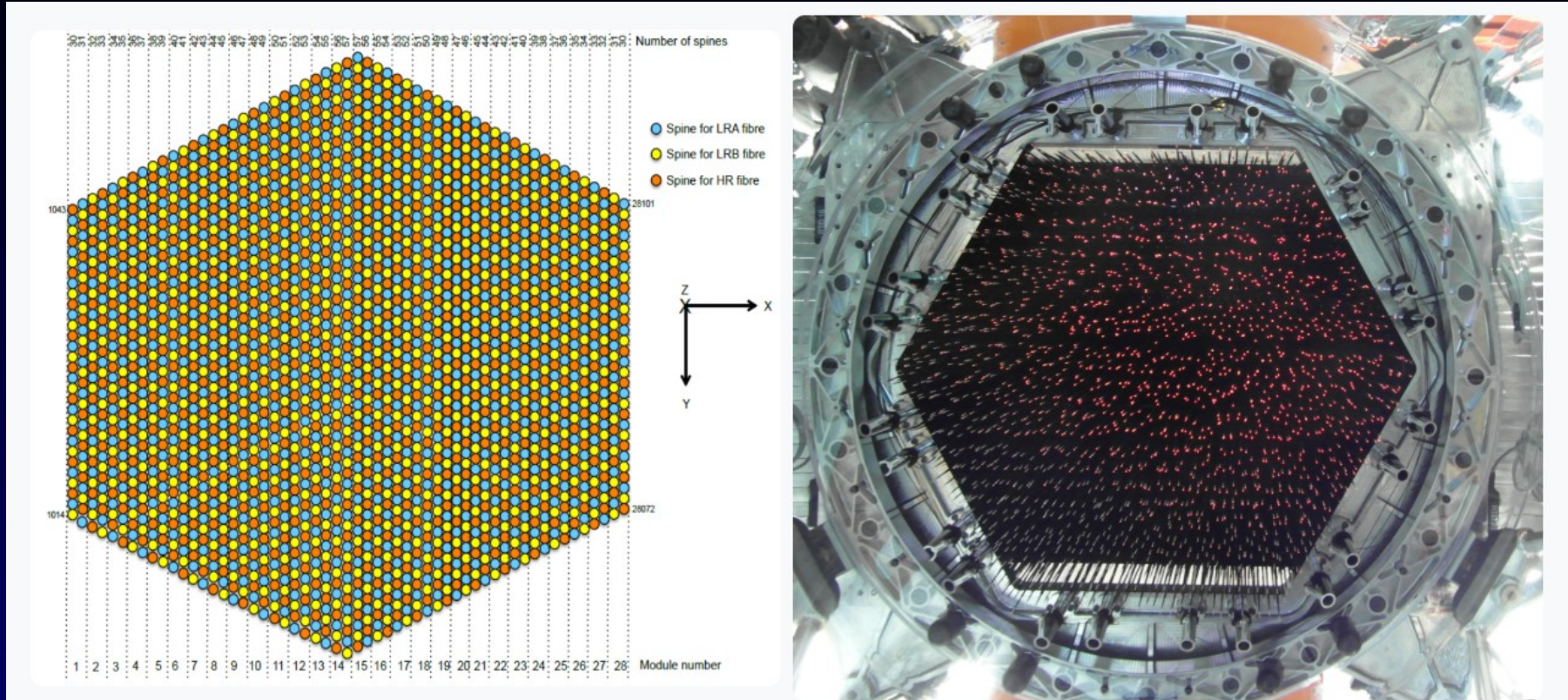
34944 fibers , FOV 22 arcmin, 3500-5500 A, R=800

# 4MOST

- March 2026 ended verification phase, start soon regular
- Paranal, Chile, driven by MPI Garching
- 4 m mirror VISTA
- 2436 fibres (2x812+812)
- 3 spectrographs
- Low resolution 2 spg 3700-9500Å R~4000-7000
- High resolution 3926-6790Å, R~18000-21000
- Plan 20mil sources low res, 3mil high res

<https://www.4most.eu/>

# 4MOST



<https://www.4most.eu/>

# Gaia

## SKY-SCANNING COMPLETE FOR ESA'S MILKY WAY MAPPER GAIA

From 24 July 2014 to 15 January 2025, Gaia made more than three trillion observations of two billion stars and other objects, which revolutionised the view of our home galaxy and cosmic neighbourhood.

**3 TRILLION**  
Observations

**2 BILLION**  
Stars & other objects observed

**938 MILLION**  
Camera pixels on board

**15 300**  
Spacecraft 'pirouettes'

**55 KG**  
Cold nitrogen gas consumed

**3827**  
Days in science operations

**50 000 HOURS**  
Ground station time used

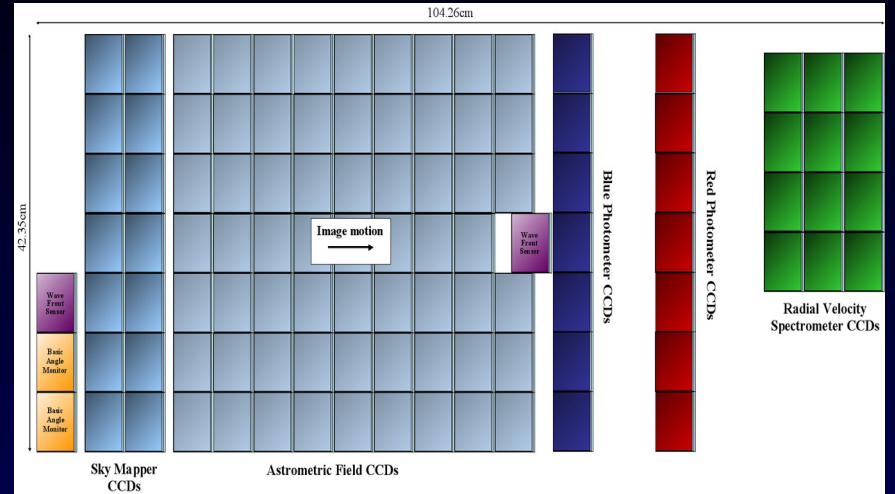
**580 MILLION**  
Accesses of Gaia catalogue so far

**13 000**  
Refereed scientific publications so far

**2.8 MILLION**  
Commands sent to spacecraft

**142 TB**  
Downlinked data (compressed)

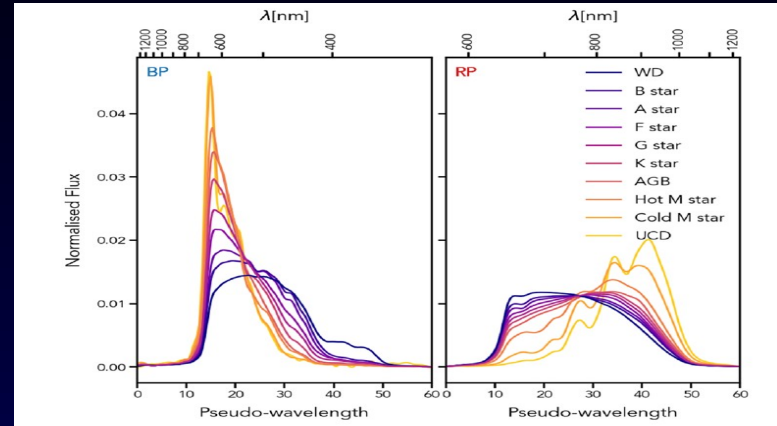
**500 TB**  
Volume of data release 4  
(5.5 years of observations)



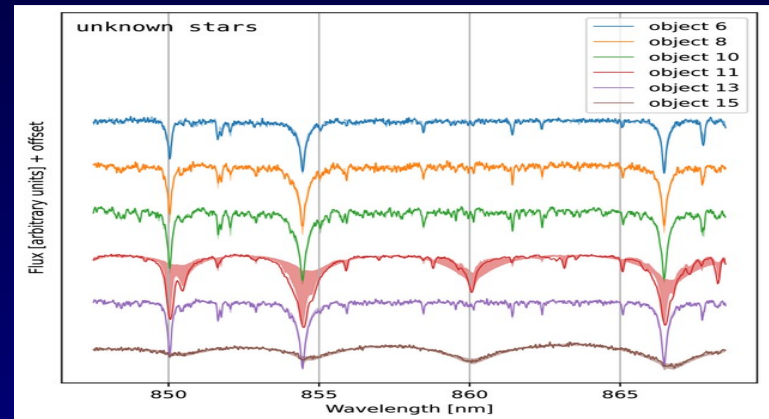
Source: ESA

# Gaia Spectrographs

XP (BP+RP) R~30-100 3300-10500 A



RVS R~11500 8450-8740 A



# Gaia Spectra Collections

- DR3 (June 2022)
  - 220 mil XP (Gauss-Hermite polynomials, GAVO SSAP)
  - 1 mil RVS
- DR4 (Dec 2026) tens of millions RVS
- DR5 planned (> 2030)
  - 150 mil RVS

# Euclid

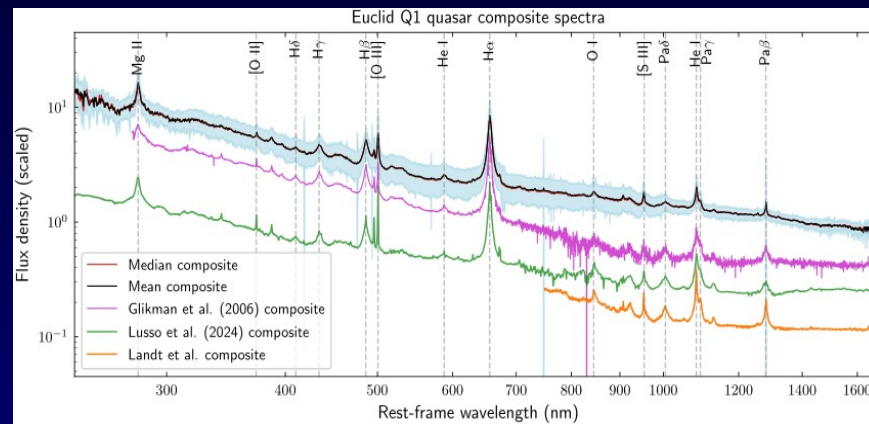
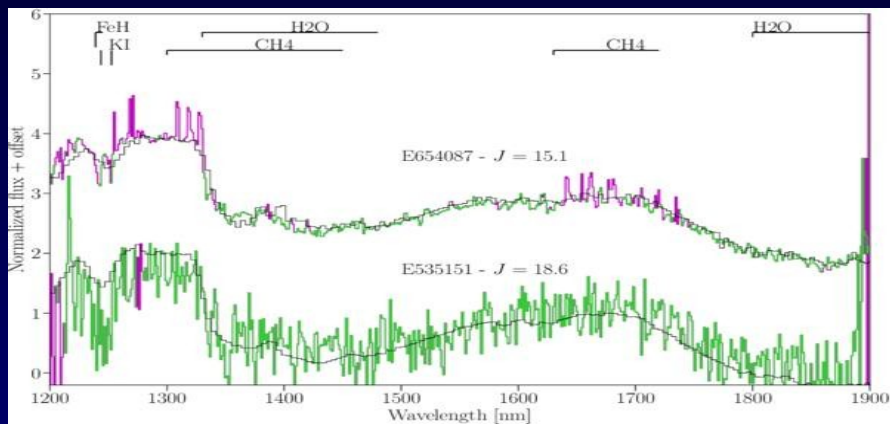


1.2m Korsch tel.  
Dark matter  
10B sources  
1B weak lensing  
600 MP camera 7filt  
Spectra VIS/NIR  
DR1 on June 2026  
100 PB processed  
26PB/year



# Euclid Spectra

- NISP 1100-2000Å  $R \sim 250$
- DR Q1 **3 mil** spectra
- Planned **35 mil** spectra (redshifts)



# *Virtual Observatory*

# IVOA (established 2002)



## 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.

# Ecosystem of VO

LEVEL 1  
empty

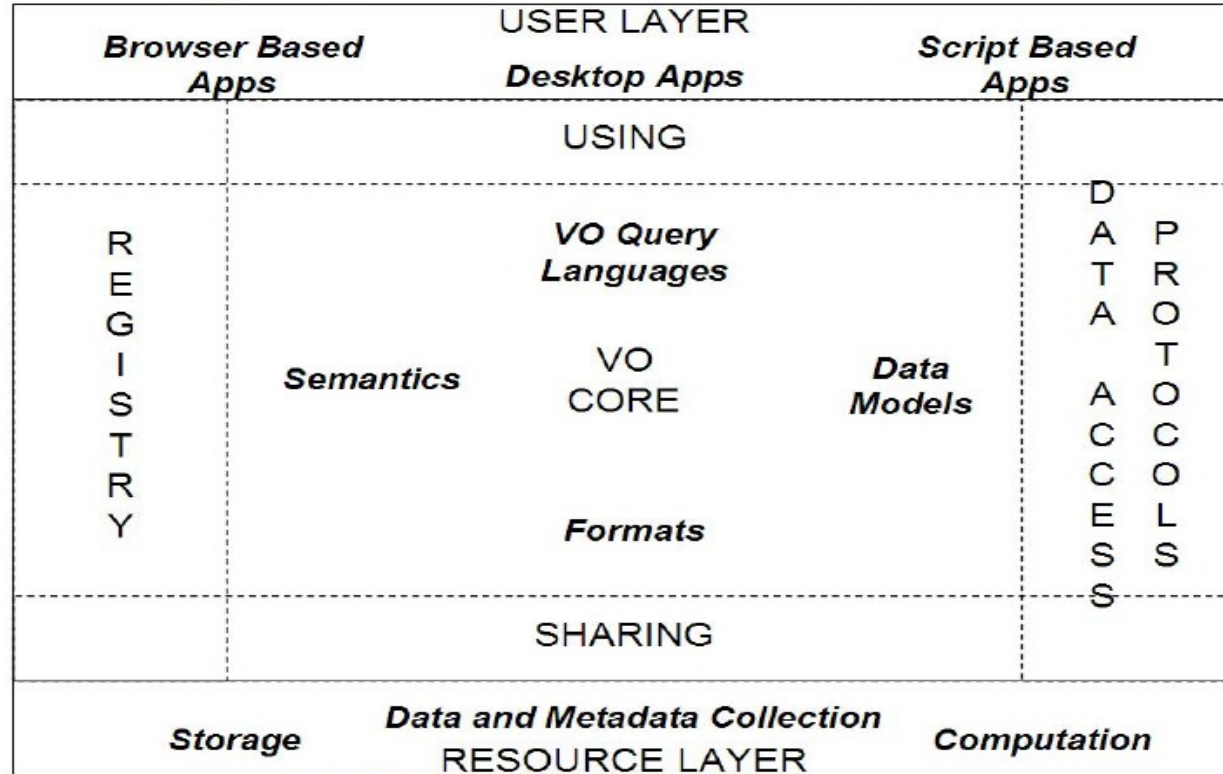
USERS



COMPUTERS

REC

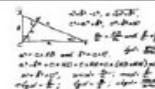
InProgress



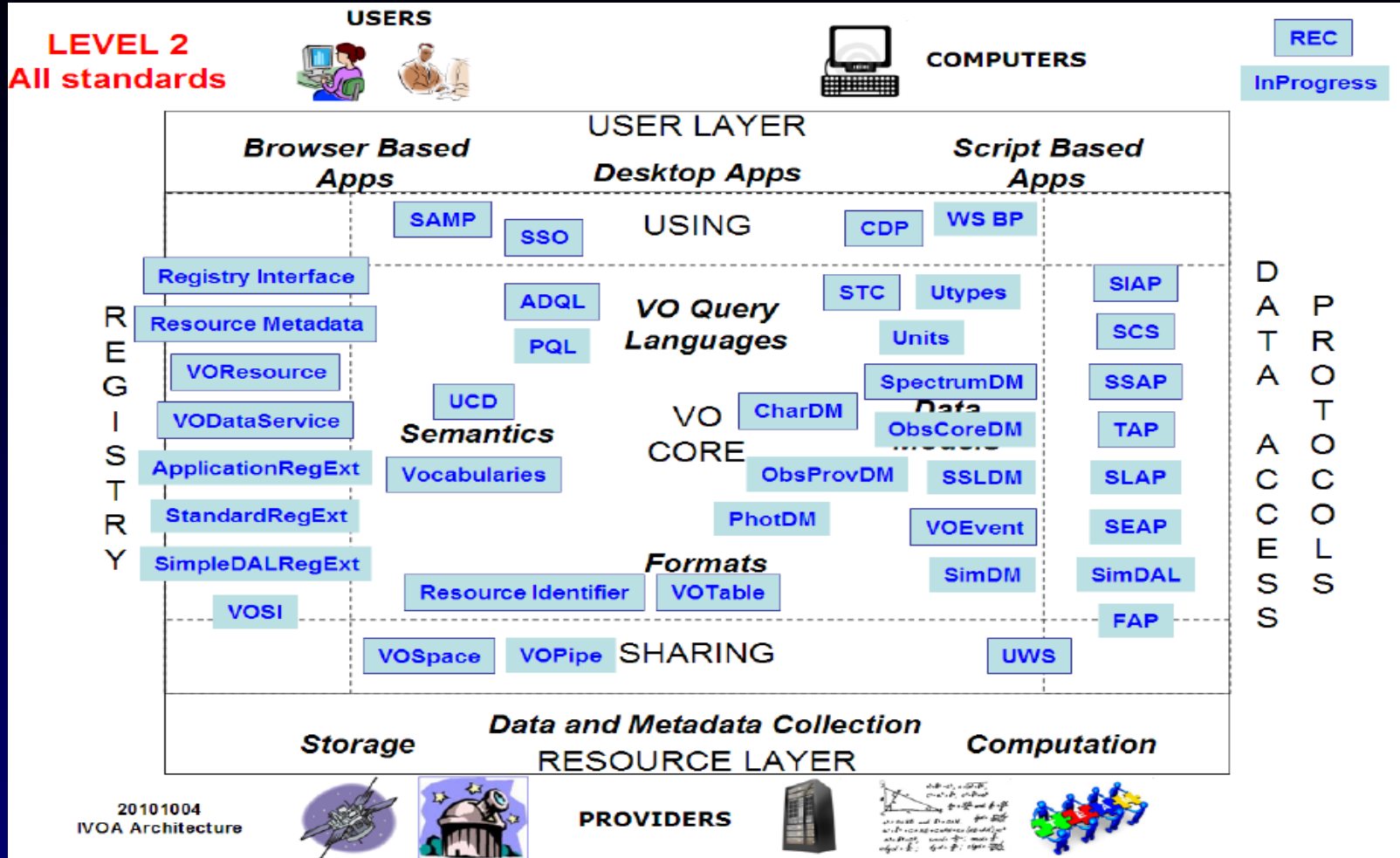
20101004  
IVOA Architecture



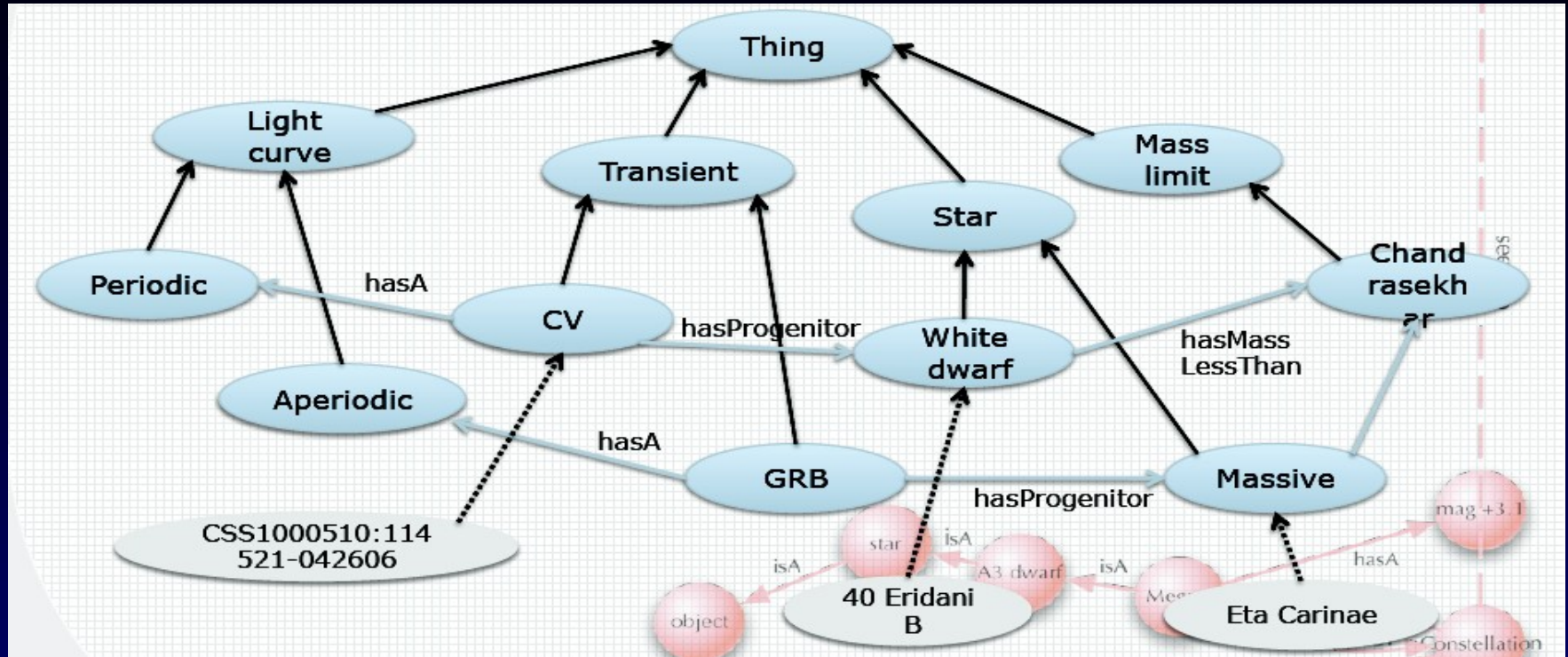
PROVIDERS



# Ecosystem of VO - level 2



# Ontologies in Astronomy



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

From Graham, M. AI2010

# 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)

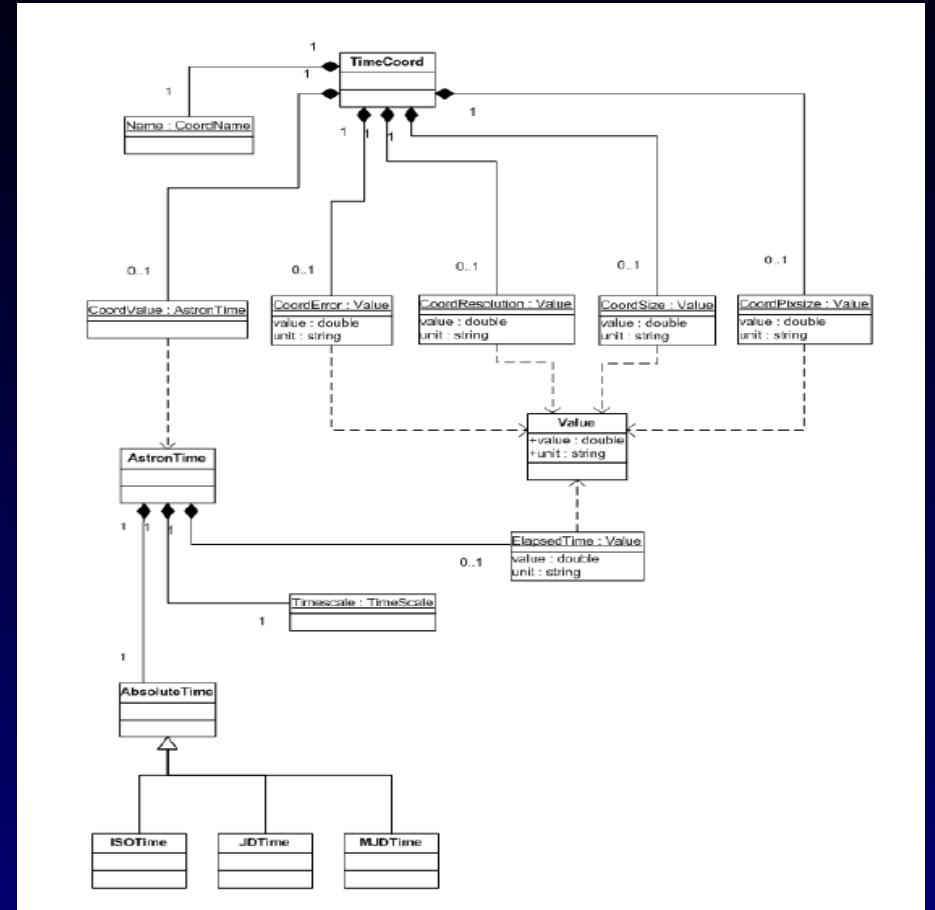
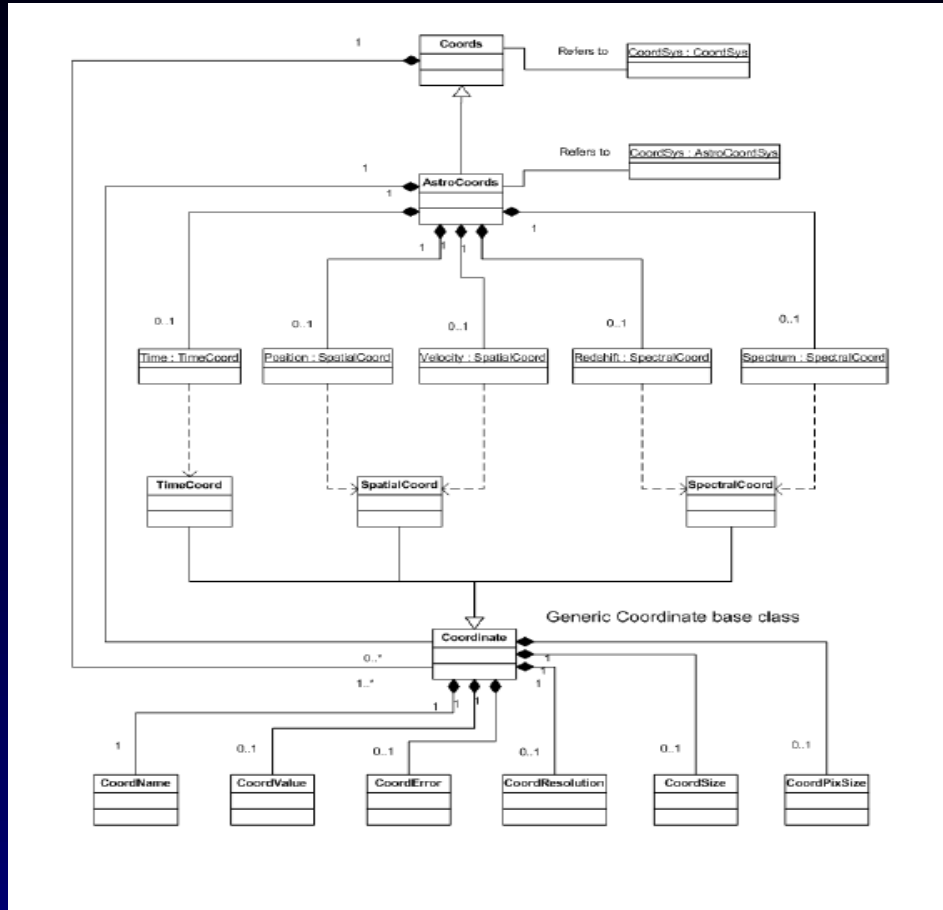
# Universal Content Descriptors

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

# Space-Time-Coordinate Data Model



# Technology of VO

Unified data format– **VOTable**, **UCD** (Vizier)

Transparent transport (**VOunits**)

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) - VAMDC

**TAP** (Table Access Protocol) – query e.g. whole SDSS

**VOEVENT** (transients, robotic telescopes, Sun

**DATALINK** (related data products, e.g. raw, mosaics..)

**SODA** Server-side Operations for Data Acces

# Technology of VO

**ADQL** (Astronomical Data Query Language)

**XMATCH, REGION** (2 catalogues – shifted)

Application interoperability – **SAMP**

Allows develop applications as bricks

sending **VOTABLES** (catalogue-spectra-images)

Surveys visualization

**HIPS** (Hierarchical Progressive Survey) - allsky zoom

**MOC** (Multi order coverages) time, space, spectral (FoV)

# Analysis





# Aladin

Aladin v9.6 \*\*\* PROTOTYPE VERSION (based on v9.623) \*\*\*

File Edit Image Catalog Overlay Coverage Tool View Interop Help

Data access → 24 / 19697      Location 05:43:27.24 -01:54:27.1      Frame ICRS      Projection Aitoff

▼ Collections → 24 / 19697  
  ▼ Data base → 1 / 2  
    SIMBAD Astronomical Database  
  ▼ Catalog → 23 / 17224  
    IX+High-Energy data → 1 / 69  
      ROSAT All-Sky Bright Source  
      IRXS Correlation tool  
  ▼ Journal table → 22 / 15736  
    A+A → 11 / 4151  
      Field star positions (from  
      Cluster positions and  
      X-ray sources detect  
      X-ray sources detect  
      NGC 6192 star positions  
      Catalog of objects from  
      Star positions (from  
      The cross-identification  
      Simbad position for a  
      Sources extracted with  
      Additional compact sources  
    A+AS → 2 / 723  
      Source positions (from  
      Summary of stars with  
    A1 → 2 / 2615  
      EBs that appear in  
      Cluster positions (from  
    Ap1 → 1 / 3206  
      Simbad positions for  
    Ap1S → 4 / 1429  
      The CVs and WDs and  
      RASS/SIMBAD class  
      SIMBAD stars with A  
      SNe positions from SIMBAD  
    MNRAS → 1 / 2549  
      Cluster positions (from  
    PASP → 1 / 141

select simbad  
from -- All collections --

filter coll inside zoom

▼ DSS ★ SDSS ★ 2MASS ★ WISE ★ GALEX ★ PLANCK ★ AKARI ★ XMM ★ Fermi ★ Gaia ★ Simbad ★ NED +

CDS/P/DSS2/color

15'      1.522" x 1.20"

select  
pan  
dist  
phot  
draw  
tag  
merge  
split  
filter  
cross  
copy  
epoch  
size  
rgb  
opac  
zoom  
crop  
cont  
pixel  
prop  
del

Stack controls:  
• the icon: show/hide a plane  
• size: change object size  
• zoom: adjust field size.  
• Opacity: adjust transparency.

The view is drawn according to the projection of a reference plane.  
For changing the reference, click on its check box.

filter CDS/Simbad  
 CDS/P/DSS2/color

epoch -  
size -  
rgb -  
opac -  
zoom -

MAIN_ID	OTYPE	RA	DEC	COO...	COO...	C...	PMR
NAME Flame Nebula	MoCl10	05 41 42.7	-01 54 44				
* zet Dri A	**	05 40 45.527	-01 56 33.26	1	1	90	4
NGC 2023	RfNeb	05 41 37.9	-02 15 52	3200...	7000...	51	
* zet Dri	**	05 40 45.52...	-01 56 33.2...	5.189	2.289	90	3
HD 37903	Ea*	05 41 38.38...	-02 15 32.4...	7.651	3.437	90	-1
* zet Dri B	Star	05 40 45.571	-01 56 35.59	8	5	90	4
HD 38087	**	05 43 00.57...	-02 18 45.3...	11.084	5.986	90	
DMCERR1 NGC 2024 FIR 5	dens	05 41 44.6	-01 55 38				

# VO Science Portals

Stellarium + VirGo (ESO, unsupported)

ESASky

<https://sky.esa.int/>

ESO Archive Science Portal

<https://archive.eso.org/scienceportal/home>

IRSA IPAC archive (Firefly)

<https://exoplanetarchive.ipac.caltech.edu/firefly/>

WWT (original MS, now AAS, web client)

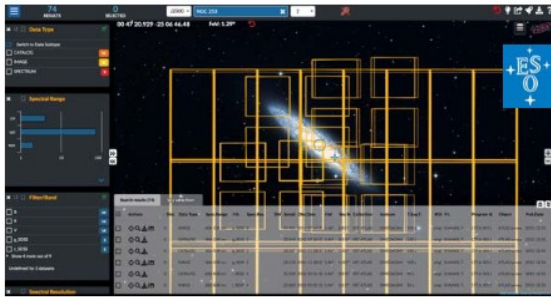
<http://worldwidetelescope.org/webclient>

GoogleSky

<https://www.google.com/sky/>

# VO Science Portals

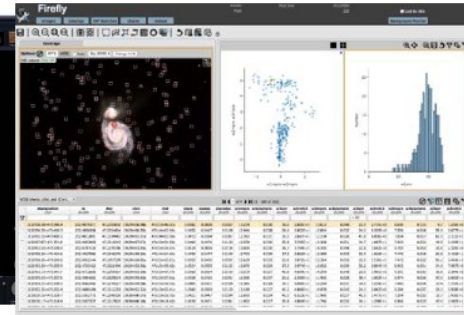
## VO embedded in astronomy services



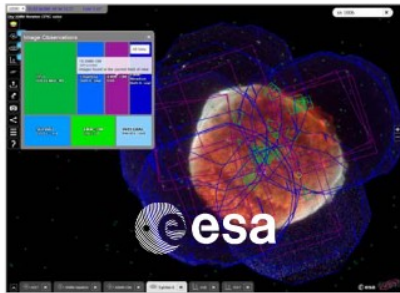
ESO Science Portal



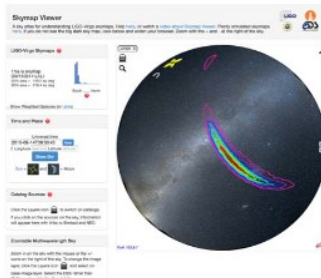
WWT



Firefly  
Caltech-IPAC



ESA Sky



Gravwaves



CDS reference data service

SVO Filter Profile service

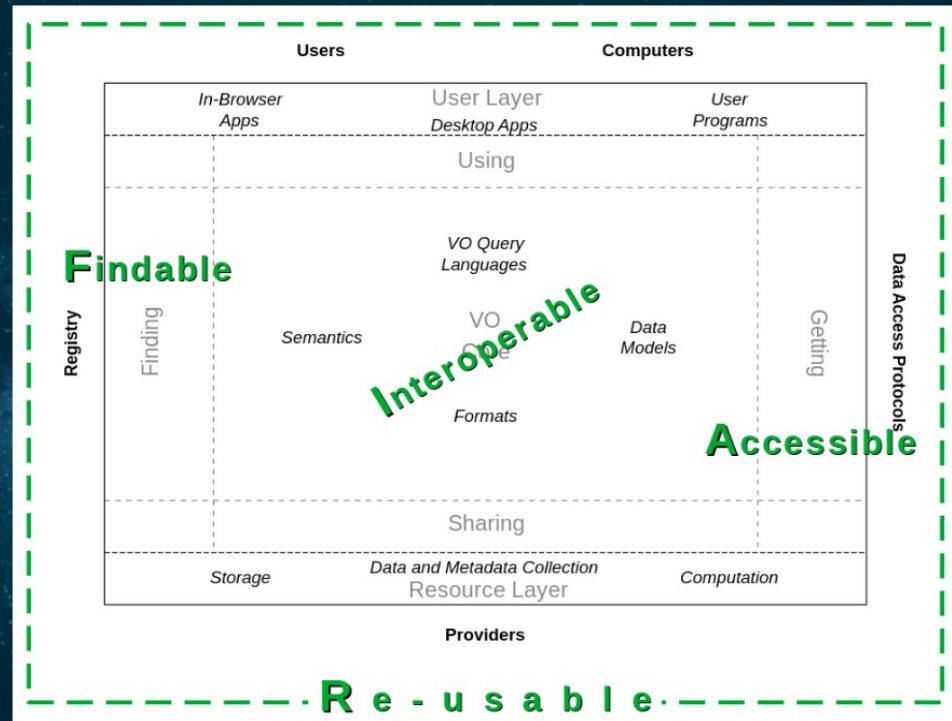
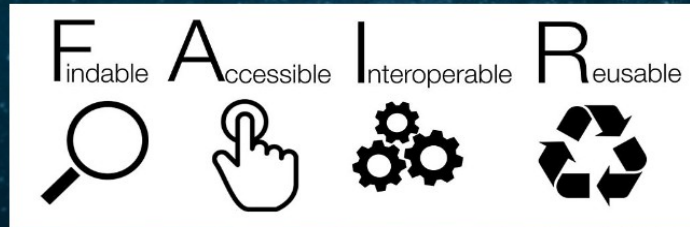
# Open Science – EOSC

EURO-VO DCA, ICE, CoSADIE, ASTERICS, ESCAPE - Astroparticles

## IVOA Architecture – FAIR data management



VO has been FAIR from the beginning!



# 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)
- SIM-DB        Simulations, theory data

Science platforms    for **BD analysis and ML**

*(SciServer JHU, NOAO DataLab, CANFAR, Gaia, Euclid , Rubin x Pangeo )*

# EUROPLANET VESPA (EPN-TAP)

Aladin v10.0 \*\*\* BETA VERSION (based on v10.041) \*\*\*

File Edit Image Catalog Overlay Coverage Tool View Interop Help

Command 16:12:32.01 -50:02:48.5

Frame ICRS Projection Spheric

ALADIN

Welcome to Aladin,  
your professional sky atlas.

- Discover all astronomical data available over the net!
- Compare them with your own data.
- Prepare your observation missions.

To start, type any object name, such as M1, and press ENTER...

Or easier, clic in the main frame and enjoy the sky...

select  
pan  
dist  
phot  
draw  
tag  
meas  
spect  
filter  
cross  
xy  
rgb  
zoom  
epoch  
size  
dems  
opac  
prop  
del

Available data

- isa\_zfiter.it → 4
- lap → 1
- irsa.ipac → 1
- jacobsuni → 4
  - CRISM
  - Mars\_craters
  - epn1.TAP service
  - USGS\_WMS
- jvo → 13
- latmos.ipd → 3
- imd.jussieu → 3
- madrigal.haystack.mit.edu → 1
- mosl.ud.ac.uk → 30
- nasa.heasarc → 31
- nd.org.au → 3
- oca → 1
- org.gavo.dc → 31
- purx → 2
- spectrum.laa → 3
- swinburne → 1
- tohoku.univ.jp → 12
- uni-heidelberg.de → 1
- vo-plasma.oebw.ac.at → 1
- vopdc.obspm → 14
- voxastro.org → 2
- wfau.roe.ac.uk → 2
- xaovo → 1
- xcatdb → 2

Problematic → 1

Planet → 11

- Earth → 1
  - Blue Marble next generation
- Mars → 5
  - Mars MGS MOLA Elevation Model 463m (M)
  - Mars MOLA Shaded Relief / Colorized E
  - Mars THEMIS Day IR Global Mosaic 100
  - Mars Viking-MDIM21-color
- Panorama → 1
  - Mars Stinson panorama
- Venus → 5
  - Venus Magellan C3-MOIR-2025m
  - Venus Magellan C3-MOIR-CrTopo-6600m
  - Venus Magellan MeterScaleSlope-4641m
  - Venus Magellan Microwave-Emissivity-464
  - Venus Magellan Topography-4641m

106.4° x 180°

16.08° x 24.4°

25-000004 - Click on it to get details

granule_	granule_	obs_id	dataproduct	target_n	target_c	time_min	time_max	time_sam	time_
25-000004	25	000004	c1	Mars	planet				
25-000007	25	000007	c1	Mars	planet				
25-000010	25	000010	c1	Mars	planet				
25-000012	25	000012	c1	Mars	planet				
25-000028	25	000028	c1	Mars	planet				
25-000031	25	000031	c1	Mars	planet				
25-000034	25	000034	c1	Mars	planet				
25-000039	25	000039	c1	Mars	planet				

358.91954 +49.37469 ICRS

16:03:23.95 -48:43:06.0  
14.05° x 24.41°

© 2017 Université de Strasbourg/CNRS - by CDS - Distributed under GNU OPL v3

17 sep / 2001 pro 437Mb

# VO in IAU



INTERNATIONAL  
ASTRONOMICAL  
UNION

[Home](#) | [About IAU](#) | [IAU Values](#) | [Donate](#) | [Member Directory](#) | [Site Map](#) | [Contact Us](#) | [Login](#)

News

Science

Publications

Administration

Training in  
Astronomy

Astronomy for  
Education

Astronomy for  
Development

Astronomy for  
the Public

[Home](#) / [Science](#) / [Scientific Bodies](#) / [Working Groups](#) » [Division B WG Virtual Observatory](#)

## Division B WG Virtual Observatory — Functional

### Description

This functional Working Group is intended to provide the liaison between the International Virtual Observatory Alliance (IVOA) and the International Astronomical Union (IAU). The IVOA is an organization, composed of 22 international member initiatives, that develops and maintains the technical standards needed to find, access, interoperate and re-use astronomy data (according to the FAIR principles), thus realizing the Virtual Observatory (VO). IVOA also acts as a framework for discussing and sharing VO ideas and technology, for engaging astronomy projects, missions and researchers, and as a body for promoting and publicizing the VO. The IVOA processes for the development of interoperability standards includes the possibility for IAU endorsement. Since its beginnings in 2002 the VO is now a mature framework for the interoperability of astronomy data, with IVOA compliant services operated by astronomy data centers worldwide. This functional working group will provide a communication channel between the IVOA and the IAU on topics of FAIR-compliant standards, use of the VO for scientific research, and for promoting best practices for interoperability of data and services in Astronomy.

### Links:

- [WG Annual Report \(2022\) - Virtual Observatory](#)

### Search Scientific Bodies

Follow the IAU on social media



## **(Simple) Spectra in VO**

# 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:

<i>Parameter</i>	<i>Sample value</i>	<i>Physical unit</i>	<i>Datatype</i>
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
CREATORID	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

The spatial, spectral, and time resolution of the data may all be used as query parameters.

## VO Tools

- SPLAT-VO (GAVO – Ondrejov)
- CASSIS (OMP)
- IPAC Firefly (IRSAViewer)
  - All NASA data, LSST (Rubin)
  - Spectra in echelle – started
  - <https://github.com/Caltech-IPAC/firefly>
- Not Supported:
  - SpecView (HST)
  - VOSpec (ESA)

# Spectra in SPLAT-VO - query

Starlink SPLAT-VO: Query VO for Spectra

File Options Resolver Interop Help

Service selection options  
Data Source  
 Observed data  Theoretical data

Wave Band  
 Radio  Millimeter  Infrared  
 Optical  UV  EUV  
 X-ray  Gamma-ray  ALL

Tags

Search parameters:  
**Simple Query**  
 Object: psi per    
 RA: 03:36:29.38 Dec: +48:11:33.48  
 Radius: 10.0 MAXREC:  
 Band: 6530e-10 / 6580e-10  
 Time:  
 Query Format: votable  
 Wavelength calibration: none  
 Flux calibration: none

Optional Parameters

Use	Name	Value	UCD
<input type="checkbox"/>	REDSHIFT		src.redshift
<input type="checkbox"/>	TARGETCLASS		src.class
<input type="checkbox"/>	MTIME		
<input type="checkbox"/>	SPECPR		spect.resolution;em.wl
<input type="checkbox"/>	SPATRES		pos.angResolution
<input type="checkbox"/>	PUBDID		
<input type="checkbox"/>	CREATORDID		meta.id

Query: <SERVER>?REQUEST=queryData&POS=54.122416666666666,48.192633333333333&FORMAT=votable&SIZE=0.166666666666666&BAND=6530e-10/6580e-10

Query results:  
 CCD700-voarchive

l...	ssa_specstart	ssa_specend	ssa_dstitle	ssa_targname	ssa_dateObs	ssa_timeExt	ssa_snr	ssa_length	accref	
12	6.26183E-7	6.77426E-7	ccd700/data/psiper/6255-676...	Psi Per	53216.0049	1800.		1997	http://voarchive.asu.cas.cz/ge...	appli
27	6.24978E-7	6.76217E-7	ccd700/data/psiper/6255-676...	psi Per	54701.97209	1333.19		1997	http://voarchive.asu.cas.cz/ge...	appli
13	6.25854E-7	6.77081E-7	ccd700/data/psiper/6255-676...	psi Per	52982.77964	1300.		1997	http://voarchive.asu.cas.cz/ge...	appli
31	6.26359E-7	6.73614E-7	ccd700/data/psiper/6260-673...	psi Per	56884.01015	1201.		2047	http://voarchive.asu.cas.cz/ge...	appli
30	6.26359E-7	6.73614E-7	ccd700/data/psiper/6260-673...	psi Per	56884.02565	901.		2047	http://voarchive.asu.cas.cz/ge...	appli
2	6.26234E-7	6.77466E-7	ccd700/data/psiper/6255-676...	psi Per	53451.88616	899.564		1997	http://voarchive.asu.cas.cz/ge...	appli
28	6.26587E-7	6.77822E-7	ccd700/data/psiper/6255-676...	psi Per	54209.80019	899.214		1997	http://voarchive.asu.cas.cz/ge...	appli
3	6.25267E-7	6.76496E-7	ccd700/data/psiper/6255-676...	Psi Per	55902.84843	800.		1997	http://voarchive.asu.cas.cz/ge...	appli
10	6.25163E-7	6.76402E-7	ccd700/data/psiper/6255-676...	Psi Per	54757.88729	730.		1997	http://voarchive.asu.cas.cz/ge...	appli
6	6.26183E-7	6.77426E-7	ccd700/data/psiper/6255-676...	Psi Per	53216.02811	600.		1997	http://voarchive.asu.cas.cz/ge...	appli
14	6.24978E-7	6.76217E-7	ccd700/data/psiper/6255-676...	psi Per	54701.96266	600.		1997	http://voarchive.asu.cas.cz/ge...	appli
9	6.25809E-7	6.77051E-7	ccd700/data/psiper/6255-676...	psi Per	52897.07116	556.599		1997	http://voarchive.asu.cas.cz/ge...	appli
18	6.26587E-7	6.77822E-7	ccd700/data/psiper/6255-676...	psi Per	54209.81274	450.		1997	http://voarchive.asu.cas.cz/ge...	appli
20	6.25854E-7	6.77081E-7	ccd700/data/psiper/6255-676...	psi Per	52982.79788	350.		1997	http://voarchive.asu.cas.cz/ge...	appli
16	6.26235E-7	6.77492E-7	ccd700/data/psiper/6255-676...	psi Per	53475.85182	300.043		1997	http://voarchive.asu.cas.cz/ge...	appli
19	6.26401E-7	6.77640E-7	ccd700/data/psiper/6255-676...	Psi Per	53249.01697	300.		1997	http://voarchive.asu.cas.cz/ge...	appli
23	6.25161E-7	6.76378E-7	ccd700/data/psiper/6255-676...	psi Per	54519.80405	300.		1997	http://voarchive.asu.cas.cz/ge...	appli
24	6.25161E-7	6.76378E-7	ccd700/data/psiper/6255-676...	psi Per	54519.7984	300.		1997	http://voarchive.asu.cas.cz/ge...	appli
7	6.26434E-7	6.77677E-7	ccd700/data/psiper/6255-676...	psi Per	53228.0436	218.373		1997	http://voarchive.asu.cas.cz/ge...	appli
11	6.25641E-7	6.76875E-7	ccd700/data/psiper/6255-676...	psi Per	55443.86444	200.		1997	http://voarchive.asu.cas.cz/ge...	appli
22	6.25161E-7	6.76378E-7	ccd700/data/psiper/6255-676...	psi Per	54519.81162	200.		1997	http://voarchive.asu.cas.cz/ge...	appli
25	6.25908E-7	6.77147E-7	ccd700/data/psiper/6255-676...	psi Per	52901.91749	194.574		1997	http://voarchive.asu.cas.cz/ge...	appli
15	6.25937E-7	6.77178E-7	ccd700/data/psiper/6255-676...	psi Per	52904.03609	180.		1997	http://voarchive.asu.cas.cz/ge...	appli
1	6.26100E-7	6.77340E-7	ccd700/data/psiper/6255-676...	psi Per	55396.07196	150.		1997	http://voarchive.asu.cas.cz/ge...	appli

SSA Servers

short name	title
BeSS	Be Stars Spectra
califa ssa	CALIFA DR2
castor	Espadons/Narval leg...
castor+	Espadons/NARVAL le...
CCD700-voarchive	ccd700 OND
ccd700-vos2	
CDFS SSAP	Optical Spectroscop...
CENCOS-VWDS_DEEP	CENCOS-VWDS_DEEP...
CENCOS-VWDS_DEEP+	CENCOS-VWDS_DEEP...
Cfa Hectospec	Cfa Hectospec Spec...
dk154	
dk154-extr15	
dk154-extr15jan	dk154
dk154-extr16jan	dk154
DK154-SSA	DK154 SSA
ELODIE	ELODIE archive
ELODIEinterp	Spectrum interpolat...
ESO SAF SSA	ESO Science Archive...
EUVE	Extreme Ultraviolet ...
extract_jul16	DK154
F/H Orders SSAP	Flash/Heros Split-Or...
FEROS SSAP	FEROS Public Spectra
Flash/Heros SSAP	Flash/Heros SSAP
FUSE	Far Ultraviolet Spect...
GALEX	Galaxy Evolution Exp...
GIRAFFE	GIRAFFE archive of r...
H.E.S.S.	High Energy Stereos...
HEAVENS @ ISDC	Mining the HEAVENS...
HEROS OND	HEROS archive of On...
HEROS OND CUTOUT	cutout server of HE...

# LAMOST Query Ha region

Starlink SPLAT-VO: Query VO for Spectra

File Options Resolver Interop Help

Service selection options  
Data Source  
 Observed data  Theoretical data

Wave Band  
 Radio  Millimeter  Infrared  
 Optical  UV  EUV  
 X-ray  Gamma-ray  ALL

Tags

SSAP Servers

short name	title
HFA	HyperLeda FITS Archive...
INES ARCHIVE	INES: The IUE Newly Ext...
IUE	International Ultraviolet...
ISO SSAP	ISO ESA Archive SSAP
IUE highly processed s...	IUE highly processed s...
LAMOST.DR1_SSAP	LAMOST DR1 SPECTRUM...
M4AST	M4AST - Modeling for A...
Magic SSAP	Magic Public Spectra S...
MAGIC	MAGIC Spectrum Service
HEAVENS @ ISDC	Mining the HEAVENS wi...
NOVA Spectral Catalog	NOVA Spectral Catalog
OMC SSA	OMC: The INTEGRAL Opt...
CCD700 OND voarchive	Ondrejov CCD700 spectr...
CCD700 OND vos2	Ondrejov CCD700 spectr...
LAMOST DR1 OND voarc...	Ondrejov copy of LAMO...
LAMOST DR1 OND vos2	Ondrejov copy of LAMO...
LAMOST DR3 OND vos2	Ondrejov copy of LAMO...
HEROS OND voarchive	Ondrejov HEROS spectr...
HEROS OND vos2	Ondrejov HEROS spectr...
XIDResult SSA	Optical spectra of the X...
CDFS SSAP	Optical Spectroscopy in...
Polarbase SSAP	Polarbase SSAP service...
mlqso bidi ssa	Spectra of lensed QSO...
ELODIEinterp	Spectrum interpolator f...
GAUDIVO	SSAP for GAUDI
RCSED_SSAP	SSA service for RCSED ...
SubaruHDS	Subaru HDS Spectrum ...
SVO HERBIG AeBe	SVO EXPORT HERBIG Ae...
TBL Narval	TBL Narval legacy
COROT ARCHIVE	The COROT PUBLIC ARC...
The Mark-I solar spectr...	The Mark-I solar spectr...
NED/SED	The NASA/IPAC Extragal...
PCSLG SSAP	The Panchromatic High...
VIPERS Spectra	The VIMOS Public Extra...
TUES	Tubingen Echelle Spect...
Ultraviolet photometry...	Ultraviolet photometry...
VizieR SSA	VizieR SSA service
VUDS COSMOS DR1	VUDS-COSMOS (DR1) - ...

Select all Deselect all

Search parameters:  
Simple Query  
 Object:     
 RA:  Dec:   
 Radius:  MAXREC:   
 Band:  Time:   
 Query Format:   
 Wavelength calibration:   
 Flux calibration:

Optional Parameters

Use	Name	Value	UCD
<input type="checkbox"/>	SPECRP		spect.resolution:em.wl
<input type="checkbox"/>	SPATRES		pos.angResolution
<input type="checkbox"/>	PUBIDD		
<input type="checkbox"/>	CREATORID		meta.id
<input type="checkbox"/>	WILDTARGET		
<input type="checkbox"/>	WILDTARGETCASE		
<input type="checkbox"/>	TOP		

Select all Deselect all Update

Query: <SERVER>?REQUEST=queryData&POS=119.2666416666667,2.9508444444444444&FORMAT=ALL&SIZE=0.008333333333333333

SEND QUERY

Query results:

CCD700 OND vos2 LAMOST DR1 OND vos2

l...	ssa_specstart	ssa_specend	ssa_dstitle	ssa_targname	ssa_dateObs	ssa_timeExt	ssa_snr	ssa_length	accref	
1	3.69999E-7	9.07403E-7	lamost_dr1/data/fits/GAC118N...	LAMOST J075704.05+025652.5	56315.65115	1200.		3897	http://vos2.asu.cas.cz/getpro...	im
2	3.69999E-7	9.07403E-7	lamost_dr1/data/fits/GAC118N...	LAMOST J075704.05+025652.5	56315.65115	1200.		3897	http://vos2.asu.cas.cz/getpro...	ap
3	3.69999E-7	9.07403E-7	lamost_dr1/data/fits/GAC118N...	LAMOST J075704.21+025655.6	56315.67766	1200.		3897	http://vos2.asu.cas.cz/getpro...	im
4	3.69999E-7	9.07403E-7	lamost_dr1/data/fits/GAC118N...	LAMOST J075704.21+025655.6	56315.67766	1200.		3897	http://vos2.asu.cas.cz/getpro...	ap
5	3.69999E-7	9.07821E-7	lamost_dr1/data/fits/GAC121N...	LAMOST J075703.80+025653.6	56341.66307	1200.		3899	http://vos2.asu.cas.cz/getpro...	im
6	3.69999E-7	9.07821E-7	lamost_dr1/data/fits/GAC121N...	LAMOST J075703.80+025653.6	56341.66307	1200.		3899	http://vos2.asu.cas.cz/getpro...	ap

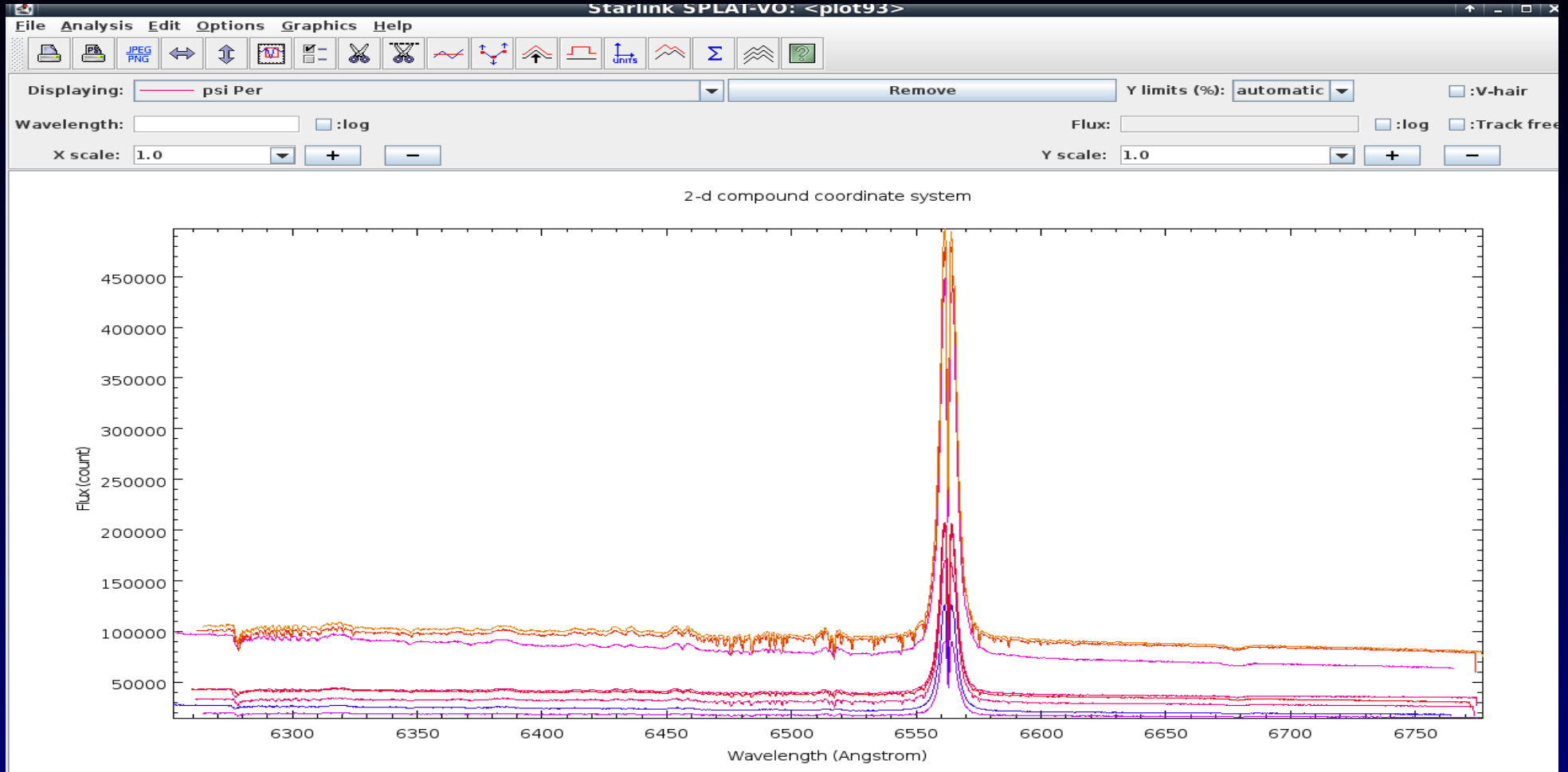
Parameters for Server-Generated data processing

FLUXCALIB :   
 REDSHIFT :   
 BAND :  /  [ m]  
 FORMAT :

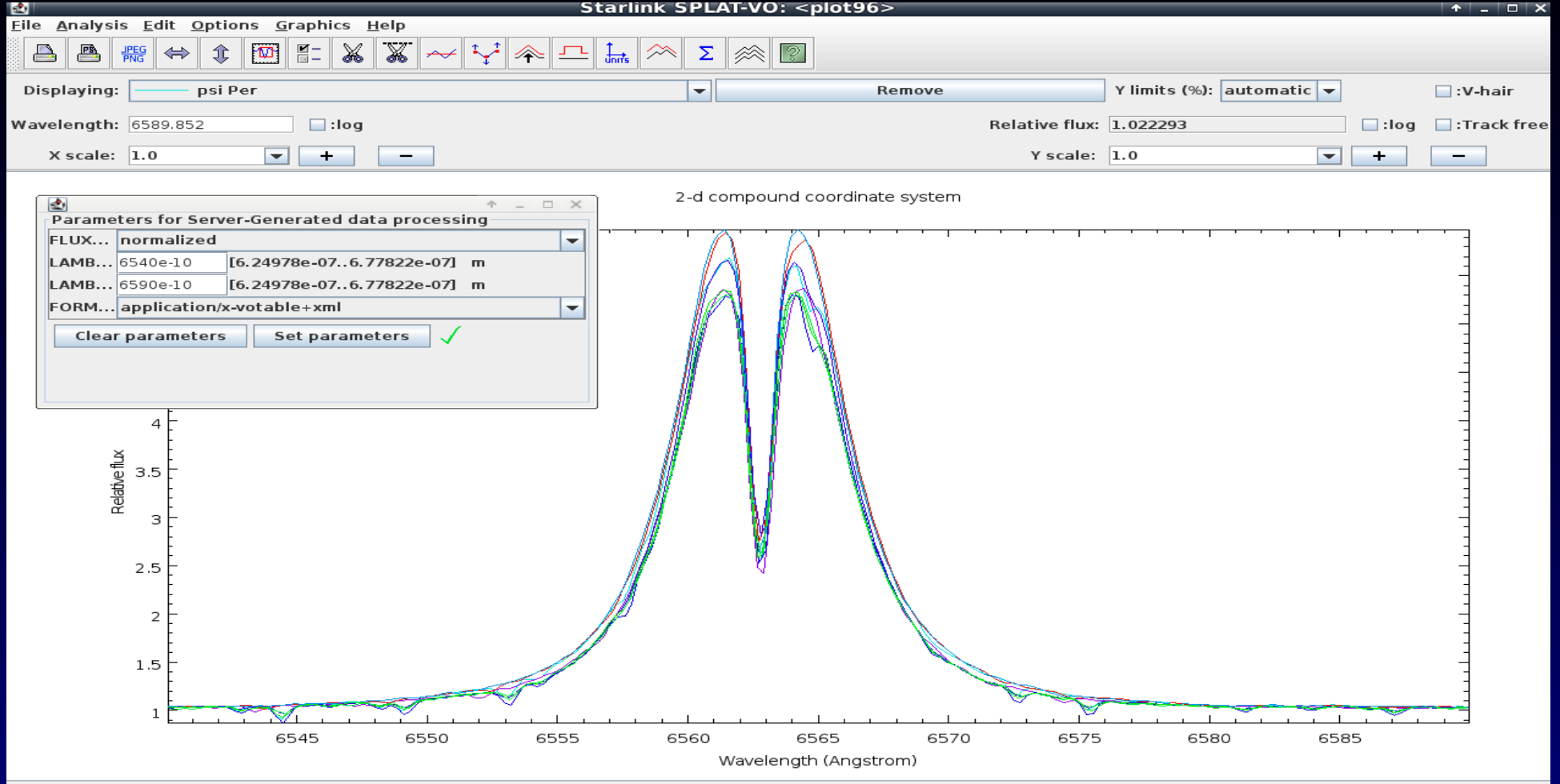
Display selected Display all Download selected Download all Deselect table Deselect all DataLink Services

Save query results Restore query results Close

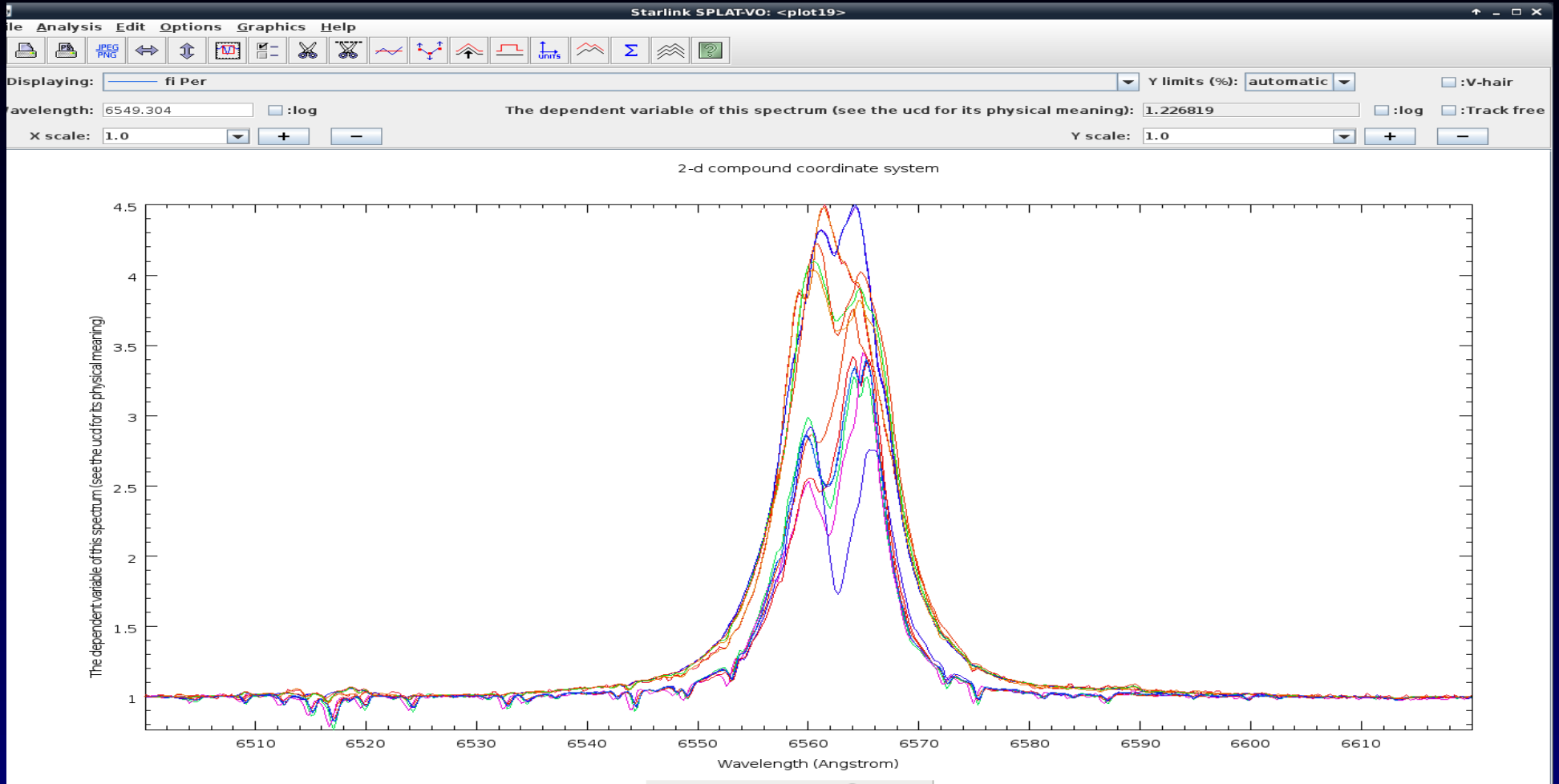
# Spectra in SPLAT-VO direct access



# Spectra in SPLAT-VO - DataLink



# Variability of Line Profile on Be star



# Series of Spectra - LPV

Starlink SPLAT-VO: <plot3>

File Analysis Edit Options Graphics Help



Displaying: 4 Her

Remove

Y limits (%): automatic

:V-hair

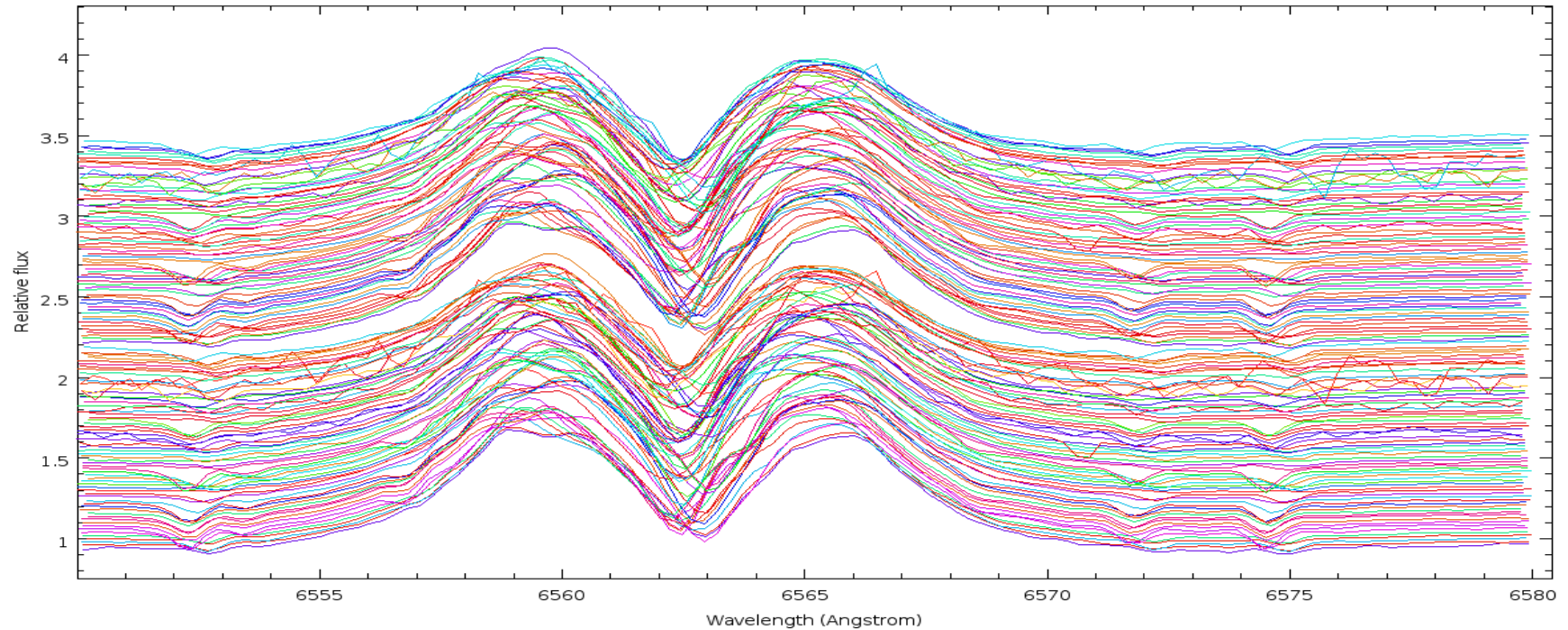
Wavelength: 6551.436  :log

Relative flux: 0.9117984  :log  :Track free

X scale: 1.0

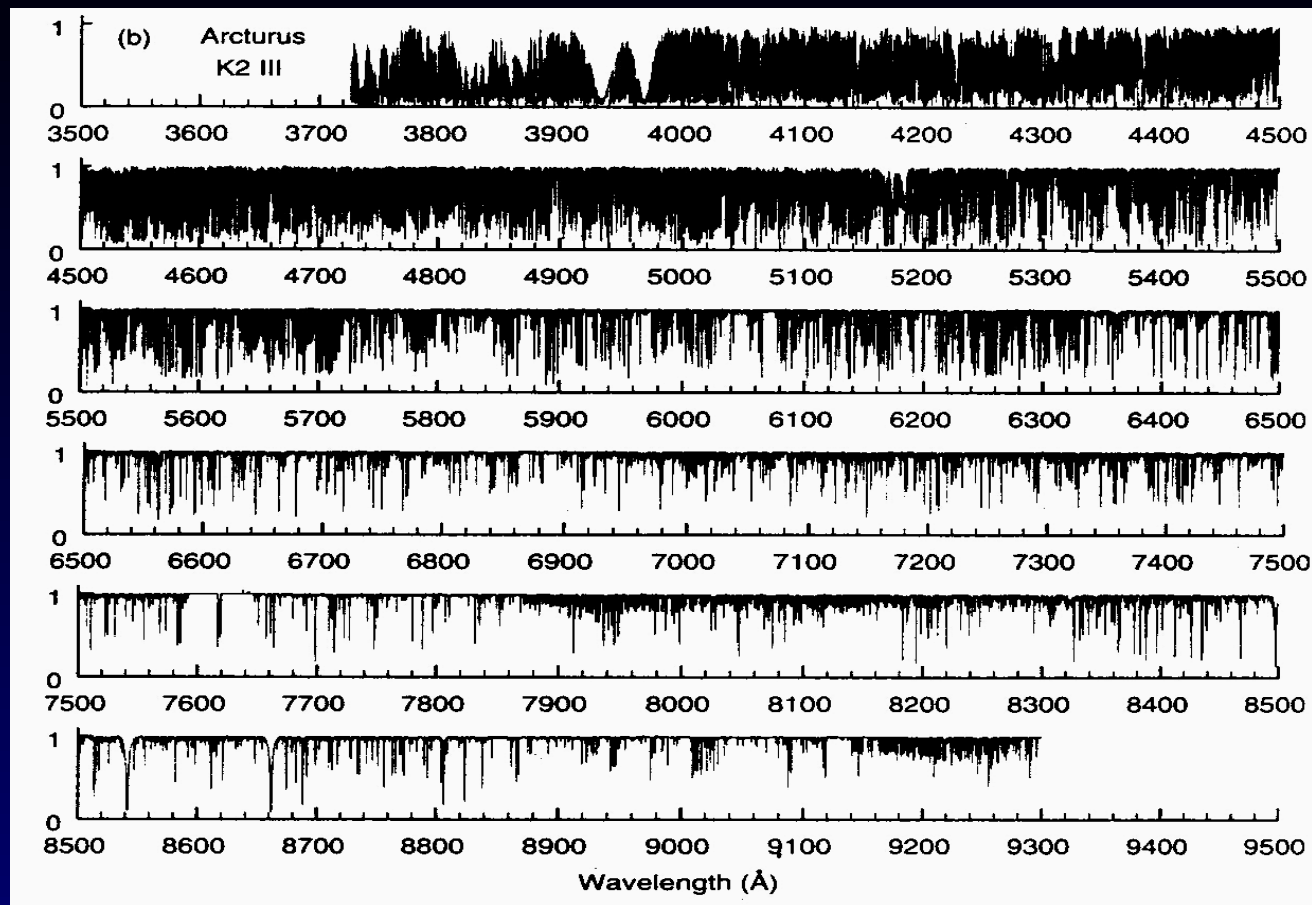
Y scale: 1.0

2-d compound coordinate system



# Echelle Spectra

# Atlas of Arcturus – 1D Normalized



# Orders Define Geometry

## Multiple orders

- Many orders to cover desired  $\lambda\lambda$ :

*Free spectral range*

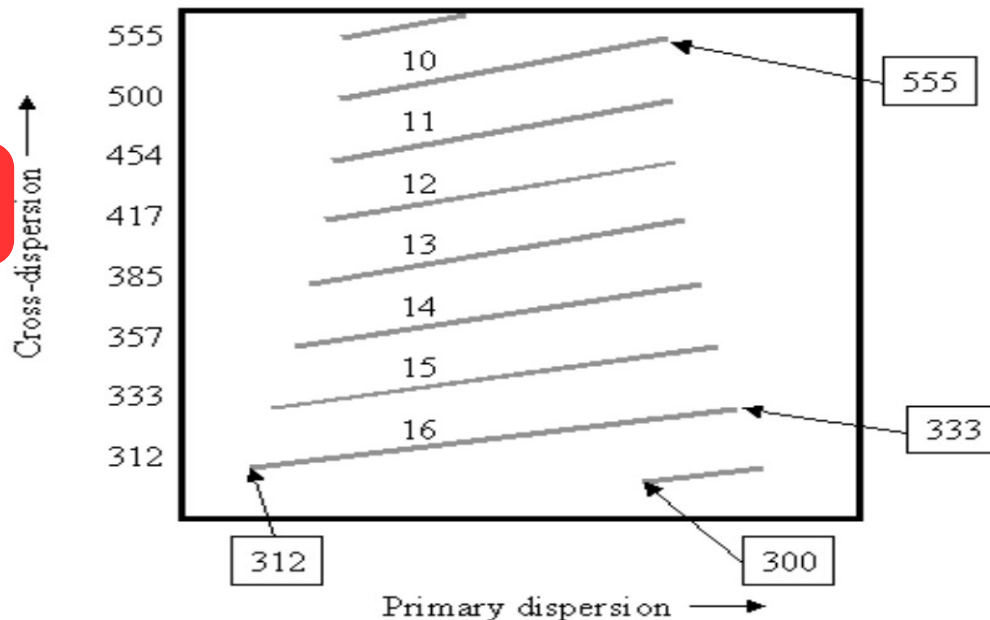
$$\Delta\lambda = \lambda/m$$

- Orders lie on top of each other:

$$\lambda(m) = \lambda(n) \times (n/m)$$

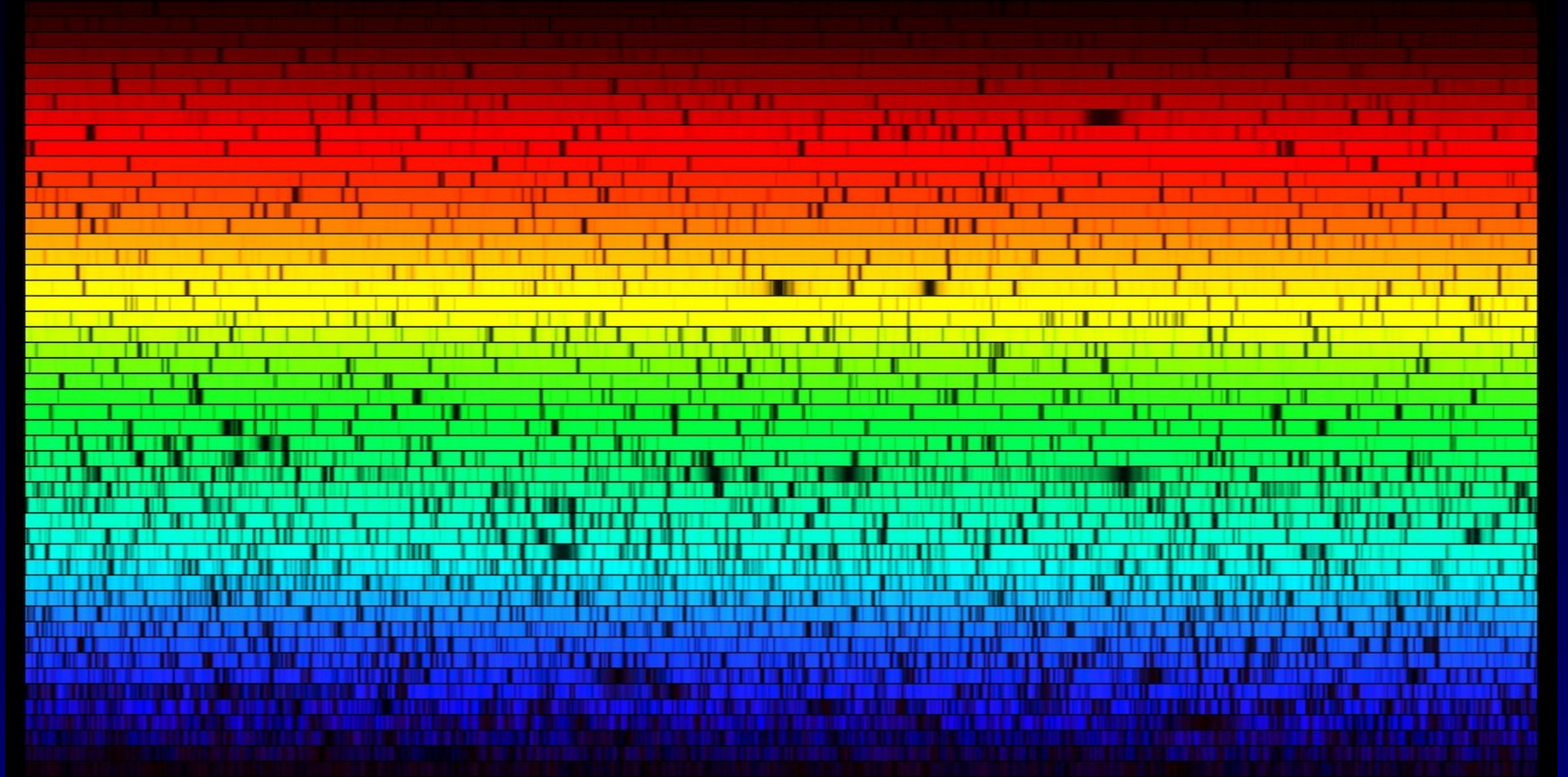
- Solution:

- use narrow passband filter to isolate one order at a time
- cross-disperse to fill detector with many orders at once

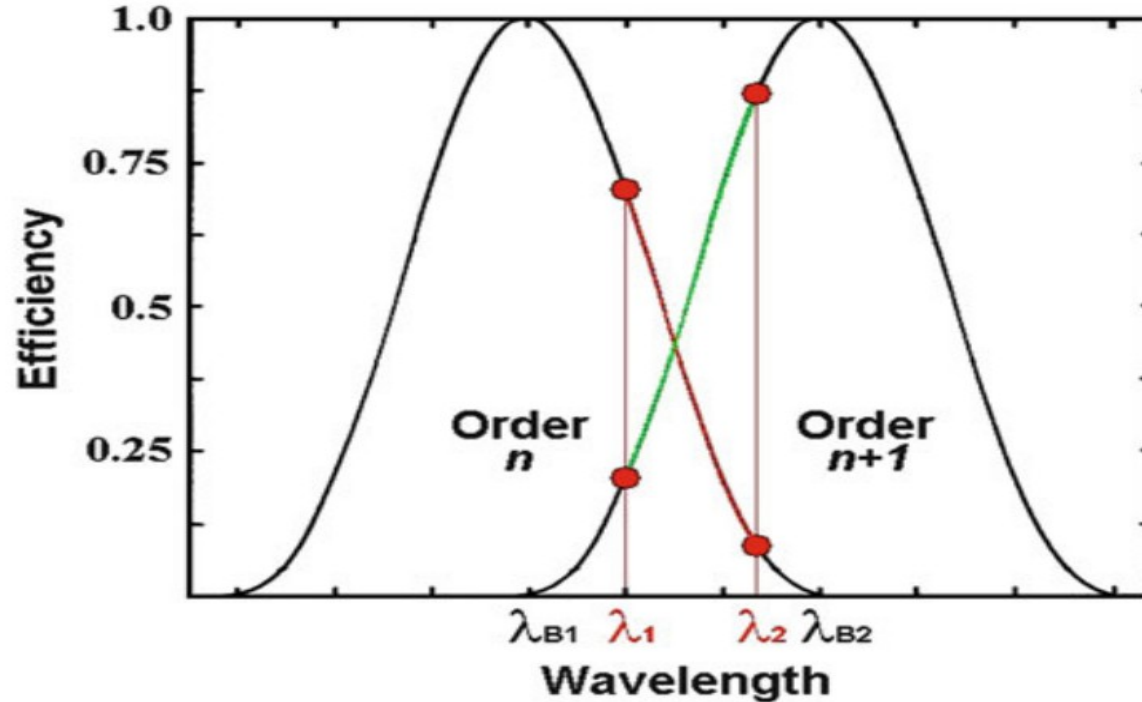


Cross dispersion may use prisms or low dispersion grating

# Spectrum of Arcturus (pixel-order)



# Overlapping Orders



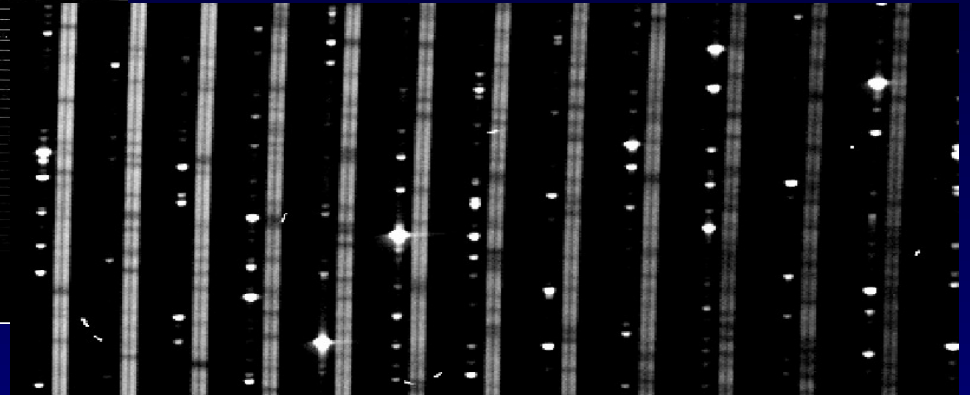
**Fig. 5.15** Sketch of the efficiencies in two adjacent orders. Identical wavelength intervals in the orders are indicated in *red* and *green*. Both wavelengths  $\lambda_1$  and  $\lambda_2$  are imaged in the two orders  $n$  and  $n + 1$ . The two blaze peaks of the orders are positioned at  $\lambda_{B1}$  and  $\lambda_{B2}$

# HARPS-S



71 orders (89-161) – 115 not  
3780-6910 Å  
R~115000  
dRV~3m/s

Simultaneous ThAr/Laser  
comb/Fabry Perot



<https://cdn.eso.org/images/large/eso0308c.jpg>

PlatoSpec simcal ThAr

# ESO X-Shooter

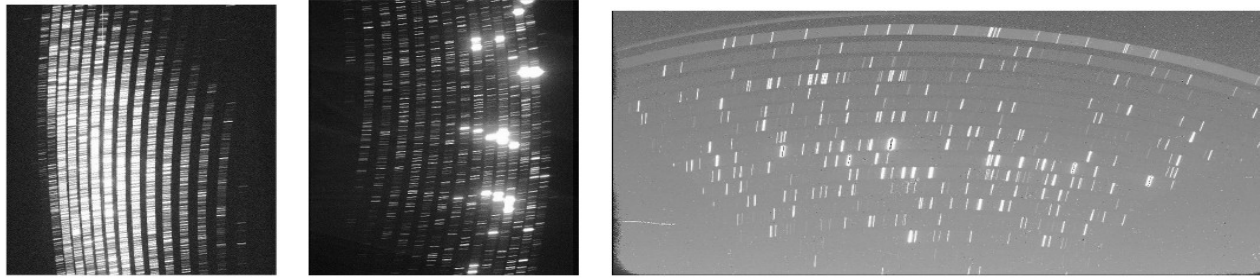
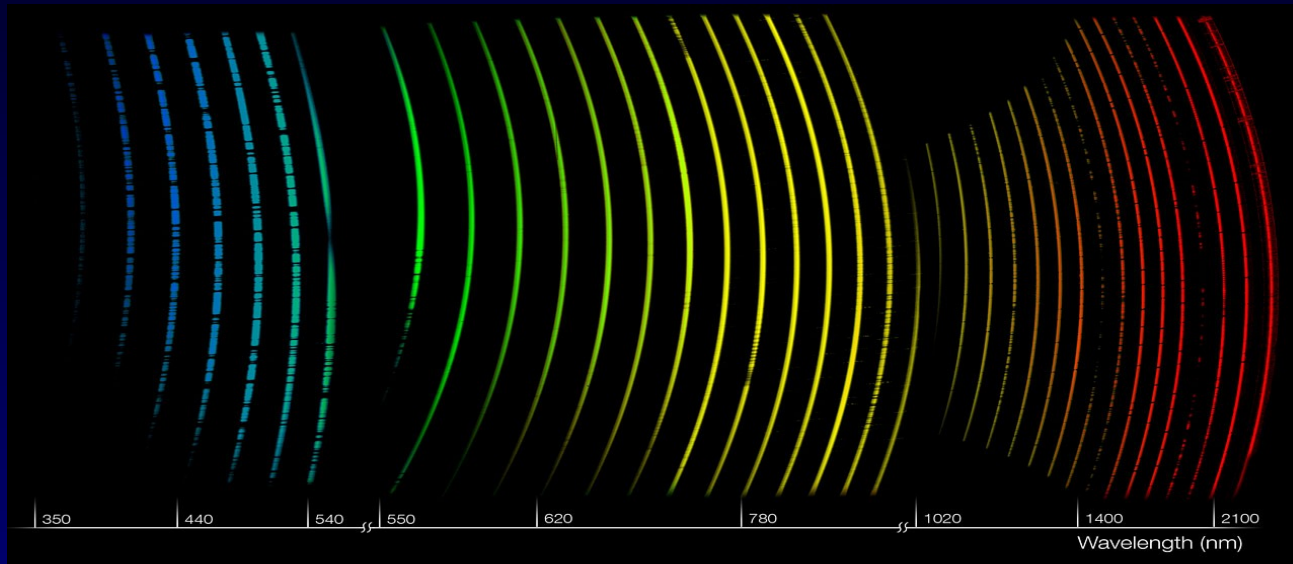
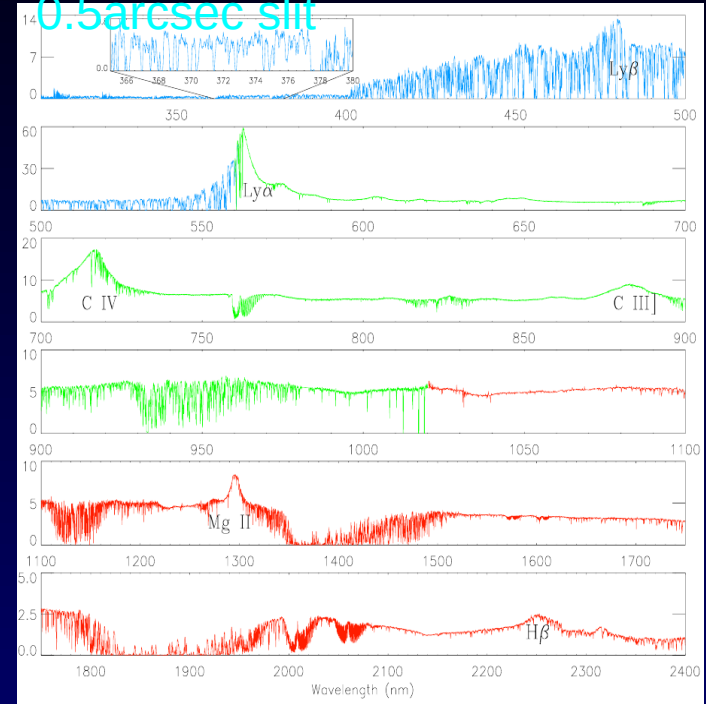


Figure 1. This image displays left to right the XSH UVB, VIS and NIR spectral formats with different image scales.



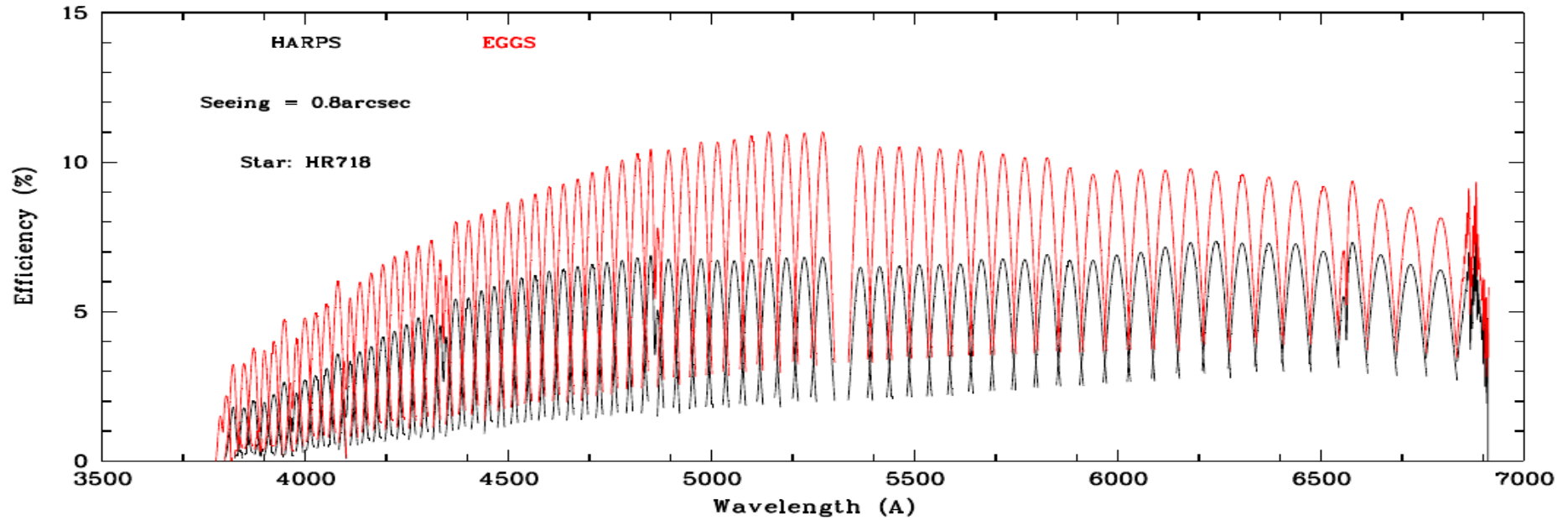
Modigliani et al. 2010

2936-24807 A,  
R ~9100,17400,11300  
0.5arcsec slit

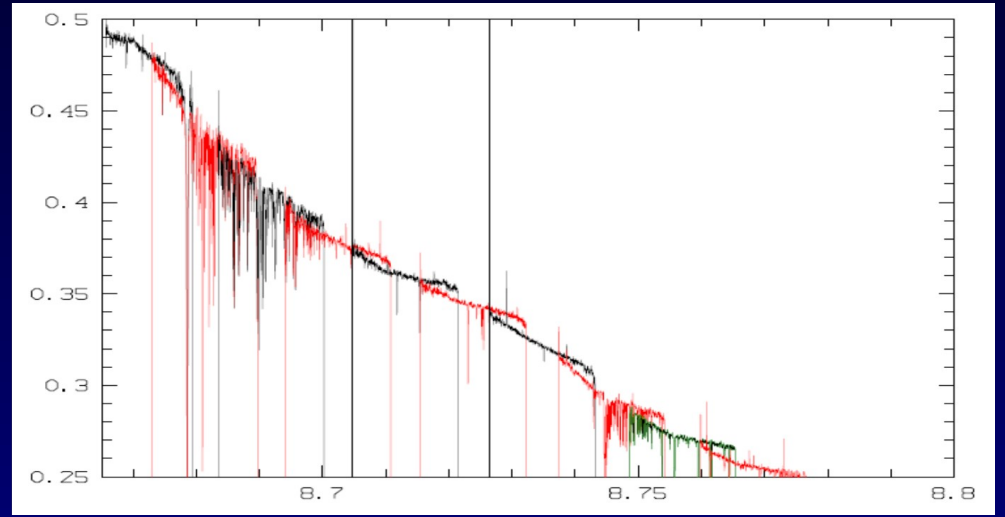
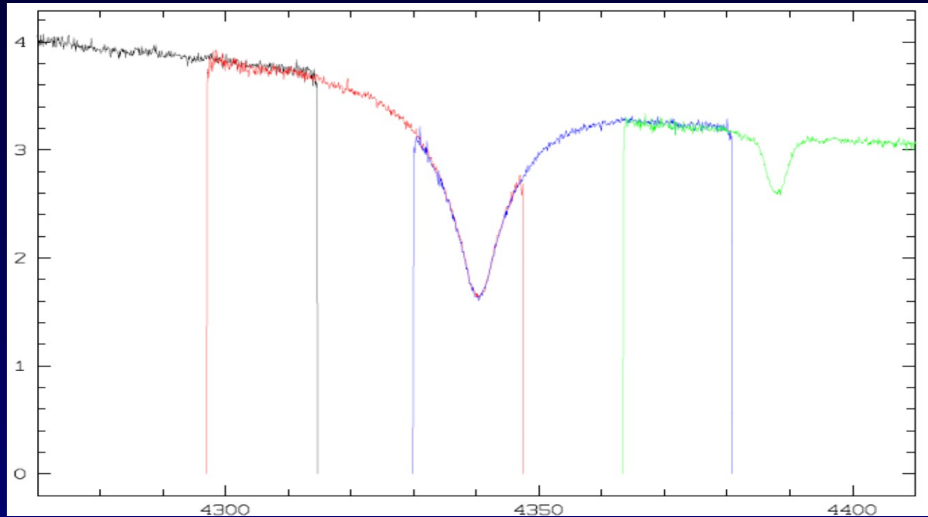
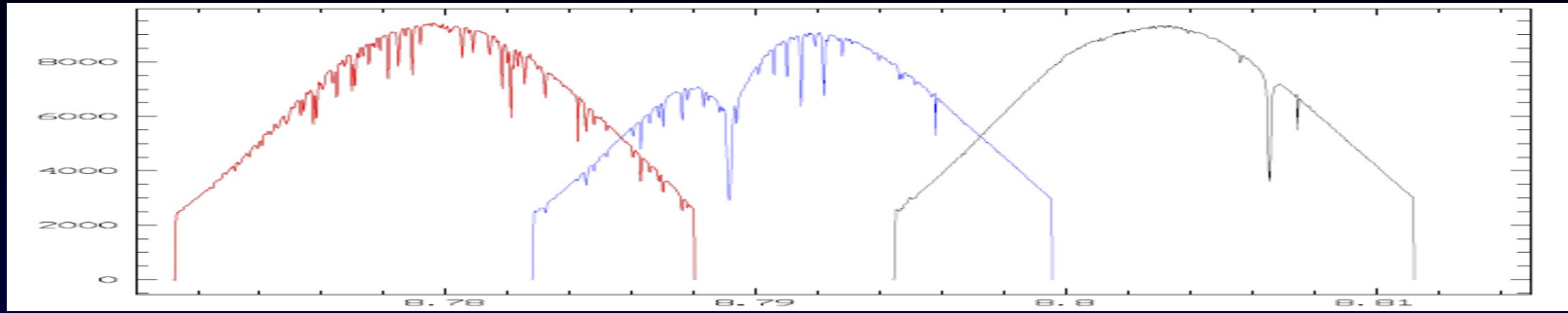


High-z QSO  
Vernet et al. 2011

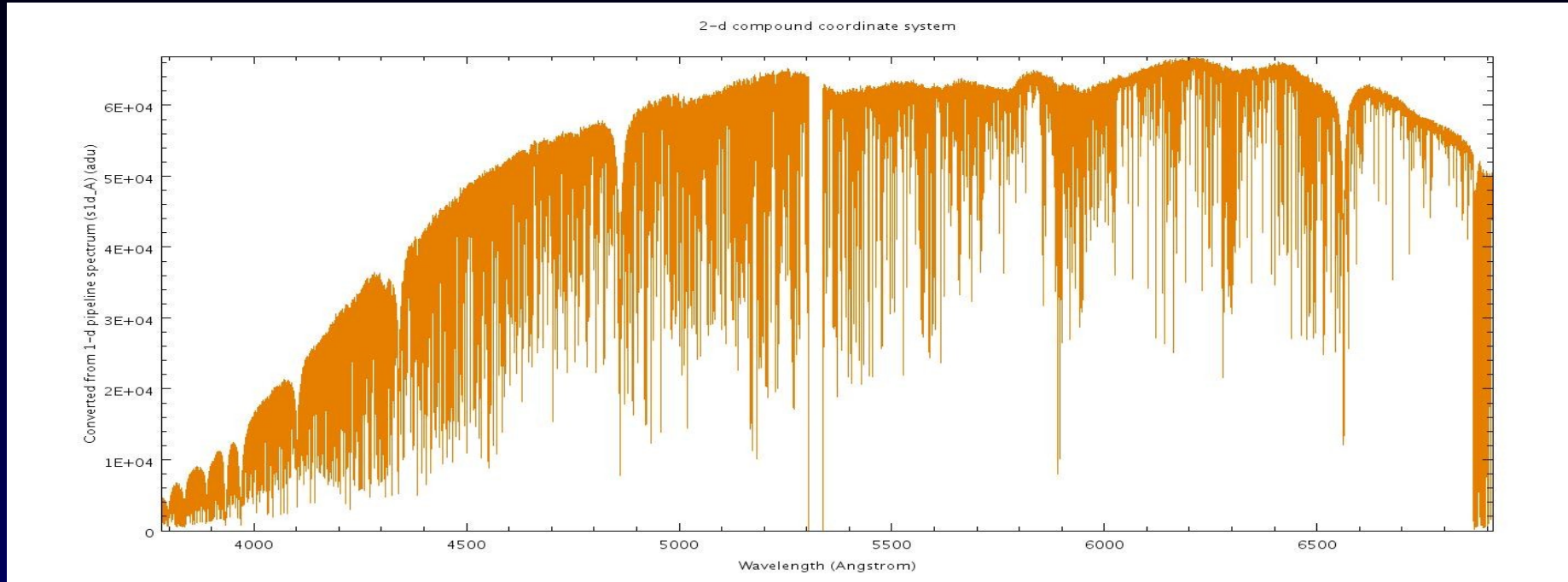
# HARPS-S



# Echelle Spectra Problems in Hot Stars



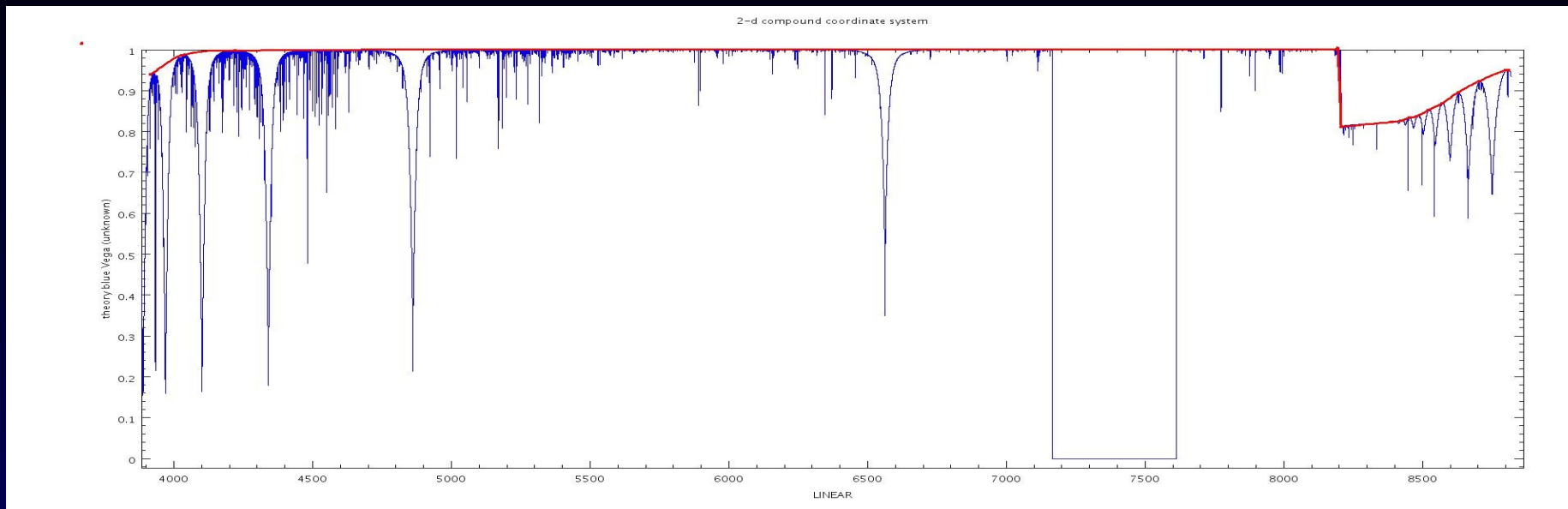
# HARPS-S in SPLAT



313138 points

Procyon

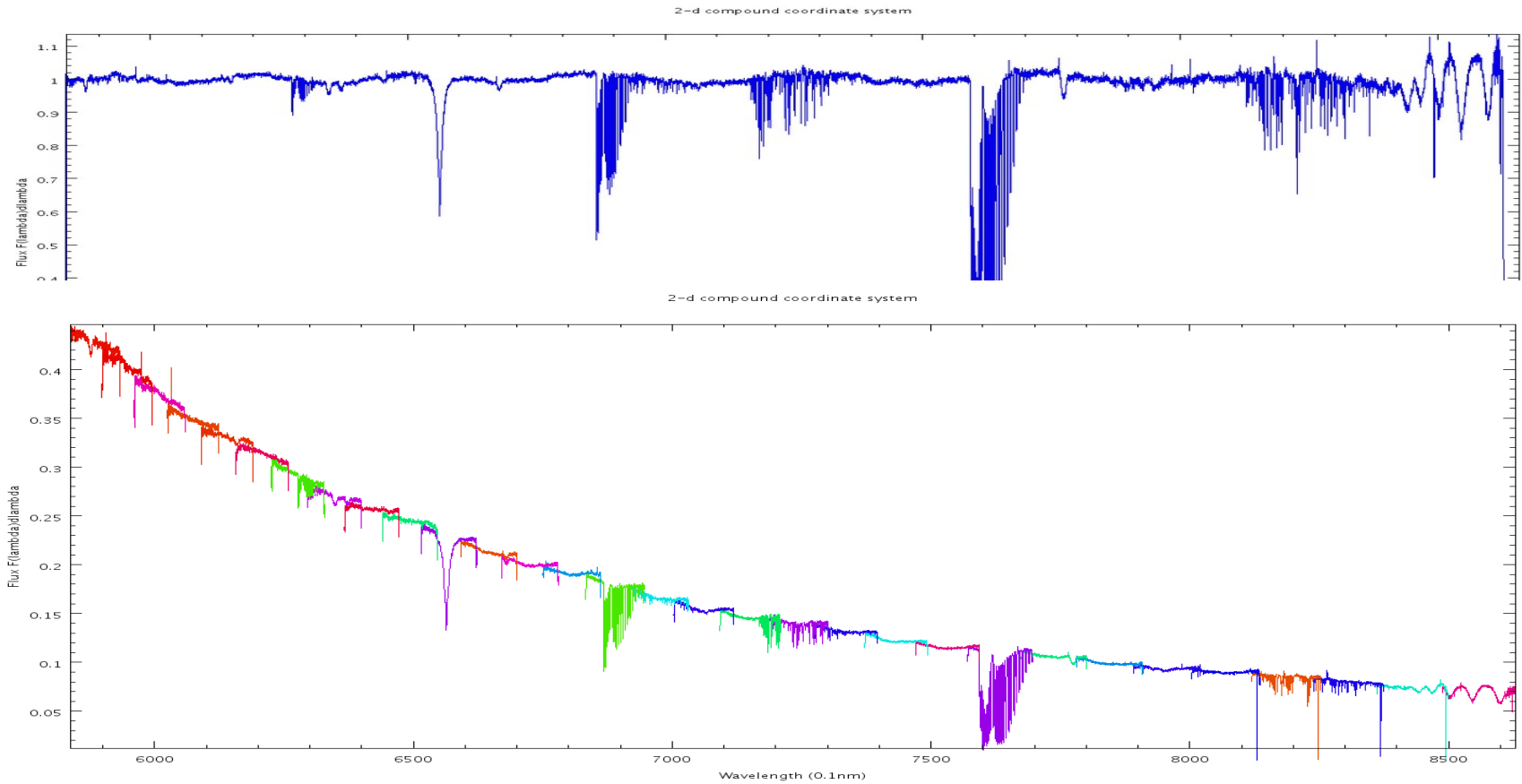
# Continuum Normalisation



Theoretical spectrum of Vega

The continuum is NOT ALWAYS at 1.0 !

# DaCHS Testbed Split-Order SSAP

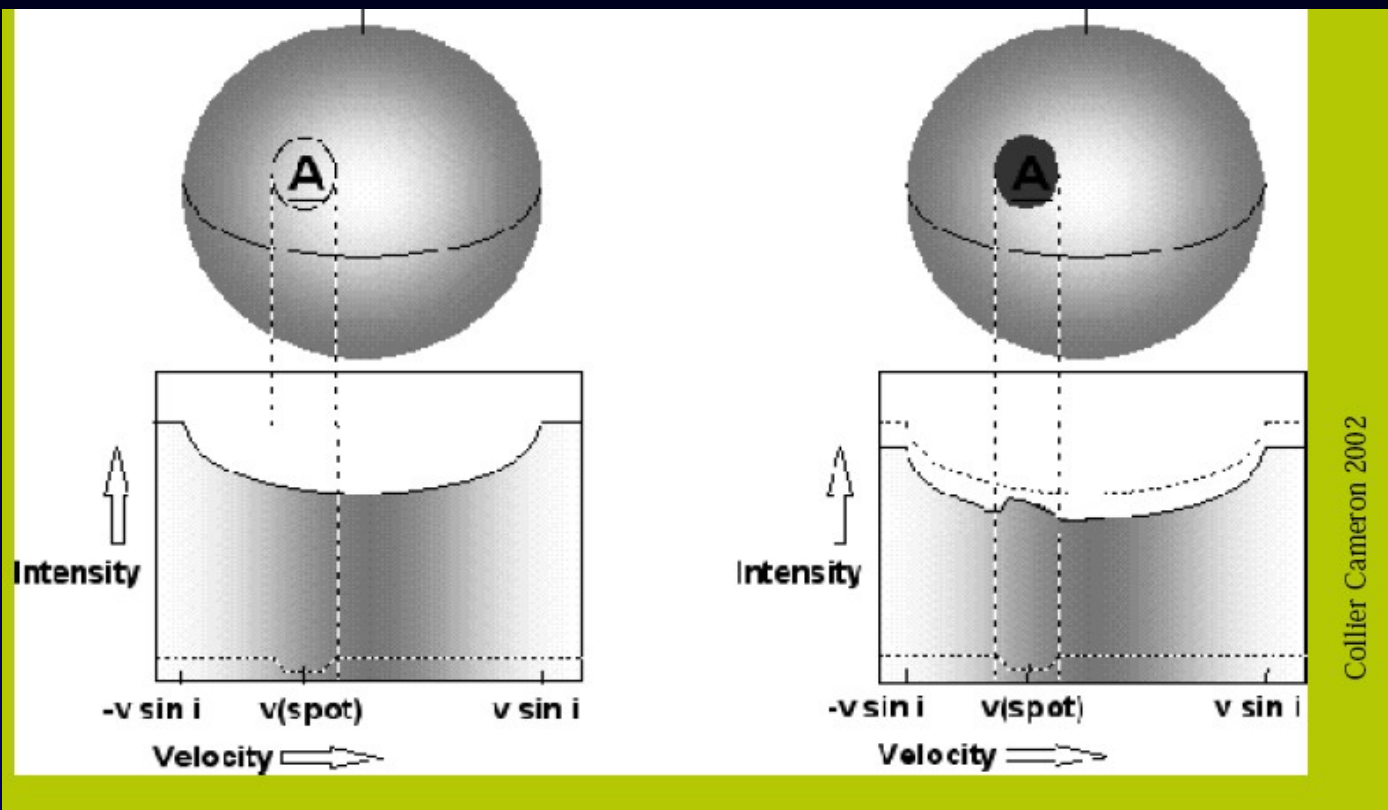


nu Pup HEROS red

**Products derived from spectra**

# Doppler Imaging

From LPV due to rotation  
stellar Spots - darker, brighter – chemical patch

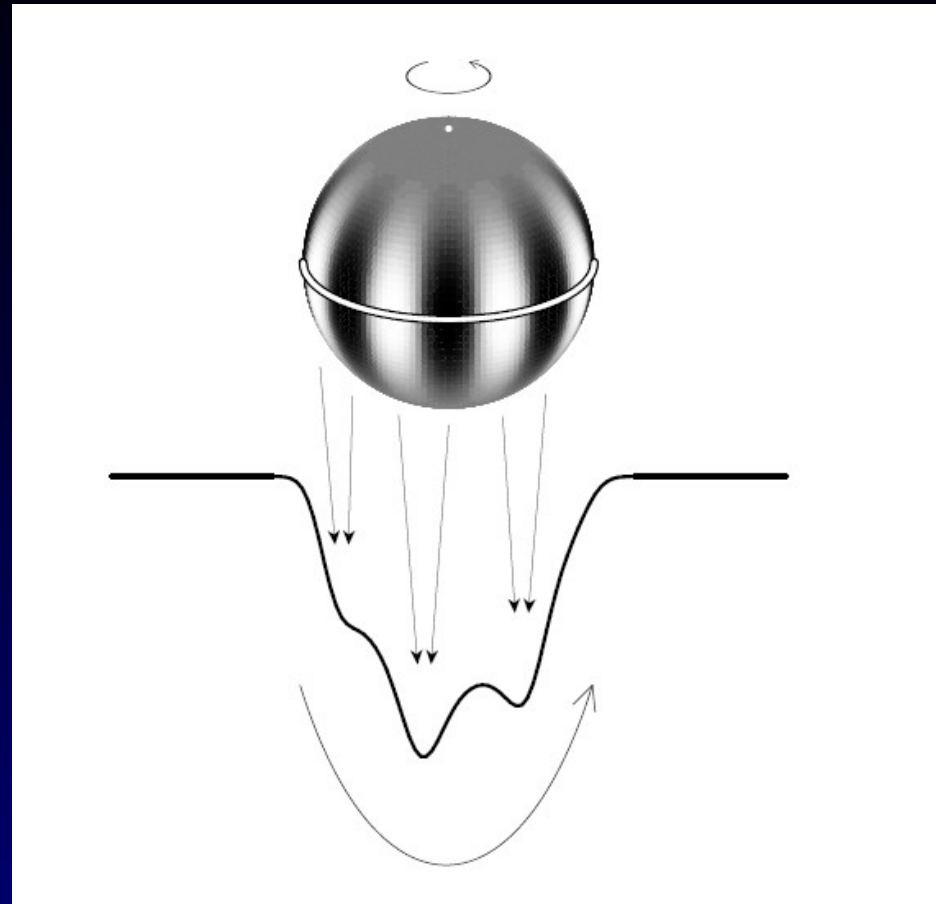


Collier Cameron 2002

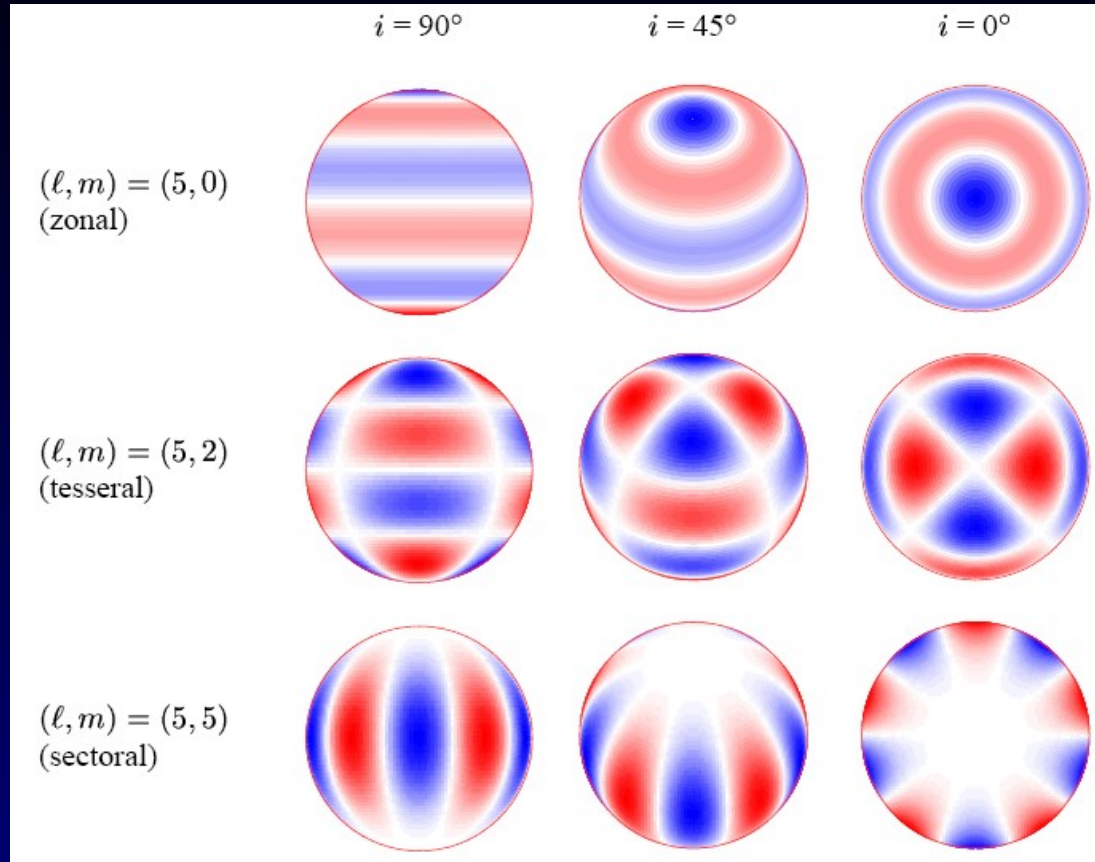


# Doppler Imaging - NRP

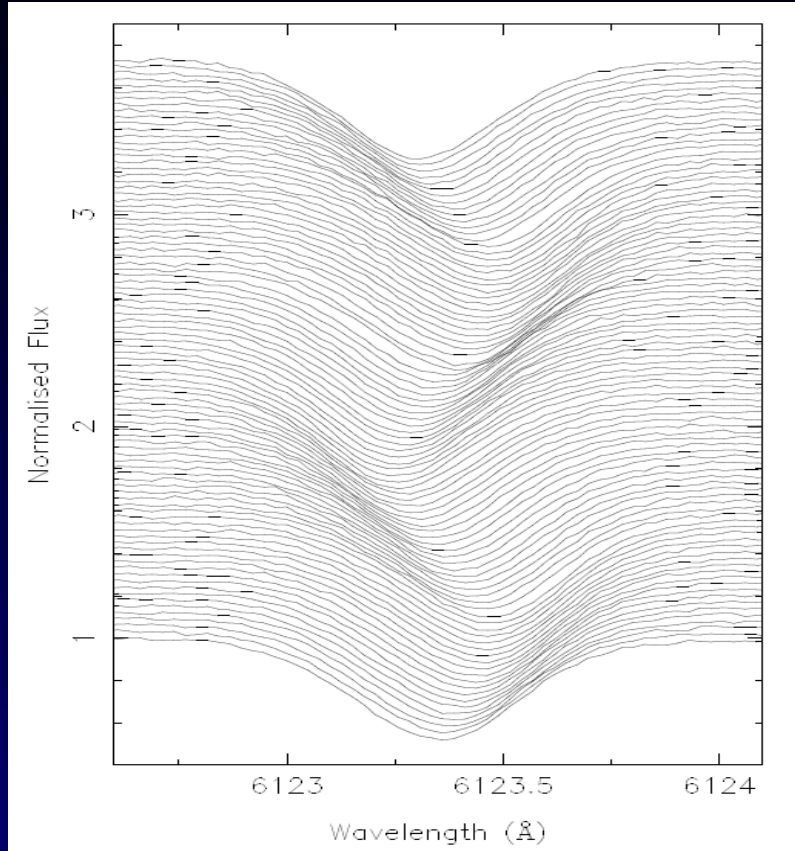
Vogt & Penrod -80s  
Zet Oph



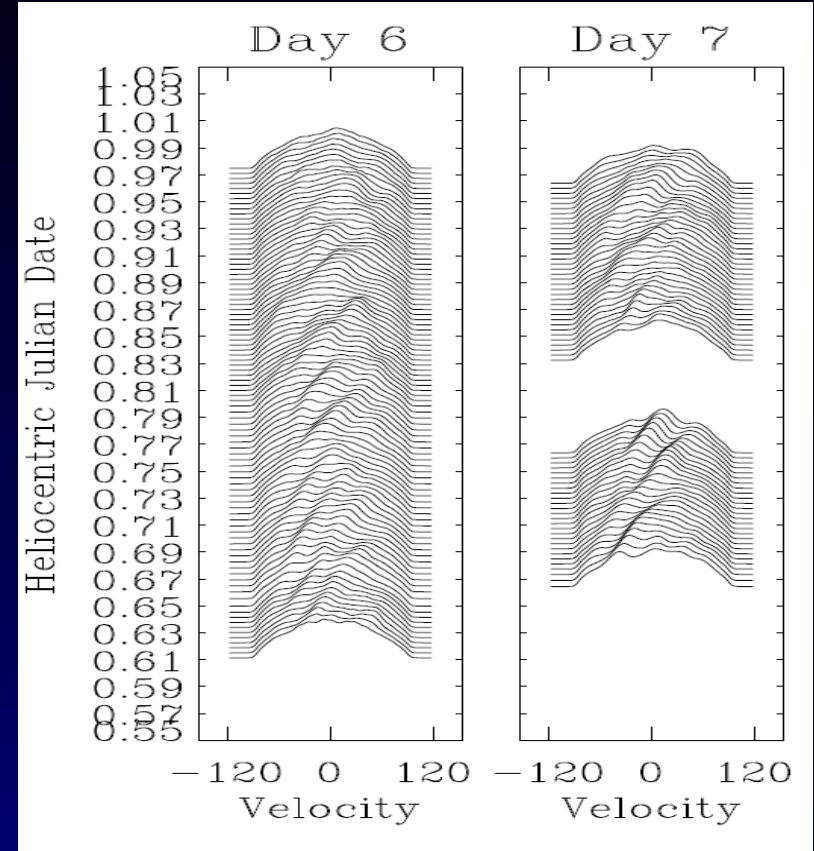
# Non Radial Pulsations Modes



# Measured Pulsations

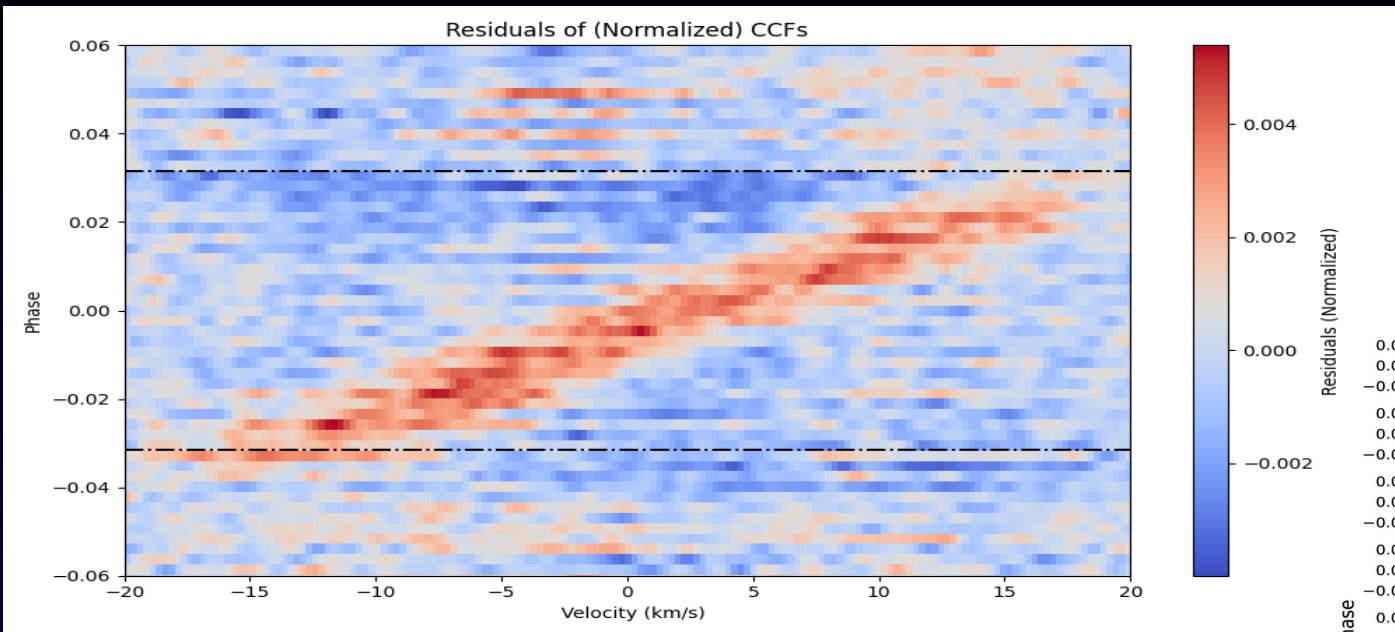


Rho Pup – del Sct type

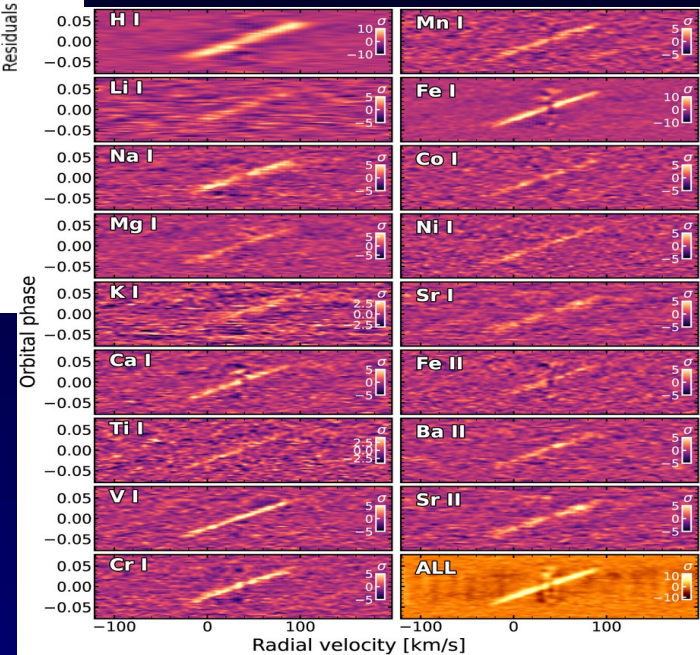


Eps Cep - del Sct type

# Doppler Shadow

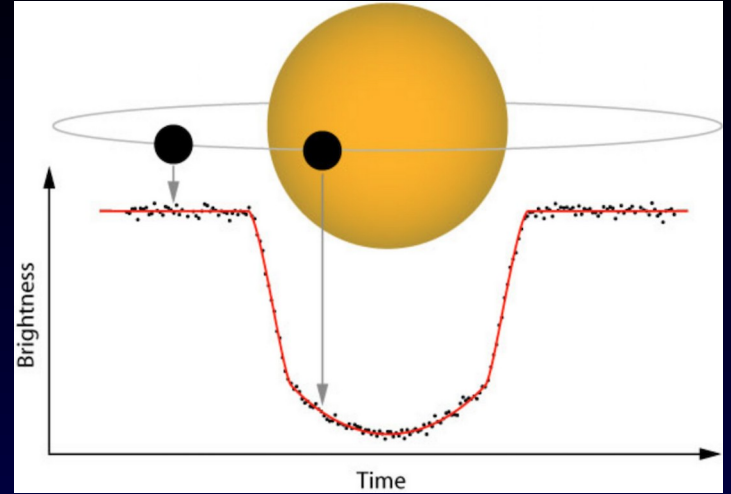
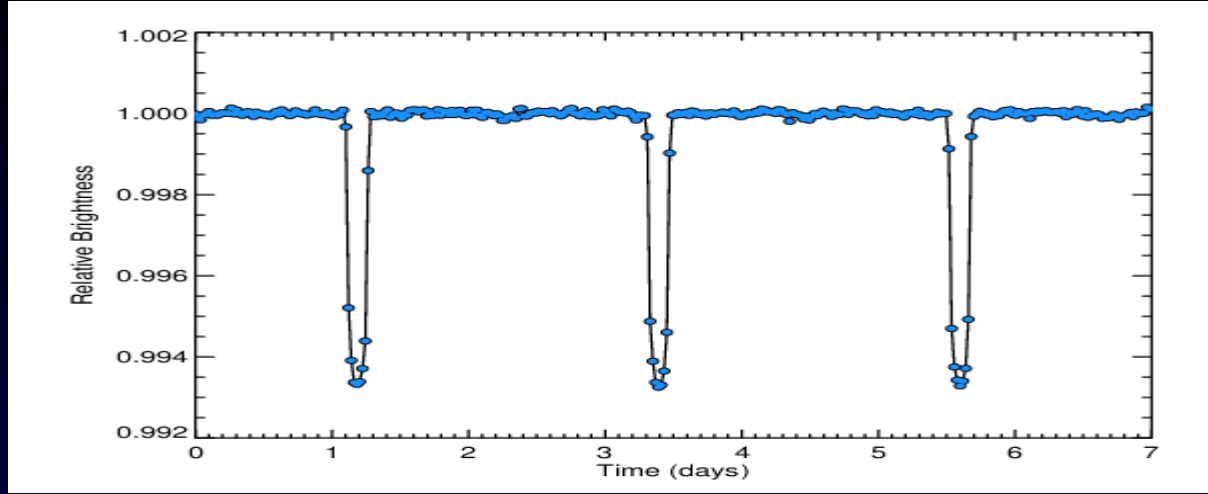


HARPS-N archival CCF of HAT-P-41A. Balkoová 2025

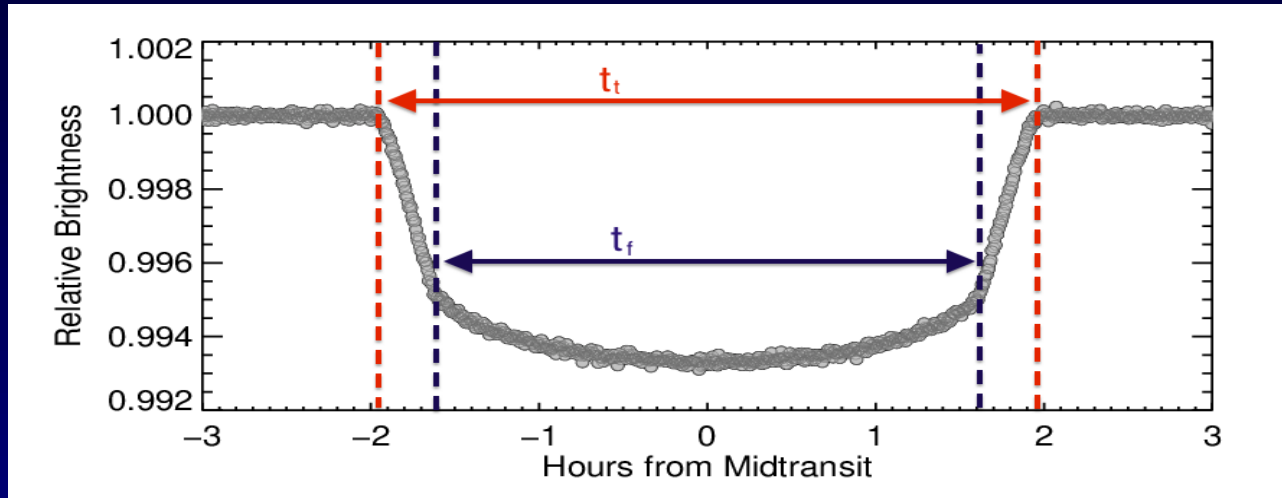


# Light Curve

# Light Curve HAT-P-7b

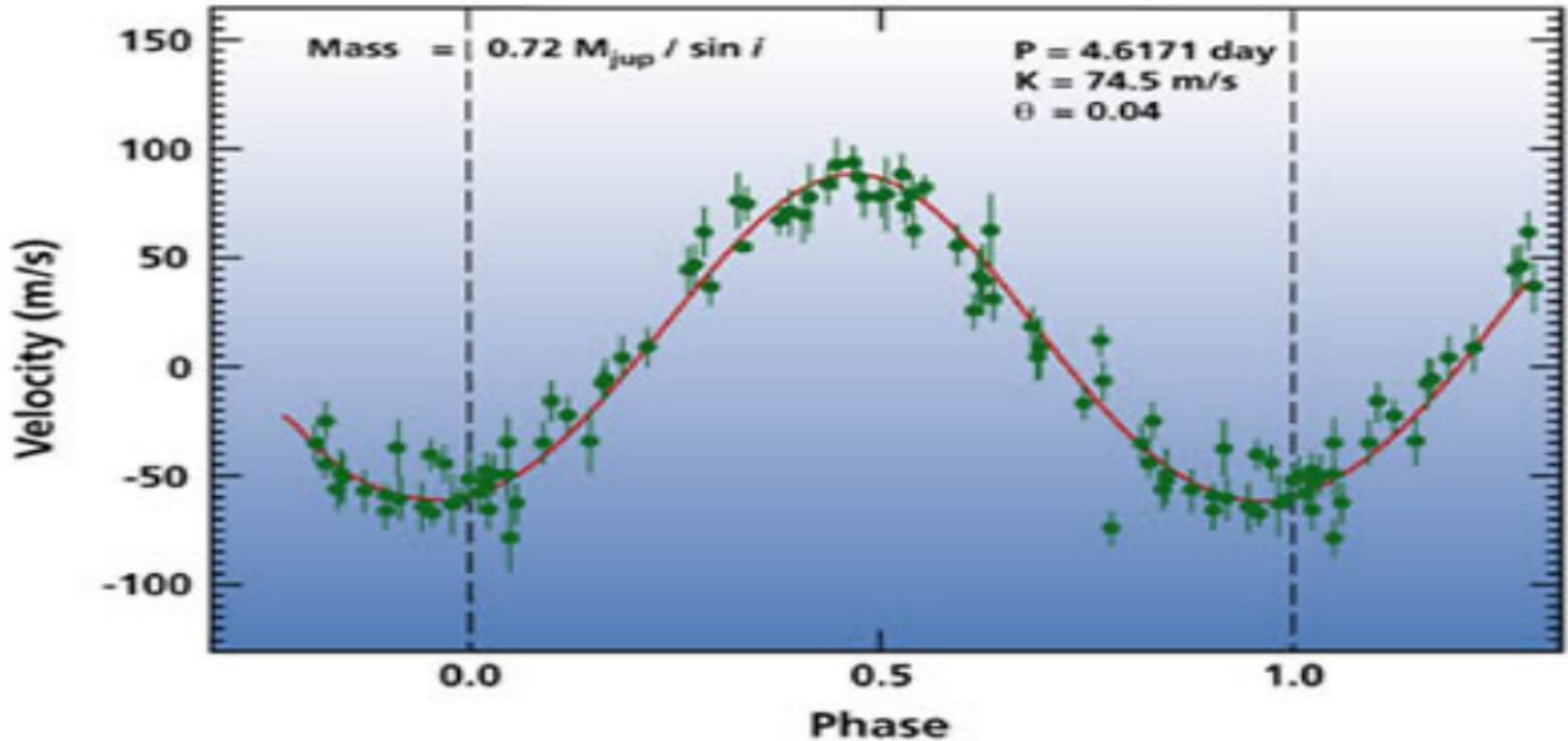


Vanderburg tutorial



# Radial Velocity Curve

# Radial Velocity Curve



# Trappist-1 RV Curve - Carmanes

Starlink SPLAT-VO: Query VO for Spectra

File Options Resolver Interop Help

Service selection options

Data Source  
 Observed data  Theoretical data

Wave Band  
 Radio  Millimeter  Infrared  
 Optical  UV  EUV  
 X-ray  Gamma-ray  ALL

Data type  
 Spectrum  Time series

Tags

SSAP Servers

short name	title
SMUSES	5 mJy Unbiased Spitzer ...
6dF Spectra	6dF DR3 Simple Spectr...
AGN Arch. SSAP	Archive of AGN spectral...
AKARI_IRC_SPEC_DSP	AKARI/IRC NIR Low-resol...
AKARI_IRC_SPEC_GALPN	AKARI/IRC NIR Spectral ...
AKARI_IRC_SPEC_V1	AKARI/IRC MIR-S slit-less...
ARCHES SED	ARCHES fully calibrated...
BEFS	Berkeley Extreme and ...
BeSS	Be Stars Spectra
BGDS TS SSAP	Bochum Galactic Disk S...
califa ssa	CALIFA DR3
CARMENES_Reiners	CARMENES_Reiners2018
carmenes SSAP	CARMENES DR1: Spectra
carmenes TS SSAP	Carmenes DR1 Radial V...
CaT library	CaT library, Empirical C...
CCD700N	CCD700 ver new
CCD700 SSAP	CCD700 Ondrejov Spec...
CDFS SSAP	Optical Spectroscopy l...
CfA Hectochelle	CfA Hectochelle Spectra
CfA Hectospec	CfA Hectospec Spectra
Chandra	Chandra Observations
Chiu2006	L and T dwarf (Chiu et ...
COROT ARCHIVE	The COROT PUBLIC ARC...
CSIRO ASKAP SSA	CSIRO ASKAP Science D...
DCSSA	Data Central SSA
DFBS SSAP	Digitized First Byurakan...
DIGIT	Herschel Dust, Ice, and...
Disks_SH_Spectra	Spitzer Disks Short-Hig...
DK-154 LC	DK-154 objects
DK-154 LCs	DK-154 objects
ELODIE	ELODIE archive
ELODIEinterp	Spectrum interpolator f...
ESO SSAP	ESO 1D reduced spectra
EUVE	Extreme Ultraviolet Exp...
F/H Orders SSAP	Flash/Heros Split-Order...

Search parameters:

Simple Query

Object: trappist-1

RA: 23:06:29.368 Dec: -05:02:29.04

Radius: 5 MAXREC:

Band:

Time:

Query Format: None

Wavelength calibration: None

Flux calibration: None

Optional Parameters

Use	Name	Value	UCD
-----	------	-------	-----

Query: <SERVER>?REQUEST=queryData&POS=346.6223666666666,-5.0414&SIZE=0.08333333333333333

Query results:

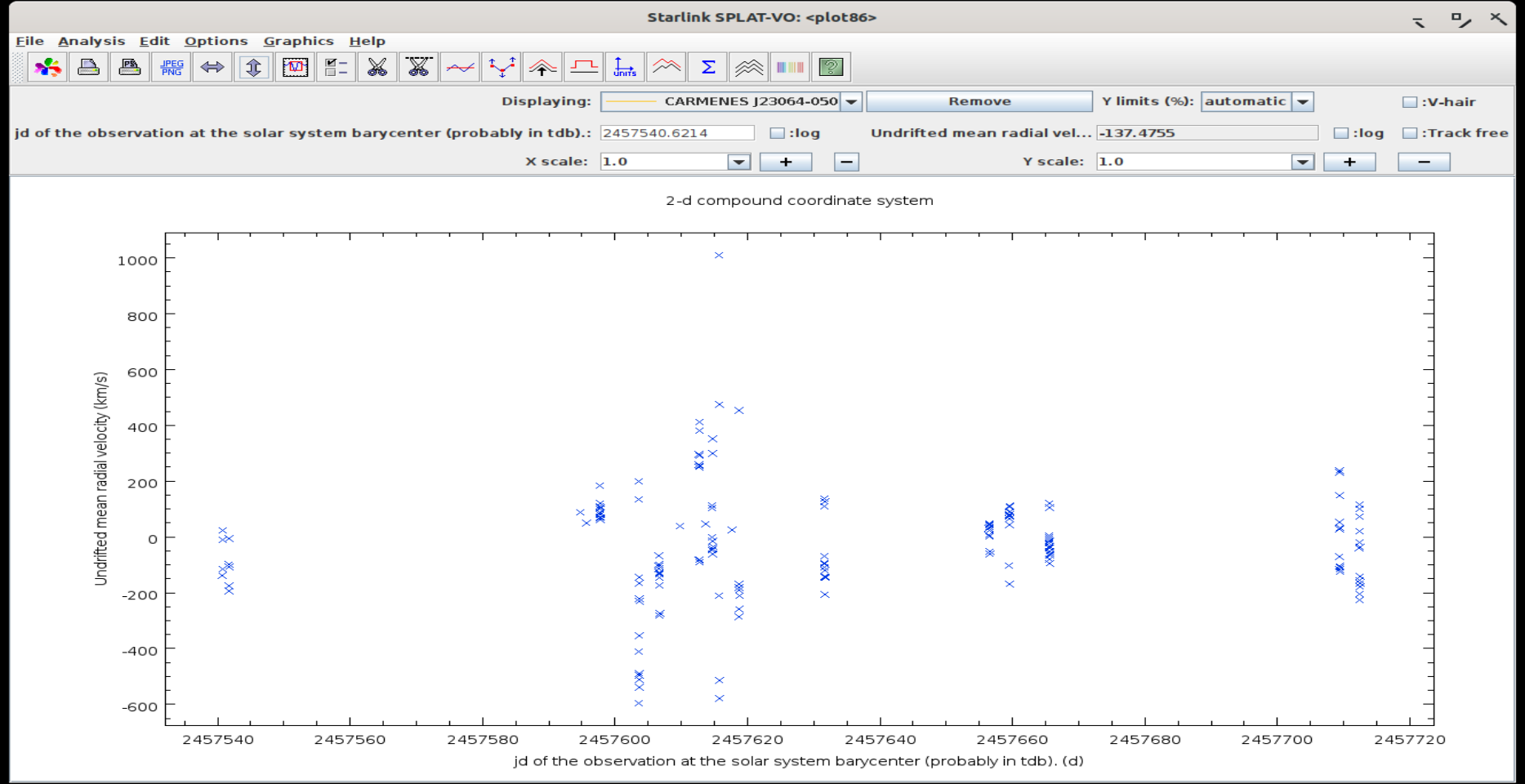
carmenes TS SSAP

...	accref	ssa_dstitle	ssa_targname	ssa_dstype	ssa_timeExt	ssa_length	ssa_snr	ssa_dateObs
1	http://dc.g-vo.org/getproduct/...	CARMENES DR1 J23064-050 RVs	CARMENES J23064-050	timeseries	1.48386E7	163		

SSAP Service Information

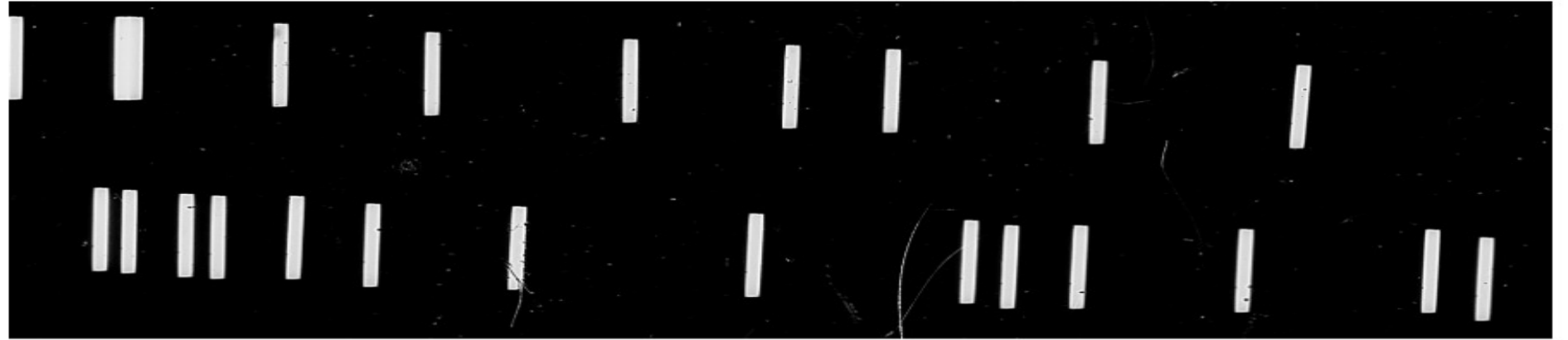
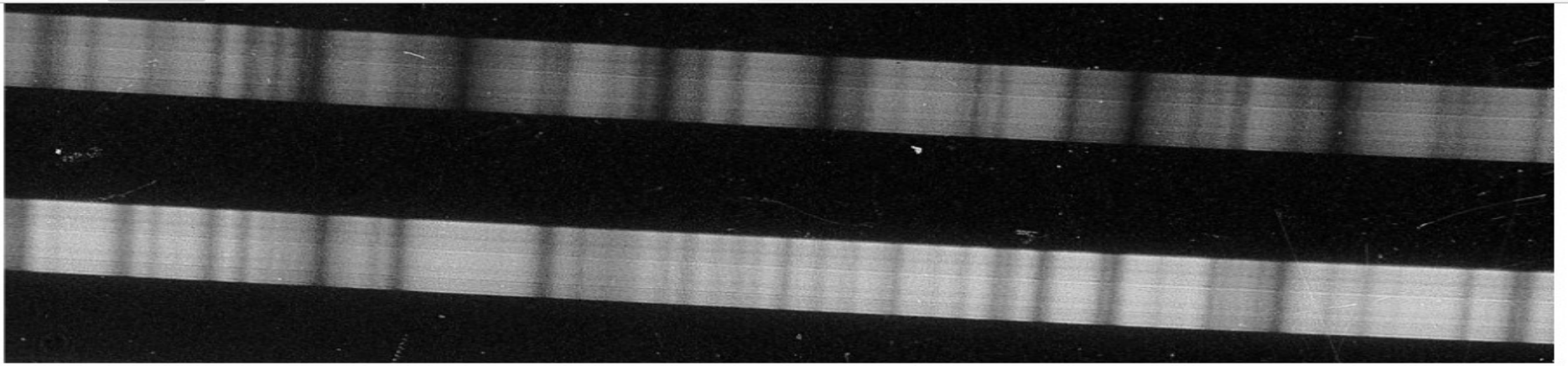
**short name:** carmanes TS SSAP  
**title:** Carmanes DR1 Radial Velocity Time Series  
**description:**  
**identifier:** ivo://org.gavo.dc/carmenes/t/ssa  
**publisher:** The GAVO DC team  
**contact:** GAVO Data Centre Team (gavo@ari.uni-heidelberg.de) (gavo@ari.uni-heidelberg.de)  
**access URL:** http://dc.g-vo.org/carmenes/t/ssa/ssap.xml?  
**reference URL:** http://dc.g-vo.org/carmenes/t/ssa/info  
**waveband:** infrared, optical  
**content type:** archive  
**data source:** pointed  
**creation type:** archival

# Trappist-1 RV Curve - Carmones



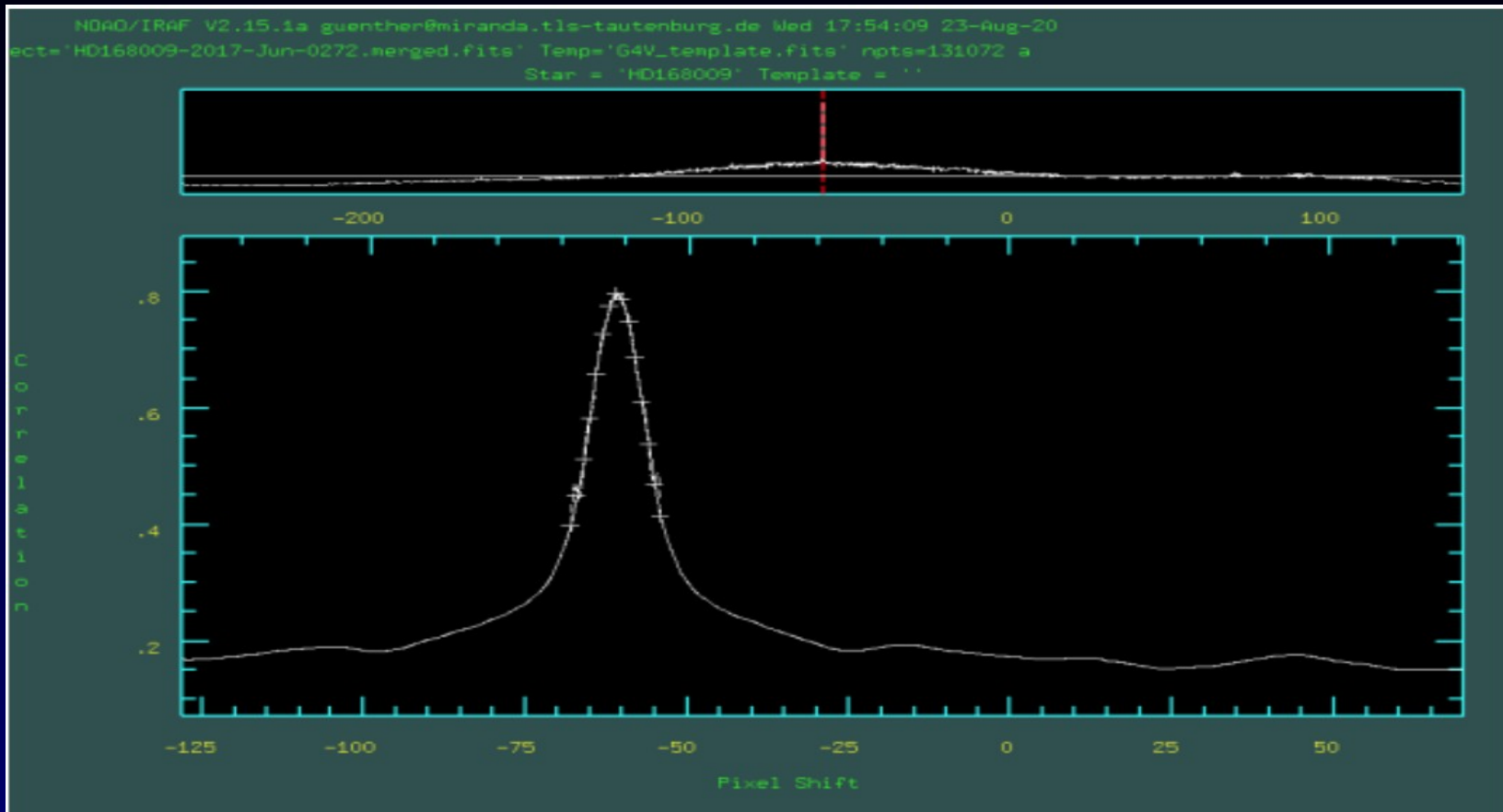
# Cross Correlation Function

# Analogue Technique



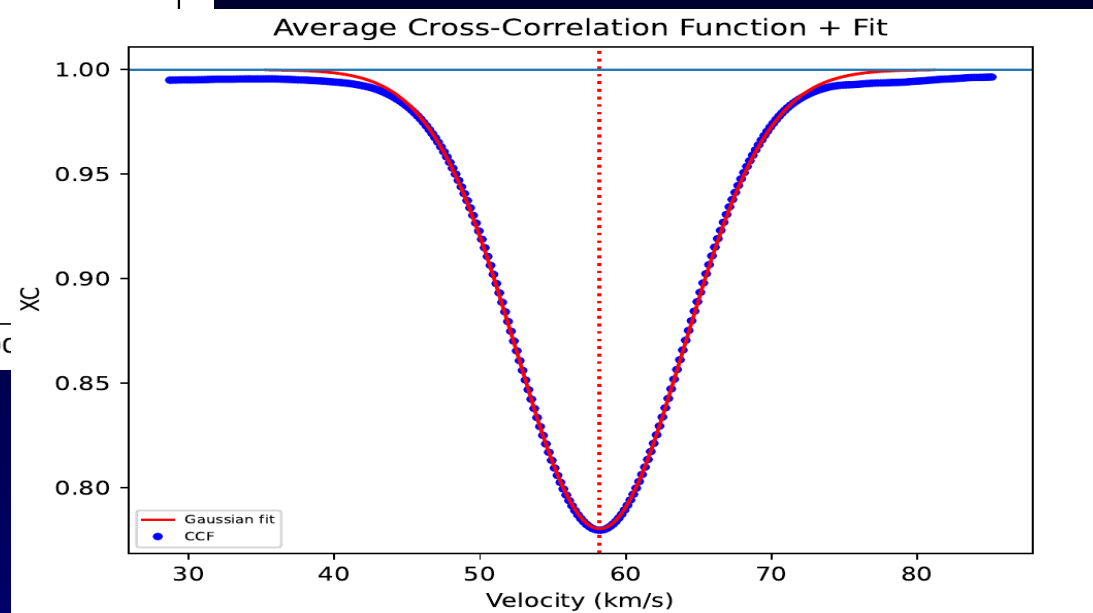
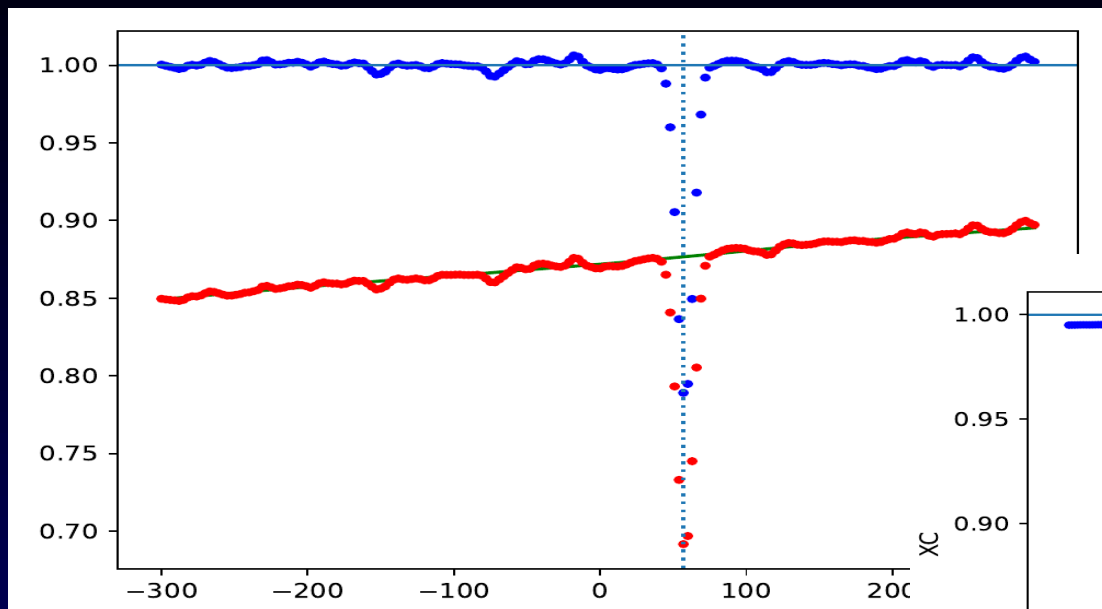
First principle Griffin 1967, DAO 122, CORAVEL ESO D154 1981

# Cross Correlation Function



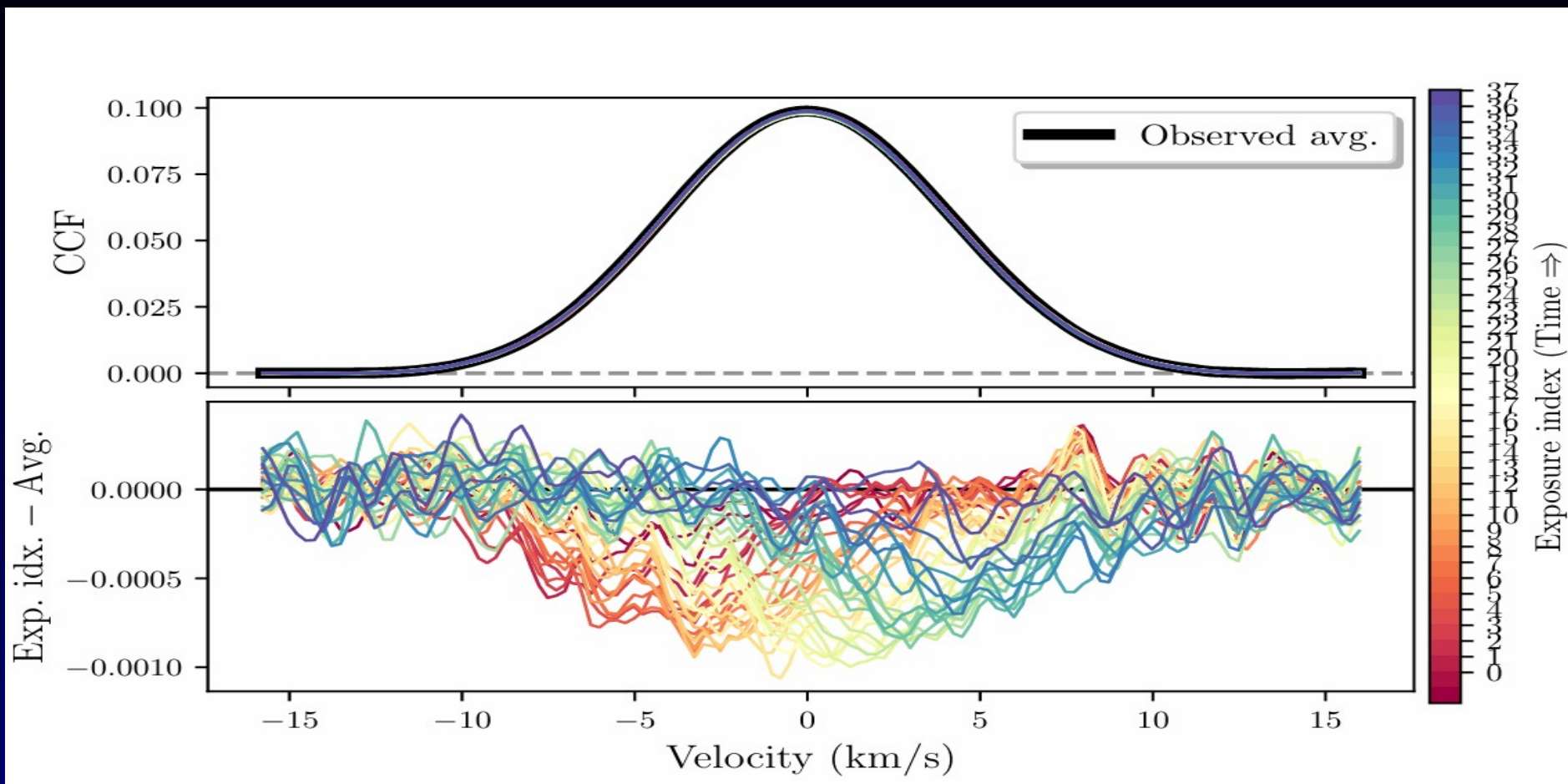
IRAF fxcor task

# Cross Correlation Function - CCF

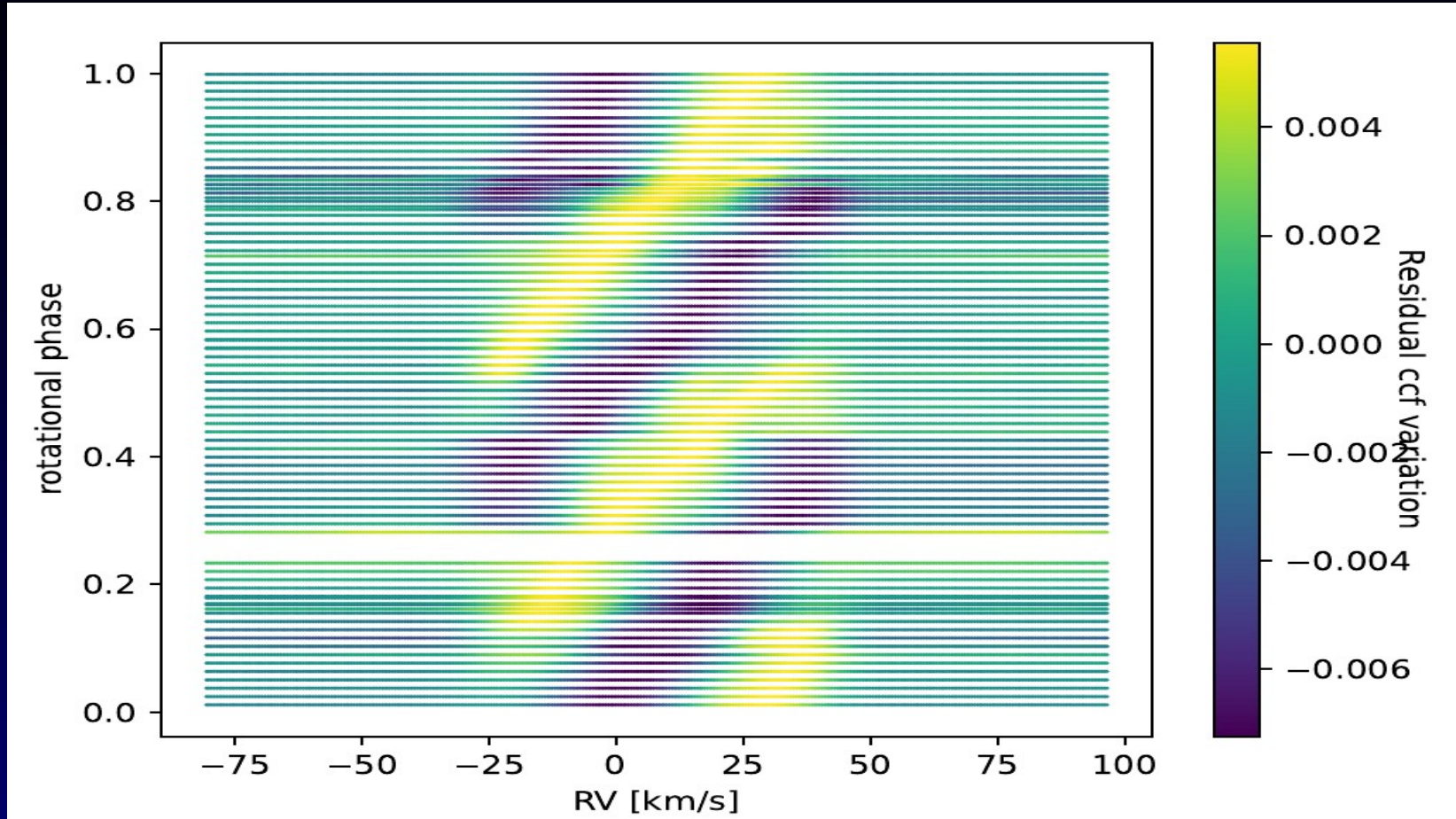


Example of PlatoSpec data reduced by CERES+

# Residual CCF



# Dynamic spectra – CCF



# EPRV Standard

About 15 spectrographs ( ~ 5 passively following)

- Translators from native format
- Started 2024
- NASA JPL (EPRV project - WG )

<https://github.com/EPRV-RCN/RVData>

<https://eprv-data-standard.readthedocs.io/en/latest>

# EPRV L2 – Order by Order Unblazed

HDU	Name	DataType	MinBetDepth	Multiplicity	Required	Description
0	PRIMARY	PrimaryHDU		False	True	EPRV Standard FITS HEADER (no data)
1	INSTRUMENT_HEADER	ImageHDU		False	True	Inherited instrument header (no data)
2	RECEIPT	BinTableHDU		False	True	Table of operations that have been performed on this file
3	DRP_CONFIG	BinTableHDU		False	True	Pipeline details (settings etc) to go from native data to L2
4	EXT_DESCRIPTOR	BinTableHDU		False	True	Table describing contents of each extension
5	ORDER_TABLE	BinTableHDU		TRUE	TRUE	Table capturing the wavelength extent of each order
6	TRACE1_FLUX	ImageHDU		True	True	Flux in trace 1
7	TRACE1_WAVE	ImageHDU	64	True	True	wavelength solution for trace 1
8	TRACE1_VAR	ImageHDU		True	True	variance for trace 1
9	TRACE1_BLAZE	ImageHDU		True	True	blaze for trace 1
10	BARYCORR_KMS	ImageHDU		False	True	barycentric correction in km/s
11	BARYCORR_Z	ImageHDU		False	True	barycentric correction in redshift
12	BJD_TDB	ImageHDU	64	False	True	Photon weighted midpoint
13	DRIFT	ImageHDU		False	False	Drift measurement map in delta lambda
14	TRACE1_QUALITY	ImageHDU	UINT8	True	False	Quality of each pixel in trace 1
15	EXPMETER	BinTableHDU		False	False	Table of exposure meter counts timeseries over the exposure
16	TELEMETRY	BinTableHDU		False	False	Table of telemetry collected during the exposure
17	TRACE1_TELLURIC	ImageHDU		True	False	Telluric model for trace 1
18	TRACE1_SKYMODEL	ImageHDU		True	False	Sky model for trace 1
19	ANCILLARY_SPECTRUM	ImageHDU		True	False	Extension(s) that store ancillary spectra
20	IMAGE	ImageHDU		True	False	Extension(s) that store useful support images
21	CUSTOM1_TRACE1_FLUX	ImageHDU		True	False	Additionally corrected flux in trace 1
22	CUSTOM1_TRACE1_WAVE	ImageHDU		True	False	Additionally corrected wavelength solution for trace 1
23	CUSTOM1_TRACE1_VAR	ImageHDU		True	False	Additionally corrected variance for the flux in trace 1

# EPRV L3 – Merged orders normalized

This is what we already know – SSAP, “distorted data low precision”

HDU	Name	DataType	Multiplicity	Required	Description
0	PRIMARY	PrimaryHDU	False	True	EPRV Standard FITS HEADER (no data)
1	INSTRUMENT_HEADER	ImageHDU	False	True	Inherited instrument header (no data)
2	RECEIPT	BinTableHDU	False	True	Table of operations that have been performed on this file
3	DRP_CONFIG	BinTableHDU	False	True	Pipeline details (settings etc) to go from native data to L2
4	ORDER_TABLE	BinTableHDU	TRUE	TRUE	Table capturing the wavelength extent of each order
5	STITCHED_CORR_TRACE1_FLUX	ImageHDU	True	True	Order stitched blaze-corrected flux in trace 1
6	STITCHED_CORR_TRACE1_WAVE	ImageHDU	True	True	Order stitched BC- and drift-corrected wavelength solution for trace 1
7	STITCHED_CORR_TRACE1_VAR	ImageHDU	True	True	Order stitched variance for the flux in STITCHED_CORR_TRACE1_FLUX
8	COMBINED_STITCHED_CORR_FLUX	ImageHDU	True	True	Order stitched and blaze-corrected flux co-added across all traces
9	COMBINED_STITCHED_CORR_WAVE	ImageHDU	True	True	Order stitched BC- and drift-corrected wavelength solution
10	COMBINED_STITCHED_CORR_VAR	ImageHDU	True	True	Order stitched variance for the combined flux in COMBINED_STITCHED_CORR_FLUX
11	STITCHED_CUSTOMCORR1_TRACE1_FLUX	ImageHDU	True	False	Additional corrections made to STITCHED_CORR_TRACE1_FLUX
12	STITCHED_CUSTOMCORR1_TRACE1_WAVE	ImageHDU	True	False	Wavelength solution corresponding to STITCHED_CUSTOMCORR1_TRACE1_FLUX
13	STITCHED_CUSTOMCORR1_TRACE1_VAR	ImageHDU	True	False	Variance corresponding to STITCHED_CUSTOMCORR1_TRACE1_FLUX
14	COMBINED_STITCHED_CUSTOMCORR1_FLUX	ImageHDU	True	False	Additional corrections made to COMBINED_STITCHED_CORR_FLUX
15	COMBINED_STITCHED_CUSTOMCORR1_WAVE	ImageHDU	True	False	Wavelength solution corresponding to COMBINED_STITCHED_CORR_WAVE
16	COMBINED_STITCHED_CUSTOMCORR1_VAR	ImageHDU	True	False	Variance corresponding to COMBINED_STITCHED_CORR_VAR

# EPRV L4 – Derived core exo science

Radial Velocity and Cross Correlation Function (various masks ....)

It is the most important output of pipeline

But exoscientists need Time Series (RV curves , Doppler shadows)

VO can do it on the fly (select all observation, apply SODA extract , create time series)

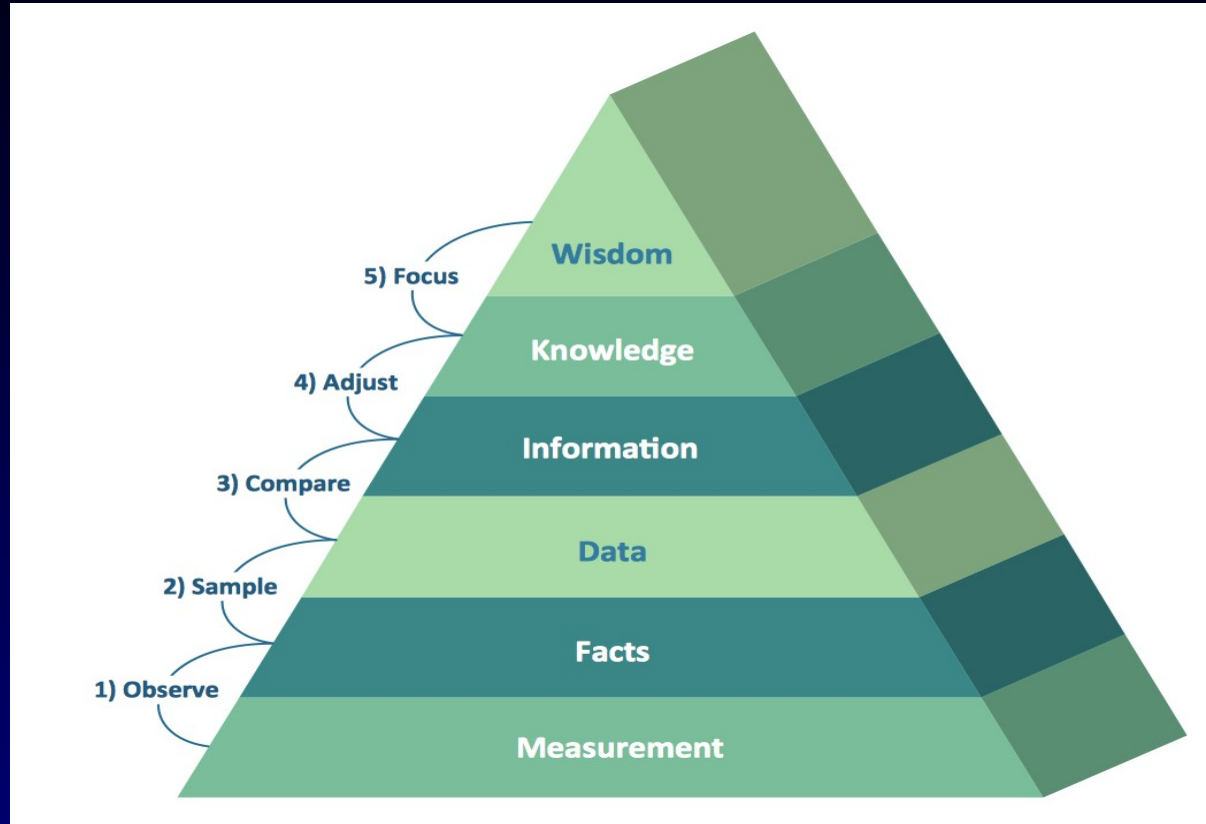
HDU	Name	DataType	Multiplicity	Required	Description
0	PRIMARY	PrimaryHDU	FALSE	TRUE	EPRV Standard FITS HEADER (no data)
1	INSTRUMENT_HEADER	ImageHDU	FALSE	TRUE	Inherited instrument header (no data)
2	RECEIPT	BinTableHDU	FALSE	TRUE	Table of operations that have been performed on this file
3	DRP_CONFIG	BinTableHDU	FALSE	TRUE	Pipeline details (settings etc) to go from native data to L2
4	RV1...N	BinTableHDU	TRUE	TRUE	Derived Radial Velocity Measurement
5	CCF1...N	ImageHDU	TRUE	FALSE	Array with same dimensionality of RV1...N, that contains the CCF that produced each RV1...N
6	DIAGNOSTICS1...J	BinTableHDU	TRUE	FALSE	Activity indicators, CCF metrics, etc
7	CUSTOM_CCF1...N	ImageHDU	TRUE	FALSE	Additional CCFs from (e.g.) different masks
8	CUSTOM_RV1...N	BinTableHDU	TRUE	FALSE	Derived Radial Velocity Measurement from CUSTOM_CCF1...N

***Astroinformatics***

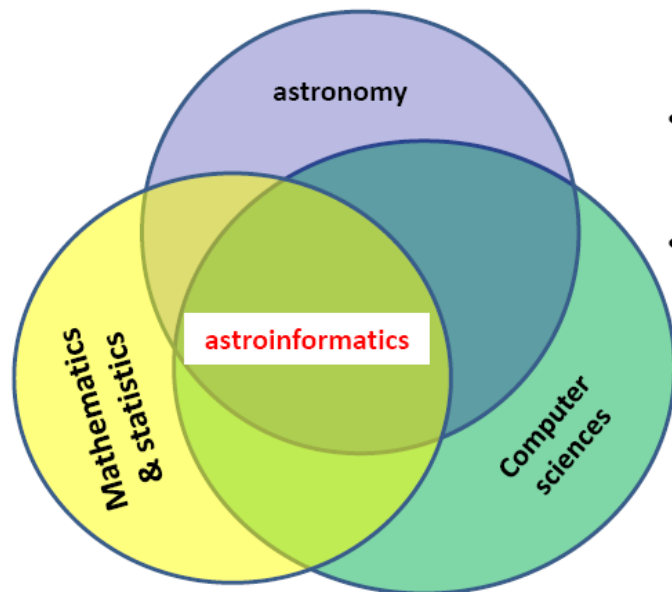
# Data-Knowledge-Wisdom Pyramid

## CRISP-DM

Cross-Industry Standard Process for Data Mining



## 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

# 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*

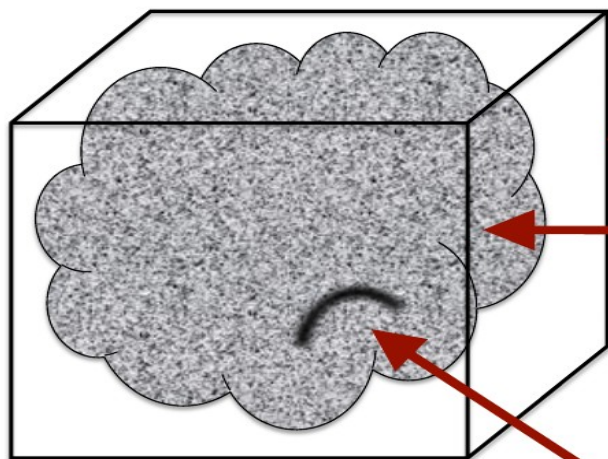
- A great increase in the *information complexity*

➔ *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

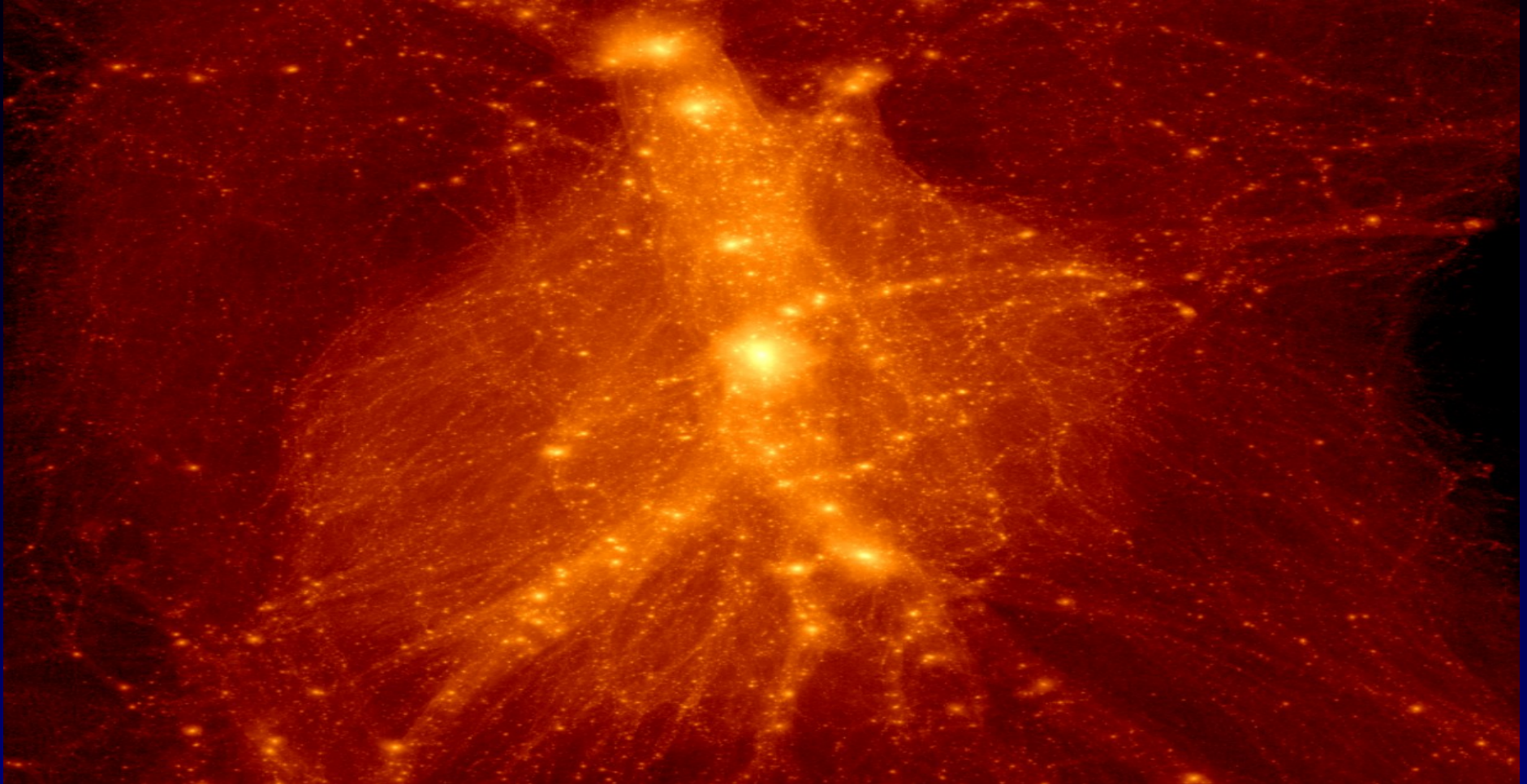


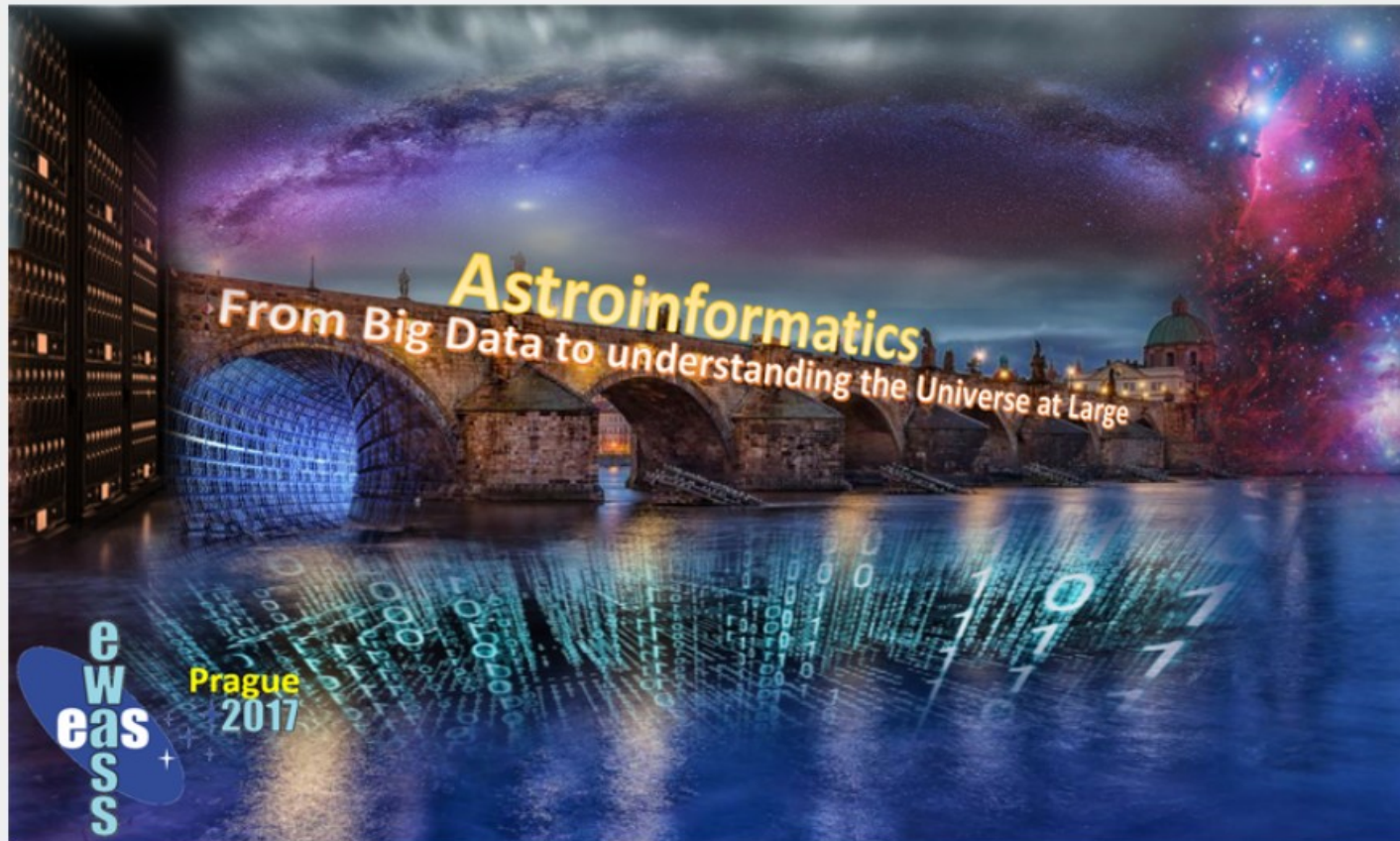
$D \gg 3$  parameter space hypercube

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 of Big Data



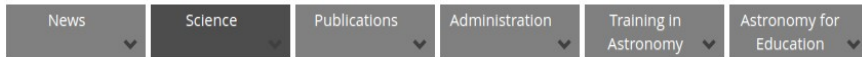


# Astroinformatics in IAU



INTERNATIONAL  
ASTRONOMICAL  
UNION

[Home](#) | [About IAU](#) | [IAU Values](#) | [Donate](#) | [Member Directory](#) | [Site Map](#) | [Contact Us](#)



[Home](#) / [Science](#) / [Scientific Bodies](#) / [Commissions](#) / [Commission B3 Structure](#) » [Commission B3 Astroinformatics and Astrostatistics](#)

## B3 – Commission B3 Astroinformatics and Astrostatistics

### Description

In most of the 20th century, astronomers investigated cosmic phenomena by careful study of individual objects or small samples of planets, stars, galaxies and diffuse media. Datasets were often modest in size with zero (photometry), one (spectra, light curves), or two (images) dimensions. But in the 21st century, increasing resources are devoted to wide-field astronomical surveys, three- or multi-dimensional data, and high-throughput instruments that produce peta-scale datasets and giga-scale samples. In addition to the growing tasks of data reduction, science analysis is becoming more complex. Astronomical insights require characterizing structure in images, spectra or time series. Astrophysical insights require fitting nonlinear, sometimes high-dimensional models to data. Modeling involves both small and large datasets.

IAU Commission B3 focuses on the statistical, computational methodological challenges arising in the various fields of astronomy. It assists the astronomical community in learning existing, and developing new, advanced methodologies to accomplish its goals in this changing environment. The Commission encourages liaison with professional communities in the fields of statistics, applied mathematics and computer science, and with private enterprises. It sponsors meetings and discussions to promulgate advanced methodologies to seek the best scientific insights from the growing flow of data.

[Commission Web Page](#)

[Commission Members \(270\)](#)



INTERNATIONAL  
ASTRONOMICAL  
UNION

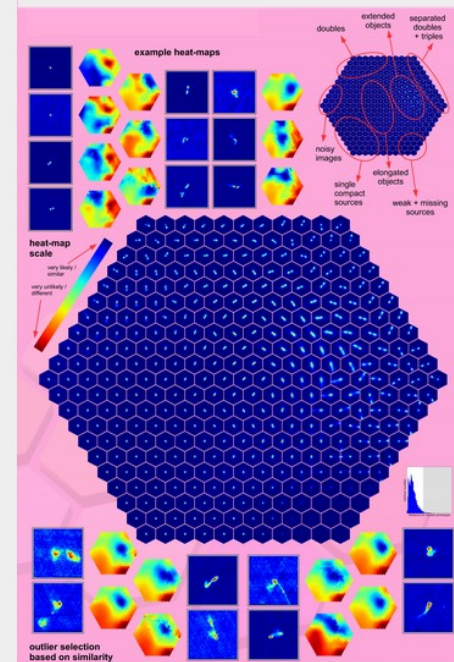
[Home](#) | [About IAU](#) | [IAU Values](#) | [Donate](#) | [Member Directory](#) | [Site Map](#) | [Contact Us](#) | [Login](#)



[Home](#) / [Science](#) / [Scientific Bodies](#) / [Commissions](#) / [Commission B3 Structure](#) / [Commission B3 Homepage](#)

## Commission B3 Astroinformatics and Astrostatistics

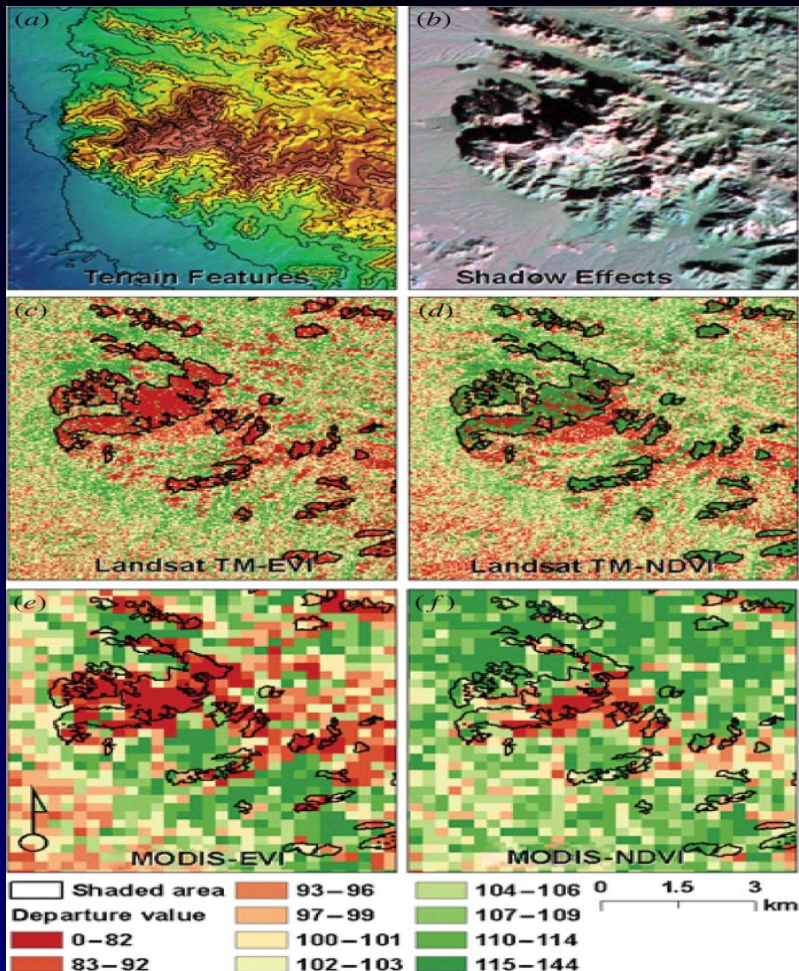
[Scientific Objectives](#) | [Members](#) | [News](#) | [Meetings](#) | [Documents](#) | [Useful Resources](#)



The unsupervised morphological classification of 200 000 Radio-Galaxy Zoo images using self-organizing Kohonen map (Polsterer, Geske, Igel, 2015ASPC..495..81P). Credit: Kai Lars Polsterer.

Please send any updated email addresses or changes of institutes to the IAU Secretariat at: [iauisfco@iap.fr](mailto:iauisfco@iap.fr).

# Remote sensing – Big Data Machine Learning



Precise farming

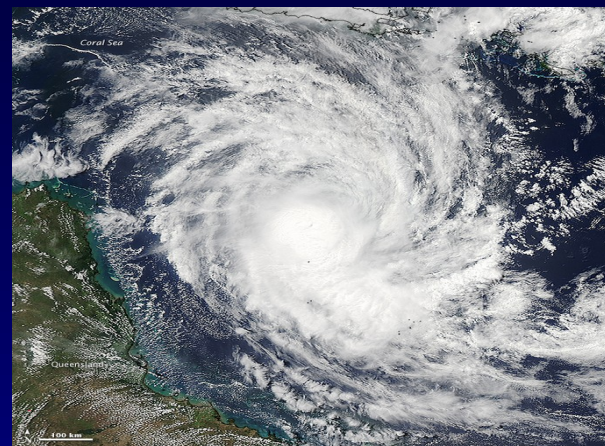
Forestry

Ore mining

Water resources monitoring

Automatic classification of terrain

Resistance of buildings (Aquila)



# Visualization of Big Data

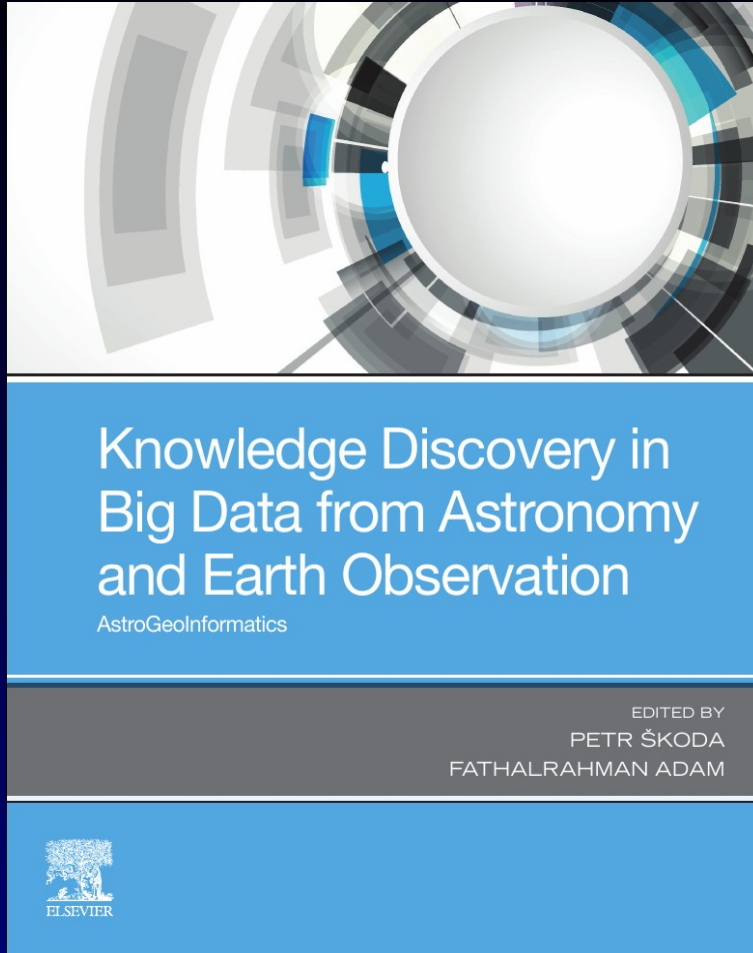


# CAVE2 Monash University AU



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

# AstroGeoInformatics



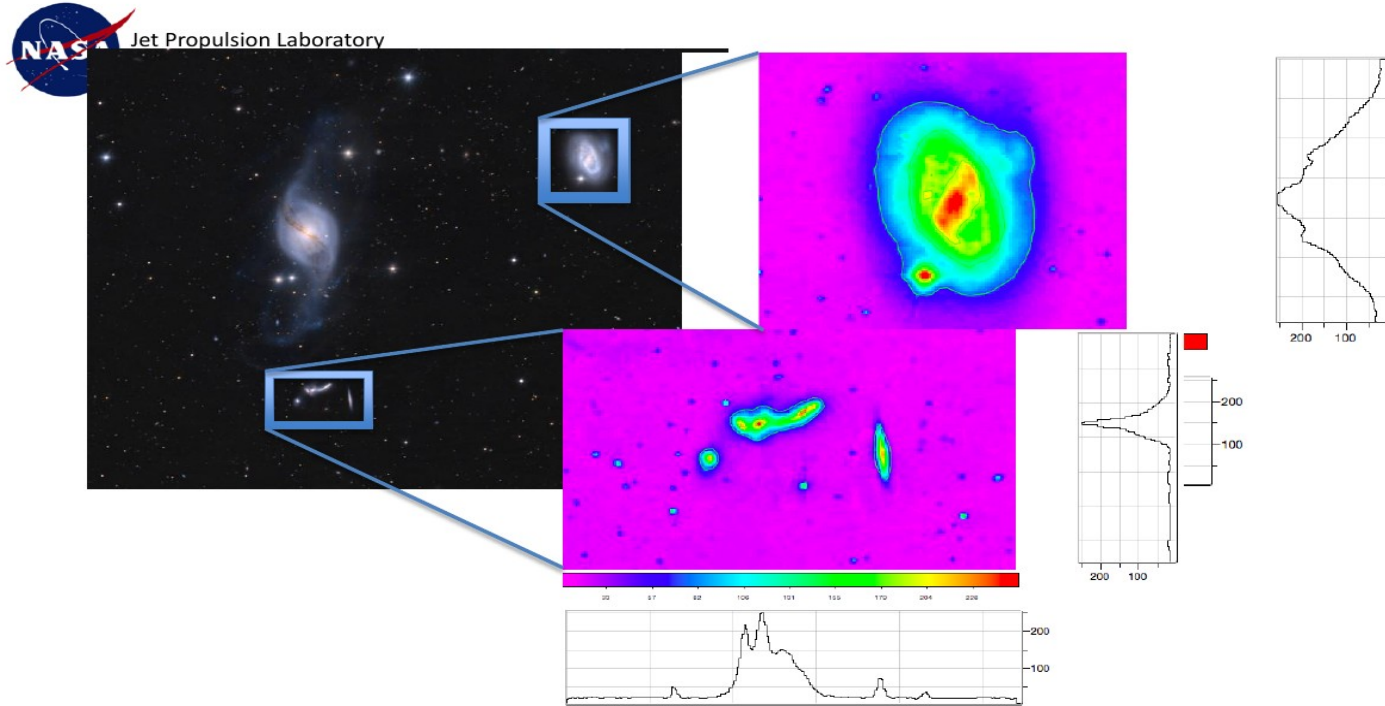
COST Transdisciplinary

BIG-SKY-EARTH TD1403

~55 authors

472 pages

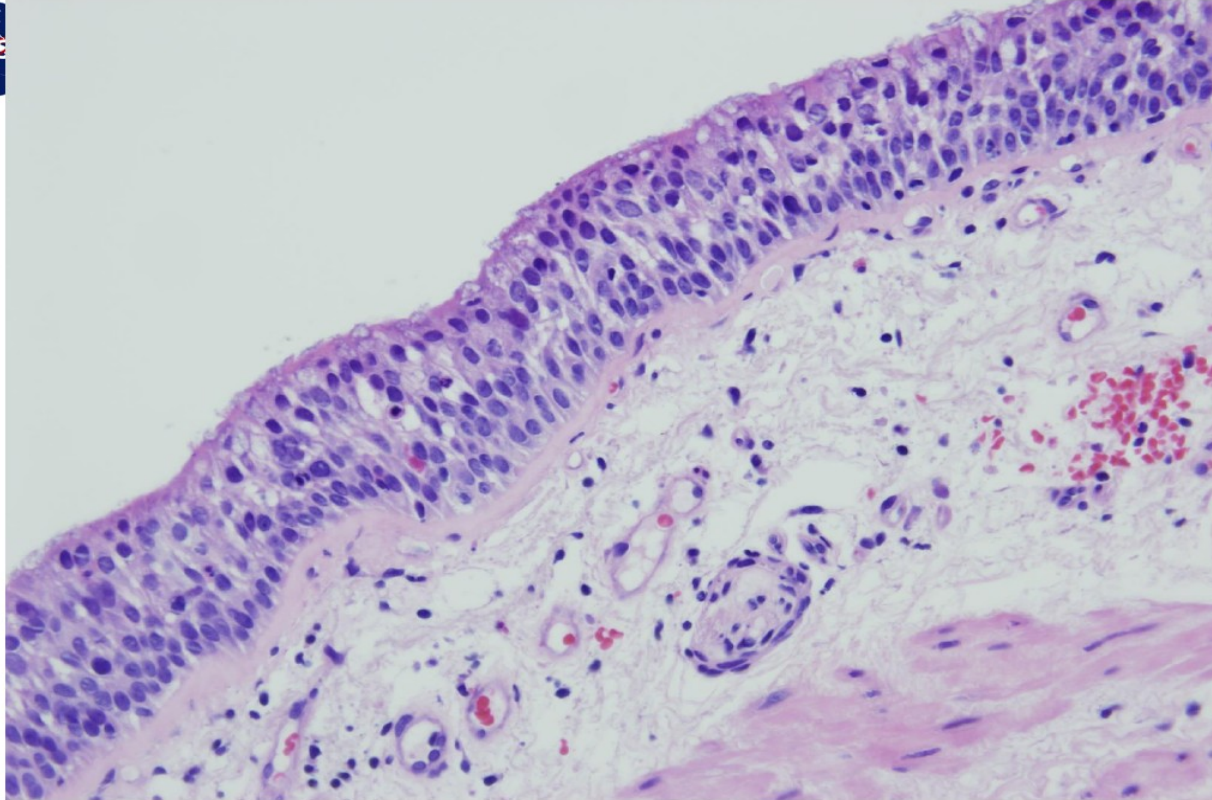
# Finding Galaxies by Shape NASA



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.

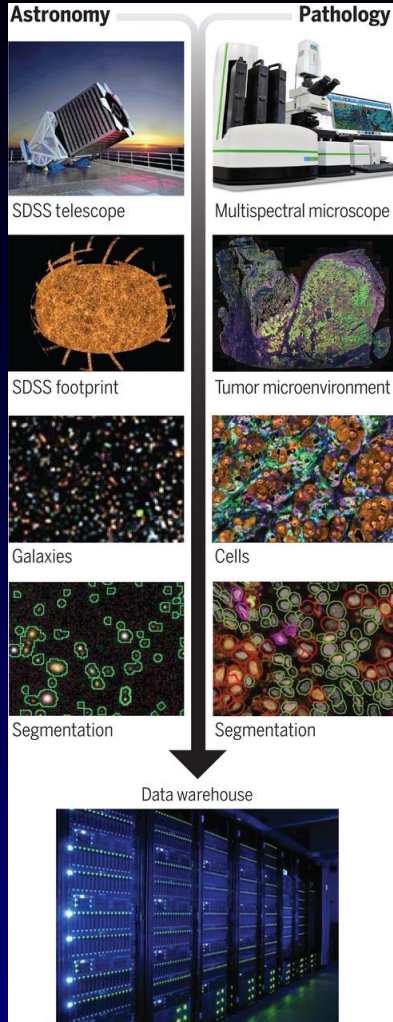
# Finding Cancer Signatures NASA



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

# Digital Pathology based on SDSS



## CLINICAL CANCER RESEARCH

ABOUT ▾ ARTICLES ▾ FOR AUTHORS ▾ ALERTS NEWS CANCER HALLMARKS WEBINARS

Volume 28, Issue 16  
15 August 2022

REVIEW | AUGUST 15 2022

### Data-Rich Spatial Profiling of Cancer Tissue: Astronomy Informs Pathology

FREE

Alexander S. Szalay ; Janis M. Taube

Check for updates

+ Author & Article Information

*Clin Cancer Res* (2022) 28 (16): 3417–3424.

<https://doi.org/10.1158/1078-0432.CCR-19-3748> [Article history](#)

Science

HOME > SCIENCE > VOL. 3

RESEARCH ARTICLE



## Analysis of multispectral imaging with the AstroPath platform informs efficacy of PD-1 blockade

SNEHA BERRY , NICOLAS A. GIRALDO , BENJAMIN F. GREEN, TRICIA R. COTTRELL , JULIE E. STEIN , ELIZABETH L. ENGLE, HAIYING XU,

ALEKSANDRA OGURTSOVA, CHARLES ROBERTS, [...], AND JANIS M. TAUBE [+27 authors](#) [Authors Info & Affiliations](#)

SCIENCE · 11 Jun 2021 · Vol 372, Issue 6547 · DOI: 10.1126/science.aba2609

9,621 56



### Astronomy accelerates tumor imaging



# *Artificial Intelligence*

# Human – Machine Collaboration

## From the Information Technology to the Cognition Technology: Towards a Human-Computer Collaborative Discovery

**AS WE MAY THINK**  
A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLD  
IN WHICH MAN-MADE MACHINES WILL START TO THINK  
by VANNEVAR BUSH



Man-Computer Symbiosis

J.C.R. Licklider  
(1960)



Vannevar Bush  
(1945)

MEMEX

„Hypertext“  
Associations

Global network  
Interactivity  
Graphics

# Turing Test

VOL. LIX. NO. 236.]

[October, 1950

M I N D

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY



**I.—COMPUTING MACHINERY AND  
INTELLIGENCE**

BY A. M. TURING

1. *The Imitation Game.*

# Imitation Game

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end

Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be

'My hair is shingled, and the longest strands are about nine inches long.'

question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as 'I am the woman, don't listen to him!' to her answers, but it will avail nothing as the man can make similar remarks.

# Turing Test – Core Idea

We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?'

Strong AI (passes TT)   x   Weak AI  
Searle 1980 – Chinese Room

Turing prize ~ Nobel prize in IT  
1954 suicide by apple with cyanide



# Origins of term „Artificial Intelligence“

IN THIS BUILDING DURING THE SUMMER OF 1956

JOHN McCARTHY (DARTMOUTH COLLEGE), MARVIN L. MINSKY (MIT)  
NATHANIEL ROCHESTER (IBM), AND CLAUDE SHANNON (BELL LABORATORIES)  
CONDUCTED

THE DARTMOUTH SUMMER RESEARCH PROJECT  
ON ARTIFICIAL INTELLIGENCE

FIRST USE OF THE TERM "ARTIFICIAL INTELLIGENCE"

FOUNDING OF ARTIFICIAL INTELLIGENCE AS A RESEARCH DISCIPLINE

"To proceed on the basis of the conjecture  
that every aspect of learning or any other feature of intelligence  
can in principle be so precisely described that a machine can be made to simulate it."

IN COMMEMORATION OF THE PROJECT'S 50th ANNIVERSARY  
JULY 13, 2006

## 1956 Dartmouth Conference: The Founding Fathers of AI



John McCarthy



Marvin Minsky



Claude Shannon



Ray Solomonoff



Alan Newell



Herbert Simon



Arthur Samuel



Oliver Selfridge

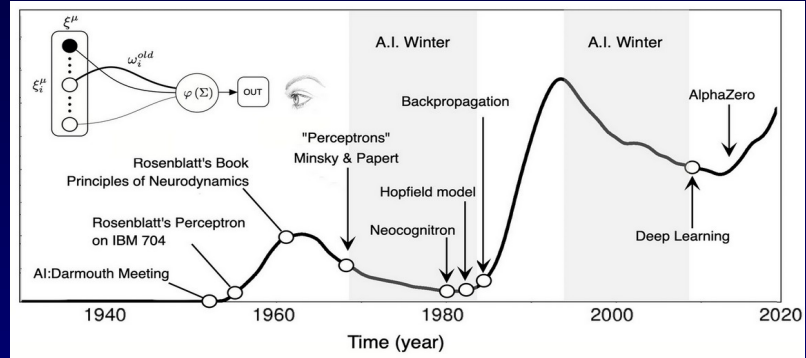


Nathaniel Rochester



Trenchard More

Founding fathers of AI. Courtesy of [scienceabc.com](https://www.scienceabc.com)



# First Chatbot - ELIZA

Weizenbaum MIT 1964-67

```
Did you come to me because you are not ?
```

```
>i dont know  
What does that suggest to you?
```

```
>that i am unintelligent  
How long have you been unintelligent?
```

```
>since i was born  
I see.
```

```
>arent you supposed to cheer me up  
You're not really talking about me, are you?
```

```
>no im talking about me  
Why not?
```

```
>mrh  
I'm not sure I understand you fully.
```

```
>rg  
Come come elucidate your thoughts.
```

```
>what is elucidate  
What do you think?
```

```
>to see  
Can you elaborate on that?
```

```
>
```

Welcome to

```
EEEEEE LL      IIII  ZZZZZZ  AAAAA  
EE      LL      II     ZZ     AA  AA  
EEEEEE LL      II     ZZZ     AAAAAAA  
EE      LL      II     ZZ     AA  AA  
EEEEEE LLLLLL  IIII  ZZZZZZ  AA  AA
```

Eliza is a mock Rogerian psychotherapist.

The original program was described by Joseph Weizenbaum in 1966.  
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?

YOU: Men are all alike.

ELIZA: What is the connection, do you suppose ?

YOU: They're always bugging us about something or other.

ELIZA: Can you think of a specific example ?

YOU: Well, my boyfriend made me come here.

ELIZA: Is it important to you that your boyfriend made you come here ?

YOU: He says I'm depressed much of the time.

ELIZA: I am sorry to hear that you are depressed.

YOU: It's true. I am unhappy.

ELIZA: Can you explain what made you unhappy ?

YOU:

# Recent Advanced Neural Networks

Reinforcement Learning

Active learning (oracle)

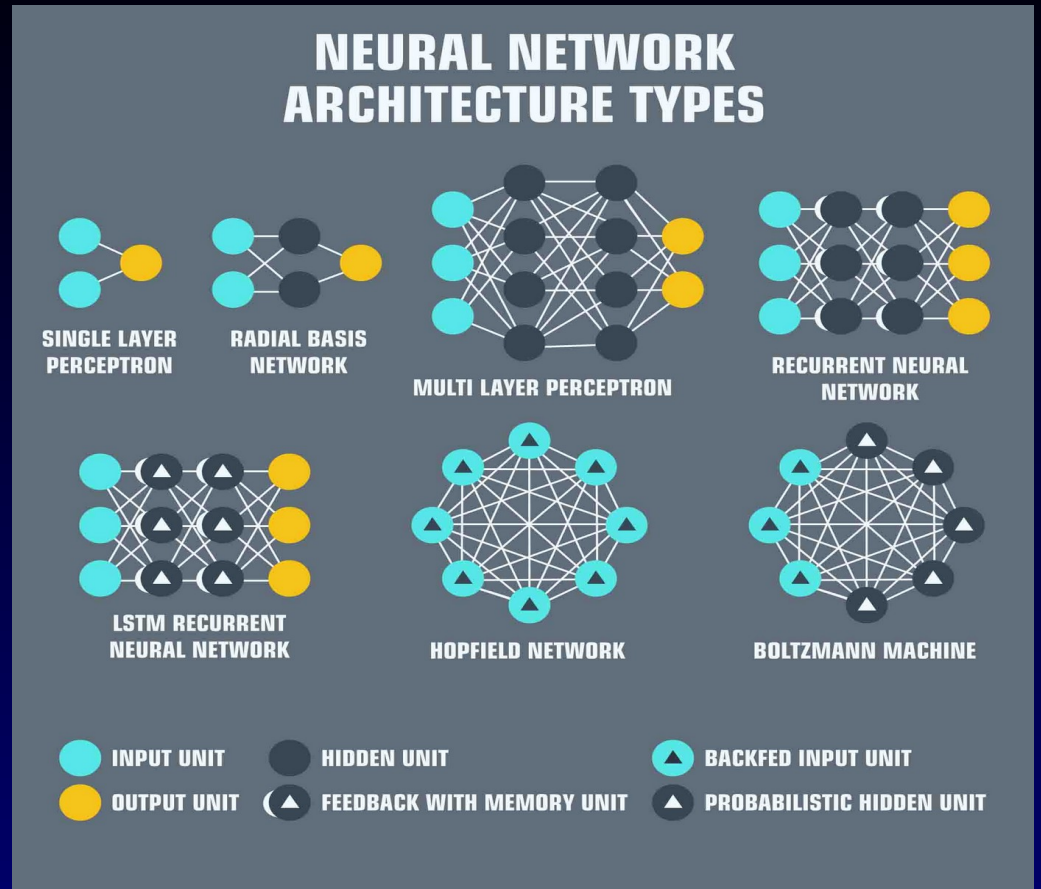
Bayesian Deep Learning

Physics Informed (Aware) NN

GANs (image creation)

Diffusion models

Foundation models



# Reviews of ML/AI in Astronomy

DRAFT VERSION APRIL 17, 2019  
Typeset using L<sup>A</sup>T<sub>E</sub>X preprint2 style in AASTeX61

## MACHINE LEARNING IN ASTRONOMY: A PRACTICAL OVERVIEW

DALYA BARON<sup>1</sup>

<sup>1</sup>*School of Physics and Astronomy  
Tel-Aviv University  
Tel Aviv 69978, Israel*

To appear in: *Artificial Intelligence for Science*,  
eds. A. Choudhary, G. Fox and T. Hey  
Singapore: World Scientific, in press (2023)

## Applications of AI in Astronomy

S. G. Djorgovski\*, A. A. Mahabal\*, M. J. Graham\*, K. Polsterer<sup>†</sup>,  
A. Krone-Martins<sup>‡</sup>

Experimental Astronomy (2022) 53:1–43  
<https://doi.org/10.1007/s10686-021-09827-4>

REVIEW ARTICLE



## Astronomical big data processing using machine learning: A comprehensive review

Snigdha Sen<sup>1,2</sup> · Sonali Agarwal<sup>1</sup> · Pavan Chakraborty<sup>1</sup> ·  
Krishna Pratap Singh<sup>1</sup>

Received: 15 July 2021 / Accepted: 27 December 2021 / Published online: 14 January 2022  
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Astronomy and Computing 48 (2024) 100851

Contents lists available at ScienceDirect



ELSEVIER

## Astronomy and Computing

journal homepage: [www.elsevier.com/locate/ascom](http://www.elsevier.com/locate/ascom)

## A review of unsupervised learning in astronomy

S. Fotopoulou

*School of Physics, HH Wills Physics Laboratory, University of Bristol, Tyndall Avenue, Bristol, BS8 1TL, United Kingdom*

# Idea of LLMs

Tomáš Mikolov

Dissertation FIT VUT Brno 2012,  
MS Research, Google Brain, Facebook, CIIRC

**word2vec**      sentence as a vector in big-N space  
improves Google Translate

**Fasttext.cc**

<https://www.seznamzpravy.cz/clanek/jeho-figl-v-usa-nakopl-vyvoj-strojoveho-uceni-proc-se-expert-vratil-176156>

# Origin of the Current AI Boom

---

## Attention Is All You Need

---

**Ashish Vaswani\***  
Google Brain  
avaswani@google.com

**Noam Shazeer\***  
Google Brain  
noam@google.com

**Niki Parmar\***  
Google Research  
nikip@google.com

**Jakob Uszkoreit\***  
Google Research  
usz@google.com

**Llion Jones\***  
Google Research  
llion@google.com

**Aidan N. Gomez\* †**  
University of Toronto  
aidan@cs.toronto.edu

**Lukasz Kaiser\***  
Google Brain  
lukaszkaizer@google.com

**Illia Polosukhin\* ‡**  
illia.polosukhin@gmail.com

# Attention

She is eating a green apple



High attention

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

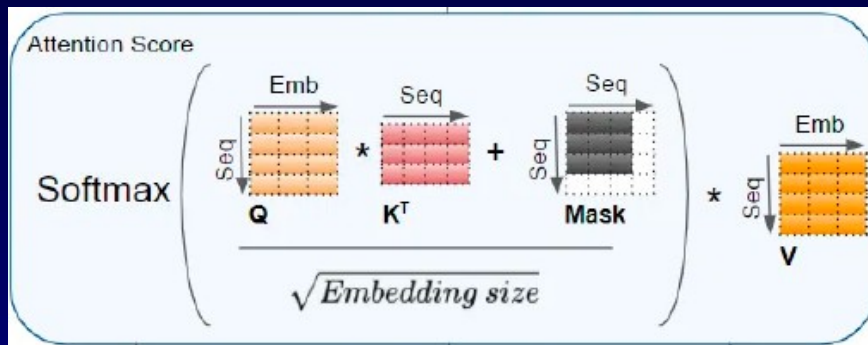
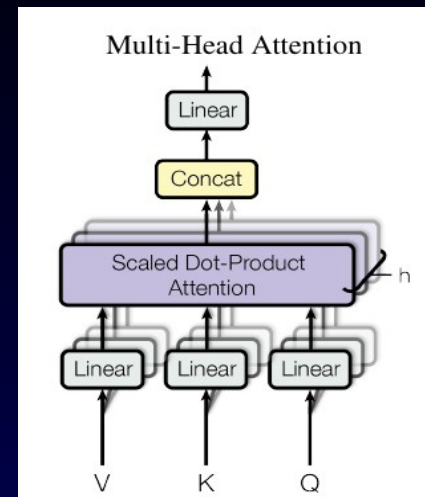
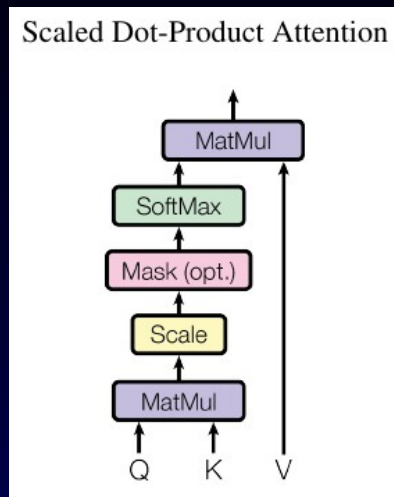
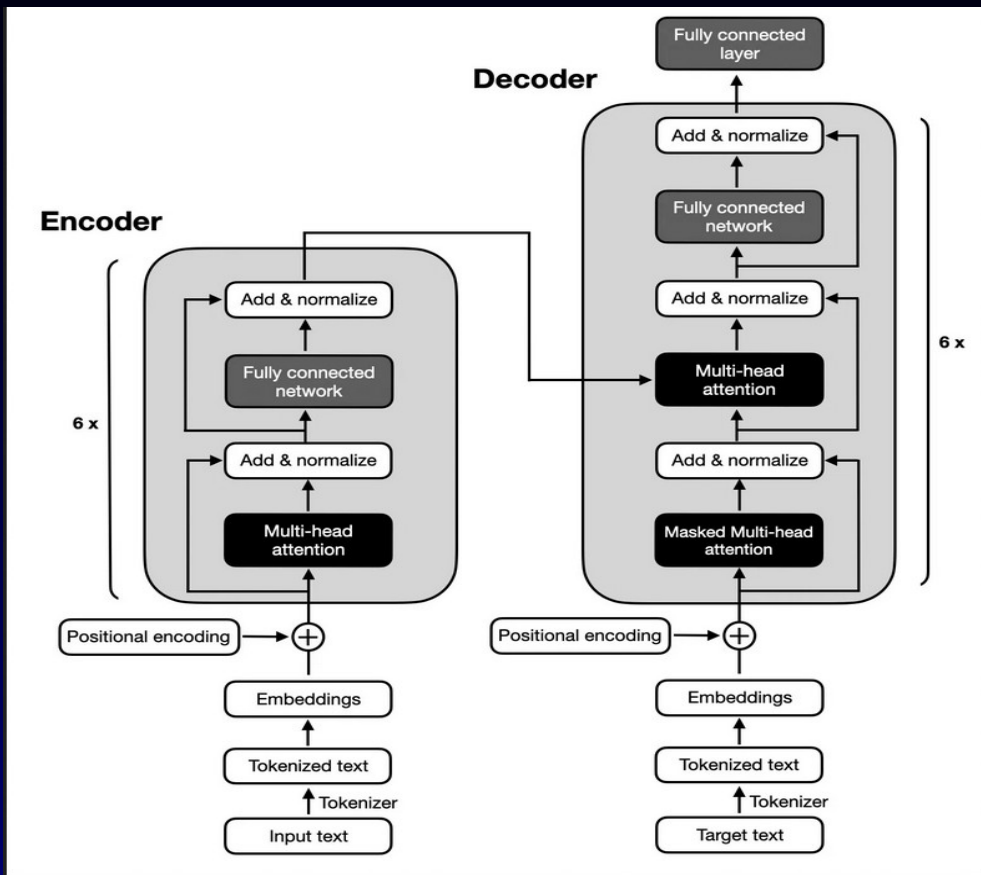
“The cat drank the milk because it was hungry”

“The cat drank the milk because it was sweet”



# Transformer

## (GPT=Generative Pre-trained Transformer)



# LLMs – Recent 2026

2026 LifeArchitect.ai data (shared) - NEW

Soubor Upravit Zobrazit Vložit Formát Data Nástroje Rozšíření Nápověda

100% Pouze zobrazení

A1 `=("(&COUNTIFS('Models Table'!M3:M, "<>TBA", 'Models Table'!M3:M, "<>")&") Permalink:"`

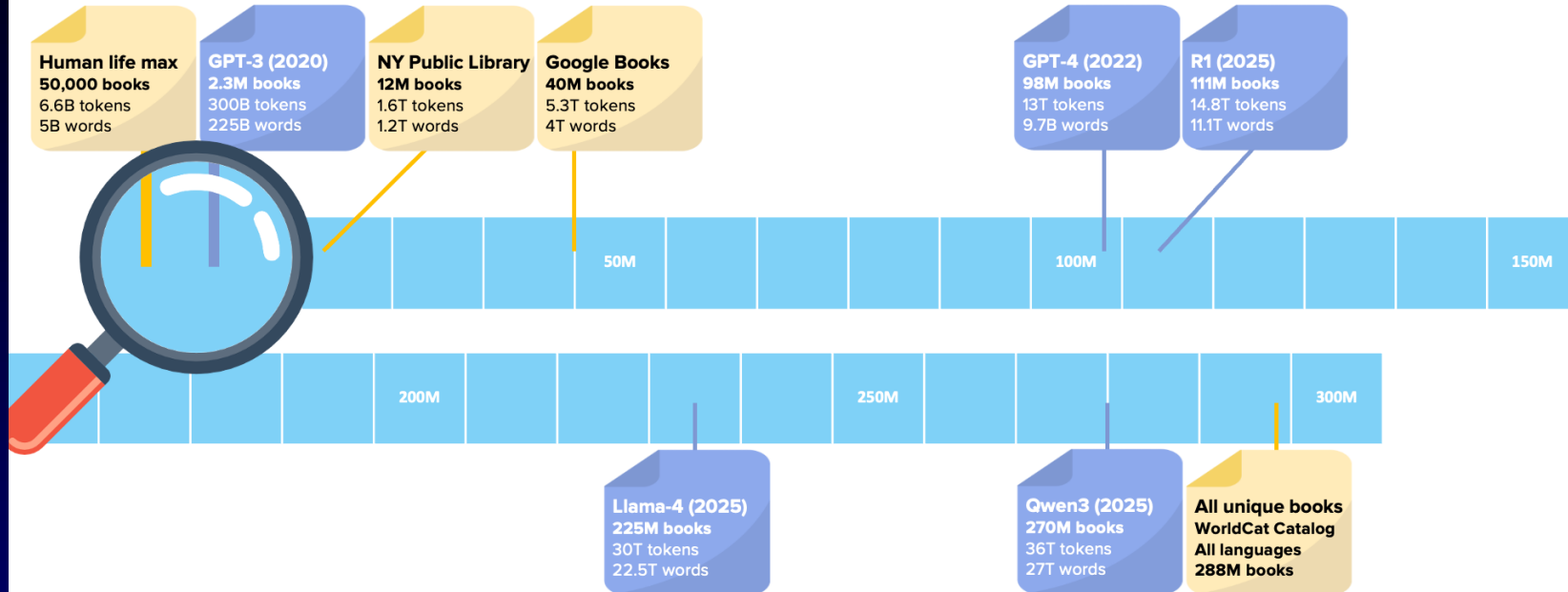
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	(829) Permalink: <a href="https://lifearchitect.ai/models-table">https://lifearchitect.ai/models-table</a>	Timeline view: <a href="https://lifearchitect.ai/timeline">https://lifearchitect.ai/timeline</a>		The Memo: <a href="https://lifearchitect.ai/memo">https://lifearchitect.ai/memo</a>		Methodology:								
2	Model	Lab	Playground	Parameters (B)	Tokens trained (B)	Ratio Tokens:Params Chinchilla scaling≥20:1	ALScore "ALScore" i Sqr Root of	MMLU	MMLU -Pro	GPQA	HLE	Training dataset	Announced	Public?
3	AuroraGPT (ScienceGPT)	Argonne National Lab	<a href="https://lifearchitect.ai/models-table">https://lifearchitect.ai/models-table</a>	2000	30000	15:1	25.8						TBA	●
4	GPT-6	OpenAI	<a href="https://lifearchitect.ai/models-table">https://lifearchitect.ai/models-table</a>										TBA	
5	Grok-5	xAI	<a href="https://lifearchitect.ai/models-table">https://lifearchitect.ai/models-table</a>	10000	200000	20:1	149.1						TBA	
6	Nemotron 3 Ultra	NVIDIA		500	25000	50:1	11.8				synthetic, web-scale		TBA	●
7	Nemotron 4	Consortium	<a href="https://www.tom">https://www.tom</a>										TBA	●
8	MAMMAL	IBM	<a href="https://huggingface">https://huggingface</a>	0.458	700	1,529:1	2.3				special		May/2026	●
9	Granite-4.1-30B	IBM	<a href="https://huggingface">https://huggingface</a>	30	15000	500:1	2.3	80.16	64.09	45.76		synthetic, web-scale	Apr/2026	●
10	Mistral Medium 3.5	Mistral	<a href="https://chat.mistral">https://chat.mistral</a>	128	12000	94:1	4.1					synthetic, web-scale	Apr/2026	●
11	Nemotron 3 Nano Om	NVIDIA	<a href="https://huggingface">https://huggingface</a>	30	25700	857:1	2.9		77.3	72.2		synthetic, web-scale	Apr/2026	●
12	Laguna XS.2	Poolside	<a href="https://huggingface">https://huggingface</a>	33	10000	304:1	1.9					synthetic, web-scale	Apr/2026	●
13	Laguna M.1	Poolside	<a href="https://platform">https://platform</a>	225	10000	45:1	5.0					synthetic, web-scale	Apr/2026	●
14	DeepSeek-V4-Pro	DeepSeek-AI	<a href="https://huggingface">https://huggingface</a>	1600	33000	21:1	24.2	90.1	87.5	90.1	37.7	synthetic, web-scale	Apr/2026	●
15	talkie-1930-13b	Independent	<a href="https://talkie-lm">https://talkie-lm</a>	13	260	20:1	0.2					history only	Apr/2026	●
16	Hy3 preview	Tencent	<a href="https://huggingface">https://huggingface</a>	295	40000	136:1	11.5	87.42	65.76		30	synthetic, web-scale	Apr/2026	●
17	Ling-2.6-1T	Inclusion AI	<a href="https://huggingface">https://huggingface</a>	1000	20750	21:1	15.2					synthetic, web-scale	Apr/2026	●
18	GPT-5.5	OpenAI	<a href="https://chatgpt">https://chatgpt</a>	3000	114000	38:1	61.6			93.6	57.2	synthetic, web-scale	Apr/2026	●
19	Marul V7	Independent	<a href="https://marulai">https://marulai</a>	0.258	1000	3,876:1	0.1					web-scale	Apr/2026	●
20	Qwen3.6-27B	Alibaba	<a href="https://huggingface">https://huggingface</a>	27	36000	1,334:1	3.3		86.1	87.8	24.3	synthetic, web-scale	Apr/2026	●
21	MiMo-V2.5-Pro	Xiaomi	<a href="https://huggingface">https://huggingface</a>	1020	27000	27:1	17.5	89.4	68.5	66.7	48	synthetic, web-scale	Apr/2026	●
22	Ling-2.6-Flash	Inclusion AI	<a href="https://huggingface">https://huggingface</a>	104	20750	200:1	4.9					synthetic, web-scale	Apr/2026	●
23	Granite-4.1-8B	IBM	<a href="https://huggingface">https://huggingface</a>	8	15000	1,875:1	2.3	73.84	55.99	41.96		synthetic, web-scale	Apr/2026	●
24	OpenMythos	Independent	<a href="https://github">https://github</a>	0.77	30	39:1	18.4					web-scale	Apr/2026	●

# LLMs – Computing Requirements

Model	Training end	Chip type	TFLOP/s (max)	Chip count	Wall clock time (days)	Total time (hours)	Total time (years)	Retail cost (\$US)	MMLU ▼
GPT-3	Apr/2020	V100	130	10,000	15 days	3,552,000	405 years	\$9M	43.9
Llama 1	Jan/2023	A100	312	2,048	21 days	1,032,192	118 years	\$4M	63.4
Llama 2	Jun/2023	A100	312	2,048	35 days	1,720,320	196 years	\$7M	68.0
Titan	Apr/2023	A100	312	13,760	48 days	11,558,400	1,319 years	\$45M	70.4
GPT-4	Aug/2022	A100	312	25,000	95 days	57,000,000	6,503 years	\$224M	86.4
Gemini	Nov/2023	TPUv4	275	57,000	100 days	136,800,000	15,606 years	\$440M	90.0
Llama 3 70B	Apr/2024	H100	989	24,576	11 days	6,300,000	719 years	\$7M	82.0
Llama 3 405B	Apr/2024	H100	989	24,576	50 days	29,491,200	3,364 years	\$125M	88.6
GPT-5	Mar/2024	H100	989	50,000	120 days	144,000,000	16,428 years	\$612M	
Olympus	Aug/2024	H100	989						
Grok 2	Jun/2024	H100	989	20,000	50 days	57,600,000	6,571 years	\$245M	
Gemini 2	Nov/2024	TPUv6	1847						
Grok 3	Dec/2024	H100	989	100,000	50 days	288,000,000	32,855 years	\$1.2B	

# LLMs – Size of Training Data (equiv. books)

## AI TRAINING DATA COMPARED TO TOTAL BOOKS (2025)



Training data sizes converted from tokens. 133K tokens = 100K words = 1 book equiv. Text datasets only. For simplicity, books are shown here in place of web data, code data, other data. Alan D. Thompson, 2025.



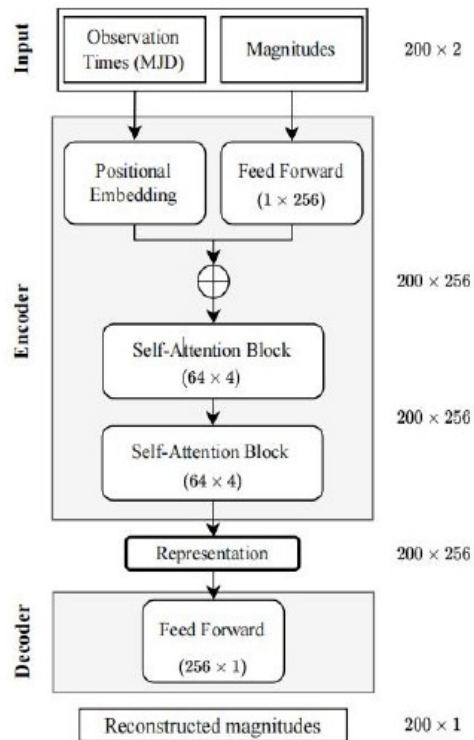
# Transformers in Astronomy - Light Curves

- ASTROMER: A transformer-based embedding for the representation of light curves
  - pre-trained on millions of light curves from different surveys (MACHO, OGLE, ATLAS)
  - representation to create informative light curves embeddings
  - finetuned for solving downstream tasks, e.g. classification of variable stars, predicting physical parameters



<https://www.stellardnn.org/projects/astromer/index.html>

C. Donoso-Oliva et al. ASTROMER: A transformer-based embedding for the representation of light curves.



# Transformers in Astronomy – Foundation Model

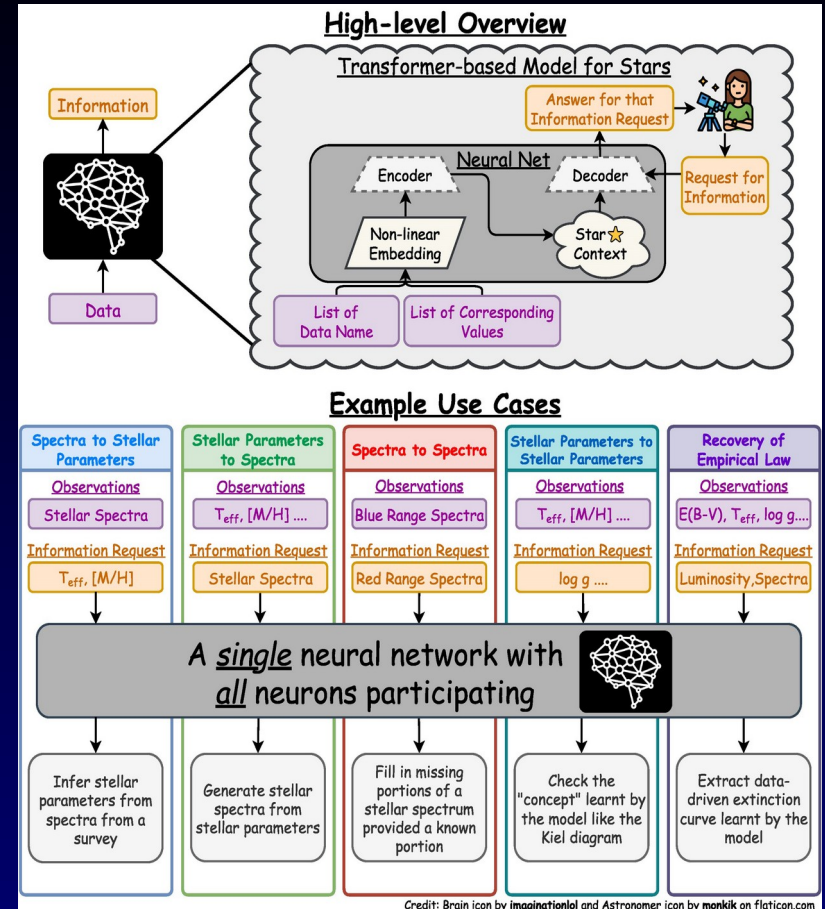
JOURNAL ARTICLE

## Towards an astronomical foundation model for stars with a transformer-based model

Henry W Leung ✉, Jo Bovy

Monthly Notices of the Royal Astronomical Society, Volume 527, Issue 1, January 2024, Pages 1494–1520, <https://doi.org/10.1093/mnras/stad3015>

Published: 05 October 2023 Article history ▾



# LLMs in Astronomy

An ESO workshop 23-27 September 2024

Hey GPT! Can you help me understand the Universe? ■

A synoptic view at the impact of chatGPT-like technologies on the future of astronomy

#ESOGPT24



Image: © ESO/P. Horálek

Scientific Organising Committee  
Giacomo Becuzzi – Henri Boffin – Sandra Grinschgl – Citta Kutyniok – Sangeeta Mysore – Maria Jose Rain Sepulveda – Suzanne Ramsay – Paula Sanchez Saaz – Sean Saporiti

Online

SEPTEMBER 23-27



## ESO Cosmic Duologues

<https://www.youtube.com/@ESOCosmicDuologues/featured>

AstroLLaMA : 7B pararameters,  
300 000 abstracts from ADS

Cosmology – FIT SED

ESO User Man

Proposal rewiews

Project assessments

Ethical problems

# Review of LLMs in Astronomy

ROYAL SOCIETY  
OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Review



**Cite this article:** Smith MJ, Geach JE. 2023  
Astronomia ex machina: a history, primer and  
outlook on neural networks in astronomy. *R. Soc.  
Open Sci.* **10**: 221454.  
<https://doi.org/10.1098/rsos.221454>

Received: 9 November 2022  
Accepted: 28 April 2023

**Subject Category:**  
Astronomy

**Subject Areas:**  
astrophysics/artificial intelligence

**Keywords:**  
neural networks, astrophysics, machine learning

## Astronomia ex machina: a history, primer and outlook on neural networks in astronomy

Michael J. Smith and James E. Geach

Department of Physics, Astronomy and Mathematics, School of Physics, Engineering and  
Computer Science, University of Hertfordshire, Hatfield AL10 9AB, UK

MJS, 0000-0003-0220-5125; JEG, 0000-0003-4964-4635

In this review, we explore the historical development and future prospects of artificial intelligence (AI) and deep learning in astronomy. We trace the evolution of connectionism in astronomy through its three waves, from the early use of multilayer perceptrons, to the rise of convolutional and recurrent neural networks, and finally to the current era of unsupervised and generative deep learning methods. With the exponential growth of astronomical data, deep learning techniques offer an unprecedented opportunity to uncover valuable insights and tackle previously intractable problems. As we enter the anticipated fourth wave of astronomical connectionism, we argue for the adoption of GPT-like foundation models fine-tuned for astronomical applications. Such models could harness the wealth of high-quality, multimodal astronomical data to serve state-of-the-art downstream tasks. To keep pace with advancements driven by Big Tech, we propose a collaborative, open-source approach within the astronomy community to develop and maintain these foundation models, fostering a symbiotic relationship between AI and astronomy that capitalizes on the unique strengths of both fields.

License: arXiv.org perpetual non-exclusive license  
arXiv:2405.20389v1 [astro-ph.IM] 30 May 2024

## Designing an Evaluation Framework for Large Language Models in Astronomy Research

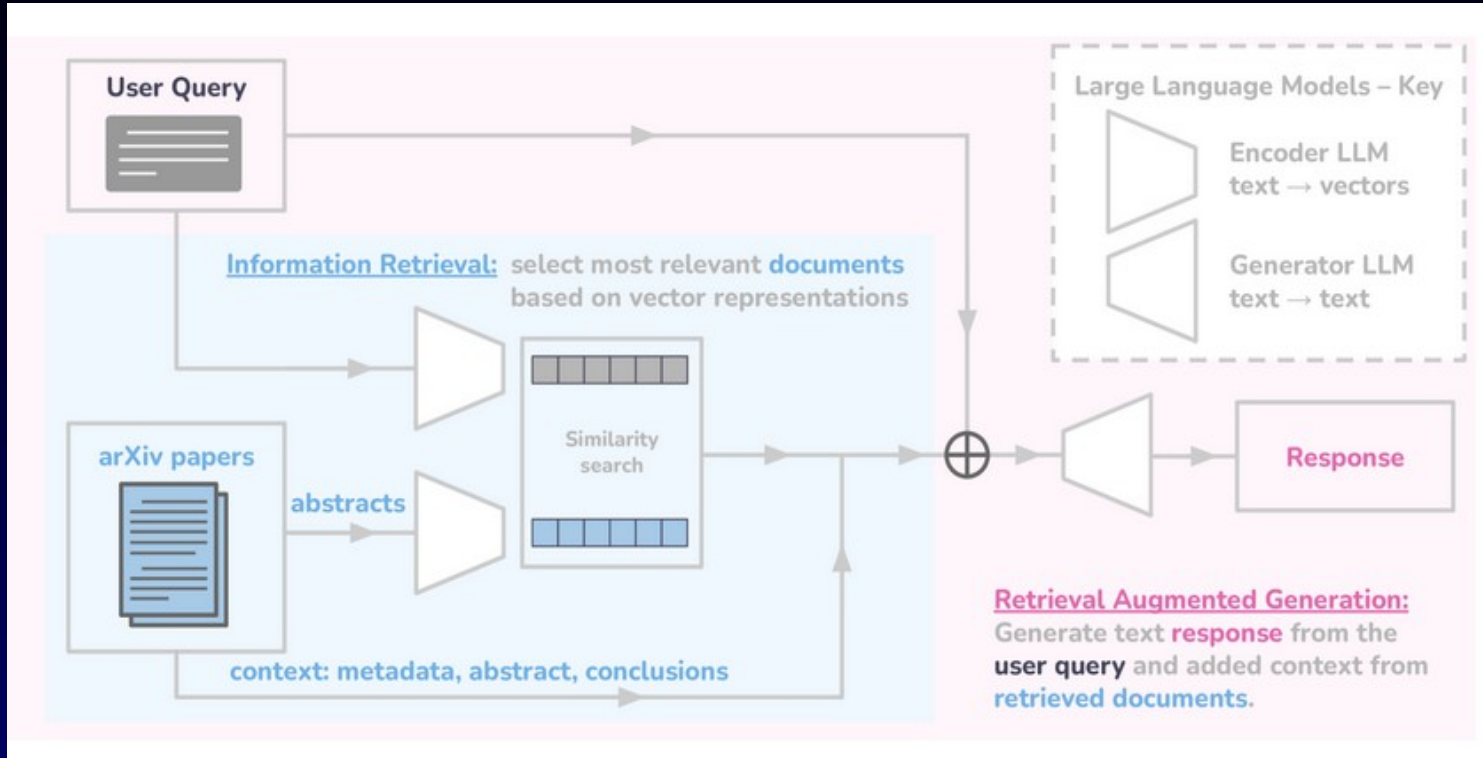
John F. Wu	Alina Hyk	Kiera McCormick
Christine Ye	Simone Astarita	Elina Baral
Jo Ciuca	Jesse Cranney	Anjalie Field
Kartheik Iyer	Philipp Koehn	Jenn Kotler
Sandor Kruk	Michelle Ntampaka	Charles O'Neill
Joshua E.G. Peek	Sanjib Sharma	Mikaeel Yunus

License: CC BY 4.0  
arXiv:2402.06196v2 [cs.CL] 20 Feb 2024

## Large Language Models: A Survey

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu  
Richard Socher, Xavier Amatriain, Jianfeng Gao

# Retrieval Augmented Generation (RAG)



# Vision Transformers in Spectroscopy

*Astronomy & Astrophysics* manuscript no. output  
March 30, 2026

©ESO 2026

## Applying vision transformers to the spectral analysis of astronomical objects

Luis Felipe Strano Moraes<sup>1</sup>, Ignacio Becker<sup>2</sup>, Pavlos Protopapas<sup>2</sup>, and Guillermo Cabrera-Vives<sup>3,4,5,6,7</sup>

<sup>1</sup> Harvard Extension School, Harvard University, Cambridge, MA, 02138, USA

<sup>2</sup> John A. Paulson School of Engineering and Applied Science, Harvard University, Cambridge, MA, 02138, USA

<sup>3</sup> Department of Computer Science, Universidad de Concepción, Edmundo Larenas 219, Concepción, Chile

<sup>4</sup> Center for Data and Artificial Intelligence, Universidad de Concepción, Edmundo Larenas 310, Concepción, Chile

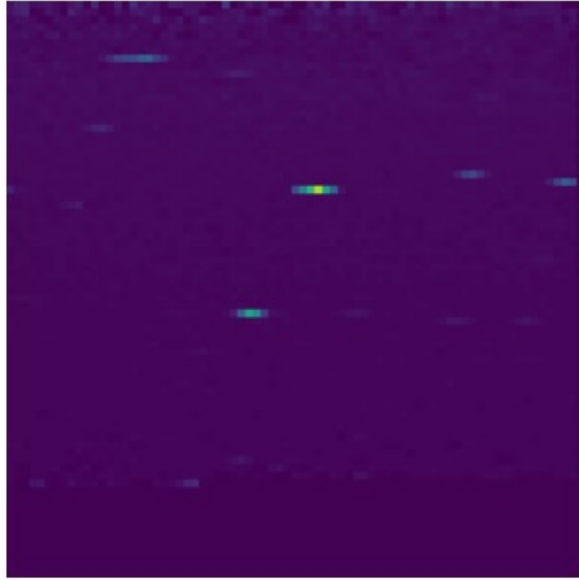
<sup>5</sup> Millennium Institute of Astrophysics (MAS), Nuncio Monseñor Sotero Sanz 100, Of. 104, Providencia, Santiago, Chile

<sup>6</sup> Millennium Nucleus for Galaxies (MINGAL), Chile

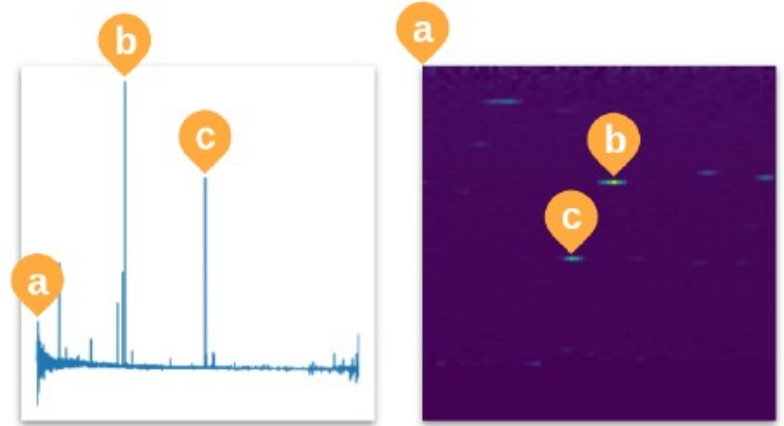
<sup>7</sup> Heidelberg Institute for Theoretical Studies, Heidelberg, Baden-Württemberg, Germany

Received May 30, 2025; Accepted Mar 01, 2026

# Vision Transformers in Spectroscopy



**Fig. 8.** 2D map plot type of the spectra, where we associate each individual wavelength with a square of 3x3 pixels in the final plot.



**Fig. 9.** Overview of how each individual flux is mapped to the final 2D image in the 2D map design. Labels *a*, *b*, and *c* can be seen on the left in the standard flux plot, and in the right side with intensity set as the color of a given region in the image.

# OLLAMA - CLI

```
skoda@betelgeuse: ~  
skoda@betelgeuse: ~ 80x24  
skoda@betelgeuse:~$ ollama  
Usage:  
  ollama [flags]  
  ollama [command]  
  
Available Commands:  
  serve      Start ollama  
  create     Create a model from a Modelfile  
  show       Show information for a model  
  run        Run a model  
  stop       Stop a running model  
  pull       Pull a model from a registry  
  push       Push a model to a registry  
  list       List models  
  ps         List running models  
  cp         Copy a model  
  rm         Remove a model  
  help       Help about any command  
  
Flags:  
  -h, --help      help for ollama  
  -v, --version    Show version information  
  
Use "ollama [command] --help" for more information about a command.
```

```
skoda@betelgeuse: ~  
skoda@betelgeuse: ~ 80x24  
skoda@betelgeuse:~$ ollama list  
NAME          ID          SIZE      MODIFIED  
llama3.2:latest a80c4f17acd5 2.0 GB    15 months ago  
llama3.2:3b    a80c4f17acd5 2.0 GB    15 months ago  
llama3.3:latest a6eb4748fd29 42 GB     15 months ago  
deepseek-r1:32b 38056bbcbb2d 19 GB     15 months ago  
deepseek-r1:7b  0a8c26691023 4.7 GB    15 months ago  
deepseek-r1:1.5b a42b25d8c10a 1.1 GB    15 months ago  
skoda@betelgeuse:~$
```

```
skoda@betelgeuse: ~  
skoda@betelgeuse: ~ 80x24  
skoda@betelgeuse:~$ ollama run llama3.2:3b  
>>> what are the basic Newton's laws of physics ?  
The three laws of motion, also known as Newton's laws of physics, were formulated by Sir Isaac Newton in the late 17th century. These laws describe the relationship between a body and the forces acting upon it. Here are the three basic laws:  
  
**1. The Law of Inertia (First Law)**  
  
* An object at rest will remain at rest, unless acted upon by an external force.  
* An object in motion will continue to move with a constant velocity, unless acted upon by an external force.  
  
In other words, an object tends to maintain its state of motion unless a force is applied to it. This law describes the concept of inertia, which is the tendency of an object to resist changes in its motion.  
  
**2. The Law of Acceleration (Second Law)**  
  
* The force applied to an object is equal to the mass of the object multiplied by its acceleration.  
*  $F = ma$  (Force = mass x acceleration)
```

# JAN - GUI

**Experimental Mode**

Enable new features that may be unstable.

**GPU Acceleration**

Enable to enhance model performance by utilizing your GPU devices for acceleration. Read troubleshooting guide for further assistance.

Choose device(s)

NVIDIA GeForce GTX 980

Nvidia

NVIDIA GeForce GTX 9804096MB VRAM for better model performance (reload needed).

**Jan Data Folder** /home/skoda/.config/jan/data

Default location for messages and other user data.

**HTTPS Proxy**

Specify the HTTPS proxy or leave blank (proxy auto-configuration and SOCKS not supported).

http://<user>:<password>@<domain or IP>:<port>

**Ignore SSL Certificates**

Allow self-signed or unverified certificates - may be required for certain proxies.

**Jan Quick Ask**

Enable Quick Ask to be triggered via the default hotkey **Ctrl + J** (reload needed).

System Monitor Jan v0.5.14

Qwen2.5 Coder 14B Instruct Q4	8.37GB	Slow on your device	Download
Qwen2.5 14B Instruct Q4	8.37GB	Slow on your device	Download
Codestral 22B Q4	12.42GB	Slow on your device	Download
Qwen2.5 Coder 32B Instruct Q4	18.53GB	Slow on your device	Download
TinyLlama Chat 1.1B Q4	638.01MB		Use
Deepseek Coder 1.3B Instruct Q8	1.33GB		Use
Opus4.7-Distill-GODSGhost-Codex-4B-Q4_K_M.gguf	2.52GB		Use
LlamaCorn 1.1B Q8	1.09GB		Download
Gemma 1.1 2B Q4	1.52GB		Download
Gemma 2 2B Q4	1.59GB		Download

System Monitor Jan v0.5.14

# JAN - GUI

General

My Models

Preferences

Keyboard Shortcuts

Privacy

Advanced Settings

Extensions

Model Providers

Anthropic

Cohere

Cortex

Groq

Martian

MistralAI

TensorRT-LLM

NVIDIA NIM

OpenAI

OpenRouter

Triton-TRT-LLM

Core Extensions

Model Management

Search or paste Hugging Face URL

Import Model

Llama.cpp (Cortex)

Model Name	Size	Status
Deepseek Coder 1.3B Instruct Q8 deepseek-coder-1.3b	1.33GB	Inactive
Llama 3.2 1B Instruct Q8 llama3.2-1b-instruct	1.23GB	Active
Llama 3.2 3B Instruct Q8 llama3.2-3b-instruct	3.19GB	Inactive

Anthropic

Cohere

Groq

Martian

Nvidia

Mistral 7B

System Monitor Jan v0.5.14

what are the Newton...

Title Telling Jokess

New Thread

New Thread

New Thread

what are the newton...

what are the newton...

Im sorry but as an AI...

New Thread

Tetris Game Develop...

recursive calls and hence increases performance.

The time complexity is still  $O(n \log n)$ , but the constant factors are reduced due to these optimizations, leading to significant speedups.

User 09:28:58 AM

write an optimized numpy accelerated quick sort algorithm

Assistant 09:29:25 AM

Here's a more optimized version of NumPy-accelerated Quick Sort Algorithm that uses Python builtin functions and makes use of the `array_split` function in Numpy which is faster for large arrays. This method divides our original list into its original list into two halves, then it sorts each half individually using recursion until they are all sorted together:

```
python
1 import numpy as np # For working with NumPy Arrays (for efficiency) and built-in
2 from math import floor # for integer division to avoid float result. This is
3 def quicksort_optimized(arr): # Define Our QuickSort Function. Input is an Array
4     if len(arr) <= 1: # If array has 0 or one element then return same because
5         return arr
6     else: # If array has more than 1 element then we will use pivot as middle i
7         mid = arr[0] # Define Middle Element for Pivoting - Assuming first elem
8         less = np.empty(0) # Define a empty list for elements that are lesser t
9         more = np.empty(0) # Define a empty list for elements that are greater
10        equal = np.empty(0) # Define a empty list for elements that are exactly
11        mid_index = None # Index where array is divided into two parts. We will
12        for ix, val in enumerate(arr): # Iterate through our array/list - This
13            if mid_index is None: # If first iteration then choose middle as p
14                mid_index = ix # Set the pivot position (mid) by taking current
15            if val < mid: # If current element less than 'mid' (pivot) then we
16                np.append(less, val) # Append this element into "less" numpy ar
```

Ask me anything

Deepseek Coder 1.3B Inst...

System Monitor Jan v0.5.14

# Multimodal Universe

Astrophysics > Instrumentation and Methods for Astrophysics

[Submitted on 3 Dec 2024]

## The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TB of Astronomical Scientific Data

The Multimodal Universe Collaboration. Eirini Angeloudi, Jeroen Audenaert, Micah Bowles, Benjamin M. Boyd, David Chemaly, Brian Cherinka, Ioana Ciucă, Miles Cranmer, Aaron Do, Matthew Grayling, Erin E. Hayes, Tom Hehir, Shirley Ho, Marc Huertas-Company, Kartheik G. Iyer, Maja Jablonska, Francois Lanasse, Henry W. Leung, Kaisey Mandel, Juan Rafael Martínez-Galarza, Peter Melchior, Lucas Meyer, Liam H. Parker, Helen Qu, Jeff Shen, Michael J. Smith, Connor Stone, Mike Walmsley, John F. Wu

We present the MULTIMODAL UNIVERSE, a large-scale multimodal dataset of scientific astronomical data, compiled specifically to facilitate machine learning research. MULTIMODAL UNIVERSE contains hundreds of millions of astronomical observations, constituting 100, TB of multi-channel and hyper-spectral images, spectra, multivariate as a wide variety of associated scientific measurements and "metadata". In addition, we include a range of benchmark tasks representative of standard practices for machine learning in astrophysics. This massive dataset will enable the development of large multi-modal models specifically targeted towards scientific applications. All codes used to compile MULTIMODAL UNIVERSE and a description of how to access the data is available at [this https URL](https://doi.org/10.48550/arXiv.2412.02527).

Comments: Accepted at NeurIPS Datasets and Benchmarks track

Subjects: **Instrumentation and Methods for Astrophysics (astro-ph.IM)**; Astrophysics of Galaxies (astro-ph.GA); Solar and Stellar Astrophysics (astro-ph.SR)

Cite as: [arXiv:2412.02527](https://arxiv.org/abs/2412.02527) [astro-ph.IM]

(or [arXiv:2412.02527v1](https://arxiv.org/abs/2412.02527v1) [astro-ph.IM] for this version)

<https://doi.org/10.48550/arXiv.2412.02527> 

The Multimodal Universe currently contains data from the following surveys/modalities:

Survey	Modality	Science Use Case	# samples
Legacy Surveys DR10	Images	Galaxies	124M
Legacy Surveys North	Images	Galaxies	15M
HSC	Images	Galaxies	477k
BTS	Images	Supernovae	400k
JWST	Images	Galaxies	300k
Gaia BP/RP	Spectra	Stars	220M
SDSS-II	Spectra	Galaxies, Stars	4M
DESI	Spectra	Galaxies	20M
APOGEE SDSS-III	Spectra	Stars	716k
GALAH	Spectra	Stars	325k
Chandra	Spectra	Galaxies, Stars	129k
VIPERS	Spectra	Galaxies	91k
MaNGA SDSS-IV	Hyperspectral Image	Galaxies	12k
PLAsTiCC	Time Series	Time-varying objects	3.5M
TESS	Time Series	Exoplanets	160k
CfA Sample	Time Series	Supernovae	1k
YSE	Time Series	Supernovae	2k
PS1 SNe Ia	Time Series	Supernovae	369
DES Y3 SNe Ia	Time Series	Supernovae	248
SNLS	Time Series	Supernovae	239
Foundation	Time Series	Supernovae	180
CSP SNe Ia	Time Series	Supernovae	134
Swift SNe Ia	Time Series	Supernovae	117
Gaia	Tabular	Stars	220M
PROVABGS	Tabular	Galaxies	221k
Galaxy10 DECaLS	Tabular	Galaxies	15k
GALEX All-Sky Imaging	Tabular	Stars	83M
SAGES DR1	Tabular	Stars	48.5M
2MASS PSC	Tabular	Galaxies, Stars	471M
AllWISE Source Catalog	Tabular	Galaxies, Stars	747M

# Multimodal Universe

← → ↻ huggingface.co/MultimodalUniverse/datasets 80% ☆

🔖 Import bookmarks... 🌐 Getting Started 📄 Read the chip data fr... 🗂 archive-program/GUI... 🛠 Explorer | Skyviewer 📄 CERES: A Set of Auto... 📄 backup - How to writ... >> 📁 Other Bookmarks

🏠 AI & ML interests


Machine Learning for Astrophysics

🔔 Recent Activity

- 👤 Smith42 authored a paper 26 days ago  
A Survey on Hypothesis Generation for Scientific Discovery in...
- 👤 Smith42 authored a paper 26 days ago  
The Platonic Universe: Do Foundation Models See the Same S...
- 👤 EffL authored a paper 6 months ago  
AION-1: Omnimodal Foundation Model for Astronomical Scien...

👁 View all activity

👤 Team members 13



< MultimodalUniverse 's datasets 24 🔍

⌵ Sort: Recently updated

- MultimodalUniverse/jwst**  
👁 Viewer · Updated Dec 6, 2024 · 📄 85.6k · ⬆ 252 · ❤ 3
- MultimodalUniverse/legacysurvey**  
👁 Viewer · Updated Dec 4, 2024 · 📄 90.6k · ⬆ 749 · ❤ 5
- MultimodalUniverse/plasticc**  
👁 Viewer · Updated Dec 4, 2024 · 📄 7.85k · ⬆ 211 · ❤ 1
- MultimodalUniverse/hsc**  
👁 Viewer · Updated Dec 4, 2024 · 📄 73k · ⬆ 200
- MultimodalUniverse/manga**  
👁 Viewer · Updated Dec 4, 2024 · 📄 12 · ⬆ 26
- MultimodalUniverse/ssl\_legacysurvey**  
👁 Preview · Updated Dec 4, 2024 · ⬆ 31 · ❤ 1
- MultimodalUniverse/ps1\_sne\_ia**  
👁 Viewer · Updated Dec 4, 2024 · 📄 369 · ⬆ 22
- MultimodalUniverse/desi**  
👁 Viewer · Updated Dec 4, 2024 · 📄 100k · ⬆ 373
- MultimodalUniverse/snls**  
👁 Viewer · Updated Dec 4, 2024 · 📄 10 · ⬆ 30
- MultimodalUniverse/tess**  
👁 Viewer · Updated Dec 4, 2024 · 📄 100k · ⬆ 46
- MultimodalUniverse/swift\_sne\_ia**  
👁 Viewer · Updated Dec 4, 2024 · 📄 117 · ⬆ 5
- MultimodalUniverse/foundation**  
👁 Viewer · Updated Dec 4, 2024 · 📄 180 · ⬆ 25
- MultimodalUniverse/des\_y3\_sne\_ia**  
👁 Viewer · Updated Dec 4, 2024 · 📄 248 · ⬆ 8
- MultimodalUniverse/gz10**  
👁 Viewer · Updated Dec 4, 2024 · 📄 17.7k · ⬆ 31 · ❤ 1
- MultimodalUniverse/yse**  
👁 Viewer · Updated Dec 3, 2024 · 📄 2k · ⬆ 11
- MultimodalUniverse/chandra**  
👁 Viewer · Updated Dec 3, 2024 · 📄 100k · ⬆ 86
- MultimodalUniverse/vipers**  
👁 Viewer · Updated Dec 3, 2024 · 📄 91.5k · ⬆ 27
- MultimodalUniverse/gaia**  
👁 Viewer · Updated Dec 3, 2024 · 📄 100k · ⬆ 19.7k · ❤ 1
- MultimodalUniverse/desi\_provabgs**  
👁 Viewer · Updated Dec 3, 2024 · 📄 100k · ⬆ 130 · ❤ 1
- MultimodalUniverse/btsbot**  
👁 Viewer · Updated Dec 3, 2024 · 📄 100k · ⬆ 30 · ❤ 1
- MultimodalUniverse/sdss**  
👁 Viewer · Updated Dec 3, 2024 · 📄 100k · ⬆ 526 · ❤ 2
- MultimodalUniverse/cfa**  
👁 Viewer · Updated Dec 3, 2024 · 📄 185 · ⬆ 17
- MultimodalUniverse/csp**  
👁 Viewer · Updated Dec 2, 2024 · 📄 134 · ⬆ 10
- MultimodalUniverse/apogee**  
👁 Viewer · Updated Jun 13, 2024 · 📄 44 · ⬆ 25

# Foundation Models in Astronomy

Monthly Notices

of the  
ROYAL ASTRONOMICAL SOCIETY

MNRAS **531**, 4990–5011 (2024)

Advance Access publication 2024 June 21

<https://doi.org/10.1093/mnras/stae1450>



## AstroCLIP: a cross-modal foundation model for galaxies

Liam Parker<sup>1,†,★</sup>, Francois Lanusse<sup>1,2</sup>, Siavash Golkar<sup>1</sup>, Leopoldo Sarra<sup>1</sup>, Miles Cranmer<sup>3</sup>, Alberto Bietti<sup>1</sup>, Michael Eickenberg<sup>1</sup>, Geraud Krawezik<sup>1</sup>, Michael McCabe<sup>1,4</sup>, Rudy Morel<sup>1</sup>, Ruben Ohana<sup>1</sup>, Mariel Pettee<sup>1,5</sup>, Bruno Régaldo-Saint Blancard<sup>1</sup>, Kyunghyun Cho<sup>6,7,8</sup>, Shirley Ho<sup>1,6,9</sup> and The Polymathic AI Collaboration

<sup>1</sup>The Flatiron Institute, 162 5th Ave, New York, NY 10010, USA

<sup>2</sup>CEA, CNRS, AIM, Université Paris-Saclay, Université Paris Cité, Paris 91190, France

<sup>3</sup>Department of Astronomy, University of Cambridge, Madingley Rd, Cambridge CB3 0HA, UK

<sup>4</sup>Department of Computer Science, University of Colorado, Boulder, 430 UCB, 1111 Engineering Dr, Boulder, CO 80309, USA

<sup>5</sup>Lawrence Berkeley National Laboratory, Berkeley, 1 Cyclotron Rd, CA 94720, USA

<sup>6</sup>Center for Data Science, New York University, 60 5th Ave, New York, NY 10011, USA

<sup>7</sup>Prescient Design, Genentech, 149 5th Ave, New York, NY 10010, USA

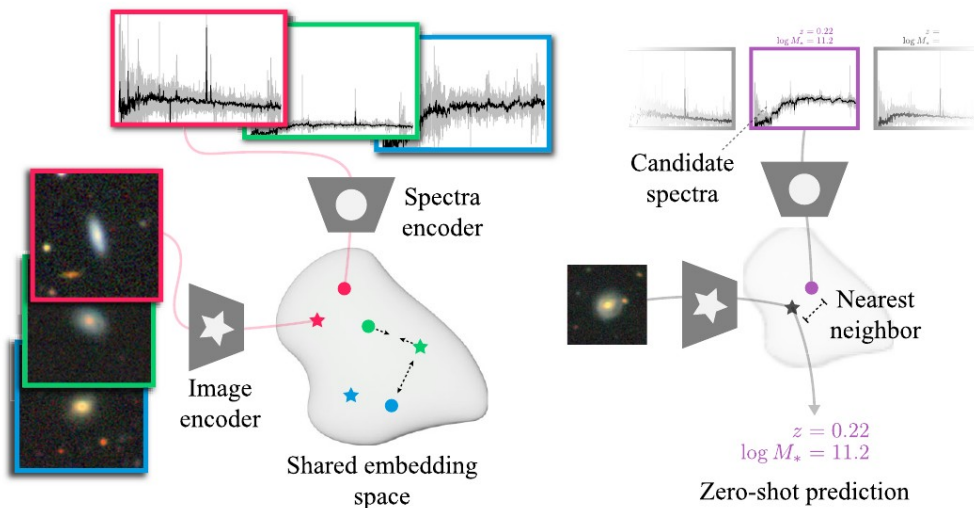
<sup>8</sup>CIFAR Learning in Machines and Brains Fellow, Toronto, ON M5G 1M1, Canada

<sup>9</sup>Department of Astrophysics, Princeton University, 4 Ivy Lane, Princeton, NJ 08544, USA

Accepted 2024 May 30. Received 2024 May 20; in original form 2024 March 15

### ABSTRACT

We present AstroCLIP, a single, versatile model that can embed both galaxy images and spectra into a shared, physically meaningful latent space. These embeddings can then be used – without any model fine-tuning – for a variety of downstream tasks including (1) accurate in-modality and cross-modality semantic similarity search, (2) photometric redshift estimation, (3) property estimation from both images and spectra, and (4) morphology classification. Our approach to implementing As



**Figure 1.** Illustration of the *AstroCLIP* cross-modal training strategy. This approach consists of two steps. First, galaxy images and spectra are embedded separately by pre-training both an image and a spectrum encoder in a SSL setting (left). Then, these encoders are aligned using a cross-modal contrastive loss. Once aligned, these embeddings allow us to connect and compare cross-modal representations (right). Moreover, they possess physically meaningful high-level information which can be used for a variety of downstream tasks on which the model was neither trained nor fine-tuned.

# Omnimodal Foundation Model in Astronomy

Astrophysics > Instrumentation and Methods for Astrophysics

[Submitted on 20 Oct 2025]

## AION-1: Omnimodal Foundation Model for Astronomical Sciences

Liam Parker, Francois Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, Leopoldo Sarra, Lucas Meyer, Micah Bowles, Sebastian Wagner-Carena, Helen Qu, Siavash Golkar, Alberto Bietti, Hatim Bourfoune, Nathan Casserau, Pierre Cornette, Keiya Hirashima, Geraud Krawezik, Ruben Ohana, Nicholas Lourie, Michael McCabe, Rudy Morel, Payel Mukhopadhyay, Mariel Pettee, Bruno Regaldo-Saint Blancard, Kyunghyun Cho, Miles Cranmer, Shirley Ho

While foundation models have shown promise across a variety of fields, astronomy still lacks a unified framework for joint modeling across modalities. We present AION-1, a family of large-scale multimodal foundation models for astronomy. AION-1 integrates heterogeneous imaging, spectroscopic, and scalar data modalities. The model is pretrained on a broad suite of astronomical data modalities. It achieves strong performance across various downstream tasks, including galaxy and stellar property estimation, galaxy morphology classification, similarity-based retrieval, galaxy image segmentation, and spectral super-resolution. We release AION-1 model variants ranging from 300M to 3.1B parameters. Beyond astronomy, AION-1 provides a scalable blueprint for multimodal scientific foundation models that can seamlessly integrate noisy, instrument-specific observations. All code, tokenizers, pretrained weights, and a lightweight evaluation suite are released under an open-source license.

Comments: Accepted at Neural Information Processing Systems (2025)

Subjects: **Instrumentation and Methods for Astrophysics (astro-ph.IM)**; Cosmology and Nongalactic Astrophysics (astro-ph.CO)

Cite as: [arXiv:2510.17960](https://arxiv.org/abs/2510.17960) [astro-ph.IM]

(or [arXiv:2510.17960v1](https://arxiv.org/abs/2510.17960v1) [astro-ph.IM] for this version)

<https://doi.org/10.48550/arXiv.2510.17960>

The screenshot shows the Hugging Face interface for the paper 'AION-1: Omnimodal Foundation Model for Astronomical Sciences'. The page includes the title, authors (Liam Parker, Francois Lanusse, Jeff Shen, Ollie Liu, Tom Hehir, Leopoldo Sarra, Lucas Meyer, Micah Bowles, Sebastian Wagner-Carena, Helen Qu, Siavash Golkar, Alberto Bietti, Hatim Bourfoune, Nathan Casserau, Pierre Cornette, Keiya Hirashima, Geraud Krawezik, Ruben Ohana, Nicholas Lourie, Michael McCabe, Rudy Morel, Payel Mukhopadhyay), and a list of models citing the paper. The abstract is also visible, describing the model's architecture and performance.

**AION-1: Omnimodal Foundation Model for Astronomical Sciences**

Published on Oct 20, 2025 • Submitted by [Francois Lanusse](#) on Oct 21, 2025 • [Polymathic AI](#)

Authors: Liam Parker, [Francois Lanusse](#), Jeff Shen, [Ollie Liu](#), Tom Hehir, Leopoldo Sarra, Lucas Meyer, Micah Bowles, Sebastian Wagner-Carena, Helen Qu, Siavash Golkar, Alberto Bietti, Hatim Bourfoune, Nathan Casserau, Pierre Cornette, Keiya Hirashima, Geraud Krawezik, Ruben Ohana, Nicholas Lourie, Michael McCabe, Rudy Morel, Payel Mukhopadhyay +5 authors

**Abstract**

AION-1, a family of large-scale multimodal foundation models, integrates diverse astronomical data using tokenization and transformer-based modeling, achieving strong performance across various downstream tasks.

While foundation models have shown promise across a variety of fields, astronomy still lacks a unified framework for joint modeling across its highly diverse data modalities. In this paper, we present AION-1, a family of large-scale multimodal foundation models for astronomy. AION-1 integrates heterogeneous imaging, spectroscopic, and scalar data using a two-stage architecture: modality-specific tokenization followed by transformer-based masked modeling of cross-modal token sequences. The model is pretrained on five large-scale surveys: Legacy Survey, Hyper Suprime-Cam (HSC), Sloan Digital Sky Survey (SDSS), Dark Energy Spectroscopic Instrument (DESI), and Gaia. These span more than 200 million observations of stars, galaxies, and quasars. With a single frozen encoder, AION-1 achieves strong results on a broad suite of downstream tasks, including galaxy and stellar property estimation, galaxy morphology classification, similarity-based retrieval, galaxy image segmentation, and spectral super-resolution. We release AION-1 model variants ranging from 300M to 3.1B parameters. Beyond astronomy, AION-1 provides a scalable blueprint for multimodal scientific foundation models that can seamlessly integrate noisy, instrument-specific observations. All code, tokenizers, pretrained weights, and a lightweight evaluation suite are released under an open-source license.

**Models citing this paper**

- [polymathic-ai/aion-base](#)  
Any-to-Any • 0.3B • Updated 27 days ago • 5.03k • 19
- [polymathic-ai/aion-large](#)  
Any-to-Any • 0.3B • Updated 27 days ago • 151 • 1
- [polymathic-ai/aion-xlarge](#)  
Any-to-Any • 3B • Updated 27 days ago • 123 • 1

**Datasets citing this paper**

- [astronolan/lens-retrieval-ls-embeddings](#)  
Viewer • Updated Feb 11 • 565k • 244
- [kshitiijd/mini-mmu-diverse-fatima-fellowship](#)  
Updated Feb 26 • 209
- [astronolan/galaxy10-aion](#)  
Viewer • Updated Feb 13 • 7.92k • 120

**Spaces citing this paper**

Ask HuggingChat about this Paper

# Omnimodal Foundation Model in Astronomy

README MIT license

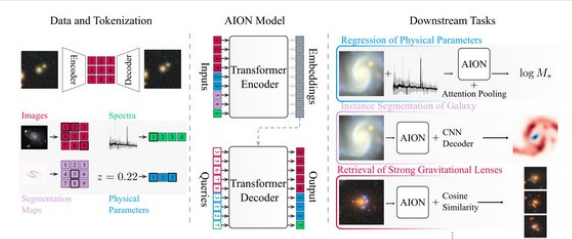
## AION-1: Astronomical Omnimodal Network

License MIT PyTorch >2.4.0 Tests failing arXiv 2510.17960 Open in Colab Hugging Face Model

Polymathic's Large Omnimodal Model for Astronomy

[Quick Start](#) • [Tutorials](#) • [Scientific Overview](#) • [Advanced Installation](#)

### Overview



AION-1 is a cutting-edge large omnimodal model specifically designed for astronomical surveys. It seamlessly integrates multiple data modalities, and enables simple adaptation to a wide range of astronomical tasks.

### Quick Start

Assuming you have PyTorch installed, you can install AION trivially with:

```
pip install polymathic-aion
```

Then you can load the pretrained model and start analyzing astronomical data:

```
import torch
from aion import AION
from aion.codecs import CodecManager
from aion.modalities import LegacySurveyImage

# Load model and codec manager
model = AION.from_pretrained('aion-base').to('cuda') # or 'aion-large', 'aion-xlarge'
codec_manager = CodecManager(device='cuda')
```

README MIT license

### Supported Modalities

AION-1's tokenizers cover 39 distinct data types, grouped by survey and data category

Category	Description	Token Name(s)
Imaging (2)	Legacy Survey, HSC Wide	tok_image_ls, tok_image_hsc
Catalog (1)	Legacy Survey catalog entries	catalog
Spectra (2)	SDSS, DESI	tok_spectrum_sdss, tok_spectrum_desi
Gaia (4)	BP/RP spectra, parallax, sky coords	tok_xp_bp, tok_xp_rp, tok_parallax, tok_ra, tok_dec
Gaia Photometry (3)	G/BP/RP flux	tok_flux_g_gaia, tok_flux_bp_gaia, tok_flux_rp_gaia
Legacy Survey (9)	g,r,i,z bands & WISE W1-W4 flux, E(B-V)	tok_flux_g_..., tok_flux_w4, tok_ebv
Legacy Shape (3)	Ellipticity components & effective radius	tok_shape_e1, tok_shape_e2, tok_shape_r
HSC Photometry (5)	g,r,i,z,y magnitudes	tok_mag_g_..., tok_mag_y
HSC Extinction (5)	g,r,i,z,y extinctions	tok_a_g_..., tok_a_y
HSC Shape (3)	Shape components 11,22,12	tok_shape11, tok_shape22, tok_shape12
Other (1)	Spectroscopic redshift	tok_z

### Model Variants

Variant	Encoder Blocks	Decoder Blocks	Model Dim	Heads	Total Params	Model
Base	12	12	768	12	300 M	<a href="#">aion-base</a>
Large	24	24	1024	16	800 M	soon
XLarge	24	24	2048	32	3 B	soon

Pretraining - Global batch size: 8 192 - Steps: Base (1.5 days on 64 H100), Large (2.5 days on 100 H100), XLarge (3.5 days on 288 H100) - Optimizer: AdamW, peak LR  $2 \times 10^{-4}$ , linear warmup + cosine decay

***Near Future  
Challenges***

# Fourier Neural Operators – Idea

Computer Science > Machine Learning

[Submitted on 18 Oct 2020 (v1), last revised 17 May 2021 (this version, v3)]

## Fourier Neural Operator for Parametric Partial Differential Equations

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, Anima Anandkumar

The classical development of neural networks has primarily focused on learning mappings between finite-dimensional Euclidean spaces. Recently, this has been generalized to neural operators that learn mappings between function spaces. For partial differential equations (PDEs), neural operators directly learn the mapping from any functional parametric dependence to the solution. Thus, they learn an entire family of PDEs, in contrast to classical methods which solve one instance of the equation. In this work, we formulate a new neural operator by parameterizing the integral kernel directly in Fourier space, allowing for an expressive and efficient architecture. We perform experiments on Burgers' equation, Darcy flow, and Navier-Stokes equation. The Fourier neural operator is the first ML-based method to successfully model turbulent flows with zero-shot super-resolution. It is up to three orders of magnitude faster compared to traditional PDE solvers. Additionally, it achieves superior accuracy compared to previous learning-based solvers under fixed resolution.

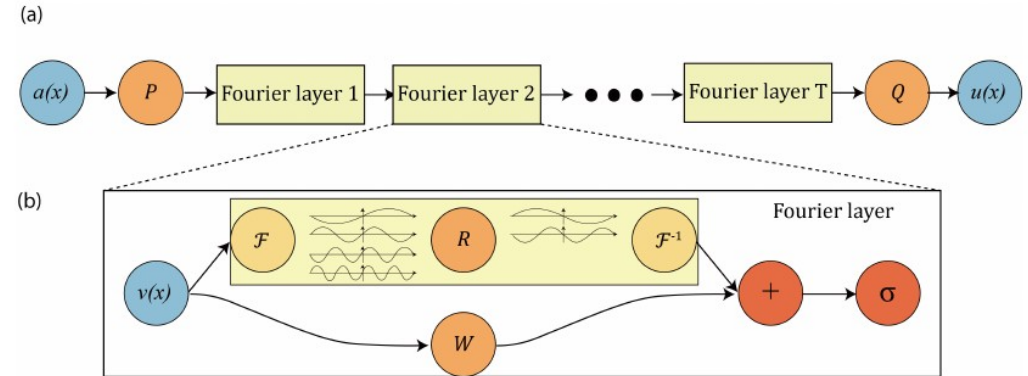
Subjects: **Machine Learning** (cs.LG); Numerical Analysis (math.NA)

Cite as: [arXiv:2010.08895](https://arxiv.org/abs/2010.08895) [cs.LG]

(or [arXiv:2010.08895v3](https://arxiv.org/abs/2010.08895v3) [cs.LG] for this version)

<https://doi.org/10.48550/arXiv.2010.08895>

Published as a conference paper at ICLR 2021



(a) The full architecture of neural operator: start from input  $a$ . 1. Lift to a higher dimension channel space by a neural network  $P$ . 2. Apply four layers of integral operators and activation functions. 3. Project back to the target dimension by a neural network  $Q$ . Output  $u$ . (b) Fourier layers: Start from input  $v$ . On top: apply the Fourier transform  $\mathcal{F}$ ; a linear transform  $R$  on the lower Fourier modes and filters out the higher modes; then apply the inverse Fourier transform  $\mathcal{F}^{-1}$ . On the bottom: apply a local linear transform  $W$ .

Figure 2: **top:** The architecture of the neural operators; **bottom:** Fourier layer.

# Fourier Neural Operators - Applications

## Earth and Space Science

Research Article | Open Access |

### Fourier Neural Operator for Moonquake Detection

[Basem Al-Qadasi](#) [Umar Bin Waheed](#)

First published: 18 March 2026 | <https://doi.org/10.1029/2025EA004792> | [VIEW METRICS](#)

SECTIONS

PDF CITE TOOLS SHAI



Volume 13, Issue 3  
March 2026  
e2025EA004792

THE ASTROPHYSICAL JOURNAL, 993:146 (12pp), 2025 November 1  
© 2025. The Author(s). Published by the American Astronomical Society.

**OPEN ACCESS**

<https://doi.org/10.3847/1538-4357/ae0f90>



### A Grid-constrained Fourier Neural Operator for Magnetic Field Extrapolation of the Solar Corona

Hao Yang , Zhenhua Liu, Pingbing Zuo , Peng Zou , Ruize Zhang, and Yanshi Huang

State Key Laboratory of Solar Activity and Space Weather, School of Aerospace, Harbin Institute of Technology, Shenzhen 518055, People's Republic of China; [pbzuo@hit.edu.cn](mailto:pbzuo@hit.edu.cn)

Received 2025 July 8; revised 2025 October 3; accepted 2025 October 4; published 2025 October 30

#### Abstract

for coronal magnetic field extrapolation modeling. This (s) with grid-based finite-difference methods used to impose NO (GC-FNO). The GC-FNO trains an FNO with nonlinear le, learning the mapping of the magnetic fields between two It then iteratively applies this FNO to reconstruct the three- el via divergence loss and force-freeness loss computed from The GC-FNO excels in processing data with nonperiodic cs of NLFFF data to minimize error accumulation during the analyses indicate that the GC-FNO achieves reasonably ms well in universality tests across 30 active regions as well

#### Computer Science > Machine Learning

[Submitted on 6 Oct 2025]

### Physics-informed Attention-enhanced Fourier Neural Operator for Solar Magnetic Field Extrapolations

[Jinghao Cao](#), [Qin Li](#), [Mengnan Du](#), [Haimin Wang](#), [Bo Shen](#)

We propose Physics-informed Attention-enhanced Fourier Neural Operator (PIANO) to solve the Nonlinear Force-Free Field (NLFFF) problem in solar physics. Unlike conventional approaches that rely on iterative numerical methods, our proposed PIANO directly learns the 3D magnetic field structure from 2D boundary conditions. Specifically, PIANO integrates Efficient Channel Attention (ECA) mechanisms with Dilated Convolutions (DC), which enhances the model's ability to capture multimodal input by prioritizing critical channels relevant to the magnetic field's variations. Furthermore, we apply physics-informed loss by enforcing the force-free and divergence-free conditions in the training process so that our prediction is consistent with underlying physics with high accuracy. Experimental results on the ISEE NLFFF dataset show that our PIANO not only outperforms state-of-the-art neural operators in terms of accuracy but also shows strong consistency with the physical characteristics of NLFFF data across magnetic fields reconstructed from various solar active regions. The GitHub of this project is available [this https URL](#)

Comments: 10 pages; accepted as workshop paper in ICDM 2025; [this https URL](#)

Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI)

Cite as: [arXiv:2510.05351](https://arxiv.org/abs/2510.05351) [cs.LG]  
(or [arXiv:2510.05351v1](https://arxiv.org/abs/2510.05351v1) [cs.LG] for this version)  
<https://doi.org/10.48550/arXiv.2510.05351>

# Fourier Neural Operators - Tutorial

Computer Science > Machine Learning

[Submitted on 1 Dec 2025 (v1), last revised 22 Jan 2026 (this version, v2)]

## Fourier Neural Operators Explained: A Practical Perspective

Valentin Duruisseaux, Jean Kossaifi, Anima Anandkumar

Partial differential equations (PDEs) govern a wide variety of dynamical processes in science and engineering, yet obtaining their numerical solutions often requires high-resolution discretizations and repeated evaluations of complex operators, leading to substantial computational costs. Neural operators have recently emerged as a powerful framework for learning mappings between function spaces directly from data, enabling efficient surrogate models for PDE systems. Among these architectures, the Fourier Neural Operator (FNO) has become the most influential and widely adopted due to its elegant spectral formulation, which captures global correlations through learnable transformations in Fourier space while remaining invariant to discretization and resolution. Despite their success, the practical use of FNOs is often hindered by an incomplete understanding among practitioners of their theoretical foundations, practical constraints, and implementation details, which can lead to their incorrect or unreliable application. This work presents a comprehensive and practice-oriented guide to FNOs, unifying their mathematical principles with implementation strategies. We provide an intuitive exposition to the concepts of operator theory and signal-processing that underlie the FNO, detail its spectral parameterization and the computational design of all its components, and address common misunderstandings encountered in the literature. The exposition is closely integrated with the NeuralOperator 2.0.0 library, offering modular state-of-the-art implementations that faithfully reflect the theory. By connecting rigorous foundations with practical insight, this guide aims to establish a clear and reliable framework for applying FNOs effectively across diverse scientific and engineering fields.

Comments: 96 pages, 27 figures

Subjects: **Machine Learning (cs.LG)**

Cite as: [arXiv:2512.01421](https://arxiv.org/abs/2512.01421) [cs.LG]

(or [arXiv:2512.01421v2](https://arxiv.org/abs/2512.01421v2) [cs.LG] for this version)

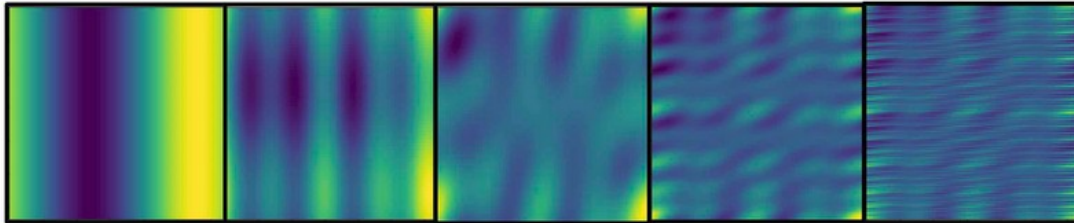
<https://doi.org/10.48550/arXiv.2512.01421> 

# Fourier Neural Operator - Libraries

## Convolutional Filters in Fourier Space



Filters in CNN



Fourier Filters

Filters in convolution neural networks are usually local. They are good for capturing local patterns such as edges and shapes. Fourier filters are global sinusoidal functions. They are better for representing continuous functions.

NeuralOperator Install Theory Guide User Guide API Examples Developer's Guide [GitHub](#)

### Neural Operators in PyTorch

NeuralOperator is a comprehensive PyTorch library for learning neural operators, containing the official implementation of Fourier Neural Operators and other neural operator architectures.

NeuralOperator is part of the PyTorch Ecosystem, check the [PyTorch announcement!](#)

# Global Filter Networks

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 1 Jul 2021 (v1), last revised 26 Oct 2021 (this version, v2)]

## Global Filter Networks for Image Classification

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, Jie Zhou

Recent advances in self-attention and pure multi-layer perceptrons (MLP) models for vision have shown great potential in achieving promising performance with fewer inductive biases. These models are generally based on learning interaction among spatial locations from raw data. The complexity of self-attention and MLP grows quadratically as the image size increases, which makes these models hard to scale up when high-resolution features are required. In this paper, we present the Global Filter Network (GFNet), a conceptually simple yet computationally efficient architecture, that learns long-term spatial dependencies in the frequency domain with log-linear complexity. Our architecture replaces the self-attention layer in vision transformers with three key operations: a 2D discrete Fourier transform, an element-wise multiplication between frequency-domain features and learnable global filters, and a 2D inverse Fourier transform. We exhibit favorable accuracy/complexity trade-offs of our models on both ImageNet and downstream tasks. Our results demonstrate that GFNet can be a very competitive alternative to transformer-style models and CNNs in efficiency, generalization ability and robustness. Code is available at [this https URL](https://github.com/yongmingrao/gfnet).

Comments: Accepted to NeurIPS 2021. Project page: [this https URL](https://github.com/yongmingrao/gfnet)

Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Artificial Intelligence (cs.AI); Machine Learning (cs.LG)

Cite as: [arXiv:2107.00645 \[cs.CV\]](https://arxiv.org/abs/2107.00645)

(or [arXiv:2107.00645v2 \[cs.CV\]](https://arxiv.org/abs/2107.00645v2) for this version)

<https://doi.org/10.48550/arXiv.2107.00645>

Alternative to transformers

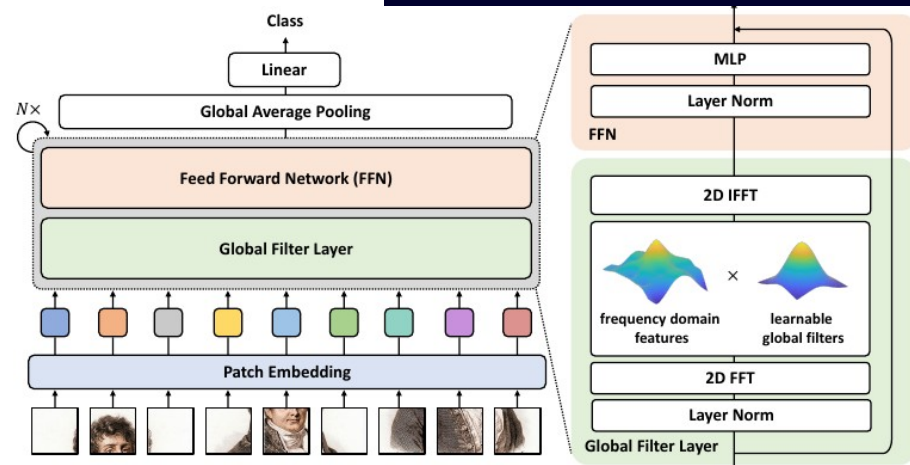


Figure 1: **The overall architecture of the Global Filter Network.** Our architecture is based on Vision Transformer (ViT) models with some minimal modifications. We replace the self-attention sub-layer with the proposed *global filter layer*, which consists of three key operations: a 2D discrete Fourier transform to convert the input spatial features to the frequency domain, an element-wise multiplication between frequency-domain features and the global filters, and a 2D inverse Fourier transform to map the features back to the spatial domain. The efficient fast Fourier transform (FFT) enables us to learn arbitrary interactions among spatial locations with log-linear complexity.

# Double Descent Learning Curve

DETAILS RELATIONS

PNAS  
Cyanobacterial symbionts and cryptic lineage  
Aug 2019

Proceedings of the National Academy of Sciences  
Vol. 116, No. 32  
Aug 2019

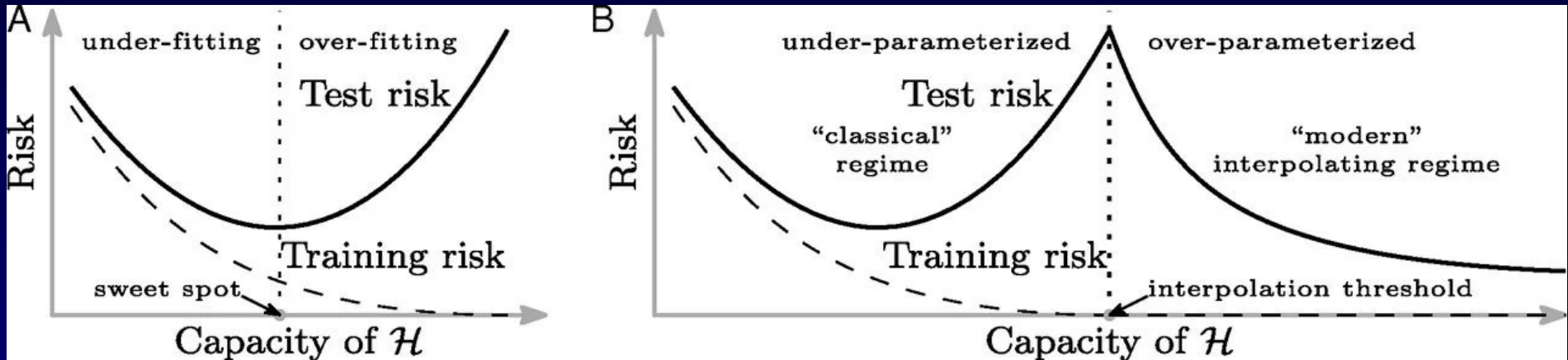
## Reconciling modern machine-learning practice and the classical bias–variance trade-off

Mikhail Belkin<sup>a,b,1</sup>, Daniel Hsu<sup>c</sup>, Siyuan Ma<sup>a</sup>, and Soumik Mandal<sup>a</sup>

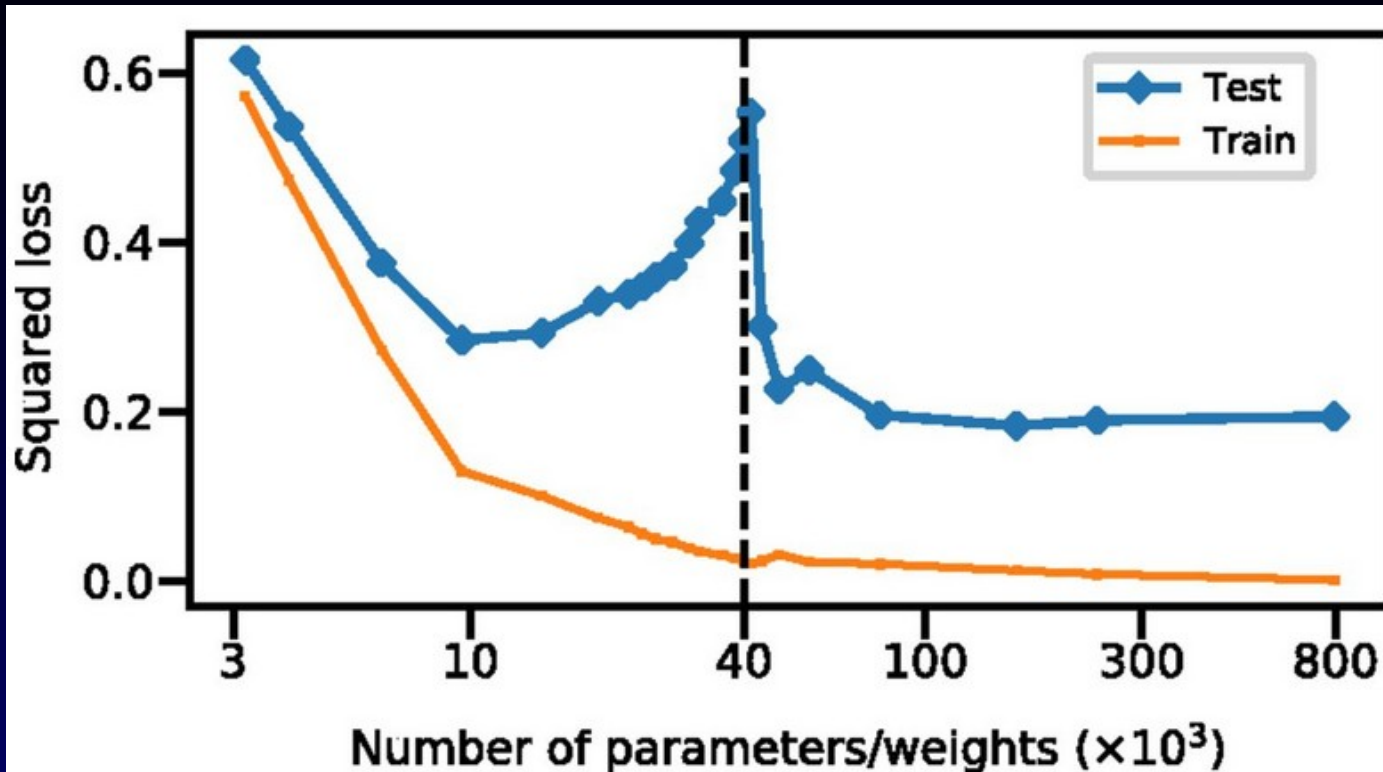
<sup>a</sup>Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210; <sup>b</sup>Department of Statistics, The Ohio State University, Columbus, OH 43210; and <sup>c</sup>Computer Science Department and Data Science Institute, Columbia University, New York, NY 10027

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved July 2, 2019 (received for review February 21, 2019)

Check for updates



# Double Descent Learning Curve



Double-descent risk curve for a fully connected neural network on MNIST. Shown are training and test risks of a network with a single layer of  $H$  hidden units, learned on a subset of MNIST ( $n = 4 \cdot 10^3$ ,  $d = 784$ ,  $K = 10$  classes). The number of parameters is  $(d + 1) \cdot H + (H + 1) \cdot K$ . The interpolation threshold (black dashed line) is observed at  $n \cdot K$ .

# Nested Learning

Computer Science > Machine Learning

[Submitted on 31 Dec 2025]

## Nested Learning: The Illusion of Deep Learning Architectures

Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, Vahab Mirrokni

Despite the recent progresses, particularly in developing Language Models, there are fundamental challenges and unanswered questions about how such models can continually learn/memorize, self-improve, and find effective solutions. In this paper, we present a new learning paradigm, called Nested Learning (NL), that coherently represents a machine learning model with a set of nested, multi-level, and/or parallel optimization problems, each of which with its own context flow. Through the lenses of NL, existing deep learning methods learn from data through compressing their own context flow, and in-context learning naturally emerges in large models. NL suggests a philosophy to design more expressive learning algorithms with more levels, resulting in higher-order in-context learning and potentially unlocking effective continual learning capabilities. We advocate for NL by presenting three core contributions: (1) Expressive Optimizers: We show that known gradient-based optimizers, such as Adam, SGD with Momentum, etc., are in fact associative memory modules that aim to compress the gradients' information (by gradient descent). Building on this insight, we present other more expressive optimizers with deep memory and/or more powerful learning rules; (2) Self-Modifying Learning Module: Taking advantage of NL's insights on learning algorithms, we present a sequence model that learns how to modify itself by learning its own update algorithm; and (3) Continuum Memory System: We present a new formulation for memory system that generalizes the traditional viewpoint of long/short-term memory. Combining our self-modifying sequence model with the continuum memory system, we present a continual learning module, called Hope, showing promising results on long-context reasoning tasks.

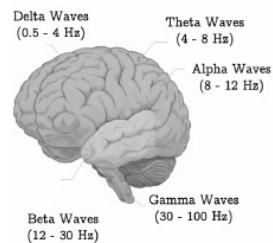
Comments: A version of this work is published at Neural Information Processing Systems (NeurIPS)

Subjects: **Machine Learning (cs.LG)**; Artificial Intelligence (cs.AI)

Cite as: [arXiv:2512.24695](https://arxiv.org/abs/2512.24695) [cs.LG]

(or [arXiv:2512.24695v1](https://arxiv.org/abs/2512.24695v1) [cs.LG] for this version)

<https://doi.org/10.48550/arXiv.2512.24695>



### Uniform and Reusable Structure

Neuroplasticity is the brain's ability to reorganize itself via mechanisms like forming new synapses, strengthening/weakening existing ones, rerouting signals through alternate pathways, etc. Such ability requires uniform and reusable structure across brain.

In Nested Learning (NL), architectures are decomposed into a set of neurons (i.e., linear or locally deep MLPs), each of which with its own context flow and objective. This design provides a uniform and reusable structure for learning.

### Multi Time Scale Update

Brain oscillations (or brain waves) critical for the brain to coordinate its activity. Notably, the brain does not rely on a single centralized clock to synchronize every neuron: the earlier layers update their activity quickly in high-frequency cycles, whereas later layers integrate information over longer, slower cycles.

In NL, parameters in each "level" are updated with their own specific frequency and does not rely on a single centralized clock. The HOPE's design allows the earlier layers update their activity quickly in high-frequency cycles, whereas later layers integrate information over longer, slower cycles.

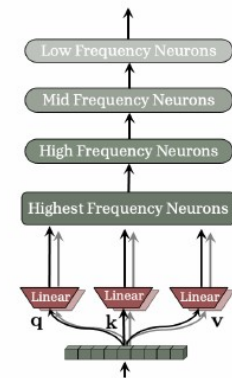


Figure 1: The uniform and reusable structure as well as multi time scale update in the brain are the key components to unlock the continual learning in humans. Nested Learning (NL) allows for multi time-scale update for each component of the brain, while showing that well-known architectures such as Transformers are in fact linear layers with different frequency updates.

# Nested Learning

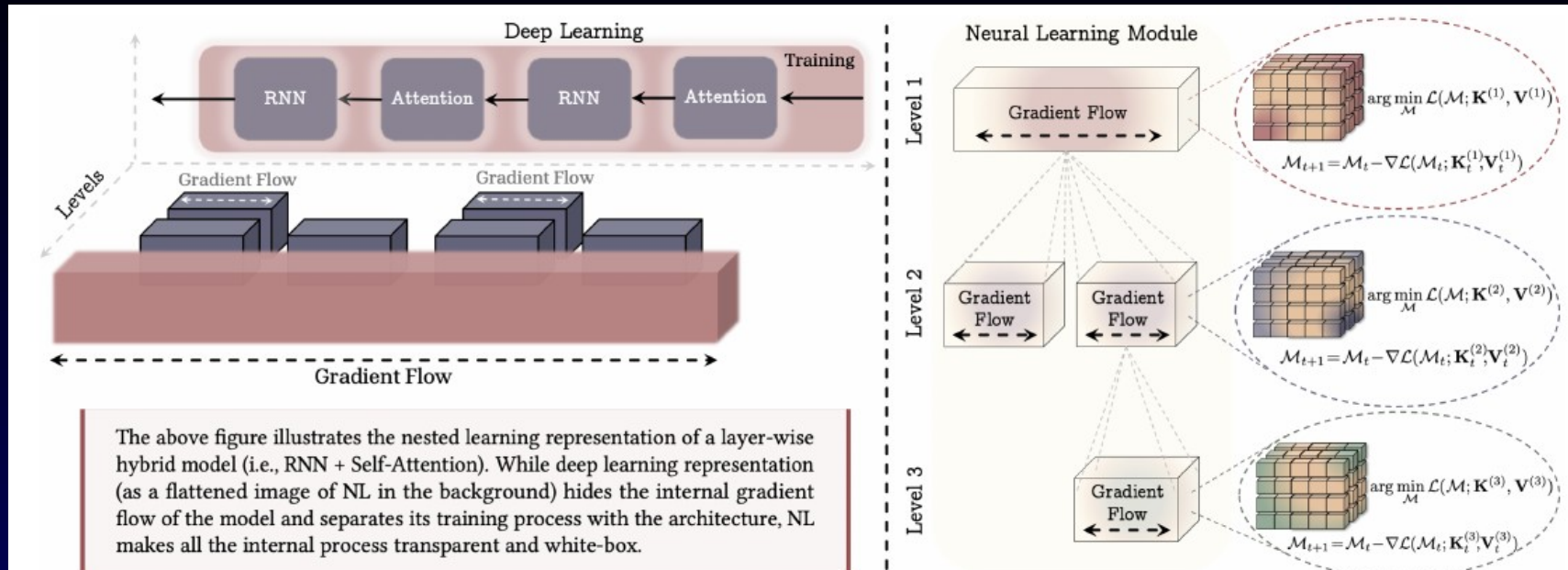


Figure 2: Nested Learning Paradigm that represent a machine learning model and its training procedure as a set of nested optimization problems. **(Left)** An example of Hybrid architecture. While deep learning perspective, as the flattened image of NL, does not provide insight about the depth of computation in the blocks, NL transparently represent all the inner gradient flows. **(Right)** A Neural Learning Module: A computational model that learns how to compress its own context flow. For example, the first level corresponds to the model's most outer-loop training, often refer to as "pre-training" step.

# Quantum Machine Learning (QML)

Bloch sphere for qubit representation

Noisy Intermediate-Scale Quantum era (NISQ)

Quantum gates (CNOT,X,Y,Z,Hadamart=entangling)

Quantum circuits = Ansatz

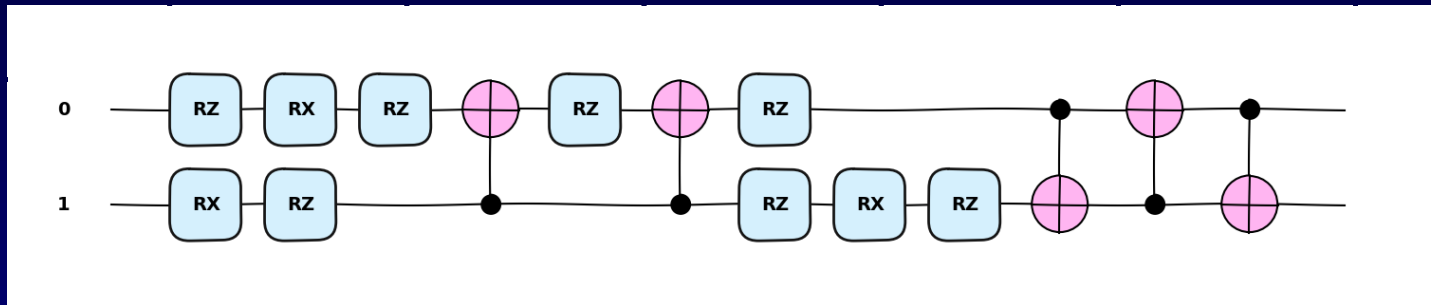
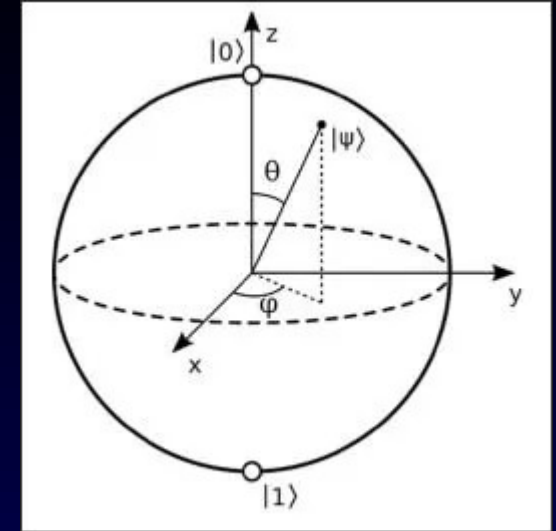
IBM

Qiskit (Josephson)

Xanadu

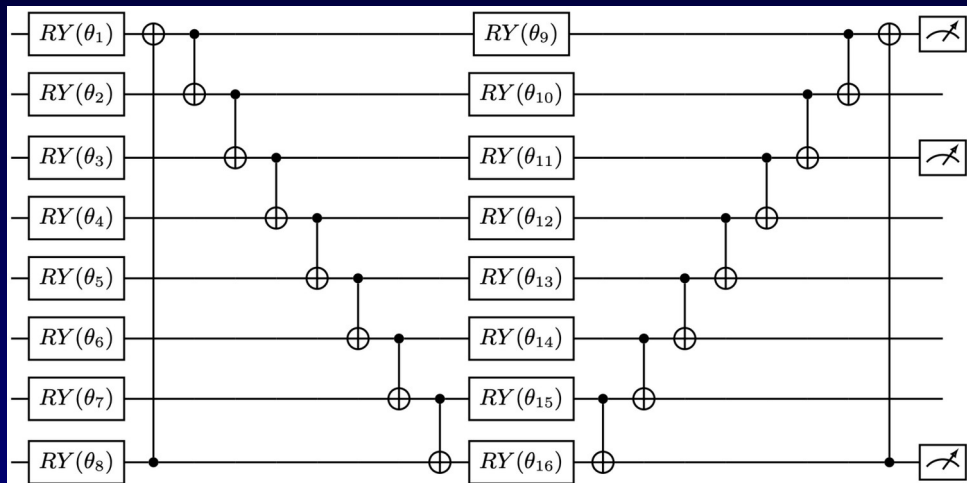
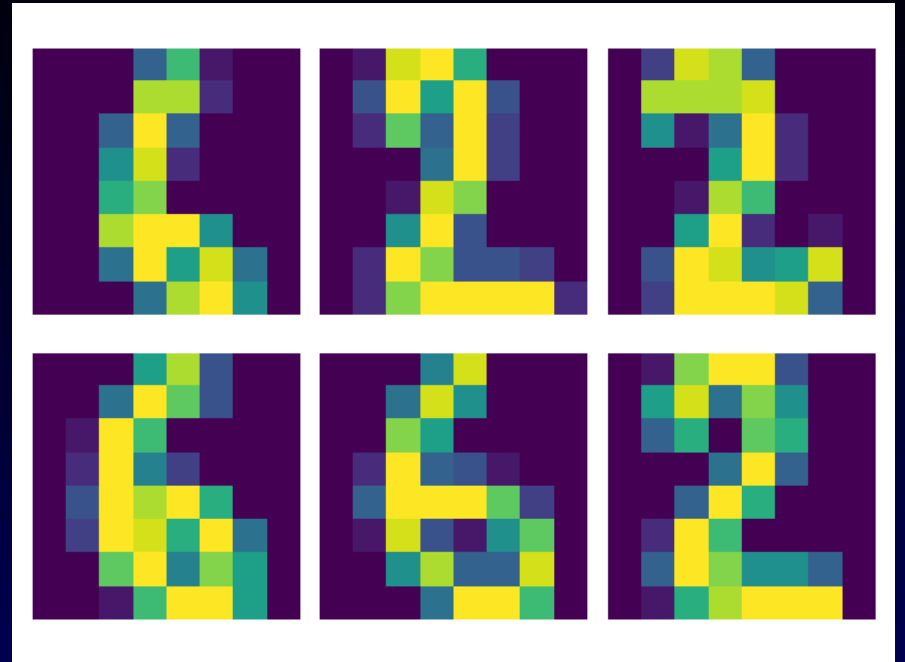
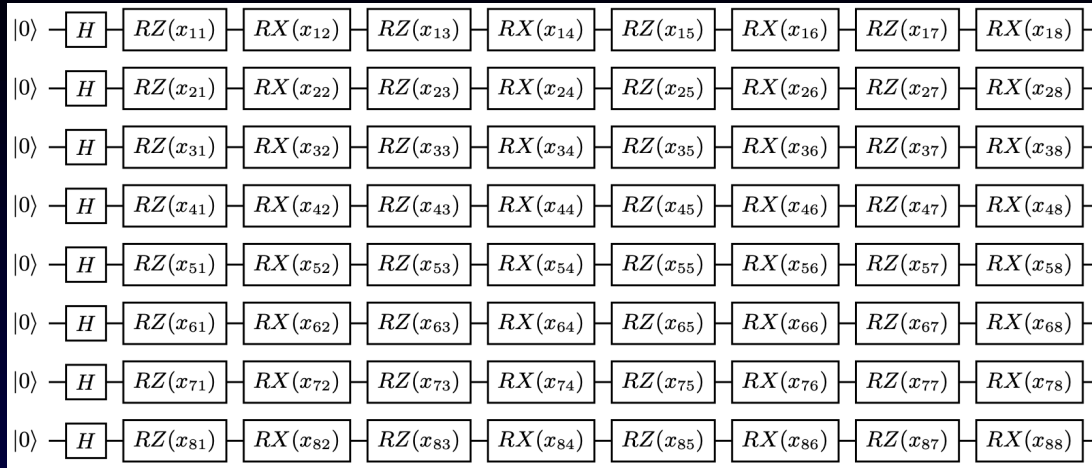
Strawberry Fields (photonic device)

Penny Lane (general quantum device simulator+API to HW)



QFT  
Quantum Fourier Transform

# QML: MNIST-Like Classification



# Conclusions

Astronomical Spectroscopy is the key to physics

Analysis of surveys requires VO + using ML (AI)

AI helps us to discover the secrets of the Universe

ML is returning to the physics theory

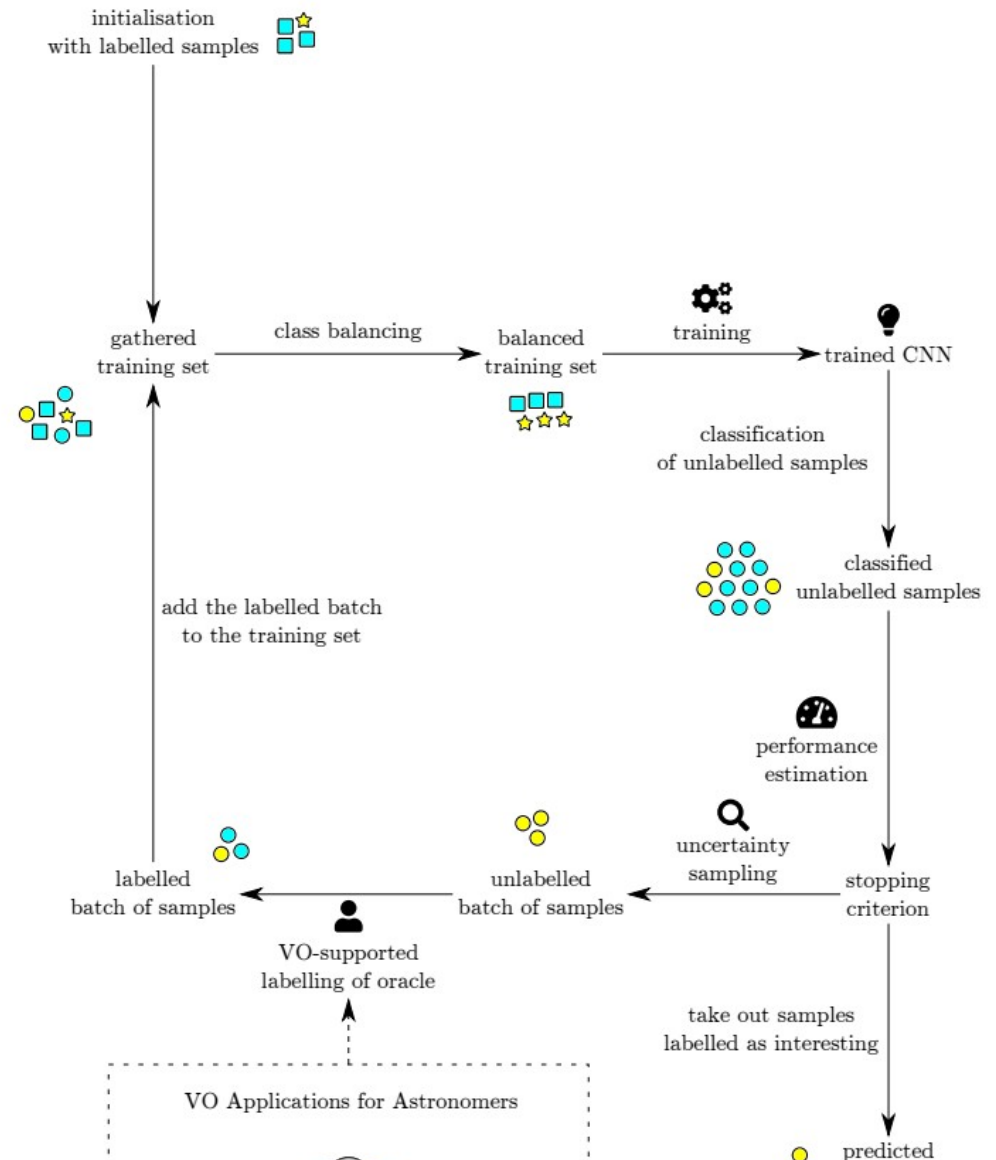
The future is exciting !

## **Backup Slides – Our ML Experiments**

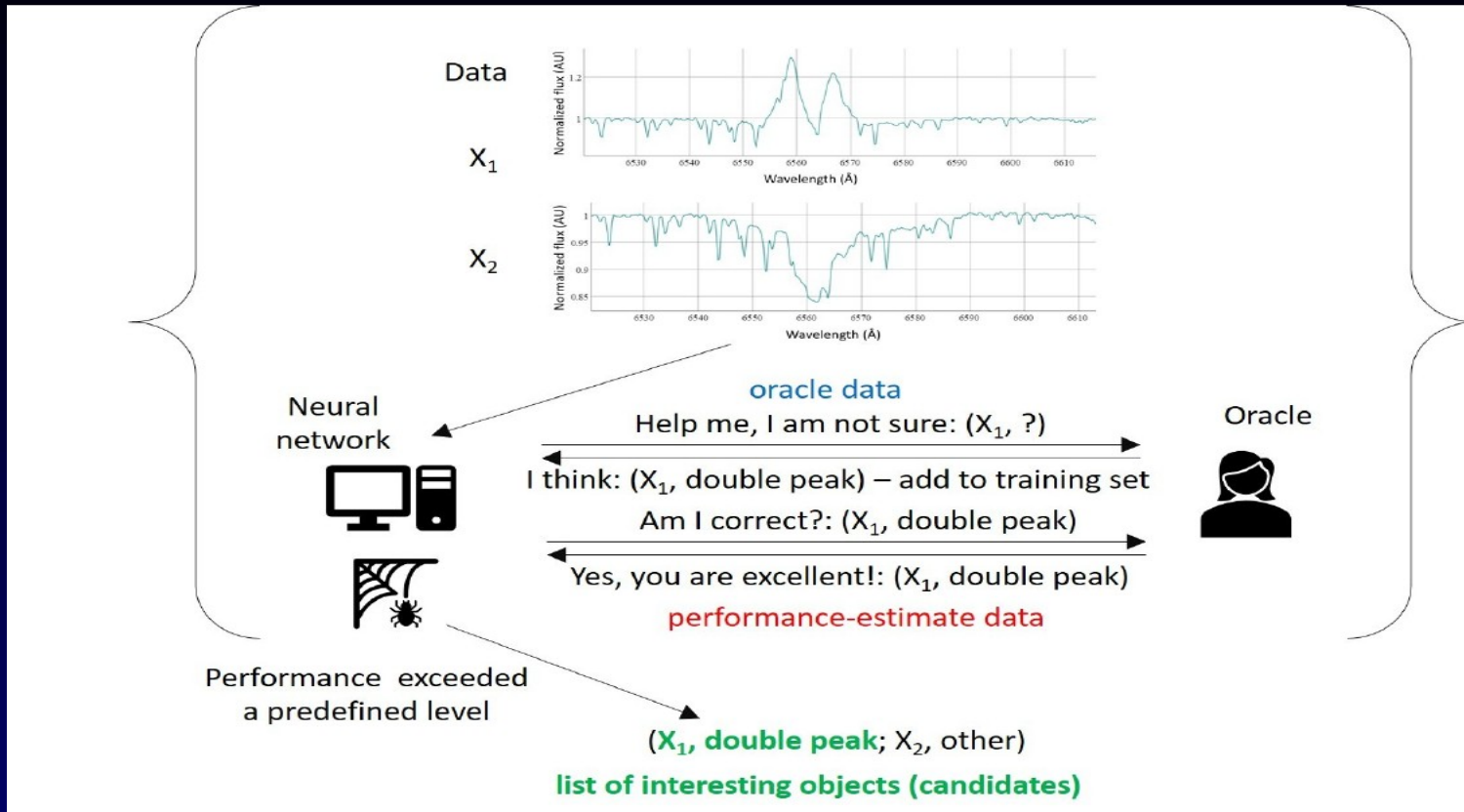
*Thank you*

# Active Deep Learning: CNN Chooses Data for Its Training

- Oracle classification (domain expert knowledge)
- Uncertainty sampling (entropy)
- From predicted target classes selected batch of 100

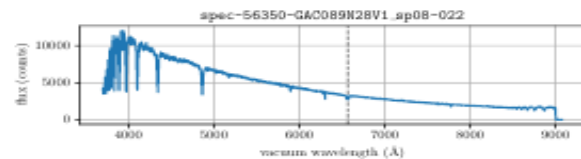
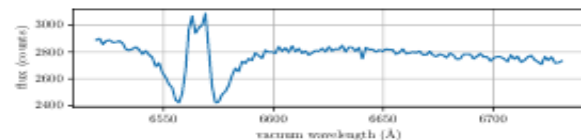
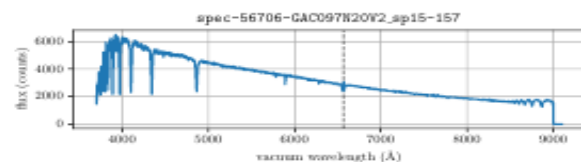
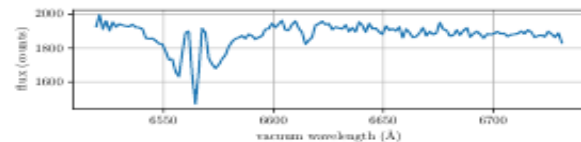
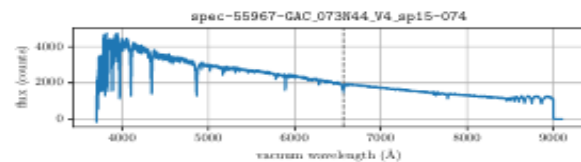
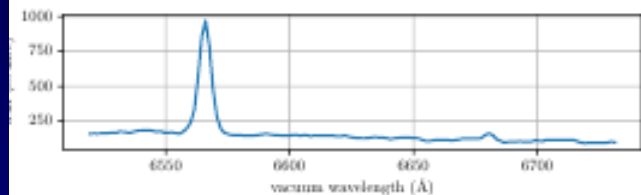
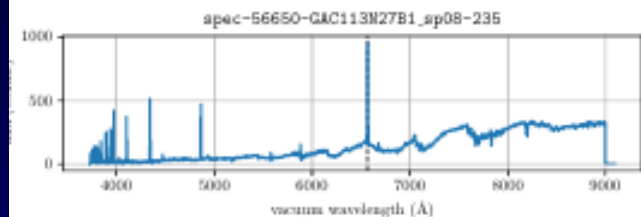
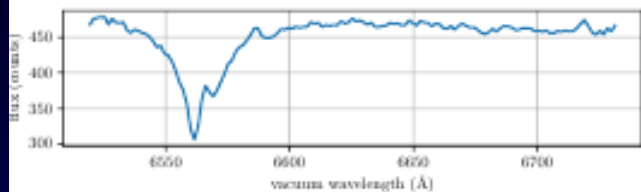
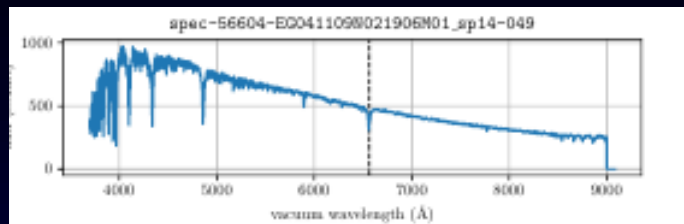


# Active Learning (insufficient labels)



Oracle : Human – Machine Interaction

# Single-peak & Double-peak Candidates



# Publication in A&A

Free Access

Issue	A&A Volume 643, November 2020
Article Number	A122
Number of page(s)	14
Section	Catalogs and data
DOI	<a href="https://doi.org/10.1051/0004-6361/201936090">https://doi.org/10.1051/0004-6361/201936090</a>
Published online	11 November 2020

A&A 643, A122 (2020)

## Active deep learning method for the discovery of objects of interest in large spectroscopic surveys<sup>\*,\*\*</sup>

 P. Škoda<sup>1,2</sup>,  O. Podsztavek<sup>2</sup> and P. Tvrđík<sup>2</sup>



\* Catalogues of our emission-line candidates are only available at the CDS via anonymous ftp to [cdsarc.u-strasbg.fr](https://cdsarc.u-strasbg.fr) (130.79.128.5) or via <http://cdsarc.u-strasbg.fr/viz-bin/cat/J/A+A/643/A122>

\*\* Based on spectra obtained with 2 m Perek Telescope of Ondřejov observatory, Czech Republic and archival LAMOST DR2 spectra.

# New Catalogue on VizieR



Portal Simbad VizieR Aladin X-Match Other ~ Help



## Active deep learning in large spectros. surveys : J/A+A/643/A122

Access to



FTP

ReadMe



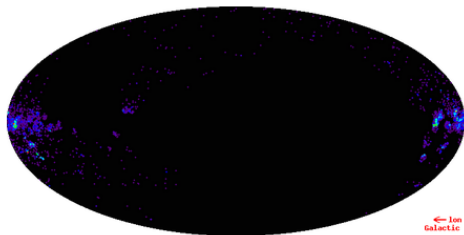
Authors : Skoda P. , Podsztavek O., Tvrdik P.

VizieR DOI : [10.26093/cds/vizier.36430122](https://cds.u-strasbg.fr/vizier/36430122) [Cite](#)

Bibcode : [2020A&A...643A.122S \(ADS\)](#)

UAT : Emission line stars, Surveys, Spectroscopy

Compilation (CCC)



Inserted into VizieR : 11-Nov-2020

Last modification : 02-Feb-2021

Article Origin

Description

See also

Prov

FTP

VizieR

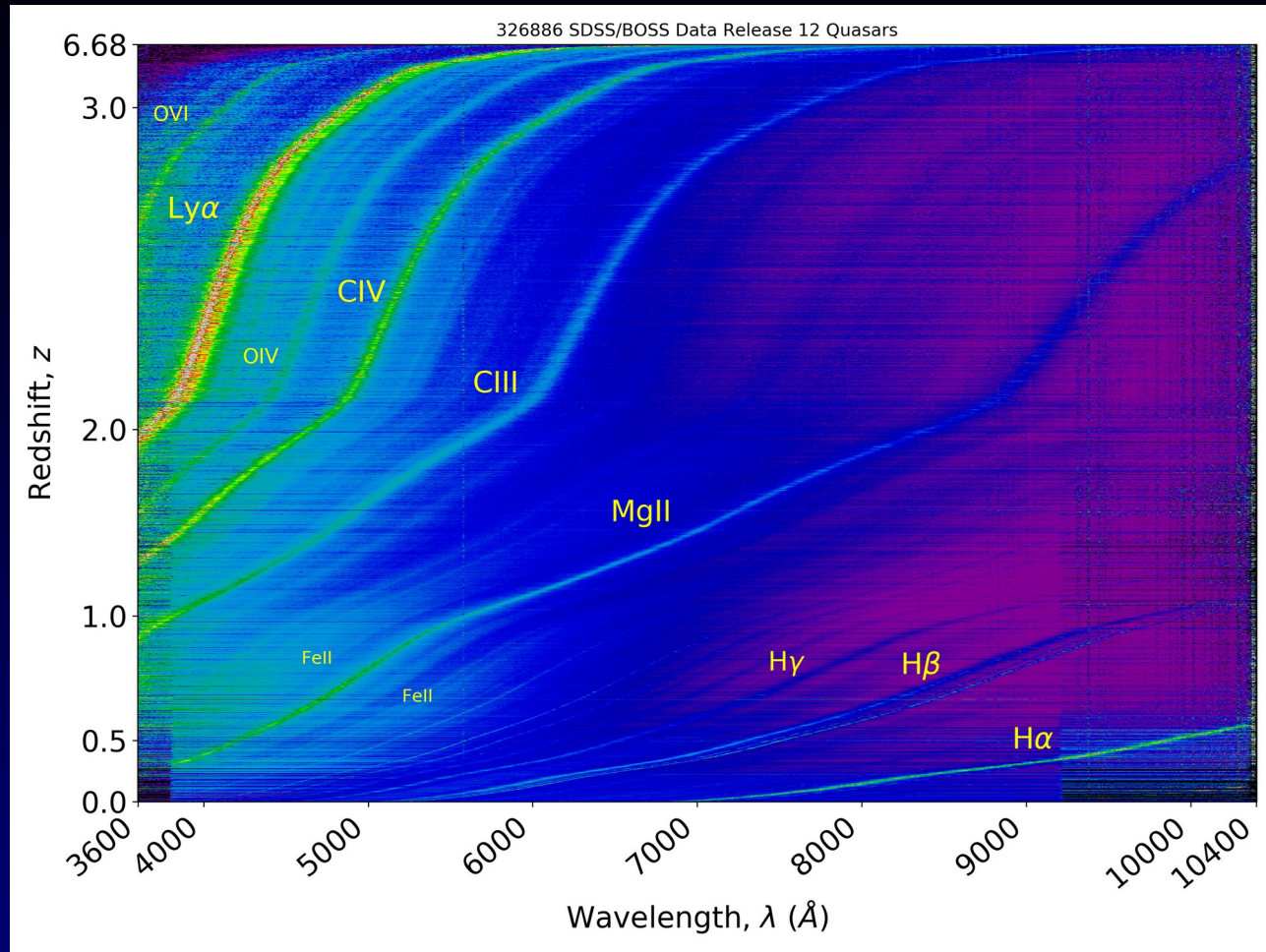
**Active deep learning method for discovery of objects of interest in large spectroscopic surveys.** (2020)

[Go to the original article \(10.1051/0004-6361/201936090\)](#)

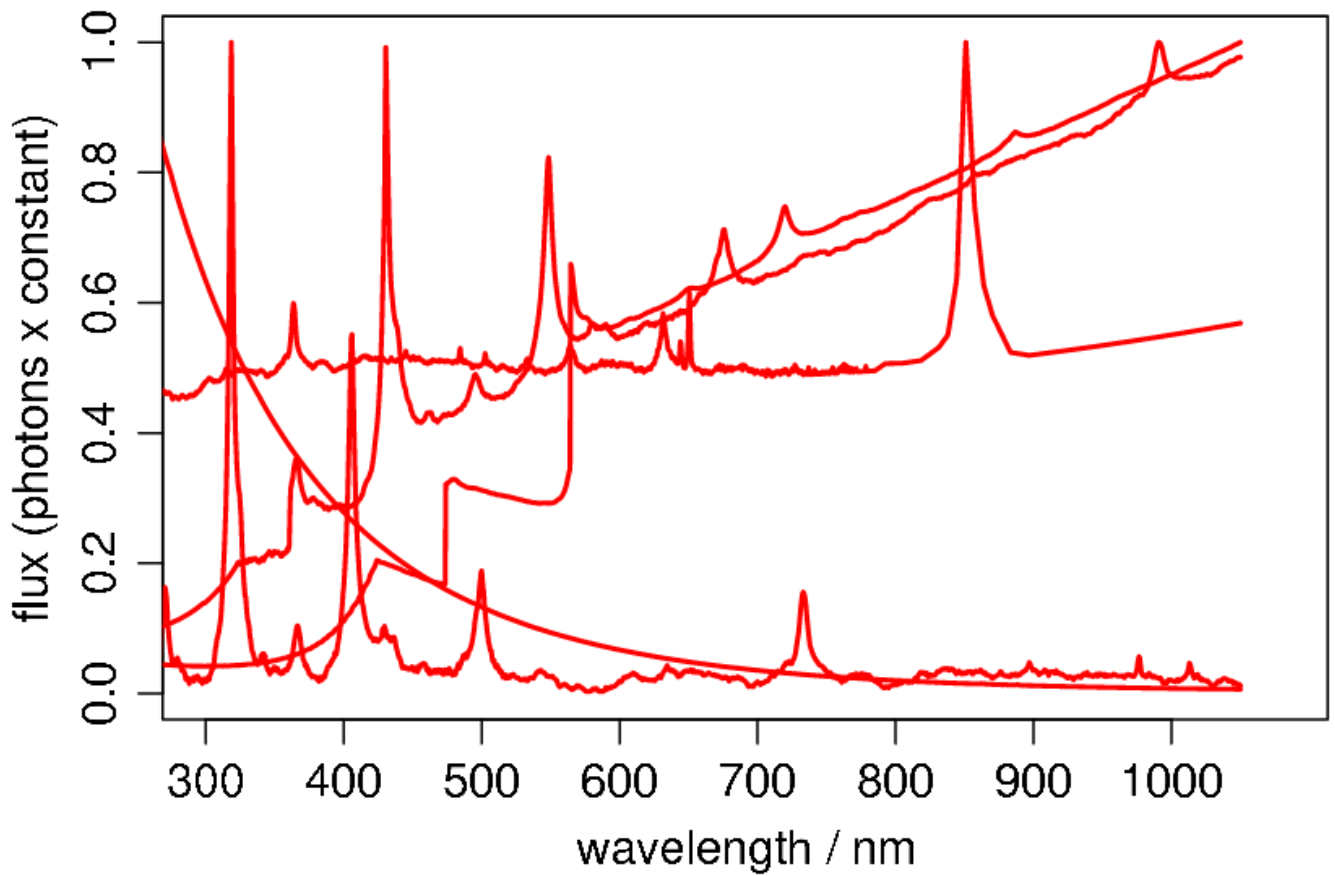
**Keywords :** surveys - virtual observatory tools - methods statistical - techniques: spectroscopic - stars: emission-line, Be - line: profiles

**Abstract:**Current archives of the LAMOST telescope contain millions of pipeline-processed spectra that have probably never been seen by human eyes. Most of the rare objects with interesting physical properties, however, can only be identified by visual analysis of their characteristic spectral features. A proper combination of interactive visualisation with modern machine learning techniques opens new ways to discover such objects. We apply active learning classification methods supported by deep convolutional neural networks to automatically identify complex emission-line shapes in multi-million spectra archives. We used the pool-based uncertainty sampling active learning method driven by a custom-designed deep convolutional neural network with 12 layers. The architecture of the network was inspired by VGGNet, AlexNet, and ZFNet, but it was adapted for operating on one-dimensional feature vectors. The unlabelled pool set is represented by 4.1 million spectra from the LAMOST data release 2 survey. The initial training of the network was performed on a labelled set of about 13000 spectra obtained in the 400Å wide region around H $\alpha$  by the 2m Perek telescope of the Ondrejov observatory, which mostly contains spectra of Be and related early-type stars. The differences between the Ondrejov intermediate-resolution and the LAMOST low-resolution spectrographs were compensated for by Gaussian blurring and wavelength conversion. After several iterations, the network was able to successfully identify emission-line stars with an error smaller than 6.5%. Using the technology of the Virtual Observatory to visualise the results, we discovered 1013 spectra of 948 new candidates of emission-line objects in addition to 664 spectra of 549 objects that are listed in SIMBAD and 2644 spectra of 2291 objects identified in an earlier paper of a Chinese group led by Wen Hou. The most interesting objects with unusual spectral properties are discussed in detail. ([hide](#))

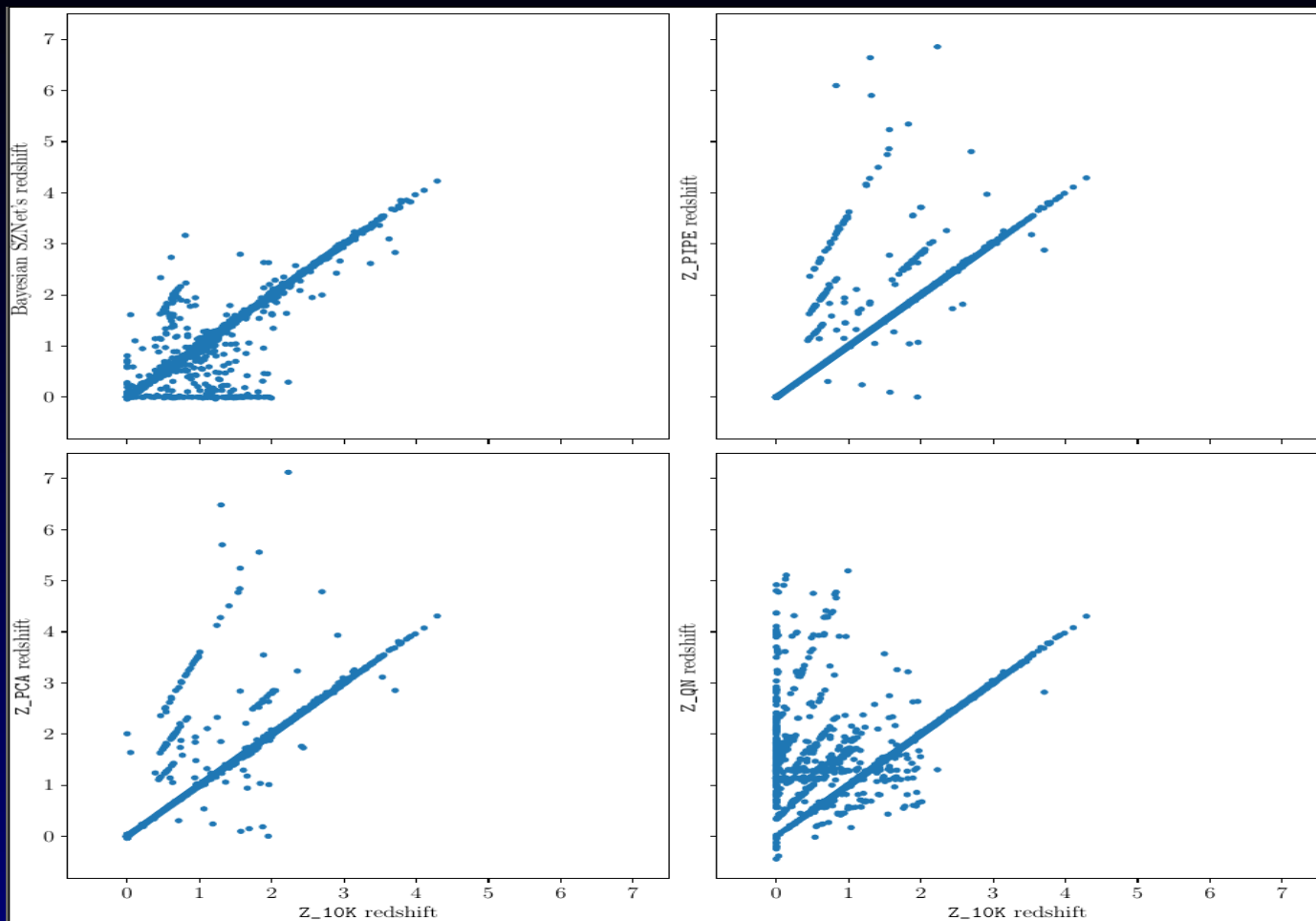
# Redshift



# SDSS Template Library



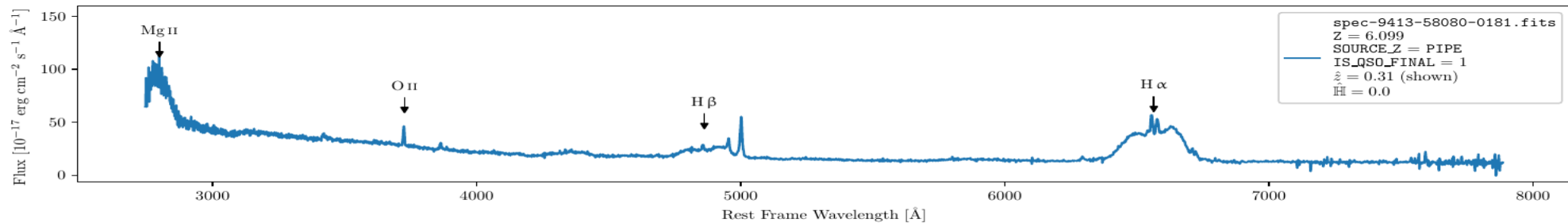
# Systematic Errors in Pipelines



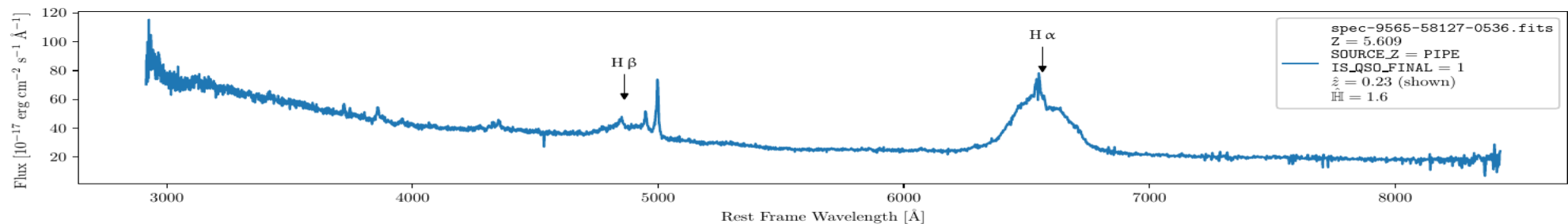
Z\_10K:  
~10000 randomly  
selected spectra  
visually checked

Used to evaluate SDSS  
DR16Q pipeline

# BNN corrects the SDSS pipeline

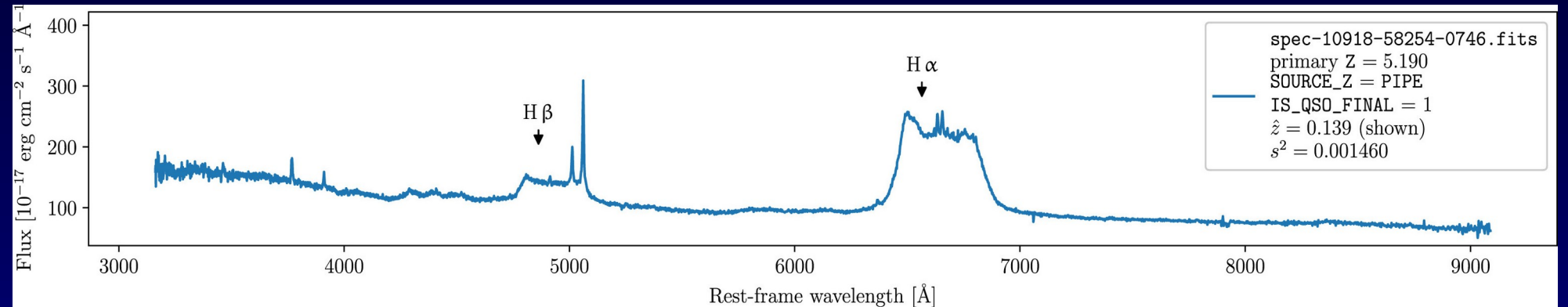
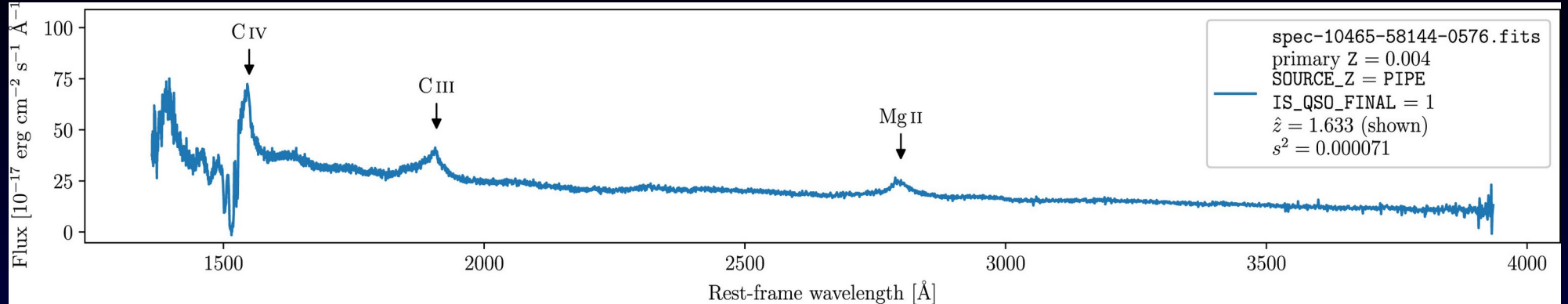


**Figure B6.** Spectrum with incorrectly high redshift prediction by the pipeline. The Bayesian CNN correctly predicted  $\hat{z} = 0.31$  with  $\hat{H} = 0$ .

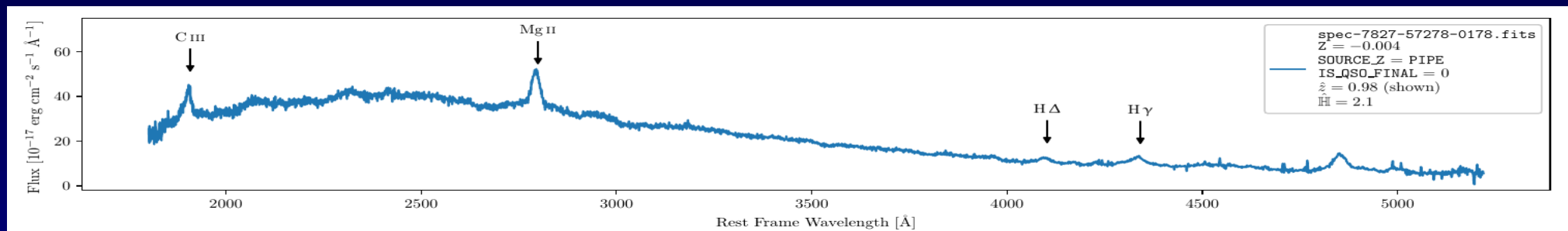
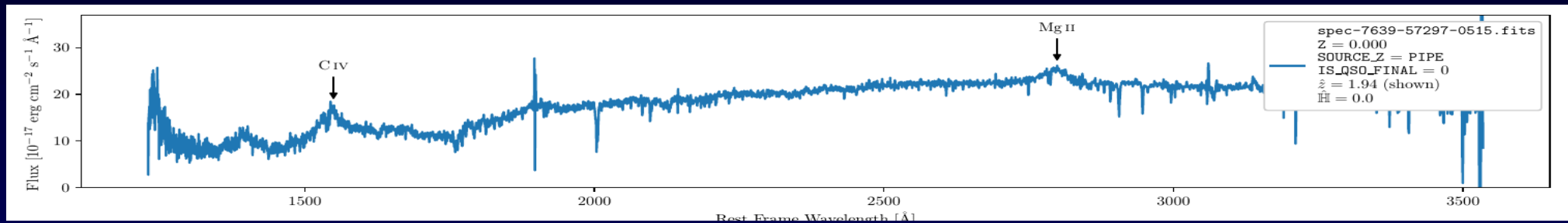
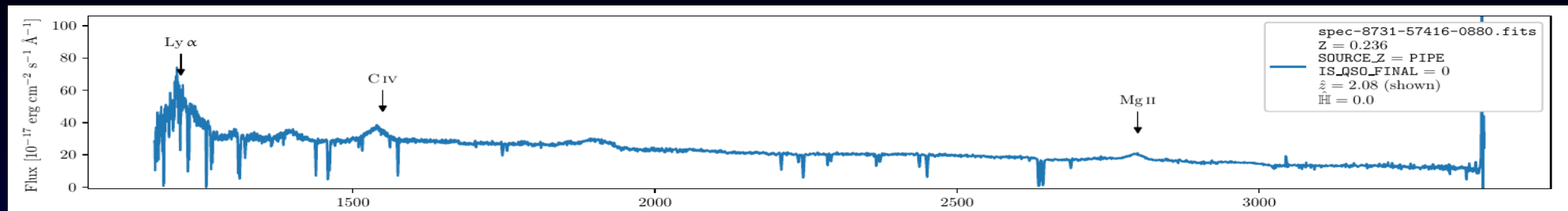


**Figure B7.** Spectrum with incorrectly high redshift prediction by the pipeline. The Bayesian CNN correctly predicted  $\hat{z} = 0.23$  with  $\hat{H} = 1.6$ .

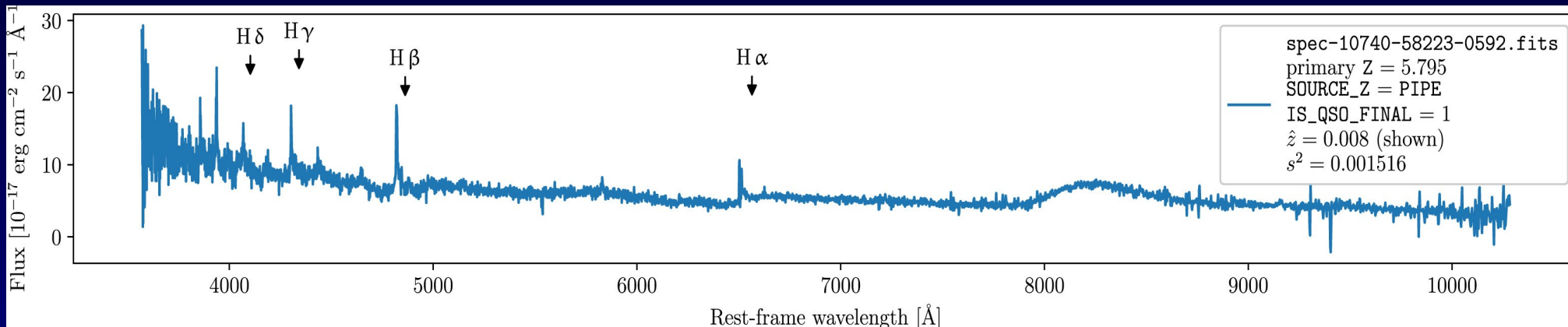
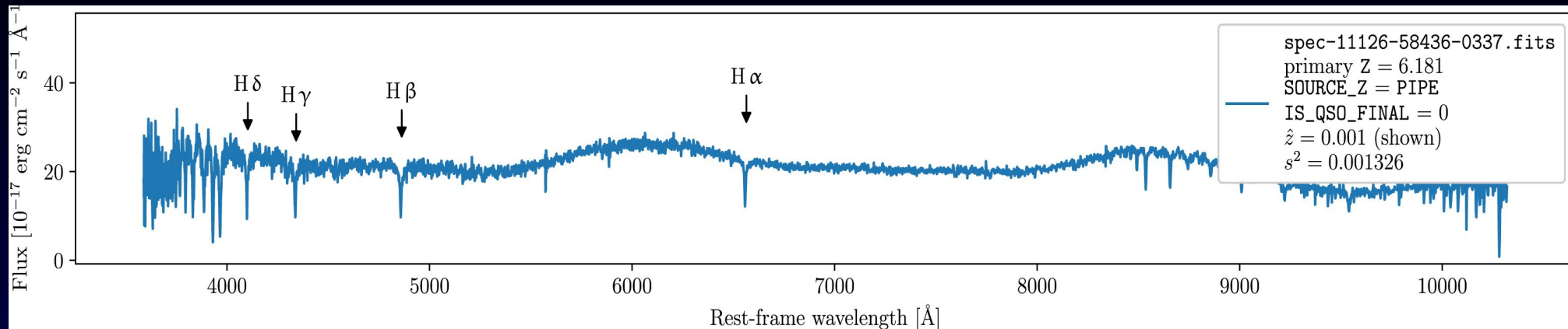
# BDN Corrects the SDSS Pipeline



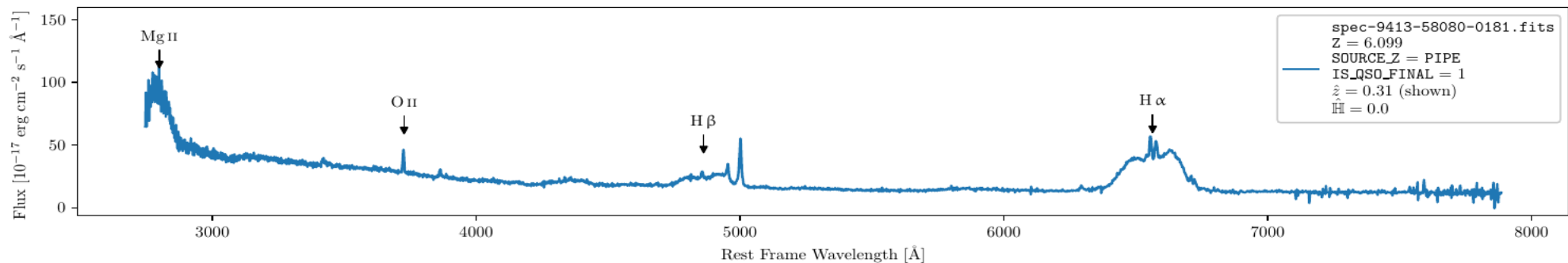
# QSOs missing due to SDSS pipeline error



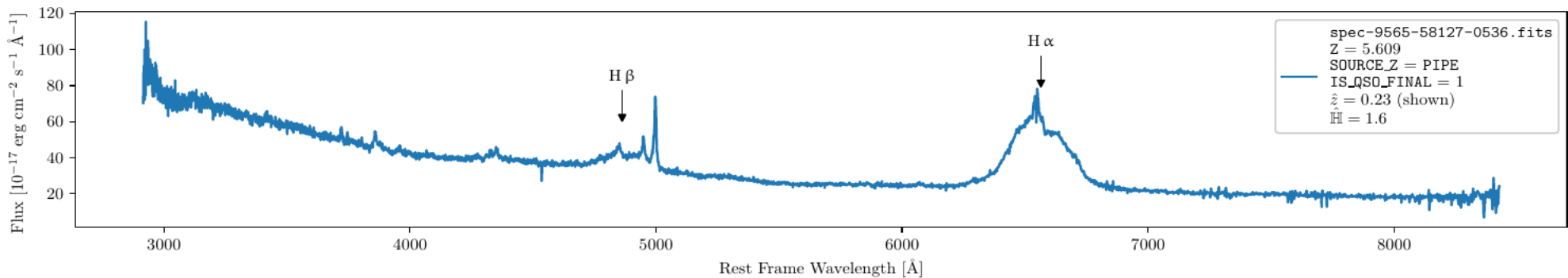
# SDSS Predicts QSO But It Is a Star



# Eliminating false high-z QSO

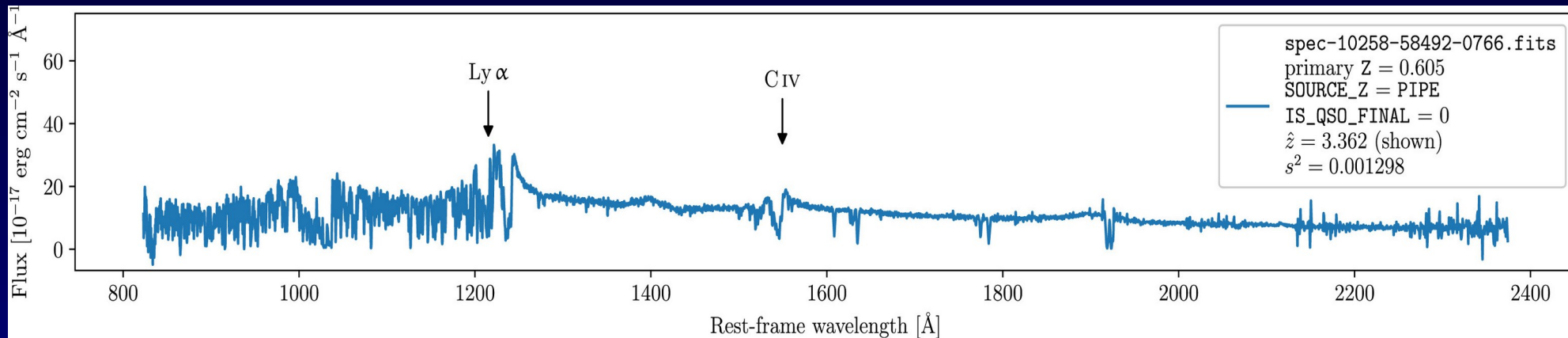
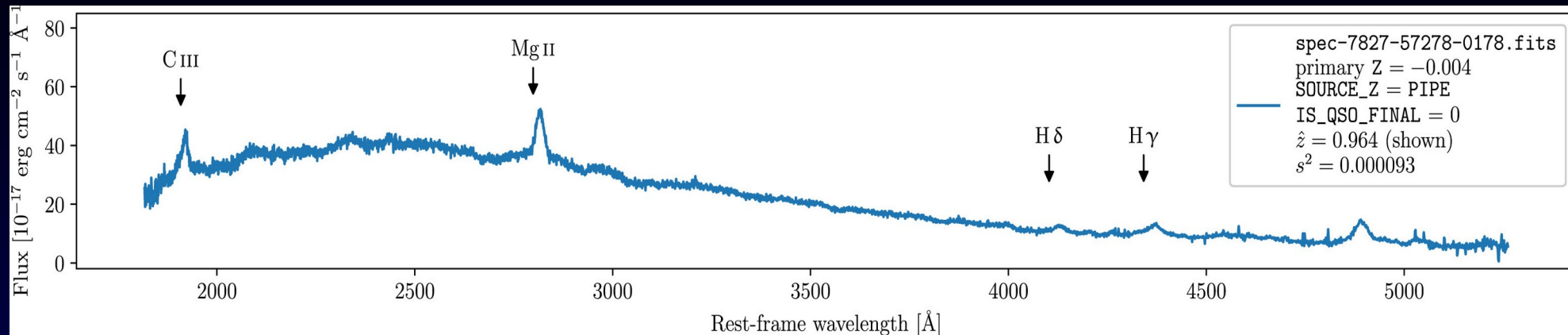


**Figure B6.** Spectrum with incorrectly high redshift prediction by the pipeline. The Bayesian CNN correctly predicted  $\hat{z} = 0.31$  with  $\hat{H} = 0.0$ .

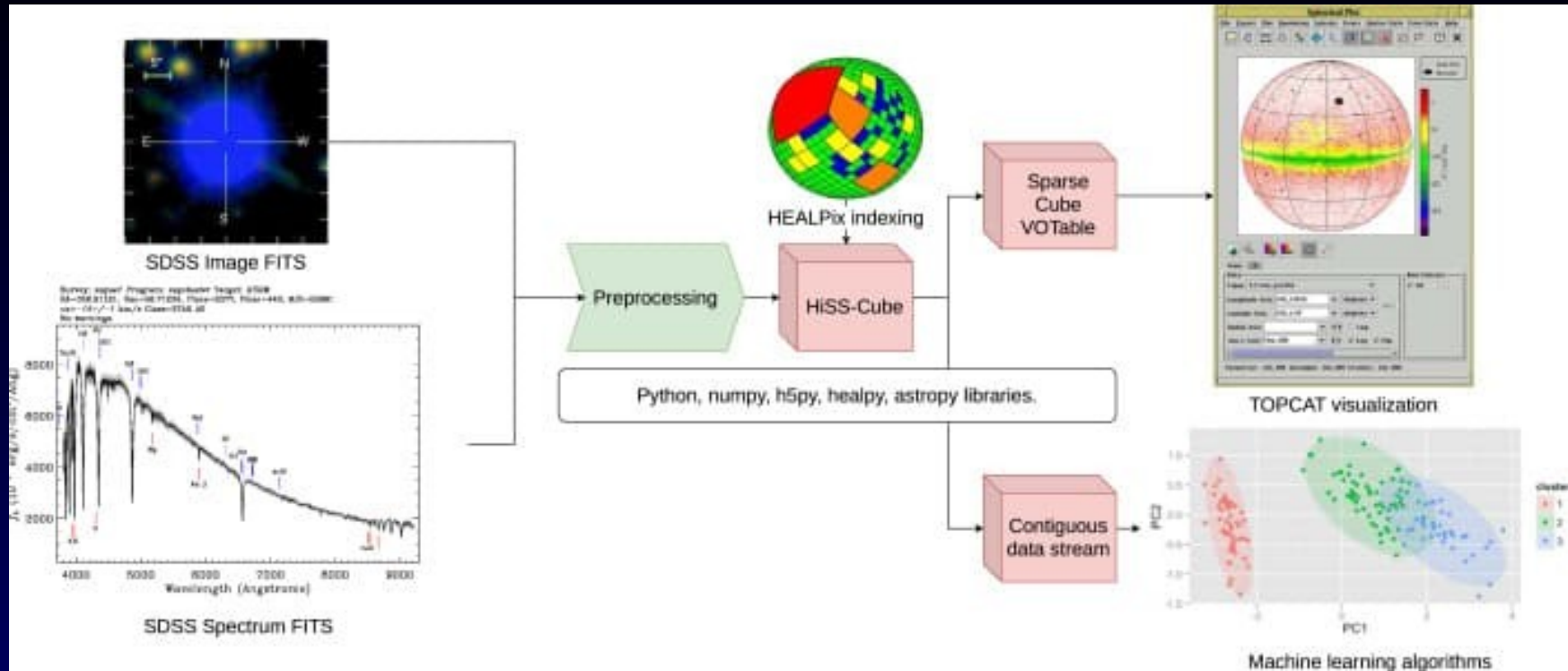


**Figure B7.** Spectrum with incorrectly high redshift prediction by the pipeline. The Bayesian CNN correctly predicted  $\hat{z} = 0.23$  with  $\hat{H} = 1.6$ .

# QSOs missing due to SDSS pipeline error

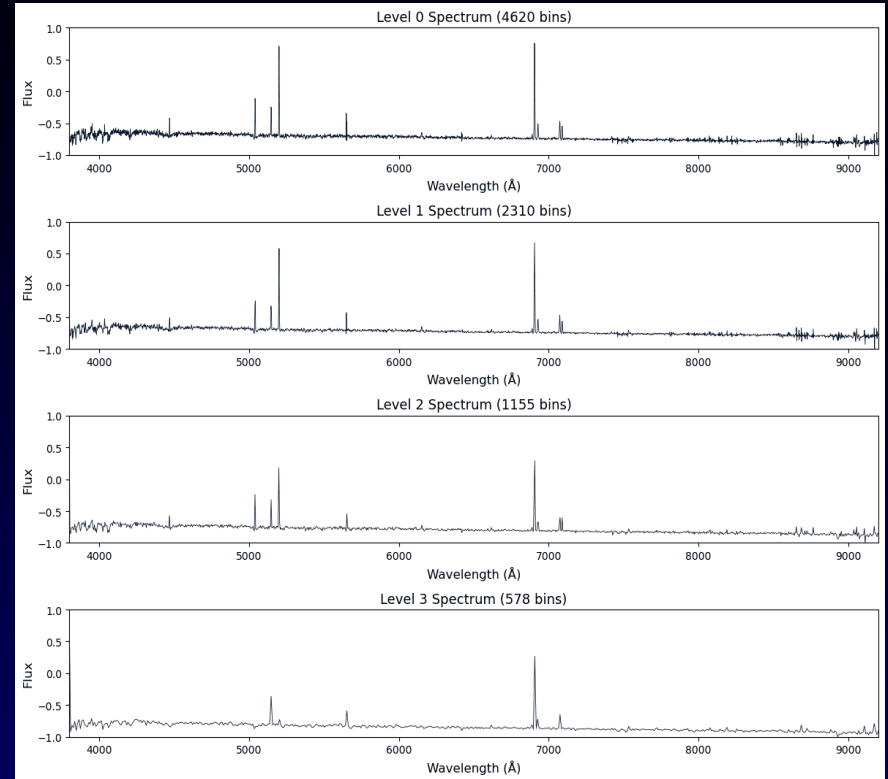
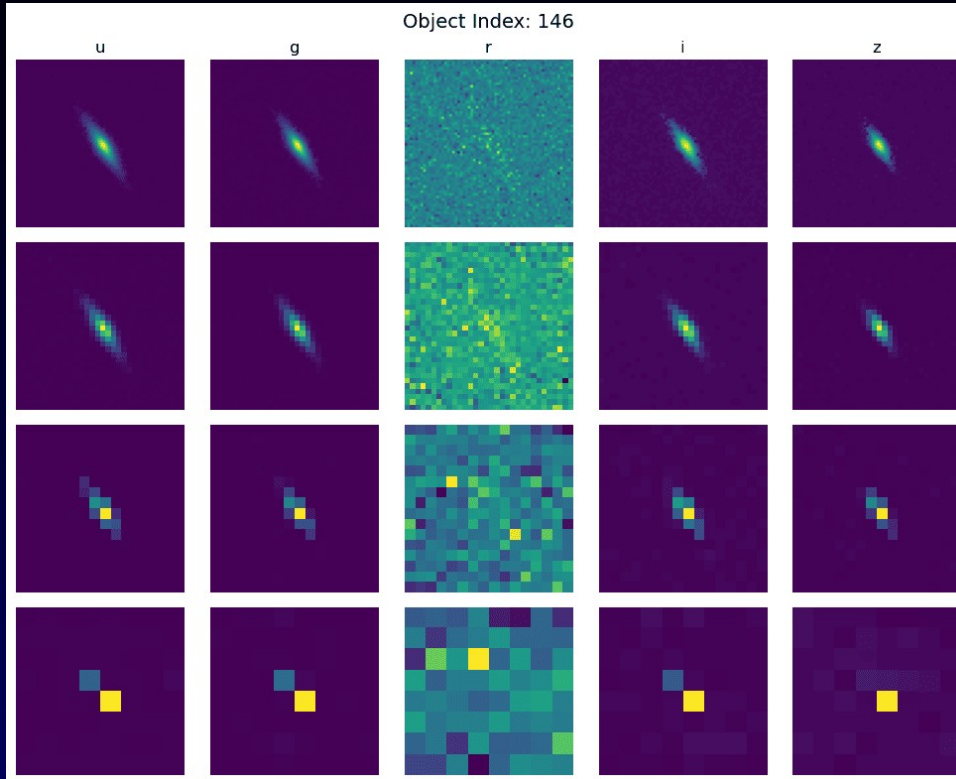


# Hierarchical Sparse Data Cube from SDSS



200TB HiSSCube – Czech national supercomputer Karolina

# SFR by Multimodal Learning of SDSS



Best performance – highest photo resolution, spectra L2, early fusion