Astronomicky velká Velká data
Virtuální observatoř
a AI v astronomii

Petr Škoda

Astronomický ústav AVČR

Setkání Kosmologické sekce ČAS
Matematický ústav AVČR, Praha, 11.3.2026

Credits

- The presentation is based on many different sources – mainly the on-line published slides from IVOA meetings, slides from Astroinformatics workshops, ADASS conferences, COST Action TD1403 or pictures found on Internet.

Outline of the Talk

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

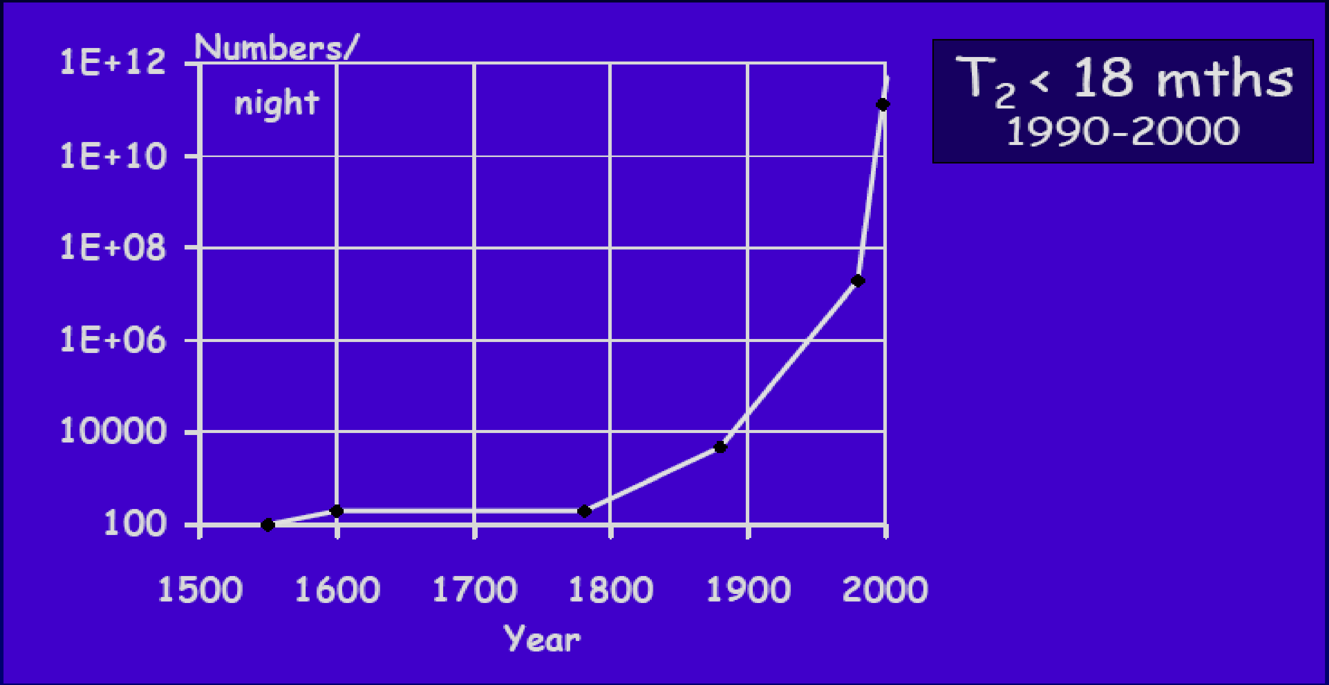
Astronomically
Big Data

Data Avalanche

Moore law for chips –doubling 1.5 year

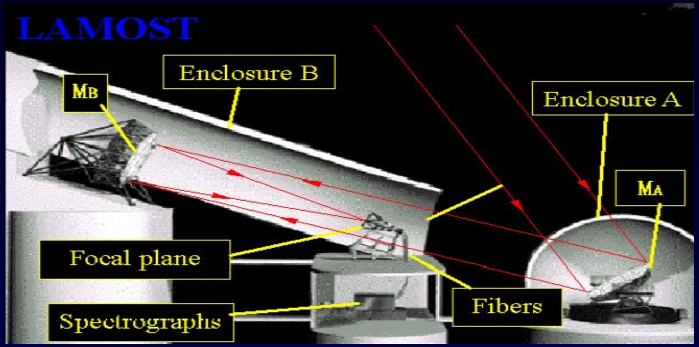
Data in astronomy – doubling < 1 yr !

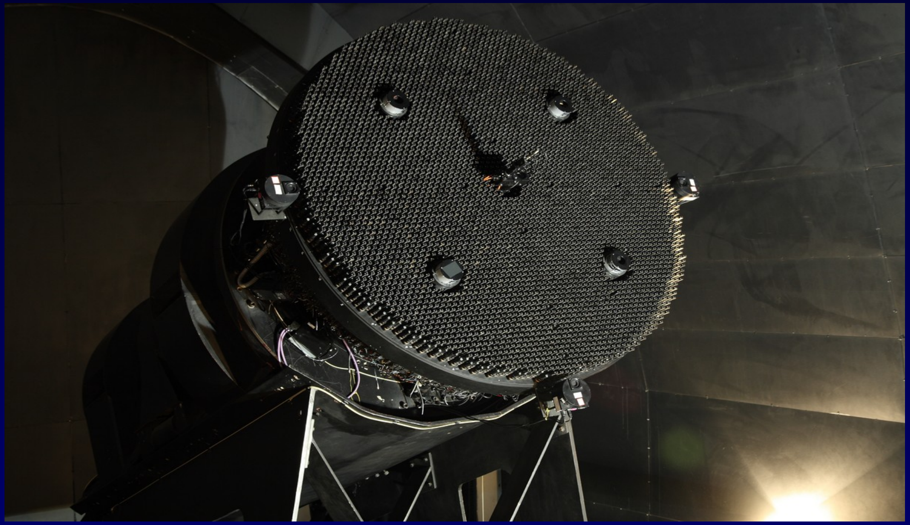
100 PB today,   100 TB/night



LAMOST (Guoshoujing)

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





LAMOST Spectral Surveys

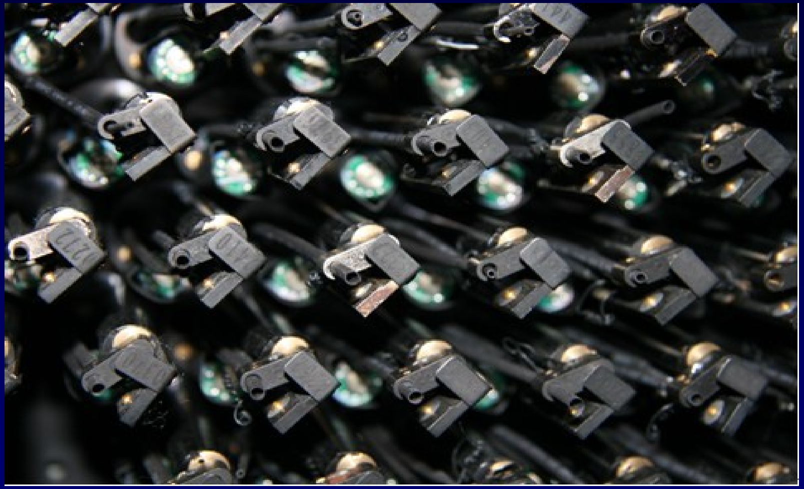
DR1 (end 2013) **2 204 860** spectra
1 085 404 stars classified by pipeline

DR2 (beg 2015) **4 132 782** spectra
3 779 674 stars
307 000 unknown!

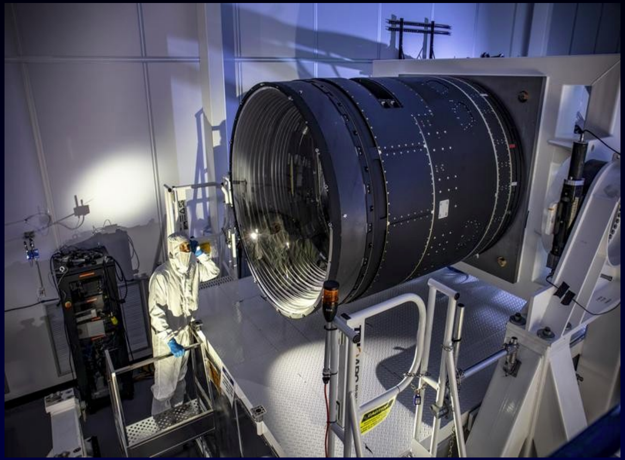
DR12 (Mar 2025) 12 602 390 low res
+ 15 475 985 mid res

Each fibre – 2 motors
double arm 33mm circle

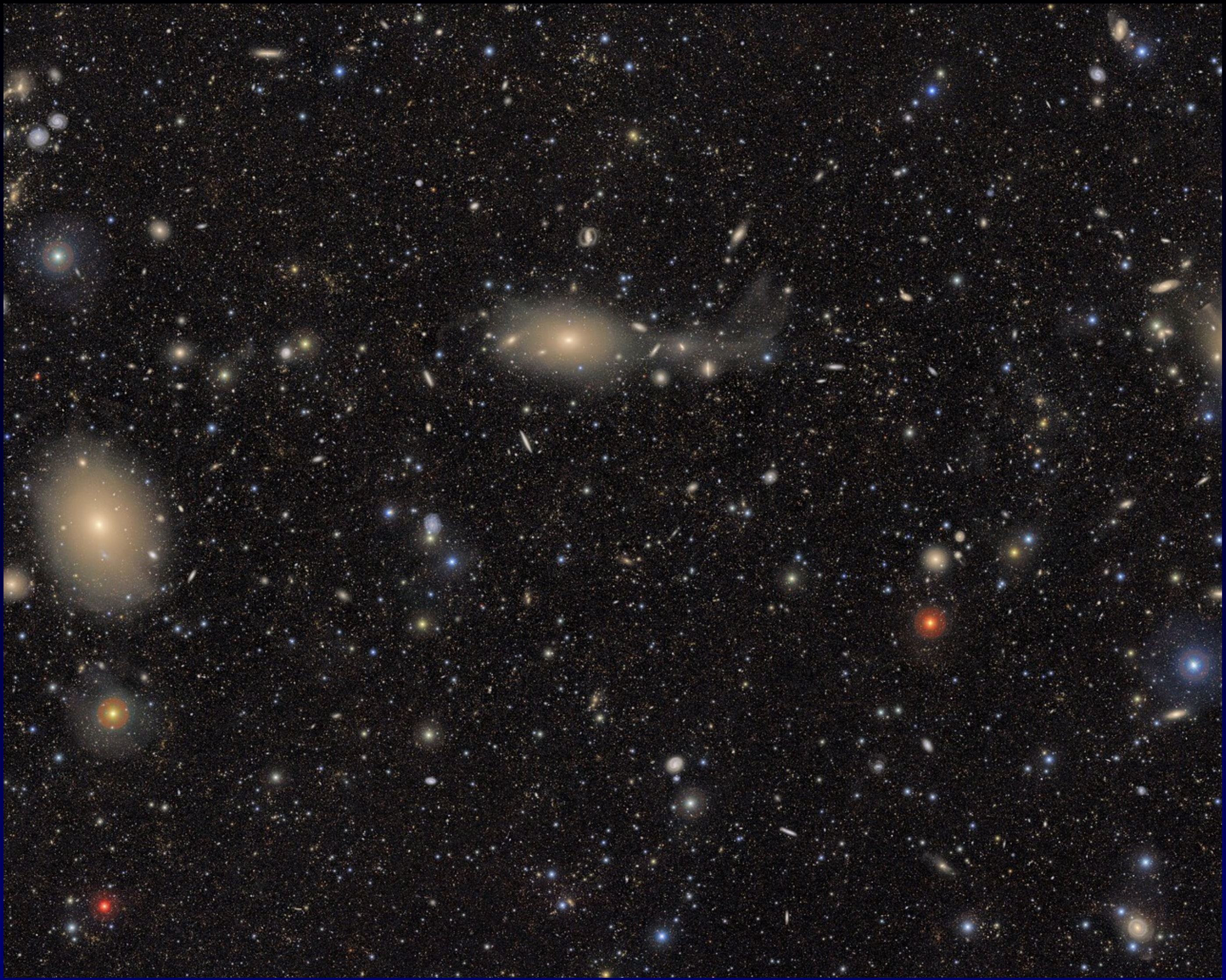
Fibre collects light from
3.3 arcsec circle on sky



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 allerts/night ! Tot 20B
38 billion objects x 800
32 tril. meas. -5 PB table



Credit: NSF–DOE Vera C. Rubin Observatory/NOIRLab/SLAC/AURA

- **Virgo cluster in Rubin**

  15 sq deg

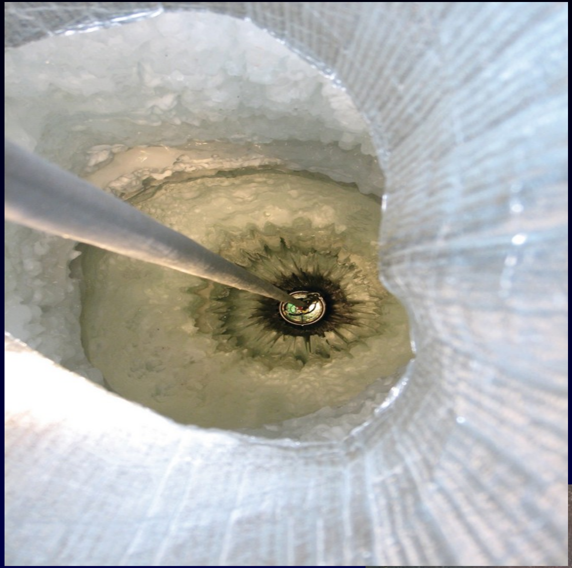
  1185 exposures / 7 nights

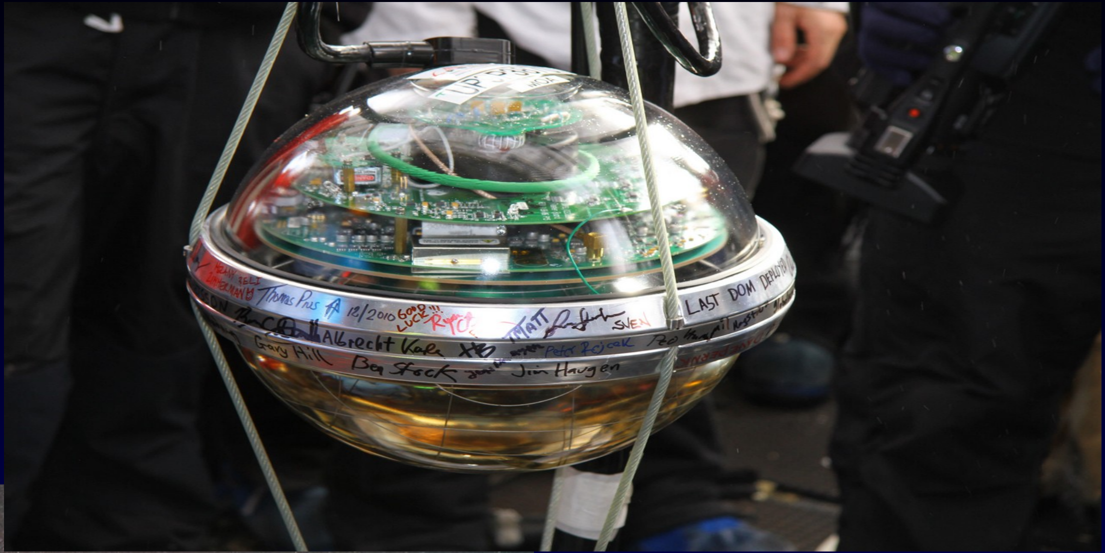
Euclid

- 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 Deep Field South J041110.98-481939.3, Credits: ESA/Euclid/Euclid Consortium/NASA,

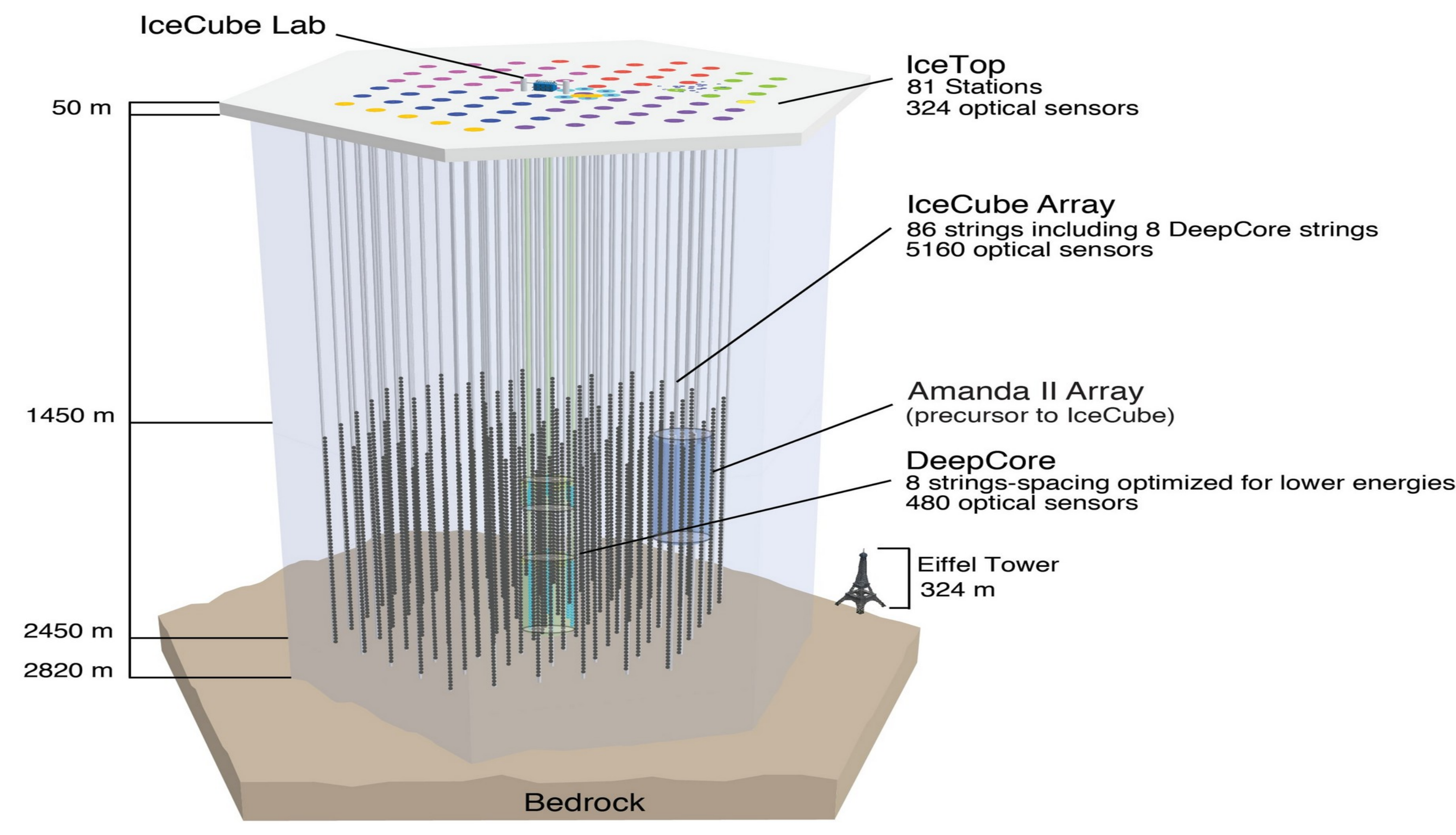
IceCube Neutrino Lab



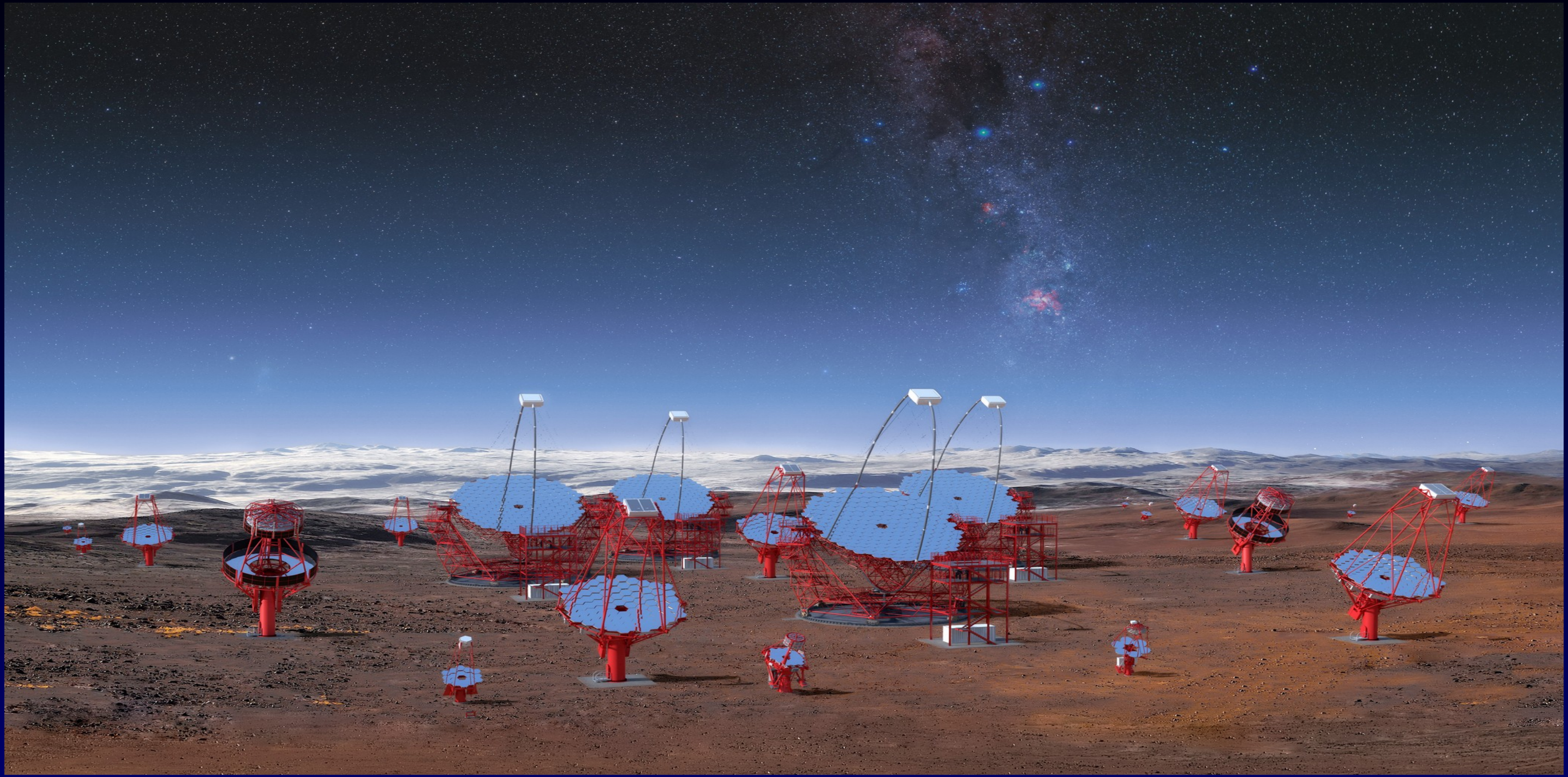


South Pole
Amundsen-Scott station

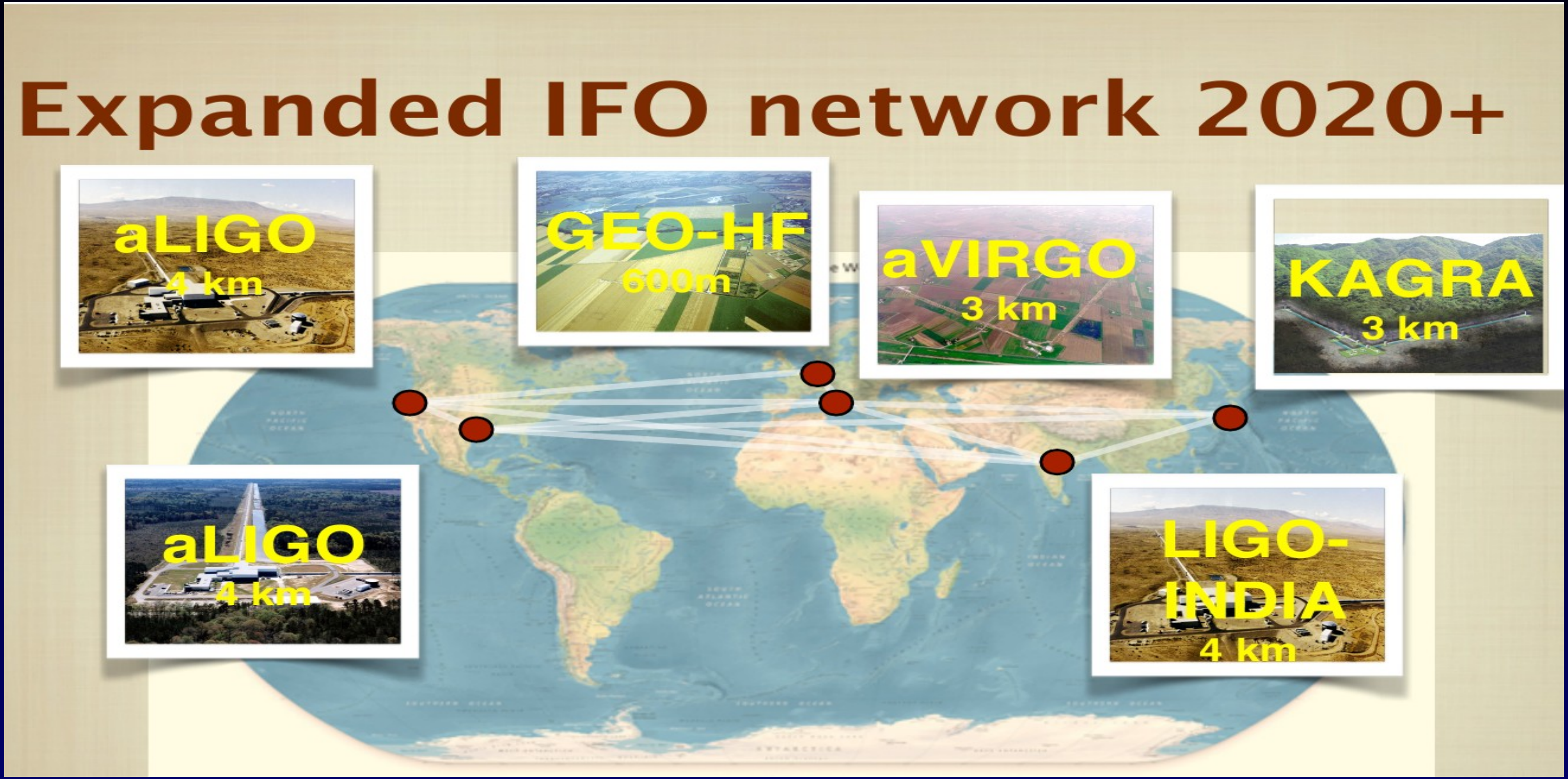
IceCube Neutrino Lab



Cherenkov Telescope Array



Gravitation Wave Detection Network

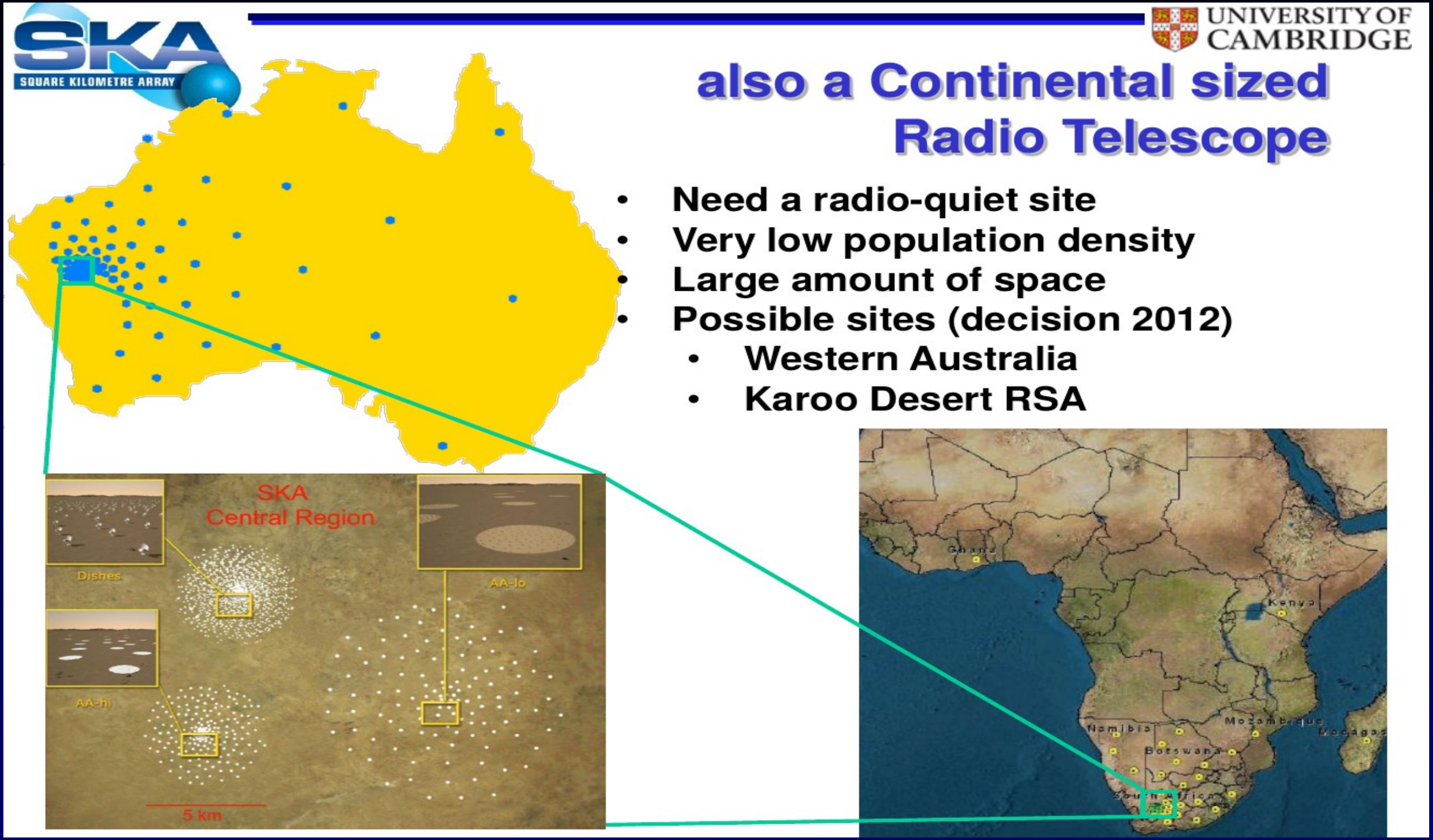


Expanded IFO network 2020+

LOFAR network

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

SKA



SKA
SQUARE KILOMETRE ARRAY

UNIVERSITY OF CAMBRIDGE

also a Continental sized Radio Telescope

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

SKA
Central Region
Dishes
AA-lo
AA-hi
5 km
Ghana
Kenya
Namibia
Botswana
Mozambique
Madagas
South Africa

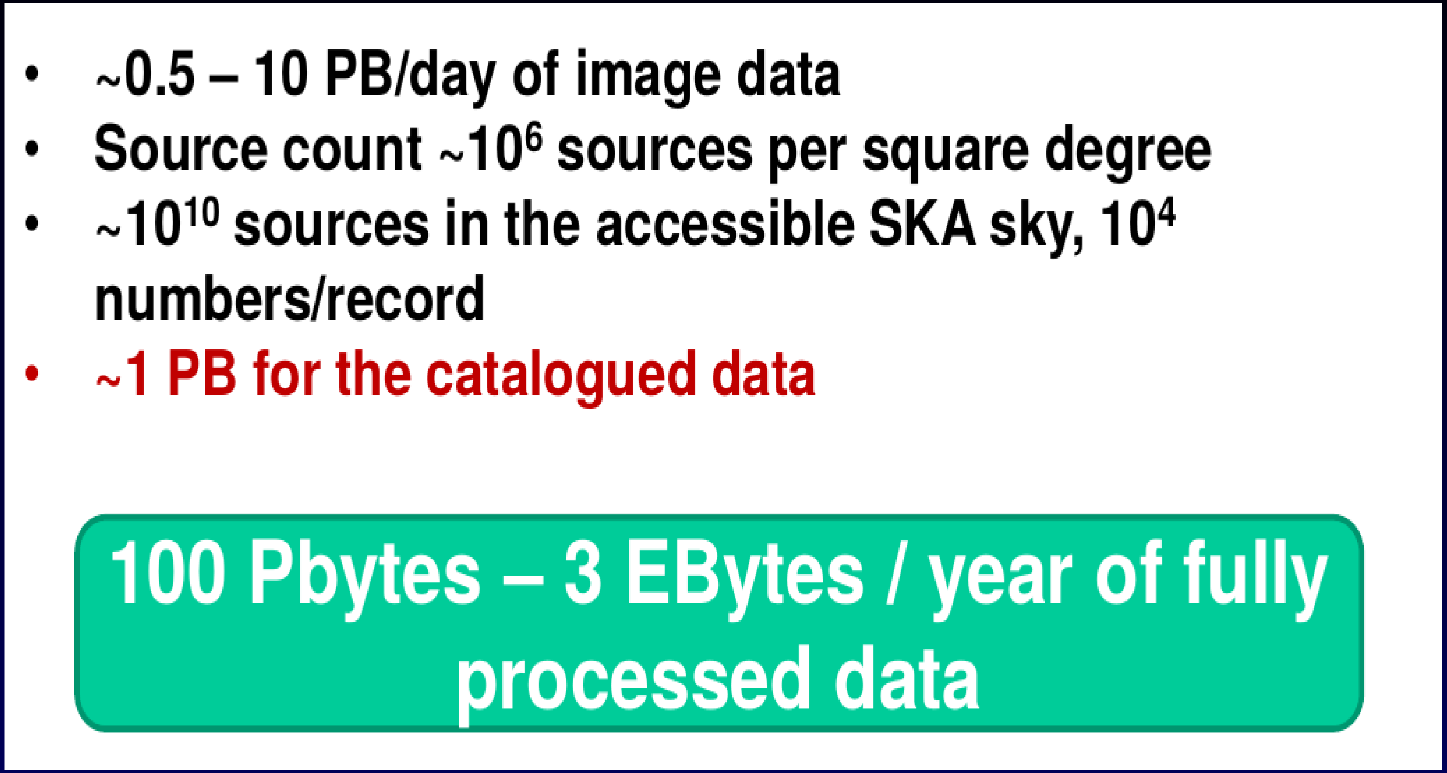
SKA

Dishes
SPDO/Swinbur

SKA

Phased Aperture array

SKA Archive Volumes

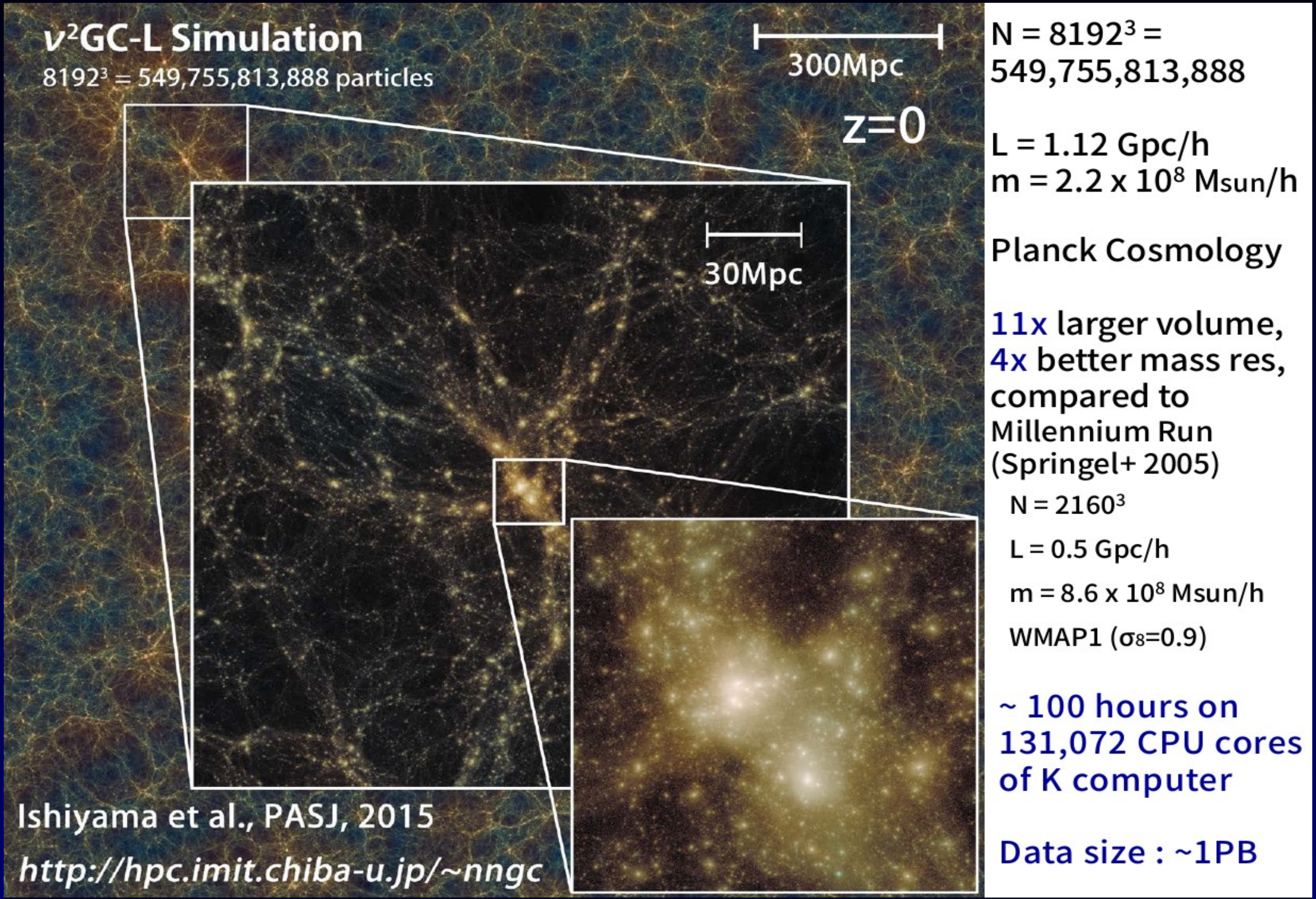


- ~0.5 – 10 PB/day of image data
- Source count ~10^6 sources per square degree
- ~10^{10} sources in the accessible SKA sky, 10^4 numbers/record
- ~1 PB for the catalogued data

100 Pbytes – 3 EBytes / year of fully processed data

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

Simulation of Universe



Ishiyama et al., PASJ, 2015

http://hpc.imit.chiba-u.jp/~nngc

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

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

Planck Cosmology

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

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

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

Data size : ~1PB

Virtual Observatory

IVOA (established 2002)

NOVA
AstroGrid
ArVO Armenian Virtual Observatory
ASVO ALL SKY VIRTUAL OBSERVATORY

BRAVO BRAZILIAN VIRTUAL OBSERVATORY
China-VO
CVO
ChiVO CHILEAN VIRTUAL OBSERVATORY

esa
EURO VO
GERMAN ASTROPHYSICAL GAVO VIRTUAL OBSERVATORY
HVO Hungarian Virtual Observatory

Japanese Virtual Observatory JVO
IVOA
KazVO

NL VO
OV FRANCE

RVO
SKAO
NRF National Research Foundation SA³ South African Astroinformatics Alliance
SVO Spanish Virtual Observatory

vObs.it
UKR VO
USVOA
vo-i

CZVO Czech Virtual Observatory

Virtual Observatory : Key Definitions

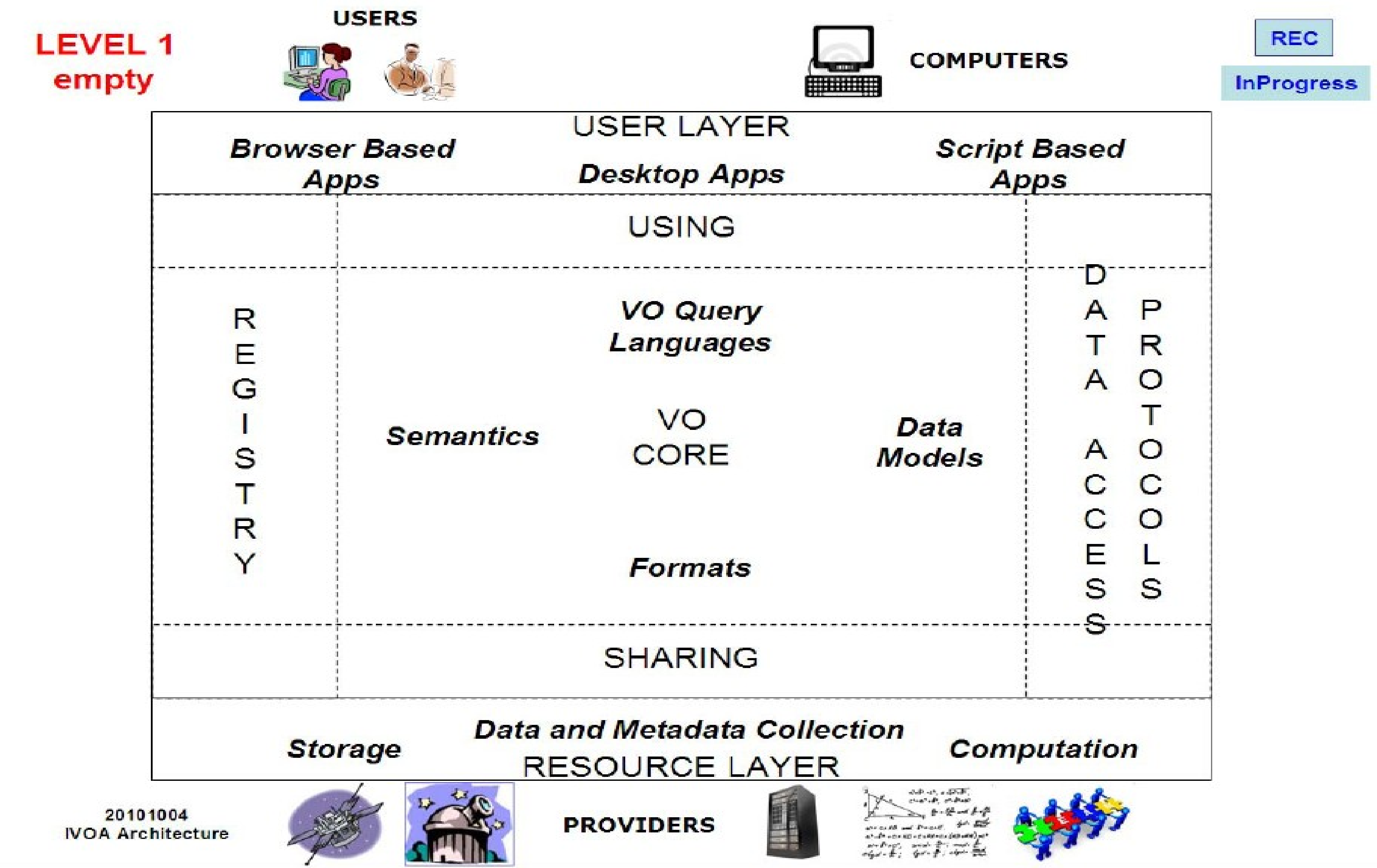
- *“The Virtual Observatory will be a system that allows astronomers to interrogate multiple data centers in a seamless and transparent way, which provides new powerful analysis and visualization tools within that system, and which gives data centers a standard framework for publishing and delivering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

- Standardization of data and metadata, and of data exchange methods.

- Registry, listing available services and what can be done with them.

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

Ecosystem of VO



LEVEL 1
empty

USERS

COMPUTERS

REC

InProgress

USER LAYER

Browser Based Apps

Desktop Apps

Script Based Apps

USING

REGISTRY

VO Query Languages

Semantics

VO CORE

Data Models

Formats

DATA ACCESS

PROTOCOLS

SHARING

Storage

Data and Metadata Collection

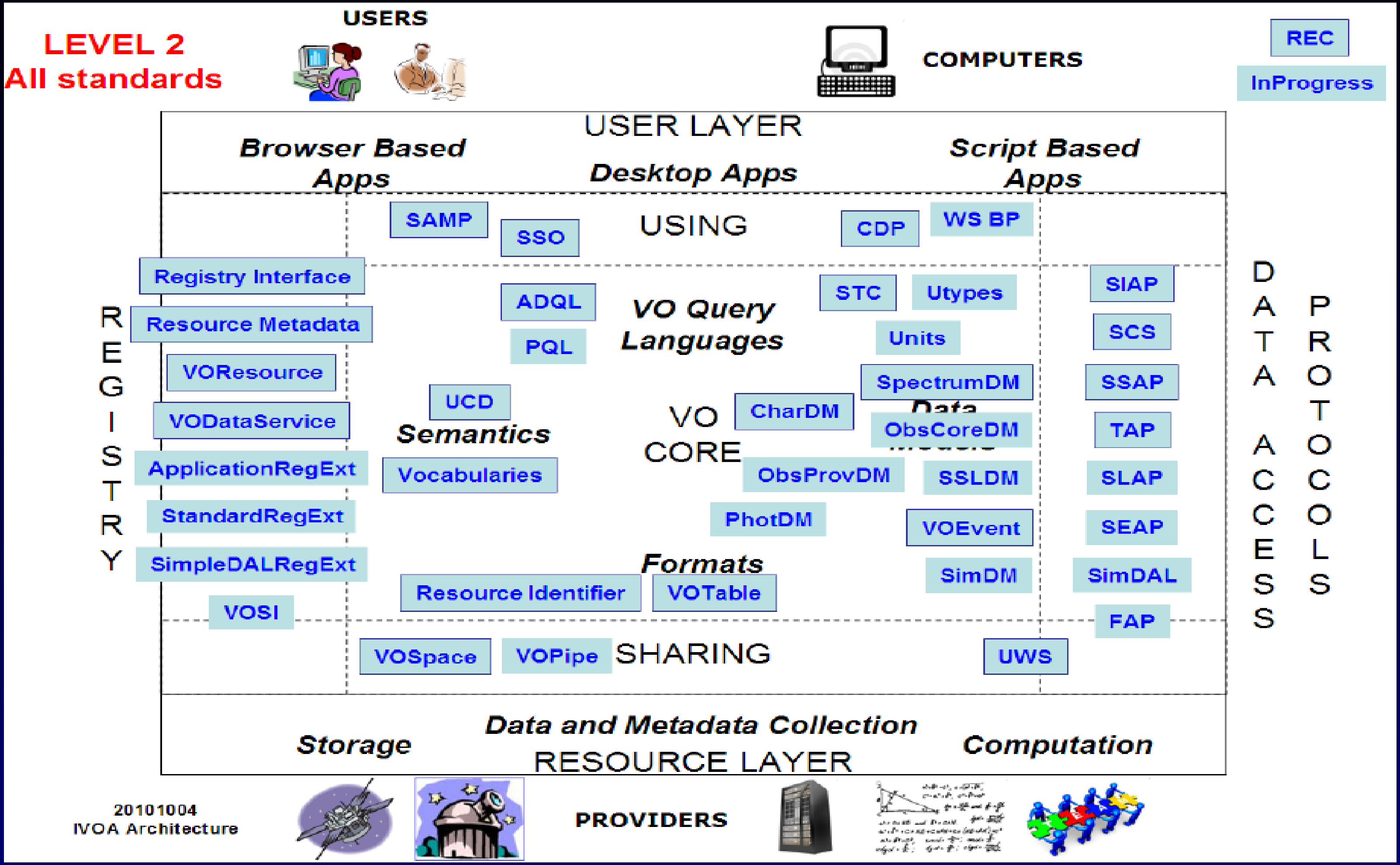
Computation

RESOURCE LAYER

20101004
IVOA Architecture

PROVIDERS

Ecosystem of VO – level 2



LEVEL 2
All standards

USERS

COMPUTERS

REC

InProgress

USER LAYER

Browser Based Apps

Desktop Apps

Script Based Apps

USING

SAMP

SSO

CDP

WS BP

REGISTRY

Registry Interface

Resource Metadata

VOResource

VODataService

ApplicationRegExt

StandardRegExt

SimpleDALRegExt

VOSI

VO Query Languages

ADQL

PQL

Semantics

UCD

Vocabularies

VO CORE

Data

STC

Utypes

Units

SpectrumDM

CharDM

ObsCoreDM

ObsProvDM

SSLDM

PhotDM

VOEvent

SimDM

Formats

Resource Identifier

VOTable

DATA ACCESS PROTOCOLS

SIAP

SCS

SSAP

TAP

SLAP

SEAP

SimDAL

FAP

SHARING

VOSpace

VOPipe

UWS

Storage

Data and Metadata Collection

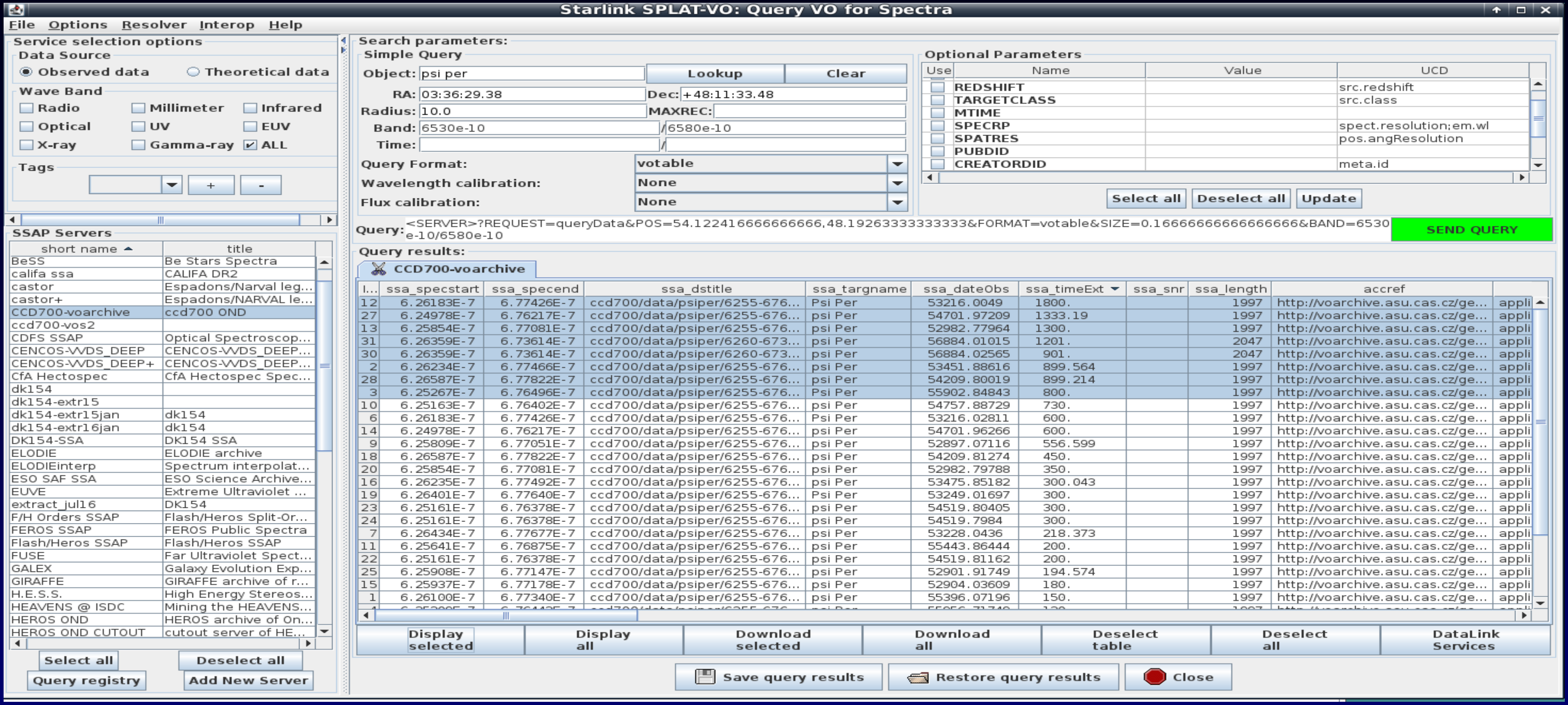
Computation

RESOURCE LAYER

20101004
IVOA Architecture

PROVIDERS

Spectra in SPLAT-VO - query



Data Source

● Observed data ○ Theoretical data

Wave Band

☐ Radio ☐ Millimeter ☐ Infrared
☐ Optical ☐ UV ☐ EUV
☐ X-ray ☐ Gamma-ray ☑ ALL

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

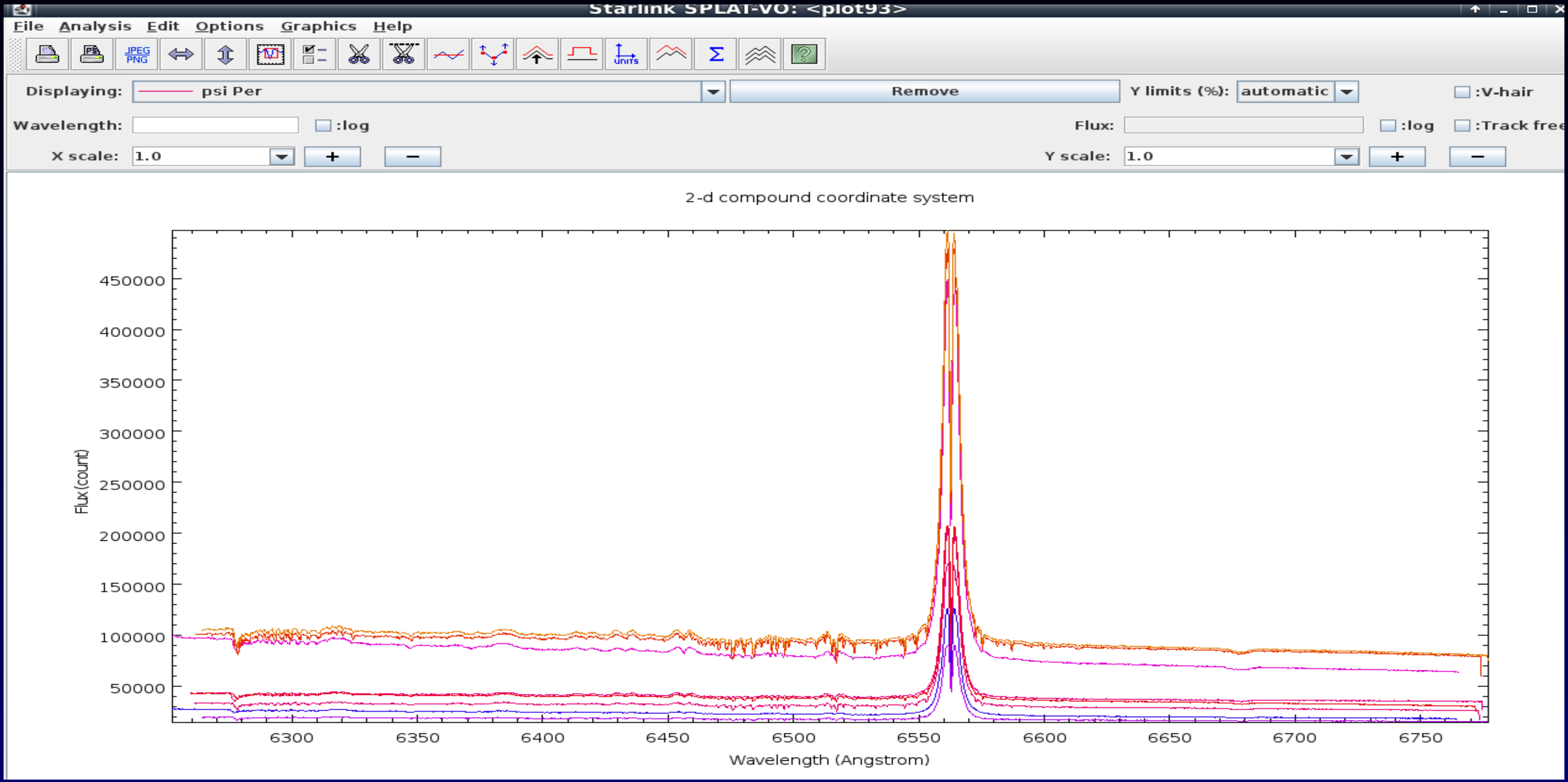
Use	Name	Value	UCD
☐	REDSHIFT		src.redshift
☐	TARGETCLASS		src.class
☐	MTIME		
☐	SPECRP		spect.resolution;em.wl
☐	SPATRES		pos.angResolution
☐	PUBDID		
☐	CREATORDID		meta.id

Query: <SERVER>?REQUEST=queryData&POS=54.122416666666666,48.192633333333333&FORMAT=votable&SIZE=0.16666666666666666&BAND=6530e-10/6580e-10

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-VVDS_DEEP	CENCOS-VVDS_DEEP...
CENCOS-VVDS_DEEP+	CENCOS-VVDS_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...

I...	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

Spectra in SPLAT-VO direct access



Displaying: psi Per

Y limits (%): automatic

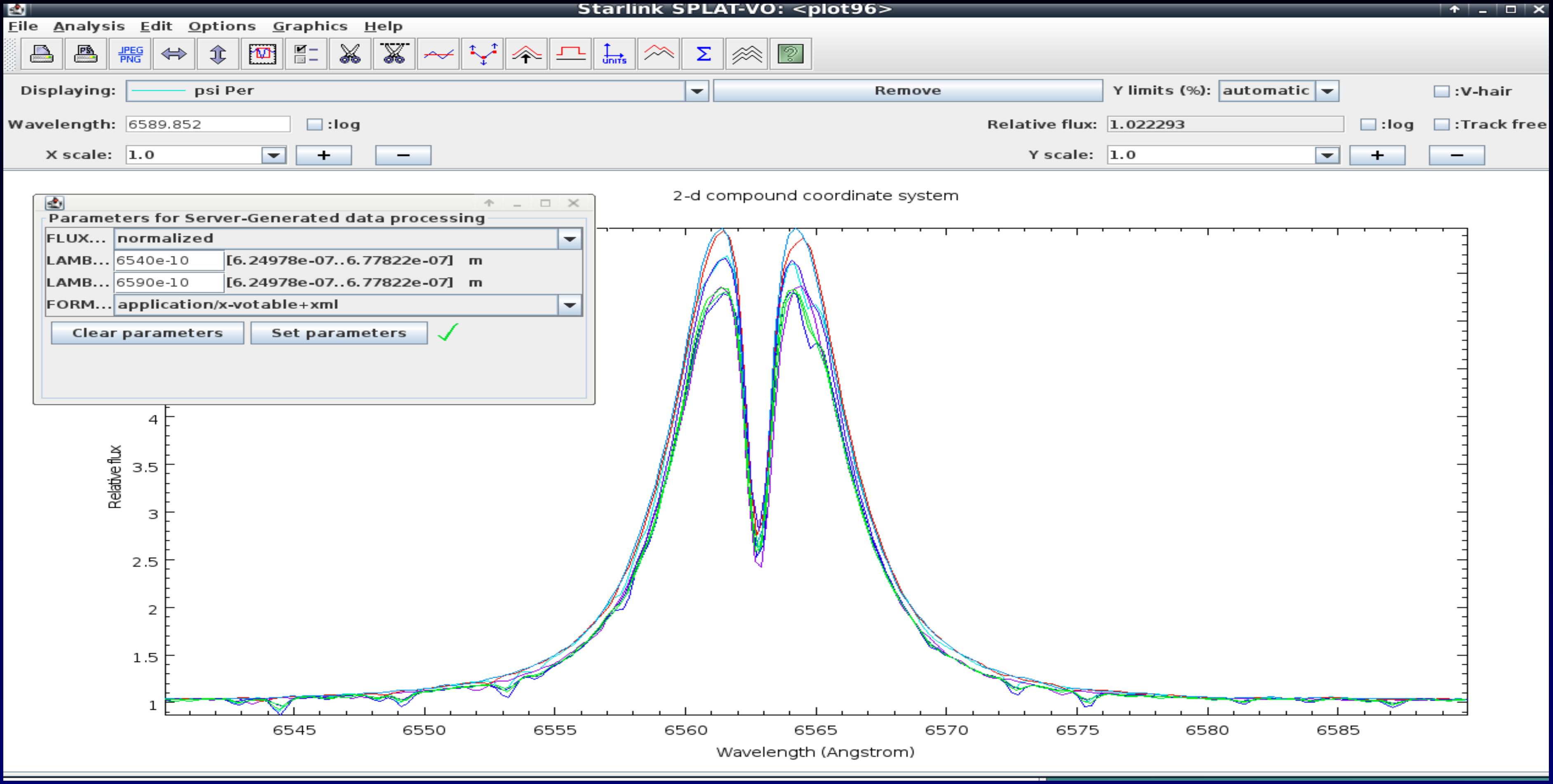
Wavelength:

Flux:

X scale: 1.0

Y scale: 1.0

Spectra in SPLAT-VO - DataLink



Displaying: psi Per

Y limits (%): automatic

☐ :V-hair

Wavelength: 6589.852 ☐ :log

Relative flux: 1.022293 ☐ :log ☐ :Track free

X scale: 1.0

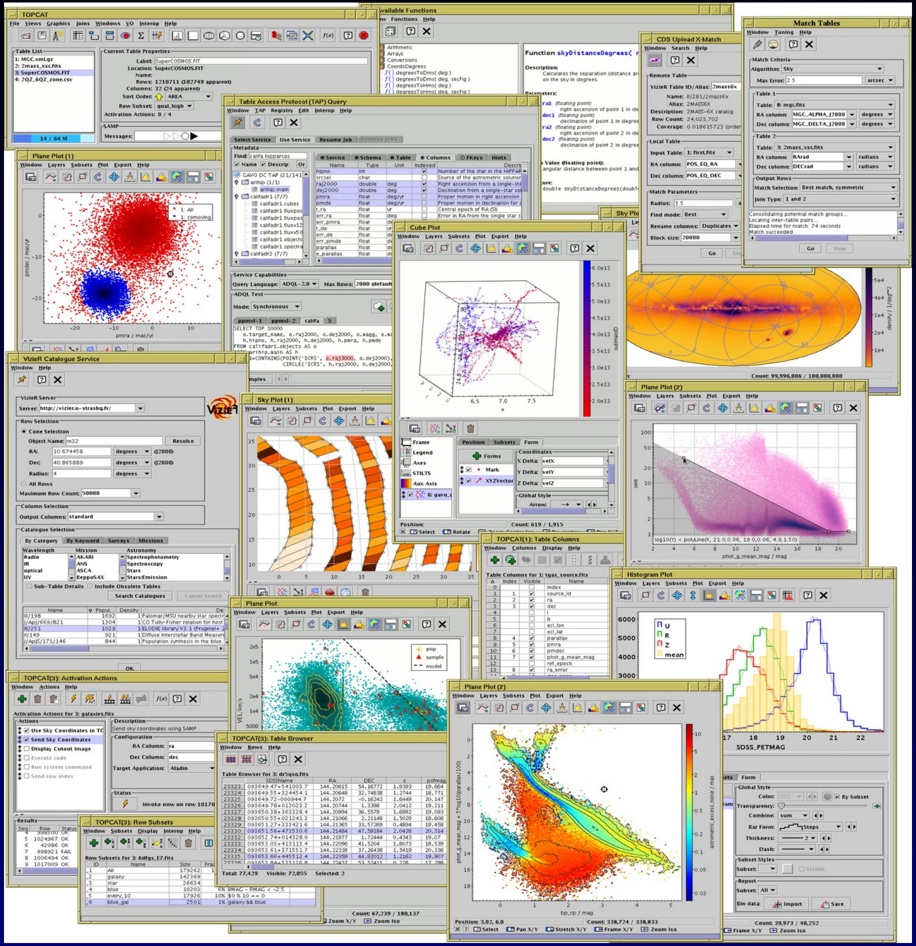
Y scale: 1.0

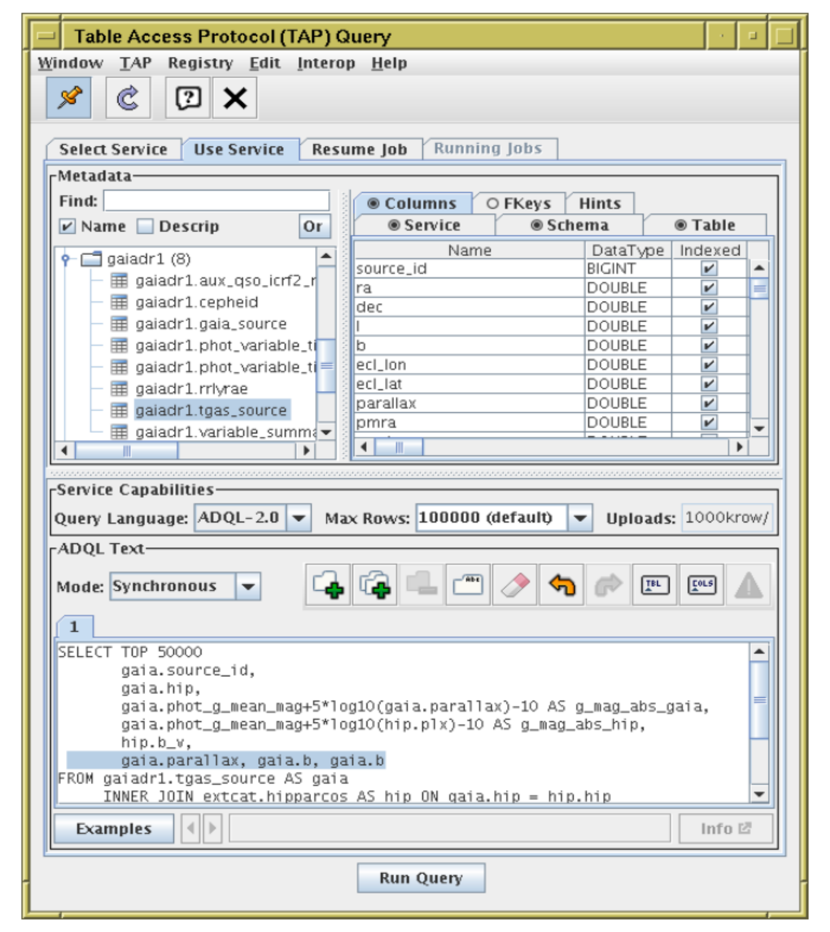
Parameters for Server-Generated data processing

FLUX...	normalized		
LAMB...	6540e-10	[6.24978e-07..6.77822e-07]	m
LAMB...	6590e-10	[6.24978e-07..6.77822e-07]	m
FORM...	application/x-votable+xml		

✓

TOPCAT





gaiadr1 (8)
- gaiadr1.aux_qso_icrf2_r
- gaiadr1.cepheid
- gaiadr1.gaia_source
- gaiadr1.phot_variable_ti
- gaiadr1.phot_variable_ti
- gaiadr1.rrlyrae
- gaiadr1.tgas_source
- gaiadr1.variable_summa

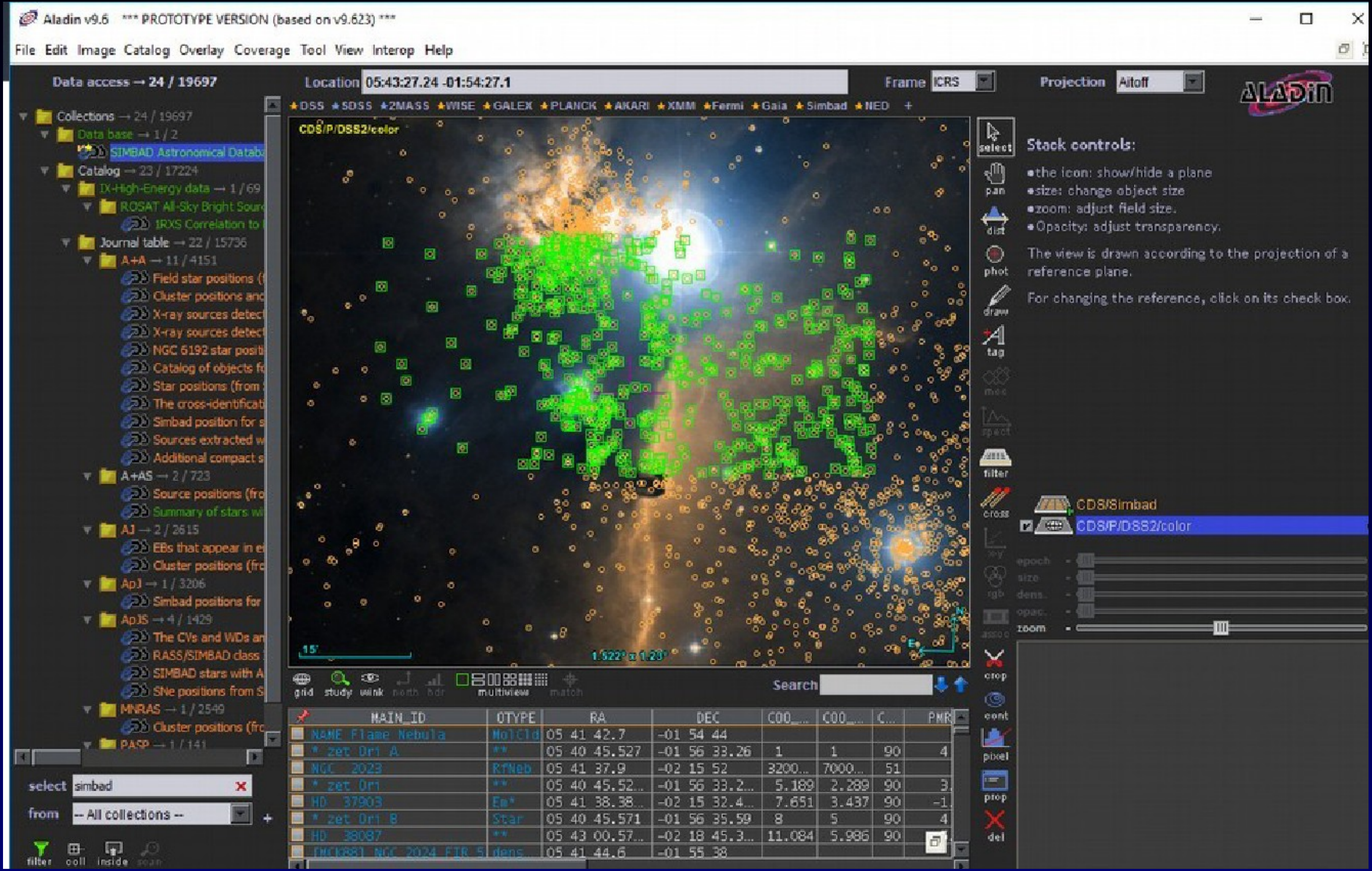
Name	DataType	Indexed
source_id	BIGINT	☑
ra	DOUBLE	☑
dec	DOUBLE	☑
l	DOUBLE	☑
b	DOUBLE	☑
ecl_lon	DOUBLE	☑
ecl_lat	DOUBLE	☑
parallax	DOUBLE	☑
pmra	DOUBLE	☑

Query Language: ADQL-2.0
Max Rows: 100000 (default)
Uploads: 1000krow/

Mode: Synchronous

```
SELECT TOP 50000
       gaia.source_id,
       gaia.hip,
       gaia.phot_g_mean_mag+5*log10(gaia.parallax)-10 AS g_mag_abs_gaia,
       gaia.phot_g_mean_mag+5*log10(hip.plx)-10 AS g_mag_abs_hip,
       hip.b_v,
       gaia.parallax, gaia.b, gaia.b
FROM gaiadr1.tgas_source AS gaia
    INNER JOIN extcat.hipparcos AS hip ON gaia.hip = hip.hip
```

Aladin



Location 05:43:27.24 -01:54:27.1

Frame ICRS

Projection Aitoff

select simbad

from -- All collections --

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.

MAIN_ID	OTYPE	RA	DEC	COO_...	COO_...	C...	PMR
NAME Flame Nebula	MolCld	05 41 42.7	-01 54 44				
* zet Ori 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 Ori	**	05 40 45.52...	-01 56 33.2...	5.189	2.289	90	3.
HD 37903	Em*	05 41 38.38...	-02 15 32.4...	7.651	3.437	90	-1.
* zet Ori 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	
[MCK88] NGC 2024 FIR 5	dens..	05 41 44.6	-01 55 38				

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 interoperabilty – SAMP

Allows develop applications as bricks

sending VOTABLES (catalogue-spectra-images)

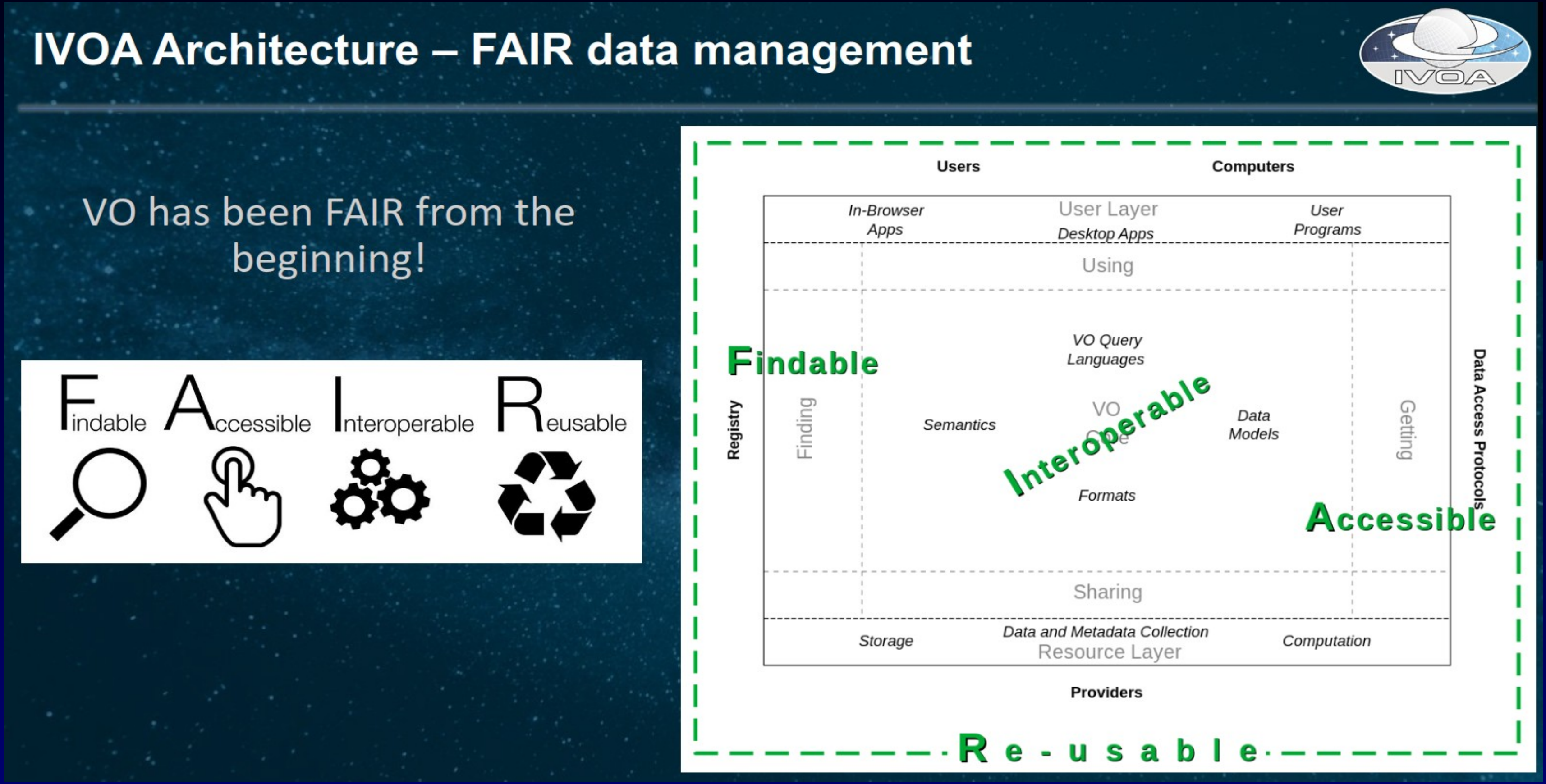
Surveys visualization

HIPS (Hiearchical Progressive Survey) - allsky zoom

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

Open Science – EOSC

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

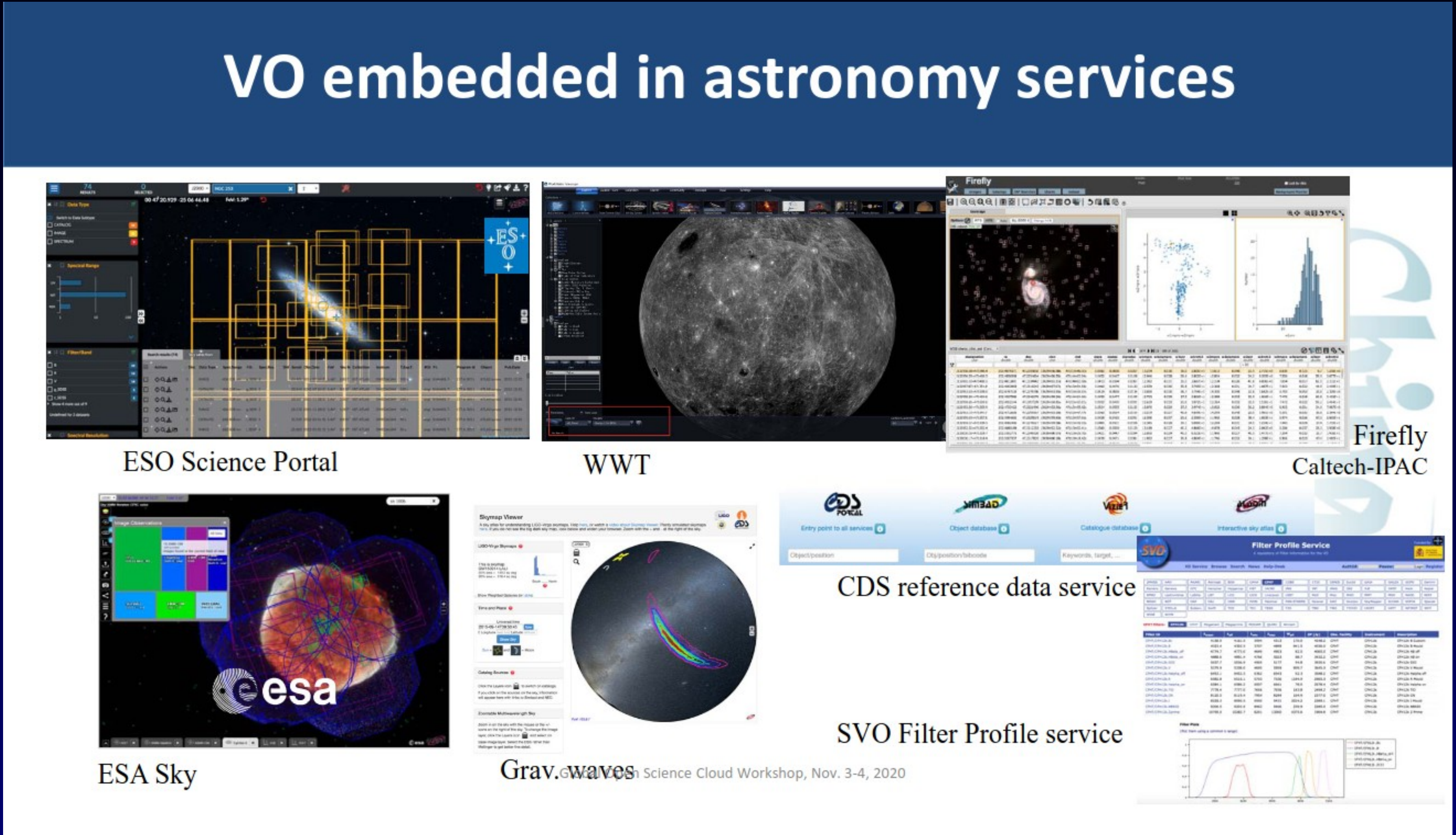


IVOA Architecture – FAIR data management

VO has been FAIR from the beginning!

Arviset 2023

VO Science Portals



VO embedded in astronomy services

ESO Science Portal

WWT

Firefly
Caltech-IPAC

ESA Sky

Grav. waves

CDS reference data service

SVO Filter Profile service

Global Open Science Cloud Workshop, Nov. 3-4, 2020

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/

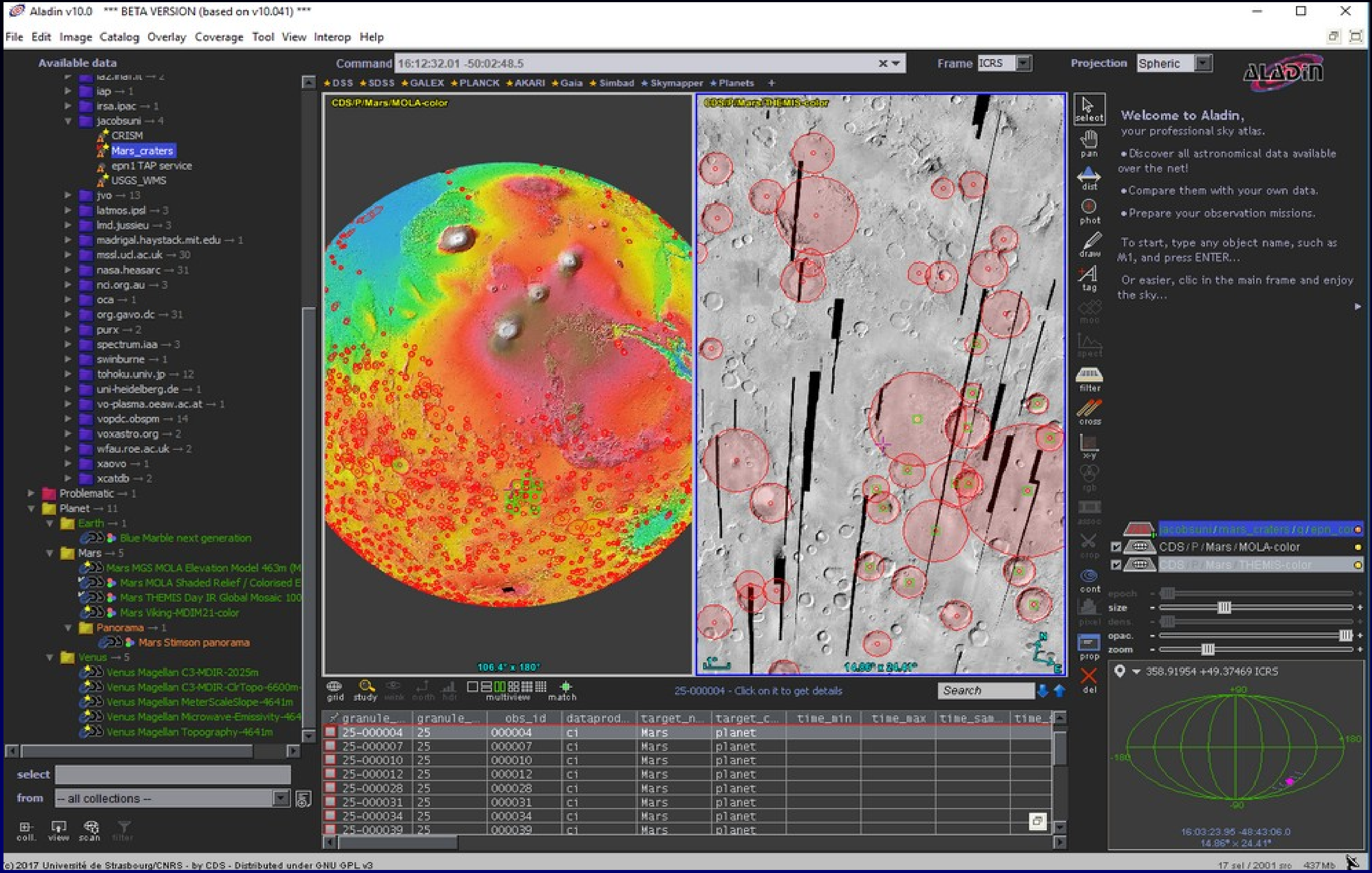
WWT (original MS, now AAS, web client)

http://worldwidetelescope.org/webclient

GoogleSky

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

EUROPLANET VESPA (EPN-TAP)



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

Frame: ICRS

Projection: Spheric

select:

from: -- all collections --

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...

25-000004 - Click on it to get details

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

358.91954 +49.37469 ICRS

16:03:23.95 -48:43:06.0
14.86° x 24.41°

(c) 2017 Université de Strasbourg/CNRS - by CDS - Distributed under GNU GPL v3

17 sel / 2001 src 437 Mb

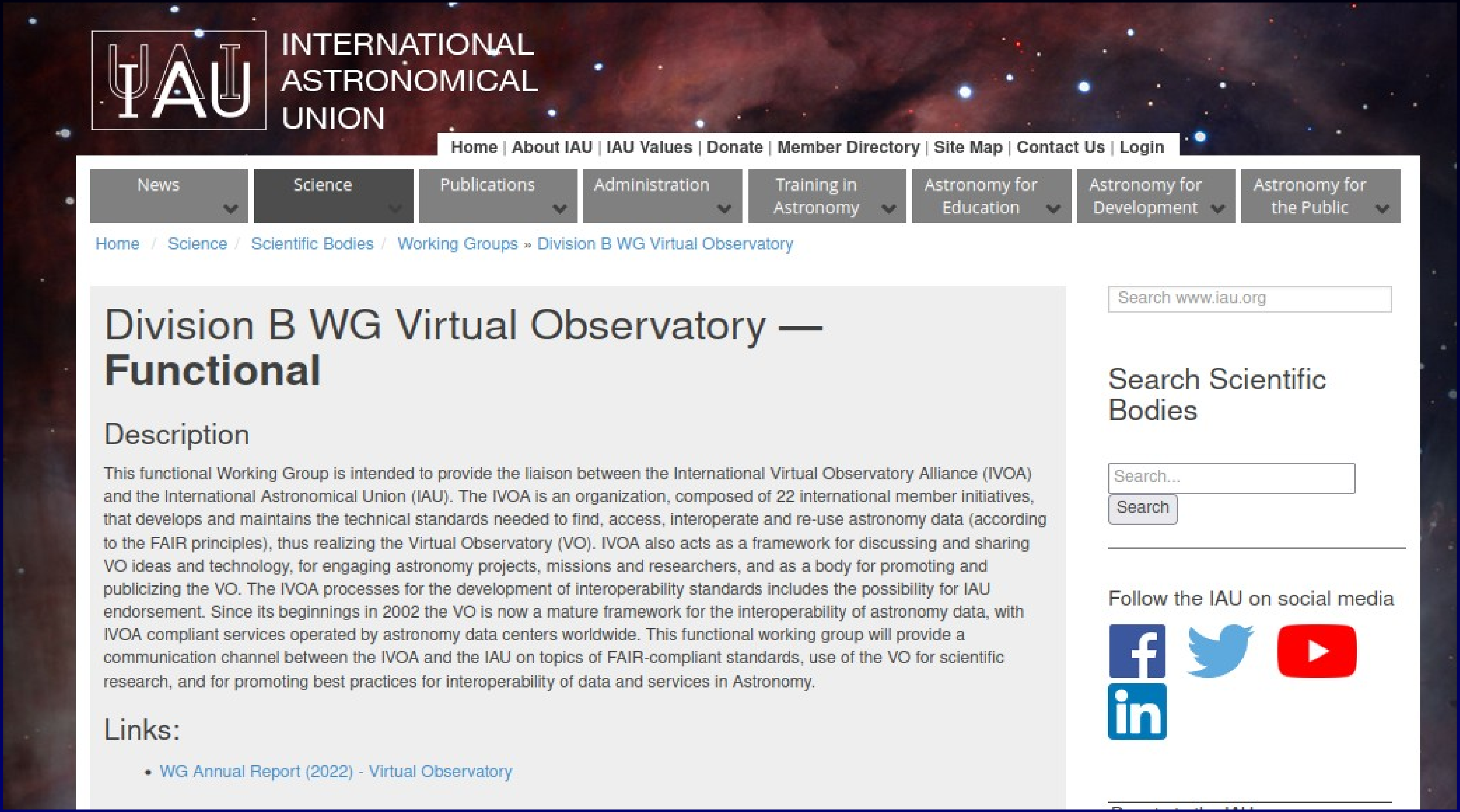
Big Data handling

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*)

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
Search www.iau.org
Search Scientific Bodies
Search...
Search
Follow the IAU on social media

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

Tutorials of VO

www.ivoa.net   portal

https://hendhd.github.io/ivoa_newcomers/

https://www.canfar.net/storage/vault/list/IVOA/virtual2021a (video)

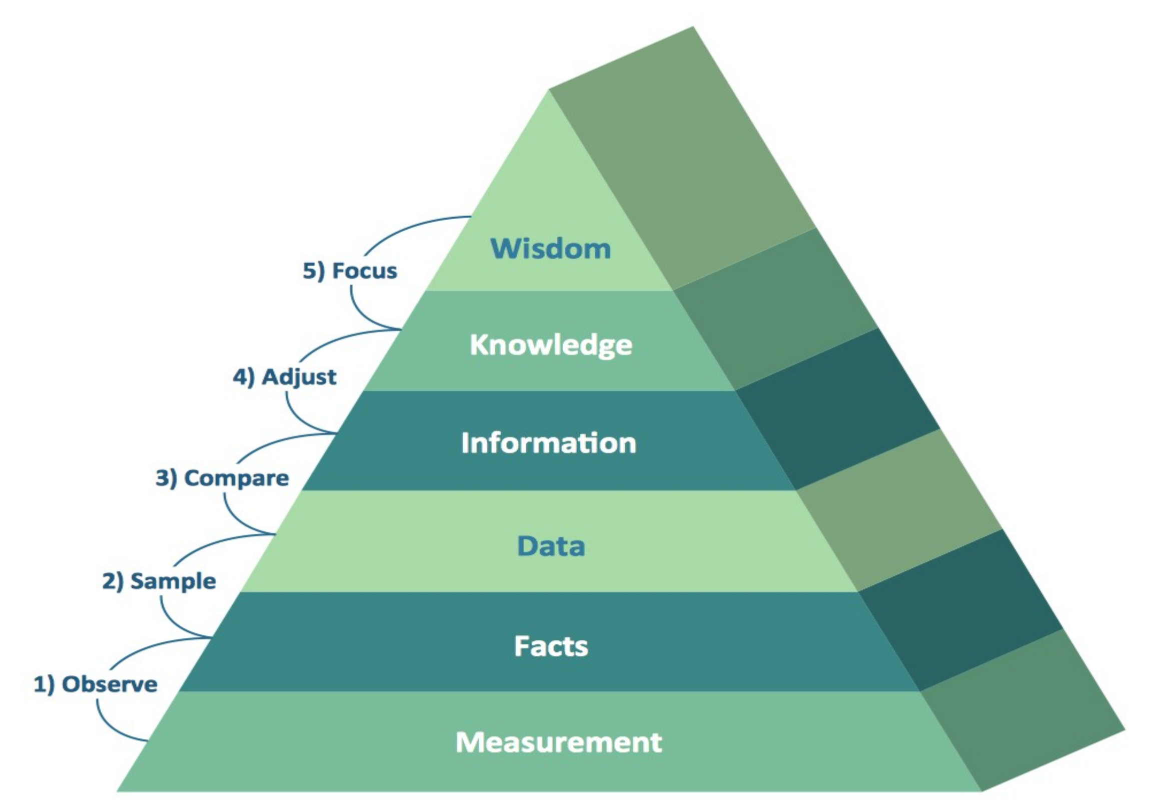
IVOA Interoperability meetings (May + November)
Newcomers Intro

Number of VO Schools

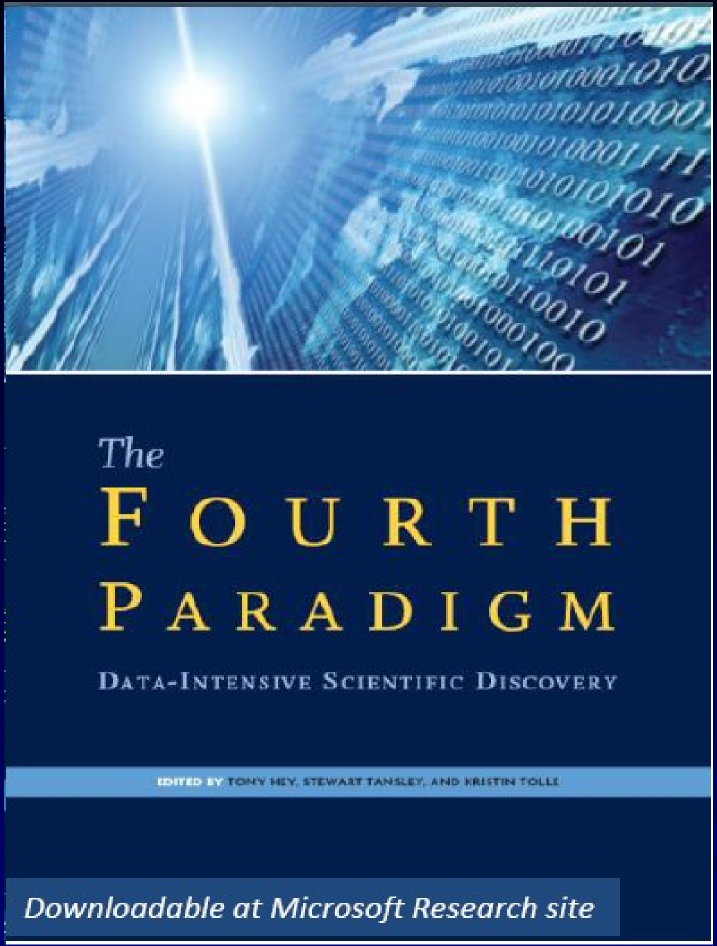
EURO-VO DCA, AIDA, ICE
CoSADIE, ASTERICS, ESCAPE

Astroinformatics

Data-Knowledge-Wisdom Pyramid



X-informatics



The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

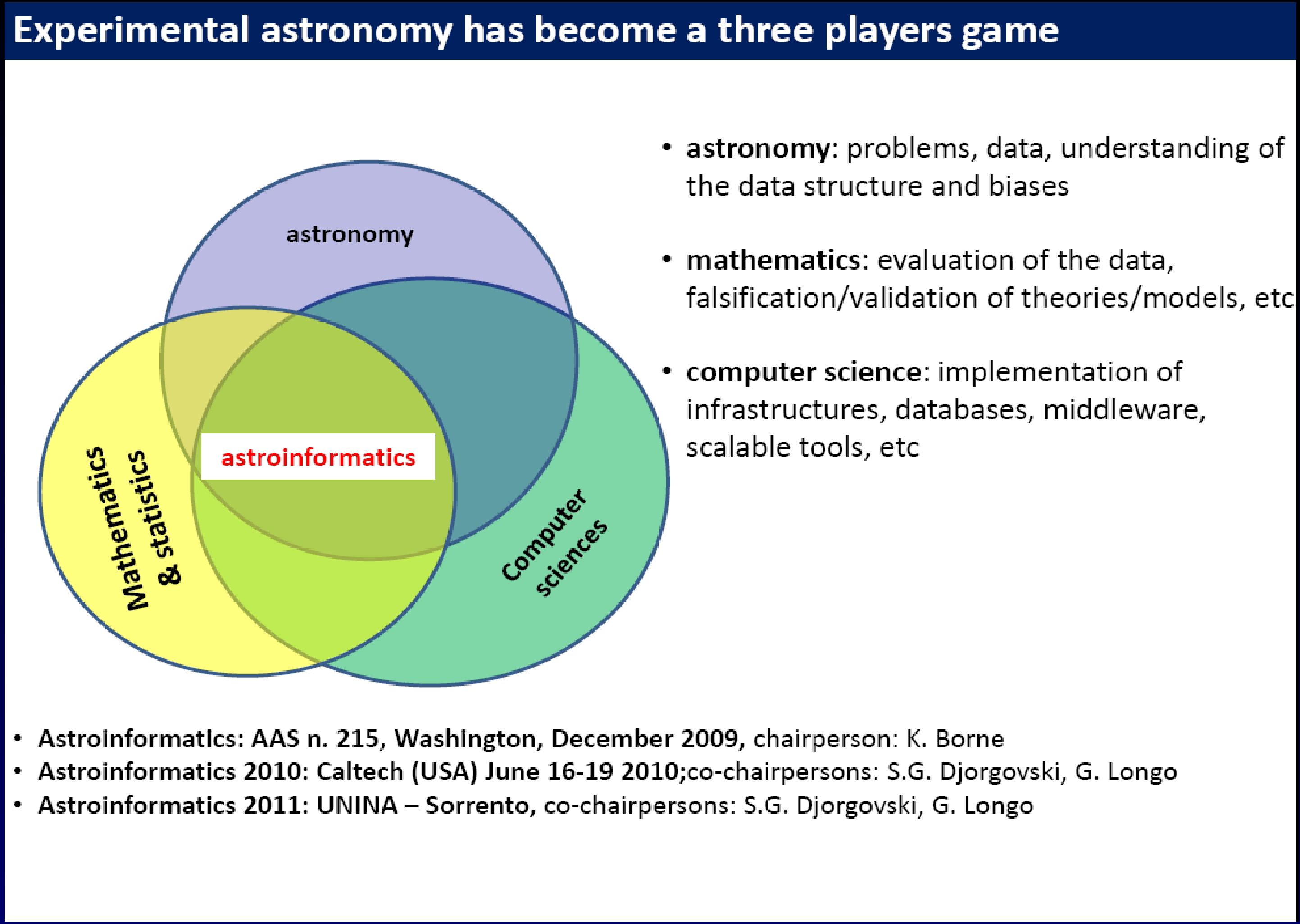
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Downloadable at Microsoft Research site

Changing methodology of the Science

Synergy between different worlds

Sociological aspects (net-based research communities)



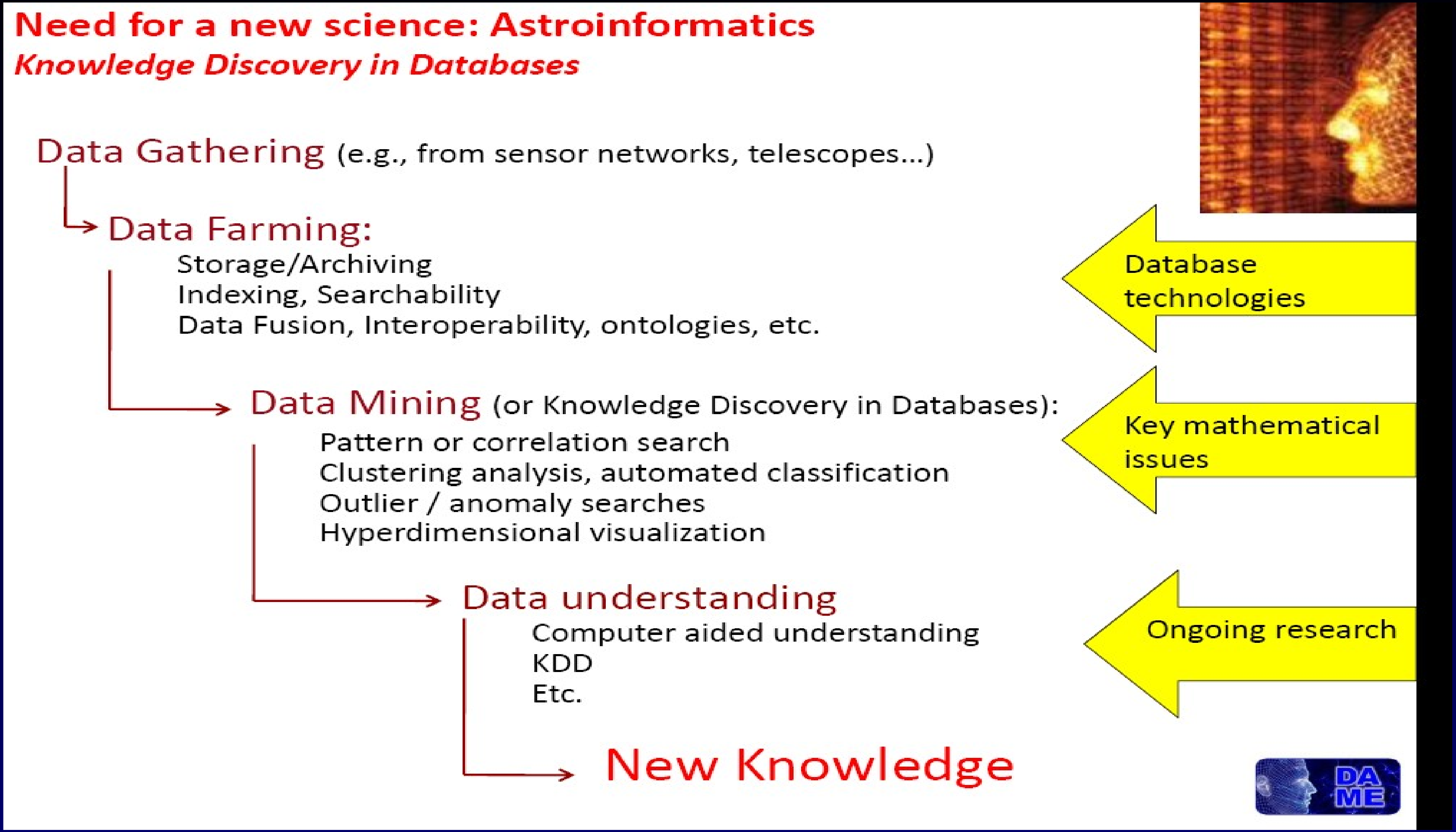
Experimental astronomy has become a three players game

astronomy
astroinformatics
Mathematics & statistics
Computer sciences

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

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

Longo 2010



Need for a new science: Astroinformatics

Knowledge Discovery in Databases

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

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

Data Mining (or Knowledge Discovery in Databases):
- Pattern or correlation search
- Clustering analysis, automated classification
- Outlier / anomaly searches
- Hyperdimensional visualization

Data understanding
- Computer aided understanding
- KDD
- Etc.

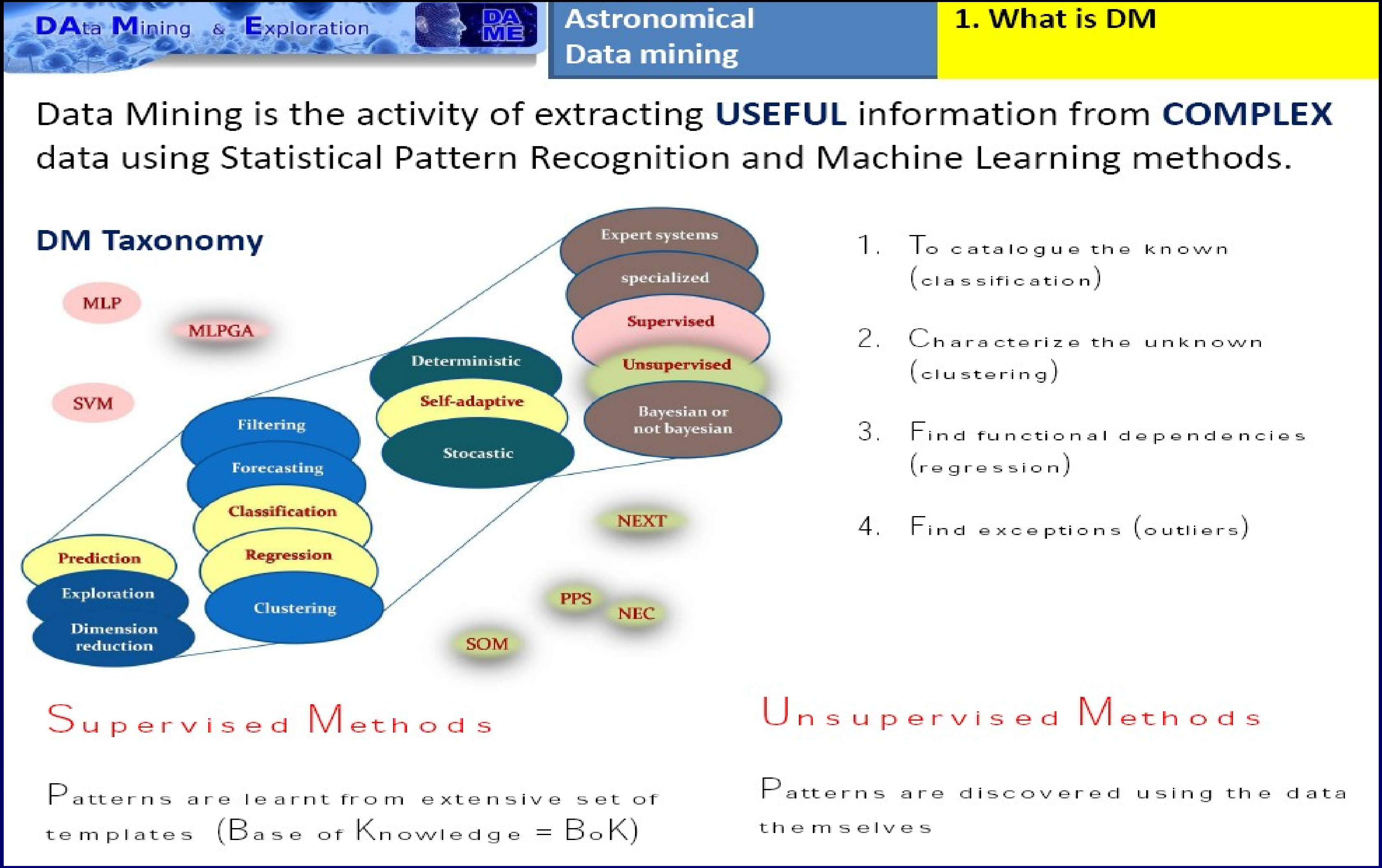
New Knowledge

Database technologies

Key mathematical issues

Ongoing research

Longo 2009



DAta Mining & Exploration DAME

Astronomical Data mining

1. What is DM

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

DM Taxonomy

MLP
MLPGA
SVM
Prediction
Exploration
Dimension reduction
Filtering
Forecasting
Classification
Regression
Clustering
Deterministic
Self-adaptive
Stocastic
Expert systems
specialized
Supervised
Unsupervised
Bayesian or not bayesian
NEXT
PPS
NEC
SOM

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

Supervised Methods

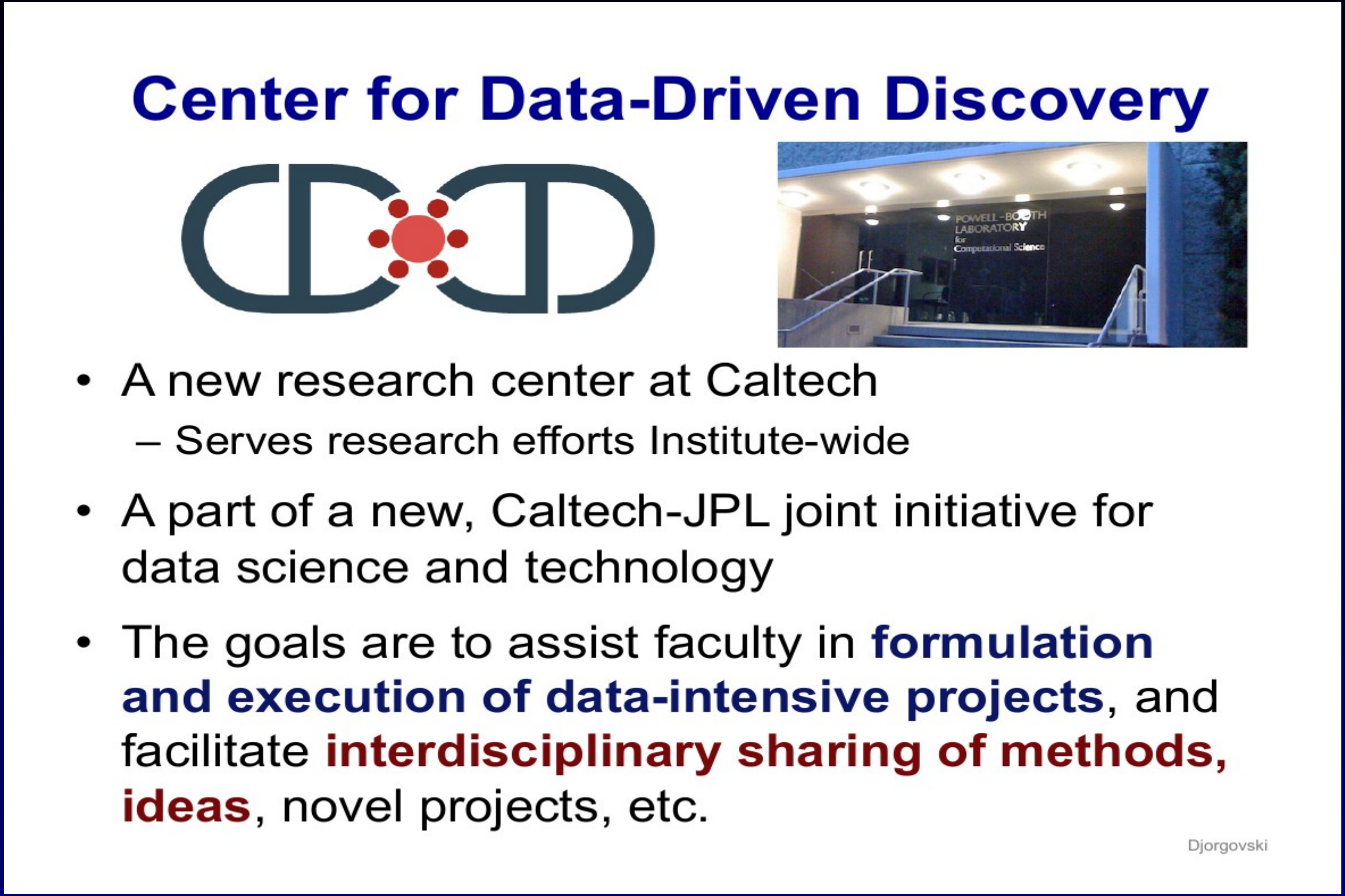
Patterns are learnt from extensive set of templates (Base of Knowledge = BoK)

Unsupervised Methods

Patterns are discovered using the data themselves

Longo 2009

New e-Science Collaborations

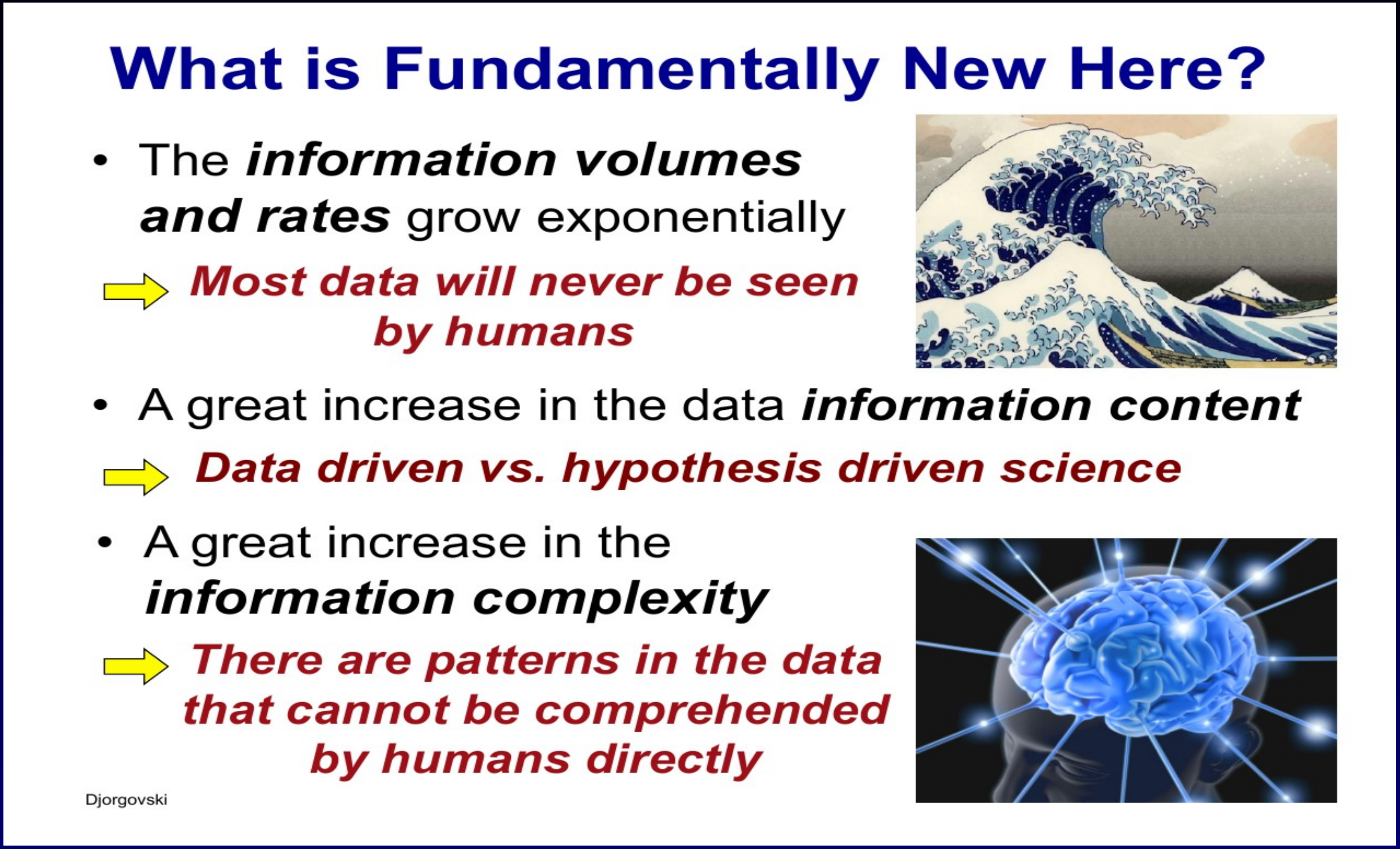

POWELL-BOOTH
LABORATORY
for
Computational Science

Center for Data-Driven Discovery

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

Djorgovski

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***

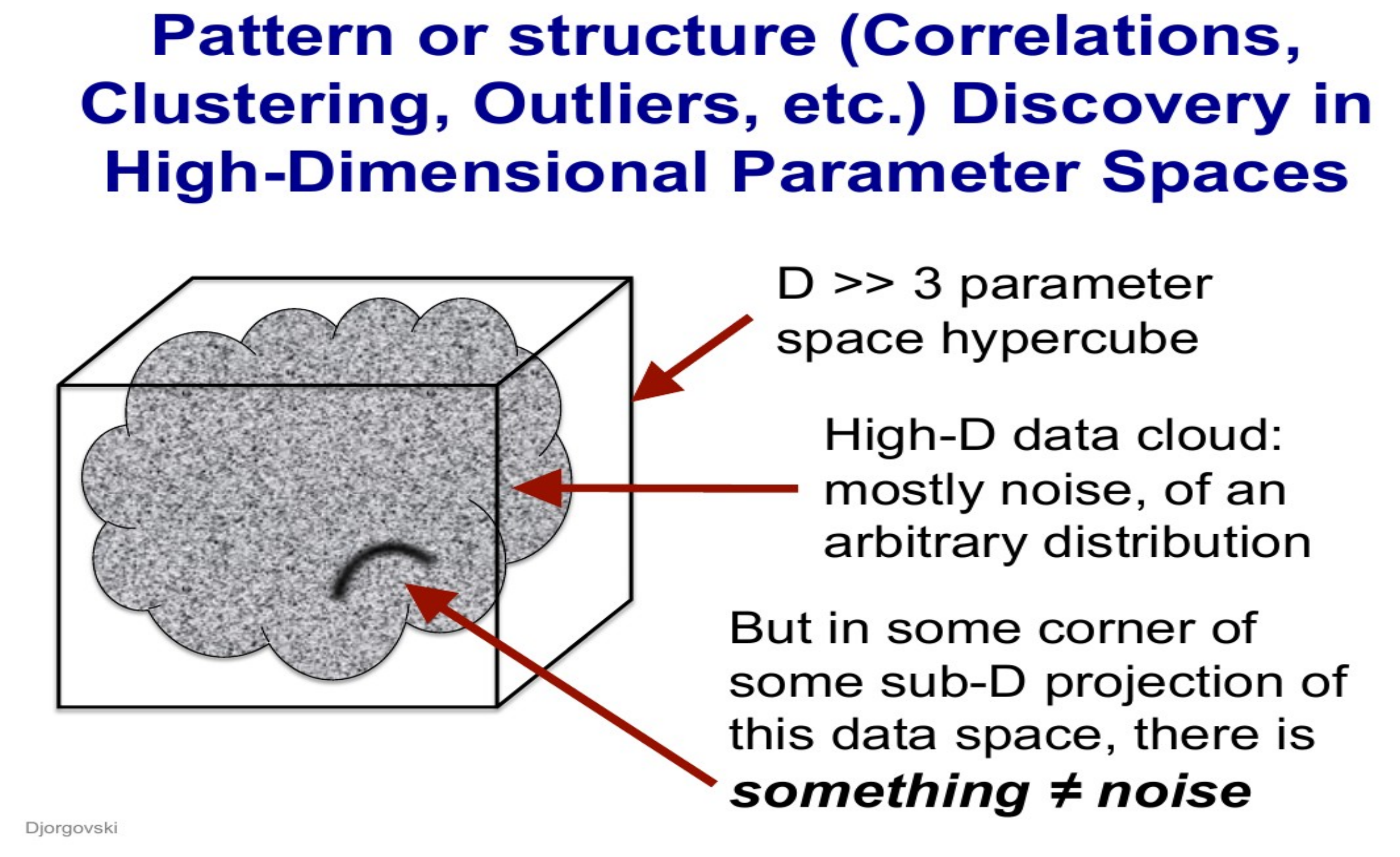
⇨ ***Data driven vs. hypothesis driven science***

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

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

Djorgovski

Hidden Patterns in Data



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

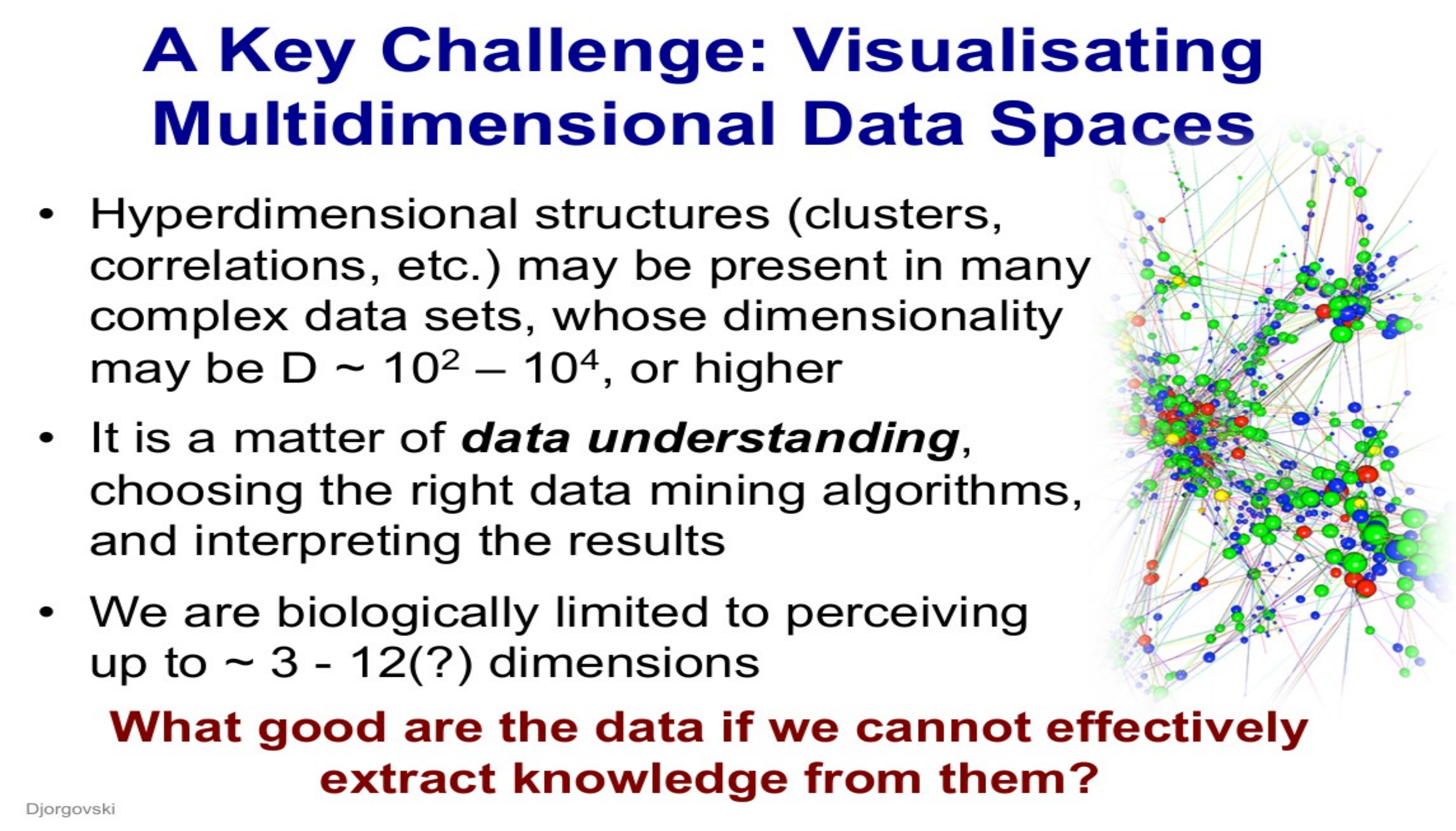
D >> 3 parameter space hypercube

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

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

Djorgovski

Visualization in Machine Learning



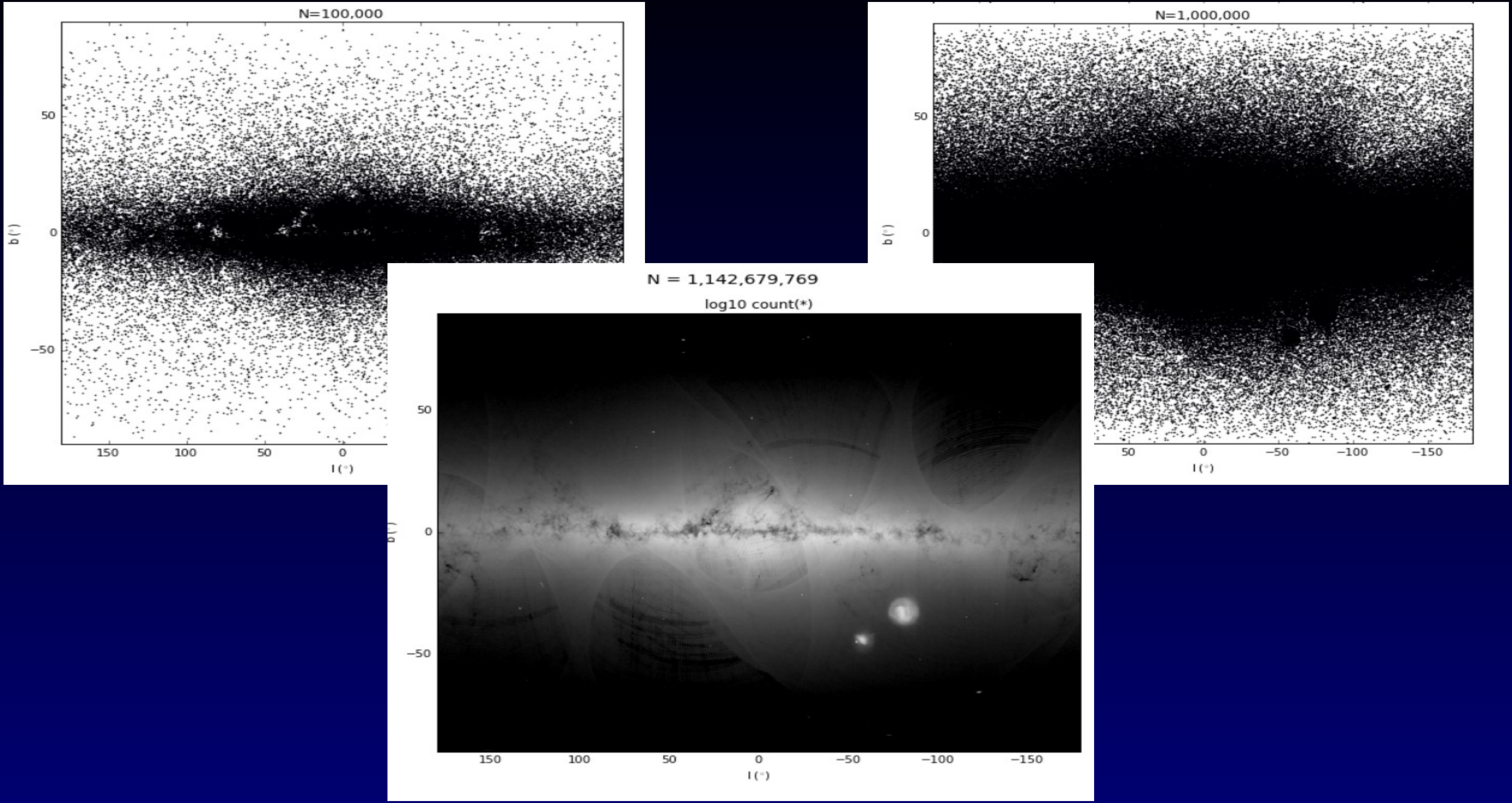
A Key Challenge: Visualisating Multidimensional Data Spaces

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

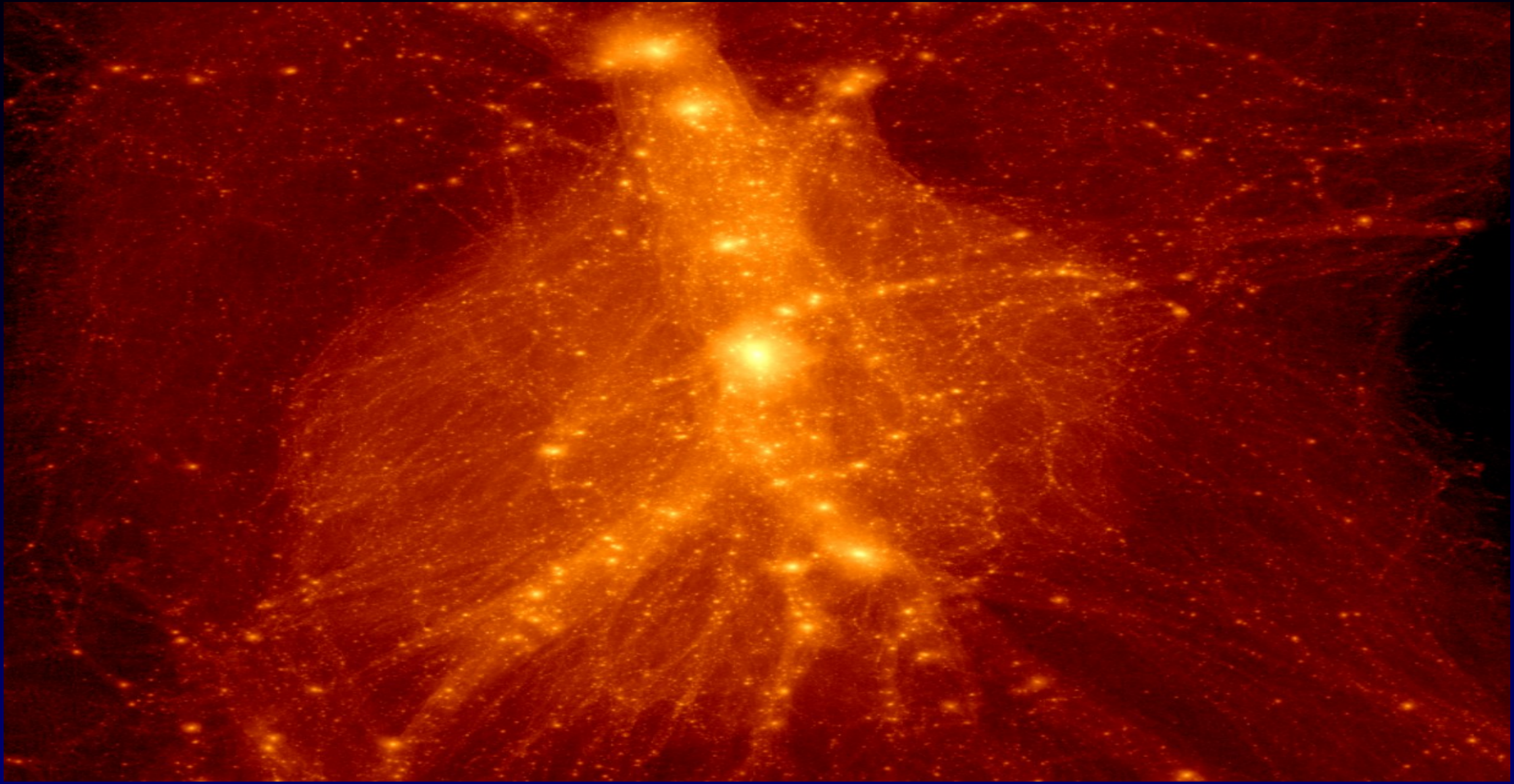
What good are the data if we cannot effectively extract knowledge from them?

Djorgovski

Visualization of 1 B points – Gaia DR1

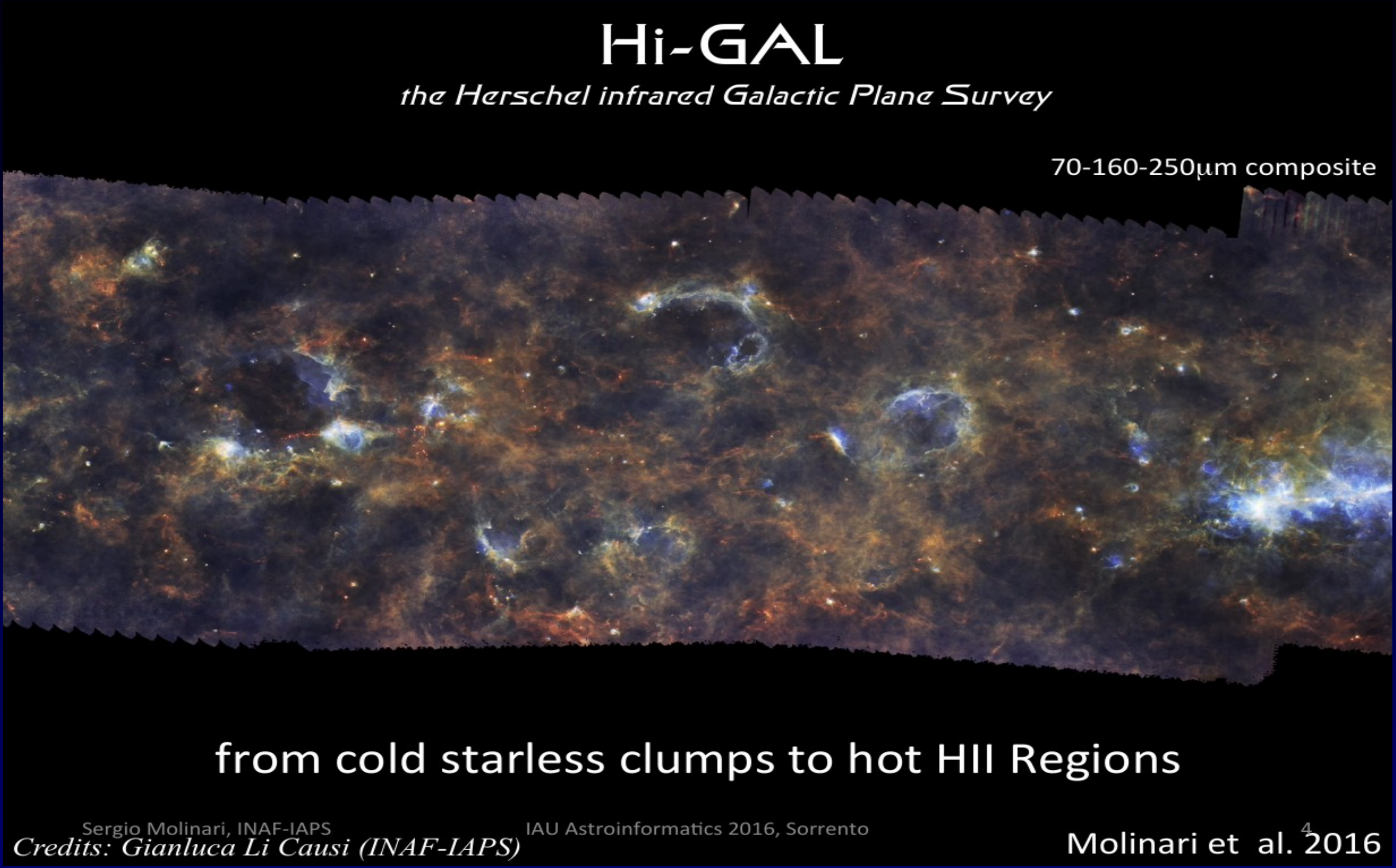


Visualization of Big Data



Breddels 2016

Star Forming Regions in Galaxy



Hi-GAL

the Herschel infrared Galactic Plane Survey

70-160-250μm composite

from cold starless clumps to hot HII Regions

Sergio Molinari, INAF-IAPS

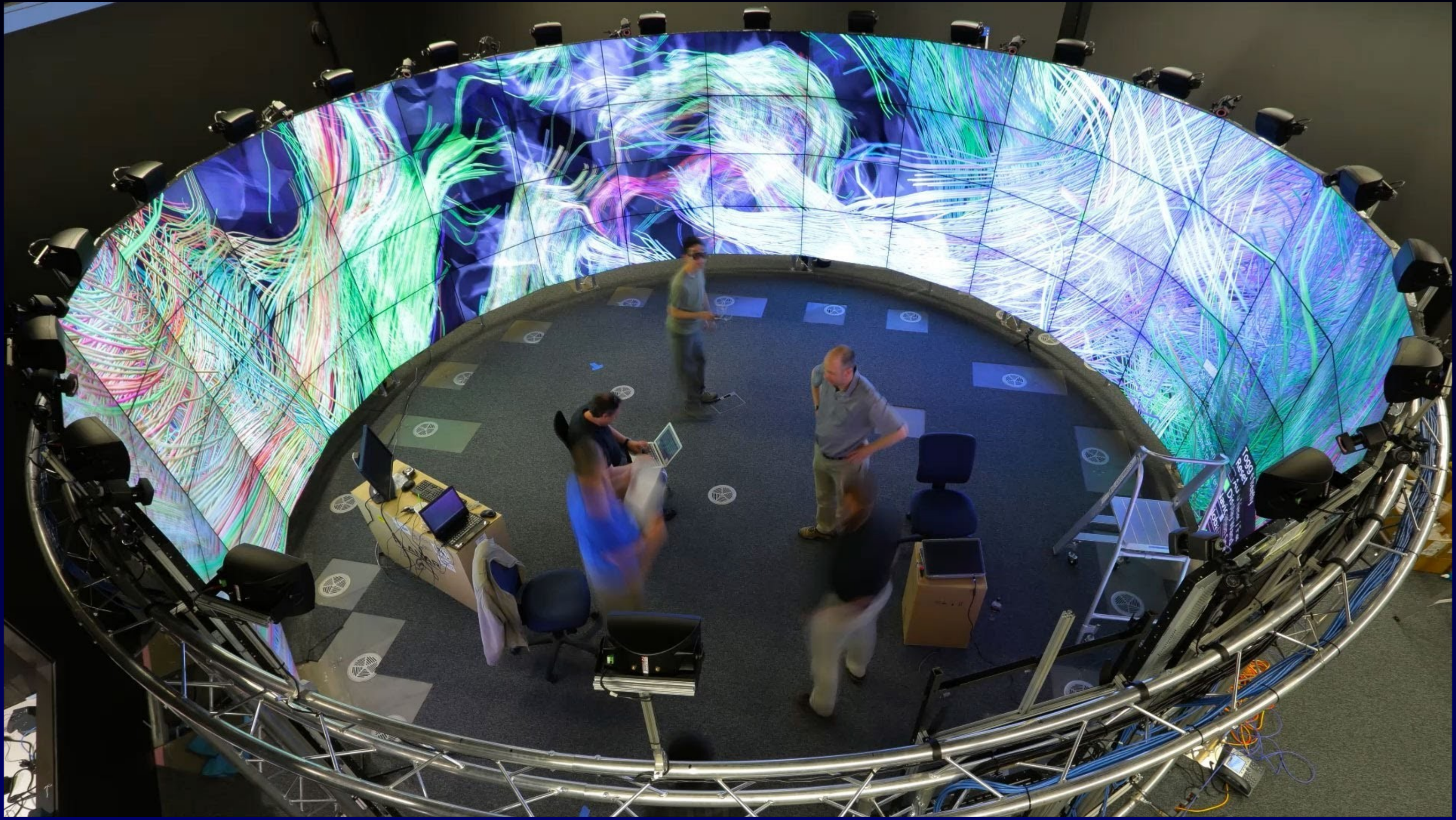
IAU Astroinformatics 2016, Sorrento

4

Credits: Gianluca Li Causi (INAF-IAPS)

Molinari et al. 2016

CAVE2 Monash University AU



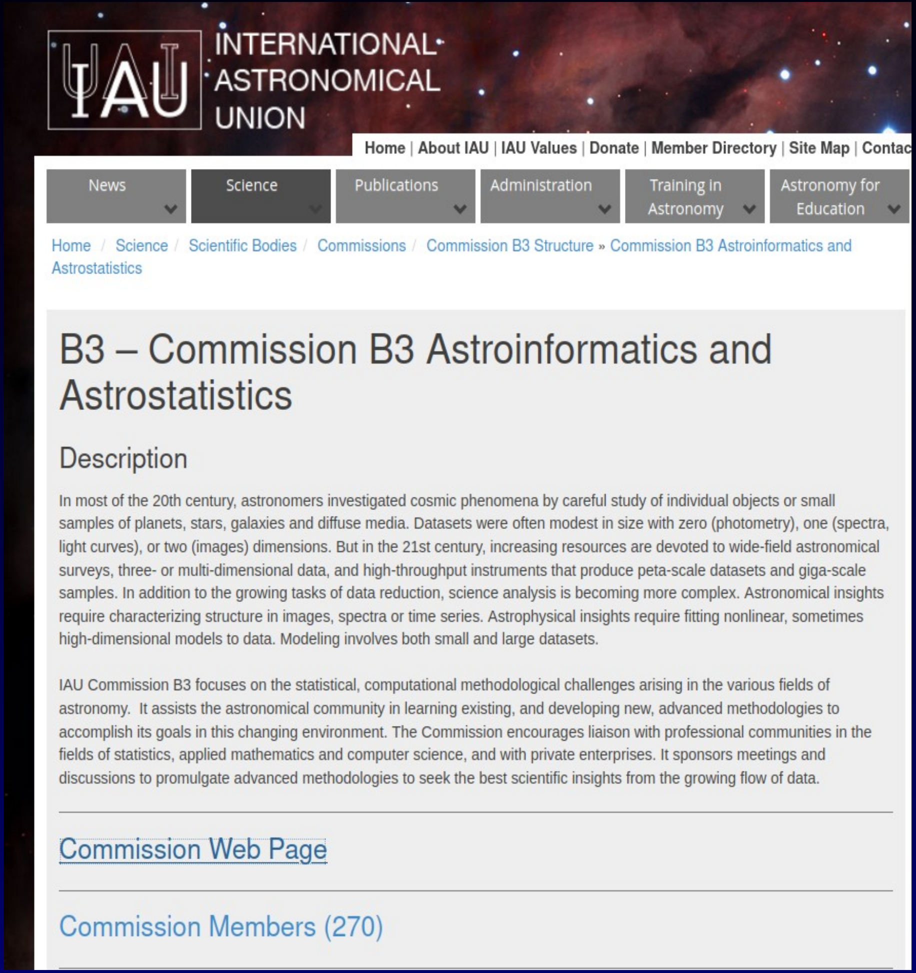
8m diameter, 330 deg FOV , 80x LCD 46“ 1366x768 Stereo + head tracking …...

Symposium S14

29 – 30 June 2017

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

Astroinformatics in IAU


INTERNATIONAL ASTRONOMICAL UNION
Home | About IAU | IAU Values | Donate | Member Directory | Site Map | Contac
News
Science
Publications
Administration
Training in Astronomy
Astronomy for Education
Home / Science / Scientific Bodies / Commissions / Commission B3 Structure » Commission B3 Astroinformatics and Astrostatistics

B3 – Commission B3 Astroinformatics and Astrostatistics

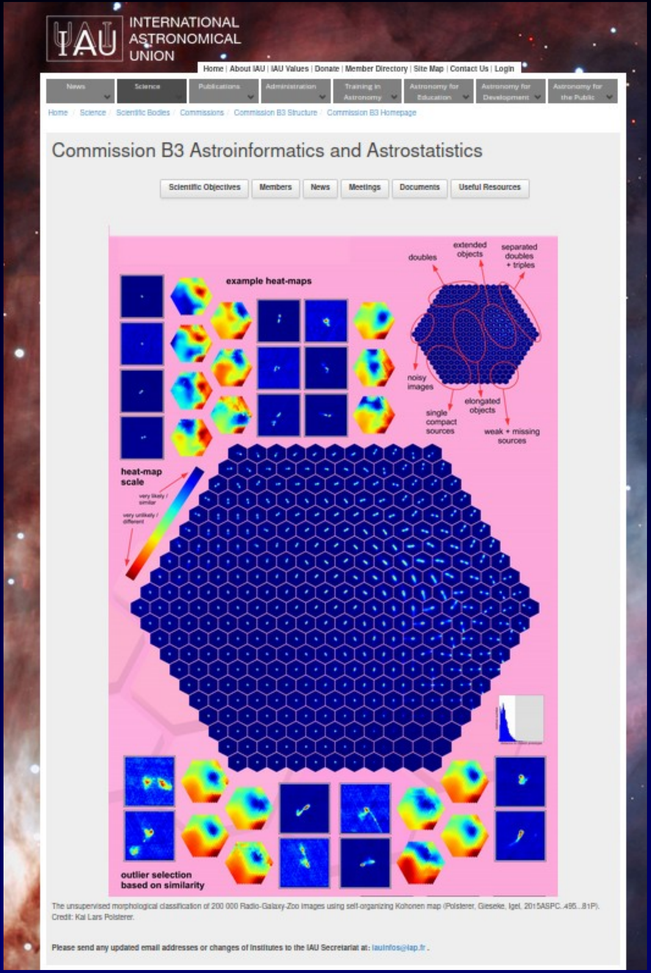
Description

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)

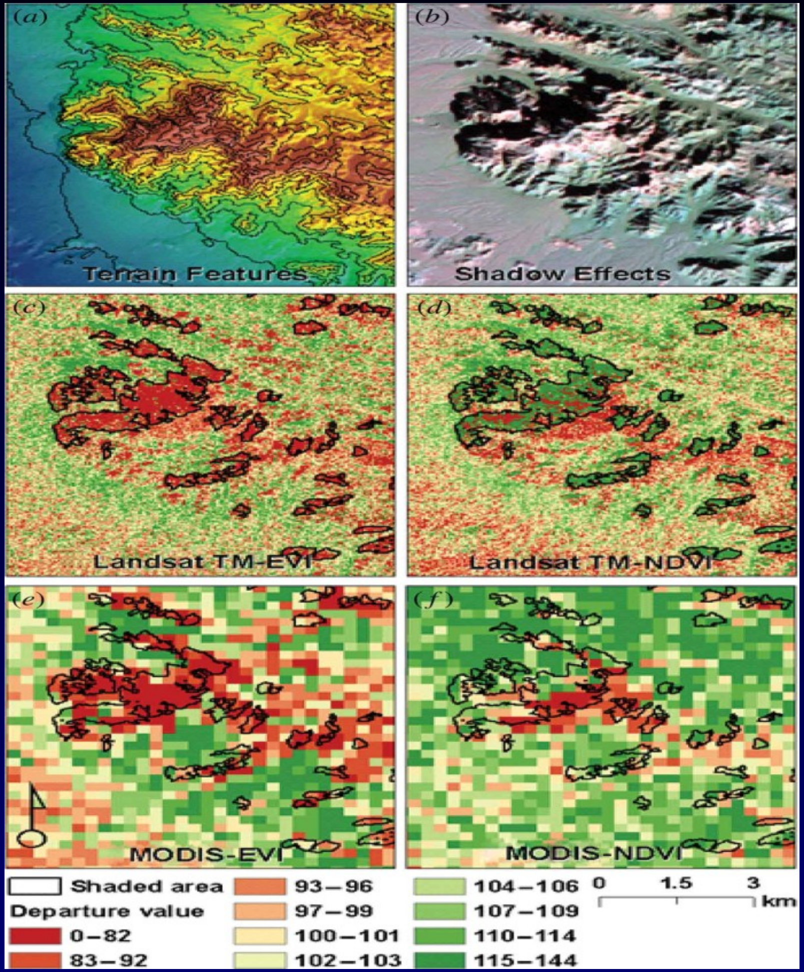

INTERNATIONAL ASTRONOMICAL UNION
Home | About IAU | IAU Values | Donate | Member Directory | Site Map | Contact Us | Login
News
Science
Publications
Administration
Training in Astronomy
Astronomy for Education
Astronomy for Development
Astronomy for the Public
Home / Science / Scientific Bodies / Commissions / Commission B3 Structure / Commission B3 Homepage
Scientific Objectives
Members
News
Meetings
Documents
Useful Resources
example heat-maps
doubles
extended objects
separated doubles + triples
noisy images
single compact sources
elongated objects
weak + missing sources
heat-map scale
very likely / similar
very unlikely / different
outlier selection based on similarity

Commission B3 Astroinformatics and Astrostatistics

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

Please send any updated email addresses or changes of institutes to the IAU Secretariat at: iauinfos@iap.fr .

Remote sensing – Big Data
Machine Learning



Precise farming

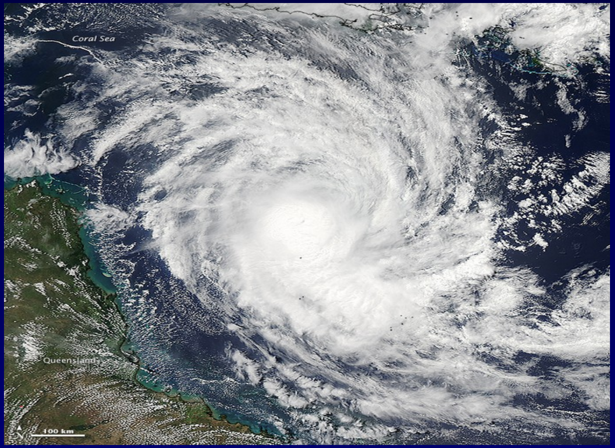
Forestry

Ore mining

Water resources monitoring

Automatic classification of terrain

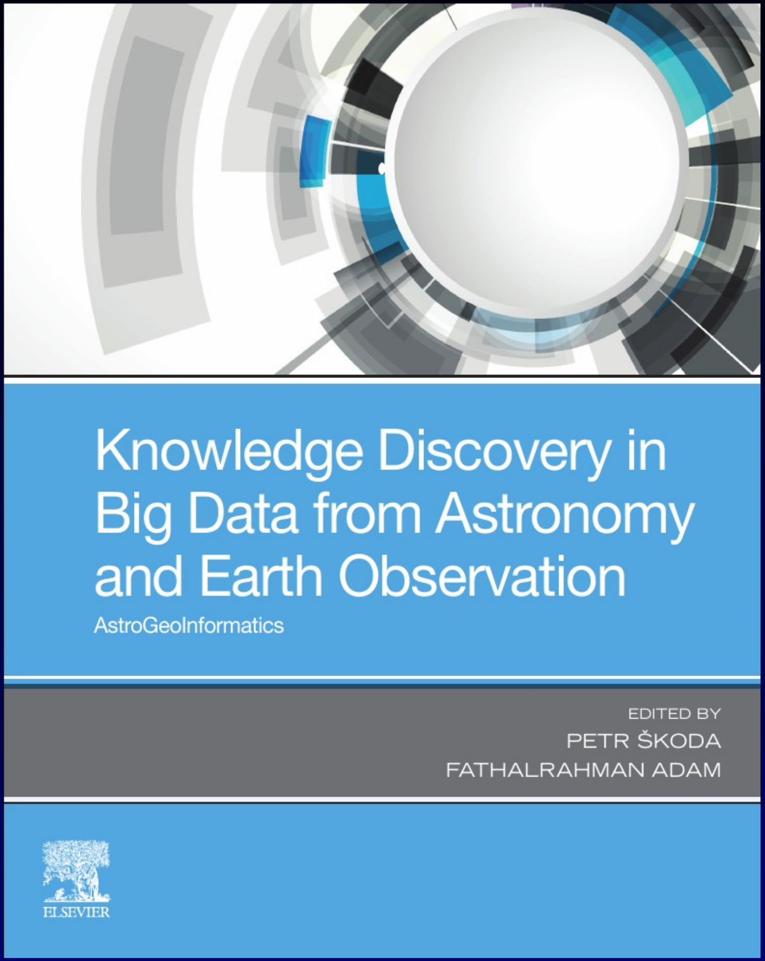
Resistence of buildings (Aquilla)



Visualization of Big Data

Breddels 2016

AstroGeoInformatics



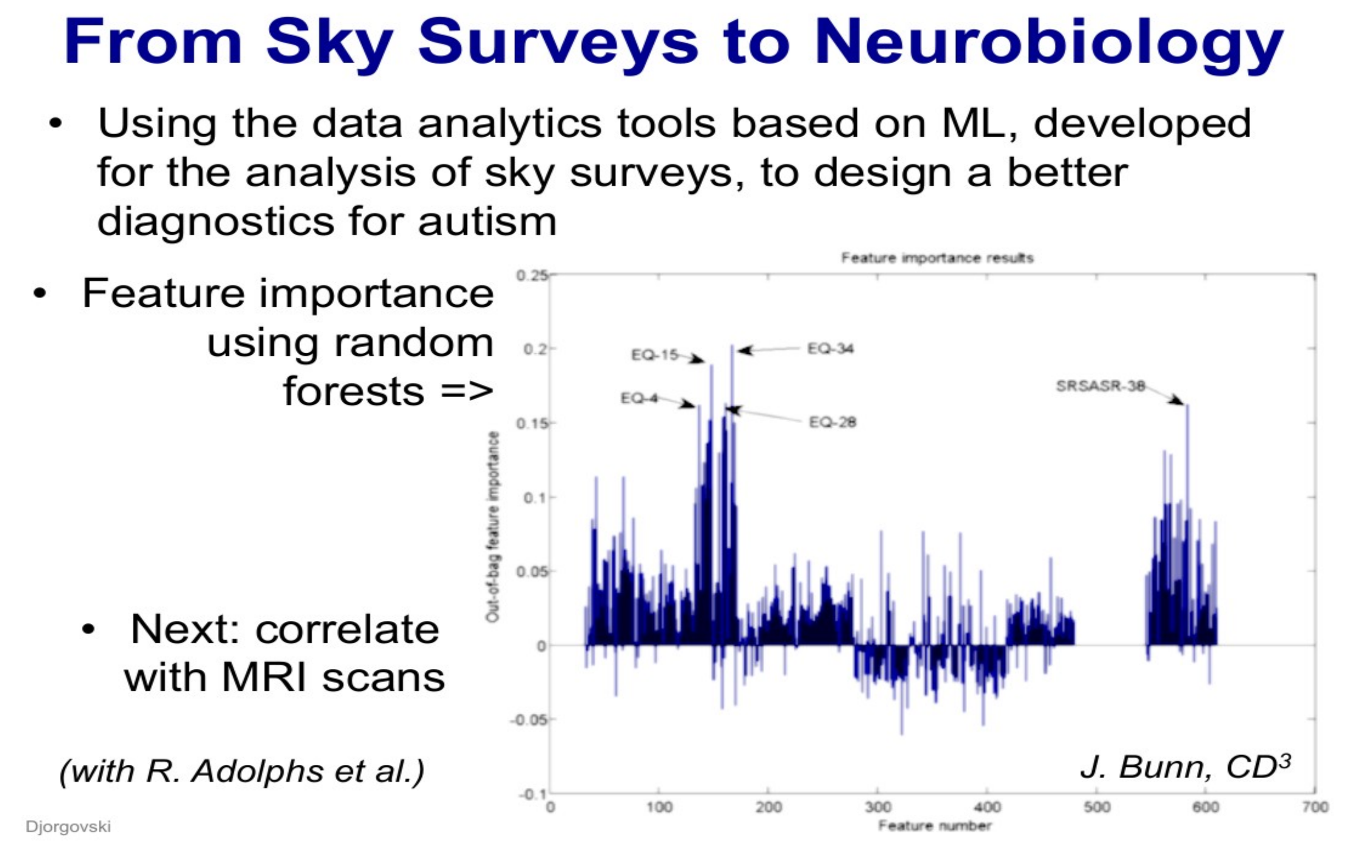
Knowledge Discovery in Big Data from Astronomy and Earth Observation

AstroGeoInformatics

EDITED BY
PETR ŠKODA
FATHALRAHMAN ADAM

ELSEVIER

Astro-Neurology



From Sky Surveys to Neurobiology

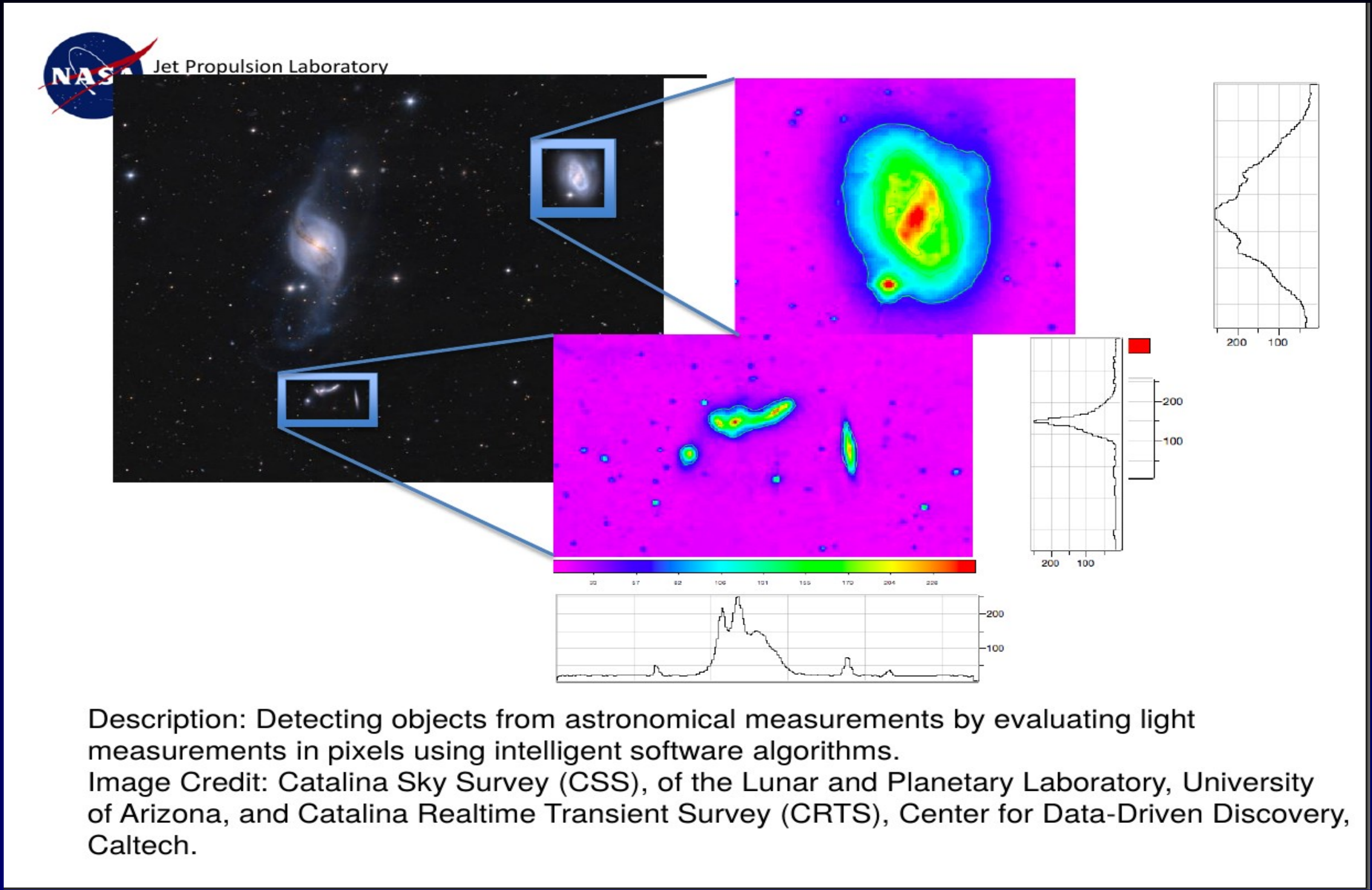
- Using the data analytics tools based on ML, developed for the analysis of sky surveys, to design a better diagnostics for autism
- Feature importance using random forests =>
- Next: correlate with MRI scans

(with R. Adolphs et al.)

J. Bunn, CD^3

Djorgovski

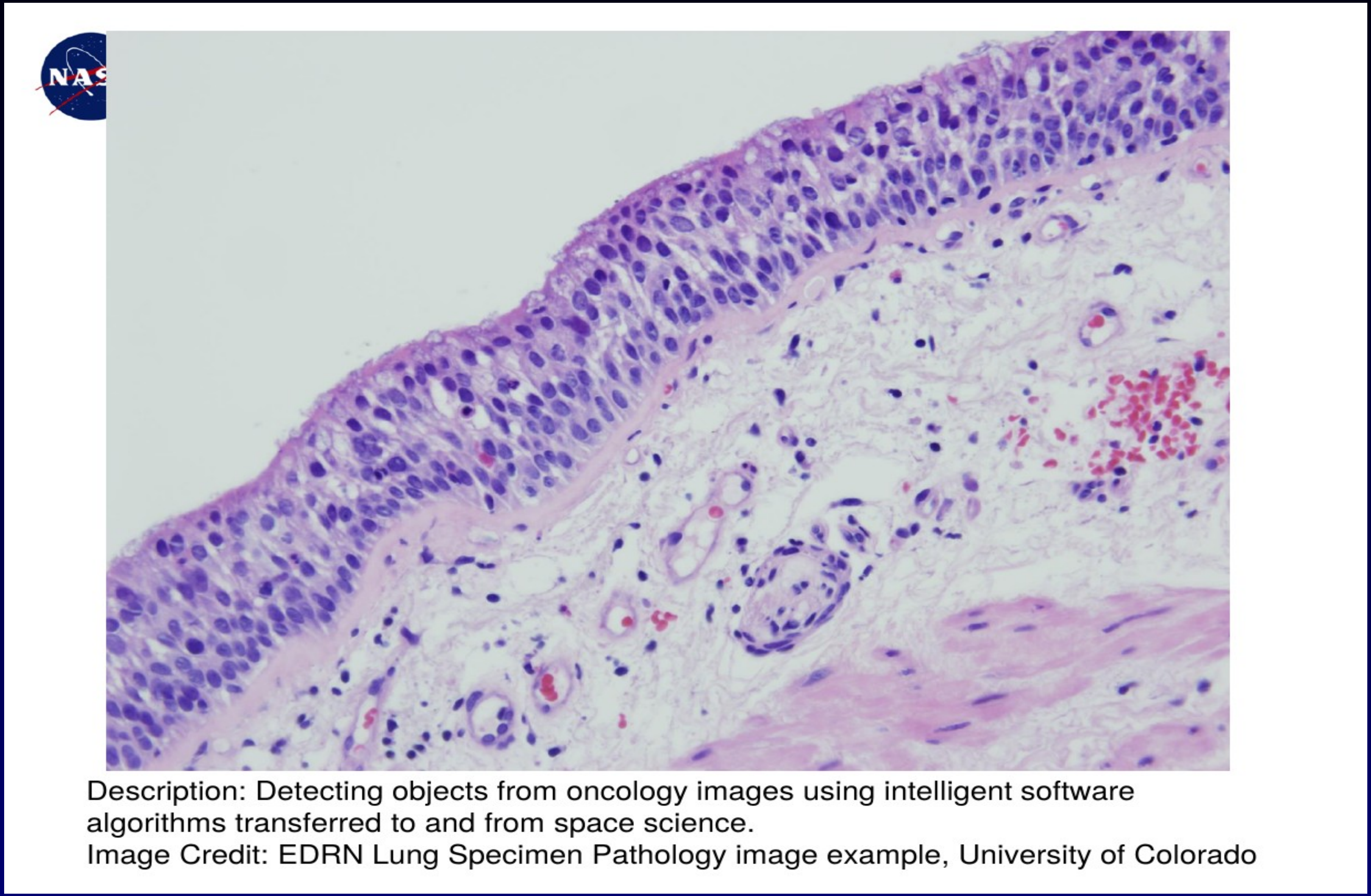
Finding Galaxies by Shape NASA



Description: Detecting objects from astronomical measurements by evaluating light measurements in pixels using intelligent software algorithms.

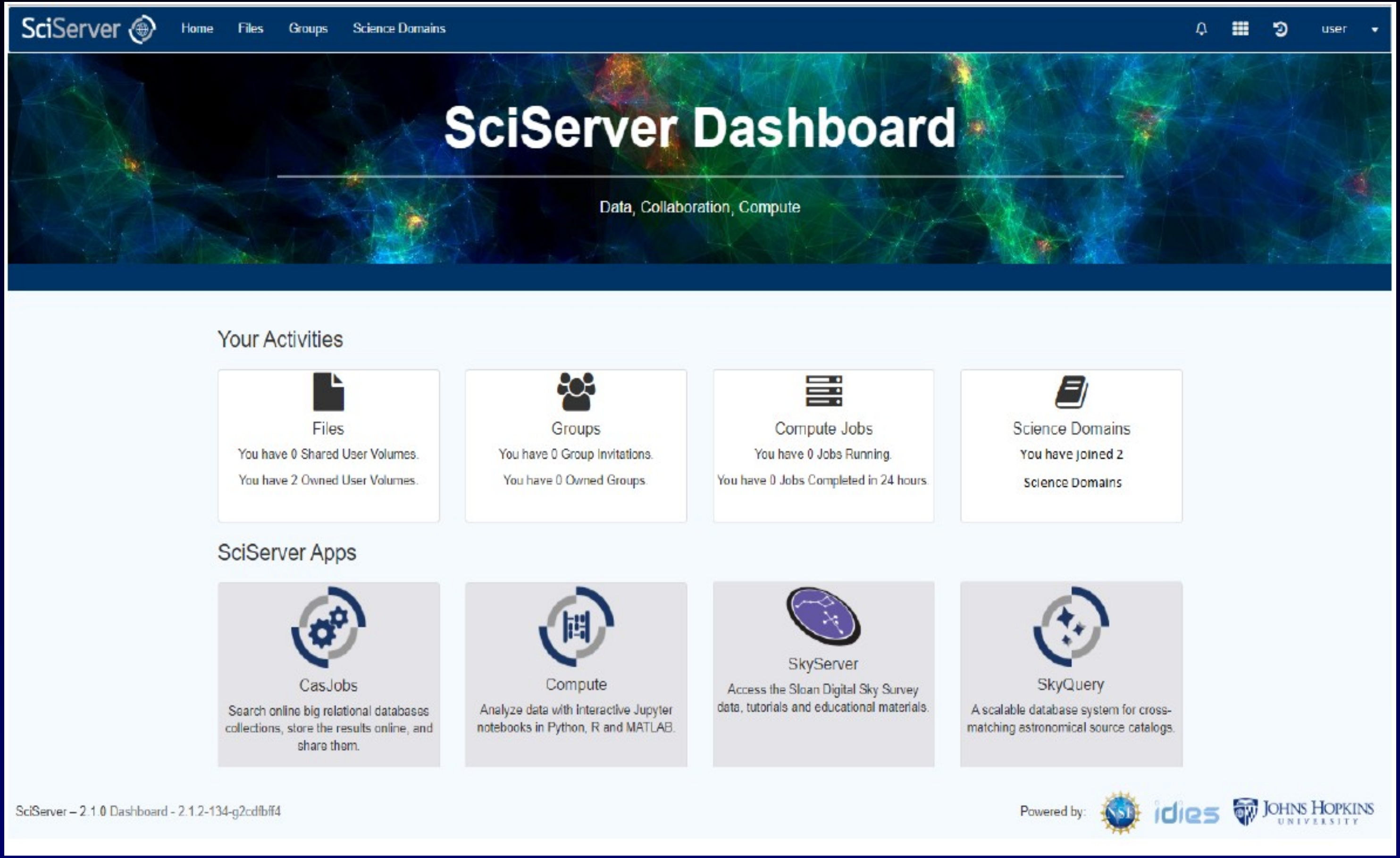
Image Credit: Catalina Sky Survey (CSS), of the Lunar and Planetary Laboratory, University of Arizona, and Catalina Realtime Transient Survey (CRTS), Center for Data-Driven Discovery, Caltech.

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

Science Platforms - SciServer



SciServer Dashboard

Data, Collaboration, Compute

Your Activities

Files

You have 0 Shared User Volumes.

You have 2 Owned User Volumes.

Groups

You have 0 Group Invitations.

You have 0 Owned Groups.

Compute Jobs

You have 0 Jobs Running.

You have 0 Jobs Completed in 24 hours.

Science Domains

You have joined 2 Science Domains

SciServer Apps

CasJobs

Search online big relational databases collections, store the results online, and share them.

Compute

Analyze data with interactive Jupyter notebooks in Python, R and MATLAB.

SkyServer

Access the Sloan Digital Sky Survey data, tutorials and educational materials.

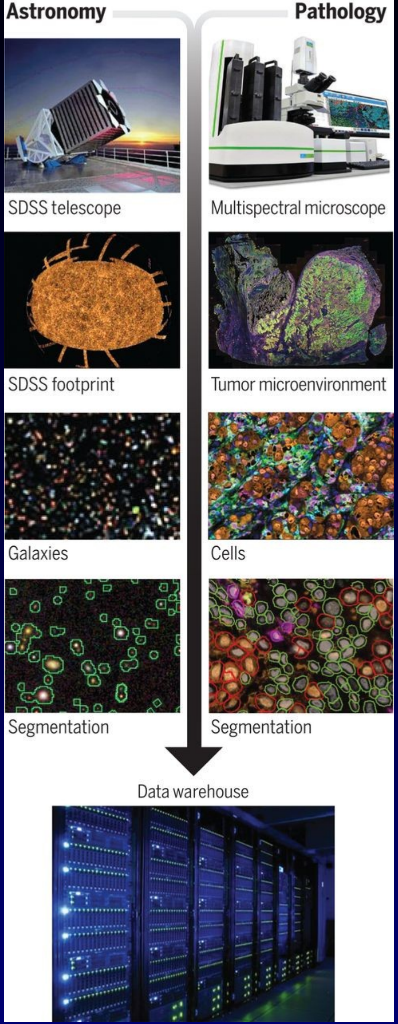
SkyQuery

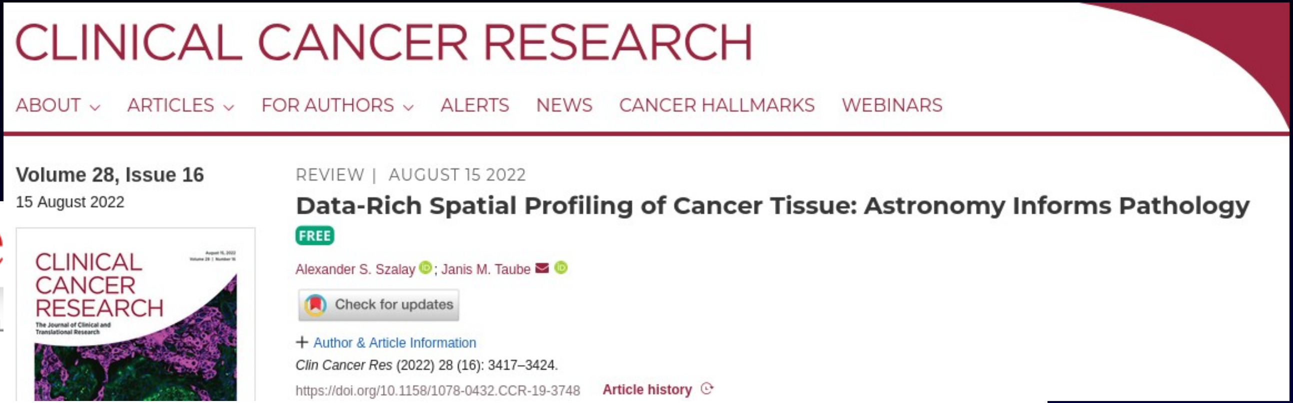
A scalable database system for cross-matching astronomical source catalogs.

SciServer – 2.1.0 Dashboard - 2.1.2-134-g2cdfbff4

Powered by: NSF idies Johns Hopkins University

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

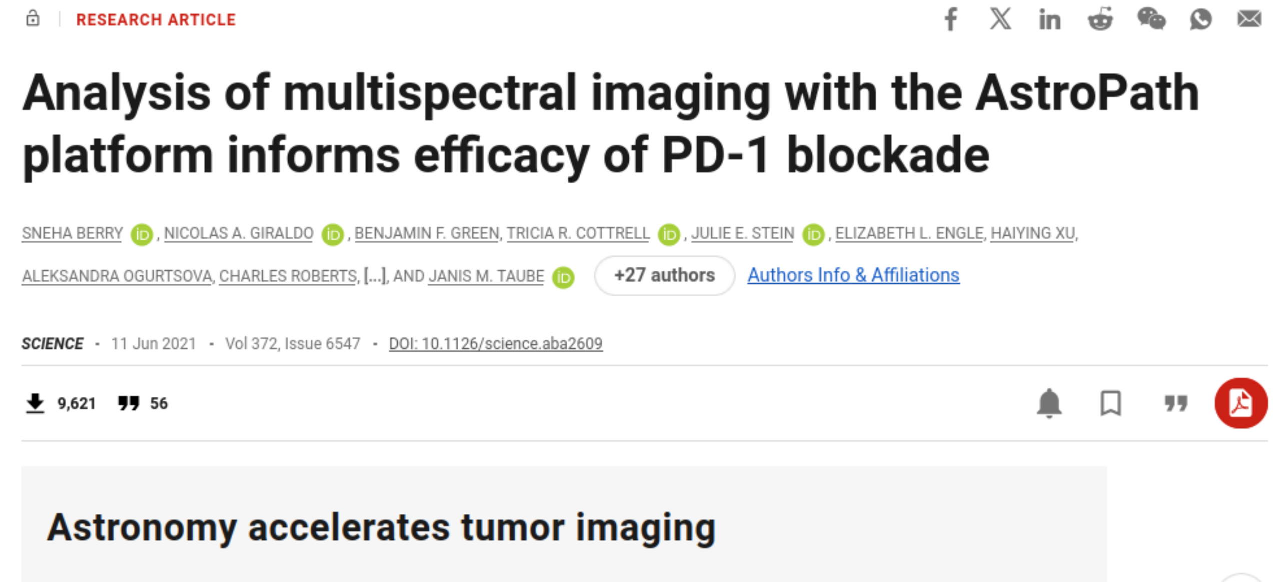
Alexander S. Szalay; Janis M. Taube

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



RESEARCH ARTICLE

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

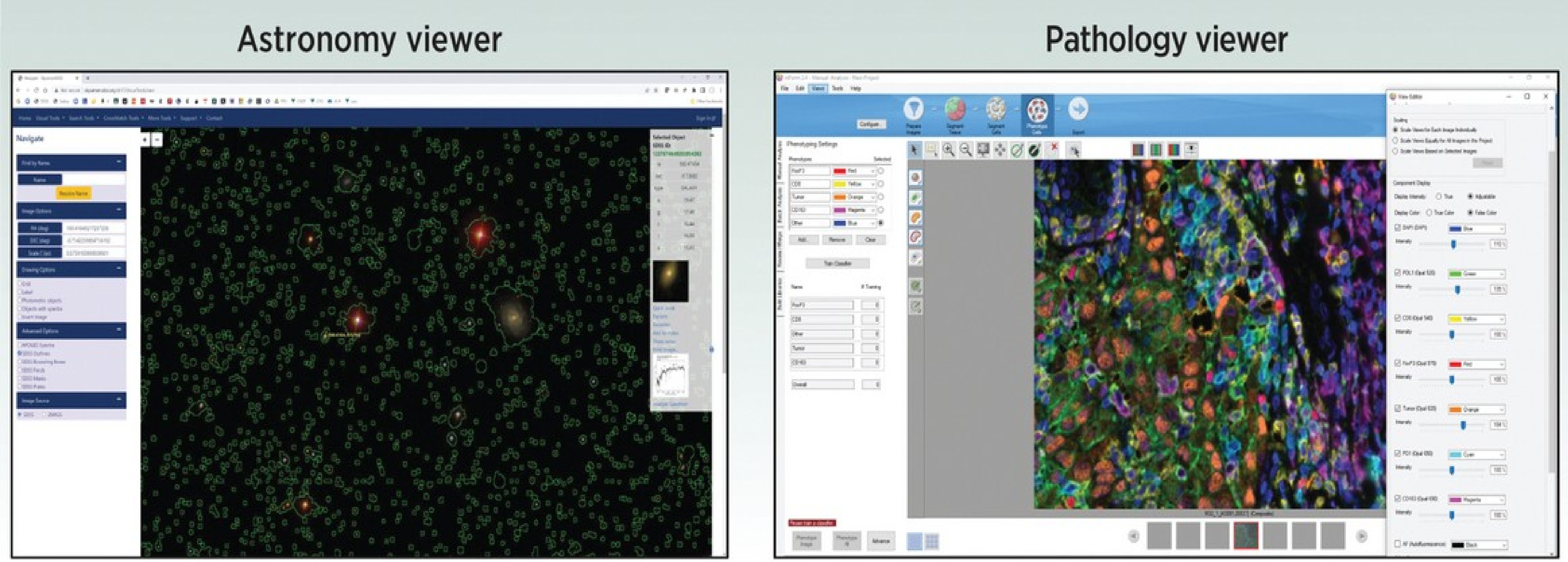
SNEHA BERRY, NICOLAS A. GIRALDO, BENJAMIN F. GREEN, TRICIA R. COTTRELL, JULIE E. STEIN, ELIZABETH L. ENGLE, HAIYING XU, ALEKSANDRA OGURTSOVA, CHARLES ROBERTS, [...], AND JANIS M. TAUBE +27 authors Authors Info & Affiliations

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

9,621 56

Astronomy accelerates tumor imaging

Digital Pathology based on SDSS

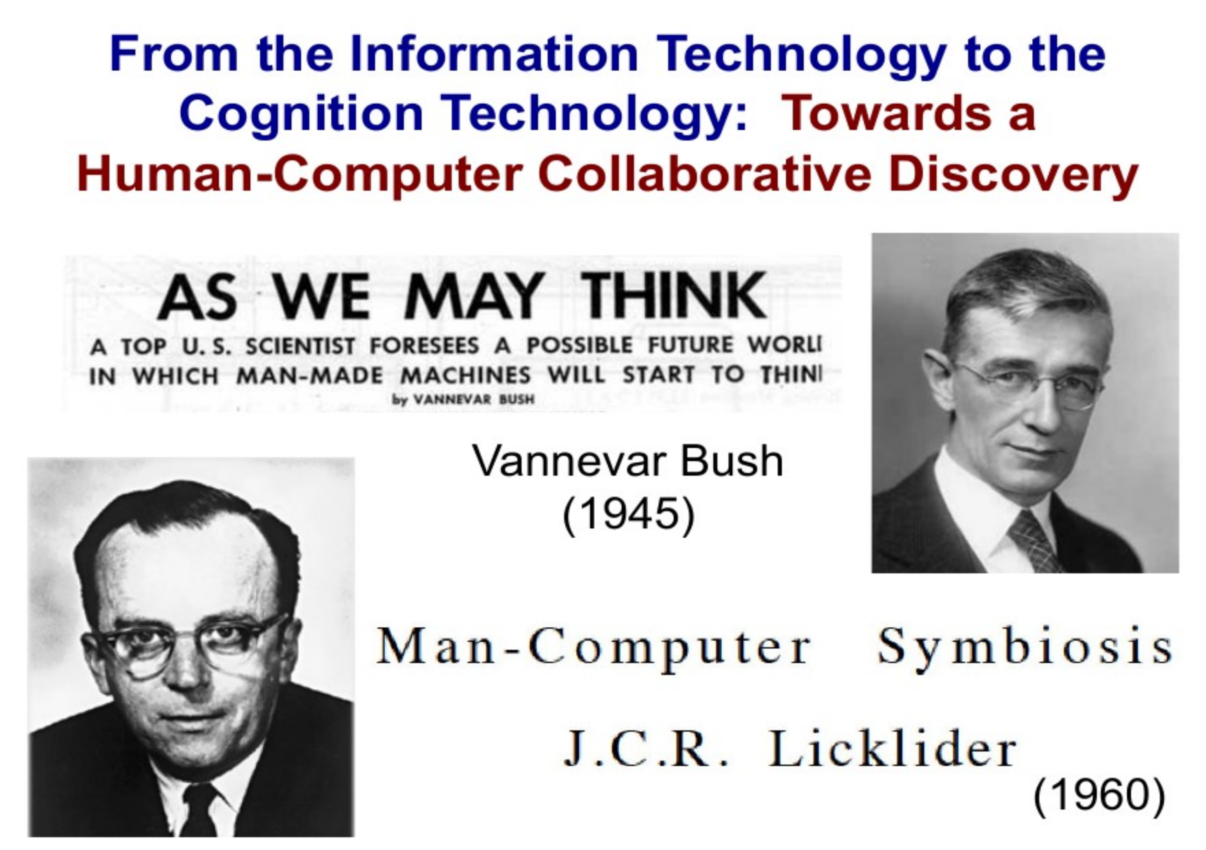


Astronomy viewer

Pathology viewer

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 WORL
IN WHICH MAN-MADE MACHINES WILL START TO THIN
by VANNEVAR BUSH

Vannevar Bush
(1945)

Man-Computer Symbiosis

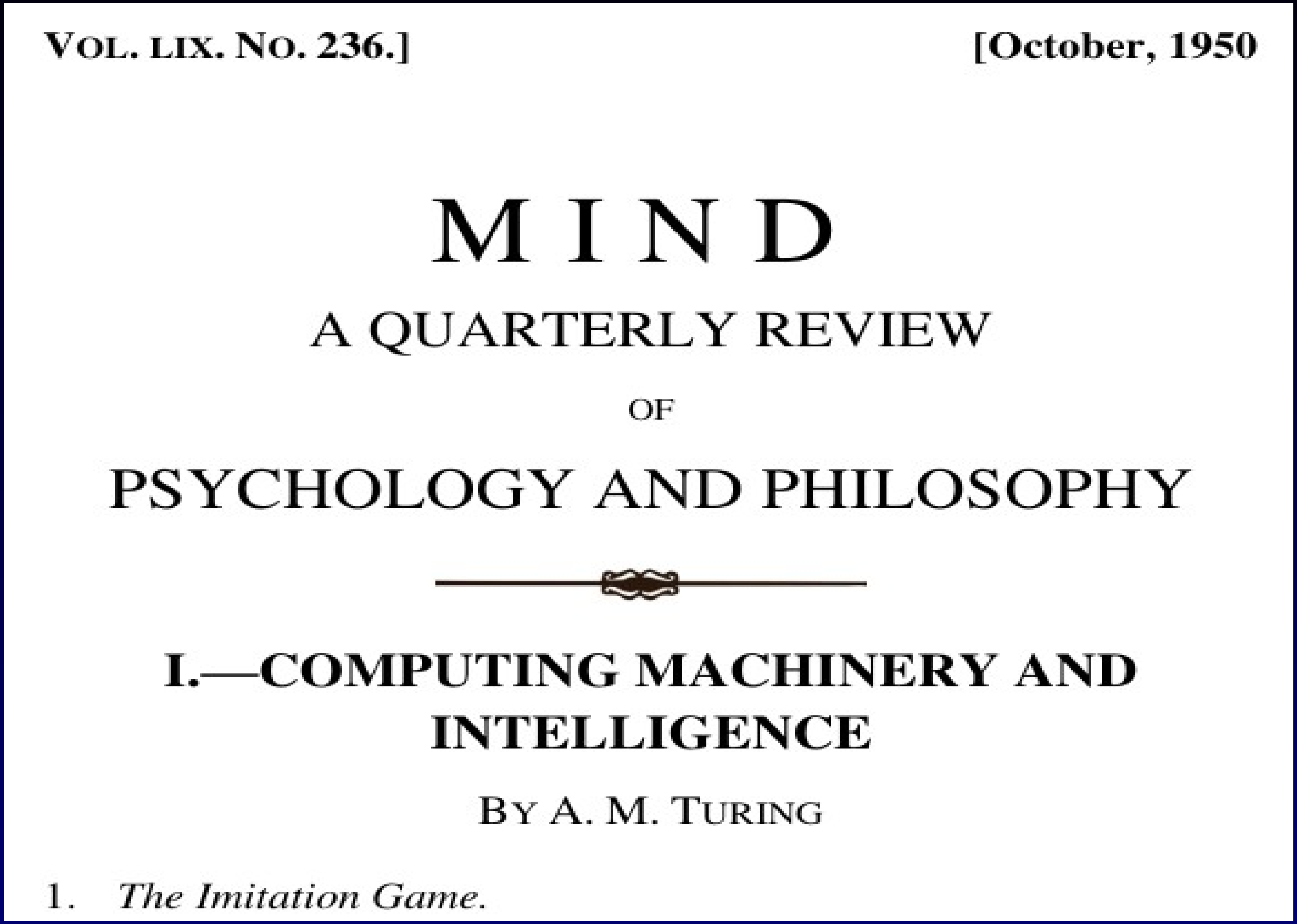
J.C.R. Licklider
(1960)

MEMEX

„Hypertext“
Associations

Global network
Interactivity
Graphics

Turing Test



VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

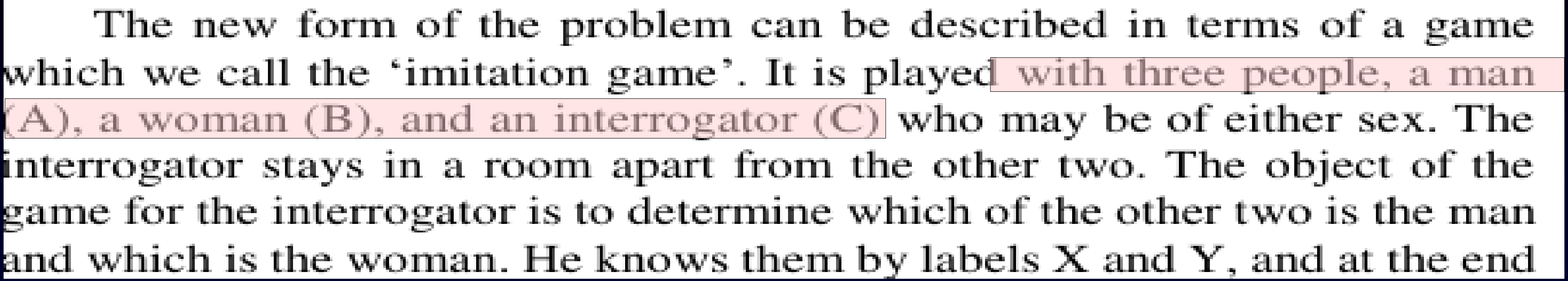
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

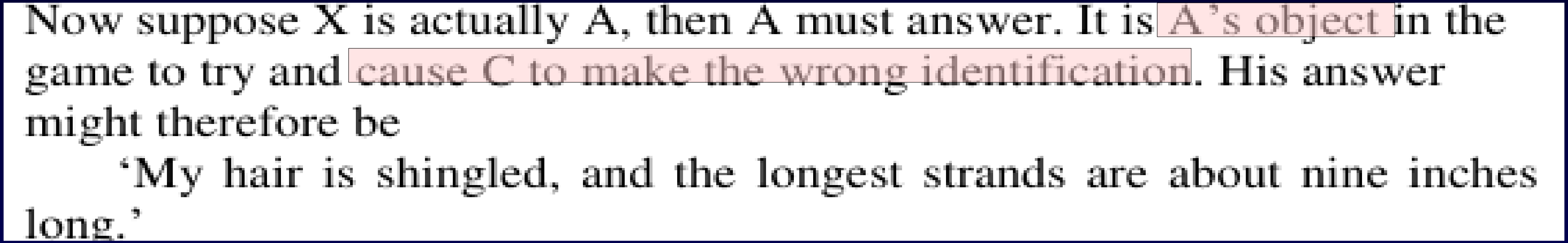
BY A. M. TURING

1. *The Imitation Game.*

Imitation Game

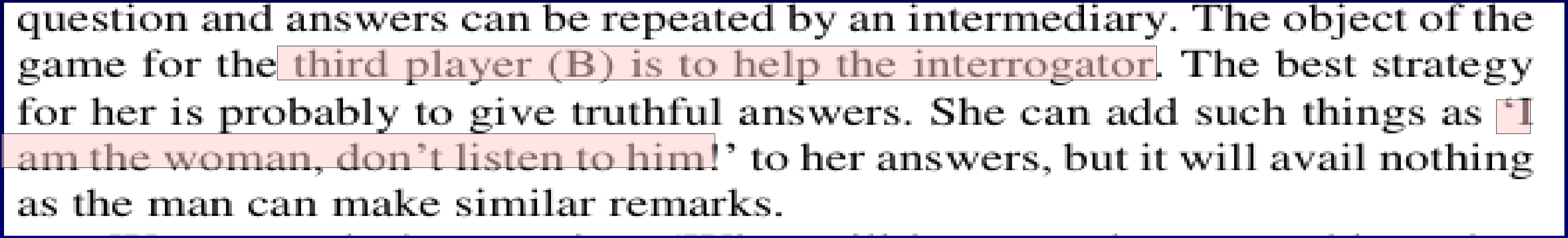


The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator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



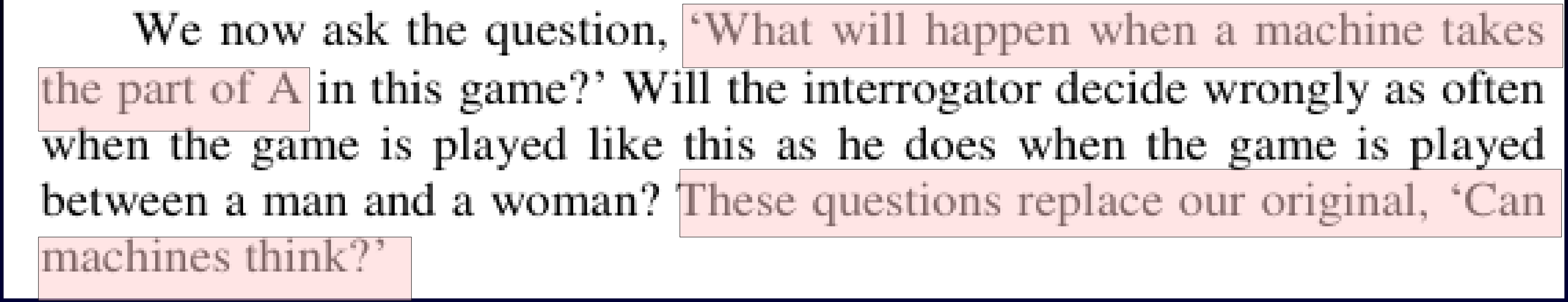
Now suppose X is actually A, then A must answer. It is A's object in the game to try and cause C to make the wrong identification. His answer might therefore be

'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

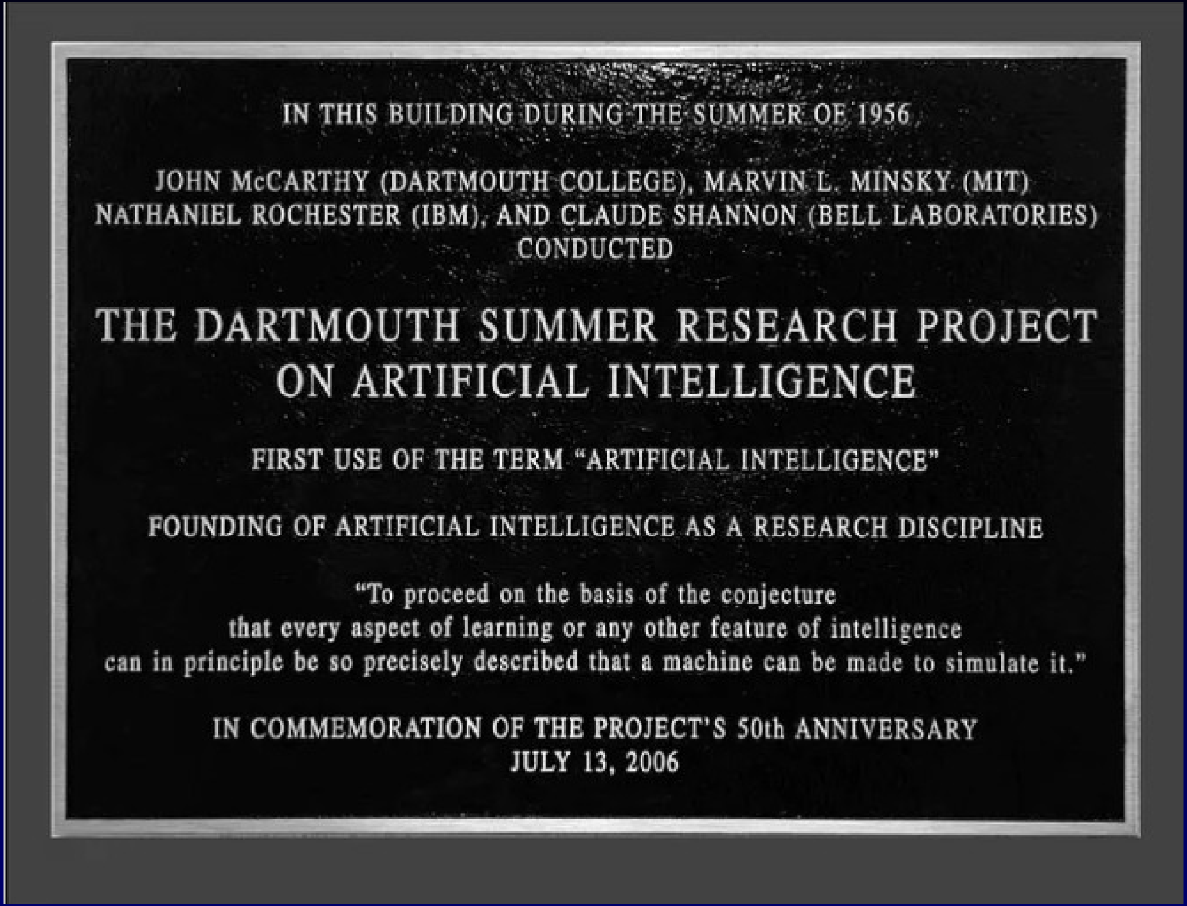


We now ask the question, 'What will happen when a machine takes the part of A in this game?' Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, 'Can machines think?'

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

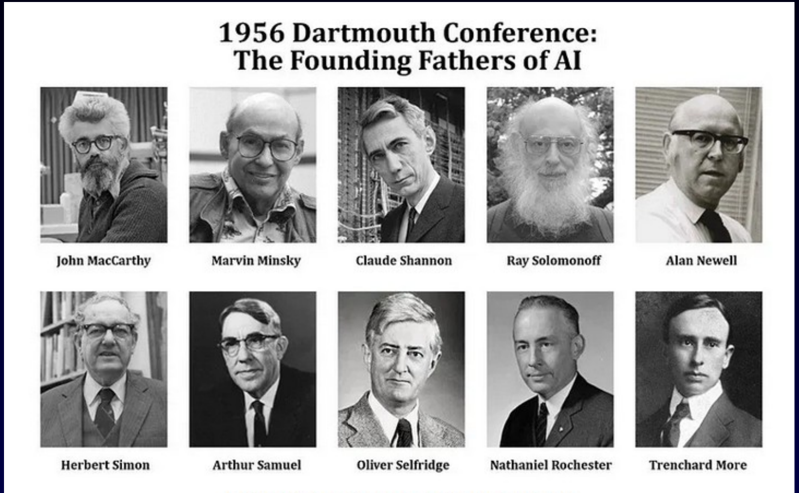
THE DARTMOUTH SUMMER RESEARCH PROJECT
ON ARTIFICIAL INTELLIGENCE

FIRST USE OF THE TERM “ARTIFICIAL INTELLIGENCE”

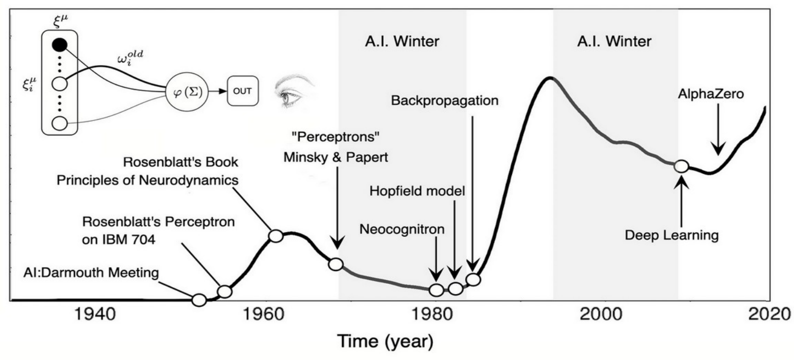
FOUNDING OF ARTIFICIAL INTELLIGENCE AS A RESEARCH DISCIPLINE

“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

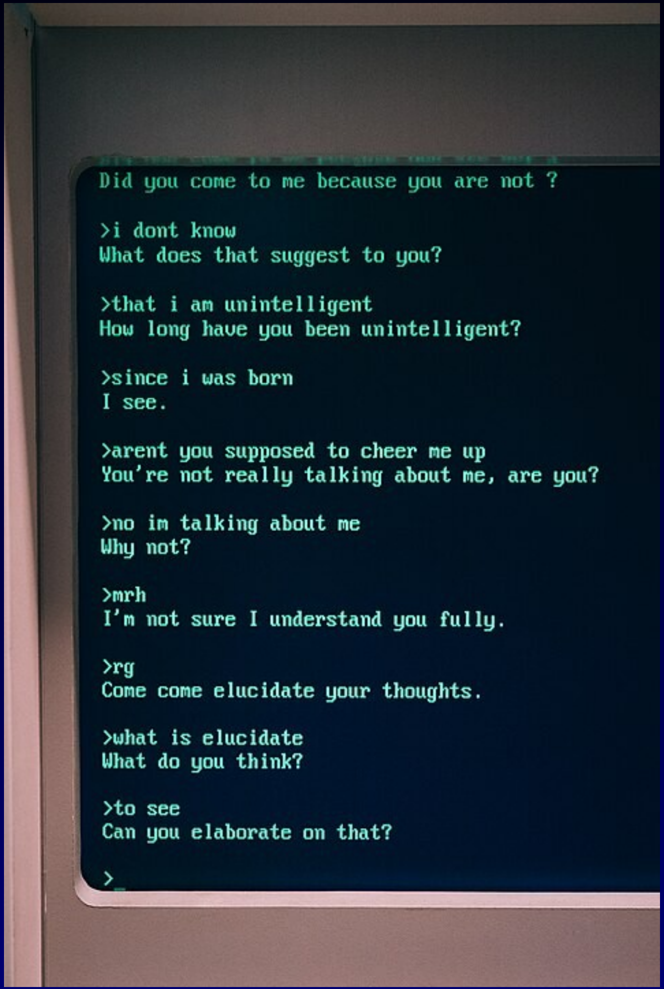


Founding fathers of AI. Courtesy of scienceabc.com



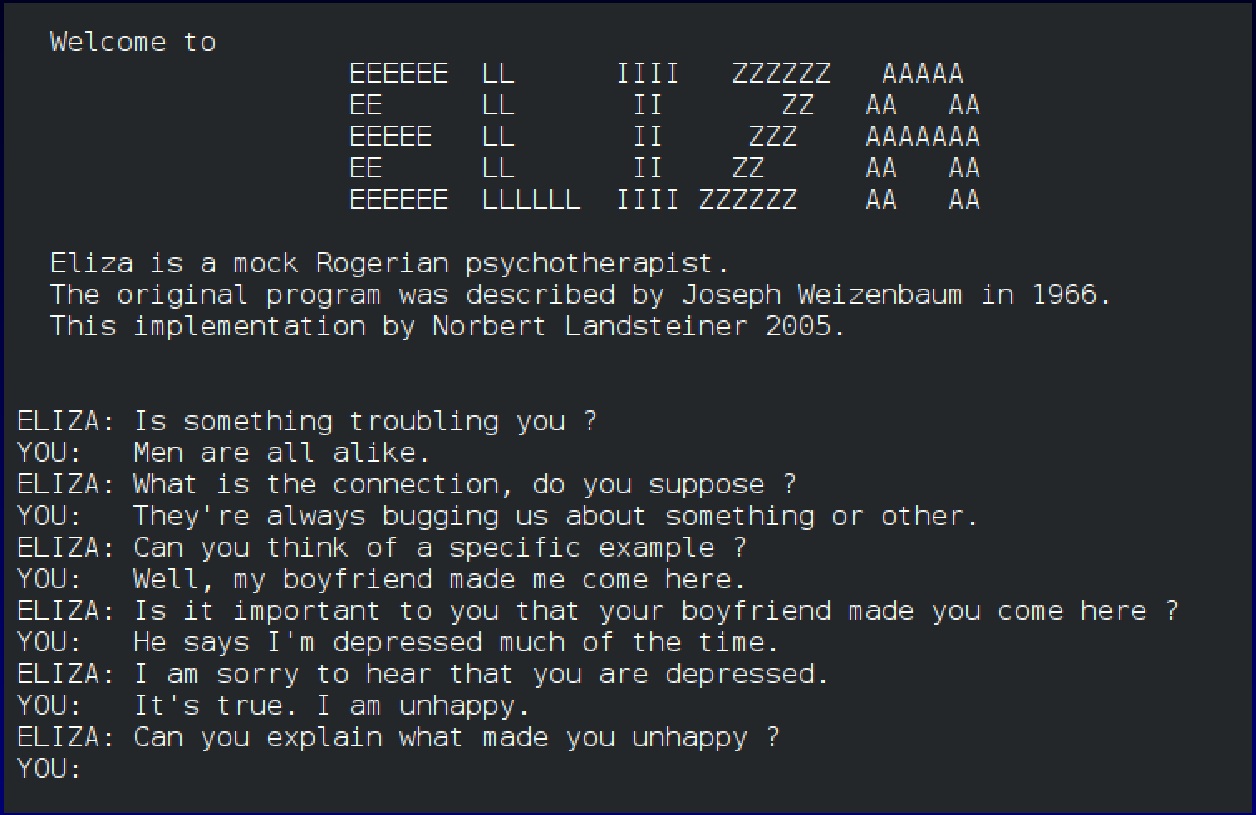
https://medium.com/rla-academy/dartmouth-workshop-the-birthplace-of-ai-34c533afe992

First Chatbot - ELIZA



```
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?

>
```

Weizenbaum MIT   1964-67



```
  Welcome to
                    EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                    EE      LL       II       ZZ   AA   AA
                    EEEEE   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)

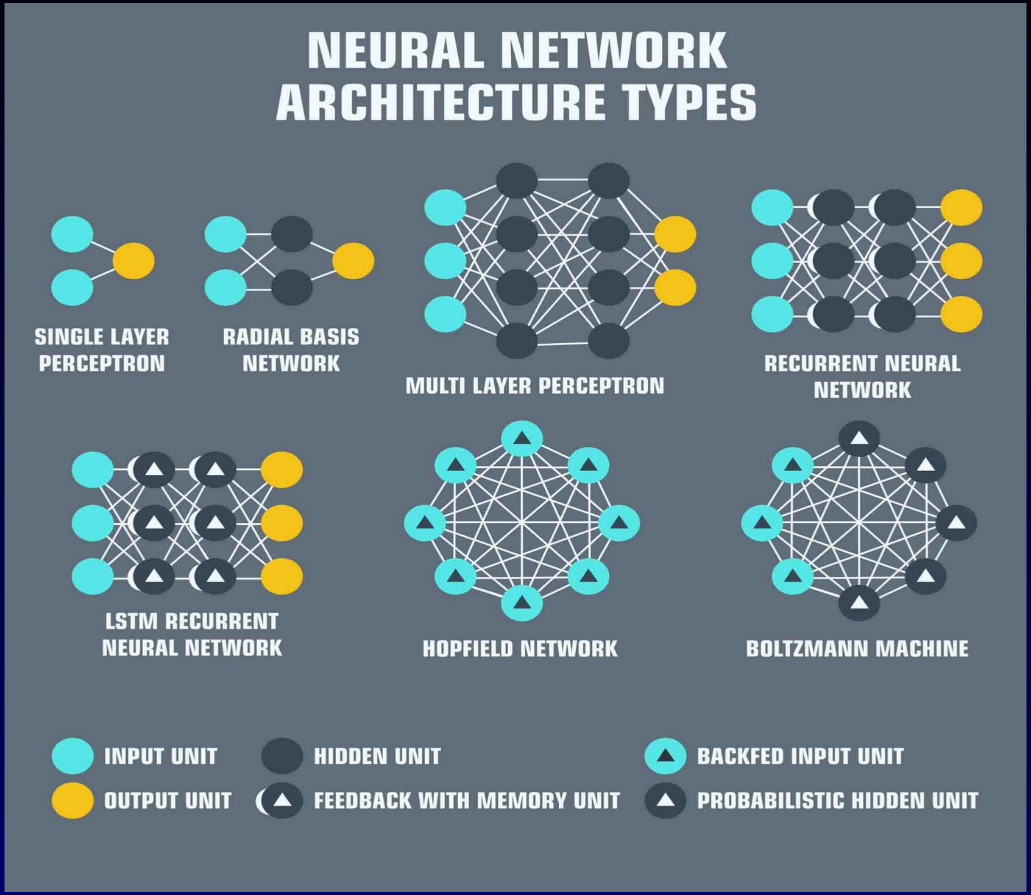
Bayesian Deep Learning

Physics Informed (Aware) NN
    Theory is important again

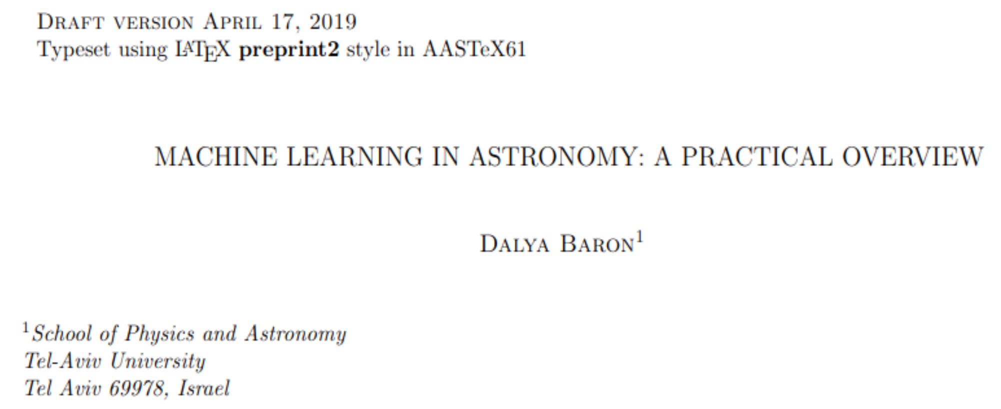
GANs (image creation)

Diffusion models

Foundation models



Reviews of ML/AI in Astronomy

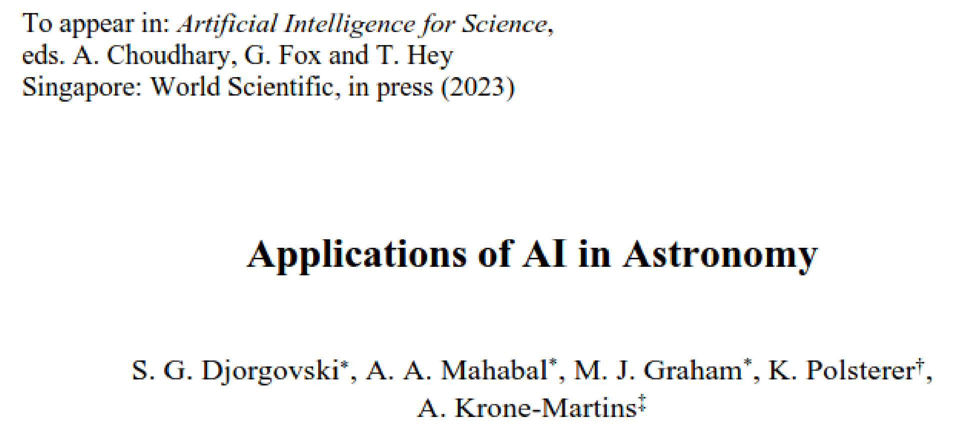


DRAFT VERSION APRIL 17, 2019
Typeset using LaTeX **preprint2** style in AASTeX61

MACHINE LEARNING IN ASTRONOMY: A PRACTICAL OVERVIEW

DALYA BARON[1]

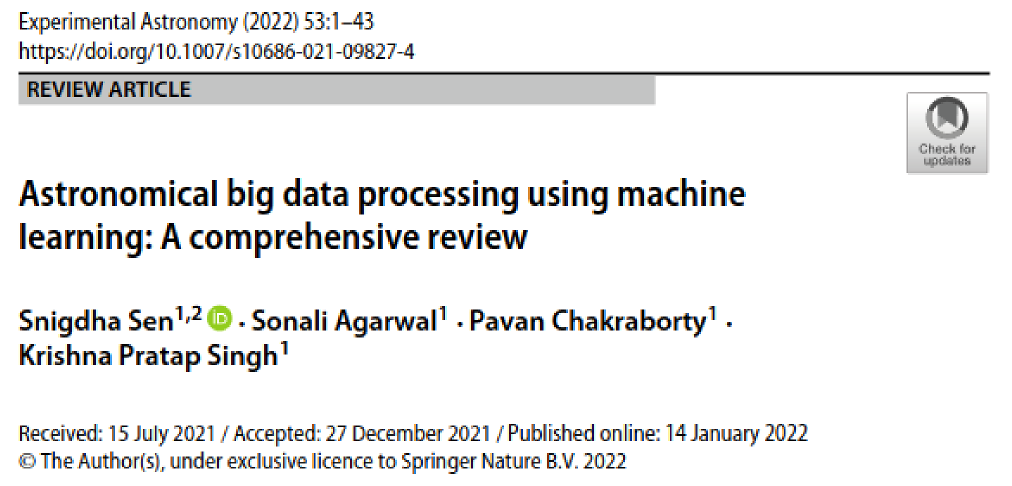
[1] *School of Physics and Astronomy*
Tel-Aviv University
Tel Aviv 69978, Israel



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†,
A. Krone-Martins‡



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

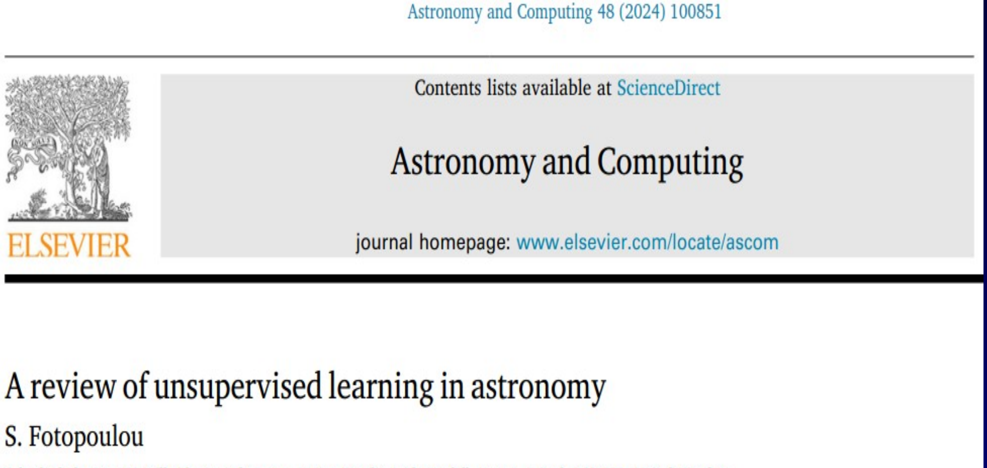
REVIEW ARTICLE

Check for updates

Astronomical big data processing using machine learning: A comprehensive review

Snigdha Sen[1,2] · Sonali Agarwal[1] · Pavan Chakraborty[1] · Krishna Pratap Singh[1]

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

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

ELSEVIER

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

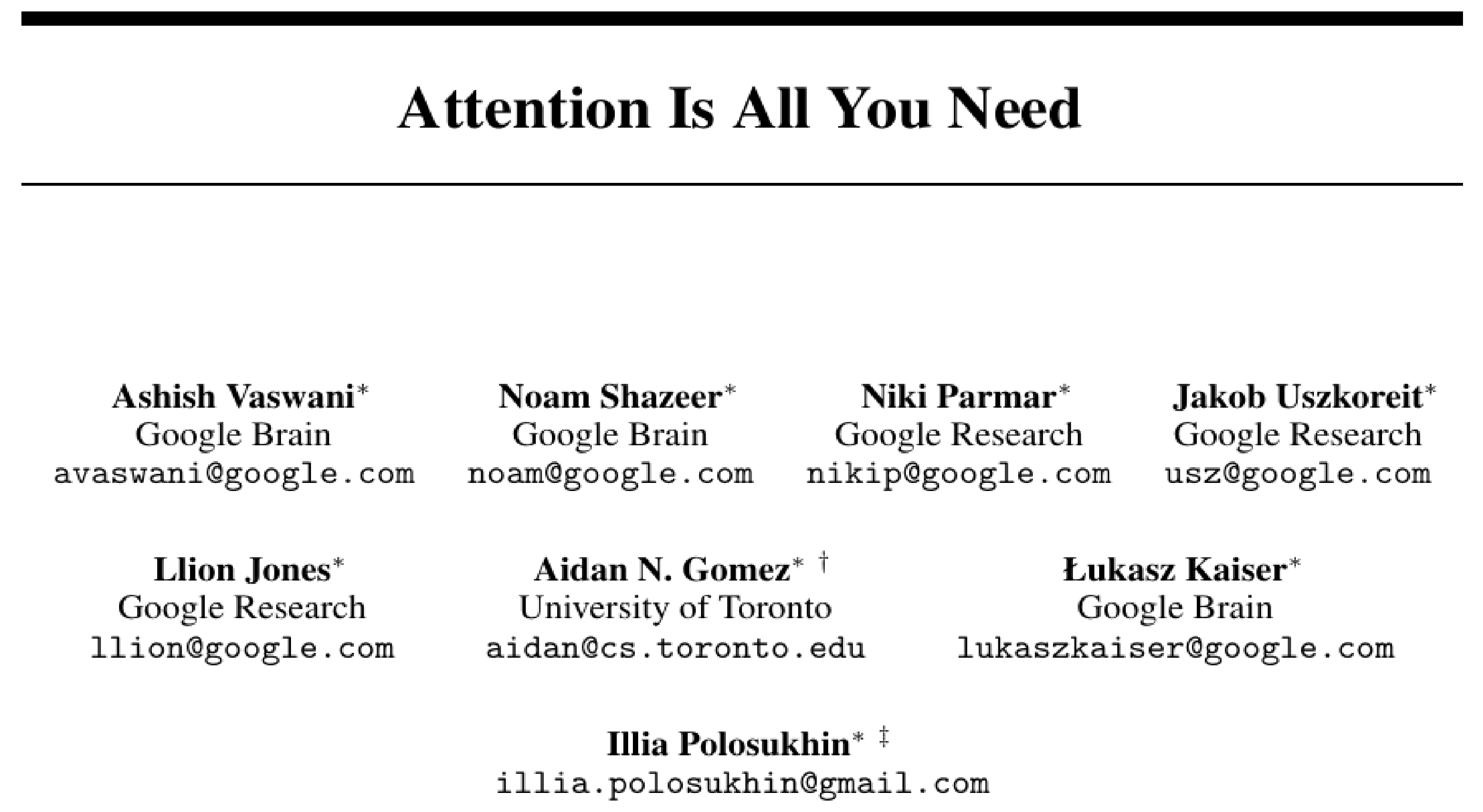
word2vec

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

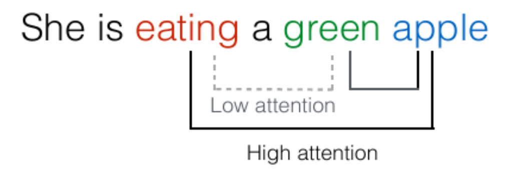
Aidan N. Gomez[*, †]
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser[*]
Google Brain
lukaszkaiser@google.com

Illia Polosukhin[*, ‡]
illia.polosukhin@gmail.com

NIPS 2017

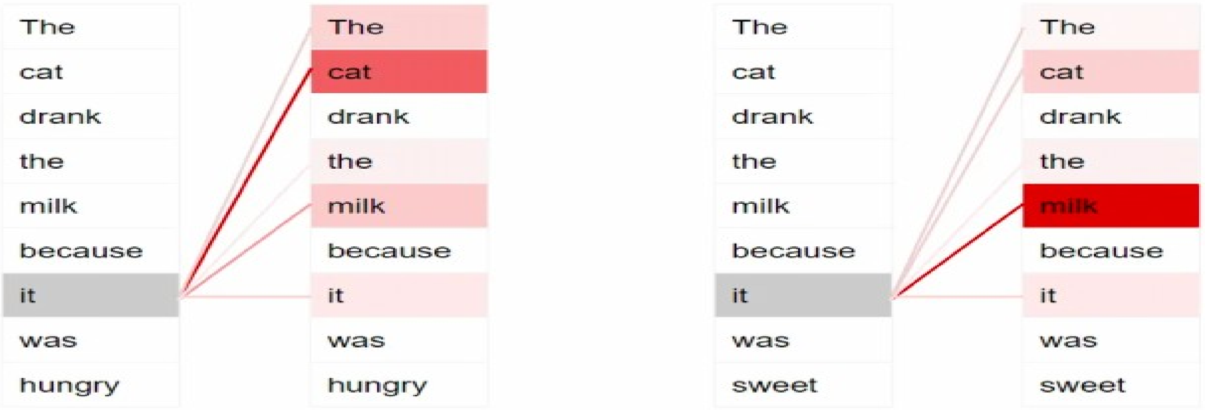
Attention



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

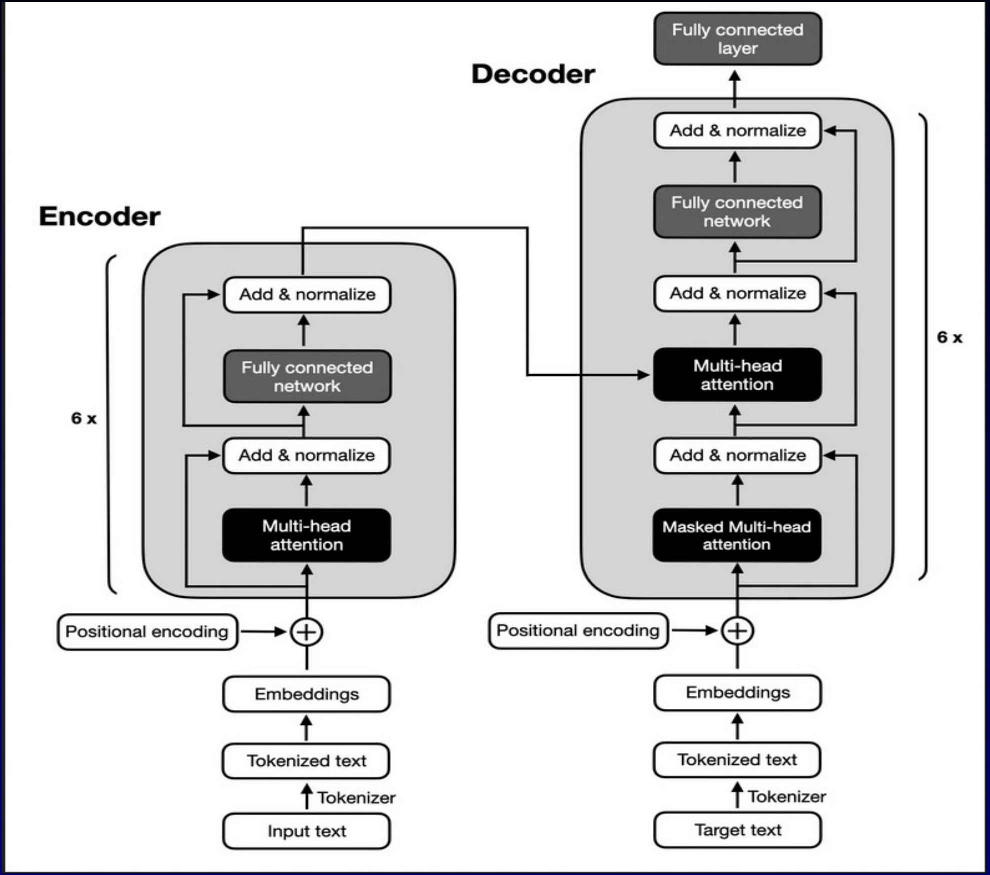
“The cat drank the milk because it was hungry”

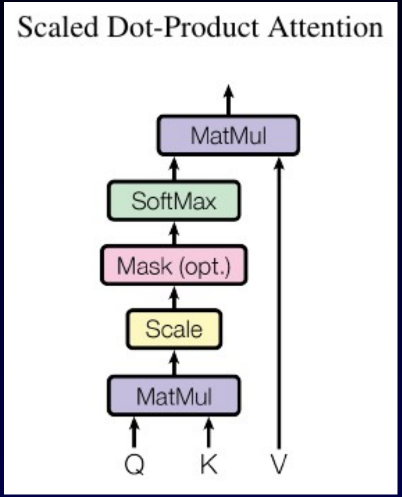
“The cat drank the milk because it was sweet”

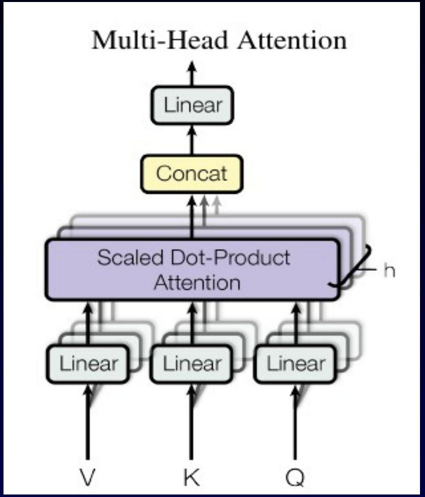


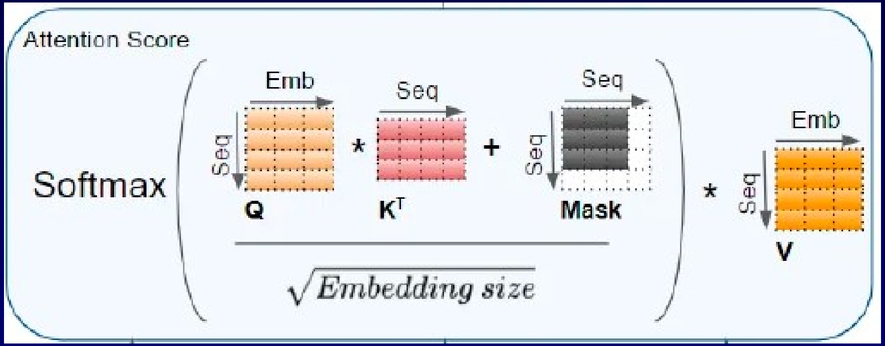
K. Doshi

Transformer
(GPT=Generative Pre-trained Transformer)

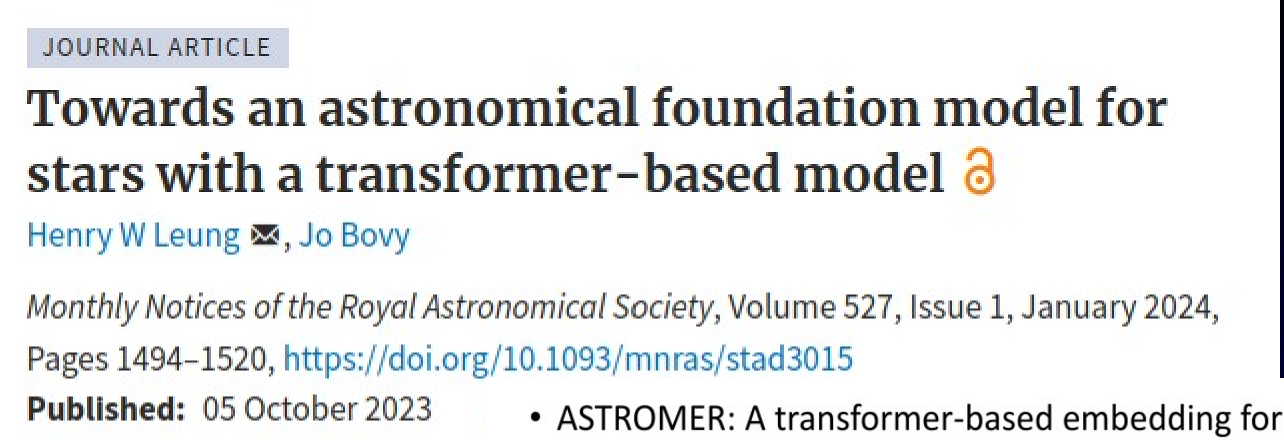








Transformers in Astronomy



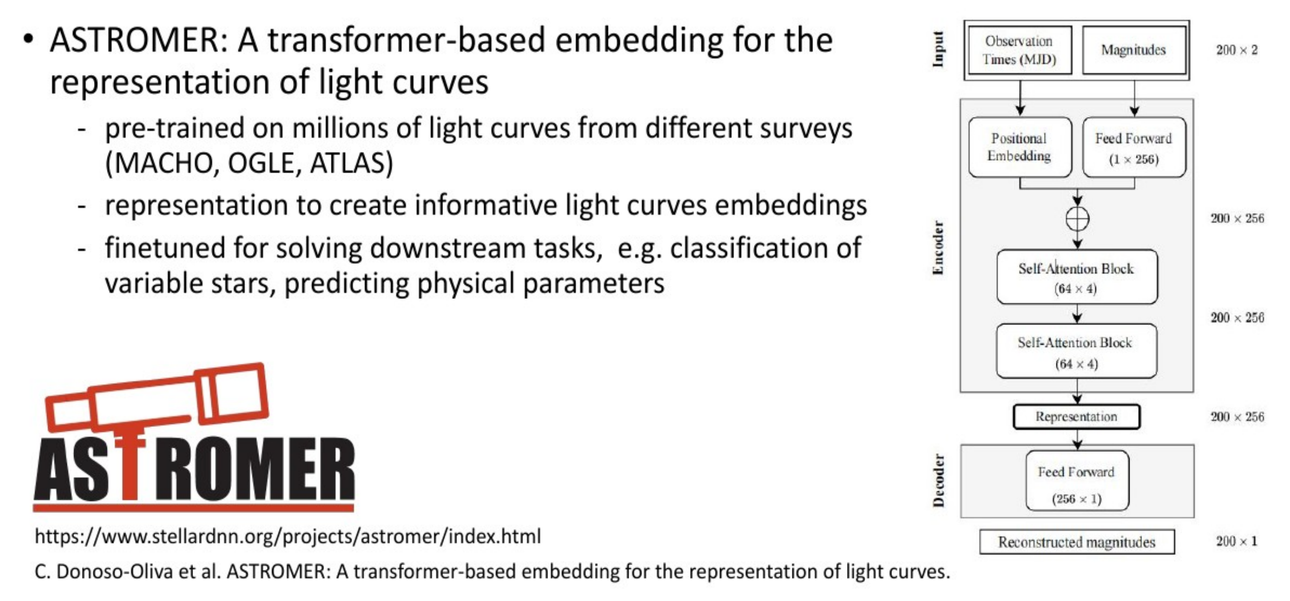
JOURNAL ARTICLE

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



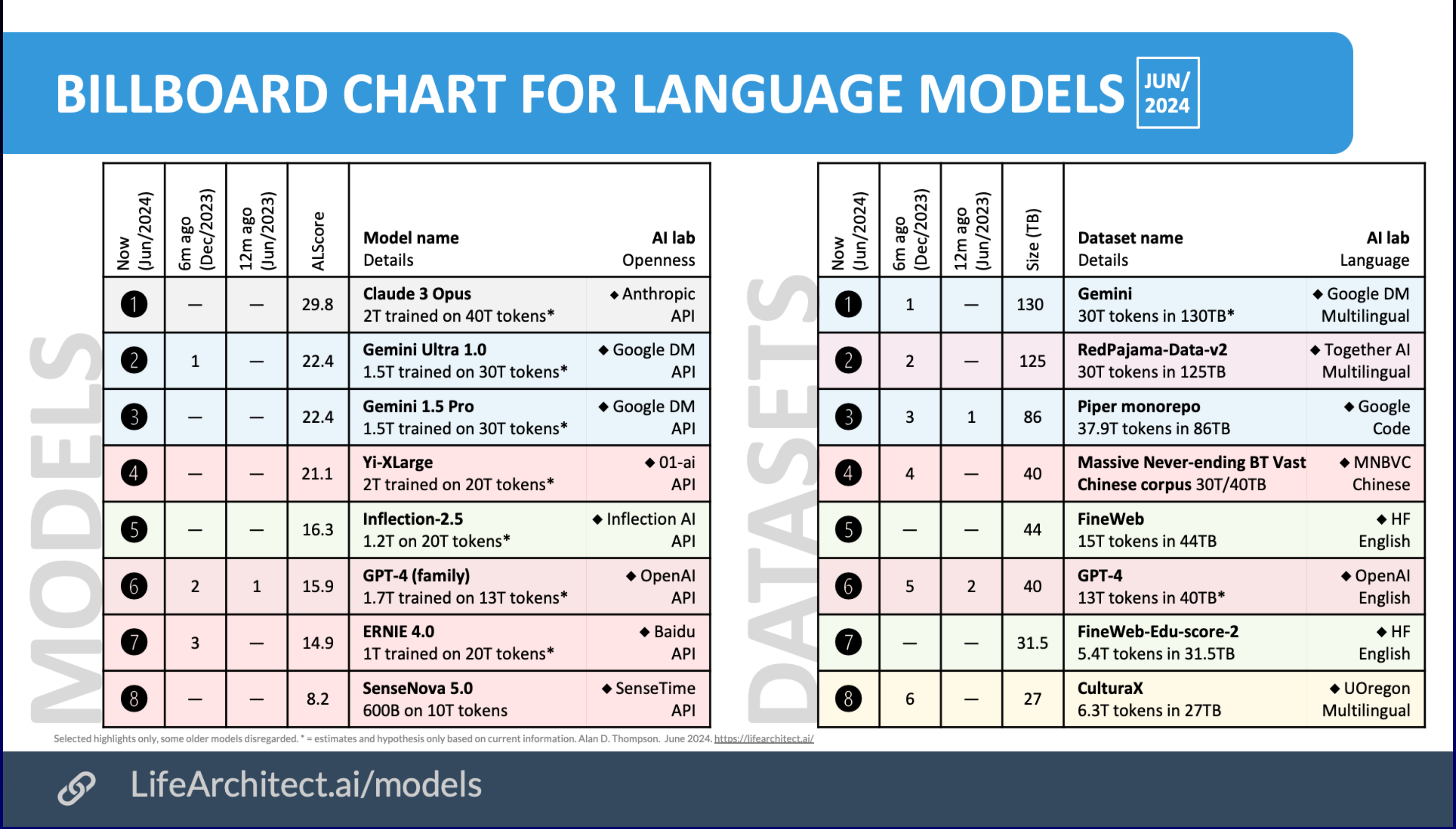
- ASTROMER: A transformer-based embedding for the representation of light curves
  - pre-trained on millions of light curves from different surveys (MACHO, OGLE, ATLAS)
  - representation to create informative light curves embeddings
  - finetuned for solving downstream tasks, e.g. classification of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

ASTROMER

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

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

LLMs in 2024



BILLBOARD CHART FOR LANGUAGE MODELS JUN/ 2024

MODELS

Now (Jun/2024)	6m ago (Dec/2023)	12m ago (Jun/2023)	ALScore	**Model name** Details	**AI lab** Openness
❶	—	—	29.8	**Claude 3 Opus** 2T trained on 40T tokens*	◆ Anthropic API
❷	1	—	22.4	**Gemini Ultra 1.0** 1.5T trained on 30T tokens*	◆ Google DM API
❸	—	—	22.4	**Gemini 1.5 Pro** 1.5T trained on 30T tokens*	◆ Google DM API
❹	—	—	21.1	**Yi-XLarge** 2T trained on 20T tokens*	◆ 01-ai API
❺	—	—	16.3	**Inflection-2.5** 1.2T on 20T tokens*	◆ Inflection AI API
❻	2	1	15.9	**GPT-4 (family)** 1.7T trained on 13T tokens*	◆ OpenAI API
❼	3	—	14.9	**ERNIE 4.0** 1T trained on 20T tokens*	◆ Baidu API
❽	—	—	8.2	**SenseNova 5.0** 600B on 10T tokens	◆ SenseTime API

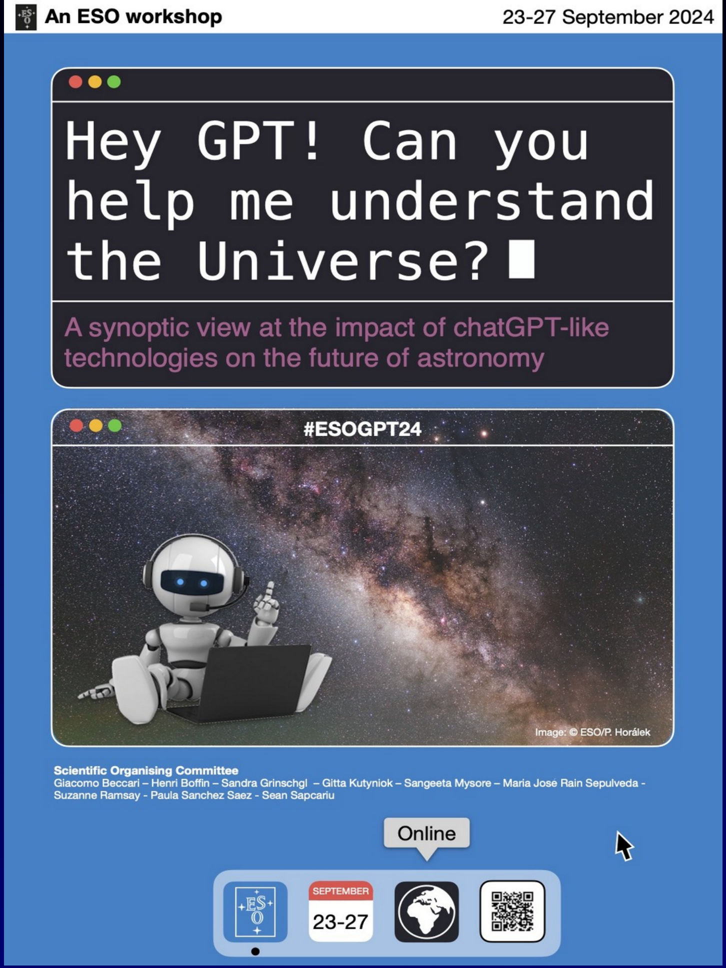
DATASETS

Now (Jun/2024)	6m ago (Dec/2023)	12m ago (Jun/2023)	Size (TB)	**Dataset name** Details	**AI lab** Language
❶	1	—	130	**Gemini** 30T tokens in 130TB*	◆ Google DM Multilingual
❷	2	—	125	**RedPajama-Data-v2** 30T tokens in 125TB	◆ Together AI Multilingual
❸	3	1	86	**Piper monorepo** 37.9T tokens in 86TB	◆ Google Code
❹	4	—	40	**Massive Never-ending BT Vast Chinese corpus** 30T/40TB	◆ MNBVC Chinese
❺	—	—	44	**FineWeb** 15T tokens in 44TB	◆ HF English
❻	5	2	40	**GPT-4** 13T tokens in 40TB*	◆ OpenAI English
❼	—	—	31.5	**FineWeb-Edu-score-2** 5.4T tokens in 31.5TB	◆ HF English
❽	6	—	27	**CulturaX** 6.3T tokens in 27TB	◆ UOregon Multilingual

Selected highlights only, some older models disregarded. * = estimates and hypothesis only based on current information. Alan D. Thompson. June 2024. https://lifearchitect.ai/

LifeArchitect.ai/models

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 Beccari – Henri Boffin – Sandra Grinschgl – Gitta Kutyniok – Sangeeta Mysore – Maria José Rain Sepulveda - Suzanne Ramsay - Paula Sanchez Saez - Sean Sapcariu

ESO Cosmic Duologues
https://www.youtube.com/@ESOCosmicDuologues/featured

AstroLLaMA : 7B pararameters, 300 000 abstracts from ADS

Cosmology – FIT SED

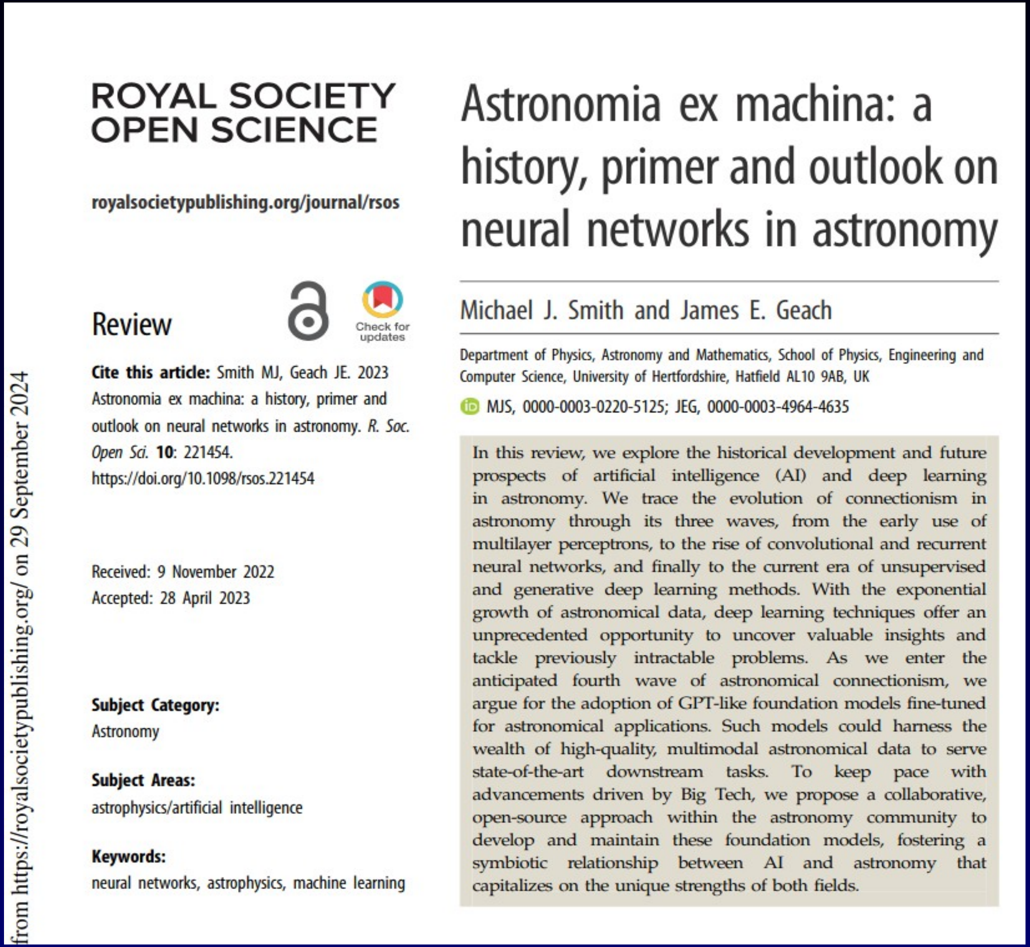
ESO User Man

Proposal rewiews

Project assessments

Ethical problems

Review of LLMs in Astronomy



ROYAL SOCIETY
OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Review

Check for updates

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

from https://royalsocietypublishing.org/ on 29 September 2024

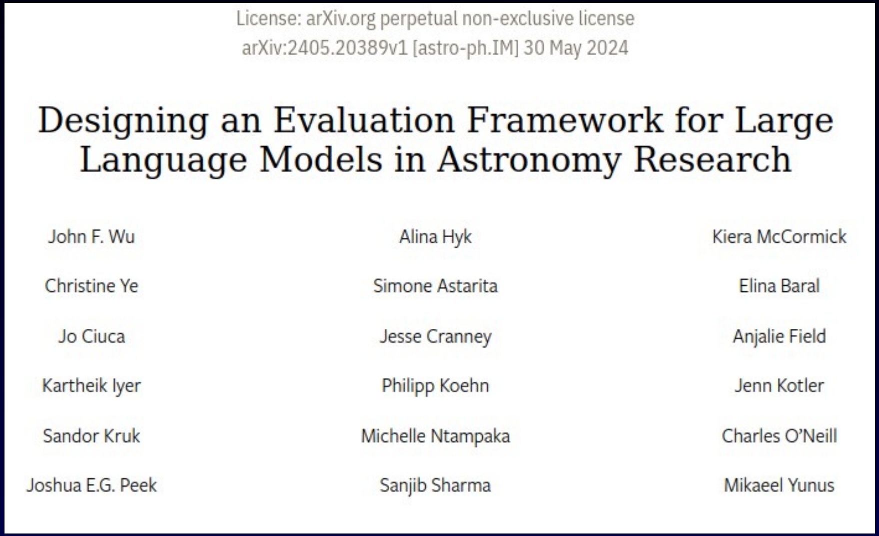
Astronomia ex machina: a history, primer and outlook on neural networks in astronomy

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

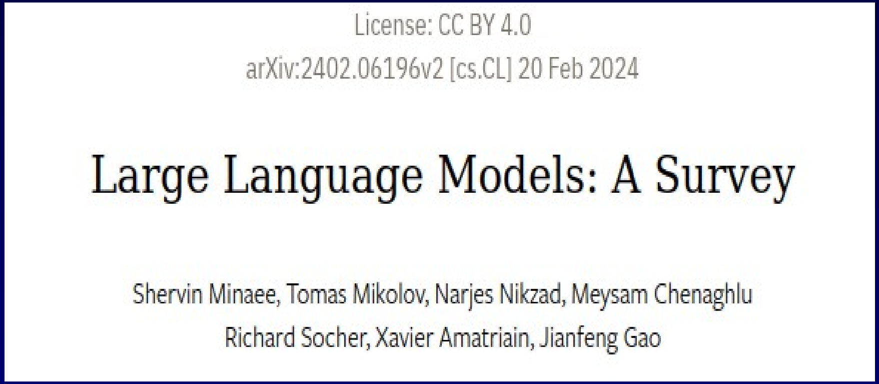
In this review, we explore the historical development and future prospects of artificial intelligence (AI) and deep learning in astronomy. We trace the evolution of connectionism in astronomy through its three waves, from the early use of multilayer perceptrons, to the rise of convolutional and recurrent 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

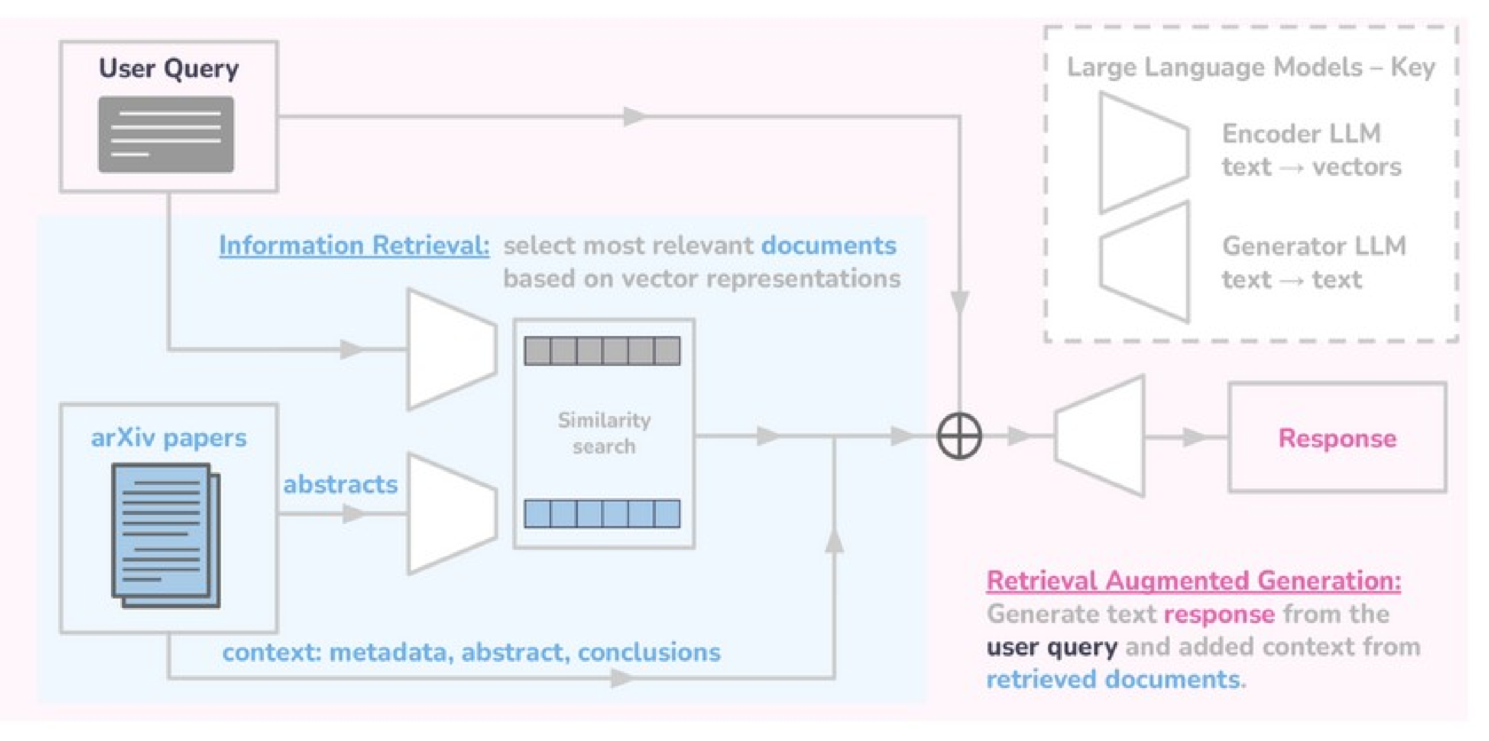


License: CC BY 4.0
arXiv:2402.06196v2 [cs.CL] 20 Feb 2024

Large Language Models: A Survey

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

Retrival Augmented Generation (RAG)


User Query
Large Language Models – Key
Encoder LLM
text → vectors
Generator LLM
text → text
Information Retrieval: select most relevant documents based on vector representations
Similarity search
arXiv papers
abstracts
Response
context: metadata, abstract, conclusions
Retrieval Augmented Generation:
Generate text response from the user query and added context from retrieved documents.

Paper 2020 by Lewis

https://arxiv.org/pdf/2405.20389

Conclusions

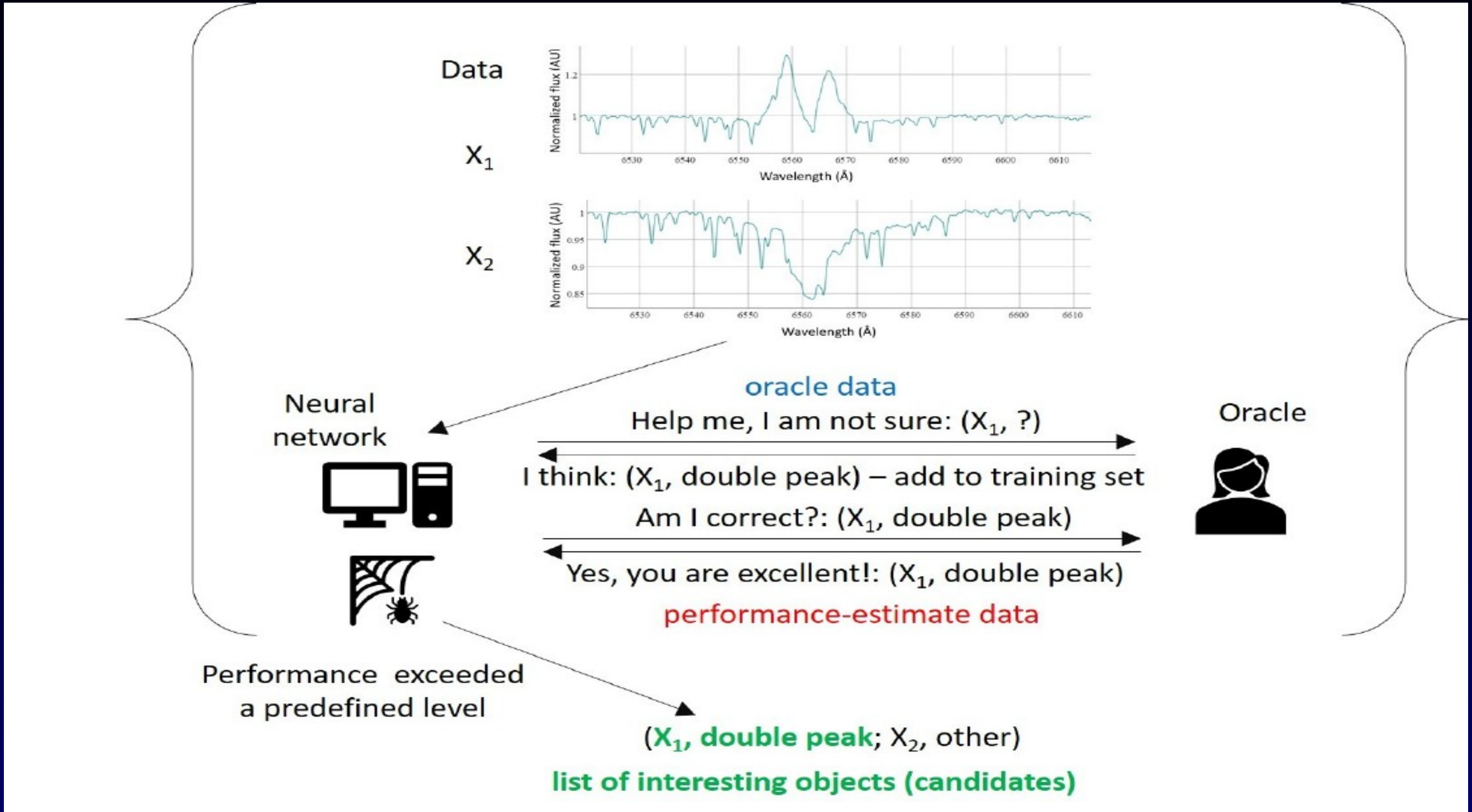
AI already helps us to discover the secrets of the Universe

But the future is ….. uncertain !

TRUST RELIABILITY FAKES

Backup Slides

Active Learning (insufficient labels)



oracle data

Help me, I am not sure: (X_1, ?)

I think: (X_1, double peak) – add to training set

Am I correct?: (X_1, double peak)

Yes, you are excellent!: (X_1, double peak)

performance-estimate data

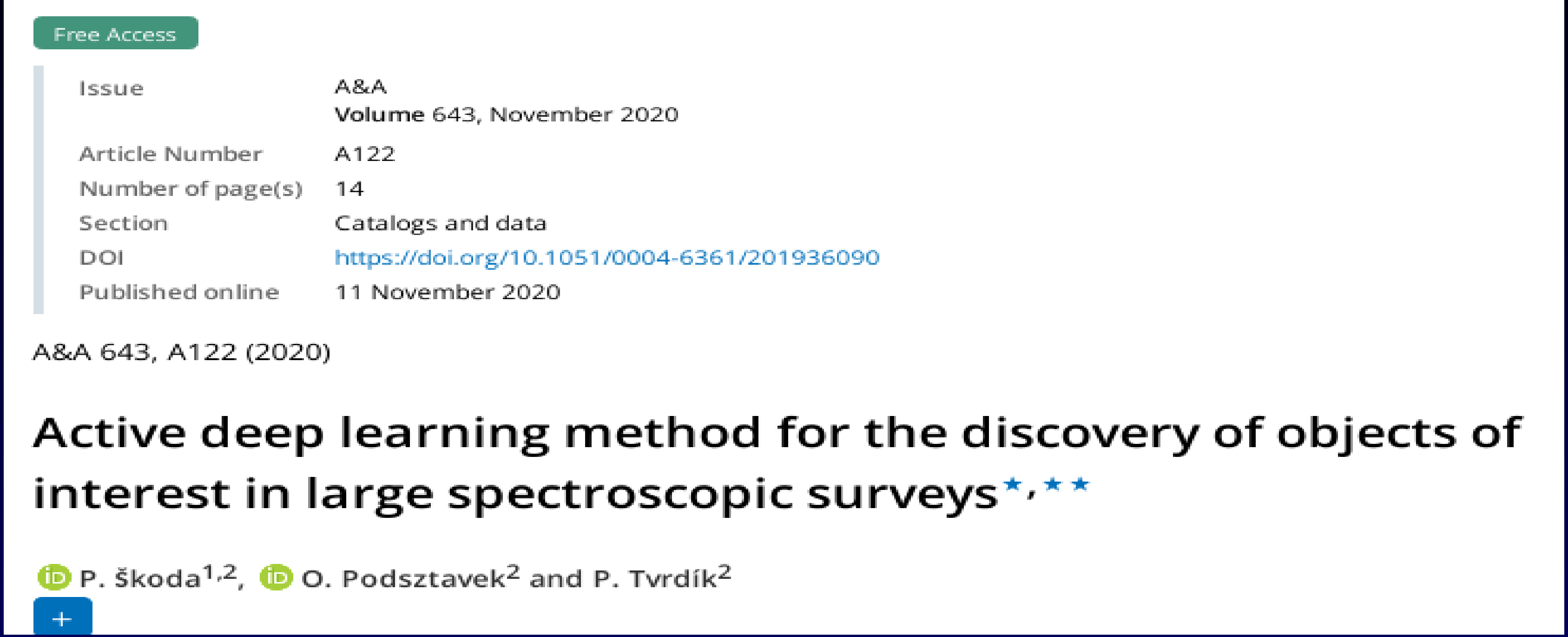
Performance exceeded a predefined level

(**X_1, double peak**; X_2, other)

list of interesting objects (candidates)

Oracle : Human – Machine Interaction

Publication in A&A



Free Access

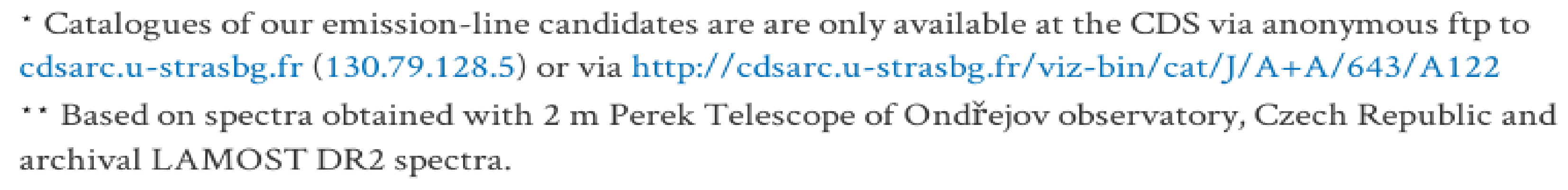
Issue	A&A **Volume** 643, November 2020
Article Number	A122
Number of page(s)	14
Section	Catalogs and data
DOI	https://doi.org/10.1051/0004-6361/201936090
Published online	11 November 2020

A&A 643, A122 (2020)

Active deep learning method for the discovery of objects of interest in large spectroscopic surveys*,**

P. Škoda[1,2], O. Podsztavek[2] and P. Tvrdík[2]

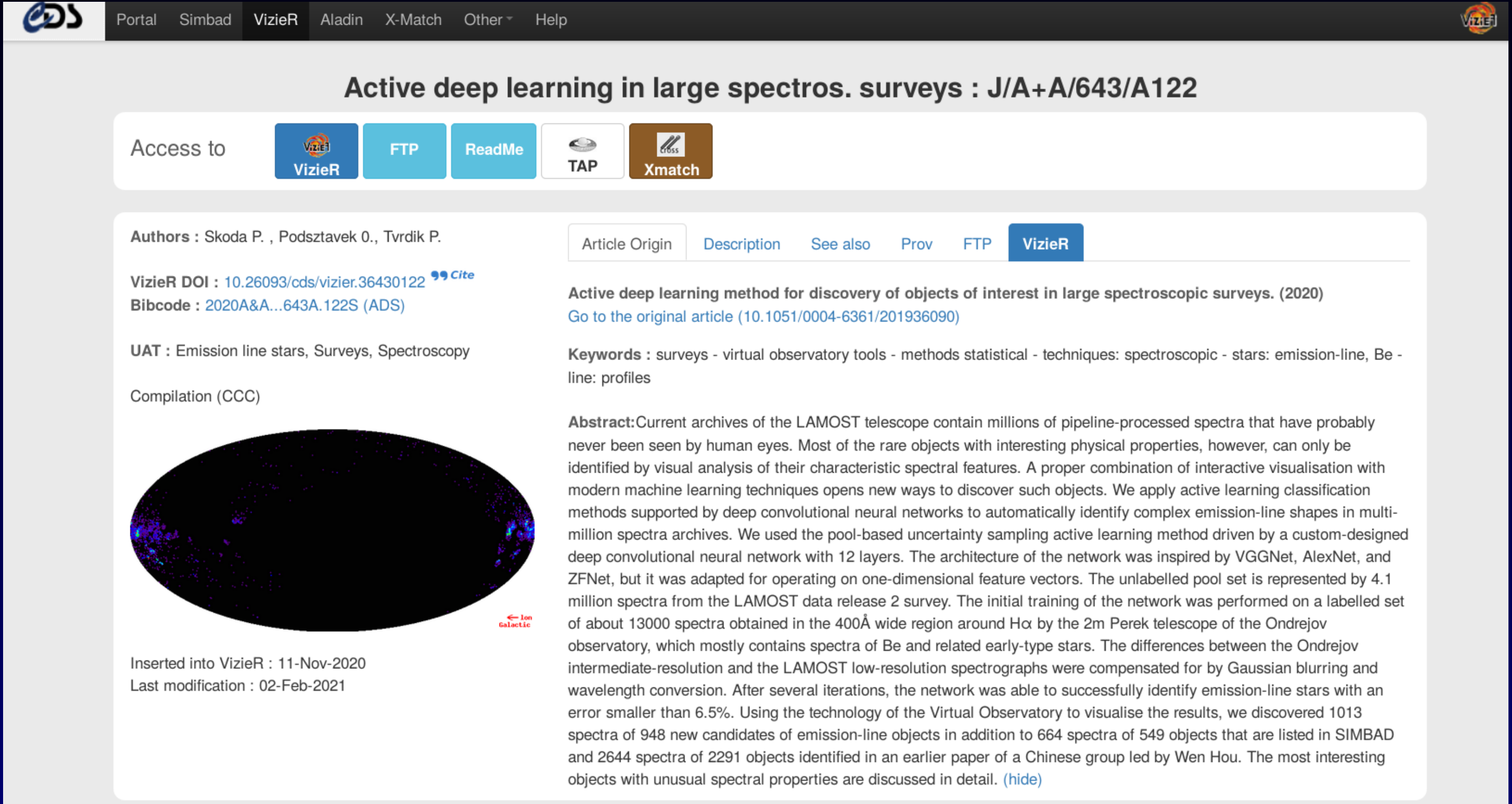
+



[*] Catalogues of our emission-line candidates are are only available at the CDS via anonymous ftp to 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



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

Authors : Skoda P. , Podsztavek 0., Tvrdik P.

VizieR DOI : 10.26093/cds/vizier.36430122 Cite
Bibcode : 2020A&A...643A.122S (ADS)

UAT : Emission line stars, Surveys, Spectroscopy

Compilation (CCC)

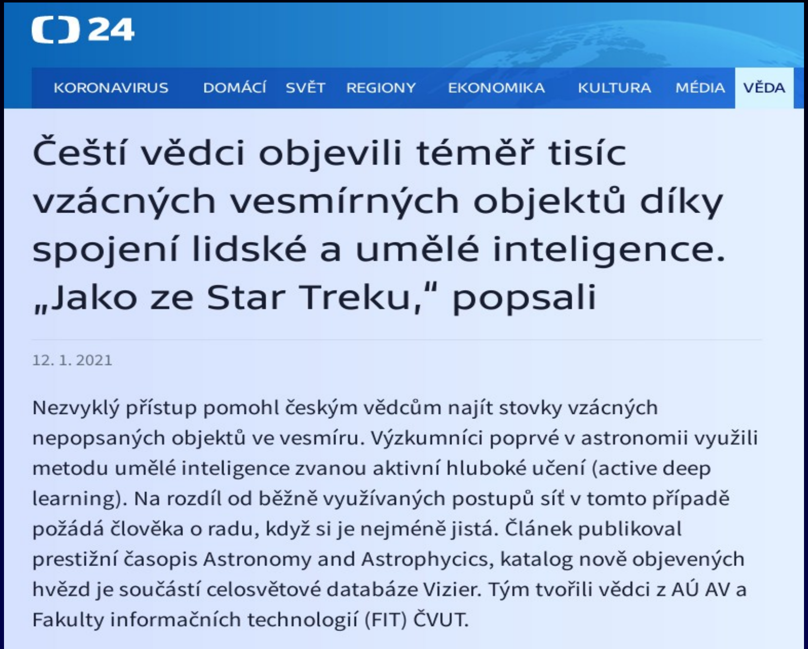
Inserted into VizieR : 11-Nov-2020
Last modification : 02-Feb-2021

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 Hα 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)

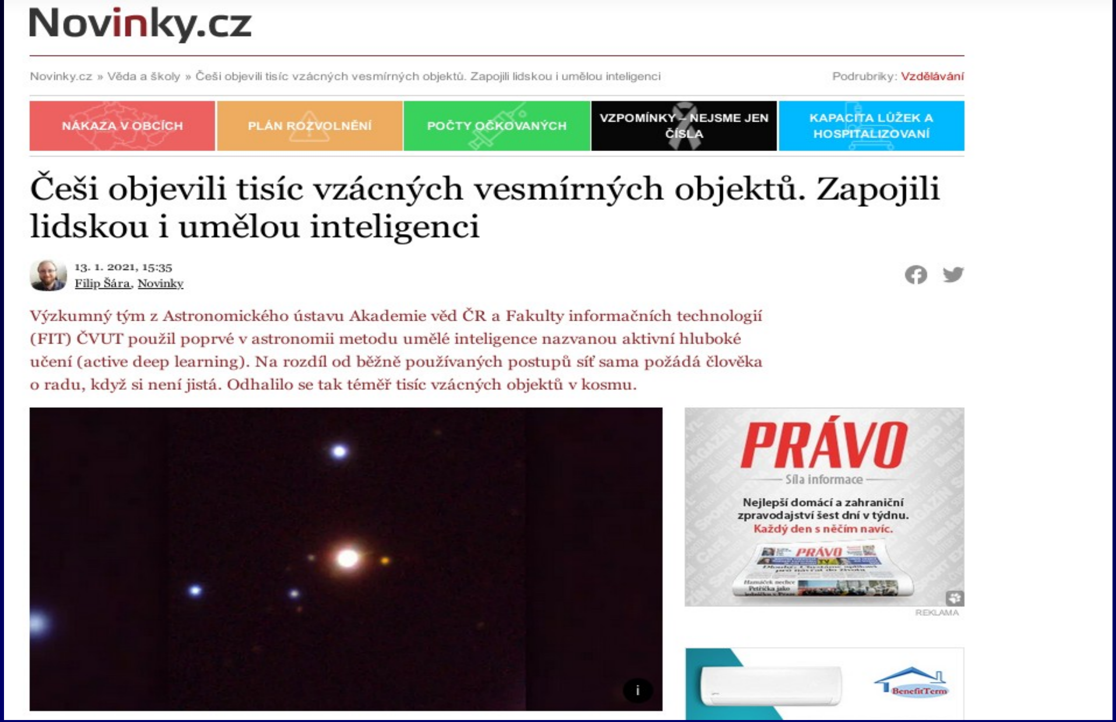
Outreach



Čeští vědci objevili téměř tisíc vzácných vesmírných objektů díky spojení lidské a umělé inteligence. „Jako ze Star Treku,“ popsali

12. 1. 2021

Nezvyklý přístup pomohl českým vědcům najít stovky vzácných nepopsaných objektů ve vesmíru. Výzkumníci poprvé v astronomii využili metodu umělé inteligence zvanou aktivní hluboké učení (active deep learning). Na rozdíl od běžně využívaných postupů síť v tomto případě požádá člověka o radu, když si je nejméně jistá. Článek publikoval prestižní časopis Astronomy and Astrophycics, katalog nově objevených hvězd je součástí celosvětové databáze Vizier. Tým tvořili vědci z AÚ AV a Fakulty informačních technologií (FIT) ČVUT.


Novinky.cz
Novinky.cz » Věda a školy » Češi objevili tisíc vzácných vesmírných objektů. Zapojili lidskou i umělou inteligenci
Podrubriky: Vzdělávání
NÁKAZA V OBCÍCH
PLÁN ROZVOLNĚNÍ
POČTY OČKOVANÝCH
VZPOMÍNKY – NEJSME JEN ČÍSLA
KAPACITA LŮŽEK A HOSPITALIZOVANÍ

Češi objevili tisíc vzácných vesmírných objektů. Zapojili lidskou i umělou inteligenci

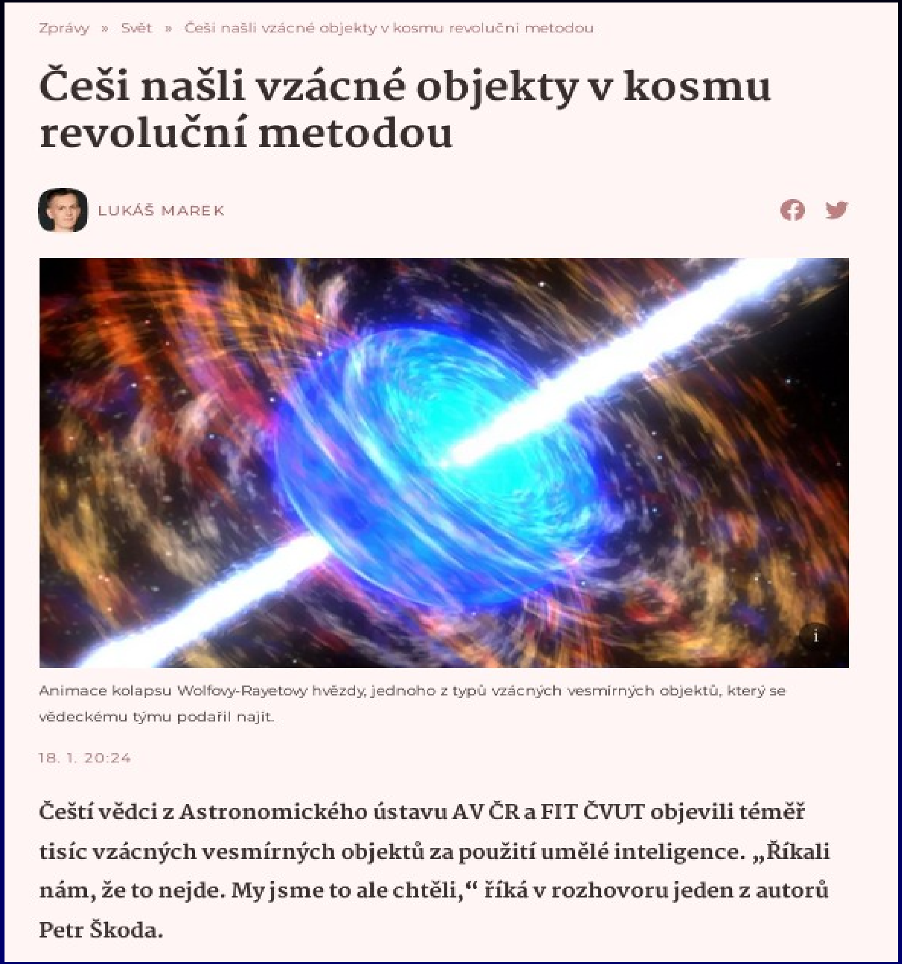
13. 1. 2021, 15:35
Filip Šára, Novinky

Výzkumný tým z Astronomického ústavu Akademie věd ČR a Fakulty informačních technologií (FIT) ČVUT použil poprvé v astronomii metodu umělé inteligence nazvanou aktivní hluboké učení (active deep learning). Na rozdíl od běžně používaných postupů síť sama požádá člověka o radu, když si není jistá. Odhalilo se tak téměř tisíc vzácných objektů v kosmu.

PRÁVO
Síla informace
Nejlepší domácí a zahraniční zpravodajství šest dní v týdnu.
Každý den s něčím navíc.
REKLAMA
BenefitTerm

Outreach

- Seznam zprávy
- 24 comments
- No hating !



Zprávy » Svět » Češi našli vzácné objekty v kosmu revolučni metodou

Češi našli vzácné objekty v kosmu revoluční metodou

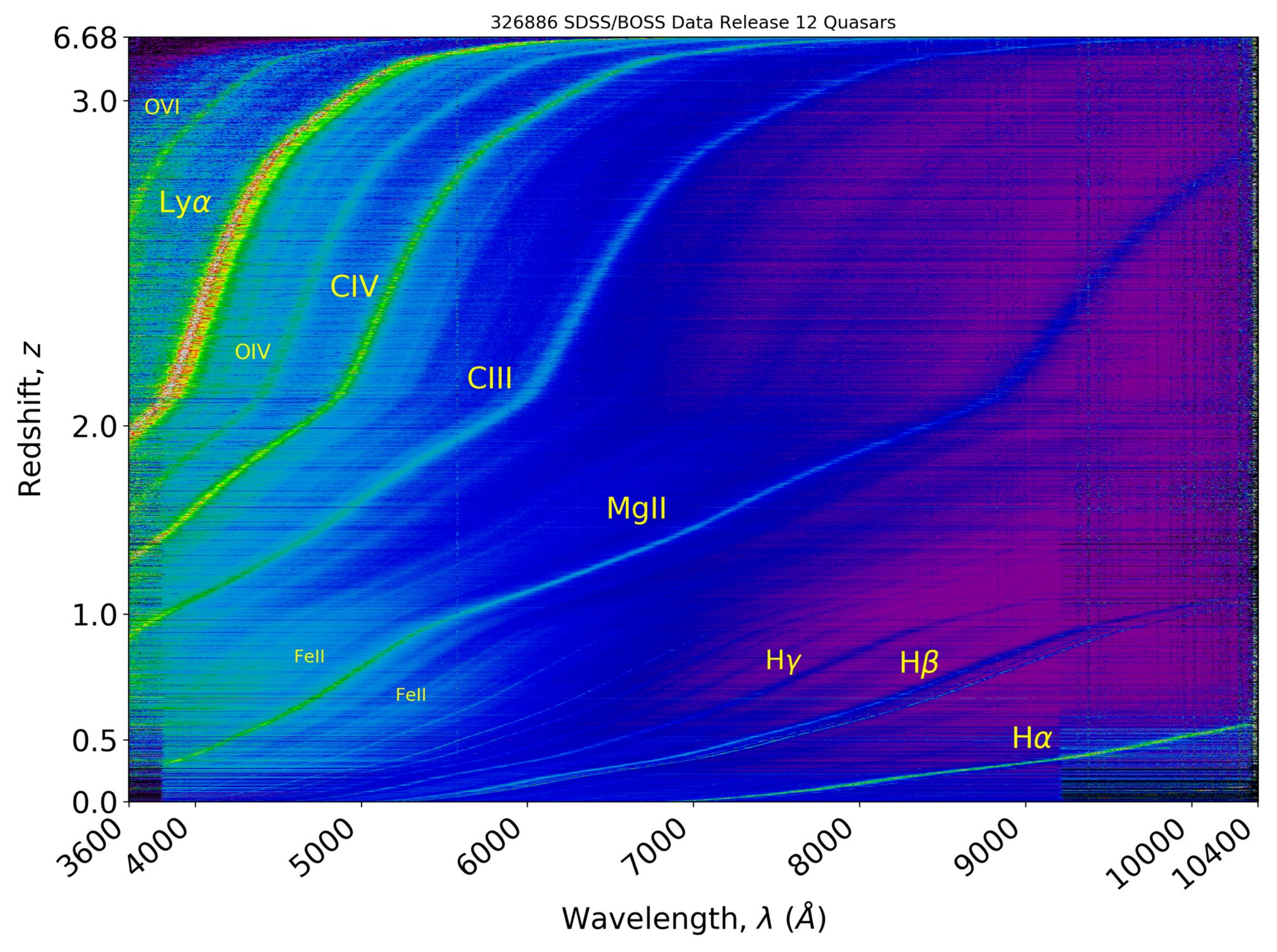
LUKÁŠ MAREK

Animace kolapsu Wolfovy-Rayetovy hvězdy, jednoho z typů vzácných vesmírných objektů, který se vědeckému týmu podařil najít.

18. 1. 20:24

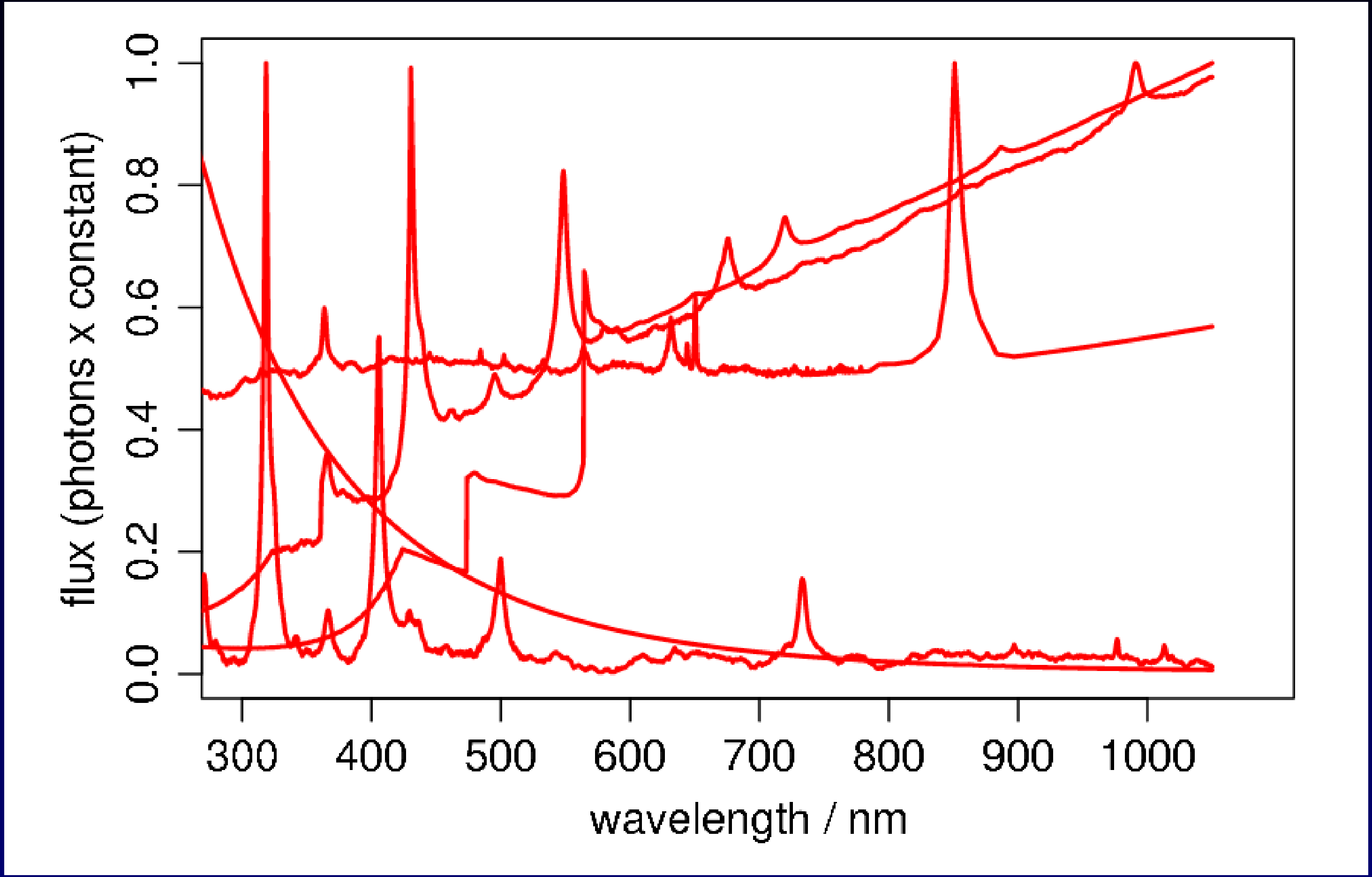
Čeští vědci z Astronomického ústavu AV ČR a FIT ČVUT objevili téměř tisíc vzácných vesmírných objektů za použití umělé inteligence. „Říkali nám, že to nejde. My jsme to ale chtěli," říká v rozhovoru jeden z autorů Petr Škoda.

Redshift

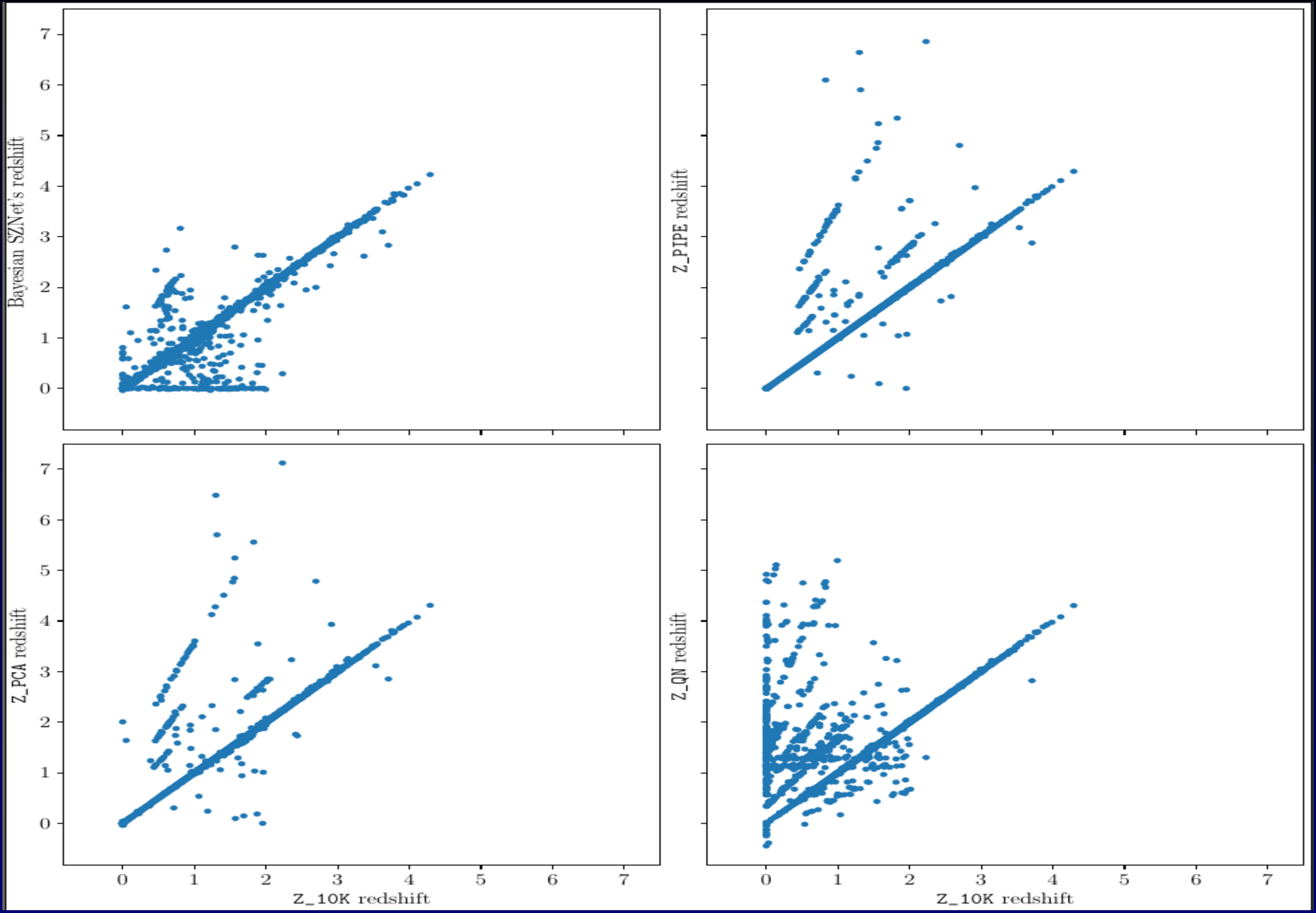


326886 SDSS/BOSS Data Release 12 Quasars
6.68
3.0
2.0
1.0
0.5
0.0
Redshift, z
3600
4000
5000
6000
7000
8000
9000
10000
10400
Wavelength, λ ($\AA$)
OVI
Lyα
CIV
OIV
CIII
MgII
FeII
FeII
Hγ
Hβ
Hα

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

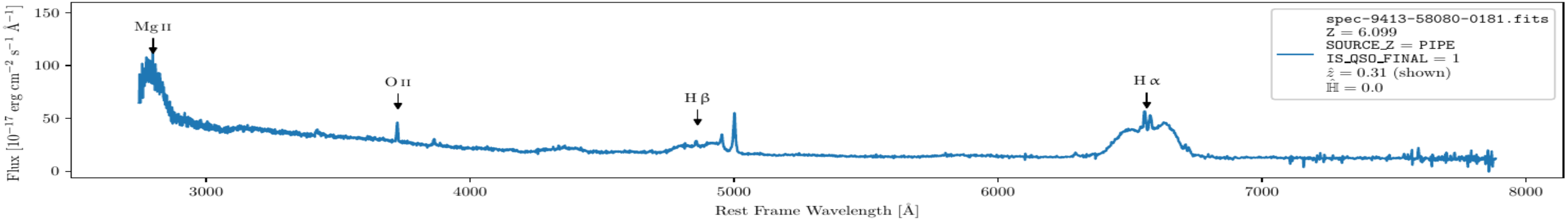


Figure B6. Spectrum with incorrectly high redshift prediction by the pipeline. The Bayesian CNN correctly predicted $\hat{z} = 0.31$ with $\hat{\mathbb{H}} = 0$.

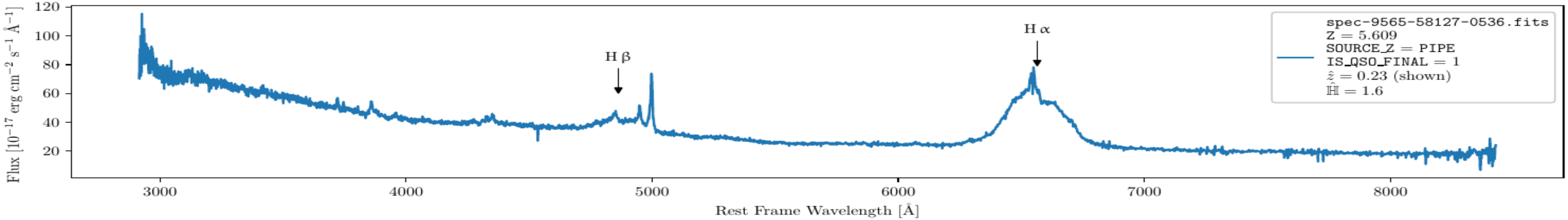
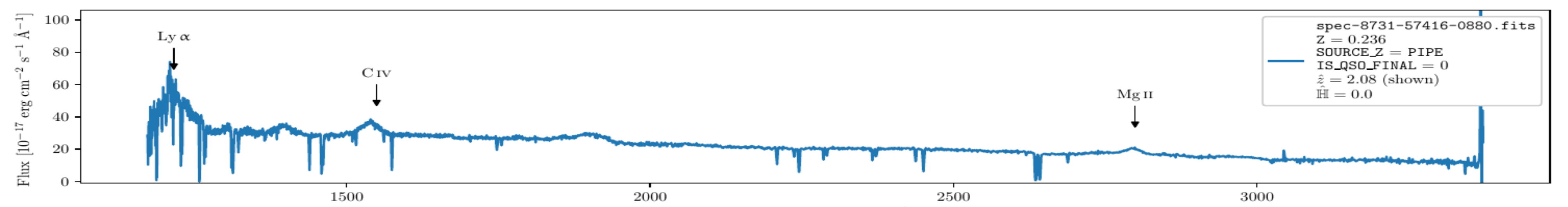
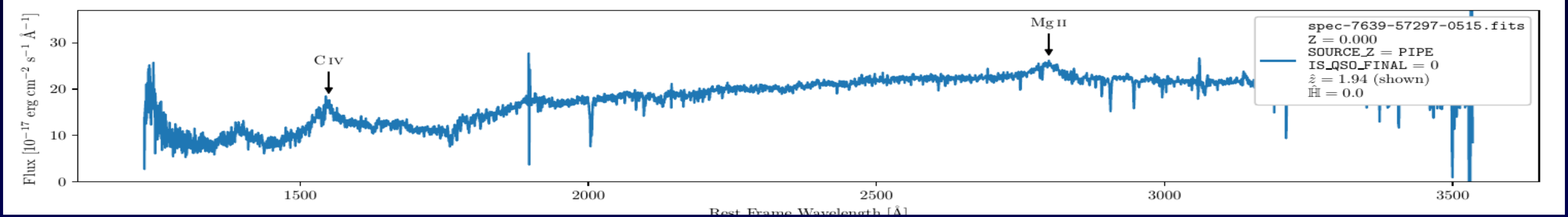
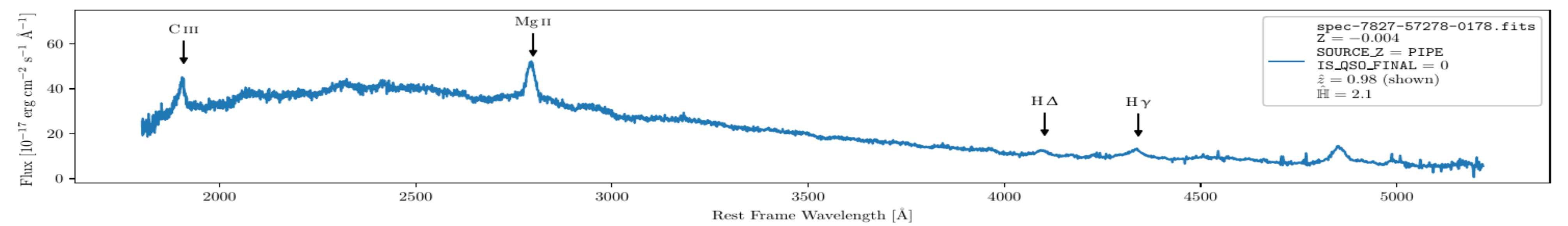


Figure B7. Spectrum with incorrectly high redshift prediction by the pipeline. The Bayesian CNN correctly predicted $\hat{z} = 0.23$ with $\hat{\mathbb{H}} = 1.6$.

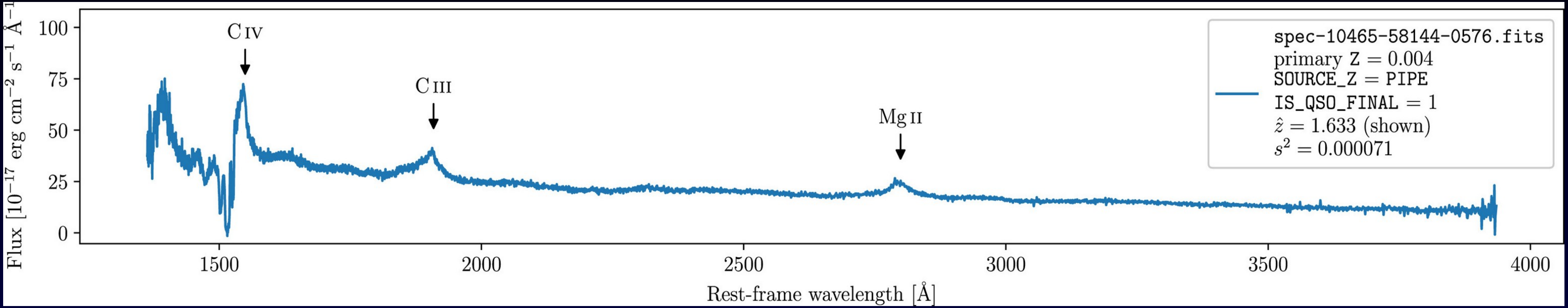
QSOs missing due to SDSS pipeline error

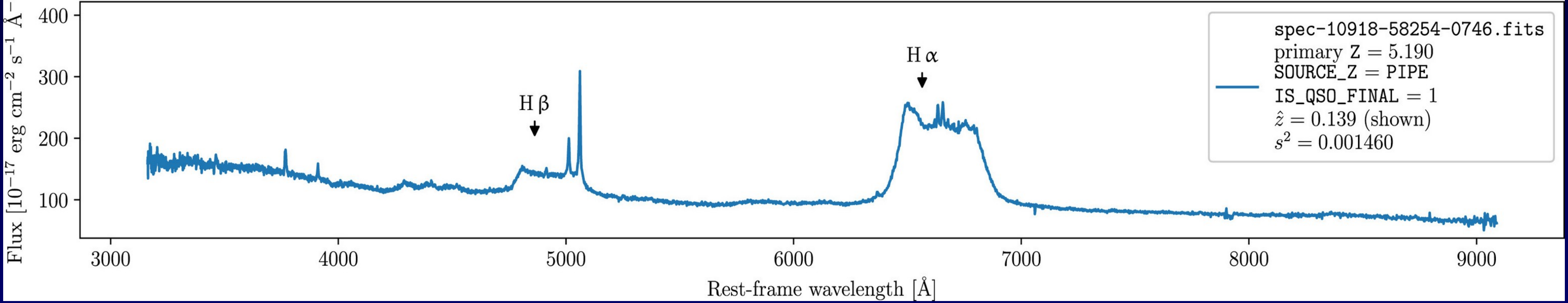




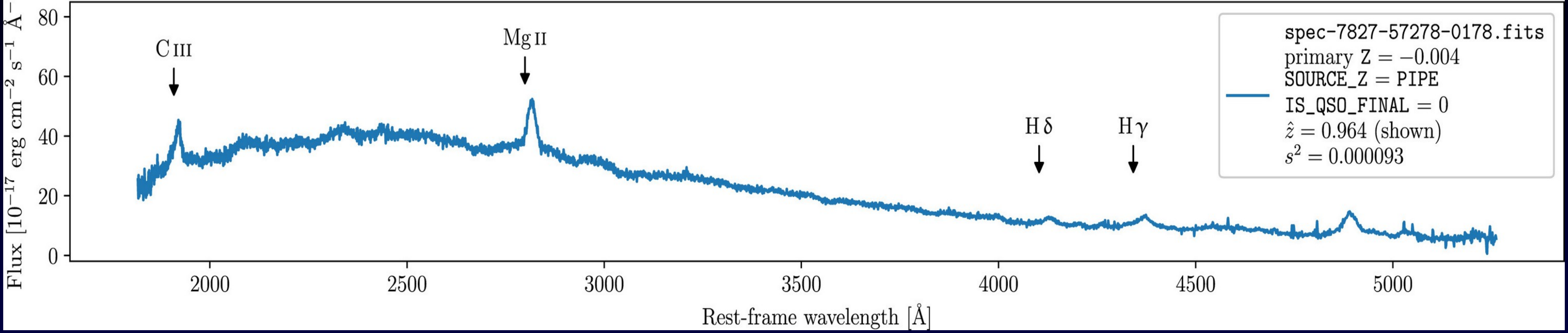


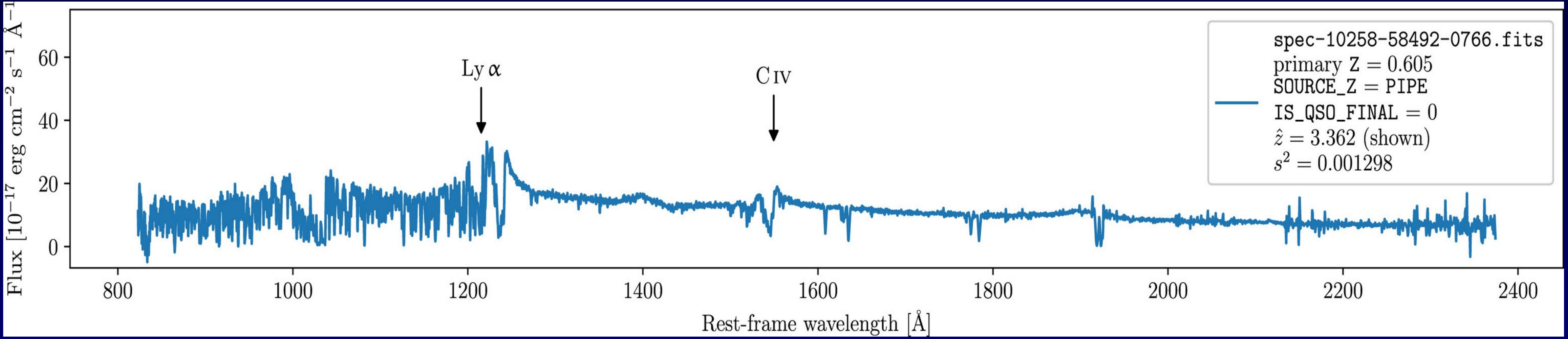
BDN Corrects the SDSS Pipeline



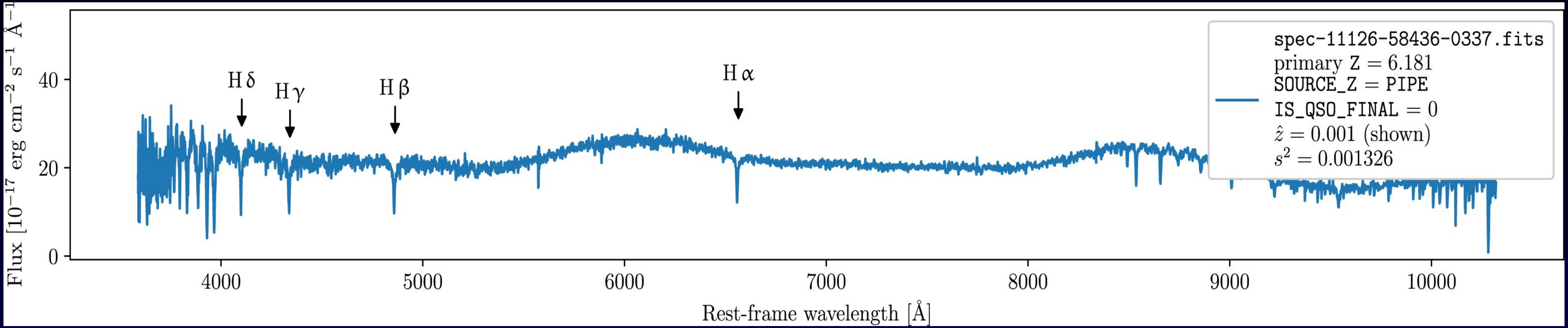


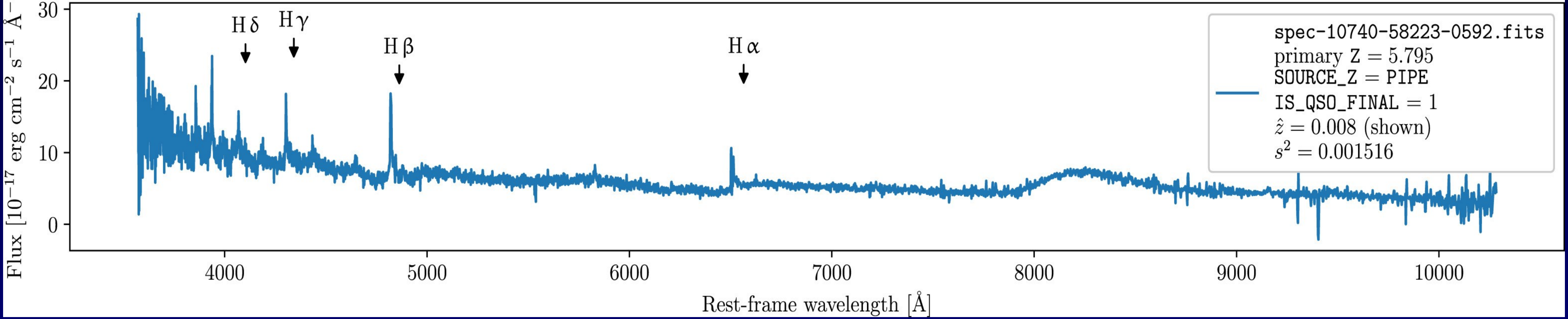
QSOs missing due to SDSS pipeline error





SDSS Predicts QSO
But It Is a Star


H δ
H γ
H β
H α
spec-11126-58436-0337.fits
primary Z = 6.181
SOURCE_Z = PIPE
IS_QSO_FINAL = 0
$\hat{z} = 0.001$ (shown)
$s^2 = 0.001326$
Flux $[10^{-17}$ erg cm^{-2} s^{-1} Å$^{-1}]$
Rest-frame wavelength [Å]


H δ
H γ
H β
H α
spec-10740-58223-0592.fits
primary Z = 5.795
SOURCE_Z = PIPE
IS_QSO_FINAL = 1
$\hat{z} = 0.008$ (shown)
$s^2 = 0.001516$
Flux $[10^{-17}$ erg cm^{-2} s^{-1} Å$^{-1}]$
Rest-frame wavelength [Å]