VO TOOLS AND BASIC DATA MINING CONCEPTS THROUGH PRACTICAL EXAMPLES

Dr Edi Bon
Astronomical Observatory, Beograd

What we need?

- DEFINING THE PROBLEM
- TEORETICAL PREPARATION
- MAKING THE MODEL
- COMPARATION OF RESULTS OF THE MODEL WITH OBSERVATIONS
  - MAKING THE SAMPLE (USE OF DATABASES OR OBSERVING)
  - ANALYSYS OF THE RESULTS
  - PUBLISHING THE RESULTS

WHAT IS VO

- It is a global project
- Large numbrer of computers involved
- Superclcuster for grid computing
- Access through internet

GRID

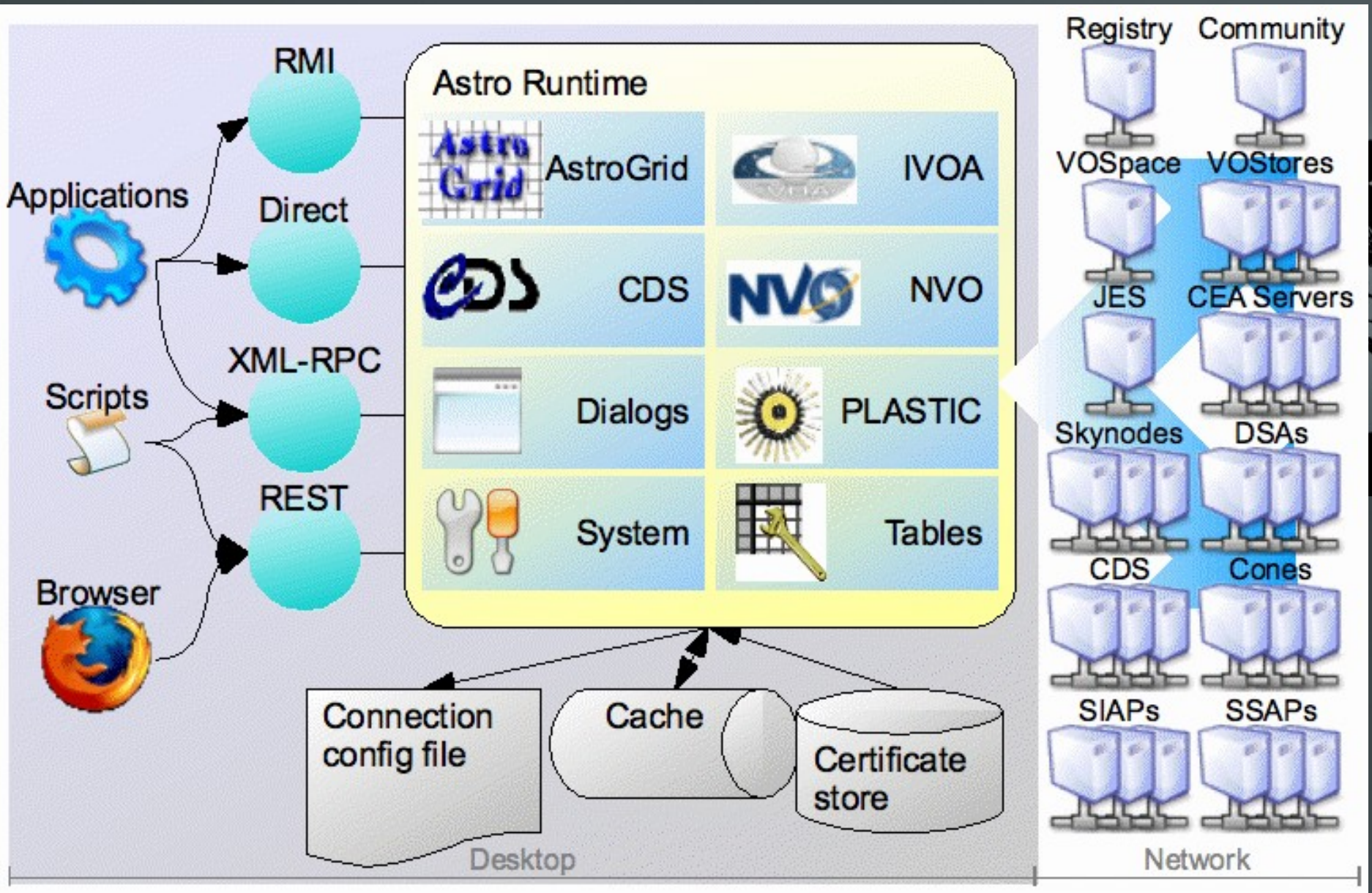
- What are supercomputers today?
- What is grid computing?
- Computer cluster=>

- The idea for the VO came out from grid computing
- GRID – analogy for electrical grid system

15 Member Organizations

http://ivoa.net

VO


RMI
Direct
XML-RPC
REST
Applications
Scripts
Browser
Astro Runtime
AstroGrid
IVOA
CDS
NVO
Dialogs
PLASTIC
System
Tables
Connection config file
Cache
Certificate store
Desktop
Registry
Community
VOSpace
VOStores
JES
CEA Servers
Skynodes
DSAs
CDS
Cones
SIAPs
SSAPs
Network

VO tools (alati)

- Stand alone
  - Aladin
  - Splat
  - VO Plot
  - Topcat
  - Google Earth/sky
  - ...
- Web based
  - Cassjobs
  - Aladin
  - Open sky query
  - ...

How to aproach the problem

- Defining the problem
- Theoretical preparation
- DATA (XML, ASCII, VOT)
  - Defining the criteria for making a sample
  - Making a search (SQL query)
  - x-matching
  - Visualization – analysis – filtering
  - Grouping / clustering
  - Making conclusion from results

Modeling

- Making the model
- Synthesis of artificial data from the model
- Analysis
  - the parameter space and finding the which changes of parameters make a big influence
  - Approximations
  - Reducing the number of parameters
  - Analysis of the correlations between parameters
- Results

Scientific Data Mining in Astronomy

Some key astronomy problems

- Some key astronomy problems that can be addressed with data mining techniques:
  - Cross-Match objects from different catalogues
  - The distance problem (e.g., Photometric Redshift estimators)
  - Star-Galaxy Separation
  - Cosmic-Ray Detection in images
  - Supernova Detection and Classification
  - Morphological Classification (galaxies, AGN, gravitational lenses, ...)
  - Class and Subclass Discovery (brown dwarfs, methane dwarfs, ...)
  - Dimension Reduction = Correlation Discovery
  - Learning Rules for improved classifiers
  - Classification of massive data streams
  - Real-time Classification of Astronomical Events
  - Clustering of massive data collections
  - Novelty, Anomaly, Outlier Detection in massive databases

Some Data Mining Software & Projects

- General data mining software packages:
  - Weka (Java): http://www.cs.waikato.ac.nz/ml/weka/
  - Weka4WS (Grid-enabled): http://grid.deis.unical.it/weka4ws/
  - RapidMiner: http://www.rapidminer.com/
- Astronomy-specific software and/or user clients:
  - VO-Neural: http://voneural.na.infn.it/
  - AstroWeka: http://astroweka.sourceforge.net/
  - OpenSkyQuery: http://www.openskyquery.net/
  - ALADIN: http://aladin.u-strasbg.fr/
  - MIRAGE: http://cm.bell-labs.com/who/tkh/mirage/
  - AstroBox: http://services.china-vo.org/
- Astronomical and/or Scientific Data Mining Projects:
  - GRIST: http://grist.caltech.edu/
  - ClassX: http://heasarc.gsfc.nasa.gov/classx/
  - LCDM: http://dposs.ncsa.uiuc.edu/
  - F-MASS: http://www.itsc.uah.edu/f-mass/
  - NCDM: http://www.ncdm.uic.edu/

Basic Concepts = Key Steps

- The key steps in a data mining project usually invoke and/or follow these basic concepts:
  - Data browse, preview, and selection
  - Data cleaning and preparation
  - Feature selection
  - Data normalization and transformation
  - Similarity/Distance metric selection
  - ... Select the data mining method
  - ... Apply the data mining method
  - ... Gather and analyze data mining results
  - Accuracy estimation
  - Avoiding overfitting

Key Concept for Data Mining: Data Previewing

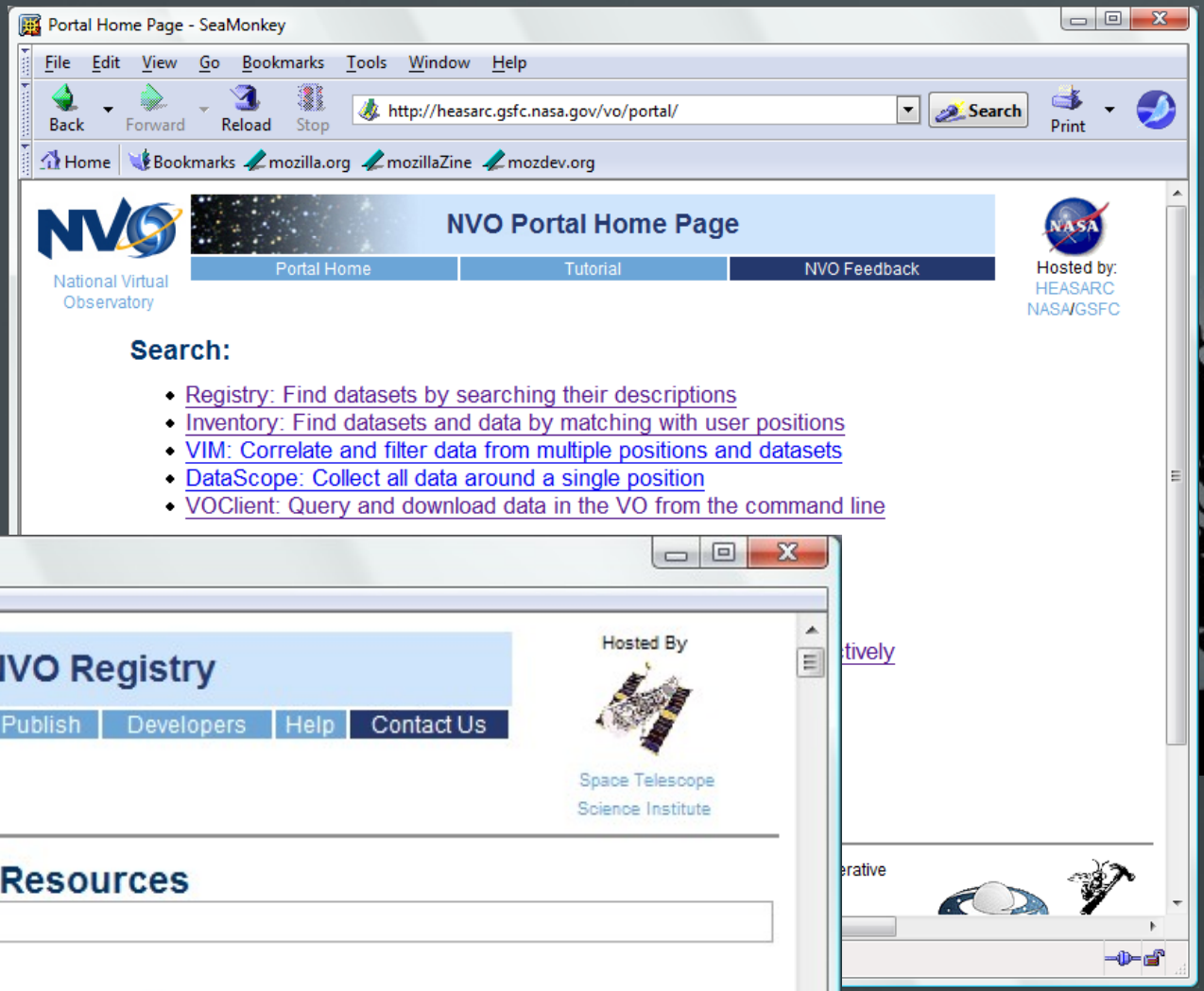
- Data Previewing allows you to get a sense of the good, bad, and ugly parts of the database
- This includes:
  - Histograms of attribute distributions
  - Scatter plots of attribute combinations
  - Max-Min value checks (versus expectations)
  - Summarizations, aggregations (GROUP BY)
  - SELECT UNIQUE values (versus expectations)
  - Checking physical units (and scale factors)
  - External checks (cross-DB comparisons)
  - Verify with input DB

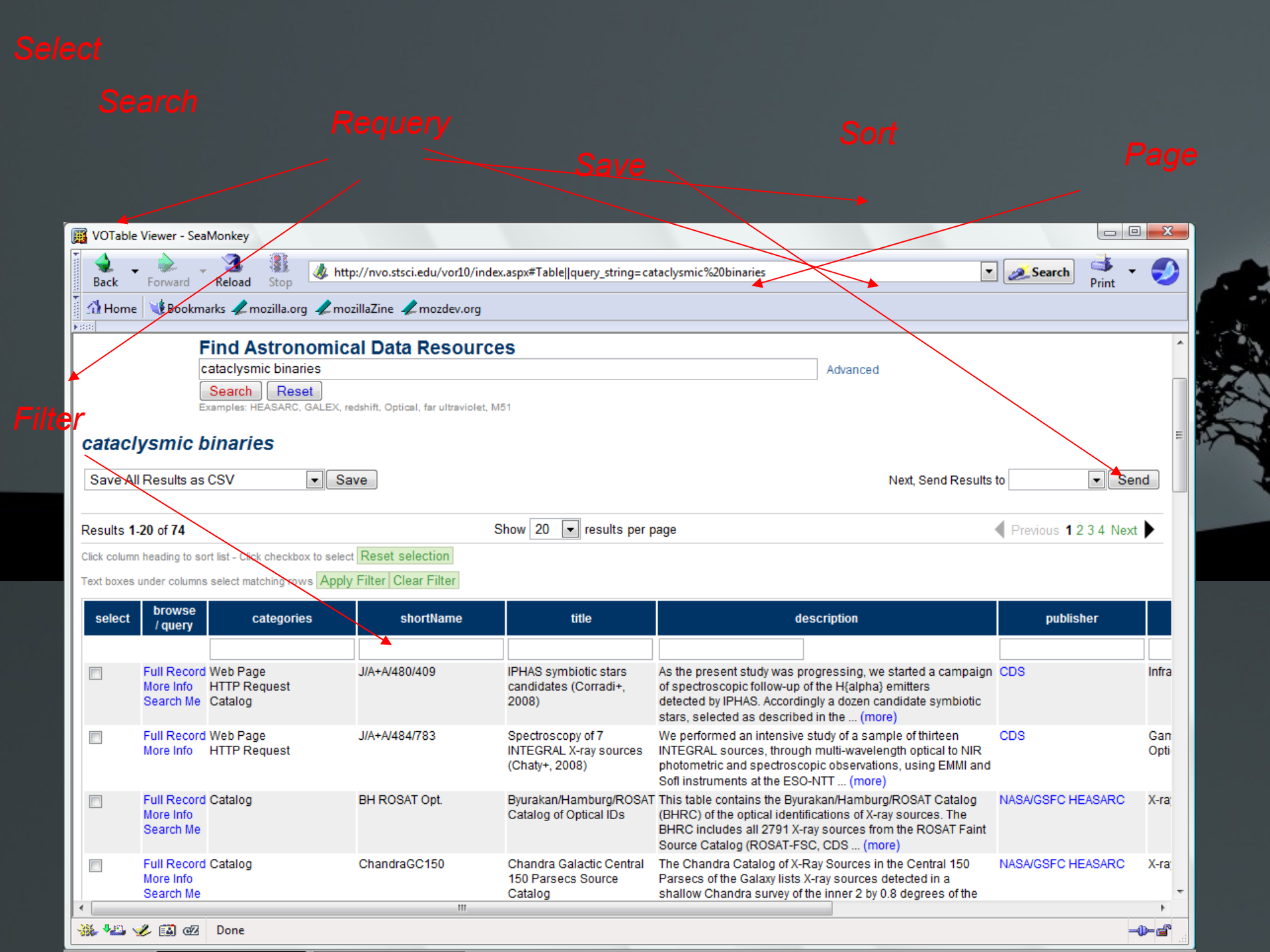
Key Concept for Data Mining: Data Preparation = Cleaning the Data

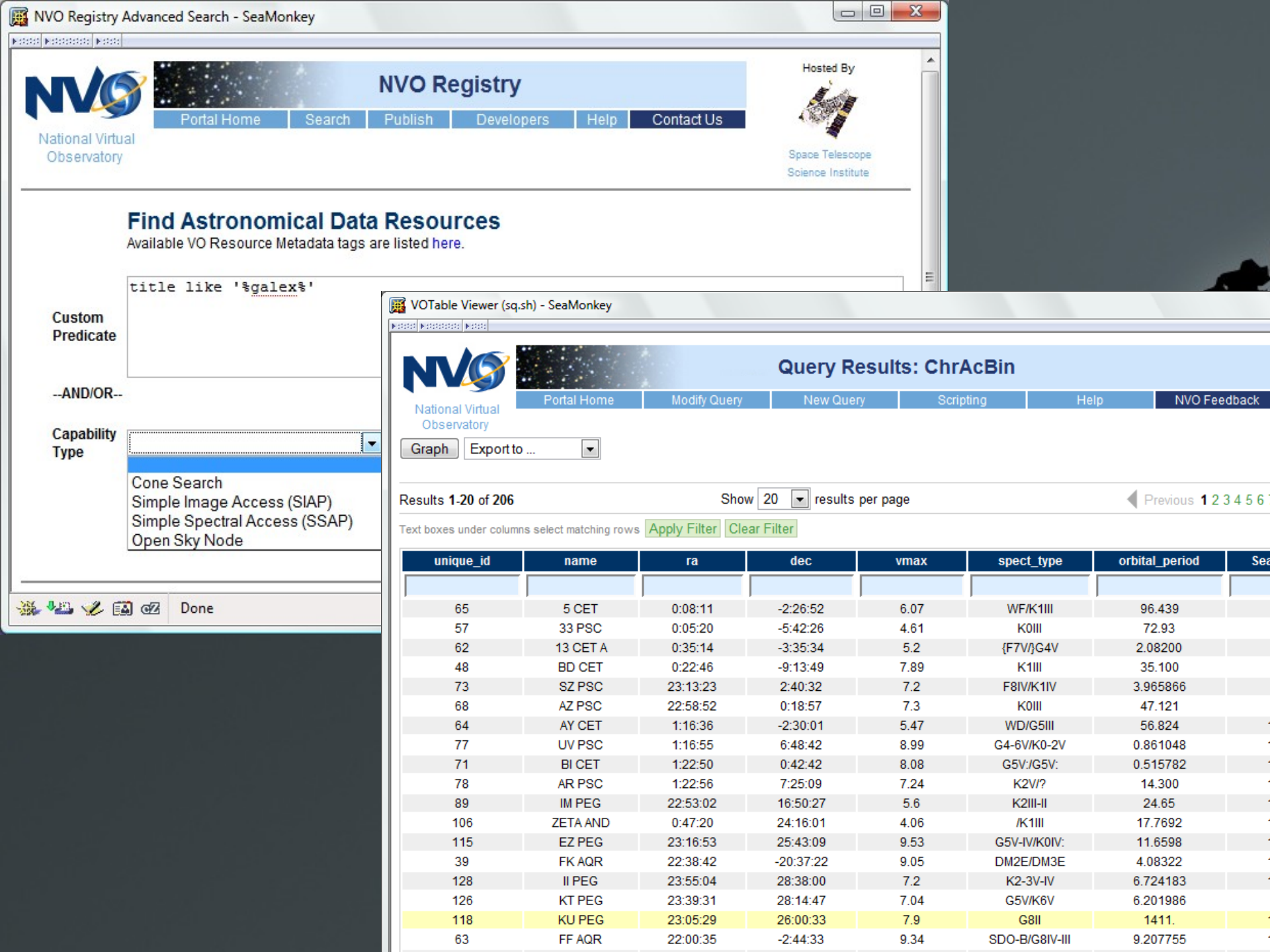
- Data Preparation can take 40-80% (or more) of the effort in a data mining project

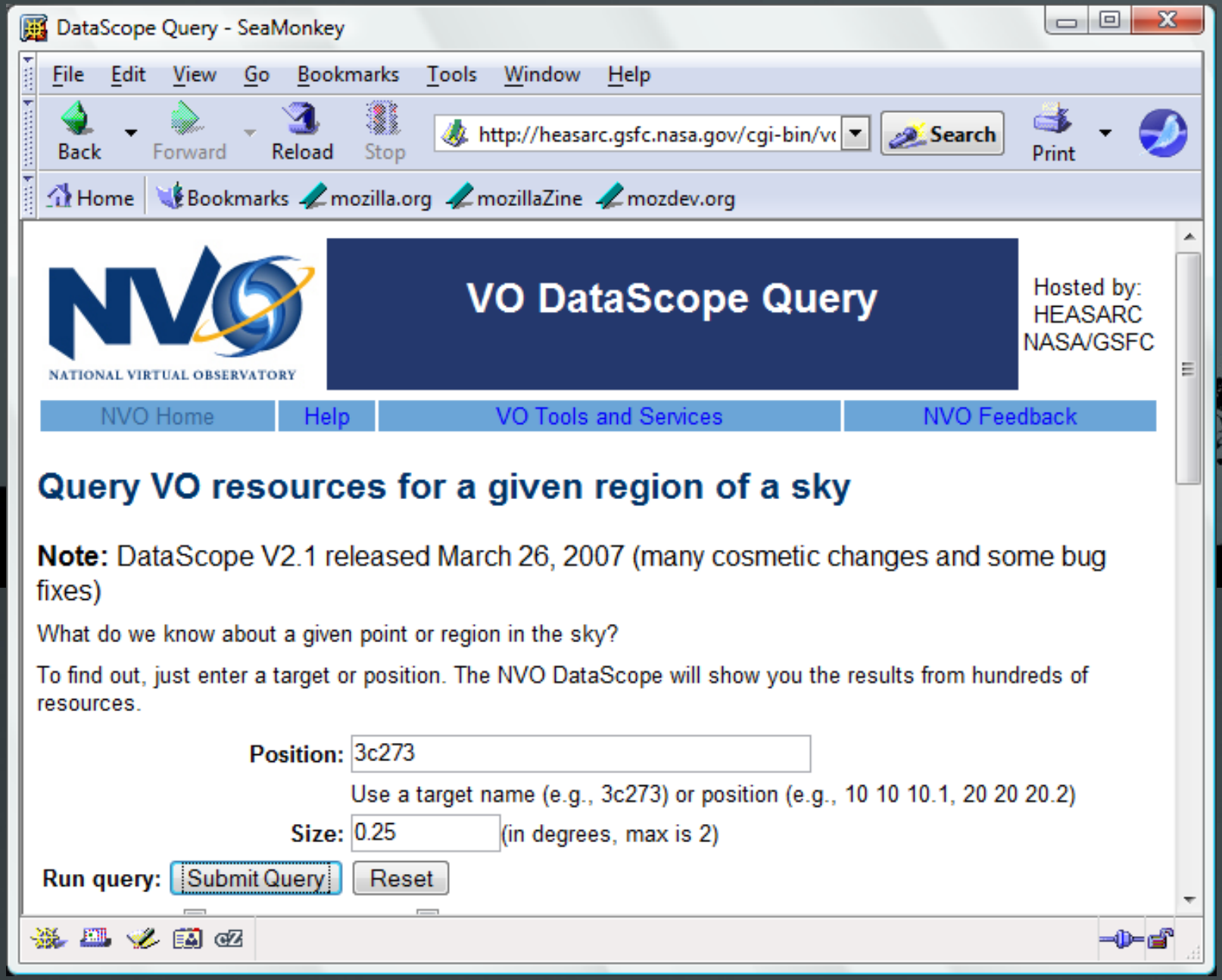
- This includes:
  - Dealing with NULL (missing) values
  - Dealing with errors
  - Dealing with noise
  - Dealing with outliers (unless that is your science!)
  - Transformations: units, scale, projections
  - Data normalization
  - Relevance analysis: Feature Selection
  - Remove redundant attributes
  - Dimensionality Reduction

Finding the data (search)




Select
Search
Requery
Save
Sort
Page
Filter
VOTable Viewer - SeaMonkey
Back
Forward
Reload
Stop
http://nvo.stsci.edu/vor10/index.aspx#Table||query_string=cataclysmic%20binaries
Search
Print
Home
Bookmarks
mozilla.org
mozillaZine
mozdev.org
Find Astronomical Data Resources
cataclysmic binaries
Advanced
Search
Reset
Examples: HEASARC, GALEX, redshift, Optical, far ultraviolet, M51
cataclysmic binaries
Save All Results as CSV
Save
Next, Send Results to
Send
Results 1-20 of 74
Show 20 results per page
Previous 1 2 3 4 Next
Click column heading to sort list - Click checkbox to select
Reset selection
Text boxes under columns select matching rows
Apply Filter
Clear Filter
select
browse / query
categories
shortName
title
description
publisher
Full Record
More Info
Search Me
Web Page
HTTP Request
Catalog
J/A+A/480/409
IPHAS symbiotic stars candidates (Corradi+, 2008)
As the present study was progressing, we started a campaign of spectroscopic follow-up of the H{alpha} emitters detected by IPHAS. Accordingly a dozen candidate symbiotic stars, selected as described in the ... (more)
CDS
Infra
Full Record
More Info
Web Page
HTTP Request
J/A+A/484/783
Spectroscopy of 7 INTEGRAL X-ray sources (Chaty+, 2008)
We performed an intensive study of a sample of thirteen INTEGRAL sources, through multi-wavelength optical to NIR photometric and spectroscopic observations, using EMMI and SofI instruments at the ESO-NTT ... (more)
CDS
Gan
Opti
Full Record
More Info
Search Me
Catalog
BH ROSAT Opt.
Byurakan/Hamburg/ROSAT Catalog of Optical IDs
This table contains the Byurakan/Hamburg/ROSAT Catalog (BHRC) of the optical identifications of X-ray sources. The BHRC includes all 2791 X-ray sources from the ROSAT Faint Source Catalog (ROSAT-FSC, CDS ... (more)
NASA/GSFC HEASARC
X-ra
Full Record
More Info
Search Me
Catalog
ChandraGC150
Chandra Galactic Central 150 Parsecs Source Catalog
The Chandra Catalog of X-Ray Sources in the Central 150 Parsecs of the Galaxy lists X-ray sources detected in a shallow Chandra survey of the inner 2 by 0.8 degrees of the
NASA/GSFC HEASARC
X-ra
Done


NVO Registry Advanced Search - SeaMonkey
NVO Registry
National Virtual Observatory
Portal Home
Search
Publish
Developers
Help
Contact Us
Hosted By
Space Telescope Science Institute
Find Astronomical Data Resources
Available VO Resource Metadata tags are listed here.
title like '%galex%'
Custom Predicate
--AND/OR--
Capability Type
Cone Search
Simple Image Access (SIAP)
Simple Spectral Access (SSAP)
Open Sky Node
Done
VOTable Viewer (sq.sh) - SeaMonkey
Query Results: ChrAcBin
National Virtual Observatory
Portal Home
Modify Query
New Query
Scripting
Help
NVO Feedback
Graph
Export to ...
Results 1-20 of 206
Show 20 results per page
Previous 1 2 3 4 5 6
Text boxes under columns select matching rows
Apply Filter
Clear Filter
unique_id
name
ra
dec
vmax
spect_type
orbital_period
65 5 CET 0:08:11 -2:26:52 6.07 WF/K1III 96.439
57 33 PSC 0:05:20 -5:42:26 4.61 K0III 72.93
62 13 CET A 0:35:14 -3:35:34 5.2 {F7V/}G4V 2.08200
48 BD CET 0:22:46 -9:13:49 7.89 K1III 35.100
73 SZ PSC 23:13:23 2:40:32 7.2 F8IV/K1IV 3.965866
68 AZ PSC 22:58:52 0:18:57 7.3 K0III 47.121
64 AY CET 1:16:36 -2:30:01 5.47 WD/G5III 56.824
77 UV PSC 1:16:55 6:48:42 8.99 G4-6V/K0-2V 0.861048
71 BI CET 1:22:50 0:42:42 8.08 G5V:/G5V: 0.515782
78 AR PSC 1:22:56 7:25:09 7.24 K2V/? 14.300
89 IM PEG 22:53:02 16:50:27 5.6 K2III-II 24.65
106 ZETA AND 0:47:20 24:16:01 4.06 /K1III 17.7692
115 EZ PEG 23:16:53 25:43:09 9.53 G5V-IV/K0IV: 11.6598
39 FK AQR 22:38:42 -20:37:22 9.05 DM2E/DM3E 4.08322
128 II PEG 23:55:04 28:38:00 7.2 K2-3V-IV 6.724183
126 KT PEG 23:39:31 28:14:47 7.04 G5V/K6V 6.201986
118 KU PEG 23:05:29 26:00:33 7.9 G8II 1411.
63 FF AQR 22:00:35 -2:44:33 9.34 SDO-B/G8IV-III 9.207755


DataScope Query - SeaMonkey
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://heasarc.gsfc.nasa.gov/cgi-bin/v
Search
Print
Home Bookmarks mozilla.org mozillaZine mozdev.org
NATIONAL VIRTUAL OBSERVATORY
VO DataScope Query
Hosted by:
HEASARC
NASA/GSFC
NVO Home Help VO Tools and Services NVO Feedback
Query VO resources for a given region of a sky
Note: DataScope V2.1 released March 26, 2007 (many cosmetic changes and some bug fixes)
What do we know about a given point or region in the sky?
To find out, just enter a target or position. The NVO DataScope will show you the results from hundreds of resources.
Position: 3c273
Use a target name (e.g., 3c273) or position (e.g., 10 10 10.1, 20 20 20.2)
Size: 0.25 (in degrees, max is 2)
Run query: Submit Query Reset

Aladin v5.0

File Edit Image Catalog Overlay Tool Help

Command ICRS Pixel full

select pixel pan prop zoom del dist draw tag text filter cross rgb assoc cont mglss

Zoom 1x

grid multiview match Search

Aladin Sky Atlas - v5.0

ALADIN is an interactive software sky atlas.
It allows one to visualize digitized images of any part of the sky,
to superimpose entries from astronomical catalogs,
and to interactively access related data and information.

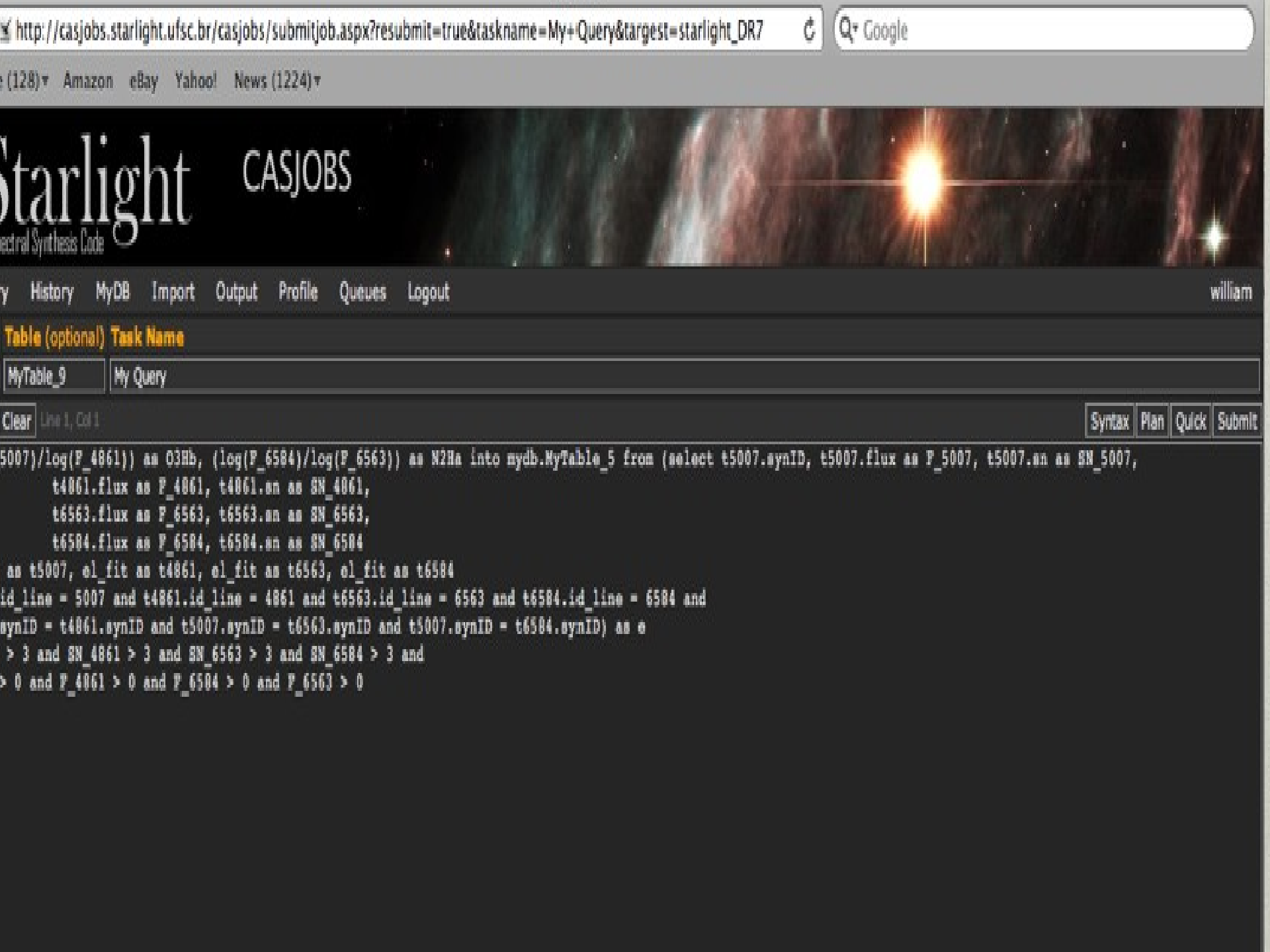
Quick start...

Just type your target in the "Command" field above
(ex: M1 or 13:29:53 +47:11:48)

CENTRE DE DONNÉES ASTRONOMIQUES DE STRASBOURG

Aladin is developed by Pierre Fernique,
Thomas Boch and François Bonnarel.
(c) ULP/CNRS 1999-2008

0 sel / 0 src 1Mb


http://casjobs.starlight.ufsc.br/casjobs/submitjob.aspx?resubmit=true&taskname=My+Query&target=starlight_DR7
Google
(128)
Amazon
eBay
Yahoo!
News (1224)
Starlight
Spectral Synthesis Code
CASJOBS
History
MyDB
Import
Output
Profile
Queues
Logout
william
Table (optional)
Task Name
MyTable_9
My Query
Clear
Line 1, Col 1
Syntax
Plan
Quick
Submit
5007)/log(P_4861)) as O3Hb, (log(P_6584)/log(P_6563)) as N2Ha into mydb.MyTable_5 from (select t5007.synID, t5007.flux as P_5007, t5007.sn as SN_5007,
t4861.flux as P_4861, t4861.sn as SN_4861,
t6563.flux as P_6563, t6563.sn as SN_6563,
t6584.flux as P_6584, t6584.sn as SN_6584
as t5007, el_fit as t4861, el_fit as t6563, el_fit as t6584
id_line = 5007 and t4861.id_line = 4861 and t6563.id_line = 6563 and t6584.id_line = 6584 and
synID = t4861.synID and t5007.synID = t6563.synID and t5007.synID = t6584.synID) as e
> 3 and SN_4861 > 3 and SN_6563 > 3 and SN_6584 > 3 and
> 0 and P_4861 > 0 and P_6584 > 0 and P_6563 > 0

Practical example using VO

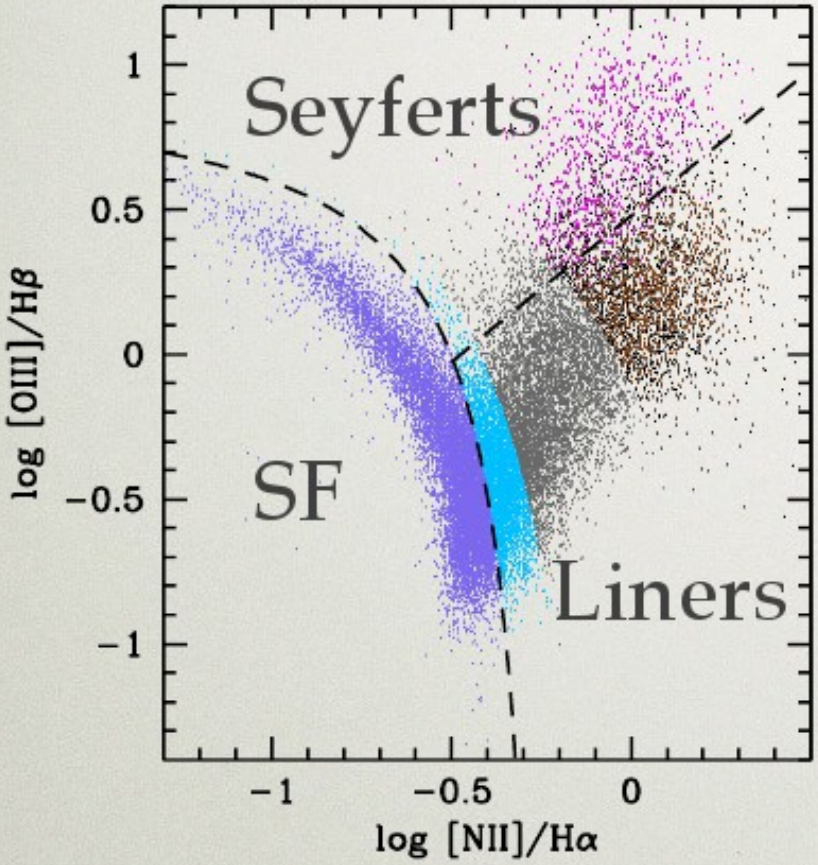
- Probelem: OIII (**4959,5008**) emission line ratio - over type and age of galaxy

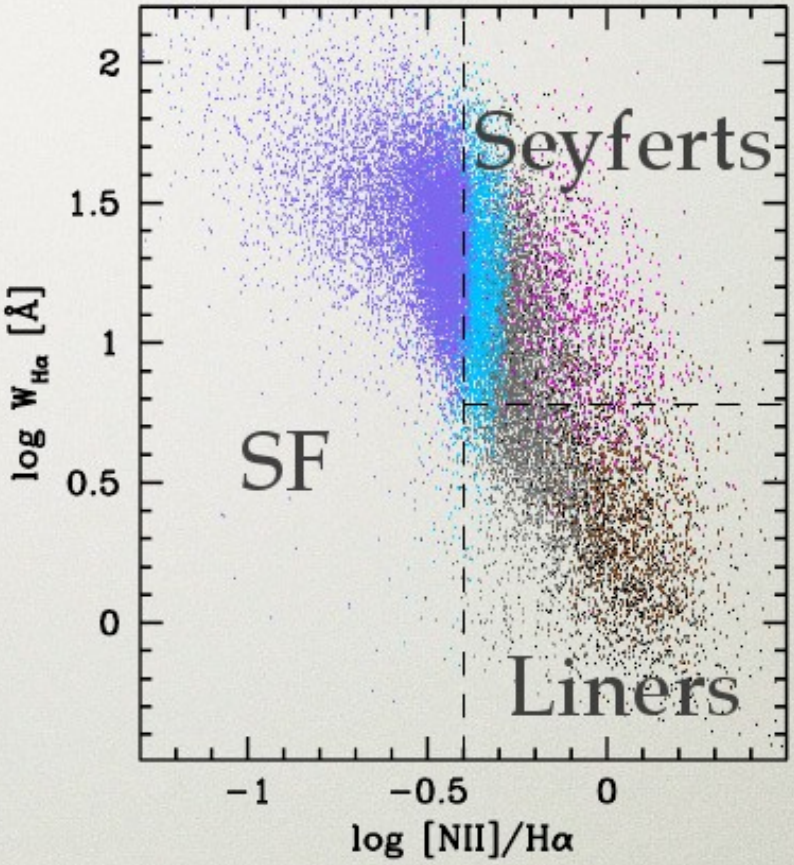
  - Collecting the data
  - Theoretical part
  - Filtering clustering the data
  - Analysis

Ligth from the galaxy :

- Stars (different ages, geometries and kinematics, star formation history, chemical evolution...)
- Gas (different densities...)
- Dust
- AGN (if there is any in galaxy)
- Emission from different types of processes (like shock waves, explosion, supernovae...)
- ...

How to make a BPT diagnostic diagram






```
select (log(F_5007)/log(F_4861)) as O3Hb, (log(F_6584)/log(F_6563)) as N2Ha from
(select t5007.synID, t5007.flux as F_5007, t5007.sn as SN_5007,
                     t4861.flux as F_4861, t4861.sn as SN_4861,
                     t6563.flux as F_6563, t6563.sn as SN_6563,
                     t6584.flux as F_6584, t6584.sn as SN_6584
   from el_fit as t5007, el_fit as t4861, el_fit as t6563, el_fit as t6584
  where t5007.id_line = 5007 and t4861.id_line = 4861 and t6563.id_line = 6563 and t6584.id_line = 6584 and
        t5007.synID = t4861.synID and t5007.synID = t6563.synID and t5007.synID = t6584.synID) as e
 where SN_5007 > 3 and SN_4861 > 3 and SN_6563 > 3 and SN_6584 > 3 and
       F_5007 > 0 and F_4861 > 0 and F_6584 > 0 and F_6563 > 0
```

- Links:
  - http://casjobs.sdss.org
  - http://www.openskyquery.org/Sky/SkySite/OSQForm/default.aspx
- Query OpenSkyQuery:

  Go to ADVANCED QUERY.
- make query and save as file: OIII.xml

```
SELECT o.objid, o.ra, o.dec, o.g, s.height, s.sigma, s.wave, s.restWave
FROM
   SDSSDR3:PhotoPrimary o, MyData:SDSS_C4_BCG t,
SDSSDR3:specline s
WHERE XMATCH(o, t) < 3.5 AND o.specobjid = s.specobjid AND s.lineID =
   5008
```

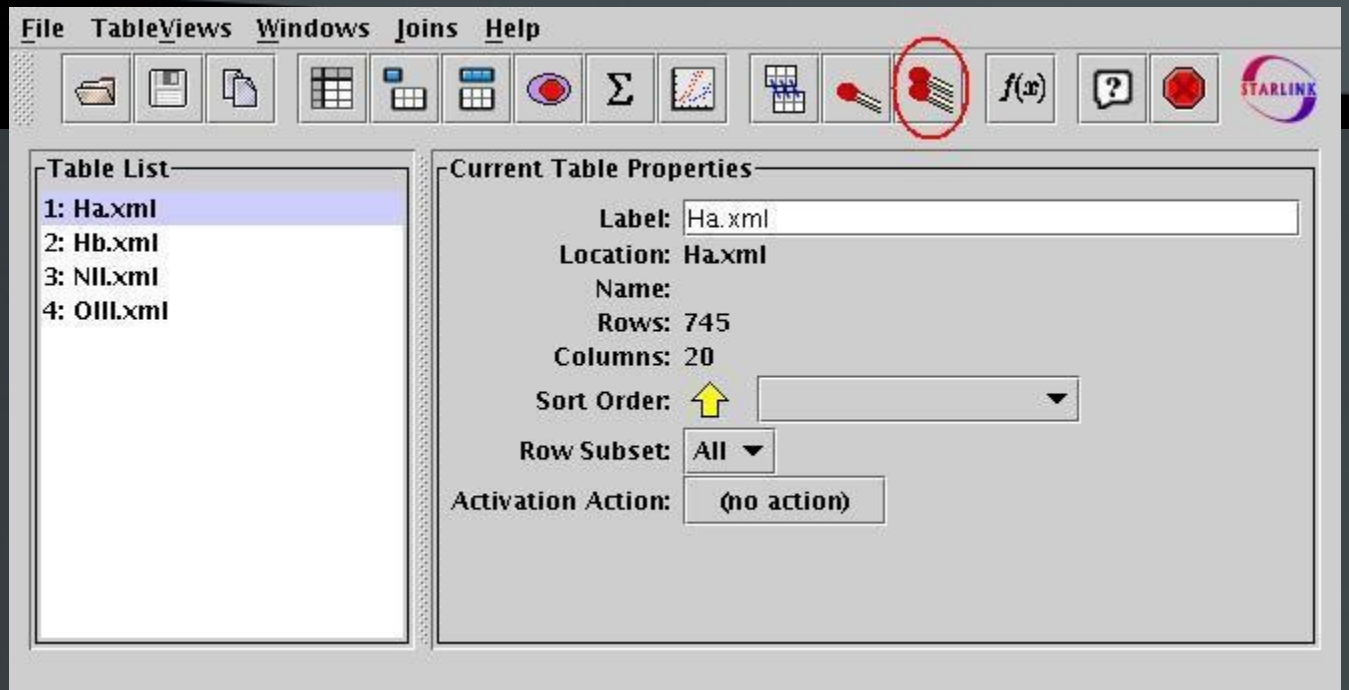
_____for others just chage 5008 with:

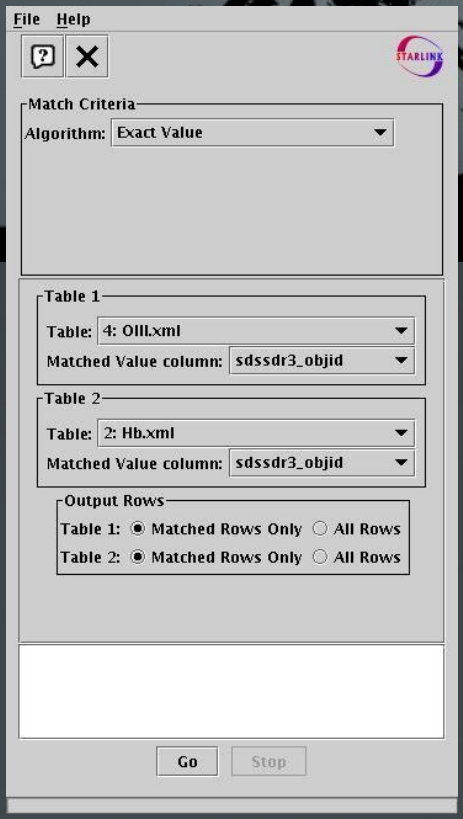
4863 save as file: Hb.xml; 6585 save as file: NII.xml; 6565 save as file: Ha.xml

Get all data with one query

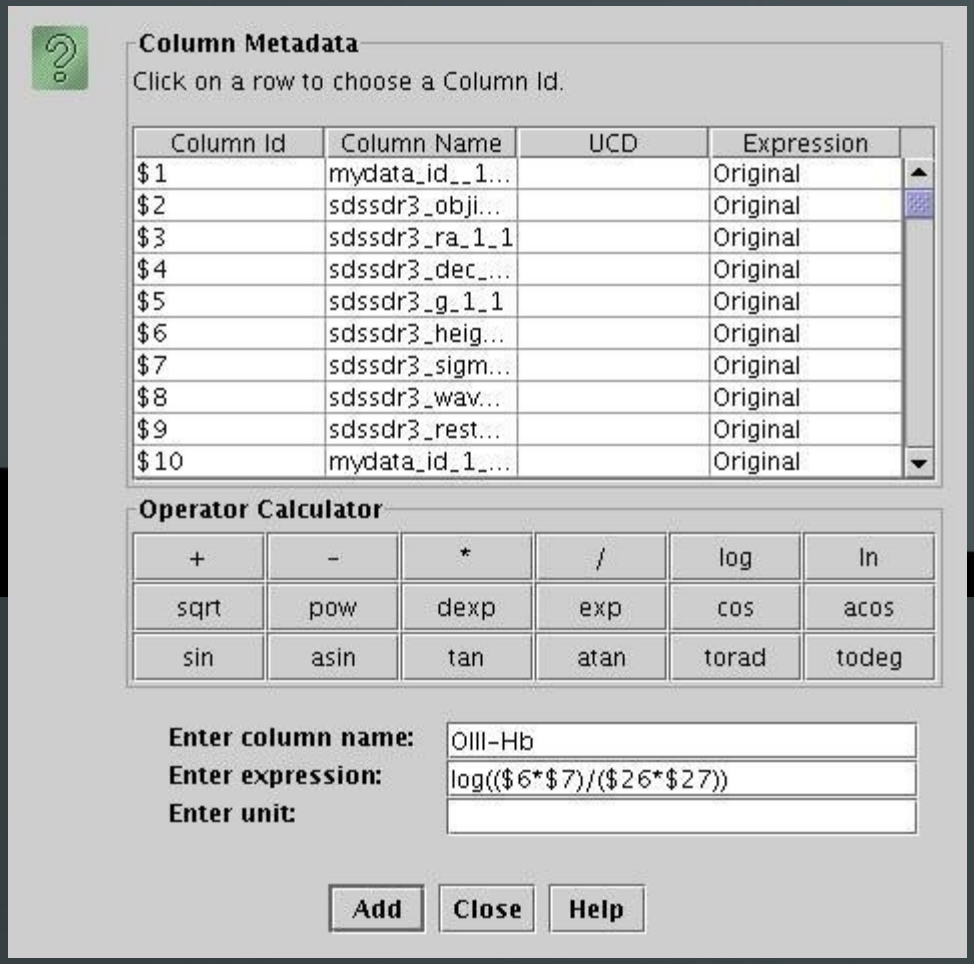
```
SELECT o.objid, o.ra,
    o.dec, o.g, s.height,
    s.sigma, s.wave, s.restWave,
    s.lineID
FROM
    SDSSDR6:PhotoPrimary o, TWOMASS:PhotoPrimary t,
    SDSSDR6:specline s
WHERE XMATCH(o, t) < 3.5 AND
    o.specobjid = s.specobjid AND
    (s.lineID = 5008 OR s.lineID = 4863 OR s.lineID = 6585 OR s.lineID = 6565)
```

- Merge Tables
  - Now we will use the Starlink Topcat tool to merge the tables into one table with all the information we need to create the diagnostic diagram. On my linux machine, I start Topcat with the following command. This will vary slightly with platform.
- $>topcat [a-z]*.xml



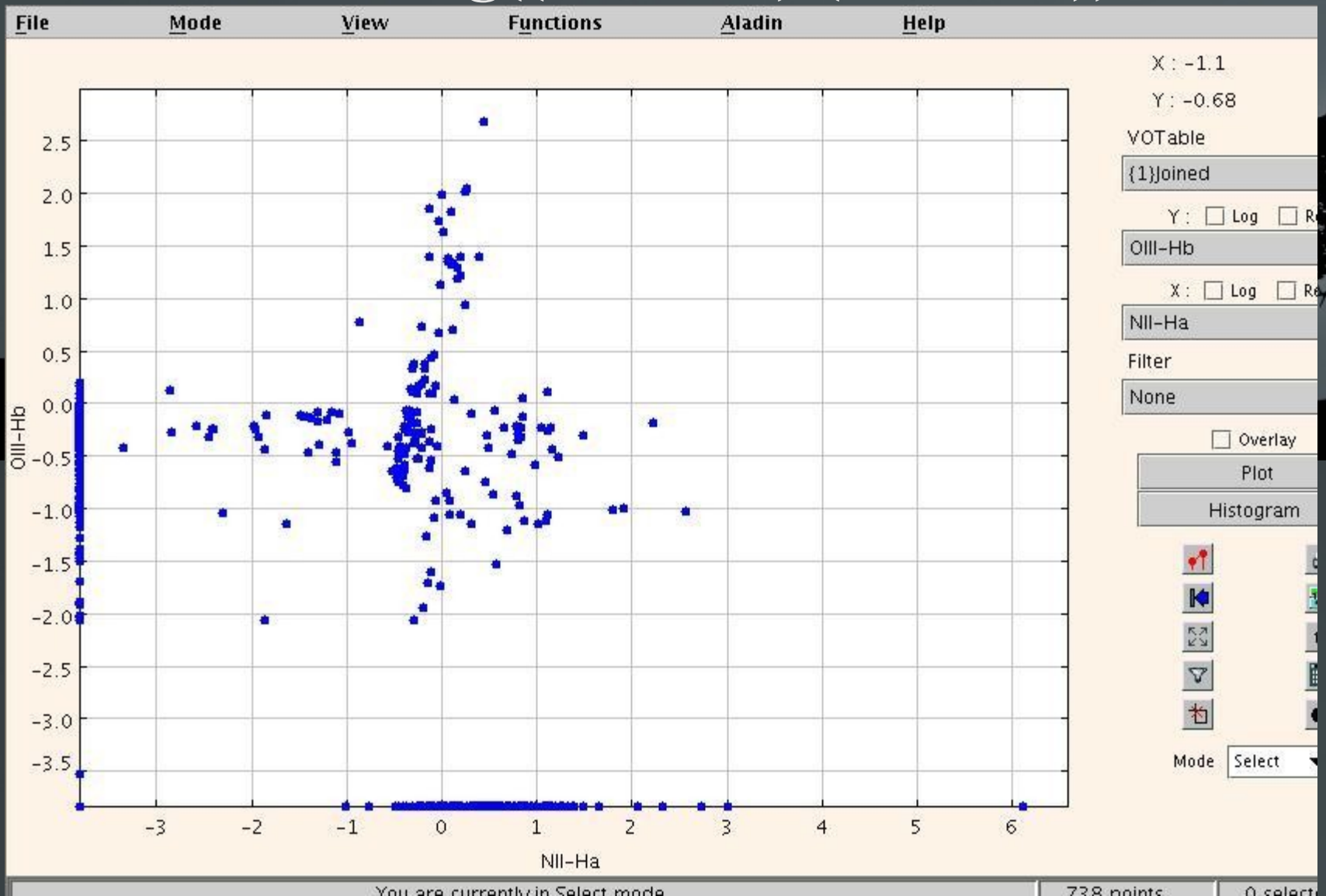


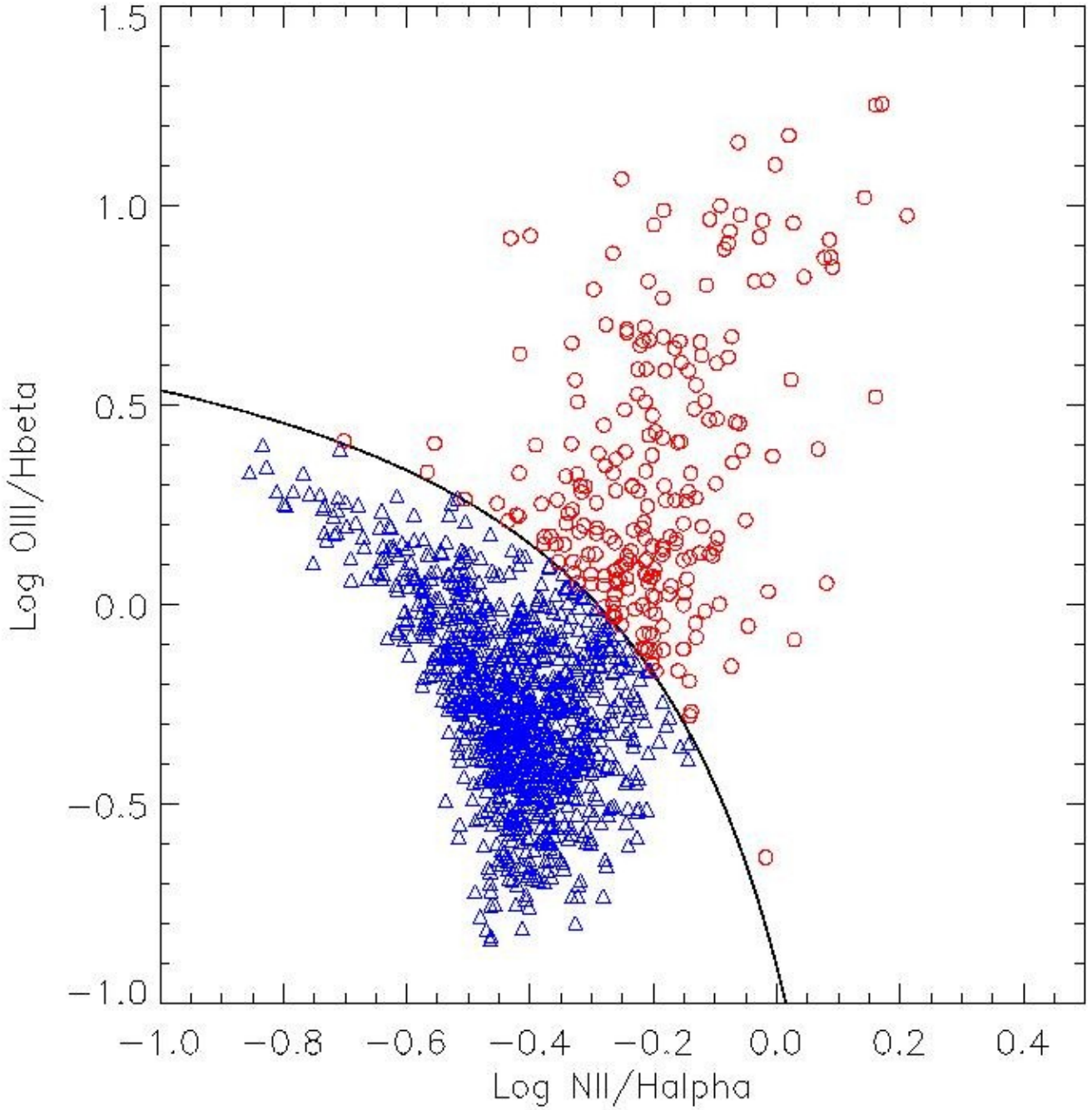
$>voplot


Column Metadata
Click on a row to choose a Column Id.
Column Id
Column Name
UCD
Expression
$1 mydata_id__1... Original
$2 sdssdr3_obji... Original
$3 sdssdr3_ra_1_1 Original
$4 sdssdr3_dec_... Original
$5 sdssdr3_g_1_1 Original
$6 sdssdr3_heig... Original
$7 sdssdr3_sigm... Original
$8 sdssdr3_wav... Original
$9 sdssdr3_rest... Original
$10 mydata_id_1_... Original
Operator Calculator
+ - * / log ln
sqrt pow dexp exp cos acos
sin asin tan atan torad todeg
Enter column name: OIII-Hb
Enter expression: log(($6*$7)/($26*$27))
Enter unit:
Add
Close
Help

OIII/Hb: log(($6*$7)/($26*$27))

NII/Ha: log(($47*$48)/($67*$68))




1.5
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
Log OIII/Hbeta
Log NII/Halpha

Another practical example using VO

Integrate optical and UV Galactic data through VO Services (STARLIGHT + SDSS)

Select suitable objects from ~ 700k objects

Use BPT (Baldwin, Phillips & Terlevich) graphics analysis to distinguish the different AGN and SF types and check theoretical predictions for [OIII]4959,5008 emission line-ratios

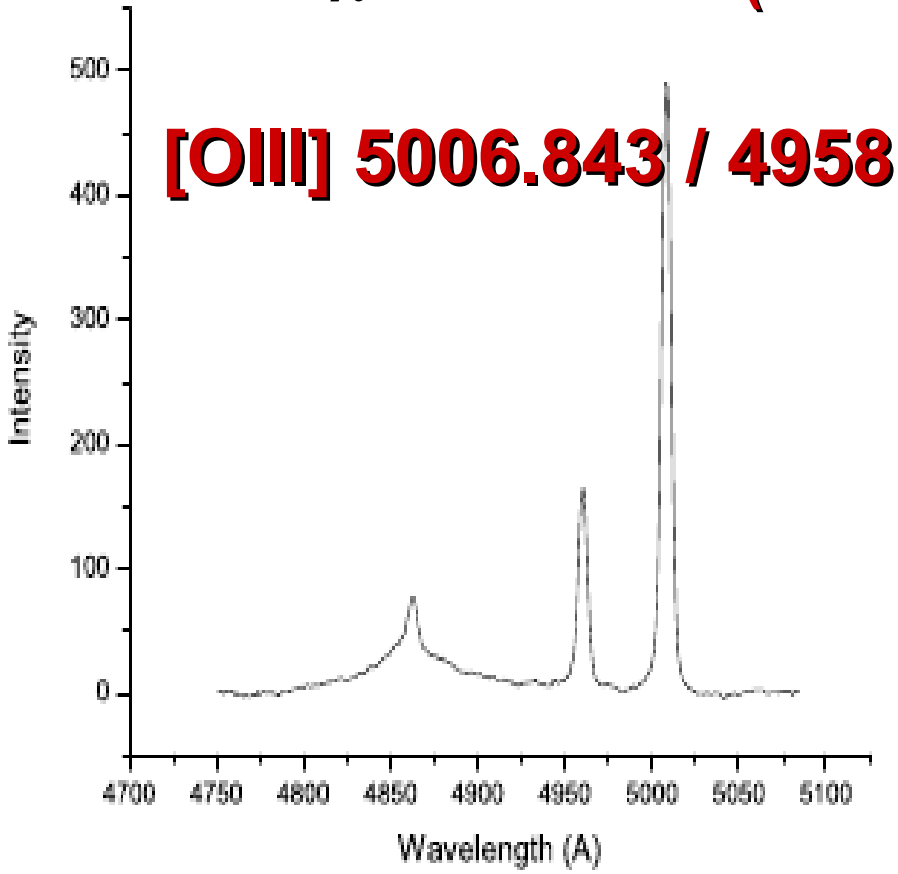
Theoretical Background

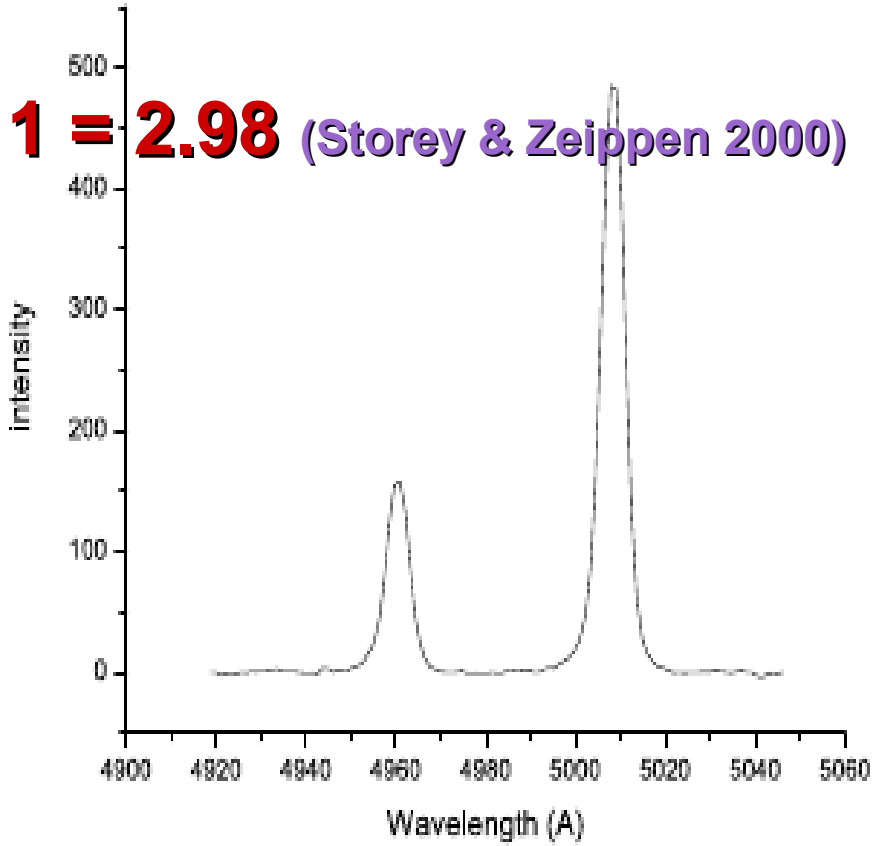
Theoretical work for the [OIII] lines ratios:

λ 4958.911 Å ($2s^2\ 2p^2\ {}^1D_2 - 2s^2\ 2p^2\ {}^3P_1$)

λ 5006.843 Å ($2s^2\ 2p^2\ {}^1D_2 - 2s^2\ 2p^2\ {}^3P_2$)

[OIII] 5006.843 / 4958.911 = 2.98 (Storey & Zeippen 2000)





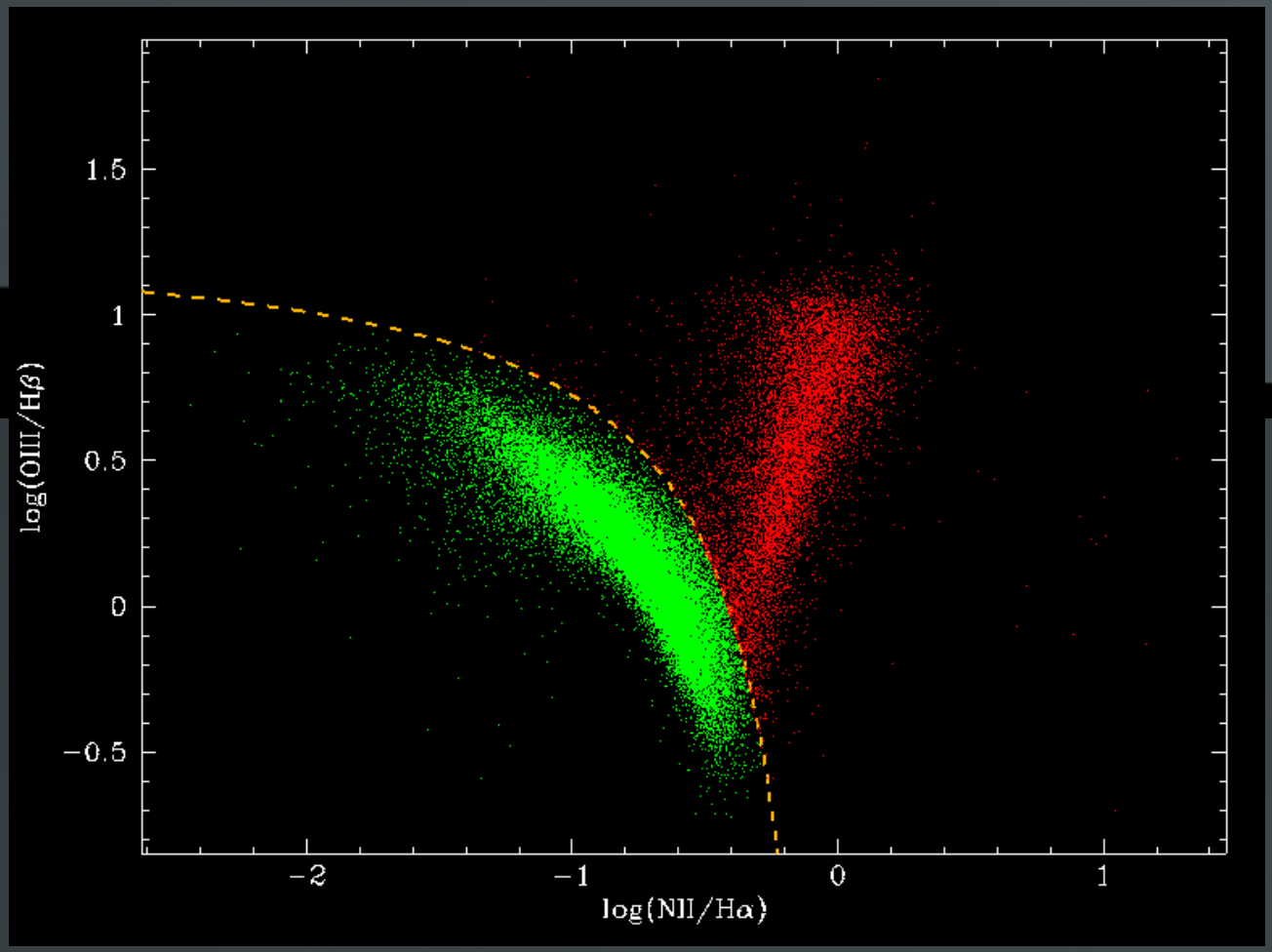
VO Services

X-match-catalogs

Optical and NIR data from SDSS

STARLIGHT Synthesis Code

SEAgal Synthesis @ BRAVO



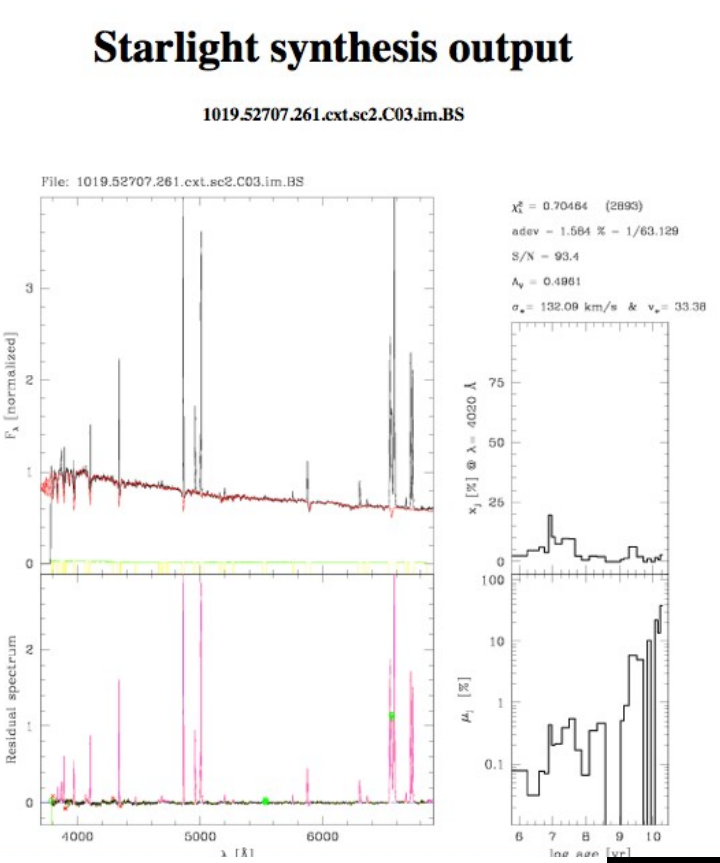
40422 objects from SDSS:

- 9167 AGN
- 31255 SF

SN > 12 on [OIII] Emission Lines

STARLIGHT Synthesis Code

STARLIGHT Database Screenshots



Parameters for **1019.52707.261.cxt.sc2.C03.im.BS**			
IAU Name	SDSS J121409.46+543136.6	**Base**	Base.bc03.S
RA	183.53944336	**DEC**	54.526839066
chi2	0.7046	**adev**	1.58
v0	33.4	**vd**	132.1
SN_w	76.7	**Mcor_gal**	9.0771
RedLaw	CCM	**AV**	0.496
at_flux	7.5856	**at_mass**	9.9914
am_flux	0.5132	**am_mass**	0.9412

SDSS parameters table (?)

z	eClass	m_u	m_g	m_r	m_i	m_z	fm_u	fm_g	fm_r	fm_i
0.008144	0.373874	15.99392	16.23726	15.62477	15.60672	15.40808	15.77365	15.15583	14.75961	14.72063

fm_z	Mu	Mg	Mr	Mi	Mz	SB_50_r	CI_r	petrorad_r	petroR50_r
14.43532	-16.75155	-16.4915	-17.10905	-17.11978	-17.31842	17.33306	2.10169	1.76766	0.89347

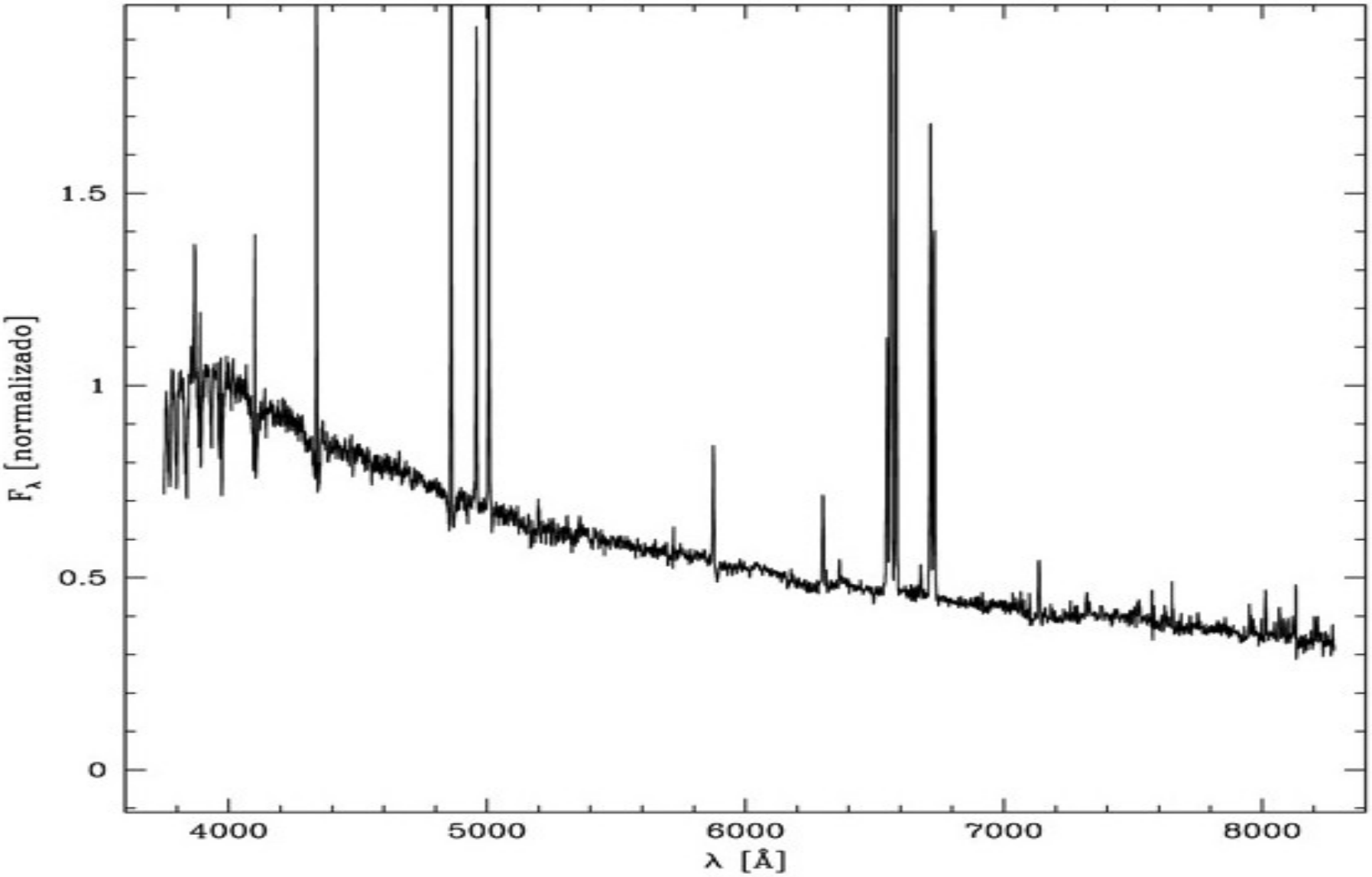
petroR90_r	expAB_r	deVAB_r	D	DA	R50	R90	DL	log_L	Mr_fiber	log_L_fiber
1.87779	0.62251	0.6004	34.81588	0.29596	0.14959	0.3144	35.09943	8.69162	-17.97421	9.03768

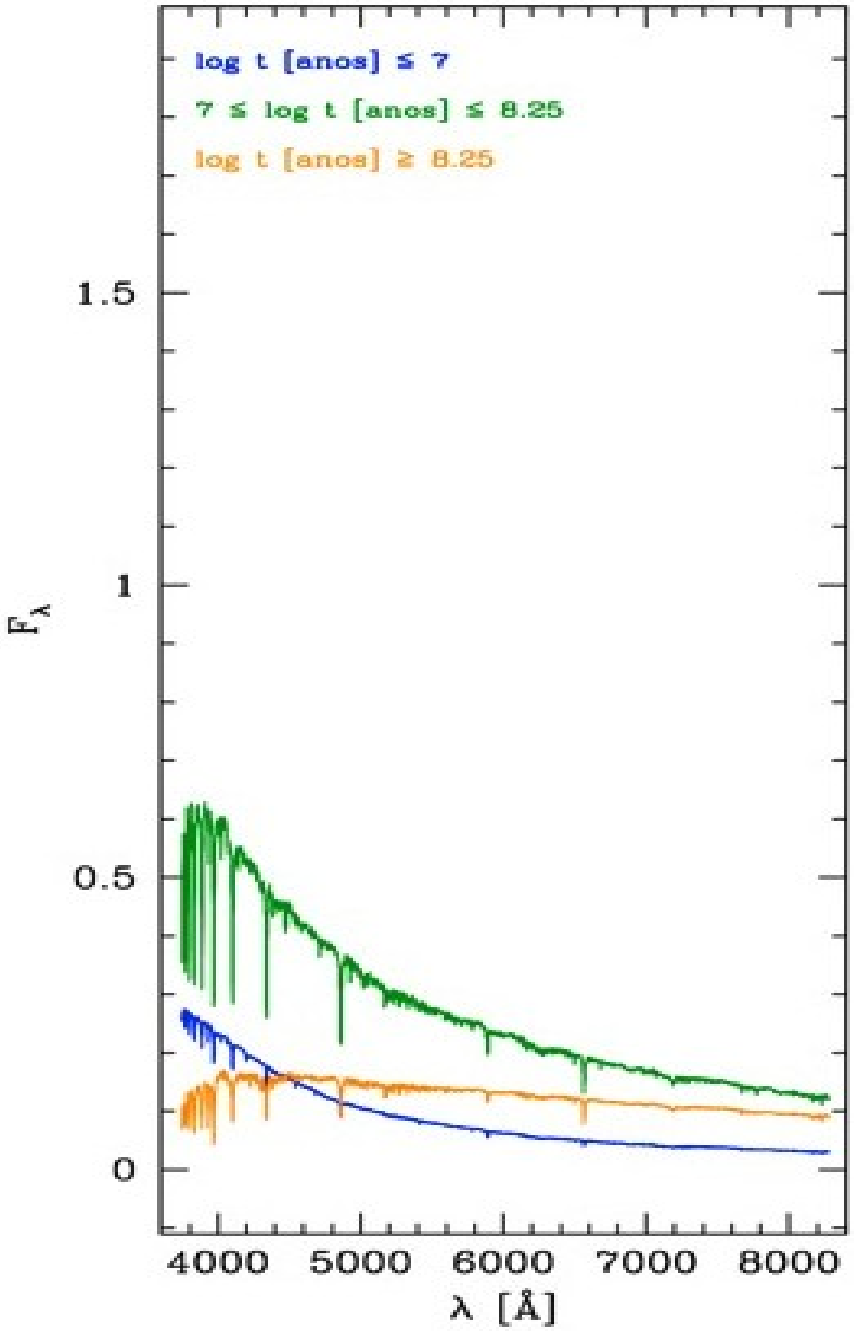
Mz_fiber	log_L_fiber_z	petrorad_z	petroR50_z	petroR90_z	DA_z	z50	z90
-18.29118	9.10847	1.77379	0.89732	1.86632	0.29698	0.15024	0.31248

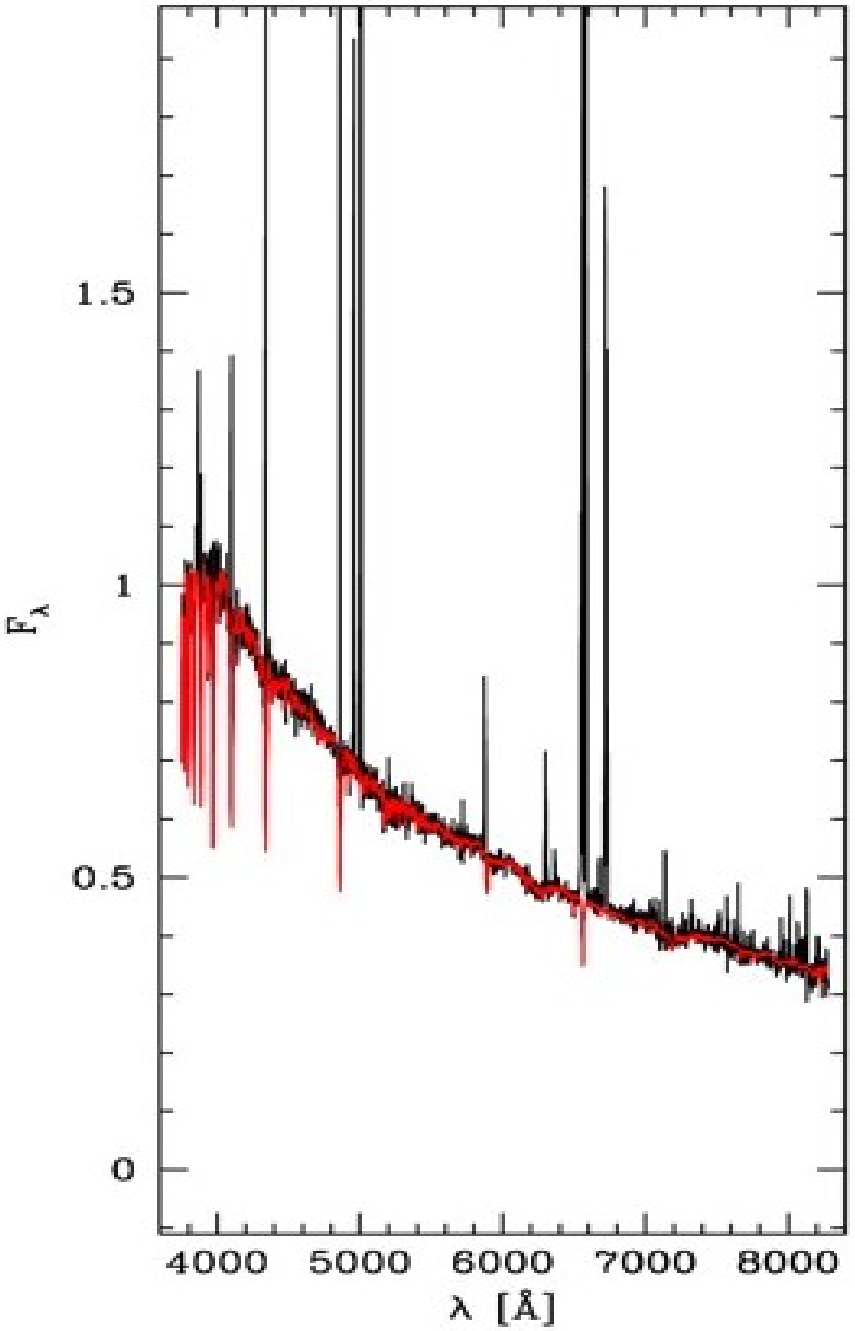
Emission line measurements for **1019.52707.261.cxt.sc2.C03.im.BS**

El	CentralWl	LowBlueCon	UppBlueCon	LowRedCon	UppRedCon	Flux	FluxErr	EqWidth	EqWidthErr	VelDisp	VelDispErr	Displ	DisplErr	SN	ConFlux	ConFluxErr
[OII]3727	3727.42	3653	3713	3741	3801	-999	-999	-999	-999	-999	-999	-999	-999	0	361.464	12.35
[NeIII]3869	3869.06	3845	3860	3900	3920	594.028	50.193	1.442	0.123	166.546	9.828	-70.235	9.828	8.762	412.038	4.694
Hdelta	4101.73	4050	4080	4120	4150	1426.042	47.221	3.418	0.133	120.636	2.917	7.849	2.917	27.107	417.163	8.573
Hgamma	4340.46	4300	4330	4385	4450	2870.216	43.53	7.824	0.171	120.636	1.335	7.849	1.335	56.507	366.834	5.792
[OIII]4363	4363.21	4300	4350	4385	4450	-999	-999	-999	-999	-999	-999	-999	-999	0.021	349.545	21.673

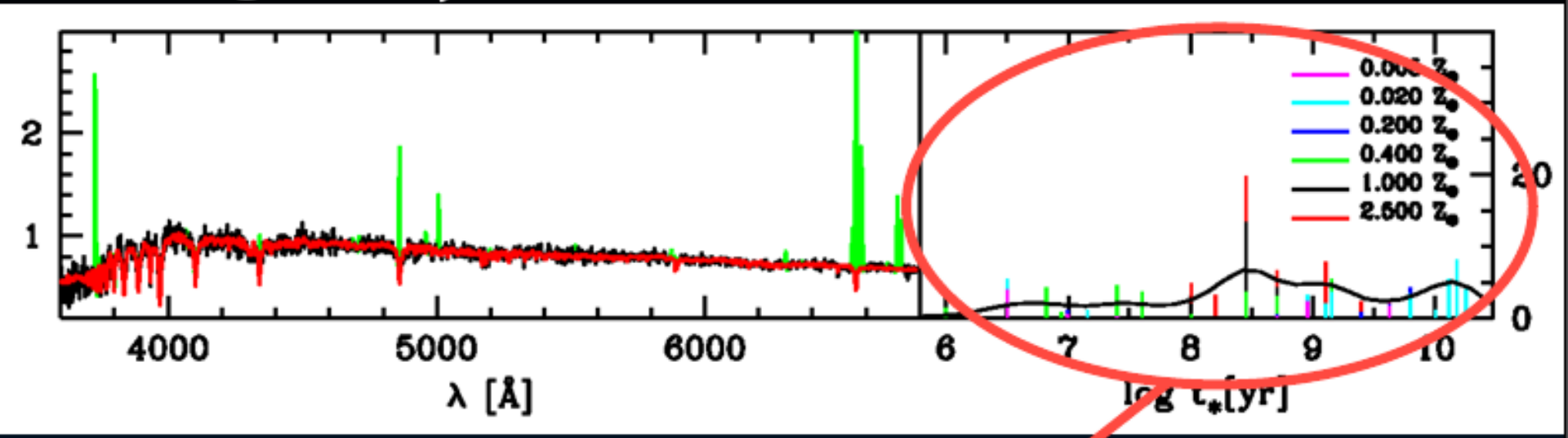
Galaxy spectar




log t [anos] ≤ 7
7 ≤ log t [anos] ≤ 8.25
log t [anos] ≥ 8.25
1.5
1
0.5
0
F_λ
4000
5000
6000
7000
8000
λ [Å]


1.5
1
0.5
0
F_λ
4000
5000
6000
7000
8000
λ [Å]

A galaxy modeled with STARLIGHT



Observables:

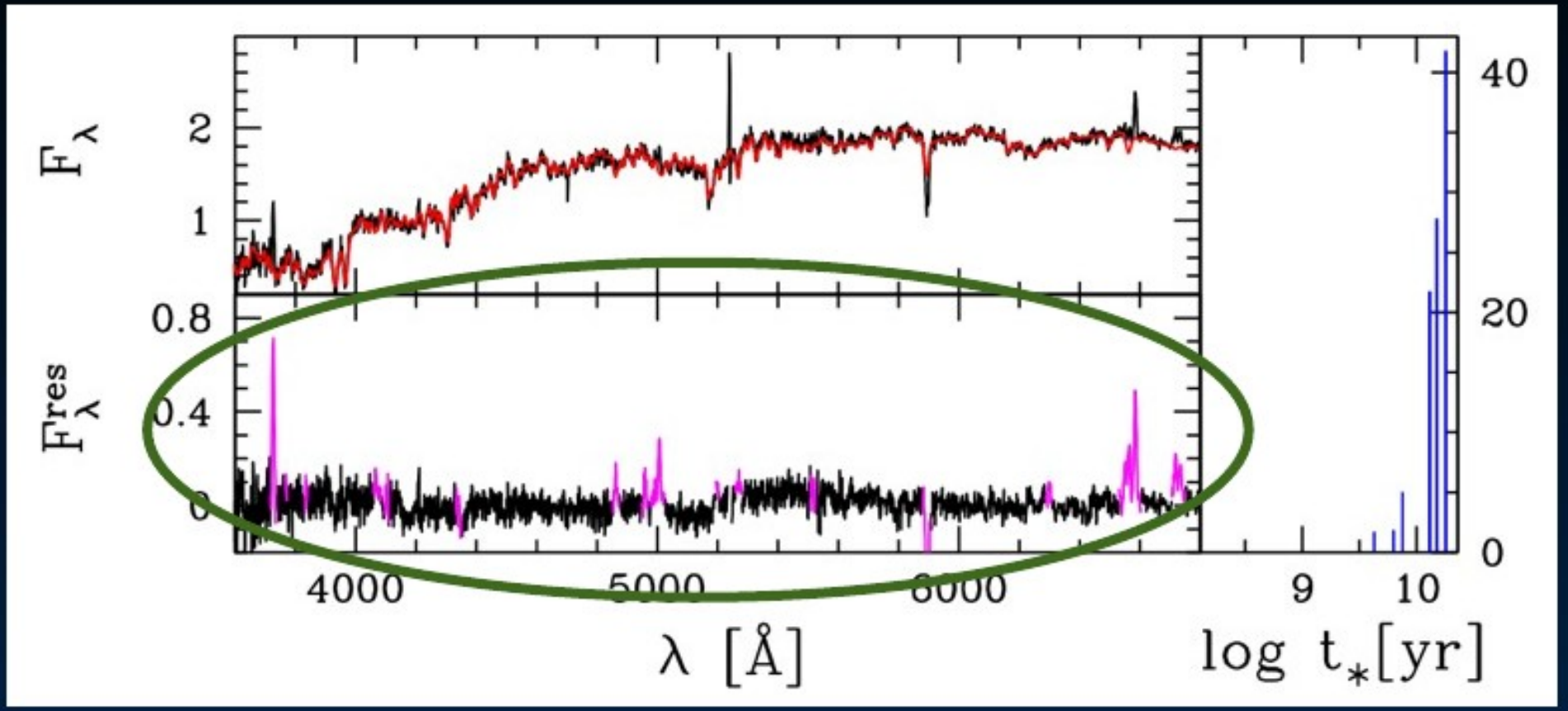
full spec (4000 pixels)

Star formation histories and chemical evolution

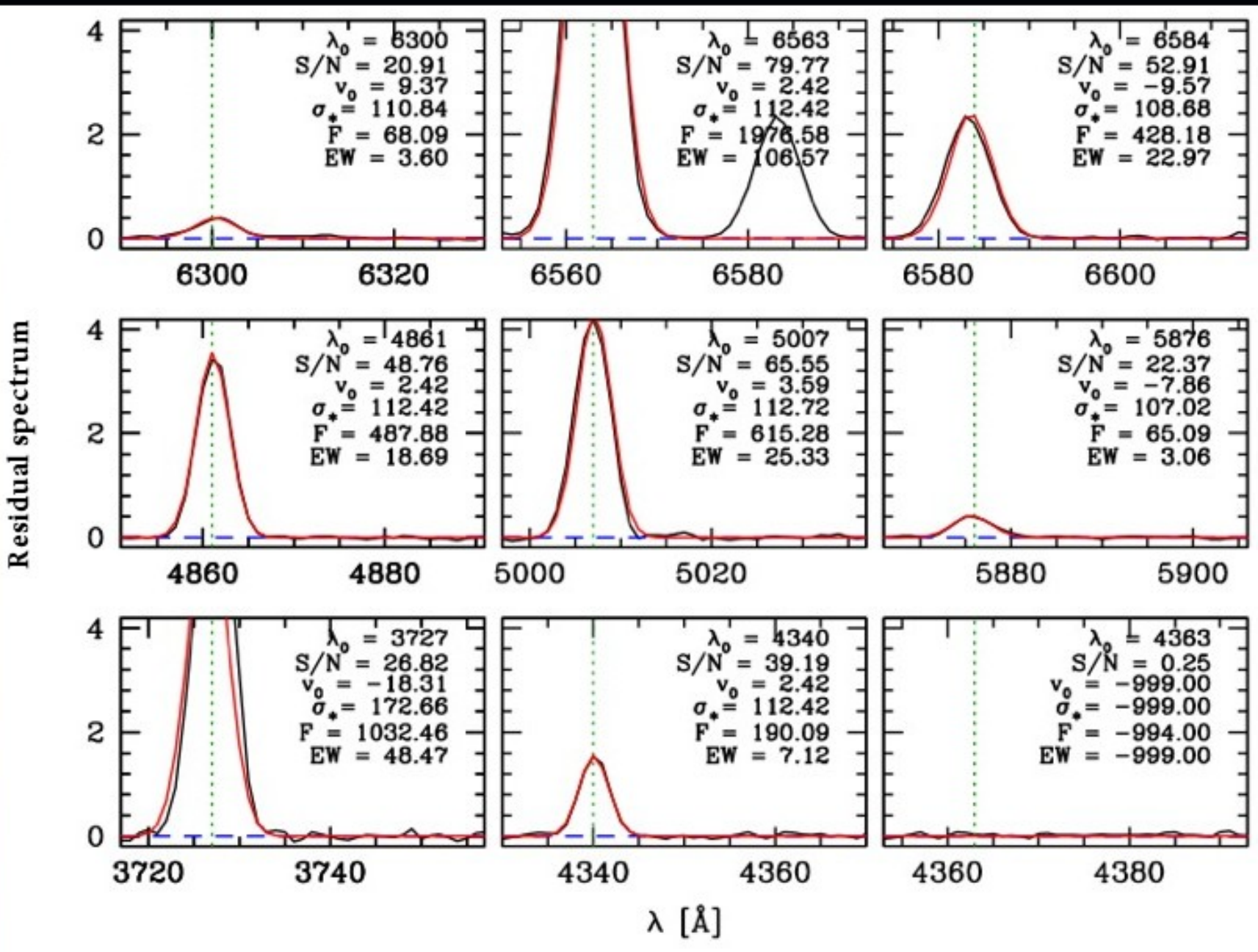
Base:

25 ages x 6 Zs = 150 BC03 SSPs

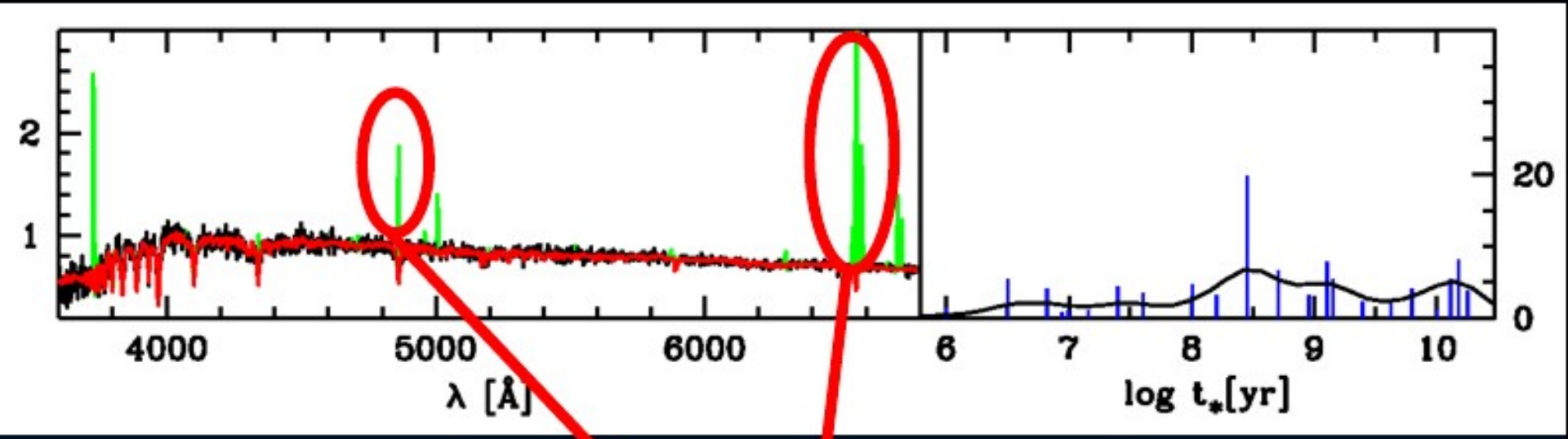
Residual spectra

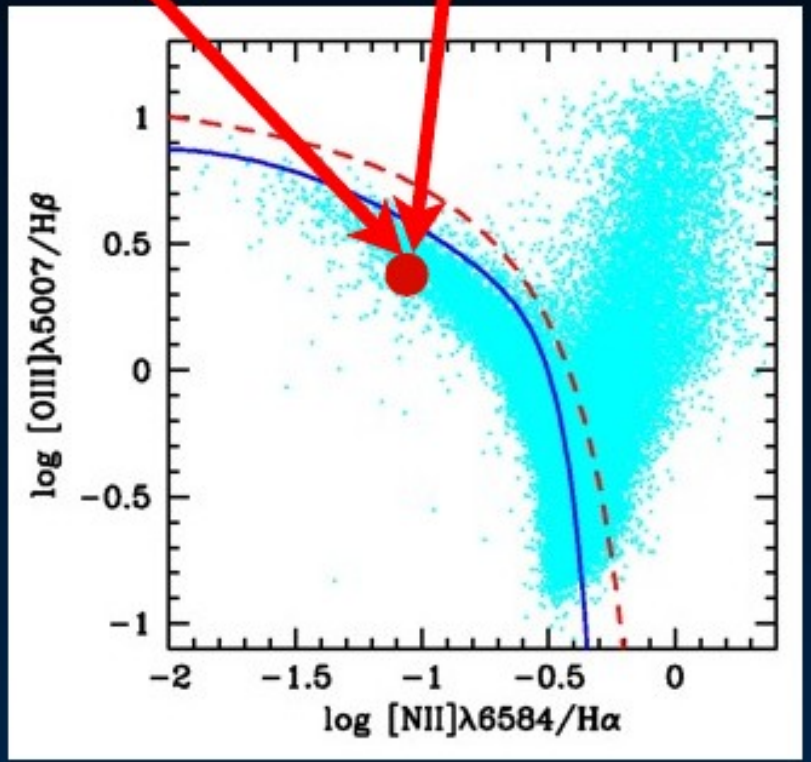


Gas: Emission line fitting code

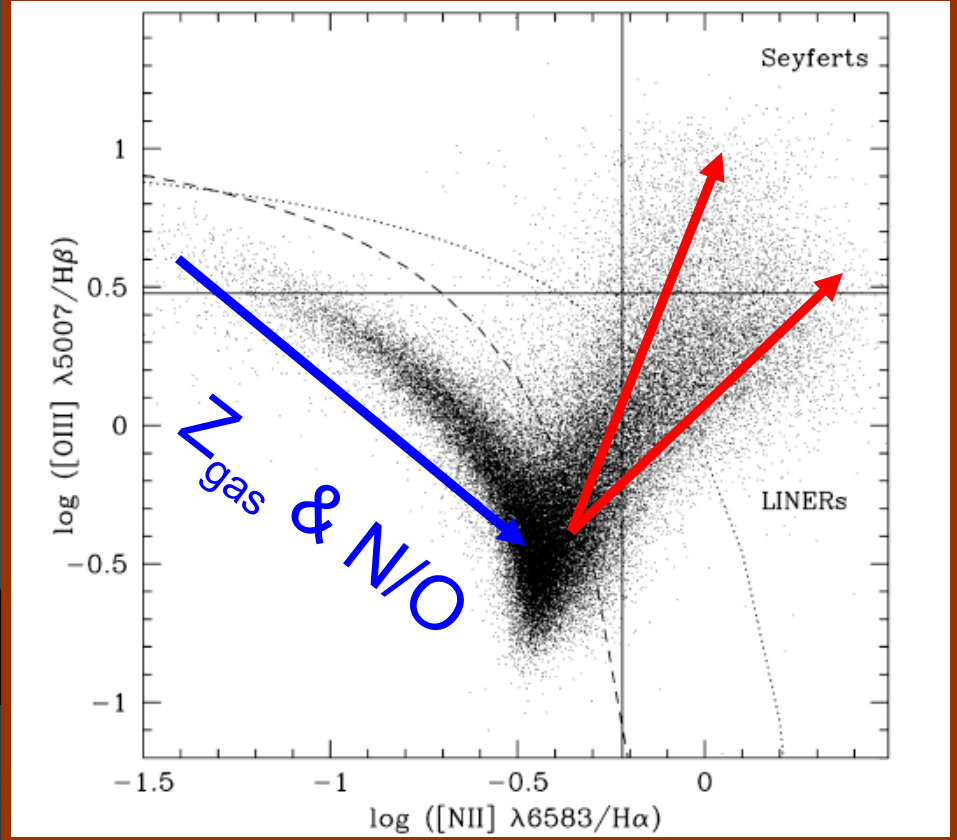


Emission lines ⟶ diagnostic diagrams





Diagnostic Diagrams: Gas physics



L_{AGN}?
U?
AGN/SB?
M_{BH}?
Z_{gas}?
N/O?

Galaxy Properties?

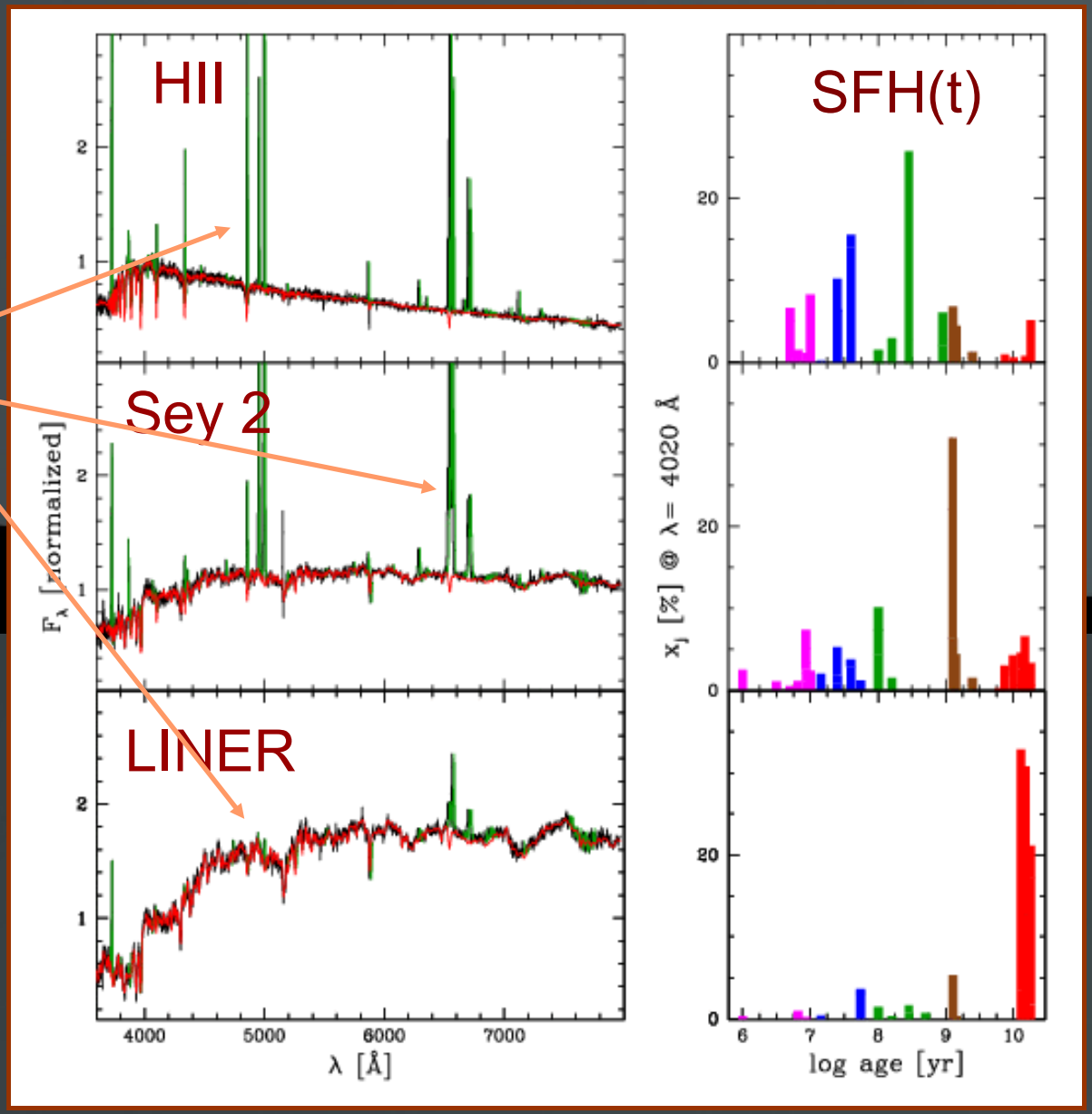
[Forbidden lines] ~ Thermometers: $\langle k T_e \rangle$

- In Starbursts: $T_e \Rightarrow Z_{gas}$
- In AGN: $T_e \Rightarrow \langle h\nu \rangle_{AGN}$ & Z_{gas} & ...

Dissecting galaxy spectra: stars + gas

Emission lines

[NII]/Hα
[OIII]/Hβ
[OI]/Hα
[SII]/Hα
[OII]/[OIII]
...



Galaxy Properties

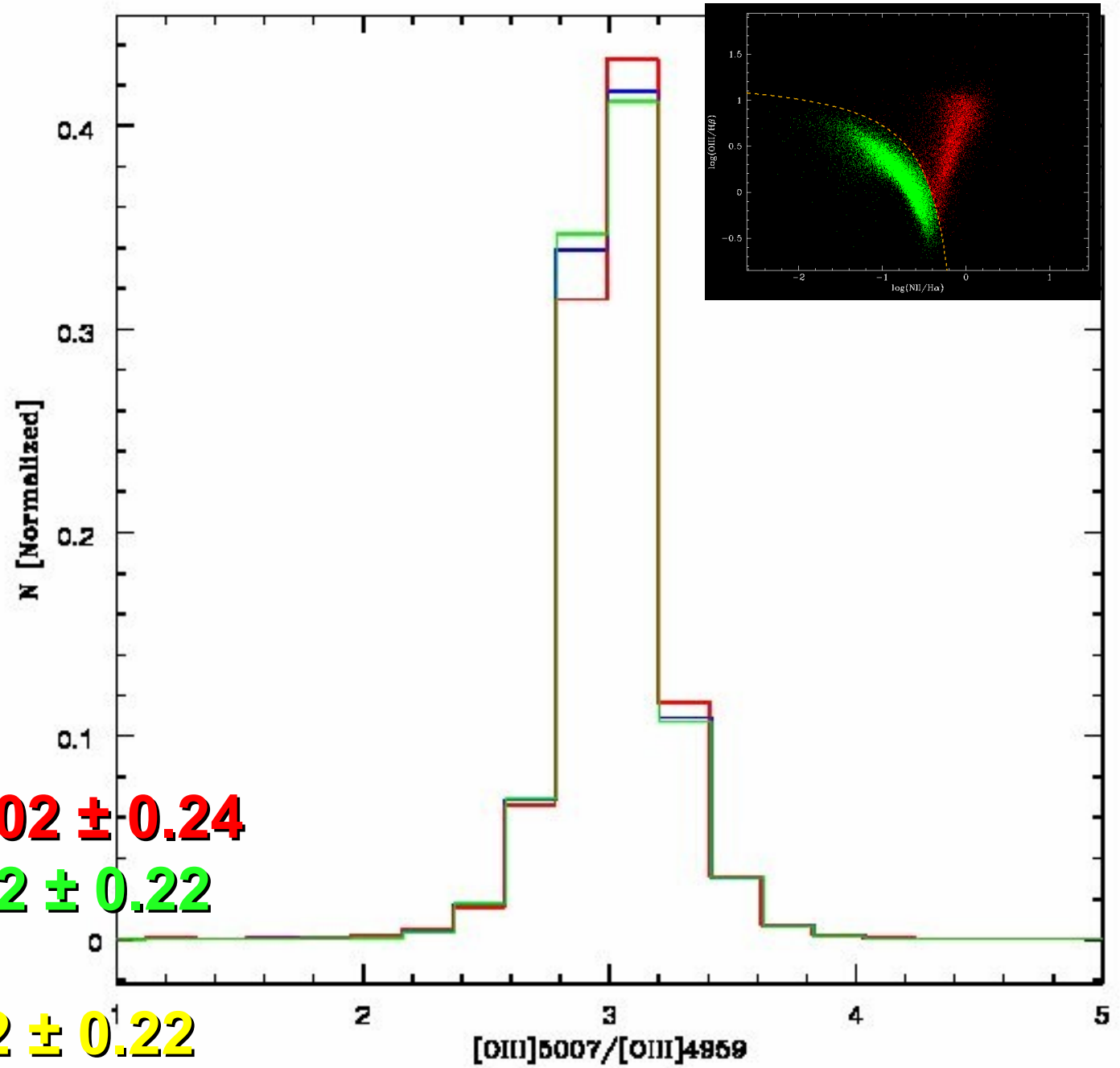
M_*
SFR(t)
SSFR(t)
σ_*
A_V
$\langle \log t_* \rangle$
$\langle Z_* \rangle$
...

Using 40422 objects from SDSS, with BPT we separated:

- 9167 AGN
- 31255 SF

filter: SN > 12

@ [OIII] Emission Lines



OIII intensity ratio for:

AGN (9167) = 3.02 ± 0.24

SF (31255) = 3.02 ± 0.22

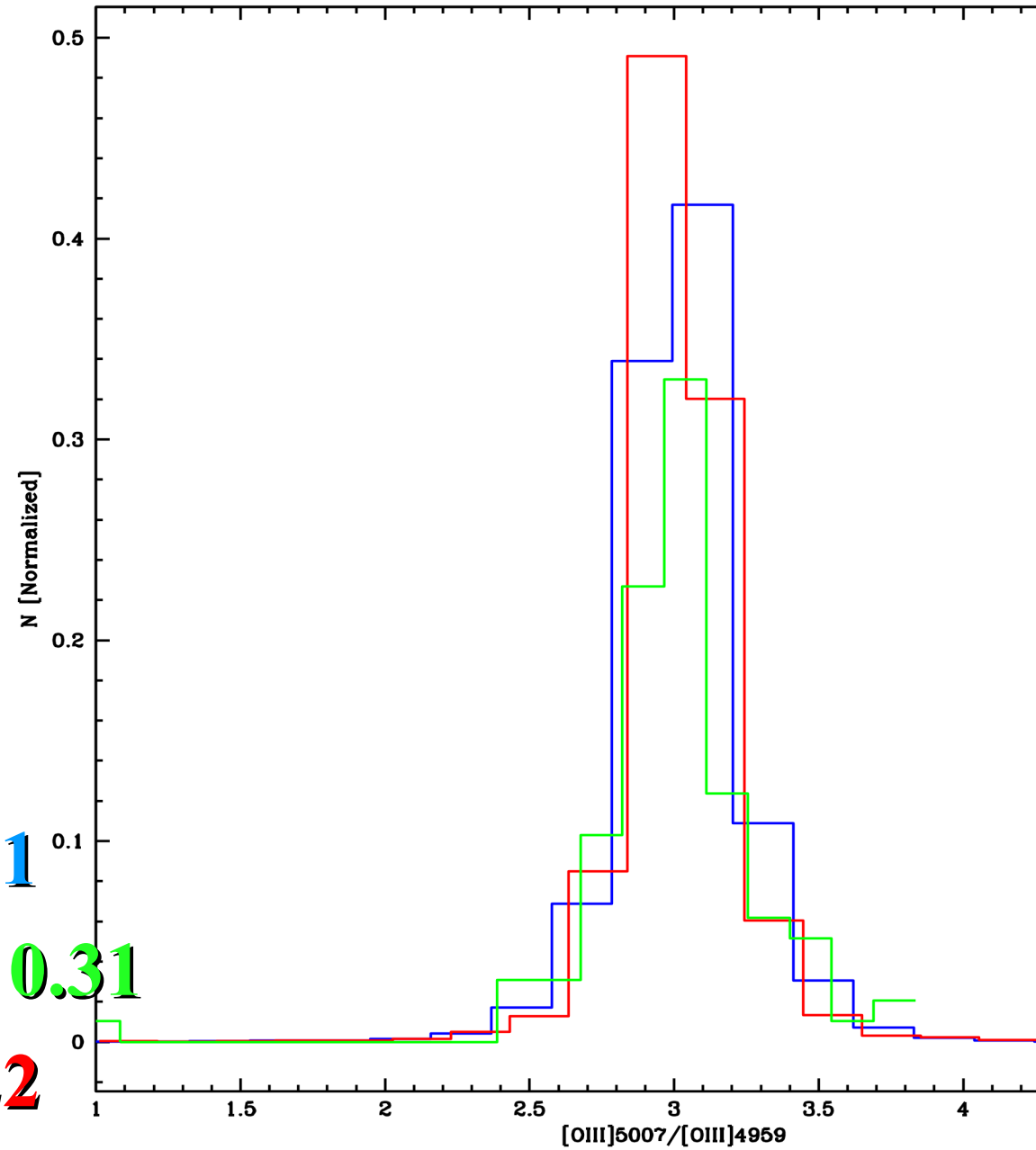
All (41044)= 3.02 ± 0.22

Using 40422 objects from SDSS, with BPT we separated:

- 9167 AGN
- 31255 SF

filter: SN > 12

@ [OIII] Emission Lines



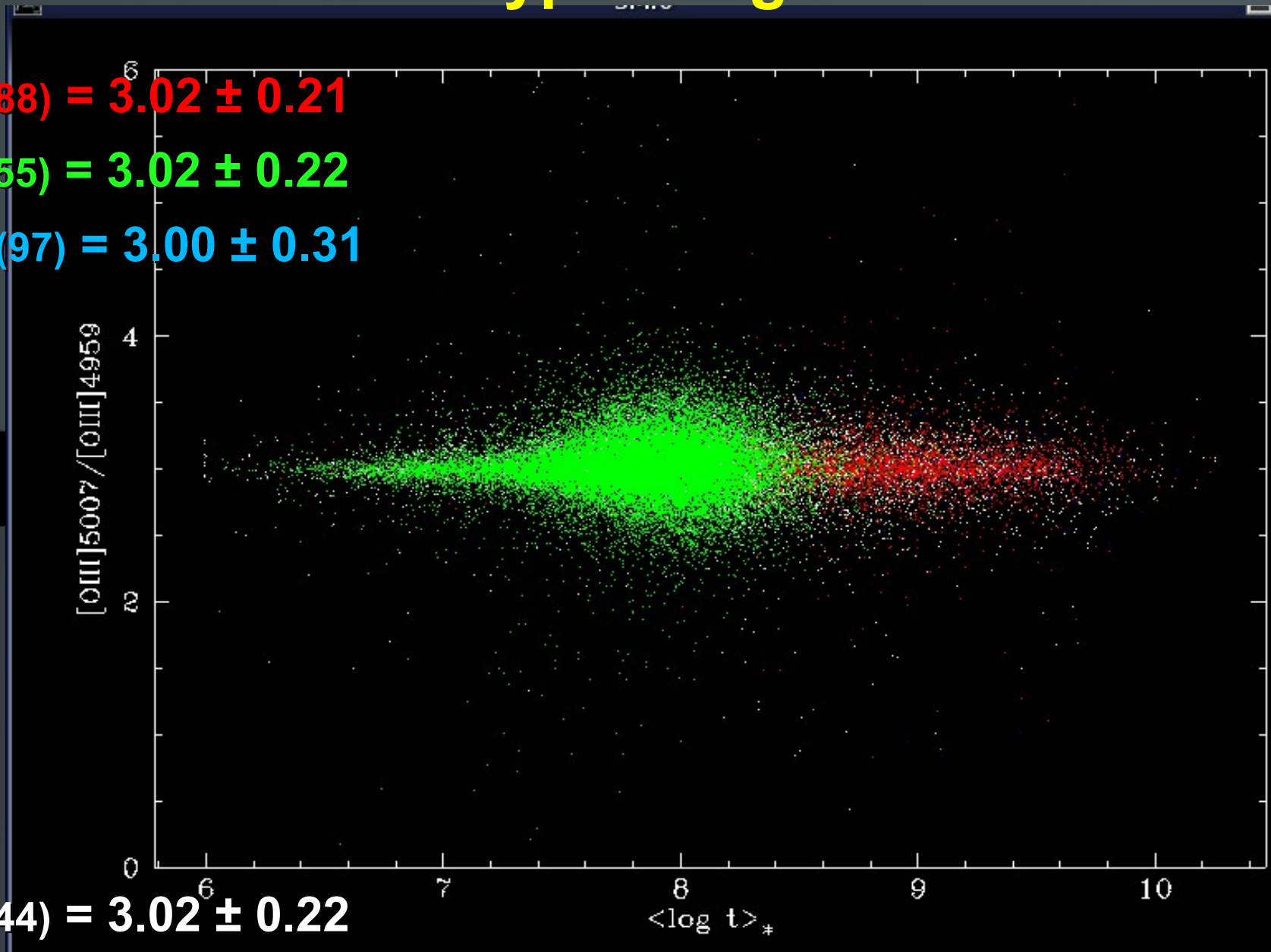
OIII intensity ratio for:

Sy2 (4788) = 3.02 ± 0.21

LINERs (97) = 3.00 ± 0.31

All (41044) = 3.02 ± 0.22

Star fomation hystory over OIII intensity ratio for different types of galaxies



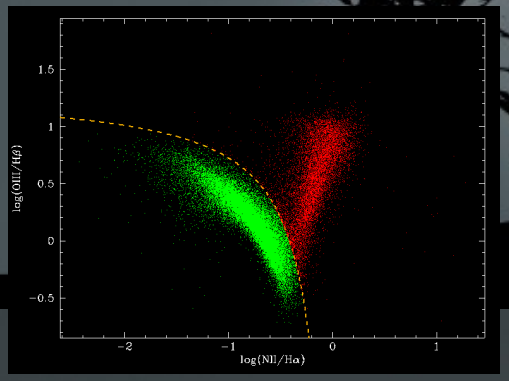
Results

Before:

Dimitrijevíc et al. 2007 → for 62 AGN spectra, from SDSS

Sy1 (34) = 2.993 ± 0.014

now, for 41044 objects from SDSS



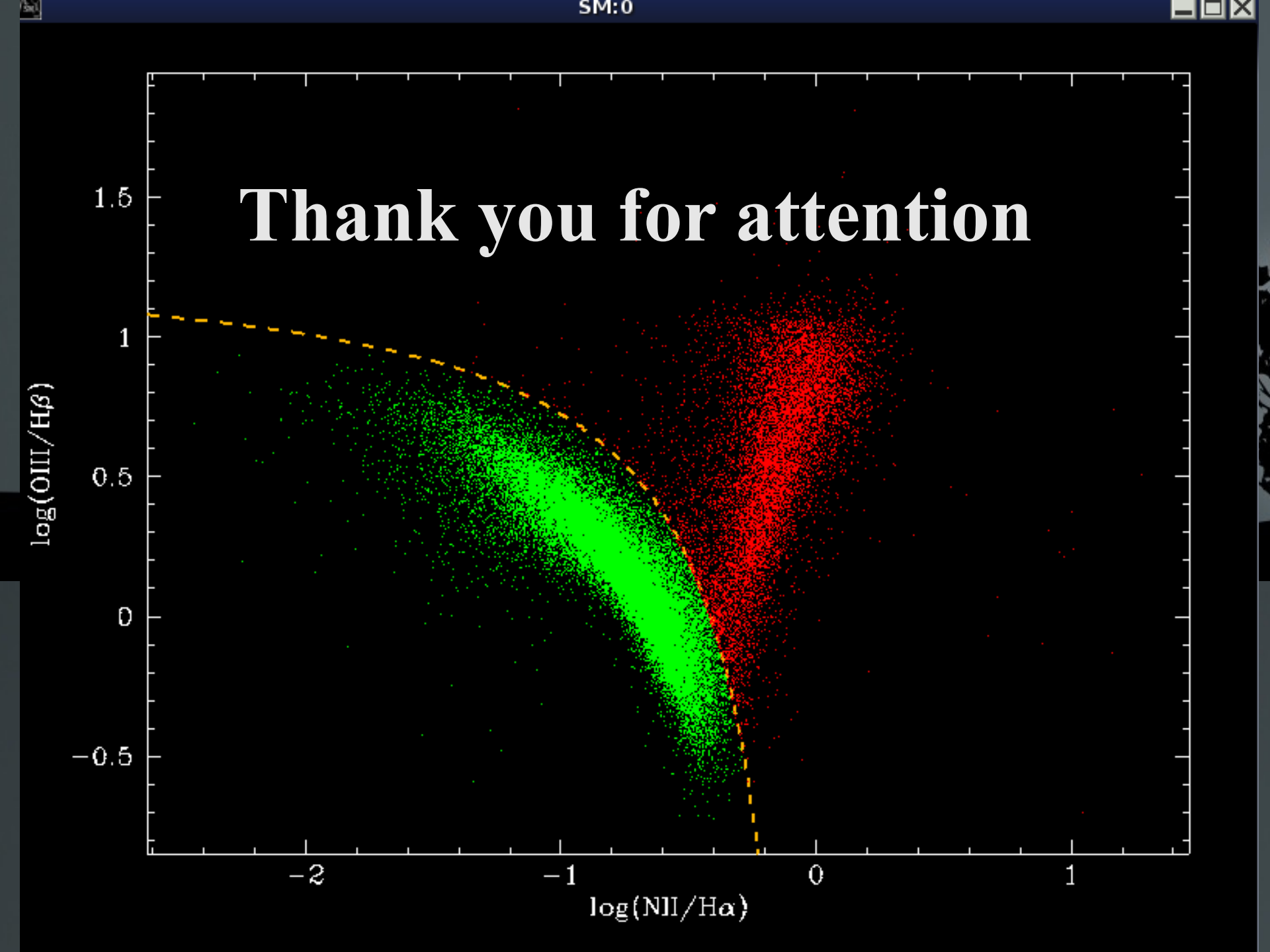
AGN (9167) = 3.02 ± 0.24

SF (31255) = 3.02 ± 0.22

Sy2 (4788) = 3.02 ± 0.21

LINER (97) = 3.00 ± 0.31

All objects (41044) = 3.02 ± 0.22


SM:0
Thank you for attention
1.5
1
0.5
0
−0.5
log(OIII/Hβ)
−2
−1
0
1
log(NII/Hα)