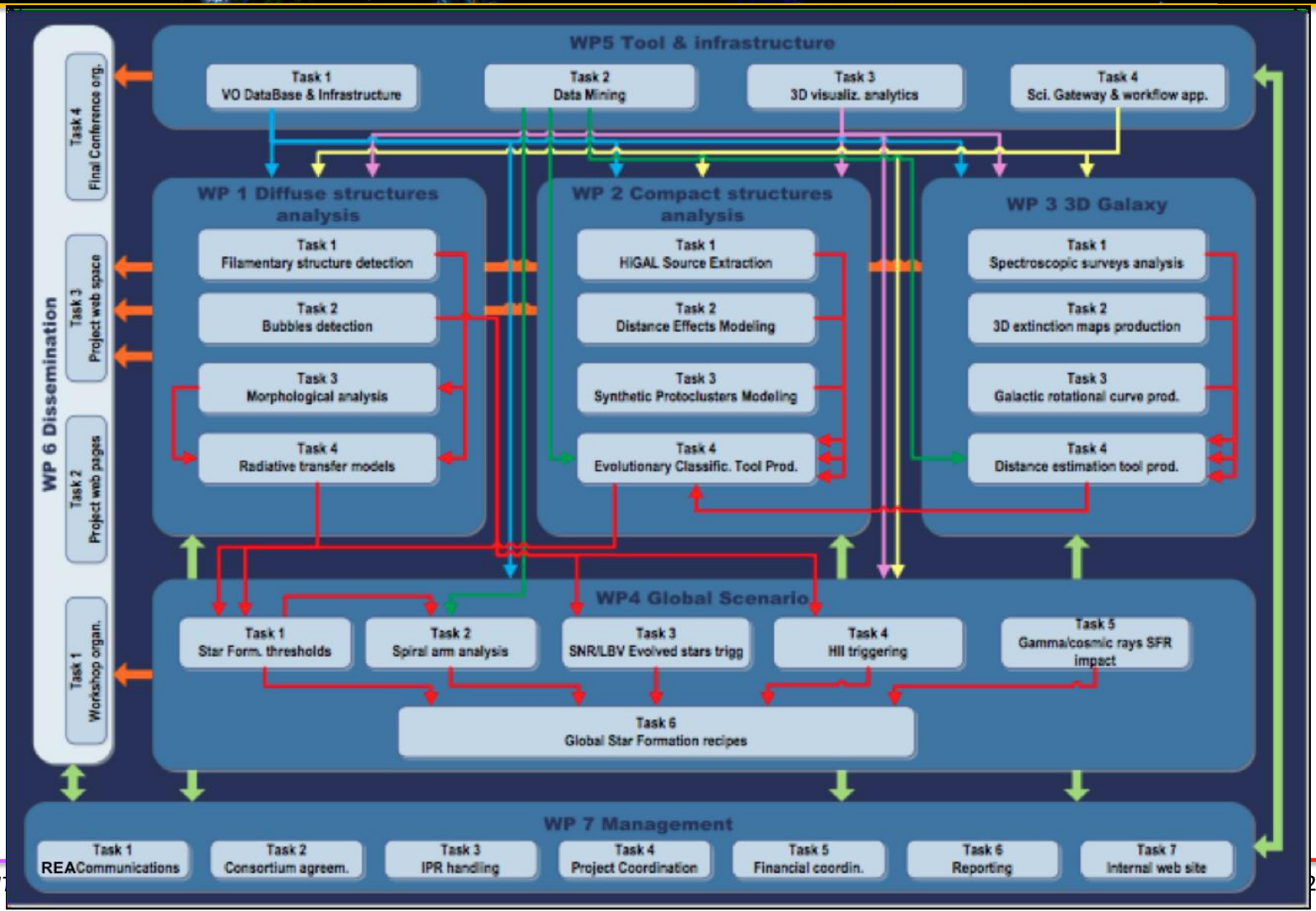
An ICT and data mining framework for knowledge discovery in the ViaLactea project

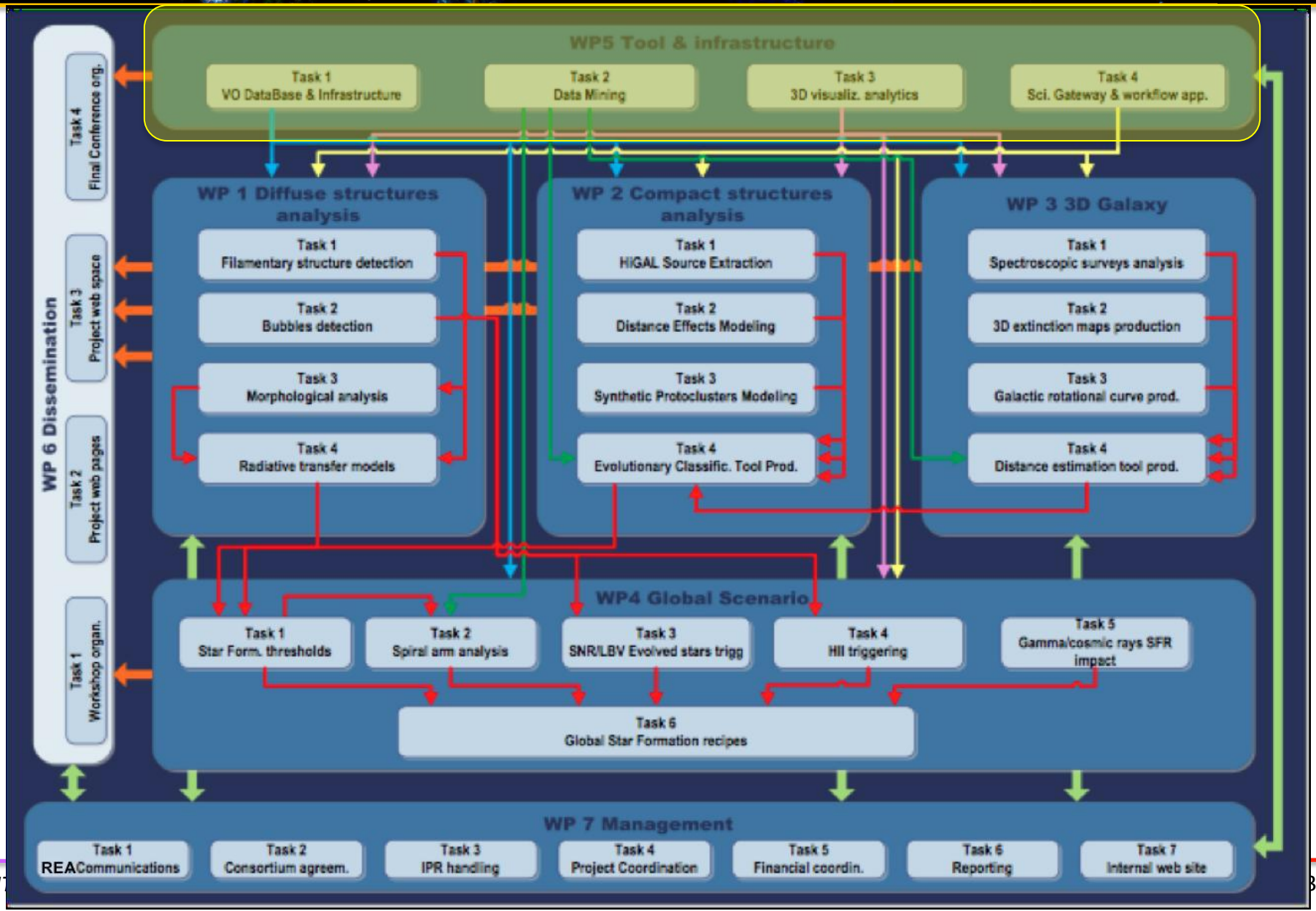
Giuseppe Riccio[1], Ugo Becciani[2], Massimo Brescia[1], Robert Butora[3], Stefano Cavuoti[1], Alessandro Costa[2], Anna Maria Di Giorgio[4], Akos Hajnal[5], Gabor Hermann[5], Peter Kacsuk[5], Istvan Marton[5], Amata Mercurio[1], Sergio Molinari[4], Marco Molinaro[3], Riccardo Smareglia[3], Fabio Vitello[2]

(1) INAF – Astronomical Observatory of Capodimonte, Via Moiariello 16, I-80131 Napoli, Italy
(2) INAF – Astronomical Observatory of Catania, Via S. Sofia 78, I-95123 Catania, Italy
(3) INAF – Astronomical Observatory of Trieste, Via Tiepolo 11, I-34131 Trieste, Italy
(4) INAF – IAPS, Via del Fosso del Cavaliere 100, I-00133 Roma, Italy
(5) MTA SZTAKI, Kende u. 13-17, Budapest, Hungary

The role of WP5 in the project



The role of WP5 in the project



WP5 main tasks

- **TASK 1:** INAF – Astronomical Observatory of Trieste

Database and Virtual Observatory Infrastructure .

This will ensure the ***integration and interoperability*** of all new data and tools products by enduring their compliance to ***VO standards***. The usage and test of VO standards and tools can ***increase the scientific productivity*** and encourage the development of automatic pipelines to explore existing VO-compatible DB/archives.

- **TASK 2:** INAF – Astronomical Observatory of Capodimonte, Napoli

Data Mining Systems

Data Mining System are intelligent integrated systems directly ***supporting scientific decision making*** and situation awareness by dynamically integrating, correlating, fusing and analysing extremely large volumes of disparate data resources and streams. All these systems are ***based on the machine learning paradigms*** (both supervised and unsupervised), enabling the self-adapting, generalization and automation capabilities ***to explore and mine data***.

Overview of the WP5 activities

WP5 main tasks

- **TASK 3:** INAF – Astronomical Observatory of Catania

3D Visual Analytics Systems

This task will implement a ***3D-aided visual analytics environment*** allowing the astronomer to easily conduct research activities using methods for multidimensional data and information visualization real-time data interaction to carry out complex tasks for multi-criteria data/metadata queries for ***subsample selection and further analysis***, or real-time ***control of data fitting to theoretical models***

- **TASK 4:** MTA – SZTAKI

Science Gateway

According to the needs of the astrophysics user community the Science Gateway with its features will be deployed. Workflows and portlets are also provided by this task with the ***support of the whole collaboration***.

Science & Technology

Most of the first period has been spent to find a common language among members…

How astronomers see astroinformaticians

How astroinformaticians see astronomers

…with doubtful but promising results

Overview of the WP5 activities

Desktop Client

INAF-OATS

Knowledge base DB

INAF-IAPS

SZTAKI

Science Gateway

Data Mining

INAF-OACN

INAF-IAPS

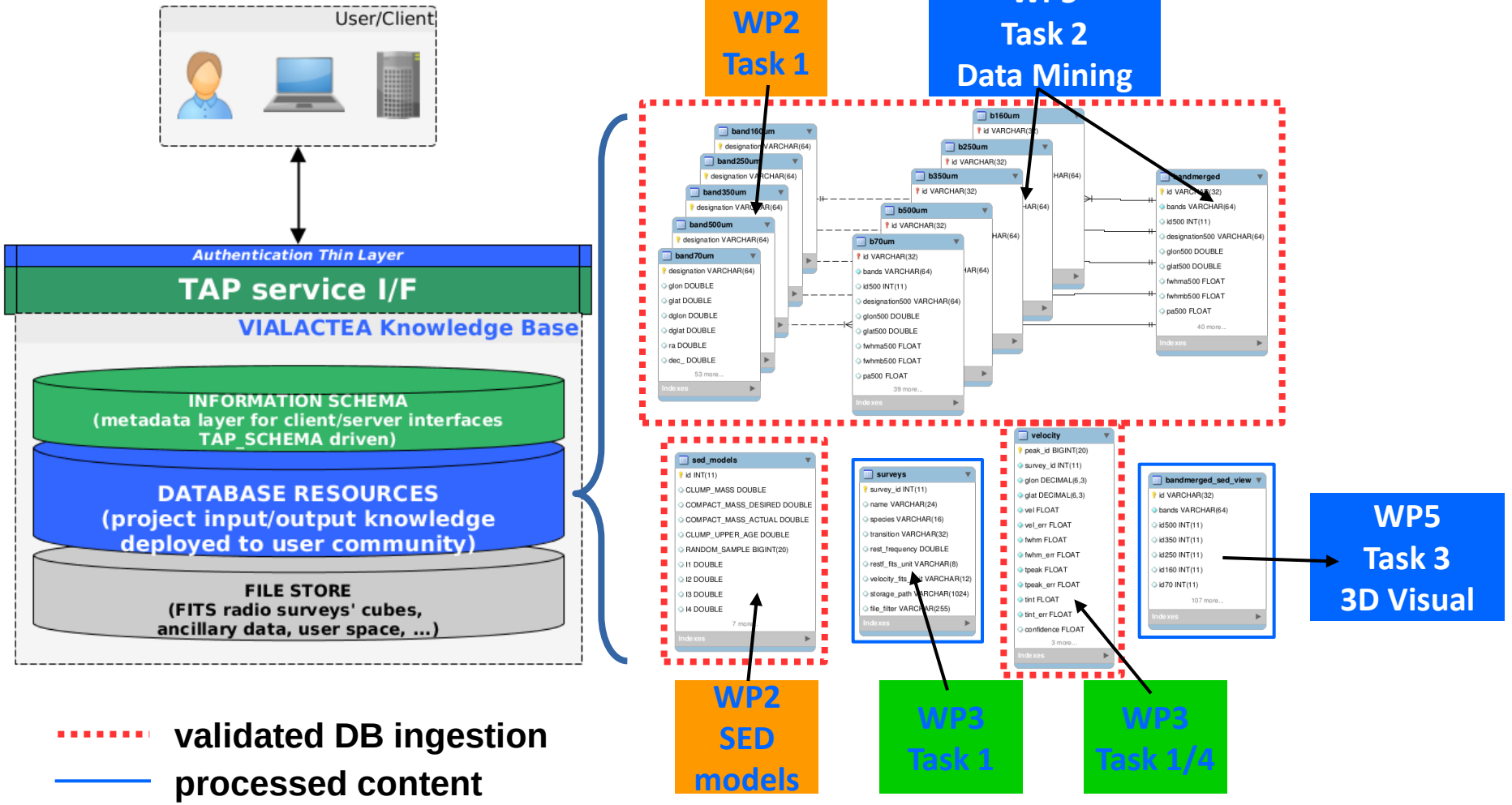
INAF-OACT

Computing Resources

Overall system architecture

The data flow and management system

VLKB main interface (I/F): IVOA Table Access Protocol (TAP) service



http://palantir19.oats.inaf.it:8080/vlkb

The Science Gateways

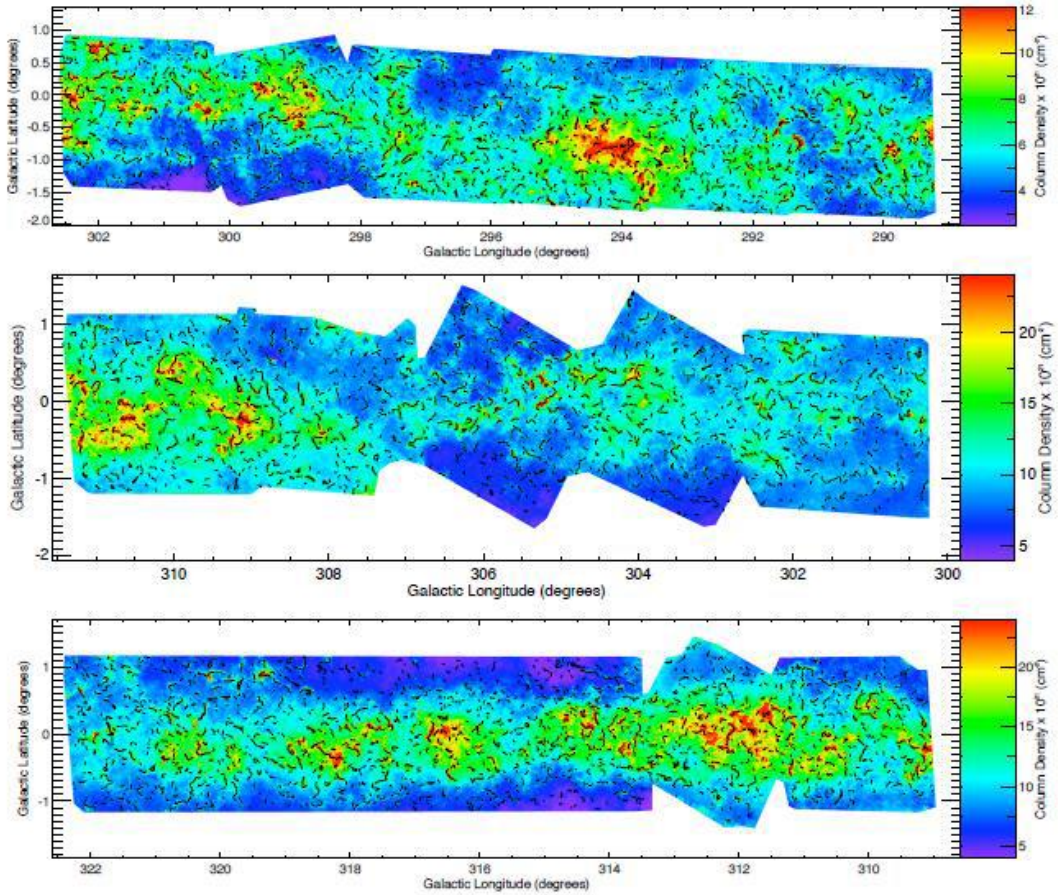
- ✓ Two science gateways had been set up and operated:
  - VIALACTEA Project Science Gateway – v0 in Catania
  - VIALACTEA Project Science Gateway – v1 in Rome
- ✓ Both conform to STARnet alliance (common authentication, resource sharing, etc.)
- ✓ Connected to PBS clusters of Catania and Rome
- ✓ VIALACTEA Project Science Gateway – v1 is based on the latest WS-PGRADE/gUSE release (series 3.7)
- ✓ Improvements of gUSE are continuously integrated
- ✓ Monitor and test: a new plan will be implemented

Let's see a video showing some details about the Science Gateway infrastructure

WP1-task1 filamentary structure detection

Task 1: Filamentary structure detection

- ❑ Production of column density maps of entire galactic plane
- ❑ Automated filament extraction workflow for Hi-GAL survey



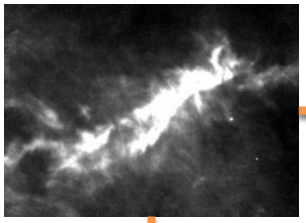
The filament extraction code was run on the column density maps covering the region between Galactic longitude 290° -- 320°, with different threshold levels equal to 2.5, 3. and 3.5 times the local standard deviation of the minimum eigenvalue (*Schisano et al.*, 2014)

OACN Data Mining goal:

- ❖ To refine the edges;
- ❖ To extend filament regions.

The right calculation of physical quantities related to filaments strongly depends on their dimensions, so the correct definition of edges is crucial.

Overview of the filament areas of interest



Original image

EXECUTION

TRAINING + EXECUTION

Intermediate mask

TRAINING

Traditional thresholding gradients

FilExSeC

Filament detection

Filament optimization

Work in progress

Beam curve detection

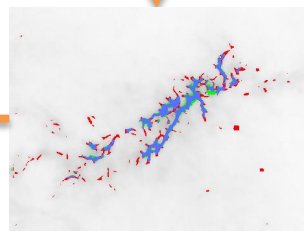
Structured Forest

SVM

Fuzzy detection

OR

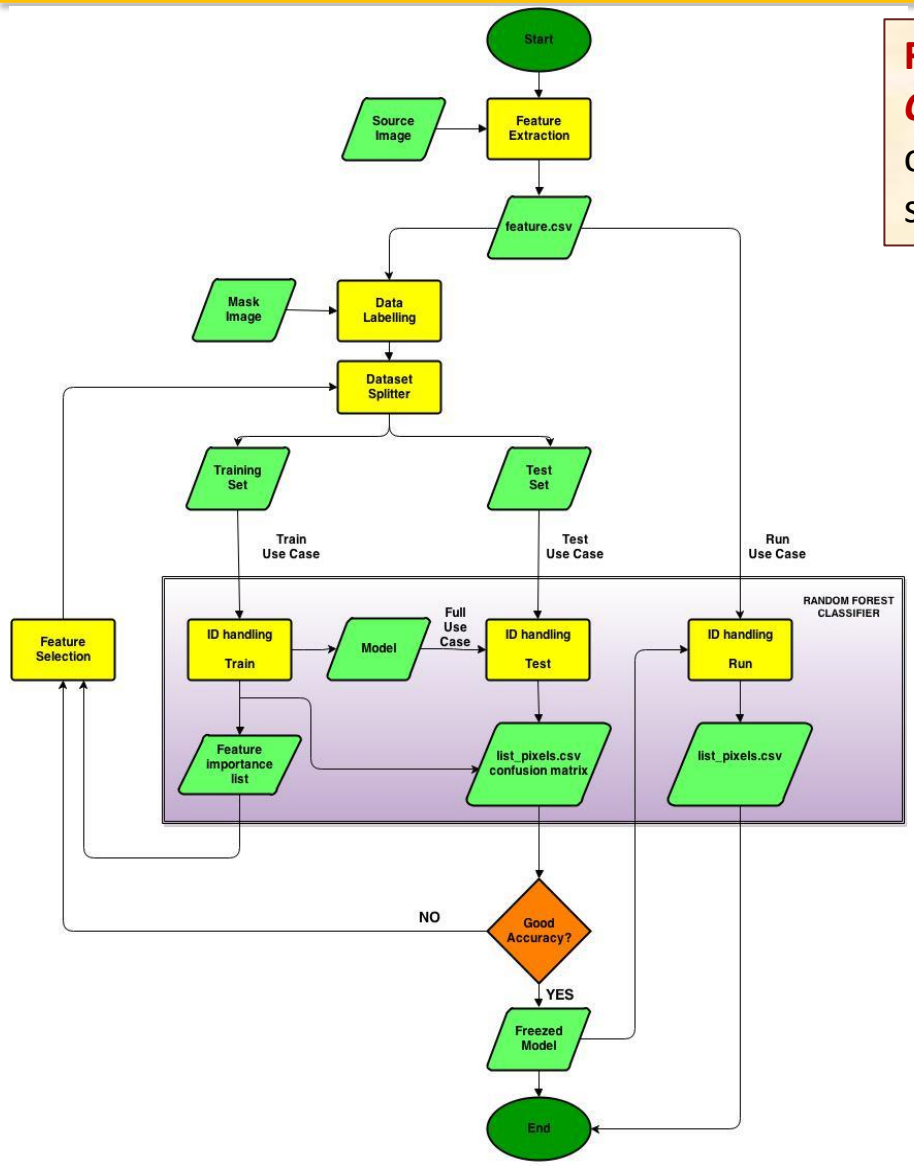
To combine different filament detection methods to improve the global performances



Spine Extraction Based on Z-S method

Filament Flux Calculation

FilExSeC algorithm



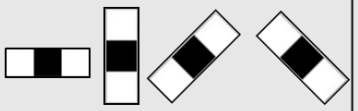
FilExSeC (*Filaments Extraction, Selection and Classification*), a data mining tool to refine and optimize the detection of the edges of filamentary structures.

The method consists in two main phases:

- ***Feature Extraction***: a set of features depending by its neighbors is associated to each pixel of the input image
- ***Classification***: image pixels are classified as filamentary or background, by using a supervised Machine Learning method, trained by these features

A further phase, ***Feature Selection***, finds the most relevant features. By reducing the initial data parameter space, it is possible indeed to improve the execution efficiency of the model, without affecting its performances.

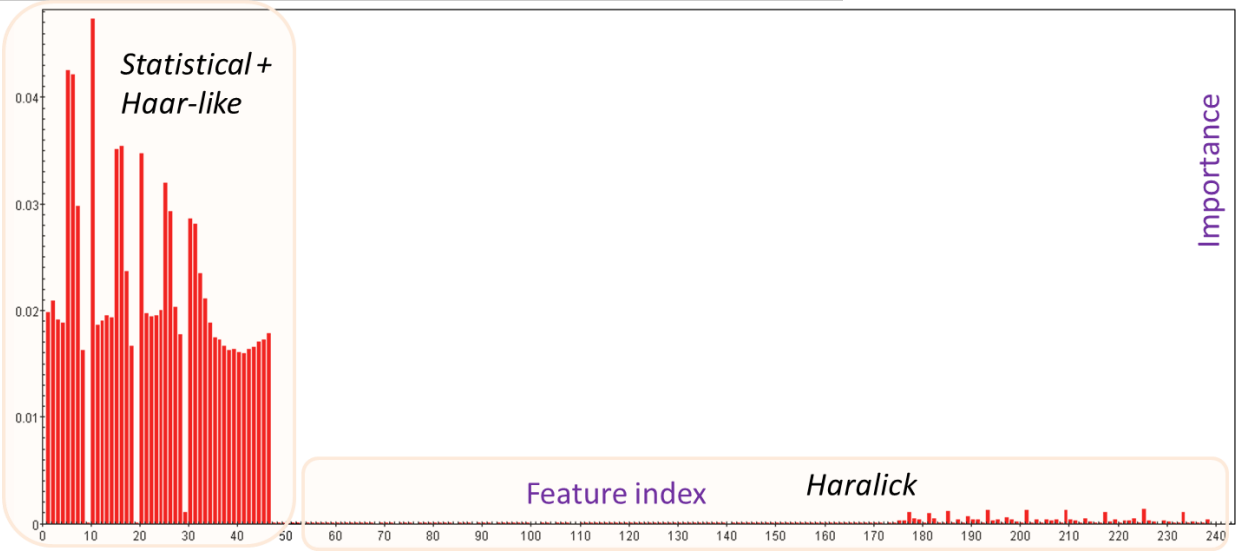
FilExSeC – pixel feature analysis

Type	Parameters			Features
Haar-like (158)	Name	Template	Dimensions	Difference between "black" and "white" rectangles
	Black rectangle		2x2 to 24x24	
	Black rectangle		2x4 to 12x24	
	Black rectangle		1 to 24	
Haralick (192)	$\lvert\vec{d}\rvert$ =1,2,3,4, directions = 0°, 45°, 90°, 135° windows = 5x5 – 7x7 – 9x9			Contrast, Energy, Entropy, Correlation
Statistical (41)	windows = 3x3 – 5x5 – 7x7 – 9x9			Gradients (vert., horiz., diag.), Mean, Stdev, Skewness, Kurtosis, Entropy, Range
	windows = 1x1			Pixel Value

Features extracted from each pixel and its neighbors

Feature selection:

Haralick type excluded (no information lost and improved computing time)

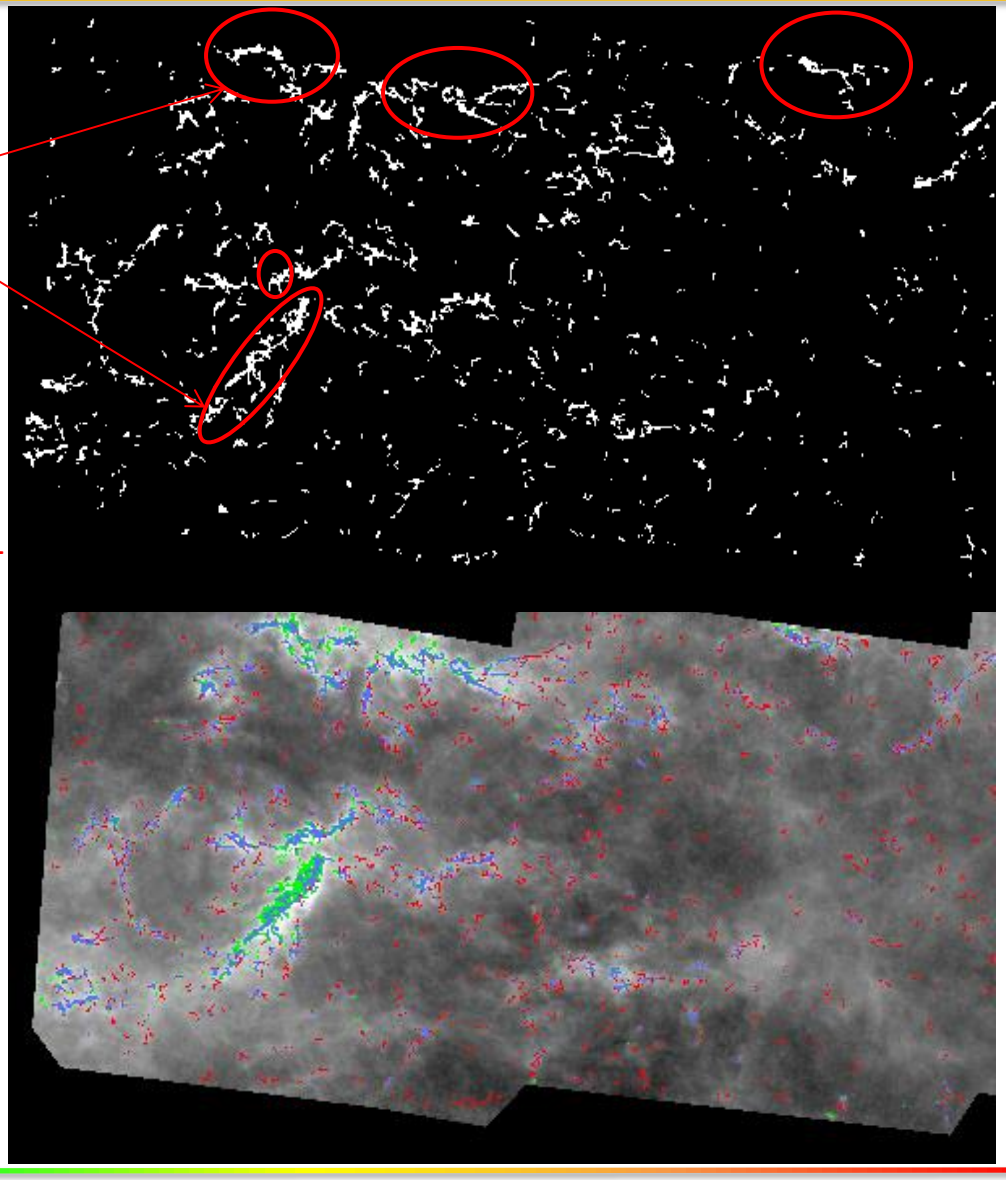


FilExSeC – Filament Connections

FilExSeC is able to connect, by means of NFPs, filaments that in the traditional method are tagged as disjointed objects.
By joining filaments as a unique structure total mass and mass per length change, inducing a different physics of the filamentary structure.

Detected Filaments	**668**
Confirmed Filaments	**298**
New Filaments	**196**
Extended Filaments	**169**
Joined Filaments	**5**

EXAMPLE:
TEST tiles 1+2
of Hi-GAL l217-l224

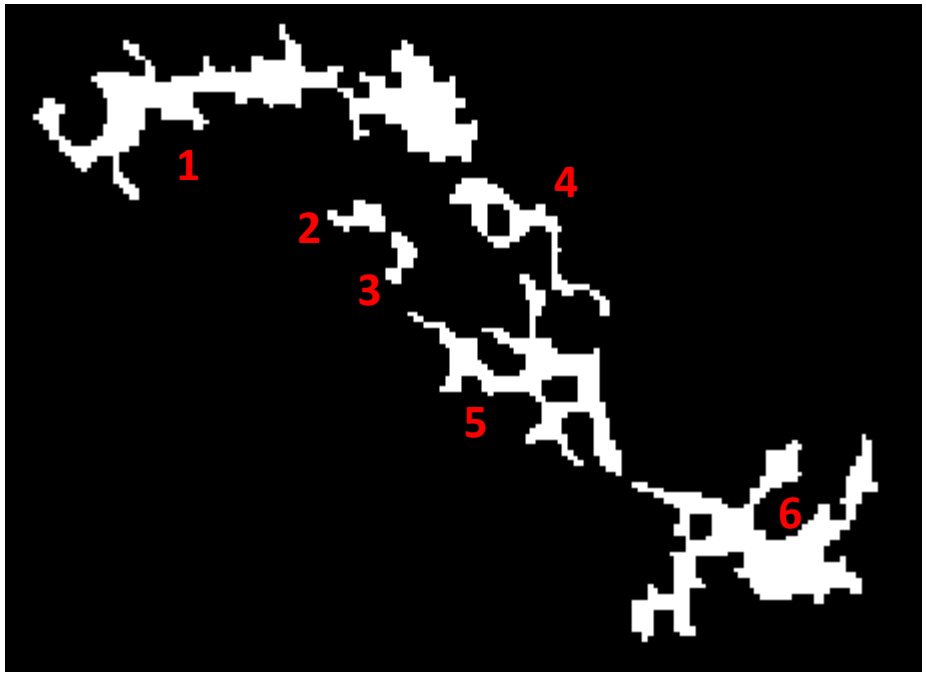


A further analysis is required to verify the correctness of the reconstruction of interconnections between different filaments, to evaluate the contribution of FilExSeC to the knowledge of the physics of the filaments.

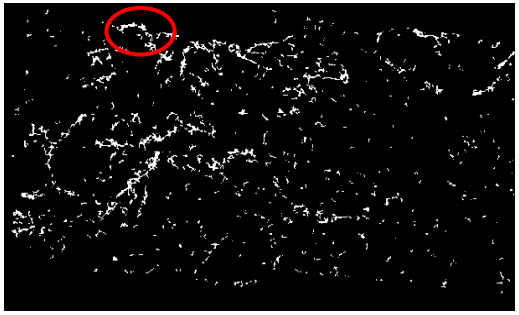
- Confirmed Filament Pixels
- New Filament Pixels
- Confirmed Background
- Undetected Filament Pixels

FilExSeC – Example of Joined Filaments

IAPS



FilExSeC



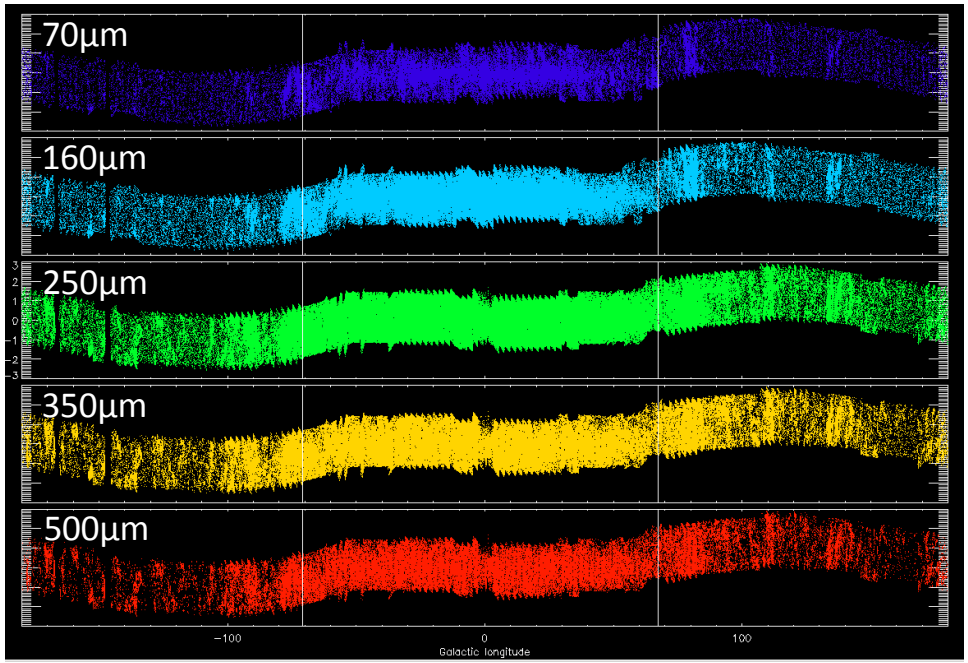
Connection of 6 filaments identified by IAPS

IAPS total number of pixels: 1852
Pixel added by FilExSeC : 858
New Total number of pixels: 2710
% NFP: +46.33%

WP2 – task 1 – Band merging

Task 1: Compact Source Extraction and band-merging

- ❑ Hi-GAL Source extraction and photometry
- ❑ Band-merging with ancillary information (from near-IR to radio)



A first result from OACN of a **band-merged catalogue** using a data-mining approach has been implemented for the Herschel bands

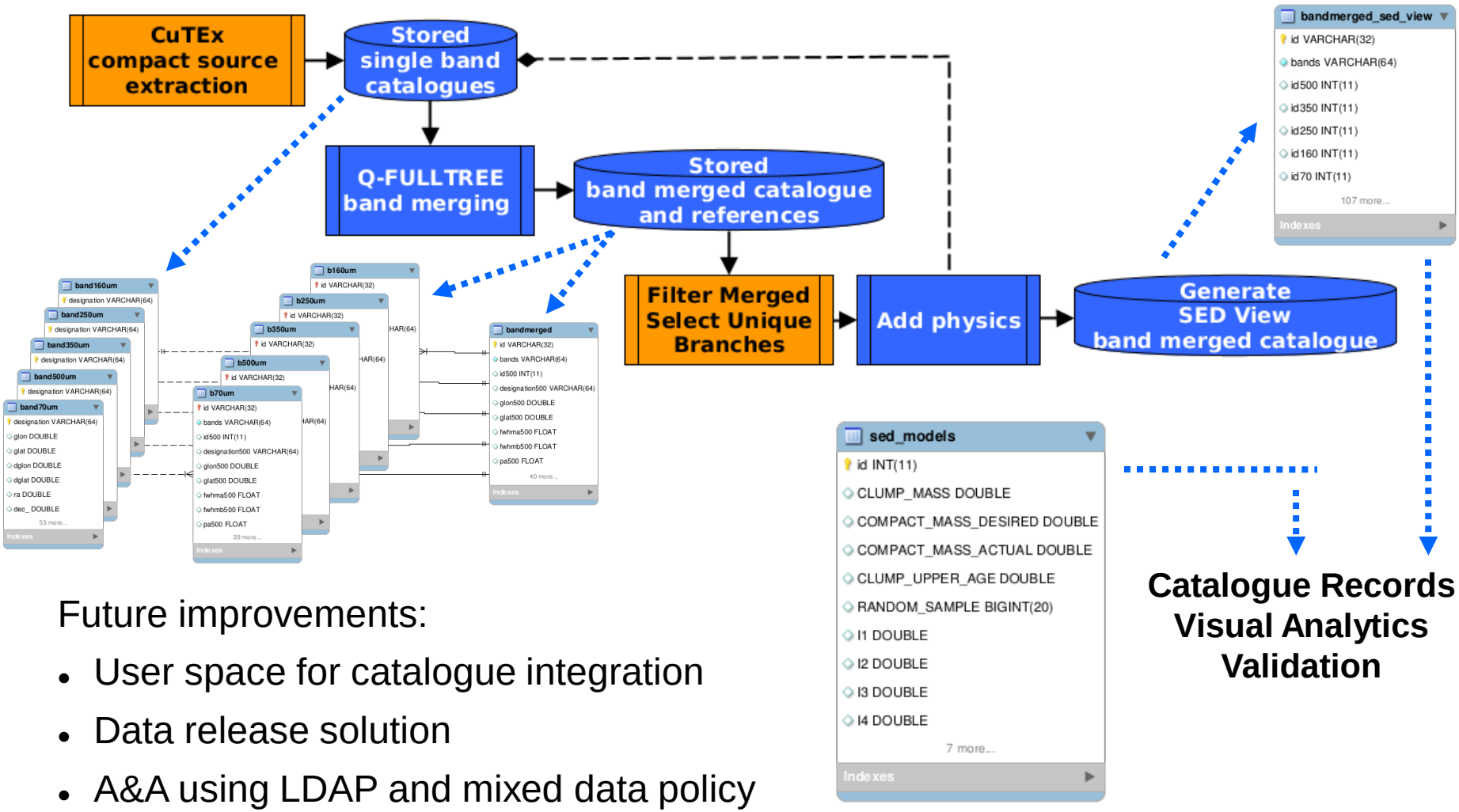
The **source extraction** with CuTEx (***Molinari et al.*, 2010a**) has been run over the entire Galactic plane.

The -71° < *l* < 67.5° portion of the HERSHEL/Hi-GAL photometry lists should be band-merged, filtered and complemented with distances and ancillary photometry : MIPSGAL, UKIDSS, WISE, MSX; ATLASGAL, BGPS …

- ❖ Captures and maintains multiple counterpart associations;
- ❖ Topological quality flagging;
- ❖ Ingested into a VO-like database so that complex queries are possible;
- ❖ Interfaced with Visualization tools;
- ❖ Massively based on multi-threading parallelization.

The workflow for compact source analysis

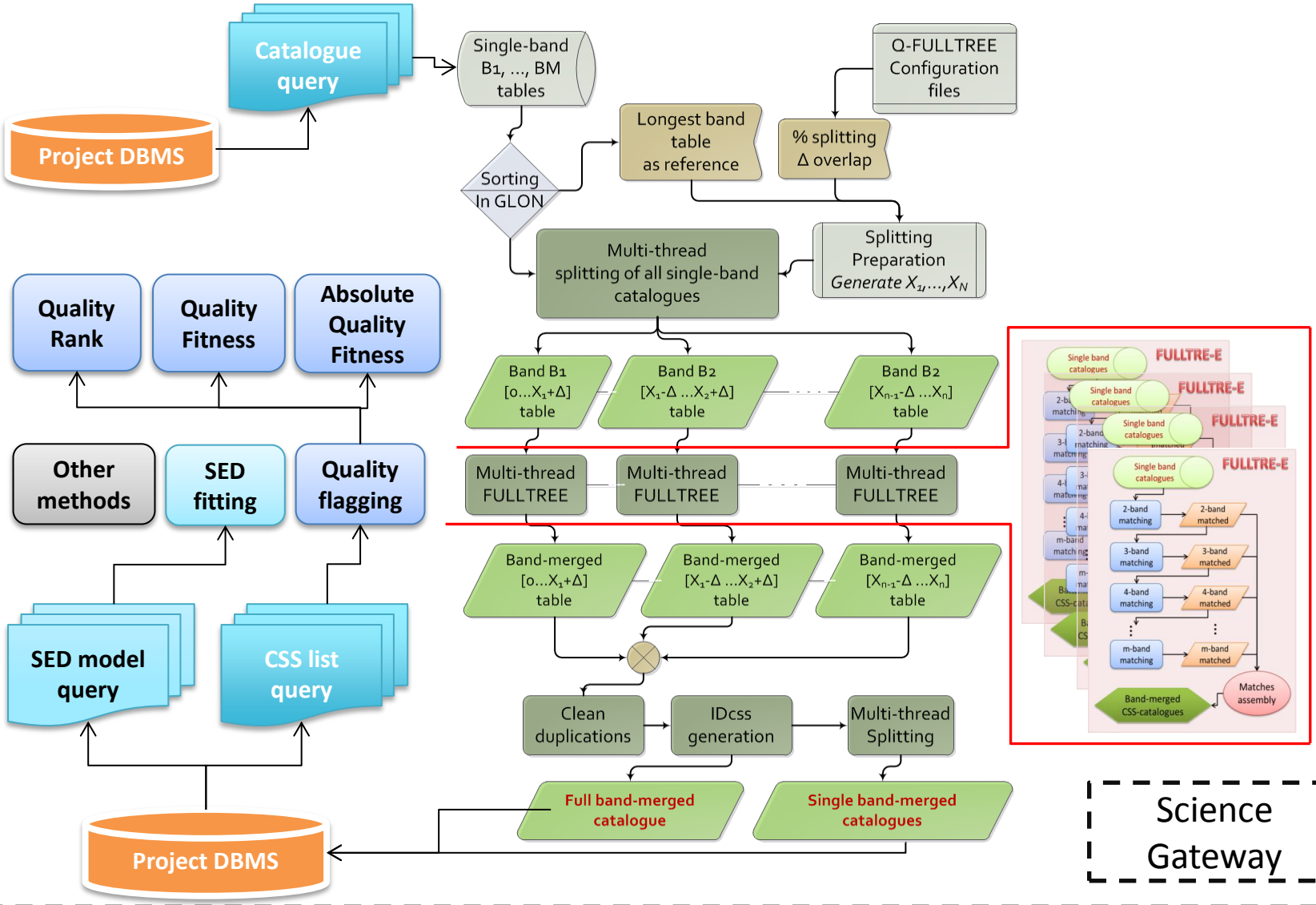
Compact Sources Band Merged Catalogue w/ Sources SED View



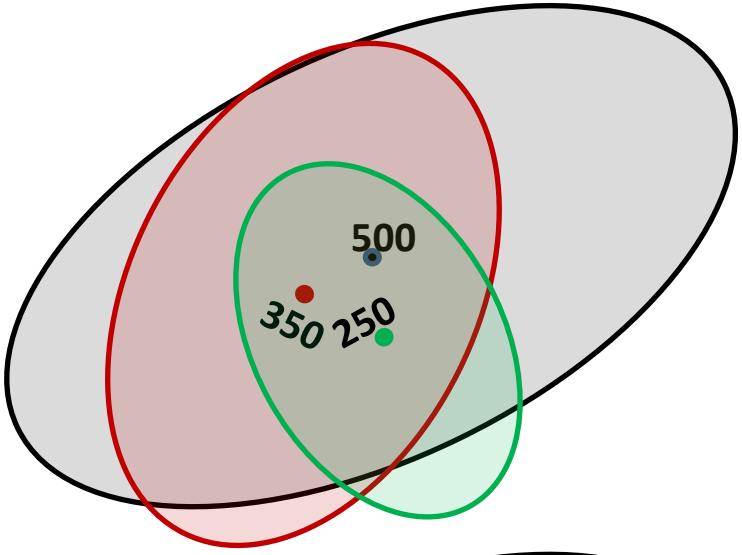
Future improvements:

- User space for catalogue integration
- Data release solution
- A&A using LDAP and mixed data policy

Q-FULLTREE processing flow



Q-FULLTREE Example



CSS_1

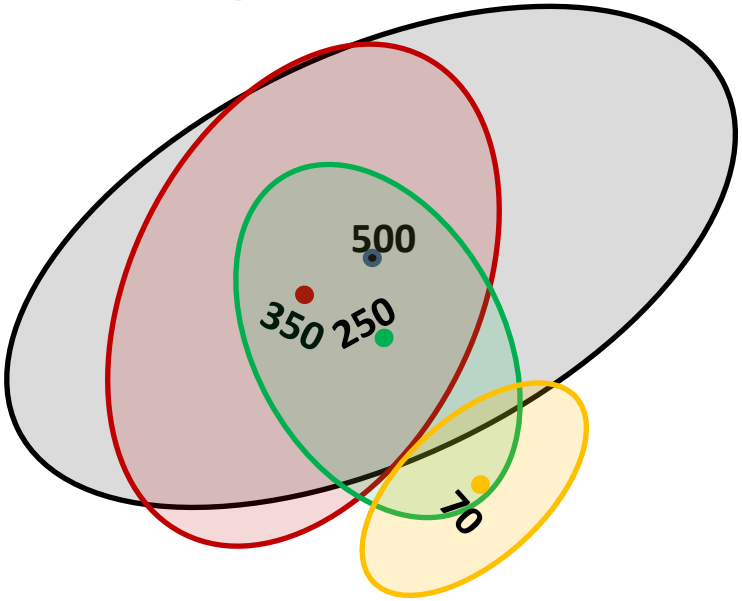
$CL_{500-350} = 0.91$
$CL_{500-250} = 0.87$
$CL_{350-250} = 0.89$

$\frac{NE}{TNE} = \frac{3}{3} = 1$

$MS_1 = 2.67$

$MS_2 < MS_1$

$MS_2 = 1.79$

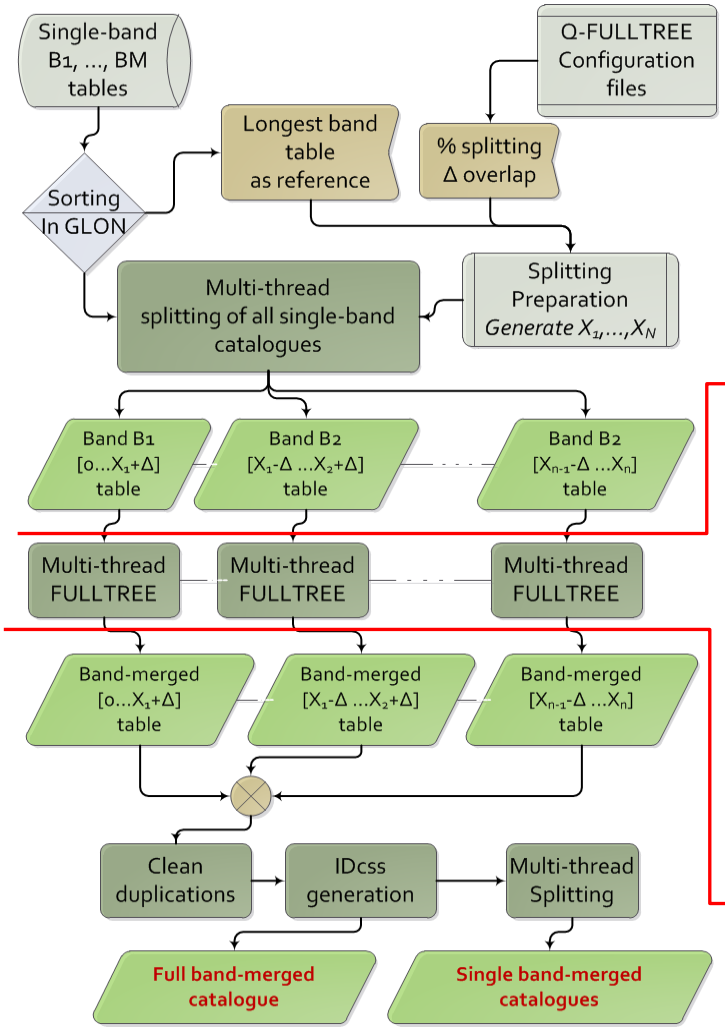


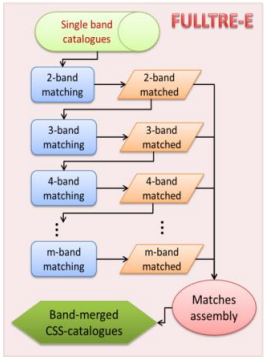
CSS_2

$CL_{500-350} = 0.91$
$CL_{500-250} = 0.87$
$CL_{350-250} = 0.89$
$+$
$CL_{250-70} = 0.02$

$\frac{NE}{TNE} = \frac{4}{6} = \frac{2}{3}$

WP2-T1 Q-FULLTREE infrastructure aspects





5 bands:
on a bi-CPU 1.6GHz, 16 cores:
FULLTREE 27 days → Q-FULLTREE 3.3 hours
On a quad-CPU 2.4GHz, 32 cores:
FULLTREE 23 days → Q-FULLTREE 1.3 hours
On CT cluster (1 CPU 2.4 GHz, 12 cores):
FULLTREE 29 days → Q-FULLTREE 3,15 hours

PERFORMANCES
Worst gain in speedup when compared with single-thread FULLTREE: 200x (mostly higher)

WP2–T4 Evolutionary classification

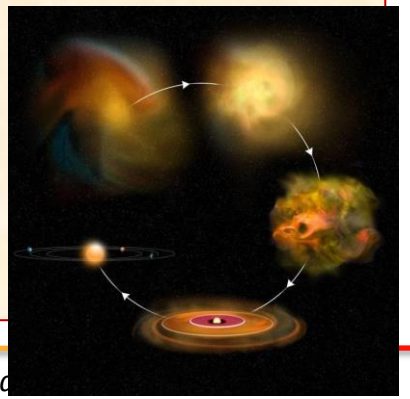
An evolutionary classification tool for ViaLactea, will catalogue “clumps” in terms of the evolutionary stage and mass regime of the ongoing star formation. There are two components that need to be developed at the foundation of the classification tool:

1. an evolutionary classification toolbox
2. a set of star-forming clumps in known stages of evolution to be used as a training/test-set for machine-learning algorithms… ...and adopt some kind of evolutionary scheme

Data-mining approaches to source classification
FOREIGN (Forming Region Exploring Intelligent Gated Network)

Weak Gated Classification
We know nothing about the sources evolutionary stage;
Identify over-densities in the given parameter space (e.g., built on the evolutionary toolbox, plus any other available evidence);
Data are then grouped into clusters: groups of data entries sharing common but *a priori* unknown correlations among parameter space features.

Supervised Classification
For a subsample of points, its category/class is well known;
Need order of 10^3 objects to be used as a training set;
Balanced population of classes in the training set.



WP2-T4 Evolutionary Classification (1st)

Parameter space

- Luminosity/mass diagram
- IR color distribution
- Mass/radius relationship
- Association with masers
- Association with radio continuum
- Outflow indicators
- YSO IR-color spatial association
- Others…?

class partitioning inferred by science experts

supervised classification

ambiguity

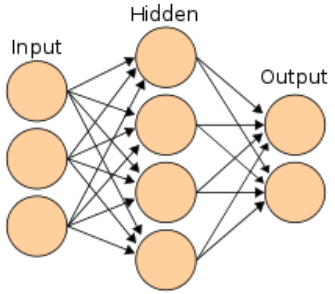
SF clumps – source class prediction – Knowledge Base

Quiescent / pre-stellar | Proto-stellar | Hot molecular cores | Ultra-compact HII regions | Evolved HII regions

Train set

Validation set

Test set



estimation

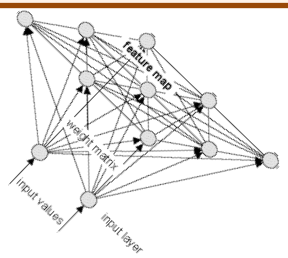
Fuzzy/cross-entropy

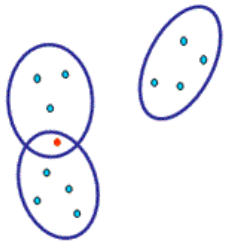
WP2-T4 Evolutionary Classification (2nd)

Parameter space

- Luminosity/mass diagram
- IR color distribution
- Mass/radius relationship
- Association with masers
- Association with radio continuum
- Outflow indicators
- YSO IR-color spatial association
- Others…?

unsupervised clustering





validated by science experts

ambiguity

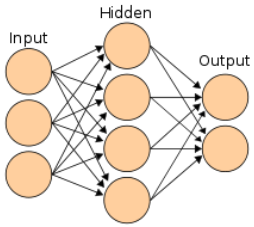
SF clumps – source class prediction – Knowledge Base

Quiescent / pre-stellar | Proto-stellar | Hot molecular cores | Ultra-compact HII regions | Evolved HII regions

Train set

Validation set

Test set



estimation

Fuzzy/cross-entropy

Weak gated classification

Conclusions

The whole project has successfully passed the mid-term official EU commission review

The initial inertia due to interaction problems between technology and science communities is going to be successfully overcome

The data and computing infrastructures and visual analytics solutions started to host and integrate the planned scientific workflows, matching the expected capabilities

The data mining paradigms are demonstrating their expected benefit to help the scientific problem solving automation as well as to manage the foreseen amount and complexity of data

In other words

The project at mid-term stage (April 2015) is respecting the initial goals, among which the WP5 expectation to release a useful resource for the wide scientific community, which will remain available also after the project closure (October 2016).