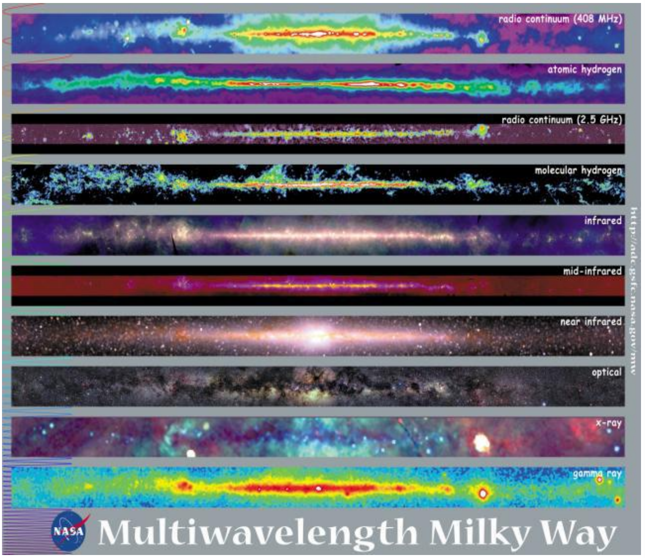
Machine learning based data mining for Milky Way filamentary structures recognition

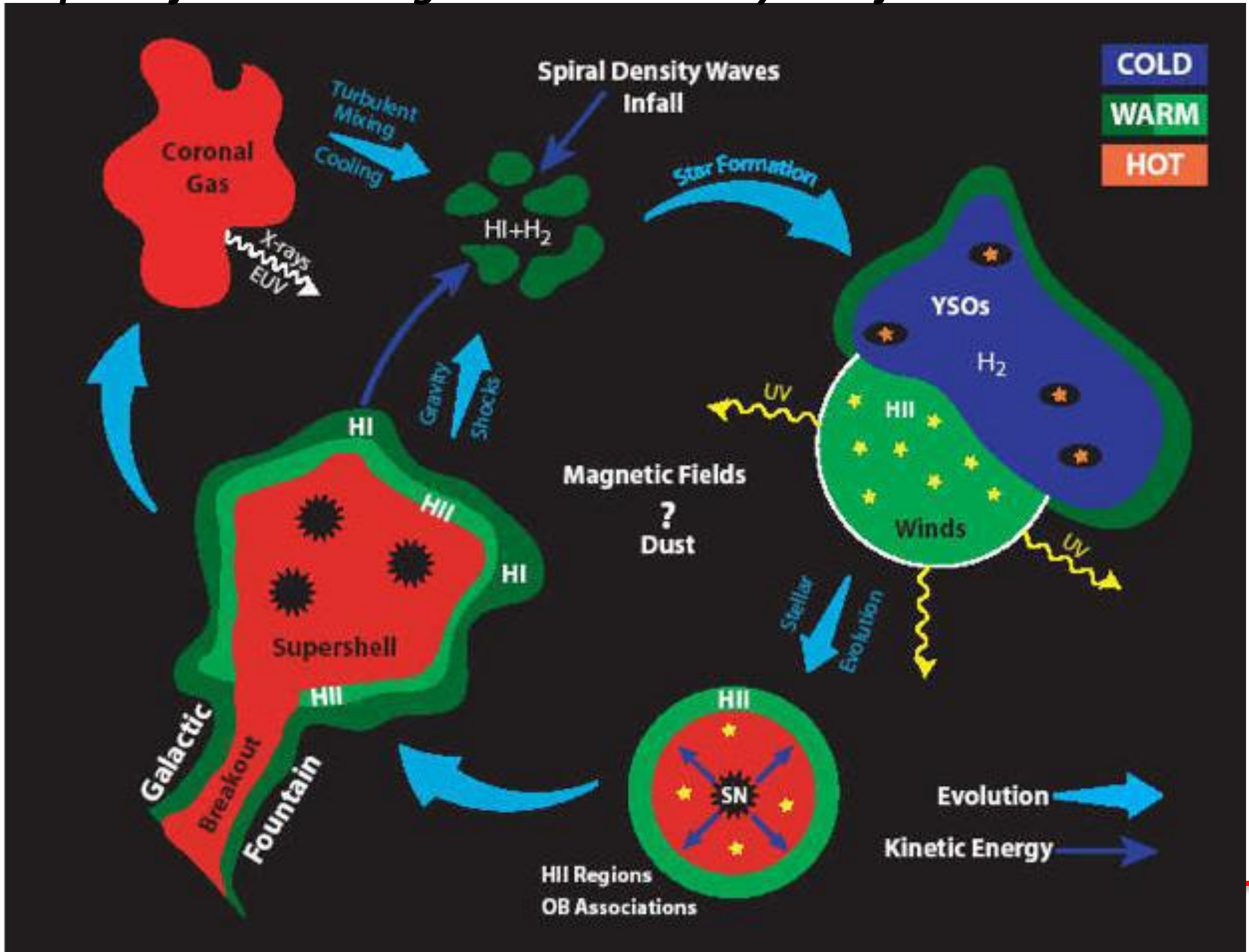
Giuseppe Riccio[1], **Stefano Cavuoti[1]**, Massimo Brescia[1], Eugenio Schisano[2], Amata Mercurio[1], Davide Elia[2], Milena Benedettini[2], Stefano Pezzuto[2], Sergio Molinari[2], Anna Maria Di Giorgio[2]

(1) INAF – Astronomical Observatory of Capodimonte, Via Moiariello 16, I-80131 Napoli, Italy
(2) INAF – IAPS, Via del Fosso del Cavaliere 100, I-00133 Roma, Italy

The role of WP5 in the project

The goal is to exploit the combination of all the new-generation Infrared → Radio surveys of the Galactic Plane from space missions and ground-based facilities, using a novel data and science analysis paradigm based on 3D visual analytics and data mining framework, to build and deliver a quantitative 3D model of our VIALACTEA Galaxy as a star formation engine that will be used as a template for external galaxies and study star formation across the cosmic time

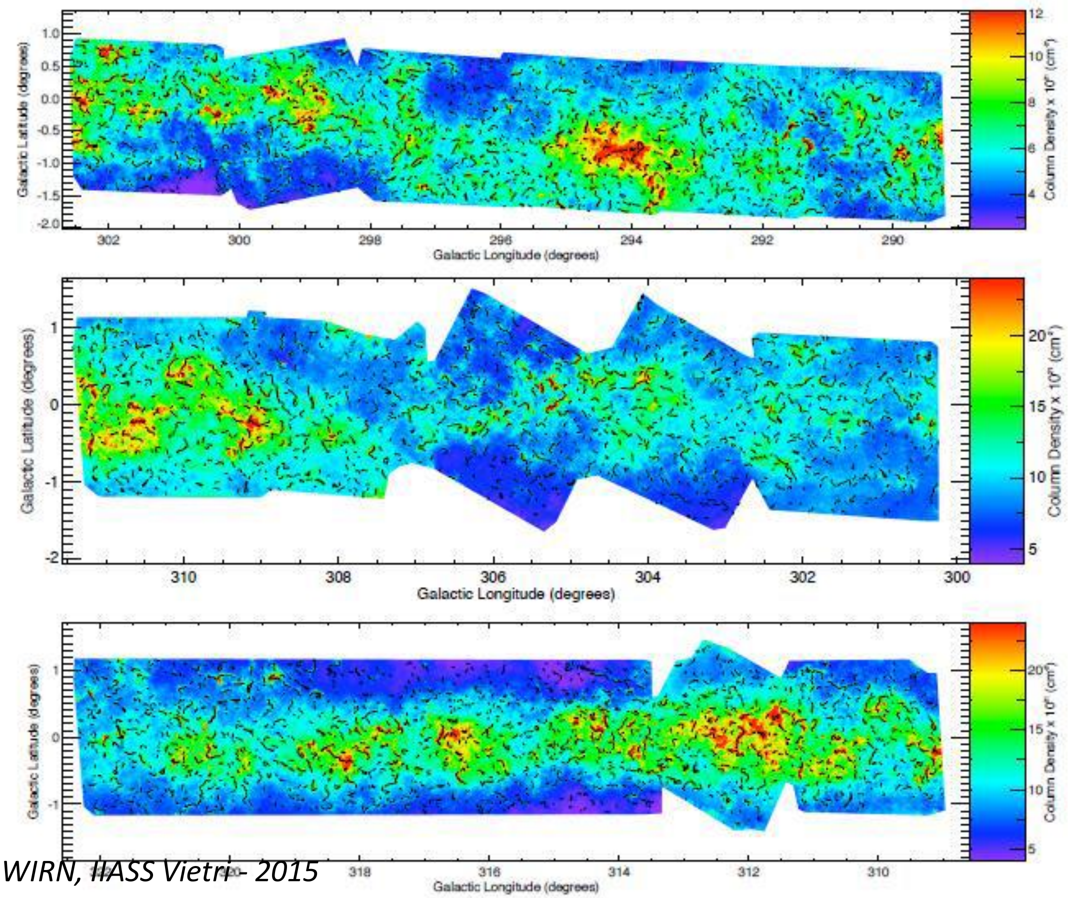




WP1-task1 filamentary structure detection

Task 1: Filamentary structure detection

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



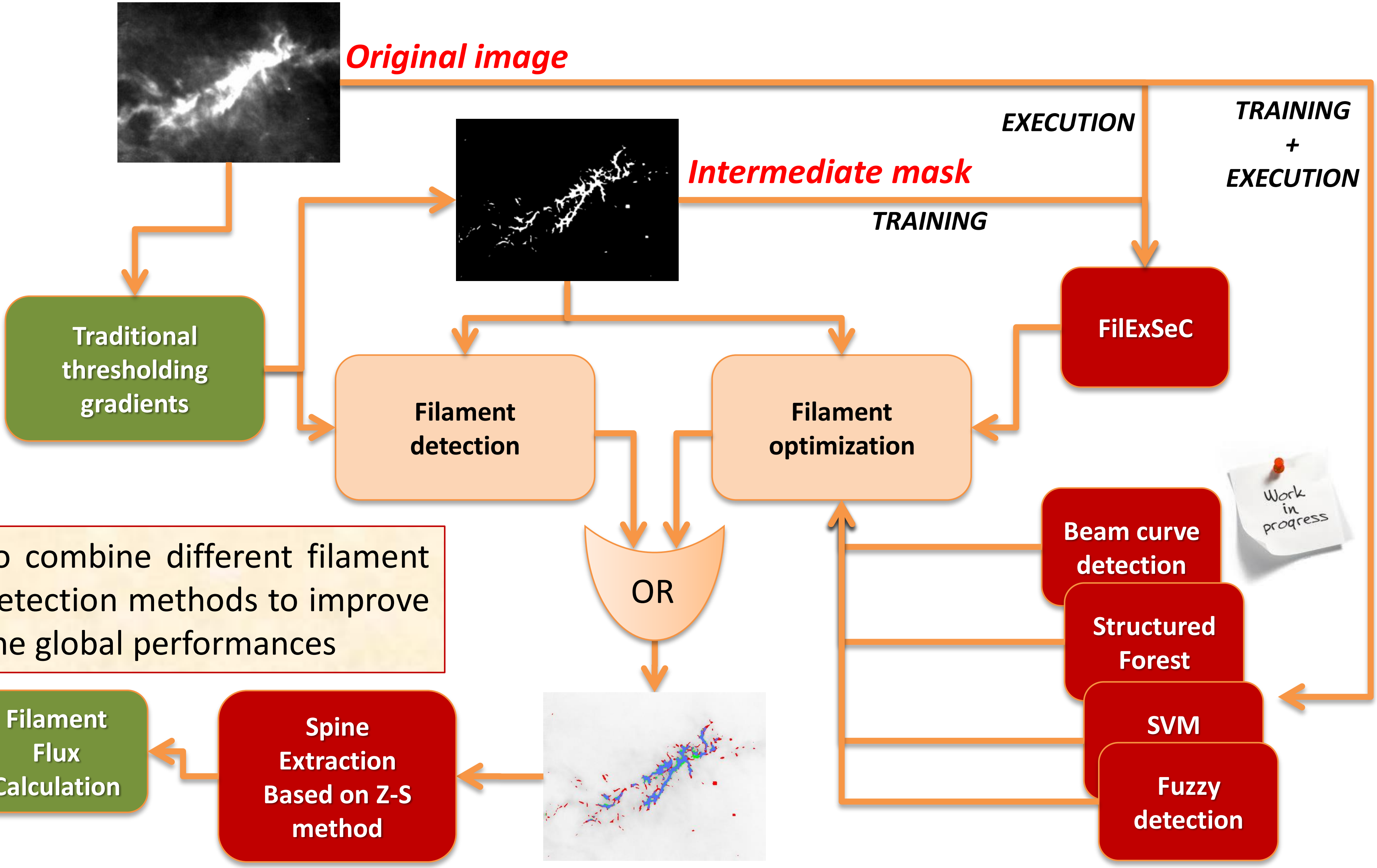
The filament extraction code was run on the column density maps covering the region between Galactic longitude 290° -- 320°, with different threshold levels equal to 2.5, 3. and 3.5 times the local standard deviation of the minimum eigenvalue (*Schisano et al.*, 2014)

OACN Data Mining goal:

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

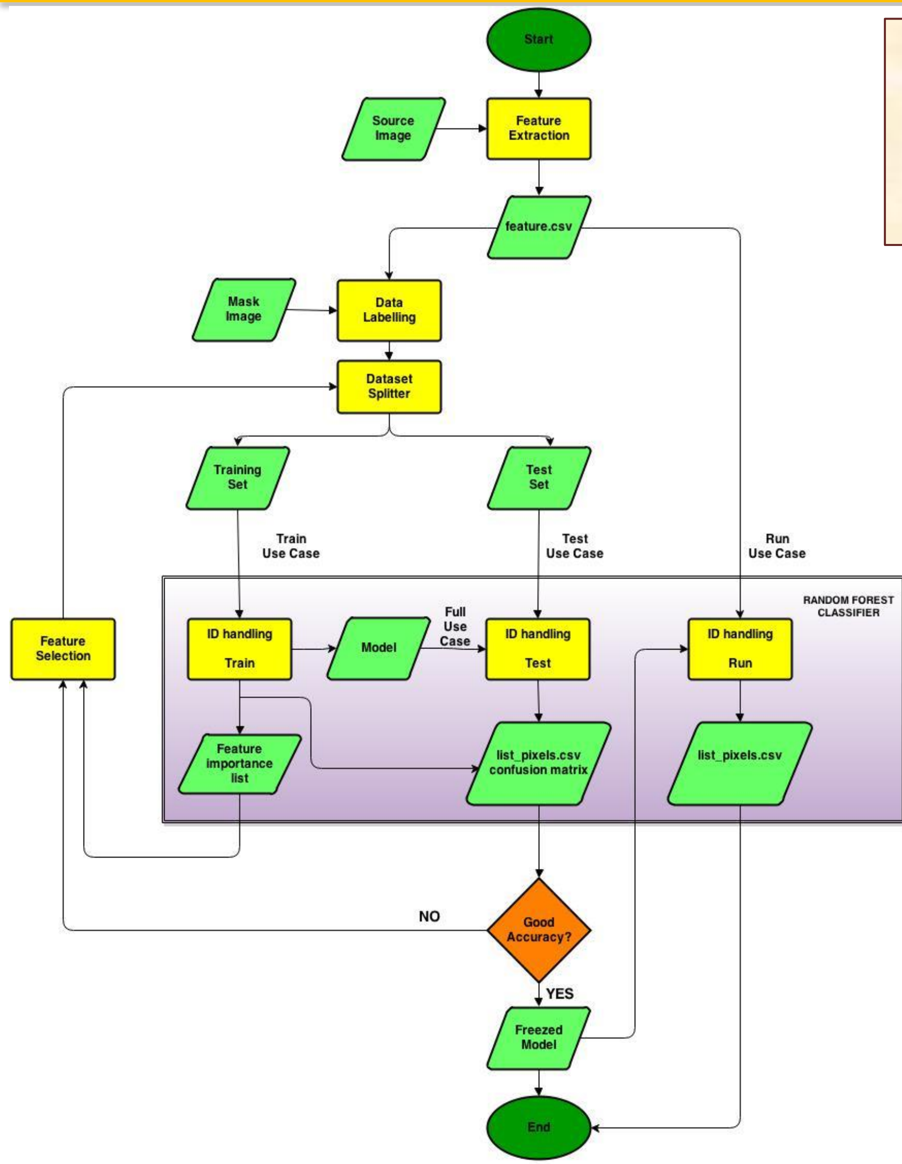
The right calculation of physical quantities related to filaments strongly depends on their dimensions, so the correct definition of edges is crucial.

Overview of the filament areas of interest



To combine different filament detection methods to improve the global performances

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.

Haralick feature space

Haralick Features (1979)

- Based on co-occurrence matrix (GLCM)
- Element $C_{i,j}$ represents, for a fixed distance and direction, the probability to have two pixels in the image at that distance, with grey level Z_i and Z_j, respectively

I

1	1	5	6	8
2	3	5	7	1
4	5	7	1	2
8	5	1	2	5

GLCM

	1	2	3	4	5	6	7	8
1	1	2	0	0	1	0	0	0
2	0	0	1	0	1	0	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	0
5	1	0	0	0	0	1	2	0
6	0	0	0	0	0	0	0	1
7	2	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0

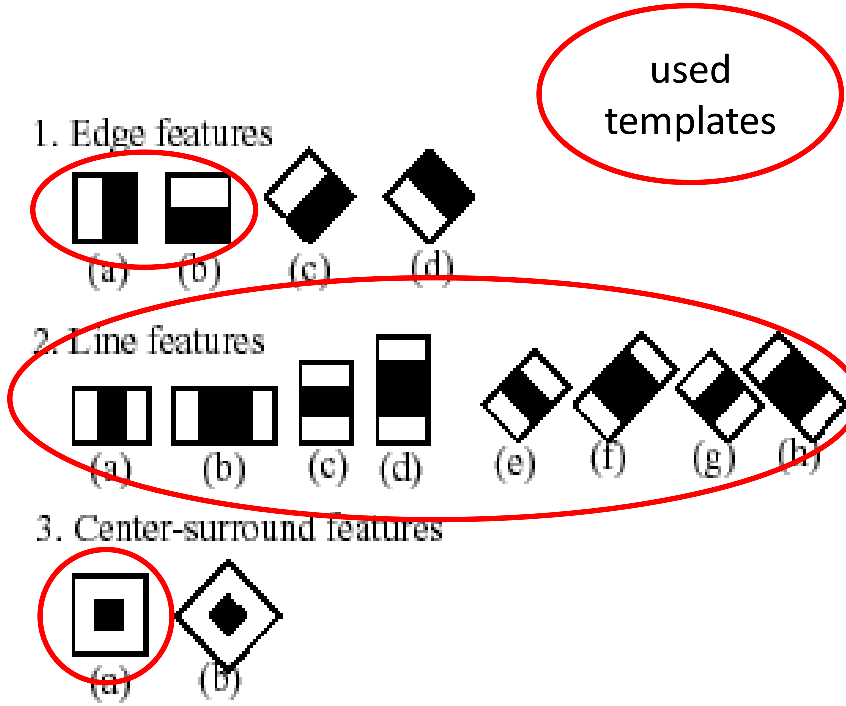
Haralick features extracted from $C_{i,j}$/(number of pairs)

Contrast	$m = \sum_i \sum_j (i-j)^2 C_{i,j}$
Energy	$\sum_i \sum_j C_{i,j}^2$
Entropy	$-\sum_i \sum_j C_{i,j} \log C_{i,j}$
Correlation	$\frac{\sum_i \sum_j (i-\mu_i)(j-\mu_j) C_{i,j}}{\sigma_i \sigma_j}$

Robert Haralick

Haar-like and statistical feature space

Alfred Haar



Haar-like Features (2001)

- Each Haar-like variable involves 2 or 3 interconnected black and white rectangles (masks or templates)
- Values of each feature are obtained by sliding masks on the image and calculating:

$$f = \sum_{\substack{black \\ rectangle}} \binom{grey\ level}{image\ pixel} - \sum_{\substack{white \\ rectangle}} \binom{grey\ level}{image\ pixel}$$

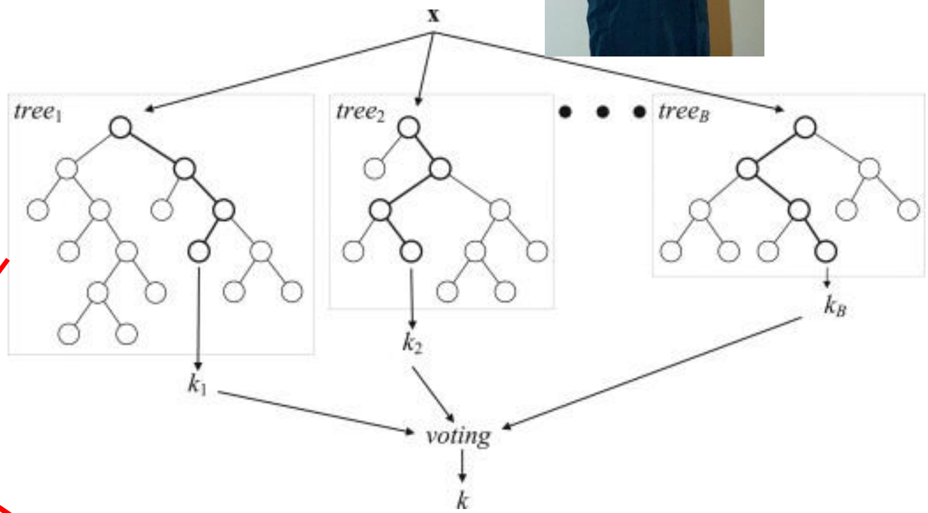
Statistical Features

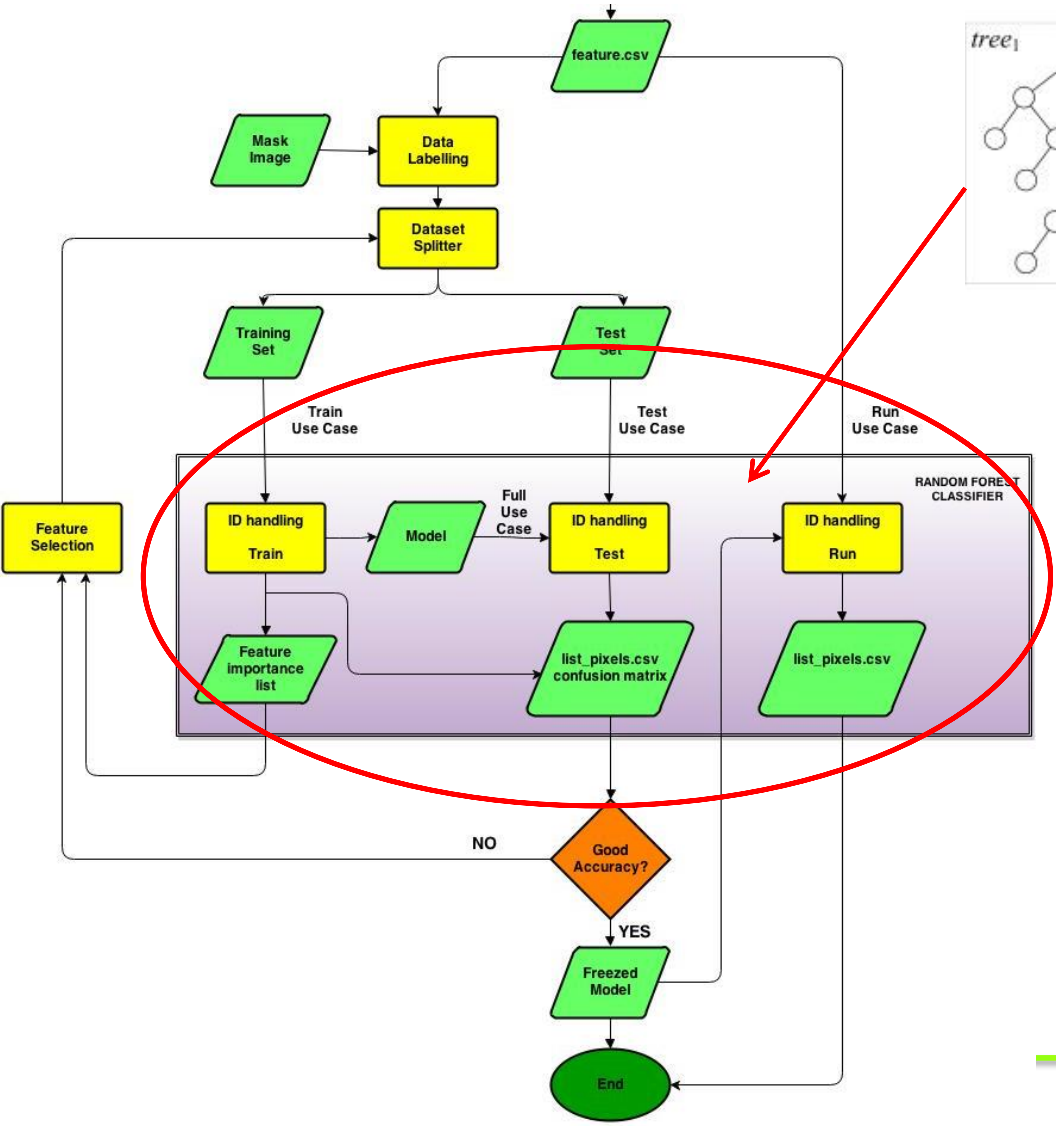
For each pixel, the following features are calculated in a centered window:

- gradients (horizontal, vertical, main diagonal, secondary diagonal)
- Mean, standard deviation, skewness, kurtosis, entropy, range

Moreover, the pixel value is considered as a Statistical Feature too

Leo Breiman





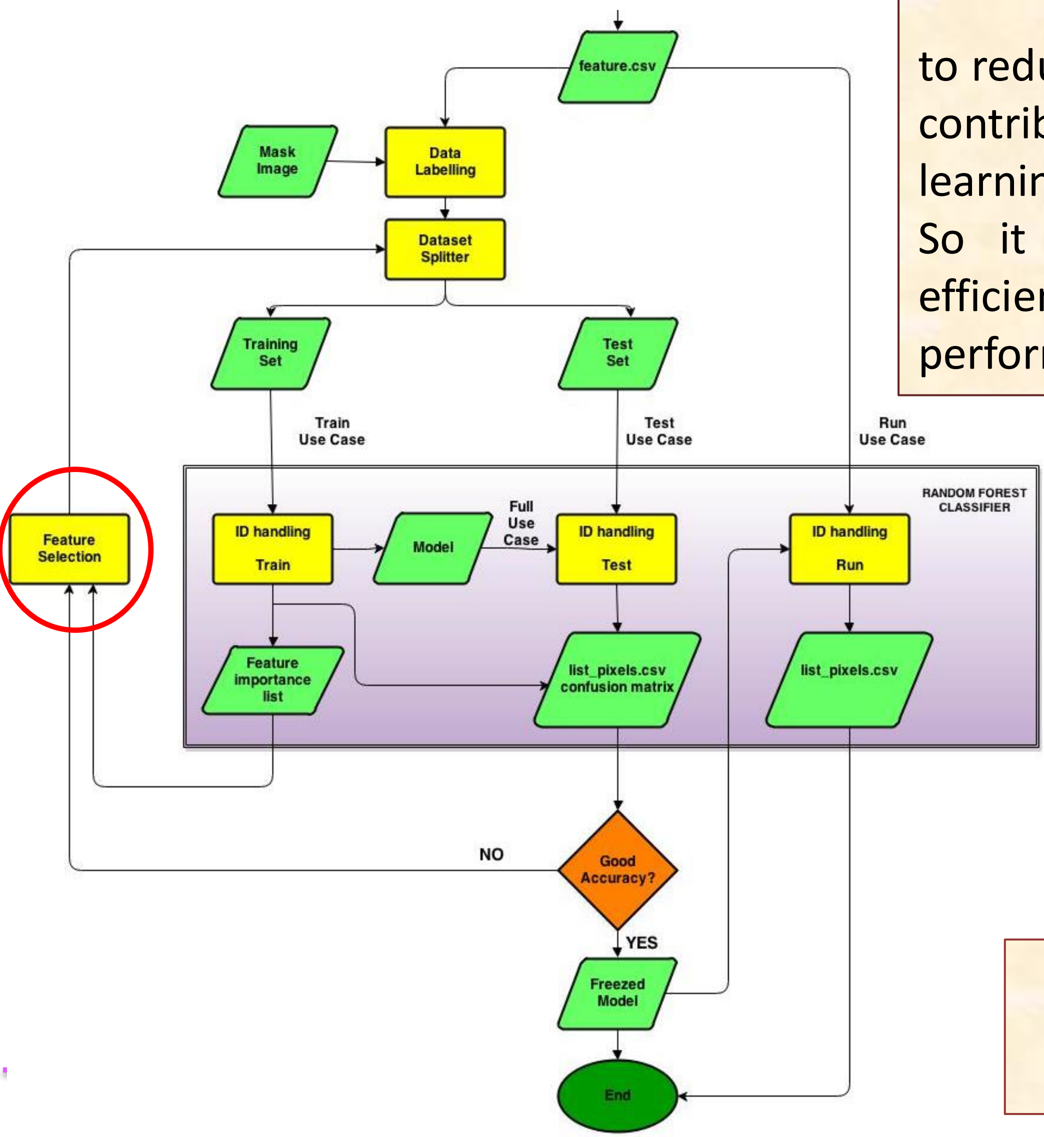
Random Forest Classifier (2001)

The classifier prefigures 4 use cases:

- **Train**: the classifier is trained to discriminate between filamentary and background pixels;
- **Test**: the classifier is evaluated by a blind test dataset;
- **Run**: the trained and validated classifier is used on new real images.

Overview of the WP5 activities

Mark A. Hall



Feature Selection (Backward Elimination 1999)

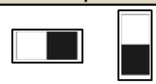
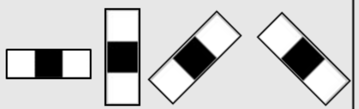
to reduce the parameter space, by weighting the contribution carried by each feature to the learning capability of the classifier.
So it is possible to improve the execution efficiency of the model, without affecting its performances.

At the end of this phase, a subset of features having higher weight (defined as *importance*) in recognizing filament pixels is considered.

This subset is then used to definitely train and test the classifier with new training and testing subsets.

Tests revealed that Haralick features are useless

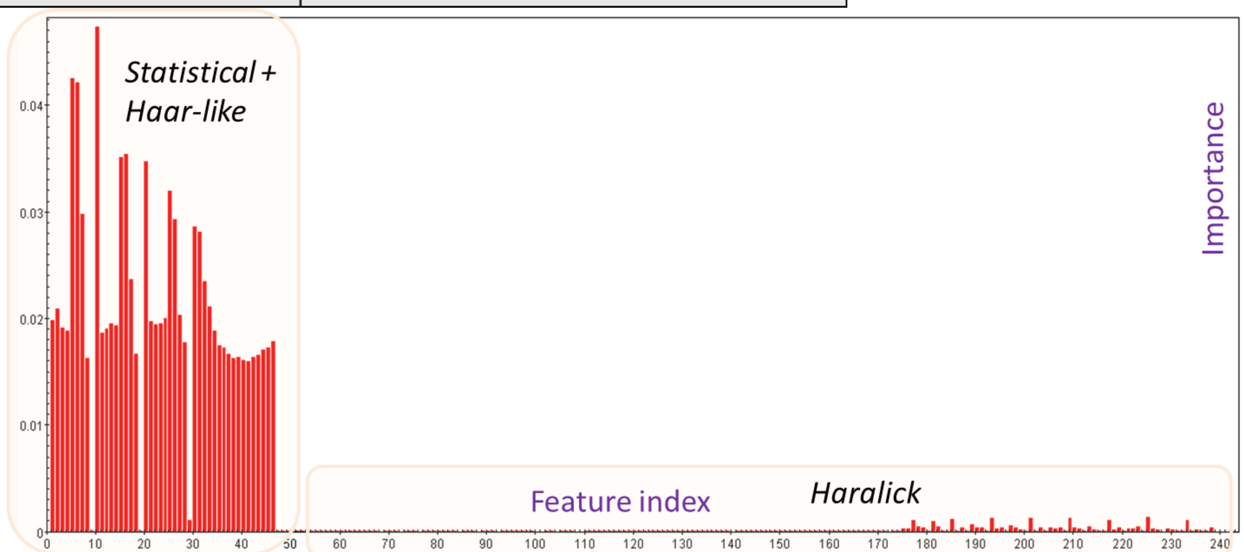
FilExSeC – pixel feature analysis

Type	Parameters			Features
	Name	Template	Dimensions	
Haar-like (158)	Black rectangle		2x2 to 24x24	Difference between "black" and "white" rectangles
	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)



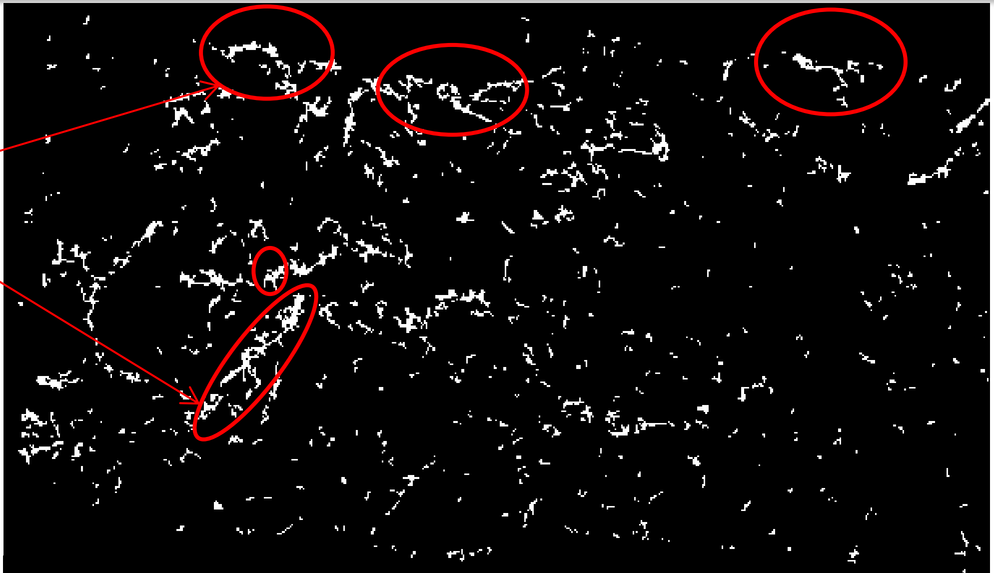
FilExSeC – Filament Connections

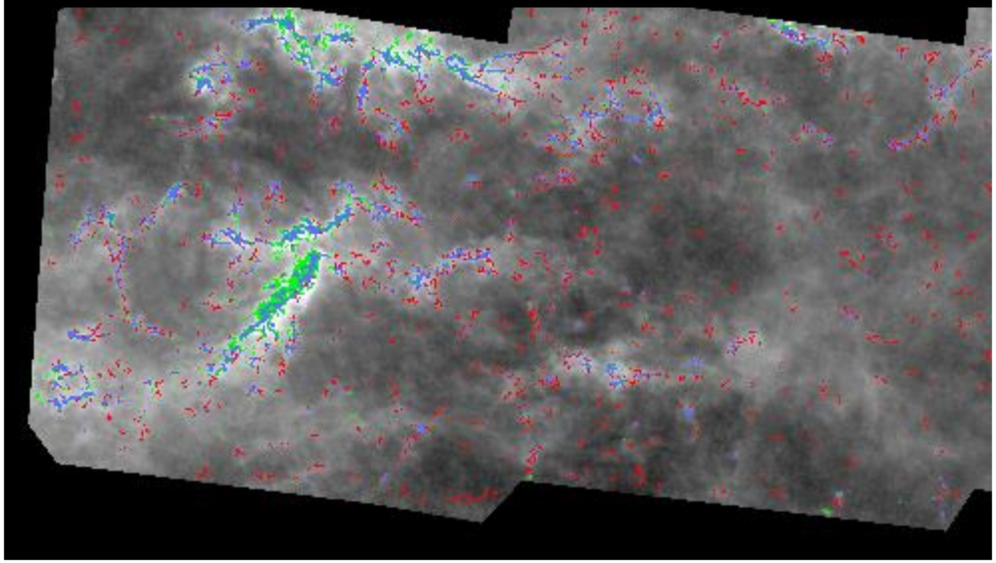
FilExSeC is able to connect, by means of NFPs, filaments that in the traditional method are tagged as disjointed objects.
By joining filaments as a unique structure total mass and mass per length change, inducing a different physics of the filamentary structure.

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.

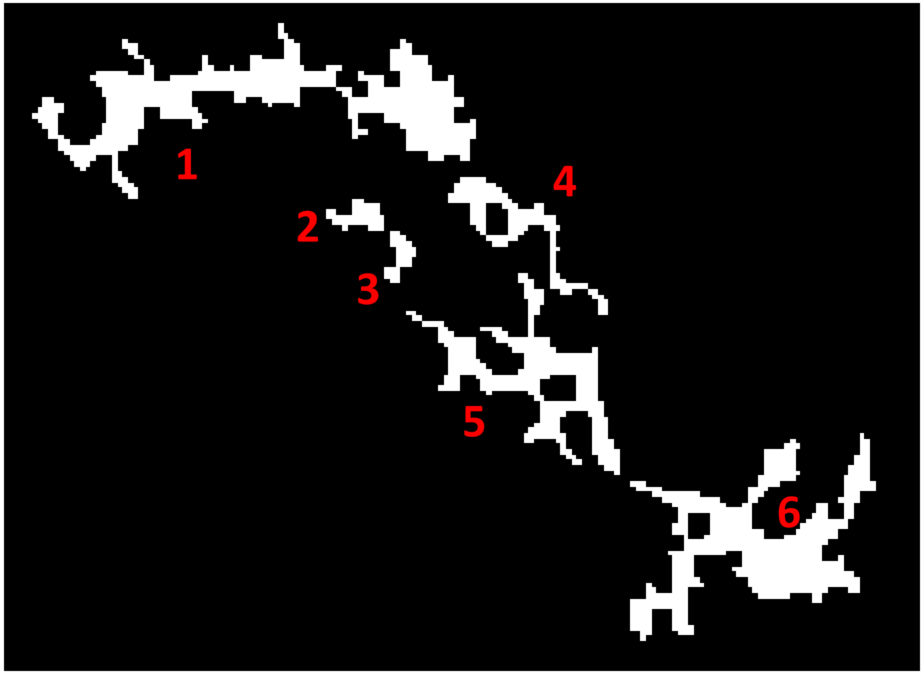




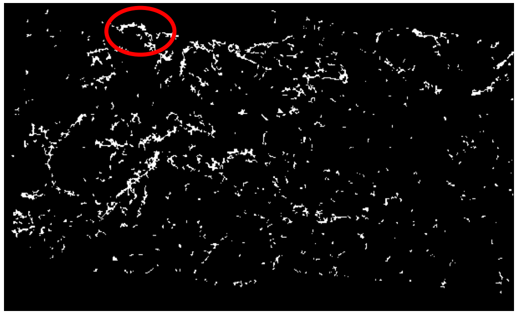
- Confirmed Filament Pixels
- New Filament Pixels
- Confirmed Background
- Undetected Filament Pixels

FilExSeC – Example of Joined Filaments

IAPS



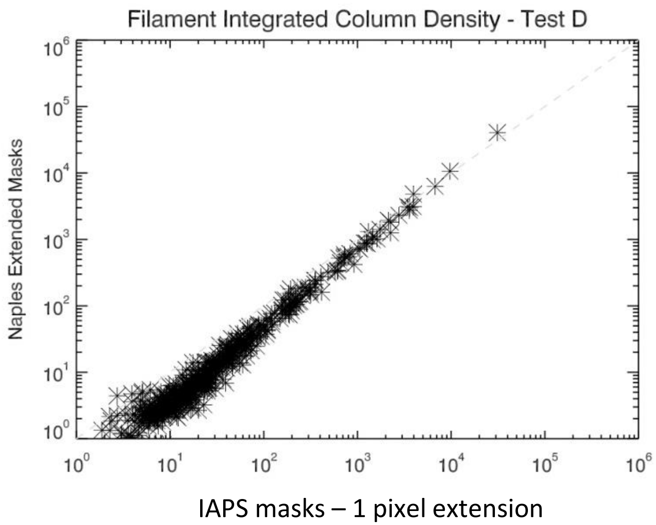
FilExSeC

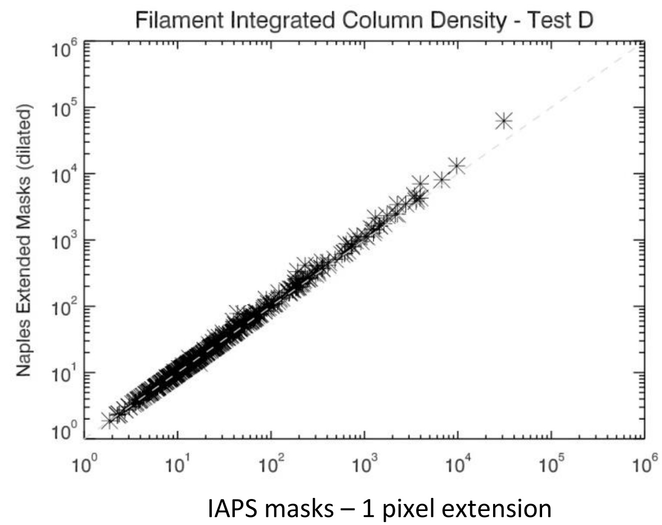


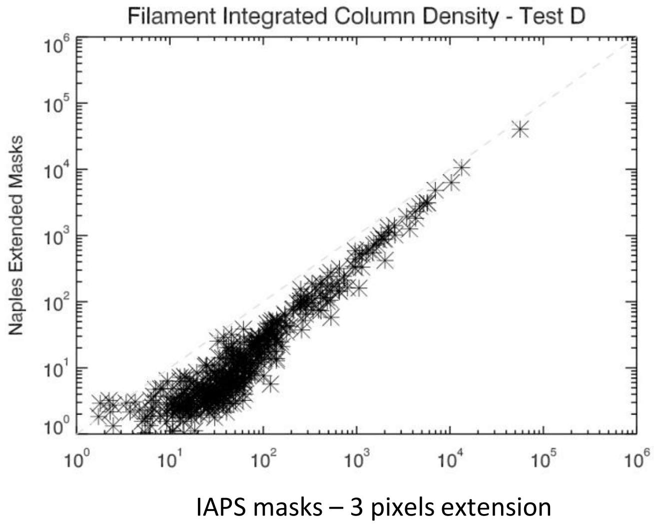
Connection of 6 filaments identified by IAPS

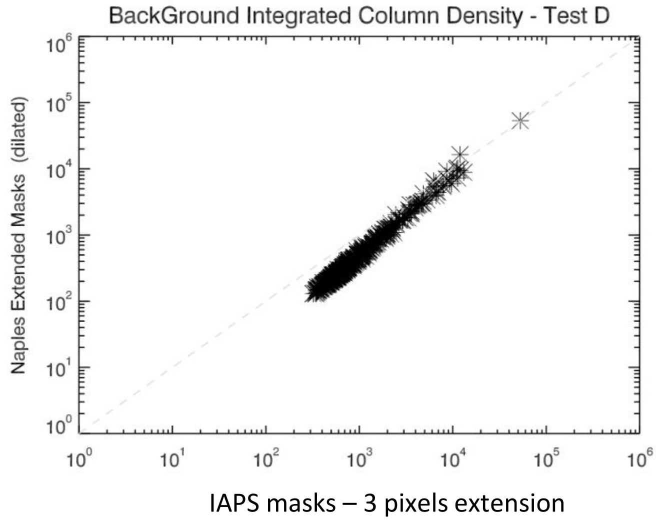
IAPS total number of pixels: 1852
Pixel added by FilExSeC : 858
New Total number of pixels: 2710
% NFP: +46.33%

Filament physical analysis



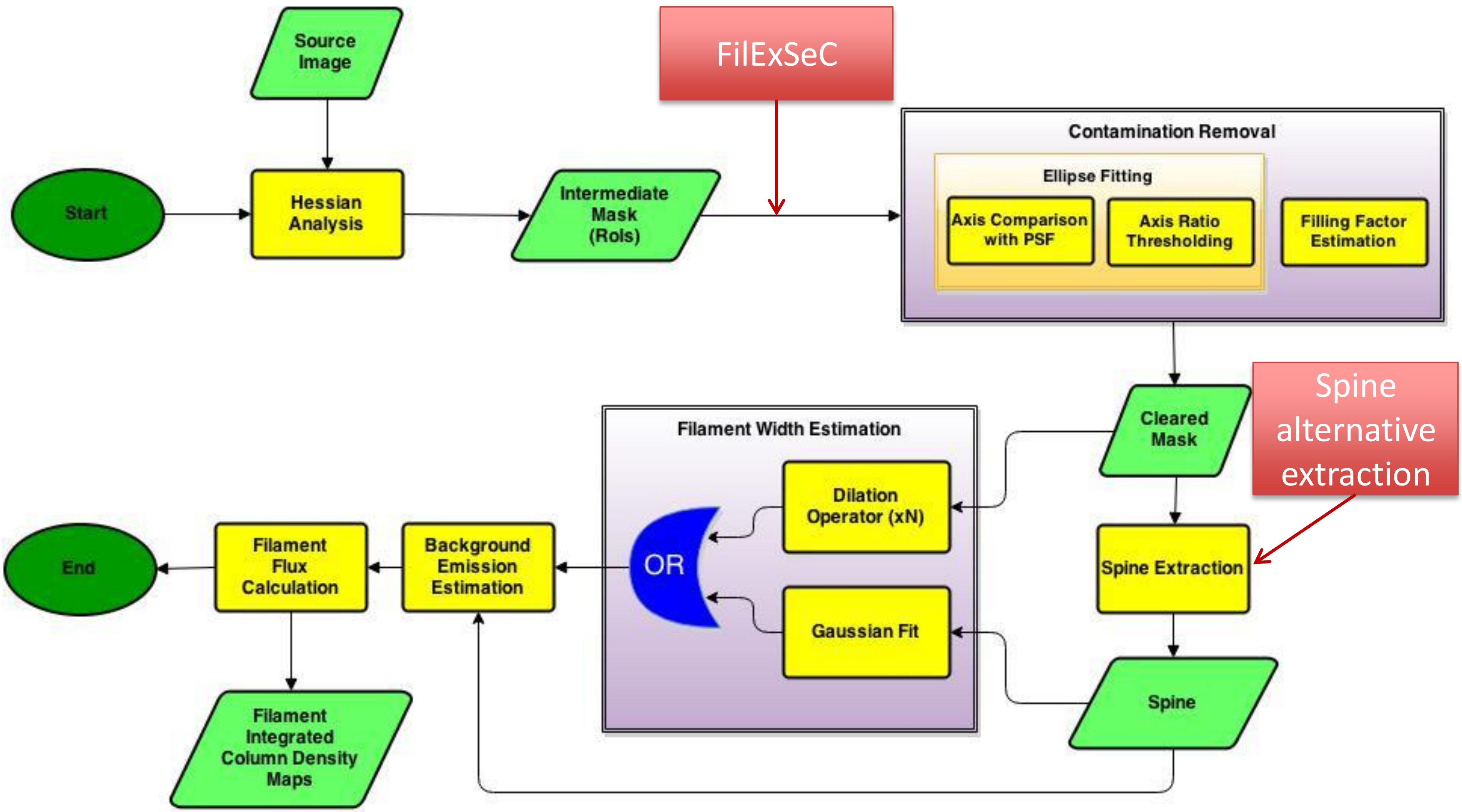






Tests on Hi-GAL l217-l224 confirm that the FilExSeC method is able to justify a 1-pixel extension of the IAPS results, but it is always an underestimation of the IAPS masks extended with 3 pixels

Spine detection

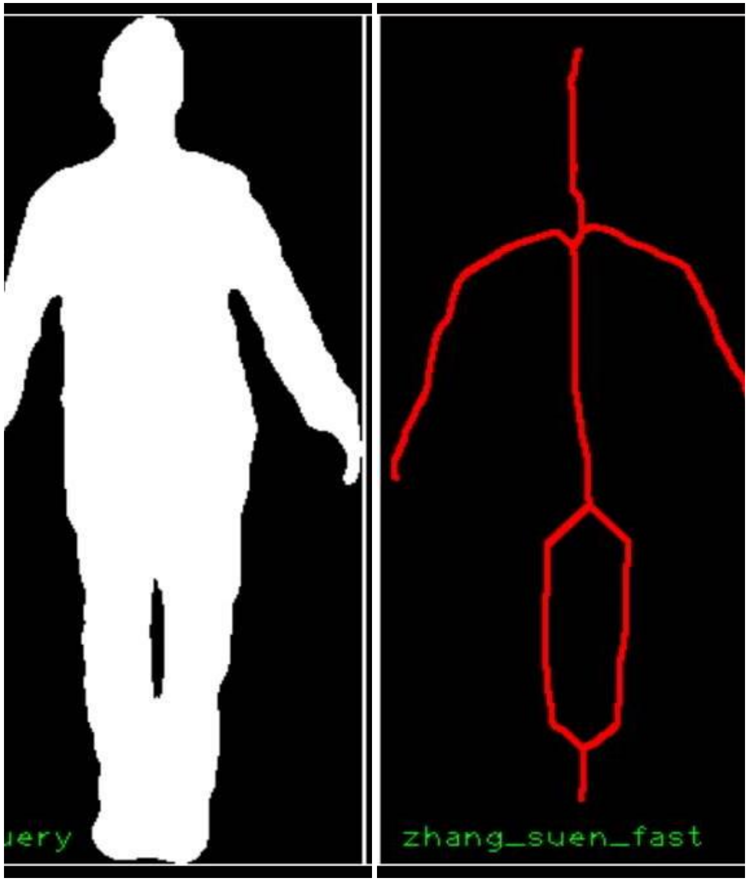


Spine detection

Z-S algorithm

Based on a modified version of Zhang and Suen algorithm (1984)

It is a fast parallel thinning algorithm that consists of two sub-iterations: one aimed at deleting the south-east boundary points and the north-west corner points while the other one is aimed at deleting the north-west boundary points and the south-east corner points. **End points and pixel connectivity are preserved**. Each pattern is thinned down to a "skeleton" of **unitary thickness**.



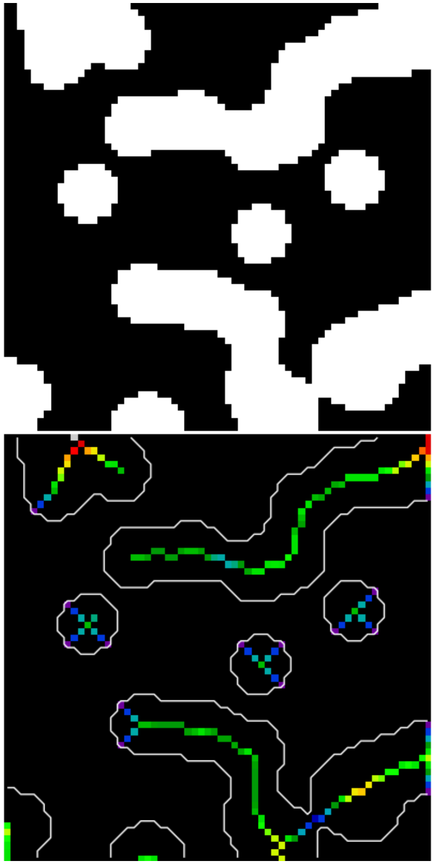
Spine detection

MAT - Medial Axis Transform (Blum 1967)

The **medial axis** of an object is the set of all points having more than one closest point on the object's boundary.

In 2D, the medial axis of a subset *S* which is bounded by planar curve *C* is the locus of the centers of circles that are tangent to curve *C* in two or more points, where all such circles are contained in *S*.

The medial axis together with the associated radius function of the maximally inscribed discs is called the **medial axis transform** (**MAT**).

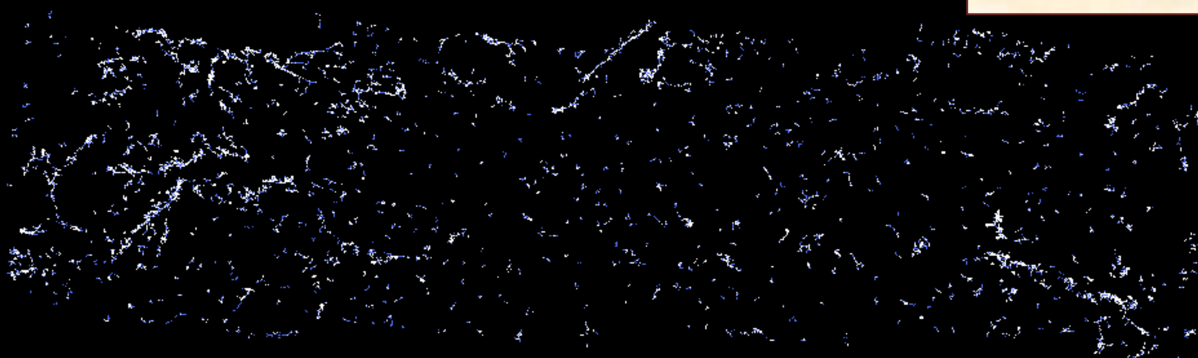


Both methods have been preliminarily validated on old simulations and tested on Lupus masks. In the following slides we report results.

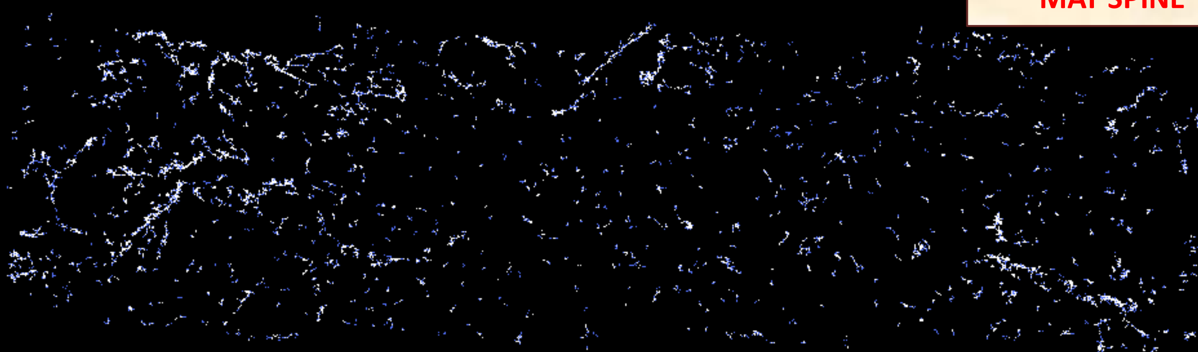
Spine detection test Hi-GAL l217-l224

IAPS SPINE

Z-S SPINE

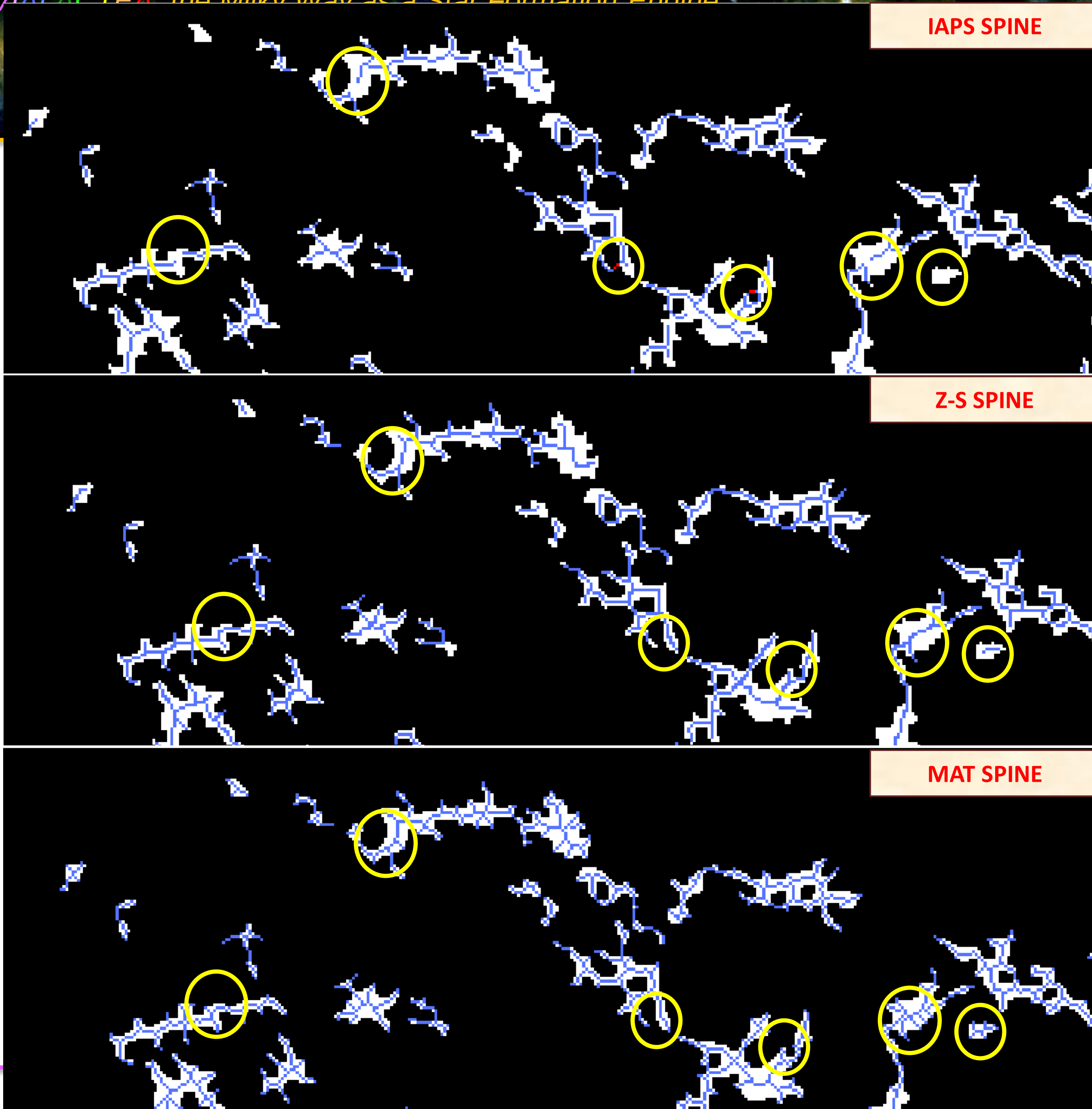


MAT SPINE



IAPS mask filament (white) pixels	78,341
Z-S Spine (blue) pixels:	30,285
MAT spine (blue) pixels:	33,949
IAPS Spine (blue) pixels:	21,146
IAPS out-of-mask (red) pixels:	98

In the following slides we highlight some interesting area.

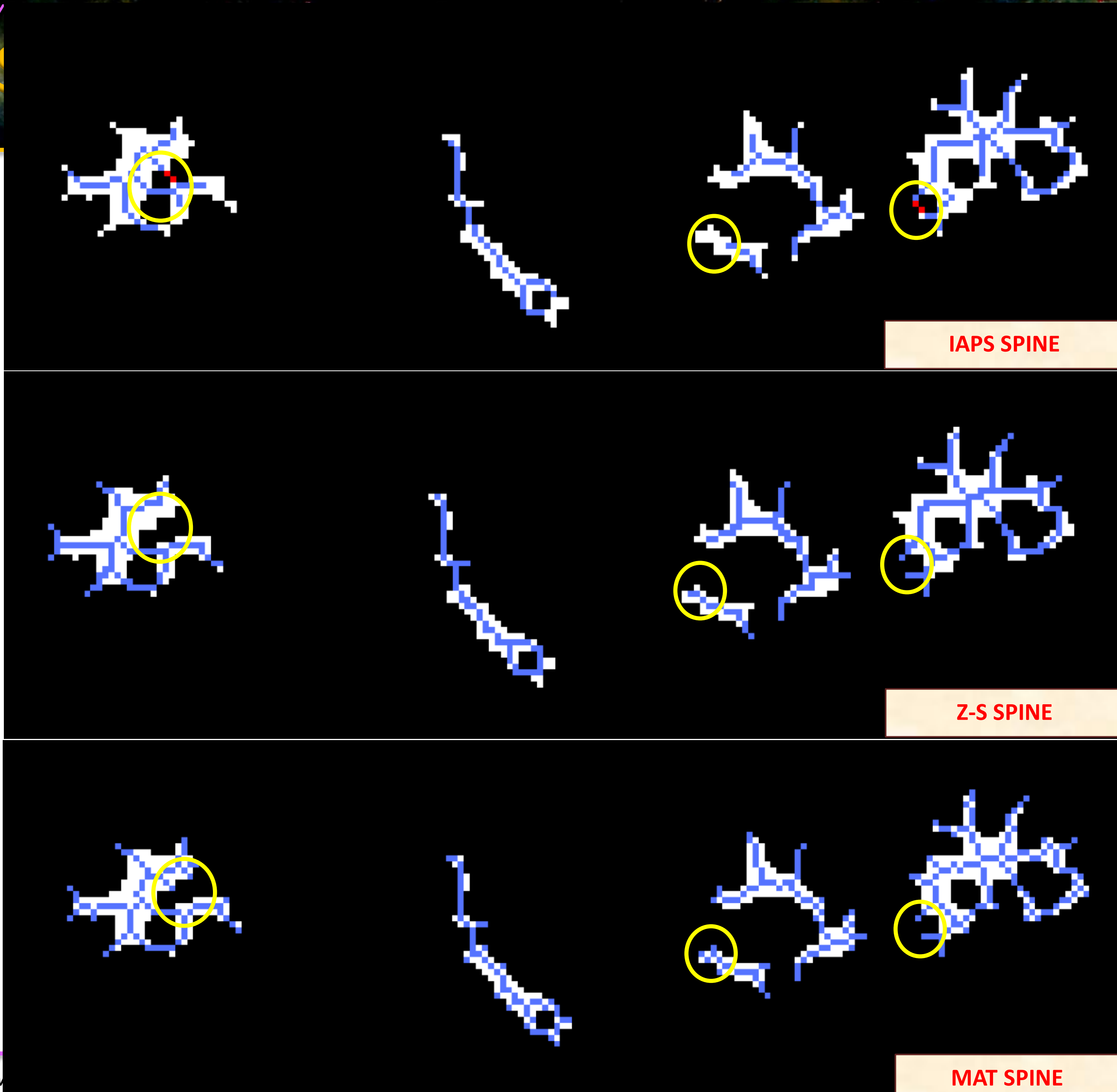


Example 1

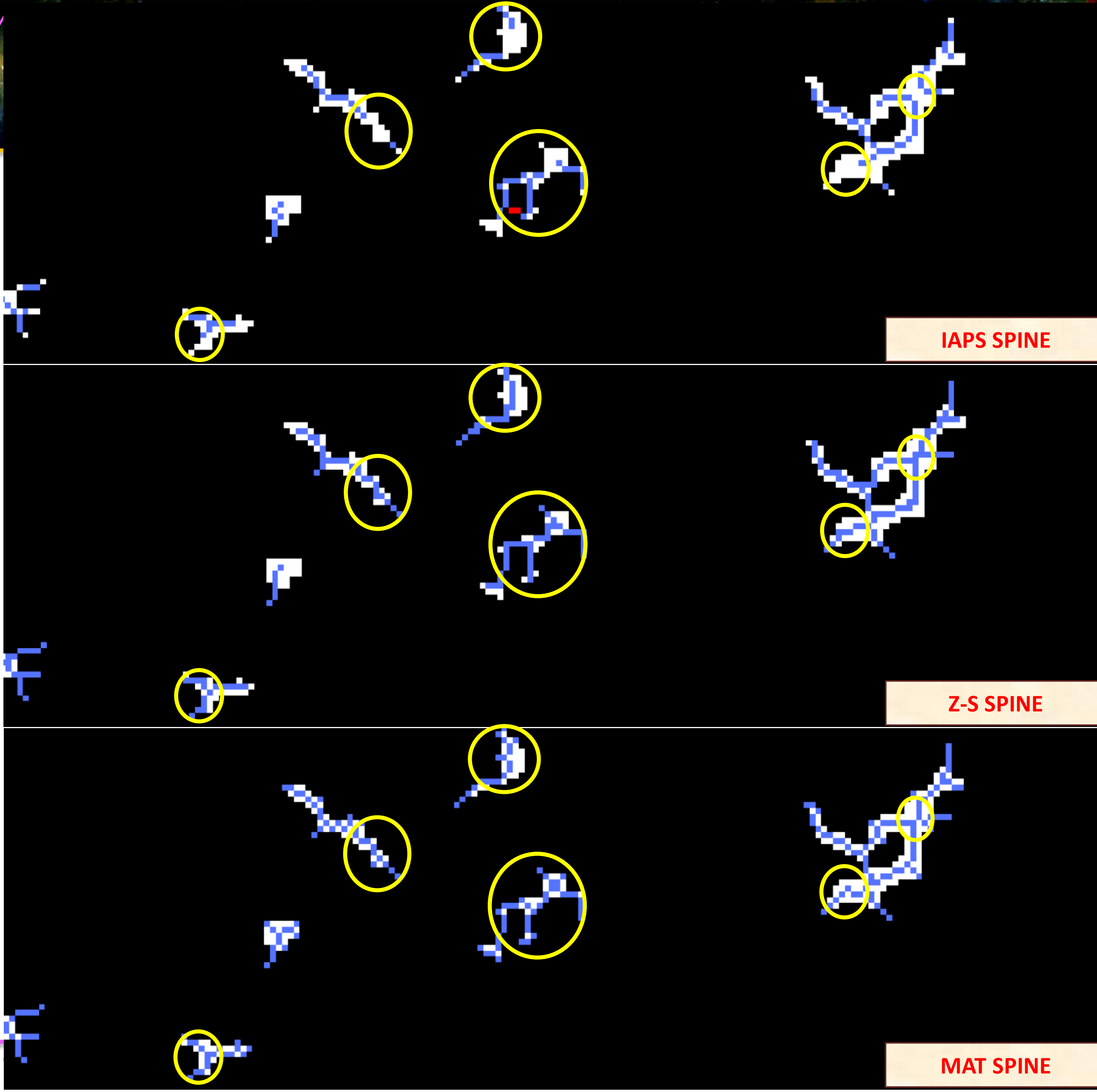
White Pixels:
Filament Mask Pixels

Blue Pixels:
Filament Spine Pixels

Red Pixels:
Filaments Spine Pixels out the mask region



Example 2



Example 3

The workflow for compact source analysis

The main limitation of FilExSec is that it works with masks obtained by traditional methods. This causes a bias on our performances.

It is necessary to find a method directly working on original images without any priors

At this time, we are under investigation on new edge detectors:

- **Boosted Edge Learning** (*Dollar et al. 2006*)
- **gPb (global Probability of boundary)** (*Arbelaez et al. 2011*)
- **Beam-curve Pyramid based edge detector** (*Alpert et al. 2010*)
- **Curvelets and Wavelets** (*Starck et al. 2002* and *Mallat 1998*)
- **Fuzzy Logic Edge Detectors** (*Becerikli et al. 2005*)
- **Canny and Sobel filters enhancement** (*Canny 1986* and *Sobel 2014*)

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 European project at mid-term stage (April 2015) is respecting the initial goals, among which the data mining and machine learning expectation to release useful resources and solutions for the wide scientific community, which will remain available also after the project closure (October 2016).


THANKS!