



Preprocessing, Management, and Analysis of Mass Spectrometry Proteomics Data

Mario Cannataro

*University “Magna Græcia” of
Catanzaro, Italy*

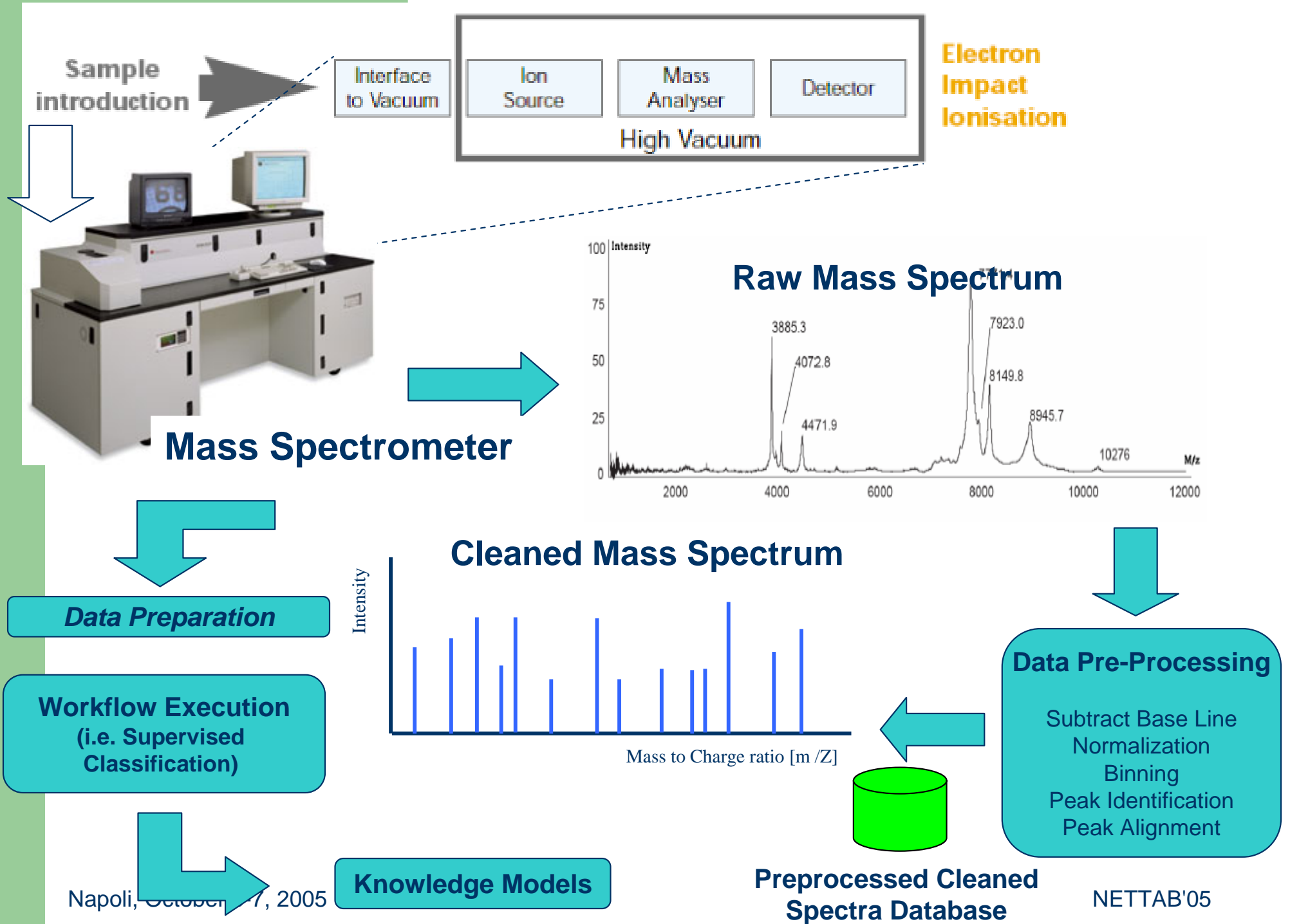
cannataro@unicz.it

*** Joint work with P. H. Guzzi, T. Mazza, P. Veltri**

Outline

- Mass Spectrometry-based Proteomics
 - MS data
- Preprocessing of MS Data
 - Noise Reduction
 - Binning
 - Peaks Alignment
- MS-Analyzer, a platform for spectra analysis
 - Ontology+Workflow
- Conclusions and future work

Mass Spectrometry-based Proteomics



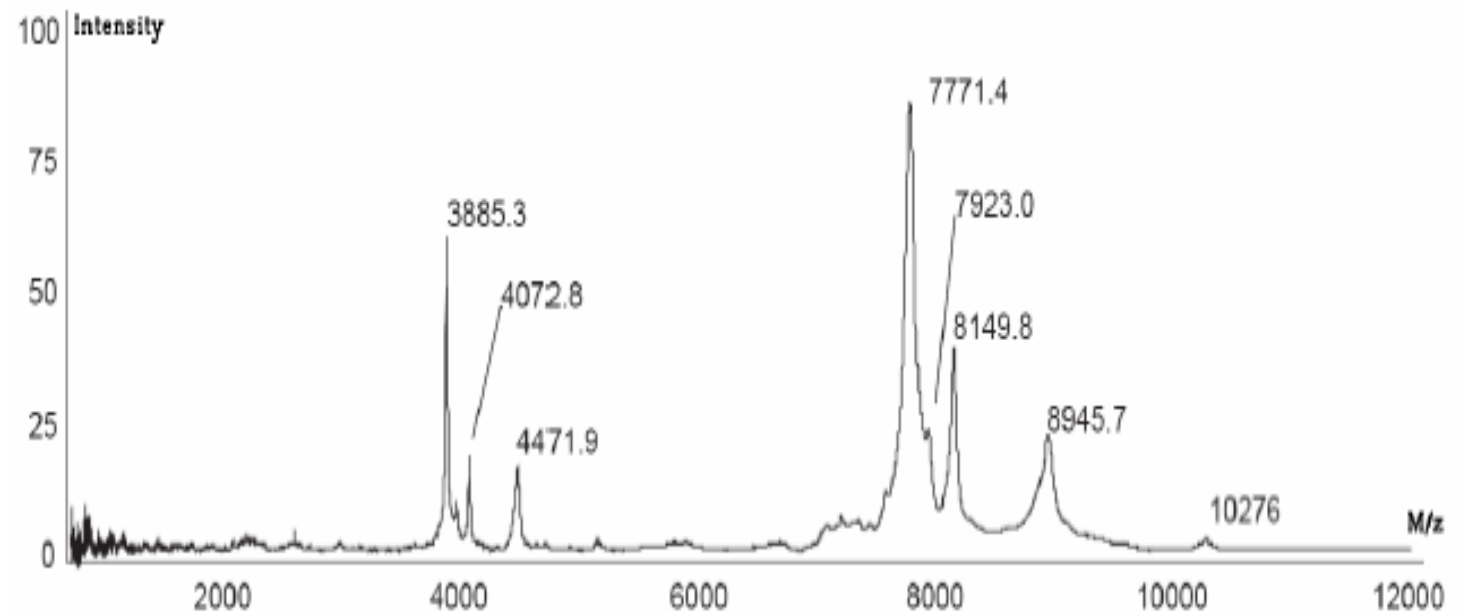
- **A spectrum is a large sequence of value pairs ($I, m/Z$):**
 - I (*intensity*) depends on the quantity of the detected biomolecules,
 - m/Z (*mass to charge ratio*) depends on the molecular mass of detected biomolecules

MALDI-TOF spectrum

...

700.003	9.000
700.015	3.000
700.026	2.000
700.038	3.000
700.050	1.000
700.062	2.000
700.073	2.000
700.085	4.000
700.097	3.000
700.109	2.000
700.121	5.000
700.132	8.000
700.144	7.000
700.156	12.000

...

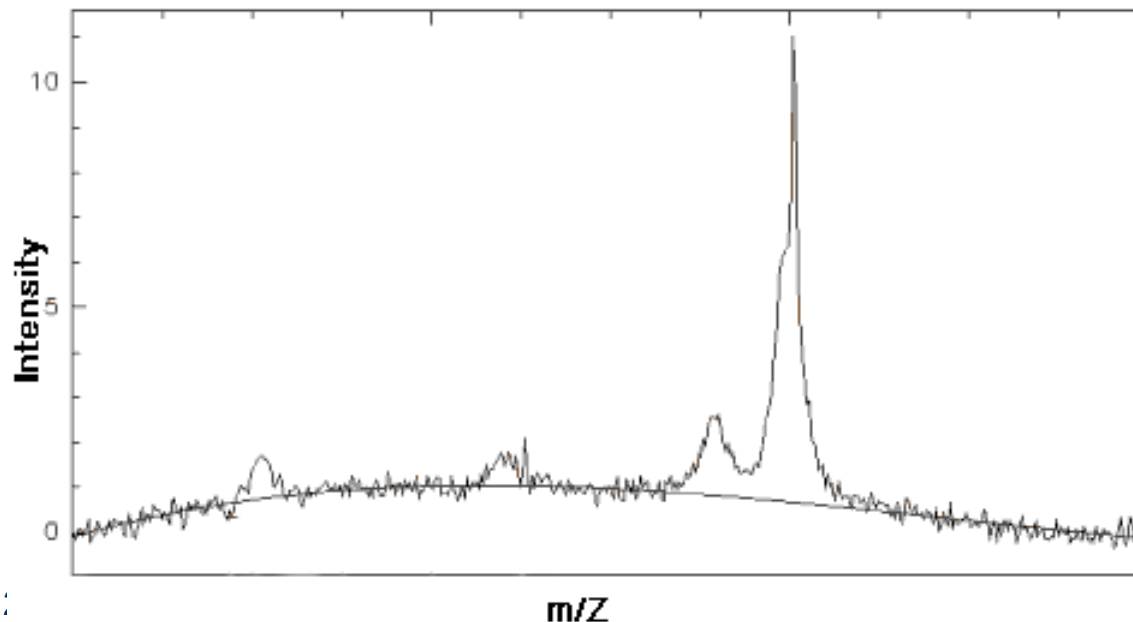


Preprocessing of MS Data

- Each spectrum data is the result of two measurements ($I, m/Z$) that are corrupted by noise
- Preprocessing aims to correct *intensity* and m/Z values in order to:
 - (i) reduce noise,
 - (ii) reduce amount of data, and
 - (iii) make spectra comparable (Data calibration or Peaks Alignment).

Noise Reduction

- Each mass spectrum exhibits a base intensity level (a baseline) which varies from fraction to fraction and consequently needs to be identified and subtracted.
 - **Base line subtraction** flattens the base profile of a spectrum
 - **Smoothing** reduces the noise level in the whole spectrum.



Normalization

- Normalization aims to make intensity comparable across different spectra. It is usually applied to each spectrum.
- Different normalization scheme

– Direct
$$I_{j_norm} = \frac{I_j - I_{min}}{I_{max} - I_{min}}$$

– Inverse
$$I_{j_norm} = 1 - \frac{I_j - I_{min}}{I_{max} - I_{min}}$$

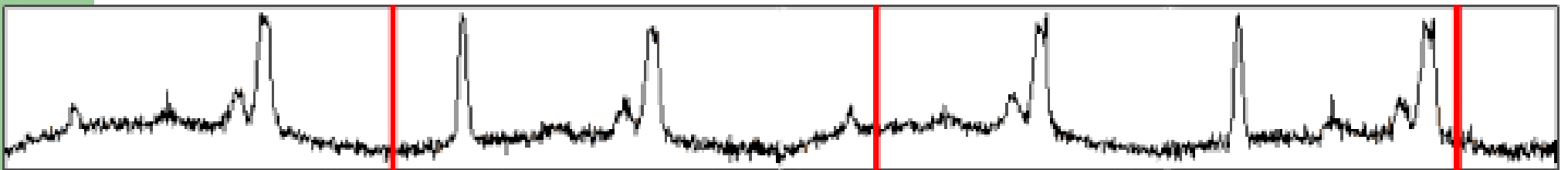
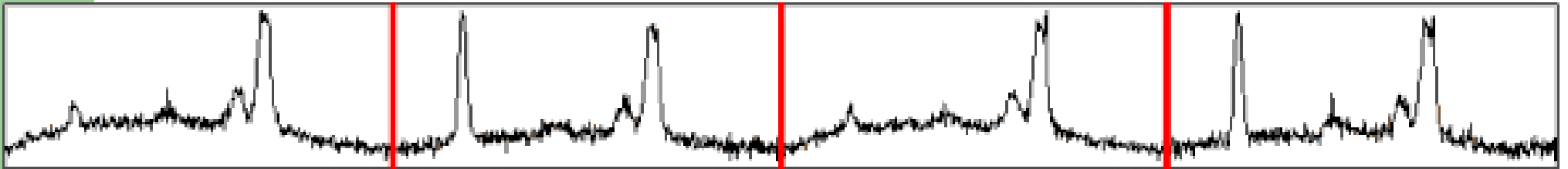
– Canonical
$$I_{j_norm} = \frac{I_j}{\sum I_j}$$

– Logarithmic

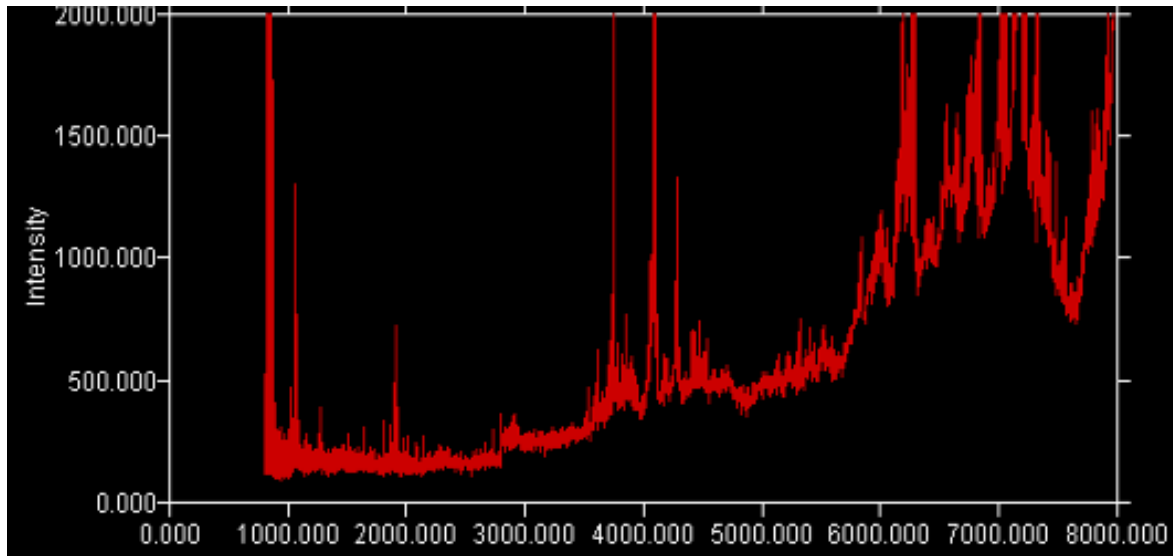
- This method performs a logarithmic transformation on a set of K spectra. The intensity measure I_{hj} for sample h , ($h = 1, \dots, K$), at m/Z value j , is transformed because of its skewed distribution.

Data Reduction: Binning

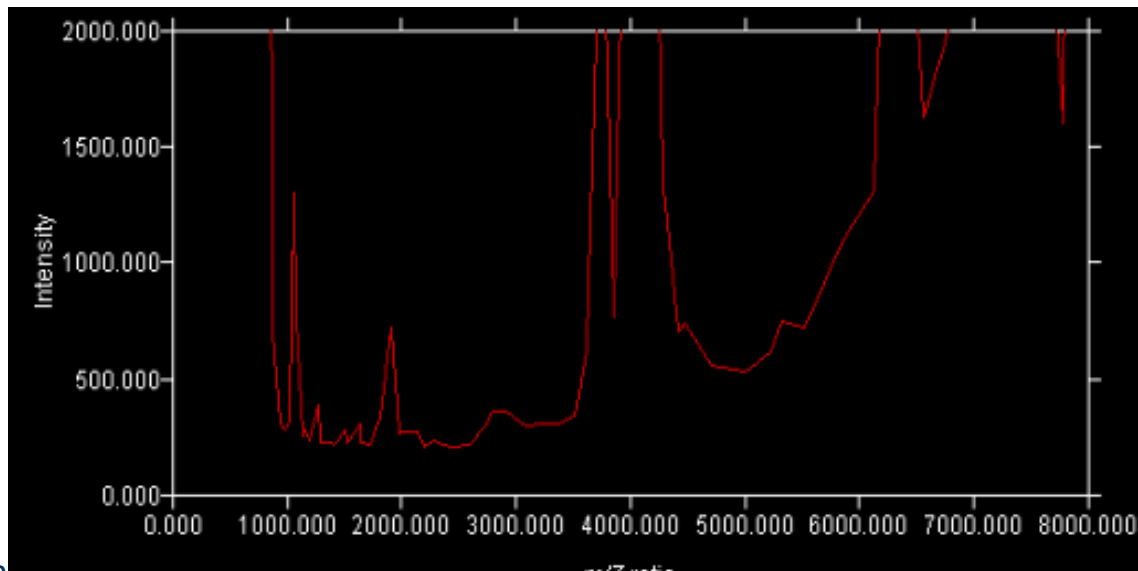
- Binning performs data dimensionality reduction by grouping measured data into *bins*.
- It calculates, for each interval of m/Z values (bin),
 - an aggregate intensity (e.g. the sum of the intensities in the bin)
 - and a representative m/Z value (e.g. the median or the one with maximum intensity)
- and elects this couple as representative of the bin.



Data Reduction: Binning



Raw Spectrum



Binned Spectrum:

- Window Size: 100
- Type: fixed

Peaks Alignment

- Without alignment, the same peak (e.g. the same peptide) may have different m/Z values across samples
- This method finds a common set of peak locations (i.e. m/Z values) in a set of spectra, so that all spectra have common m/Z values for the same biological entities.
 - Applied to an entire dataset.
- Window in which m/Z can be shifted is defined as *window of potential shift* (intrinsically related to instrument)
 - A precise definition of this value has to consider instrument data sheet or a set of measurement (a sort of calibration).
 - For MALDI-TOF MS instrument this values can be set as 0.5% of current m/Z values (variable window shift).

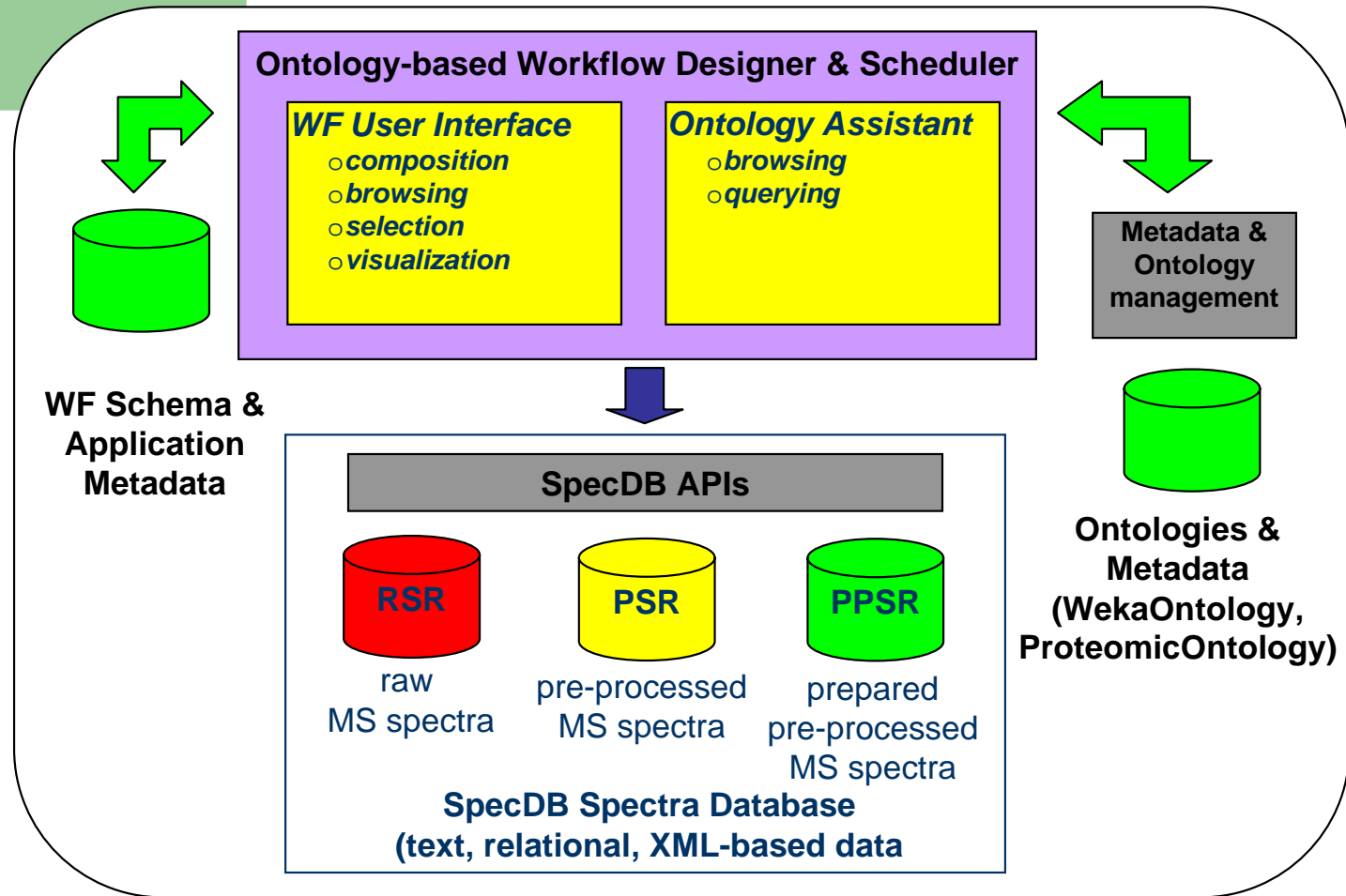
MS Data Mining Steps

1. **Loading** of the raw spectra produced by mass spectrometer,
2. **Preprocessing** of the raw spectra data,
3. **Preparation** of the data mining input file (i.e. the Weka ARFF file),
4. Data Mining **analysis** (i.e. classification) of mass spectra,
5. Knowledge Models **visualization** (e.g. decision tree)
 - represented in a standard language as PMML

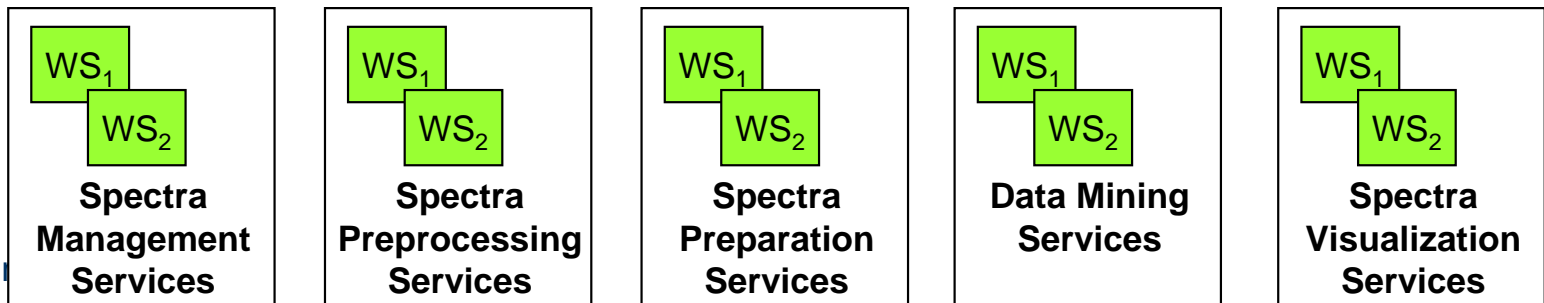
MS-Analyzer: a software platform supporting MS Analysis Workflows

- MS-Analyzer is a Grid-based Problem Solving Environment for the preprocessing, management and data mining analysis of MS data. It uses
 - domain ontologies to model
 - software tools (e.g. preprocessing and mining) and their relationships
 - data sources (e.g. MALDI-TOF, LC-MS/MS spectra datasets)
 - constraints (e.g. binning cannot be applied twice)
 - and workflow techniques to design complex in silico experiments.
- MS-Analyzer is able to:
 - interfacing mass spectrometers (i.e. spectra data sources)
 - acquiring, preprocessing and managing MS data on the Grid
 - offering composable filtering, preprocessing, and DM services
 - sharing experiments (data, workflows and knowledge)
- enabling a **Virtual Proteomics Laboratory**

MS-Analyzer



Grid Infrastructure (Globus)



MS-Analyser Ontologies: WekaOntology & ProteomicOntology

- owl:Thing
- ▶ ● Algorithm
- ▶ ● Method
- ▶ ● Spectrum
- ▶ ● Software
- ▶ ● Knowledge_Model
- ▶ ● Task

- ▼ ● Software
 - ▶ ● Classification_Software
 - ▶ ● Clustering_Software
 - ▶ ● Data_Preparing_Softwares
 - ▼ ● PreProcessing_Softwares
 - ▼ ● Data_Dimensionality_Reduction_Softwares
 - ▼ ● Binning_Softwares
 - ▶ ● Linear_Binning_Softwares
 - ▶ ● Variable_Window_Software
 - ▶ ● Peak_Extraction_Software
 - ▼ ● Noise_Reduction_Softwares
 - ▶ ● Smoothing_Softwares
 - ▶ ● Base-Line_Subtraction_Softwares
 - ▼ ● Normalization_Softwares
 - ▶ ● Canonical_Normalization
 - ▶ ● Direct_Normalization
 - ▶ ● Inverse_Normalization
 - ▶ ● Alignment_Softwares
 - ▶ ● Selection_Attribute_Software
 - ▶ ● Visualization_Software

- ▼ ● Activities
 - ▶ ● Data_Generation
 - ▼ ● Data_Preprocessing
 - ▼ ● Data_DeNoising
 - ▶ ● Smoothing
 - ▶ ● Data_Alignment
 - ▼ ● Data_Dimensionality_Reduction
 - ▶ ● Data_Binning
 - ▶ ● Peak_Extraction
 - ▶ ● Data_Preparation

Loading Spectra in different Formats

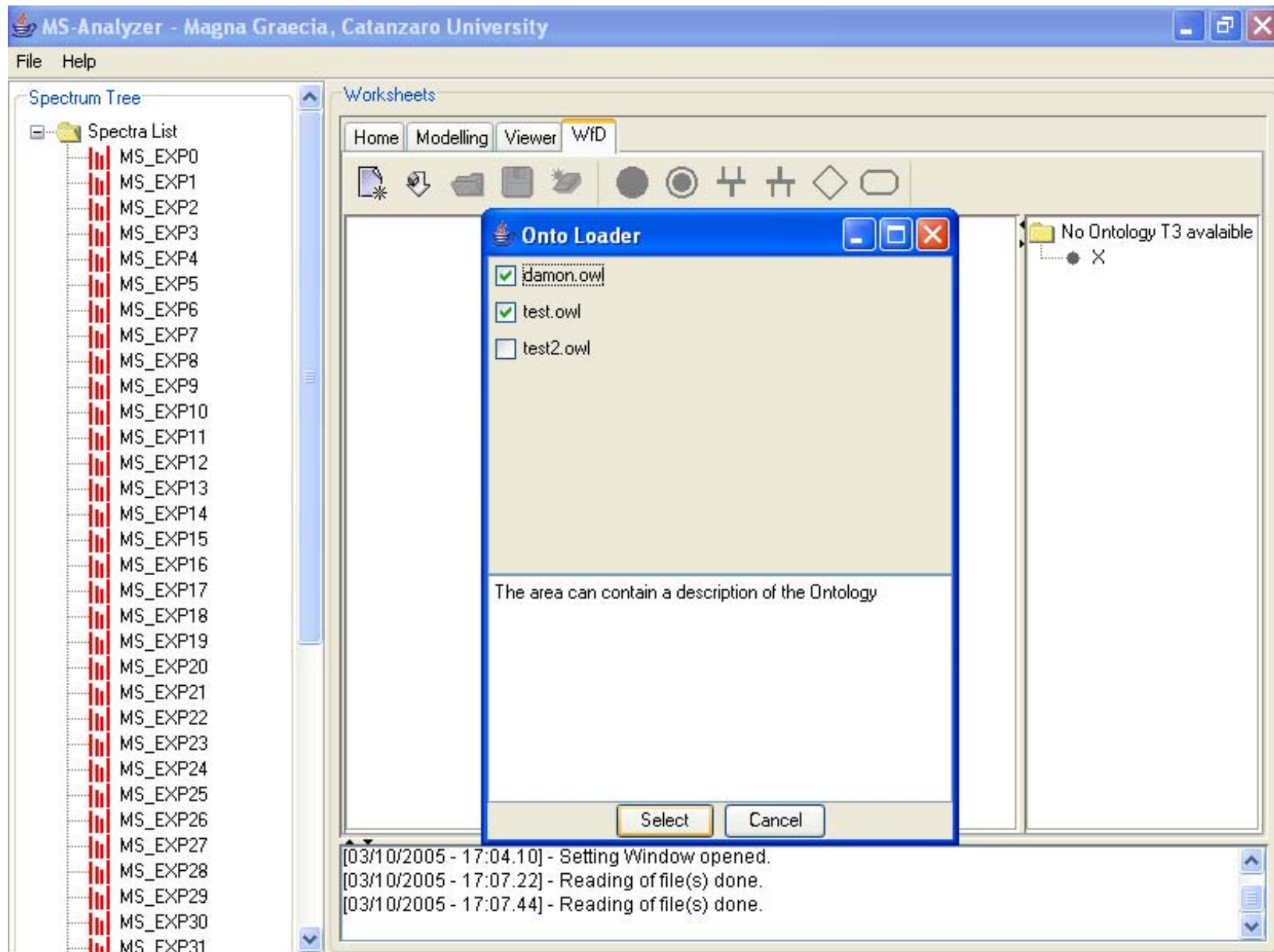
The screenshot displays the MS-Analyzer software interface with the 'Default settings' dialog box open. The dialog box is titled 'Default settings' and contains the following sections:

- Default Separator (IN-OUT):** Radio buttons for 'comma (,)', 'tab', 'semicolon (;)', 'blank', 'two points (:)', 'pipe (|)', and 'other' (with an empty text box).
- Parsing preferences (OUT):** Checkboxes for 'Allows Neg. intensities' and 'Allows 0.0.....E values', both of which are checked.
- Parsing from (IN):** Radio buttons for 'first line' and 'line number...' (with a text box containing '1').
- Local default folder:** Text boxes for 'Input data' (containing ':ettings\Alex\Desktop\Desktop\Ms_AnalyzerFinale\Data') and 'Output data' (containing ':ettings\Alex\Desktop\Desktop\Ms_AnalyzerFinale\output'), each with an 'Open' button.

At the bottom of the dialog box are 'Ok' and 'Close' buttons. The background application window shows a 'Spectra List' on the left and a 'Worksheets' tab labeled 'WfD' at the top. A status bar at the bottom of the application window displays the following log messages:

```
[03/10/2005 - 17:03.54] - MS-Analyzer v1.0 - Welcome!  
[03/10/2005 - 17:04.10] - Setting Window opened.
```

Loading Domain Ontologies



Ontology-based Workflow Design

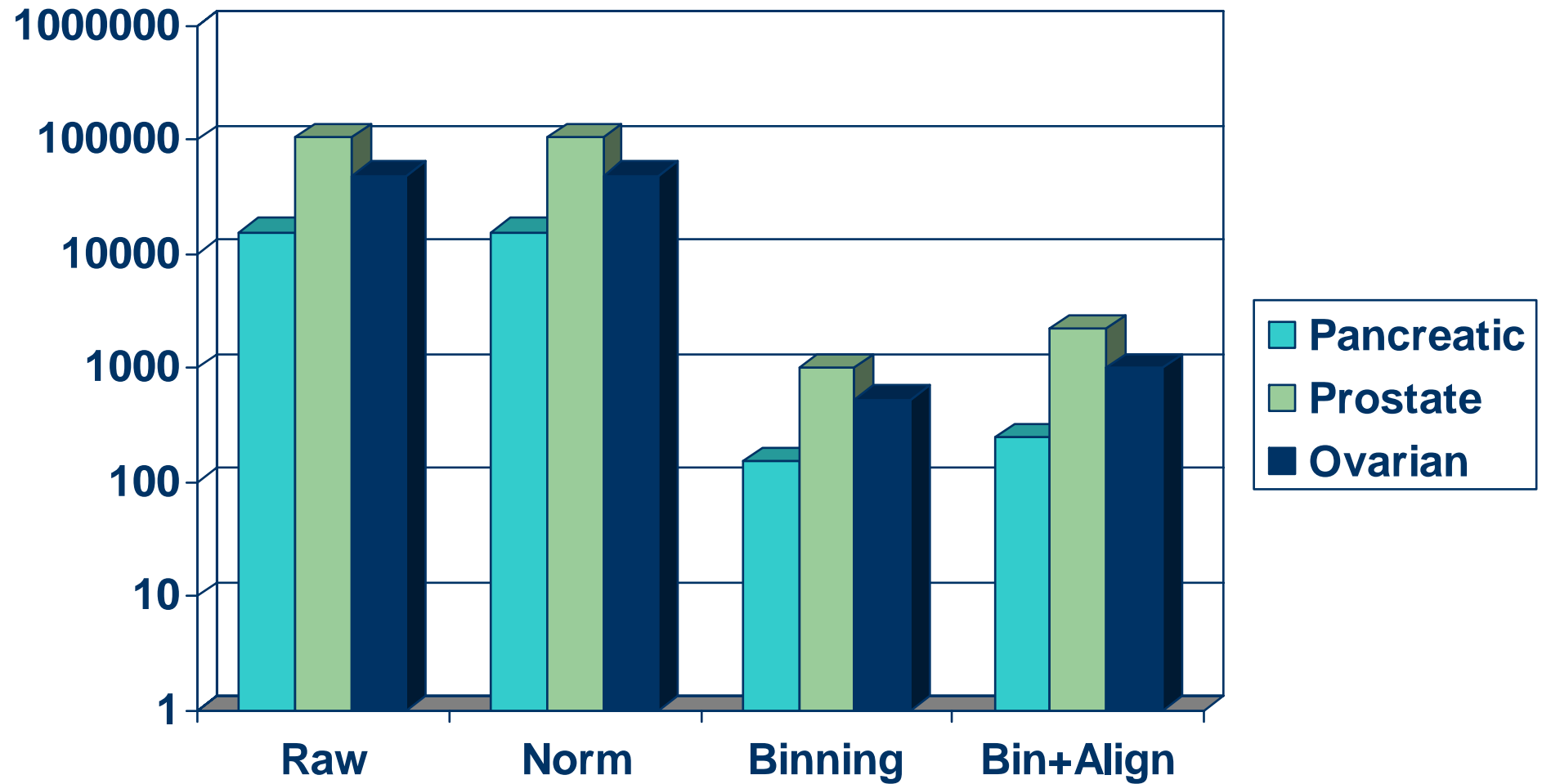
The screenshot displays the MS-Analyzer software interface. The title bar reads "MS-Analyzer - Magna Graecia, Catanzaro University". The interface is divided into several panels:

- Spectra List:** A vertical list of 31 spectra labeled MS_EXP0 through MS_EXP31.
- Worksheets:** A central workspace with tabs for "Home", "Modelling", "Viewer", and "WfD". It contains a workflow diagram with nodes: "Start", "Binning", "Alignment", "Node 3", and "End". Red arrows labeled "Source" connect these nodes. A context menu is open over "Node 3", listing "Connect Operation", "Join", "Fork", "Merge", "Node", and "Delete".
- Algorithm List:** A vertical list of algorithms on the right side, including "HBT", "ID3", "Standard_Voting", "Fuzzy_Algorithms", "SVM_by_Platt", "Gini", "Simple_Algorithm", "Entropy", "SVMlight-Algorithm", "Quest", "SLIQ", "Inferences_Algori", "Wrapper_Algorith", "SMO_by_Platt", "Naive_Bayes", "Batch_Classifier", "Belief_Functions", "SPI_Algorithms", "C4.5_Algorithm", "ConditionalQuerie", "Hierarchical_Influ", "Regression_Algorithm", and "Sequential-Pattern_A".
- Log:** A bottom panel showing system messages: "[03/10/2005 - 17:07.22] - Reading of file(s) done.", "[03/10/2005 - 17:07.44] - Reading of file(s) done.", and "[03/10/2005 - 17:07.59] - Status: Ready".

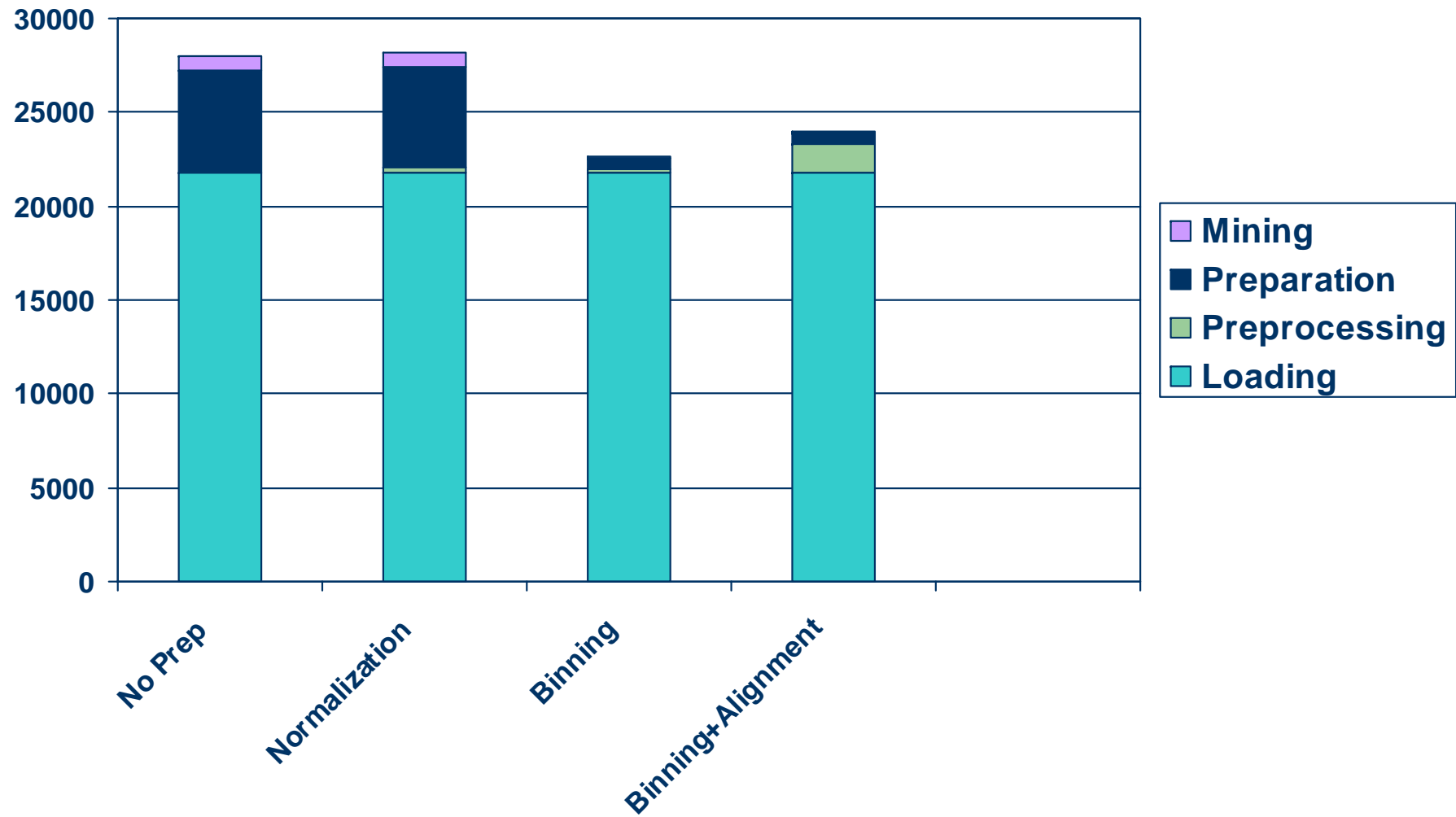
Performance evaluation of preprocessing techniques

- We considered three different mass spectra datasets publicly available on Internet:
 - (i) *Pancreatic Cancer dataset*
 - 142 spectra each one with 6,772 (m/Z, intensity) measurements,
 - two classes: **healthy** and **diseased** patients;
 - (ii) *Prostate Cancer dataset*
 - 322 spectra each one with 15,154 measurements,
 - four classes: **no disease**, **benign cancer**, **pc410**, **pcg10** patients;
 - (iii) *Ovarian Cancer dataset*
 - 49 spectra each one with 59,386 measurements,
 - two classes, 25 **control** and 24 **disease**.
- We measured: execution times, memory occupancy, classification quality

Spectra size [Kbytes]



Execution time vs Preprocessing



Quality of classification

Table 7. Classification indexes with none pre-processing

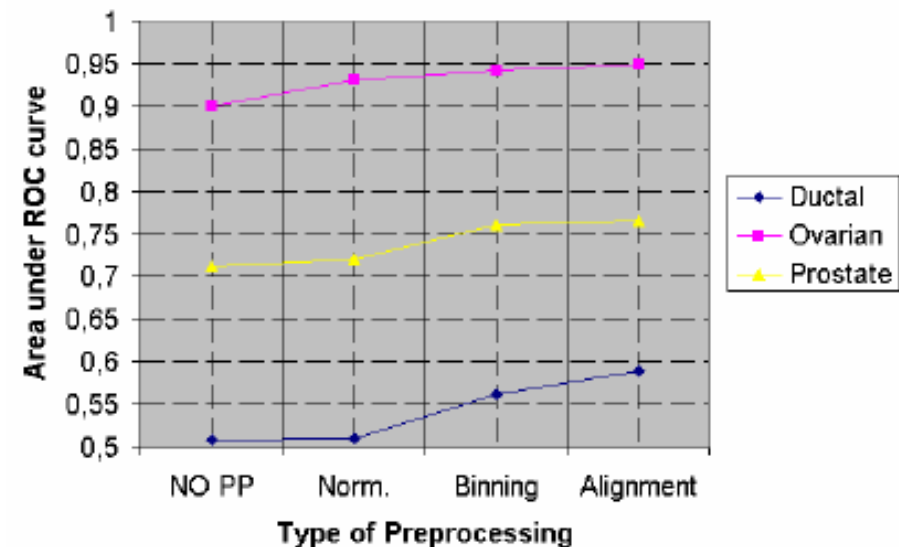
<i>No Prep.</i>	TP Rate	Precision	Recall
Pancreatic C11	0.513	0.5	0.513
Pancreatic C12	0.488	0.5	0.488
Prostate C11	0.774	0.786	0.774
Prostate C12	0.794	0.704	0.794
Prostate C13	0.269	0.269	0.269
Prostate C14	0.395	0.447	0.395
Ovarian C11	0.84	0.913	0.84
Ovarian C12	0.917	0.846	0.917

Table 9. Classification indexes with binning

<i>Binning</i>	TP Rate	Precision	Recall
Pancreatic C11	0.536	0.578	0.536
Pancreatic C12	0.614	0.573	0.614
Prostate C11	0.847	0.885	0.847
Prostate C12	0.825	0.813	0.825
Prostate C13	0.577	0.405	0.577
Prostate C14	0.558	0.615	0.558
Ovarian C11	0,849	0,913	0,845
Ovarian C12	0,934	0,944	0,917

Table 10. Classification indexes with binning and alignment

<i>Bin.+Alig.</i>	TP Rate	Precision	Recall
Pancreatic C11	0.613	0.563	0.613
Pancreatic C12	0.525	0.575	0.525
Prostate C11	0,937	0,922	0,937
Prostate C12	0,937	0,967	0,937
Prostate C13	0,577	0,556	0,577
Prostate C14	0,698	0,732	0,698
Ovarian C11	0.96	0.96	0.96
Ovarian C12	0,958	0,958	0,958



Conclusions and future work

- We surveyed some preprocessing techniques and presented a first prototype of MS-Analyzer
 - ontology-based workflow design simplifies workflow building and helps enforcing constraints
 - selective use of preprocessing techniques may improve execution times, memory occupancy and quality of data mining
- Future work will regard
 - the implementation of MS-Analyzer functions as Grid services and
 - the integration of a spectra relational database to allow “inside database” preprocessing

Questions?

- Contact author
 - Mario Cannataro
 - Informatics and Biomedical Engineering,
 - University “Magna Græcia” of Catanzaro, Italy
 - cannataro@unicz.it
- MS-Analyzer (standalone version) can be downloaded at
 - www.icar.cnr.it/proteus