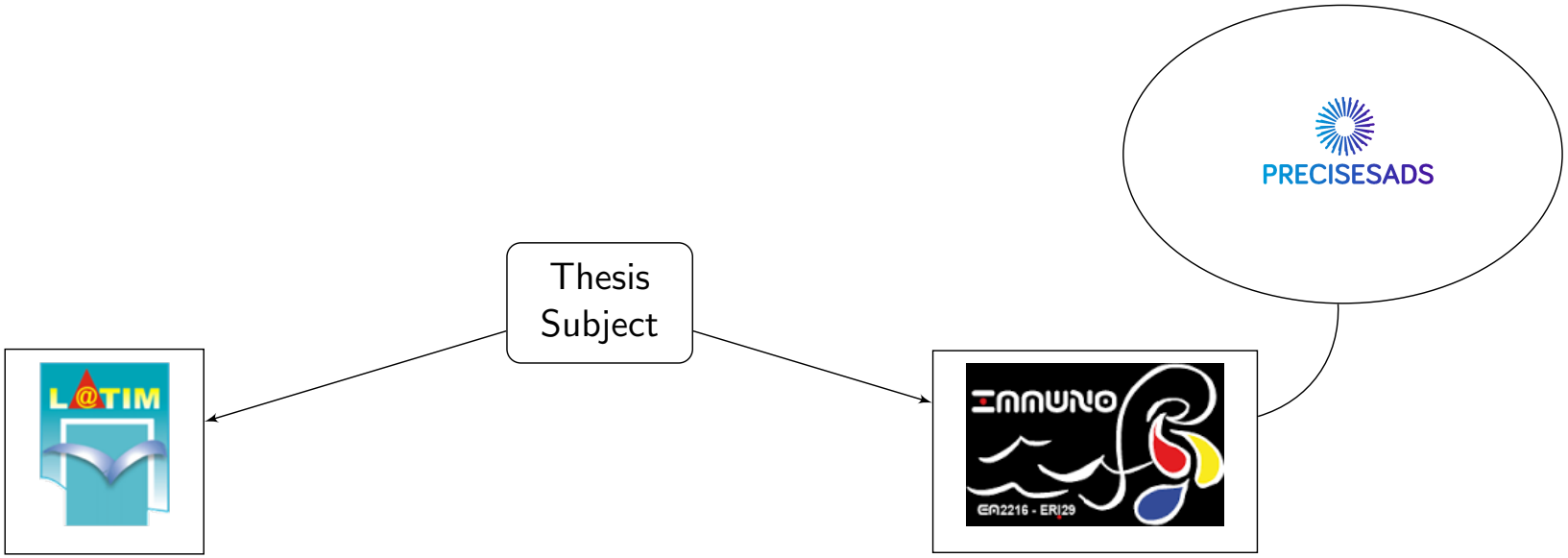
Systemic simulation for the extraction of a patient's biological signature

Nathan Foulquier

LaTIM
CHU Morvan, UMR 1227

Seminar, May 2017

Context



Pascal Redou

Codirection

LaTIM | Lipab

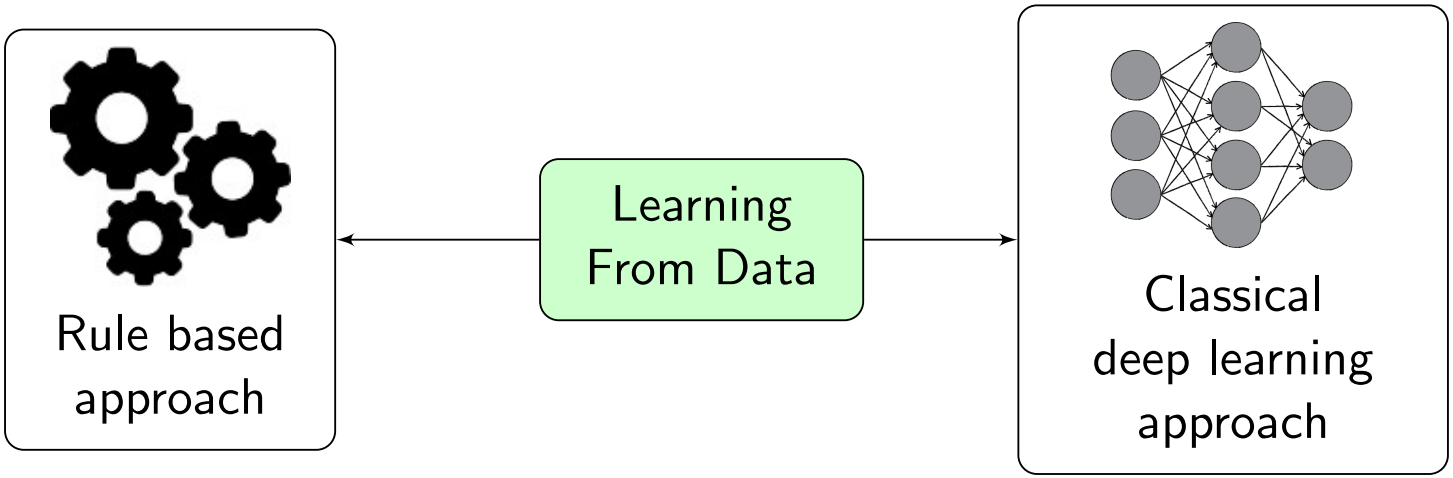
Alain Saraux

Christophe Le Gall

supervisors

Laurent Gobert

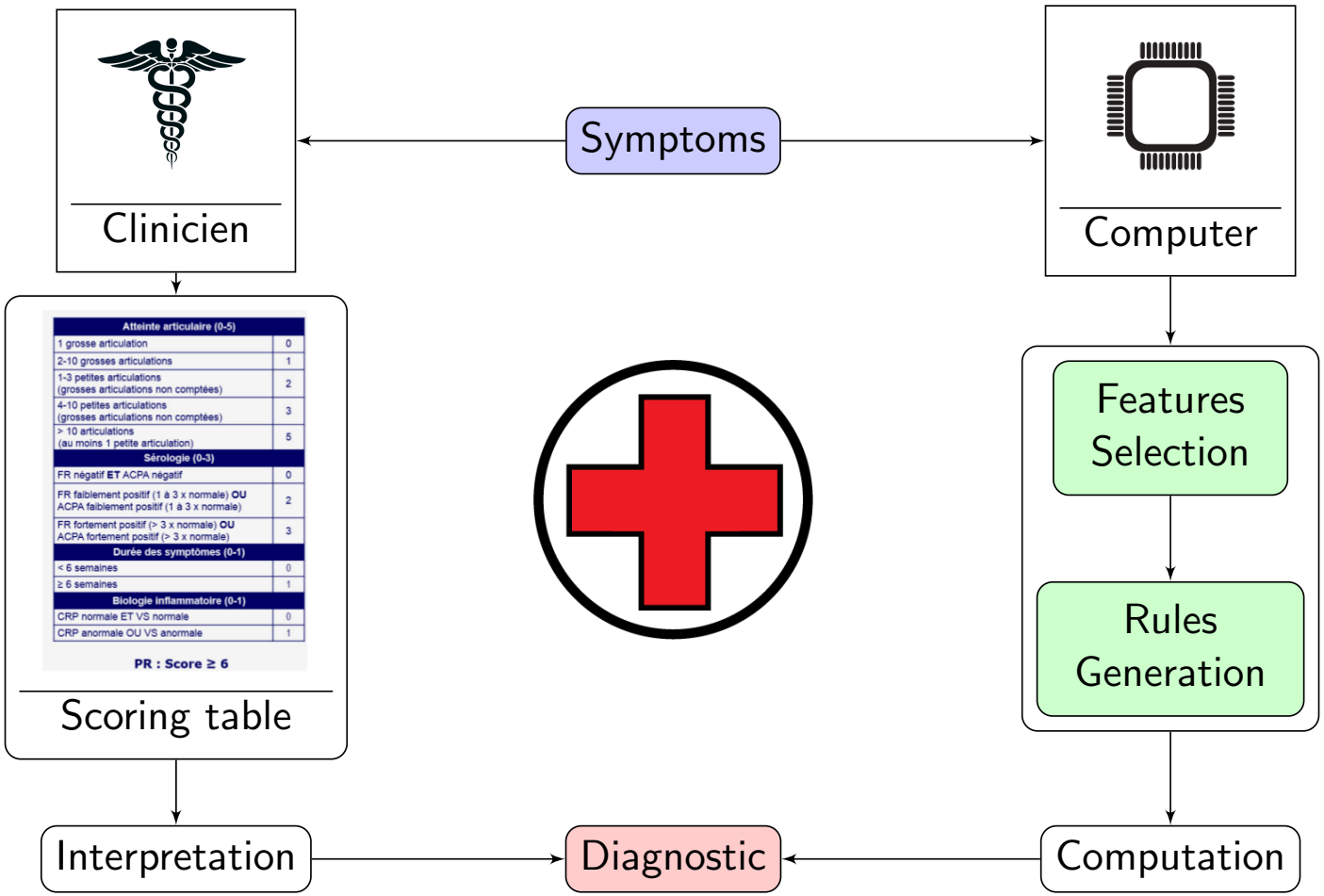
The main idea



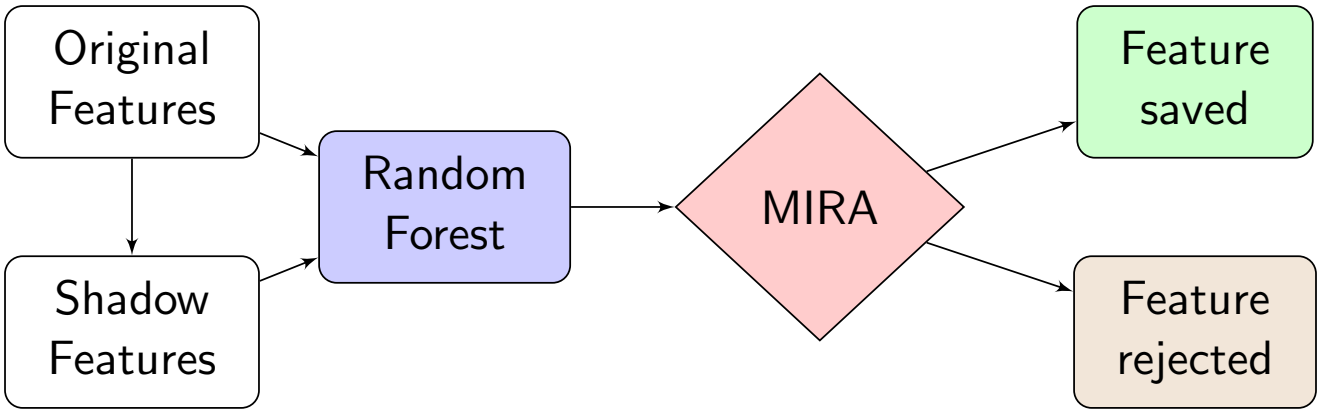
Objectives

- → modelisation
- → sementics
- → signature

The diagnostic



The Boruta Algorithm

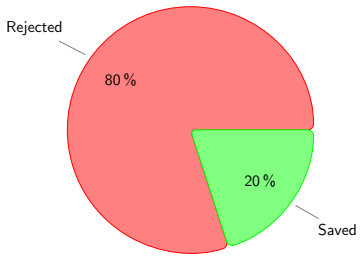


Shadow Feature: random permutation of an original feature
MIRA: maximum importance of all shadow features

→ Rank the originals features
→ Select the "important" features

Results

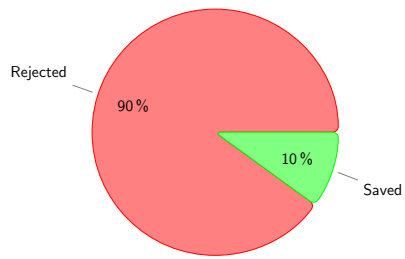
RA



27 features selected

Autoantibody CCP2
Arthrisis
RF

SjS



12 features selected

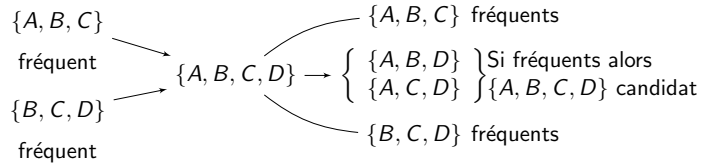
AESICCA
Autoantibody SSA
Autoantibody SSB

→ 80% of irrelevant features for RA
→ 90% of irrelevant features for SjS

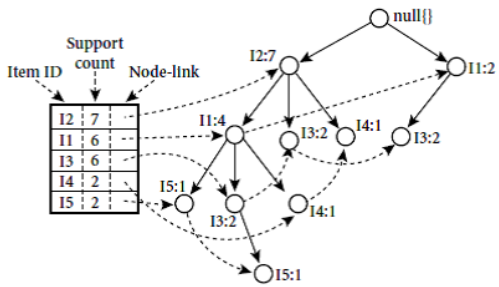
Pattern Mining

$$support(E) = \frac{card(\{p \in P | E \subset p\})}{card(P)}$$

Pattern E is frequent if $support(E) \geq minsup$



A priori Generation



FP-tree Mining

Rules Extraction

$$r : (e - h) \rightarrow h$$

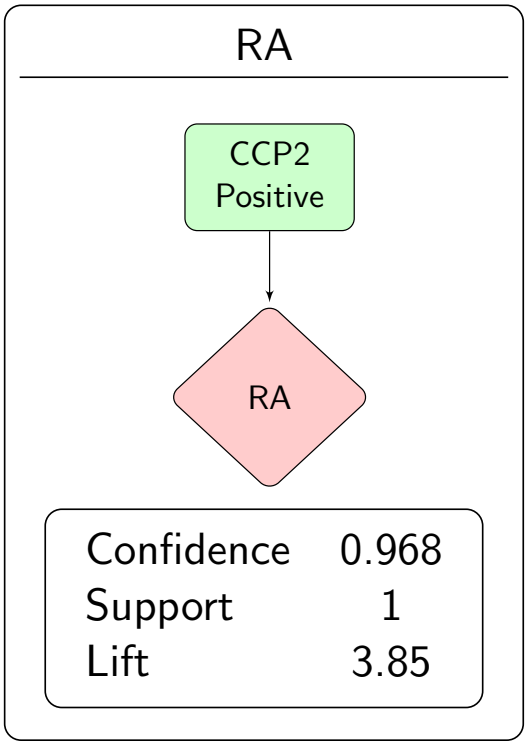
$$e \in \{items\}, card(e) \geq 2$$

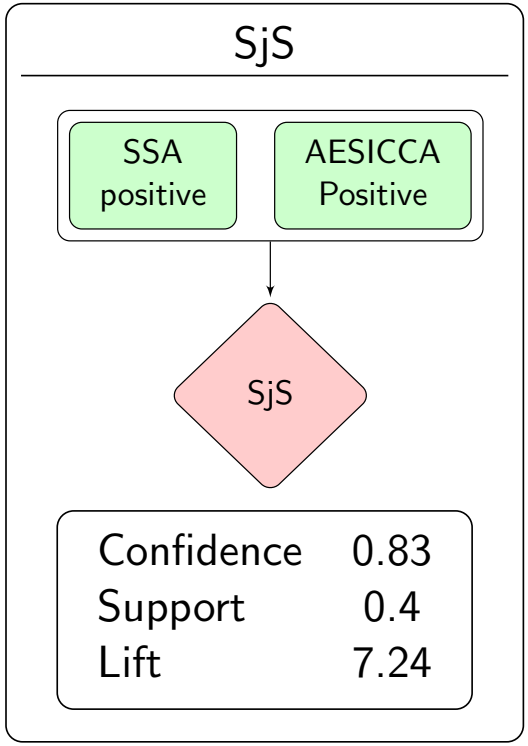
Generate h with $h \neq \emptyset, h \neq e$

$$confidence(r) = \frac{support(e)}{support(e - h)}$$

a rule r is valid if $confiance(r) \geq minconf$

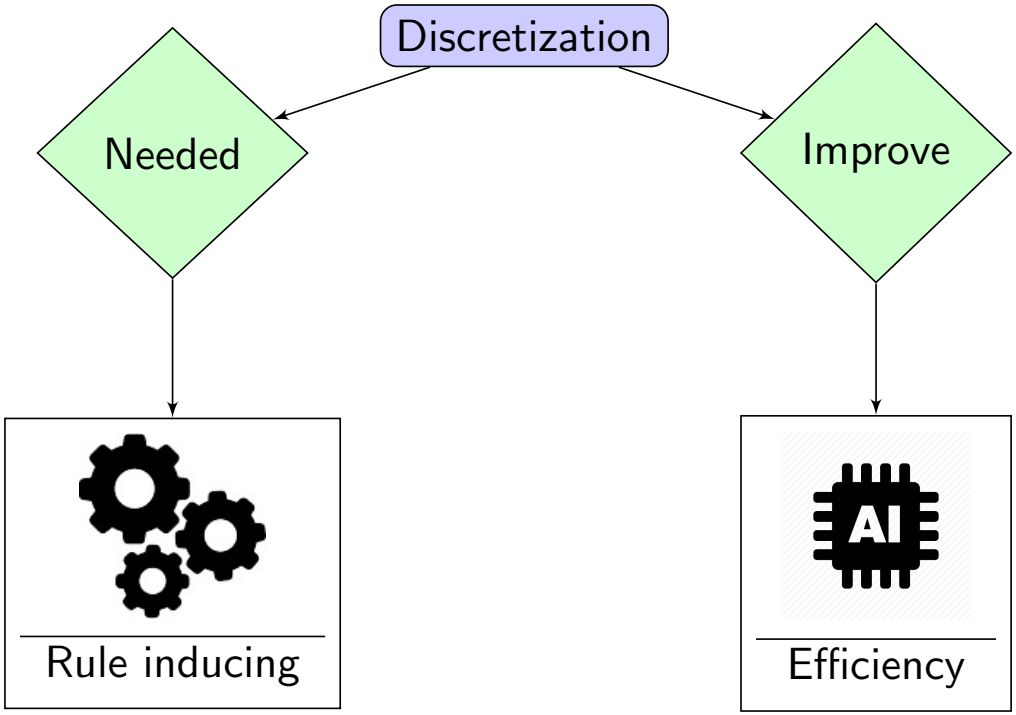
Results





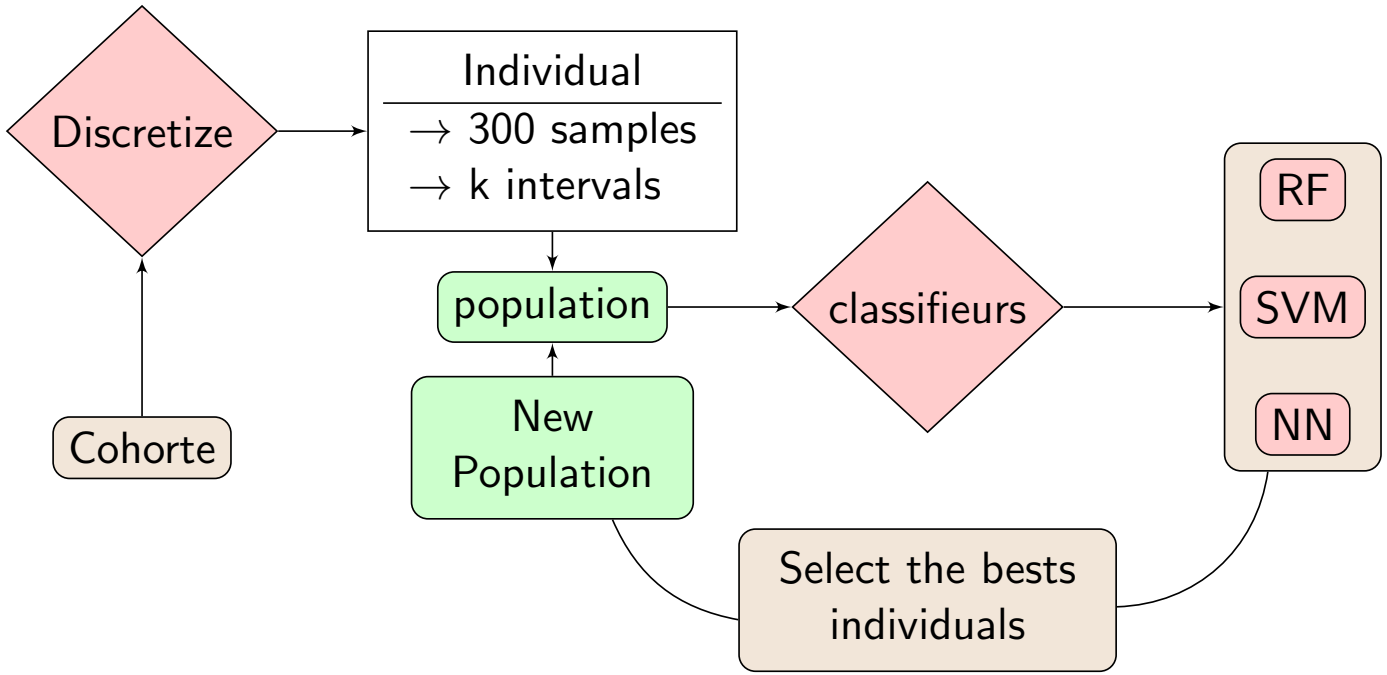
→ Easiest to identify RA than SjS

Why?



→ Some algorithms can handle discrete attributes only
→ correct skewed distribution
→ reduce the influence of outliers

First Attempt



⇒ Learn the optimal value of k
Where k $\in \{2, \ldots, max_k\}$

The Ameva Algorithm

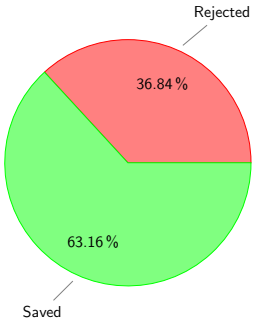
$$Ameva(k) = \frac{\chi^2(k)}{k(I-1)} \qquad \chi^2(k) = N(-1 + \sum_{i=1}^{I}\sum_{j=1}^{k}\frac{n_{ij}^2}{n_i n_j})$$

k:number of discrete intervals, I: number of classes

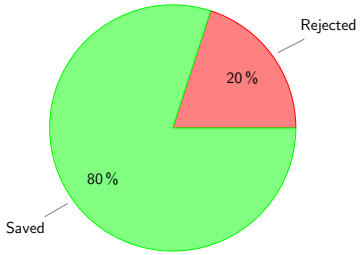
Maximize the dependency relationship
between the class labels and the continuous-values attribute

Minimize the number of intervals k

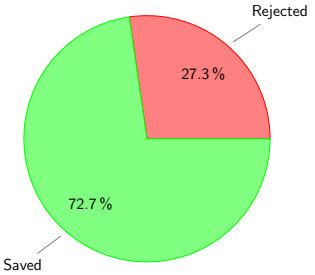
Flow cytometry



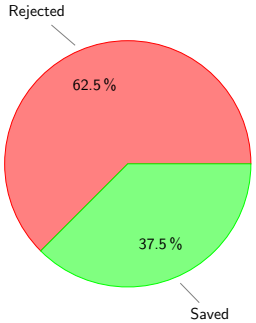
P1, 12 variables saved



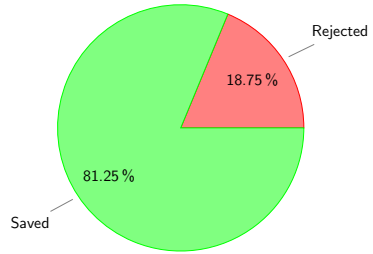
P2, 4 variables saved



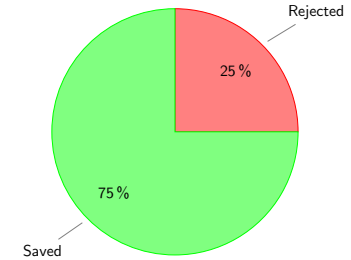
P3, 8 variables saved



P4, 3 variables saved

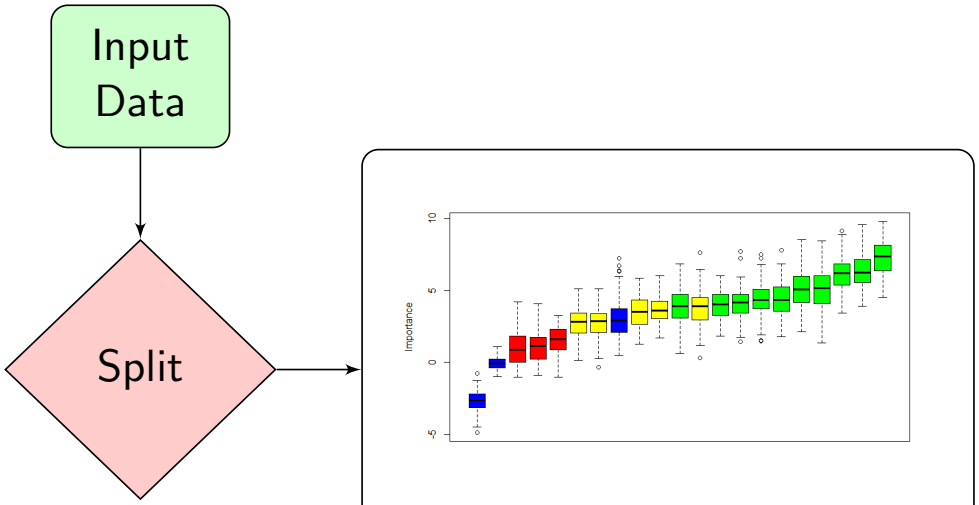


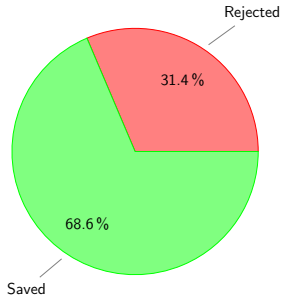
P5, 13 variables saved



P6, 6 variables saved

Flow cytometry

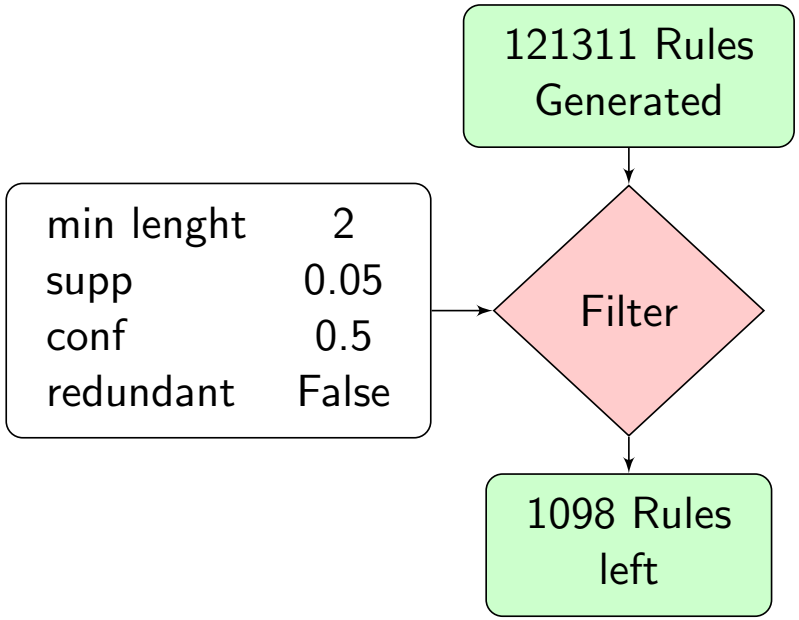


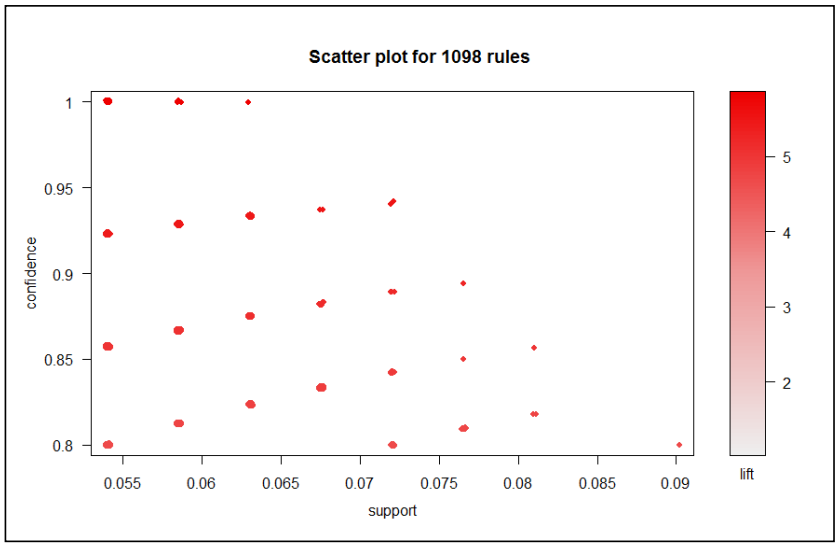


All Panels

→ 21 rejected features
→ 46 saved features

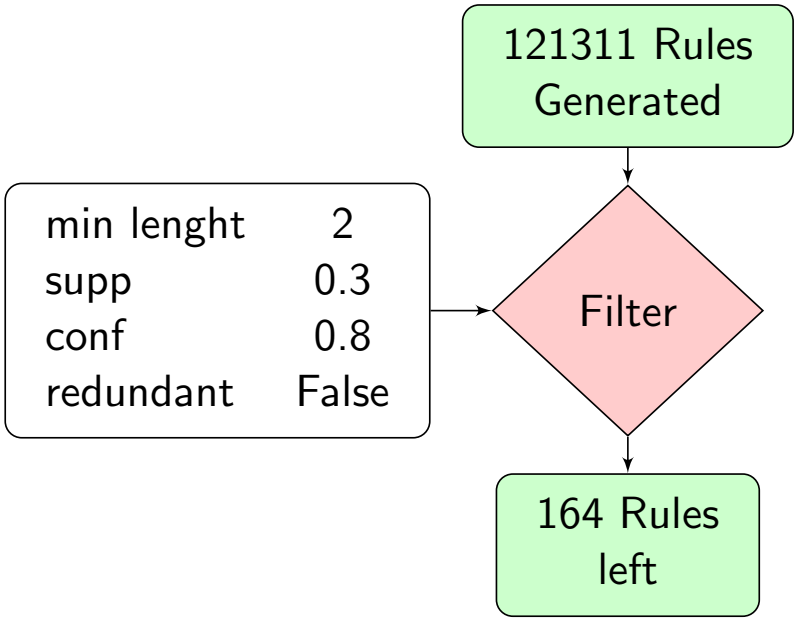
Flow Cytometry

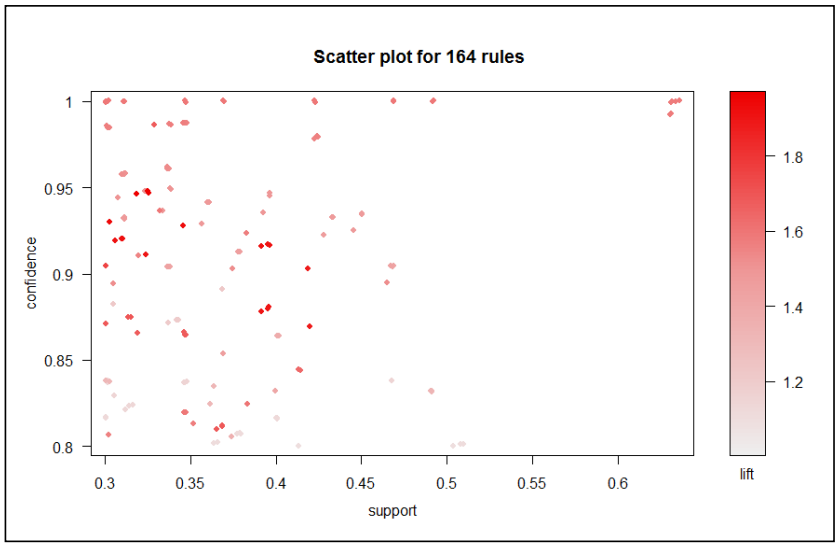




→ High Confidence
→ Low Support

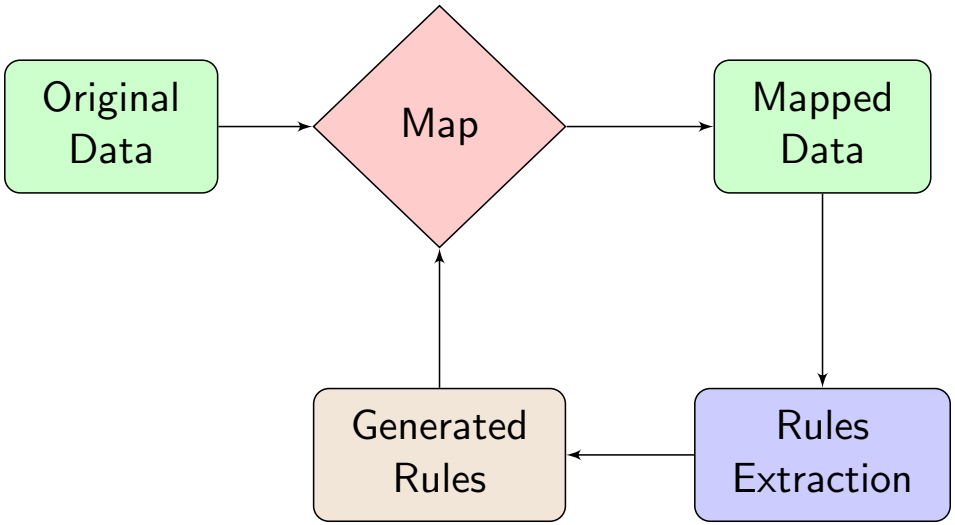
Flow Cytometry





→ High Confidence
→ Better Support

Improvements



Rule based discrimination

$X_i \rightarrow [0, 1]$
Probablistic approach

Perspectives

Perspectives

- Refine the rules
- Implement the inference engine
- More data
- Include Inception