



# Contribution of the **xtractis**<sup>®</sup> methodology to the automatic extraction of robust fuzzy models

Application to the prediction of consumer liking and sensory evaluation  
and to the optimization of product formulation

**Agrostat 2008**

**Z. Zalila, J. Cuquemelle, A. Chikh, C. Penet, B. Lorentz, D. Deschamps**

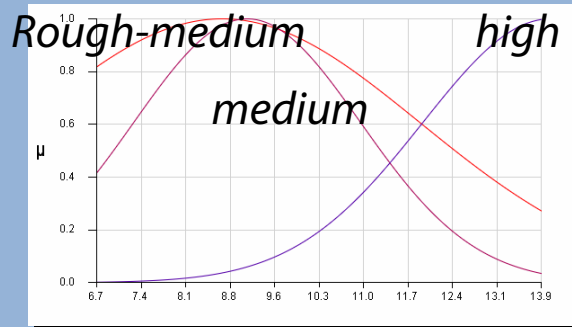
January 24, 2008

# Fuzzy model

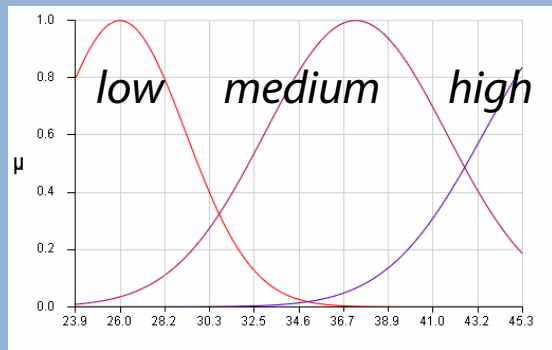
fusion of symbolic and numerical approaches



## Partitions



Input 1: Total Acidity



Input 2: Sum of Sugars

## Rules

### Rule ①

If *Total Acidity* is *rough-medium*  
And *Sum of Sugars* is *low*  
then *Sweet* equals *3.1*

### Rule ②

If *Total Acidity* is *medium*  
And *Sum of Sugars* is *medium*  
then *Sweet* equals *8.1*

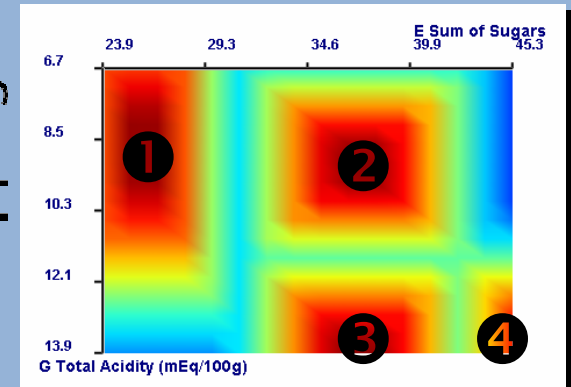
### Rule ③

If *Total Acidity* is *high*  
And *Sum of Sugars* is *medium*  
then *Sweet* equals *3.3*

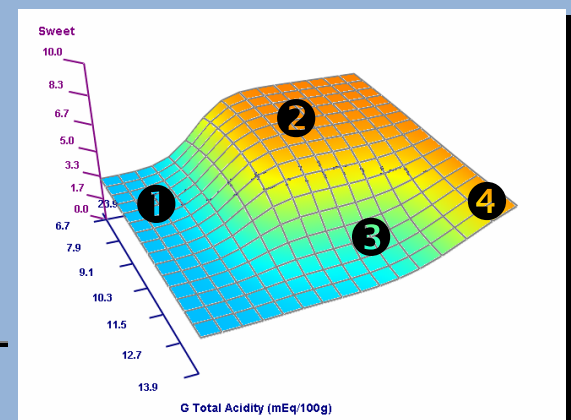
### Rule ④ ...

## Inference

Mapping



Response surface





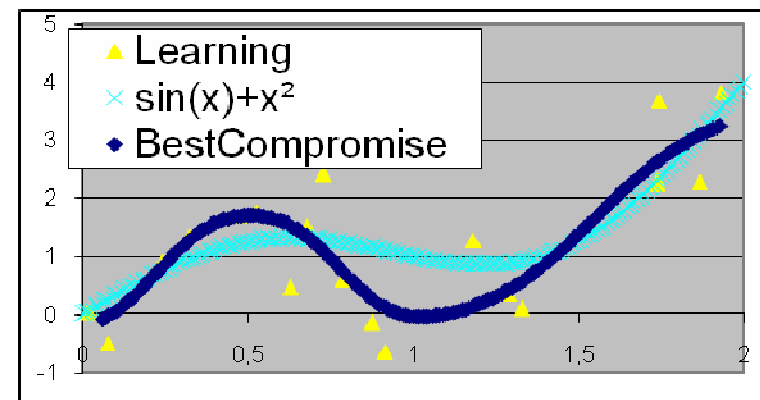
1/2

## Interface between linguistic and numerical computation

- Linguistic and approximate qualification of variables
- Interpretability thanks to a linguistic rule-based structure
- Approximate reasoning algorithms link a numerical function with the linguistic structure
- Unification of qualitative and quantitative approaches

## Handling of low quality data

- Fuzzy data (imprecise, uncertain, subjective)
- Incomplete data (no imputation needed for missing values)
- Noisy data





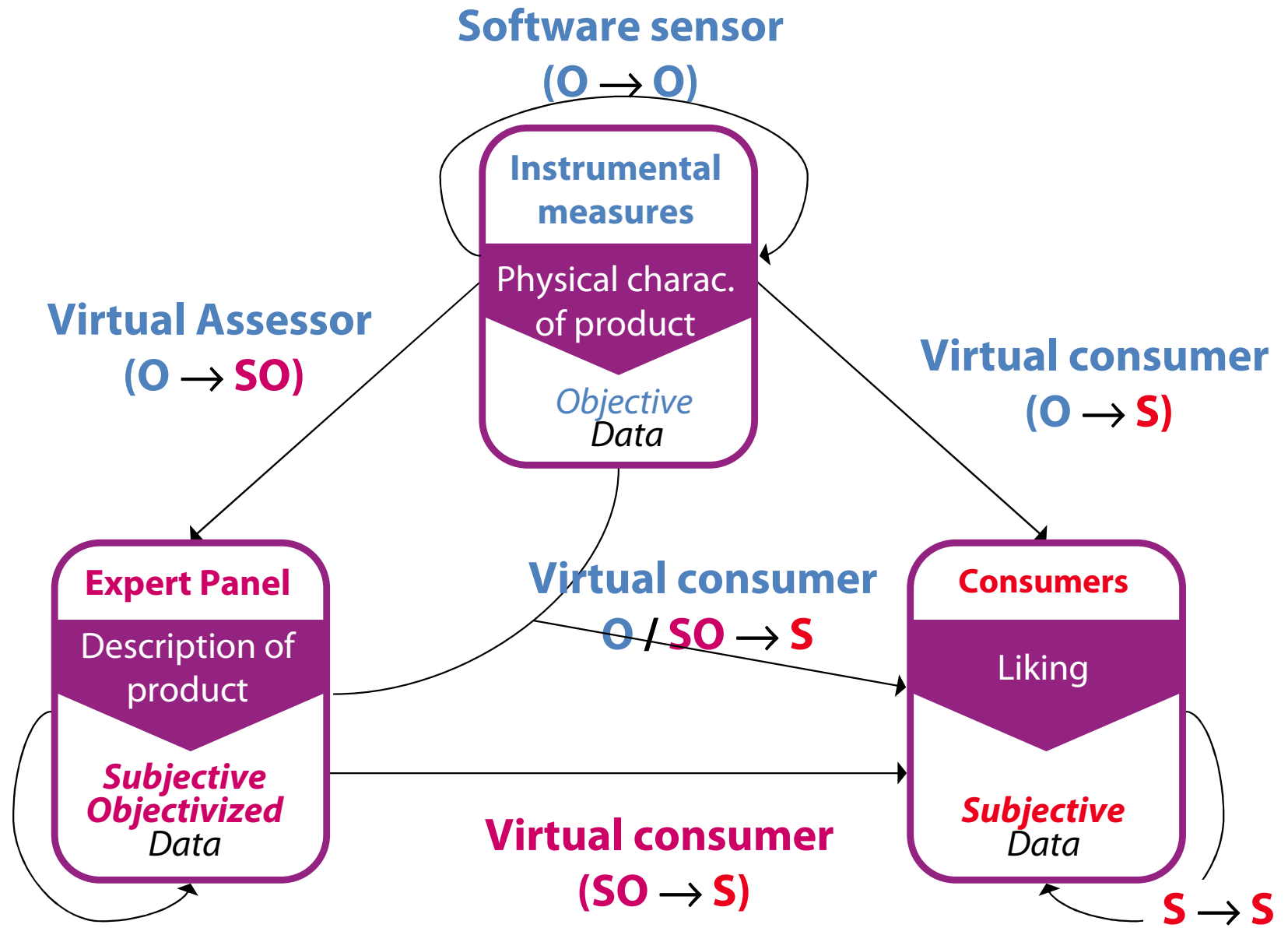
2/2

## Mathematical properties of fuzzy models

- Universal approximator of non linear functions [Kosko, 1992]
- Locality: each rule defines a relationship between a fuzzy area in the input space and a value of the output space
- Granularity: the area of influence of each rule may be as small / large as needed
- Locality and granularity allow a compact representation in large input spaces
- Possibility theory (efficient replacement of probability theory when data is sparse / of low quality) [Dubois & Prade, 1994]

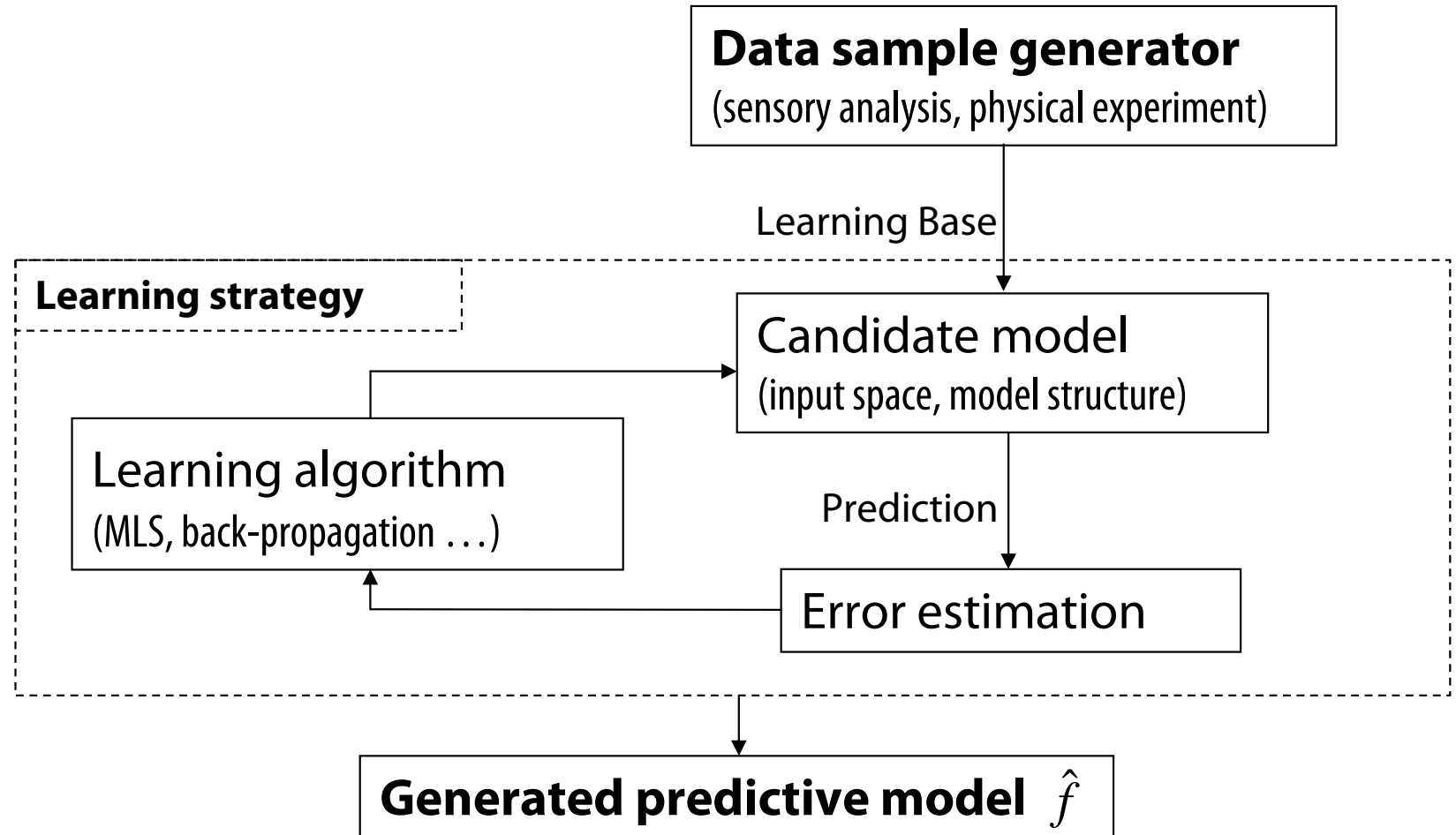
Probability theory	Possibility theory
$P(A) = \int_{u \in A} p(u) du$ Crisp event	$\Pi(A) = \sup_{u \in U} (\min(\mu_A(u), \pi(u)))$ Fuzzy event
	$\Pi(A) = \sup_{u \in A} (\pi(u))$ Crisp event
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	$\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$

# Data and model types



# Knowledge extraction

## model induction from data



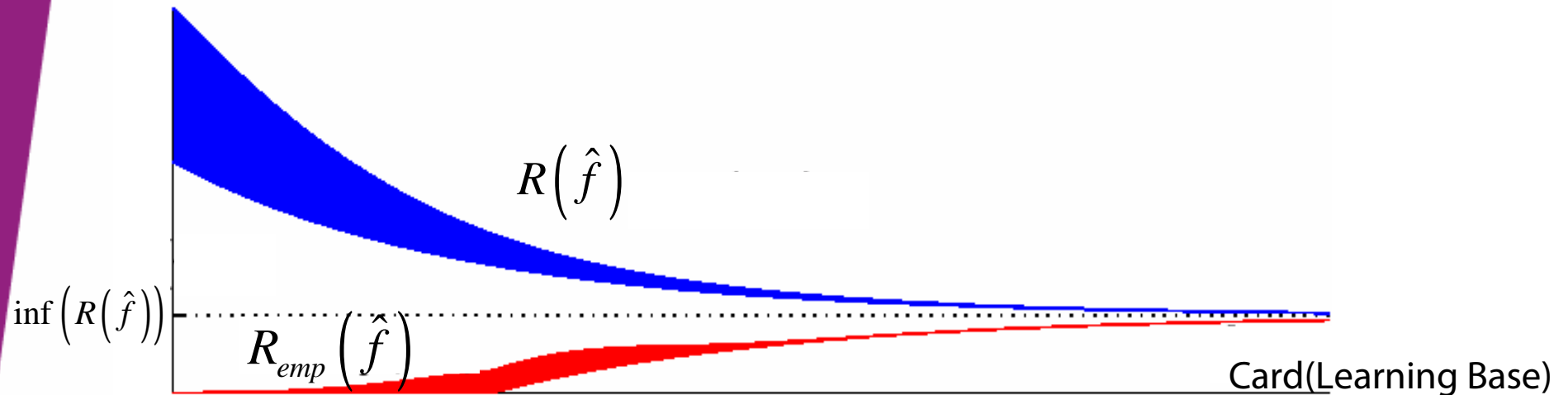
### Important remark

- Every data-driven computation (including input space transformation / simplification) is part of the learning strategy



# Actual Performance of the generated model

- Empirical risk  $R_{emp}$
- ○ Prediction error computed on the learning sample distribution
- Actual risk  $R$
- ○ Prediction error on the whole population (usually impossible to compute)
- ○ Several available estimators



Evolution of  $R$  and  $R_{emp}$  with the size of learning samples (for a consistent learning strategy)  
[Vapnik, 2000]

# Overfitting management implemented in *xtractis*<sup>®</sup>



## Dimensionality reduction

- Only uses original variables (no creation of new axes)
- Learning algorithms handling problems with high dimensionality (hundreds/thousands of variables) allow a top-down procedure for variable selection
- This procedure preserves the synergies between groups of variables

## Regulation methods for overfitting reduction

- Control the complexity of the candidate (removal and merging of similar fuzzy classes)
- Smooth the response of the candidate (noise injection on inputs to avoid strong non linearities)
- Supervise the learning algorithm with several constraints

## Validation method for robustness assessment

- Insufficiency of a single validation sample
- Leave One Out (LOO)
- Monte-Carlo Cross Validation (MC-CV, performed with different sizes of validation samples)

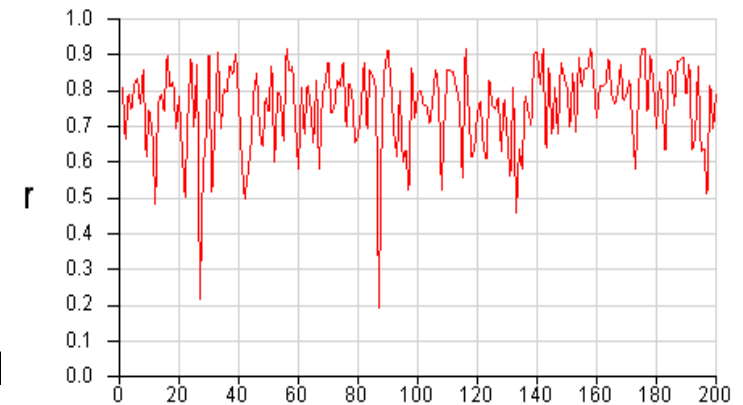
# Validation methods

1/3



## Single validation sample

- Several drawbacks of “2/3 - 1/3 split”:
  - 1/3 data not used for learning
  - High sensitivity to choice of validation set
- *Example:* estimation for the same learning strategy applied on 200 randomly selected learning sets (leave 30% validation points out)

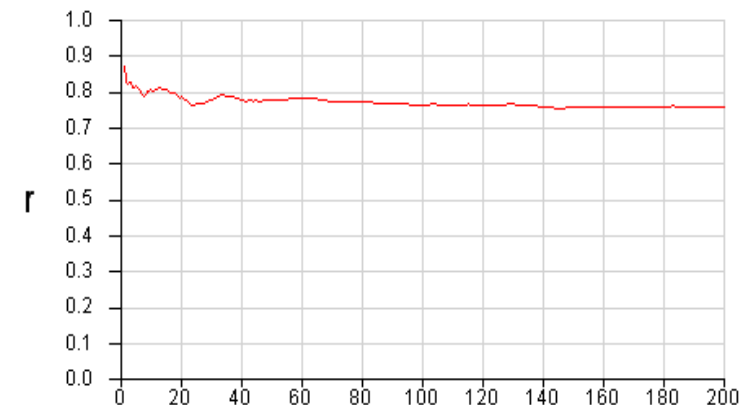


Variability of validation with single sample

➔ **Impossibility to assess robustness of the model, except when a lot of data is available**

## Cross-Validation

- Aggregates the outcome of the learning strategy on a large number of different learning/validation splits
- *Example:* robustness of the previous learning strategy on 200 random splits (leave 30% validation points out)



Convergence of Cross-Validation

# Validation methods



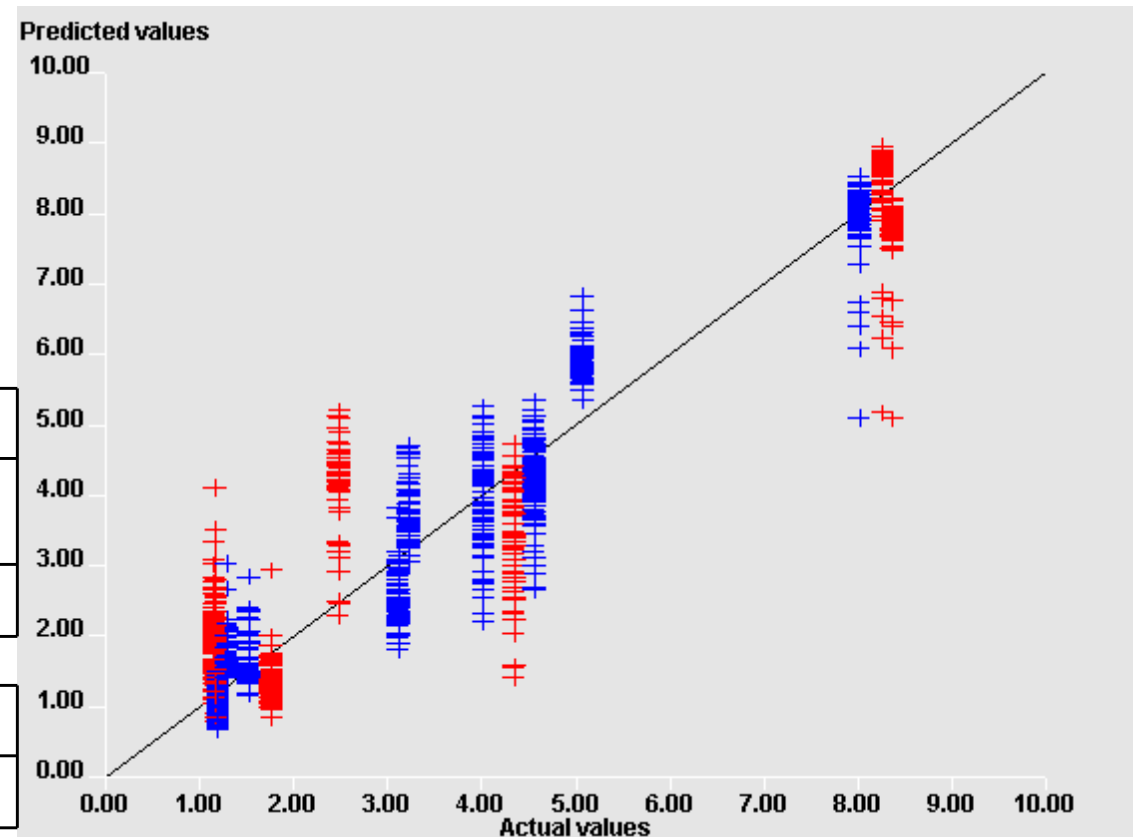
2/3

Example of MC-30%

200 random splits of 6 validation points → 1 200 predictions in the scatter

<b>Correlation coefficient</b>	<b>0.943</b>
<b>Mean error (predicted - actual)</b>	<b>0.109</b>
<b>Error standard deviation</b>	<b>0.804</b>

<b>Hamming's error</b>	<b>6.12 %</b>
<b>Maximum error</b>	<b>32.7 %</b>



# Validation methods

## 3/3 - key ideas on Cross-Validation



### Main characteristics

- Whole database used for model generation AND model validation
- Robustness estimation computed on the scatter of validation points of all models (convergence ~ 50 to 300 iterations)
- Insensitiveness to an arbitrary choice of learning and validation base
- Computationally intensive (each iteration involves a model generation)

### Variance of estimates

- High variance of estimate if  $N$  is low ( $N = \text{Card}(\text{Learning Base})$ )
- High variance of estimate with few data splits (LOO or few iterations of Monte Carlo)
- Single sample > Disjoint sets > LOO > MC ( $p$  small) > MC ( $p$  large)

### Bias of estimates

- $R(N)$ : actual risk of a model generated with  $N$  learning points
- Any Cross Validation leaving  $p$  points out is an unbiased estimator of  $R(N-p)$  [Shao, 1993], and thus a pessimistic estimator of  $R(N)$

# Heuristics of modelling with xtractis<sup>®</sup>



- Insufficiency of  $R_{emp}$  as performance evaluation
- Difficulty to select a good estimator of generalization error
  - ○ Strong variance of least biased estimators (p small)
  - ○ Strong bias of more robust estimators (p large)
- **xtractis<sup>®</sup>** approach
  - ○ Evaluate a large number of learning strategies (involving regulation techniques)
  - ○ Evaluation performed with several CV methods (LOO, MC-10%, -20%, -30% ...)
  - ○ Select models that give the best compromise between all estimators of  $R$ , low complexity and interpretability criteria
- **Benefits**
  - ○ Relying on several robust estimators helps to reduce the effects of variance of one single estimator
  - ○ Models ranking not perturbed by the bias of the estimators
  - ○ This heuristic extracts models among the best possible given the available information

# Modelling results

## 1/4 - tomato liking prediction



### Segment 1

- Better performance of "O/SO → S" modelling over standard "SO → S", but at the price of a higher model complexity

Type	Nb. of		Variables	Error			Remarks
	var.	rules		Correlation	Hamming	Max.	
SO-S	4	2	Firm_inside Juicy Melly Tomato_flavor	0.984	3.57%	10.90%	Gen. 1568
				0.954	5.55%	23.65%	Monte Carlo 15 %
				0.943	6.12%	32.75%	Monte Carlo 30 %
OSO-S	6	2	Average_weight G_Total_acidity Firm_inside Juicy Sweet Tomato_flavor	0.999	0.81%	2.94%	Gen. 697
				0.971	4.40%	19.02%	Monte Carlo 15 %
				0.938	5.88%	37.97%	Monte Carlo 30 %

# Modelling results

## 2/4 - tomato liking prediction



### Segment 2

- Slightly better performance of "O/SO → S" modelling over standard "SO → S", but biggest benefit is a complexity reduction

Type	Nb. of		Variables	Error			Remarque
	var.	rules		Correlation	Hamming	Max.	
SO-S	5	3	Firm_inside	0.909	5.03%	14.39%	Gen. 2144
			Tomato_odor	0.81	7.42%	30.10%	Monte Carlo 15 %
			Firm Mealy Tomato_flavor	0.76	8.44%	38.60%	Monte Carlo 30 %
OSO-S	4	2	G_Total_acidity	0.946	4.37%	10.46%	Gen. 4445
			Ext_color	0.802	8.14%	23.65%	Monte Carlo 15 %
			Firm_inside Mealy	0.786	8.44%	28.08%	Monte Carlo 30 %

# Modelling results

## 3/4 - tomato liking prediction



### Segment 3

No better "O/SO → S" model than "SO → S"

Output	Nb. of		Variables	Error			Remarque
	var.	rules		Correlation	Hamming	Max.	
SO-S	2	2	Juicy Skin width	0.897	7.20%	31.73%	Gen. 1325
				0.874	8.77%	32.73%	Monte Carlo 15 %
				0.808	10.42%	42.31%	Monte Carlo 30 %

# Modelling results

## 4/4 - tomato liking prediction



### Segment 4

- Similar performance of "O/SO → S" modelling over standard "SO → S", but complexity reduction

Type	Nb. of		Variables	Error			Remarque
	var.	rules		Correlation	Hamming	Max.	
SO-S	3	3	Firm_inside Firm Skin_width	0.964	5.33%	11.70%	Gen. 1591
				0.867	10.16%	40.00%	Monte Carlo 15 %
				0.747	13.58%	57.23%	Monte Carlo 30 %
OSO-S	3	2	M_Total_acidity G_Sum_of_sugars Mealy	0.958	6.41%	14.03%	Gen. 9050
				0.83	11.07%	43.44%	Monte Carlo 15 %
				0.784	11.81%	62.54%	Monte Carlo 30 %

# Model exploitation

## example of product optimization



ELEMENTARY REQUESTS				
Search	liking(S1)-02_C1_4v_2r_SO-S	max		ER 1
and	liking(S2)-03_C1_4v_2r_OS0-S	max		ER 2
and	liking(S3)-04_C1_2v_2r_SO-S	≥	~8.00 (~8.00 = tri(8.00 8.00 8.61))	ER 3
and	liking(S4)-05_C1_3v_3r_OS0-S	max		ER 4

INPUT CONSTRAINTS				
Where	M Total Acidity (mEq/100g)	∈	[3.87 ; 7.07]	IC 1

**Inputs**  
Request satisfaction degree:

M Total Acidity (mEq/100g)	<input type="text" value="6.92"/>	3.87		7.07
G Total Acidity (mEq/100g)	<input type="text" value="13.60"/>	6.80		13.60
G Sum of Sugars	<input type="text" value="22.1"/>	22.1		41.2
Ext_Color	<input type="text" value="8.43"/>	3.50		8.43
Firm_Inside	<input type="text" value="4.97"/>	4.96		7.77
Juicy	<input type="text" value="8.20"/>	2.45		8.21
Melty	<input type="text" value="3.88"/>	2.15		7.88
Mealy	<input type="text" value="5.12"/>	0.50		7.44
Skin_Width	<input type="text" value="3.97"/>	3.28		7.00
Tomato_Flavor	<input type="text" value="7.90"/>	3.69		8.00

**Outputs**

				Mapping	
liking(S1)-02_C1_4v_2r_SO-S	<input type="text" value="8.21"/>	0.00		10.00	<input type="text" value="0.82"/>
liking(S2)-03_C1_4v_2r_OS0-S	<input type="text" value="8.86"/>	0.00		10.00	<input type="text" value="0.91"/>
liking(S3)-04_C1_2v_2r_SO-S	<input type="text" value="8.60"/>	0.00		10.00	<input type="text" value="0.57"/>
liking(S4)-05_C1_3v_3r_OS0-S	<input type="text" value="8.66"/>	0.00		10.00	<input type="text" value="0.43"/>

Fuzzy multi objective request

Optimal solution

Constraints on inputs

Verification of the request

# Conclusion

## widened perspectives



### Innovative model extraction method

- Benefits of fuzzy theory, efficient learning algorithms and robustness estimation
- Merging of quantitative and qualitative approaches
- Robustness analyses ensure confidence in the model validity

### New insights for data analysis

- Understand complex processes or behaviours
- Instantly predict potential success of new formulations
- Automatically identify candidate formulations matching a set of objectives
- Explore novel dataset combinations and relationships (Sensory / Instrumental / Liking):  
More robust or simpler models of tomato liking prediction using both instrumental (**0**) and sensory (**S0**) datasets

### Universal approach of Automatic Knowledge Discovery from Data (AKDD)



**xtractis<sup>®</sup>**

Universal Solution for Modelling and Optimization  
of Complex Processes

**For more information, visit us at:  
[www.xtractis.fr](http://www.xtractis.fr)**