Workshop virtuel - Nouvelles méthodes pour l'analyse descriptive et prédictive de données massives et structurées
25 septembre 2020

Towards the Automation of Data Analysis for Large Scale Relational Data

Marc Boullé, Orange Labs

September 25, 2020

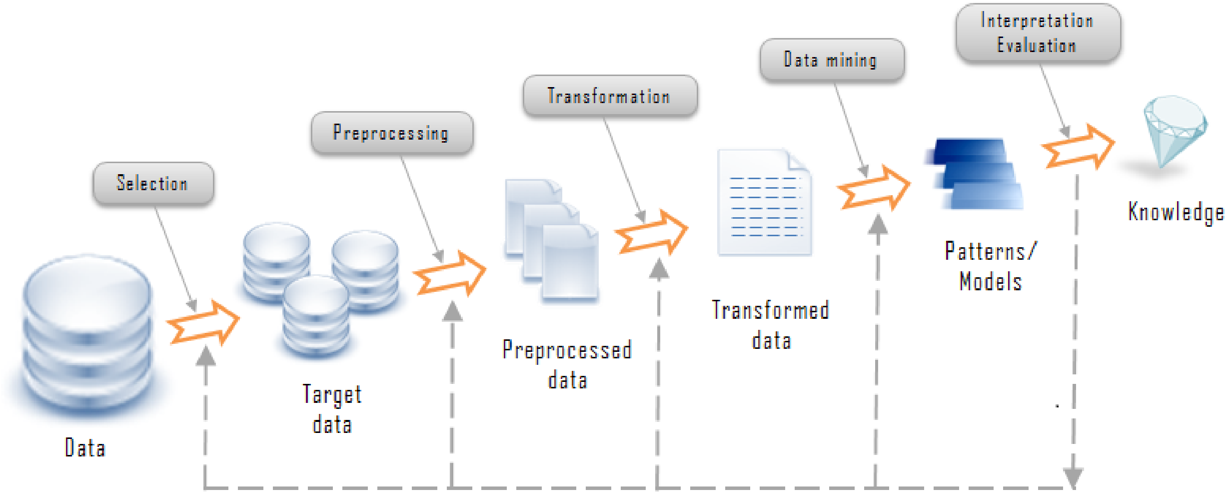
Data Mining in Orange

Example of use case

- Marketing campaigns
  - Objective: scoring
    - churn, appetency, up-selling…
  - Many domains
    - Marketing, Text mining, Web mining,
      Traffic classification, Sociology, Ergonomics…
  - Millions of instances
  - Multiple tables source data
    - Customer contracts
    - Call detail records (**billions**)
    - Multi-channel customer support
    - External data
    - …
  - Train sample
    - 100 000 instances
    - 10 000 variables (based on expertise)
    - Heavily unbalanced
    - Missing values
    - Thousands of categorical values
    - …
  - Challenge: industrial scale
    - Hundred of scores every month

Data Mining in Orange

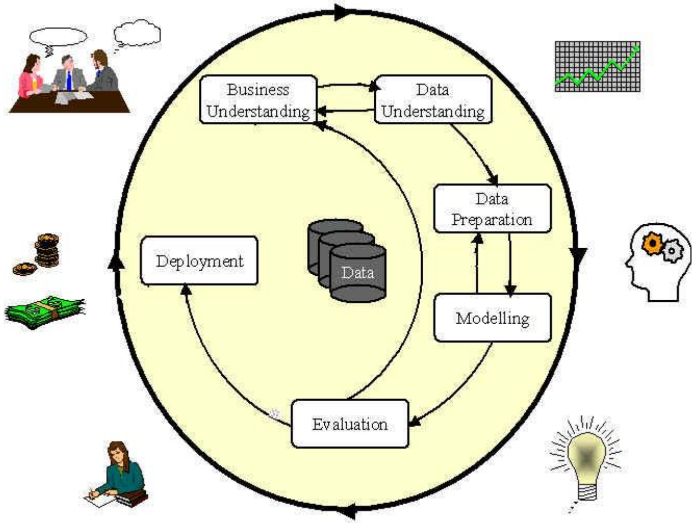
How to efficiently apply data mining techniques in an industrial context?



Objective

- Towards an effective automation of data mining

- Evaluation criterions
  - Genericity
  - No parameter
  - Robustness
  - Accuracy
  - Understandability
  - Scalability



Lift the brakes to the dissemination
With a high-quality tool

Schedule

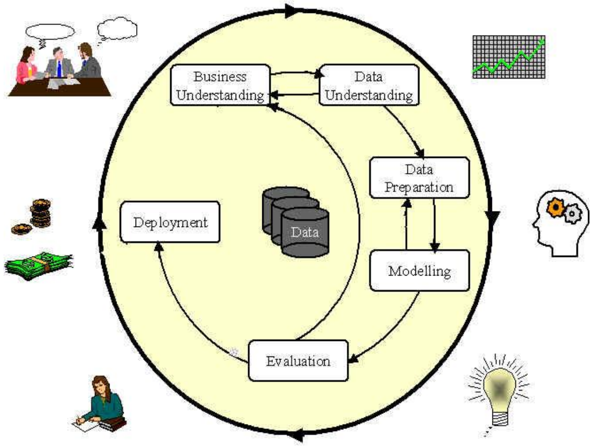
- Automatic data preparation
- Multi-tables data mining
- Automatic variable construction
- Conclusion and future work

Context

- Statistical learning
  - Objective: train a model
    - Classification: the output variable is categorical
    - Regression: the output variable is numerical
    - Clustering: no output variable

- Data preparation
  - Variable selection
  - Search for a data representation

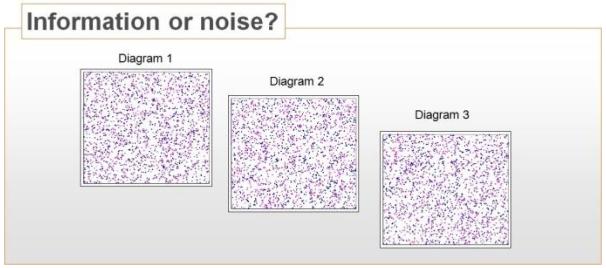
- Data preparation is critical
  - 80% of the process time
  - Requires skilled data analysts



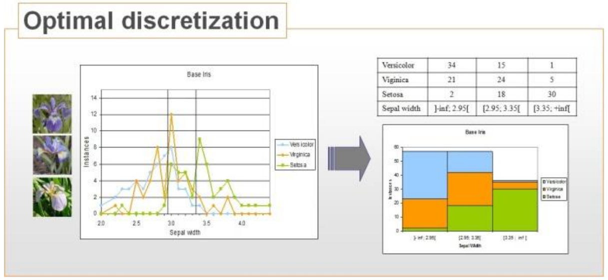
Single-table datasets
instances x variables

Age	Education	Education Num	Marital status	Occupation	Race	Sex	Hours Per week	Native country	...	Class
39	Bachelors	13	Never-married	Adm-clerical	White	Male	40	United-States	...	less
50	Bachelors	13	Married-civ-spouse	Exec-managerial	White	Male	13	United-States	...	less
38	HS-grad	9	Divorced	Handlers-cleaners	White	Male	40	United-States	...	less
53	11th	7	Married-civ-spouse	Handlers-cleaners	Black	Male	40	United-States	...	less
28	Bachelors	13	Married-civ-spouse	Prof-specialty	Black	Female	40	Cuba	...	less
37	Masters	14	Married-civ-spouse	Exec-managerial	White	Female	40	United-States	...	less
49	9th	5	Married-spouse-absent	Other-service	Black	Female	16	Jamaica	...	less
52	HS-grad	9	Married-civ-spouse	Exec-managerial	White	Male	45	United-States	...	more
31	Masters	14	Never-married	Prof-specialty	White	Female	50	United-States	...	more
42	Bachelors	13	Married-civ-spouse	Exec-managerial	White	Male	40	United-States	...	more
37	Some-college	10	Married-civ-spouse	Exec-managerial	Black	Male	80	United-States	...	more
30	Bachelors	13	Married-civ-spouse	Prof-specialty	Asian	Male	40	India	...	more
23	Bachelors	13	Never-married	Adm-clerical	White	Female	30	United-States	...	less
32	Assoc-acdm	12	Never-married	Sales	Black	Male	50	United-States	...	less
...	...	...	...	...	...	...	...	...	...	...

Proposed approach: data grid models



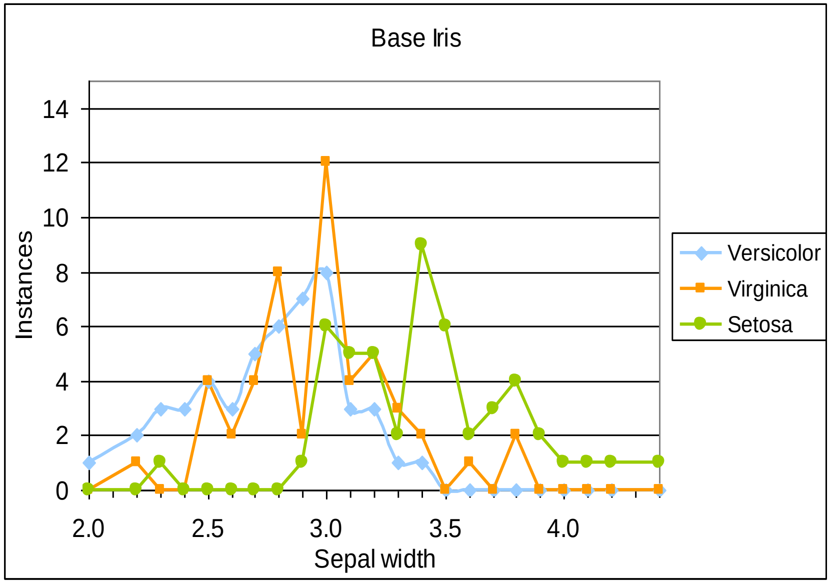
- Objective
  - Evaluate the informativeness of variables
- Data grid models for non parametric density estimation
  - Discretization of numerical variables
  - Value grouping of categorical variables
  - Data grid are the cross-product of the univariate partitions, with a piecewise constant density estimation in each cell of the grid



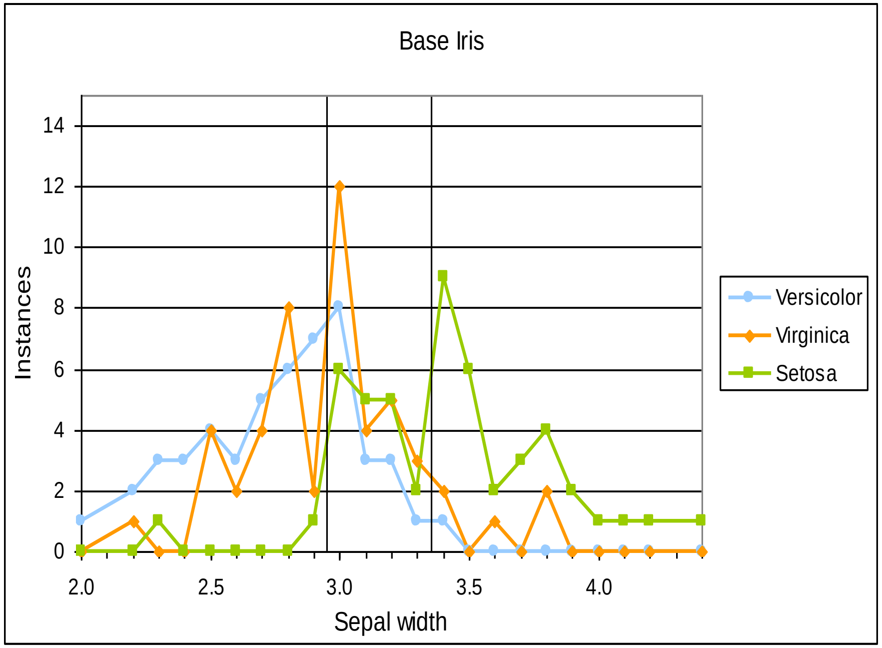
- Modeling approach: MODL
  - Bayesian approach for model selection
    - Minimum Description Length
  - Efficient optimization algorithms

Numerical variables
Univariate analysis using supervised discretization

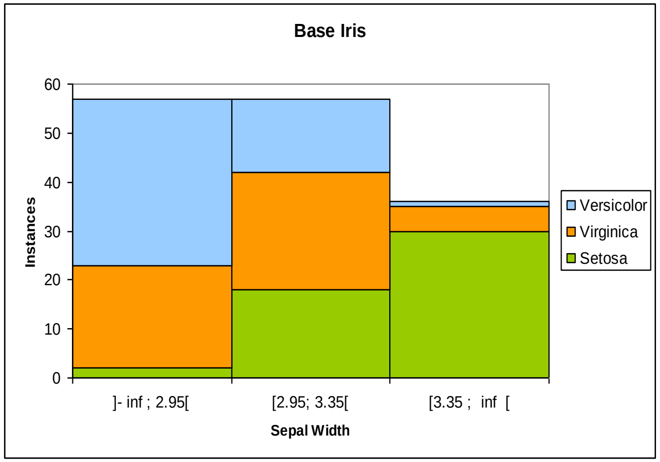
- Discretization:
  - Split of a numerical domain into a set of intervals
- Main issues:
  - Accuracy:
    - Good fit of the data
  - Robustness:
    - Good generalization



Supervised discretization
Model for conditional density estimation



Versicolor	34	15	1
Virginica	21	24	5
Setosa	2	18	30
Sepal width	]- inf ; 2.95[	[2.95; 3.35[	[3.35 ; inf [



How to select the best model?

Formalization

- **Definition:** A discretization model is defined by:
  - the number of input intervals,
  - the partition of the input variable into intervals,
  - the distribution of the output values in each interval.

- **Notations:**
  - N: number of instances
  - J: number of classes
  - I: number of intervals
  - $N_{i.}$: number of instances in the interval i
  - N_{ij}: number of instances in the interval i for class j

Bayesian approach for model selection

- Best model: the most probable model given the data
- Maximize $P(M \mid D) = \dfrac{P(M)P(D \mid M)}{P(D)}$
- Using a decomposition of the model parameters

$$P(M)P(D \mid M) = P(I)P(\{N_i\} \mid I)P(\{N_{ij}\} \mid I, \{N_i\})P(D \mid M)$$

- Assuming independence of the output distributions in each interval

$$P(M)P(D \mid M) = P(I)P(\{N_{i.}\} \mid I)\prod_{i=1}^{I} P(\{N_{ij}\} \mid I, \{N_{i.}\})\prod_{i=1}^{I} P(D_i \mid M)$$

- We now need to evaluate the prior distribution of the model parameters

Prior distribution of the models

- **Definition:** We define the hierarchical prior as follows:
  - the number of intervals is uniformly distributed between 1 et N,
  - for a given number of intervals I, every set of I interval bounds are equiprobable,
  - for a given interval, every distribution of the output values are equiprobable,
  - the distributions of the output values on each input interval are independent from each other.

- Hierarchical prior, uniformly distributed at each stage of the hierarchy

Optimal evaluation criterion MODL

- **Theorem**: A discretization model distributed according the hierarchical prior is Bayes optimal for a given set of instances if the following criterion is minimal:

$$\underbrace{\log(N) + \log\binom{N+I-1}{I-1} + \sum_{i=1}^{I} \log\binom{N_{i.}+J-1}{J-1}}_{\text{prior}} + \underbrace{\sum_{i=1}^{I} \log\left(N_{i.}!/N_{i1}!N_{i2}!\ldots N_{iJ}!\right)}_{\text{likelihood}}$$

N: number of instances
J: number of classes
I: number of intervals
$N_{i.}$: number of instances in the interval i
N_{ij}: number of instances in the interval i for class j

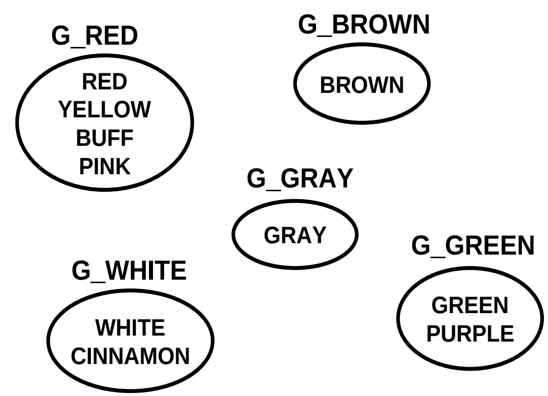
  - 1° term: choice of the number of intervals
  - 2° term: choice of the bounds of the intervals
  - 3° term: choice of the output distribution Y in each interval
  - 4° term: likelihood of the data given the model

- Combinatorial heuristics to retrieve the best model in O(N log N)

Categorical variables
Univariate analysis using value grouping

Cap color	EDIBLE	POISONOUS	Frequency
BROWN	55.2%	44.8%	1610
GRAY	61.2%	38.8%	1458
RED	40.2%	59.8%	1066
YELLOW	38.4%	61.6%	743
WHITE	69.9%	30.1%	711
BUFF	30.3%	69.7%	122
PINK	39.6%	60.4%	101
CINNAMON	71.0%	29.0%	31
GREEN	100.0%	0.0%	13
PURPLE	100.0%	0.0%	10

Cap color	EDIBLE	POISONOUS	Frequency
G_RED	38.9%	61.1%	2032
G_BROWN	55.2%	44.8%	1610
G_GRAY	61.2%	38.8%	1458
G_WHITE	69.9%	30.1%	742
G_GREEN	100.0%	0.0%	23



MODL approach

- Density estimation using data grids
  - Discretization of numerical variables
  - Value grouping of categorical variables
  - Density estimation based on data grid models, with piecewise constant density per cell
  - Strong expressiveness
- Model selection
  - Bayesian approach for model selection
  - Hierarchical prior for the model parameters
  - Exact analytical criterion
- Optimization algorithm
  - Combinatorial algorithms
  - Heuristic exploiting the sparseness of the data grids and the additivity of the criterion
  - Efficient implementation

Genericity of the data grid models

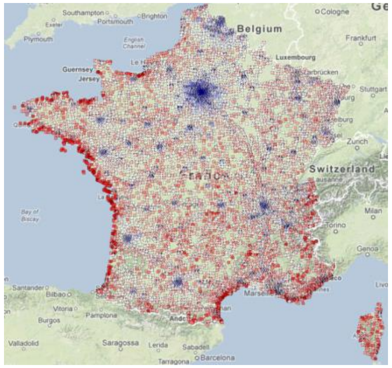
	Univariate	Bivariate	Multivariate
Classification *Y* categorical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \ldots, X_K)$
Regression *Y* numerical	$P(Y \mid X)$	$P(Y \mid X_1, X_2)$	$P(Y \mid X_1, X_2, \ldots, X_K)$
Clustering	–	$P(Y_1, Y_2)$	$P(Y_1, Y_2, \ldots, Y_K)$

K-coclustering of variables

Joint density estimation: $P(Y_1, Y_2, \ldots, Y_K)$

- **Bi-clustering:** $P(Y_1, Y_2)$
  - Text clustering
    - Y_1: texts, Y_2: words
  - Graph clustering
    - Y_1: source nodes, Y_2: target nodes
  - Web mining
    - Web usage mining (logs)
    - Web structure mining
  - Market basket analysis
    - Y_1: customers, Y_2: products
  - Spatial data
    - ex: geographical distribution of industries
    - Y_1: code NAF, Y_2: code Iris

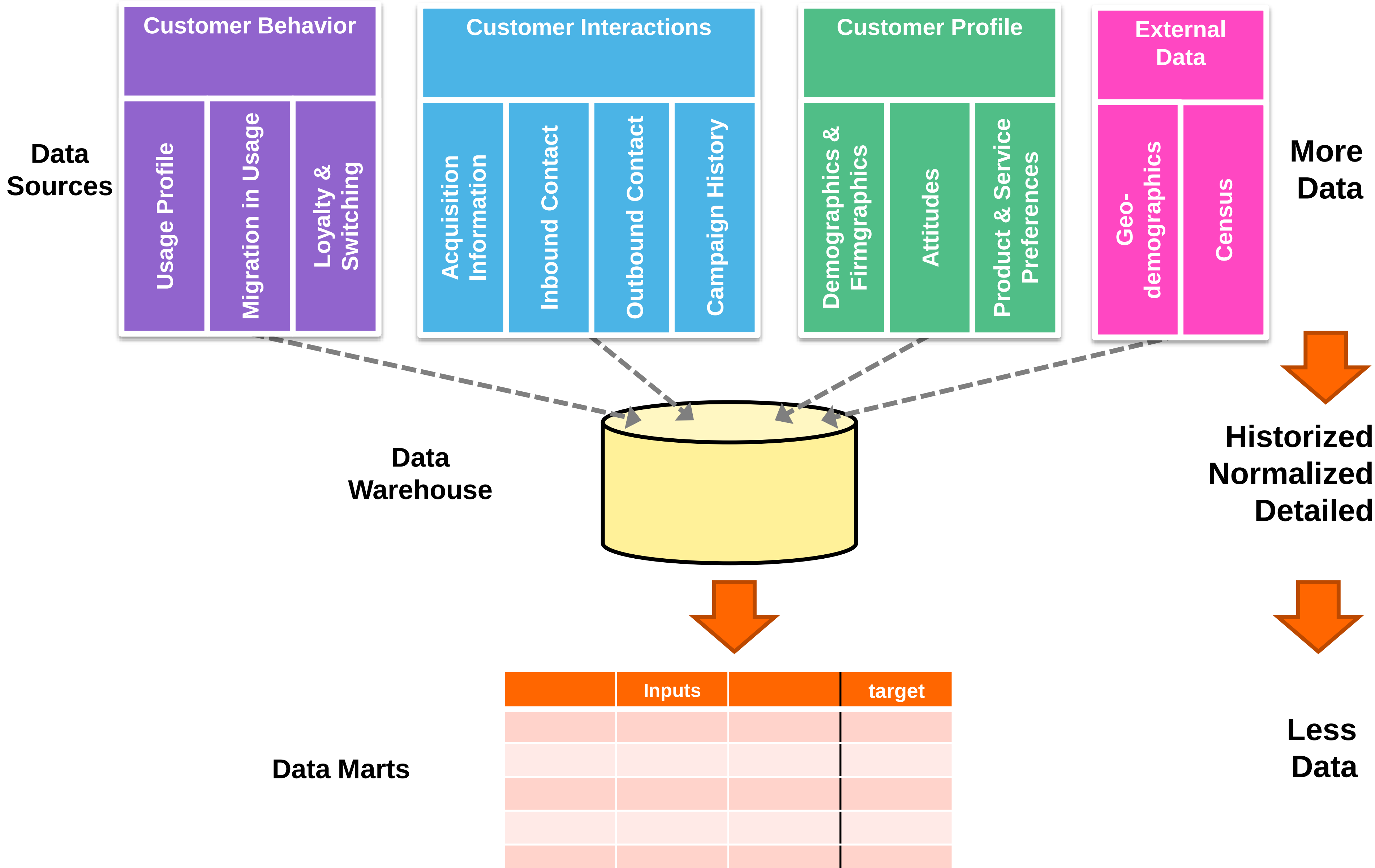
- **Tri-clustrering:** $P(Y_1, Y_2, Y_3)$
  - Temporal graph clustering
    - Y_1: source nodes, Y_2: target nodes
    - Y_3 : timestamp
  - Curve clustering, time series
    - Y_1: curve ID
    - (Y_2, Y_3): (X, Y) curve point
  - Spatio-temporal data
    - ex: Rental bike service
    - ex: Call detail records



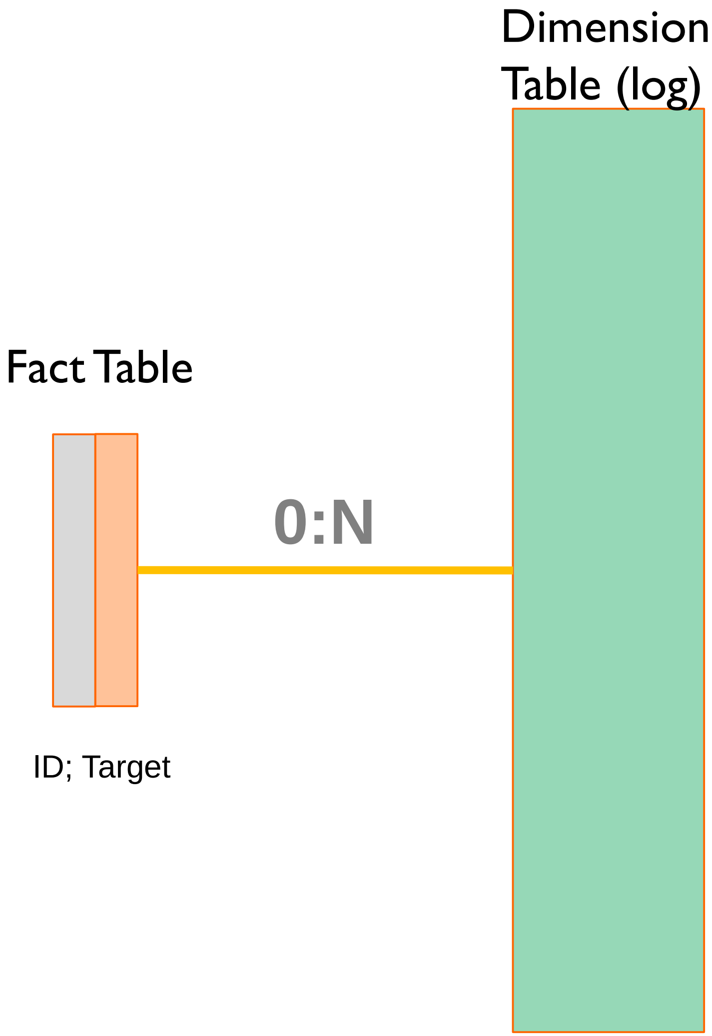
Schedule

- Automatic data preparation
- Multi-tables data mining
- Automatic variable construction
- Conclusion and future work

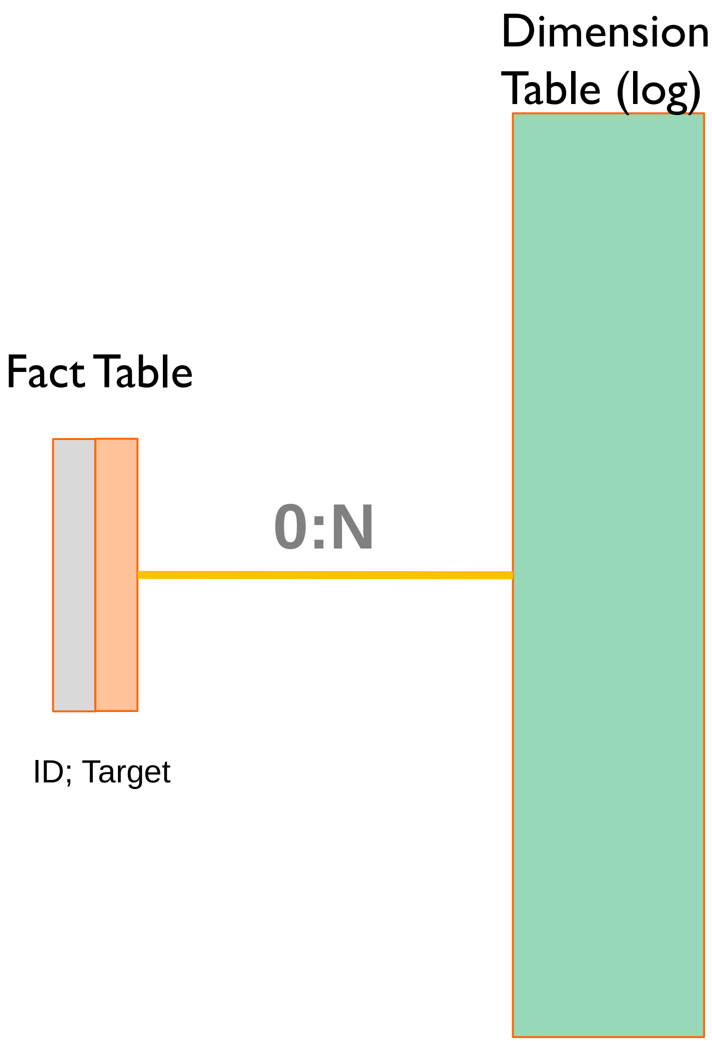
Where does data come from?



Big Data = relational data!



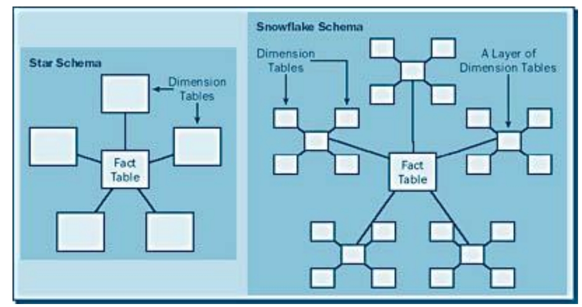
Big Data = relational data!



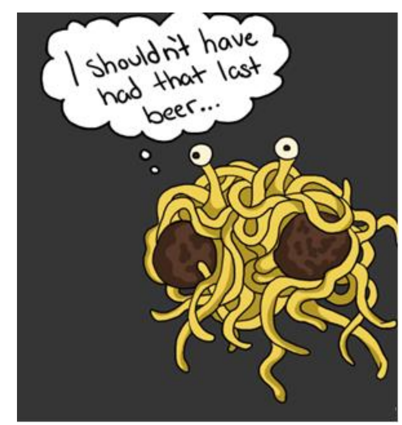
Generalization

Star Schema

Snowflake Schema

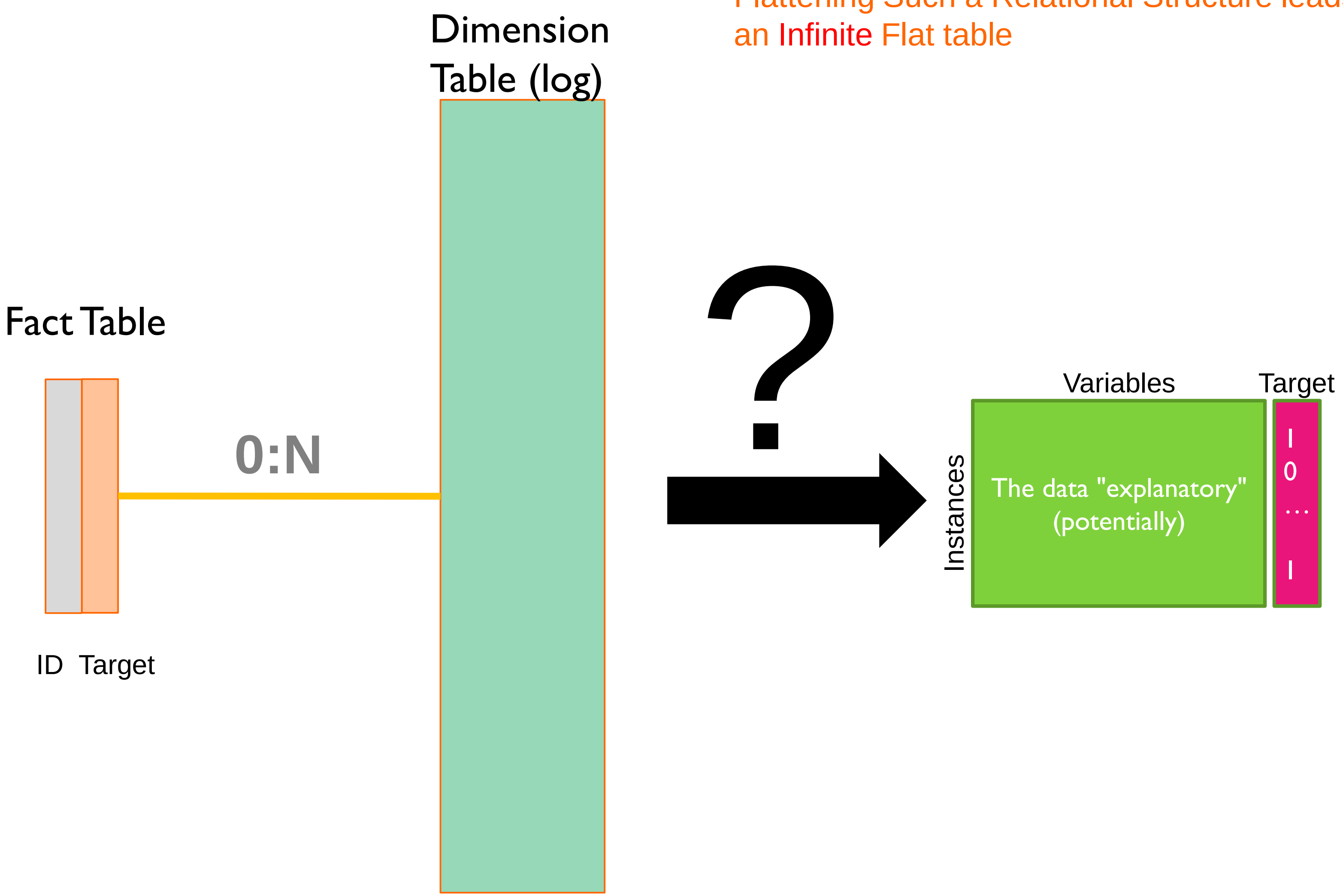


More complex structures are not considered (Yet):

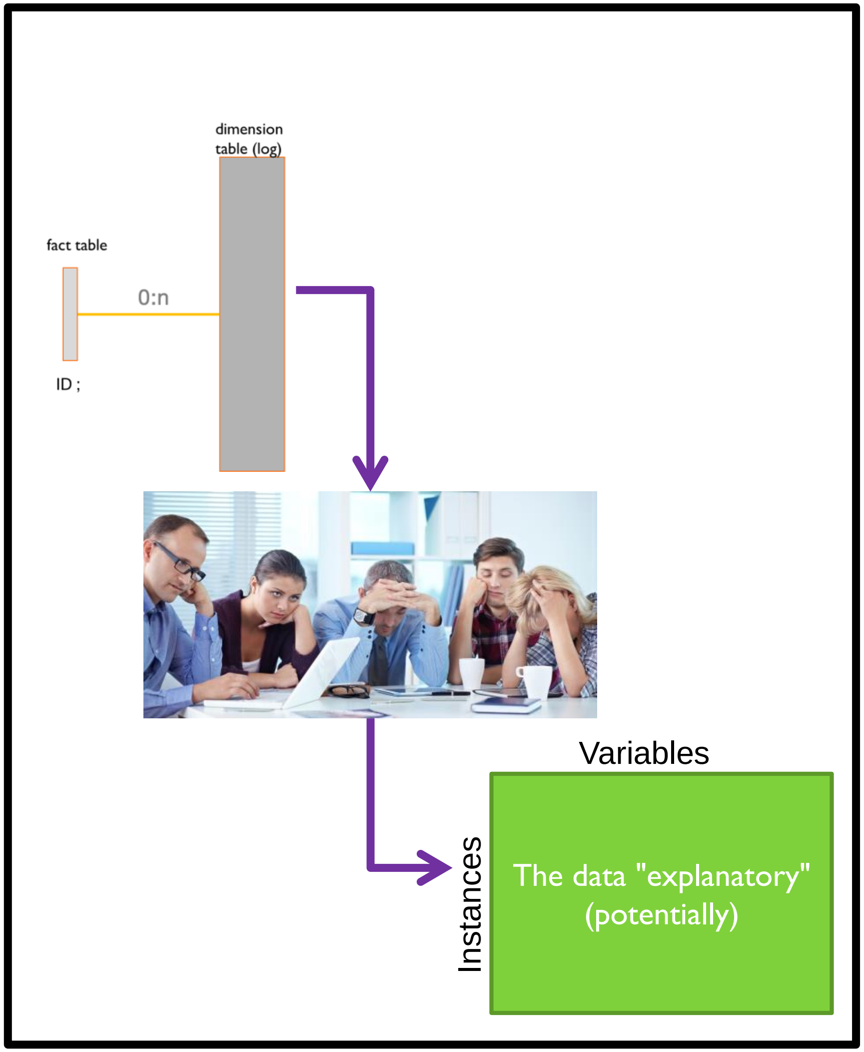


Big Data = relational data!

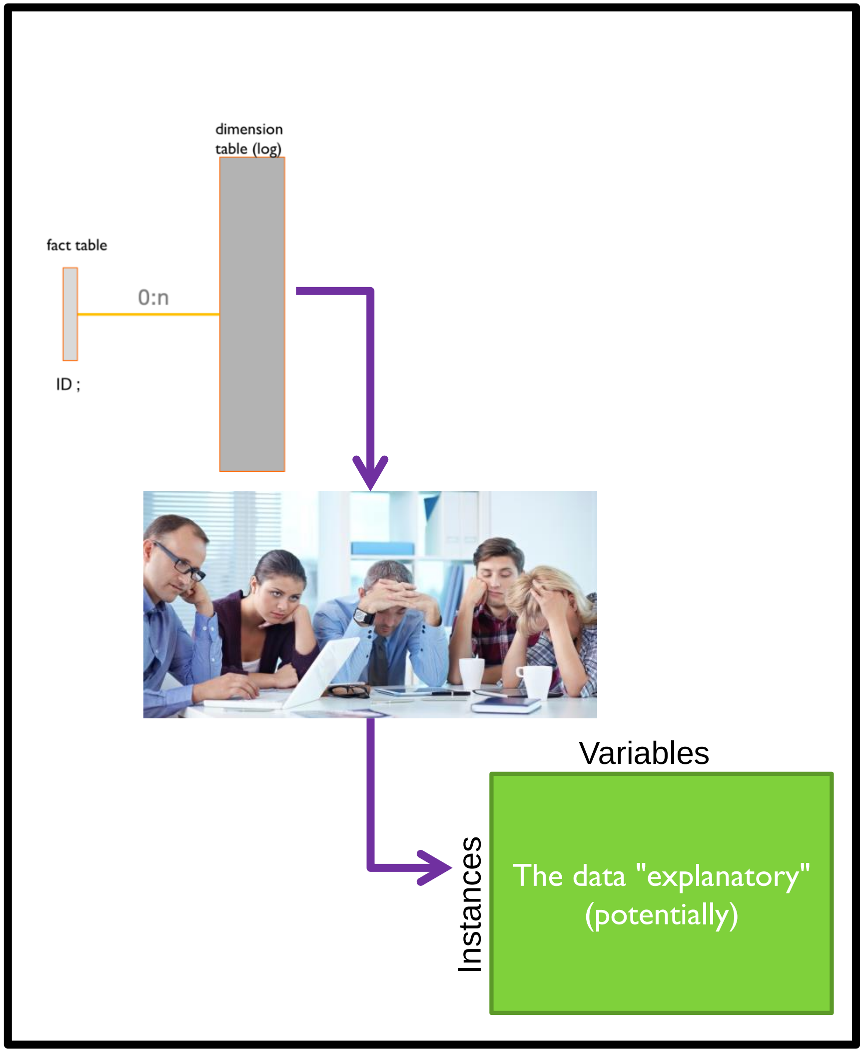
Flattening Such a Relational Structure leads to an Infinite Flat table



Creation of aggregates



Creation of aggregates



- Long
  - Time expensive process to get a flat table usable for data analysis
- Costly
  - Expert knowledge necessary to constructed new variables
- Risky
  - Risk of missing informative variables
  - Risk of constructing and selecting irrelevant variables

- Data-mart specified once for all from business knowledge from a History …
- … and it is hoped valid for a whole range of Future problems

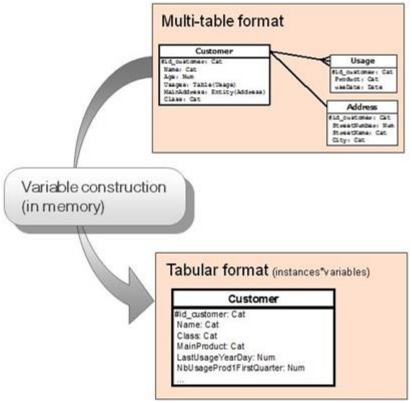
- (a little caricature, the specification of the data mart evolves in the course of the time but always a posteriori)

Schedule

- Automatic data preparation
- Multi-tables data mining
- Automatic variable construction
- Conclusion and future work

Automatic variable construction

- Search for an efficient data representation
  - Context: supervised analysis
    - especially, in the multi-tables settings
  - Data preparation:
    - automatic variable selection
    - next step: automatic variable construction (propositionalisation)



- Objective:
  - Explore numerous data representations using variable construction
  - Select the best representation

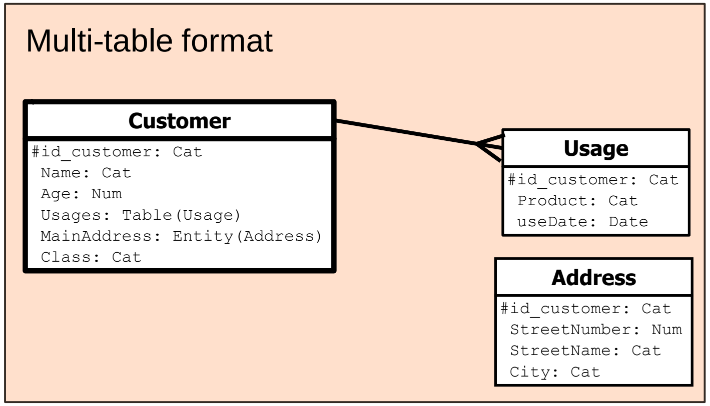
- Challenges
  - The number of constructed variables is infinite
    - it is a subset of all computer programs
  - How to specify domain knowledge in order to control the space of constructed variables?
  - How to efficiently exploit this domain knowledge in order to reach the objective?
    - Explore a very large search space
    - Prevent the risk of over-fitting

Schedule

- Automatic data preparation
- Multi-tables data mining
- Automatic variable construction
  - Specification of domain knowledge
  - Evaluation of constructed variables
  - Sampling a subset of constructed variables
  - Experiments
- Conclusion and future work

Specification of data format

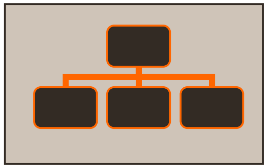
- Table
  - Two kinds of tables
    - Root table: statistical unit of the studied problem
    - Secondary table: sub-part of the statistical unit
  - Variables of simple type
    - Numerical (Num)
    - Categorical (Cat)
  - Variables of advanced type
    - Date, Time, Timestamp…
  - Variables of relation type
    - Simple composition: sub-entity with 0-1 relation (Entity)
    - Multiple composition: sub-entity with 0-n relation (Table)



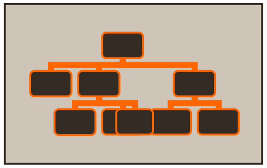
Multi-table schemas

- Mono-table

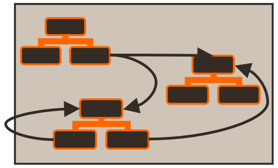
- Multi-tables
  - Star schema



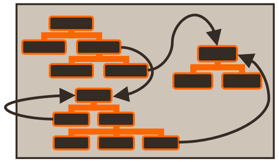
  - Snowflake schema



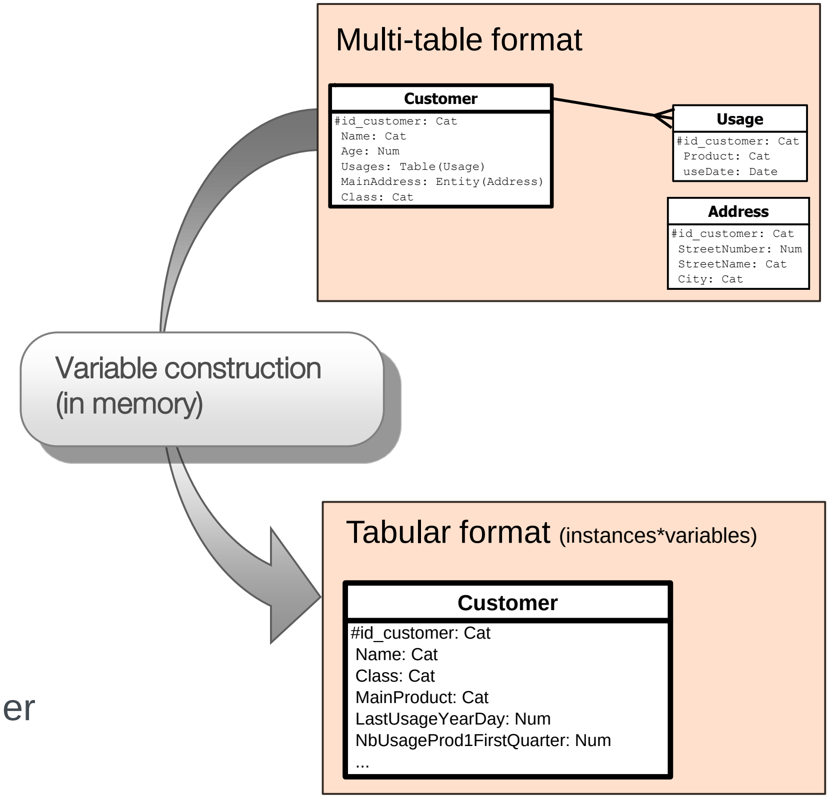
  - External data



  - Multiple snowflake schema



Specification of a variable construction language



- **Construction rule**
  - Program function
    - Input: one or several values
    - Output: one value
  - Type of values
    - Simple: Numerical, Categorical
    - Advanced: Date, Time, Timestamp…
    - Relation: Entity or Table
- **Constructed variable**
  - Output of a construction rule
  - Rule operands
    - Value
    - Variable
    - Output of another rule
- **Examples:**
  - New variables constructed in table Customer
    - *MainProduct = Mode(Usages, Product)*
    - *LastUsageYearDay = Max(Usages, YearDay(useDate))*
    - *NbUsageProd1FirstQuarter = Count(Selection(Usages, YearDay(useDate) in [1 ;90] and Product = “Prod1”))*
    - *…*

Variable construction language
List of construction rules

Name	Return type	Operands	Label
Count	Num	Table	Number of records in a table
CountDistinct	Num	Table, Cat	Number of distinct values
Mode	Cat	Table, Cat	Most frequent value
Mean	Num	Table, Num	Mean value
StdDev	Num	Table, Num	Standard deviation
Median	Num	Table, Num	Median value
Min	Num	Table, Num	Min value
Max	Num	Table, Num	Max value
Sum	Num	Table, Num	Sum of values
Selection	Table	Table, (Cat, Num…)	Selection from a table given a selection criterion
YearDay	Num	Date	Day in year
WeekDay	Num	Date	Day in week
DecimalTime	Num	Time	Decimal hour in day
…	…	…	…

Schedule

- Automatic data preparation
- Multi-tables data mining
- Automatic variable construction
  - Specification of domain knowledge
  - Evaluation of constructed variables
  - Sampling a subset of constructed variables
  - Experiments
- Conclusion and future work

MODL approach: evaluation of one variable

- Definition of modeling space M_C of constructed variables
  - Exploit the domain knowledge
  - Exploit the multi-table format of the input data
  - A constructed variable X is a formula
    - it is a « small » computer program

- Definition of a prior distribution on all constructed variables

$$L\left(M_C\left(X\right)\right) = -\log p\left(M_C\left(X\right)\right)$$

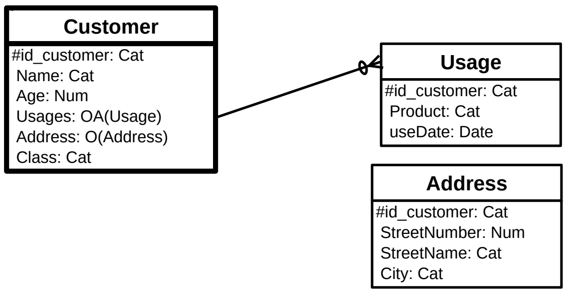
- Evaluation criterion of a constructed variable

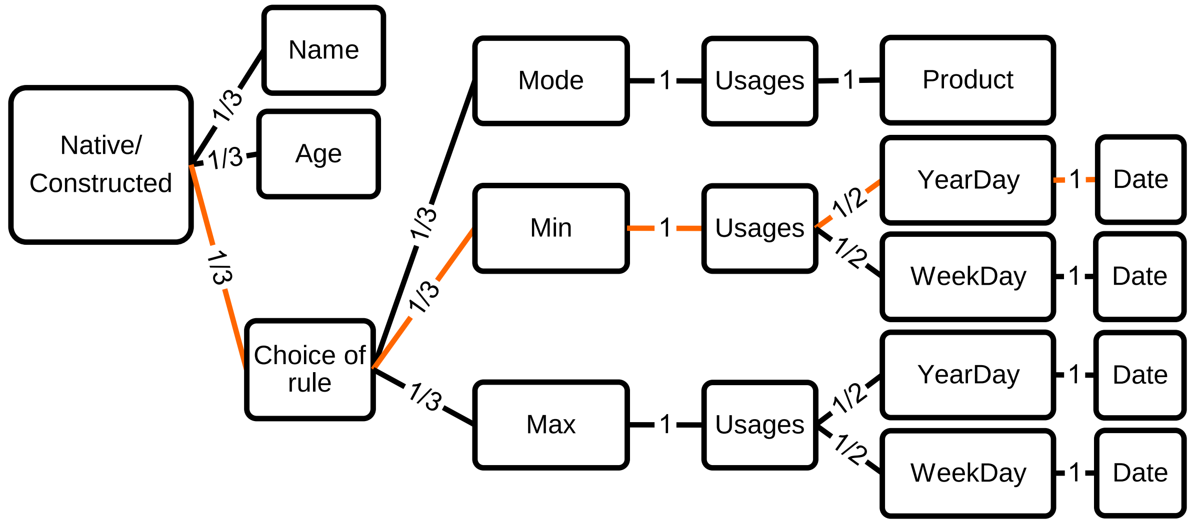
$$c\left(X\right) = \underbrace{\boxed{L\left(M_C\left(X\right)\right)}}_{\text{construction prior}} + \underbrace{L\left(M_P\left(X\right)\right)}_{\text{preprocessing prior}} + \underbrace{L\left(D_Y \mid M_P\left(X\right), D_X\right)}_{\text{likelihood}}$$

Penalization of complex constructed variables

Prior distribution on all constructed variables

Example





- Rules
  - YearDay
  - Weekday
  - Mode
  - Min
  - Max

Hierarchy of Multinomial Distributions with potentially Infinite Depth (HMDID) prior

- Cost of Name $L(M_C(X)) = \log(3)$
  - Choice of variable : log(3)

- Cost of Min(Usages, YearDay(Date)) $L(M_C(X)) = \log(3)+\log(3)+\log(1)+\log(1)+\log(2)+\log(1)$
  - Choice of constructing a variable: log(3)
  - Choice of rule Min: log(3)
  - Choice of first operand (Usages) of Min: log(1)
  - Choice of constructing a variable for second operand of Min: log(1)
  - Choice of rule YearDay: log(2)
  - Choice of operand of YearDay (Date): log(1)

Schedule

- Automatic data preparation
- Multi-tables data mining
- Automatic variable construction
  - Specification of domain knowledge
  - Evaluation of constructed variables
  - Sampling a subset of constructed variables
  - Experiments
- Conclusion and future work

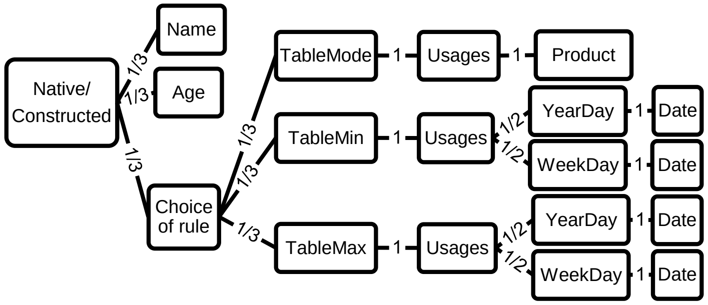
Exploitation of domain knowledge

How to draw a sample from the space of variable construction?

- Objective: draw a sample of *K* variables
  - At this step, the problem of selecting the informative variables is ignored

- Principle
  - Draw the variables one by one according to the HMDID prior

- Naive algorithm: successive random draws
  - Input: *K* {Number of draws}
  - Sortie: $\mathbb{X}=\{X\}$,$|\mathbb{X}|\leq K$ {Sample of constructed variables}
    - 1: $\mathbb{X}=\emptyset$
    - 2: **for** k = 1 to *K* **do**
    - 3: Draw *X* according to HMDID prior
    - 4: Add *X* into $\mathbb{X}$
    - 5: **end for**



Exploitation of domain knowledge

The naive algorithm is neither efficient not computable

- The naive algorithm is not efficient
  - Most draws do not produce new variables
  - Few constructed variables are drawn in case of numerous native variables

- The naive algorithm is not computable
  - Example:
    - Variable v de type Num, rule f(Num, Num) -> Num
    - Example: f = Sum(., .)
    - Family of constructed variables

Size	Example	Coding	Coding length	Prior	Number of variables
1	x	0	1	2^{-1}	1
2	f(x,x)	100	3	2^{-3}	1
3	f(f(x,x), x)	11000	5	2^{-5}	2
4	f(f(x,f(x,x)), x)	1101000	7	2^{-7}	5
5	f(f(x,f(x,x)), f(x,x))	110100100	9	2^{-9}	14
...					
n			2n-1	$2^{-(2n-1)}$	C(n-1)

  - Catalan number C_n
    - C_n is the number of different ways n + 1 factors can be completely parenthesized
    - C_n is also the number of full binary trees with n+1 leaves

  - Expectation of the size of formula: infinite

$$E\left(s\left(X\right)\right)=\sum_{n=1}^{\infty} n2^{-(2n-1)}C_{n-1}=\infty$$

Exploitation of domain knowledge

Draw many constructed variables simultaneously

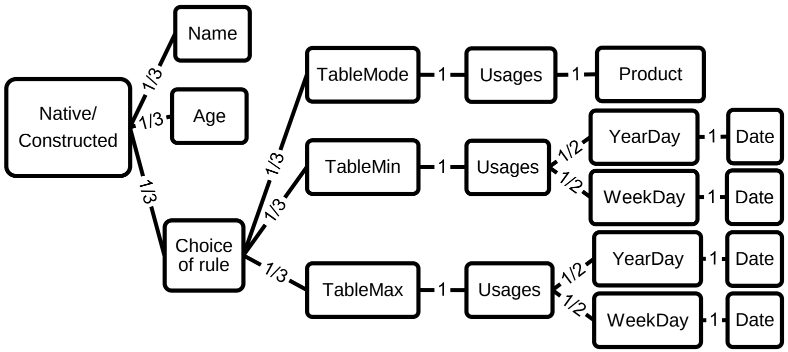
- Principle
  - Draw directly a sample of variables according to prior HMDID
  - Exploit the multinomial maximum likelihood of the whole sample

$$p(D) = \frac{n!}{n_1!n_2!\ldots n_K!} p_1^{n_1} p_2^{n_2} \ldots p_K^{n_K}$$

M L reached with frequencies $n_k = p_k n$

- Whole sample algorithm: simultaneous random draws
  - Input: K {Number of draws}
  - Output: $\mathbb{X}=\{X\}$,$|\mathbb{X}|\leq K$ {Sample of constructed variables}
    - 1: $\mathbb{X}=\emptyset$
    - 2: Start from root node of hierarchy of HMDID prior
    - 3: Compute number of draws K_i per child node of the prior (native variable, rule, operand...)
    - 4: **for all** child node in current node of the prior **do**
    - 5: **if** leaf node of the prior (constructed variable with complete formula) **then**
    - 6: Add X into $\mathbb{X}$
    - 7: **else**
    - 8: Propagate construction recursively by distributing the K_i draws on each child node according to the multinomial distribution
    - 9: **end if**
    - 10: **end for**

- The whole sample algorithm is both efficient and computable

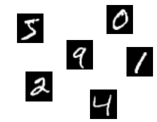


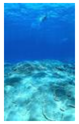
Schedule

- Automatic data preparation
- Multi-tables data mining
- Automatic variable construction
  - Specification of domain knowledge
  - Evaluation of constructed variables
  - Sampling a subset of constructed variables
  - Experiments
- Conclusion and future work

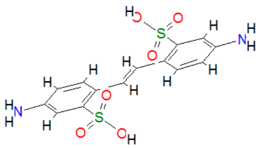
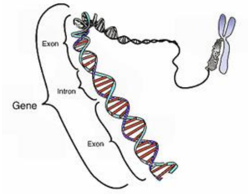
Benchmark

Datasets





- 14 benchmark multi-tables datasets
  - Various domains
    - Handwritten digit
    - Pen tip trajectory character
    - Australian sign language
    - Image
    - Speaker recognition
    - Molecular chemistry
    - Genomics
    - …
  - Various sizes and complexity
    - 100 to 5000 instances
    - 500 to 5000000 records in secondary tables
    - Numerical and categorical variables
    - 2 to 96 classes
    - Unbalanced class distribution

Dataset	Instances	Records	Cat. var	Num. var	Classes	Maj.
Auslan	2565	146949	1	23	96	0.011
CharacterTrajectories	2858	487277	1	4	20	0.065
Diterpenes	1503	30060	2	1	23	0.298
JapaneseVowels	640	9961	1	13	9	0.184
MimlDesert	2000	18000	1	15	2	0.796
MimlMountains	2000	18000	1	15	2	0.771
MimlSea	2000	18000	1	15	2	0.71
MimlSunset	2000	18000	1	15	2	0.768
MimlTrees	2000	18000	1	15	2	0.72
Musk1	92	476	1	166	2	0.511
Musk2	102	6598	1	166	2	0.618
Mutagenesis	188	10136	3	4	2	0.665
OptDigits	5620	5754880	1	3	10	0.102
SpliceJunction	3178	191400	2	1	3	0.521

Benchmark results

Synthesis

- Our method: MODL
- Genericity
  - Useful in a large variety of domains
  - Also applied to classification of time series
- Accuracy
  - Underfit in tiny datasets (Musk)
  - Performance increases with the number of variables
  - Best accuracy overall
- Automation
  - One single parameter: number of features
- Scalability
  - Several orders of magnitude faster that other accurate methods

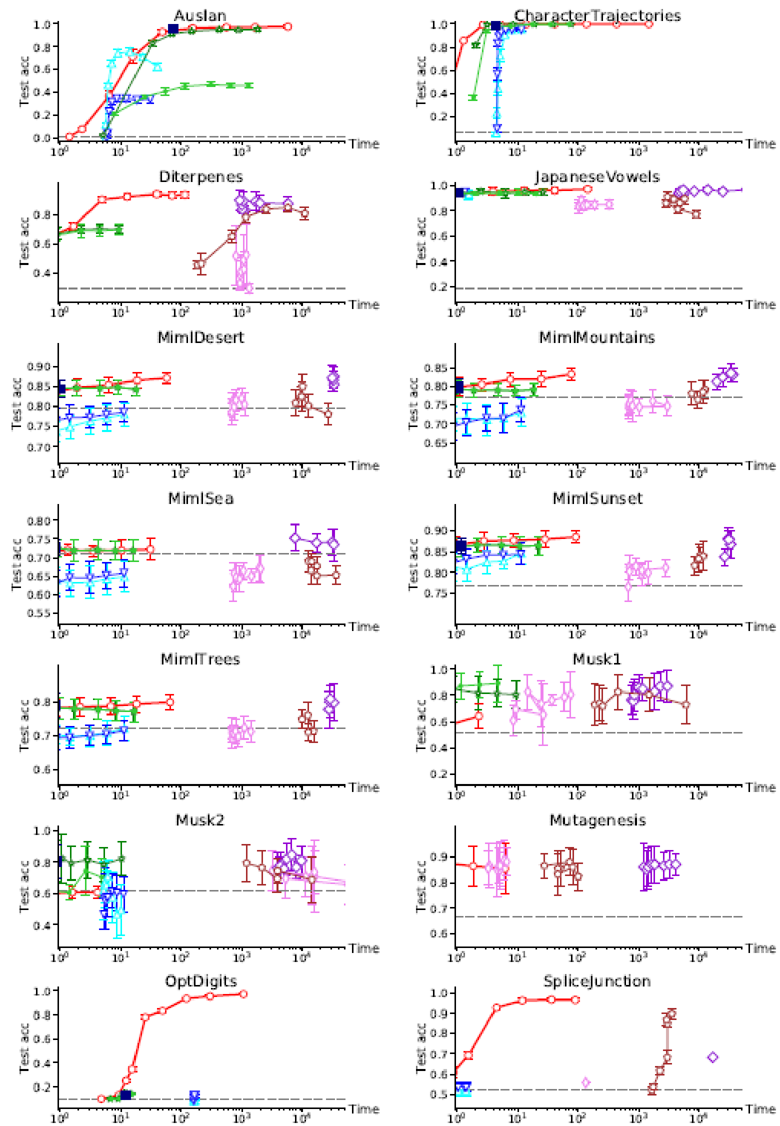


Fig. 11 Test accuracy versus training time per dataset.
MODL: ○ red RELAGGS: ■ navy Cardinal.: ● lime Quantil.: ⋆ green
1BC: △ cyan 1BC2: ▽ blue NFOIL: ⬠ brown Tilde: ◇ pink FORF: ◇ violet

Benchmark: robustness

- Protocol
  - Random shuffle of class values in each dataset
  - Experiments repeated in 10 cross-validation
    - 10000 constructed variables per dataset in each fold
    - 1.4 million of variables evaluated overall
- Results
  - With construction regularization
    - Not one single wrongly selected variable, among the 1.4 million
    - Highly robust approach

Use cases in Orange

- Experiments on large datasets
  - 100 000 customers
    - up to millions in main table
  - 50 millions call detail records
    - up to billions in secondary tables
    - up to hundreds of GB
  - Up to 100 000 automatically constructed variables

- Results
  - Genericiy
  - Parameter-free
    - Rely on domain knowledge description: multi-table specification and choice of construction rules
  - Reliability
  - Accuracy
  - Interpretability:
    - Constructed variables may be numerous, redundant and some of them complex
  - Efficicency

- Use cases and methodology: need to be explored
  - Automatic evaluation of additional data sources
  - Fast automatic solution to many data mining problems
  - Help to suggest new variables to construct
  - …

Schedule

- Automatic data preparation
- Multi-tables data mining
- Automatic variable construction
- Conclusion and future work

Summary

- Variable selection using data grid models
  - Discretization/value grouping
  - Conditional/joint density estimation
- Specification of domain knowledge
  - Multi-table format, advanced data types (Date, Time…)
  - Construction variable language
- Specification of a prior distribution on the space of variable construction
  - Hierarchy of Multinomial Distributions with potentially Infinite Depth
- Sampling algorithm on this infinite variable construction space
  - Concept of maximum likelihood of a whole sample of variables
- Experiments with accurate results, on many relational data mining domains
  - Now widely used on large Orange datasets: effective automation of variable construction

Khiops tool available at www.khiops.com

Future work

- Future work: numerous open problems
  - Design of more parsimonious prior
  - Extension of the specification of domain knowledge
  - Large scale parallelization for exploration of the space of variable construction
  - Sampling constructed variable according to their posterior (*vs*. prior) distribution
  - Any time variable construction, jointly with multivariate classifier training
  - …

thank you for your attention!

References

- Data preparation

  - M. Boullé. A Bayes optimal approach for partitioning the values of categorical attributes. Journal of Machine Learning Research, 6:1431-1452, 2005
  - M. Boullé. MODL: a Bayes optimal discretization method for continuous attributes. Machine Learning, 65(1):131-165, 2006
  - M. Boullé. Data grid models for preparation and modeling in supervised learning. In Hands-On Pattern Recognition: Challenges in Machine Learning, volume 1, I. Guyon, G. Cawley, G. Dror, A. Saffari (eds.), pp. 99-130, Microtome Publishing, 2011

- Modeling

  - M. Boullé. Compression-Based Averaging of Selective Naive Bayes Classifiers. Journal of Machine Learning Research, 8:1659-1685, 2007

- Feature construction

  - M. Boullé. Towards Automatic Feature Construction for Supervised Classification. In ECML/PKDD 2014, Pages 181-196, 2014
  - M. Boullé, C. Charnay, N. Lachiche. A scalable robust and automatic propositionalization approach for Bayesian classification of large mixed numerical and categorical data. Machine Learning, 108(2):229-266, 2019.