# Report on Preliminary Experiments with Data Grid Models in the Agnostic Learning vs. Prior Knowledge Challenge

Marc Boullé

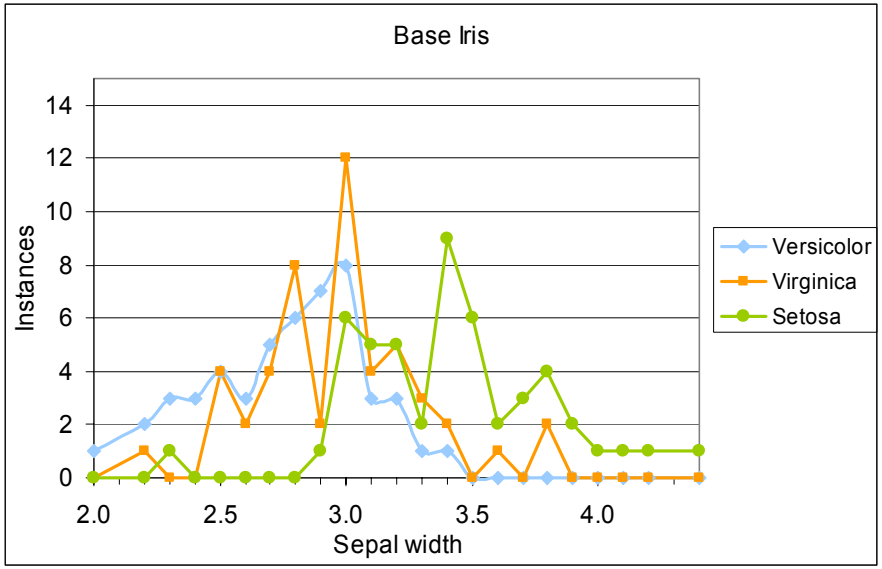
IJCNN 2007 – 08/2007

# Outline

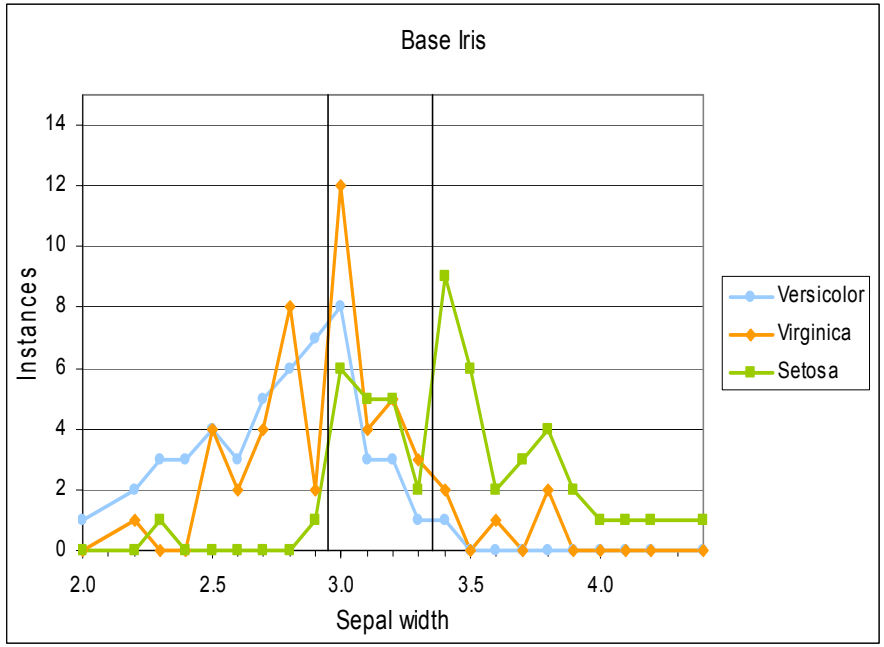
- From univariate to multivariate supervised data grids
- Evaluation on the challenge
- Conclusion

# Numerical variables
## Univariate analysis owing to supervised discretization

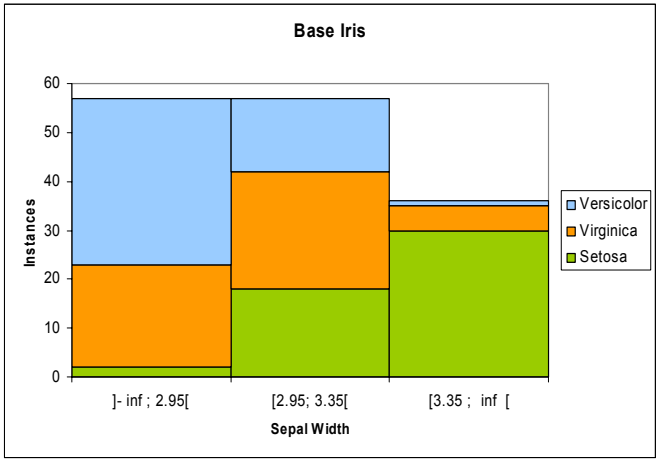
- Context
  - Supervised classification
  - Data preparation
- Discretisation
  - Slice a numerical domain into intervals
- Main issues
  - Informational quality
    - Good fit of the data
  - Statistical quality:
    - Good generalization



# Supervised discretization
## 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 [ |




Which model is the best one?

# MODL approach: principles

- Definition of the model space
  - Definition of the parameters ($I$, $N_{i.}$, $N_{ij}$)
- Definition of a prior distribution on the model parameters
  - Hierarchical prior
  - Uniform at each stage of the hierarchy of the parameters
  - Independence asumption of the distribution between the intervals
- Exact analytical criterion to evaluate the models

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

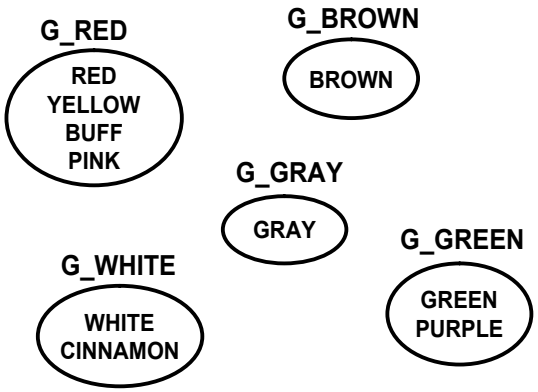
- Efficient optimization: O($N$ log($N$))

# Categorical variables

## Univariate analysis owing to 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 |




Which model is the best one?

# Supervised value grouping

## Same approach as for discretization

- A value grouping model is defined by:
  - The number of groups of values,
  - The partition of the input values into groups,
  - The distribution of the target values for each group.
- Model selection
  - Bayesian approach for model selection
  - Hierarchical prior for the model parameters
  - Exact analytical criterion to evaluate the models

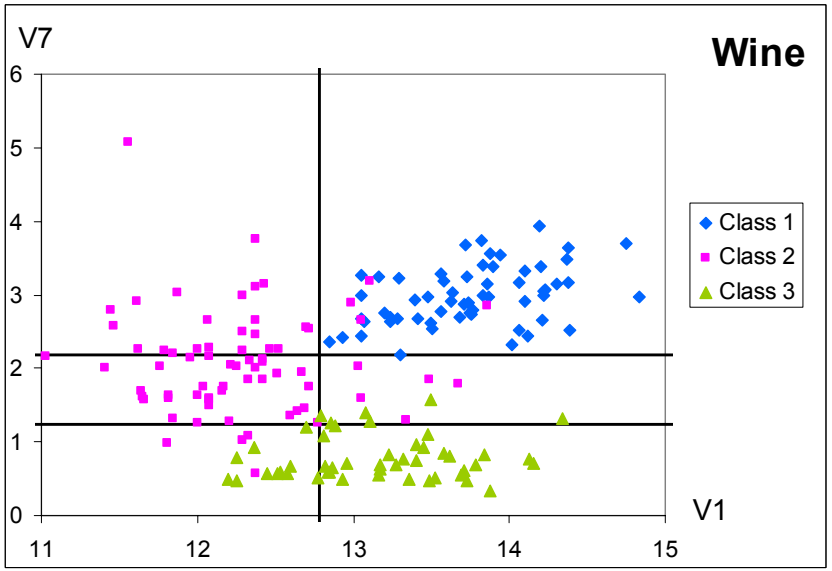
$$\underbrace{\log(V)+\log\left(B(V,I)\right)+\sum_{i=1}^{I}\log\left(C_{N_{i.}+J-1}^{J-1}\right)}_{\text{prior}}+\underbrace{\sum_{i=1}^{I}\log\left(N_{i.}!/N_{i1}!N_{i2}!\ldots N_{iJ}!\right)}_{\text{lilelihood}}$$

- Efficient optimization: O($N$ log($N$))

# Pair of numerical variables

## Bivariate discretization to evaluate the conditional density

- Each variable is discretized
- We get a bivariate data grid
- In each cell of the data grid, estimation of the conditional density




| | | |
|---|---|---|
| ]2.18;+inf[ | (0, 23, 0) | (59, 0, 4) |
| ]1.235;2.18] | (0, 35, 0) | (0, 5, 6) |
| ]-inf;1.235] | (0, 4, 11) | (0, 0, 31) |
| V7xV1 | ]-inf;12.78] | ]12.78;+inf[ |

# Application of the MODL approach

- Definition of the model space
  - Definition of the parameters $I_1, I_2, N_{i_1..}, N_{.i_2.}, N_{i_1 i_2 j}$
- Definition of a prior distribution on the model parameters
  - Hierarchical prior
  - Uniform at each stage of the hierarchy of the parameters
  - Independence asumption of the class distribution between the cells
- Exact analytical criterion to evaluate the models

prior

likelihood

$$\log(N) + \log\left(C_{N+I_1-1}^{I_1-1}\right) + \log(N) + \log\left(C_{N+I_2-1}^{I_2-1}\right) + \sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} \log\left(C_{N_{i_1 i_2.}+J-1}^{J-1}\right) +$$

$$\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2} \log\left(N_{i_1 i_2.}!/N_{i_1 i_2 1}!N_{i_1 i_2 2}!\ldots N_{i_1 i_2 J}!\right)$$

# Optimization algorithms

- Main algorithm: greedy top-down merge heuristic
- Post-optimization by alterning univariate optimizations
  - Freeze the partition of $X_1$ and optimize the partition of variable $X_2$
  - Freeze the partition of $X_2$ and optimize the partition of variable $X_1$
- Global search:
  - Variable Neighborhood Search (VNS) meta-heuristic
- Algorithmic complexity: $O\left(N \log\left(N\right)\right)$
  - Exploiting sparseness of data grids
    - At most $N$ non-empty cells among $N^2$ potential cells
  - Exploiting the additivity the evaluation criterion

# Illustration: noise or information?

Diagram 1

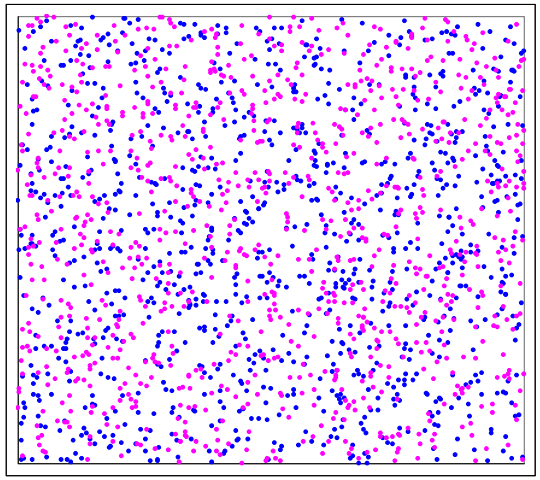


Diagram 2

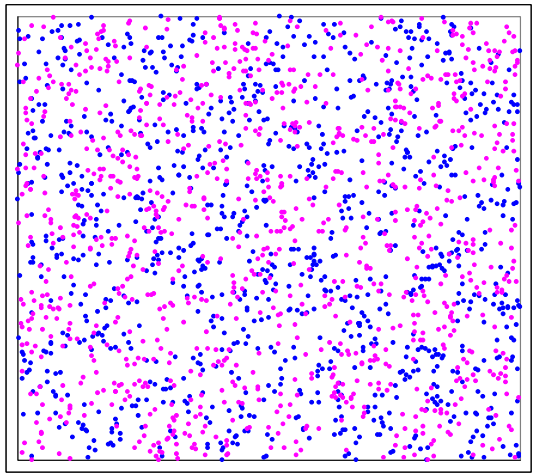
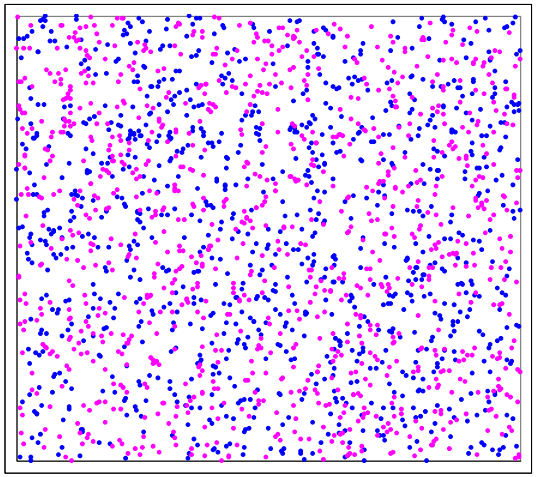
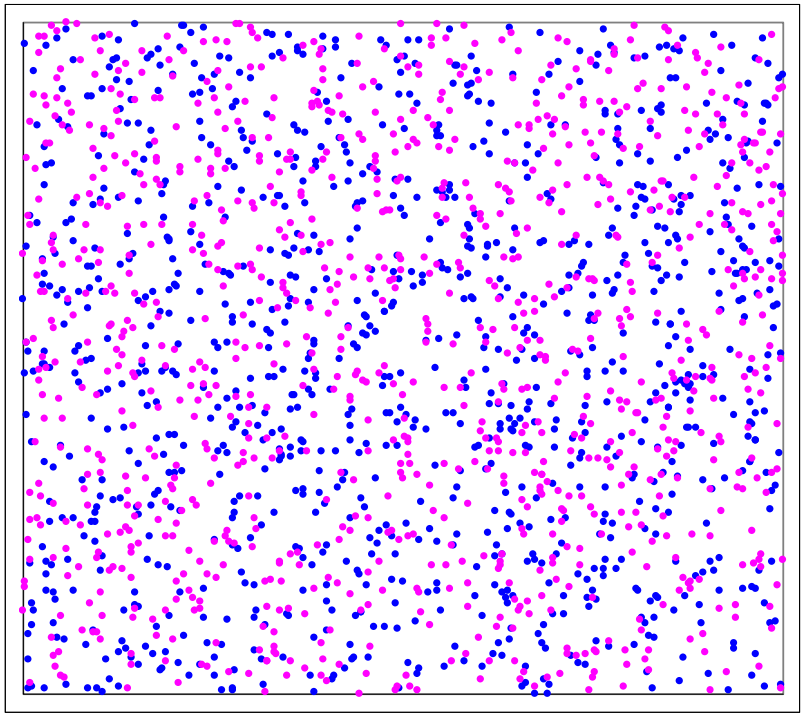


Diagram 3

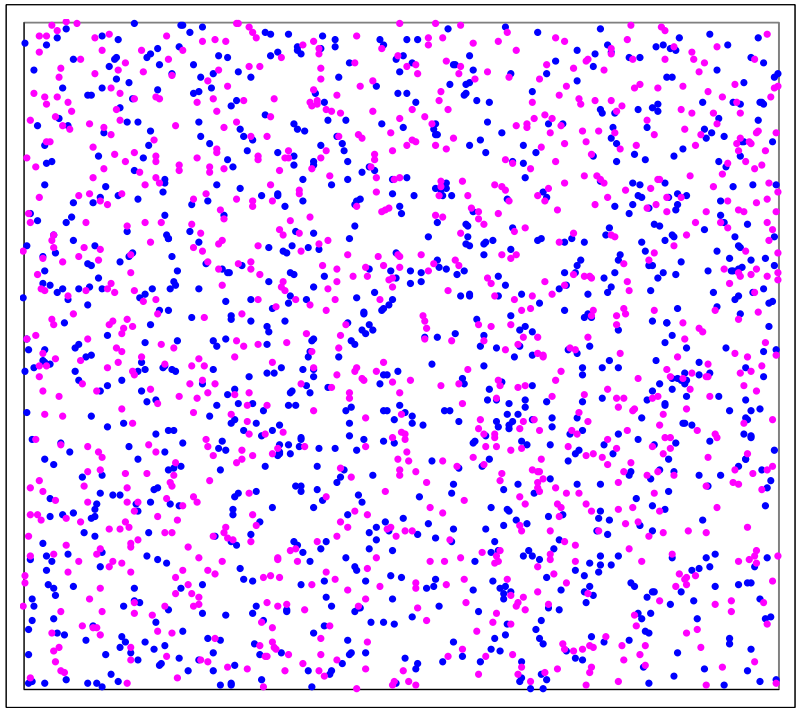


# Diagram 1
## Noise

Diagram 1 (2000 points)



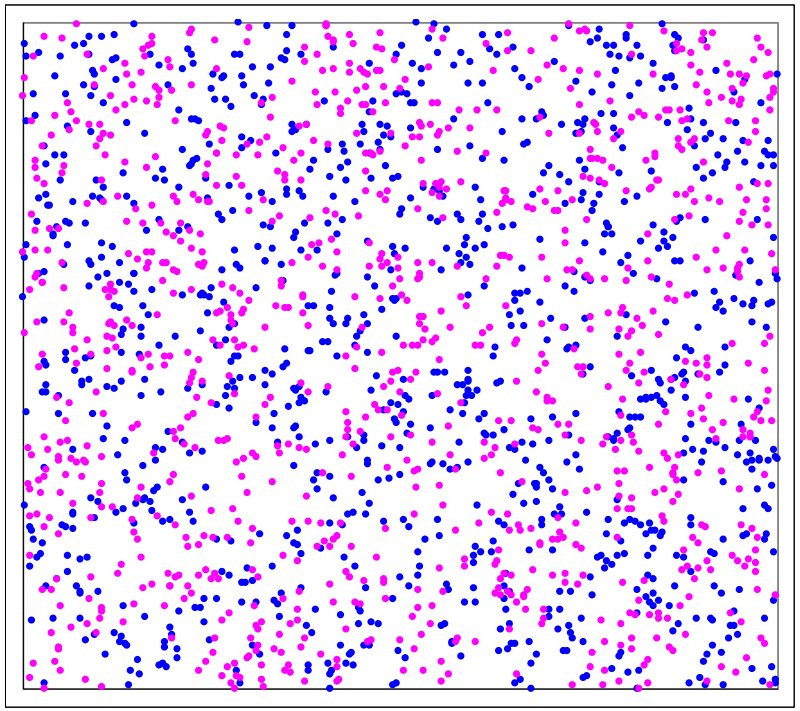
Data grid 1 x 1 (1 cell)



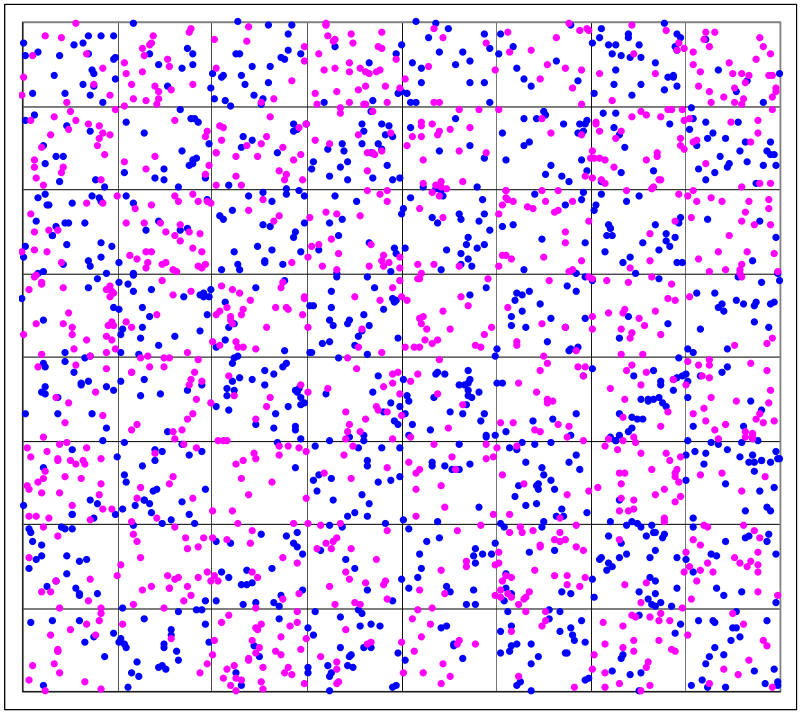
Criterion value = 2075

# Diagram 2
## Chessboard 8 x 8 with 25% noise

Diagram 2 (2000 points)



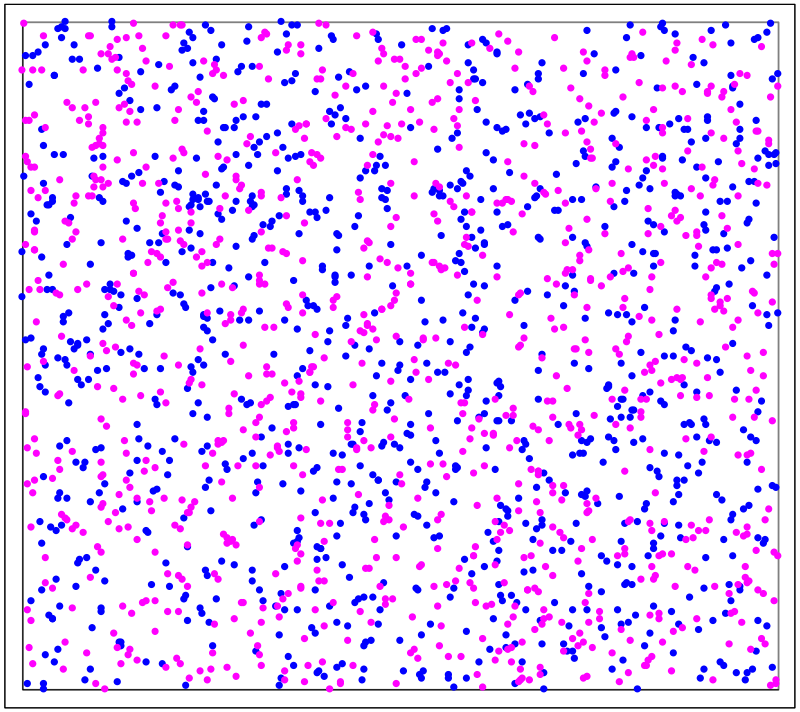
Data grid 8 x 8 (64 cells)



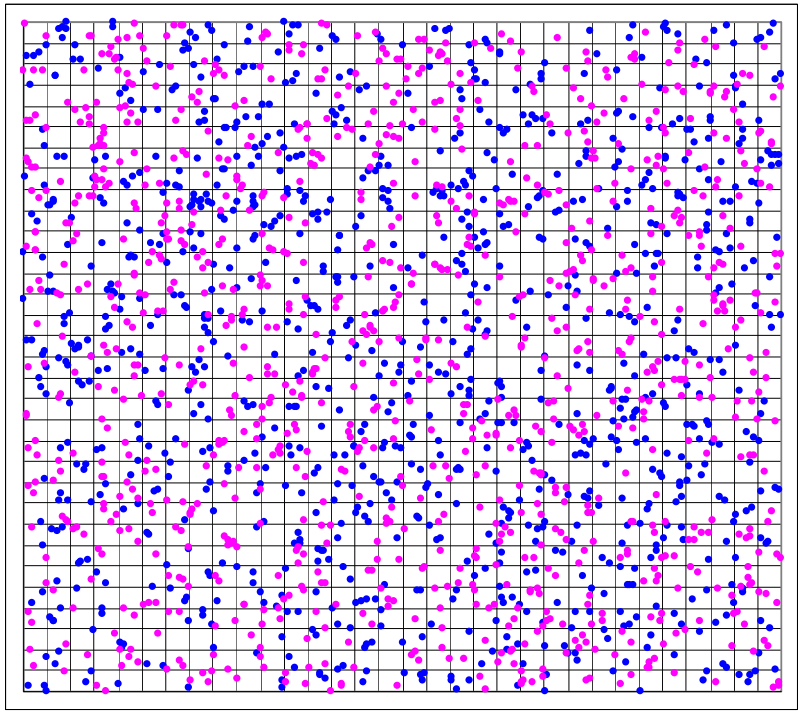
Criterion value = 1900

# Diagram 3
## Chessboard 32 x 32, no noise

Diagram 3 (2000 points)



Data grid 32 x 32 (1024 cells)



Criterion value = 1928

# Multivariate data grid models for conditional density estimation

- Data grid model
  - Variable selection ($k \in \mathbf{K}_S$)
  - Discretization of numerical variables ($k \in \mathbf{K}_1$)
  - Value grouping of categorical variables ($k \in \mathbf{K}_2$)
  - For each cell of the resulting data grid, definition of the distribution of the output values
- Model selection
  - Bayesian approach
  - Hierachical prior on the parameters
  - Exact analytical criterion
- Advanced optimization heuristics
  - $K$: number of variables
  - $N$: number of instances

$$O\left(KN\sqrt{N}\log(N)\max(K,\log(N))\right)$$

prior

$$\log(K+1)+\log\left(C_{K+K_S-1}^{K_S-1}\right)+$$

$$\sum_{k\in \mathrm{K}_S\cap \mathrm{K}_1}\left(\log(N)+\log\left(C_{N+I_k-1}^{I_k-1}\right)\right)+$$

$$\sum_{k\in \mathrm{K}_S\cap \mathrm{K}_2}\left(\log(V_k)+\log\left(B(V_k,I_k)\right)\right)+$$

$$\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_K=1}^{I_K}\log\left(C_{N_{i_1 i_2\ldots i_K}+J-1}^{J-1}\right)+$$

likelihood

$$\sum_{i_1=1}^{I_1}\sum_{i_2=1}^{I_2}\cdots\sum_{i_K=1}^{I_K}\left(\log\left(N_{i_1 i_2\ldots i_K}!\right)-\sum_{j=1}^{J}\log\left(N_{i_1 i_2\ldots i_K j}!\right)\right)$$

# Outline

- From univariate to multivariate supervised data grids
- Evaluation on the challenge
- Conclusion

# Building classifiers from data grids

- Data grid classifier (DG)
  - Train the MAP data grid on all the available data (no validation set)
  - For a test instance:
    - retrieve the related train cell
    - predict the empirical target distribution of that cell
- Data grid ensemble classifier (DGE)
  - Collect the train data grids encountered during optimization
  - Compression-based averaging
    - Weight according to the posterior probability (with a logarithmic smoothing)
- Coclustering of instances and variables (DGCC)
  - See paper

# Challenge results

| Dataset | My best entry | Entry ID | Test BER | Test AUC | Score | Track | Best test BER (agnostic) | Best test BER (prior) |
|---|---|---|---|---|---|---|---|---|
| **ADA** | Data Grid (CMA) | 920 | 0.1756 | 0.8464 | 0.0245 | Prior | 0.166 | 0.1756 |
| **GINA** | Data Grid (Coclustering) | 921 | 0.0516 | 0.9768 | 0.3718 | Prior | 0.0339 | 0.0226 |
| **HIVA** | Data Grid (Coclustering) | 921 | 0.3127 | 0.7077 | 0.5904 | Agnos | 0.2827 | 0.2693 |
| **NOVA** | Data Grid (Coclustering) | 921 | 0.0488 | 0.9813 | 0.141 | Agnos | 0.0456 | 0.0659 |
| **SYLVA** | Data Grid (CMA) | 918 | 0.0158 | 0.9873 | 0.6482 | Agnos | 0.0062 | 0.0043 |
| **Overall** | Data Grid (Coclustering) | 921 | 0.1223 | 0.8984 | 0.3813 | Prior | 0.1117 | 0.1095 |

# Analysis of the results

- Predictive performance
  - Data grids classifier perform well on dense datasets
  - Building ensemble of classifiers  is better
  - Coclustering looks appealing for sparse datasets

- Interpretation
  - Few variables are selected (for example in the prior track)
    - Ada: 6 variables (Test BER = 0.2058)
    - Gina: 7 variables (Test BER = 0.1721)
    - Sylva: 4 variables (Test BER=0.0099)

# MAP data grid for Ada dataset (prior)

## Six selected variables

- occupation
  - {Prof-specialty, ...}
    - Prof-specialty
    - Exec-managerial
    - Sales
    - Adm-clerical
    - Tech-support
    - Protective-serv
    - Armed-Forces
  - {Craft-repair, ...}
    - Craft-repair
    - Other-service
    - Machine-op-inspct
    - Transport-moving
    - Handlers-cleaners
    - Farming-fishing
    - Priv-house-serv
- relationship
  - {Not-in-family, ...}
    - Not-in-family
    - Own-child
    - Unmarried
    - Other-relative
  - {Husband, …}
    - Husband
    - Wife
- educationNum
  - ]-inf;12.5]
  - ]12.5;+inf[
- age
  - ]-inf;27.5]
  - ]27.5;+inf[
- capitalGain
  - ]-inf;4668.5]
  - ]4668.5;5095.5]
  - ]5095.5;+inf[
- capitalLoss
  - ]-inf;1805]
  - ]1805;2001.5]
  - ]2001.5;+inf[

# Adult data grid cells

## 12 most frequent cells (90% of the instances)

| relationship | occupation | educationNum | age | capitalGain | capitalLoss | Class -1 | Class 1 | Frequency |
|---|---|---|---|---|---|---|---|---|
| {Husband, ...} | {Craft-repair, ...} | ]-inf;12.5] | ]27.5;+inf[ | ]-inf;4668.5] | ]-inf;1805] | 78% | 22% | 736 |
| {Not-in-family, ...} | {Craft-repair, ...} | ]-inf;12.5] | ]27.5;+inf[ | ]-inf;4668.5] | ]-inf;1805] | 97% | 3% | 577 |
| {Not-in-family, ...} | {Prof-specialty, ...} | ]-inf;12.5] | ]27.5;+inf[ | ]-inf;4668.5] | ]-inf;1805] | 94% | 6% | 531 |
| {Husband, ...} | {Prof-specialty, ...} | ]-inf;12.5] | ]27.5;+inf[ | ]-inf;4668.5] | ]-inf;1805] | 59% | 41% | 489 |
| {Husband, ...} | {Prof-specialty, ...} | ]12.5;+inf[ | ]27.5;+inf[ | ]-inf;4668.5] | ]-inf;1805] | 31% | 69% | 445 |
| {Not-in-family, ...} | {Craft-repair, ...} | ]-inf;12.5] | ]-inf;27.5] | ]-inf;4668.5] | ]-inf;1805] | 100% | 0% | 425 |
| {Not-in-family, ...} | {Prof-specialty, ...} | ]-inf;12.5] | ]-inf;27.5] | ]-inf;4668.5] | ]-inf;1805] | 99% | 1% | 316 |
| {Not-in-family, ...} | {Prof-specialty, ...} | ]12.5;+inf[ | ]27.5;+inf[ | ]-inf;4668.5] | ]-inf;1805] | 79% | 21% | 268 |
| {Not-in-family, ...} | {Prof-specialty, ...} | ]12.5;+inf[ | ]-inf;27.5] | ]-inf;4668.5] | ]-inf;1805] | 99% | 1% | 112 |
| {Husband, ...} | {Craft-repair, ...} | ]-inf;12.5] | ]-inf;27.5] | ]-inf;4668.5] | ]-inf;1805] | 95% | 5% | 96 |
| {Husband, ...} | {Prof-specialty, ...} | ]12.5;+inf[ | ]27.5;+inf[ | ]5095.5;+inf[ | ]-inf;1805] | 0% | 100% | 93 |
| {Husband, ...} | {Craft-repair, ...} | ]12.5;+inf[ | ]27.5;+inf[ | ]-inf;4668.5] | ]-inf;1805] | 76% | 24% | 50 |

# Outline

- From univariate to multivariate supervised data grids
- Evaluation on the challenge
- Conclusion

# Data grids for supervised learning

- Evaluation of class conditional density $P(Y \mid X_1, X_2, \ldots, X_K)$
  - Each input variable is partitioned into intervals or groups of values
  - The cross-product of the univariate partitions forms a data grid, used to evaluate the correlation with the class variable
- MODL approach
  - Bayesian model selection approach
  - Analytical evaluation criterion
  - Efficient combinatorial optimization algorithms
  - **High quality data grids**
- Experimental evaluation
  - Good prediction in dense datasets
  - Variable selection (limited to $\log_2 N$ variables)
  - Reliable interpretation
  - **Efficient technique for data preparation**

# Future work

- Explore the potential and limits of data grids
  - For all tasks of machine learning: supervised and unsupervised
  - For data preparation, prediction, explanation

- Combine data grids with the naive Bayes approach
  - Use data grids to build informative multivariate features
  - Use naive Bayes classification to combine features
  - Use ensemble of classifiers to further improve the results