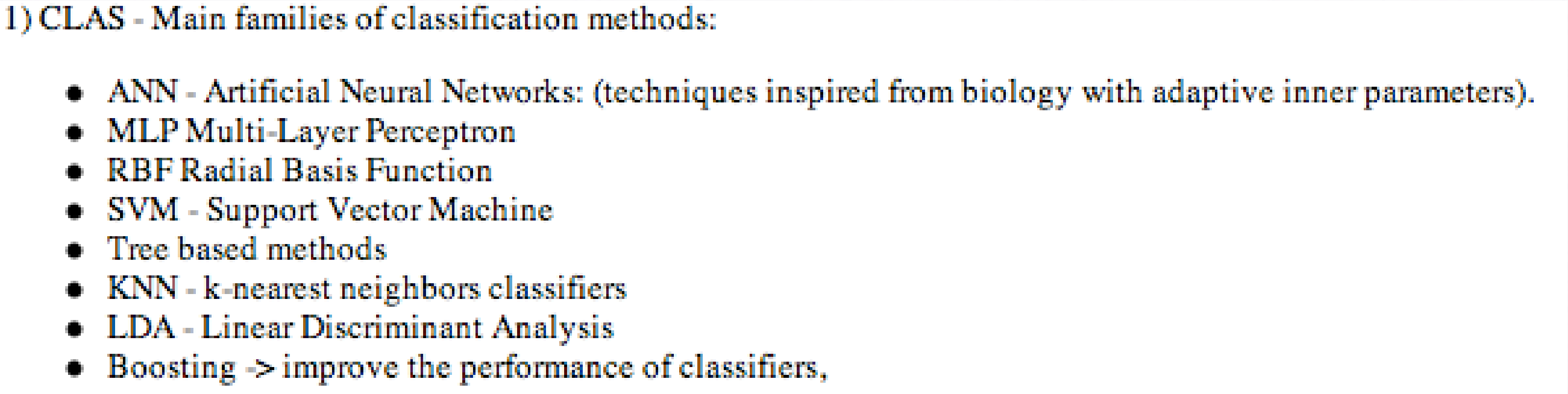
Classification

Laurent Eyer (Observatoire de Genève)

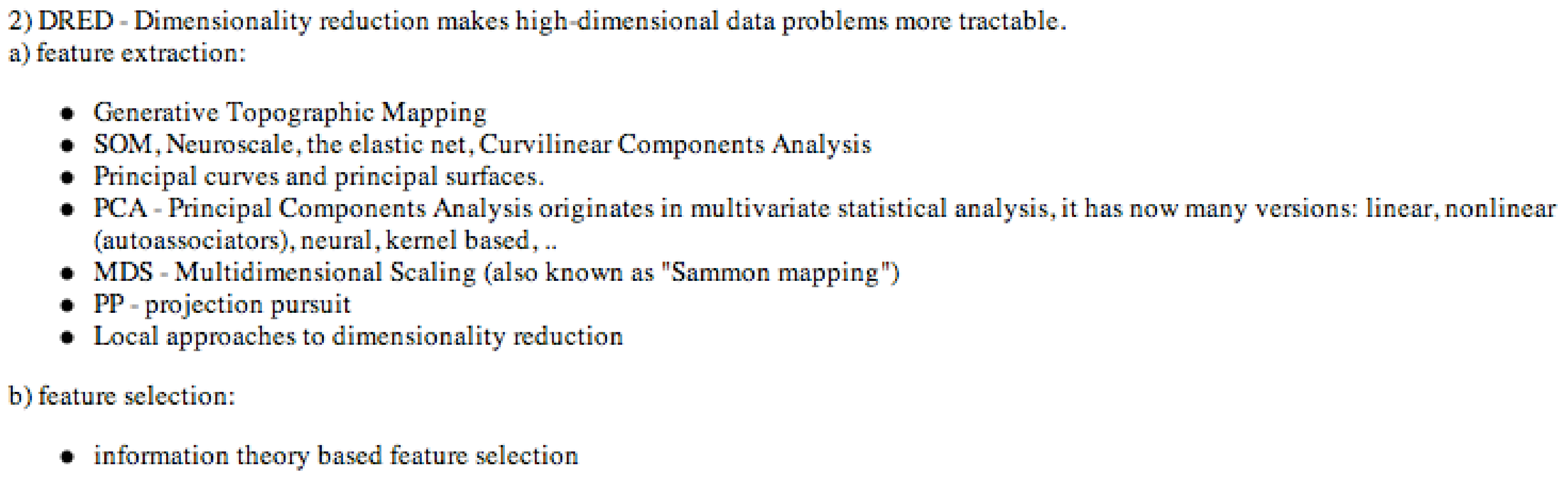
Cambridge, April 16 2004

Some data mining methods (Antoinue Naud)



1) CLAS - Main families of classification methods:

- ANN - Artificial Neural Networks: (techniques inspired from biology with adaptive inner parameters).
- MLP Multi-Layer Perceptron
- RBF Radial Basis Function
- SVM - Support Vector Machine
- Tree based methods
- KNN - k-nearest neighbors classifiers
- LDA - Linear Discriminant Analysis
- Boosting -> improve the performance of classifiers,

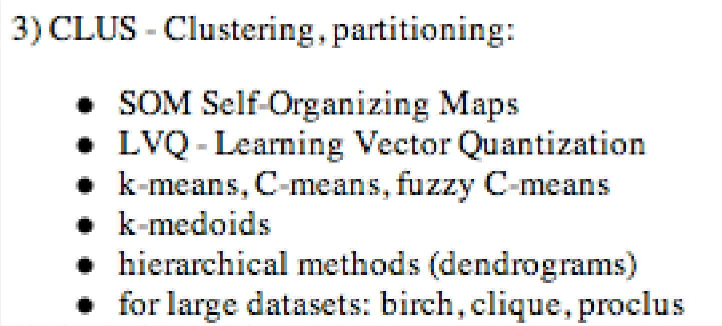


2) DRED - Dimensionality reduction makes high-dimensional data problems more tractable.
a) feature extraction:

- Generative Topographic Mapping
- SOM, Neuroscale, the elastic net, Curvilinear Components Analysis
- Principal curves and principal surfaces.
- PCA - Principal Components Analysis originates in multivariate statistical analysis, it has now many versions: linear, nonlinear (autoassociators), neural, kernel based, ..
- MDS - Multidimensional Scaling (also known as "Sammon mapping")
- PP - projection pursuit
- Local approaches to dimensionality reduction

b) feature selection:

- information theory based feature selection



3) CLUS - Clustering, partitioning:

- SOM Self-Organizing Maps
- LVQ - Learning Vector Quantization
- k-means, C-means, fuzzy C-means
- k-medoids
- hierarchical methods (dendrograms)
- for large datasets: birch, clique, proclus

Works that have been done

- Self Organising Maps (Belokurov, Naud (Eyer): Hipparcos, Belokurov: AGAPE)
- Bayesian classifier (Eyer: ASAS, Hipparcos)
- Discriminant analysis (Waelkens, Aerts, Kestens (Eyer) 2 studies on Hipparcos)
- Wavelets and k-means on Hipparcos data (H. Gu, D. Campbell (Eyer))
- Neural network (Belokurov, Evans & Le Du: MACHO)
- Classical methods (Pojmanski: ASAS)

Bayesian Classifier

work already presented

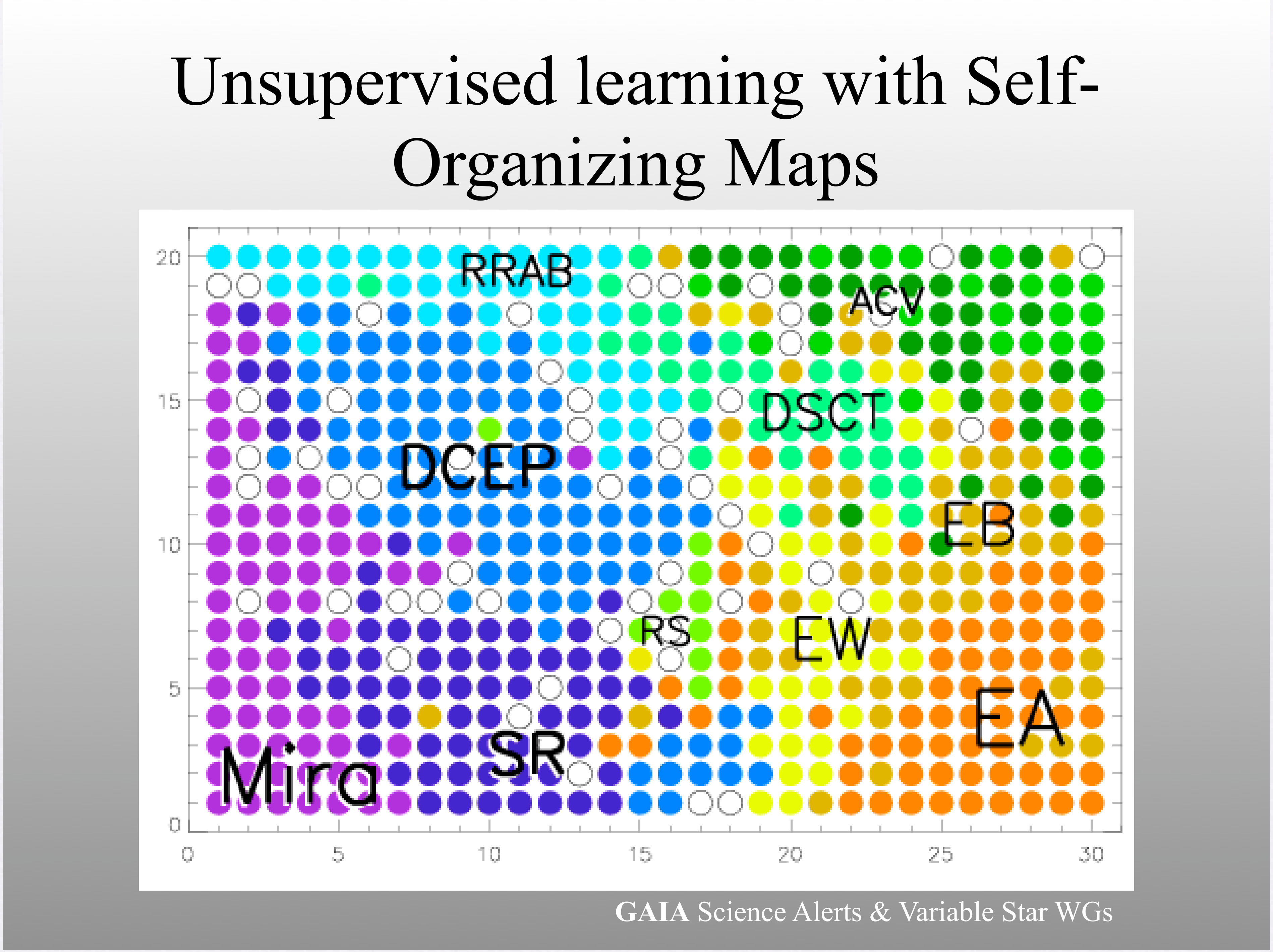
- Hipparcos
- ASAS: All Sky Automated Survey
- Data model (no cyclic variables)
- Given the data and a data model, search for a classification (number of class) which is most probable

Vasily’s Work on SOMs

- 52 input parameters:
  - Lomb periodogram in 40 bins
  - 5, 15, ..., 85, 95 percentiles
  - Ratio magnitudes above/below median
  - V-I

→ 2D representation

Vasily's work on Hipparcos

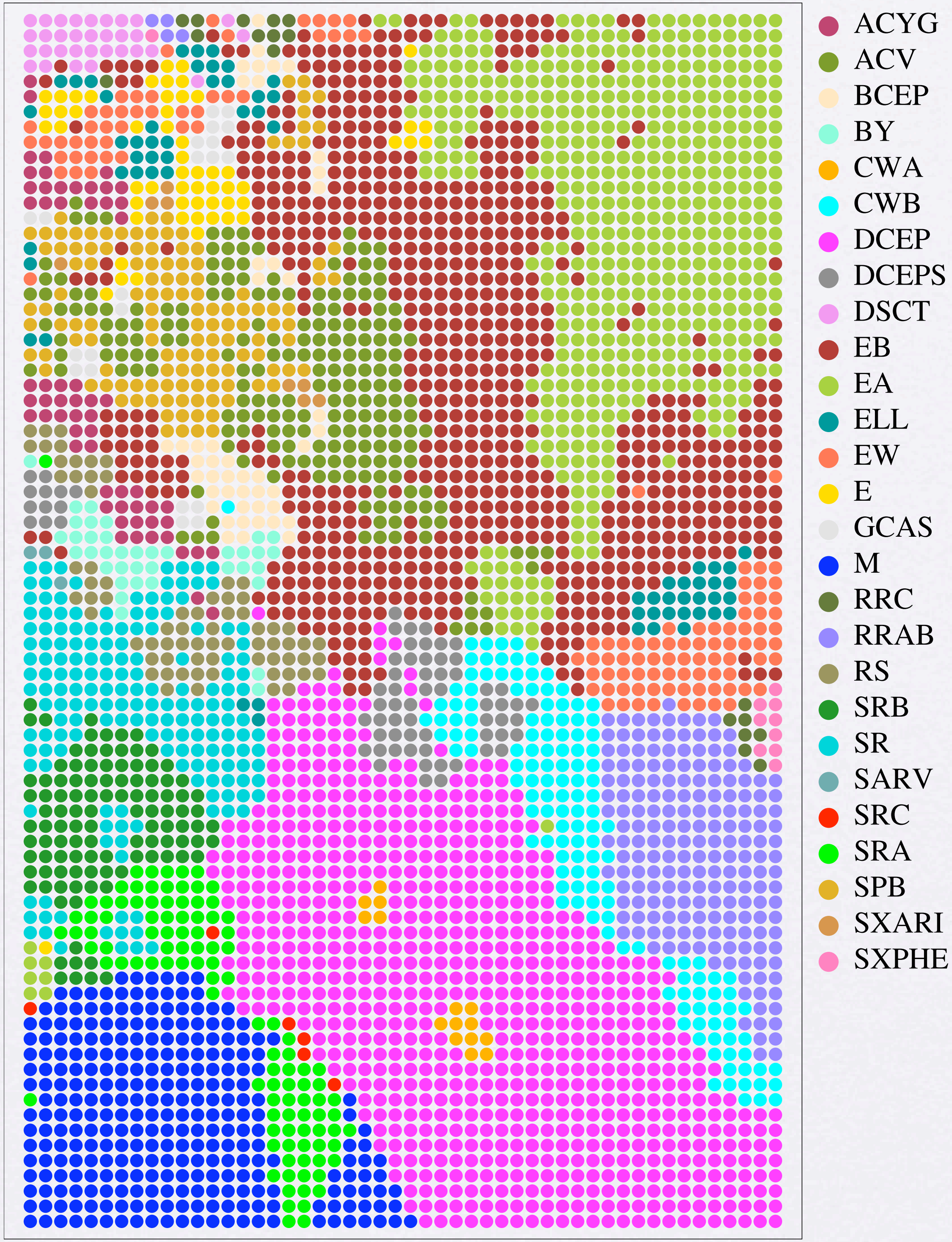


Unsupervised learning with Self-Organizing Maps

GAIA Science Alerts & Variable Star WGs

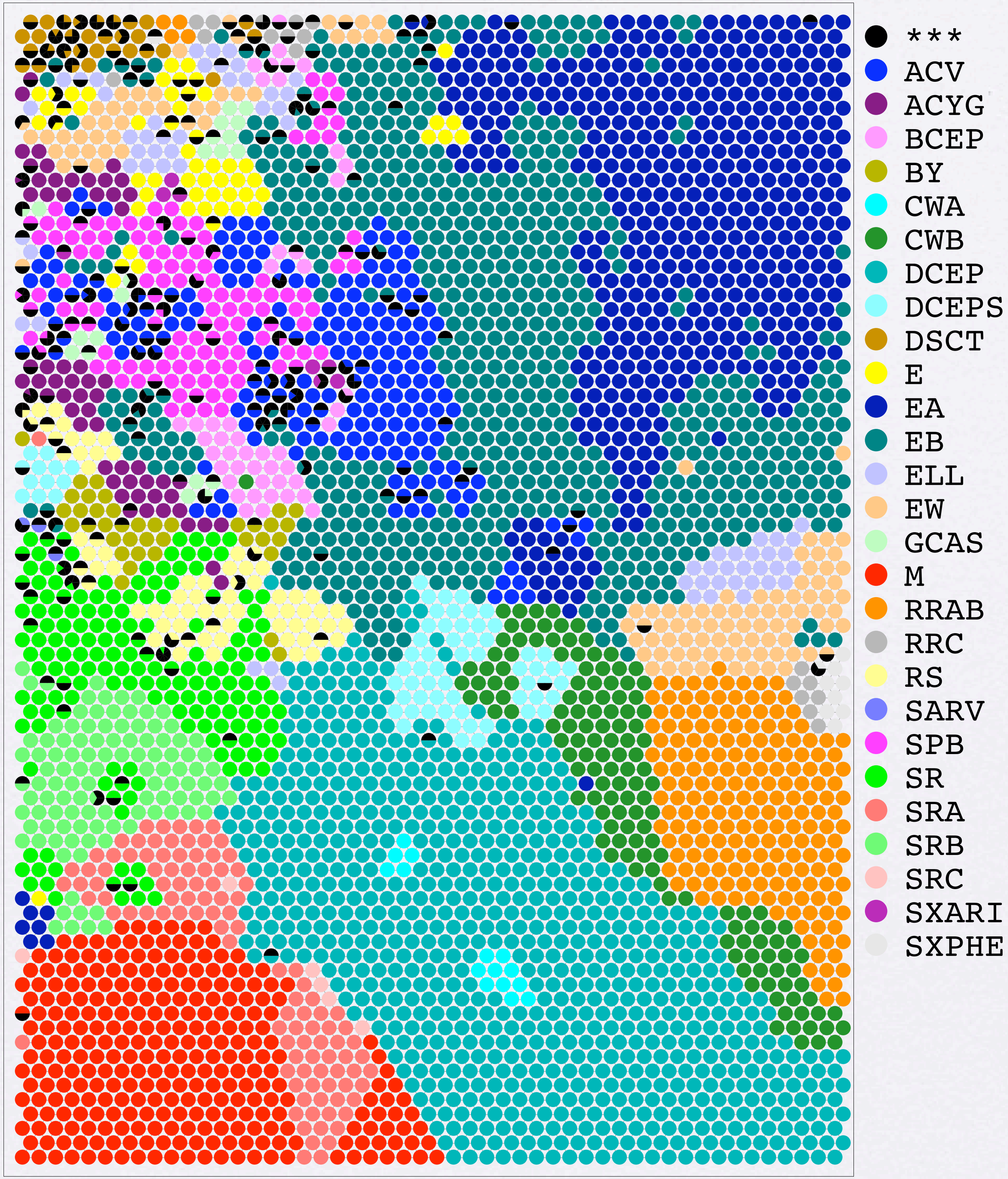
Anoine Naud's work with Hipparcos

- Features:
    - period, amplitude
    - V-I, (Mv)
    - skewness


ACYG
ACV
BCEP
BY
CWA
CWB
DCEP
DCEPS
DSCT
EB
EA
ELL
EW
E
GCAS
M
RRC
RRAB
RS
SRB
SR
SARV
SRC
SRA
SPB
SXARI
SXPHE

Classification of the Hipparcos unclassified objects

The *** objects are from the periodic catalogue, i.e fairly well behaved



Future work

- Antoine Naud, Darek Graczyk (Torun):
    - Hipparcos, OGLE

Discriminant Analysis

- With Hipparcos: 267 new B variable stars
    - Initially: classification by hand looking at spectral type and light curve, then done with discriminant analysis
- Same method applied to select gamma Doradus stars

Principle of Discriminant analysis

- Definition of classes (with already known objects)
- determine the centre of the cluster classes
- For a new object, compute Mahalanobis distance to the class centres
- determine most likely membership

Discriminant Analysis: One application

- SPBs
  - freq, 3 colours of Geneva photometry(173) 3 calibrating classes: beta cephei, SPB, CP
    - 4 new beta Cep
    - 72 new SPB
    - 34 new CP stars
    - 32 new alpha Cyg
    - 7 new eclipsing binaries
    - 17 unclassed

Conclusion: TO WORK

- Add summary about classification on VSWG webpage
- To build knowledge:
    - First: experiences with real database!
    - simulations
- To compare methods (blind testing procedure?)