Classificação de Padrões

Carlos Eduardo Pedreira

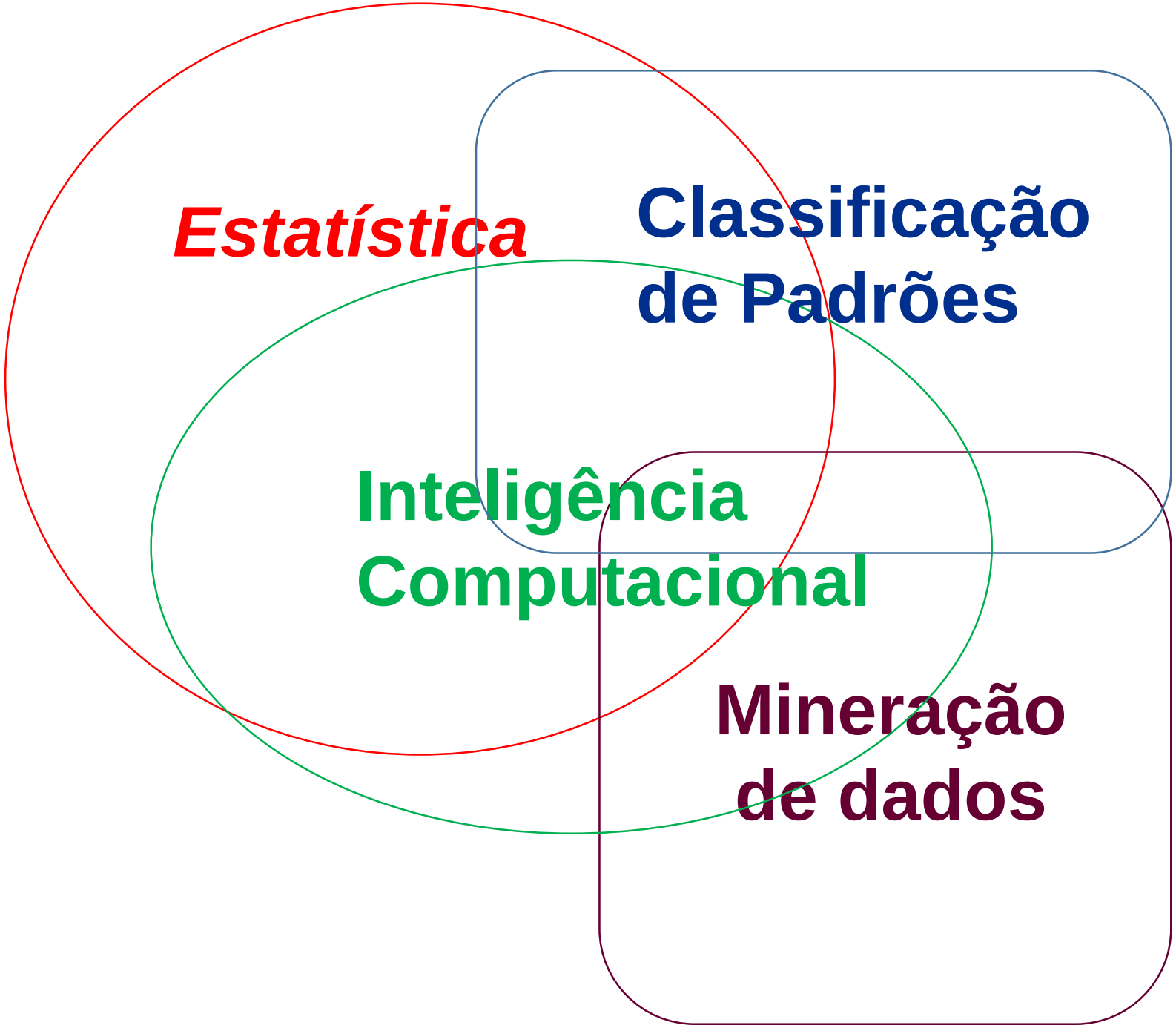
PESC - COPPE

2018

É importante classificar padrões?

Onde se aplica classificação de padrões?

- Diagnostico médico
- Risco de crédito
- Classificação facial
- Fraudes dem cartões de crédito
- . . .


Estatística
Classificação
de Padrões
Inteligência
Computacional
Mineração
de dados

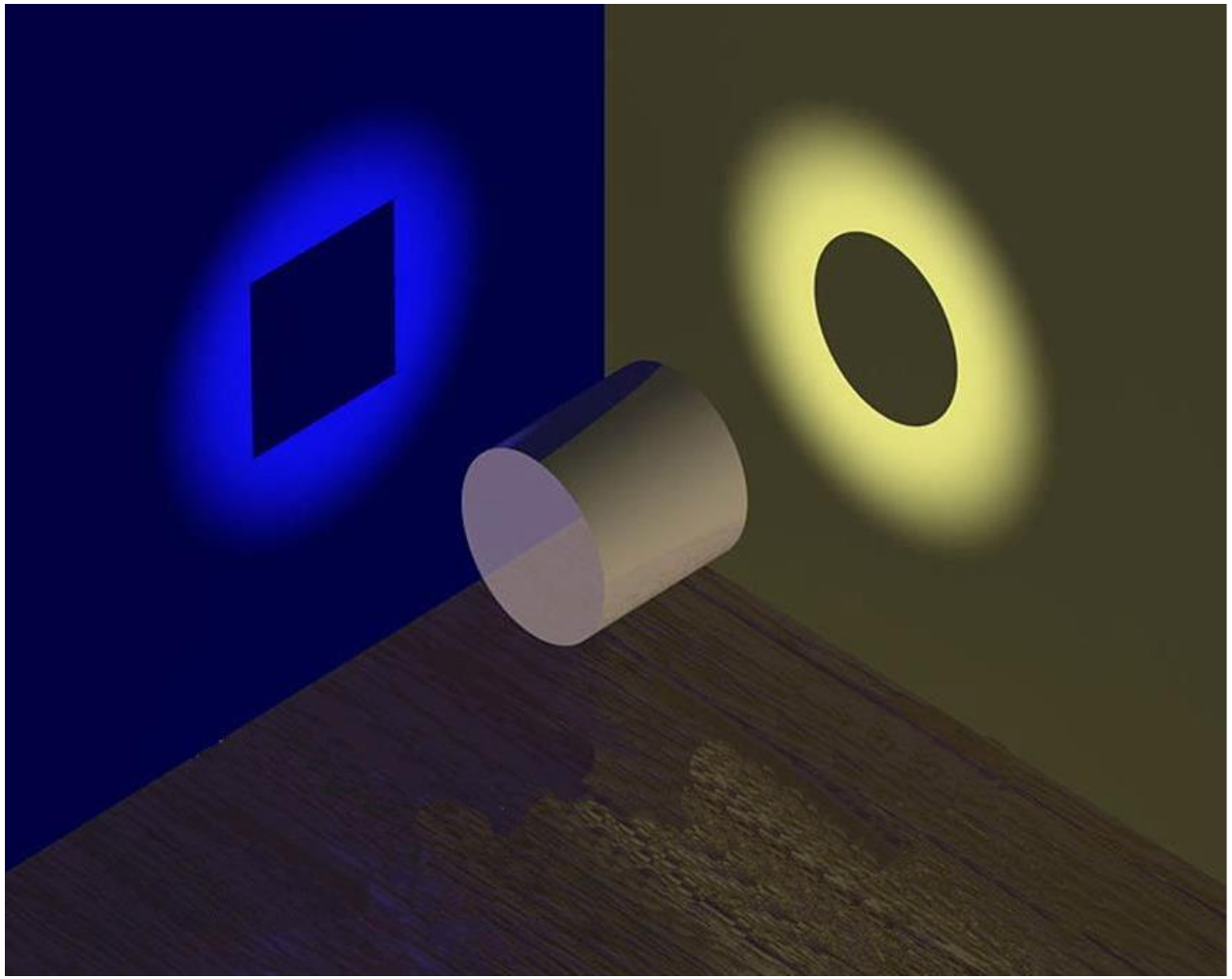
Onde estão os dados? Em R^n

Temos então 2 posibilidades:

- Ir ao R^n clasificar
- Trazer os dados para R^2

Projetando em 2-D

The way one projects = The way one sees



Porque ‘ver’ em R^n?

Porque:

Frequentemente, é interessante ter uma ferramenta de suporte a decisão para auxiliar na tarefa de classificação. **Busca-se que a decisão final seja tomada pelo usuário e não pelo ‘sistema’**.

Quando:

- Não se quer classificar automaticamente por **razões éticas ou legais** e.g. diagnósticos médicos.
- Existe **informação adicional** difícil de ser modelada mas relevante de ser incluída.

O problema de projeção em 2D

Dado um conjunto de observações X em $\Re^n$, encontre um mapeamento $y = f(x)$ $f : \Re^n \rightarrow \Re^2$ tal que a **informação** (ou a estrutura) existente no espaço original **se preserva** (na medida do possível) em $\Re^2$.

*Mas, como definir '**o que' deve ser preservado**?*

Alguns critérios para trazer dados de R^n para R^2

- Buscar preservar a topologia ou a estrutura de distância no espaço projetado $\Re^2$.

- Produzir agrupamentos concentrados e bem separados no espaço projetado.

Critério de separabilidade

Queremos agrupamentos que sejam:

1) O mais separados possível

2) O mais concentrados possível

Existem muitas posibilidades para trazer dos dados de Rn para R^2:

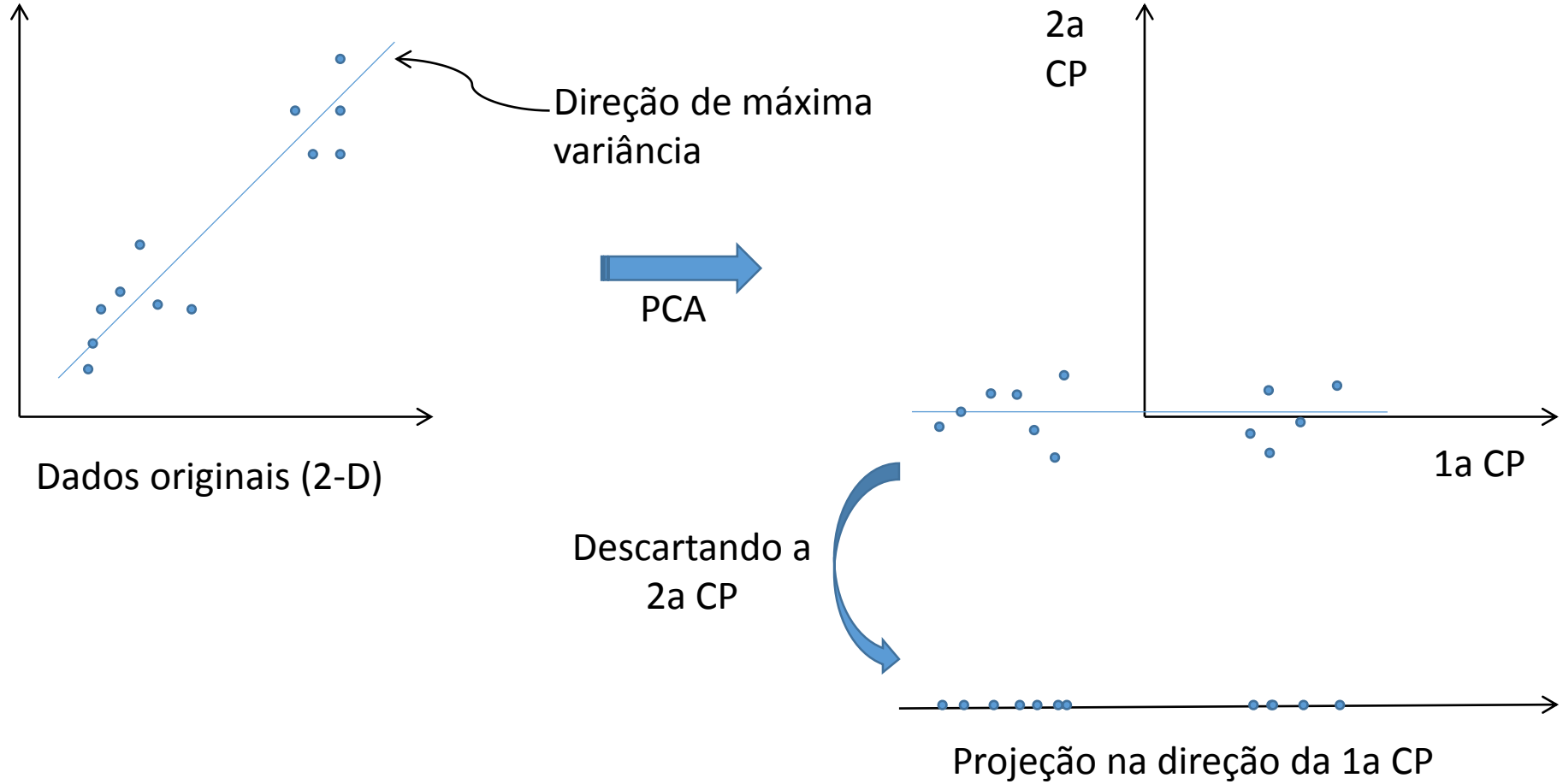
- PCA -Principal Component Analisys
- MDS - Multidimensional Scaling
- t-SNE Stochastic Neighborhood Environment
- Manifold Learning

PCA

Projeções nas componentes pricipais (transformada de Karhunen) **retêm o máximo da variação** presente nos dados no espaço original ($\Re^n$).

Como estamos interessados em '**visualização**', iremos direcionar a atenção à **primeira e segunda componentes**.

Vamos, por simplicidade, considerar uma projeção $\Re^2 \rightarrow \Re$ (normalmente estaríamos interessados em reduzir de $\Re^n \rightarrow \Re^2$)



Porque usar PCA ?
(dispersão como critério)

- **Porque a solução do problema de otimização envolvido é bem conhecida. Existem alguns algoritmos bastante testados para esta finalidade.**
- **Porque funciona bastante bem em muitas situações.**

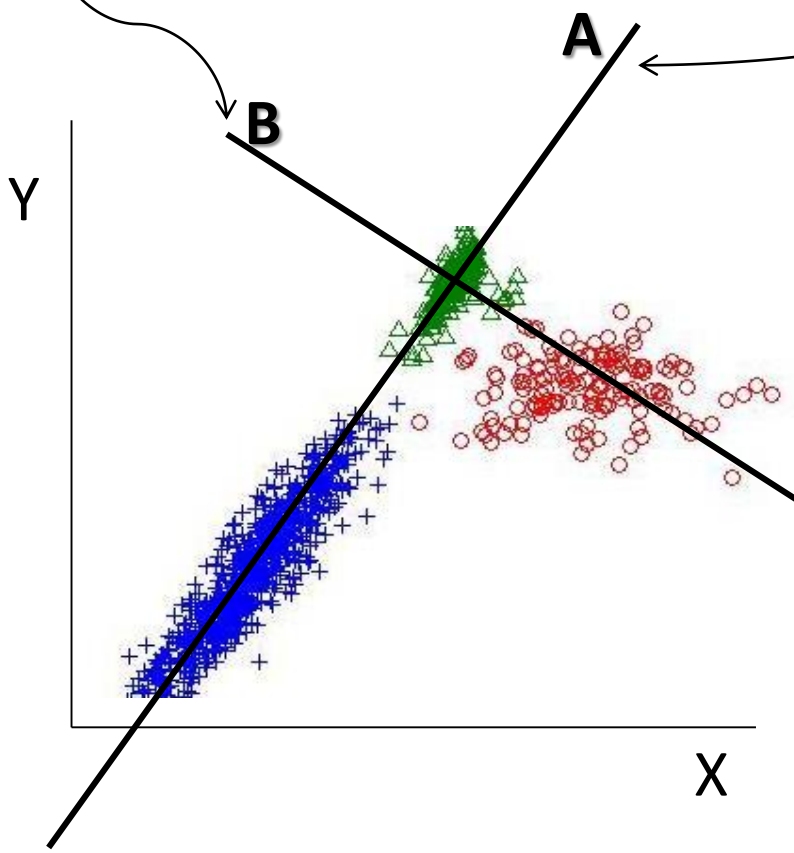
Mas não tão bem quanto gostaríamos ...

Porque?

Quando PCA vai mal para classificação

A direção **B** seria um desastre para agrupamentos azul e verde

Agrupamentos azul e verde se separam muito bem na direção **A**



A direção **A** também é boa para azul e vermelho

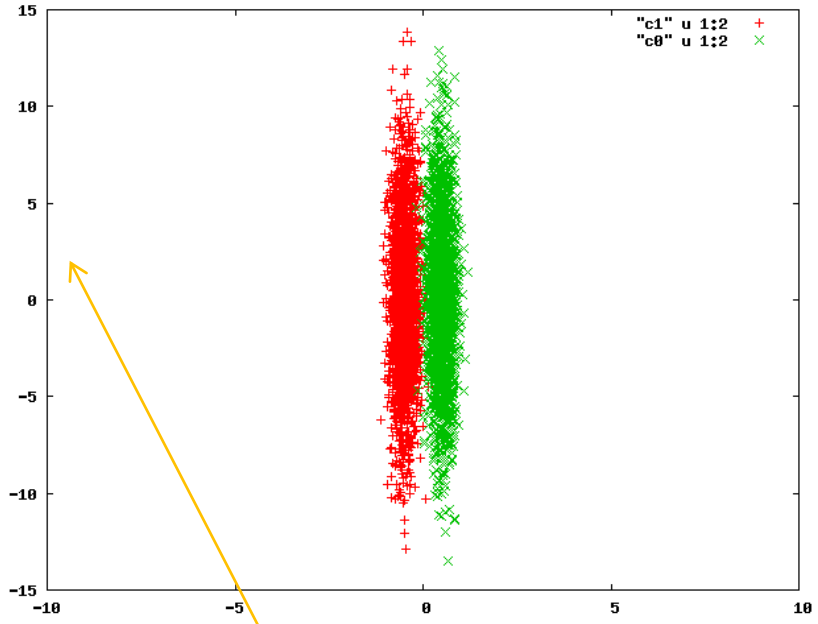
Mas não tão boa para agrupamentos verde e vermelho

B

Estes seriam melhor separados na **B**

Quando PCA vai mal para classificação

Esta direção seria melhor



A direção de máxima variância não separa os dados de nenhuma maneira.

MDS - Muldimensional Scaling

Dado um conjunto de observações em $\Re^n$, busca-se a melhor representação em 2-D tal que a **estrutura original de distância** seja prevervada.

Note-se que este problema em geral não tem uma solução perfeita.

Vamos então buscar uma solução otimizada.

MDS Multidimensional Scaling

Seja, D_{ij} a distância entre as observações x_i e x_j .

O Problema é encontrar o conjunto de pontos $z_i \in \Re^2$ (i=1…n) tal que o seguinte critério é minimizado:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(D_{ij} - \left\|z_i - z_j\right\|_2\right)^2$$

Indo ao R^n

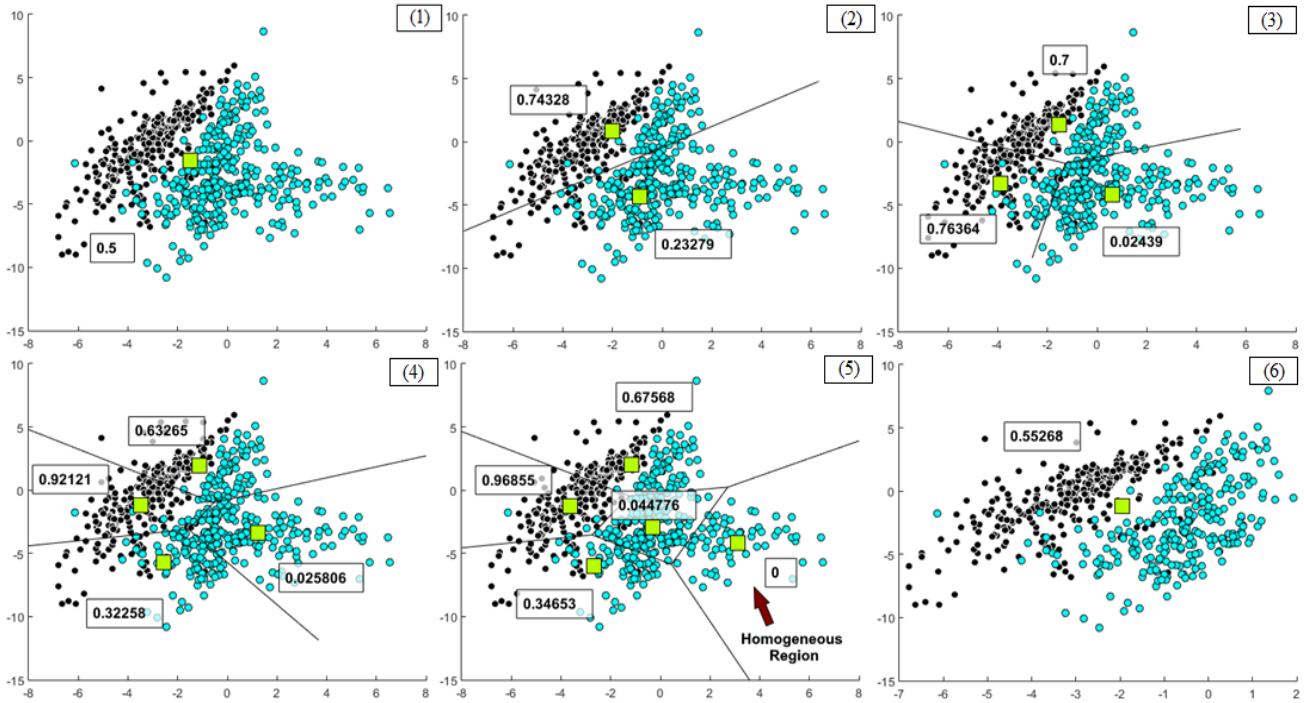
- SVM - Support Vector Machine
- Redes Neurais
- Métodos Local-Global

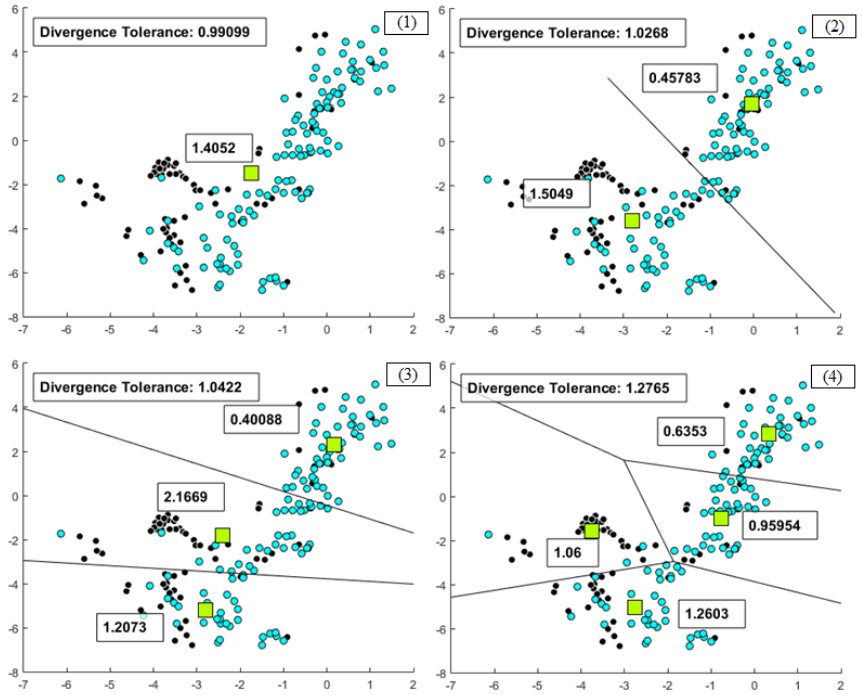
A New Partitioning Algorithm for Classification

C.E. Pedreira[a], C.G. Marcelino[a], L.M. da Costa[a], R.T. Peres[b]and E.V. Leite[a]

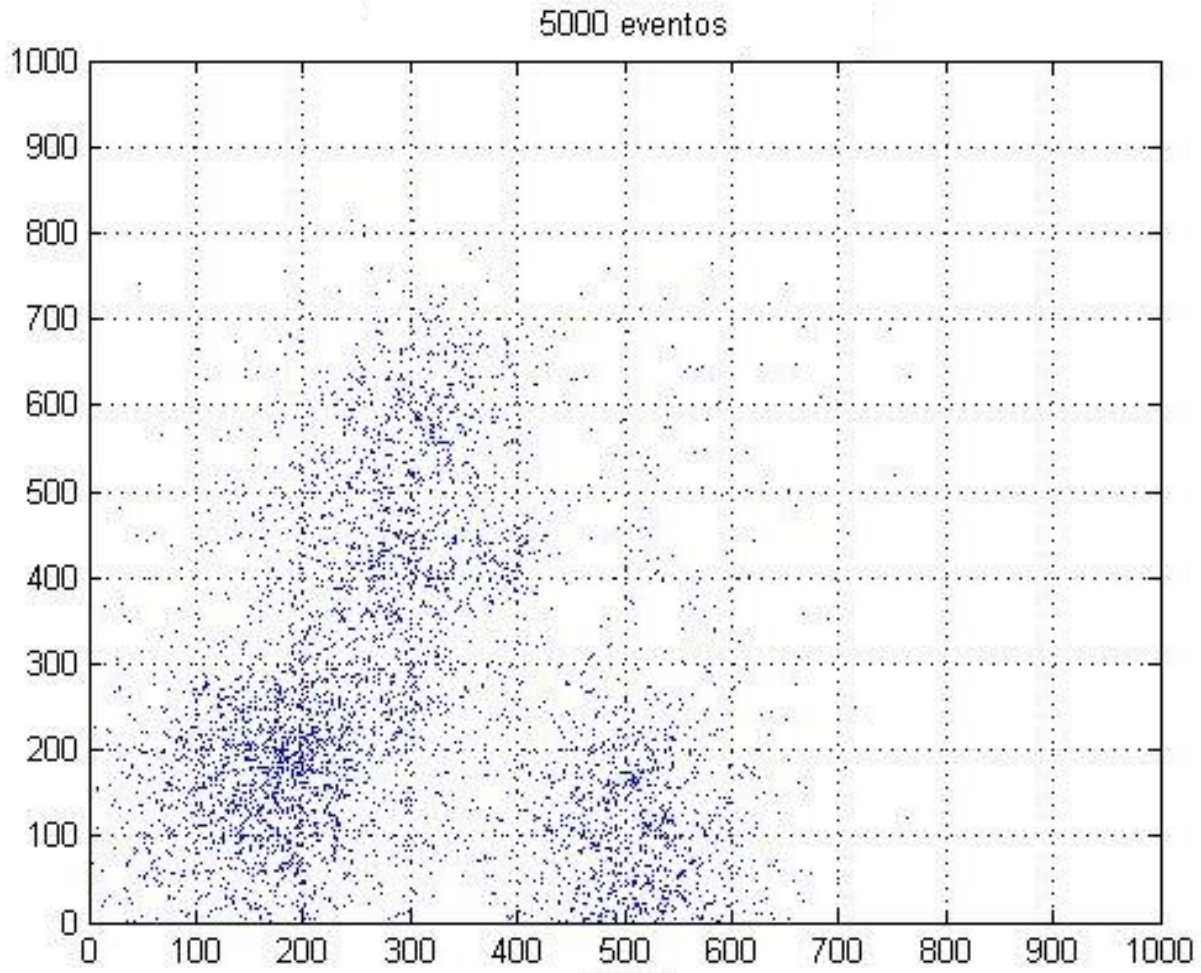
[a]*COPPE – PESC – Systems Engineering and Computer Science Program, Federal University of Rio de Janeiro (UFRJ), Brazil*

[b]*Mathematics Department (DEMAT), Federal Center of Technological Education of Rio de Janeiro (CEFET/RJ), Brazil*

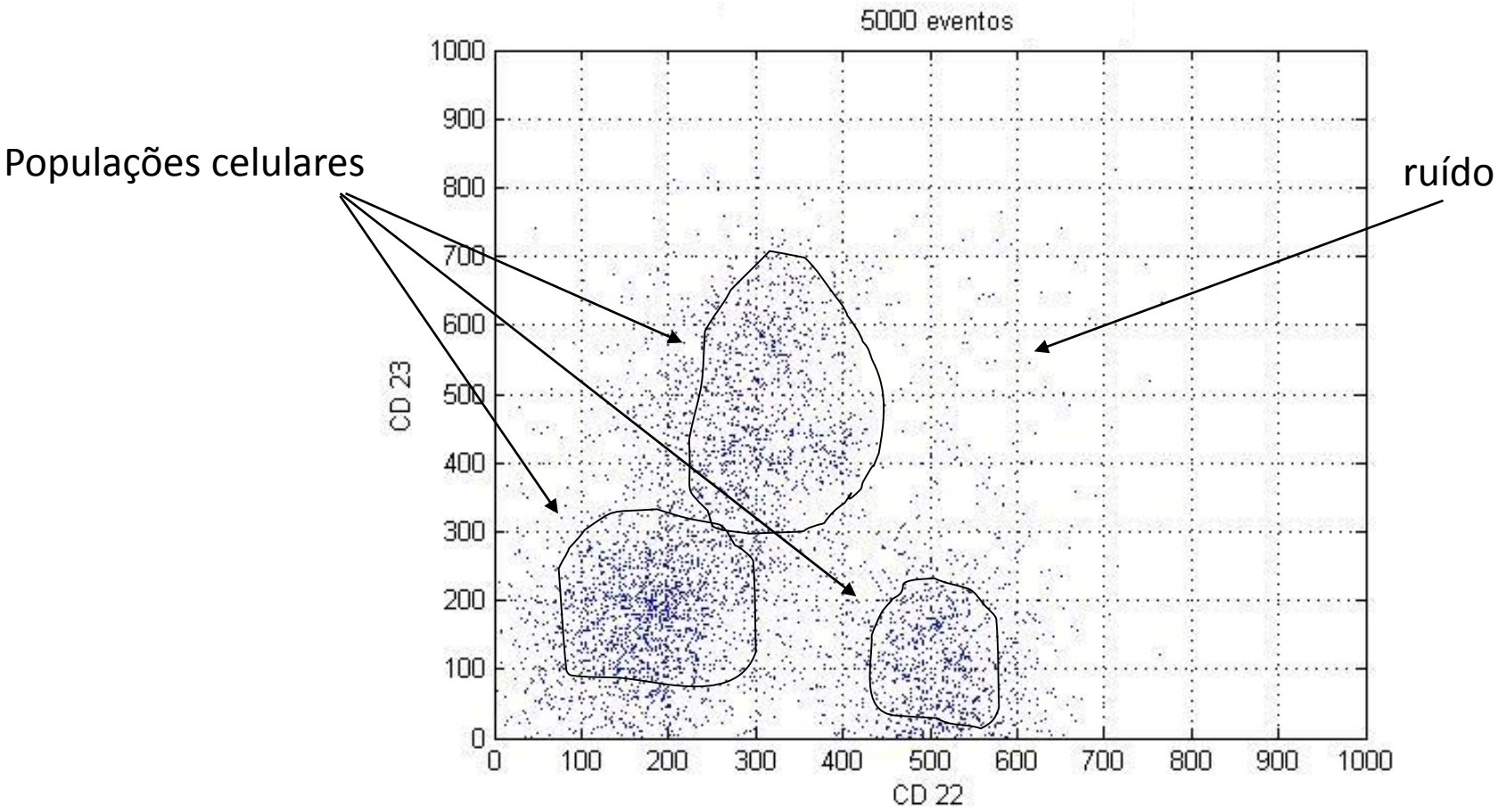




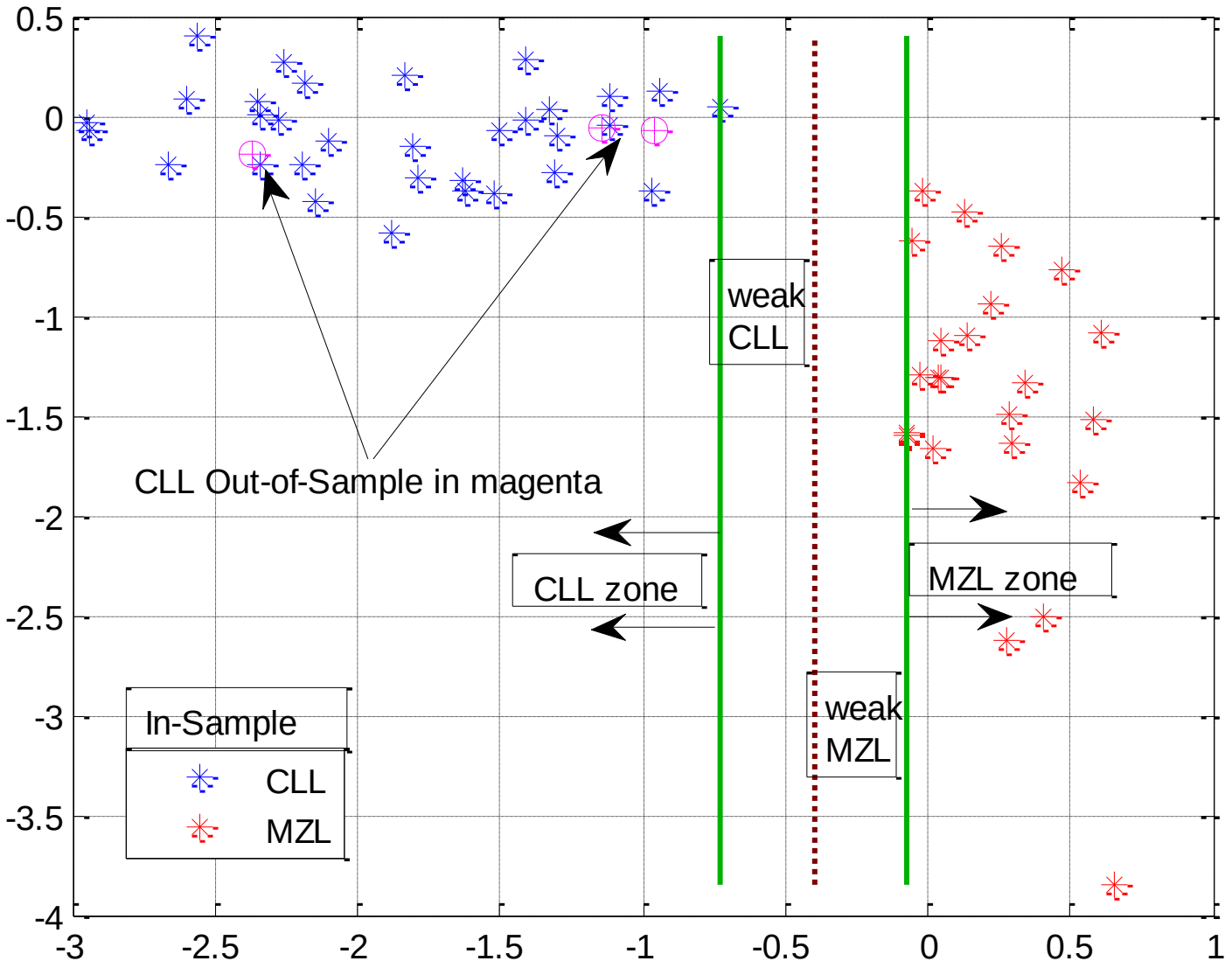
Um problema de classificação populações biologicas



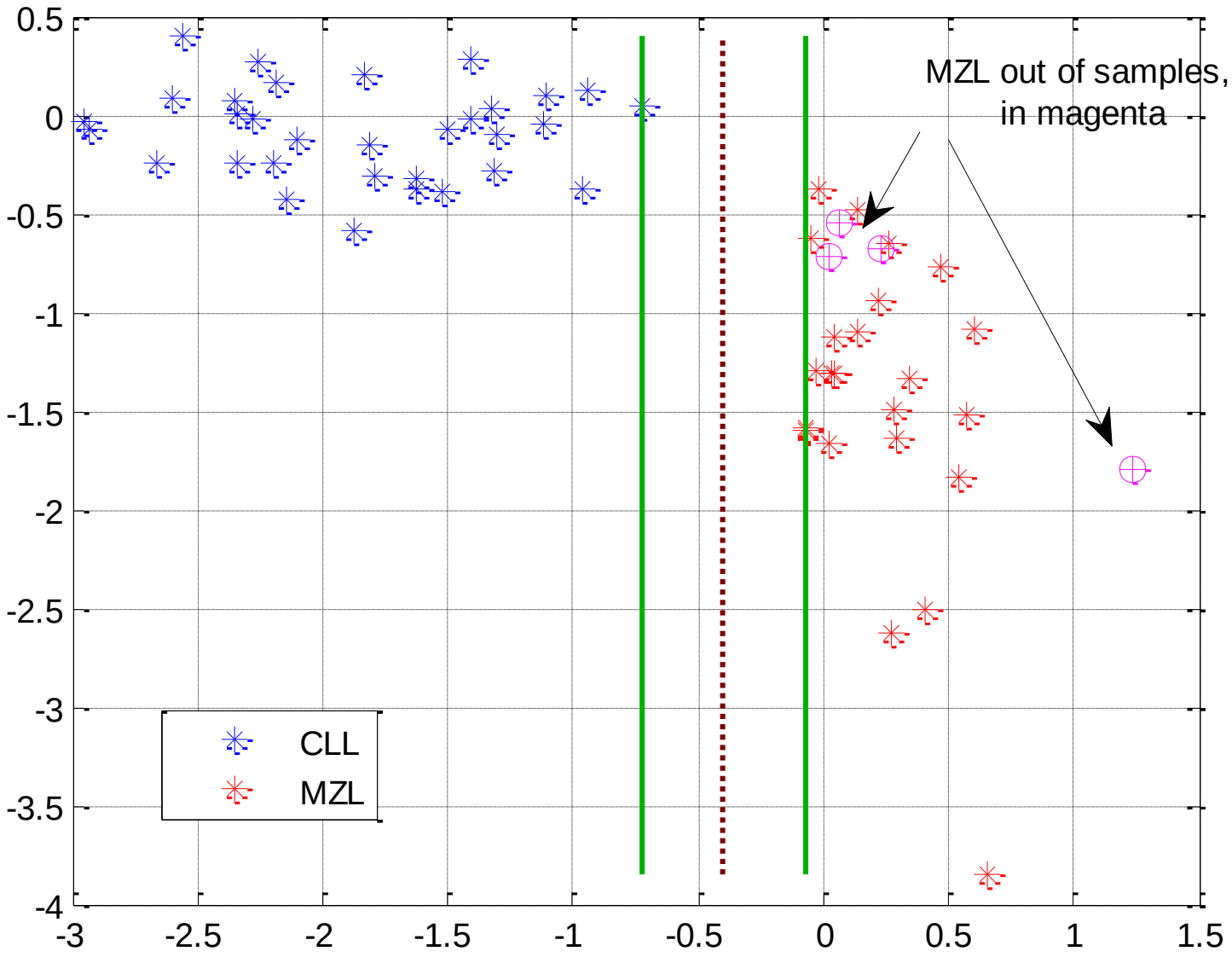
mas onde estão os grupos?



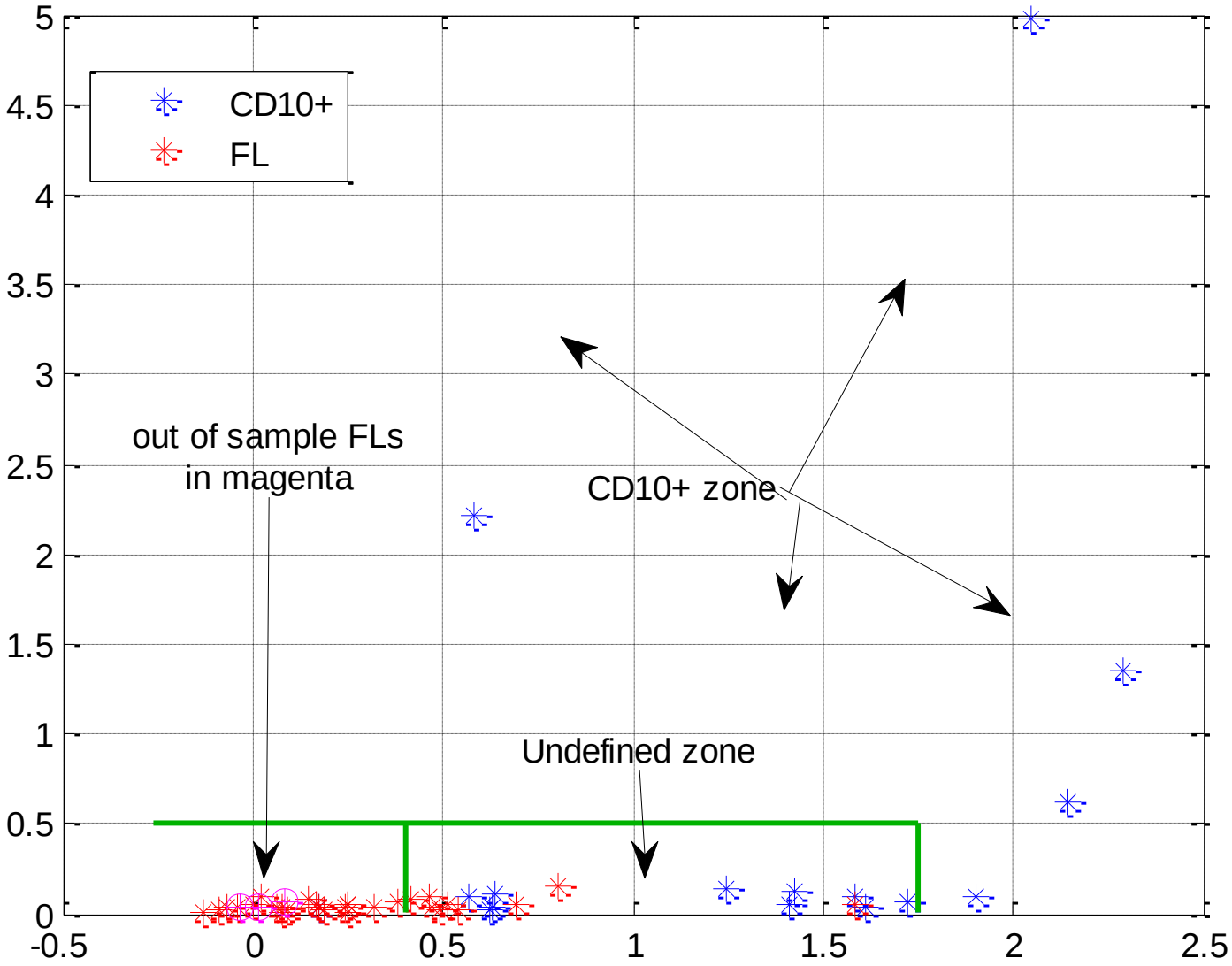
Out of Sample CLL X MZL



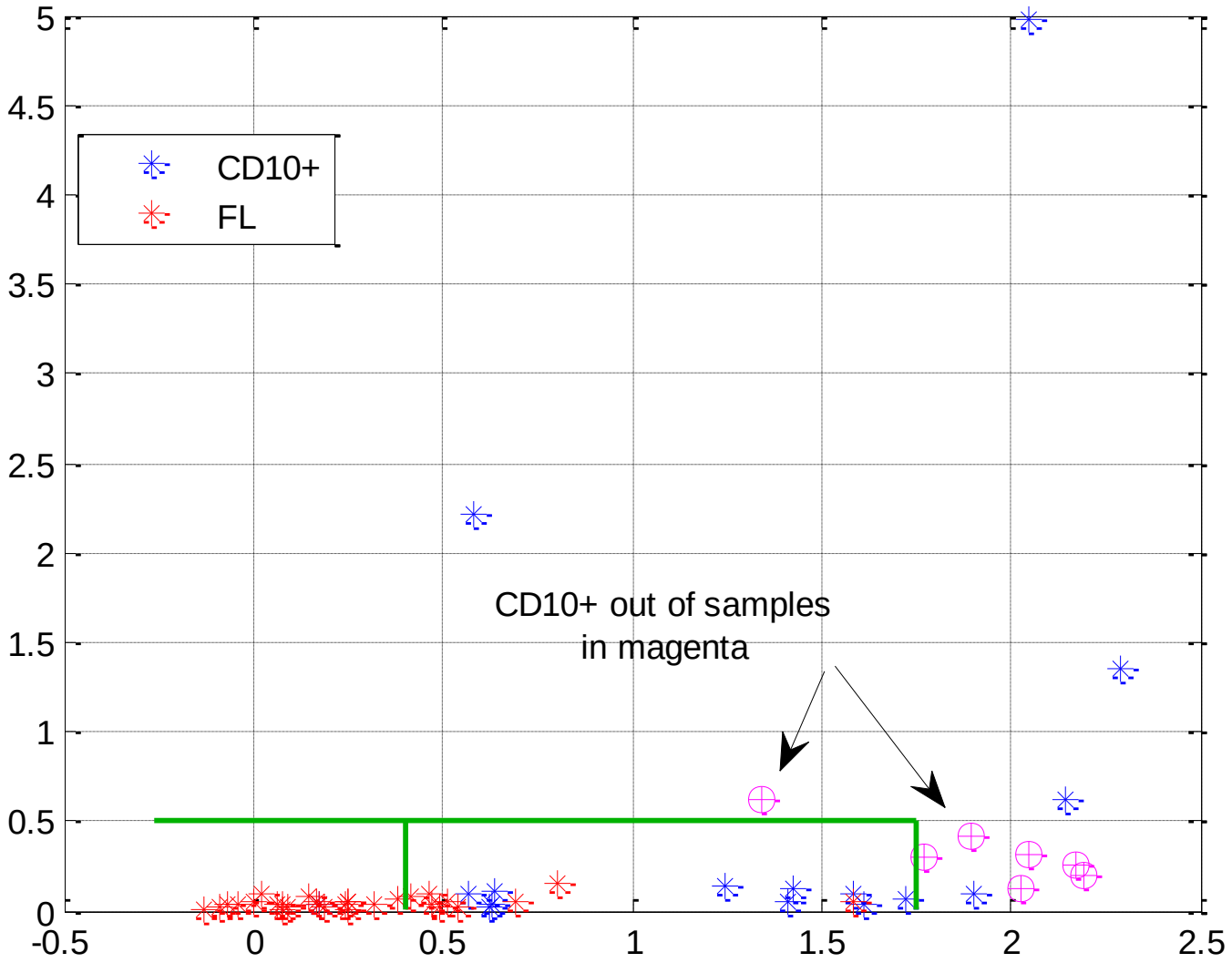
Out of Sample CLL X MZL



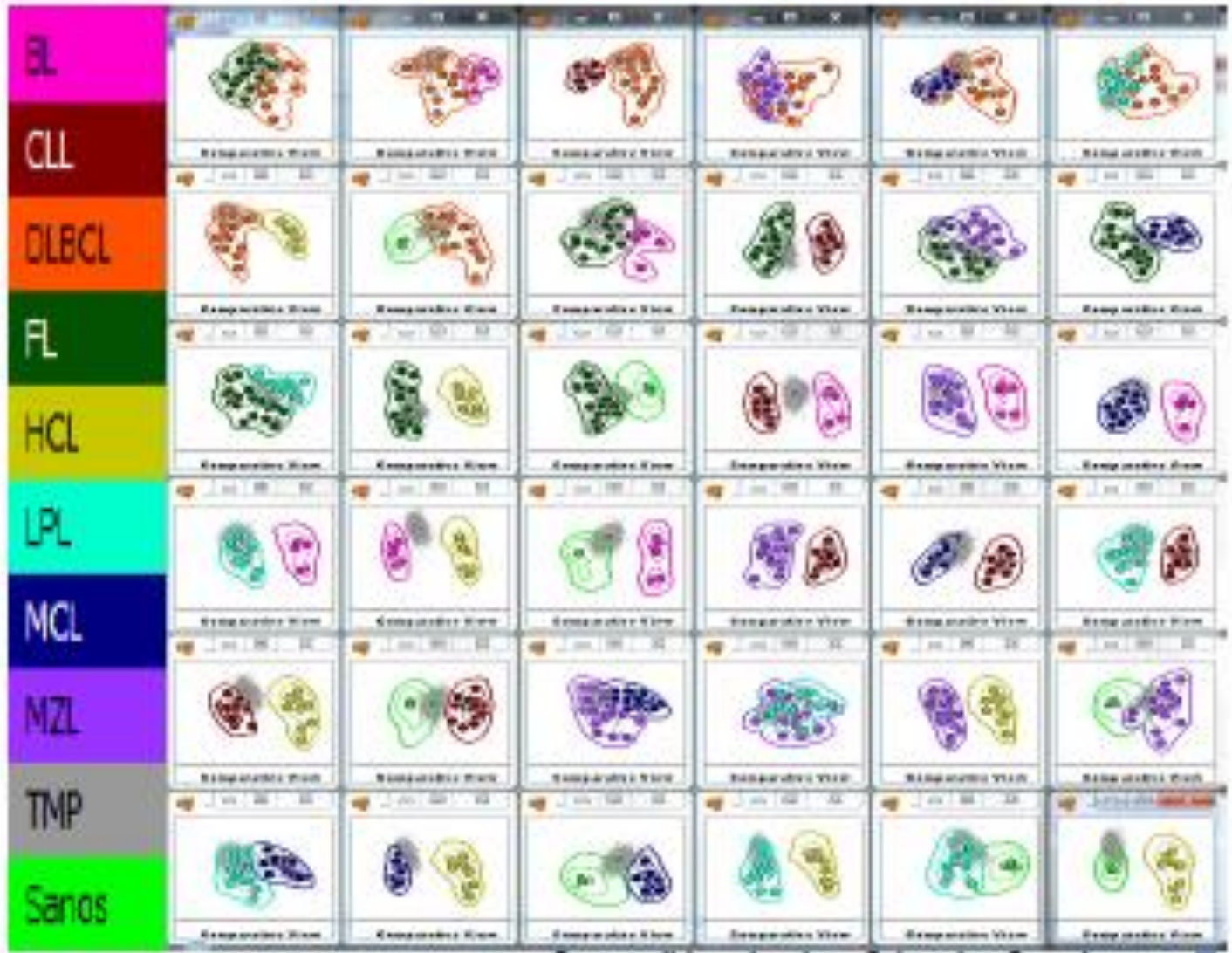
Out of Sample dificult case CD10+ X FL



Out of Sample dificult case CD10+ X FL



BCLPD panel: classification of an atypical case vs the reference WHO diagnostic groups



Responsible scientist: Sebastian Bottcher