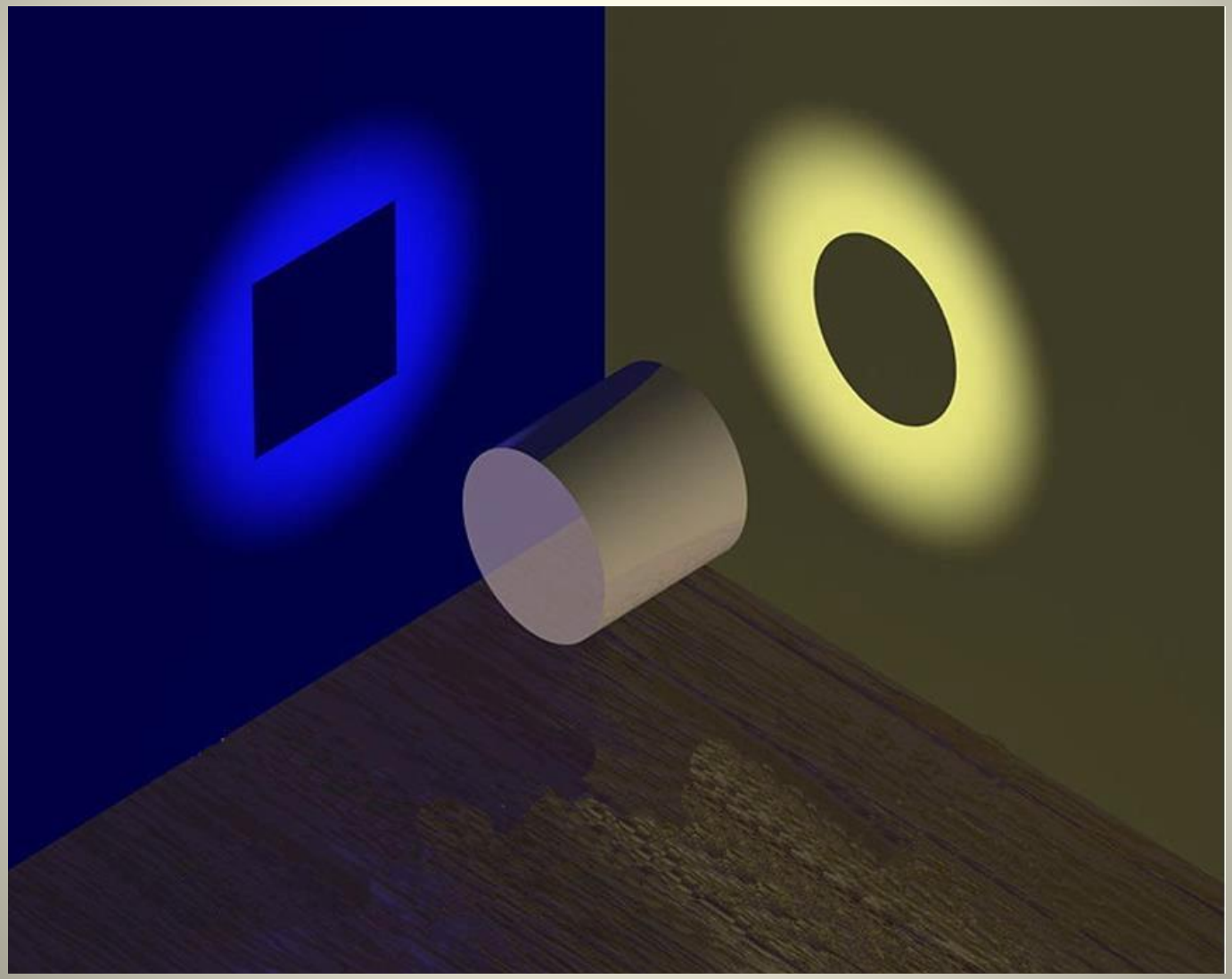
Podemos ver em R^n ?

Carlos Eduardo Pedreira

PESC - COPPE

2018

Como se projeta = Como se vê



Porque (e quando) queremos ‘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 $\mathfrak{R}^n$, encontre um mapeamento $y = f(x)$ $f : \mathfrak{R}^n \rightarrow \mathfrak{R}^2$

tal que a **informação** (ou a estrutura) existente no espaço original **se preserva** (na medida do possível) em $\mathfrak{R}^2$.

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

Escolhendo critérios para projetar os dados

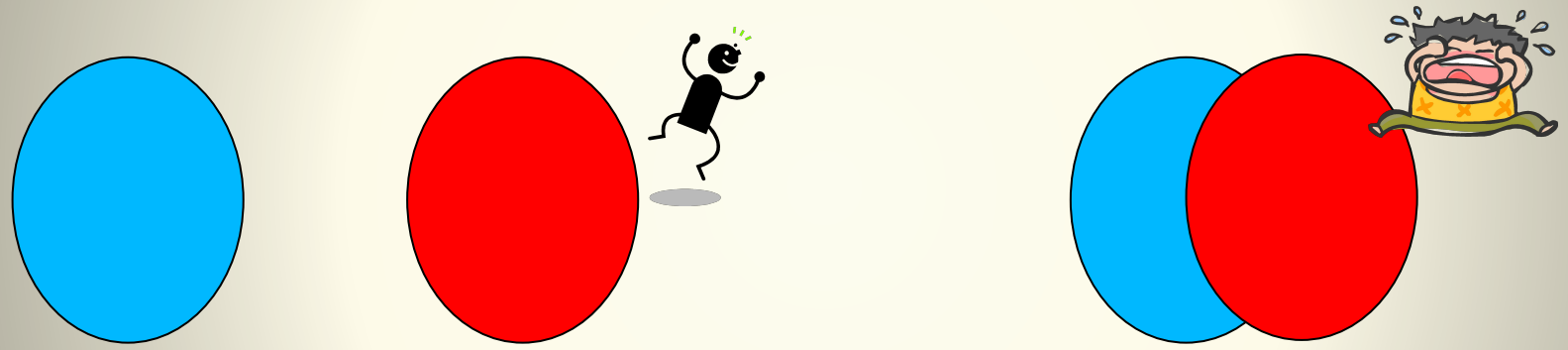
- Minimizar o *erro médio quadrático de reconstrução*.
- Buscar preservar a topologia ou a estrutura de distância no espaço projetado $\Re^2$.
- Produzir agrupamentos concentrados e bem separados no espaço projetado.

Bom para classificação!

Critério de separabilidade

Queremos agrupamentos que sejam:

1) O mais separados possível



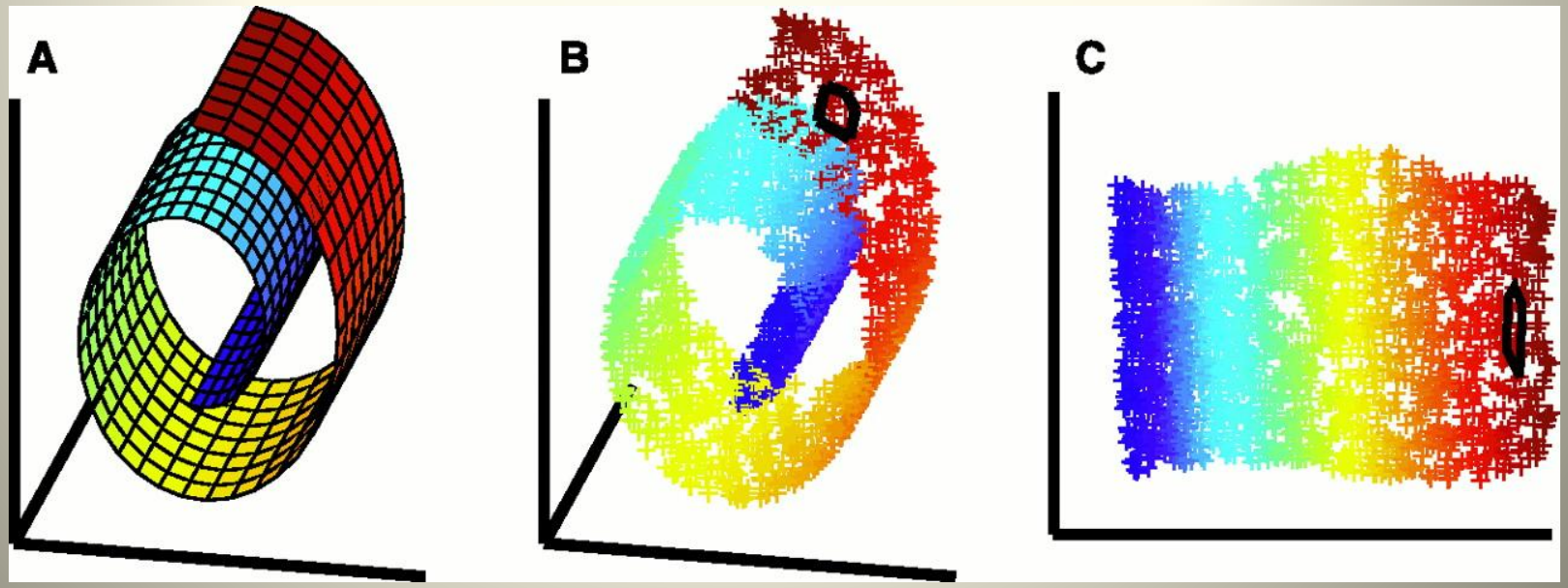
2) O mais concentrados possível

Exstem muitas possibilidades:

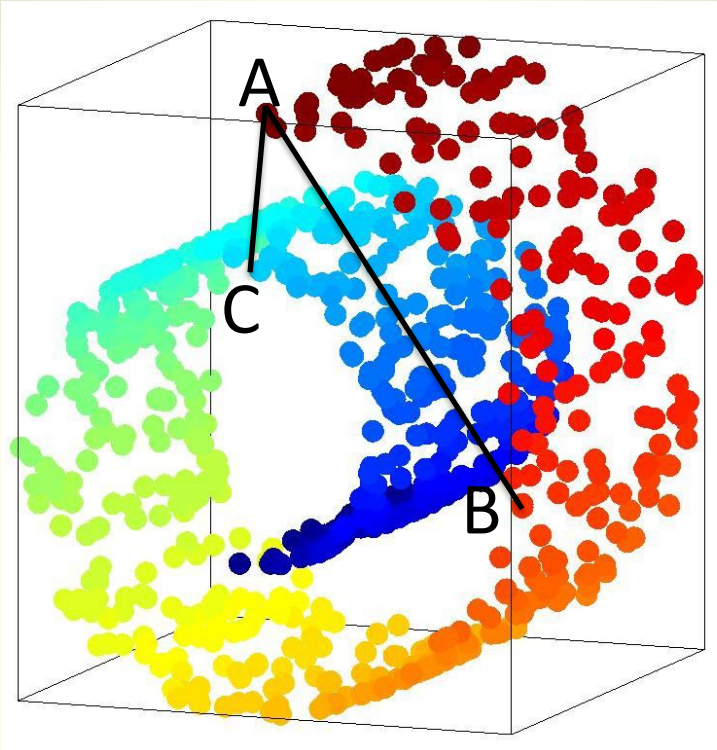
- Manifold Learning
- PCA -Principal Component Analisys
- MDS - Multidimensional Scaling Supervisionado
- Outros métodos (alguns em desenvolvimento)

Manifold Learning: Desenrolando o Rocambole

A ideia central é **revelar uma ´dimensão intrínseca´** dos dados usando uma **métrica** baseada no menor caminho em um **grafo de vizinhos mais próximos**.



Se usarmos a distância Euclidiana, $D_{AC} < D_{AB}$



a estrutura real dos dados seria ocultada

Manifold Learning

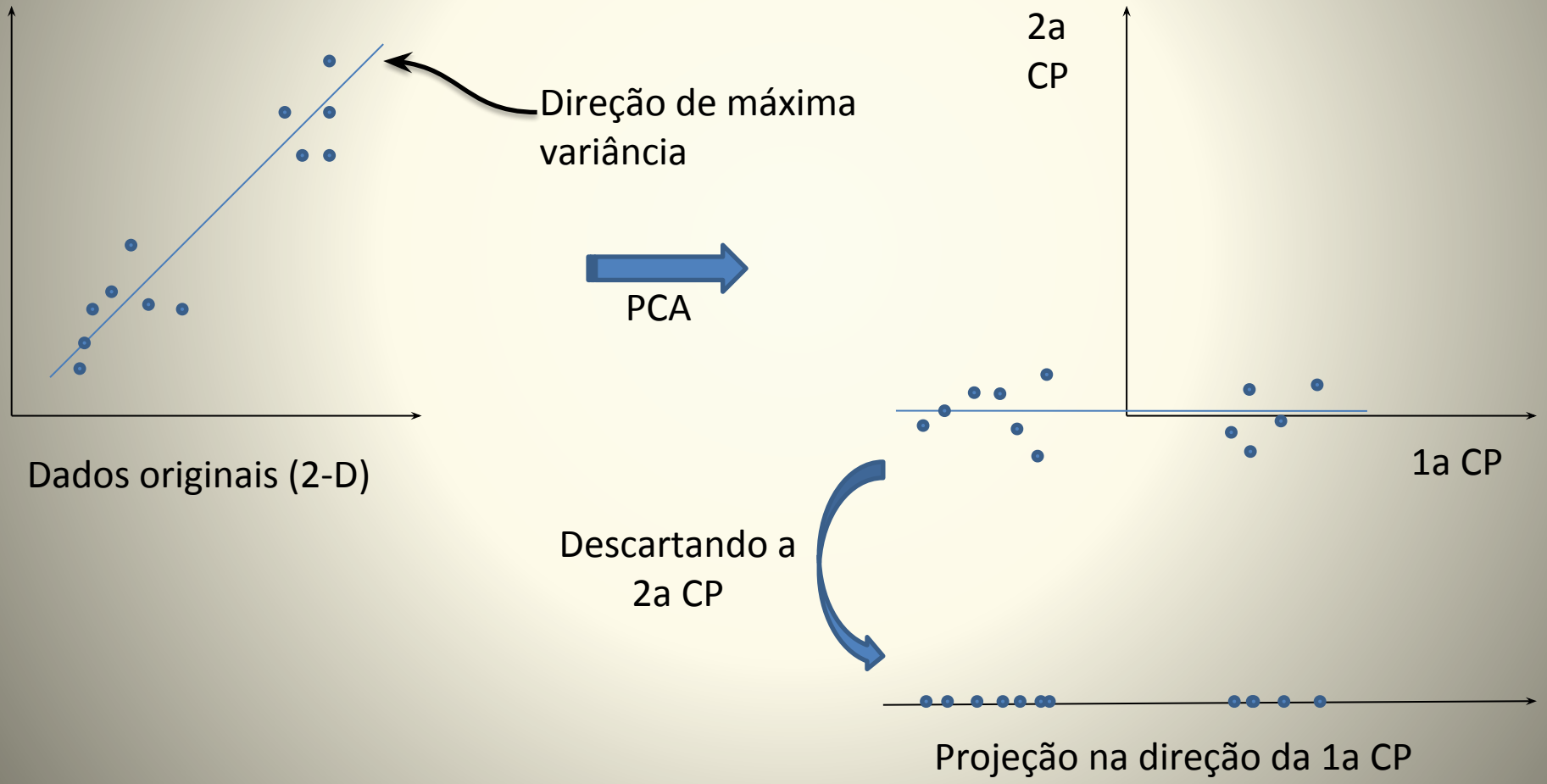
- Atrativo do ponto de vista teórico
- Dificil mostrar que essa estrutura de rocambole de fato existe em problemas reais
- Há vários algoritmos disponíveis na literatura
- Computacionalmente caro
- Sensível à ruído

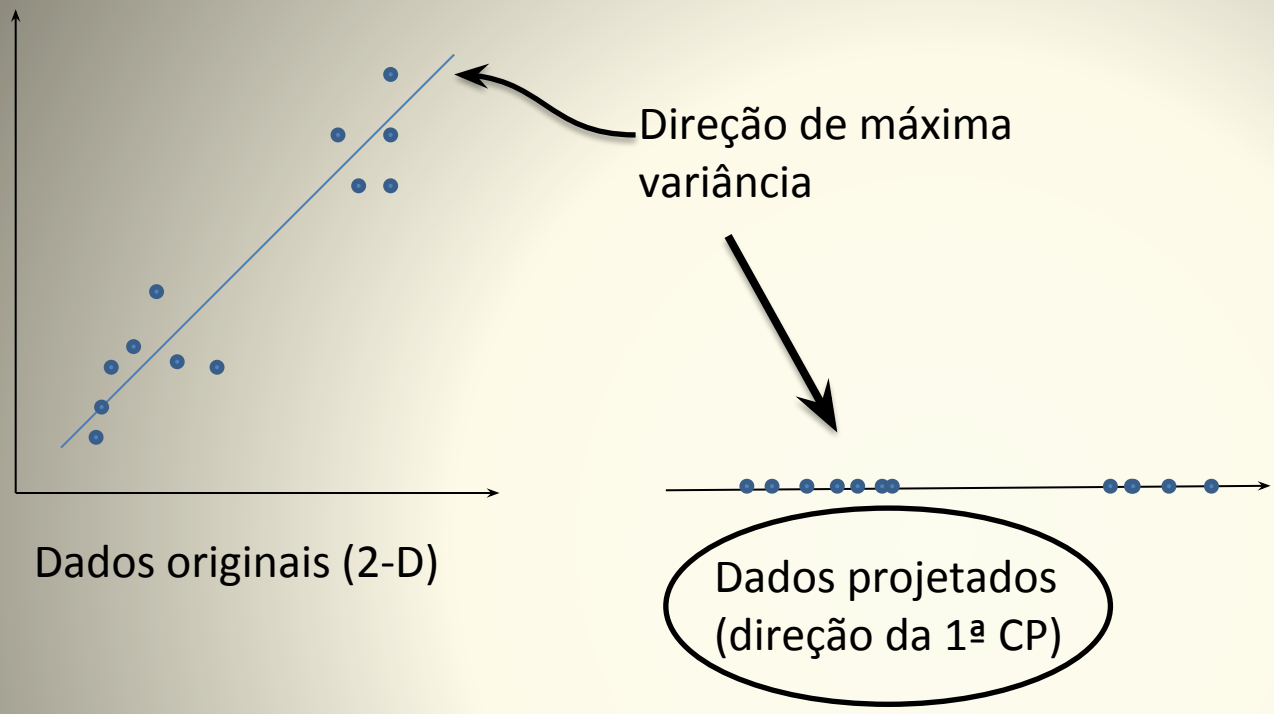
PCA

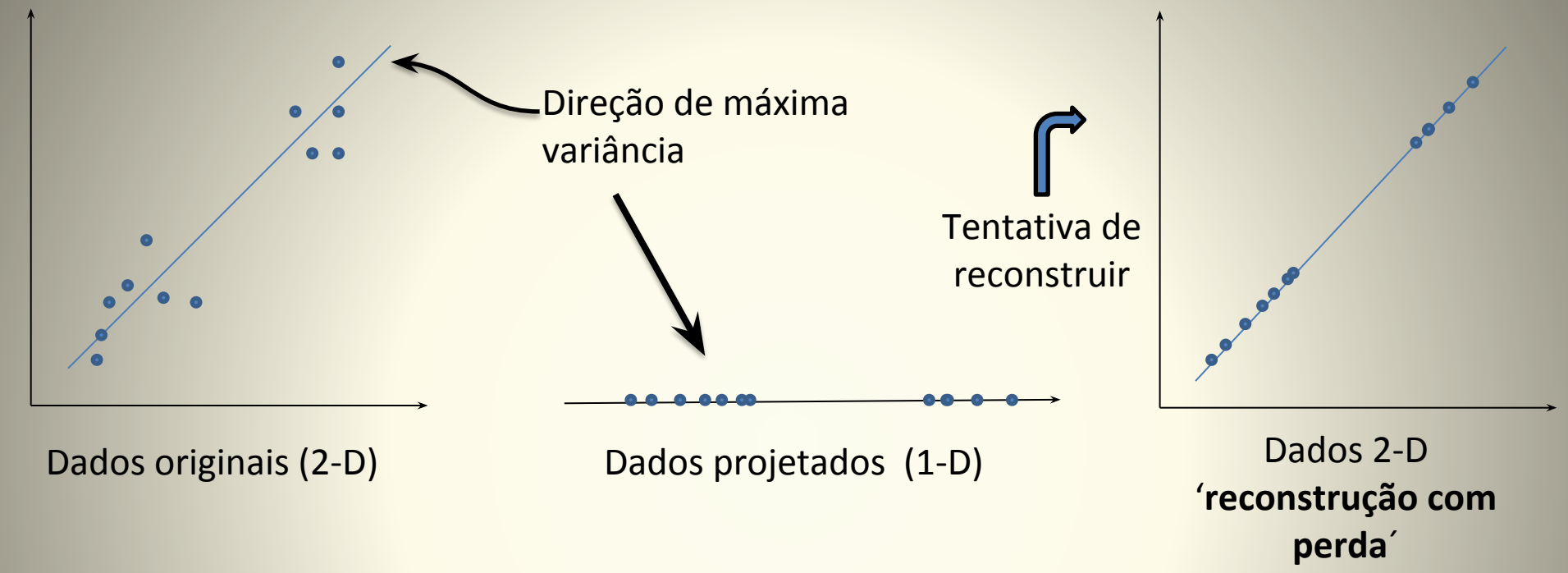
Projeções nas componentes pricipais (transformada de Karhunen) **retêm o máximo da variação** presente nos dados no espaço original ($\Re^n$).

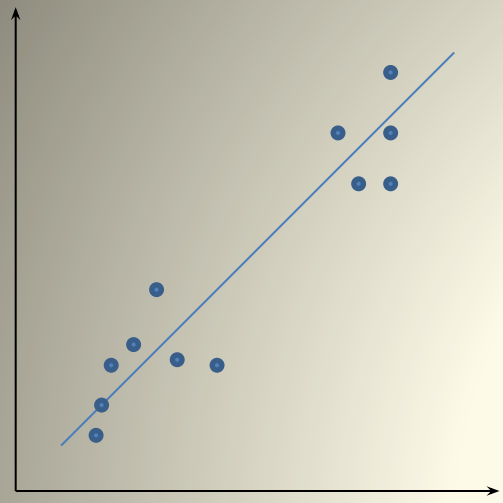
Como estamos interessados em '**visualização**', iremos direcionar a atenção à **primeira e segunda componentes**.

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



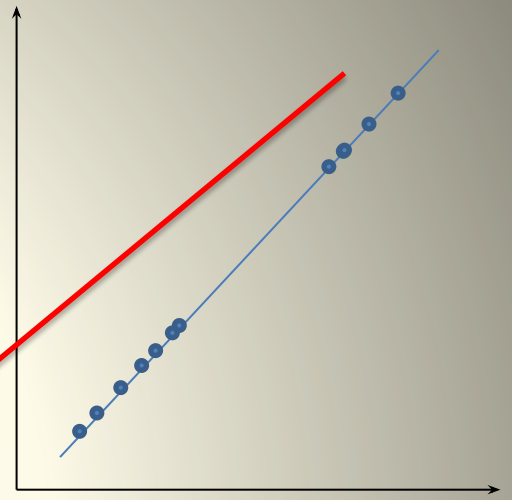

Direção de máxima
variância
Dados originais (2-D)
Dados projetados
(direção da 1ª CP)


Direção de máxima variância
Tentativa de reconstruir
Dados originais (2-D)
Dados projetados (1-D)
Dados 2-D
‘reconstrução com perda´



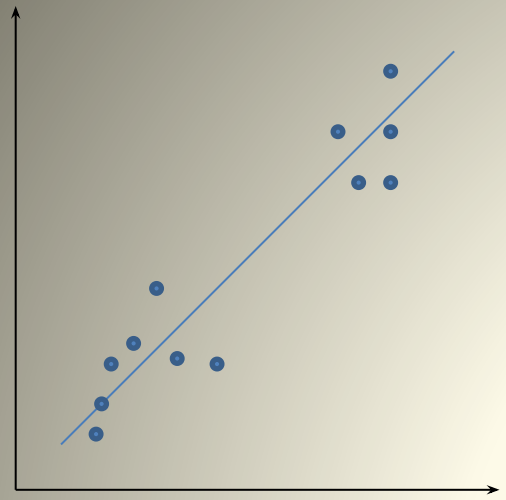
Dados originais (2-D)

Reconstruindo

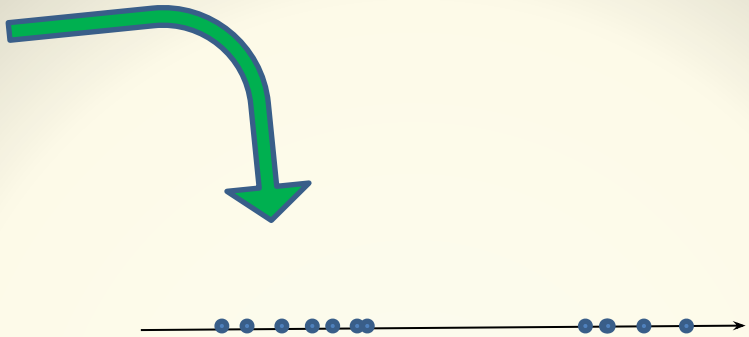


Dados 2-D
‘reconstrução com perda´

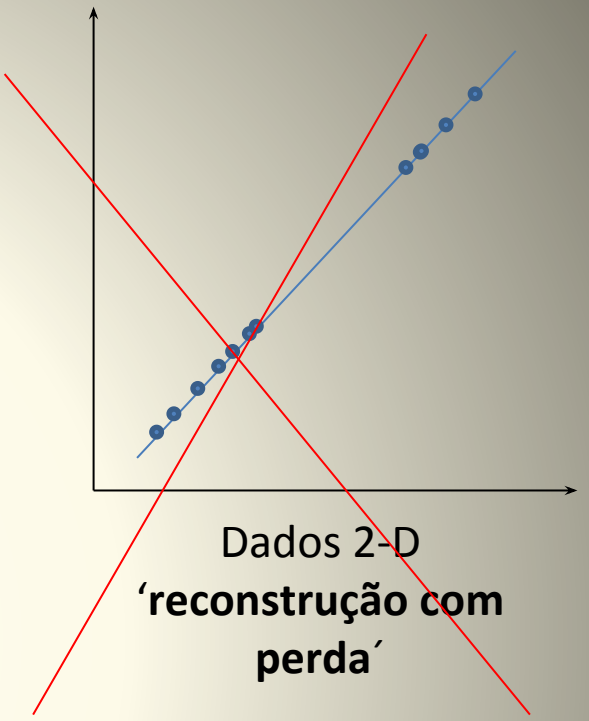
Mas não estamos interesados em reconstruir!



Dados originais (2-D)

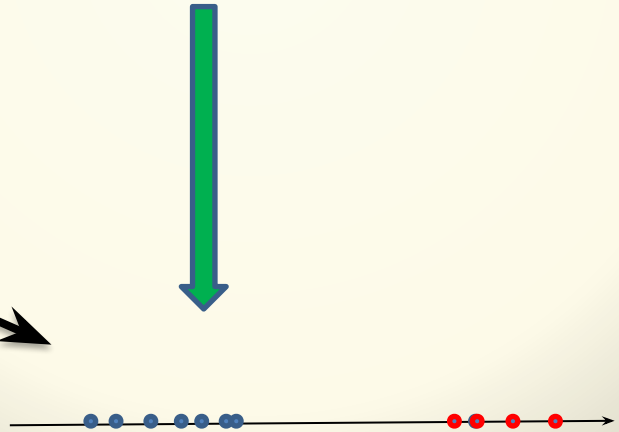


Dados projetados (1-D)



Dados 2-D
'reconstrução com perda'

Queremos classificar



Dados projetados (1-D)

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.**

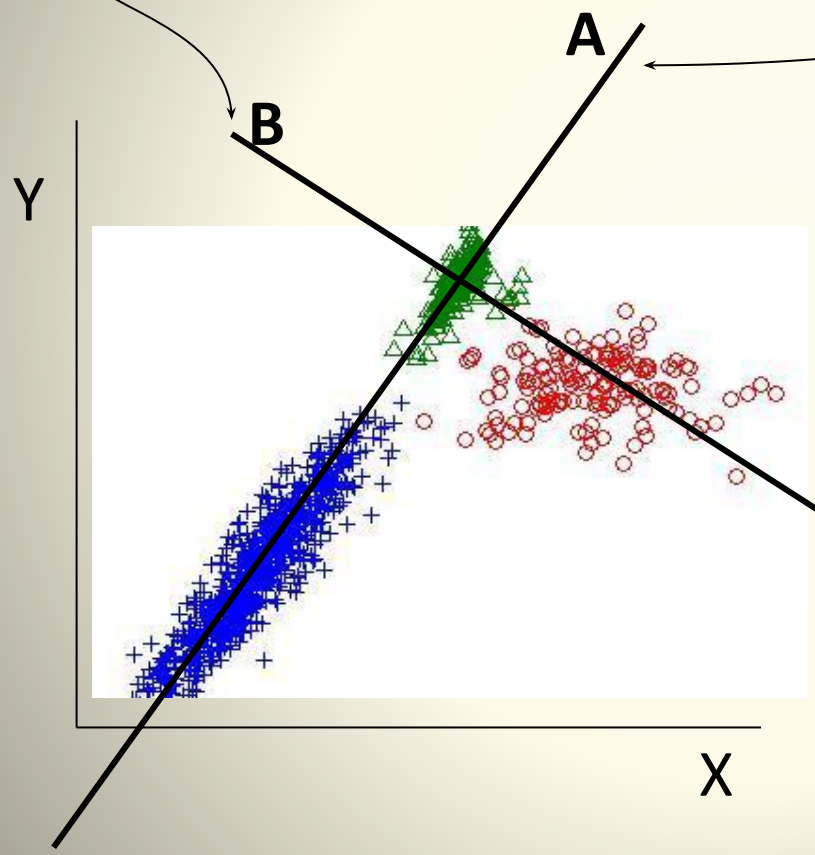
Mas não tão bem quanto gostaríamos ...

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

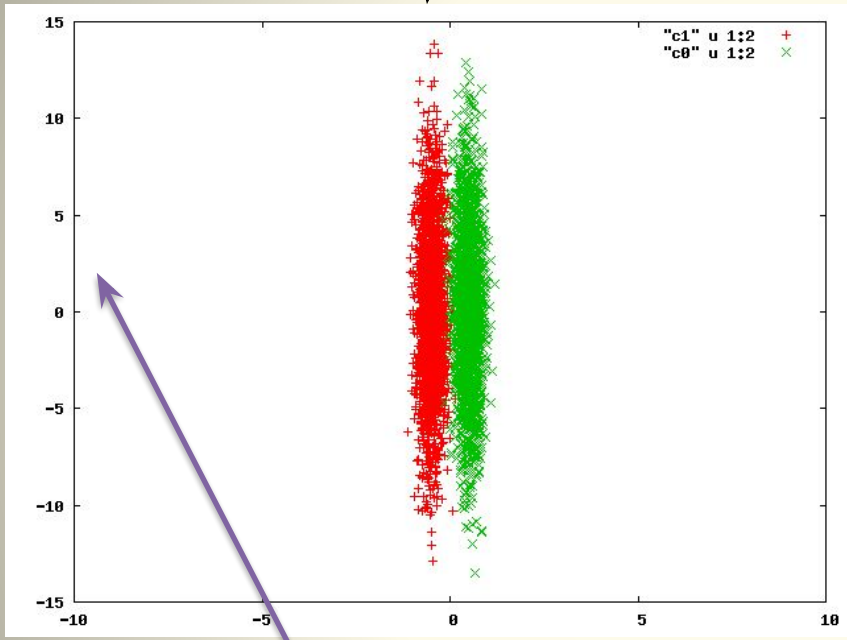
Mas não tão boa para agrupamentos verde e vermelho

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.

SMDS - Supervised Multidimensional Scaling

- É supervisado, utiliza-se portanto os rótulos das classes
- A ‘quantidade’ de supervisão pode ser controlada pelo usuário.

Consideremos um conjunto de 'n' observações $X = \{ x_1, x_2, \dots , x_n\}$, $x_i \in \Re^p$. Seja $D \in \Re^{n\,x\,n}$ uma matriz simétrica que contém informação sobre as dissimilaridades entre os pares de observações. **Ou seja, D_{ij} é a distância (Euclidiana) entre as observações x_i e x_j .**

<u>O Problema</u>: Encontrar o conjunto de pontos $z_i \in \Re^2$ (i=1...n) tal que o seguinte critério é minimizado:

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

Introduzindo Supervisão:

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(D_{ij}-\|\mathbf{z}_i-\mathbf{z}_j\|_2)^2$$

o critério que tínhamos

$$\underset{\mathbf{z}_1,\ldots,\mathbf{z}_n\in\mathbb{R}^S}{\text{minimize}}\left\{\frac{1}{2}(1-\alpha)\sum_{i=1}^{n}\sum_{j=1}^{n}(D_{ij}-\|\mathbf{z}_i-\mathbf{z}_j\|_2)^2+\alpha\sum_{i:y_i=1}\sum_{j:y_j=2}\sum_{s=1}^{2}\left(\frac{D_{ij}}{\sqrt{S}}-(z_{js}-z_{is})\right)^2\right\}$$

α É usado para controlar a supervisão

Quanto maior for α, mais importancia é dada ao 2º termo, e consequentemente mais supervisado é o algoritmo.

Entendendo o custo que deve ser minimizado

2 dimensões

$$\underset{\mathbf{z}_1,\ldots,\mathbf{z}_n \in \mathbb{R}^S}{\text{minimize}} \left\{ \frac{1}{2}(1-\alpha) \sum_{i=1}^{n} \sum_{j=1}^{n} (D_{ij} - \|\mathbf{z}_i - \mathbf{z}_j\|_2)^2 + \alpha \sum_{i:y_i=1} \sum_{j:y_j=2} \sum_{s=1}^{2} \left(\frac{D_{ij}}{\sqrt{S}} - (z_{js} - z_{is}) \right)^2 \right\}$$

Supervisão está aqui: precisamos saber $y_i=1$ ou $y_j=2$ (agrupamentos 1 ou 2)

$$\underset{z_1,\ldots,z_n\in\mathbb{R}^S}{\text{minimize}} \left\{ \frac{1}{2}(1-\alpha)\sum_{i=1}^{n}\sum_{j=1}^{n}(D_{ij}-\|z_i-z_j\|_2)^2 + \alpha \sum_{i:y_i=1}\sum_{j:y_j=2}\sum_{s=1}^{2}\left(\frac{D_{ij}}{\sqrt{S}}-(z_{js}-z_{is})\right)^2 \right\}$$

ZOOM

$$\left\{ \sum_{j=1}^{n}(D_{ij}-\|z_i-z_j\|_2)^2 + \alpha \sum_{i:y_i=1}\sum_{j:y_j=2}\sum_{s=1}^{2}\left(\frac{D_{ij}}{\sqrt{S}}-(z_{js}-z_{is})\right)^2 \right\}$$

Quanto maior ($z_j - z_i$), menor será o termo de α , assim os "z_j's" (população 2) tendem a se localizar a direita em comparação com os "z_i's" (população 1)

Este critério é não-convexo, e é preciso usar um enfoque iterativo de majoração para resolver o problema de minimização que resulta em:

$$z_{ks} \leftarrow \frac{1}{(n-1)(1-\alpha)+n_2\alpha}\left[(1-\alpha)\sum_{j\neq k} z_{js} + (1-\alpha)\sum_{j\neq k} D_{jk}\frac{\tilde{z}_{ks}-z_{js}}{\|\tilde{\mathbf{z}}_k - \mathbf{z}_j\|_2} + \alpha\sum_{j:y_j=2} z_{js} - \frac{\alpha}{\sqrt{S}}\sum_{j:y_j=2} D_{kj}\right]$$

Iterar para calcular uma melhor localização em 2-D

Escolher um ponto incial de localização da representação da observação em 2-D

Projeção Baseada em Separabilidade

A ideia central é calcular, em um ambiente supervisionado, projeções (lineares) X & Y usando um **critério de separabilidade** (ao invés da variância como no caso de PCA).

Agora, precisamos definir o que precisamente queremos dizer com ´**separabilidade**´ (nosso critério).

1) **Melhor separação**

Usamos o *Divergente para medir o quanto ´próximas estão as distribuições de probabilidade*.

2) **Agrupamentos bem concentrado s**

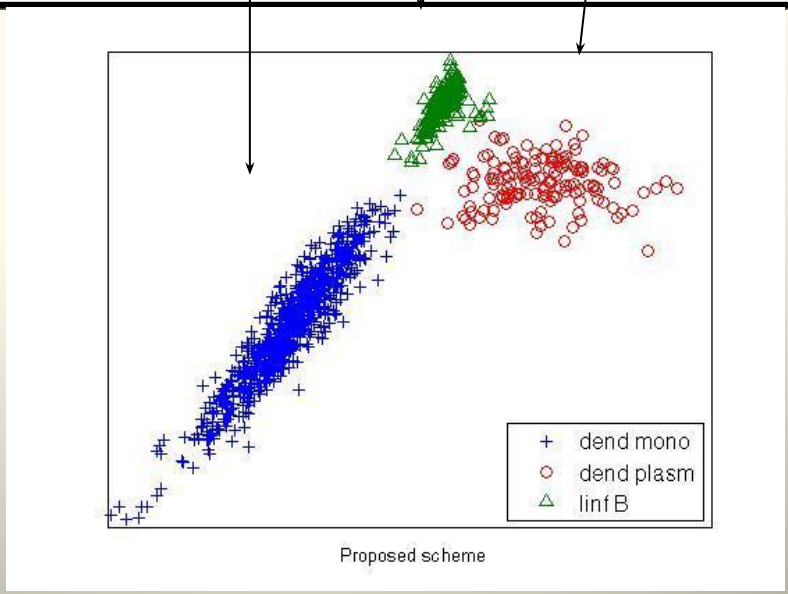
Usamos a *Entropia* para medir a ‘concentração’ dos agrupamentos (quanto menor a entropia, mais concentrados).

Sobre o critério de separação

Queremos maximizar a seguinte função de custo:

Dc-s = CET(C1,C2,C3) – H(C1) – H(C2) – H(C3), **onde:**

- CET(X) é o ***Divergente*** entre agrupamentos;
- H(X) é a **entropia de *Renyi*** dos agrupamentos (mede o quanto concentrados estão);



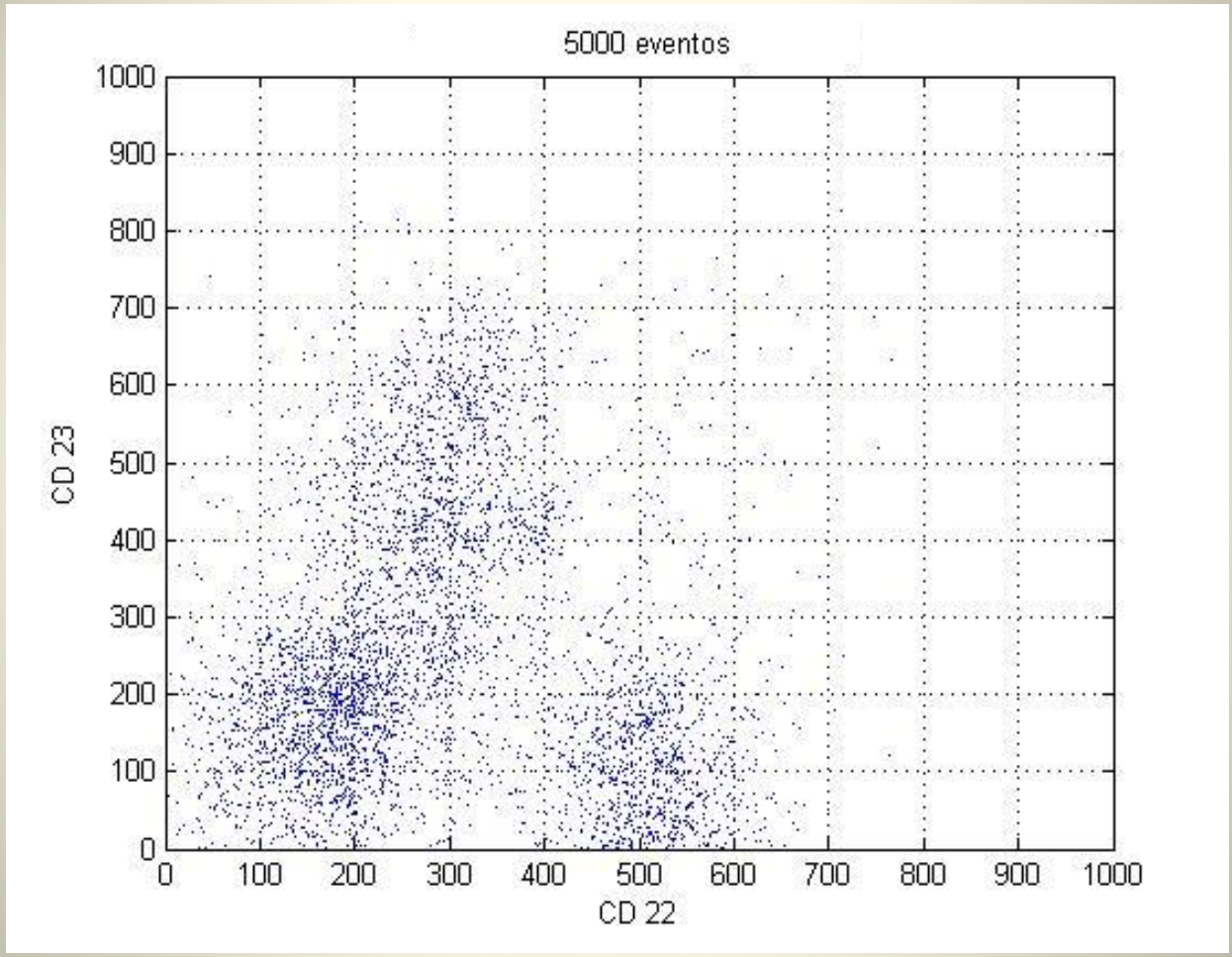
Usamos um **algoritmo baseado em computação evolucionaria** para resolver este `problema de otimização'.

Queremos encontrar as direções X e Y tal que o critério Dc-s seja maximizado. Buscamos então coeficientes A1, A2, A3, B1, B2, B3 … etc em:

$$X = A1 * \alpha_x + A2 * \beta_x + A3 * \delta_x + \ldots$$

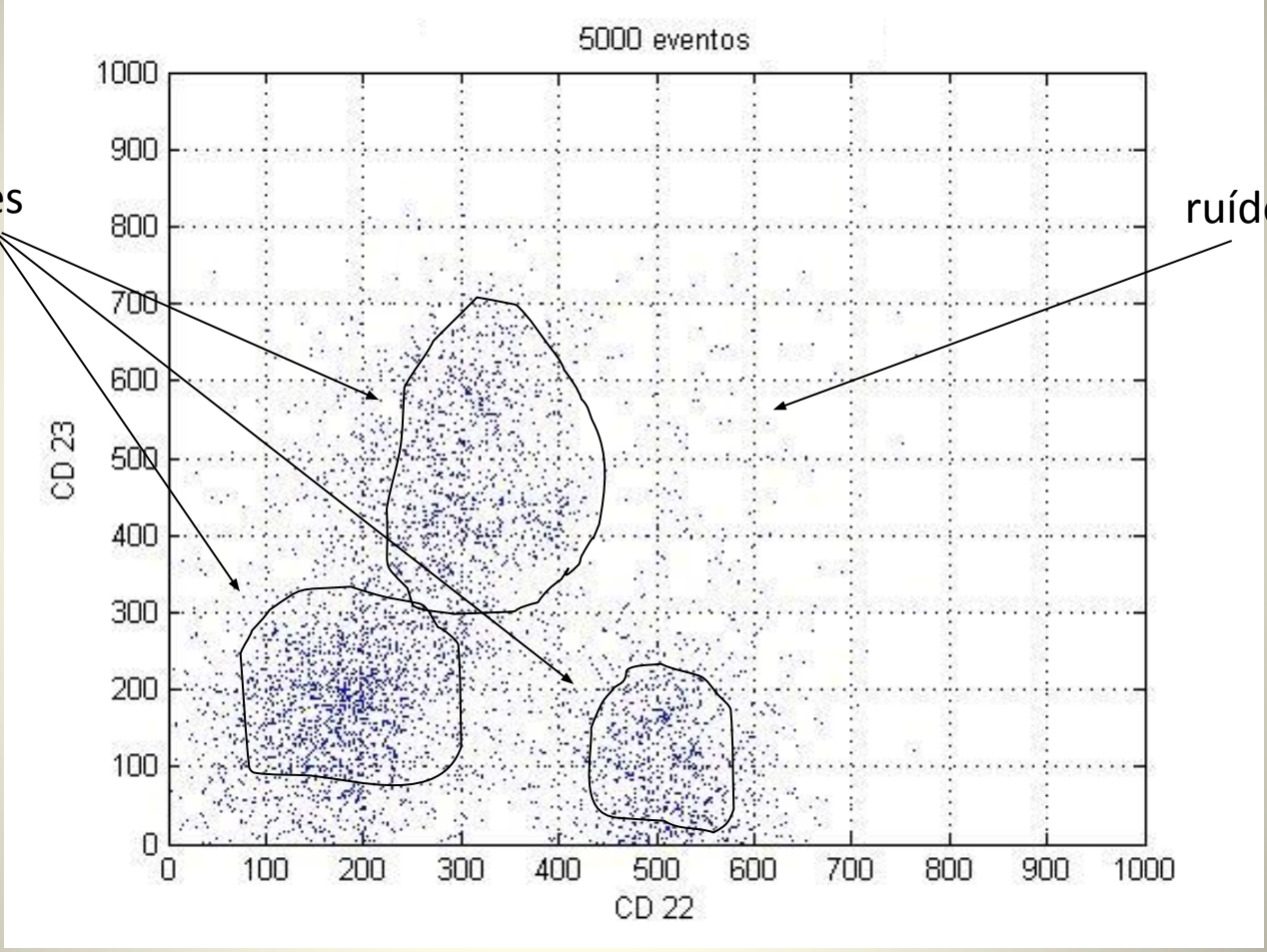
$$Y = B1 * \alpha_y + B2 * \beta_y + B3 * \delta_y + \ldots$$

Um problema de classificação



mas onde estão os grupos?

Populações celulares



ruído

A solução é trivial ?

- ✓ Podemos ter 30 ou mais dimensões (estavamos vendo apenas uma projeção em duas dimensões)
- ✓ Não sabemos quais destas dimensões são relevantes para separabilidade dos agrupamentos
- ✓ Podem existir agrupamentos com apenas 10 ou 20 observações em uma amostra de milhões

Temos um problema relevante?

Sim, temos um **problema complexo** onde a aplicação de **inteligência computacional** e outras ferramentas avançadas de **estatística**, se justificam.

As figuras acabamos de ver são dados gerados através de ***citometria de fluxo***

O que é citometria de fluxo?

Citometria de Fluxo

É a principal ferramenta na caracterização fenotípica de enfermidades infecciosas como a infecção pelo **HIV**, e de doenças neoplásicas - **leucemias**, **linfomas**, e **tumores sólidos** - ao diagnóstico e também durante o tratamento.

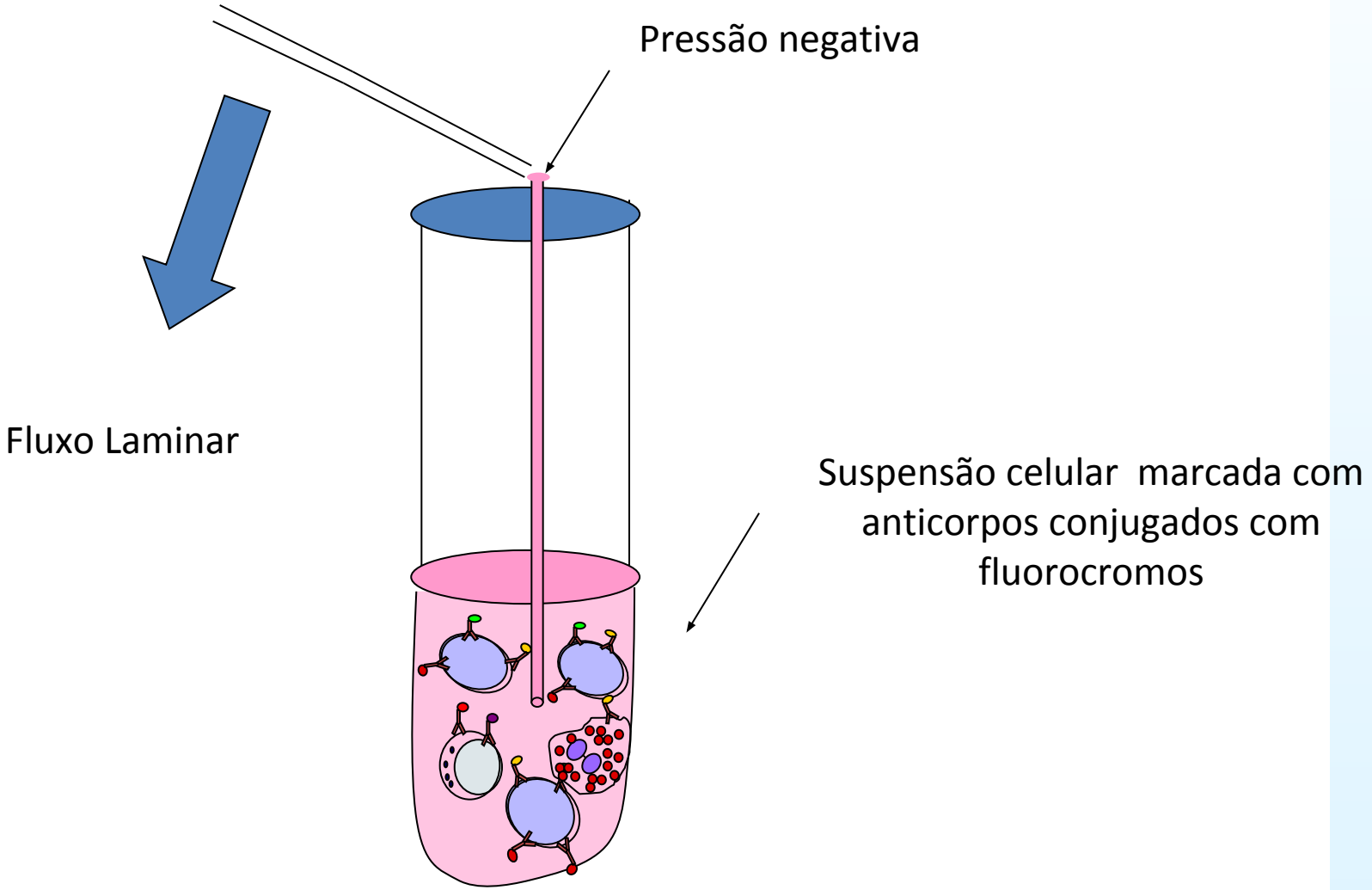
Esta caracterização tem fundamental importância prognóstica, dela depende decisivamente o tratamento aplicado ao paciente.

Citometria de Fluxo

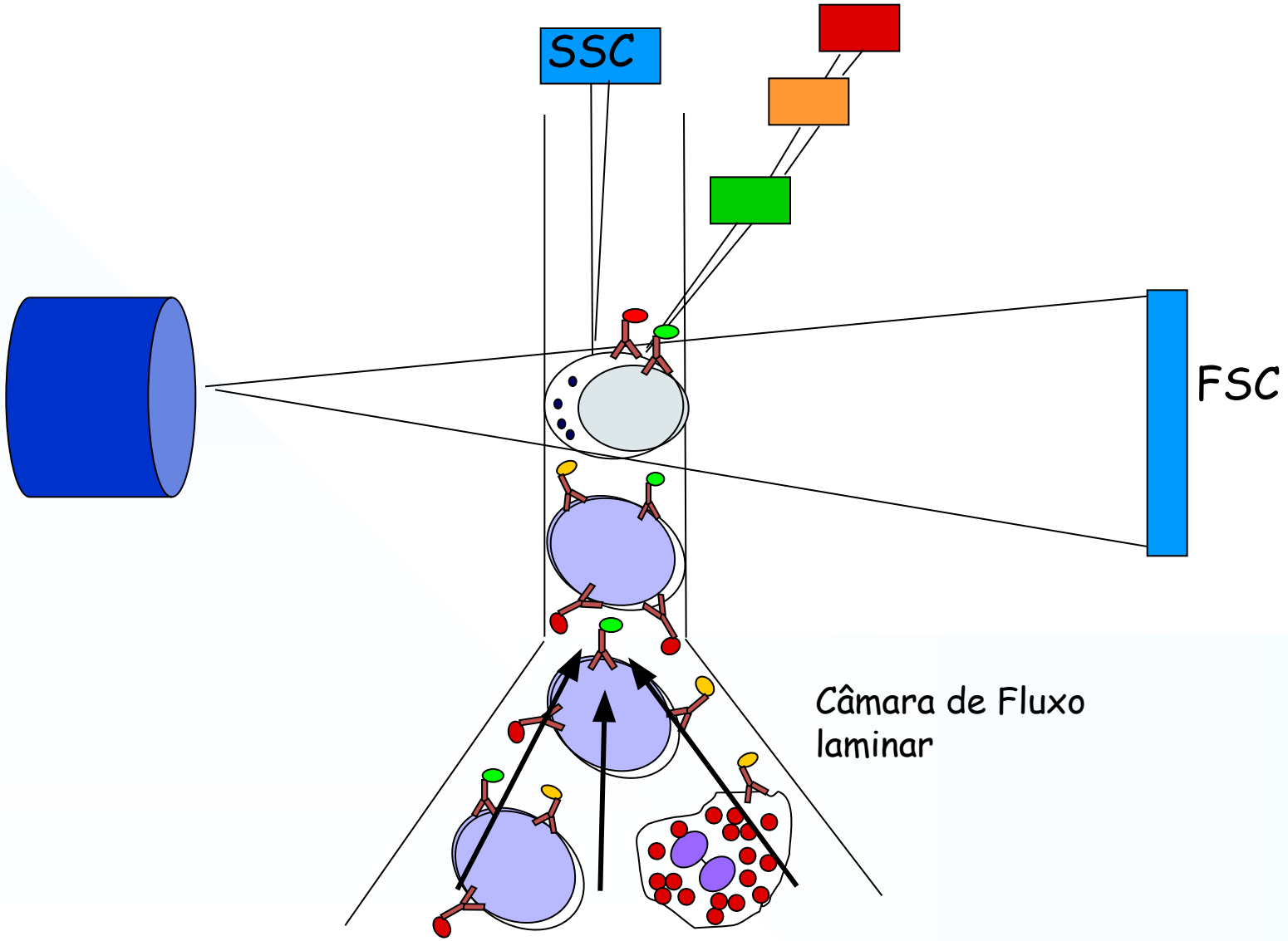
A citometria de fluxo multiparamétrica é capaz de medir simultaneamente **diversos parâmetros de milhares de células** por segundo.

É possível avaliar milhões de células em suspensão e obter diversas **informações individualizadas de cada célula**.

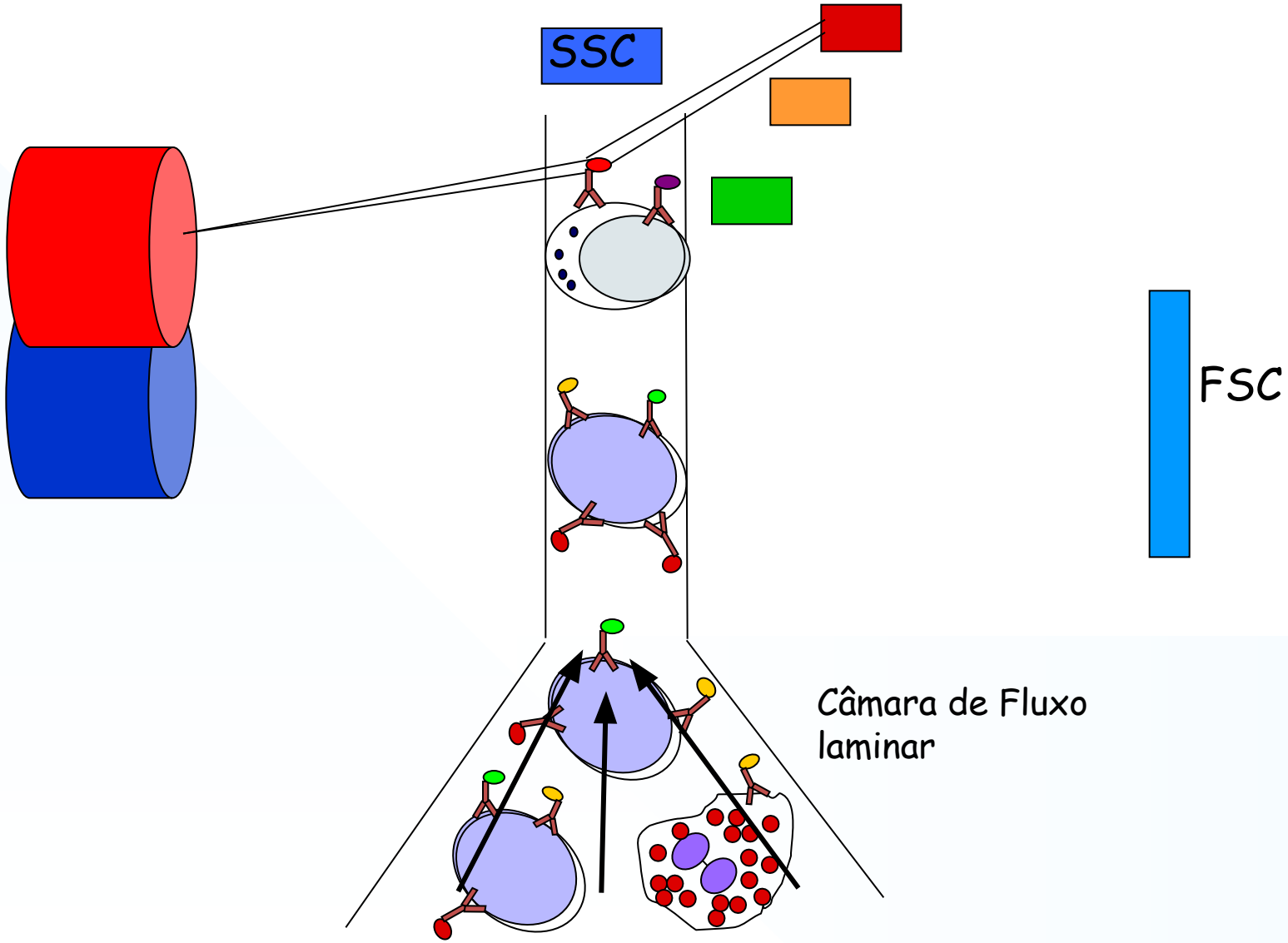
Citometria de Fluxo Multiparamétrica:



Citometria de Fluxo Multiparamétrica:

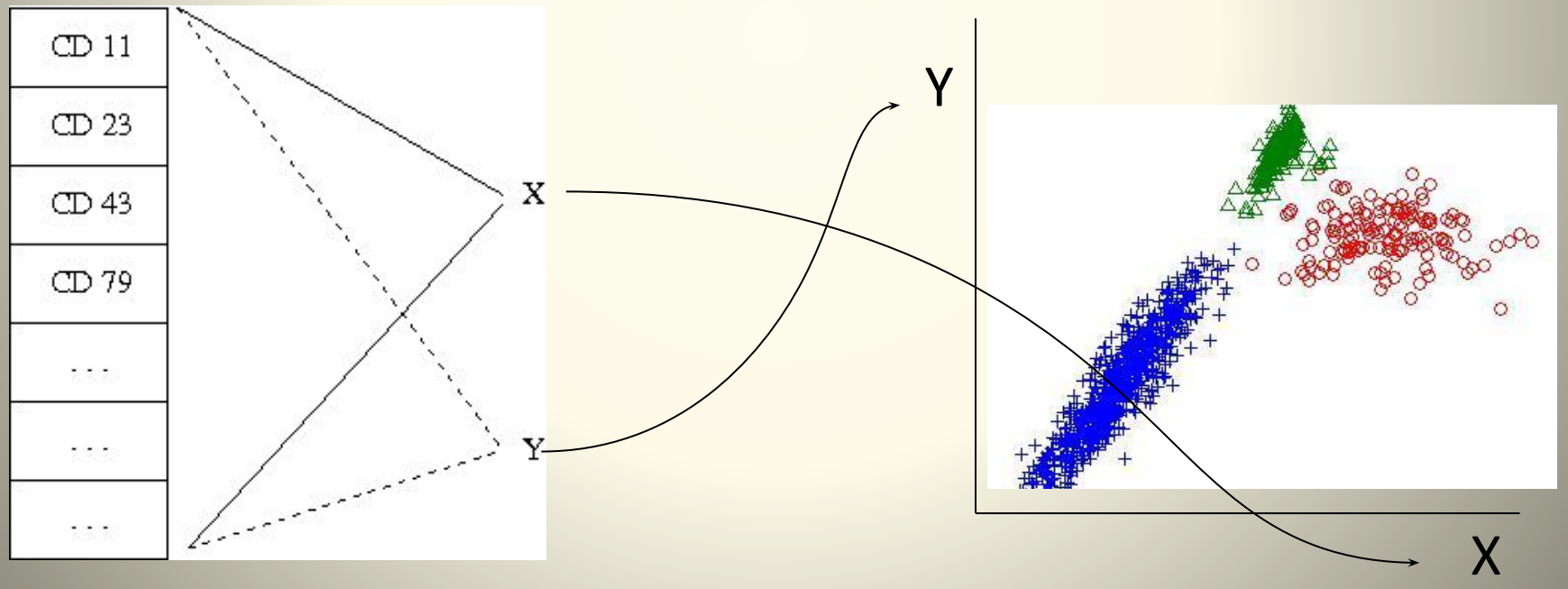


Citometria de Fluxo Multiparamétrica:



Classificação de doenças hematológicas

O objetivo é representar a informação (dos atributos relevantes) nos eixos x e y, de modo que **se possa ver a informação** em um só gráfico.



Um exemplo:

Para as doenças (linfomas) BL X FL, os pesos seriam 77, 15 & -7, resultando em:

$$X = 77*CD38 + 15*CD43 - 7*CD95$$

e

$$Y = -57*CD81 + 20*CD45 - 14*CD31 + 9*CD39$$

CD31 CD 43 etc são anticorpos monoclonais (nossos atributos)

O procedimento

1. **Selecionar os marcadores relevantes** para cada par de doenças

2. **Calcular os coeficientes**, x1, x2…. y1, y2…. Que geram a melhor separação

3. **Testar**, fora-da-amostra, com novos casos

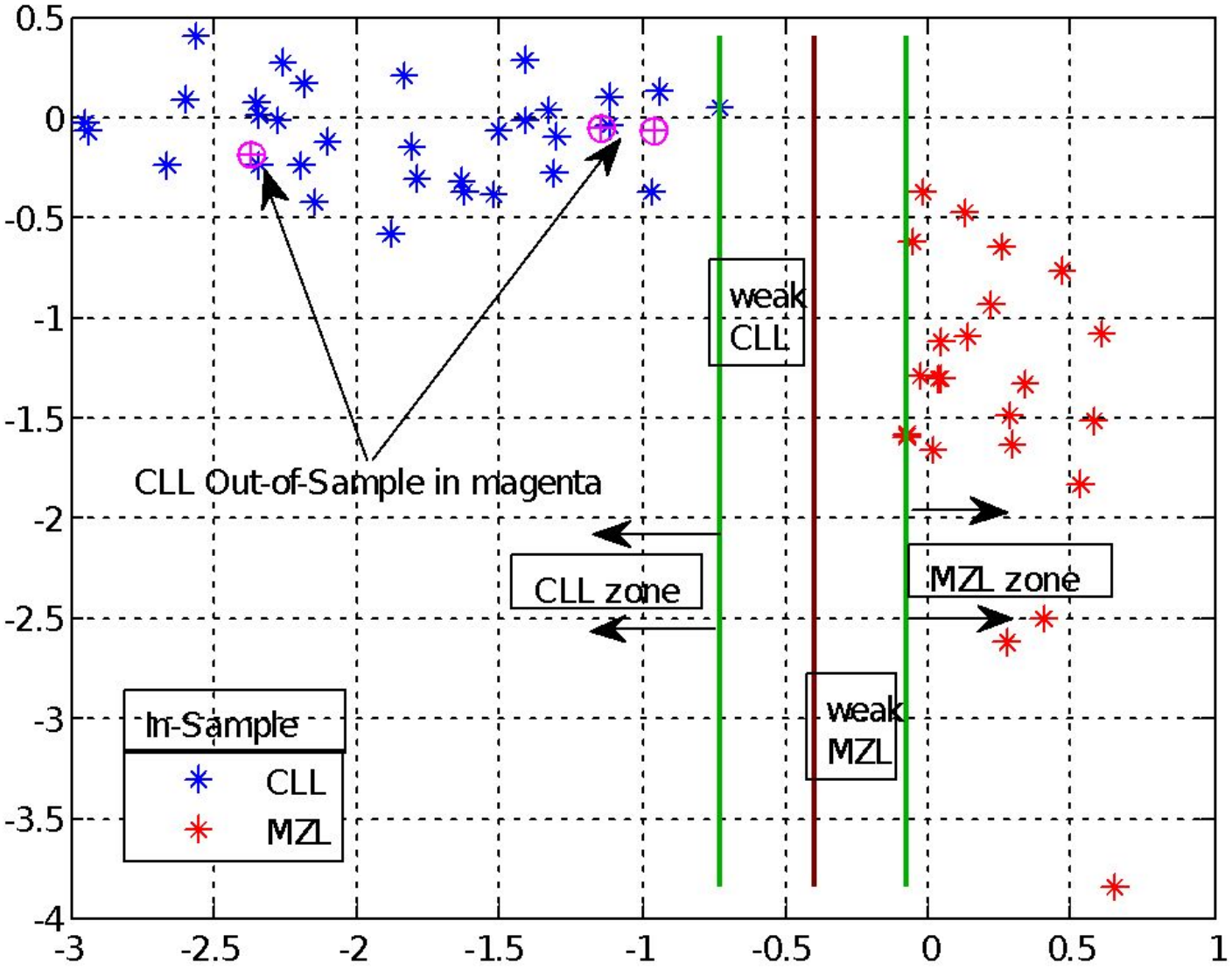
Selecionando os atributos relevantes

1. Selecionar um par doenças, e.g. BL X FL

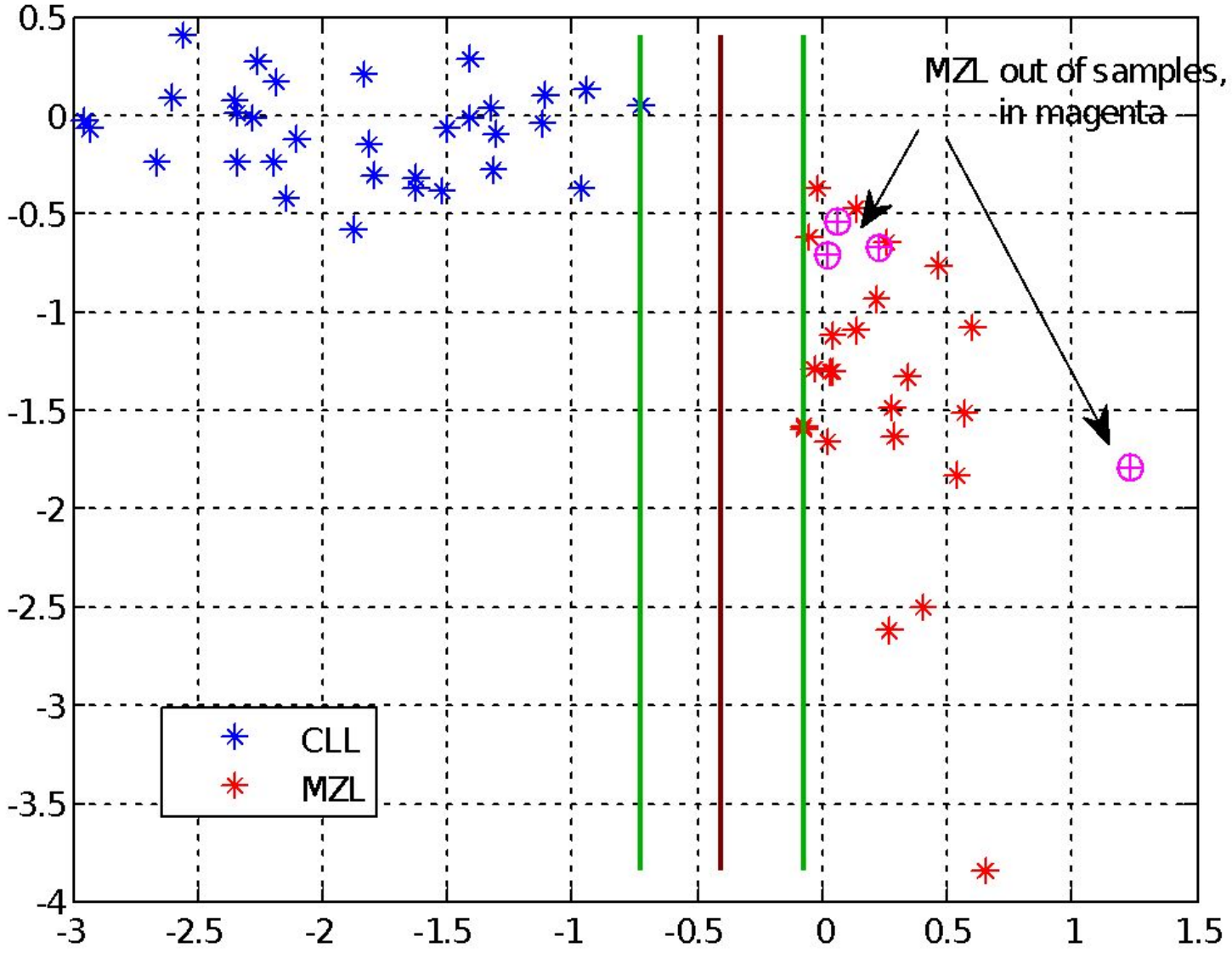
2. Rodar o algoritmo de otimização 20 vezes. Selecionar para o eixo-X os atributos que retêm > 5% dopeso total em pelo menos 50% das rodadas.

3. Excluir os atributos selecionados do 'pool' e re-rodar para o eixo-Y.

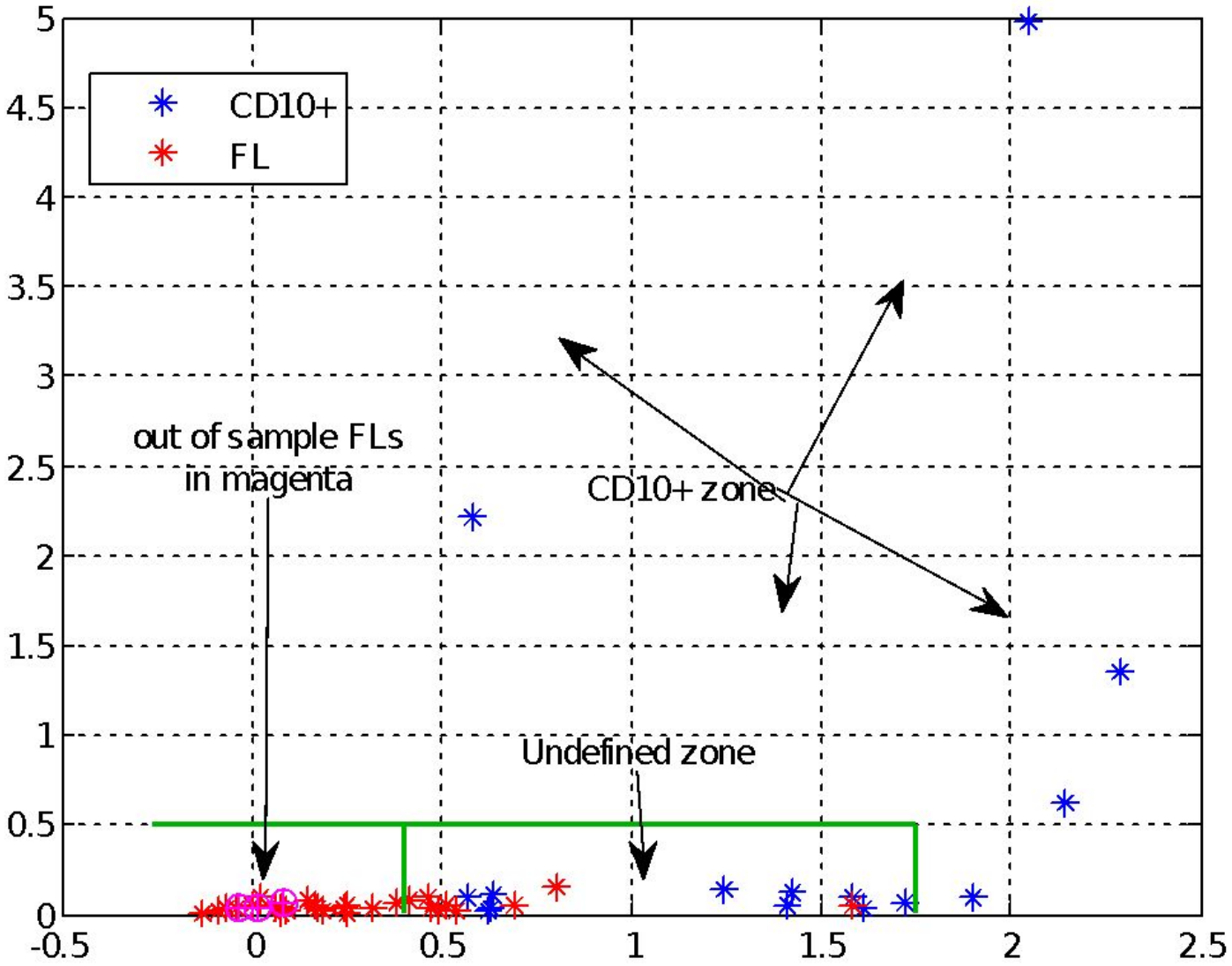
Out of Sample CLL X MZL



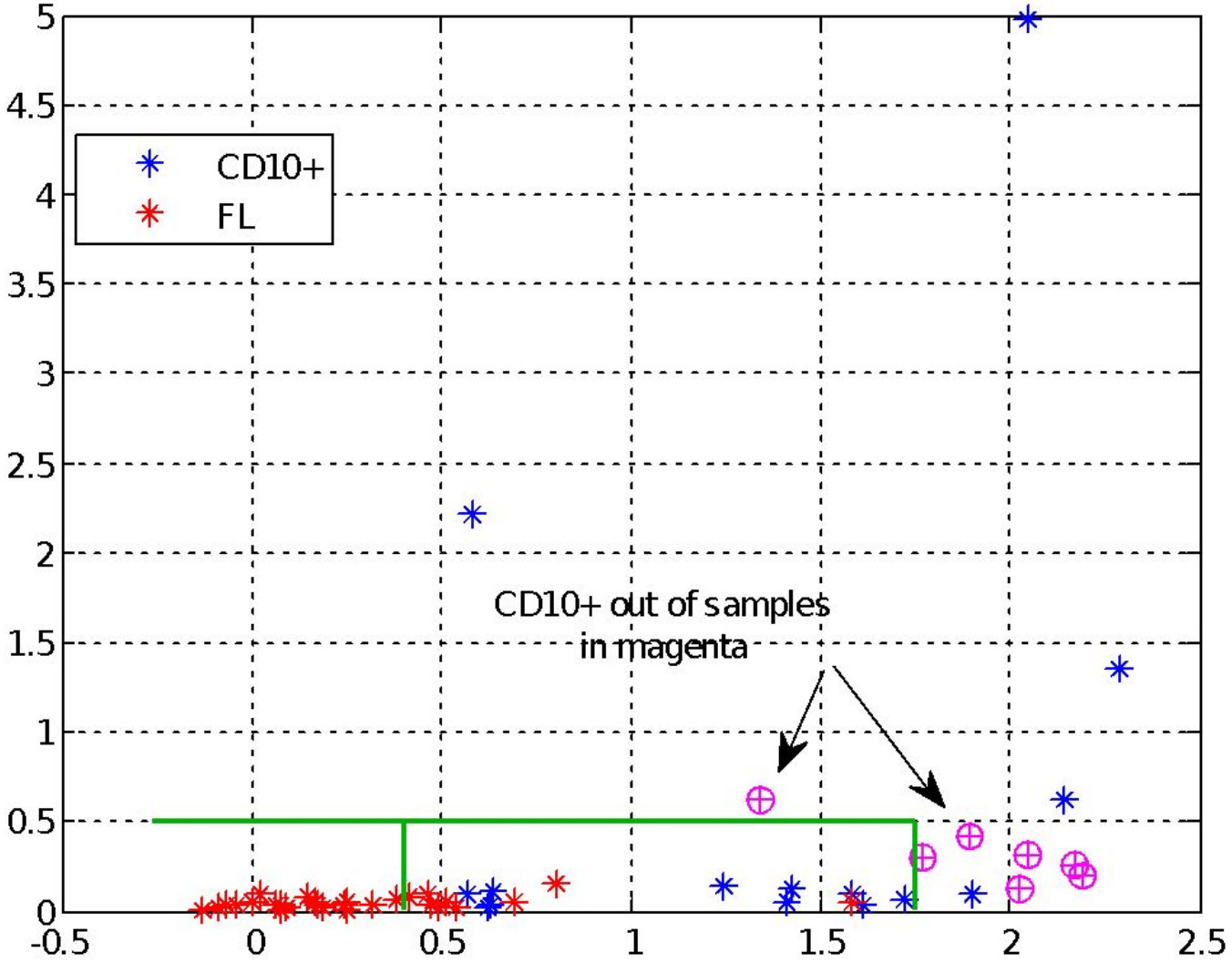
Out of Sample CLL X MZL



Out of Sample dificult case CD10+ X FL



Out of Sample dificult case CD10+ X FL



Alguns artigos e patentes relacionados:

- Pedreira CE; “Automating flow cytometry”. **Cytometry A** , v. 81A, p.110-111, (2012).
- Ayuso MM; Costa ES Pedreira CE; et al. “EuroFlow strategies and tools for data analysis. In: EuroFlow standardization of flow cytometer instrument settings and immunophenotyping protocols”. (in press, on line published), **Leukemia**, (2012).
- Peres RT, Aranha CC, and Pedreira CE, Optimized Bi-Dimensional Data Projection For Clustering Visualization” **Information Sciences (to appear)**
- Costa ES; Pedreira CE; Flores J; Lecrevisse Q; Quijano S; Barrena S; Almeida, J; Böttcher S; Van Dongen JJM; Orfao A; on behalf of EuroFlow Consortium . “*Automated Pattern-Guided Principal Component Analysis versus Expert-Based Immunophenotypic Classification of Hematological Malignancies*” **Leukemia,** 24(11):1927-33, (2010).
- Peres RT e Pedreira CE; “A New Local-Global Approach for Classification. **Neural Networks** , v. 23, p. 887-891, (2010).

- **Patente nos Estados Unidos da América nº US 7,507,548B2.** “Multidimensional detection of aberrant phenotypes to be used to monitor minimal disease levels using flow cytometry measurements”. Inventores: Alberto Orfao de Matos**,** Carlos Eduardo Pedreira e Elaine Sobral da Costa. (2009). Licença cedida a Cytognos SL .
- **Patente nos Estados Unidos da América nº US 7,321,843B2** “Method for generating flow cytometry data files containing an infinite number of dimensions based on data estimation” (2008). Inventores: Alberto Orfao de Matos, Carlos Eduardo Pedreira e Elaine Sobral da Costa. Licença cedida a Cytognos SL.

Apoio:

CNPq

FAPERJ

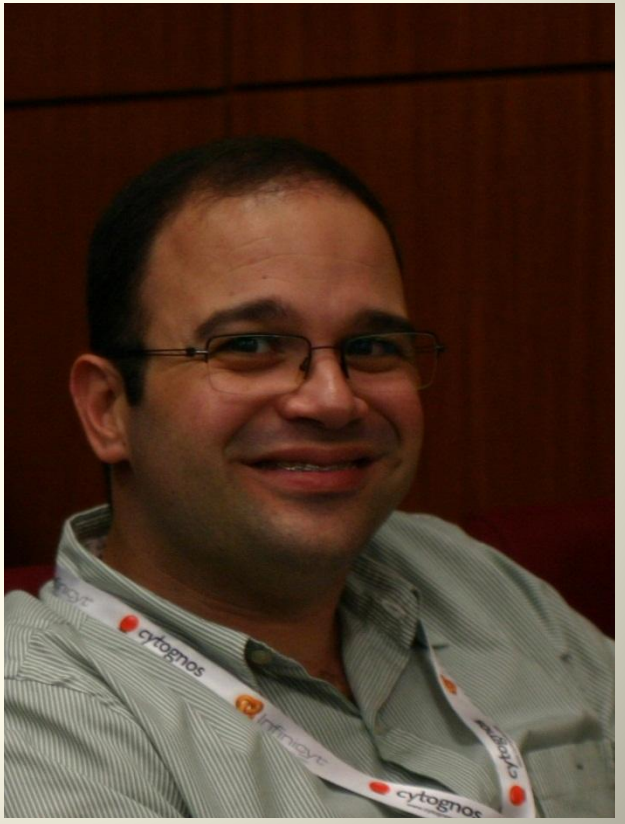
CAPES

Meus Colaboradores:

Claus Aranha - Universidade de Tsukuba (Japão)

Rodrigo Peres

CEFET-RJ / UERJ



Alberto Orfao

Centro de Investigação do Câncer da
Universidade de Salamanca - Espanha

Elaine Sobral da Costa
IPPMG e Dept. Clinica Médica UFRJ

Quentin Lécrevisse

Universidade de Salamanca - Espanha

Julia Almeida

Centro de Investigação do Câncer da ***Universidade de Salamanca - Espanha***

Obrigado!

sites.google.com/site/pedreira56

pedreira@ufrj.br