Big Bio Data: Inteligência Computacional a Serviço da Medicina

Carlos Eduardo Pedreira

COPPE-UFRJ

PESC - Programa de Sistemas e Computação

pedreira@ufrj.br carlosp@centroin.com.br

www.cos.ufrj.br/~pedreira

Agosto de 2014

Patentes

Localizar estado da técnica | Discutir esta patente

Apparatus and method for detecting cancer in tissue

US 3789832 A

RESUMO

An apparatus and method in which a tissue sample is positioned in a nuclear induction apparatus whereby selected nuclei are energized from their equilibrium states to higher energy states through nuclear magnetic resonance. By measuring the spin-lattice relaxation time and the spin-spin relaxation time as the energized nuclei return to their equilibrium states, and then comparing these relaxation times with their respective values for known normal and malignant tissue, an indication of the presence and degree of malignancy of cancerous tissue can be obtained.

Número da publicação	US3789832 A
Tipo de publicação	Concessão
Data de publicação	5 fev. 1974
Data de depósito	17 mar. 1972
Data da prioridade	17 mar. 1972
Também publicado como	CA1004297A1
Inventores	Damadian R
Cessionário original	Damadian R
Exportar citação	BiBTeX, EndNote, RefMan

Citações de patente (3), Citações de não patente (1), Citada por (70), Classificações (9), Eventos legais (2)

Links externos: USPTO, Cessão do USPTO, Espacenet

Em **1977** realiza-se o **primeiro exame de ressonância magnética em humanos**. São necessárias 5 horas para gerar a imagem.

O **primeiro aparelho comercial** é produzido em **1980**.

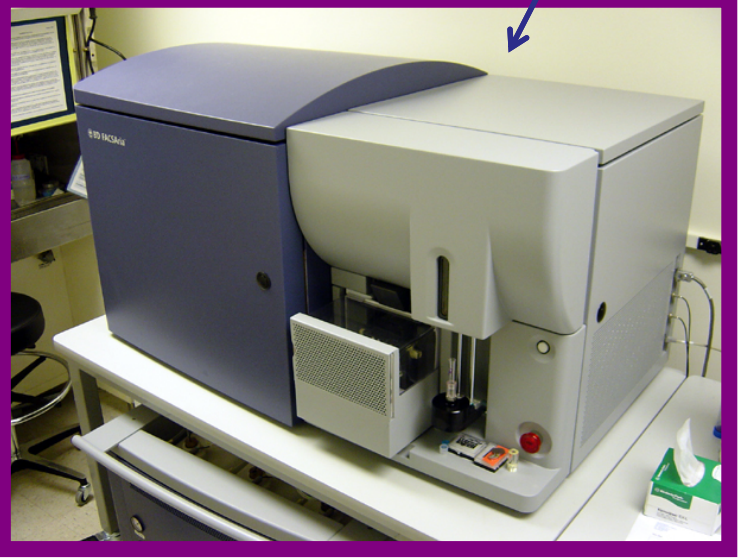
No início da década de 70, a empresa Becton Dickinson Immunocytometry Systems colocou no mercado os primeiros citômetros

Diagnósticos de Leucemias e Linfomas

Atualmente:

3 a 4 detectores de fluorescência

8 detectores de fluorescência



Criam aparelhos

Engenheiros, físicos, computeiros etc

DADOS que precisam ser
procesados de modo inteligente

Médicos e ‘outros Bios’

Engenheiros, físicos, computeiros etc

Uma ENORME quantidade de dados passa a ser rotineiramente gerada, abrindo novas pespectivas e a necessidade de procesar estes dados de forma INTELIGENTE para obter a informação desejada.

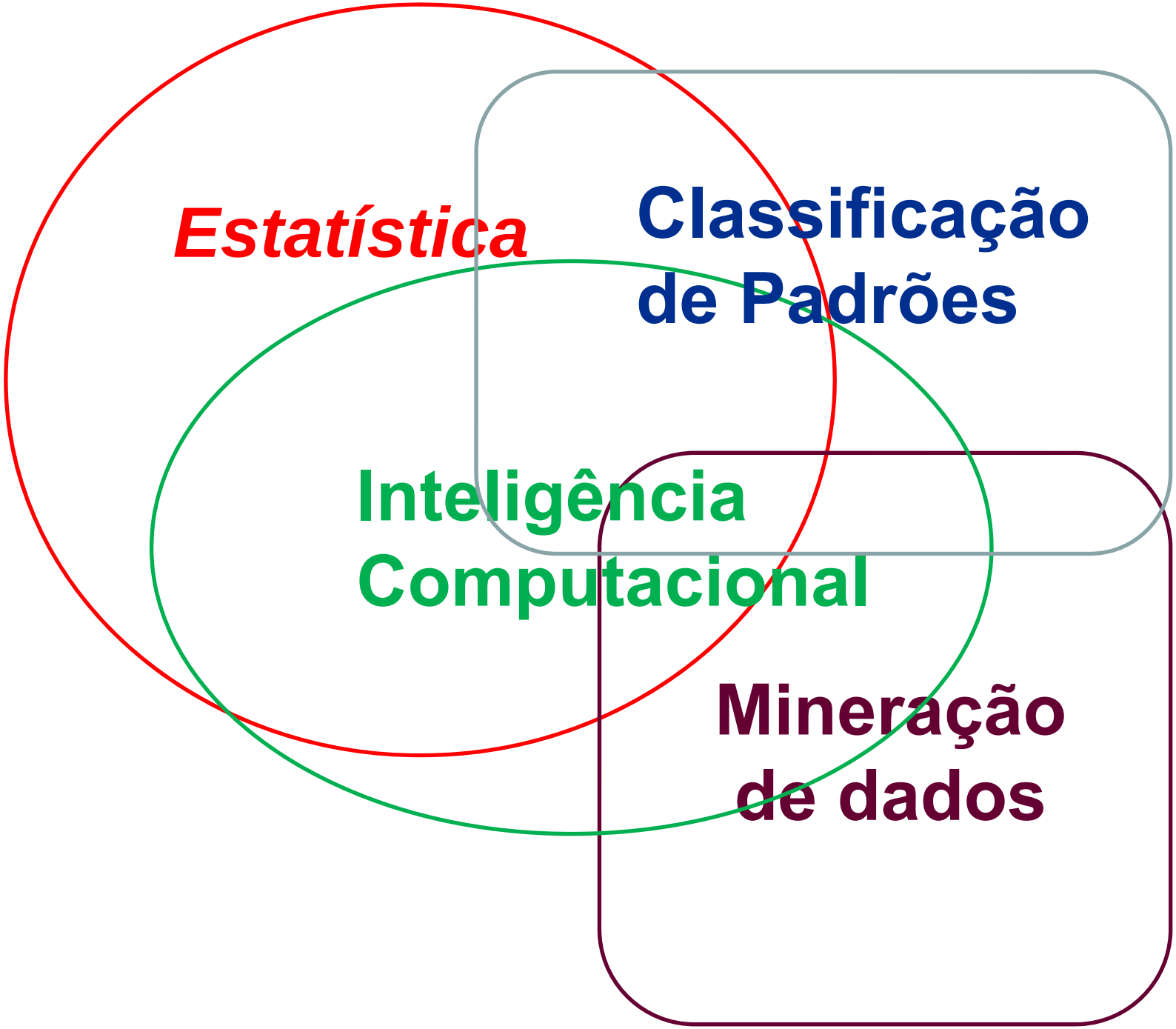
Inteligência Computacional

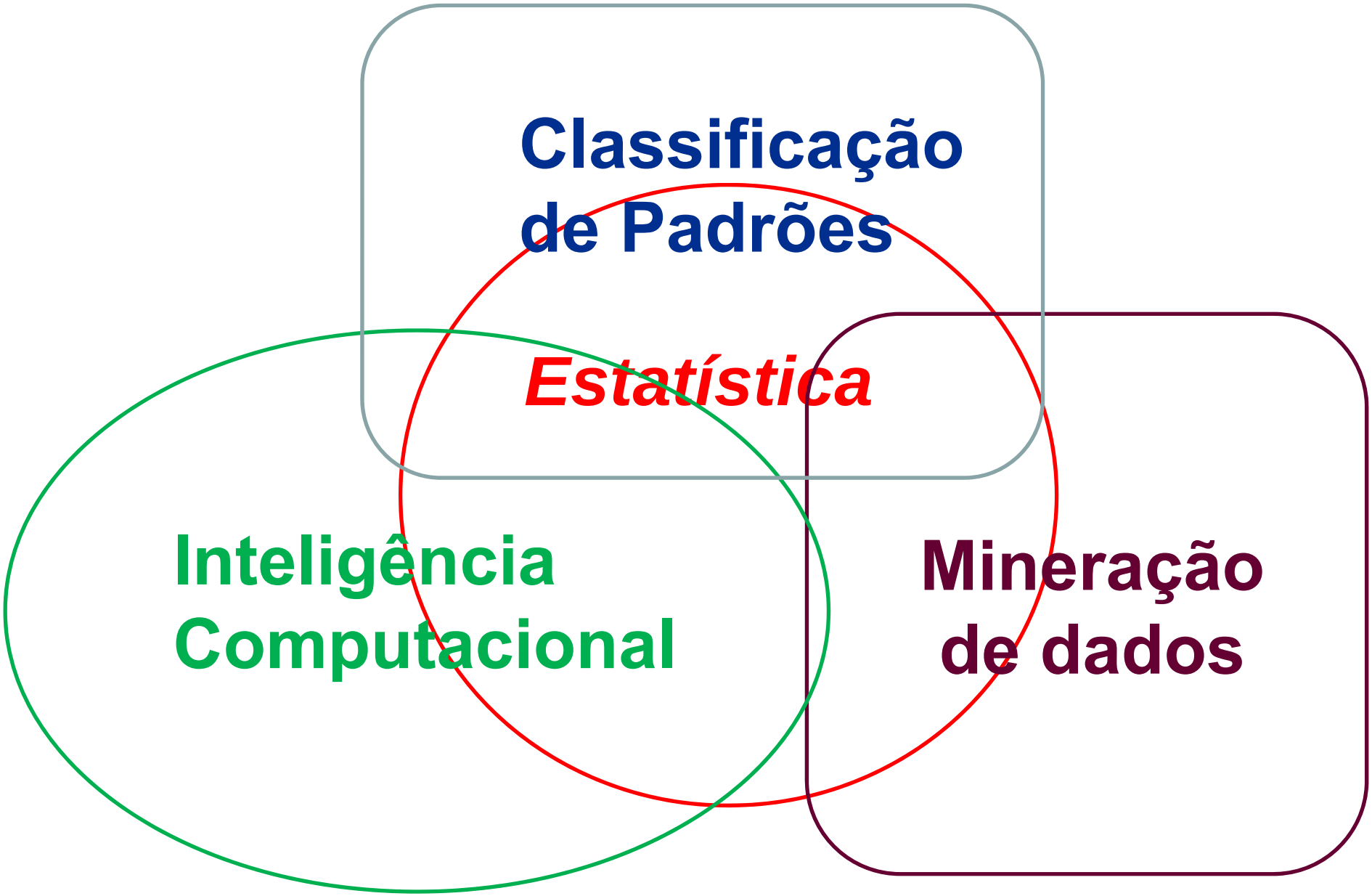
Objetivo:

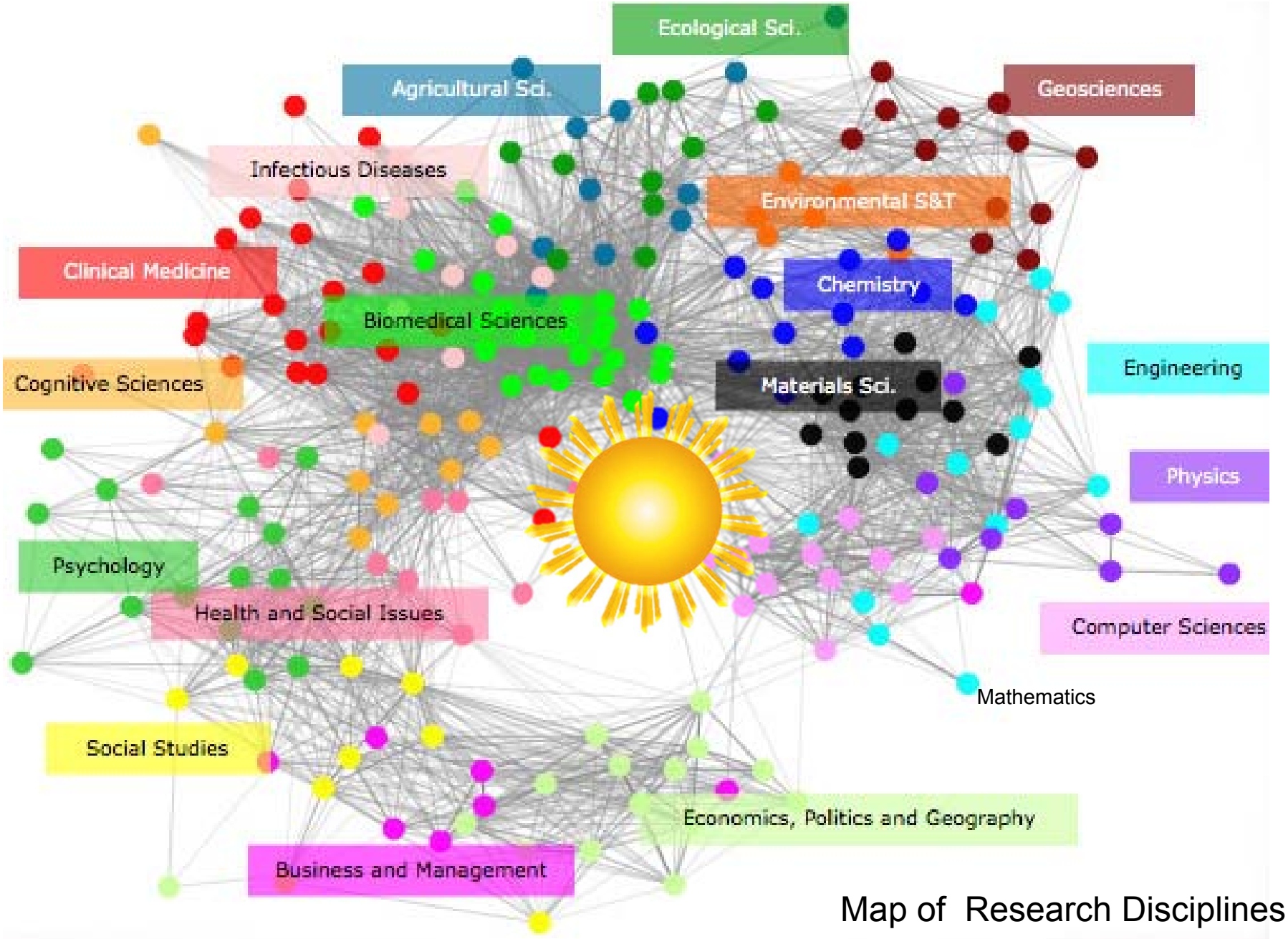
Usar um conjunto de observações para inferir informação (desfecho) sobre uma população.

Pontos essenciais:

- Existe a informação (ou padrão) a ser descoberta
- Não há como obter a informação diretamente por um método matemático.
- Há dados.


Estatística
Classificação
de Padrões
Inteligência
Computacional
Mineração
de dados


Classificação
de Padrões
Estatística
Inteligência
Computacional
Mineração
de dados



Map of Research Disciplines

http://idr.gatech.edu/

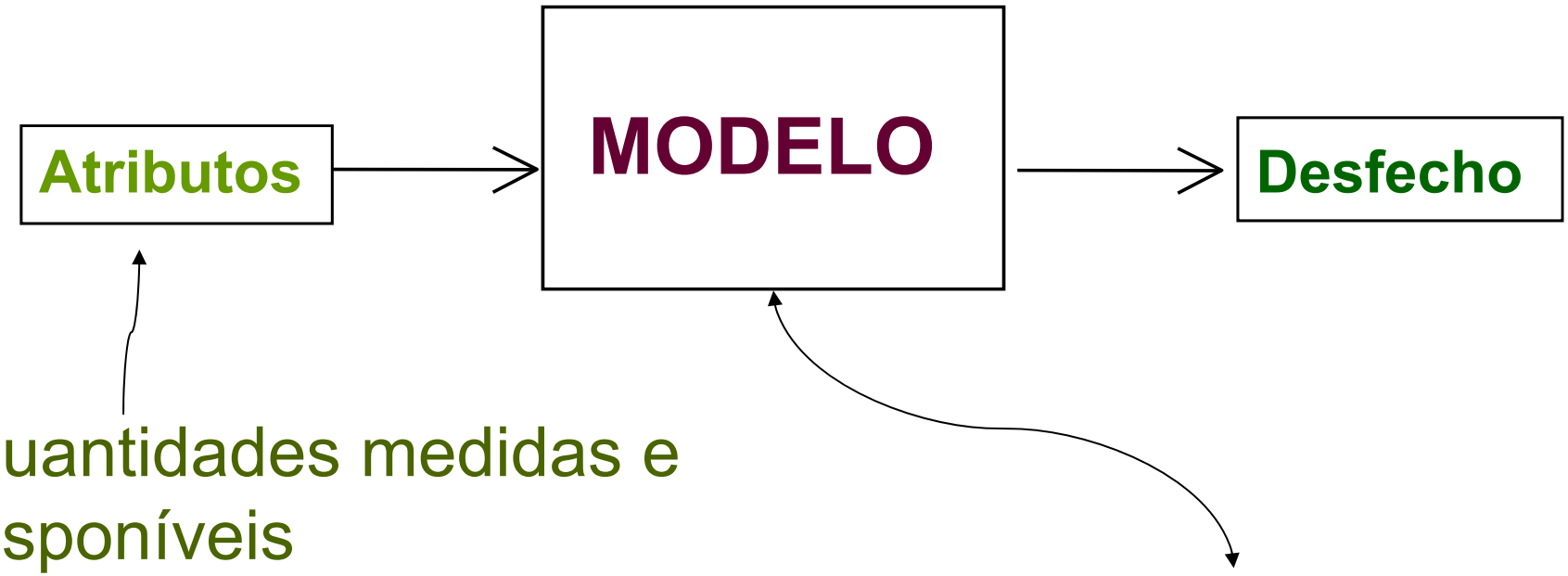
Inteligência Computacional

Existem muitos modelos :

- Redes Neurais
- Máquinas de Vetor de Suporte (SVM)
- Aprendizado por reforço
- Etc etc ….

Não se deve **‘começar’ pela escolha do modelo** mas sim pelas necessidades da aplicação.

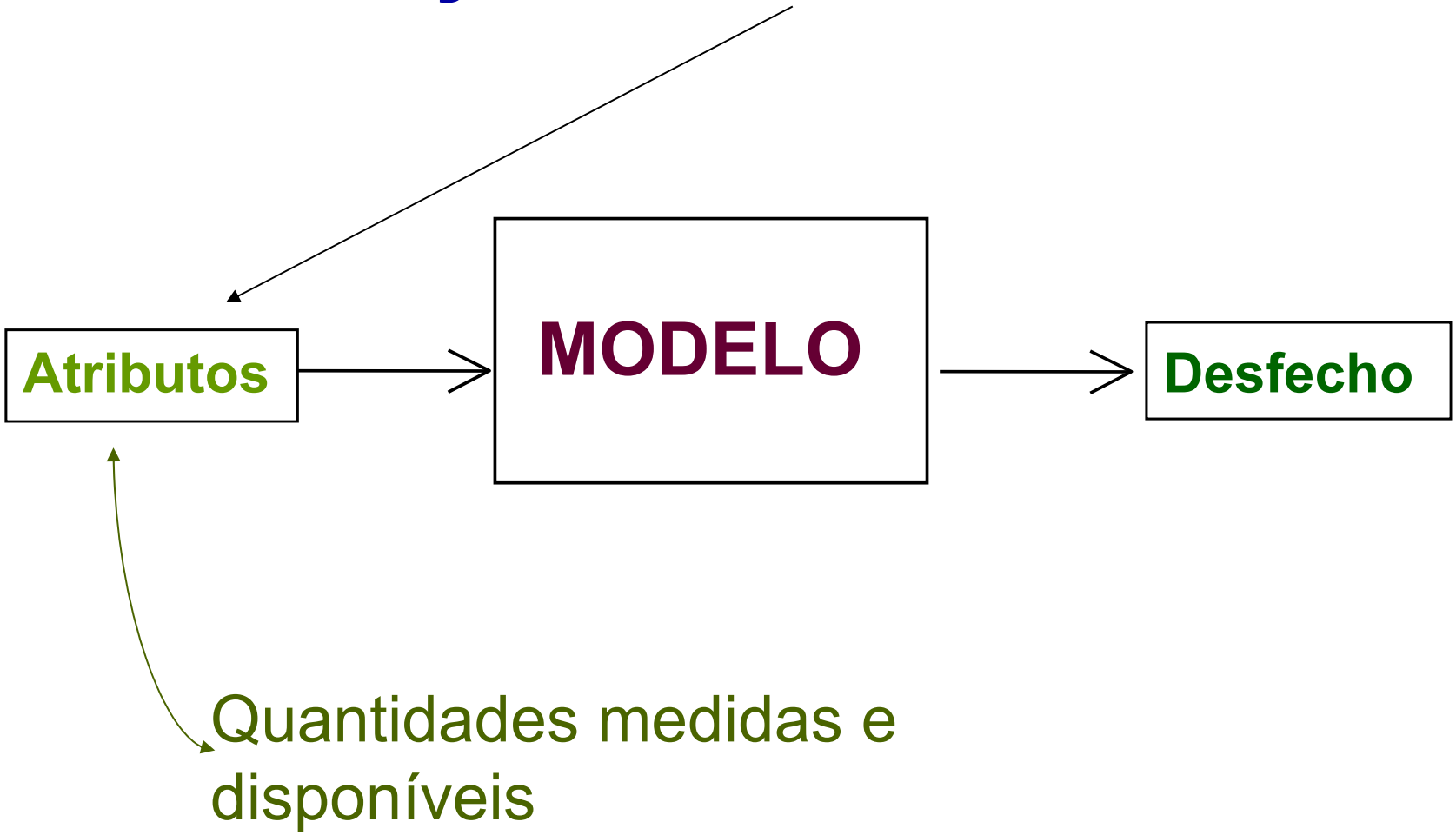
Normalmente, o sucesso na aplicação não está relacionado a escolha do modelo mas ao seu uso apropriado.



Quantidades medidas e disponíveis

Os parâmetros do modelo precisam ser ajustados

Seleção de Atributos



Porque é preciso selecionar atributos ?

- Por que queremos usar os atributos que tem **maior capacidade em explicar o desfecho**
- Por que muitas vezes temos muitos atributos e poucas observações, é preciso **construir modelos com parcimônia** (quanto mais parâmetros mais sofisticado é o modelo.


Qual é o meu?

Alguns dos grandes desafios dos próximos anos

- Proteômica
- Apoio a decisão em diagnósticos
- Estimação de risco
- Construção de modelos explicativos
- Jogos para reabilitação
- …. Muitas outras

Apoio a decisão: um caminho é projetar em 2D?

Porque:

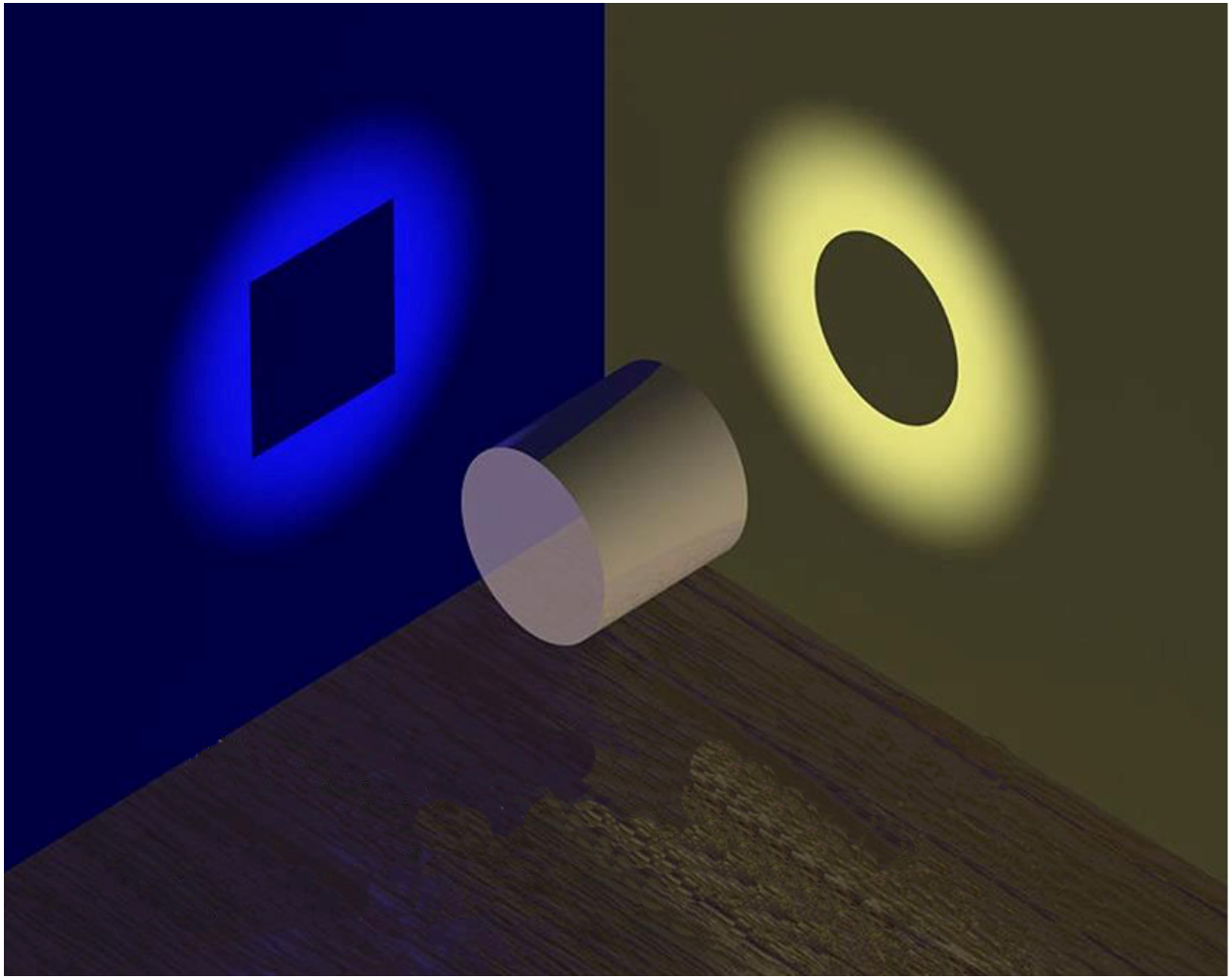
Frequentemente, é interessante ter uma ferramenta de suporte a decisão para auxiliar na tarefa de classificação ao invés de um algoritmo para classificação automática. **A decisão final deve ser tomada pelo usuário e não pelo 'sistema'**.

Quando:

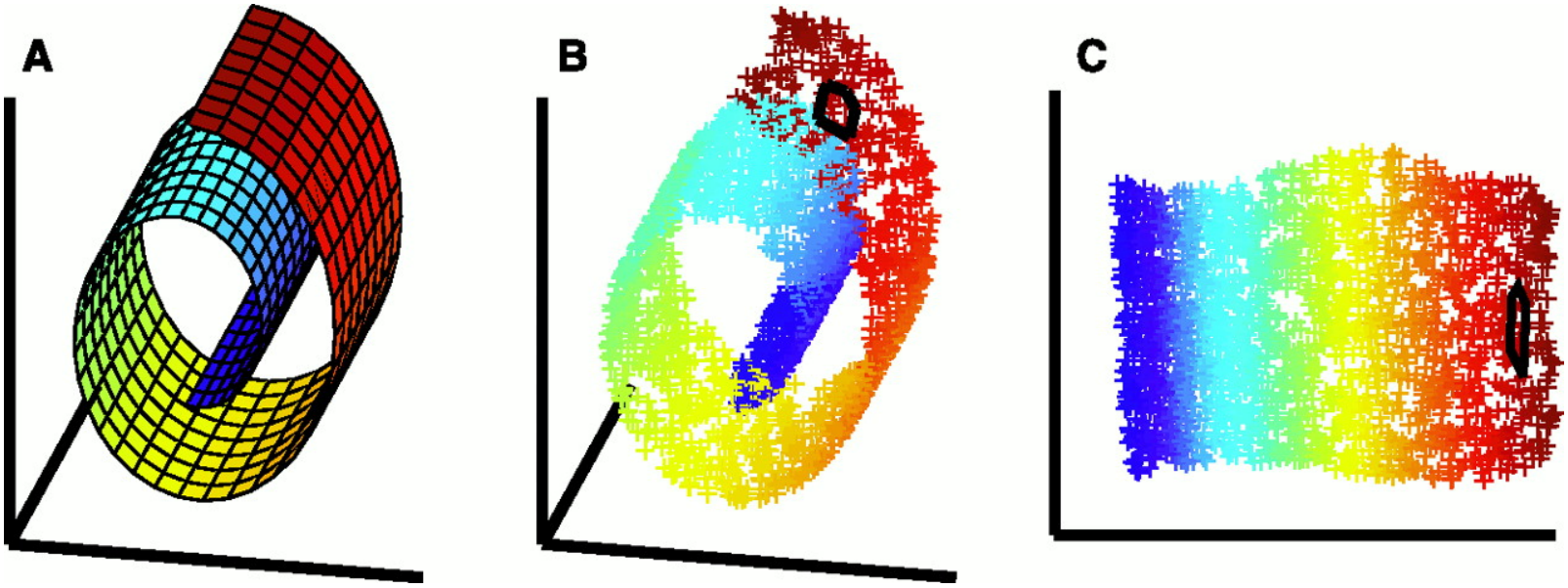
- Não se quer classificar automaticamente por **razões éticas ou legais** e.g. diagnósticos médicos.
- Existe **informação adicional** difícil de ser modelada mas relevante de ser incluída.

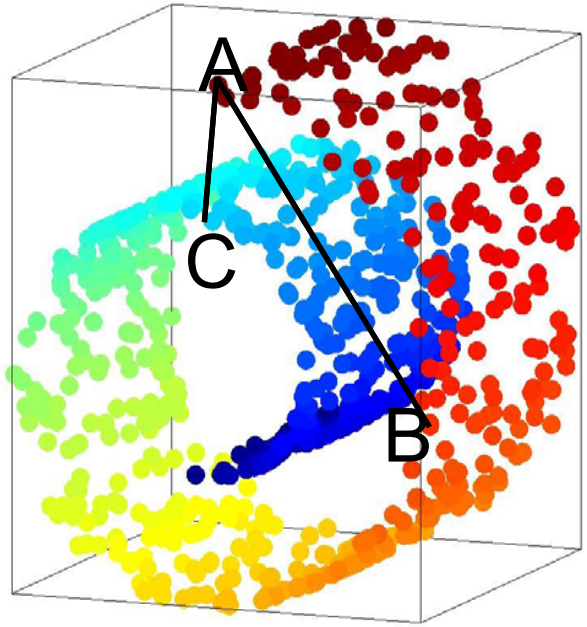
Projetando em 2-D

The way one projects = The way one sees



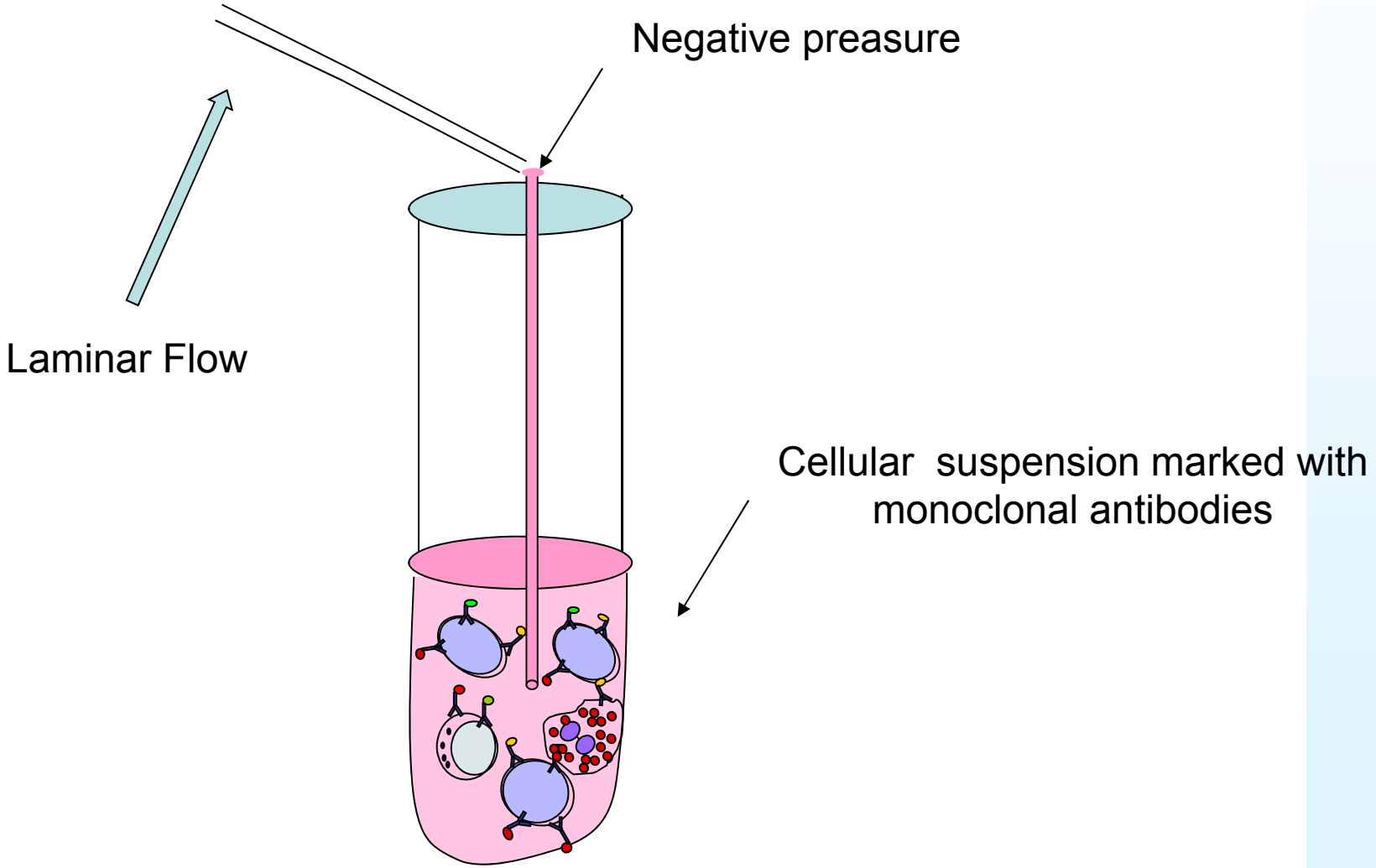
Desenrolando o rocambole



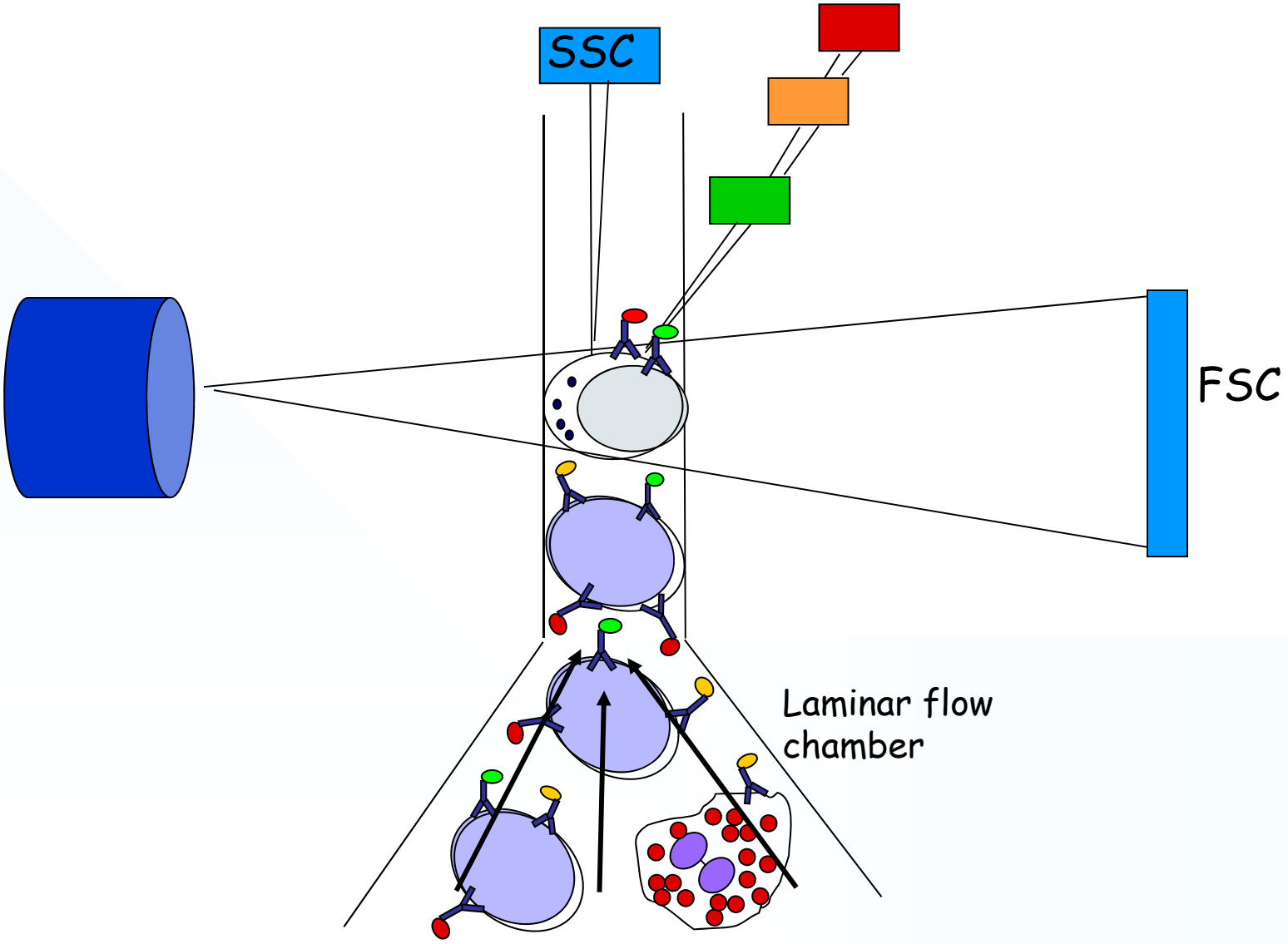


Let´s go BIO

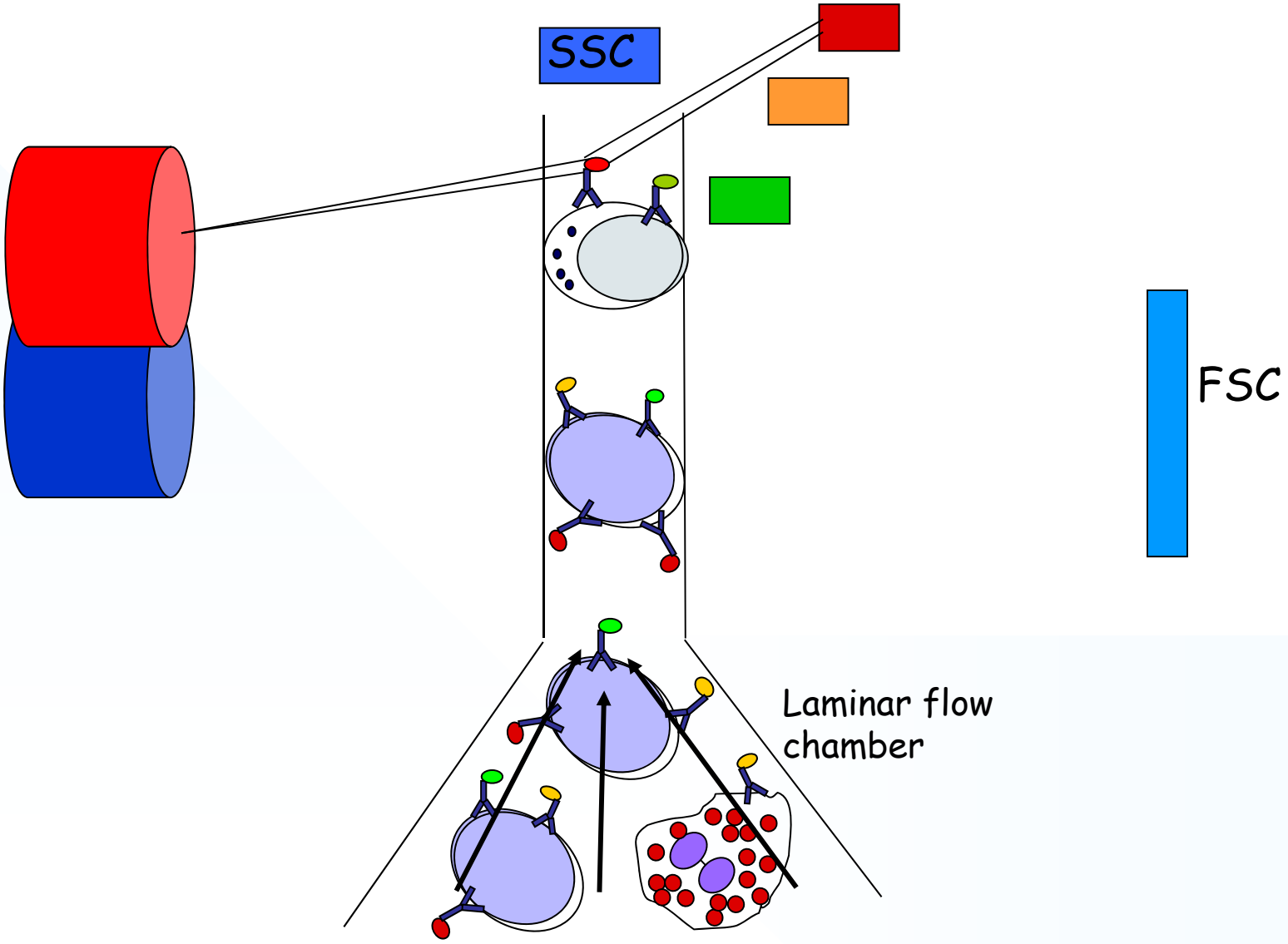
Análise de dados de citometria de fluxo

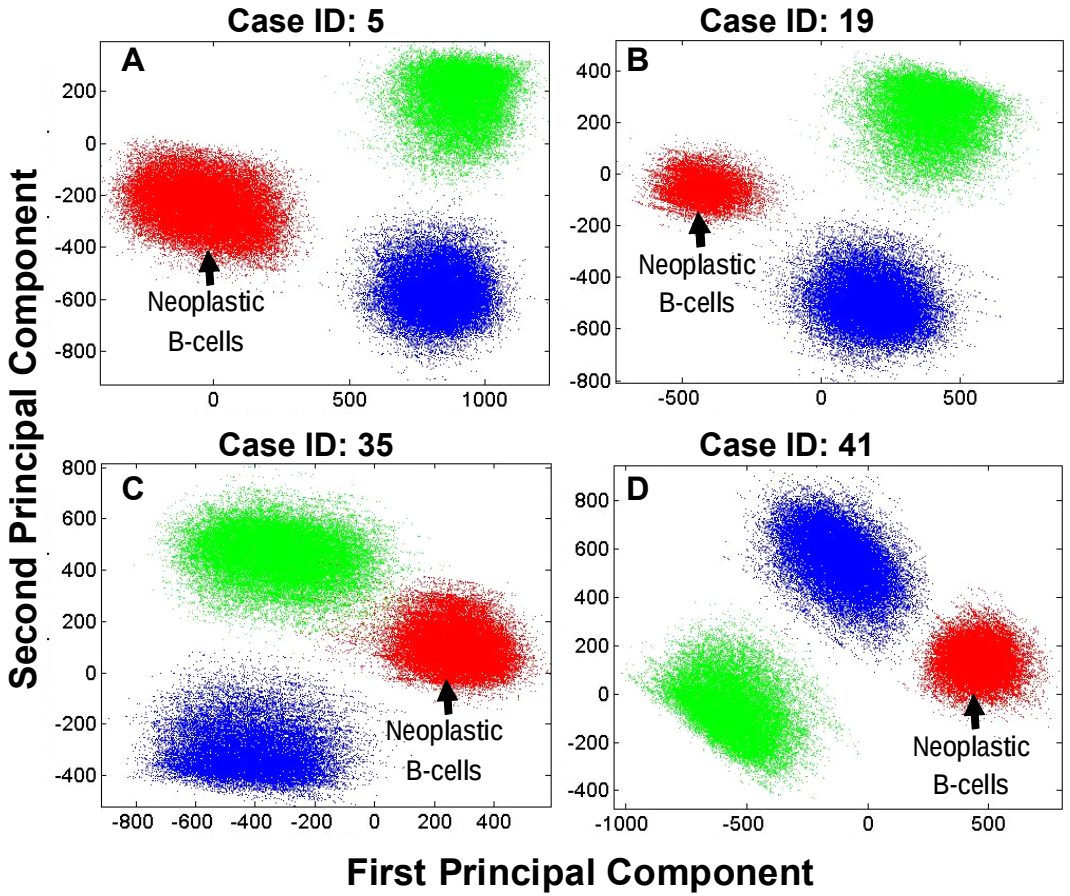


Multiparametric Flow Cytometry:



Multiparametric Flow Cytometry:




Case ID: 5
A
Neoplastic
B-cells
Case ID: 19
B
Neoplastic
B-cells
Case ID: 35
C
Neoplastic
B-cells
Case ID: 41
D
Neoplastic
B-cells
Second Principal Component
First Principal Component

Computação em Citometria: Desafios

- **Estimação de Doença Residual Mínima**
- **Diagóstico diferencial de linfomas**
- **Desenho de novos paineis**
- **…..muitas outras possibilidades**

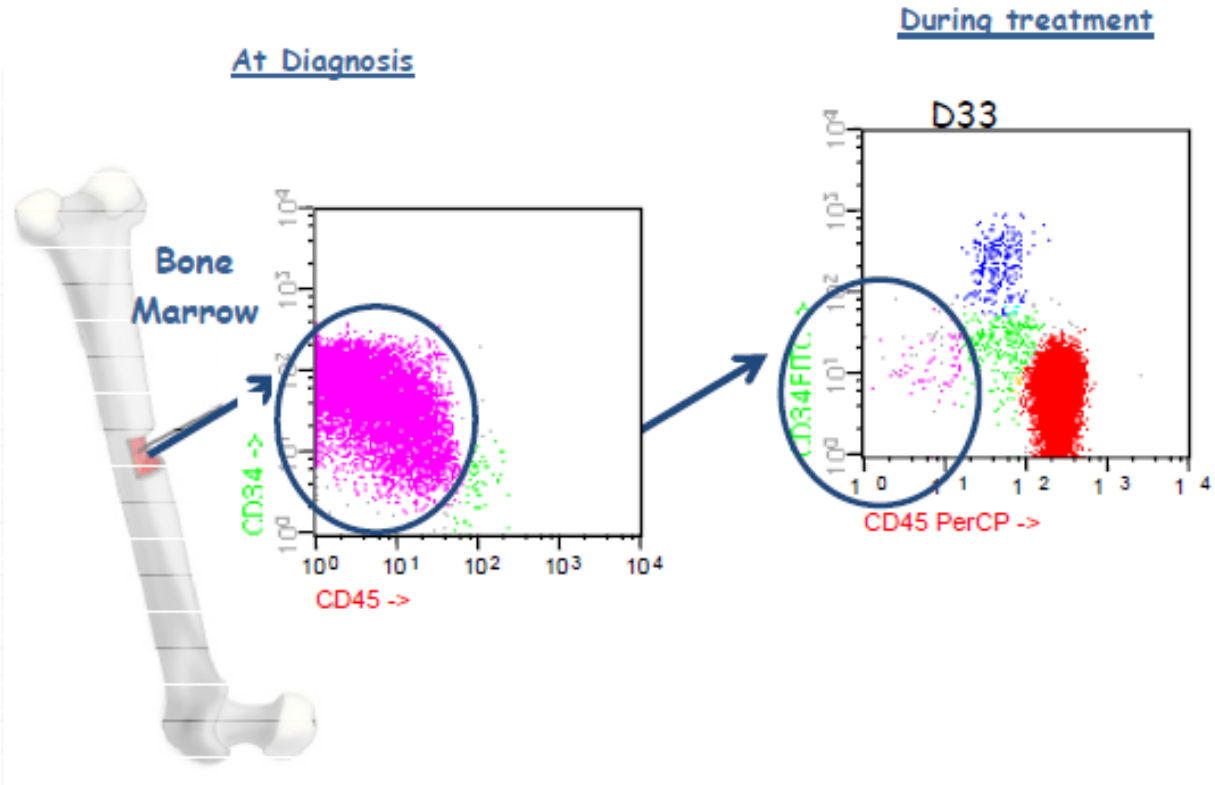
- Pedreira, C.E. ; Costa,E.S; Lecrevisse Q.; van Dongen J.J.M.; Orfao A. "*Overview of Clinical Flow Cytometry Data Analysis: Recent Advances and Future Challenges*" **Trends in Biotechnology**, v.31 n.7, p.415-427, (2013).

Uma estratégia probabilística para detecção de Doença Residual Mínima (DRM)

A idéia central é atribuir uma **probabilidade** A CADA CÉLULA de pertencer a população **normal** ou à **patológica**.

Pedreira CE, Costa ES, Almeida J, Fernandez C, Quijano S, Flores J, Barrena S, Lecrevisse Q, van Dongen JJ, Orfao A; "*A Probabilistic Approach For The Evaluation Of Minimal Residual Disease By Multiparameter Flow Cytometry In Leukemic B-Cell Chronic Lymphoproliferative Disorders*" **Cytometry A**, (2008) 12; pp 1141- 1150 .

Doença Residual Mínima



- DRM é um **fator prognóstico em diversas doenças hematológicas.** É um critério para **mudanças no protocolo de tratamento.**

A Probabilistic Approach to Identify MRD

Step 1:

Build artificial **"diagnostic-files"** for each patient (50 patients), by **mixing** events corresponding to **neoplastic B-cells** from the patient, with events corresponding **to normal B-cells** from the "normal-B-cell-pool file" at a 1:1 proportion.

Step 2:

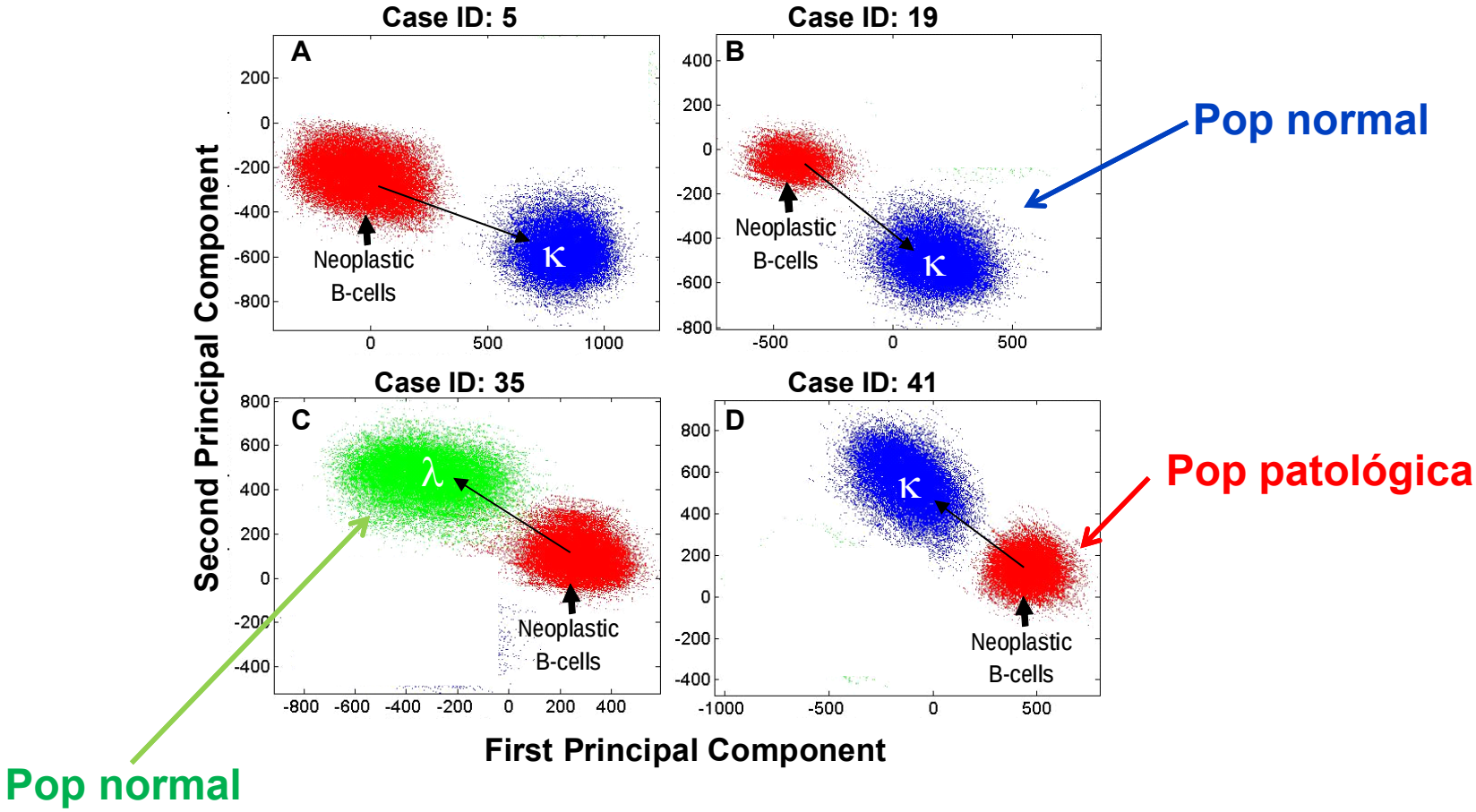
Apply **Principal Component Analysis (PCA)** to each of these artificial "diagnostic-files". **Restrict attention** to the data projection into the space defined by the **first versus second principal components** (we projected data into a 2-D space).

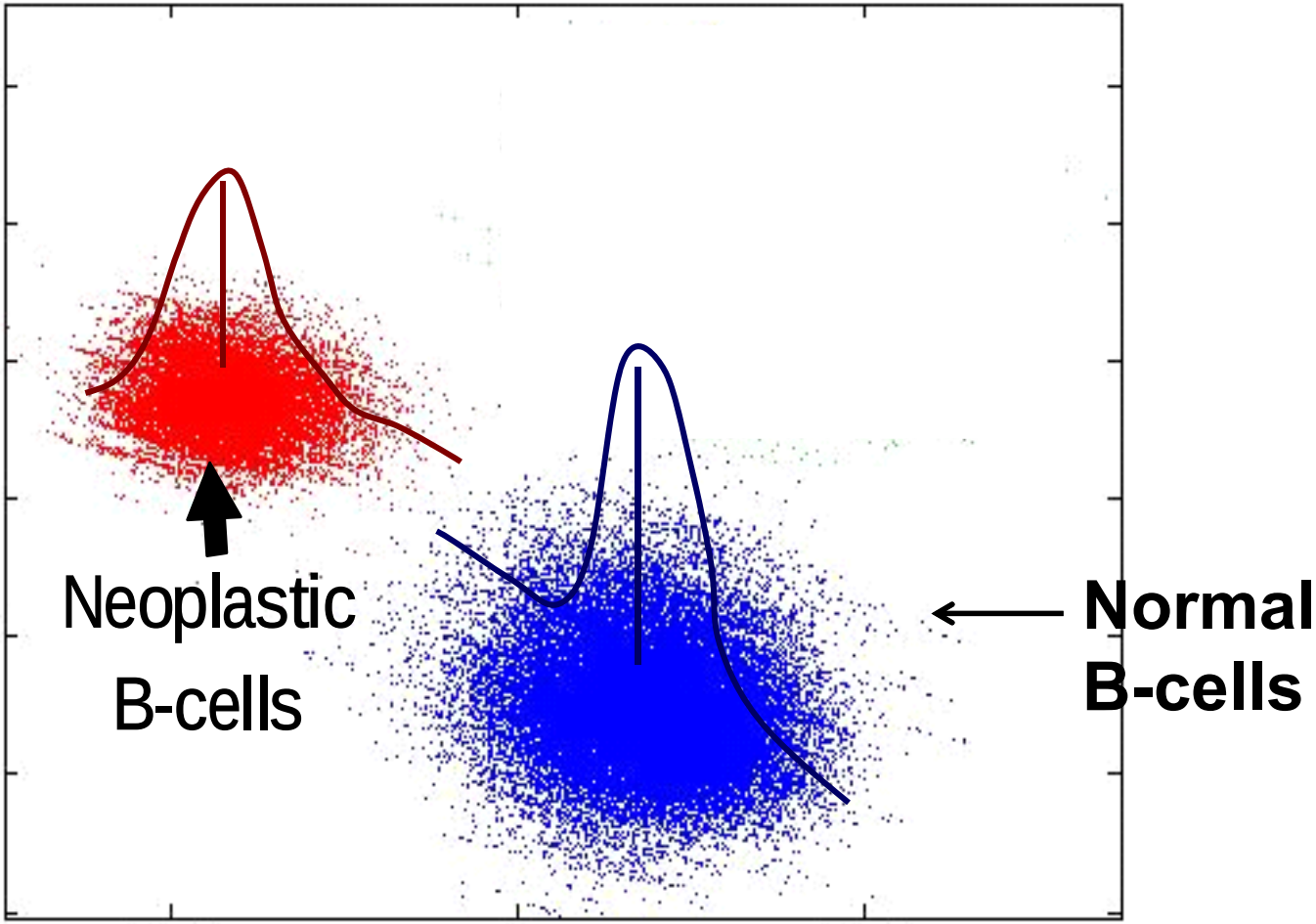
Calculamos a média e a matriz de covariância para a estimativa da projeção das populações normal e para CADA UMA (para cada paciente) das populações neoplasicas.

Assumindo Gaussianidade podemos estimar :

- **p(x | normal)** - a pdf associada a um evento assumir o valor x dado que sabemos que a população é normal, e
- **p(x | neoplastic)**

Dados reais de 4 pacientes




Neoplastic
B-cells
Normal
B-cells

Mas o que queremos de fato, é **P(normal | x)**, i.e. a probabilidade de que um evento pertença a população normal, uma vez que medidos (sabemos) os atributos deste evento.

This goal may be achieved by applying the *Bayes theorem* as follows:

(i) For the normal B-cell population:

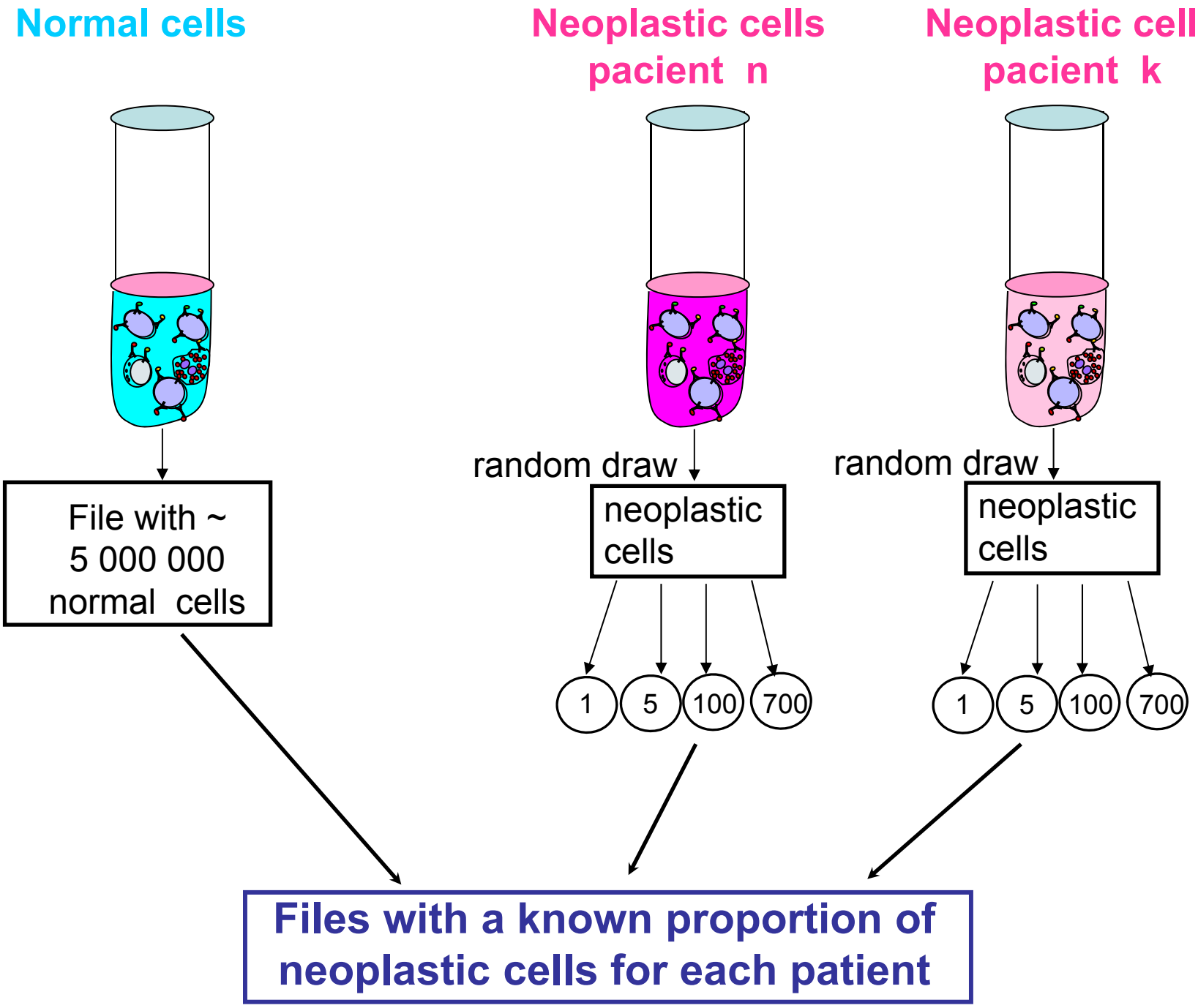
$$P(\text{normal} \mid x) = \frac{p(\text{normal}) \times p(x \mid \text{normal})}{K}$$

(ii) For the normal B-cell population:

$$P(\text{neoplastic} \mid x) = \frac{p(\text{neoplastic}) \times p(x \mid \text{neoplastic})}{K}$$

Here, K is a constant to make:

$$P(\text{normal} \mid x) + P(\text{neoplastic} \mid x) = 1$$


Normal cells
Neoplastic cells pacient n
Neoplastic cells pacient k
File with ~ 5 000 000 normal cells
random draw
neoplastic cells
1
5
100
700
random draw
neoplastic cells
1
5
100
700
Files with a known proportion of neoplastic cells for each patient

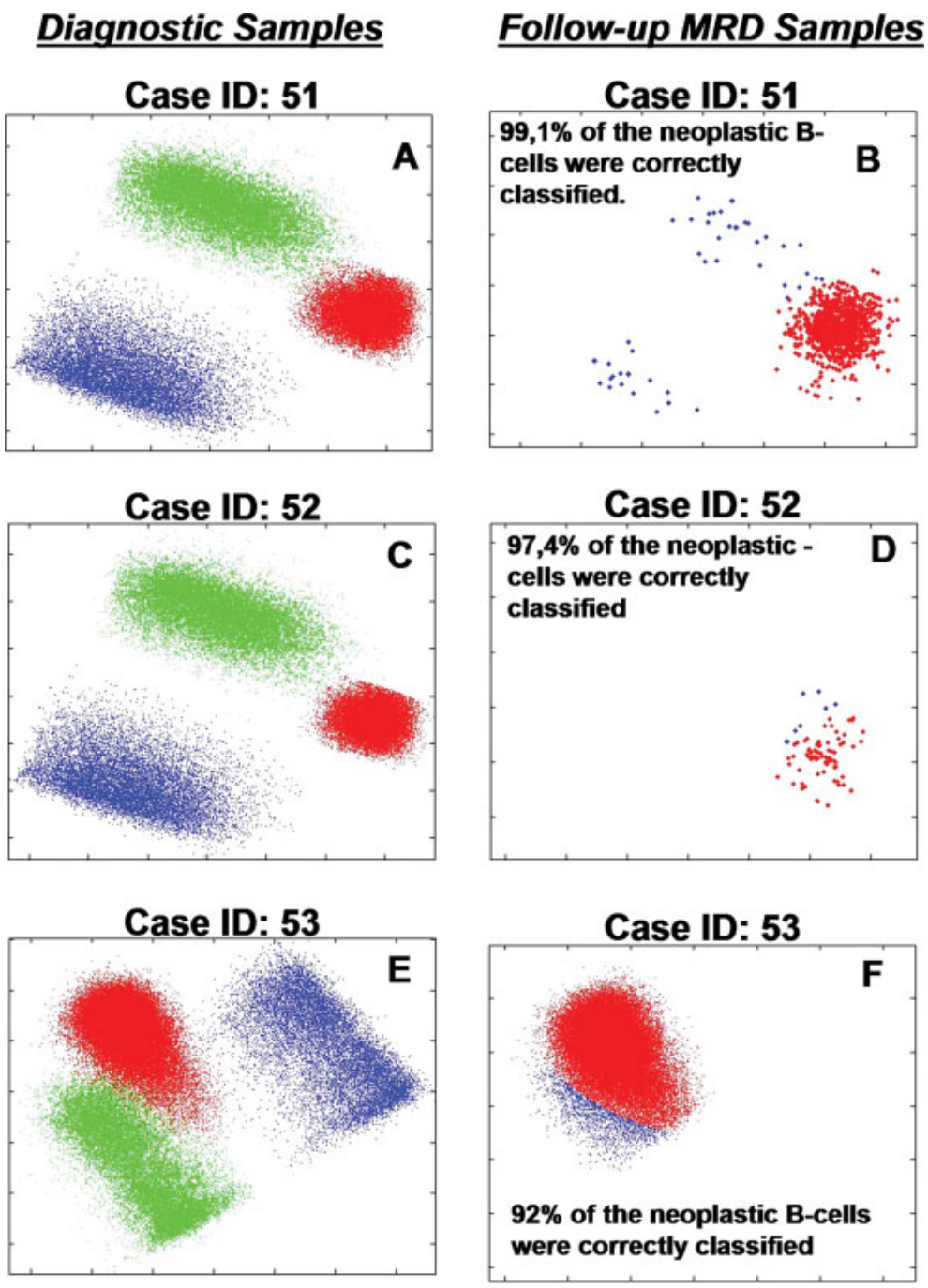
Results

Sensitivity:

In 80 % of the cases (**40/50**), the method was able to detect just 1 patological event in **5 x 10^6** normal cells.

Level of agremment:

For 90% of the pacients (**45/50**), the **correlation coeficient (r^2) was greater than 0.999**. The other 10% (5 cases) reached **0.964 ≤ r^2 ≤ 0.999**.


Diagnostic Samples
Follow-up MRD Samples
Case ID: 51
A
Case ID: 51
99,1% of the neoplastic B-cells were correctly classified.
B
Case ID: 52
C
Case ID: 52
97,4% of the neoplastic - cells were correctly classified
D
Case ID: 53
E
Case ID: 53
F
92% of the neoplastic B-cells were correctly classified

Dados de intensidades de proteína corrigida segundo a aproximação de LaBaer

	Pacientes ---->	1	2	3	4	5	6	7	8	9	10	11	12	13	14
		PAT 7657	PAT 7938	PAT 7942	PAT 8014	PAT 8015	PAT 8062	PAT 8063	PAT 8099	PAT 8136	PAT 8151	PAT 8253	PAT 8264	PAT 8286	PAT 8291
Evolução ---»»	Al diagnóstico ->	1	1	1	1	1	1	1	1	1	1	1	1	2	2
(Legendas ao final)	Evolución ->	1	1	1	1	1	1	1	1	1	1	1	1	3	3
	Final ->	1	1	1	1	1	1	1	1	1	1	1	1	1	1
proteinas ↓															
FGF13	1	4.50	6.47	2.05	3.55	4.20	2.24	1.05	2.79	0.87	0.14	1.07	1.71	1.19	0.37
TNF	2	4.20	6.69	2.28	4.28	7.34	5.24	2.25	3.08	2.30	0.23	1.58	1.95	4.41	0.69
PRKCA	3	3.61	5.99	1.83	3.82	4.30	4.85	0.99	2.94	1.07	0.17	0.95	0.85	2.22	0.31
WNT5A	4	4.02	5.57	2.61	5.06	4.81	6.42	1.24	2.57	2.49	0.25	1.50	4.18	5.95	0.45
CTSZ	5	3.74	5.71	2.40	5.82	4.01	6.22	1.28	5.49	2.52	0.20	1.26	4.46	2.66	0.51
PI3	6	3.95	6.95	2.21	6.56	5.05	5.14	1.36	4.92	3.16	0.21	1.62	5.30	3.14	0.67
IL9	7	2.09	2.73	1.09	5.64	4.39	5.41	0.69	3.11	0.81	0.06	0.65	1.96	1.73	0.29
CDKN1A	8	4.91	6.90	2.41	2.41	4.62	0.00	2.45	4.01	2.14	0.29	1.58	1.35	2.88	0.63
VEGFB	9	3.45	6.95	2.24	3.12	3.50	4.56	1.22	4.38	1.91	0.24	1.11	3.58	3.37	0.59
ARAF1	10	3.79	8.23	2.46	3.79	4.08	6.53	1.14	2.95	1.27	0.21	1.04	3.88	1.95	0.56

proteinas

Perguntas:

1) Quais proteinas diferenciam ‘ sãos’ de ‘patológicos’

2) Quais proteínas diferenciam ‘ metastásicos’ de ‘ Não metastasicos’

3) Quais proteinas podem predizer evolução

~~BIG~~ Bio Data

Apoio a decisão para estimação de risco de crianças diagnosticadas com LLA

Usa-se dados do **diagnóstico** para estimar o risco e modular o tratamento.

- Pedreira CE, Macrini L; Land M; Costa ES; “A New Decision Support Tool for Treatment Intensity Choice in Childhood Acute Lymphoblastic Leukemia”, **IEEE Transactions on Information Technology in Biomedicine**, v.13, p.284-290, (2009).

SMALL (but quite relevant) Bio Data

Apoio a decisão para estimação de risco de crianças diagnosticadas com LLA

Usa-se dados do **diagnóstico** para estimar o risco e modular o tratamento.

- Pedreira CE, Macrini L; Land M; Costa ES; “A New Decision Support Tool for Treatment Intensity Choice in Childhood Acute Lymphoblastic Leukemia”, **IEEE Transactions on Information Technology in Biomedicine**, v.13, p.284-290, (2009).

Patentes e Software

- **United States Patent nº US 7,321,843B2** "Method for generating flow cytometry data files containing an infinite number of dimensions based on data estimation". Inventors: Alberto Orfao de Matos, **Carlos Eduardo Pedreira** and Elaine Sobral da Costa. **License assigned to Cytognos SL**.
- **Internacional Patent nº WO 2010/140885 A1** (Provisional) "Methods, reagents and kits for flow cytometric immunophenotyping". Inventors: JJM van Dongen, JA Orfao de Matos Correia e Vale, JA Montero Flores, JM Almeida Parra, VHJ van der Velden, S Bottcher, AC Rawstron, RM de Tute, LBS Lhermitte, V Asnafi, E Mejstrikova, T Szczepanski, PJ Monteiro da Silva Lucio, M Martin Ayuso, **CE Pedreira**. **License assigned to Becton/Dickinson Biosciences and to Cytognos SL**.

✓ ***Software*** *'INFINICYT'* www.infinicyt.com que usa alguns dos resultados que mostramos é hoje uma ferramenta chave para o diagnostico de leucemias e linfomas. Esta licenciado e em uso em mais de **50 países** (> 1000 licenças).

Obrigado !

www.cos.ufrj.br/~pedreira

pedreira@ufrj.br
carlosp@centroin.com.br