Introdução à Mineração de Dados com Aplicações em Ciências Espaciais

Escola de Verão do Laboratório Associado de Computação e Matemática Aplicada

Rafael Santos

Programa

- ***Dia 1:*** Apresentação dos conceitos de mineração de dados, motivação e alguns exemplos.
- ***Dia 2:*** Algoritmos de classificação supervisionada e aplicações.
- ***Dia 3:*** Algoritmos de classificação não-supervisionada e aplicações. Algoritmos de mineração de associações.
- ***Dia 4:*** Visualização e mineração de dados. Outros algoritmos e idéias. Onde aprender mais.

Visualização

Visualização

- Pode ser usada no início do processo de mineração...
  - Para ter uma idéia da distribuição dos dados ou de relações entre os dados para formulação de hipóteses.
  - Para selecionar atributos ou regiões de dados.
  - Para ter uma idéia de que tipos de algoritmos podem trazer resultados para estes dados.
- Pode ser usada no final do processo de mineração...
  - Para ver as informações/regras/grupos/etc. obtidos: sumarização do conhecimento.
  - Para ver distribuições contextualizadas (isto é, com conhecimento adicional adquirido integrado).
  - Análise Explorativa / Análise Confirmativa / Apresentação

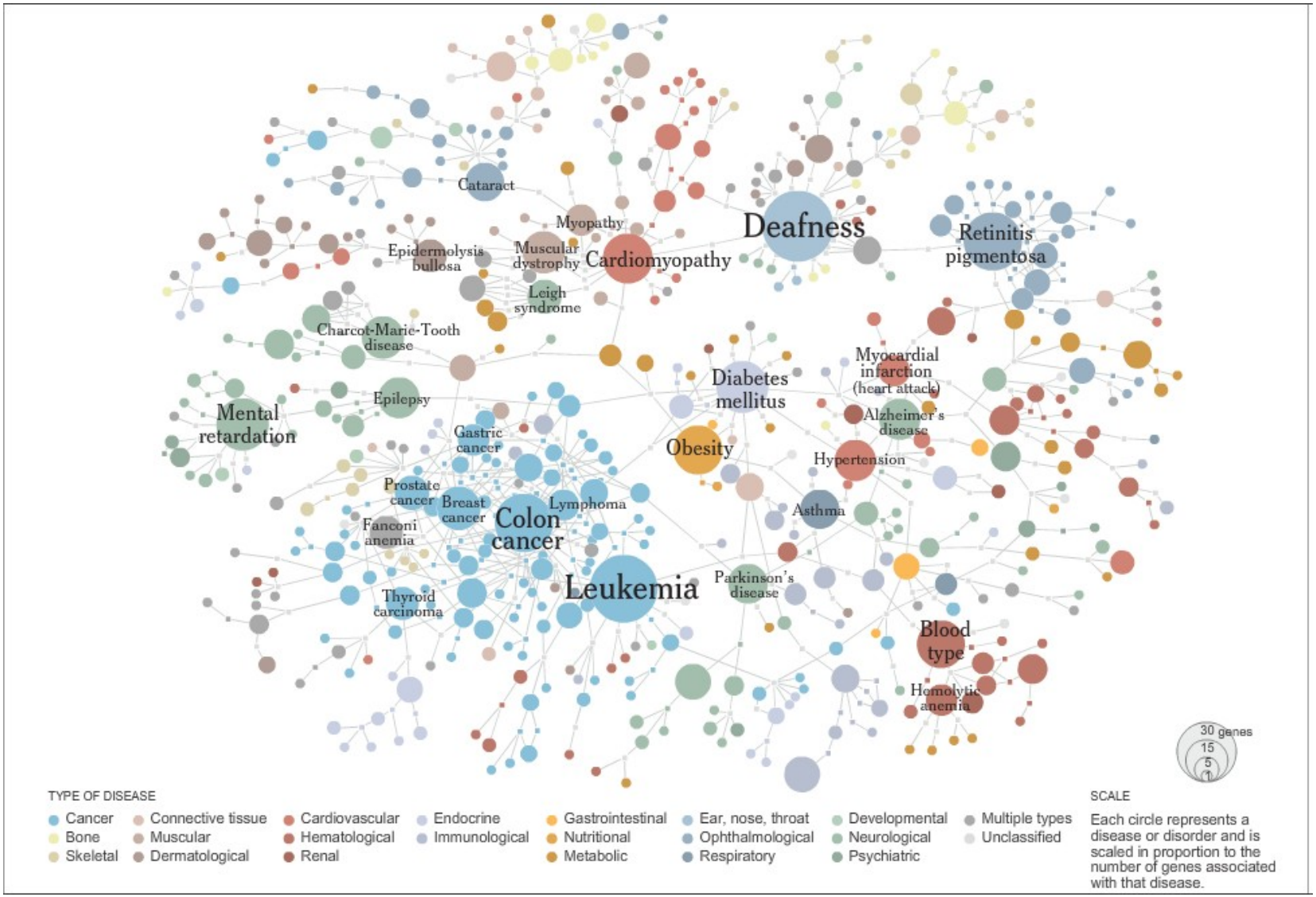
Objetivos de Visualização

- Análise exploratória:
  - Temos os dados, não temos hipótese sobre os mesmos.
  - Busca visual por padrões, estruturas, etc.
- Análise para confirmação:
  - Temos os dados e hipótese sobre os mesmos.
  - Busca visual para confirmar ou rejeitar.
- Apresentação
  - Técnica adequada deve ser usada!

Conceitos de Visualização

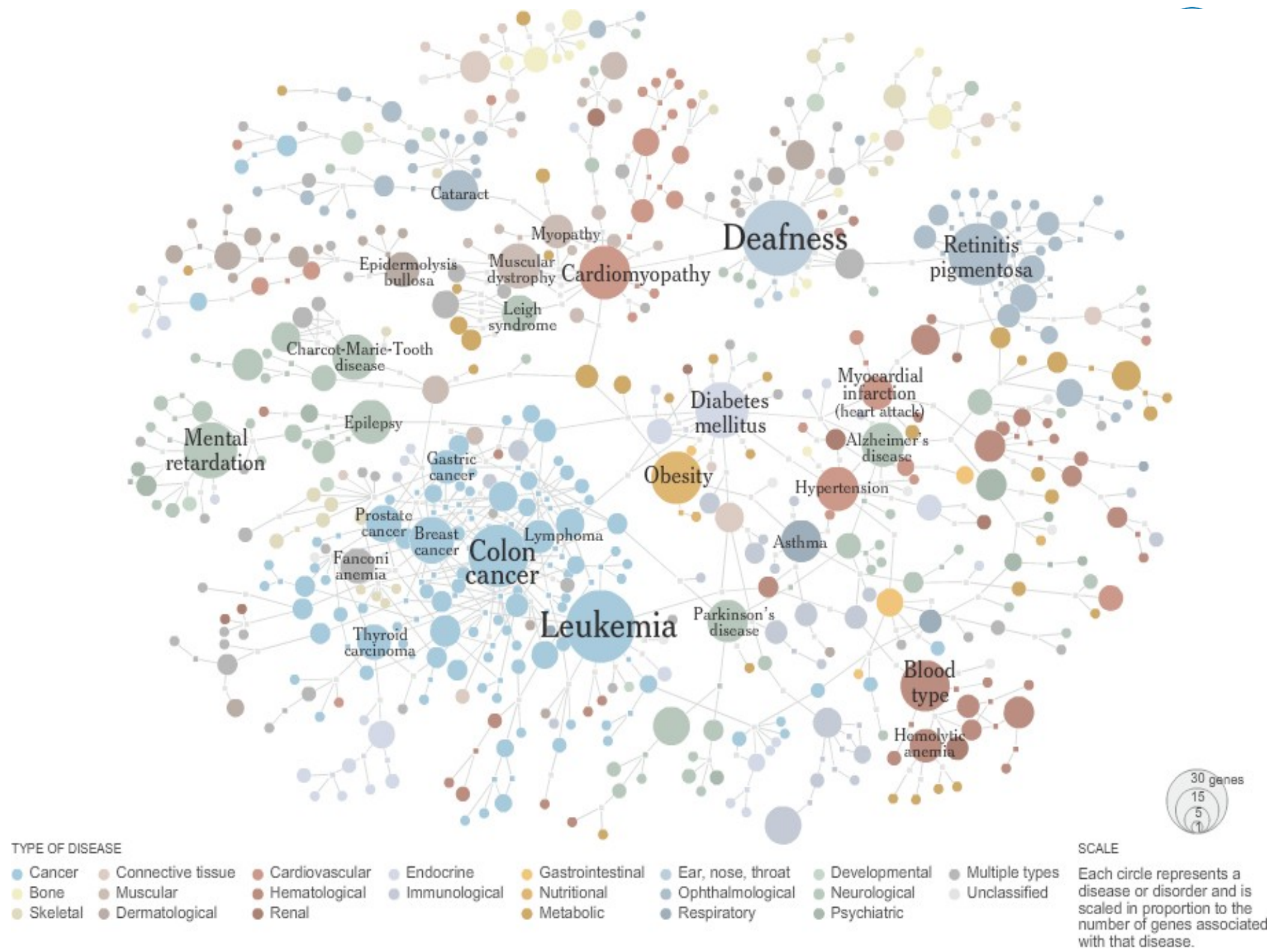
- Edward Tufte, *The Visual Display of Quantitative Information:*
  - "... gráficos sobre dados podem fazer muito mais do que simplesmente ser substitutos para pequenas tabelas estatísticas. Na sua melhor concepção, gráficos são instrumentos para compreender informação quantitativa."
  - "Frequentemente a forma mais efetiva de descrever, explorar e sumarizar um conjunto de números – mesmo um conjunto com muitos números – é **ver figuras destes números**."
  - "Adicionalmente, de todas as formas de analisar e comunicar informação estatística, gráficos bem feitos sobre dados são geralmente ao mesmo tempo a mais simples e mais poderosa".

Supergráficos

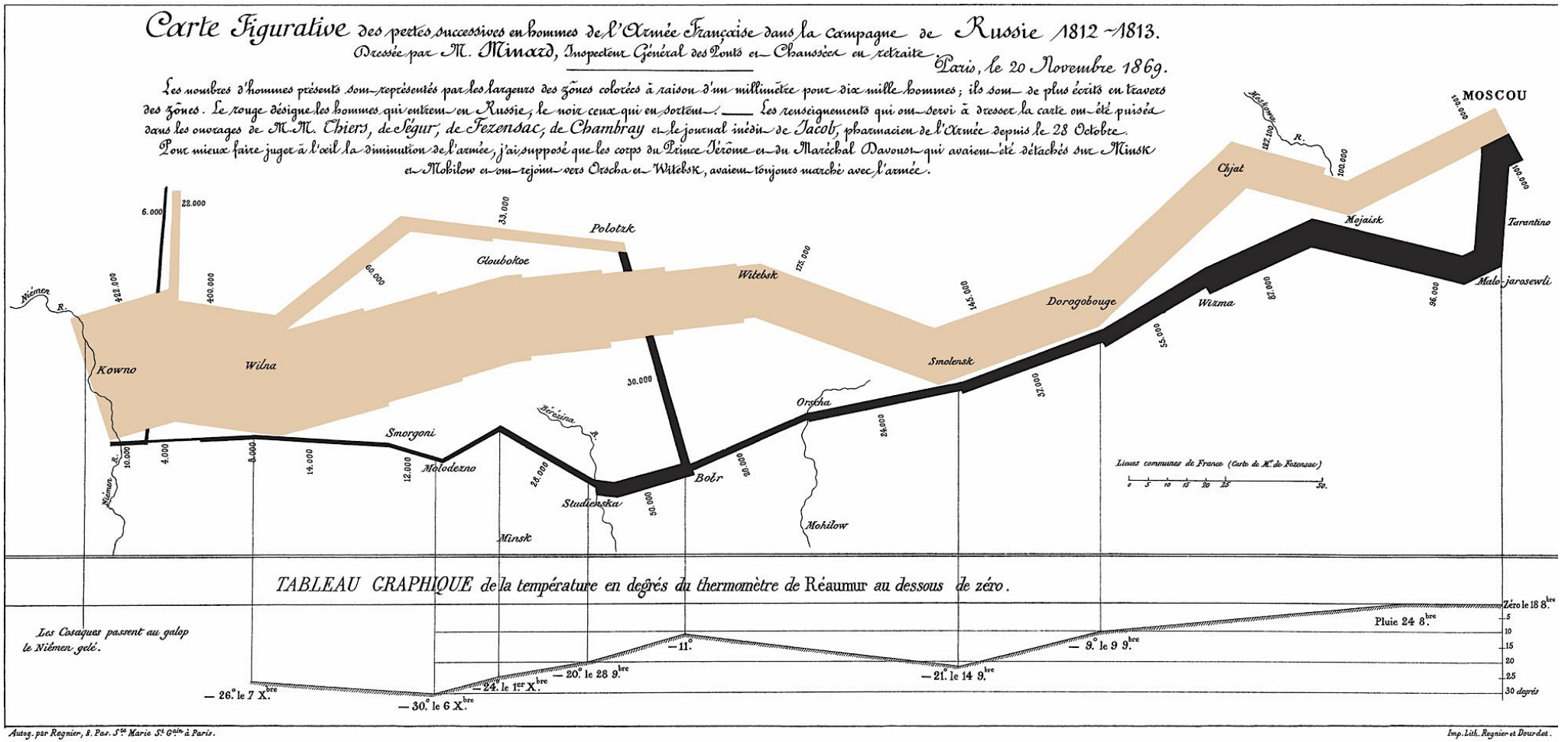


http://visualthinkmap.ning.com/

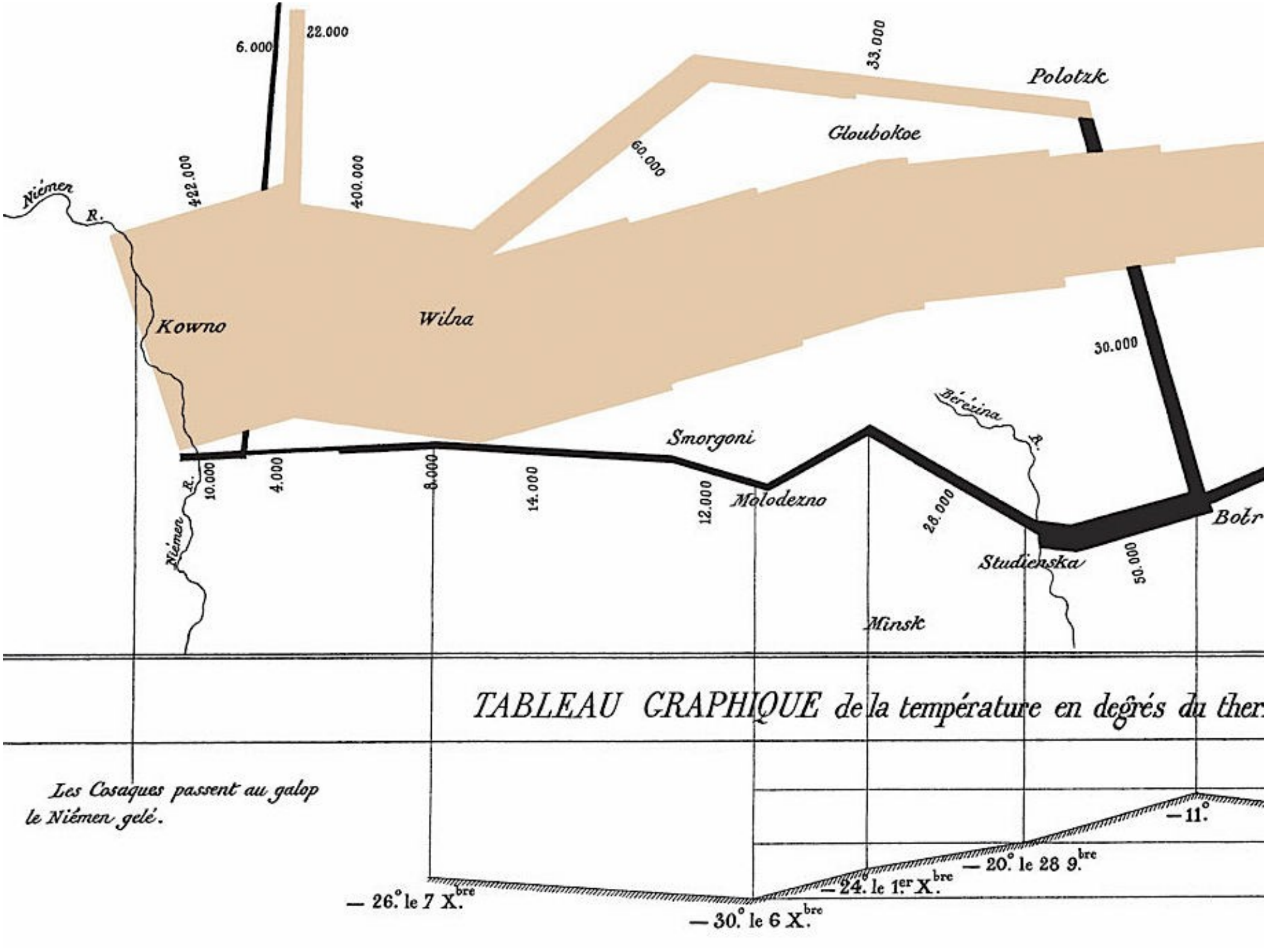
Mapping the Human 'Diseasome' by Marc Vidal, Albert-Laszlo Barabasi and Michael Cusick: ligação entre doenças e genes em comum.


Cataract
Myopathy
Deafness
Retinitis pigmentosa
Epidermolysis bullosa
Muscular dystrophy
Cardiomyopathy
Leigh syndrome
Charcot-Marie-Tooth disease
Myocardial infarction (heart attack)
Diabetes mellitus
Epilepsy
Mental retardation
Alzheimer's disease
Gastric cancer
Obesity
Hypertension
Prostate cancer
Breast cancer
Lymphoma
Asthma
Fanconi anemia
Colon cancer
Parkinson's disease
Leukemia
Thyroid carcinoma
Blood type
Hemolytic anemia
30 genes
15
5
1
TYPE OF DISEASE
Cancer
Bone
Skeletal
Connective tissue
Muscular
Dermatological
Cardiovascular
Hematological
Renal
Endocrine
Immunological
Gastrointestinal
Nutritional
Metabolic
Ear, nose, throat
Ophthalmological
Respiratory
Developmental
Neurological
Psychiatric
Multiple types
Unclassified
SCALE
Each circle represents a disease or disorder and is scaled in proportion to the number of genes associated with that disease.

Supergráficos



Marcha de Napoleão para Moscou na Guerra de 1812 (Charles Joseph Minard)


6.000
22.000
33.000
Polotzk
Glouboko̊e
60.000
Niémen R.
422.000
400.000
Kowno
Wilna
30.000
Bérézina R.
Smorgoni
Niémen R.
10.000
4.000
8.000
14.000
12.000
Molodezno
28.000
Bobr
Studienska
50.000
Minsk
TABLEAU GRAPHIQUE de la température en degrés du ther
Les Cosaques passent au galop le Niémen gelé.
— 11.°
— 20.° le 28 9.bre
— 24. le 1.er X.bre
— 26.° le 7 X.bre
— 30.° le 6 X.bre

Subjetividade em Visualização

Subjetividade em Visualização

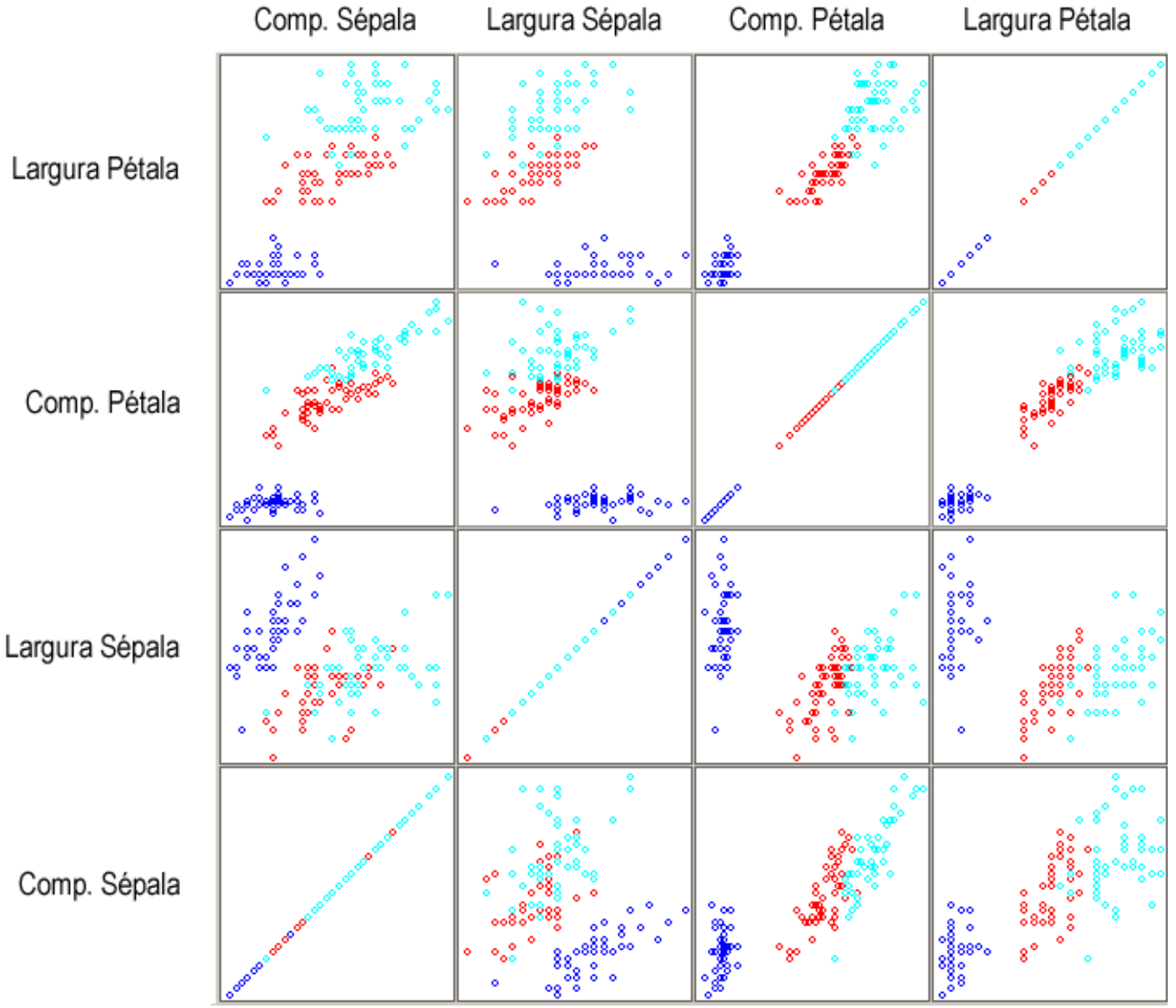
Visualização

- Desafios:
  - Subjetividade / Interpretabilidade.
  - Métodos e técnicas específicos.
  - Limitações de hardware (humano e máquina!)
  - Número de dimensões (atributos) dos dados.
  - Número de instâncias para visualização.
  - “Empilhamento” e ordenação.
- Vantagens:
  - Inerentemente exploratório.
  - Padrões detectados mesmo que não sejam explicáveis!

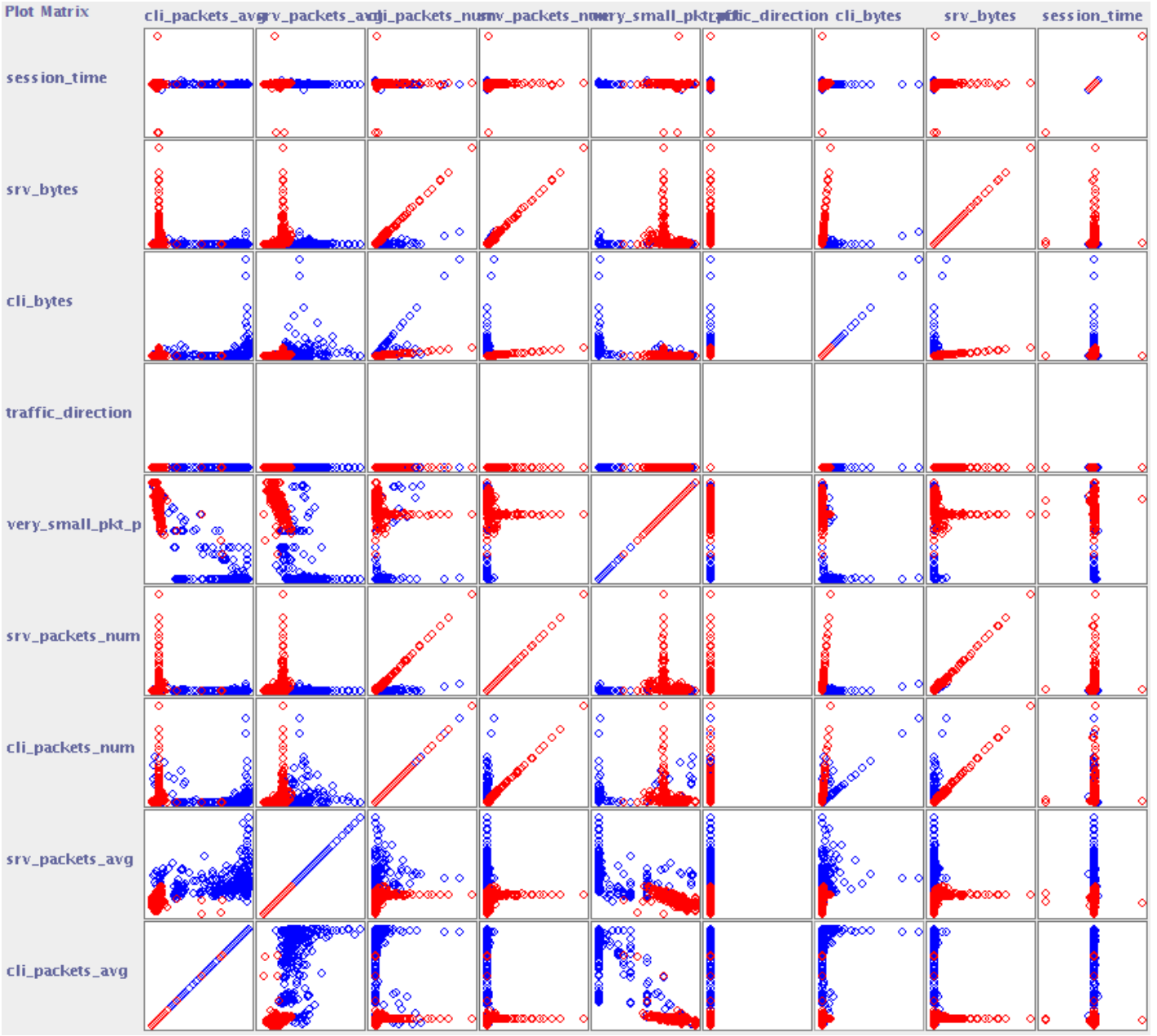
Visualização: Técnicas Geométricas

- Idéia básica: transformações e projeções usando arranjos em um número menor de dimensões.
  - *Scatterplot Matrices*: *K* atributos em grade *KxK*.
  - *Prosection Views*: *Scatterplot Matrices* com mecanismos de seleção (*drill-down*).
  - *Parallel Coordinates*: muito bom para dados mistos, requer exploração e rearranjos.
  - Visualização com Mapas de Kohonen (*SOMs*).

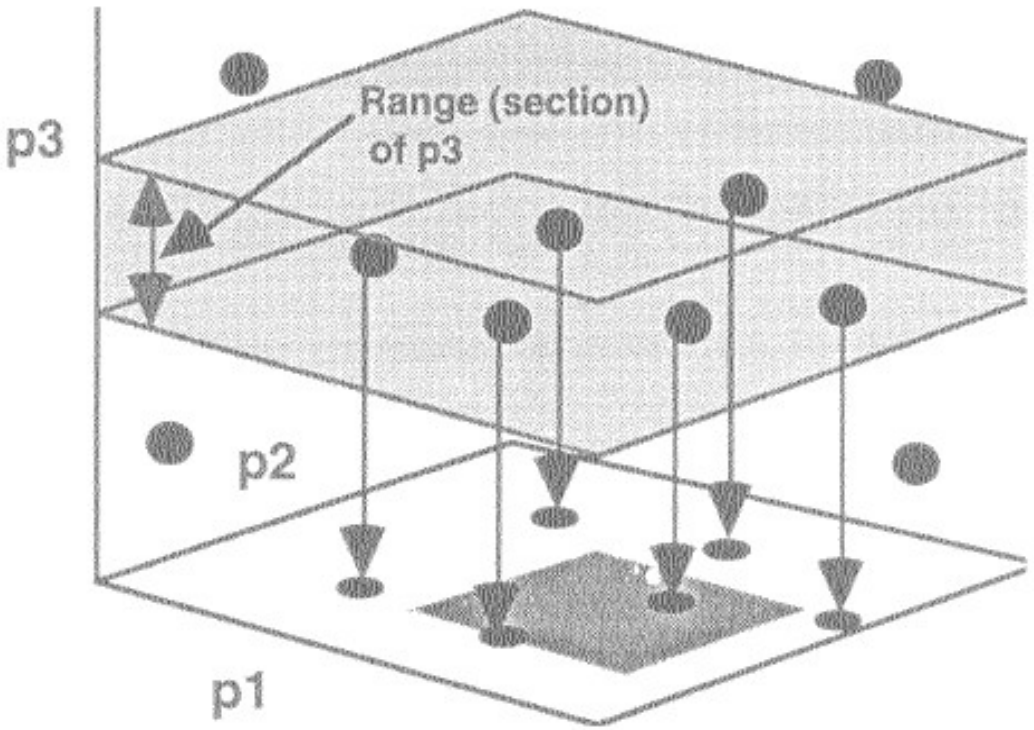
Visualização: *Scatterplot Matrices*

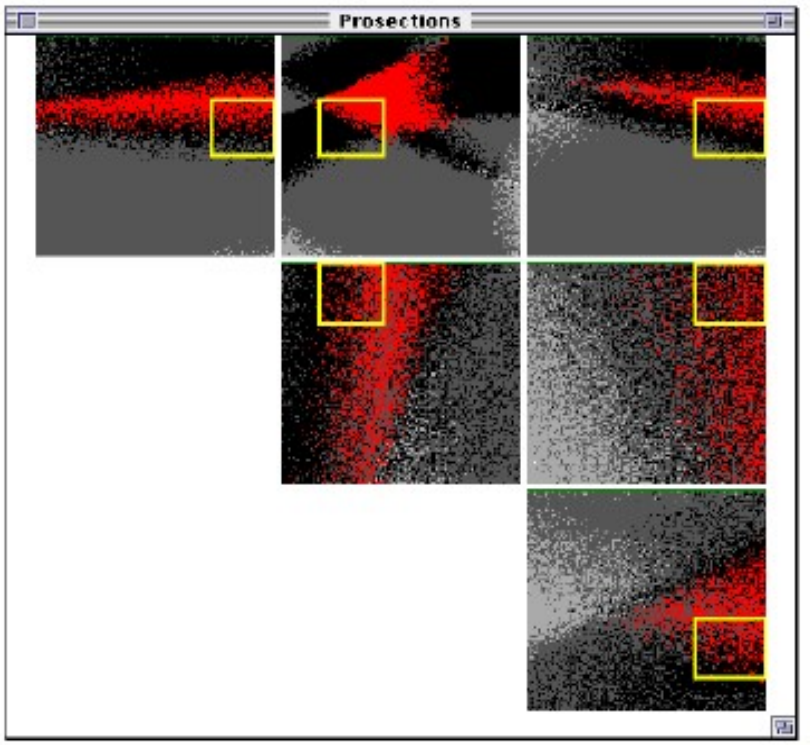


Visualização: *Scatterplot Matrices*



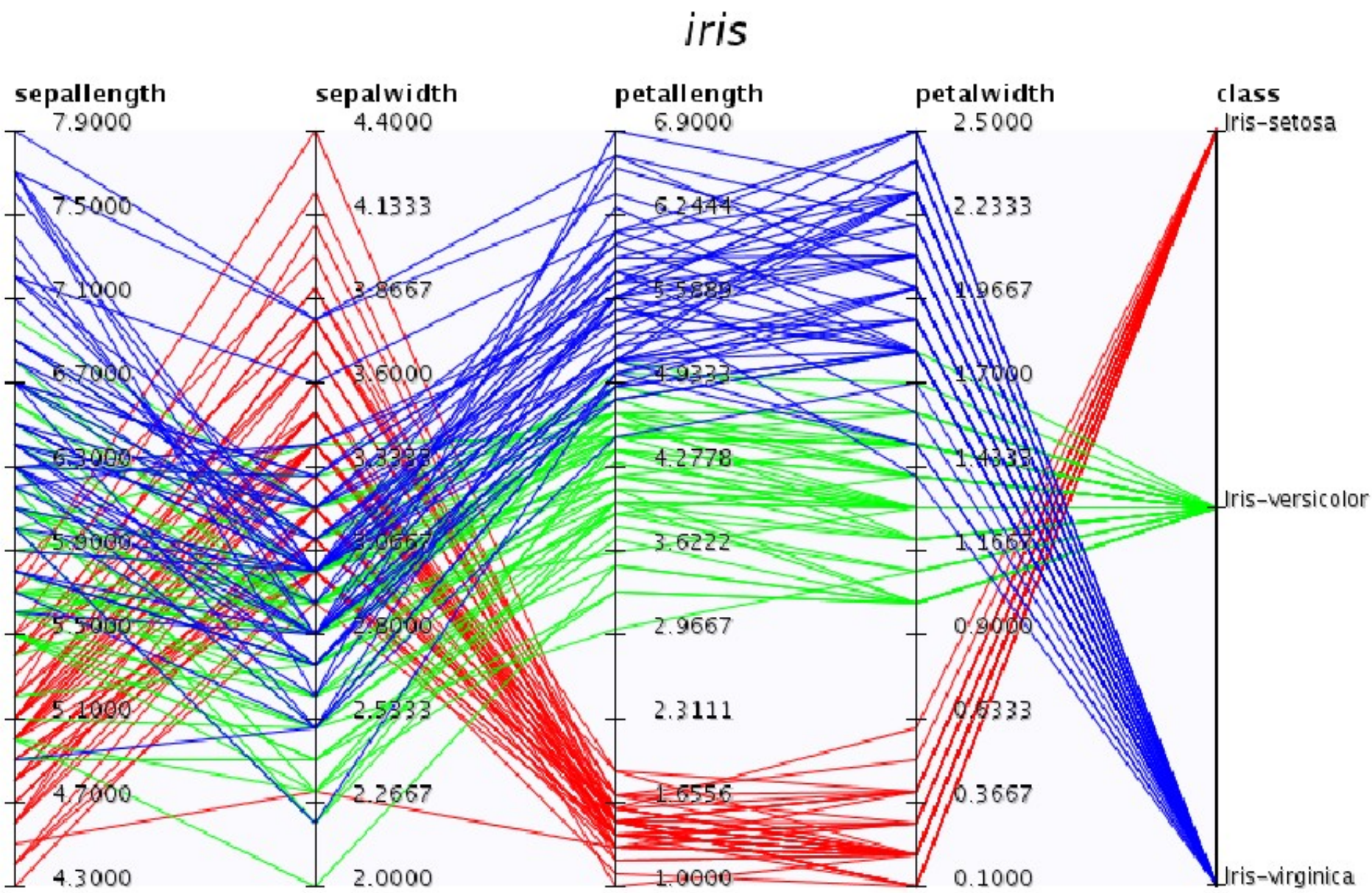
Visualização: *Prosection Views*



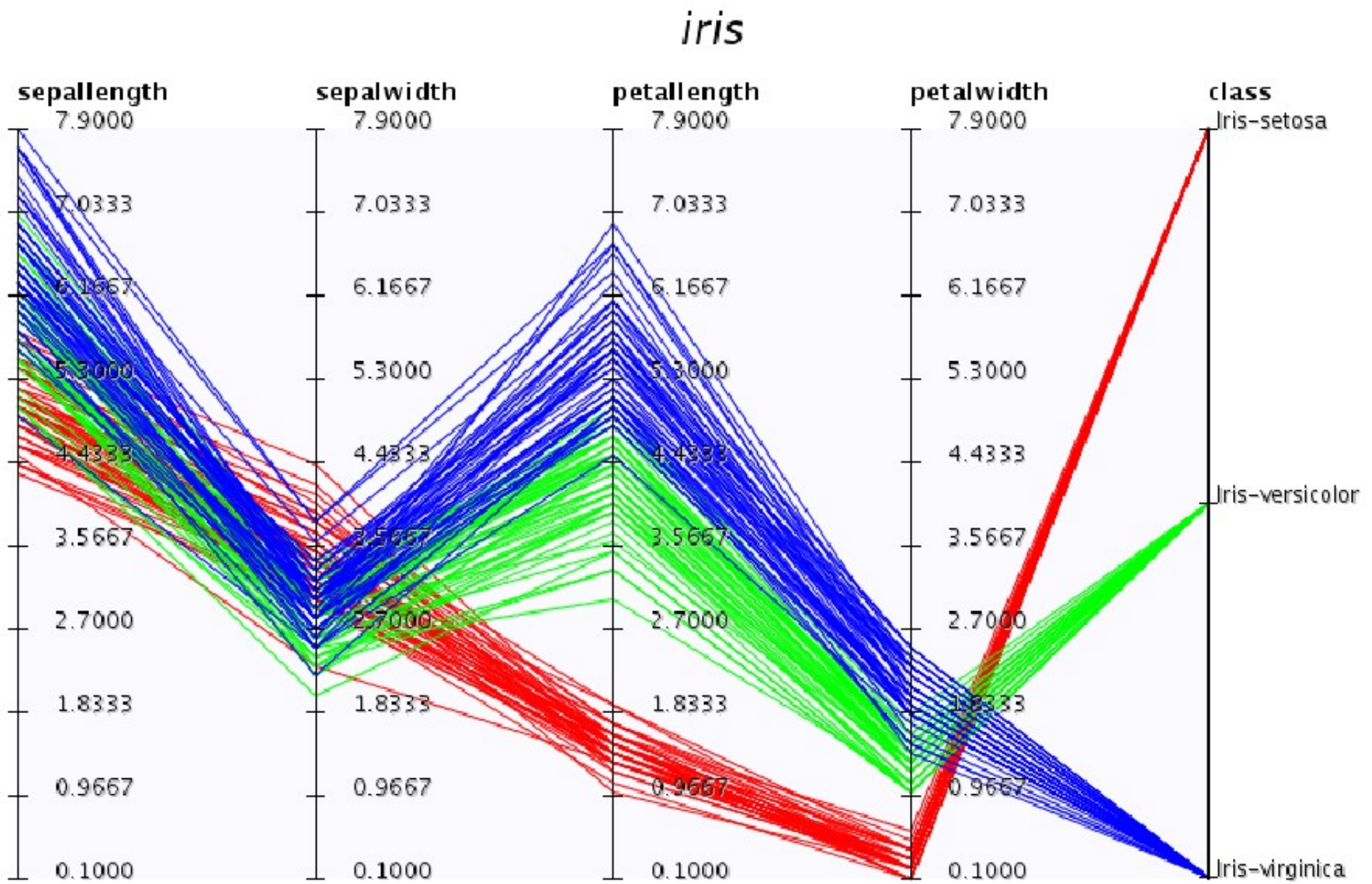


Exemplo de R. Spence, ilustrado no tutorial de Daniel Keim.

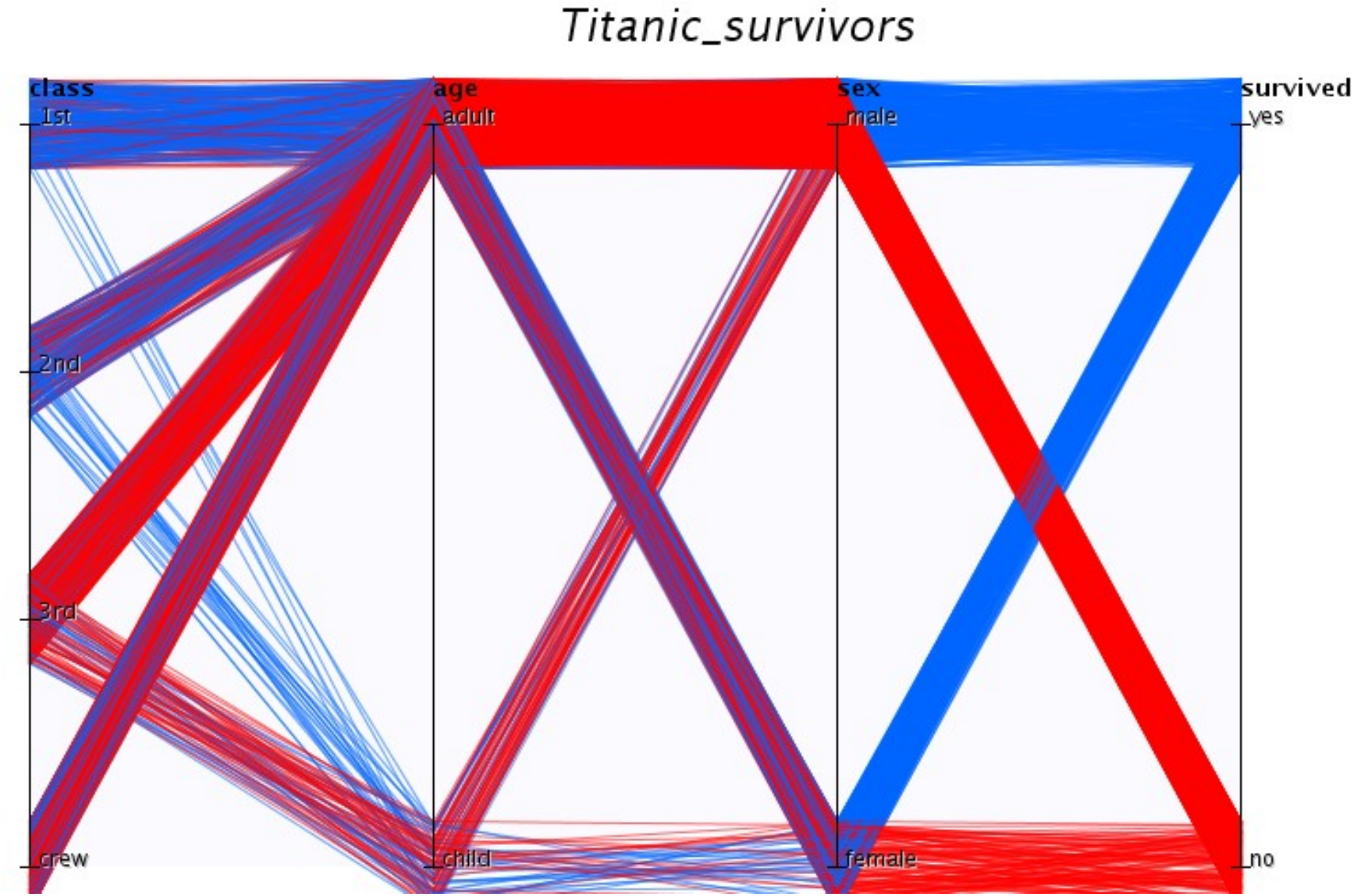
Visualização: *Parallel Coordinates*



Visualização: *Parallel Coordinates*



Visualização: *Parallel Coordinates*

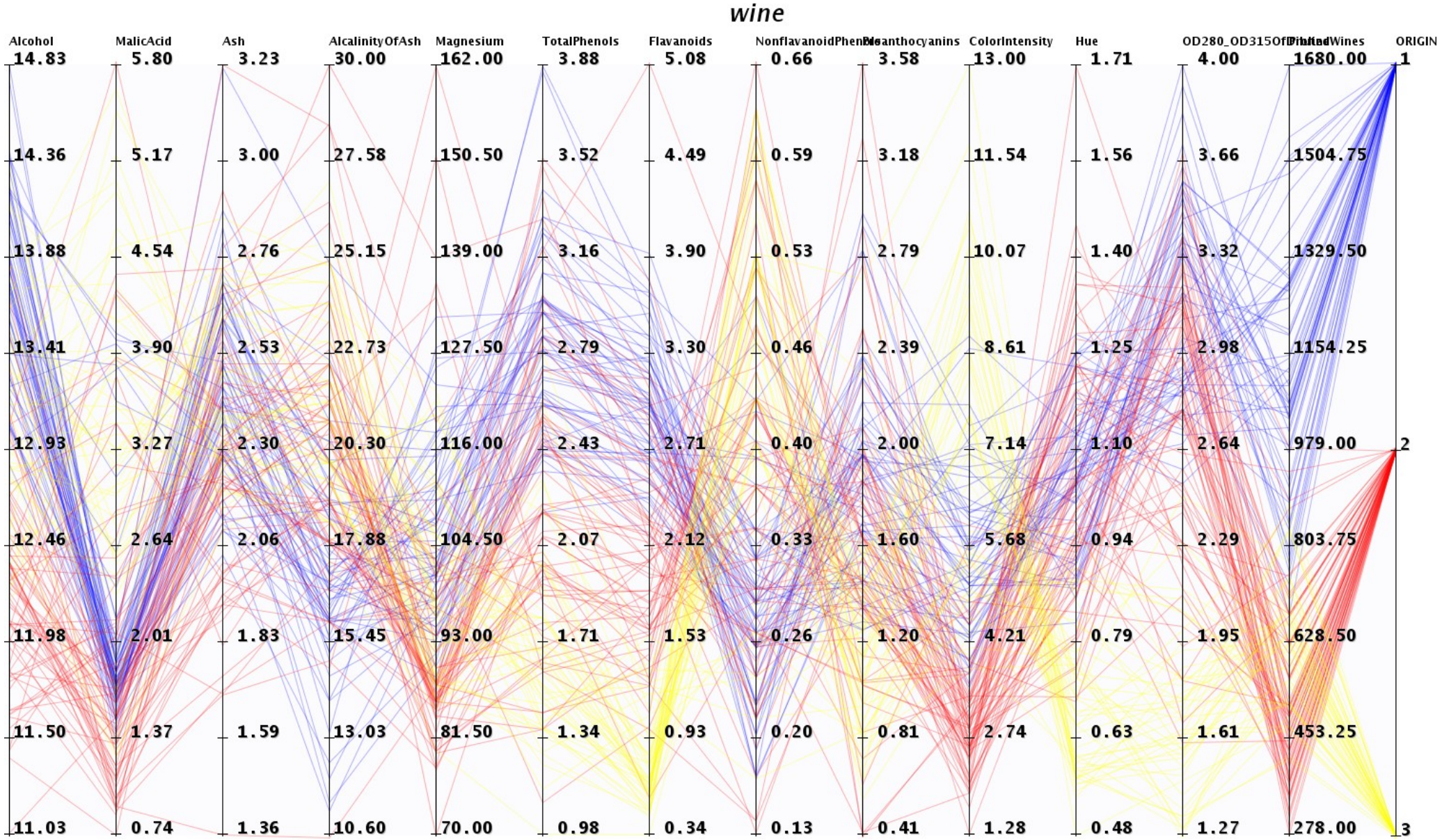


Visualização: *Parallel Coordinates*

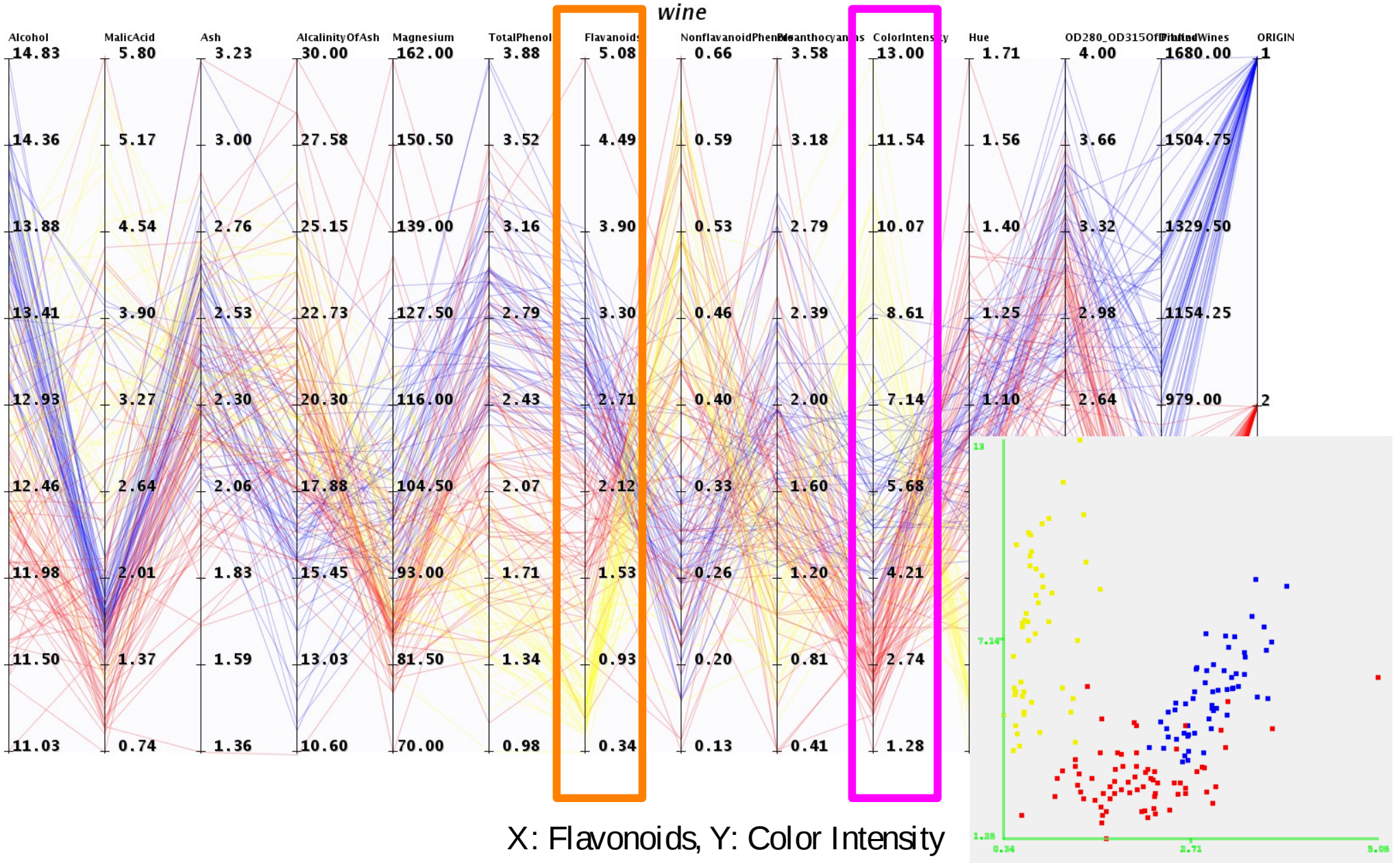
- Origem do vinho a partir de conteúdo físico-químico (13 atributos)
  http://archive.ics.uci.edu/ml/datasets/Wine (nomes de atributos originais)

No.	Alcohol Numeric	MalicAcid Numeric	Ash Numeric	AlcalinityOfAsh Numeric	Magnesium Numeric	TotalPhenols Numeric	Flavanoids Numeric	NonflavanoidPhenols Numeric	Proanthocyanins Numeric	ColorIntensity Numeric	Hue Numeric	OD280_OD315OfDilutedWines Numeric	Proline Numeric	**ORIGIN** Nominal
1	14.23	1.71	2.43	15.6	127.0	2.8	3.06	0.28	2.29	5.64	1.04	3.92	106...	1
2	13.2	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	3.4	105...	1
3	13.16	2.36	2.67	18.6	101.0	2.8	3.24	0.3	2.81	5.68	1.03	3.17	118...	1
4	14.37	1.95	2.5	16.8	113.0	3.85	3.49	0.24	2.18	7.8	0.86	3.45	148...	1
5	13.24	2.59	2.87	21.0	118.0	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735.0	1
6	14.2	1.76	2.45	15.2	112.0	3.27	3.39	0.34	1.97	6.75	1.05	2.85	145...	1
7	14.39	1.87	2.45	14.6	96.0	2.5	2.52	0.3	1.98	5.25	1.02	3.58	129...	1
8	14.06	2.15	2.61	17.6	121.0	2.6	2.51	0.31	1.25	5.05	1.06	3.58	129...	1
9	14.83	1.64	2.17	14.0	97.0	2.8	2.98	0.29	1.98	5.2	1.08	2.85	104...	1
10	13.86	1.35	2.27	16.0	98.0	2.98	3.15	0.22	1.85	7.22	1.01	3.55	104...	1
11	14.1	2.16	2.3	18.0	105.0	2.95	3.32	0.22	2.38	5.75	1.25	3.17	151...	1
12	14.12	1.48	2.32	16.8	95.0	2.2	2.43	0.26	1.57	5.0	1.17	2.82	128...	1
13	13.75	1.73	2.41	16.0	89.0	2.6	2.76	0.29	1.81	5.6	1.15	2.9	132...	1
14	14.75	1.73	2.39	11.4	91.0	3.1	3.69	0.43	2.81	5.4	1.25	2.73	115...	1
15	14.38	1.87	2.38	12.0	102.0	3.3	3.64	0.29	2.96	7.5	1.2	3.0	154...	1
16	13.63	1.81	2.7	17.2	112.0	2.85	2.91	0.3	1.46	7.3	1.28	2.88	131...	1
17	14.3	1.92	2.72	20.0	120.0	2.8	3.14	0.33	1.97	6.2	1.07	2.65	128...	1
18	13.83	1.57	2.62	20.0	115.0	2.95	3.4	0.4	1.72	6.6	1.13	2.57	113...	1
19	14.19	1.59	2.48	16.5	108.0	3.3	3.93	0.32	1.86	8.7	1.23	2.82	168...	1
20	13.64	3.1	2.56	15.2	116.0	2.7	3.03	0.17	1.66	5.1	0.96	3.36	845.0	1
21	14.06	1.63	2.28	16.0	126.0	3.0	3.17	0.24	2.1	5.65	1.09	3.71	780.0	1
22	12.93	3.8	2.65	18.6	102.0	2.41	2.41	0.25	1.98	4.5	1.03	3.52	770.0	1
23	13.71	1.86	2.36	16.6	101.0	2.61	2.88	0.27	1.69	3.8	1.11	4.0	103...	1
24	12.85	1.6	2.52	17.8	95.0	2.48	2.37	0.26	1.46	3.93	1.09	3.63	101...	1
25	13.5	1.81	2.61	20.0	96.0	2.53	2.61	0.28	1.66	3.52	1.12	3.82	845.0	1
26	13.05	2.05	3.22	25.0	124.0	2.63	2.68	0.47	1.92	3.58	1.13	3.2	830.0	1
27	13.39	1.77	2.62	16.1	93.0	2.85	2.94	0.34	1.45	4.8	0.92	3.22	119...	1
28	13.3	1.72	2.14	17.0	94.0	2.4	2.19	0.27	1.35	3.95	1.02	2.77	128...	1
29	13.87	1.9	2.8	19.4	107.0	2.95	2.97	0.37	1.76	4.5	1.25	3.4	915.0	1
30	14.02	1.68	2.21	16.0	96.0	2.65	2.33	0.26	1.98	4.7	1.04	3.59	103...	1
31	13.73	1.5	2.7	22.5	101.0	3.0	3.25	0.29	2.38	5.7	1.19	2.71	128...	1
32	13.58	1.66	2.36	19.1	106.0	2.86	3.19	0.22	1.95	6.9	1.09	2.88	151...	1
33	13.68	1.83	2.36	17.2	104.0	2.42	2.69	0.42	1.97	3.84	1.23	2.87	990.0	1
34	13.76	1.53	2.7	19.5	132.0	2.95	2.74	0.5	1.35	5.4	1.25	3.0	123...	1
35	13.51	1.8	2.65	19.0	110.0	2.35	2.53	0.29	1.54	4.2	1.1	2.87	109...	1
36	13.48	1.81	2.41	20.5	100.0	2.7	2.98	0.26	1.86	5.1	1.04	3.47	920.0	1
37	13.28	1.64	2.84	15.5	110.0	2.6	2.68	0.34	1.36	4.6	1.09	2.78	880.0	1
38	13.05	1.65	2.55	18.0	98.0	2.45	2.43	0.29	1.44	4.25	1.12	2.51	110...	1
39	13.07	1.5	2.1	15.5	98.0	2.4	2.64	0.28	1.37	3.7	1.18	2.69	102...	1

Visualização: *Parallel Coordinates*

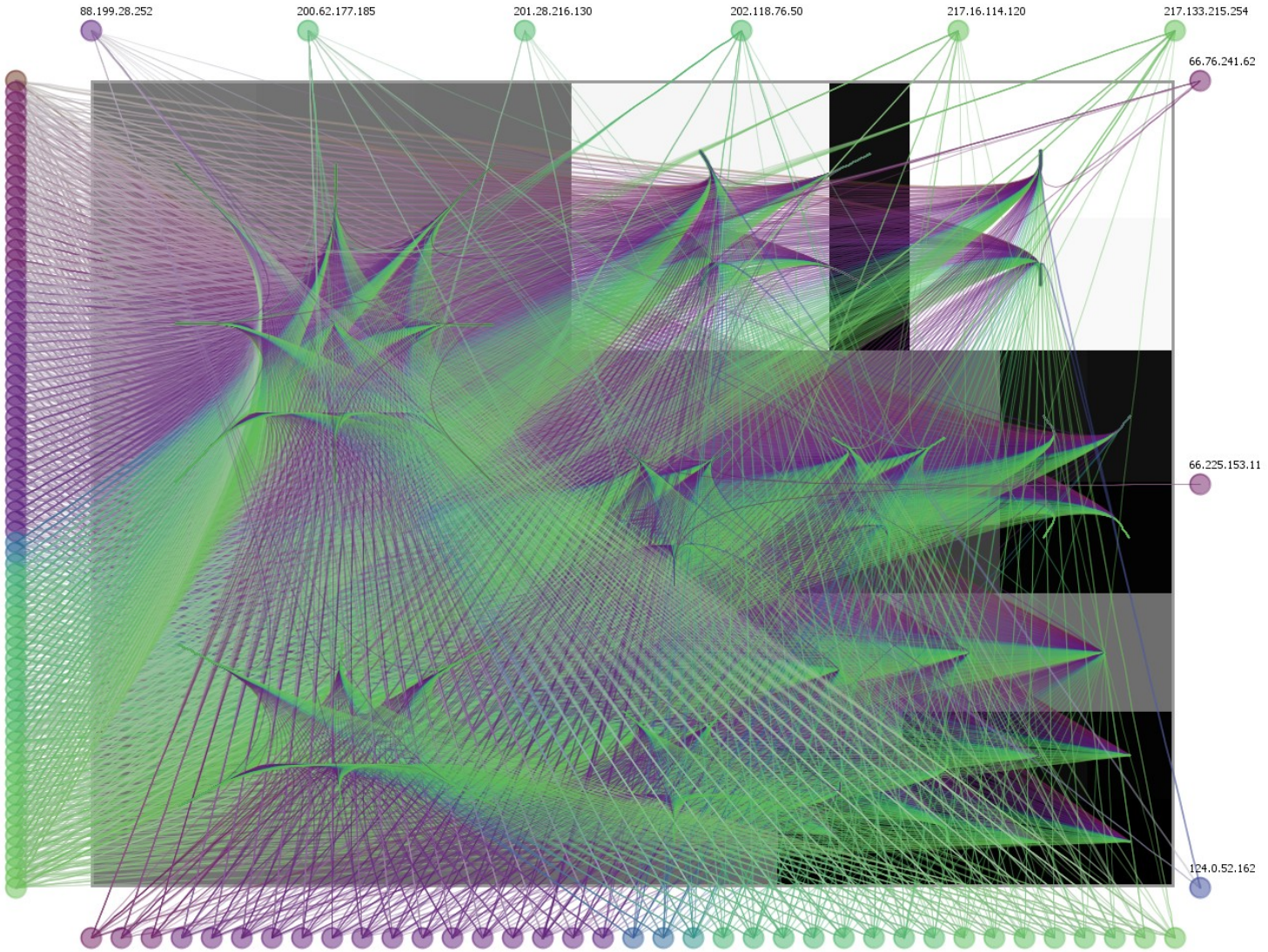


Visualização: *Parallel Coordinates*

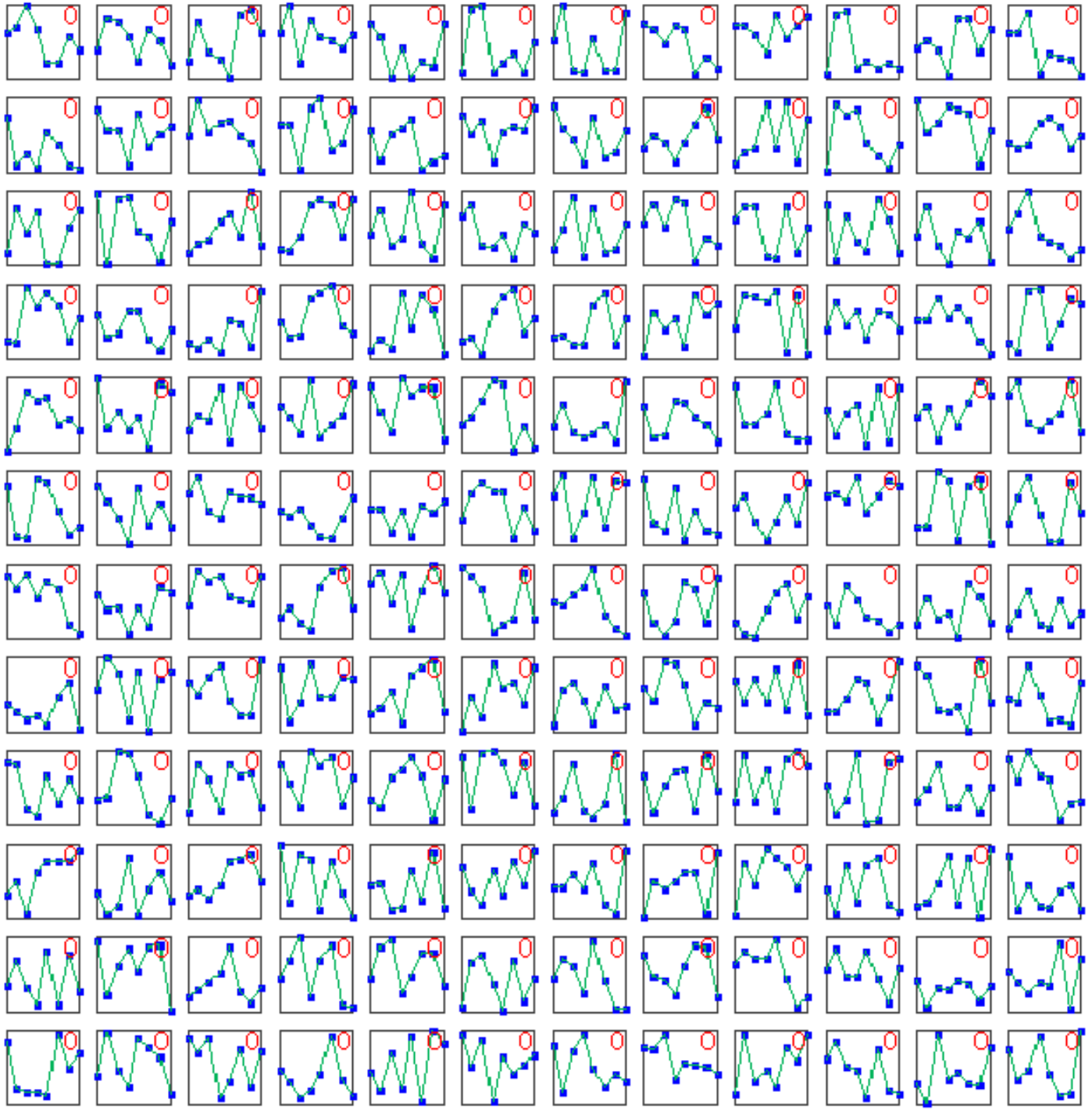


X: Flavonoids, Y: Color Intensity

Visualização: *Parallel Coordinates*



Fabian Fischer, Florian Mansmann, Daniel A. Keim, Stephan Pietzko, and Marcel Waldvogel. Large-Scale Network Monitoring for Visual Analysis of Attacks. VizSec 2008 (LNCS 5210)



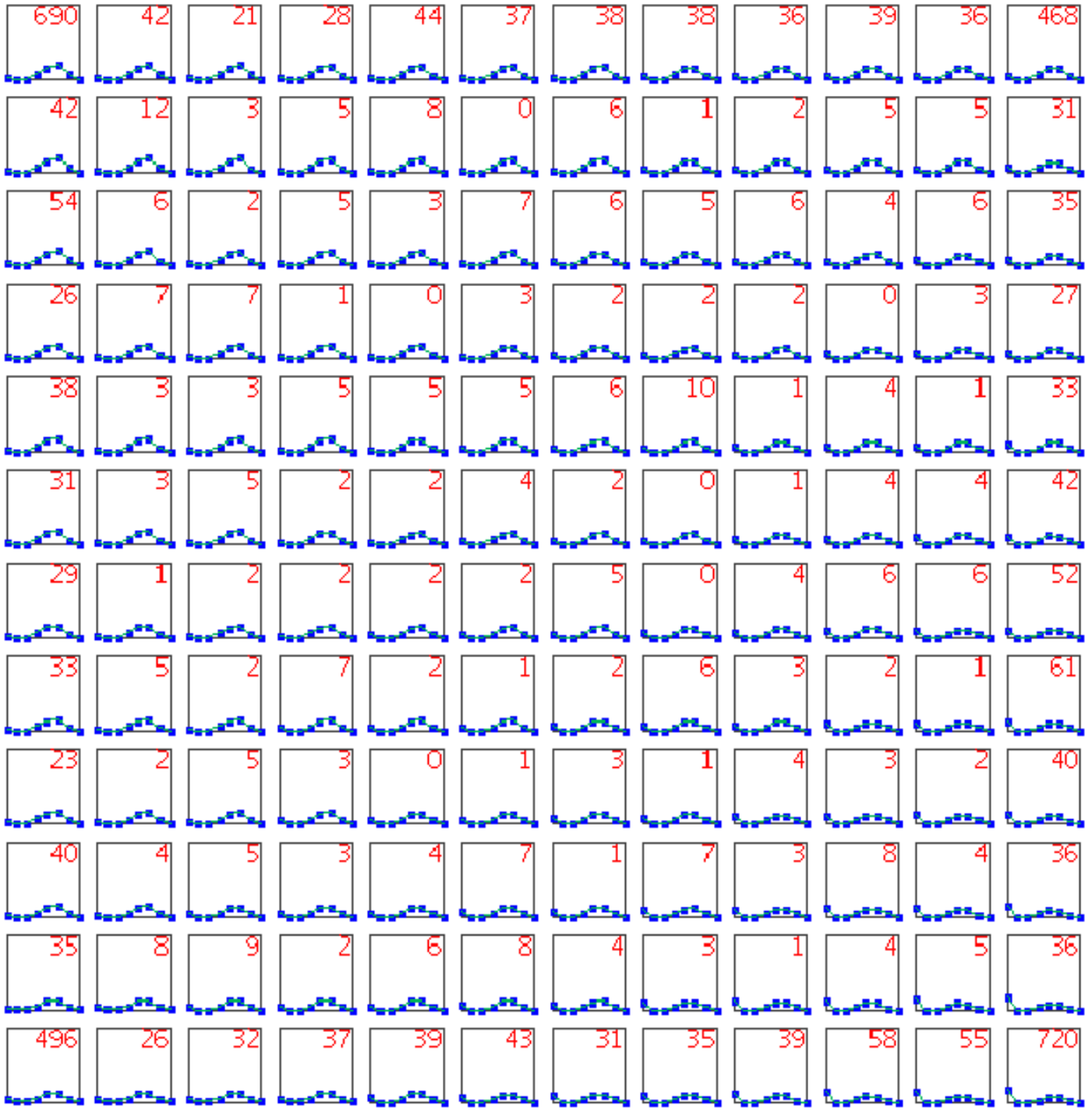
12x12 SOM,

Dados em 8 dimensões.

T=0

R=25

Lr=0.9



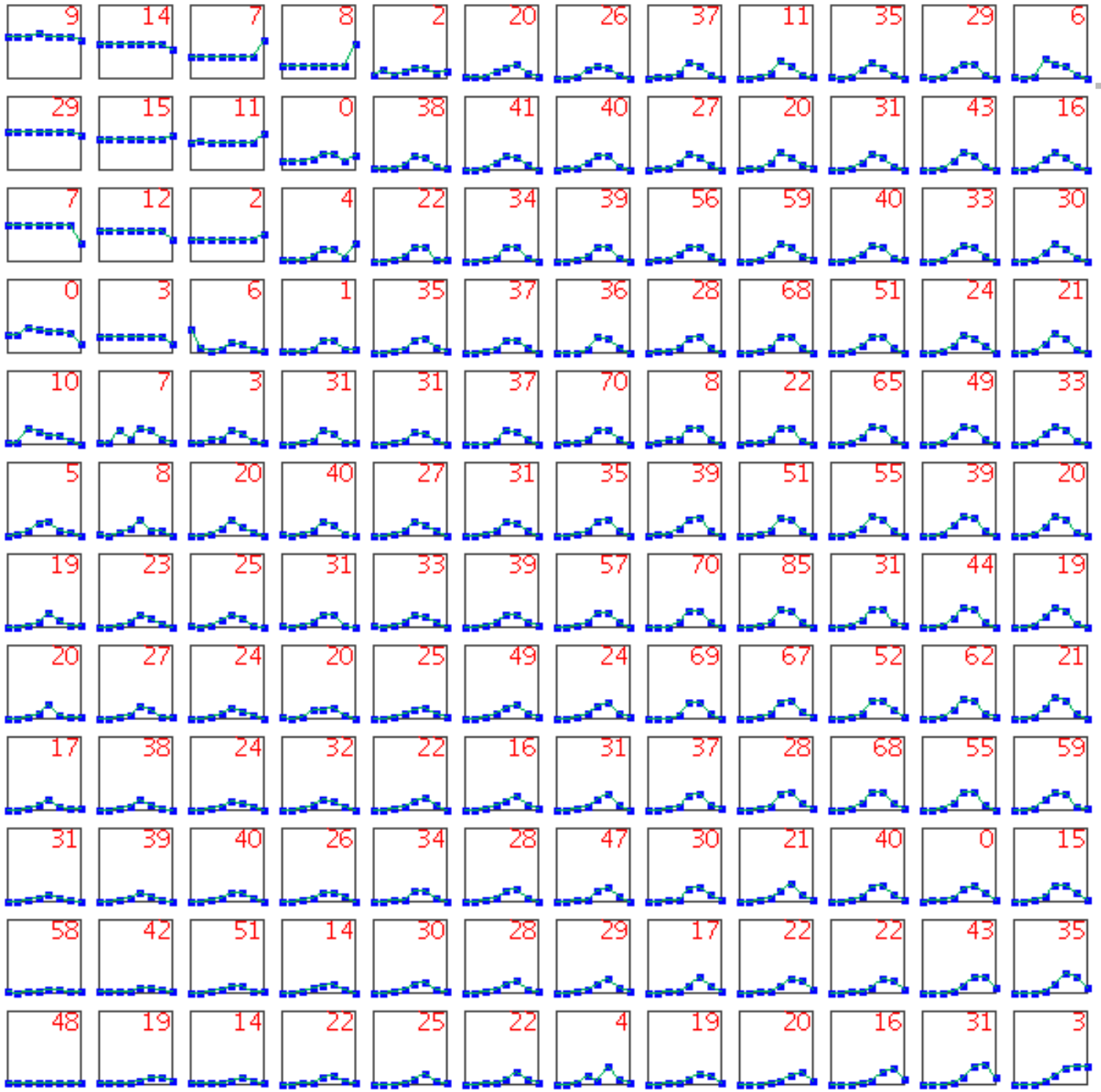
12x12 SOM,

Dados em 8 dimensões.

T=40

R=16.7

Lr=0.74



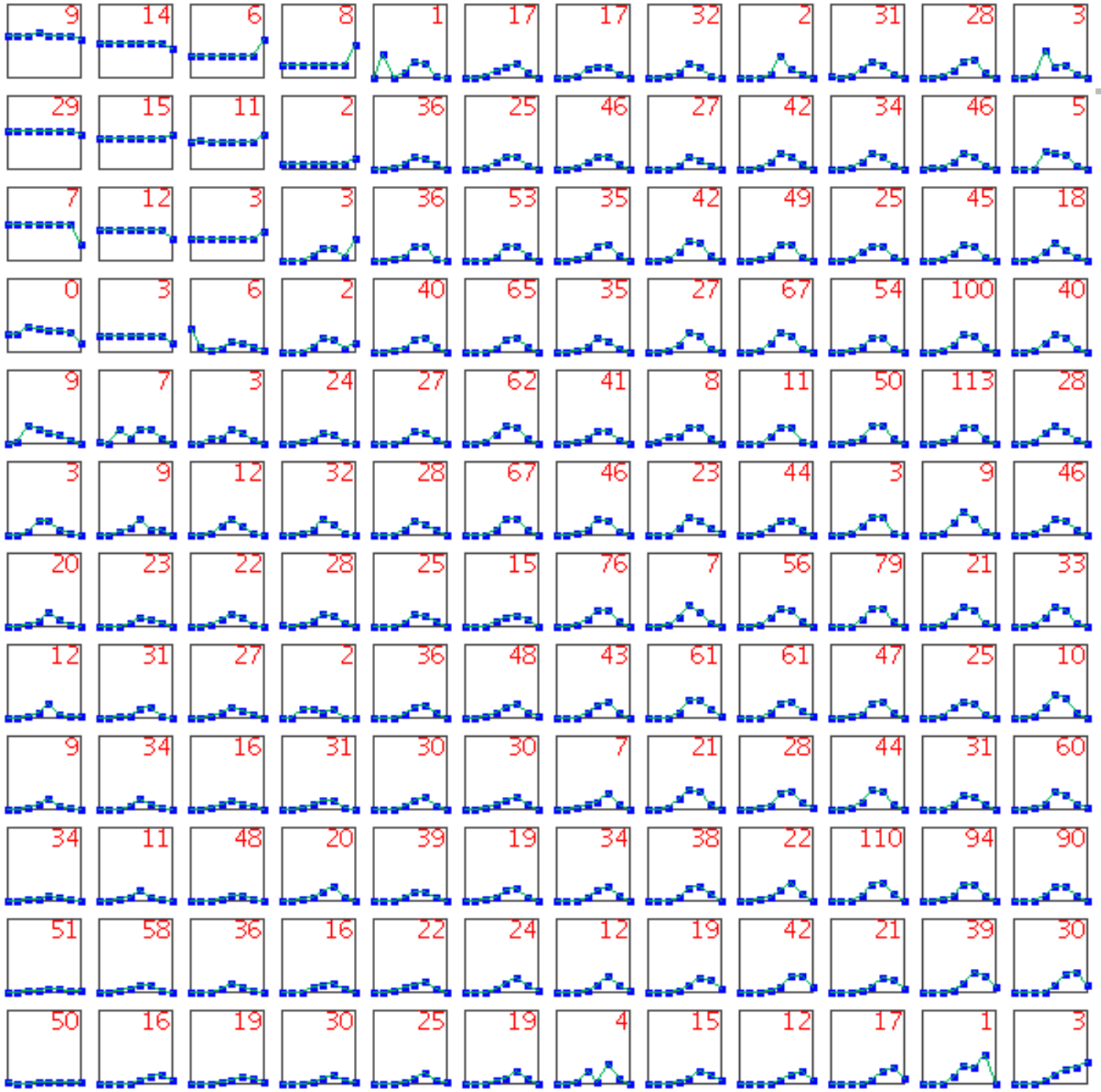
12x12 SOM,

Dados em 8 dimensões.

T=320

R=1

Lr=0.18



12x12 SOM,

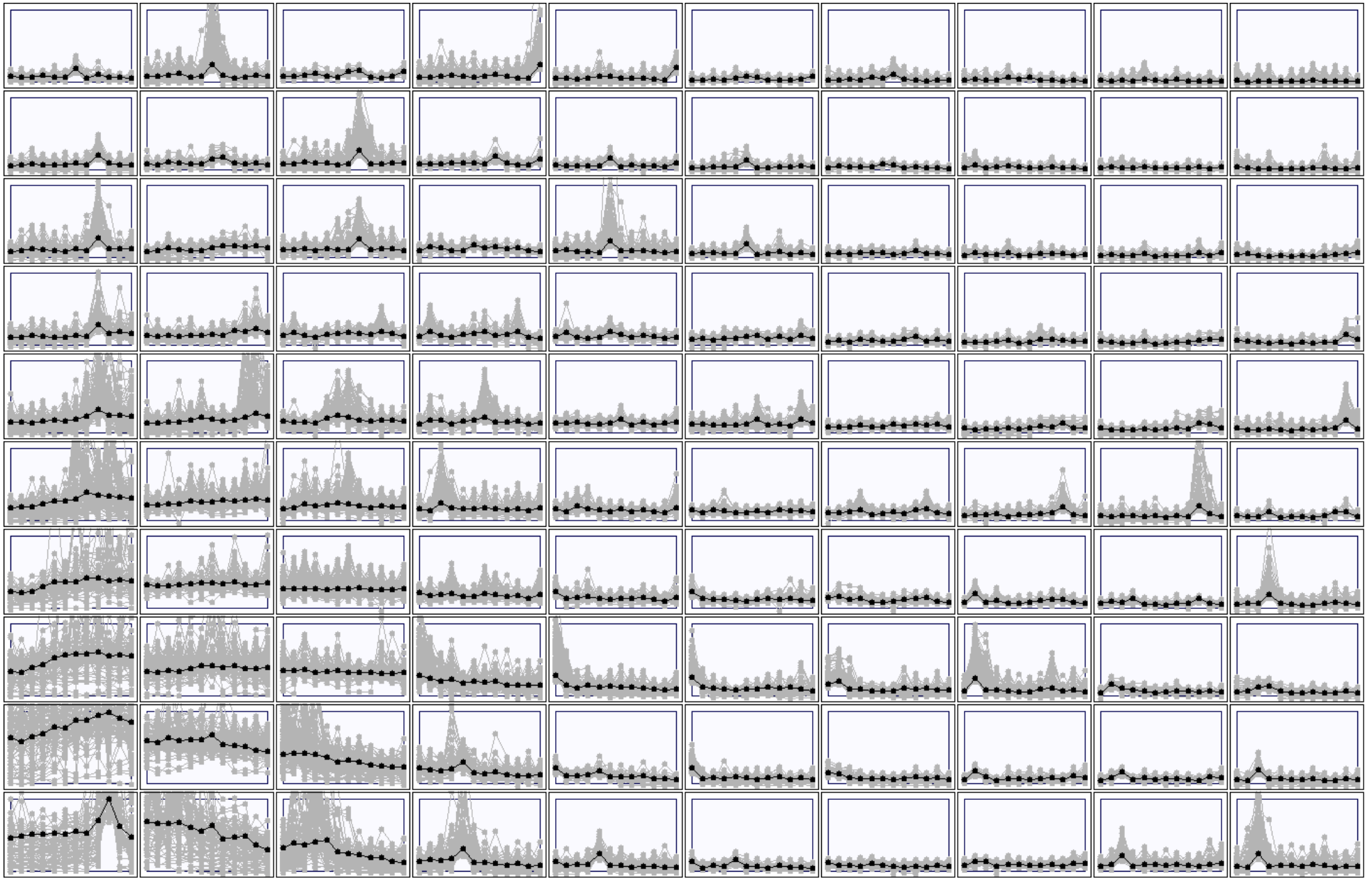
Dados em 8 dimensões.

T=480

R=1

Lr=0.1

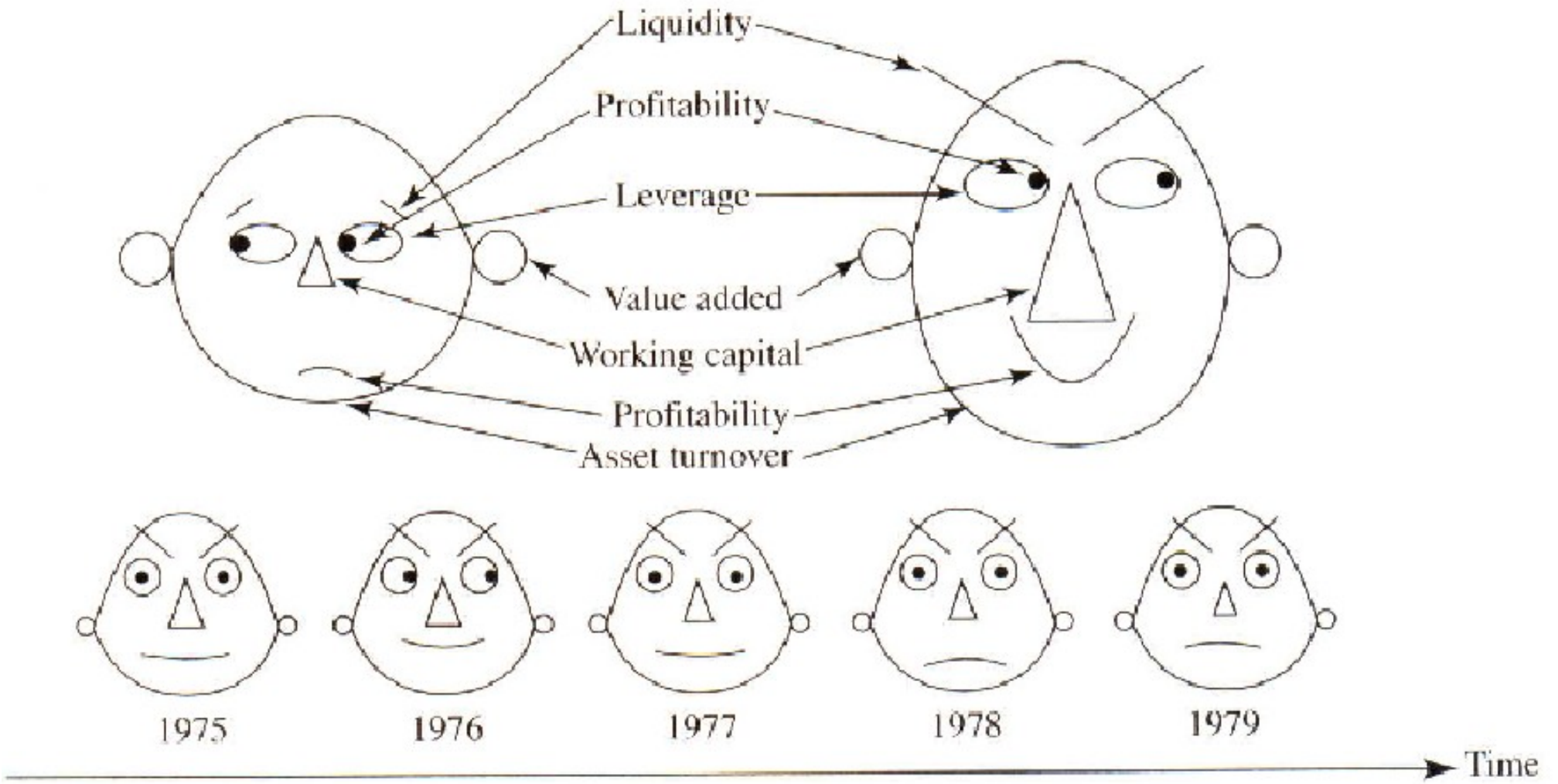
Visualização: *Self-Organizing Maps*



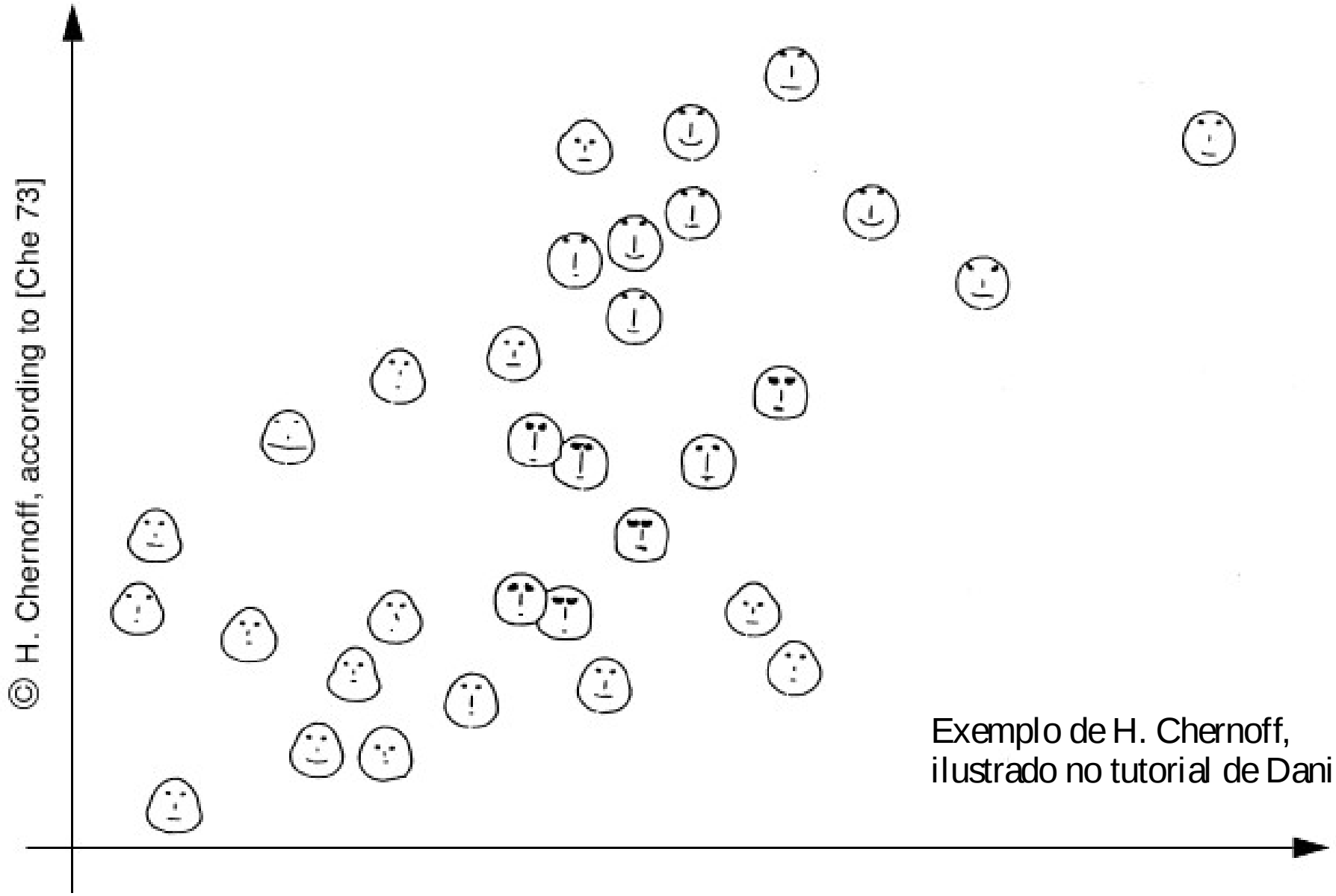
Visualização: Técnicas Baseadas em Ícones

- Idéia básica: usamos duas dimensões para mostrar ícones que representam outras dimensões adicionais.
  - Interpretação deve ser feita com legendas!
  - *Chernoff faces*: atributos das faces (geometria, olhos, excentricidade, curvaturas, etc.) representam outras dimensões.
  - *Stick figures*: dimensões adicionais mapeadas para ângulos e comprimentos de segmentos de retas.

Visualização: *Chernoff Faces*

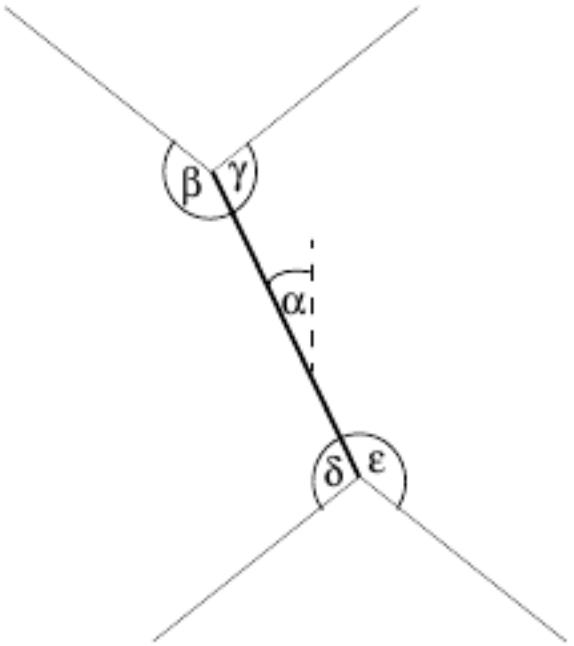


Visualização: *Chernoff Faces*

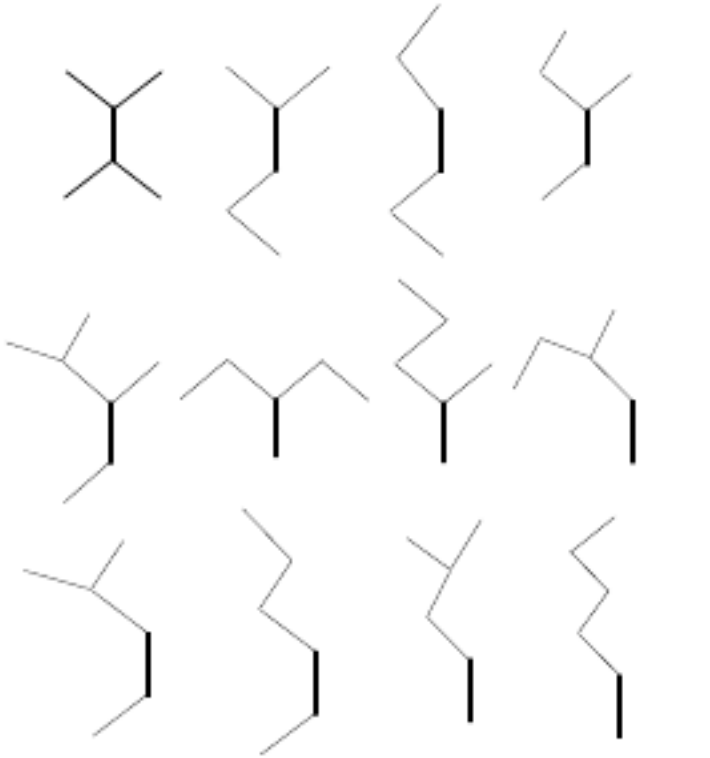


Exemplo de H. Chernoff,
ilustrado no tutorial de Daniel Keim.

Visualização: *Stick Figures*



Stick Figure Icon

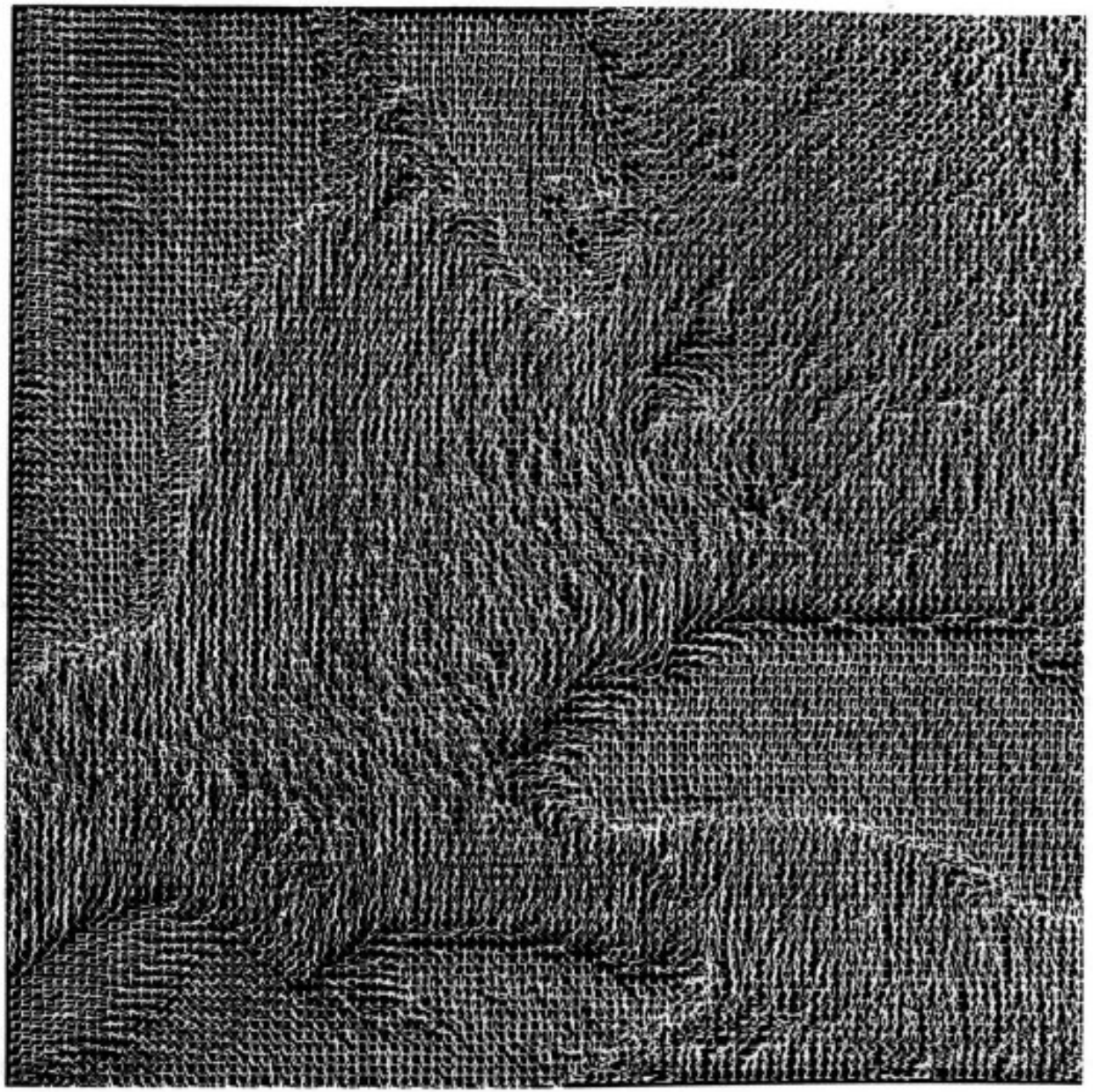


A Family of Stick Figures

- Uso de duas dimensões mais textura

Fonte: Tutorial de Daniel Keim.

Visualização: *Stick Figures*



5-dim. image
data from the
great lake region

Fonte: Tutorial de Daniel Keim.

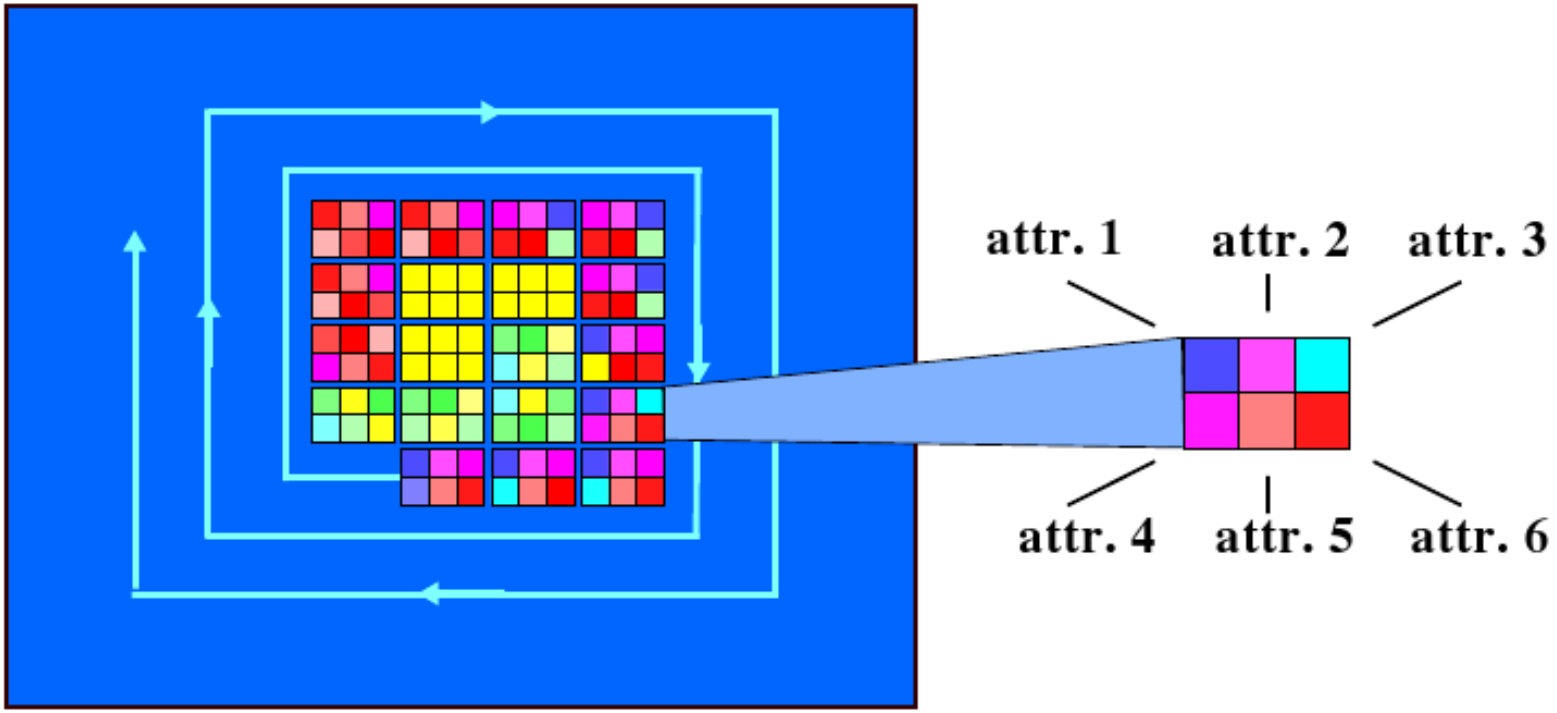
Visualização: Técnicas Baseadas em Pixels

- Idéia básica: ícones pequenos, uso de cores, geometria simples.
  - Interpretação mais instintiva, menos uso de legendas.
  - Distribui pixels em duas dimensões que podem ou não ser índices (podendo ou não causar artefatos!).
  - Existem várias maneiras de organizar pixels em duas dimensões.

Visualização: *Grouping Technique*

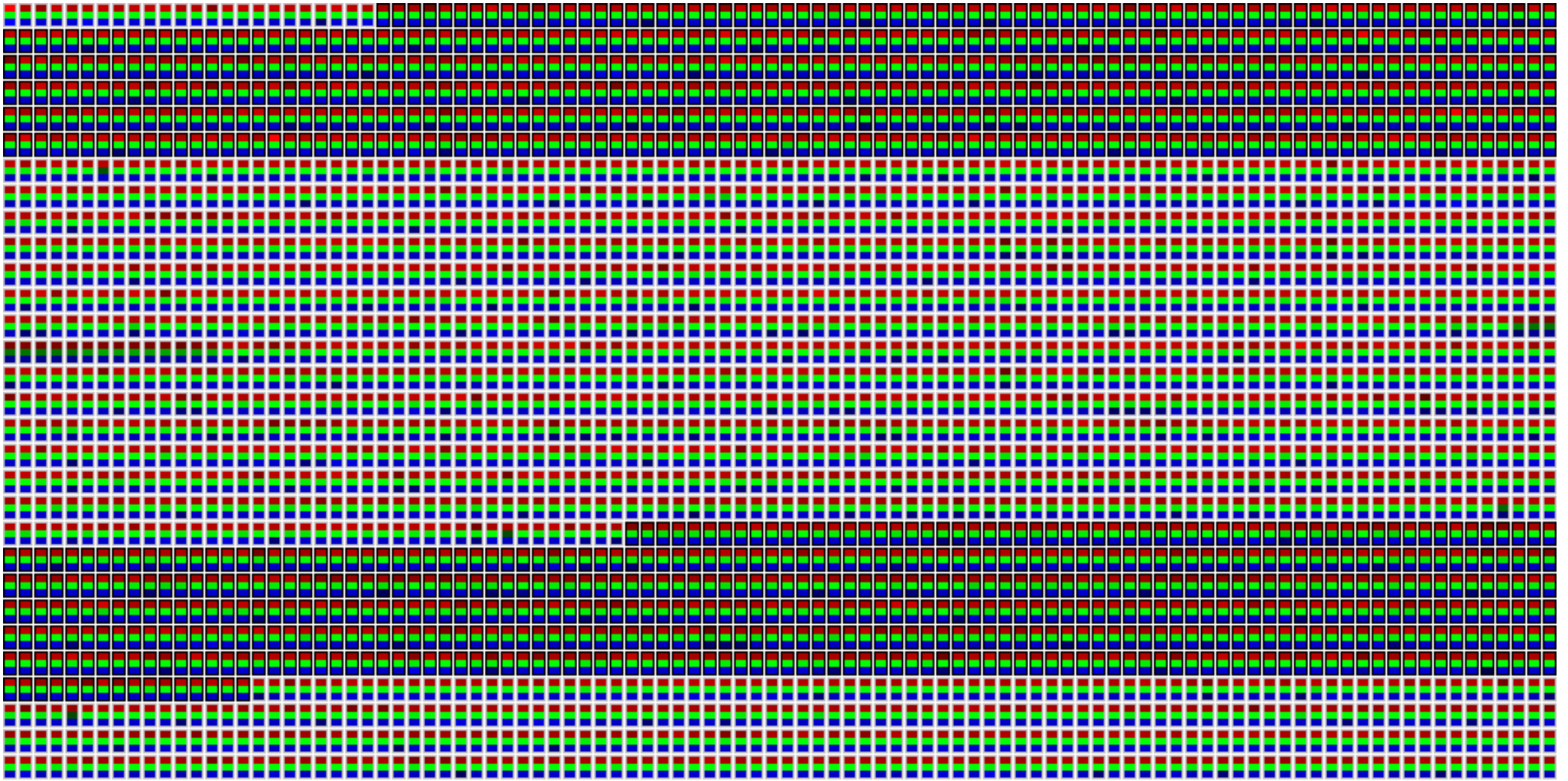
- ⇨ visualization of the data using color icons
- ⇨ color icons are array of color fields representing the attribute values
- ⇨ arrangement is query-dependent (e.g., spiral)

schematic representation of 6-dim. data



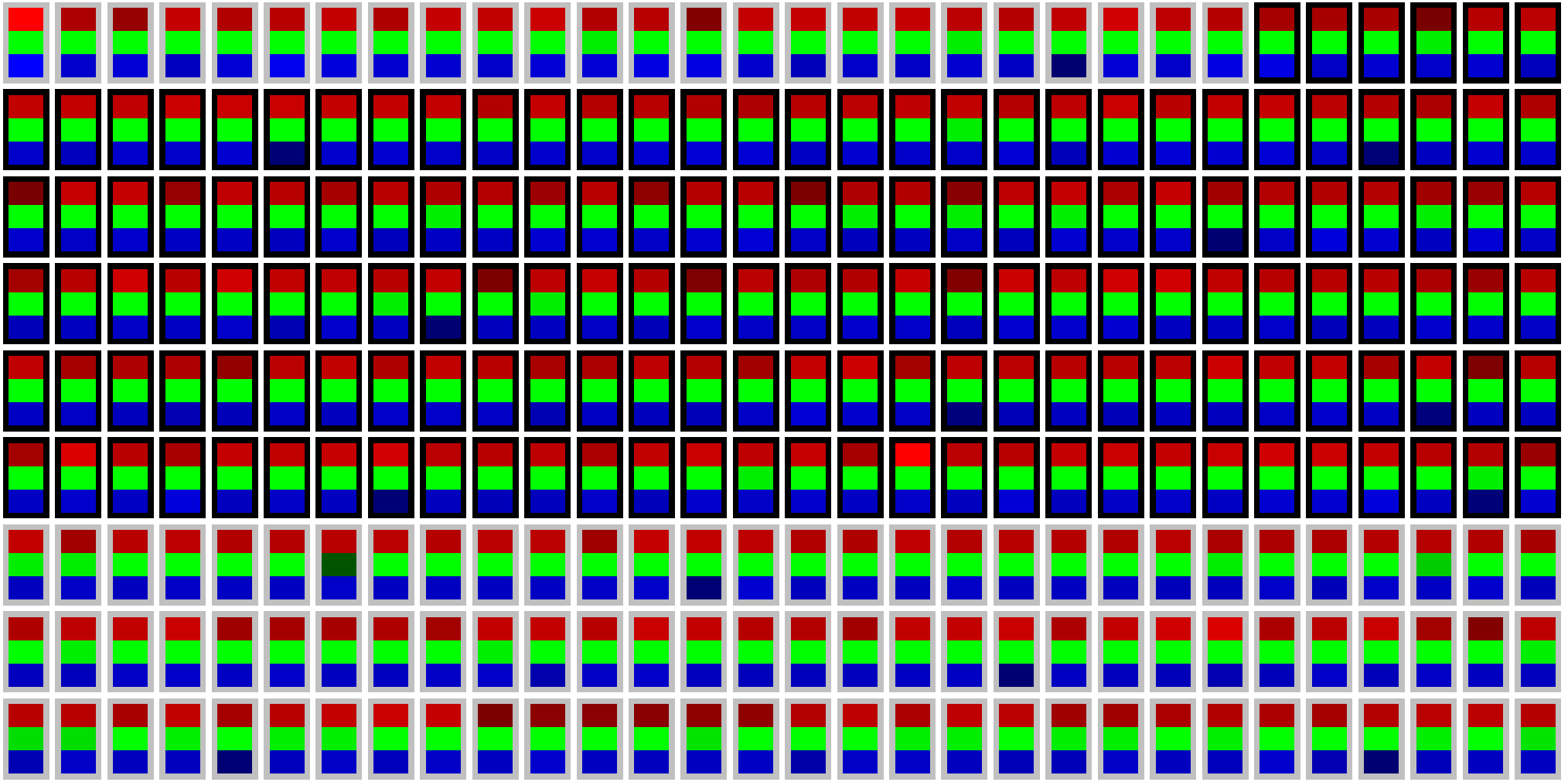
Fonte: Tutorial de Daniel Keim.

Técnicas de Visualização: *Grouping Technique*



Pacotes TCP, UDP e ICMP recebidos por honeypots em 10 dias (a cada 20 minutos).

Técnicas de Visualização: *Grouping Technique*

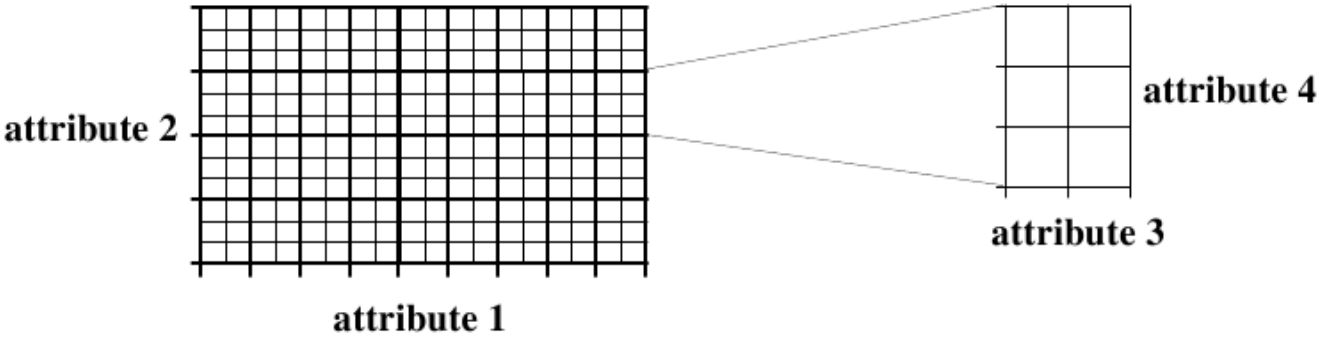


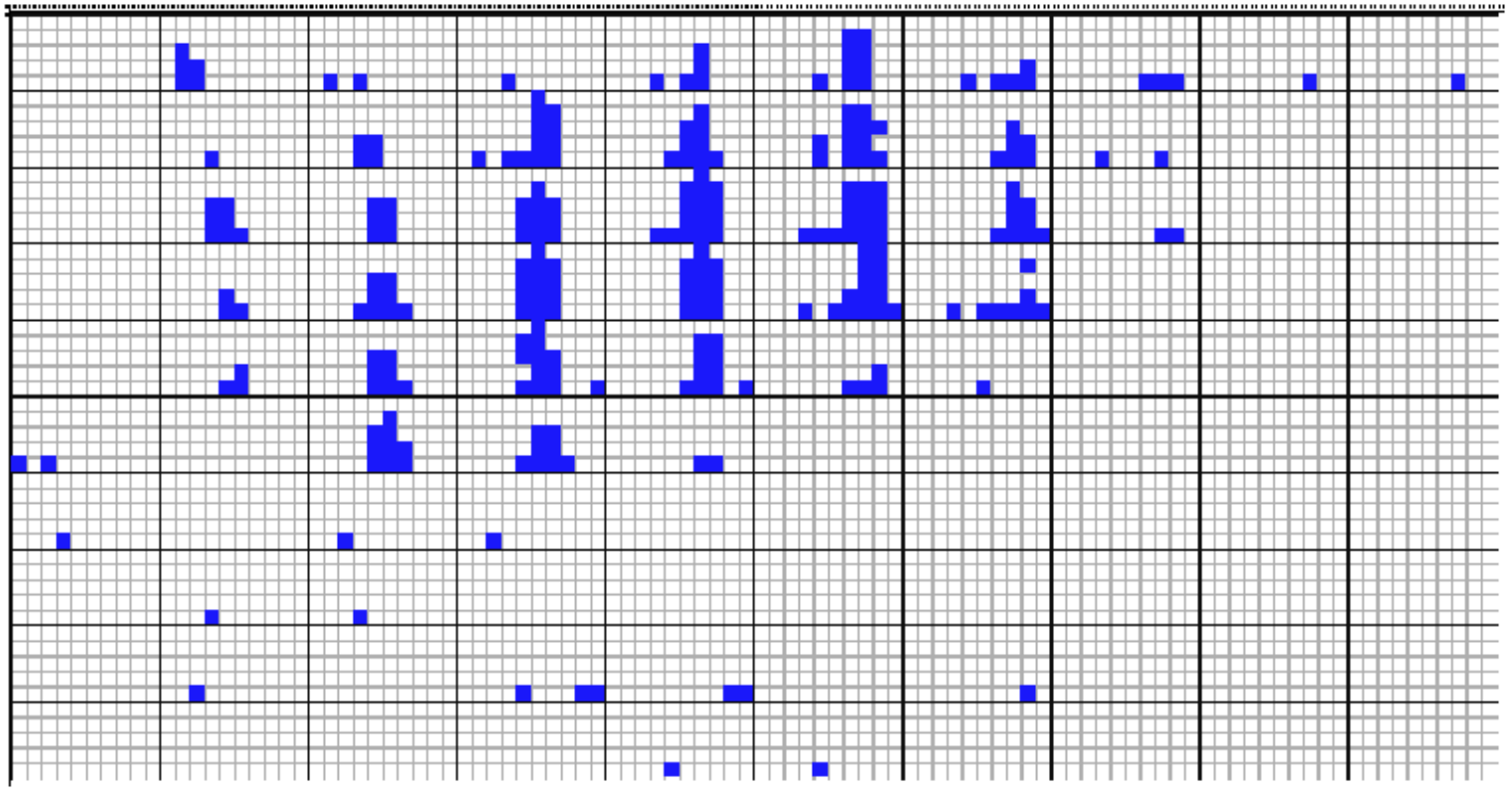
Pacotes TCP, UDP e ICMP recebidos por honeypots em 10 dias (a cada 20 minutos).

Visualização: Técnicas Hierárquicas

- Idéia básica: particionamento das dimensões em subdimensões.
  - *Dimensional Stacking:* Particionamento de N dimensões em conjuntos de 2 dimensões.
  - *Worlds-within-Worlds:* Particionamento de N dimensões em conjuntos de 3 dimensões.
  - *Treemap:* Preenche área de visualização alternando eixos X e Y.
  - *Cone Trees:* Visualização interativa de dados hierárquicos.
  - *InfoCube:* Visualização hierárquica com 3D e transparência.

Visualização: *Dimensional Stacking*



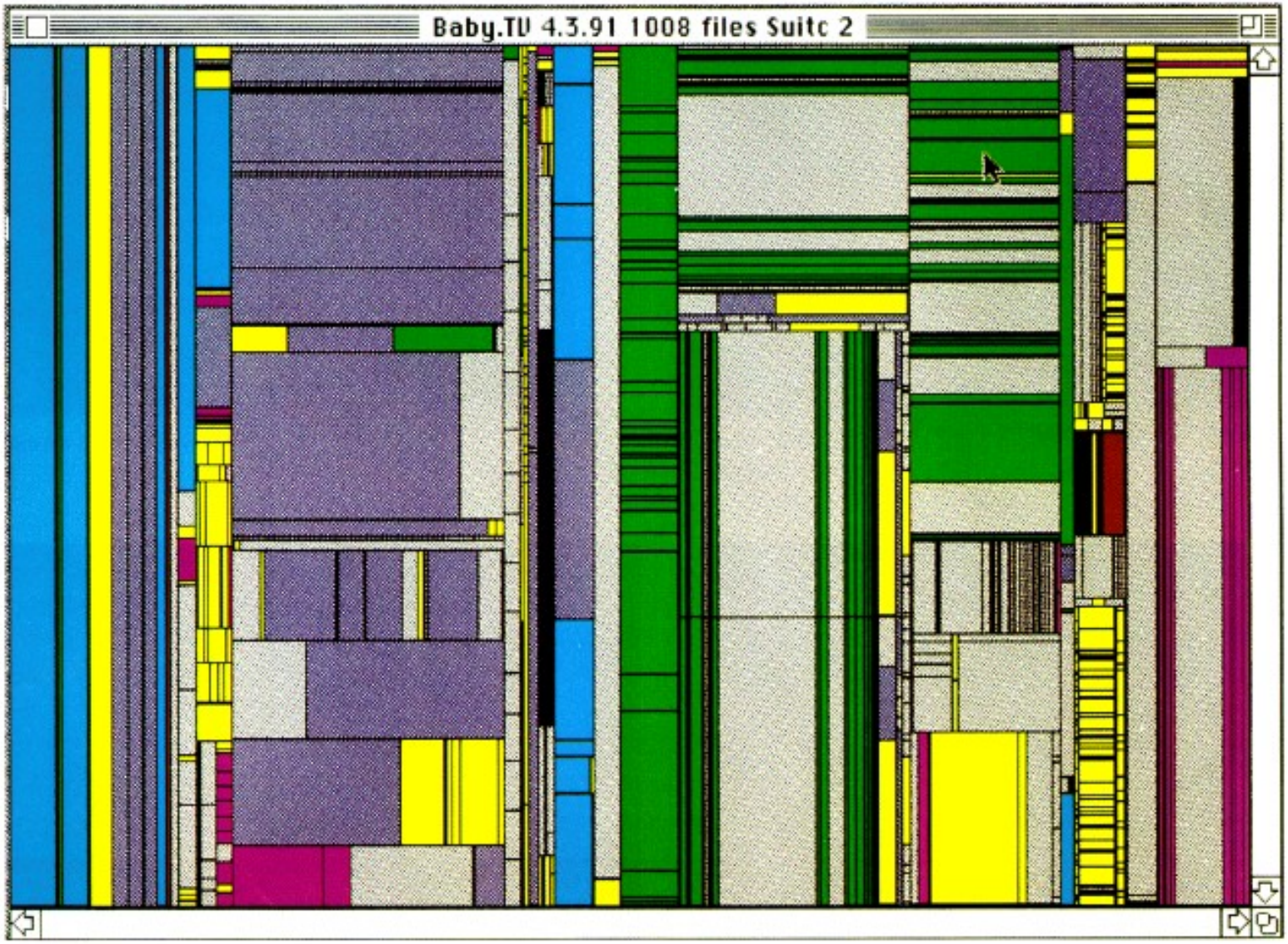


used by permission of M. Ward, Worcester Polytechnic Institute

visualization of oil mining data with longitude and latitude mapped to the outer x-, y- axes and ore grade and depth mapped to the inner x-, y- axes

Fonte: Tutorial de Daniel Keim.

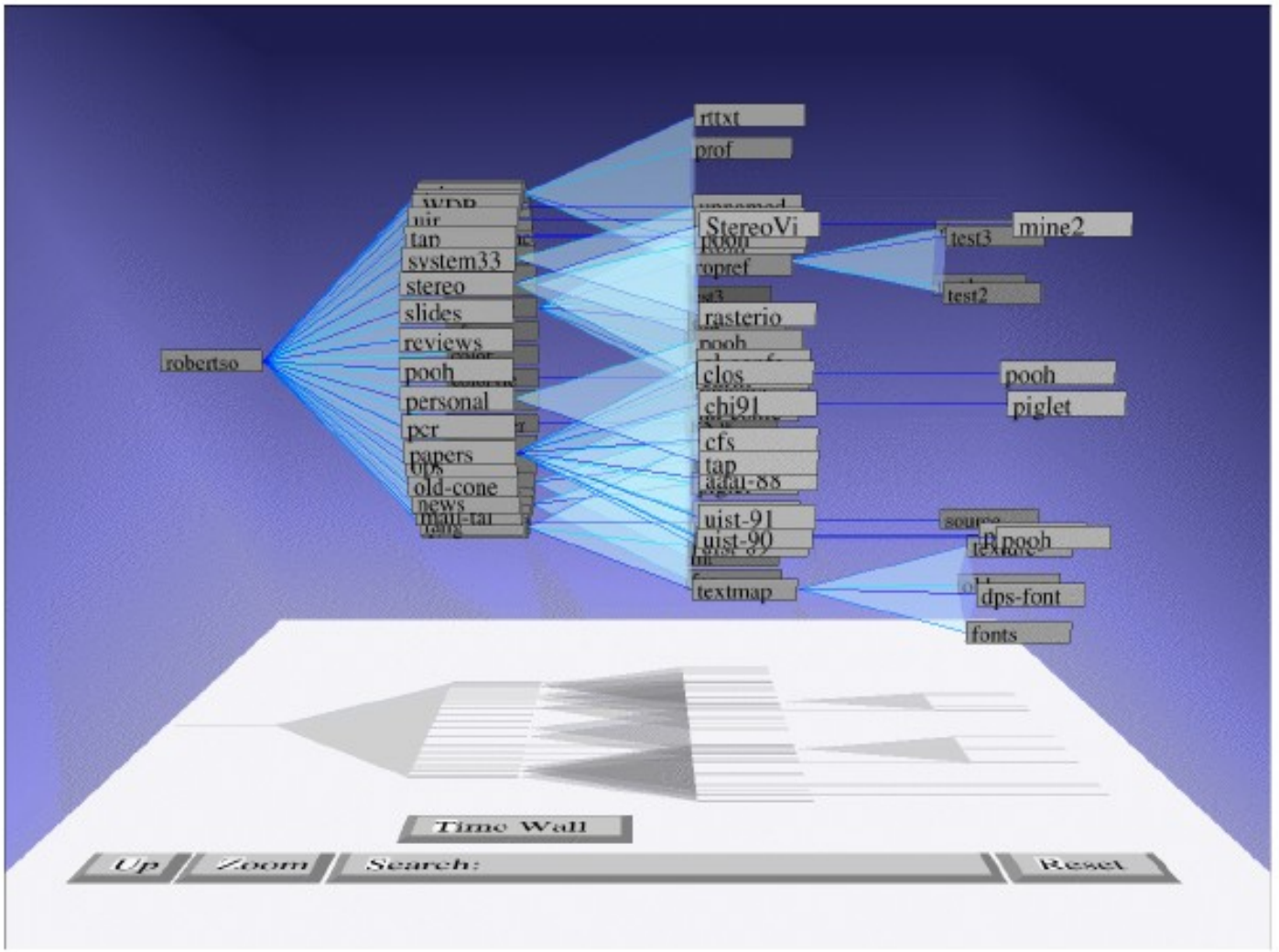
Visualização: *Treemap*



treemap of a file system containing about 1000 files

Fonte: Tutorial de Daniel Keim.

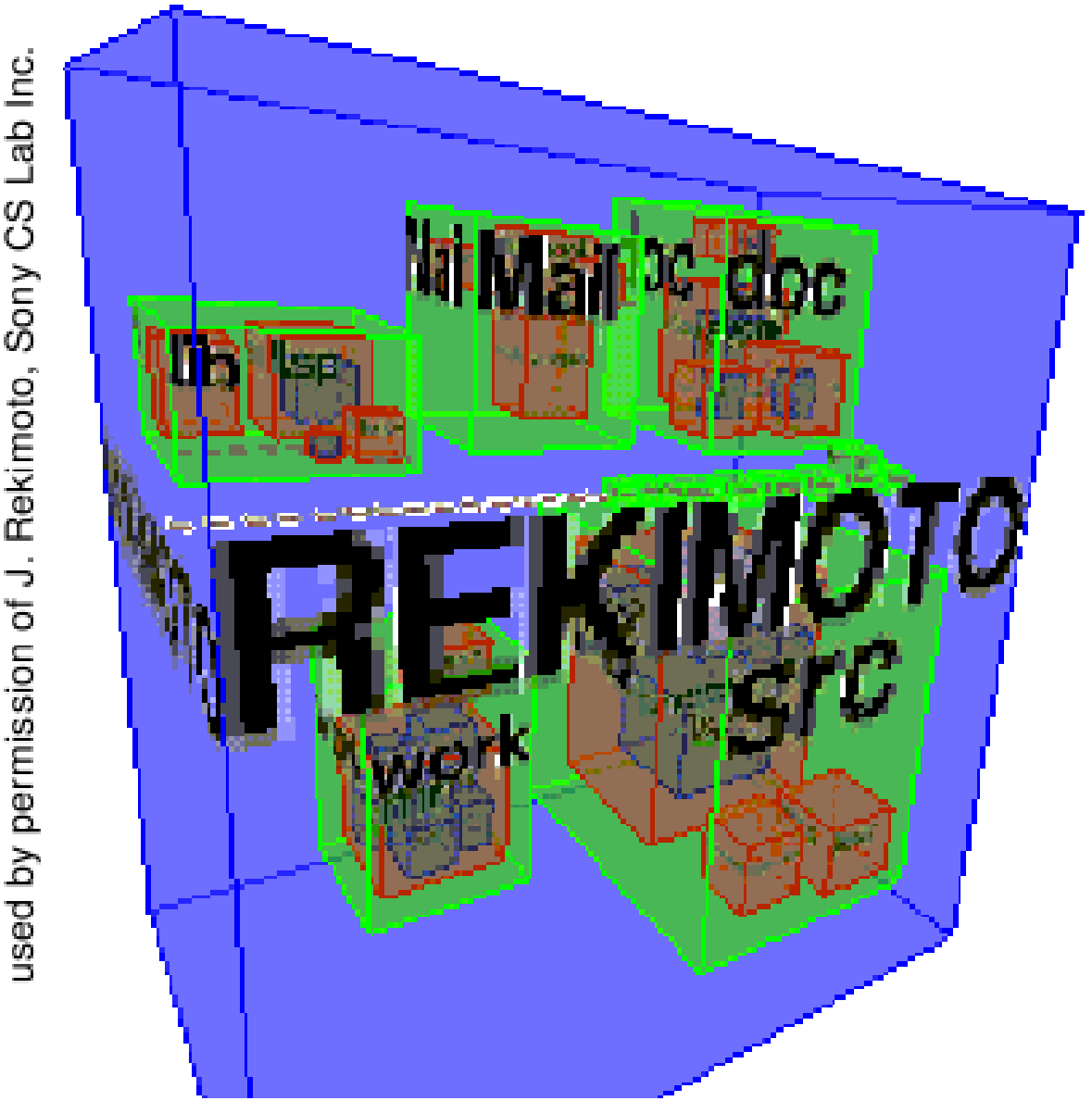
Visualização: *Cone Trees*



file system structure visualized as a cone tree

Fonte: Tutorial de Daniel Keim.

Visualização: *InfoCube*



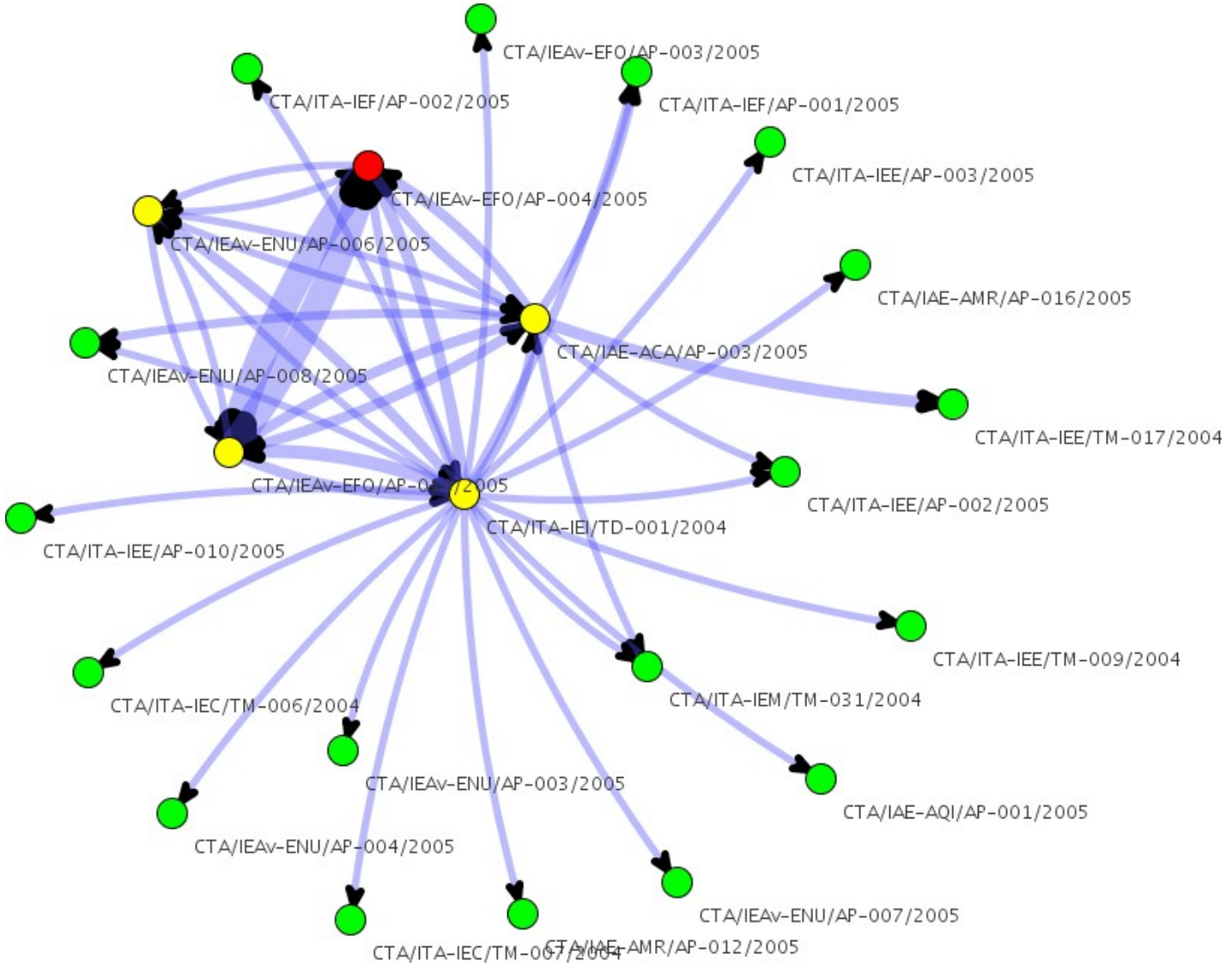
Fonte: Tutorial de Daniel Keim.

Visualização: Técnicas Baseadas em Grafos

- Idéia básica: conjunto de pontos (vértices) ligados por linhas (as arestas).
  - Representam conexões ou ligações de alguma forma.
  - Enorme variabilidade na organização geométrica dos vértices e arestas.
  - Representações gráficas diferentes para vértices e arestas.

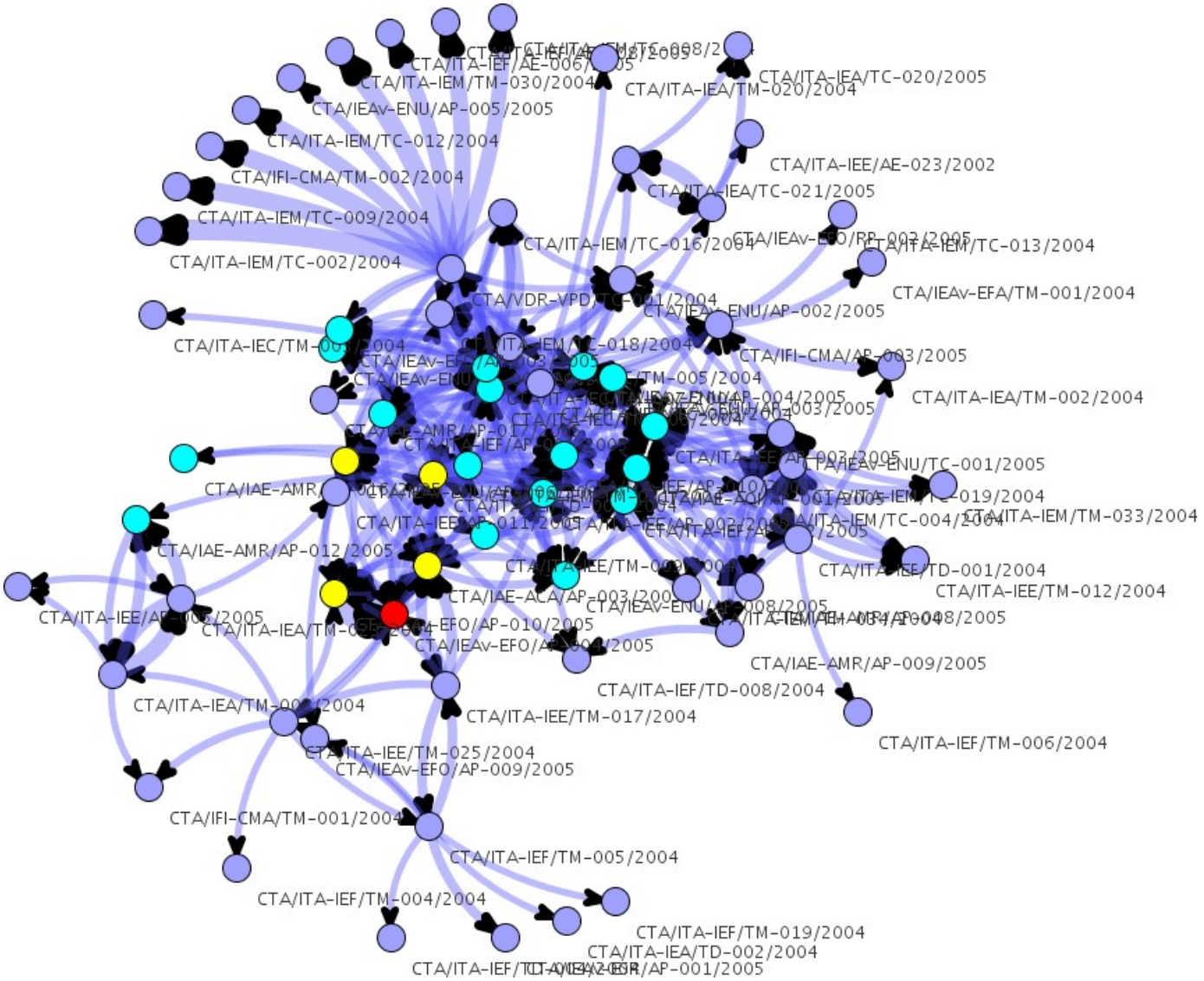
- Representação para visualização → *mineração de grafos*.

Técnicas de Visualização: Grafos



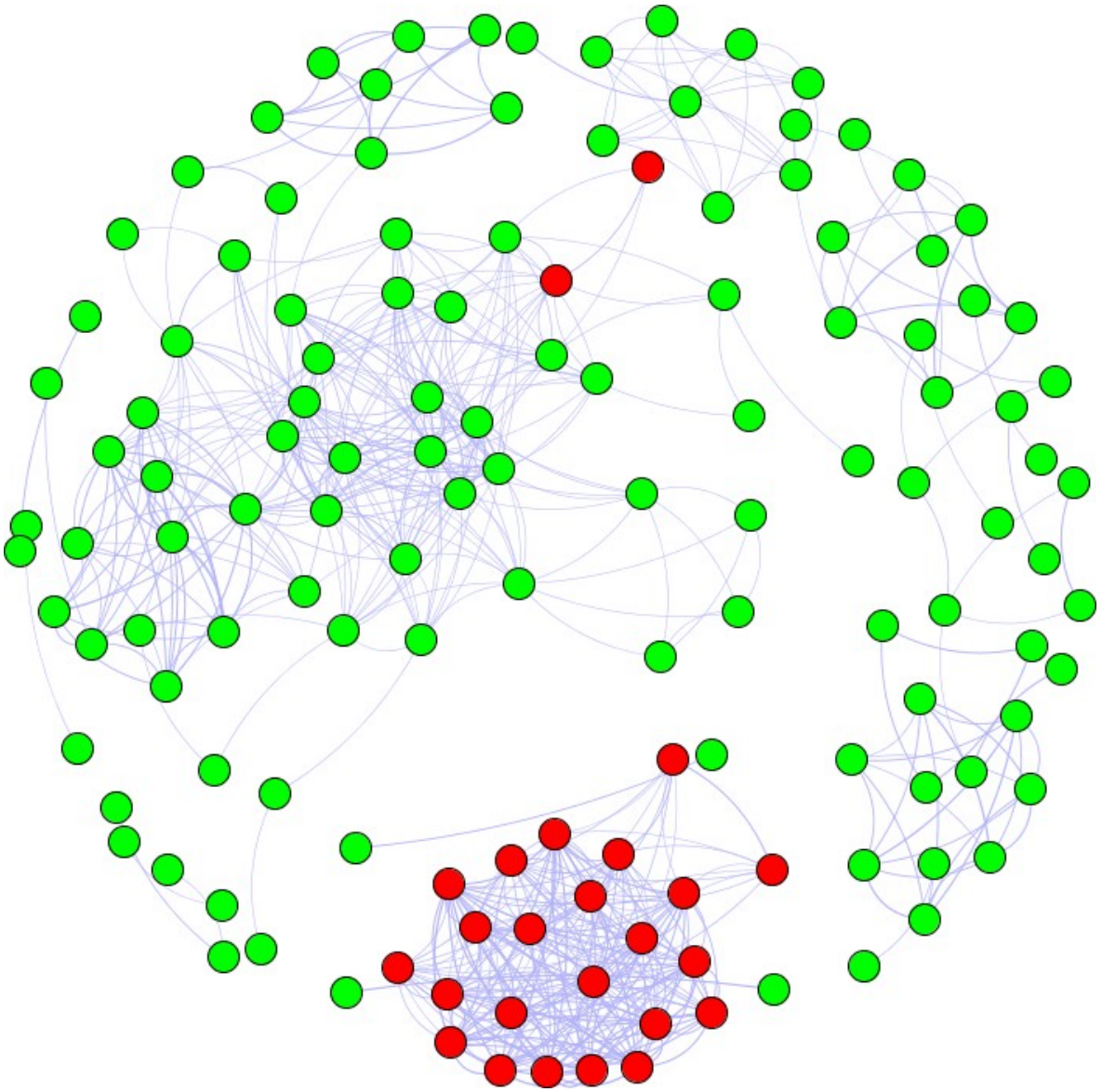
Um Sistema de Recomendação de Publicações Científicas Baseado em Avaliação de Conteúdo, Relatório Final de Alessandro Oliveira Arantes, disciplina CAP-359, INPE.

Técnicas de Visualização: Grafos



Um Sistema de Recomendação de Publicações Científicas Baseado em Avaliação de Conteúdo, Relatório Final de Alessandro Oliveira Arantes, disciplina CAP-359, INPE.

Técnicas de Visualização: Grafos

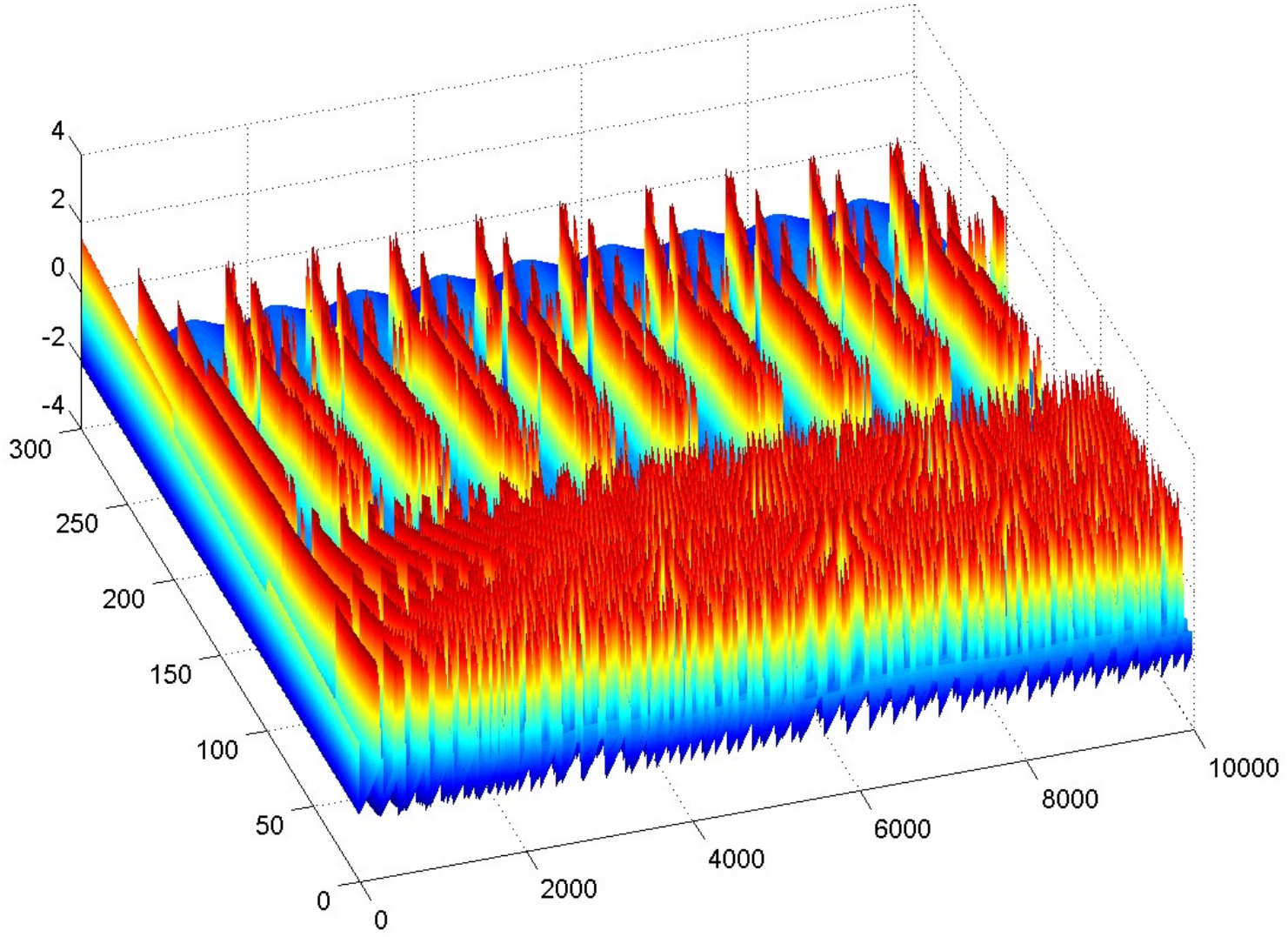


Visualização de similaridade entre *malware*.

Visualização: Técnicas Tridimensionais

- Idéia básica: recursos de computação gráfica para usar dimensão adicional na exibição dos gráficos.
  - Muito mais efetivo para *display* do que para impressão.
  - Devem ser interativos (*pan*, *zoom*, rotação, etc.)

Técnicas de Visualização: 3D

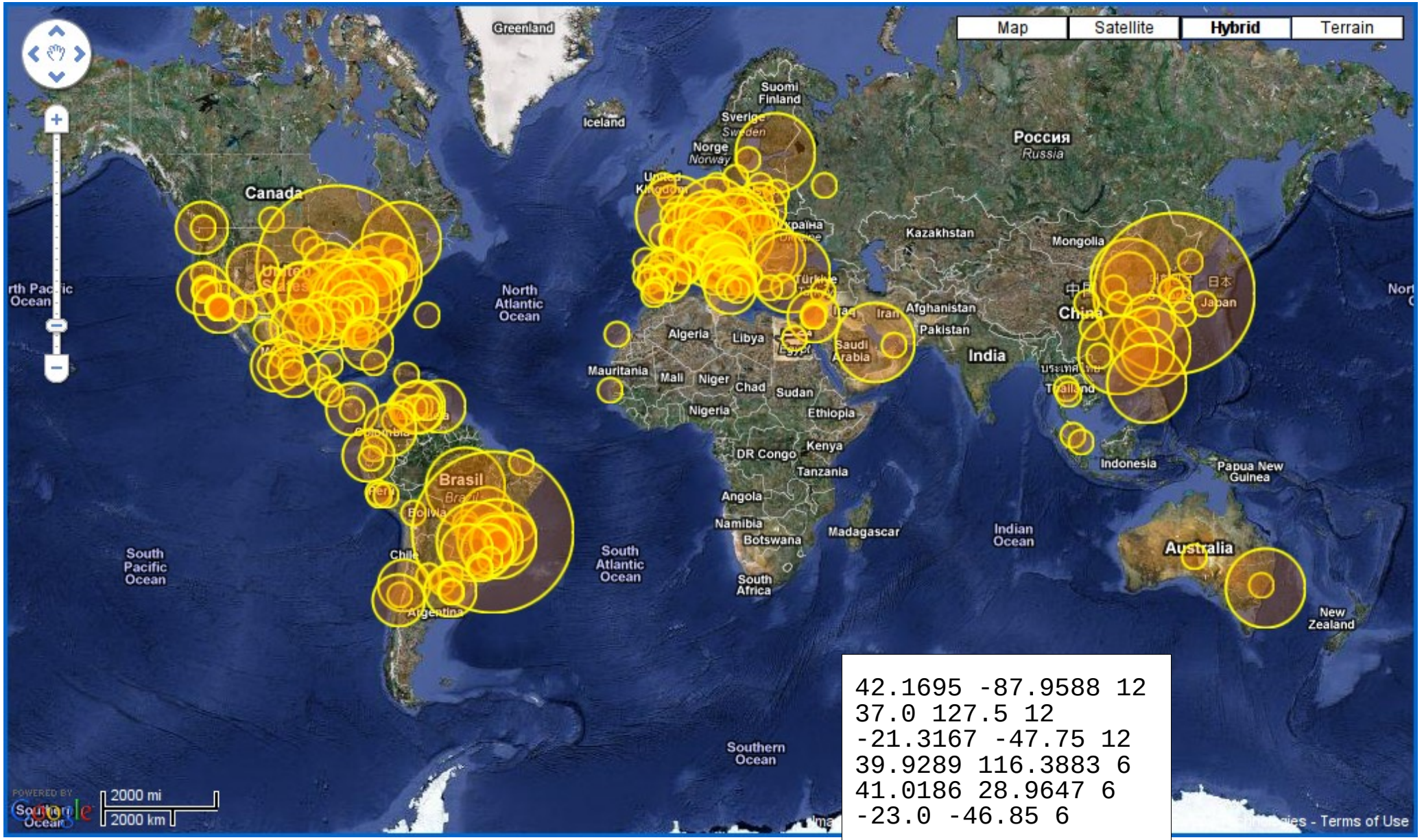


Mineração de Dados para Encontrar Motifs em Séries Temporais, Relatório Final de Rosângela Follmann Bageston, disciplina CAP-359, INPE.

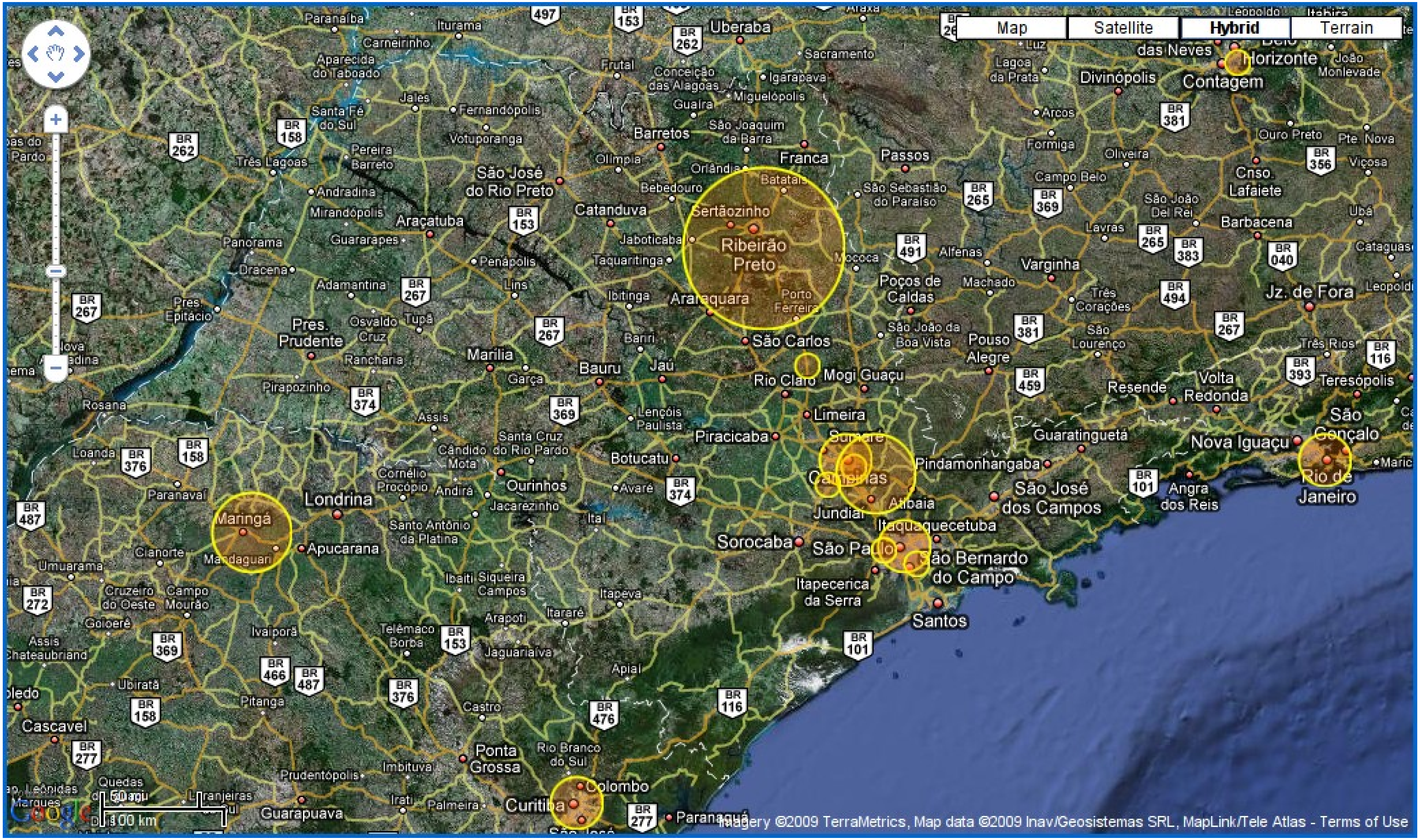
Visualização: Mapas

- Idéia básica: plotagem de elementos sobre coordenadas geográficas.
  - Valores, categorias, etc. podem ser representados como ícones, pixels, etc.
  - Devem ser interativos (*pan*, *zoom*).

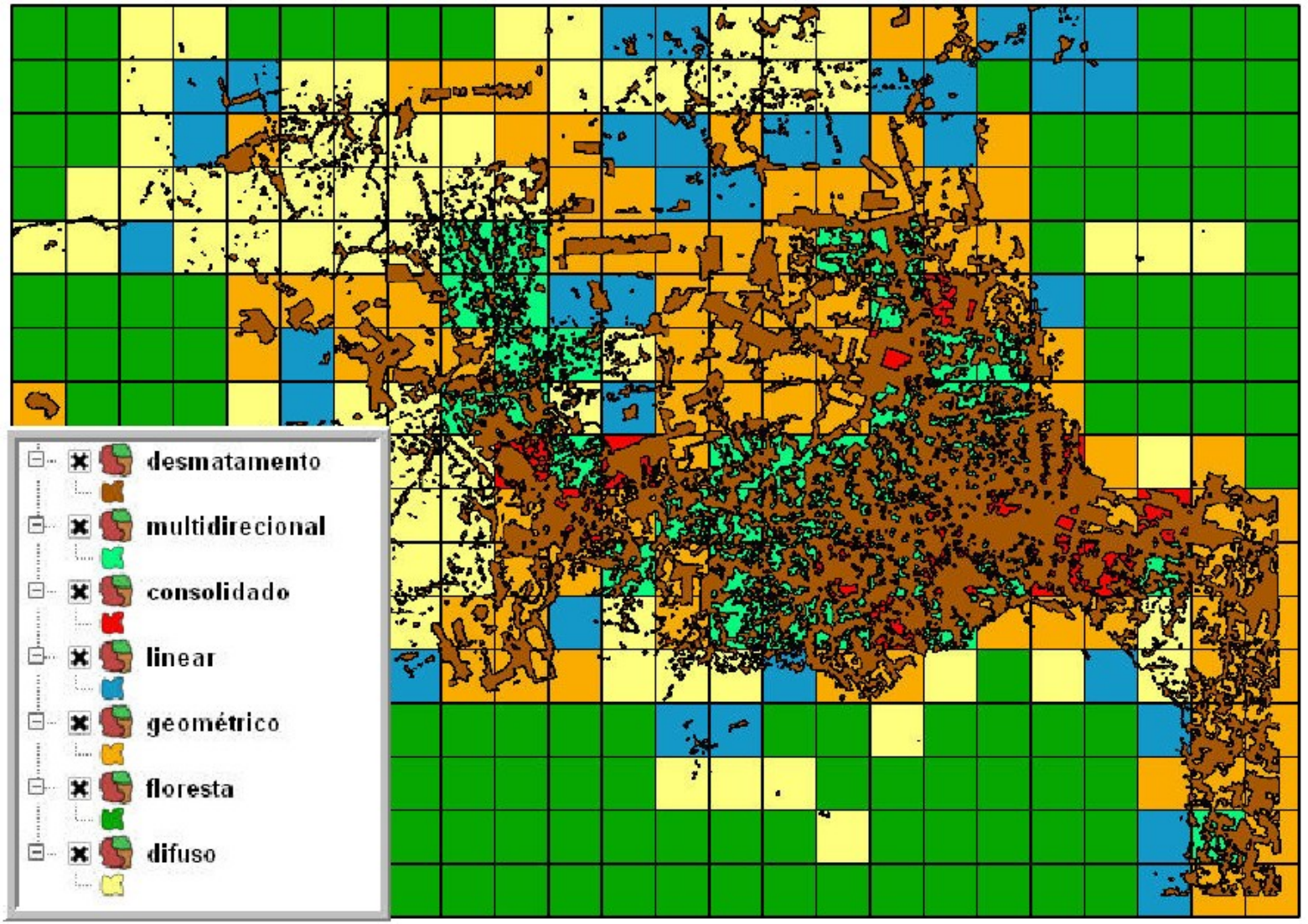
Técnicas de Visualização: Mapas



Técnicas de Visualização: Mapas

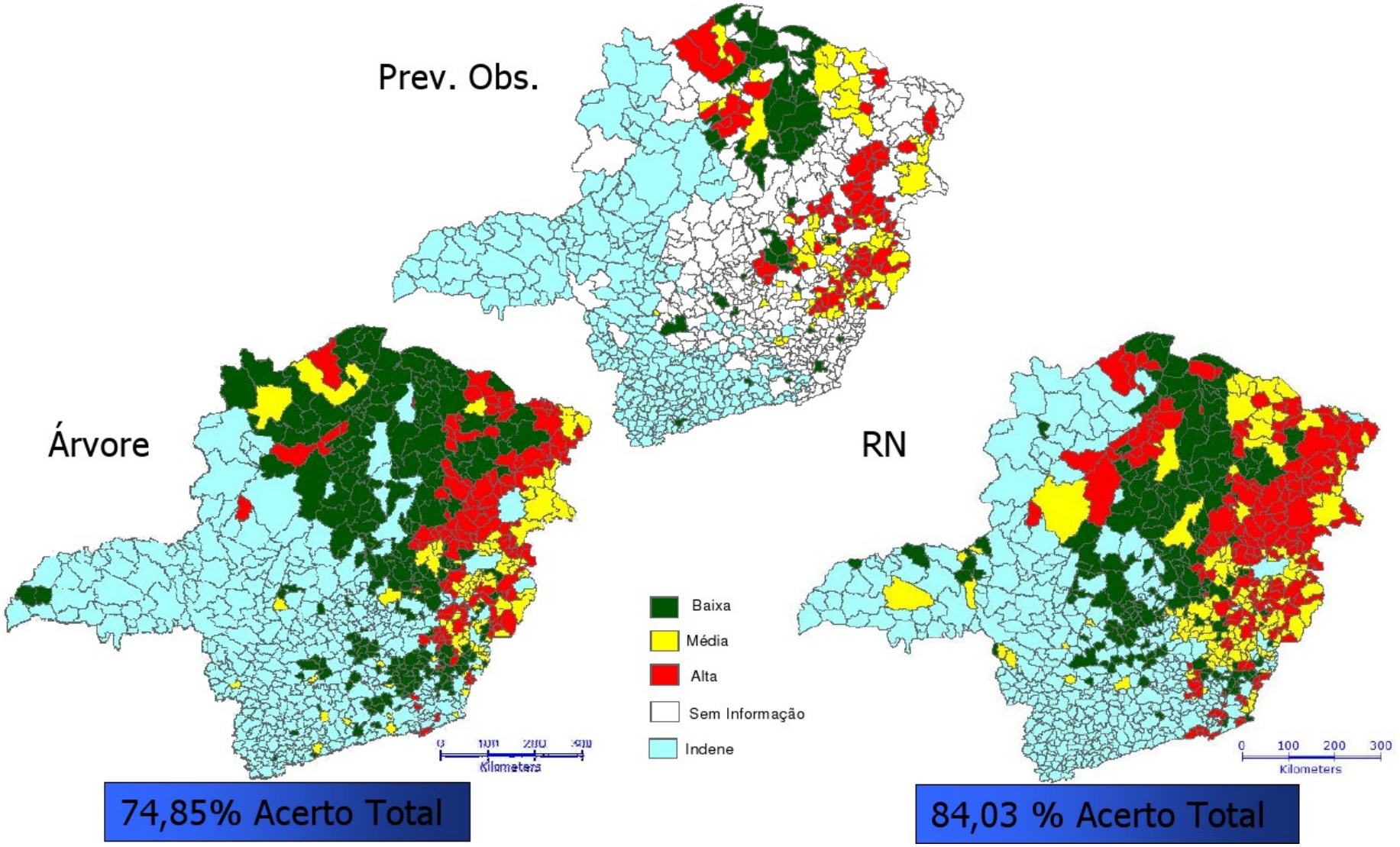


Técnicas de Visualização: Mapas



Mineração de Dados Espaciais Utilizando Métricas de Paisagem, Relatório Final de Márcio Azeredo, disciplina CAP-359, INPE.

Técnicas de Visualização: Mapas



Classificação do risco da esquistossomose no estado de Minas Gerais, Relatório Final de Flávia de Toledo Martins, disciplina CAP-359, INPE.

Técnicas de Visualização: Comentários

- Esta taxonomia é incompleta e imperfeita.
- Técnicas podem pertencer a mais de uma categoria ou mesmo usar elementos de várias.
- Implementação de técnicas deve considerar também:
  - Interatividade com o gráfico em si (seleção, *drill-down*).
  - Interatividade com os dados usados para o gráfico (filtros, *queries*).

Próximos passos...

- ***Dia 1:*** Apresentação dos conceitos de mineração de dados, motivação e alguns exemplos.
- ***Dia 2:*** Algoritmos de classificação supervisionada e aplicações.
- ***Dia 3:*** Algoritmos de classificação não-supervisionada e aplicações. Algoritmos de mineração de associações.
- ***Dia 4:*** Visualização e mineração de dados. Outros algoritmos e idéias. Onde aprender mais.

Mais informações em...

- http://www.lac.inpe.br/~rafael.santos
  - http://www.lac.inpe.br/~rafael.santos/dmapresentacoes.jsp
  - http://www.lac.inpe.br/~rafael.santos/cap359-2010.jsp
- http://www.lac.inpe.br/ELAC/index.jsp