Análise de *Logs*: Abordagens Tradicionais e por *Data Mining*

SSI 2006

8o. Simpósio Segurança em Informática

(http://www.ssi.org.br/)

Rafael Santos

Data Mining: Introdução

Data Mining

- *We are drowning in information, but starving for knowledge* – *John Naisbett*
- Crescimento explosivo na capacidade de gerar, coletar e armazenar dados:
  - Científicos: imagens, sinais.
  - Sociais: censos, pesquisas.
  - Econômicos e comerciais: transações bancárias e comerciais, compras, ligações telefônicas, transações com código de barras e RFID.
  - **Sistemas: *logs*.**

Data Mining

- Alguns exemplos de grandes volumes dados coletados:
  - **SLAC (*Stanford Linear Accelerator Center*)**: 200 terabytes/ano, 10 megabytes/segundo por 10 anos.
    2 petabytes = 2.097.152 gigabytes =∼440.000 DVDs = pilha de ∼4.4km de altura.
    Estimativa que o CERN precisará de armazenamento duas ordens de magnitude superior.
- **Walmart**: 20 milhões de transações por dia (em 1994), banco de dados de 24 terabytes.
- **INPE:** 130 terabytes de dados de imagens.

Data Mining

- ***Europe's Very Long Baseline Interferometry (VLBI)***: 16 telescópios, cada um produz 1 gigabit/segundo de dados astronômicos em sessões de 25 dias =∼ 4 terabytes.
- ***Wayback machine***: 1 petabyte, 20 terabytes/mês (55 bilhões de páginas em março de 2006).
- **Yahoo!**: 100 terabytes.
- **AT&T**: 93 terabytes.
- **Amazon**: 24 terabytes.
- **Verizon Communications**: 7 terabytes.

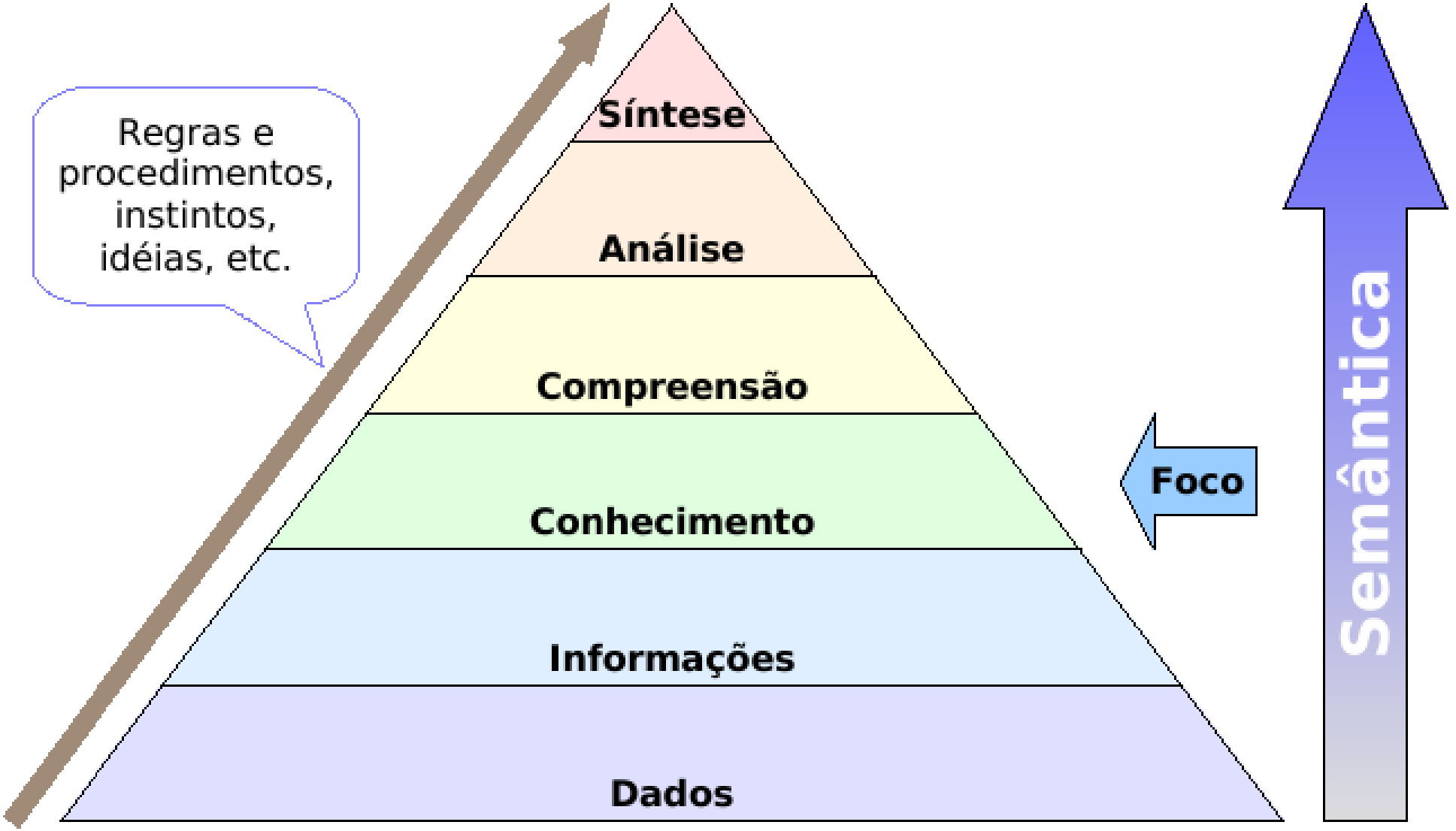
Data Mining

- Justificativas para este aumento:
  - Barateamento de componentes computacionais.
  - Exigências científicas/sociais.
  - Mudança de paradigmas!
- Temos um crescimento correspondente na capacidade de processar e analisar estes dados?
  - Quem vê todos estes dados? Alguém?
  - Análise “visual” é viável? É automatizável?
- Dados por si não valem nada!

Data Mining

- Como identificar...
  - Padrões (“X” acontece se...)
  - Exceções (isto é diferente de... por causa de...)
  - Tendências (ao longo do tempo, “Y” deve acontecer...)
  - Correlações (se “M” acontece, “N” também deve acontecer)
- O que existe de interessante nestes dados? Como definir “interessante”?
- Informação, e não dados, valem dinheiro/tempo/conhecimento!

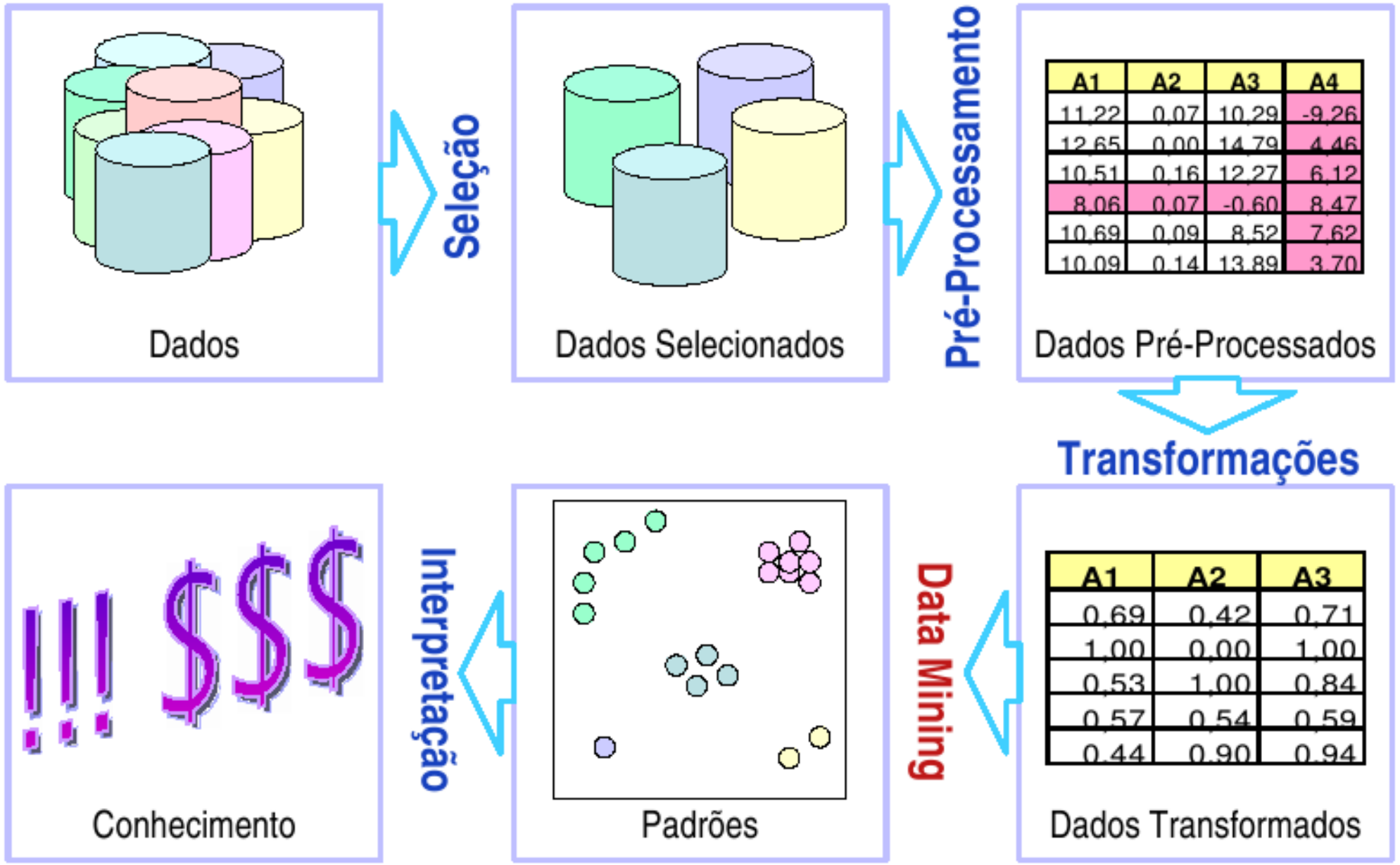
Dados, Informações e Conhecimento



Data Mining: definição

- Parte do processo de descoberta de conhecimentos em bancos de dados (*Knowledge Discovery in Databases, KDD*).
- **KDD**: Processo geral de descoberta de conhecimentos úteis previamente desconhecidos a partir de grandes bancos de dados (adaptado de Fayyad, 1996)

KDD


Dados
Seleção
Dados Selecionados
Pré-Processamento
A1 A2 A3 A4
11.22 0.07 10.29 -9.26
12.65 0.00 14.79 4.46
10.51 0.16 12.27 6.12
8.06 0.07 -0.60 8.47
10.69 0.09 8.52 7.62
10.09 0.14 13.89 3.70
Dados Pré-Processados
Transformações
Conhecimento
Interpretação
Padrões
Data Mining
A1 A2 A3
0.69 0.42 0.71
1.00 0.00 1.00
0.53 1.00 0.84
0.57 0.54 0.59
0.44 0.90 0.94
Dados Transformados

KDD

- Processo de *KDD*:
  1. Compreender o domínio da aplicação, entender as expectativas do usuário final do processo.
  2. Criar/selecionar uma coleção de dados para aplicação.
  3. Pré-processar e limpar os dados (eliminar impurezas e dados irrelevantes).
  4. Transformar (reduzir e reprojetar) os dados (encontrar atributos úteis e interessantes).
  5. Escolher a tarefa, métodos, modelos, parâmetros etc. do processo de mineração de dados e executar este processo.
  6. Interpretar os resultados, iterar se necessário.
  7. Consolidar o conhecimento adquirido, resolver conflitos, iterar se necessário.

Data Mining: definição

- É um dos passos do processo de *KDD*.
- Envolve:
  - Estatística e Matemática.
  - Computação aplicada (inteligência artificial, reconhecimento de padrões, aprendizado por máquina).
  - Visualização de dados, computação gráfica.
  - Bancos de dados.
  - Sistemas distribuídos, algoritmos paralelos, alta performance.
- O processo de *KDD* é iterativo, portanto *Data Mining* também o é.
- *Data Mining* não é mágica: ***Garbage in, garbage out***.

Data Mining: Tarefas

- Classificação:
  - Aprendizado de uma função que mapeia um dado em uma de várias classes conhecidas.
- Regressão ou Predição:
  - Aprendizado de uma função que mapeia um dado em um ou mais valores reais.
- Agrupamento (clustering):
  - Identificação de grupos de dados onde os dados tem características semelhantes com os do mesmo grupo e onde os grupos tenham características diferentes entre si.

Data Mining: Tarefas

- Sumarização:
  - Descrição do que caracteriza um conjunto de dados (ex. conjunto de regras).
- Detecção de desvios ou *outliers*:
  - Identificação de dados que deveriam seguir um padrão mas não o fazem.
- Identificação de regras de associação:
  - Identificação de fenômenos que apresentam co-ocorrência (ex. cesta de compras).

- Alguns casos de sucesso:
  - **Amazon.com**: melhoria da customização da interface com o usuário (melhoria de vendas por indicação), eliminação de fraudes.
  - **U.S. Census Bureau**: análise de dados espaciais de ensino público para determinar políticas para melhoria na educação.
  - **Columbia Interactive/Columbia University**: Análise de visitas a sites, coletando "trilhas" de usuários (como usam o site, que páginas são mais atraentes para usuários, quando usuários deixam o site) para melhorar interatividade e planejar conteúdo.
  - **Verizon Wireless**: redução de churn de 2 para menos de 1.5 por cento: de 34.6 milhões de usuários, aproximadamente 170.000 foram retidos.

- Alguns casos de sucesso:
  - **argonauten360º**: Consultoria para empresa de telecomunicações que provê serviços “call-by-call” de telefonia móvel. Estudou volume de tráfego minuto a minuto e identificou possíveis faixas de uso onde melhor competitividade podia ser alcançada.
  - **IMS America**: Empresa de pesquisa de mercado farmacêutico do mundo, mantém um banco de dados de 1.5 bilhões de prescrições de 600.000 médicos, usadas em 33.000 farmácias para identificar mudança de padrão em prescrições.
  - **Harrah’s Entertainment Inc.**: Dobrou lucros usando informações de cartões de “jogadores freqüentes”, identificando grupo de jogadores que gastavam pouco mas geravam muito lucro, criou promoções diferenciadas.

- Alguns casos de fracasso:
  - **Gazelle.com (caso simulado)**: Tentativa de segmentar e caracterizar clientes; custo de DM excedia lucro possível.
  - ***Drinking diet drinks leads to obesity***, de Piero Bonissone: confusão entre causa e conseqüência.
  - **Total Information Awareness**: Projeto do departamento de defesa dos EUA para detetar atividade terrorista: não é tecnicamente um fracasso, mas causou forte reação pública.
  - **Caso apócrifo**: *data mining* de cartões de fidelidade achou padrão específico de consumo de mulheres divorciadas (muitas divorciadas não indicavam este *status* nos cadastros).
- Existem poucos casos de fracasso publicados.

Data Mining: Técnicas

Conceitos

ARFF-Viewer - /home/rafael/Pesquisa/Andre/dataset_fev2005_ATAQTRIVIAIS.arff

File Edit View

dataset_fev2005_ATAQTRIVIAIS.arff

Relation: network_logs

No.	cli_packets-avg Numeric	srv_packets-avg Numeric	cli_packets-num Numeric	srv_packets-num Numeric	very_small_pkt_percent Numeric	traffic_direction Numeric	cli_bytes Numeric	srv_bytes Numeric	session_time Numeric	class Nominal
1	76.8	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	1.644915...	ataque
2	76.8	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	0.592979...	ataque
3	76.8	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	0.090904...	ataque
4	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	1.312824...	ataque
5	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	0.631538...	ataque
6	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	0.769995...	ataque
7	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	0.631952...	ataque
8	95.75	200.8571428...	4.0	7.0	81.8181818181818	-1.0	383.0	1406.0	7888.692...	ataque
9	101.833333...	191.625	6.0	8.0	78.5714285714286	-1.0	611.0	1533.0	7913.886...	ataque
10	93.4	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	0.925822...	ataque
11	77.8333333...	25.0	0.0	2.0	100.0	-1.0	0.0	50.0	4.697239...	ataque
12	78.0	208.8333333...	6.0	6.0	83.3333333333333	-1.0	468.0	1253.0	12.06194...	ataque
13	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	1.486647...	ataque
14	78.0	25.0	0.0	2.0	100.0	-1.0	0.0	50.0	1.697881...	ataque
15	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	0.736618...	ataque
16	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	0.615198...	ataque
17	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	1.001753...	ataque
18	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	0.636913...	ataque
19	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	0.499177...	ataque
20	78.0	25.0	0.0	1.0	100.0	-1.0	0.0	25.0	5.795408...	ataque

Colunas

Entradas, campos, registros, etc.

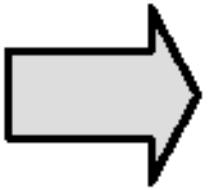
Pré-Processamento

- Etapa crucial de mineração de dados!
- No caso particular de análise de *logs*, é indispensável, custosa e problemática!
- Potencialmente deve ser feita e refeita juntamente com a aplicação das diversas técnicas.
- Algumas técnicas:
  - Transformação dos atributos
  - Enriquecimento dos dados
  - Redução do conjunto de atributos
  - Redução de dados

Pré-Processamento: Transformação

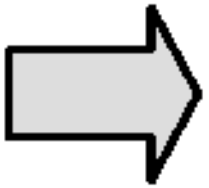
- Manipulação dos tipos dos atributos.
  - Conversões, normalizações, etc.

Outlook
sunny
sunny
rainy
rainy
sunny
rainy
overcast



isSunny	isOvercast	isRainy
TRUE	FALSE	FALSE
TRUE	FALSE	FALSE
FALSE	FALSE	TRUE
FALSE	FALSE	TRUE
TRUE	FALSE	FALSE
FALSE	FALSE	TRUE
FALSE	TRUE	FALSE

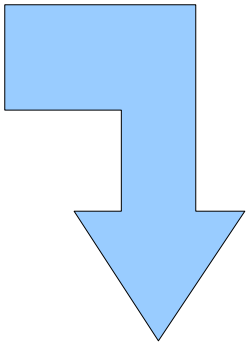
Temperature
85
80
68
65
69
75
75



>60	>70	>80	>90
TRUE	TRUE	TRUE	FALSE
TRUE	TRUE	TRUE	FALSE
TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE
TRUE	TRUE	FALSE	FALSE
TRUE	TRUE	FALSE	FALSE

Pré-Processamento: Transformação

Loja	Caixa	Transação	Compras
03	05	011672	'PAO FRANCES'
03	05	011673	'PAO FRANCES'
03	06	010169	'PAO FRANCES'
03	05	011674	'PAO FRANCES', 'FLV PIMENTAO VERDE', 'LEITE PAST. SERRAMAR S', 'DANONE DANETTE CHOCOLA', 'ADOCANTE DOCE MENOR LI'
01	14	003752	'PAO FRANCES', 'LEITE PAST. SERRAMAR S'
01	14	003758	'BEB. REF.COCA COLA 1L', 'PAO FRANCES', 'LEITE PAST. PAULI TIPO'
01	13	003001	'LEITE PAST. PAULI TIPO', 'PAO FRANCES'
03	05	011685	'PAO FRANCES', 'PAO FRANCES'
01	14	003764	'ACUCAR REFINADO UNIAO', 'FEIJAO PRETO TARUMA 1K', 'PAO FRANCES'
03	05	011688	'PAO FRANCES'
01	14	003765	'BISC. TRIUNFO C.CRACKE', 'BISC. BAUD.GULOSOS 170', 'PAO FRANCES', 'ACUCAR REFINADO A. ALE', 'MORTADELA MARBA'
03	06	010188	'PAO FRANCES', 'ACUCAR REFINADO A. ALE'



Loja	Caixa	Trans.	PF	FLVPV	LPSS	DDC	ADML	RFCC1	LPPT	ARU	FPT1	...
03	05	011672	T	F	F	F	F	F	F	F	F	...
03	05	011673	T	F	F	F	F	F	F	F	F	...
03	06	010169	T	F	F	F	F	F	F	F	F	...
03	05	011674	T	T	T	T	T	F	F	F	F	...
01	14	003752	T	F	T	F	F	F	F	F	F	...
01	14	003758	T	F	F	F	F	T	T	F	F	...
01	13	003001	T	F	F	F	F	F	T	F	F	...
03	05	011685	T	F	F	F	F	F	F	F	F	...
01	14	003764	T	F	F	F	F	F	F	T	T	...
03	05	011688	T	F	F	F	F	F	F	F	F	...
01	14	003765	T	F	F	F	F	F	F	F	F	...
03	06	010188	T	F	F	F	F	F	F	F	F	...

Pré-Processamento

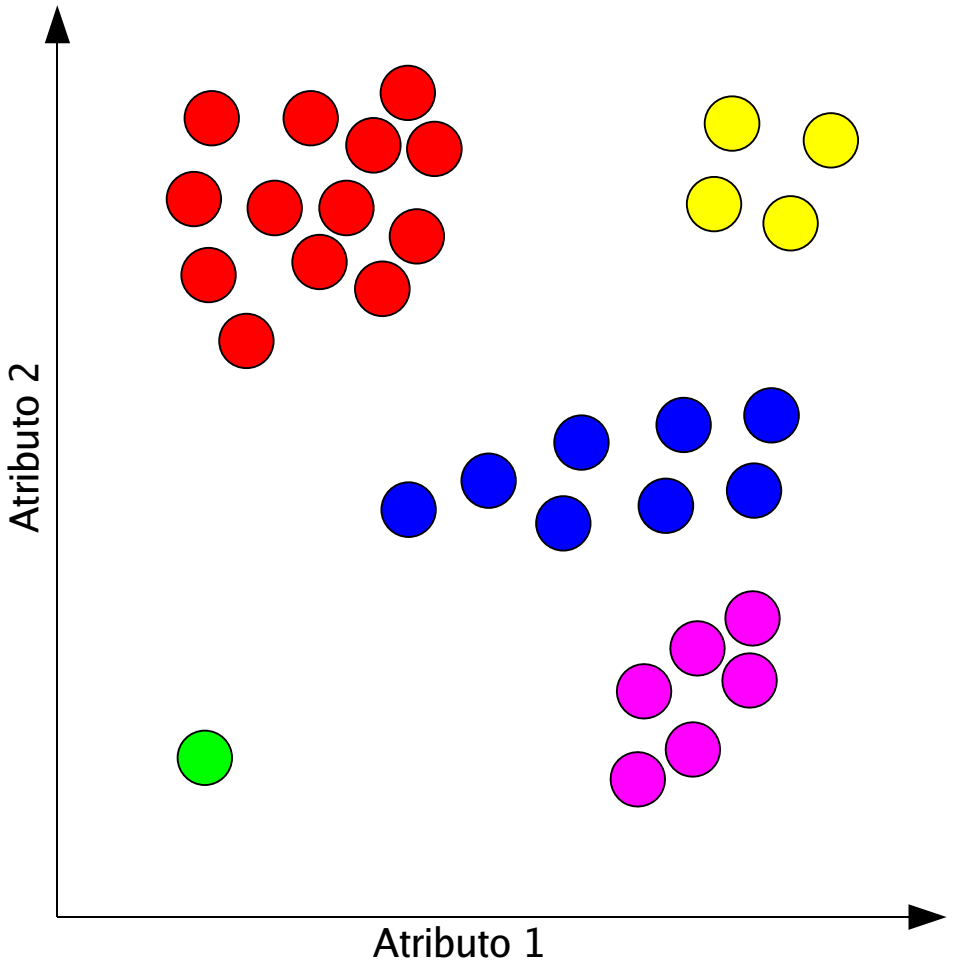
- Enriquecimento: alteração/aumento do número de atributos.
  - Exemplo: IPs podem ser mapeados para regiões geográficas.
  - Documentos em um *log* de *httpd* podem ser verificados para informações adicionais.
  - Informações auxiliares podem ser integradas, *logs* podem ser relacionados.
- Redução do conjunto de atributos.
  - Atributos irrelevantes podem ser retirados ou transformados.
  - Exemplo: diferentes representações de tempo.
  - Pode ser **muito** complexo determinar que atributos são relevantes ou não (usar conhecimento do domínio da aplicação!)

Pré-Processamento

- Redução do conjunto de dados.
  - Remover dados redundantes ou irrelevantes.
  - Como determinar o que é irrelevante?
  - Uma alternativa quando temos muitas entradas com atributos incompletos ou pouco confiáveis.
  - Pode facilitar aplicação de alguns algoritmos: segmentação do problema e dos dados relacionados.

Alguns conceitos

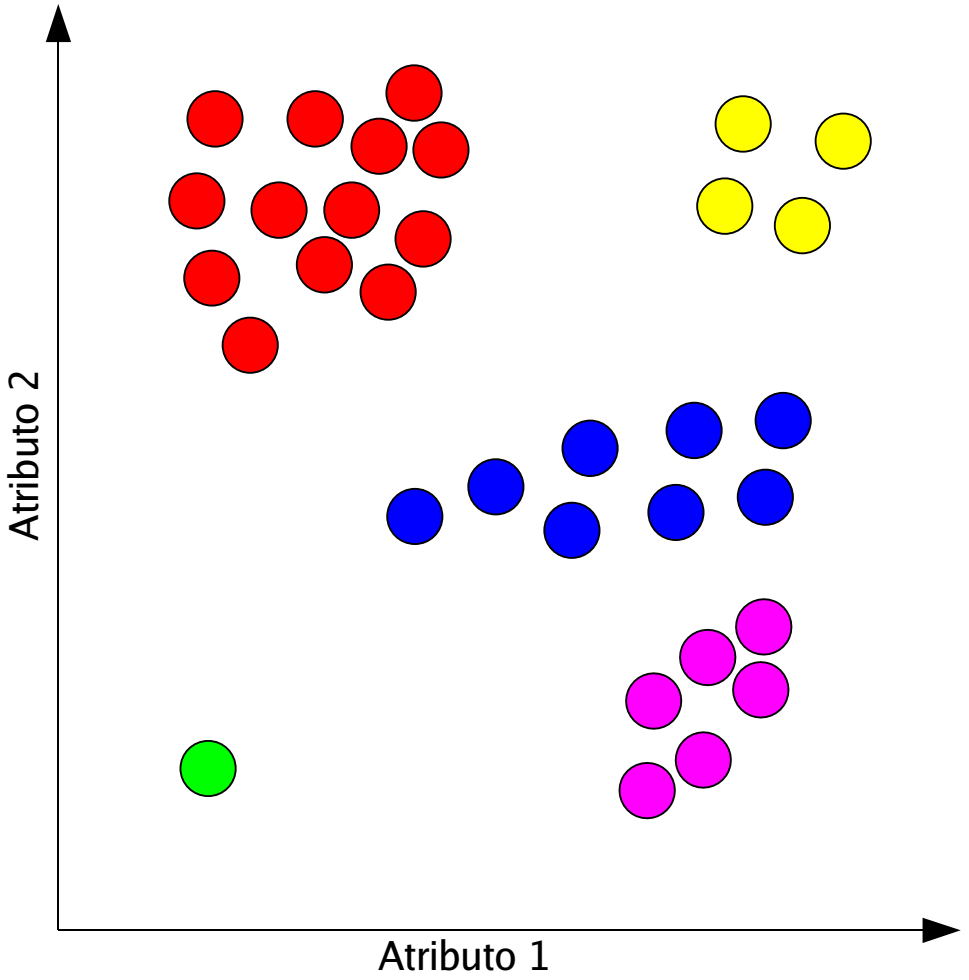
- Espaço de atributos:
  - Para cada evento ou entrada em um *log* temos medidas (numéricas ou não) de atributos.
  - Estas medidas podem ser visualizadas como pontos em gráficos N-dimensionais, onde os eixos são os atributos.
  - Visualização é um artifício, usaremos o conceito de **proximidade como semelhança**.

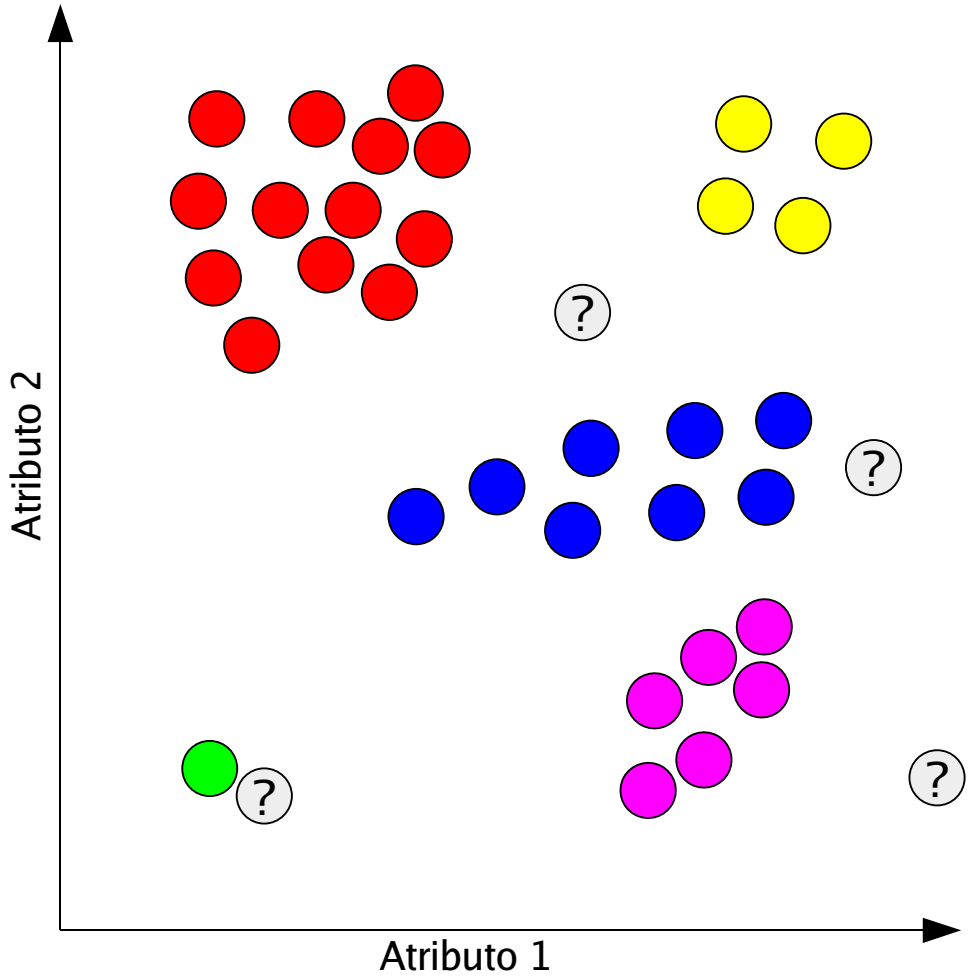


Técnicas: Classificação

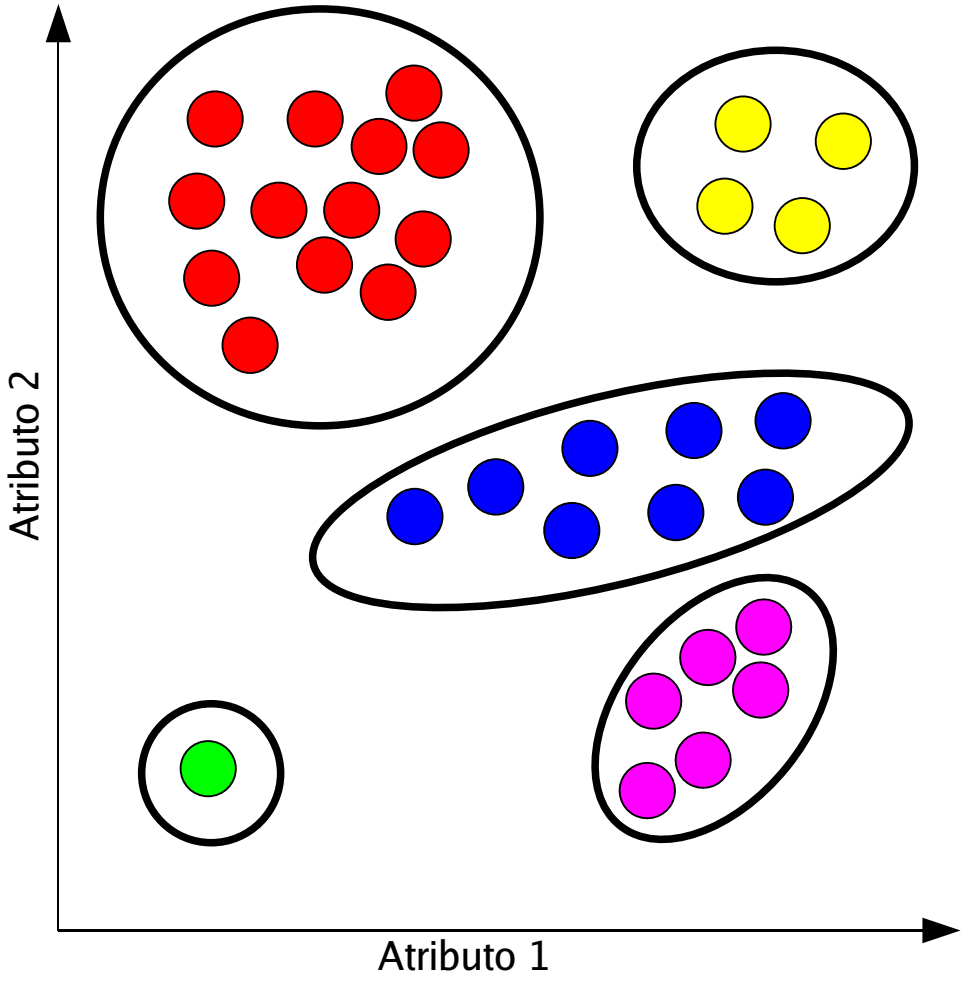
- Predição de uma categoria ou classe discreta.
  - Como entrada, temos muitos dados para os quais as classes são conhecidas.
  - Criamos um classificador ou modelo (fase de treinamento).
  - Como entrada em uma segunda fase, temos vários dados para os quais as classes não são conhecidas.
  - Usamos o classificador para indicar classes para estes dados. Assumimos que dados desconhecidos “próximos” de dados conhecidos terão a mesma classe dos dados conhecidos.

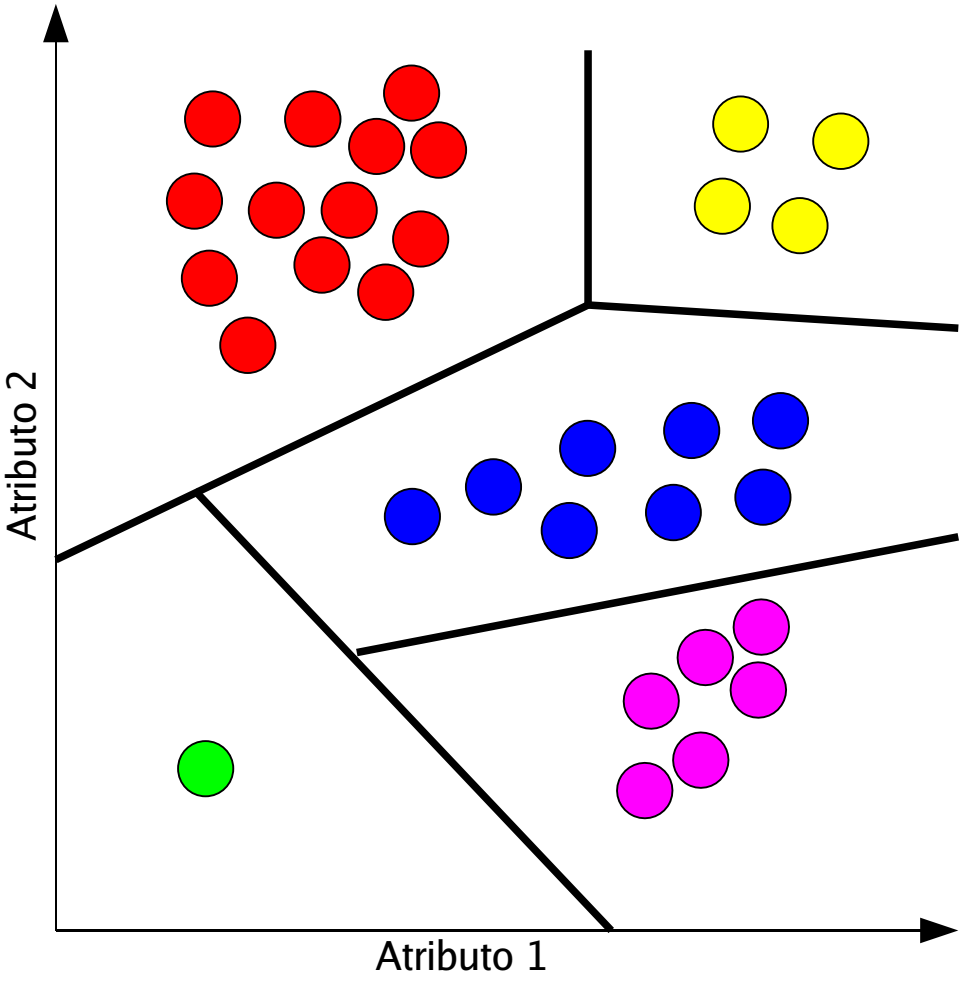
Técnicas: Classificação



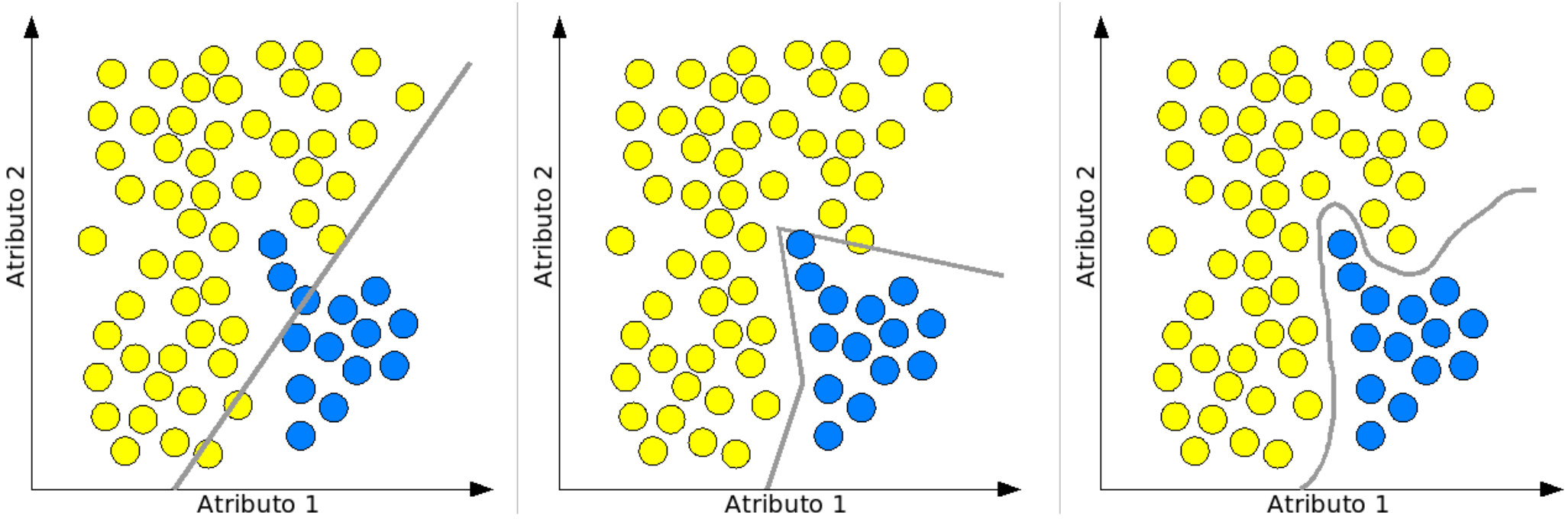


Técnicas: Classificação





Técnicas: Classificação



Técnicas: Associação

- **Descoberta de elementos que ocorrem (ou não!) em comum em coleções de dados.**
  - Dados de entrada: estruturas com associações (ex. lista de artigos comprados, pequenas séries temporais multivariadas, etc.).
  - Algoritmo identifica a existência de elementos em comum e suporte para esta existência.

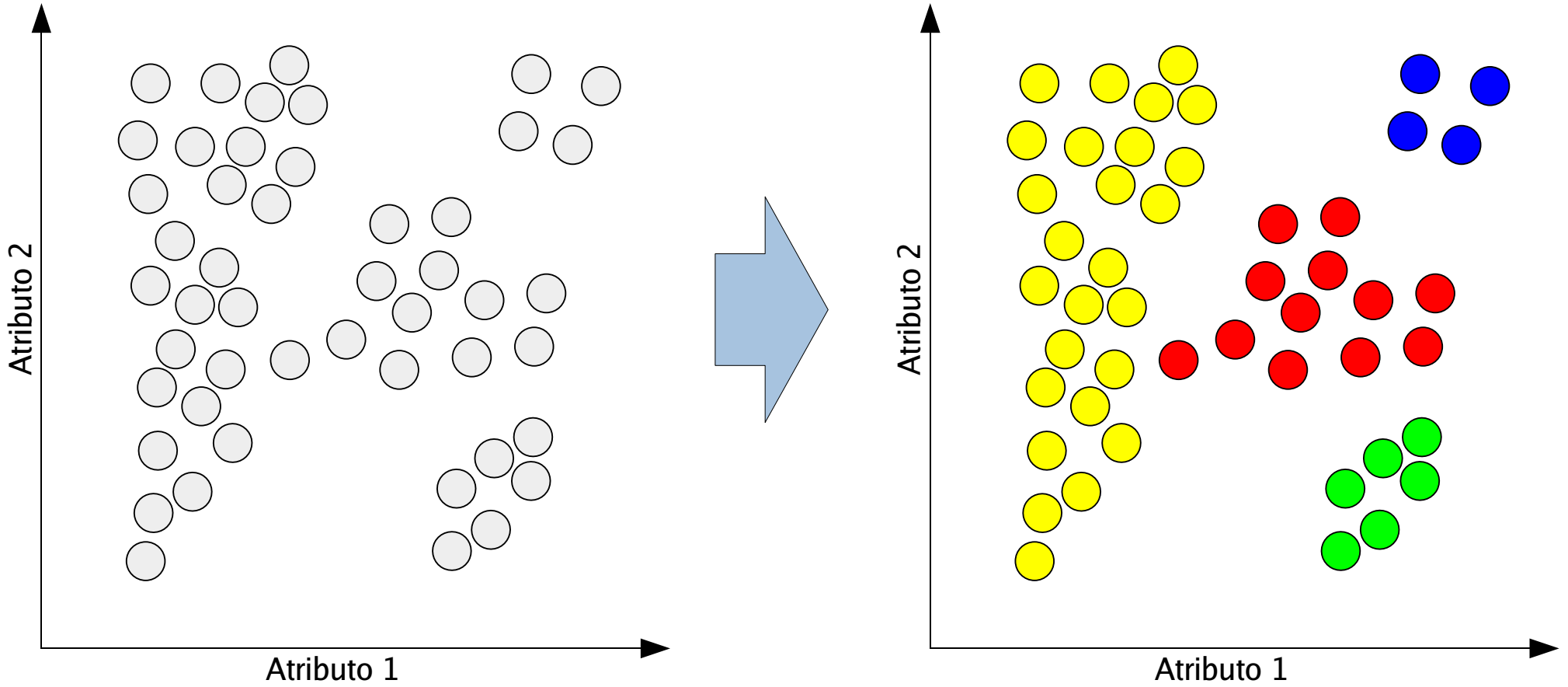
Técnicas: Associação

Loja	Caixa	Transação	Compras
03	05	011672	'PAO FRANCES'
03	05	011673	'PAO FRANCES'
03	06	010169	'PAO FRANCES'
03	05	011674	'PAO FRANCES', 'FLV PIMENTAO VERDE', 'LEITE PAST. SERRAMAR S', 'DANONE DANETTE CHOCOLA', 'ADOCANTE DOCE MENOR LI'
01	14	003752	'PAO FRANCES', 'LEITE PAST. SERRAMAR S'
01	14	003758	'BEB. REF.COCA COLA 1L', 'PAO FRANCES', 'LEITE PAST. PAULI TIPO'
01	13	003001	'LEITE PAST. PAULI TIPO', 'PAO FRANCES'
03	05	011685	'PAO FRANCES', 'PAO FRANCES'
01	14	003764	'ACUCAR REFINADO UNIAO', 'FEIJAO PRETO TARUMA 1K', 'PAO FRANCES'
03	05	011688	'PAO FRANCES'
01	14	003765	'BISC. TRIUNFO C.CRACKE', 'BISC. BAUD.GULOSOS 170', 'PAO FRANCES', 'ACUCAR REFINADO A. ALE', 'MORTADELA MARBA'
03	06	010188	'PAO FRANCES', 'ACUCAR REFINADO A. ALE'

Técnicas: Agrupamentos (*Clustering*)

- **Identificação de grupos semelhantes.**
  - Dados em um grupo devem apresentar alguma semelhança entre si.
  - Dados em um grupo devem ser diferentes de dados em outros grupos.
  - Dados de entrada: não precisam ter indicações de classe.
  - Algoritmo identifica determinado número de grupos de dados e calcula a associação dos dados de entrada aos grupos de saída.
  - Adicionalmente estatísticas e outras informações sobre os grupos podem ser criadas.

Técnicas: Agrupamentos (*Clustering*)



Técnicas: Predição ou Regressão

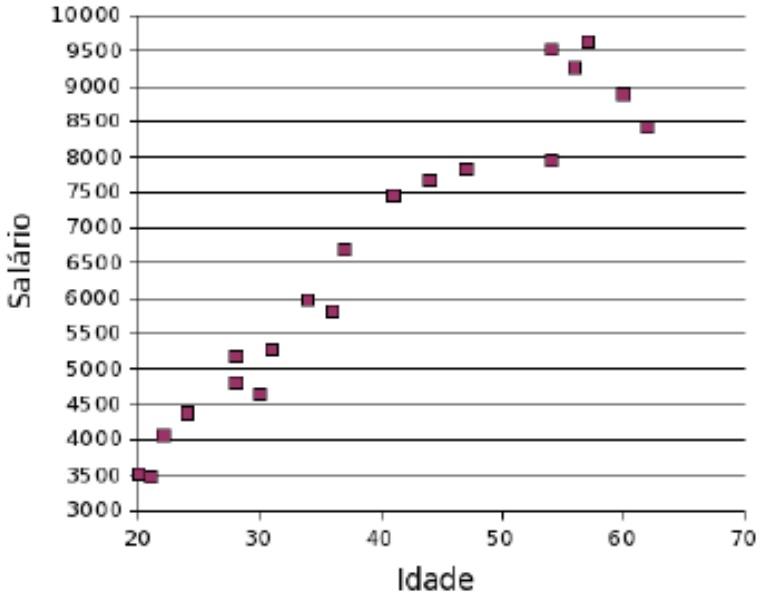
- **Predição de uma categoria ou classe com valores contínuos.**
  - Como entrada, temos muitos dados para os quais as classes (valores contínuos) são conhecidas. Com isso criamos um classificador ou função de regressão.
  - Como entrada em uma segunda fase, temos vários dados para os quais os valores das classes não são conhecidos. Usaremos a função de regressão para identificar a classe.
  - Alguns pontos em comum com classificação (treinamento, criação de modelo, uso, verificação e avaliação).

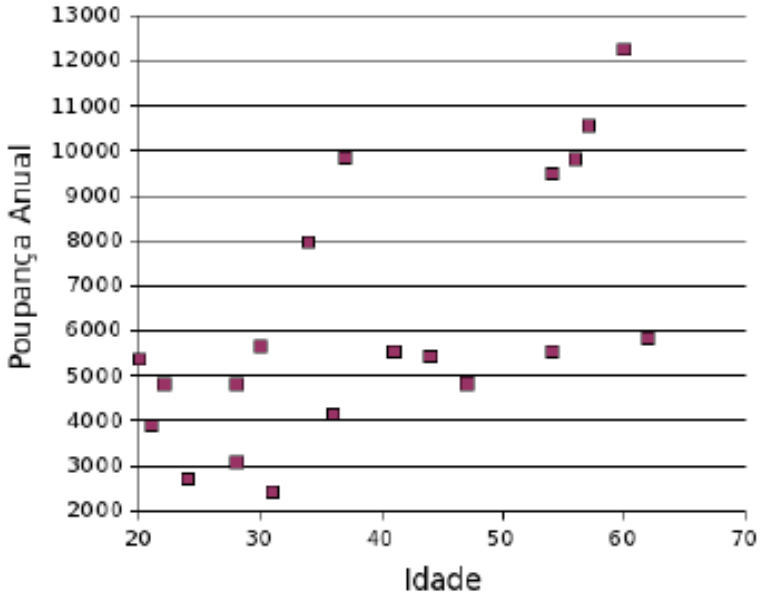
Técnicas: Predição ou Regressão

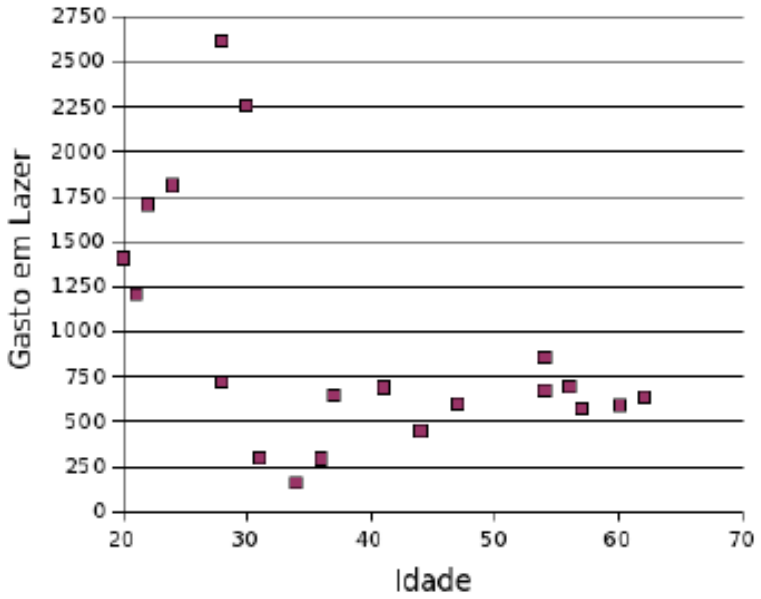
Idade	Salário	Poupança	Lazer
47	7837,41	4825,55	604,54
34	5982,92	7961,95	160,33
44	7659,56	5443,05	448,61
21	3483,75	3901,89	1205,73
37	6678,65	9851,69	650,58
54	7945,37	5549,21	675,82
28	5186,72	4799,23	2612,87
31	5270,25	2407,88	297,52
20	3514,11	5355,26	1406,72
41	7450,90	5533,70	689,04
28	4816,61	3074,03	723,88
22	4058,80	4798,99	1708,94
56	9270,28	9824,40	691,66
57	9636,85	10554,63	572,28
60	8877,44	12237,84	594,79
54	9521,86	9490,23	860,10
62	8418,57	5827,67	639,48
36	5818,98	4131,59	291,97
30	4652,78	5658,90	2256,84
24	4377,10	2718,81	1810,74

- Salário =
  141.65 * idade + 955.95
  (erro 29.3%)
- Poupança =
  130.8 * idade + 1056.85
  (erro 82.6%)
- Lazer =
  -24.53 * idade + 1909.07
  (erro 88.3%)

Técnicas: Predição ou Regressão



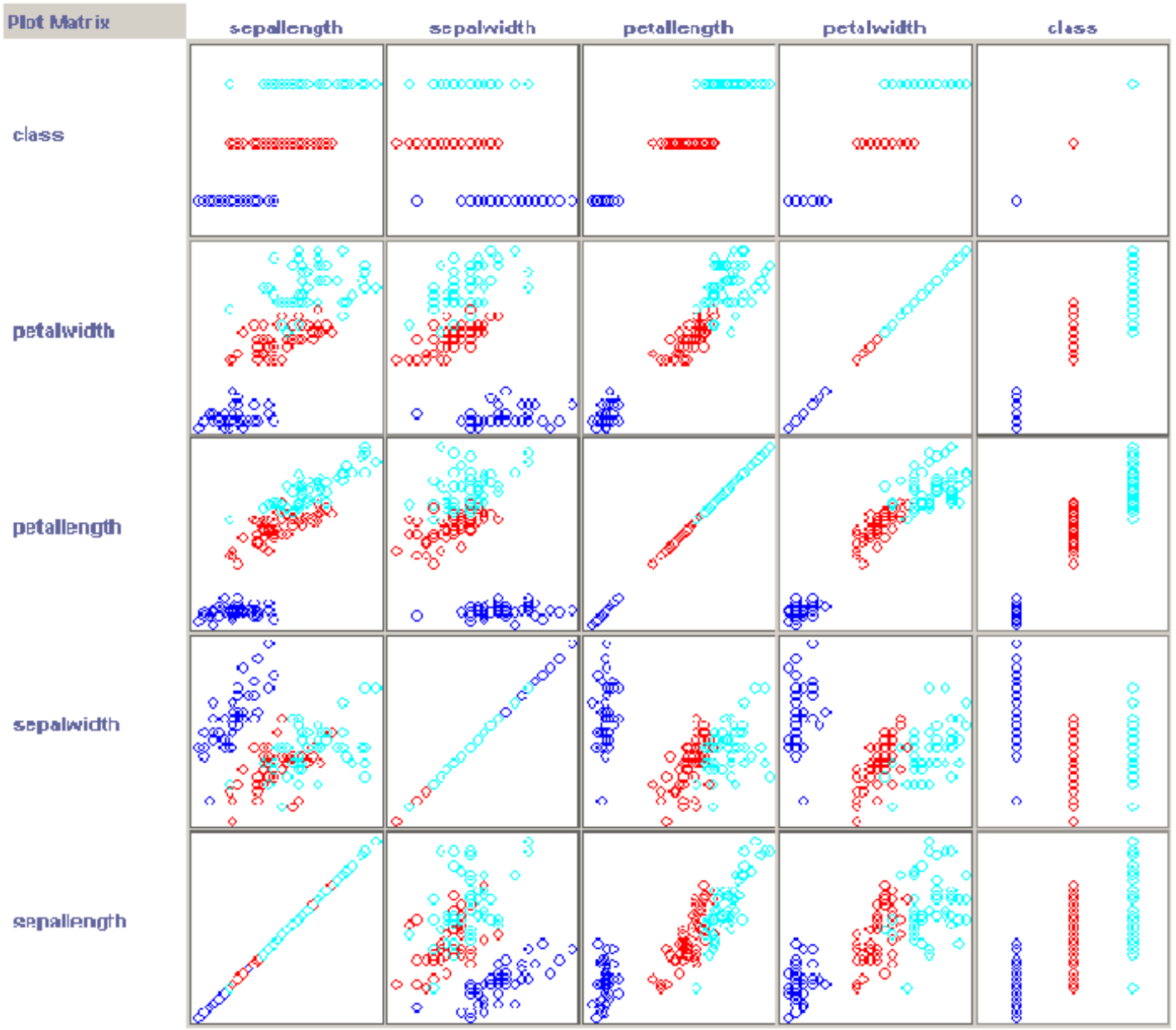




Visualização

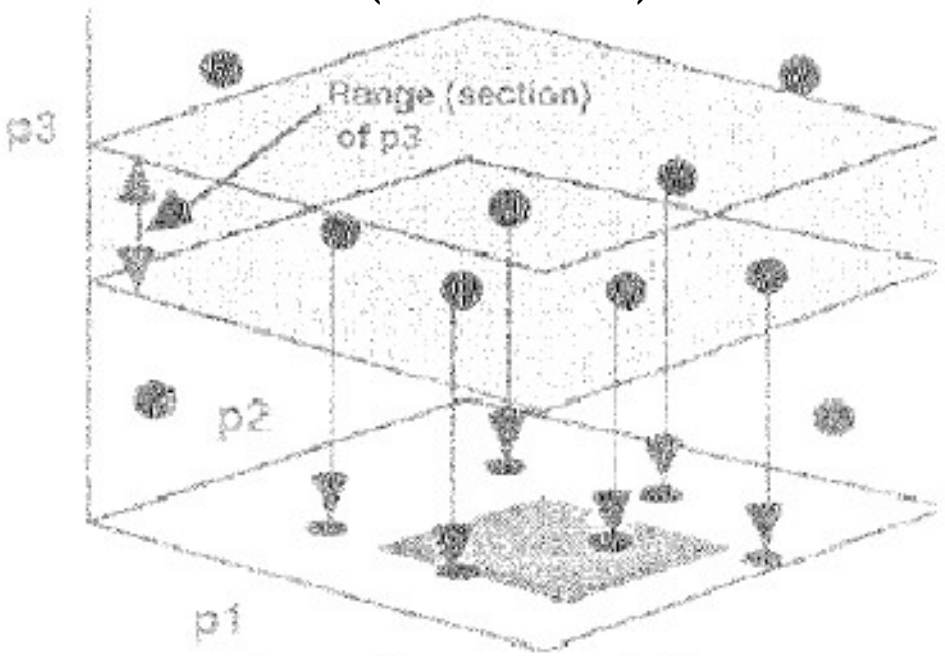
- Etapa importante: apresentação gráfica de resultados.
- Pode ser usada também como ferramenta no pré-processamento
- Permite visualização de estruturas dos dados (separabilidade, distribuições).
- Tópico interessante, porém complexo (dimensões, tipos de atributos, etc.)

Visualização

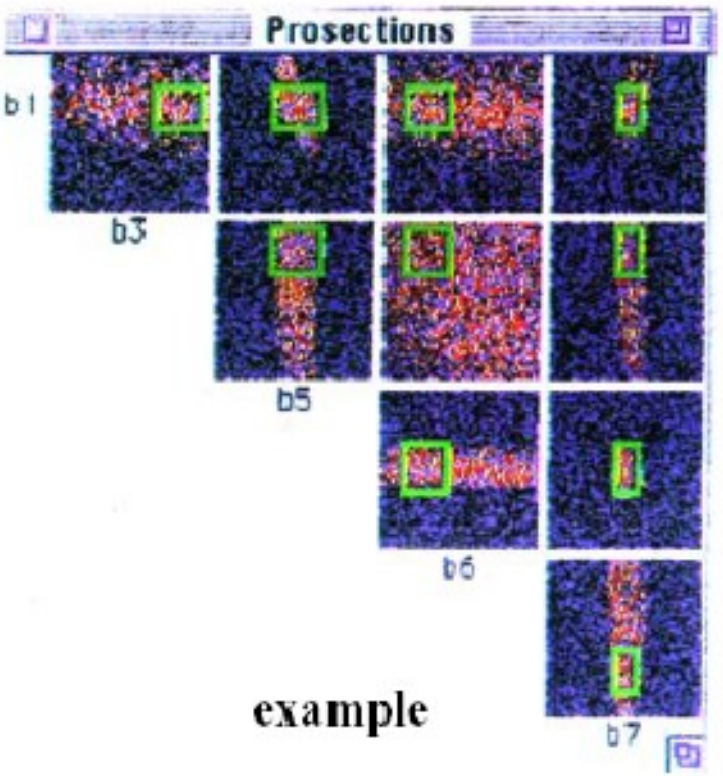


Visualização

Prosection Views (Daniel Keim)

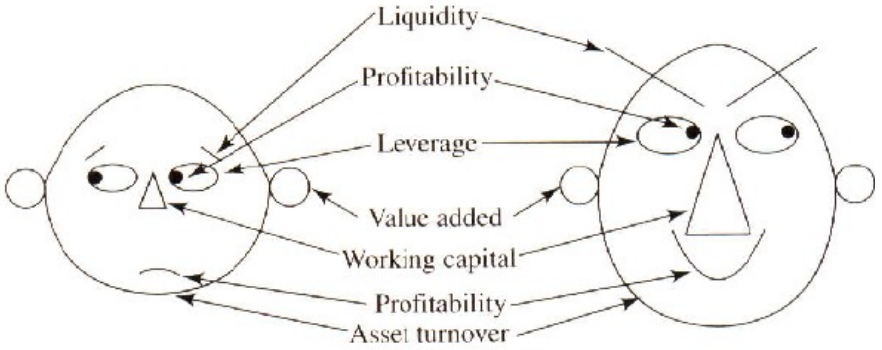


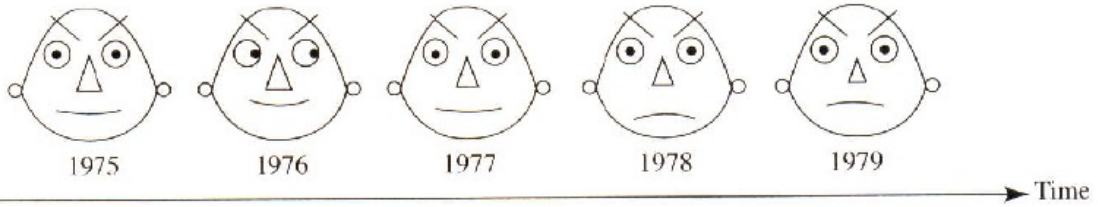
schematic representation



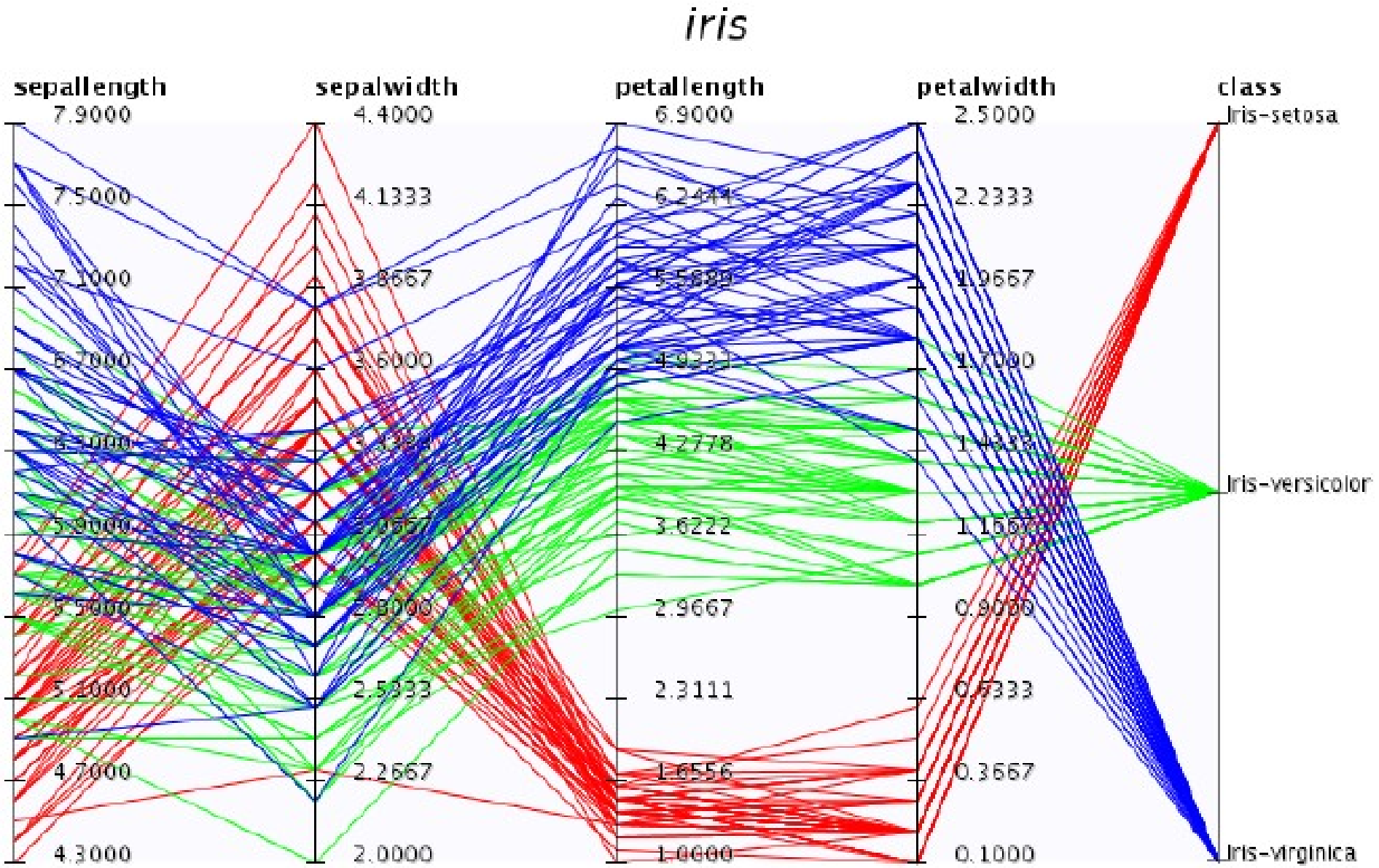
example

Chernoff Faces

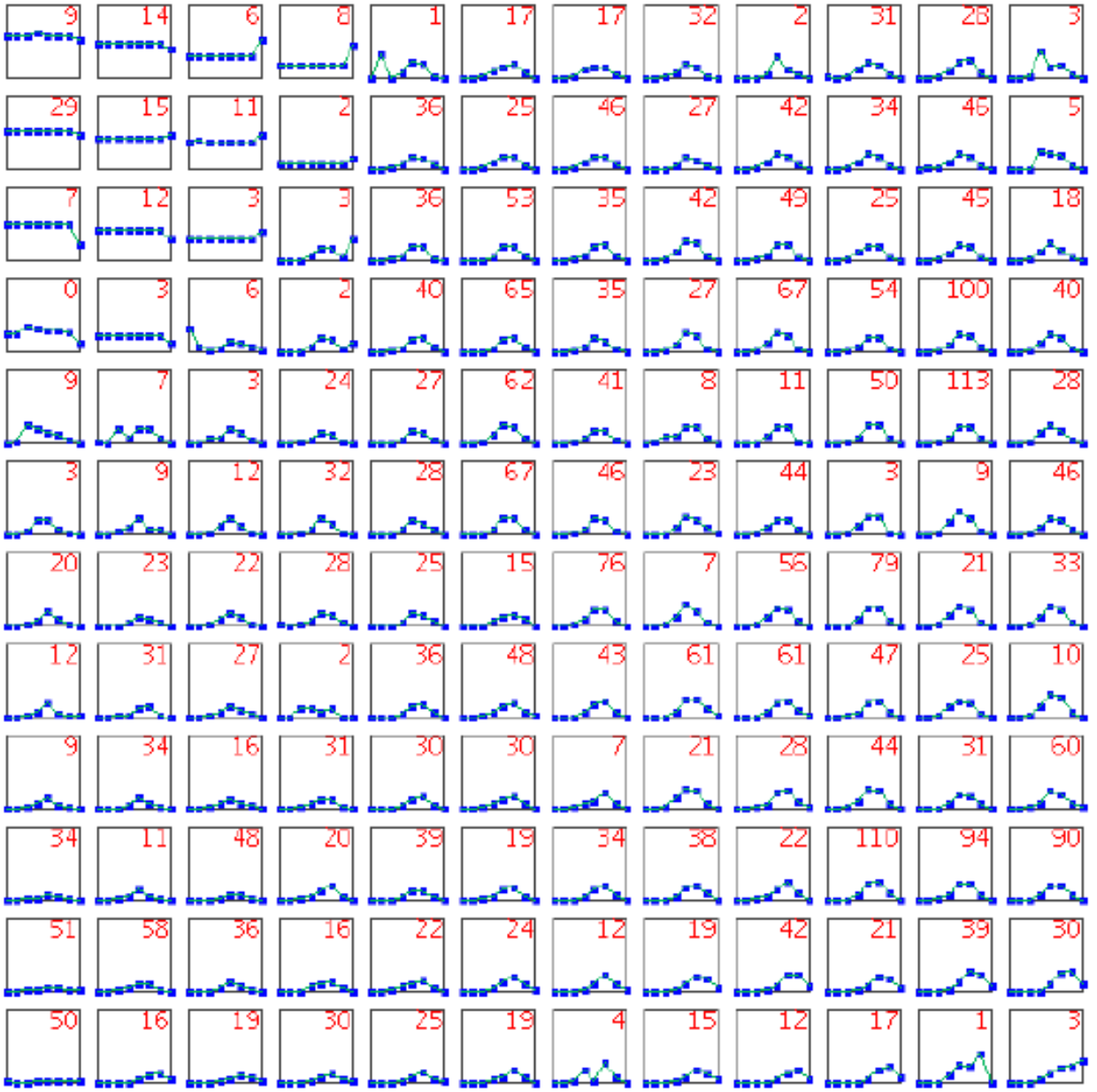




Visualização



Visualização

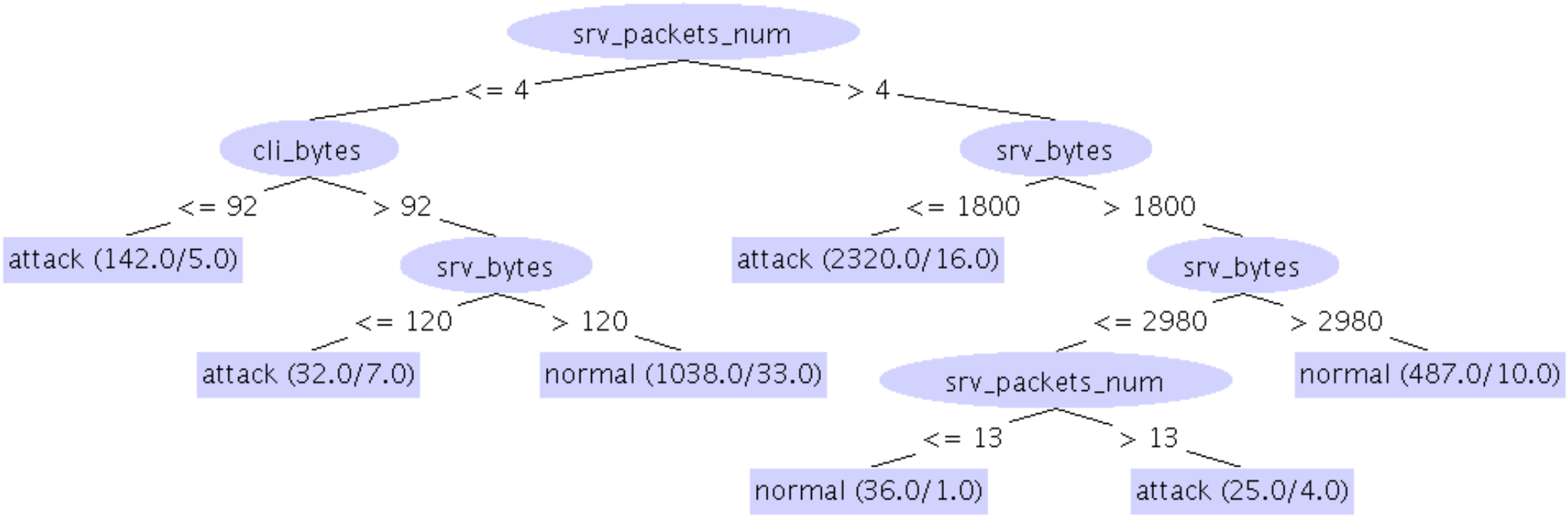


Data Mining: Alguns Algoritmos e Implementações

Árvores de Decisão

- Um dos tipos de algoritmo de classificação supervisionada mais popular.
  - A partir de dados rotulados de entrada, cria árvores que permitem tomar decisões usando atributos.
  - Divisões na árvore são criadas considerando a minimização da entropia dos dados em um galho.
  - Criação recursiva.

Árvores de Decisão



Árvores de Decisão

- Cnodições:
  - Para determinar que nós serão criados temos que ter instâncias com classes definidas.
  - Devemos saber também qual é o atributo a ser usado como classe.
- Classificação: testes sobre uma base de dados que indica a classe a partir dos valores dos atributos de entrada.
  - Nós em uma árvore de decisão: testes sobre os atributos.
  - Folhas: determinação das classes.
- Semelhança com sistemas especialistas.

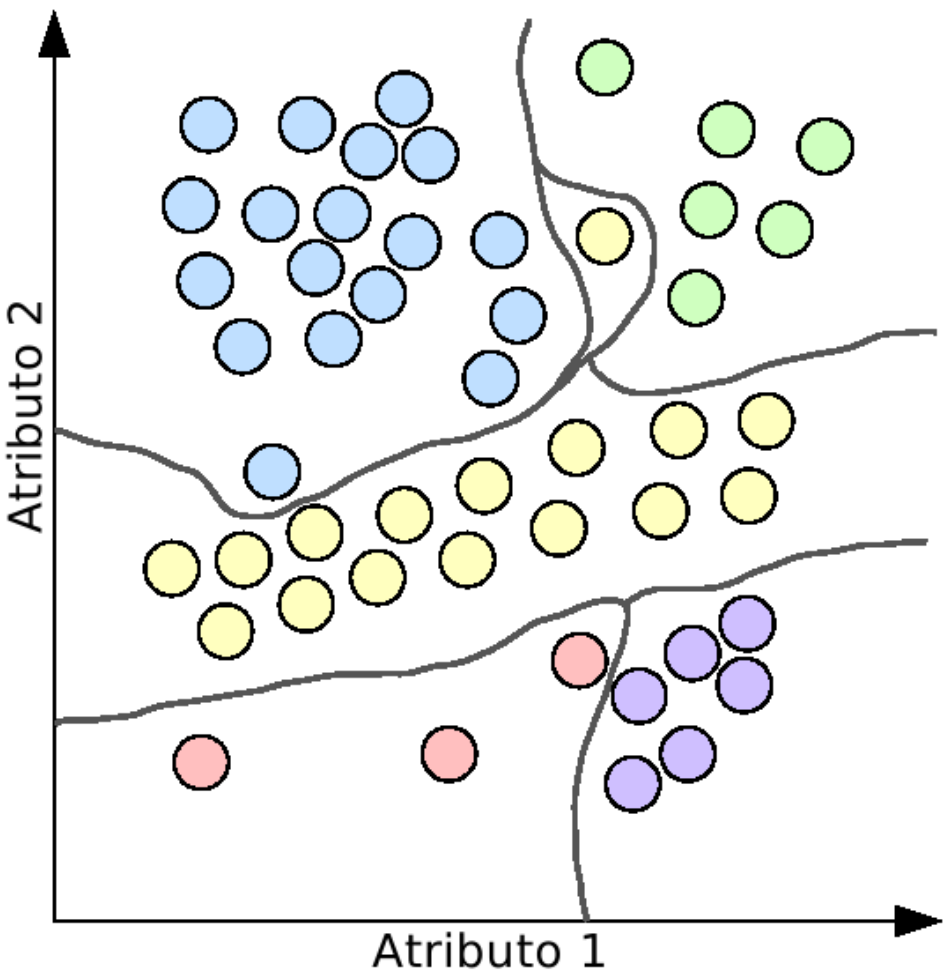
Vizinhos Mais Próximos

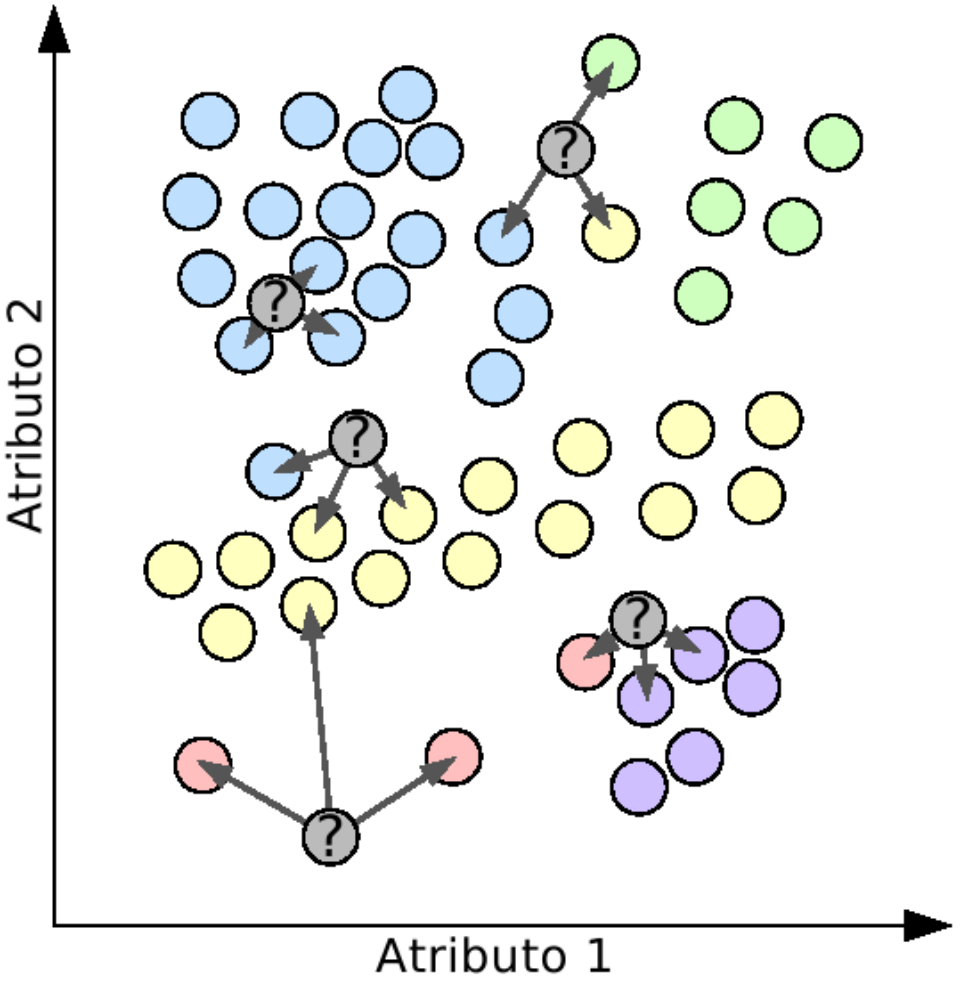
- Método simples de classificação supervisionada.
  - Bastante intuitivo: se uma amostra de classe desconhecida estiver bem próxima de uma de classe conhecida, as classes devem ser as mesmas.
  - Não criamos protótipos ou assinaturas para as classes conhecidas: usamos as amostras com classes conhecidas como protótipos.
- Algoritmo básico: para cada amostra com classe desconhecida, comparamos a distância dela para cada amostra com classe conhecida. Usamos a classe da amostra mais próxima.
- Cria *hipersuperfícies* de separação.

Vizinhos Mais Próximos

- Extensão mais aplicável: **K-Vizinhos mais próximos**.
- Vantagens:
  - Permite boa classificação independentemente do modelo de distribuição dos dados.
  - Não sofre de problemas de criação de assinaturas causados por exemplo, por número pequeno de amostras.
- Desvantagens:
  - Resultado difícil de sumarizar (ex. comparado com regras ou árvores de decisão).
  - Sujeito à influência de *outliers*.
  - Problemas para comparação de valores não-numéricos.

Vizinhos Mais Próximos

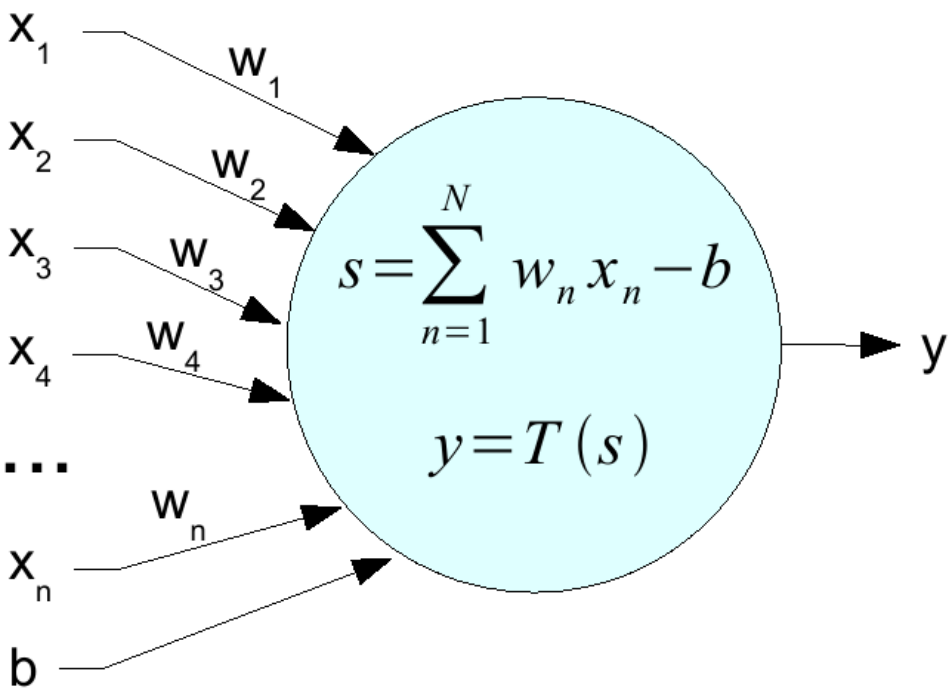




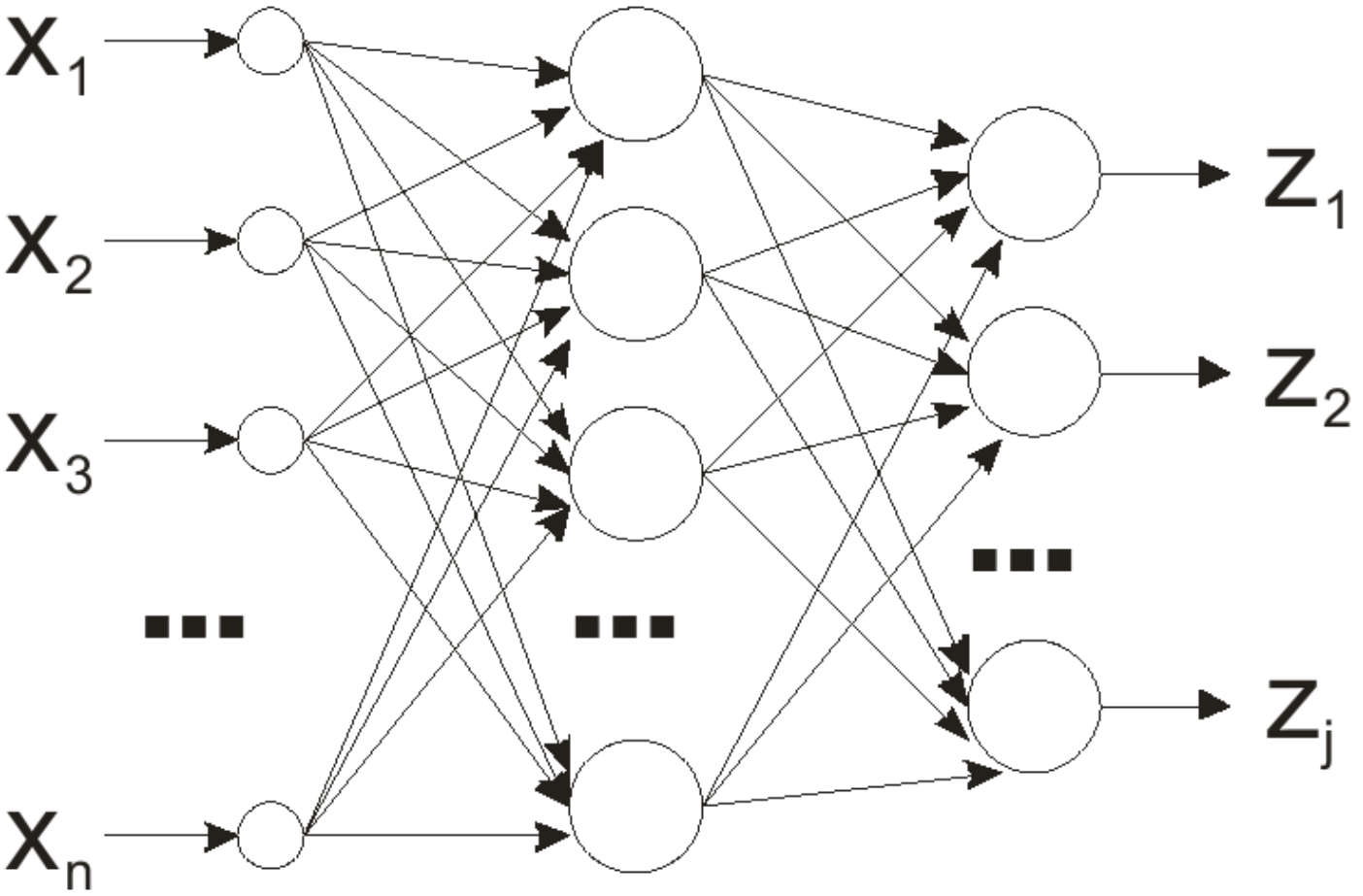
Redes Neurais Artificiais - *MLP*

- Inspirada no funcionamento (simplificado) de neurônios naturais.
- *Multi-Layer Perceptrons* (MLPs):
  - Neurônios processam valores de entrada e apresentam um de saída (classe).
  - Vários neurônios artificiais conectados com várias camadasscompõem uma *rede neural*.

- Esquema básico de um neurônio artificial (*perceptron*):
  - Cada entrada $\mathbf{x}_n$ tem um peso multiplicativo associado $\mathbf{w}_n$. Podemos ter uma entrada adicional (**b** ou bias)
  - O neurônio calcula combinação linear das entradas e pesos e aplica um limiar (função **T**) que determina a saída do neurônio.
  - O treinamento é feito através da apresentação de entradas e resultados conhecidos (classificação supervisionada) e ajuste dos pesos com algoritmos específicos.



Redes Neurais Artificiais - *MLP*



Redes Neurais Artificiais - *MLP*

- Vantagens:
  - Capacidade de separar bem classes não linearmente separáveis (com múltiplas camadas).
  - Existem várias implementações, muitos parâmetros para ajuste fino.
- Desvantagens:
  - Complexidade do mecanismo de classificação (“caixa preta”) e interpretação dos resultados.
  - Treinamento pode ser complexo (computacionalmente caro), definição da arquitetura também.
  - Existem várias implementações, muitos parâmetros para ajuste fino.

Clustering – K-Médias

- Algoritmo simples, iterativo, para agrupamento (não-supervisionado).
  - Entrada: instâncias, medida de distância, número de grupos (K).
  - Saída: centróides dos grupos, pertinência das instâncias aos grupos, métricas.
  - O algoritmo tenta minimizar o erro quadrático calculado entre as instâncias e os centróides dos grupos.
-

- Algoritmo:
  1. Inicializamos os centróides dos *K* grupos.
  2. Marcamos cada instância como pertencente ao grupo (centróide) mais próximo.
  3. Recalculamos os centróides dos grupos considerando as pertinências.

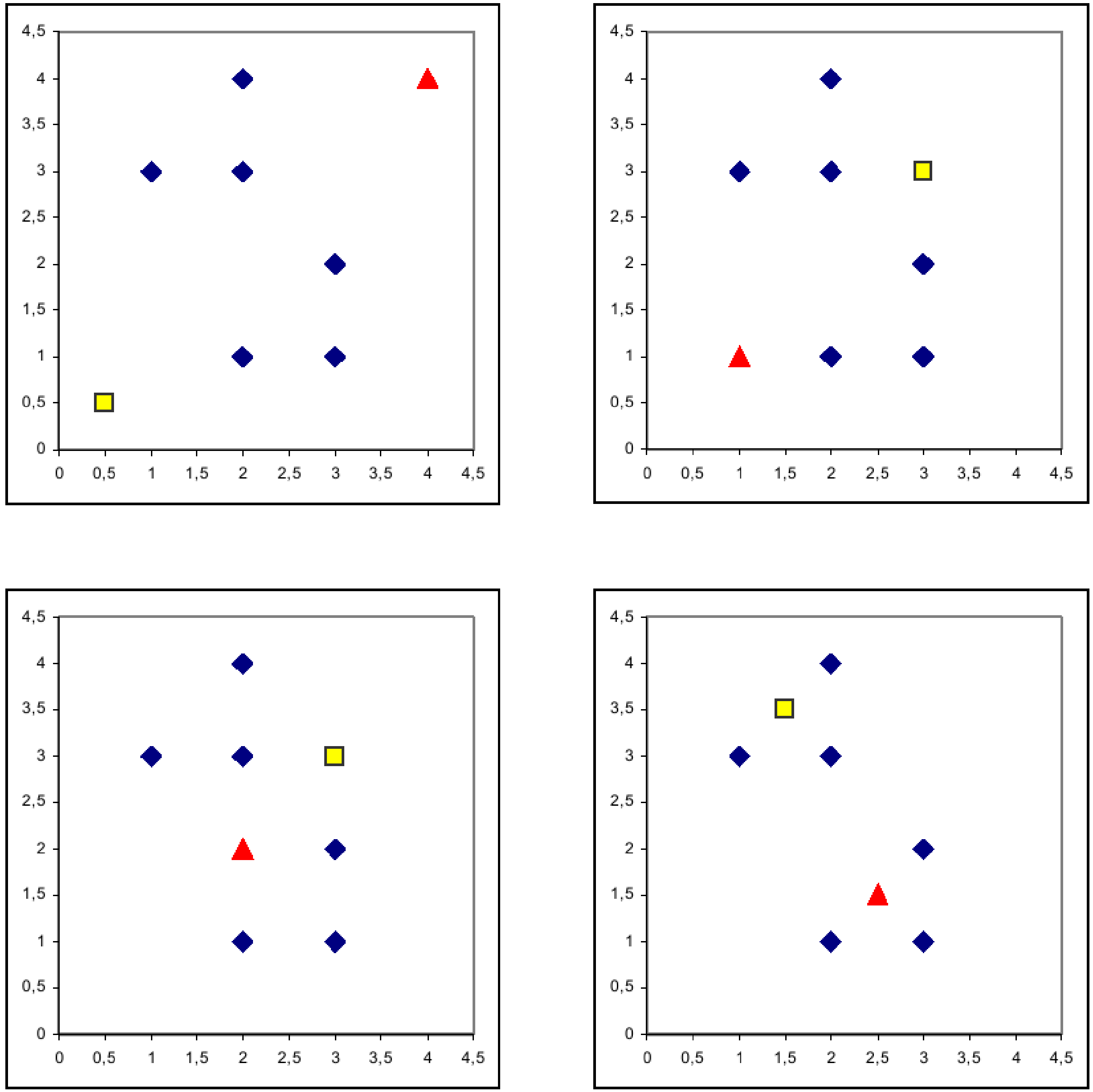
$$v_i = \frac{1}{n_i} \sum_{x_k \in C_i} x_k$$

  4. Recalculamos o erro quadrático total.

$$J = \sum_{k=1}^{n} \sum_{x_k \in C_i} |x_k - v_i|^2$$

  5. Verificamos as condições de parada e repetimos a partir do passo 2.

Clustering – K-Médias



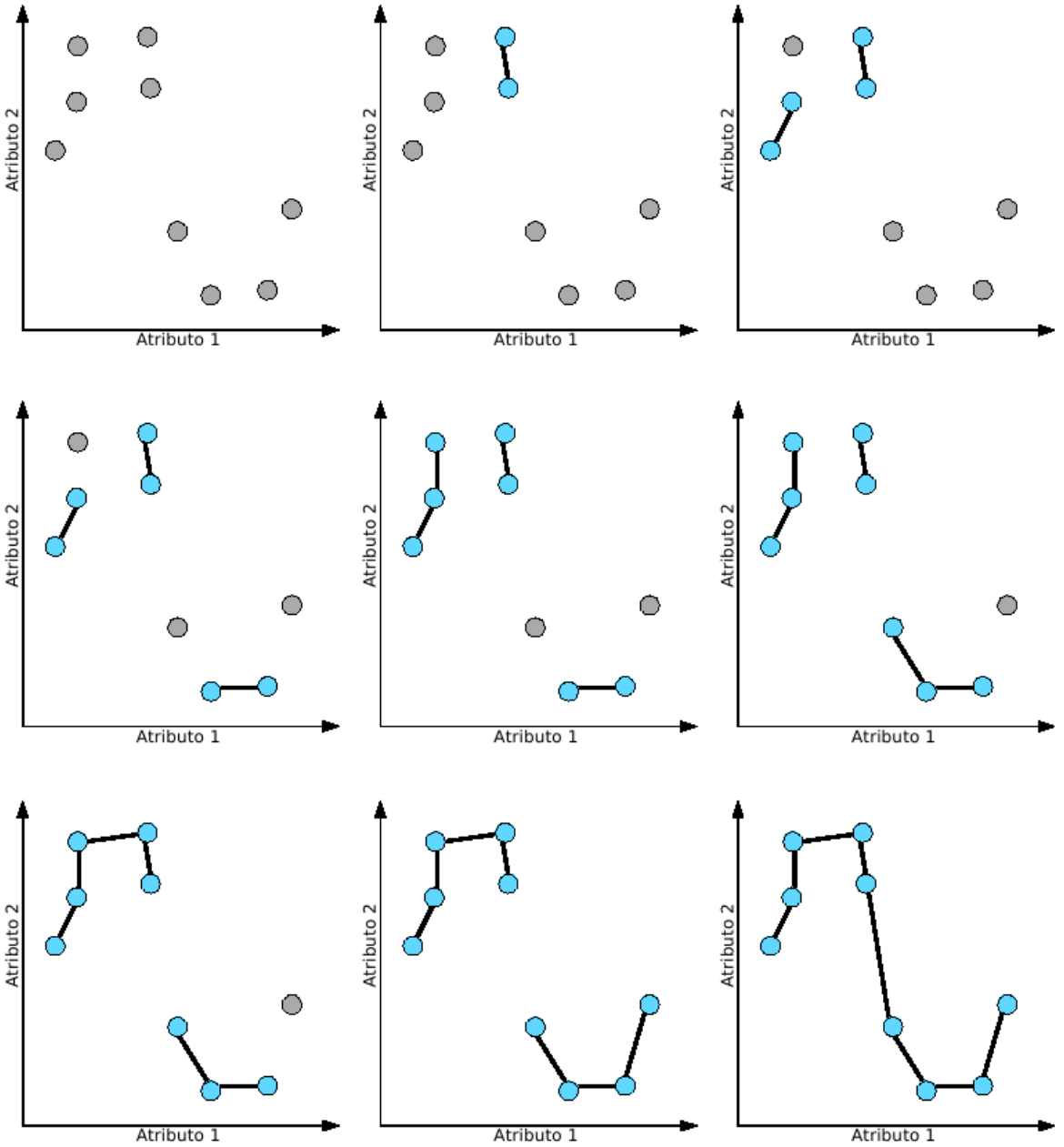
Clustering – K-Médias

- Problemas:
  - Somente dados numéricos!
  - Múltiplas iterações com todos os dados: problemas de performance.
  - Inicialização: como escolher centróides iniciais (impacto na convergência).
  - Converge para um mínimo local: pode ser bom o suficiente.
  - Singularidades: grupos sem instâncias relacionadas.
    - Não podemos calcular seus centróides.
    - Podemos resolver com a eliminação de grupos vazios (complexo, caro?)
  - Apresentação dos dados: ordem de apresentação de instâncias pode alterar resultados.
    - Escolha de K?
    - Existe um K' melhor do que o K?

Clustering Hierárquico

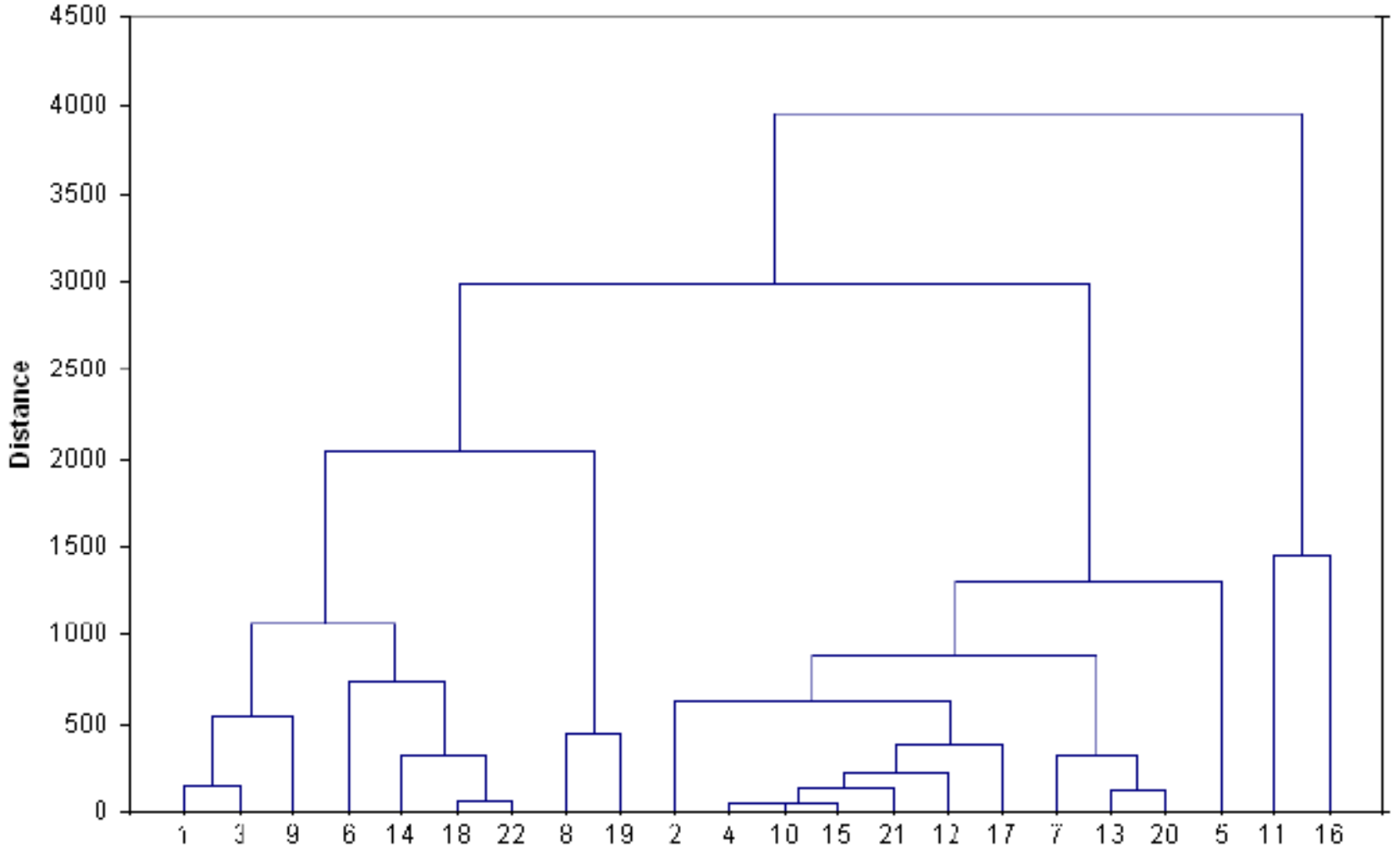
- Algoritmos hierárquicos de agrupamento formam uma série de partições dos dados onde cada partição contém um número menor de grupos do que a partição anterior.
- Passos do algoritmo:
  1. Considere todas as instâncias como grupos. Os centros destes grupos são os valores da própria instância.
  2. Crie uma matriz de distâncias que indique a distância de cada grupo a cada outro grupo.
  3. Localize, nesta matriz, os dois grupos com menor distância entre eles, e efetue a união destes grupos.
  4. Se ainda houver dois ou mais grupos, volte ao passo 2.

Clustering Hierárquico



Clustering Hierárquico

- Resultado: *dendograma*.



Regras de Associação

- Regras sobre relações e co-ocorrências em bases de dados:
  - Se *X* ocorre na base de dados, então *Y* também ocorre (com alguma relação a *X*).
  - Co-ocorrência: se *X*, *Y* e *Z* ocorrem na base de dados então *A* também ocorre (com alguma relação à *X*, *Y* e *Z*).
    - *X*, *Y* e *Z* são os antecedentes da associação; *A* é o conseqüente.
- Muito usado para verificar associações em tabelas de transações (“carrinhos de compra”)

Regras de Associação

- Regras devem ter métricas que indiquem usabilidade:
- Significância em uma associação: ela pode existir mas ser muito rara em uma base de dados.
  - Suporte $X \rightarrow Y$: número de casos que contém X e Y dividido pelo número total de registros.
- Confiança em uma associação: o antecedente pode ocorrer várias vezes na base de dados mas nem sempre com o mesmo conseqüente associado.
  - Confiança $X \rightarrow Y$: número de registros que contém X e Y dividido pelo número de registros que contém X.

Regras de Associação

- Algoritmo *APriori*
- Um dos mais conhecidos, base de muitos outros.
- Definições necessárias:
  - *K-itemsets* são conjuntos com *K* itens que podem aparecer na base de dados.
  - Suporte mínimo é o valor mínimo do suporte para que um *K-itemset* seja considerado.
  - Confiança mínima é um limite para filtragem das associações descobertas pelo algoritmo.

Regras de Associação

- Passos (simplificados) do algoritmo *Apriori*
  1. Dados de entrada: coleção de dados associados, suporte mínimo, confiança mínima.
  2. Considerar *K = 1* para criação de *K-itemsets.*
  3. Criar uma tabela de *K-itemsets* com suporte acima do suporte mínimo.
  4. Criar com os *itemsets* filtrados um conjunto de candidatos a *(K+1) itemsets*.
  5. Usar propriedades do *Apriori* para eliminar itemsets infreqüentes.
  6. Repetir desde o passo 3 até que o conjunto gerado seja vazio.
  7. Listar regras de associação (com permutações) e aplicar limite de confiança.

Regras de Associação

- Vantagens:
  - Um dos poucos algoritmos de processamento simbólico, aplicável para detecção de co-ocorrências.
  - Algoritmo simples, várias implementações.
- Desvantagens:
  - Tempo de processamento e uso intensivo de memória.
  - Difícil decidir valores dos parâmetros.
  - Não pode usar diretamente valores numéricos.
  - Freqüentemente requer análise das regras encontradas!

Estudo de caso: Classificação de Sessões em Servidores Web

Estudo de Caso: Classificação de Sessões

- Logs em servidores HTTP foram coletados em um servidor normal e em um servidor em um Honeypot.
  - O objetivo da análise é verificar se é possível construir um mecanismo que separe o mais corretamente possível sessões correspondentes a um acesso normal de sessões correspondentes a ataques ou tentativas de invasão.
  - Pré-processamento já foi feito (sessões já foram reconstituídas).
  - Usaremos técnicas de classificação supervisionada.

Estudo de Caso: Classificação de Sessões

- Atributos usados:
  - **session_time**: Duração da sessão em segundos;
  - **cli_packets**: Número de pacotes enviados pelo cliente;
  - **cli_bytes**: Quantidade de bytes enviados pelo cliente;
  - **srv_packets**: Número de pacotes enviados pelo servidor;
  - **srv_bytes**: Quantidade de bytes enviados pelo servidor;
  - **class**: Se a sessão foi pré-classificada como suspeita ou não. A classe pode ter dois valores: *normal* ou *attack*.
- Temos 1549 dados com classe *normal* e 2531 com classe *attack*.

Estudo de Caso: Classificação de Sessões

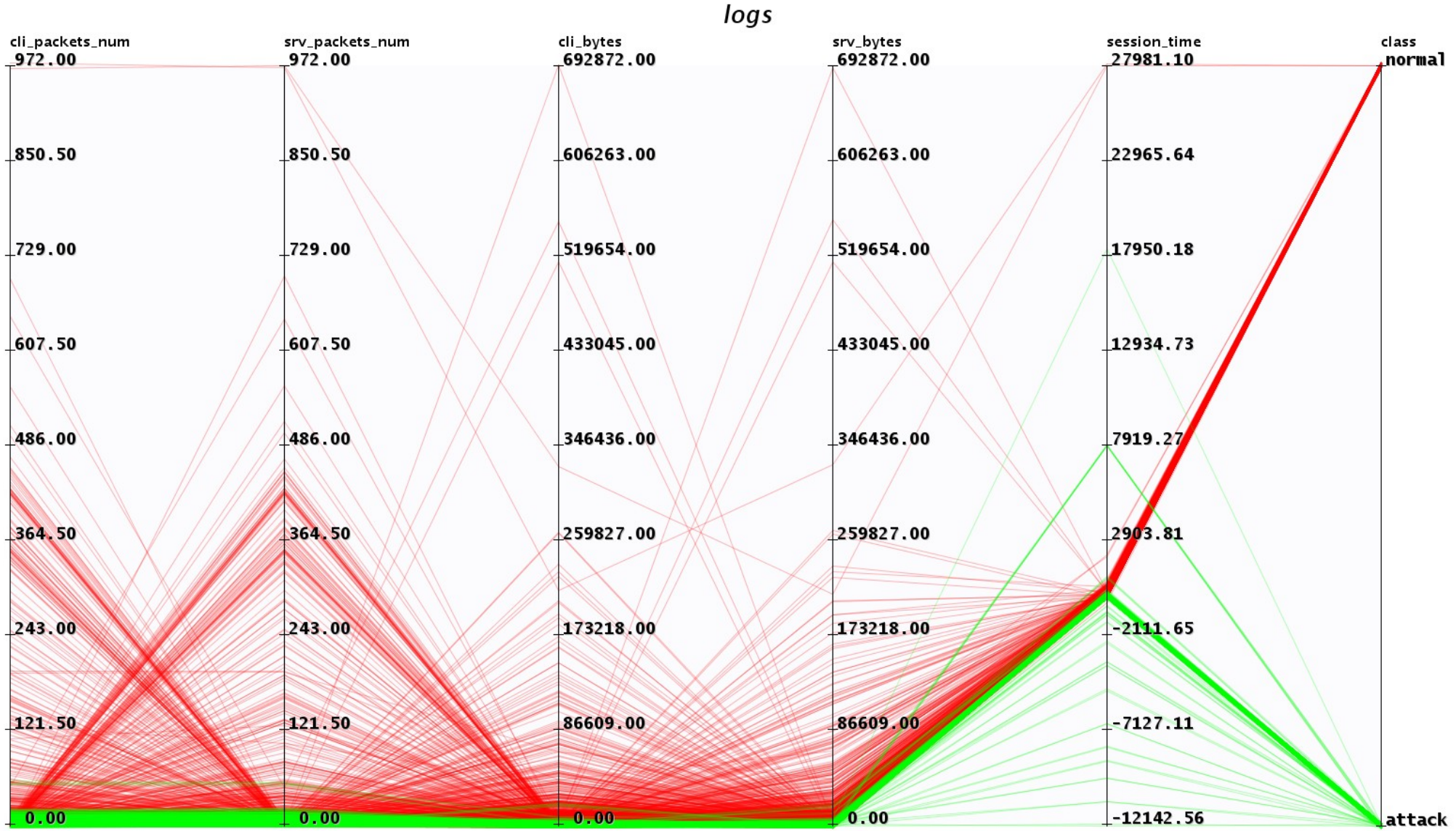
- Arquivo .arff (formato do software Weka)

```
@relation logs
@attribute cli_packets_num numeric
@attribute srv_packets_num numeric
@attribute cli_bytes numeric
@attribute srv_bytes numeric
@attribute session_time numeric
@attribute class {normal,attack}

@data
44,2,63432,591,460.92804,normal
2,2,2626,599,13.671535,normal
5,0,5162,0,0.514798,normal
35,40,5861,11882,89.268859,normal
108,128,22749,38478,894.124939,normal
.....
5,5,744,418,6.139626,attack
5,5,742,418,1.935146,attack
5,5,738,418,2.671999,attack
5,5,756,418,2.565365,attack
6,6,783,457,5.65425,attack
```

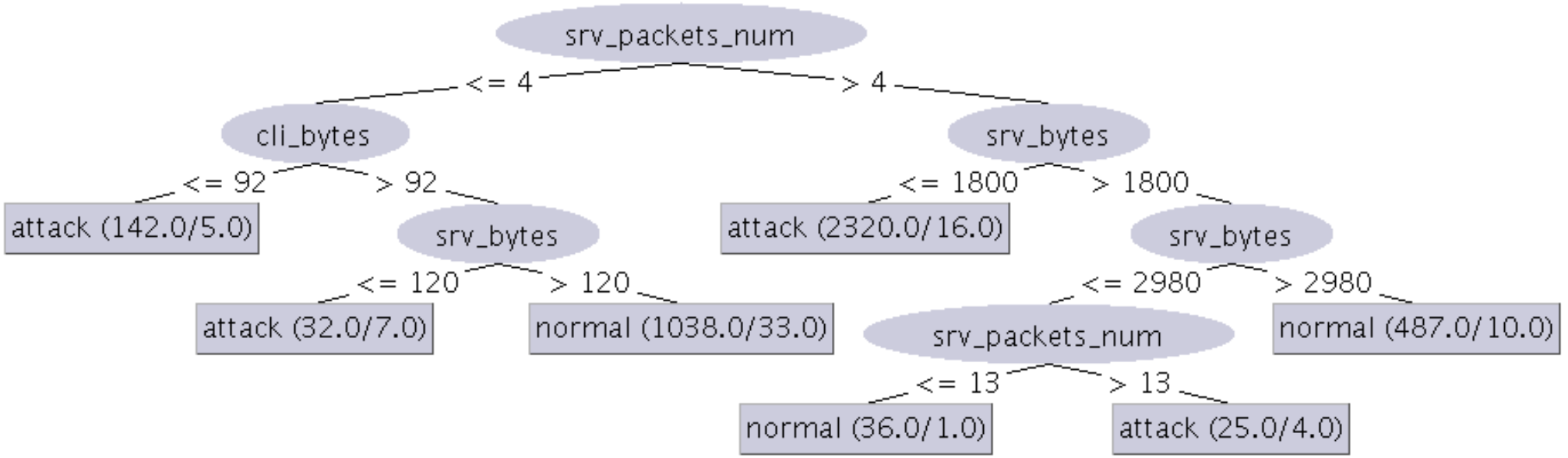
Estudo de Caso: Classificação de Sessões

- Como é o comportamento destes dados?



Estudo de Caso: Classificação de Sessões

- Classificação com uma árvore de decisão (bastante simplificada)




```
Correctly Classified Instances        3999               98.0147 %
Incorrectly Classified Instances        81                1.9853 %

    a       b   <-- classified as
 1513      36 |     a = normal
   45    2486 |     b = attack
```

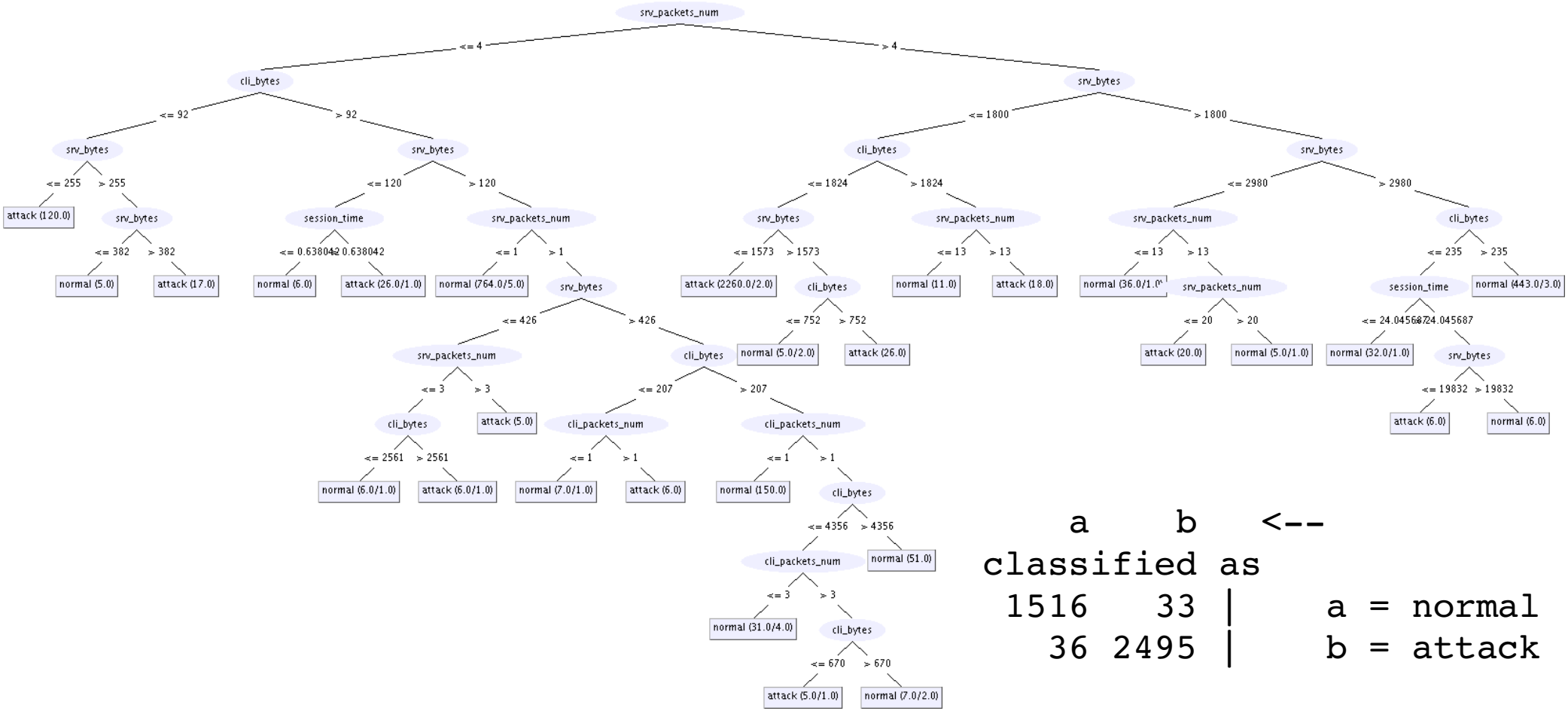
Estudo de Caso: Classificação de Sessões

- Classificação com uma árvore de decisão (bastante simplificada)

```
srv_packets_num <= 4
|   cli_bytes <= 92: attack (142.0/5.0)
|   cli_bytes > 92
|   |   srv_bytes <= 120: attack (32.0/7.0)
|   |   srv_bytes > 120: normal (1038.0/33.0)
srv_packets_num > 4
|   srv_bytes <= 1800: attack (2320.0/16.0)
|   srv_bytes > 1800
|   |   srv_bytes <= 2980
|   |   |   srv_packets_num <= 13: normal (36.0/1.0)
|   |   |   srv_packets_num > 13: attack (25.0/4.0)
|   |   srv_bytes > 2980: normal (487.0/10.0)
```

Estudo de Caso: Classificação de Sessões

- Classificação com uma árvore de decisão (mais detalhada)

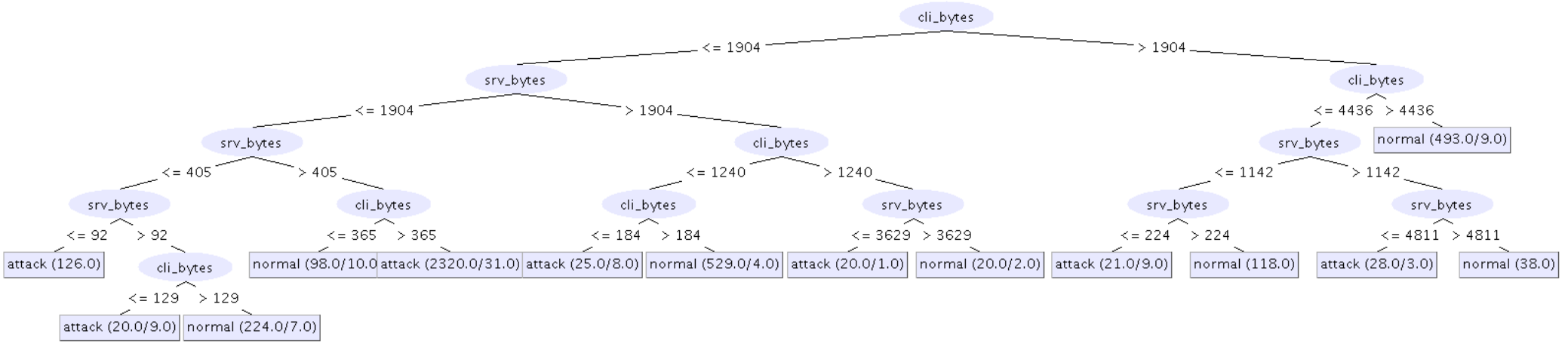



```
    a    b   <--
classified as
 1516   33 |    a = normal
   36 2495 |    b = attack
```

```
Correctly Classified Instances        4011       98.3088 %
Incorrectly Classified Instances        69        1.6912 %
```

Estudo de Caso: Classificação de Sessões

- Exploração: quais atributos podem ser desconsiderados?
  - Usamos somente *cli_bytes*, *srv_bytes* e *class*.




```
Correctly Classified Instances        3954               96.9118 %
Incorrectly Classified Instances       126                3.0882 %

                 a    b   <-- classified as
              1482   67 |    a = normal
                59 2472 |    b = attack
```

Estudo de Caso: Classificação de Sessões

- Classificação com uma rede neural MLP (arquitetura 5x**5**x2)

```
Correctly Classified Instances        3844               94.2157 %
Incorrectly Classified Instances       236                5.7843 %
    a    b   <-- classified as                          1 minuto
 1452   97 |    a = normal
  139 2392 |    b = attack
```

- Classificação com uma rede neural MLP (arquitetura 5x**15**x2)

```
Correctly Classified Instances        3880               95.098  %
Incorrectly Classified Instances       200                4.902  %
    a    b   <-- classified as                          2 minutos
 1504   45 |    a = normal
  155 2376 |    b = attack
```

- Classificação com uma rede neural MLP (arquitetura 5x**50**x2)

```
Correctly Classified Instances        3876               95      %
Incorrectly Classified Instances       204                5      %
    a    b   <-- classified as                          6 minutos
 1502   47 |    a = normal
  157 2374 |    b = attack
```

Estudo de Caso: Classificação de Sessões

- Classificação com K-vizinhos mais próximos (K=1)

```
Correctly Classified Instances        4026               98.6765 %
Incorrectly Classified Instances        54                1.3235 %
    a    b   <-- classified as
 1521   28 |    a = normal
   26 2505 |    b = attack
```

- Classificação com K-vizinhos mais próximos (K=5)

```
Correctly Classified Instances        3992               97.8431 %
Incorrectly Classified Instances        88                2.1569 %
    a    b   <-- classified as
 1508   41 |    a = normal
   47 2484 |    b = attack
```

- Classificação com K-vizinhos mais próximos (K=25)

```
Correctly Classified Instances        3978               97.5    %
Incorrectly Classified Instances       102                2.5    %
    a    b   <-- classified as
 1515   34 |    a = normal
   68 2463 |    b = attack
```

Estudo de caso: Co-ocorrência de termos de busca

Estudo de caso: Co-ocorrência de termos

- **America On-Line**: provedor de acesso.
- Tem aplicação com interface própria e mecanismo de busca integrado.
- Funcionários da empresa fizeram uma pesquisa:
  - Termos usados por mais de 650.000 usuários foram coletados entre março e maio de 2006, totalizando quase onze milhões de entradas.
  - Os usuários estavam anonimizados, isto é, somente números aparentemente seqüenciais mas fora de um padrão foram usados para identificar quem fez as buscas.
  - Por um descuido, a base de dados (um arquivo de texto de aproximadamente 2.5 gigabytes, descompactado) acabou sendo disponibilizada temporariamente na Internet

Estudo de caso: Co-ocorrência de termos

- Apesar da anonimização dos usuários, padrões de comportamento interessantes (e até mesmo bizarros!) puderam ser observados.
- É possível observar termos co-ocorrentes freqüentes nesta base de dados?

Estudo de caso: Co-ocorrência de termos

- Pré-processamento:
  - Anonimização total: nem mesmo o número de usuário será usado.
  - Todas as buscas de cada usuário foram consideradas. Procuramos a existência de ao menos uma das palavras-chave nas buscas efetuadas.
  - Palavras-chave: *car*, *family*, *travel*, *loan*, *school*, *debt*, *mortgage*, *college*, *savings*, *marriage*, *wedding*, *children*, *work*, *money*, *dog*, *party*, *beer*, *pizza*, *wine* e *liquor*.
  - Somente 65.000 usuários foram considerados.

Estudo de caso: Co-ocorrência de termos

```
@RELATION queries
@attribute car {sim, não}
@attribute family {sim, não}
@attribute travel {sim, não}
@attribute loan {sim, não}
@attribute school {sim, não}
@attribute debt {sim, não}
@attribute mortgage {sim, não}
@attribute college {sim, não}
@attribute savings {sim, não}
@attribute marriage {sim, não}
@attribute wedding {sim, não}
@attribute children {sim, não}
@attribute work {sim, não}
@attribute money {sim, não}
@attribute dog {sim, não}
@attribute party {sim, não}
@attribute beer {sim, não}
@attribute pizza {sim, não}
@attribute wine {sim, não}
@attribute liquor {sim, não}
@DATA
 ? , ? , ? , ? , ? , ? ,sim,sim,sim, ? , ? , ? , ? , ? ,sim, ? , ? , ? , ? , ?
 ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ?
 ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ?
 ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ?
sim, ? , ? , ? ,sim, ? , ? ,sim, ? ,sim, ? , ? , ? , ? , ? ,sim, ? , ? , ? , ?
 ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ? , ?
 ? , ? , ? , ? ,sim, ? , ? , ? , ? , ? , ? ,sim, ? ,sim, ? ,sim, ? ,sim, ? , ?
```

Estudo de caso: Co-ocorrência de termos

- Algoritmo APriori foi aplicado.

```
Best rules found:
 1. college=sim 2518 ==> school=sim 793     conf:(0.31)
 2. family=sim 1325 ==> school=sim 399     conf:(0.3)
 3. car=sim 2524 ==> school=sim 519     conf:(0.21)
 4. school=sim 3955 ==> college=sim 793     conf:(0.2)
 5. school=sim 3955 ==> car=sim 519     conf:(0.13)
 6. school=sim 3955 ==> family=sim 399     conf:(0.1)
```

- Uma interpretação simples da primeira regra é:
  - "Se um usuário procurou por *college* também procurou por *school*. 2518 usuários procuraram por *college* e 793 por *school*, e a confiança nesta regra é 0.31".
- É interessante ver que as regras não são reflexivas:
  - A regra 4 mostra quantas pessoas procuraram por *school* e também por *college* e a confiança é diferente (já que o antecedente e o conseqüente trocaram de posição).

Estudo de caso: Co-ocorrência de termos

- Apesar do algoritmo ser muito bom para encontrar *itemsets* ele pode ficar inaplicável se usarmos muito mais palavras-chave e/ou número de usuários.
- Busca por palavras-chave (tanto no exemplo quanto em mecanismos de busca) é ineficiente:
  - Questão da semântica (faculdade, universidade).
  - Representação não-hierárquica (álcool, bebida, cachaça)

Estudo de caso: E-mails da Enron

Estudo de caso: E-mails da Enron

- **Enron** era uma companhia americana de energia...
  - Considerada uma das maiores empresas do mundo (faturamento de 111 bilhões de dólares em 2000).
  - Eleita por seis anos consecutivos como “Companhia mais inovadora dos EUA” pela revista Fortune.
  - Altamente envolvida com financiamento de candidados políticos.
- Foi à falência em 2001, depois de um escândalo causado por declarações contábeis fraudulentas (perdas não declaradas).
  - Preço das ações caiu de Us$ 90.00 para Us$ 0.30.
  - Causou *enorme* tumulto social!

Estudo de caso: E-mails da Enron

- Em Outubro de 2004 a Comissão Federal de Regulamentação de Energia disponibilizou mensagens de e-mail de executivos da Enron como parte da investigação.
  - Diferentes versões com diferentes características foram disponibilizadas.
  - Houve *algum* esforço de proteção dos inocentes.
- Usaremos como exemplo a versão “raw”.
  - Aprox. 400 megabytes (.tar.gz)
  - ~1.8 gigabytes (.tar)
  - ~520.000 mensagens, organizadas por recipientes (153) e por *folders*.

Estudo de caso: E-mails da Enron

Message-ID: <32213030.1075858183820.JavaMail.evans@thyme>
Date: Wed, 25 Apr 2001 04:59:00 -0700 (PDT)
From: mheffner@carrfut.com
To: mike.maggi@enron.com
Subject: margin financing
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: MHeffner@carrfut.com
X-To: mike.maggi@enron.com
X-cc:
X-bcc:
X-Folder: \Michael_Maggi_Jun2001\Notes Folders\All documents
X-Origin: Maggi-M
X-FileName: mmaggi.nsf

we maybe slow, but we eventually get there,,

as you know Carr has been trying to get approval from within and from Enron finance people to create margin financing to execute & clear Nymex (and e-nymex too) business for Enron. Well we are finally there..
We would love the opportunity to renew our realtionship of executing and/or clearing for you again.

Estudo de caso: E-mails da Enron

- O que queremos encontrar?
- Existe algum padrão no uso de palavras-chave nas mensagens em função do tempo?
- Que informações e conhecimento podemos esperar de um processo automatizado?

Estudo de caso: E-mails da Enron

- *Existe algum padrão no uso de palavras-chave nas mensagens em função do tempo?*
- Pré-Processamento:
  - Ler todas as mensagens, descartando folders.
  - Armazenar nome do recipiente, mês/ano, ocorrência das palavras-chave *business, investigation, sell, president, government, corruption, millions, bankruptcy, profit, scandal.*
  - Descartar entradas que não contém nenhuma das palavras-chave.
    - Das 517431 mensagens selecionamos 86396

Estudo de caso: E-mails da Enron

- Arquivo no formato .arff (para o software Weka):

@relation enron
@attribute recipiente {stepenovitch-j, steffes-j, lavorato-j, parks-j, martin-t, hodge-j, lay-k, shapiro-r, scholtes-d, haedicke-m, keavey-p, corman-s, farmer-d, pereira-s, salisbury-h, gilbertsmith-d, quigley-d, **...** , staab-t}
@attribute data {Jun2002, Jun2001, Jun2000, Dec2043, Jan2000, Jan2001, Apr2002, Jan2002, Apr2001, Apr2000, Oct0001, May2024, Nov2000, Jan0002, Aug1999, Oct2002, Nov0001, Mar1997, Sep1997, Mar1999, Sep1998, Nov2012, Jul0001, Oct1997, Dec2000, Dec2001, Dec2002, Jul2001, Jul2000, May1986, Dec0001, Feb0002, Apr1986, Feb2007, Mar2002, Sep2002, Feb2004, Feb2002, Feb2001, Feb2000, Mar0002, Sep0001, **...** Dec2020, Feb1999, Jan2044, May2000, May2001, May0001, Dec1979, Jan1997, Jan1998, Aug0001, Jun0001, Nov1998, Aug2001, Aug2000}
@attribute business {Y,N}
@attribute investigation {Y,N}
@attribute sell {Y,N}
@attribute president {Y,N}
@attribute government {Y,N}
@attribute Bush {Y,N}
@attribute Clinton {Y,N}
@attribute corruption {Y,N}
@attribute millions {Y,N}
@attribute bankruptcy {Y,N}
@attribute profit {Y,N}
@attribute illegal {Y,N}
@attribute scandal {Y,N}
@data
wolfe-j,Feb2001,?,?,?,?,?,?,?,Y,?,?,?,?,?,?,?
wolfe-j,Feb2001,?,?,?,?,Y,Y,?,?,?,?,?,?,?,?,?
wolfe-j,Feb2001,Y,?,?,?,Y,?,?,?,?,?,?,?,?,?,?
...

Estudo de caso: E-mails da Enron

1. recipiente=lay-k millions=Y bankruptcy=Y profit=Y 1021 ==> data=Jan2002 1021 conf:(1)
2. recipiente=lay-k data=Jan2002 bankruptcy=Y profit=Y 1022 ==> millions=Y 1021 conf:(1)
3. recipiente=lay-k millions=Y bankruptcy=Y 1122 ==> data=Jan2002 1120 conf:(1)
4. recipiente=lay-k bankruptcy=Y profit=Y 1024 ==> data=Jan2002 1022 conf:(1)

...

15. recipiente=lay-k data=Jan2002 profit=Y 1029 ==> millions=Y bankruptcy=Y 1021 conf:(0.99)
16. data=Jan2002 millions=Y profit=Y 1069 ==> bankruptcy=Y 1056 conf:(0.99)

...

20. sell=Y government=Y millions=Y 914 ==> business=Y 885 conf:(0.97)
21. data=Jan2002 millions=Y bankruptcy=Y profit=Y 1056 ==> recipiente=lay-k 1021 conf:(0.97)
22. business=Y sell=Y government=Y Bush=Y 1049 ==> president=Y 1008 conf:(0.96)
23. data=Jan2002 millions=Y profit=Y 1069 ==> recipiente=lay-k 1026 conf:(0.96)
24. recipiente=lay-k millions=Y 1177 ==> data=Jan2002 1128 conf:(0.96)

...

27. sell=Y president=Y millions=Y 1011 ==> business=Y 964 conf:(0.95)
28. recipiente=lay-k millions=Y 1177 ==> bankruptcy=Y 1122 conf:(0.95)
29. sell=Y president=Y government=Y profit=Y 1086 ==> business=Y 1034 conf:(0.95)

...

33. sell=Y president=Y government=Y bankruptcy=Y 914 ==> business=Y 868 conf:(0.95)

...

37. sell=Y president=Y government=Y Bush=Y 1065 ==> business=Y 1008 conf:(0.95)
38. president=Y government=Y millions=Y 966 ==> business=Y 914 conf:(0.95)

Estudo de caso: E-mails da Enron

54. president=Y government=Y profit=Y 1325 ==> business=Y 1210 conf:(0.91)
55. president=Y government=Y bankruptcy=Y 1135 ==> business=Y 1035 conf:(0.91)
72. sell=Y president=Y bankruptcy=Y 1256 ==> business=Y 1126 conf:(0.9)
73. Bush=Y profit=Y 1154 ==> president=Y 1031 conf:(0.89)
78. recipiente=beck-s 2741 ==> business=Y 2438 conf:(0.89)
79. sell=Y president=Y Bush=Y 1367 ==> business=Y 1215 conf:(0.89)
110. recipiente=mcconnell-m 1192 ==> business=Y 996 conf:(0.84)
131. investigation=Y president=Y 1444 ==> business=Y 1150 conf:(0.8)
172. recipiente=jones-t 2262 ==> business=Y 1644 conf:(0.73)
197. data=Jul2000 1996 ==> business=Y 1366 conf:(0.68)
358. government=Y 11845 ==> business=Y 5499 conf:(0.46)
372. recipiente=shapiro-r 2188 ==> government=Y 973 conf:(0.44)
470. Bush=Y 5559 ==> president=Y government=Y 1754 conf:(0.32)
681. Bush=Y 5559 ==> government=Y profit=Y 865 conf:(0.16)
705. profit=Y 7099 ==> president=Y Bush=Y 1031 conf:(0.15)
777. government=Y 11845 ==> business=Y president=Y profit=Y 1210 conf:(0.1)
778. government=Y 11845 ==> business=Y bankruptcy=Y 1209 conf:(0.1)
779. government=Y 11845 ==> data=Oct2001 1200 conf:(0.1)
780. president=Y 17943 ==> business=Y profit=Y 1813 conf:(0.1)
781. business=Y president=Y 9274 ==> Bush=Y profit=Y 936 conf:(0.1)
782. government=Y 11845 ==> recipiente=dasovich-j business=Y 1187 conf:(0.1)

Estudo de caso: E-mails da Enron

- Podemos observar que...
  - Várias regras aparecem em combinações ligeiramente diferentes.
  - Muitas regras foram geradas por uma campanha de e-mails dirigida a um dos diretores da empresa.
- Algumas regras são aparentemente óbvias...
  - “president” e “Bush”.
  - “business” e alguns destinatários.
- Foi solicitado um número grande de regras, mas tivemos que praticamente minerá-las!

Estudo de caso: E-mails da Enron

- Problemas com mineração de dados de e-mails:
  - Conteúdo não textual pode ser relevante, mas mineração multimídia é difícil!
  - Formatos alternativos de texto podem também complicar.
- Ainda problemas com semântica e hierarquia de conceitos...
- Problemas sérios com falta de normalização (ex. “Jun2001” e “Jun0001”).
- Que outras buscas genéricas podiam ser feitas?
  - Aprecie a diferença entre uma busca com *grep* ou via SQL.

Estudo de caso: E-mails da Enron

- Referências:
  - O conjunto de dados (~400M!) pode ser copiado de http://www.cs.cmu.edu/~enron/ (cortesia de William Cohen)
  - Ron Bekkerman da UMASS fez uma caracterização da distribuição de mensagens em *folders*. *Datasets* parciais, comentários e resultados podem ser vistos em http://www.cs.umass.edu/~ronb/enron_dataset.html
  - Jitesh Shetty e Jafar Adibi (USC) tem informações para enriquecimento em http://www.isi.edu/~adibi/Enron/Enron.htm
- Questões a considerar:
  - O conjunto original tem problemas de integridade!
  - Privacidade!

Conclusões

Problemas potenciais com *Data Mining*

- O problema pode não ser corretamente interpretado.
- Padrões encontrados podem não ser interessantes.
- Padrões encontrados podem não ser explicáveis.
- Padrões podem ser incorretos (modelos construídos de dados inadequados), explicando o fenômeno medido de forma incorreta.
- Pré-processamento dos dados pode ser feito de forma incorreta, por exemplo, com relações redundantes ou ruídos, seleção de dados inadequada, etc.

Problemas potenciais com *Data Mining*

- Dados podem não ser confiáveis ou ser pouco representativos das classes/categorias/fenômenos existentes.
- Modelos, algoritmos, parâmetros podem ser estimados incorretamente!
- Modelos, algoritmos, parâmetros podem não representar a realidade adequadamente (ex. *underfitting* e *overfitting*).
- Lembrando: *Data Mining* não é mágica!: ***Garbage in, Garbage out***.

Problemas potenciais com *Data Mining*

- O formato e tipo de dados coletados em *logs* é extremamente complexo.
  - Como adequar para uso com algoritmos mais poderosos?
  - Considere somente *timestamp*: como verificar periodicidade ou freqüência?
- Pré-processamento é **crucial** e deve ser feito por um especialista no domínio.
- Mineração de conteúdo (texto) é **extremamente** complexa.
  - Considere a semântica!

Problema maior...

- Privacidade.
- Cada vez mais dados coletados, sem conhecimento ou consentimento de usuários.
- Mesmo com anonimização é possível inferir comportamentos de grupos!
- Problemas legais e sociais em potencial!

Obrigado!

Perguntas?