MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

Mineração de Dados

CAP 349: Bancos de Dados Geográficos 2014

Adeline Marinho
Doutorado – CAP – INPE
adelsud6@gmail.com

8 setembro 2014

Introdução

- Crescimento explosivo das bases de dados
  - Governo, corporações, institutos científicos
  - *Internet*
- Demanda por informações estratégicas
  - Meios convencionais para examinar dados
  - Análise automática e inteligente de grandes BDs
- Disponibilidade de dados
  - Armazenamento eletrônico

Cenário

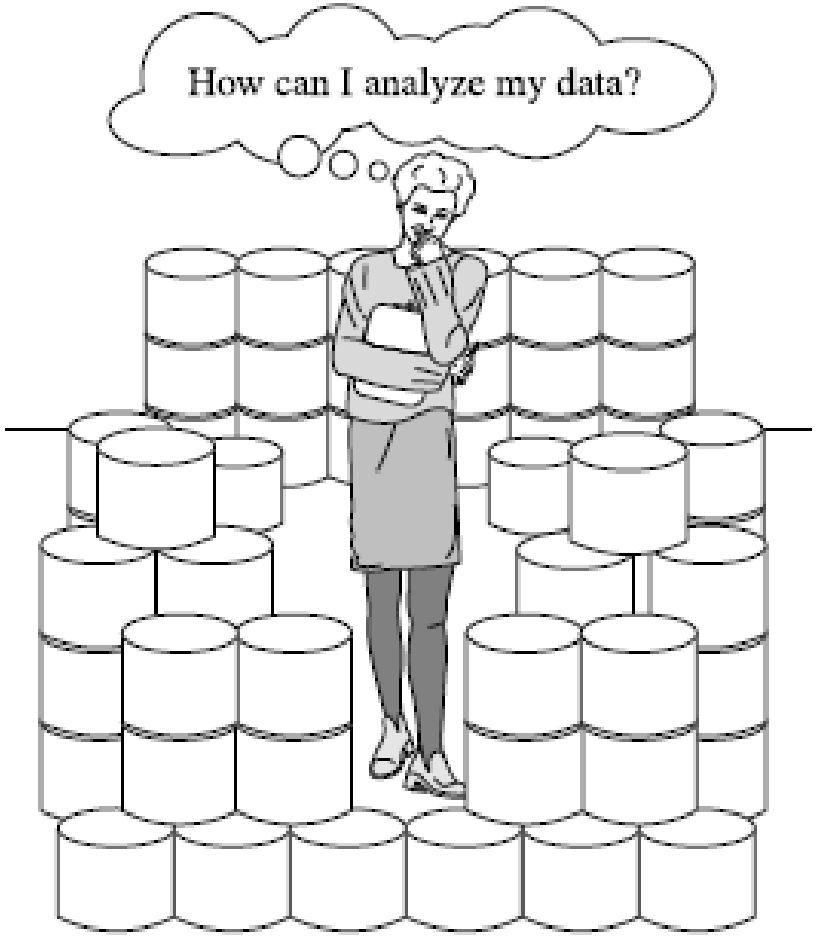
Hoje, a maioria das organizações produz mais informações em uma semana do que muitas pessoas poderiam ler em toda vida.

"MAIS DADOS PODE SIGNIFICAR MENOS INFORMAÇÕES"

Cenário

- Dificuldade para analisar dados
  - métodos manuais tradicionais
- A informação não está explicitada
- Necessidade
  - técnicas que facilitem a extração da informação
  - dado operacional não oferece grande valor quando estudado isoladamente.

O que fazer?



Dados ricos, porém pobres em informação

Fonte: (HAN e KAMBER, 2006)

Informação e Conhecimento

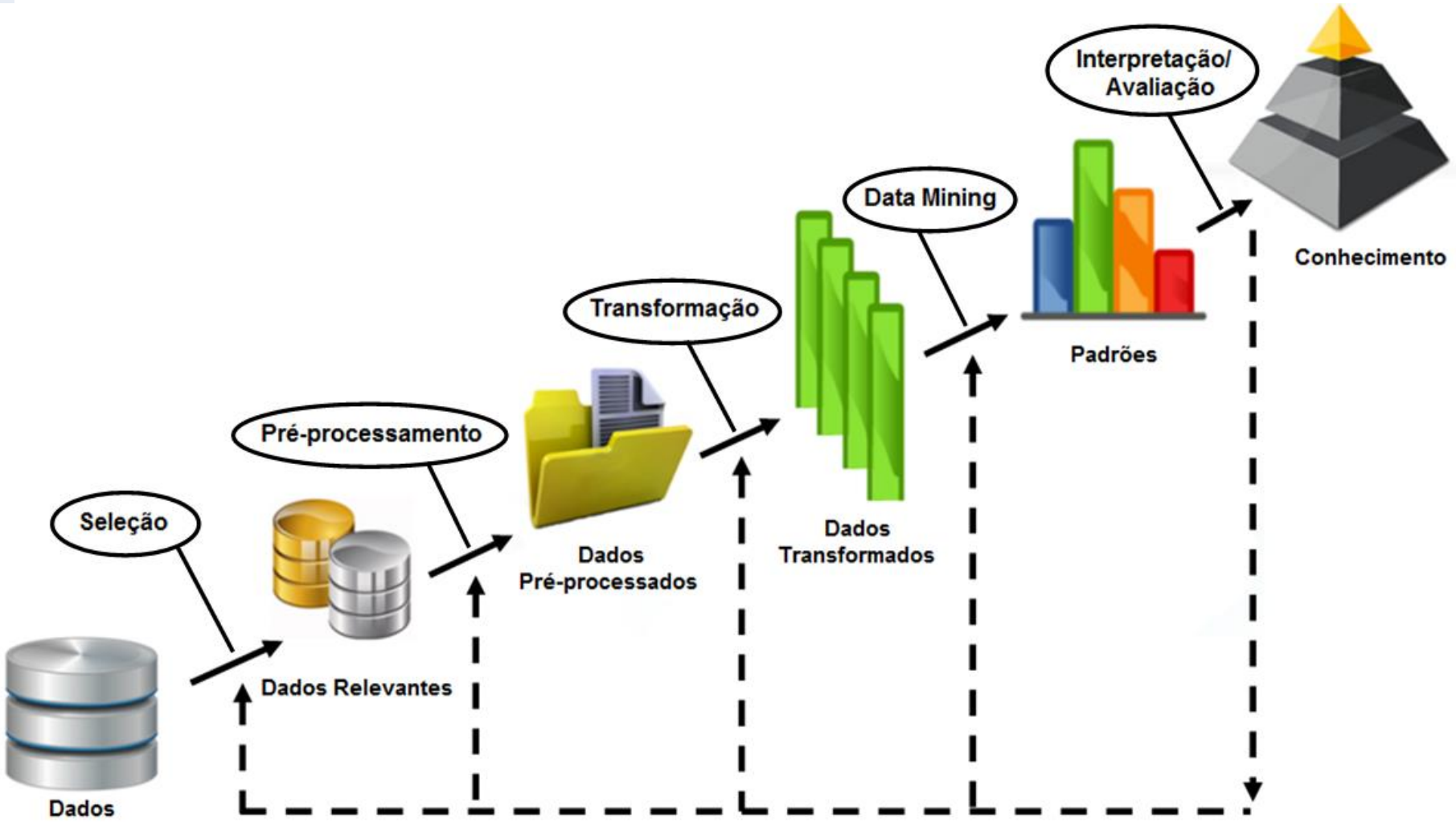
- Segundo Han e Kamber (2006)
  - A abundância de dados
  - A necessidade de ferramentas de análise de dados

- É descrita como uma situação de dados ricos, mas pobres em informação.
  - Os dados tornam-se "dados túmulos" arquivos de dados que raramente são visitados.

Processo de KDD

- Descoberta de Conhecimento em BDs
  - *Knowledge Discovery in Databases* (KDD) (FAYYAD et al., 1996)
  - Ferramentas e técnicas empregadas para análise automática e inteligente de imensos repositórios
- Processo não trivial de identificar em dados padrões que sejam
  - Válidos
  - Novos (previamente desconhecidos)
  - Potencialmente úteis
  - Compreensíveis
- Visando melhorar o entendimento de um problema ou um procedimento de tomada de decisão.

Etapas do KDD



Etapas do processo KDD

Etapas do KDD

- Fases/Etapas do KDD
  - Seleção.
  - Pré-processamento.
  - Transformação.
  - Mineração de dados
  - Interpretação e Avaliação.

Etapas do KDD

- **Seleção**
  - Selecionar ou segmentar dados de acordo com critérios definidos:
    - Ex.: Todas as pessoas que são proprietárias de carros é um subconjunto de dados determinado.

Etapas do KDD

- **Pré-processamento**
  - Estágio de limpeza dos dados, onde informações julgadas desnecessárias são removidas.
    - Ex. : O sexo de um paciente, gestante
  - Reconfiguração dos dados para assegurar formatos consistentes (identificação)
    - Ex. : sexo = “F” ou “M”
      sexo = “M” ou “H”
  - Agregar dados externos

Etapas do KDD

- **Transformação**
  - Transforma-se os dados em formatos utilizáveis.
  - Depende da técnica *data mining* usada.
    - converter valor literal em valor numérico
  - Normalização de valores
  - Agregar semântica ao dado
  - Disponibilizar os dados de maneira usável e navegável.

Etapas do KDD

- **Mineração de Dados**
  - É a verdadeira extração dos padrões de comportamento dos dados
  - Utilizando a definição de fatos, medidas de padrões, estados e o relacionamento entre eles.

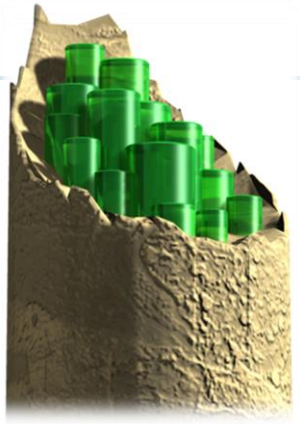
Etapas do KDD

- **Interpretação e Avaliação**
  - Identificado os padrões pelo sistema, estes são interpretados em conhecimentos, os quais darão suporte a tomada de decisões humanas
  - Ex.: Tarefas de previsões e classificações

Processo de KDD

- Alguns possíveis resultados são:
  - Confirmação do óbvio
  - Conhecimento novo
  - Nenhum relacionamento encontrado (dados aleatórios)
- Problemas:
  - Identificação dos dados relevantes
  - Representação dos dados
  - Busca por modelos ou padrões válidos

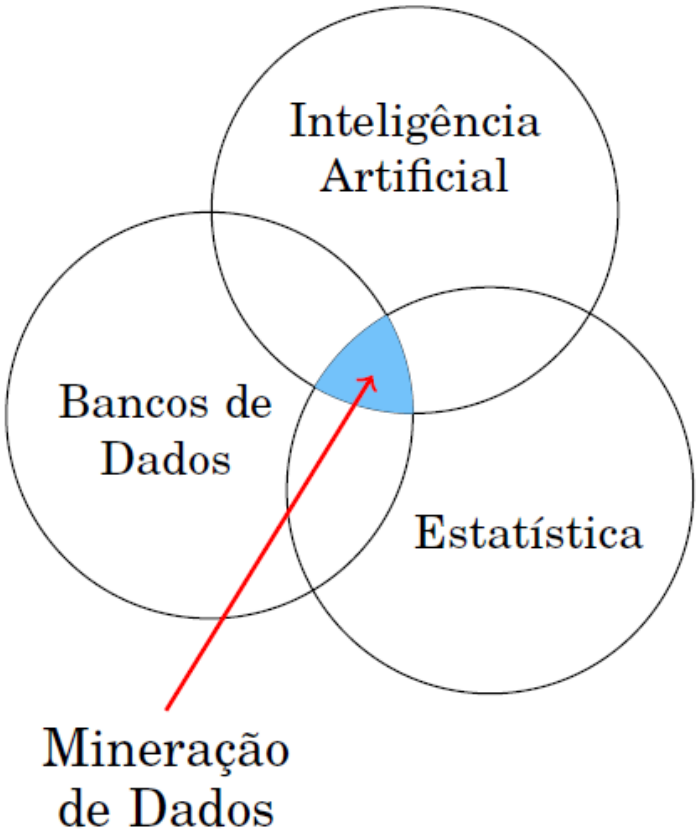
Mineração de Dados



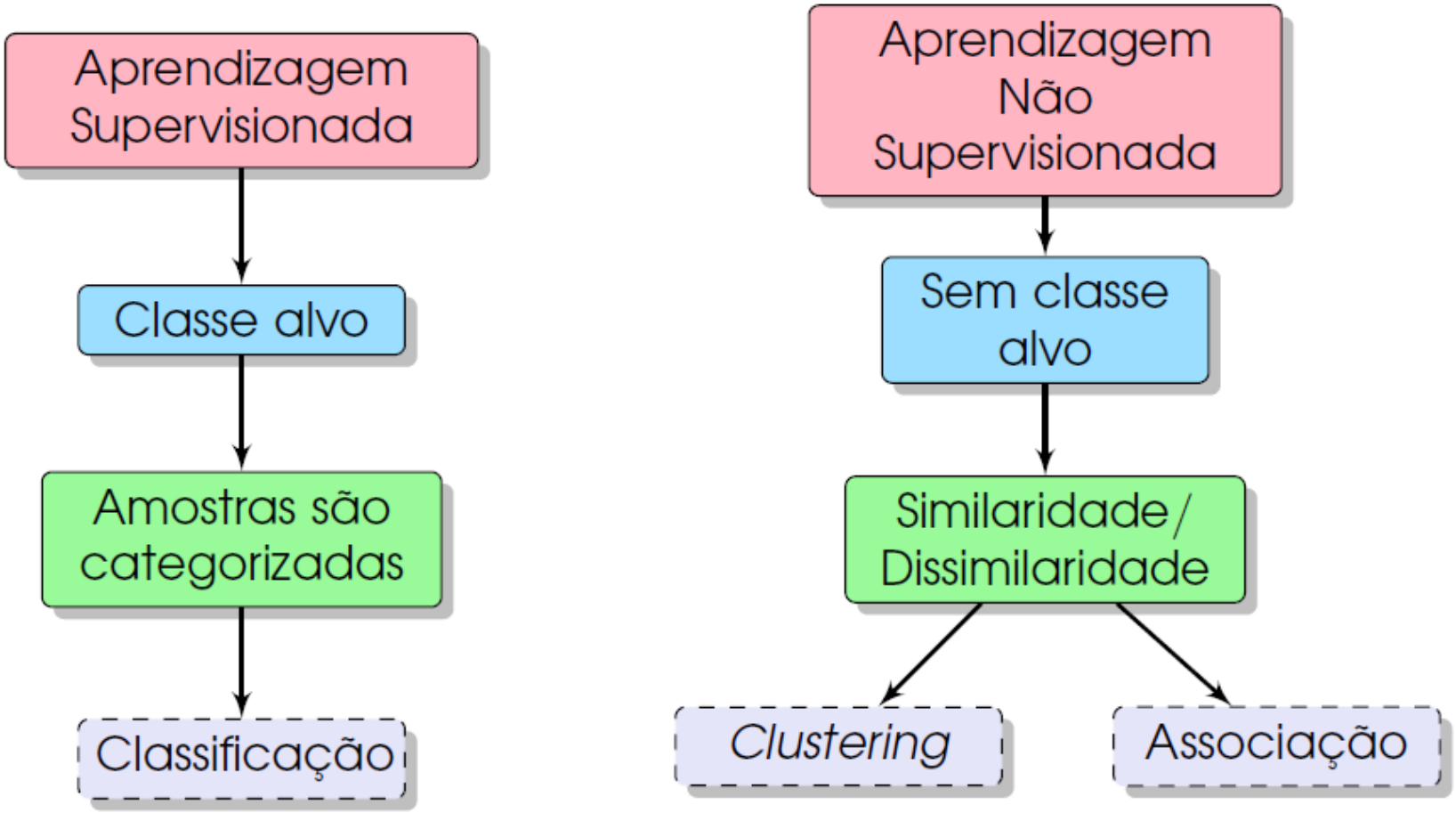
- É uma das etapas no KDD
- A mineração de dados preocupa-se em buscar conhecimento compreensível em grandes conjuntos de dados.

Mineração de Dados

- Utiliza técnicas e algoritmos de diferentes áreas do conhecimento
  - inteligência artificial
    - aprendizagem de máquina
  - banco de dados
    - recursos para manipular grandes bases de dados
  - Estatística
    - avaliação e validação de resultados



Técnicas de Mineração de Dados



Aprendizagem Supervisionada

- Aprendizagem do modelo é supervisionada
  - É fornecida uma classe à qual cada amostra no treinamento pertence

- Algoritmos preditivos
  - suas tarefas de mineração desempenham inferências nos dados;
  - fornecem previsões ou tendências, obtendo informações não disponíveis a partir dos dados disponíveis.

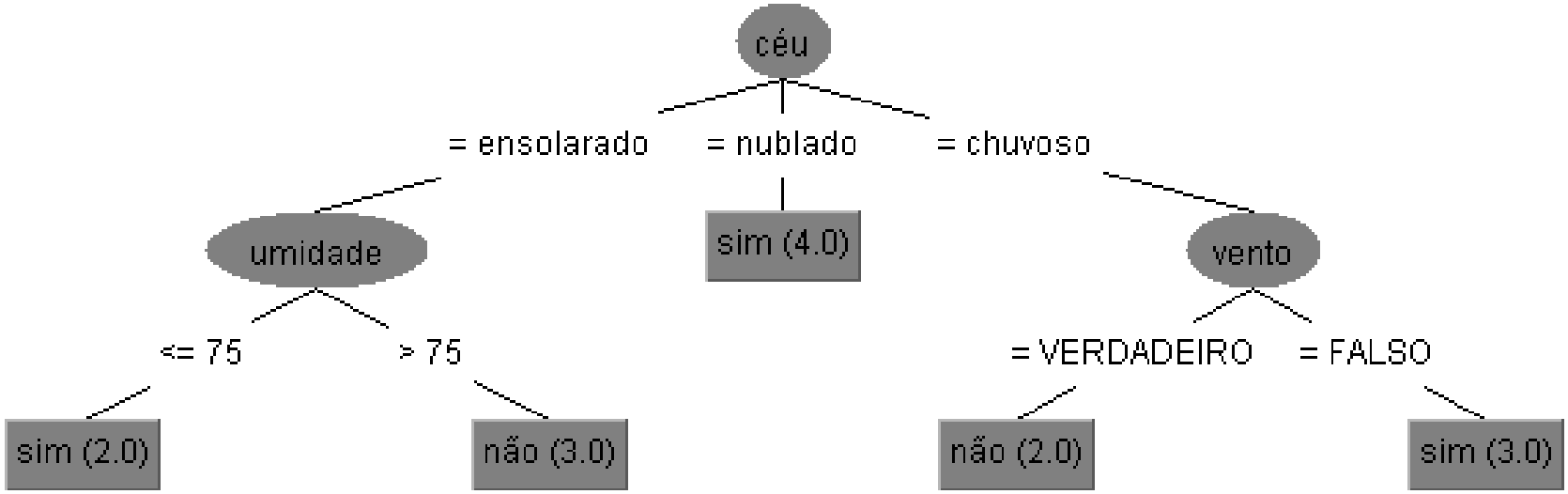
Aprendizagem Supervisionada

- **Classificação**
  - determina o valor de um atributo *(classe)* através dos valores de um subconjunto dos demais atributos.

- Ex. Qual o perfil dos clientes que consomem cosméticos importados?
  - Inferir (Prever) - "clientes do sexo feminino, com renda superior a R$ 1.500,00 e com idade acima de 30 anos compram cosméticos importados "
  - Atributo compra cosmético importado = classe
    - Atributo alvo (Sim ou Não)

Aprendizagem Supervisionada

- **Classificação**
  - Formas mais comuns de representação: regras e árvores.



Árvore de decisão

Aprendizagem Supervisionada

- **Seleção de atributos**
  - Alguns atributos têm um peso maior ou até determinante nas tarefas de mineração de dados
    - ex.: o atributo renda é determinante nos hábitos de consumo do cliente
  - Com os algoritmos é possível determinar os atributos relevantes para a mineração separando-os dos atributos irrelevantes
    - ex.: nome do cliente caso não influenciem nos hábitos de consumo

Aprendizagem Não-Supervisionada

- O rótulo da classe de cada amostra do treinamento não é conhecida
  - Número ou conjunto de classes a ser treinado pode não ser conhecido a priori

- Descritivos
  - Descreve de forma concisa os dados disponíveis
  - Fornece características das propriedades gerais dos dados minerados.

Aprendizagem Não-Supervisionada

- **Associação**
  - A classe da tarefa de mineração não é determinada
  - Gera regras do tipo:
    - clientes do sexo masculino, casados, com renda superior a R$ 1.800,00 têm o seguinte hábito de consumo: *roupas de grife, perfumes nacionais, relógios importados*
  - Revela associações entre valores dos atributos

Aprendizagem Não-Supervisionada

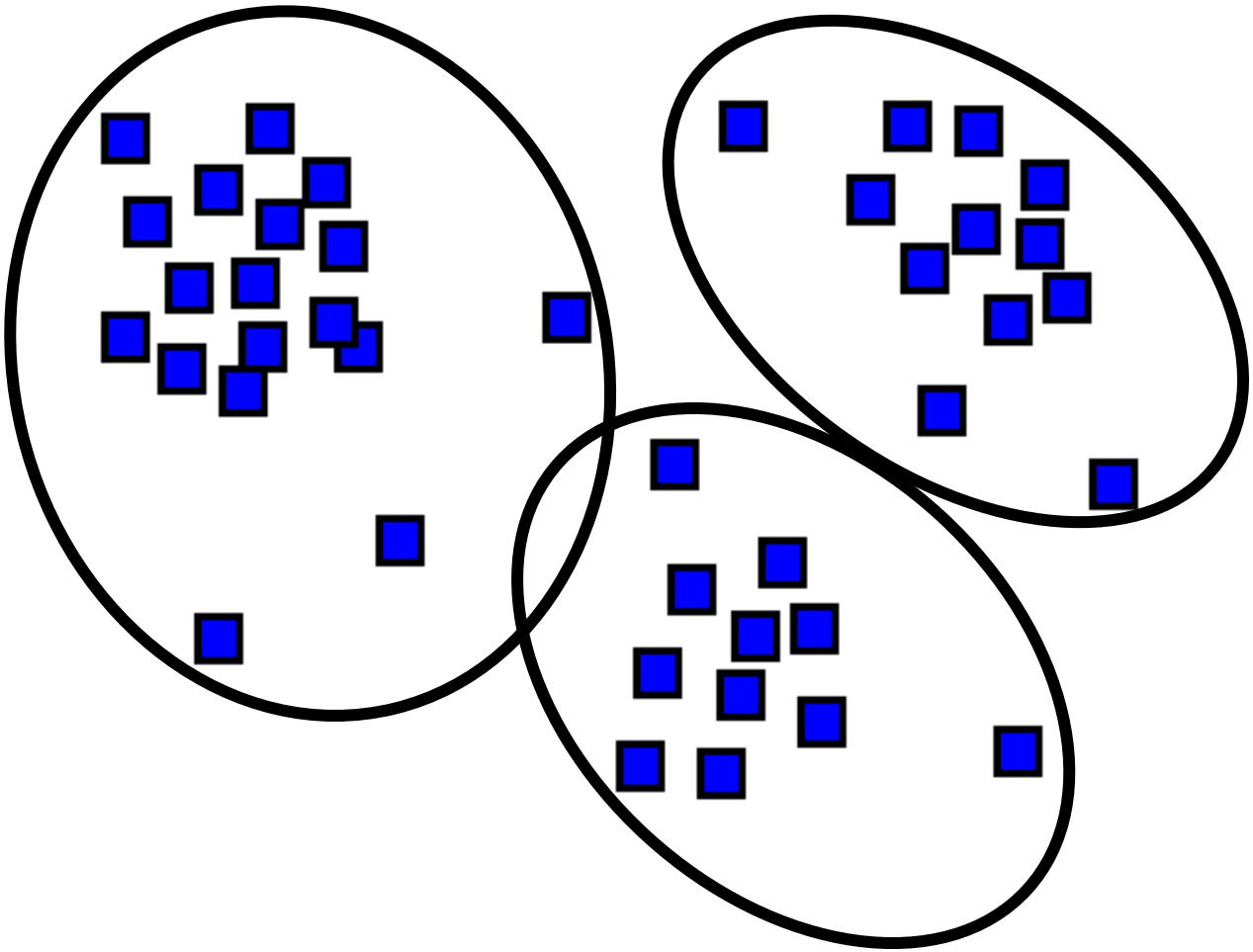
- Exemplo de Regras de associação

```
 1. classe=tripulação 885 ==> idade=adulto 885    conf:(1)
 2. classe=tripulação sexo=masculino 862 ==> idade=adulto 862    conf:(1)
 3. sexo=masculino sobreviveram=não 1364 ==> idade=adulto 1329    conf:(0.97)
 4. classe=tripulação 885 ==> sexo=masculino 862    conf:(0.97)
 5. classe=tripulação idade=adulto 885 ==> sexo=masculino 862    conf:(0.97)
 6. classe=tripulação 885 ==> idade=adulto sexo=masculino 862    conf:(0.97)
 7. sobreviveram=não 1490 ==> idade=adulto 1438    conf:(0.97)
 8. sexo=masculino 1731 ==> idade=adulto 1667    conf:(0.96)
 9. idade=adulto sobreviveram=não 1438 ==> sexo=masculino 1329    conf:(0.92)
10. sobreviveram=não 1490 ==> sexo=masculino 1364    conf:(0.92)
```

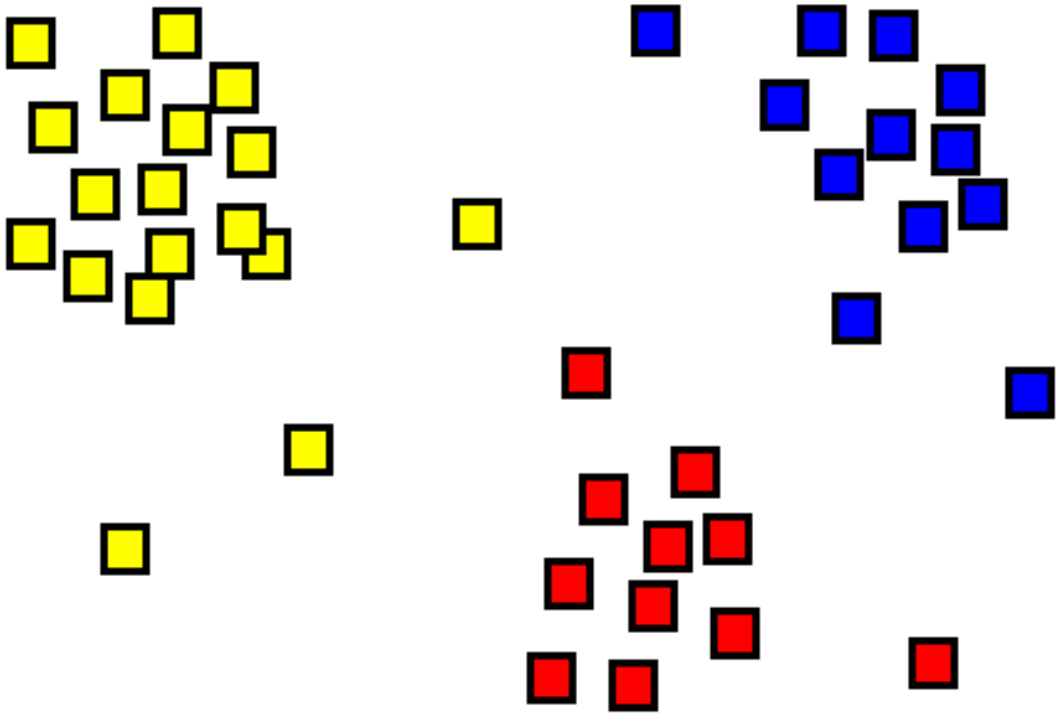
Aprendizagem Não-Supervisionada

- *Clustering*
  - Verifica como as instâncias de uma determinada base de dados se agrupam
    - considera características intrínsecas de seus atributos, sem que seja definida uma classe para a tarefa
  - A partir de uma métrica de similaridade
    - objetos são agrupados com base no princípio da maximização da similaridade intraclasse e da minimização da similaridade interclasse
  - ex.: identificar subgrupos homogêneos de clientes

Clustering



Clustering



Validação de Resultados

- É importante que os resultados e modelos possam ser avaliados e comparados
- Teste e validação
  - fornecem parâmetros de validade e confiabilidade nos modelos gerados
    - *cross validation*, *suplied test set*, *training set*, *percentage split*
  - indicadores estatísticos para auxiliar a análise dos resultados
    - matriz de confusão, estatística kappa, erro médio absoluto, precisão...

Medidas de interesse

Artigo (item)	número que o representa
Pão	1
Leite	2
Açúcar	3
Papel Higiênico	4
Manteiga	5
Fralda	6
Cerveja	7
Refrigerante	8
Iogurte	9
Suco	10

Representação numérica de cada artigo do supermercado

Um banco de dados de transações de clientes

TID	Itens comprados
101	{1,3,5}
102	{2,1,3,7,5}
103	{4,9,2,1}
104	{5,2,1,3,9}
105	{1,8,6,4,3,5}
106	{9,2,8}

Medidas de Interesse - Confidência

- Grau de confiança
  - conf(A=>B)
  - A porcentagem das transações que suportam B dentre todas as transações que suportam A

$$conf(A \rightarrow B) = \frac{\text{número de transações que suportam } (A \cup B)}{\text{número de transações que suportam } A}$$

  - Ex. O grau de confiança da regra
  - {cerveja}=>{manteiga}, isto é {7}=>{5}, com relação ao banco de dados é 1{100%}

Confidência

Artigo (item)	número que o representa
Pão	1
Leite	2
Açúcar	3
Papel Higiênico	4
Manteiga	5
Fralda	6
Cerveja	7
Refrigerante	8
Iogurte	9
Suco	10

Representação numérica de cada artigo do supermercado

Um banco de dados de transações de clientes

TID	Itens comprados
101	{1,3,5}
102	{2,1,3,7,5}
103	{4,9,2,1}
104	{5,2,1,3,9}
105	{1,8,6,4,3,5}
106	{9,2,8}

Conf {7}->{5} = 1/1 = 1

Confidência

- Seja uma regra "A => B".
- A confidência da regra é dada por:

$$\text{Confidência}\ (A \Rightarrow B) = \frac{\#\_\text{Tuplas\_Contendo\_Tanto\_A\_Como\_B}}{\#\_\text{Tuplas\_Contendo\_A}}$$

- Exemplo:
  - Uma confidência de 85% (0,85) da regra:

compra(mulher, computadores) =>compra(mulher, software)

  - *S*ignifica que 85% das mulheres que compram computadores também compram software.

Medidas de Interesse - Suporte

- Função do suporte
  - determinar a frequência que ocorre um *itemset* dentre todas as transações da base de dados
    - é a porcentagem de transações onde este *itemset* aparece
  - Um *itemset* será frequente se seu suporte for maior ou igual a um suporte mínimo estabelecido previamente.
- Forma:
  - X =>Y (lê-se X implica em Y)
    - onde X é o antecessor e Y o consequente
    - X e Y são dois *itemsets* distintos na Base de Dados

Suporte

- O suporte da regra "A => B" é dado por:

$$\text{Suporte}\ (A \Rightarrow B) = \frac{\#\_\text{Tuplas\_Contendo\_Tanto\_A\_Como\_B}}{\#\_\text{Total\_De\_Tuplas}}$$

$$\text{Sup}\ (X\ U\ Y) = \frac{\text{Nº de registros com}\ (X\ U\ Y)}{\text{Nº total de transações do BD}}$$

  - Um suporte de 5% significa que de todas as transações comerciais realizadas, 5% são efetuadas por *mulheres que comprando computador também compram softwares.*

Suporte

Um banco de dados de transações de clientes

TID	Itens comprados
101	{1,3,5}
102	{2,1,3,7,5}
103	{4,9,2,1}
104	{5,2,1,3,9}
105	{1,8,6,4,3,5}
106	{9,2,8}

Artigo (item)	número que o representa
Pão	1
Leite	2
Açúcar	3
Papel Higiênico	4
Manteiga	5
Fralda	6
Cerveja	7
Refrigerante	8
Iogurte	9
Suco	10

Representação numérica de cada artigo do supermercado

Itemset	Suporte
{1,3}	0,6666
{2,3}	0,3333
{1,2,7}	0,16666
{2,9}	0,5

Suporte de alguns *itemsets*

Sup **{7}->{5} = 1/6 = 0.16...**

Critérios de Comparação

- Critérios para comparar métodos e resultados de mineração de dados permitem avaliar e optar pelo melhor custo/benefício
  - Precisão avaliativa ou preditiva
    - habilidade do modelo para avaliar ou prever corretamente classes, agrupamentos, regras.
  - Velocidade
    - custo computacional da geração e utilização do modelo

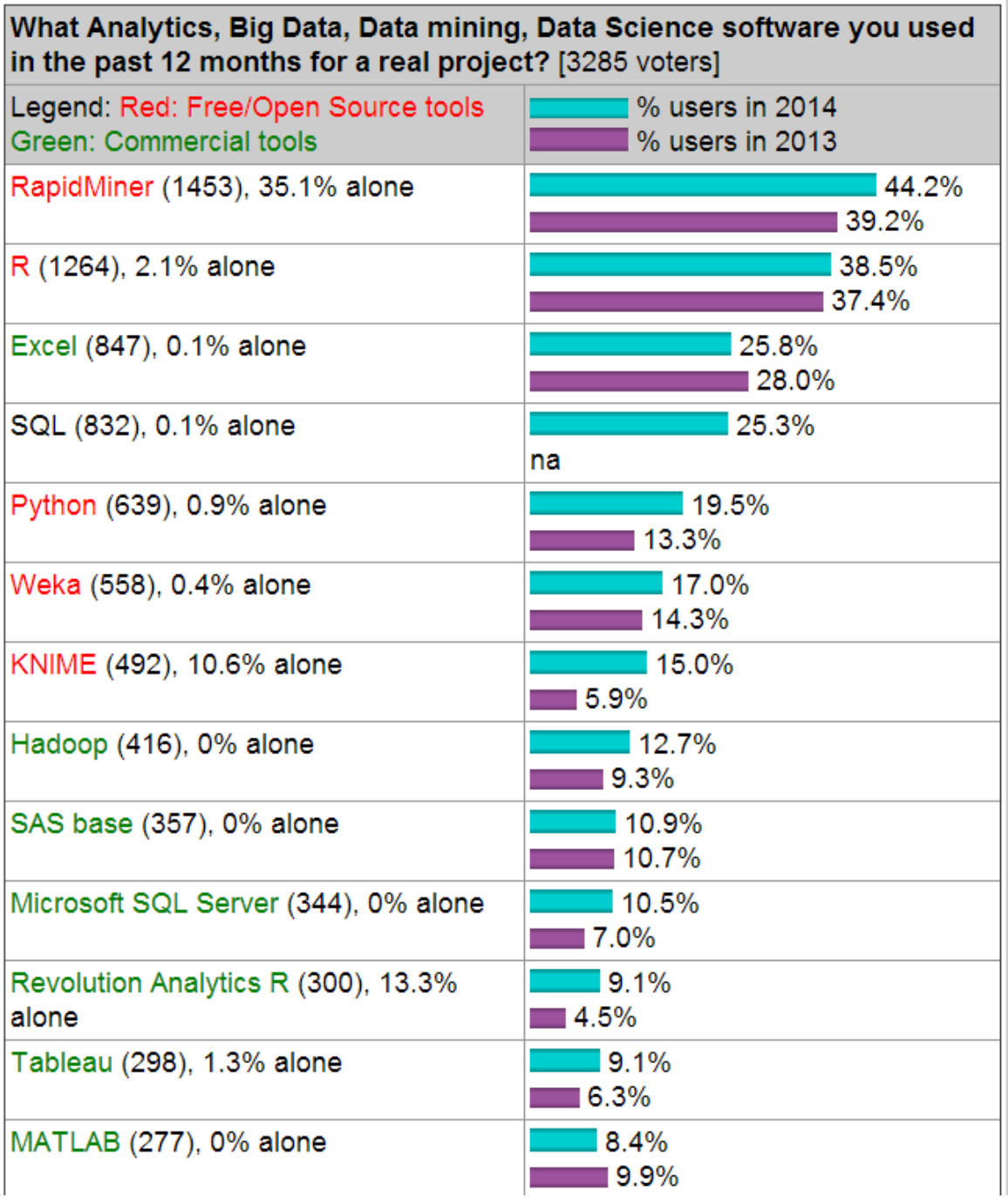
Critérios de Comparação

- Robustez
  - habilidade do modelo para avaliar ou prever corretamente utilizando dados ruidosos ou com valores ausentes
- Escalabilidade
  - capacidade de construir modelos eficientemente a partir de grandes volumes de dados
- Interpretabilidade
  - nível de compreensão fornecido pelo modelo

Softwares

KDnuggets Polls

May 20 - Jun 2, 2014



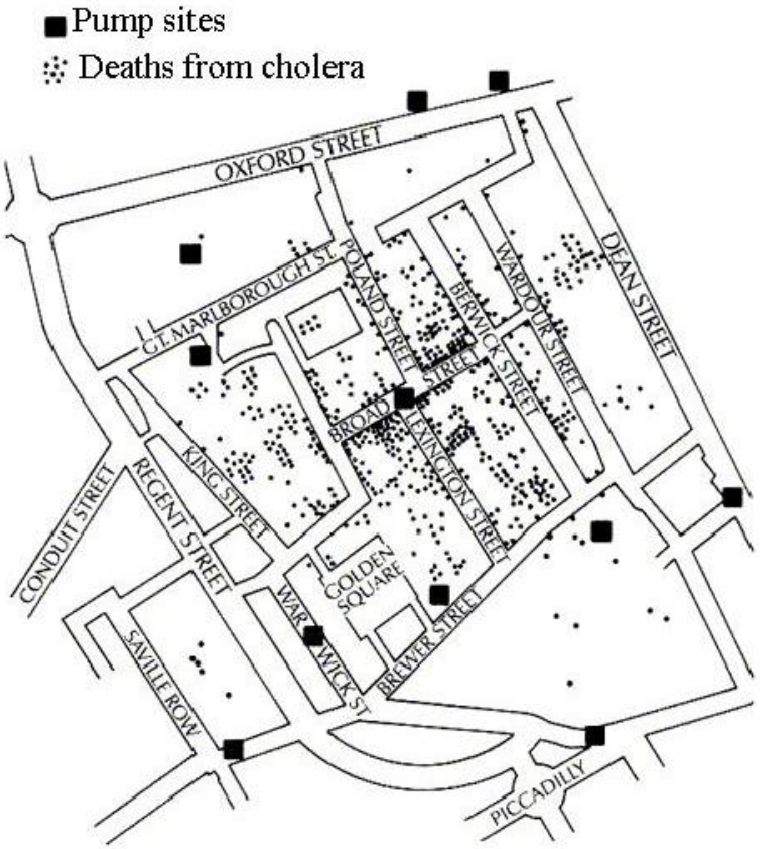
Fonte: http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-software-used.html

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

Mineração de Dados Espaciais

Exemplo Histórico

- Cólera Asiática em Londres (1855): um poço identificado como a fonte do problema

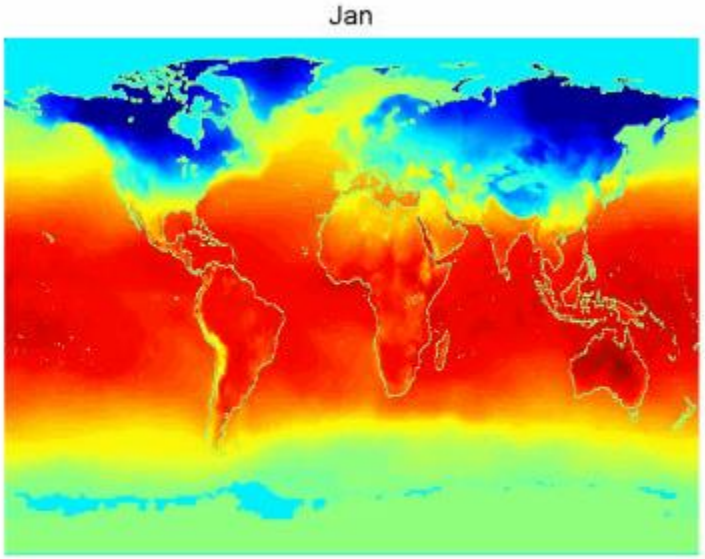


Exemplos Modernos

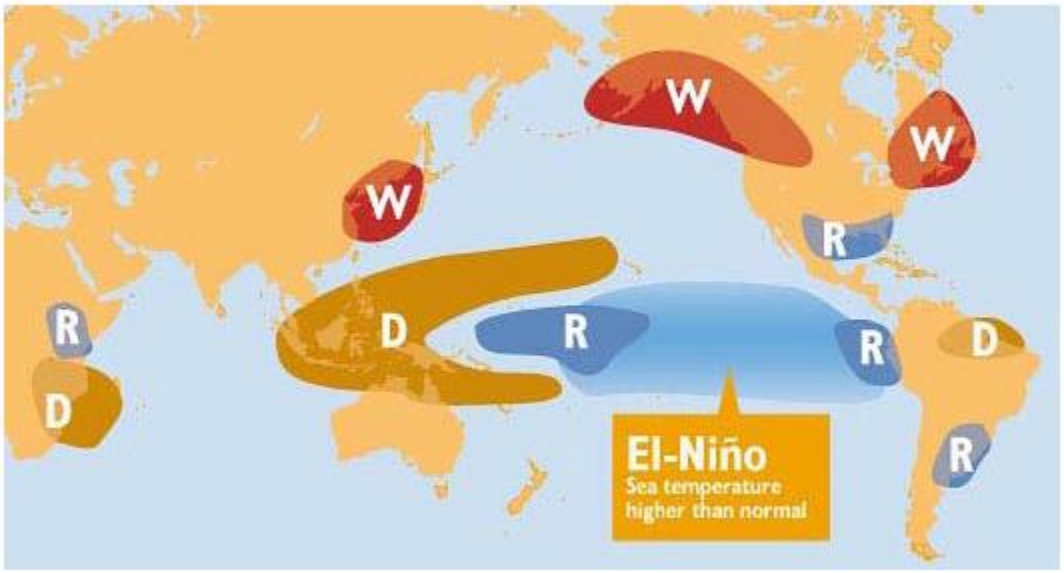
- *Clusters* de pessoas com cancer para investigar a influência do ambiente
- Locais de concentração de crimes para planejar as rotas de patrulha da polícia
- Identificação de características onde certos tipos de águias fazem os ninhos

Exemplos Modernos

- Aquecimento anormal de região do oceano Pacífico (El Niño) afeta o clima



Average Monthly Temperature
(Courtsey: NASA, Prof. V. Kumar)



Global Influence of El Nino during the Northern Hemisphere Winter
(D: Dry, W: Warm, R: Rainfall)

O que é um padrão espacial?

- O que não é um padrão?
  - Aleatório, ao acaso, acidental
  - Casual

- O que é um padrão (*pattern*)?
  - Um arranjo frequente, configuração, regularidade
  - Uma regra, lei, método
  - Uma direção ou tendência importante
  - Uma irregularidade espacial importante

O que é descoberta de conhecimento em dados espaciais?

- Definindo *Data Mining* Espacial
  - Procura por padrões espaciais
  - Procura não-trivial tão automática quanto possível
    - reduzindo o esforço humano
  - Padrões espaciais interessantes, úteis e inesperados
    - Desconhecidos
  - Auxílio especialista do domínio

O que é *data mining* espacial?

- Busca não trivial por padrões espaciais **interessantes** e **desconhecidos**
- Busca não trivial
  - Grande (ex. exponencial) espaço de busca de hipóteses plausíveis
  - Ex. Cólera Asiática - causas plausíveis: água, alimento, ar, insetos, …;
- Interessante
  - Útil em algum domínio de aplicação
  - Ex. Desativando o poço identificado => salvar vidas humanas
- Inesperado
  - O padrão não é conhecimento comum
  - Pode levar a um novo entendimento do mundo
  - Ex. A conexão Poço - Colera levou a teoria do "germe"

O que NÃO é *data mining* espacial?

- Consultas simples a dados espaciais
  - Encontre os vizinhos de Porto Alegre dados os nomes e limites de todas as cidades
  - Encontre o menor caminho do RS a SP na malha de rodovias
  - O espaço de busca não é grande (não é exponencial)

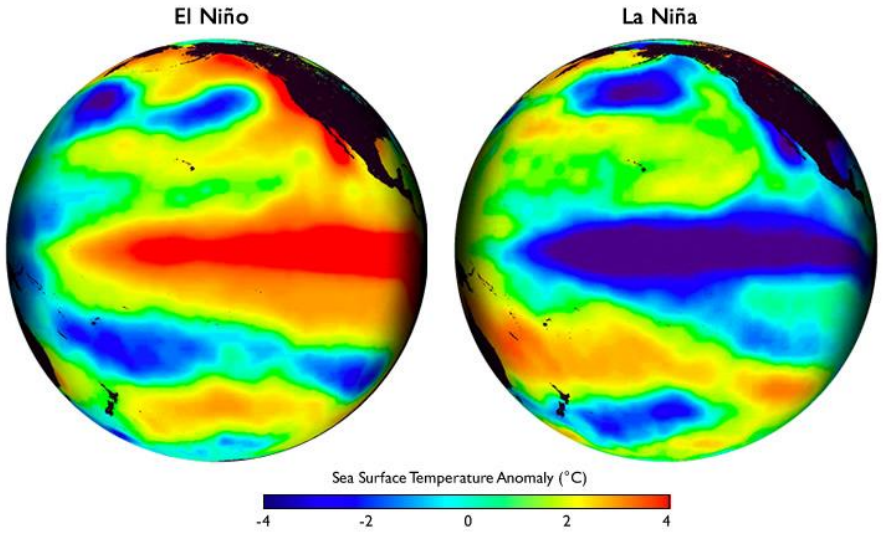
- Testar uma hipótese através de uma análise simples de dados
  - Ex. O território das chimpanzes femeas é menor do que o dos machos
  - O espaço de busca não é grande!

O que NÃO é *data mining* espacial?

- Padrões espaciais não interessantes ou óbvios
  - Muita chuva em Porto Alegre está correlacionada com muita chuva em Canoas, dado que as duas cidades são vizinhas.
  - Conhecimento comum: lugares próximos tem precipitações similares
- Mineração de dados não espaciais
  - As vendas de fraldas e cervejas são correlacionadas nas sextas-feiras

Porque estudar mineração de dados espaciais?

- Novo conhecimento dos processos geográficos para questões críticas
  - Ex. Como está a saúde do planeta Terra?
  - Ex. Caracterizar os efeitos da atividade humana para o ambiente e a ecologia
  - Ex. Predizer o efeito do El Niño no clima e na economia



Porque estudar mineração de dados espaciais?

- Abordagem tradicional: gerar e testar hipóteses manualmente
  - Mas os dados espaciais estão crescendo rápido demais para uma análise manual
    - Imagens de satélite, trajetórias geradas por GPS, sensores em rodovias, …
  - Número de hipóteses geográficas possíveis é grande demais para uma análise manual
    - Grande número de objetos geográficos
    - O número de relacionamentos entre os objetos cresce exponencialmente
    - Ex. Encontre correlação entre eventos climáticos oceânicos e em terra firme
- *Data Mining* Espacial pode reduzir o conjunto de hipóteses plausíveis

Escolha dos métodos

- Duas abordagens:
  - Uso de técnicas específicas para mineração de dados espaciais
  - Obtenção dos dados ou relacionamentos espaciais de interesse para uso com métodos de DM clássicos

- Abordagem possível:
  - Defina o problema: obtenha as necessidades particulares
  - Analise os dados usando mapas e outras técnicas de visualização
  - Tentar usar métodos clássicos de *data mining*
    - Se não obtiver resultados satisfatórios, tente novos métodos
  - Avalie os métodos escolhidos rigorosamente

Famílias de padrões espaciais

- Famílias comuns de padrões espaciais
  - Classificação
  - *Clustering*
  - Detecção de *Outliers*
  - Co-location
  - Trajetórias
  - ….
- Outras famílias podem ser definidas

Classificação

- Dado um conjunto de instâncias, a função da classificação é descobrir as classes das instâncias

- Objetos podem ser caracterizados, classificados, por diferentes tipos de informações
  - Atributos não espaciais (população)
  - Atributos espacialmente relacionados com valores não-espaciais (população total que vive a menos de 100 metros das antenas celulares);

Remote Sensing Data Mining (Silva *et al.* 2008)

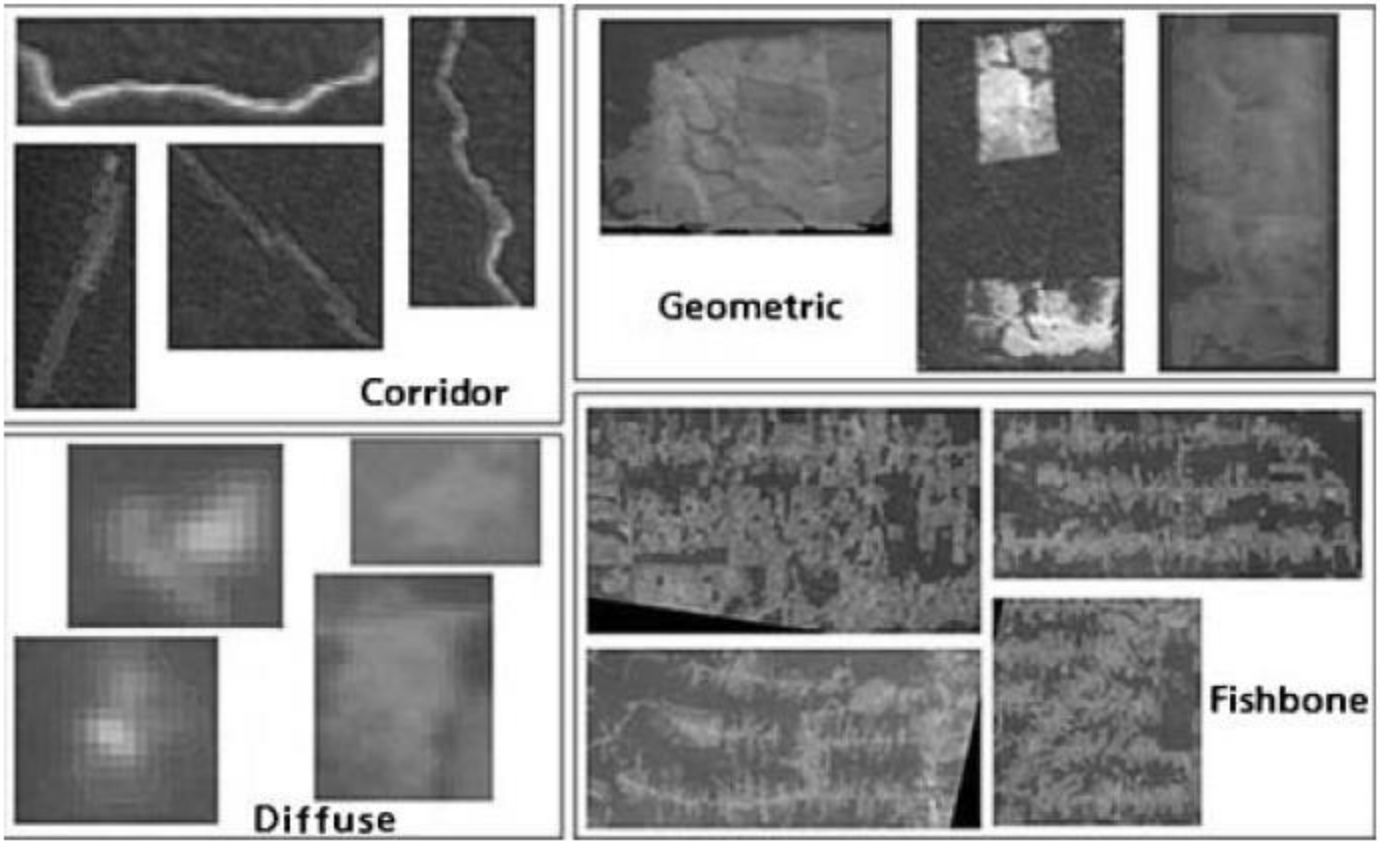


Figure 2. Examples of patterns of tropical deforestation proposed by Mertens and Lambin (1997) in the Brazilian Amazonia: corridor, diffuse, fishbone, and geometric.

Método

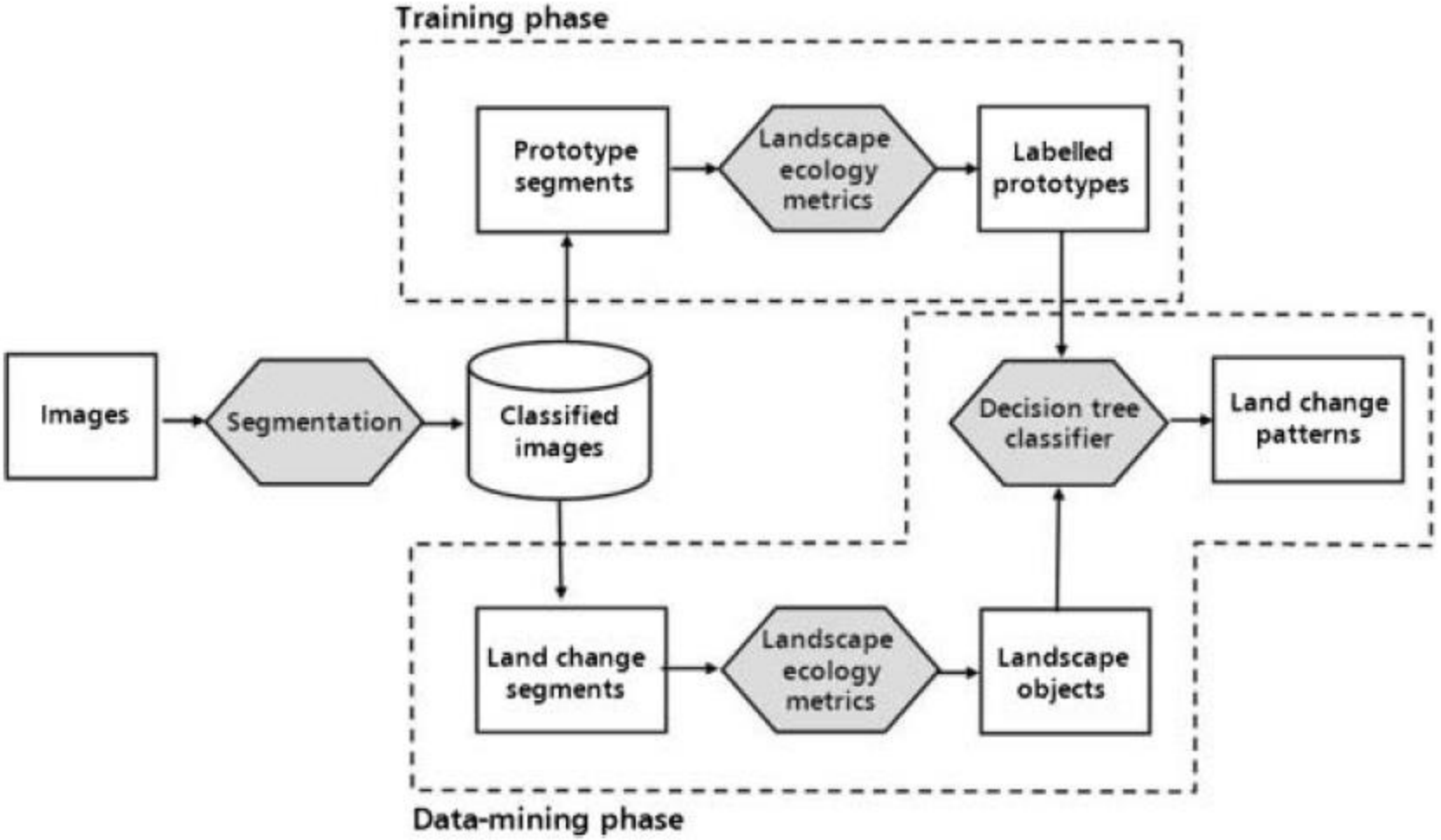


Figure 1. Proposed method for remote sensing image mining.

Métricas de Ecologia da Paisagem

- Perimeter (m):

$$\text{PERIM} = p_{ij}. \tag{1}$$

- Area (ha):

$$\text{AREA} = \left(a_{ij}/10\,000\right). \tag{2}$$

- PARA, perimeter–area ratio, a measure of shape complexity:

$$\text{PARA} = \frac{p_{ij}}{a_{ij}}. \tag{3}$$

- Shape, shape compactness index, calculated by the patch perimeter p_{ij} divided by $p_{ij\ \min}$, which is the minimum perimeter possible for a maximally compact patch of the matching patch area. It is equal to 1 when the region is a square and grows according to the region's irregularity.

$$\text{SHAPE} = \frac{p_{ij}}{p_{ij\ \min}}. \tag{4}$$

Árvore de Decisão

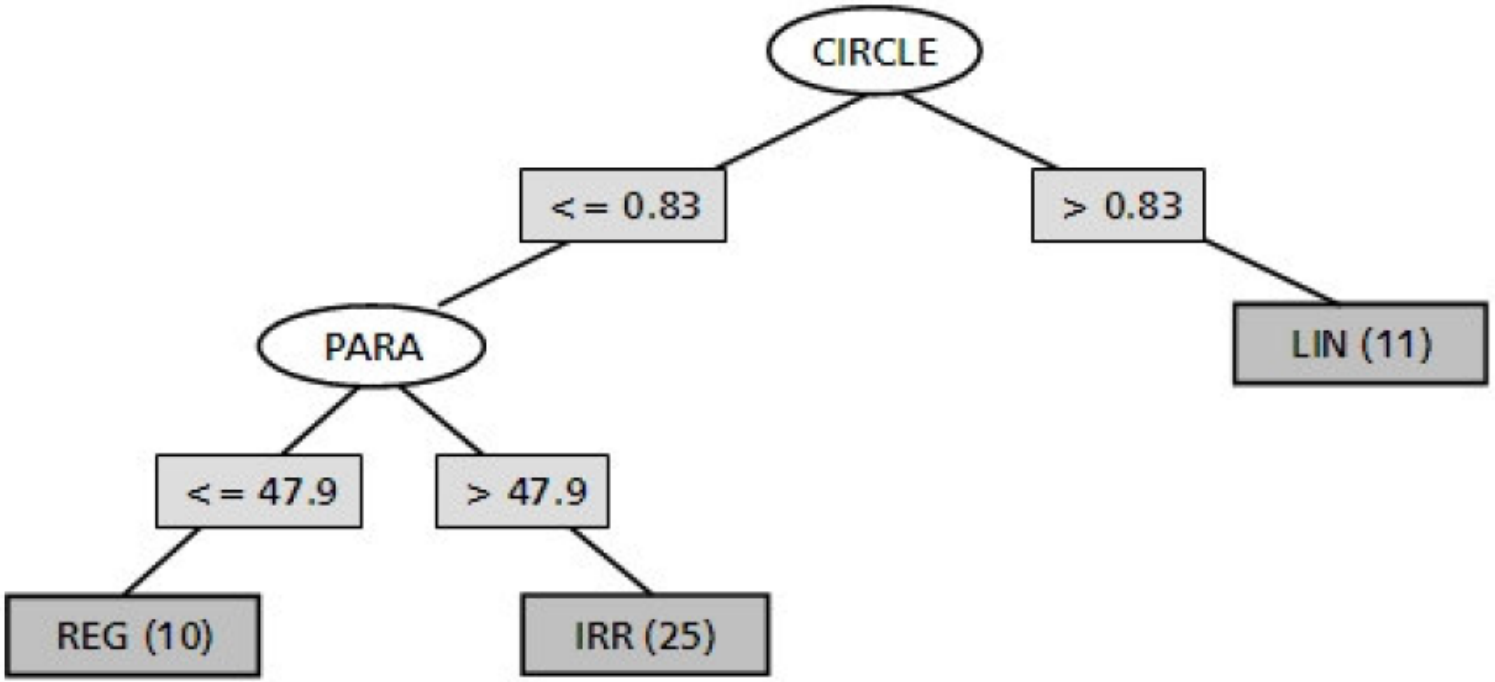
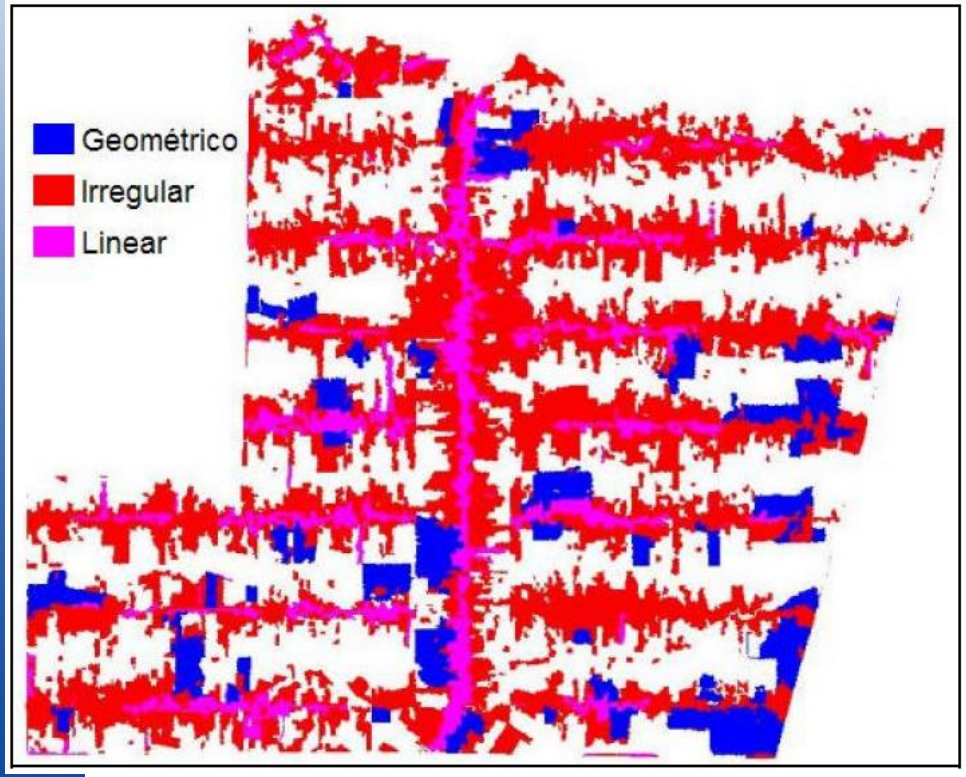
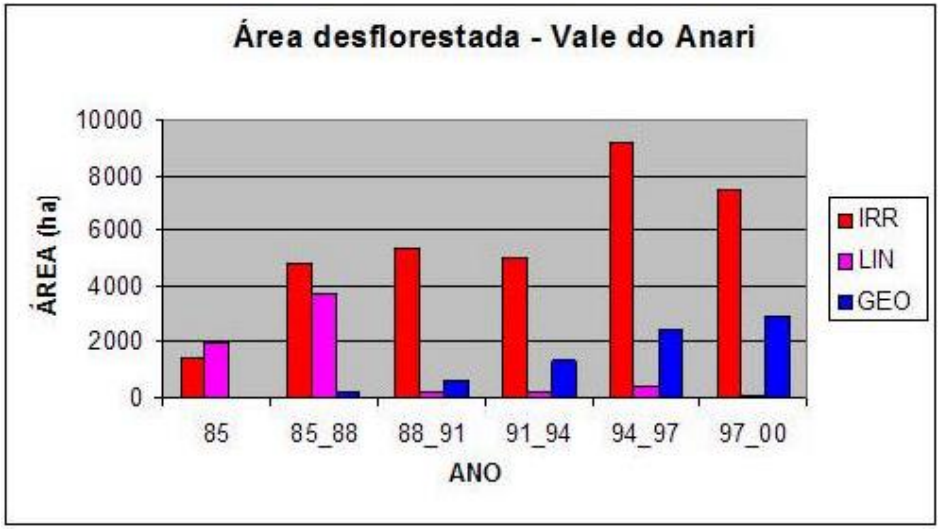


Figure 11. Decision tree for Vale do Anari spatial patterns. The metrics are: perimeter/area ratio (PARA) and circumscribing circle (CIRCLE).

Resultados


Geométrico
Irregular
Linear


Área desflorestada - Vale do Anari
ÁREA (ha)
10000
8000
6000
4000
2000
0
85
85_88
88_91
91_94
94_97
97_00
ANO
IRR
LIN
GEO

Clustering (cluster analysis)

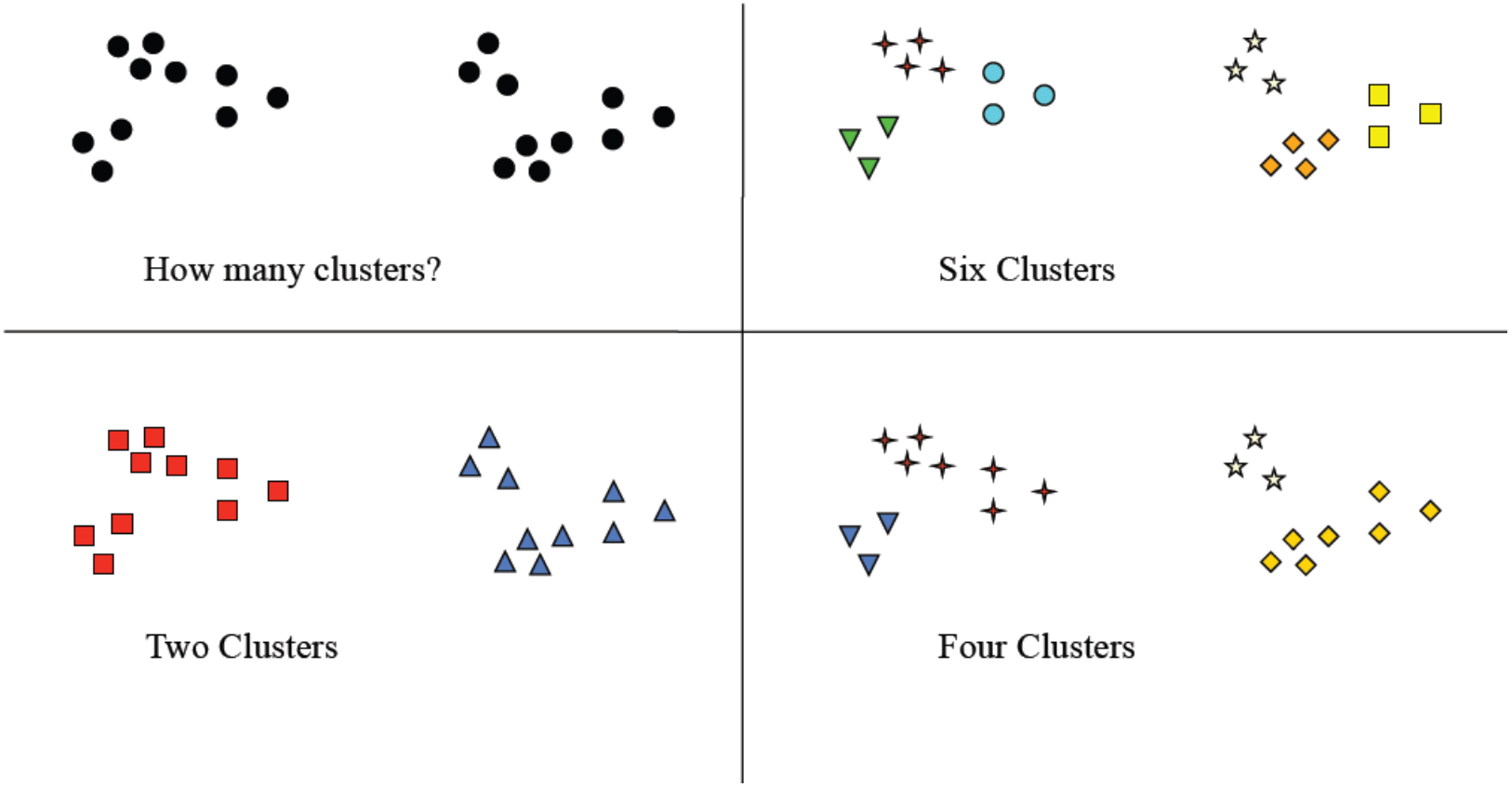
- *Clustering* é um processo de particionamento de um conjunto de dados em um conjunto de grupos chamados *clusters*

- Um cluster é um conjunto de dados (objetos) com características similares
  - que podem ser tratadas coletivamente como um grupo

- *Clustering* é um método não supervisionado
  - não há classes pré-definidas

Clustering Analysis (Kumar 2005)

- Diferentes formas de agrupamento para o mesmo conjunto de pontos



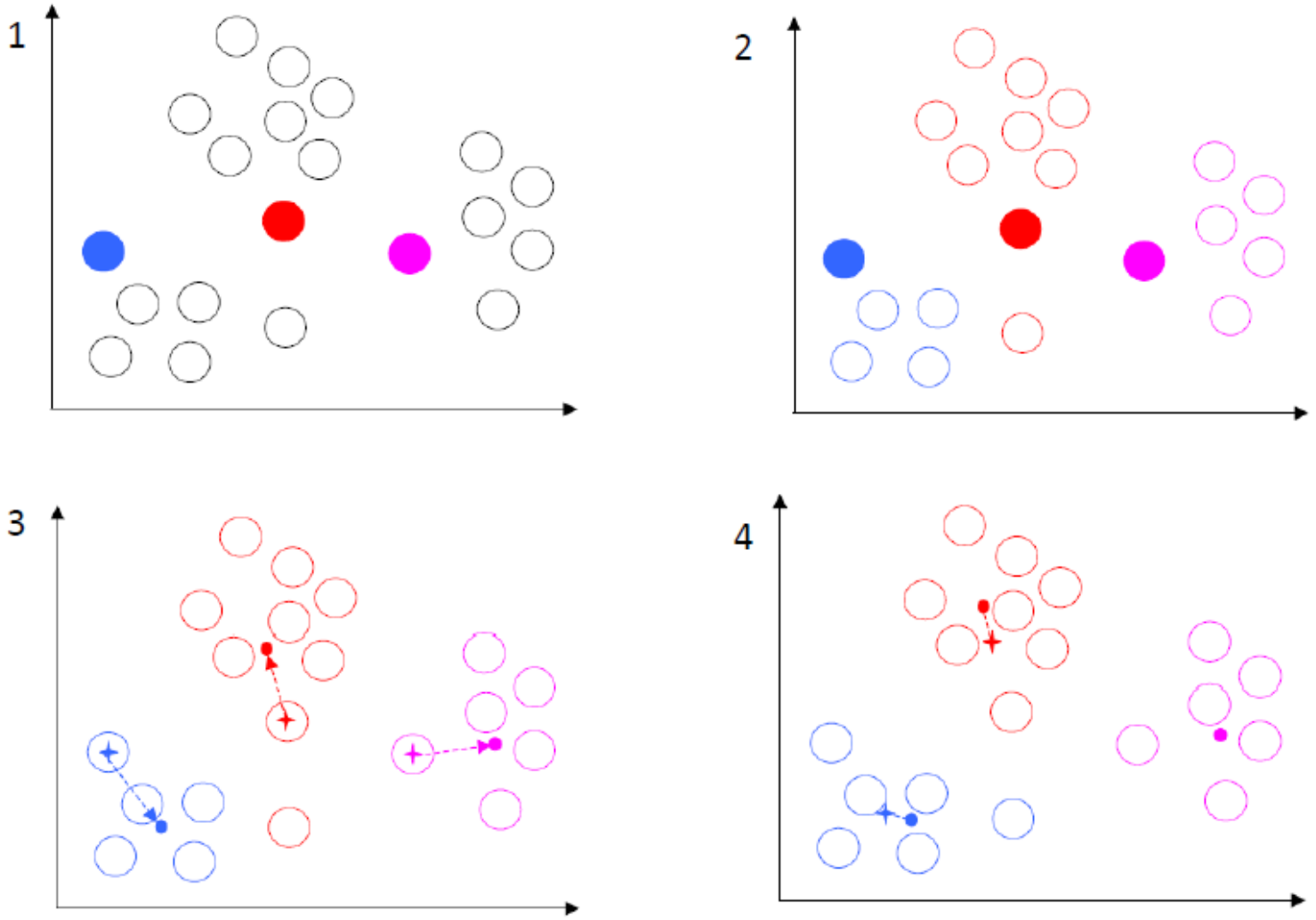
Principais categorias de *clustering*

- Métodos de Particionamento
  - A divisão de dados em subconjuntos (*clusters*)
  - Dados n objetos, o método constrói k partições, onde cada partição representa um grupo, e $k \leq n$.
- Métodos Hierárquicos
  - Um conjunto de clusters aninhados organizado como uma árvore hierárquica
- Métodos Baseados em Densidade
  - Encontra grupos com base na densidade das regiões
- Baseado em Grades
  - Encontrar grupos com base no número de pontos em cada célula

K-Means

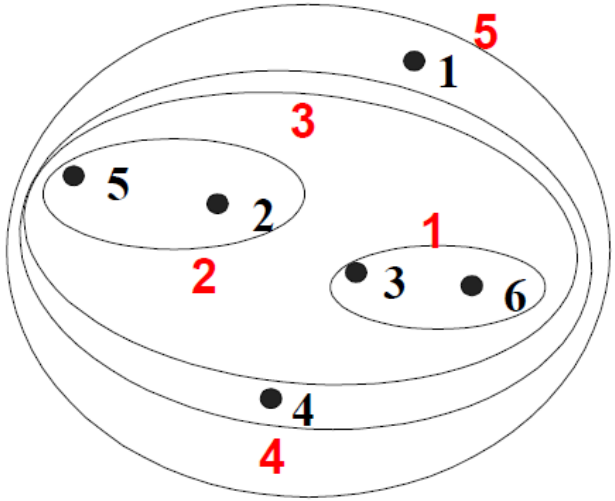
- Métodos de Particionamento
- Cada conjunto está associado com um centroide
- Cada ponto é atribuído ao grupo com o centroide mais próximo
- A desvantagem das k-means é que o número de *clusters* K é um parâmetro de entrada

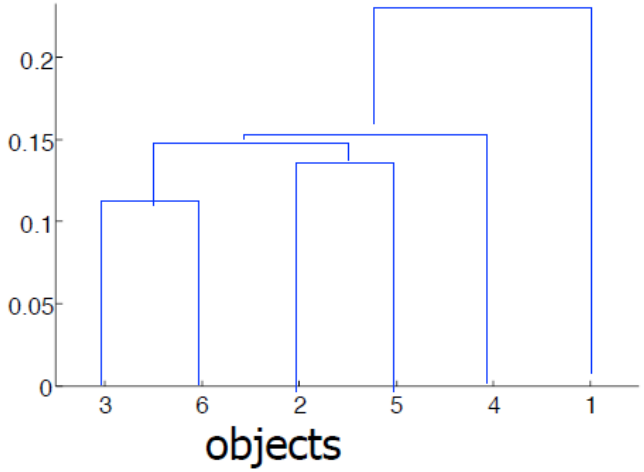
K-Means



Hierarchical Clustering

- Existem dois tipos principais: Divisivo e Aglomerativo
- Aglomerativo
  - Começa com todos os objetos como aglomerados individuais
  - A cada passo, uni os dois *clusters* mais semelhantes
  - Até restar um *cluster* (ou k *clusters*)

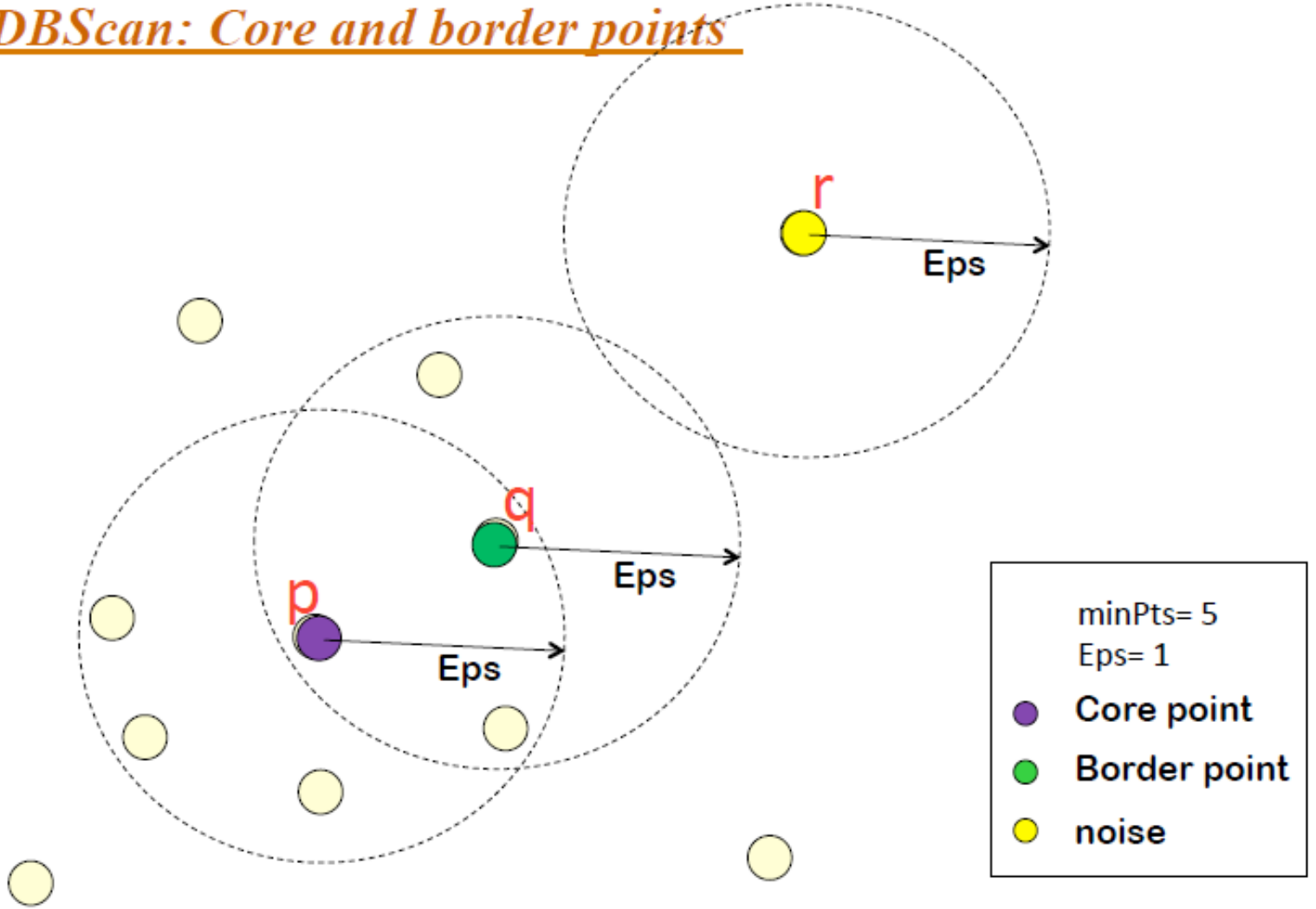




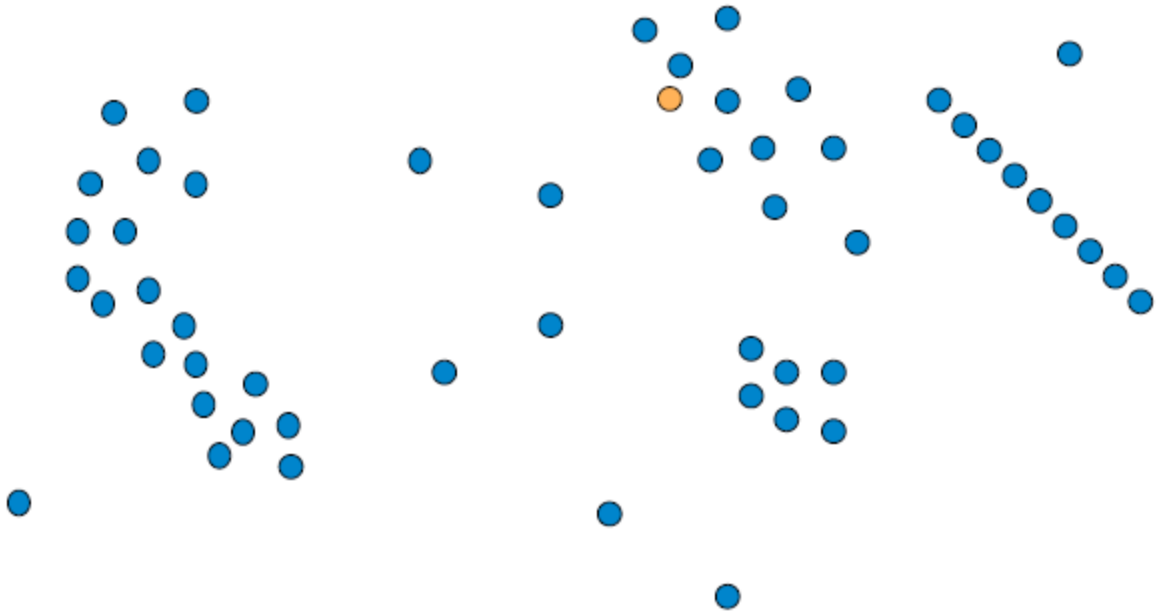
DBSCAN (Ester 1996)

- DBSCAN é um algoritmo baseado em densidade
- Densidade = número de pontos dentro de um raio específico (EPS)
- Um ponto é um ponto central
  - se tem mais do que um determinado número de pontos (MinPts) dentro Eps
- Um ponto de fronteira
  - tem menos que MinPts dentro Eps, mas está na vizinhança de um ponto central
- Um ponto de ruído é qualquer ponto que não é um ponto central ou um ponto de fronteira.

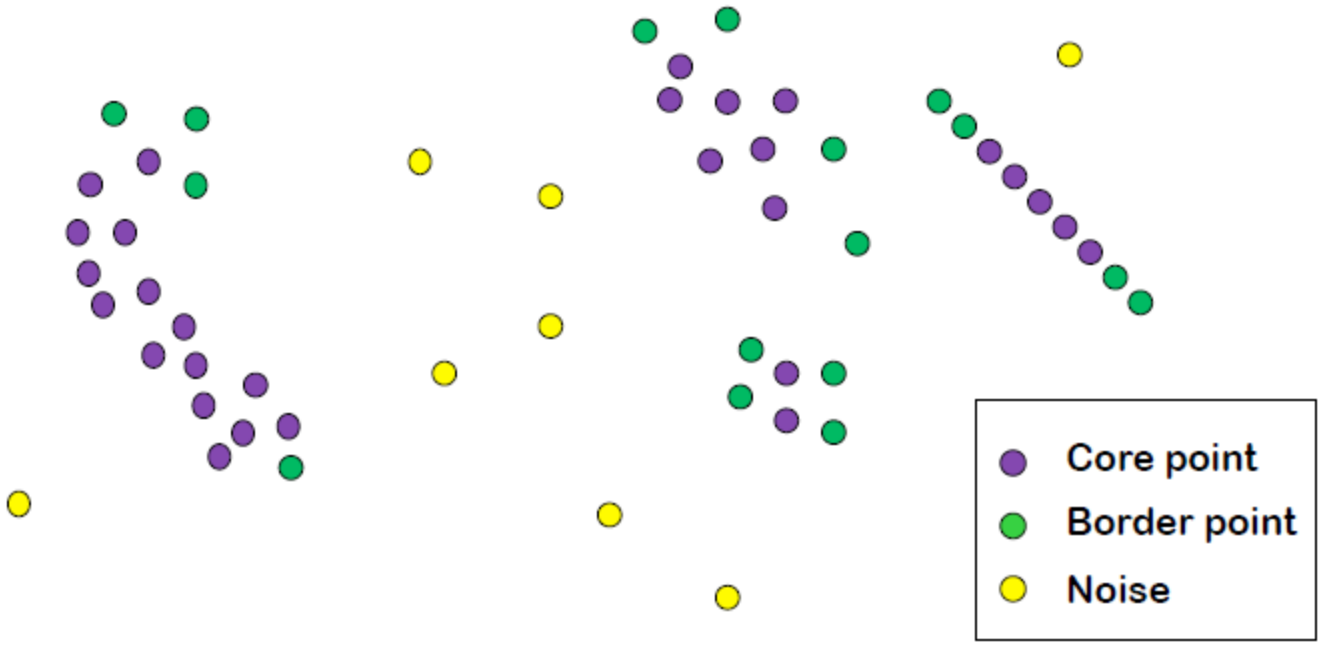
DBScan: Core and border points



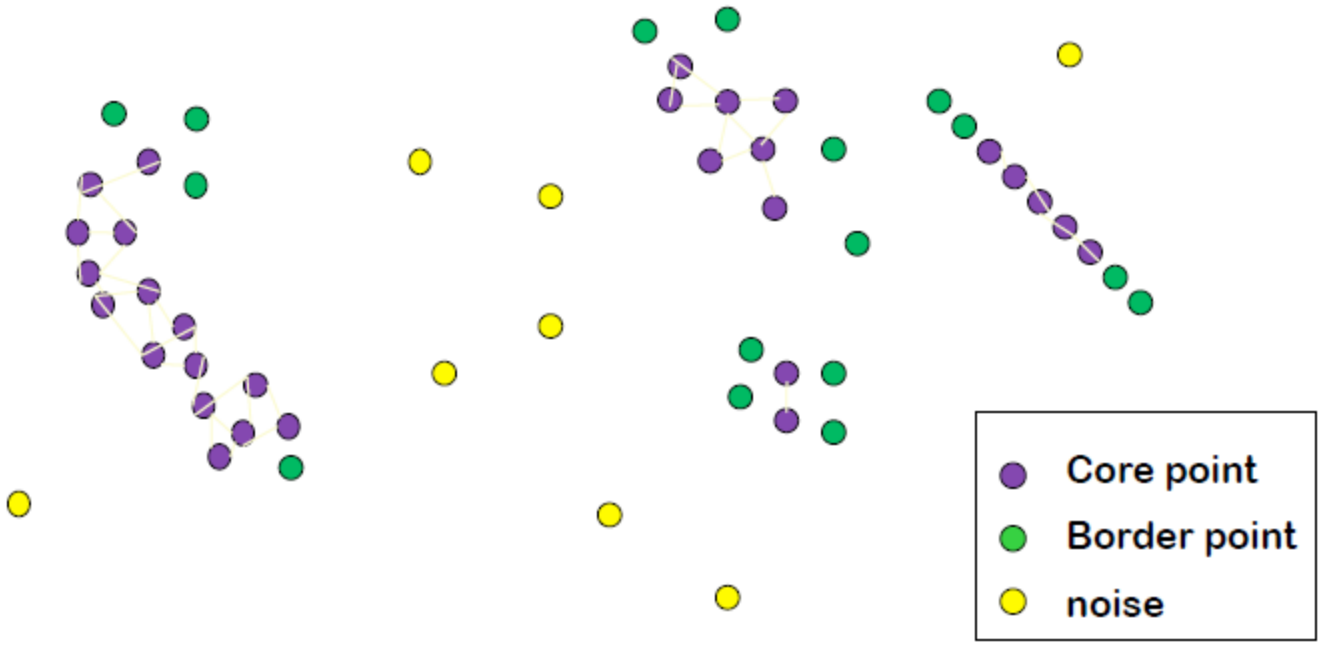
DBSCAN example



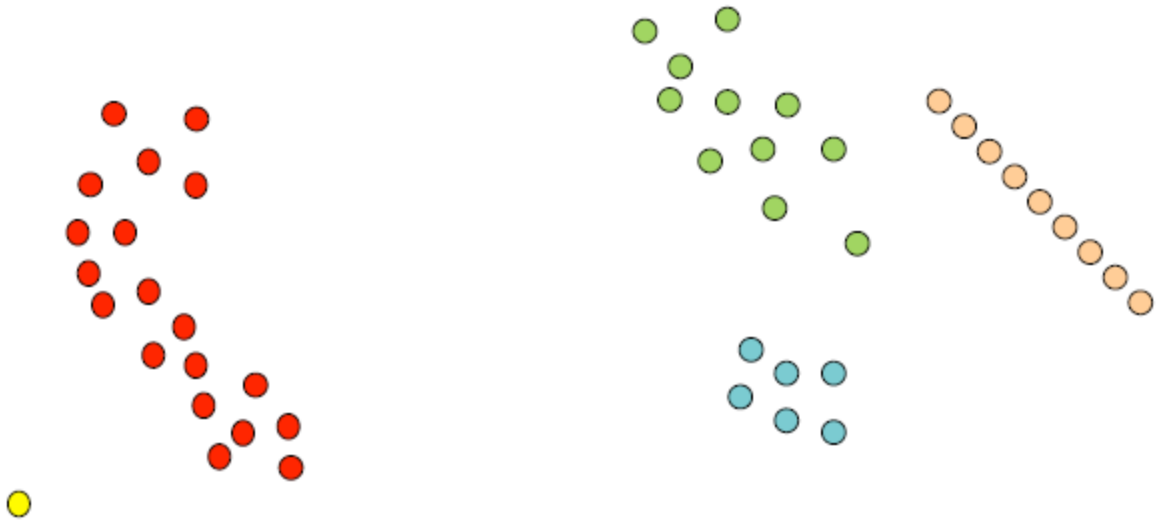
Identifying core, border and noise points



Computing distance



Final Clusters



Detecção de *Outliers*

- O que é um *outlier*?
  - Observações inconsistentes com o resto do conjunto de dados
- O que é um *outlier* espacial?
  - Observações inconsistentes com sua vizinhança
  - Uma instabilidade ou descontinuidade locais

Outliers – Examples *(Shekhar 2003)*

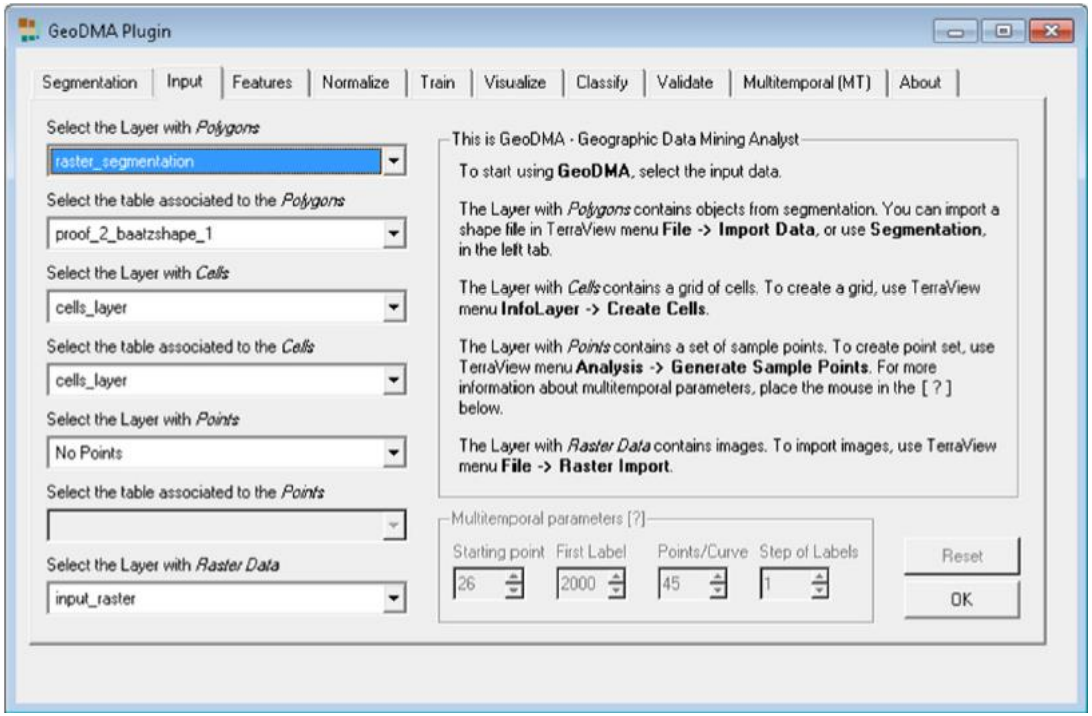
- Map Production
  - Error identification
  - E.g., spatial object violation

Ferramentas

- GeoMiner (Han 1997)
- INGENS (Malerba 2001)
- Weka-GDPM (Bogorny 2006)
  - http://www.inf.ufsc.br/~vania/software.html
- GeoDMA (Korting 2013)
  - http://www.dpi.inpe.br/menu/Projetos/geodma.php
- Spatial Data Mining and Visual Analytics Lab
  - http://www.spatialdatamining.org
- GeoImageRMP - Rapid Miner extension (Guyet 2013)
  - http://geoimagermp.gforge.inria.fr
- Lista com outros *softwares*
  - http://www.spatial.cs.umn.edu/sdm_software.htm
  - http://www.rdatamining.com/resources/tools

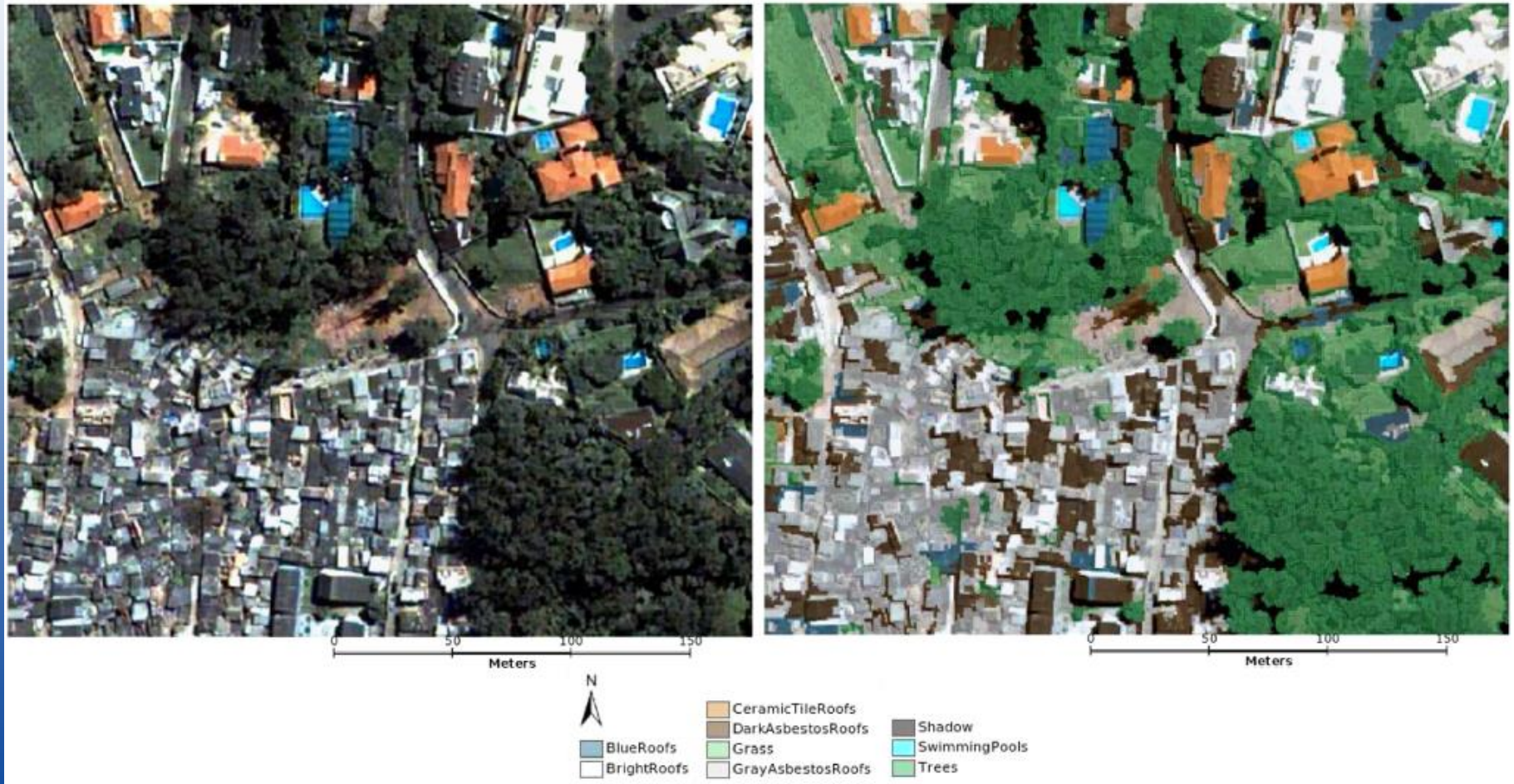
Geographic Data Mining Analyst (GeoDMA)

- Sistema que permite a aplicação de técnicas de mineração de dados espaciais sobre unidades de paisagem definidas por processos de segmentação



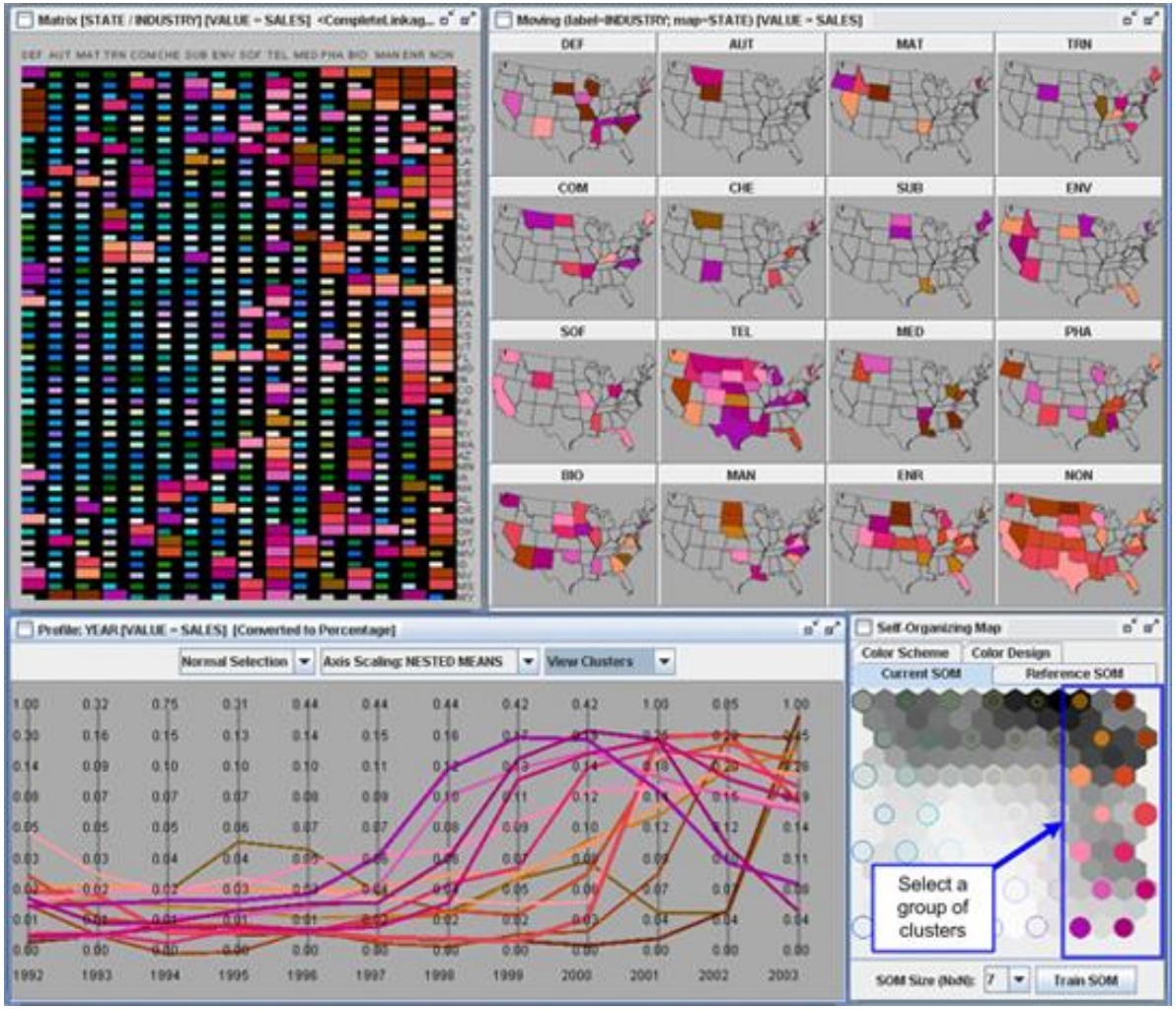
Fonte: Korting et al. (2013)

GeoDMA



Fonte: Korting et al. (2013)

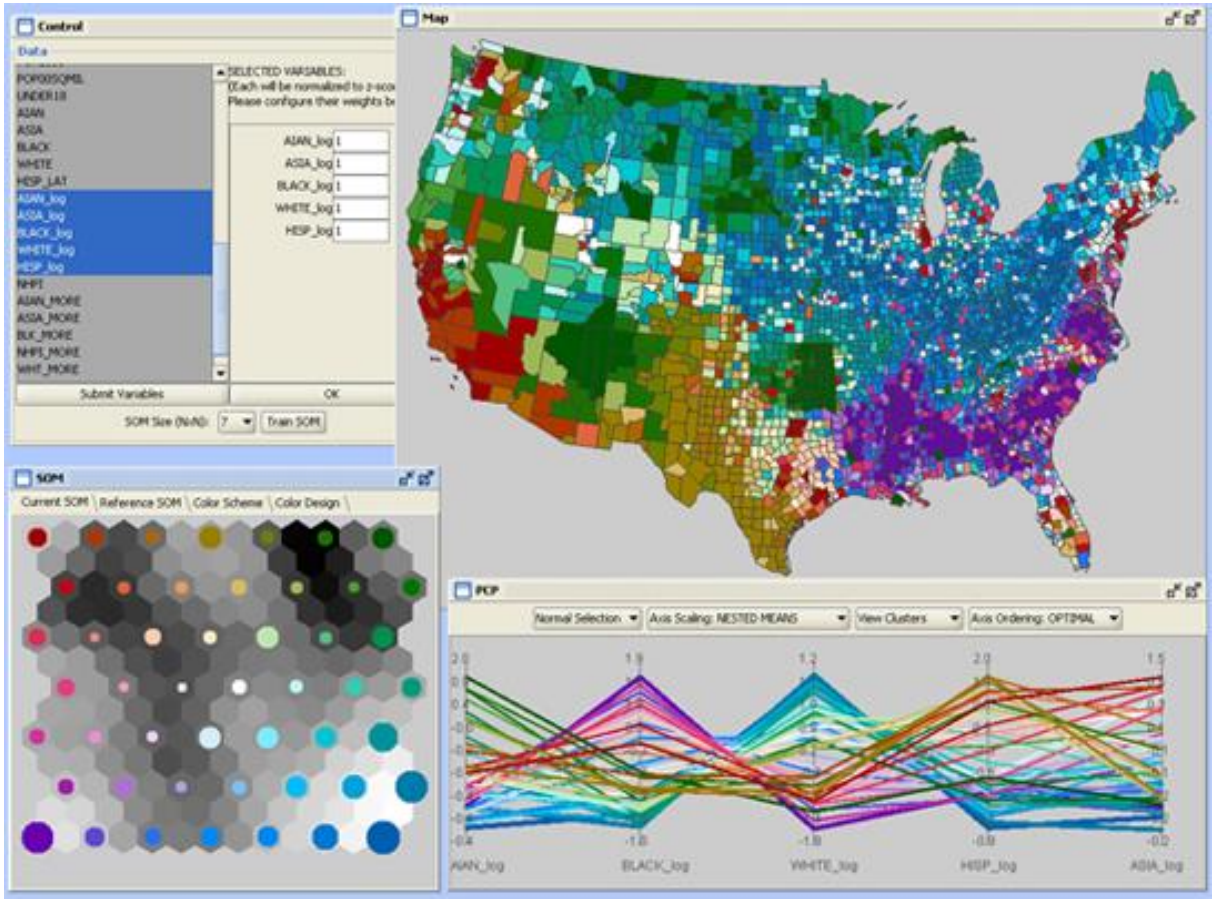
VIS-STAMP: A Visualization System for Space-Time and Multivariate Patterns



Spatial Data Mining and Visual Analytics Lab

Fonte: Guo *et al.* (2006)

SOMVIS: Multivariate Mapping and Visualization



Spatial Data Mining and Visual Analytics Lab

Fonte: Guo *et al.* (2006)

Referências

- Adaptado de material elaborado pelo Prof. Shashi Shekhar, University of Minnesota www.cs.umn.edu/~shekhar
- SHEKHAR, S., CHAWLA, S. **Spatial databases: a tour**. Upper Saddle River, NJ: Prentice Hall, 2003.
- Bogorny, V. and Shekhar S. **Tutorial on Spatial and Spatio-Temporal Data Mining** In: SBBD 2008. (http://www.inf.ufsc.br/~vania/tutorial.zip).
- FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, AAAI Press, v. 17, p. 37–54, 1996. Disponível em: <http://www.aaai.org/aitopics/assets/PDF/AIMag17-03-2-article.pdf>.
- HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd ed.. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.