

Sistemas de Apoio à Tomada de Decisão

Prof^a. Dr^a. Silvia Inês Dallavalle de Pádua

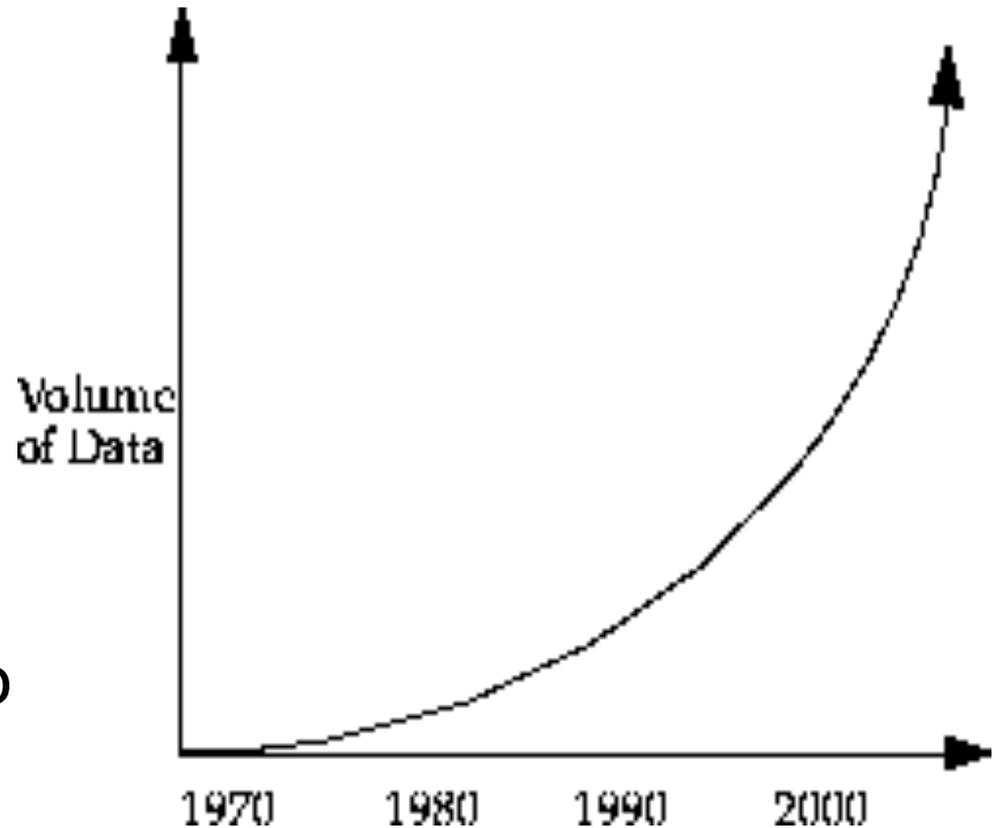
Conteúdo

- O que é Data mining
- Data mining X Data warehouse
- Evolução até o Data mining
- O processo Data mining
- Arquitetura Data mining
- Aplicações do Data mining



Data Mining - Motivação

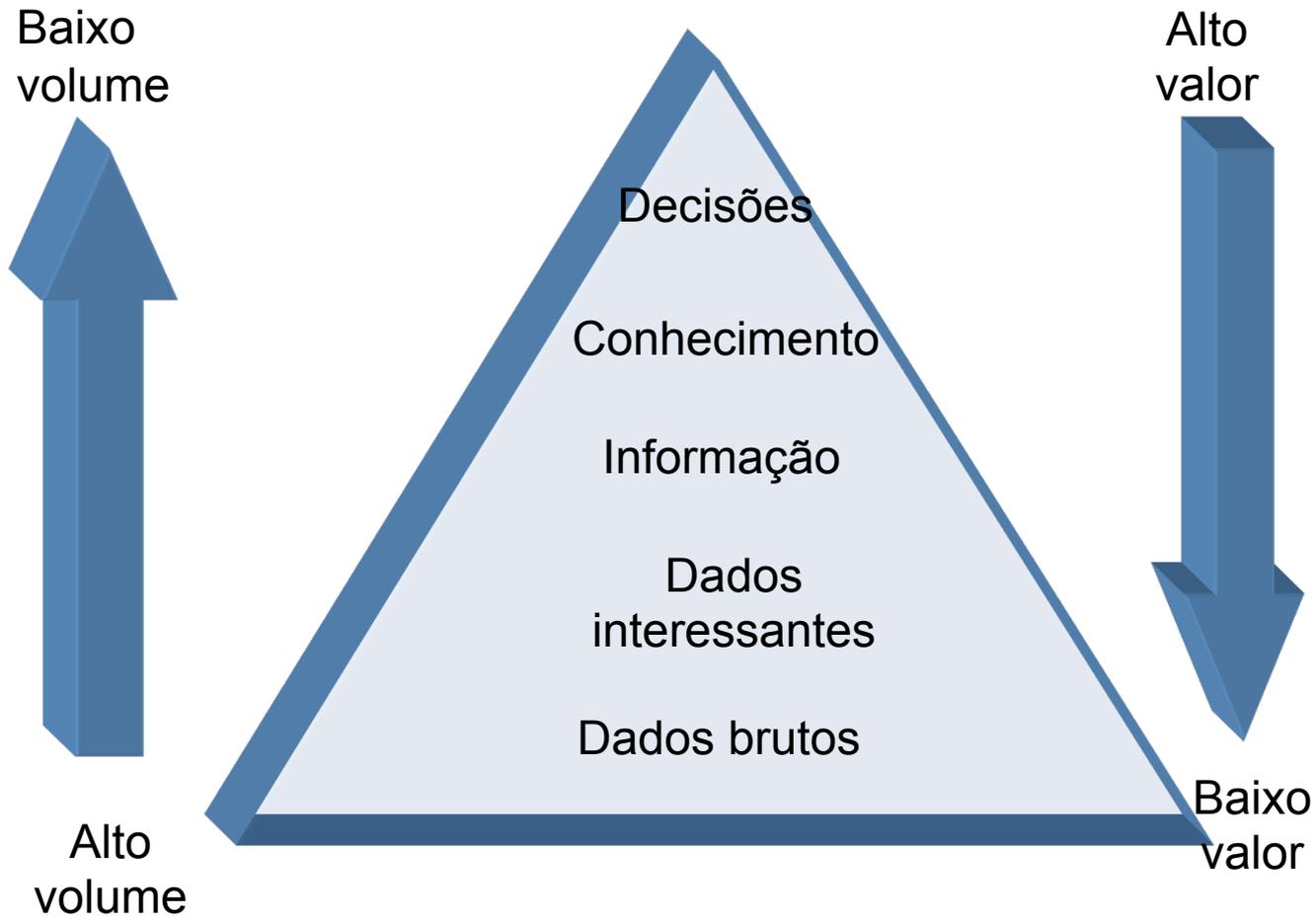
- Grande disponibilidade de dados armazenados eletronicamente
- Existem informações úteis, invisíveis, nesses grandes volumes de dados
- Aproveitar para prever um conhecimento futuro (ir além do armazenamento **explícito de dados**).



[01] Queens University - Belfast

Capacidade de Processamento & Capacidade de Armazenamento

*A combinação dessas capacidades produz um **gap crescente** entre nossa habilidade de **gerar dados** e nossa habilidade de **fazer uso dele**.*



O que é Data mining?

- “Data mining (mineração de dados), é o processo de **extração de conhecimento de grandes bases de dados, convencionais ou não**.
- Utiliza técnicas de **inteligência artificial** que procuram relações de **similaridade** ou **discordância** entre dados.
- Seu objetivo é encontrar, automaticamente, **padrões, anomalias e regras** com o propósito de transformar dados, aparentemente **ocultos**, em **informações úteis** para a tomada de decisão e/ou avaliação de resultados.

Data Mining

- Usado para prever **tendências** e **comportamentos**.
- Identifica **padrões** previamente desconhecidos.
- A análise tende ser da **base para o topo**, e as melhores técnicas têm seu desenvolvimento orientado para grandes volumes de dados



*A base de dados da companhia é **um ativo tão valioso quanto uma marca registrada, uma patente ou como os bens de capital de uma indústria.***

Data mining e OLAP

- Data mining
 - é mais **orientado por descoberta**.
 - Fornece **percepções dos dados corporativos** que não podem ser obtidas com o OLAP.
 - **Descobre padrões e relacionamentos ocultos** em grandes bancos de dados e inferindo regras a partir deles para prever comportamentos futuros.
 - Esses modelos e regras podem **então ser utilizados para guiar o processo de decisão** e prever o efeito dessas decisões.

Data mining & Data warehouse

- Data mining X Data warehouse:
 - Data mining ⇒ **extração** inteligente de dados;
 - Data warehouse ⇒ **repositório** centralizado de dados;
 - Data mining **não é** uma evolução do Data warehouse;
 - Data mining **não depende** do Data warehouse, mas obtém-se **melhores resultados** quando aplicados em conjunto;
 - Cada empresa deve saber escolher o que **é importante** para o seu negócio;
 - Data warehouse aliado a ferramentas estatísticas desempenham papel semelhante ao data mining, **mas não descobrem novos padrões de comportamento** (a não ser empiricamente).

Data mining & Data warehouse

Data warehouse: a **memória** da empresa

Data mining: a **inteligência** da empresa

KDD & DM

- KDD – Knowledge Discovery (in databases):
Descoberta de Conhecimento
- DM – Data Mining: Mineração de Dados
- Um campo multidisciplinar:
 - Banco de dados e data warehouse
 - Métodos de modelagem e visualização de dados
 - Aprendizado de máquina
 - **Estatística**
 - Sistemas Especialista e Aquisição de conhecimento

De **forma simplista**, pode-se dizer que um processo de data mining efetivamente automatiza o processo estatísticos, aliviando desta forma o usuário final de um grande trabalho e desenvolvimento (PINHEIRO, 2008).

PINHEIRO, C.A. R Inteligência Analítica. Ed. Moderna. São Paulo, 2008.

KDD & DM

- KDD é o processo de selecionar e processar os dados que permitam identificar estruturas interessantes:
 - Pré-processamento
 - Preparação dos dados
 - Redução dos dados.
 - **Mineração de dados**
 - Pós-processamento ou análise da solução
- **Mineração de Dados é uma etapa no processo de KDD**
 - Descoberta automática de padrões
 - Desenvolvimento de modelos preditivos e explicativos

Exemplos

- O governo dos USA se utiliza do data mining para identificar padrões de transferência de fundos internacionais que se parecem com lavagem de dinheiro do narcotráfico.
- Também outros crimes, como o atentado na cidade de Oklahoma, foram analisados através do data mining varrendo registros de compra de explosivos e outros acessórios adquiridos para a fabricação de bombas.

Exemplos

- A MasterCard **processa diariamente cerca de 12 milhões de transações e utiliza o KDD** para extrair estatísticas sobre os portadores de cartões.
- Isto permite a visualização de diferentes **classes de portadores de cartões** e a análise de como estes utilizam seus cartões para desenvolver promoções especializadas e deteções de fraude.

Exemplos

- Na medicina já é possível a criação e manutenção de grandes bancos-de-dados com **informação sobre sintomas, resultados de exames, diagnósticos, tratamento e curso das doenças** para cada paciente.
- A mineração destes dados pode fornecer conhecimento novo como, por exemplo, **a relação entre algumas doenças e certos perfis** profissionais, sócios culturais, hábitos pessoais e local de moradia

Data mining na empresa

- Minerar dados não se resume em analisar grandes massas de informação e descobrir novas relações.
- O processo de **data mining precisa ser incorporado à empresa e passar a ser uma etapa** natural dos outros processos empresariais.

Descoberta: Não-Supervisionada e Supervisionada

- **Descoberta não-supervisionada**
 - Permite que os algoritmos de mineração de dados encontrem “**padrões**” nos dados independente de um alvo (objetivo) pré-estabelecido.
- **Descoberta supervisionada**
 - Requer a análise dos dados para identificar um campo alvo
 - **Hipóteses**

KDD (Pinheiro, 2008)

- **Knowledge Discovery in Databases – KDD** pode ser descrito como sendo composto por seis etapas distintas, descritas a seguir:
 1. Entendimento do problema
 2. Extração dos dados
 3. Limpeza dos dados
 4. Redução dos dados
 5. Escolha do método de mineração de dados
 6. Interpretação dos resultados

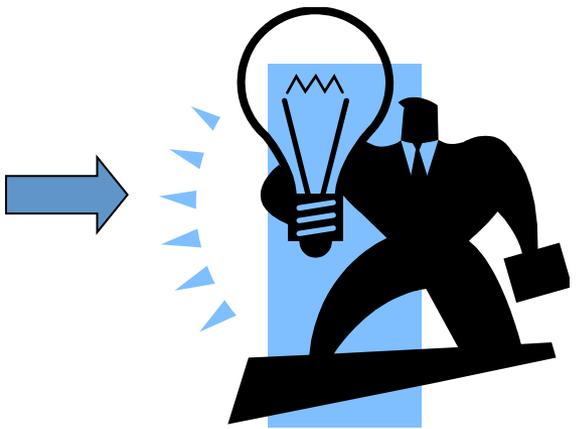
Etapas KDD



Entendimento do problema



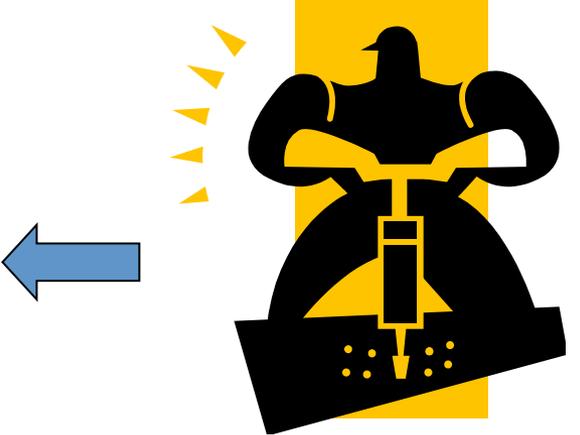
Extração dos dados
Limpeza dos dados



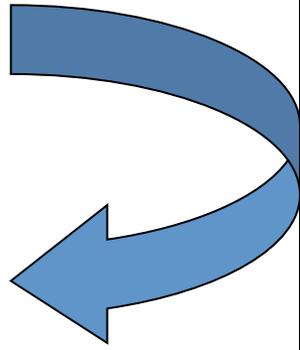
Redução dos dados



Interpretação dos resultados



Escolha do método de mineração



Etapa – Entendimento do problema

- **Entendimento do problema:** é quando se busca a compreensão dos objetivos do projeto e suas necessidades, bem como, o conhecimento desejado pelo usuário final.
- Essa etapa é fundamental para **se ter um bom conhecimento do negócio**, necessário para uma clara definição do problema e um plano preliminar das ações a serem executadas.

- **Entendimento do problema:** “O departamento financeiro da rede notou pelos seus bancos-de-dados que o faturamento **sofreu uma queda significativa nos últimos meses**. Isso ocorreu pela queda das vendas em suas lojas que estão espalhadas por todo país. Uma **análise humana simples** constatou que a queda das vendas se deu em quase todas as regiões do país com predominância **nas regiões sudeste, nordeste e norte**”.

Várias soluções poderiam ser propostas pelos departamentos:

Nova campanha publicitária, aumento de oferta de cartão de crédito, mudança no formato da rede, mala-direta

“Outra análise mais profunda buscou encontrar pontos comuns entre os dados das vendas de cada uma das regiões, concluindo que **a queda do volume de vendas** das regiões norte e nordeste se deu a partir, aproximadamente, **do mês de julho**, recuperando-se um pouco a partir de outubro mais ainda assim permanecendo, daí para frente **em níveis mais baixos**. Enquanto que a queda das vendas nas regiões sudeste e sul não foi tão acentuada, como já mencionado, porém progressiva ao longo de todo o ano. Outra nova relação descoberta foi **a diminuição das vendas nos cartões de crédito da financiadora da empresa**, em todas as regiões do país. Uma análise mais aprofundada por data mining, desta nova relação permitiu visualizar um elemento comum entre os clientes que deixaram de comprar no cartão-de-crédito, a saber, **um alto percentual de reclamações destes clientes no serviço de atendimento ao cliente.**”

Etapa – Extração dos Dados

Extração dos dados: é responsável pela criação de um conjunto de dados alvo. Nessa fase, seleciona-se um conjunto de dados ou de instâncias em que a descoberta deverá ser efetuada.

Etapa – Limpeza dos dados

- **Limpeza dos dados:** pré-processamento dos dados.
 - É de grande importância, pois é executado o tratamento dos dados ausentes, inconsistentes, ou fora dos padrões normais
 - Assegurar a qualidade dos dados **pode consumir entre 20-40%** de todo processo de KDD

Etapa – Redução dos Dados

- **Redução dos dados:** uma vez conhecida a fonte de informações e **eliminadas as inconsistências**, será necessário selecionar os **dados que podem influenciar** nos resultados do modelo a ser construído.
- Informações dúbias devem ser eliminadas, assim como, as variáveis que não se mostram sensíveis ao modelo.

Etapa – Escolha do método de mineração de dados

- **Escolha do método de mineração de dados:** envolve as fases de levantamento dos objetivos do processo e identificação da melhor técnica a ser aplicada e a abordagem para a aplicação do modelo. Podem-se distinguir os objetivos da descoberta de conhecimento de dois tipos básicos:
 - Supervisionada
 - Não-supervisionada

Descoberta - Supervisionada

Trabalha com **hipóteses formuladas** pelos usuários.

- Para estas hipóteses **são realizados testes de validação** contra o banco de dados.
- A ênfase desse modelo é que o usuário obtém uma **resposta afirmativa ou negativa** às suas hipóteses.
- O problema é que **nenhuma informação nova é criada** ao longo deste processo.
- Este processo de procura é iterativo, onde a saída gerada é revisada e **um novo conjunto de hipóteses é formulado** para refinar a procura, sendo que o processo deve ser repetido várias vezes.

Descoberta – não supervisionada

Difere da verificação por ter uma ênfase na qual o sistema procura descobrir **de forma autônoma**, os novos padrões nas bases de dados.

Os dados são analisados na procura de ocorrências freqüentes de determinados padrões e tendências, **generalizações são feitas** sobre os dados **sem a intervenção do usuário**.

Etapa: Interpretação dos resultados

- **Apresentação das descobertas obtidas, determinação da melhor forma de utilizar as informações na tomada de decisão.**
- **Definição das vantagens e desvantagens dos modelos, reavaliações do processo como um todo e, sendo necessária, a criação de novos modelos** ou mesmo, processos adicionais de mineração.

Continuação do caso...

A ocorrência de um inverno muito brando no ano analisado, não tendo a empresa oferecido opções de vestuário mais leve em sua rede, o que seria imperioso nas regiões mais quentes, norte e nordeste. Isto levou os clientes tradicionais **compradores da rede para outras lojas** que ofereciam o vestuário mais apropriado. Certamente, estas outras lojas concorrentes souberam manter a preferência da nova clientela de alguma forma ignorada e que resultou no retorno, terminada a estação fria, de apenas uma parcela dos clientes que a rede na qual estavam acostumados a comprar.

A nova relação descoberta de que as vendas nos cartões-de-crédito da rede caíram progressivamente ao longo do ano representavam o segundo fator que se sobrepõe ao primeiro, piorando ainda mais as vendas nas regiões norte e nordeste e diminuindo, menos acentuadamente, as vendas nas regiões sul e sudeste.

A análise por data mining forneceu a informação adicional de que o serviço de cartões-de-crédito vem funcionando mal, cometendo erros de cobrança e tendo como reclamações principais o não lançamento de pagamento já realizados em outra filial diferente da qual a compra se realizou, além da não emissão dos extratos em tempo hábil, fazendo com que o cliente tenha que correr para pagar a parcela ou mesmo acabe pagando-a com multa. Isto explicaria a queda progressiva ao longo do ano das vendas em todas as lojas independentemente de região.

Neste caso, poderia ser o oferecimento de vestuário mais leve em conjunto com vestuário apropriado ao inverno em todas as lojas da rede nos meses de julho a setembro e **não a fixação de todo o vestuário à estação do ano, como fora realizado**. Isto implica no **planejamento de encomendas antecipadas** deste vestuário que ficaria estocado aguardando a estação mais fria.

O quanto e o que ser encomendado seria objeto de **novas análises de data mining** para estabelecer preferências da clientela.

Em relação à má administração do cartão-de-crédito, possíveis soluções seriam a mudança do sistema computacional ou da empresa prestadora do serviço, de forma a ganhar novamente a confiança da clientela neste importante meio de compra.

A etapa Interpretação dos resultados busca verificar se as causas do problema foram sanadas ou o objetivo da empresa alcançado.

No exemplo apresentado, sem análise dos resultados, a empresa teria investido nas soluções genéricas apresentadas por cada um dos setores da empresa. Este investimento poderia trazer retorno porém a real causa da evasão da clientela não teria sido resolvida e o risco de novas evasões estaria sempre presente.

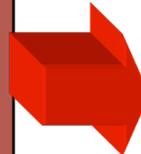
Dentre as aplicações mais usuais de mineração de dados podem-se citar primariamente dois grandes focos:

Identificação de segmentos com características semelhantes, agrupando os casos ou as ocorrências, de acordo com as informações descritivas dos mesmos.



Modelagem não supervisionada, onde não se sabe previamente as respostas esperadas, ou seja, não se pode ensinar, ou treinar o modelo baseado em determinada premissa.

Predição de eventos ou acontecimentos com prévia ciência dos resultados.



Modelagem supervisionada, quando se conhece a resposta de um determinado cenário e, com base nessa premissa, treina-se o modelo para uma futura interpolação da variável resposta ou alvo.

Um ambiente de **descoberta de conhecimento** envolve bem mais do que apenas atividades de mineração de dados.

Podem ser incluídos neste cenário:

- **processos de extração de informações** dos sistemas operacionais, **criação de base de dados analíticas**, aglutinando informações unificadas sobre clientes, produtos, transações etc,
- **geração de bases gerenciais e analíticas** de acordo com os objetivos de negócio da empresa e os contextos empresariais envolvidos.

- As etapas específicas de construção dos modelos de mineração de dados, existem diversas técnicas possíveis, por exemplo:
 - Redes Neurais Artificiais
 - Árvore de decisão
 - Algoritmos genéticos
 - Modelos de regressão
 - Regras de associação

Técnicas

- **Redes Neurais Artificiais:** possuem **capacidade intrínseca de aprendizado** a partir de **dados de entrada** e permitem a possibilidade de generalização posterior.
- **Algoritmos genéticos:** são métodos de otimização combinatória baseado em **processos de evolução biológica**. A idéia básica é que, ao longo do tempo, a evolução tenha sido selecionada pela espécie mais adequada.

Técnicas

- **Modelos de regressão:** estima uma variável dependente utilizando um conjunto de variáveis independentes e um conjunto de constantes.
- **Regras de associação:** procuram por associação de itens em um banco de dados transacional que possua alguma dependência estatística

Tarefas de mineração de dados

- Tarefas podem ser entendidas como tipos de relacionamentos entre dados que vamos estabelecer para obtermos conhecimento, passível de interpretação humana.
- Tipos de tarefas:
 - Associações
 - Seqüências
 - Classificações
 - Aglomerações
 - Prognósticos.

Técnicas de Mineração de dados

As **técnicas de mineração de dados** podem ser consideradas como os **fundamentos computacionais**, ligados intimamente a área de inteligência artificial, **que propiciam a construção de algoritmos** que possibilitem a busca por **padrões escondidos nos dados**.

Associações

São ocorrências ligadas a **um único evento**.

Procura registros que tenham similaridades associativas

Por exemplo: um estudo de modelos de compra em supermercados pode revelar que quando se compram salgadinhos de milho, compra-se também um refrigerante tipo cola em 65% das vezes; mas quando há uma promoção, o refrigerante é comprado em 85% das vezes.

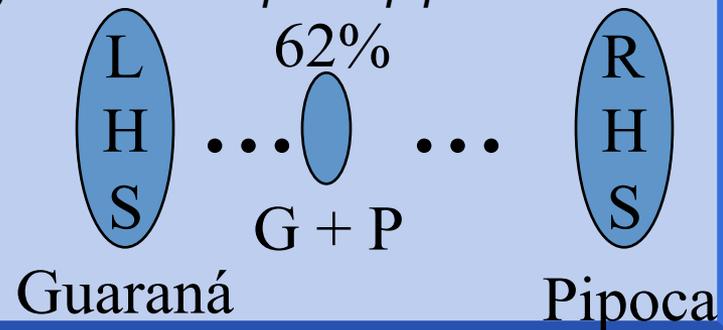
Podem ser expressados por regras

Ex: 62% dos compradores de guaraná compram pipoca

62% fator de confiança

LHS (left hand side)

RHS (right hand side)



Seqüência

Na **seqüência os evento estão ligados ao longo do tempo**. Pode-se descobrir, por exemplo, que quando se compra **uma casa**, em 65% das vezes se compra também **uma nova geladeira** no período de duas semanas; e que, em 45% das vezes, **um fogão** também é comprado um mês após a compra da residência.

Classificação

A classificação reconhece modelos que descrevem o grupo ao qual o item pertence, por meio do exame dos itens já classificados e pela inferência de um conjunto de regras.

Por exemplo, empresas como operadoras de cartões de crédito e companhias telefônicas preocupam-se com a perda de clientes regulares.

A **classificação** pode ajudar a **descobrir as características de clientes** que provavelmente virão a abandoná-las e oferecer um modelo para ajudar os gerentes a prever quem são eles, de modo que se possam elaborar **antecipadamente** campanhas especiais para reter esses clientes.

Classificação

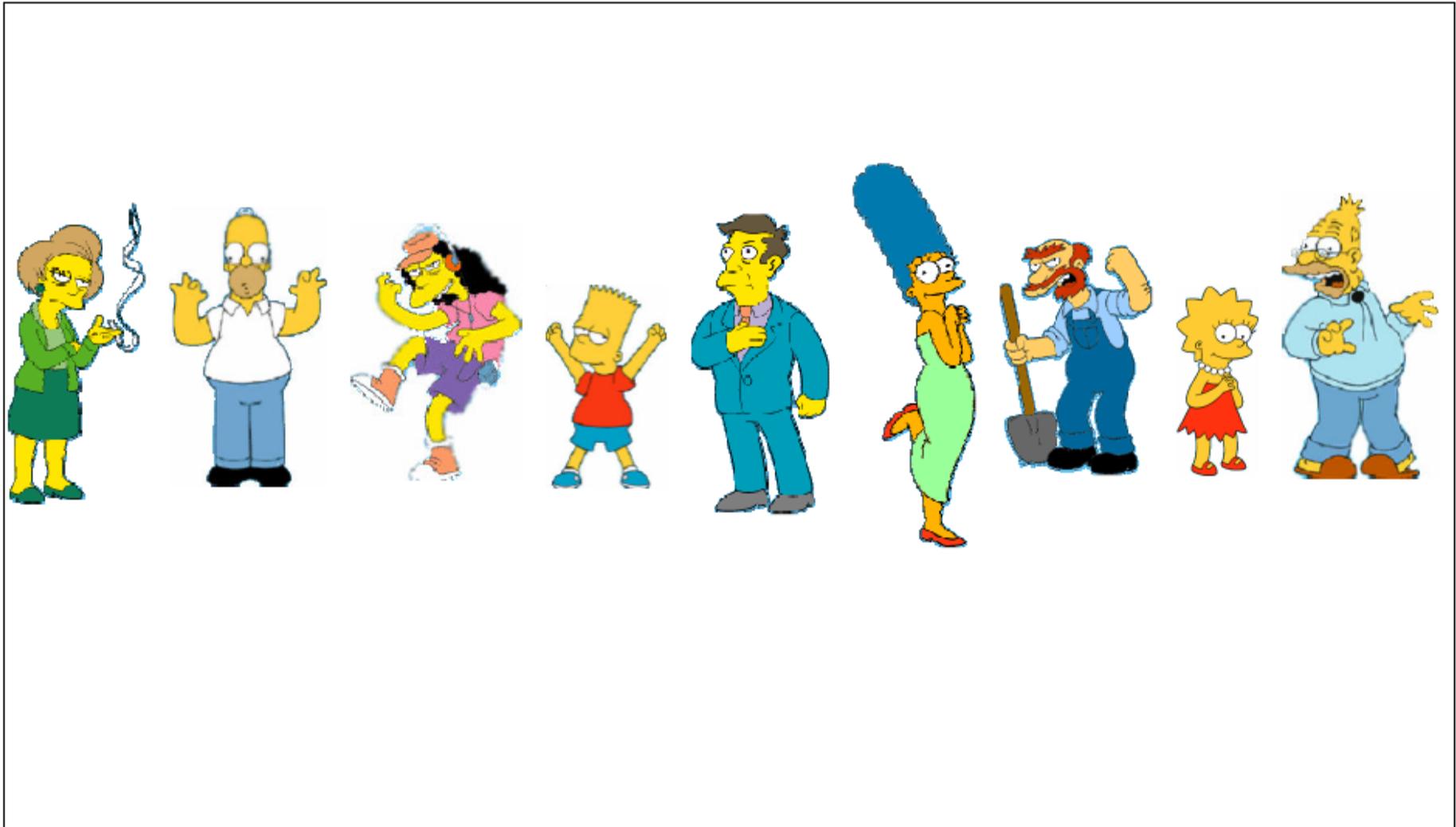


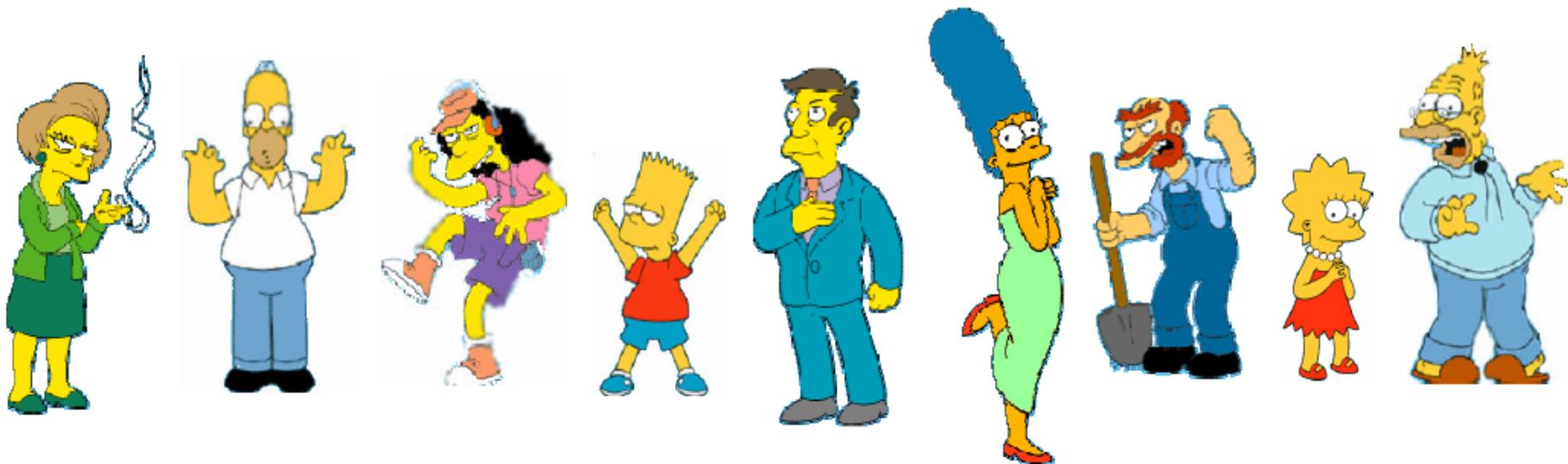
Aglomeraco (Clustering)

A **aglomeraco** (*clustering*) funciona de maneira semelhante à classificaco **quando ainda no foram definidos grupos**.

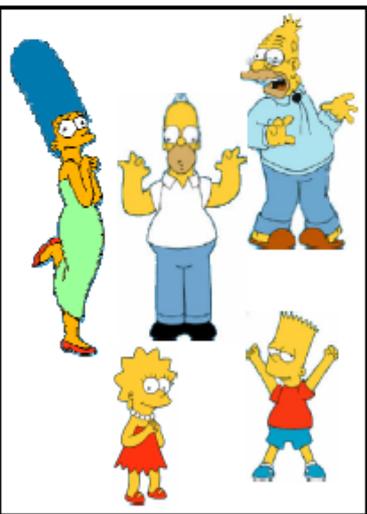
Uma ferramenta de data mining descobrir diferentes agrupamentos dentro da massa de dados, por exemplo ao **encontrar grupos de afinidade para cartes bancrios** ou ao dividir o banco de dados em grupos de clientes com base na demografia e em investimento pessoais.

Qual é o agrupamento natural entre estes objetos?





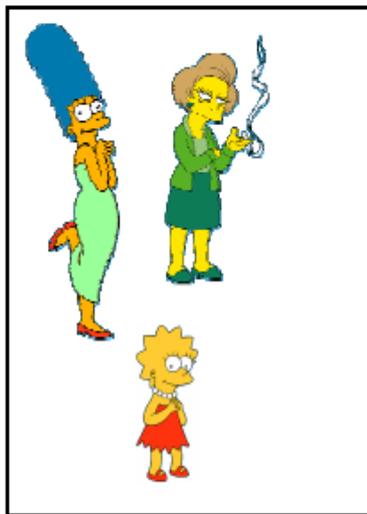
Clustering é subjetivo!



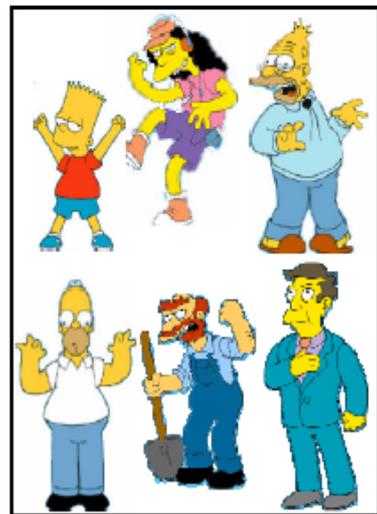
Simpson's Family



School Employees



Females



Males

Prognósticos

Embora todas essas aplicações anteriores envolvam previsões, os **prognósticos** as utilizam de modo diferente. Partem de uma série de valores existentes para prever quais serão os outros valores.

Por exemplo, um prognóstico **pode descobrir padrões** nos dados que ajudam os gerentes a estimar o valor futuro de variáveis como números de vendas.

Esses sistemas que realizam uma análise de alto nível quanto a padrões ou tendências, mas também podem esmiuçar os dados para revelar mais detalhes, se necessário.

Existem aplicações de data mining para todas as **áreas funcionais da empresa**, bem como para **o trabalho científico ou governamental**.

É comum usar o data mining para analisar **detalhadamente padrões em dados sobre consumidores** e, a partir disso montar campanhas de marketing um-a-um ou identificar clientes lucrativos.

Tendências Tecnológicas para Ambientes de Mineração

- **Inclusão de processos de mineração** de dados nos sistemas gerenciadores de banco de dados.
- Fornecedores como Oracle, Teradata, Microsoft e IBM estão cada vez mais, incluindo objetos de banco de dados com **capacidades de mineração de dados**.

Mitos (Turban et al, 2009)

Mito	Realidade
O data mining fornece previsões imediatas como bola de cristal	O data mining é um processo com várias etapas que exige projeto e uso proativos e calculados
O data mining ainda não é viável para aplicações de negócios	A tecnologia atual está pronta para escolher aproximadamente qualquer negócio
O data mining exige um banco de dados dedicado e distinto	Devido aos avanços na tecnologia de banco de dados, um banco de dados dedicado não é necessário, embora seja desejado.
Somente aqueles com formação avançada podem utilizar o data mining	Ferramentas baseadas na Web mais recentes permitem que gerentes de todos os níveis educacionais realizem o data mining

Data Mining

A força impulsionadora que torna o **data mining** atraente para os usuários finais é o **processamento de linha de pensamento** em que a resposta estimula outra pergunta.

FIM

Glossário - Algoritmo

- Algoritmo: é uma seqüência finita de instruções bem definidas e não ambíguas, cada uma das quais pode ser executada mecanicamente num período de tempo finito e com uma quantidade de esforço finita.
- O conceito de algoritmo é freqüentemente ilustrado pelo exemplo de uma receita, embora muitos algoritmos sejam mais complexos. Eles podem repetir passos (fazer iterações) ou necessitar de decisões (tais como comparações ou lógica) até que a tarefa seja completada. Um algoritmo corretamente executado não irá resolver um problema se estiver implementado incorretamente ou se não for apropriado ao problema.

Glossário – Inteligência Artificial

- *“Inteligência Artificial (IA) é a área da ciência da computação orientada ao entendimento, construção e validação de sistemas inteligentes, isto é, que exibem, de alguma forma, características associadas ao que chamamos inteligência”.*

Bibliografia

- CARVALHO, L.A.V. Datamining - A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. Rio de Janeiro: Ed. Ciência Moderna Ltda.,2005.
- LAUDON, K.C., LAUDON, J.P Sistemas de Informação Gerenciais. 7ª.ed. São Paulo: Pearson Prentice Hall, 2007.
- LEME FILHO, Trajano. Business Intelligence no Microsoft Excel. Axcel books, 2004.
- LUCAS Jr.,H. C. Tecnologia da informação – tomada de decisão estratégica para administradores. Ed. LTC, 2006.
- PINHEIRO, C.A.R. Inteligência analítica. Mineração de Dados e Descoberta de Conhecimento. Ed. Ciência Moderno. 2008.