Data Warehouse & Data Mining

Data Warehouse

Data Warehouse

- Histórico
  - Criado pela IBM na década de 60 com o nome Information Warehouse
  - Relançado diversas vezes sem grande sucesso
  - O nome Data Warehouse foi dado por William H. Inmon, considerado o inventor desta tecnologia
  - Tornou-se viável com o surgimento de novas tecnologias para armazenar e processar uma grande quantidade de dados

Data Warehouse

- Conceito
  - Sistema que armazena dados históricos usados no processo de tomada de decisão
  - Integra os dados corporativos de uma empresa em um único repositório
- Funcionalidade
  - Criar uma visão única e centralizada dos dados que estavam dispersos em diversos BDs
  - Permitir que usuários finais executem consultas, gerem relatórios e façam análises

Data Warehouse

- BDs usados nas aplicações de negócio são chamados BDs operacionais
- DW é um BD informacional alimentado com dados dos BDs operacionais da empresa
  - Disponibiliza dados atuais e dados históricos
  - Dados podem ser sumarizados (condensados) para que sejam analisados
  - Contém também metadados, que são dados sobre os dados armazenados no DW

BD Operacional X Data Warehouse

	BD Operacional	Data Warehouse
Usuários	Funcionários	Alta administração
Utilização	Tarefas cotidianas	Decisões estratégicas
Padrão de uso	Previsível	Difícil de prever
Princípio de funcionamento	Com base em transações	Com base em análise de dados
Valores dos dados	Valores atuais e voláteis	Valores históricos e imutáveis
Detalhamento	Alto	Sumarizado
Organização dos dados	Orientado a aplicações	Orientado a assunto

Principais Características de um DW

- Para que seja considerado um Data Warehouse, um banco de dados deve:
  - Coletar dados de várias fontes
  - Dados coletados devem ser transformados para que haja uma visão única dos dados
  - Dados devem ser usados por aplicativos para obter informações que dêem apoio à decisão
- Um Data Warehouse também deve ser:
  - Orientado a assunto
  - Integrado
  - Não-volátil
  - Variável com o tempo

Principais Características de um DW

Orientação a assunto

- Os dados em um DW são organizados de modo a facilitar a análise dos dados
- Dados são organizados por assunto e não por aplicação, como em BDs operacionais

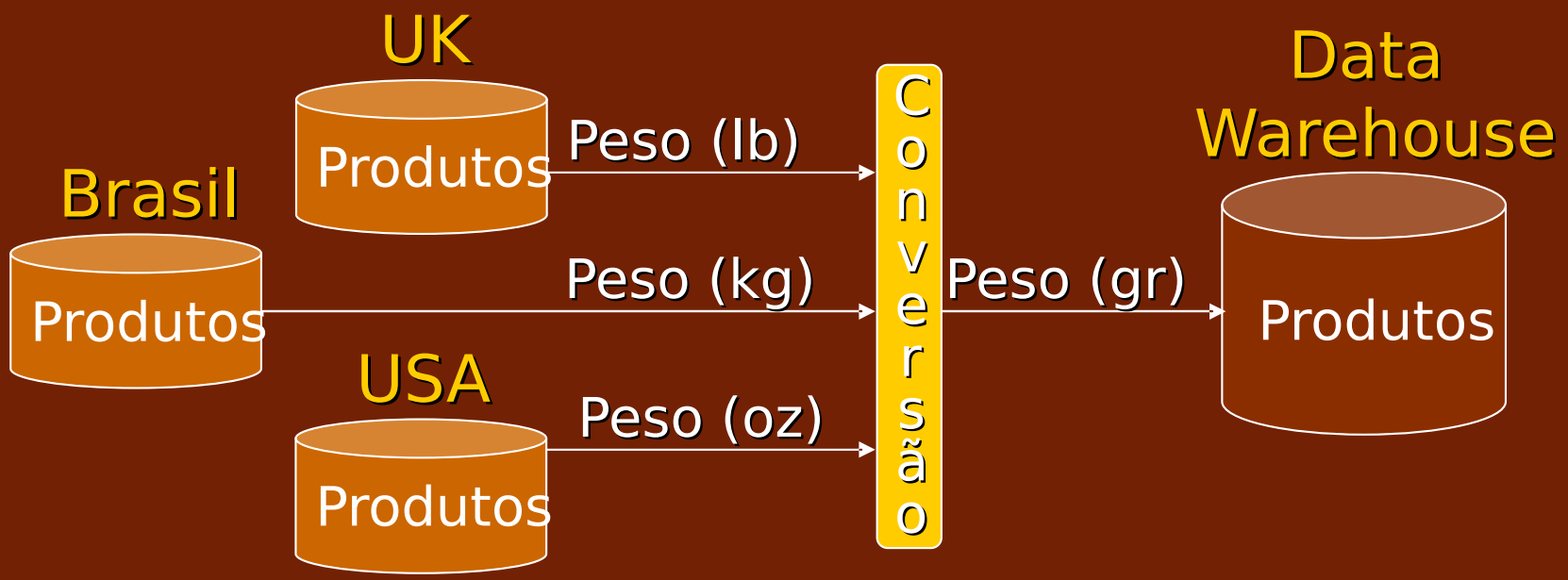
Aplicação de Venda

Análise de Vendas

Principais Características de um DW

Integração

- Dados de um DW provém de diversas fontes
- Dados podem ser sumarizados ou eliminados
- Formato dos dados deve ser padronizado para uniformizar nomes, unidades de medida, etc.



Principais Características de um DW

- Não-Volátil
  - Dados não são mais alterados depois de incluídos no DW
  - Operações no DW
    - Em um BD operacional é possível incluir, alterar e eliminar dados
    - Já no DW é possível apenas incluir dados
  - Garante que consultas subseqüentes a um dado produzirão o mesmo resultado

Principais Características de um DW

Variável com o Tempo

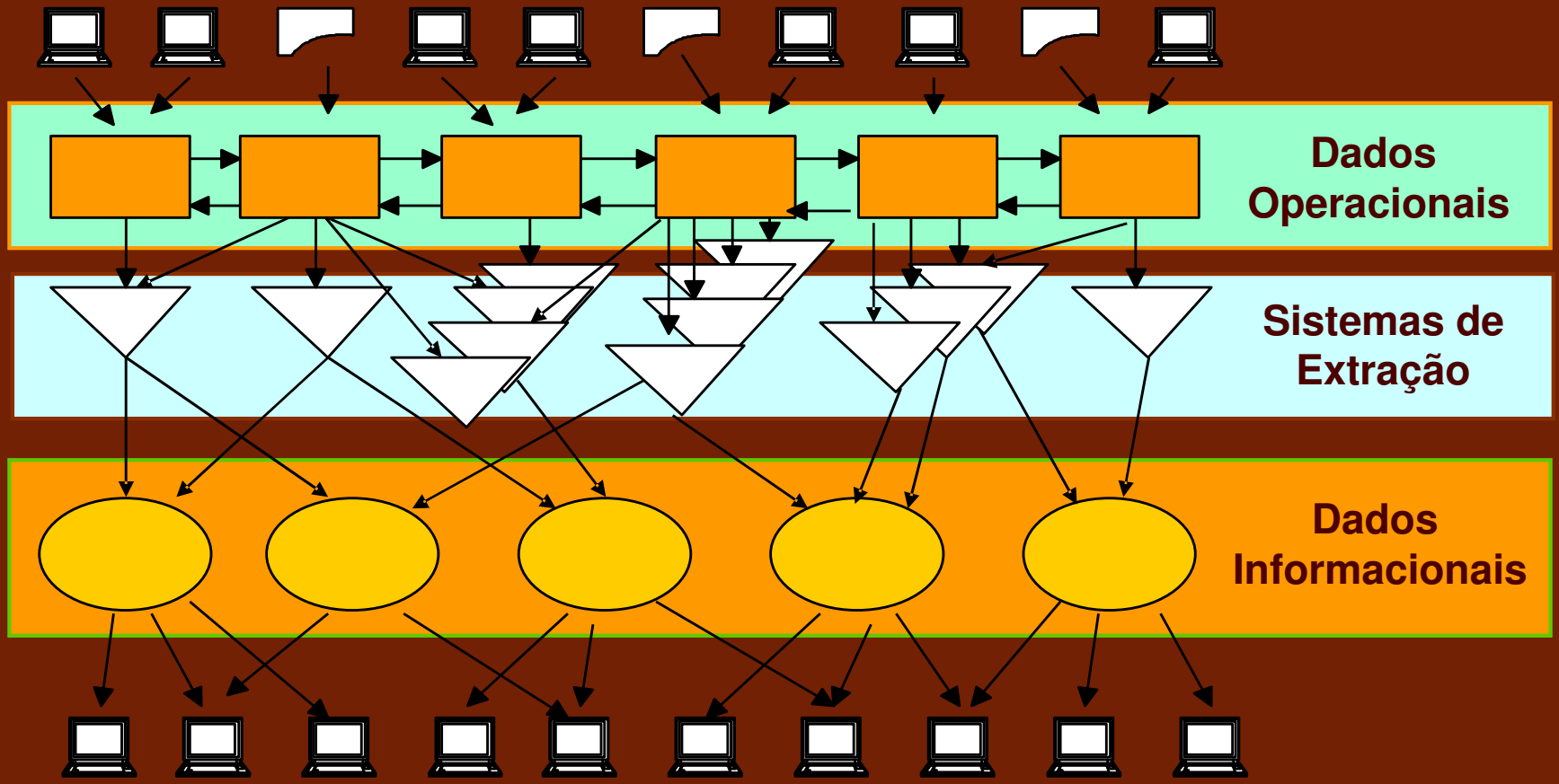
Os dados no DW são relativos
a um determinado instante de tempo

Produto	Preço
Caneta Azul	0,50
Lápis Preto	0,30
...	...

Produto	Jan/03	Fev/03	Mar/03
Caneta Azul	0,40	0,45	0,50
Lápis Preto	0,25	0,28	0,30
...	...	...	...

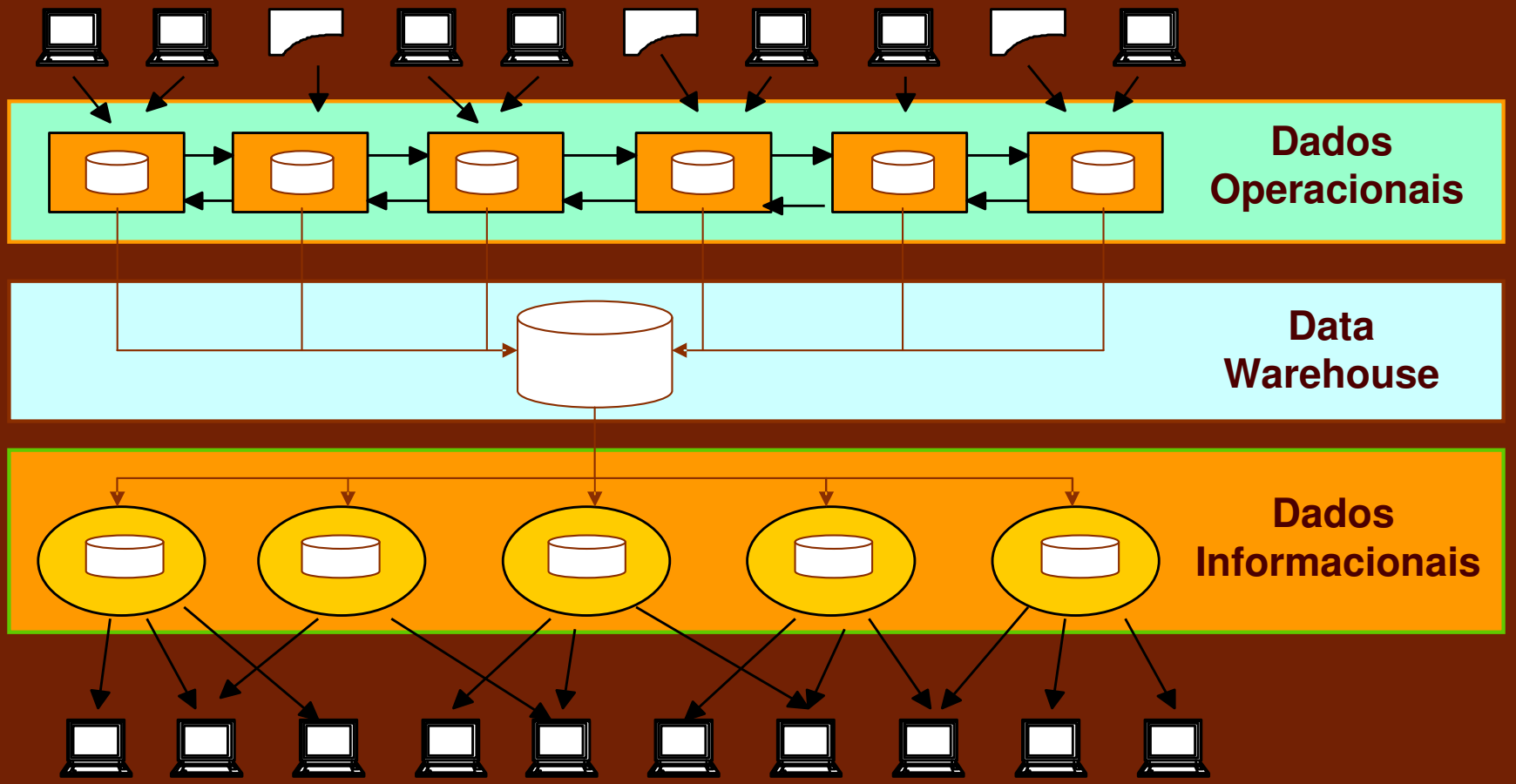
Arquitetura de um DW

Sistemas de Extração Tradicionais



Arquitetura de um DW

Sistemas baseados em Data Warehouse



Arquitetura de um DW

- Principais tarefas efetuadas pelo DW
  - Obter dados dos BDs operacionais e externos
  - Armazenar os dados
  - Fornecer informações para tomada de decisão
  - Administrar o sistema e os dados
- Principais componentes do DW
  - Mecanismos para acessar e transformar dados
  - Mecanismo para armazenamento de dados
  - Ferramentas para análise de dados
  - Ferramentas de gerência

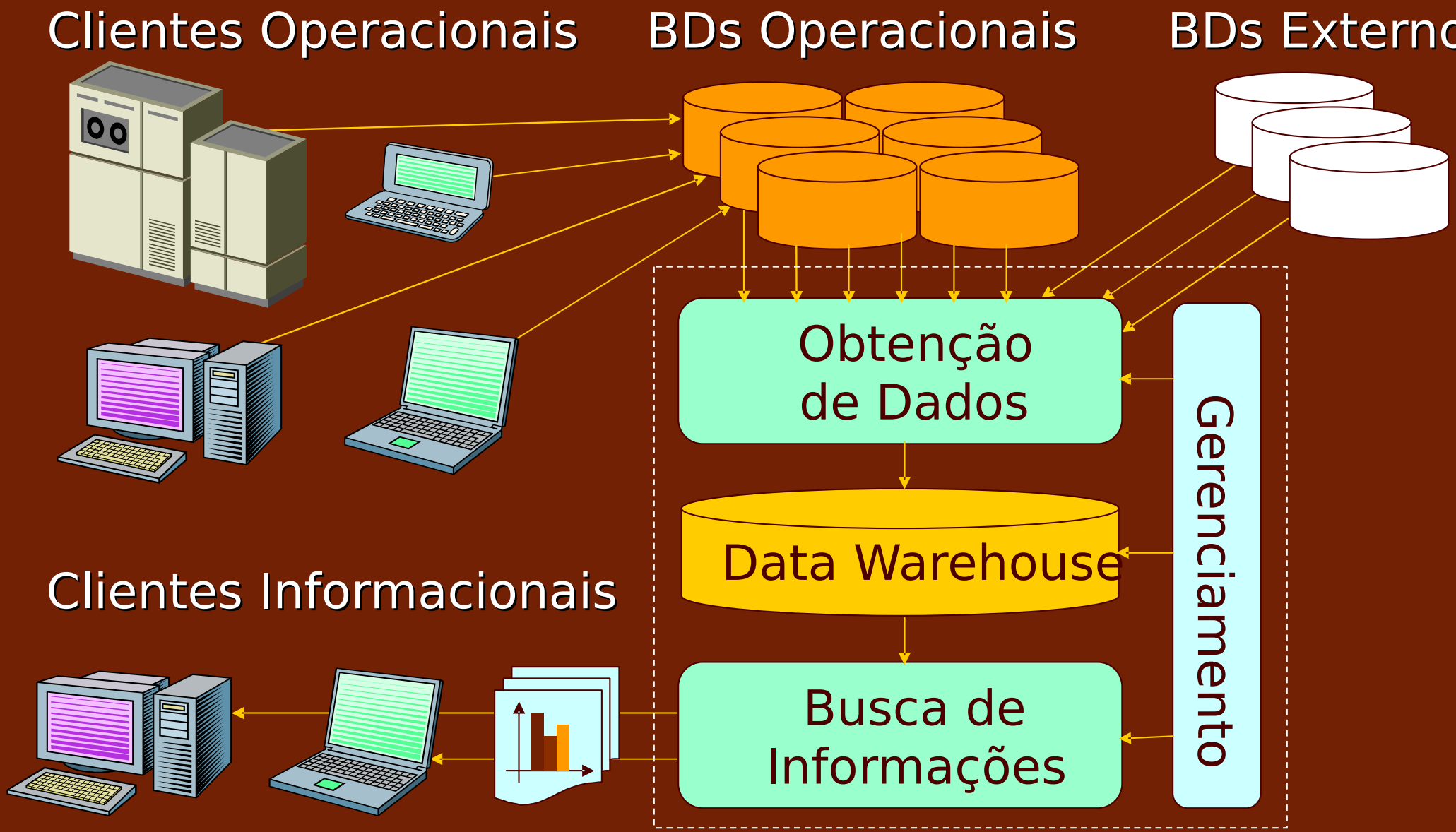
Estrutura Interna de um DW

- Requisitos do DW
  - Eficiente
    - Grande volume de dados imutáveis
    - Processamento paralelo e/ou distribuído
  - Confiável
    - Funcionamento do sistema
    - Resultado das análises
  - Expansível
    - Crescente volume de dados
    - Maior número de fontes de dados

Estrutura Interna de um DW

- Em geral são usados BDs relacionais para armazenar os dados do DW
  - Capazes de manter e processar grandes volumes de dados
  - Otimizados para lidar com dados imutáveis
- As ferramentas de análise empregam:
  - Técnicas de mineração de dados
  - Inteligência artificial: redes neurais, fuzzy, etc.
  - A Internet: Web mining, agentes móveis, etc.

Estrutura Interna de um DW

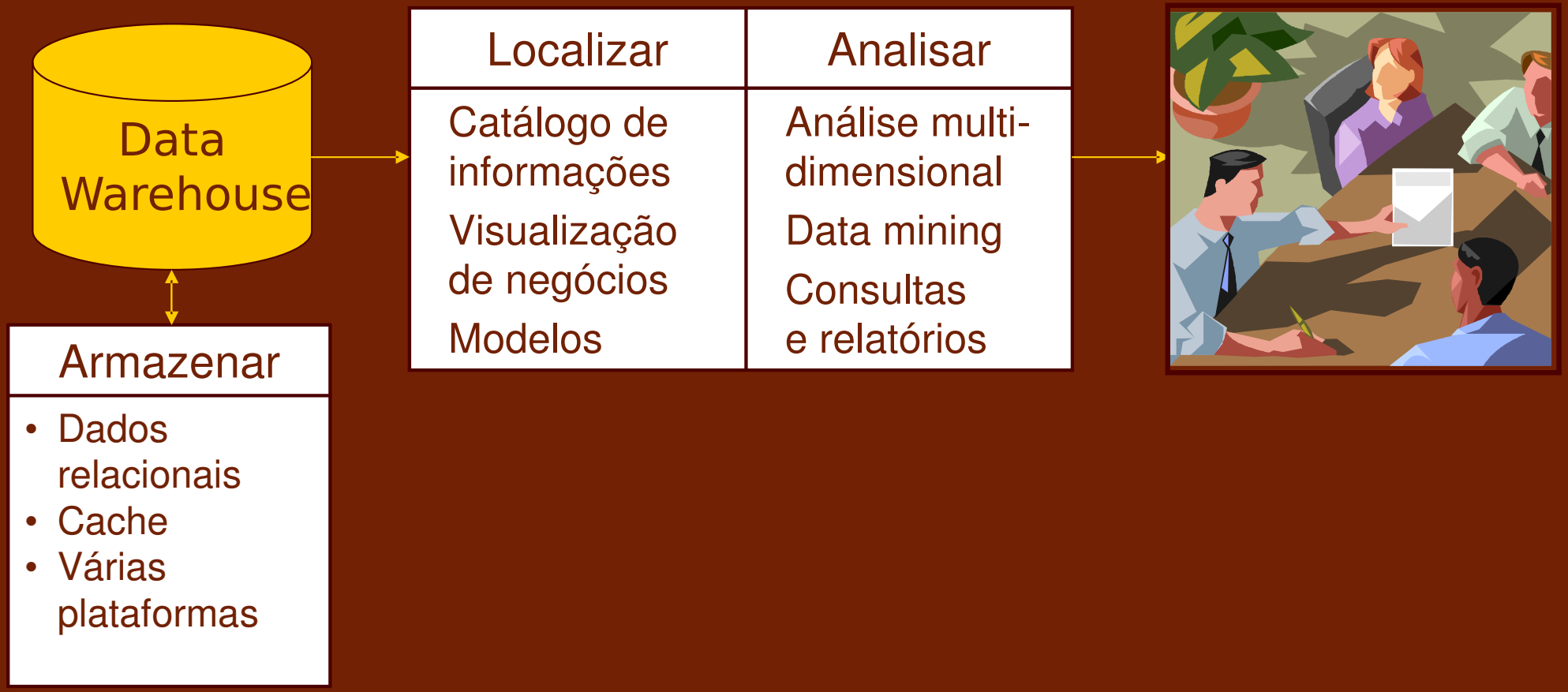


Estrutura Interna de um DW

Obtenção de Dados

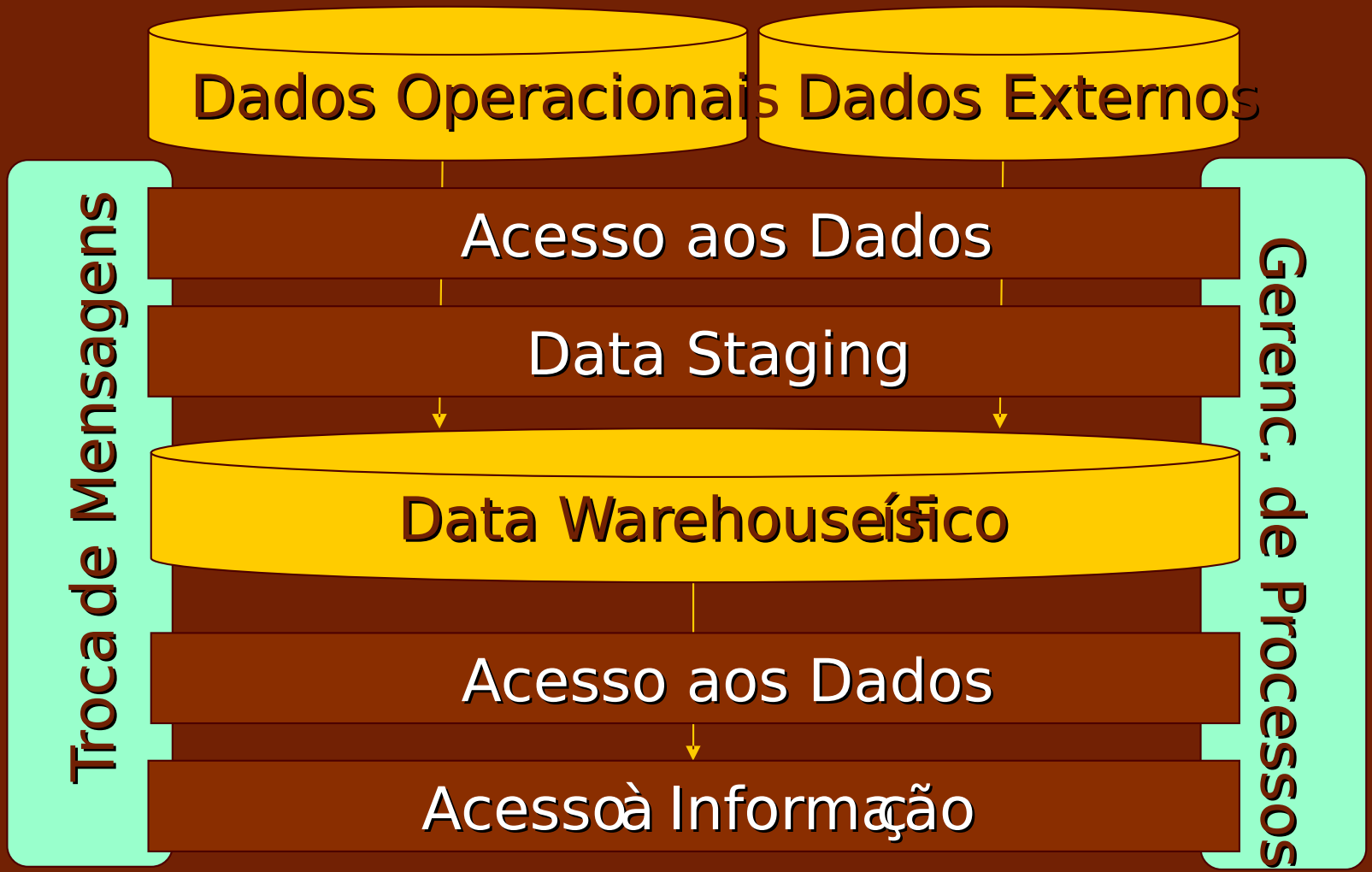
Estrutura Interna de um DW

Busca de Informações



Estrutura Interna de um DW

Modelo de Camadas



Estrutura Interna de um DW

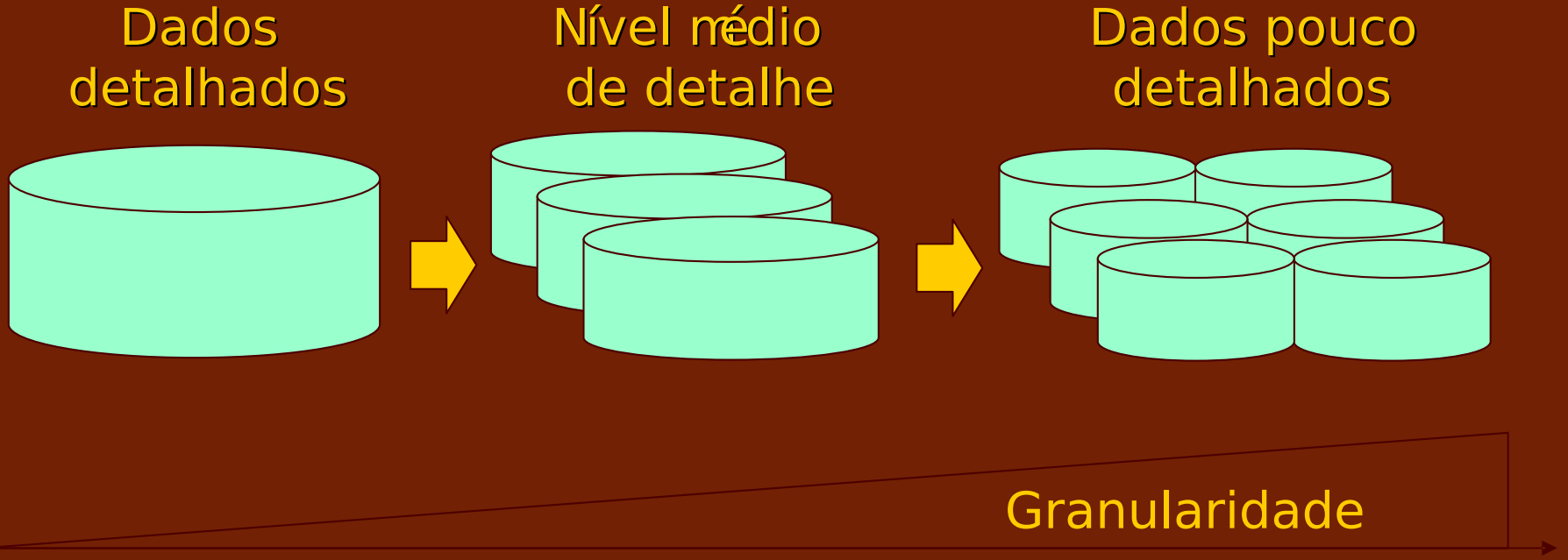
Funções das Camadas do DW

- Dados Operacionais/Externos: fontes de dados
- Acesso aos Dados: extrair dados dos BDs
- Data Staging: transformar e carregar dados
- Data Warehouse Físico: armazenar dados
- Acesso aos Dados: localizar dados para análise
- Acesso à Informação: analisar dados
- Troca de Mensagens: transportar dados
- Gerenc. de Processos: controlar atividades

Granularidade

Granularidade

- Nível de detalhe dos dados
- De extrema importância no projeto do DW



Granularidade

- Definir a granularidade adequada é vital para que o DW atenda seus objetivos
  - Mais detalhes Mais dados Análise mais longa Informação mais detalhada
  - Menos detalhes Menos dados Análise mais curta Informação menos detalhada
- Para evitar que se perca informação são criados vários níveis de granularidade

Granularidade

- Dados x Granularidade
  - Dados Atuais
    - Refletem acontecimentos recentes
    - Alto nível de detalhe (baixa granularidade)
  - Dados Sumarizados
    - Dados históricos condensados
    - Menor nível de detalhe (maior granularidade)
  - Dados Antigos
    - Dados históricos mantidos em fita, CD, etc
    - Alto nível de detalhe (baixa granularidade)

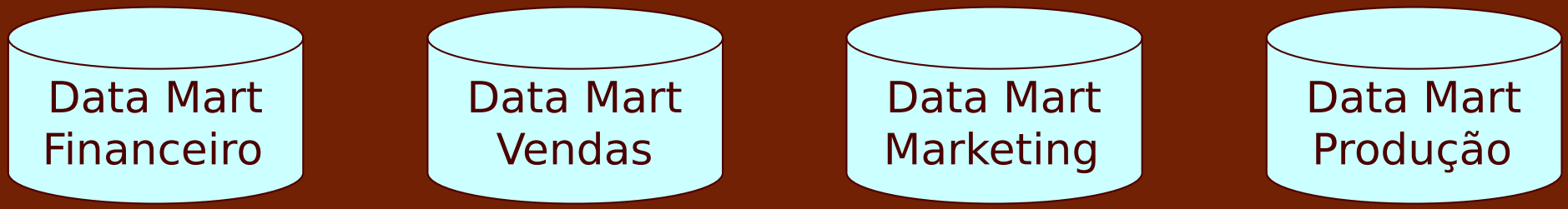
Granularidade

- Processo de sumarização
  - Aplica um novo esquema de modo a condensar os dados
  - Ex.: armazenar totais, médias, etc.
- Processo de envelhecimento
  - Transfere os dados antigos do HD para fita, CD, etc.
  - Mantém o nível de detalhe para que nenhuma informação seja perdida

Granularidade

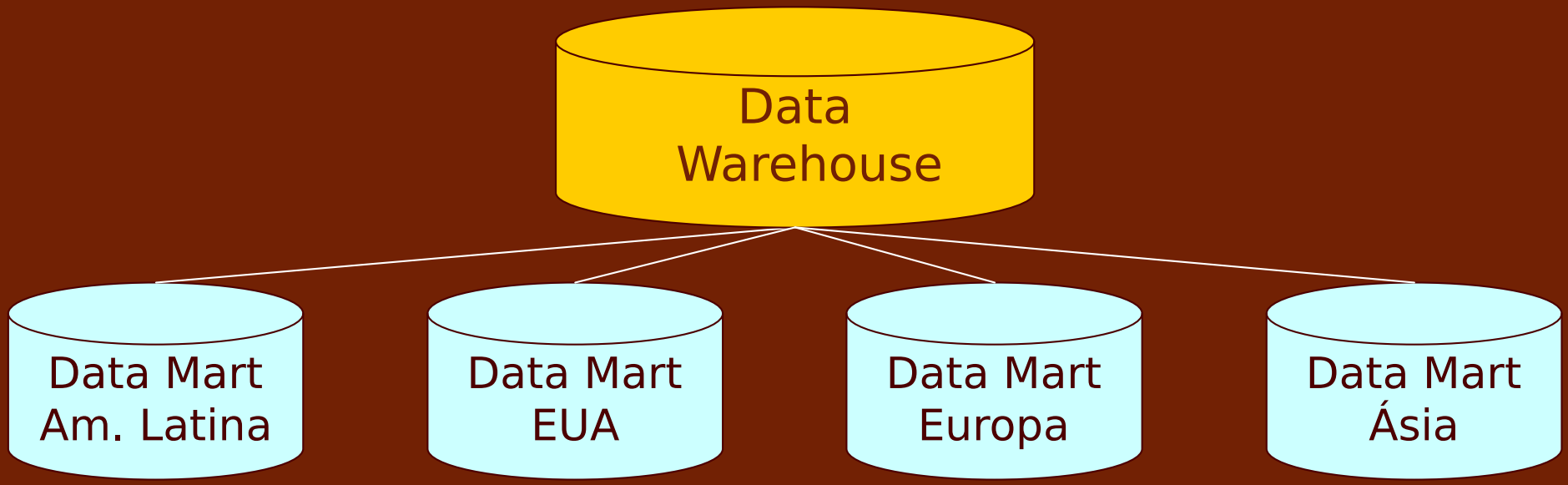
Data Marts

- Dados mantidos no DW são separados por assunto em subconjuntos de acordo com:
  - A estrutura interna da empresa
  - O processo de tomada de decisão
- Estes subconjuntos dos dados são chamados de Data Marts



Data Marts

- Um Data Mart desempenha o papel de um DW departamental, regional ou funcional
- Uma empresa pode construir seus Data Marts gradativamente a partir do DW



Data Marts

- Dados podem ser repetidos em dois ou mais Data Marts
- Os mesmos dados podem estar representados com granularidade diferente
- Ex:

Vendas detalhadas

Vendas totais mensais

Metadados

- Os Metadados são dados sobre os dados
  - Para cada atributo mantido no DW há uma entrada no dicionário de dados
  - Os dados são processados, atualizados e consultados partindo dos metadados
  - Usuários ficam conhecendo a estrutura e o significado dos dados
  - No BD operacional, a estrutura e o significado dos dados estão embutidos nas aplicações

Metadados

Camadas de Metadados

Metadados Operacionais

- Definem a estrutura dos dados operacionais

Metadados do DW

- Orientados por assunto
- Informam como os dados do DW foram calculados e como devem ser interpretados

Metadados do Usuário

- Organizam os metadados do DW com base em conceitos familiares ao usuário final

Metadados

- Classificação em função dos dados descritos
  - Metadados de Mapeamento
    - Como BDs operacionais são mapeados no DW
  - Metadados de Sumarização
    - Como os dados foram sumarizados no DW
  - Metadados Históricos
    - Como a estrutura dos dados vem mudando
  - Metadados de Padrões de Acesso
    - Como os dados do DW vem sendo acessados
  - Metadados de Miscelânea

Metadados

- Fontes de Metadados
  - Código fonte dos SBDs operacionais
  - Diagramas CASE de BDs operacionais e do DW
  - Documentação dos BDs operacionais e do DW
  - Entrevistas com usuários, administradores e programadores dos BDs e do DW
  - O ambiente de DW
    - Freqüência de acesso aos dados, tempo de resposta, controle de usuários, etc.

Acesso aos Dados

Acesso em Duas Camadas

Acesso em Três Camadas

Tipos de Data Warehouse

- DW baseado em Servidor
  - Mainframe ou servidor de rede local (LAN)
- DW Virtual
  - Reúne dados operacionais e dados históricos mantidos em BDs – não há um DW central
- DW Distribuído
  - DW global reúne dados de vários DWs locais
- DW baseado na Web
  - Dados provenientes da World Wide Web

Data Mining

Data Mining

Motivações

- 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).

Data Mining

- Definição
  - 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

Uma empresa utilizando data mining é capaz de:

- Criar parâmetros para entender o comportamento do consumidor
- Identificar afinidades entre as escolhas de produtos e serviços
- Prever hábitos de compras
- Analisar comportamentos habituais para detectar fraudes

Data Mining

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 qual das técnicas 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.

Evolução

Evolução	Perguntas	Tecnologia disponível	Características
Coleção de dados 1960	"Qual foi meu rendimento total nos últimos cinco anos ?"	Computadores, Fitas, discos	Retrospectiva, Dados estáticos como resposta
Acessos aos dados 1980	"Qual foi meu rendimento no Brasil no último janeiro ?"	RDBMS, SQL, ODBC	Retrospectiva, dados dinâmicos a nível de registros como resposta
Data warehousing & suporte a decisão 1990	"Qual foi meu rendimento no Brasil no último janeiro? Do sul até o nordeste"	Processamento analítico on-line, banco de dados multidimencionais, data warehousing	Retrospectiva, dados dinâmicos em múltiplos níveis como resposta
Data Mining Atualmente	"Porque alguns produtos são mais vendidos na região sul ?"	Algoritmos avançados, computadores multiprocessados, B.D. grandes e poderosos	Prospectivo, Informações (perspectivas) como resposta.

O Processo Data Mining

- Fases / Etapas.
  - Seleção.
  - Pré-processamento.
  - Transformação.
  - Data mining.
  - Interpretação e Avaliação.

O Processo Data Mining

- Seleção
  - Selecionar ou segmentar dados de acordo com critérios definidos
- Pré-processamento
  - Estágio de limpeza dos dados, onde informações julgadas desnecessárias são removidas
  - Reconfiguração dos dados para assegurar formatos consistentes (identificação)

O Processo Data Mining

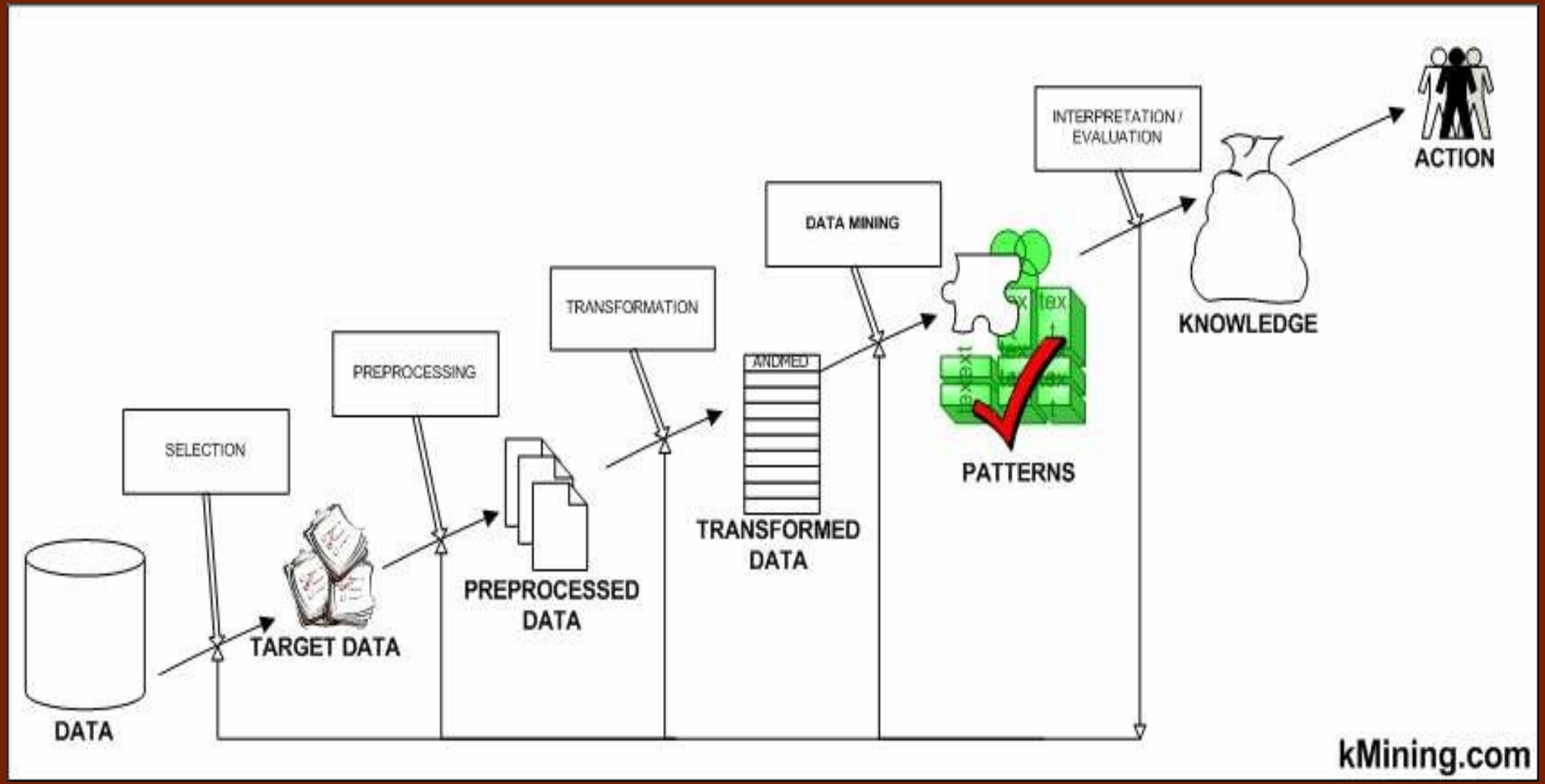
- Transformação
  - Transforma-se os dados em formatos utilizáveis. Esta depende da técnica data mining usada
    - Disponibilizar os dados de maneira usável e navegável
- Data mining
  - É a verdadeira extração dos padrões de comportamento dos dados
  - Utilizando a definição de fatos, medidas de padrões, estados e o relacionamento entre eles

O Processo Data Mining

Interpretação e Avaliação

Identificado os padrões pelo sistema, estes são interpretados em conhecimentos, os quais darão suporte a tomada de decisões humanas

Uma arquitetura data mining



Aprendizagem para data mining

Aprendizagem computacional

- Automação do processo de aprendizagem, através da construção de regras baseadas em observações dos estados e transações do ambiente.
- Examina os exemplos e seus resultados e aprende como reproduzi-los e como fazer generalizações sobre novos casos

Aprendizagem para data mining

- Aprendizagem indutiva:
  - Faz análise nos dados para encontrar padrões
  - Agrupa objetos similares em classes
  - Formula regras

  - Aprendizagem supervisionada
    - Aprende baseando-se em exemplos (“professor” ajuda a construir um modelo definido de classes e fornecendo exemplos de cada classe ⇒ formular a descrição e a forma da classe)
  - Aprendizagem não supervisionada
    - Aprende baseando-se em observações e descobertas (não se define classes, deve-se observar os exemplos e reconhecer os padrões por si só ⇒ uma descrição de classes para cada ambiente).

Funções do data mining

- Modelo de verificação
  - Aprende baseando-se em exemplos pré-classificados (+/-)
  - Objetivo: formular descrições consistentes e gerais de classes em função de seus atributos.
- Modelo de descoberta
  - Aprende baseando-se em observações e descobertas
  - Descoberta automática de informações ocultas
  - Procura ocorrências de padrões, tendências e generalizações sobre os dados sem a intervenção do usuário
  - Agrupar elementos similares

Funções do data mining

- Modelo de classificação :
  - Atributos mais significativos definidos um classe
  - O usuário define as atributos para cada classe
  - Aplica regras para criar modelos de ações futuras
- Associação:
  - Procura registos que tenham similaridades associativas
  - Podem ser expressados por regras

Funções do data mining

- Padrões temporais/seqüenciais :
  - Analisa registros num período de tempo, procurando encontrar padrões (eventos/compras) de comportamento.
  - Identificar o perfil do cliente
  - Identificar padrões que precedem outros padrões
- Segmentação/agrupamento:
  - Segmenta a base de dados em grupos por suas similaridade e diferenças
  - O sistema tem que descobrir por si próprio as similaridade e diferenças

Técnicas

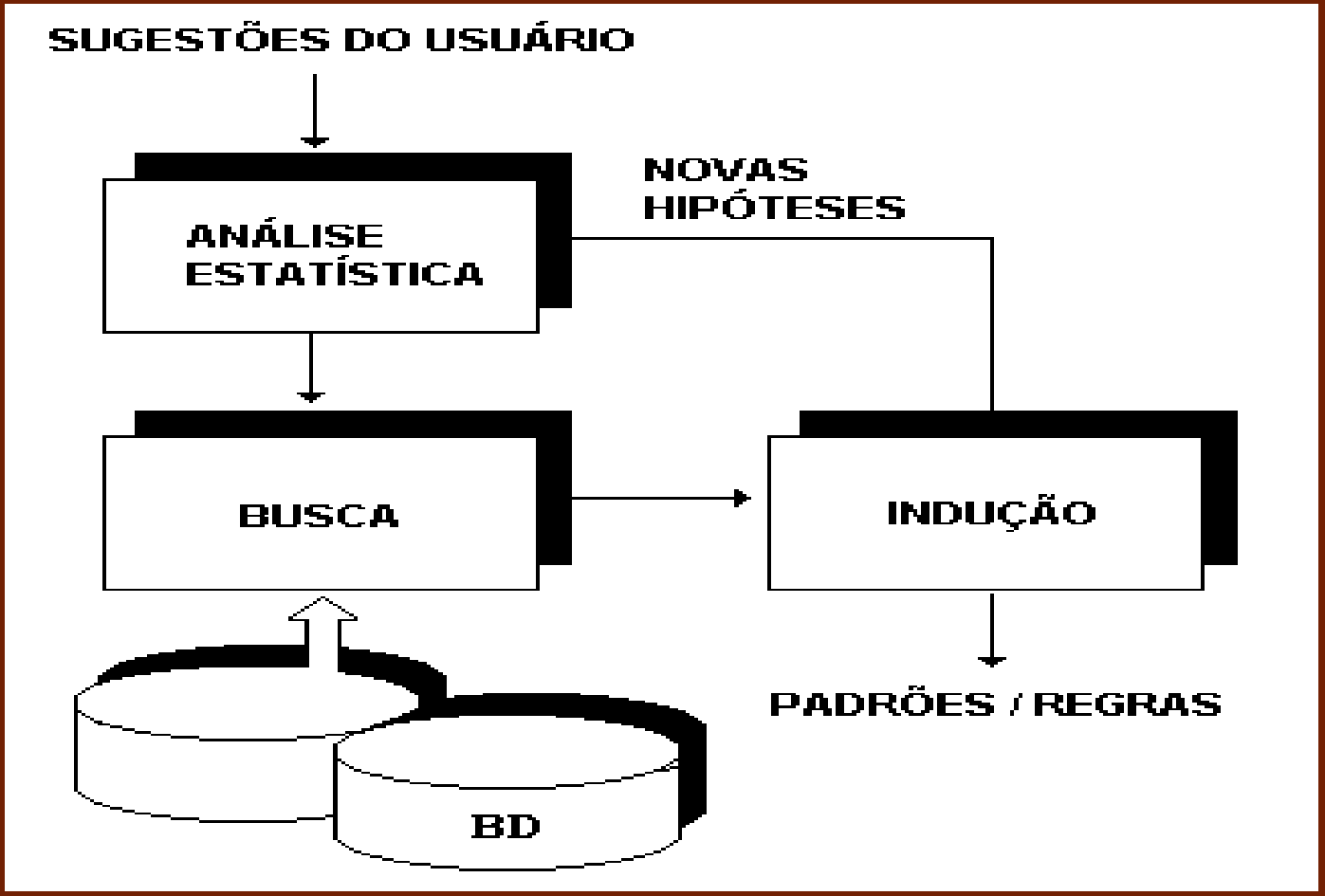
Indução

Regras indutivas (rule induction)

- Regra indutiva é o processo de olhar uma série de dados e, a partir dela, gerar padrões.
- Pode-se trabalhar com dados numéricos ou não
- Pelo fato de explorar uma série de dados, o sistema indutivo cria hipóteses que conduzem a padrões
- Regras cobertas ⇒ comportamentos estáveis
- Regras inexatas ⇒ margem de precisão “fixada” (%)

Técnicas

Indução:



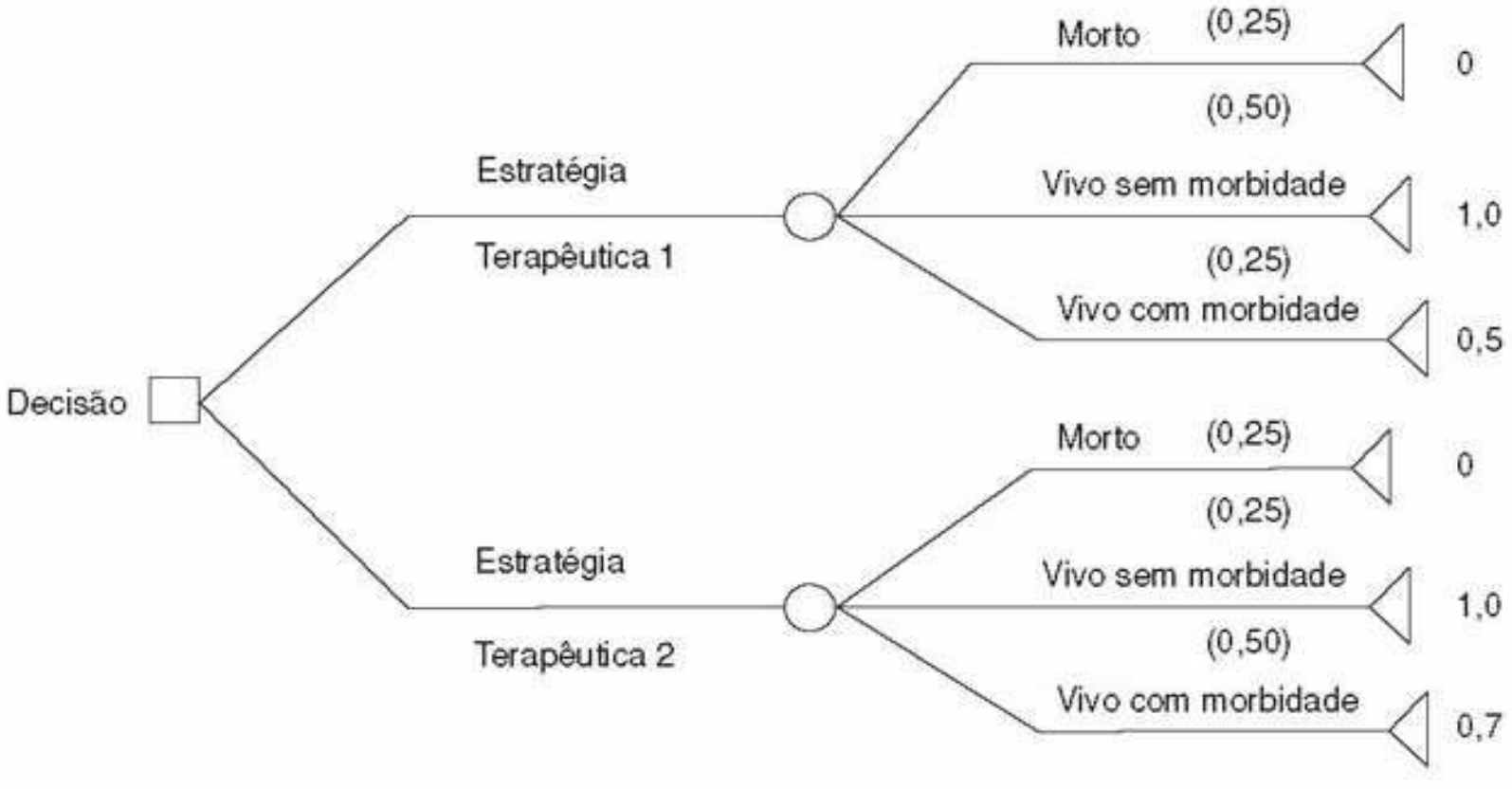
Técnicas

- Árvores de decisão:
  - Representações simples do conhecimento
  - Utilização de regras condicionais
  - A partir de um conjunto de valores decide SIM ou NÃO
  - Mais rápida e mais compreensível que redes neurais

Técnicas

Árvores de decisão:

FIGURA 2. Árvore de decisão utilizada para análise de decisões em saúde[a]



[a] O quadrado na figura indica um ponto de decisão; os círculos indicam os pontos de chance; e os triângulos indicam os desfechos quantitativos (medidos em utilidades). Os números entre parênteses indicam a probabilidade de ocorrência de cada desfecho possível.

Técnicas

Redes Neurais:

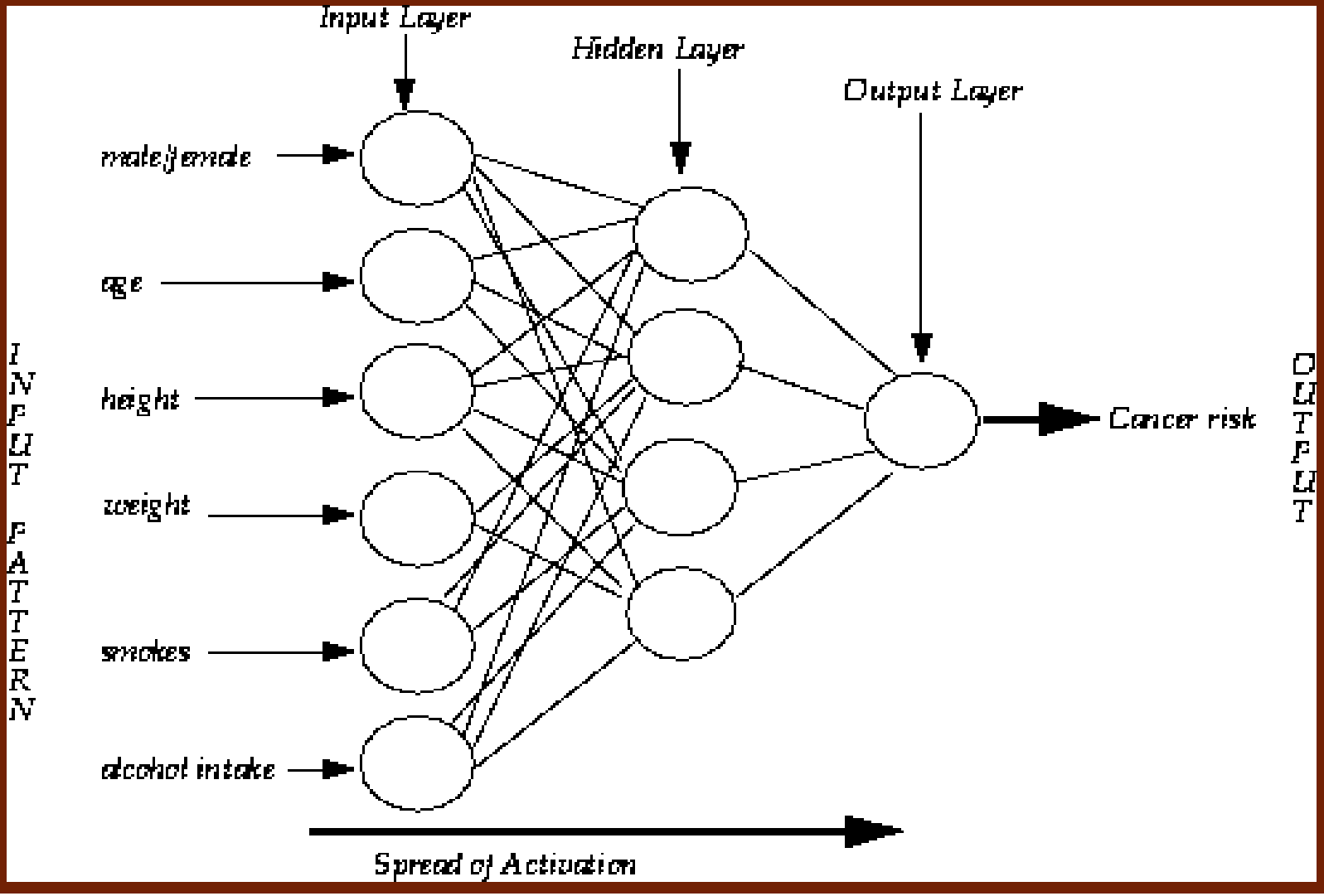
- É uma abordagem computacional que envolve desenvolvimento de estruturas matemáticas com a habilidade de aprender
- Estruturalmente, uma rede neural consiste em um número de elementos interconectados (chamados neurônios/nós), que possuem entrada, saída e processamento
- São organizados em camadas que aprendem pela modificação da conexão

Técnicas

- Redes Neurais:
  - Para construir um modelo neural, nós primeiramente "adestramos" a rede em um dataset de treinamento e então usamos a rede já treinada para fazer predições
  - Problemas:
    - Não retorna informação a priori
    - Não pode ser treinada em uma grande base de dados
    - Entrada não pode ser dados alfa-numéricos
    - Nenhuma explanação dos dados é fornecida

Técnicas

Redes Neurais:



Área de atuação

- Áreas de aplicações potenciais:
  - Vendas e Marketing
    - Identificar padrões de comportamento de consumidores
    - Associar comportamentos à características demográficas de consumidores
    - Campanhas de marketing direto
    - Identificar consumidores “leais”

Área de atuação

- Áreas de aplicações potenciais:
  - Bancos
    - Identificar padrões de fraudes (cartões de crédito)
    - Identificar características de correntistas
    - Mercado Financeiro
  - Médica
    - Comportamento de pacientes
    - Identificar terapias de sucessos para diferentes tratamentos
    - Fraudes em planos de saúdes
    - Comportamento de usuários de planos de saúde

Exemplos

- Fraldas e cervejas
  - O que as cervejas tem a ver com as fraldas ?
  - Homens casados, entre 25 e 30 anos, compravam fraldas e/ou cervejas às sextas-feiras à tarde no caminho do trabalho para casa;
  - Wal-Mart otimizou às gôndolas nos pontos de vendas, colocando as fraldas ao lado das cervejas
  - Resultado: o consumo cresceu 30%

Exemplos

- Lojas Brasileiras
  - Aplicou 1 milhão de dólares em técnicas de data mining
  - Reduziu de 51000 produtos para 14000 produtos oferecidos em suas lojas
  - Exemplo de anomalias detectadas:
    - Roupas de inverno e guarda chuvas encalhadas no nordeste
    - Batedeiras 110v a venda em SC onde a corrente é 220v

Conclusões

- Data Warehouse é um sistema de aquisição de informação que mantém um histórico para avaliação posterior, além de manter de forma unificada e padronizada os dados que são adquiridos
- Data Mining é um processo que permite compreender o comportamento dos dados
- Tem um suporte muito forte em I. A.
- Pode ser bem aplicado em diversas áreas de negócios
- Só será eficiente se o valor das informações extraídas exceder o custo do processamento dos dados brutos

Bibliografia

www.db-book.com

www.the-data-mine.com

http://www.intelliwise.com/reports/i2002.htm

www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

http://www.thearling.com/text/dmwhite/dmwhite.htm

Bibliografia

http://www.cs.bham.ac.uk/~anp/TheDataMine.html

http://www.satafe.edu/~kurt/index.shtml

http://pt.wikipedia.org/wiki/Data_Warehouse

www.baguete.com.br/artigosDetalhes.php?id=154

www.ica.ele.puc-rio.br/cursos/download/DM-**apostila**1.pdf