Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Sistemas de Apoio à Decisão

Técnicas e Algorítmos

Prof. Doutor Victor Lobo

Mestrado em Estatística e Gestão de Informação

Objectivo desta disciplina

- Dar uma visão geral sobre os SAD
  - Enquadramento na organização, e tipo que tarefas que podemos esperar destes sistemas
  - **Principais técnicas** disponíveis
  - Tendências actuais
- Aprender algumas técnicas mais avançadas
  - Sistemas difusos, Algoritmos genéticos, Sistemas auto-organizados, etc…

Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Programa (tarços gerais)

- **1. Introdução aos Sistemas de Apoio à Decisão**
- **2. Principais áreas de SAD**

- **3. Teoria da decisão e sistemas Bayesianos**
- **4. Pré-processamento, projecções, e métricas para dados. Estimativas de erro.**
- **5. Mapas auto-organizados (SOM)**
- **6. Aprendizagem e classificação baseada em instâncias**
- **7. Sistemas Fuzzy (Lógica Difusa)**
- **8. Algoritmos Genéticos**
- **9. Sistemas Periciais**
- **10. Redes Neuronais (para além de MLP)**
- **11. Estudo de casos**

Programa (Detalhado) 1/5

- **1. Introdução aos Sistemas de Apoio à Decisão**
  - 1.1. Sistemas de Apoio à Decisão (SAD).
  - 1.2. Processo de tomada de decisão.
  - 1.3. Indicadores para tomada de decisão.
  - 1.5. Tendências: Internet e Gruopware
- **2. Principais áreas de SAD**
  - 2.1. Organização de dados e datawarehousing.
  - 2.2. Visualização de dados.
  - 2.3. Geração de relatórios, indicadores, e OLAP.
  - 2.4. Modelação de incerteza.
  - 2.5. Técnicas de previsão –Visão geral, árvores e redes neuronais
  - 2.6. Técnicas de agrupamento – Visão geral, árvores e k-médias
  - 2.7. Heurísticas de Optimização.
  - 2.8. Pesquisa de soluções e sistemas periciais.

Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Programa (detalhado) 2/5

- **3. Teoria da decisão e sistemas Bayesianos**
  - 3.1 Conceitos gerais
  - 3.2 Decisões óptimas Bayesianas.
- **4. Pré-processamento, projecções, e métricas para dados. Estimativas de erro.**
  - 4.1 Técnicas de normalização
  - 4.2 Métricas para dados numéricos e categóricos
  - 4.3 Estimativas de erro de sistemas de classificação e regressão
  - 4.4 O problema dos valores em falta
  - 4.5 Técnicas para extracção de características e projecções

Programa (detalhado) 3/5

- **5.Mapas auto-organizados (SOM)**
  - 5.1. Conceitos fundamentais.
  - 5.2. Formalização dos SOM.
  - 5.3. Matrizes U e sua interpretação.
  - 5.4. Utilização e Variantes de SOM.
- **6. Aprendizagem e classificação baseada em instâncias**
  - 6.1 Algoritmo do vizinho mais próximo
  - 6.2 Variantes do vizinho mais próximo
  - 6.3 Escolha selectiva

Programa (detalhado) 4/5

- **7. Sistemas Fuzzy (Lógica Difusa)**
  - 7.1. Representação de incerteza.
  - 7.2. Funções de pertença.
  - 7.3. Operadores difusos.
  - 7.4. Clustering difuso.
  - 7.5 Outras abordagens: probabilidades e rough sets
- **8. Algoritmos Genéticos**
  - 8.1. Conceitos e definições.
  - 8.2 Problemas de codificação e espaço de busca
  - 8.3. Operadores de cruzamento e mutação.
  - 8.4. Operadores de selecção.

Programa (detalhado) 5/5

- **9.Sistemas Periciais**
  - 9.1 Arquitectura geral de sisteas periciais
  - 9.2 Lógica como paradigma de programação
  - 9.3 Estratégias para exploração do espaço de soluções
  - 9.4 Sistemas de forward chaining vs backward chaining
- **10. Redes Neuronais**
  - 10.1 Perceptrões Multi-camada (MLP)
  - 10.2 Redes de RBF
  - 10.3 Redes de Hopfield
  - 10.4 Support Vector Machines
  - 10.5 Outros tipos de redes
- **11. Estudo de casos**

Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Bibliografia

- Livros de texto (nenhum é seguido "à risca")
  - **Decision Support and Business Intelligence Systems**, Turban, E., J. E. Aronson, *et al.*, Prentice Hall, 2007
  - **Sistemas de Suporte à Decisão**, Bruno Cortes, FCA, 2005.
  - **Decision Support Systems in the 21st Century**, George Marakas, Prentice-Hall, 2002.

- Para os "topicos avançados"
  - Textos de apoio e referências próprias

Resolução de problemas práticos

- MS-Excel
- SAS Enterprise Miner
- Microsoft SQL server
- Alguns programas dedicados
  - WEKA
  - Matlab

Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Avaliação

- Exame Final
  - Obrigatório para todos (60 a 100% da nota)
- Trabalhos
  - Trabalho prático de grupo (opcional, 20%)
  - Trabalhos de Casa (opcional, até 20%)
  - Trabalho individual de pesquisa e síntese
    - Ler, apresentar, e comentar um artigo sobre aplicações práticas de SAD.
    - Avaliado em conjunto com os trabalhos de casa
  - NOTA MÍNIMA EM TODAS AS PROVAS – 10 valores

Horário de dúvidas e contactos

- Email: vlobo@isegi.unl.pt
  - Dúvidas
    - 5ª Feira às 21:15, 6ª Feira às 16:00
    - Por mail em qualquer altura
    - Sempre que estiver no ISEGI (!)
  - Material de apoio
    - www.isegi.unl.pt/docentes/vlobo

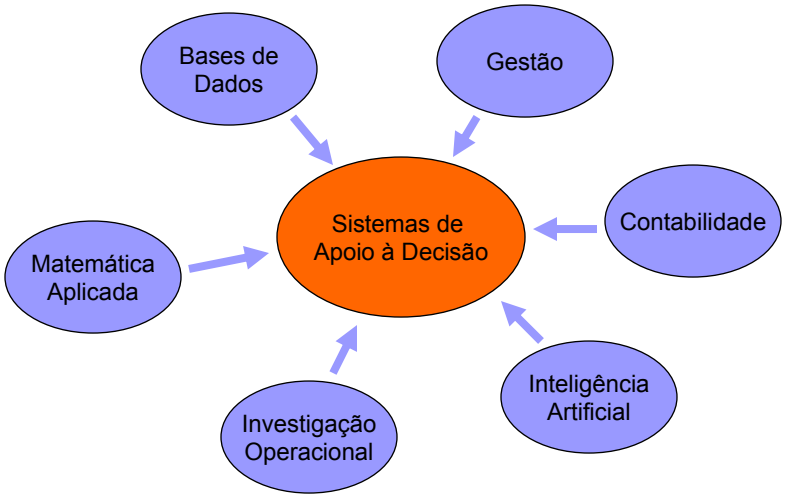
Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

O que é ? Exemplos ?

- No dia a dia…

Origens



Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Factores importantes para os SAD

- Quantidade de dados disponíveis
  - Dados operacionais, sensores de baixo custo
- Poder de cálculo e armazenamento de dados
- Sistemas computacionais de baixo custo
- Desenvolvimento científico e tecnológico
  - Inteligência Artificial e Aprendizagem Máquina (Machine Learning), e convergência com as técnicas mais clássicas da área da estatística, da investigação operacional, e do reconhecimento de padrões
- Software de fácil utilização
  - SAS, SAP, etc.

O que se espera obter de um SAD ?

- Informação útil e relevante
  - Permite decisões informadas
- Conselhos sobre a acção correcta a tomar
  - Não substitui o decisor, aconselha-o. Identifica situações inesperadas. Optimiza as acções necessárias para um dado objectivo
- Gestão e acompanhamento das decisões
  - Possibilidade de agir e medir consequências
- Ferramentas para trabalho em grupo
  - Trabalho colaborativo e comunicação interna
- Armazenamento e gestão de "conhecimento"
  - Capacidade para superar as limitações humanas de processamento de informação. "Lembra" lições aprendidas
- Ferramentas para obter vantagem competitiva
  - Decidir melhor e mais depressa que a concorrência, detectar oportunidades e falhas

Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Confusão na nomenclatura

- Sinónimos, “trademarks”, “partes” de SAD, diferentes perspectivas
  - DSS – Decision Suport Systems
  - BI – Buisiness Intelligence
  - KMS - Knowlege Management Systems
  - EIS – Enterprise Information Systems
  - ERP – Enterprise Resource Planning
  - BA- Buisiness Analytics
  - CRM – Customer Relation Management
  - Data Warehouse
  - Expert Systems
  - Intelligent Agents
  - Datamining
  - Groupware, GSS - Group Support Systems
  - SCM, EIP, OPAP, ERM, Etc, etc, etc,etc,

Funcionalidades (não são todas mandatórias)

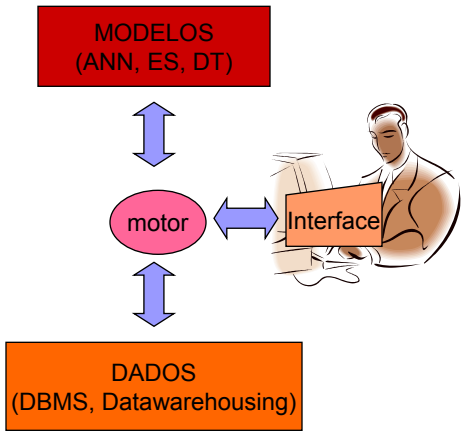
- Recolha de informação
- Visualização da informação
- Construção de modelos
- Fazer previsões, detectar situações anómalas (outliers)
- Resolução e optimização de problemas
- Suporte para decisões individuais ou em grupo
- Capacidade para lidar com problemas mal definidos e pouco estruturados
- Evolutivos, *i.e.*, capacidade para se adaptarem a novas situações…

Introdução a SAD

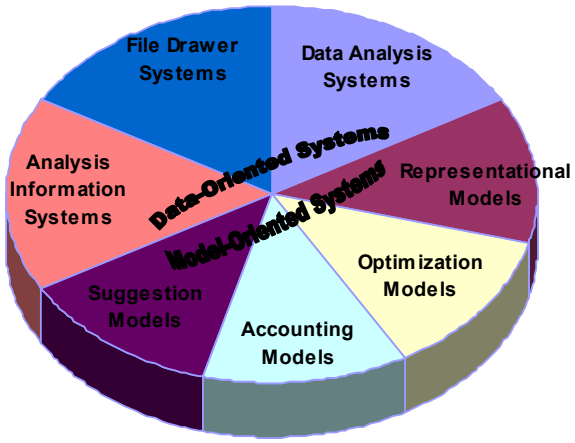
V 1.3, V.Lobo, EN/ISEGI, 2009

Componentes comuns

- Sistema de gestão de dados
- Sistema de gestão de modelos
- Motor de inferência
- Interface com o utilizador



Componentes (segundo Alter)



Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Tipos e ênfases dos SAD

- Centrados nos Dados *versus* Modelos
- Dedicados *versus* Generalistas
- Formais *versus Ad Hoc*
- Dirigidos (ou operativos) *versus* não-dirigidos (ou descritivos)
- Baseados na WEB

Evolução histórica

- Sempre houve "suporte à decisão"
  - Decidir com *razão* vs *coração*
  - Mais informação → melhor decisão
- Origem do termo "Decision Suport System"
  - Início dos anos 70 (Little, G. & S. Morton)
  - Usar modelos informáticos em gestão, produzindo software de fácil utilização
- Cada vez mais…amigáveis…potentes…abrangentes….
- Não há (nem pode haver…) o "SAD universal"

Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Diferentes perspectivas

- Teoria da decisão
  - O que é uma BOA decisão ?
- Engenharia e Informática
  - O que são as ferramentas que permitem uma BOA decisão ?
  - Como se fazem essas ferramentas ?
- Gestão
  - **Como se usam** essas ferramentas ?
- Como interpretar e usar um SAD ?
  - Compreender as ferramentas
  - Compreender o processo de tomada de decisão

Objectivos nesta cadeira

- **Compreender a importância** que os SAD têm para as organizações, e o modo como se integram nessas mesmas organizações.
- Compreender o **tipo de tarefas** que é executado pelos SAD.
- Compreender os problemas associados ao **armazenamento**, tratamento, e disponibilização ou **visualização** de grandes volumes de dados.
- Conhecer e compreender as principais **técnicas de previsão**.
- Conhecer e compreender as principais **técnicas de agrupamento**.
- Conhecer e compreender as principais **técnicas de pesquisa** e **optimização heurística**.
- Reconhecer a técnica **mais adequada a cada problema**, aplicá-la, e **compreender os resultados**.

Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Dominar algumas técnicas

- Visão mais geral
  - Conhecer as diversas técnicas disponíveis
- Particular ênfase
  - Mapas auto-organizados para clustering
  - Algoritmos genéticos
  - Sistemas “Fuzzy”
  - Instance Based Learning
  - Pré-processamento dos dados
  - … …

Software (para esta cadeira e para DSS)

- Excel !
  - Resolve muitos problemas.
  - Teste de métodos para “poucos” dados
- SAS - Enterprise Miner
  - Escalável para problemas “a sério”
  - Grande variedade de ferramentas
  - Pouca informação detalhada sobre métodos
  - Bom interface visual mas programação “pouco amigável”
  - www.sas.com – Muita informação sobre aplicações

Nosso patrocinador ! Disponível nas salas

Software (pacotes comerciais genéricos)

- SPSS – Clementine
  - Muito difundido nalgumas universidades
  - Versão de educação brevemente disponível
  - www.spss.com
- IBM - Intelligent Miner
  - Tem uma versão para dowload gratuito
  - http://www-306.ibm.com/software/data/iminer/
- SAP - Módulos de Business Intelligence
  - Grande variedade de módulos
  - http://www.sap.com/platform/netweaver/components/bi/index.epx

Software (pacotes facilmente disponíveis)

- WEKA
  - Para Datamining e "Machine Learning"
  - "open source" em Java
  - Corre em muitos ambientes, bastante completo (v3)
  - http://www.cs.waikato.ac.nz/ml/weka/
- Matlab (ou Octave e SciLab que são GNU)
  - Toolboxs de NN, DT, GA, ML, etc
    - SOMTOOLBOX (som), NETLAB (machine learning)
  - www.mathworks.com (site comercial da mathworks)
  - http://www.gnu.org/software/octave/
  - http://www.scilab.org/
- R
  - Package estatístico com muito suporte para datamining
  - Parecido com Matlab (mas diferente)
  - http://www.r-project.org/
- Outros – "Statistica Neural Networks", SOM_PAK, C4.5(original), SNNS, plug-ins para Excel, etc, etc, etc, etc.

Outros sites interessantes…

- DSS Resources
  - Prof. Daniel Power, livros, referências, etc
  - http://dssresources.com/
- Decisionarium
  - Software GNU, referências, etc
  - http://www.decisionarium.tkk.fi
- Machine Learning Network
  - www.mlnet.org
  - Software, dados, conferências, projectos, etc.
- Repositório de Irvine
  - www.ics.uci.edu/~mlearn
  - Dados, software, artigos
- Fabricantes de soluções "dedicadas"
  - Para gestão de terrenos, para marketing, etc, etc

Existem decisões "óptimas" ?

- Optimalidade
  - Definida em função de um objectivo
    - sem função objectivo não há um óptimo !
  - Exige informação completa
    - Há sempre incerteza num caso real
    - Incerteza aumenta com a "não estruturação"
  - Matematicamente é encontrar o máximo de uma função
    - Exemplo da decisão óptima de Bayes

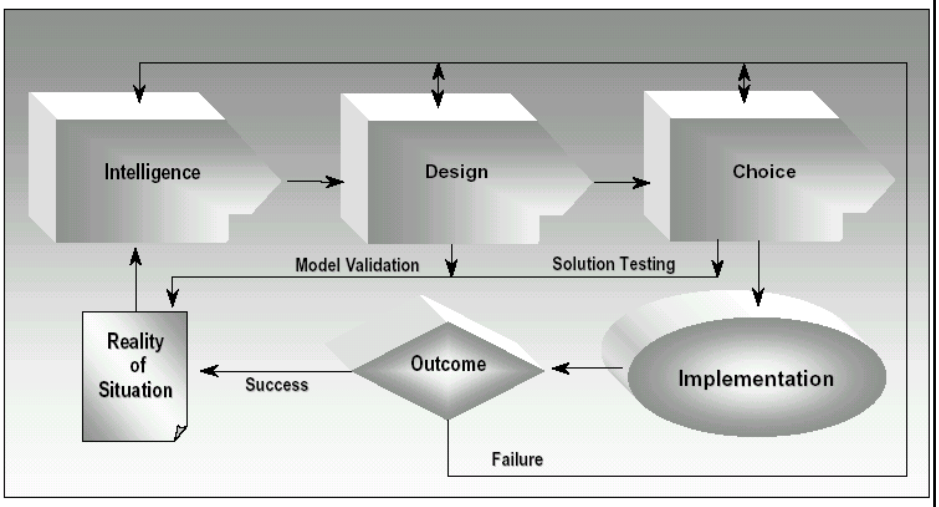
Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Processo de tomada de decisão

Enquadramento organizacional

Ciclo de tomada de decisão

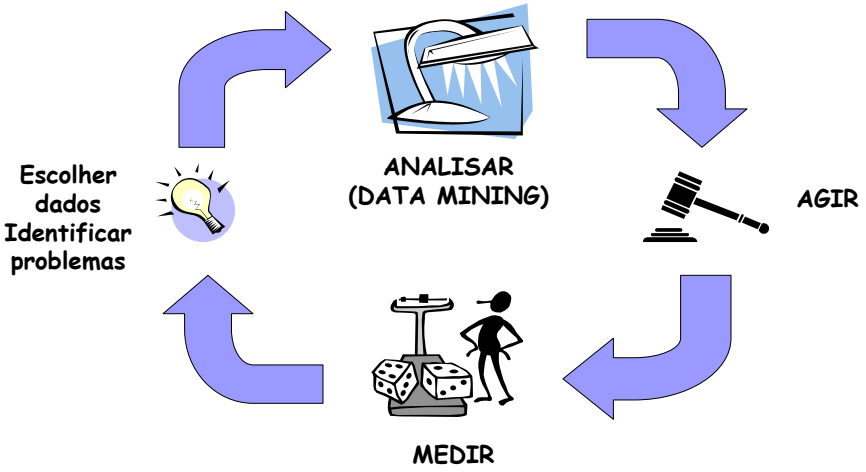


[Marakas 03]

Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Forma simplificada…



Tipos e níveis de decisão

	Nível da decisão		
Tipo de decisão	Controlo **Operacional**	Controlo de **Gestão**	Planeamento **Estratégico**
Estruturada (programada)	**Registo contabilístico, processamento de encomendas**	**Análise de orçamento, previsões de curto prazo, Relatórios**	**Investimentos, Localização de lojas e armazéns**
Semi-estruturada	**Escalonamento da produção, Controlo de inventário**	**Avaliação de crédito, preparação de orçamento, escalonamento de projectos, incentivos**	**Fusões e aquisições, planeamento novos produtos, planeamento de políticas**
Não estruturada (não programada)	**Aquisição de software, help desk, etc**	**Recrutamento, negociações, aquisição de máquinas**	**Planeamento de I&D, desenvolvimento de tecnologia, programas sociais**

Tecnologias para os diversos tipos de decisão

Tipo de decisão	Tipo de tecnologia
Estruturada (Programada)	MIS, Management Science Models, Transaction Processing
Semi-Estruturada	DSS, KMS, GSS, CRM, SCM
Não Estruturada (Não programada)	GSS, KMS, ES, NN, DT

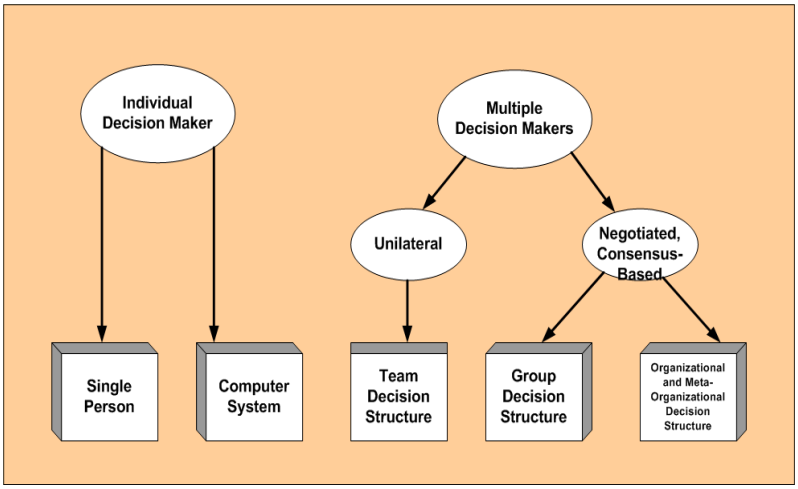
A envolvente organizacional

- Cultura organizacional
  - Afecta o processo de tomada de decisão
  - Afecta a utilização e enquadramento dos SAD
- Factores importantes
  - Estrutura da organização
    - Interacção entre actores
  - Estilos de liderança

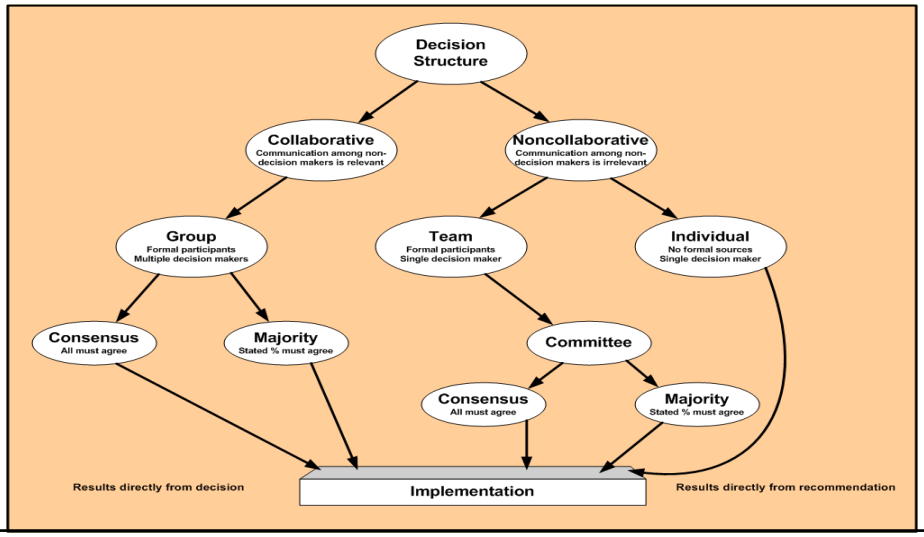
Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Diversos modelos para descrever a tomada de decisão-individual vs grupo



Diversos modelos para descrever a tomada de decisão-individual vs grupo



Neste contexto, quais as vantagens em partilhar o SAD ?

- Explora múltiplas perspectivas de uma decisão
- Gera alternativas múltiplas e de maior qualidade
- Explora múltiplas estratégias
- Facilita o brainstorming
- Fornece orientação e reduz possíveis desvios
- Aumenta a capacidade de lidar com problemas complexos
- Melhora o tempo de resposta
- Desencoraja a decisão prematura
- Permite controlar múltiplas fontes de dados

Tarefas típicas (1 a 7)

- 1 - **Organização** dos dados
  - Recolha, “limpeza”, normalização, armazenamento, dados heterogénios…
- 2 – **Visualização**
  - Apresentar os dados, compreendê-los, ter “insights” sobre os dados, explorá-los
- 3 – **Representação** de conhecimento e incerteza
  - Dados->Informação->Conhecimento, ser “mais ou menos”, ser “provável”, etc.

Tarefas típicas (1 a 7)

- 4 - **Previsão**
  - Estimadores estatísticos, regressões, redes neuronais, árvores de decisão, sistemas periciais, "case based reasoning"
- 5 – **Agrupamento**
  - Clustering, detectar "outliers", detectar grupos de interesse
- 6 – **Pesquisa de soluções**
  - Encontrar uma solução possível. Heurísticas de busca, simuladores, GA, etc
- 7 - **Optimização**
  - Encontrar a melhor solução possível. Técnicas de IO, heurísticas, GA, SA, etc

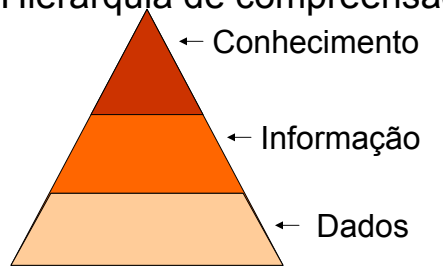
Organização dos dados

Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Informação é poder...

- "Água é vida"...
  - Todos os anos morre gente afogada...
- É necessário "trabalhar" a informação
- Hierarquia de compreensão e utilidade



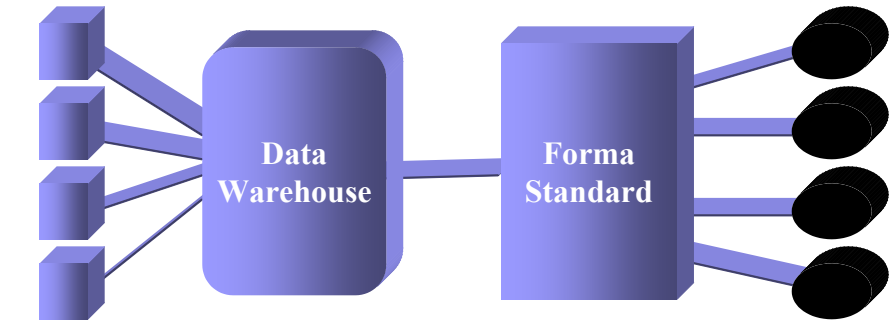
SI Operacional vs Analítico

- Sistema de Informação Operacional
  - Ligado directamente aos processos
  - Processamento em tempo real, contínuo
  - Muitos dados, pouco processamento
  - Constante mutação
- Sistema de Informação Analítico
  - Ligado aos decisores
  - Processamento "off-line", em tempo diferido
  - Muitos dados e MUITO processamento
  - Maior estabilidade

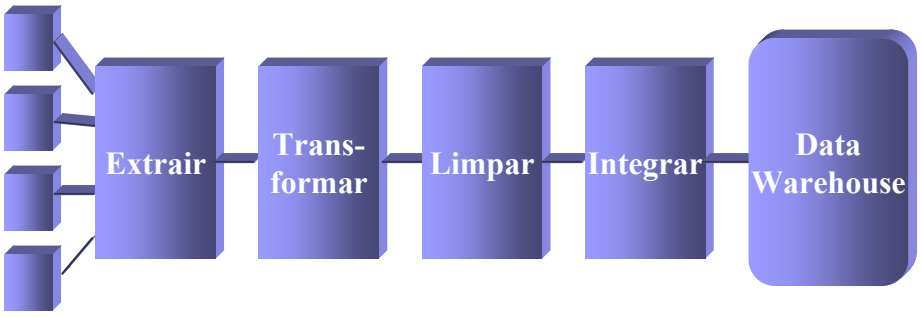
Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

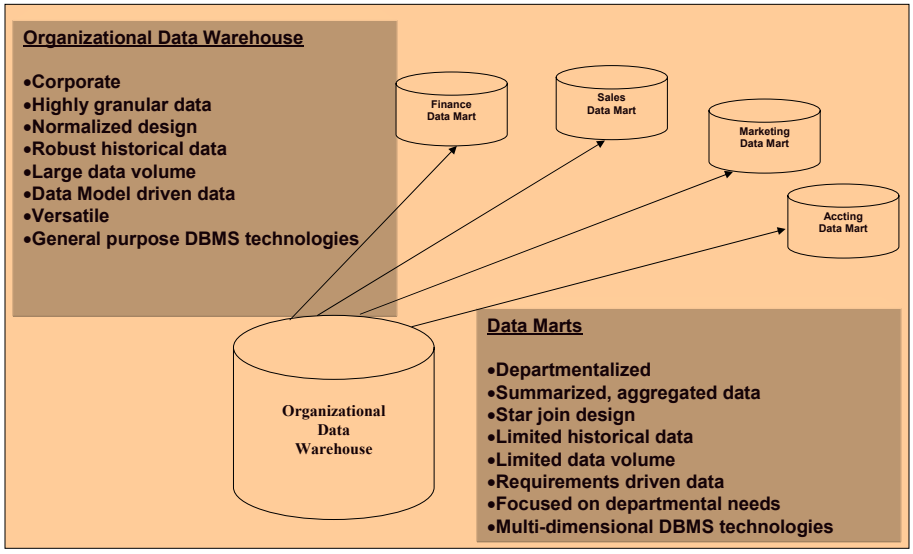
O modelo de "data warehouse"



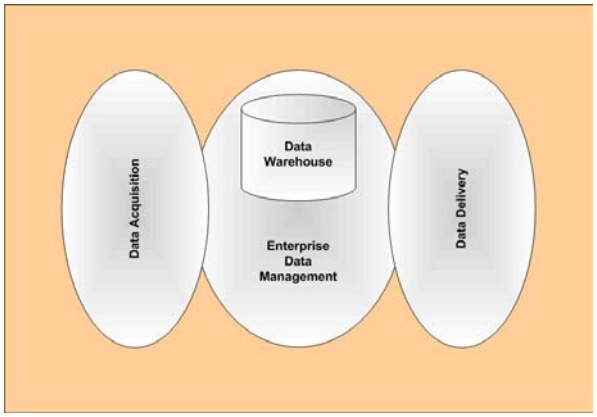
Passos para construir a "data warehouse"



Datawarehouse & data-marts



Outras perspectivas….



Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

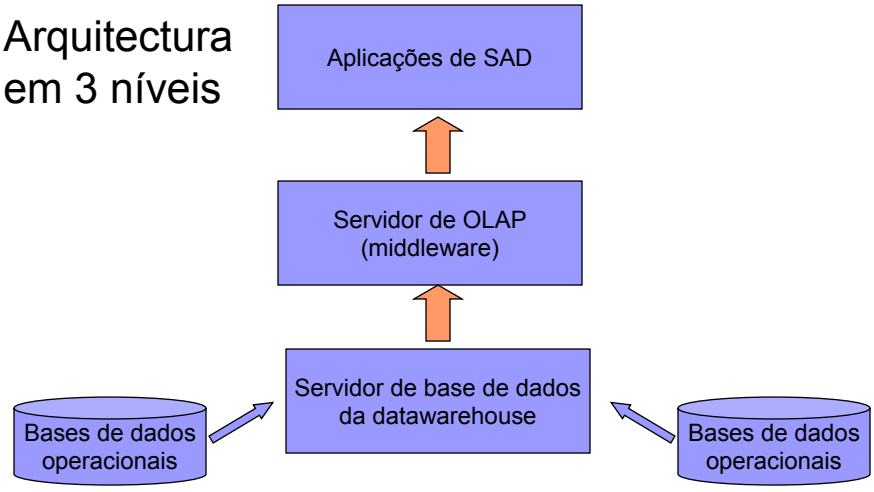
Medição, indicadores, visualização

- Relatórios “tradicionais”
  - Relatórios contabilísticos, tabelas de resultados
- Dashboards
  - Conceito de “tableau de bord”
  - Um (ou mais) números que indicam a “saúde” da empresa

- Scorecards
  - Metodologias para medir “o que é importante” num dado negócio
  - Técnicas para elaboração de “balanced scorecards”

Acesso à datawarehouse

- Arquitectura em 3 níveis

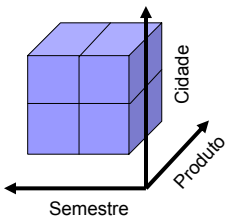


Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Sistemas de OLAP

- OLAP- On-Line Analytical Processing
  - Disponível para muitos sistemas de bases de dados
  - Conjunto de ferramentas de "reporting": fáceis e flexíveis
- Conceito de **hipercubo de dados**
  - Agrupar segundo **diversas dimensões**
    - Tempo, Local, Produto, Cliente, etc.
  - **Cortes** (slices) e **vistas**
    - Ver o hipercubo sob uma dada perspectiva
    - "Colapsar" (ou não) algumas dimensões
  - **Roll-up**:
    - Consolidar ou agregar em dados mais gerais
  - **Drill-down**:
    - Separar em nódulos mais específicos
  - Outras:
    - Ranking, Filtering, Dicing, estruturas ROLAP, HOLAP



Exemplo de um cubo de dados

- dados de vendas por semestre, por produto e por cidade:

Semestre	*Vendas*
Primeiro	16.000,00
Segundo	16.000,00

Produto	*Vendas*
Banana	16.000,00
Laranja	16.000,00

Cidade	*Vendas*
Lisboa	16.000,00
Porto	16.000,00

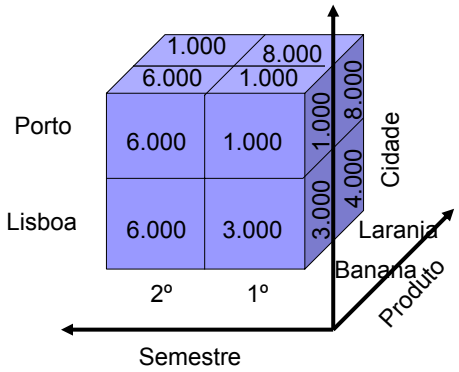
Exemplo de um cubo de dados

- Dados mais detalhados: numa tabela

Semestre	*Produto*	*Cidade*	*Valor*
Primeiro	Banana	Lisboa	3.000,00
		Porto	1.000,00
	Laranja	Lisboa	4.000,00
		Porto	8.000,00
Segundo	Banana	Lisboa	6.000,00
		Porto	6.000,00
	Laranja	Lisboa	3.000,00
		Porto	1.000,00

Exemplo de um cubo de dados

- Dados mais detalhados: num cubo



Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Bibliografia

- George Marakas, Modern Data Warehousing, Mining, and Visualization, Prentice-Hall 2003
- Barry Devlin, Data Warehouse – from Architecture to Implementation, Addison-Wesley, 1997

Tipos de dados e operações básicas

Dados numéricos

- Inteiros ou reais
- Precisão e gama dinâmica
  - Número de bits
  - Tipo de representação
    - Vírgula fixa, vírgula flutuante, números astronómicos
- Operações
  - Relações de ordem, operações aritméticas
- Exemplos
  - Temperaturas, nº de pessoas, etc
  - 34, 24.5, 20.4×10^{-15}, 32144152353, …
- Dados numéricos multidimensionais
  - Vectores numéricos

Dados numéricos

- Como comparar vectores numéricos ?
  - Distâncias $d(x,y)$
    - 3 condições formais:
      - $d(x,y) \geq 0$, $\forall x,y$, *e* $d(x,y) = 0, \Rightarrow x=y$
      - $d(x,y) = d(y,x)$, $\forall x,y$
      - $d(x,y) \leq d(x,z) + d(z,y)$, $\forall x,y,z$
    - Exemplos
      - Distância Euclideana

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2}$$

Distâncias entre vectores

- Distâncias de Minkowski de ordem *p*

$$d(x,y) = \left(\sum (x_i - y_i)^p\right)^{1/p}$$

  - *Ordem 1 – Distância de manhatten, ou "city block"*

$$d(x,y) = \sum |x_i - y_i|$$

  - *Ordem 2 – Distância Eulideana*
  - *Ordens mais altas*
    - *Dependem cada vez mais da componente mais diferente*
    - *Úteis para evitar "outliers"*

Distâncias entre vectores

- Distâncias ponderadas
  - Dão pesos diferentes a componentes diferentes

$$d(x,y) = \left(\sum \varphi_i (x_i - y_i)^p\right)^{1/p}$$

  - Se o factor de ponderação fôr a matriz de correlação e a ordem fôr 2, teremos a distância de Mahalanobis, ou distância euclideana normalizada

$$d(x,y) = \sqrt{(x-y)^T \Sigma^{-1} (y-x)}$$

- Produto interno
  - São uma medida de correlação entre os vectores
  - São a projecção de um vector sobre o outro

$$d(x,y) = \sum x_i y_i$$

Distâncias entre vectores

- Máxima correlação

$$d(x,y) = \max_k \sum x_i y_{i-k}$$

- Cosenos directores
  - É sensível à relações entre as componentes e não à sua magnitude

$$d(x,y) = \cos\theta = \frac{\sum x_i y_i}{\|x\| \times \|y\|}$$

- Outras
  - Menor diferença
  - Maior diferença
  - Tanimoto (aplicado a reais)

$$d(x,y) = \frac{\sum x_i y_i}{\|x\|^2 + \|y\|^2 - \sum x_i y_i}$$

Dados categóricos

- Booleanos
  - Só têm valor 0 ou 1
  - Exemplos
    - Tem a altura mínima, tem um curso, tem...
- Ordinais
  - Têm um número finito de valores
  - Os valores têm uma relação de ordem (mas não podem ser feitas operações aritméticas)
  - Exemplos
    - Escalões de vencimentos, Escalas de comportamento
    - Mau/Suficiente/Bom/Muito Bom, Alto/médio/baixo…
- Categóricos (puros)
  - Não têm relação de ordem
  - Exemplos
    - Naipes de cartas, raças,
    - Paus/Ouros/Espadas/Copas, Marinha/Administração Naval/Fuzileiros/…

Introdução a SAD

V 1.3, V.Lobo, EN/ISEGI, 2009

Distâncias entre vectores categóricos

- Distância de Hamming
  - Número de bits diferentes
  - Equivalente à distância de manhatten ou ao quadrado da distância euclideana
  - Exemplo
    - D(0010, 1010)=1, D(0010,1101)=4
- Distância de edição ou de Levenshtein
  - Número de alterações (apagar um valor ou acrescentar um valor)
  - Exemplo
    - D(ABC,AB)=1, D(ABC,AD)=3

Distâncias entre vectores categóricos

- Tabela de contingência entre valores dos vectores

		Object x		
		1	0	sum
Object y	1	***a***	***b***	a+b
	0	***c***	***d***	c+d
	sum	a+c	b+d	a+b+c+d

- Métricas:

Coefficients	Equation	Range
Simple Matching (Sokal and Michener 1958)	$\frac{a+d}{a+b+c+d}$	[0,1]
Russel and Rao (Russel and Rao 1940)	$\frac{a}{a+b+c+d}$	[0,1]
Rogers and Tanimoto (Rogers and Tanimoto 1960)	$\frac{a+d}{a+d+2(b+c)}$	[0,1]
Hamann (Hamann 1961)	$\frac{(a+d)-(b+c)}{a+b+c+d}$	[-1,1]
Ochiai II (Ochiai 1957)	$\frac{ad}{\sqrt{(a+d)(a+c)(d+b)(d+c)}}$	[0,1]
Sokal and Sneath (Sokal and Sneath 1963)	$\frac{2(a+d)}{2(a+d)+b+c}$	[0,1]

Coefficients	Equation	Range
Jaccard (Jaccard 1901)	$\frac{a}{a+b+c}$	[0,1]
Anderberg (Anderberg 1973)	$\frac{a}{a+2(b+c)}$	[0,1]
Czekanowsky / Sorensen-Dice (Dice 1945)	$\frac{2a}{2a+b+c}$	[0,1]
Kulczynski I (Kulczynski 1927)	$\frac{a}{b+c}$	[0,+∞]
Kulczynski II (Kulczynski 1927)	$\frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right)$	[0,1]
Ochiai (Ochiai 1957)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	[0,1]

Medidas de semelhança/dissemelhança

- Não obedecem às 3 condições das distâncias
  - Podem não ser simétricas
  - Podem ser o inverso de uma distância
  - Podem não respeitar a desigualdade triangular
- Exemplos
  - Algumas das métricas do acetato anterior
  - "Distância" de Kullback–Leibler

$$d(x,y) = \sum x_i \log \frac{x_i}{y_i}$$

Outros tipos de dados

- Conjuntos
  - Podem ser semelhantes a dados categóricos
    - Representados e manipulados como categóricos
  - Podem ser conjuntos de pontos
    - Representados como listas
    - Distância de Hausdorff
      - Maior das menores distâncias de um conjunto ao outro

$$d(x,y) = \max(\min_j d(x_i, y_j)$$

- Árvores ou outros grafos
- Mapas
- Etc,etc,etc…