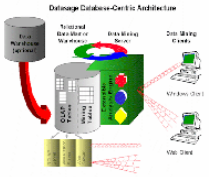
# Introdução à Análise de Dados e Tomada de Decisão







## Universo de Dados em Expansão

**Numa história chamada "Biblioteca de Babel" é descrita uma livraria infinita. Qualquer livro que possa ser imaginado (pode até não ter significado) existe algures na biblioteca. Formulam-se hipóteses: "algures deve haver um catálogo central", "a biblioteca deve ser uma estrutura sem fim que se repete indefinidamente". Nenhuma destas hipóteses pode ser verificada: a biblioteca contém uma quantidade infinita de dados, mas nenhuma informação.**

**A biblioteca pode ser comparada à situação em que as pessoas se encontram actualmente: temos um universo em expansão de dados, no qual há muitos dados e pouca informação.**

- **Assiste-se a um crescimento explosivo da capacidade de gerar e reunir dados, através de:**
  - sensores remotos, satélites, códigos de barras, cartões de crédito, etc.
- **Também muita informação aparece agora sob a forma não estruturada**

## Universo de Dados em Expansão

- **O desenvolvimento de novas técnicas para encontrar a informação necessária de entre enormes quantidades de dados é um dos maiores desafios de hoje.**
- **A quantidade de dados duplica todos os anos; paradoxalmente a quantidade de informação significativa diminui rapidamente.**
  - é cada vez mais difícil encontrar os factos significativos que procuramos
  - também, em grande medida o crescimento de informação é devido à produção de textos
- **A produção e reprodução mecânica de dados força-nos a adoptar estratégias e desenvolver métodos mecânicos de filtrar, seleccionar e interpretar dados.**





## Necessidade p/ a Análise de Dados (factores)

- **Alterações no ambiente de negócio**
  - mercado mais competitivo
    - conhecimento dos padrões de comportamento dos clientes
    - saturação de mercado
    - novos nichos de mercado
    - pesquisa de novos canais de mercado, dada a dificuldade de diferenciação
    - as abordagens tradicionais de marketing não são eficazes
    - "*time to market*"
    - ciclos de vida dos produtos mais curtos
    - aumento da competição e riscos de negócio

# Necessidade p/ a Análise de Dados (factores)

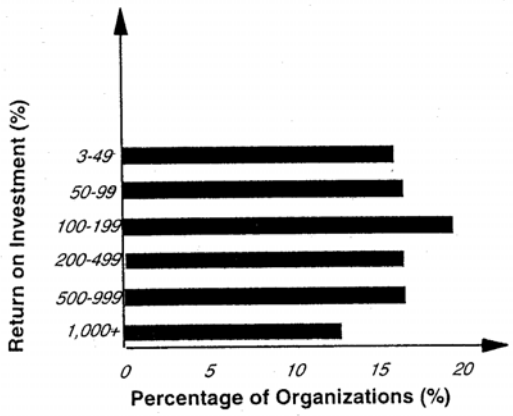
- **Reacção ao marketing de massas, tentativa da criação de loja do bairro**
  - **focagem nos clientes**
    - **que classes de clientes tenho?**
    - **como posso vender mais aos meus actuais clientes ?**
    - **há um padrão reconhecível nas aquisições?**
    - **que clientes são bons e quais são aqueles que me custam dinheiro?**
    - **posso prever que clientes poderão entrar em ruptura de pagamentos ou cometer fraudes?**
  - **focagem nos competidores**
    - **prever as estratégias ou planos de negócio dos meus principais competidores**
    - **previsão dos movimentos tácticos (ex. abertura de novas lojas ou novos serviços)**
    - **descoberta de sub-populações dos meus clientes que possam ser particularmente vulneráveis às ofertas da competição;**
  - **focagem nos dados**
    - **crescente evidência do retorno exponencial do investimento em estratégias de tomada de decisão, baseadas em técnicas conduzidas por dados (ex. Data warehousing, query, OLAP e data mining)**
    - **crescente disponibilidade de histórias de sucesso**





# Necessidade p/ a Análise de Dados (factores)

- **Retorno exponencial do investimento em estratégias de tomada de decisão**




*Retirado de "The Foundations of Wisdom: A Study of Finacial Impact of Data Warehousing"*

## Análise de Dados (como?)

- **Permitida a análise de dados pois que há:**
  - Manancial de dados (40 anos de tecnologia de informação)
  - Crescimento do Data Warehousing
    - disponibilização de bases de dados limpas e bem documentadas
  - Novas soluções em tecnologias de informação
    - soluções de armazenamento baratas, escaláveis
    - capacidade de processamento quase ilimitada, utilizando arquitecturas paralelas
  - Novas pesquisas em *machine learning*
    - rápida utilização comercial de algoritmos originados na comunidade científica
    - algoritmos melhores e escaláveis
    - *joint ventures* entre centros de pesquisa e empresas comerciais





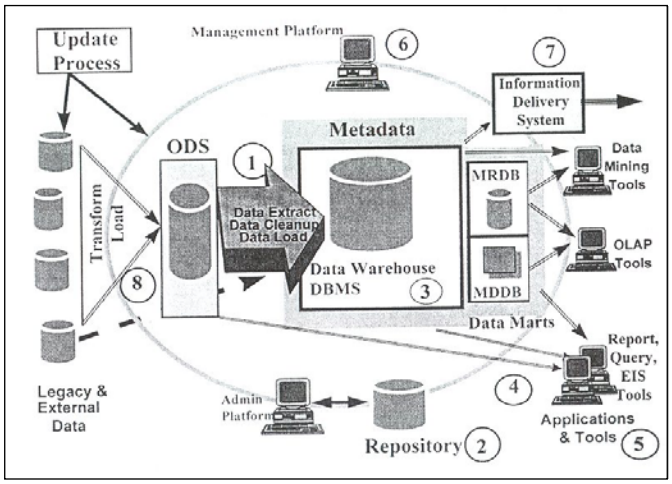
## A Informação como Factor de Produção

***A chave para o sucesso nos negócios é conhecer algo que mais ninguém sabe.***

***Aristotle Onassis***

- **Organizações com nível de excelência na extracção de informação e conhecimento, terão uma melhor hipótese de sobreviver. Devido a isto, a própria informação tornou-se um importante factor de produção.**

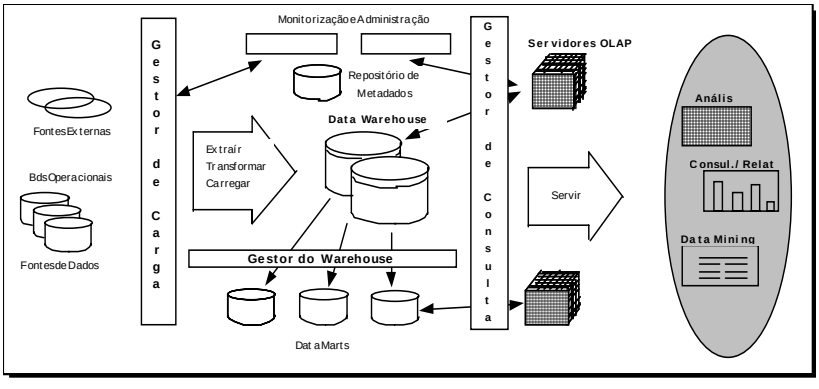
# Referencial Informacional e de Conhecimento








# Referencial Informacional e de Conhecimento



## SQL, OLAP e KDD

- **Com SQL podemos descobrir os dados rasos, ou seja informação que é facilmente acessível do conjunto de dados.**
- **Os dados multidimensionais podem ser analisados utilizando OLAP, mas é importante perceber que o que se consegue com OLAP poderia ser atingido com SQL, só que este tipo de ferramentas foi optimizado para a análise e pesquisa multidimensional.**
- **Quase 80% da informação interessante de uma base de dados pode ser extraída utilizando SQL. Os restantes 20%, constituem os dados escondidos e profundos.**

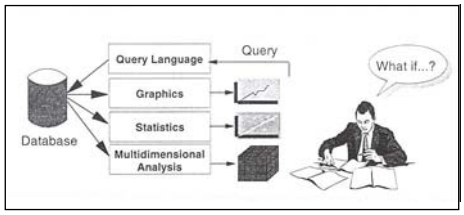




## SQL, OLAP e KDD

- **Dados escondidos e profundos (20 %)**
  - **requerem técnicas mais avançadas, no domínio do KDD**
  - **para organizações conduzidas por marketing são de importância vital**
  - **principais técnicas:**
    - **estatísticas**
    - **visualização**
    - **semelhança e distância**
    - **árvores de decisão e regras de associação**
    - **redes neuronais e algoritmos genéticos**

# SQL e OLAP




- **Exploração de dados**
  - compreende:
    - análise de dados tradicional (linguagem query, gráficos, estatísticas )
    - análise multidimensional
  - disponibiliza representações dos dados adequadas à obtenção de informação
  - com vista
    - à extracção de informação para apoio à decisão
  - necessária formulação prévia de hipóteses
    - deve saber-se previamente o que procurar





# KDD

- **Extracção de Conhecimento em Bases de Dados (Knowledge Discovery in Databases - KDD)**
  - compreende
    - técnicas e ferramentas para a análise inteligente e automática de bases de dados
  - com vista à
    - obtenção de conhecimento não óbvio e de valor para o negócio a partir de grandes bases de dados
  - descoberta de informação sem formulação prévia de hipóteses
    - não é necessário conhecer-se previamente o que procurar
    - natureza exploratória
- **Minagem dos Dados (data mining)**
  - Algoritmos de detecção de padrões nos dados (constitui um das fases da extracção de conhecimento em BD)

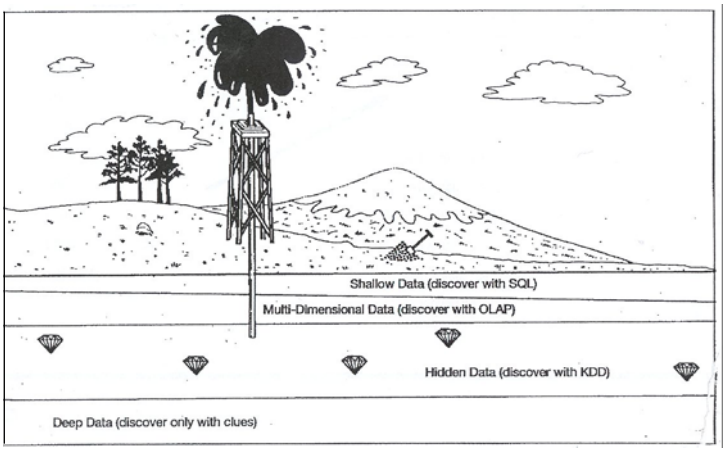
## SQL / OLAP x KDD

- **Em resumo, pode dizer-se que:**
  - **se for sabido exactamente do que estamos à procura,**
    - **utilizar SQL;**
    - **para dados multidimensionais, o OLAP;**
  - **mas se apenas soubermos vagamente o que procuramos, utilizar data mining.**
- **A abordagem inicial é normalmente vaga, não se sabendo exactamente o que se pretende. Daí que:**
  - **seja enorme a motivação e o renovado interesse no data mining;**
  - **o paradigma do mundo de informação na mão, indicado na figura, torna-se, com estas ferramentas, mais próximo.**

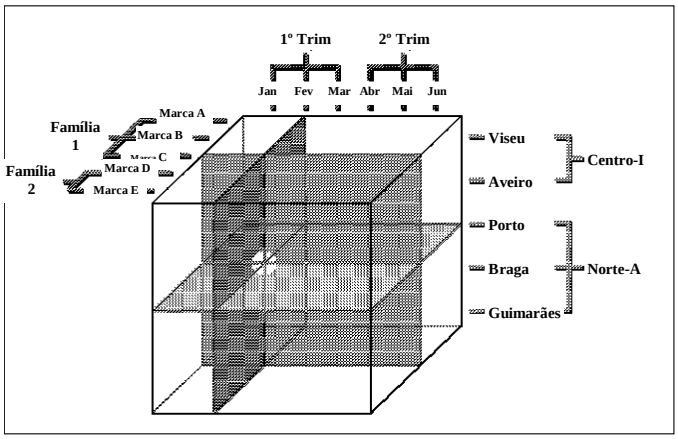




## SQL, OLAP e KDD



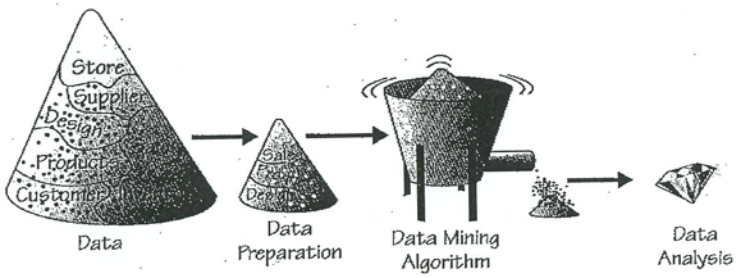
# OLAP







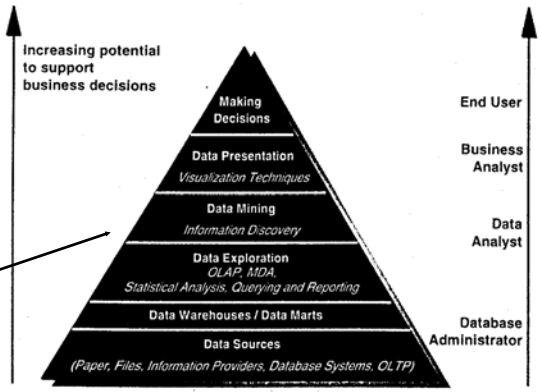

# Descoberta de Conhecimento em bases de dados



# Business Intelligence

- **Business Intelligence - Termo global para todos os processos, técnicas e ferramentas que suportam a tomada de decisões de negócio baseadas em tecnologias de informação.**
- **Pode ir desde:**
  - **uma simples folha de cálculo**
  - **queries**
  - **OLAP**
  - **data mining**
  - **visualização**

**O valor da informação para a tomada de decisões aumenta do fundo para o topo**




Retirado de Discovering Data Mining, PH, 1998





# Conceitos para ECBD

- **Padrões e Modelos**
  - **O que é um padrão?**
  - **O que é um modelo?**
  - **Onde se utilizam os modelos?**
  - **O que é um modelo correcto?**
  - **Preditores e predição.**
  - **Amostragem.**

## Padrões e Modelos

**Padrão:**

- **Expressão E numa linguagem L, descrevendo factos num subconjunto $F_E$ pertencente a F.**
- **E constitui um padrão se constitui uma descrição mais simples do que a enumeração de todos os factos em $F_E$**
  - **A aprendizagem pode ser descrita, na maioria dos casos, de um ponto de vista matemático, como a compressão de conjuntos de dados.**
  - **Se tivermos um algoritmo que crie uma descrição dum conjunto de dados que seja efectivamente menor do que o conjunto de dados original, pode dizer-se que se aprendeu algo.**

**Definição de Padrão centrado em BD e DW:**

- **Evento ou combinação de eventos numa base de dados que ocorrem mais vezes do que seria de esperar.**
- **Tipicamente isto quer dizer que a ocorrência actual é significativamente diferente da que seria de esperar aleatoriamente.**





## Padrões e Modelos

**Problema: determinar o próximo número da seguinte sequência:**

**1212121 ....? R: 2**

**Fácil: o padrão 12 é encontrado vezes suficientes para haver confiança de que existe um modelo predictivo que diz: "Se 1, então 2 seguir-lhe-à;**

**Também: Se 2, então 1 seguir-lhe-á"**

**Mas, pode ser mais complicado:**

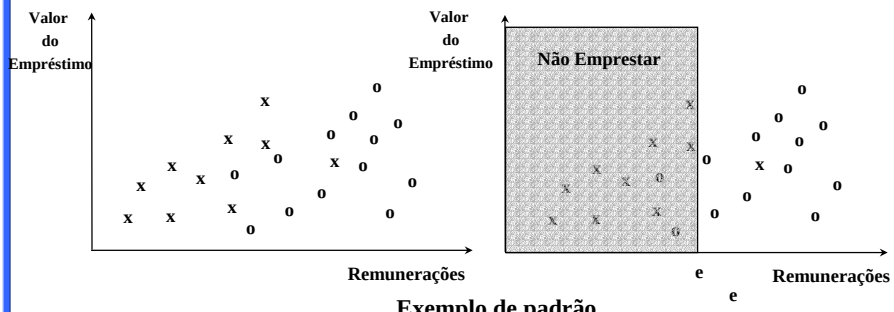
**Se o conjunto for 121?**

**Se for 1212123121212?**

**Um modelo representa alguma característica importante da coisa maior que está a ser modelada, não a descreve completamente.**

**Para aplicações de negócio, um modelo pode ser algo como uma equação matemática, um conjunto de regras que descrevem segmentos de clientes, representações computacionais duma arquitectura de redes neuronais.**

# Padrões e Modelos




**Exemplo de padrão**

**Dados: Conjunto F de factos. No caso da figura, é a colecção de casos, cada um com 3 campos - valor do empréstimo, remuneração e situação do empréstimo (o - normal, x - em falta).**

**Neste caso o padrão será "se a remuneração do cliente é < e euros, então este não cumpre as condições do contrato de crédito".**

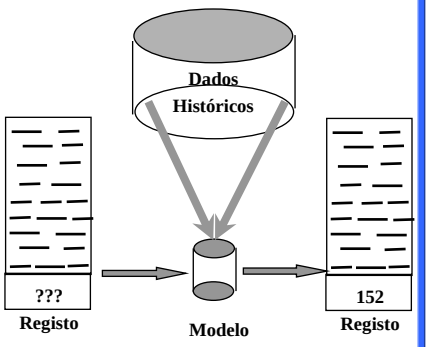




# Padrões e Modelos

## Definição de Modelo centrado em BD e DW:

- **Descrição da base de dados histórica original a partir da qual foi construído, que pode ser aplicado com sucesso a novos dados, por forma a fazer predições acerca de valores em falta ou fazer declarações acerca dos valores esperados.**




**Visão de alto nível do processo de modelação**

## Padrões e Modelos

**Diferença entre padrões e modelos:**

- **Os padrões são conduzidos pelos dados e geralmente reflectem os próprios dados.**
- **O modelo geralmente reflecte um propósito e pode não ser necessariamente conduzido pelos dados, sendo uma descrição de alto nível.**
  - **Ex. um modelo do mundo físico utilizando as equações da física Newtoniana, poderá explicar a rapidez da queda de qualquer objecto ou ainda o quanto poderá voar.**
- **Os modelos são mais complexos do que os padrões, usualmente há muitos. Um modelo contém, em regra, muitos padrões.**
  - **Ex. o modelo de comportamento de clientes pode ser muito complexo e conter centenas de padrões que foram encontrados na base de dados.**





## Padrões e Modelos

**Onde são utilizados os modelos?**

- **Vejamos alguns exemplos de problemas de negócio que poderão beneficiar da existência de modelos:**
  - **Selecção**
  - **Aquisição**
  - **Retenção**
  - **Extensão.**

## Utilização de Modelos

- **Selecção - o negócio tenta seleccionar novos clientes**
  - **a organização tem uma lista de possíveis candidatos a clientes, mas não sabe quais serão os desejáveis; é necessário concentrar-se nos clientes que se tornarão bons clientes;**
  - **a lista pode ser adquirida a partir de várias fontes: lista de endereços, endereços para cupons, base de dados de censos ou aleatoriamente a partir da lista telefónica;**
  - **há informação limitada acerca dos clientes, é um desafio à construção de um modelo de predição;**
  - **há que recorrer à informação histórica da própria base de dados e, a partir desta, detectar os padrões e construir um modelo dos hábitos dos próprios clientes; depois extrapolar para os possíveis novos clientes.**
    - **Ex. no sector das telecomunicações, seleccionar possíveis novos bons clientes, para chamadas de longa distância. Criar um modelo da rentabilidade dos próprios clientes, dados um conjunto de campos que a determinam (chamados predictores ou variáveis independentes), a partir dos próprios dados históricos. Utilizar este modelo com seus padrões, para avaliação dos possíveis futuros clientes.**





## Utilização de Modelos

- **Aquisição - depois de seleccionados os clientes, há que efectivá-los.**
  - **normalmente efectuada através de alguma oferta ou produto em que o cliente poderá estar interessado (desconto, simplificação de facturação, amostra de produto, etc.);**
  - **há que notar que nem todos os clientes seleccionados terão o mesmo perfil, assim, para os mais lucrativos poderão ser utilizadas estratégias mais caras, ao contrário de outros;**
  - **o desafio é modelar a táctica que resulte em esforço mínimo (e despesa) mas que resulte no sim à oferta;**
  - **o modelo poderá ser do tipo: probabilidade de aquisição x táctica x atributos de cliente; o cliente terá 90% de probabilidade de ser conquistado se lhe for oferecida um bónus de 50€, 60%, se tiver um desconto de 10% e de 1%, caso lhe seja simplesmente enviados prospectos por correio;**
  - **estes modelos podem ser baseados em experiências passadas de outras ofertas feitas.**

## Utilização de Modelos

- **Retenção - reter os clientes que foram conquistados**

**Dada a competitividade do mercado actual, é grande a facilidade com que um competidor pode contactar e roubar um cliente; a lealdade do cliente é algo que deve ser activamente encorajada e seguida.**

- **no mercado bancário e de comunicações móveis, quase 1 em cada 3 clientes são perdidos para os competidores, em cada ano, sendo perdidos normalmente os mais lucrativos; o custo de aquisição é, normalmente bastante alto;**
- **ter um modelo dos clientes que estamos em risco de perder será de grande valia: corrigir os motivos de insatisfação do cliente ou adiantar-se numa oferta antes do contacto dum competidor, é muito mais eficaz do que reagir, já depois do cliente ter decidido por outro fornecedor;**
- **este modelo terá duas partes:**
  - **um modelo para saber quais os clientes em risco**
  - **outro modelo para determinar que estratégia de retenção será a mais eficaz.**
- **mais uma vez, há que recorrer à informação histórica da própria base de dados, saber quais os clientes descontentes e que estratégias foram bem sucedidas na sua retenção.**





## Utilização de Modelos

- **Extensão - estender os serviços ou produtos que se vendem aos clientes, para além dos originais.**
  - **nesta fase, tal como na anterior, já temos disponível muitos dados acerca do cliente - facto que não se verificava nas duas primeiras fases - e que os nossos competidores não têm (eles poderão estar na fase 1 e 2). Com esta informação, e desde que utilizada eficazmente, estaremos em vantagem perante os nossos concorrentes;**
  - **a extensão é também denominada, em terminologia anglo-saxónica, cross-selling;**
  - **exemplo: um banco, vender a um cliente que contraiu um empréstimo para habitação, um seguro de vida;**
  - **a modelação dos clientes que poderão estar interessados em outros produtos é importante, pois que o cliente pode facilmente ser inundado com ofertas de produtos, para os quais não tenha o mínimo interesse e não responder também a outras ofertas der produtos que até desejaria; claro que se não for sugerido ao cliente algo de que este necessite, deixaremos a porta aberta aos nossos competidores.**

## O que é um Modelo Correcto?

"Se for possível conhecer precisamente o estado actual de tudo no universo num dado momento, será então possível criar um modelo que prediga rigorosamente todos os eventos futuros"

Pierre Laplace

A afirmação acima foi já temperada pela mecânica quântica, princípio de incerteza de Heisenberg e os desenvolvimentos mais recentes relativos aos sistemas caóticos. Apesar disso, muitos perseguem a ideia que debaixo de toda a complexidade de eventos que ocorrem, muitas vezes mesmo sem significado, pode existir um modelo bem definido que, sendo descoberto, explicaria e poderia predizer muito do que é observado. Sem mais filosofias, poderemos focar-nos, dada a relevância e acuidade para o nosso estudo, em questões como:

- existe um "modelo perfeito"?
- pode um modelo ser melhor do que outro?
- como poderemos avaliar que modelo será melhor?





## Modelo Perfeito

- O modelo perfeito, se é que tal coisa existe, deveria ter várias características importantes:
  - poderia ser sempre utilizado para fazer as previsões correctas;
  - não se degradaria com o tempo;
  - poderia ser utilizado com os dados mais à mão, não requerendo um volume de dados extraordinário;
  - deveria ser mais simples e pequeno do que os dados utilizados para a modelação.
- Não há modelo perfeito:

⇨ no mundo real: há sempre dados relevantes que não puderam ser recolhidos, ou os dados contêm erros ou valores em falta e quase todos os modelos construídos são susceptíveis de alterar-se com o tempo.

## Dados em Falta

- **Um dos maiores problemas na recolha dos dados do mundo real, a partir dos quais construir modelos predictivos, é não ter os dados certos, na quantidade devida.**
  - **pessoas com dieta pobre em gorduras desde a nascença**
    ⇨ **taxa muito baixa de ataques do coração**
    **Mas ...? Como conseguir estes dados? (no exemplo, nos USA?)**
  - **Solução: recolha pró-activa dos dados em campanhas de marketing de teste**
  - **o volume de informação disponível pode ser insuficiente para a construção de um bom modelo**
    - **ex. predizer o nome de alguém, sabendo o seu número de telefone?**
- **Recordar: não importa quão grande seja a base de dados ou o esforço feito, sucederá sempre faltar algo que poderia melhorar o desempenho do modelo que se está a construir. Podem faltar predictores, outras vezes faltam registos. O segredo está no reconhecimento de que no mundo real o modelo será baseado sempre em dados em falta e fazer compensações para isso.**





## Registos, Predictores e Predição em Modelação Predictiva

- **Registo - Estrutura de dados ao nível atómico que suporta os dados pertinentes aos indivíduos na base de dados. Um registo corresponde a uma linha de uma tabela numa base de dados desnormalizada. Cada registo é feito de valores para cada campo que contém, incluindo os campos predictores e o campo predição.**
- **Variável Independente, Campo Predictor ou de Entrada - Campo que pode ser utilizado para construir um modelo de predição. Alguma função dos valores do predictor do registo produzem o valor de predição para esse registo.**
  - **Geralmente, chamamos predictores aos campos quando são utilizados para exploração ou predição.**
- **Variável Dependente, Campo Predição, de alvo ou de saída - É o campo que contém o resultado conhecido, passado à técnica de Data Mining para que o modelo seja construído - o valor que eventualmente esperamos predizer. Geralmente, trata-se dum campo semelhante a qualquer outro, excepto na forma como é manejado pelo processo de criação do modelo de predição.**

## Correspondência Entre Vários Conceitos de Modelação Predictiva

| Data Mining | BD Relacionais | OLAP | Estatística | I.A. |
|---|---|---|---|---|
| Data Set (conjunto de dados) | Tabela, base de dados | BD multidimensional, cubo | Data set, amostra | Data set, conjunto de treino |
| Registo | Linha | ND | Registo, datum | Exemplo, registo |
| Campo | Coluna | Variável, dimensão, medida | Variável | Campo, característica, dimensão |
| Predictor, variável independente | Coluna | Variável, dimensão, medida | Variável independente | Campo, característica, dimensão |
| Predição, variável dependente | Coluna | Variável, dimensão, medida | Variável dependente | Alvo de classificação |





## Tipos de Predictores

- Em qualquer base de dados há diversos tipos de colunas (aqui chamadas variáveis dependentes ou independentes e predictores). Os tipos de dados são os permitidos pela base de dados e, usualmente, podem ser uma dezena ou mais. No entanto, para as técnicas de data mining, teremos, como em sinal, dois grandes tipos: contínuos e descontínuos (aqui denominados de categóricos).
- Dependendo do tipo das variáveis, especialmente da possibilidade de ordenação, poderemos aplicar ou não alguns algoritmos de data mining.
- O tipo de predictor pode ter também um impacto importante na forma de como pré-processar os dados.
  - Ex. informação extra acerca da distância entre dois valores seguidos de um predictor poderá conduzir a melhores modelos predictivos do que se a distância ou ordem não for possível.

## Tipos de Predictores

- **Contínuos** - **hipoteticamente, podem ter um número infinito de valores ou categorias. A idade de uma pessoa pode ser medida em dias, horas, segundos, milissegundos, microssegundos, ...**
- **Categóricos** - **podem ter um número finito de valores ou categorias.**
  - **Nominais (deriva de nome) - cujos valores ou categorias que não têm qualquer relacionamento particular uns com os outros (ex. cores de um sapato). Não se pode estabelecer qualquer ordem.**
  - **Ordinais - podem ser ordenados, como o nome indica. Ex. sapatos de criança, adolescente e de adulto. Não permitem, de qualquer modo, saber quão maior ou menor é a ordem relativa.**
  - **Intervalo - tem aqui sentido uma distância numérica entre valores. Ex. Sapato de tamanho 42 é 2 números acima do 40.**
- **Hierarquia:** **Contínuos, intervalo, ordinais e nominais.**
- **É também possível, e muitas vezes desejável ou obrigatório, converter os tipos de predictores.**





## Amostragem

**"os padrões existentes nos dados que procuramos podem provavelmente ser reconhecidos sem ter de olhar para todos exemplos de cada uma das combinações possíveis de predictores"**

- **Independentemente do tamanho da base de dados, não encontraremos decerto um exemplo de cada possível cliente e de todos os predictores que descrevam essa pessoa.**
  - **Ex. classificar, do ponto de vista de uma companhia de cartões de crédito, os clientes de risco. Avaliaremos os rendimentos das pessoas x limites de crédito. Se tivermos limites de crédito entre 100 € e 10000 € e rendimentos de clientes entre 5000 € e 500000 €, se utilizaremos valores de 100 € como incremento, teremos 100 * 500 possibilidades, ou seja 50000 tipos diferentes de clientes. Mesmo com uma grande base de dados, já seria muito difícil que tivéssemos pelos menos 1 registo de cada tipo, mas estamos a considerar apenas 2 predictores. Necessariamente teremos de adicionar mais. Consequentemente o nº de possibilidades aumentará enormemente. Aliás, se tivéssemos todas as combinações possíveis, não necessitaríamos de modelo: bastaria procurar o caso adequado e analisá-lo. Mas aí, poderíamos defrontar-nos com outro problema...**
- **Outras vezes, não podemos utilizar todos os dados disponíveis por ser difícil processá-los ou armazená-los: temos de efectuar uma amostragem. Mesmo com uma pequena amostra é, muitas vezes, e surpreendemente, possível extrair um padrão.**

## Problemas com a Amostragem: Polarização

- Quando se faz a amostragem, é importante reconhecer certas diferenças na forma como a mostra pode ser obtida, verificar se a amostra é feita num processo verdadeiramente aleatório.
  - Muitas vezes a polarização está presente na maneira como a amostra é colhida. Há que verificar qual o universo de amostragem e de análise. É o clássico problema das sondagens eleitorais... Contactam-se pessoas com telefone e... temos aí já a polarização (tendência) do processo: nem todos os eleitores terão telefone.
  - Há, claro, a tendência de minorar o trabalho ou custos do processo. Mas, deveremos ter em mente que na qualidade dos dados a analisar derivará directamente a qualidade do modelo e padrões obtidos.
- Como efectuar então a amostragem?





## Técnicas para a Amostragem

- Round Robin - Forma mais simples de amostragem: buscar todos os n-ésimos registos da base de dados.
  - Problema: a selecção da amostra dependerá da forma como os dados residem na base de dados. A amostra pode ser polarizada se houver um padrão na forma como os dados estão armazenados consecutivamente na base de dados.
    - Imaginemos que, num sistema MPP, os dados são distribuídos de acordo com um determinado padrão, por forma a balancear a carga pelos diversos nós. Neste caso, a nossa amostra poderá ser polarizada.
- Amostragem Estratificada - Em casos em que temos um valor de uma coluna utilizada na predição, que é particularmente importante.
  - Ex. numa campanha de mailing: o predictor relativo ao resultado do contacto. Normalmente teremos valores usuais de 1% ou menos. Se a amostra for perfeitamente aleatória, teremos um número de registos com resposta positiva muito pequeno, tornando difícil extrair padrões relativos a características dos clientes que responderam positivamente.
  - Melhor efectuar uma amostra com número de registos sensivelmente idêntico. Depois de construído o modelo, há que corrigi-lo para as concentrações originais.

## Técnicas para a Amostragem

- **Amostragem Grupo (Cluster) - Para assegurar que todos os subgrupos importantes na base de dados são representados.**
  - **A base de dados original é dividida em grupos e um número equivalente de registos de cada grupo é retirado.**
- **Ex. com clientes, poderemos dividi-los em grupos por qualidades sócio-económicas similares. Depois bastará seleccionar alguns registos de cada um dos grupos para termos a certeza de que todos os grupos mais importantes estarão representados no modelo.**
- **Normalmente, a amostragem aleatória é adequada, se a amostra e a base de dados for suficientemente grande. Em casos em que alguns subgrupos importantes tiverem poucos registos, será necessário utilizar esta abordagem para termos a certeza de que haverá uma cobertura adequada dos grupos na amostra.**





## Mais Alguns Conceitos Estatísticos Importantes

- **Aprendizagem e Conteúdo Informacional**
- **Probabilidade**
- **Independência**
- **Causalidade e colinearidade**
- **Teste do Qui-Quadrado**

## Aprendizagem como Compressão de Conjuntos de Dados

*"Na maioria dos casos, a aprendizagem pode ser descrita, de um ponto de vista matemático, como a compressão de conjunto de dados."*

Retirado de Data Mining, Pieter Adrians and Dolf Zantinge

**Se um algoritmo cria uma descrição do conjunto de dados que é efectivamente menor do que os dados originais, podemos dizer que se aprendeu alguma coisa.**

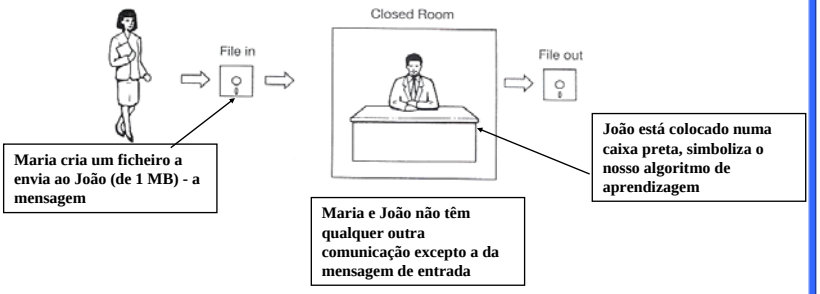
**Há uma relação entre a complexidade dos dados e a capacidade de aprendizagem:**

- **em geral, conjuntos de dados complexos são difíceis de comprimir e assim de se perceberem**
- **conjuntos de dados pouco complexos, podem ser facilmente comprimidos e aprendidos**
- **mas nem todos os dados compressíveis são de fácil aprendizagem (dados encriptados)**





## Exemplo Ilustrativo (Aprendizagem x Compressão)




**Maria envia uma mensagem ao João, que este tem de descodificar**

- **quando poderemos dizer que João encontrou a chave certa para descodificar a mensagem?**
- **como poderemos saber que João encontrou algo de real valor na mensagem da Maria?**

## Exemplo Ilustrativo(Aprendizagem x Compressão)

**Mensagem 1: mensagem aleatória ou não estruturada**

- **João não consegue criar nenhuma descrição curta para a mensagem, não há compressão possível, dado que não encontra quaisquer padrões: qualquer descrição da mensagem deve conter a própria mensagem**

**Mensagem2: mensagem altamente estruturada, com padrões simples de encontrar (tudo 0's ou tudo 1's)**

- **Descrição curta "um megabyte de 1's" - altamente comprimida**

**Mensagem 3: mensagem altamente estruturada nas difícil de decifrar (ex. 3.141592653 ....)**

- **Descrição curta, mas um padrão escondido muito profundamente: valor de pi.**

**Mensagem 4: conjunto de dados parcialmente estruturados - não são completamente aleatórios, mas também não são muito compressíveis - a maioria dos conjuntos de dados que encontraremos em DMining são desta natureza.**

- **Pode encontrar algumas regularidades que possam ser algo compressíveis.**





## Conclusões (Aprendizagem x Compressão)

**Compressibilidade e Capacidade de aprendizagem têm algo em comum**

- **Em geral um objecto terá baixa complexidade se houver um programa muito simples que o produza num computador (lembremo-nos do megabyte de 1's)**

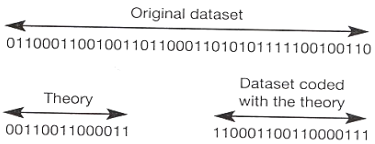
**A complexidade deste tipo é traduzida matematicamente pela denominada complexidade Kolmorogov, que pode ser enunciada, do nosso ponto de vista como:**

***"A complexidade Kolmorogov de um objecto é o tamanho em bits do mais pequeno programa que produza esse objecto num computador"***

## Conclusões (Aprendizagem x Compressão)

**O relacionamento entre complexidade e capacidade de aprendizagem é formulado pelo princípio de Rissanen, chamado de princípio da descrição de tamanho mínimo:**

***"A melhor teoria para explicar um conjunto de dados é aquela que minimize a soma do comprimento, em bits, da descrição da teoria, mais o tamanho, em bits, dos dados, quando codificados com a ajuda da teoria"***




**Por outras palavras: *"se for encontrada alguma regularidade num conjunto de dados e a descrição dessa regularidade em conjunto com a descrição das excepções for ainda menor do que o conjunto de dados original, então encontrou-se algo de valor"***





## Conteúdo Informacional de uma Mensagem

**A Maria envia uma mensagem ao João.**

**Quando é que a mensagem terá alguma informação para o João?**

**Suponhamos que João perguntou à Maria se quer ir às compras à tardinha:**

**1. se o João souber antecipadamente que a Maria quer ir às compras, o sim não conterá qualquer nova informação**

**2. se João não souber da apetência da Maria de ir às compras, o sim ou não já conterão informação real**

**Mas agora se a questão for: "O que queres fazer à noite?"**

**aqui o leque de respostas será maior**

- **poderá querer ir às compras (resposta que conterá pouca informação, pois que é algo que o João quase sabe já)**
- **mas poderá ser "quero ir numa expedição aos Himalaias" (aqui teremos algo inesperado - conterá muito mais informação)**

## Conteúdo Informacional de uma Mensagem

**Bits necessários para contar o número de mensagens possíveis:**

- **A resposta sim / não - pode ser codificada em 1 bit**
- **Uma resposta com 256 valores possíveis necessitará de 8 bits**
- **para n valores possíveis teremos bits = $\log_2 n$**

**A intuição anterior - o valor informacional da mensagem depende inversamente da sua probabilidade**

**e o número de bits necessários para codificar a mensagem, levam-nos à noção de conteúdo informacional de Shannon:**

***"Se tivermos n mensagens cada uma das quais com igual probabilidade de ocorrência, cada mensagem terá probabilidade 1/n de ocorrer, então o conteúdo informacional de cada mensagem será $\log_2 1/n = -\log_2 n$"***





## Ruído e Redundância

**O ruído pode criar problemas nas operações de DMining.**

**Nos conjuntos de dados o ruído manifesta-se por:**

- **erros em valores de campos**
- **falta de valores em campos**
- **inconsistências**
- **transformações indevidas**

**Há algo de positivo no ruído, pode conter informação: se há muito ruído numa dada base de dados, há que concluir que deverá haver razões para ele estar lá. Pode limpar-se, mas isso não resolve o problema base: a forma como a organização está a lidar com a informação - como as aplicações estão construídas e como são utilizadas. O ruído indica que há que alterar a forma como se está a trabalhar com os sistemas de informação nesta organização.**

# Probabilidade

**Conceito crítico em estatística e em todas as técnicas de data mining.**

- **Apesar de familiar, não deve ser depreciado, pois através dele é possível efectuar predições e detectar padrões.**

- **Probabilidade a priori - Aquela que existe antes de qualquer informação ser conhecida.**
  - **Ex. Para predizer a cotação de uma acção no dia seguinte, iremos dar como valor mais provável o de fecho do dia anterior.**

- **Probabilidade Condicional - Temos mais informação disponível. Desta forma, podem ser colocadas condições para o evento que alterarão a probabilidade deste ocorrer.**
  - **Ex. Poderemos ter uma probabilidade a priori de 1 / 1 000 000 de ocorrer uma transacção fraudulenta com cartão de crédito. Mas, se colocarmos a condição de só olharmos para transacções de equipamento electrónico (com alto valor e fácil revenda), encontraremos uma taxa 10 vezes maior de transacções fraudulentas, ou seja de 1 / 100 000.**





# Independência

**Em estatística dois eventos são considerados <u>independentes</u> um do outro se a probabilidade de ambos ocorrerem for igual à probabilidade de um multiplicada pela probabilidade do outro.**

**Ex. pessoa com camisa e gravata de duas cores e insensível às cores. Haverá 50% de probabilidade de vestir cada cor de camisa ou gravata. Tb. haverá 25 % de probabilidade de vestir qualquer combinação de cores de camisa / gravata.**

**Se os fenómenos não forem independentes, quer dizer que há <u>relacionamento</u> entre predictores, podendo haver relações de <u>causalidade</u> ou <u>colinearidade</u>.**

**Ex. se a pessoa anterior do exemplo acima, já for sensível às cores e tiver sentido de moda.**

## Relacionamentos

- **Causalidade - A ocorrência de um fenómeno causa o outro.**
  - **importante pois que se trata de um relacionamento mais previsível ao longo do tempo, em locais diferentes e sob uma variedade de condições diferentes.**
  - **No exemplo anterior, a escolha de uma cor de camisa causa a escolha da cor da gravata.**
- **Colinearidade - Efeito no qual um predictor parece andar de mãos dadas com outro, mas não é realmente a causa.**
  - **No nosso país, o advento do frio parece não ser independente de aumento maciço de vendas nos supermercados, especialmente de brinquedos.**

    **Poderemos dizer que o abaixamento de temperatura é causa directa do aumento de vendas?**

    **Provavelmente, não. É a época do Natal a causa real, apesar de a temperatura parecer ser também um bom predictor.**





## Predictor Colinear: Usar ou não?

- **Se não estiver disponível o real, será de utilizar.**
- **Problemas:**
  - **Suponhamos que utilizamos o modelo relativo às temperaturas e vendas no hemisfério sul?**
  - **E se num ano, tivermos um tempo particularmente ameno?**
- **Mesmo se utilizarmos o predictor correcto...**
  - **E se o evento Natal não for importante no país para onde exportarmos o modelo?**

# Teste do Qui-Quadrado

- **Utilizado largamente para verificar se há relacionamento entre duas colunas de uma base de dados. Faz uso do enunciado do princípio de independência.**
- **Este teste mede a diferença entre o número de ocorrências esperadas de uma combinação de predictores, supostos independentes e o número de ocorrências que efectivamente ocorrem. Na realidade não mede a diferença, mas o quadrado das diferenças.**