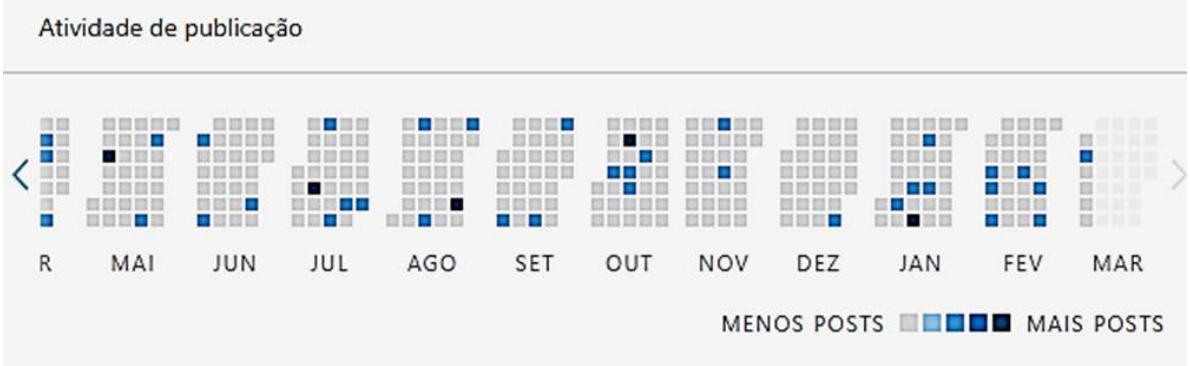
Mineração de Dados (Data Mining)

O tempo todo geramos dados, muitos dados!

Seja no rastro que deixamos ao fazer compras de produtos ou serviços, seja nas plataformas sociais, trocas de mensagens, etc.

Nós também passamos a ter sede de dados. A simples manutenção de um blog já gera dados interessantes:



Cada vez é mais forte o potencial de visualização das informações

As grandes empresas sabem muito a nosso respeito, ainda que muitas vezes não tenhamos consciência disso.

Empresas e o que sabem sobre nós:

clario.

The companies that know most about you

#	Company	% of personal data collected	Email	Name	Age	Gender/Sex	Sexual Orientation	Marital Status	Race	Religious Belief	Live Location	Home Address	Employment Status	Job Title	Pet/Animal Ownership	Mobile Number	Landline Number	Type of Phone/Device	Hobbies	Interests	Height	Weight	Next of Kin	Mother's Maiden Name	Current Employers	Past Employers	Bank Account Details	Salary	Social Profile (Friends)	Social Profile (Hobbies)	Social Profile (Interests)	Country of Birth	Allergies/Intolerances	Health & Lifestyle Info
1	Facebook	70.59%	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•		•	•		•	•	•				
2	Instagram	58.82%	•	•	•	•					•	•	•	•	•	•		•	•	•	•	•			•	•		•	•				
3	Tinder	55.88%	•	•	•	•	•	•			•		•	•	•	•		•	•	•	•						•	•	•	•			
4	Grindr	52.94%	•	•	•	•	•	•			•		•	•	•	•		•	•	•	•							•	•	•			
5	Uber	52.94%	•	•	•	•					•	•	•	•		•	•	•	•	•						•	•	•	•	•			
6	Strava	41.18%	•	•	•	•					•	•				•		•			•	•					•		•	•			
7	Tesco	38.24%	•	•	•	•		•				•				•	•	•									•		•	•			
8	Spotify	35.29%	•	•	•							•				•	•	•		•							•		•	•			
9	MyFitnessPal	35.29%	•	•	•							•				•	•	•			•	•					•					•	
10	Jet2	35.29%	•	•	•	•			•	•		•				•	•	•									•				•		
11	Credit Karma	32.35%	•	•	•								•	•		•		•			•	•					•	•					
12	Lidl Plus	32.35%	•	•	•						•	•			•	•	•	•									•					•	
13	Netflix	26.47%	•	•	•							•				•	•	•			•						•						
14	Nike	26.47%	•	•	•	•						•						•			•	•					•						
15	Asos	26.47%	•	•	•							•				•		•			•	•					•						
16	Depop	26.47%	•	•	•							•				•		•			•	•					•						
17	Ryanair	26.47%	•	•	•							•				•	•	•									•						
18	Ocado	26.47%	•	•	•							•				•	•	•									•					•	
19	Airbnb	26.47%	•	•	•	•						•			•	•		•									•						
20	American Airlines	26.47%	•	•	•							•				•	•	•									•						
21	Ikea	23.53%	•	•	•							•				•	•	•									•						
22	Trainline	23.53%	•	•	•						•	•				•		•									•						
23	Amazon	23.53%	•	•	•							•	•			•	•	•									•						
24	PayPal	23.53%	•	•	•							•				•		•							•		•						
25	eBay	23.53%	•	•	•							•				•	•	•									•						
26	Walmart	23.53%	•	•	•	•					•	•				•											•						
27	Deliveroo	20.59%	•	•							•	•				•	•										•					•	
28	Twitter	20.59%	•	•	•	•					•							•											•					
29	NHS COVID-19	20.59%	•	•	•						•	•						•															•	
30	SlimmingWorld	20.59%		•	•	•												•			•	•											•	
31	Google Maps	20.59%	•	•							•	•						•							•				•					
32	CVS Pharmacy	20.59%	•	•	•	•						•				•								•									
33	Amtrak	20.59%	•	•	•						•		•			•											•						
34	Sleepcycle	20.59%	•		•	•												•			•	•					•						
35	JustEat	17.65%	•	•								•						•									•					•	
36	Offerup	17.65%	•	•							•	•				•											•						
37	Doordash	17.65%	•	•							•	•				•											•						
38	McDonalds (USA)	14.71%	•	•							•					•											•						
39	TikTok	14.71%	•	•	•													•											•					
40	Protect Scotland	14.71%		•							•	•						•															•	
41	CoStar	14.71%		•	•											•		•																
42	Bet365 USA	14.71%	•	•	•							•				•																		
43	Wetherspoon	14.71%	•	•							•							•									•						
44	Skybet	11.76%	•	•	•																						•						
45	Flo Period Tracker	11.76%	•		•													•															•	
48	Whatsapp	11.76%	•	•												•		•																
47	Facetune	5.88%	•															•																
48	Pornhub	5.88%																•		•														

Em outras palavras, tanto nós quanto as empresas querem dados para analisar

Análise de dados

Dados são fatos brutos que representam eventos que ocorrem nas organizações ou no ambiente físico, antes de serem organizados e arranjados de uma forma que as pessoas possam entendê-los e usá-los. Dados precisam ser lapidados para que se tornem informações úteis.

Por que queremos dados?

Por que dados aparentemente sem muito sentido podem resultar em informações valiosas. Se dispomos de sistemas computacionais adequados, a seguinte transformação pode ocorrer:

A ideia de big data vem exatamente daí, da necessidade de extração de conhecimento a partir de muitos dados.

Junto ao big data está a ideia de mineração de dados (data-mining).

- Data-mining ou mineração de dados é o processo de explorar grandes quantidades de dados à procura de padrões para detectar novos relacionamentos entre variáveis.

- Procura descobrir padrões, tendências e correlações ocultas nos dados,

- Geralmente ainda está identificada com **algoritmos.**

Mineração de dados é usada para :

- ***Explicar:*** *... Por que caiu a venda de sorvetes no Rio de Janeiro?*

- ***Confirmar uma hipótese:*** *uma companhia de seguros pode querer examinar os registros de seus clientes para determinar se famílias de duas rendas têm mais probabilidade de adquirir um plano de saúde do que famílias de uma renda.*

- ***Explorar:*** *analisar os dados buscando relacionamentos novos e não previstos. Uma companhia de cartão de crédito pode analisar seus registros históricos para determinar que fatores estão associados a pessoas que representam riscos para créditos.*

O que é BI?

Baseia-se na capacidade de disseminar informações de forma rápida e segura auxiliando em processos de tomada de decisões.

As organizações recolhem informações a fim de avaliar o ambiente de negócio e converter em campos significativos, tais como pesquisa de mercado, de indústria, de marketing e análise de competidores

Utiliza técnicas de recuperação da informação, inteligência artificial, reconhecimento de padrões, estatísticas...

DM é usual em grandes bancos de dados e o resultado final pode ser exibido por meio de regras, hipóteses, árvores de decisão, etc.

DM evolui como evolui o conhecimento e a inteligência empresarial.

Aplicações do data-mining:

- Lucratividade
- Retenção de clientes
- Segmentação de clientes
- Propensão de clientes
- Otimização de canais
- Marketing dirigido
- Gerenciamento de risco
- Prevenção de fraudes
- Análise de cesta de compras
- Previsao de demanda
- Otimização de preços

Wal- Mart

- A cadeia de lojas Wal-Mart, que identificou um hábito curioso dos consumidores:

- *Ao procurar eventuais relações entre o volume de vendas e os dias da semana, o software de data mining apontou que, às sextas-feiras, as vendas de cervejas cresciam na mesma proporção que as de fraldas.*

- *Ao comprar fraldas para seus bebês, os pais aproveitavam para abastecer o estoque de cerveja para o final de semana.*

Data mining opera de acordo com diferentes métodos:

Método de visualização

Método estatístico

Método de dedução

Método de indução

Método de estruturas de decisão (árvores)

Método de Redes Neurais

Método de Associação

Método da Cesta de Compras

Método de algoritmos genéticos

Método de indução de regras

- Indução de regras: conjunto de regras que classificam os conjuntos de dados

- Ex: Se Renda > 60.000 e Débito < 10%,

então Risco "bom" => **Aprovar!**

Sistemas de apoio à decisão / modelagem analítica

- **Análise de Sensibilidade**

Observar como mudanças repetidas em uma única variável afetam outras variáveis.

Exemplo:

Vamos reduzir a propaganda em 1000 reais repetidamente de forma que possamos entender sua relação com as vendas.

Sistemas de apoio à decisão / modelagem analítica

- **Análise de Otimização**

Encontrar um valor ótimo para variáveis selecionadas, dadas certas restrições.

Exemplo:

Qual o melhor montante de propaganda considerando nosso orçamento e escolha de mídia?

Sistemas de apoio à decisão / modelagem analítica

- **Análise de Otimização**

Encontrar um valor ótimo para variáveis selecionadas, dadas certas restrições.

Exemplo:

Qual o melhor montante de propaganda considerando nosso orçamento e escolha de mídia?

Técnica de cesta de compras para gerar regras de associação

- As regras de associação são bastante utilizadas em banco de dados de tamanho grande e o resultado depende do algoritmo usado.

- O nome “análise da cesta de compras” tem origem no algoritmo que começou classificando o tipo de cliente de um supermercado de acordo com a composição da sua cesta de compras.

Técnica de cesta de compras para gerar regras de associação (cont.)

- Associação sequencial:

- Descobre associações que ligam eventos ao longo do tempo (ou seja, identifica padrões sequenciais).

- Exemplo:

Clientes que abrem uma conta corrente e após três meses abrem uma conta poupança e abrirão uma conta de cartão de crédito dentro de seis meses em 24% dos casos.

Método de associação

- Relações significativas entre itens de dados armazenados.

- O objetivo é encontrar tendências a partir de um grande número de transações.

- *Exemplo:* varrer registros de terminais de pontos de venda e descobrir que itens são vendidos juntos para redefinir disposição a campanhas.

Exemplo: Quando são comprados salgadinhos de milho, em 55% dos casos é comprado um refrigerante tipo coca-cola, a menos que haja uma promoção, caso em que a coca-cola é comprada em 75% dos casos durante a promoção.

Algoritmo de regra de associação

Passo 1 – Procurar a combinação de atributos, formando grupos com dois, três, quatro ou mais itens.

Passo 2 – Gerar regras de associação dessas combinações e calcular o nível de certeza de cada regra.

Exemplos

Exemplo do método regras de associação

- If (bebida=cerveja e comida=salame), then (cliente = homem e idade> 40 anos) com nível de certeza = 70%.

- If (bebida=champagne e comida=caviar), then (cliente =homem e idade> 60 anos) com nível de certeza =90%

- If (bebida=água mineral e roupa =casaco para bebê), then (cliente=mulher e idade > 30 anos), com nível de certeza =90 %

Número	Sexo	Atributos		Lipídios	Colesterol	Doença
		Idade (anos)	Peso (kg)			
1	masc.	25 ou menos	60 ou menos	normal	normal	Não
2	masc.	25 ou menos	60 a 80	normal	alto	Não
3	masc.	25 ou menos	60 a 80	alto	normal	Sim
4	masc.	25 ou menos	80 ou mais	alto	normal	Não
5	masc.	25 a 45	60 ou menos	alto	alto	Sim
6	masc.	25 a 45	60 a 80	normal	normal	Não
7	masc.	45 ou mais	60 ou menos	alto	alto	Sim
8	masc.	45 ou mais	80 ou mais	normal	normal	Sim
9	fem.	25 ou menos	60 ou menos	alto	normal	Não
10	fem.	25 ou menos	80 ou mais	normal	alto	Sim
11	fem.	25 a 45	60 ou menos	alto	normal	Não
12	fem.	25 a 45	80 ou mais	normal	alto	Sim
13	fem.	45 ou mais	60 a 80	alto	alto	Sim
14	fem.	45 ou mais	80 ou mais	normal	normal	Não

Grupos de dois itens:

(sexo=masc, idade=25 ou menos) → 3x

(sexo fem, peso = 80 kg ou mais) → 3 x

Grupos de quatro itens:

(sexo=masc, idade=25 ou menos, peso=60 a 80, colesterol normal)
→ 1 x

Grupos de 5 itens:

(masc, 60kg ou menos, lipídios=alto, colesterol=alto, doença=sim)

→ 2 x

Por exemplo, vimos que o grupo de cinco itens ocorre duas vezes e gera as seguintes regras, entre outras:

1 – Um item na parte IF e quatro na parte THEN:

if (sexo=masc), then (peso=60kg ou menos, lipidios = alto, coleterol = alto, doença=sim)

O antecedente (sexo=masc) ocorre em oito regras do banco de dados. Sendo assim, o nível de certeza desta regra é 2/8 ou 25%.

2 – Três itens na parte IF e dois na parte THEN

if (sexo=masc, peso=60kg ou menos e colesterol=alto) then (lipidios=alto e doença=sim)

O antecedente é 2/3

Importante

- É muito provável que o modelo inicial não atenda os objetivos do exercício de mineração de dados, sendo necessárias muitas repetições, especialmente entre as fases de projeto e de análise de dados.

- Isso envolve tentativas de diferentes técnicas de mineração de dados ou parâmetros em diferentes subconjuntos de dados antes de chegar a um resultado bem-sucedido!

Exemplo de algoritmo

Vamos considerar o seguinte cenário para a utilização do algoritmo. Um sistema de contas a receber de um clube esportivo envia para um banco no início de cada mês um boleto contendo a mensalidade do clube a ser paga pelos associados. O banco então envia pelo correio a fatura para os clientes e espera os recebimentos. No final do mês, o banco retorna para o sistema do clube quais clientes pagaram o boleto, quais não pagaram e quais clientes pagaram com atraso, dentre outras informações. Com o objetivo de diminuir a quantidade de clientes que pagam o boleto com atraso, foi feita uma **mineração de dados** na base de associados para identificar o perfil de quem paga com atraso o boleto.

Exemplo – Inadimplentes

Temos um conjunto de dados formado por 14 amostras.

Cada ponto representa alguém que contraiu um empréstimo no passado.

Eixo X: salário

Eixo Y: débito mensal (hipoteca, carro...)

- Base de dados relativos a empréstimos pessoais.

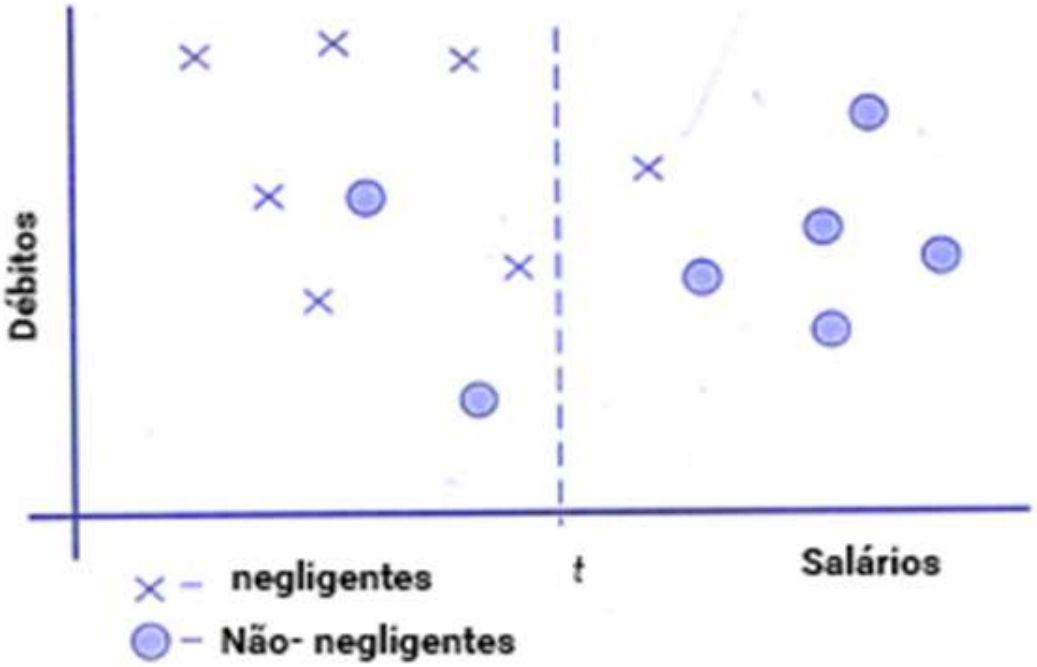
- O conhecimento que queremos extrair é:
  - *Como identificar os mutuários negligentes?*

- Há um consenso de que os atributos mais importantes são:
  - Salário

  - Débito

  - Regularidade de pagamento

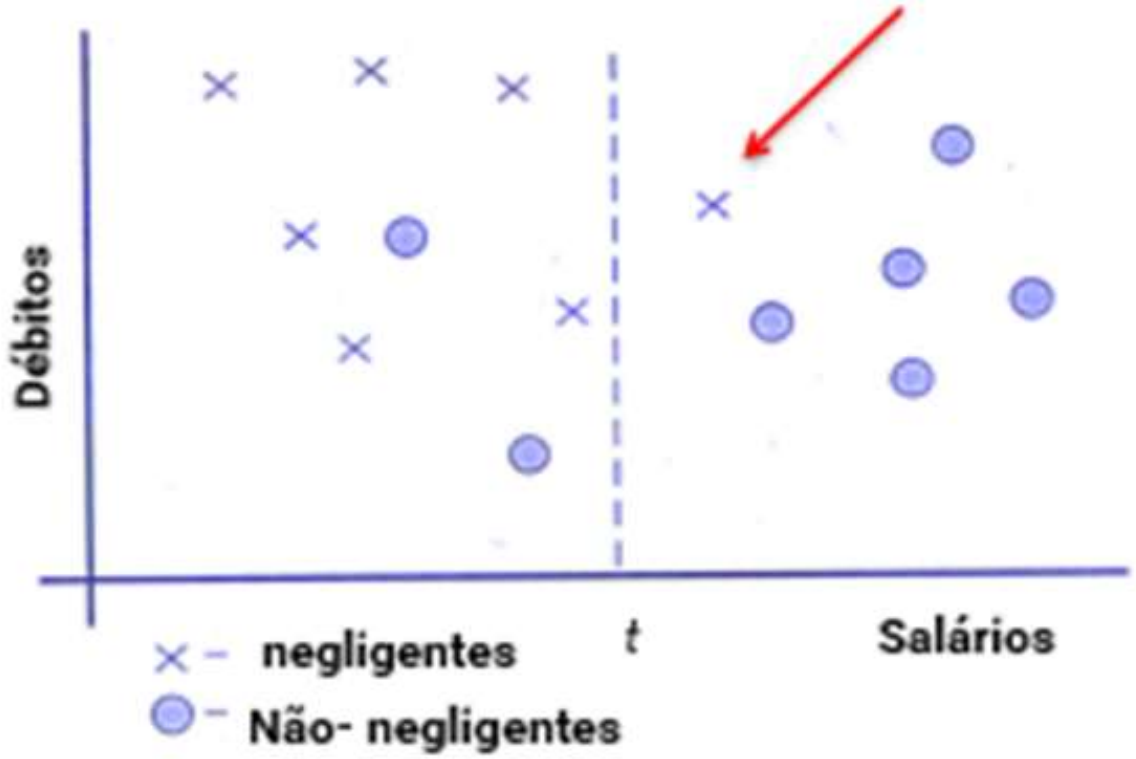
Como identificar inadimplentes?



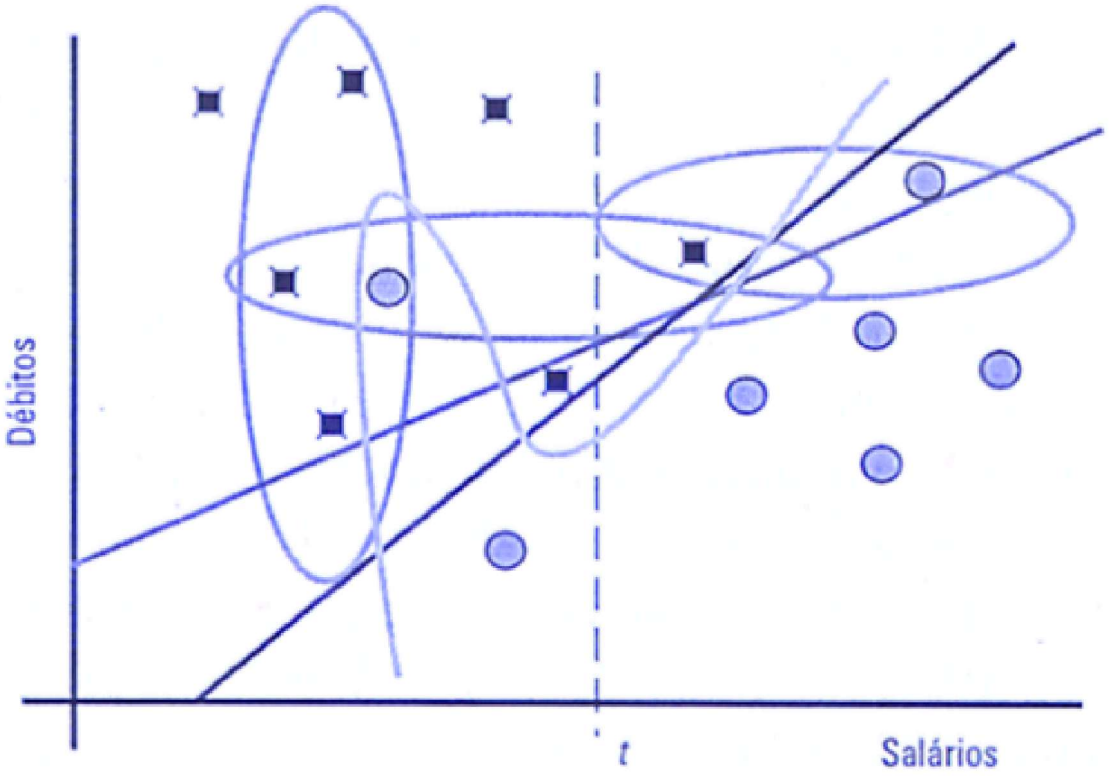
- t representa um padrão linear (parte da realidade)

- “Se x> t então o mutuário é bom pagador”

- Isso nem sempre é verdadeiro – observe:



Quando o padrão extraído do exemplo for definido, pode-se dizer que o conhecimento existente na base de dados foi descoberto!



Desafios

- Os softwares são trabalhosos e frustram quem espera respostas rápidas.
- Não há mágica...
- Os softwares de DM estão muito longe de descobrirem conhecimento sozinhos.

Como selecionar a ferramenta de data-mining?

1 - Que algoritmos suporta?

2 - Que saídas gera?

3 - Que formatação de entrada exige?

4 - Como a ferramenta/serviço adquire os dados para sua utilização?

5 – Como o analista interage com a ferramenta?

- É gráfica? – É linha de comando?

6 – Que nível de experiência é exigido desse profissional? Que pré-requisitos necessita?

7 – A ferramenta suporta valores contínuos ou discretos?

8 – Qual o limite de carga?

9 – É focada em problemas específicos?

10 – É uma ferramenta que "aprende"?

- Ex: possui um modelo de aprendizagem interno?

11 – Que organizações estão utilizando essa ferramenta? É exclusiva de um modelo de empresa?

12 – Que resultado já foram obtidos por essa ferramenta? Qual a relação de custo/benefício projetada?

Escolha da ferramenta:

- **Acesso a fonte de dados heterogêneas** (s/n)

- **Integração de conjuntos de dados** (s/n)
- **Facilidade de incluir novas operações** (s/n)
- **Facilidade para incluir novos métodos** (s/n)
- **Recursos para planejamento de ações** (s/n)
- **Processamento paralelo/distribuído** (s/n)
- **Métodos disponíveis....**(outro slide)

- No setor público é possível fazer o cruzamento de dados entre o estado civil de um funcionário e o salário que ele ganha, para verificar se isso tem influência na sua vida pessoal.

- Empresas como cadeias de supermercados podem recorrer a cruzamentos de dados para determinar que produtos são comprados em conjunto. Se um cliente X também compra o produto Y, talvez seja uma boa ideia posicionar os dos produtos perto, para facilitar a compra por parte do cliente.

- Ferramentas altamente especializadas.
- Elevado custo, em geral, pouco amigável.
- Voltada para grandes volumes de dados.
- Não há ferramentas melhores do que outras, depende da aplicação e interesses.
- O fator humano é preponderante.

RESULTADO:

- Escolhas de softwares/ferramentas são quase empíricas, demandam elevada ação humana.

Referência Bibliográfica

BROOKSHEAR, J.G. **Ciência da Computação: uma visão abrangente.** Porto Alegre: Bookman, 2013.

NOSENGO, N. **A extinção dos tecnossauros: histórias de tecnologias que não emplacaram.** Campinas: Editora da Unicamp, 2008.

Ir para exercício