Data Mining – Podemos passar sem ela?

Carla Henriques e Manuel Reis
Área Científica de Matemática

“Those who cannot remember the past are condemned to repeat it”

G. Santayana

Para evoluir é preciso aprender com a informação obtida da experiência passada.

Isto, claro, não é novidade e aplica-se a qualquer área do conhecimento.

Mas o que é que isto tem a ver com ***Data Mining***?

Data Mining faz exatamente isto: usa os dados recolhidos para aprender com a experiência passada.

O que é então *Data Mining*?

Vivemos numa era em que uma enorme quantidade de dados são recolhidos diariamente.

Data Mining

- É o processo de escrutínio de grandes conjuntos de dados para extrair conhecimento;
- Transforma um emaranhado de dados em informação útil para os decisores.

Data Mining

Pode ser usada em enumeras aplicações para:

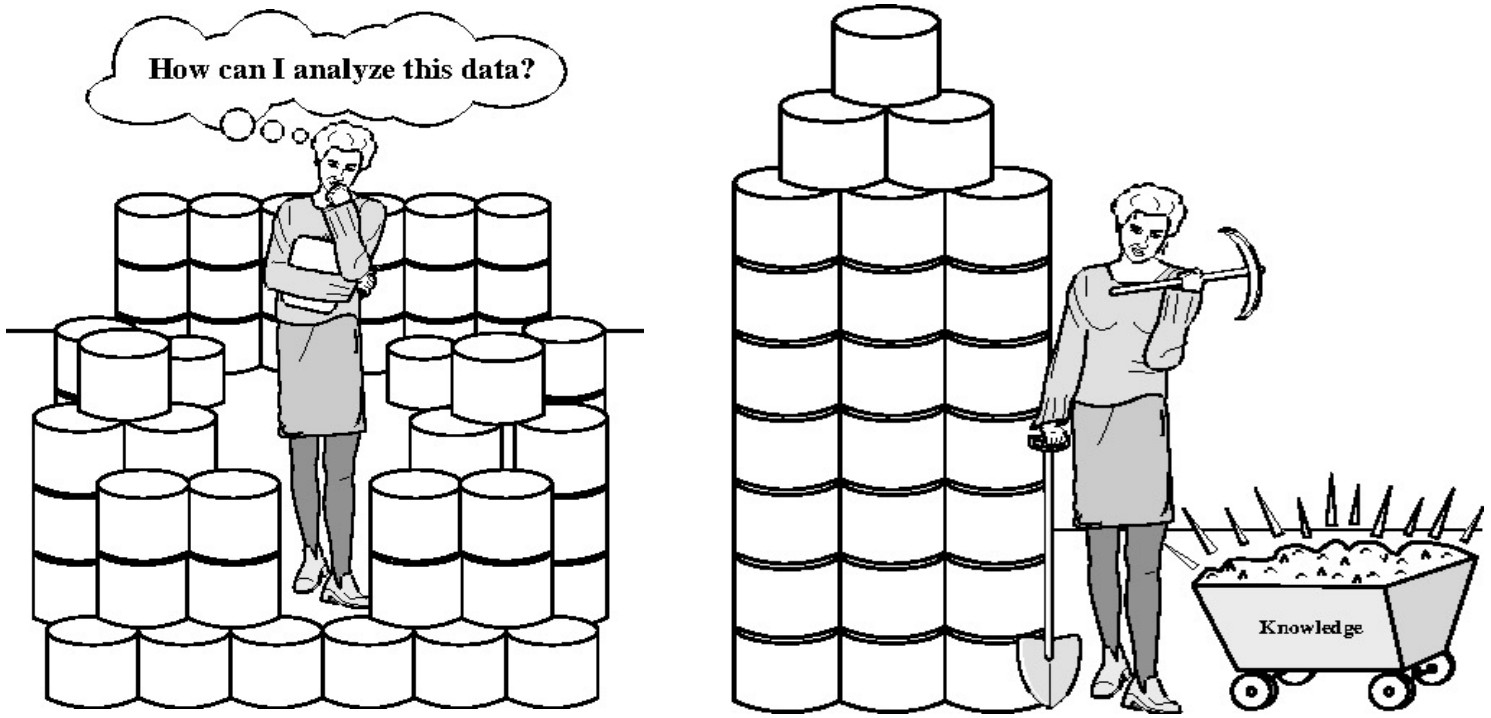
- Construir modelos – modelação;
- Classificar;
- Fazer previsões;
- Procurar anomalias;
- Identificar padrões;
- Identificar relacionamentos/correlações;
- Etc.

Data Mining

- Envolve um conjunto de metodologias da estatística, *machine learning*, inteligência artificial, …
- Nos dias de hoje a sua aplicação está amplamente generalizada: bancos, finanças, indústria, *marketing*, seguros, deteção de fraude, saúde, etc.

Data Mining

A Mineração de Dados (em Português), como o próprio nome indica, envolve minerar bases de dados de grande dimensão à procura de conhecimento que se revele útil.



Fonte: J Han & Kamber (2001).

Data Mining

- O recurso à mineração de dados na **indústria** remonta a 1990 e gradualmente foi recebendo mais e mais atenção por parte dos agentes e profissionais da indústria (Harding et al., 2006).
- Segundo a revista Forbes, num artigo de dezembro de 2017 de Louis Columbus, à data da publicação do artigo, 53% das empresas aplicavam *Data Mining* e esta percentagem estava em crescimento (Columbus, 2017).
- A mineração de dados é agora usada em diferentes áreas da engenharia de manufatura.

Data Mining

Aplicações de mineração de dados em engenharia de manufatura:

- Processos de produção
- Operações
- Deteção de falhas
- Manutenção preditiva – ajuda a identificar anomalias; aponta a necessidade de intervenção preventivamente
- Suporte à decisão
- Melhoria da qualidade do produto
- Gestão de relacionamento com o cliente
- *Engineering design*

Data Mining

Uma grande vantagem da mineração de dados é que os dados necessários para análise podem ser recolhidos durante as operações normais do processo de fabricação e, portanto, geralmente não é necessário introduzir processos dedicados à colheita de dados *(Harding, et al., 2006)*.

Data Mining - Exemplos

Publicidade/Marketing

Data mining é usada para estudar padrões de comportamento e necessidades associadas, de modo a direcionar as publicidades certas a determinados públicos;

Analisando relações existentes entre caraterísticas de consumidores, como idade, género, gostos, compras, etc., é possível:

- Prever o comportamento do consumidor;
- Direcionar campanhas especificas para determinados segmentos de mercado;
- Prever quais os consumidores que são mais propícios a adquirir certo produto/serviço;

Data Mining - Exemplos

Retalho

Supermercados estudam padrões de compras para identificar produtos que frequentemente se compram conjuntamente e, assim, decidir como dispor os produtos nos corredores e nas prateleiras.

Também é usada para estudar os dados de compras nas filas das caixas e decidir que produtos lá colocar para aumentar as vendas.

O mesmo exemplo se aplica a sites de vendas online para sugerir aos consumidores que procuram um produto, muitos outros que são muitas vezes comprados ao mesmo tempo. Já estamos habituados à frase: *"Quem viu este produto, viu estes também"* ou "*Produtos similares*" ...

Data Mining - Exemplos

Email – escolha das mensagens que vão para a caixa de spam dos nossos emails;

Empresas fornecedoras de serviços (luz, telecomunicações, etc) – os técnicos fazem manutenção preventivamente – *Data Mining* é usado para estudar os padrões associados a determinadas anomalias e conseguir prever essa anomalias antes de elas ocorrerem.

Data Mining - Exemplos

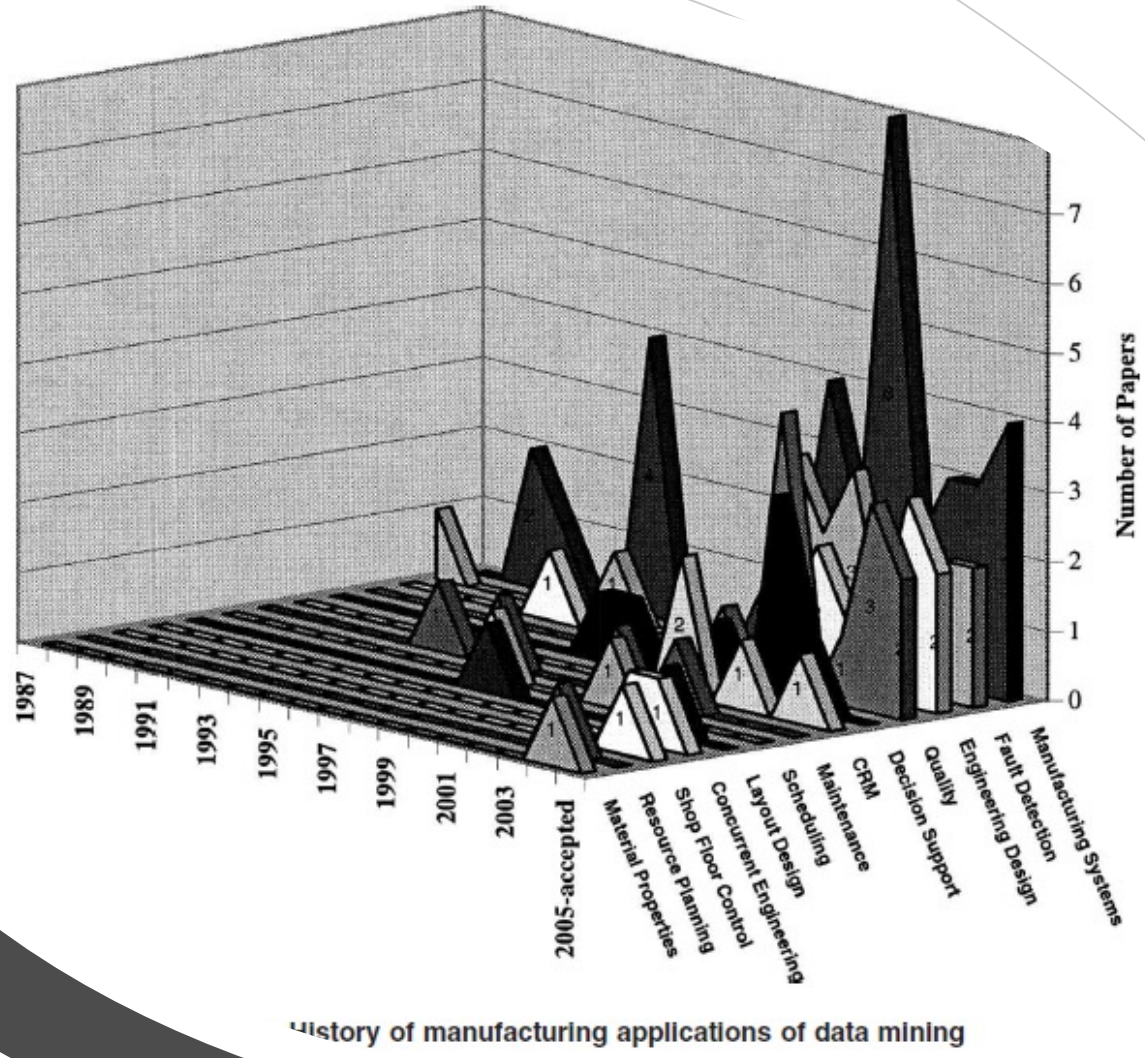
Bancos e seguradoras – usam *data mining* para avaliar riscos de modo a melhor definir pacotes de seguros, preços, etc.

Também é muito utilizado em sistemas inteligentes anti-fraude – analisar transações, pagamentos por cartão, padrões de compra, etc., de forma a traçar o perfil de transações duvidosas ou fraudulentas.

Data Mining - Exemplos

Saúde

- *data mining* ajuda a identificar padrões (de comportamentos/caraterísticas) associados a certas doenças; essa identificação contribui para um melhor conhecimento da evolução da doença e ajuda o médico nas decisões clínicas que perspetivam um cenário evolutivo favorável.
- Também permite uma gestão mais eficiente dos recursos de saúde, identificando riscos, prevendo doenças em determinados segmentos da população ou prevendo o tempo de internamento hospitalar.



History of manufacturing applications of data mining

Data Mining - Exemplos

Manufatura

CRISP-DM

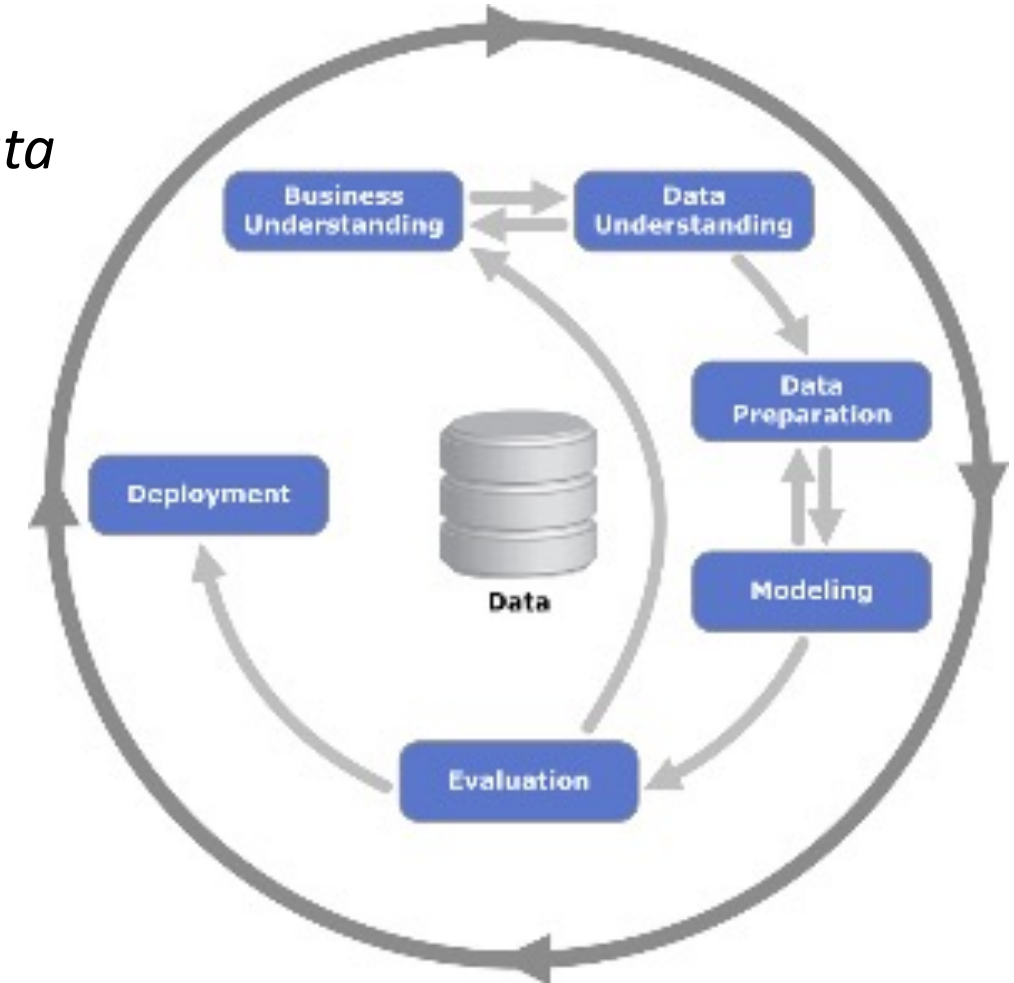
CRoss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (*CRISP-DM*)

- É um modelo de processo de *data mining* que permite a qualquer pessoa – desde novatos/inexperientes a especialistas – conduzir um projeto de mineração de dados.
- Foi uma iniciativa da Daimler Chrysler, SPSS e NCR (1996). Em 1997, a união europeia promoveu e financiou a criação de um modelo para tarefas de *data mining*.
- Objetivos: incentivar a interoperabilidade entre diversas ferramentas em todo o processo; eliminar o mistério/preço excessivo associado a tarefas simples de *data mining*.
- Em 1999 foi publicado o CRISP-DM

CRISP-DM: fases

O CRISP-DM estrutura o projeto de *data mining* em 6 fases:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment



CRISP-DM: fases

1. **Business Understanding (Entender o Negócio)**
   - Foco em entender o objetivo do projeto a partir de uma perspetiva de negócios definindo um plano preliminar para atingir os objetivos. Especificação do tipo de problema de *data mining* (e.g., classificação, predição, *clusters*).
   - Definição do plano do projeto: discussão da sua viabilidade com as pessoas envolvidas no projeto, criar procedimentos coerentes para os objetivos e técnicas selecionada, Identificar os recursos necessários e identificação dos passos críticos.

CRISP-DM: fases

2. Data Understanding (Entender os Dados)

Início de atividades para familiarização com os dados.

- Recolha de dados;
- Descrição dos dados: volume de dados, acessibilidade e disponibilidade dos atributos, compreensão do significado de cada atributo e do importância do atributo para o negócio;
- Exploração dos dados: análise minuciosa das propriedade de atributos interessantes (distribuição, relação entre pares de atributos, análises estatísticas simples);
- Análise da qualidade dos dados: identificação de problemas/erros.

Nesta fase emerge já alguma informação relevante para formular hipóteses.

CRISP-DM: fases

3. **Data Preparation (Preparação dos Dados)**

   Construção do conjunto de dados final a partir dos dados iniciais. Consome cerca de 90% do tempo do projeto.

   - Seleção dos dados: qual o conjunto de dados a usar, recolha de dados adicionais, explicar a exclusão/inclusão de determinados dados;
   - Limpeza dos dados: corrigir dados, remover dados, decidir como lidar com determinados valores e o seu significado (e.g., 99 para o estado civil), identificar valores omissos, *outliers*, etc..
   - Construção do conjunto de dados final

   Normalmente ocorre várias vezes no processo.

CRISP-DM: fases

4. **Modeling (Modelação)**

   As técnicas de *data mining* são selecionadas e aplicadas. Assim, é comum retornar à Preparação dos Dados durante essa fase.

   - Selecionar técnica;
   - Gerar design do teste: procedimento para testar a qualidade e validade do modelo;
   - Construção do modelo;
   - Avaliação do modelo: Seleção do melhor modelo entre os vários considerados, interpretação dos resultados na perspetiva do negócio.

CRISP-DM: fases

5. **Evaluation (Avaliação)**
   - Avaliar os resultados: avaliar os impactos para os objetivos da *data mining*, verificar se há informação nova e útil, ordenar os resultados de acordo com os critérios do negócio, etc.
   - Rever os processos: Obter uma visão global do processo de *data mining*, identificar falhas e possíveis ações alternativas, etc.
   - Determinar os próximos passos e tomar uma decisão.

   É necessário verificar se o modelo cumpre os objetivos do negócio.

CRISP-DM: fases

6. **Deployment (Implantação)**
   - Colocar o modelo obtido em prática;
   - Configurar para a mineração contínua de dados;
   - O conhecimento adquirido pelo modelo é organizado e apresentado de forma a que o cliente o possa utilizar.

Exemplo: Aviação

Descobrir os antecedentes de acidentes na aviação

- Tarefa: Descobrir os antecedentes dos acidentes na aviação de forma a permitir uma gestão proactiva do risco de segurança;
- Desafios: extrair informação fidedigna e útil de uma grande e heterogénea fonte de dados com a menor intervenção humana e integração da informação;
- Solução: Processamento de linguagem natural;
- Resultado: Aceite pela industria da aviação e pela FAA com beneficio potencial para milhões de pessoas.

Fonte: https://ntrs.nasa.gov/citations/20070018015

Exemplo – SI.mobil

Vamos ver um exemplo concreto de aplicação de soluções *Data Mining* numa empresa de Telecomunicações da Eslovénia, a SI.Mobil.

https://www.spssanalyticspartner.com/case-study-si-mobil/

Si.mobil

Predictive analytics cuts churn and reduces hardware investments, saving millions of euros per year

Exemplo – SI.mobil

When you are competing in a crowded market, how can you foster loyalty among your existing customers and attract new ones? By investing in state-of-the-art data modeling software, Slovenian telecommunications provider Si.mobil is now able to predict which handsets customers are likely to choose, and which customers are likely to switch provider.

Finding the key in analytics

Having already built a data warehouse to store data from its business systems, Si.mobil decided to purchase IBM® SPSS® Modeler Server, an advanced analytics platform that would help to model and predict future customer behavior.

Exemplo – SI.mobil

Overview

The need

To prosper in a crowded market, telecom provider Si.mobil wanted to find new ways to cut costs and boost revenues. It identified customer retention and handset investment as key areas for improvement.

Exemplo – SI.mobil

The solution

Si.mobil deployed powerful data modeling software to reveal deep insights into customer behavior – predicting whether customers are likely to churn and which handsets they are likely to choose.

Exemplo – SI.mobil

The benefit

Predicting churn helps Si.mobil boost retention by 10 percent, saving EUR1.1 million a year. Predicting customer handset choices enables the company to cut annual hardware investment by EUR1 million.

Exemplo – SI.mobil

Em resumo, com as soluções adotadas conseguiram:

- Diminuir o abandono (*churn*) dos clientes (aumentar a retenção), identificando os clientes com maior probabilidade de trocar de operadora;

  "The modeling team established 53 key performance indicators that correlate with customers who are likely to switch to another provider – including factors such as where a customer lives, how often they call or text, and whether they have called one of the company's competitors. The team then built two churn models – one for customers who have monthly contracts and one for customers with pay-as-you-go phones – that categorize customers as high, medium or low risk of churn. "

Exemplo – SI.mobil

Em resumo, com as soluções adotadas conseguiram:

- Prever quais os telemóveis que os clientes vão adquirir ao longo do período do contrato. A SI.mobil tem de os comprar antecipadamente e assume o risco de o cliente não cumprir o contrato. Para a empresa faz sentido incentivar os cientes a comprar equipamentos mais económicos de forma a minimizar o investimento e o risco.

"Si.mobil set out to predict which handset each customer was likely to purchase, looking at factors such as how much they are likely to spend, which handset manufacturer they would select, which operating system they would choose, whether they would purchase the handset outright or via a monthly payment on their bill, and how often they call, text or use mobile data.
By examining these factors, Si.mobil can establish a shortlist of phones that each customer is likely to purchase – and offer this shortlist to the customer, rather than the entire range. "

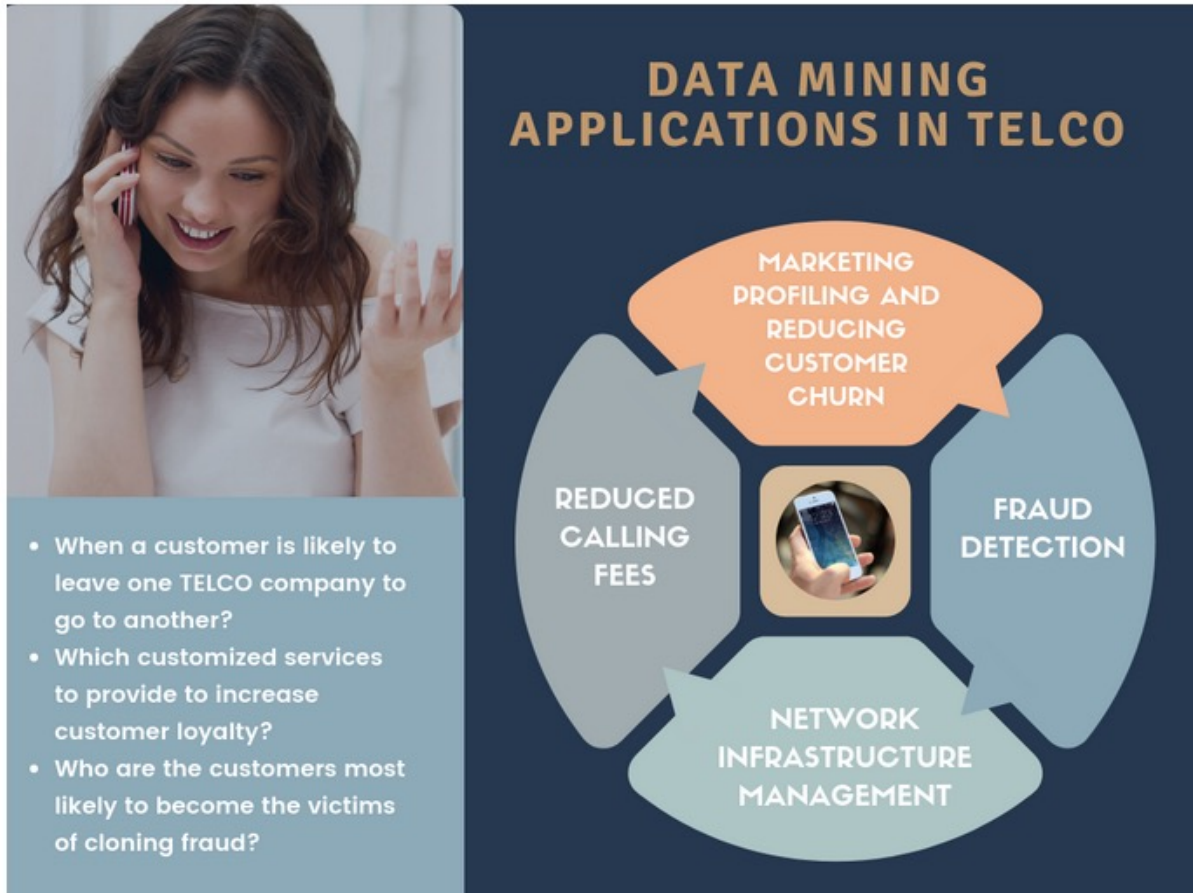
Exemplo – SI.mobil

A notícia dá ainda conta de futuras planos da empresa:

- Avaliando a satisfação dos clientes no que diz respeito à qualidade e cobertura da rede permite à empresa identificar zonas com pior receção e cobertura. Ao combinar a informação do estado da rede com as falhas reportadas pelos clientes e com dados de análise comportamental é possível identificar erros que causam verdadeiros problemas aos utilizadores.

> *Andreja Stirn explains: "Call quality and network coverage are hugely influential in securing and maintaining the satisfaction of our customers.(...)*
> *The SPSS platform is a key enabler for this type of large-scale analysis, which will require the company to match seven million daily call records with hundreds of millions of network events. "*

Exemplo:
Telco (Empresas de telecomunicações)



Possible Question	
When a customer is likely to leave one TELCO company to go to another?	Customer retention management
Which customized services to provide to increase customer loyalty?	Sales and customer loyalty management
Who are the customers most likely to become the victims of cloning fraud?	Fraud detection

Data Mining

Empresas como a IBM ou a Samsung, entre muitas outras, disponibilizam há muito tempo soluções para que as empresas possam lidar de forma eficiente com os dados que recolhem tirando partido da informação neles contida.

A título de curiosidade poderão espreitar os vídeos seguinte:

https://www.youtube.com/watch?v=31W0EzcfE74

https://www.youtube.com/watch?v=utxb6bbq820

Data Mining – Software

- Angoss knoeledge seeker
- Right point datacruncher
- WEKA (freeware)
- RapidMiner
- IBM SPSS Modeler (pode usufruir gratuitamente por 30 dias)