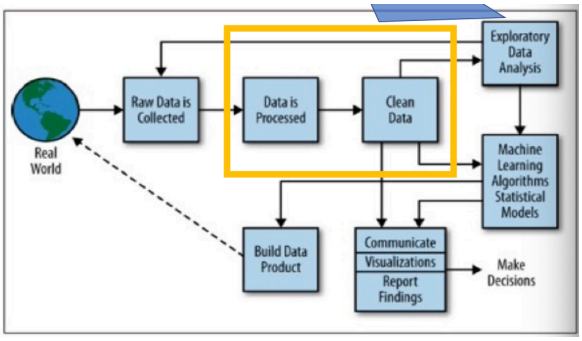
Descoberta e Extração de Dados

2 - Dados

Ciência de Dados: fases

- Mundo real
- Dados não tratados (raw data)
- Tratamento ou Preparação dos dados



Mas o que são dados?

- Uma fotografia de férias tirada com o telemóvel?
- A chapa de matrícula de um automóvel?
- O saldo da conta bancária?
- O número de contribuinte fiscal?
- Uma sequência "0000, 0001, 0001, …" é um dado?

… tudo são (podem ser) dados!

Dados: Definições

- Data: informação factual (tal como medidas ou estatísticas) usadas como base de raciocínio, discussão ou cálculo.

- Definição especializada no modelo OAIS - Open Archive Information System, ISO 14721:2012:
  - **Dados são uma representação, reinterpretável, de informação, formalizada para comunicação, interpretação ou processamento**
- O modelo OAIS faz ainda uma distinção interessante entre **dados**, **informação** e **conhecimento**:
  - Informação é qualquer tipo de conhecimento que se pode partilhar. Na partilha, a informação é representada por por dados. Por exemplo, **uma sequência de bits acompanhada por uma descrição de como interpretar essa sequência** como números que representam observações térmicas medidas em graus Celsius.

- Dados são factos descontextualizados: palavras, números, datas, imagens, sons, etc.
- Exemplos: 42, 86, laranjas, 35€, porto, 11:00, 963456221…

-> Para dados constituírem informação é necessário contextualizar esses dados, dar significado.
-> Informação são factos compostos por dados contextualizados por uma interpretação/significado

Informação: Dados com estrutura e…

-> Para transformar dados em informação necessitamos de os estruturar **e fornecer significado**.
-> Não basta organizar (tabela), é necessário contextualizar (dar significado a cada coluna).

Dados estruturados

- Dados como informação de vários tipos (médicos, clientes, questionários, sensores…)
- Dados que estão organizados em exemplos (ou observações) por variáveis ou dimensões (ou atributos ou características)
- Uma linha na tabela é um ponto no conjunto de dados (dataset)

Conhecimento

- Conhecimento é formado de informação, a partir da compreensão dessa informação de modo a formar julgamentos, opiniões, previsões e tomar decisões

Exemplo

- Informação

Produção de Laranjas					
Anos	2015	2016	2017	2018	2019
Toneladas	23	26	31	33,5	38

- Conhecimento
  - *Durante os 5 anos anteriores, a produção de laranjas tem crescido, em toneladas, uma média de 10% ao ano. A mesma tendência de crescimento é esperada este ano, pelo que será necessário garantir escoamento de produção para mais 10% de toneladas este ano relativamente ao ano anterior.*

Dados não estruturados

- Dados como informação de vários tipos não organizados
- Exemplos: texto, imagens, video, audio,…
- Existe estrutura nestes dados, mas não está organizada como no exemplo anterior
- É necessário extrair elementos de informação e perceber como estão relacionados

Dados heterogéneos

- Exemplo I : Saúde

- Análises médicas — estruturados
- Notas do médico
- História familiar
- Imagens médicas (radiografias, Ecografias, …)
- Histórico do paciente
- Sintomas atuais — estruturados (ou não …)

(Notas do médico, História familiar, Imagens médicas, Histórico do paciente: não estruturados)

Questão possível a responder

- O paciente está em risco de contrair uma dada doença?

- Exemplo II: Fornecimento de eletricidade

- Respostas a um questionário sobre necessidades eléctricas de diferentes (amostras de) populações — estruturados
- Custos laborais — estruturados
- Imagens de satélite — não estruturados
- Custos de materiais (paineis fotovoltaicos, baterias, …) — estruturados

Questão possível a responder

- Onde deve ser localizada uma instalação fotovoltaica num dado país?

- Exemplo III: Educação a distância

- Estatísticas sobre os videos (acessos, # de visualizações, …) — estruturados
- *Posts* no forum do curso — não estruturados
- Questões sobre os materiais — não estruturados
- *Clickstream* no sítio do curso — não estruturados
- Submissão de TPCs — estruturados e não estruturados

Questão possível a responder

- O que deve ser alterado num curso on-line para melhorar o sucesso dos alunos?

Contexto e Metadados

- Dados não são objetos naturais com essência própria
- A informação encontra-se no contexto
- A sua semântica (significado) depende do contexto e da perspetiva do seu utilizador (estando sujeito a parcialidade/ "bias")

- O contexto pode ser (parcialmente) expresso por dados acerca de dados (metadados)
- A riqueza desses metadados influencia a possibilidade de transferência e re-utilização
- Exemplos de metadados incluem:
  - fonte, data de criação, propriedade, formato, codificação, …
- Metadados podem interligar-se, criando grafos de conhecimento ("knowledge graphs") - ontologias de informação
- "Knowlegde graphs" foram popularizados pela Google que os usa para enriquecer os resultados do seu motor de busca

Bias (enviesamento) - não intencional

- O que é o Sampling Bias?
  - O sampling bias ocorre quando a amostra de dados numa investigação sistemática não representa com precisão o que é possível obter no ambiente de investigação
  - Acontece quando se recolhem dados de uma forma que alguns membros da população pretendida têm uma

probabilidade de amostragem menor ou maior do que outros

- Subrepresentação
  - A subrepresentação é um tipo comum de sampling bias
  - Algumas das variáveis da população estão mal representadas (ou não estão de todo representadas) na amostra do estudo
  - Causa comum
    - Amostragem de conveniência
      - só se recolhem amostras de dados de fontes facilmente acessíveis

Dados Abertos (Open Data)

- A socialização dos dados (a ideia de que deverão ser l**ivremente acessíveis para qualquer um os poder explorar sem restrições** de direito de cópia, patentes ou outros mecanismos de controlo) tem ganho adesão crescente;
- O movimento Open Data é comparável ao movimento Open Source para o software
- Apesar do acesso a dados já se colocar há vários anos, nomeadamente no domínio científico, o tema ganhou popularidade com a internet
- A questão de open data é principalmente relevante no domínio da administração pública, onde a transferência é exigidas pelos cidadãos

Ética

- Bias
  - Assegurar que os conjuntos de treino, os algoritmos e parâmetros que utilizamos não são, à partida, enviesados.
- Desigualdades
  - Apontar sempre para a (tentativa de) redução das desigualdades no planeta, nomeadamente a divisão digital desigual
  - A Ciência de Dados tem um papel importante na democratização da educação e oportunidades de desenvolvimento: posiciona-se como fulcral para a concretização do Objetivos para o Desenvolvimento Sustentável (ODS)
- Segurança e Privacidade
  - O RGPD (GDPR) - Regulamento Geral de Proteção de dados visa proteger o direito à privacidade e esquecimento,…

Qual o impacto da falta de Privacidade?

- 25% empresas em Portugal foram alvo de um ataque no último ano
- 229: média de dias que as empresas demoram a detetar ataques
- 62% : total de ataques a pequenas e médias empresas
- 59% : ex-empregados admitem roubar dados quando saem das empresas

Normas de direito internacional

- Protegendo a privacidade em geral
  - Art. 12º da Declaração Universal dos Direitos do Homem
  - Art. 17º do Pacto Internacional relativo aos direitos civis e políticos
  - Art. 8º das Convenção Europeia dos Direitos do Homem

- Protegendo os dados pessoais, em especial
  - Convenção 108 do Conselho da Europa (convenção para a Proteção das Pessoas relativamente ao Tratamento Automatizado de Dados de Carácter Pessoal), aprovada em 1981

- Carta dos Direitos Fundamentais da União Europeia

Necessidade de um novo enquadramento jurídico?

- A quantidade de dados armazenados e transacionados, estruturados e não estruturados, aumentou e aumentará exponencialmente nos próximos anos

- O desafio da tecnologia
  - 92% dos europeus estão preocupados com aplicações móveis que recolhem os seus dados sem o seu consentimento
  - 89% das pessoas dizem que querem saber quando os seus dados no smartphone são partilhados com terceiros. Querem a opção de dar ou recusar a permissão
  - 3 em cada 4 cidadãos não sentem que controlam os seus dados

- Poderá a economia continuar a crescer sem a confiança dos cidadãos?

- A recolha e utilização dos dados pessoais é uma preocupação cada vez maior dos titulares de dados e dos responsáveis pelo tratamento de dados

- A enorme quantidade de regulamentos na área da proteção de dados no EEE, força um regime jurídico mais rigoroso

- A salvaguarda de "novos direitos" aos titulares dos dados, como aceder, alterar, transferir, apagar ou solicitar a sua informação na qualidade de consumidores, fornecedores ou colaboradores

- O RGPD visa responder aos desafios colocados pela revolução tecnológica ocorrida nas últimas décadas e aumentar a proteção das pessoas singulares no que diz respeito aos tratamentos de dados pessoais e à livre circulação desses dados

O que traz de novo?

- Um regulamento (agregador de normas dispersas em tratados, cartas, convenções e legislação nacional)
- Todos os países com o mesmo regime (evita que as empresas procurem os países mais permeáveis)
- Reforço dos direitos dos cidadãos como titulares dos dados (the right to be left alone)
- Facilita fluxos de dados internacionais assegurando uma proteção adequada
- Orientações, regulação e aplicação efetiva de coimas
- Novas obrigações para empresas
- Extensão do âmbito de aplicação

A quem se aplica?

- Com a finalidade de contribuir para a harmonização da legislação de todos os países do Espaço Económico Europeu, o RGPD aplica-se a todas as pessoas singulares e coletivas que efetuem tratamento de dados pessoais a residentes do EEE
- A maior alteração às leis de privacidade nos últimos vinte anos
- Vai obrigar a uma mudança significativa na forma como todas as empresas recolhem e tratam os dados pessoais, obrigando à implementação de mecanismos de controlo e de capacidades de gestão para garantir a privacidade dos dados pessoais que sejam recolhidos
- Coimas que podem ir de 20 milhões de Euros até 4% do volume de negócios e sanções de natureza civil e criminal

Um "novo" glossário

dados pessoais	tratamento transfronteiriço	limitação do tratamento	definição de perfis
pseudonimização	autoridade de controlo	**responsável pelo tratamento**	subcontratante
destinatários	objeção pertinente e fundamentada	consentimento	anonimização
dados genéticos	dados biométricos	dados sensíveis	sociedade de informação
representante	organização internacional	**titular dos dados**	apagamento

Mudanças mais relevantes

- Mudança do paradigma de regulação externa, para **auto-regulação**
- Deveres de informação e de obtenção de **consentimento explícito**
- Reforço dos **direitos dos titulares dos dados**
- **Dever de notificação** em caso de violações de dados pessoais
- Imposição do tratamento dados numa lógica privacy by design e by default
- Obrigação de **conservação de um registo** das atividades de tratamento
- Designação de encarregado da proteção de dados (**Data Protection Officer**)

O que são dados pessoais?

- **Qualquer informação**, de qualquer natureza e independentemente do respetivo suporte, incluindo som e imagem, **relativa a uma pessoa singular** identificada ou identificável

O que é o tratamento de dados pessoais?

- Uma operação ou um conjunto de operações efetuadas sobre dados pessoais ou sobre conjuntos de dados pessoais, por meios automatizados ou não automatizados, tais como a recolha, o registo, a organização, a estruturação, a conservação, a adaptação ou alteração, a recuperação, a consulta, a utilização, a divulgação por transmissão, difusão ou qualquer outra forma de disponibilização, a comparação ou interconexão, a limitação, o apagamento ou a destruição
- Todas as atividades que refletem o ciclo de vida da informação, desde a sua recolha até à destruição

A Reter…

- Ciência de Dados: fases
- Dados vs Informação vs Conhecimento
- Dados estruturados vs Dados não estruturados
- Dados heterogéneos
- Bias
- Privacidade