NOME DO DOCUMENTO

**Data_Science.pdf**

NÚMERO DE PALAVRAS

**2447 Words**

NÚMERO DE CARACTERES

**13374 Characters**

NÚMERO DE PÁGINAS

**8 Pages**

TAMANHO DO ARQUIVO

**716.4KB**

DATA DE ENVIO

**Jan 15, 2024 6:38 PM GMT**

DATA DO RELATÓRIO

**Jan 15, 2024 6:39 PM GMT**

## 17% geral de similaridade

O total combinado de todas as correspondências, incluindo fontes sobrepostas, para cada banco de

- 16% Banco de dados da Internet
- Banco de dados do Crossref
- 7% Banco de dados de trabalhos enviados
- 5% Banco de dados de publicações
- Banco de dados de conteúdo publicado no Cross

# Data Science

TOMÁS GOMES, Universidade da Beira Interior, Portugal
TIAGO RISCADO, Universidade da Beira Interior, Portugal

O estudo de dados para a extração de *insights* de uma quantidade exorbitante de informação através de coleta, limpeza, processamento e análise dos dados de forma a tranformá-los em informação útil.

Additional Key Words and Phrases: Dados, *BIG DATA*, Python, Organização, *Insights*, Processamento, Utilidade, Recolha

Contents



## 1 INTRODUÇÃO

Com a exponencial evolução da tecnologia, foi necessário criar um mecanismo que armazenasse e tratasse de gigantescas quantidades de informação, mecanismo esse chamado *DATA SCIENCE* que utiliza métodos, processos, algoritmos e sistemas científicos para extrair conhecimento e *insights* de dados em várias formas.

Inicialmente a informação é armazenada em *BIG DATA* que se refere a conjuntos de dados volumosos que são analisados para compreender as tendências dos dados. No entanto, apenas depois é utilizada a *DATA SCIENCE* que cria

Authors' addresses: Tomás Gomes, Universidade da Beira Interior, Castelo Branco, Portugal, tomygomes7@gmail.com; Tiago Riscado, Universidade da Beira Interior, Covilhã, Portugal, tiago.d.riscado@gmail.com.


Manuscript submitted to ACM

algoritmos de aprendizagem automática para desenhar e criar métodos estatísticos com o objetivo de gerar informações a partir de *big data*.

O objetivo fundamental da *DATA SCIENCE* é analisar e interpretar dados, para resolver problemas complexos e descobrir padrões posteriormente armazenados que possam ser úteis.

## 2 BIG DATA

O tema *Big Data* é relativamente recente e consiste na recolha de grandes volumes e variedades de dados, o que torna o processo desafiante pois estes não podem ser trabalhados e processados por bancos de dados tradicionais.

Cada dia mais de 2.5 exabytes (2,500,000,000 GB) são criados e armazenados criando a necessidade de organizá-los e processar a informação útil para os utilizadores separando-a do resto.

### 2.1 Envolução

Ao longo dos anos, a tecnologia tem criado novas maneiras de armazenar os dados. Inicialmente na decada de 70 e 80 eram armazenadas em bancos de dados relacionais com capacidade limitada, com o avanço da tecnologia por volta do ano 2000 foi criado o armazenamento em nuvem onde se tornou possível armazenar grandes volumes de dados de forma eficiente. Na decada de 2010 o *Big Data* tornou-se crucial para a inteligencia artificial e o *machine learning*, com uma forte integração nessas tecnologias .

### 2.2 Estruturação dos Dados

Todos os dados têm uma classificação e para entendermos de forma clara o assunto do *Big Data* precisamos de entender a base da classificação da estrutura dos dados. Existem três tipos de estrutura de dados estruturados, dados não estruturados e dados semiestruturados.

Os dados estruturados diz respeito a estruturas rígidas, ou seja , antes mesmo da presença dos dados, aquele ambiente já foi pensado para eles. Por exemplo - bancos de dados, arquivos CSV, arquivos XML, arquivos JSON .

Dados não estruturados são os mais comuns, equivalem a cerca de 80% dos dados do mundo e estão presentes em aplicações do dia a dia como - YouTube, WhatsApp, estes são dinâmicos e flexíveis e têm uma complexidade um pouco maior para análise com difícil processamento e recuperação.

Sem falta também temos os dados semiestruturados como o nome indica é uma misturas dos dois tipos falados em cima possuem características definidas mas não são rígidos.

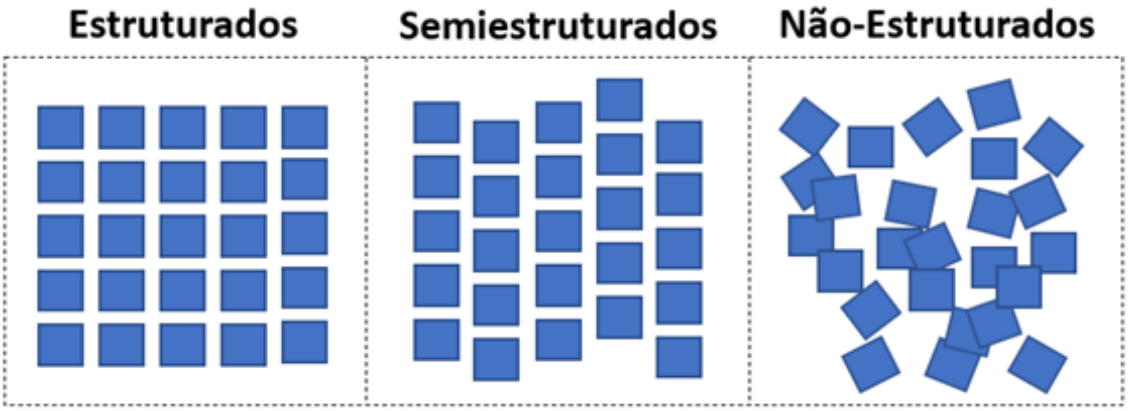


Fig. 1. Estruturas

### 2.3 5 V's do Big Data

Os 5V's do *Big Data* - velocidade, volume, valor, variedade e veracidade - são a características mais importantes e inatas do *Big Data*.

A combinação de dados não estruturados, estruturados formam o *Big Data*. Os dados podem ser explorados para obter *insights* e usados para desenvolvimento da *machine learning* para tal é necessário :

(1) Volume:
O volume é o ponto de partida para entender o *Big Data*, cada objeto ou pessoa conectado á Internet gera dados constantemente. É preciso saber onde buscar as informações necessárias nessa quantidade incontável de dados para devolver ao utilizador consoante a necessidade do mesmo.

(2) Velocidade:
Refere-se à rapidez com que os dados são gerados e com que rapidez eles se movem. Isto é importante quando existe a necessidade de que os dados fluam rapidamente.
Por exemplo uma organização que armazene uma grande quantidade de dados terá um fluxo grande de dados entre máquinas, redes,etc. Na área da saúde, o uso de grandes quantidades de dispositivos médicos para monitorização de pacientes e recolher dados que necessitam de ser analisados e enviados rapidamente para o seu destino.

(3) Valor:
Numa empresa não adianta ter uma boa quantidade de dados processados de forma rápida se estes não puderem agregar valor á empresa. A analise desses dados e entender com podem ser úteis é importante para que sejam valorizados.

(4) Variedade:
Refere-se a diversidade de tipo de dados, podemos ter dados de diversas fontes, esse dados coletados podem ser não estruturados, semiestruturados ou estruturados.

(5) Veracidade:
Corresponde à qualidade, precisão e credibilidade dos dados coletados. Nem toda a informação é verdadeira, existem dados falsos gerados com segundas intenções como também existem dados desatualizados, o desafio do *Big Data* é separar as informações importantes do resto.

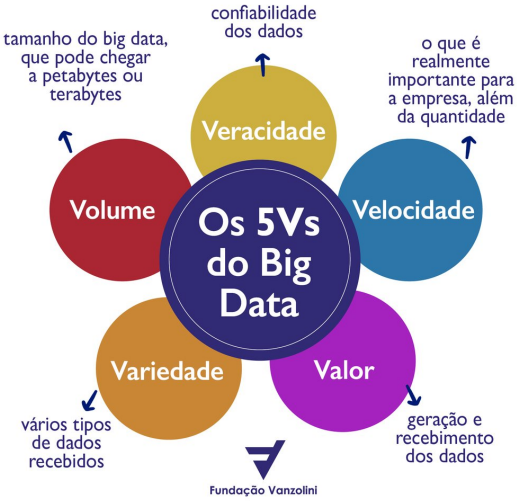


Fig. 2. 5 V's

## 3 METODOLOGIA

### 3.1 Conjuntos de Dados utilizados

Em *DATA SCIENCE* lidamos com uma grande variadade de dados que podem ser agrupados em três secções consoante a sua origem:

| | Origem | Exemplos |
|---|---|---|
| | Dados gerados por seres humanos | Registros de experiências humanas, anteriormente gravados em livros e obras de arte e, posteriormente, em fotografias, áudio e vídeo, disponibilizados em redes sociais, blogs, documentos pessoais, aplicativos de mensagens instantâneas, e-mails etc. |
| | Dados mediados por processos | Registros produzidos por sistemas de negócios de agências públicas ou privadas, como registros médicos, transações comerciais e financeiras, dados de e-commerce etc. |
| | Dados gerados por máquinas | Registros derivados da Internet das Coisas, de sensores fixos (de clima, poluição, tráfego, segurança, automação doméstica etc.) ou de sensores móveis (geolocalização de aparelhos celulares, carros e imagens de satélites), além de registros de sistemas computacionais (como logs e Web logs). |

Fig. 3. Variadade de Dados

Podemos observar que as redes sociais geralmente geram dados provenientes de interações humanas, enquanto a Internet das Coisas (IoT) é baseada em informações recolhidas por máquinas. Além disso, os dados provenientes dos processos dos sistemas de negócios, como registos de transações, constituem uma terceira fonte de dados, sendo as organizações uma parte significativa na geração desses dados.

Na educação, as atividades dos alunos e professores geram uma variedade de dados educacionais (frequência, participação, interação, produção, etc.), os quais, quando registados digitalmente, podem ser utilizados desde o planeamento de ações didáticas específicas até avaliações em grande escala, seja a nível nacional ou internacional.

Tanto nos dados de origem humana quanto nos provenientes de organizações, existe sempre uma interação com as máquinas. Em alguns casos, pode ser desafiador distinguir se os dados são gerados por pessoas ou por máquinas.

## 2.2 Ferramentas e linguagens de programação utilizadas

Existe uma grande variedade de linguagens de programação que nos permitem trabalhar na aréa da *DATA SCIENCE*, entre elas temos alguns nomes bastante conhecidos: **Python, Linguagem R, Java, Linguagem C**.

Falemos um pouco sobre cada uma:

.

- **Python**:
  A linguagem Python apresenta uma sintaxe simples e é reconhecida como uma das mais acessíveis para aprendizagem. Além disso, é altamente apreciada pelas empresas, visto que oferece oportunidades que vão para além da análise de dados. Isso ocorre porque ela não se limita apenas à manipulação de dados, mas também se destaca no que diz respeito ao desenvolvimento web e telemóveis (sites e aplicações), elaboração de protótipos, automatização de scripts diferentes , entre outras aplicações.
  Esta linguagem de programação demonstra capacidades em capturar, organizar e manipular extensos conjuntos de dados, contando ainda com uma ampla gama de bibliotecas para aqueles que desejam utilizar os dados de forma estruturada.
  Os pacotes que possui foram desenvolvidos especificamente para Data Science sendo essenciais para o trabalho com dados, abrangendo desde a manipulação até à sua visualização. Dentre eles, destacam-se o **scikit-learn** (software aprendizagem automatica), **pandas** (manipulação de dados) e **Matplotlib** (representação gráfica), todos amplamente utilizados nesta área.
- **Linguagem R**:
  A Linguagem R é uma linguagem de programação para Data Science bastante completa e com muitos recursos, conhecida por ser a mais robusta para a área de dados. Por isso, muitos preferem utilizar R em vez de Python.
  O seu maior diferencial, sem dúvidas, é o fato de ter sido projetada, especialmente, para o uso em cálculos e análises estatísticas para o trabalho com dados. Assim, foi pensada para que auxiliasse na manipulação, análise e visualização de dados. Para atender a essas necessidades, a linguagem R inclui diferentes pacotes de cálculos estatísticos e matemáticos, que contribuem na construção da análise e da probabilidade, além de também ser utilizada para machine learning.
- **Java**:
  É uma linguagem "clássica" da computação em alta performance, por isso, é vista como uma linguagem de propósito geral, incluindo manipulação de dados.
  Já dentro da *DATA SCIENCE*, esta linguagem pode ser utilizada principalmente na criação de modelos de machine learning e na manipulação de dados em Big Data. Também é usado nos produtos de Internet das Coisas (IoT).
- **Linguagem C**

1 ma das pioneiras para a computação moderna como conhecemos hoje. A linguagem C tem grande potencial para a área de *DATA SCIENCE* pois apresenta funcionalidades que tornam a estrutura de dados menos complexa e mais facilmente de ser "organizada".

A linguagem C é um tipo de linguagem de programação bastante mais complexa que as outras referidas, como por exemplo python. Por outro lado no momento em que se aprende C facilmente se domina as restantes linguagens.

:

## 4 ETAPAS DE UTILIZAÇÃO

Nos podemos interpretar a ciência de dados como um ciclo de funções, ferramentas processos para a obtenção dos *insights*. Para essa obtenção o projeto tem de passar pelas seguintes etapas:

.

(1) Coleta de dados:
A base para qualquer tipo de projeto de *data science*, a coleta de dos dados de fontes relevantes, considerando a natureza dos dados (estruturados e não estruturados) e a sua quantidade, dados inconsistente ou desatualizados podem invalidar os relutados.

10

(2) Armazenamento:
Após a coleta começa a fase de armazenamento, considerando o volume de dados e velocidade de acesso é feita a escolha das estratégias e plataformas de armazenamento. Para os dados estruturados, bancos de dados relacionais e bancos *NoSQL* por exemplo para dados não estruturados.
A segurança e muito importante quando se fala em dados, logo a necessidade de manter as praticas de segurança robustas como implementado criptografia e controle acesso adequados é importante para a proteção dos dados sensíveis .

(3) Tratamento dos dados:
Numa quantidade extensa de dados e preciso garantir a qualidade, consistência e relevância dos mesmos tornando esta etapa uma etapa mais critica. E necessário a limpeza, transformação e integração de dados para que estejam prontos para as analises e estatísticas. Começa pela identificação e tratamento de valor ausentes (*outliers*) e inconsistência dos dados. Pode ser feita normalização onde são lidados com novas *features* baseadas nos dados para evitar problemas nos algoritmos de *machine learning*. É aqui que são usadas também ferramentas de programação como o Python para facilitar as manipulações e transformações.

(4) Comunicar:
Por fim os *insights* obtidos são apresentados de maneira clara em relatórios ,apresentações e visualizações das descobertas mais importantes. A comunicação eficaz dos relutados é crucial para garantir que as descobertas sejam compreendidas e utilizadas.

## 5 TECNOLOGIAS DE CIÊNCIA DE DADOS

(1) Inteligência artificial:
Desempenha um papel fundamental na ciência de dados,contribui para a analise avançada, tomada de decisões automatizada e *insights*. Temos como exmplo modelos de *machine learning* onde são criados algoritmos treinados com dados para fazer previsões ou classificações. Outro exemplo é o *Deep Learning* que é uma área que utiliza redes neutrais profundas par aprender padrões complexos em conjuntos de dados.

(2) Computação nuvem:
Um modelo que disponibiliza recursos computacionais, como servidores e armazenamento, o uso de uma nuvem tem como beneficio permitir o acesso aos recursos computacionais de forma remota dividida em modelos de serviço (IaaS, PaaS, SaaS) e modelos de implantação (pública, privada, híbrida).

(3) Ambientes de Desenvolvimento Integrado:
Ambientes como o "Jupyter Notebooks", "VSCode" , entre outros são usados para facilitar a analise e visualização de dados, assim com desenvolvimento de códigos. A escolha varia consoante a necessidade do projeto.

(4) *Big Data Frameworks*:
È um conjunto de ferramentas projetadas para liderar com o processamento , armazenamento e anailse de grandes volumes de dados.

Existem muitas outras tecnologias do que aquelas aqui referidas.

## 6 ÉTICA

Esta área também requer atenção em relação a questões éticas e de privacidade, ou seja, a coleta deve estar sempre em conformidade com regulamentações e padrões éticos de forma a não envolver dados sensíveis ou informações de pessoas. Os indivíduos devem ter o direito de acederem aos seu dados, de saber o que lhes acontece, de terem preferência em relação ao facto do que pode ou não ser armazenado e ao respetivo processamento e de saber de que maneira a "Inteligência Artificial" lida e trata os dados.

Um caso bastante conhecido em Portugal foi o de Rui Pinto, um "hacker" que criou um site chamado Football Leaks no qual publicava documentos privados de personalidades e clubes como Cristiano Ronaldo e Real Madrid.

Rui pinto é considerado para alguns um hacker e deve ser punido mas para outros é uma pessoa que divulgou os podres de quem expôs, mas na realidade e de facto uma coisa é certa, este acedeu de forma ilegítima a computadores, contas de e-mail e divulgou todos os documentos a que teve acesso.

Assim conseguimos perceber que até todos nós estamos vulneráveis perante o armazenamento e tratamento dos nossos dados.

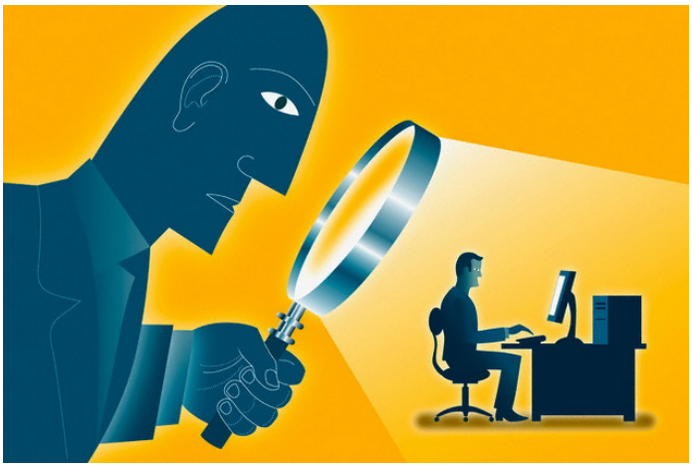

Fig. 4. Etica na internet

## 7 CONCLUSÃO

Com esta pesquisa pudemos perceber a origem da *data science*, a sua evolução, como é trabalhada e a responsabilidade necessária para trabalhar com esta área. Com a realização deste trabalho é nos possível demonstrar a importância e o cuidado necessário deste campo.

## 8 REFERÊNCIAS

ttps://www.datascienceformanagers.com/importance-of-data-science-in-society/
https://aws.amazon.com/pt/what-is/data-science/
https://awari.com.br/linguagens-de-programacao-para-ciencia-de-dados/
ttps://www.ibm.com/topics/data-science
https://pt.linkedin.com/pulse/dados-estruturados-semiestruturados-e-não-fernanda-ministerio
ttps://www.unisys.com/pt/glossary/what-is-cloud-computing/

## 17% geral de similaridade

As principais fontes encontradas nos seguintes bancos de dados:

- 16% Banco de dados da Internet
- Banco de dados do Crossref
- 7% Banco de dados de trabalhos enviados
- 5% Banco de dados de publicações
- Banco de dados de conteúdo publicado no Cross

PRINCIPAIS FONTES

As fontes com o maior número de correspondências no envio. Fontes sobrepostas não serão exibidas.

| | | |
|---|---|---|
| 1 | **awari.com.br**<br>Internet | 9% |
| 2 | **research.autodesk.com**<br>Internet | 5% |
| 3 | **Colorado State University Fort Collins on 2023-12-09**<br>Submitted works | <1% |
| 4 | **automatic-bad-pie.rockstage.io**<br>Internet | <1% |
| 5 | **The New Art College on 2006-01-05**<br>Submitted works | <1% |
| 6 | **quieora.ink**<br>Internet | <1% |
| 7 | **Denise Gomes Silva Morais Cavalcante. "Modelagem semântica de sis...**<br>Crossref posted content | <1% |
| 8 | **uri.gbv.de**<br>Internet | <1% |

| | | |
|---|---|---|
| 9 | Alex Andrade de Paula e Silva. "A voz verbal em kimbundu", Universida...<br>Crossref posted content | <1% |
| 10 | IPS Instituto Politécnico de Setubal on 2023-12-29<br>Submitted works | <1% |