Mundo Tech

PRINCÍPIOS DE *DATA SCIENCE*

SENAI <LAB365>

SUMÁRIO

"– Informações! Quero informações! – exclamava com impaciência. – Não posso fabricar tijolos sem barro." — Arthur Conan Doyle (2011, p. 266).

E assim é a *data science*, os dados são nossa matéria-prima, não podemos realizar uma análise sem os dados. Porém, os dados sozinhos, sem o cientista de dados, não são nada!

Como mencionou a revista Forbes, em um de seus artigos, "dados são o novo petróleo" — Bhageshpur/Forbes (2019, online).

ANÁLISE EM *DATA SCIENCE*

Você sabe o que é análise?

Análise consiste em uma avaliação sobre determinada matéria ou assunto, observando todos os mínimos detalhes.

A análise de dados pode oferecer algumas vantagens. Conheça algumas:

1. Antecipar necessidades da empresa, ou seja, resolver as situações antes que eles aconteçam;
2. Reduzir riscos. Como no item 1, podemos antecipar riscos e já fazer as correções necessárias;
3. Criar estratégias de serviços e vendas de produtos mais efetivas – isso olhando para nossos clientes e concorrentes;
4. Melhorar a experiência do cliente. Conhecendo melhor nosso cliente e seu perfil, poderemos oferecer melhores resultados.

Existem quatro tipos principais de análise de dados, e seu aprendizado aqui está concentrado no primeiro tipo.

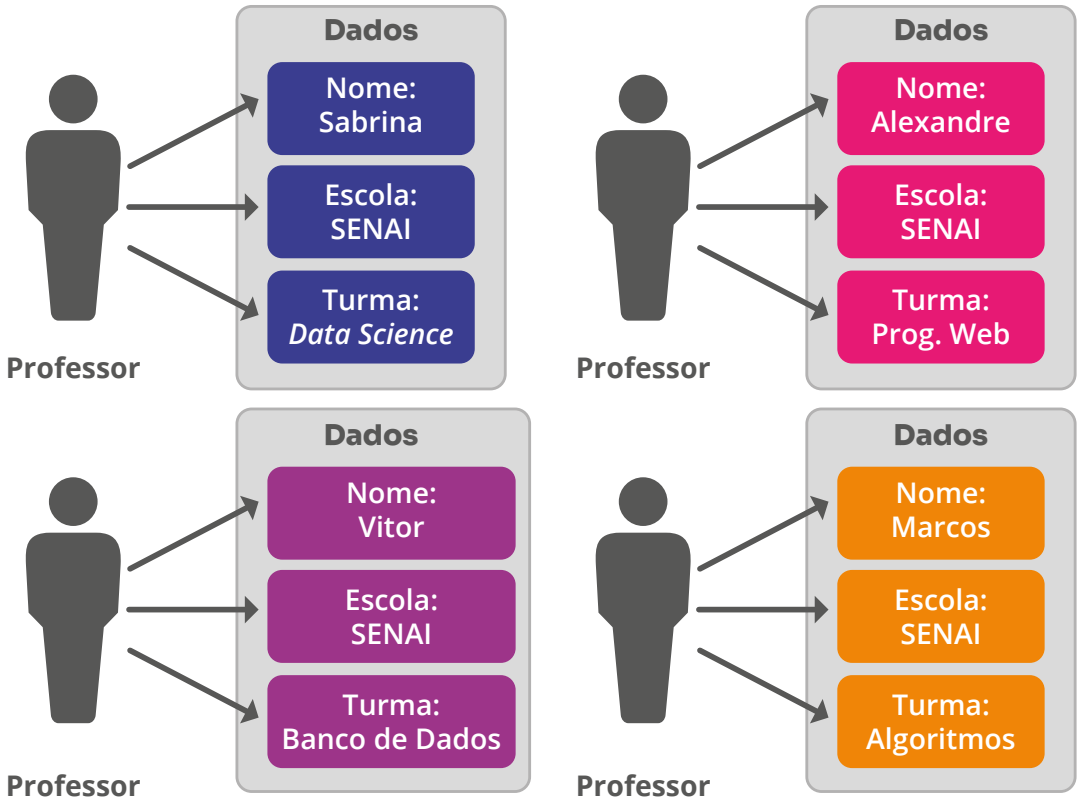
Análise Descritiva	Responde o que já aconteceu
Análise Diagnóstica	Descobre o por que isso já aconteceu
Análise Preditiva	Indica o que irá acontecer
Análise Prescritiva	Indica o que deve ser feito

DATA SCIENCE E *BIG DATA*

Atenção

Big data e *data science* não são a mesma coisa – lembre-se de que *big data* é o nome de todo o conjunto de dados que você irá utilizar em *data science*.

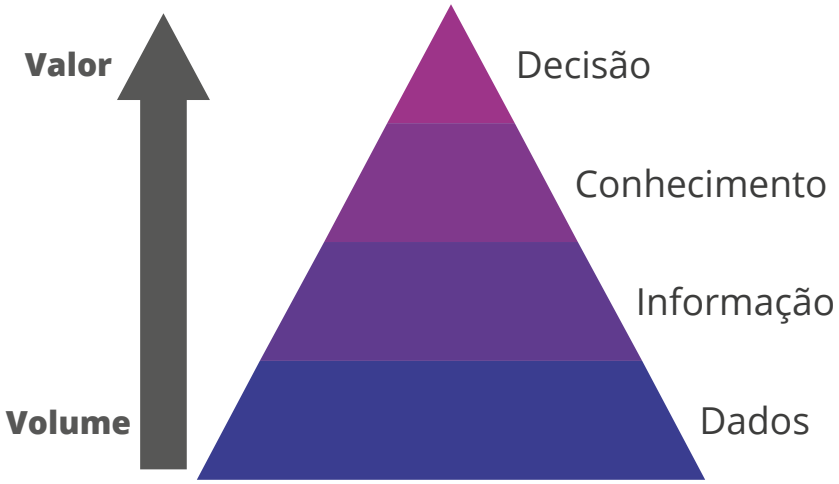
Dados são coleções de fatos, tais como números, palavras, medições, observações ou mesmo apenas descrições de coisas. São um conjunto de valores sobre pessoas, objetos ou eventos. Os dados são desorganizados – é algo ainda bruto, e seu significado é muito isolado. Confira, no exemplo a seguir, dados referentes a pessoas. Neste caso específico, são dados de professores. Na figura a seguir, é possível ver o nome da pessoa, escola e turma em que cada professor trabalha. Se você observar, cada dado é um objeto isolado.



Curiosidade

De acordo com a IDC, empresa de inteligência de mercado e consultoria em serviços estratégicos, são gerados 2,5 quintilhões de bytes de dados todos os dias.

Porém, os dados isolados não nos dizem muita coisa. Diariamente, são gerados uma enxurrada de dados; e, se você não souber o que fazer com eles, serão apenas lixo virtual. Por isso, a importância do cientista de dados – transformar tudo isso em informação; a partir dessa informação, gerar conhecimento; e, com o conhecimento, decisões podem ser tomadas. É exatamente isso que demonstra a pirâmide a seguir.



Vamos voltar ao exemplo da figura anterior, o que apresenta dados de professores: Que informações é possível buscar nesta base, mesmo que ela seja extremamente pequena? Mas, o que poderia ser informação? Informação é algo que não está visível ali; eu preciso observar bem esses dados para descobrir informações.

Por exemplo, posso perceber que o SENAI possui quatro professores, sendo um do sexo feminino e três do sexo masculino, e todos são da área de tecnologia. Essas informações são simples, porém nos trazem um retrato deste cenário reduzido.

Como levar essa informação para o grau de conhecimento? Se você observar as informações que acabou de identificar, é possível perceber que as mulheres são a minoria neste cenário de professores. Se olhar o mercado de trabalho, os

estudos apontam que as mulheres na tecnologia representam 20% do mercado – isso significa que as informações levantadas estão de acordo com o mercado. Isso é conhecimento!

E a decisão, onde está? Exatamente! O que você vai fazer com o conhecimento que acabou de criar? Como gestores de uma empresa de educação, é possível criar ações que incentivem mais mulheres para o mercado de tecnologia. Quem sabe promover cursos de TI só para mulheres? Ou, então, abrir uma seleção de professores e priorizar o ingresso de mulheres? São algumas possíveis ideias.

Mas de onde surgem os dados?

Vamos começar com um exemplo: a maioria das pessoas hoje possui uma rede social. Quando você se conecta em sua rede social, com certeza, a partir de seus dados de login, tais como data, horário e local de login, cada imagem que você clica, uma foto que você curte ou quando interage com algum post, são gerados dados e esses dados são armazenados. As músicas que você ouviu, os filmes a que assistiu... tudo gera dados que, posteriormente, serão analisados e utilizados para melhorar sua experiência, como a recomendação de um filme.

Outro exemplo é na área da saúde, com os exames e consultas médicas; tudo pode se transformar em dados.

Os dados podem vir de todos os lugares, tudo que podemos imaginar que tenha interação com um computador ou dispositivo eletrônico pode gerar dados. E esses dados que irão alimentar seu processo de análise.

Se pararmos para pensar, todos nós deixamos “rastros” ao usar a rede social, assistindo a um vídeo na internet, fazendo compras online. São muitos dados sobre o nosso consumo, sobre o que gostamos etc.

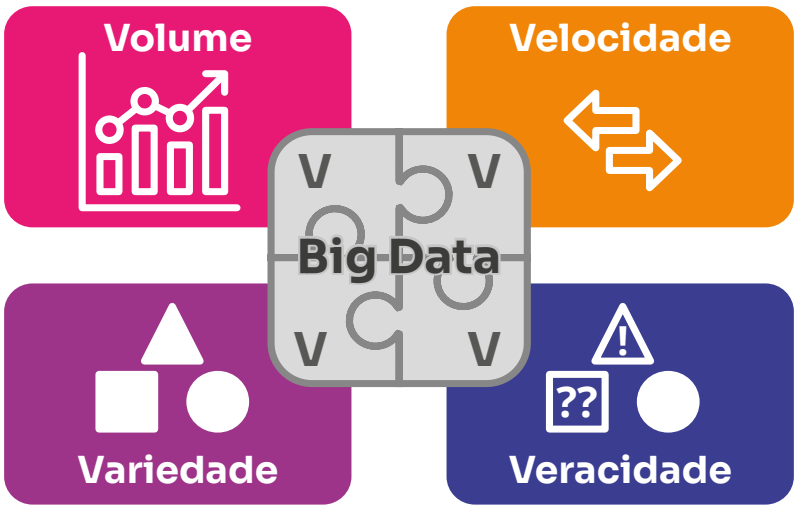
Se você multiplicar todos esses dados por bilhões de vezes, terá o *big data*!

Redes Sociais **Comércio Eletrônico** **Sistemas Diversos**

O *big data* é uma coleção de conjuntos de dados, grandes e complexos, que não podem ser processados por bancos de dados ou aplicações tradicionais.

O *big data* é definido por 4 Vs:

- **Volume:** é a quantidade de dados;
- **Variedade:** são os tipos e formatos dos dados (estruturados e não estruturados);
- **Velocidade:** os dados são gerados cada vez em maior velocidade;
- **Veracidade:** os dados devem ser dados reais; não podem ser dados fictícios.



Curiosidade

O que podem ser os dados?

Os dados podem ser qualquer caractere, texto, palavras, números, imagem, som ou vídeo.

Formato *versus* Fontes de Dados

O *big data* pode ser composto por dados de diferentes fontes e diferentes formatos. As fontes de dados podem vir de sistemas internos da sua empresa ou de fora da sua empresa (externos). As fontes de dados podem ser estruturadas – isso significa que são dados organizados; já os não estruturados normalmente são informações textuais – um e-mail, por exemplo. Confira, na imagem a seguir, exemplos de fontes de dados *versus* formatos de dados.

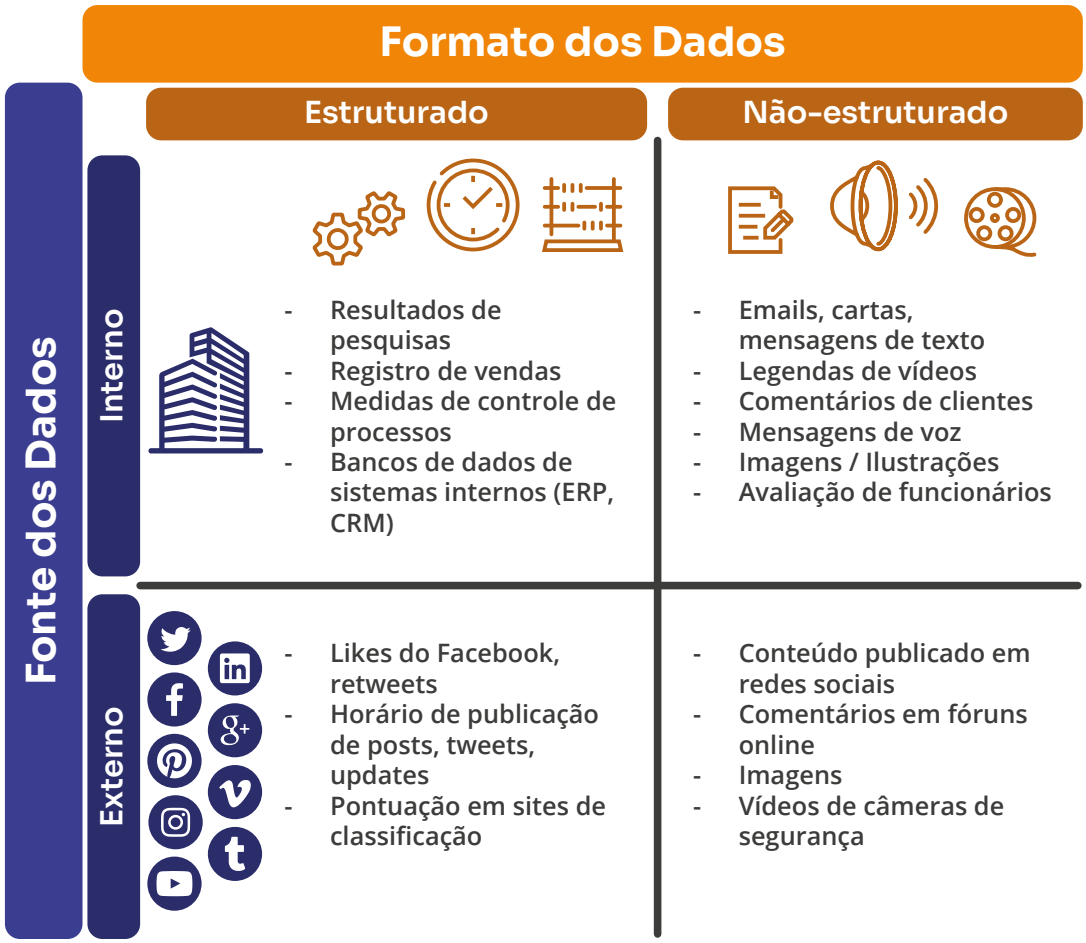


Figura 1 - Fontes de dados *versus* Formatos de dados

Fonte: Da autora (2022)

Então, agora está claro para você que *big data* foi um dos fenômenos responsáveis pelo crescimento da *data science*, pois, sem esse volume de dados, não seria possível aplicar técnicas estatísticas e gerar resultados mais precisos para um negócio.

DATASETS

Os *datasets* são conjuntos de dados estruturados. Eles estão organizados como uma tabela – ou em formato tabular – e são o principal insumo para o cientista de dados. Em um *dataset,* as linhas são os registros de acontecimentos e as colunas representam o que significa cada uma, ou seja, são os atributos. O *dataset* pode conter 1 registro (linha) e 1 atributo (coluna), e não há limite máximo, sendo que na verdade o limite está na capacidade de armazenamento e do processamento do computador.

O formato de um *dataset* varia entre CSV, TXT, XML e até XLS.

O quadro a seguir mostra um exemplo de *dataset* de uma clínica de saúde fictícia. Observe que cada coluna representa alguma característica de um determinado paciente. É muito importante observar que cada coluna tem um padrão no seu tipo de dados (conforme legenda).

COD	Nome	Idade	Sexo	Peso	Temperatura	Internações	Bairro	Diagnóstico
01	Amanda	37	F	73,5	38,0	2	Centro	Doente
02	Henrique	18	M	79,3	39,5	4	Centenário	Saudável
03	Isabela	40	F	68,0	38,0	0	Guarujá	Doente
04	João	18	M	90,1	38,5	11	Centenário	Doente
05	Mateus	50	M	78,9	37,6	2	Centro	Saudável
06	Laura	54	F	56,4	38,4	3	Guarujá	Saudável
07	Pedro	72	M	67,0	39,0	0	Centro	Doente

Legenda

TIPOS DE DADOS		
	Qualitativo	São informações com texto ou caracteres
	Quantitativo discreto	São informações com números inteiros
	Quantitativo contínuo	São informações com números decimais

Como encontrar *datasets*?

Antes de começar um projeto de desenvolvimento de *data science*, as pessoas geralmente passam por um processo de busca dos *datasets* ideais. Para isso, é preciso atender a alguns critérios de procura.

O primeiro critério é saber se você precisa de uma base pública ou de dados privados de sua empresa. Em aplicações corporativas, por exemplo, é comum que profissionais de *data science* coletem dados de sistemas internos, como ERPs[1] (*Enterprise Resource Planning)*, CRMs[2] (*Custumer Relationship Management)* ou ferramentas de marketing, atendimento e vendas. Já em projetos pessoais, estudantes tendem a buscar bases públicas.

[1]ERP: vem do termo em inglês *Enterprise Resource Planning* que significa Planejamento de Recursos Empresariais. O ERP é um sistema de gestão de empresas os quais são compostos de diversos módulos que oferecem suporte a diferentes setores dentro de uma empresa.

[2]CRM: é a sigla de *Customer Relationship Management* que significa Gestão de Relacionamento com o cliente. É uma ferramenta utilizada para o processo de vendas, que gerencia, por exemplo, diferentes pontos de contato entre vendedores e clientes. Ela tem como objetivo práticas e estratégias focadas, principalmente, em aumentar as vendas e personalizar o atendimento.

No caso de bases públicas, é possível encontrar diversas fontes interessantes na internet.

Temos, inclusive, *datasets* brasileiros e vários outros *datasets* em outros idiomas (principalmente em inglês). A vantagem de o *dataset* ser nacional é não conter termos específicos de segmentos que a pessoa cientista desconhece. Contudo, esse problema com *datasets* estrangeiros é diminuído quando há uma documentação explicando as colunas e as características (embora essa documentação nem sempre esteja disponível).

Geralmente, essas plataformas oferecem dados em formato CSV, JSON, PDF e outros. Então, a pessoa que cuida dessa área pode baixar os arquivos e fazer upload com as funções devidas.

Confira, no quadro a seguir, algumas fontes públicas de *datasets*:

Descrição	Link
Dados do Governo do Brasil	http://dados.gov.br
Instituto de Pesquisa Econômica Aplicada	http://www.ipeadata.gov.br
Dados do Governo dos EUA	http://data.gov
Dados sobre as cidades americanas	http://datasf.org
Dados da NASA	https://data.nasa.gov
Kaggle	https://www.kaggle.com/
Dados do Banco Mundial	http://data.worldbank.org
Dados sobre a saúde	http://www.healthdata.gov
Dados públicos da Amazon	http://aws.amazon.com/datasets
Dados sobre diversos países (incluindo o Brasil)	http://knoema.com
Google Trends	https://www.google.com/trends
Dados sobre milhões de músicas	https://aws.amazon.com/datasets/million-song-dataset
Dados sobre os mais diversos assuntos	http://www.freebase.com
DBpedia	http://wiki.dbpedia.org/
Open Data Network	http://www.opendatanetwork.com

Você pode perceber que *data science* é um assunto muito amplo – estamos apenas começando nesse universo. Neste material, você viu conceitos importantes para seus próximos passos. Entendeu a diferença entre *big data* e *data science*, a diferença entre dados e informação, e viu também o termo *dataset*. Aguardo você nos próximos estudos!

REFERÊNCIAS

BHAGESHPUR, K. **Data Is The New Oil** -- And That's A Good Thing. Forbes Technology Council, Jersey City (NJ, USA), 15 Nov. 2019. Disponível em: https://www.forbes.com/sites/forbestechcouncil/2019/11/15/data-is-the-new-oil-and-thats-a-good-thing. Acesso em: 20 ago. 2022.

DOYLE, A. C. **As Aventuras de Sherlock Holmes**. São Paulo: Martin Claret, 2011.

GRUS, J. **Data Science do Zero**: primeiras regras com Python. Tradução de Welington Nascimento. Rio de Janeiro: Alta Books, 2016.

HOPPEN, J.; PRATES, W. **Datasets, o que são e como utilizá-los**. Aquarela Analytics, Florianópolis (SC), 23 abr. 2018. Disponível em: https://www.aquare.la/datasets-o-que-sao-e-como-utiliza-los/. Acesso em: 13 jul. 2022.

SENAI <LAB365>