Apontamentos do Curso Data Science

INTRODUÇÃO - CIÊNCIA DE DADOS

Eduardo Alves Fonseca Queirós
GitHub: code36u4r60

Entidade Formadora: Data Science Academy
e-mail: suporte@datascienceacademy.com.br
Website: datascienceacademy.com.br/blog

Índice

1 – Introdução 1
Ciência de Dados 1
O que é Ciência de Dados? 1
Qual a importância da Ciência de Dados? 1
Por que a Ciência de Dados está crescendo? 1
Praticamos Ciência de Dados o tempo todo 1
Quem é o Cientista de Dados? 1
O que faz o Cientista de Dados? 1
Nunca o mundo gerou tantos dados 1
2 – O que é Ciência de Dados 3
Definição de Ciências de Dados 3
Um Mercado em Expansão 3
Ciência de dados 3
Resumo: 4
Áreas de Conhecimento 4
Matemática e Estatística 4
Conhecimento das áreas de negócio 4
Ciência da Computação 4
Definição: 5
Dados e Tomada de Decisão 6
O que é Data Science afinal 7
Resumo 8
Questionário 9
3 – Ciência de Dados e Big Data 11
O que são dados 11
O que são dados? 11
O que são Datasets 11
Big Data 12
Ciência de Dados e Big Data 13
Questionário 15
O que são dados? 15
4 – Ciência de Dados e Estatística 17
Definindo Estatística 17
Ciência de Dados e Estatística 17
Ciência de Dados é uma Arte 18

Aprendizado de Estatística 18
Questionário 20
5 – Aprendizado de Máquinas (Machine Learning) 21
O que é Machine Learning 21
Algoritmos de Machine Learning 22
Supervisionado 22
Não Supervisionado 23
Reinforcement Learning 23
Classificação 23
Regressão 23
Algoritmos de Machine Learning 24
Aprendizagem Profunda (Deep Learning) 24
Aplicações do Aprendizado de Máquina 24
Exemplos de aplicações de Machine Learning 24
Questionário 27
6 – Aplicações da Ciência de Dados 29
Onde aplicar a Ciência de Dados 29
Exemplos de Aplicações da Ciência de Dados 30
Walmart 30
Amazon 30
Citibank 30
Outros exemplos 30
Educação 31
Varejo 31
Telecomunicações 31
Saúde 32
Financeiro 32
Governos 32
Business Intelligence x Ciência de Dados 32
DataOps 33
Data Lake 33
Enterprise Data Hub 34
Open Data 35
Questionário 37
7 – Ciclo de Vida de Projetos de Data Science 39
Projetos de Ciência de Dados 39

Ciclo de Vida 41
Fases do projeto 41
Produtos Gerados 42
Cultura Orientada a Dados 42
Exemplos de Projetos de Data Science 43
Questionário 44
8 – Carreiras em Data Science 45
O Mercado de Ciência de Dados e Big Data 45
Carreiras em Data Science 47
Analista de Negócios 47
Analista de Dados 47
Arquiteto de Dados 48
Engenheiro de Dados 48
Administrador de Bancos de Dados 48
Estatístico 48
Cientista de Dados 49
Funções gerenciais em Data Science 50
Questionário 51
9 – Como se tornar um Cientista de Dados 52
A Profissão de Cientista de Dados 52
Quem é o Cientista de Dados 52
Características do Cientista de Dados 53
O que faz um Cientista de Dados 54
Como se preparar para a Profissão do Futuro 55
Como se tornar um Cientistas de Dados 55
O que o Cientista de Dados precisa saber? 55
Como se preparar 55
Questionário 56

1 – Introdução

Ciência de Dados

O que é Ciência de Dados?

Vivemos na era da informação. E o mundo nunca gerou tanta informação como nos dias atuais. Informação é gerada a partir de dados. Dados produzidos por cada um dos seres humanos no planeta, por máquinas, sistemas, celulares, dispositivos e muito em breve até mesmo pela sua geladeira. E a Ciência de Dados nos traz as ferramentas, métodos e tecnologias para analisar, visualizar e tomar decisões a partir dos dados.

Qual a importância da Ciência de Dados?

Imagine a quantidade de dados geradas a cada dia por uma empresa. Pedidos, vendas, pagamentos, relacionamento com os clientes, processos internos, auditorias, contabilidade, finanças, marketing, bancos de dados, e-mails, sistemas, redes sociais… Como estes dados se relacionam? E como se relacionam com o mundo externo à empresa? Como tomar melhores decisões a partir dos dados? E como fazer isso com dados gerados em tempo real?

Por que a Ciência de Dados está crescendo?

A Ciência de Dados cresce na mesma velocidade com que os dados são gerados! Novos métodos, tecnologias e processas são necessários para que se possa extrair informação valiosa da imensidão de dados.

A Ciência de Dados tem o desafio de ajudar aqueles que precisam responder as perguntas que ainda não foram feitas!

Praticamos Ciência de Dados o tempo todo

Estima-se que no futuro, todos serão Cientistas de Dados. Na prática, já somos. Recebemos toneladas de dados todos os dias, das mais variadas fontes e formatos. E cada um de nós decide o que fazer com estes dados e como transformá-los em informação útil. Nossa vida já é baseada em dados.

Precisamos apenas aprender as melhores técnicas para fazer com que os dados ajudem a tomar melhores decisões.

Quem é o Cientista de Dados?

O Cientista de Dados é o profissional responsável por aplicar técnicas, modelos, tecnologias e processos e extrair informação relevante dos dados. Quando você acede a um site de compras e recebe ofertas de produtos baseados no seu gosto, algum Cientista de Dados aplicou conhecimentos para criar um sistema que avaliasse suas preferências de acordo com os cliques que você efetuou no site.

O que faz o Cientista de Dados?

O Cientista de Dados é responsável por extrair grandes volumes de dados de múltiplas fontes internas e externas e empregar tecnologias sofisticadas de análise, aprendizado de máquina e métodos estatísticos para preparar os dados para uso em modelagem preditiva e prescritiva, além de explorar e analisar dados de uma variedade de ângulos a fim de determinar fraquezas escondidas, tendências e/ou oportunidades.

Nunca o mundo gerou tantos dados

Neste exato momento, uma verdadeira enxurrada de dados, ou 2.5 quintilhões de bytes por dia, estão sendo gerados por indivíduos, empresas e governos – e está dobrando a cada dois anos.

A demanda por profissionais de dados está decolando

É preciso pensar fora da caixa

> Até 2018, haverá um déficit superior a 200 mil profissionais com habilidades de análise de dados e mais de 1,5 milhão de gerentes e analistas que saibam usar Big Data e Ciência de Dados de forma efetiva para tomada de decisões.
>
> -McKinsey Global Institute “Big Data and Data Science Report 2015”

2 – O que é Ciência de Dados

Definição de Ciências de Dados

Enquanto nossas vidas continuam **migrando para a internet**, produzimos um fluxo constante e exaustivo de informação digital.

Estima-se que 90% dos dados armazenados no mundo foram produzidos apenas nos últimos dois anos e os rastros desses dados continuam duplicando a cada ano.

O termo "Ciência de Dados" ou Data Science no termo em inglês, tem sido muito usado em notícias recentes e por uma boa razão.

É uma das áreas com maior crescimento atualmente

Um Mercado em Expansão

Um estudo realizado pela OBS (Online Business School) revela que, no período de 2004 a 2014, mais dados foram gerados do que em toda a história da humanidade. Logo, informação tornou-se a moeda mais poderosa no mundo dos negócios e exige que o mercado saiba interpretá-la a seu favor.

Em pesquisa divulgada pela Computerworld, 20% das empresas afirmam planejar a contratação de profissionais de Big Data e Data Science neste ano – gerando nada menos que 4,4 milhões de empregos em todo mundo. Segundo previsão do Gartner, 500 mil vagas para profissionais de Big Data e Data Science serão abertas no Brasil em 2016.

"Decisões baseadas em emoções não são decisões. São instintos"

Ciência de dados

É o termo usado para definir a extração de insights de dados que são coletados de várias fontes.

Utilizando várias técnicas, incluindo modelagem preditiva, a Ciência de Dados ajuda a analisar e interpretar grandes quantidades de dados.

As pessoas que trabalham com Ciência de Dados são chamadas Cientistas de Dados, porém muitas outras carreiras estão associadas a Data Science, como veremos mais adiante.

O termo Ciência de dados surgi-o na EMC (www.emc.com) e vem ganhando popularidade a cada dia.

Ciência de Dados é o processo para extrair informação valiosas a partir de "dados". Como estamos vivendo na era do Big Data, a Ciência de Dados está se tornando um campo muito promissor para explorar e processar grandes volumes de dados gerados a partir de várias fontes e em diferentes velocidades.

Ciência de dados é uma grade disciplina em si e consiste em conjunto de habilidades especializadas, tais como: estatística, matemática, programação, computação e conhecimento de negócios, além de técnicas e teorias, como a análise preditiva, modelagem, engenharia e mineração de dados e visualização.

A Ciência de Dados tem estado entre nós há bastante tempo, sob a forma de análise de negócios ou inteligência competitiva, mas somente agora o seu verdadeiro potencial foi percebido e isso se deve em parte ao Big Data.

O principal objetivo da Ciência de Dados é extrair e interpretar os dados de forma eficaz e apresenta-lo em uma linguagem simples e não técnica para os utilizadores finais e tomadores de decisão.

Assim, a Ciência de Dados é tudo aquilo relacionado sobre a construção de informações úteis e a capacidade de converter etas informações em produtos “data-driven”!

Resumo:

Em poucas palavras, a Ciência de Dados é um conjunto de métodos, técnicas usados para analisar vasto volume de dados e transformá-lo em insights significativos e acionáveis.

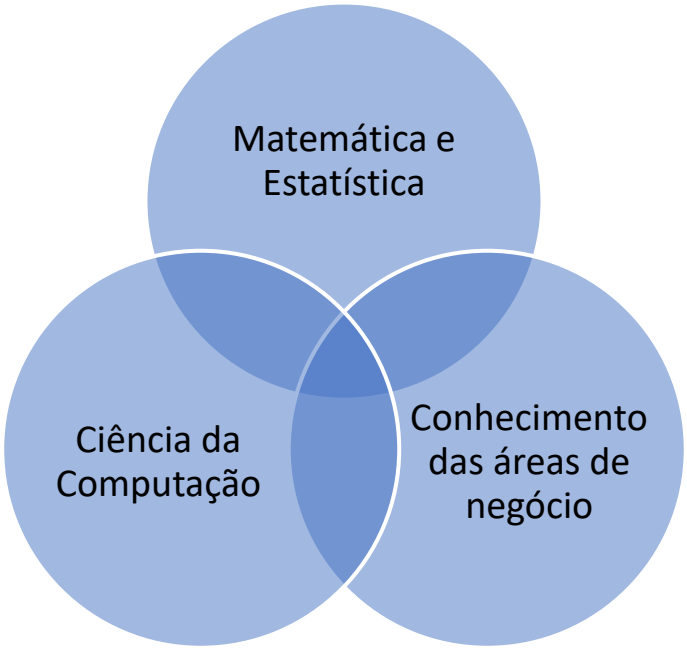
Os dados vêm das mais variadas fontes – literalmente cada assunto no planeta Terra.

A Ciência de Dados usa a mineração e análise de dados para gerar inteligência de negócios, o que é muito interessante para empresas no mundo todo.

Não apenas interessante. Atualmente, uma questão de sobrevivência!

Áreas de Conhecimento

A Ciência de Dados em si é uma área interdisciplinar, que envolve uma serie de áreas de conhecimento.



Matemática e Estatística

A matemática e estatística criam a base necessária para a criação de modelos de dados.

Conhecimento das áreas de negócio

É daqui que normalmente surge a necessidade da solução para problemas específicos.

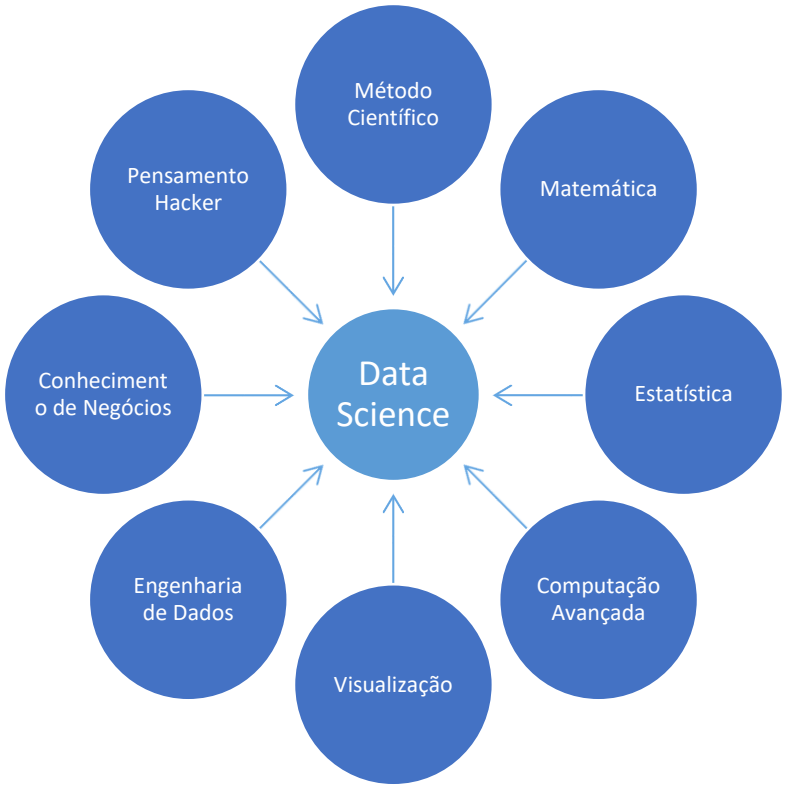
Ciência da Computação

Neste caso estamos a falar de programação de computadores, infraestruturas de banco de dados, armazenamento, segurança e governança de dados. A Ciência de Computação,

principalmente a programação de computadores, vais fornecer as ferramentas necessárias para analise além de permitir automatizar o processo de analise de dados.

Definição:

Na sua essência, a Ciência de Dados envolve o uso de métodos automatizados (ciência da computação) para analisar enormes quantidades de dados (estatística) e para extrair conhecimento (áreas de negócio) a partir deles.



Ciência de Dados é a exploração e análise de todos os dados disponíveis, sejam eles estruturados ou não, com o objetivo de desenvolver compreensão, extrair conhecimento e formular ações que gerem resultados.

A Ciência de Dados está ajudando a criar novos ramos da ciência e influenciando áreas das ciências sociais e humanas.

E esta tendência deve acelerar nos próximos anos com dados de sensores móveis, instrumentos sofisticados, o crescimento da internet e muito mais.

Cada vez mais, veremos as universidades oferecendo disciplinas relacionadas a análise de dados, desde de cursos de Engenharia e Ciência da Computação, até Direito e Pedagogia.

A análise de dados estará presente em todas as áreas de conhecimento.

Dados e Tomada de Decisão

Existe um número cada vez um maior numero de empresa que já perceberam que a tomada de decisões baseadas em dados, é o caminho que pode levar a uma vantagem competitiva num cenário cada vez mais desafiador.

O principal objetivo da Ciência de Dados é extrair informação e conhecimento dos dados. Mas o passo seguinte, a tomada de decisão, é que vai definir o sucesso ou fracasso de Data Science. Por mais que se fale muito nas decisões orientadas a dados existe um processo ainda mais além que é a execução, ou seja, a ação em si.

Exemplo:

Problema identificado: Clientes cancelam seus planos de telefonia com frequência de 2 a 3 meses após assinatura do plano.

1. Dados e Análise – Dados coletados e analisados, levaram à conclusão que existe uma tendência entre os clientes de operadoras telefónicas em trocar de operadora após o terceiro contato com entendimento ao cliente.
2. Decisão – Os executivos da empresa, baseados em dados, decidiram rever todos os processos de atendimento ao cliente.
3. Ação – a empresa contratou uma consultoria para fornecer treinamento especializado de atendimento ao cliente, passou a monitorar as ligações, criou programa de incentivo aos melhores operadores, etc…

> O principal objetivo é extrair informações dos dados e transformá-las em conhecimento que possa ser usado para tomada de decisões

Entre a coleta dos dados e as ações existe um caminho ser percorrido. Neste caminho esta a analise dos dados, que não pode ser algo sem direção. Para isso devemos fazer perguntas aos dados. Tais como:

- O que aconteceu?
- Por que aconteceu?
- Acontecerá novamente?
- O que deve ser feito?

Estas perguntas ajudaram a definir que decisões precisam ser tomadas.

O valor do processo de tomada de decisão esta em todas as etapas. Por mais que se tenha processos avançados de ciência de dados, as decisões e ações que serão tomadas poderão

decretar o sucesso ou fracasso de qualquer projeto. Por isso o Cientista de Dados deve entender do modelo de negócio para o qual esta a fazer o trabalho de analise.

A Gestão de decisões orientadas por dados é uma abordagem para a governança empresarial, que valoriza decisões que possam se apoiadas totalmente por dados.

E a abordagem orientada a dados está ganhando popularidade dentro das empresas, à medida que a quantidade de dados disponíveis aumenta, em conjunto com as pressões do mercado.

A Gestão de decisões orientadas por dados é normalmente um meio de ganhar uma vantagem competitiva.

Um estudo do **MIT Center for Digital Business** descobriu que as organizações data-driven (orientados por dados) para tomada de decisões, tiveram índices de produtividade 4% mais altos e os lucros 6% mais elevados. Em mercados de milhões ou bilhões esses números fazem muita diferença.

Mas o sucesso da abordagem orientada a dados é dependente da qualidade dos dados coletados e da eficácia da sua análise e interpretação.

O que é Data Science afinal

- Identificar o problema da área de negócio

Todo o projeto de Data Science deve começar com uma questão de negócio a ser resolvida. É este o objetivo primário de qualquer projeto de Data Science. Por isso é tão importante que os profissionais de dados possam ir além do conhecimento técnico. É fundamental o conhecimento de negócio quem pretende trabalhar como cientista de dados.

- Compreender o problema (entidades e atributos)

Que entidades e atributos estão envolvidos no problema a ser resolvido. Uma compreensão clara da questão ajuda na coleta correta dos dados e na seleção das tecnologias a ser usadas. Dependendo do problema, uma tecnologia poderá ser melhor que outra e gerar resultados mais eficientes.

- Coletar conjuntos de dados (data sets), que representem a entidade

Os projetos Data Science começam com as questões de negócio a serem resolvidas. Sem um objetivo claro a ser alcançado fica difícil saber que dados devem ser usados, e a partir daí, todo o processo estar comprometido. O volume de dados disponíveis é muito grande, a compreensão do problema ira ajudar a definir que dados são relevantes.

- Limpar e transformar os dados

Uma vez que os dados tenham sido obtidos é hora de limpa-los e transforma-los. Não espere que os dados cheguem prontos, não chegaram. Uma boa parte do tempo da analise de dados é gasto em limpeza e transformação, e este trabalho é crucial. É bem provável que uma boa parte dos dados estejam em falta.

- Compreender os relacionamentos entre os dados

É fundamental uma boa compreensão do relacionamento entre os dados. É onde a experiencia de negócio mostra o seu deferencial.

- Criar modelos que representem os relacionamentos

É nesta fase que se utiliza bastante analise estatística e modelos matemáticos.

- Utilizar os modelos para fazer predições

É aqui que começa a fase de automação do processo. A utilização da tecnologia correta, seleção de uma linguagem de programação ou ferramenta, permitirá transformar o modelo de predição em algo automático que poderá ser colocado em produção.

- Entregar valor e resultado

É preciso medir se o objetivo foi alcançado e se o problema identificado no inicio agora possui uma solução. Que representa não apenas o passado, mas que seja capaz de determinar comportamentos futuros a medida que novos dados vão chegando.

Resumo

Normalmente, o processo compreende várias etapas, começando por uma questão de negócio a ser resolvida. Uma vez que você sabe o que se quer analisar, é preciso obter os dados corretos, limpá-los, explorá-los, criar e avaliar um modelo, repetir este ciclo algumas vezes, e finalmente, você está pronto para começar a procurar uma maneira de como comunicar adequadamente seus resultados.

Questionário

1. O que é ciência de dados?
    a. é o termo usado para definir a extração de insights de dados que são coletados de várias fontes.
    b. inclui modelagem preditiva e ajuda a analisar e interpretar grandes quantidades de dados
    c. é o processo para extrair informações valiosas a partir de dados.
    d. seus profissionais são chamados Cientistas de Dados.
    e. todas as afirmativas acima estão corretas.

2. Ciência de dados é uma grande disciplina em si e consiste em conjuntos de habilidades especializadas, tais como.
    a. estatística, história da filosofia, computação e conhecimento de negócios, além de técnicas e teorias, como a análise preditiva, modelagem, engenharia e mineração de dados e visualização.
    b. estatística, matemática, programação, computação e conhecimento de negócios, além de técnicas e teorias, como a análise preditiva, modelagem, engenharia e mineração de dados e visualização.
    c. matemática espacial, programação de negócios, computação, além de técnicas e teorias, como a análise esportiva, modelagem de espaços, engenharia e mineração de dados e visualização.
    d. estatística inferencial, matemática, física quântica, programação, computação e conhecimento de negócios, mas não inclui técnicas e teorias, como a análise preditiva, modelagem, engenharia e mineração de dados e visualização.

3. Marque a alternativa correta em relação à Ciência de Dados:
    a. a Ciência de Dados é um conjunto de métodos, técnicas e teorias usados para analisar apenas grande volume de dados e transformá-lo em insights não significativos e direcionáveis.
    b. a Ciência de Dados é um conjunto de métodos, técnicas e teorias usados para analisar apenas um pequeno volume de dados e transformá-lo em insights significativos e acionáveis.
    c. a Ciência de Dados é um conjunto de métodos, técnicas e teorias usados para analisar apenas um pequeno volume de dados e transformá-lo em insights não significativos e acionáveis.
    d. a Ciência de Dados é um conjunto de métodos, técnicas e teorias usados para analisar vasto volume de dados e transformá-lo em insights significativos e acionáveis.

4. Como estamos vivendo na era do Big Data, a Ciência de dados está se tornando um campo muito promissor para explorar e processar grandes volumes de dados gerados a partir de várias fontes e em diferentes velocidades.
    a. Verdadeiro
    b. Falso

5. A Ciência de Dados surgiu recentemente, sob a forma de análise de negócios ou inteligência competitiva, o seu verdadeiro potencial foi percebido e isso se deve em parte ao Big Data.

a. Verdadeiro
b. Falso - A Ciência de Dados tem estado entre nós há bastante tempo, sob a forma de análise de negócios ou inteligência competitiva, mas somente agora o seu verdadeiro potencial foi percebido e isso se deve em parte ao Big Data.

3 – Ciência de Dados e Big Data

O que são dados

Existe diversos elementos geradores de dados, basicamente qualquer coisa gera dados hoje em dia. Exemplos:

- Procedimentos Médicos
- Mídias Sociais
- Noticias de Jornais
- Imagens por satélite
- E-commerce
- Tv & Vídeo
- Sensores e Sistemas de Monitoramento

Esses dados precisão ser coletados. A **corelação** entre eles precisa ser identificada. Realiza-se então a **análise** e gera-se **Insights** para resolver problemas de negocio.

A *Oracle* estima que já em 2015 atingimos 300 Exabytes de dados, dos quais 88% dos dados são não estruturados.

Em 2012 havia no planeta 2.8 Zettabytes de dados até 2020 será superior a 140 Zettabytes.

O que são dados?

Dados são coleções de fatos, tais como números, palavras, medições, observações ou mesmo apenas descrições de coisas.

A maioria das pessoas acredita que o termo **dados** e **informação** são intercambiáveis e significa a mesma coisa, no entanto, há uma diferença distinta entre as duas palavras. Os dados podem ser qualquer personagem, texto, palavras, números, imagens, som ou vídeo e, se não colocados em contexto, significam pouco ou nada para um ser humano.

No entanto, a informação é útil e formatada de uma maneira geral, permite que seja entendida por um ser humano.

E então, o que são dados?

- Dados representam **Entidades**.
- Dados possuem **Características**.
- Dados estão dentro de um **Ambiente**.
- Dados são alvo de **Eventos**.
- Dados possuem **Comportamento** específicos.
- Dados normalmente geram algum tipo de **Resultado**.
- Todo isso é coletado através da **Observação**.

O que são Datasets

Datasets são conjuntos de dados. Um dataset normalmente é uma coleção de algumas observações realizadas sobre o conjunto maior de dados.

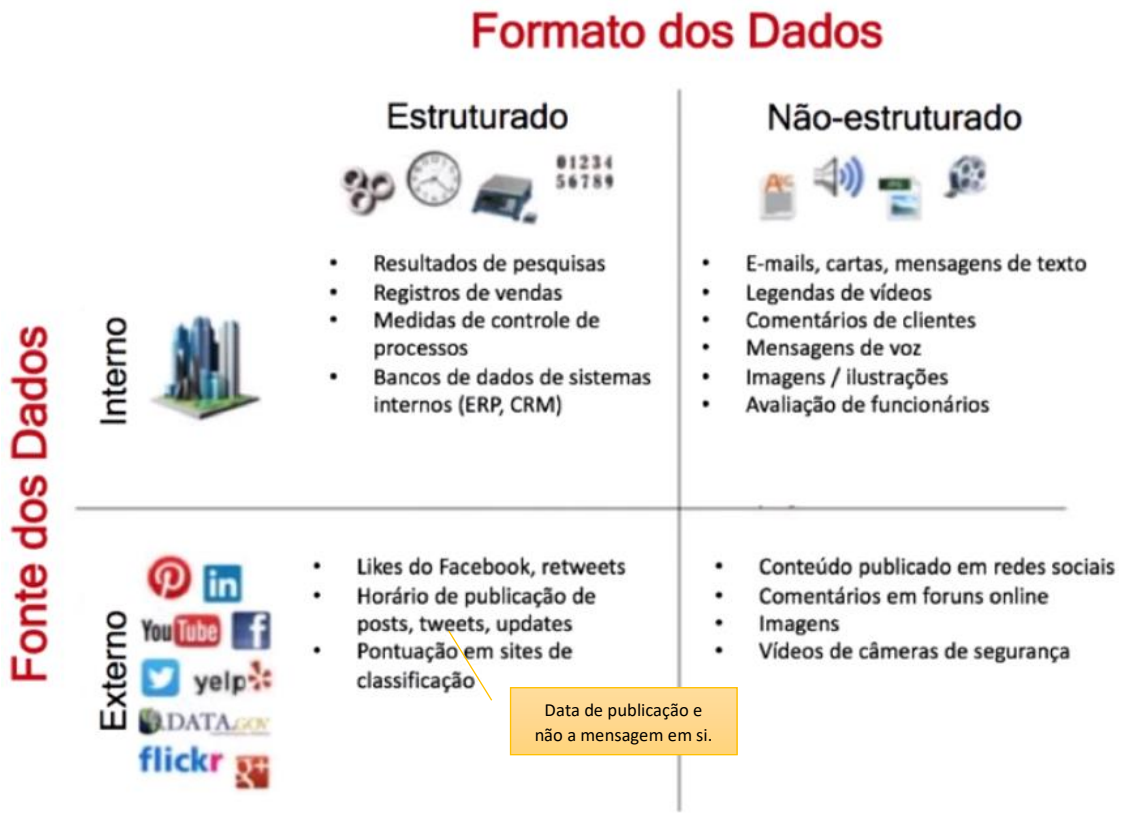
Cada observação é tipicamente chamada de registro.

Cada registro tem um conjunto de atributos que apontam características, comportamentos ou resultados.

Os **datasets** podem ser classificados em três tipos principais:

- Estruturados. Exemplo: Base de dados
- Semiestruturados. Exemplo: mail
- Não estruturados. Exemplo: Twitter

Cientistas de dados coletam e utilizam datasets para aprender sobre entidades e prever seu comportamento futuro e possíveis resultados!



Big Data

Big Data é mais que uma palavra da moda no mundo dos negócios. Todo mundo deixa "rastro" quando cria uma conta no Facebook, abre conta em banco, faz compras em supermercado, etc. São dados como frequência de compra, consumo médio, número de curtidas em posts de determinado assunto, tweets. Multiplique isso bilhões de vezes e você começa a perceber o que é o Big Data.

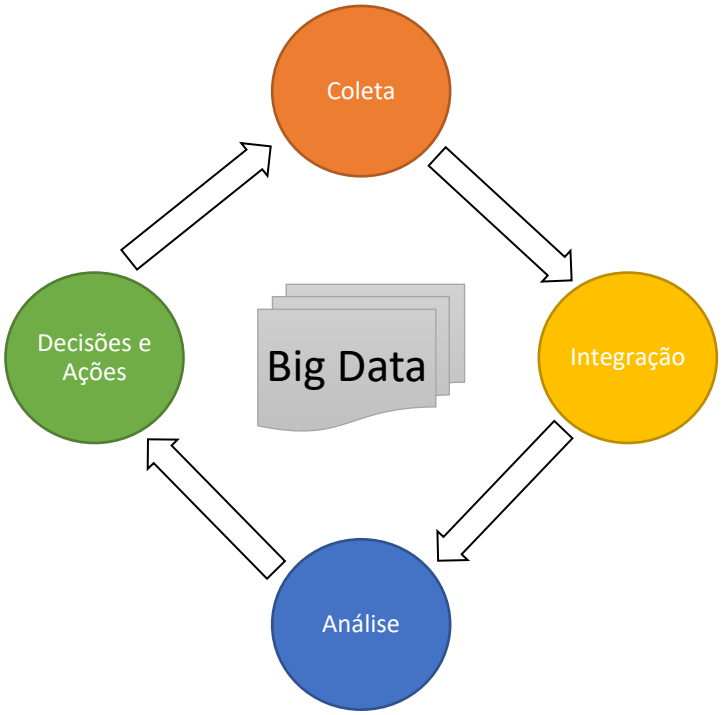
Principalmente devido à internet, a geração de dados cresceu de maneira exponencial nos últimos anos. Esses dados, podem ser armazenados e utilizados de diversas formas no varejo (para criar promoções mais direcionadas, por exemplo), por empresas telefónicas, bancos e até mesmo para previsão do tempo e de fenômenos naturais.

Big Data é uma coleção de conjuntos de dados, grandes e complexos, que não podem ser processados por bancos de dados ou aplicações de processamento tradicionais.

Empresas de diversos portes estão descobrindo novas maneiras de coletar e usar mais dados e os esforços para tornar o Big Data mais popular serão cada vez mais intensos.

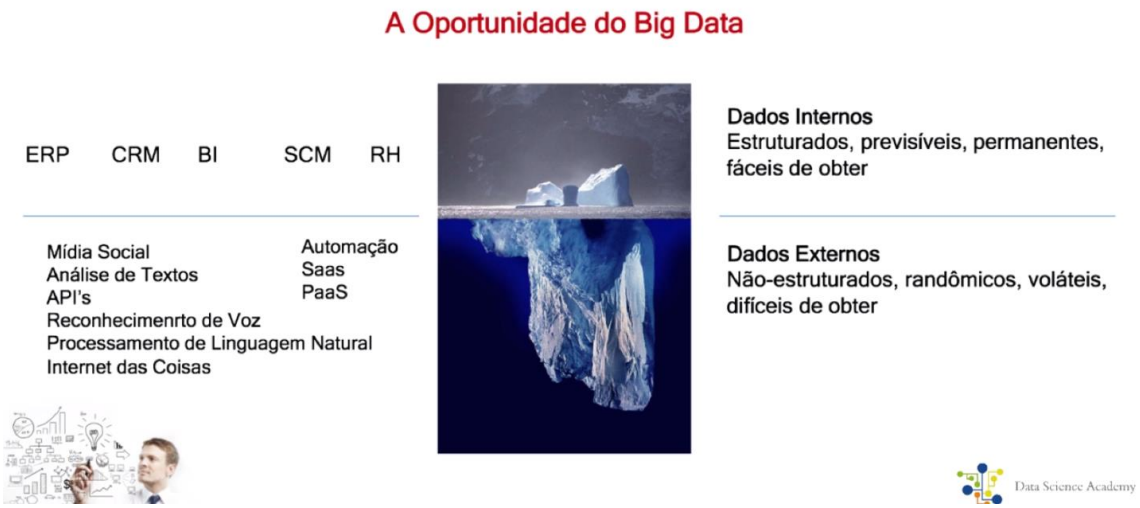
Agora que as organizações já têm mais conhecimentos do que é o Big Data, as questões chaves mudaram para quais são as estratégias e as habilidades necessárias e como é possível medir o ROI (Retorno Sobre Investimento).

Normalmente o Big Data envolve quarto etapas fundamentais.



Ciência de Dados e Big Data

A melhor forma de explicar a relação entre Ciência de Dados e Big Data é exatamente esta, Big Data é o combustível para Ciência de Dados.



E por que é importante aprender a analisar o Big Data, utilizando a Ciência de Dados?

> Até 2018, haverá um deficit de 140 a 190 mil profissionais com habilidades em análise de dados e mais de 1,5 milhões de gerentes e analistas que saibam usar Big Data de forma efetiva para tomada de decisões. - McKinsey Global Institute “Big Data Report 2015”

Nós tendemos a superestimar o efeito de uma nova tecnologia no curto prazo e subestimar o feito no longo prazo.

Questionário

O que são dados?

a. Dados são coleções de fatos, tais como números, palavras, medições, observações ou mesmo apenas descrições de coisas.
b. Dados são coleções de fatos, tais como números, palavras, medições, informações ou mesmo descrições de coisas.
c.
d. Dados são coleções de informações, tais como números, palavras, medições, observações ou mesmo apenas descrições de coisas.
e. Dados são coleções de relatórios com números, palavras, medições, observações e descrições de coisas.

2. Que tipos de dados fazem parte do nosso dia a dia atualmente?
    a. notícias de jornais e imagens por satélite.
    b. e-commerce, TV & vídeo.
    c. medias sociais e procedimentos médicos.
    d. sensores e sistemas de monitoramento.
    e. todas as alternativas acima estão corretas.

3. Segundo a Oracle, quantos por cento dos dados, hoje em dia, são do tipo "não estruturados"?
    a. 90%
    b. 12%
    c. 88%
    d. 68%

4. Qual a estimativa da quantidade de dados gerados até o ano de 2020?
    a. 14 Zettabytes.
    b. 140 Zettabytes.
    c. 2.8 Zettabytes.
    d. 280 Zettabytes.

5. Resultados de pesquisas, registro de vendas, banco de dados de sistemas internos são classificados como:
    a. Formato de dados estruturado e interno.
    b. Formato de dados não estruturado e interno.
    c. Formato de dados estruturado e externo.
    d. Formato de dados não estruturado e interno.

6. Procedimentos médicos, registro de prontuário, informação sobre medicamento, dados de cirurgia, são exemplos de dados gerados todos os dias.
    a. Verdadeiro
    b. Falso

7. As medias sociais não geram dados.
    a. Verdadeiro
    b. Falso

4 – Ciência de Dados e Estatística

Definindo Estatística

O termo Estatística provém da palavra Estado e foi utilizada originalmente para denominar levantamentos de dados, cuja finalidade era orientar o Estado em suas decisões.

Neste sentido, a Estatística é utilizada desde épocas remotas para determinar o valor dos impostos cobrados dos cidadãos ou para determinar a estratégia de uma nova batalha em guerras que se caracterizavam por uma sucessão de batalhas (era fundamental aos comandantes saber de quantos homens, armas, cavalos, etc… dispunham após a última batalha).

Estatística envolve a coleta, classificação, resumo, organização, análise e interpretação dos dados.

Estatística é um conjunto de métodos e processos quantitativos que serve para estudar e medir fenômenos coletivos.

A Estatística tem importante papel no pensamento crítico, seja no trabalho, na pesquisa ou no dia-a-dia.

A Estatística trata dados. Todo dado se refere a uma varável. Portanto, a Estatística trabalha com variáveis, que podem assumir diferentes valores, em diferentes unidades.

Ciência de Dados e Estatística

As teorias estatísticas servem como base para tudo, desde a criação de perfis de passageiros em uma era de ameaças terroristas até à eficácia de novos programas para reduzir a taxa de erros hospitalares.

A Estatística pode ajudar a avaliar se o sucesso de um fundo de investimentos é genuíno ou devido ao acaso, pode ajudar a prever se um determinado assinante vai cancelar sua assinatura este ano ou se uma reivindicação de seguro é fraudulenta.

Vivemos uma era em que a ciência deve prevalecer sobre o empirismo, em que a lógica deve prevalecer sobre o "achismo".

A Ciência de Dados difere das Análise Estatísticas em seu método, que é aplicado a dados coletados usando princípios científicos.

A razão para a necessidade crescente desta nova abordagem está relacionada ao Big Data, que demanda o uso de diferentes tecnologias à análise estatística.

A **American Statistical Association (ASA)** divulgou recentemente uma declaração sobre o papel da Estatística na Ciência de Dados. O Presidente da ASA, David Morganstein, deu esta declaração no seu comunicado de imprensa:

"Através desta declaração, a ASA e seus membros reconhecem que a ciência de dados abrange mais do que estatísticas, mas ao mesmo tempo também reconhece que a ciência estatística desempenha um papel fundamental no rápido crescimento deste campo. É nossa esperança que esta declaração possa reforçar a relação de estatística para a ciência de dados e ainda fomentar relacionamentos mútuos de colaboração entre todos os contribuintes na ciência de dados."

A declaração evidencia que Estatística é fundamental para a Ciência de Dados, juntamente com gestão de banco de dados, sistemas distribuídos e paralelos, computação, matemática e

programação. A sua Utilização neste campo emergente, capacita pesquisadores para extrair conhecimento e obter melhores resultados de grandes projetos. A declaração também incentiva a colaboração máxima e multifacetada entre estatísticos e cientistas de dados para maximizar o potencial da ciência de dados.

Em resumo, a Estatística desempenha um papel fundamental dentro da Ciência de Dados. Porém, a Ciência de Dados compreende outras áreas de conhecimento, como já vimos anteriormente.

A Ciência de Dados, utiliza Estatística, aprendizado de máquina e gerenciamento de banco de dados para criar um novo conjunto de ferramentas para aqueles que trabalham com dados.

Os Cientistas de Dados possuem 3 características principais:

1. Eles têm um forte conhecimento de Estatística e aprendizado de máquina, pelo menos o suficiente para evitar má interpretação de correlação e causalidade.
2. Eles têm habilidades de informática para usar uma linguagem de programação (como R ou Python) para facilitar o trabalho de análise.
3. Eles podem visualizar e resumir seus dados e sua análise de uma maneira que seja significativa para alguém menos familiarizado com os dados, baseado em sua experiência de áreas de negócio.

A Estatística é parte fundamental do trabalho do Cientista de Dados

Ciência de Dados é uma Arte

A Ciência de Dados não é apenas uma ciência ou uma técnica de ouvir as suas intuições, enquanto enfrenta enorme quantidade de dados, classificando, avaliando e buscando as conclusões.

Cientistas de Dados precisam ser realmente criativos em visualizar os dados de várias formas gráficas e apresentar os dados altamente complexos de uma forma muito simples e amigável!

Se um Cientista de Dados é capaz de converter aterrorizantes Petabytes de dados estruturados, bem como dados não-estruturados (imagens, vídeos, arquivos de log, etc…) em um formato mais fácil e simples, ele é um – **“artista!”**

Afinal, apenas um Cientista de Dados mais hábil pode gerenciar o banco de dados do McDonald’s, gerar insights dos bilhões de vídeos carregados no Youtube, gerenciar grandes volumes de dados da GE ou ainda gerenciar os dados relativos aos milhares de amostras de sangue de pacientes ou dados não estruturados gerados a partir de raios-x!

Aprendizado de Estatística

Para muitas pessoas, o aprendizado de Estatística pode ser meio obscuro. Isso se deve aos métodos de ensino, muitas vezes artificiais e focados apenas em fórmulas matemáticas, sem qualquer ligação com o mercado de trabalho ou com as nossas vidas diárias.

Entretanto, o rápido crescimento da Ciência de Dados e da análise de dados em geral, colocou a Estatística no centro das atenções, trazendo novas formas de ensinar e aprender Estatística.

Se você não entender o mundo da estatística em torno de você, você realmente não sabe como as coisas funcionam. A Estatística pode ajudar a compreender questões reais como:

- Assistir TV realmente provoca comportamento violento?
- Como o sexo de uma pessoa pode afetar seus ganhos?
- Qual o melhor momento do dia para exercitar e perder peso?

Em uma análise estatística o contexto é vital e muito interessante.

Você precisa entender o problema que deu origem à investigação e coleta de dados (por isso o conhecimento de áreas de negócios é tão importante). O contexto é o que faz cada investigação estatística diferente.

Os Cientistas de Dados muitas vezes trabalham ao lado de outros pesquisadores em áreas como Medicina, Psicologia, Biologia, Geologia, Marketing, Finanças, Contabilidade, E-Commerce, etc…, que fornecem o plano de fundo contextual para o problema.

Questionário

1. Qual a definição de estatística?
    a. Estatística envolve a distribuição, reclassificação, resumo, organização, análise e interpretação dos dados. Bem como ela é um conjunto de métodos e processos qualitativos que serve para estudar e medir fenômenos coletivos.
    b. Estatística envolve dados matemáticos e numéricos, utilizado para realizar um resumo histórico, e fazer a análise e interpretação dos dados qualitativos. Bem como ela é um segmento de métodos e processos administrativos que serve ampliar os recursos da Ciência de Dados.
    c. Estatística envolve somente dados estatístico e não tem nenhuma correlação com Ciência de dados.
    d. Estatística envolve a coleta, classificação, resumo, organização, análise e interpretação dos dados. Bem como ela é um conjunto de métodos e processos quantitativos que serve para estudar e medir fenômenos coletivos.

2. Sobre tratamento de dados no âmbito da Estatística é correto afirmar:
    a. A Estatística trabalha com variáveis, que podem assumir diferentes valores, em diferentes unidades.
    b. A Estatística trata dados.
    c. Os dados estatísticos fazem referência a uma variável.
    d. Todas as afirmações estão corretas.

3. As estatísticas sempre “mentem” apresentando dados errados ou, no mínimo, mal interpretados.
    a. Verdadeiro
    b. Falso

5 – Aprendizado de Máquinas (Machine Learning)

O que é Machine Learning

O avanço da automação e, mais recentemente, do conjunto de tecnologias que chamamos de "Machine Learning", vai mudar em muito as profissões atuais.

O impacto da robotização chegando às áreas de conhecimento muda nossa perceção sobre a automação. Antes era consenso que automação afetaria apenas as atividades operacionais, como as linhas de produção.

Mas agora percebemos que podemos vê-la atuando em atividades mais intelectuais do que manuais, que envolvem tomadas de decisão e que tradicionalmente abrange pessoas com formação universitária.

Machine Learning (ou Aprendizado de Máquina) é uma das tecnologias atuais mais fascinantes.

Você provavelmente usa algoritmos de aprendizado várias vezes por dia sem saber.

Sempre que você usa um site de busca como "Google" ou "Bing", uma das razões para funcionarem tão bem é um algoritmo de aprendizado.

Toda a vez que você usa o aplicativo para "marcar" pessoas nas fotos do "Facebook", e ele reconhece as fotos dos seus amigos, isto também é Machine Learning.

Toda a vez que o filtro de spam dos eu email filtra toneladas de mensagens indesejadas, isto também é um algoritmo de aprendizado.

Basicamente a ideia do Machine Learning é extrair conhecimento a partir dos dados. Ensinamos a maquinas a gerar inteligência a partir da analise de dados.

Machine Learning é uma coleção de métodos de modelagem estatística aplicada nos mais diversos campos.

Machine Learning pode ser aplicada em problemas quantitativos (números) e qualitativos (grupos de dados).

Machine Learning é um subcampo dentro da Inteligência Artificial que constrói algoritmos que permite que os computadores possam aprender a executar tarefas a partir de dados, ao invés de serem programados de forma explícita.

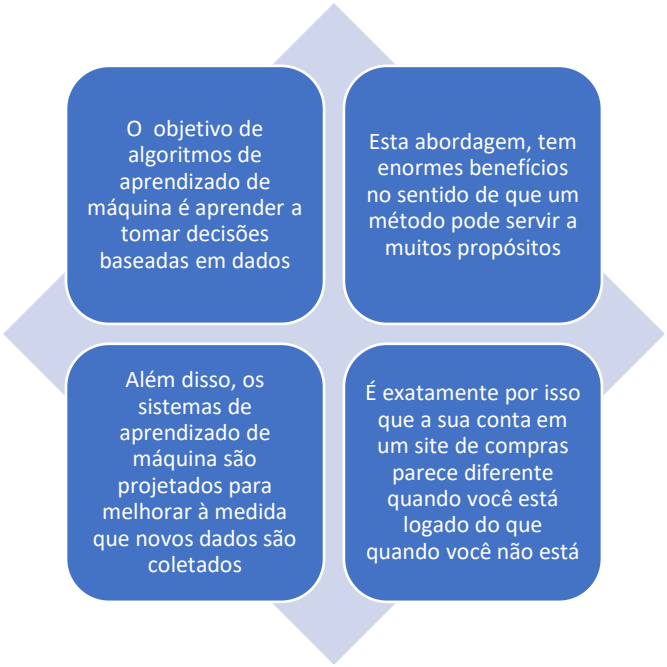
Nós podemos fazer com que as máquinas aprendam a fazer as coisas! Isso significa que podemos programar computadores para aprender por si mesmo!

A capacidade de aprendizagem é um dos aspetos mais importantes da inteligência.

Traduzindo esse poder para máquina, soa como um enorme passo no sentido de torná-las mais inteligentes.

E, de facto, Machine Learning é a área que mais tem conseguido progresso em Inteligência Artificial atualmente.

Machine Learning é o que diferencia a Ciência de Dados de todos os outros modelos tradicionais de análise de dados.



Quando você está navegando por um catálogo e clicando nos itens que gostaria de comprar, um algoritmo está aprendendo as suas preferências e aquele tipo de produto passa a ser oferecido para você.

Algoritmos de Machine Learning

Um algoritmo de aprendizagem de máquina é uma função matemática que, aplicada a uma massa de dados consegue identificar padrões ocultos e prever o que poderá ocorrer.

Basicamente existem 2 tipos de algoritmos de Machine Learning: Supervisionado e Não Supervisionado.

No supervisionado, numa fase chamada de treinamento do modelo, os dados de entrada e de saída são apresentados juntos. O treinamento do algoritmo dura até que o modelo aprenda a mapear os dados e a identificar padrões entre as entradas e as saídas. Como exemplo deste modelo temos as redes neurais e a árvore de decisão.

O modelo não supervisionado só recebe os dados de entrada e sua função é descobrir relacionamentos entre os dados apresentados.

Supervisionado

É o termo usado sempre que o programa é “treinado” sobre um conjunto de dados pré-definido. Baseado no treinamento com os dados pré-definidos, o programa pode tomar decisões precisas quando recebe novos dados. Usado para criação de modelos preditivos.

Exemplo:

Pode-se usar um conjunto de dados de recursos humanos para treinamento da Machine Learning, que tenha tweets marcados como positivos, negativos e neutros e assim treinar um classificador de análise de sentimento.

Não Supervisionado

Termo usado quando um programa pode automaticamente encontrar padrões e relações em um conjunto de dados. Os modelos não supervisionados só recebem os dados de entrada e sua função é descobrir os relacionamentos entre os dados apresentados.

Exemplo:

Análise de um conjunto de dados de mails e agrupamento automático de mails relacionados a um termo, sem que o programa possua qualquer conhecimento sobre qual deveria ser a saída dos dados.

Existe um terceiro tipo de aprendizagem, que ainda não é tão popular, mas que vale a pena mencionar:

Reinforcement Learning

Esse algoritmo é treinado para tomar decisões específicas. Ele funciona da seguinte maneira: o algoritmo é exposto a um ambiente onde ele treina a si mesmo usando tentativa e erro. O algoritmo aprende com a experiência do passado e tenta capturar o melhor conhecimento possível para tomar decisões precisas.

Dependendo do tipo de saída, existem 2 importantes subtipos de **Aprendizado Supervisionado:**

- Classificação
- Regressão

Classificação

A classificação é uma subcategoria de aprendizagem supervisionada, quando o resultado pertence a um conjunto finito de possibilidades.

Classificação é o processo de tomar algum tipo de entrada e atribuir um rotulo a este tipo de entrada. Sistema de classificação são utilizados geralmente quando as previsões são de natureza destinta, ou seja, um simples “sim ou não”.

Texto	Sentimento
“Excelente curso! Bem explicativo, claro e objetivo”	Positivo
“Não fiquei neste hotel. Os quartos cheiram a mofo”	Negativo

Regressão

Outra subcategoria de aprendizado supervisionado usada quando o valor que está sendo previsto difere de um “sim ou não” e que siga um espectro contínuo, como por exemplo uma probabilidade.

Ocupação	Salário	Score
Cientista de Dados	>R$ 15.000,00	0.91
Técnico de Suporte	<R$ 15.000,00	0.28

Aprendizagem supervisionada é a categoria mais popular de algoritmos de aprendizagem de máquina. A desvantagem de usar essa abordagem é que, para cada exemplo de treinamento, temos de fornecer a saída correta e, em muitos casos, isso tem um custo.

Por exemplo, no caso de análise de sentimentos, se precisamos de 10.000 exemplos de treinamento (tweets), teríamos de marcar cada tweet com o sentimento correto (positivo, negativo ou neutro).

Um Algoritmo de Aprendizado de Máquina é quando a matemática e lógica entram em ação, a fim de transformar uma entrada em uma saída desejada.

Os diferentes algoritmos de Aprendizado de Máquina usam diferentes paradigmas ou técnicas para fazer o processo de aprendizagem e representam o conhecimento do que eles aprenderam.

Mas, antes de ir em frente e falar sobre cada algoritmo, um princípio comum é que algoritmo de Aprendizado de Máquina tentam fazer generalizações. Isto é, eles tentam explicar algo com a teoria mais simples. Cada algoritmo de aprendizagem de máquina, independentemente do paradigma que utiliza, irá tentar criar a hipótese mais simples que explica a maioria dos exemplos de treinamento.

Algoritmos de Machine Learning

- Árvore de Decisão
- Naive Bayes
- KNN (K – Nearest Neighbors)
- K-Means
- Random Forest
- Redução de Dimensionalidade
- SVM – Support Vector Machines

Aprendizagem Profunda (Deep Learning)

Este tem sido um tema muito discutido recentemente. Basicamente, a aprendizagem profunda refere-se a uma categoria de algoritmos de Machine Learning, que muitas vezes usam redes neurais artificiais para gerar modelos.

Técnicas de Deep Learning, por exemplo, forma muito bem-sucedidas na resolução de problemas de reconhecimento de imagem, devido a sua capacidade de escolher as melhores características bem como para expressar camadas de representação.

Aplicações do Aprendizado de Máquina

Qualquer área onde dados estejam disponíveis, pode beneficiar de algoritmos de Machine Learning.

Machine Learning não é a mesma coisa que Data Science

Machine Learning é um conjunto de algoritmos que recebe um conjunto de dados e retorna uma previsão.

Exemplos de aplicações de Machine Learning

Processamento de Imagens

O processamento de imagem, basicamente, tem que analisar as imagens para obter dados ou fazer algumas transformações. Vejamos:

Marcação de imagem, como no Facebook – quando o algoritmo deteta automaticamente que o seu rosto ou o rosto dos seus amigos aparece em uma foto. Basicamente um algoritmo de aprendizagem de máquina aprende com as fotos que você marca manualmente.

Reconhecimento Ótico de Caracteres (OCR) – quando algoritmo aprendem a transformar um manuscrito ou documento de texto digitalizado em uma versão digital. O algoritmo tem de aprender a transformar uma imagem de um caracter escrito em uma versão digital correspondente.

Veículos self-driving – parte dos mecanismos que permitem carros conduzir por si mesmos usam o processamento de imagem. Um algoritmo de Aprendizado de Máquina descobre onde está à beira da estrada, se há um sinal de STOP ou um carro está se aproximando, analisado cada foto tirada por uma câmara de vídeo.

Análise de texto

A análise de texto é o processo onde se extraem ou classificam as informações de texto, como tweets, mails, chats, documentos, etc. Alguns exemplos comuns são:

Filtragem de Spam – um dos aplicativos mais conhecidos e utilizados de classificação de texto (atribuir uma categoria a um texto). Filtros de spam aprendem a classificar um mail como spam ou não, dependendo do conteúdo e do objeto.

Análise de Sentimentos – outra aplicação da classificação de texto em que um algoritmo deve aprender a classificar se um determinado texto possui conotação positiva, neutra ou negativa, dependendo do humor expresso nas palavras escritas pelo escritor.

Extração de Informação – a partir de um texto, aprender a extrair uma determinada parte da informação ou dos dados, por exemplo, extrair endereços, entidades, palavras-chave, etc.

Data Mining

A mineração de dados é o processo de descobrir padrões ou fazer previsões a partir de dados. A definição é um pouco genérica, mas pense nisso como mineração de informações úteis a partir de uma enorme tabela e um banco de dados.

Cada linha seria uma instância de informação e cada coluna uma característica. Podemos estar interessados na previsão de uma nova coluna na tabela com base no restante das colunas ou descobrir padrões para agrupar as linhas.

Por exemplo:

Deteção de anomalias – detetar valores atípicos, por exemplo, para a deteção de fraudes de cartão de crédito. Você pode detetar quais das transações são discrepantes do padrão de comprar habitual de um usuário.

Regras de associação – por exemplo, em um site de compras online, você pode descobrir os hábitos de compra do cliente, procurando que produtos são adquiridos em uma única transação de compra. Esta informação pode ser usada para aprimorar as campanhas de marketing.

Previsões – prever uma variável (coluna numa base de dados) do restante das variáveis. Por exemplo, você poderia prever a pontuação de crédito de novos clientes de um banco, a partir da pontuação e perfis de crédito dos clientes atuais.

Games e Robótica

Games e robótica têm sido os campos onde mais se aplica Machine Learning. Vejamos alguns exemplos:

Em geral, temos um agente (personagem do jogo ou um robô), que deve se mover dentro de um ambiente (um ambiente virtual ou físico).

Aprendizagem de máquina, em seguida, pode ser usada para permitir que o agente possa executar tarefas, como mover-se no ambiente, evitando obstáculos ou inimigos.

Uma técnica de Aprendizado de Máquina muito popular nestes casos é **Reinforcement Learning**, dentro do ambiente (o reinforcement é negativo se atinge um obstáculo ou positivo se ele atinge a meta).

A fim de permitir que o algoritmo possa aprender a transformar a entrada em um a saída desejada, você tem que fornecer o que é chamado de instâncias de treinamento ou exemplos de treinamento.

Um conjunto de treino é um conjunto de instâncias (dados) que trabalham como exemplos a partir do qual o algoritmo de Aprendizado de Máquina vai aprender a executar a tarefa desejada.

Questionário

1. O que é Machine Learning?
    a. Machine Learning, é uma tecnologia de Big Data que permite realizar programação em maquinas virtuais.
    b. Machine Learning, ou aprendizado de máquina, é um subcampo dentro de Inteligência Artificial que constrói algoritmos que permitem que os computadores possam aprender a executar tarefas a partir de dados, ao invés de serem programados de forma explícita.
    c. Machine Learning, ou aprendizado de máquina, ou seja, uma linguagem de programação para ensinar novos programados a construírem seus sistemas em ambientes operacionais.
    d. Nenhuma definição acima.

2. Como é conhecida a tecnologia que os sites de busca como “Google” e “Bing” utilizam?
    a. Algoritmo de dados
    b. Algoritmo de ensinamento
    c. Algoritmo estruturado.
    d. Algoritmo de aprendizado.

3. Existem 2 categorias de algoritmos de Machine Learning. Marque as categorias corretas:
    a. Positiva e Negativa
    b. Artificial e Natural
    c. Supervisionada e Não-Supervisionada
    d. Neural e Científica

4. Sobre a aprendizagem Supervisionada é incorreto afirmar:
    a. aprendizagem supervisionada é a categoria mais popular de algoritmos de aprendizagem de máquina.
    b. a desvantagem de usar essa abordagem é que, para cada exemplo de treinamento, temos de fornecer a saída correta e, em muitos casos, isso é muito caro.
    c. necessitam somente ser a entrada para o algoritmo, mas não a saída desejada.
    d. são exemplos de algoritmos de aprendizado de máquina: Support Vector Machines, Naive Bayes, Árvores de Decisão ou Deep Learning.

5. Sobre Redes Neurais Artificiais é incorreto afirmar:
    a. Uma Rede Neural Artificial é um paradigma de processamento de informação que é inspirado na forma como o cérebro humano processa as informações.
    b. O elemento chave das redes neurais é a nova estrutura do sistema de processamento de informações.
    c. É composto por um grande número de elementos de processamento altamente interligados (neurônios) que trabalham em uníssono para resolver problemas específicos.
    d. Aprendizagem em sistemas biológicos envolve adaptações das conexões sinápticas que existem entre os neurônios.
    e. É um tipo de tecnologia que ainda está em fase inicial e tem pouca aplicabilidade para problemas de negócios do mundo real.

6. Machine Learning pode ser aplicada em problemas qualitativos (números) e quantitativos (grupos de dados).
    a. Verdadeiro
    b. Falso - Machine Learning pode ser aplicada em problemas quantitativos (números) e qualitativos (grupos de dados).

6 – Aplicações da Ciência de Dados

Onde aplicar a Ciência de Dados

O potencial de aplicações da Ciência de Dados é imenso.

E este é um dos fatores que tem feito com que Data Science tenha atraído tanto o interesse de empresas, profissionais e do mercado em geral.

Vejamos alguns exemplos:

- Deteção de fraudes
- Carros automatizados
- Melhores sistemas de monitoramento
- Deteção e prevenção de epidemias
- Deteção de terremotos
- Educação customizada, por demanda e online
- Medicamentos customizados, baseados no histórico de cada paciente
- Processo otimizado de iluminação residencial, industrial e pública
- Sistemas de buscas mais eficientes
- Geo-marketing através de smartphones
- Marketing personalizado
- Combate ao crime e ao terrorismo

O que todas estas áreas têm em comum é, a geração de dados em grandes quantidades, variedades e velocidades, características que definem Big Data. Alias todas estas áreas são fontes de Big Data, e por isso, são terrenos férteis para a Ciência de Dados, pois permitem que inúmeras analises sejam feitas, gerando Insights e oportunidades.

Temos visto nos últimos anos, investimento em toda a infraestrutura de negócios, o que tem gerado habilidade de coletar todos os tipos de dados através das empresas.

Atualmente, cada aspeto de um negócio está aberto para a coleta de dados.

- Operações
- Manufatura
- Gestão de Logística
- Comportamento do Cliente
- Processos e Procedimentos
- Performance de Campanhas de Marketing
- Etc…

Em paralelo, a informação está vastamente disponível, como tendência de mercado, novas tecnologias e movimento dos competidores.

Esta disponibilidade de dados, tem aumentado o interesse por métodos de extração de informação útil e conhecimento dos dados.

«E este é o reino da ciência de dados»

A Ciência de Dados pode ser utilizada em praticamente todas as atividades humanas, desde que dados sejam gerados e possam se coletados.

- Finanças
- Comércio e Varejo
- Call Center
- Saúde
- Tecnologia
- Astronomia
- Etc...

Na prática qualquer atividade que gera dados pode-se beneficiar de Data Science.

Exemplos de Aplicações da Ciência de Dados

Walmart

- Servidores de dados com 4 Petabytes de dados
- Cada venda é registrada
- Aproximadamente 267 milhões de transações por dia, nas 6000 lojas em todo o mundo
- Análise de Dados ficada na avaliação da efetividade de estratégias de preço e campanhas de marketing
- Busca de melhoria na sua gestão logística e de inventário

Amazon

- Personalização da experiência de compra online
- Cada cliente possui sua própria loja, baseada nas suas preferências
- Influência das avaliações de outros usuários, nas decisões de compra

Citibank

- A análise de cada uma das transações realizadas pelo banco, nos mais de 100 países em que opera, permite a geração de insights relacionados a investimentos, mudanças de mercada, padrões de operações e condições econômicas

Outros exemplos

Netflix

O trabalho básico de análise é ajudar as empresas a obterem insights sobre os seus clientes, em seguida as empresas podem otimizar a sua comercialização e entregar um produto melhor, a verdade é que sem analises as empresas estão no escuro sobre os seus clientes. Analytics dá as empresas dados quantitativos que podem ser usados par tomarem decisões melhores e com isso melhorar os seus serviços.

Redes te televisão não tem o privilégio de obter retorno dos telespectadores em tempo real. Os dados obtidos em audiência, por exemplo, são apenas aproximações.

Na Netflix cada clique, avaliação ou pesquisa é armazenado e posteriormente analisado permitindo experiencias personalizadas para cada cliente. É como cada um tivesse a sua própria Netflix.

Media Social (Facebook, LinkedIn, Twitter)

Media sociais são como alto volume de dados permitem análises especificas e precisas. O Facebook, por exemplo, possui um modelo de vendas de anúncios totalmente baseado em estatística, usando como referencia o seu alto volume de dados.

Web Apps (Uber, AirBnB)

Planeamento Urbano (Cidades Europeias)

Astrofísica (NASA)

Saúde Pública (Hospitais Americanos)

Diversos hospitais Americanos, estão a conhecer a coletar dados dos seus pacientes e assim garantir um tratamento personalizado. Algumas redes farmacêuticas já começam a desenvolver medicamentos personalizados de acordo com a característica de cada paciente.

Desporto (NFL)

A NFL já esta a utilizar a analise de dados em tempo real. O equipamento de cada um dos atletas da equipa de Futebol Americano possui uma serie de censores, que coletam dados em tempo real sobres o batimento cardíaco, desempenho do atleta e enviam esses dados para uma base de dados, e essa analise em tempo real permite que os treinadores modifiquem a estratégia da sua equipa baseado no que esta acontecendo naquele momento.

Educação

- Orientação em tempo real para que o estudante não abandone a escola
- Personalização do processo de aprendizagem
- Monitoramento do estudante na sua vida acadêmica

Imagine se escolas e faculdades conhecessem melhor seus estudantes, suas limitações, forças e fraquezas e com isso pudessem direcionar e instruir cada um deles com formas personalizadas de aprendizado mais desafiador.

Imagine se estes dados pudessem ser cruzados com indicadores socioeconómicos, renda familiar, previdência social, etc…, gerando com isso informação para governos e instituições.

Varejo

Esta talvez seja uma das áreas mais beneficiadas pela Ciência de Dados, devido a diversidade de dados gerados.

- Análise de sentimentos da marca
- Recomendações individualizada de produto
- Retenção e satisfação do cliente

Telecomunicações

Desenvolver novos produtos já que os dispositivos móveis estão produzindo uma grande quantidade de dados sobre como, por que, onde e quando estão sendo usados. Estes dados são

extremamente valiosos, mas devido ao volume e variedade fica difícil ingerir, armazenar e analisar estes dados que podem resultar em novos grandes insights.

- Análise de registos de ligações
- Geolocalização de ligações
- Ofertas personalizadas de produtos
- Racionalização dos custos

Saúde

- Monitoramento de sinais vitais
- Redução de taxa de retorno de pacientes
- Medicamentos personalizados

Financeiro

- Deteção de fraudes
- Análise de risco
- Análise de perfil
- Geolocalização

Governos

- Fornecimento personalizado de serviços públicos
- Mapeamento de segurança pública
- Redução e racionalização de gastos

Business Intelligence x Ciência de Dados

Business Intelligence não é a mesma coisa que Data Science.

Business Intelligence e Data Science tem muita coisa em comum e Cientistas de Dados e Analistas de Negócios que trabalham com Business Intelligence são como primos.

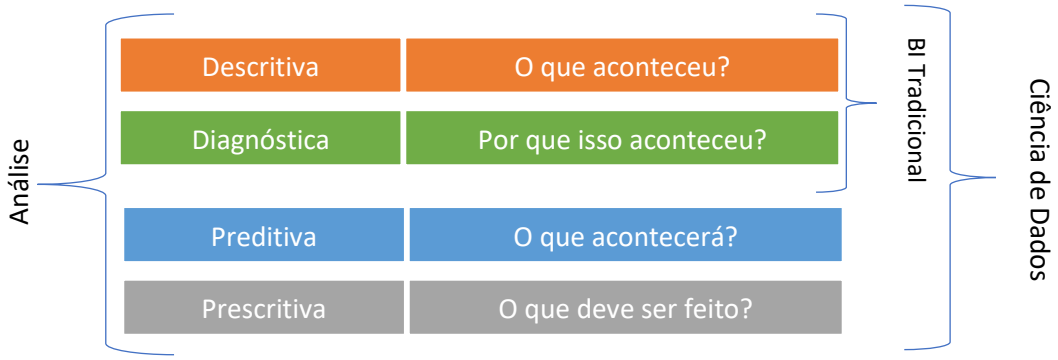
Ambos usam dados com o mesmo objetivo, mas a sua abordagem, tecnologia e função diferem de diversas maneiras.

Business Intelligence

- O objetivo do BI é converter dados brutos em insights de negócio, que os líderes empresariais e gestores possam usar para tomar decisões.

Ciência de Dados

- A Ciência de Dados também converte dados brutos em insights de negócio, mas aplica metodologia científica para exploração dos dados, teste de hipótese, modelagem estatística e aprendizado de máquina.

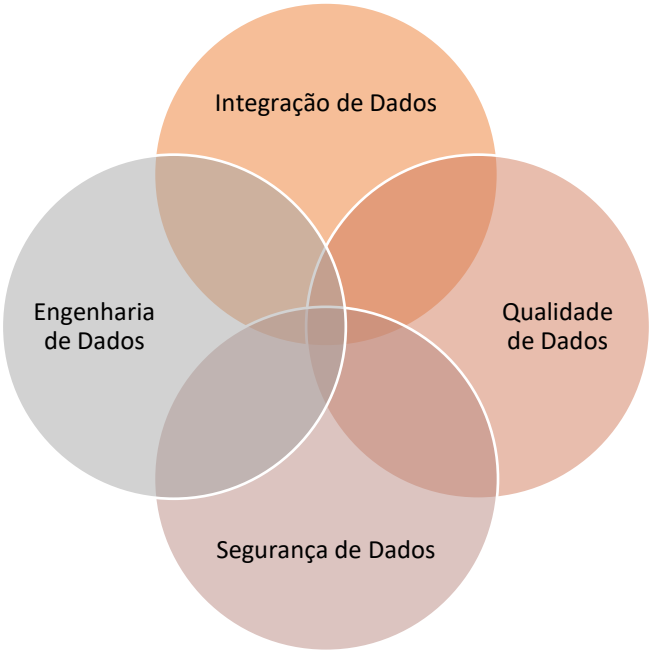


DataOps

DataOps é a operação da infraestrutura necessária para suportar a quantidade, velocidade e variedade de dados disponíveis na empresa, que hoje é radicalmente diferente das abordagens de gerenciamento de dados tradicionais.

A natureza do DataOps abraça a necessidade de gerenciar fontes de dados diferentes e gerados em tempo real, permitindo que o Cientista de Dados tenha a sua disposição a infraestrutura necessária para fazer o seu trabalho.

Basicamente o DataOps é composto por 4 áreas principais



DataOps reconhece a natureza interligada da engenharia de dados, integração de dados, qualidade de dados e segurança de dados e tem como objetivo ajudar uma empresa a entregar rapidamente os dados que poderão acelerar a análise e permitir resultados antes impossíveis.

Em resumo, DataOps, é um conjunto de melhores práticas que visam tornar eficiente a coordenação entre a Ciência de Dados e as operações de dados e tornou-se assim um assunto fundamental para qualquer organização de TI que queira sobreviver e prosperar em um mundo onde inteligência de negócios em tempo real é uma necessidade competitiva.

A Ciência de Dados é uma disciplina extremamente importante atualmente. Mas essa disciplina só é útil na medida em que ela possa ser executada de forma confiável e eficiente. E para que isso aconteça, você precisa de DataOps para os seus DBAs, Cientistas de Dados, Desenvolvedores, Infraestrutura e Operações estejam em harmonia e focados na mesma direção.

Data Lake

Data Lake é um termo recente, criado pelo CTO do Pentaho, James Dixon, para descrever um componente importante no universo da Ciência de Dados e do Big Data.

A ideia é ter um único repositório dentro da empresa, para que todos os dados brutos estejam disponíveis a qualquer pessoa que precise fazer análise sobre eles.

Normalmente utiliza-se o Haddop para trabalhar com os Data Lakes, mas o conceito é bem mais amplo do que apenas Haddop.

A ideia de Data Lake como recurso corporativo ainda está no começo. O conceito de um repositório central, relativamente de baixo custo, que possa armazenar todos os tipos de dados da empresa, ainda é um sonho, apesar de soluções comerciais já disponíveis no mercado.

Os Data Lakes armazenam os dados em seu formato bruto, sem qualquer processamento e sem governança.

Aliás, apesar das soluções comerciais, Data Lake é um conceito e não uma tecnologia. Podem ser necessárias várias tecnologias para criar um Data Lake.

O Data Lake em essência, é uma estratégia de armazenamento de dados.

Quando se ouve falar sobre um ponto único para reunir todos os dados que uma organização deseja analisar, imediatamente se imagina a noção de Data Warehouse e Data Mart.

Mas há uma distinção fundamental entre Data Lake e Data Warehouse.

O Data Lake armazena dados brutos, sob qualquer forma do jeito que foram coletados na fonte de dados. Não há suposições sobre o esquema dos dados e cada fonte de dados pode usar qualquer esquema. Cabe aqueles que vão analisar os dados, dar sentido a esses dados para o propósito ao qual a análise se destina.

Em contrapartida, o Data Warehouse tende a usar a noção de um único esquema para todas as necessidades de análise, o que se torna impraticável em muitas situações relacionadas a Big Data. Os dados são limpos e organizados antes do armazenamento, fazendo com que os dados estejam disponíveis para uso e análise, assim que são armazenados.

Ao mudar o foco para o armazenamento dos dados brutos, isso coloca a responsabilidade sobre os Cientistas de Dados.

Enterprise Data Hub

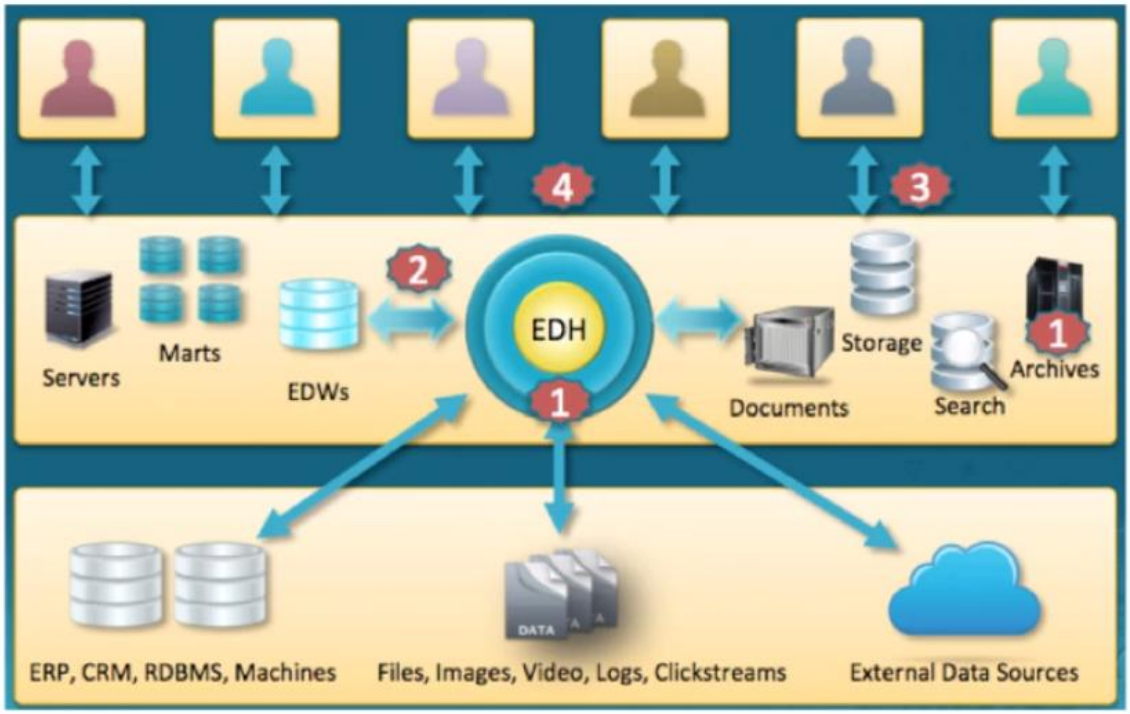
Um Enterprise Data Hub é um modelo de gerenciamento de gerenciamento de Big Data que usa uma plataforma Hadoop como o repositório de dados central

O objetivo de um Enterprise Data Hub é permitir que a empresa tenha uma fonte de dados centralizada e unificada que possa fornecer rapidamente informações a diversos usuários de negócio, apoiando a tomada de decisão.

Todos os aplicativos de analytics se conectam ao EDH para extrair as informações para análise.

A introdução de um Entreprise Data Hub no cerne da arquitetura de informação de uma empresa, promove a centralização de todos os dados, em todos os formatos, disponíveis para todos os usuários de negócios, com total fidelidade e segurança e custo até 99% menor por Terabyte, em comparação com um Data Warehouse tradicional.

Um Enterprise Data Hub serve como um repositório flexível para coletar e manter dados de forma ilimitada, seja para fins de conformidade ou para aplicações sofisticadas, como deteção de anomalias em tempo real.



O conceito de EDH (Enterprise Data Hub), não chega a ser exatamente novo, pois já existem os EDW (Enterprise Data Warehouse).

O desafio está em mover a empresa de modelo tradicional de gestão de dados, para um modelo voltado ao Big Data e suas particularidades como volume, velocidade e variedade, permitindo assim a plena utilização da Ciência de Dados.

O Entreprise Data Hub inclui:

- Reservatório de Dados ou “Data Lake”
- Exploração do Big Data
- Fácil acesso aos dados
- Armazenamento de dados em formato nativo
- Fonte para projetos de Data Science

Open Data

Abaixo, algumas fontes de open data:

Dados do Governo do Brasil: http://dados.gov.br

IPEA: http://www.ipeadata.gov.br

Banco Central do Brasil: https://www3.bcb.gov.br

Dados do Governo dos EUA: http://data.gov

Dados sobre as cidades americanas: http://datasf.org

Dados do Governo do Canadá (em inglês e francês): http://open.canada.ca

Dados do Governo do Reino Unido: https://data.gov.uk

Dados da União Europeia: http://open-data.europa.eu/en/data

Dados do Censo dos EUA (dados da população americana e mundial): http://www.census.gov

Dados da NASA: https://data.nasa.gov

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

Dados sobre diversas áreas de negócio e finanças: https://www.quandl.com

Google Trends: https://www.google.com/trends

Google Finance: https://www.google.com/finance

Gapminder: http://www.gapminder.org/data

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 Monitor: http://opendatamonitor.eu

Open Data Network: http://www.opendatanetwork.com

R Datasets: http://www.stats4stem.org/data-sets.html

R Datasets packages: R Dataset packages

Datasets: http://www.statsci.org/datasets.html

Portal de Estatística: http://www.statista.com

Data 360: http://www.data360.org

Reconhecimento de Faces: http://www.face-rec.org/databases

Stanford Large Network Dataset Collection: http://snap.stanford.edu/data

Datahub: http://datahub.io/dataset

Questionário

1. São exemplos de onde pode ser aplicada a Ciência de Dados?
    a. Deteção de fraudes Carros automatizados, melhores sistemas de monitoramento, deteção e prevenção de epidemias.
    b. Deteção de terremotos, educação customizada, por demanda e online.
    c. Medicamentos customizados, baseados no histórico de cada paciente.
    d. Processo otimizado de iluminação residencial, industrial e pública, sistemas de buscas mais eficientes.
    e. Geo-marketing através de smartphones, marketing personalizado, combate ao crime e ao terrorismo
    f. Todos os exemplos se aplicam a Ciência de Dados.

2. A Ciência de Dados pode ser aplicada à Educação. Indique a afirmação incorreta.
    a. Conjunto muito grande de informação que não pode ser cruzado com indicadores sócio-econômicos.
    b. Orientação em tempo real para que o estudante não abandone a escola.
    c. Personalização do processo de aprendizagem
    d. Monitoramento do estudante na sua vida acadêmica.

3. Sobre Business Intelligence e Ciência de Dados é incorreto afirmar:
    a. Ambos usam dados com o mesmo objetivo, mas a sua abordagem, tecnologia e função diferem de diversas maneiras.
    b. O objetivo do BI é converter dados brutos em insights de negócio, que os líderes empresariais e gestores possam usar para tomar decisões.
    c. Business Intelligence é a mesma coisa que Ciência de Dados.
    d. A Ciência de Dados também converte dados brutos em insights de negócio, mas aplica metodologia científica para exploração dos dados, testes de hipótese, modelagem estatística e aprendizado de máquina.

4. O que é DataOps?
    a. é um conjunto operações de dados independente da Ciência de Dados e tornou-se assim uma ferramenta de dados eletrônicos e funcionais, fundamental para qualquer organização de TI que queira sobreviver e prosperar em um mundo onde inteligência de negócios em tempo compartilhado é uma necessidade competitiva.
    b. é um conjunto de melhores práticas que visam tornar eficiente a coordenação entre a Ciência de Dados e as operações de dados e tornou-se assim um assunto fundamental para qualquer organização de TI que queira sobreviver e prosperar em um mundo onde inteligência de negócios em tempo real é uma necessidade competitiva.
    c. é um conjunto de linguagens de programação para desenvolvedores em C++, dev, java e scripts operacionais.
    d. é um conjunto de técnicas de levantamento de dados para análise científica, para impressão de relatórios em tempo real.

5. O objetivo de um Enterprise Data Hub é permitir que a empresa tenha uma fonte de dados centralizada e unificada que possa fornecer rapidamente informações a diversos usuários de negócio, apoiando a tomada de decisão.
    a. Verdadeiro
    b. Falso

6. A área de Varejo é a menos beneficiadas pela Ciência de Dados, devido a diversidade de dados gerados. Tais como: Análise de sentimento da marca; Recomendação individualizada de produtos e Retenção e satisfação do cliente.
    a. Verdadeiro
    b. Falso - A área de Varejo é uma das mais beneficiadas pela Ciência de Dados, devido a diversidade de dados gerados. Tais como: Análise de sentimento da marca; Recomendação individualizada de produtos e Retenção e satisfação do cliente.

7 – Ciclo de Vida de Projetos de Data Science

Projetos de Ciência de Dados

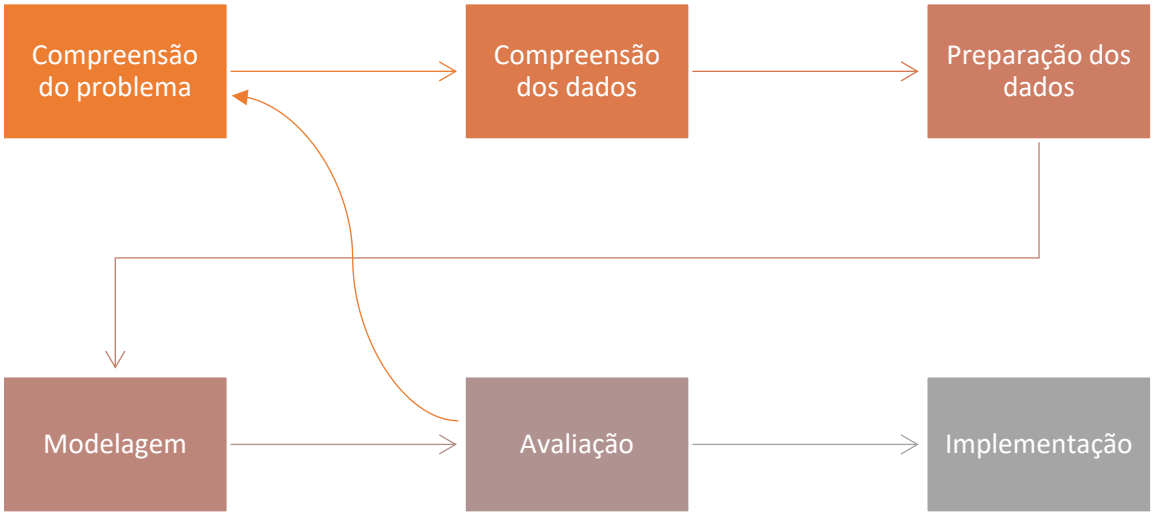
Todo projeto de Data Science deve começar com o objetivo, ou seja, as questões que precisam se respondidas.

Depois de formuladas as questões, busca-se os dados que ajudarão a respondê-las.

Uma vez que o problema tenho sido identificado, temos condições de iniciar o projeto de Ciência de Dados.

A maturidade da disciplina de Gestão de Projetos, permite que boas práticas de gerenciamento de projetos, como as definidas pelo PMI (Project Management Institute), possam ser utilizadas em projetos de Data Science.

Não há porque reinventar a roda. Conhecimento e experiência adquiridos em Gestão de Projetos, serão fundamentais para o sucesso de projetos de Data Science.



Compreensão do problema – uma compreensão clara do problema fara com que se busque os dados e as tecnologias corretas para a analise.

Compreensão dos dados – Os dados estão armazenados num Enterprise Data Hub? Qual é a qualidade dos dados? Existe restrições acesso ou preocupações relativas a segurança dos dados? São dados externos ou internos? Estruturados ou não estruturados? Todas essas questões precisão ser respondidas durante esta etapa.

Preparação dos Dados – Aqui é onde se realizam os processos de limpeza e transformação. Esta é uma faze critica, pois, os dados serão de alguma forma transformados e preciso assegurar que não existe perdas ou inconsistências.

Modelagem – A criação de modelos e aplicação dos mais elevados conceitos de Data Science, Machine Learning, Algoritmos de Aprendizagem, Estatística. É hora de criar um modelo que interprete os dados, de forma permitir extrair informação útil e previsões para o futuro.

Avaliação – Consistem a compreender se o modelo criado se o modelo criado a partir dos dados ajuda realmente a responder as questões e a resolver os problemas levantados no inicio do processo.

Implementação – É onde a empresa vai então obter os resultados do projeto e automatizar a analise subsequente de dados. Nesta fase pudera ser criado um produto, uma PP por exemplo, ou um processo automatizado de analise de dados.

É claro que todas essas etapas girão em torno dos dados coletados.

Podemos chamar todo este processo de **Big Data Analytics** se a fonte de dados for Big Data.

Inicialmente, é vital compreender o problema a ser resolvido. Isso pode parecer óbvio, mas projetos de negócios raramente vem empacotados de forma clara. Avaliar e reavaliar o problema e desenhar uma solução, é um processo interativo de descoberta. Na prática, o maior desafio em Ciência de Dados, está na identificação do problema.

Identificar o problema é uma das tarefas principais em um projeto de Data Science. A identificação correta do problema, vai permitir selecionar a melhor solução!!!

Diversas técnicas analíticas podem ser usadas no processo:

- Estatística
- Busca em bancos de dados (Database Query)
- Data Warehousing
- Análise de Regressão
- Machine Learning e Data Mining
- Etc…

Exemplo:

Vamos elaborar algumas perguntas para identificar o problema e a possível solução a ser abotada.

Quais são os clientes mais rentáveis?

- Esta pergunta poderia ser respondida com uma consulta simples a um banco de dados, usando linguagem SQL

Existem diferenças entre os clientes mais rentáveis e a média dos clientes?

- Aqui poderíamos usar a Estatística, realizando um teste de hipótese para confirmar ou não a nossa tese de que os clientes mais rentáveis possuem deferença em relação à média dos clientes

Agum cliente em particular estará no grupo dos mais rentáveis? Quanto de faturamento posso esperar vindo deste Cliente?

- Aqui poseríamos utilizar Data Mining, para examinar o histórico dos clientes e criar modelos preditivos de geração de lucro por cliente. Estes modelos poderiam ser aplicados automaticamente usando Machine Learning

Perceba que a solução adotada será aquela que permita buscar respostas para as perguntas.

Tudo começa com um problema a ser resolvido. A identificação do problema a ser resolvido, vai impactar todo o projeto da Data Science, não só com relação ao custo, mas também com relação aos recursos que serão adotados.

E quanto mais complexas as perguntas, mais a Ciência de Dados pode ajudar nas Respostas!!

Não existe a melhor tecnologia. Existe aquela que melhor atende o requisito do cliente.

Ciclo de Vida

Basicamente existem 4 grades fases:

Preparação:

- Definição do objetivo
- Compreensão do problema
- Conhecimento dos dados

Engenharia dos Dados

- Aquisição dos dados
- Limpeza dos dados
- Transformação …
- Enriquecimento …
- Persistência …

Analytics

- Análise Exploratória de Dados
- Inferência
- Modelagem
- Predição
- Comunicação

Produção

- Construção de produtos de dados
- Operacionalização de alimentação de dados
- Melhoria continua

Fases do projeto

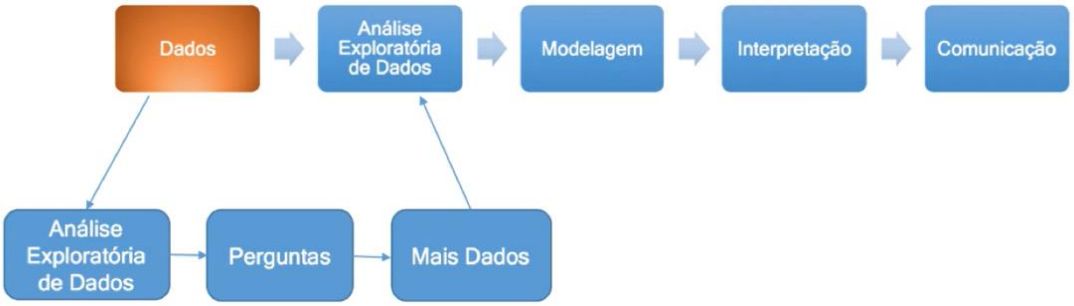
Vamos ver agora outra perspetiva das fases de projetos de Data Science.

Em vez de começar com a definição do problema a ser resolvido, começamos com uma Análise Exploratória do Dados. Este é um trabalho de descoberta diferente do modelo apresentado anteriormente. Neste modelo não temos um objetivo claro, o cientista de dados fica livre para explorar os dados a fim de encontrar respostas as perguntas que ainda não forma feitas. A vantagem desta perspetiva é o que o cientista de dados não fica condicionado a buscar algo especifico, permitindo que ele tenha a liberdade de explorar os dados, detetar relacionamentos inimagináveis. As etapas são basicamente as mesmas que vimos anteriormente, a única diferença é que não se define um objetivo prévio e a exploração é livre. A experiência do cientista de dados será determinante para o sucesso desta abordagem.

Basicamente todo o processo começa com os dados em vez de começar com um objetivo especifico.

Parte-se então para uma Análise Exploratória de Dados e a partir da identificação dos dados começa-se então a gerar perguntas, que podem levar a coleta de mais dados, e então a uma nova fase de Análise Exploratória de Dados. Esse processo normalmente é iterativo, quanto mais analise mais dados provavelmente serão necessários.

A partir daí parte-se para a modelagem, Interpretação, e por fim, Comunicação.



A qualidade dos seus outputs será determinada pela qualidade dos seus inputs. Bons resultados dependem da formulação de boas questões.

Produtos Gerados

Normalmente gera-se no final de projetos de Data Science:

- Relatórios
- Narrativas
- Apresentações
- Web Sites
- Aplicativos
- Etc…

Entretanto algumas características devem ser observadas em qualquer produto gerado através de Data Science:

- Facilidade de uso
- Reprodutibilidade
- Documentação
- Conclusões concisas

Cultura Orientada a Dados

Projetos de Data Science e Big Data ainda são novidades para muitas empresas, pelo menos aquelas empresas que ainda perceberam as possibilidades que dados são capazes de gerar. E apara aproveitar essas oportunidades será necessário rever a cultura da empresa. OS dados estão em todos os lugares, gerados em grande velocidade e em diferentes formatos. A única forma de identificar e aproveitar as oportunidades trazidas pelo Big Data é transformar a cultura da empresa é uma cultura orientada a dados. Veremos cada vez mais empresas a adotar esta abordagem nos seus negócios.

Já ouviu falar da cultura orientada por dados, ou Data-Driven Culture?

Estamos percorrendo uma transformação permanente no modo em que dirigimos os nossos negócios e principalmente, as nossas vidas.

Neste exato momento, uma verdadeira enxurrada de dados, ou quintilhões de bytes por dia, é gerada para nortear indivíduos, empresas e governos – e está dobrando a cada dois anos.

Toda a vez que fazemos uma compra, uma ligação ou interagimos nas redes sociais, estamos produzindo dados.

E com a recente conectividade em objetos, tal como relógios, carros e até geladeiras, as informações capturadas se tornam massivas e podem ser cruzadas para criar roadmaps cada vez mais elaborados, apontando e até prevendo o comportamento de empresas e clientes.

Imagine uma geladeira avisando que a sua bebida favorita está acabando e que o mercado mais próximo de sua casa p está vendendo com desconto.

Agora, pense que um dispositivo RFID pode identificar a sua chegada ao mercado e cruzar um perfil de compras, sugerindo outras marcas de bebidas e produtos similares pelo smartphone.

Quando pensamos na análise de todos esses dados, de diversas fontes conectadas, estamos descrevendo as bases fundamentais de Big Data e da Internet das Coisas.

Esse conceito deixou de ser uma projeção de futuro para se transformar em uma indústria que movimentará, segundo o IDC, US$ 1.7 trilhões em 2020.

Serão mais de 50 bilhões de dispositivos conectados.

Uma empresa com cultura orientada a dados, reconhece a importância do Big Data e como a análise destes dados, pode levar a empresa a criar produtos e serviços personalizados e totalmente aderentes as necessidades do público.

A mudança de paradigma será inevitável. Aliás, já estamos vivenciando esta mudança. Terá sucesso, quem melhor se adaptar a elas.

Os dados já fazem parte do negócio.

Exemplos de Projetos de Data Science

- Análise de Texto das avaliações de itens comprados online.
- Previsão de comportamento de usuário.
- Previsão de desempenho de desportistas ou clubes.
- Análise de Marketing e Media Social para redes de hotéis.
- Previsão de crimes e locais de ocorrências.
- Análise e previsão de taxas de inadimplência.
- Sistemas de recomendação
- Previsão de atrasos em voos
- Análise e Classificação de arquivos potencialmente maliciosos

Questionário

1. Por onde devemos começar um projeto de Data Science?
    a. Pelo levantamento de dados.
    b. Pela reunião de início de gerenciamento do projeto.
    c. Pelo objetivo, ou seja, as questões que precisam ser respondidas.
    d. Pela preparação dos dados do projeto.

2. Quais as fases do ciclo de vida de projeto de Data Science?
    a. Compreensão do problema --> compreensão dos dados --> preparação dos dados --> modelagem --> avaliação --> implementação.
    b. Compreensão dos dados--> compreensão dos problemas --> preparação dos dados --> modelagem --> avaliação --> implantação.
    c. Compreensão do problema --> compreensão dos dados --> preparação dos dados --> modelagem --> avaliação --> implantação.
    d. Preparação dos dados --> compreensão dos dados --> identificação do problema--> modelagem --> avaliação --> implementação

3. A cultura orientada por dados é conhecida por:
    a. Data-Science culture.
    b. Data-driven culture.
    c. Data-lake culture.
    d. DataOps culture.

4. A identificação do problema a ser resolvido, pode impactar todo o projeto de Data Science, não só com relação ao custo, mas também com relação aos recursos que serão adotados.
    a. Verdadeiro
    b. Falso

5. Big Data e a Internet das Coisas deixou de ser uma projeção de futuro para se transformar em uma indústria que movimentará, segundo o IDC, US$ 1,7 trilhão em 2020.
    a. Verdadeiro
    b. Falso

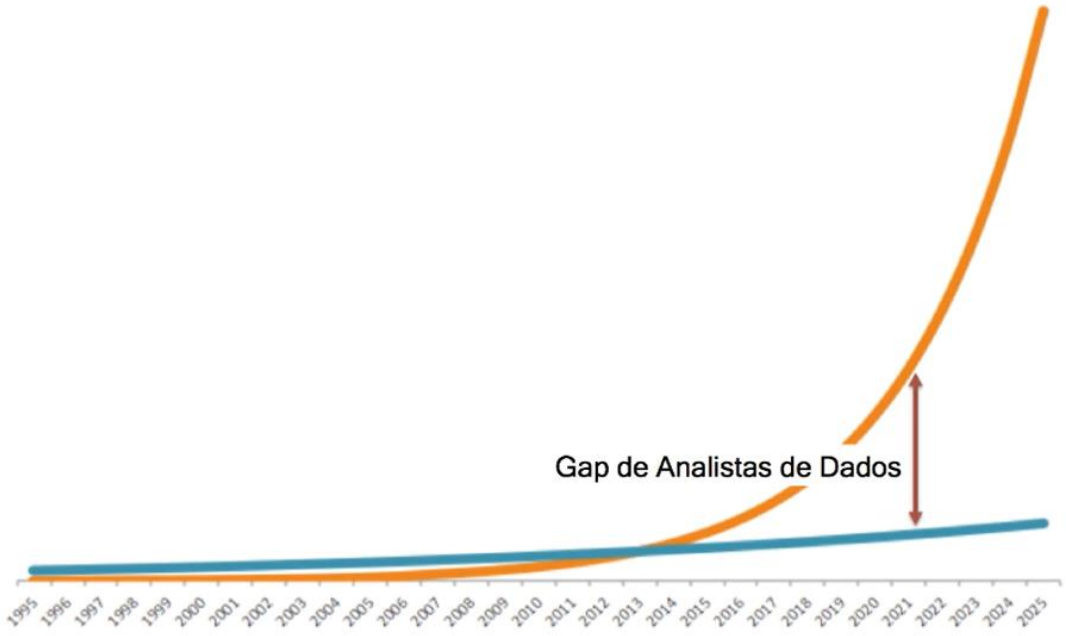
8 – Carreiras em Data Science

O Mercado de Ciência de Dados e Big Data

Com a explosão da análise de dados e do Big Data, a buscar por profissionais capazes de extrair, analisar e gerar insights dos dados, não para de crescer.

Veja este gráfico do site de empregos Indeed

Crescimento do volume de dados VS Crescimento da formação de Analistas de Dados



O armazenamento de dados aumenta 28% ao ano, enquanto a formação de profissionais de Análise de Dados aumenta a 5,7% ao ano.

Até 2024, estima-se que 35 Zettabytes (1ZB = 10^{21} bytes) de dados estejam circulando na rede.

Ou seja, as empresas carecem cada vez mais da experiência daqueles que trabalham com tecnologias como o Big Data – leia-se profissionais experts em ciência da computação, estatística, matemática e domínio de negócios.

A partir daí, extrair o máximo de informação relevantes e decifrá-las com velocidade, torna-se a chave para transformar dados em novos negócios.

A explosão da busca por profissionais de dados tem levado a criação de novas carreiras, bem como a reinvenção de outras.

Por mais que os conceitos de dados existam a séculos, as tecnologias que surgiram recentemente, permitem fazer coisas que não eram possíveis antes, sem falar no fato que o volume de dados gerados pela humanidade nunca foi tão grande.

Na prática, todas as profissões envolvem a coleta e análise de dados.

Vejamos as carreiras que mais estão sendo requisitadas, principalmente pelas empresas que já começaram seus projetos de Big Data e que precisam de profissionais para a Ciência de Dados.

De forma geral, podemos identificar têS perfis básicos de profissionais engajados em Data Science e Big Data:

1. **Cientistas de dados**
   Profissionais capacitados em estatística, ciência da computação e/ou matemática capazes de analisar grandes volumes de dados e extrair deles insights que criem novas oportunidades de negócio.
2. **Analistas de negócio**
   Conhecendo bem o negócio em que atuam conseguem formular as perguntas corretas, analisar as respostas e tomar decisões estratégicas e táticas que alavanquem novos negócios ou aumentem a lucratividade da empresa. Esta função tende a ser acoplada às funções do Cientista de Dados.
3. **Profissionais de tecnologia**
   Cuidarão da infraestrutura e do suporte técnico para suportar Big Data. O aparato tecnológico de Big Data não é muito comum em empresas tipicamente comerciais, pois demanda expertise em gerenciar hardware em clusters de alta performance.

Entretanto, nos próximos anos viveremos uma escassez destes profissionais em todo o mundo.

Esta escassez ao mesmo tempo em que abre muitas perspetivas profissionais para os que abraçam a função, também atuará como um entrave, pois dificultará às empresas usarem Big Data com eficiência.

Recentes pesquisas estimam que por volta de 2018 Big Data demandará cerca de 4,4 milhões de profissionais em todo o mundo e que apenas 1/3 destes cargos poderá ser preenchido com as capacitações disponíveis hoje em dia.

Uma pesquisa mundial da IBM corrobora estes dados, mostrando que apenas uma em dez organizações acredita que tenha profissionais com as capacitações necessárias e que três em cada quatro estudantes e professores reportam que existe um gap de moderado a grande entre o que é ensinado hoje e o que o mercado de trabalho necessita.

Carreiras em Data Science

Não existe uma classificação formal para as carreiras em Data Science, entretanto, alguns perfis são facilmente identificados em qualquer projeto de Data Science e Big Data.

Analista de Negócios

- Estabelecer os objetivos e o âmbito de sistemas de negócios e de TI
- Identificar problemas organizacionais e conceber soluções orientadas a dados
- Realizar análises estatísticas, pesquisa, oficinas de formação e testes
- Recomendar mudanças nos processos, pessoal ou ofertas de produtos para tomar os departamentos internos mais eficientes
- Conceber novos sistemas ou alterar os existentes
- Fazer recomendações específicas de TI e apoiar a sua implementação
- Agir como um elo de ligação entre os gestores e equipas técnicas
- Propor suas decisões baseadas em dados

Analista de Dados

- Trabalhar comas equipas de TI, gestão e/ou Cientistas de Dados para determinar os objetivos organizacionais
- Coletar dados de fontes primárias e secundárias
- Realizar limpeza nos dados e descartar informações irrelevantes
- Analisar e interpretar os resultados utilizando ferramentas estatísticas e técnicas convencionais
- Identificar tendências, correlações e padrões em conjunto de dados complexos
- Identificar novas oportunidades para a melhoria de processos
- Fornecer relatórios de dados concisos e visualização de dados claros para a gestão
- Conceção, criação e manutenção de bancos de dados relacionais e NoSQL e sistemas de dados

- Resolver problemas de código e questões relacionadas a dados
- Dominar linguagens (R, Python, SQL) e software de análise de dados (SAS, Tableau, Qlik)

Arquiteto de Dados

- Colaborar com as equipas de TI e gestão para elaborar uma estratégia de dados que atenda os requisitos da empresa
- Criar um inventário de dados necessários para implementar a arquitetura
- Pesquisar novas oportunidades da aquisição de dados
- Identificar e avaliar as atuais tecnologias de gerenciamento de dados
- Criar um fluxo de dados dentro da empresa
- Desenvolver modelos de dados
- Projetar, documentar, contruir e implementar arquiteturas e aplicações de bancos de dados (por exemplo, grandes bancos de dados relacionais e NoSQL)
- Integrar a funcionalidade técnica (por exemplo, escalabilidade, segurança, desempenho, recuperação de dados, confiabilidade, etc.)
- Implementar medidas para assegurar a precisão dos dados e acessibilidade
- Monitorar constantemente, aperfeiçoas e apresentar um relatório sobre o desempenho dos sistemas de gerenciamento de dados

Engenheiro de Dados

- Projetar, contruir, instalar, testar e manter sistemas de gerenciamento de dados altamente escaláveis
- Construir algoritmos de alto desempenho, protótipos, modelos preditivos e provas de conceito
- Pesquisar a aquisição de dados e novos usos para os dados existentes
- Desenvolver processos de conjuntos de dados para modelagem de dados, mineração e produção
- Integrar novas tecnologias de gerenciamento de dados e ferramentas de engenharia de software nas estruturas existentes
- Criar componentes personalizados de software e aplicações analíticas
- Empregar uma variedade de linguagens e ferramentas
- Instalar e atualizar os procedimentos de recuperação de desastres
- Recomendar formas de melhorar a confiabilidade dos dados, eficiência e qualidade
- Dominar tecnologias como Hadoop, Spark e Cassandra

Administrador de Bancos de Dados

- Suporte técnico aos bancos de dados existentes
- Personalização de bancos de dados comerciais para necessidades específicas
- Planejamento e projeto de banco de dados para necessidades específicas
- Solução de problemas para atender às necessidades dos clientes
- Desenvolvimento de Bancos de Dados para uma ampla variedade de aplicações
- Supervisão da instalação de novos SGBD
- Criar procedimentos de backup, restauração e recuperação de desastres
- Atuar com bancos de dados relacionais e não relacionais

Estatístico

- Aplicar teorias e métodos estatísticos para resolver problemas práticos de negócio, engenharia, ciência ou outras áreas de conhecimento

- Decidir quais dados são necessários para responder a perguntas ou problemas específicos
- Determinar métodos para encontrar ou coletar dados
- Realizar pesquisas de opinião para coletar dados
- Coletar dados ou treinar outras pessoas a fazê-lo
- Analisar e interpretar dados
- Relatar conclusões a partir de suas análises
- Relatar sobre uma estratégia adequada para coletar dados
- Decidir sobre uma estratégia adequada para coletar dados
- Extrair dados de fontes existentes ou instigar novos procedimentos (por exemplo, pesquisa com clientes, experiências cientificas, sondagens de opinião, etc.)
- Analisar e interpretar dados usando ferramentas, algoritmos, modelos estatísticos e software (por exemplo R, SAS, SPSS, etc.)
- Projetar novos modelos estatísticos e ferramentas de coleta de dados, se necessário

Cientista de Dados

- Comunicar previsões e resultados para a gestão e departamento de TI através de visualização de dados eficazes
- Extrair grande volumes de dados de múltiplas fontes internas e externas
- Empregar os programas de análise sofisticadas, aprendizado de máquina e métodos estatísticos para preparar os dados para uso em modelagem preditiva e prescritiva
- Explorar e analisar dados de uma variedade de ângulos para determinar fraquezas escondidas, tendências e /ou oportunidades
- Conceber solução orientadas a dados para os principais desafios da empresa
- Criar novos algoritmos para resolver problemas e criar novas ferramentas para automatizar o trabalho
- Recomendar mudanças econômicas aos procedimentos e estratégias existentes
- Dominar técnicas de análise e armazenamento de dados

O desafio do Cientista de Dados é capturar informações em tempo real, sendo a maioria não-estruturados, como as publicações em redes sociais e websites.

É preciso filtrar tudo, cruzar, analisar com os diversos bancos de dados internos da empresa e entregar relatórios valiosos que possam apoiar as estratégias de negócios.

Porém, mais do que isso, o Cientista de Dados deve buscar nos dados, respostas para perguntas que ainda não foram feitas. Esse talvez seja um dos seus maiores desafios.

Porém como se trata de uma profissão nova, achar gente com esse tipo de capacitação não é tarefa fácil.

Segundo o Vice-Presidente Sênior do Gartner, ter profissionais especializados para dar suporte a Data Science e Big Data é um desafio global. Ele constata que os sistemas de educação tanto o publico como privado não têm como formar essa quantidade de profissionais na velocidade que as empresas precisam.

O Cientista de Dados tem de saber programação, ser capaz de criar modelos estatísticos e ter o conhecimento e domínio apropriado de negócios. Precisa também compreender as deferentes plataformas de Big Data e como elas funcionam.

Existem várias definições disponíveis para os Cientistas de Dados.

Em palavras simples, um Cientista de Dados é aquele que pratica a arte da Ciência de Dados.

Cientistas de Dados são capazes de trabalhar com complexos problemas de dados, com sua forte experiência em determinadas disciplinas científicas.

Eles trabalham com vários elementos relacionados com matemática, estatística, ciência da computação, etc… (embora não tenham que ser especialistas em todos estes domínios).

Funções gerenciais em Data Science

Apesar de muitas denominações possíveis, 3 posições chave em liderança de projetos de Data Science são facilmente identificáveis:

- Líder de Equipa de Data Science
- Gerente de Dados e Analytics
- Chief Data Officer

Embora o volume de dados aumente a cada dia, o que requer investimento em armazenamento e análise, a má gestão dos dados ainda tem sido o que mais se vê no ambiente corporativo.

Gerenciar dados custa caro e de acordo com o *Gartner*, estima-se um prejuízo na ordem de 13 bilhões de dólares com o gerenciamento ineficaz dos dados.

Como os dados são armazenados eletronicamente, muitas empresas deixam sua gestão para o departamento de Tecnologia de Informação (TI). No entanto, não há uma função em TI que seja voltada para o gerenciamento de dados e ninguém é oficialmente responsável pelos dados.

As pessoas esperam que os dados sejam armazenados com precisão, mas na maioria das vezes não é o que acorre.

Entra em cena então o Chief Data Officer – CDO (Executivo Chefe de Dados). Esta função é relativamente nova e muitas empresas ainda não estão dando atenção a isso. O papel do CDO é trazer ordem para o caos e proteger o investimento da empresa, seja na coleta, armazenamento ou análise de dados.

Tão ruim quanto armazenar dados de forma imprecisa, é analisar estes dados armazenados de forma imprecisa, o que pode levar a decisões de negócio catastróficas.

Questionário

1. Quais os três perfis básicos de profissionais engajados em Ciência de Dados e Big Data?
    a. Cientista de Requisitos, Analista de Sistemas, Profissionais de tecnologia.
    b. Cientista de Dados, Analista de sistemas e Estatísticos de Inferência.
    c. Cientista de Dados, Analista de Negócios, Profissionais de Tecnologia.
    d. Analista de Dados, Analista de Negócios, Analista de Requisitos.

2. Quais dessas carreiras não fazem parte da Ciência de Dados?
    a. Cientista de dados, Engenheiro de Dados, Analista de Dados, Estatístico.
    b. Arquiteto de Dados, Analista de Negócios, Administrador de Banco de Dados.
    c. Gerente de Recursos humanos, Programador e Cientista de Ensaios
    d. Todas essas carreiras listadas acima fazem parte da Ciência de Dados.

3. Coletar dados de fontes primárias e secundárias, realizar limpeza nos dados e descartar informações irrelevantes; analisar e interpretar os resultados utilizando ferramentas estatísticas e técnicas convencionais e identificar tendências, correlações e padrões em conjuntos de dados complexos. São tarefas do:
    a. Analista de negócios
    b. Analista de Dados
    c. Engenheiro de Dados
    d. Cientista de Dados

4. Comunicar previsões e resultados para a gestão e os departamentos de TI através de visualizações de dados eficazes; extrair grandes volumes de dados de múltiplas fontes internas e externas; empregar os programas de análise sofisticadas, aprendizado de máquina e métodos estatísticos para preparar os dados para uso em modelagem preditiva e prescritiva, são tarefas do:
    a. Analista de negócios
    b. Analista de Dados
    c. Cientista de Dados
    d. Arquiteto de Dados

5. Nos próximos anos viveremos uma escassez de profissionais de Ciência de Dados. Esta escassez ao mesmo tempo em que abre muitas perspetivas profissionais para os que abraçarem a função, também atuará como um entrave, pois dificultará às empresas usarem Big Data com eficiência.
    a. Verdadeiro
    b. Falso

6. O Estatístico tem por tarefa realizar o suporte técnico aos bancos de dados existentes, bem como, realizar a personalização de bancos de dados comerciais para necessidades específicas.
    a. Verdadeiro
    b. Falso - É o Administrador de Banco de Dados que tem por tarefa realizar o suporte técnico aos bancos de dados existentes, bem como, realizar a personalização de bancos de dados comerciais para necessidades específicas.

9 – Como se tornar um Cientista de Dados

A Profissão de Cientista de Dados

O campo é novo, mas está a crescendo rapidamente.

Há uma grande procura pelos Cientistas de Dados – remuneração media nos EUA já é superior a 100 mil dólares por ano.

O déficit de competência de Ciência de Dados significa que muitas pessoas estão aprendendo ou tentando aprender Data Science.

O primeiro passo para aprender a Ciência de dados é geralmente perguntando “Por onde começar?”

A resposta a esta pergunta tende a ser uma longa lista de cursos e livros para ler, começando com álgebra linear ou estatística. É semelhante a um professor entregando-lhe uma pilha de livros e dizendo “leia tudo isso”

Esse caminho tende a ser pouco motivador e na maioria das vezes, fracassa.

Algumas pessoas aprendem melhor com livros, outras com vídeos, mas a maneira mais eficaz de aprender, é fazendo.

A motivação aparece quando você começa a ver o resultado do seu aprendizado.

O melhor de tudo, quando você aprende desta forma, você adquire habilidades imediatamente úteis.

Se você quer aprender a Ciência de Dados, seu primeiro objetivo deve ser aprender a gostar de dados.

Quem é o Cientista de Dados

Um Cientista de Dados é alguém que é curioso, que analisa os dados para detetar tendências.

Cientistas de Dados são uma nova geração de especialistas analíticos que têm as habilidades técnicas para resolver problemas complexos.

E a curiosidade explorar quais são os problemas ser resolvidos.

Eles são parte matemáticos, parte cientistas da computação e parte analistas de tendências.

E por transitarem entre o mundo dos negócios e de TI, eles são muito procurados e bem pagos.

Eles também são um sinal dos tempos modernos. Cientistas de Dados não estavam no radar há uma década, mas sua popularidade repentina reflete como as empresas agora pensam sobre Big Data.

Essa imensidão de informações não estruturadas já não pode mais ser ignorada e esquecida. É uma mina de ouro virtual que ajuda a aumentar receitas – desde que haja alguém que escave e desenterre insights empresariais que ninguém havia pensado em procurar.

Entra em cena o Cientista de Dados.

Quem é o Cientista de Dados?

«Pessoa que é melhor em estatística que o engenheiro se software e que é melhor em engenharia de software que qualquer estatístico.»

- Profissional que pratica Ciência de Dados
- Experiência em engenharia de dados, analytics, estatística e áreas de negócio
- Investiga problemas complexos de negócio e provê soluções a partir dos dados

Características do Cientista de Dados

- Curiosidade
- Intuição
- Comunicação
- Capacidade de apresentação
- Criatividade
- Conhecimento de Negócios

E como as características de um Cientista de Dados devem ser balanceadas?

1. Curiosidade – para descobrir novas informações a partir dos dados.
2. Engenharia de Dados – é fundamental que se conheça um pouco da infraestrutura necessária para armazenar e disponibilizar esses dados para analise.
3. Programação – é um conhecimento importante, uma linguagem de programação vai permitir que se automatize o trabalho.
4. Machine Learning e Estatística – este é o grande deferencial da Ciência de Dados comparada a outras técnicas de analise.
5. Aplicação de Métodos Científicos – como já vimos anteriormente o que diferencia o Business Intelligence da Ciência de Dados é aplicação do método cientifico. Por isso é importante que o Cientista de Dados tenha uma base em matemática, estatística ou ciência da computação.
6. Contador de Historias – lembre-se comunicação é uma das características principais de um Cientista de Dados. De nada adianta realizar um excelente trabalho de analise se não for capaz de mostrar o seu trabalho. Por isso o Cientista de Dados deve ter a habilidade necessária para contar uma historia a partir dos dados. O publico alvo poderá ser formando por pessoas sem perfil técnico. Essas pessoas não querem saber se foi usada a linguagem de programação X ou Z, o que elas querem saber é como o problema será resolvido a partir dos dados.
7. Conhecimento de Negócio – aqui é onde reside a diferença entre um bom Cientista de Dados e um Cientista de Dados que não conseguira espaço no mercado. É fundamental que se conheça área de negocio onde vai atuar, para que com isso, seja mais fácil identificar problemas e fazer novas perguntas a partir dos dados coletados.

Business	**Machine Learning / Big Data**	**Matemática**	**Programação**	**Estatística**
Desenvolvimento de Produtos	Dados não estruturados	Otimização	Programação Back e Front-end	Estatística temporais
Áreas de Negócio	Dados estruturados	Modelos gráficos		Estatística espacial
	Machine Learning	Algoritmos		Manipulação de dados
	Big Data e dados distribuídos	Simulação		Estatística clássica

É preciso profundo conhecimento em pelo menos uma das áreas abaixo:

- Estatística
- Machine Learning
- Big Data
- Business

O que faz um Cientista de Dados

O trabalho de Cientista de Dados começa com a identificação do problema.

Uma vez que o problema tenha sido identificado e compreendido o próximo passo e coletar os dados necessários para que se possa começar uma analise para selecionar o problema.

Uma que os dados tenham sido coletados começa-se o processamento. É onde se faz a limpeza, transformação e normalmente se busca por novos dados se necessário.

Após o processamento os dados são divididos subconjuntos menores, chamados de Datasets, para facilitar a analise e com isso criar produtos, que será o ressoltado do Cientista de Dados.

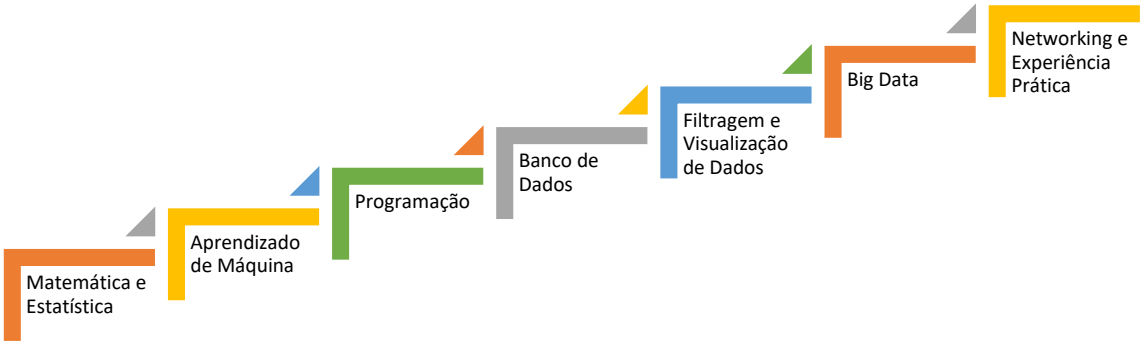
A partir dos Datasets se aplica o Machine Learning ou Análise / Modelos Estatísticos, ou até mesmo os dois dependendo do objetivo.

Por fim gera-se o resultado, que serão os Relatórios, visualizações ou Apps orientadas a dados.

Basicamente é este o trabalho de um cientista de dados.

Como se preparar para a Profissão do Futuro

Como se tornar um Cientistas de Dados



O que o Cientista de Dados precisa saber?

Área de Conhecimento	Habilidade
Matemática e Estatística	Álgebra Linear, Estatística Descritiva, Teste de Hipótese, Análise Bayesiana
Aprendizado de Máquina	Aprendizagem Supervisionada e Não Supervisionada, Classificação, Regressão, Clustering
Programação	Python, R, Scala, Julia, Java, SAS, SQL, C++
Banco de Dados	Banco Relacionais e Bancos No-SQL como MongoDB
Filtragem e Visualização de Dados	D3.js, Tableau, Infovis, ggplot2
Big Data	Hadoop, Spark, Storm, Cassandra
Área de Negócio	Finanças, Marketing, Varejo, Astronomia, Saúde, Tecnologia, etc…

Como se preparar

Existem diversos sites onde se pode praticar o que se esta a aprender.

Kaggle – competições especificas de analise de dados

CrowdAnalytix

KDD Cup

HackerRank

DataKind

Data Science For Social Good

Questionário

1. Há uma grande procura pelos Cientistas de dados - remuneração média nos EUA já é superior a 100 mil dólares por ano.
    a. Verdadeiro
    b. Falso

2. Cientistas de dados fazem parte de uma geração antiga de especialistas analíticos que têm as habilidades técnicas para resolver problemas complexos.
    a. Verdadeiro
    b. Falso - Cientistas de dados são uma nova geração de especialistas analíticos que têm as habilidades técnicas para resolver problemas complexos.

3. Quais dessas características não se aplica ao Cientista de Dados?
    a. Curiosidade e intuição.
    b. Comunicação e Capacidade de apresentação.
    c. Criatividade e Conhecimento de Negócios.
    d. Introspeção e avesso à Tecnologia.

4. Conhecimentos em matemática e estatística, aprendizado de máquina, programação, banco de dados, filtragem e visualização de dados, big data e experiência prática representam os degraus que um profissional preciso trilhar para se tornar um Cientista de Dados.
    a. Verdadeiro
    b. Falso