

— O QUE É —

DATA SCIENCE?

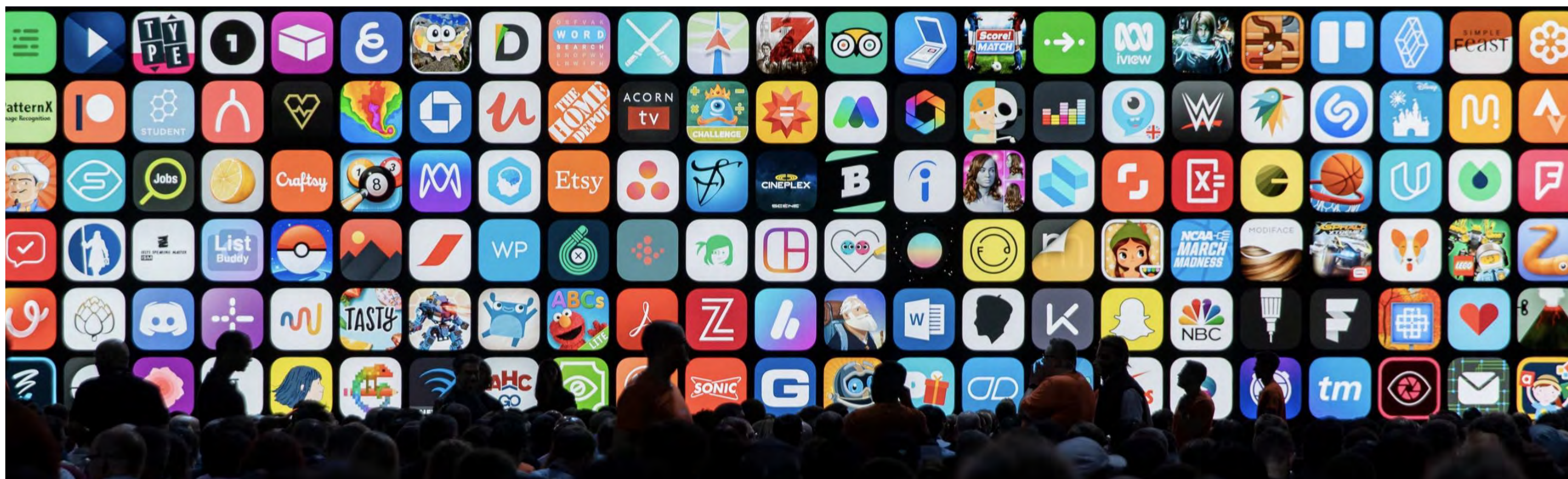
Uma introdução sobre o assunto



Introdução

Todas as novas plataformas online como Youtube, Netflix, Snapchat, Facebook e Instagram geram grandes massas dados. São as chamadas data-driven companies, empresas orientadas a dados e que utilizam de **data science** para tomar decisões.

Neste eBook, explicaremos **o que é data science**, qual sua origem e quais são as suas principais aplicações. Além disso, também abordaremos as suas principais subdivisões e quais são as profissões possíveis de se seguir nessa área.



Sumário

04 A revolução dos dados

05 Aceleração dos dados

06 O início da ciência de dados

07 Qual a diferença entre dado e informação?

08 Conceito de Insight

09 O que é data science?

11 Pilares da Ciência de dados

11 Qual a diferença entre Data Science e BI?

13 E afinal, o que é machine learning?

14 Aplicações de data science

14 Carreiras de Data Science

16 O Futuro do Ciência de Dados

A revolução dos dados

A inovação tecnológica gera avanços que transformam o mundo à nossa volta e nos empoderam como indivíduos. Um dos importantes efeitos da transformação digital foi a democratização do conhecimento, hoje virtualmente gratuito.

Entretanto, usufruir do potencial pleno dessa transformação digital só é possível se explorarmos a capacidade dos dados gerados por essas inovações. Além da abundância de dados disponíveis, o que impulsiona a revolução no setor de dados são as tecnologias que alteram a forma como coletamos, armazenamos, analisamos e transformamos as informações.

Por toda a história da humanidade, os marcos da nossa civilização foram caracterizados pelos progressos em nossa capacidade de observar e coletar dados. Nossos ancestrais distantes desenvolveram ferramentas e métodos práticos para medir distância, peso, volume, temperatura, tempo e localização.



Toda essa experiência foi fundamental para a sua transformação de caçadores-coletores em agricultores e, posteriormente, em habitantes de cidades com organizações e divisões mais complexas de trabalho.

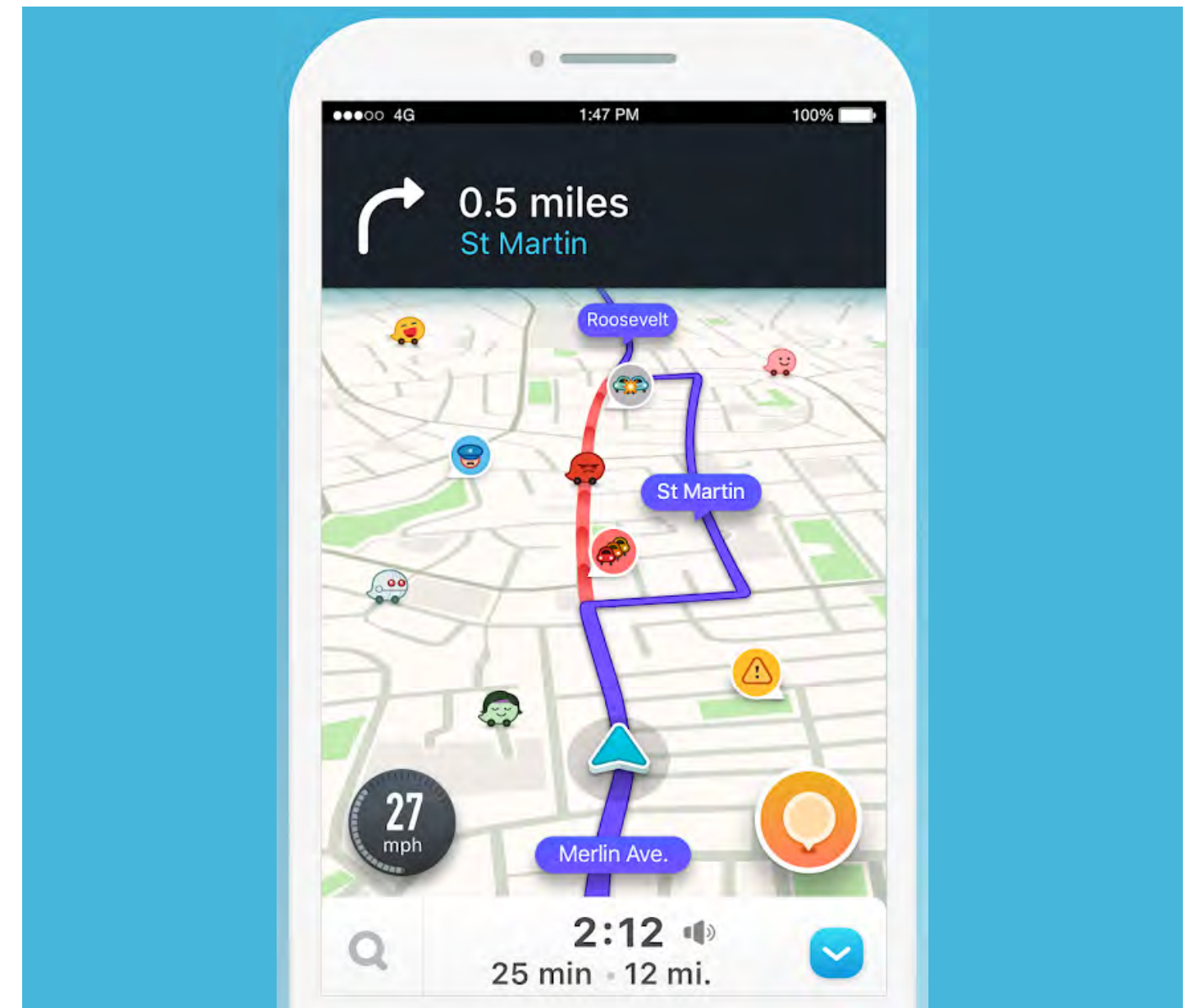
Ao longo da história moderna, até mesmo pequenas quantidades de dados nos ofereceram informações importantes na busca de soluções para alguns dos nossos maiores desafios. O registro das informações em pedra, papiro, livros impressos e, posteriormente, computadores tem sido um dos principais motivadores do progresso humano.

Aceleração dos dados

Neste século, vivemos uma rápida aceleração de todo esse processo. Com a maior abundância dos dados e a redução de seus custos de armazenamento e processamento, os cientistas de dados se beneficiam da tecnologia para revelar insights valiosos.

Com ferramentas avançadas de análise de dados, os profissionais de data science conseguem realizar previsões que resolvem grandes problemas e melhoram

a nossa vida cotidiana. Hoje, as tecnologias de transporte, por exemplo, estão sofrendo uma disrupção causadas pelas novas aplicações baseadas em grandes volumes de dados, como o Uber e o Waze.

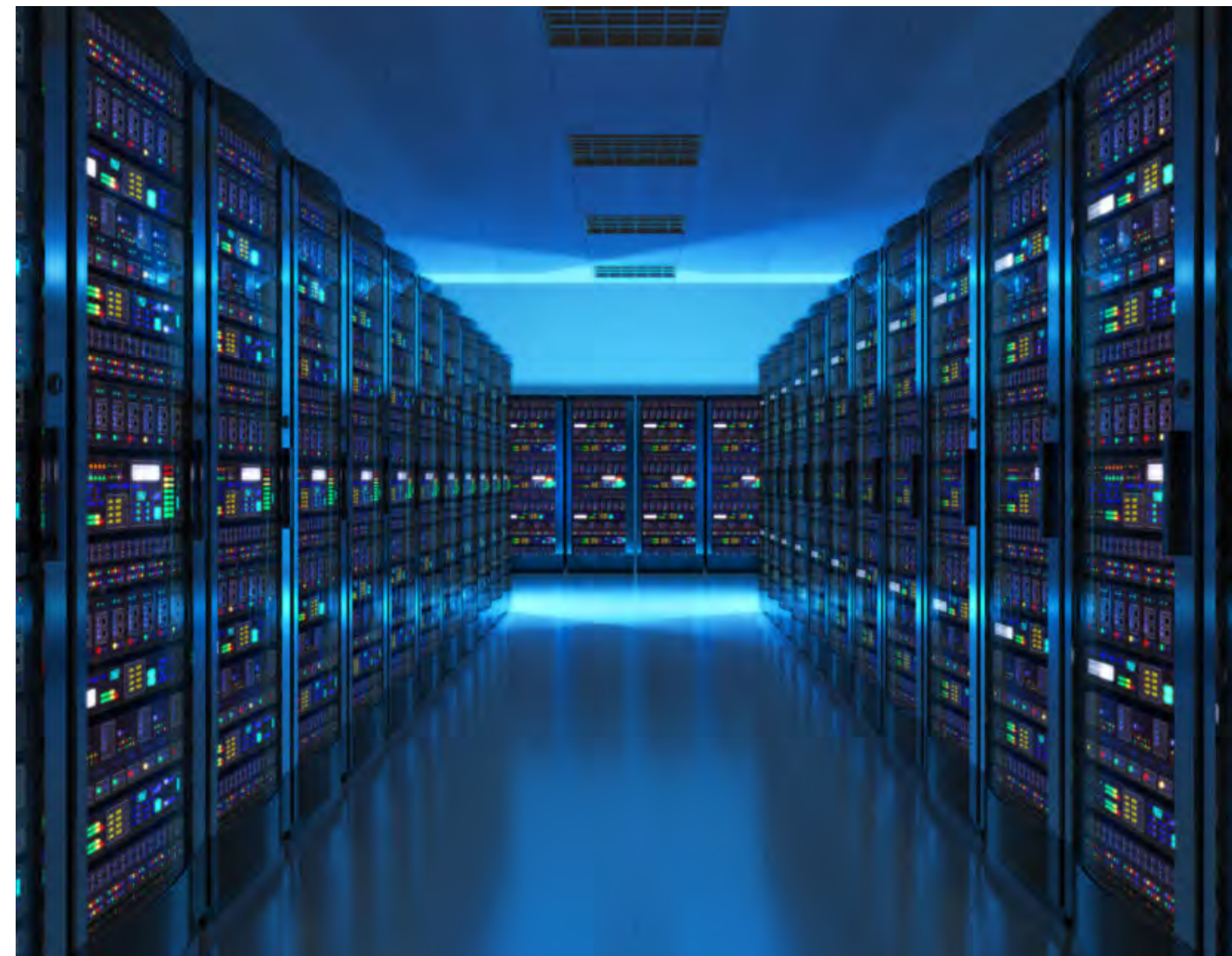


O início da ciência de dados

Alguns fatores culminaram na existência da ciência de dados. O principal deles é o aumento de **dados não estruturados** disponíveis, a partir da digitalização da informação. Esse grande volume de dados não estruturados também é conhecido como Big Data.

O segundo fator importante foi o avanço na capacidade de processamento em nuvem, por meio de processamento horizontal com clusters. Sem esse aumento de capacidade de processamento a ciência de dados certamente não existiria. Isso ocorre porque o processamento vertical tradicional é caro e ineficiente para grandes quantidades de dados.

Esse problema foi resolvido, principalmente, a partir da especialização de capacidade computacional disponibilizada por



forneadores de computação em nuvem, como Amazon (AWS), Google (GCP) e Microsoft (Azure). Com a possibilidade de locação de hardware sob demanda e a sua redistribuição para atingimento de máxima eficiência, muitos projetos passaram a ser viabilizados com a computação em nuvem.

Qual a diferença entre dado e informação?

Conceitualmente, dados são conteúdos quantificáveis que não possuem nenhum valor, sendo considerados uma unidade básica de valor. Já a informação é o resultado do processamento dos dados, ou seja, a interpretação dos dados ou o seu significado.

No esquema abaixo, é possível compreender como funciona o fluxo de transformação entre dado e informação:

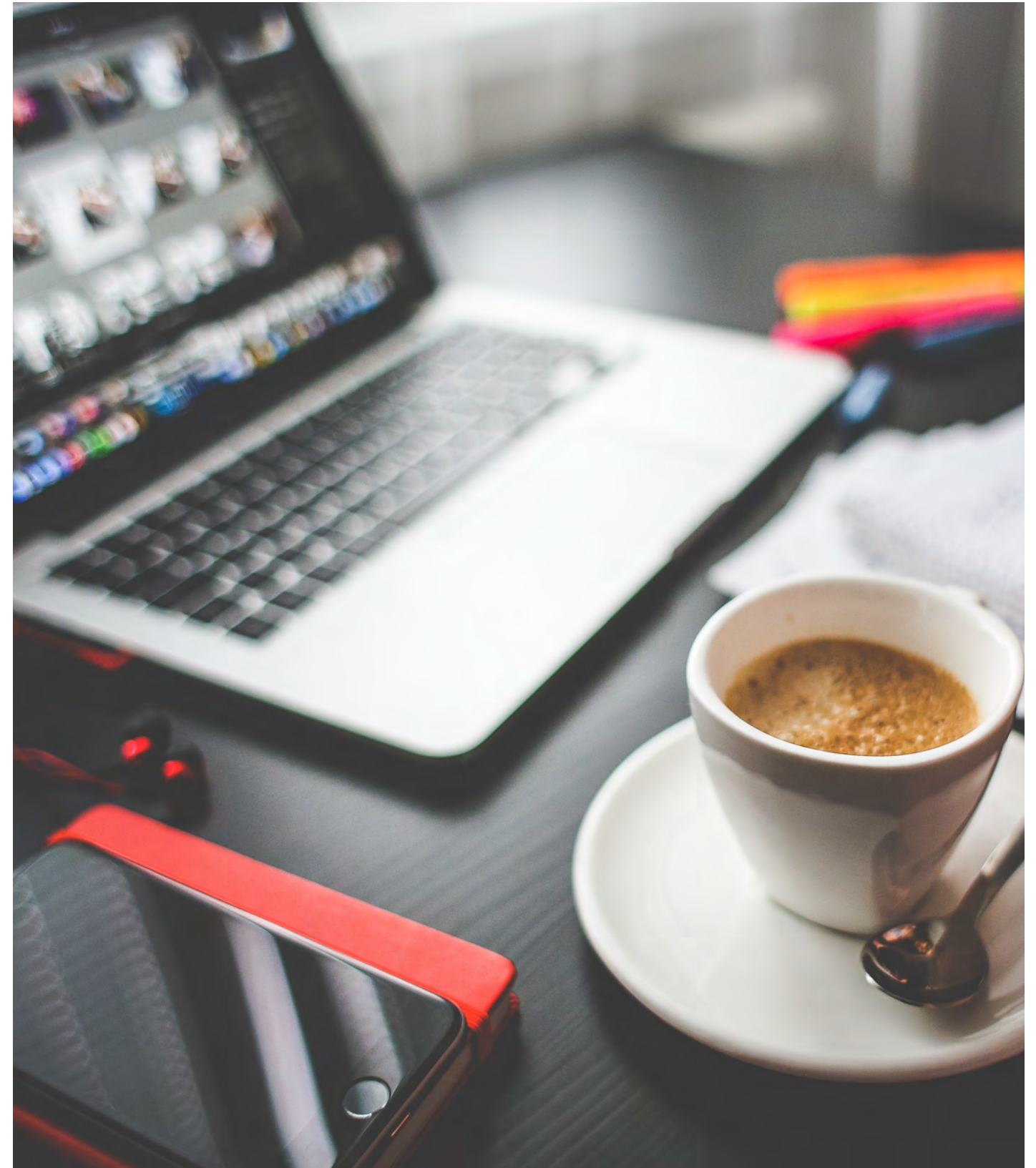


No mundo atual, quanto maior a quantidade de dados, maior é a quantidade de informações disponíveis e, por consequência, maior o conhecimento adquirido pela humanidade. Tudo isso gera a necessidade de pessoas para interpretar dados e, a partir disso, criar previsões. Esse profissional é o cientista de dados.

No passado, a maior parte dos dados era não processada, ou seja, não era transformada em informação. Hoje, com a capacidade de processamento em nuvem, as empresas estão buscando transformar dados em informações para interpretá-las e gerar insights importantes para seus negócios.

Conceito de Insight

Normalmente, os negócios apresentam problemas que necessitam de soluções. Por sua vez, as soluções necessitam de decisões tomadas em cima de dados.



Insight é a solução ou a conclusão acerca de algo. Se eu tenho um problema eu posso concluir algo sobre esse problema, que seria o insight. Sob o ponto de vista dos negócios, todo processo de decisão deveria ser baseada em dados, por isso a importância dos insights.

Com os conceitos de dado, informação e insight solidificados podemos partir para tentar entender a ciência de dados.



O que é data science?

Hoje, a profissão de data science é a que mais cresce no mundo. Muito disso é causado a partir da necessidade que as empresas têm de tratar dados não estruturados e transformá-los em informações úteis.

Segundo especialistas, estima-se que em torno de 90% dos dados armazenados na web tenham sido gerados somente nos últimos 2 anos. Além disso, no máximo 20% destes dados estão estruturados em linhas e colunas para serem analisados por ferramentas tradicionais. Um vídeo subido no Youtube, por exemplo, é considerado um dado não estruturado, pois é composto por imagem e áudio. Ou seja, não possui informações organizadas em categorias (etiquetadas).

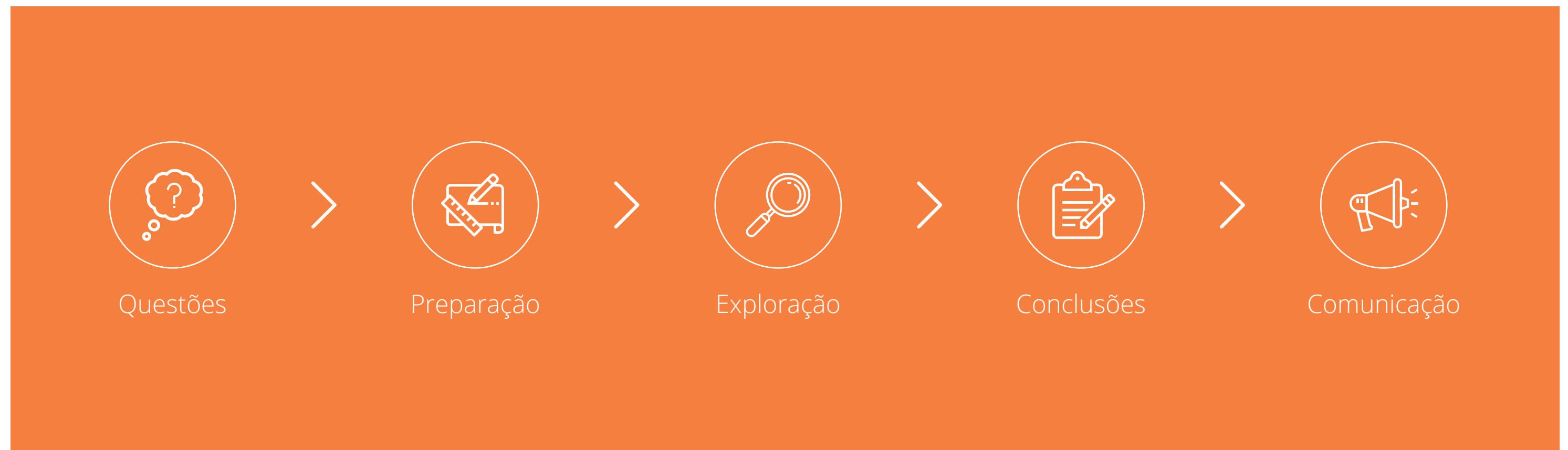
Data science é a coleta de dados de diversas fontes para analisar e subsidiar a tomada de decisões, de forma preditiva, em grandes quantidades e gerando insights.

É importante lembrar que a predição não garante o futuro, é apenas uma ferramenta para melhorar o processo de decisão. Ou seja, o planejamento não é certeza, pois não está imune a falhas.

Ciência de dados, como é conhecida em português, é o processo que extrai dados de diversas fontes, em diferentes velocidades, processando grandes quantidades (big data) e gerando

valor. De modo algum pode ser entendida como uma ferramenta, mas sim como um conjunto de métodos, assim como big data e o business intelligence.

Geralmente o processo de data science é composto por definição dos problemas ou questões, preparação, exploração, conclusão e comunicação. Veja o esquema:



Pilares da Ciência de dados

Entre os principais pilares da ciência de dados estão a matemática, a estatística, a área de negócios, a mineração e a visualização de dados, a programação e a computação. Embora essa área seja essencialmente multidisciplinar, a estatística e a matemática são a base da ciência de dados e o diferencial de métodos anteriores, pois é através delas que são construídos os modelos de análise de dados para predição futura, também conhecidos como algoritmos.

Qual a diferença entre Data Science e BI?

Existe uma certa confusão a respeito da diferença entre data science e business intelligence (BI). O que não é de se estranhar,

uma vez que as duas disciplinas são bastante parecidas e utilizam como insumo grandes quantidades de dados. Entretanto, apesar de trabalharem com o mesmo objetivo, suas abordagens, tecnologias e funções diferem de várias maneiras.



O objetivo do business intelligence é converter dados brutos em insights de negócio para que líderes empresariais possam tomar decisões. O profissional de BI, o analista de negócios, utiliza ferramentas para criar produtos de apoio à gestão, como dashboards e relatórios.

Já a ciência de dados emprega o método científico para a exploração dos dados, formação de hipóteses e testes de hipóteses, por meio de simulação e modelagem estatística. Dentro da ciência de dados ainda se utiliza o machine learning como ferramenta para automatizar a transformação de dado em informação.

A principal diferença entre os dois é que o business intelligence trata de dados do passado enquanto o data science vai tratar do futuro, a partir da análise preditiva. Às vezes, o profissional do BI até pode fazer algumas previsões acerca do futuro, mas elas são baseadas em extrapolações do passado, ou seja, não utilizam base científica.



E afinal, o que é machine learning?

O aprendizado de máquinas é uma subárea da **inteligência artificial** onde se programa uma máquina para aprender com dados coletados. Essa aprendizagem pode ser supervisionada ou não supervisionada.

Um exemplo de aprendizado de máquinas é o reconhecimento de imagens que permite que inteligências artificiais atinjam probabilidades satisfatórias e alto grau de precisão para categorizar fotografias. Um bom exemplo para isso é o aprendizado supervisionado pelos usuários do Google ao preencher um captcha que ajuda inteligências artificiais a identificar padrões e diferenciar objetos em imagens.

```
val training_df = sqlContext.createDataFrame(training_pca)
val test_df = sqlContext.createDataFrame(test_pca)

val tokenizer = new Tokenizer()
val hashingTF = new HashingTF()
val lr1 = new LogisticRegression()
val pipeline = new Pipeline()

val mod = pipeline.fit(training_df)
```


Aplicações de data science

A ciência de dados possui diversas aplicações práticas. Algumas delas são a recomendação de produtos no varejo online, o reconhecimento de voz (deep learning), o tratamento de doenças a partir de correlações de dados e o reconhecimento facial.

Hoje, diversos fabricantes de tecnologia estão investindo pesado em tecnologias de deep learning para reconhecimento de voz. Cortana (Microsoft), Siri (Apple) e Alexa (Amazon) são alguns exemplos de tecnologias conversacionais, que permitem que o usuário interaja com uma inteligência artificial por meio de comandos de voz. Essa tecnologia revela de forma bastante compreensiva como funciona a transformação entre dados não estruturados (voz) em informações úteis (comandos computacionais).



Carreiras de Data Science

A carreira de data science é uma das mais promissoras neste momento. Ainda que existam inúmeras possibilidades de profissões a se seguir, há 3 perfis macros que agregam todas as possibilidades atuais da profissão.

1. Cientista de dados

Profissional com um forte background em exatas, como ciência da computação, matemática e estatística. É capaz de analisar grandes quantidades de dados e chegar a conclusões (insights). Certamente é um perfil mais completo, equivalente ao analista de Business Intelligence, que mistura conhecimentos de negócios com exatas.



2. Analista de negócios

Tal como na área de BI, a área de Data Science também precisa do profissional com perfil de negócios, que seja capaz de entender o coração da empresa, sugerindo novas práticas ou negócios de forma a gerar mais valor.

Muitas empresas têm buscando profissionais com diferentes conhecimentos aliados à ciência de dados. Profissionais que dominam determinada área de negócio podem gerar resultados incríveis aliados à programação (R e Python).

3. Analista de tecnologia

A área de Data Science, tal como BI, também precisa de um profissional com perfil tecnológico e de infraestrutura. Devido à grande quantidade de dados que esse profissional trabalhará, é necessária a administração de clusters para processamento paralelo dos dados, sejam eles estruturados ou não estruturados.

O Futuro do Ciência de Dados

O data science, por meio de suas previsões, nos informa com antecedência se devemos sair de casa com um guarda-chuva, qual o melhor caminho para chegar ao trabalho e de qual filme temos maior possibilidade de gostar baseado em nossas preferências anteriores. À medida que deixam de ser um recurso escasso, passando a ser cada vez mais abundantes, os dados tornam-se uma fonte essencial de benefícios sociais e econômicos.

Com o custo de armazenamento e processamento de dados em queda e com o aumento do número de sensores que capturam cada vez mais informações, a quantidade de dados disponíveis será cada vez maior, assim como as possibilidades de uso desses dados. Vivemos cercados por oportunidades geradas pelos dados, que podem nos dar respostas a alguns dos maiores desafios globais, como a maior eficiência dos recursos

de saúde ou a reestruturação dos sistemas de transporte ou a melhor alocação de recursos governamentais.

Cabe aos profissionais desse novo campo da ciência criar modelos para potencializar a produtividade de todas as áreas. Não existe restrição em nenhuma área para o trabalho dos cientistas de dados, o que é uma grande oportunidade para tornar o esforço humano cada vez mais eficiente.



GOSTOU DO NOSSO CONTEÚDO E
QUER SABER MAIS SOBRE AS NOSSAS
TECNOLOGIAS E SERVIÇOS?

Somos especializados em desenvolver projetos para a coleta e visualização de dados de clientes de médio e grande porte.

ACESSE O NOSSO SITE

Sobre a OpServices

Empresa especializada em **monitoração e visualização de dados**, com quase 20 anos de mercado. É reconhecida por desenvolver soluções e serviços de BI orientados às necessidades específicas de cada cliente.

marisa

LEAR

Grendene

SantaCruz

ENGIE

Zaffari HIPERMERCADOS **BOURBON**

Unimed 
Curitiba

 **Klabin**

Unimed 
Porto Alegre

 **Banrisul**

FINDES
PELO FUTURO DA INDÚSTRIA

 **UNIGEL**

Conheça nossas redes sociais:   