CLEI LACLO 2018

Análise de dados com R

uma visão inicial das atividades de um cientista de dados

Rosangela de Fátima Pereira Marquesone – rpereira@larc.usp.br
Francisco Pereira Junior – fpereira@utfpr.edu.br
Tereza Cristina Melo de Brito Carvalho – terezacarvalho@usp.br

02/10/2018

Tópicos

- Introdução à análise de dados
- Etapas do processo de análise de dados
- Casos de uso
- O papel do cientista de dados

Vivemos em um mundo **conectado**

Big Data faz referência ao grande
volume, variedade e velocidade de dados
que demandam formas inovadoras e rentáveis
de processamento da informação,
para melhor percepção e tomada de decisão.

Gartner

Big Data faz referência ao grande volume, variedade e velocidade de dados

que demandam formas inovadoras e rentáveis de processamento da informação, para melhor percepção e tomada de decisão.

Gartner

VOLUME

40 milhões de artigos

Wikipedia

1.23 bi usuários ativos/dia

Facebook

80 milhões de fotos por dia

Instagram

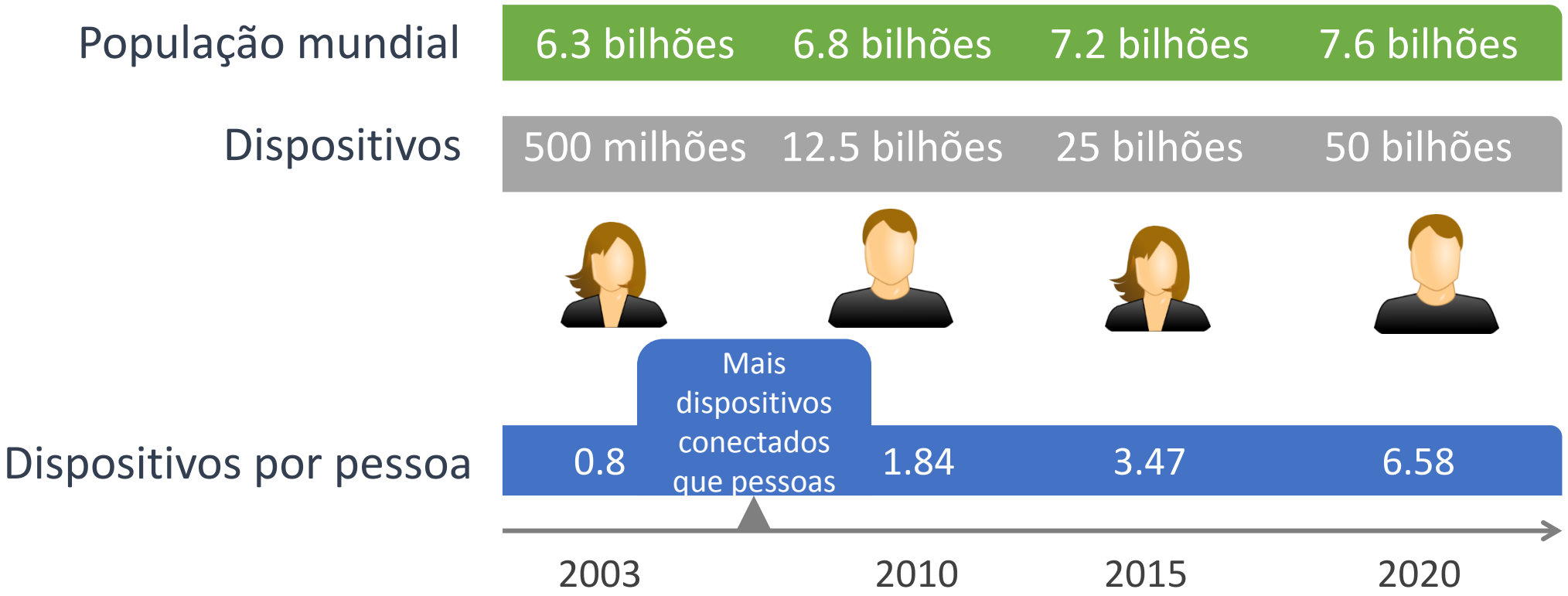
1 bilhão usuários

Whatsapp

4 bilhões Visualizações por dia

Youtube

VOLUME



Fonte: https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

VARIEDADE

80%

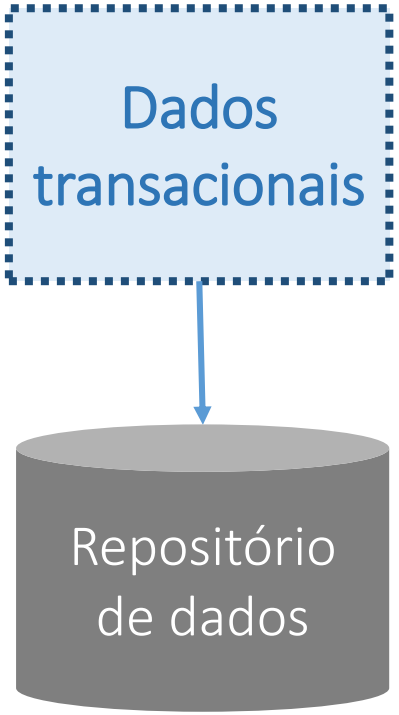
dos dados globais atualmente são

não estruturados

Fonte: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

CENÁRIO TRADICIONAL

VARIEDADE

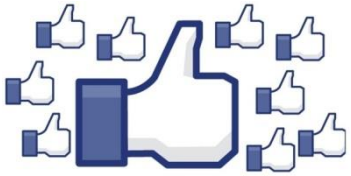


CENÁRIO ATUAL

VARIEDADE

Dados transacionais

Repositório de dados



O que acontece em 30 segundos na Internet?

VELOCIDADE

VELOCIDADE

O valor dos dados é reduzido com o passar do tempo

Análise de dados (Data analytics)

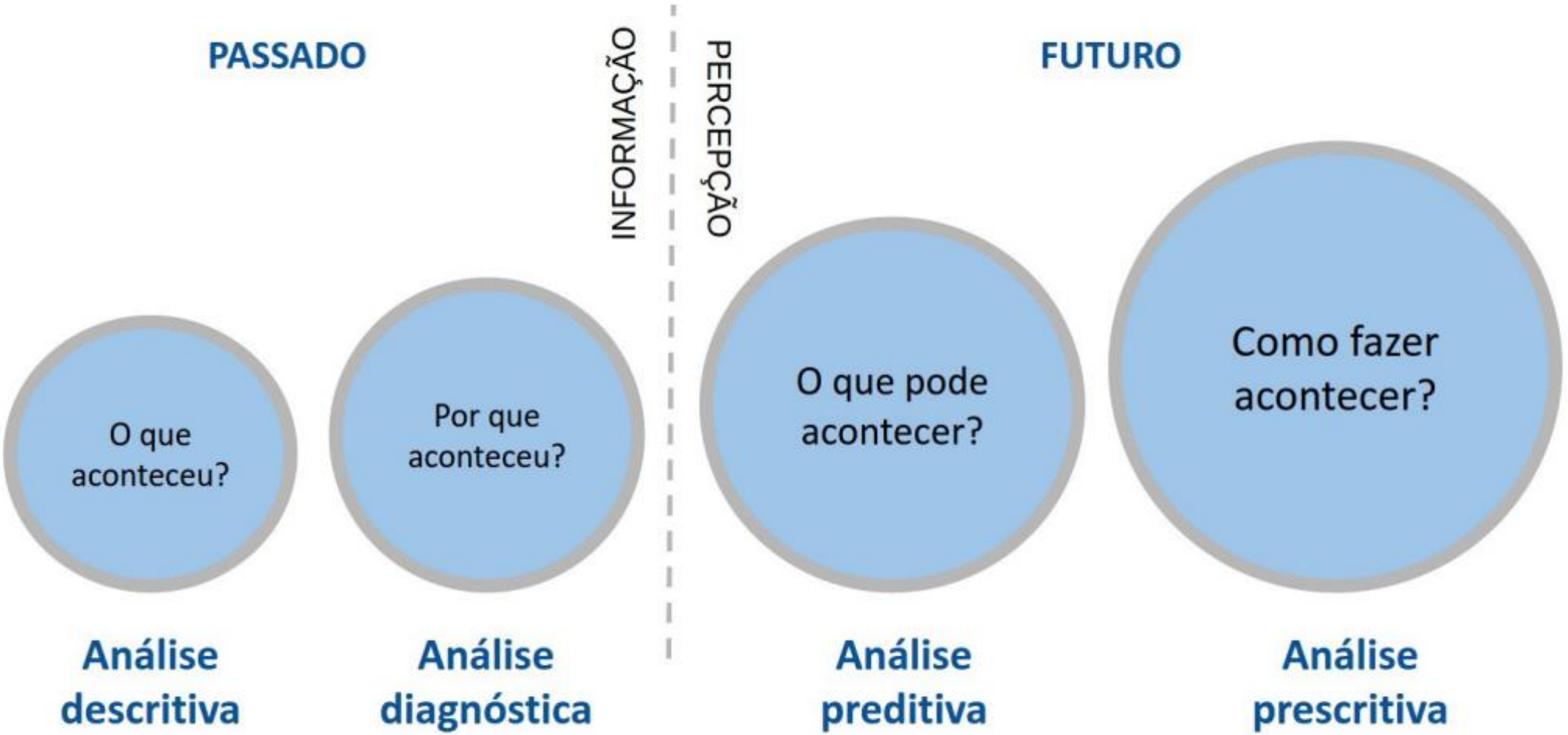
Analytics = aplicação da tecnologia computacional e de estatística para solucionar problemas de negócios e da indústria

- Tomada de decisões baseadas em dados existentes
  - Não somente na intuição humana

- Oferece meios para obter vantagem competitiva

- Permite otimizar novas iniciativas

Categorias de análise de dados



Descritiva

- Foco em sumarizar **fatos ocorridos**
- Permite compreender como a organização está sendo operada
- Uso de ferramentas de *Business Intelligence* (BI)
  - Alertas
  - Dashboards
  - Relatórios

O que aconteceu?

Descritiva

- Baseado em métricas
  - Visualização de tendências
  - Descoberta de padrões

- Análise mais adotada pelas organizações
  - Estima-se que mais de 80% das análises de negócios são descritivas*

- *Exemplos:*
  - Relatório diário/semanal de vendas
  - Lista de cancelamento de assinantes

*Fonte: https://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279

Diagnóstica

- Foco em determinar o **motivo de um evento ter ocorrido**
- Demonstra variações de desempenho positivas e negativas
  - Ex.: segmentação de clientes
- Gráfico de controle é uma das técnicas mais utilizadas
- Uso de outros métodos estatísticos
  - Análise de variância
  - Testes de hipóteses

Por que aconteceu?

Preditiva

- Foco em **predições de eventos futuros**
- Extrai padrões encontrados em dados históricos
- Utilizado para diferentes aplicações
  - Detecção de fraude
  - Gerenciamento de risco
  - Fidelização de clientes

O que acontecerá?

Preditiva

- Requer uso de diversos métodos e ferramentas
  - Análises estatísticas
  - Técnicas de simulação
  - Mineração de dados
  - Aprendizado de máquina

Análise descritiva

Relatório climático

Análise preditiva

Previsão do tempo

Prescritiva

- Foco em **prever ações futuras e possíveis consequências**

- Sugere ações baseadas no conhecimento extraído dos dados

- Envolve regras de negócios, conhecimentos matemáticos, mineração de dados

Como fazer acontecer?

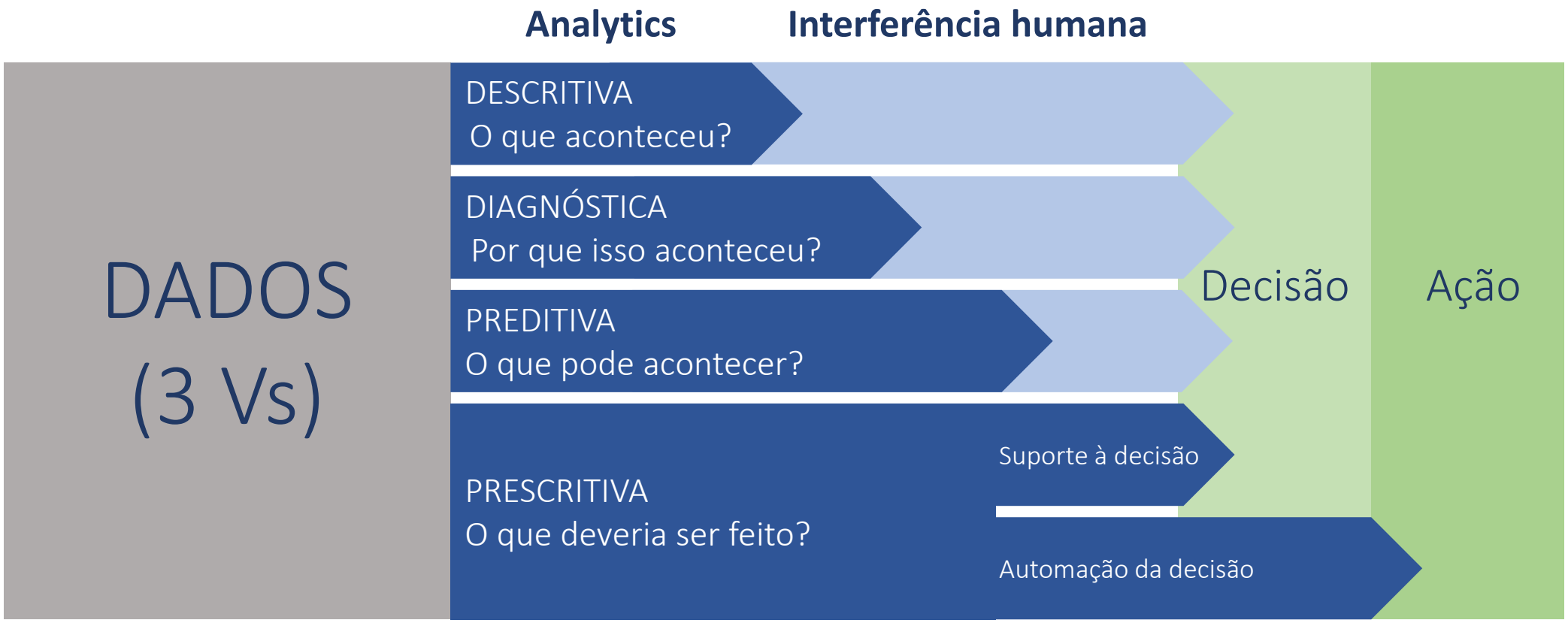
Prescritiva

- Segundo a consultora Gartner, somente 3% das empresas utilizam análise prescritiva

- Necessário grandes quantidades de dados

- Exemplos:
  - Ofertas promocionais baseadas na segmentação do cliente
  - Veículos autônomos
  - Manutenção inteligente

Fonte: https://searchbusinessanalytics.techtarget.com/feature/Prescriptive-analytics-takes-analytics-maturity-model-to-a-new-level



FONTE: Adaptado de http://www.datasciencecentral.com/profiles/blogs/prescriptive-versus-predictive-analytics-a-distinction-without-a

Tópicos

Introdução à análise de dados

Etapas do processo de análise de dados

Casos de uso

O papel do cientista de dados

1.
SELEÇÃO DOS DADOS

2.
PRÉ-PROCESSAMENTO DOS DADOS

3.
TRANSFORMAÇÃO DOS DADOS

4.
MINERAÇÃO DOS DADOS

5.
INTERPRETAÇÃO E VALIDAÇÃO

Seleção de dados

- Um dos desafios existentes na análise de dados é a identificação de fontes de dados apropriadas para responder questões feitas aos dados.

- Ou seja, é importante ter uma **definição clara** do que se deseja extrair de informação, pois é essa clareza que irá guiar a etapa de seleção, bem como as etapas seguintes.

- Para realizar a seleção de dados é importante compreender suas categorias, pois essa compreensão permitirá identificar e priorizar as estratégias de aquisição, gerenciamento e extração dos dados selecionados.

Seleção de dados

Os dados podem ser classificados de diferentes formas

- Interno
- Externo
- Estruturado
- Não estruturado
- Semiestruturado
- Gerados por humanos
- Gerados por máquinas
- Dados biométricos
- Dados numéricos
- Dados categóricos

Seleção de dados

Os dados podem ser classificados de diferentes formas

- **Interno**
- Externo
- Estruturado
- Não estruturado
- Semiestruturado
- Gerados por humanos
- Gerados por máquinas
- Dados biométricos
- Dados numéricos
- Dados categóricos

- Dados gerados dentro de uma organização.
- Dados que estão sob controle do negócio.
- Comumente utilizados no processo de tomada de decisão.

Seleção de dados

Os dados podem ser classificados de diferentes formas

- Interno
- **Externo**
- Estruturado
- Não estruturado
- Semiestruturado
- Gerados por humanos
- Gerados por máquinas
- Dados biométricos
- Dados numéricos
- Dados categóricos

- Dados gerados fora da organização.
- Utilizados para fornecer novos inputs ao negócio.
- Não estão sob o controle da organização.

dados.gov.br
PORTAL BRASILEIRO DE DADOS ABERTOS

Seleção de dados

Os dados podem ser classificados de diferentes formas

- Interno
- Externo
- **Estruturado**
- Não estruturado
- Semiestruturado
- Gerados por humanos
- Gerados por máquinas
- Dados biométricos
- Dados numéricos
- Dados categóricos

- Dados que possuem um esquema pré-definido rígido.
- Armazenados em banco de dados relacionais.
- Normalmente uma mudança na estrutura de um registro requer a mudança de estrutura de todos os demais.

Seleção de dados

Os dados podem ser classificados de diferentes formas

- Interno
- Externo
- Estruturado
- **Não estruturado**
- Semiestruturado
- Gerados por humanos
- Gerados por máquinas
- Dados biométricos
- Dados numéricos
- Dados categóricos

- Dados que não seguem um formato específico.
- Normalmente a estrutura é definida após o armazenamento dos dados, durante sua utilização.
- Não são indicados para o armazenamento em banco de dados relacionais.

Seleção de dados

Os dados podem ser classificados de diferentes formas

- Interno
- Externo
- Estruturado
- Não estruturado
- **Semiestruturado**
- Gerados por humanos
- Gerados por máquinas
- Dados biométricos
- Dados numéricos
- Dados categóricos

- Dados que possuem uma estrutura pré-definida, porém não tão rígida quanto os dados estruturados.
- Estrutura geralmente baseada em *tags*
- Permite que os dados se tornem auto-descritivos

Seleção de dados

Os dados podem ser classificados de diferentes formas

- Interno
- Externo
- Estruturado
- Não estruturado
- Semiestruturado
- **Gerados por humanos**
- Gerados por máquinas
- Dados biométricos
- Dados numéricos
- Dados categóricos

- Dados gerados diretamente por humanos na interação com computadores.
- São normalmente não estruturados.
- Incluem propriedade intelectual.

Seleção de dados

Os dados podem ser classificados de diferentes formas

- Interno
- Externo
- Estruturado
- Não estruturado
- Semiestruturado
- Gerados por humanos
- **Gerados por máquinas**
- Dados biométricos
- Dados numéricos
- Dados categóricos

- Dados gerados automaticamente por um processo computacional, aplicação ou outro mecanismo que não requer a interferência humana.

Seleção de dados

Os dados podem ser classificados de diferentes formas

- Interno
- Externo
- Estruturado
- Não estruturado
- Semiestruturado
- Gerados por humanos
- Gerados por máquinas
- **Dados biométricos**
- Dados numéricos
- Dados categóricos

- Dados gerados a partir da identificação automática de uma pessoa, baseando-se em características anatômicas e/ou comportamentais.



Seleção de dados

Os dados podem ser classificados de diferentes formas

- Interno
- Externo
- Estruturado
- Não estruturado
- Semiestruturado
- Gerados por humanos
- Gerados por máquinas
- Dados biométricos
- **Dados numéricos**
- Dados categóricos

- Dados quantitativos, expressos em termos numéricos.
- Podem ser discretos ou contínuos.
- Dados discretos representam dados contáveis e adotam apenas valores inteiros.
- Dados contínuos são variáveis numéricas (inteiras ou decimais).

Seleção de dados

Os dados podem ser classificados de diferentes formas

- Interno
- Externo
- Estruturado
- Não estruturado
- Semiestruturado
- Gerados por humanos
- Gerados por máquinas
- Dados biométricos
- Dados numéricos
- **Dados categóricos**

- Dados que podem ser divididos em grupos.
- As variáveis nominais não apresentam uma relação de maior/menor entre elas, como por exemplo a variável gênero.
- Nas variáveis ordinais há uma relação de ordem que as define, permitindo a realização de comparações entre elas.

Seleção de dados

Em resumo, algumas das principais atividades no processo de seleção de dados, são:

1. Identificar perguntas que se deseja responder a partir da análise dos dados.
2. Identificar quais fontes de dados podem auxiliar na extração de informação útil.
3. Categorizar os dados de acordo com sua estrutura.
4. Identificar em qual aspecto os dados selecionados podem contribuir com a análise.

1.
SELEÇÃO DOS DADOS

2.
PRÉ-PROCESSAMENTO DOS DADOS

3.
TRANSFORMAÇÃO DOS DADOS

4.
MINERAÇÃO DOS DADOS

5.
INTERPRETAÇÃO E VALIDAÇÃO

Pré-processamento dos dados

- Na maioria das situações, os dados a serem utilizados podem conter as seguintes características:

Quantidade significativa de outliers	Dados duplicados
Dados incorretos e/ou irrelevantes	Dados ausentes

- Considerada uma das etapas primordiais no processo de análise de dados, o pré-processamento compreende um conjunto de técnicas para a limpeza e a preparação dos dados para a análise.

Pré-processamento dos dados

- Esse processo investigativo, na qual os dados são avaliados e compreendidos com detalhes é chamado de **análise exploratória de dados**.

- Nesse processo, inúmeras medidas, tais como média, mediana e desvio padrão são geradas com o objetivo de aumentar a compreensão dos dados.

- Além disso, a análise exploratória também faz uso de recursos visuais para auxiliar na investigação. Inúmeros gráficos podem ser gerados nesse processo, tais como o *boxplot*, histograma e *scatterplot*.

Pré-processamento dos dados

Em resumo, algumas das principais atividades no processo de pré-processamento de dados, são:

1. Explorar e visualizar os dados em busca de conhecê-los em detalhes, identificando necessidades de melhoria.
2. Desenvolver e aplicar técnicas de limpeza dos dados, aumentando a confiabilidade dos dados para a análise.
3. Gerar e armazenar novos conjuntos de dados mediante ao pré-processamento das fontes de dados anteriormente selecionadas.

1. SELEÇÃO DOS DADOS

2. PRÉ-PROCESSAMENTO DOS DADOS

3. TRANSFORMAÇÃO DOS DADOS

4. MINERAÇÃO DOS DADOS

5. INTERPRETAÇÃO E VALIDAÇÃO

Transformação dos dados

- O pré-processamento dos dados permite aumentar a confiabilidade dos dados selecionados.

- Porém, é comum que, mesmo após o pré-processamento, os dados ainda não estejam preparados de acordo com a necessidade da análise a ser realizada.

Transformação dos dados

- Para estas situações, surge a etapa de transformação dos dados e as principais ações realizadas nessa etapa são:

Normalização dos dados

Discretização dos dados

Enriquecimento dos dados

Agregação dos dados

Redimensionamento dos dados

Transformação dos dados

- Para estas situações, surge a etapa de transformação dos dados e as principais ações realizadas nessa etapa são:

Normalização dos dados

Discretização dos dados

Enriquecimento dos dados

Agregação dos dados

Redimensionamento dos dados

- Técnica utilizada para organizar os atributos em um intervalo específico, caso esses possuam valores em intervalos diferentes.

Transformação dos dados

- Para estas situações, surge a etapa de transformação dos dados e as principais ações realizadas nessa etapa são:

Normalização dos dados

Discretização dos dados

Enriquecimento dos dados

Agregação dos dados

Redimensionamento dos dados

- Técnica que visa o mapeamento do domínio de um atributo contínuo para um discreto.
- Nessa técnica ocorre um processo de divisão de um conjunto de atributos contínuos em intervalos categóricos.

Transformação dos dados

- Para estas situações, surge a etapa de transformação dos dados e as principais ações realizadas nessa etapa são:

Normalização dos dados

Discretização dos dados

Enriquecimento dos dados

Agregação dos dados

Redimensionamento dos dados

- Técnica que refere-se à complementação e atualização de informações, e até mesmo a inserção de novos atributos em um conjunto de dados.
- Visa melhorar a qualidade dos dados.

Transformação dos dados

- Para estas situações, surge a etapa de transformação dos dados e as principais ações realizadas nessa etapa são:

Normalização dos dados

Discretização dos dados

Enriquecimento dos dados

Agregação dos dados

Redimensionamento dos dados

- Técnica que visa a combinação de dados de múltiplas fontes em um único conjunto de dados.

Transformação dos dados

- Para estas situações, surge a etapa de transformação dos dados e as principais ações realizadas nessa etapa são:

Normalização dos dados

Discretização dos dados

Enriquecimento dos dados

Agregação dos dados

Redimensionamento dos dados →

- Técnica que tem como objetivo encontrar uma projeção linear de um conjunto de dados correlacionados, gerando um conjunto substancialmente menor de variáveis não correlacionadas, mantendo a informação do conjunto original.

80%

do processo de análise de dados é normalmente gasto preparando os dados

(pré-processamento + transformação)

Transformação dos dados

Em resumo, algumas das principais atividades no processo de transformação dos dados, são:

1. Identificar, a partir do modelo que será utilizado na mineração de dados, quais dados necessitam ser normalizados.
2. Avaliar e gerar novos atributos a partir dos dados existentes, visando o enriquecimento dos dados.
3. Explorar outras fontes de dados existentes, avaliando a possibilidade de integrá-las aos dados selecionados.

1. SELEÇÃO DOS DADOS

2. PRÉ-PROCESSAMENTO DOS DADOS

3. TRANSFORMAÇÃO DOS DADOS

4. MINERAÇÃO DOS DADOS

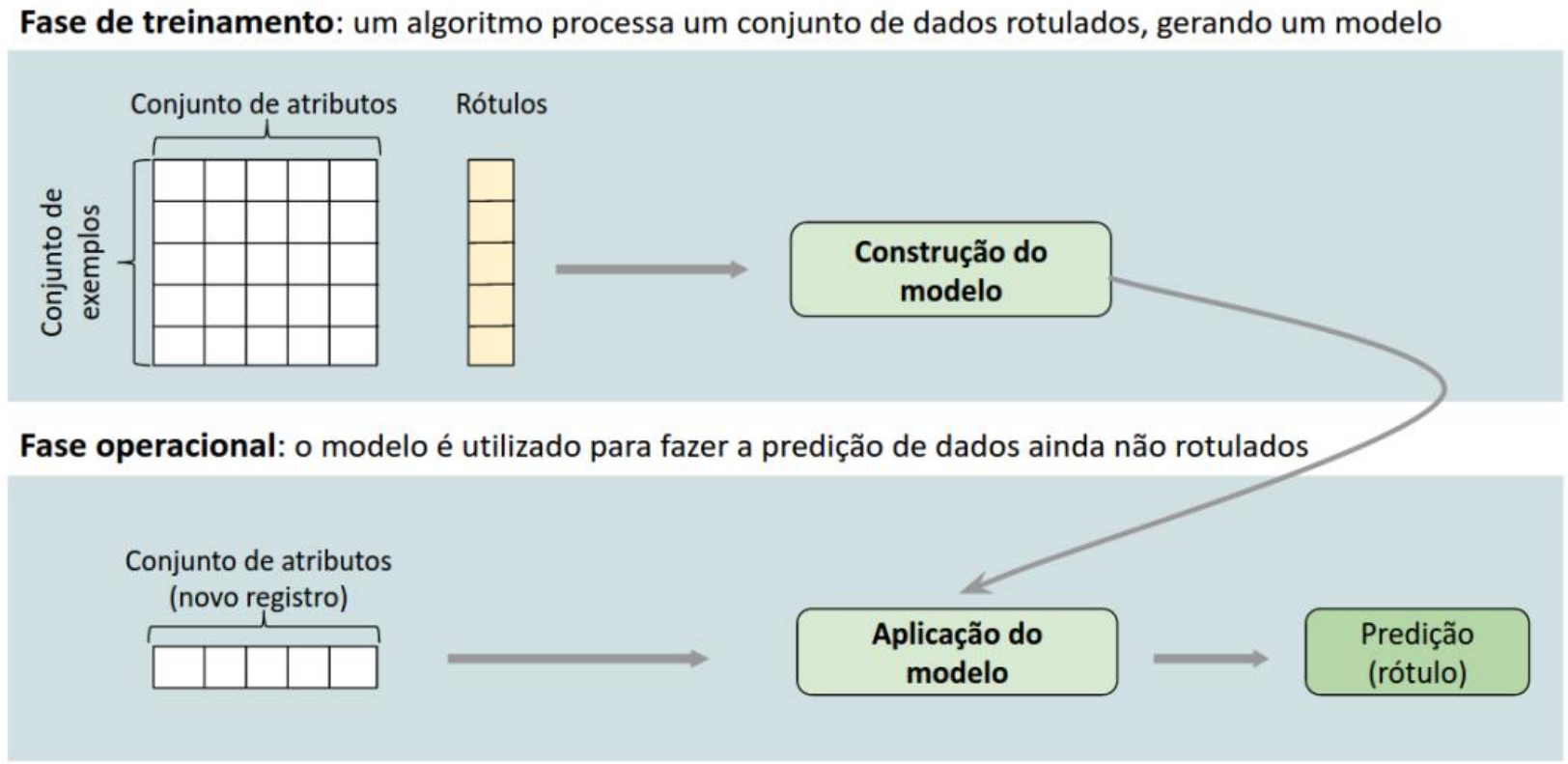
5. INTERPRETAÇÃO E VALIDAÇÃO

Mineração de dados

- Etapa considerada como fase de aprendizado, com objetivo de reduzir a incerteza sobre uma predição, melhorando a capacidade de se prever algo por meio dos dados.
- Uma das primeiras atividades da etapa de mineração de dados consiste em selecionar o algoritmo a ser utilizado para a mineração e análise dos dados.
- Esses algoritmos podem ser divididos nas seguintes categorias: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado por reforço.

Mineração de dados

- Nos algoritmos de **aprendizado supervisionado** há uma variável de saída (rótulo) para todas as observações que serão utilizadas na construção do modelo.



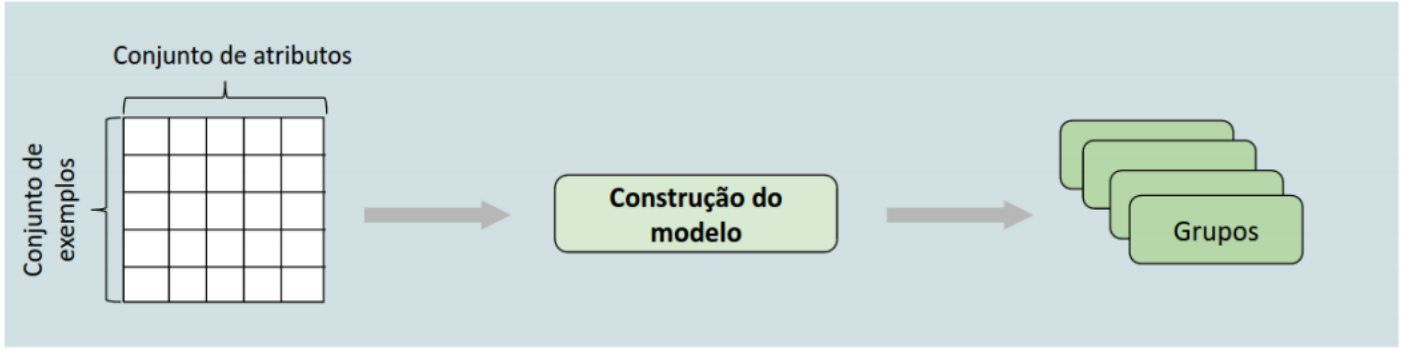
Mineração de dados

- Para prever valores numéricos, diversos algoritmos são disponibilizados, tais como a regressão linear simples e regressão logística.
- Caso o objetivo seja prever um valor a partir de um conjunto finito de classes, deve-se utilizar modelos de classificação, com algoritmos tais como Random Forest e SVM (Support Vector Machine).
- Exemplos de análises por meio de classificação são:
  - Prever o risco de ceder o crédito a um determinado cliente;
  - Prever o diagnóstico de um paciente com base nos resultados dos exames;
  - Prever a possibilidade de um e-mail ser considerado spam ou não-spam.

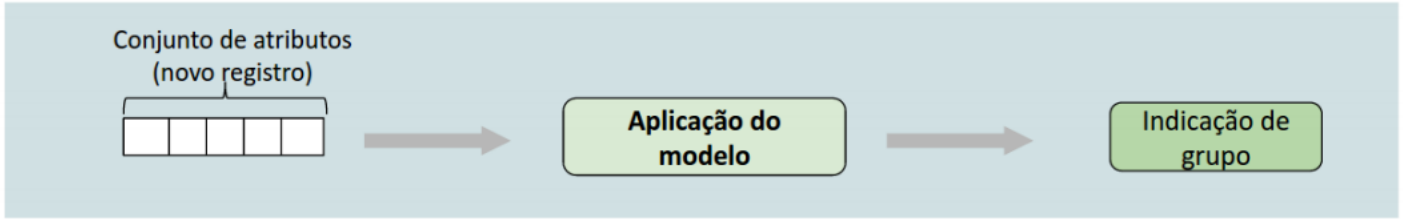
Mineração de dados

- Não havendo a variável de saída discriminada, uma outra estratégia é a utilização de algoritmos de **aprendizado não-supervisionado**. Nessa categoria, não há uma clareza sobre as possíveis saídas do modelo, espera-se que o algoritmo seja capaz de identificar padrões nos dados de entrada.

Fase de treinamento: um modelo é construído para detectar padrões/grupos sobre dados não rotulados



Fase operacional: um novo registro é aplicado ao modelo, que deverá inferir à qual grupo ele pertence



Mineração de dados

- São exemplos de mineração de dados por meio de agrupamento:
  - Segmentar clientes a partir de uma base de dados, para campanhas de marketing;
  - Identificar ações fraudulentas de ações legítimas, de acordo com o comportamento de uma compra;
  - Identificar imagens similares em bases de dados de imagens médicas.
- Entre as técnicas existentes nesse contexto, pode-se citar: k-means, agrupamento fuzzy e agrupamento hierárquico.

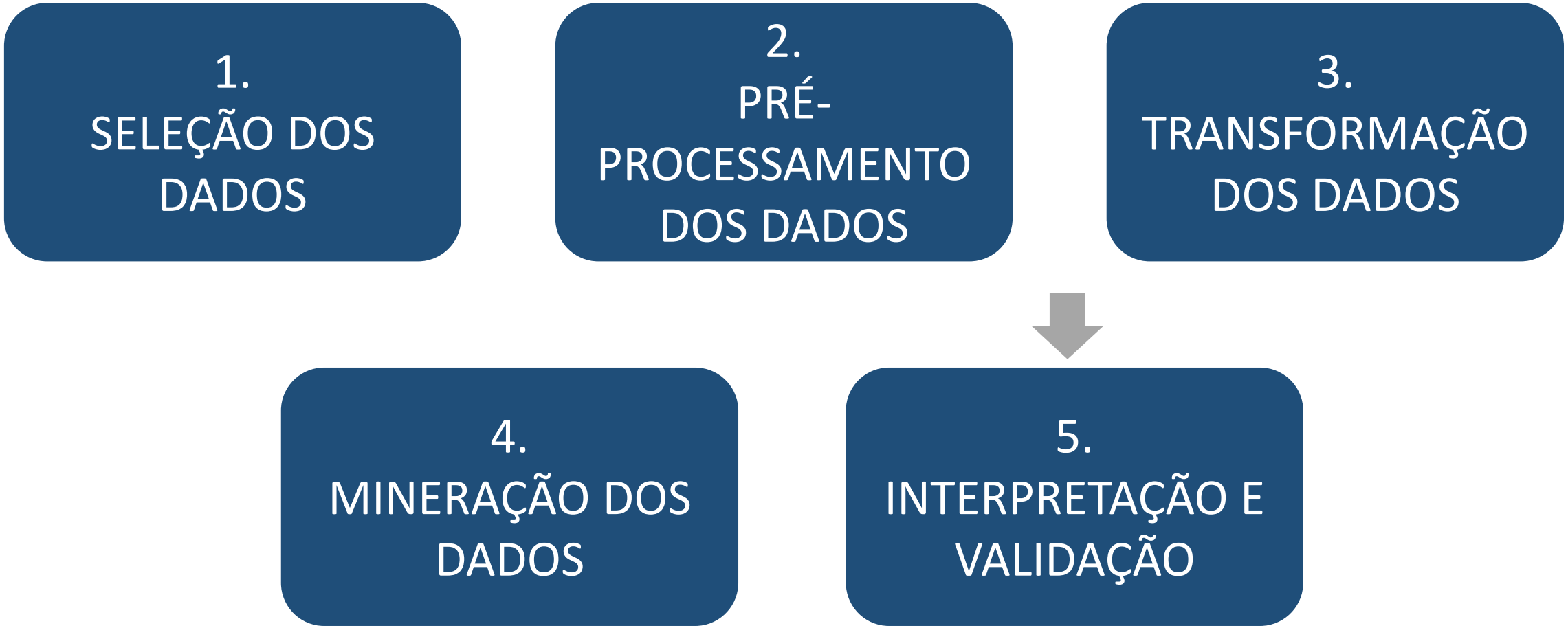
Mineração de dados

- Os algoritmos de **aprendizado por reforço** são utilizados em cenários nos quais os dados das variáveis de saída não estão disponíveis, porém, o modelo recebe um conjunto de informações ligadas à essa variável.
- Como exemplo pode-se citar o controle de movimentos de um robô e a definição de passos em uma partida de xadrez.
- OBS.: Cabe ressaltar também que há ainda algoritmos que podem ter o comportamento tanto de aprendizado supervisionado quanto de aprendizado não-supervisionado, como é o caso das redes neurais.

Mineração dos dados

Em resumo, algumas das principais atividades no processo de mineração dos dados, são:

1. Separar bases de dados entre conjuntos de treinamento e de teste.
2. Selecionar o melhor algoritmo para a construção do modelo.
3. Construir o modelo a partir do conjunto de treinamento.


1.
SELEÇÃO DOS DADOS
2.
PRÉ-PROCESSAMENTO DOS DADOS
3.
TRANSFORMAÇÃO DOS DADOS
4.
MINERAÇÃO DOS DADOS
5.
INTERPRETAÇÃO E VALIDAÇÃO

Interpretação e validação

- Uma vez que os modelos não chegam a um nível de acerto de 100%, é necessário identificar qual o limite mínimo de acerto do modelo que está sendo proposto.
- Essa etapa refere-se a um conjunto de técnicas para a interpretação e validação do modelo desenvolvido.
- Uma baixa taxa de erro de validação indica que o modelo treinado contém uma boa representação da relação de entrada/saída dos dados.
- Nessa etapa também insere-se a validação cruzada, um método utilizado para validar os modelos gerados, evitando a ocorrência de overfitting.

Interpretação e validação dos dados

Em resumo, algumas das principais atividades no processo de interpretação e validação dos dados, são:

1. Executar o modelo com um conjunto de teste.
2. Avaliar as métricas de validação, identificando se a qualidade do modelo está de acordo com o esperado.
3. Ajustar o modelo ou os dados e construí-lo novamente, caso a qualidade do modelo não tenha atingido o desempenho esperado.

Tópicos

Introdução à análise de dados

Etapas do processo de análise de dados

Casos de uso

O papel do cientista de dados

Medicina

Photo by rawpixel on Unsplash

Medicina

- Área composta por uma variedade de dados, extraídos de diferentes fontes.
- Um conjunto de dados muito utilizado refere-se aos dados clínicos, extraídos de registros eletrônicos de saúde, compostos por informações como anotações médicas, informações de tratamentos, medicamentos e procedimentos laboratoriais de um paciente.
- Devido ao formato não estruturado, as análises de tais dados utilizam, em geral, técnicas de processamento de linguagem natural para extrair informações dos dados textuais.

Medicina

Como exemplos de como a análise de dados pode proporcionar avanços na medicina, pode-se citar:

- Análise de genes;
- Diagnóstico de câncer;
- Aceleração no desenvolvimento de medicamentos;
- Monitoramento de doenças;
- Apoio na tomada de decisões clínicas.

Caso de uso - Medicina

Projeto "Google flu trends"

- Alternativa para o monitoramento de doenças infecciosas guiados por dados.
- Utilizou dados de consulta de pesquisa na Internet para prever tendências de disseminação de doenças.
- Como resultado, a aplicação desenvolvida foi capaz de identificar surtos de gripe na população cerca de duas semanas antes dos sistemas convencionais de saúde.

Fonte: https://www.google.org/flutrends/about/

Marketing

Photo by rawpixel on Unsplash

Marketing

- Muitos dos dados utilizados nessa área são oriundos atualmente de mídias sociais (redes sociais online, blogs, comunidades, fóruns), por essas produzirem volumes de dados em larga escala.
- Isso permite aos especialistas realizar um grande número de análises, tais como análises comportamentais, de segmentação e de fidelização, além de gerar um vasto número de indicadores de desempenho.
- A área de marketing em que as ações são apoiadas por dados é atualmente denominada ***data-driven marketing.***

Marketing

Entre as atividades de um cientista de dados em marketing, pode-se citar:

- Geração de sumarizações que revelem diferenças entre grupos de clientes.
- Análise de cesto de compras, na qual se identifica a partir de uma base de dados histórica, quais itens são comprados em conjunto em uma única compra.
- Análise de sentimentos, permitindo compreender a atitude ou a reação emocional de um sujeito diante de um tópico ou marca específica.
- Análise comportamental, permitindo compreender o comportamento dos clientes em cada canal e ponto de interação, para assim identificar quais aspectos influenciam suas ações.

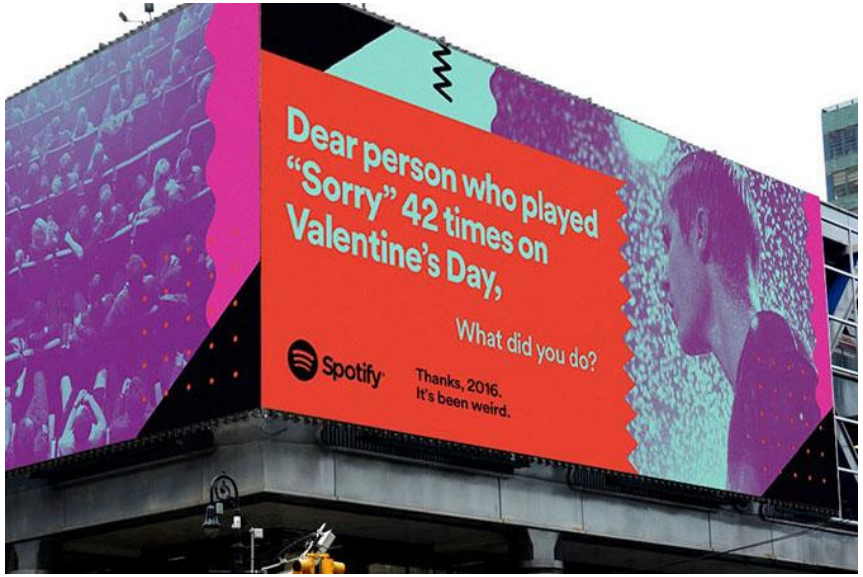
Caso de uso - Marketing

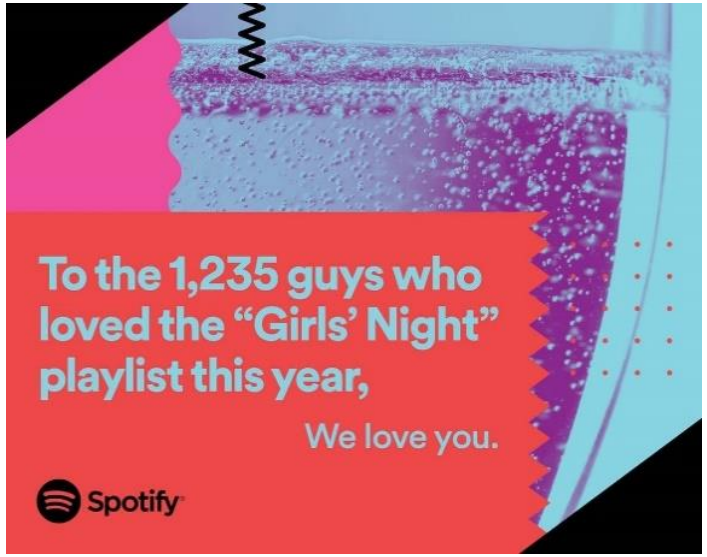
Estratégias de marketing do serviço de streaming de músicas Spotify.

- Para compreender melhor seu cliente, oferecer uma melhor experiência, e realizar campanhas de marketing efetivas, a empresa realiza análise a partir de 3 bases principais: (1) de playlists; (2) histórico de audições individuais e (3) likes, dislikes e saltos de músicas.

- Em 2016, por meio de uma equipe formada por profissionais de marketing e cientistas de dados, a empresa identificou comportamentos considerados relevantes ou engraçados de seus usuários, utilizando tais informações para uma campanha global chamada *"Thanks, 2016. It's been weird".*

Caso de uso - Marketing





Fonte: http://newronio.espm.br/spotify-e-a-criatividade-do-big-data/

Varejo

Photo by rawpixel on Unsplash

Varejo

Área que categoriza as estratégias de análises de dados em 4 níveis, sendo eles:

- Mercado: análises referentes à precificação, expansão de mercado, *marketing share* e publicidade.
- Empresa: análises relacionadas ao marketing multi-canais, alianças e marcas de loja.
- Loja: análises e estratégias associadas à localização, incluindo, por exemplo, precificação dinâmica e promoções.
- Cliente: análises que permitem identificar aspectos sobre a experiência, fidelização, satisfação e engajamento do usuário.

Caso de uso - Varejo

Empresa varejista notória no uso de dados: Walmart

- Já em 2012, estimava-se que a empresa coletava mais de 2.5 petabytes de dados a cada hora, referentes às transações de seus clientes.
- Além desse volume em larga escala, a empresa também revela ser capaz de pesquisar bilhões de documentos de mídias sociais, identificando fatores como tendências, produtos populares, e sentimentos dos clientes, inclusive separado por localização geográfica.

Fonte: https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109

Telecomunicações

Photo by rawpixel on Unsplash

Telecomunicações

- Os dados coletados no setor de telecomunicações são capazes de evidenciar tendências e comportamentos dos clientes de uma operadora.
- Exemplos incluem os dados de registro de chamada (CDR - *Call Detail Record*), dados de comportamento do usuário em mídias sociais, dados de histórico de bilhetagem, histórico de compras e de avaliações de serviço.
- Além desse objetivo, tais dados também podem ser utilizados internamente nas empresas de telecomunicações, permitindo otimizar atividades de gerenciamento financeiro, de recursos de infraestrutura, cadeia de suprimentos, entre outros.

Caso de uso - Telecomunicações

Exemplo notório no uso de dados: Telefônica – Plataforma Smart Steps

- Construída com o objetivo de extrair informações referentes à movimentação da população no espaço e no tempo.
- Contém funcionalidades para inferir o movimento de um conjunto aglomerado de pessoas a partir da infraestrutura de telefonia móvel.
- Os dados foram capturados em nível nacional, obtendo informações de registros de celulares de aproximadamente 80 milhões de pessoas.
- Essa vasta quantidade de informações permitiu identificar padrões e obter novas percepções em cenários como a previsão de congestionamentos de veículos no trânsito e a necessidade de maior demanda de infraestrutura de rede.

Fonte: https://www.business-solutions.telefonica.com/en/products/big-data/business-insights/smart-steps/

Tópicos

Introdução à análise de dados

Etapas do processo de análise de dados

Casos de uso

O papel do cientista de dados

Desafio atual:

Encontrar **profissionais capacitados**

Cientista de dados

Para os cientistas de dados não há desemprego | EXAME
https://exame.abril.com.br/ciencia/para-os-cientistas-de-dados-nao-ha-desemprego/
5 de mar de 2016 - Mesmo em época de desemprego em alta, é grande a **demanda** por **cientistas de dados** — profissionais que extraem informações valiosas do ...

Brasil sofre com a falta de profissionais para o big data
Ana Paula Lobo e Pedro Costa ... 12/05/2015 ... Convergência Digital

Por que Cientistas de Dados Continuam em Alta Demanda? - Data ...
datascienceacademy.com.br/.../por-que-cientistas-de-dados-continuam-em-alta-deman...
9 de mai de 2017 - Além disso, a ferramenta de tendências no trabalho, do site Indeed, que mostra a **demanda** por **Cientistas de Dados** em todo mundo, revela ...

Oportunidades aumentam para quem estuda ciência de dados - Folha
www1.folha.uol.com.br/.../1971998-oportunidades-aumentam-para-quem-estuda-cienci...
17 de jun de 2018 - A **demanda** por extrair conclusões de bancos de dados sempre existiu ... Renato Vicente, **cientista de dados** da Serasa Experian; segundo ele, ...

PREVISÃO DA IBM: DEMANDA POR CIENTISTAS DE DADOS AUMENTARÁ 28% ATÉ 2020

PROFISSÃO DO ANO DE 2016: **CIENTISTA DE DADOS**

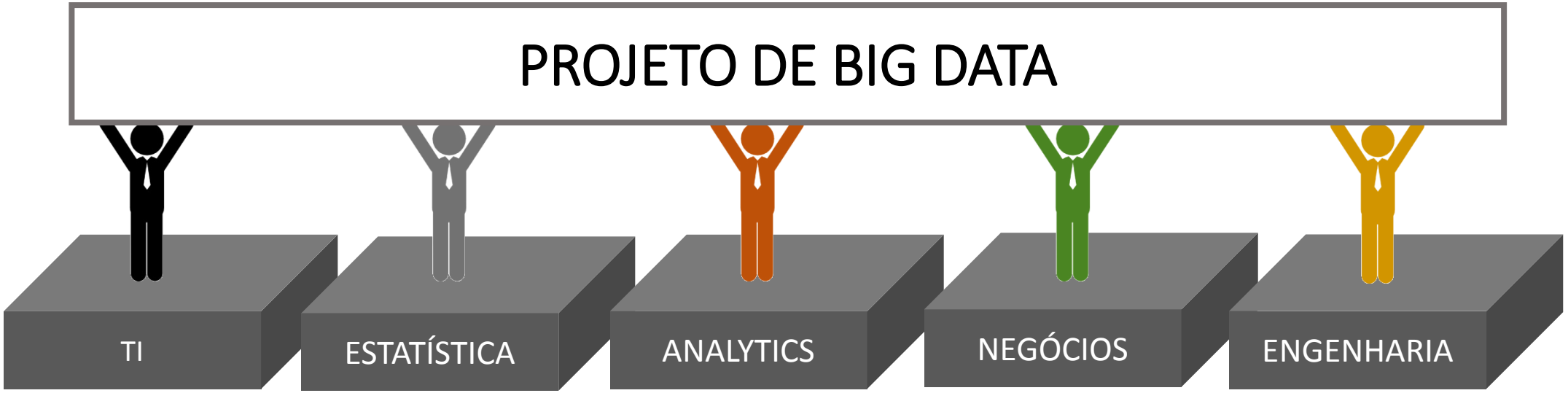
salário médio anual: US$ 128.240

Eleita a profissão do ano de 2016 pelo site Americano de empregos CareerCast.com, considerando critérios como: ambiente de trabalho, renda, nível de stress e perspectiva de contratação

Profissão nº 2: estatístico
salário médio anual: US$ 79.990

Fonte: http://g1.globo.com/economia/concursos-e-emprego/noticia/2016/05/veja-lista-com-10-melhores-e-piores-profissoes-para-2016.html

Um projeto de Big Data deve ser formado por uma **equipe multidisciplinar** altamente qualificada



"Data Science is a team sport"

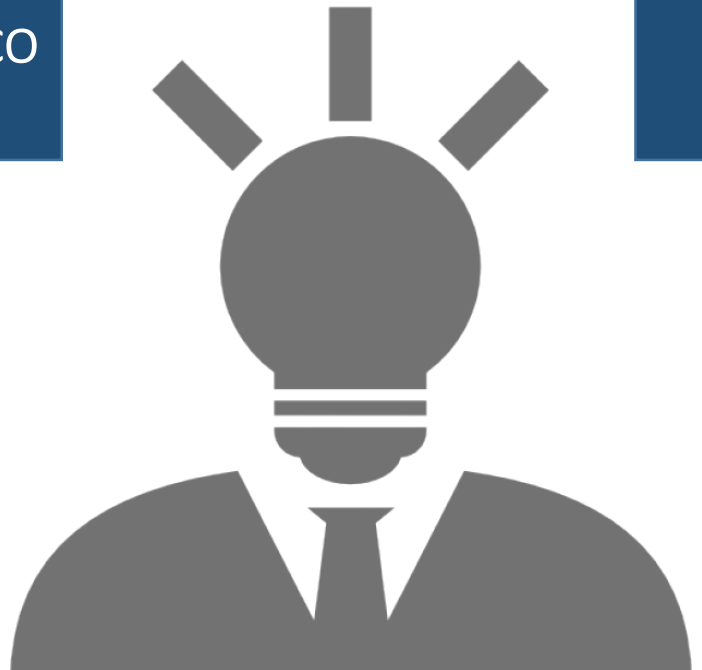
Barack Obama

Perfil do profissional

CURIOSIDADE

PENSAMENTO ANALÍTICO

CRIATIVIDADE

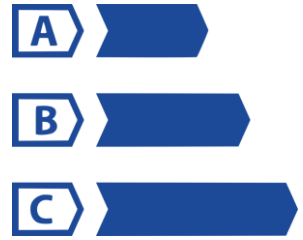


Sugestões para se tornar um cientista de dados

HABILIDADE COM R OU PYTHON

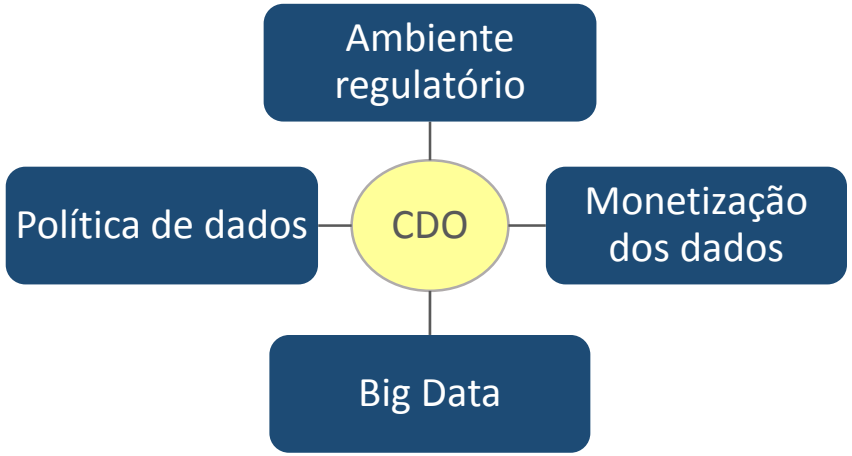
HABILIDADE COM TECNOLOGIAS DE BIG DATA

CONHECIMENTO EM ESTATÍSTICA



Chief Data Officer - CDO

- Responsável pelo ciclo de vida dos dados da empresa, pelas pessoas envolvidas com os dados e pela parte orçamentária.
- Embora esse papel ainda esteja evoluindo, ele está se tornando cada vez mais necessário nas empresas.
- A **governança de dados** é uma função crítica desse profissional.




Doug Cutting
@cutting
Seguindo
As I talk with companies about digital transformation, by far the biggest challenges they face are cultural, not technical.
Traduzir do inglês
12:42 - 20 de set de 2017

Chief Data Officer - CDO

Dicas do DJ Patil, ex-CDO da Casa Branca, para uma carreira em Big Data:

- Seja sempre curioso
- Atenha-se sempre à ética e à segurança
- Faça parte de uma equipe
- Resolva um problema local

Considerações finais

O que se falava antes...

"Big Data pode oferecer vantagem competitiva para as empresas"

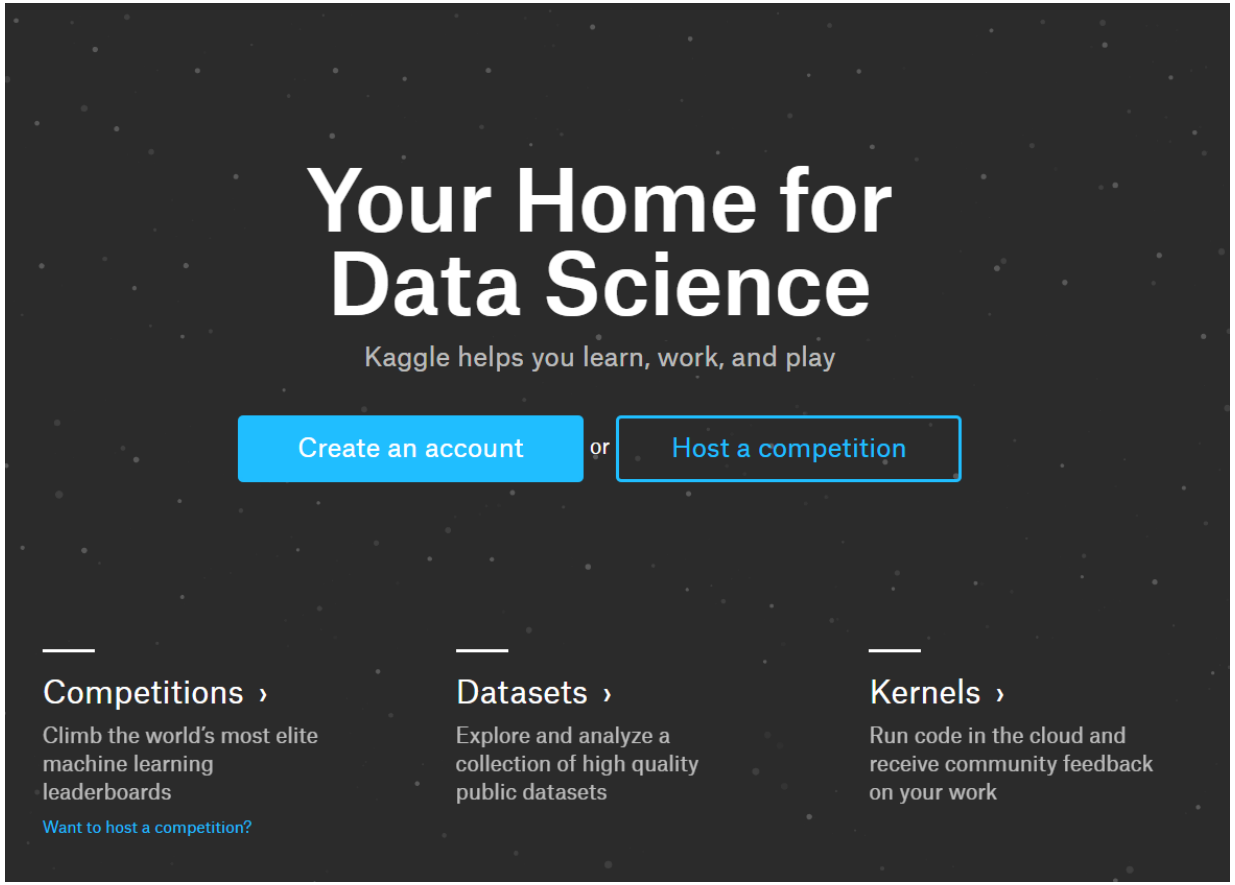
O que se fala agora...

“Big Data é necessário para que as empresas se **mantenham competitivas**”

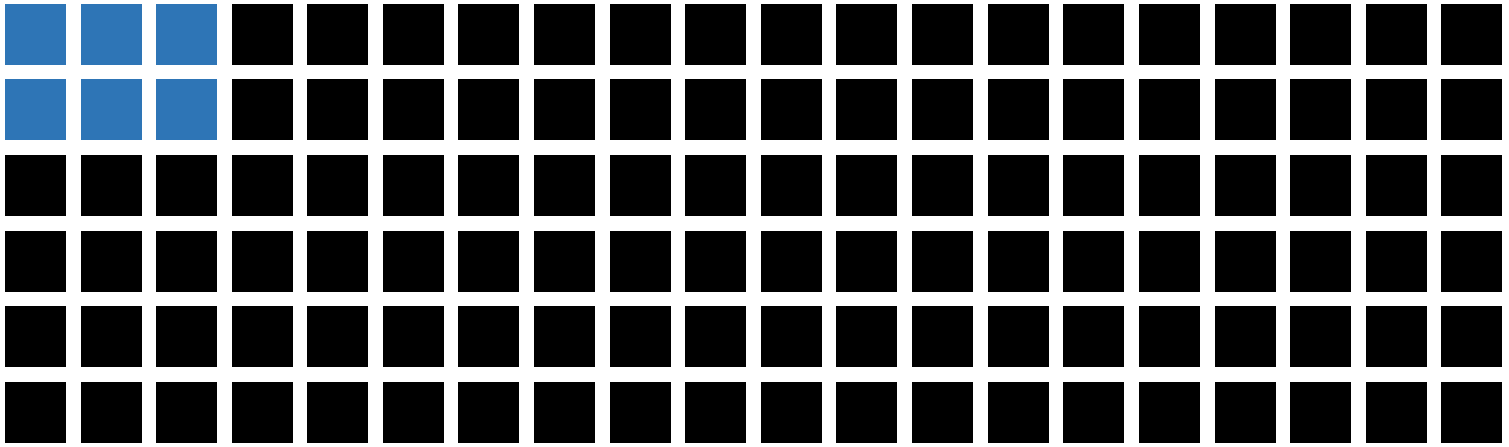
Sugestão de site

kaggle

https://www.kaggle.com



Big Data está apenas em seu início



Estima-se que somente 0.5%

dos dados globais são analisados

Fonte: http://www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume

Um **BIG** Obrigada!

Referências

MARQUESONE, R. **Big Data - Técnicas e Tecnologias para Extração de Valor dos Dados**. Casa do Código, São Paulo, 2016.

GOLDMAN, Alfredo et al. **Apache Hadoop: conceitos teóricos e práticos, evolução e novas possibilidades**. XXXI Jornadas de atualizaçoes em informatica, p. 88-136, 2012.

PROVOST, Foster; FAWCETT, Tom. **Data Science for Business: What you need to know about data mining and data-analytic thinking**. " O'Reilly Media, Inc.", 2013.

WHITE, Tom. **Hadoop: The definitive guide**. " O'Reilly Media, Inc.", 4ed, 2015.

SOARES, Sunil. **The chief data officer handbook for data governance**. Mc Press, 2015.

ABEDIN, Jaynal. **Data Manipulation with R**. Apress, 2018.

DE MAURO, Andrea et al. **Beyond Data Scientists**: a Review of Big Data Skills and Job Families. Proceedings of IFKAD 2016 Towards a New Architecture of Knowledge: Big Data, Culture and Creativity, p. 1844-1857, 2016.

FISCHETTI, Tony. **Data Analysis with R**. Packt Publishing Ltd, 2015.

GANDOMI, Amir; HAIDER, Murtaza. **Beyond the hype: Big data concepts, methods, and analytics**. International Journal of Information Management, v. 35, n. 2, p. 137-144, 2015.

SHEIKH, Nauman. **Implementing Analytics: A Blueprint for Design, Development, and Adoption**. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Sobre os autores

Rosangela de Fátima Pereira Marquesone Pesquisadora no Laboratório de Arquitetura e Redes de Computadores (LARC-USP), atuando nas áreas de computação em nuvem e Big Data. Atua como professora e palestrante de cursos de Big Data para empresas e programas de MBA, tendo ministrado mais de 500 horas de aula sobre o tema. Também atua como revisora de código no Nanodegree Analista de Dados da rede de cursos on-line Udacity. É Mestre e doutoranda em Engenharia de Computação pela Escola Politécnica da Universidade de São Paulo (EPUSP). É Bacharel em Administração de Empresas pela Universidade Estadual do Norte do Paraná (UENP) (2007), Tecnóloga em Análise e Desenvolvimento de Sistemas pela Universidade Tecnológica Federal do Paraná (UTFPR) (2011) e Especialista em Tecnologia Java pela UTFPR (2010). É autora do livro "Big Data – Técnicas e tecnologias para extração de valor dos dados", publicado pela editora Casa do Código. Seus principais interesses de pesquisa são: Big Data, computação em nuvem e people analytics. Também se interessa por temas como design thinking, mulheres na tecnologia e empreendedorismo social.

Sobre os autores

Francisco Pereira Junior possui graduação em Tecnologia em Processamento de Dados pelo Centro de Estudos Superiores de Londrina (1998) e mestrado em Ciência da Computação pela Universidade Estadual de Maringá (2006). É professor efetivo do Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná (UTFPR) atuando em cursos de graduação e pós-graduação (Engenharia de Computação, Engenharia de Software, Análise e Desenvolvimento de Sistemas, Tecnologia Java e Informática Aplicada à Educação). Também é representante Institucional da Universidade Tecnológica Federal do Paraná Câmpus Cornélio Procópio (UTFPR/CP) junto a Sociedade Brasileira de Computação (SBC). Participa de projetos e tem interesse em pesquisas com ênfase em Processamento de Alto Desempenho (Cluster / Grid / Nuvem / Programação Paralela - MPI), Big Data, Hadoop e todo seu ecossistema.

Sobre os autores

Tereza Cristina Melo de Brito Carvalho Graduada em 1980 em Engenharia Eletrônica, em 1988 como mestre e em 1996 como doutora na área de redes de computadores pela Escola Politécnica da USP (Poli). Concluiu o Sloan Fellows Program em 2002 pelo MIT – Massachusetts Institute of Technology, Boston – EUA. Já trabalhou na Siemens, Nuremberg – Alemanha, e na France Telecom, Every – França. Recebeu diversos prêmios, como: Prêmio InfoExame Inovação em Iniciativa Verde (2010), Prêmio e Menção Honrosa do Prêmio Governador Mário Covas em Inovação (2009 e 2008) do Governo de Estado de São Paulo pelo projeto do CEDIR (Centro de Descarte e Reuso de Resíduos de Informática) e de Criação do Selo Verde da USP, Personalidade em Tecnologia pela InfoExame (2005 e 2007) e Executiva em TI pela ABACO (2006). Foi diretora do CCE (Centro de Computação Eletrônica) da USP de 2006-2010. Atualmente, é Assessora para Projetos Especiais da CTI (Coordenadoria de Tecnologia de Informação) da USP, coordenadora do CEDIR, coordenadora geral do LASSU (Laboratório de Sustentabilidade em TIC) e membro do conselho diretor do LARC (Laboratório de Arquitetura e Redes de Computadores), ambos laboratórios de pesquisa do PCS (Departamento de Engenharia de Computação e Sistemas Digitais) da Escola Politécnica da USP. É professora assistente do PCS/Poli. Atua em projetos de pesquisa e desenvolvimento nas áreas de Sistemas de Informação, redes de comunicação, gerenciamento, segurança, Governança e Sustentabilidade em TIC. Possui mais de 100 artigos científicos e tecnológicos publicados em revistas indexadas, conferências internacionais e nacionais.