
ecossistema
ânima

TRANS_
FORMAR
O PAÍS PELA
EDUCAÇÃO
É O QUE
NOS MOVE

ecossistema
ănima

Jorge Werner

Big Data e Análise de Dados

ă 2024.01

Bem vindos !!!

MINERAÇÃO DE DADOS

Limpeza de Dados

O que vimos até agora?

➔Coleta de Dados

MINERAÇÃO DE DADOS

PARA! PARA!

PARA! PARA! PARA!

PARA! PARA!

ă O que é Mineração de Dados?

Mineração de dados, também conhecida como descoberta de conhecimento em bancos de dados (KDD, do inglês Knowledge Discovery in Databases), é um processo de extração de padrões interessantes e úteis ou conhecimento oculto de grandes conjuntos de dados. O objetivo da mineração de dados é descobrir informações valiosas e insights que podem ajudar na tomada de decisões e na solução de

ă Qual o Objetivo da Mineração de Dados?

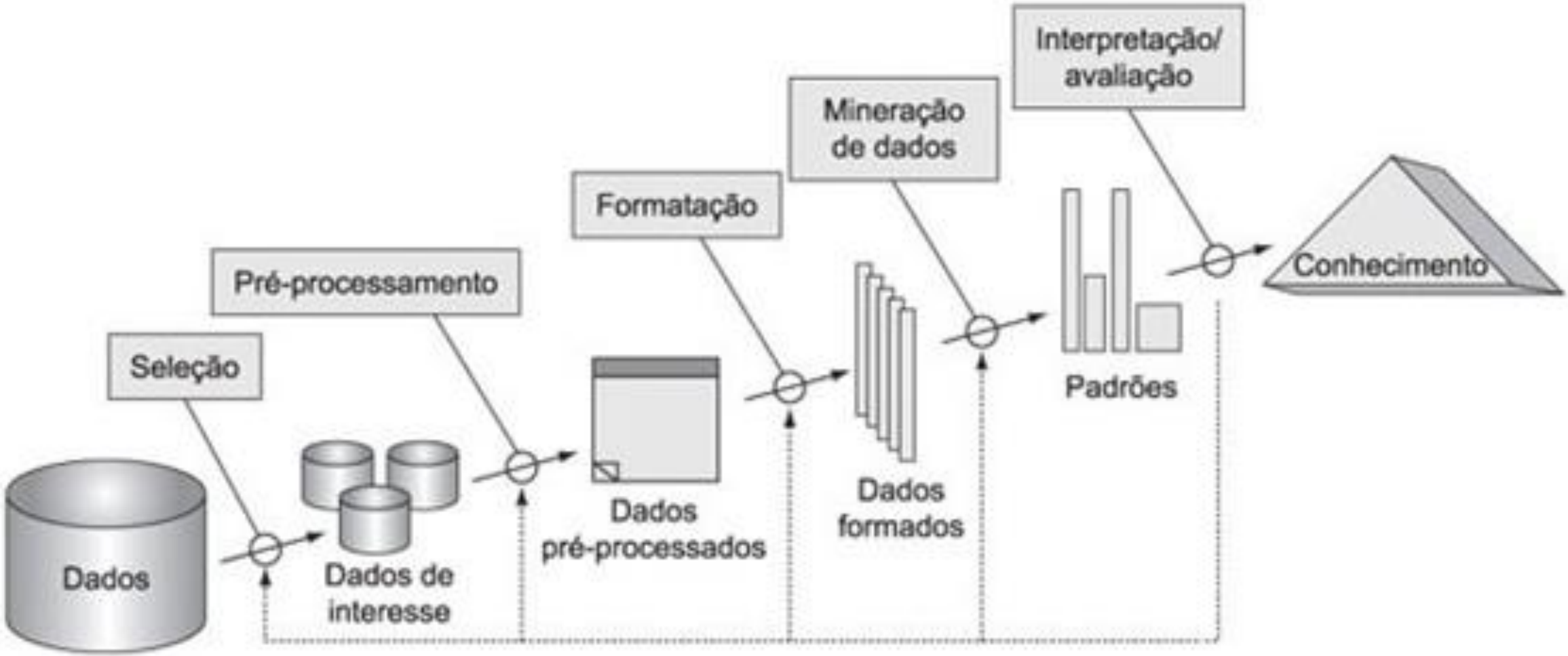
O objetivo da mineração de dados é descobrir informações valiosas e insights que podem ajudar na tomada de decisões e na solução de problemas em diversas áreas, como negócios, ciência, saúde, entre outras.

ă Processo de Mineração de Dados

O processo de mineração de dados geralmente envolve várias etapas:

- **Seleção dos dados**: Escolha dos conjuntos de dados relevantes para a análise, levando em consideração os objetivos específicos do projeto.

- **Pré-processamento dos dados**: **Limpeza e preparação dos dados para análise**, o que pode incluir a remoção de valores ausentes, correção de erros, normalização de dados e outras técnicas de preparação.
- …


Interpretação/
avaliação
Mineração
de dados
Formatação
Pré-processamento
Seleção
Conhecimento
Padrões
Dados
formados
Dados
pré-processados
Dados de
interesse
Dados

O que é Limpeza de Dados?

ă O que é Limpeza de Dados?

ă O que é Limpeza de Dados?

A limpeza de dados, refere-se ao **processo de identificar e corrigir erros, inconsistências e valores ausentes nos conjuntos de dados**.

É crucial porque os conjuntos de dados em big data podem ser vastos e complexos, muitas vezes compostos por dados provenientes de diversas fontes e em diferentes formatos.

ă Limpeza de dados

A limpeza de dados geralmente envolve várias etapas, incluindo:

Detecção de erros: Identificação de dados incorretos, como valores discrepantes, duplicados ou inconsistentes.

Exemplo: Suponha que você tenha um conjunto de dados que contém informações de idade de clientes. Durante a entrada de dados, alguns valores foram digitados incorretamente, como "150" anos de idade. Você pode identificar e corrigir esses erros, considerando que a idade de um cliente não pode ser superior a um limite razoável, como 100 anos.

ă Limpeza de dados

A limpeza de dados geralmente envolve várias etapas, incluindo:

Tratamento de valores ausentes: Identificação e preenchimento de lacunas nos dados. Isso pode envolver a imputação de valores ausentes com base em métodos estatísticos, interpolação, exclusão de registros com valores ausentes ou outras técnicas.

Exemplo: Em um conjunto de dados de vendas, pode haver registros sem informação sobre o preço de um produto. Para tratar esses valores ausentes, você pode decidir imputar o preço médio dos produtos similares na mesma categoria.

ă Limpeza de dados

A limpeza de dados geralmente envolve várias etapas, incluindo:

Padronização e normalização: Garantir que os dados estejam em um formato consistente e uniforme, o que pode incluir a conversão de unidades, normalização de datas e formatos de texto, entre outros.

Exemplo: Suponha que você tenha um conjunto de dados de registros de produtos em um e-commerce. Nesse conjunto de dados, a unidade de medida do peso dos produtos é inconsistente. Alguns produtos têm o peso em gramas (g), enquanto outros estão em quilogramas (kg) e alguns até mesmo em libras (lb). Isso dificulta a comparação direta entre os pesos dos produtos.

ă Limpeza de dados

A limpeza de dados geralmente envolve várias etapas, incluindo:

Correção de erros de formato: Correção de erros de formatação, como erros de digitação, erros de codificação ou outros problemas que possam afetar a interpretação dos dados.

Exemplo: Digamos que você tenha um conjunto de dados de endereços de clientes, onde a coluna "Código Postal" está misturando diferentes formatos de códigos postais. Alguns estão no formato "XXXXX" (como nos EUA), enquanto outros estão no formato "XXXX-XXX" (como no Brasil). Isso pode causar problemas na interpretação dos dados e dificultar a análise geográfica.

ă Limpeza de dados

A limpeza de dados geralmente envolve várias etapas, incluindo:

Deduplicação: Identificação e remoção de entradas duplicadas nos conjuntos de dados.

Exemplo: Você tem um banco de dados de clientes onde o mesmo cliente pode ter sido cadastrado várias vezes devido a erros no processo de entrada de dados. Você pode identificar e remover entradas duplicadas, considerando critérios como nome, endereço e número de telefone.

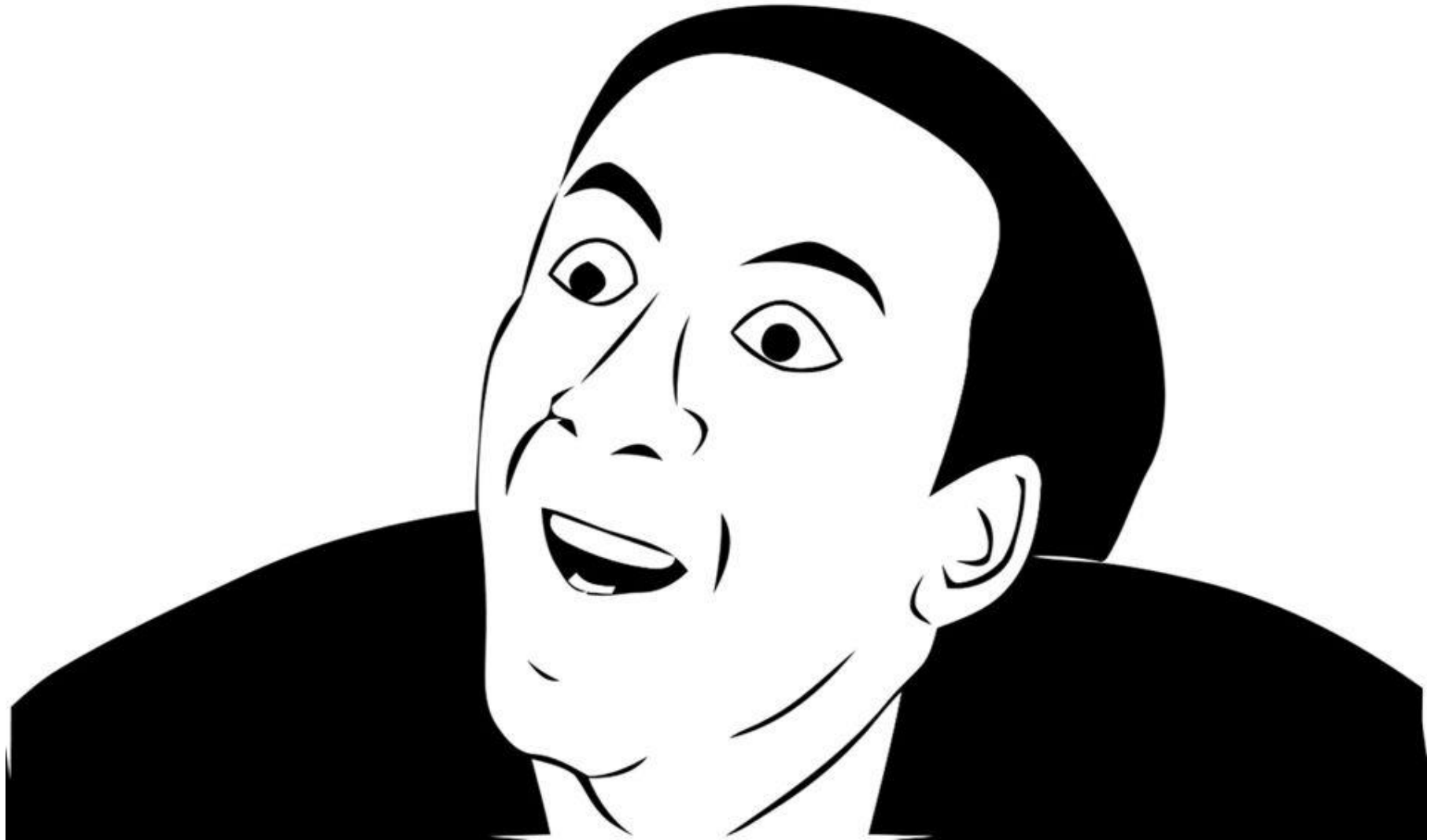
ă Limpeza de dados

A limpeza de dados geralmente envolve várias etapas, incluindo:

Verificação de integridade referencial: Garantir que as relações entre os diferentes conjuntos de dados (se aplicável) estejam corretas e que não haja referências a dados inexistentes.

Exemplo: Em um banco de dados relacional que contém tabelas de clientes e pedidos, você pode verificar se todos os pedidos estão associados a clientes existentes na tabela de clientes. Se houver pedidos sem um cliente associado, isso indica uma falta de integridade referencial que precisa ser corrigida.

Tá, mas... E daí?



Limpeza de Dados

A limpeza de dados é uma **etapa essencial no processo de análise de dados em big data**, pois dados sujos ou inconsistentes podem levar a conclusões imprecisas ou tendenciosas.

Além disso, em ambientes de big data, onde os conjuntos de dados podem ser enormes e complexos, a limpeza de dados muitas vezes envolve **o uso de técnicas automatizadas** e algoritmos de aprendizado de máquina para lidar com grandes volumes de dados de forma eficiente.

DÚVIDAS

ecossistema
ănima