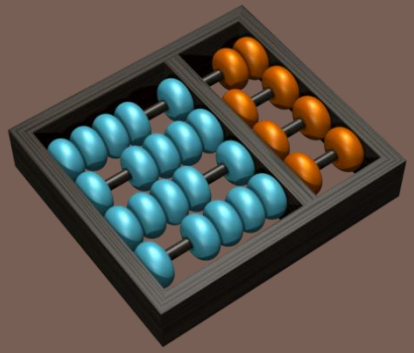


UNICAMP

CONCEITOS BÁSICOS EM MINERAÇÃO DE DADOS

Prof. Julio Cesar dos Reis

jreis@ic.unicamp.br

www.ic.unicamp.br/~jreis

CT-0611

Vídeo

Objetivos da aula

- Apreender as motivações e objetivos da mineração de dados
- Estudar o processo de descoberta de conhecimento

Motivação

Usuários na internet - Brasil

G1

ECONOMIA

TECNOLOGIA

Uso da internet no Brasil cresce, e 70% da população está conectada

Segundo pesquisa TIC Domicílios, 126,9 milhões de pessoas usaram a rede regularmente em 2018. Metade da população rural e das classes D e E agora têm acesso à internet.

Por Thiago Lavado, G1

28/08/2019 11h01 · Atualizado há 8 meses

https://g1.globo.com/economia/tecnologia/noticia/2019/08/28/uso-da-internet-no-brasil-cresce-e-70percent-da-populacao-esta-conectada.ghtml

Usuários na internet - Mundo

Worldwide Internet users

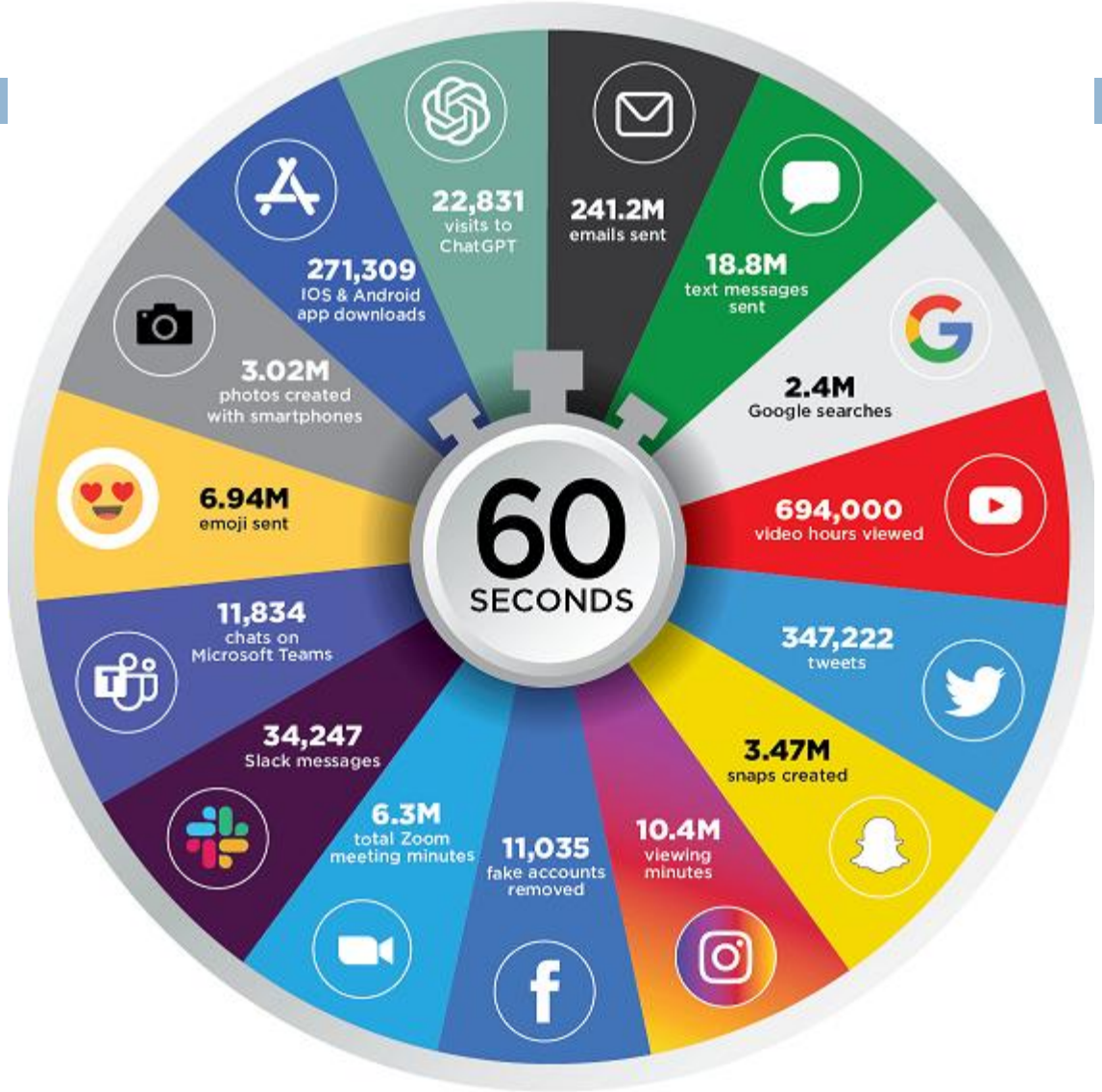
	2005	2010	2017	2019[a]
World population[6]	6.5 billion	6.9 billion	7.4 billion	7.75 billion
Users worldwide	16%	30%	48%	53.6%
Users in the developing world	8%	21%	41.3%	47%
Users in the developed world	51%	67%	81%	86.6%

[a] Estimate.

Source: International Telecommunications Union.[7]

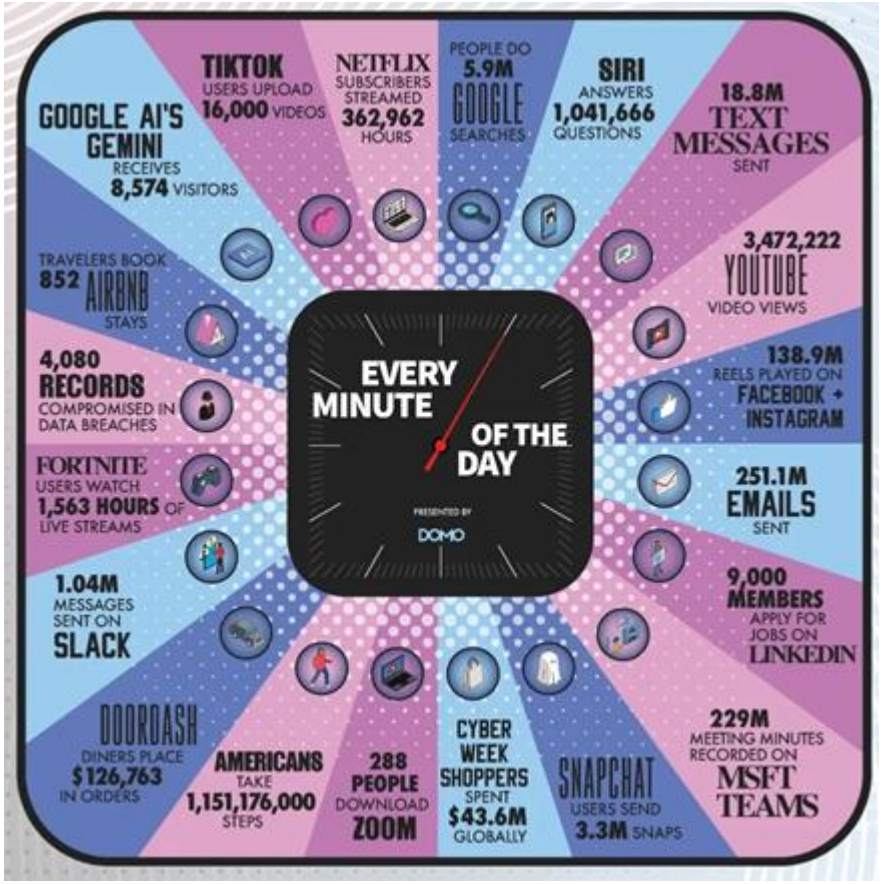
https://en.wikipedia.org/wiki/Global_Internet_usage

THE INTERNET IN 2023 EVERY MINUTE



Created by: eDiscovery Today & LTMG

2024 – 1 minuto na web



https://www.socialmediatoday.com/news/what-happens-online-every-minute-2024/735990/

Dilúvio de dados

- Crescimento explosivo na capacidade de gerar, coletar e armazenar dados

- Máquinas e pessoas continuamente
  - Coletam dados
  - Geram dados
  - Processam dados
  - Transmitem dados

De onde vem os dados?

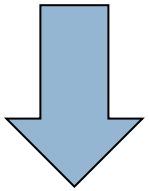
- Científicos: imagens, sinais
- Sociais: censos, pesquisas, redes sociais
- Econômicos e comerciais: transações bancarias e comerciais, compras, ligações telefônicas, acessos a web, transações com código de barras e RFID.

De onde vem os dados?

- Científicos: imagens, sinais
- Sociais: censos, pesquisas, redes sociais
- Econômicos e comerciais: transações bancarias e comerciais, compras, ligações telefônicas, acessos a web, transações com código de barras e RFID.
- Segurança: acessos a sistemas em rede (*logs*), e-mails corporativos, registro de atividades.
- Sensores de dados
  - Climáticos, reservatórios de água, corpo humano
- Imagens e vídeos
  - Câmeras de monitoramento de trânsito, de segurança

Cenário atual

Avanços recentes nas tecnologias para **aquisição, transmissão, armazenamento** e **processamento** de dados



Big Data

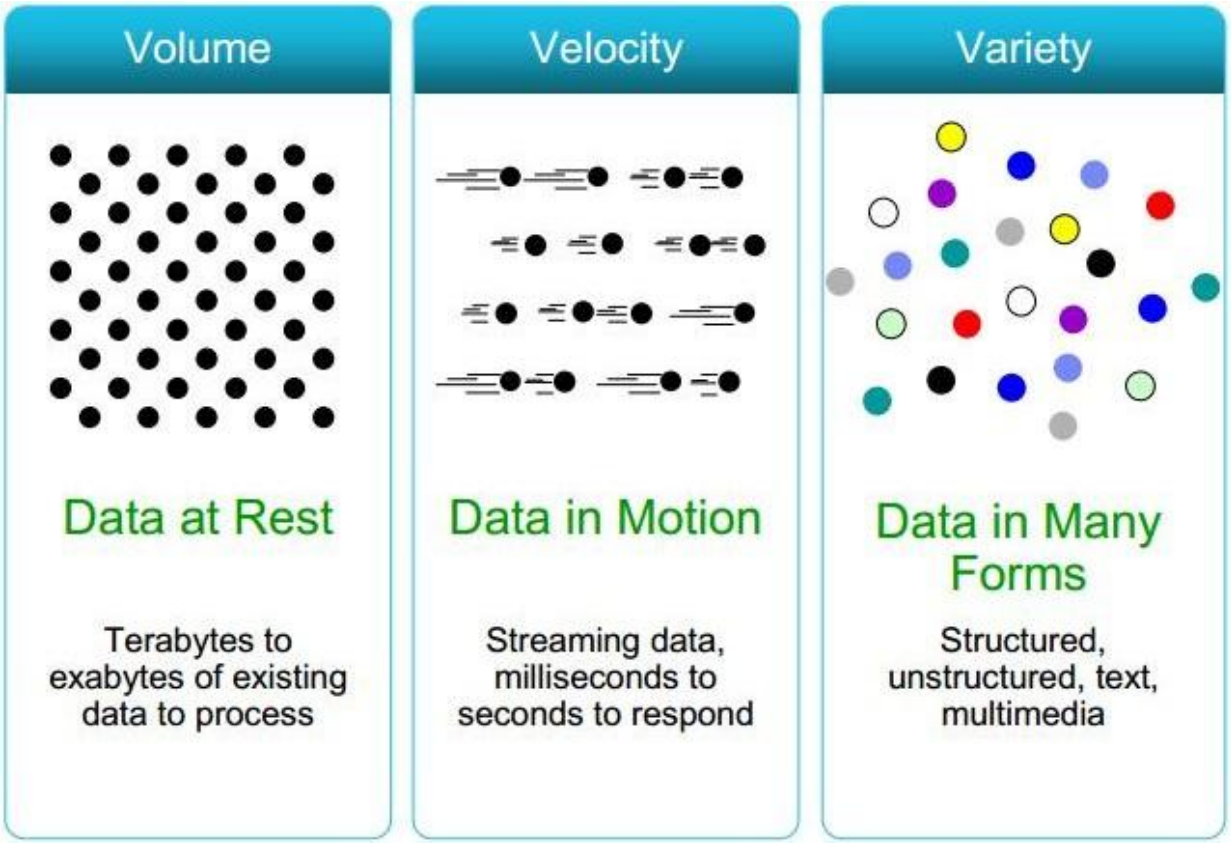
O que é Big Data?

- Várias definições
  - Dados que são grandes demais para sistemas tradicionais de processamento de dados
  - Dados que precisam de novas técnicas para serem processados
  - Dados que são muito complexos
  - Dados que são importantes
  - Coletar dados agora para entendê-los depois

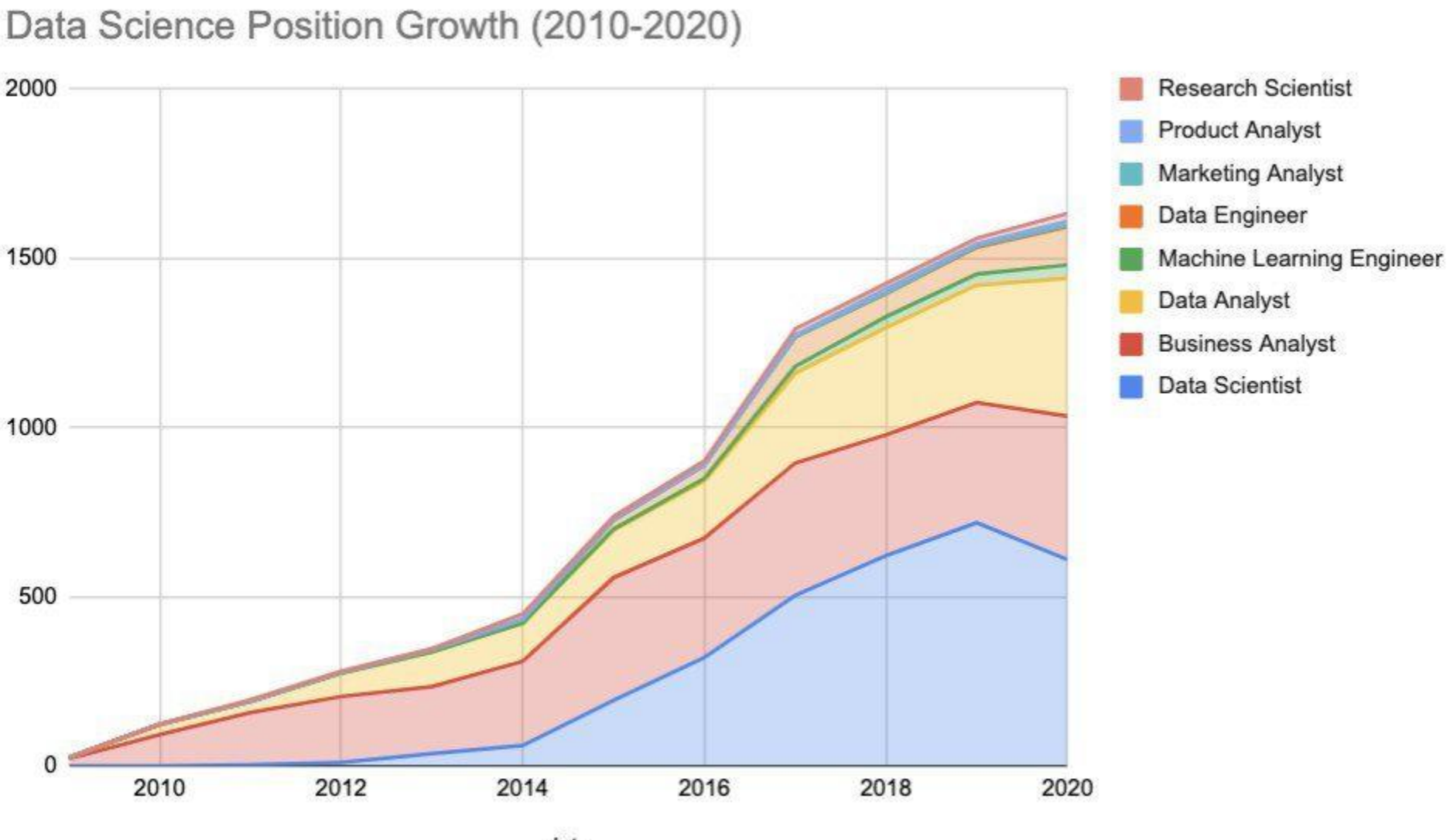
Características de Big Data

- Grande **volume** de dados, gerados a uma grande **velocidade** e com uma grande **variedade** (3 Vs)
  - Volume: tanto de dados estruturados quanto de não estruturados
  - Variedade: vindos de fontes diferentes e que precisam ser integrados
  - Velocidade: gerados em fluxos cada vez mais rápidos

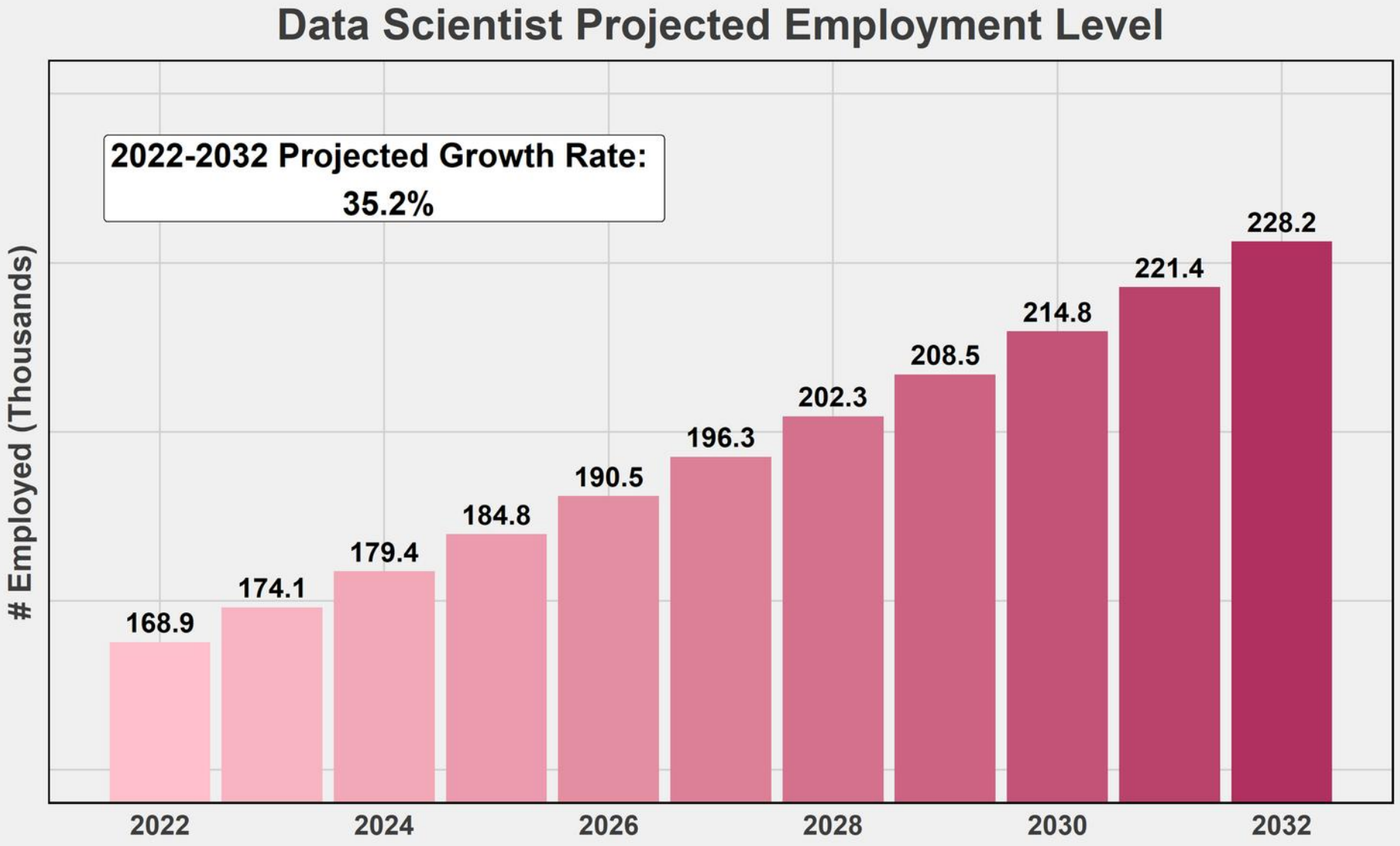
3Vs


Volume
Data at Rest
Terabytes to exabytes of existing data to process
Velocity
Data in Motion
Streaming data, milliseconds to seconds to respond
Variety
Data in Many Forms
Structured, unstructured, text, multimedia

Mercado profissional



Mercado profissional



https://www.visualizecuriosity.com/posts/data-science-job-market-2024

Necessidades

1. Muitos dados disponíveis em diversos domínios de aplicação

2. Necessário técnicas para extrair conhecimento de grande volume de dados

3. Transformação de dados em conhecimento

4. Descoberta de conhecimento em banco de dados

O que é um dado?

- Coleção de objetos e seus atributos

Atributos

Objetos

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

O que é um dado?

- Coleção de objetos e seus atributos
- Um atributo é uma propriedade ou característica de um objeto
  - Exemplos: cor dos olhos, temperatura, etc.
  - Atributo é também conhecido como variável, campo, característica (*characteristic*, ou *feature*)

Atributos

Objetos

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

O que é um dado?

- Coleção de objetos e seus atributos
- Um atributo é uma propriedade ou característica de um objeto
  - Exemplos: cor dos olhos, temperatura, etc.
  - Atributo é também conhecido como variável, campo, característica (*characteristic*, ou *feature*)
- Uma coleção de atributos descreve um objeto
  - Objeto é também conhecido como registro, ponto, caso, amostra, entidade, ou instância

Atributos

Objetos

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

De dados para conhecimento

- O que é feito desses dados?
- O que existe de interessante nesses dados?
  - Como definir “interessante”?
- Como analisar esses dados?
  - Por que não analisá-los para descobrir **novas informações** e utilizá-las de forma **estratégica**?

Dados vs. informação

- Os dados brutos são "inúteis"
- São necessárias técnicas que automaticamente extraiam informação deles
- Informação: padrões nos dados

Informação é essencial

- Exemplo 1: fertilização em vidro
  - Embriões descritos por 60 características
  - Problema: selecionar os embriões que vão sobreviver
  - Dados: registros históricos de embriões
- Exemplo 2: Seleção de gado
  - Gado descrito por 700 características
  - Problema: selecionar o gado
  - Dados: registros históricos com a decisão dos fazendeiros

De dados para conhecimento

- **Informação**, e não dados, valem dinheiro / tempo / conhecimento!

- Aproveitamento da informação permite ganho de competitividade

Conhecimento

O que se pode identificar

- Como identificar?
  - Padrões (“X” acontece se...)
  - Exceções (isto é diferente de... por causa de...)
  - Tendências (ao longo do tempo, “Y” deve acontecer...)
  - Correlações (se “M” acontece, “N” também deve acontecer)

Limites de técnicas tradicionais

- Empresas mantêm bancos de dados com bilhões ou trilhões de registros históricos de suas transações
  - Centenas de atributos precisam ser analisados simultaneamente
- Os recursos de análise de dados tradicionais são inviáveis para acompanhar essa evolução
  - Métodos tradicionais (SQL, Planilhas, investigação manual)

Base de dados vs. mineração de dados

- Consulta
  - Bem definida
  - SQL
- Saída
  - Subconjunto da base

- Consulta
  - Não se define a priori
  - Sem linguagem para consulta
- Saída
  - Não é um subconjunto da base

Exemplos de consultas

- Base de dados
  - Encontre todos os clientes que bebam cerveja do tipo X
  - Encontre todos os clientes com parcelas em atraso

Exemplos de consultas

- Base de dados
  - Encontre todos os clientes que bebam cerveja do tipo X
  - Encontre todos os clientes com parcelas em atraso

- Mineração de dados
  - Encontre todos clientes que podem ter uma parcela em atraso (**classificação**)
  - Agrupe os clientes por hábitos de compra (**agrupamento**)
  - Liste todos os itens que são frequentemente comprados com bicicletas (**regras de associação**)

Aplicações práticas em negócios

- Entender o perfil dos clientes
  - Quais são os clientes típicos da empresa?
- Desenvolvimento de novos produtos
- Controle de estoque em postos de distribuição
- Propaganda mal direcionada gera maiores gastos e desestimula o possível interessado a procurar as ofertas adequadas

Extraindo informações úteis

- Extração
  - Implícita
  - Previamente desconhecida
  - Potencialmente útil
- Necessidades
  - Programas que detectem padrões e regularidades em dados
- Padrões fortes $\Rightarrow$ boas predições
  - Problema 1: a maior parte dos padrões não são interessantes
  - Problema 2: os padrões podem não ser exatos
  - Problema 3: os dados podem estar truncados ou faltantes

Técnicas e ferramentas são necessárias

- Ferramentas de automatização das tarefas repetitivas e sistemática de análise de dados
- Ferramentas de auxílio para as tarefas cognitivas da análise (e.g., visualização de dados)
- Integração de ferramentas em sistemas apoiando o processo completo de descoberta de conhecimento para tomada de decisão

Área de Mineração de dados

Mineração de dados

- Processo realizado através de **metodologias** automatizadas e **algoritmos eficientes** que tem por objetivo a descoberta de conhecimento valioso em grandes bases de dados

- Obter um "*diamante de informação*" a partir de um grande volume de dados

Propriedades de “diamante de informação”

- O conhecimento descoberto através de processos de mineração de dados é considerado **interessante** quando apresenta certas **propriedades**
  - Validade
  - Inesperabilidade
  - Interpretabilidade
  - Novidade
  - Utilidade

Exemplo clássico

- A mineração de um banco de dados de uma grande loja dos EUA revelou:
  - "Grande parte dos consumidores que fazem compras nas **noites de quinta-feira** costumam adquirir os dois produtos: **fraldas** e **cerveja**"

Propriedades da regra encontrada

- Representou uma informação **nova**
  - Não era conhecida pelos analistas da empresa
- Foi uma associação **inesperada**
  - Os analistas imaginavam que as vendas de cerveja estivessem associadas apenas a produtos como salgados, carne para churrasco e outras bebidas alcoólicas, mas nunca a produtos de higiene infantil
- Foi considerada **válida**
  - Possuía expressividade estatística
  - Uma porcentagem considerável das compras realizadas nas noites de quinta-feira continha ambos os produtos

Propriedades da regra encontrada

- É **interpretável**
  - Pôde ser entendida e explicada pelos analistas
  - Sugere que nas noites de quinta-feira, os **casais jovens** se preparam para o fim-de-semana estocando fraldas para os bebês e cerveja para o papai

- **A regra descoberta era útil**
  - Os gerentes da loja de departamentos puderam tomar ações capazes de aumentar as vendas de cerveja
    - E.x.: os produtos foram colocados em prateleiras próximas

Objetivos da mineração de dados

- Extrair conhecimento de grandes volumes de dados
- Formada por um conjunto de ferramentas e técnicas para evidenciar padrões nos dados e auxiliar a descoberta de conhecimento
- Conhecimento descoberto pode ser apresentado por essas ferramentas de diversas formas
  - Agrupamentos, hipóteses, regras, árvores de decisão, ou grafos

Uso de algoritmos

- A mineração de dados baseia-se na utilização de **algoritmos**

- Algoritmos adequados para revelar **padrões** escondidos nos dados

Ensino de mineração de dados

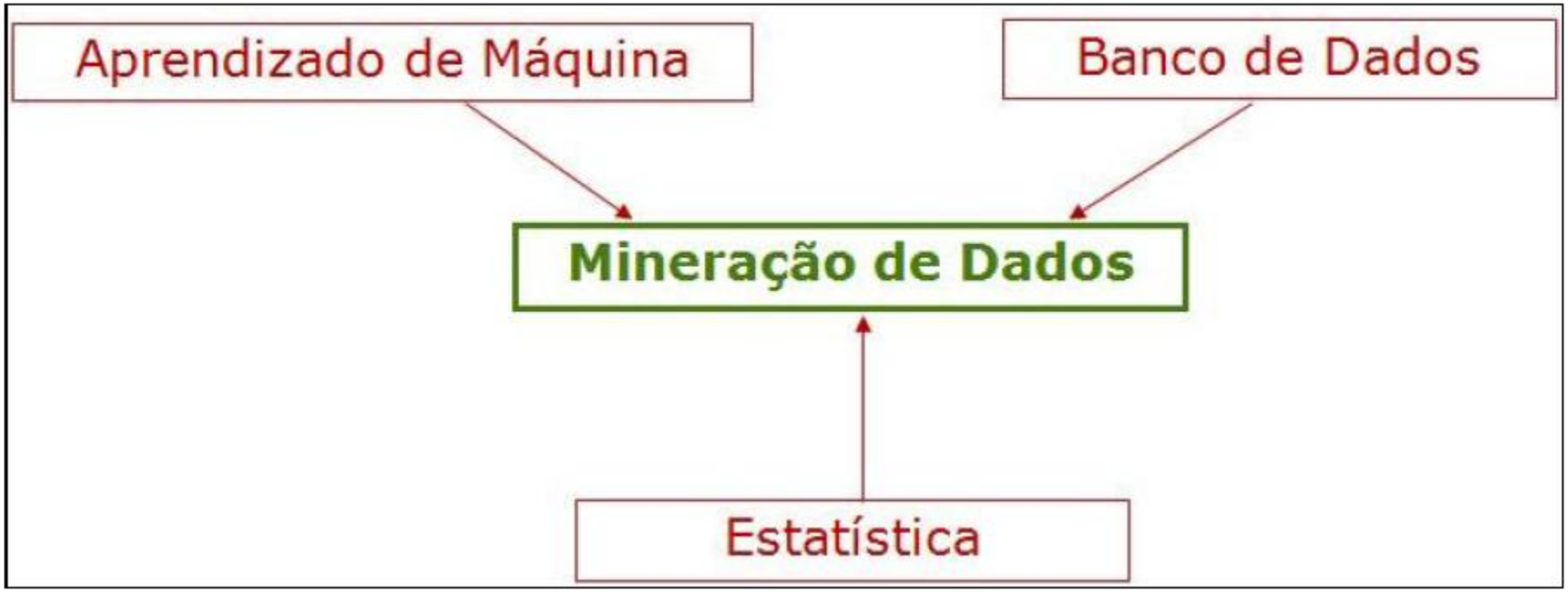
- Neste curso, os fundamentos de mineração de dados serão apresentados, bem como a aplicação dessa tecnologia
- Teremos um enfoque prático e aplicado
- Atividades de mineração serão realizadas com ferramentas visando resolver problemas reais de mineração de dados
- As atividades permitirão a fixação dos conceitos apresentados, assim como uma melhor percepção do potencial dessa desafiadora área

Ensino de mineração de dados

- Falamos sobre *terabytes* e *petabytes*, mas não podemos mostrar exemplos práticos nessa escala
- Falamos sobre dezenas ou centenas de atributos de diversos tipos, mas não é simples demonstrar algoritmos usando-os
- Focaremos em "problemas didáticos"
  - Uso de duas dimensões numéricas, focando mais em características de um algoritmo do que em performance e escalabilidade

Área multidisciplinar

- Adaptou conceitos provenientes de diferentes áreas com o intuito de resolver o problema da descoberta de conhecimento escondido em grandes bases de dados



Processo de descoberta do conhecimento

Descoberta de conhecimento em bancos de dados

- "Processo não trivial de extração de informações implícitas, anteriormente desconhecidas, e potencialmente úteis de uma fonte de dados"
  - Conhecido como *Knowledge Discovery in Databases* - ***KDD***

- **KDD**: Processo geral de descoberta de conhecimentos úteis previamente desconhecidos a partir de grandes bancos de dados

Posicionamento do KDD

KDD vs. mineração de dados

- Mineração de dados é o passo do processo de KDD que produz um conjunto de padrões sob um custo computacionalmente aceitável

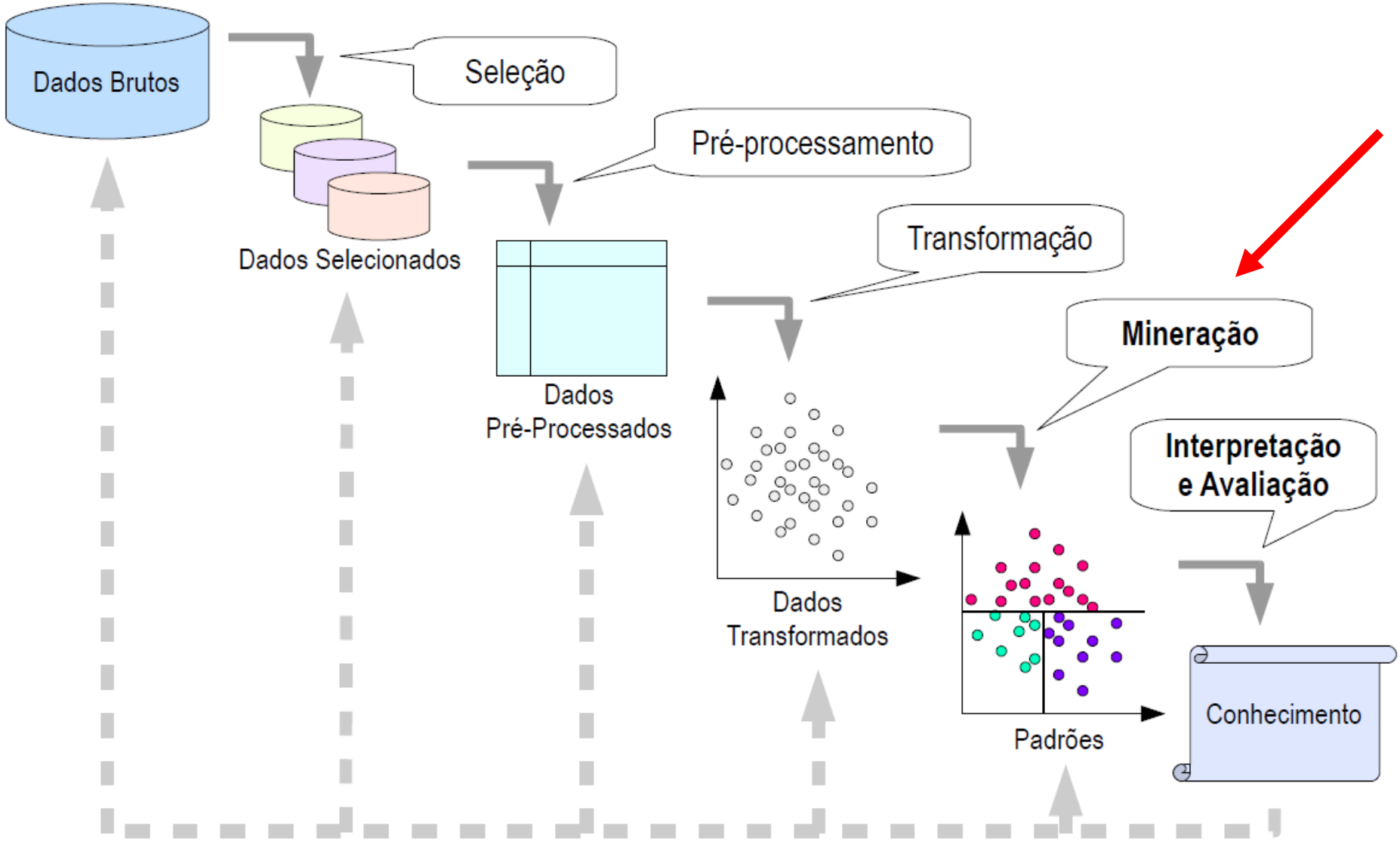
- *KDD* utiliza algoritmos de mineração para extrair padrões classificados como "conhecimento"
  - Padrão interessante? (válido, novo, útil e interpretável)

- Incorpora igualmente tarefas como escolha do algoritmo adequado, processamento e amostragem de dados e interpretação de resultados

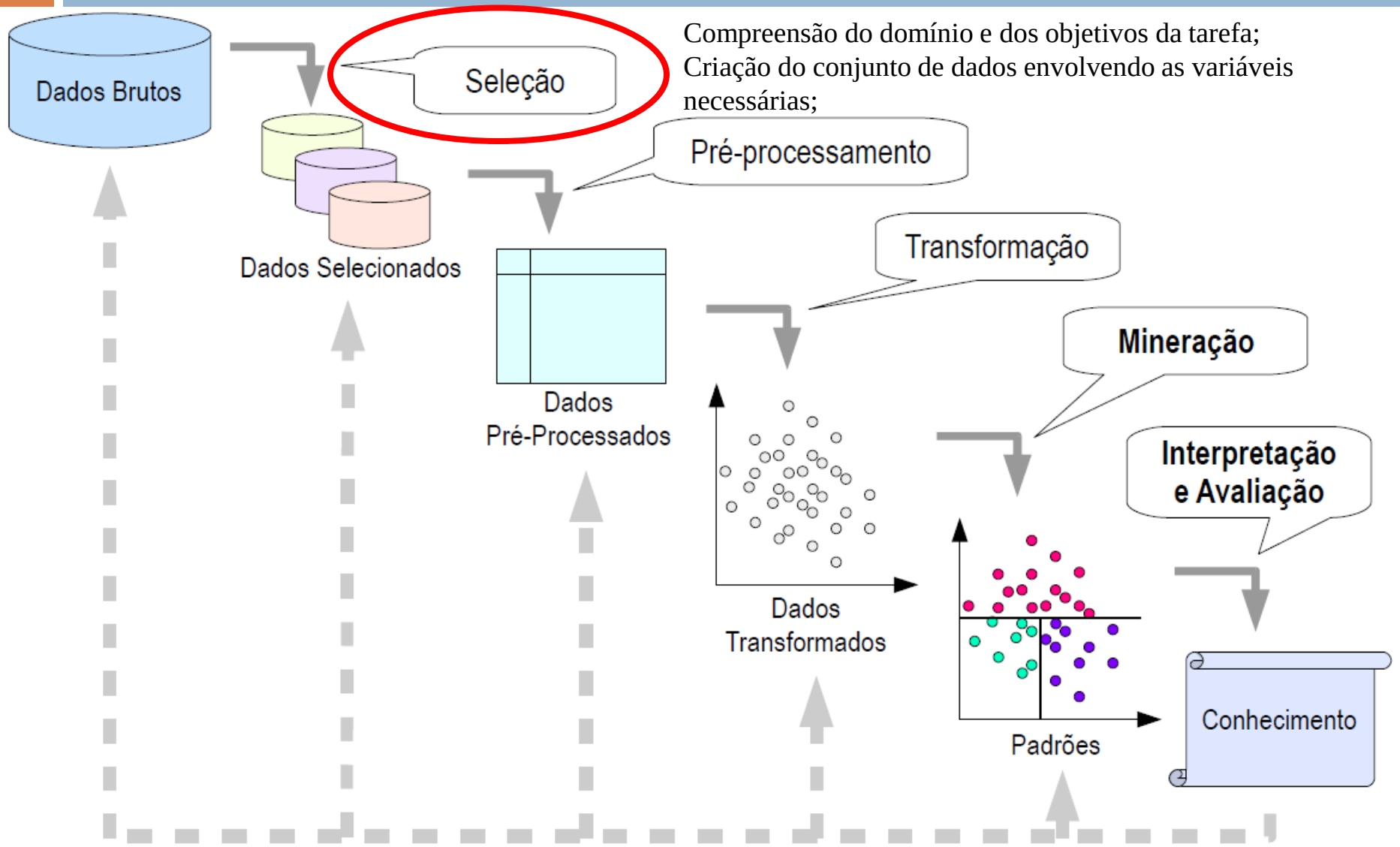
Passos no KDD

1. Compreensão do domínio da aplicação
2. Criação de conjunto de dados para descoberta
3. Limpeza e pré-processamento dos dados
4. Redução e projeção (atributos)
5. Escolha da tarefa de mineração de dados
6. Escolha dos algoritmos de mineração e seus parâmetros
7. **Mineração de dados**
8. Interpretação de resultados
9. Consolidação e avaliação

Processo de descoberta de conhecimento



Processo de descoberta de conhecimento

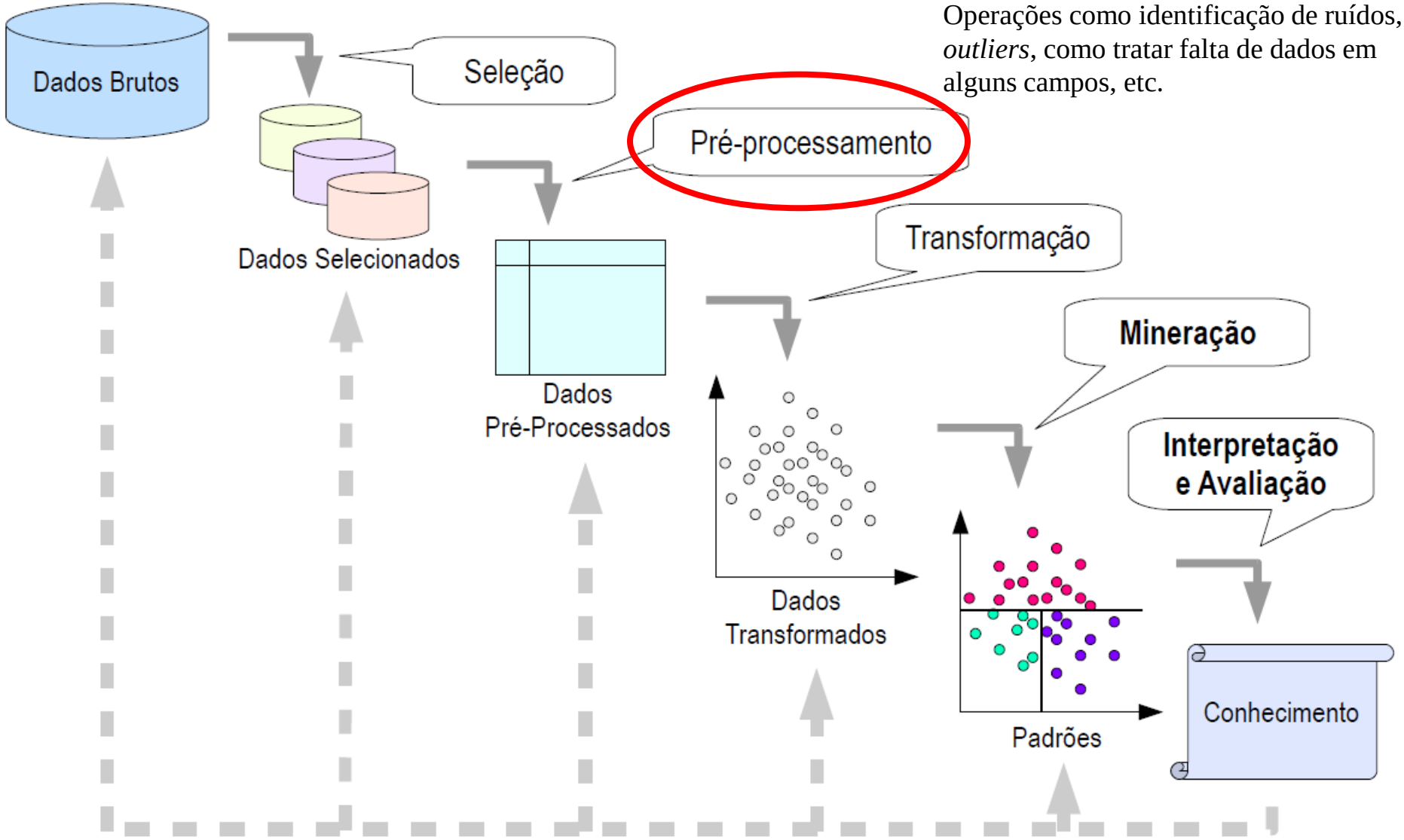


Compreensão do domínio e dos objetivos da tarefa;
Criação do conjunto de dados envolvendo as variáveis necessárias;

Seleção de dados

- Selecionar ou segmentar dados de acordo com critérios definidos
  - Ex.: Todas as pessoas que são proprietárias de carros é um subconjunto de dados determinado

Processo de descoberta de conhecimento



Operações como identificação de ruídos, *outliers*, como tratar falta de dados em alguns campos, etc.

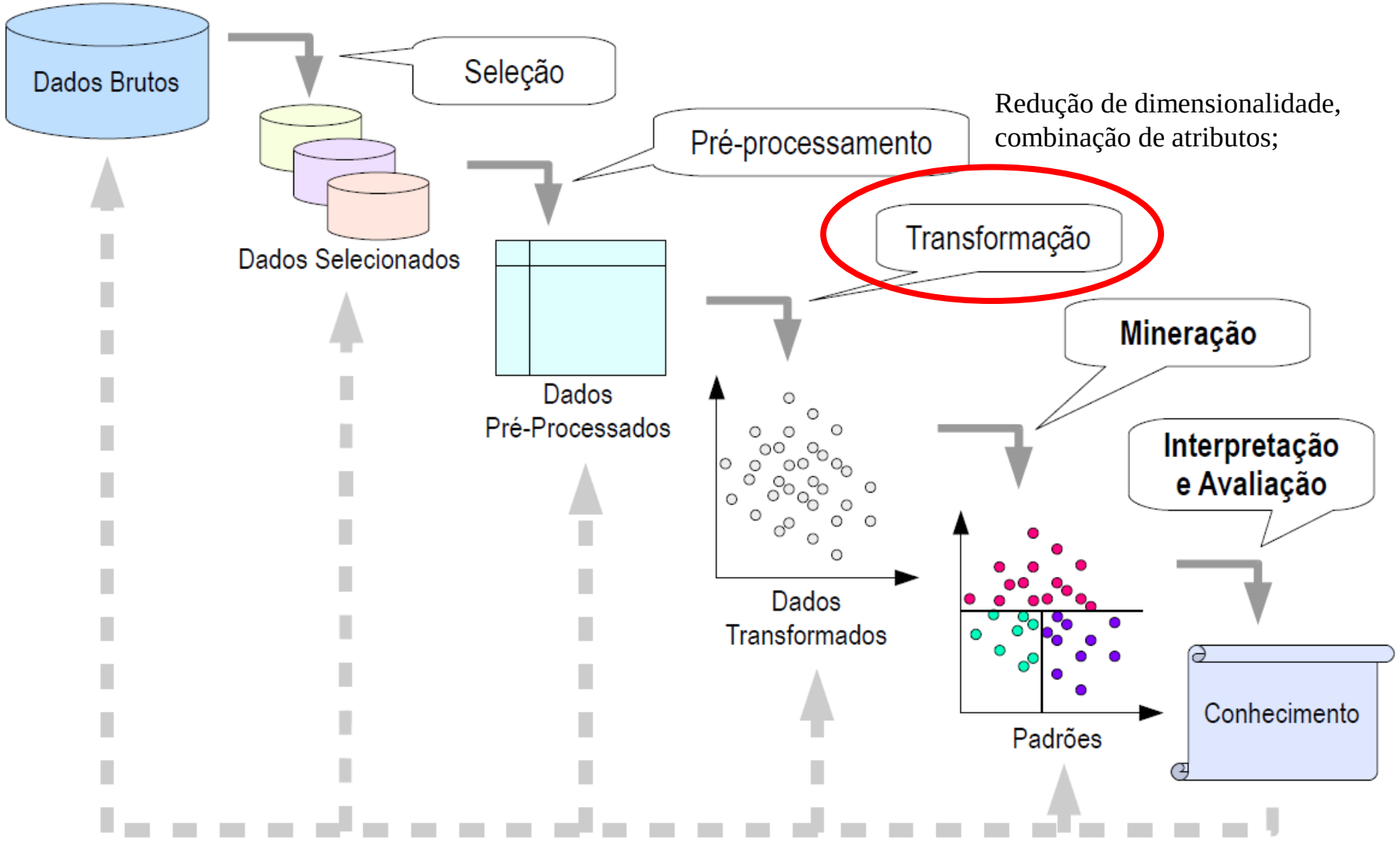
Pré-processamento de dados

- Estágio de limpeza dos dados
- Informações julgadas desnecessárias são removidas
- Reconfiguração dos dados para assegurar formatos consistentes (identificação)
  - Ex. sexo = “F” ou “M”
  - sexo = “M” ou “H”

Pré-processamento de dados

- Rotinas de limpeza de dados visam
  - Suprir valores ausentes
  - Reduzir discrepâncias de valores ruidosos
  - Corrigir inconsistências

Processo de descoberta de conhecimento



Transformação

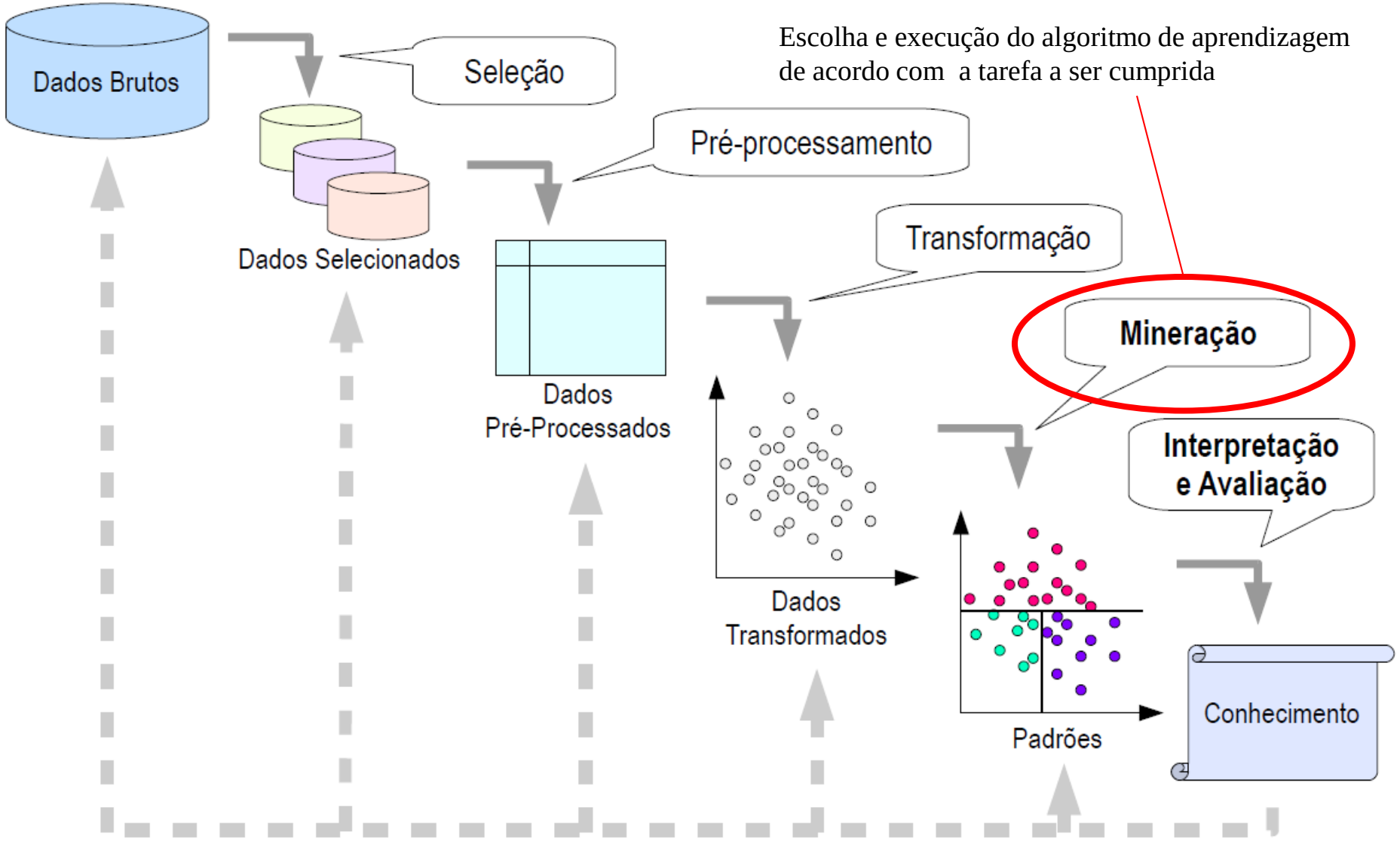
- Transformam-se os dados em formatos utilizáveis
- Esta etapa depende da técnica mineração de dados adotada
- Disponibilizar os dados de maneira usável e navegável

Considerações sobre pré-processamento e transformação

- São fases aplicadas para aumentar a qualidade e o poder de expressão dos dados a serem minerados

- Essas fases tendem a consumir a maior parte do tempo dedicado ao processo de KDD (aproximadamente 70%)

Processo de descoberta de conhecimento

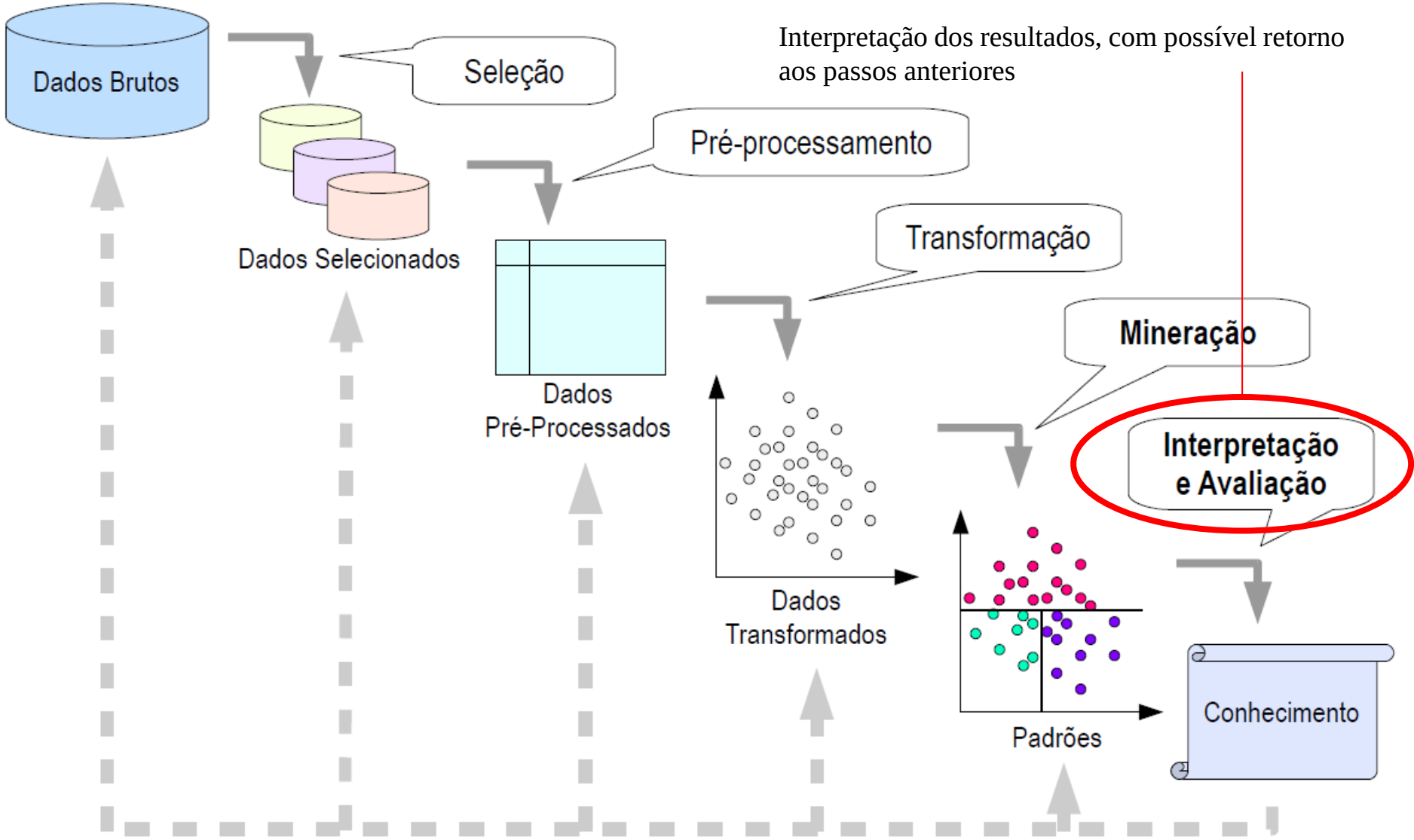


Etapa de mineração de dados

- É a verdadeira extração dos padrões de comportamento dos dados

- Tipos de tarefa
  - Classificação, predição numérica, agrupamento, associação

Processo de descoberta de conhecimento

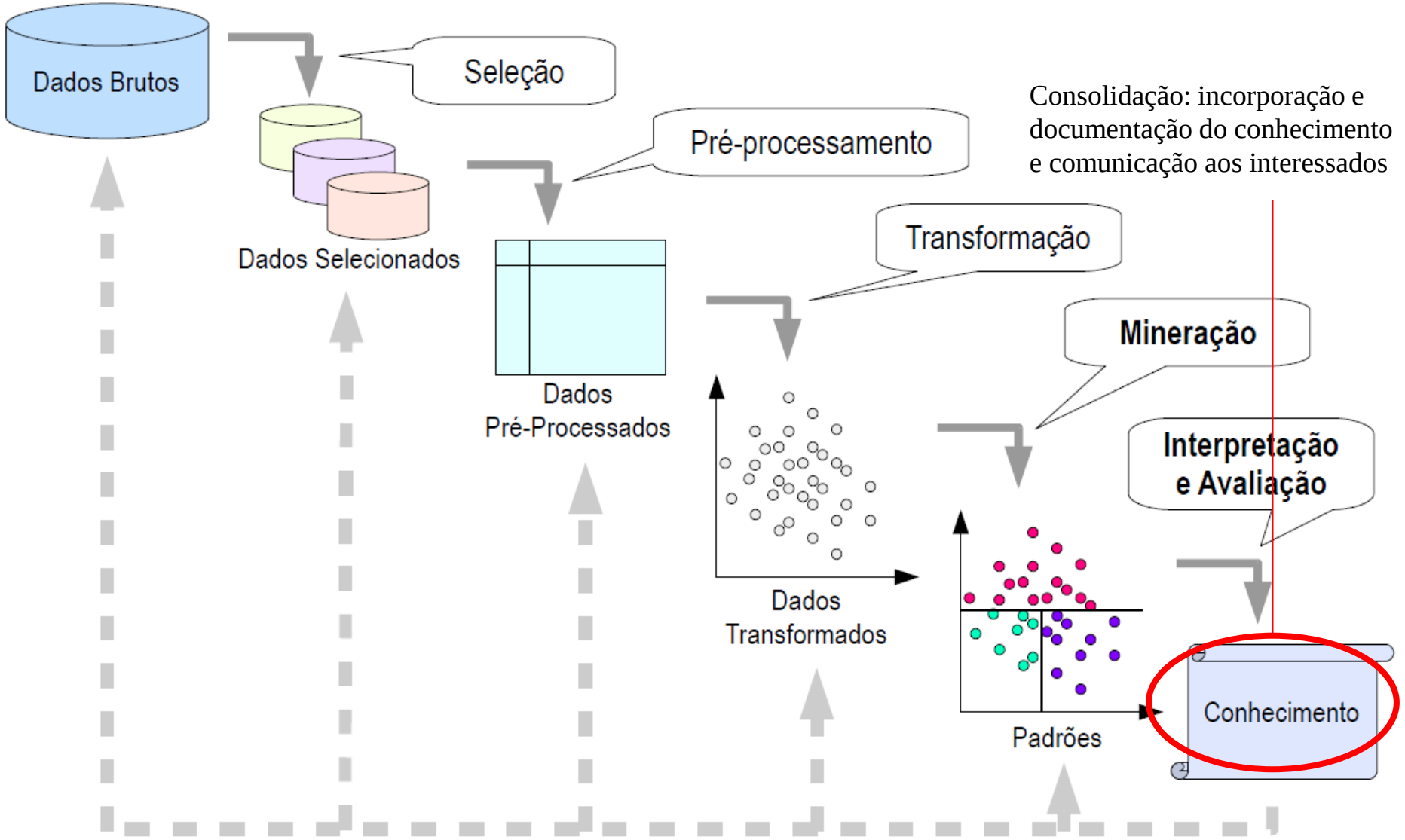


Interpretação e avaliação

- Identificado os padrões pelo sistema, esses são interpretados em conhecimentos

- Padrões darão suporte a tomada de decisões humanas

Processo de descoberta de conhecimento



Considerações sobre o processo

- O processo de *KDD* é interativo, iterativo, cognitivo e exploratório, envolvendo vários passos

- Muitas decisões são feitas pelos analistas do domínio
  - Especialistas do domínio dos dados

Ferramentas

- **PYTHON**
  - Coleção de algoritmos para mineração de dados
  - Usaremos nesse curso

- **R**

Síntese da Aula

- Mineração de dados usam algoritmos para aprender padrões sobre dados
- Etapa chave do processo de descoberta de conhecimento em bases de dados
- Essencial garantir a qualidade dos dados a serem minerados (etapas de pré-processamento e transformação)

Referências

- Charu C. Aggarwal (2015) Data Mining: The Textbook. Springer International Publishing, 1st edition.