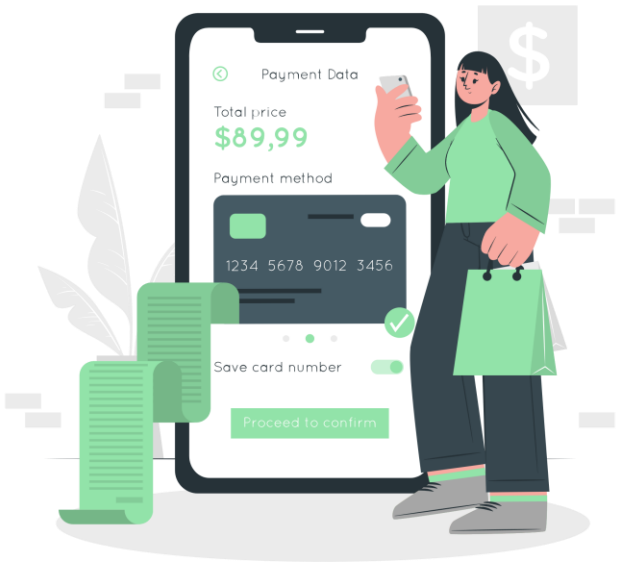
Mineração de Dados e CRISP-DM

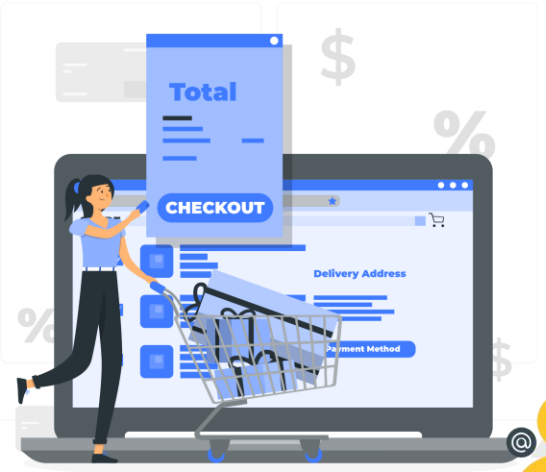
PUCRS
Pontifícia Universidade Católica do Rio Grande do Sul

– Aula 02 –
Coleta, Preparação e Análise de Dados

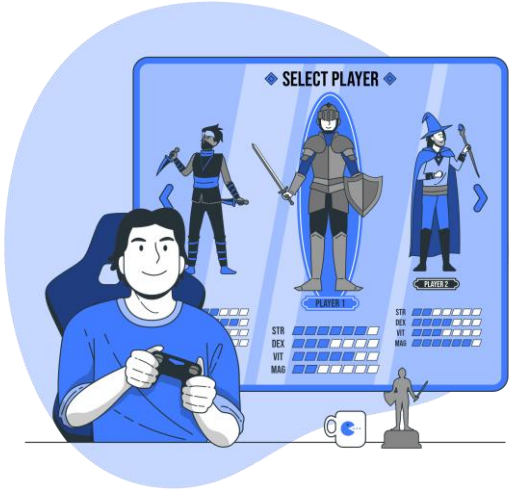
Prof. Me. Lucas R. C. Pessutto

Adaptação dos slides dos profs. Luan Garcia e Lucas Kupssinskü





Há dados em todo lugar!



Dados em larga escala

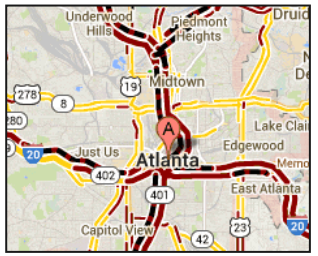
* Crescimento enorme de dados tanto em bases comerciais quanto científicas graças aos avanços em tecnologias de geração e coleta de dados

* Novo mantra
  → Coletar qualquer dado sempre que possível de qualquer lugar possível.

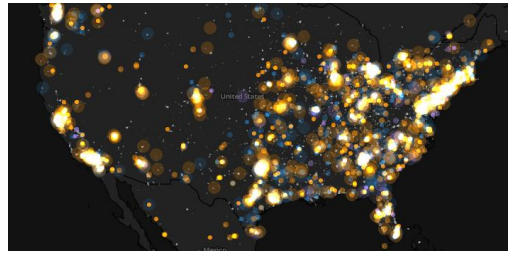
* Expectativa
  → Dados coletados terão valor seja pelo motivo de serem coletados seja por um motivo ainda não visto.

Cyber Security

E-Commerce

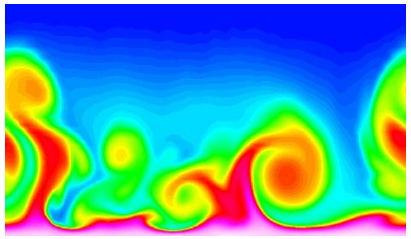


Traffic Patterns



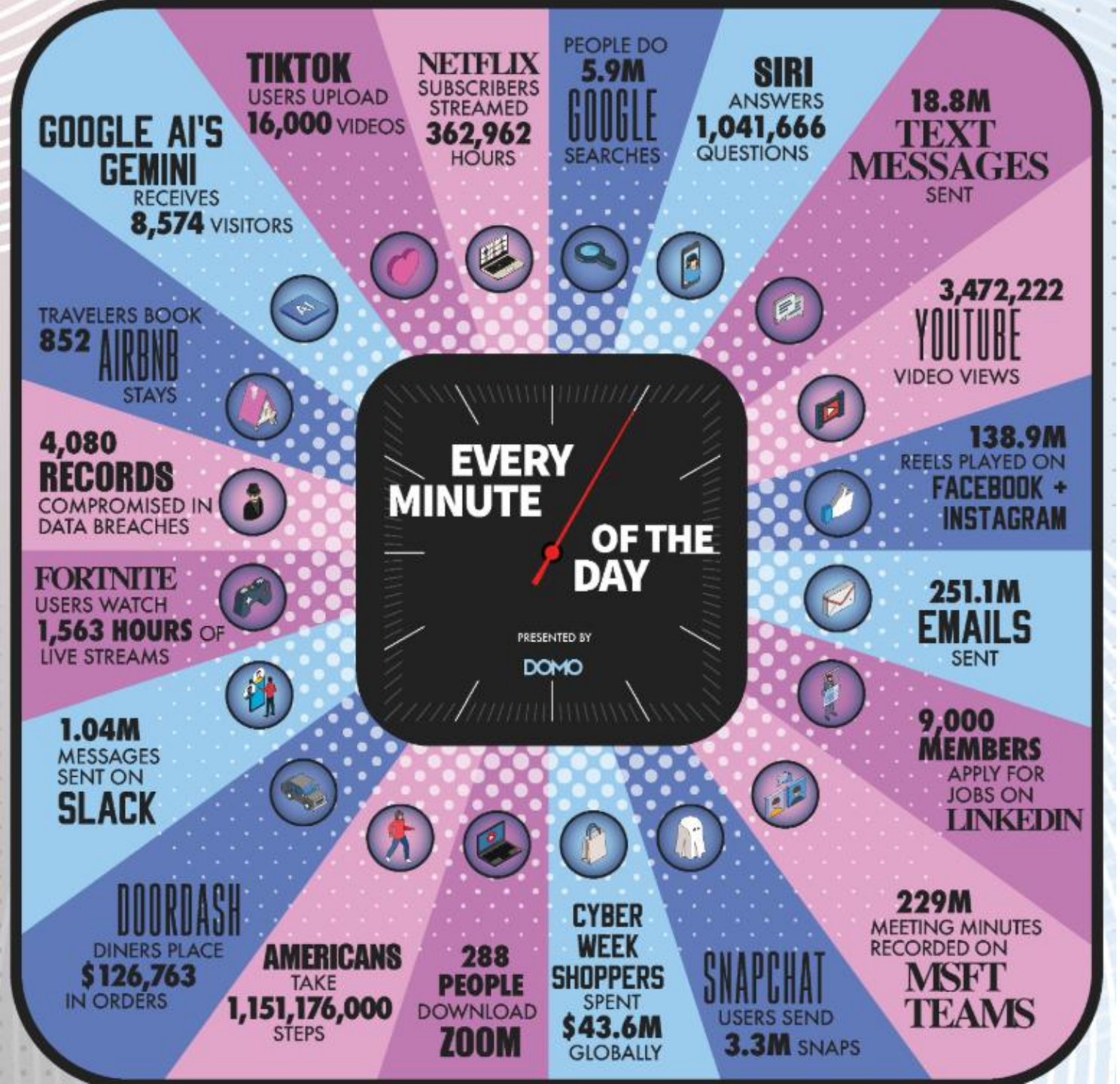
Social Networking: Twitter

Sensor Networks

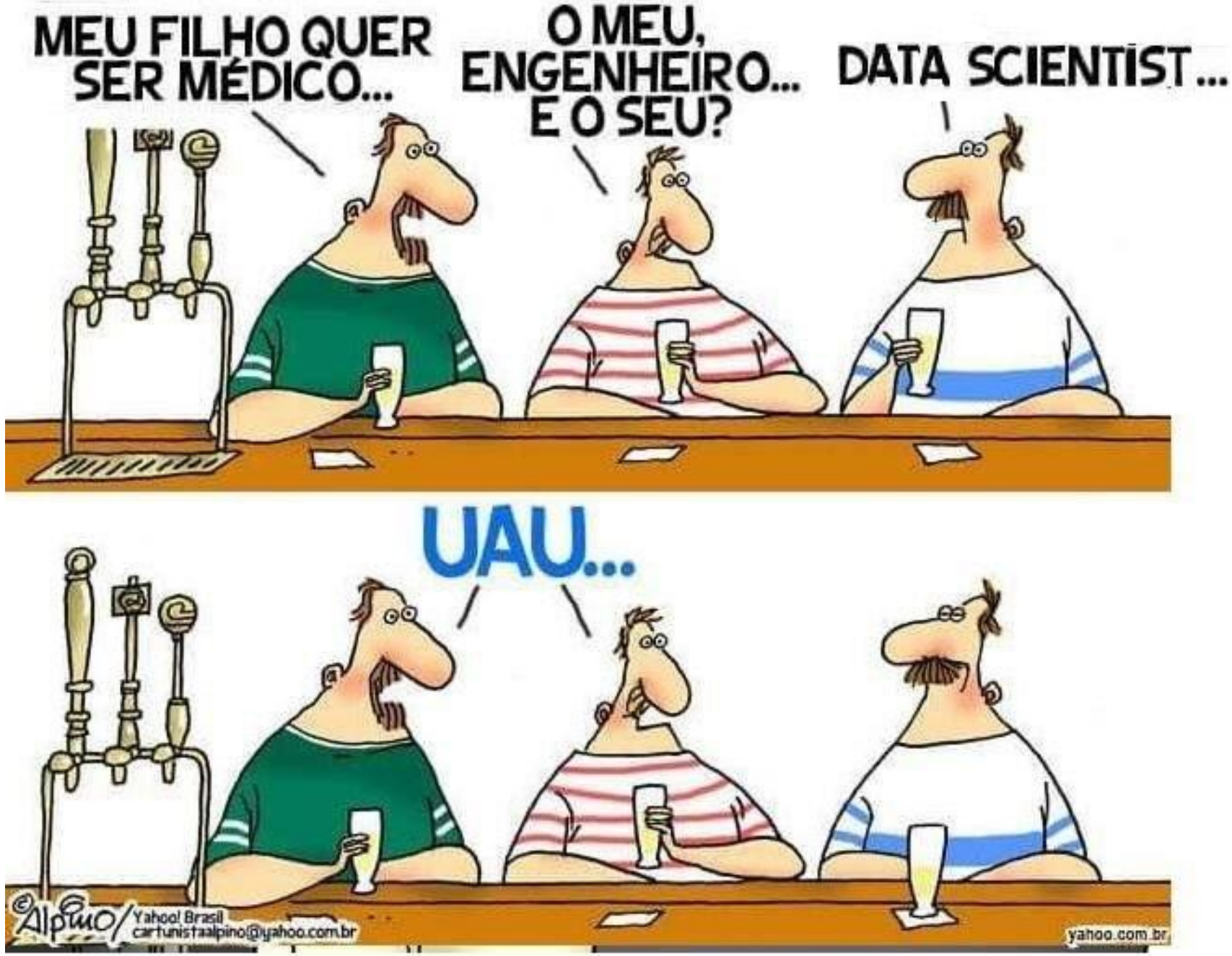


Computational Simulations

Dados referentes à 2024



https://www.domo.com/learn/infographic/data-never-sleeps-12


MEU FILHO QUER SER MÉDICO...
O MEU, ENGENHEIRO... E O SEU?
DATA SCIENTIST...
UAU...
©Alpino
Yahoo! Brasil
cartunistaalpino@yahoo.com.br
yahoo.com.br

Analista e cientista de dados serão as profissões mais relevantes no Brasil até 2030, indica pesquisa

As funções foram apontadas como as mais relevantes por 33% dos 477 participantes do levantamento; também foram citados os profissionais especializados em saúde mental (30%) e especialistas em inteligência artificial e machine learning (30%)

Por Redação Época NEGÓCIOS

22/09/2023 11h29 · Atualizado há 9 meses

[Analista e cientista de dados serão as profissões mais relevantes no Brasil até 2030, indica pesquisa | Futuro do trabalho | Época NEGÓCIOS (globo.com)]

Ciência de Dados

"Campo interdisciplinar que usa métodos científicos, processos, algoritmos e sistemas para extrair conhecimentos e insights de uma quantidade massiva de dados de vários tipos."

—Han, Pei & Tong (2023)

Dado, Informação e Conhecimento

* Dado:
  - → Conjuntos de símbolos que podem ser quantificados
  - → Não possuem significado relevante e não conduzem a nenhuma compreensão

100 milibares

poucas

30º

5.1 m/s; 95º

1000 m

Dado, Informação e Conhecimento

* Informação:
  → Ordenação e organização dos dados de forma a transmitir significado e compreensão dentro de um determinado contexto

Pressão atmosférica: 100 milibares

Nuvens: poucas

Temperatura: 30º

Velocidade e direção do vento: 5.1 m/s; 95º

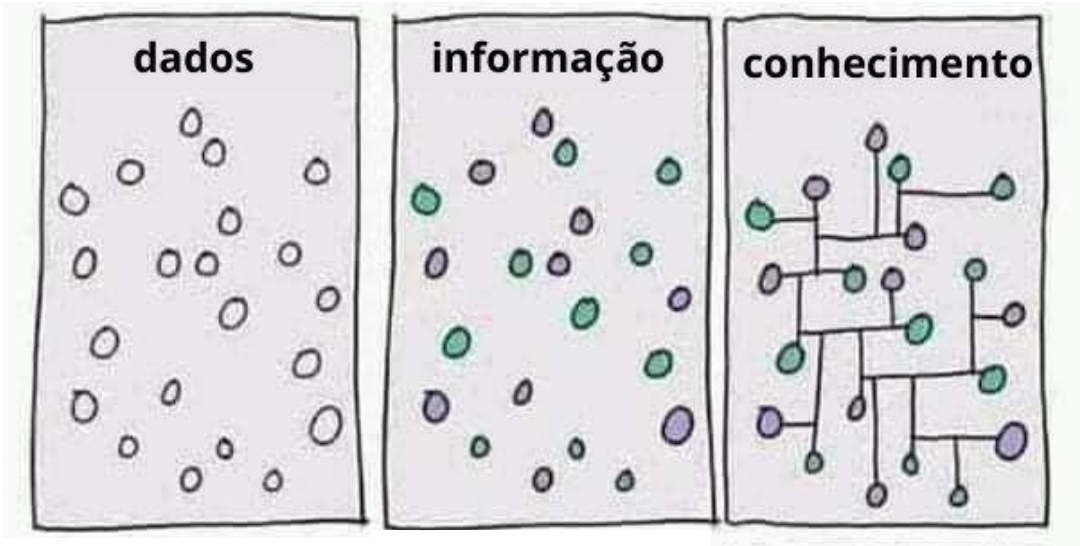
Visibilidade: 1000 m

Dado, Informação e Conhecimento

* Conhecimento:
  - → Comparação e combinação de informações que possuem alguma utilidade e um significado, e que permitem que sejam aferidas novas informações a partir dessas comparações
  - → O conhecimento constitui um saber. Produz ideias e experiências que as informações por si só não será capaz de mostrar.

A probabilidade de chuva é baixa. Posso ir à praia!

Dado, Informação e Conhecimento



Tomada de Decisão Orientada por Dados

* Tomada de decisão orientada por dados (DOD) refere-se à prática de basear as decisões na análise dos dados, em vez de apenas na intuição.

* Exemplo: Decisão sobre marketing em uma empresa
  - → Tradicional: Decidir uma política de anúncios baseados puramente em sua longa experiência na área e em sua intuição de que funcionará.
  - → DOD: Basear sua escolha na análise dos dados sobre a forma como os consumidores reagem a diferentes anúncios.

* No mundo real é comum utilizar uma combinação dessas abordagens.

* DOD não é uma prática do tipo “tudo ou nada”

Caso 1: Walmart

O furacão Frances estava a caminho, avançando pelo Caribe, ameaçando atingir a costa atlântica da Flórida. Os residentes se mudaram para terrenos mais elevados, porém distantes, em Bentonville, Arkansas. Executivos das lojas Walmart decidiram que a situação oferecia uma grande oportunidade para uma de suas mais recentes armas orientadas em dados: a tecnologia preditiva.

Uma semana antes de a tempestade atingir a costa, Linda M. Dillman, diretora executiva de informação, pressionou sua equipe para trabalhar em previsões baseadas no que havia acontecido quando o furacão Charley apareceu, várias semanas antes. Com o apoio dos trilhões de bytes de histórico de compras contidos no banco de dados do Walmart, ela sentiu que a empresa poderia "começar a prever o que aconteceria, em vez de esperar que acontecesse". (Hays, 2004)

Fonte: New York Times.

Caso 1: Walmart

* Porque previsões orientadas em dados são úteis neste cenário?

* Resposta 1: Para prever que as pessoas na trilha do furacão comprariam mais garrafas de água.
  → Óbvio demais! Será que precisa de ciência de dados para descobrir isso?

* Resposta 2: Para projetar o aumento nas vendas devido ao furacão, assegurando que os Walmarts locais estejam bem abastecidos.
  → Útil mas muito genérico! Um produto pode ter esgotado naquela semana em todas as lojas do Walmart do país, não somente na época do furacão!

Caso 1: Wallmart

* Seria mais valioso descobrir padrões não tão óbvios causados pelo furacão.
  → Por exemplo, examinar um grande volume de dados do Walmart em situações prévias semelhantes para identificar demanda local incomum de produtos.
* De fato, foi o isso que aconteceu! A matéria do NYT relata:

"... especialistas exploraram os dados e descobriram que as lojas realmente precisariam de certos produtos — e não apenas das habituais lanternas. 'Não sabíamos, no passado, que havia tido um aumento nas vendas de Pop-Tarts de morango, sete vezes acima do normal, antes de um furacão', disse a Sra. Dillman em uma entrevista recente. 'E o principal produto pré-furacão mais vendido era a cerveja."

Caso 2: Target

* Todo varejista preocupa-se com os hábitos de compra dos seus consumidores.

* A chegada de um novo bebê na família é um momento em que as pessoas mudam significativamente seus hábitos de compras. Nas palavras do analista da Target, “assim que percebemos que estão comprando nossas fraldas, eles comprarão todo o resto também”.

* Como a maior parte dos registros de nascimento é pública, os varejistas obtêm informações sobre nascimentos e enviam ofertas especiais para os novos pais.

Caso 2: Target

* A Target desejava sair na frente da concorrência, prevendo quando uma pessoa estava esperando um bebê.

* Usando técnicas de ciência de dados, a Target analisou dados históricos sobre os clientes que souberam posteriormente que estavam grávidas, e foi capaz de obter informações que poderiam predizer quais consumidores estavam esperando um bebê.

* Por exemplo, mulheres grávidas costumam mudar a dieta, o guarda-roupa, as vitaminas e assim por diante. Esses indicadores podem ser extraídos dos dados históricos!

Caso 2: Target

* Algumas descobertas da Target:
  - → Loções sem cheiro são compradas por mulheres no início de seu segundo trimestre de gravidez.

  - → Em algum momento nas primeiras 20 semanas, mulheres grávidas tendem a comprar em suplementos como cálcio, magnésio e zinco.

  - → Quando lotes de sabão sem cheiro e sacos extragrandes de bolas de algodão, além de desinfetantes para as mãos e panos, sinaliza que a data do parto está chegando.

Caso 2: Target

Forbes

FORBES > TECH

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Kashmir Hill Former Staff

Welcome to The Not-So Private Parts where technology & privacy collide

Feb 16, 2012, 11:02am EST

TARGET

https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

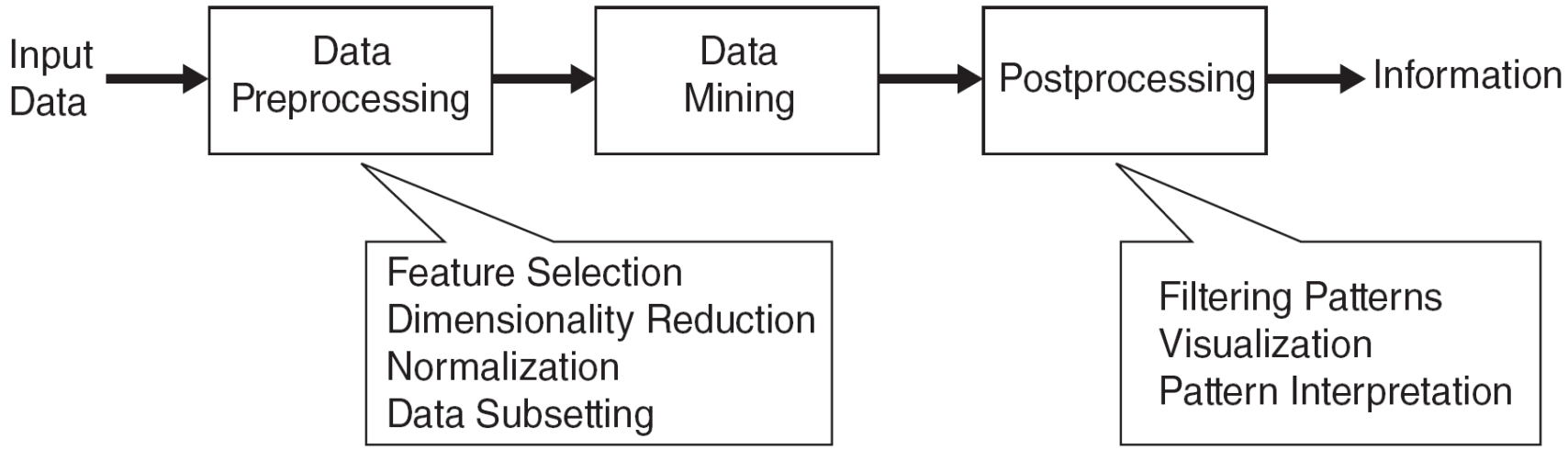
Mineração de Dados

"We are drowning in information,
but starving for knowledge."

– John Naisbitt (1982)

Mineração de Dados

* Extração não trivial de informação implícita, desconhecida e potencialmente útil a partir de dados.
* Exploração e análise de dados, utilizando meios automáticos ou semiautomáticos, para descobrir padrões significativos.



Mineração de Dados

* Termo tem origem em analogia à mineração de metais preciosos
  → Encontrar metais de maior valor entre pedras e areia sem muito valor

* Mineração de dados
  → Encontrar informação potencialmente valiosa no meio de uma grande quantidade de dados

* Geralmente se aceita que faz parte de um processo maior, chamado de Knowledge Discovery in Databases (KDD)

Por que Minerar? (Comercial)

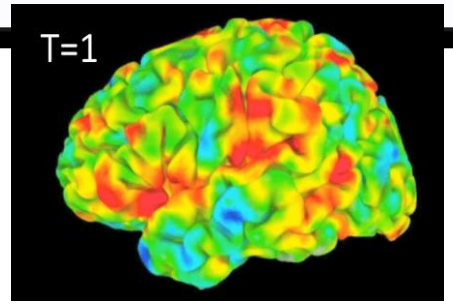
* Muitos dados estão sendo coletados e armazenados
  - → Dados na web
  - → Google tem petabytes de dados web
  - → Facebook tem bilhões de usuários
  - → Compras online
  - → Amazon tem milhões de visitas por dia

* Transações bancárias
* Computadores e armazenamento cada vez mais potentes e baratos
* Competitividade no mercado é grande
  - → Fornecer serviços melhores e customizados traz uma vantagem competitiva

Por que Minerar? (Científico)

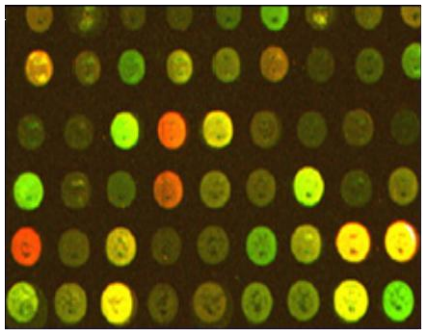
* Dados coletados e armazenados em uma velocidade muito alta
  - → Sensores remotos em satélites
    - Dados climáticos, controle de florestas
  - → Telescópios varrendo o céu
    - Dados do espaço
  - → Simulações científicas
    - Diversos modelos em poucas horas

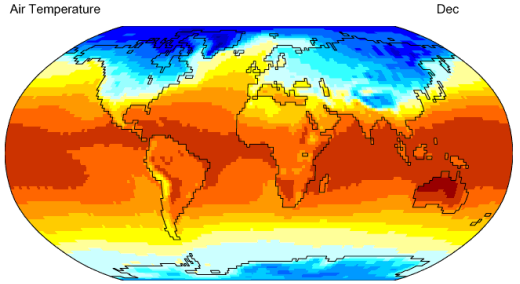
* Data mining ajuda os cientistas na
  - → Análise de datasets gigantescos
  - → Formulação e validação de hipóteses



fMRI Data from Brain


Sky Survey Data

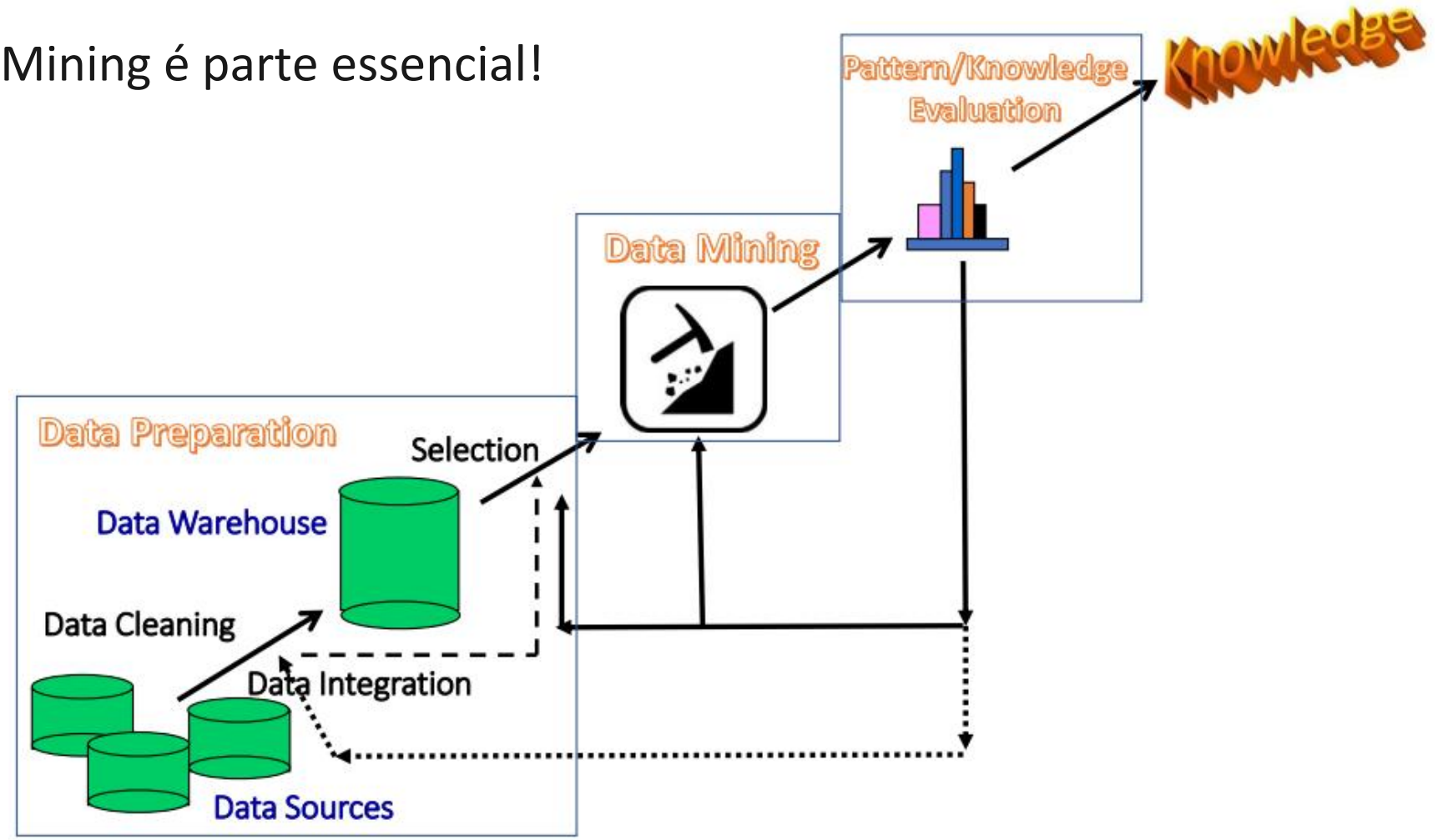

Gene Expression Data



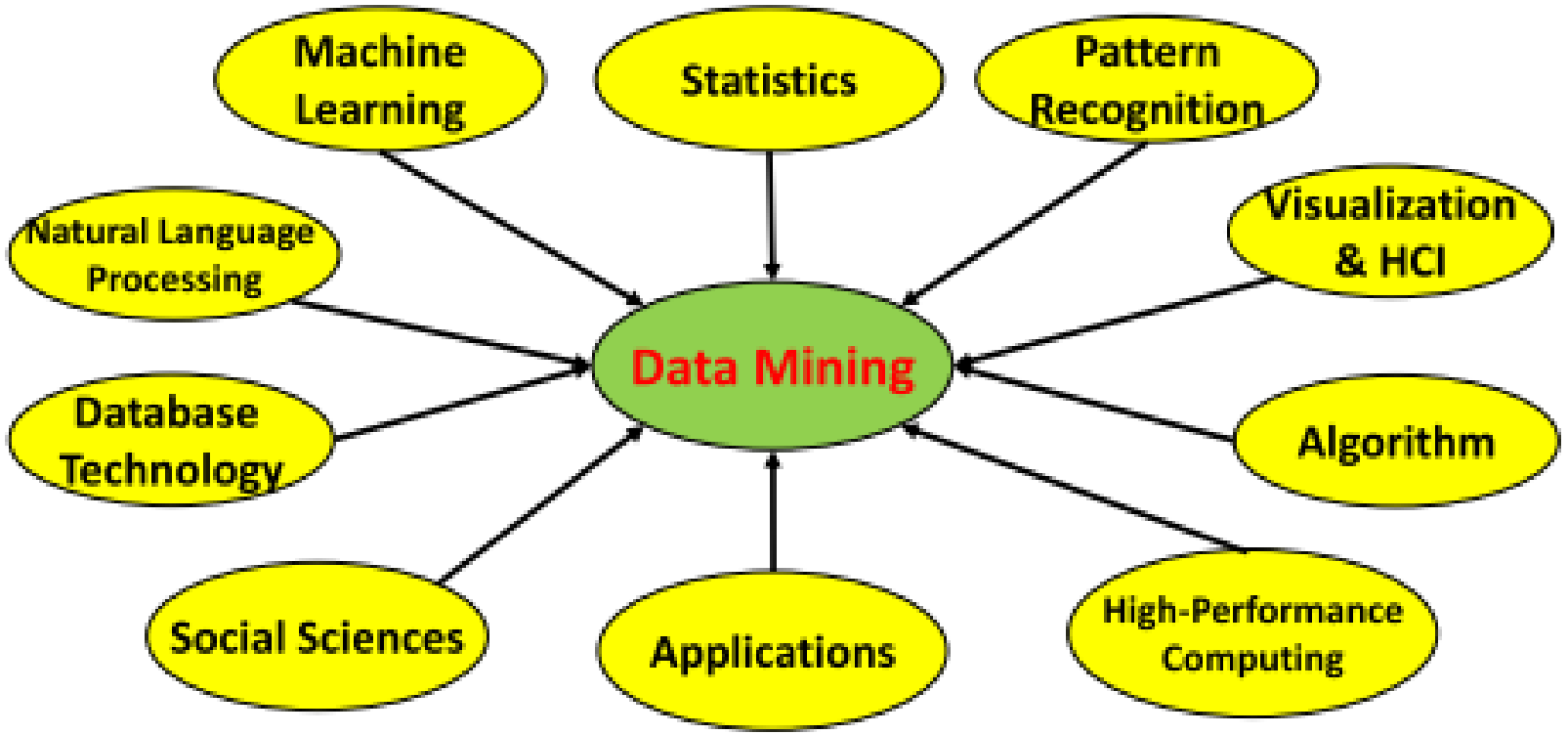
Surface Temperature of Earth

Mineração de Dados

* Data Mining é parte essencial!



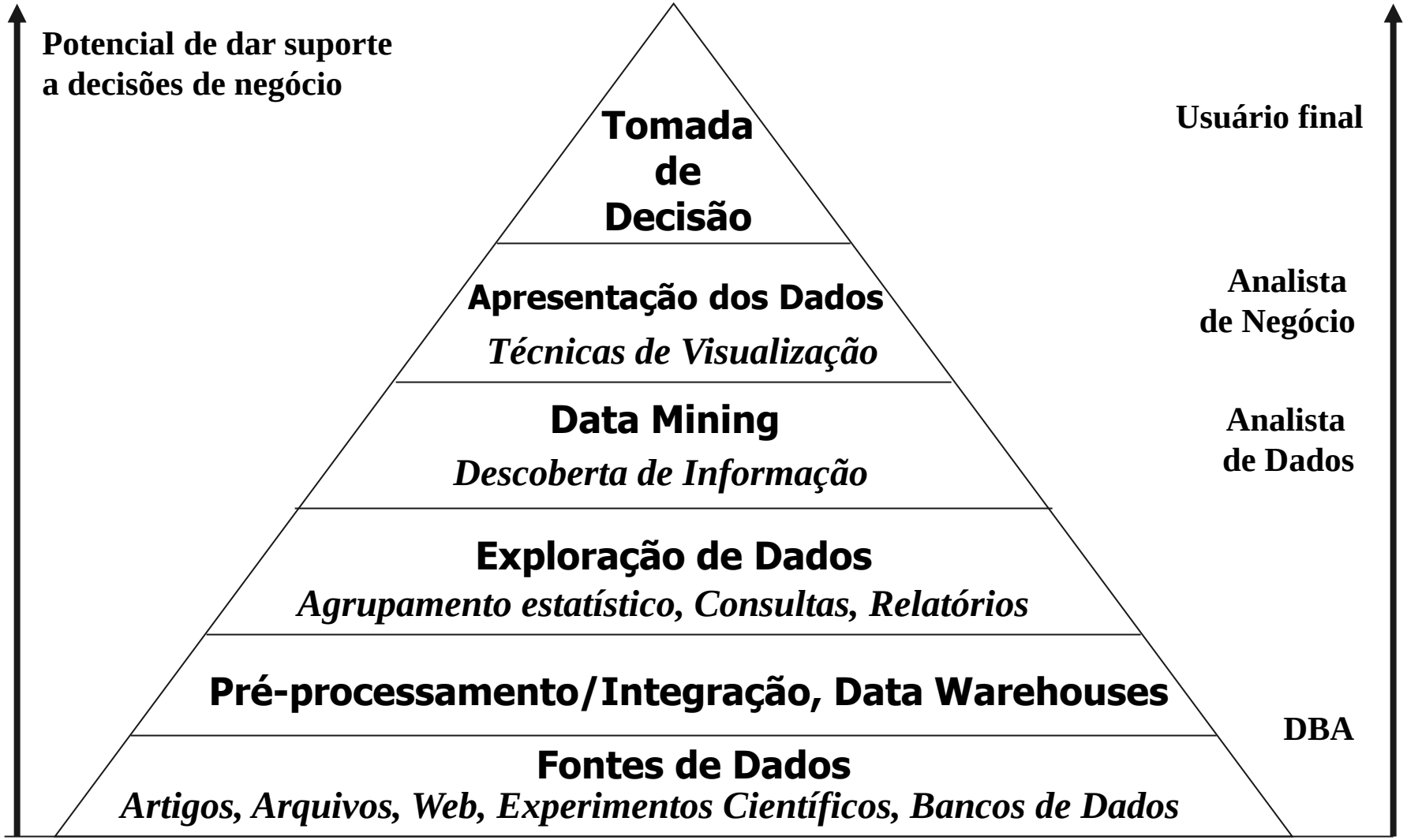
Confluência de Diversos Campos



Por que diversos campos?

* Quantidade gigantesca de dados
  → Algoritmos precisam ser escaláveis para lidar com petabytes de dados
* Dados com muitas dimensões
  → Algumas vezes milhares de dimensões
* Dados extremamente complexos
  → Data streams e dados de sensoriamento
  → Séries temporais, dados espaciotemporais.
  → Vários tipos de dados.
  → Base de dados heterogêneas e bancos legados.
  → Aplicações comerciais, simulações científicas.

Data Mining e Business Intelligence



Processo de Mineração

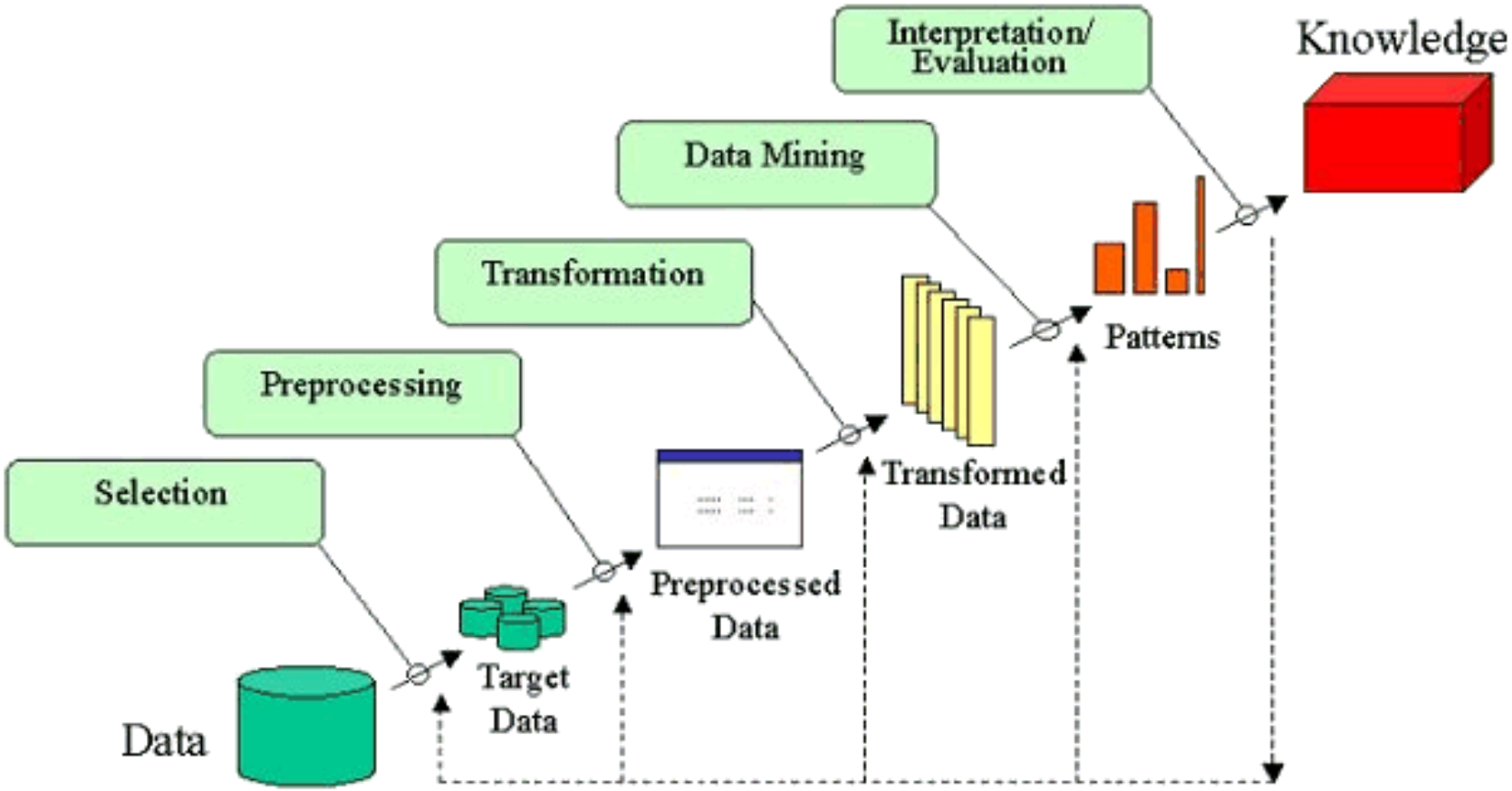
* Conjunto de etapas necessários para executar o processo completo de mineração de dados.

* Similar ao processo de engenharia de software

* Diversas metodologias propostas ao longo dos anos.

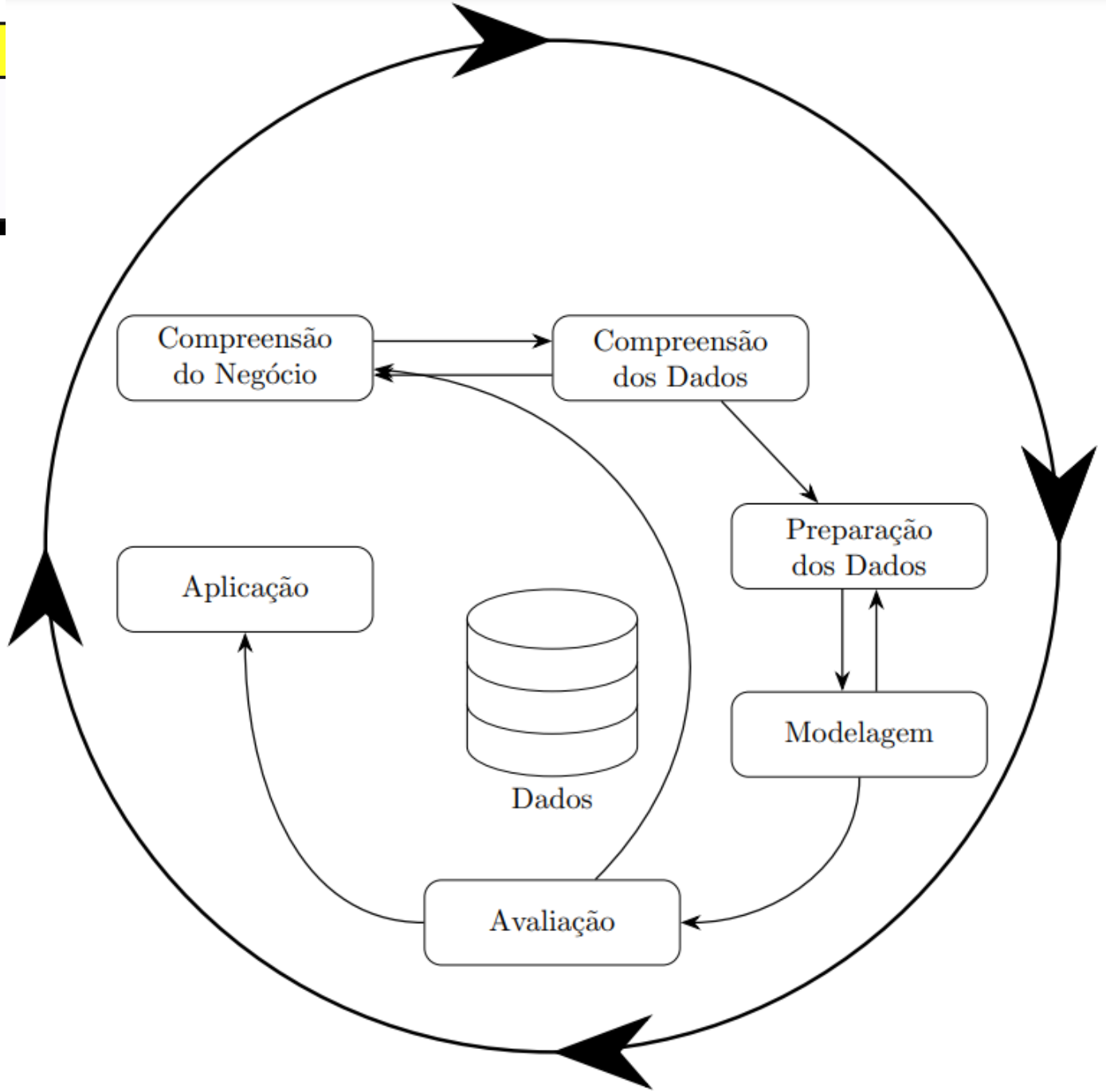
* Vejamos alguns exemplos...

Knowledge Discovery in Databases - KDD



Processo de Análise de Dados da Google

CRISP-DM



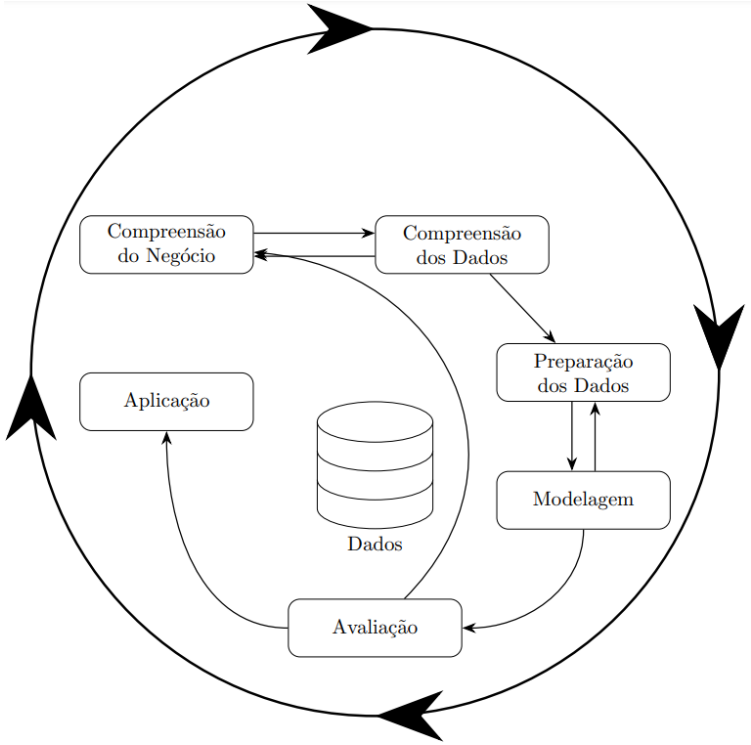
Metodologia usada nesta disciplina

(e no curso!)

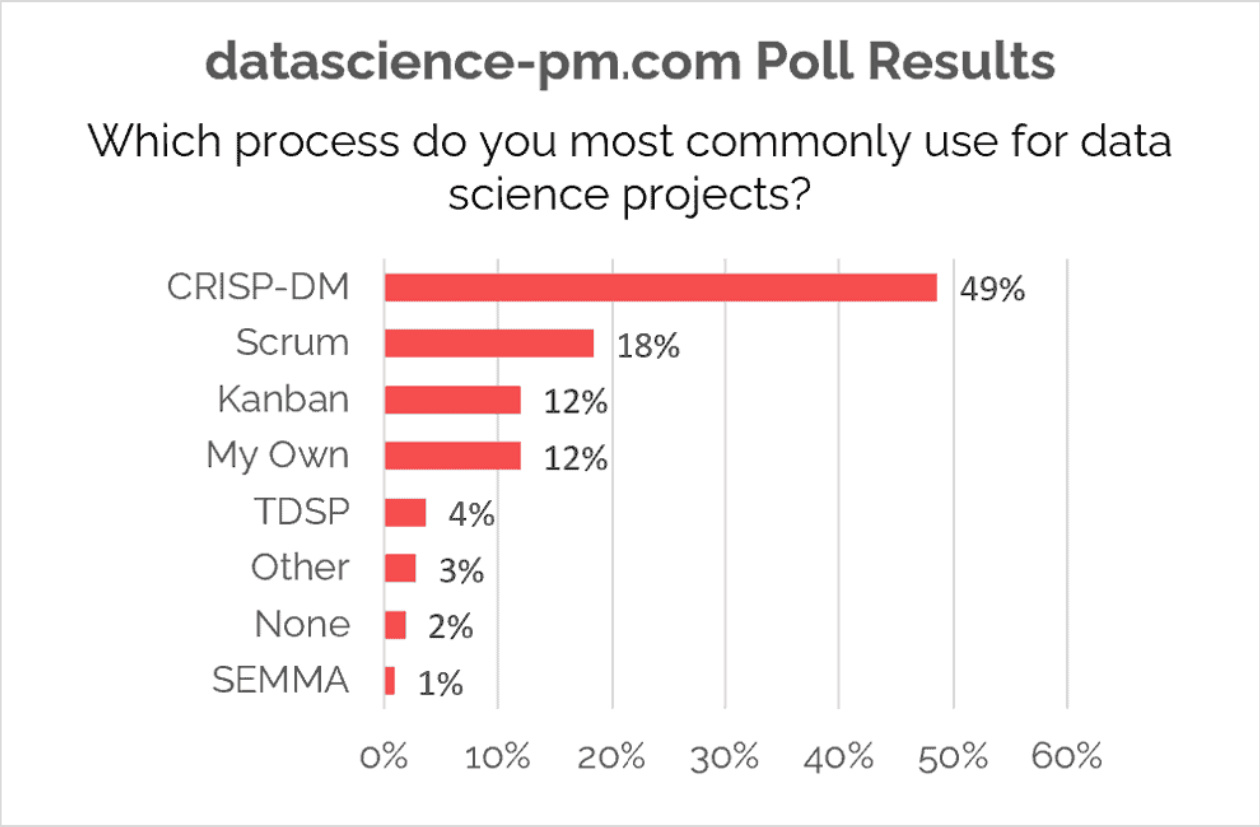
CRISP-DM

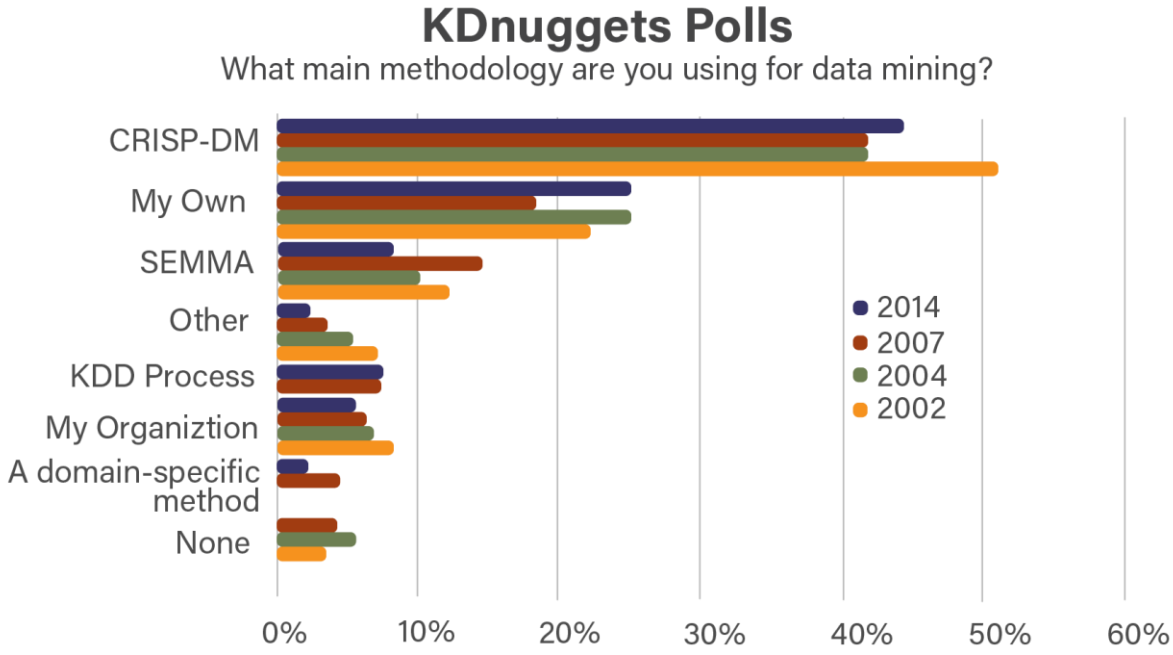
* CRISP-DM: **CR**oss **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining

* É um processo padronizado de descoberta de conhecimento que consiste de seis fases:
  - → Compreensão do Negócio
  - → Compreensão dos Dados
  - → Preparação dos Dados
  - → Modelagem
  - → Avaliação
  - → Aplicação



CRISP-DM





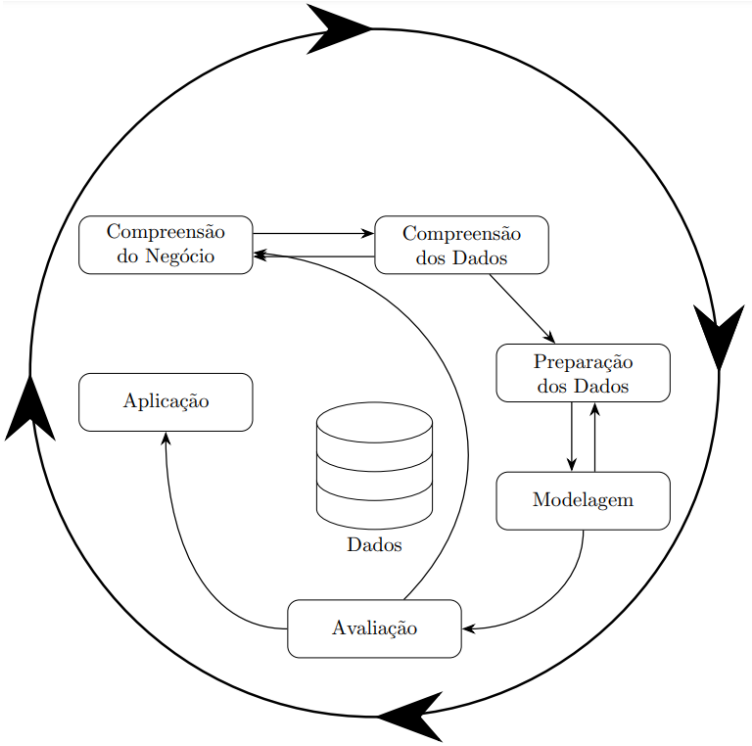
Abril de 2024 – 109 participantes

CRISP-DM

* A sequência das fases do processo não é estrita. Avançar e retroceder pelas fases do processo é sempre necessário, dependendo do resultado de uma fase ou tarefa.

* As setas da figura indicam as principais (e mais importantes) dependências entre as fases

* O círculo externo simboliza a natureza cíclica do processo de mineração de dados
  - → A mineração de dados não termina quando uma solução é entregue!
  - → As lições que foram aprendidas durante o processo e do uso da solução desenvolvida podem levar a novas questões.



Características do CRISP-DM

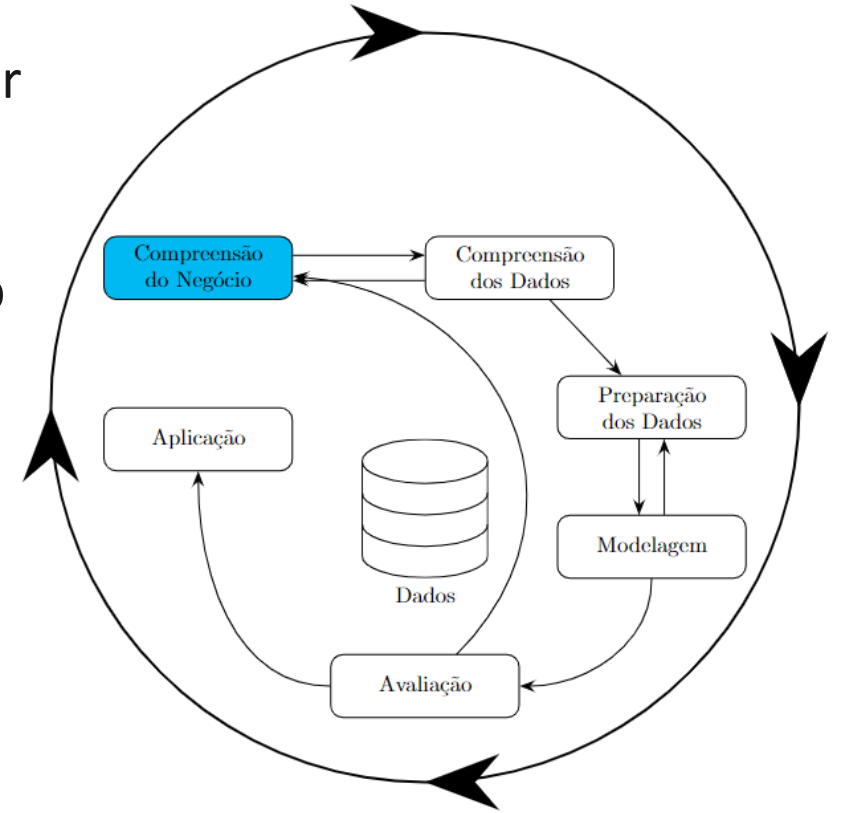
* Metodologia que foca no registro de experiências
* Permite que projetos sejam facilmente replicáveis
* Ajuda no planejamento e na gestão de projetos
* Facilmente aprendido por novos utilizadores
* Demonstra a maturidade do Processo de Descoberta de Conhecimento

Compreensão do Negócio

* Fase que foca em identificar o problema, entender como podemos resolve-lo e se a mineração de dados ajudará a resolver este problema

* Envolve compreender os objetivos do projeto e os requisitos sob a perspectiva de negócio

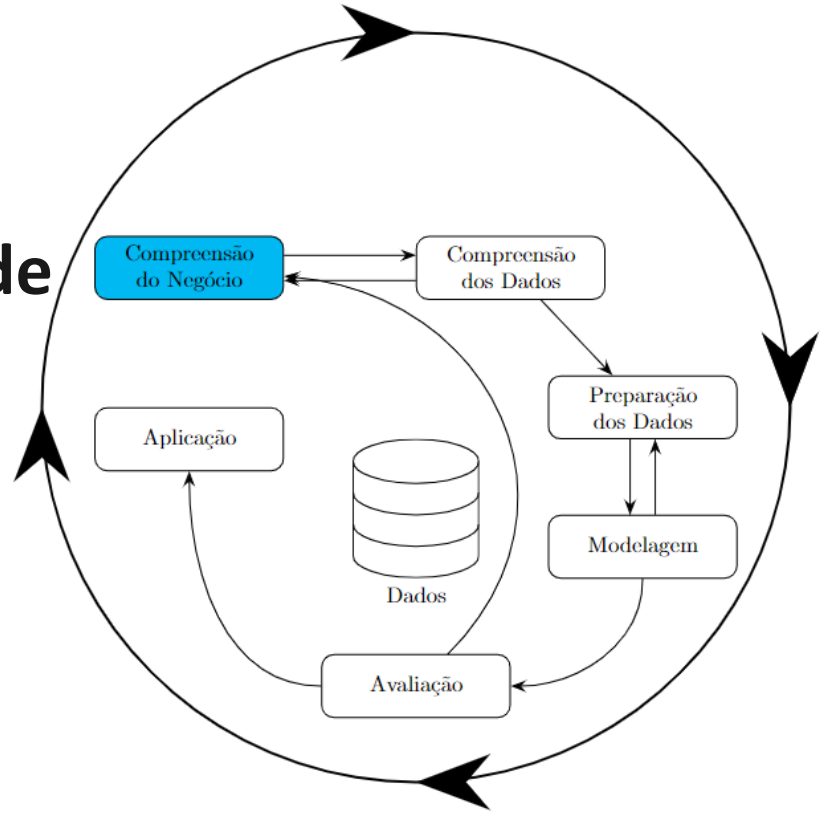
* Engloba quatro fases:
  - → Determinar os objetivos do negócio
  - → Avaliar a situação
  - → Determinar os objetivos da mineração de dados
  - → Produzir o plano do projeto



Compreensão do Negócio

O que se pretende atingir com a mineração de dados?

Qual o critério de sucesso?

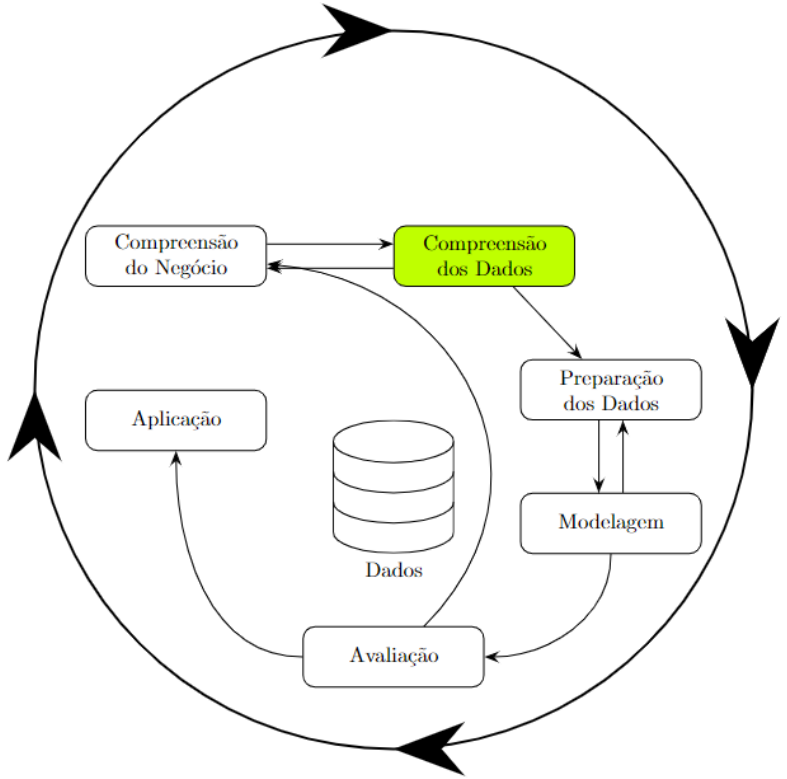


Compreensão dos Dados

* Nesta fase, avaliamos as fontes de dados que estão disponíveis e determinamos se é necessário coletar mais dados

* É possível que as descobertas dessa fase influenciem no objetivo estabelecido na etapa anterior, sendo necessário retroceder uma etapa e ajustar o objetivo de acordo com as descobertas feitas nessa etapa

* Envolve quatro fases:
  - → Coletar os dados iniciais
  - → Descrever os dados
  - → Explorar os dados
  - → Verificar a qualidade dos dados

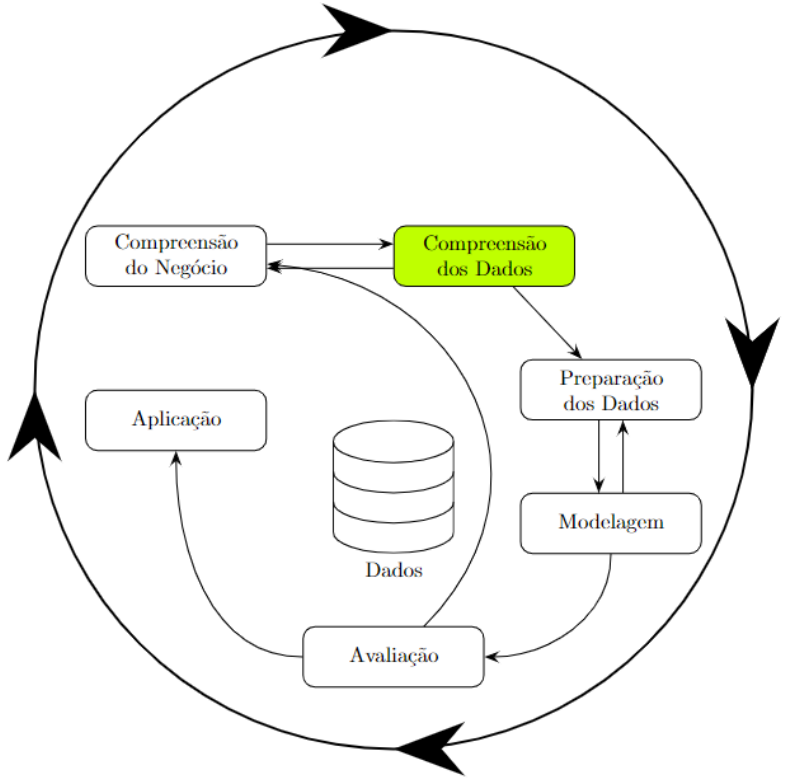


Compreensão dos Dados

Coletar e organizar os dados que serão analisados

Garantir a qualidade dos dados

Realizar o entendimento básico dos dados obtidos

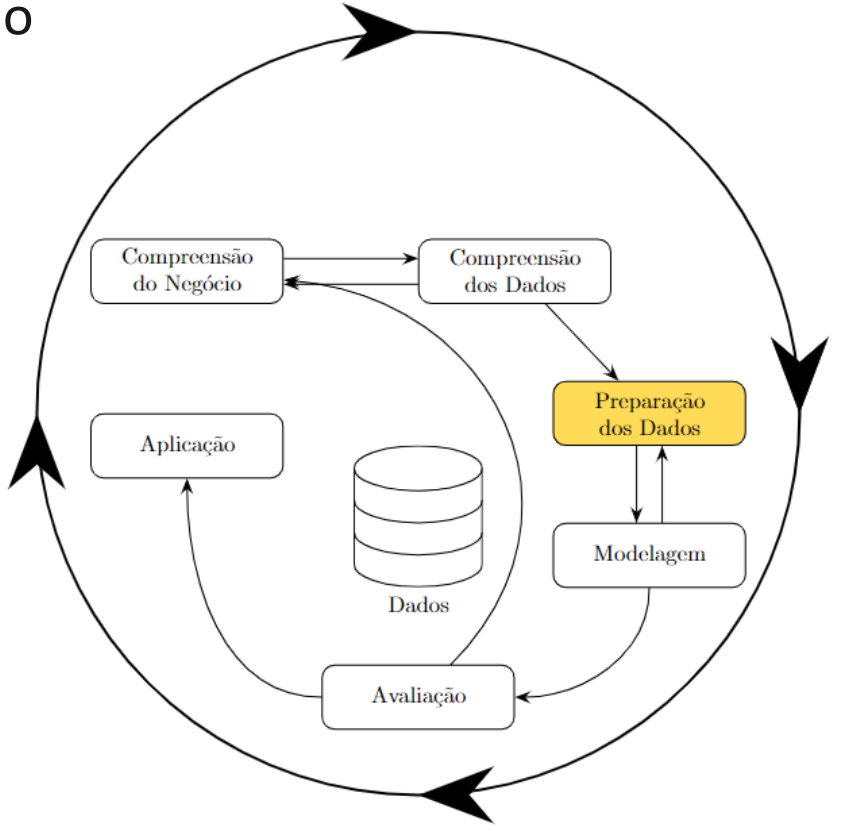


Preparação dos Dados

* Nesta etapa realizamos a limpeza dos dados, isto é, preparamos o dataset que será a utilizado na fase de modelagem.

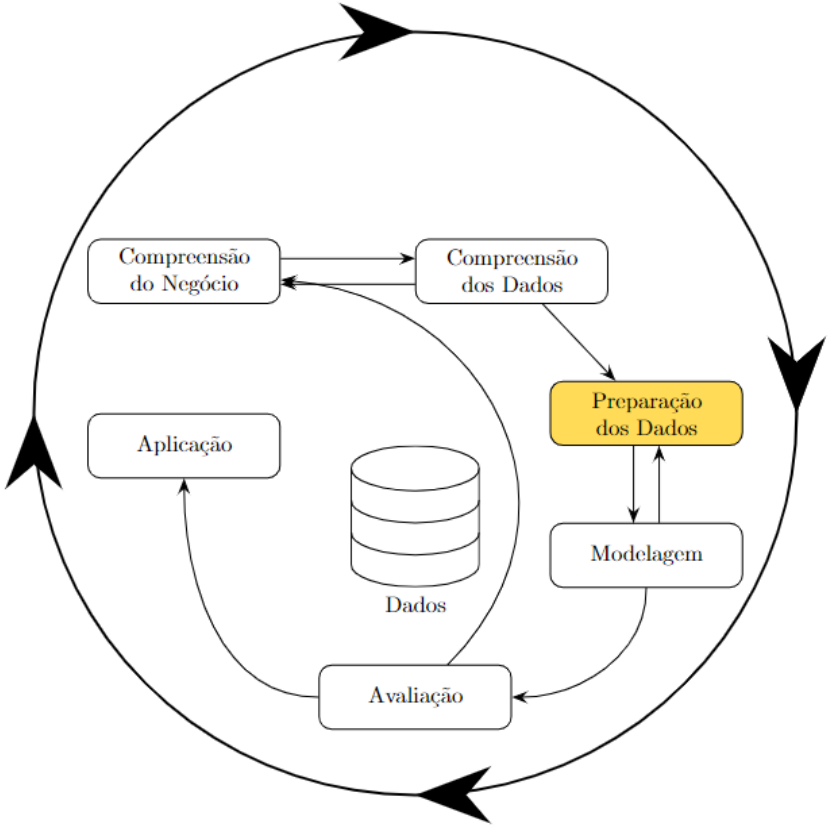
* É a fase mais trabalhosa de todo o processo!

* Envolve seis etapas:
  - → Selecionar os dados
  - → Limpeza dos dados
  - → Construção de dados
  - → Integração de dados
  - → Formatar os dados
  - → Criação do dataset



Preparação dos Dados

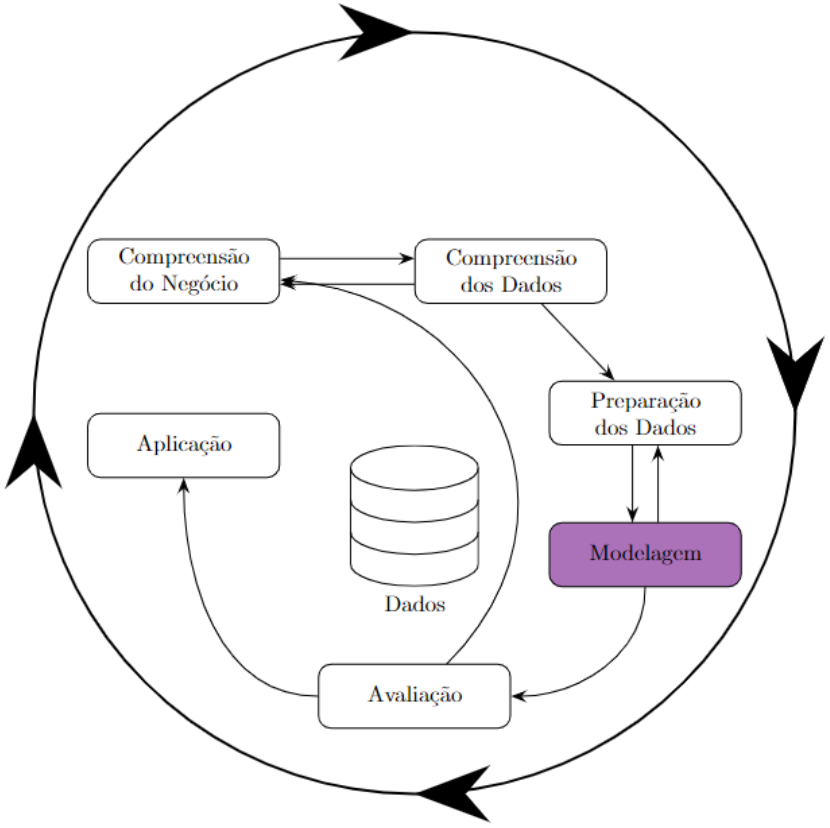
Garantir que os dados estão prontos para a etapa de modelagem



Modelagem

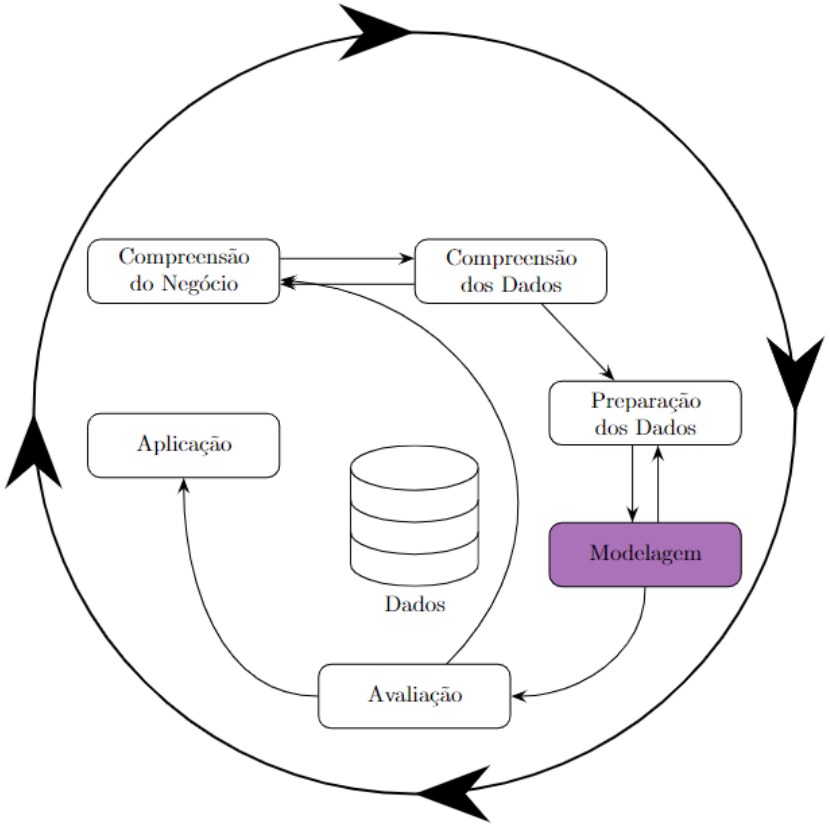
* Esta fase envolve a escolha do modelo que será utilizado para resolver o problema, como obter o melhor desse modelo com os dados disponíveis e da performance dos modelos gerados.

* Quatro etapas são seguidas nessa fase:
  - → Seleção das técnicas de modelagem
  - → Projetar os testes de performance
  - → Construção dos modelos
  - → Avaliação dos modelos



Modelagem

Construção de modelos de Aprendizagem de Máquina

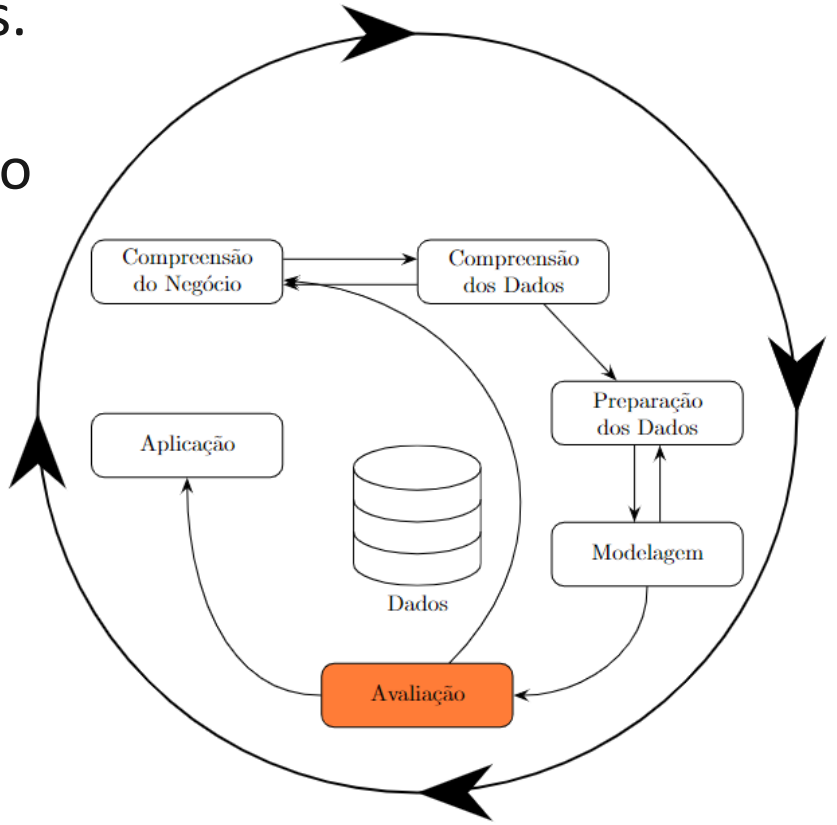


Avaliação

* Consiste em ver se o modelo gerado atende às expectativas.

* De maneira objetiva, consiste em ver se o modelo atende ao objetivo definido na etapa de compreensão do negócio.

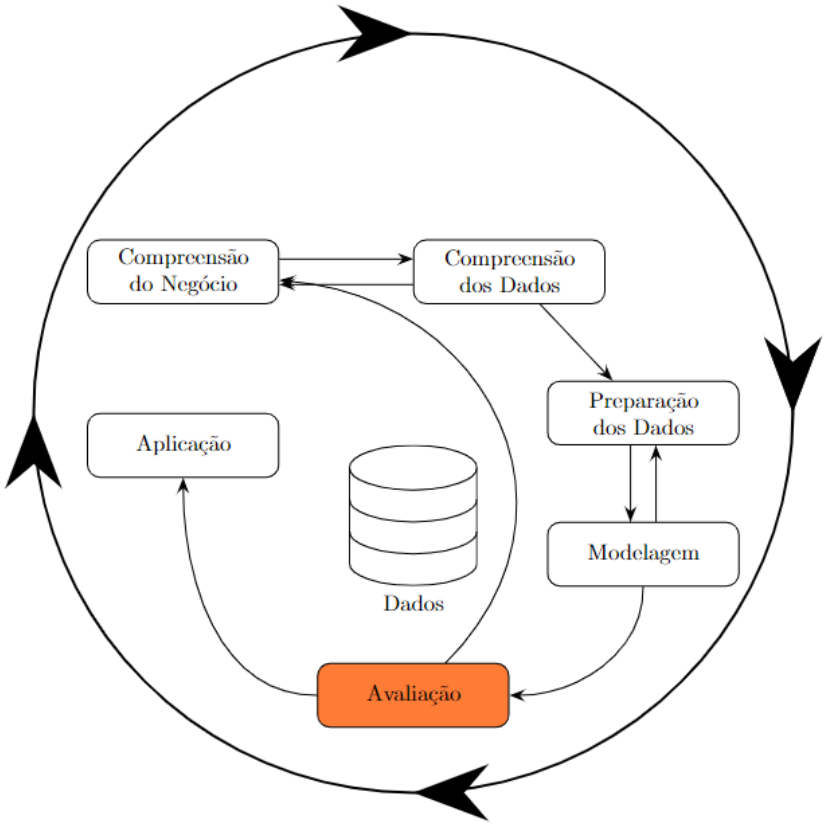
* Consiste de três etapas:
  - → Avaliar resultados
  - → Revisão do processo
  - → Determinar os próximos passos



Avaliação

Análise crítica de resultados

Os critérios de sucesso foram atingidos?

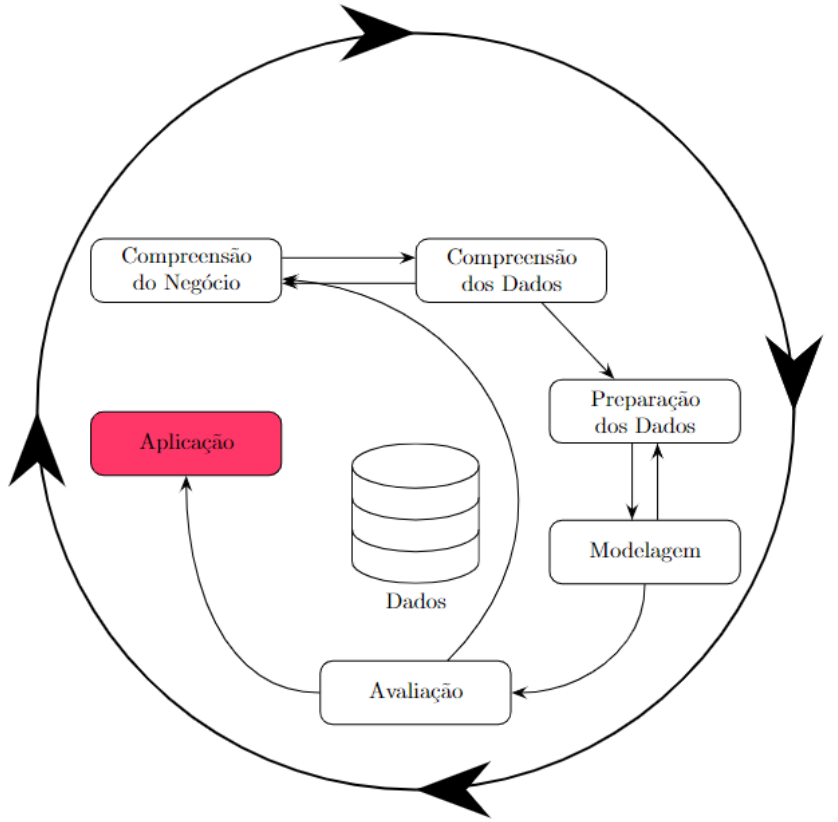


Aplicação (Deploy)

* Entrega do modelo gerado para o cliente e checar se ele atendeu aos requisitos do negócio

* Ao final desta etapa, voltamos para a etapa inicial e refletimos, com base no conhecimento adquirido no processo, sobre o que foi atingido no projeto (ou não)

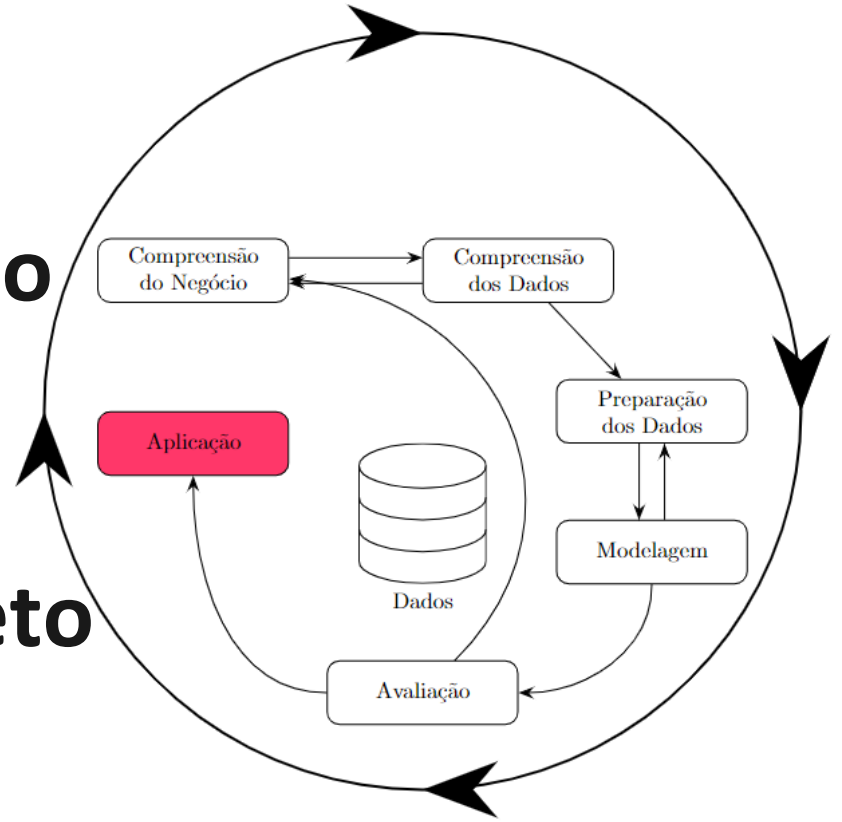
* Essa fase envolve:
  - → Planejar o deployment
  - → Planejar o monitoramento e a manutenção
  - → Produzir o relatório final
  - → Revisão do projeto



Aplicação (Deploy)

Colocar em prática o conhecimento obtido.

Revisar, avaliar e monitorar o projeto como um todo



Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Select Data
Rationale for Inclusion/ Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Dataset
Dataset Description

Modeling

Select Modeling Techniques
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Descriptions

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Para saber mais...

* Data Mining: Concepts and Techniques. (Han, J., Pei, J., Tong, H.) 4th edition. 2023.

* Introduction to Data Mining. (Tan, Steinbach, Karpatne, Kumar) 2nd Edition

* Machine Learning Bookcamp. (Alexey Grigorev) 1st edition, 2021.

* Data Science para Negócios (Foster Provost, Tom Fawcett) 1st edition, 2016

* Exploratory Data Analysis. (Marc A.T. Teunis, Jan-Willem Lankhaar), 2022.

* Documentação oficial do CRISP-DM