Ciência de Dados

Profa. Solange Kanso
solange.kanso@uni9.pro.br

UNINOVE/SP
1º/2024

O QUE É CIÊNCIA DE DADOS?

- “A ciência de dados é o estudo dos dados para extrair insights significativos para os negócios. Ela é uma abordagem multidisciplinar que combina princípios e práticas das áreas de matemática, estatística, inteligência artificial e engenharia da computação para analisar grandes quantidades de dados. Essa análise ajuda os cientistas de dados a fazer e responder perguntas como o que aconteceu, por que aconteceu, o que acontecerá e o que pode ser feito com os resultados” (AWS)

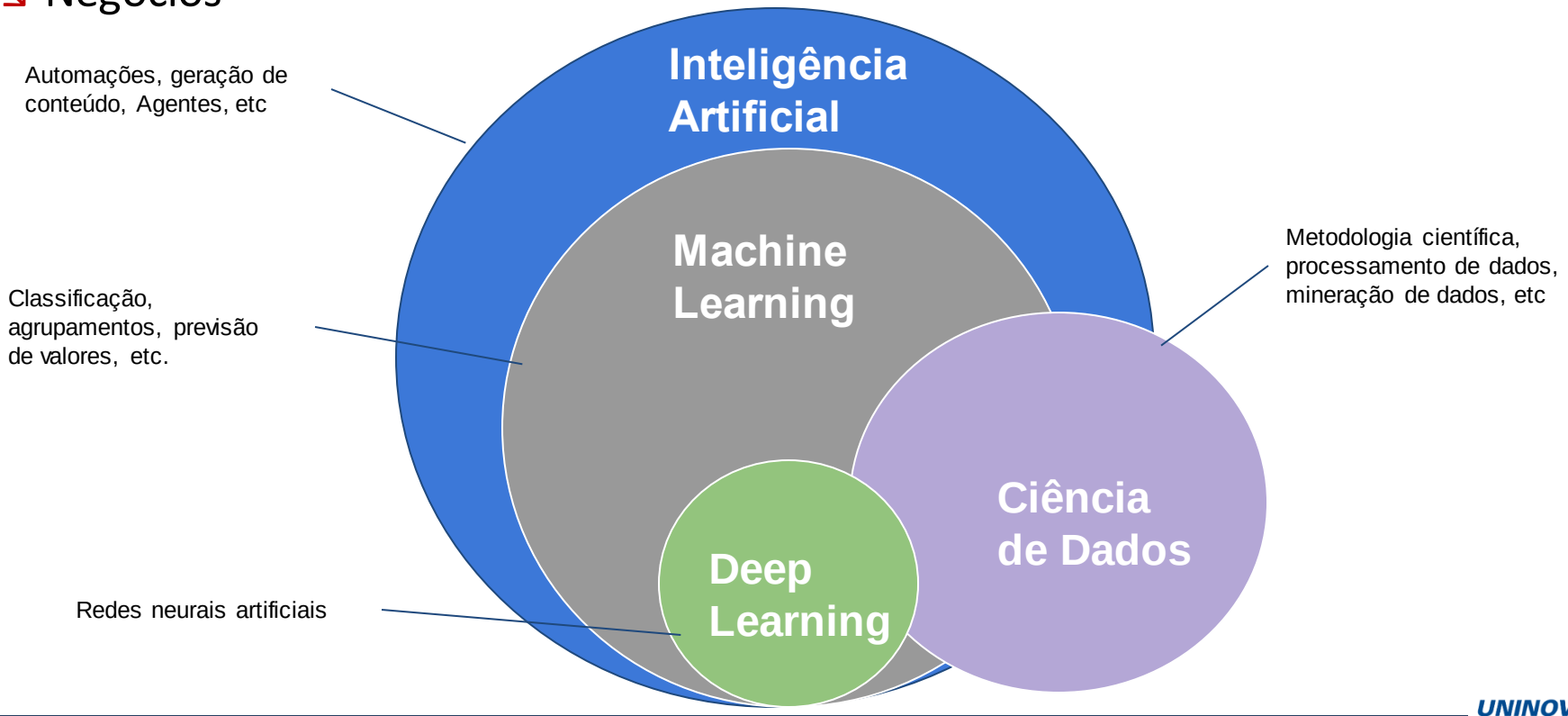
- “A ciência de dados combina matemática e estatística, programação especializada, análise avançada, inteligência artificial (IA) e machine learning com conhecimento em assuntos específicos para descobrir insights práticos, ocultos nos dados de uma organização. Esses insights podem ser usados para orientar a tomada de decisões e o planejamento estratégico” (IBM)

- “O estudo de dados e algoritmos para a resolução de problemas” (EBAC online)

Pilares da Ciência de Dados

Interdisciplinaridade

- ↘ Matemática e Estatística
- ↘ Ciência da computação
- ↘ Negócios



O QUE É CIÊNCIA DE DADOS?

- Conhecida também como *Data Science;*
- Área interdisciplinar que tem como objetivo o estudo e análise de dados para extração de informações;
- As fontes de dados são variadas;
- Estruturadas e não estruturadas;

Slide cedido gentilmente pelo Prof. Wanderley Jr.

CRIANDO SENTIDO PARA QUE TODOS ENTENDAM

DADOS

ORDENADOS

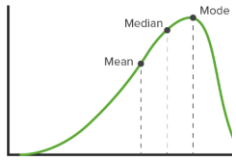
ESTRUTURADOS

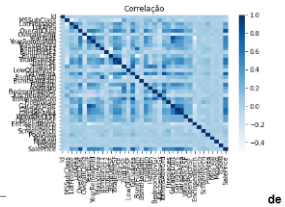
VISUALMENTE APRESENTADOS

EXPLICADOS EM UMA HISTÓRIA

Foto: Branden Rossen | Designer: Karyn Lurie

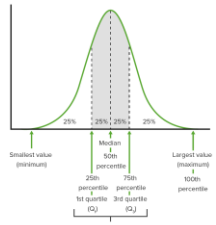
UTILIZAMOS ESTATÍSTICA DESCRITIVA E INFERENCIAL

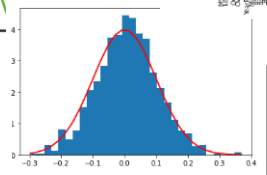


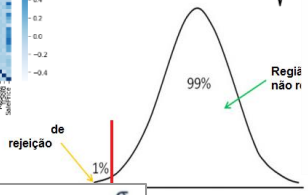


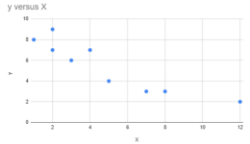
$$D_P = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M_A)^2}{n}}$$

Grau de confiança	α	Valor Crítico Zα/2
90%	0,1	1,645
95%	0,05	1,96
99%	0,01	2,575







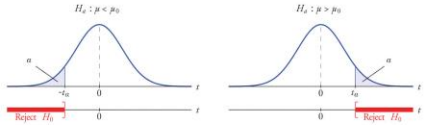


$$n = \frac{N \cdot \sigma^2 \cdot (Z_{\alpha/2})^2}{(N-1) \cdot E^2 + \sigma^2 \cdot (Z_{\alpha/2})^2}$$

LIMITE INFERIOR: $\bar{x} - z \times \frac{\sigma}{\sqrt{n}}$

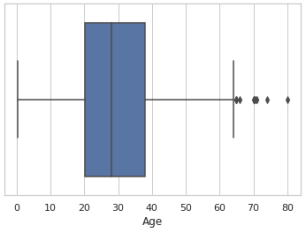
LIMITE SUPERIOR: $\bar{x} + z \times \frac{\sigma}{\sqrt{n}}$

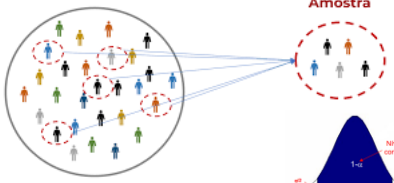
$$n = \left(\frac{Z_{\alpha/2}}{E}\right)^2 = \left(\frac{1{,}96 \cdot 6250}{500}\right)^2 = 600{,}25$$

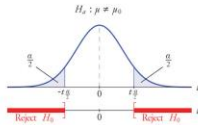


$$c_v = \frac{\sigma}{\bar{x}}$$

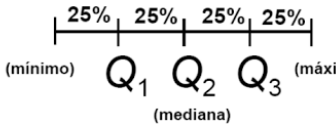
$$n = \frac{Z_{\alpha/2}^2 \cdot 0{,}25}{E^2}$$



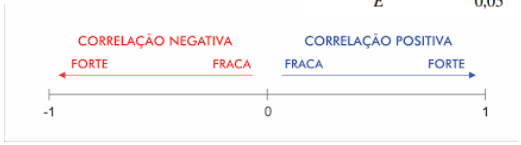




$$n = \frac{N \cdot \hat{p} \cdot \hat{q} \cdot (Z_{\alpha/2})^2}{\hat{p} \cdot \hat{q} \cdot (Z_{\alpha/2})^2 + (N-1) \cdot E^2}$$



$$n = \frac{[Z_{\alpha/2}]^2 \cdot 0{,}25}{E^2} = \frac{1{,}645^2 \cdot 0{,}25}{0{,}05^2} = 270{,}6$$



Nível de Confiança	Valor de Z*-
80%	1.28
90%	1.645 (convencional)
95%	1.96
98%	2.33
99%	2.58

		Valor Predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

DADOS ESTRUTURADOS E NÃO ESTRUTURADOS

Dados estruturados

- Possuem organização;
- Separação clara e bem definida da informação que está sendo gravada em cada porção da estrutura de armazenamento;
- Organizados por linhas e colunas;
- Em geral, cada linha representa um novo registro nesta tabela e/ou planilha eletrônica;

Slide cedido gentilmente pelo Prof. Wanderley Jr.

DADOS ESTRUTURADOS E NÃO ESTRUTURADOS

Dados NÃO estruturados

- Referem-se a dados que não podem ser organizados em linhas e colunas;
- Não existe padronização (organização de linhas e colunas);
- Exemplo de dado não estruturado: uma reclamação feita por um cliente em uma plataforma digital (Reclame Aqui, Facebook etc...);

Slide cedido gentilmente pelo Prof. Wanderley Jr.

DADOS ESTRUTURADOS E NÃO ESTRUTURADOS

Dados NÃO estruturados

- Não existe uma padronização de quais são as informações que o cliente citará e muito menos onde tais informações serão citadas;
- Também não existe uma organização de colunas onde cada dado está muito bem localizado;
- Vídeos, áudios são outros exemplos clássicos de dados não estruturados;

Slide cedido gentilmente pelo Prof. Wanderley Jr.

DADOS ESTRUTURADOS E NÃO ESTRUTURADOS

Dados NÃO estruturados

↘ Redes sociais (Facebook, Instagram, X etc);

- Os comentários feitos nas redes sociais não possuem padronização;
- Cada membro da rede social tem liberdade para postar a informação que deseja e da forma como deseja;

Slide cedido gentilmente pelo Prof. Wanderley Jr.

Origem da Ciência de Dados

- A origem da Ciência de Dados é antiga. Inicialmente era citada como análise de negócios, inteligência competitiva etc...;
- O grande impulso na Ciência de dados se deu devido ao surgimento do Big Data. O Big Data permitiu que a Ciência de Dados mostrasse todo o seu verdadeiro potencial;
- O Big Data permite o armazenamento de grandes **volumes** de dados - para não dizer gigantescos

Slide cedido gentilmente pelo Prof. Wanderley Jr.

Origem da Ciência de Dados

- Big Data possui a **Veracidade** dos dados, ou seja, a verificação dos dados coletados para adequação e relevância ao propósito da análise.
- Os dados presentes no Big Data são verídicos e confiáveis;
- Temos também o **Valor** associado ao Big Data, nos dias atuais temos uma imensidão de dados. Portanto, é importante definir desta imensidão de dados quais são aqueles dados que agregam valor ao negócio ou não.
- 5Vs do Big Data (Volume, Variedade, Velocidade, Veracidade e Valor)

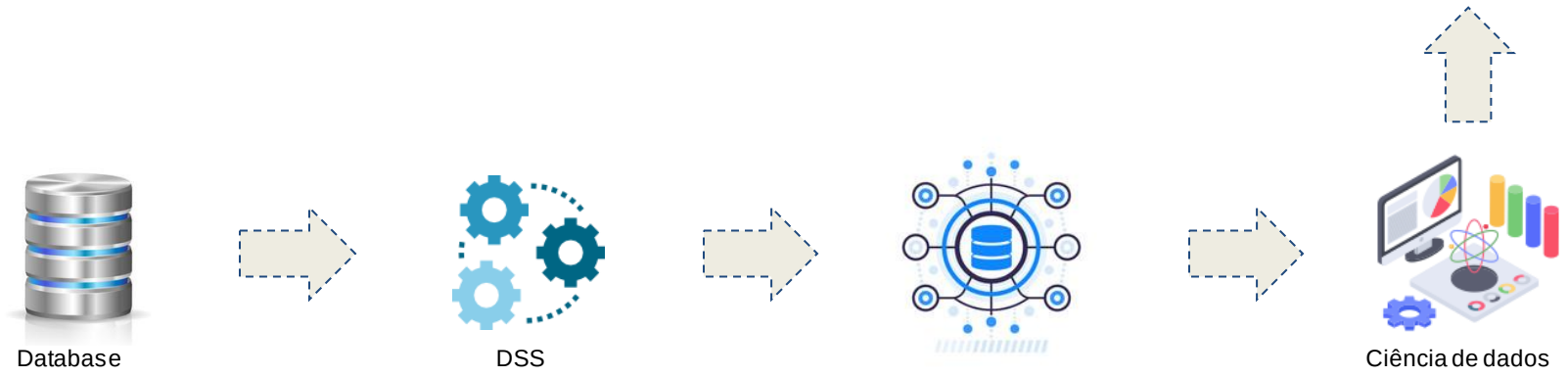
Slide cedido gentilmente pelo Prof. Wanderley Jr.

Tomada de decisão

TOMADA DE DECISÃO

↘ Dados na tomada de decisão

- Os dados para uma tomada de decisão não são algo recente, o próprio termo Business Intelligence remete ao ano de 1865, e os primeiros sistemas voltados à tomada de decisão os DSS (Decision Support System) remetem a década de 70, chegando ao BI moderno na década de 90.

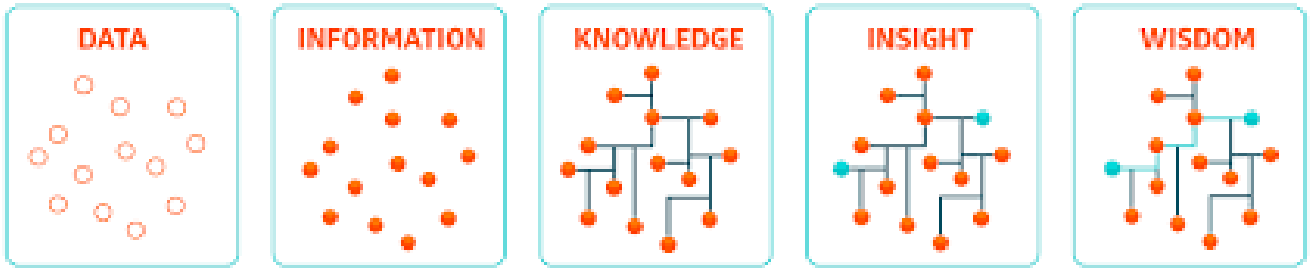


Data Mining

DATA MINING

↘ Data mining – extraindo valor dos dados

- Data mining consiste em processo analítico projetado para explorar conjuntos de dados na busca de padrões consistentes, relacionamento sistemático entre variáveis e testes em um novo conjunto de dados em busca de validação do padrão e/ou relacionamento identificado.



Source: https://hotmart.com/en/blog/what-is-data-science

DATA MINING

↘ Data mining – dados brutos

- Os dados brutos podem ser provenientes de diversos sistemas transacionais. Alguns exemplos são: ERPs, CRMs, Sistemas de fraudes, Sistemas de cobranças, Sistemas de suporte ao cliente, Avaliações de crédito, entre outros.

Source: https://hotmart.com/en/blog/what-is-data-science

DATA MINING

↘ Data mining – informação

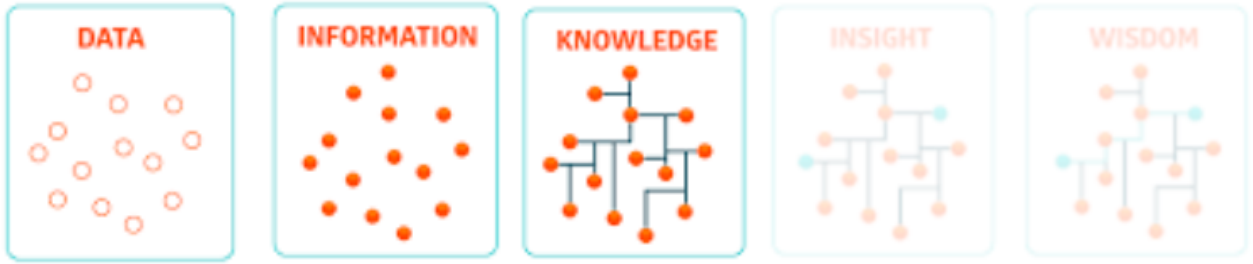
- A informação é o segundo passo, enriquecer os dados brutos com outros dados e obter uma informação mais rica sobre a companhia como um todo.
- Um exemplo pode ser saber que clientes de determinado segmento costumam comprar usando orçamento de opex e não de capex.

Source: https://hotmart.com/en/blog/what-is-data-science

DATA MINING

↘ Data mining – conhecimento

- Neste passo obtemos conhecimento com as bases de dados, isso é possível com o cruzamento de diversas fontes de dados, a informação que tínhamos de produto contratado de um cliente nesta etapa é enriquecida com quantas ligações ele faz para o suporte, qual nota ele deu no NPS, quanto tempo levou para ele fechar o contrato e saber que clientes com determinado produto contratado fazem mais ligações para o suporte por uma necessidade específica do produto.

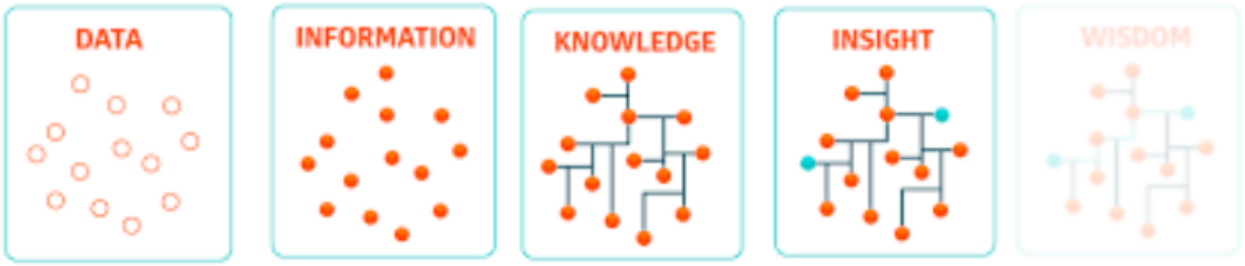


Source: https://hotmart.com/en/blog/what-is-data-science

DATA MINING

↘ Data mining – insights

- Na etapa dos insights, geramos um conhecimento que não é sabido até então, geramos um conhecimento que pode ser chave para algum gatilho na jornada do cliente, ou tratativa que deve ser adicionada no fechamento de contratos novos. Clientes de determinado segmento compram usando opex e costumam pedir muito prazo para pagamento, podemos usar isso para gerar uma fidelidade deste cliente dando mais prazo para pagamento



Source: https://hotmart.com/en/blog/what-is-data-science

DATA MINING

↘ Data mining – wisdom (sabedoria)

- A sabedoria é a etapa onde conhecemos bem nossos dados, seja ele sobre clientes, logs de uma máquina, um processo na empresa e sabemos o que e quando um evento pode ocorrer. Por exemplo: Sei que um cliente que está mês a mês vem caindo seu acesso na minha plataforma de streaming, tem um NPS baixo, possui contratado outros 3 streamings e vem reclamando de preço é um potencial churn

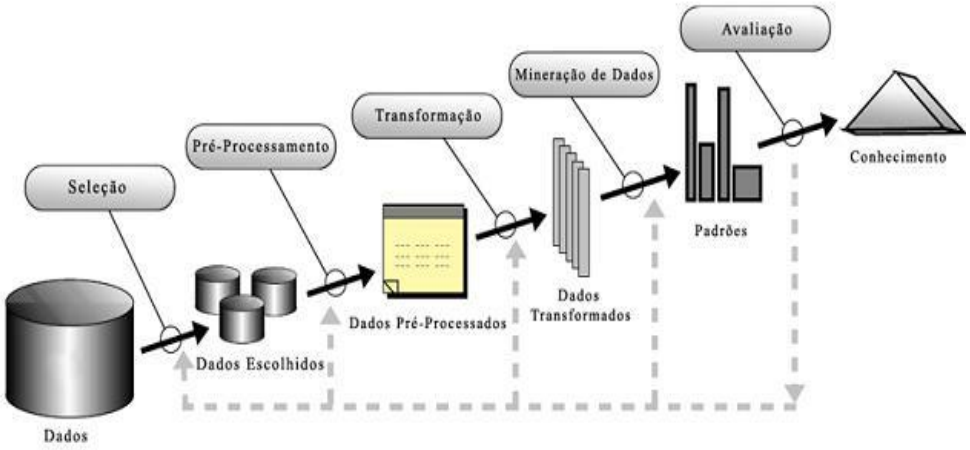
Source: https://hotmart.com/en/blog/what-is-data-science

KDD - Knowledge Discovery in Database

KNOWLEDGE DISCOVERY IN DATABASE (KDD)

↘

- A mineração de dados é parte de um processo conhecido como KDD, Knowledge Discovery in Database.
- Este processo consiste na execução de diversas etapas sequenciais para análise de um conjunto de dados e a cada etapa concluída é gerado insumos para a próxima etapa.

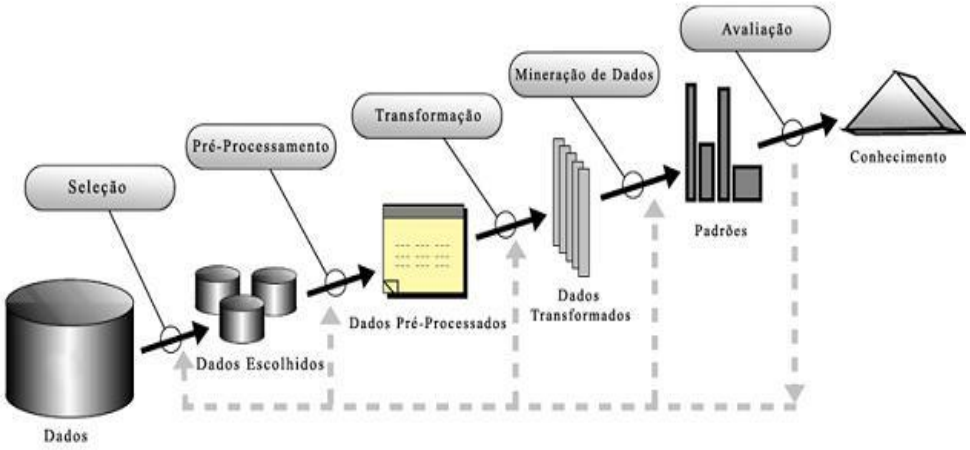


Source: https://www.researchgate.net/figure/Figura-1-Processo-de-KDD-O-processo-de-KDD-consiste-em-uma-sequencia-de-etapas-que-devem_fig1_308995146

KNOWLEDGE DISCOVERY IN DATABASE (KDD)

↘ Etapas - seleção

- Por mais que tenhamos muitos dados, nem todos são necessários ou relevantes para o projeto, por isso a seleção de quais dados serão utilizados é muito importante. Um modelo com muitas variáveis pode se tornar complexo demais de interpretar suas decisões.

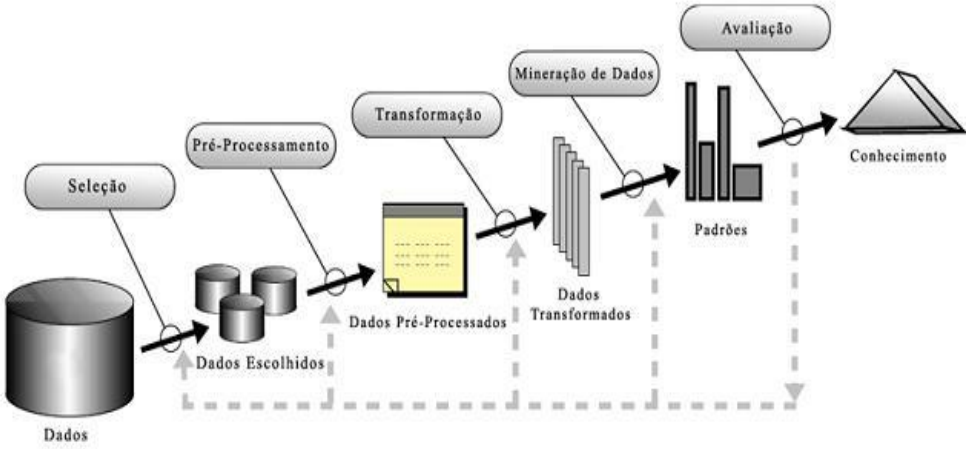


Source: https://www.researchgate.net/figure/Figura-1-Processo-de-KDD-O-processo-de-KDD-consiste-em-uma-sequencia-de-etapas-que-devem_fig1_308995146

KNOWLEDGE DISCOVERY IN DATABASE (KDD)

↘ Etapas – pré-processamento

- Hora de limpar os dados, imputar dados ausentes, corrigir data *types*, deixar os dados prontos para uso.
- Pode parecer simples mas este é um dos maiores desafios do projeto de dados

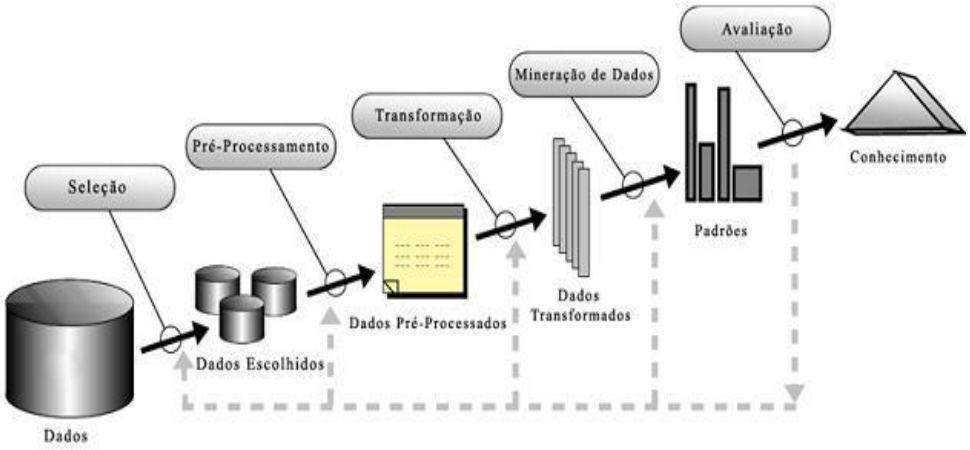


Source: https://www.researchgate.net/figure/Figura-1-Processo-de-KDD-O-processo-de-KDD-consiste-em-uma-sequencia-de-etapas-que-devem_fig1_308995146

KNOWLEDGE DISCOVERY IN DATABASE (KDD)

↘ Etapas - transformação

- Nossos dados vieram de um sistema transacional, isso significa que ele veio no seu formato raw, temos variáveis de textos, outliers, etc. Precisamos criar um cenário que melhor represente o que estamos buscando, variáveis categóricas podem sofrer transformações para variáveis dummies e se tornar mais relevantes ao modelo por exemplo.

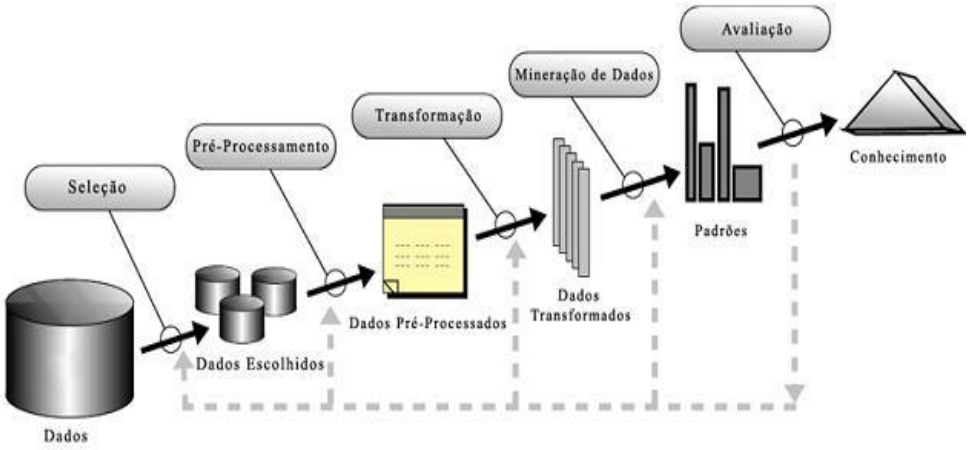


Source: https://www.researchgate.net/figure/Figura-1-Processo-de-KDD-O-processo-de-KDD-consiste-em-uma-sequencia-de-etapas-que-devem_fig1_308995146

KNOWLEDGE DISCOVERY IN DATABASE (KDD)

↘ Etapas – mineração de dados (data mining)

- Etapa onde vamos explorar estes dados a fim de encontrar insights, padrões, formas de resolver um problema de negócio, e coisas que ninguém nunca pensou que existiria nestes dados.

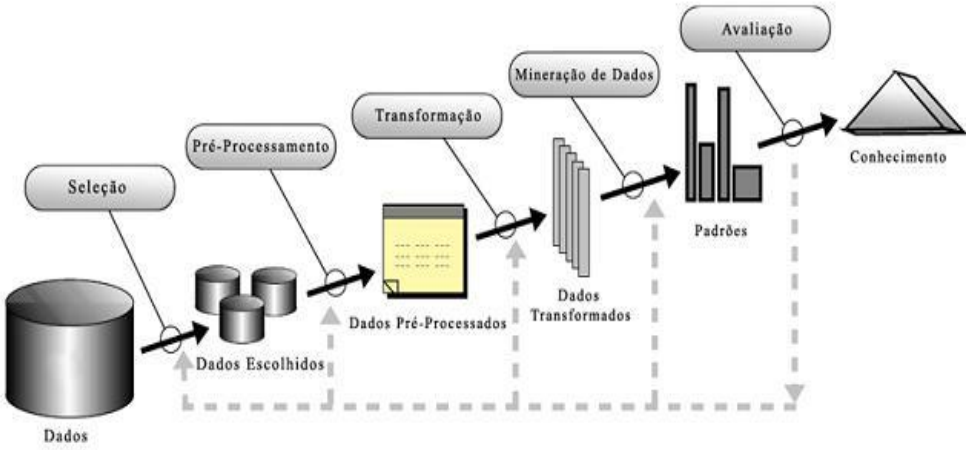


Source: https://www.researchgate.net/figure/Figura-1-Processo-de-KDD-O-processo-de-KDD-consiste-em-uma-sequencia-de-etapas-que-devem_fig1_308995146

KNOWLEDGE DISCOVERY IN DATABASE (KDD)

↘ Etapas – avaliação dos resultados

- Agora é o momento onde avaliamos os resultados obtidos e geramos conhecimento sobre os problemas levantados, ou modelos desenvolvidos.
- Não existe projeto falho, mesmo que esteja avaliando uma hipótese de negócio e chegamos na etapa final ela foi rejeitada, é uma dúvida a menos na tomada de decisão.



Source: https://www.researchgate.net/figure/Figura-1-Processo-de-KDD-O-processo-de-KDD-consiste-em-uma-sequencia-de-etapas-que-devem_fig1_308995146

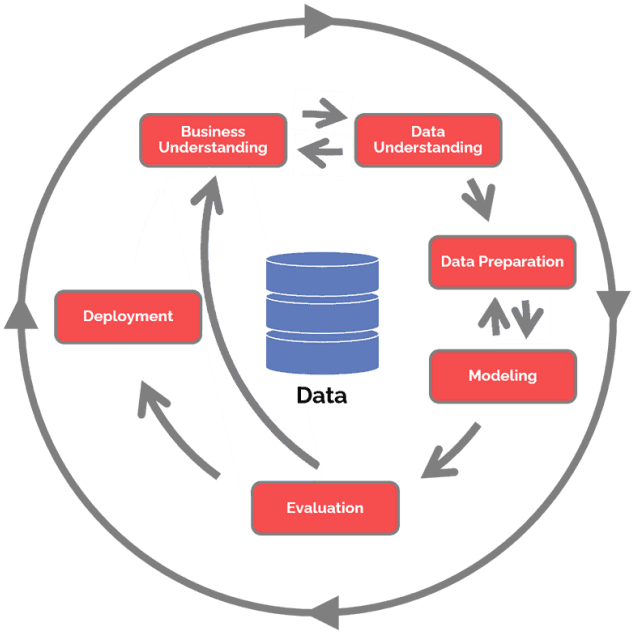
Framework Crisp – DM

(Cross Industry Standard Process for Data Mining)

FRAMEWORK CRISP - DM

↘

- Há um framework que serve como referência na execução de projetos de análise de dados (metodologia). O Crisp-DM é formado por seis etapas essenciais para o sucesso do projeto de dados

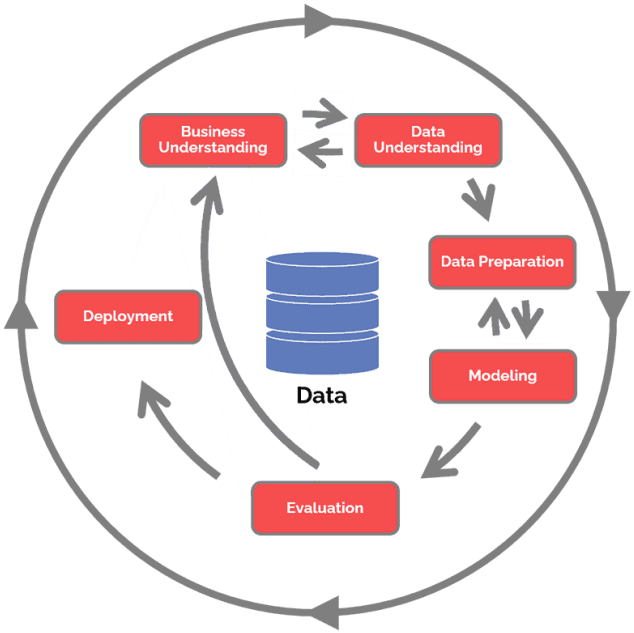


Source: https://www.datascience-pm.com/crisp-dm-2/

FRAMEWORK CRISP - DM

↘ Etapa 1 – entendimento do negócio

- Esta fase consiste em compreender as necessidades, motivações e requisitos do projeto de dados.
- É importante entender claramente o que é um entregável de sucesso para o negócio, definir indicadores que ajudem a validar este objetivo, construa um roadmap de projeto alinhado com o negócio.

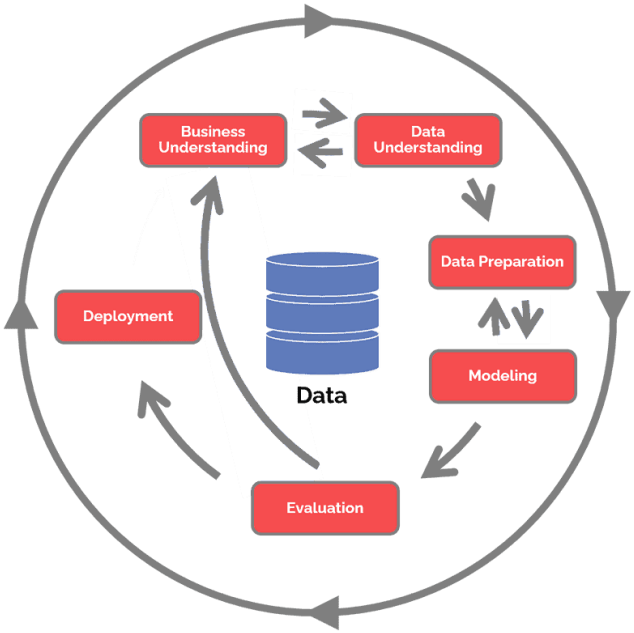


Source: https://www.datascience-pm.com/crisp-dm-2/

FRAMEWORK CRISP - DM

↘ Etapa 2 – entendimento dos dados

- Com a compreensão de negócio feita, agora é preciso entender os dados recebidos se eles satisfazem a necessidade do projeto, seja por uma amostra de dados de período inadequado, quantidade insatisfatória de dados, dados ausentes, os problemas nesta fase podem ser muitos.

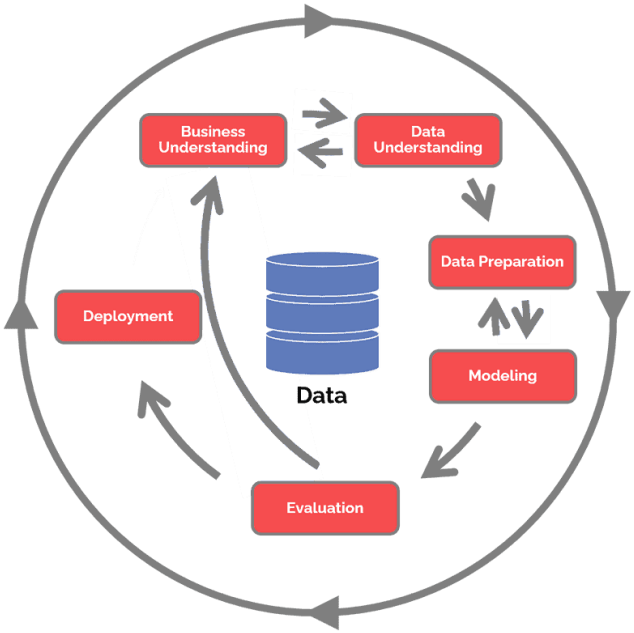


Source: https://www.datascience-pm.com/crisp-dm-2/

FRAMEWORK CRISP - DM

↘ Etapa 3 – preparação dos dados

- Problema nos dados é algo muito comum, preparar os dados pode ser o processo que mais leva tempo em um projeto, como tratar dados ausentes, outliers, criação de novas features que representem melhor o cenário do projeto para o modelo, todos estes pontos demandam muito tempo em um projeto

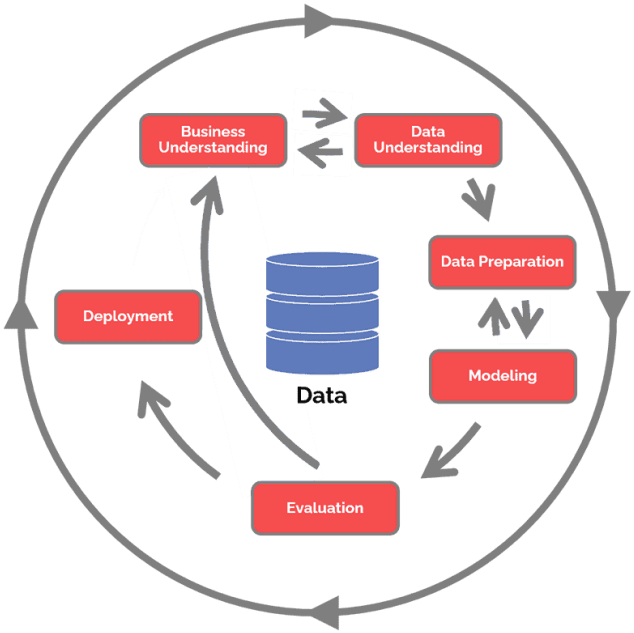


Source: https://www.datascience-pm.com/crisp-dm-2/

FRAMEWORK CRISP - DM

↘ Etapa 4 – modelagem dos dados

- A modelagem é o processo de desenvolver o modelo matemático que vai nos ajudar a responder as perguntas de negócio, pode ser um modelo que responda o melhor preço para um cliente, o melhor produto para determinado cliente, qual melhor região para fazer um piloto com um novo produto até a probabilidade de um cliente deixar de ser cliente no próximo mês.

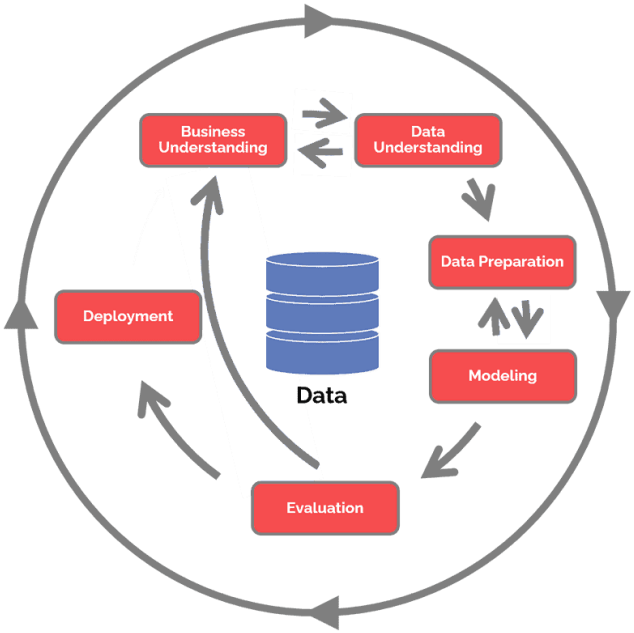


Source: https://www.datascience-pm.com/crisp-dm-2/

FRAMEWORK CRISP - DM

↘ Etapa 5 – avaliação dos resultados

- Geralmente mantemos uma amostra de dados de validação para validar nosso modelo da etapa anterior e analisar se o comportamento analisado/compreendido na etapa anterior se mantém em uma outra amostra de dados. Isso garante ou não que nosso modelo está pronto para deploy.

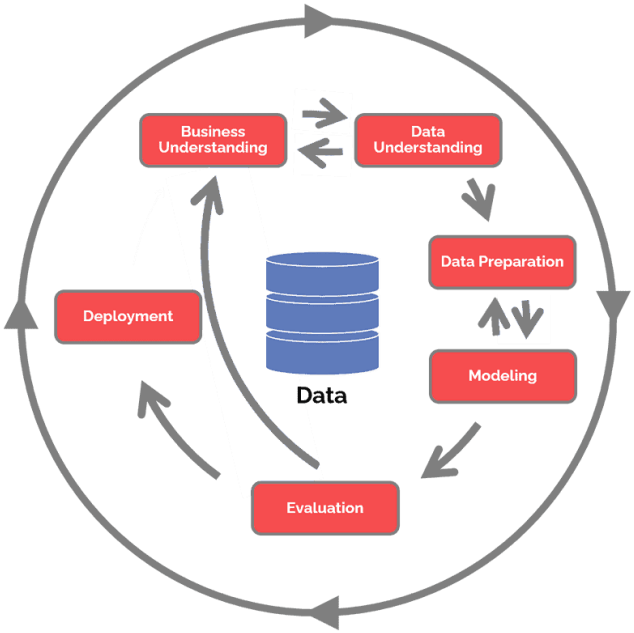


Source: https://www.datascience-pm.com/crisp-dm-2/

FRAMEWORK CRISP - DM

↘ Etapa 6 – deploy

- Com o modelo validado e com resultados aceitáveis para ser entregue em ambiente produtivo chegamos a etapa 6 onde o modelo de fato vai para produção



Source: https://www.datascience-pm.com/crisp-dm-2/

PROJETOS UTILIZANDO CIÊNCIA DE DADOS - SITUAÇÕES REAIS

↘ **Projetos**

- Previsão de vendas de roupas femininas
- Probabilidade de clicks em anúncios na mídia
- *Chatbot* para ONG
- Geração de dashboard com indicadores de acompanhamento e insights para empresa de depilação
- *Process Mining* para empresa de fornecimento de gás
- Cesta de ofertas ou sistema de recomendação para empresa de ferragens
- Clusterização de contratos e clientes para empresa de depilação
- Elasticidade de preços para empresa de software
- Previsão de acidentes ferroviários para empresa de transporte ferroviário
- Range de preços para softwares

↘

Obrigada!

Até a próxima aula ☺