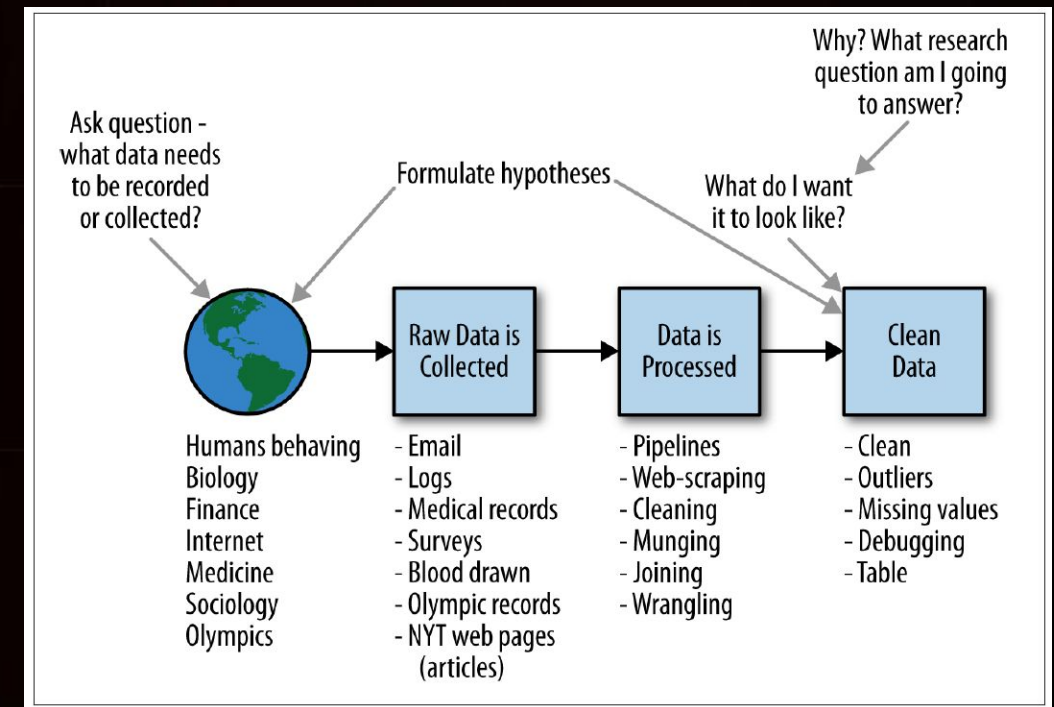
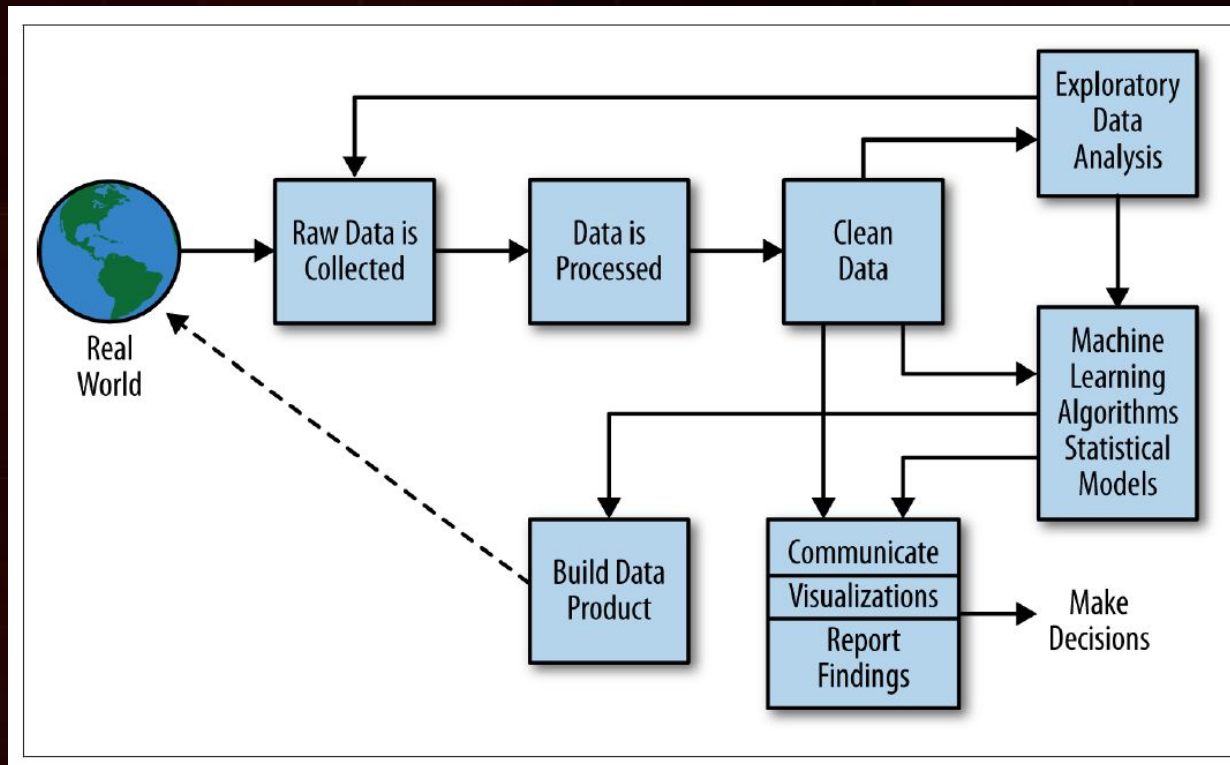


# **INTRODUÇÃO À CIÊNCIA DE DADOS**

**KDD e Análise de Dados**

# PROCESSO DS

- Compreensão do todo! Mas vamos ver primeiro a análise exploratória para depois estudar processamento dos dados.



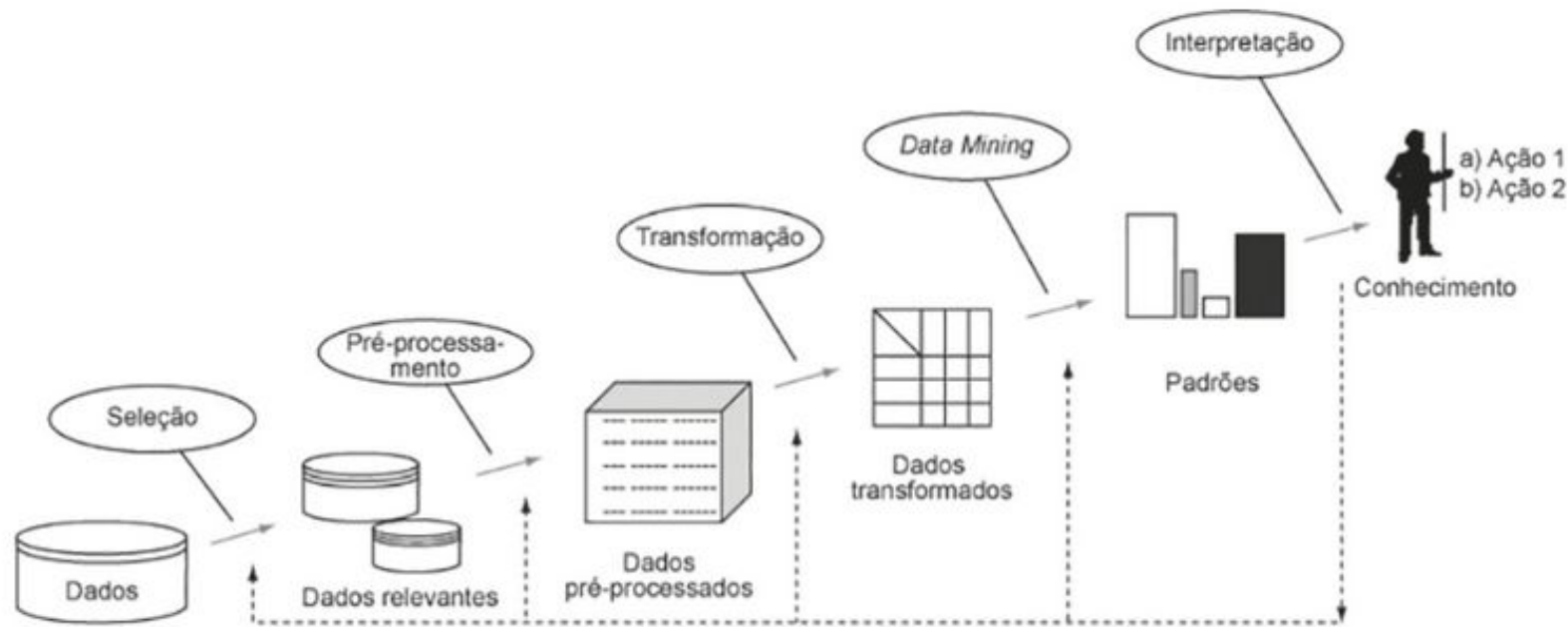
Fonte: Shutt, R. and O'Neil, C.; Doing Data Science, 2014

# KDD (Knowledge Discovery in Databases)

- KDD ou "Processo de Descoberta de Conhecimento", segundo Fayyad, Piatetsky-Shapiro e Smyth, é um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de conjuntos de dados.
- A característica “não trivial” diz respeito à complexidade existente na execução e manutenção dos processos de KDD; “interativo” representa a relevância de possuir um elemento que controle o processo; “iterativo” indica a possibilidade de repetições em qualquer uma das etapas do processo; e “conhecimento útil” aponta para a indicação de que o objetivo foi alcançado.

# KDD (Knowledge Discovery in Databases)

- Geralmente é dividido em 5 fases: Seleção, Pré-Processamento, Transformação, Data Mining e Interpretação.



**Figura 3.** Etapas do KDD.

*Fonte:* Fayyad, Piatetsky-Shapiro e Smyth (1996 apud BRITO, 2012, documento on-line).

# KDD (Knowledge Discovery in Databases)

- 5 Fases:
  - **Seleção:** consiste em selecionar um conjunto ou subconjunto de dados que farão parte da análise. As fontes de dados podem ser variadas (planilhas, bancos de dados, data warehouses) e possuir dados com formatos diferentes (estruturados, semiestruturados e não-estruturados).
  - **Pré-Processamento:** consiste em fazer a verificação da qualidade dos dados. Exceções e ruídos são removidos. Limpeza, correção, remoção de dados inconsistentes, identificação de dados ausentes, incompletos ou não íntegros são parte do processo.

# KDD (Knowledge Discovery in Databases)

- 5 Fases:
  - **Transformação:** consiste em aplicar técnicas de transformação como: normalização, agregação, criação de novos atributos, redução e sintetização dos dados. Busca-se identificar atributos úteis nos dados para alcançar os objetivos pretendidos.
  - **Mineração de Dados:** consiste na aplicação de algoritmos e técnicas para identificar padrões nos dados e verificar hipóteses. Geralmente as descobertas podem ser descritivas ou preditivas, com os seguintes objetivos: regressão (uma função que faça o mapeamento dos dados), clusterização (identificar um conjunto finito de categorias ou clusters), sumarização (buscar descrição compacta para subconjunto dos dados), dependências ou associações (identificar dependências significativas entre as variáveis) e divergências (identificar alterações significativas a partir dos valores medidos).



# KDD (Knowledge Discovery in Databases)

- 5 Fases:
  - **Interpretação:** consiste em avaliar o desempenho do modelo, ocorrendo a consolidação do conhecimento descoberto. A validação pode ser feita baseada em análise de profissionais ou mesmo em comparação com dados coletados anteriormente.

# AED (Análise Exploratória de Dados)

- A Análise Exploratória de Dados é um termo bastante usado por profissionais de DS. A AED tem como finalidade principal examinar os dados previamente à aplicação de qualquer técnica estatística. Desta forma o analista consegue um entendimento básico de seus dados e das relações existentes entre as variáveis analisadas.
- Na AED é muito comum a análise descritiva, que de forma detalhada permite ao cientista de dados familiarizar-se com os dados, organizá-los e sintetizá-los de forma a obter as informações necessárias do conjunto de dados para responder as questões que o problema de DS está tentando resolver.
- A AED pode ser comparada com as três primeiras fases do KDD, e pode ser entendida como a primeira e importantíssima observação sobre os dados!



# AED (Análise Exploratória de Dados)

- Para realizar a análise exploratória é determinante conhecer tecnicamente o que seus dados representam e como eles são classificados. Veja a tabela apresentada:

Tabela 2.1    Conjunto de dados <code>hospital</code> com seus atributos									
Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente

# QUANTIDADE DE VARIÁVEIS

- Dados **unidimensionais** (univariados) são dados nos quais você tem apenas uma coleção de números, por exemplo a temperatura de pessoas de uma ala (como na tabela do slide anterior), a quantidade de gols que seu time fez por jogo no campeonato ou a média de minutos diários que você usa olhando seu instagram.
- Um primeiro passo inevitável é computar alguma estatística, saber o dia que gastou mais minutos no instagram, que gastou menos, a média de minutos, a soma deles.

# QUANTIDADE DE VARIÁVEIS

- Agora imagine que seus dados possam ter mais de uma dimensão (**mutidimensionais** ou multivariados), por exemplo o sexo, a idade, o peso (como na tabela já vista), os gols feitos e os gols sofridos pelo seu time, ou então a quantidade de minutos no instagram e também a quantidade de posts realizados.
- Em muitos casos é importante conhecer cada dimensão individualmente, mas também é necessário dispersar os dados e entender a relação entre eles, se elas existirem.

# TIPO

O tipo define se o atributo representa quantidades, sendo então chamado de **quantitativo ou numérico**, ou qualidades, quando é chamado de **qualitativo, simbólico ou categórico**, pois os valores podem ser associados às categorias.

Dados qualitativos: { pequeno, médio, grande}. Eles podem ter seus valores ordenados, mas nunca podem receber operações aritméticas.

Dados quantitativos são valores numéricos, que podem ser ordenados e usados em operações aritméticas.

Obs: Às vezes vai escutar que um dado é **escalar** (dado único), que representa um dado que não é um array (vetor ou matriz) ou objeto (dict)!

# QUANTITATIVOS

As variáveis quantitativas podem ser **contínuas** ou **discretas**:

- Variáveis discretas: normalmente são representadas por valores que contêm um número finito ou infinito contável de valores. Casos de atributos contáveis são valores (0/1), idade, número de peças com defeito.
- Variáveis contínuas: normalmente são representadas por valores que podem assumir um número infinito de valores. Geralmente resultados de medidas (por instrumento). Atributos que representam peso, tamanho, distância.

# QUANTITATIVOS

As variáveis quantitativas também podem ser categorizadas como intervalar (valores dentro de um intervalo, sem zero absoluto) e racional (com zero absoluto).

Exemplos são:

- Intervalar: temperatura, datas de um calendário.
- Racional: quantidade de vezes que uma pessoa foi ao hospital (o zero é parâmetro)



# QUALITATIVOS

As variáveis qualitativas podem ser **nominais** ou **ordinais**:

- Variáveis nominais: os valores são nomes diferentes, carregando a menor quantidade de informação possível. Não existe relação de ordem entre seus valores. Exemplos como CPF, RG, cor dos olhos, sexo.
- Variáveis ordinais: os valores refletem uma ordem das categorias representadas, desta forma operadores de comparação (maior, menor) podem ser utilizados. Exemplos: escolaridade, patente militar, classificação no campeonato.

# EXPLORAÇÃO

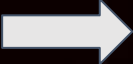
- Uma grande quantidade de informações úteis pode ser extraída a partir do conhecimento sobre tipos de dados e, principalmente, sobre a exploração de um conjunto de dados.
- A estatística descritiva resume de forma quantitativa as principais características de um conjunto de dados.

# DICA!

- Geralmente, partir de um tipo de análise quando se tem o conhecimento sobre o tipo de dado é algo muito interessante para se identificar características dos dados.
- Na tabela a seguir apresentamos uma sugestão de possíveis representações para cada tipo de dados.

Escala	Representação	Medida de Tendência Central
Nominal	Diagramas de barras, linhas e pizza	Moda
Ordinal	Boxplot	Mediana
Intervalar	Histograma e polígonos de frequência	Média
Racional		Média Geométrica

# DICA 2!

- Vamos demonstrar rapidamente algumas ideias sobre análise exploratória.
- Demonstração 

# FINALIZANDO

- KDD (que ainda seguiremos compreendendo nas próximas aulas) e AED são processos fundamentais para que o cientista de dados seja capaz de compreender o conjunto de dados que estão disponíveis e, principalmente, ser capaz de determinar os tipos de análises que são possíveis.
- É por meio da Análise Exploratória que se compreende o caminho para responder às questões de um projeto de DS.

# REFERÊNCIAS

- Grande parte do conteúdo dessa aula é baseado no livro:

CARVALHO, André Carlos Ponce de Leon Ferreira et al.  
*Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina*. Disponível em: Minha Biblioteca, (2nd edição).  
Grupo GEN, 2021.



# **INTRODUÇÃO À CIÊNCIA DE DADOS**

**KDD e Análise de Dados**