Análise Exploratória de Dados

Noções Gerais sobre o Fluxo de Data Science

João Pedro Albino

Departamento de Computação / Faculdade de Ciências

PPG-MiT / Faculdade de Artes, Arquitetura, Comunicação e Design

Introdução

- Análise exploratória de dados (EDA – Exploratory Data Analysis)
    - utilizada por cientistas de dados para **analisar e investigar conjuntos de dados** (datasets)
    - resumir as principais características de um **dataset** empregando métodos de visualização de dados.
- EDA: técnica definida pelo estatístico americano John Tukey

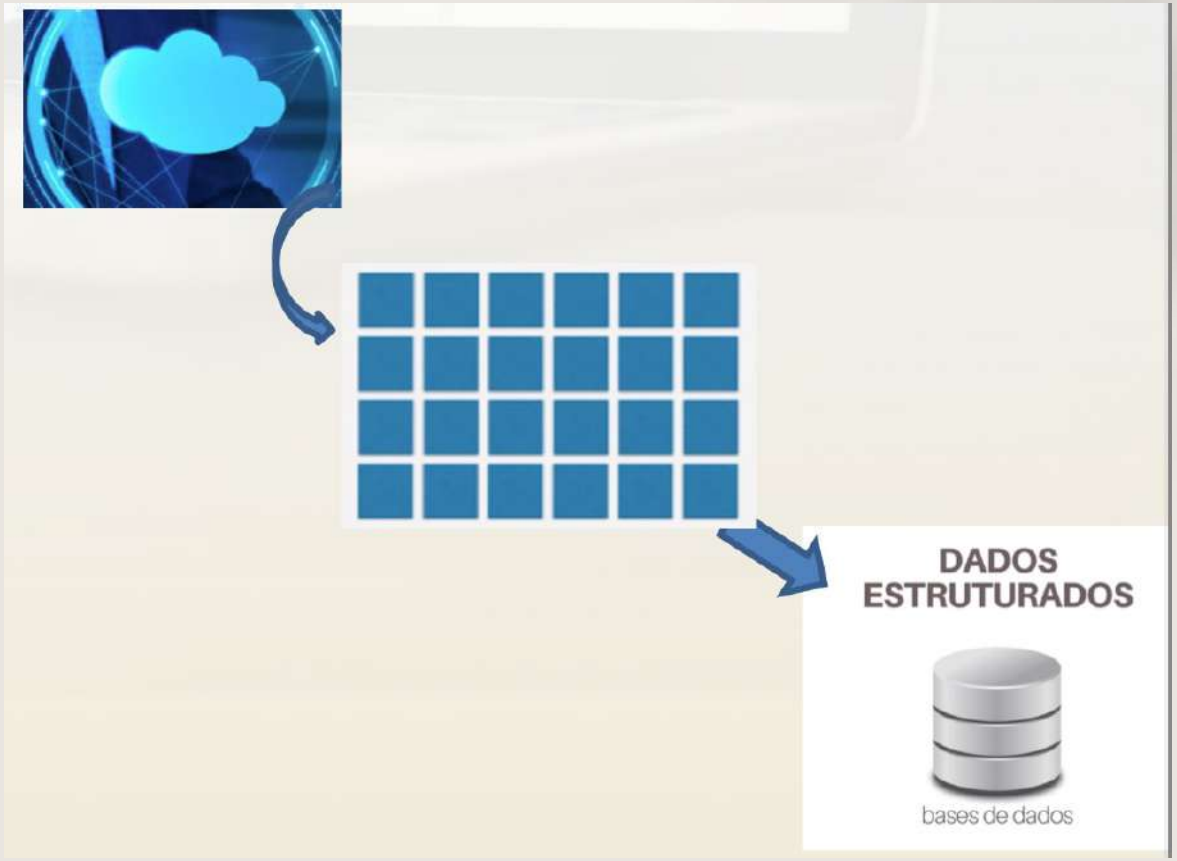
https://www.ibm.com/cloud/learn/exploratory-data-analysis

Introdução

- Principais objetivos da AED - Análise Exploratória de Dados
  - ajudar a analisar os dados antes de fazer quaisquer suposições
  - ajudar a identificar erros óbvios
  - detectar **outliers** ou eventos anômalos
  - compreender melhor os padrões dentro dos dados
  - encontrar relações interessantes entre as variáveis

Processo básico

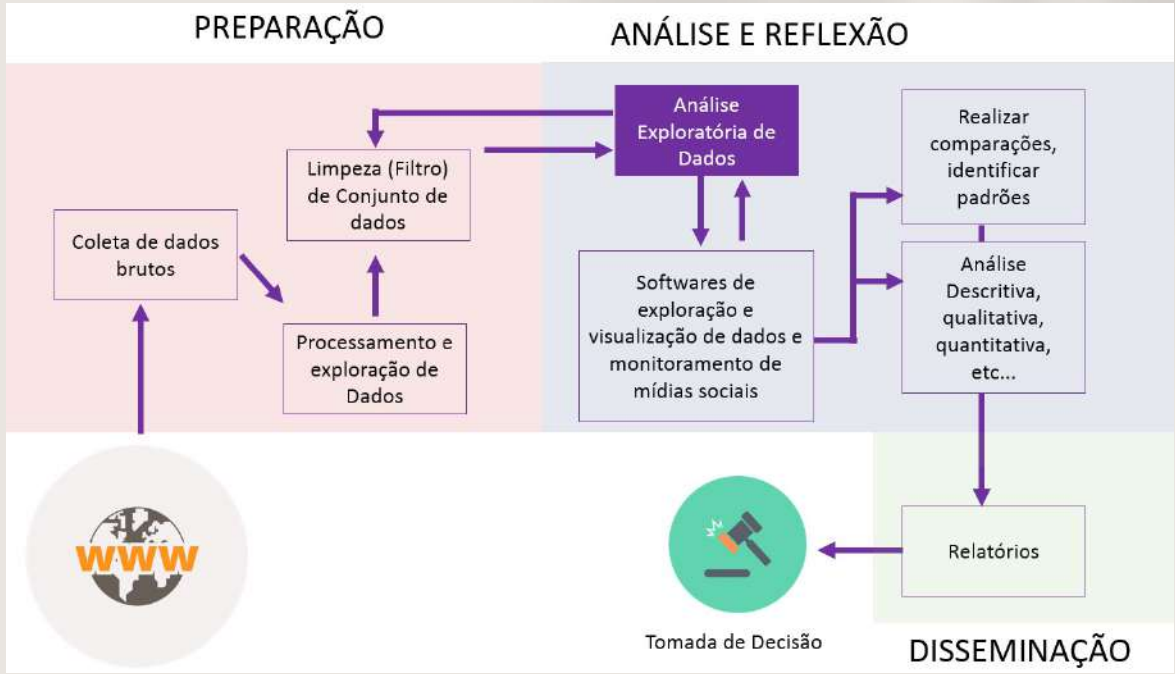
- Após a coleta e a digitação de dados em um banco de dados apropriado, o próximo passo é a análise descritiva.
- A **análise descritiva** detalhada permite
  - familiarizar-se com os dados
  - organizá-los
  - sintetizá-los para obter as informações necessárias do dataset
- **Responder as questões que estão sendo estudadas**.



Etapas da AED

- Preparar os dados para serem acessíveis a qualquer técnica estatística;
- Realizar um **exame gráfico** da natureza das **variáveis individuais** a analisar e **uma análise descritiva** que permita quantificar alguns aspectos gráficos dos dados;
- Realizar um **exame gráfico** das **relações entre as variáveis analisadas** e uma **análise descritiva** que **quantifique o grau de inter-relação** entre elas;
- Identificar os possíveis **casos atípicos** (**outliers**);
- Avaliar, se for necesário, a **presença de dados ausentes** (**missing data**);
- Avaliar, se for necesário, algumas suposições básicas, como **normalidade**, **linearidade** e **homocedasticidade**.

http://www.each.usp.br/lauretto/SIN5008_2011/aula01/aula1/



https://www.ibpad.com.br/aula/importancia-da-analise-exploratoria-de-dados-mms-2ed/

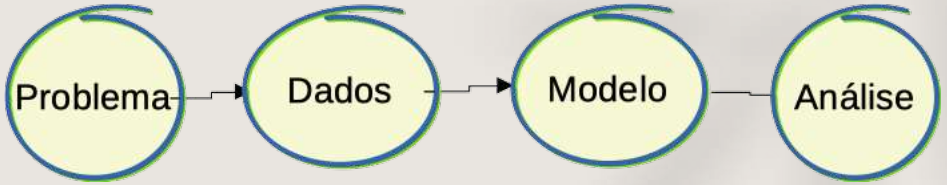
Características da AED

- AED extrai informações de um dataset sem suposições de um modelo.
- As técnicas gráficas desempenham um importante papel em AED.
  - Obter insights sobre os dados!
  - Inclusive sobre distribuição, tendência central, extensão (escopo), modalidade e outliers (valores que fogem à normalidade)

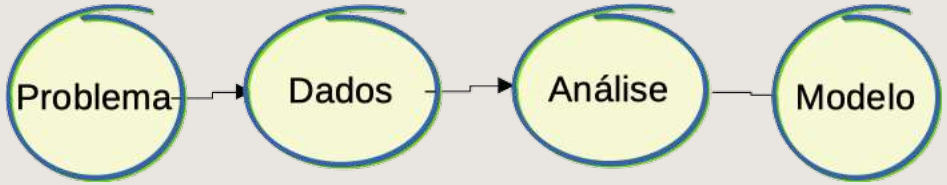
Abordagem dos dados em AED

- Na Análise Exploratória de Dados não há a imposição de um modelo aos dados
- Mineração nos dados para *eventualmente* indicar qual modelo

Estatística Clássica



Análise Exploratória de Dados



Tópicos de Estatística Básica

Definição básica de estatística

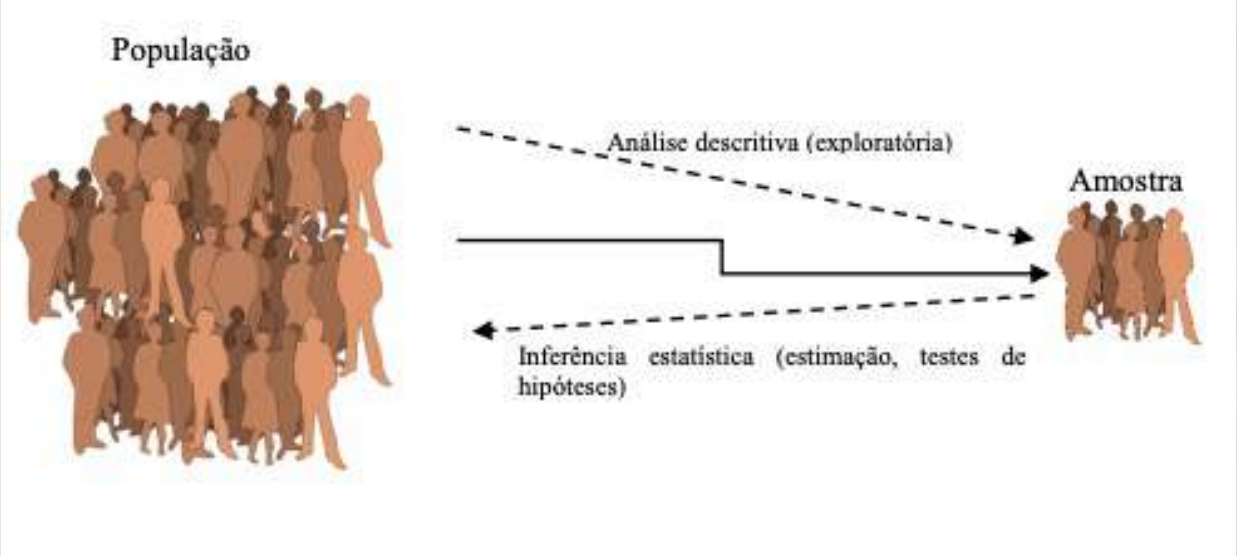
Conjunto de técnicas para, sistematicamente:

- Planejar a coleta de dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento.
- Descrever, analisar e interpretar dados
- Extrair informações para subsidiar decisões ou conclusões

Conjunto de técnicas úteis para a tomada de decisão sobre um processo ou população, baseada na análise da informação contida em uma amostra desta população.

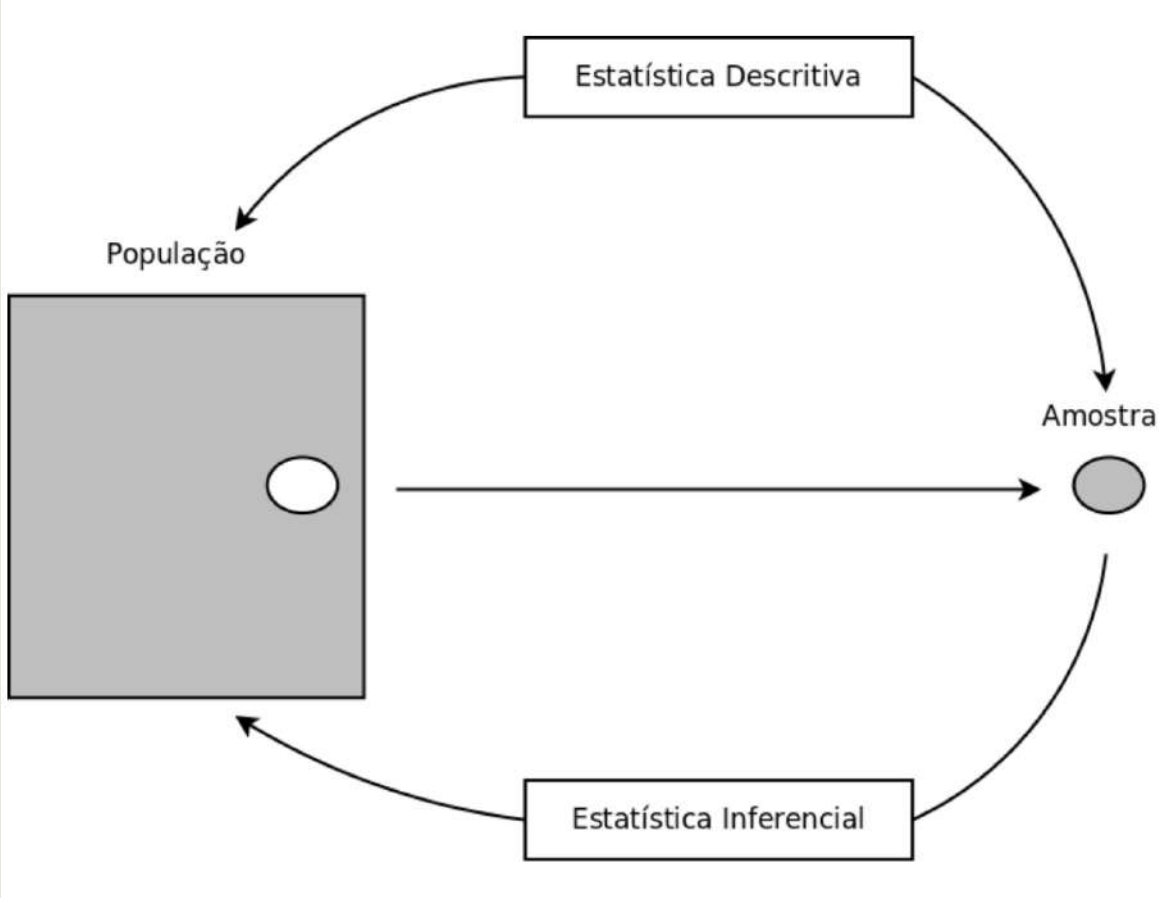
Pensamento estatístico

- Raciocínio analítico que enfoca a onipresença da variação;
  - Investigação das fontes de variação;
  - Planejamento de coleta de dados com a variação em mente;
  - Quantificação da variação;
  - Explicação da variação.



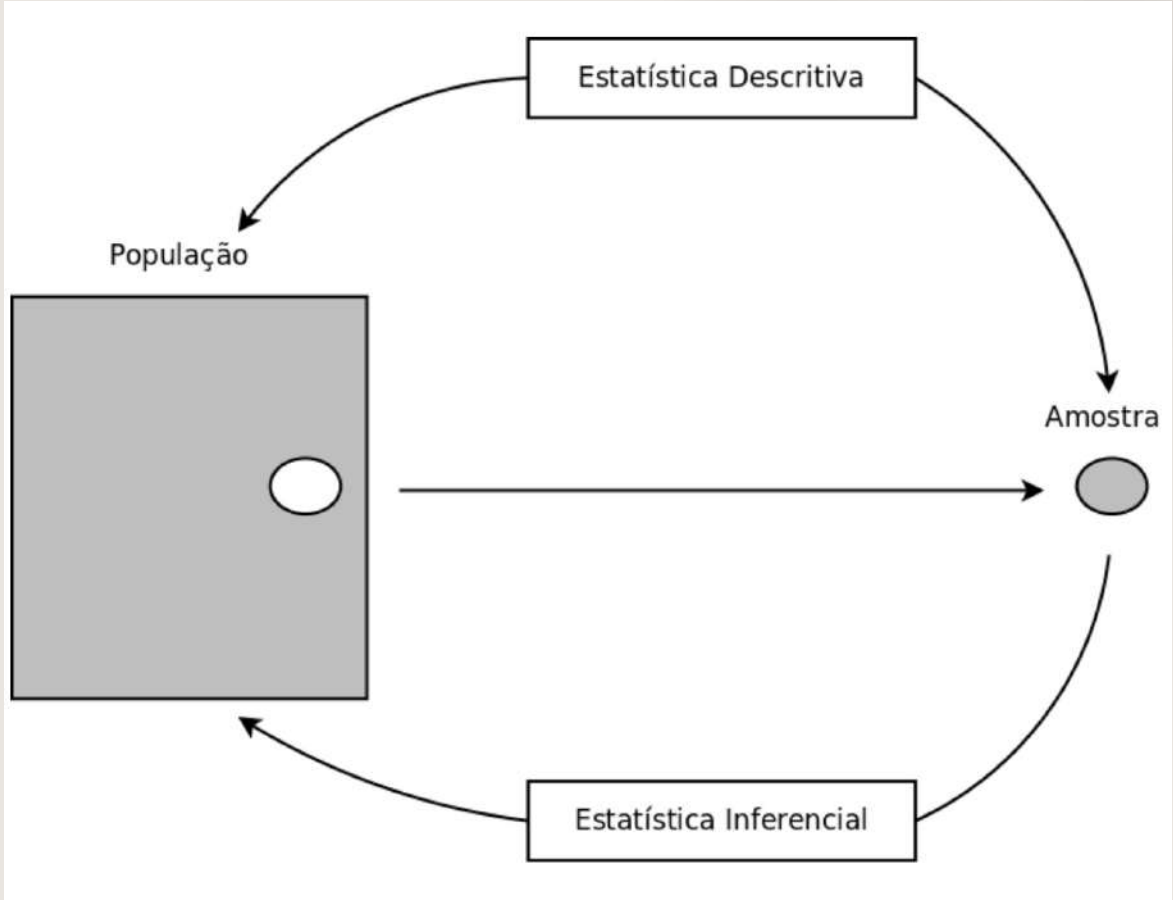
Conceitos essenciais de estatística básica

- Estatística descritiva
  - Ramo da estatística que aplica várias *técnicas* para **descrever e sumarizar** um conjunto de dados (dataset).
  - A disponibilidade de uma grande quantidade de dados e de métodos computacionais muito eficientes revigorou está área da estatística
  - Fornece resumos simples sobre a amostra e sobre as observações que foram feitas.
  - O resumo pode ser quantitativo ou visual.
- Recentemente coleção de técnicas de resumos ➡ análise exploratória de dados



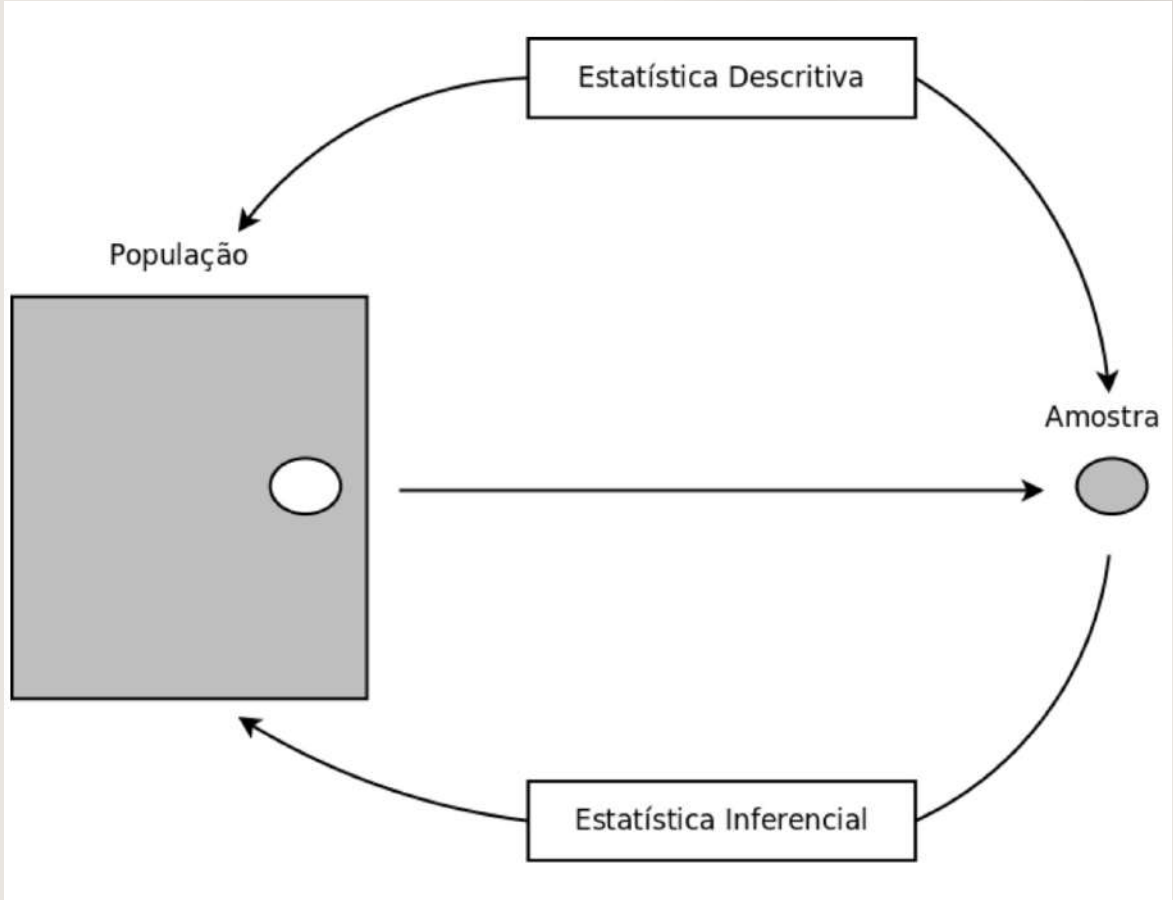
Conceitos essenciais
Estatística descritiva

- Análise univariada
  - Descreve a distribuição de uma **única variável**
    - **medida central** da variável:
      - *Média*
      - *mediana*
      - *moda*
    - **dispersão** da variável:
      - **diferença** entre **maior** e **menor** valor
      - **quantil/quartil**
      - **Variância**
      - **desvio padrão**



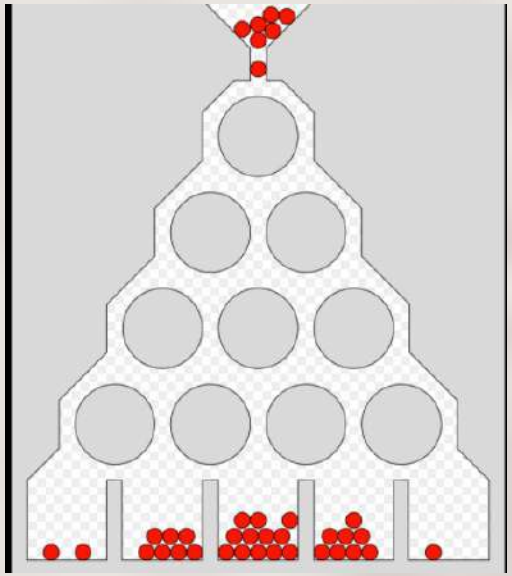
Conceitos essenciais
Estatística descritiva

- Análise bivariada
  - Quando uma amostra consiste de **mais de uma variável**
  - Não é só análise descritiva simples, mas também *o relacionamento entre duas variáveis diferentes*
  - Medidas quantitativas de **dependência** incluem:
    - **Correlação**
      - **Pearson**: quando ambas variáveis são contínuas
      - **Spearman**: quando as variáveis são descontínuas
    - **covariância**



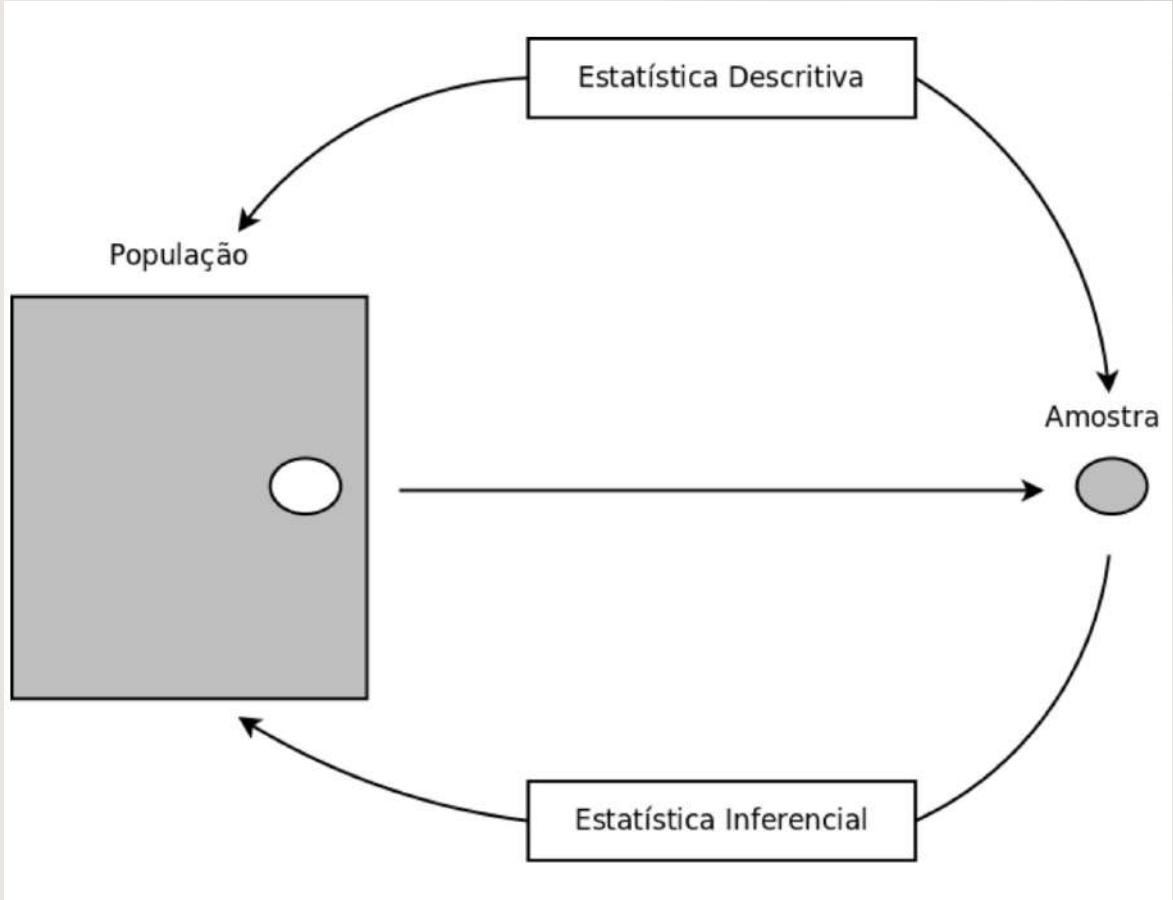
Conceitos essenciais Probabilidade

- Área da Matemática que calcula as chances de um evento ocorrer, em um determinado contexto
- Serve para obter estimativas matemáticas da possibilidade de certos eventos acontecerem ao acaso.



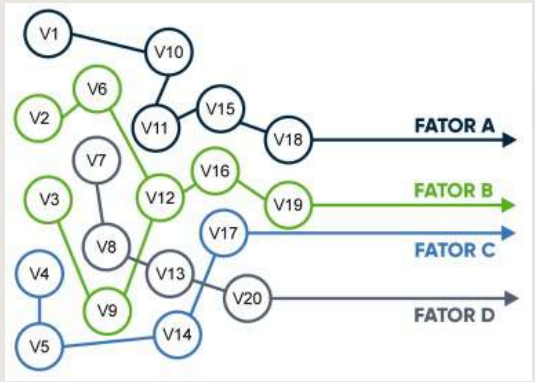
Conceitos essenciais Estatística Inferencial

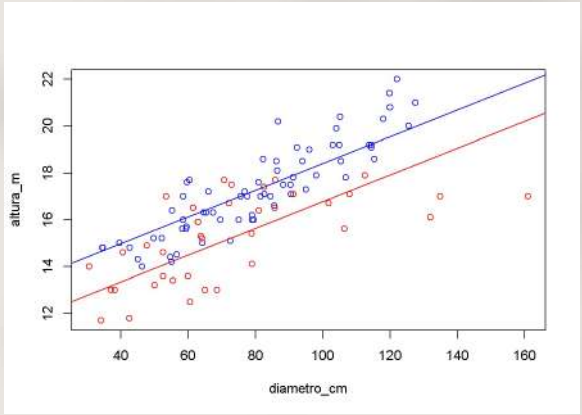
- Processo pelo qual estatísticos tiram conclusões acerca da **população** usando informação de uma **amostra**
- Assume-se que a população é muito maior do que o dataset utilizado (amostra)
- Sempre vem acompanhada de uma medida de precisão sobre sua veracidade
- Principais tipos de inferência
  - Frequencista (ou clássica)
    - enfatiza a frequência ou proporção dos dados
    - metodologias: testes de hipóteses e intervalos de confiança
  - Inferência bayesiana
    - avaliação de hipóteses pela máxima verossimilhança (MLE - maximum-likelihood estimation)
    - utilizada em métodos computacionais relacionados à IA, mineração de dados, linguística

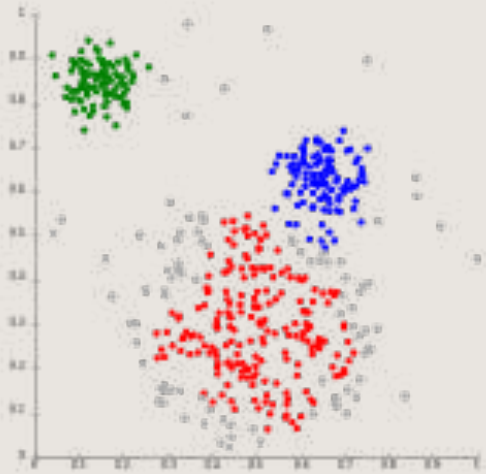


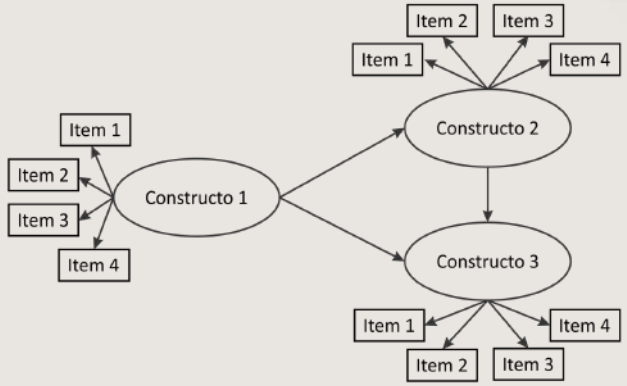
Estendendo conceitos estatísticos
Análise Multivariada

- Conjunto de métodos estatísticos utilizados em situações em que várias variáveis são medidas simultaneamente em cada elemento amostral
- As variáveis são correlacionadas entre si e quanto maior o número de variáveis, mais complexa torna-se a análise por métodos comuns
- Utilizada com o propósito de simplificar ou facilitar a interpretação do fenômeno que está sendo estudado.
- Métodos mais utilizados:
  - Análise de correspondência
  - **Análise de componentes principais**
  - **Análise fatorial**
  - **Análise de cluster**
  - **Análise de regressão múltipla**
  - **Modelagem de equações estruturais**

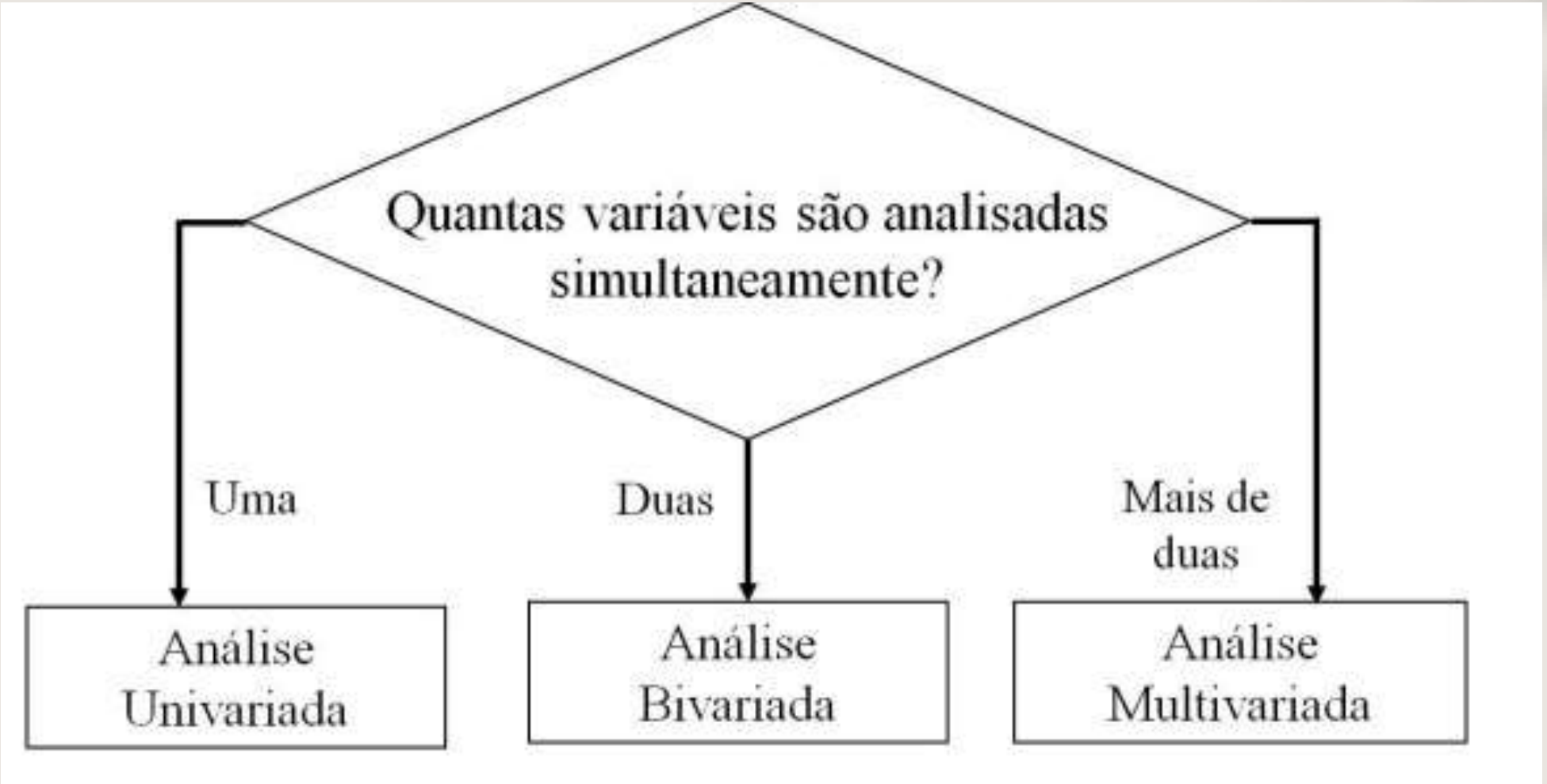




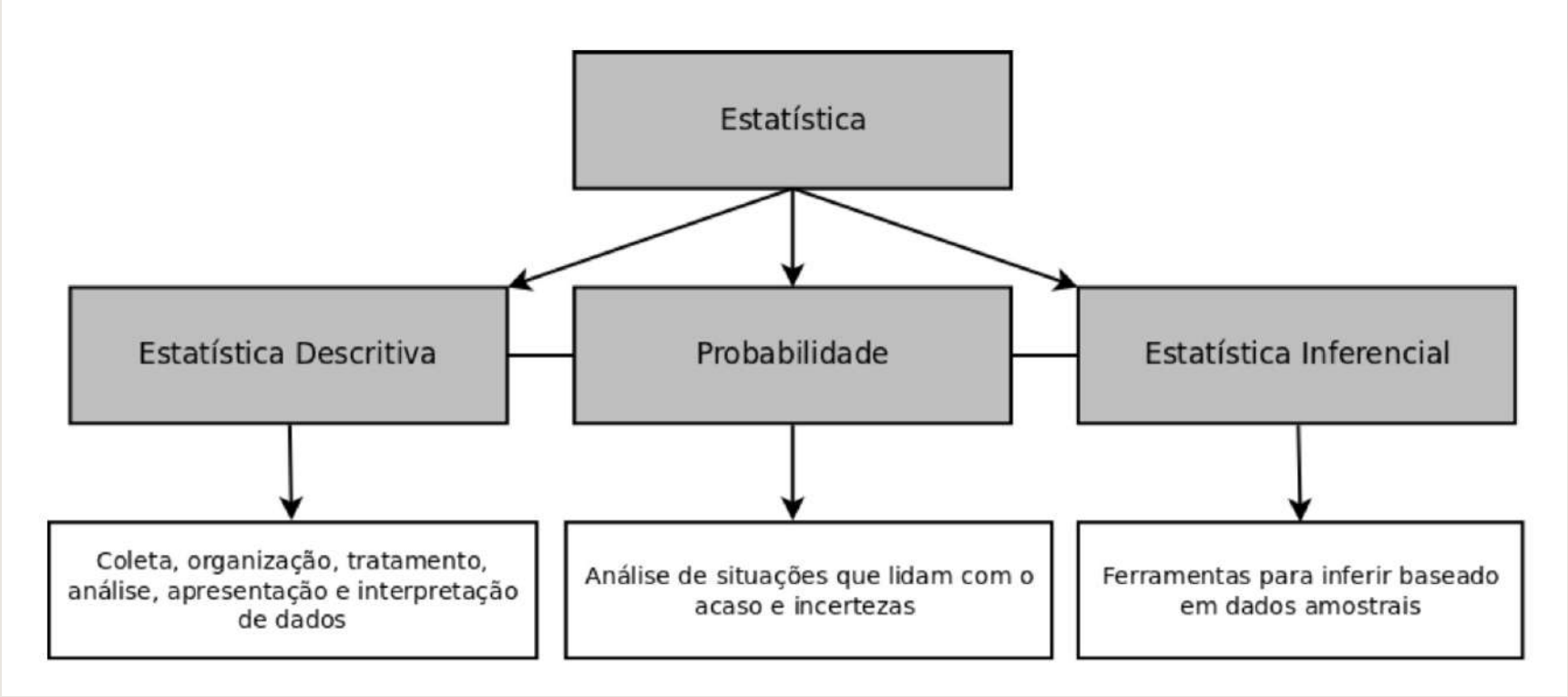




Resumo


Quantas variáveis são analisadas simultaneamente?
Uma
Duas
Mais de duas
Análise Univariada
Análise Bivariada
Análise Multivariada

Resumo:
Etapas da Análise Estatística



Escalas de Mensuração

As técnicas a serem utilizadas na análise dependem da natureza de mensuração das variáveis de interesse

Nominal: as variáveis são medidas em classes discretas, mas não é possível estabelecer ordem.
Ordinal: as variáveis são medidas em classes discretas entre as quais é possível definir uma ordem, segundo uma relação descritível mas não quantificável.
Intervalar: as variáveis assumem valores quantitativos, não possuem zero absoluto, i.e. não possuem uma medida de ausência de atributo.
Razão: as variáveis assumem valores quantitativos, cuja relação exata entre estes é possível definir porque esta escala possui um zero absoluto.

Tipos de Dados Estruturados

- Estrutura rígida/fixa
- Previamente planejados
- Volume pequeno
- Banco de Dados

Nome	CPF	Endereço	Telefone
Marcela Freitas	11111	Rua A, nº 1	101010
João Augusto	22222	Rua B, nº 2	202020
Pablo Silva	33333	Rua C, nº 3	303030
André Mendes	44444	Rua D, nº 4	404040
Juliana Freitas	55555	Rua E, nº 5	505050

Dados Estruturados – Tabela

Tipos de Dados Semi Estruturados

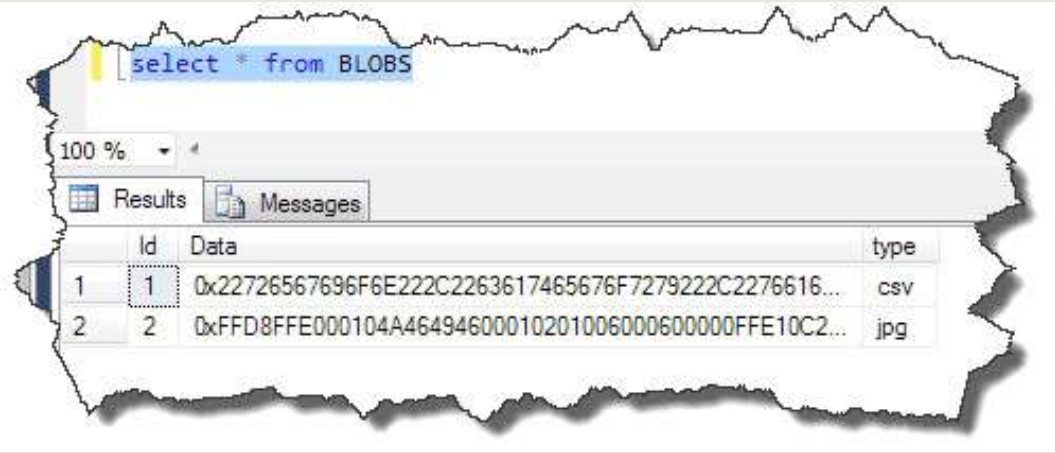
- Dados em que uma parte tem estrutura, outra não
- O esquema da parte estruturada está contido junto como dado
- Exemplo
  - E-mail
    - Estrurada: <to>, <subject>, <date>, <cc>
    - Não estrurada: o corpo da mensagem (texto, imagem, etc.)

A Quick Look At Kappa And Lambda Architectures

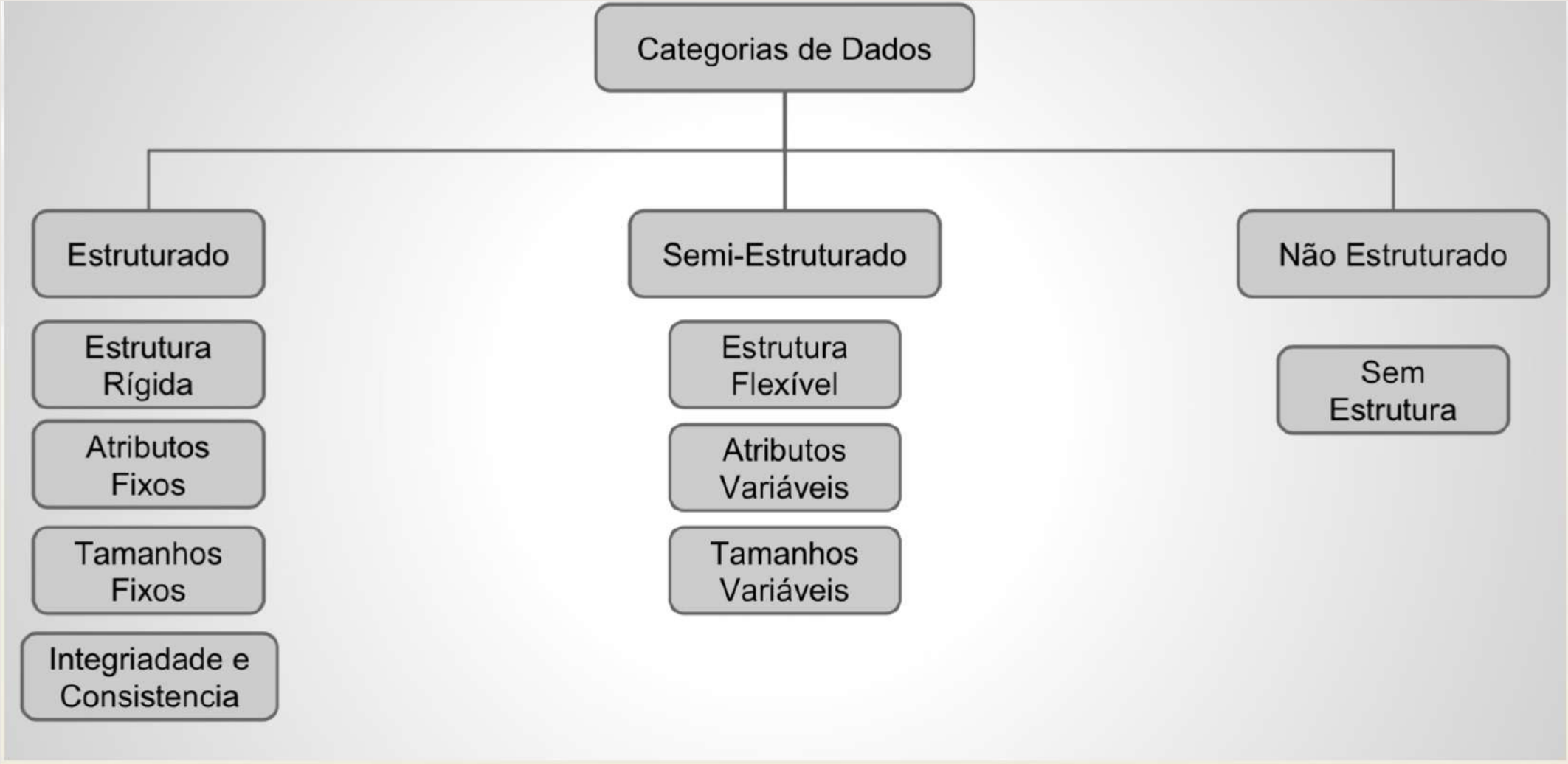
type	TechArticle
id	https://www.iunera.com/kraken/big-data-science-intelligence/a-quick-look-at-kappa-and-lambda-architectures/#techarticle
url	https://www.iunera.com/kraken/big-data-science-intelligence/a-quick-look-at-kappa-and-lambda-architectures/
inLanguage	en-US
mainEntityOfPage	https://www.iunera.com/kraken/big-data-science-intelligence/a-quick-look-at-kappa-and-lambda-architectures/#webpage
headline	A Quick Look At Kappa And Lambda Architectures
description	This article will explain the Kappa and Lambda Architectures in Stream processing and how they are designed to handle real-time streaming data

Tipos de Dados
Não Estruturados

- Estrutura flexível e dinâmica
- Sem uma estrutura pré-definida
- 80% do conteúdo digital não é estruturado
  - Texto em .pdf
  - Imagem .jpg
  - Vídeo ou áudio
- Banco de dados escaláveis (noSQL)



Resumo
Categoria de Dados



Tipos de Variáveis

- *São os valores que assumem determinadas características*
- *Variável estatística é uma característica que admite diferentes valores por cada unidade estatística*
- *É a característica dos elementos da amostra que nos interessa averiguar estatisticamente*

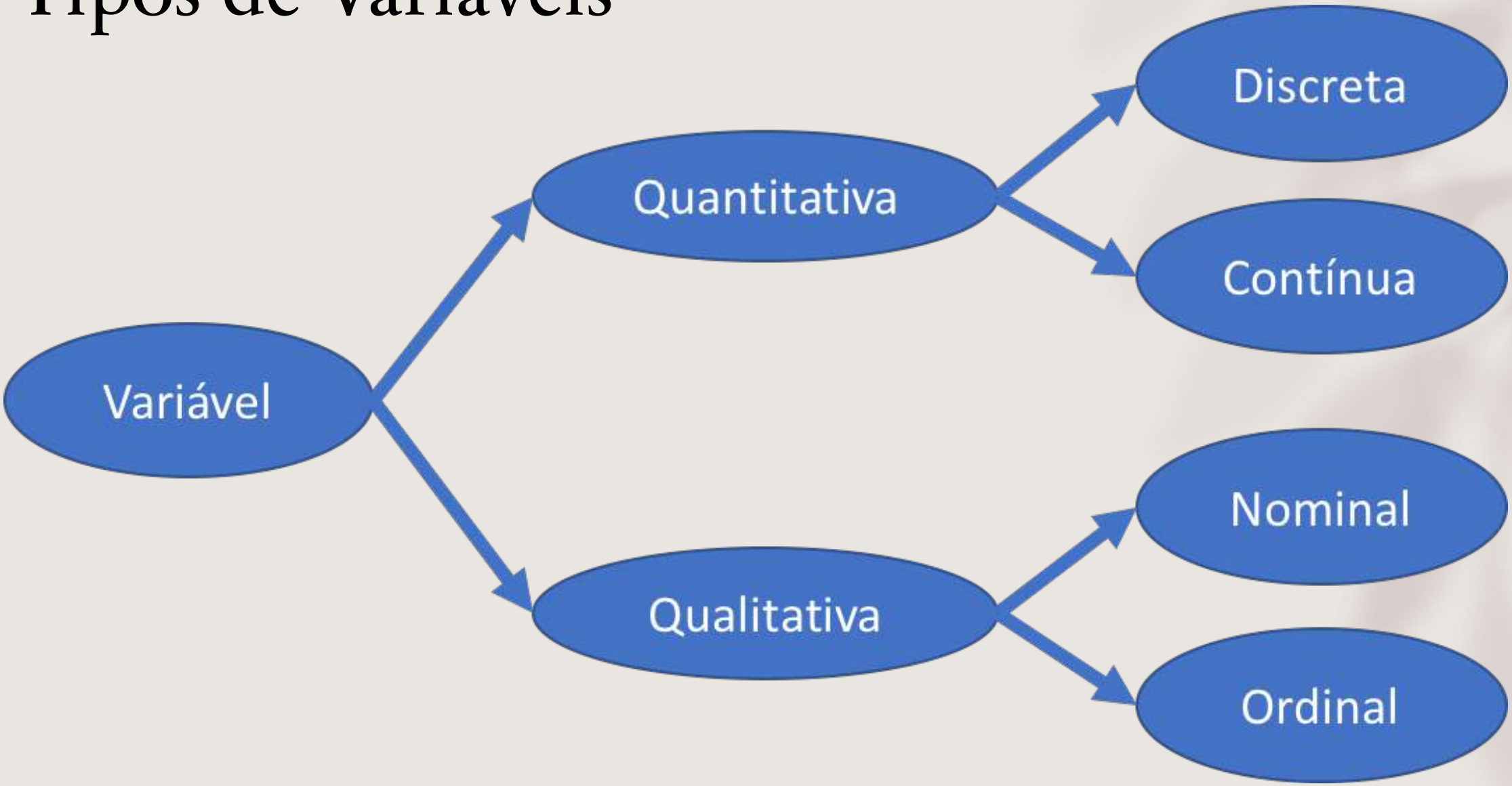
Qualitativa

- Variáveis que apresentam como possíveis realizações uma qualidade ou atributo do indivíduo pesquisado
  - **Nominal**: sexo, cor dos olhos
  - **Ordinal**: classe social, grau de instrução

Quantitativa

- Variáveis que apresentam como possíveis realizações números resultantes de uma contagem ou mensuração
  - **Contínua**: peso, altura
  - **Discreta**: número de filhos, número de carros

Tipos de Variáveis



Tipos de Variáveis
Exemplos

Escalas de Mensuração

O tipo da análise que pode ser realizado depende da escala de medida da variável analisada.
Tabela: sugestões de representações gráficas e resumos descritivos numéricos mais recomendáveis para o tipo de análise.

Escala de medida	Representações Gráficas	Medidas de tendência central	Medidas de dispersão
Nominal	Diagrama de barras Diagrama de linhas Diagrama de pizza	Moda	
Ordinal	Diagrama de caixa (boxplot)	Mediana	Intervalo interquartílico
Intervalo	Histogramas Polígono de frequências	Média	Desvio padrão
Razão		Média Geométrica	Coeficiente de Variação

Fluxo de Ciência de Dados

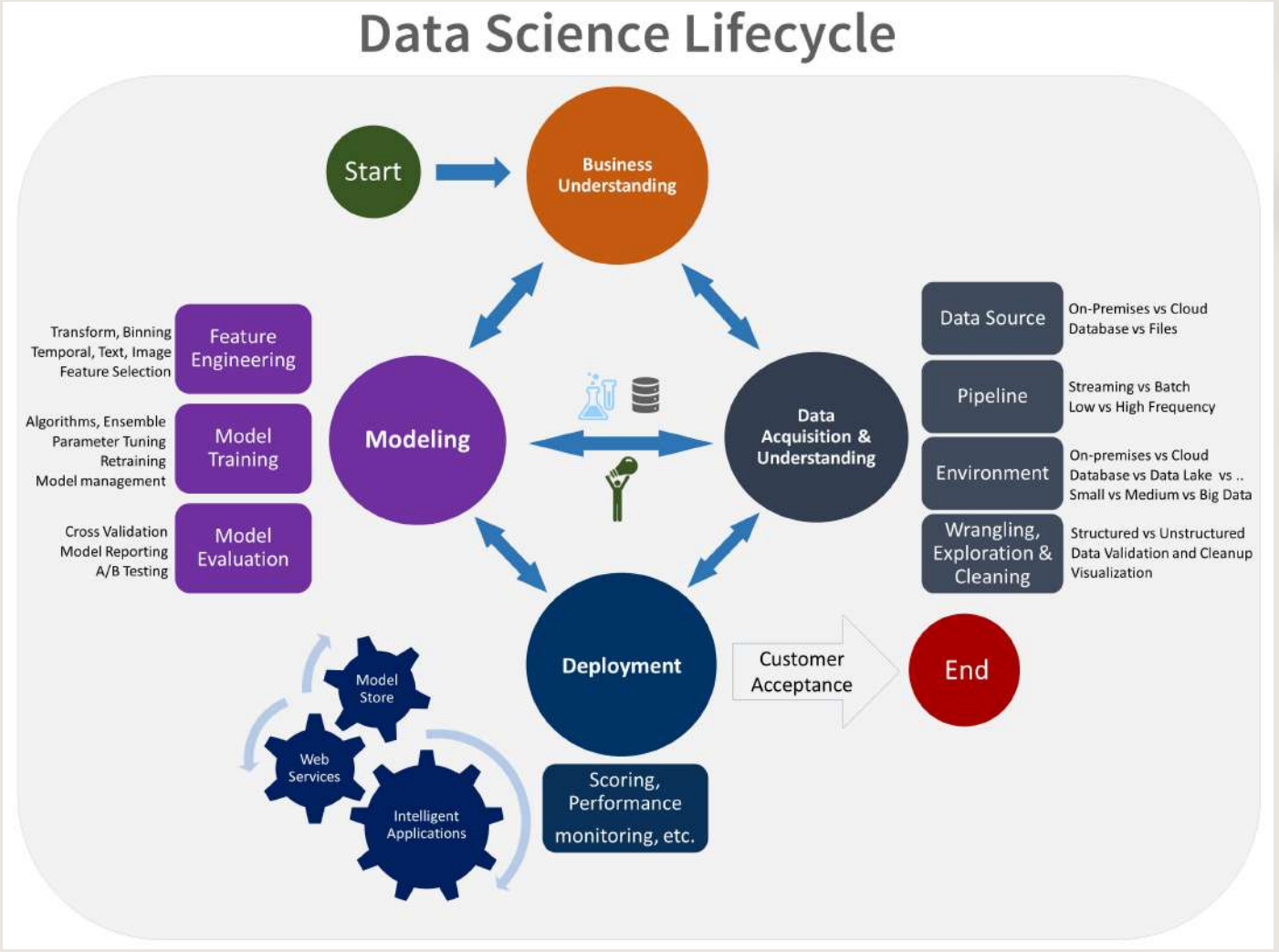
Apresentando alguns “frameworks” básicos

Algumas definições

- Ciência de Dados é uma abordagem multidisciplinar para extrair insights de grandes volumes de dados
- Seu processo envolve
  - Preparação de dados para análise e processamento
  - Execução de análises avançadas dos dados
  - Apresentação de resultados revelando padrões para tomada de decisão

https://www.ibm.com/cloud/learn/data-science-introduction#toc-the-data-s-HmAl4i7a

Cíclo de Vida de Ciência de Dados



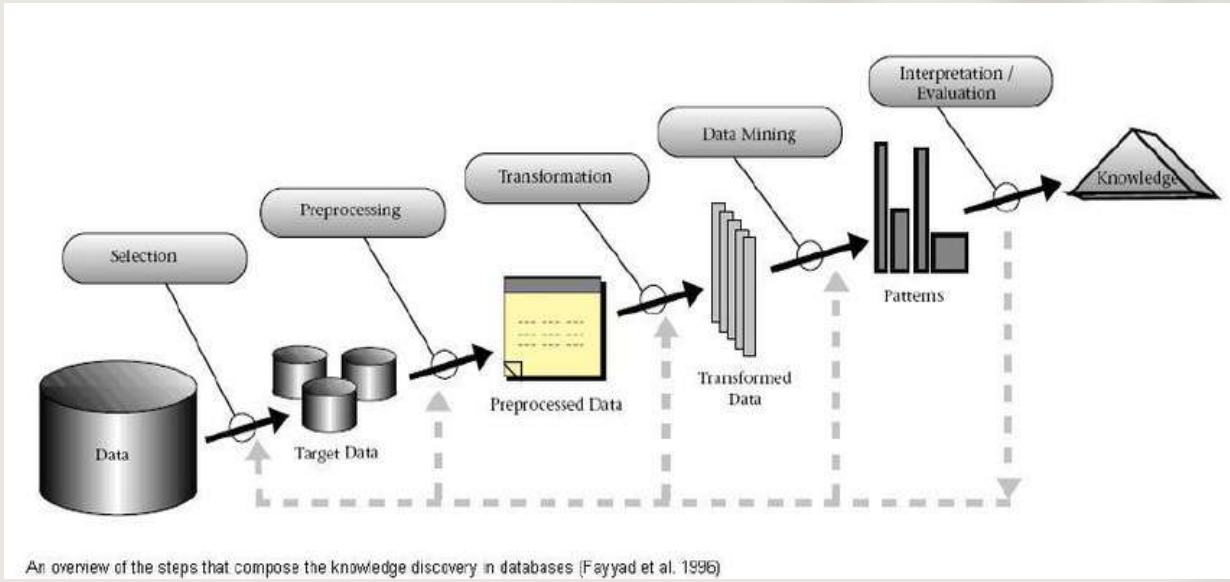
https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview

Principais atividades

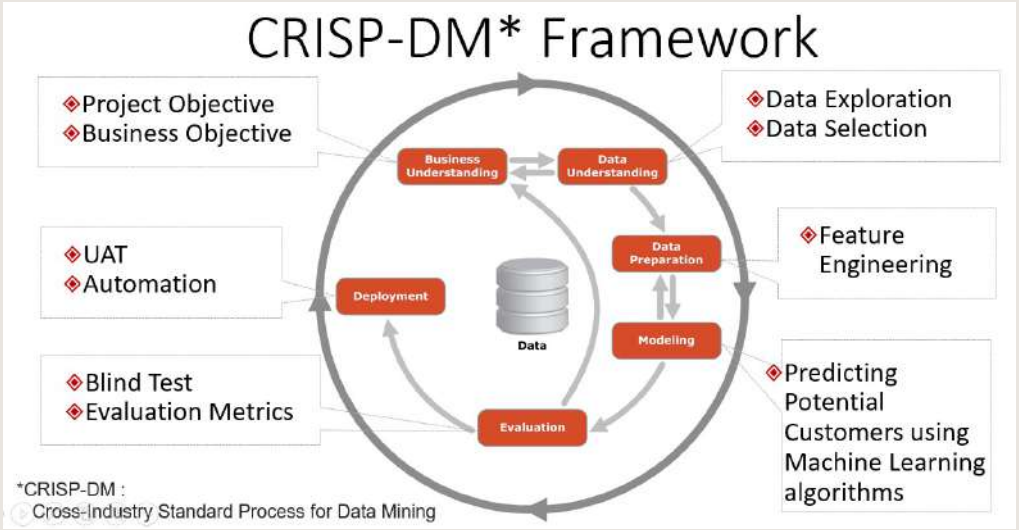
- A preparação de dados envolve limpeza (raspagem), agregação e transformação para que os dados estejam prontos para tipos específicos de processamento.
- A análise requer o desenvolvimento e o uso de algoritmos, análises e modelos de Machine Learning, IA.
- No processo de análise o software vasculha os dados para encontrar padrões internos e transformar esses padrões em previsões que apoiem a tomada de decisões de negócios.
- Os resultados devem ser compartilhados por meio do uso hábil de ferramentas de visualização de dados que possibilitem a qualquer pessoa ver os padrões e entender as tendências.

CRISP Data Mining Process

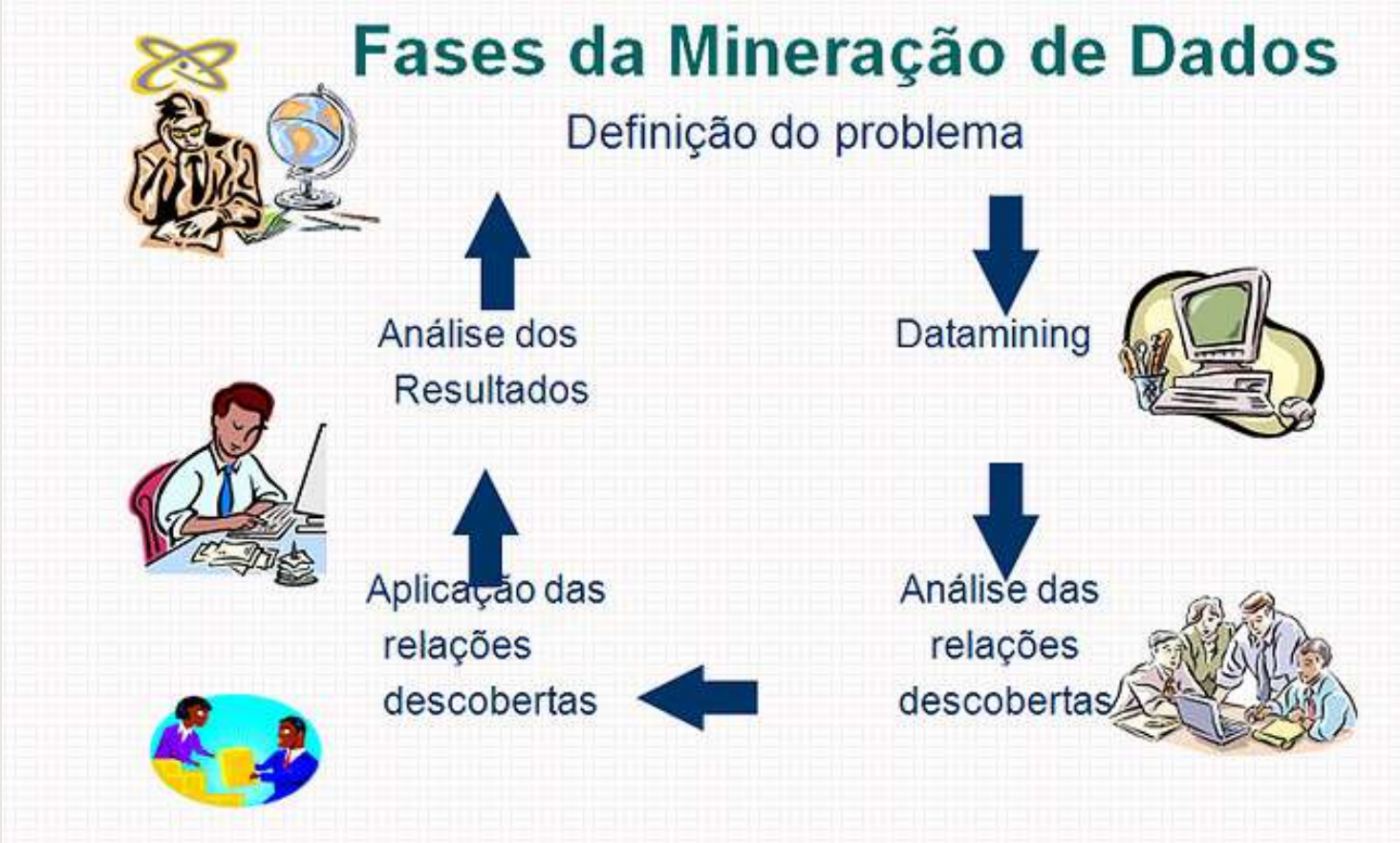
- CRoss Industry Standard Process for Data Mining (CRISP-DM
  - Modelo de processo com seis fases de mineração de dados
  - Relação muito próxima com os modelos de processo de Knowledge Discovery in Databases (KDD)
  - Início dos anos 1990
- **Mineração de dados**
  - processo de explorar dados à procura de padrões consistentes
  - detectar relacionamentos sistemáticos entre variáveis
  - descobrir regras, identificar fatores e tendências-chave, descobrir padrões e relacionamentos ocultos em grandes bancos de dados para auxiliar a tomada de decisões
  - Usada preferencialmente em **Business Intelligence** (BI)



An overview of the steps that compose the knowledge discovery in databases (Fayyad et al. 1996)



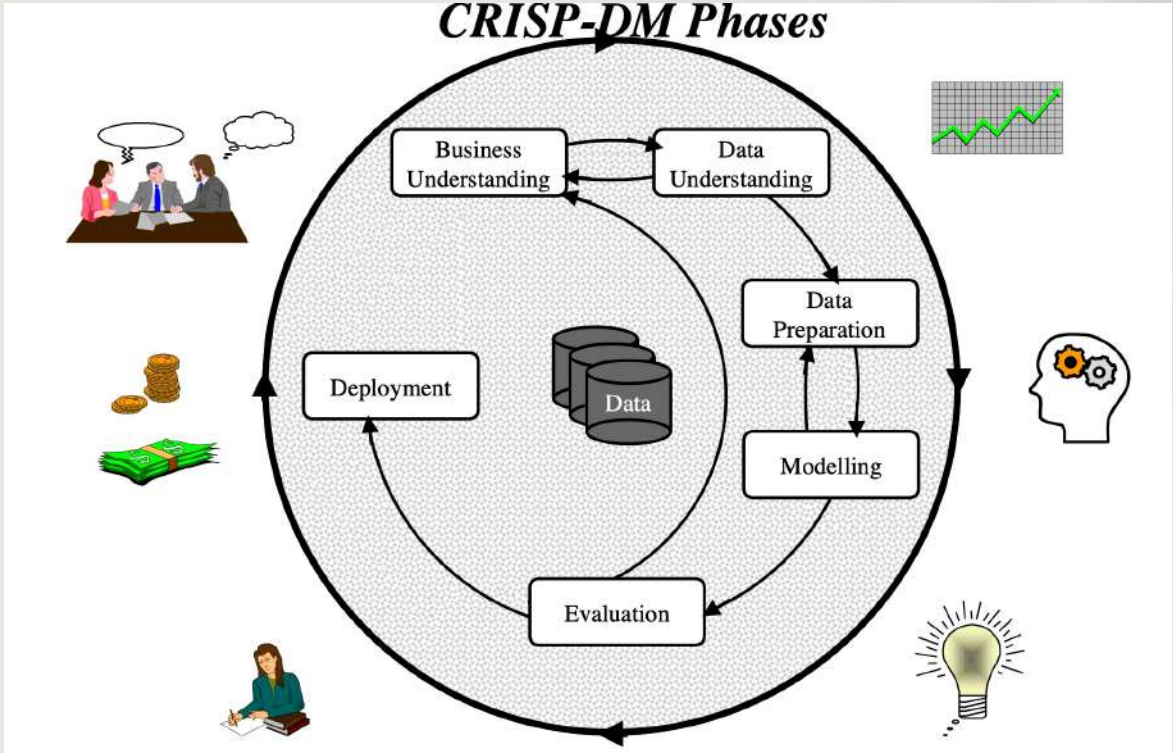
Fluxo de mineração de dados



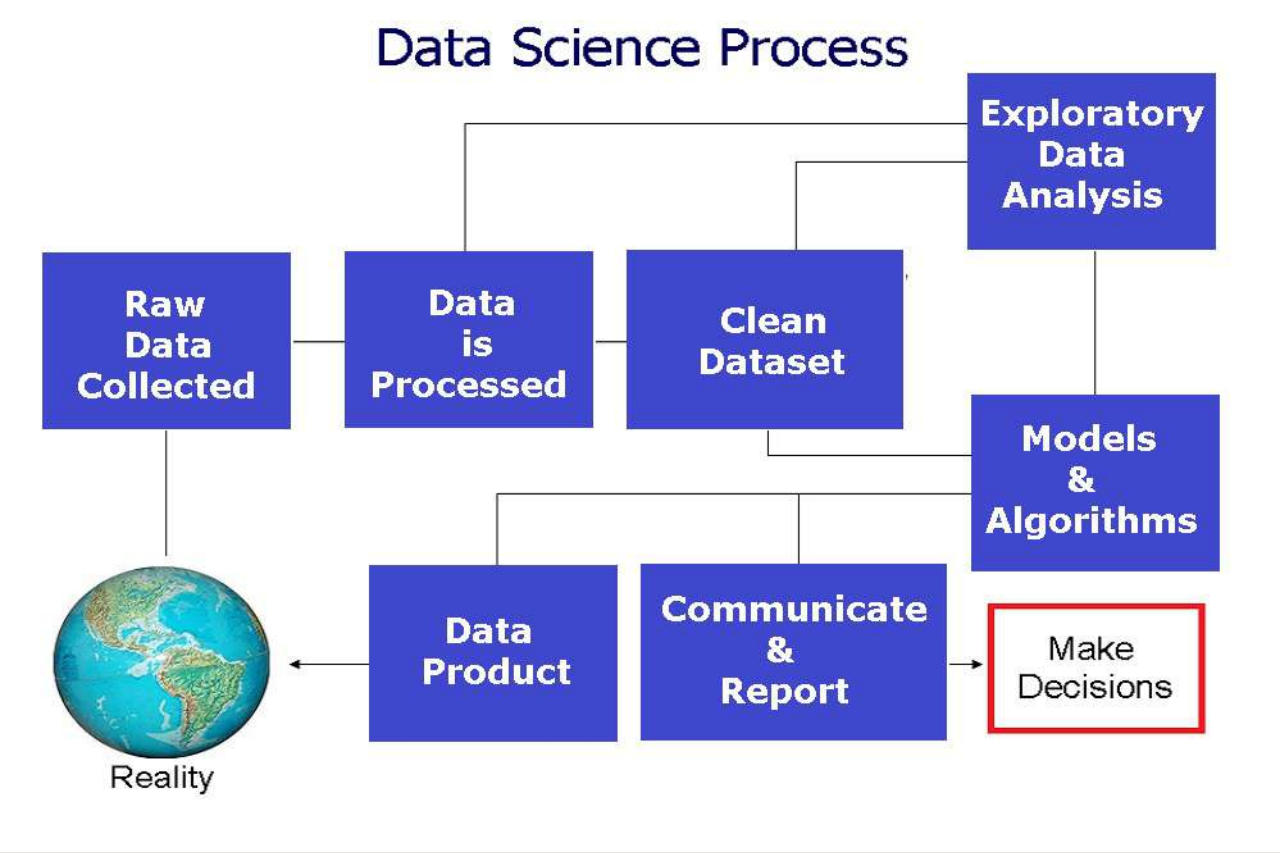
http://bicomvatapa.blogspot.com/2014/11/mineracao-de-dados-evolucao-natural-da.html

Fluxo CRISP-DM

- Publicado em 1999 para padronizar os processos de mineração de dados em todos os setores
- Tornou-se metodologia padrão para projetos de mineração de dados
- Conjunto de fases para planejar, organizar e implementar um projeto
- Etapas
  - Compreensão do negócio - O que o negócio precisa?
  - Compreensão de dados - Que dados temos / precisamos? Estão limpos?
  - Preparação de dados - como organizamos os dados para modelagem?
  - Modelagem - Quais técnicas de modelagem devemos aplicar?
  - Avaliação - Qual modelo atende melhor aos objetivos de negócios?
  - Implementação - Como as partes interessadas acessam os resultados?
- Modelo base absorvido para projetos em Data Science

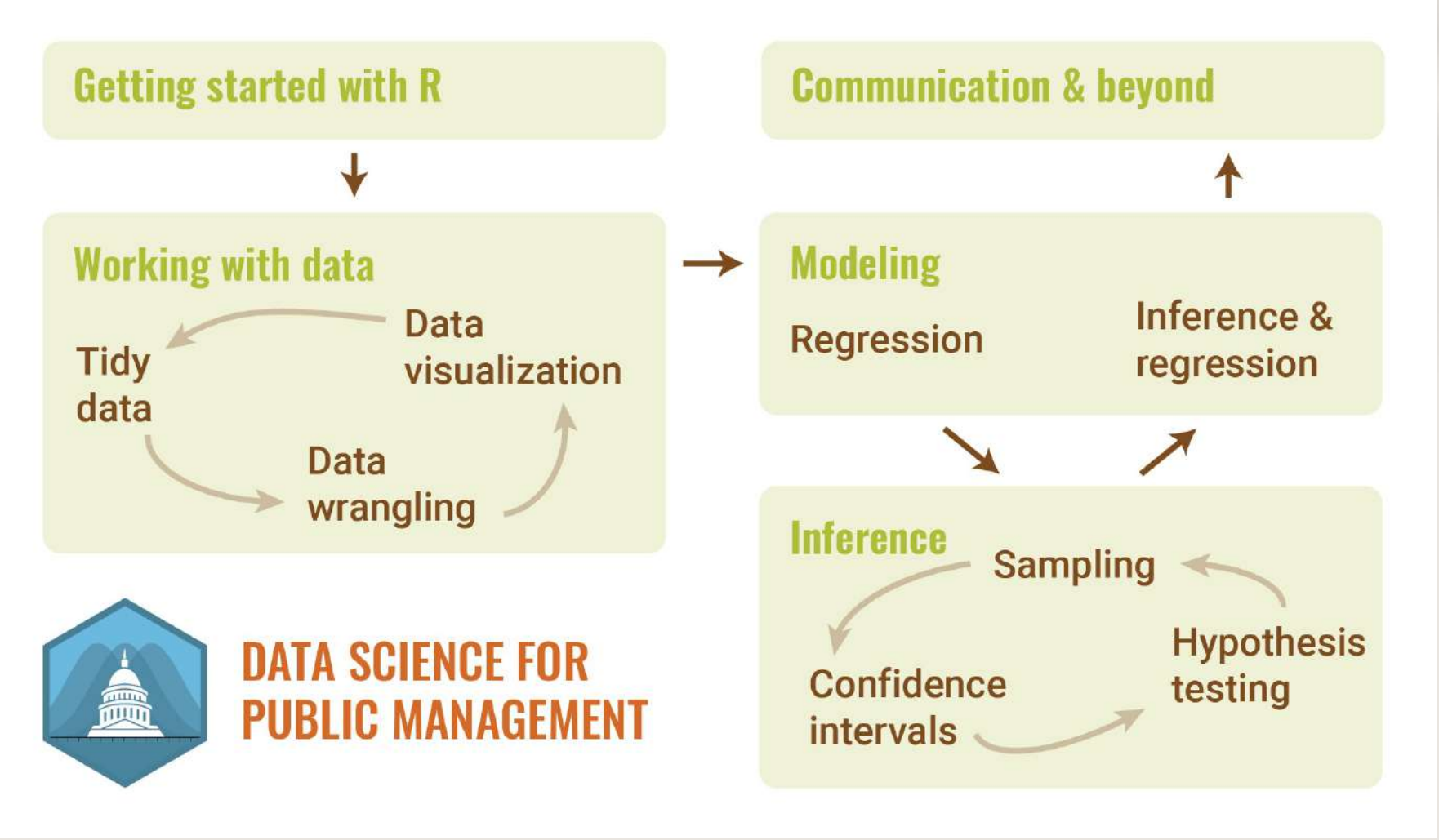


Alguns fluxos de Ciência de Dados



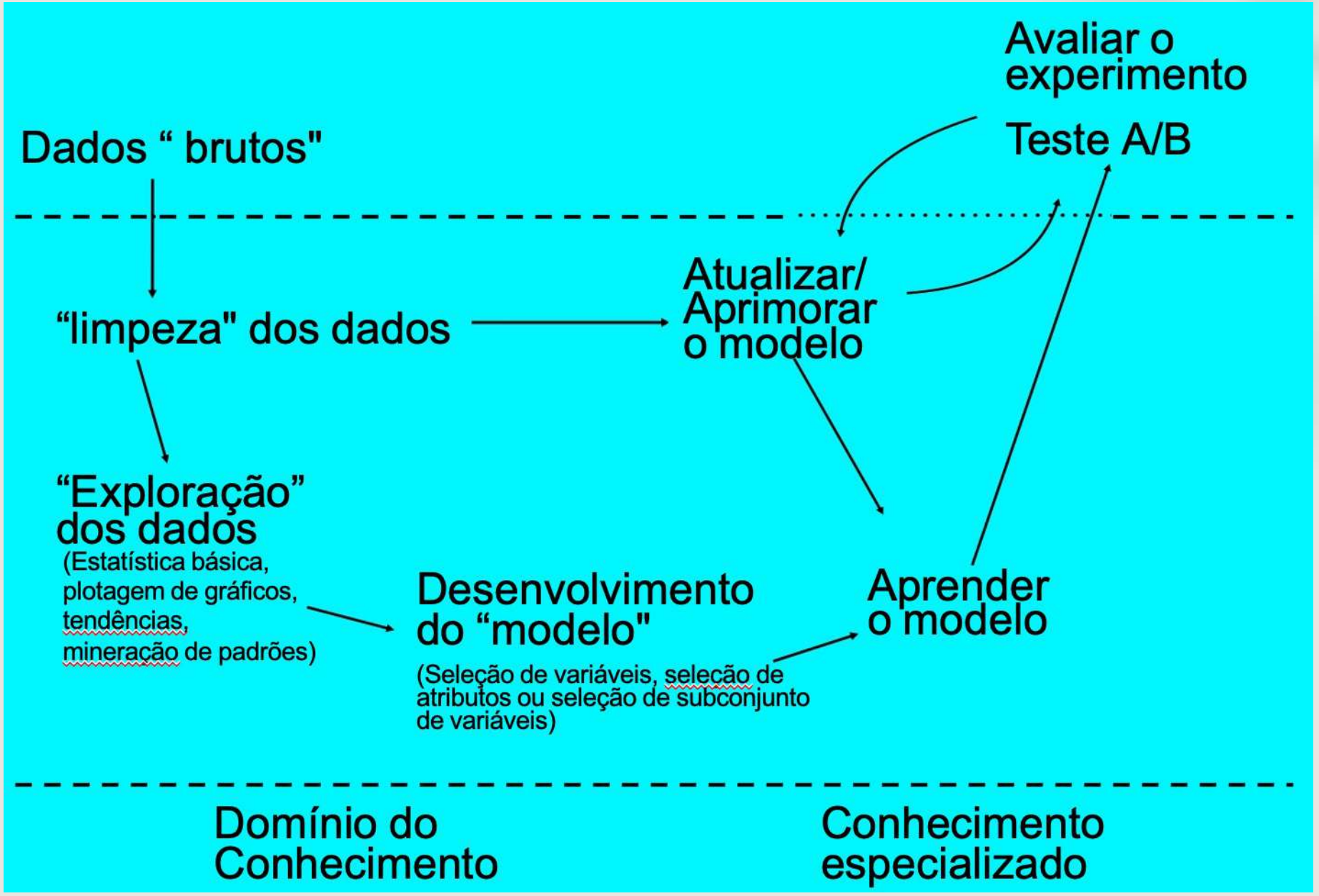
SCHUTT, R., O'NEIL, C. **Doing Data Science: Straight Talk from the Frontline**, O'Reilly Media, California, USA 2014. 408 p

Modelo proposto



https://statsf18.classes.andrewheiss.com/

Fluxo de Ciência de Dados



If you torture the data long enough, it will confess to anything.

Ronald Coase

https://i.redd.it/ikdymainj6p21.jpg

Análise Exploratória de Dados

Noções Gerais sobre o Fluxo de Data Science

João Pedro Albino

Departamento de Computação / Faculdade de Ciências

PPG-MiT / Faculdade de Artes, Arquitetura, Comunicação e Design