

MBAUSP
ESALQ

DATA WRANGLING

Prof. Dr. Wilson Tarantin Junior

Filipe Carbonera De Souza 020.163.260-85

MBAUSP ESALA

A responsabilidade pela idoneidade, originalidade e licitude dos conteúdos didáticos apresentados é do professor.

Proibida a reprodução, total ou parcial, sem autorização.

Lei nº 9610/98

Data Wrangling

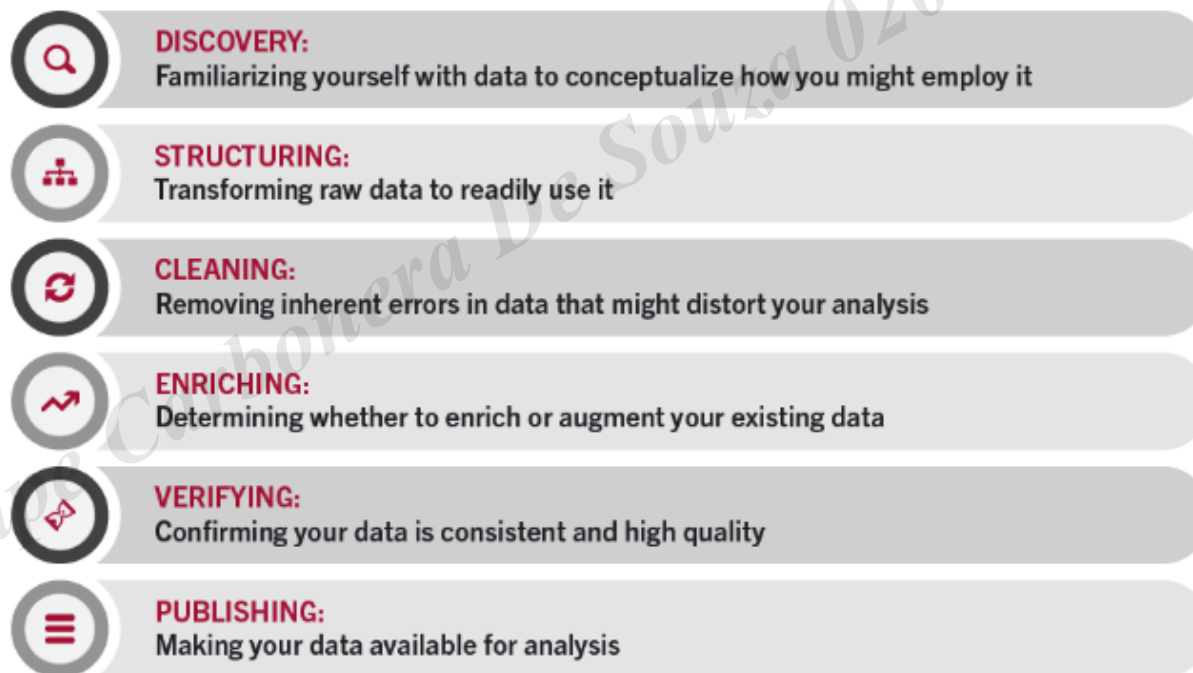
- O que é data wrangling?
 - É o processo de transformar a base de dados de sua estrutura original (dados brutos) para uma nova estrutura que permita a adequada extração de informações
 - Raramente os dados brutos estão disponíveis na estrutura mais adequada para as análises
 - Portanto, é uma etapa de preparação, organização, manipulação dos bancos de dados
 - Ocorre antes da análise exploratória dos dados, criação de gráficos e modelagem
 - Quanto melhor for a etapa de data wrangling, melhores serão os dados e maior tende a ser a qualidade das informações extraídas
 - Não é uma atividade padronizada: em função do contexto, diferentes etapas serão aplicadas

Data Wrangling

- Etapas do processo de **tratamento dos dados**
 - **Coleta e importação:** os dados podem ser provenientes de fontes diversas
 - Exemplos: planilhas de Excel e CSV, API, SQL. Em alguns casos, é possível automatizar a coleta
 - **Junção:** os dados provenientes das várias fontes precisam ser reorganizados para que seja possível a união de observações e variáveis
 - Em muitos casos, exige variáveis que sejam “chaves” para o relacionamento entre tabelas
 - **Transformação:** contempla as diversas etapas de modificação dos dados
 - Exemplos: limpeza dos dados, criação e alteração de variáveis, seleções, agregações, resumos...

Data Wrangling

6 Steps of Data Wrangling



Fonte: <https://online.hbs.edu/blog/post/data-wrangling>

Python

- Biblioteca para data wrangling
 - Embora muitas bibliotecas sejam úteis no processo de manipulação de dados, utilizaremos o **pandas** como ferramenta principal
 - Manual do usuário: https://pandas.pydata.org/docs/user_guide/index.html
 - Material de consulta: https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf

MBAUSP ESALQ

Obrigado!

Wilson Tarantin Junior

[linkedin.com/in/wilson-tarantin-junior-359476190](https://www.linkedin.com/in/wilson-tarantin-junior-359476190)