Introdução à Ciência de Dados como Prática na Pesquisa Acadêmica

Seção 02 - Ciência de Dados e a Pesquisa Acadêmica

Jadson Pessoa

Professor do DECON | Membro GAPE

Plano de Trabalho

Objetivos

Método Científico como Teória e Prática

DS e a Pesquisa Teórica-Empírica

O mundo *tidyverse*

Objetivos

- Discutir o método científico e sua relação com a sistematização de um projeto de Data Science (DS), assim como, apresentar os modelos canônicos da pesquisa acadêmica, tendo como foco pesquisas que utilizam aboragens empíricas.

Método Científico como Teória e Prática

- Ciência como prática (técnica);
- Ciência como método;
- Ciência como epistemologia.

  "Ciência é sempre um enlace de uma malha teórica com dados empíricos, é sempre uma articulação do lógico com o real, do teórico com o empírico, do ideal com o real" Severino [2017], *p. 100.*

Método Científico

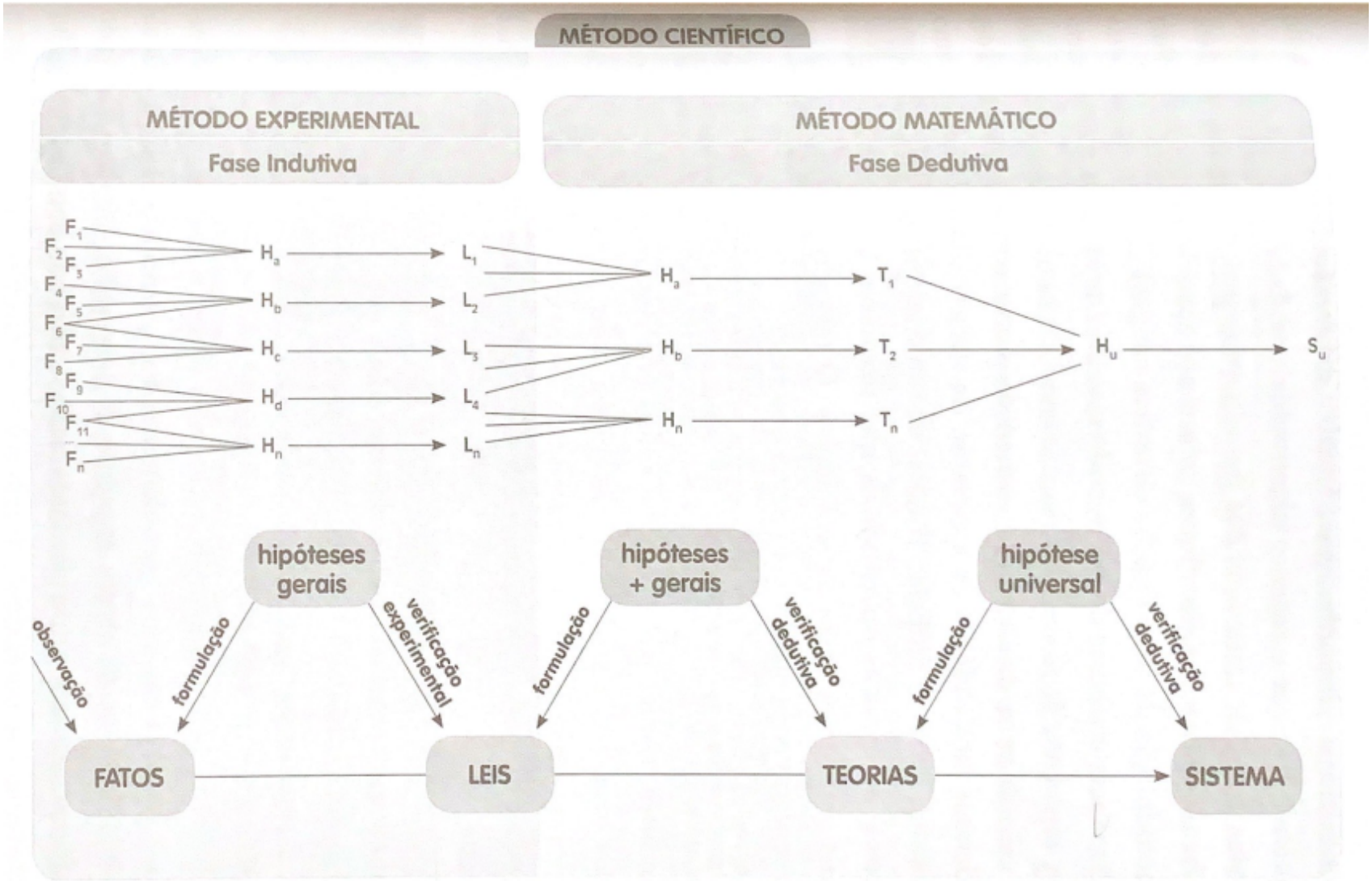

Figure 1: Estrutura Lógica do método científico Severino [2017], p.101.

Método Científico Revisitado

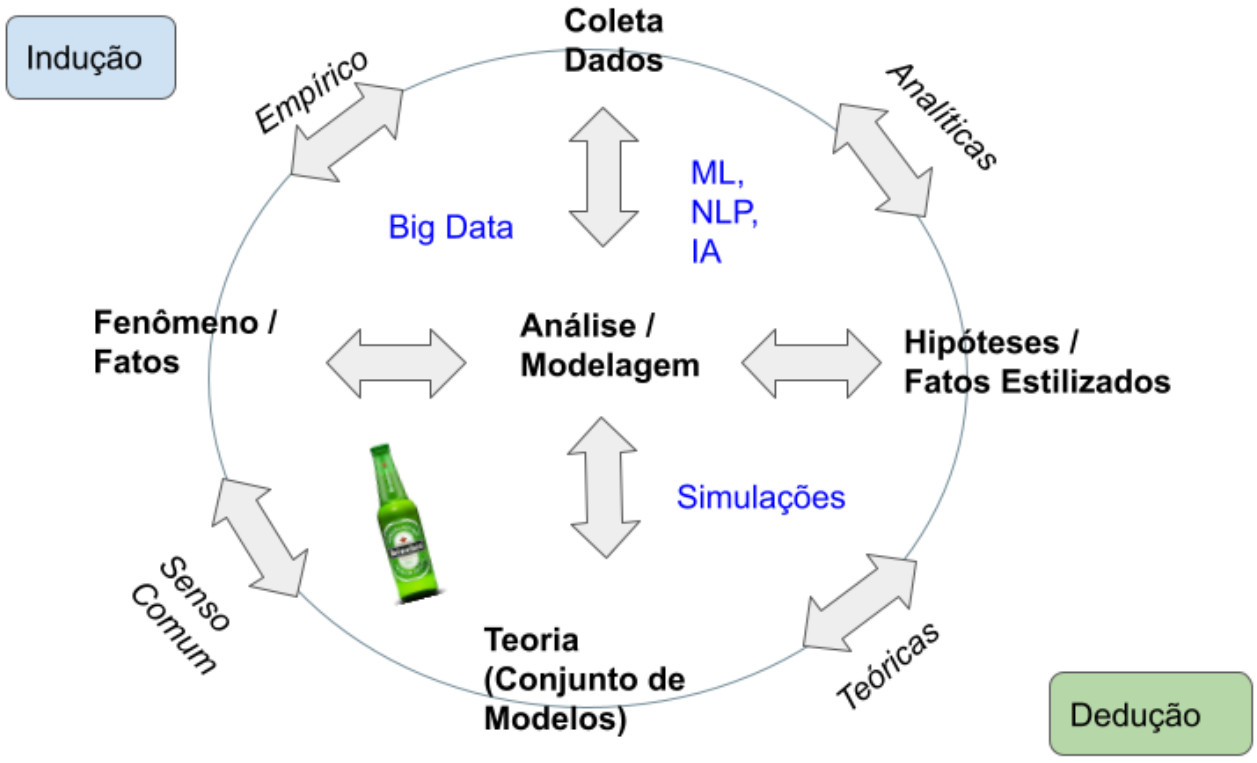

Figure 2: Computational Social Science Methods - Hilbert [2021]

Metodologias de Pesquisa Científica

Ciência:
Aplica Técnicas $\rightarrow$ Métodos $\rightarrow$ Fundamentos Epistemológicos

- Modalidades de Pesquisa Acadêmicas:
  - Pesquisa Teórica
  - Pesquisa Teórica-Empírica

Elementos de um Artigo Empírico

INTRODUÇÃO	Discussão Geral	Justificativa	Problema de Pesquisa
MÉTODO	Base de Dados	Desenho da Pesquisa	Execução
RESULTADOS E DISCUSSÃO	Resultados Obtidos	Discussão com a teoria ou com outros trabalhos	Resposta a questão
CONCLUSÃO	Achados (*Findings*)	Possíveis Contribuições	Novas questões

Figure 3: Elementos - Artigo Empírico

Exemplos: Artigos 1; Revista Economia Aplicada - USP; Journal of Applied Economics

DS e a Pesquisa Teórica-Empírica

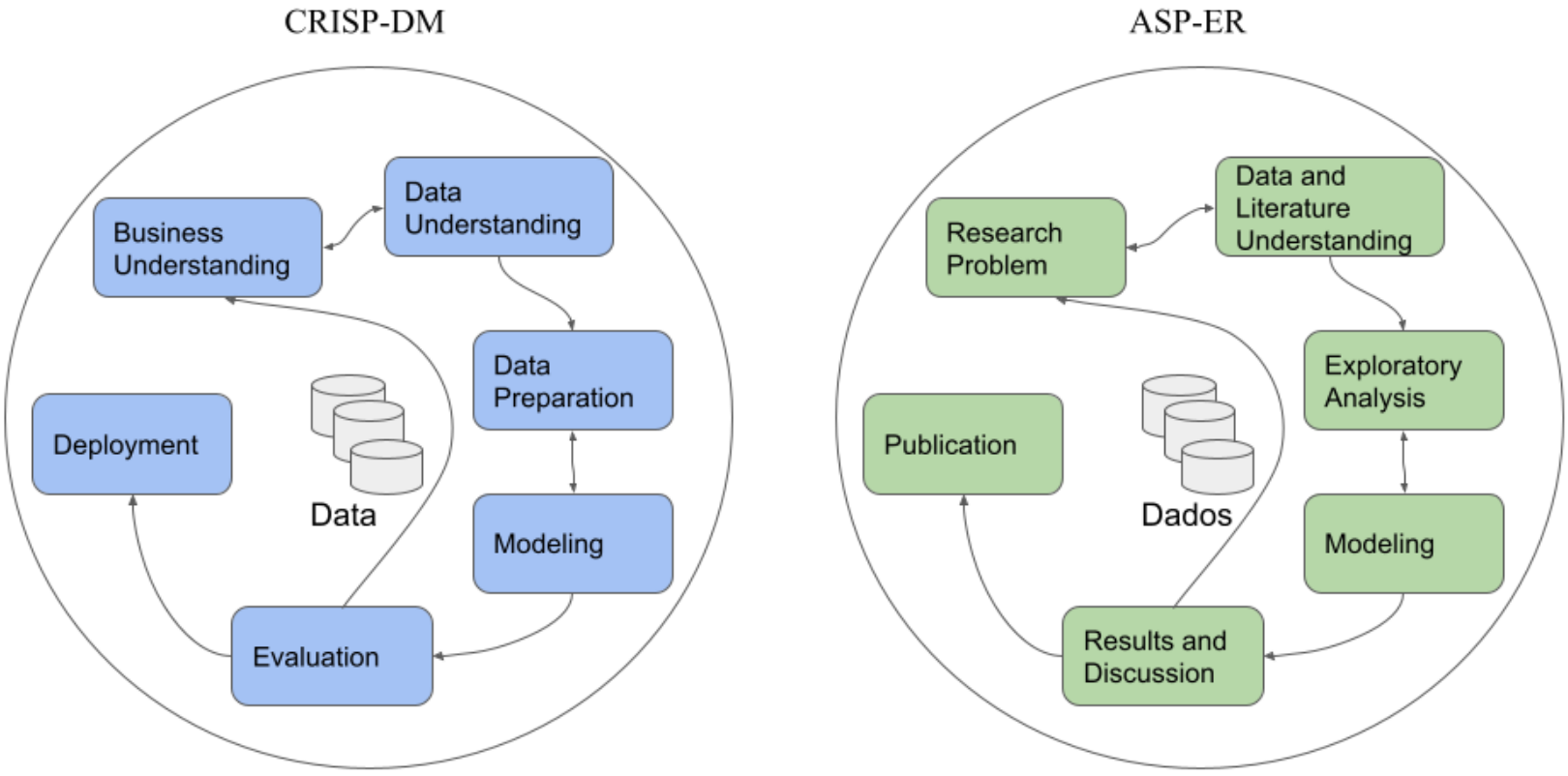

Figure 4: CRISP-DM IBM [2021] e ASP-ER

Coleta e Tratamento

A **coleta** dos dados podem em diferentes formatos:

- Excel;
- XML;
- JSON;
- txt;
- HTML;
- MySQL;
- Formatos proprietários (Stata, Minitab, SPSS, SAS, etc).

Coleta e Tratamento

O dados precisam ser **tratamentodos**:

- Limpeza de dados;
- Tratamento de *missing values* ou NA;
- Construção de números índices;
- Deflacionar valores correntes;
- Obtenção de taxas de crescimento;
- Tratando tendências;
- Dessazonalização;
- Subconjuntos (*subsetting*);
- Classificação;
- Utilização de *lags*.

Visualização

Uma vez que seus dados estejam arrumados, podemos passar para a parte de **exploração dos dados**. A exploração de dados é a arte de analisar seus dados, gerando hipóteses rapidamente, testando-os rapidamente, repetindo-os várias vezes. O objetivo da exploração de dados é gerar muitos leads promissores que você poderá explorar mais tarde com mais profundidade. Em geral, faz-se exploração de dados por meio da *visualização* desses dados.

Um bom processo de visualização de dados permite que possamos nos concentrar naquilo que realmente importa, deixando de lado relações não tão importantes.

Modelagem

Uma vez que tenhamos conseguido propor uma *hipótese de trabalho* através da etapa de exploração/visualização de dados, o próximo passo é propor um **modelo** entre as variáveis do nosso conjunto de dados. O objetivo da modelagem é capturar a essência de um conjunto de dados.

Comunicação

A última etapa do processo de *data science* é comunicar os resultados para clientes, gestores ou demais interessados. Na pesquisa acadêmica: publicar! É uma fase absolutamente crítica do projeto. Isto porque, ao menos que você consiga se comunicar com a sua audiência, de nada valeu todo o trabalho realizado nas etapas anteriores.

O mundo *tidyverse*

De modo a fazer cada uma dessas etapas dentro do R, nós vamos utilizar a família de pacotes `tidyverse`. Assim, antes de qualquer coisa, certifique-se que você tenha o tenha instalado na atual versão do R.

```
install.packages('tidyverse')
require(tidyverse)
```

O mundo *tidyverse*

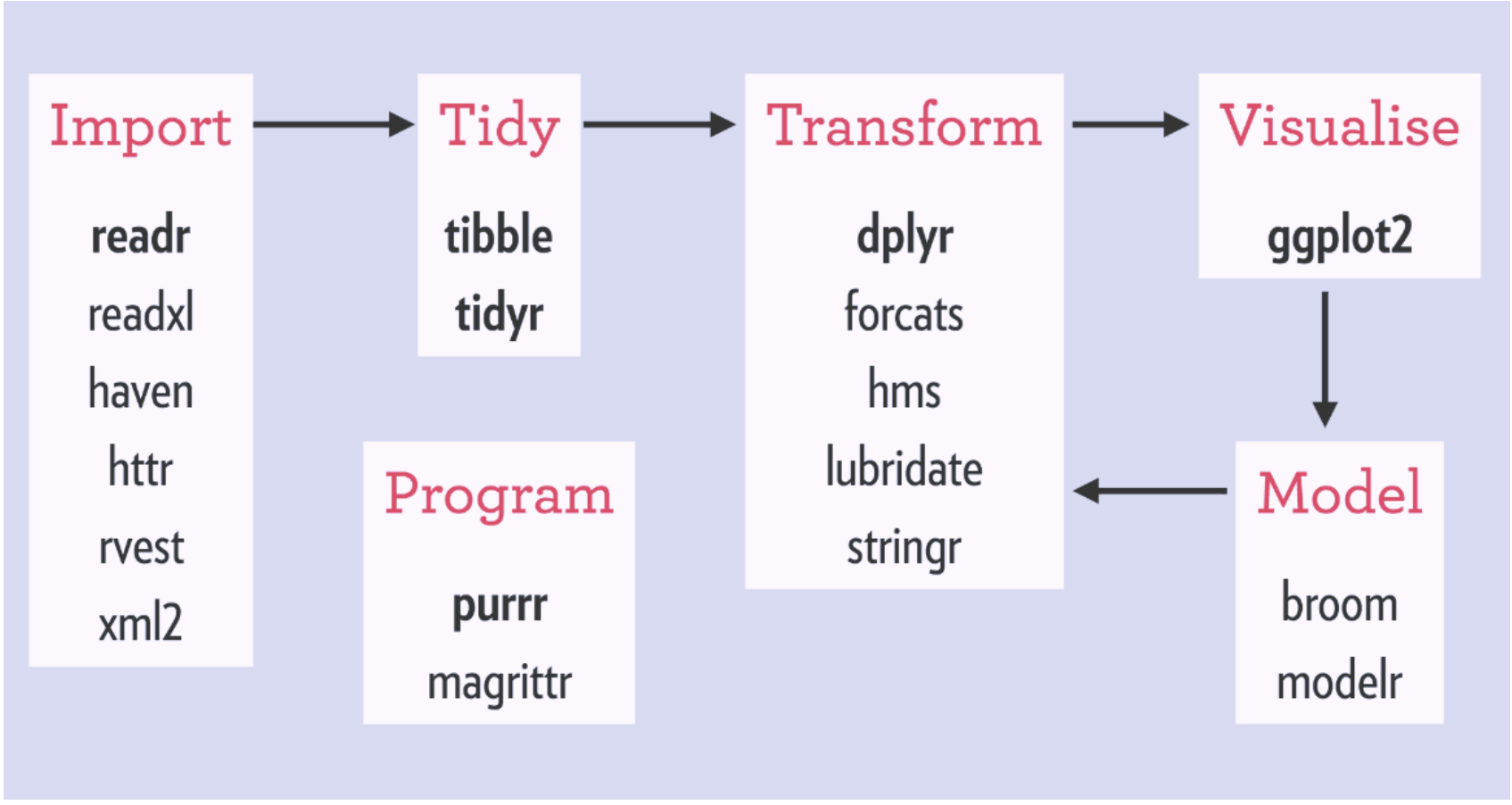

Figure 5: Os pacotes `tidyverse` (Wickham and Grolemund [2017])

Referências I

Martin Hilbert. Computational Social Science Methods, 2021. URL https://www.coursera.org/learn/computational-social-science-methods/home/welcome.

IBM. IBM SPSS Modeler CRISP-DM Guide, May 2021. URL https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/en/spss-modeler/SaaS?topic=SS3RA7_sub/modeler_crispdm_ddita/modeler_crispdm_ddita-gentopic1.html.

Antônio Joaquim Severino. *Metodologia do trabalho científico*. Cortez editora, 2017.

H. Wickham and G. Grolemund. *R for Data Science*. O'Reilly Media, 2017.