Introdução ao mundo *tidyverse*

Vítor Wilher

Cientista de Dados | Mestre em Economia

Plano de Voo

Introdução

Coleta e Tratamento

Exploração

Modelagem

Comunicação

A família *tidyverse*

Introdução

O avanço da informática e das telecomunicações possibilitou o armazenamento e a distribuição de conjuntos de dados cada vez mais complexos. Lidar com essas bases de dados exigiu a sistematização de diversas técnicas de coleta, tratamento, análise e apresentação de dados.

Essa sistematização de técnicas deu origem ao que hoje chamamos de **data science**, cujo objetivo principal é extrair informações úteis de conjuntos de dados aparentemente confusos.

Introdução

Aplicações interessantes:

- Identificar mensagens indesejáveis em um e-mail (spam);
- Segmentação do comportamento de consumidores para propagandas direcionadas;
- Redução de fraudes em transações de cartão de crédito;
- Predição de eleições;
- Otimização do uso de energia em casas ou prédios;
- etc, etc, etc...

Introdução

Ao coletar dados, introduzimos em uma plataforma de Análise de Dados (como o R) informações coletadas no mundo real. Seja vindas de base de dados prontas ou adquiridas minerando dados abertos em sites. Depois tratamos as informações para que sejam devidamente legíveis para um computador e bem formatadas para nosso próprio entendimento. Então começamos a análise propriamente dita. Visualizamos os dados, procurando perguntas interessantes e padrões, depois avaliamos nossas intuições com modelos estatísticos. Por fim, comunicamos nossos achados - afinal de pouco importa achar algo somente para si.

Introdução

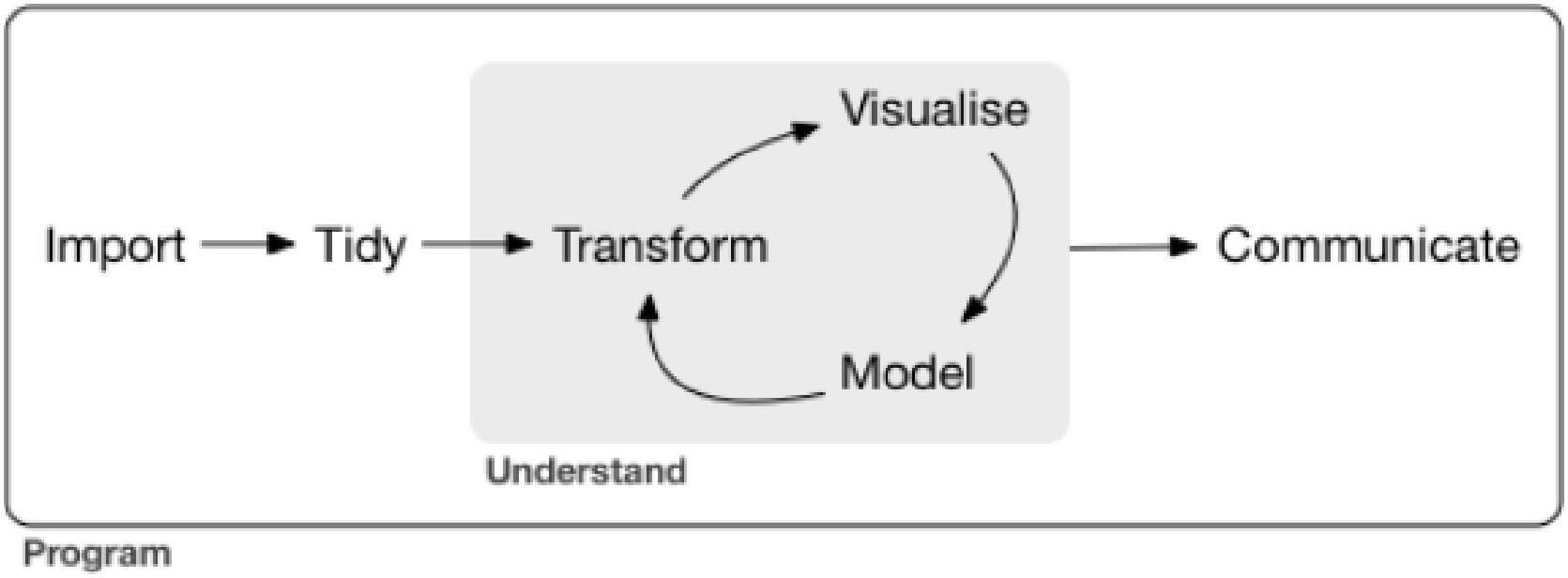

Figure 1: O processo de compreensão dos dados (Grolemund and Wickham [2017])

Coleta e Tratamento

Dados podem estar dispostos em diferentes formatos:

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

Coleta e Tratamento

Dados precisam ser tratados:

- Limpeza de dados;
- Tratamento de *missing values*;
- Construção de números índices;
- Deflacionar valores correntes;
- Obtenção de taxas de crescimento, a partir de comparações mensais, interanuais, acumuladas em 12 meses, etc;
- Tratando tendências;
- Dessazonalização;
- Obtendo subconjuntos (*subsetting*) relevantes;
- Classificando dados de acordo com algum critério;
- Transformando dados de acordo com alguma operação.

Exploração

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* dos mesmos.

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. 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.

A família *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)
```

A família *tidyverse*

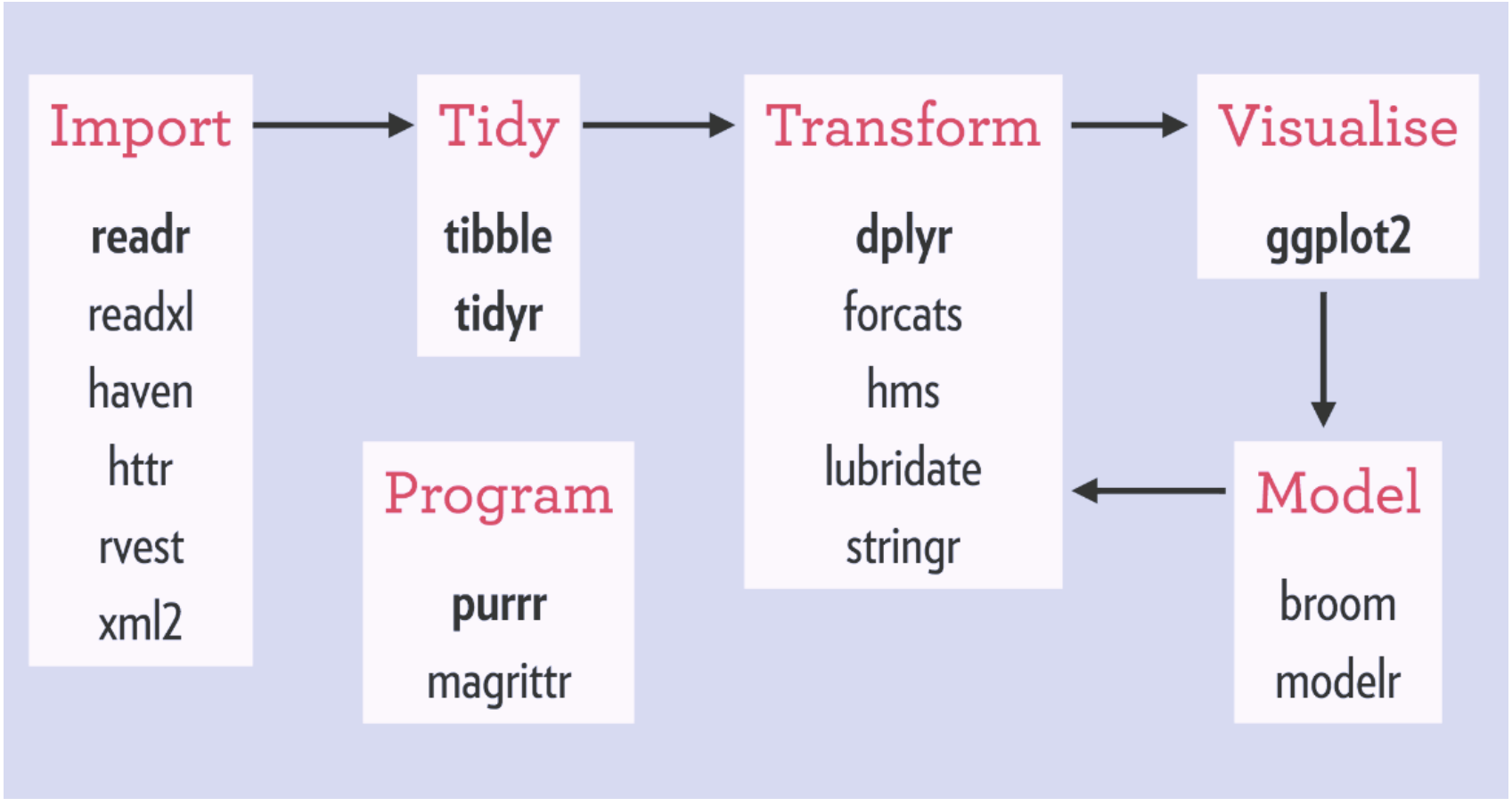

Figure 2: Os pacotes `tidyverse` (Grolemund and Wickham [2017])

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