



Data Science Academy

Formação Cientista de Dados

Módulo 1



Data Science Academy

www.datascienceacademy.com.br

Formação Cientista de Dados

Big Data Analytics com R e Microsoft Azure Machine Learning



Data Science Academy

www.datascienceacademy.com.br

Formação Cientista de Dados

Big Data Analytics com R e Microsoft Azure Machine Learning

Linguagem
R

Capítulos
2, 3, 4, e 5

Estatística

Capítulo
6

Machine
Learning

Capítulo
7

Azure
Machine
Learning

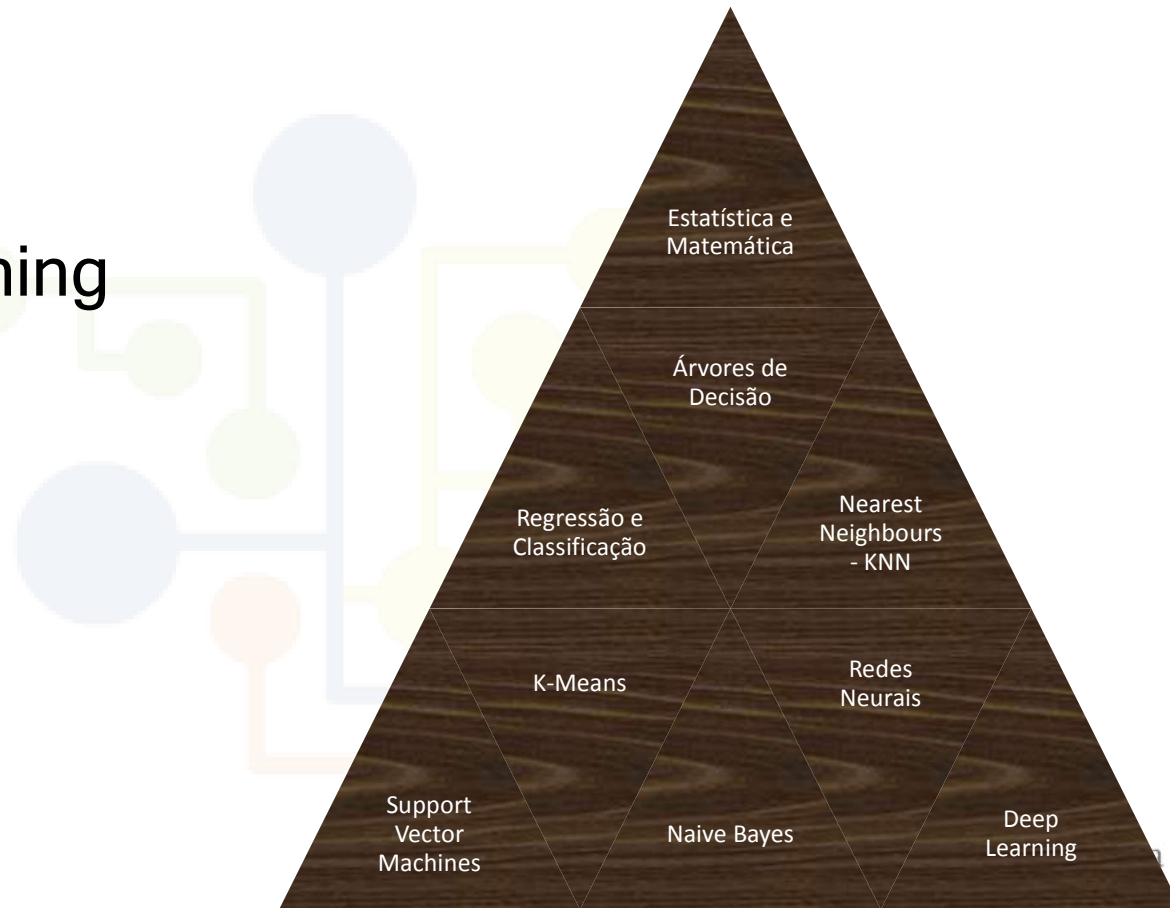
Capítulos
8, 9, 10 e 11



Data Science Academy

Formação Cientista de Dados

Machine Learning

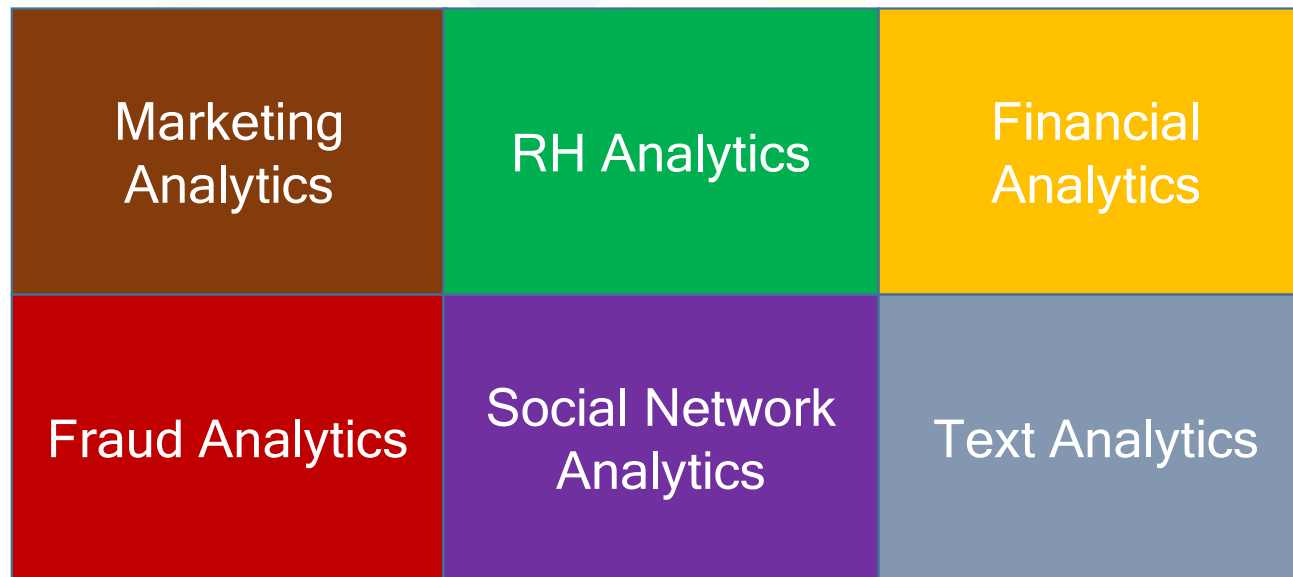


Data Science Academy

www.datascienceacademy.com.br

Formação Cientista de Dados

Business Analytics



Data Science Academy



Por que Cientistas de Dados usam R?




Data Science Academy

www.datascienceacademy.com.br



R possui diversas funções para:

- Extração de Dados
 - Limpeza de Dados
 - Carregamento e Transformação de Dados
 - Análise Estatística
 - Modelagem Preditiva
 - Machine Learning
 - Visualização de Dados
- 



Data Science Academy



Vantagens



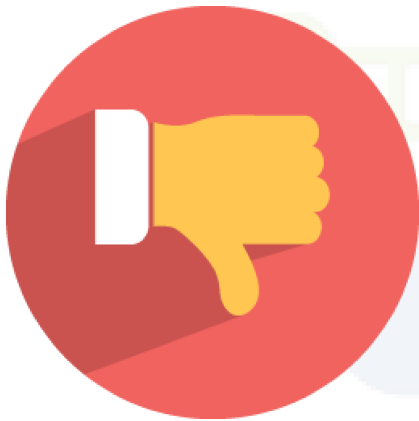
- Grande variedade de pacotes disponíveis
- Flexibilidade e Rapidez
- Machine Learning



Data Science Academy



Desvantagens



- Não há interface gráfica. Tudo é feito por linha de comando
- Limitações no uso de memória, principalmente com datasets muito grandes



Data Science Academy



Versatilidade



Data Science Academy

www.datascienceacademy.com.br



- **sqldf** - pacote que permite realizar queries SQL em dataframes no R
- **forecast** - modelar séries temporais
- **plyr** - dividir uma estrutura de dados em grupos e aplicar funções a cada grupo
- **stringr** - manipulação de strings
- **Database drivers** - RMongo, RODB, RMySQL
- **ggplot2** - visualização de dados
- **qcc** - controle de qualidade estatístico
- **randomForest** - pacote para Machine Learning



Data Science Academy

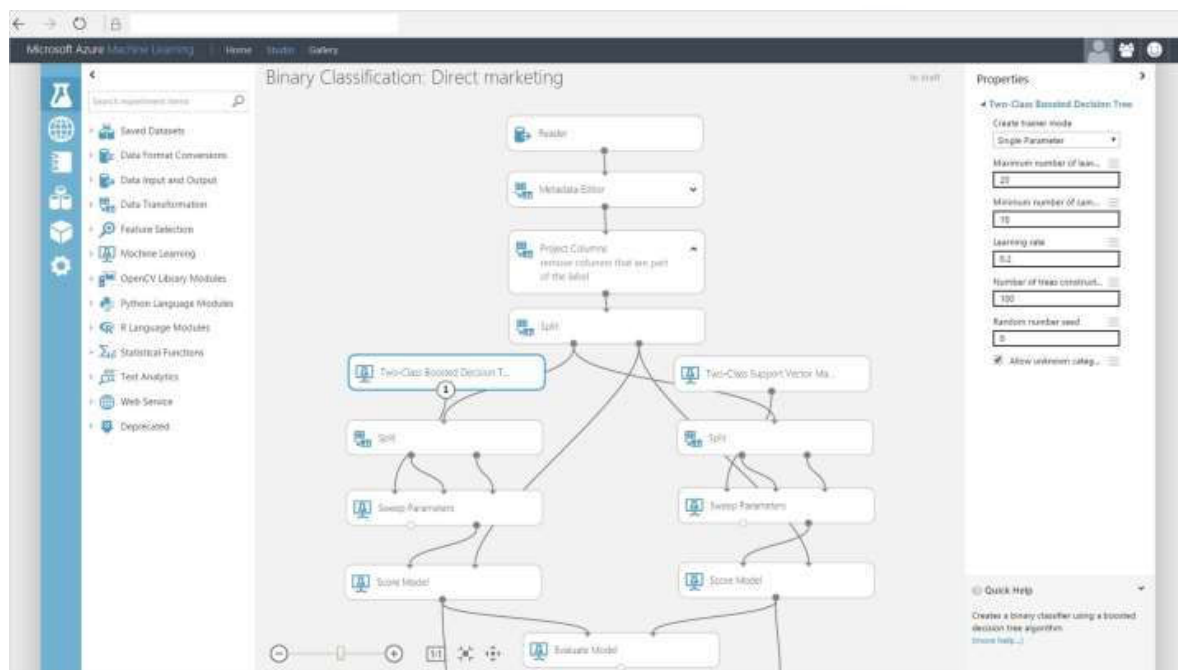


O que é o Azure Machine Learning?



Data Science Academy

www.datascienceacademy.com.br

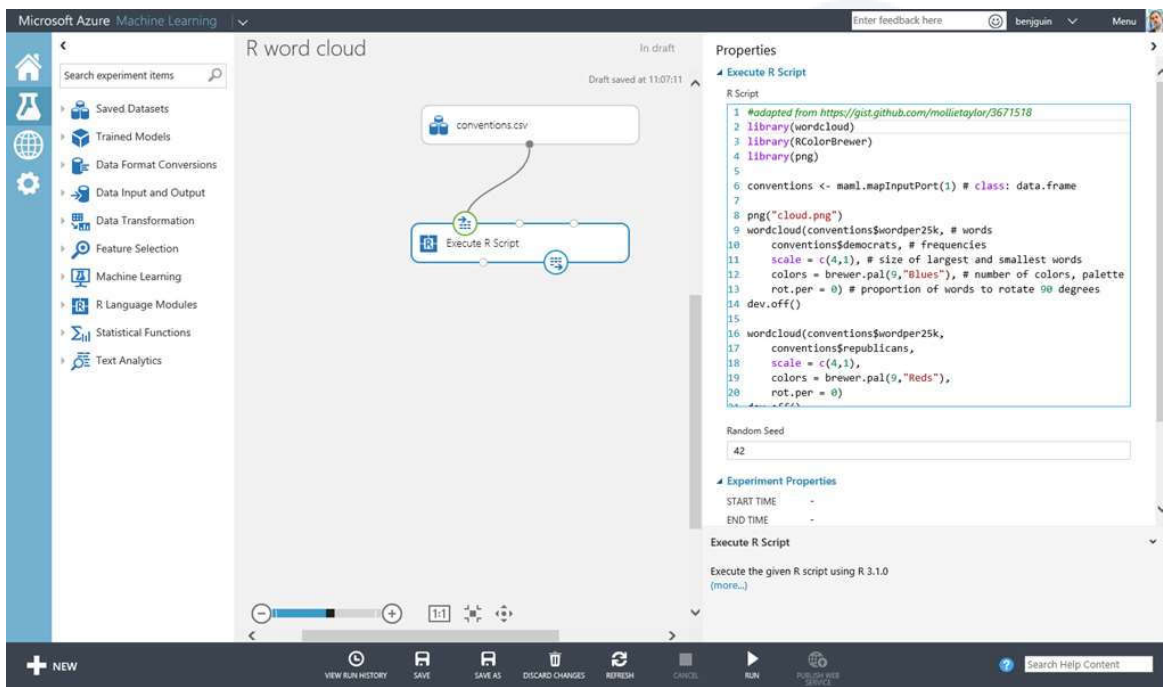


Tudo na Nuvem



Data Science Academy

www.datascienceacademy.com.br



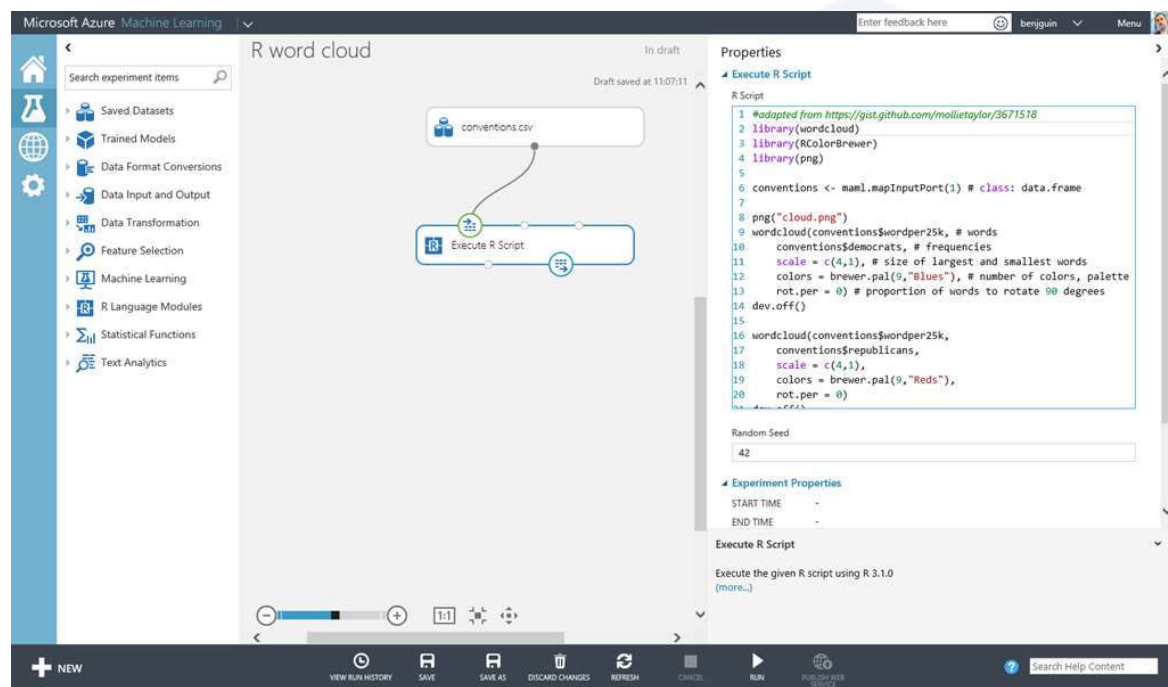
Machine Learning Studio

KNIME



Data Science Academy

www.datascienceacademy.com.br



Módulos prontos
para análises com
R, Python e SQL



Data Science Academy

www.datascienceacademy.com.br



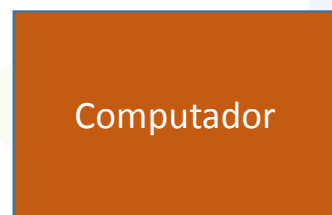
Machine Learning traz um novo paradigma



Data Science Academy

www.datascienceacademy.com.br

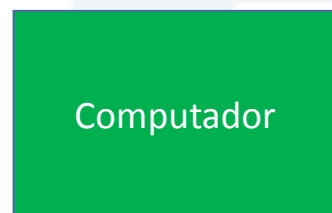
Dados
Programas



Output

Programação
Tradicional

Dados
Output



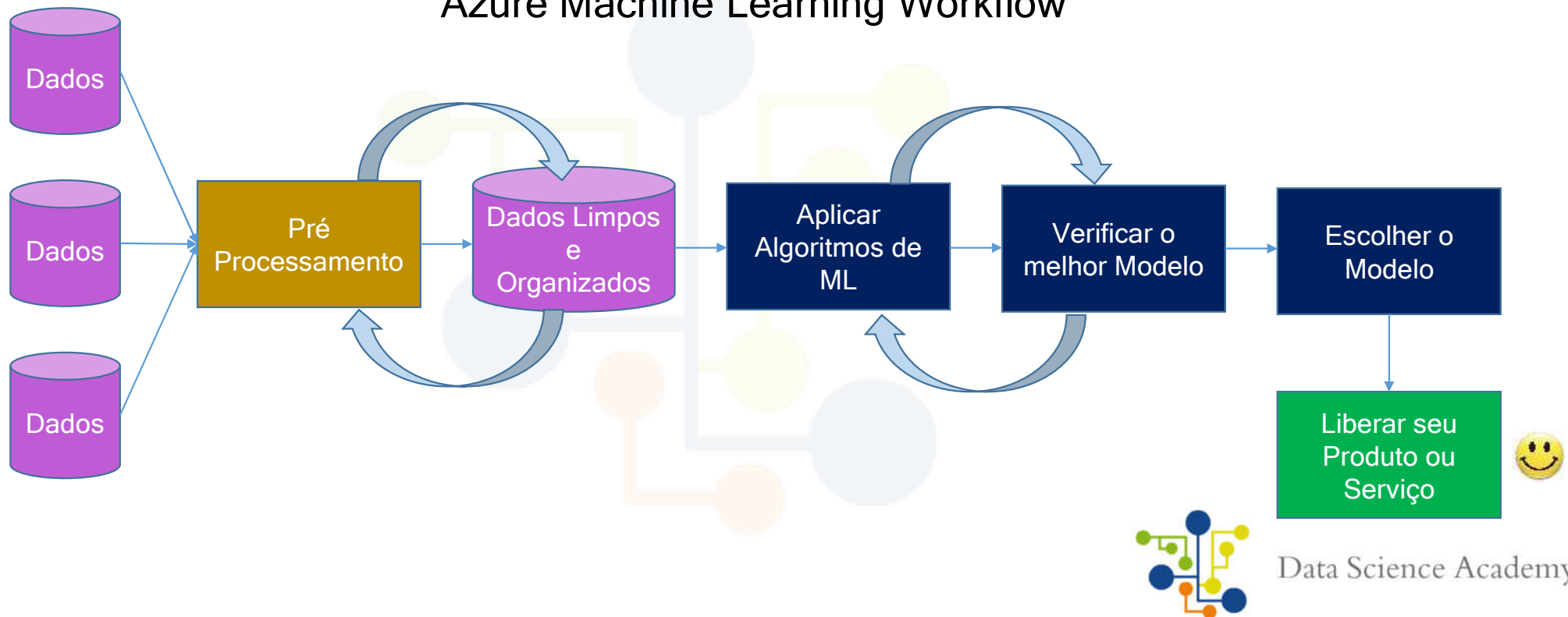
Programa

Machine
Learning

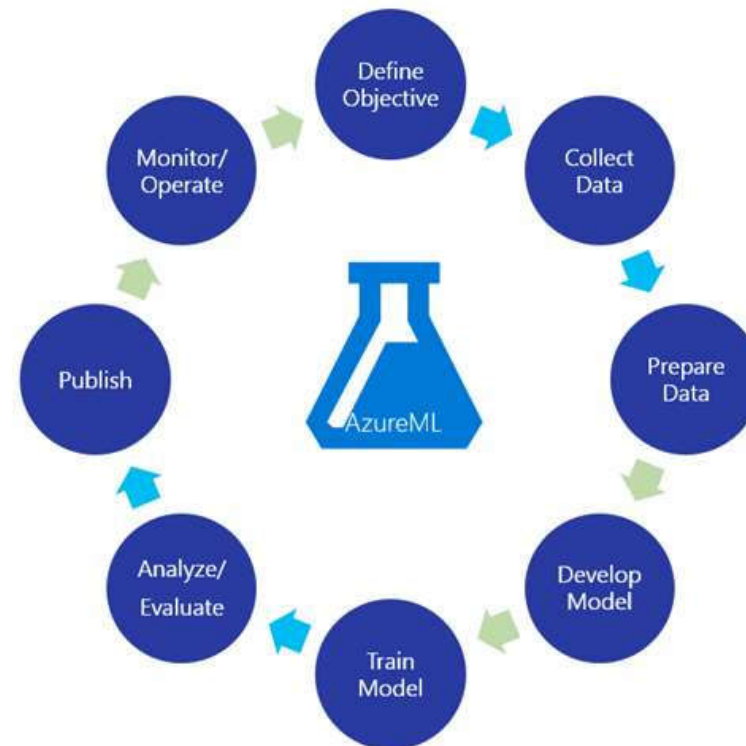


Data Science Academy

Azure Machine Learning Workflow



Azure Machine Learning Workflow



Data Science Academy



Qual a importância do Big Data Analytics?



Data Science Academy

www.datascienceacademy.com.br



O que é Big Data Analytics?

O objetivo é simples: melhorar seus processos de trabalho e adquirir insights valiosos acerca das tendências de mercado, comportamento dos consumidores e suas expectativas



Data Science Academy



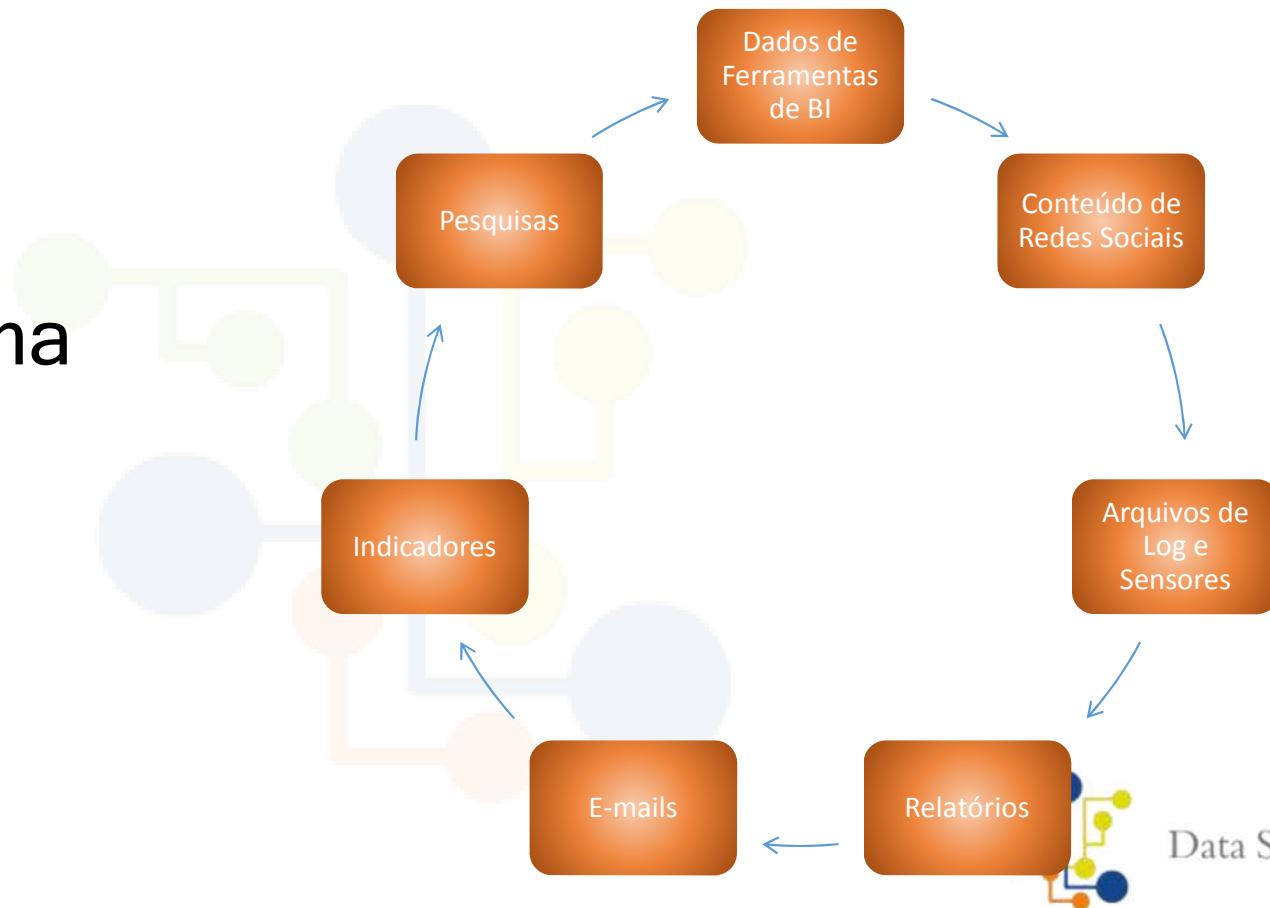
O que é Big Data Analytics?

Big Data Analytics é o trabalho analítico e inteligente em grandes volumes de dados, estruturados ou não-estruturados, que são coletados, armazenados e interpretados por softwares de altíssimo desempenho



Data Science Academy

Matéria-prima



Data Science Academy




Vantagens e Benefícios

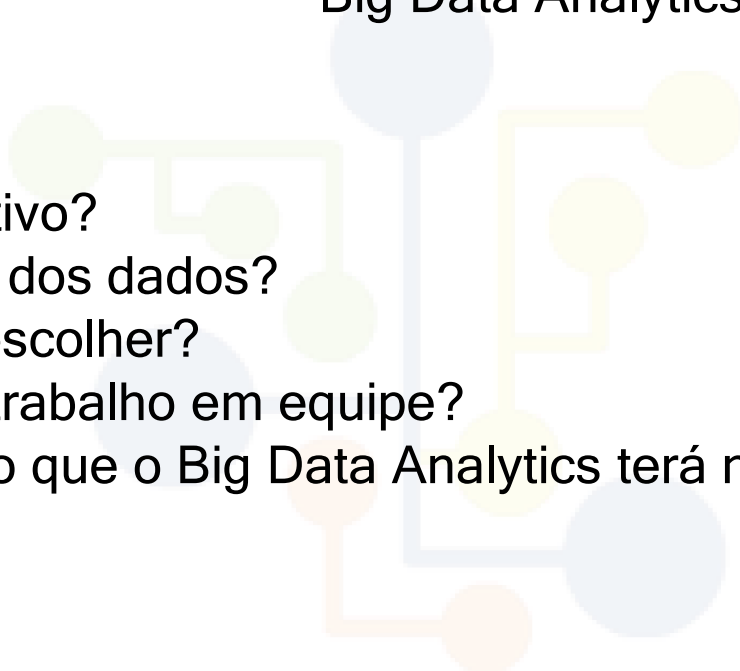
- **Direcionamento das Vendas**
- **Aperfeiçoamento do Processo de Logística**
- Atendimento mais eficiente
- Melhoria na Gestão de Recursos Humanos
- Identificação de Padrões
- Análise da Concorrência



Data Science Academy



5 Perguntas que Precisam ser respondidas antes de pensar em Big Data Analytics

- 
- 1- Qual seu objetivo?
 - 2- Qual a origem dos dados?
 - 3- Que solução escolher?
 - 4- Este será um trabalho em equipe?
 - 5- Qual o impacto que o Big Data Analytics terá no negócio?



Data Science Academy



Como o Big Data pode me ajudar a aumentar o Market Share da empresa?



Data Science Academy

www.datascienceacademy.com.br



Coletar Dados

Faturamento

Marketing

Clientes

Custos

Efetividade das Campanhas de Marketing

Concorrentes

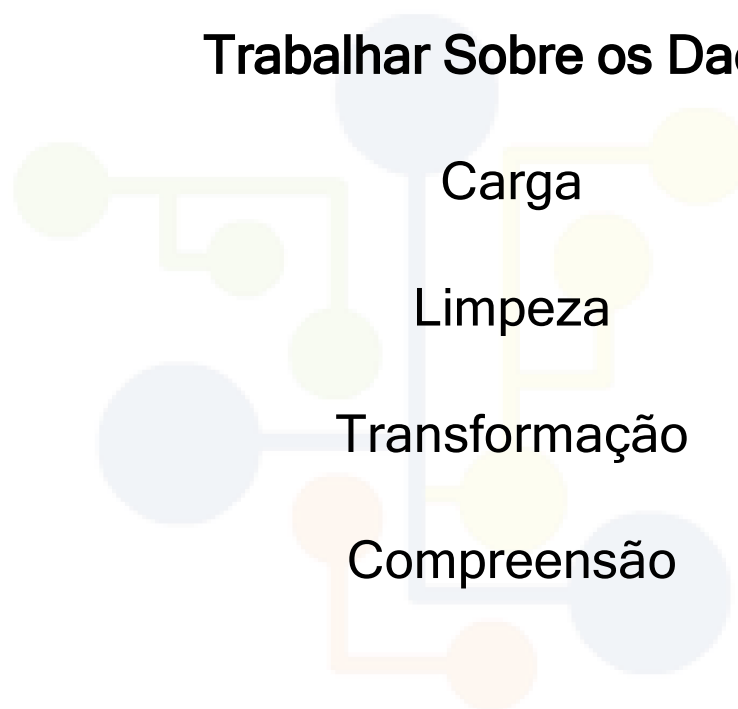
Redes Sociais



Data Science Academy



Trabalhar Sobre os Dados



Data Science Academy



Técnicas de Análise

Aplicar modelos estatísticos e compreender o relacionamento entre os dados

Definir variáveis de observação e explanatórias

Buscar correlação e causalidade



Data Science Academy



Machine Learning

Juntar tudo e criar um modelo de machine learning, prevendo como estas variáveis afetam umas às outras quando alteradas

Automatizar o processo



Data Science Academy



Apresentar seus Resultados



Data Science Academy

www.datascienceacademy.com.br



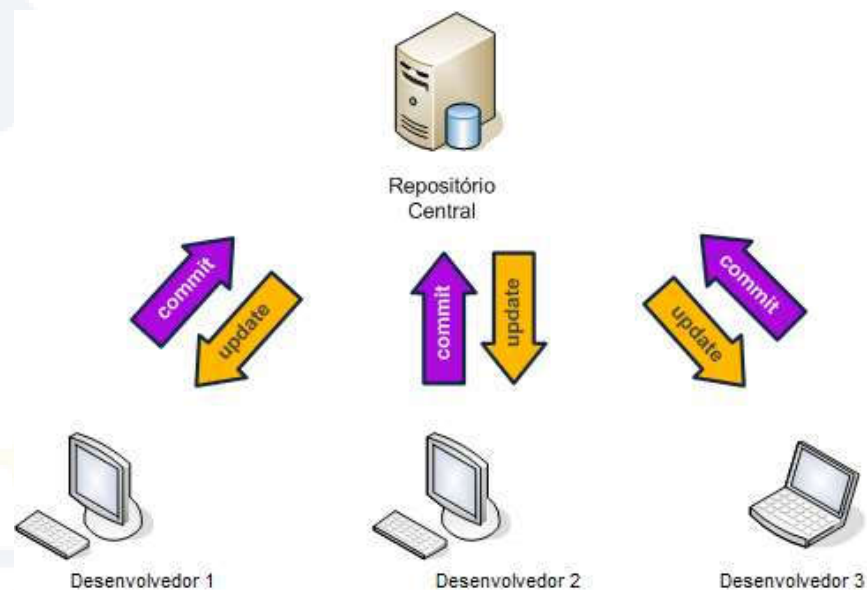
Usando o Github



Data Science Academy

www.datascienceacademy.com.br

Várias pessoas
trabalhando nos
mesmos arquivos



Data Science Academy



Sistemas de Controle de Versão

- *Concurrent Versions System (CVS)*
- *Subversion (SVN)*
- *Visual SourceSafe (VSS)*
- *Rational ClearCase*
- *Git*



Data Science Academy



git

- Não depender de um servidor central
- Dar ênfase à velocidade
- Integridade dos dados
- Potencializar o trabalho paralelo



Data Science Academy



Principais Conceitos do Git



Data Science Academy



Principais conceitos do Git

Branch

Ramificação do projeto,
cada *branch* representa uma versão do
seu projeto e podemos seguir uma linha de
desenvolvimento a partir de cada *branch*



Data Science Academy



Principais conceitos do Git

Clone

Cópia local de todos os arquivos de um repositório git



Data Science Academy



Principais conceitos do Git

Commit

Coleção de alterações realizadas, é uma espécie de *checkpoint*, sempre que necessário você pode retroceder até algum *commit* existente



Data Science Academy



Principais conceitos do Git

Fork

Uma bifurcação do projeto, uma cópia do projeto existente para seguir uma nova direção



Data Science Academy



Principais conceitos do Git

Master

Branch padrão de um repositório Git



Data Science Academy



Principais conceitos do Git

Merge

É a capacidade de incorporar alterações do git, quando acontece uma junção de diferentes *branches*



Data Science Academy



Principais conceitos do Git

Pull

Puxa as alterações do repositório remoto



Data Science Academy



Principais conceitos do Git

Push

Empurra as suas alterações para o
repositório remoto



Data Science Academy



Principais conceitos do Git

Repositório

Local onde ficam todos os arquivos do projeto, inclusive o histórico e versões



Data Science Academy

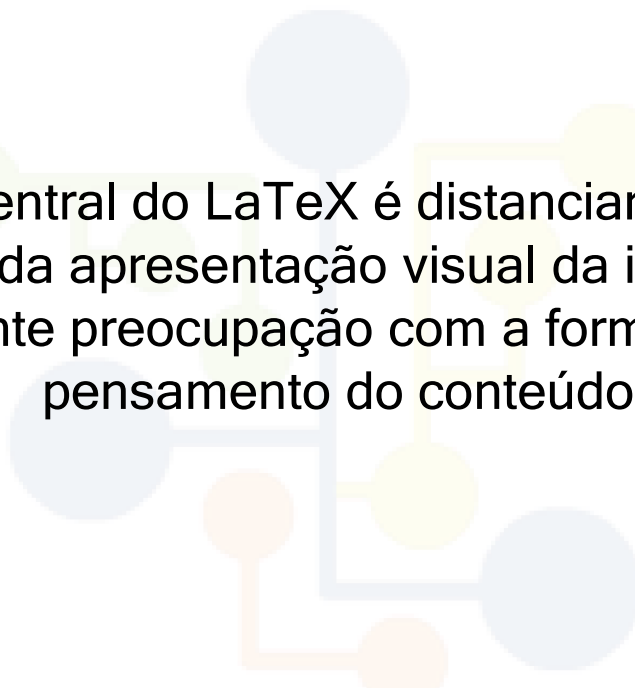



Preparação de Documentos com R e LaTeX



Data Science Academy

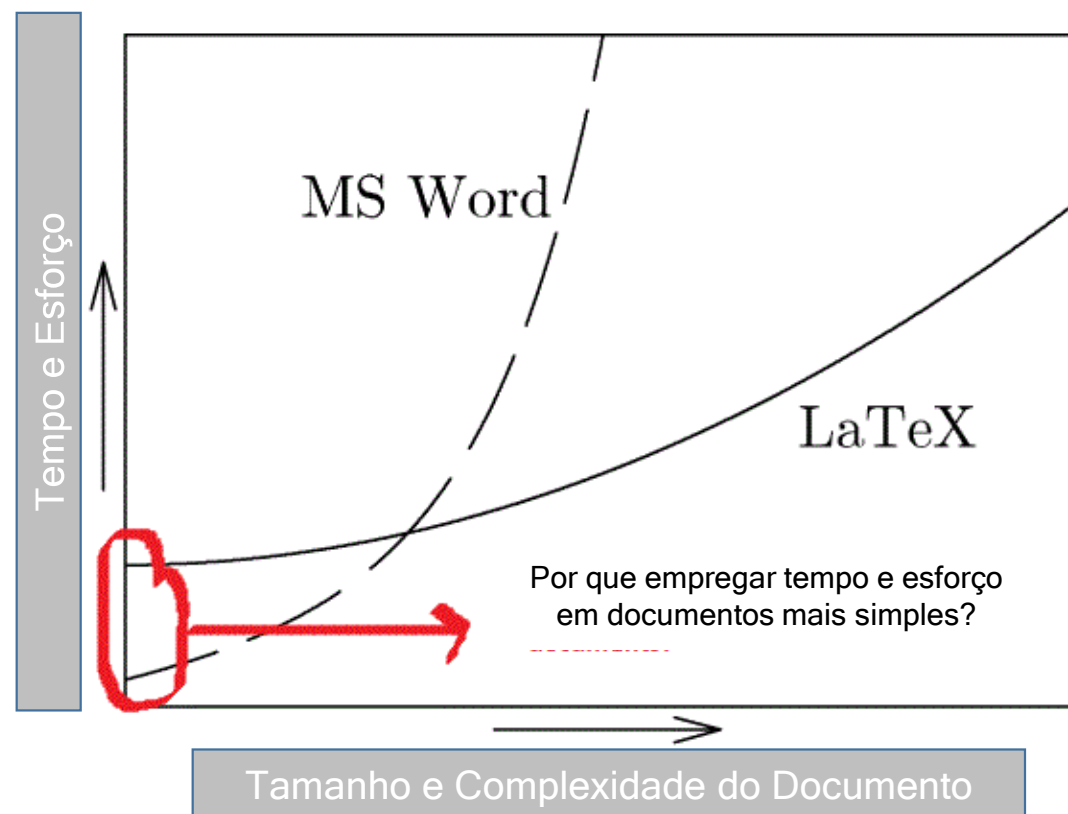
www.datascienceacademy.com.br



A ideia central do LaTeX é distanciar o autor o máximo possível da apresentação visual da informação, pois a constante preocupação com a formatação desvia o pensamento do conteúdo escrito



Data Science Academy



Data Science Academy



R e LaTeX

Arquivos com extensão .Rnw



Data Science Academy

www.datascienceacademy.com.br



R e LaTeX

Sweave()
knit()

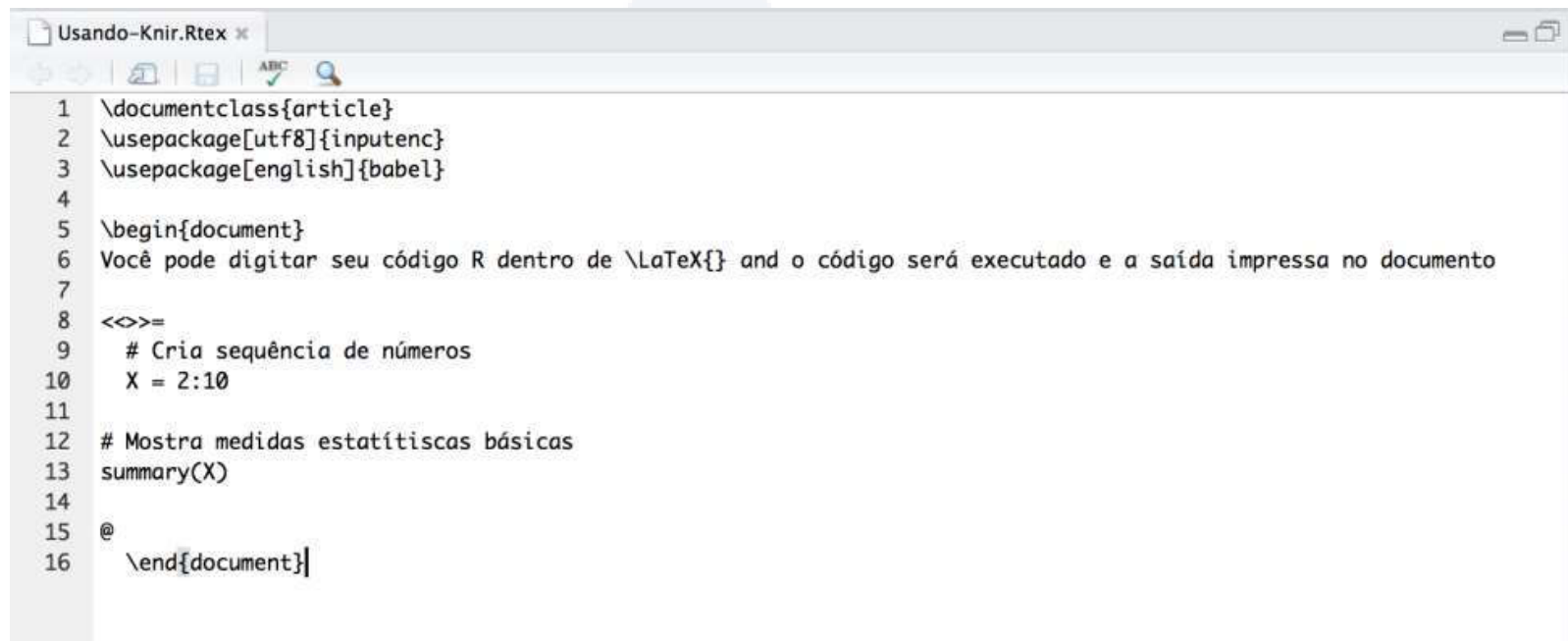
Reproducible Research



Data Science Academy

www.datascienceacademy.com.br

knitr()

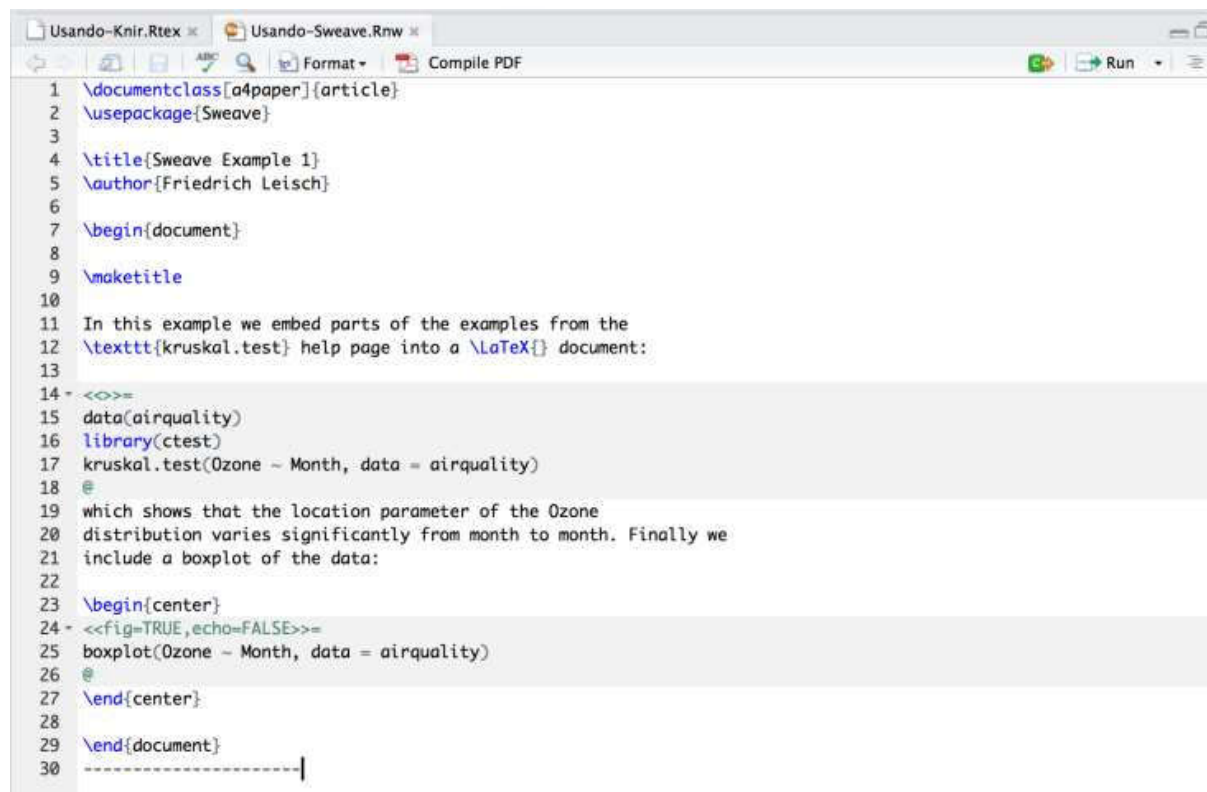


```
1 \documentclass{article}
2 \usepackage[utf8]{inputenc}
3 \usepackage[english]{babel}
4
5 \begin{document}
6 Você pode digitar seu código R dentro de \LaTeX{} and o código será executado e a saída impressa no documento
7
8 <>=
9   # Cria sequência de números
10  X = 2:10
11
12  # Mostra medidas estatísticas básicas
13  summary(X)
14
15 @
16 \end{document}
```



Data Science Academy

Sweave()



The screenshot shows a text editor window with two tabs: 'Usando-Knitr.Rtex' and 'Usando-Sweave.Rnw'. The active tab is 'Usando-Sweave.Rnw'. The editor contains a LaTeX document structure with Sweave integration. The code is as follows:

```
1 \documentclass[a4paper]{article}
2 \usepackage{Sweave}
3
4 \title{Sweave Example 1}
5 \author{Friedrich Leisch}
6
7 \begin{document}
8
9 \maketitle
10
11 In this example we embed parts of the examples from the
12 \texttt{kruskal.test} help page into a \LaTeX{} document:
13
14 <<=>
15 data(airquality)
16 library(ctest)
17 kruskal.test(Ozone ~ Month, data = airquality)
18 @
19 which shows that the location parameter of the Ozone
20 distribution varies significantly from month to month. Finally we
21 include a boxplot of the data:
22
23 \begin{center}
24 <<fig=TRUE,echo=FALSE>>=
25 boxplot(Ozone ~ Month, data = airquality)
26 @
27 \end{center}
28
29 \end{document}
30 -----|
```