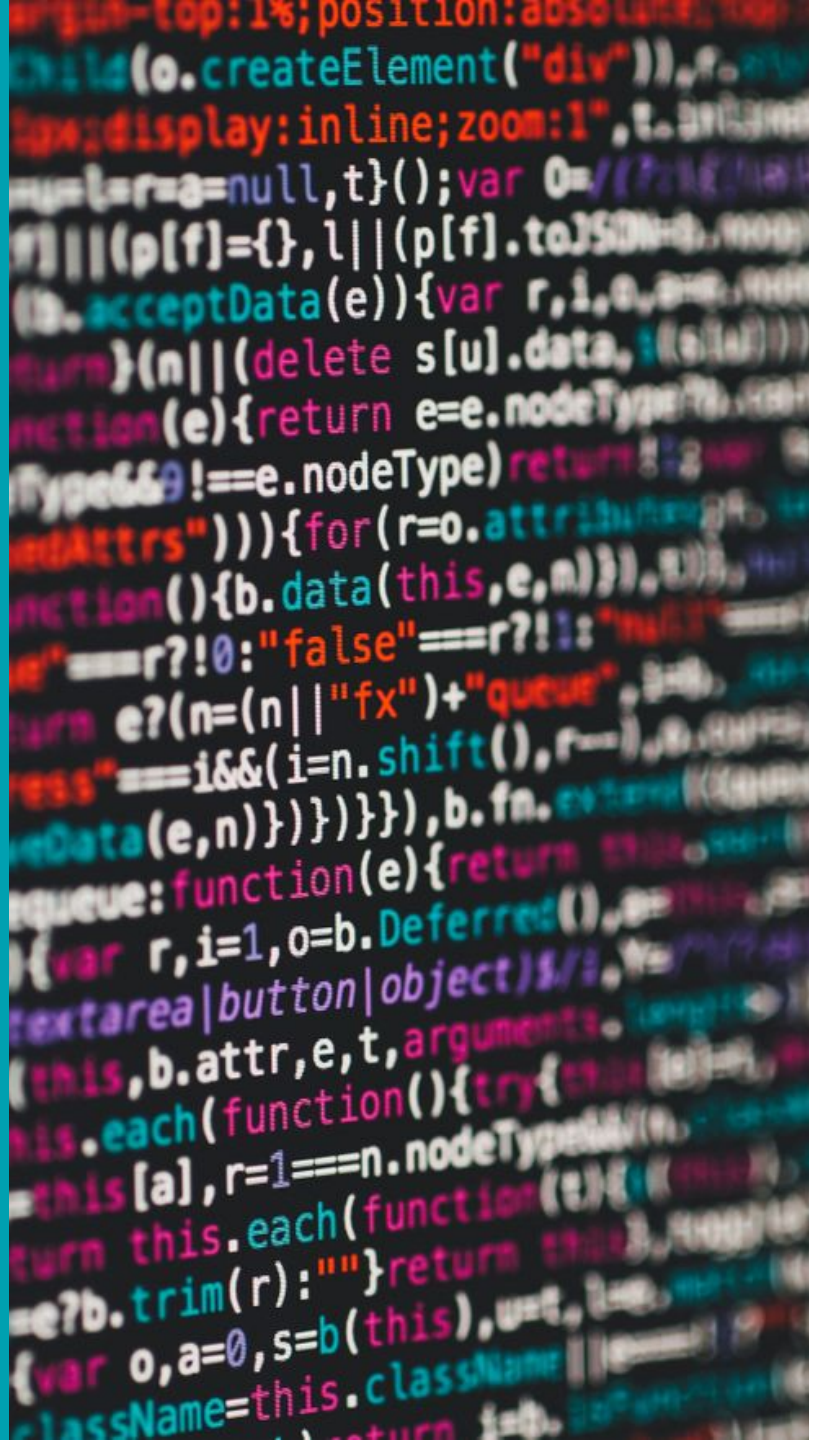


Extração e Visualização de Dados

Ramon Lopes
rlopes@ufrb.edu.br

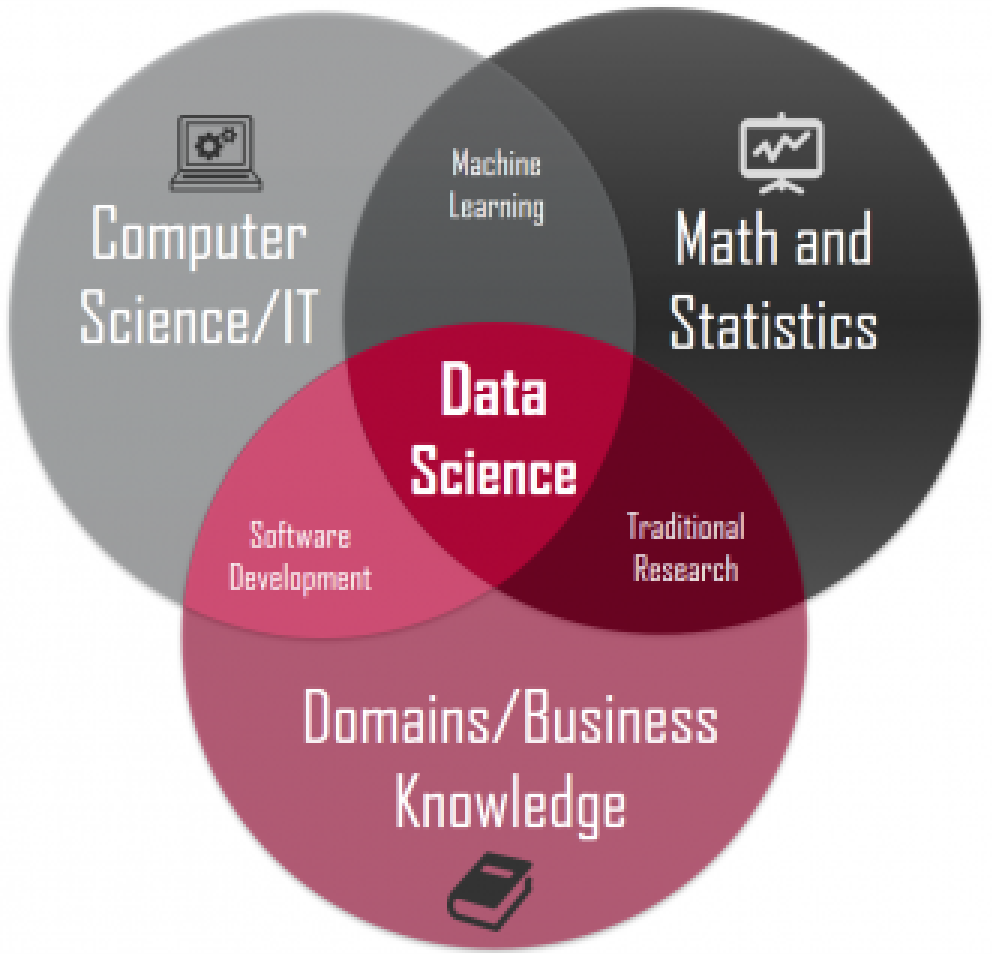
“If You're Not Paying For It, You Become The Product”

https://www.forbes.com/sites/marketshare/2012/03/05/if-youre-not-paying-for-it-you-become-the-product/#7a4eb82d5d6e

“The world’s most valuable resource is no longer oil, but data”

https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

Extração de informação e insights



Extração de Dados

Exemplos

- Extração de publicações no Lattes pra analisar rede de colaboração
- Extração de preços de produtos para construir um comparador de preços
- Coleta de preços de imóveis para alimentar base de uma imobiliária
- Coleta de postagens sobre uma marca para gerenciamento de mídia

Uma imagem vale mais do que mil palavras

Unstructured data

The university has 5600 students.
John's ID is number 1, he is 18 years old and already holds a B.Sc. degree.
David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

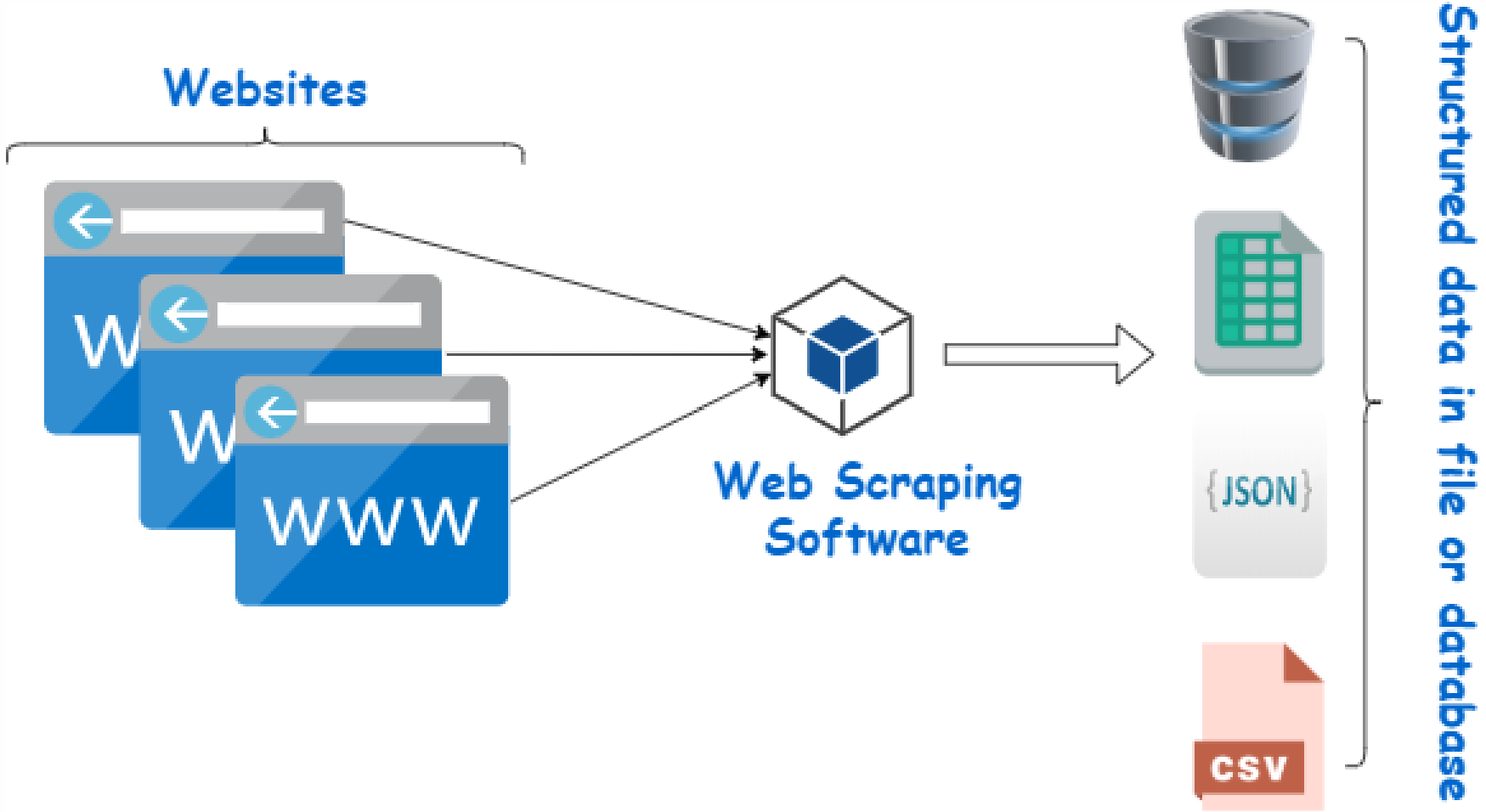
```
<University>
  <Student ID="1">
    <Name>John</Name>
    <Age>18</Age>
    <Degree>B.Sc.</Degree>
  </Student>
  <Student ID="2">
    <Name>David</Name>
    <Age>31</Age>
    <Degree>Ph.D. </Degree>
 </Student>
....
</University>
```

Structured data

ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

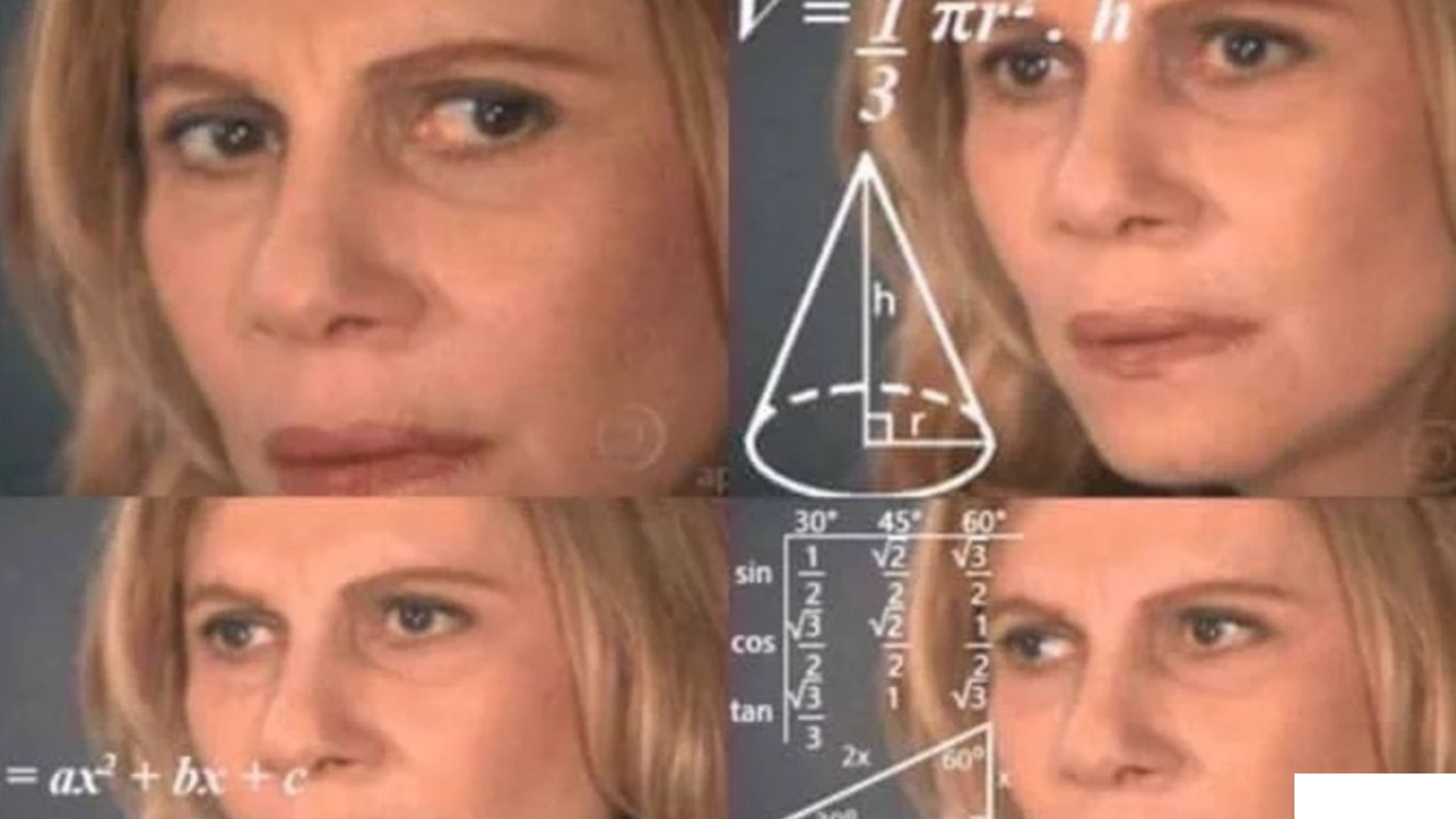
Como extrair dados da Web?

Web Scraping

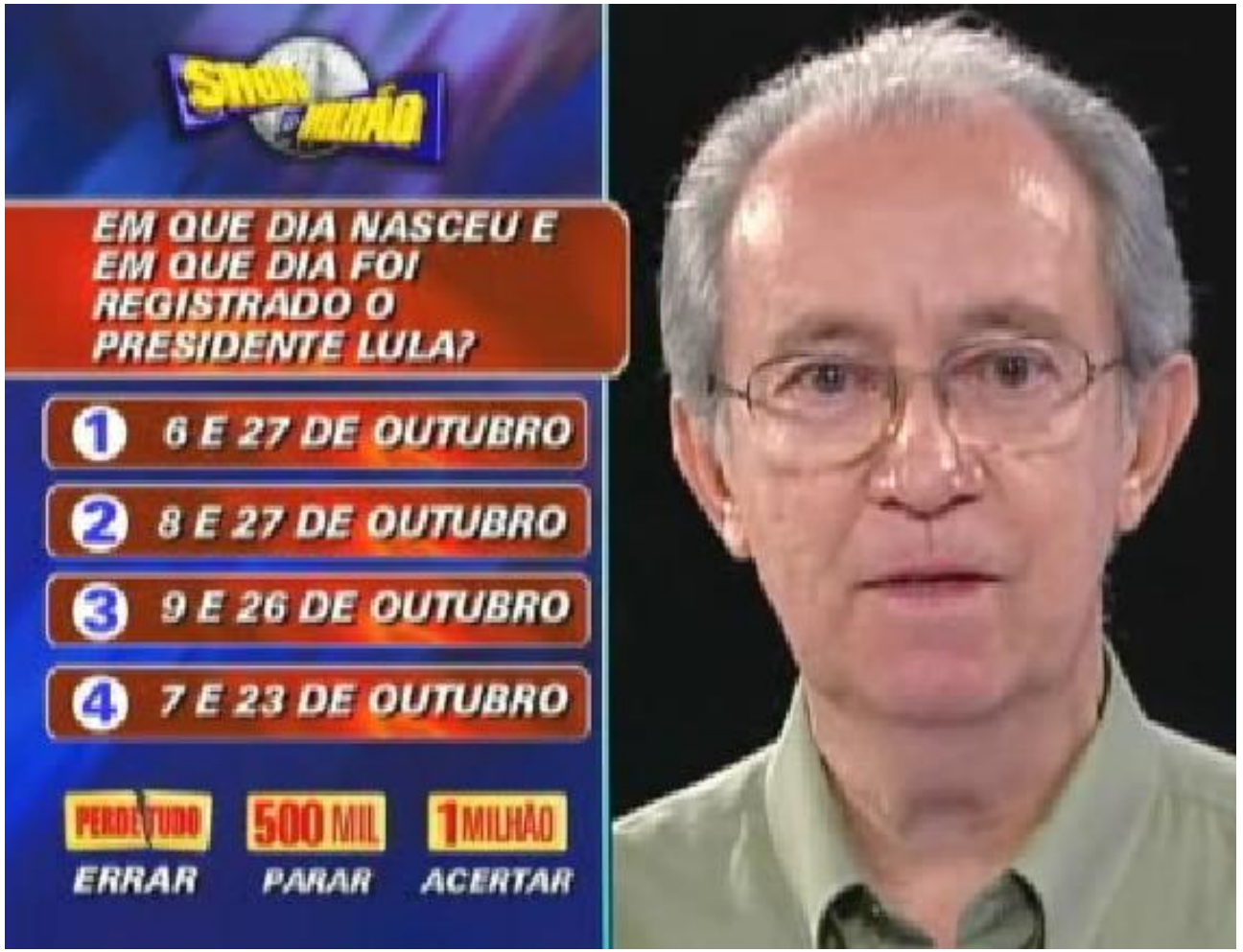


Hypertext Markup Language (HTML)

- Linguagem de *marcação* para páginas na web
- Descreve a *estrutura* de uma página web
- Consiste de um conjunto de componentes
- Estrutura aninhada de elementos


$V = \frac{1}{3}\pi r^2 \cdot h$
h
r
$= ax^2 + bx + c$
30°
45°
60°
sin
cos
tan
$\frac{1}{2}$
$\frac{\sqrt{3}}{2}$
$\frac{\sqrt{3}}{3}$
$\frac{\sqrt{2}}{2}$
$\frac{\sqrt{2}}{2}$
1
$\frac{\sqrt{3}}{2}$
$\frac{1}{2}$
$\sqrt{3}$
2x
60°

Como coletar e extrair dados de uma página HTML?



Beautiful Soup

Beautiful Soup

- Biblioteca em Python
- Navegação na estrutura HTML
- Busca de tags por atributos
- Extração de dados das tags

Buscando tags

- from bs4 import BeautifulSoup
- html = ' 20,00 5.00 '
- soup = BeautifulSoup(html)
- soup.find('div')

  20,00

- soup.findAll('div')

  [20,00 , 5.00]

Recuperando o texto

- from bs4 import BeautifulSoup
- html = ' 20,00 5.00 '
- soup = BeautifulSoup(html)
- preco_div = soup.find('div', {'id': 'preco'})

  20,00

- preco = preco_div.get_text()

  '20,00'

Hora de aplicar!

Estudo de Caso - Fundo de Investimento Imobiliário

- Queremos coletar a categoria, cotação e rendimento pago no último mês
- Vamos extrair dados do site www.fiis.com.br/
- Precisamos entender a estrutura da página

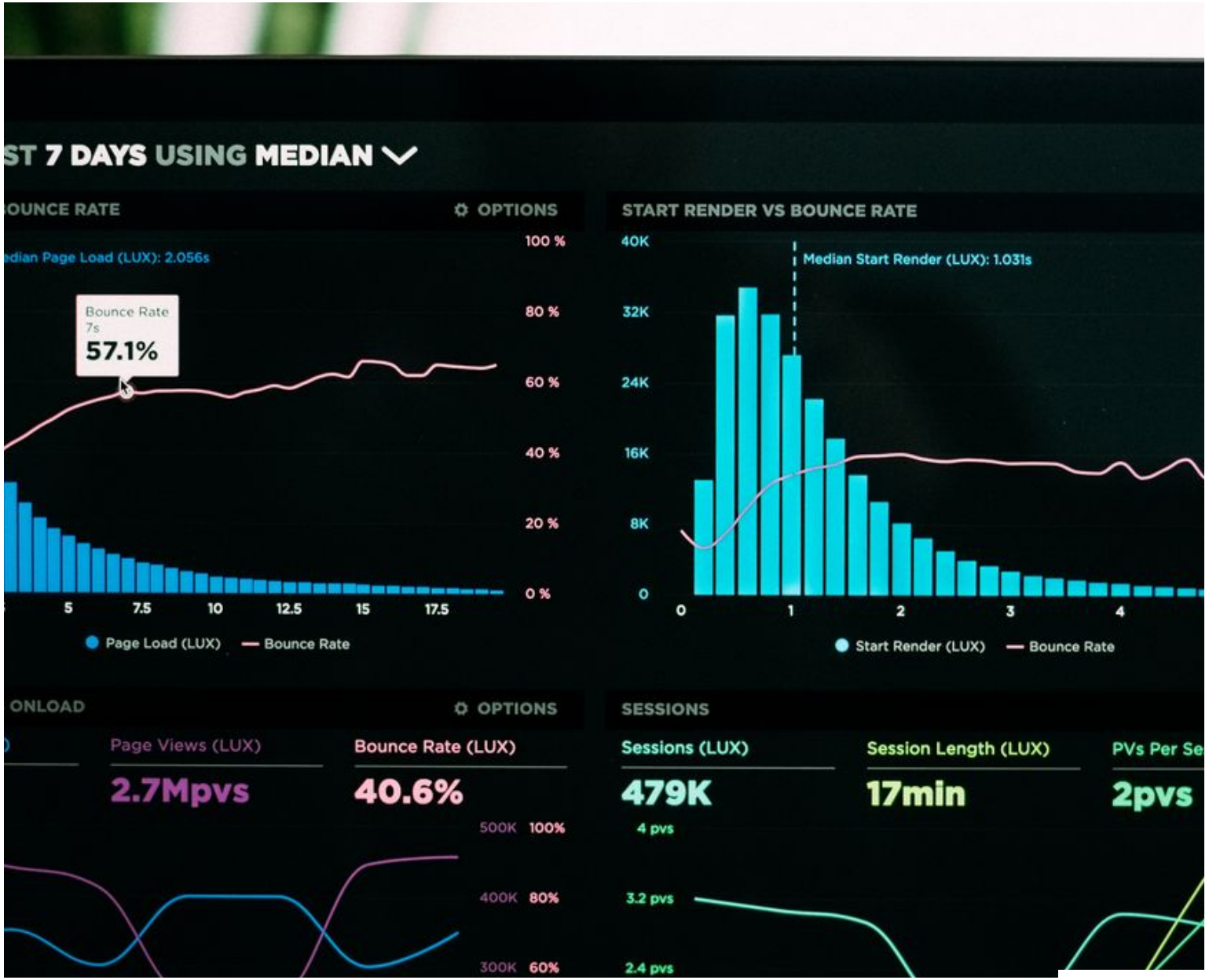
  Firefox e Chrome: F12

ÚLTIMOS REND. DO BBPO11

Data Base	Data Pagamento	Cotação Base	DY	Rendimento
30/06/20	14/07/20	R$ 164,00	0,65%	R$ 1,06

Como visualizar os dados?

Dashboard



Google Data Studio

- Ferramenta do Google para visualização de dados
- Integração de dados de várias fontes
- Fornece vários tipos de visualização

Vamos criar um dashboard para visualizar os dados coletados no Google Data Studio

Finalizando

- Web Scraping permite extrair dados de páginas Web
- Visualização de dados é fundamental para gerar insights e tomada de decisão

Próximos Passos

- **Navegar por links**
- **Selenium pode ser usado para scraping**
- **Uso de crawlers mais robustos**

  Scrapy, Apache Nutch
- **Outras ferramentas de visualização**

  Tableau, Power BI, Kibana, Metabase

Desafio

- Coletar todas cotações e rendimentos disponíveis por FII
- Apresentar os dados temporais no Google Data Studio

Muito Obrigado!

https://github.com/rlopes404