

COVID-19: Aquisição, tratamento e visualizações interativas de dados do Ministério da Saúde

Alexandre Ribeiro Cajazeira Ramos
Jonnison Lima Ferreira

VIII Escola Regional de Computação Ceará, Maranhão, Piauí
XIII Encontro Unificado de Computação do Piauí
<https://ercemapi2020.enucompi.com.br/>

10 Setembro de 2020

Cronograma de Apresentação

1 Introdução

2 Aquisição e Tratamento de Dados

3 Visualização de Dados

Cronograma de Apresentação

1 Introdução

2 Aquisição e Tratamento de Dados

3 Visualização de Dados

Pandemia Covid

- 2020 vivemos a pandemia de Covid
- Internet mais que nunca se tornou a principal fonte de informações
- Governo lança o site próprio para acompanhamento

Painel Ministério da Saúde

CORONAVÍRUS // BRASIL

COVID19

Painel Coronavírus

Atualizado em: 07/09/2020 18:30

Casos recuperados

3.355.564

Em acompanhamento

665.270

CASOS CONFIRMADOS

4.147.794

Acumulado

1973,8

Incidência*

10.273

Casos novos

Principais problemas

- Pouca interação
- Foco no dado diário

Cronograma de Apresentação

1 Introdução

2 Aquisição e Tratamento de Dados

3 Visualização de Dados

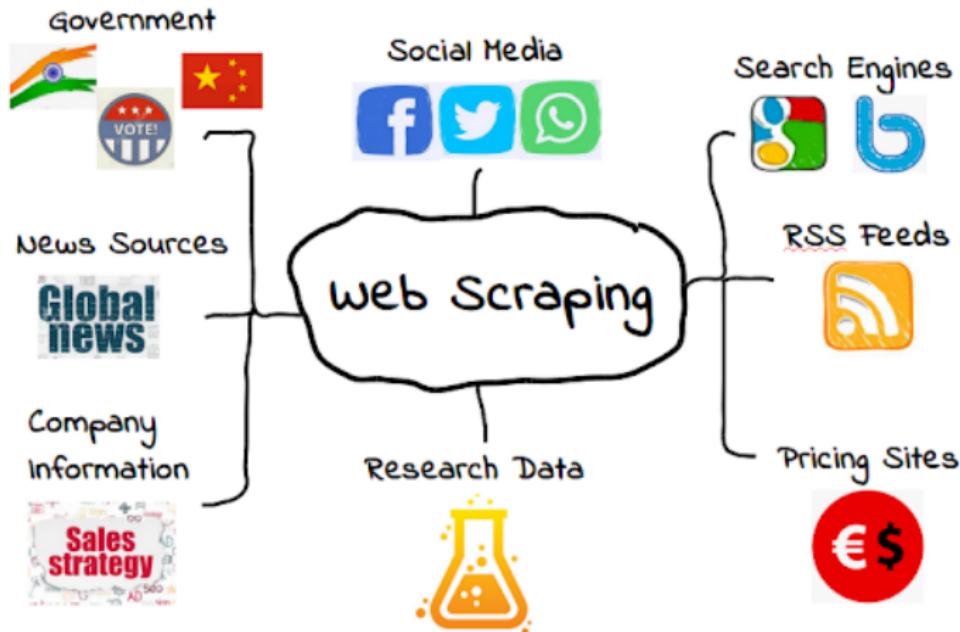
Aquisição e Tratamento de Dados

- Web Scraping
- Tratamento de dados
- Bibliotecas disponíveis
- Codificação

Web Scraping

- Você já parou para pensar como a Web mudou a maneira como consumimos informação?
- A web se tornou a maior fonte de dados para a humanidade e também para a própria computação...

Web Scraping



Web Scraping

- Web Scraping consiste na elaboração de scripts que realizam requisições HTTP que simulam um usuário acessando determinados sites, extraíndo dados e salvando de maneira organizada automaticamente.

Web Scraping - Dificuldades

- Desenvolvimento é baseado no HTML existente na página, se estrutura da página mudar, o código terá que mudar também
- O código dificilmente virá com a mesma estrutura, o desenvolvedor deverá cuidar de todas essas mudanças e métodos de processamento
- Existem dois tipos de códigos:
 - Os que não funcionam
 - Os que ainda funcionam

Web Scraping - Bibliotecas

- Existem diversas bibliotecas python que permitem a automação de requisições HTTP
 - urllib
 - requests
 - scrappy
 - httpplib2

Vamos ao código!

- Primeiro analisar o site do ministério da saúde

Vamos ao código!

- Primeiro analisar o site do ministério da saúde
- Como fazer essa requisição direta agora?

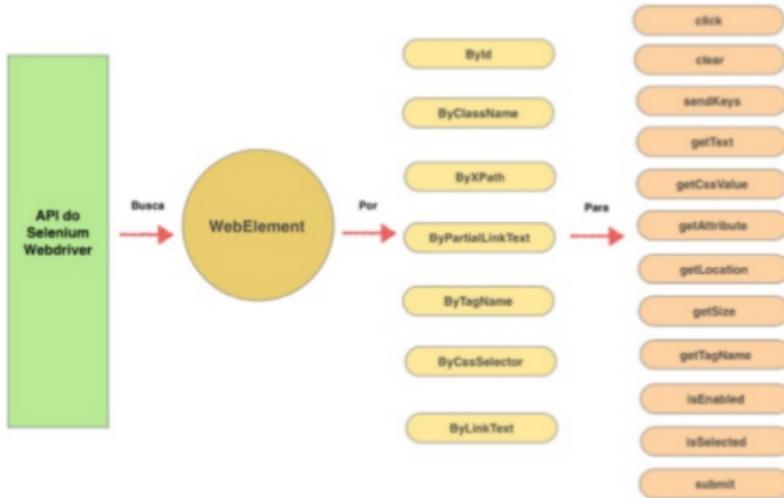
Selenium

- Ferramenta de testes automatizados de aplicações web para múltiplas plataformas
 - Windows, Linux, MAC OS
 - Firefox, Chrome, EDGE, Safari, Opera
 - Python, Java, Ruby, C, entre outras...
- Mas iremos fazer testes ou Web Scraping?

Como funciona o Selenium?



Como funciona o Selenium?



Instalando o Selenium

① pip install selenium

② wget

<https://github.com/mozilla/geckodriver/releases/download/v0.24.0/geckodriver-v0.24.0-linux64.tar.gz>

③ sudo tar -xvf geckodriver-v0.24.0-linux64.tar.gz

④ sudo mv geckodriver /usr/local/bin/

Vamos ao código?



Tratamento dos dados

- Etapa que consiste na análise dos dados para remover inconsistências
- Os dados disponibilizados pelo ministério da saúde possuem algumas inconsistências
- Formato de datas diferentes nas variadas versões
- Campos desnecessários

Pandas para tratamento dos dados

- Pandas é uma biblioteca *open source* escrita sobre o *numpy*
- Permite visualização rápida e limpeza de dados com Python
- Semelhante ao Excel
- Pode trabalhar com diversos tipos de dados
- Possui também alguns métodos de visualização
- pip install python

Vamos ao código?



Cronograma de Apresentação

1 Introdução

2 Aquisição e Tratamento de Dados

3 Visualização de Dados

Visualização de dados

- Introdução à visualização de dados
- Visualização do Min. da Saúde
- Soluções desenvolvidas
- Bibliotecas utilizadas
- Codificação

Introdução à visualização de dados

- O que é visualização?
- Por que visualizar e por que aprender visualização?

O que é visualização?



O que é visualização?

“Visualização é qualquer tipo de **representação visual** de informação projetada para permitir comunicação, análise, descoberta, exploração, etc.”

Alberto Cairo em *The Truthful Art, Cap. 1.*

O que é visualização?

- Visualização é uma área dedicada à geração de elementos visuais que auxiliem seus usuários na compreensão de dados e processos.
- Assume dois eixos primários de utilização na ciência de dados: **Exploração** e **Comunicação** de dados.
- Potencializadora da geração de *insights*, caracterizada pela interdisciplinaridade.

Fontes: *Visualization Analysis and Design* (Tamara Munzner)
Data Science do Zero (Joel Grus)

Por que visualizar e por que aprender visualização?

"A capacidade de utilizar dados – de ser capaz de entendê-los, processá-los, extrair valor deles, visualizá-los, comunicá-los
- Essa será uma habilidade importantíssima nas próximas décadas."

(Hall Varian, Google's Chief Economist)

"Gráficos ruins estão por todas as partes" e "Não somos naturalmente bons em *Storytelling* com dados"
(Cole Knaflic)

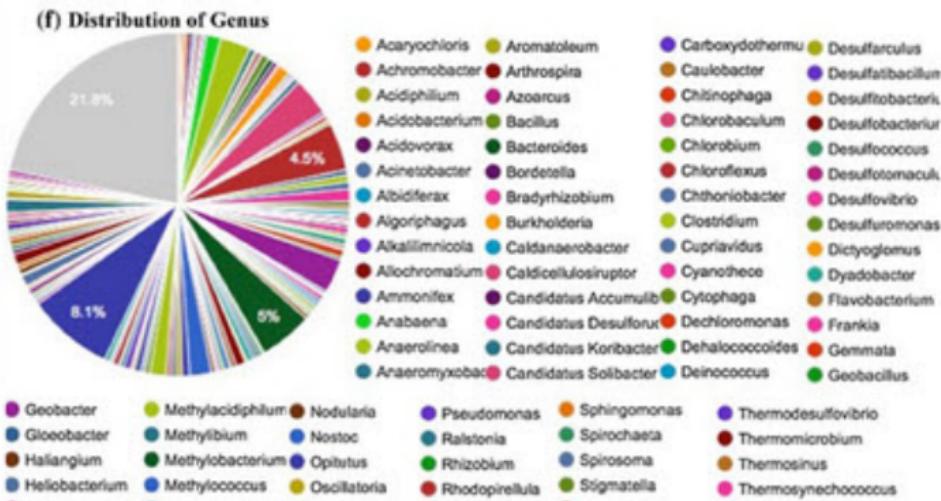
Por que visualizar e por que aprender visualização?

Analizar minuciosamente grandes conjuntos de dados é humanamente impossível.

39 State-gov	77516	Bachelors	13 Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50 Self-emp-not-inc	83311	Bachelors	13 Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38 Private	215649	HS-grad	9 Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53 Private	234721	11th	7 Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
29 Private	338409	Bachelors	13 Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37 Private	284582	Masters	14 Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49 Private	160187	9th	5 Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
52 Self-emp-not-inc	209642	HS-grad	9 Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
31 Private	45781	Masters	14 Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
42 Private	159449	Bachelors	13 Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
37 Private	280464	Some-college	10 Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
30 State-gov	141297	Bachelors	13 Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
23 Private	122272	Bachelors	13 Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
32 Private	205019	Assoc-acdm	12 Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
40 Private	121772	Assoc-voc	11 Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
34 Private	245487	7th-8th	4 Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
25 Self-emp-not-inc	176756	HS-grad	9 Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
32 Private	186824	HS-grad	9 Never-married	Machine-op-inspect	Unmarried	White	Male	0	0	40	United-States	<=50K
38 Private	28887	11th	7 Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
43 Self-emp-not-inc	292175	Masters	14 Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
40 Private	193524	Doctorate	16 Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
54 Private	302146	HS-grad	9 Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
35 Federal-gov	76845	9th	5 Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
43 Private	117037	11th	7 Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
59 Private	109015	HS-grad	9 Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K

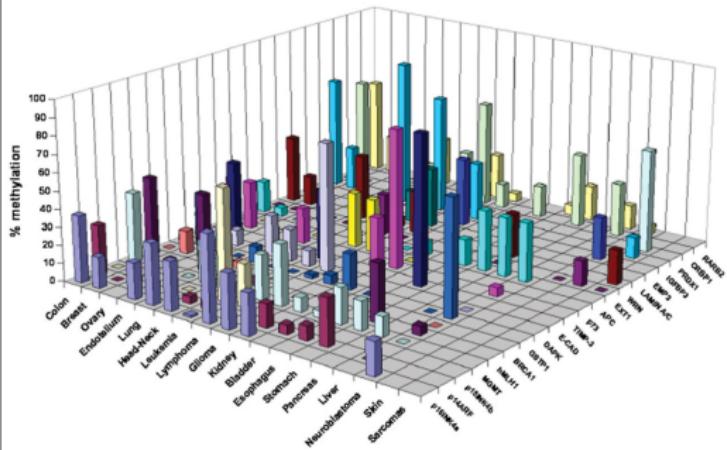
Por que visualizar e por que aprender visualização?

Visualizar por visualizar não gera contribuições, pode confundir ou desinformar.



Por que visualizar e por que aprender visualização?

A CpG Island Hypermethylation Profile of Human Cancer



Hum. Mol. Genet. (2007) 16:R50-59

Por que visualizar e por que aprender visualização?

- Quarteto de Anscombe:
- Quatro conjuntos de dados distintos com estatísticas descritivas quase idênticas, incluindo a média, variância e correlação.

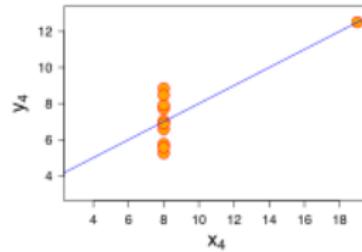
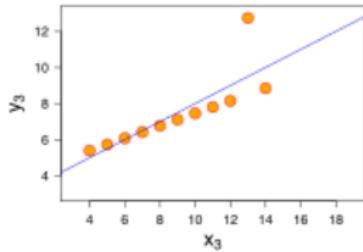
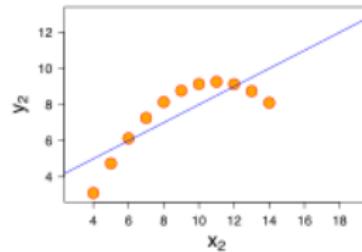
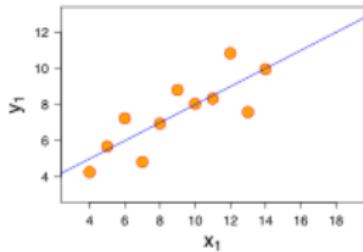
Por que visualizar e por que aprender visualização?

O que você consegue inferir sobre estes dados?

x1	y1	x2	y2	x3	y3	x4	y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Por que visualizar e por que aprender visualização?

Visualização em gráficos de dispersão do quarteto de Ascombe:



Para quem se interessar: Leitura recomendada no campo da visualização de dados

- *Storytelling with Data: A Data Visualization Guide for Business Professionals* (Cole Knaflic)
- *Truthful Art, The: Data, Charts, and Maps for Communication* (Alberto Cairo)
- Data Science do Zero: Primeiras Regras com o Python (Joel Grus)
- Como mentir com gráficos (Publicação anual do Nexo Jornal)

Painel Ministério da Saúde

CORONAVÍRUS // BRASIL

COVID19

Painel Coronavírus

Atualizado em: 07/09/2020 18:30

Casos recuperados

3.355.564

Em acompanhamento

665.270

CASOS CONFIRMADOS

4.147.794

Acumulado

1973,8

Incidência*

10.273

Casos novos

Painel Ministério da Saúde

- Exige tempo e concentração consideráveis
- Visualizações extensivas e de difícil interação
- Não permite navegação em mapas e gráficos com filtros por estado ou região
- Não é acessível à todos os públicos

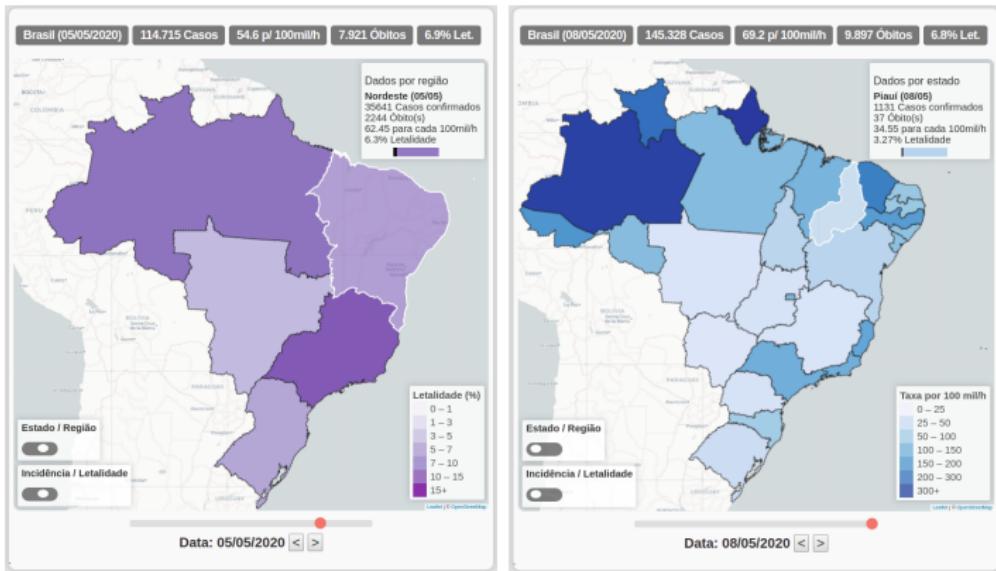
Soluções desenvolvidas: Questões-base

- "Qual o número de casos confirmados e de óbitos do meu estado?"
- "Como está a situação do meu estado em relação à sua região ou ao Brasil?"
- "Como aconteceu a evolução temporal de novos casos e óbitos em meu estado ou Região?"

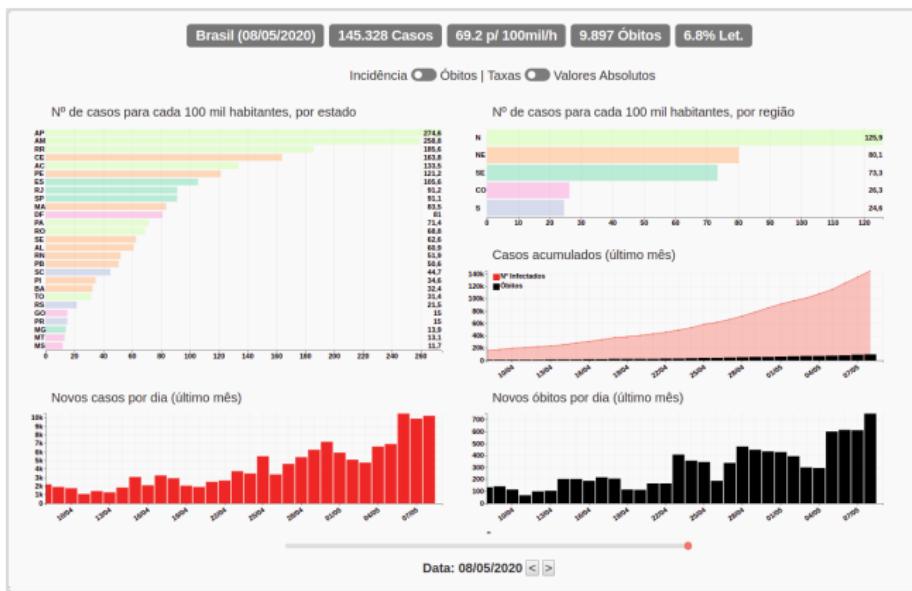
Soluções propostas

- Mapas da incidência e letalidade por regiões e estados
 - Quadro geral da pandemia nos estados e regiões
 - Evolução temporal
- *DashBoard* Interativo
 - Gráficos integrados com filtros para observação detalhada
 - Taxas e valores absolutos, casos e óbitos novos, por dia, e acumulados.
 - Evolução temporal

Mapas interativos do Coronavírus no Brasil



DashBoard Interativo do Coronavírus no Brasil



Soluções propostas: Informações adicionais

- Escala de Cores
 - <https://colorbrewer2.org/>
 - <https://color.adobe.com/>
- Mapas dos estados e regiões em GeoJson
 - <https://github.com/tbrugz/geodata-br>
- Por que utilizar a plataforma Web para desenvolver as visualizações?

Bibliotecas .js utilizadas no projeto

- D3js
 - <https://d3js.org/>
- Leaflet
 - <https://leafletjs.com/>
- DCjs e Crossfilter
 - <https://dc-js.github.io/dc.js/>
 - <https://square.github.io/crossfilter/>

Bibliotecas .js utilizadas no projeto: D3js



- Biblioteca de visualização com foco nos dados
- Variedade de exemplos disponíveis ¹
- Criação: InfoVis 2011, por Mike Bostock (@mbostock)

¹<https://observablehq.com/@d3/gallery>

Bibliotecas .js utilizadas no projeto: D3js

- Carrega dados na memória do *browser* (Não esconde dados)
- Vincula e transforma elementos (DOM) interpretando o dado associado
- Responde à interação do usuário (Transição entre estados de elementos)
- Não é responsável pela renderização de gráficos

Bibliotecas .js utilizadas no projeto: *Leaflet*

- Biblioteca *open source* para criação de mapas interativos
- Fácil utilização, documentação simples e tutoriais claros



Bibliotecas .js utilizadas no projeto: *Leaflet*

Quatro tutoriais oficiais são recomendados para a devida compreensão e replicação do painel desenvolvido:

- *Leaflet Quick Start Guide*²;
- *Using GeoJSON with Leaflet*³;
- *Interactive Choropleth Map*⁴;
- *Working with map panes*⁵.

²<https://leafletjs.com/examples/quick-start/>

³<https://leafletjs.com/examples/geojson/>

⁴<https://leafletjs.com/examples/choropleth/>

⁵<https://leafletjs.com/examples/map-panes/>

Bibliotecas .js utilizadas no projeto: *DCjs* e *Crossfilter*

- DCjs - *Dimensional Charting Javascript Library*
 - Renderização amigável (CSS/SVG) de gráficos para dados multidimensionais
 - Resposta instantânea à interação de usuários
 - Suporte Nativo Crossfilter
- Crossfilter - *Fast Multidimensional Filtering for Coordinated Views*
 - Análise e exploração extremamente rápida de conjuntos gigantescos de dados
 - "...extremely fast (-30ms) interaction with coordinated views, even with datasets containing a million or more records"

Bibliotecas .js utilizadas no projeto: *DCjs* e *Crossfilter*

- DCjs - *Dimensional Charting Javascript Library*
 - *Ordinal Line*⁶; *Ordinal Bar*⁷; *Row*⁸; *Composite*⁹
- Crossfilter - *Fast Multidimensional Filtering for Coordinated Views*
 - <https://github.com/square/crossfilter/wiki/API-Reference>

⁶<https://dc-js.github.io/dc.js/examples/ordinal-line.html>

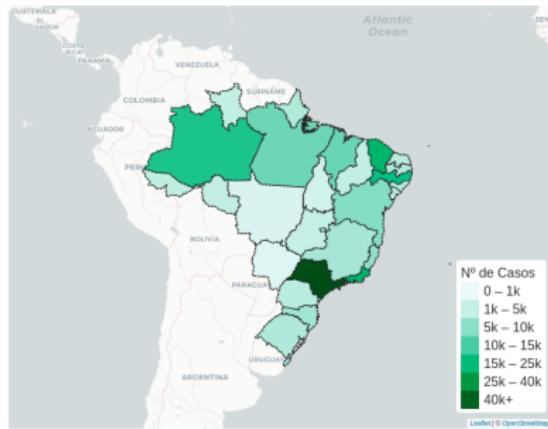
⁷ [.../examples/ordinal-bar.html](https://dc-js.github.io/dc.js/examples/ordinal-bar.html)

⁸ [.../examples/row.html](https://dc-js.github.io/dc.js/examples/row.html)

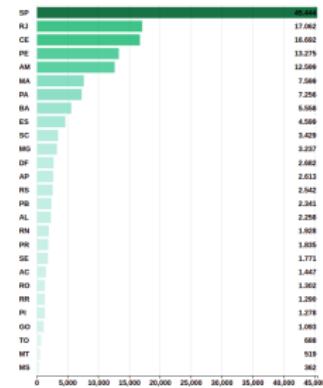
⁹ [.../examples/composite.html](https://dc-js.github.io/dc.js/examples/composite.html)

Codificação: Exemplo ilustrativo

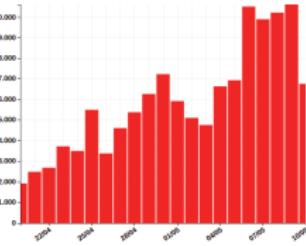
- Vamos desenvolver uma versão simplificada das visualizações para debater os principais trechos de código?



Nº de Casos por Estado (10/05/2020)



Novos casos por dia



Número acumulado de casos por Estado e novos casos por dia

Vamos ao código?



Para seguir os exemplos:

[https://github.com/
CajazeiraRamos/
ERCEMAPI2020.git](https://github.com/CajazeiraRamos/ERCEMAPI2020.git)

- Trecho de código 1: Lendo e manipulando dados com *D3js* e *Crossfilter*;
- T2: Mapa temático com o número de casos por estado *Leaflet*;
- T3: Gráficos interligados e interativos com *D3js* e *Crossfilter* e *DCjs*.

COVID-19: Aquisição, tratamento e visualizações interativas de dados do Ministério da Saúde

Alexandre Ribeiro Cajazeira Ramos
Jonnison Lima Ferreira

VIII Escola Regional de Computação Ceará, Maranhão, Piauí
XIII Encontro Unificado de Computação do Piauí
<https://ercemapi2020.enucompi.com.br/>

10 Setembro de 2020