

# Câncer de pulmão e tabagismo em tempos de mídias sociais

Vitor Horodynski, RA 206953

Carlos Marques, RA 214327

Gabriel Pereira, RA 216194

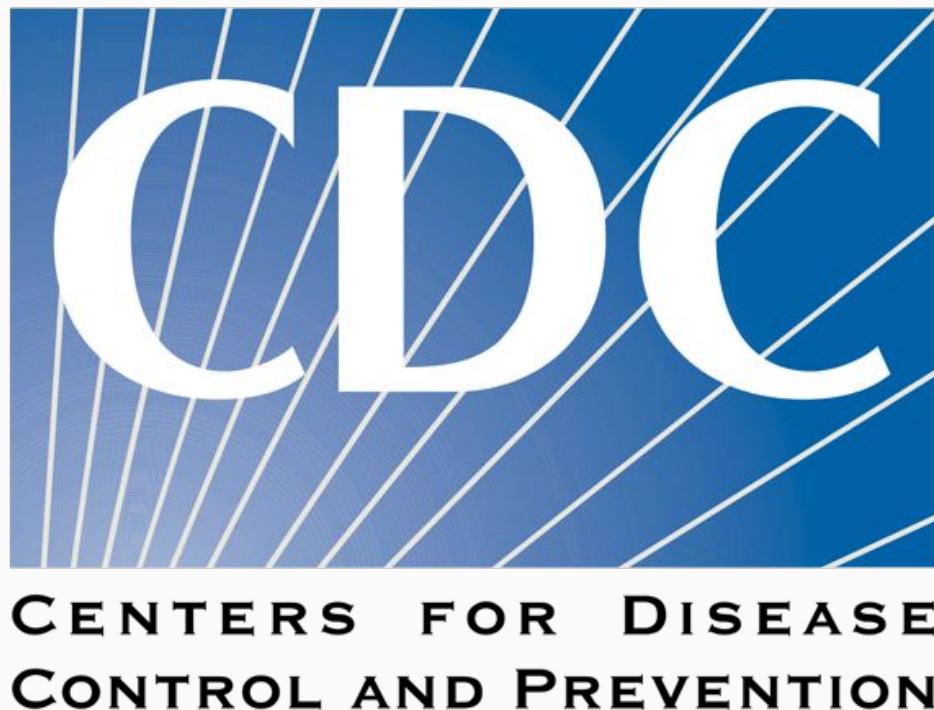
# Estágio 3

# Introdução

Tratou-se do modelo de documentos por meio da API CDC Wonder.

Esta API é um registro de mortalidade organizado por causa, ano, mês, entre outros dos EUA.

Os dados foram filtrados por ano e estado norte-americano.



# Extração de Dados

Dois modelos de extração de dados da API CDC Wonder (organizada por documentos) :

- Por meio de requests automatizados para a API:
  - Rápido, altamente configurável e reusável para outras aplicações. Entretanto, é limitada visto que só é possível retirar os dados dos últimos 10 anos e requer conversão do formato XML para CSV.
- Download manual da série histórica pelo site:
  - Necessita conversão de formato, mas fornece uma série histórica a partir de 1968.

Solução: usar os dois métodos de extração

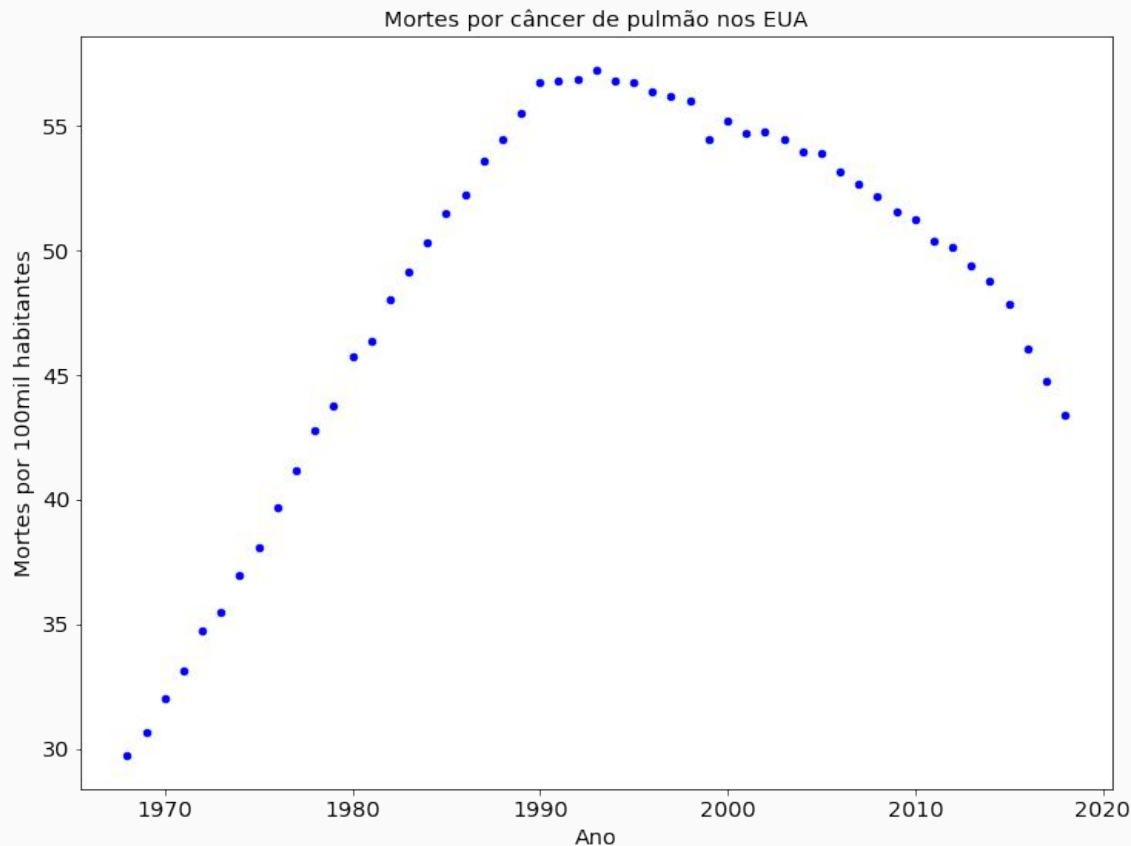
# Conversão de dados

- Requests automatizados:
  - Transformar o arquivo XML retornado pelo API em um dicionário “navegável” (dicionários contendo dicionários recursivamente) e então navegar por este dicionário escrevendo os dados relevantes para o problema
- Download manual dos dados
  - O site da API disponibiliza um arquivo de texto que é semelhante a um CSV. Como a série histórica vem separada de 10 em 10 anos aproximadamente, usou-se a biblioteca pandas para unir as tabelas já tratadas, com o objetivo de gerar apenas um arquivo com toda a série histórica de mortes por câncer de pulmão nos EUA disponível.

# Consultas SQL

A partir do tratamento de dados da API, algumas consultas com dados interessantes sobre as mortes de pulmão.

A organização dos dados anterior a este estágio possibilitou recuperar o total de mortes nacionalmente com o auxílio das consultas.



# 70

mortes por câncer de pulmão a cada 100 mil habitantes em média de 1968 a 2018 no estado de West Virginia. Em 1993, foi registrado o máximo valor desta estatística em toda a nação: 57 mortes/100 mil habitantes.

“Em 2017, 26.6% dos estudantes de ensino médio de West Virginia disseram que usam algum produto do tabaco, incluindo cigarros eletrônicos. Entre a juventude do ensino médio do estado, 14.4% disseram que fumam cigarro convencional.”

---

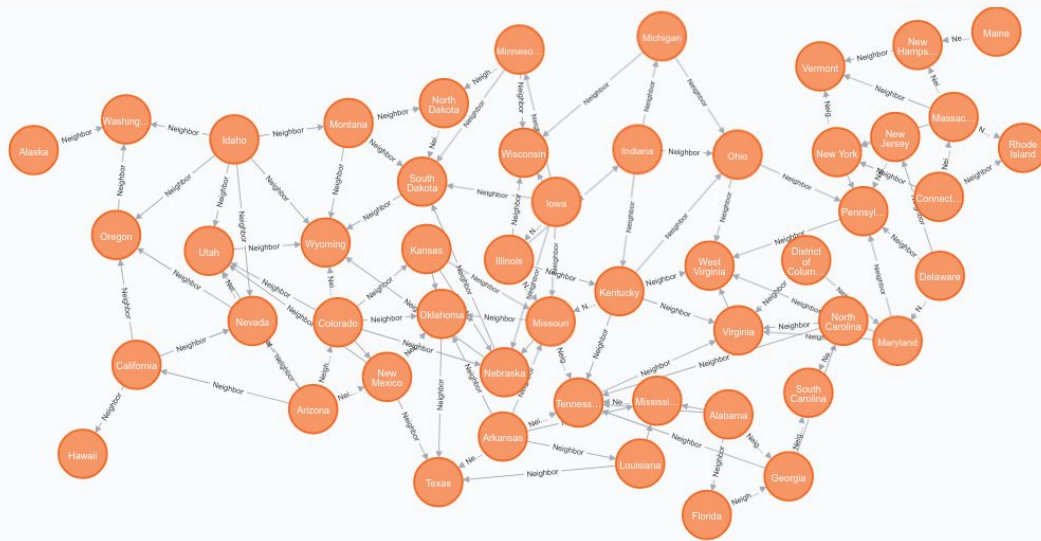
- CDC, que também considera o hábito de fumo no estado uma epidemia (<https://www.cdc.gov/tobacco/about/osh/state-fact-sheets/west-virginia/index.html>)



# Estágio 4

# Construção do Grafo

A partir de uma base de dados que lista os vizinhos dos estados norte-americanos, é possível construir um “grafo de vizinhanças” em se representa um estado por um nó e dois nós tem uma aresta se os estados que os representam são vizinhos.



# Uso do Modelo de Grafos

Apesar da possibilidade de realizar as mesmas análises sem o uso do modelo de grafos, é preciso atentar ao fato de que este modelo oferece algumas vantagens em relação ao modelo conceitual para as análises propostas:

- Facilidade de visualização dos dados, por meio da representação espacial inerente ao modelo;
- Diversas definições possíveis de relacionamento entre dois nós, como localidade, poder econômico, políticas públicas, etc;
- Conhecimento dos parâmetros de rede para um conjunto de relacionamentos escolhido, possibilitando análises posteriores.

# Meios de análise para o modelo

A partir deste grafo, e por meio dos dados de mortalidade e uso de tabaco disponíveis, é possível:

- Representar a mortalidade por câncer de pulmão em um determinado estado pelo tamanho do nó do grafo, gerando uma análise visual.
- Determinar parâmetros de conectividade da rede, como mean-betweenness, e então relacionar estes parâmetros com as redes sociais de fumantes e por pacientes de câncer de pulmão
- Percorrer o grafo em uma determinada localidade determinada por um parâmetro econômico ou social. Por exemplo, relacionar os estados do Corn Belt com suas práticas de tabagismo e presença em mídias sociais.

Obrigado!