

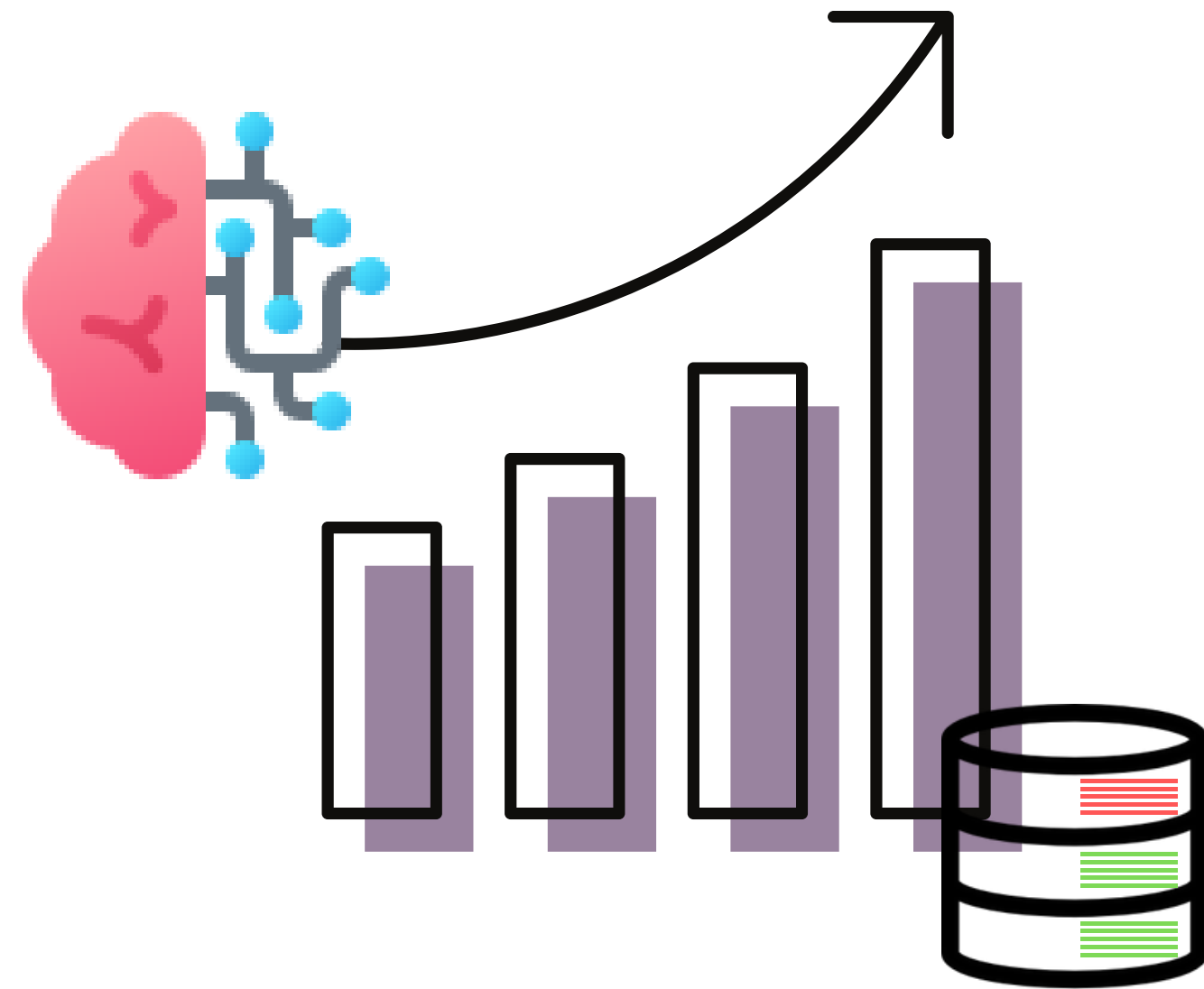
# **Detecção de Malwares Android: Levantamento Empírico da Disponibilidade e da Atualização das Fontes de Dados**

**6º Workshop Regional de Segurança da Informação e de Sistemas Computacionais**

Tainá Soares, Joner Mello, Lucas Barcellos,  
Renato Sayyed, Guilherme Siqueira,  
Karina Casola, Estevão Costa , Nicolas Gustavo,  
Diego Kreutz e Eduardo Feitosa

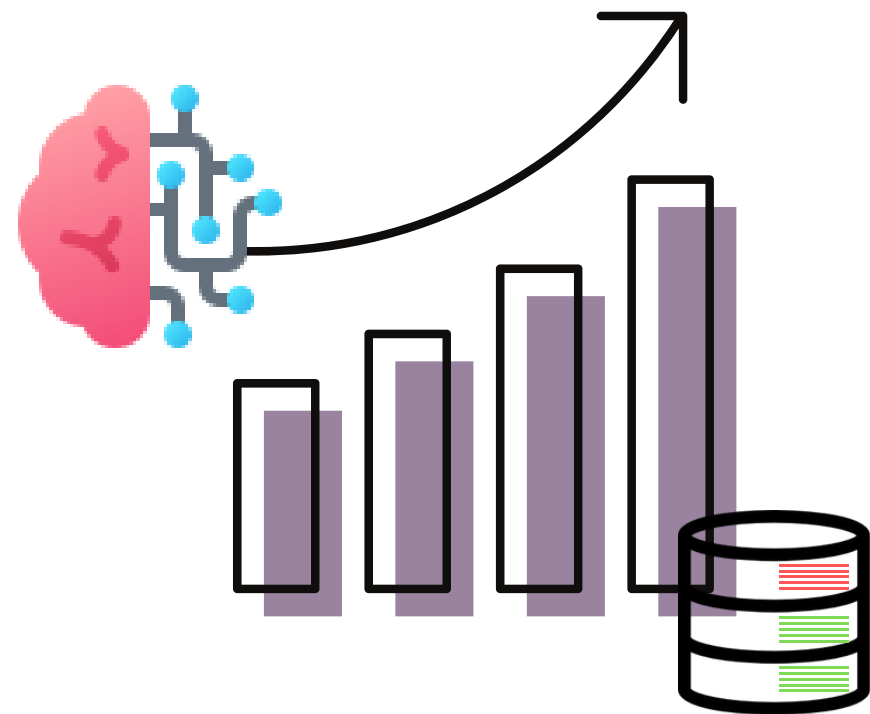


# A importância do dataset



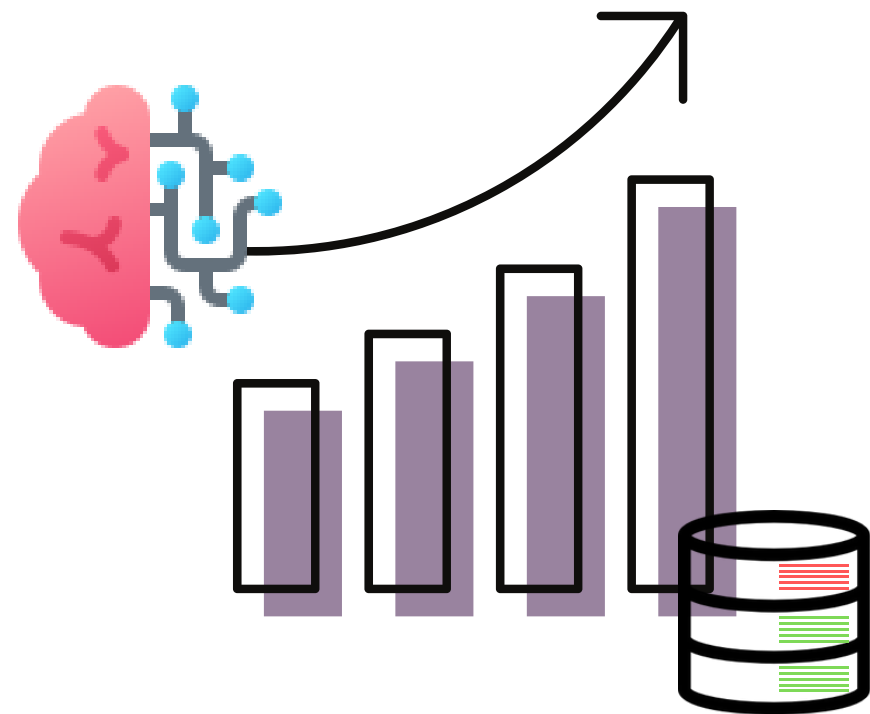
# A importância do dataset

➤ As pesquisas utilizam datasets atuais?



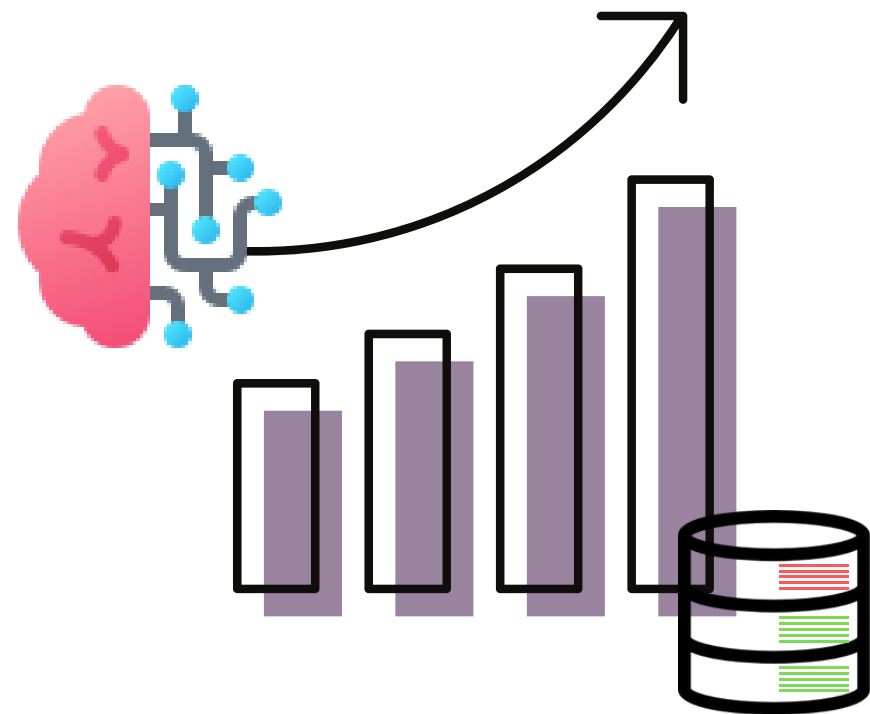
# A importância do dataset

- As pesquisas utilizam datasets atuais?
- Existem fontes de dados disponíveis e atuais?



# A importância do dataset

- As pesquisas utilizam datasets atuais?
- Existem fontes de dados disponíveis e atuais?
- É seguro considerar atual uma fonte de dados de acordo com sua data de criação ou atualização?



# Roteiro

➤ Objetivo

# Roteiro

- Objetivo
- Metodologia

# Roteiro

- Objetivo
- Metodologia
- Resultados



# Roteiro

- Objetivo
- Metodologia
- Resultados
- Conclusão

# Roteiro

- Objetivo
- Metodologia
- Resultados
- Conclusão
- Trabalhos Futuros

# Objetivos

(a) Atualidade e disponibilidade de fontes de dados

# Objetivos

- (a) Atualidade e disponibilidade de fontes de dados
- (b) Fontes de dados utilizadas na prática

# Metodologia

- Seleção dos trabalhos
- Seleção das fontes
- Análise das fontes

# Metodologia - seleção dos trabalhos

**G1** -> survey ou revisão sistemática de literatura

**G2** -> 40 citações - Google Scholar

**G3** -> conferências de segurança

**G4** -> conferências de inteligência artificial

35

# Metodologia - seleção das fontes

- > Revisão sistemática
- > 35 trabalhos
- > 100 primeiros resultados em plataformas de busca

# Metodologia - seleção das fontes

39 lojas de aplicativos

30 datasets

15 repositórios de APKs

84



# Metodologia – análise das fontes

## Atualização

- 1.** Dados informados nos sites ou trabalhos
- 2.** Trabalhos dos últimos 4 anos

# Resultados

## Lojas de Aplicativos (39)

54% disponíveis e em atualização constante

# Resultados

## Datasets e Repositórios (45)

80% disponíveis ou restritos

# Resultados




## Datasets e Repositórios (45)

36% disponíveis e atuais

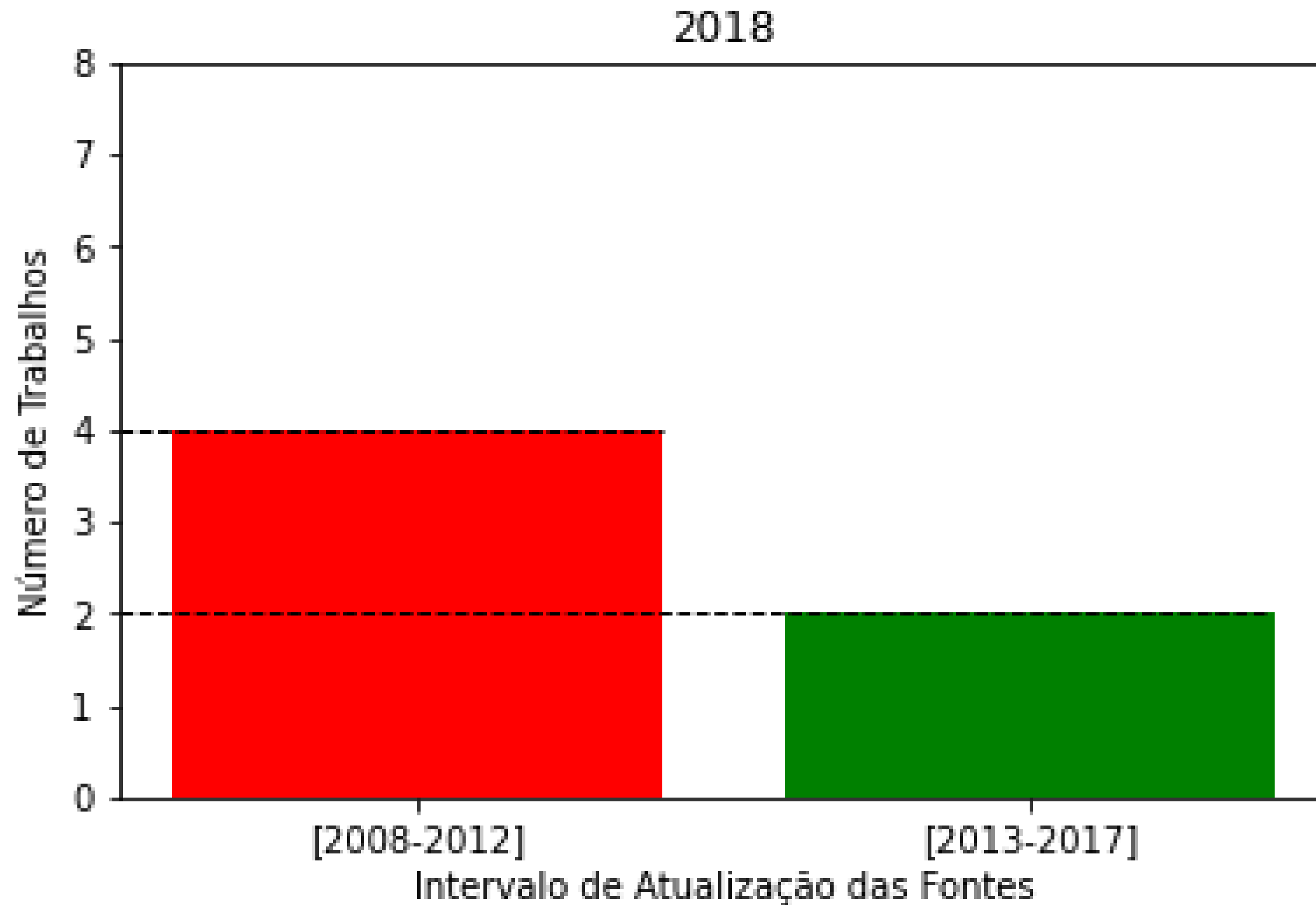
# Resultados

## Atualização

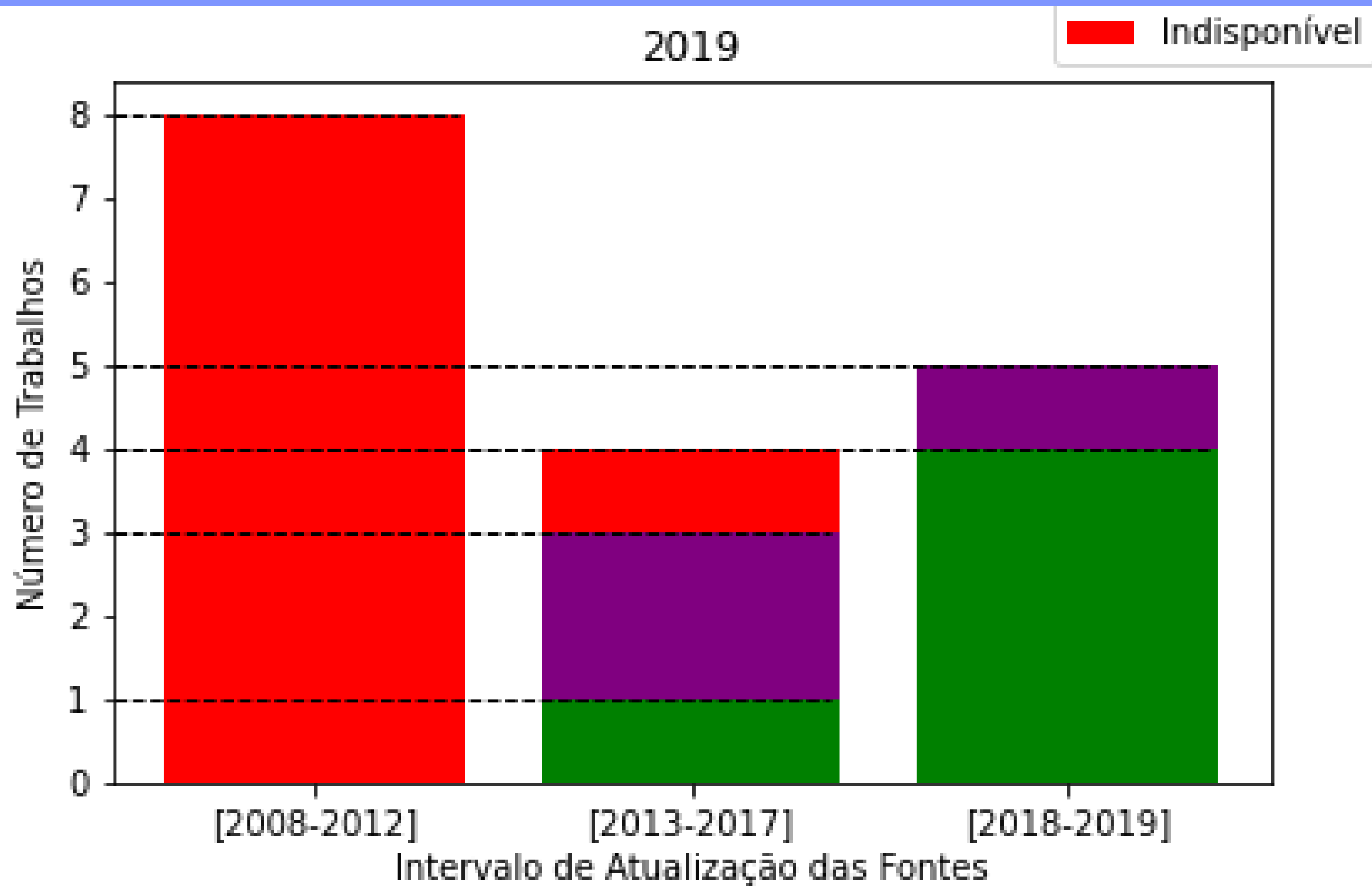
- Trabalhos dos últimos 4 anos
- Atualização das fontes

-  Disponível
-  Restrito
-  Indisponível

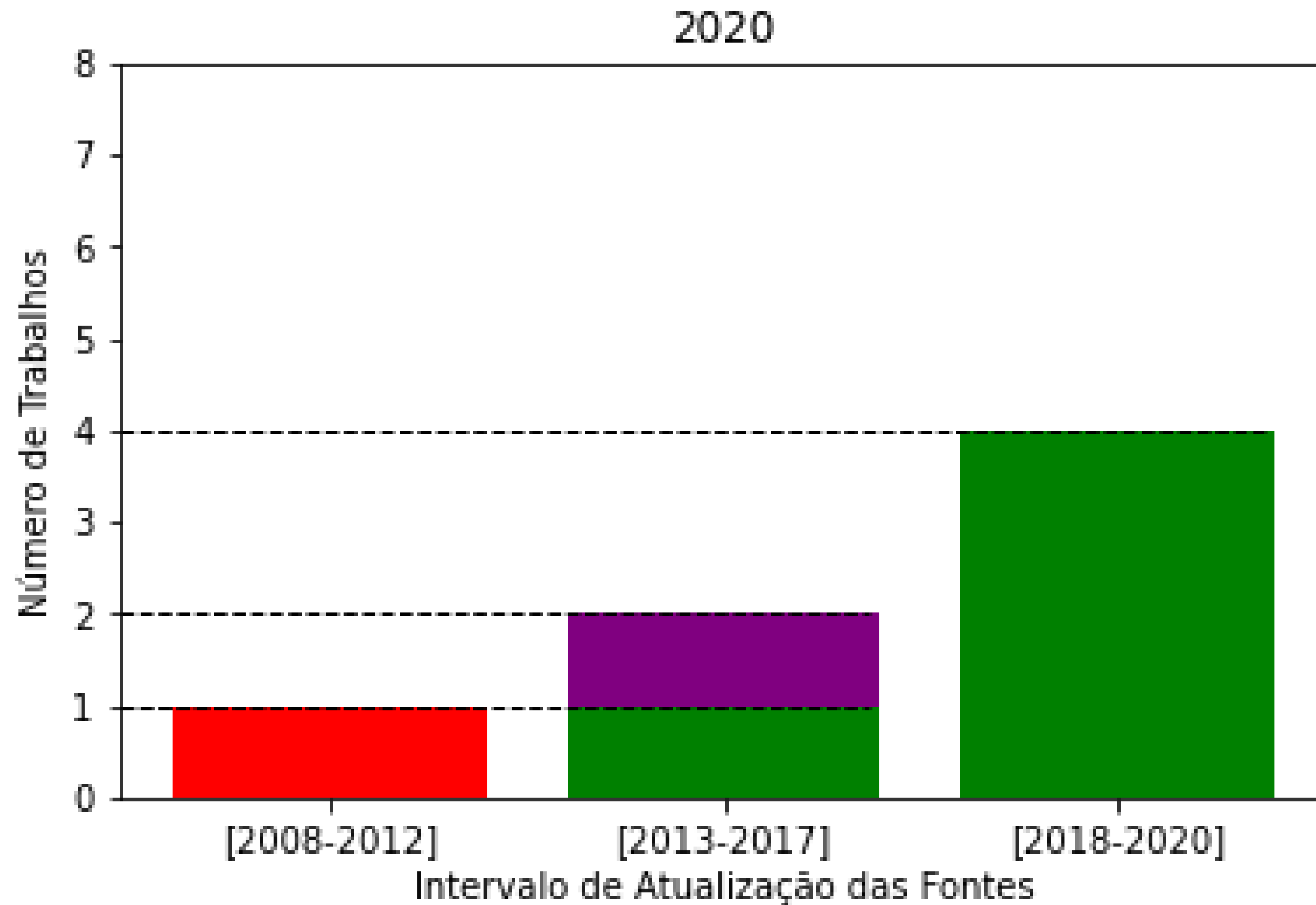
# Resultados



# Resultados

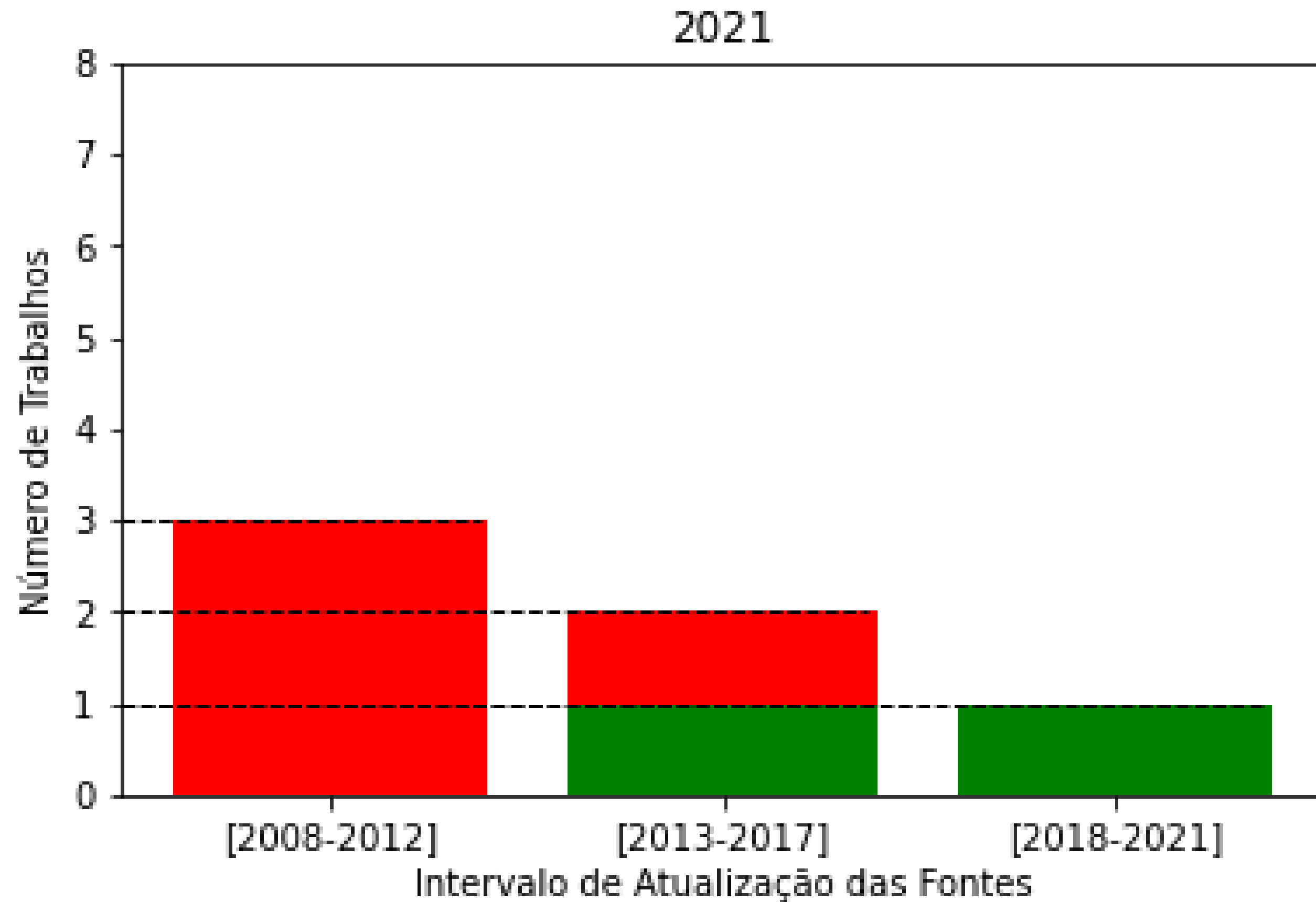


# Resultados





# Resultados



# Resultados

**100% dos trabalhos dos últimos 4 anos  
utilizam fontes antigas**

# Resultados

100% dos trabalhos dos últimos 4 anos  
utilizam fontes antigas

➤ Fontes mais utilizadas:

➤ Drebin

➤ Genome

# Resultados

100% dos trabalhos dos últimos 4 anos  
utilizam fontes antigas

➤ *"A mobile malware detection method using behavior features in network traffic" (2019)*

➤ Drebin

# Resultados

100% dos trabalhos dos últimos 4 anos  
utilizam fontes antigas

➤ *"MLDroid—framework for Android malware detection  
using machine learning techniques" (2020)*

➤ Genome

# Conclusão

- Existem fontes atuais que podem ser acessadas
- Pesquisas utilizam dados desatualizados

# Conclusão

- A atualidade dos dados deve ser melhor analisada
  - Datasets "atuais" compostos por dados antigos
    - CIC-InvesAndMal2019
      - API 25 (2016)

# Trabalhos Futuros

- *Análise APIs*
- *Mapeamento de origem dos Datasets*
- *Acesso às versões antigas*



Obrigada pela atenção!