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

Projeto de Trabalho de Conclusão de Curso

Ciência da Computação - UNIPAMPA

Tainá Oliveira Soares

Orientador Prof. Dr. Diego Luis Kreutz

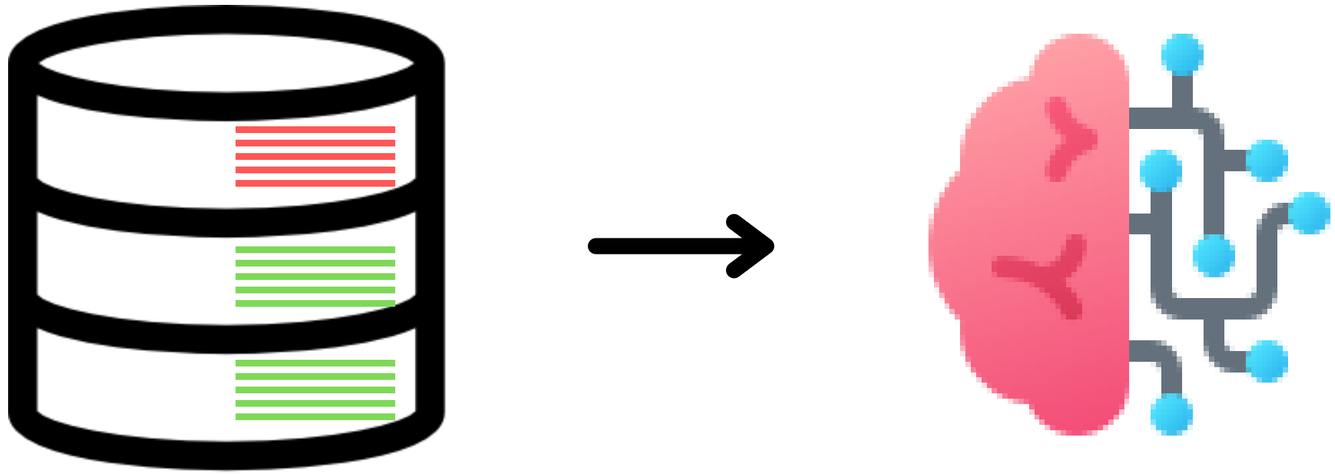
Coorientador Prof. Dr. Eduardo Luzeiro Feitosa

A importância do dataset

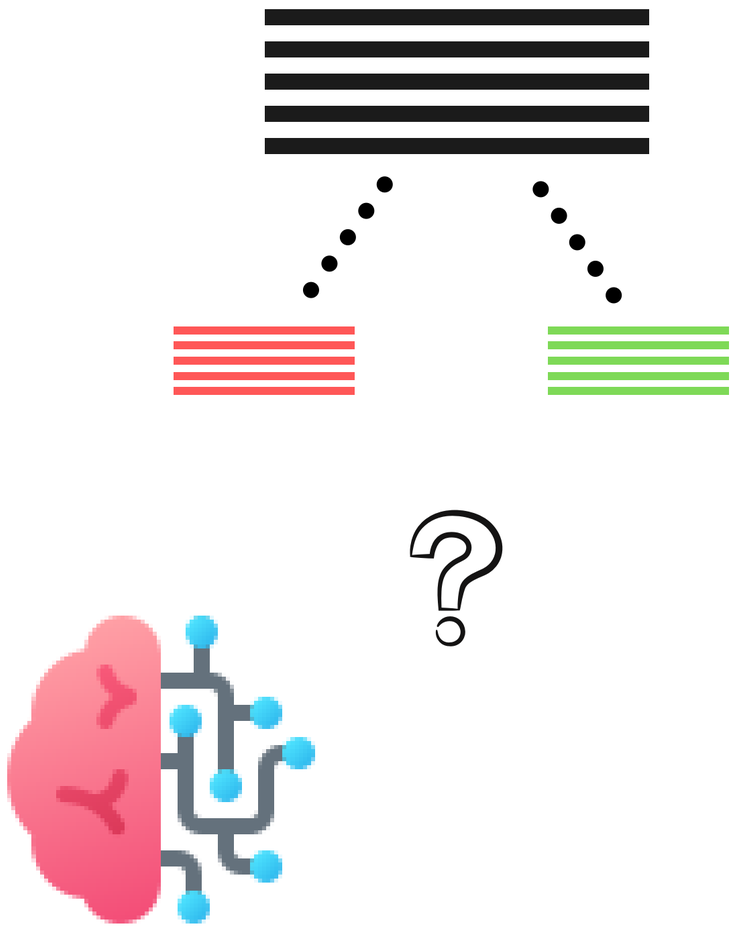
aplicações maliciosas

aplicações benignas

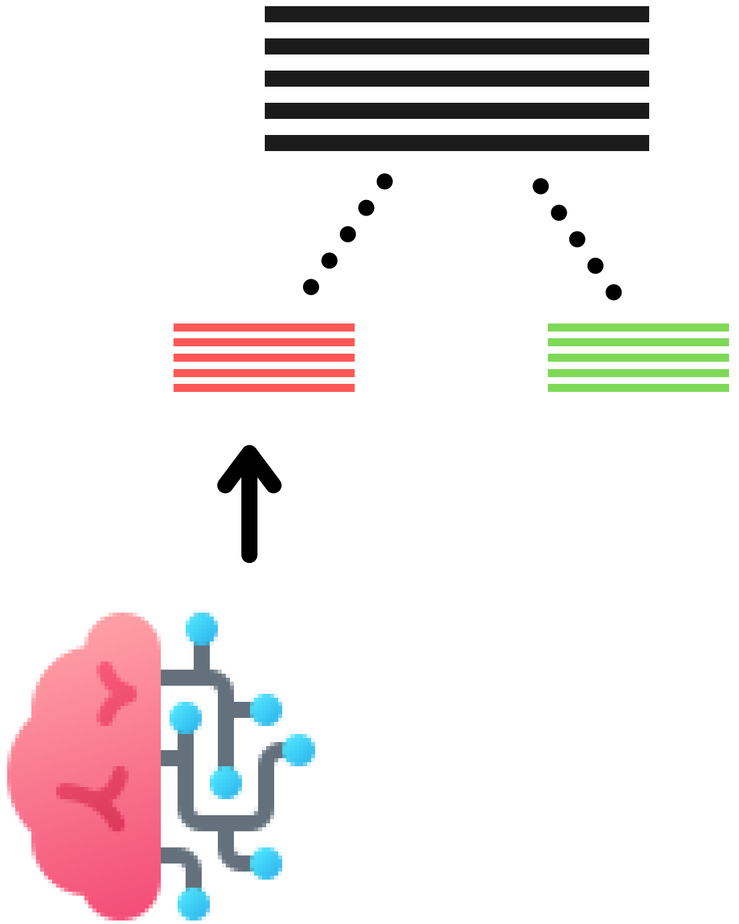
A importância do dataset



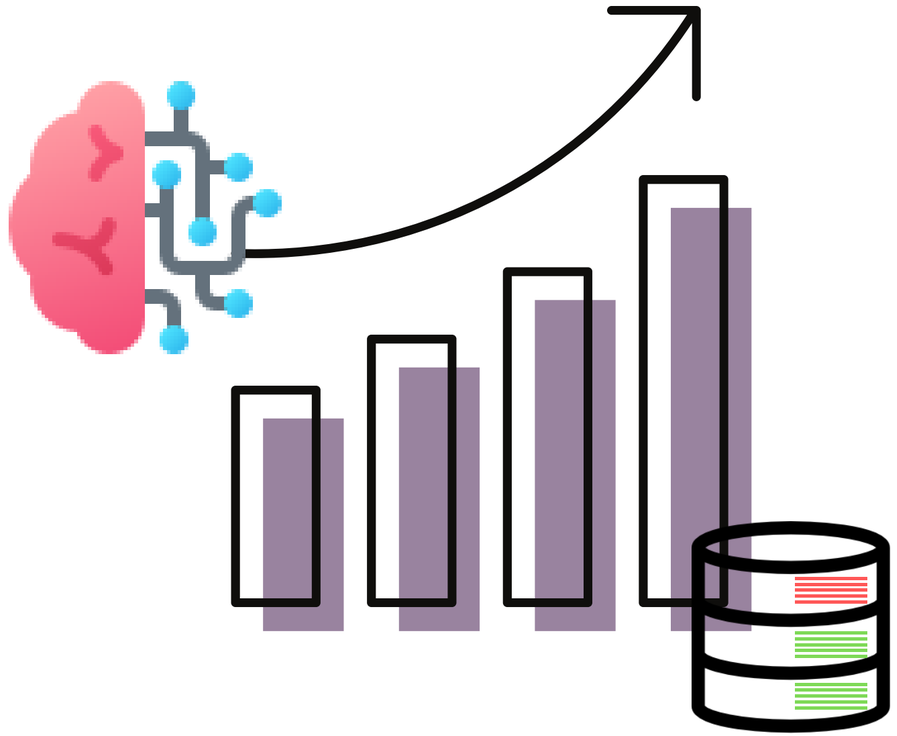
A importância do dataset



A importância do dataset

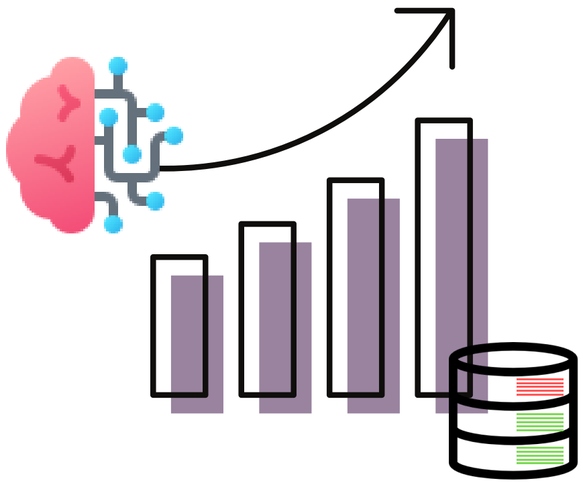


A importância do dataset



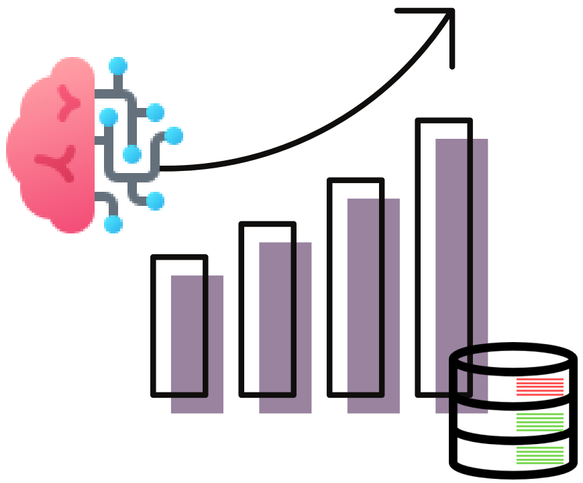
A importância do dataset

- As pesquisas utilizam datasets atuais?



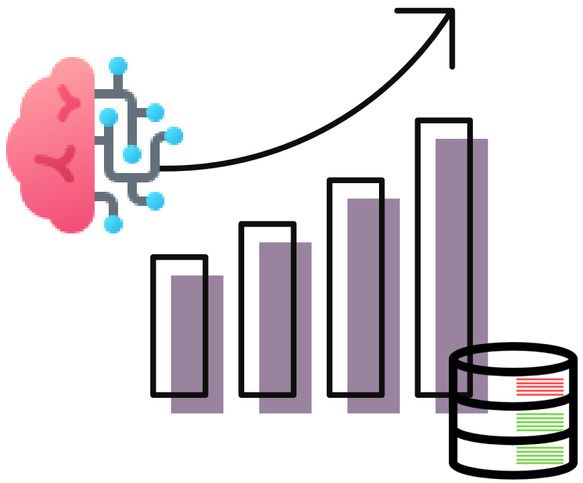
A importância do dataset

- As pesquisas utilizam datasets atuais?
- Os datasets das pesquisas são reprodutíveis?



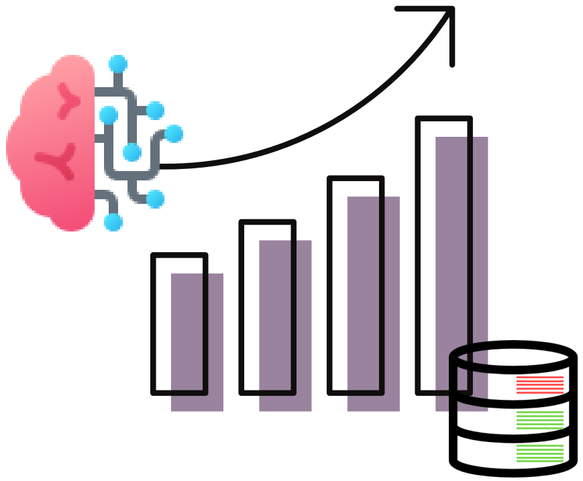
A importância do dataset

- As pesquisas utilizam datasets atuais?
- Os datasets das pesquisas são reprodutíveis?
- Existem fontes de dados disponíveis e atuais?



A importância do dataset

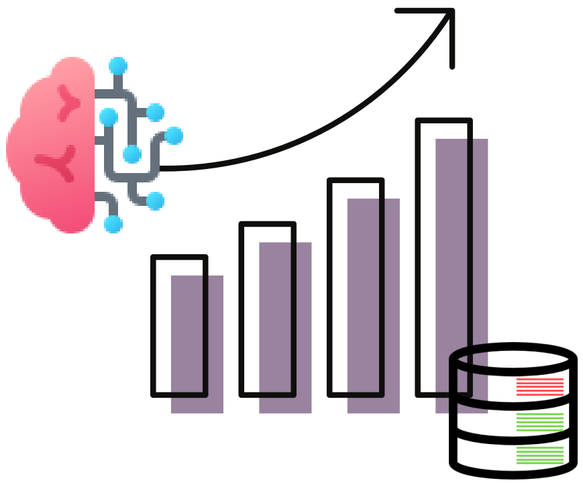
- As pesquisas utilizam datasets atuais?
- Os datasets das pesquisas são reprodutíveis?
- Existem fontes de dados disponíveis e atuais?



- É seguro considerar atual uma fonte de dados de acordo com sua data de criação ou atualização?

A importância do dataset

- As pesquisas utilizam datasets atuais?
- Os datasets das pesquisas são reprodutíveis?
- 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

- Objetivos

Roteiro

- Objetivos
- Etapas

Roteiro

- Objetivos
- Etapas
  - O que já foi realizado?

Roteiro

- Objetivos
- Etapas
  - O que já foi realizado?
- Observações

Roteiro

- Objetivos
- Etapas
  - O que já foi realizado?
- Observações
- Cronograma

Objetivo

Análise qualitativa e quantitativa das fontes de dados utilizadas na detecção de malwares

Objetivo

- Análise qualitativa e quantitativa das fontes de dados utilizadas na detecção de malwares

1. Análise da atualidade e da reprodutibilidade dos datasets utilizados em pesquisas

Objetivo

- Análise qualitativa e quantitativa das fontes de dados utilizadas na detecção de malwares

1. Análise da atualidade e da reprodutibilidade dos datasets utilizados em pesquisas

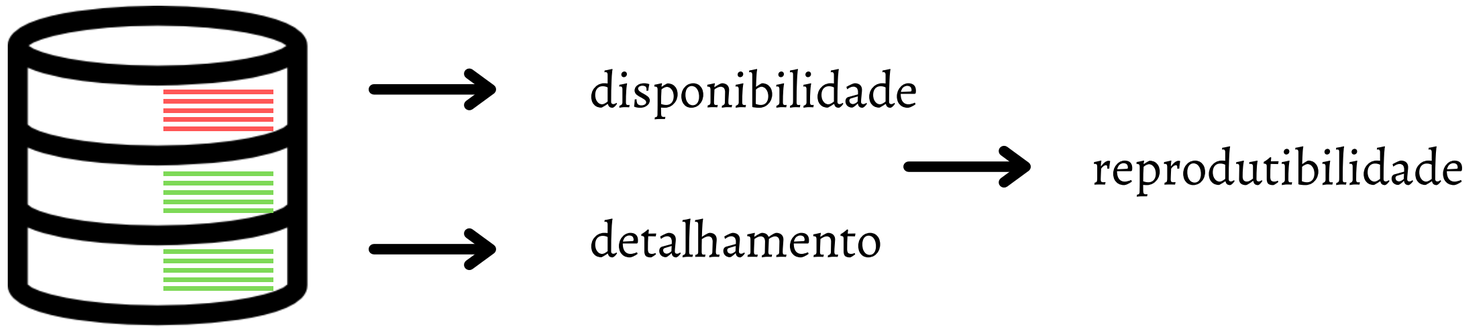
→ atualização

→ disponibilidade

Objetivo

- Análise qualitativa e quantitativa das fontes de dados utilizadas na detecção de malwares

1. Análise da atualidade e da reprodutibilidade dos datasets utilizados em pesquisas



Objetivo

- Análise qualitativa e quantitativa das fontes de dados utilizadas na detecção de malwares

1. Análise da atualidade e da reprodutibilidade dos datasets utilizados em pesquisas

→ disponibilidade

→ detalhamento

→ reprodutibilidade

- validação
- comparação

Objetivo

- Análise qualitativa e quantitativa das fontes de dados utilizadas na detecção de malwares

1.
2. Análise da atualidade e da disponibilidade de fontes de dados utilizadas na detecção de malwares

Objetivo

- Análise qualitativa e quantitativa das fontes de dados utilizadas na detecção de malwares

1.
2.
3. Proposta de dataset

Objetivo

- Análise qualitativa e quantitativa das fontes de dados utilizadas na detecção de malwares

1.
2.
3.

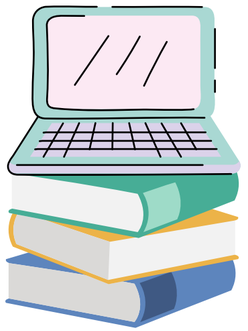
Etapas

1	2	3	4	5	6

Etapas

1 **2** **3** **4** **5** **6**

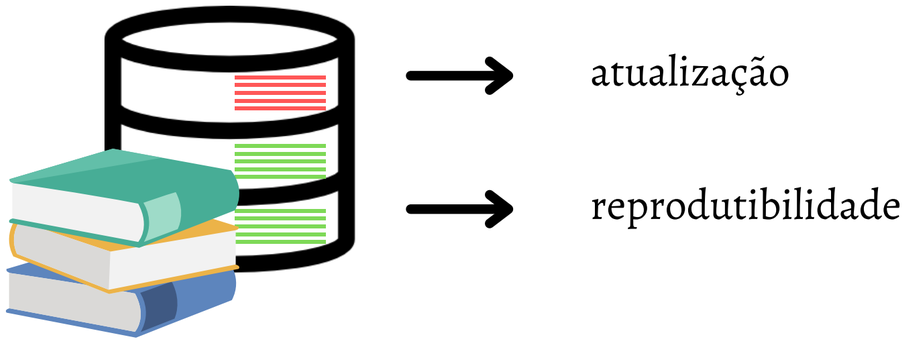
Levantamento de trabalhos que propõem métodos de detecção de malwares baseados em modelos de aprendizado de máquina



Etapas

1 2 3 4 5 6

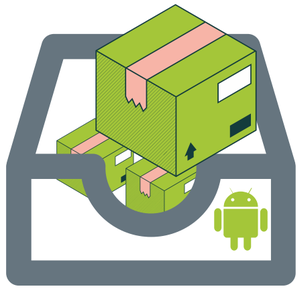
Levantamento de dados quanto a atualização e reprodutibilidade dos datasets utilizados nos estudos levantados na Etapa 1

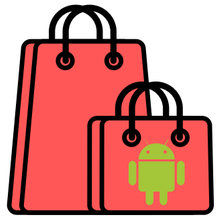


Etapas

1 **2** **3** **4** **5** **6**

Catalogação de fontes de dados utilizadas para detecção de malwares





Etapas

1 **2** **3** **4** **5** **6**

Levantamento de dados quanto a disponibilidade e a atualização das fontes catalogadas na Etapa 3

→ atualização

→ disponibilidade

Etapas

1 **2** **3** **4** **5** **6**

Investigação de aspectos relevantes nos datasets

Etapas

1 **2** **3** **4** **5** **6**

Elaboração de dataset baseado nas informações coletadas na Etapa 5

Etapas – o que já foi realizado?

- Etapa 1 
- Etapa 2 
- Etapa 3 
- Etapa 4 

Etapa 1

Levantamento de trabalhos que propõem métodos de detecção de malwares baseados em modelos de aprendizado de máquina

Etapa 1

35 trabalhos

Etapa 1

35 trabalhos

G1 -> citados por algum survey ou revisão sistemática de literatura

Etapa 1

35 trabalhos

G1 -> citados por algum survey ou revisão sistemática de literatura

G2 -> 40 (ou mais) citações segundo o Google Scholar

Etapa 1

35 trabalhos

G1 -> citados por algum survey ou revisão sistemática de literatura

G2 -> 40 (ou mais) citações segundo o Google Scholar

G3 -> publicados nos principais periódicos ou conferências da área de segurança

Etapa 1

35 trabalhos

G1 -> citados por algum survey ou revisão sistemática de literatura

G2 -> 40 (ou mais) citações segundo o Google Scholar

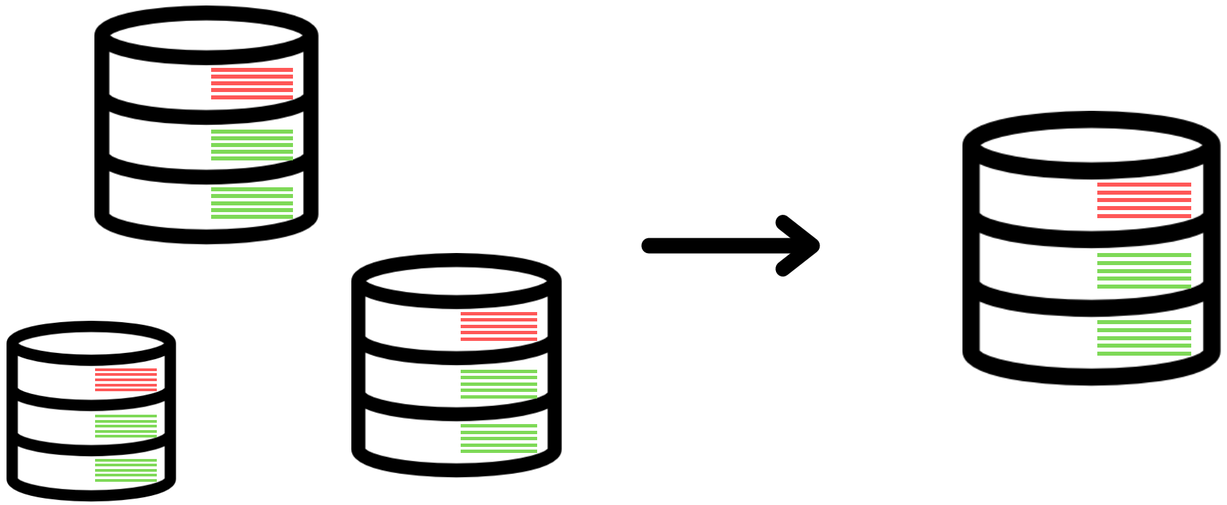
G3 -> publicados nos principais periódicos ou conferências da área de segurança

G4 -> publicados em conferências específicas da área de inteligência artificial

Etapa 2

Levantamento de dados quanto a atualização e reprodutibilidade dos datasets utilizados nos estudos

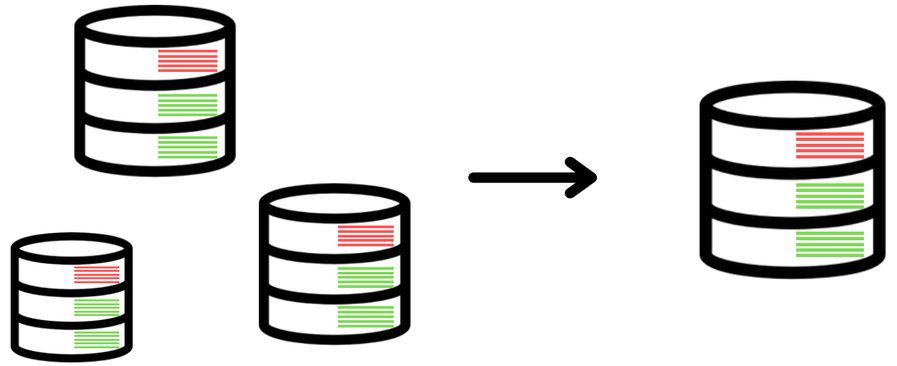
Etapa 2



Etapa 2

Reprodutibilidade

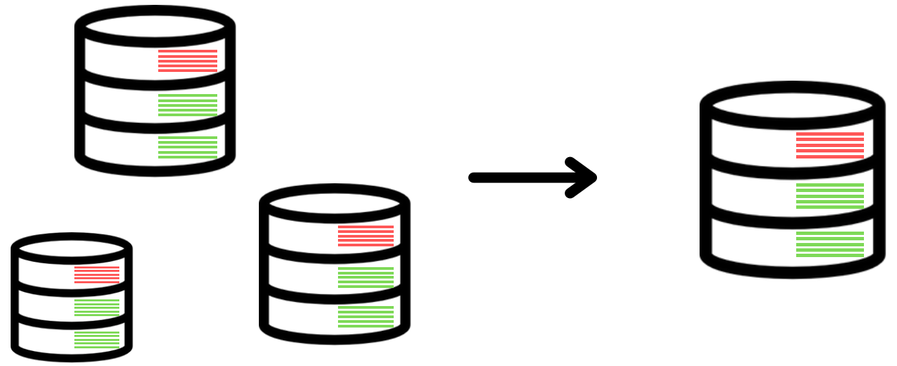
1. Referência a origem dos dados



Etapa 2

Reprodutibilidade

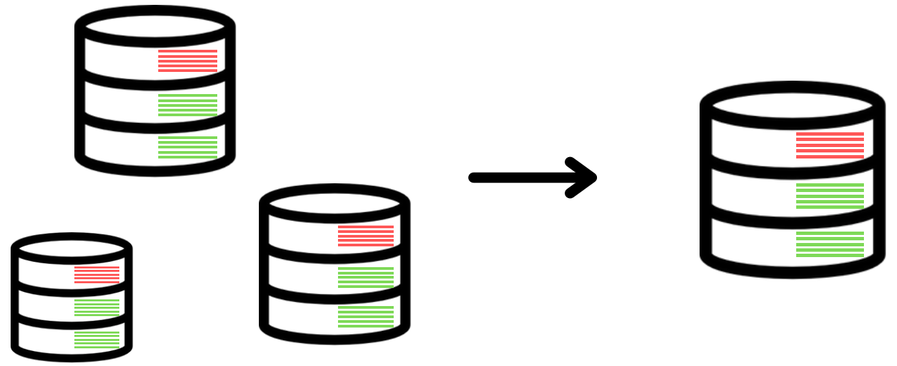
1.
2. Detalhamento da quantidade de amostras



Etapa 2

Reprodutibilidade

1.
2.
3. Descrição específica das amostras



Etapa 2

Reprodutibilidade

- 5% dos estudos apresentam fontes disponíveis + quantidade

Etapa 2

Reprodutibilidade

- 5% dos estudos apresentam fontes disponíveis + quantidade
- 34% dos estudos não apresentam quantidade retirada dos mercados

Etapa 2

Reprodutibilidade

- 5% dos estudos apresentam fontes disponíveis + quantidade
- 34% dos estudos não apresentam quantidade retirada dos mercados
- 0% dos estudos apresentam detalhamento específico

Etapa 2

Atualização

1. Datas informadas nos sites ou trabalhos

Etapa 2

Atualização

1. Datas informadas nos sites ou trabalhos

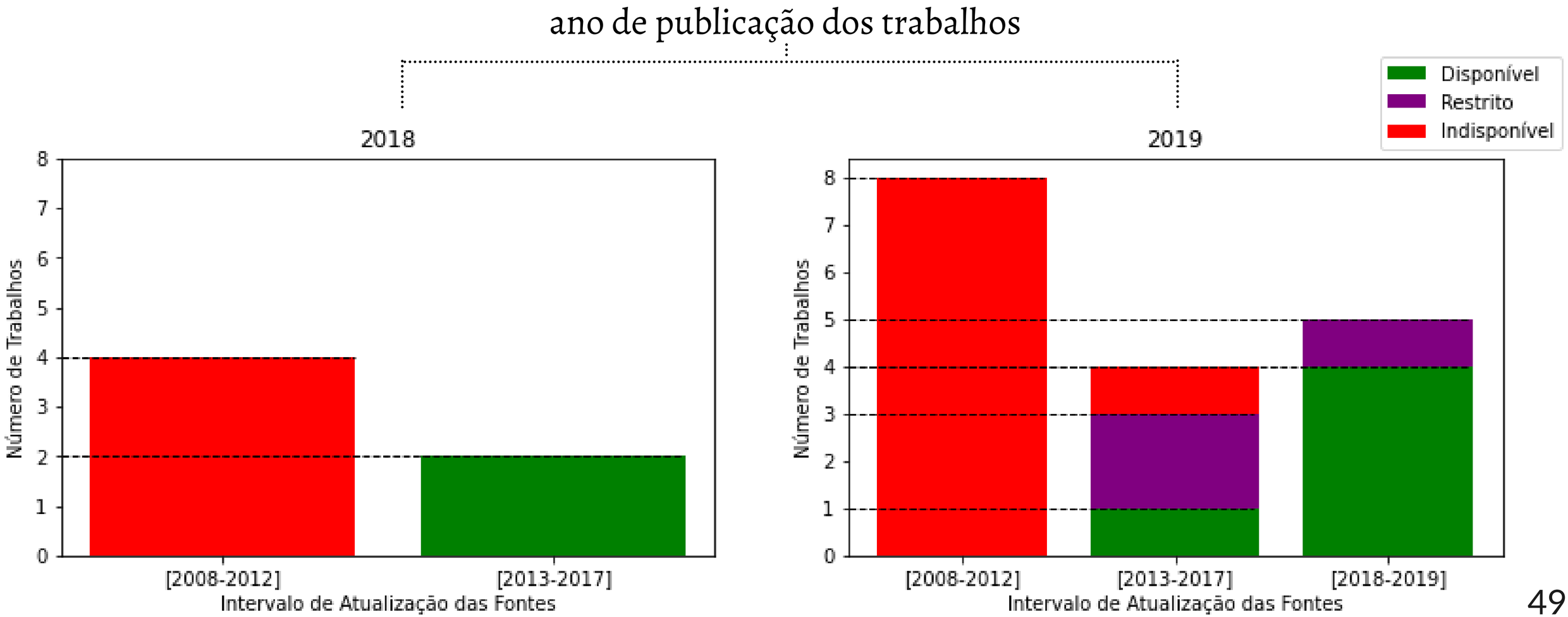
Etapa 2

Atualização

1. Datas informadas nos sites ou trabalhos
2. Trabalhos dos últimos 4 anos

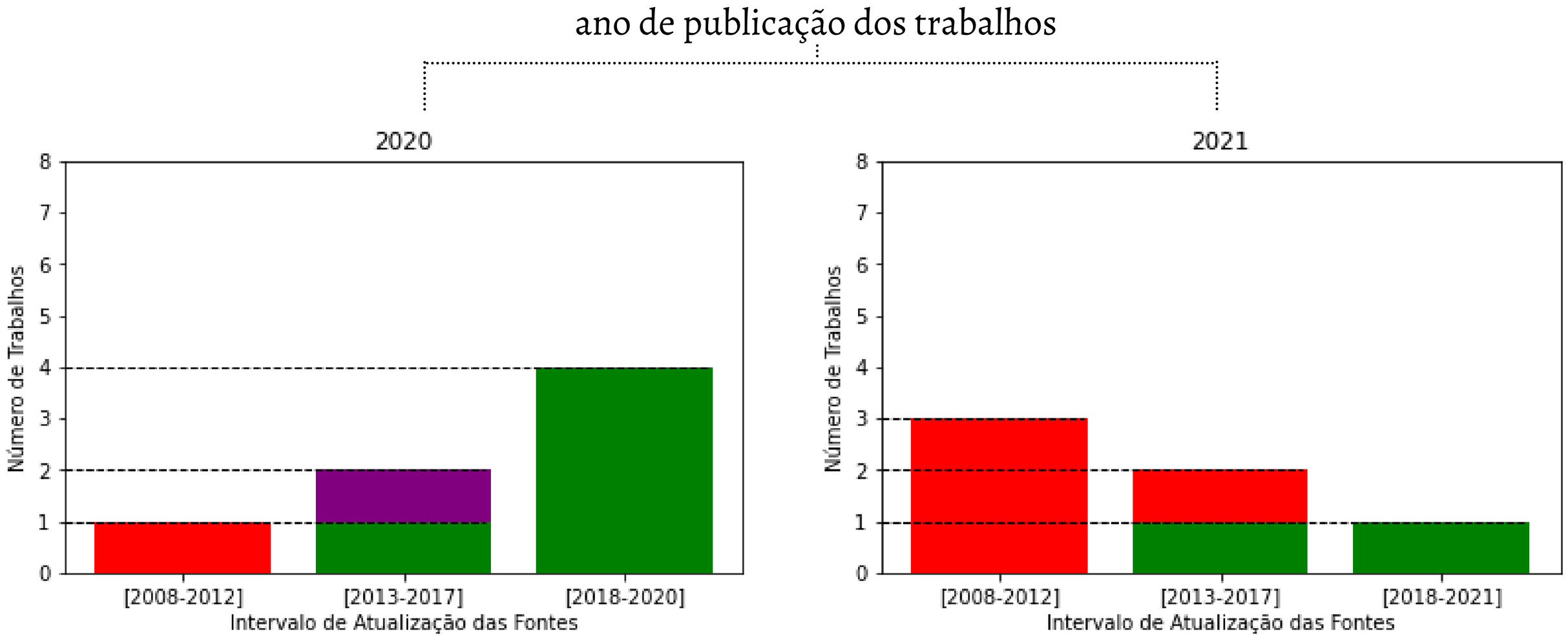
Etapa 2

Atualização



Etapa 2

Atualização



Etapa 3

Catalogação de fontes de dados utilizadas para detecção de malwares

Etapa 3

84 fontes de dados

Etapa 3

84 fontes de dados

39 mercados de aplicativos

30 datasets

15 repositórios de APKs

Etapa 3

84 fontes de dados

-> Revisão sistemática

-> 35 trabalhos analisados

-> 100 primeiros resultados em plataformas de busca

Etapa 4

Levantamento de dados quanto a disponibilidade e a atualização das fontes catalogadas

Etapa 4

Mercados de Aplicativos

54% disponíves e em atualização constante

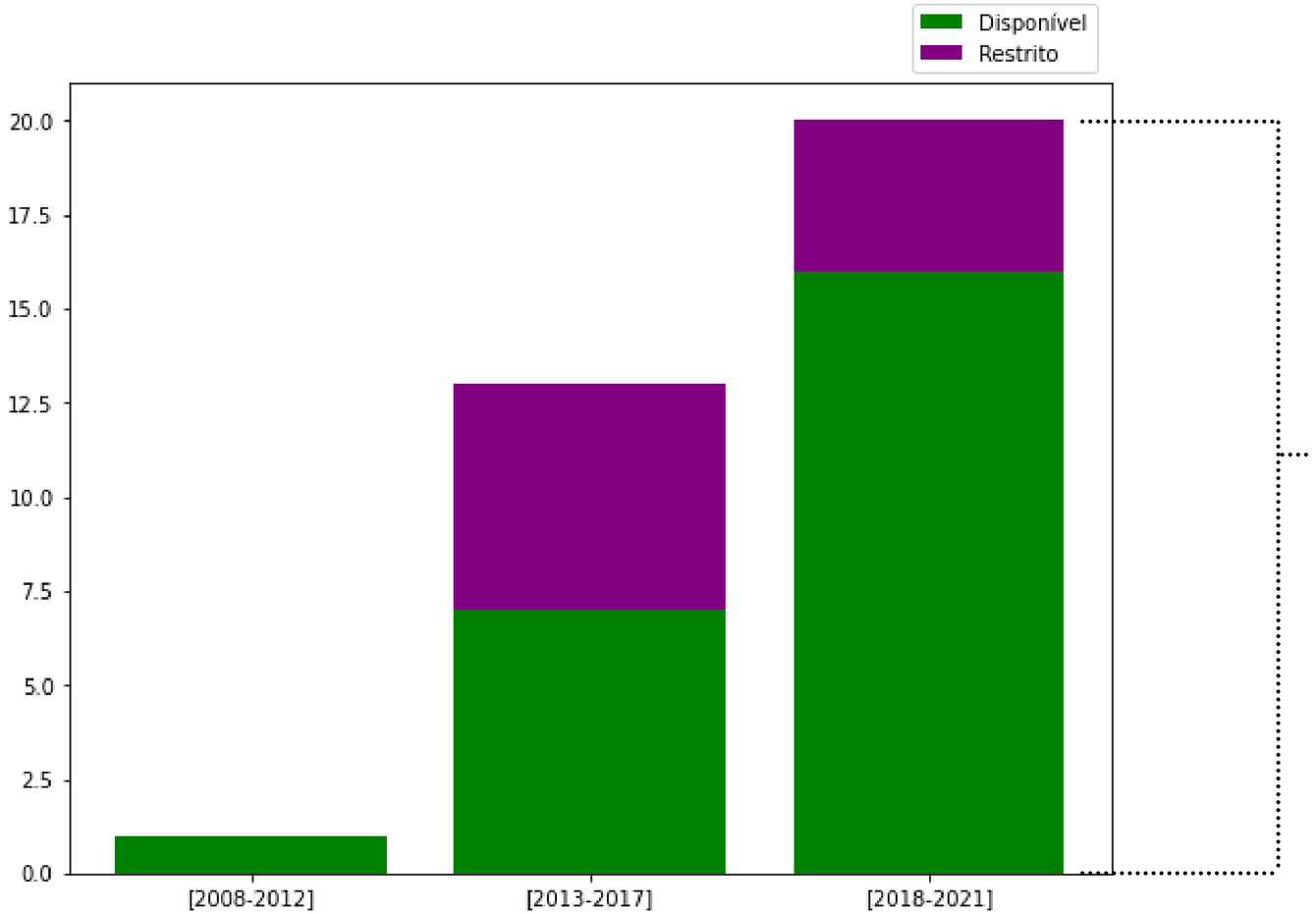
Etapa 4

Datasets e Repositórios

80% disponíveis ou restritos

Etapa 4

Datasets e Repositórios



44% dos datasets e repositórios

Observações

100% dos datasets são irreprodutíveis

Observações

100% dos datasets são irreprodutíveis

- Crowdroid: behavior-based malware detection system for Android
  - Desenvolvimento das amostras

- Machine Learning in Wavelet Domain for Electromagnetic Emission Based Malware Analysis
  - Quantidade específica de cada fonte

Observações

100% dos trabalhos dos últimos 4 anos mencionam fontes antigas

Observações

100% dos trabalhos dos últimos 4 anos mencionam fontes antigas

- The Drebin Dataset e Malware Genome Dataset

Observações

Maioria das fontes de dados são disponíveis

Observações

Maioria das fontes de dados são disponíveis

- 64% das lojas
- 55% dos datasets e repositórios

Observações

A atualidade dos dados deve ser melhor analisada

Observações

A atualidade dos dados deve ser melhor analisada

- Datasets "atuais" compostos por dados antigos
  - CIC-InvesAndMal2019
    - API 25 (2016)

Cronograma

		Nov	Dez	Jan	Fev	Mar
3	Inclusão de novas fontes no catálogo	X				
4	Árvore genealógica de datasets	X	X			
4	Análise da atualidade das fontes de dados através das versões das APIs presentes	X	X	X		
5	Categorização dos datasets			X	X	
5	Análise do impacto dos datasets nos modelos de aprendizado de máquina			X	X	
6	Construção de dataset				X	X

Obrigada!