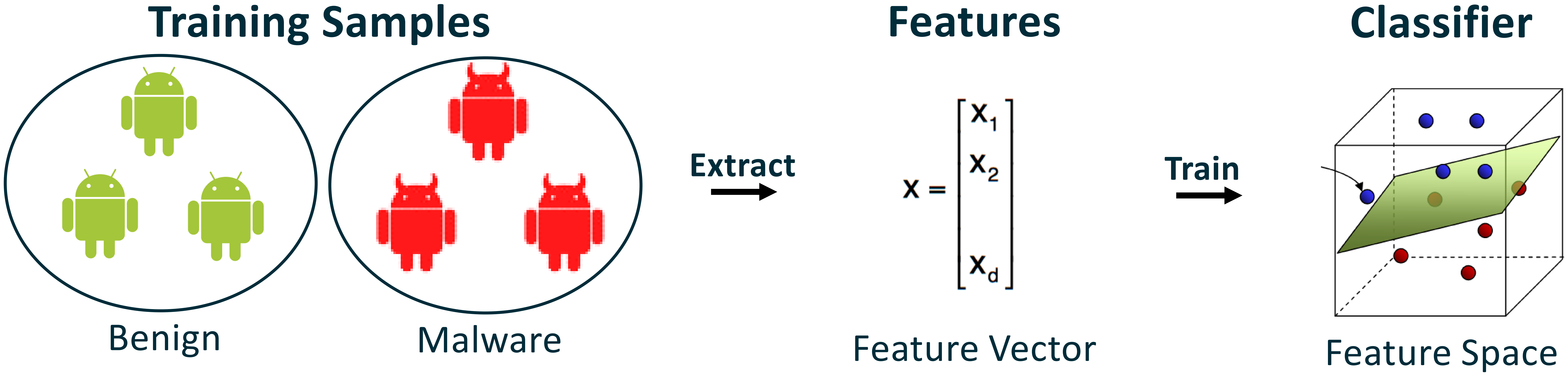
VISUALIZING ANDROID APP SIMILARITY

Michael Cao & Gabriella Xiong

OCTOBER 1ST 2020

BACKGROUND & MOTIVATION

- Cybercriminals pose a serious threat to mobile software systems
- Most techniques are machine learning based



Training Samples

Benign

Malware

Extract

Features

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_d \end{bmatrix}$$

Feature Vector

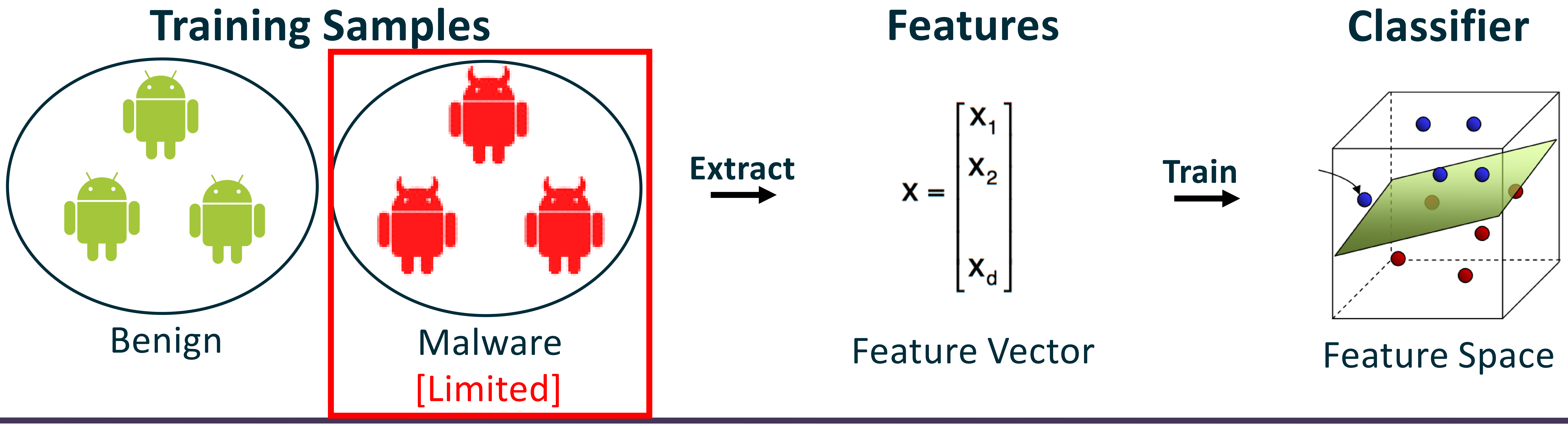
Train

Classifier

Feature Space

BACKGROUND & MOTIVATION

- Cybercriminals pose a serious threat to mobile software systems
- Most techniques are machine learning based



Training Samples

Benign

Malware

[Limited]

Extract

Features

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_d \end{bmatrix}$$

Feature Vector

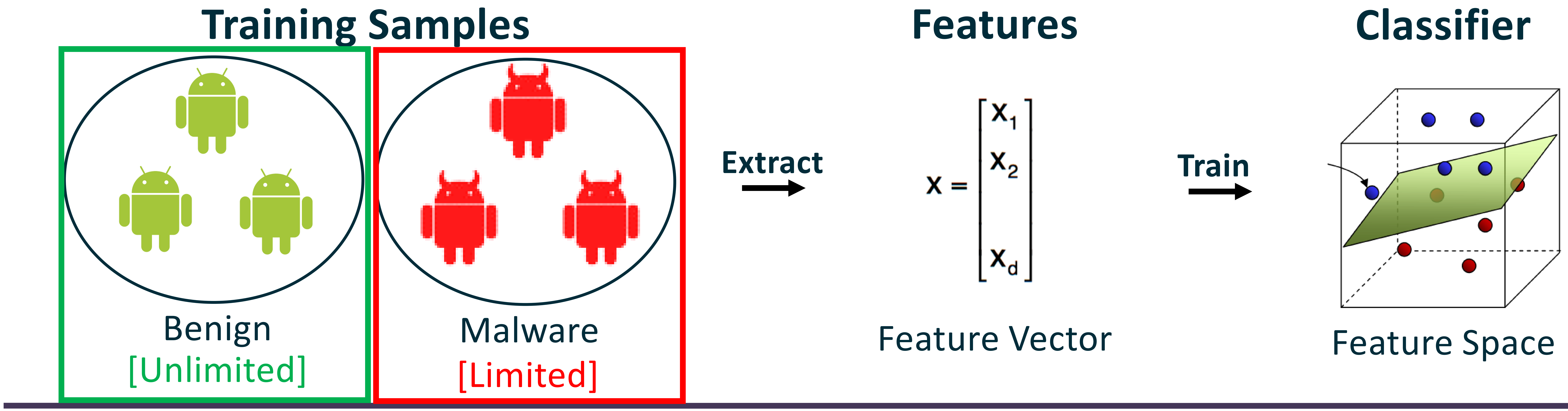
Train

Classifier

Feature Space

BACKGROUND & MOTIVATION

- Cybercriminals pose a serious threat to mobile software systems
- Most techniques are machine learning based



Training Samples

Benign
[Unlimited]

Malware
[Limited]

Extract

Features

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

Feature Vector

Train

Classifier

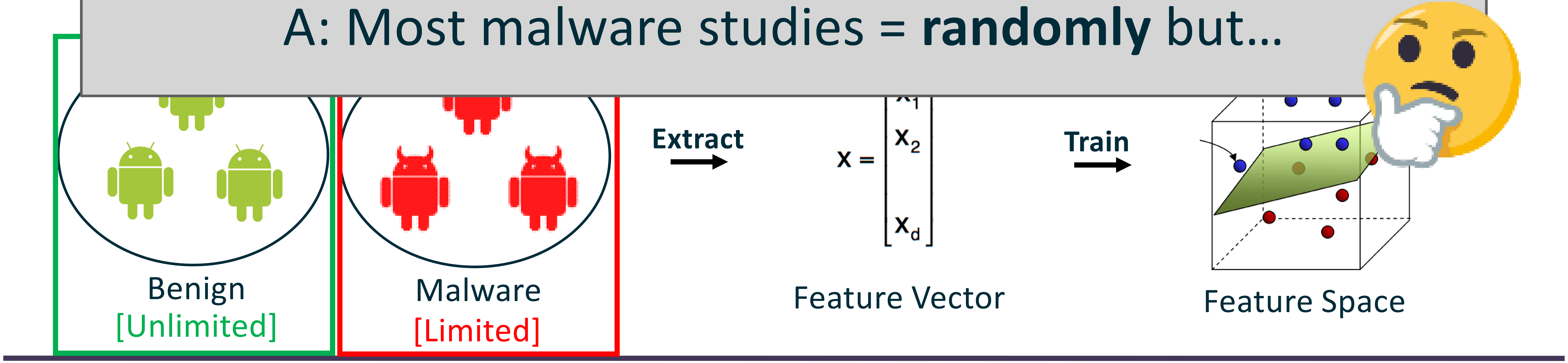
Feature Space

BACKGROUND & MOTIVATION

- Cybercriminals pose a serious threat to mobile software systems

Q: How do we select benign samples?

A: Most malware studies = **randomly** but…



Benign
[Unlimited]

Malware
[Limited]

Extract

$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$

Feature Vector

Train

Feature Space

PROBLEM

- Random benign can produce separable, vulnerable patterns [1]
  - Malware adopt benign behaviors to evade detection
- Select benign samples similar to malware → mitigate vulnerability

[1] Cao, Michael, Sahar Badihi, Khaled Ahmed, Peiyu Xiong, and Julia Rubin. "On Benign Features in Malware Detection."

PROBLEM

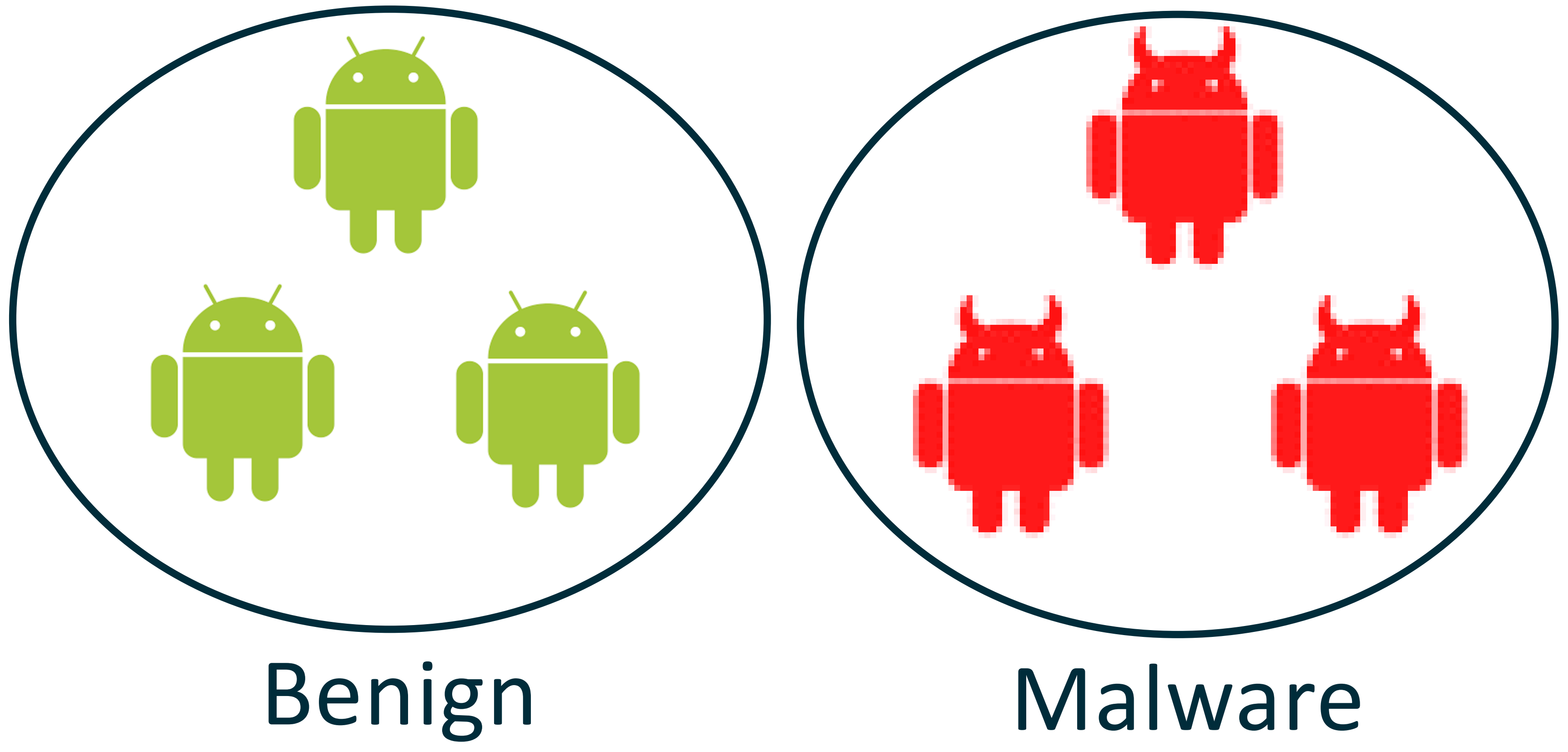
- Random benign can produce separable, vulnerable patterns [1]
  - Malware adopt benign behaviors to evade detection
- Select benign samples similar to malware → mitigate vulnerability

GOAL

Help researchers identify similar benign samples w.r.t a malware set

[1] Cao, Michael, Sahar Badihi, Khaled Ahmed, Peiyu Xiong, and Julia Rubin. "On Benign Features in Malware Detection."

OUR DATASETS



Data Source:	Google Play	VirusTotal
Data Amount:	50000+	10000
Data Time Range:	2016 ~ 2019	2016 ~ 2019

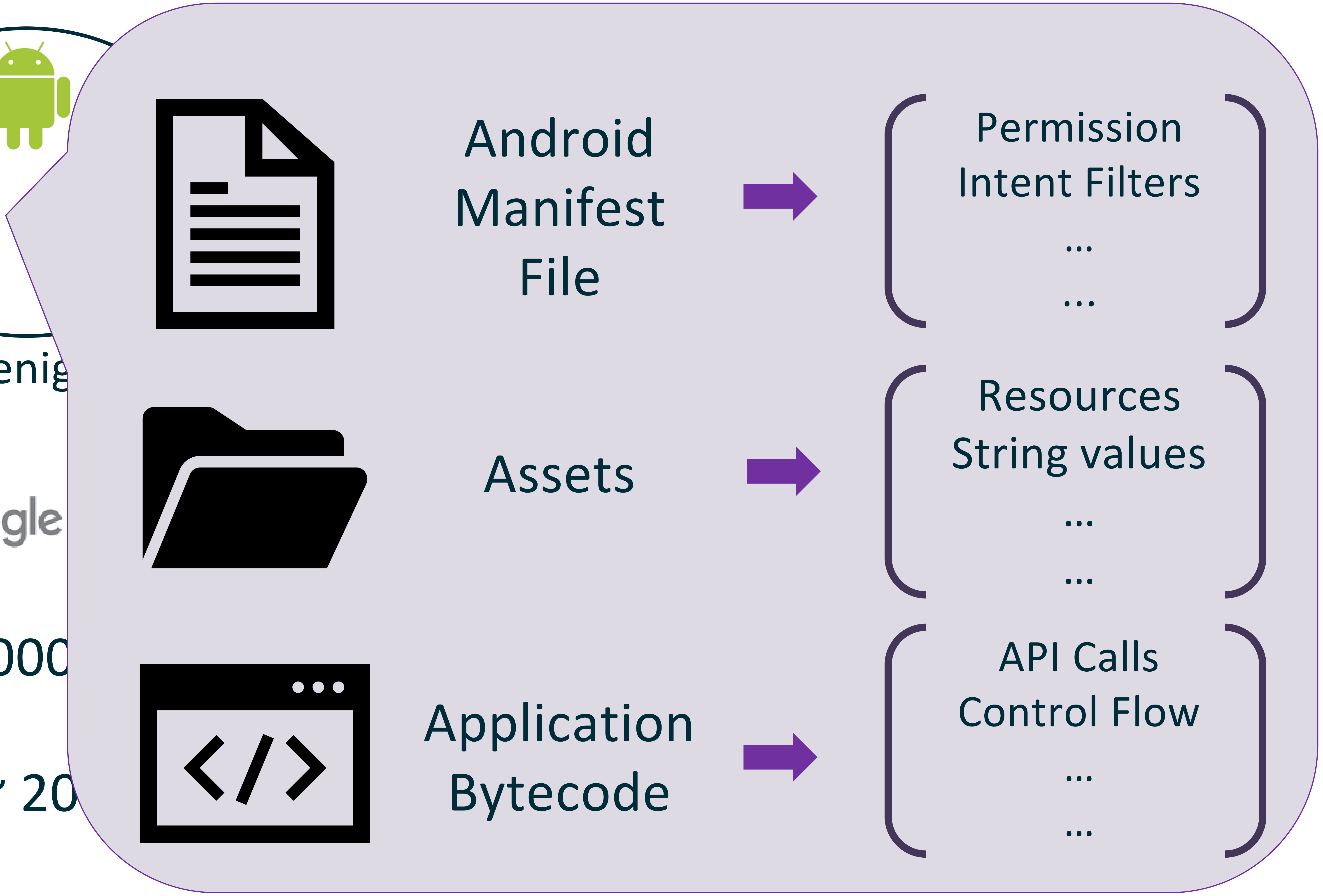
OUR DATASETS

Benig

Data Source: Google

Data Amount: 50000

Data Time Range: 2016 ~ 20



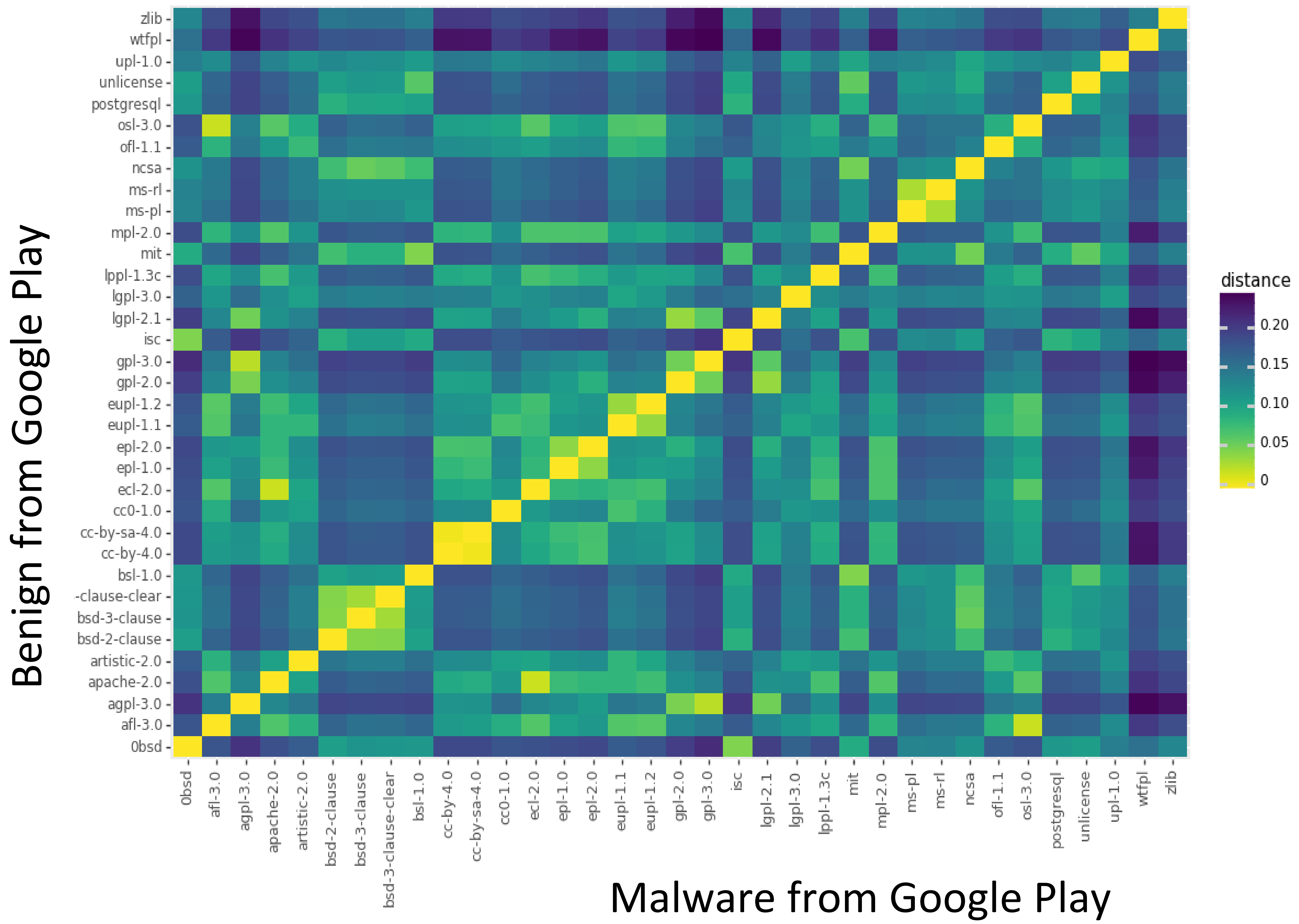
Android Manifest File → Permission Intent Filters … …

Assets → Resources String values … …

Application Bytecode → API Calls Control Flow … …

OUR PROPOSED VISUALIZATION

- Heatmap to reveal the similarity between samples



Similarity can be calculated as:

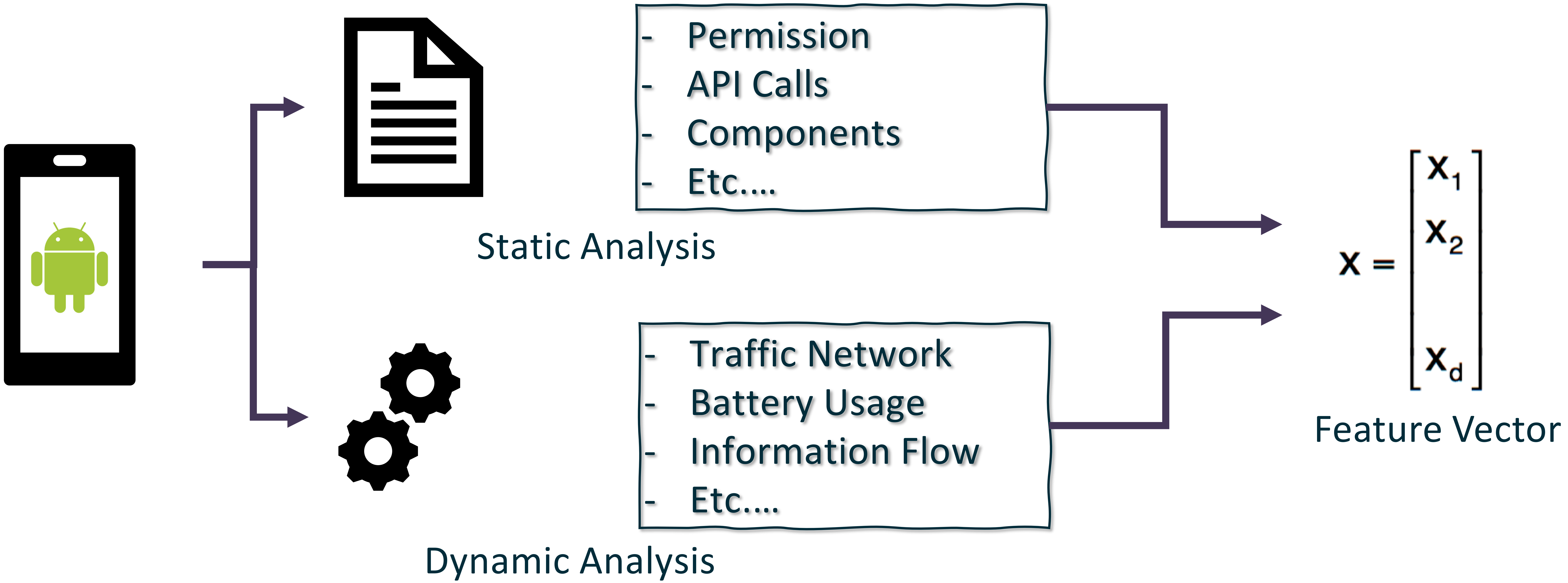
- Distance between samples feature vectors
- Cosine similarity between sample feature vectors
- Etc.

THANK YOU!

BACKUPS

DATA ATTRIBUTES

- Feature vector can be extracted through static / dynamic analysis



- Permission
- API Calls
- Components
- Etc.…

Static Analysis

- Traffic Network
- Battery Usage
- Information Flow
- Etc.…

Dynamic Analysis

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \\ \mathbf{x}_d \end{bmatrix}$$

Feature Vector

OUR PROPOSED VISUALIZATION

- Possible interaction
  - Allow user to select modify the set of features for similarity calculation
  - Allow user to select a subset of samples:
    - Automatically identify the features contribute to similarity/difference the most → the set of interesting features
    - Show the distribution of samples over the interesting features