Mind Your Weight(s): A Large-scale Study on Insufficient Machine Learning Model Protection in Mobile Apps

Zhichuang Sun
Northeastern University

Ruimin Sun
Northeastern University

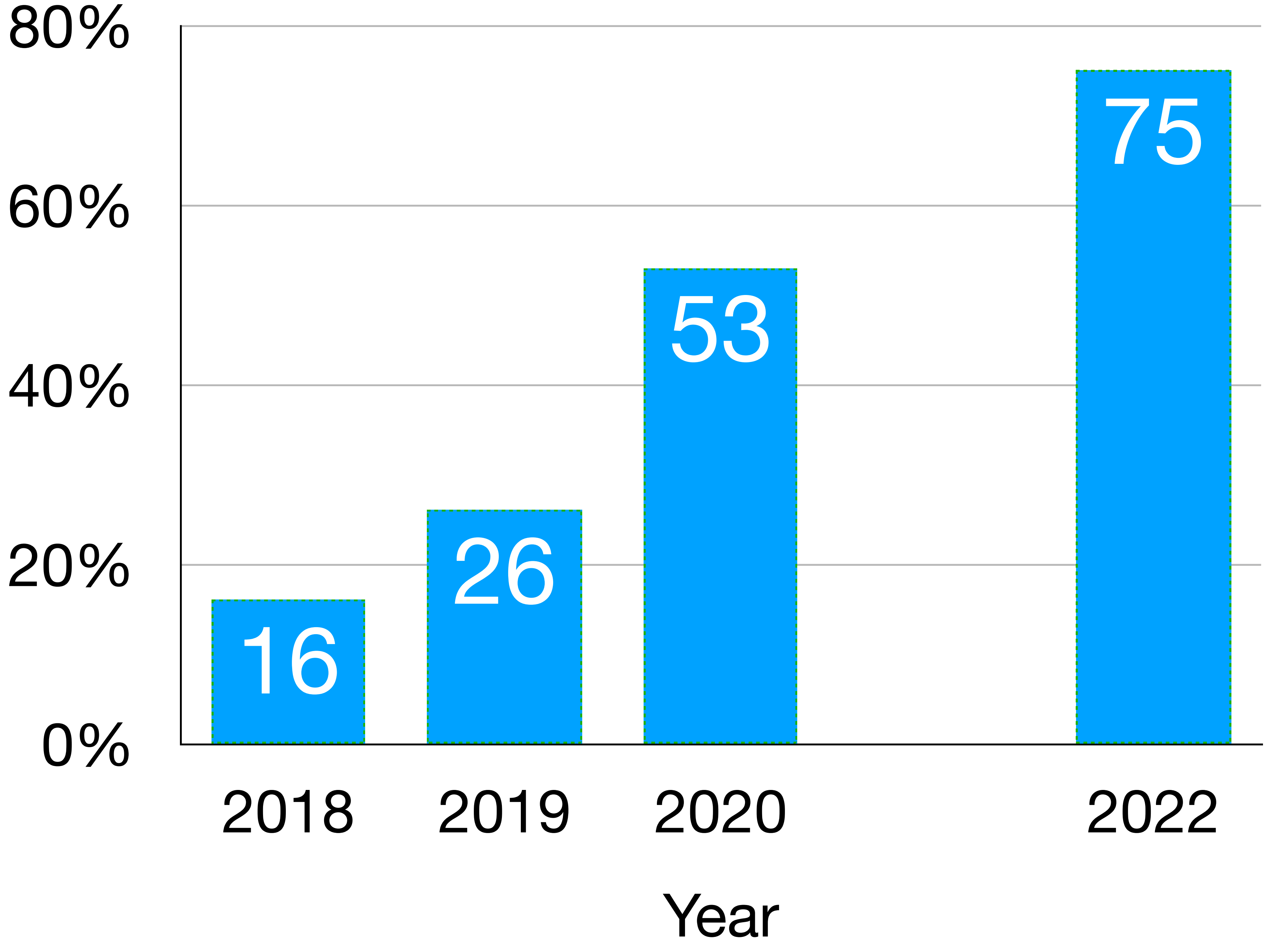
Long Lu
Northeastern University

Alan Mislove
Northeastern University

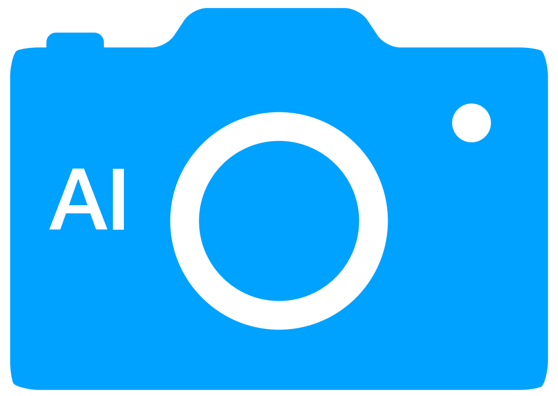
AI/ML are becoming very important for smartphones

- More phones comes with dedicated AI chips

Smartphone released with AI chips (%)



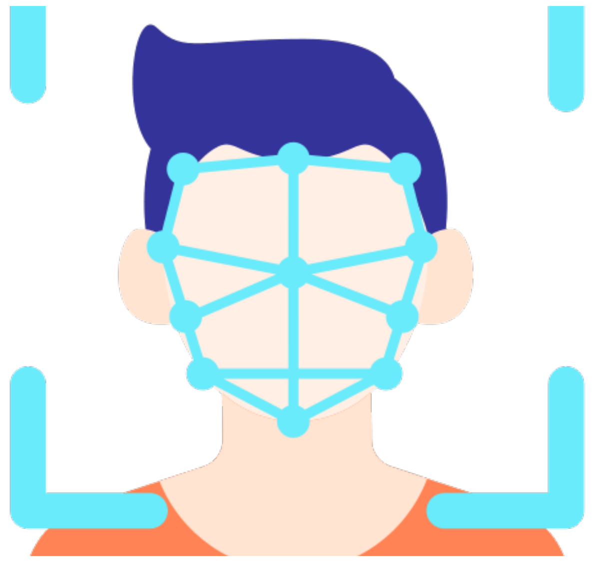
- More AI tasks on smartphones



AI camera

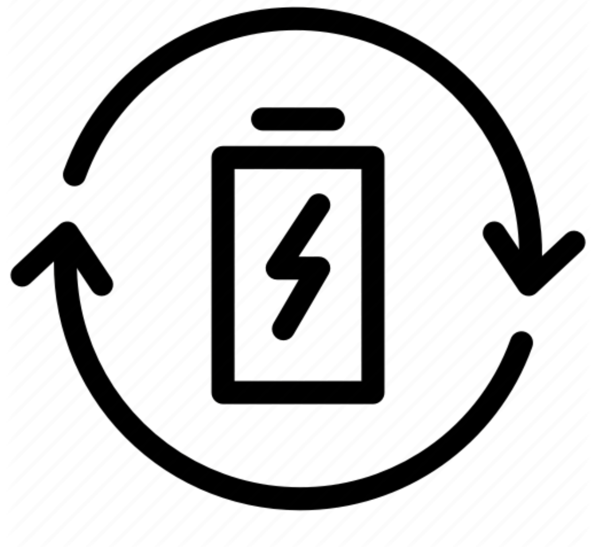
Siri

Augmented Reality



Facial Recognition

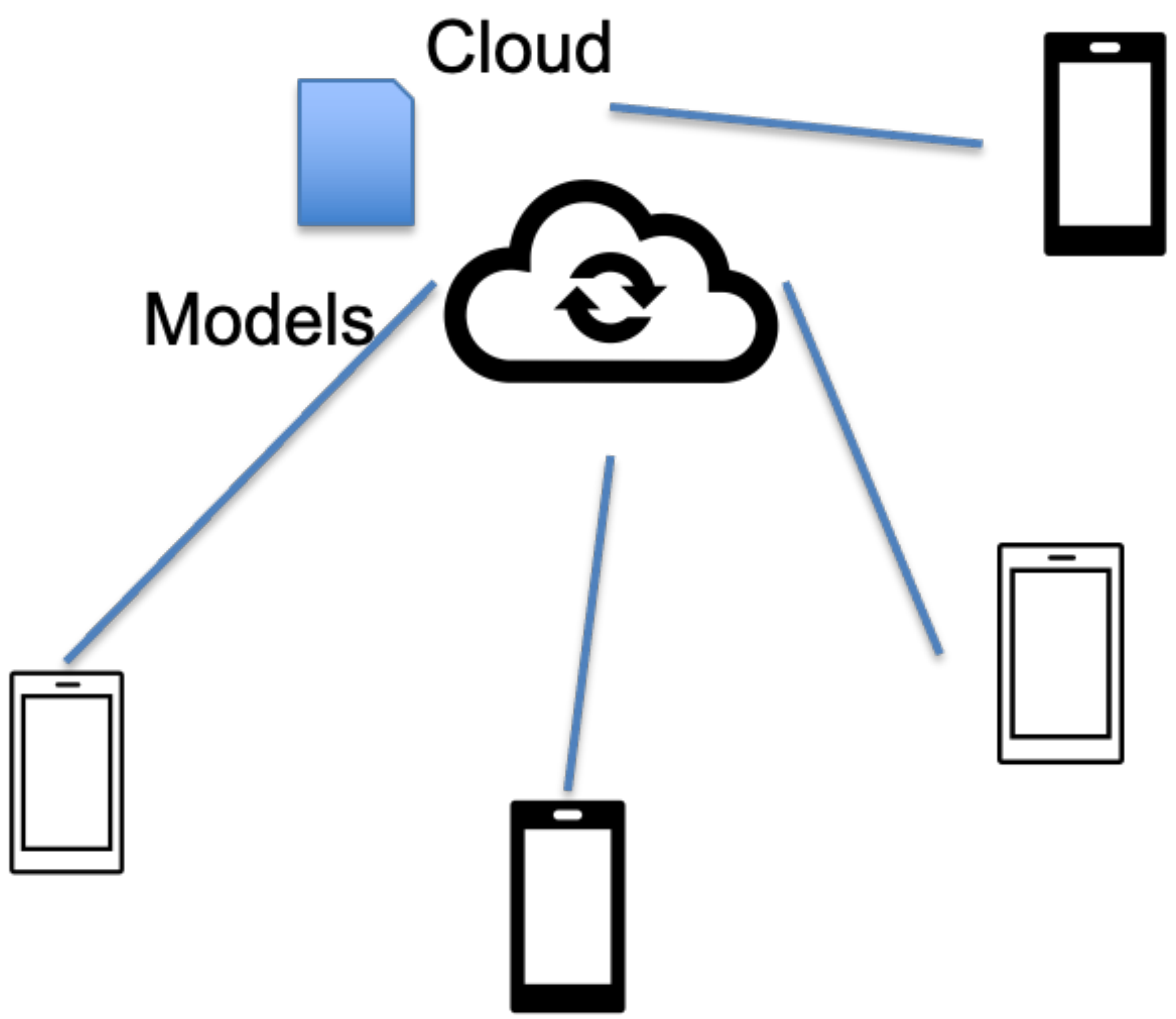
OCR



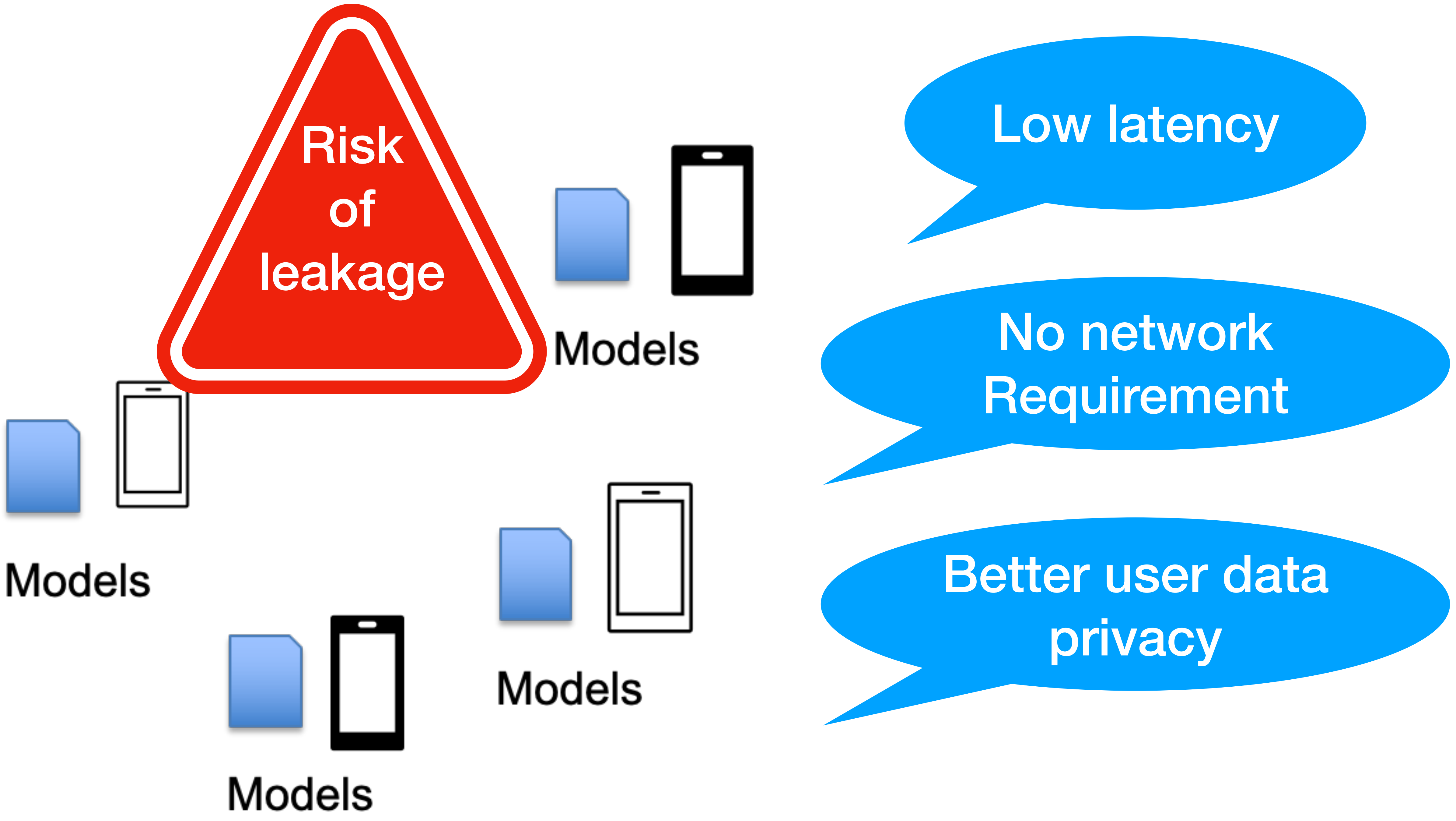
battery management

Cloud-based ML vs On-device ML

- Machine Learning(ML) models are the core IP of model vendors



Cloud-based ML



On-device ML

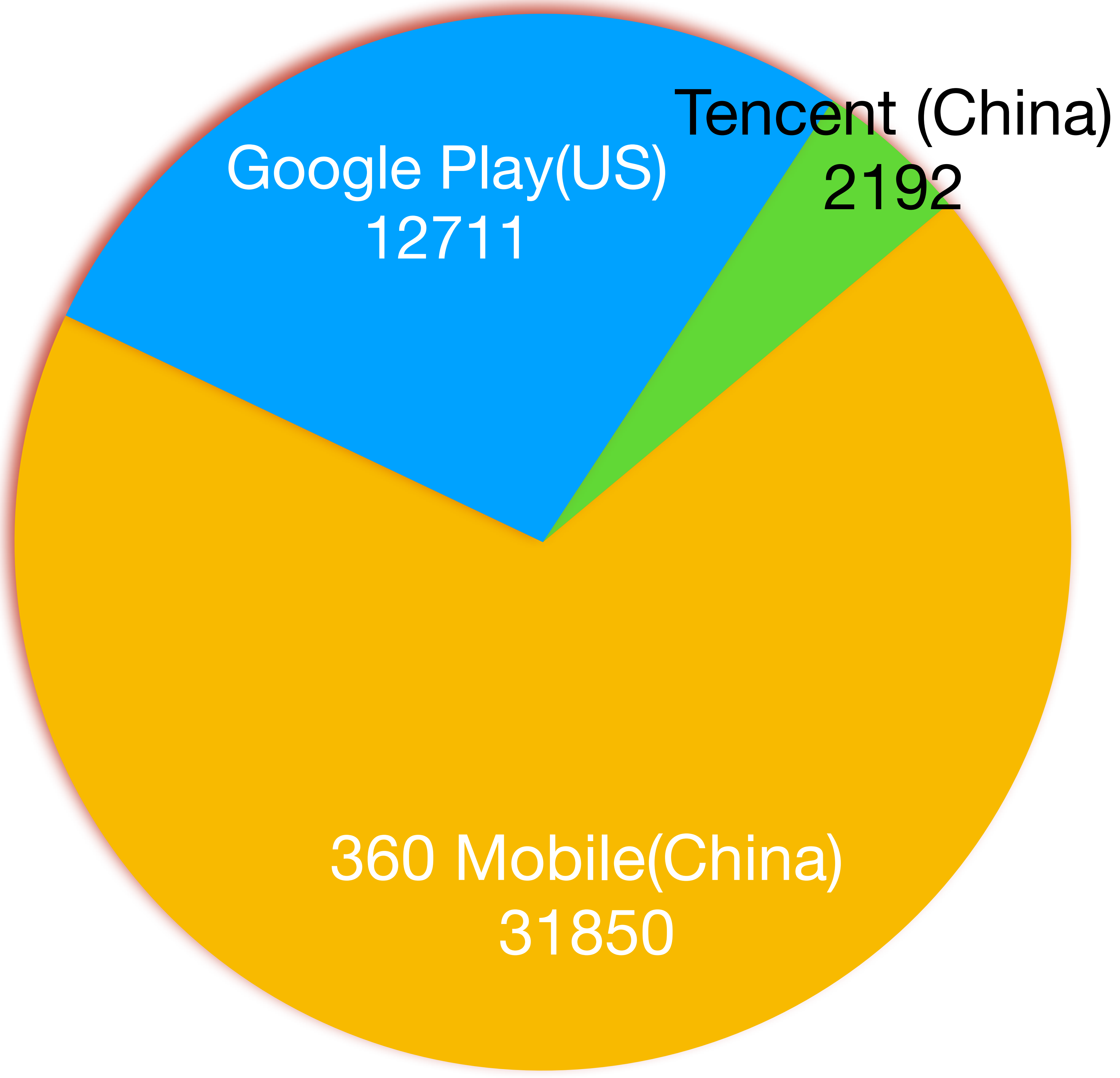
Research Questions

- Q1: How widely is model protection used in apps?
- Q2: How robust are existing model protection techniques?
- Q3: What impact can (stolen) models incur?

Data Collection

- Collect ~45,000 apps from Android App stores in both US and China

- all apps labeled NEW and TRENDING or recently updated



Methodology

Static App Analysis (ModelXRay)

- Analyze whether an app uses on-device Machine Learning (ML)
- Extract information : ML SDK libraries, model files (encrypted or not)

→

Dynamic App Analysis (ModelXtractor)

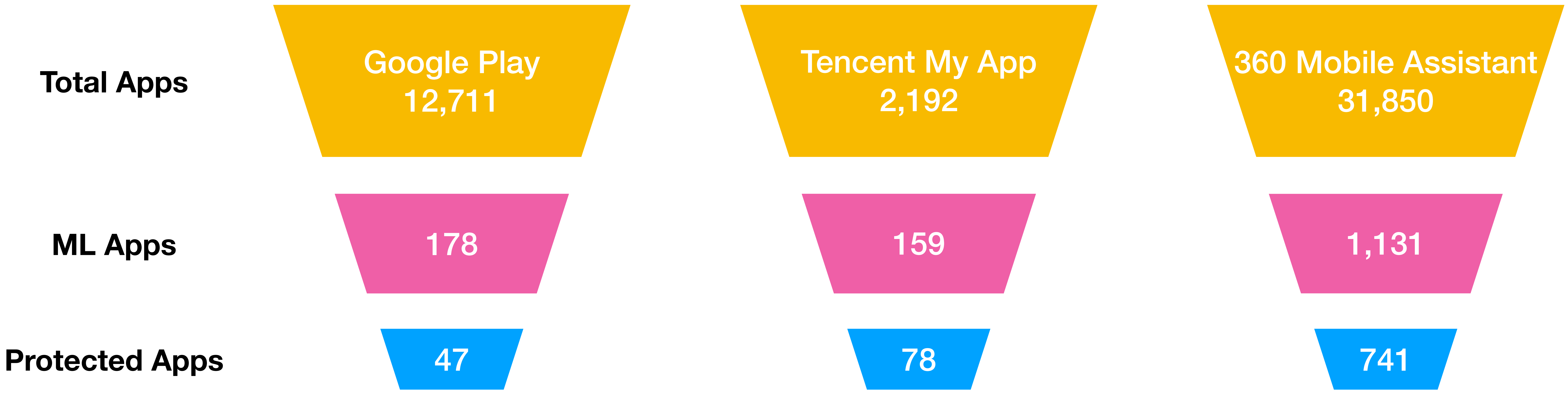
- Instrument the app and run it
- Evaluate how hard is it to steal the decrypted models

ModelXRay is Effective at Discovering ML Apps

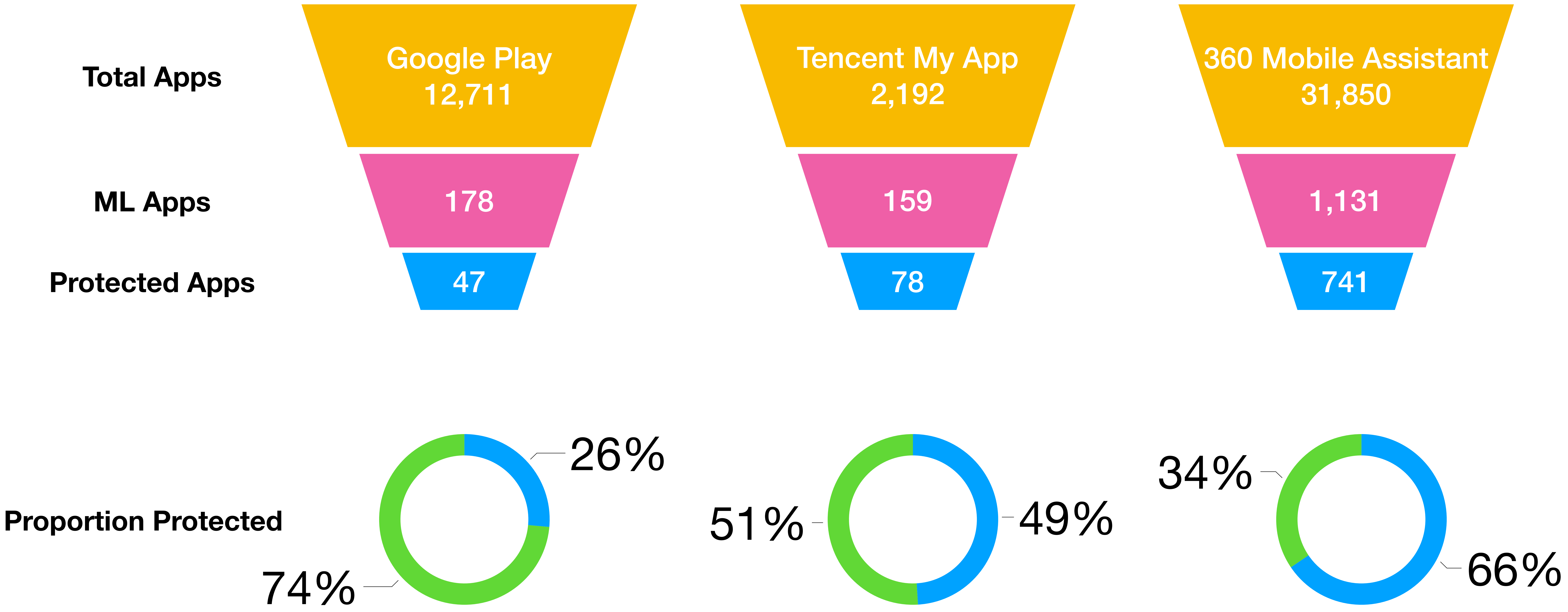
- ModelXRay is simple
  - Identify ML models and libraries with key words matching and filtering
  - Detect encrypted models with file entropy (high entropy—> encrypted ?)
- ModelXRay is effective
  - Identify ML apps (False Positive: 0%, False Negative: 6.4%)
- *Refer to our paper for accuracy analysis*

Q1: How widely is model protection used in apps?

- Among 46,753 apps, 1,468 are ML apps, 866 (59%) of them encrypt models.



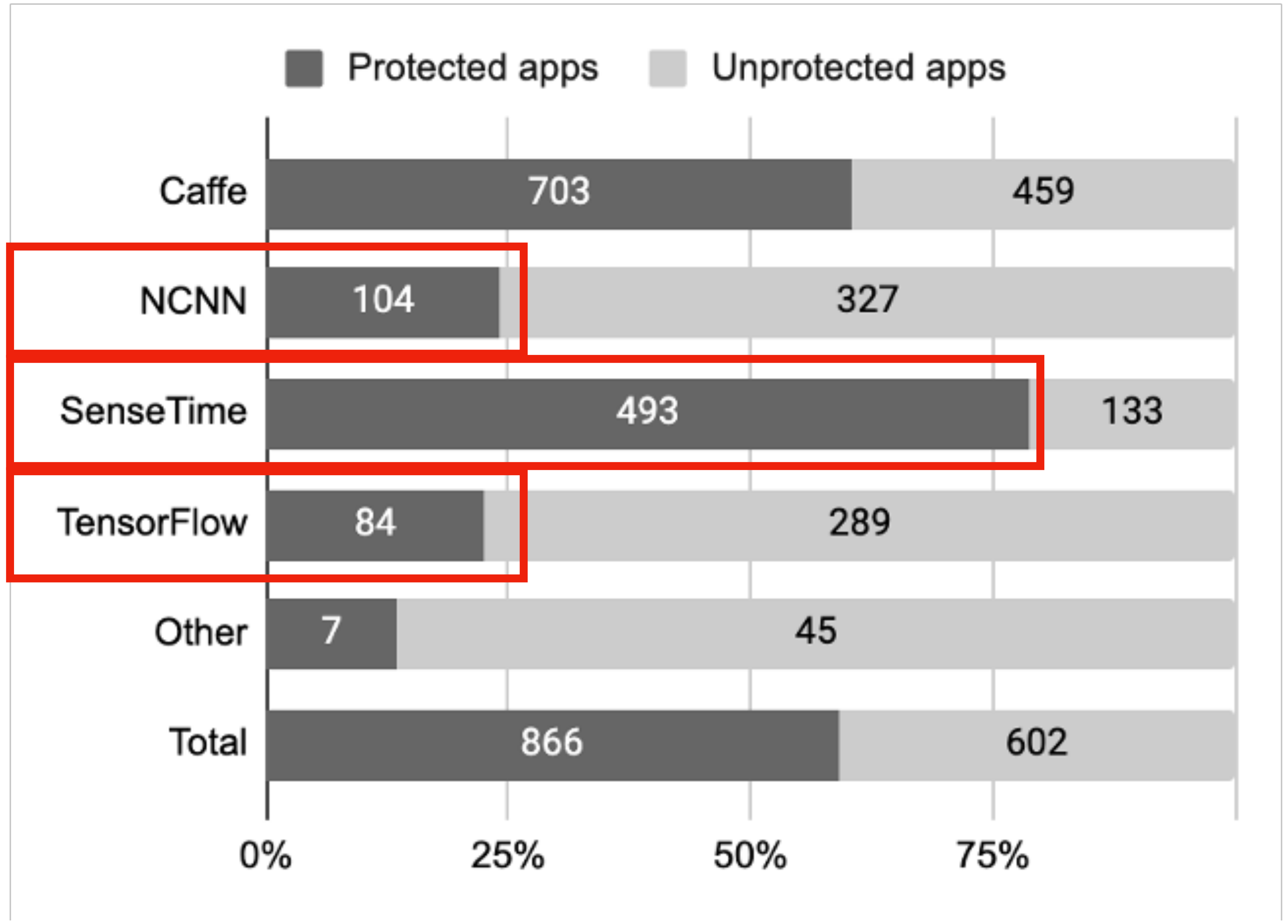
Q1: How widely is model protection used in apps?



Different ML frameworks have different model protection rate

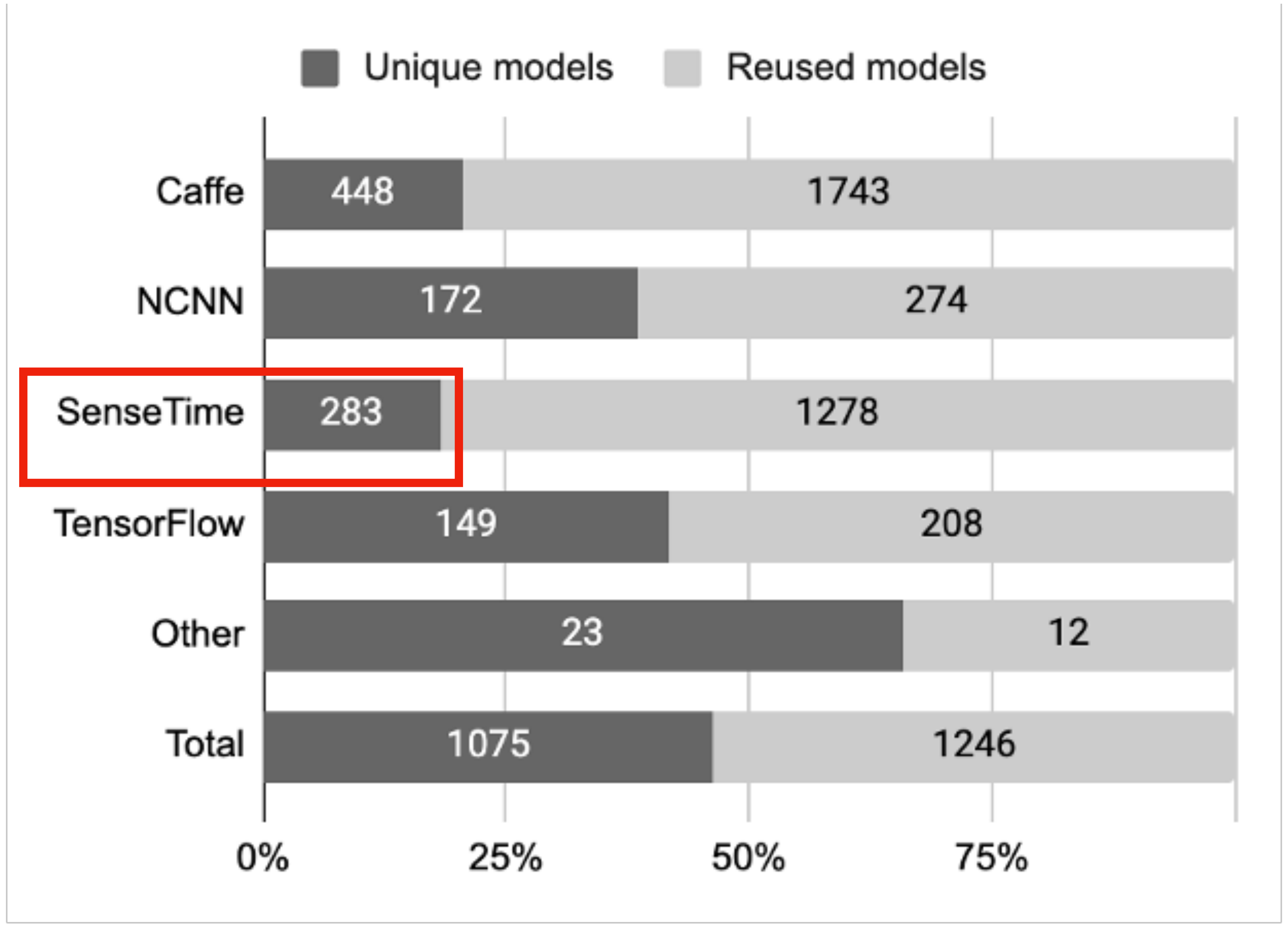
- Open-sourced frameworks like TensorFlow, NCNN have relatively low protection rate (~25%)

- Proprietary framework like SenseTime has higher protection rate (~75%)

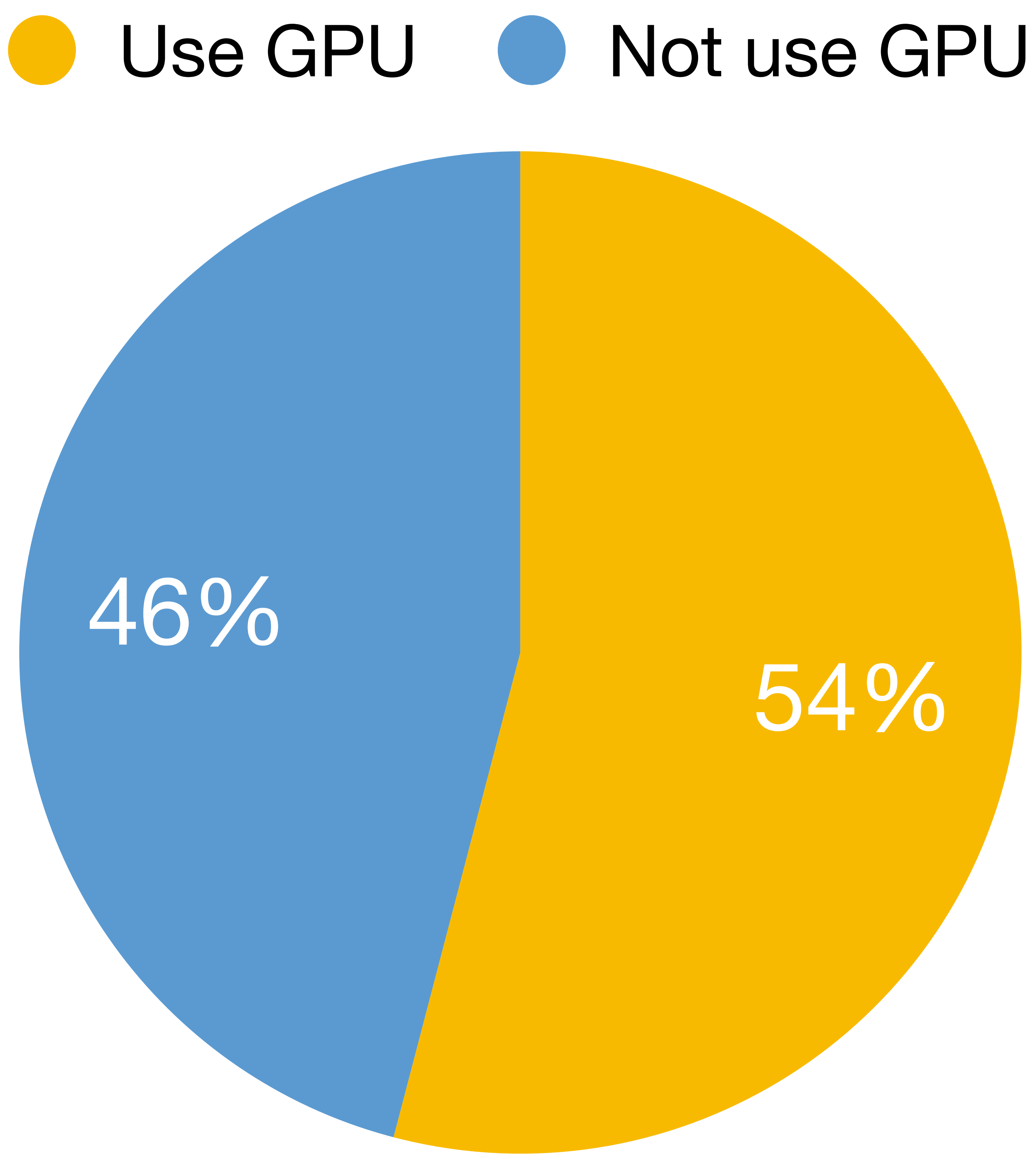


Model reuse is common among different apps

- We use model MD5 hash to identify reuse of models
- Many apps buy licenses from model vendor instead of developing their own models
- Example: for model vendor SenseTime, only ~20% of its observed models are unique

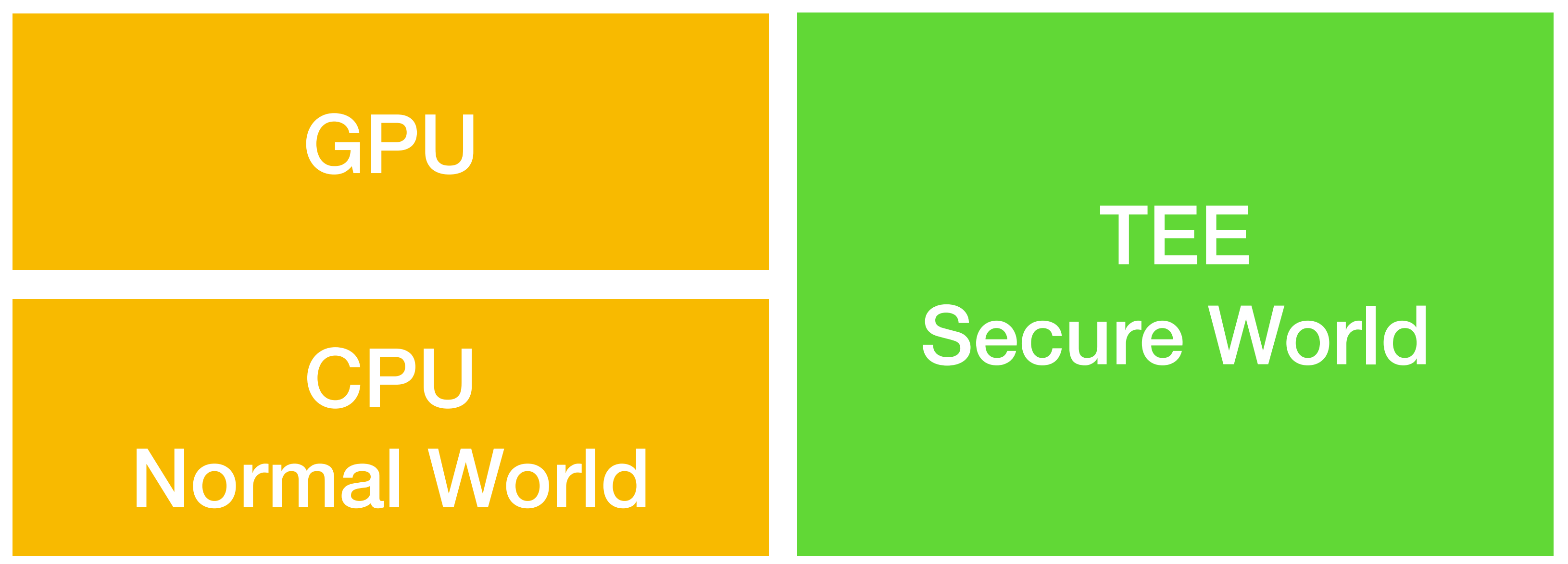


GPU acceleration usage is common for on-device ML



GPU Usage for 1,468 ML Apps

- Security Implication
  - GPU needs to be shared with Non-Secure World, thus not trusted.
  - Make it hard to protect ML models. e.g., simply moving ML into the Secure World will lose access to GPU accelerator



We identify GPU acceleration by checking ML library dependency on GPU library

Q2: How robust are existing model protection techniques?

- We developed ModelXtractor, a dynamic app analysis tool that can extract decrypted models from memory
- Assumption: encrypted models needs to be decrypted in memory before usage
  - We can instrument ML app and dump decrypted model buffers at runtime

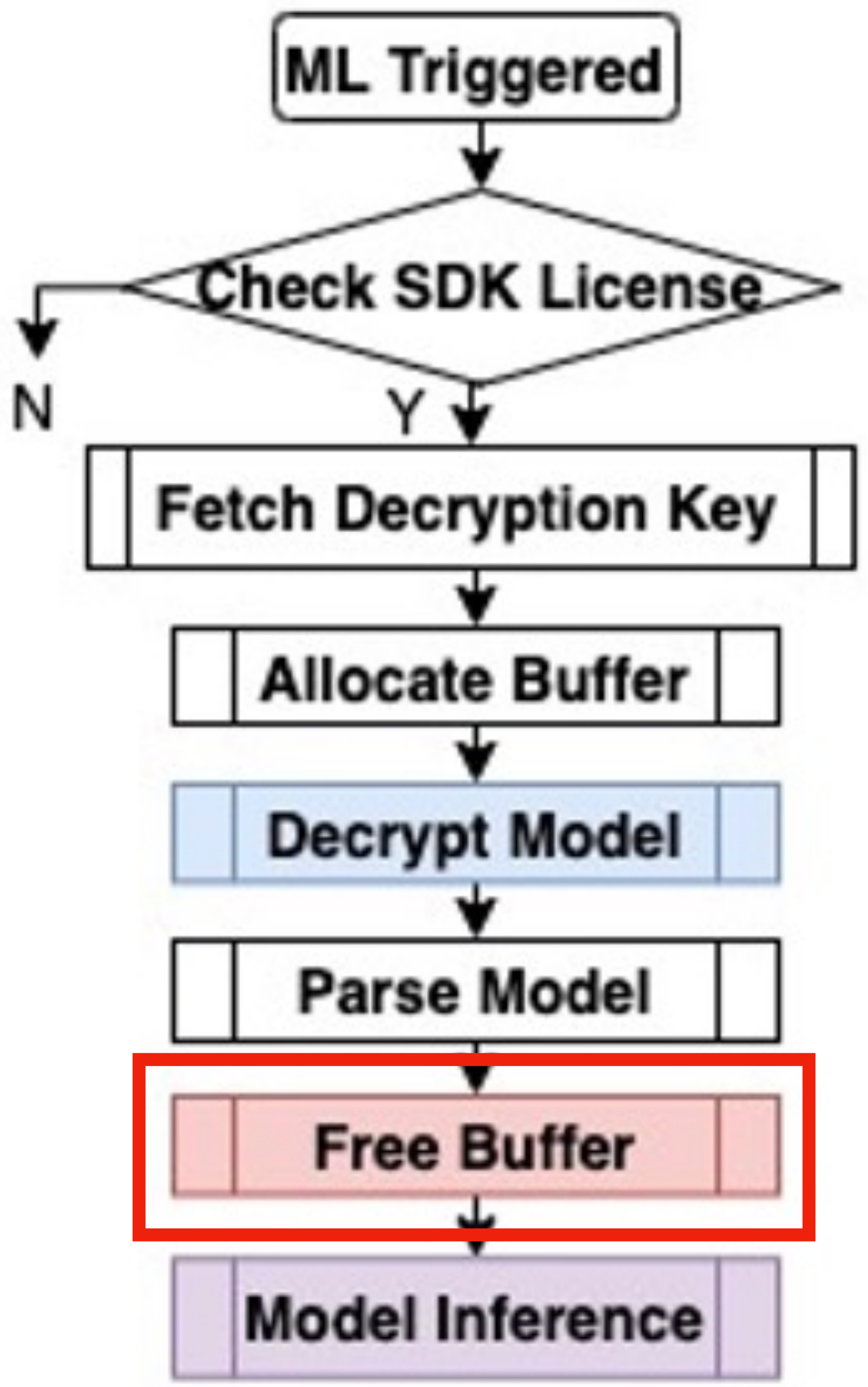
Workflow of ModelXtractor

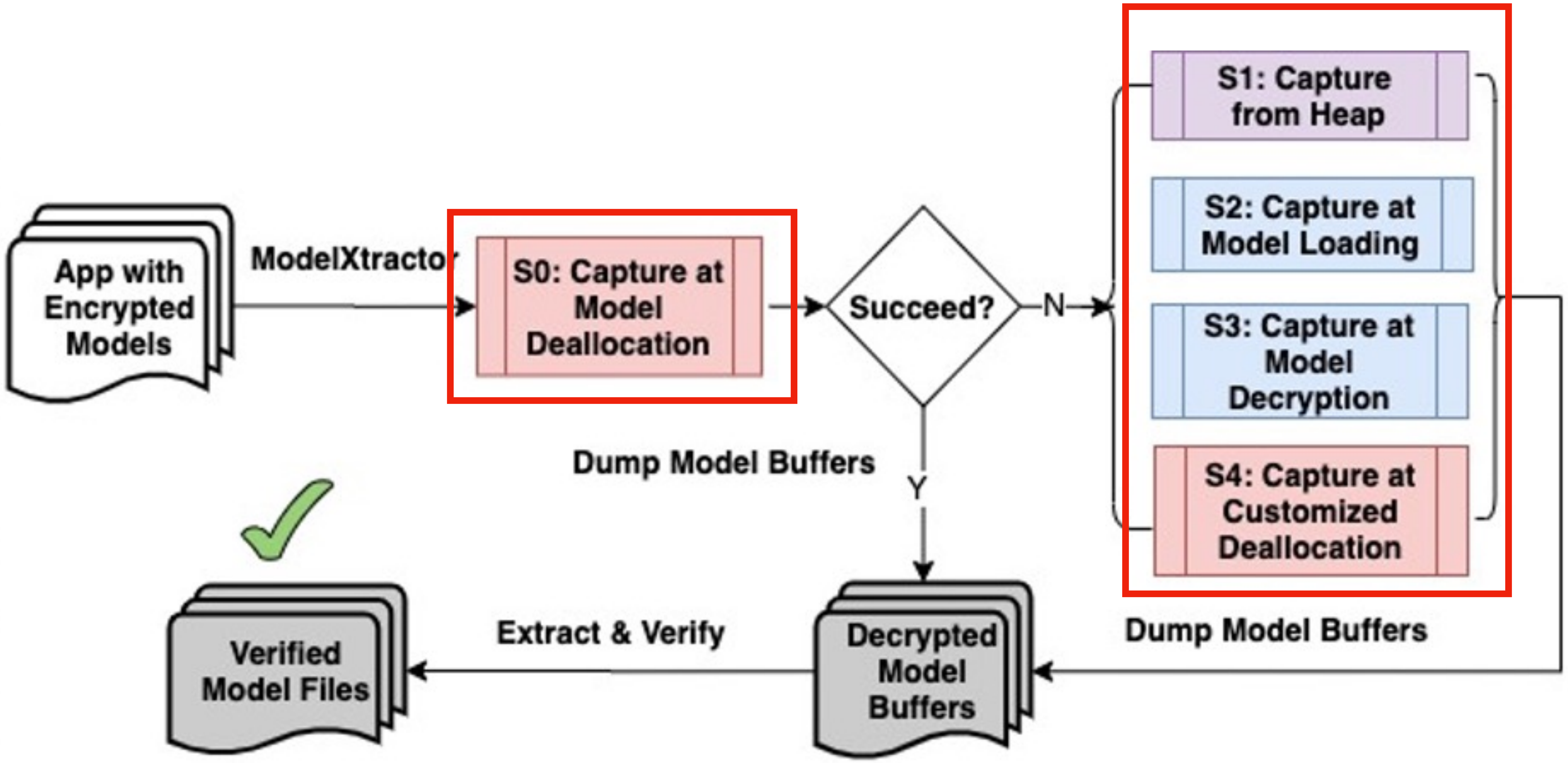
Model loading and decrypting process

Workflow of ModelXtractor

- ModelXtractor has one default strategy(S0) and four alternative strategies (S1-S4)
  - S0 is the most effective one and requires least manual effort
  - S1-S4 is selected when S0 does not work, requires more manual effort

Model loading and decrypting process





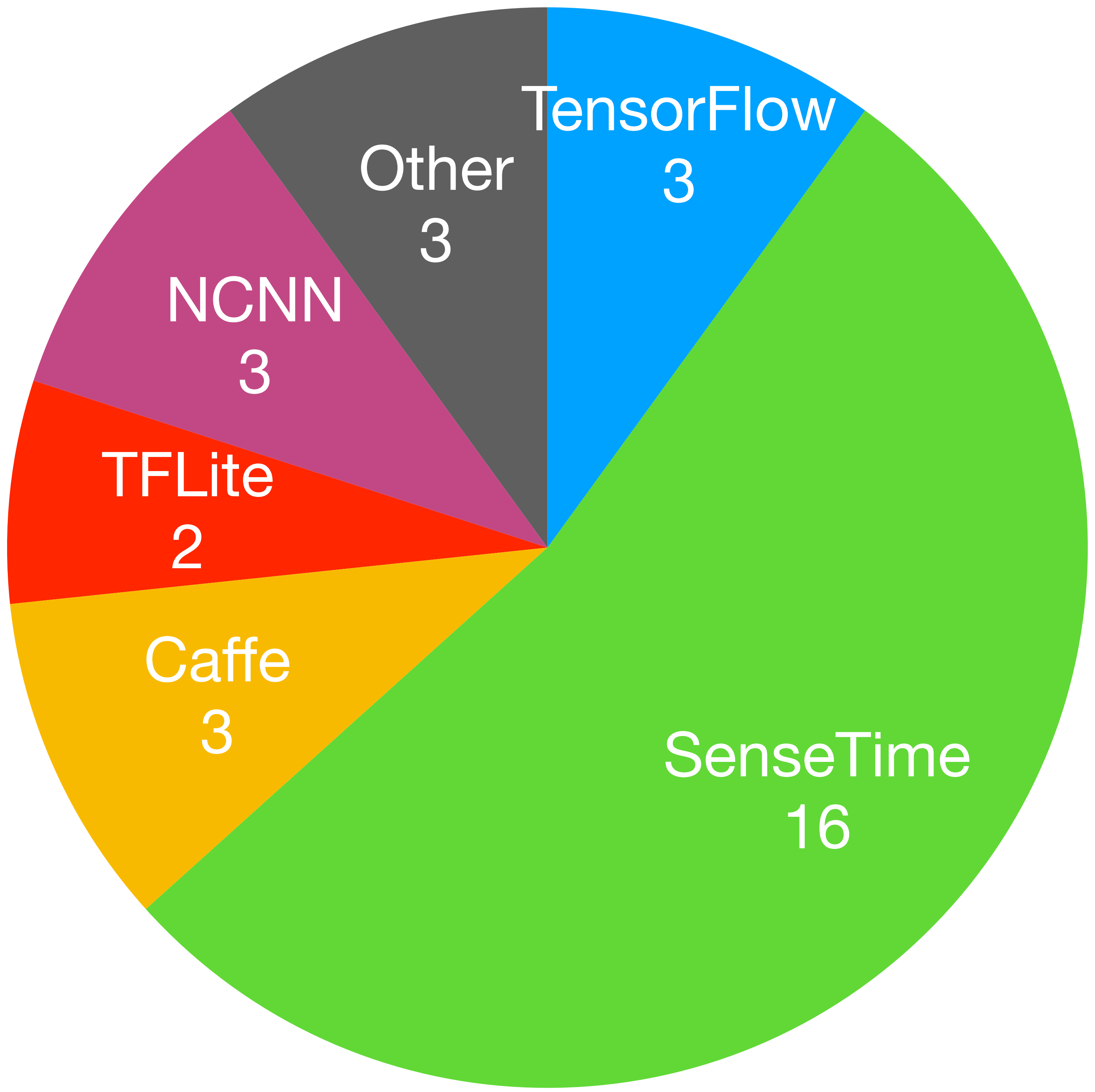
Model extraction workflow

Apply ModelXtractor on real apps

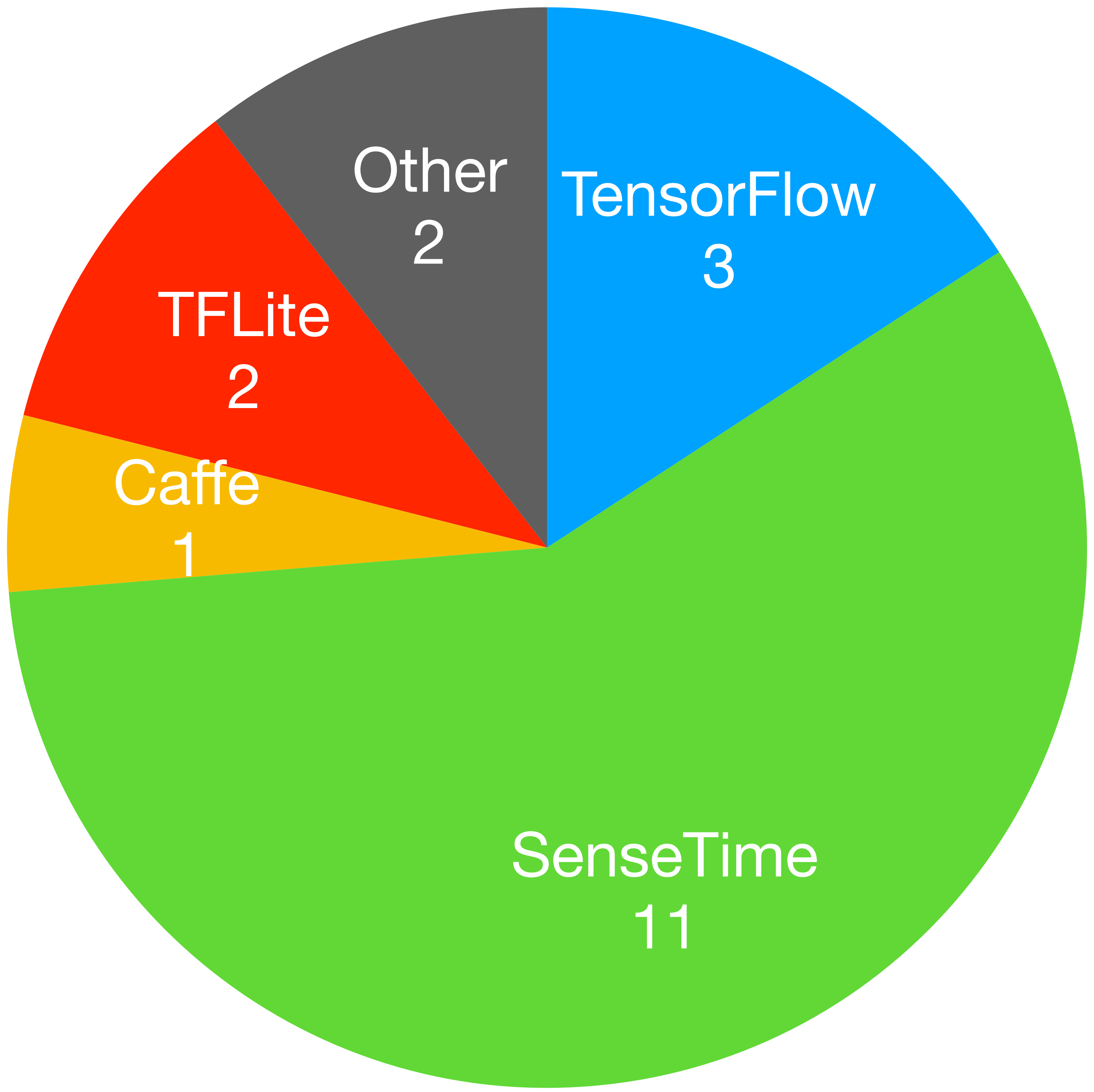
- Which app to analyze?
  - Apps with popular models (highly reused)
    - Maximize the impact of analyzed models
  - Apps that use different ML frameworks
    - Maximize our ML framework coverage

ModelXtractor is Effective at Extracting Models

- We extract models from 18 apps among 29 ML triggered apps
  - Affects 347 ML apps due to model reuse



29 tested apps with ML function triggered



18 Apps with models extracted

Highlights of apps that leaked their models

- 8 apps are downloaded more than 10 million times

App name	Downloads	Framework	Model Functionality	Size (B)	Format	Reuses	Extraction Strategy
Anonymous App 1	300M	TFLite	Liveness Detection	160K	FlatBuffer	18	Freed Buffer
Anonymous App 2	10M	Caffe	Face Tracking	1.5M	Protobuf	4	Model Loading
Anonymous App 3	27M	SenseTime	Face Tracking	2.3M	Protobuf	77	Freed Buffer
Anonymous App 4	100K	SenseTime	Face Filter	3.6M	Protobuf	3	Freed Buffer
Anonymous App 5	100M	SenseTime	Face Filter	1.4M	Protobuf	2	Freed Buffer
Anonymous App 6	10K	TensorFlow	OCR	892K	Protobuf	2	Memory Dumping
Anonymous App 7	10M	TensorFlow	Photo Process	6.5M	Protobuf	1	Freed Buffer
Anonymous App 8	10K	SenseTime	Face Track	1.2M	Protobuf	5	Freed Buffer
Anonymous App 9	5.8M	Caffe	Face Detect	60K	Protobuf	77	Freed Buffer
Anonymous App 10	10M	Face++	Liveness	468K	Unknown	17	Freed Buffer
Anonymous App 11	100M	SenseTime	Face Detect	1.7M	Protobuf	18	Freed Buffer
Anonymous App 12	492K	Baidu	Face Tracking	2.7M	Unknown	26	Freed Buffer
Anonymous App 13	250K	SenseTime	ID card	1.3M	Unknown	13	Freed Buffer
Anonymous App 14	100M	TFLite	Camera Filter	228K	Json	1	Freed Buffer
Anonymous App 15	5K	TensorFlow	Malware Classification	20M	Protobuf	1	Decryption Buffer

Note: 1) We excluded some apps that dumped the same models as reported above; 2) We anonymized name of the apps to protect the user's security; 3) Every app has several models for different functionalities, we only list one representative model for each app

Highlights of apps that leaked their models

- They are from 6 different ML frameworks including TensorFlow, Caffe, SenseTime, Face++, Baidu, etc

App name	Downloads	Framework	Model Functionality	Size (B)	Format	Reuses	Extraction Strategy
Anonymous App 1	300M	TFLite	Liveness Detection	160K	FlatBuffer	18	Freed Buffer
Anonymous App 2	10M	Caffe	Face Tracking	1.5M	Protobuf	4	Model Loading
Anonymous App 3	27M	SenseTime	Face Tracking	2.3M	Protobuf	77	Freed Buffer
Anonymous App 4	100K	SenseTime	Face Filter	3.6M	Protobuf	3	Freed Buffer
Anonymous App 5	100M	SenseTime	Face Filter	1.4M	Protobuf	2	Freed Buffer
Anonymous App 6	10K	TensorFlow	OCR	892K	Protobuf	2	Memory Dumping
Anonymous App 7	10M	TensorFlow	Photo Process	6.5M	Protobuf	1	Freed Buffer
Anonymous App 8	10K	SenseTime	Face Track	1.2M	Protobuf	5	Freed Buffer
Anonymous App 9	5.8M	Caffe	Face Detect	60K	Protobuf	77	Freed Buffer
Anonymous App 10	10M	Face++	Liveness	468K	Unknown	17	Freed Buffer
Anonymous App 11	100M	SenseTime	Face Detect	1.7M	Protobuf	18	Freed Buffer
Anonymous App 12	492K	Baidu	Face Tracking	2.7M	Unknown	26	Freed Buffer
Anonymous App 13	250K	SenseTime	ID card	1.3M	Unknown	13	Freed Buffer
Anonymous App 14	100M	TFLite	Camera Filter	228K	Json	1	Freed Buffer
Anonymous App 15	5K	TensorFlow	Malware Classification	20M	Protobuf	1	Decryption Buffer

Note: 1) We excluded some apps that dumped the same models as reported above; 2) We anonymized name of the apps to protect the user's security; 3) Every app has several models for different functionalities, we only list one representative model for each app

Highlights of apps that leaked their models

- 7 of them has models reused more than 10 times

App name	Downloads	Framework	Model Functionality	Size (B)	Format	Reuses	Extraction Strategy
Anonymous App 1	300M	TFLite	Liveness Detection	160K	FlatBuffer	18	Freed Buffer
Anonymous App 2	10M	Caffe	Face Tracking	1.5M	Protobuf	4	Model Loading
Anonymous App 3	27M	SenseTime	Face Tracking	2.3M	Protobuf	77	Freed Buffer
Anonymous App 4	100K	SenseTime	Face Filter	3.6M	Protobuf	3	Freed Buffer
Anonymous App 5	100M	SenseTime	Face Filter	1.4M	Protobuf	2	Freed Buffer
Anonymous App 6	10K	TensorFlow	OCR	892K	Protobuf	2	Memory Dumping
Anonymous App 7	10M	TensorFlow	Photo Process	6.5M	Protobuf	1	Freed Buffer
Anonymous App 8	10K	SenseTime	Face Track	1.2M	Protobuf	5	Freed Buffer
Anonymous App 9	5.8M	Caffe	Face Detect	60K	Protobuf	77	Freed Buffer
Anonymous App 10	10M	Face++	Liveness	468K	Unknown	17	Freed Buffer
Anonymous App 11	100M	SenseTime	Face Detect	1.7M	Protobuf	18	Freed Buffer
Anonymous App 12	492K	Baidu	Face Tracking	2.7M	Unknown	26	Freed Buffer
Anonymous App 13	250K	SenseTime	ID card	1.3M	Unknown	13	Freed Buffer
Anonymous App 14	100M	TFLite	Camera Filter	228K	Json	1	Freed Buffer
Anonymous App 15	5K	TensorFlow	Malware Classification	20M	Protobuf	1	Decryption Buffer

Note: 1) We excluded some apps that dumped the same models as reported above; 2) We anonymized name of the apps to protect the user's security; 3) Every app has several models for different functionalities, we only list one representative model for each app

Highlights of apps that leaked their models

- 12 of them are extracted with our default strategy from freed buffer.

App name	Downloads	Framework	Model Functionality	Size (B)	Format	Reuses	Extraction Strategy
Anonymous App 1	300M	TFLite	Liveness Detection	160K	FlatBuffer	18	Freed Buffer
Anonymous App 2	10M	Caffe	Face Tracking	1.5M	Protobuf	4	Model Loading
Anonymous App 3	27M	SenseTime	Face Tracking	2.3M	Protobuf	77	Freed Buffer
Anonymous App 4	100K	SenseTime	Face Filter	3.6M	Protobuf	3	Freed Buffer
Anonymous App 5	100M	SenseTime	Face Filter	1.4M	Protobuf	2	Freed Buffer
Anonymous App 6	10K	TensorFlow	OCR	892K	Protobuf	2	Memory Dumping
Anonymous App 7	10M	TensorFlow	Photo Process	6.5M	Protobuf	1	Freed Buffer
Anonymous App 8	10K	SenseTime	Face Track	1.2M	Protobuf	5	Freed Buffer
Anonymous App 9	5.8M	Caffe	Face Detect	60K	Protobuf	77	Freed Buffer
Anonymous App 10	10M	Face++	Liveness	468K	Unknown	17	Freed Buffer
Anonymous App 11	100M	SenseTime	Face Detect	1.7M	Protobuf	18	Freed Buffer
Anonymous App 12	492K	Baidu	Face Tracking	2.7M	Unknown	26	Freed Buffer
Anonymous App 13	250K	SenseTime	ID card	1.3M	Unknown	13	Freed Buffer
Anonymous App 14	100M	TFLite	Camera Filter	228K	Json	1	Freed Buffer
Anonymous App 15	5K	TensorFlow	Malware Classification	20M	Protobuf	1	Decryption Buffer

Note: 1) We excluded some apps that dumped the same models as reported above; 2) We anonymized name of the apps to protect the user's security; 3) Every app has several models for different functionalities, we only list one representative model for each app

Model Vendors Are Trying Hard to Protect Models

- **Encrypting both code and model**
  - An OCR SDK, code written in javascript, which is also encrypted
- **Encrypting feature vector and its sequence**
  - A malware detection app with decision tree model
  - Feature vector has 1000 features, sequence is critical to use the model
- **Encrypting models multiple times**
  - An app with several liveness detection related models
  - Several models are encrypted and packed, then encrypted again

Q3: What impact can (stolen) models incur?

- **Financial impact (millions of dollars)**
  - **Attacker save R&D cost & model license fee**
  - **Model vendors lose competition and pricing advantage**

- **Security impact**
  - **Bypass model-based authentication: liveness detection to verify real person**
  - **Private user information of training data leaked due to membership inference attacks**

Existing methods for model protection are vulnerable

- **File encryption**, which can be easily extracted from memory after decryption

- **Obfuscation**, which does not prevent reuse of the model
    - For example: MAZE ML framework can compile the model into a binary to obfuscate the model

- **Undisclosed model format**, which still suffers from documentation leakage or reverse engineering

Responsible Disclosure

- Contact major vendors
  - 12 major vendors contacted, including Google, Facebook, Tencent, SenseTime and etc.
  - 5 responded.

Vendors already protect models	Vendors do not protect models
-internal discussion on improving the model security **-Seeking advice and collaboration**	2 vendors :unaware of leakage or the impact 2 vendors :aware of impact, but no good solution

Summary

- 60% ML apps protect their models
- 2/3 analyzed apps with encrypted models suffers from our unsophisticated analysis, affecting 300+ protected ML apps.
- Model leakage has both financial and security impact.

We need more research into protecting on-device ML models to mitigate this serious privacy problem.

Thank you!

Connect with the authors!

Zhichuang Sun
sun.zhi@northeastern.edu

Ruimin Sun
r.sun@northeastern.edu

Our project is open-sourced on Github!
https://github.com/RiS3-Lab/ModelXRay

Long Lu
l.lu@northeastern.edu

Alan Mislove
amislove@ccs.neu.edu

Q&A