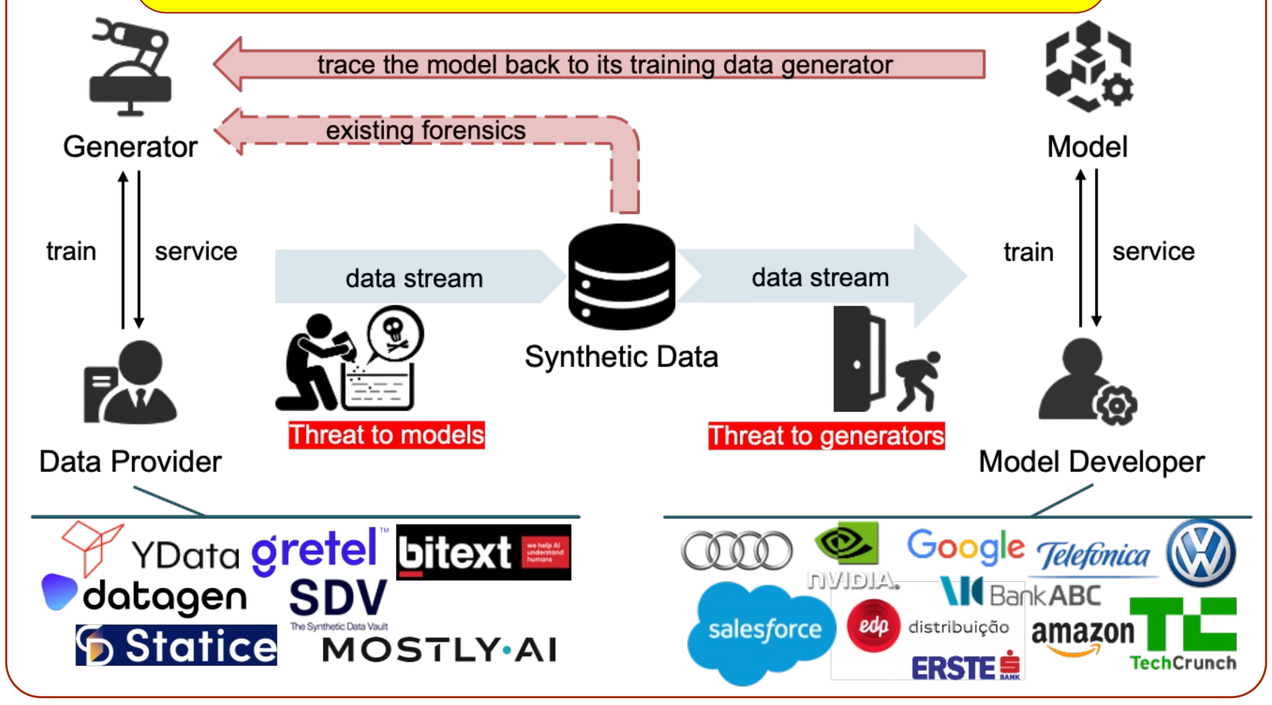
Detection and Attribution of Models Trained on Generated Data

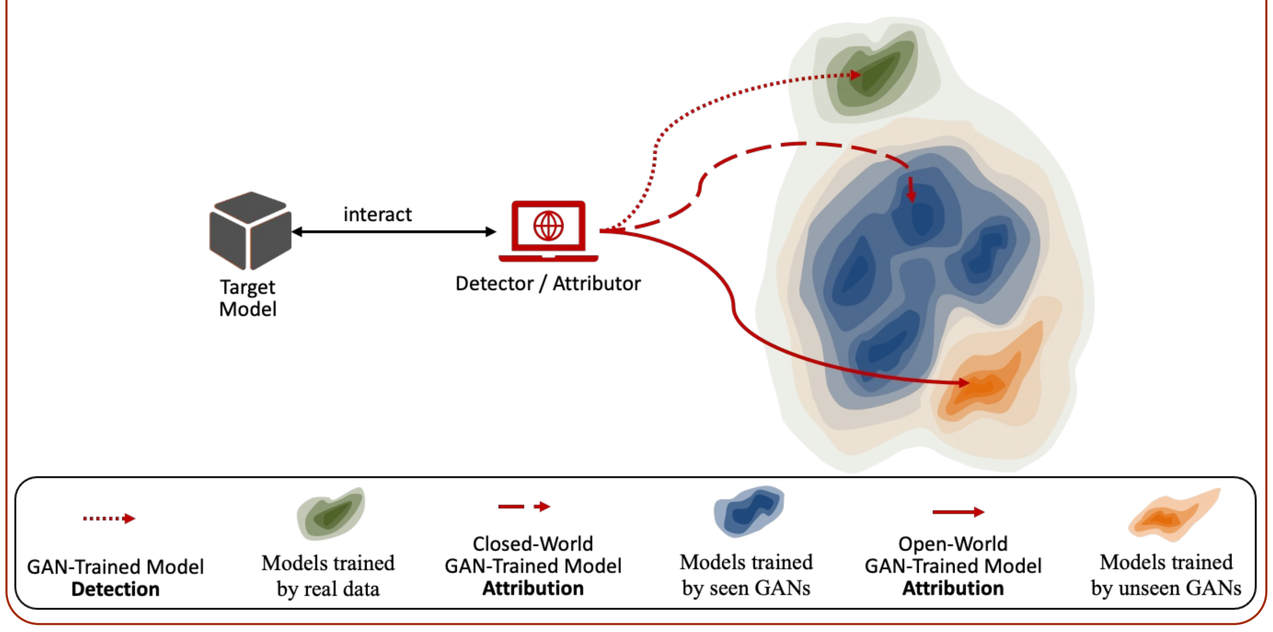
Ge Han[1], Ahmed Salem[2], Zheng Li[3], Shanqing Guo[1]

[1] Shandong University; [2] Azure Research; [3] Helmholtz Center for Information Security (CISPA)

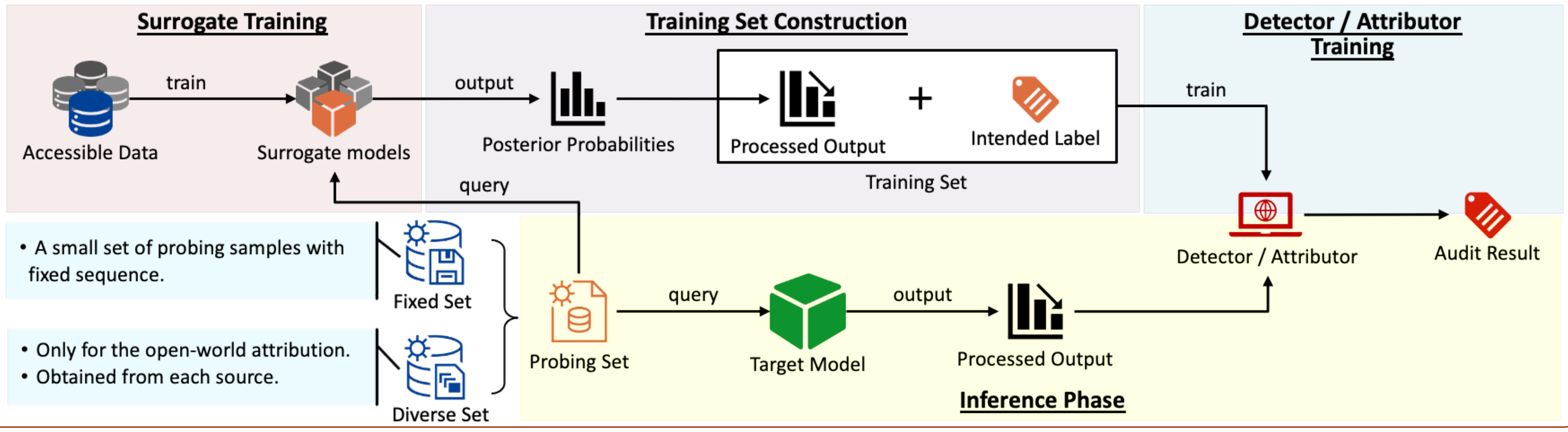
Training models on generated data?



Contributions



General Methodology



Setup

Detector / Attributor I: sorted output

Detector / Attributor II: sorted output + 1-bit correctness

Detector / Attributor III: unsorted output + ground-truth label

Why do they work?

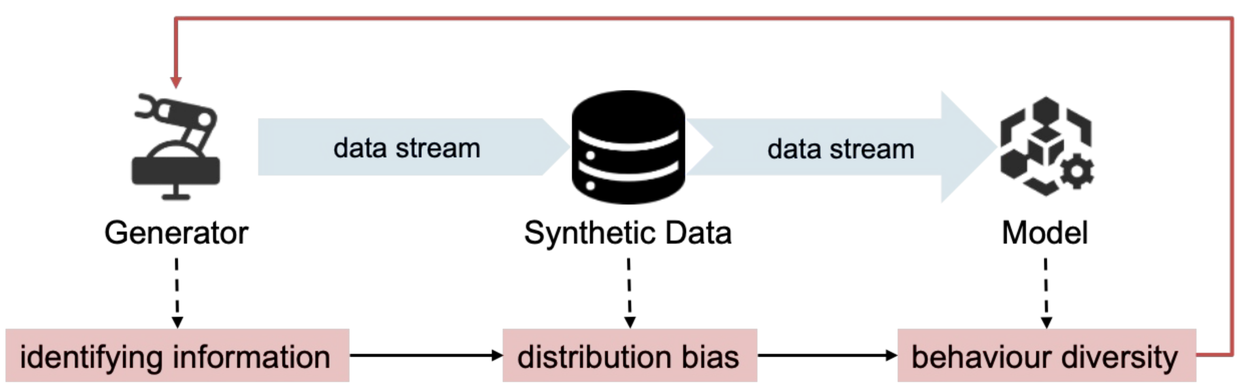
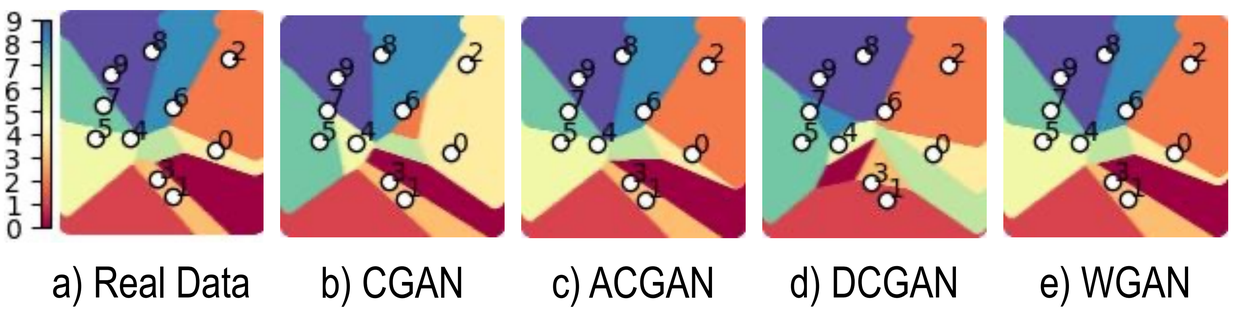


Figure I Intuitive explanation.



a) Real Data b) CGAN c) ACGAN d) DCGAN e) WGAN

Figure II The visualization of decision boundaries for VGG-9 models trained from different data sources.

Detection

Table I The Accuracy of Model Detection.

Dataset	Detector I	Detector II	Detector III
CelebA	0.705	**0.948**	0.933
FMNIST	0.691	**0.926**	0.850
SVHN	0.689	**0.942**	0.915

Attribution

Table II The Accuracy of Closed-World Model Attribution.

Dataset	Detector I	Detector II	Detector III
CelebA	0.649	0.629	**0.851**
FMNIST	0.589	0.569	**0.860**
SVHN	0.640	0.625	**0.873**

Table III The Accuracy of Open-World Model Attribution.

Dataset	Detector I	Detector II	Detector III
CelebA	0.871	0.918	**0.955**
FMNIST	0.912	**0.974**	0.967
SVHN	0.754	0.791	**0.820**

SHANDONG UNIVERSITY · Microsoft Azure · CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY · ICASSP 2024 KOREA