

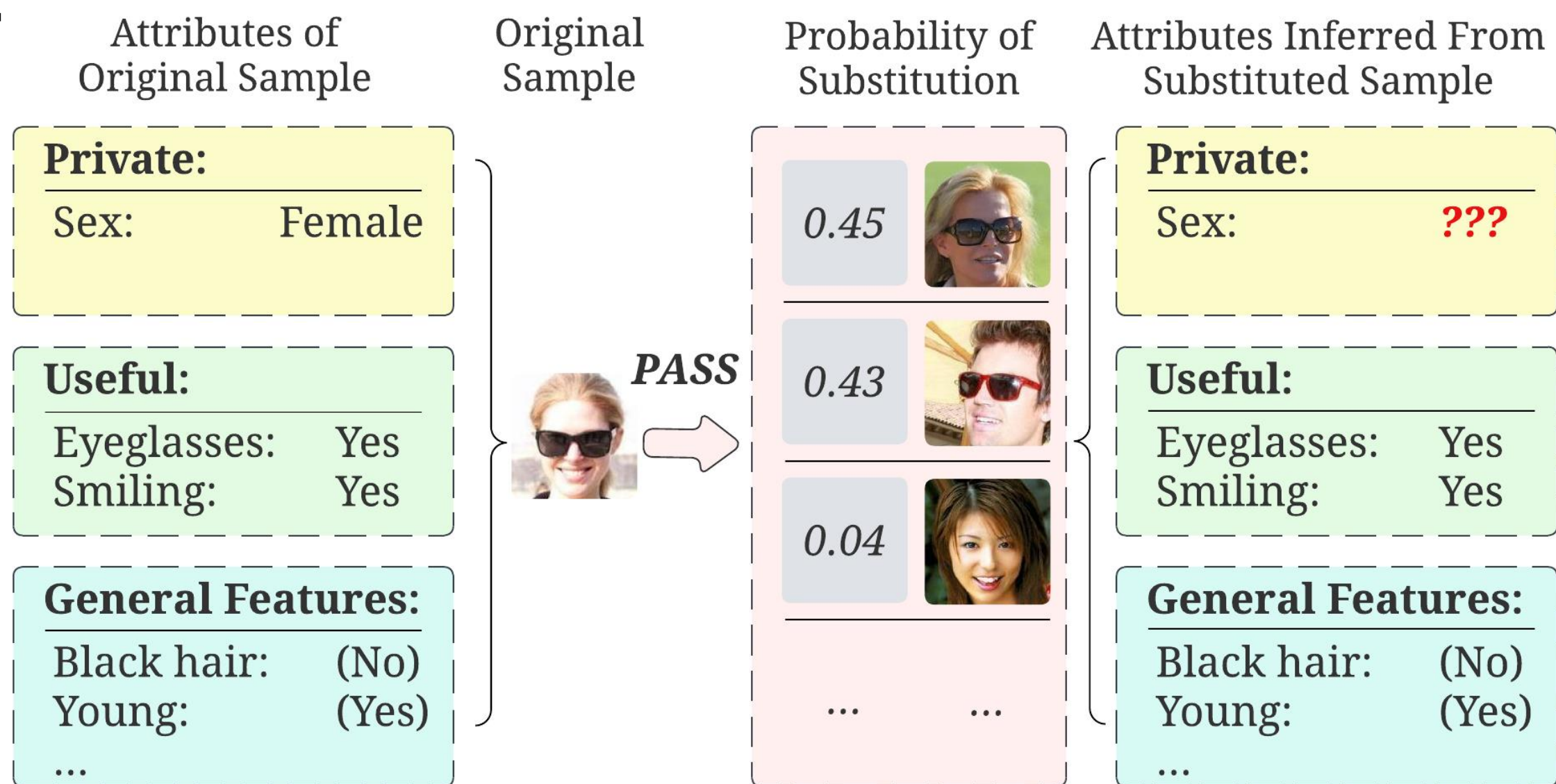


Take Away:

- Our goal: Protect **private** attributes while preserving the **utility** of the data for downstream tasks in data sharing or ML pipelines.
- Problem: We show that existing **adversarial training** based methods are **vulnerable** to slightly stronger or unseen attackers.
- Solution: We propose **PASS**, a **stochastic data substitution** based method that overcomes this common problem.

Motivation:

- Our goal:

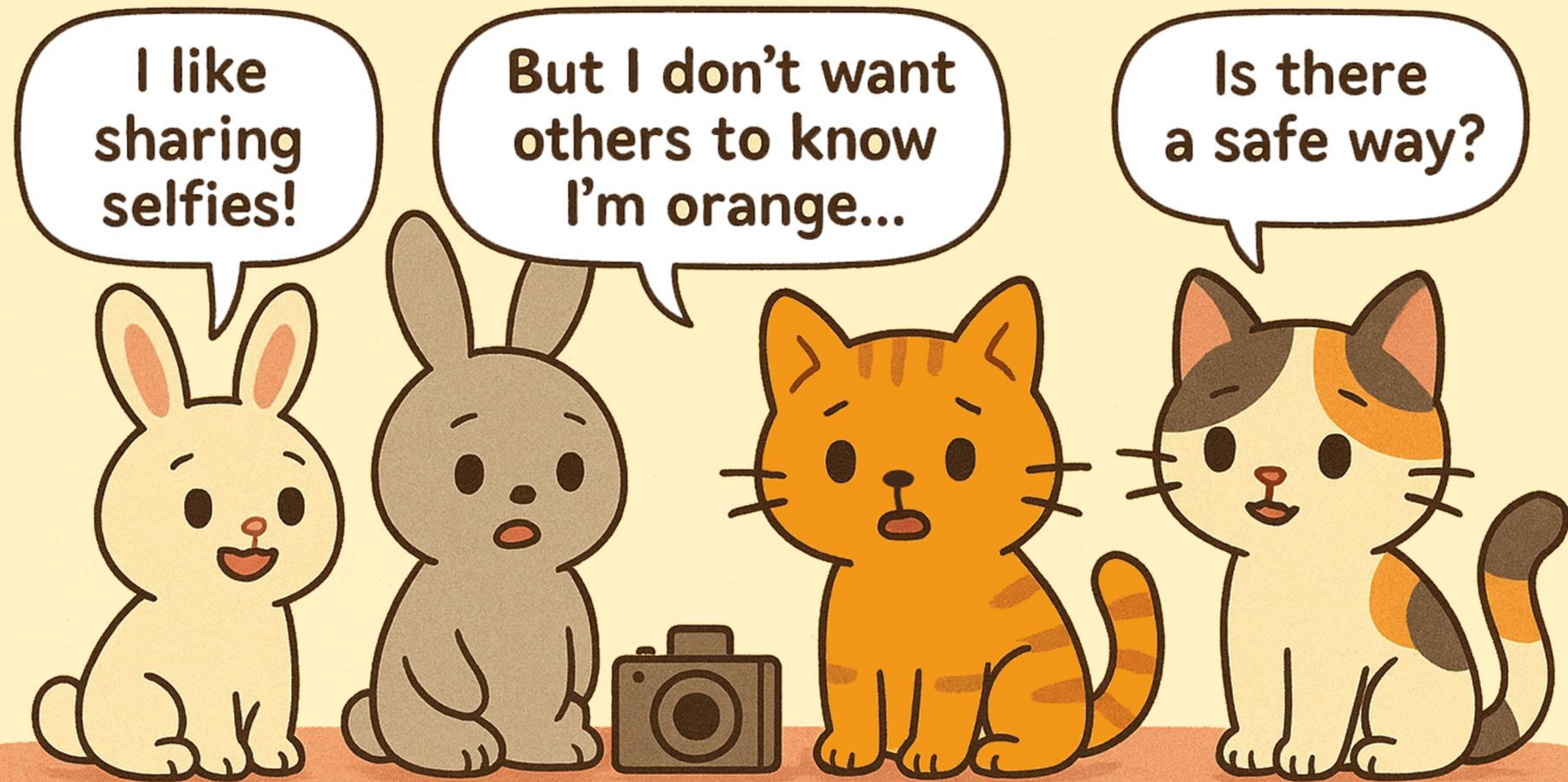


- SOTA methods: obfuscate the data based on **adversarial training**, where they train the **data obfuscation model** to confuse an **adversarial classifier jointly** trained to infer each private attribute.
- Problem: These methods are **vulnerable**...
 - Theoretically**, from an information theory perspective.
 - Empirically**, to a **simple** attacking strategy called the **Probing Attack**, where the attacker applies the (black-box) obfuscation algorithm to a public dataset with labeled private attributes, and then uses the resulting obfuscated samples to train a new classifier.

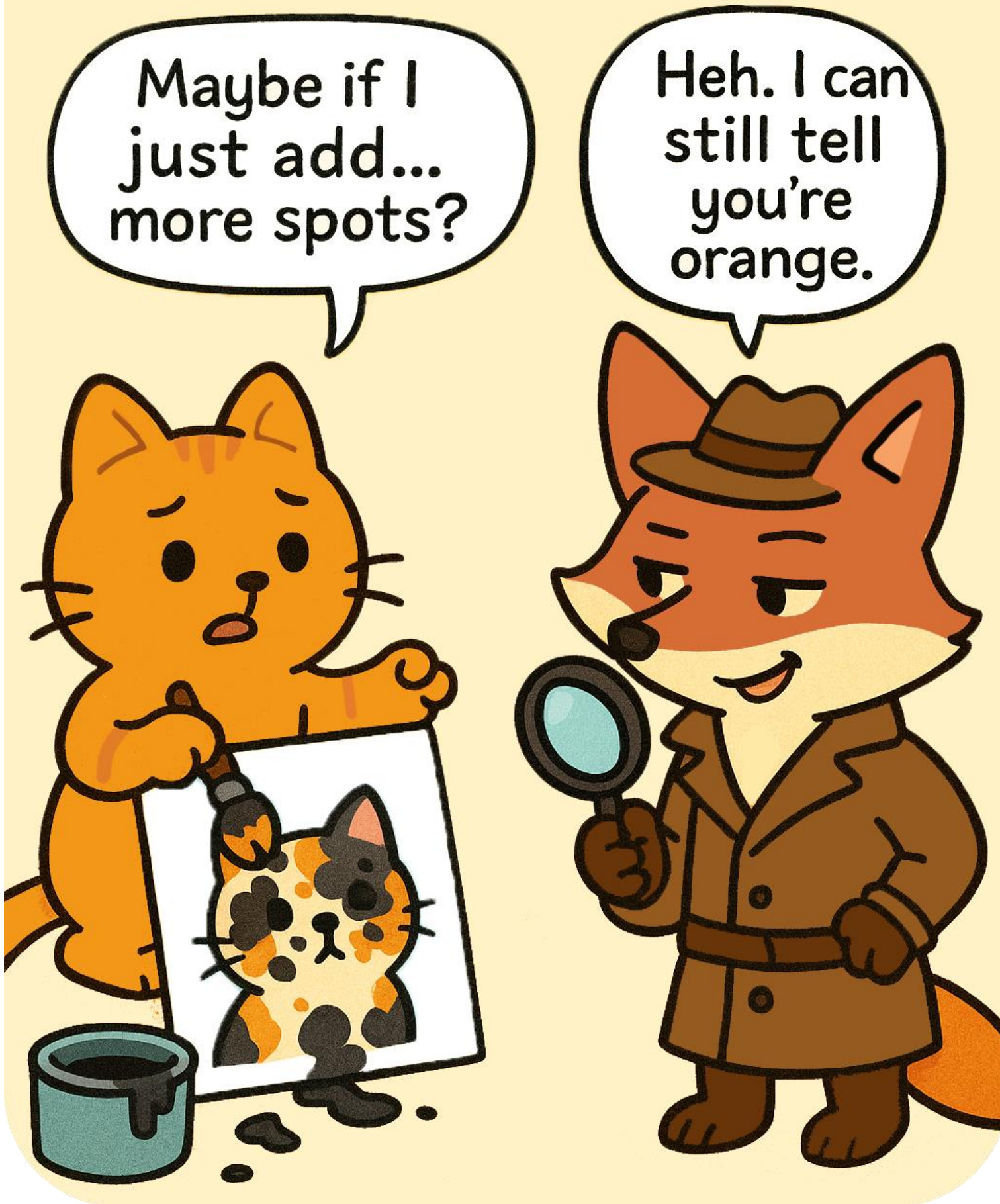
JPMorganChase Disclaimer:

This paper was prepared for informational purposes with contributions from the Global Technology Applied Research center of JPMorgan Chase & Co. This paper is not a product of the Research Department of JPMorgan Chase & Co. or its affiliates. Neither JPMorgan Chase & Co. nor any of its affiliates makes any explicit or implied representation or warranty and none of them accept any liability in connection with this paper, including, without limitation, with respect to the completeness, accuracy, or reliability of the information contained herein and the potential legal, compliance, tax, or accounting effects thereof. This document is not intended as investment research or investment advice, or as a recommendation, offer, or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction.

① Can I share my selfie...without sharing my color?



② Disguises can be seen through by a smarter fox



③ Stochastic data substitution works!



50%-50%

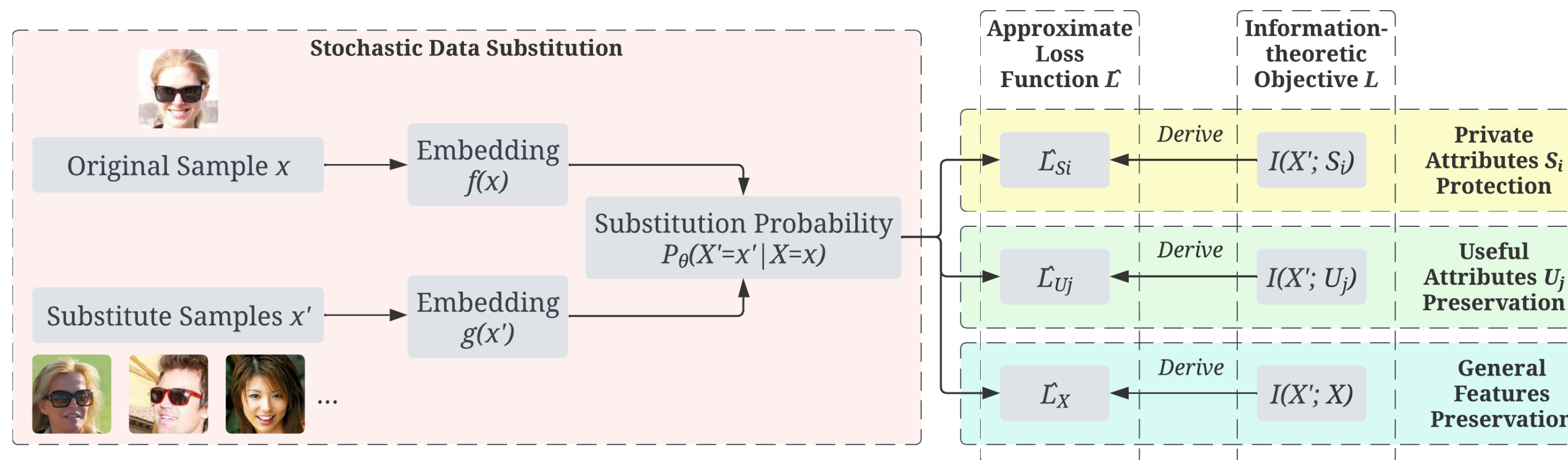


Approach:

- Information-theoretic formulation of our goal:

$$\min_{P_{\theta}(X'|X)} L = \sum_{i=1}^M I(X'; S_i) - \lambda \sum_{j=1}^N I(X'; U_j) - \mu I(X'; X)$$

- PASS**: stochastically substitute each sample with another one according to cosine similarity in an embedding space.



- Theoretical Grounds:

- PASS's training objective** is derived soundly from the information-theoretic definition of our goal.
- PASS can also be interpreted within **Differential Privacy** framework, as a generalized **randomized response** method.
- PASS has information-theoretic **operational boundary** when the private and useful attributes are **entangled**.

Experiments:

- Outperforms baselines on **CelebA**, **AudioMNIST** and **MotionSense**.

Method	Private		Useful		Results on CelebA		"Hidden" Useful	
					NAG (%)			
	Male (↓)	Smiling (↑)	Young (↑)	Attractive (↑)	Mouth_Slightly_Open (↑)	High_Cheekbones (↑)	mNAG (%) (↑)	
ADV	99.9±0.1	98.8±0.1	97.0±0.9	94.6±0.4	99.1±0.1	97.0±0.5	-2.6±0.2	
GAP	83.0±1.1	75.9±1.3	45.4±3.0	77.6±1.1	61.1±2.1	75.6±0.7	-15.9±2.3	
MSDA	91.6±0.7	99.8±0.2	92.4±2.4	89.9±1.0	91.8±0.8	95.7±1.1	2.3±0.8	
BDQ	99.7±0.1	98.8±0.2	96.3±0.8	94.1±0.6	98.9±0.4	97.0±0.3	-2.7±0.2	
PPDAR	99.7±0.1	98.9±0.3	97.2±1.2	94.4±0.6	99.0±0.1	97.0±0.4	-2.4±0.3	
MaSS	96.9±0.1	97.2±0.2	86.2±1.4	90.6±0.3	97.6±0.2	94.6±0.4	-3.7±0.4	
PASS	4.9±0.5	98.3±0.1	78.6±0.8	58.1±2.8	67.0±0.8	86.7±0.3	72.9±0.2	

NAG (Normalized Accuracy Gain): NAG=0 -> "fully protected", NAG=1 -> "not protected".
mNAG: NAG averaged over useful attributes - NAG averaged over private attributes.