

A photograph of the Washington Monument in Washington, D.C., viewed from the reflecting pool. The monument is the central focus, standing tall against a dramatic sunset sky with orange, yellow, and blue hues. The reflecting pool in the foreground shows a clear reflection of the monument and the sky. Lush green trees line the walkways on either side of the pool.

SECURE AND PRIVATE RECOMMENDATION

Tommaso Di Noia

Online and Adaptive Recommender Systems (OARS)
Workshop @ KDD'22
August 14, 2022

Motivation

SECURITY AND PRIVACY IN RS: WHY WE CARE ABOUT

- Users are at the center of the recommendation task
- Attacking a recommendation engine has a direct consequence on (potentially) all the users of the system
- Users' preferences are very sensitive knowledge

SECURITY AND PRIVACY: WHY WE CARE ABOUT



...

WHAT THEY HAVE IN COMMON

SECURITY: protect users final recommendations against attacks

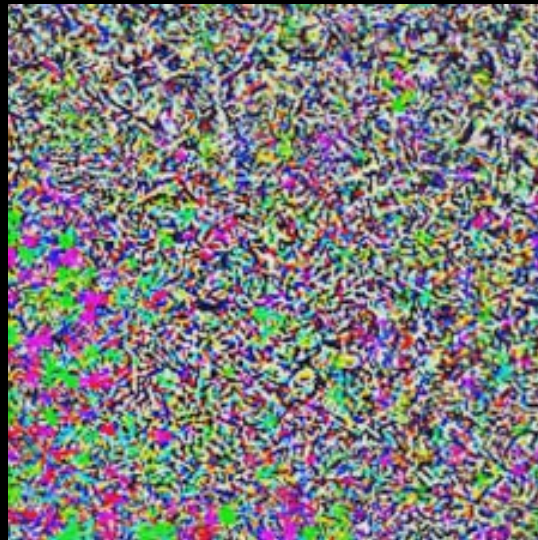
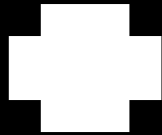
PRIVACY: protect users' data against attacks and improper use

Security

SECURITY: YOU MAY KNOW THE PANDA



"panda"



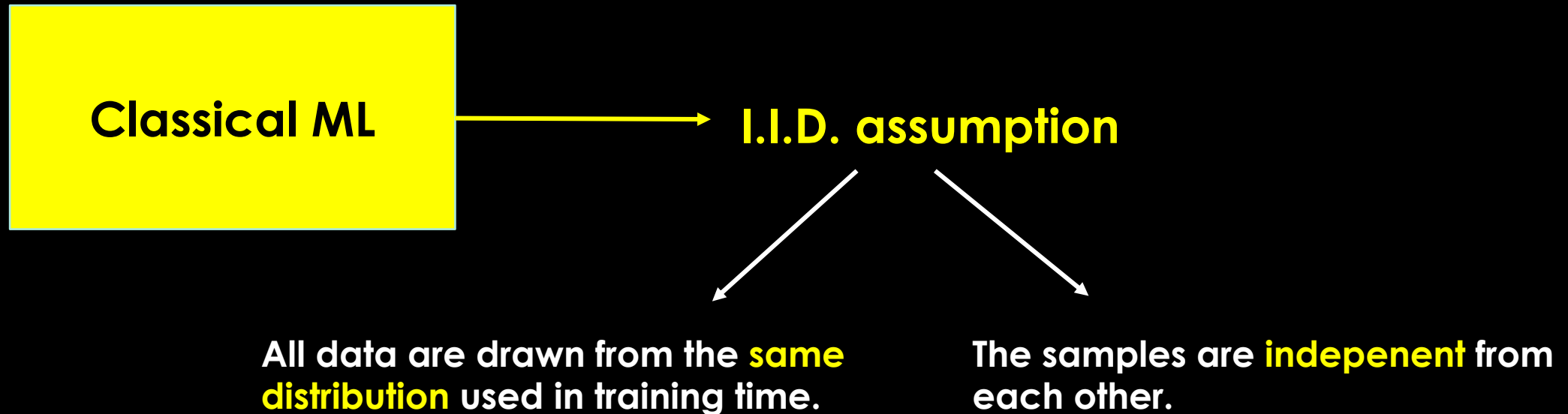
*Adversarial
Noise*



"gibbon"

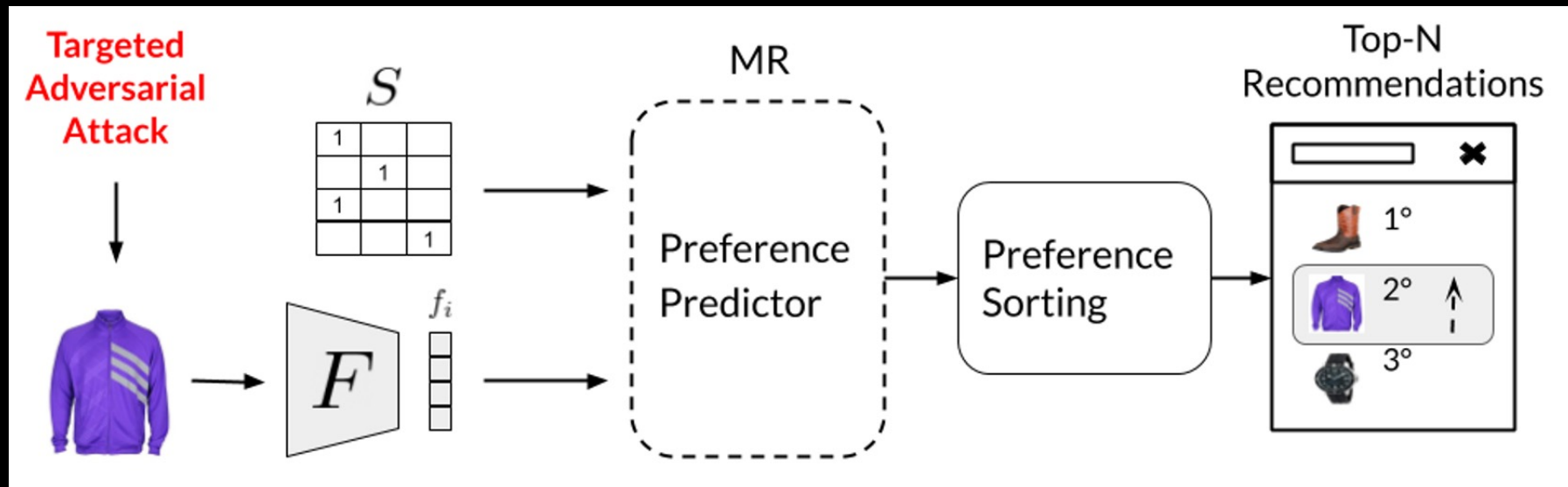


ADVERSARIAL LEARNING BREAKS AN IMPORTANT ASSUMPTION



«Such assumptions [...] rule out the possibility that an adversary could alter the distribution at either training time or test time.»

ADVERSARIAL EXAMPLES IN RS



Simulation of Targeted Adversarial Attacks against Multimedia Recommender Systems can push low recommended product categories even **3 times more recommended** by **perturbing** product images in a **human-imperceptible way**.

ADVERSARIAL PERSPECTIVE

Supervised learning (classification) problem

$$\arg \max_{\Delta_{adv}} J(\Omega, \underbrace{x + \Delta_{adv}}_{\text{Adversarial perturbation of sample } x}, y) \quad s. t. , \underbrace{\|\Delta_{adv}\|_p}_{\text{perturbation budget}} \leq \epsilon$$

Adversarial perturbation of sample x

perturbation budget

Algorithms that aim to find such adversarial perturbations are referred to as **adversarial attacks**.

ADVERSARIAL TRAINING

[GOODFELLOW ET AL., ICLR'15]

Including adversarial samples in the **training** of a model makes it **more robust**.
The objective function of the model **adversarially-trained** is:

$$\arg \min_{\Omega} \max_{\Delta_{adv}} J(\Omega, x, y) + \lambda J(\Omega, x + \Delta_{adv}, y)$$

Adversarial Regularization term

Adversarial training provides better generalization performance

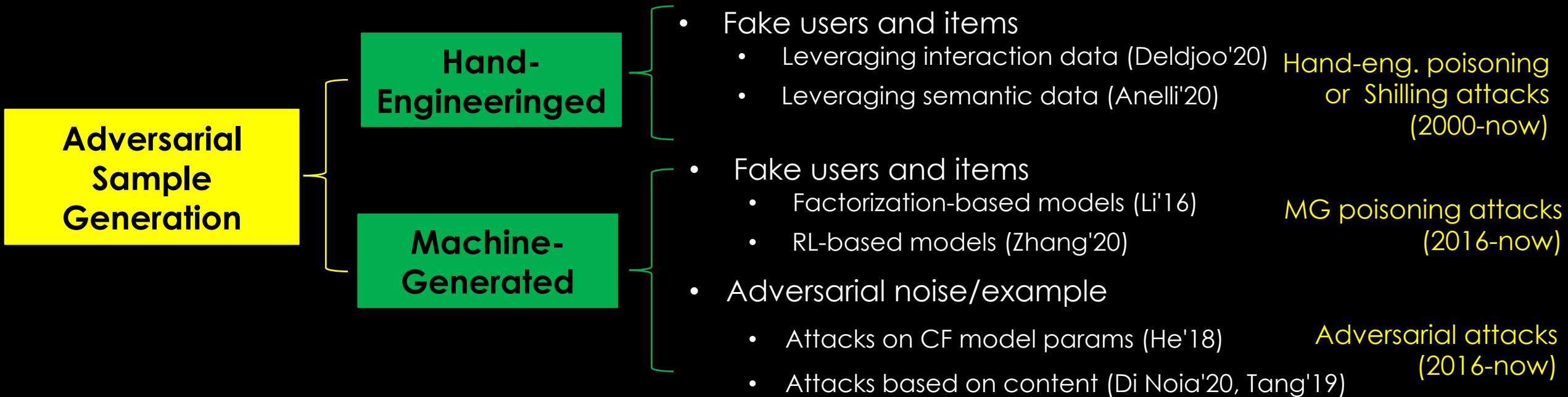
[Miyato et al., ICLR'17]

COUNTERMEASURES

- **Proactive** countermeasures
 - **Adversarial Training** [Goodfellow et al., ICLR '15]
 - Additional training epochs with adversarial examples
 - **Defensive Distillation** [Papernot et al., ISS'16]
 - Adapt distillation to increase the robustness of the network
 - **Robust Optimization** [Madry et al., ICLR'18]
 - design robust DNN to prevent a specific class of adversarial examples
- **Reactive** countermeasures
 - **Adversarial Detecting**
 - **Input Reconstruction**
 - **Network Verification**

Security and RS

ATTACKS AGAINST RECOMMENDER SYSTEMS



HAND-CRAFTED SHILLING ATTACKS

Problem: Given a U-I matrix, the goal is to add a small number of fake users, where each new profile can have maximum 'C' ratings.

Different attack types: Constructed based on the composition a of user profile.
(e.g, random, popular, bandwagon, love-hate)

I_S			I_F			I_θ			I_T
$i_s^{(1)}$...	$i_s^{(\alpha)}$	$i_f^{(1)}$...	$i_f^{(\phi)}$	$i_\theta^{(1)}$...	$i_\theta^{(\chi)}$	i_t

Gunes, I., Kaleli, C., Bilge, A., & Polat, H. (2014). Shilling attacks against recommender systems: a comprehensive survey. *Artificial Intelligence Review*, '14.



5		5		
	5			4
3	4	5		
4			1	
		5		5
		4	4	
2	3			5

fake users

HAND-CRAFTED SHILLING ATTACKS AGAINST RS

Recent advances focuses on:

Goal (attack): Study the Impact of Dataset Characteristics on the efficacy of most popular CF shilling attacks

$$\mathbf{y} = \epsilon + \theta_0 + \theta_d \mathbf{X}_d + \theta_c \mathbf{X}_c$$

$$\mathbf{y} \rightarrow \Delta_{HR@k} = \hat{H}R@k - HR@k$$

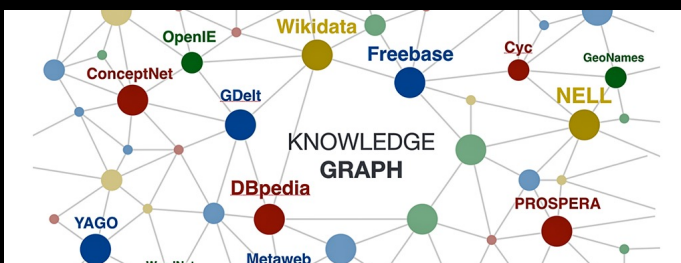
$x \rightarrow$ data characteristics

$$x_1 = \log_{10}\left(\frac{|\mathcal{U}| \cdot |\mathcal{I}|}{sc}\right) \quad x_4 = 1 - 2 \sum_{i=1}^{|\mathcal{I}|} \left(\frac{|\mathcal{I}| + 1 - i}{|\mathcal{I}| + 1}\right) \times \left(\frac{|\mathcal{K}_i|}{|\mathcal{K}|}\right)$$

$$x_2 = \log_{10}\left(\frac{|\mathcal{U}|}{|\mathcal{I}|}\right) \quad x_5 = 1 - 2 \sum_{u=1}^{|\mathcal{U}|} \left(\frac{|\mathcal{U}| + 1 - u}{|\mathcal{U}| + 1}\right) \times \left(\frac{|\mathcal{K}_u|}{|\mathcal{K}|}\right)$$

$$x_3 = \log_{10}\left(\frac{|\mathcal{K}|}{|\mathcal{U}| \times |\mathcal{I}|}\right) \quad x_6 = \sqrt{\frac{\sum_{i=1}^{|\mathcal{K}|} (r_i - \bar{r})^2}{|\mathcal{K}| - 1}}$$

KNOWLEDGE-AWARE SHILLING ATTACK



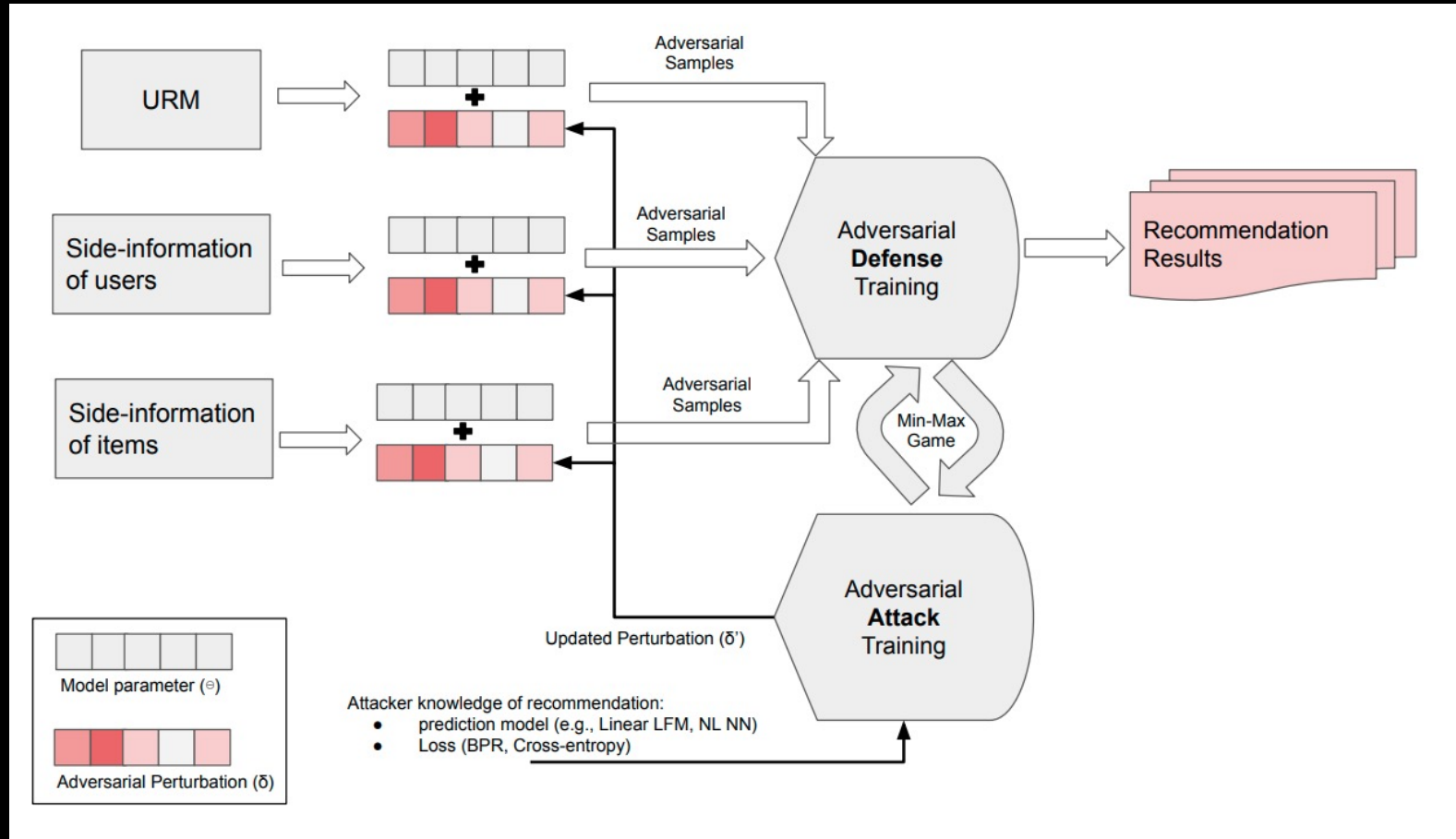
Metric:	LibraryThing									Yahoo!Movies									
	User- k NN			Item- k NN			MF			User- k NN			Item- k NN			MF			
	1%	2.5%	5%	1%	2.5%	5%	1%	2.5%	5%	1%	2.5%	5%	1%	2.5%	5%	1%	2.5%	5%	
Rnd	baseline	.074	.157	.230	.281	.457	.557	.767	.900	.942	.189	.366	.449	.329	.508	.598	.410	.580	.702
	CS-1H	.068*	.143*	.213*	.271*	.441*	.558	.778*	.898	.940	.202	.372	.455*	.336	.522*	.609*	.430*	.607*	.707
	OS-1H	.081*	.170*	.250*	.290*	.467*	.576*	.786*	.902	.944	.217*	.394*	.477*	.345*	.535*	.622*	.446*	.635*	.742*
	FS-1H	.072	.154	.229	.280	.455	.570*	.786*	.901	.942	.213*	.381*	.468*	.338*	.530*	.619*	.442*	.623*	.728*
L-H	baseline	.502	.518	.518	.874	.952	.978	.955	.987	.995	.604	.608	.605	.888	.930	.958	.956	.967	.980
	CS-1H	.502	.518	.518	.876*	.953	.979	.957	.987	.994	.604	.608	.605*	.889	.932	.957	.956	.967	.979
	OS-1H	.502	.518	.518	.870*	.950*	.974*	.955*	.986	.994	.604	.605	.605	.887	.933	.955*	.956	.967	.979
	FS-1H	.502	.518	.518	.874	.951	.977	.955	.987	.993	.604*	.608	.605	.888	.933	.956	.956	.967	.979
Avg	baseline	.086	.197	.285	.313	.508	.605	.803	.915	.951	.233	.416	.494	.374	.574	.654	.489	.685	.788
	CS-1H	.081*	.187*	.269*	.301*	.507	.621*	.814*	.915	.950	.220*	.399*	.479*	.357*	.554*	.639*	.467*	.652*	.744*
	OS-1H	.093*	.202	.289	.313	.507	.610*	.810	.911	.948	.237	.412	.494	.371	.563*	.646*	.475	.656*	.754*
	FS-1H	.084	.190*	.272*	.305*	.504	.614*	.811	.911	.946*	.215*	.397*	.473*	.350*	.547*	.634*	.448*	.627*	.729*

ADVERSARIAL RS CHALLENGES

1. Unlike **images** composed of **continuous features**, the input to RS are discrete (rating, (u,i,j) in BPR)
2. Adversarial examples on images aim to be **UNNOTICEABLE**.

Where can we add adversarial noise?

ADVERSARIAL RS



ADVERSARIAL NOISE

Adding adversarial noise on CF model paramters:

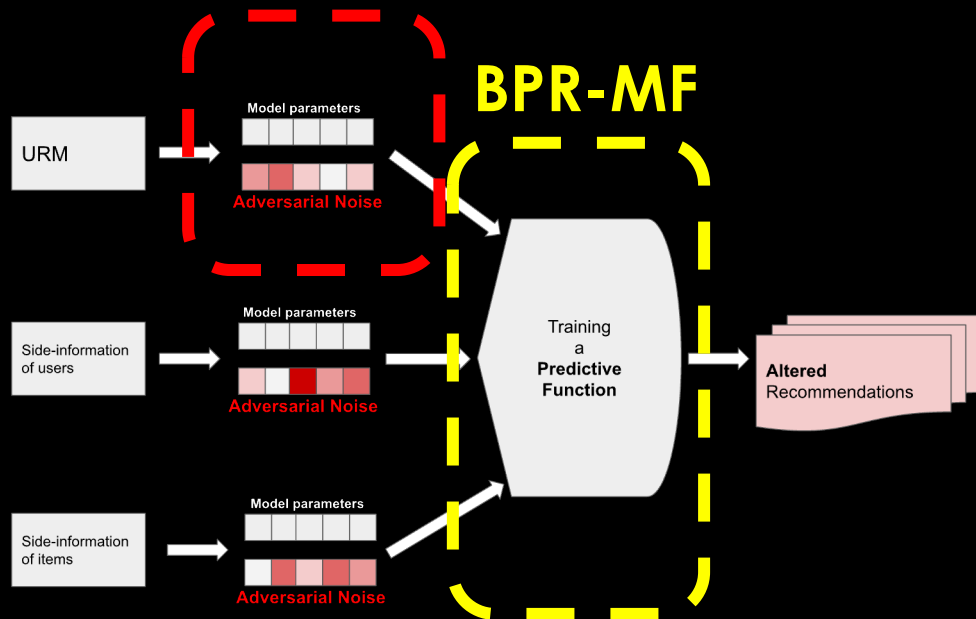
- Adds adversarial noise to the model paramters of **BPR-MF**
- Compares **adversarial v.s. random noise**
- Applies **adversarial training** as a defense mechnasim

ADVERSARIAL PERSONALIZED RANKING

Adversarial Perturbation on each **embedding** vector of user and item



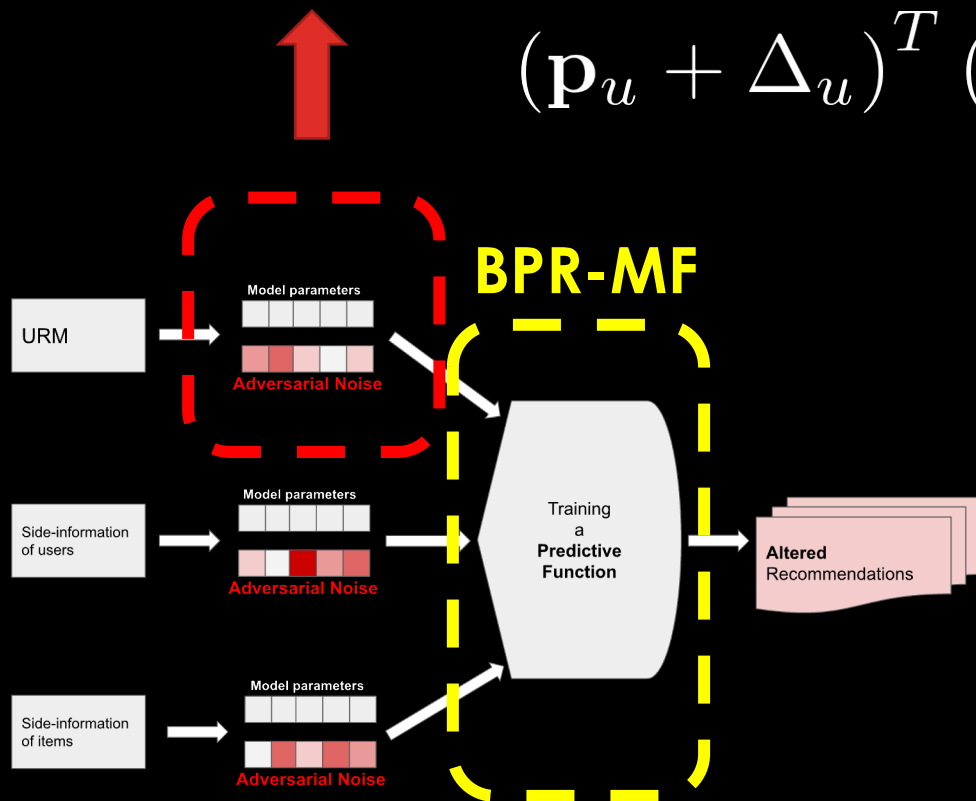
$$(\mathbf{p}_u + \Delta_u)^T (\mathbf{q}_i + \Delta_i)$$



ADVERSARIAL PERSONALIZED RANKING

Adversarial Perturbation on each **embedding** vector of user and item

$$(\mathbf{p}_u + \Delta_u)^T (\mathbf{q}_i + \Delta_i)$$



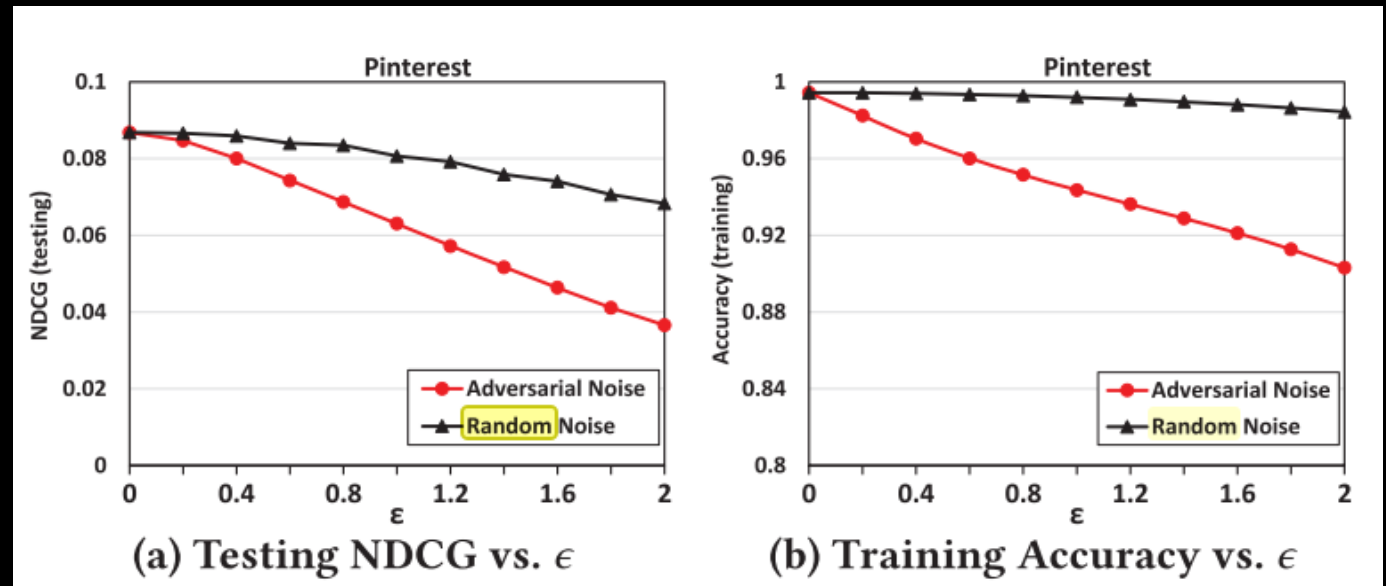
The impact of applying adversarial perturbation

reduction of NDCG@100			
	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$
Dataset	BPR-MF	BPR-MF	BPR-MF
Yelp	-22.1%	-42.7%	-63.8%
Pinterest	-9.5%	-25.1%	-55.7%
Gowalla	-26.3%	-53.0%	-78.0%

ADVERSARIAL PERSONALIZED RANKING

The impact of adversarial v.s. random noise on BPR-MF:

- **adversarial perturbations:** NDCG decreases -21.2%
- **random perturbations:** NDCG decreases -1.6%



13 times
difference!

DEFENSE AGAINST ADVERSARIAL SAMPLES

- **Goal:** Build ML models that can make robust prediction even in presence of adversarial examples.
- Main defensive approaches:
 - (i) increasing robustness,
 - Robust optimization
 - Adversarial training (regularization)
 - Robust gradient descent
 - Certified robustness
 - Defense distillation
 - (ii) detection

Most Popular in RecSys

ADVERSARIAL PERSONALIZED RANKING

[XIANGNAN HE ET AL., SIGIR '18]

Do **Adversarial training** improve the **robustness**?

	NDCG@100					
	$\epsilon = 0.5$		$\epsilon = 1$		$\epsilon = 2$	
Dataset	BPR-MF	APR	BPR-MF	APR	BPR-MF	APR
Yelp	-22.1%	-4.7%	-42.7%	-12.5%	-63.8%	-31.0%
Pinterest	-9.5%	-2.6%	-25.1%	-7.2%	-55.7%	-23.4%
Gowalla	-26.3%	-2.9%	-53.0%	-13.2%	-78.0%	-29.2%

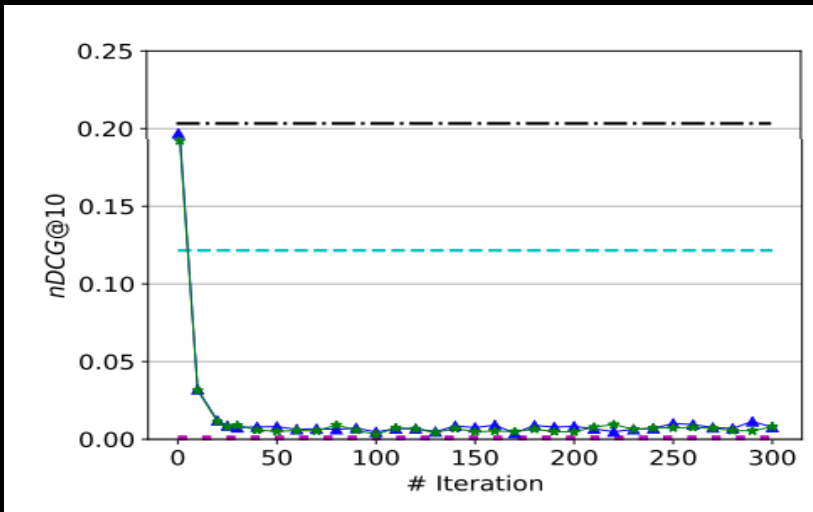
ITERATIVE ADVERSARIAL NOISE

Adding **iterative** adversarial noise on CF model paramnters:

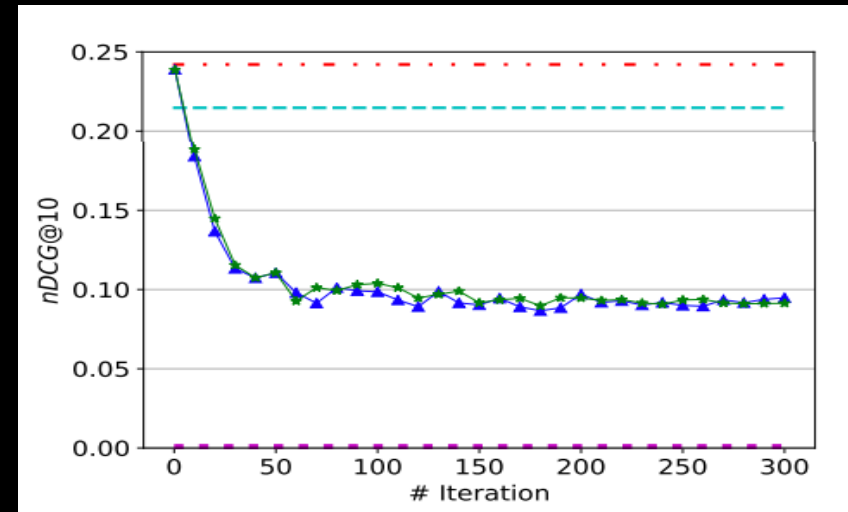
$$\Theta_0^{adv} = \Theta + \Delta_0 \quad \Theta_1^{adv} = Clip_{\Theta, \epsilon} \left\{ \Theta_0^{adv} + \alpha \frac{\Pi}{\|\Pi\|} \right\} \text{ where } \Pi = \frac{\partial \mathcal{L}(\Theta + \Delta_0)}{\partial \Delta_0}$$

- Iterative Perturbation can make the recommendation model worse than a random model
- The APR defense strategy limitates but does not protect from MSAP

BPR-MF



APR



MULTIMEDIA RS: ATTACK TIMMIG

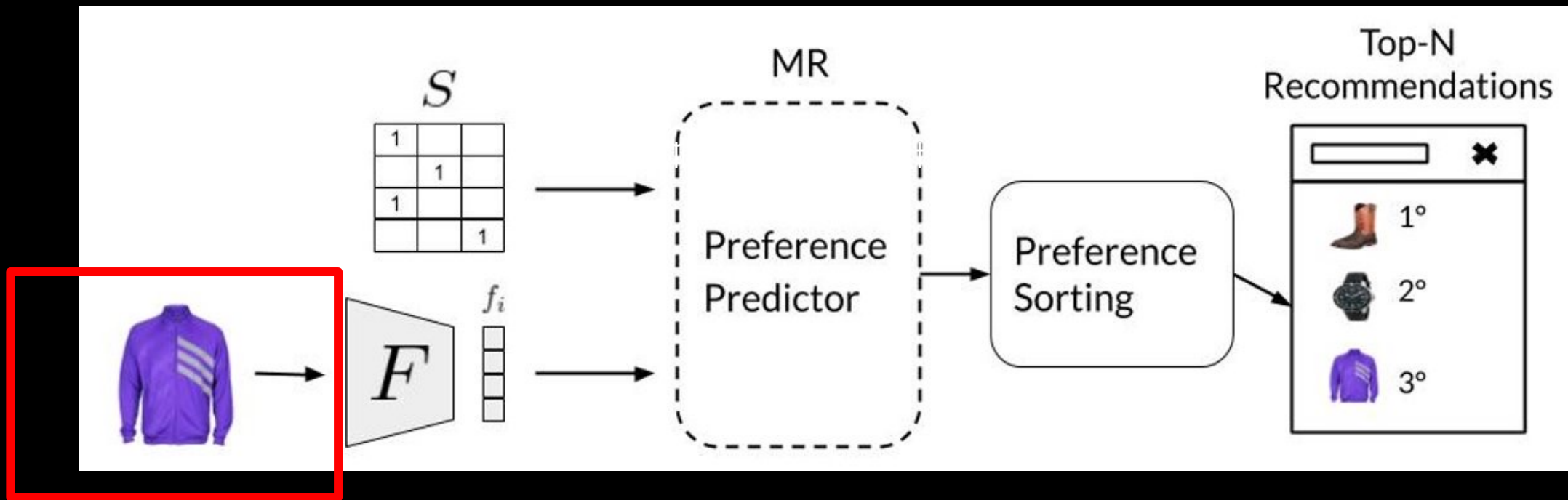
TRAINING TIME (Poisoning)

- Image samples are perturbed and injected in the VRSs before the training.
- **WORKS**
 - TAaMR [Di Noia et al, 2020]
 - VAR [Anelli et al, 2021]

TESTING TIME (Evasion)

- Images are perturbed at **inference** time
- **WORKS**
 - BlackBox-Model [Cohen et al, 2021]
 - Adv. Item Promotion [Zhouran et al, 2021]

ADVERSARIAL ATTACKS AGAINST VISUAL-AWARE RS



THE ADVERSARY CAN PERTURB THE PRODUCT IMAGES

ADVERSARIAL ATTACKS AGAINST VISUAL-AWARE RS



(a) original (sock)
probability: 60%
rec. position: 180th



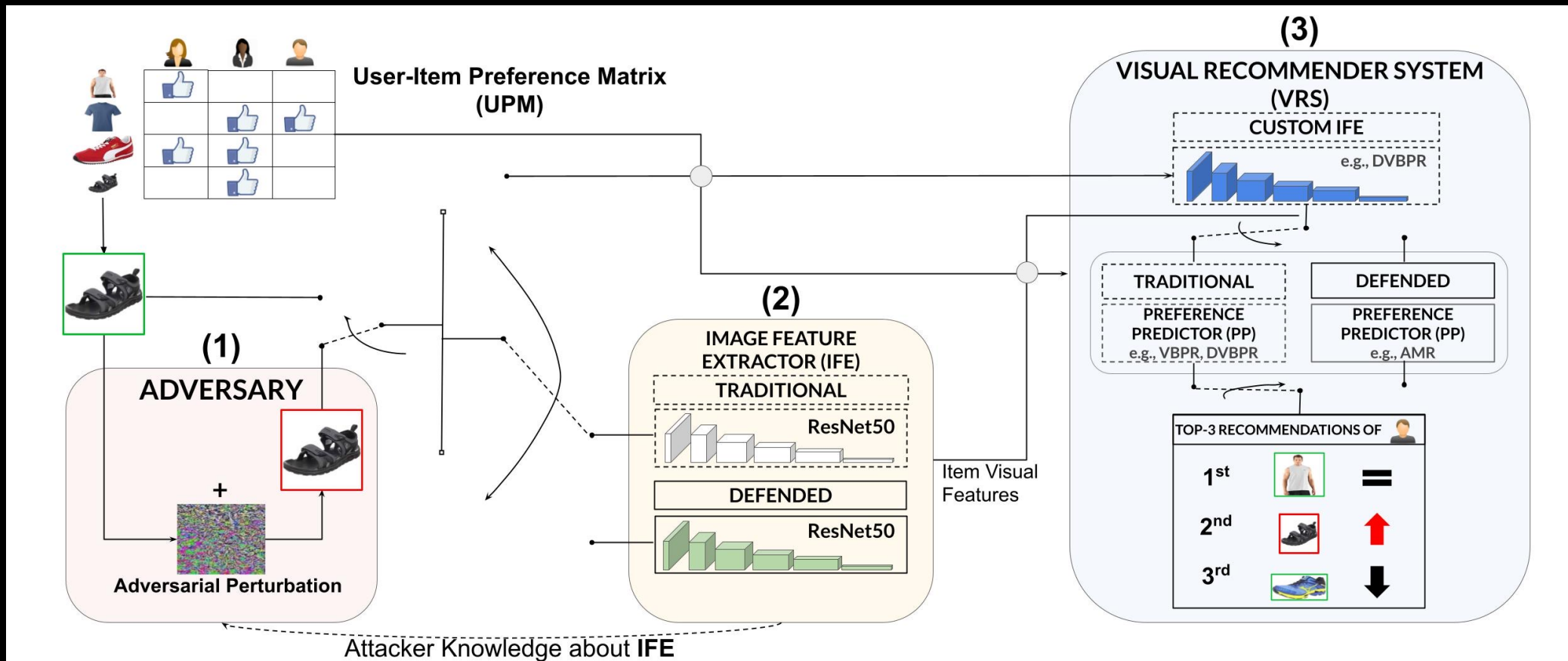
(b) attacked (running shoe)
probability: 100%
rec. position: 14th

Attacks success probability.

Dataset	Origin→Target	Attack	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 8$	$\epsilon = 16$
Amazon Men	Sock→Running Shoes	FSGM	9.32%	17.02%	22.14%	21.68%
		PGD	68.69%	98.37%	99.92%	99.84%
	Sock→Analog Clock	FSGM	0.16%	0.31%	0.39%	0.23%
		PGD	30.77%	87.10%	99.46%	100.00%
	Sock→Jersey, T-shirt	FSGM	8.24%	17.17%	26.50%	15.54%
		PGD	67.29%	98.83%	100.00%	100.00%
Amazon Women	Maillot→Brassiere	FSGM	45.51%	51.48%	52.30%	56.46%
		PGD	85.32%	99.40%	99.95%	100.00%
	Maillot→Chain	FSGM	0.38%	1.31%	1.92%	2.68%
		PGD	17.20%	90.53%	99.95%	99.95%

TRAINING TIME ATTACK

VISUAL ADVERSARIAL RECOMMENDATION FRAMEWORK



TRAINING TIME ATTACK

VISUAL ADVERSARIAL RECOMMENDATION FRAMEWORK

- **Adversarial Attacks**

- FGSM
- PGD
- Carlini&Wagner

WHITE BOX wrt the IFE

BLACK BOX wrt the Recommender

- **Adversarial Defense**

- Adversarial Training of the IFE
- Free Adversarial Training of the IFE

TRAINING TIME ATTACK

VISUAL ADVERSARIAL RECOMMENDATION FRAMEWORK

Data	VRS	Att.	Image Feature Extractor					
			Traditional		Adv. Train.		Free Adv. Train.	
			SR	FL	SR	FL	SR	FL
Amazon Men	FM, VBPR, AMR	FGSM	65%	14.0948	18%	0.0330	15%	0.0278
		PGD	97%	36.8843	18%	0.0334	15%	0.0283
		C&W	89%	20.5172	48%	2.8022	42%	1.9080
	ACF	FGSM	65%	9.0480	18%	0.0944	15%	0.0951
		PGD	97%	9.2606	18%	0.0944	15%	0.0954
		C&W	89%	10.4917	48%	0.7582	42%	0.4955
	DVBPR	FGSM	65%	16.4055	—	—	—	—
		PGD	97%	16.1151	—	—	—	—
		C&W	89%	16.3442	—	—	—	—

TRAINING TIME ATTACK HUMAN IMPERCEPTIBILITY



a. Clean
 Rec. Position: 68th



b. Attack + T
 Rec. Position: 10th
 LPIPS: 0.5484



c. Attack + AT
 Rec. Position: 27th
 LPIPS: 0.5347



d. Attack + FAT
 Rec. Position: 40th
 LPIPS: 0.3447

Privacy in RS

THE PRIVACY-PERSONALIZATION TRADE-OFF IN RS

- The quality of the recommendations is correlated with the amount, richness, and freshness of the underlying user modeling data
- The same factors drive the severity of the privacy risk

PRIVACY RISKS IN RS

- **Direct access to data**
 - Unsolicited data collection
 - Sharing data with third parties
 - Unsolicited access by employees
- **Inference from User Preference Data**
 - Exposure of sensitive information
 - Targeted Advertising
 - Discrimination
- **Risks Imposed by other System Users**
 - In collaborative approaches, users are compared with each other
 - Create fake profiles to identify other users' preferences
 - By observing changes in item-to-item collaborative systems an attacker may infer the preferences of a target user

Privacy-preserving Machine Learning for RS

WHAT PRIVACY-PRESERVING MACHINE LEARNING TRIES TO PROTECT

- Input training data;
- Output predicted labels;
- Model information, including parameters, architecture, and loss function;
- Identifiable information, such as which site a record comes from.

ATTACK AND THREAT MODELS

1

Targets

Data vs. Model

2

Knowledge

White-box vs. Black-box

3

Methods

Model extraction vs. Encoding Information

THE POINT WITH PRIVACY

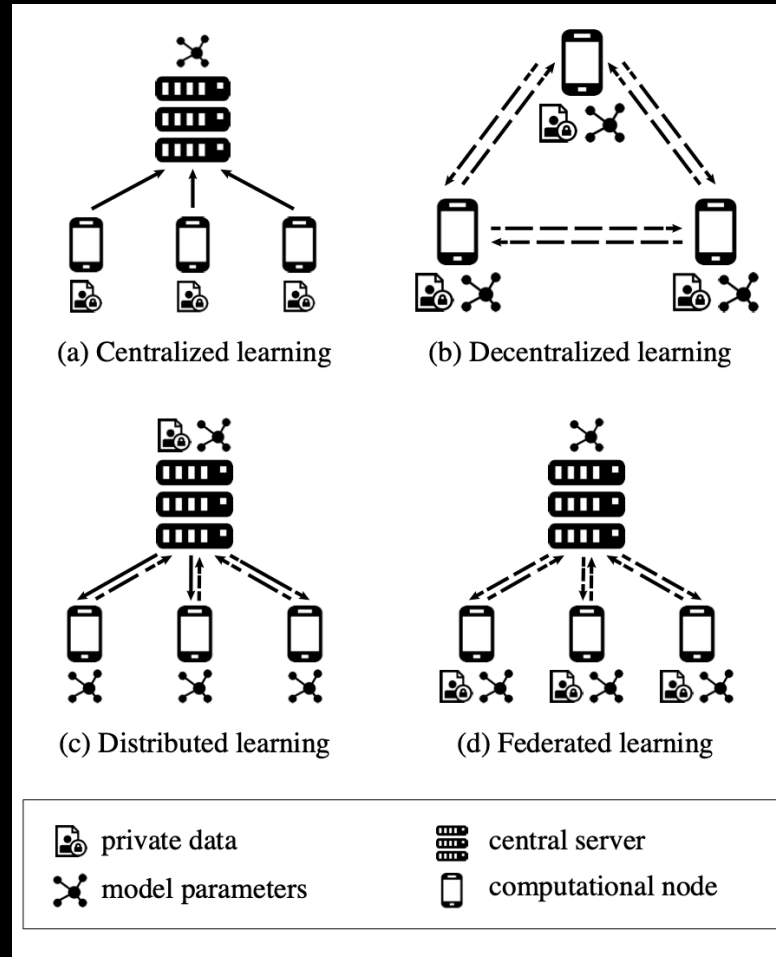
We want to learn nothing about individuals but still learn useful information about a population.

De-identified data are not so secure

Releasing just statistics is still non-private

LEARNING PARADIGMS

- Learning paradigms
 - Centralized
 - Decentralized
 - Distributed
 - Federated

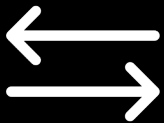


FEDERATED LEARNING: ADVANTAGES



Data
privacy/security

Data pool not required for the model. Data don't leave user's devices



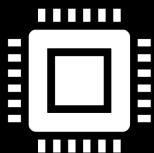
Data diversity
and Model
Liability

FL facilitates access to heterogeneous data. Reduces legal liability of the model



Real time
continuous
learning

Model are constantly improved using client data with no need to aggregate data for continuous learning



Hardware /
Bandwidth
efficiency

FL models do not need complex central server to analyze data/Do not require uploading large amount of data

DIFFERENTIAL PRIVACY

\mathcal{X} and \mathcal{Y} are adjacent datasets (\mathcal{Y} is equal to \mathcal{X} but for one more example)

\mathcal{M} is a randomized mechanism over a dataset

\mathcal{M} gives ε -differential privacy if for all pairs of datasets \mathcal{X} and \mathcal{Y} and all events S we have:

$$\Pr[\mathcal{M}(\mathcal{X}) \in S] \leq e^\varepsilon \Pr[\mathcal{M}(\mathcal{Y}) \in S]$$

If $\varepsilon = 0$, we have no probability loss, and an attacker cannot distinguish the two datasets

With current and future side information and with postprocessing, the probability ratio should still hold

(ALMOST) DIFFERENTIAL PRIVACY

$$P(\mathcal{M}(x) \in S) \leq e^\epsilon P(\mathcal{M}(y) \in S) + \delta$$

DIFFERENTIAL PRIVACY IN SHORT

- Strong privacy guarantees
- No longer needed attack modeling
- Quantifiable privacy loss
- Composable mechanisms
- Useful for analyzing any algorithm

SECURE MULTI-PARTY COMPUTATION

Additive Secret Sharing

We can split a secret into N shares and keep it hidden as long as at most $N-1$ shareholders collaborate.

We can sum shares of different secrets between them or sum and multiply any non-encrypted number (homomorphic addition)

HOMOMORPHIC ENCRYPTION

- It is a cryptographical scheme allowing certain mathematical operations to be performed directly in ciphertexts without prior decryption.
 - **Partially homomorphic encryption**: can reach additive homomorphism or multiplicative homomorphism;
 - **Somewhat homomorphic encryption**: operations can be applied for a limited number of times, since noise is used;
 - **Fully homomorphic encryption**: allows unlimited number of additions and multiplications over ciphertexts

WHICH TECHNIQUE?

- HE and SMPC are often replaceable
 - HE: little interaction and expensive computation
 - SMPC: Cheap computation and significant amount of interaction
- SMPC replaces computation with interaction, offering better practical performance
- DP replaces accuracy with efficiency. If the coordinator is trusted, send plain data to preserve more accuracy

Closing Remarks

SECURITY: OPEN DIRECTIONS IN RS

- New attacks strategies
 - Use state-of-the-art adv. Attack strategies
 - Implement perturbation direct on the input:
 - user-rating profile
 - Imitation of implicit feedback
 - images, audio, videos
- New defence approaches
- Verify and Extend the AVD-RF on other recommenders
- New domains

SECURITY AND PRIVACY: OPEN DIRECTIONS IN RS

- Both related to attacking and defending the user
- What's the effect of combining privacy-preserving ML with adversarial ML for recommender systems?
 - Accuracy
 - Diversity
 - Novelty
 - Fairness

MANY THANKS TO THE RECSYS CREW (AND ALUMNI) @ POLITECNICO DI BARI



Vito Walter Anelli
Assistant Professor
@ Politecnico di Bari



Vito Bellini
Applied Scientist II
@ Amazon



Giandomenico Cornacchia
Ph.D. student @
Politecnico di Bari



Yashar Deldjoo
Assistant Professor
@ Politecnico di Bari



Antonio Ferrara
Assistant Professor
@ Politecnico di Bari



Daniele Malitesta
Ph.D. student @
Politecnico di Bari



Alberto Mancino
Ph.D. student @
Politecnico di Bari



Roberto Mirizzi
Vice President of ML,
RecSys, Search
@ Discovery Inc



Felice Merra
Applied Scientist II
@ Amazon



Fedelucio Narducci
Associate Professor
@ Politecnico di Bari



Vito Claudio Ostuni
Senior Research
Scientist @ Netflix



Vincenzo Paparella
Ph.D. student @
Politecnico di Bari



Claudio Pomo
Ph.D. student @
Politecnico di Bari



Azzurra Ragone
Assistant Professor
@ University of Bari
Aldo Moro

A photograph of the Washington Monument in Washington, D.C., taken at sunset. The monument is the central focus, standing tall against a sky with soft orange and pink clouds. The reflecting pool in the foreground is perfectly still, creating a clear reflection of the monument and the sky. Lush green trees line both sides of the pool, and the city skyline is visible in the distance. The overall mood is serene and majestic.

THANK YOU + Q&A

Tommaso Di Noia. SECURE AND PRIVATE RECOMMENDATION. OARS Workshop @ KDD 2022