Towards purpose-aware privacy-preserving techniques for predictive Applications

Manel Slokom

18-06-2024

TU Delft

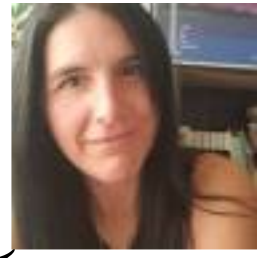
CWI
Centrum Wiskunde & Informatica

Statistics Netherlands

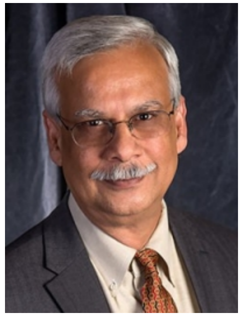
Committee members

Gillian Raab
Professor of Applied Statistics at Edinburgh Napier University
Research area: *Privacy, Synthetic data*.

Alessandro Bozzon
Professor of Computer Science at TU Delft
Research area: *Human-centred AI, recommender systems, crowdsourcing*.



Sole Pera
Associate Professor of Computer Science at TU Delft
Research area: *Information retrieval, recommender systems*.



Krish Muralidhar
Professor of Marketing and Supply Chain.
Research area: *Privacy, Synthetic data, (re-identification) attacks*

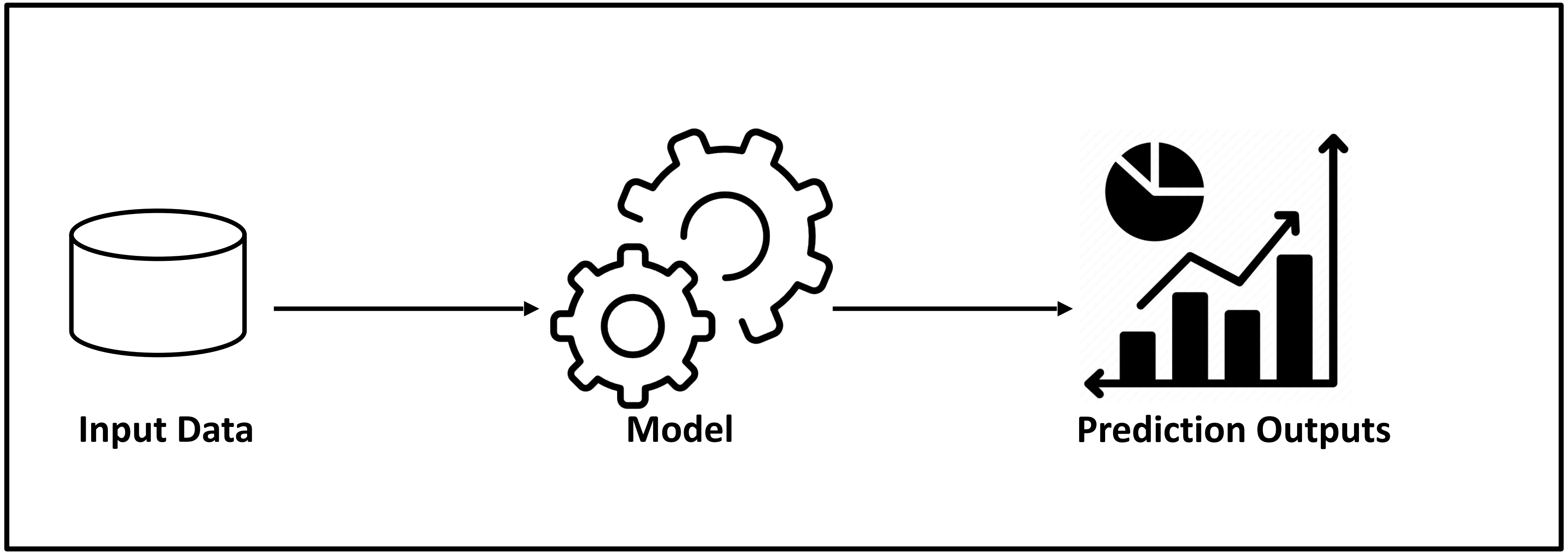
Mykola Pechenizkiy
Professor of Computer Science at TU Eindhoven
Research area: *Responsible AI/ Analytics, reinforcement learning*.

Pablo Cesar (reserved member)
Professor of Computer Science at TU Delft
Research area: *Human-centred multimedia*.

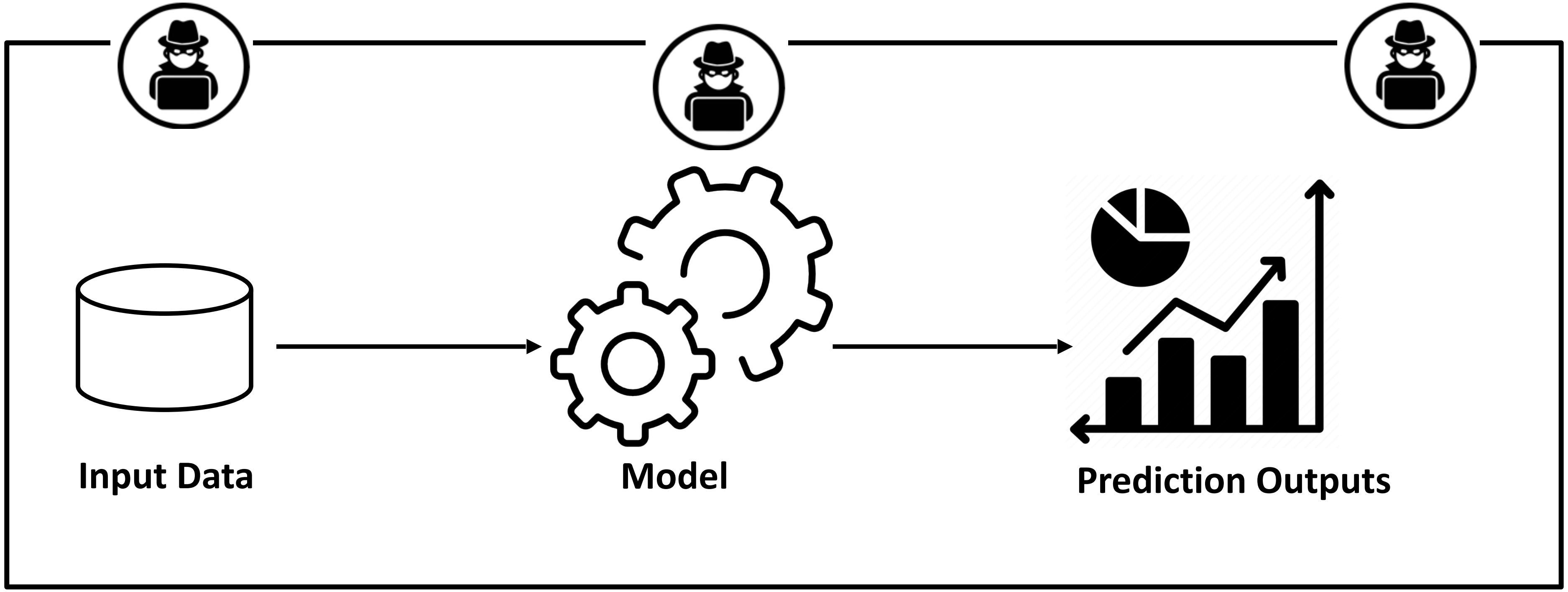
TU Delft

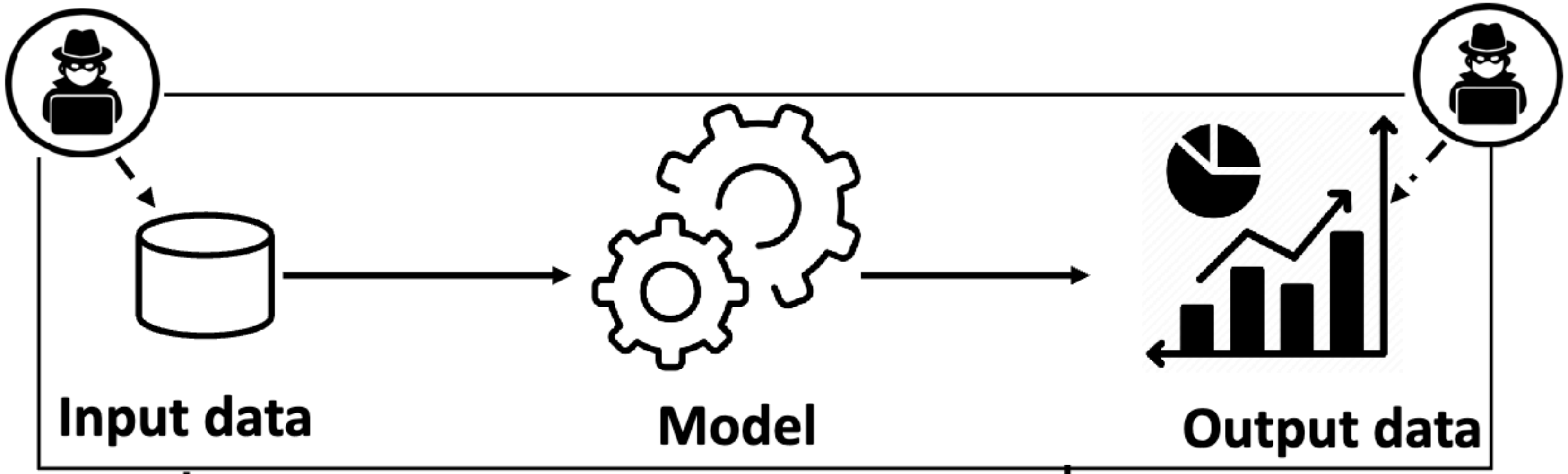
Towards **purpose-aware** privacy-preserving techniques for **predictive applications**

Introduction ~ Context



Introduction ~ Problems





Chapter 1: Introduction

How can we protect sensitive information when the attacker has access to **the input data** from inference attack while maintaining utility? (**RQ1**)

Chapter 2 : *PerBlur: A personalization-based approach to gender obfuscation for user profiles.*

Chapter 3: *Data Masking for recommender Systems: Prediction performance and rating hiding.*

How can we protect sensitive information when the attacker has access to **the model's predictions** while maintaining utility? (**RQ2**)

Chapter 4: *When Machine Learning Models Leak: An exploration of Synthetic Training Data.*

Chapter 5: *Exploring Privacy-Preserving Techniques on Synthetic Data as a Defense against Model Inversion Attacks.*

Do recommender system's **output data** leaks sensitive information about users? (**RQ3**)

Chapter 6: *A Closer Look at User Attributes in Recommendations: Implications for Privacy and Diversity.*

Chapter 7: Outlook

TU Delft

Outline

- Introduction
  - Context, problem, research questions
  - Threat model formulation
- Part 1: Attacking input data
- Part 2: Attacking model and output data
- Outlook and discussion

Threat model

- Threat model describes the adversary by looking at the ***resources*** at the adversary's disposal and the adversary's ***objective*** *[Salter, C., et (1998)]*.
  - What the attacker is capable of.
  - What the attacker goal is.

- The **vulnerability**, including the opportunity that makes an attack possible.

- The **countermeasures** that can be taken to prevent the attack.

Salter, C., Saydjari, O.S., Schneier, B., Wallner, J.: Toward a secure system engineering methodology. In: Proceedings of the 1998 Workshop on New Security Paradigms. pp. 2-10. NSPW (1998)

Part 1:

Chapter 2: **PerBlur**: Towards **User-Oriented Privacy** for Recommender System Data: A **Personalization-based** Approach to Gender **Obfuscation** for User Profiles

__Manel Slokom__, , Alan Hanjalic, and Martha Larson. Towards User-Oriented Privacy for Recommender System Data: A Personalization-based Approach to Gender Obfuscation for User Profiles. Information Processing & Management, 2021, vol. 58, no 6, p. 102722.

Threat Model

Threat model: Gender inference on user-item data used for recommender systems

Component	Description
Adversary: Resources	The attacker has a gender classifier pre-trained on unobfuscated data or has the data necessary to train one.
Adversary: Objective	The inference of users' gender attribute.
Vulnerability: Opportunity	The possession of a user-item matrix.
Vulnerability: Countermeasure	Obfuscation of the user-item matrix to block the inference of gender.

TUDelft

Obfuscation for Recommendation

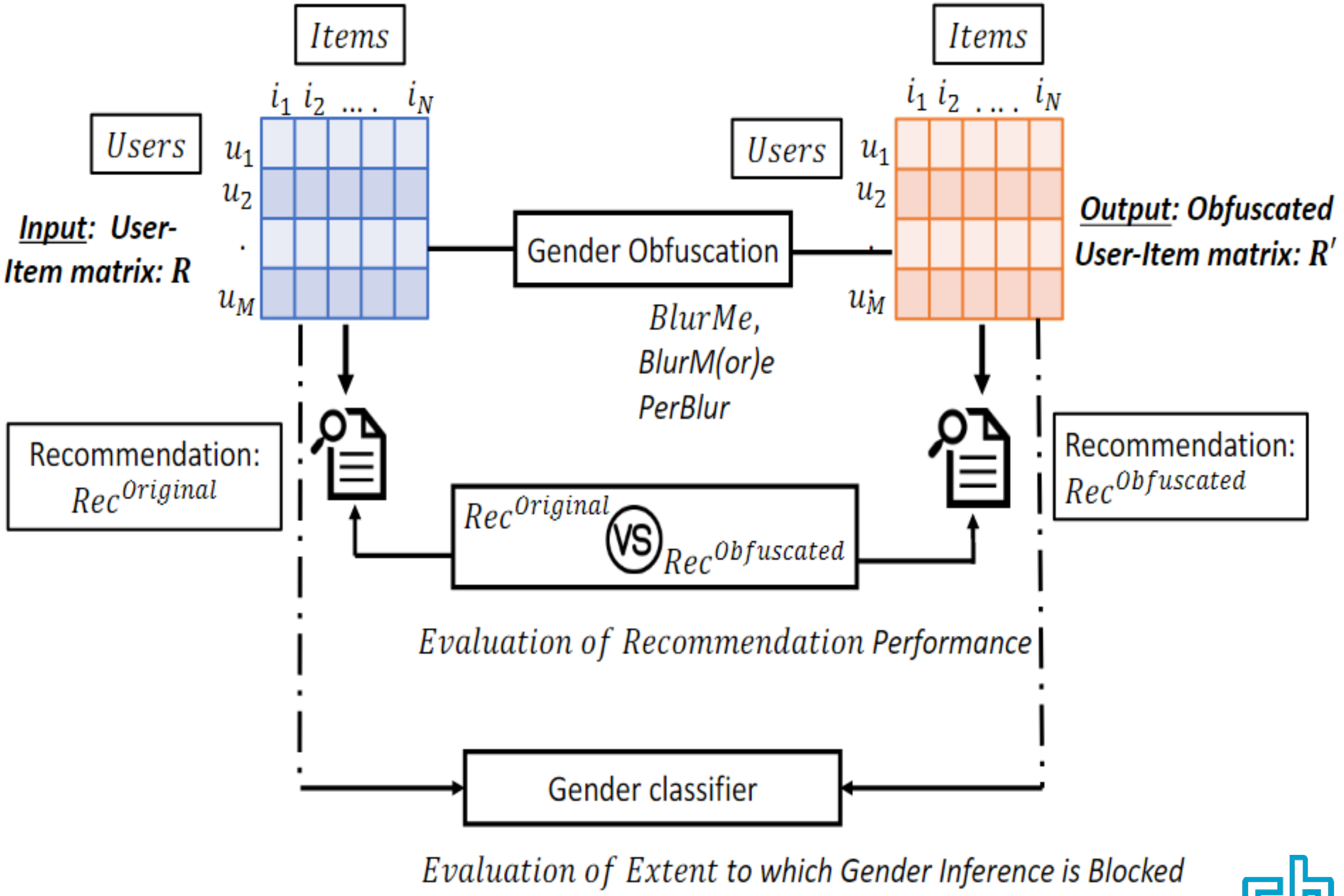
Items

Users	i_1	i_2	i_3	i_4	i_5	i_6	i_7
u_1	5	0	5	0	3	0	0
u_2	4	0	3	0	5	0	1
u_3	2	5	0	4	0	0	3
u_4	5	0	4	0	0	4	0
u_5	0	0	1	4	3	0	2

Data Obfuscation for Recommendation

Data Obfuscation:

- *Hide* implicit sensitive information by *modifying* the data.
- BlurMe (Weinsberg et al., 2012)
- BlurM(or)e (Strucks et al., 2019)



Udi **Weinsberg**, Smriti Bhagat, Stratis Ioannidis and Nina Taft (2012) BlurMe: Inferring and obfuscating user gender based on ratings. In: Proceedings of the 2012 ACM Conference on Recommender Systems. pp. 195–202.
Christopher **Strucks**, Manel Slokom, and Martha Larson (2019) BlurM(or)e: Revisiting gender obfuscation in the user-item matrix. Recommendation in Multistakeholder Environments in conjunction with 13th ACM RecSys.

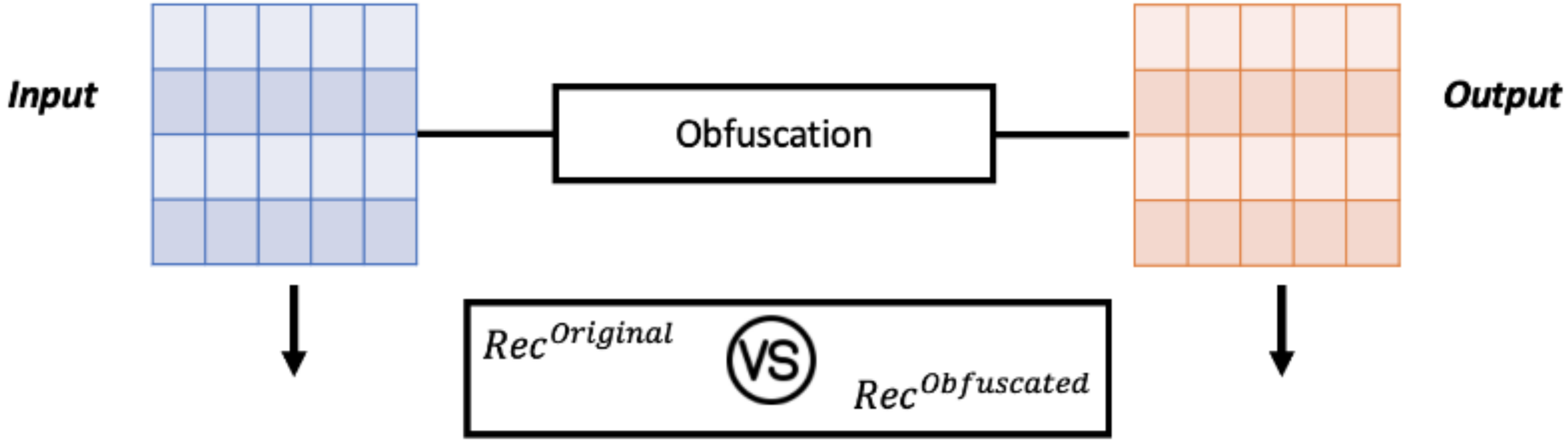
PerBlur – Personalized Blurring

- **PerBlur** creates the *personalized* lists of indicative items by intersecting:
  - Two lists of indicative items: Lm and Lf
  - A ***personalized*** list of items ranked in order of the probability that the user will have rated them.

- **Standard PerBlur**
  - Obfuscation by adding extra items from the personalized lists of indicative items
  - Level of obfuscation: Adds (p%) fake items from the **opposite** gender

- **PerBlur with removal**
  - Similar to Standard PerBlur but we also **remove** certain items.

Data obfuscation for recommendation



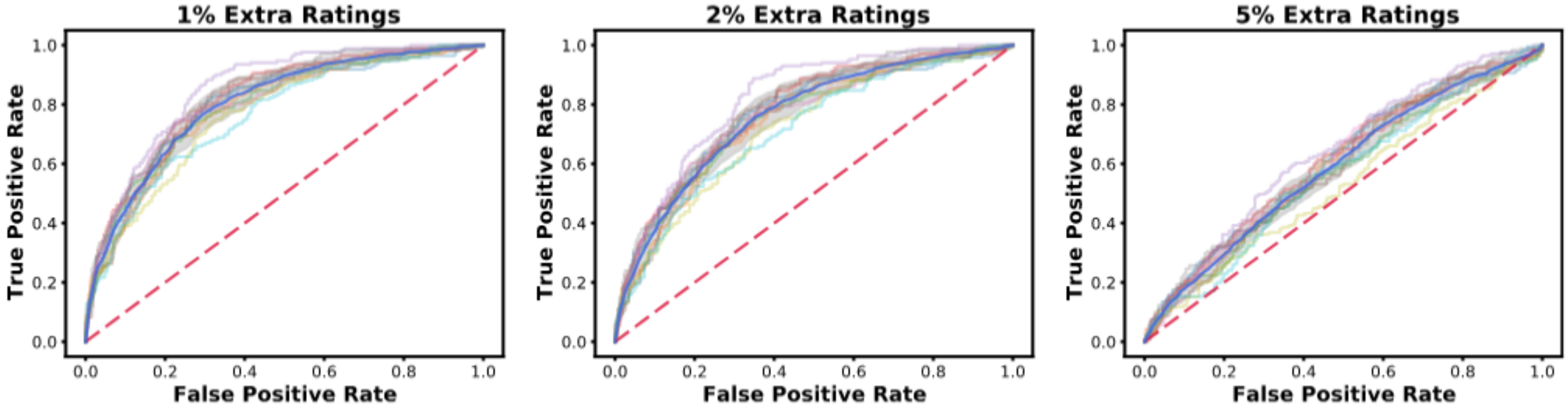
Recommendation Performance

	nDCG	HR@10
Original	0.1634	0.1712
PerBlur	0.1637	0.1704

In the table: we used ML1M data set. PerBlur is created with addition from the personalized lists of indicative items. Logistic regression classifier.

TUDelft

Results: Gender inference



- Obfuscation inhibits the inference of the gender
- PerBlur requires less obfuscation
- Transferability

TU Delft

Achieving diverse recommendation

- The proportion of correctly recommended items that are stereotypical for gender
- Three different cutoff levels (10, 20, 50)

	Obfuscation Strategy		**Gender-steretypical items**					
	Personalization	*Removal*	*top10F*	*top10M*	*top20F*	*top20M*	*top50F*	*top50M*
Original	*None*	*None*	0.0020	0.0045	0.0038	0.0069	0.0082	0.0128
PerBlur	*Personalized*	*Greedy*	**0.0003**	**0.0005**	**0.0014**	**0.0020**	**0.0051**	**0.0073**

- PerBlur is effective in lowering the proportion of TopN gender-steretypical items

TU Delft

In the table: we used ML1M data set. PerBlur is created with addition from the personalized lists of indicative items and removal from Lm or Lf

Outlook Part 1: Attacking input data

- A simple, yet effective **personalized-based** approach to gender **obfuscation** for user profiles
- A recommender system trained on the obfuscated data is able to reach performance **comparable** to what is attained when trained on the original data
- A classifier can **no longer** reliably predict the gender of users
- The ability to recommend **diverse** items.

Part 2:

Chapter 4 & 5: When Machine Learning Models **Leak**: An Exploration of **Synthetic** Training Data

__Manel Slokom__, Peter-Paul de Wolf, Martha Larson. When Machine Learning Models Leak: An exploration of Synthetic Training Data. Privacy in Statistical Databases 2022.
__Manel Slokom__, Peter-Paul de Wolf, Martha Larson. Exploring Privacy-Preserving Synthetic Data as a Defense Against Model Inversion Attacks. Information Security Conference 2023.

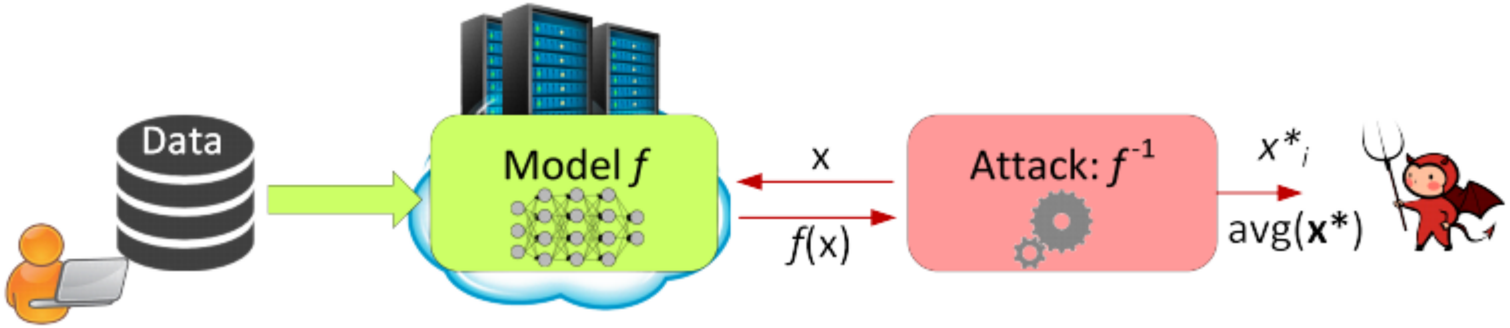
Threat model

Component	Description
Adversary: Objective	Specific sensitive attributes of the target individuals.
Adversary: Resources	A set of non-sensitive attributes of the target individuals, including the correct value for the propensity-to-move attribute, for "inclusive individuals" (in the training set) or "exclusive individuals" (not in the training set).
Vulnerability:Opportunity	Ability to query the model to obtain output plus the marginal distributions of the data that the model was trained on. Additionally, the output might include confidence scores and a confusion matrix calculated on the training data might be available.
Countermeasure	Modify the data on which the model is trained.

Threat model

- *What information is leaked from a model that is trained on original data?*
- *Does a machine learning model trained on data that has been synthesized prevent this leak?*

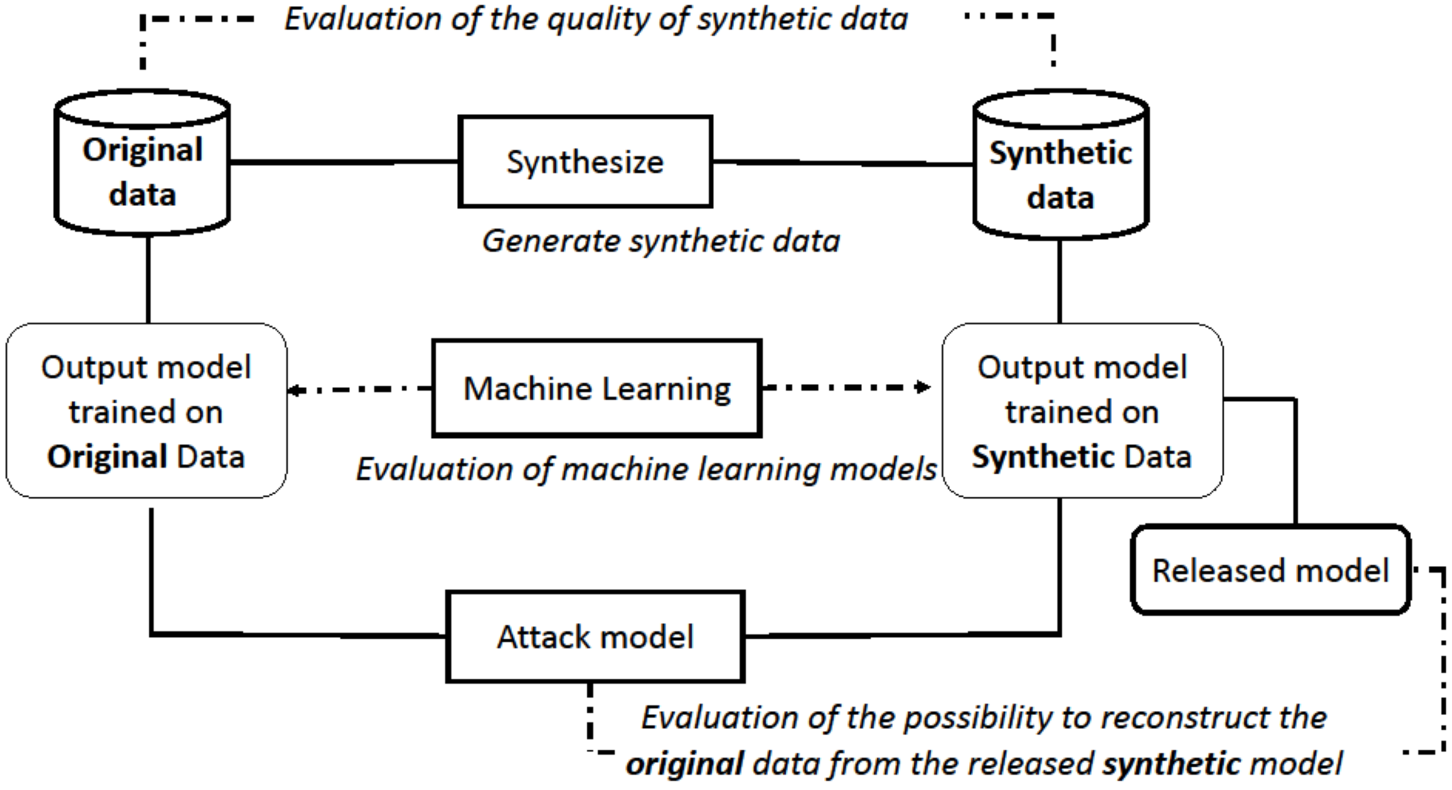
- Try to recover **sensitive** features or the **full data** sample based on **output labels** and **partial knowledge** (subset of data) of some features [Mehnaz, S et al (2022)]



Model Inversion Attack: Adversary learns certain features $x^*_i \in \mathbf{x}^*$ or statistical properties such as avg($\mathbf{x}^*$) of the training dataset

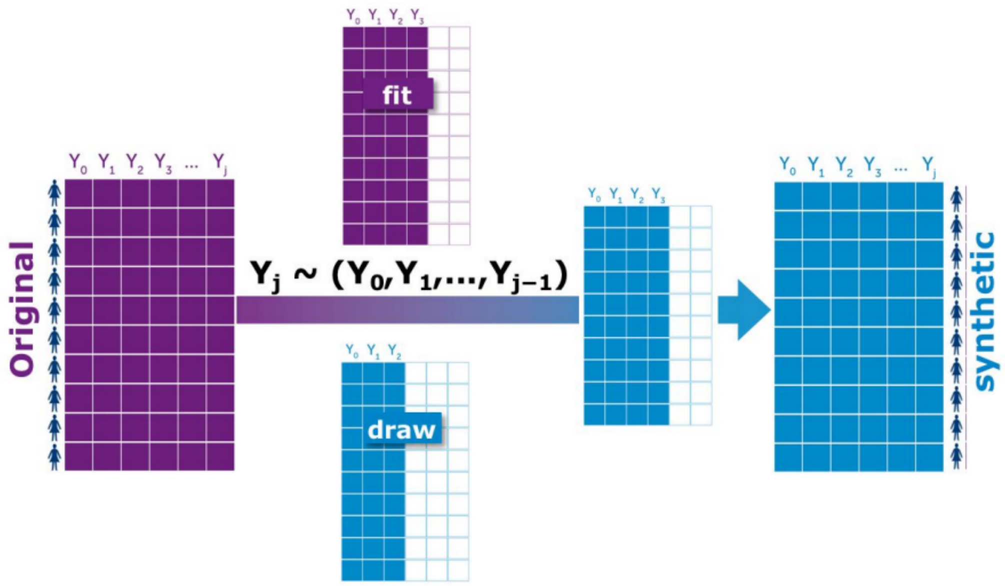
__Mehnaz__, S., Dibbo, S.V., Kabir, E., Li, N., Bertino, E.: Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In: 31st USENIX Security Symposium. pp. 4579–4596. (2022)

Machine learning using synthetic training data



Machine learning using synthetic training data

Sequentially replacing **original** values by **synthetic** values generated from **conditional probability distributions** [Beata. N et al, (2013)].



Credits picture from Synthpop an R package for generating synthetic microdata by Beata Nowok and Gillan Raab (2013).

Evaluation of machine learning algorithms

Machine Learning Algorithms		*Training and test individuals are exclusive*			*Training and test individuals are inclusive*		
		AUC	**MCC**	**F1-score**	**AUC**	**MCC**	**F1-score**
Original Data	*Random*	0.4962	-0.0105	0.2139	0.5014	0.0029	0.1633
	NaiveBayes	0.5656	-0.0328	0.5491	0.6815	0.2204	0.2992
	RandomForest	0.7061	0.3210	0.6322	0.7532	0.3121	0.4460
	DecisionTree	0.6372	0.2692	0.5376	0.6568	0.2292	0.3057
	ExtraTrees	0.7226	0.3197	0.6325	0.7597	0.3212	0.4525
	KNN	0.6304	0.2074	0.4104	0.6717	0.1744	0.2235
Synthetic Data	*Random*	0.4991	-0.025	0.2261	0.5011	0.0022	0.1657
	NaiveBayes	0.5658	0.045	0.5451	0.6822	0.2029	0.2578
	RandomForest	0.7053	0.3282	0.6343	0.7467	0.3133	0.4471
	DecisionTree	0.6489	0.2598	0.4878	0.6618	0.2125	0.3078
	ExtraTrees	0.7188	0.3185	0.6321	0.7557	0.3138	0.4464
	KNN	0.6067	0.1152	0.1857	0.6542	0.1637	0.2070

Evaluation of machine learning algorithms

Target MLs to be Released	**Data sets**	**Privacy-preserving**	**F1-Macro**	**MCC**	**G-mean**	**TN**	**FP**	**FN**	**TP**
Random Classifier	**Original data**	*None*	0.4924	0.0012	0.4924	46452	9539	17818	3686
Random Forest Classifier	**Original Data**	*None*	0.5946	0.2407	0.5779	61907	2363	10677	2548
Random Forest Classifier	**Synthetic data using CART**	*None*	0.5946	0.2426	0.5793	61848	2422	10628	2597
		Swapping	0.5881	0.2389	0.5742	62174	2096	10831	2394
		Conditional swapping	0.4654	0.0216	0.5028	63704	566	13034	191
		PRAM	0.5941	0.2415	0.5789	61844	2426	10638	2587
	Synthetic data using CTGAN	*None*	0.4586	0.0392	0.5021	64207	63	13155	70
		Differential privacy	0.4534	0.000	0.5000	64270	0	13225	0

Model inversion attribute inference attack (**Real training data**)

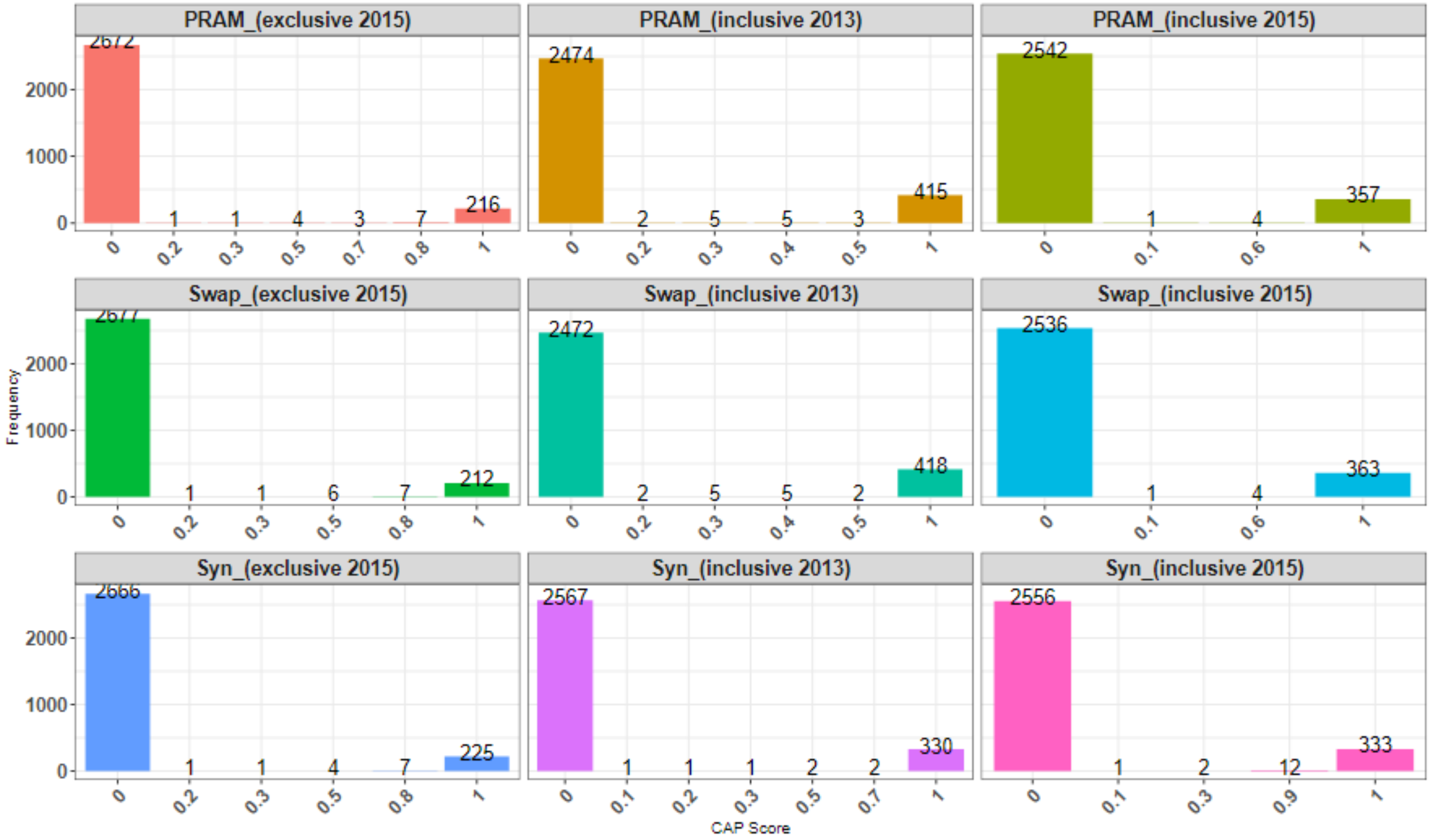
Attacker Knowledge	**Inclusive individuals (2013)**			**Inclusive individuals (2015)**			**Exclusive individuals (2015)**		
Attack models	*Gender*	*Age*	*Income*	*Gender*	*Age*	*Income*	*Gender*	*Age*	*Income*
Random Attack	0.4977	0.1238	0.1982	0.5029	0.1244	0.1991	0.5012	0.1275	0.2001
LOMIA + marginal	*0.5157*	*0.1336*	*0.2105*	**0.5035**	**0.1291**	0.1983	*0.5014*	0.1234	**0.2005**
CSMIA	0.3206	0.0105	0.0514	0.4660	0.0638	0.1581	0.4943	0.0721	0.1602
FMIA	**0.7563**	**0.6777**	**0.6898**	0.4647	0.0170	**0.2499**	**0.5205**	0.1091	0.1452

TU Delft

Model inversion attribute inference attack (**PP + synthetic training data**)

PP-Synthetic data	Attack Models	Inclusive individuals (2013)			Inclusive individuals (2015)			Exclusive individuals (2015)		
		Gender	*Age*	*Income*	*Gender*	*Age*	*Income*	*Gender*	*Age*	*Income*
Synthesis Only	*Random Attack*	0.5036	0.1228	0.2021	0.4938	0.1225	0.2033	0.4979	**0.1233**	0.1980
	LOMIA + marginal	0.4980	*0.1261*	0.1995	*0.5003*	**0.1282**	0.1972	*0.4989*	**0.1252**	0.1985
	CSMIA	0.4901	0.0675	0.1423	0.4947	0.0775	0.1544	***0.5018***	0.1012	0.1826
	FMIA	**0.5153**	0.0498	**0.3453**	**0.5007**	0.0588	**0.2772**	**0.5069**	0.1080	0.1452
Synthesis + Swapping	*Random Attack*	0.4980	0.1238	0.1974	0.4979	0.1233	0.2060	0.4975	0.1248	**0.1973**
	LOMIA + marginal	**0.5012**	**0.1280**	*0.1984*	0.4972	*0.1265*	0.1984	*0.5032*	**0.1242**	***0.1988***
	CSMIA	0.4958	0.1198	**0.2032**	**0.4996**	0.1175	0.1848	***0.5093***	**0.1457**	*0.1986*
	FMIA	0.4473	0.0901	0.0792	0.4320	**0.1362**	**0.3098**	**0.5351**	0.1020	0.1452
Synthesis + PRAM	*Random Attack*	0.5002	0.1259	0.2010	0.5063	0.1239	0.2039	0.5002	0.1255	0.2000
	LOMIA + marginal	**0.5038**	**0.1274**	0.1963	0.5004	0.1238	0.2002	0.5004	**0.1247**	**0.1987**
	CSMIA	0.4967	0.1175	0.1701	0.4913	0.1059	0.1827	0.4895	**0.1371**	**0.2070**
	FMIA	0.4827	0.0282	0.1635	**0.5286**	0.1129	0.1188	**0.5120**	0.1019	0.1452

Attribute Disclosure using Correct Attribution Probability (CAP)



Outlook: Attacking model and output data

- Investigation of an **attack** on a machine learning model
- Exploration of the ability of **privacy-preserving techniques** on **synthetic training data** to protect against model inversion attribute inference attack
- Measuring the **disclosure risk** per individuals using correct attribution probability

Take-aways

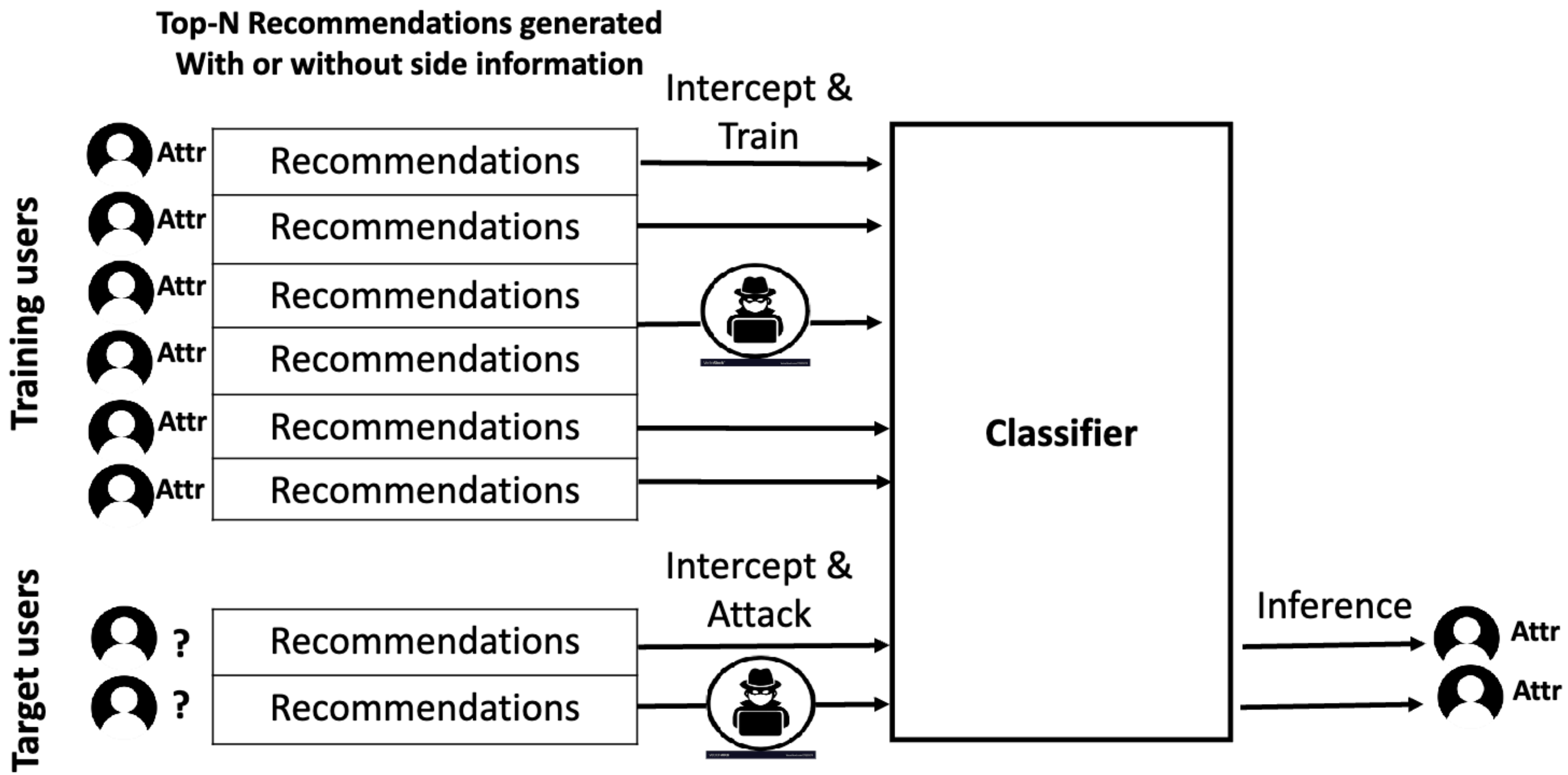
- A specific **purpose** should be always behind the creation of synthetic data
- Synthetic data does **not necessarily** protect against inference attack
- Exploration of other more **threat models**, e.g., gray-box or white-box attacks.

Part 2:

Chapter 6: A Closer Look at User Attributes in Recommendations: Implications for **Privacy** and **Diversity**

*This chapter is under preparation as **Manel Slokom**, Jesse Brons, Özlem Özgobek and Martha Larson. A closer Look at User Attributes in Recommendations: Implications for Privacy and Diversity.*

Threat model



TU Delft

Recommendation Performance of Standard Recommenders

Data Sets	ML100K		ML1M		LastFM	
Algorithms	**N = 5**	**N = 10**	**N = 5**	**N = 10**	**N = 5**	**N = 10**
MostPop	0.0484	0.0583	0.0275	0.0383	0.2135	0.2079
ItemKNN	0.0704	0.0831	0.0342	0.0479	0.2790	0.2671
UserKNN	**0.0795**	0.0898	0.0334	0.0468	0.3190	0.3036
BPRMF	0.0748	0.0848	0.0561	0.0586	**0.3436**	**0.3132**
FM	0.0771	**0.0905**	**0.0639**	**0.0687**	0.3088	0.2888

TU Delft

Leaking in the Output of Standard Recommenders

			Top-N = 5				Top-N = 10			
Data Sets	*Recommenders*	Classifiers	**Gender**	**Age**	**Occupation**	**State**	**Gender**	**Age**	**Occupation**	**State**
ML100K		*Majority-class*	0.4330	0.1381	0.0154	0.0053	0.4330	0.1381	0.0154	0.0053
	MostPop	*LogReg*	0.4428	*0.1354*	0.0412	0.0081	0.4464	0.1629	0.0280	0.0087
	UserKNN	*LogReg*	0.4865	0.1923	0.0492	**0.0165**	0.5012	0.1847	**0.0631**	0.0161
	ItemKNN	*LogReg*	0.5196	0.1789	**0.0674**	0.0095	0.4944	**0.2165**	0.0599	**0.0233**
	BPRMF	*LogReg*	**0.5334**	0.1390	0.0403	0.0145	**0.5642**	0.1631	0.0383	0.0081
	FM	*LogReg*	0.5015	**0.2012**	0.0394	0.0149	0.5470	0.1884	0.0329	0.0129
ML1M		*Majority-class*	0.4160	0.1299	0.0104	0.0058	0.4160	0.1299	0.0104	0.0058
	MostPop	*LogReg*	*0.4158*	0.2257	0.0308	0.0075	0.4327	0.2623	0.0421	0.0108
	UserKNN	*LogReg*	0.5665	0.3276	0.0587	0.0165	0.5840	0.3333	**0.0737**	0.0161
	ItemKNN	*LogReg*	0.5758	0.3330	0.0618	0.0164	0.6077	0.3354	0.0540	**0.0193**
	BPRMF	*LogReg*	0.5305	0.3364	**0.0627**	0.0103	0.5607	0.3602	0.0615	0.0182
	FM	*LogReg*	**0.6163**	**0.3671**	0.0520	**0.0171**	**0.6346**	**0.3730**	0.0723	0.0154
			Gender	**Continent**	**EU vs. Rest**		**Gender**	**Continent**	**EU vs. Rest**	
LastFM		*Majority-class*	0.3646	0.1126	0.3377		0.3646	0.1126	0.3377	
	MostPop	*LogReg*	0.5035	0.1298	0.4963		0.4990	0.1321	0.4704	
	UserKNN	*LogReg*	0.5323	0.1914	**0.5456**		0.5249	0.1897	0.5015	
	ItemKNN	*LogReg*	0.5250	**0.2092**	0.5171		0.5275	0.1776	0.4957	
	BPRMF	*LogReg*	**0.5479**	0.1721	0.5337		**0.5595**	0.1892	0.5328	
	FM	*LogReg*	0.5015	0.1719	0.5111		0.5160	**0.2205**	**0.5635**	

Recommendation Performance of Context-aware Recommenders

Factorization Machine

Data Sets	ML100K		ML1M		LastFM		
User Attributes	**Top-N = 5**	**Top-N = 10**	**Top-N = 5**	**Top-N = 10**	*User Attributes*	**Top-N = 5**	**Top-N = 10**
None	0.0771	0.0905	0.0639	0.0687	*None*	0.3088	0.2888
Gender	0.0932	**0.1097**	0.0647	0.0688	*Gender*	**0.3196**	0.3049
Age	0.0888	0.1013	0.0644	*0.0684*	*continent*	0.3125	0.2996
Occupation	0.0903	0.1025	*0.0620*	*0.0657*	*EU vs Rest*	0.3188	**0.3061**
State	**0.0933**	0.1082	**0.0665**	**0.0721**			

GNN-Pre-train

Algorithms	*Side information*	ML100K		ML1M		LastFM	
		Top-N = 5	**Top-N = 10**	**Top-N = 5**	**Top-N = 10**	**Top-N = 5**	**Top-N = 10**
FM	*None*	0.0771	0.0905	0.0639	0.0687	0.3088	0.2888
	All User & item Attributes	0.0518	0.0759	0.0444	0.0488	0.2811	0.2200
GNN (Single-P)	*All User & item Attributes*	0.0757	0.0985	0.0626	0.0800	0.4928	0.4827

TU Delft

Leaking in the Output of Context-aware Recommenders

Factorization Machine

		Top-N = 5				Top-N = 10			
Data Sets	*User Attributes*	**Gender**	**Age**	**Occupation**	**State**	**Gender**	**Age**	**Occupation**	**State**
ML100K	None	**0.5015**	**0.2012**	0.0394	0.0149	**0.5470**	0.1884	0.0329	**0.0129**
	With side information	0.4871	0.1843	**0.0476**	**0.0162**	0.5269	**0.2112**	**0.0533**	0.0128
ML1M	None	0.6163	0.3671	0.0520	**0.0171**	0.6346	0.3730	**0.0723**	0.0154
	With side information	**0.6275**	**0.4025**	**0.0531**	0.0148	**0.6520**	**0.4401**	0.0704	**0.0197**
		Gender	**Continent**	**EU vs. Rest**		**Gender**	**Continent**	**EU vs. Rest**	
LastFM	None	0.5015	0.1719	0.5111		0.5160	0.2205	0.5635	
	With side information	**0.5578**	**0.2031**	**0.6197**		**0.5427**	**0.2430**	**0.6250**	

Leaking in the Output of GNN Recommenders

GNN-Pre-train

Classifier= LogReg			Top-N = 5				Top-N = 10			
Data Sets	*Algorithms*	*Side information*	**Gender**	**Age**	**Occupation**	**State**	**Gender**	**Age**	**Occupation**	**State**
ML100K	*FM*	User attribute	0.4871	0.1843	0.0476	0.0162	0.5269	0.2112	0.0533	0.0128
		User & Item Attributes	0.5032	0.1604	0.0643	0.0127	0.5404	0.2202	0.0350	0.0100
	GNN (Single-P)	User & Item Attributes	0.4807	0.1513	0.0327	0.0145	0.4823	0.2018	0.0390	0.0224
ML1M	*FM*	User attribute	0.6275	0.4025	0.0531	0.0148	0.6520	0.4401	0.0704	0.0197
		User & Item Attributes	0.6288	0.4176	0.0696	0.0160	0.6405	0.4625	0.0834	0.0161
	GNN (Single-P)	User & Item Attributes	0.4179	0.1759	0.0289	0.0061	0.4234	0.2037	0.0341	0.0106
			Gender	**Continent**	**EU vs. Rest**		**Gender**	**Continent**	**EU vs. Rest**	
LastFM	*FM*	User attribute	0.5578	0.2031	0.6197		0.5427	0.2430	0.6250	
		User & Item Attributes	0.5781	0.1941	0.5458		0.5162	0.1820	0.4817	
	GNN (Single-P)	User & Item Attributes	0.4711	0.1366	0.5062		0.5164	0.1806	0.5592	

Diversity in the Output of Context-aware Recommenders

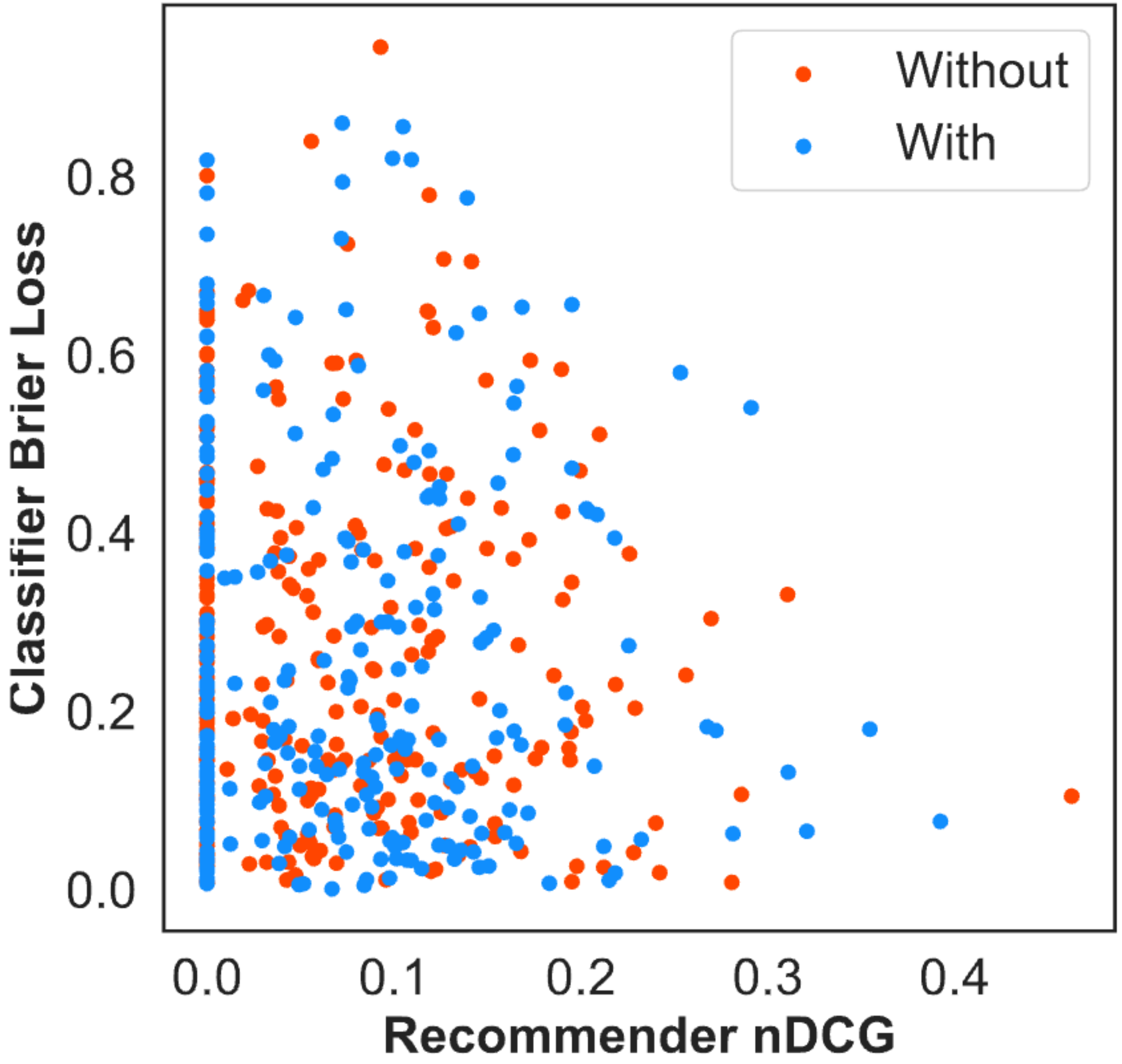
Data Sets	User Attributes	Top-N = 5			Top-N = 10		
		Item coverage	**Shannon Entropy**	**Gini index**	**Item coverage**	**Shannon Entropy**	**Gini index**
ML100K	*None*	415	7.770	0.109	546	8.097	0.136
	Gender	315↓	7.275↓	0.077↓	422↓	7.657↓	0.099↓
	Age	336↓	7.075↓	0.070↓	461↓	7.564↓	0.096↓
	Occupation	424↑	7.697↓	0.105↓	563↑	8.072↓	0.134↓
	State	369↓	7.514↓	0.092↓	507↓	7.888↓	0.117↓
ML1M	*None*	840	7.995	0.056	1110	8.376	0.072
	Gender	647↓	7.572↓	0.042↓	1181↑	8.363↓	0.074↑
	Age	687↓	7.308↓	0.037↓	902↓	7.741↓	0.049↓
	Occupation	779↓	7.657↓	0.047↓	985↓	8.058↓	0.058↓
	State	901↑	8.031↑	0.059↑	1181↑	8.363↓	0.074↑
LastFM	*None*	1180	9.173	0.041	1802	9.547	0.054
	Gender	1107↓	9.167↓	0.039↓	1625↓	9.543↓	0.051↓
	Continent	1152↓	9.246↑	0.042↑	1668↓	9.595↑	0.053↓
	EU vs Rest	814↓	8.483↓	0.025↓	1195↓	8.895↓	0.033↓

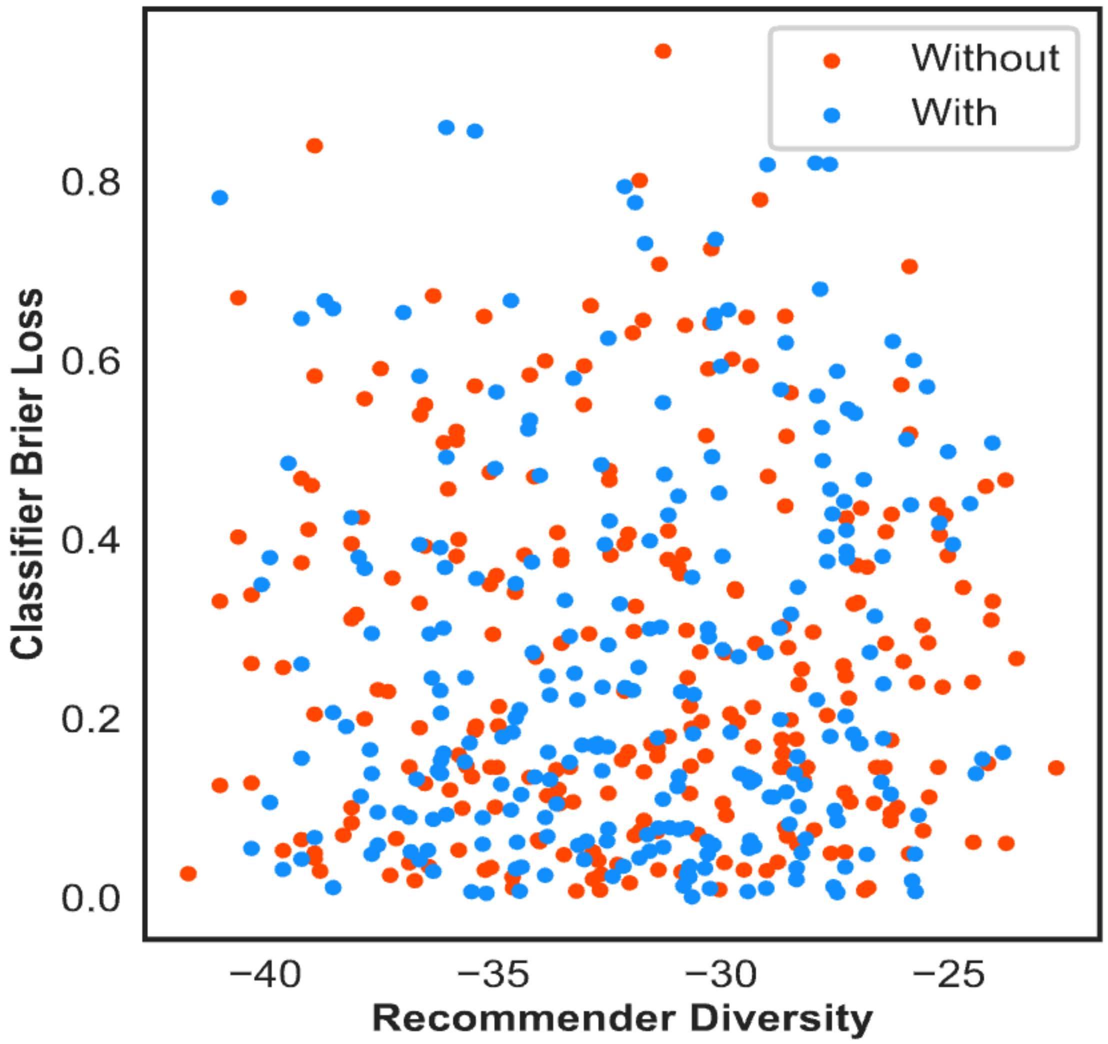
TU Delft

Diversity in the Output of GNN Recommenders

Data Sets	*Algorithms*	*Side Information*	Top-N = 5			Top-N = 10		
			Items coverage	***Shannon Entropy***	***Gini index***	***Items coverage***	***Shannon Entropy***	***Gini index***
ML100K	*FM*	*None*	415	7.770	0.109	546	8.097	0.136
		All User & item Attributes	215↓	6.357↓	0.039↓	302↓	6.853↓	0.0553↓
	GNN (Single-P)	*All User & item Attributes*	128 ↓	6.145↓	0.033↓	163↓	6.698↓	0.050↓
ML1M	*FM*	*None*	840	7.995	0.056	1110	8.376	0.072
		All User & item Attributes	431↓	6.985↓	0.027↓	575↓	7.448↓	0.037↓
	GNN (Single-P)	*All User & item Attributes*	125↓	4.749 ↓	0.005↓	220↓	5.650↓	0.010↓
LastFM	*FM*	*None*	1180	9.173	0.041	1802	9.547	0.054
		All User & item Attributes	732↓	8.365↓	0.022↓	1107↓	8.816↓	0.0296↓
	GNN (Single-P)	*All User & item Attributes*	162↓	6.034↓	0.004↓	299↓	6.879↓	0.007↓

TU Delft


Without
With
Classifier Brier Loss
0.0
0.2
0.4
0.6
0.8
0.0
0.1
0.2
0.3
0.4
Recommender nDCG


Without
With
Classifier Brier Loss
0.0
0.2
0.4
0.6
0.8
−40
−35
−30
−25
Recommender Diversity

TU Delft

Conclusion: Attacking model and output data

- Investigation of user attributes from a perspective of **privacy** and **diversity**
  - **Privacy**: **standard** recommenders leak and that using user attributes as **side information** during the training of a **context**-aware recommender system may **exacerbate this leak**.
  - **Diversity: user attributes** restricts the **coverage** of a recommender system and **lowers the diversity**.

- Recommender system platforms should consider **carefully** whether it is *advantageous* to make use of user attributes for training recommender systems.

Take-aways

- It is important to consider whether **side information** is actually bringing a substantial benefit and to ensure that there are no hidden '**side effects**'.

- We should *not assume* that making recommendation lists more indicative of a particular user attribute, i.e., 'female' will better satisfy users with that attribute.

- We should *not assume* that there is **a trade-off** between leak reduction and recommender system performance.
  - The **best of both** is worth pursuing

Outlook and Discussion

Outlook

1. Attacking input data

- Data **obfuscation** for recommender systems.
- Personalized blurring.
- From privacy to fairness and diversity.

2. Attacking model and output data

- Investigation of an **attack** on a machine learning model.
- Exploration of the ability of **privacy-preserving** techniques on **synthetic data** to protect against model inversion attribute inference attack.
- Investigation of user attributes from a perspective of **privacy** and **diversity in context-aware recommendations**

Moving forward (Self-reflections)!

1. **Trade-offs** *should* not exist
   - Bias mitigation,
   - Fair / diverse recommendations
2. Responsible predictions for: Individual vs group

3. **Averaging** scores could result in a loss of information:
   - ➢ What privacy, fairness, diversity should be in a user-level!
   - ➢ Users should be treated differently as they are different, i.e., different profile sizes, interests!

4. Diverse and fair recommendations are context dependent

Propositions

1. More data does not necessarily lead to a better model performance.
   This proposition pertains to this thesis.
2. Privacy-accuracy trade-offs should not exist.
   This proposition pertains to this thesis.
3. Every type of attack requires a careful selection of privacy protection.
   This proposition pertains to this thesis.
4. Synthetic data amplifies societal harms as much as real data do.
   This proposition pertains to this thesis.
5. Top-rated toolboxes fail to guarantee the reproducibility of results.
6. Perfection stifles productivity.
7. The potential of negative results needs more attention.
8. Social media distorts our perception of reality.
9. The path to self-discovery in life lies not in finding our passion but in finding our purpose.
10. Years of experience lose value if not paired with self-doubt.

These propositions are regarded as opposable and defendable, and have been approved

Thank You!

Martha Larson (Radboud University)
Alan Hanjalic (TU Delft)
Peter-Paul de Wolf (CBS)
Laura Hollink (CWI)