Safeguarding privacy: how to leverage synthetic data?

IPEN 2021

16 June 2021

Statice

statice.ai

Statice GmbH

Berlin-based company

Since 2017

Synthetic data and Privacy

Omar Ali Fdal
Co-founder & CEO

omar@statice.ai

Today's agenda

01 Data release and challenges of preserving privacy

- Linkage and re-identification
- Inference and Attribution
- From pseudonymization to synthetic data

02 Synthetic data as a privacy mechanism

- By design
- Combined with other techniques
- Practical risk assessment

01

Data release and challenges of preserving privacy

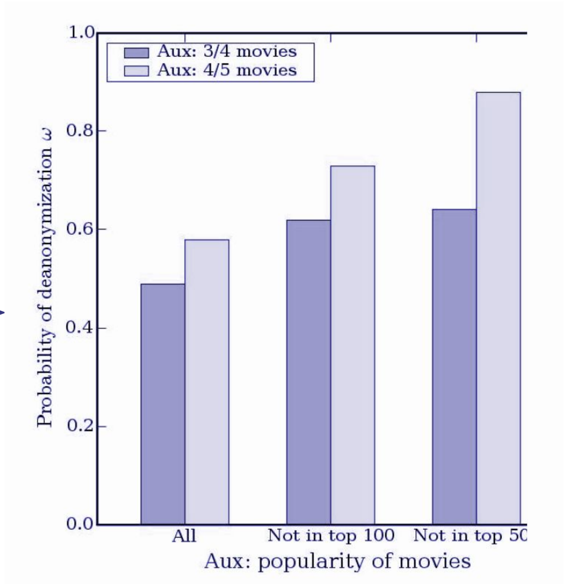
Risks and mitigation tactics

The risks related to data release

- (re-)identification and linkage
- (specific) Attribute inference

Illustration

Netflix movie preferences



Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. InSecurity and Privacy, 2008. SP 2008. IEEE Symposium on 2008 May 18 (pp. 111-125). IEEE.

Researchers **re-identified significant numbers of Netflix users and their viewing habits** by matching the **redacted viewing information with IMDb ratings**.

Types of risk

Linkage and re-identification

- Uniqueness

> *Simple Demographics Often Identify People Uniquely*
>
> Latanya Sweeny, 2000

- Background knowledge and auxiliary information

Attribute Inference

- **General inference**: Learning that “smoking causes cancer”
- **“Specific” inference**: Information that can only be learned based on the specific dataset at hand but not from the population

Common data protection techniques

- Pseudonymization
- K-anonymization
- No data?

Data protection

In the beginning was the data

phone	race	birth year	sex	zip code	medical condition	headache
015940192	white	1964	f	1203002	chest_pain	10110010110100010
010405919	white	1964	f	1203505	obesity	100000100000111010
011500159	white	1964	f	1203106	short_breath	10110010110100010
010192042	black	1965	m	5403221	heart_disease	1010010110100010
015909191	black	1965	m	5403221	heart_disease	010010110100010
015553436	black	1965	m	5403221	heart_disease	10010010110100010
016901095	white	1960	f	3003202	ovarian cancer	1111001110100010
017497297	white	1960	f	3003555	ovarian cancer	10110010000000010
018206810	white	1960	m	3003890	prostate cancer	00000011100000010

Statice

Pseudonymization: protecting “obvious identifiers”

phone	race	birth year	sex	zip code	medical condition	headache
[redacted]	white	1964	f	1203002	chest_pain	10110010110100010
[redacted]	white	1964	f	1203505	obesity	10000010000011010
[redacted]	white	1964	f	1203106	short_breath	10110010110100010
[redacted]	black	1965	m	5403221	heart_disease	1010010110100010
[redacted]	black	1965	m	5403221	heart_disease	010010110100010
[redacted]	black	1965	m	5403221	heart_disease	10010010110100010
[redacted]	white	1960	f	3003202	ovarian cancer	1111001110100010
[redacted]	white	1960	f	3003555	ovarian cancer	10110010000000010
[redacted]	white	1960	m	3003890	prostate cancer	0000001110000010

Pseudonymous data is personal data

> *… Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person.*

-- Recital 26, GDPR

Data protection

K-anonymity: protecting "quasi-identifiers"

race	birth year	sex	zip code	medical condition	headache
white	1964	f	1203002	chest_pain	10110010110100010
white	1964	f	1203505	obesity	10000010000011010
white	1964	f	1203106	short_breath	10110010110100010
black	1965	m	5403221	heart_disease	1010010110100010
black	1965	m	5403221	heart_disease	010010110100010
black	1965	m	5403221	heart_disease	10010010110100010
white	1960	f	3003202	ovarian cancer	1111001110100010
white	1960	f	3003555	ovarian cancer	1011001000000010
white	1960	m	3003890	prostate cancer	0000001110000010

Statice

K-anonymity: protecting "quasi-identifiers"

Transform the data so that unique joins that expose sensitive attributes are no longer possible.

phone	race	birth year	sex	zip code
015940192	white	1964	f	1203002

phone	race	birth year	sex	zip code
015909191	black	1965	f	5403014
018206810	white	1960	m	3003890

race	birth year	sex	zip code	medical condition
white	1964	*	1203*	chest_pain
white	1964	*	1203*	obesity
white	1964	*	1203*	short_breath
black	1965	*	5403*	heart_disease
black	1965	*	5403*	heart_disease
black	1965	*	5403*	heart_disease
white	1960	*	3003*	ovarian cancer
white	1960	*	3003*	ovarian cancer
white	1960	*	3003*	prostate cancer

P. Samarati and L. Sweeney, Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement through Generalization and Suppression

Data protection

Can we do better than no data?

phone	race	birth year	sex	zip code	medical condition	headache
[redacted]						
[redacted]						
[redacted]						
[redacted]						
[redacted]						
[redacted]						
[redacted]						
[redacted]						
[redacted]						

Statice

02

Synthetic Data as a protection mechanism

By Design and Risk-based

Privacy by Design

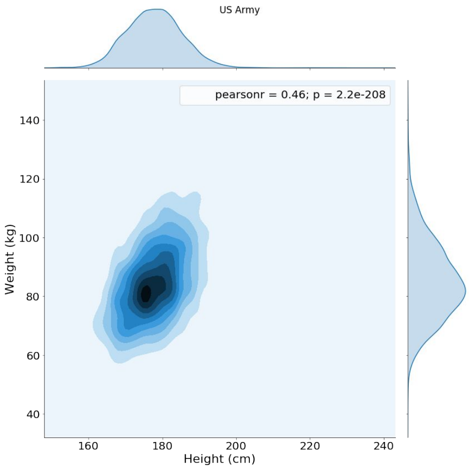
What is synthetic data?

Fully artificial, algorithmically generated data that approximate original data and that can be used for the same purposes as the original.

Principles of fully Synthetic data

Learnt distribution

Original data



synthetic data

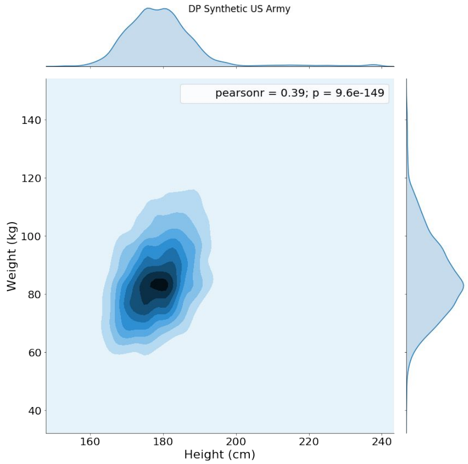
Irreversible processing: There is **no key** to retrieve the original records from the synthetic records

Privacy by Design

Synthetic data meets Differential Privacy

Original data

Learnt distribution with
Differential Privacy



Differentially Private
synthetic data

Other techniques and principles can also be combined with synthetic data

How do we measure the risks in Synthetic Data

- Linkage potential
- Attribute inference risk

Statice

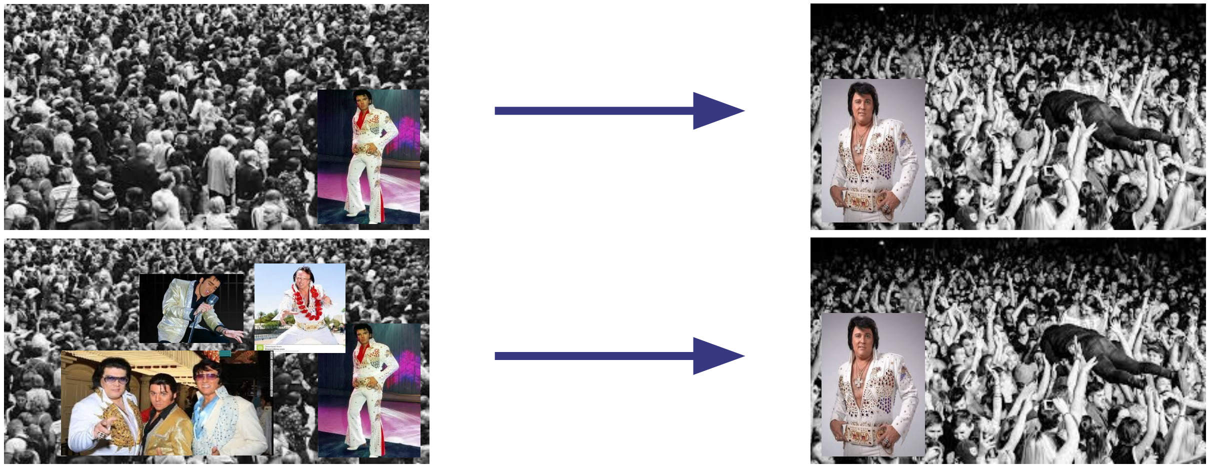
Risk Assessment

Linkage Potential

Objective: detect **suspicious records**, e.g. close matches and sensitive duplicates

Suspicious

Not suspicious



Original crowd

Synthetic crowd

Risk Assessment

Linkage Potential

Suspicious Records

185 (out of 8000 records) suspicious records found

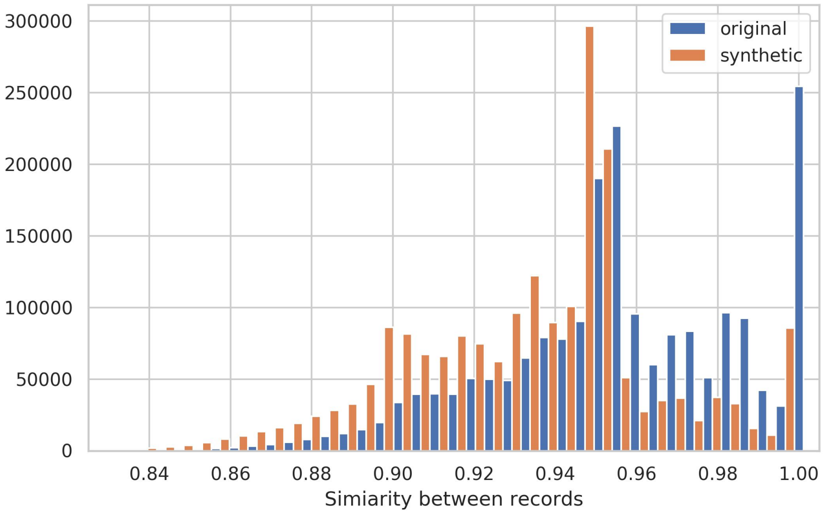
Dataset	Row	Linkage Potential	col_01	col_02	col_03	col_04	col_05	col_06	col_07	col_08
Synthetic	3273	0.786	35000	30000	1122.89	36	7.9	A	A5	Columbia University
Original	2389		33500	33500	1063.74	36	8.9	A	A5	best friends
Synthetic	590	0.786	28000	28000	708.29	60	23.63	F	F2	The Clorox Company
Original	564		30000	30000	850.55	60	23.28	F	F2	FRANZ FAMILY BAKERIES
Synthetic	4027	0.779	2800	8325	73.44	60	19.72	E	E2	Mcdean inc
Original	5084		6000	6000	226.06	36	21.0	E	E2	Nesco Service Company
Synthetic	5256	0.772	10000	15000	332.72	36	9.49	B	B2	Dept. of Navy-Fleet Readiness Cer
Original	3191		10000	10000	328.06	36	11.14	B	B2	Abbott Northwestern Hospital

Risk Assessment

Linkage Potential

A match between two rare values has a greater importance than a match between more common values.

Original records are closer to other original records, than they are to synthetic records.



Statice

Risk Assessment

Attribute Inference risk evaluator

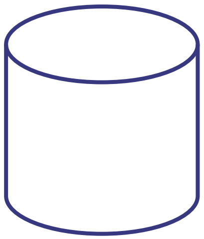
Objective: detect **specific information leaks** about the data sample

age	type_employer	education	marital	occupation	relationship	race	sex	hr_per_week	country	income
20	Self-emp-not-inc	HS-grad	Never-married	Farming-fishing	Not-in-family	White	Male	33	United-States	<=50K

1) The adversary knows **some of the attributes** of a set of target records

2) using this knowledge, they search for best matches in the **synthetic data**.

3) The results of the inference complete their knowledge of the **secret attributes**.



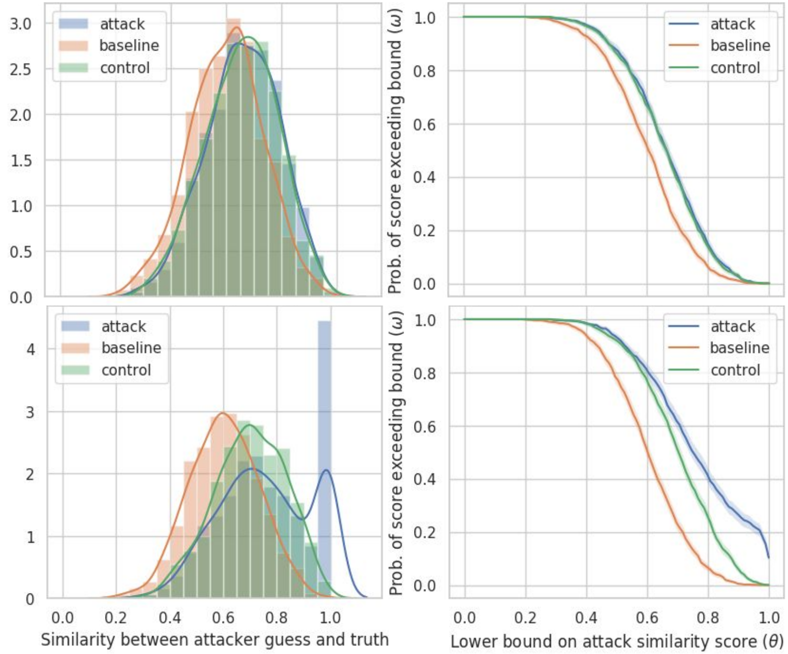
Risk Assessment

Attribute Inference risk evaluator

measure success of the attack for different amount of auxiliary knowledge, comparing training and test data.

Private synthesization

Leaky synthesization



Statice

Take-aways

- Releasing data is challenging
- Synthetic data can be both useful and private
- Understanding your risks is still crucial

statice.ai

Statice GmbH

Thank you!

Omar Ali Fdal
Co-founder & CEO

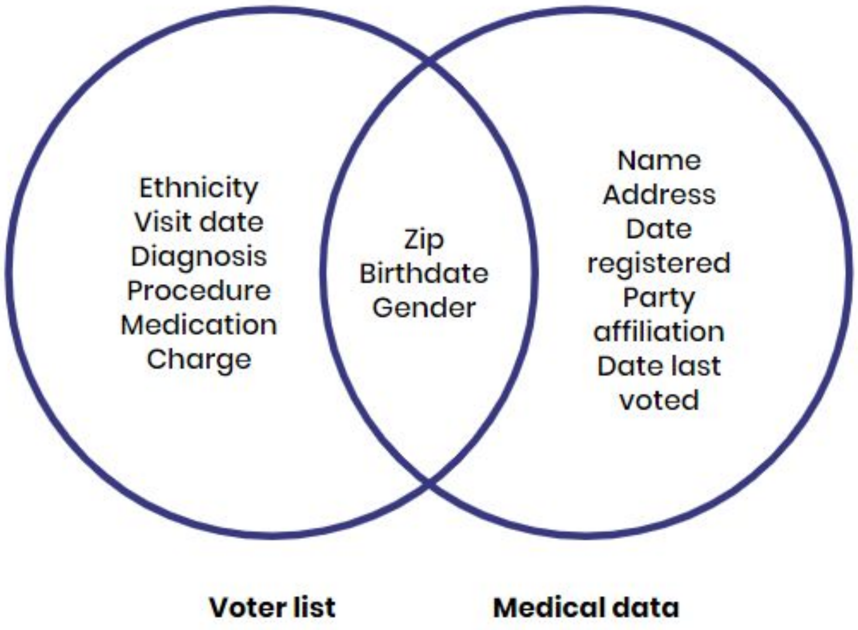
omar@statice.ai

Illustration

Massachusetts Governor health records

This privacy breach **demonstrated clearly that simply removing PII is not enough**.

Even **3 variables** available from a $25 voter registration list **were enough to be able to uniquely identify individuals** from redacted medical records.



Sweeney, Latanya. Weaving Technology and Policy Together to Maintain Confidentiality. Journal of Law, Medicine and Ethics, Vol. 25 1997, p. 98-110

Statice

01

Risks of “anonymous” data

About Linkage and Inference

03

Synthetic Data as a protection mechanism

By Design and Risk-based

02

A brief history of data protection

From pseudonymization to synthetic data

02

A brief history of data protection

From pseudonymization to synthetic data

… hopefully, we can do better!

a) We can do much better -> synthetic data: at this point, in order not to repeat what previous speakers have said, we will show a few cases showing similar performance for complex tasks (ML, forecasts, etc.)

K-anonymity (in all flavours) carries risk … and significantly reduces utility

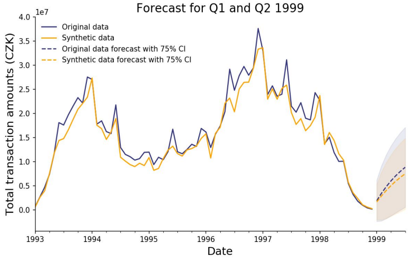
a) There are no “quasi-identifiers” : when it comes to privacy, all attributes are to be protected
b) K-anonymous data has non-negligible risk of re-identification

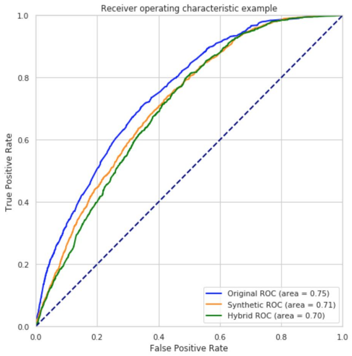
Initial sanitizing of original data

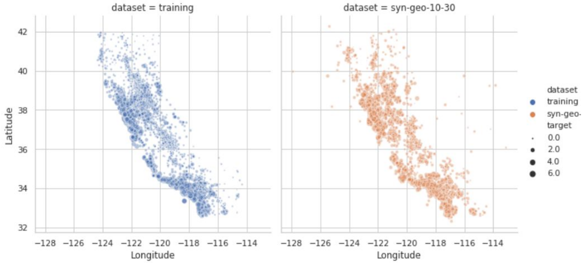
1) Risk Assessment and initial processing of original data
    i) Detection of uniques / outliers
    ii) Detection of sensitive data (we can also assume this has been done before, using other means, e.g. pseudonymization)

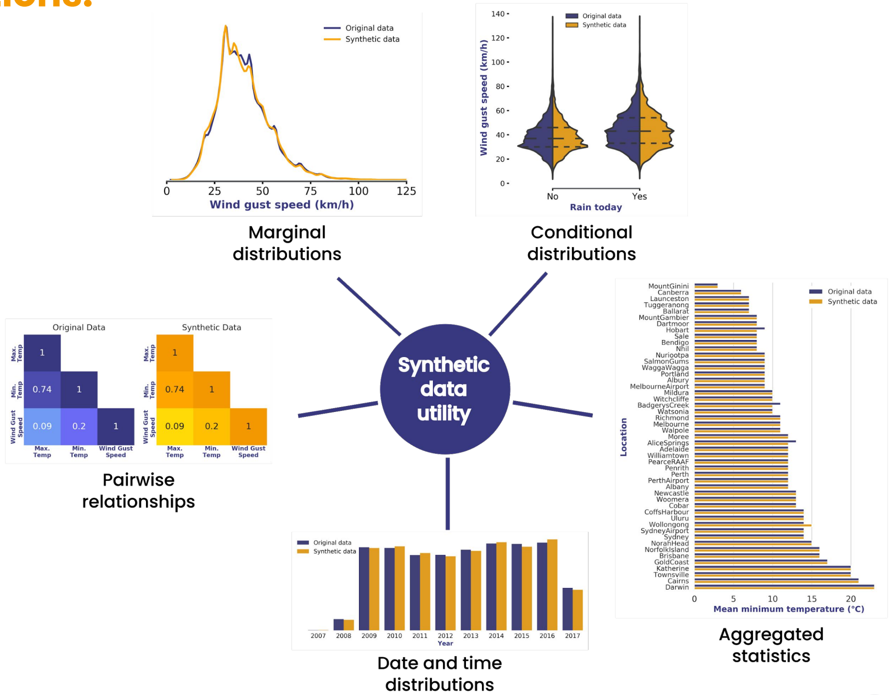
Utility evaluations

A set of **utility evaluations** assess the **quality and integrity of the synthetic data**, including for **Machine Learning applications.**







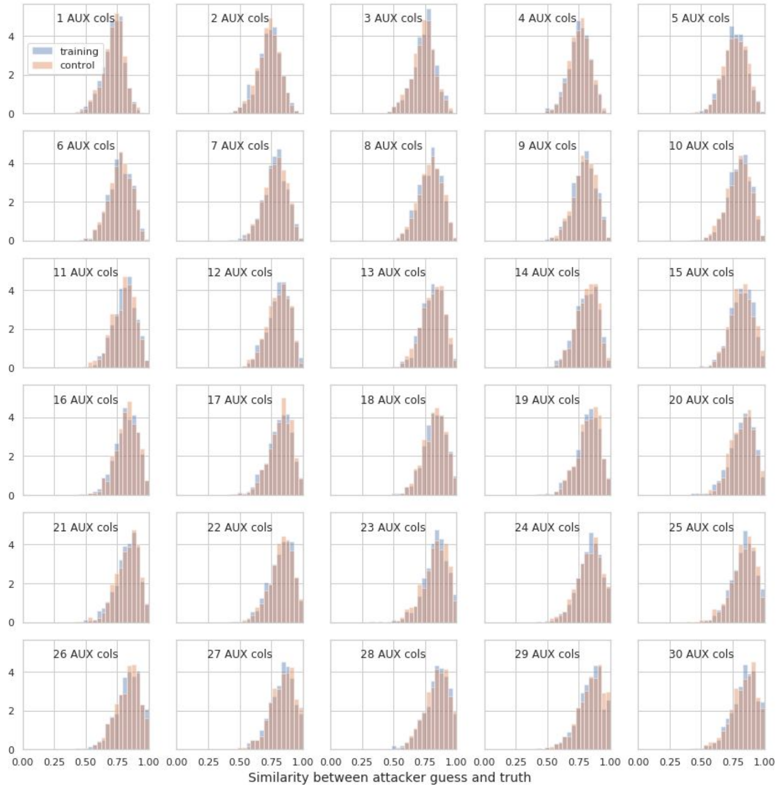


Statice

Case A: Privacy Analysis

Inference risk evaluator:

- The inference is equally successful on the control data (unseen during synthesization)
- This means that in the synthesization process, no specific information about some of the records has leaked into the synthetic data.



Risk Assessment

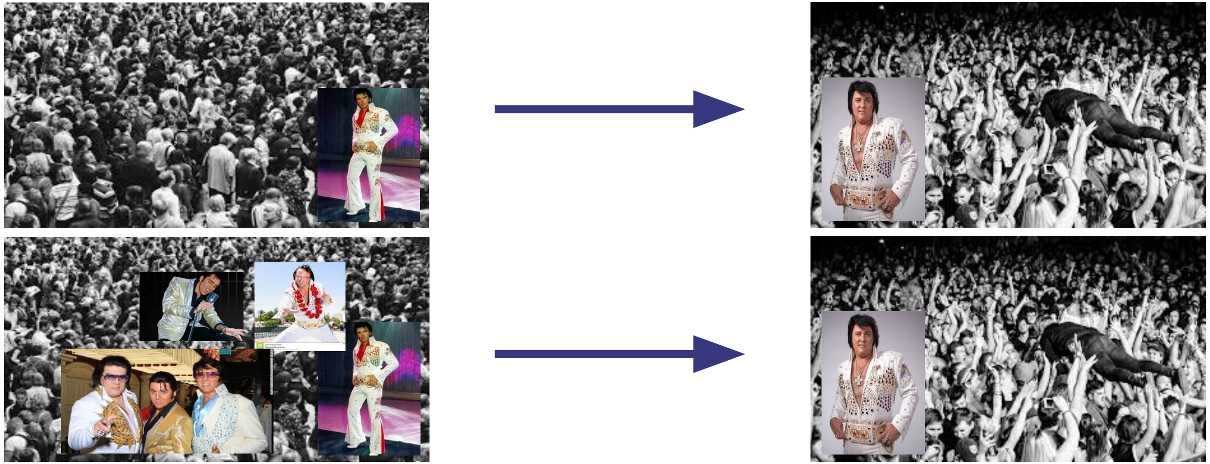
Linkage Potential

Intuition: synthetic records should not be closer to the original ones than original records are to other original records.

Objective: detect **suspicious records**, e.g. close matches and sensitive duplicates

Suspicious

Not suspicious



Original crowd

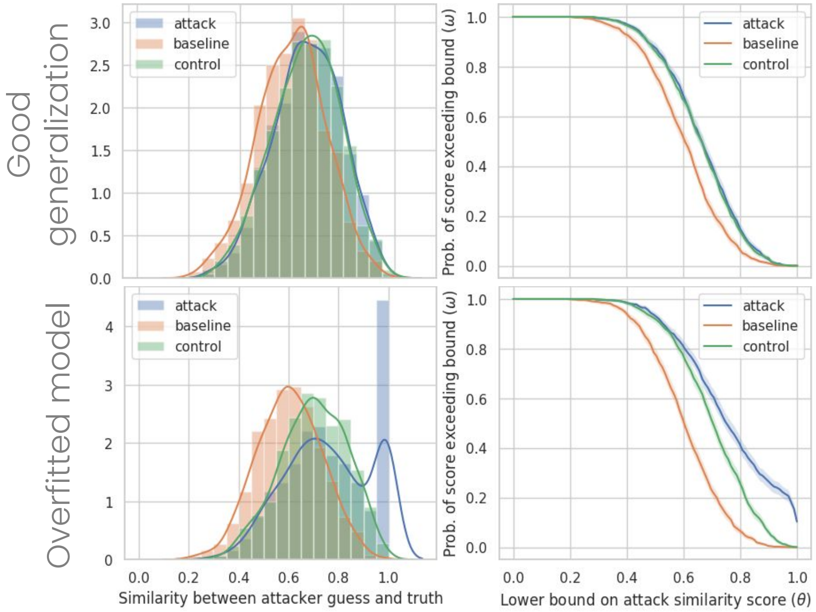
Synthetic crowd

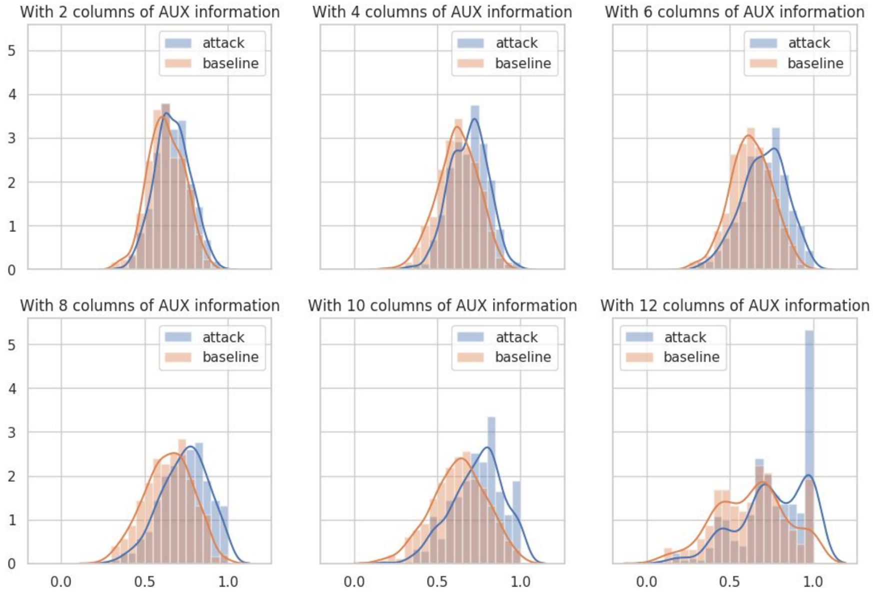
Statice

Risk Assessment

Attribute Inference risk evaluator

measure success of the attack for different amount of auxiliary knowledge, comparing training and test data.





Statice

Intro: why synthetic data?

- Synthetic data as data release mechanism
- Internal, external