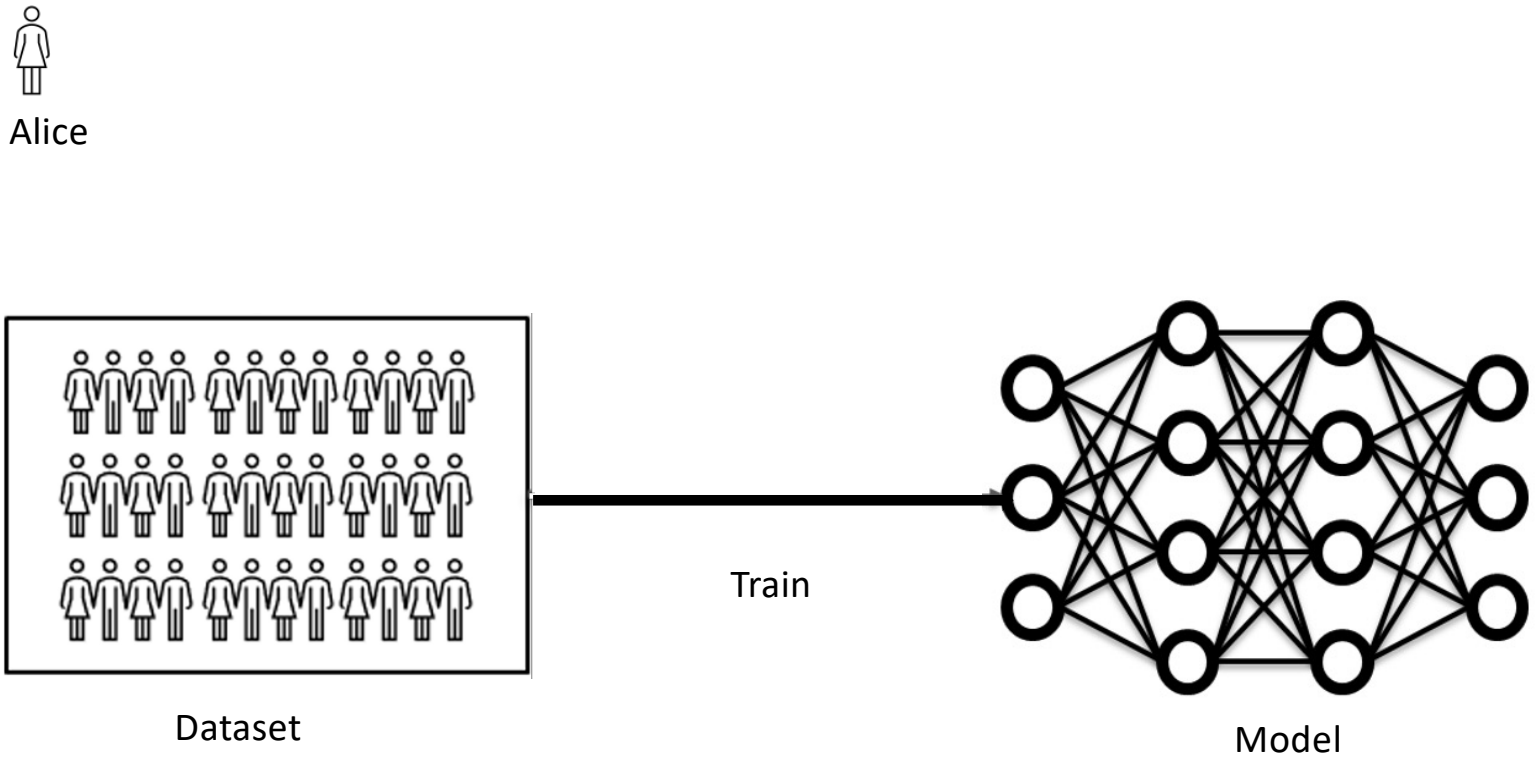
The Impact of Synthetic Data on Membership Inference Attacks

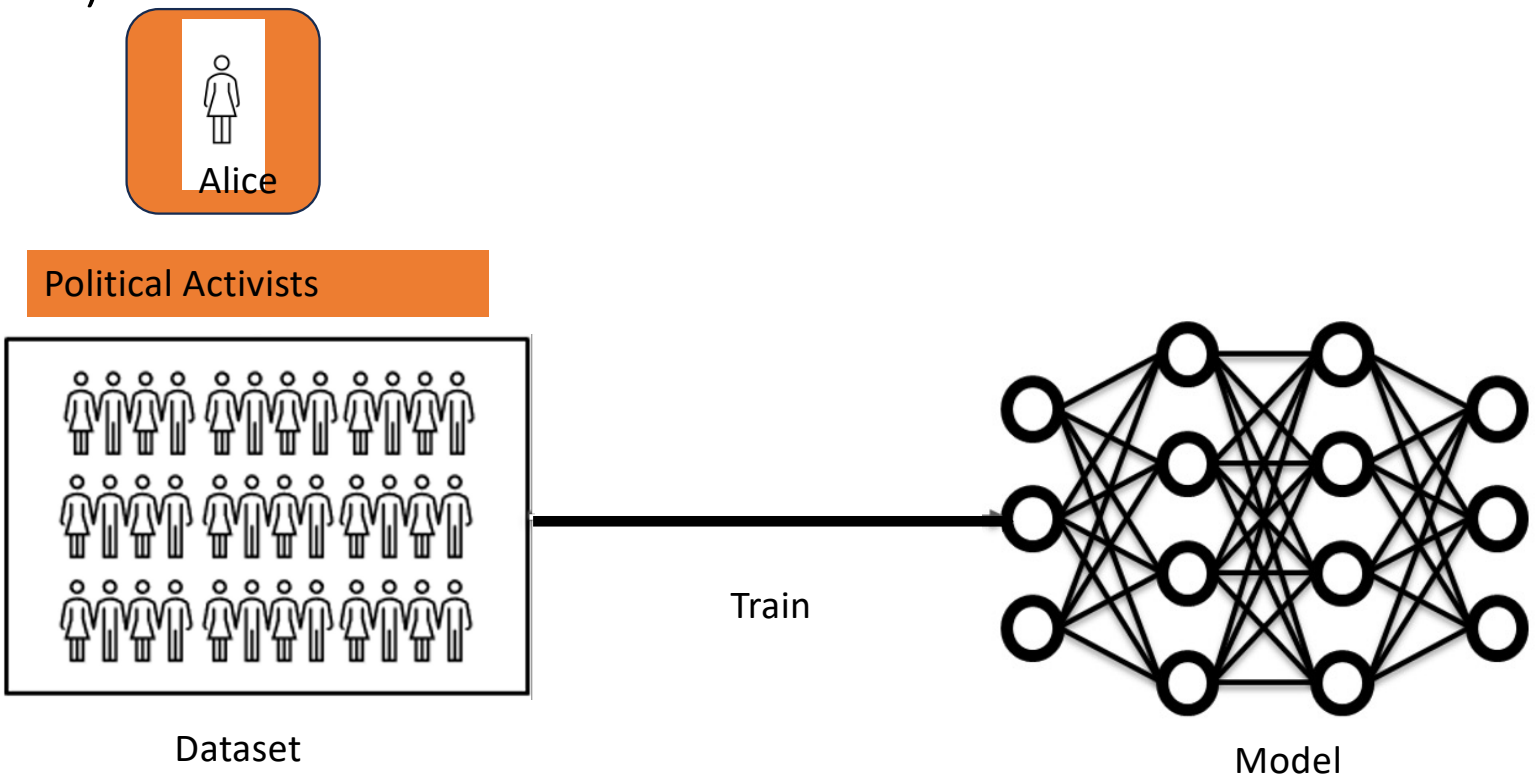
*Md Sakib Nizam Khan & **Sonja Buchegger** buc@kth.se*
KTH Royal Institute of Technology

Membership Inference: Is Alice(‘s record) part of the dataset that was used to train the model?



Membership Inference: what is the privacy threat?

Datasets usually have a purpose and common characteristics of the data records. (Also, right to be forgotten)



Membership Inference: why is it possible?

memory of training data, different behavior w.r.t. records in training dataset vs. other similar records.

Why is THAT possible?

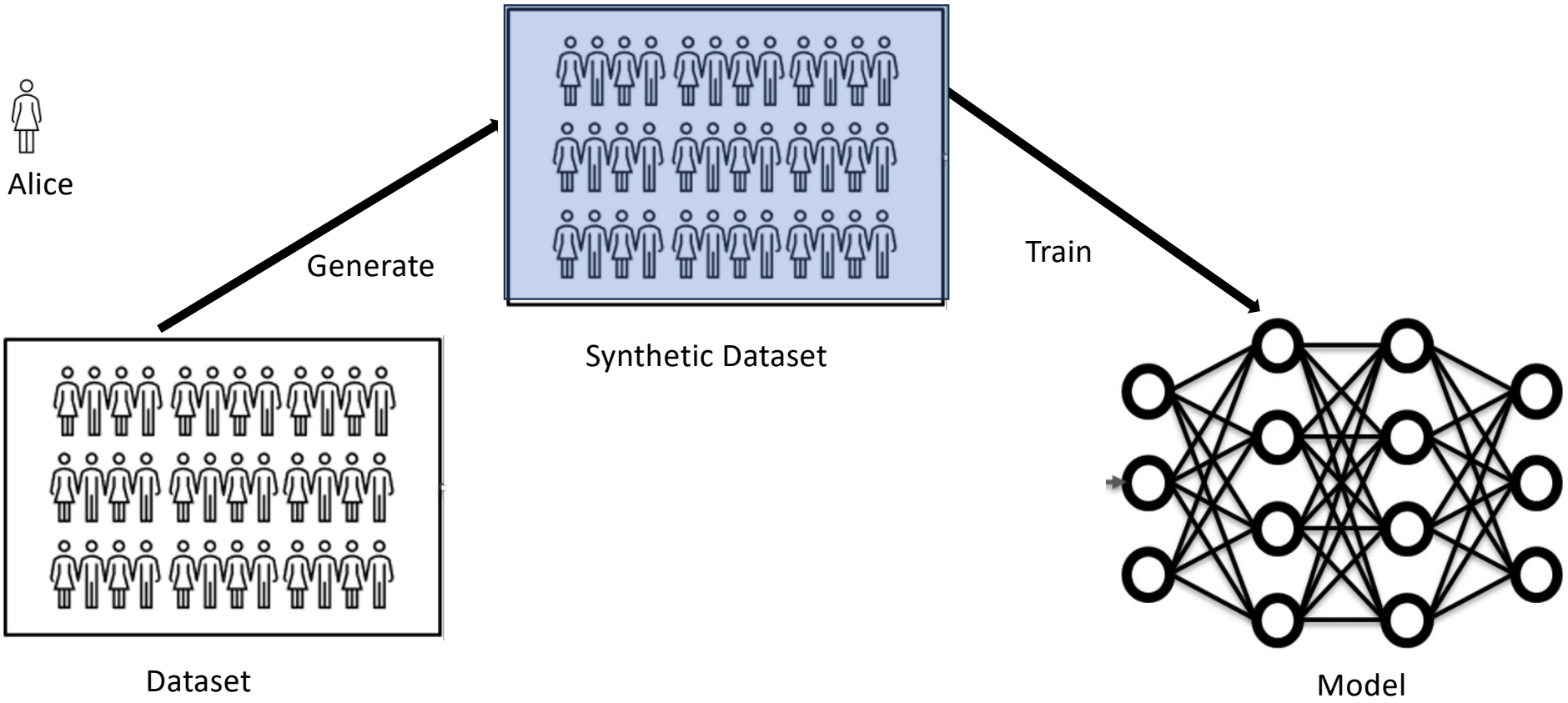
Mainly overfitting. Train accuracy significantly better than test accuracy.

Mitigation Techniques

- 2 approaches:
  1. Work on ML: reduce overfitting (i.e., increase generalization, e.g., by regularization)
  2. Work on data, response (e.g., Differential Privacy) but at cost of accuracy
     - Let's check synthetic data!

- Synthetic data
  - getting used as a tool for disclosure protection
  - better than original data straight off or "de-identified" (pseudonymous) data for linkability, but is it safe?
  - not much investigated yet in the context of MI attacks

Membership Inference with Synthetic Data: Is Alice('s record) part of the dataset that was used to generate the synthetic data that was used to train the model?



Synthetic Data vs. Membership Inference Attacks: what do we need to know?

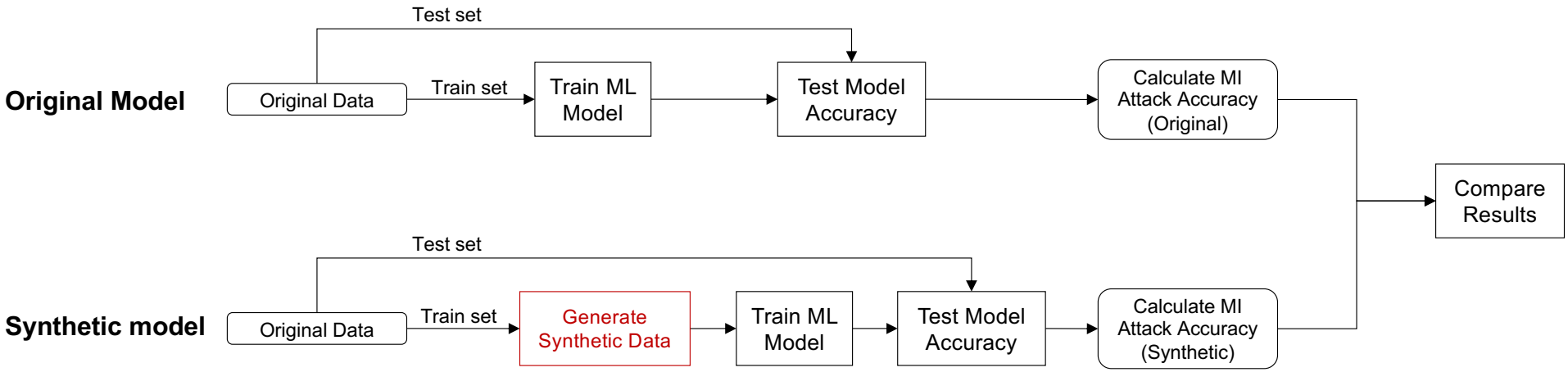
1. accuracy (if not accurate enough, no point): compare MI prediction accuracy between models trained on synthetic vs original data

2. security/privacy (does it mitigate MIA): compare MI attack accuracy between models trained on synthetic vs original data

3. what about overfitting (let's check extremes): compare the effect of overfitting on models trained on original vs synthetic data

4. And do these depend on how we generate the synthetic data? Check different synthetic data generation methods in terms of MI attack accuracy and prediction accuracy

Experimental Setup



Experimental Results – Attack Accuracy

Accuracy Comparison between Original and Synthetic Model

Dataset	Target Model	Train Accuracy	Test Accuracy	Attack Accuracy
Adult	Original	92.84	80.73	0.545
	Synthetic	93.057	83.26	0.5042
Polish	Original	97.38	55.83	0.66
	Synthetic	98.33	60.84	0.53
Location-30	Original	100	48.61	0.83
	Synthetic	100	64.44	0.54
Avila	Original	99.92	98.66	0.5108
	Synthetic	99.95	99.15	0.4991

Experimental Results – Synthesizers

Comparison of Synthetic Data Generation Methods

Dataset	Model	Train Accuracy	Test Accuracy	Generalization Error	Attack Accuracy
Polish	Original	97.38	55.83	41.55	0.66
	Synthpop CART+Catall	98.33	60.84	37.49	0.53
	Synthpop Parametric	92.85	55.96	36.89	0.51
	SDV CTGAN	99.88	61.26	38.62	0.505
	SDV Copula GAN	99.76	50.48	49.28	0.49
Location-30	Original	100	48.61	51.39	0.83
	Synthpop CART+Catall	100	64.44	35.56	0.54
	Synthpop Parametric	100	24.16	75.84	0.534
	SDV CTGAN	100	8.33	91.67	0.506
	SDV Copula GAN	100	4.44	95.56	0.491

Experimental Results – Overfitting

Effect of Overfitting

Dataset	Train Size	Model Type	Train Accuracy	Test Accuracy	Generalization Error	Attack Accuracy
Adult	7000	Original	92.84	80.73	12.11	0.545
		Synthetic	93.057	83.26	9.797	0.5042
	100	Original	98	69.99	28.01	0.63
		Synthetic	95.99	75.99	20	0.53
Location-30	840	Original	100	48.61	51.39	0.83
		Synthetic	100	64.44	35.56	0.54
	100	Original	100	28.45	71.55	1
		Synthetic	100	56	44	0.55

Synthetic Data vs. Membership Inference Attacks: what did we find out?

1. accuracy (if not accurate enough, no point): can achieve similar prediction accuracy as original data.

2. security/privacy (does it mitigate MIA): synthetic data reduces MIA accuracy (success) to near guessing.

3. what about overfitting (let's check extremes): still much reduced by synthetic training data.

4. And do these depend on how we generate the synthetic data? Generation methods vary in prediction accuracy but not much in MIA attack accuracy.

Remaining Questions

- what about other MIA attacks?
- what about other types of attribute inference?
- what about other types of data sets?
- what about uneven distribution of success probability?

- (philosophical) what does 1% of higher MIA success probability mean?