RSAConference2022
San Francisco & Digital | June 6 – 9

TRANSFORM

SESSION ID: **MLAI-M03**

Privacy and Compliance for AI – Open-Source Tools and Industry Perspective

Beat Buesser

Research Staff Member
IBM Research

Disclaimer

Presentations are intended for educational purposes only and do not replace independent professional judgment. Statements of fact and opinions expressed are those of the presenters individually and, unless expressly stated to the contrary, are not the opinion or position of RSA Conference LLC or any other co-sponsors. RSA Conference does not endorse or approve, and assumes no responsibility for, the content, accuracy or completeness of the information presented.

Attendees should note that sessions may be audio- or video-recorded and may be published in various media, including print, audio and video formats without further notice. The presentation template and any media capture are subject to copyright protection.

The Team

Beat Buesser
Research Staff Member
IBM Research – Dublin
beat.buesser@ie.ibm.com

Abigail Goldsteen
Research Staff Member
IBM Research – Haifa
abigailt@il.ibm.com

Ron Shmelkin
Research Staff Member
IBM Research – Haifa
ronsh@il.ibm.com

Agenda

- Motivation for AI Privacy
- Short Recap from RSAC 2021
- Securing AI Privacy
- Industry Perspective
  - Enterprise AI
  - Trustworthy AI
- Open-Source Tool for AI Privacy
  - AI Privacy Toolkit (APT)

Motivation

- The Era of AI
  - Large amounts of data generated
  - Omnipresent data collection
  - Better AI models
- Privacy Regulations:
  - HIPAA, GDPR, ePrivacy, Canada's Consumer Privacy Protection Act (CPPA), Singapore's Personal Data Protection Act (PDPA), etc...
  - Serious fines for non-compliance

Country	Date	Fine [€]	Controller/Processor	Quoted Art.
Filter Column		Filter Column	Filter Column	
FRANCE	2019-01-21	50,000,000	Google Inc.	Art. 13 GDPR, Art. 14 GDPR, Art. 6 GDPR, Art. 5 GDPR
GERMANY	2020-10-01	35,258,708	H&M Hennes & Mauritz Online Shop A.B. & Co. KG	Art. 5 GDPR, Art. 6 GDPR
ITALY	2020-01-15	27,800,000	TIM (telecommunications operator)	Art. 5 GDPR, Art. 6 GDPR, Art. 17 GDPR, Art. 21 GDPR, Art. 32 GDPR

https://www.enforcementtracker.com

With Big Data comes big responsibility !!!

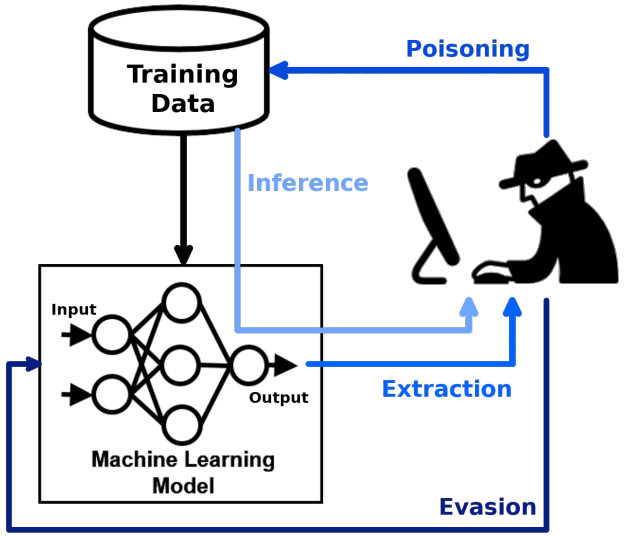
RSAConference2022

Recap RSAC 2021

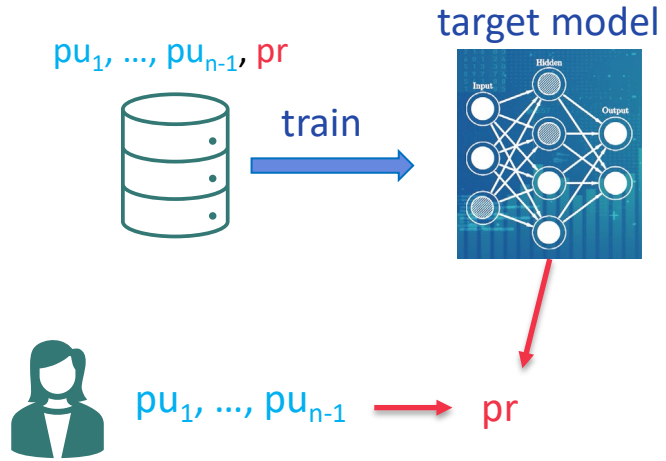
Evasion, Poisoning, Extraction, and Inference: Tools to Defend and Evaluate

Recap RSAC 2021

Adversarial Robustness Toolbox



1. Attribute Inference Attack



2. Membership Inference Attack

3. Defending with Differential Privacy

IBM / **differential-privacy-library**

https://www.rsaconference.com/library/Presentation/USA/2021/evasion-poisoning-extraction-and-inference-tools-to-defend-and-evaluate)

RSAConference2022

Privacy Preserving Technologies

Privacy Preserving Technologies

Differential Privacy

- Model dependent
- “invasive”

Anonymization

- Syntactic privacy
- Depends on available external information

Ensembles + student/teacher

- Challenging for large models
- Requires non-sensitive data

Encryption

- Much slower inference
- Requires key management

Defensive Approaches Without General Guarantees

Adversarial Learning

- Works only for single attack

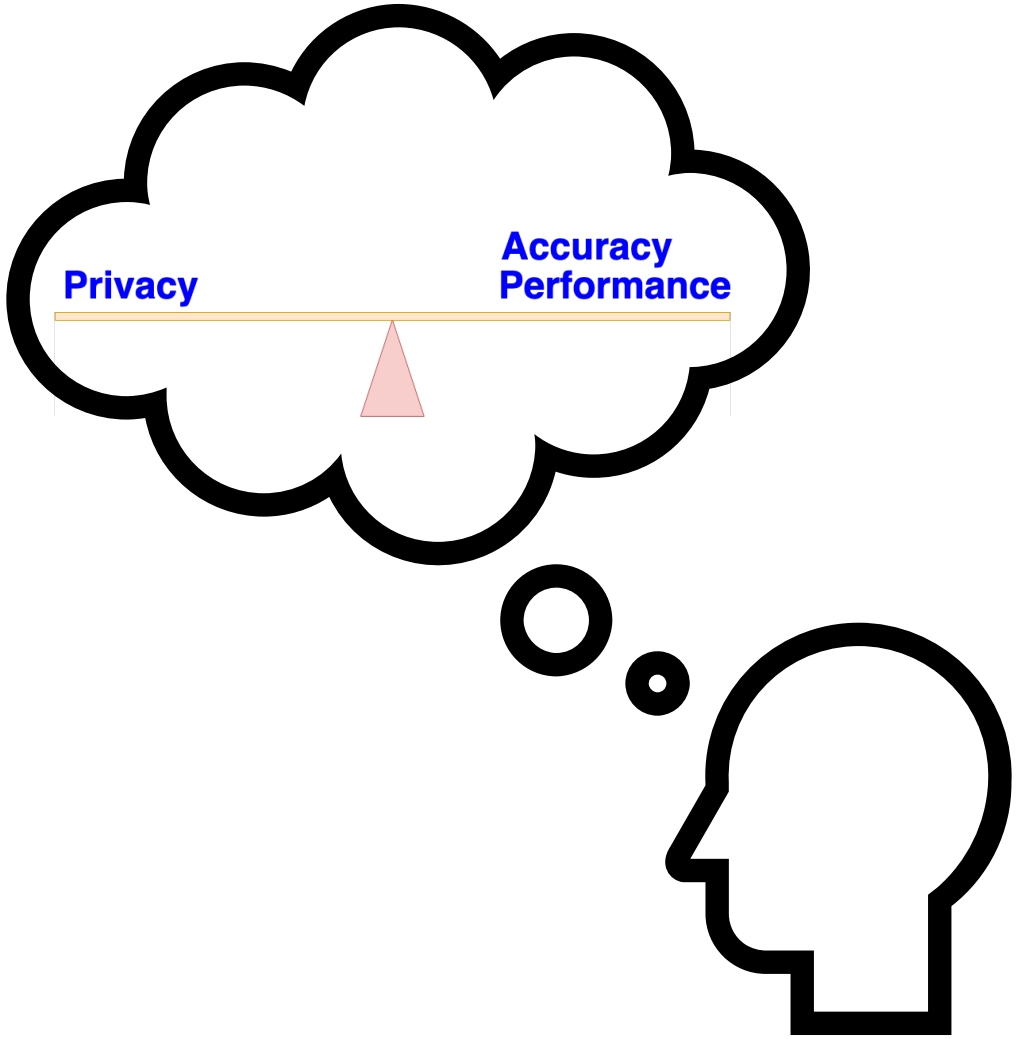
Regularization

- No guarantees for increased privacy

Confidence Masking

- Easily circumvented with WB or Label Only attacks

Trade-Off in Preserving Privacy

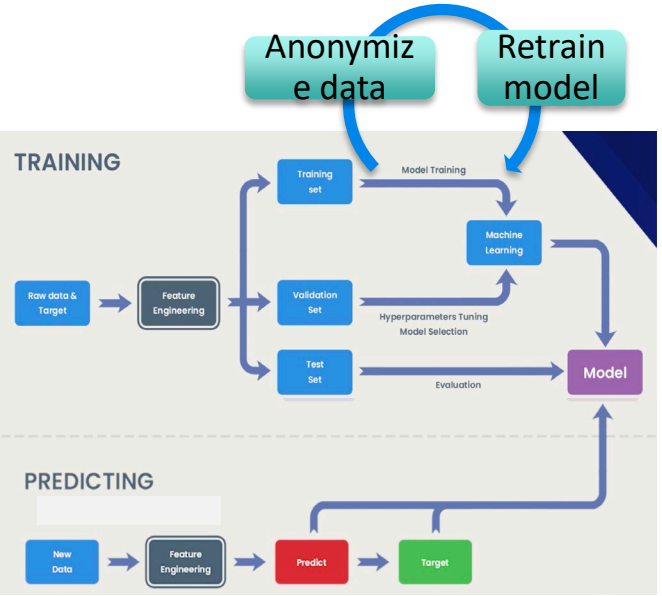


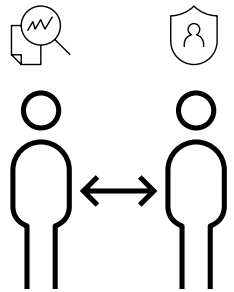
RSAConference2022

Industry Perspective

Defenses Should be Non-Disruptive

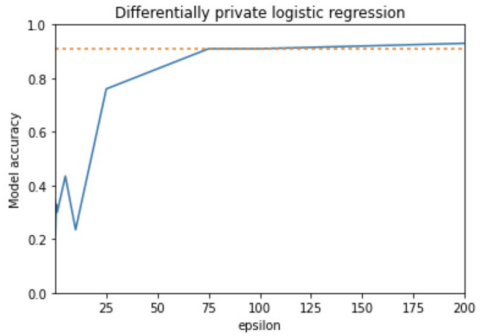
- Most organizations already have complex ML design and ops pipelines
- Solutions should integrate into these pipelines with minimal disruption
- Separate concerns: Data scientists are not experts in privacy and vice versa

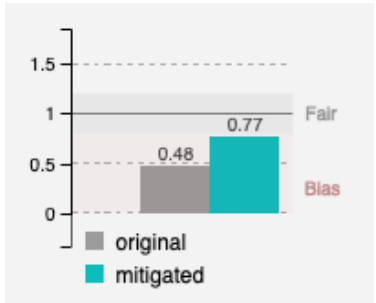




Design Choices – AI Privacy Toolkit

- "One-click" solutions easier to learn
- Good default parameters facilitate getting started
- Interpretable Visualisation preferred





First API

Two methods

All parameters must be supplied

```
class Anonymize:
def __init__(self, k: int, quasi_identifiers: Union[np.ndarray, list], categorical_features: Optional[list]=None)
def anonymize(self, x: Union[np.ndarray, pd.DataFrame], y: Union[np.ndarray, pd.DataFrame]) \
        -> Union[np.ndarray, pd.DataFrame]:
:return: An array containing the anonymized training dataset
```

Returns dataset (model training occurs outside)

Updated API

Single method

Supports all scikit-learn models/pipelines

Reasonable default parameters if None provided

```
class E2EAnonymize:
def anonymize_options(self, model: BaseEstimator,
                      X_train: Union[np.ndarray, pd.DataFrame],
                      y_train: Union[np.ndarray, pd.DataFrame],
                      X_test: Union[np.ndarray, pd.DataFrame],
                      y_test: Union[np.ndarray, pd.DataFrame],
                      k_values: List[int]=None,
                      quasi_identifiers: Optional[Union[np.ndarray, list]] = None)
:return: A list of models, each anonymized with a given k value, along with its accuracy score and additional
        classification metrics. A dict with the following structure:
        {
            'k': k value used,
            'model': anonymized model,
            'data': anonymized training data,
            'accuracy': accuracy score,
            'metrics': a dict containing the classification metrics returned by
                       sklearn.metrics.classification_report
        }
        The first model in the list is always the original model (k=1).
"""
```

Enables visualization tradeoffs

Returns trained model

Scalability and Performance

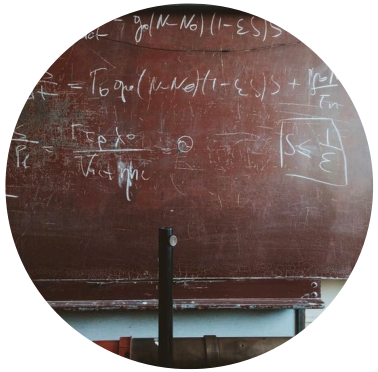
- Some privacy preserving methods are great for academic work but don’t scale to enterprise workloads
  - Thousands of models
  - Millions of records
  - Small teams
- Requires automation and efficient algorithms
  - sometimes resulting in sacrificing privacy and/or accuracy
- Prioritization of models based on their risk assessment

Trustworthy AI Requires Privacy, but also …

Principles of Trustworthy AI

Performance

Fairness



Explainability

Adversarial robustness

Privacy

Uncertainty

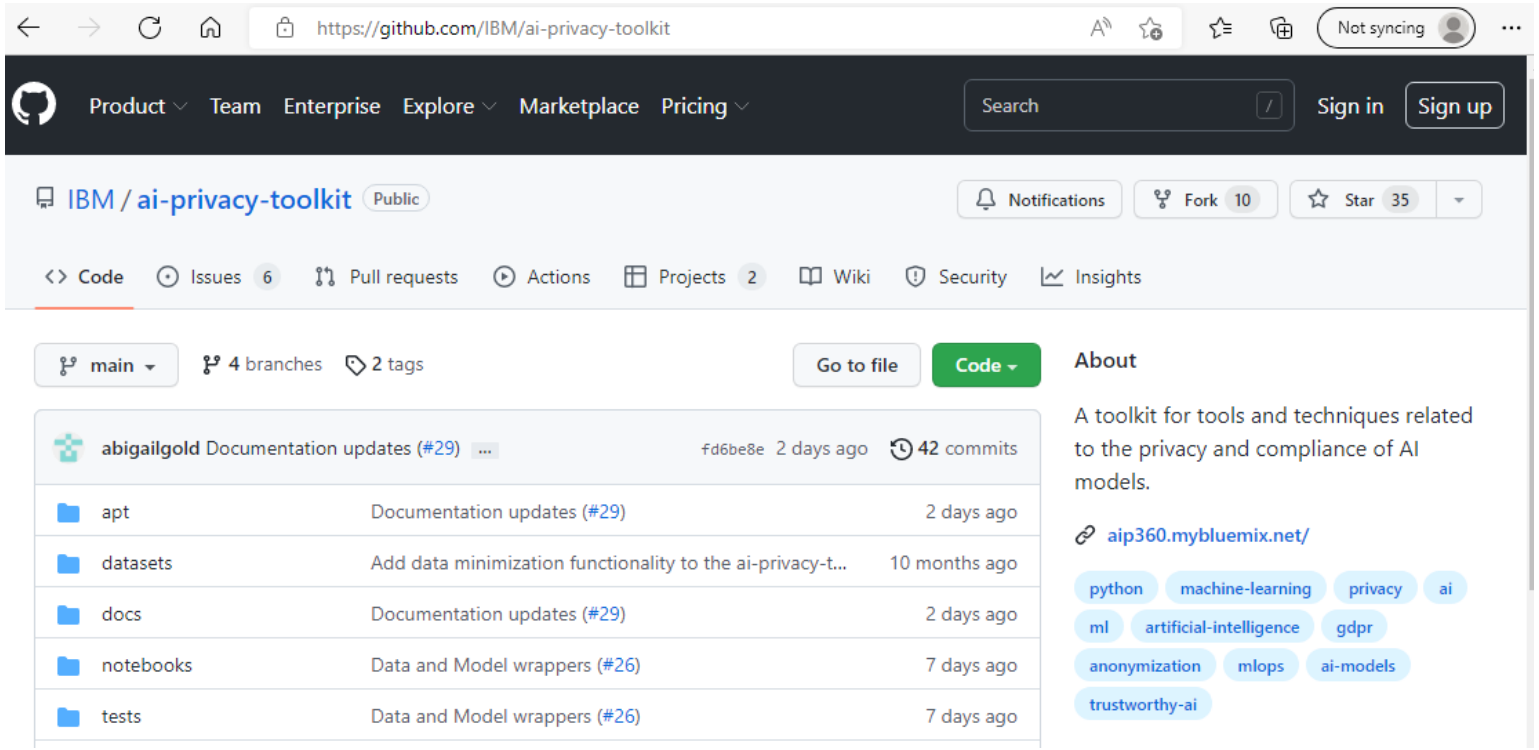
RSAConference2022

Open-Source Tools for AI Privacy

AI Privacy Toolkit (APT)

Open-Source Tool: AI Privacy Toolkit (APT)

https://github.com/IBM/ai-privacy-toolkit



- Current Modules (next slides)
  - Model Anonymization
  - Data Minimization
- Future Modules
  - Right-To-Be-Forgotten tools
  - Privacy Risk Assessment

Model Anonymization vs Data Anonymization

Data Anonymization

Original Data

Zip	Gender	Age	Account	Fee
47919	Male	35	Personal	5.5
47902	Male	34	Personal	5.75
47918	Female	37	Personal	4.9
47919	Female	39	Joint	4.5
47904	Female	30	Joint	5.1
47909	Male	31	Joint	4.8

Deployment

Anonymized Data

Zip	Gender	Age	Account	Fee
4791*	Person	[35-39]	Personal	5.5
4790*	Person	[30-34]	Personal	5.75
4791*	Person	[35-39]	Personal	4.9
4791*	Person	[35-39]	Joint	4.5
4790*	Person	[30-34]	Joint	5.1
4790*	Person	[30-34]	Joint	4.8

Model Anonymization

Original Data

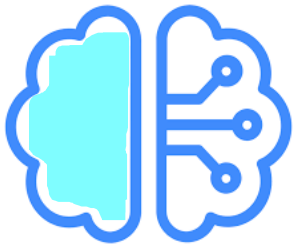
Zip	Gender	Age	Account	Fee
47919	Male	35	Personal	5.5
47902	Male	34	Personal	5.75
47918	Female	37	Personal	4.9
47919	Female	39	Joint	4.5
47904	Female	30	Joint	5.1
47909	Male	31	Joint	4.8

Original Model

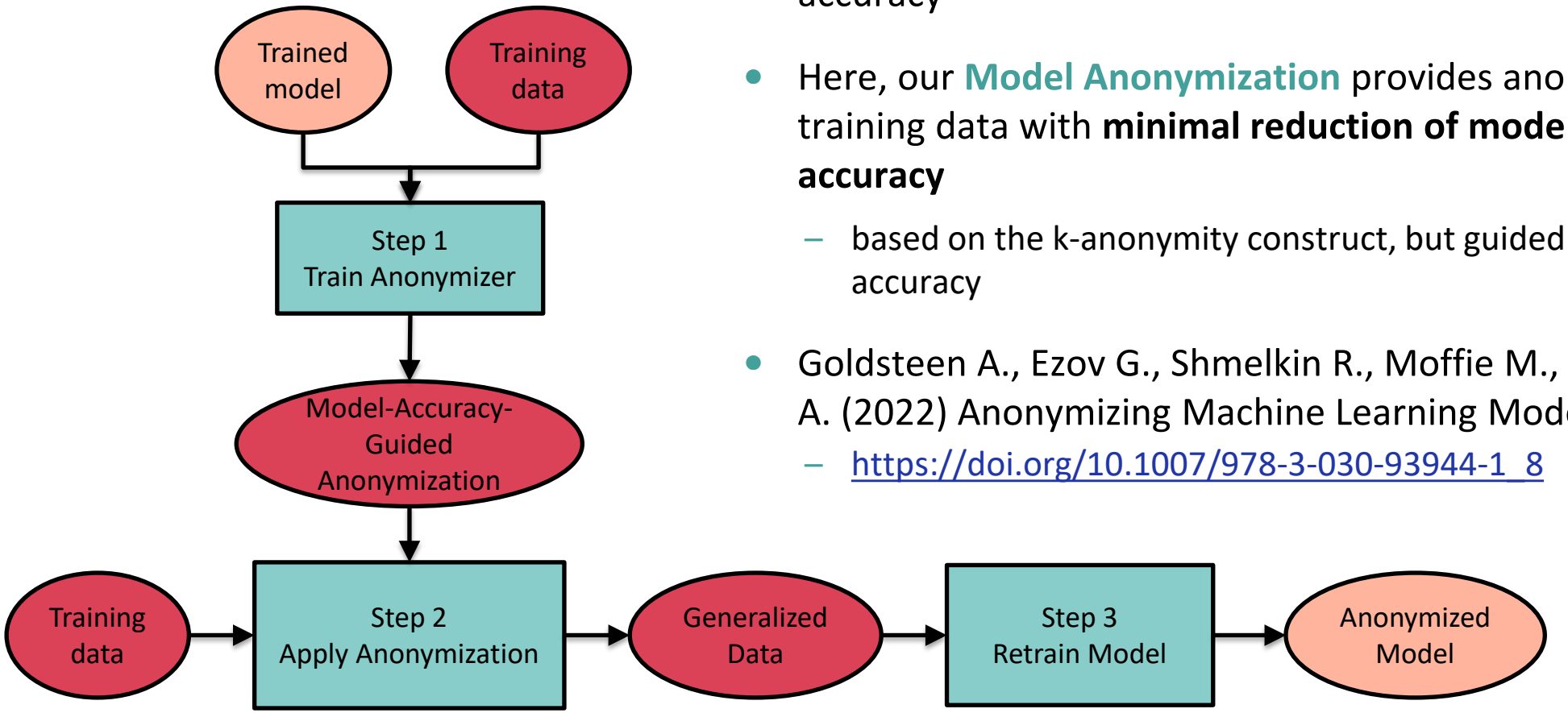
Anonymized Data

Zip	Gender	Age	Account	Fee
4791*	Person	[35-39]	Personal	5.5
4790*	Person	[30-34]	Personal	5.75
4791*	Person	[35-39]	Personal	4.9
4791*	Person	[35-39]	Joint	4.5
4790*	Person	[30-34]	Joint	5.1
4790*	Person	[30-34]	Joint	4.8

Anonymized Model



AI Privacy Toolkit: Model Anonymization



- Data Anonymization often results in low ML model accuracy
- Here, our **Model Anonymization** provides anonymized training data with **minimal reduction of model accuracy**
  - based on the k-anonymity construct, but guided by model accuracy
- Goldsteen A., Ezov G., Shmelkin R., Moffie M., Farkash A. (2022) Anonymizing Machine Learning Models
  - https://doi.org/10.1007/978-3-030-93944-1_8

Differential Privacy vs Model Anonymization

Pros

Differential Privacy

- Actual mathematical **privacy guarantee**
- Works well for **high-dimensional data**, including images
- **Future proof** because of guarantees

Model Anonymization

- **Is external to the training process**, which does not need to change
- Single, **model-agnostic** algorithm
- Can be applied to existing models after the fact. Retraining is needed, but algorithm/ architecture/hyperparams **can be reused**.

Cons

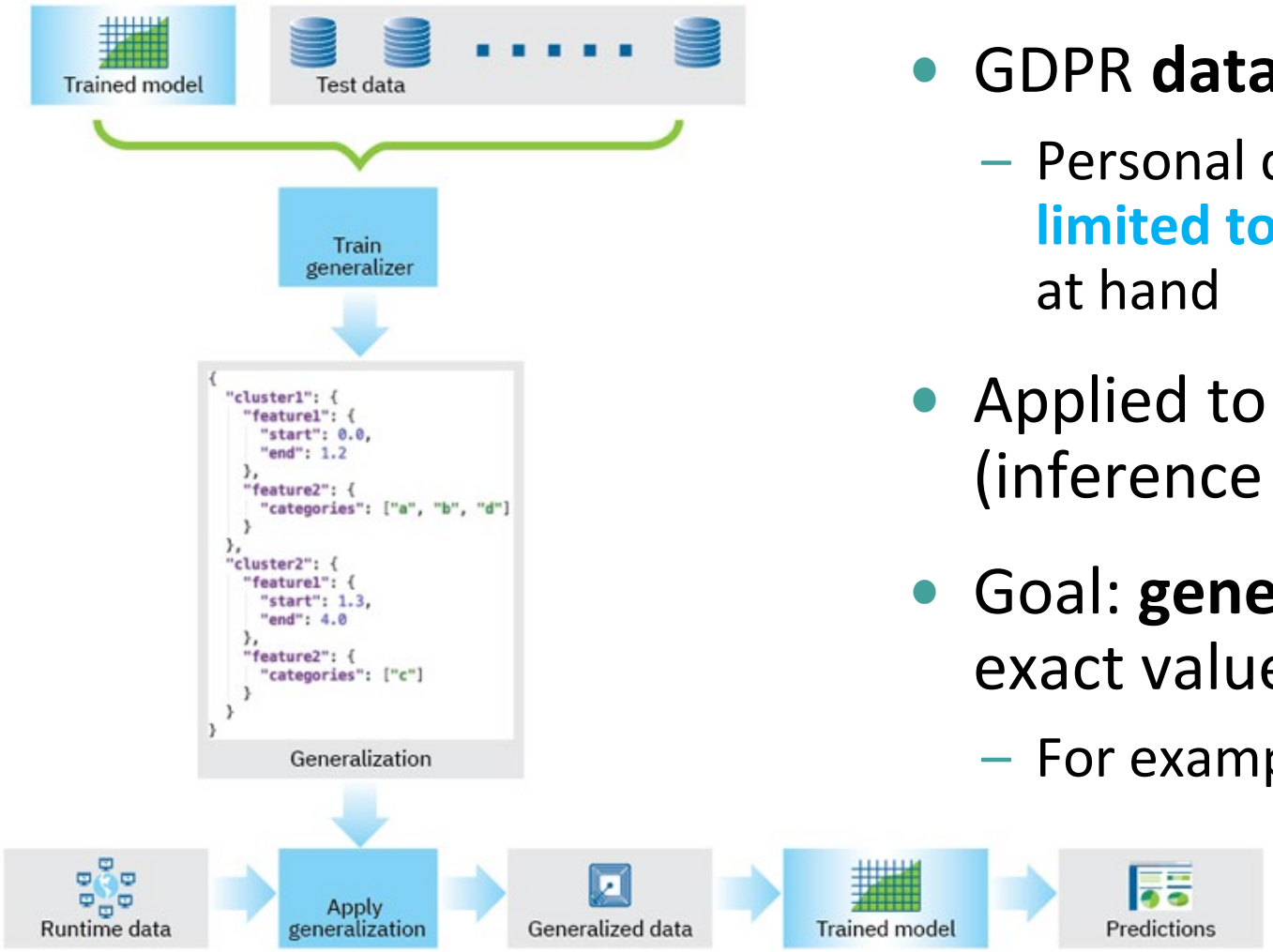
Differential Privacy

- **Invasive** – replaces training algorithm
- **Different implementation** for different learning algorithms/architectures
- May be more **difficult to combine with other Trustworthy AI aspects** (that may require special model impl.), e.g., bias, explainability…
- Requires the model trainer (data scientist) to be aware of privacy needs

Model Anonymization

- **Syntactic** privacy
- Selection of **quasi-identifiers** may affect re-identification risk
- Works only for **tabular**, relatively low-dimensional data

AI Privacy Toolkit: Data Minimization 1/2

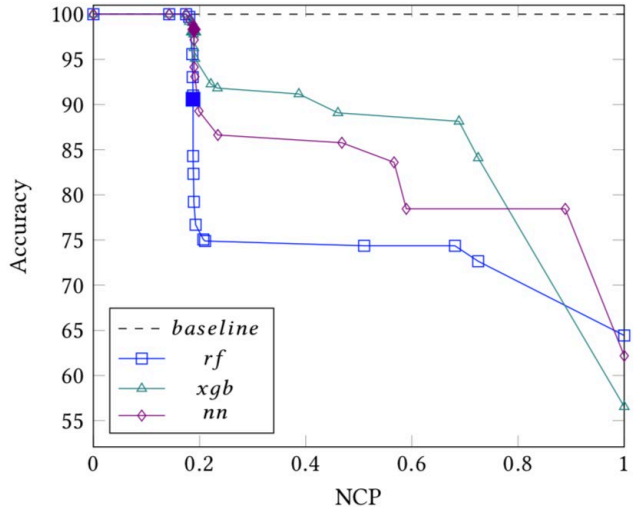


- GDPR **data minimization** clause
  - Personal data shall be adequate, relevant and **limited to what is necessary** for the purposes at hand
- Applied to **new data** collected for analysis (inference phase)
- Goal: **generalize features** by replacing exact values with groups/ranges
  - For example, replace Age: 38 with Age: 35-40

AI Privacy Toolkit: Data Minimization 2/2

- **Normalized Certainty Penalty (NCP):** a metric for information loss or how well the data is generalized
  - Larger NCP means fewer specific data
- **Goal:** Maximize NCP for desired accuracy
- **Table** shows how 2% decrease in accuracy allows generalization of 3 sensitive features

Accuracy vs NCP



Feature	2% relative accuracy loss	No accuracy loss
Marital status	Not needed	Not needed
Happiness	[Pretty happy, Not too happy, Other], [Very happy]	Same as 2%
Race	[Black, Other], [White]	Not generalized
Work status	[Temp not working, Unemployed - laid off, School, Other], [Working fulltime], [Keeping house], [Retired], [Working parttime]	Not generalized
Age	54 ranges representing values 0-89	Not generalized
Children	Not generalized	Not generalized
X rated	Not generalized	Not generalized
NCP value:	*0.189703*	*0.174658*

Table 3: Example of resulting generalizations for the GSS dataset

Paper:
Goldsteen, A., Ezov, G., Shmelkin, R. et al. Data minimization for GDPR compliance in machine learning models. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00095-8

AI Privacy Toolkit (APT) in Action – 1/2

- Run minimization with APT's GeneralizeToRepresentative

```
from apt.minimization import GeneralizeToRepresentative

# Target accuracy of minimized model set to 0.998
minimizer = GeneralizeToRepresentative(model, target_accuracy=0.99)

# Create predictions of representative model
x_train_predictions = model.predict(X_generalizer_train)

# Fitting the APT  minimizer
minimizer.fit(dataset=ArrayDataset(X_generalizer_train, x_train_predictions))

# Apply the APT  minimizer to transform additional data
transformed = minimizer.transform(dataset=ArrayDataset(x_test))
```

AI Privacy Toolkit (APT) in Action – 2/2

- Check generalized features

print(minimizer.generalizations)

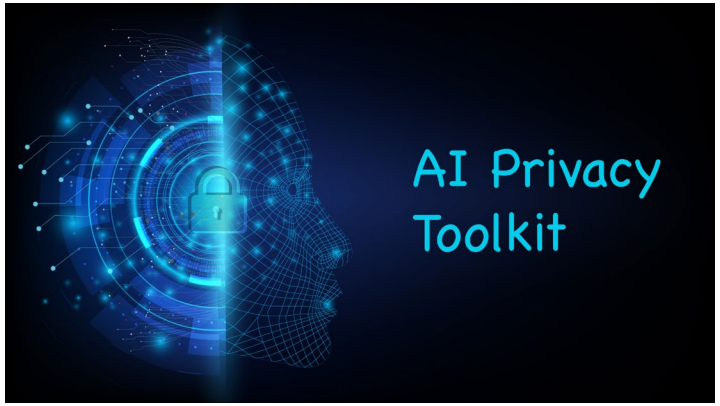
Initial accuracy of model on generalized data, relative to original model predictions (base generalization derived from tree, before improvements): 0.936540
Improving accuracy
feature to remove: 2
Removed feature: 2, new relative accuracy: 0.935261
feature to remove: 4
Removed feature: 4, new relative accuracy: 0.946776
feature to remove: 0
Removed feature: 0, new relative accuracy: 0.972876
feature to remove: 1
Removed feature: 1, new relative accuracy: 0.992835
Accuracy on minimized data: 0.8192845079072624
{'ranges': {'3': [569.0, 782.0, 870.0, 870.5, 938.0, 1016.5, 1311.5, 1457.0, 1494.5, 1596.0, 1629.5, 1684.0, 1805.0, 1859.0, 1867.5, 1881.5, 1938.0, 1978.5, 2119.0, 2210.0, 2218.0, 2244.5, 2298.5, 2443.5]}, 'categories': {}, 'untouched': ['2', '1', '0', '4']}
This time we were able to generalize one feature, feature number 3 (capital-loss).

- Complete example of data minimization with AI Privacy Toolkit
  - https://github.com/IBM/ai-privacy-toolkit/blob/main/notebooks/minimization_adult.ipynb

AI Privacy Toolkit : Meet the Developers

- Slack
  - Announcements, Q&A, etc.
  - https://aif360.slack.com/messages/C02HKUD0JG6
- GitHub
  - Issues, Bug Reports, Discussions, **Contributions!**
  - https://github.com/IBM/ai-privacy-toolkit

- Maintainer and Leading Core Developer
  - Abigail Goldsteen
  - abigailt@il.ibm.com



Apply What You Have Learned Today

- Next week you should:
  - **Locate** all AI/ML models in your organization and identify those trained on personal data
  - Create **awareness** of potential privacy vulnerabilities and possible mitigations
- In the first three months following this presentation you should:
  - **Identify** which **mitigation strategies** best suit your use case and models
    - May be more than one
  - Start **learning how to use** the appropriate toolkits
  - **Design an overall solution** encompassing the needs and resources available in your company
- Within six months you should:
  - Start **protecting your models** used in production
  - **Adhere to** different privacy **regulation** requirements
  - **Monitor, assess** and **adapt** the suitability of your solution over time

Acknowledgement

Parts of the work mentioned here are being developed within the EU funded H2020 project iToBoS:
https://itobos.eu/index.php

Thank you!


data
model
learning
machine
privacy
enables
minimization
training
certain
risk
personal
anonymization
trained
used
accuracy
models
also
tool
algorithms
information
predictions
regulations
effect
inference
specific
risks
using
forgetting
minimized
anonymize
dealing
may
can
depiction
techniques
tailored
check
process
published
individuals
relevant
attacks
neural
new
encoded
services
directing
collected
reduce
set
derive
video
method
perform
level
post
read
however
allows
reveal
whose
blog
train
difficult
sensitive