WHITEPAPER

Differential Privacy in Responsible AI

With the ubiquitous use of Artificial Intelligence and Machine Learning (AIML) based systems for decision-making, there are increasing incidents of bias and discrimination[1]. These rising concerns around the societal implications of a bias call for developing a disciplined approach called Responsible AI, which seeks to enforce principles such as transparency, fairness, and explainability. **One significant consideration of this framework is privacy and the various techniques that can be employed to address it.**

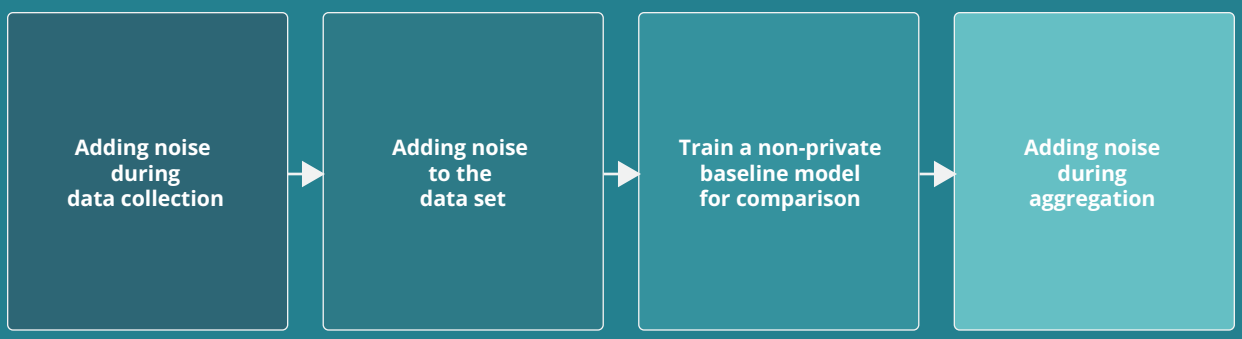
AI-based systems need large volumes of data to train and test the ML models. Datasets may contain Personally Identifiable Information (PII), such as names, SSNs, and so on, that require careful handling. Data privacy breaches[2] cost enterprises financially and cause reputational damage. For example, cybercriminals (acting alone or belonging to a criminal syndicate) getting access to health records can put lives at risk.

Governments across the globe, such as the European Union (GDPR), Brazil (BCRF), Japan (APPI), and the USA (CPRA), have laws and regulations to protect the privacy of their citizens. Data privacy laws govern how data should be collected, stored, and shared with third parties.

Privacy Enhancing Techniques: Overview

Synthetic data	Federated learning	Pseudonymization
Artificially generated data for a given use case instead of the data captured directly.	The data owner allows the system to use it for insights without sharing the actual data.	Artificial identifiers replace PTI fields within the dataset.
Homomorphic encryption	**Generative Adversarial Networks (GANs)**	**Differential privacy**
Sensitive data is converted to Ciphertext (plain text transformed using an encryption algorithm).	Competing neural networks attempt to become more accurate than others.	A degree of randomization is added to the dataset to maintain individuals' privacy. Since the amount of noise added gets controlled, generated aggregate insights are still accurate.

Options To Apply Differential Privacy to a Machine Learning Workflow



A key question in selecting the best approach is which stakeholders should be allowed to access the data in an unprotected state.

Differential Privacy in Data

One can control the randomness or noise level by adding a “privacy loss” parameter (Ɛ) to a dataset, thus maintaining the data privacy.

Differential privacy in data can be implemented:

Locally – Database noise is added before storing the data in the central repository.

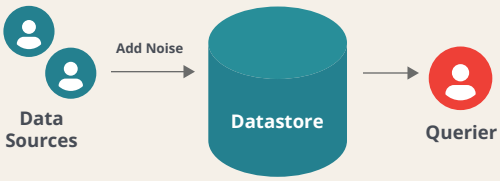


Figure 1 Local Privacy

Globally - Raw data is stored directly in the central datastore without adding any noise. The noise gets added when a user queries data.

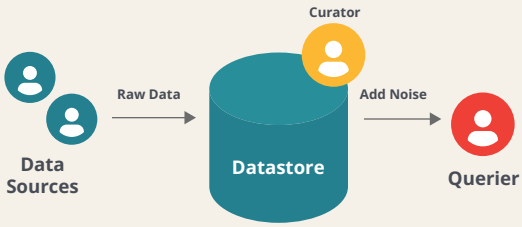


Figure 2 Global Privacy

Impact of applying differential privacy to data (statistical noise)

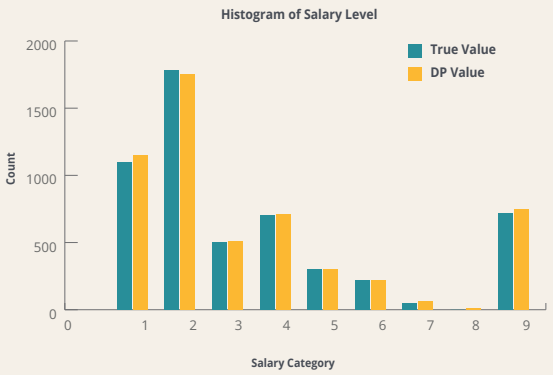


Figure 3.1

Comparison of applying differential privacy to data

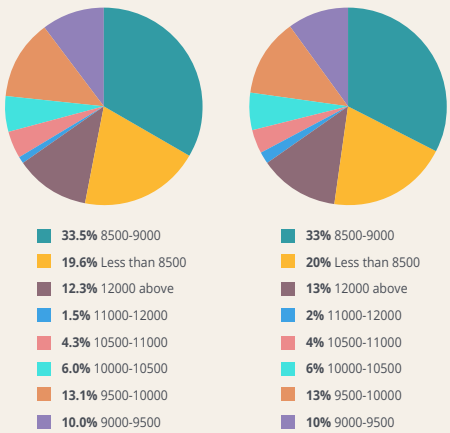


Figure 3.2

Figures 3.1 and 3.2 demonstrate how noised samples differ from the original data. Noised values are generated by different privacy budgets (controlled by the parameter Ɛ). There are almost no observable deviations between the histograms.

Differential Privacy in ML Algorithms:

In this case, whether any individual’s data is included in the actual dataset is not revealed.

ML models can be made differentially private by the following means:

- BY ADDING NOISE TO MODEL’S OBJECTIVE FUNCTION
- BY ADDING NOISE TO MODEL OUTPUT
- BY ADDING NOISE TO THE WEIGHTS OF THE OUTPUT OF MODEL’S OBJECTIVE FUNCTION

Case Study

Customer:
Leading bank in the United Kingdom

Problem:
Application of machine learning in home loan lending with personal data protection. Predictive analytics can be effectively employed to minimize human intervention and automate decision-making. The dataset used for training the ML models includes personal data. Hence strict measures to protect user privacy are needed.

Solution:
Fractal applied differentially private algorithms to protect the user's identity while using the data for analytics. Employed Random Forest Classifier (RFC) that was made differentially private by adding noise (using Exponential Mechanism) to the prediction probability of labels.

Methodology

- Use differentially private data with normal model and vice versa
- Use differentially private data with a differentially private model
- Test the model for different epsilon values
- **Implementation details:** Open-Source library Diffprivlib from IBM; scikit-learn; Pandas, numpy and Matplotlib; Python v3.9
- **Model performance metrics:**
  - **Accuracy:** 76.72% (DP Data + DP Model); 76.36% (Raw Data + DP Model); 79.42% (DP Data + Normal Model).

Note: Not applying differential privacy to data and/or model accuracy is **79.98%**

Prediction probability with different 'ε'

The charts below illustrate how a differentially private model predicts the outcome for the same customer. While iterating a DP model multiple times with the same epsilon, probabilities fluctuate slightly, but the outcome (prediction) is the same.

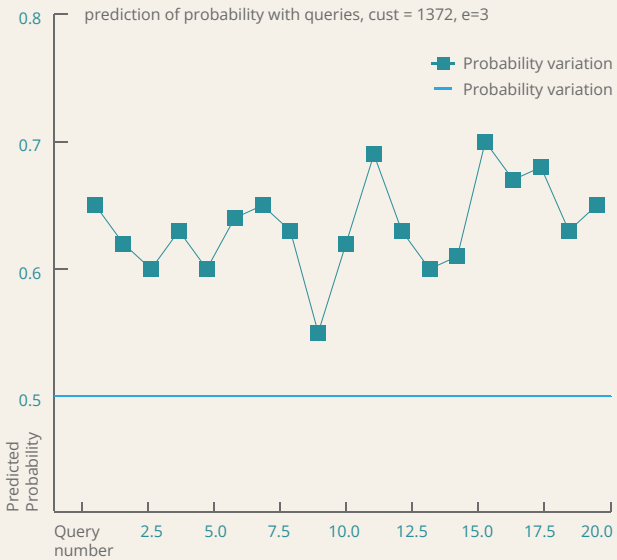


Figure 5: DP model predictions epsilon=3

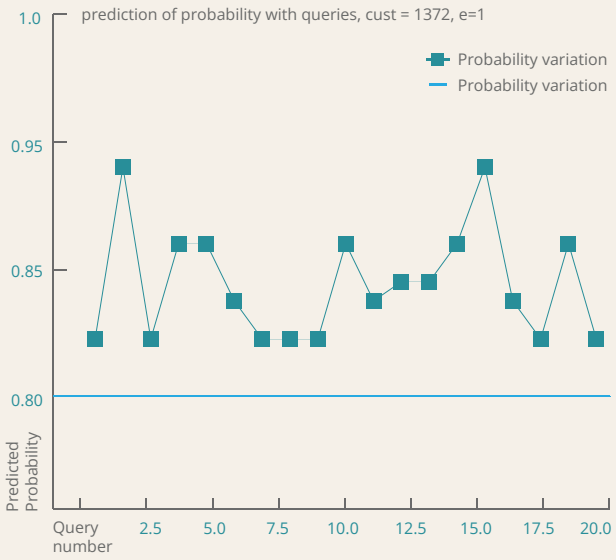


Figure 6: DP model predictions epsilon=1

Differentially Private Algorithms

For an algorithm to be differentially private, its output should not change even if a data point is excluded from the dataset. This provides confidence that even if personally identifiable information is present within the dataset, it would not be visible to the outside world.
DP algorithms are resistant to adaptive attacks since the noise introduced into the dataset makes the data imprecise.

DP Algorithms Models and Explanation

MODELS	How is a model made differentially private?	Advantages	Disdvantages
Tree based algorithms	Exponential Mechanism adds noise to the prediction probability of labels regarding the most frequent label.	High accuracy. Good starting point to solve the problem. Flexible and suitable for a variety of different data. Fast to execute. Easy to use. Can model missing values. High performing.	Slow at training. Overfitting. Not suitable for small samples. Small changes in training data change the model. Occasionally too simple for very complex problems.
Unsupervised Learnings	Noise is added to the averages of centroids calculated where noise is taken from a Laplace distribution, which is a function of the number of centroids, epsilon, sensitivity, and the number of data partitions.	Easy to learn, configure and maintain. Simple to implement. Aims toward spherical clusters (for some applications might be a con). Handles large datasets.	Inconsistent (depends on the selection of the initial seed). The "K" input requires specifying the size of the clusters. Sensitive for outliers, especially if they were used as initial seeds
Linear Models	Laplacian noise is added to the coefficients of the objective function. Noise is added to the coefficients of each feature where noise is proportional to the exponential function.	Easy to implement, the theory is simple, low computational power compared to other algorithms. Easy to interpret coefficients for analysis. Perfect for linearly separable datasets. Inclined to overfit, but can be avoided using dimensionality reduction, cross-validation, and regularization techniques.	Prone to underfitting. Sensitive to outliers. Assumes that the data is independent.

Benefits

- Resistant to privacy attacks.
- Compositional. One can add the privacy loss for multiple analyses on the same dataset.

Drawbacks

- Not suitable for small datasets.
- Repeated application of the algorithm increases privacy loss.
- Reduces accuracy with a low privacy budget.

How organizations like Apple and Google are implementing DP3

Apple uses local differential privacy, computed on individual devices before being collected by the central server.

Google shares random samples of aggregated and anonymized historical traffic statistics that are differentially privatized by noise before data transmission.

Microsoft has developed local DP mechanisms for collecting counter data for their basic analytical tasks.

Conclusion

Data privacy is often overlooked when creating a machine learning algorithm. With the ubiquitous data collection around us, extracting private information from a dataset that does not have privacy built into it is now easier than ever. Differential privacy allows organizations to customize the privacy level and leads attackers to access only partially correct data.

References

[1] Real-life Examples of Discriminating Artificial Intelligence | by Terence Shin | Towards Data Science

[2] List of Data Breaches and Cyber Attacks in May 2022 | 49.8 Million Records (itgovernance.co.uk)

[3] Book: Responsible AI by Sray Agarwal and Shashin Mishra

What is Differential Privacy? | Georgian Partners

Privacy-preserving logistic regression Kamalika Chaudhuri Information Theory and Applications University of California, San Diego

Microsoft SmartNoise Differential Privacy Machine Learning Case Studies

https://research.aimultiple.com/differential-privacy/

What is Differential Privacy and How does it Work? | Analytics Steps

We believe complex problems need to be looked at through multiple lenses simultaneously to be grasped. With the new lens new dimensions emerge, thus making complexity more evident and solvable.

How is Fractal Dimension set up to do it?

We identify complex and unstructured problem themes in the industry that are relevant. We invest in building expertise and a dimensionalized point of view around it.

We engage clients via 'slow-thinking' workshops and co-creation jams to curate our perspective for their problem. We invest in architecting an end-to-end state-change program.

We partner with client teams at Fractal to deploy cross-functional solutions and support them in helping clients realize value ROI.

Want to find out more on how our approach can help your business? Reach out today at dimension@fractal.ai

Our experts

Sray Agarwal
Principal Consultant, Dimension

Olena Fylymonova
Consultant, Dimension

Balachandra Kamat
Principal Manager, Dimension

Enable better decisions with Fractal

Fractal is one of the most prominent providers of Artificial Intelligence to Fortune 500® companies. Fractal's vision is to power every human decision in the enterprise, and bring AI, engineering, and design to help the world's most admired companies.

Fractal's businesses include Crux Intelligence (AI driven business intelligence), Eugenie.ai (AI for sustainability), Asper.ai (AI for revenue growth management) and Senseforth.ai (conversational AI for sales and customer service). Fractal incubated Qure.ai, a leading player in healthcare AI for detecting Tuberculosis and Lung cancer.

Fractal currently has 4000+ employees across 16 global locations, including the United States, UK, Ukraine, India, Singapore, and Australia. Fractal has been recognized as 'Great Workplace' and 'India's Best Workplaces for Women' in the top 100 (large) category by The Great Place to Work® Institute; featured as a leader in Customer Analytics Service Providers Wave™ 2021, Computer Vision Consultancies Wave™ 2020 & Specialized Insights Service Providers Wave™ 2020 by Forrester Research Inc., a leader in Analytics & AI Services Specialists Peak Matrix 2022 by Everest Group and recognized as an 'Honorable Vendor' in 2022 Magic Quadrant™ for data & analytics by Gartner Inc. For more information, visit fractal.ai

Corporate Headquarters
Suite 76J,
One World Trade Center, New York,
NY 10007

Get in touch