A brief introduction to Differential Privacy

Overview of Data Privacy Challenges

- **Data Explosion**: Massive amounts of data driving research and innovation
- **Privacy Risks**: Cybersecurity threats and data misuse
- **Regulatory Compliance**: Importance of adhering to GDPR and other regulations
- **Public Concerns**: Growing demand for stronger privacy measures
- **ENCRYPT Project**: Aims to enhance data security and privacy across federated data spaces.

Differential Privacy

- Differential Privacy
  - Ensures data privacy and anonymity by adding carefully calibrated noise to query responses
  - Prevents the identification of individual records
  - Still allows for accurate aggregate analysis
- Differential Privacy Addition of Noise
  - Defined by ε (epsilon) parameter
  - Smaller amounts of ε
    - Introduce **greater amounts** of noise, **stronger privacy** guarantees, may lead to **less accurate** data analytics
- Addition of noise
  - Drawn from a Laplace distribution or Gaussian distribution or other mechanisms

encrypt

Differential Privacy Variants

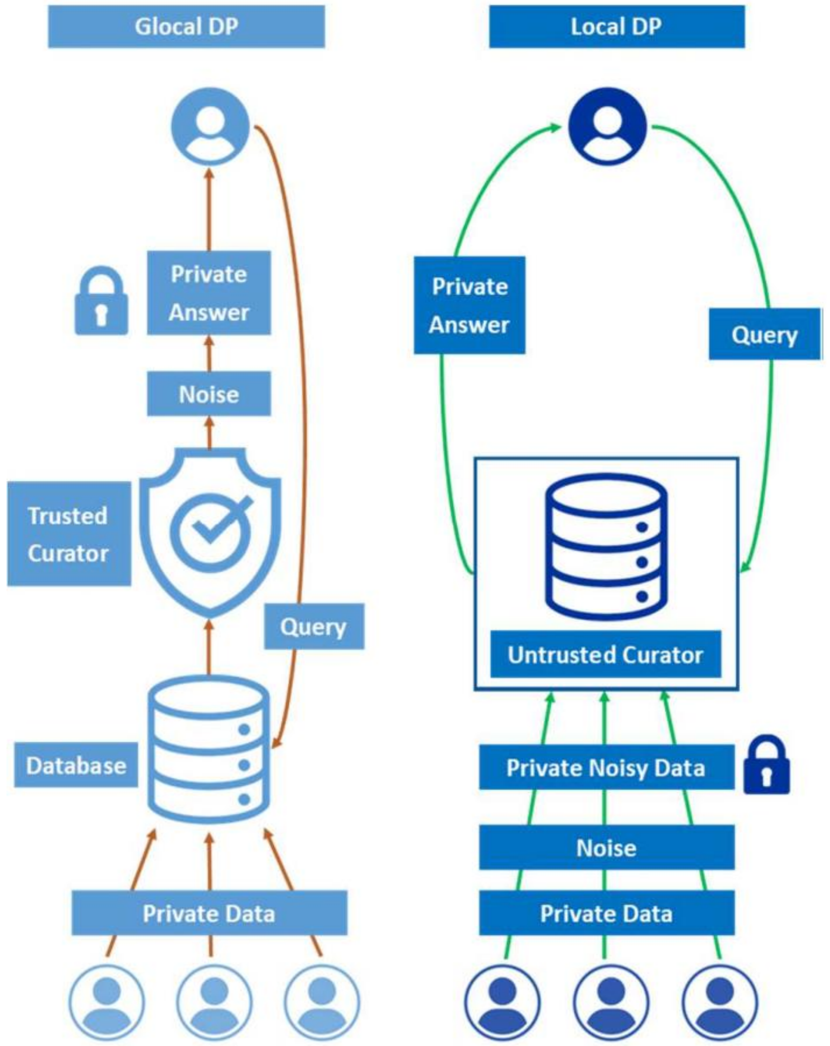
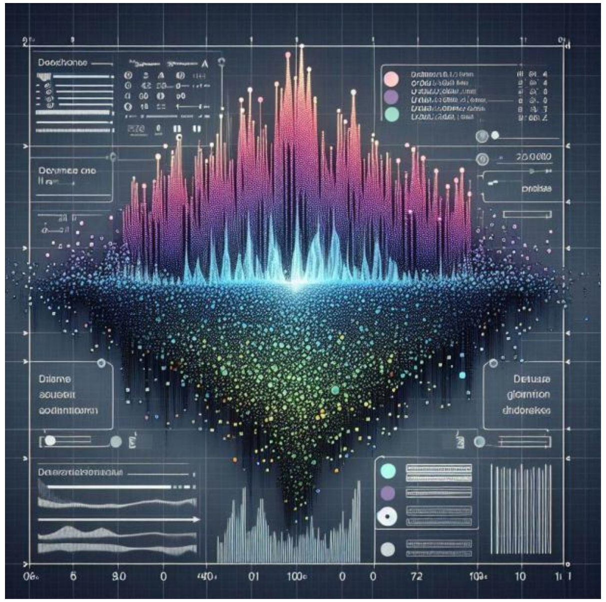


Figure 1: Global and Local Differential Privacy

- Local vs Global DP
  - ✓ Depending on whether we have a trusted curator or not
- Local Differential Privacy
  - Extends privacy guarantee by introducing noise at the source (individual data points)
  - Data contributors add noise (epsilon amount) to data before sharing it
  - Ensures privacy even when a central curator cannot be fully trusted
- Both prioritize privacy without sacrificing utility – useful for data analysis

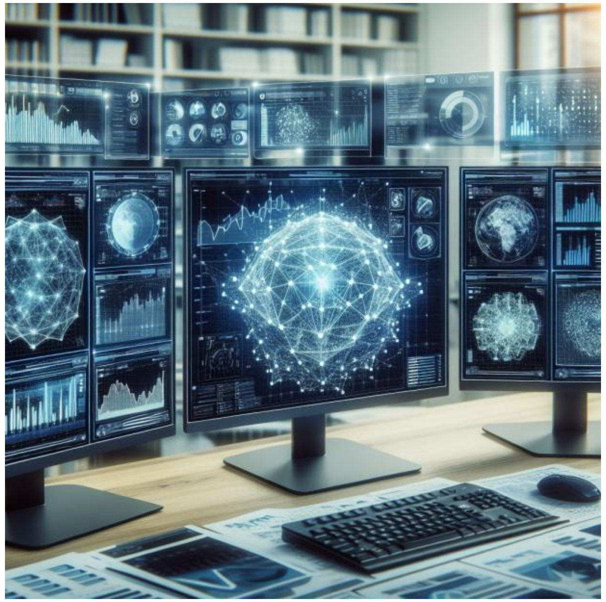
Local Differential Privacy within ENCRYPT

DP suggested by Recommendation Engine with ε-value



Users will have to locally add noise to data using the ENCYPT interface

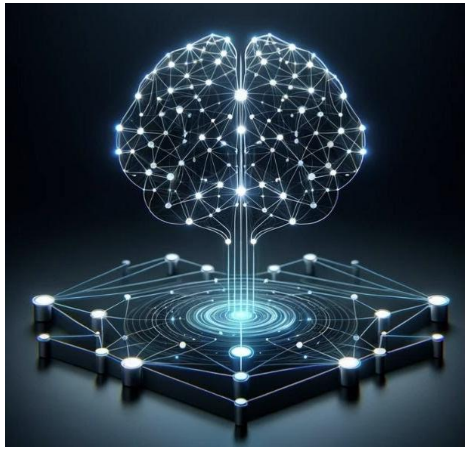
This noisy data will then be uploaded to the ENCRYPT platform



Machine Learning models can then be trained on the anonymized data

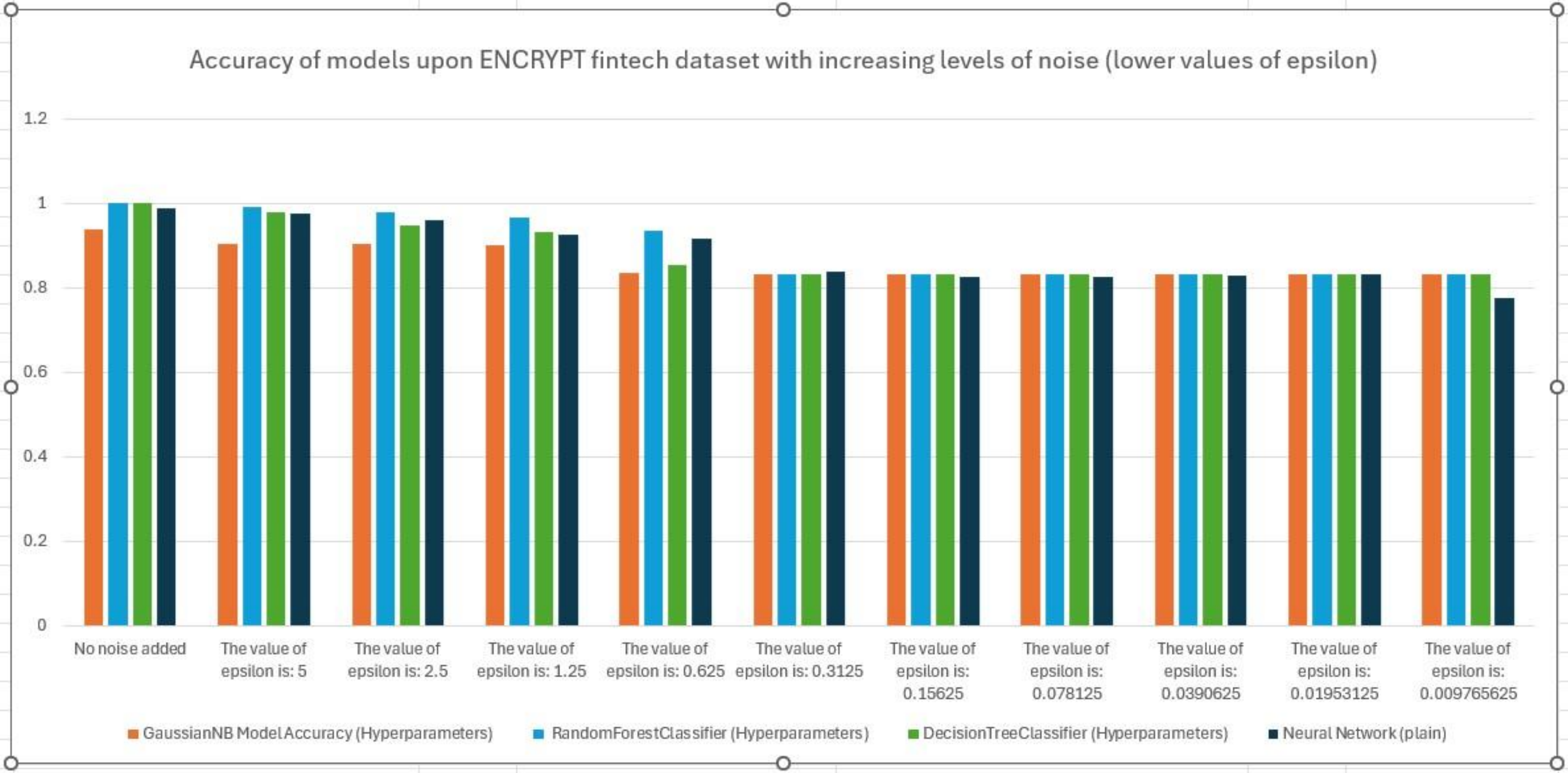
Differential Privacy – tested on USE CASE data

- Initial experiments on data provided by EXUS
  - Accuracy of Random Forest model without DP is not much greater than when DP is applied to the data
    - 91.7% vs 87.8% accuracy, *(plain Random Forest model vs D.P. Random Forest model)*
- Other experiments carried out

Decision Tree Classifier	Logistic/Linear Regression	Gaussian Naive Bayes	Artificial Neural Networks
			
87% accuracy for both plain and DP models	Random results, unable to ascertain an accuracy for either model	Depending on epsilon, comparable accuracy around 85%	- 89.5% no noise - Depends on epsilon and network, 86-90.5% accuracy

encrypt

How the value of epsilon affects model accuracy



How the value of epsilon affects model accuracy

- Experiments on dataset using 4 different models
- Each model run with different hyperparameters
  - ✓ At least 1000 iterations of each of the 4 models was run
  - ✓ The average accuracy for each model is presented

- It is clear to see that the value of epsilon affects the accuracy of models
  - ✓ Too much noise can decrease model accuracy significantly

- Finding the right balance between data utility and data randomization is most important

Thank you!

https://encrypt-project.eu/

encrypt-project

@encrypt_project