Responsible AI

Lecture 2

Today

1. Introduction to privacy
2. Differential privacy
3. Differentially private ML algorithms
4. Introduction to discrimination in ML
5. Key parameters
6. Common accuracy metrics

“‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;”

- GDPR, Article 4

> **In 2006, Netflix released a dataset containing ~100M movie ratings by ~500K users (1/8 of the Netflix user base)**

- Narayanan and Shmatikov, 2008

What is Privacy in Data?

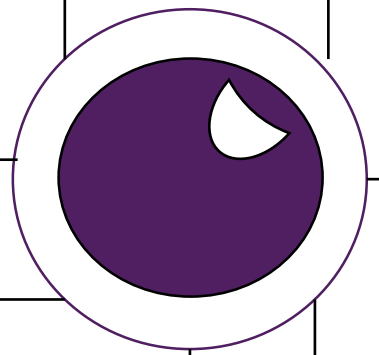
Privacy Enhancing Techniques by technology

- Differential Privacy
- Federated Analysis

Whether or how data is shared with third parties.

Privacy by design

- Data minimization
- Purpose Limitation
- Security(technical organizational)



Misuse of Personal data

- sell the information on the dark web
- to carry out financial or identify fraud
- Selling of information to Marketing firms
- charge a ransom to safely return it to the victim.

How data is legally collected or stored.

Threats to Privacy:

- Malware
- Non-secure third-party services
- Data Breaches

Regulatory restrictions such as GDPR, HIPAA, GLBA, or CCPA.

Differential Privacy (DP)

Definition

Achieved by introducing some **noise** which is enough to protect privacy and at the same time limited enough so that the provided information is still useful.

Goal

Allow data analysts to build accurate models without sacrificing the privacy of the individual data points.

Approaches to protect privacy

- **Sampling** (“just a few”) - release a small subset of the database
- **Aggregation** (e.g., k-anonymity) - each record in the release is indistinguishable from at least k-1 other records
- **De-identification** - mask or drop personal identifiers
- **Query auditing** - stop answering queries when they become unsafe

Differential privacy – a way to secure AI

The chance that the noisy published result **S** is same with/without you (***i***):

$$\frac{\boldsymbol{Pr(f(D_I = S)}}{\boldsymbol{Pr(f(DI}_{\mp\, i} = \boldsymbol{S})} \leq \boldsymbol{e^{\varepsilon}}$$

$\boldsymbol{f}$ is a Query here seeking private information

e^{ε} is very close to 1, with ε>0

If e^{ε} is large then no privacy, if e^{ε} = 1, the new data has no utility

Cynthia Dwork

Is differential privacy a solution?

A randomized algorithm M provides ε-differential privacy if, for all neighbouring databases D_1 and D_2, and for any set of outputs S:

$$Pr\,[M(D_1) \in S] \leq e^{\varepsilon}\, Pr\,[M(D_2) \in S]$$

ε (epsilon) is a privacy parameter or the privacy budget (lower ε, stronger privacy)

Cynthia Dwork

Sensitivity, noise and differential privacy

Sensitivity is the maximum amount that a query changes when removing an **individual** from a database

$$\Delta f = \max_{D1,D2} \| f(D_1) - f(D_2) \|_{L1}$$

counting the number of rows has a sensitivity of 1, because adding or removing a single row from any dataset will change the count by at most 1

If Δf is the sensitivity of a function, then adding Laplace noise with scale $\frac{\Delta f}{\varepsilon}$ preserves ε-differential privacy. The output function of f as a real valued function is $\tau_M(x) = f(x) + noise$

Apple uses an epsilon of 2 in their keyboard's differentially-private auto-correct. It adds noise to individual user inputs. That means it can track, for example, the most frequently used emojis, but the emoji usage of any individual user is masked

Cynthia Dwork

Noise and Laplacian

Laplacian is Lap (μ, b)

$$f(x|\mu, b) = \frac{1}{2b} e^{\frac{-|x-\mu|}{b}}$$

Substituting b = $\frac{\Delta f}{\varepsilon}$; μ = $f(D_1)$ in above equation we get,

$$\frac{1}{2\frac{\Delta f}{\varepsilon}} e^{\frac{-|x-f(D)|\varepsilon}{\Delta f}}$$

Private **R** = $f(D_1)$ + Lap $(\frac{\Delta f}{\varepsilon})$

Where's Next

- Federated Learning
- DP SGD
- Exponential mechanism DP
- PySyft
- TF encrypted
- TF Privacy
- Moment's Accountant

How DP Works

- Introduces privacy loss parameter (**ε**) to the dataset. This adds randomness to data.
- A high value of **ε** means more accurate but less private data
- Noise can be added to **data** and/or **algorithm**

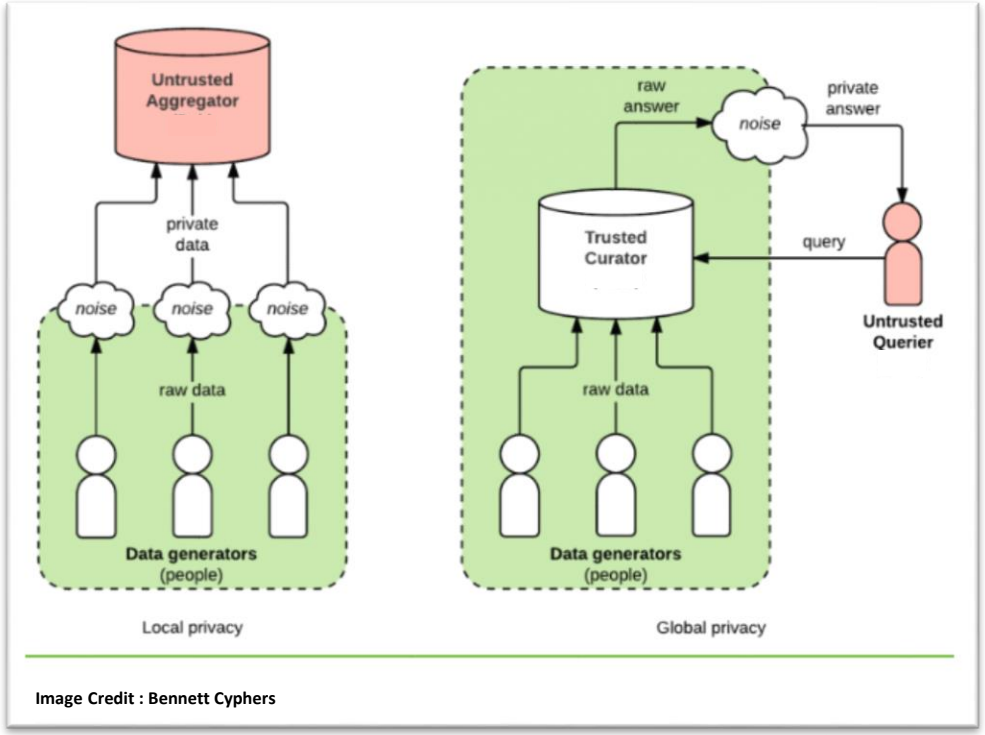


Image Credit : Bennett Cyphers

Value generation with synthetic data

Any data algorithmically generated approximating original data.

Motivation:

Safety and Privacy

Industry Collaboration

Responsible AI

Secure

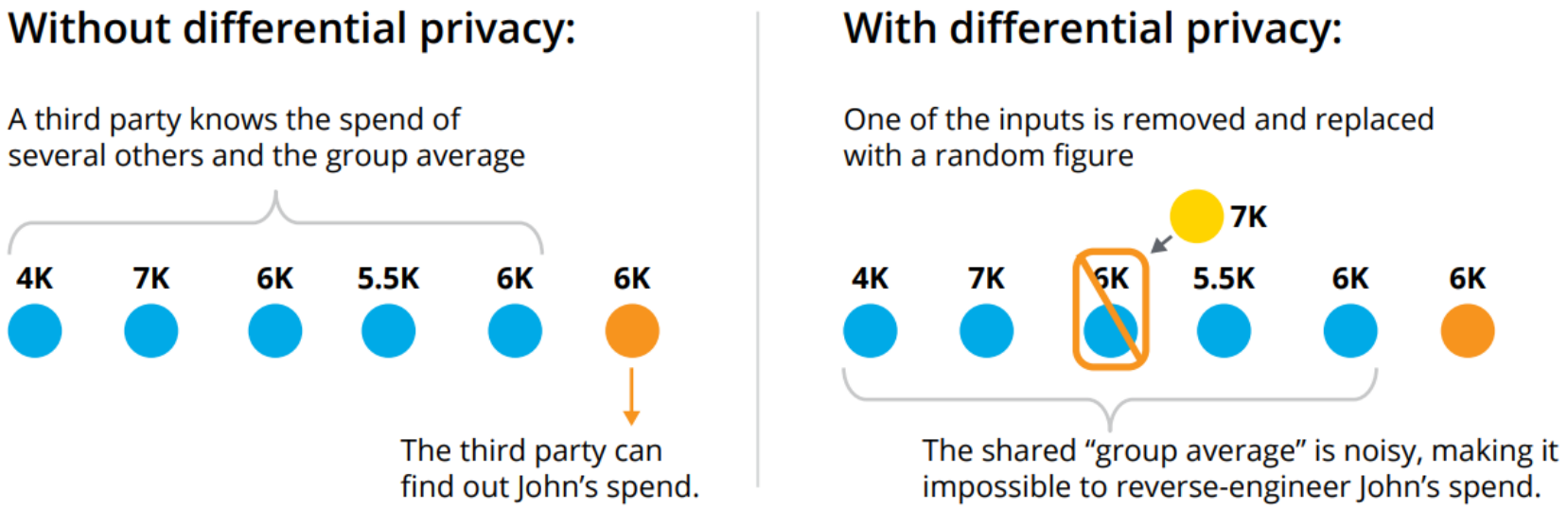
- Create safe datasets that retain the same insights and statistical integrity equivalent to original data source
- Develop cross-domain AI use cases to drive industry collaboration
- Drive responsible AI practices through balanced synthetic datasets or de-bias datasets for ML/AI model testing
- Defend against re-identification and joinability attacks.

Differential Privacy in Data

For Binary (Healthcare) Data

Data is perturbed before computation so, every individual in the data set would be able to plausibly deny that their actual response was included.

For Continuous (Financial) Data



fractal

Differential Privacy in Algorithms

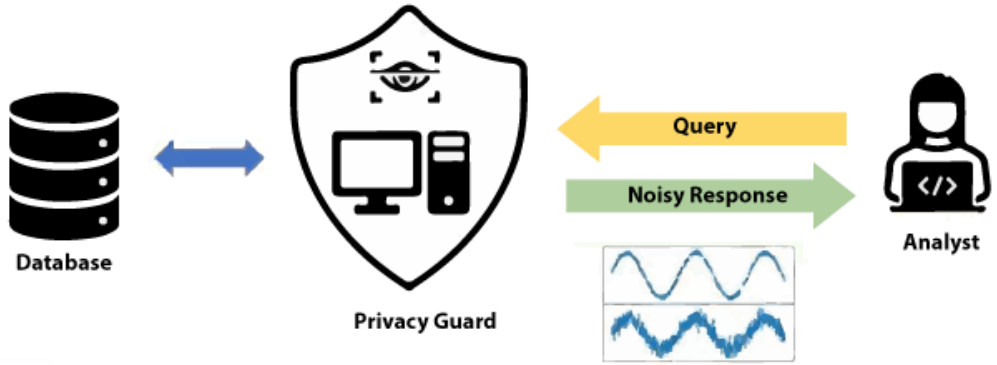
The algorithm might produce an output, on the database that contains some individual's information, is almost the same output that a database generates without having individuals' information. This assurance holds true for any individual or any dataset.

Database Shape

Rating	Count
Bad	3
Normal	1510
Good	200

Adversary wants to know the no. of bad rated movies

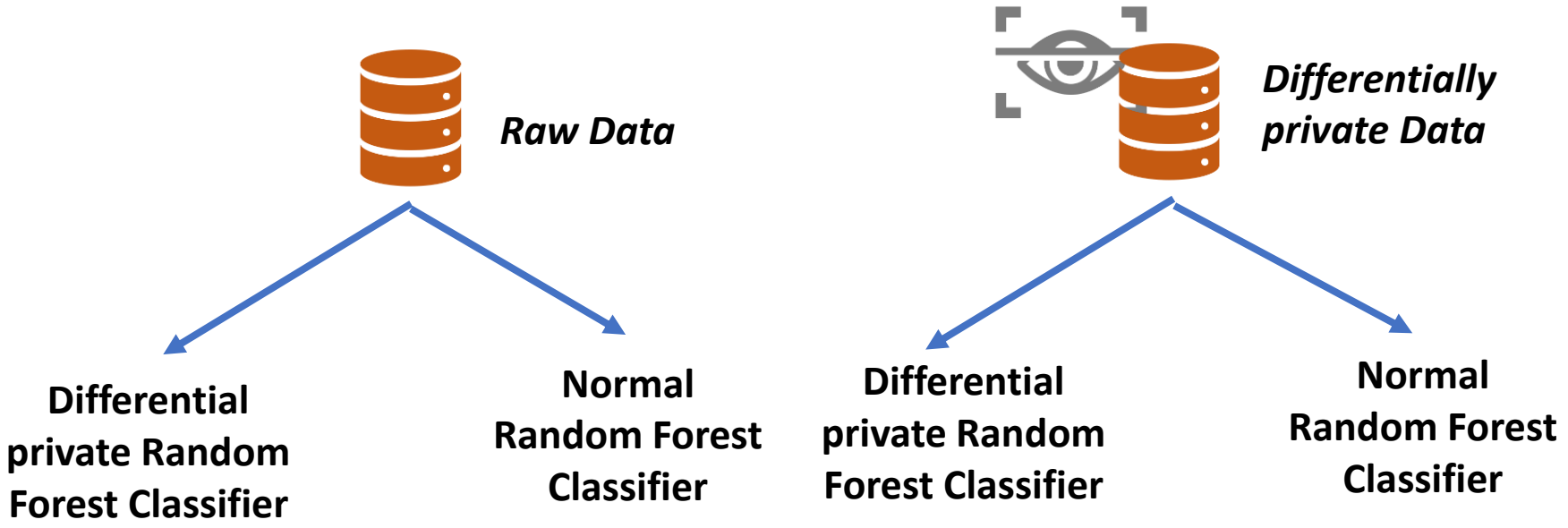
Differential Private Algo



Query Number	Response
1	2.915
2	1.882
3	1.292
4	4.026
5	5.346
Average	3.090
90% confidence interval	1.696 to 4.484

fractal

DP in ML model



- Libraries used were – Diffprivlib (IBM initiative), sklearn, pandas, numpy, matplotlib,etc.

How to make models Differential Private

BY ADDING NOISE TO MODEL'S OBJECTIVE FUNCTION

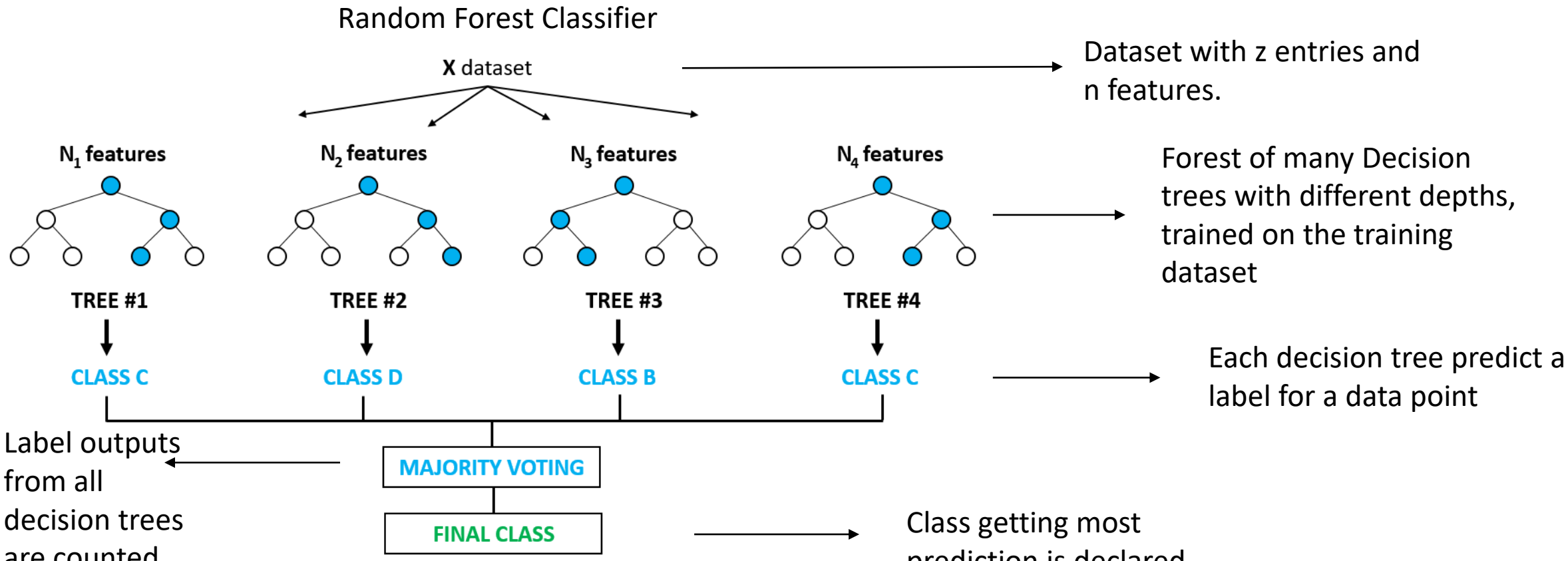
BY ADDING NOISE TO MODEL OUTPUT

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

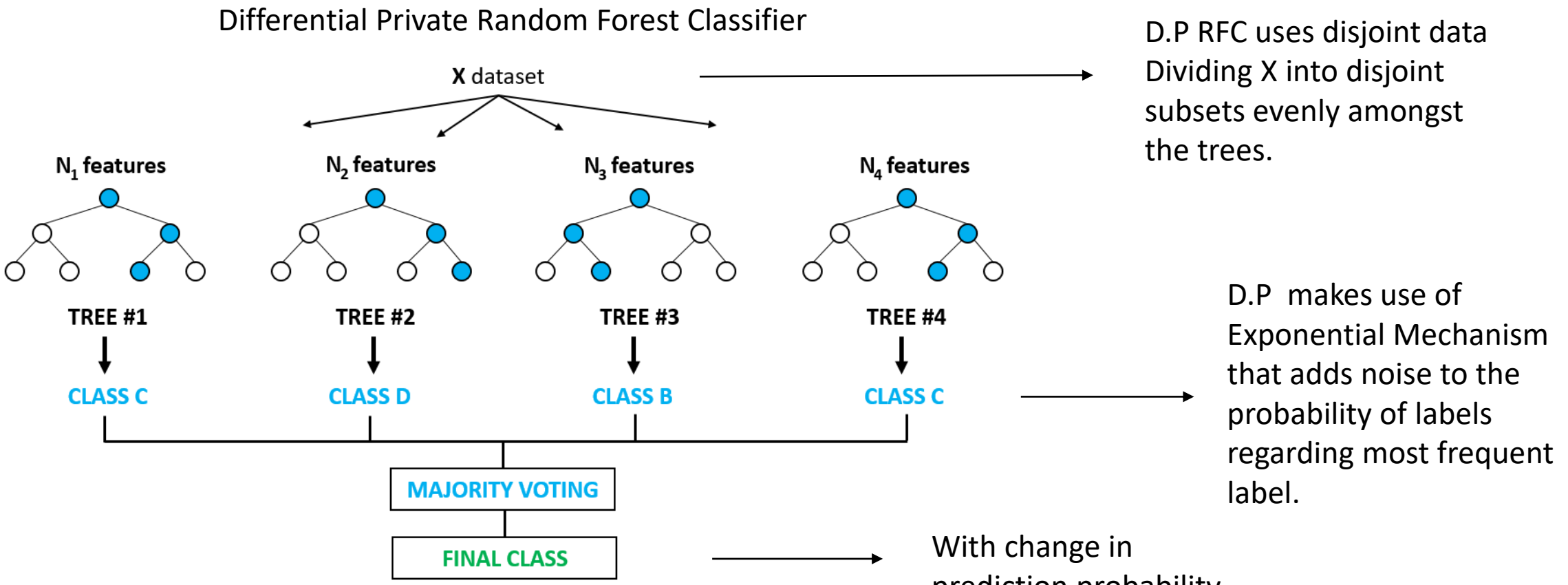
Normal model vs D.P model

Models	How is model made Differential private?
Random Forest Classifier	Exponential Mechanism adds noise to the probability of labels regarding most frequent label.
K- means Classifier	Noise is added to the averages of centroids calculated where noise is taken from a Laplace distribution which is function of number of centroids, epsilon, sensitivity, number of data partitions.
Linear Regression	By adding Laplacian noise to the coefficients of the objective function
Logistic Regression	Noise is added to the coefficients of each feature where noise is proportional to exponential function.
Naïve Bayes	Laplacian noise of the appropriate scale (and mean 0) is added to the parameters (the counts for categorical attributes, the means and standard deviations for numeric attributes).

How is Differential Privacy introduced in Random Forest Classifier?



How is Differential Privacy introduced in Random Forest Classifier?



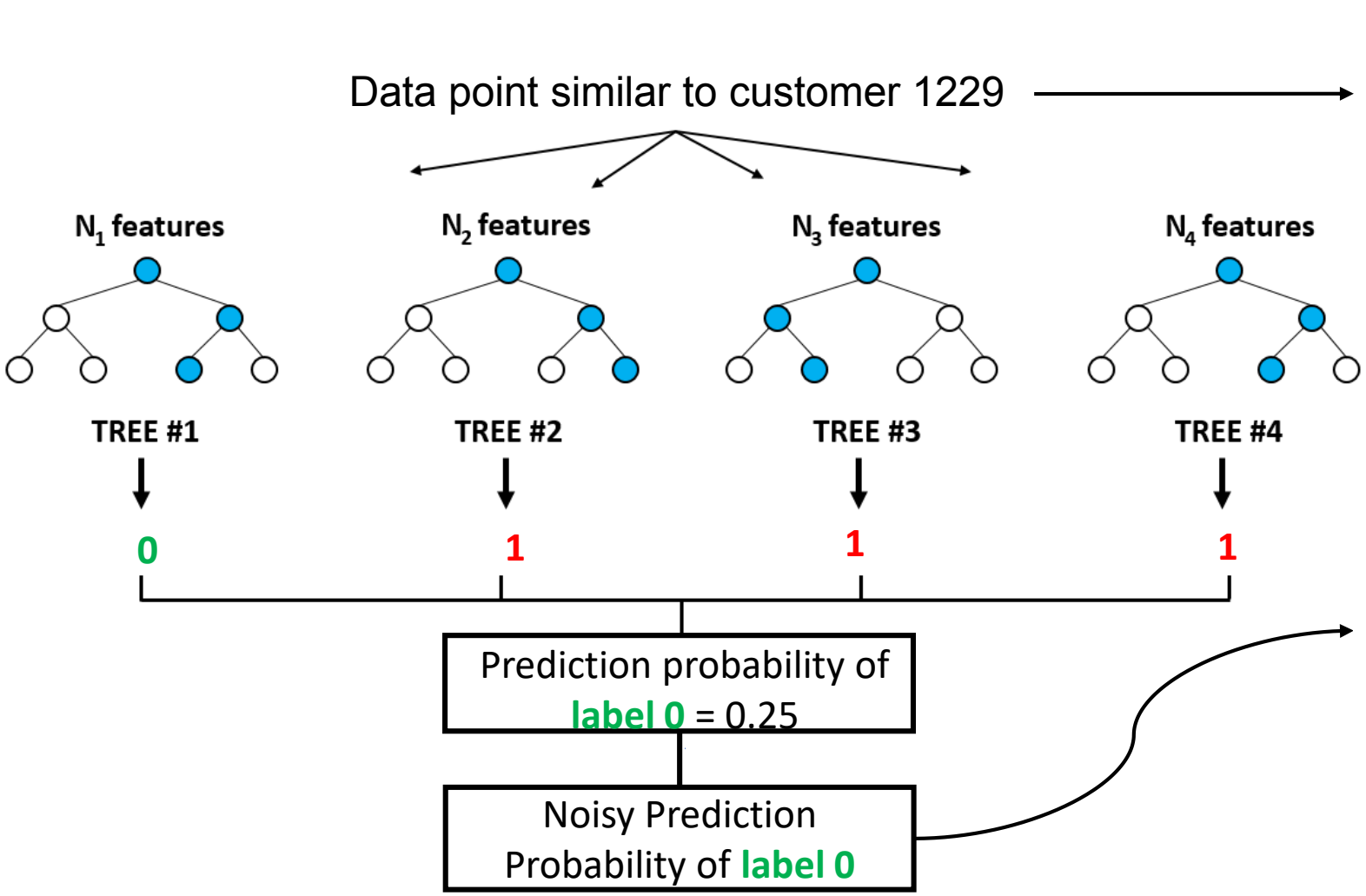
D.P RFC uses disjoint data Dividing X into disjoint subsets evenly amongst the trees.

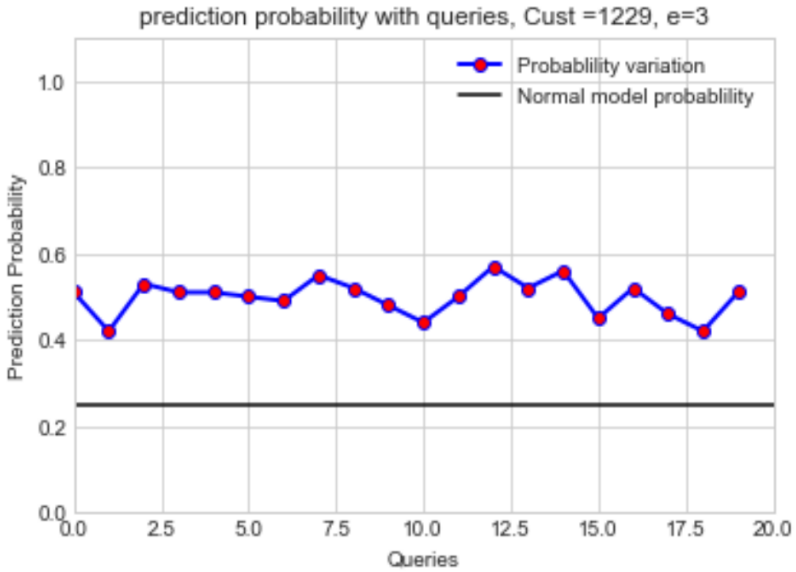
D.P makes use of Exponential Mechanism that adds noise to the probability of labels regarding most frequent label.

With change in prediction probability the output label can change.

fractal

How D.P Random Forest embeds privacy?

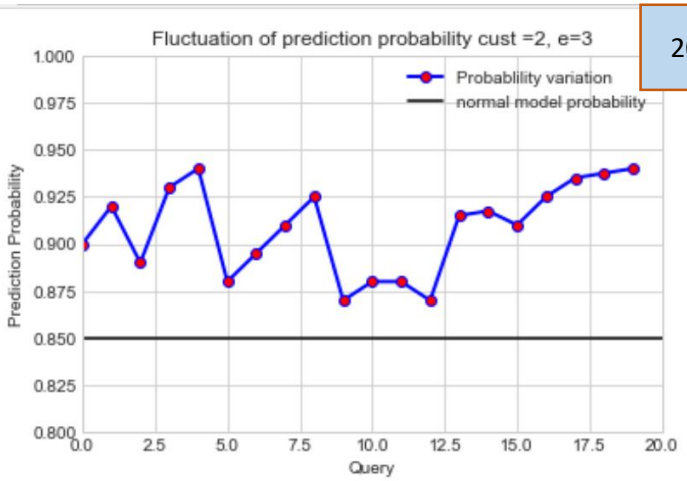




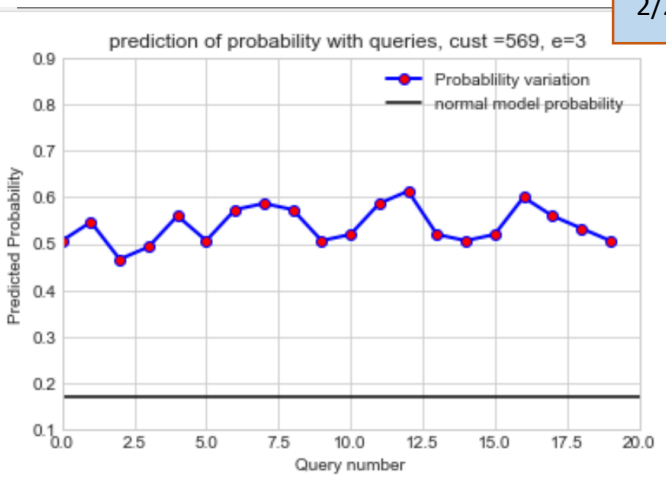
fractal

Good, Bad and Ugly

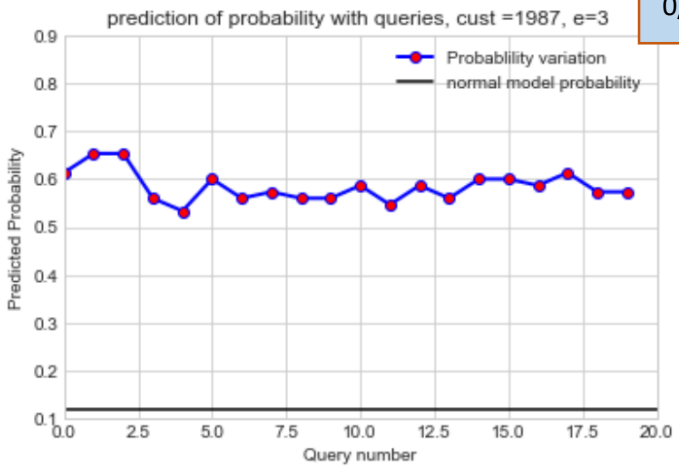
20/20



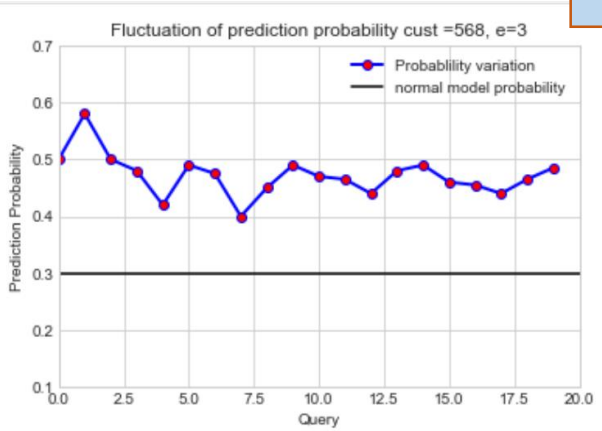
2/20



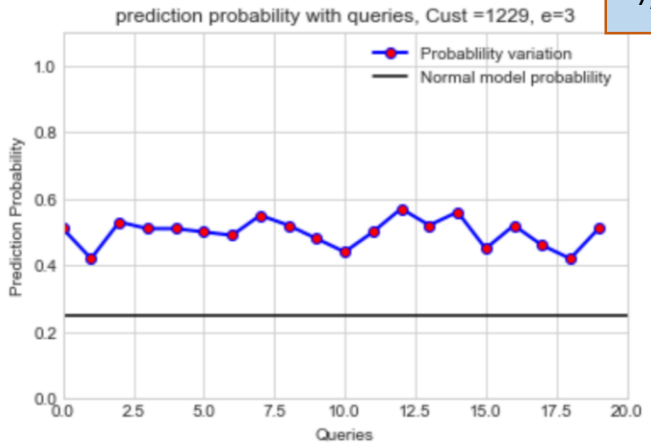
0/20



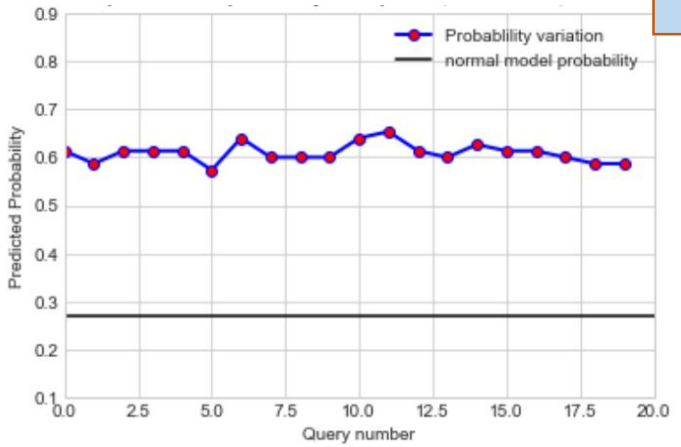
17/20



7/20

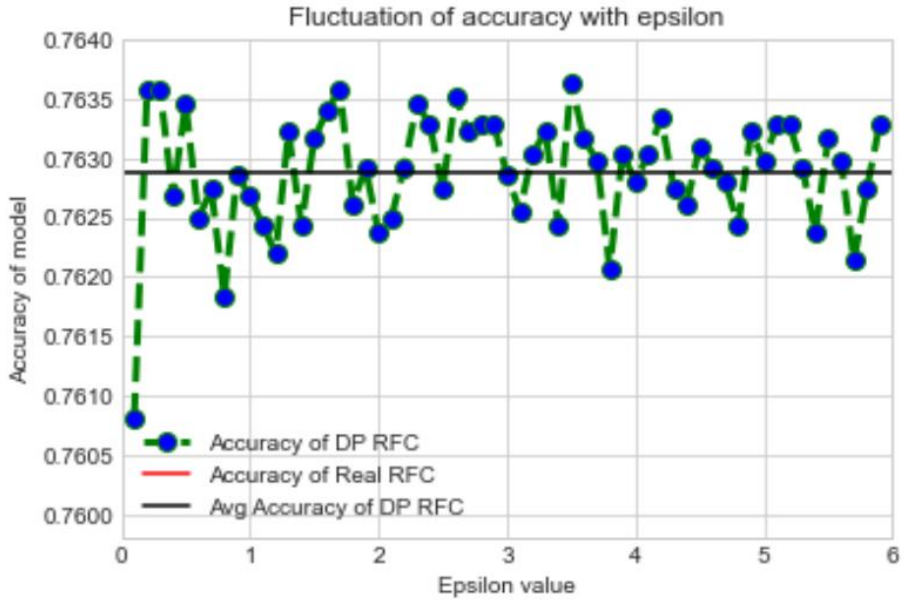


0/20

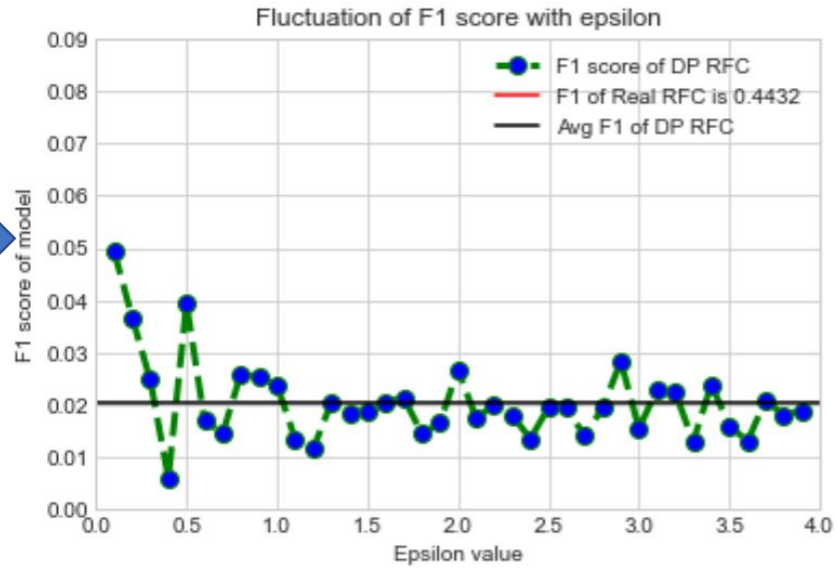


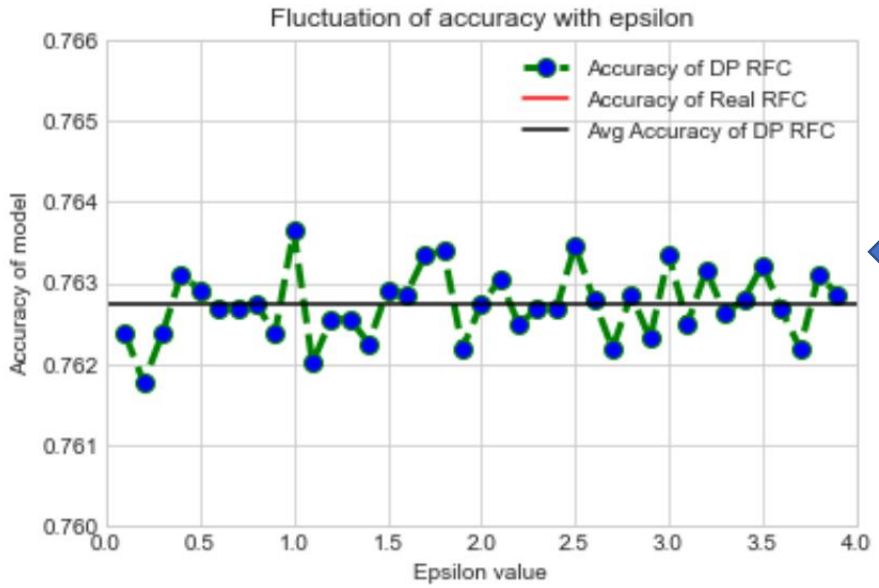
fractal

Fluctuation with different epsilons

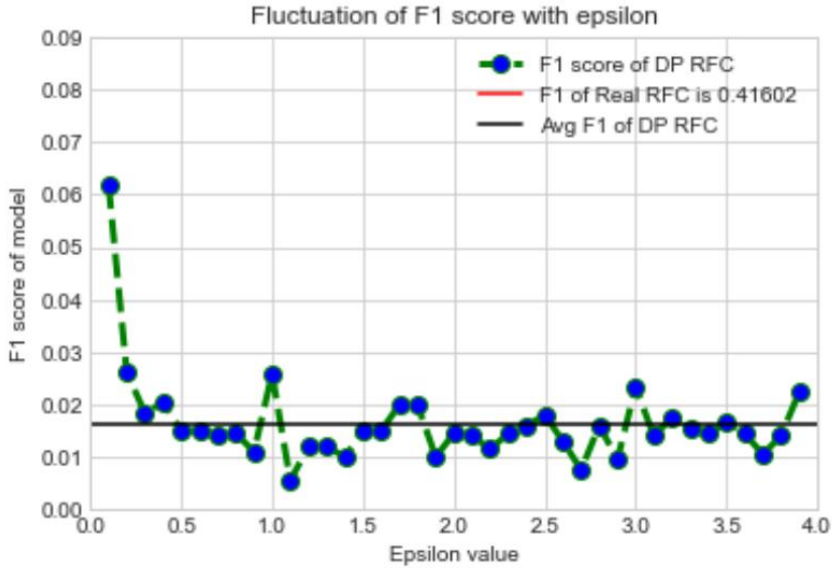


Differential private model with raw data





Differential private model with D.P data



tractable

Logistic Regression

A randomized algorithm *A*, taking a dataset as input, is said to be **ϵ**-differentially private if it holds that

$$|\log(P(A(D) \in S)) - \log(P(A(D') \in S))| \leq \epsilon$$

output perturbation

$$w' = w^* + 2/(n\lambda\epsilon)b,$$

Where training data of size *n* and dimension *d* with labels *y* and covariates *x*

objective perturbation

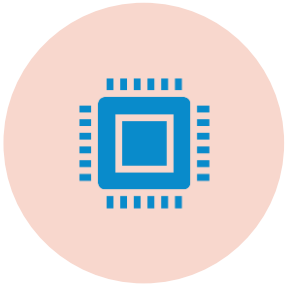
$$F(w, \lambda, \epsilon) = J(w, \lambda) + 2/(\epsilon n)b^T w$$

https://livebook.manning.com/book/privacy-preserving-machine-learning/chapter-3/v-1/1

How DP data and DP model together embeds privacy?

Differential Data ensures you the **plausible deniability** of your answer.

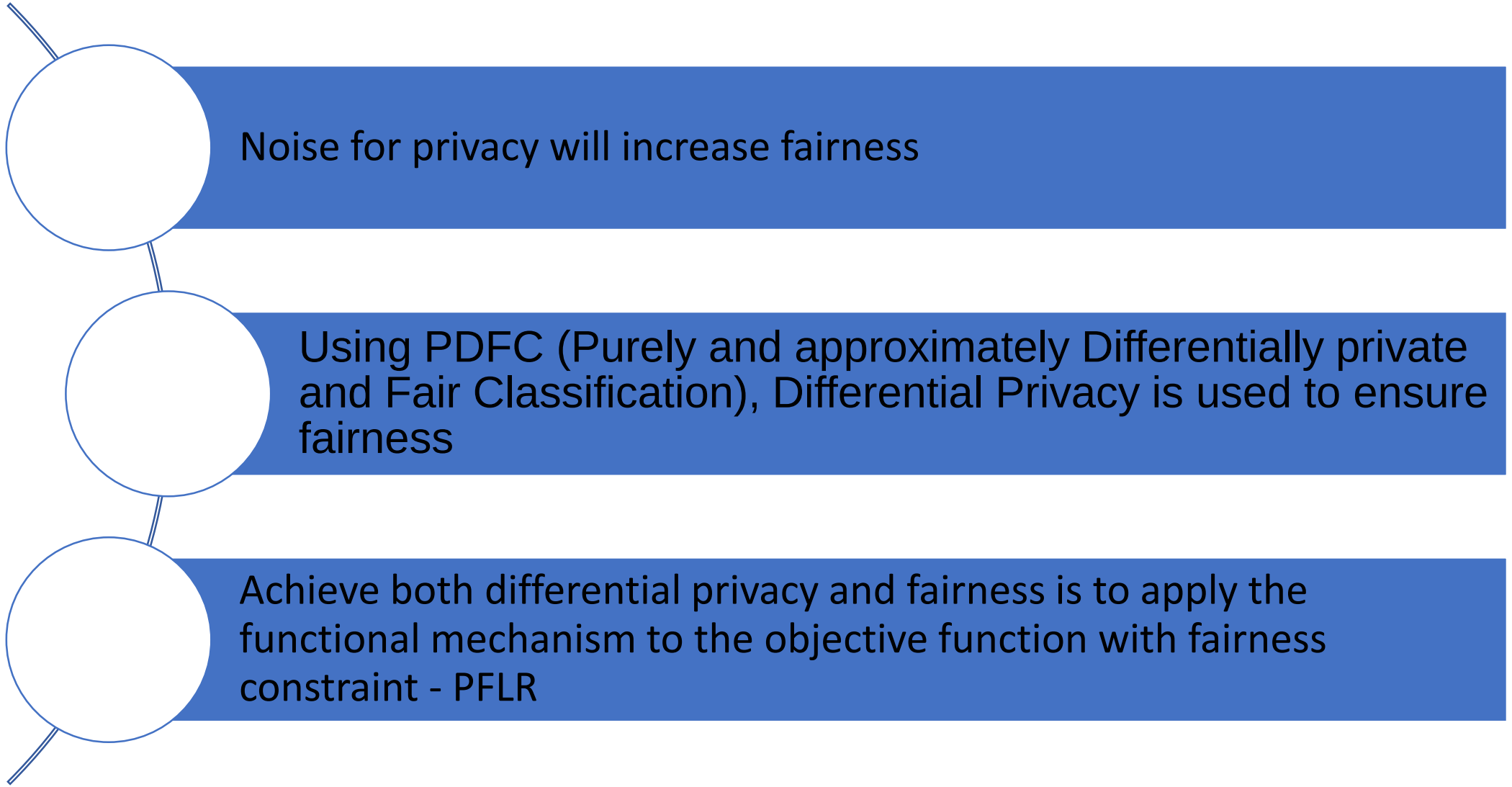
Provides privacy even at the level of data storing.



Differential Private algorithms satisfy privacy by sometimes returning false value.

Adversary makes the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis.

Privacy and Fairness



Why Privacy

In Group of three name three cases of Privacy breach

Assessment

Use any data of your choice having PII data:

1. Use DP on data (age/salary/income etc)
    1. Loop through multiple epsilon and find a value that you feel has least trade off – plot it as graph with a line representing the actual value
    2. Keep epsilon constant, and iterate it 100 times and show average of all iterations
2. DP model:
    1. Use Logistic regression DP on original Data
    2. Use RF DP on original data
    3. Report the findings
3. Model on data DP:
    1. Use Logistic regression on DP Data
    2. Use RF on DP data
    3. Report the findings
4. DP Model and DP data
    1. Logistic Regression
    2. RF

Submission on 4th lecture

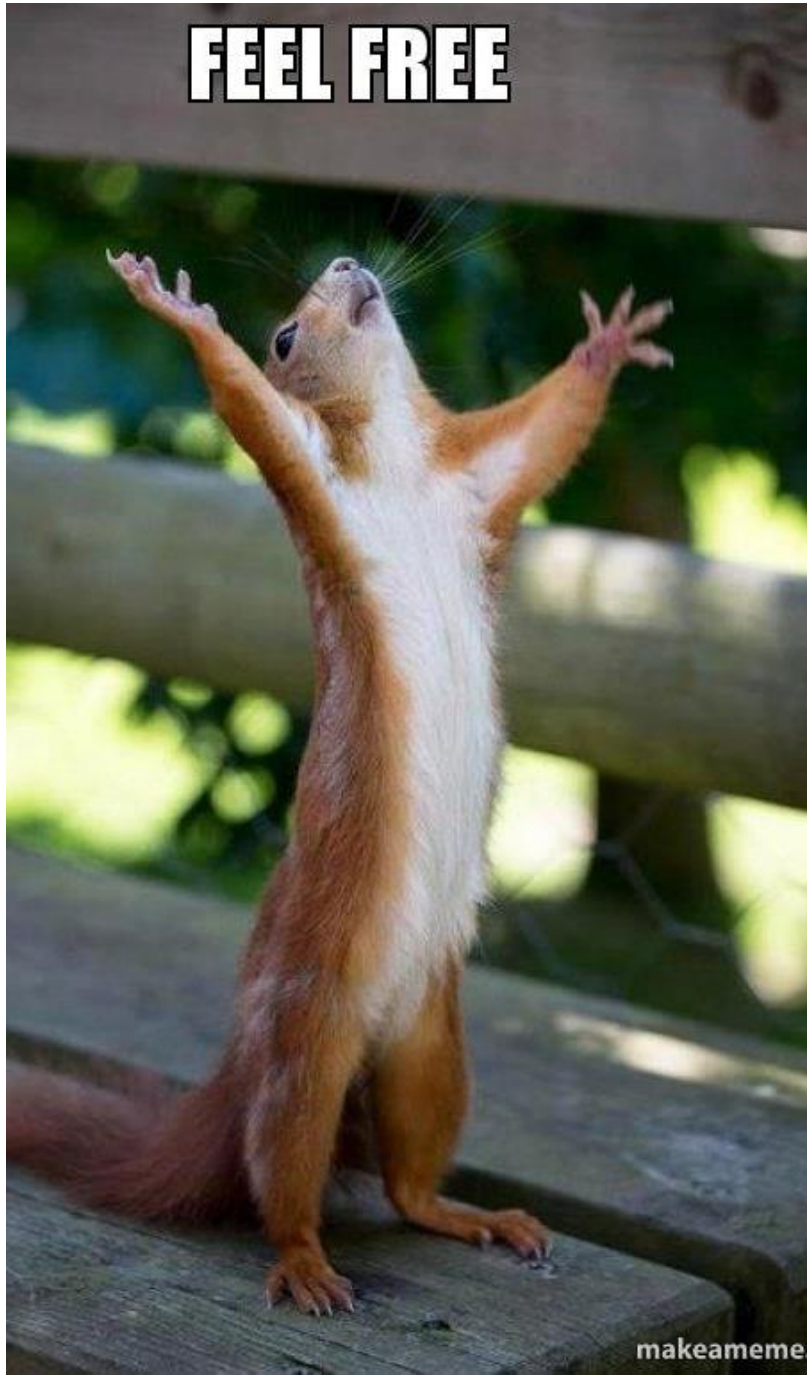
Explore

- Microsoft DP libraries
- IBM 360 DP algorithms
- White Noise (https://github.com/srayagarwal/whitenoise-core-python)
- Github (https://github.com/srayagarwal/JIO_RAI/blob/main/Ch%208%20Data%20and%20Model%20Privacy.ipynb)

Next

1. Fairness and Proxies
2. Fairness metrics
3. Proxy features
4. Methods to detect proxy features
5. Variance Inflation Factor (VIF)
6. Linear association method using variance

Read Ch 2 & 9 from RAI book
Revise: Cosine similarity, Distance method, Mutual Information


FEEL FREE
makeameme


YOU CAN CALL ME


DID YOU REALLY JUST SEND THAT EMAIL!


KNOCK, KNOCK!