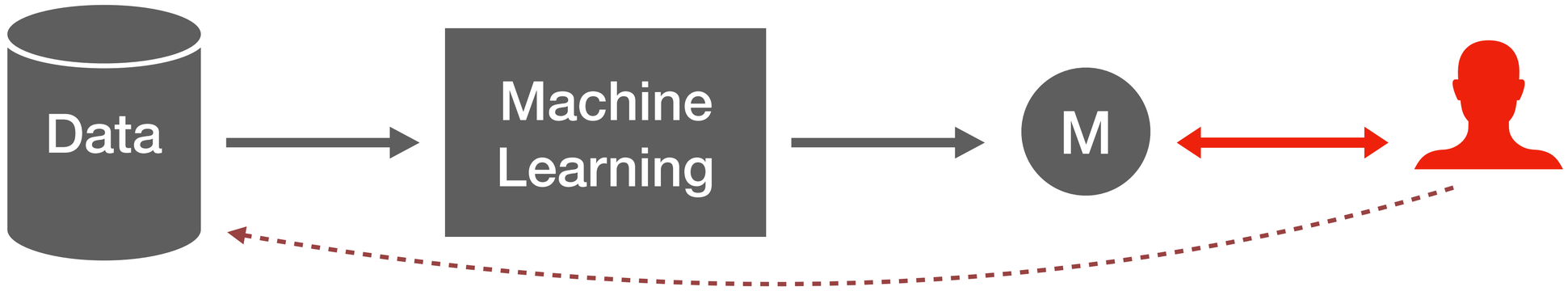
Evaluating Differentially Private Machine Learning in Practice

Bargav Jayaraman and David Evans
Department of Computer Science
University of Virginia

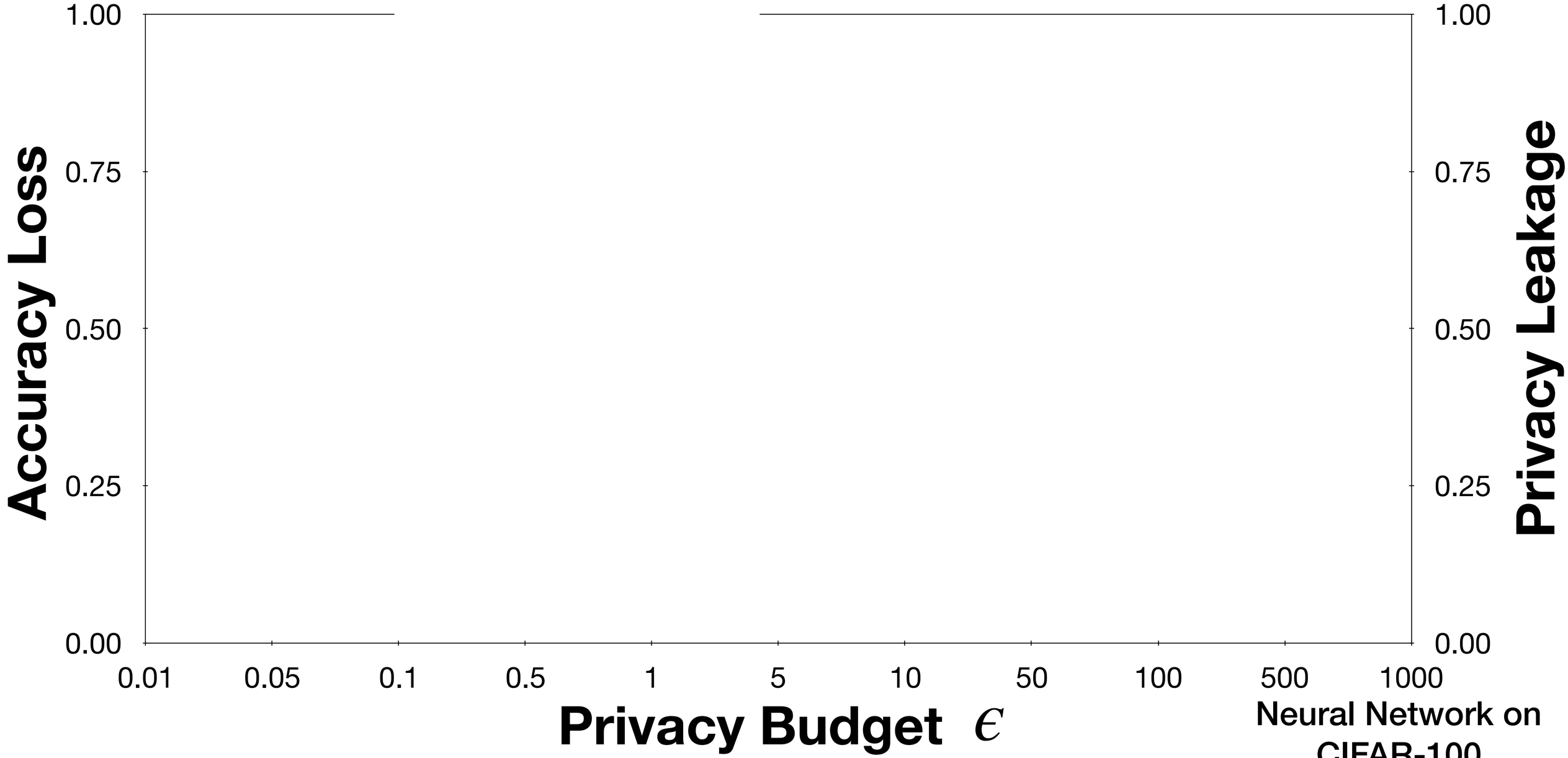
Our Objective



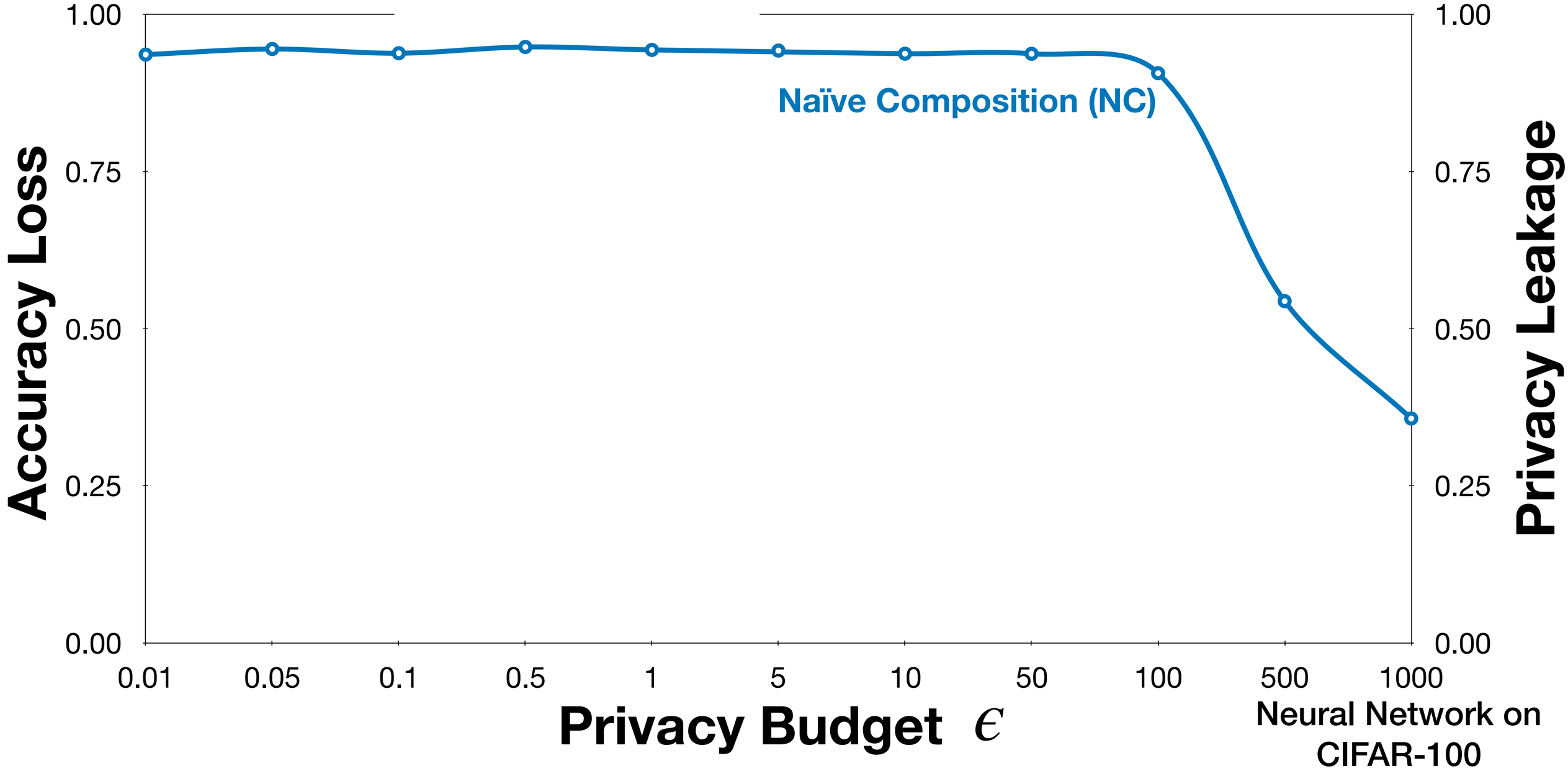
To evaluate the privacy leakage of private mechanisms

Leakage is quantified in terms of inference attacks

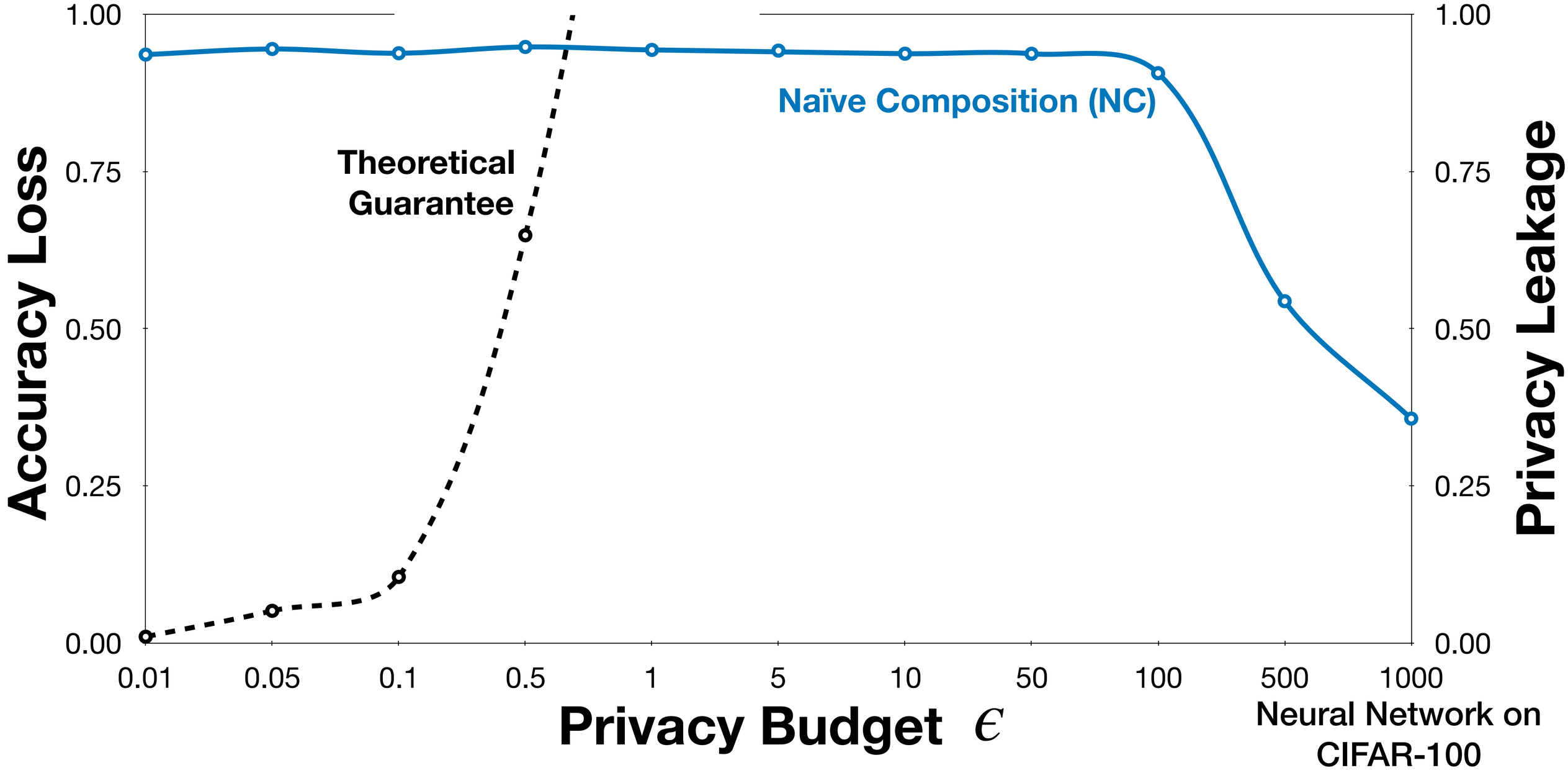
Result Highlights



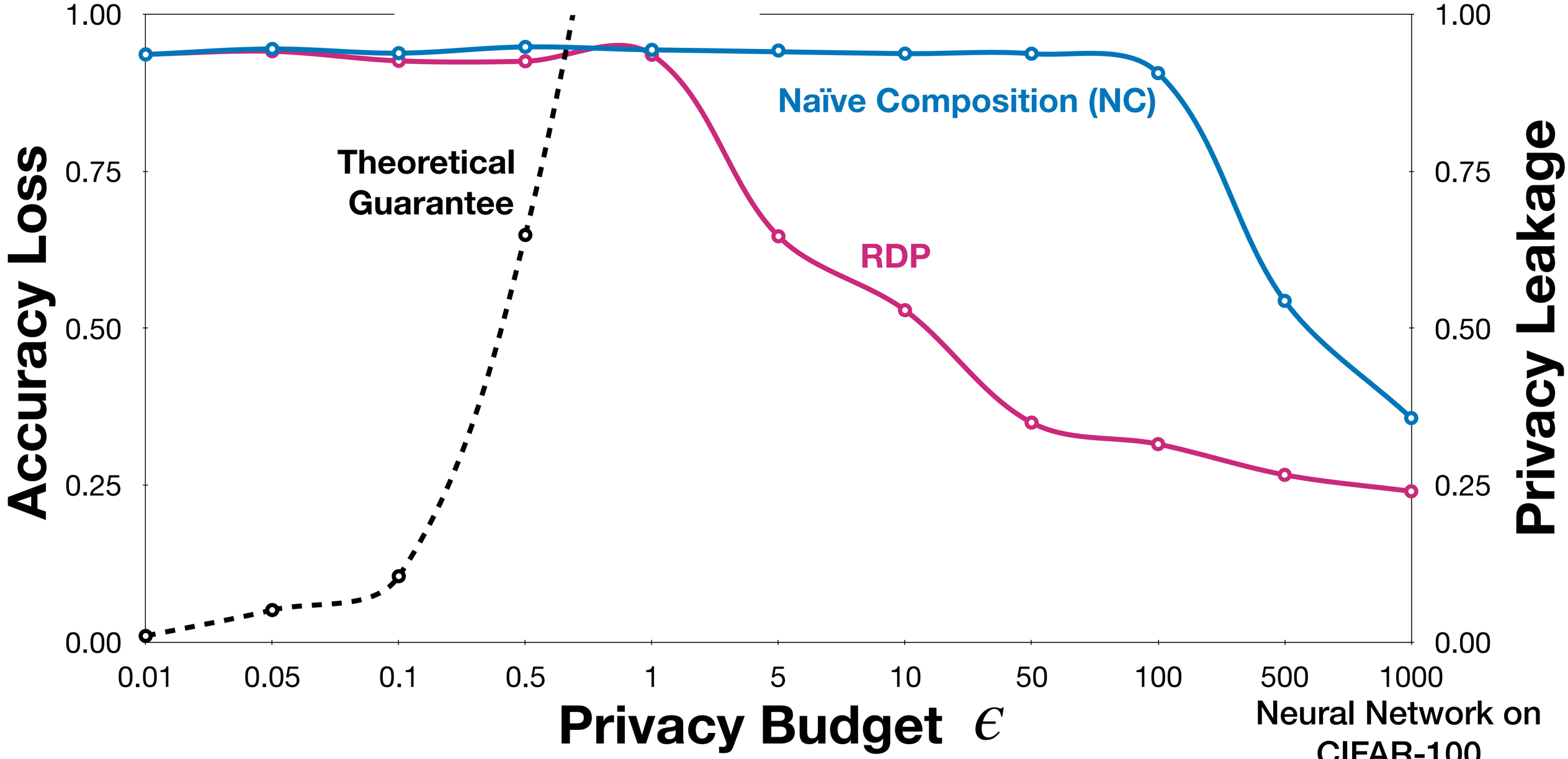
Result Highlights



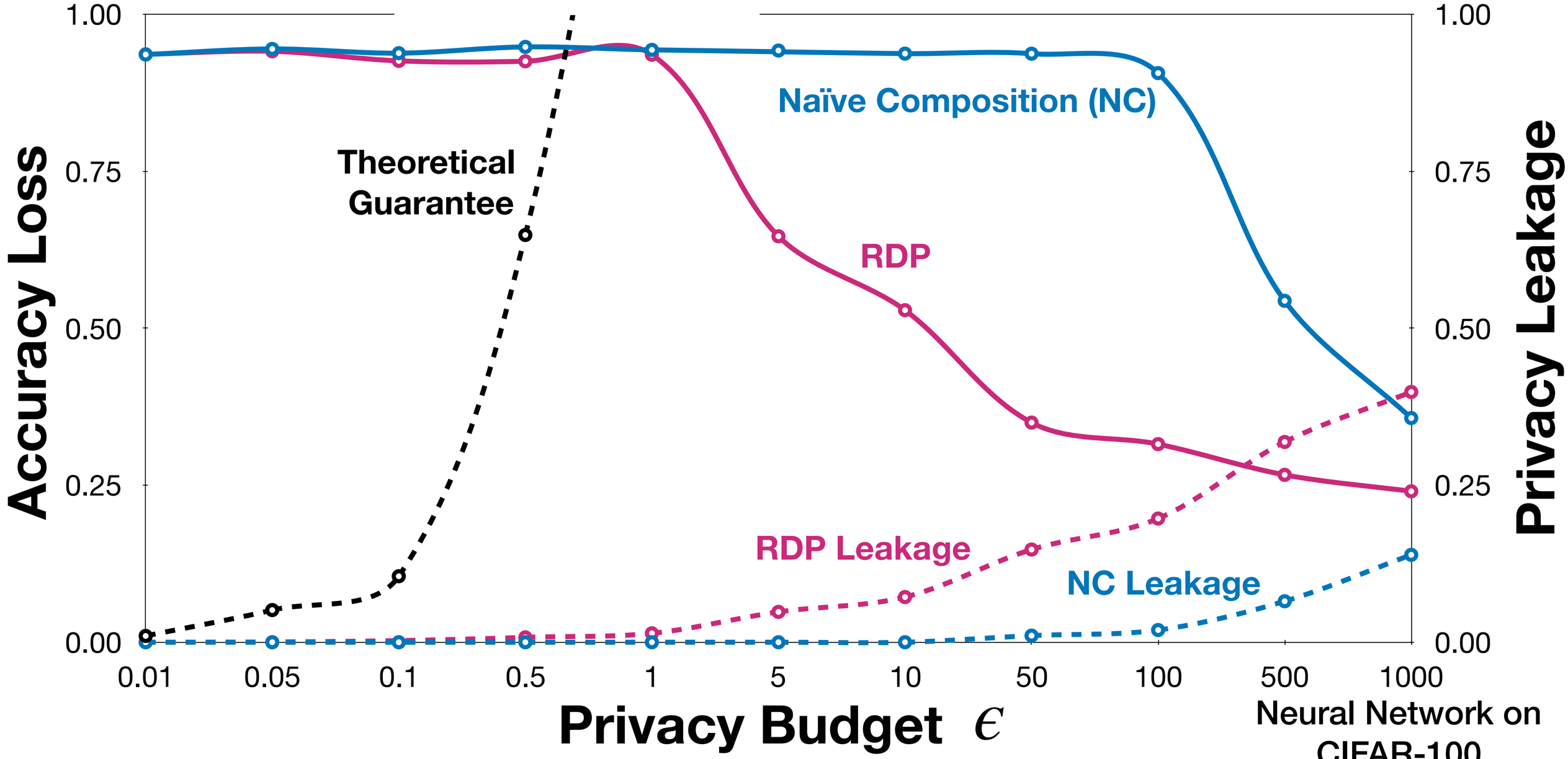
Result Highlights



Result Highlights



Result Highlights

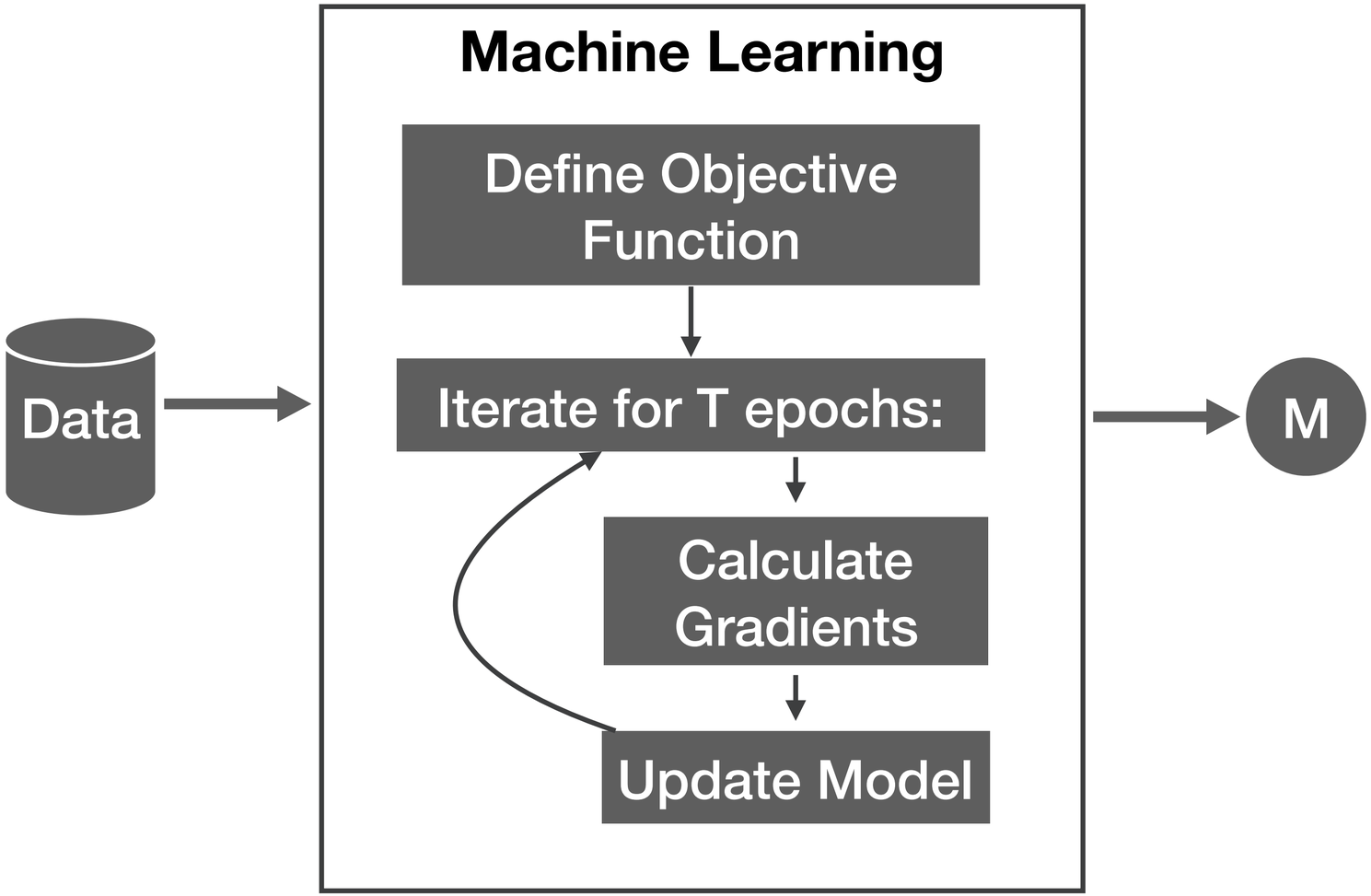


Rest of the Talk

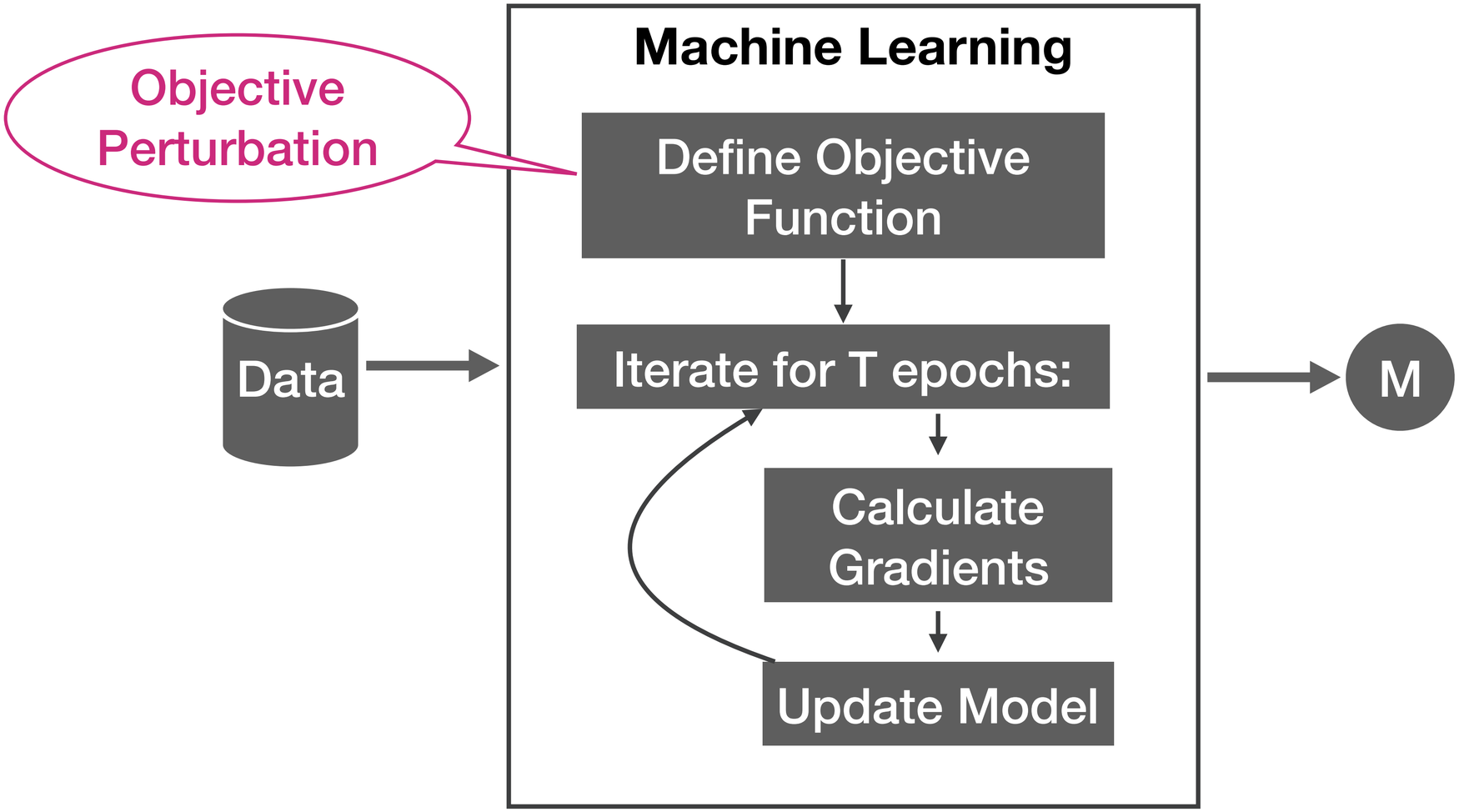
1. Background on Applying Differential Privacy to Machine Learning

2. Experimental Evaluation of Differentially Private Machine Learning Implementations

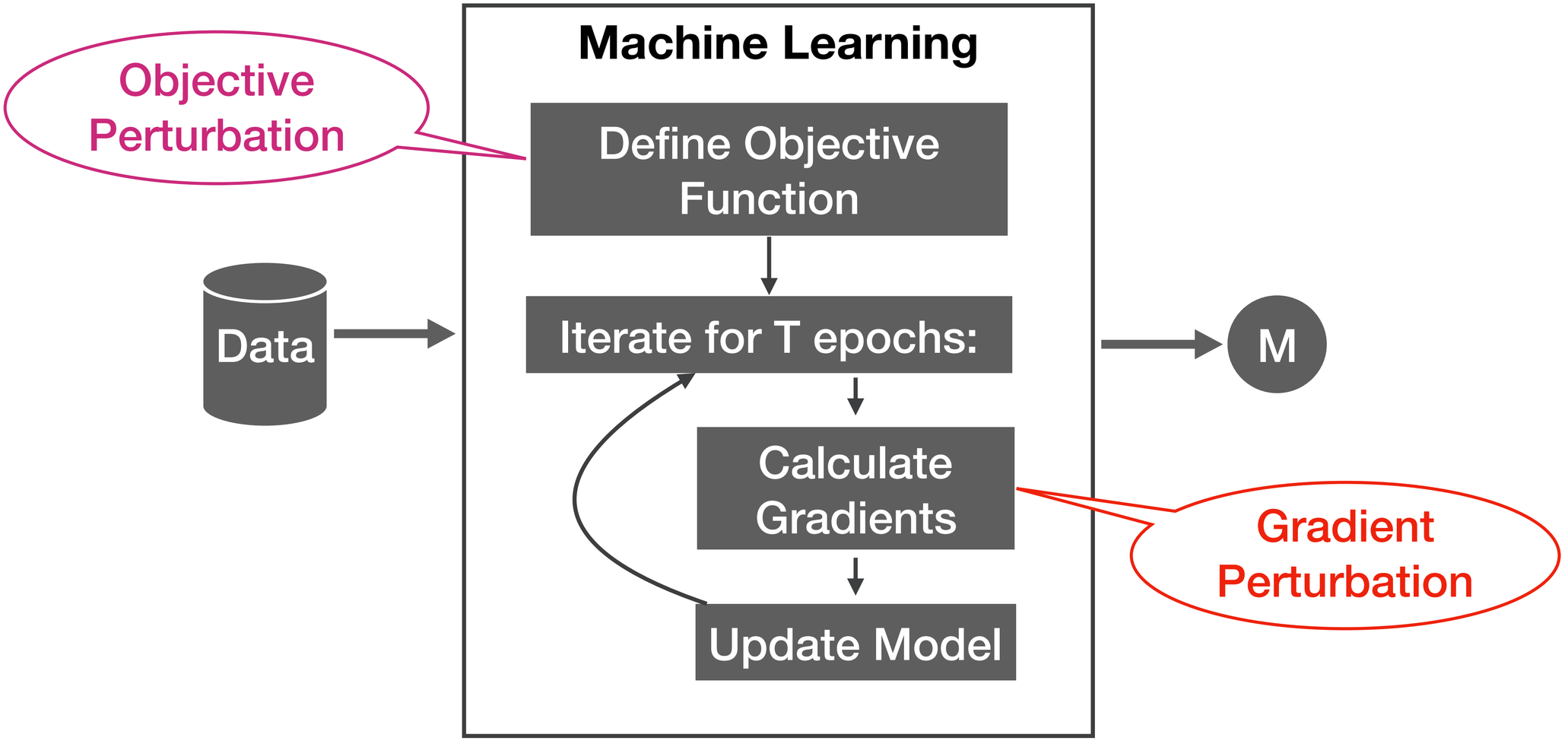
Applying DP to Machine Learning



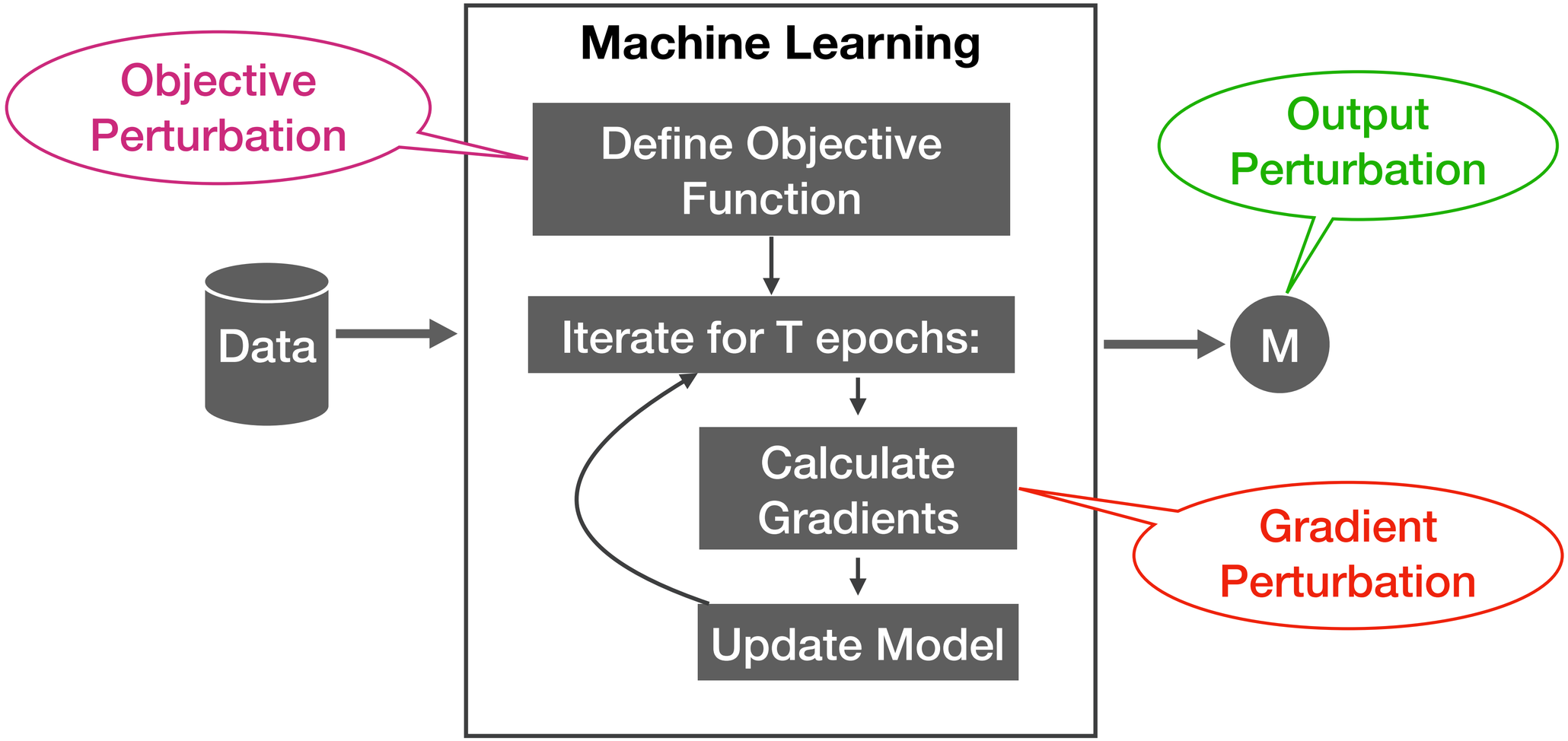
Applying DP to Machine Learning



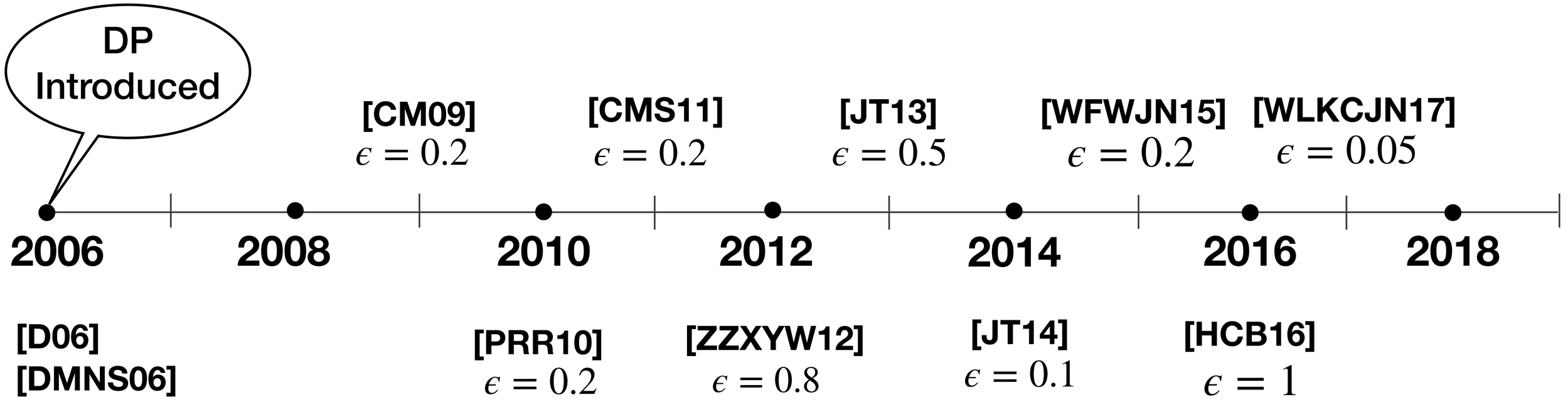
Applying DP to Machine Learning



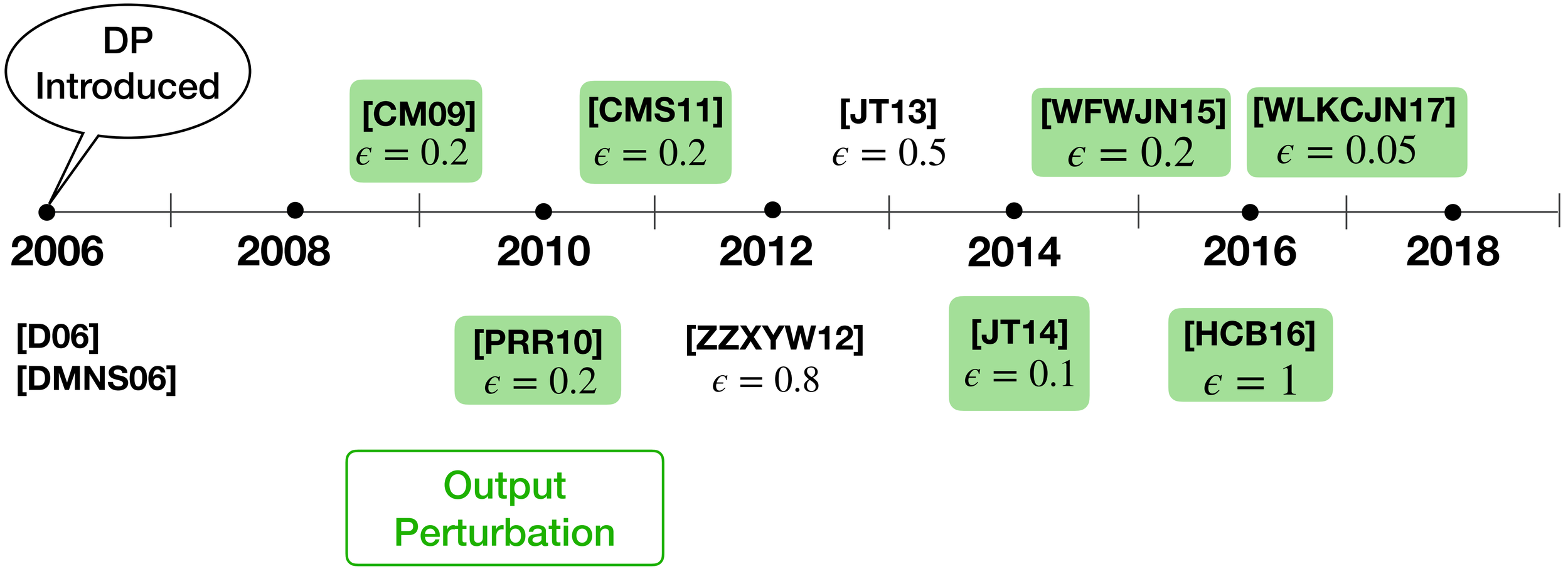
Applying DP to Machine Learning



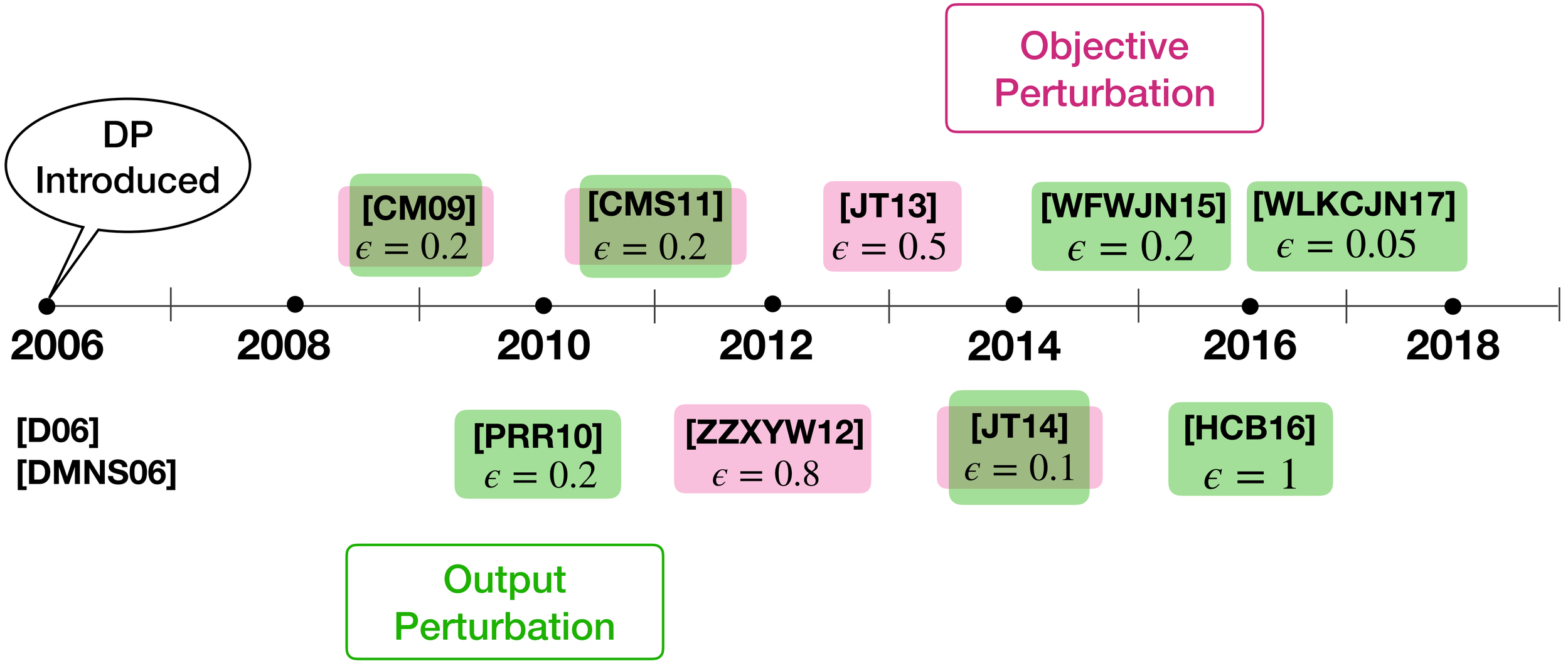
ERM Algorithms using $\epsilon \leq 1$



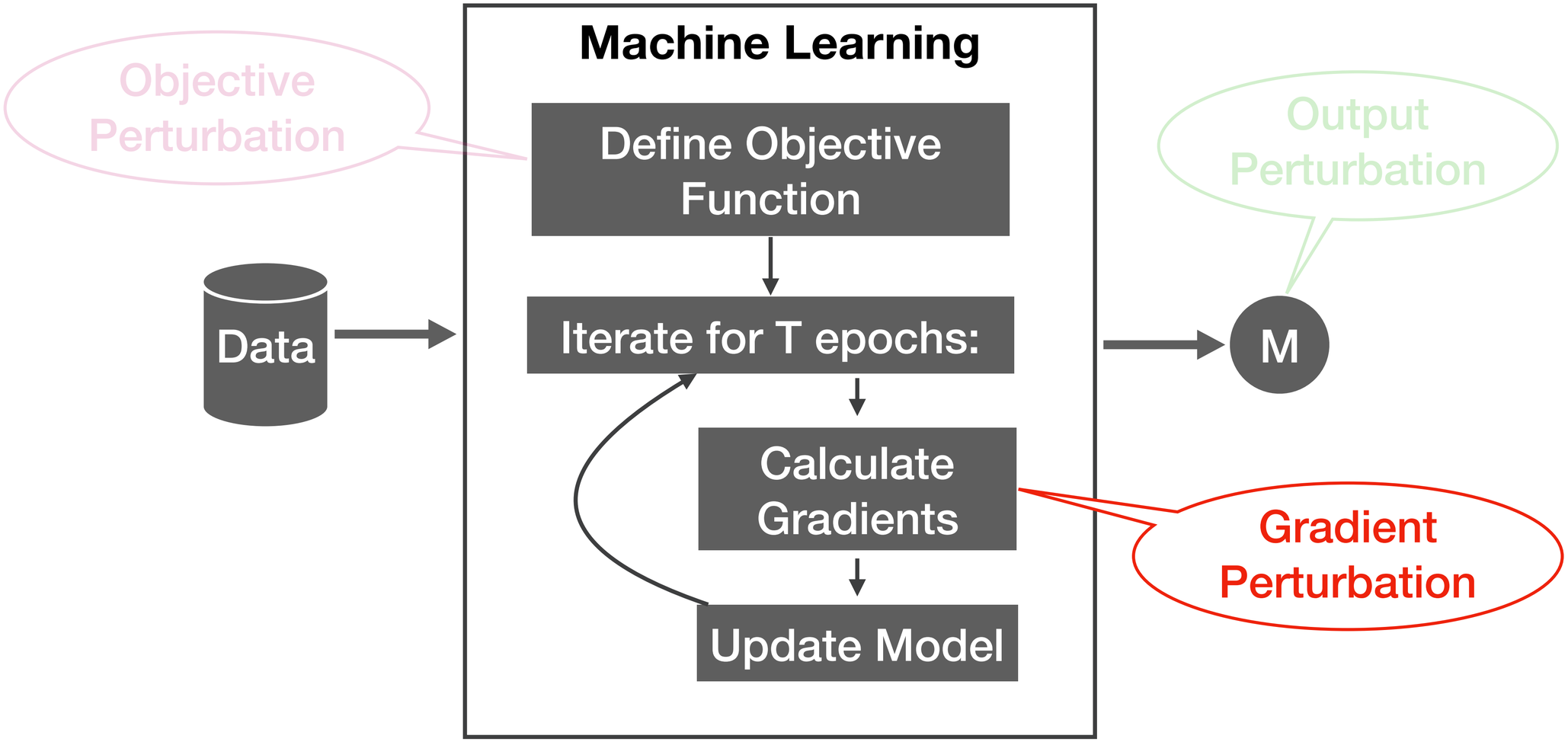
ERM Algorithms using $\epsilon \leq 1$



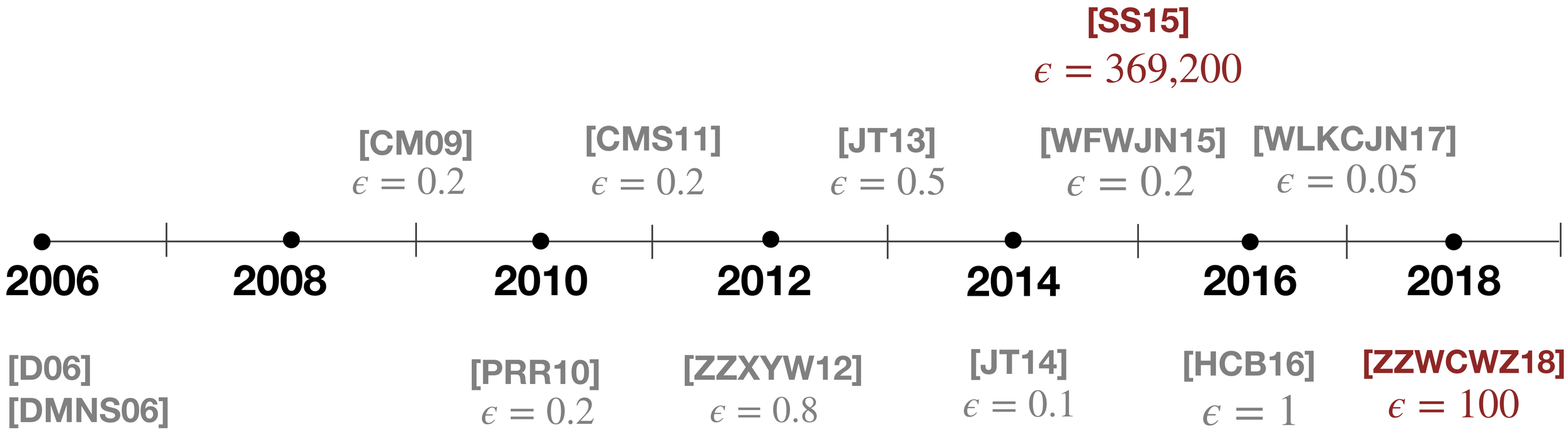
ERM Algorithms using $\epsilon \leq 1$



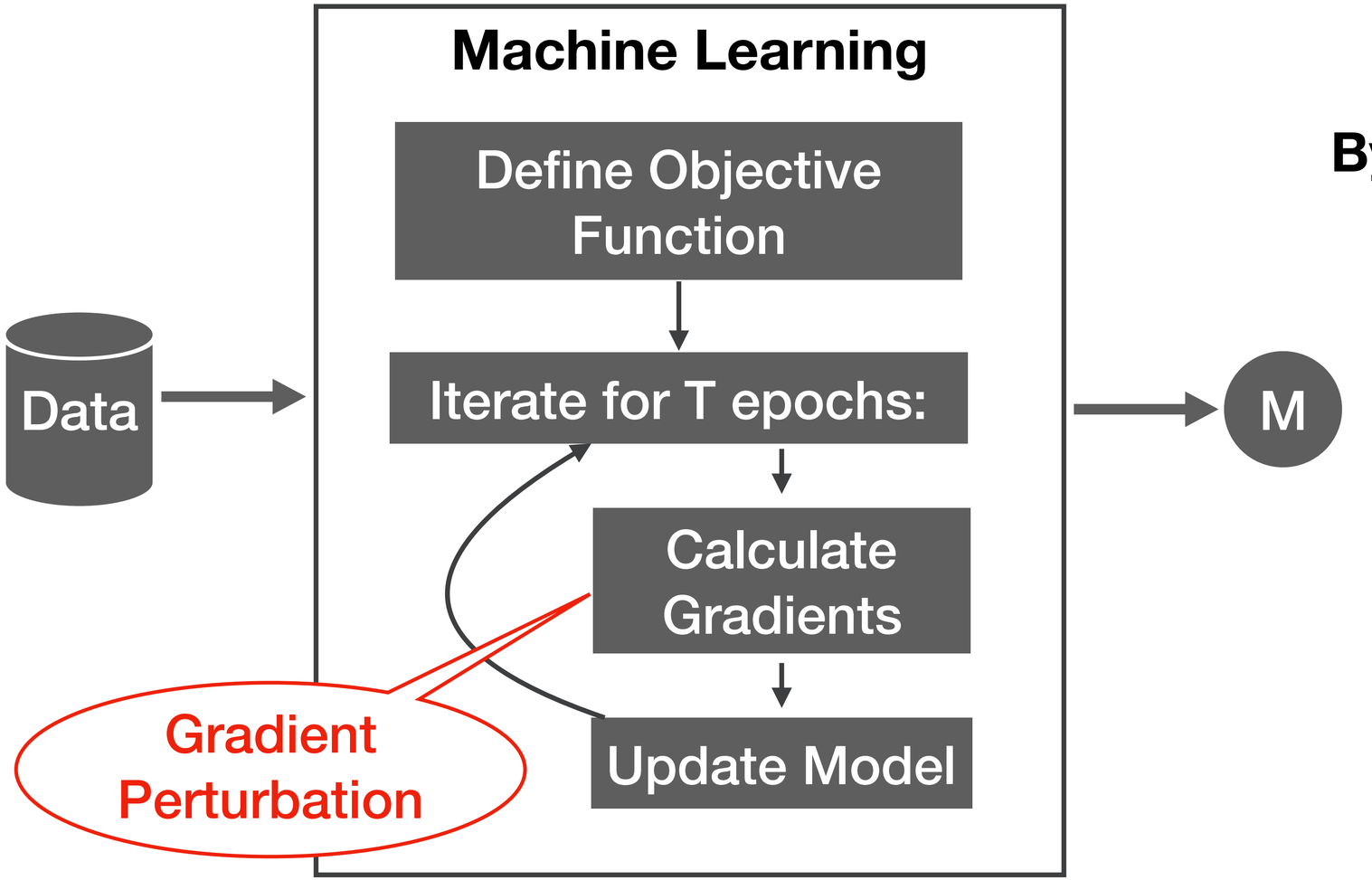
Applying DP to Deep Learning



Deep Learning requiring high ϵ value



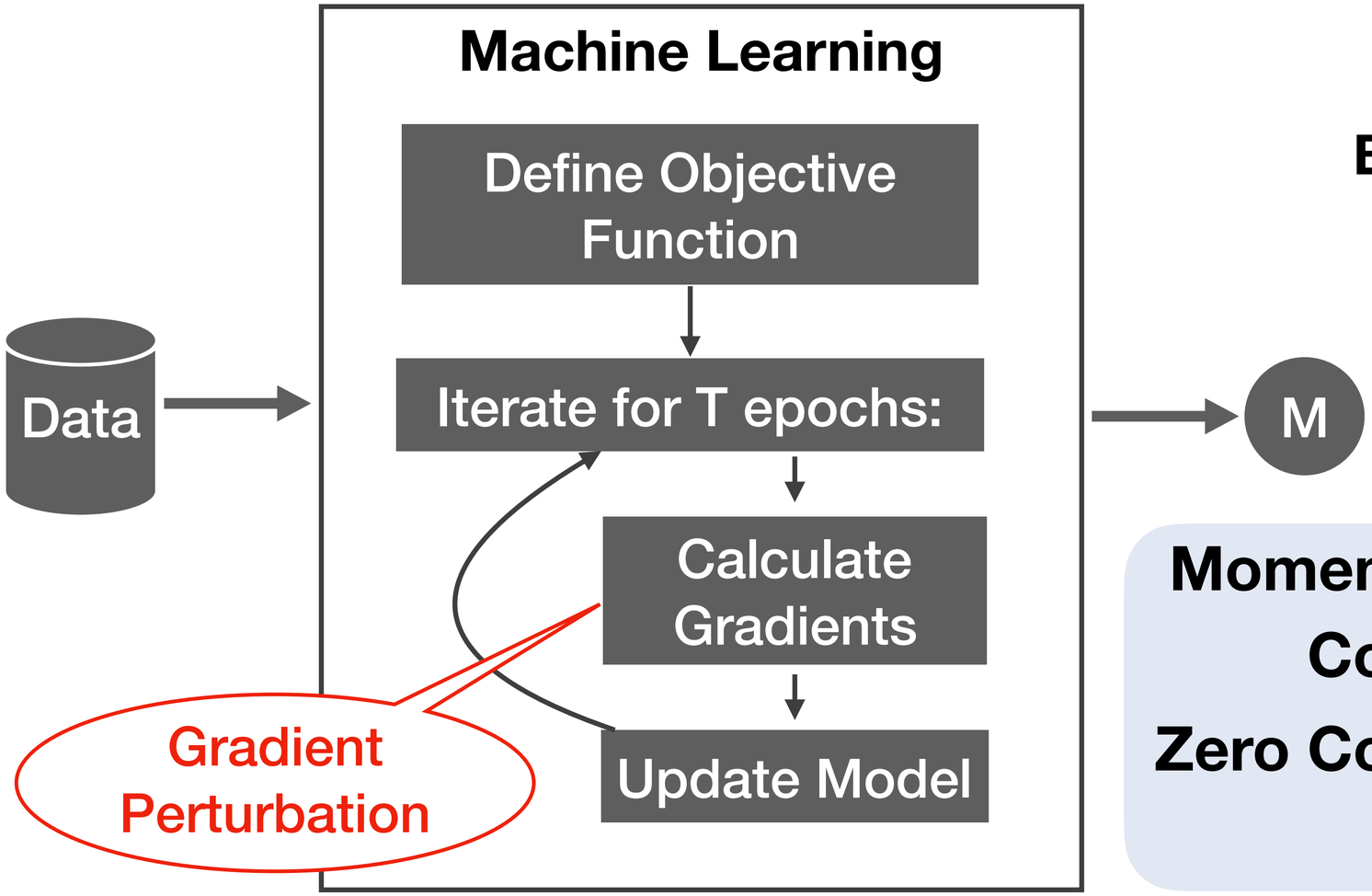
Improving Composition



If each iteration is ϵ -DP

By composition, model: $T\epsilon$ -DP

Improving Composition



If each iteration is ϵ -DP

By composition, model: $T\epsilon$ -DP

Model is: $\left(O\left(\sqrt{T}\epsilon\right), \delta\right)$ -DP

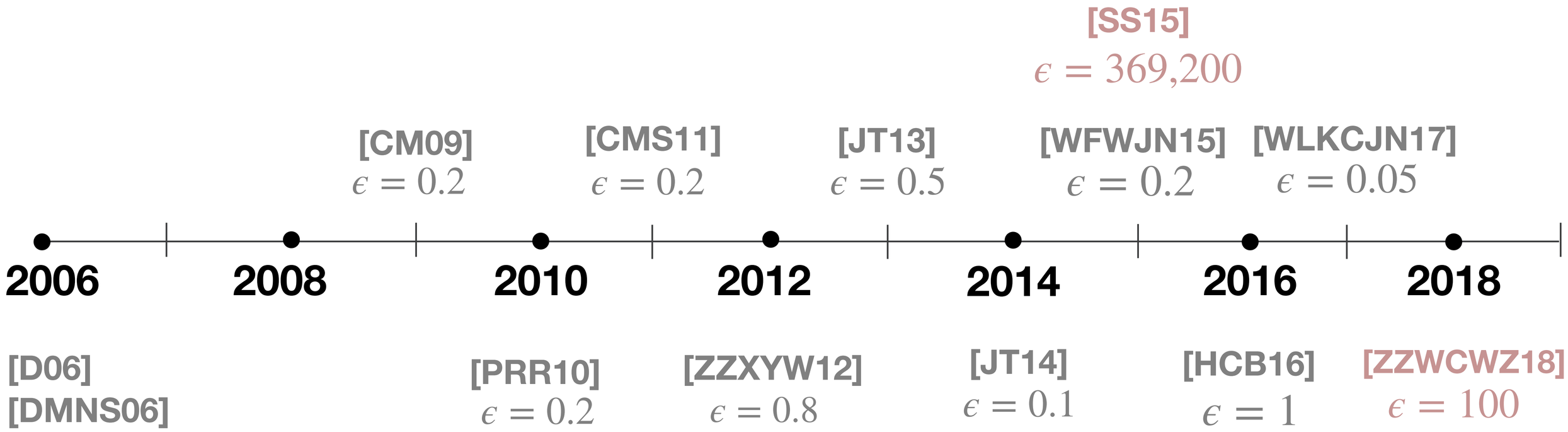
Moments Accountant [Abadi et al. (2016)]

Concentrated DP [Dwork et al. (2016)]

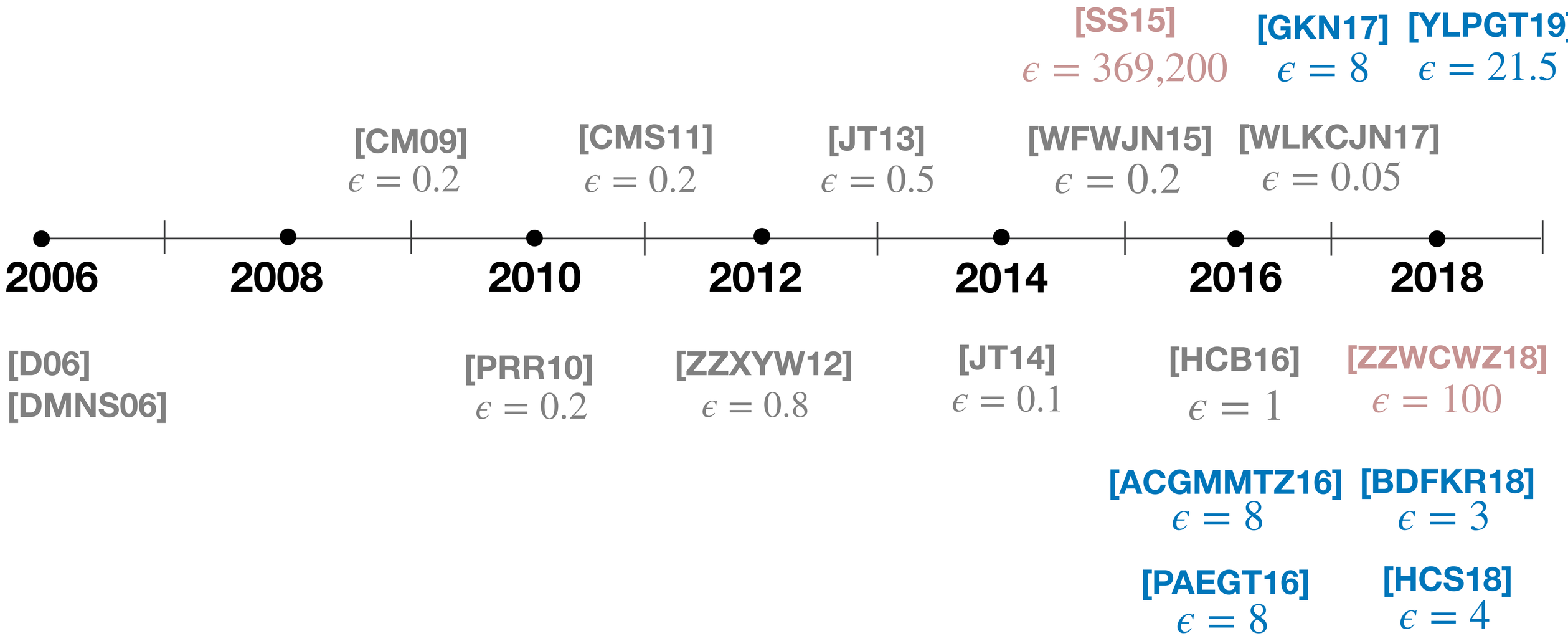
Zero Concentrated DP [Bun & Steinke (2016)]

Rènyi DP [Mironov (2017)]

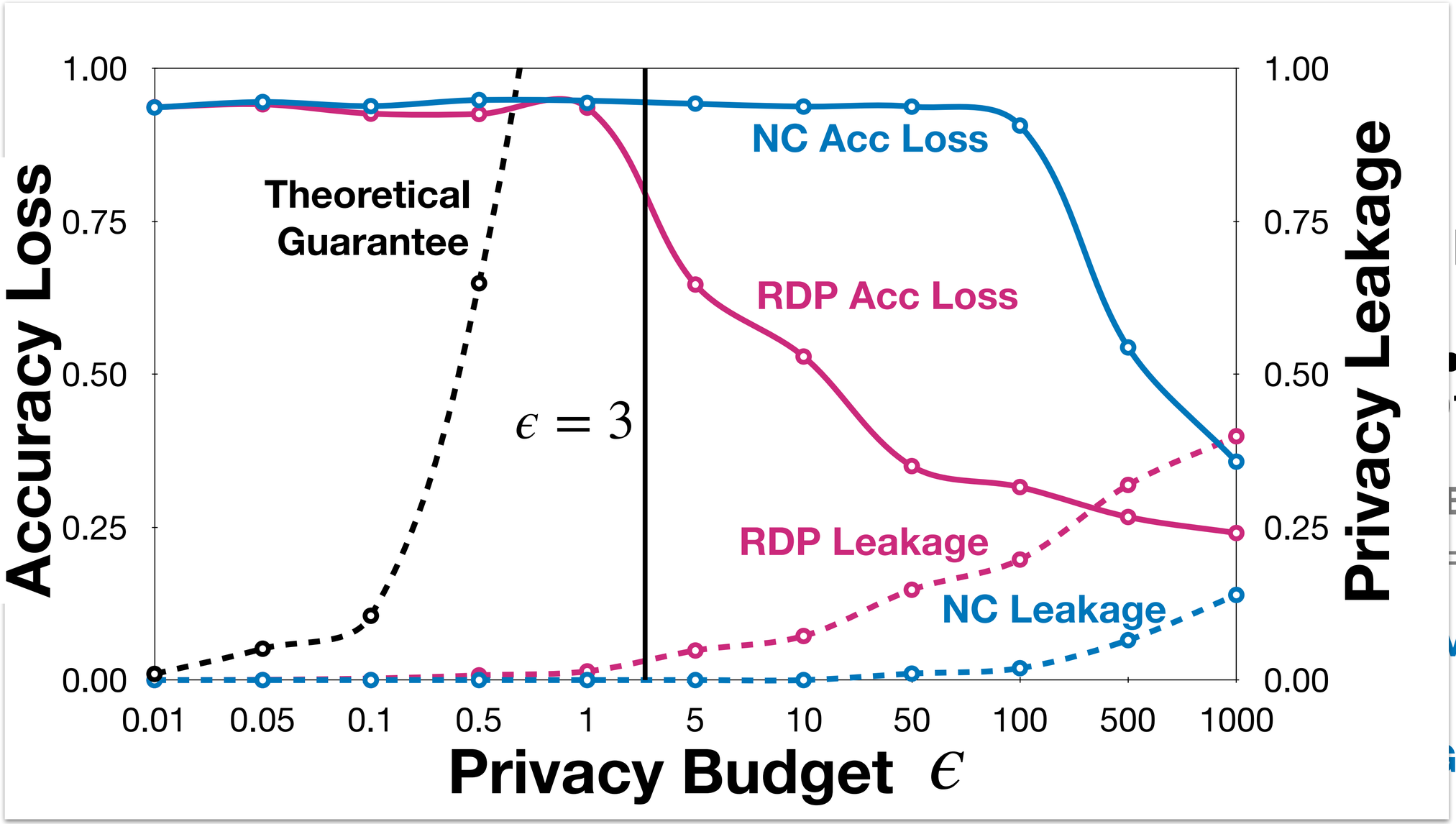
Lower ϵ value with recent DP notions



Lower ϵ value with recent DP notions



Lower ϵ value with recent DP notions



[GKN17] $\epsilon = 8$

[YLPGT19] $\epsilon = 21.5$

[WLKCJN17] $\epsilon = 0.05$

16 2018

B16] $= 1$

[ZZWCWZ18] $\epsilon = 100$

MTZ16] 8

[BDFKR18] $\epsilon = 3$

GT16] 8

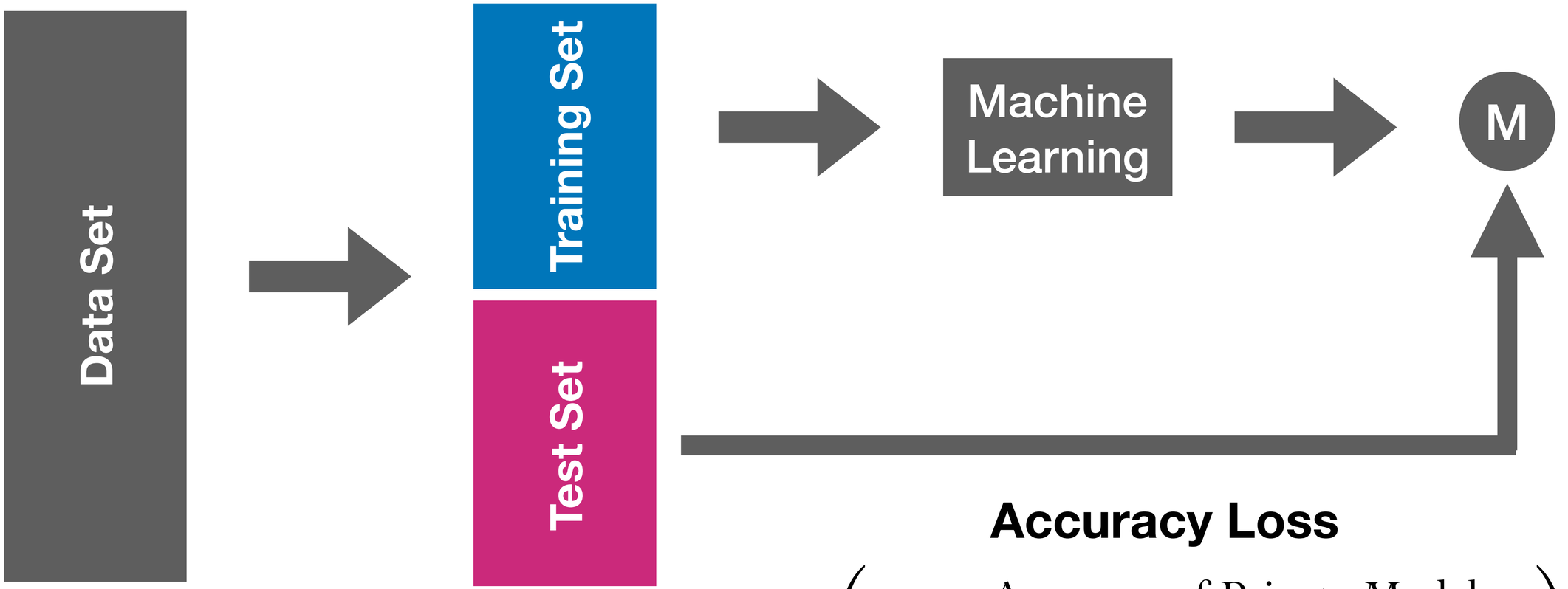
[HCS18] $\epsilon = 4$

Experiments

Model	Task	Evaluation Metric
Logistic Regression	100 class classification on CIFAR-100	Accuracy Loss
Neural Network	100 class classification on Purchase-100	Privacy Leakage

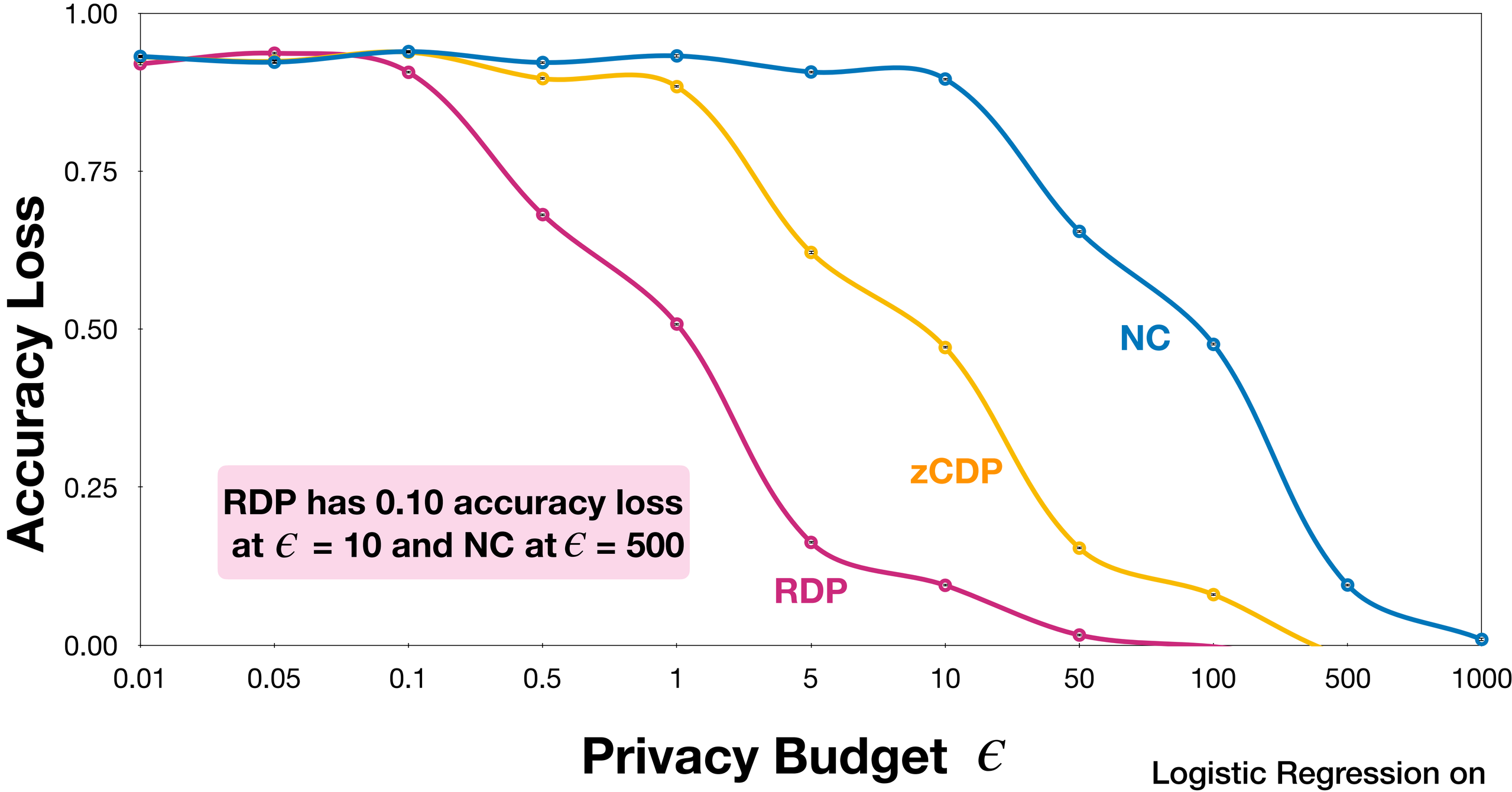
Code Available: https://github.com/bargavj/EvaluatingDPML

Training and Testing

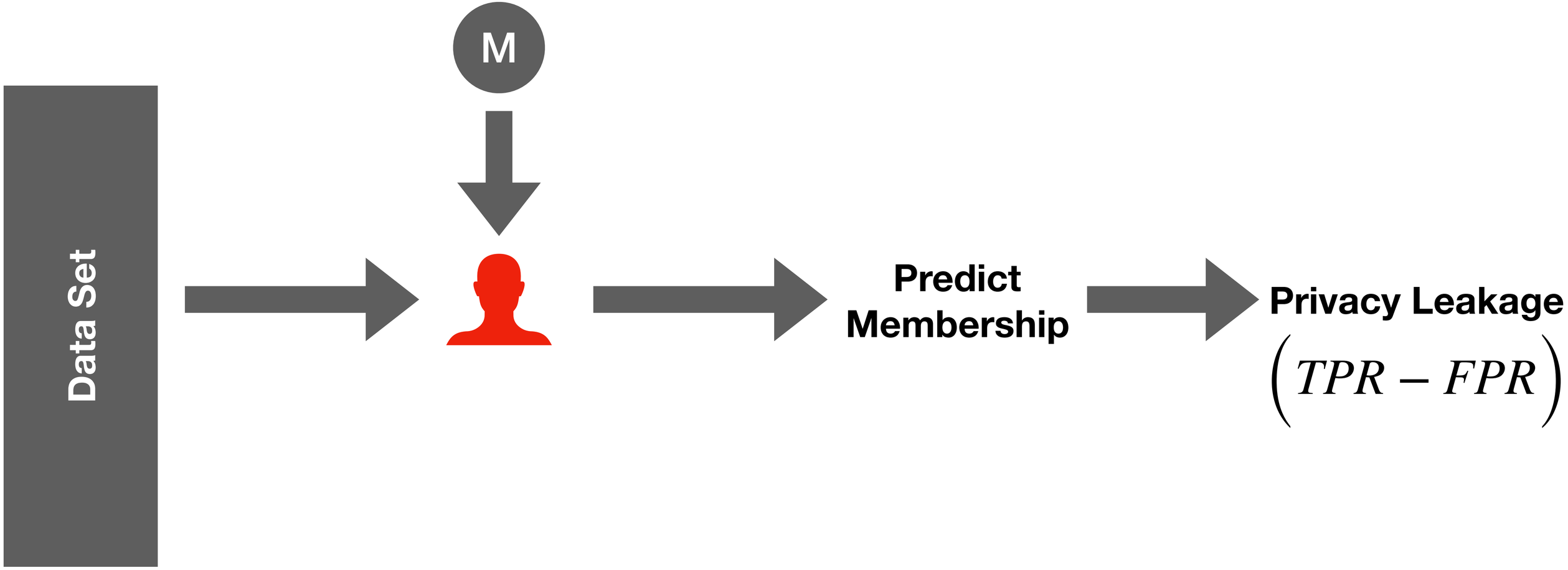


Accuracy Loss

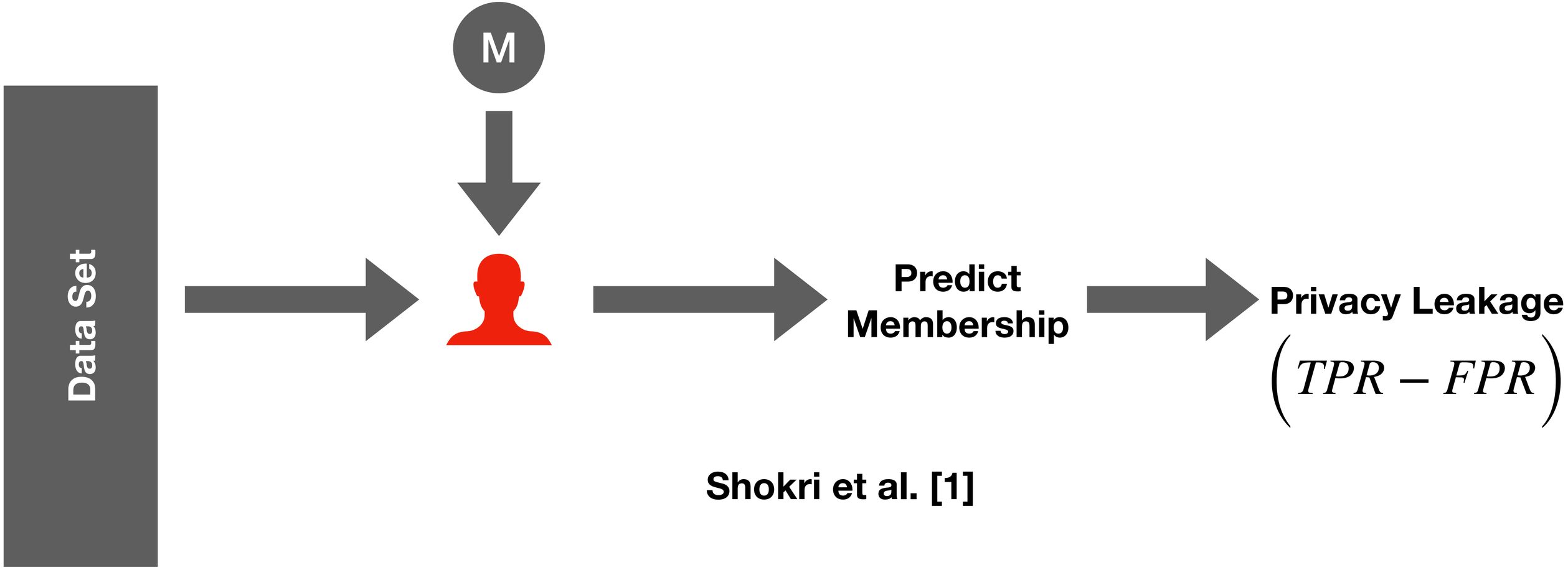
$$\left(1 - \frac{\text{Accuracy of Private Model}}{\text{Accuracy of Non-Private Model}}\right)$$


1.00
0.75
0.50
0.25
0.00
Accuracy Loss
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
NC
zCDP
RDP
RDP has 0.10 accuracy loss at ϵ = 10 and NC at ϵ = 500
Logistic Regression on CIFAR-100

Membership Inference Attacks



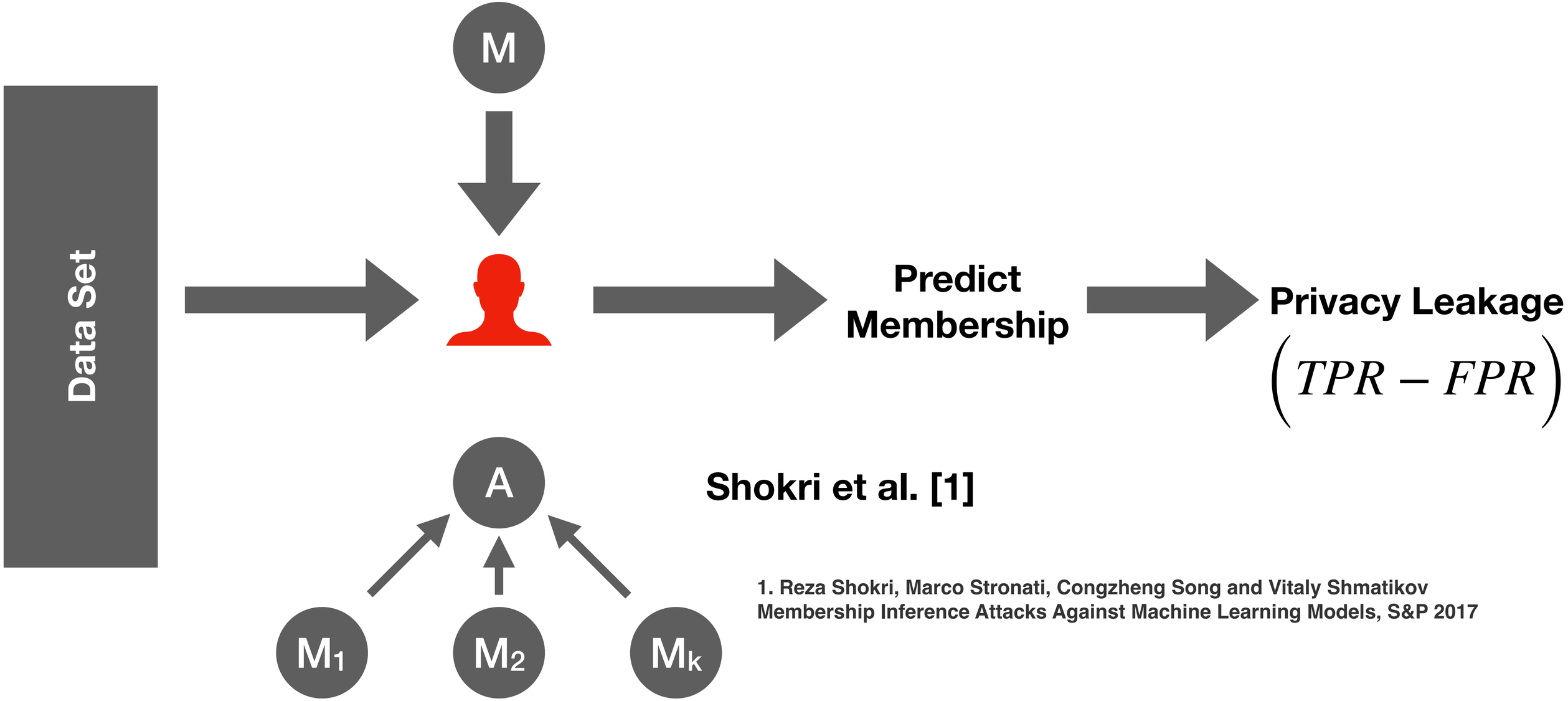
Membership Inference Attacks



Shokri et al. [1]

1. Reza Shokri, Marco Stronati, Congzheng Song and Vitaly Shmatikov
Membership Inference Attacks Against Machine Learning Models, S&P 2017

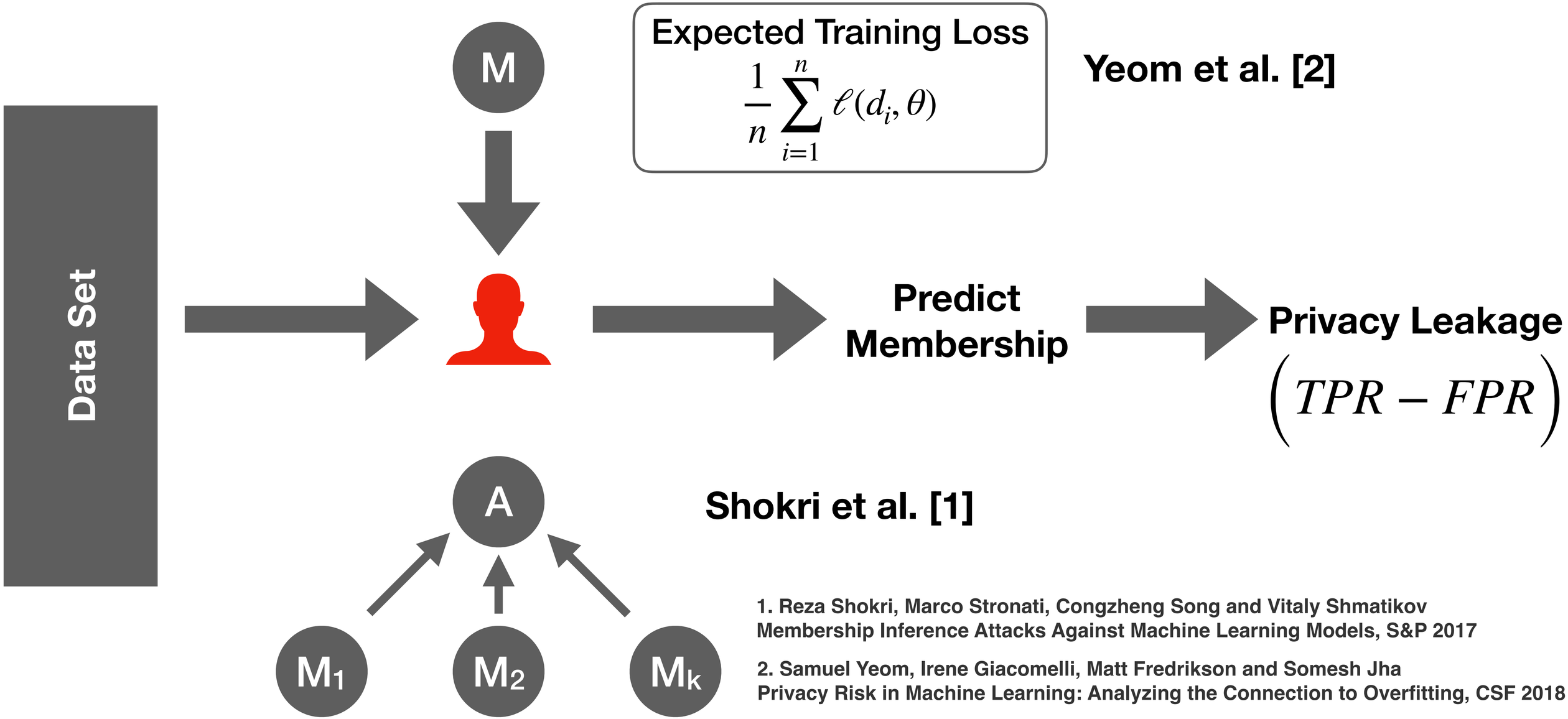
Membership Inference Attacks



Shokri et al. [1]

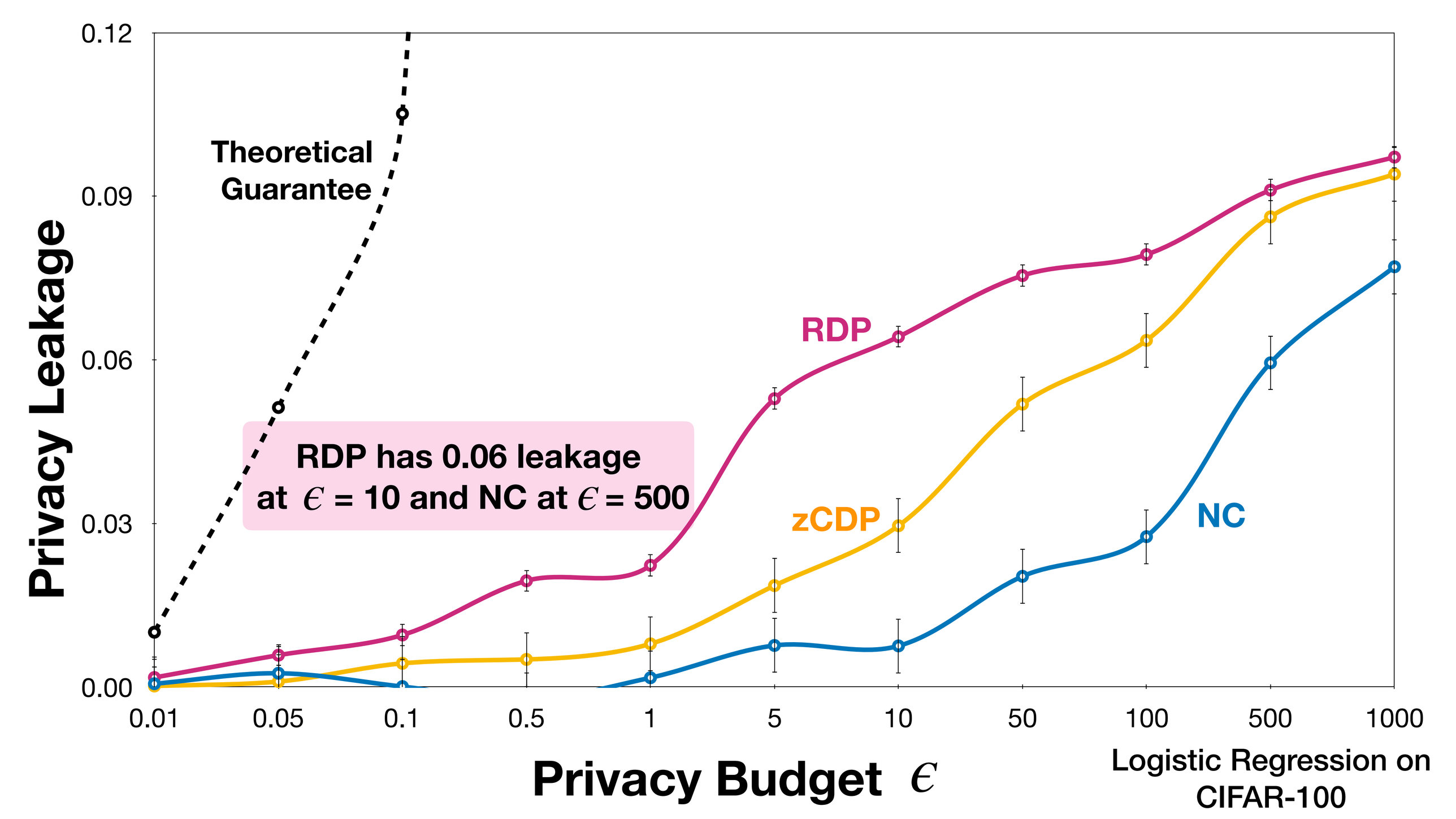
1. Reza Shokri, Marco Stronati, Congzheng Song and Vitaly Shmatikov
Membership Inference Attacks Against Machine Learning Models, S&P 2017

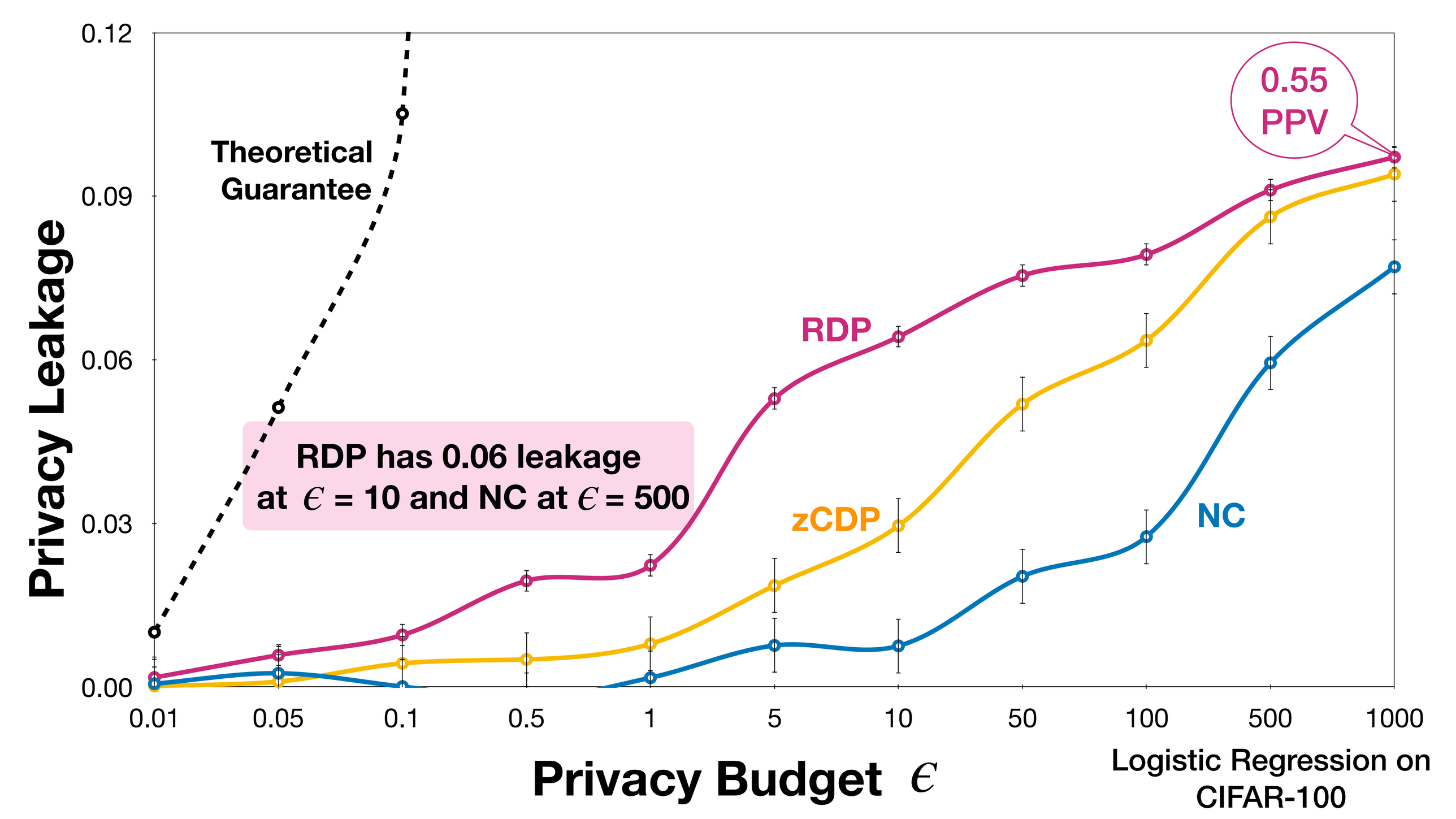
Membership Inference Attacks

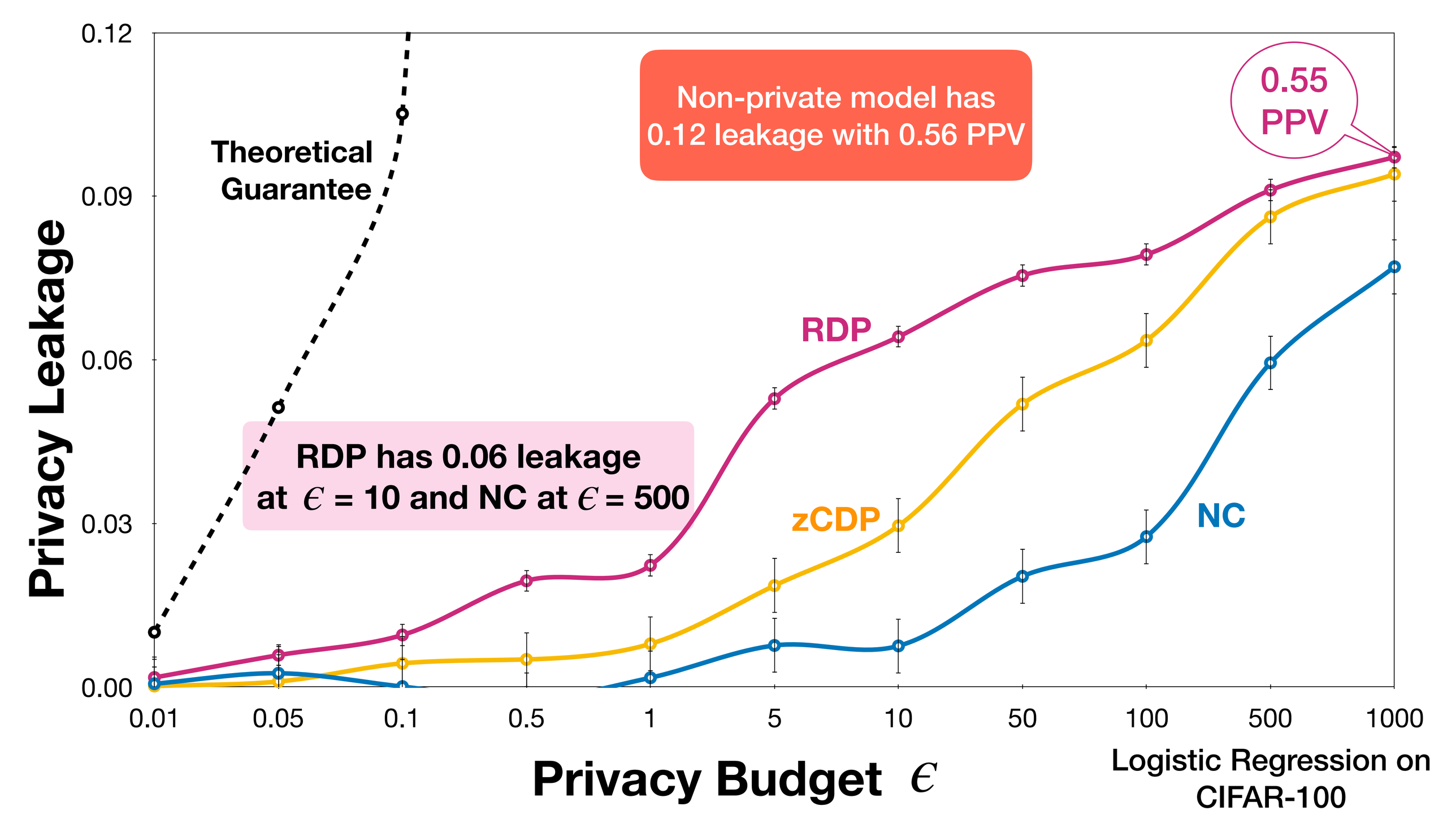


1. Reza Shokri, Marco Stronati, Congzheng Song and Vitaly Shmatikov
Membership Inference Attacks Against Machine Learning Models, S&P 2017

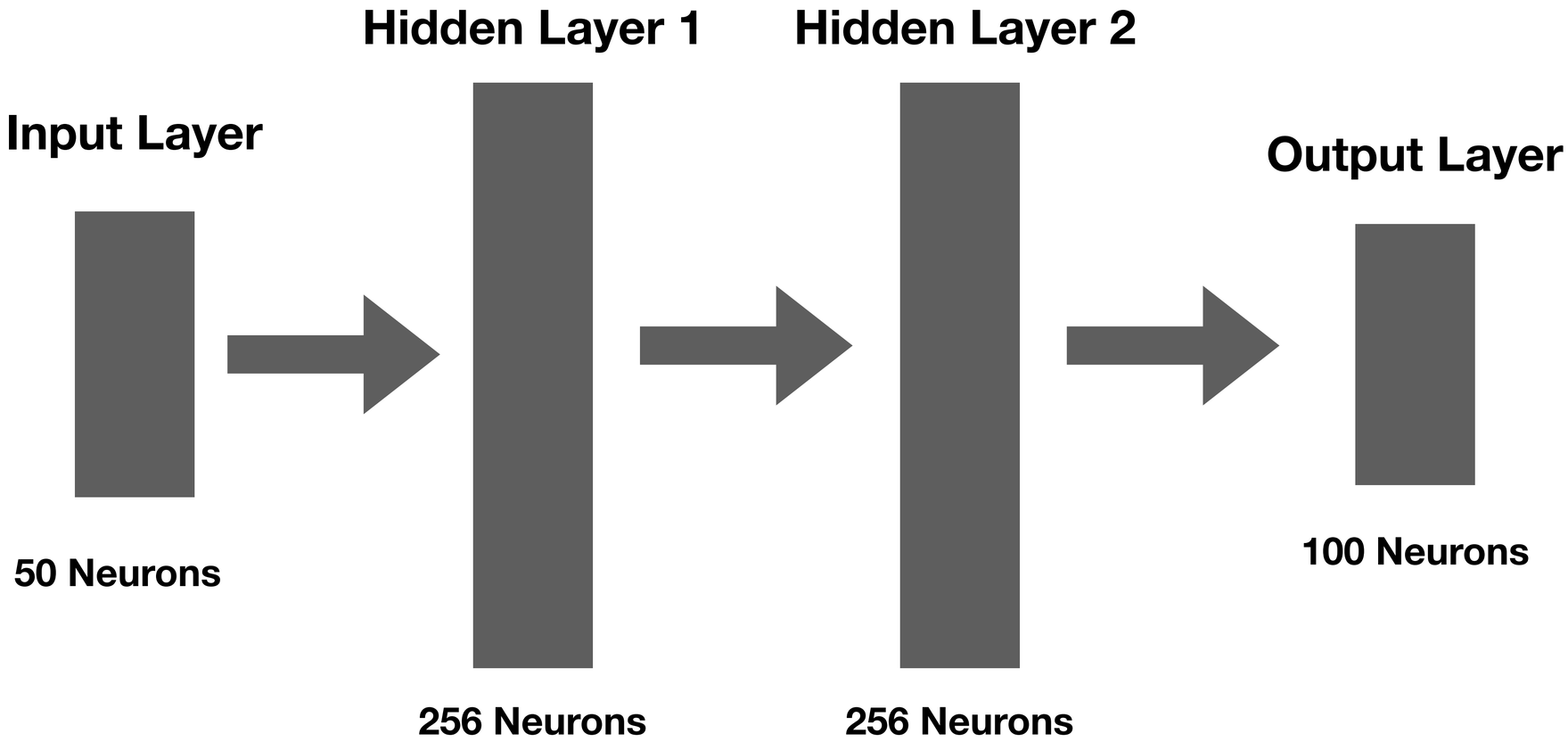
2. Samuel Yeom, Irene Giacomelli, Matt Fredrikson and Somesh Jha
Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, CSF 2018


0.12
0.09
0.06
0.03
0.00
Privacy Leakage
Theoretical Guarantee
RDP has 0.06 leakage at ϵ = 10 and NC at ϵ = 500
RDP
zCDP
NC
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Logistic Regression on CIFAR-100

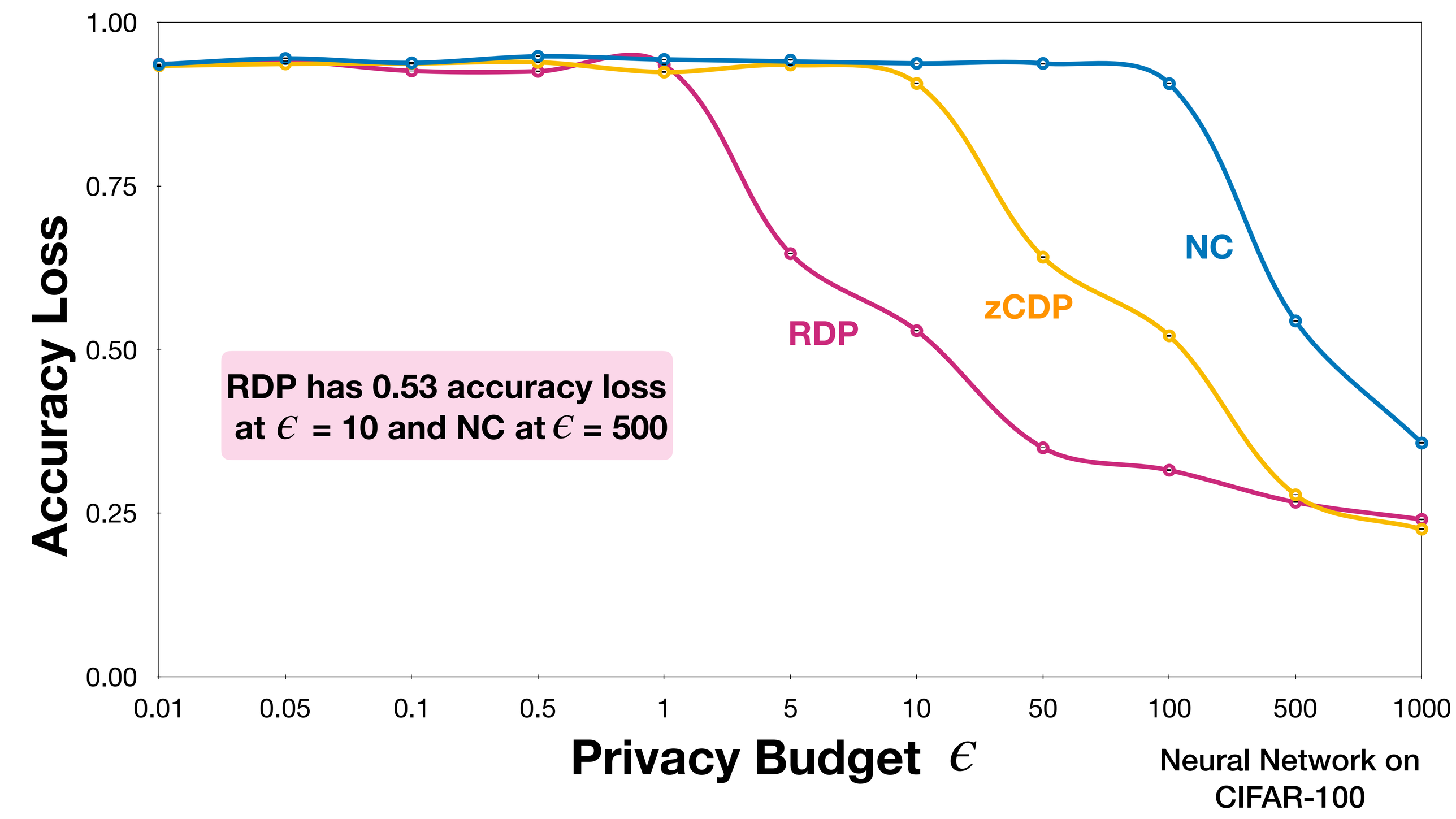

0.12
0.09
0.06
0.03
0.00
Privacy Leakage
Theoretical Guarantee
RDP has 0.06 leakage at ϵ = 10 and NC at ϵ = 500
RDP
zCDP
NC
0.55 PPV
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Logistic Regression on CIFAR-100

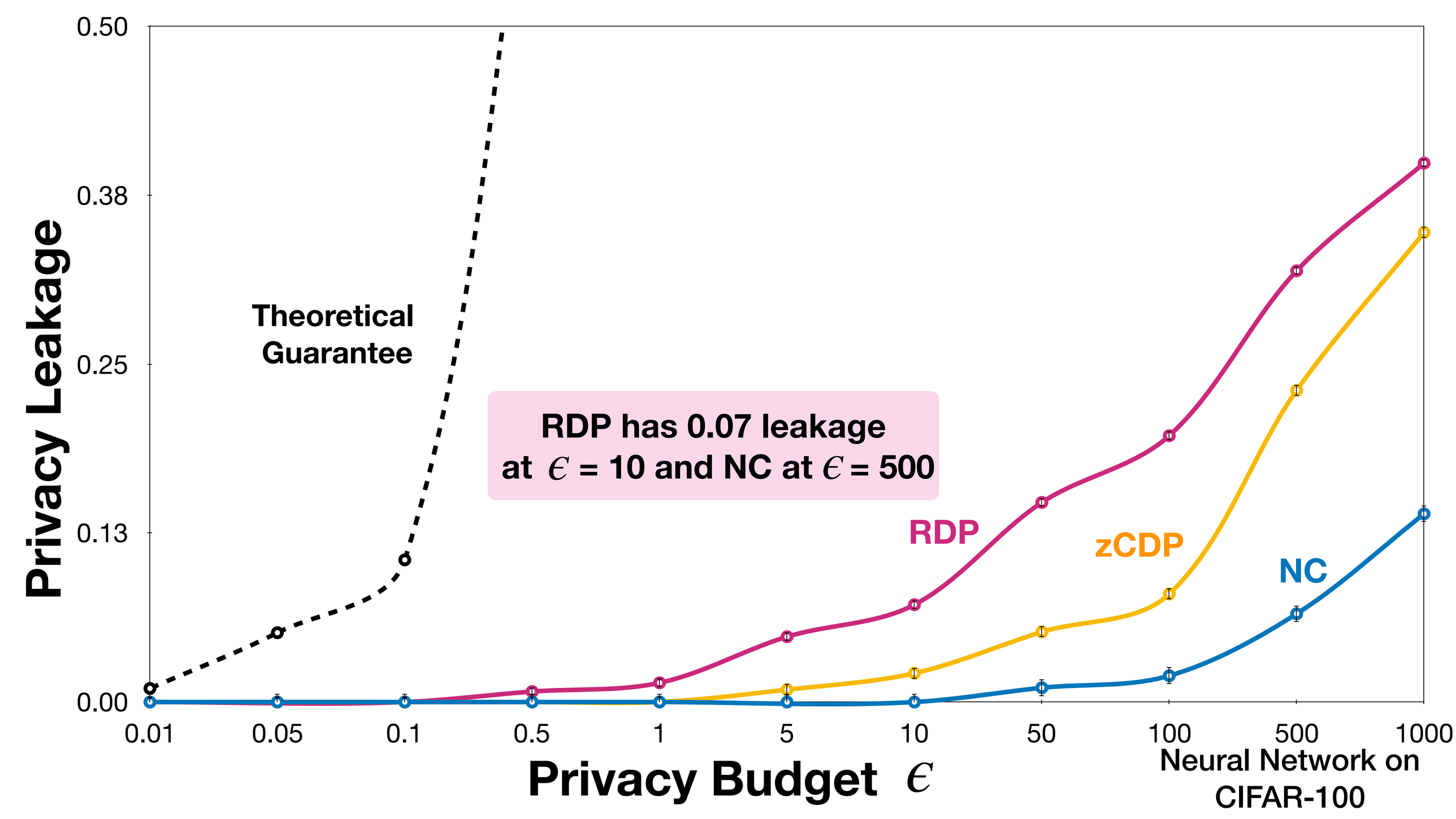

0.12
0.09
0.06
0.03
0.00
Privacy Leakage
Theoretical Guarantee
Non-private model has 0.12 leakage with 0.56 PPV
0.55 PPV
RDP
zCDP
NC
RDP has 0.06 leakage at ϵ = 10 and NC at ϵ = 500
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Logistic Regression on CIFAR-100

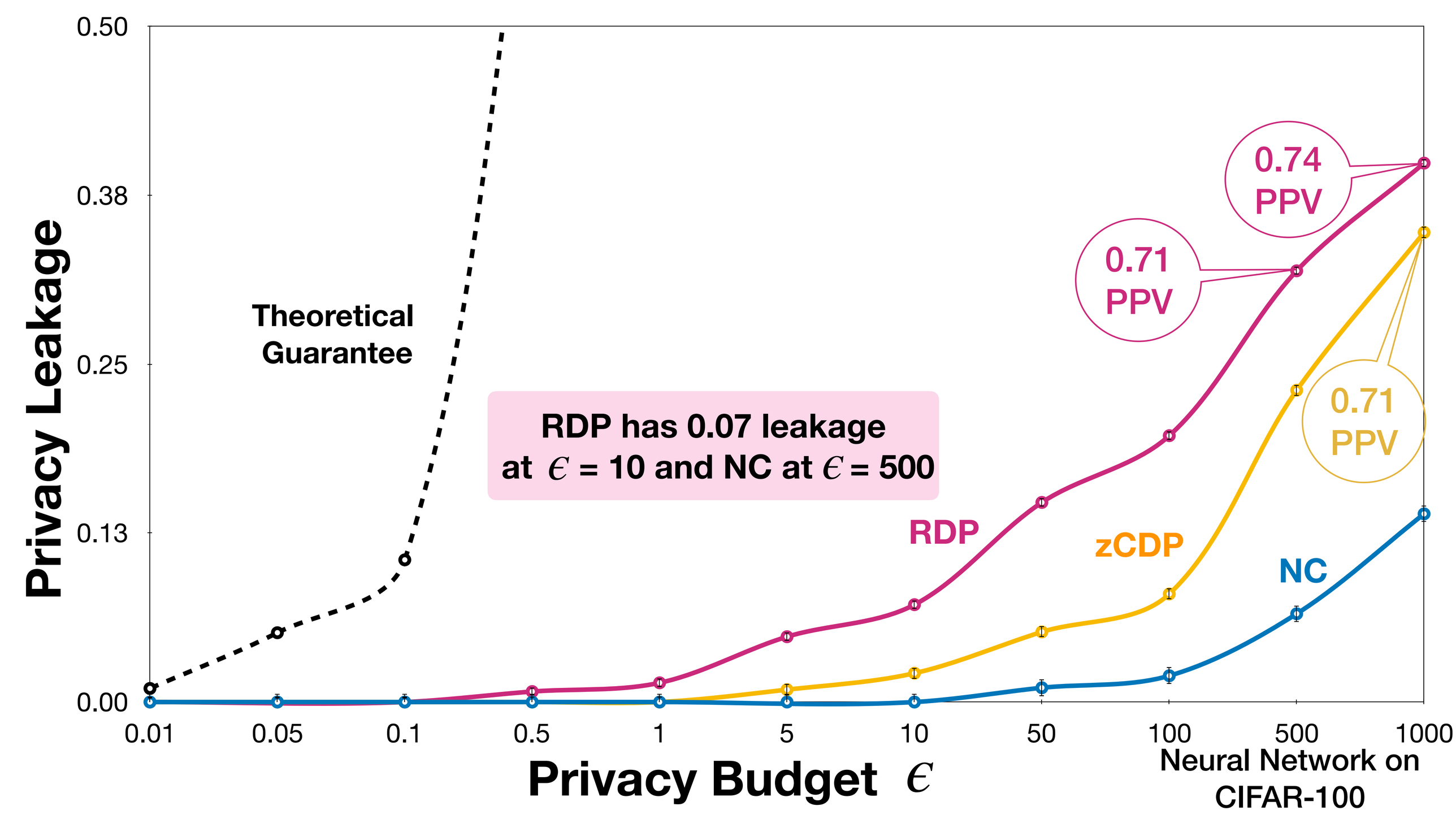
Neural Networks

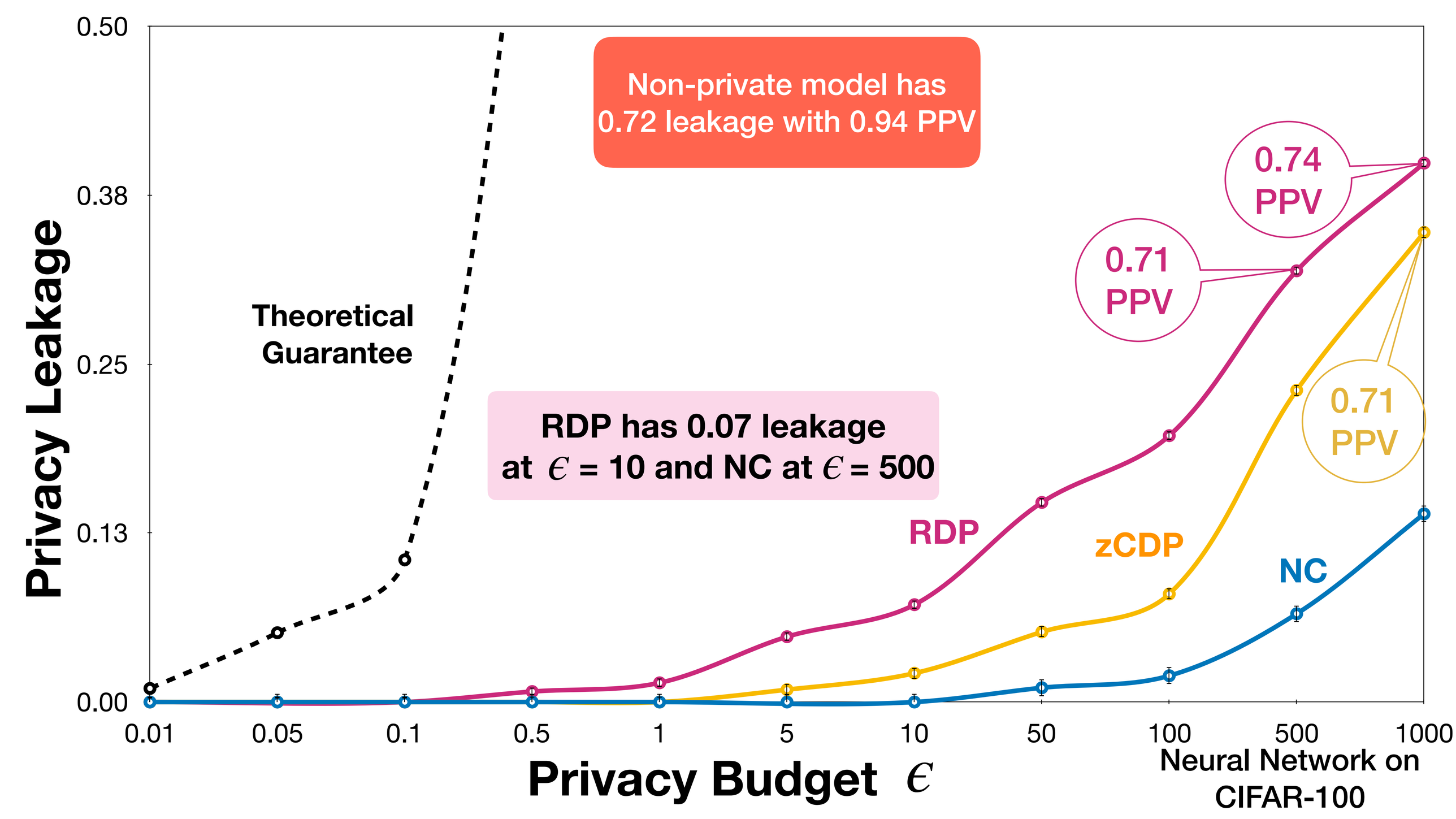


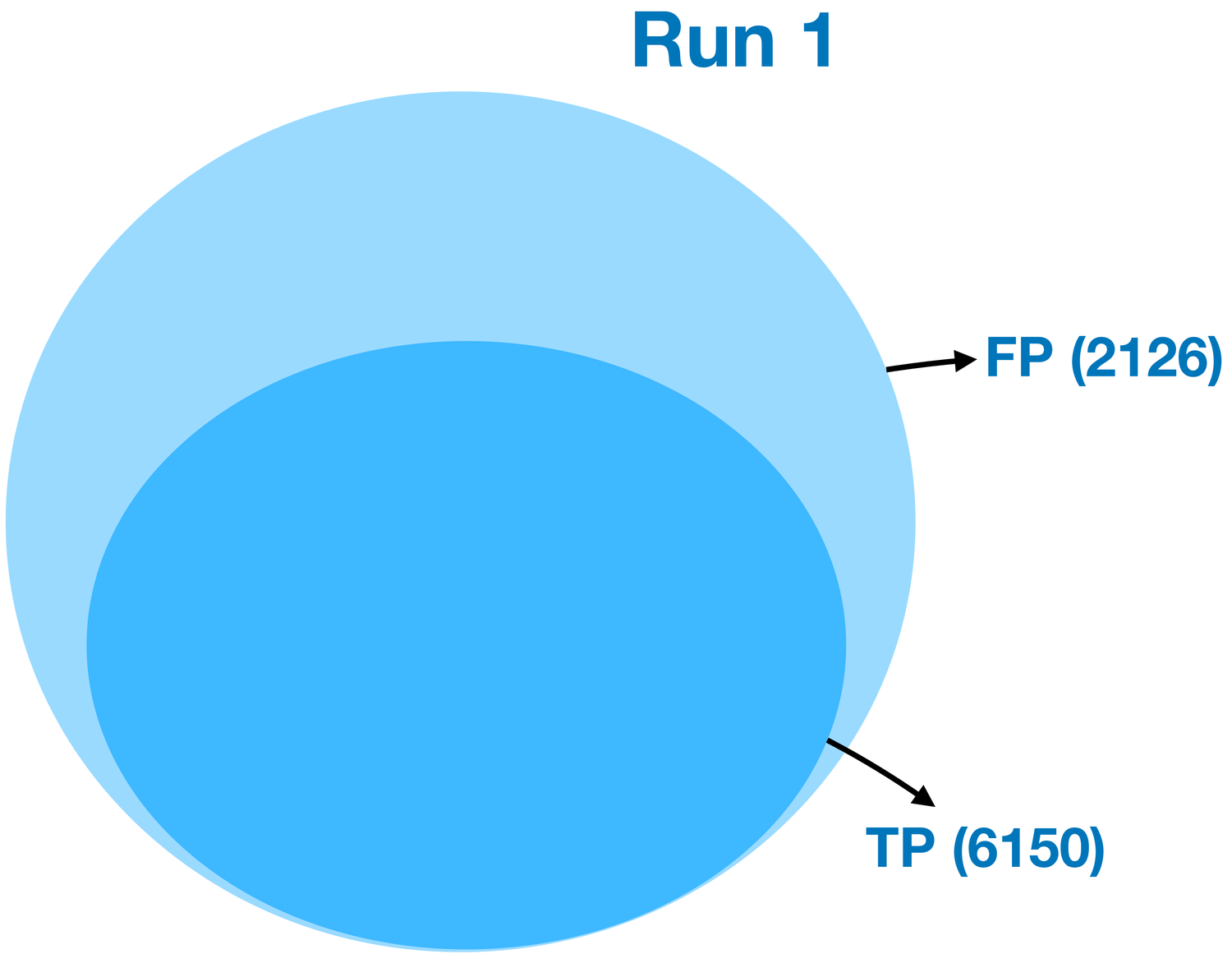
NN has 103,936 trainable parameters so it has more capacity to learn on training data


1.00
0.75
0.50
0.25
0.00
Accuracy Loss
NC
zCDP
RDP
RDP has 0.53 accuracy loss at ϵ = 10 and NC at ϵ = 500
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Neural Network on CIFAR-100


0.50
0.38
0.25
0.13
0.00
Privacy Leakage
Theoretical Guarantee
RDP has 0.07 leakage at ϵ = 10 and NC at ϵ = 500
RDP
zCDP
NC
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Neural Network on CIFAR-100

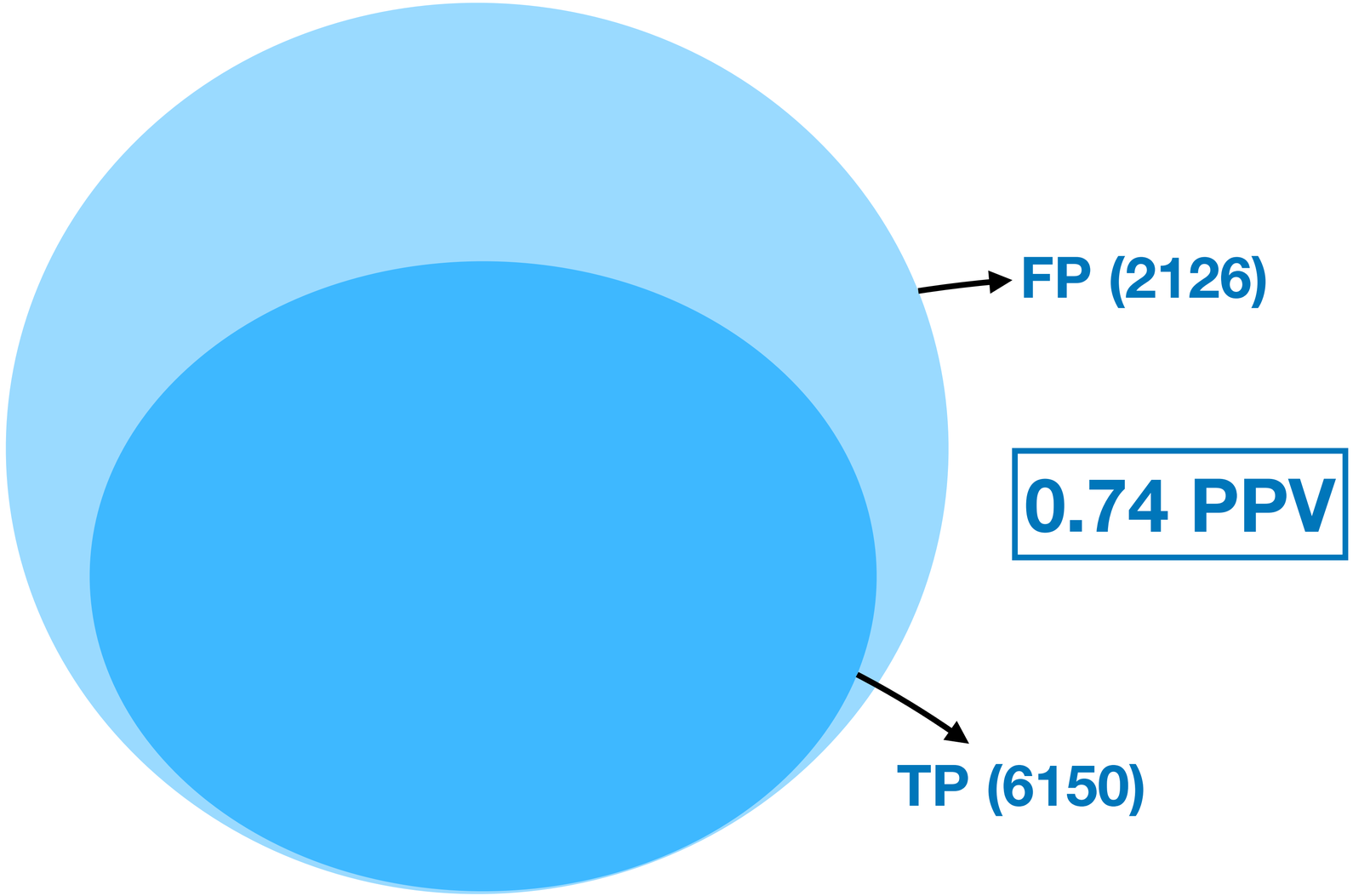

0.50
0.38
0.25
0.13
0.00
Privacy Leakage
Theoretical Guarantee
RDP has 0.07 leakage at ϵ = 10 and NC at ϵ = 500
0.74 PPV
0.71 PPV
0.71 PPV
RDP
zCDP
NC
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Neural Network on CIFAR-100


Non-private model has
0.72 leakage with 0.94 PPV
0.74
PPV
0.71
PPV
0.71
PPV
Theoretical
Guarantee
RDP has 0.07 leakage
at ϵ = 10 and NC at ϵ = 500
RDP
zCDP
NC
Privacy Leakage
0.50
0.38
0.25
0.13
0.00
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Neural Network on
CIFAR-100

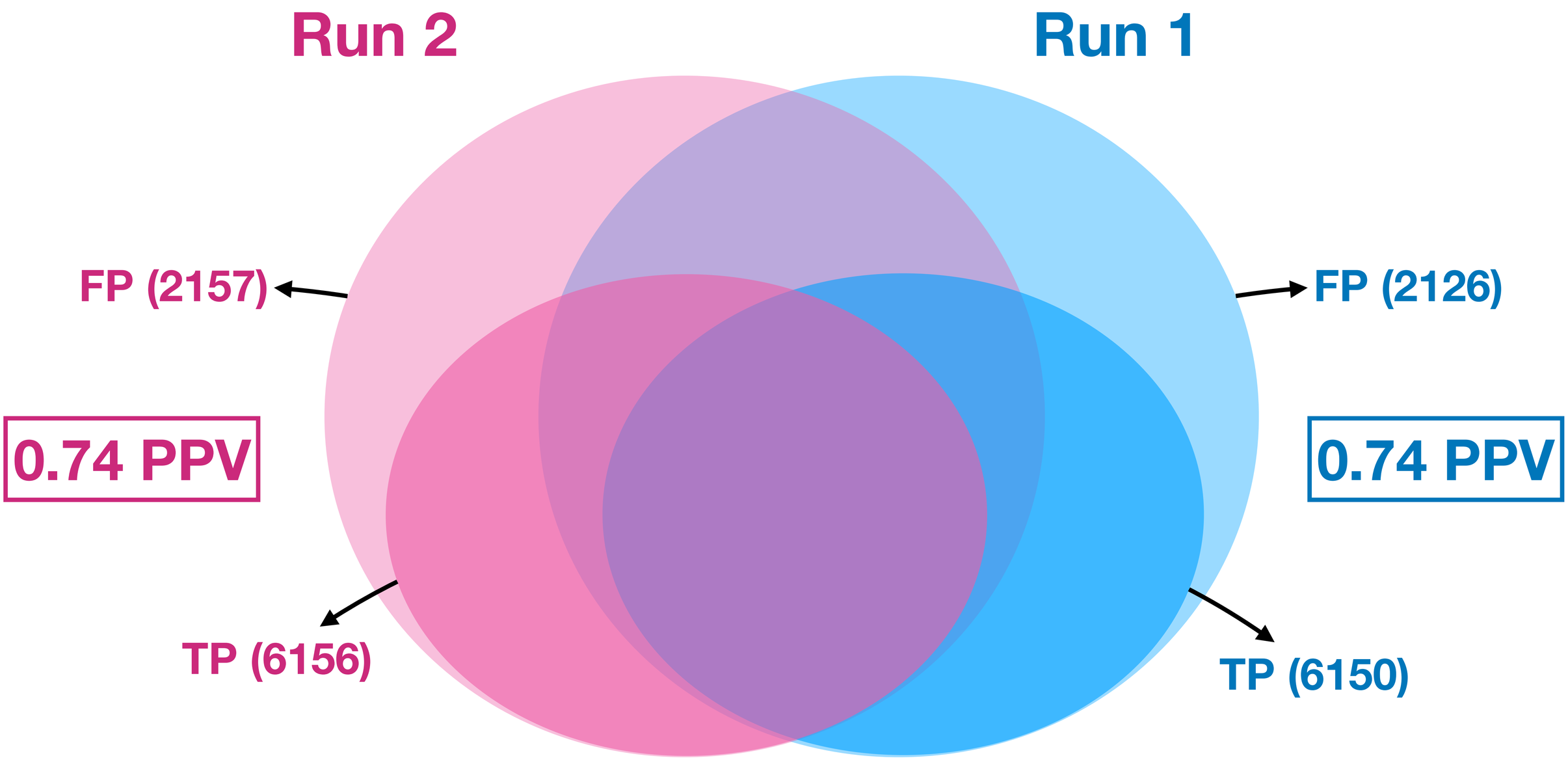


*New results, included in the updated version of the paper

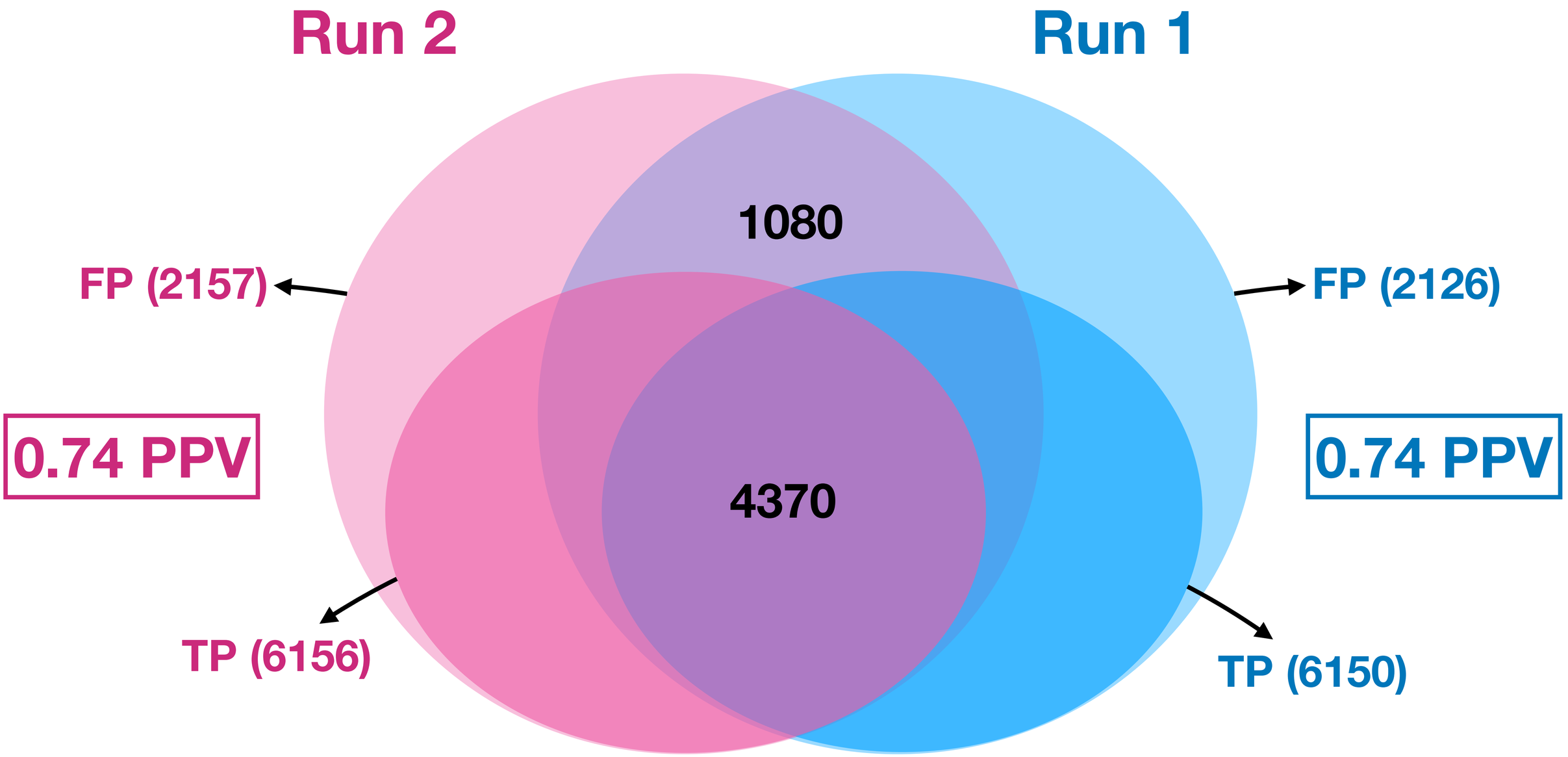
Run 1



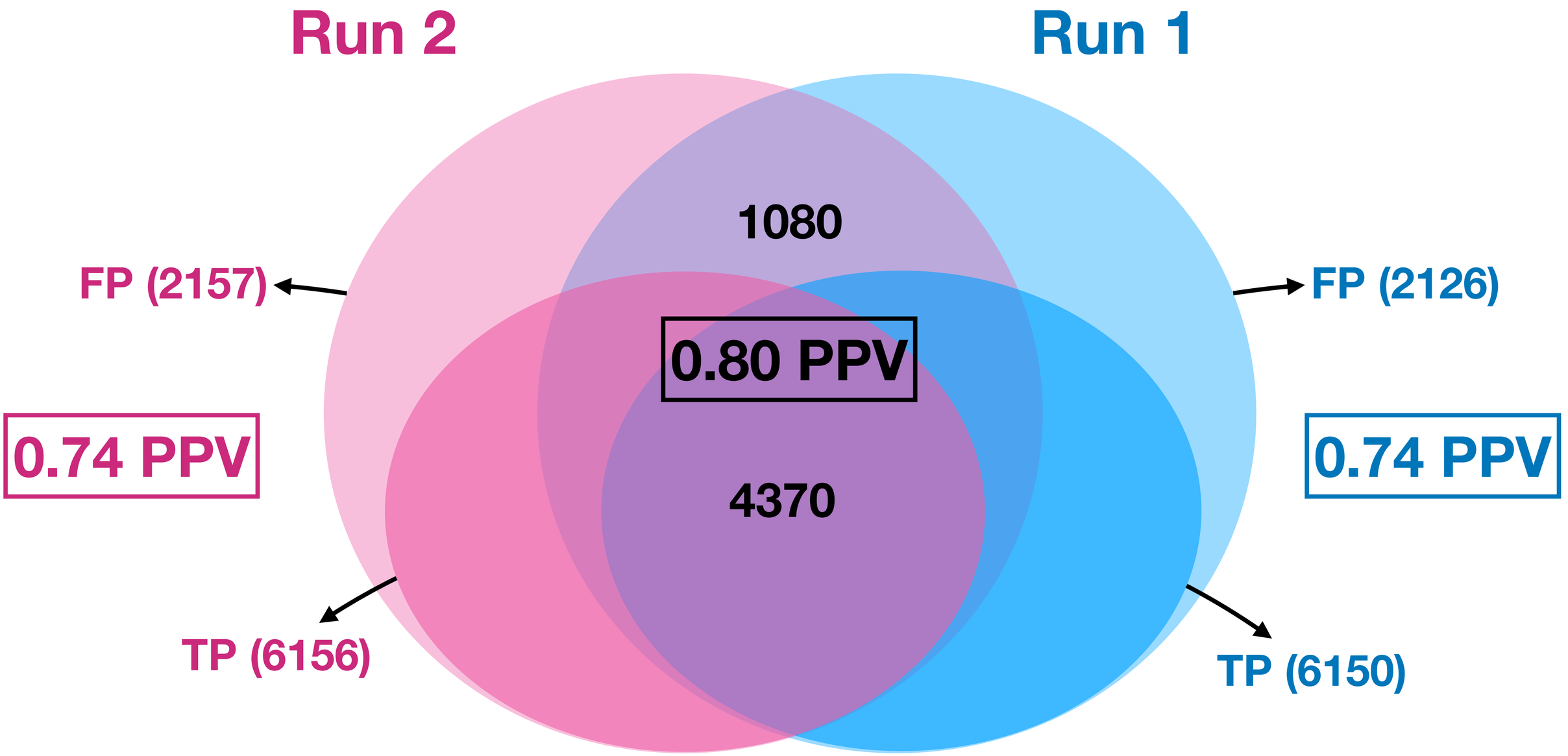
*New results, included in the updated version of the paper



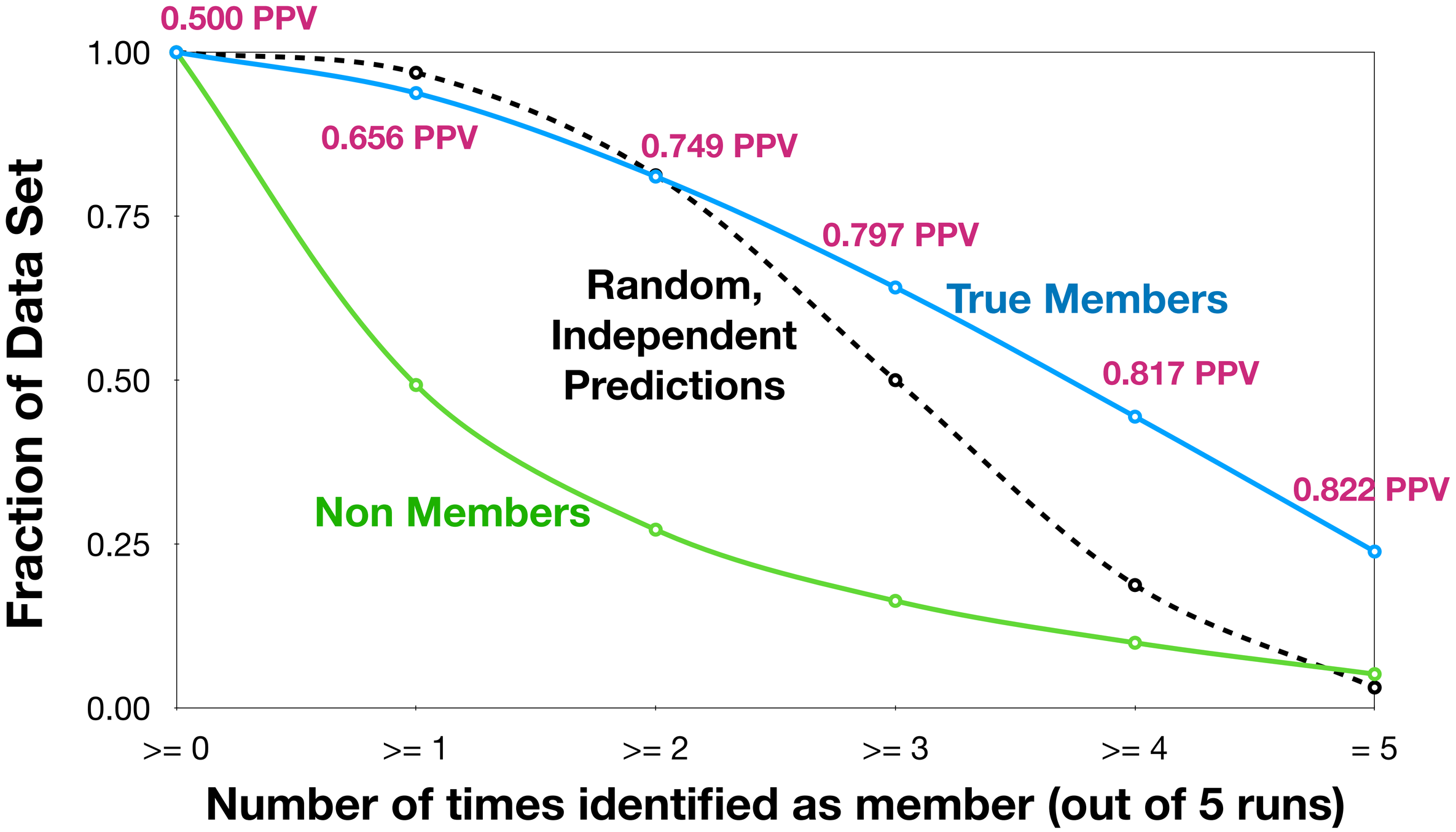
*New results, included in the updated version of the paper



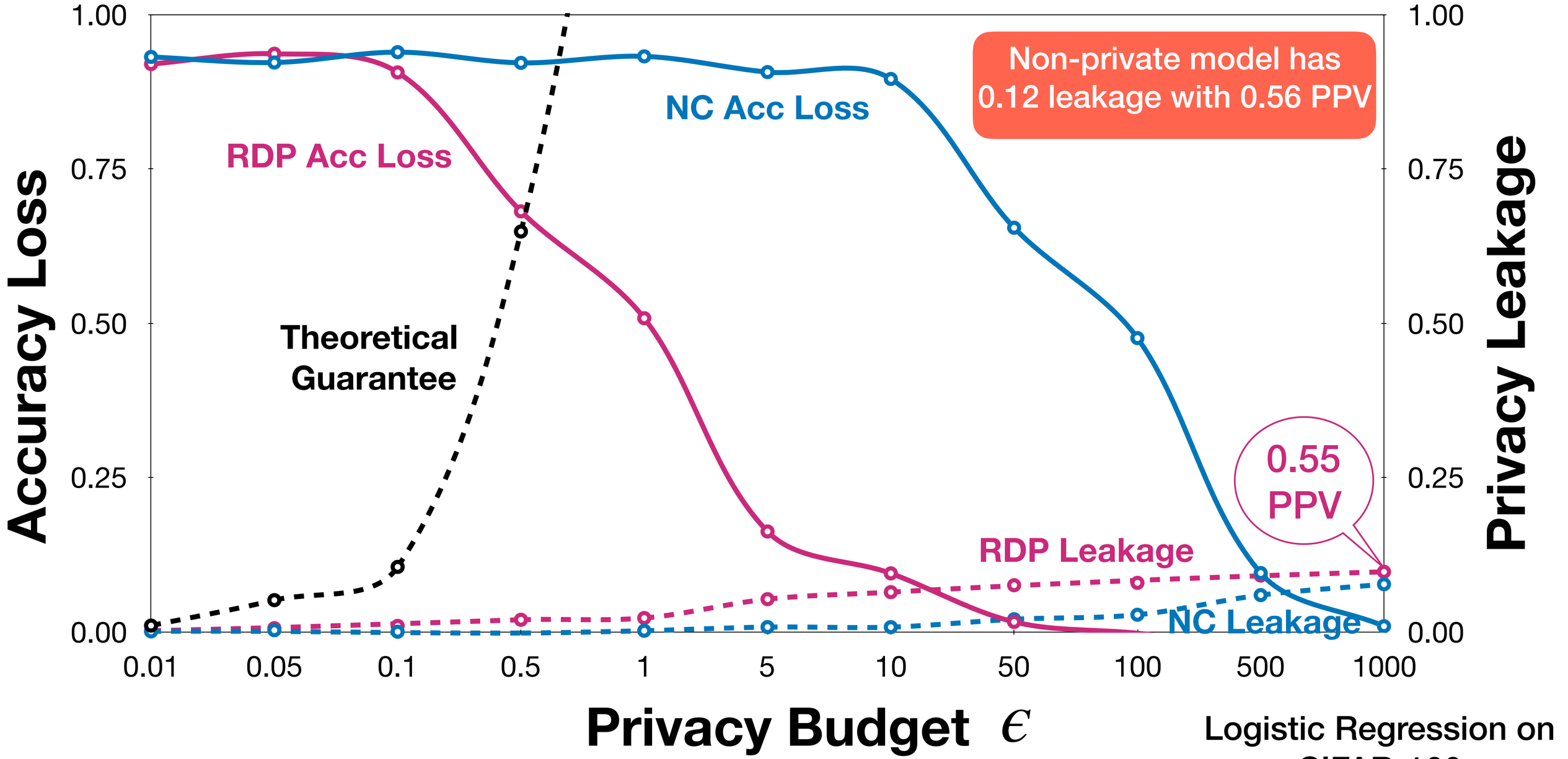
*New results, included in the updated version of the paper



*New results, included in the updated version of the paper

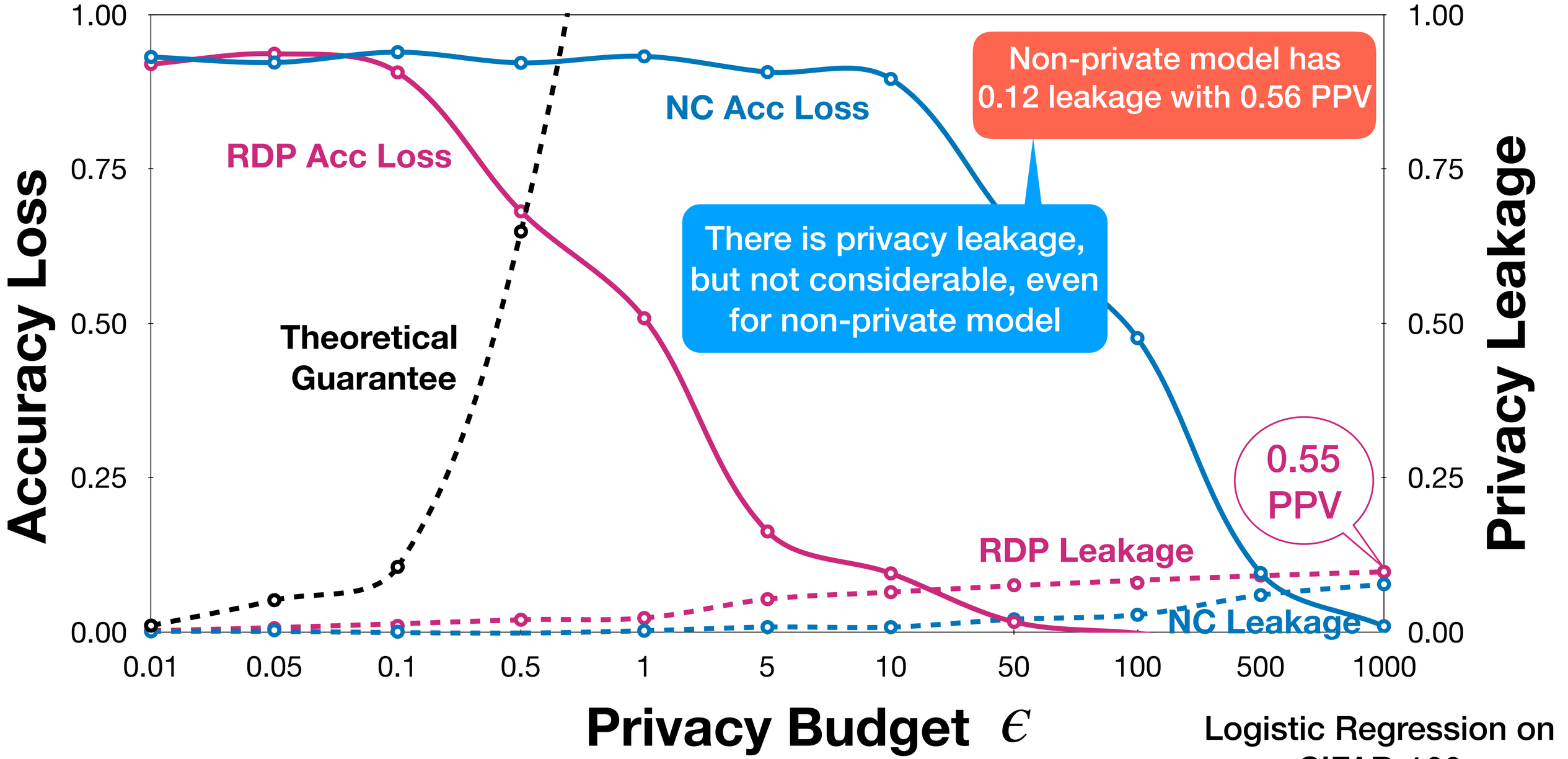

0.500 PPV
1.00
0.656 PPV
0.749 PPV
0.797 PPV
Random,
Independent
Predictions
True Members
0.817 PPV
0.822 PPV
Non Members
0.75
0.50
0.25
0.00
Fraction of Data Set
>= 0
>= 1
>= 2
>= 3
>= 4
= 5
Number of times identified as member (out of 5 runs)

Conclusion

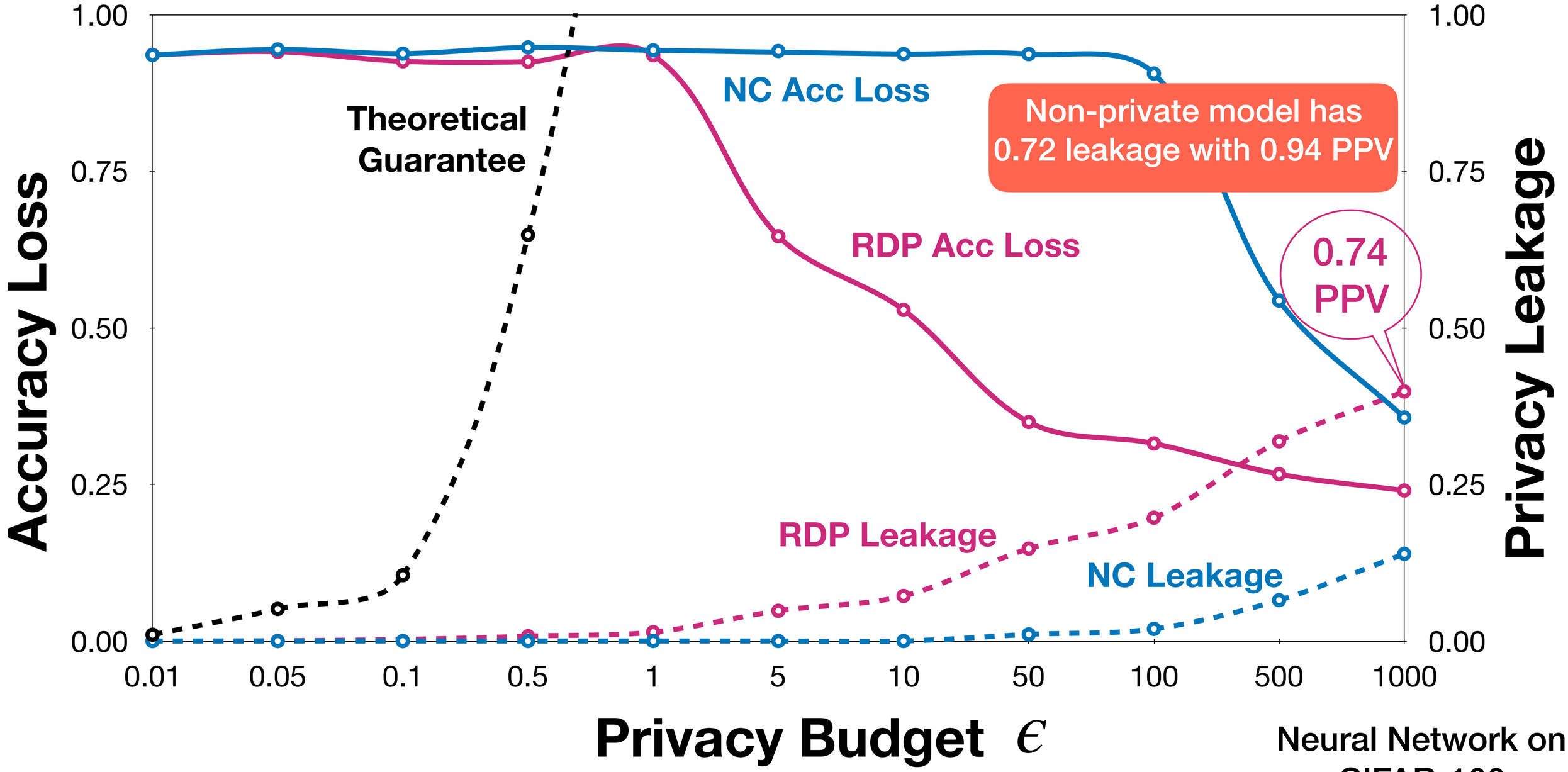


Logistic Regression on CIFAR-100

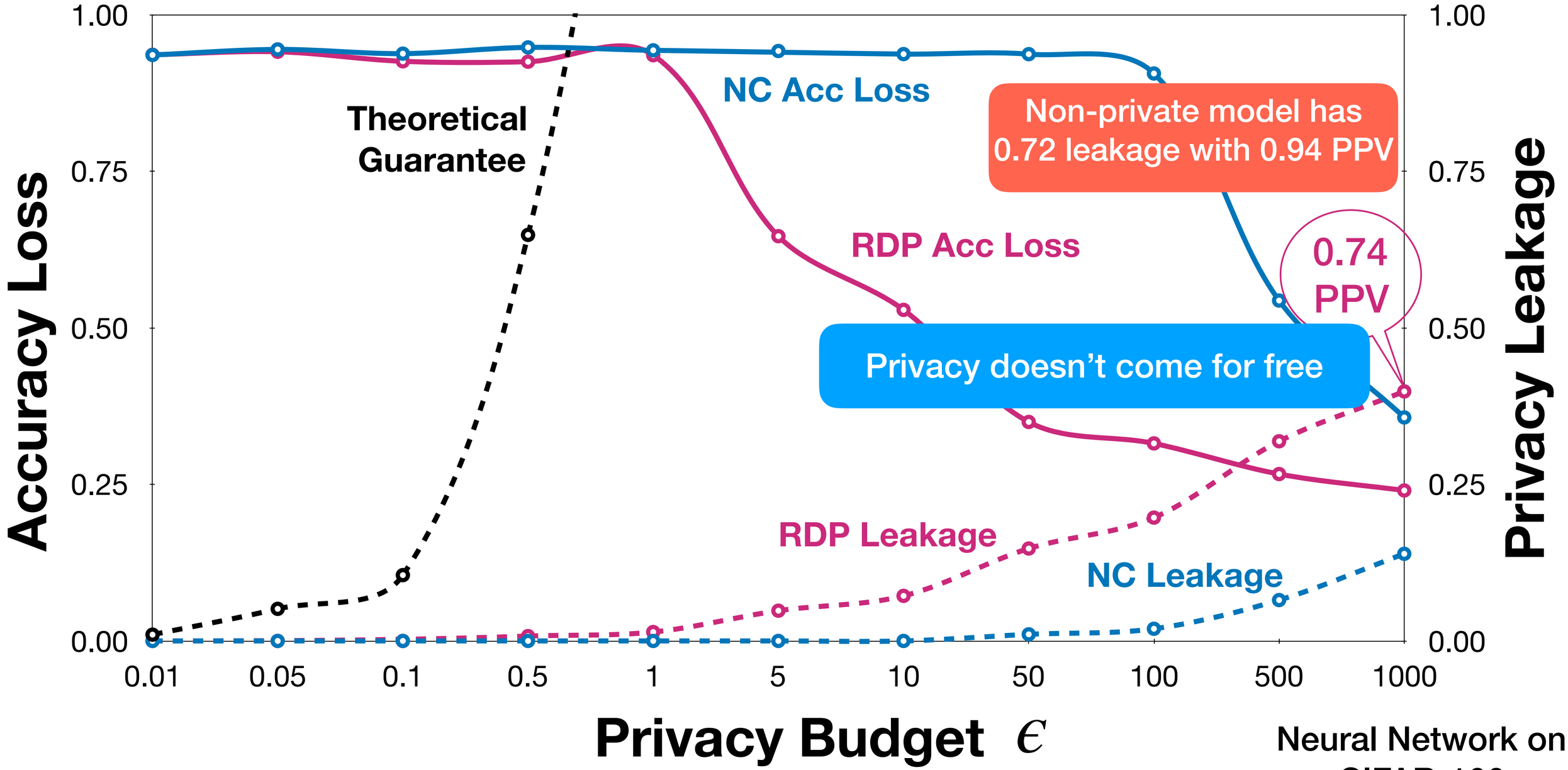
Conclusion


1.00
0.75
0.50
0.25
0.00
Accuracy Loss
Privacy Leakage
RDP Acc Loss
NC Acc Loss
Theoretical Guarantee
Non-private model has 0.12 leakage with 0.56 PPV
There is privacy leakage, but not considerable, even for non-private model
0.55 PPV
RDP Leakage
NC Leakage
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Logistic Regression on CIFAR-100

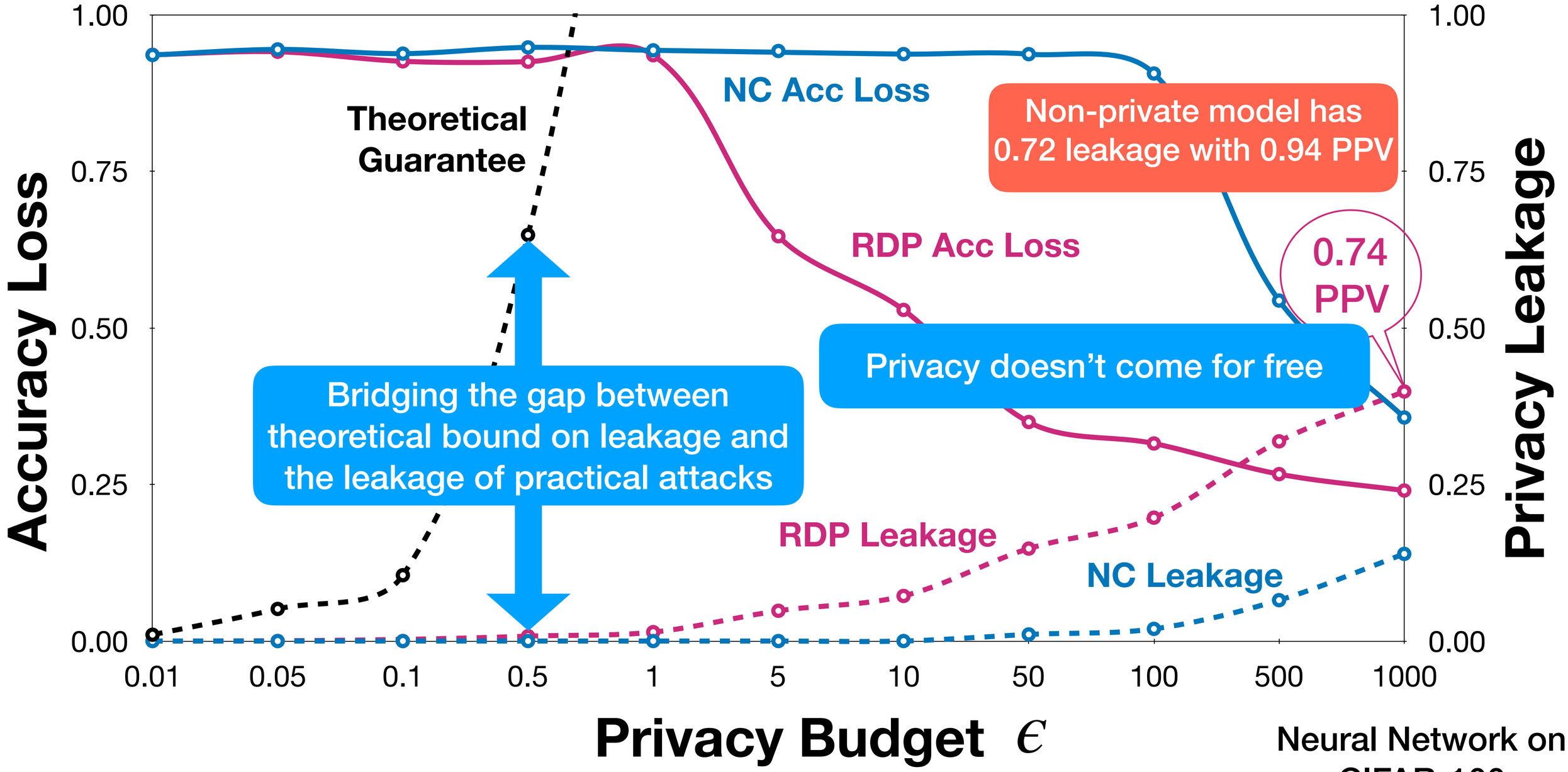
Conclusion


1.00
0.75
0.50
0.25
0.00
Accuracy Loss
Privacy Leakage
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Theoretical Guarantee
NC Acc Loss
RDP Acc Loss
RDP Leakage
NC Leakage
Non-private model has 0.72 leakage with 0.94 PPV
0.74 PPV
Neural Network on CIFAR-100

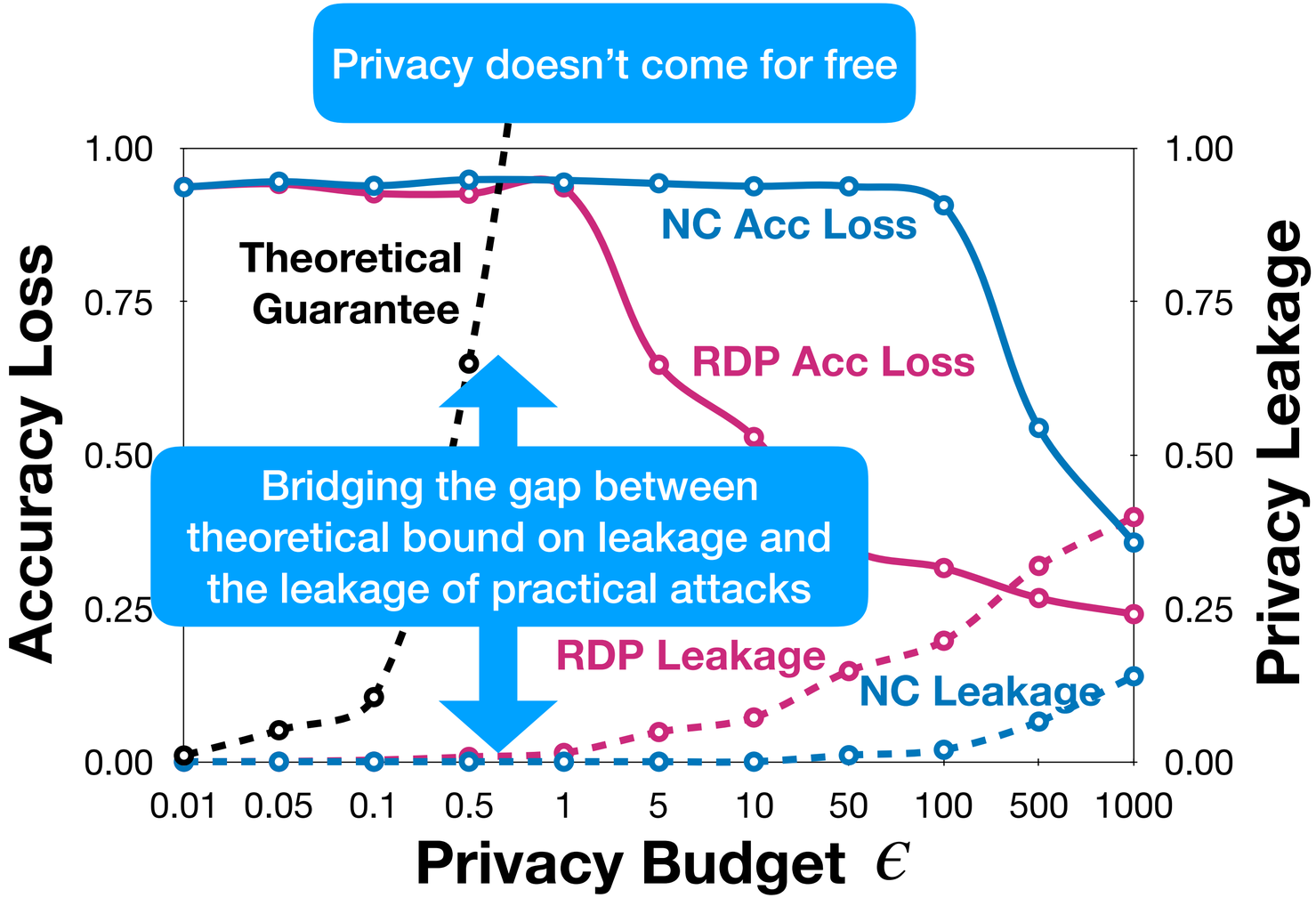
Conclusion


1.00
0.75
0.50
0.25
0.00
Accuracy Loss
Privacy Leakage
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Theoretical Guarantee
NC Acc Loss
RDP Acc Loss
RDP Leakage
NC Leakage
Non-private model has 0.72 leakage with 0.94 PPV
0.74 PPV
Privacy doesn't come for free
Neural Network on CIFAR-100

Conclusion


1.00
0.75
0.50
0.25
0.00
Accuracy Loss
Privacy Leakage
0.01
0.05
0.1
0.5
1
5
10
50
100
500
1000
Privacy Budget ϵ
Theoretical Guarantee
NC Acc Loss
RDP Acc Loss
RDP Leakage
NC Leakage
Non-private model has 0.72 leakage with 0.94 PPV
0.74 PPV
Privacy doesn't come for free
Bridging the gap between theoretical bound on leakage and the leakage of practical attacks
Neural Network on CIFAR-100

Conclusion



Thank You!

Questions?

Speaker:
Bargav Jayaraman

Project Site:
https://bargavjayaraman.github.io/project/evaluating-dpml/

Code Available: https://github.com/bargavj/EvaluatingDPML