Privacy-Preserving Data Mining

PANKAJ KUMAR

ANDREY MITYASHOV

ANTON TSITSULIN

Data mining: privacy / utility

Protect individual information while releasing accurate data aggregations

Problem:

- Exact aggregation might leak confidential data

Goal:

- Protect individual information, release aggregate

Data mining: privacy / utility

Question:

- Can we just use some fancy cryptography?

Answer:

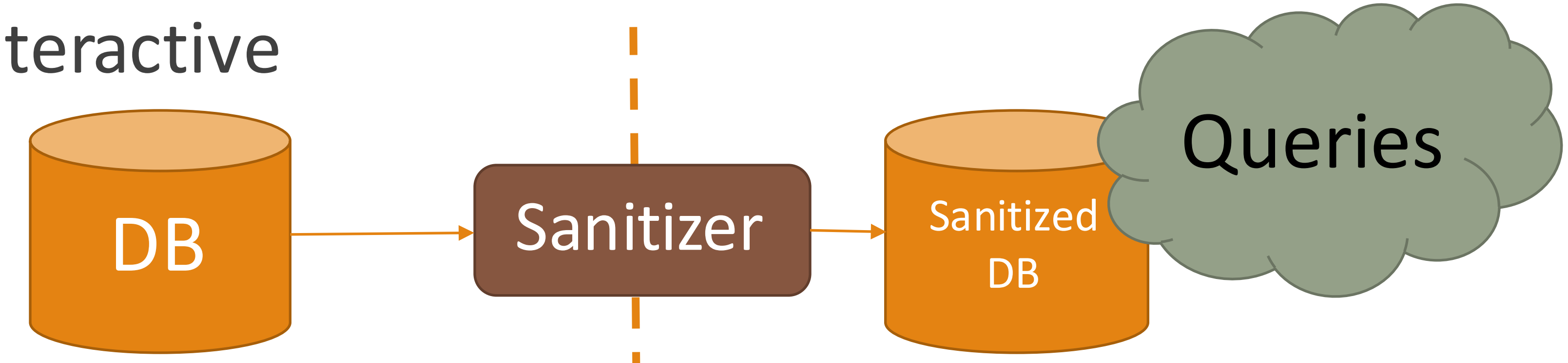
- No – the information leaks through correct answers

Solution:

- Add some noise when answering the queries

Two privacy models

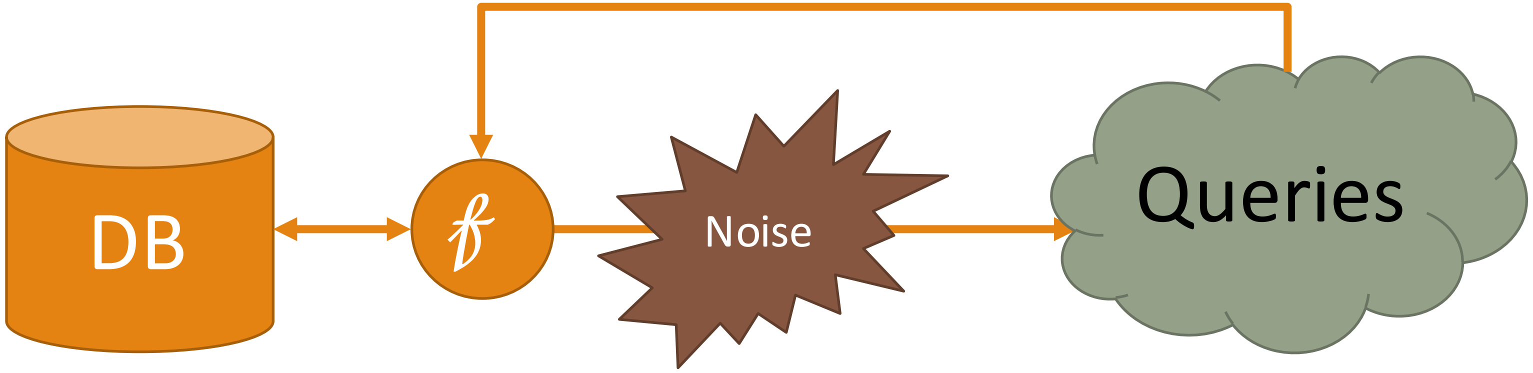
1. Non-interactive



2. Interactive

An interactive sanitizer $\mathcal{K}_f$

$\mathcal{K}_f$ applies query function f to database, and returns noisy result: $\mathcal{K}_f(DB) \equiv f(DB) + Noise$



Differential privacy

Strong privacy goal:

- Joining the database should not substantially increase or decrease the probability of any event happening

Definition:

- $\mathcal{K}_f$ provides ε-differential privacy if for any datasets A and B such that $|A \Delta B| = 1$, and all possible outcomes S,

$$\boldsymbol{P}\left(\mathcal{K}_f(A)\right) \leq \boldsymbol{P}\left(\mathcal{K}_f(B)\right) e^{\varepsilon}$$

Differential privacy

Strong privacy goal:

- Joining the database should not substantially increase or decrease the probability of any event happening

Definition:

- $\mathcal{K}_f$ provides ε-differential privacy if for any datasets A and B such that $|A \Delta B| = 1$, and all possible outcomes S,

$$\boldsymbol{P}\left(\mathcal{K}_f(A)\right) \leq \boldsymbol{P}\left(\mathcal{K}_f(B)\right) e^{\varepsilon}$$

Differential privacy

Composition:

- Sequence of ε-differential private queries $(f_1, f_1, \dots, f_n)$ provide $\mathrm{n} * \varepsilon$-differential privacy

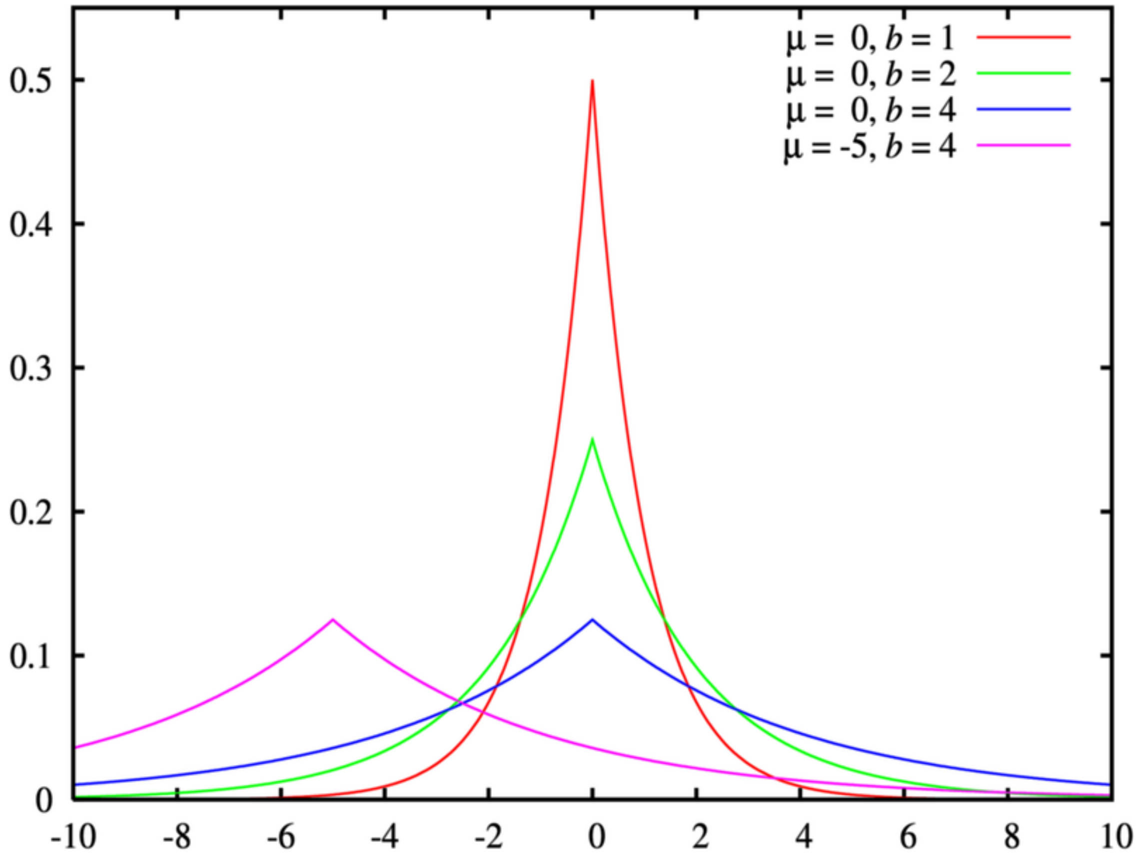
Parallel composition:

- For disjoint sets, the privacy guarantee depends on the worst privacy on the set

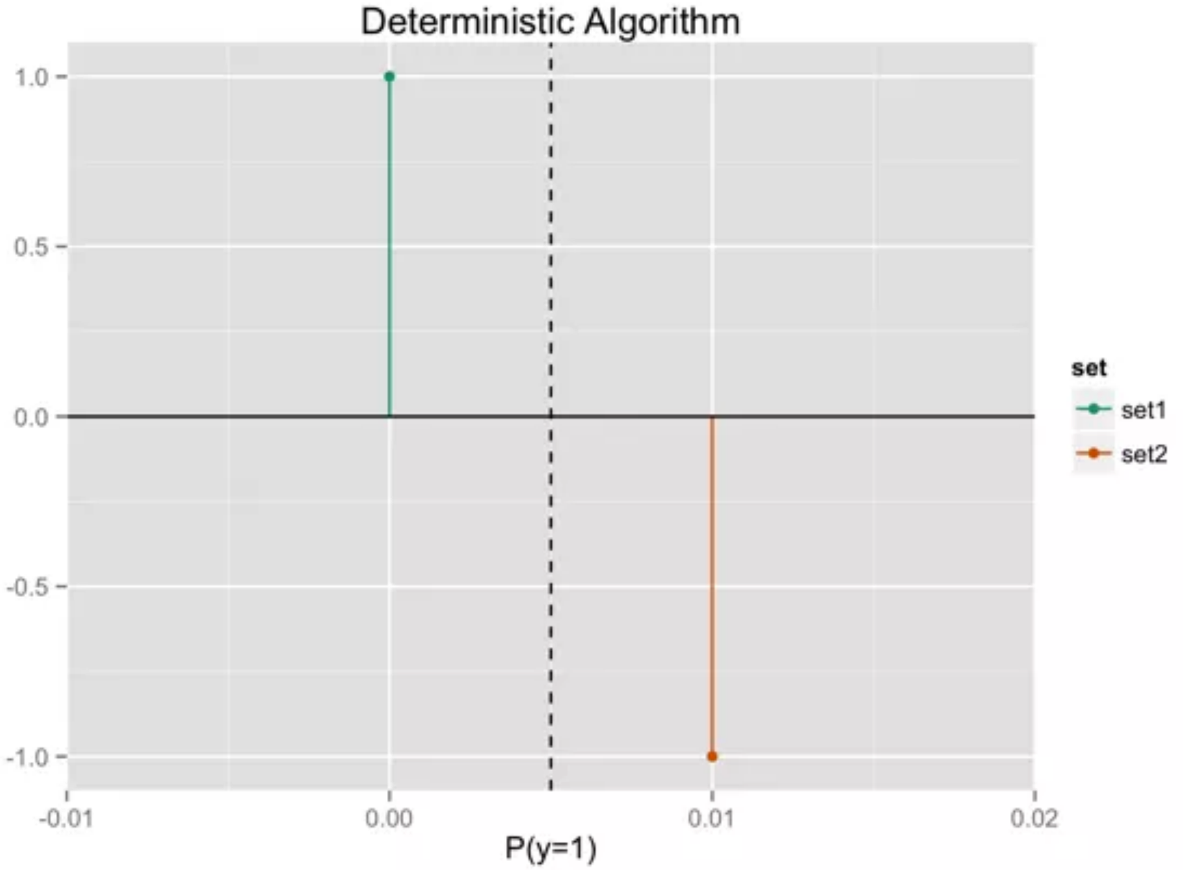
Differential privacy

Noise distribution:

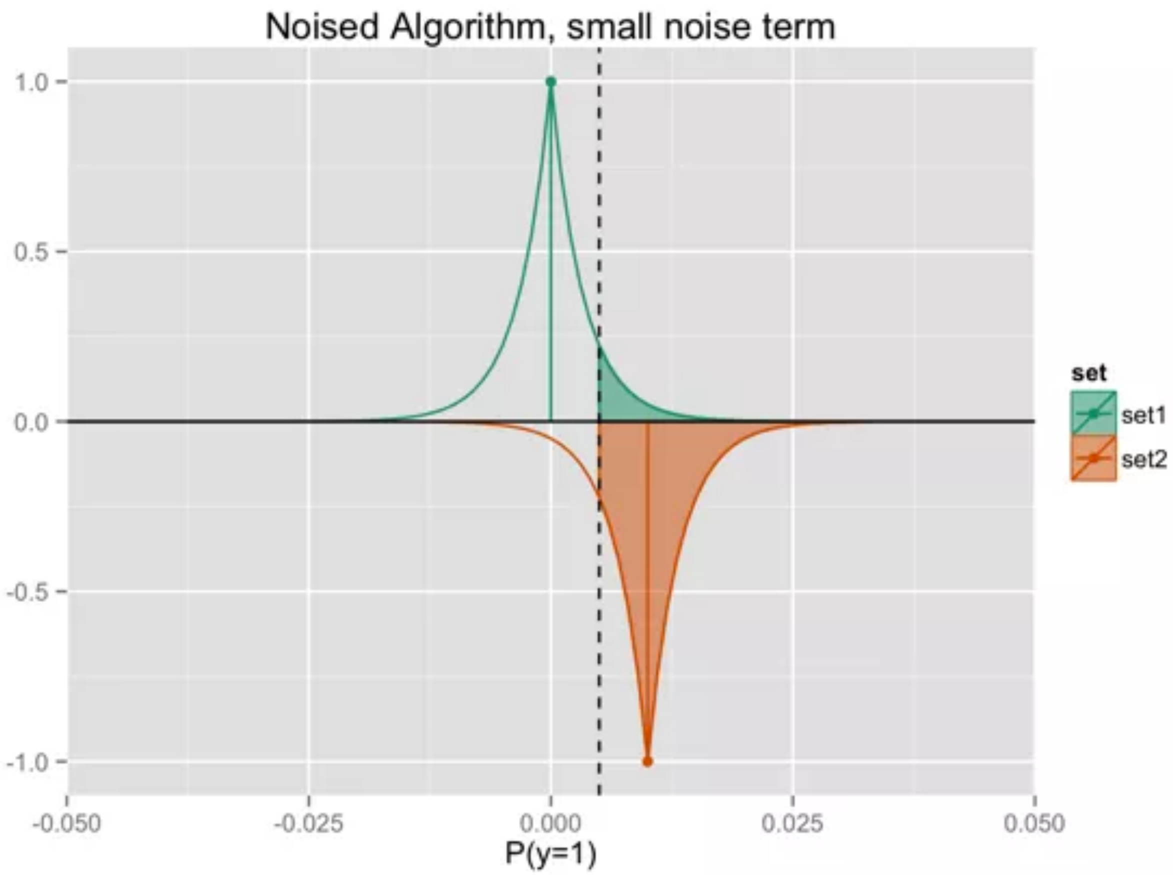
- Laplace, parameterized



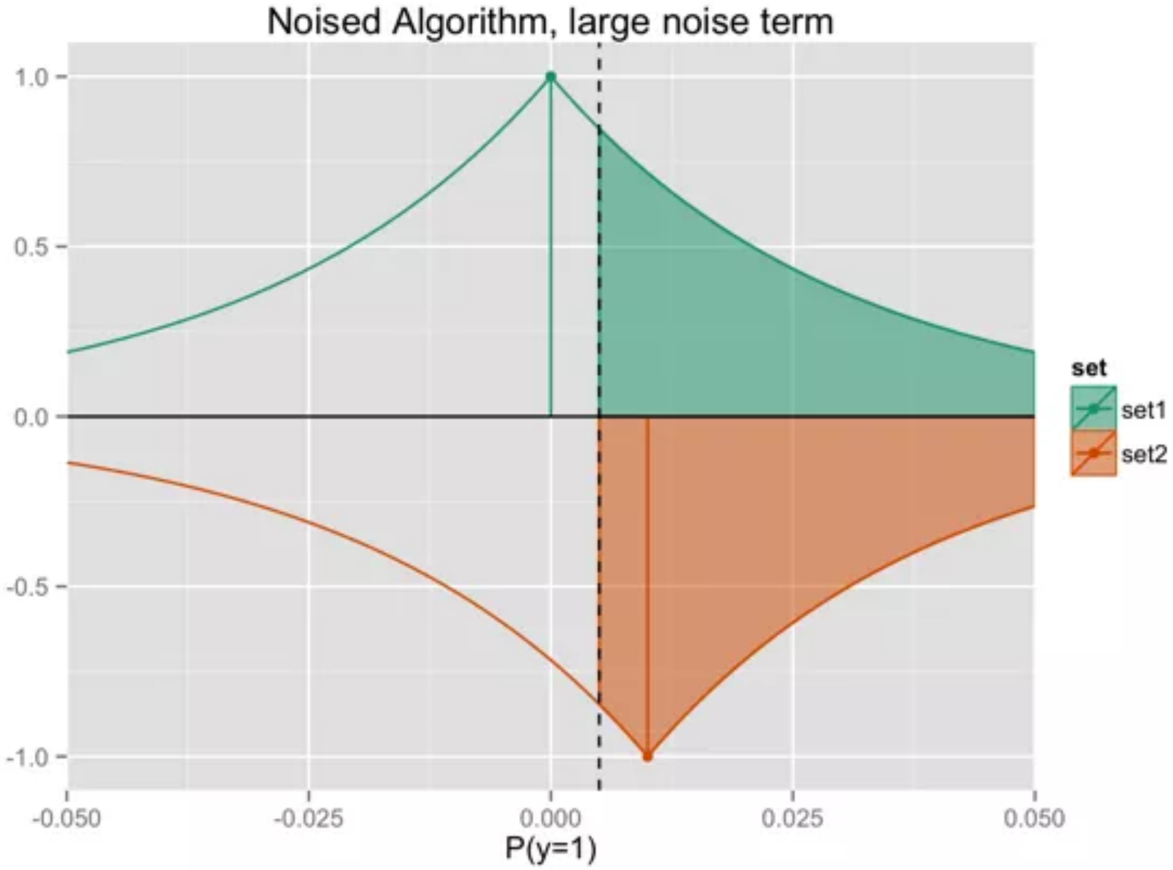
Differential privacy



Differential privacy



Differential privacy



Our question

How to select appropriate ε for machine learning?

Literature: $0.01 \leq \varepsilon \leq 100 \;\Leftrightarrow (1.01, 2.69 * 10^{43})$

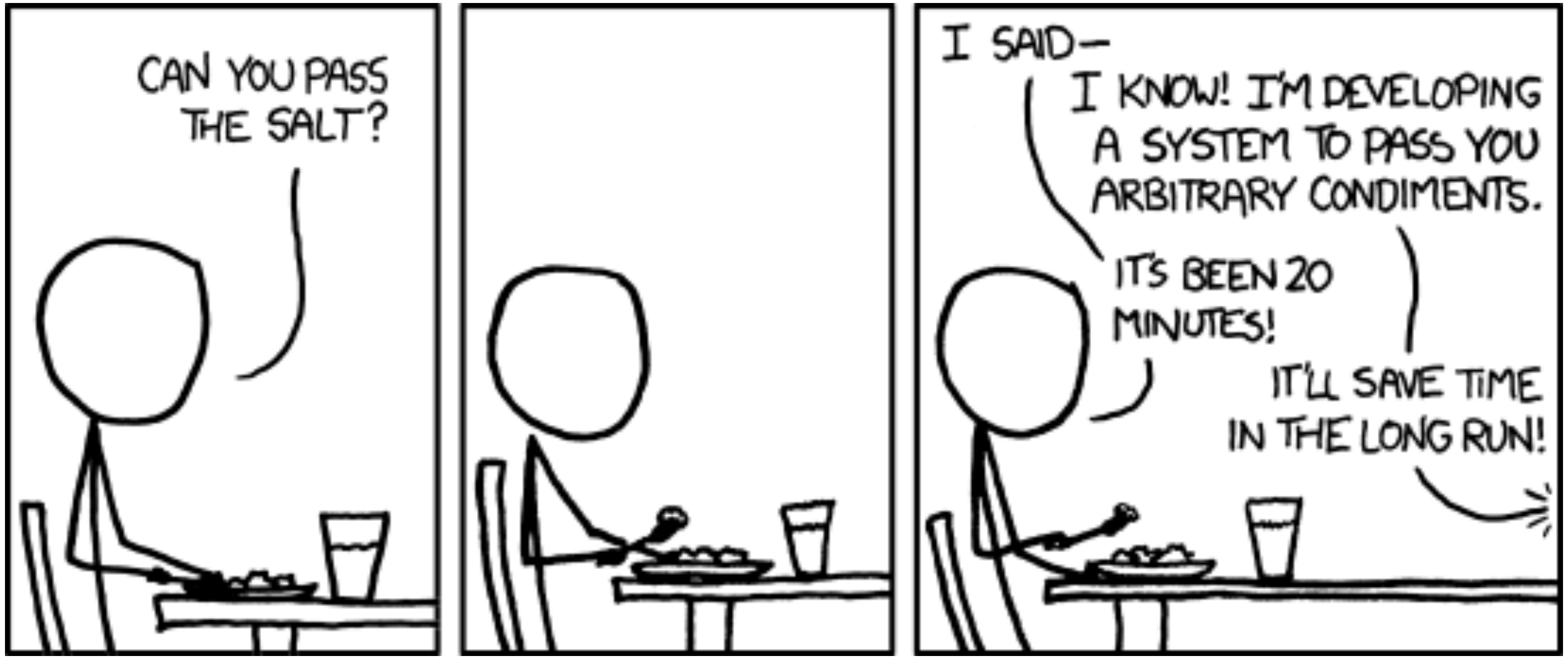
Solution:

- Tie ε to adding noise in the raw data, compare through the classifiers results

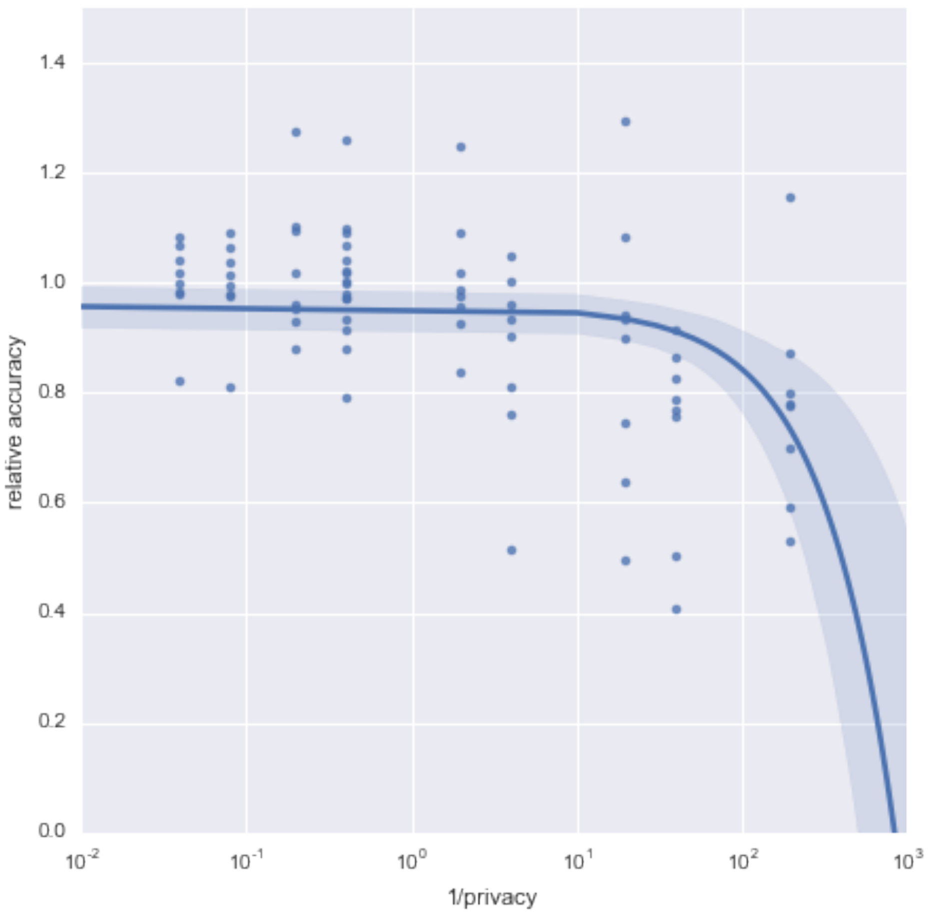
The goal

Study the actual relation between ε and norm-scaled data noise

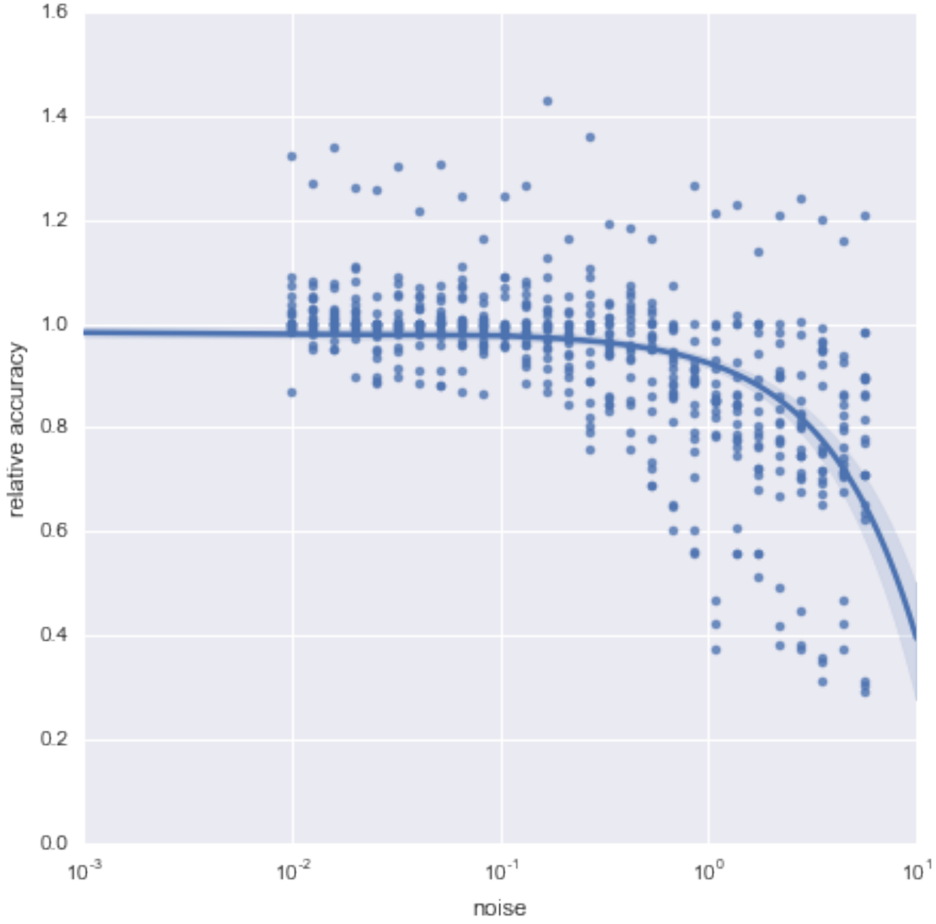
Create set of recommendations to select ε.



Results – SVM

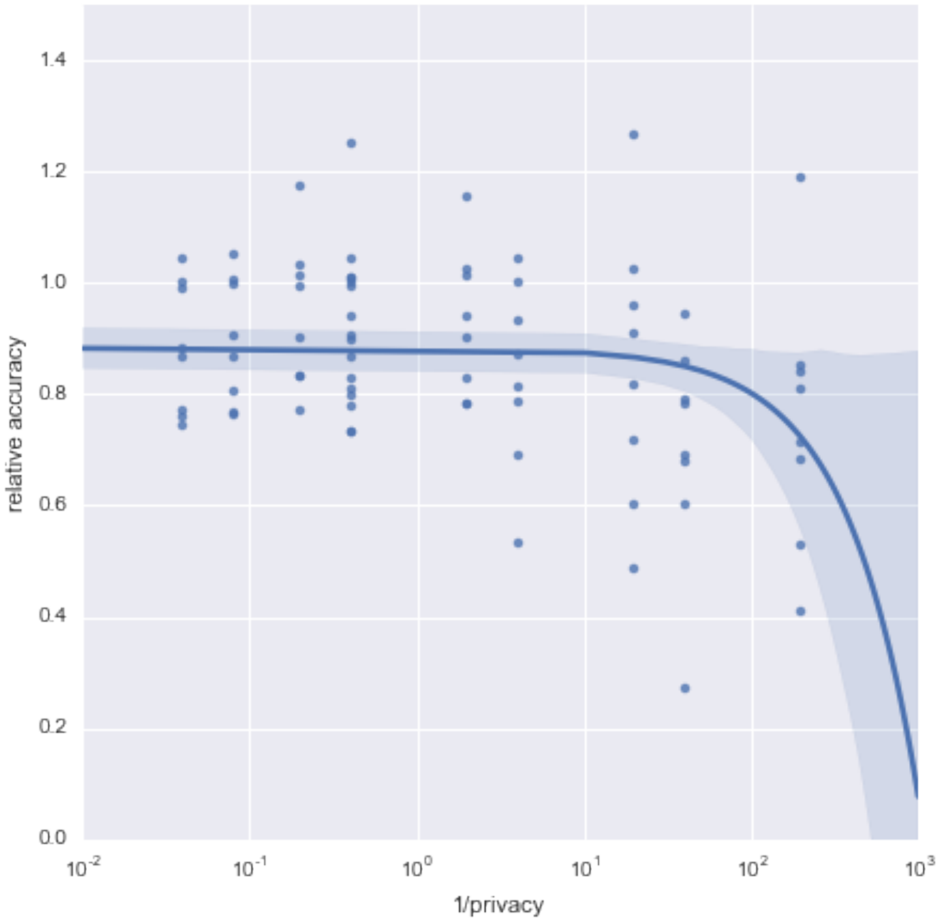


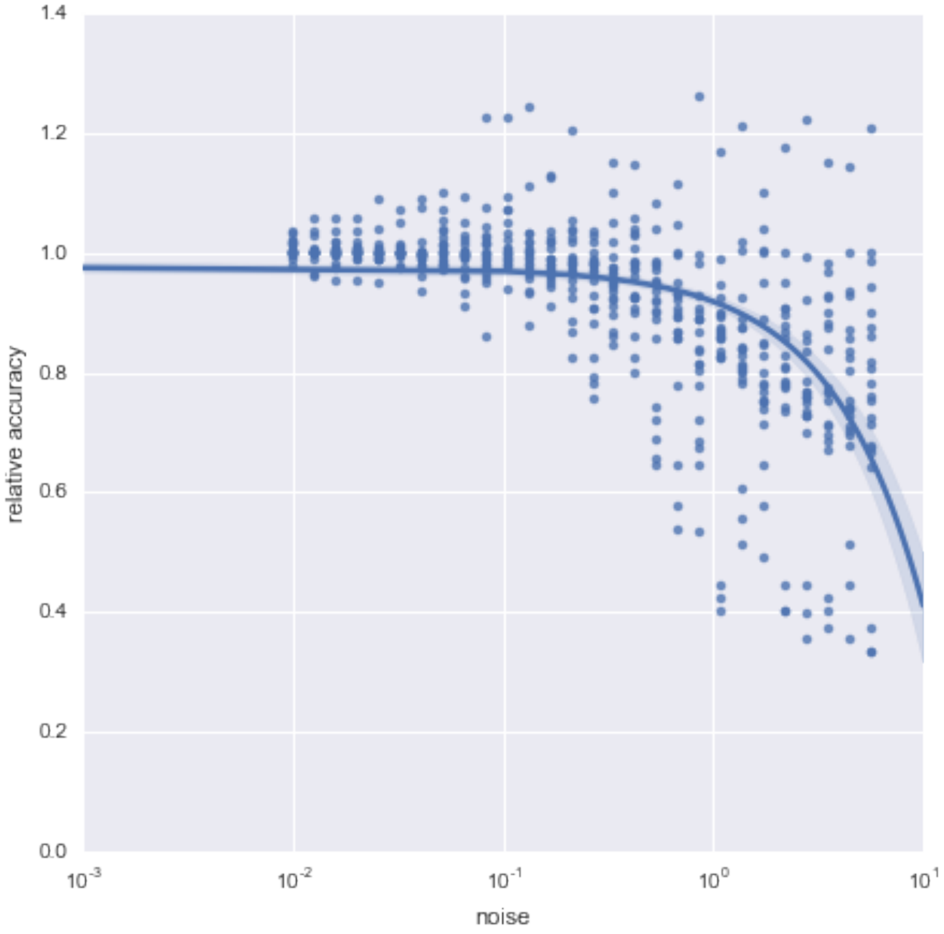
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
relative accuracy
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
1/privacy



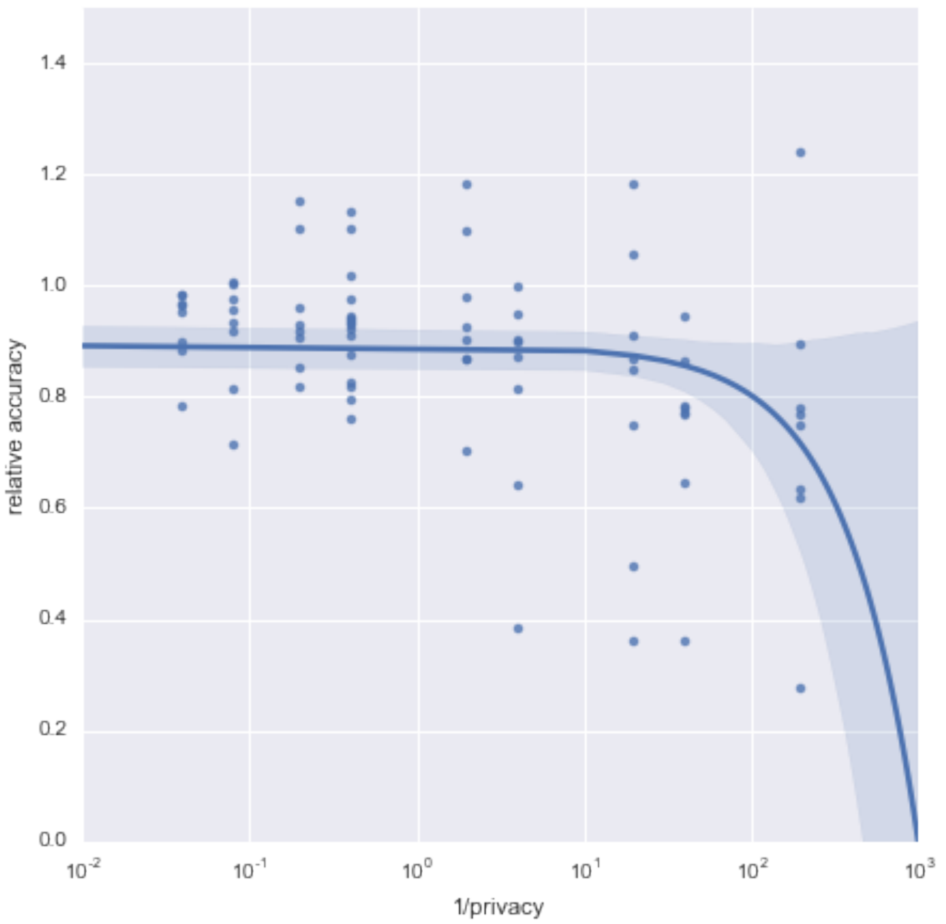
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
relative accuracy
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
noise

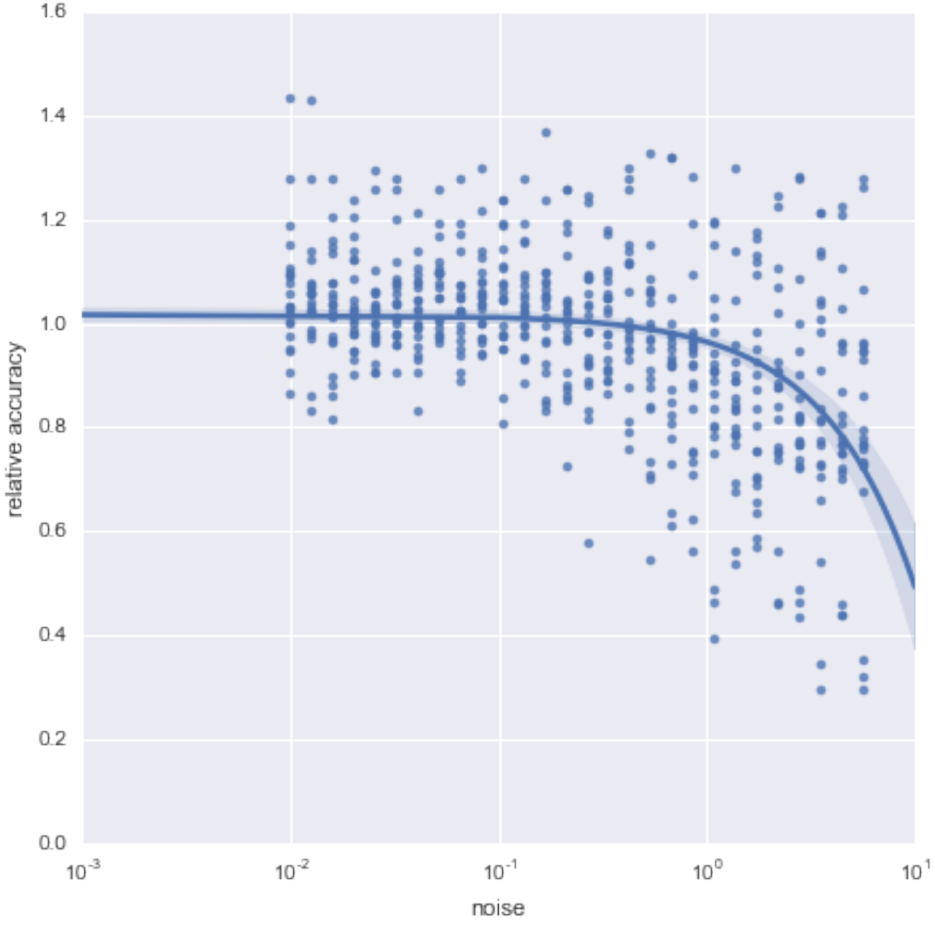
Results – Logistic regression





Results – Perceptron





Summary

Machine learning algorithms are robust to noise

There is a (notable) correlation between the data noise and privacy-induced noise

Library-adopted levels of privacy do not hurt too much

Recommended level of privacy $0.004 \leq \varepsilon \leq 0.1 \Leftrightarrow (1.004, 1.1)$ exp

Thank you for attention

PANKAJ KUMAR

ANDREY MITYASHOV

ANTON TSITSULIN