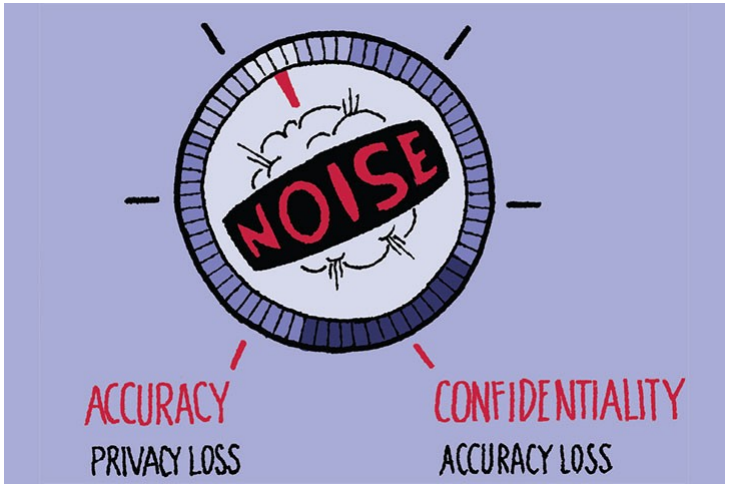
Lecture 30

Private Data Analysis II



Announcements

- Course project deliverable 2 due tonight @ 11:59pm

- Office hours every Tuesday (noon to 4pm) and Friday (11am to 1pm)
  - Link to book appointments is on the course website

Today

How to privatize statistics?

- Mean, median, covariance, etc.
- Not trivial problems!

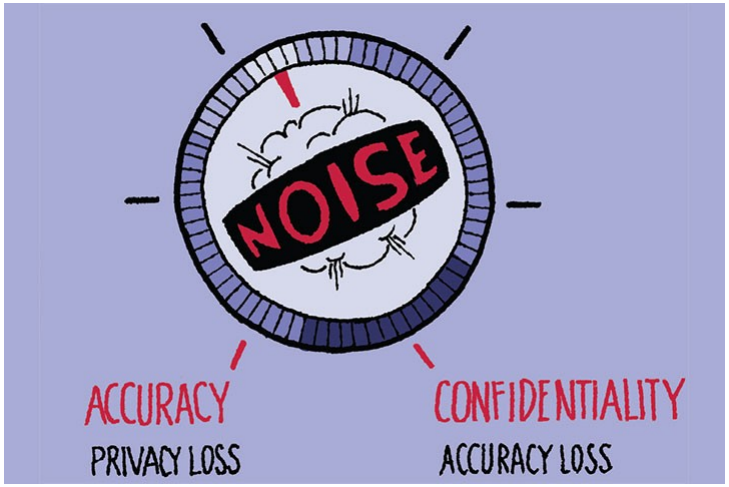
More details on differential privacy framework

Differential Privacy

Obtain privacy guarantees using framework of **differential privacy**

Idea: an attacker cannot recover any individual's data from a privatized statistic

- This is guaranteed probabilistically
- Governed by privacy parameter ρ
- ρ is correlated with the probability of successful attack

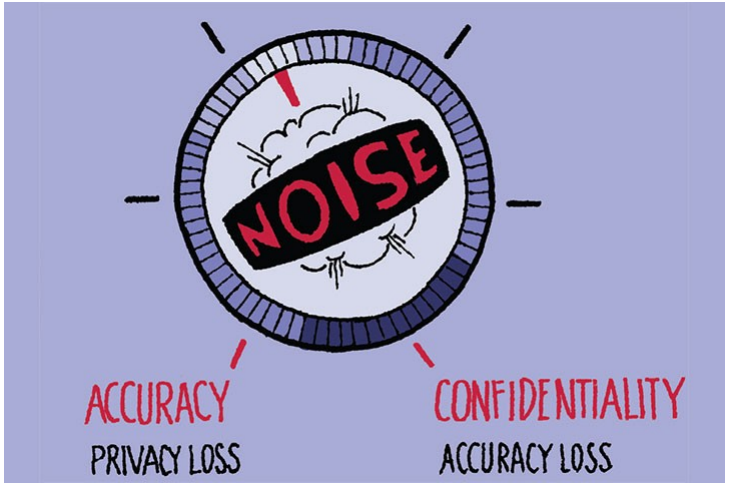


Adding Noise

Obtain privacy guarantees using framework of **differential privacy**

Strategy: inject randomization (often through adding noise) to the data analysis

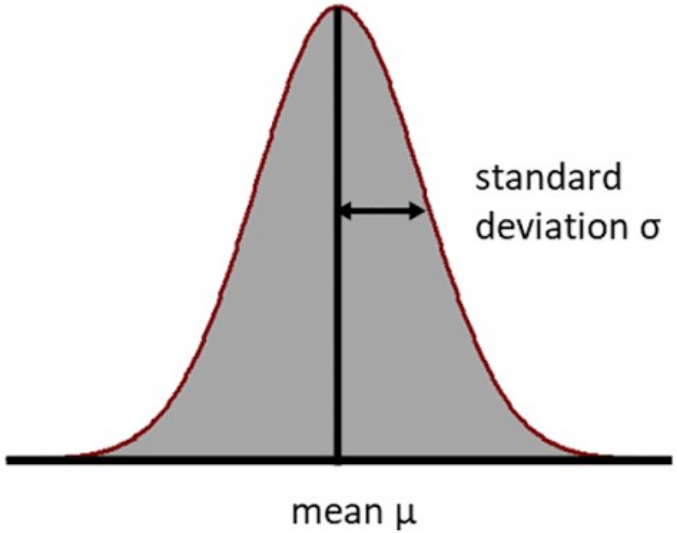
- By carefully controlling the randomization, we can provide a guarantee on how much information is leaked in the worst case
- More noise = stronger privacy guarantee
- How much noise to add for a given privacy parameter ρ?



Adding Noise

Idea: add noise to our statistic from a normal distribution

1) Calculate statistic
2) Sample noise from normal distribution
3) Release their sum



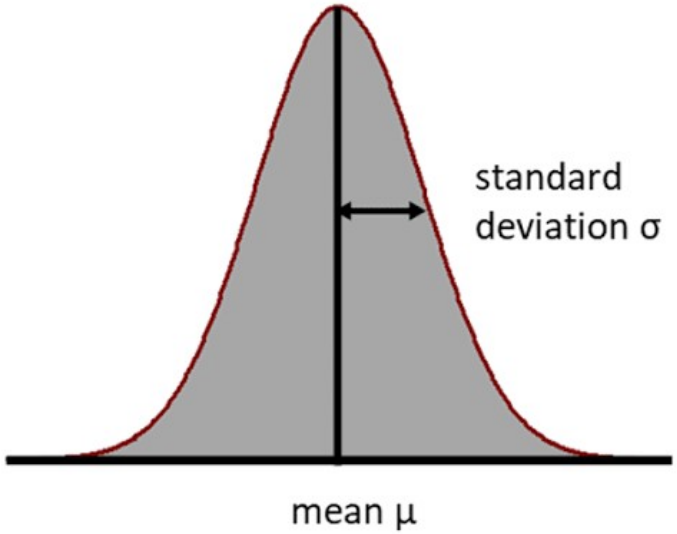
Privacy guarantee ρ is determined by

- Standard deviation of the noise σ
- Sensitivity of the statistic Δ

Adding Noise

Idea: add noise to our statistic from a normal distribution

1) Calculate statistic
2) Sample noise from normal distribution
3) Release their sum



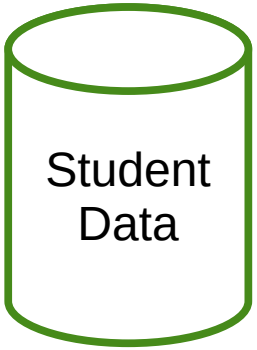
Suppose you want to measure a statistic with privacy guarantee $\rho > 0$

- Choose $\sigma = \frac{\Delta}{\sqrt{2\rho}}$

Setting

Student data

- Columns: age, height, GPA
- One row per student
- n rows total

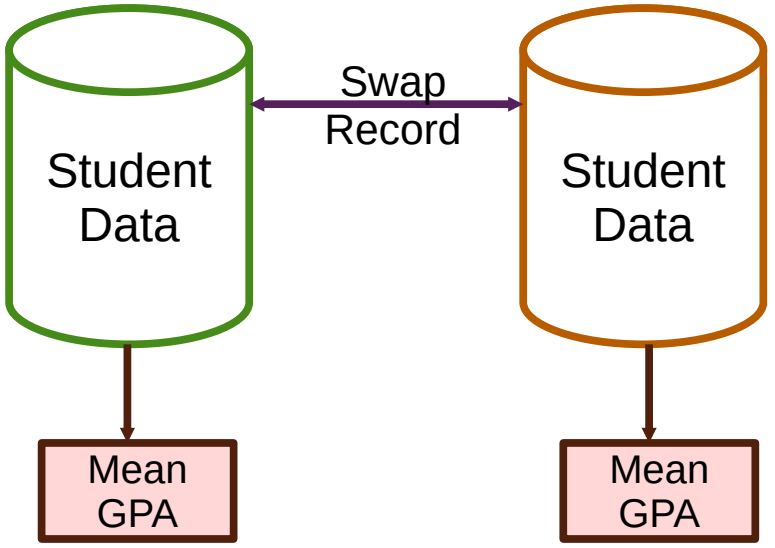


Current goal

- Privately compute the mean GPA

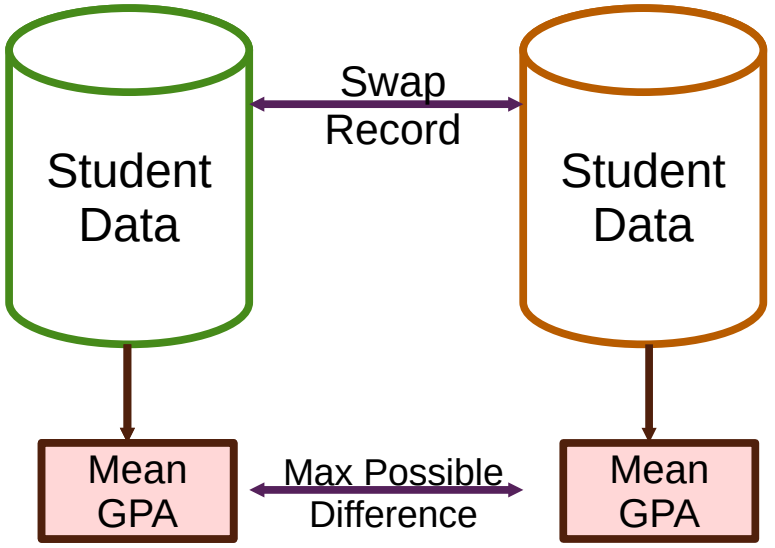
Sensitivity

Idea: how much does our statistic change on datasets that differ on one row's data



Sensitivity

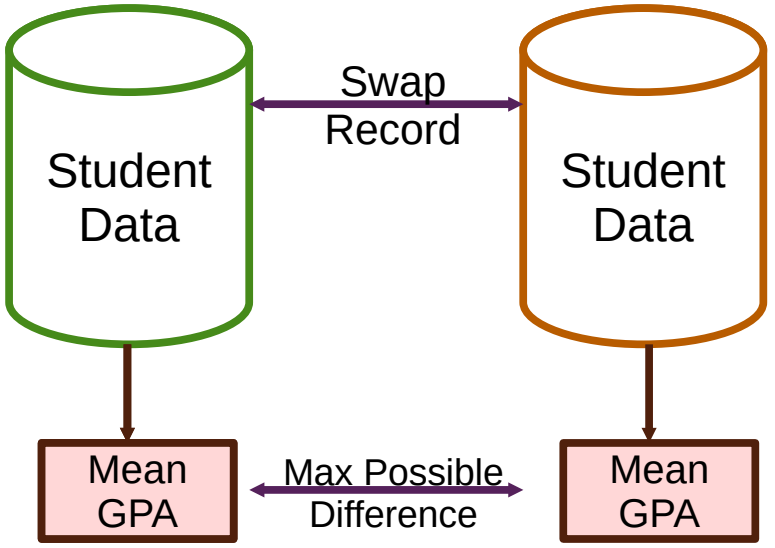
Idea: how much does our statistic change on datasets that differ on one row's data



Sensitivity

Idea: how much does our statistic change on datasets that differ on one row's data

Recall: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$



Sensitivity

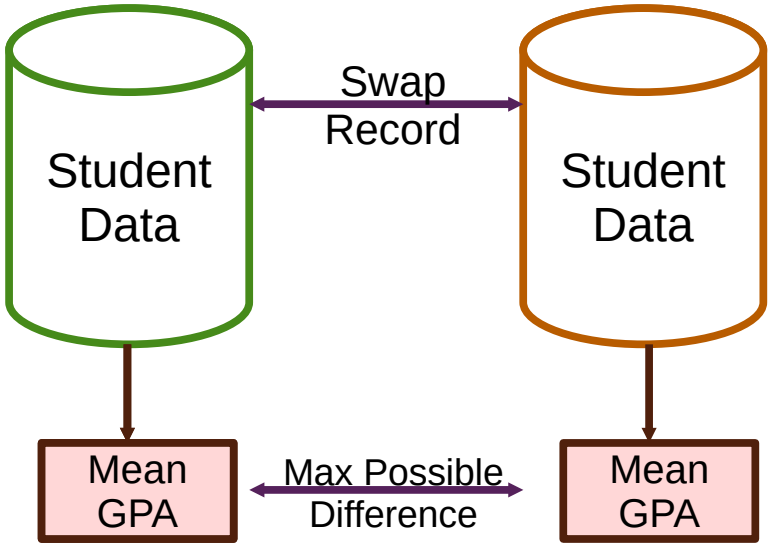
Idea: how much does our statistic change on datasets that differ on one row's data

Recall: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Maximum difference:

$$\left| \frac{1}{n} (x_j - y_j) \right| = \frac{4}{n}$$

where does the 4 come from?



Sensitivity

Idea: how much does our statistic change on datasets that differ on one row's data

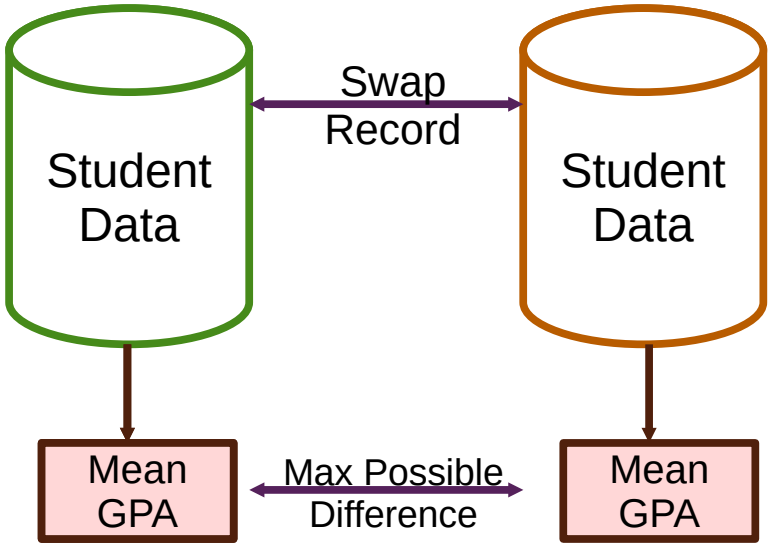
Recall: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Maximum difference:

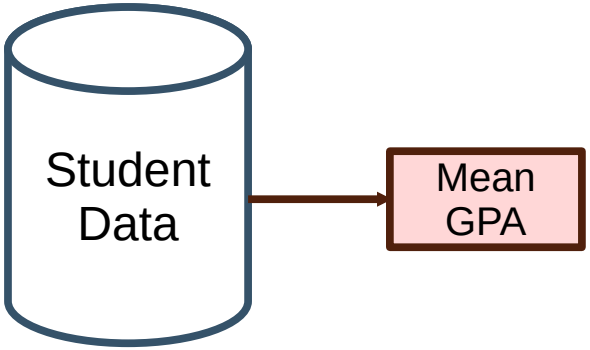
$$\left| \frac{1}{n} (x_j - y_j) \right| = \frac{4}{n}$$

where does the 4 come from?

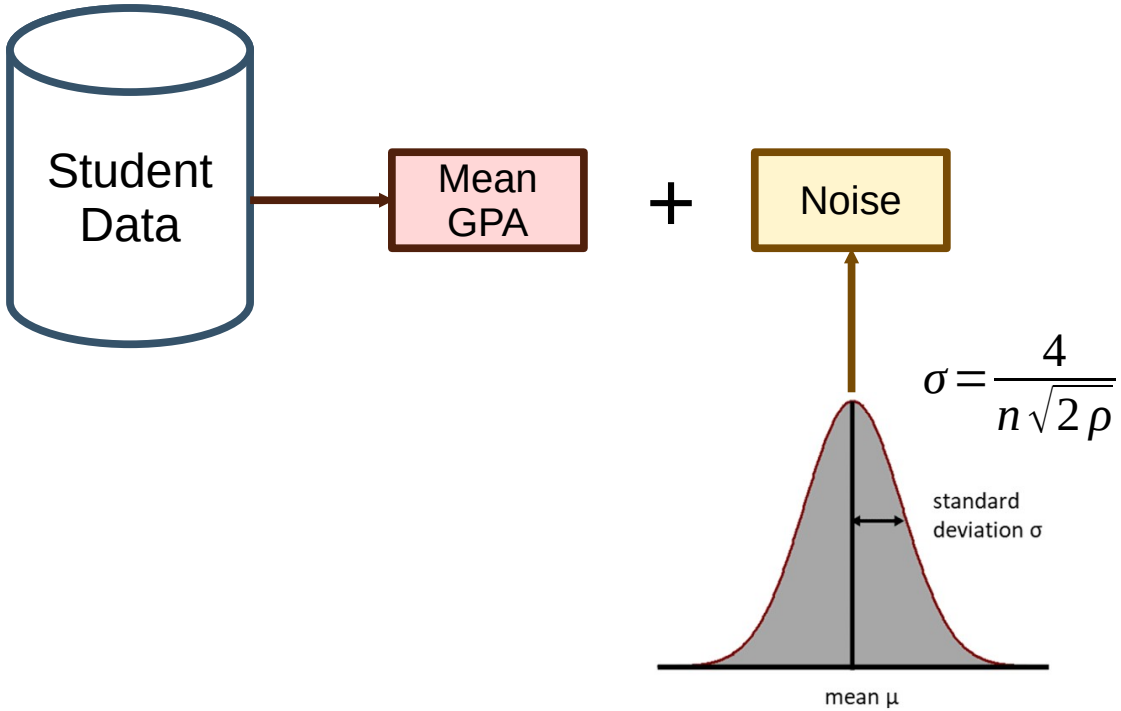
Imagine swapped a 4.0 GPA for a 0.0 GPA (or vice versa)



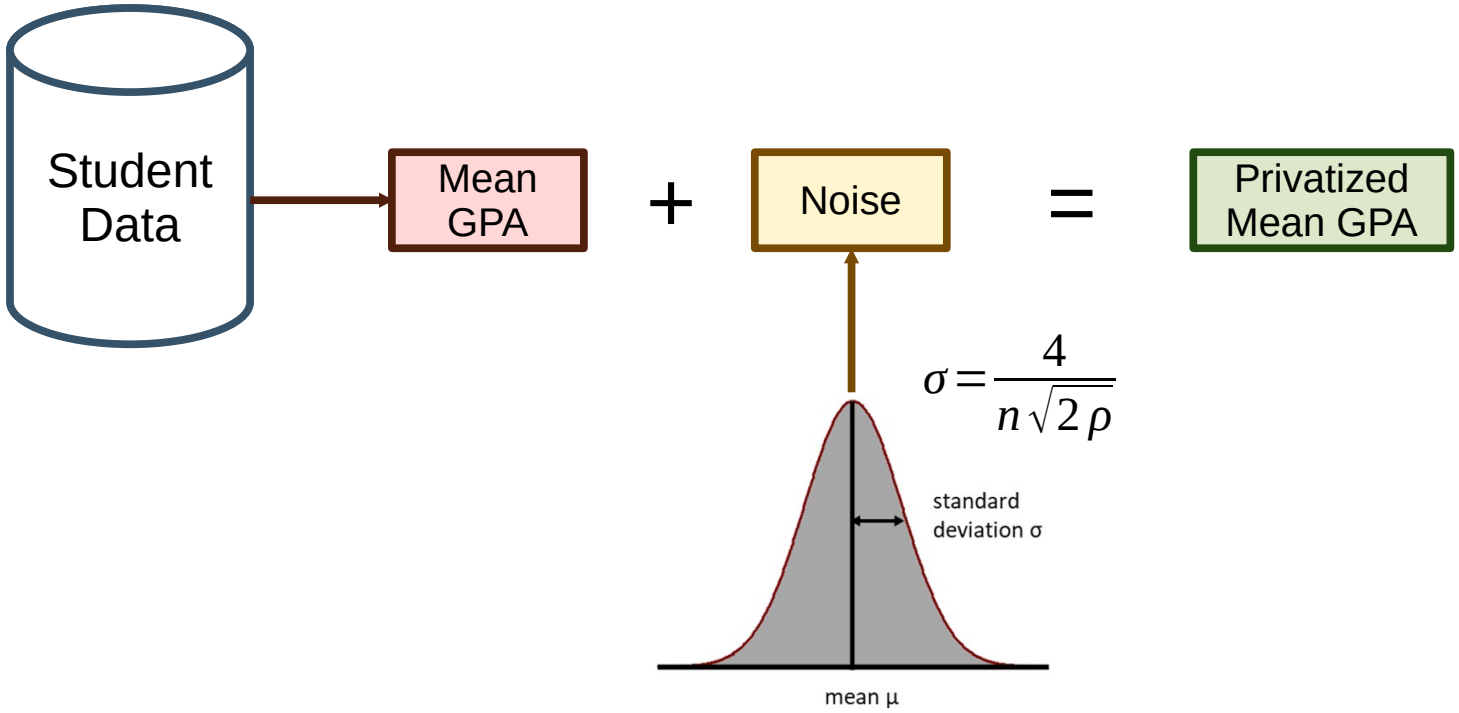
Privatizing the Mean



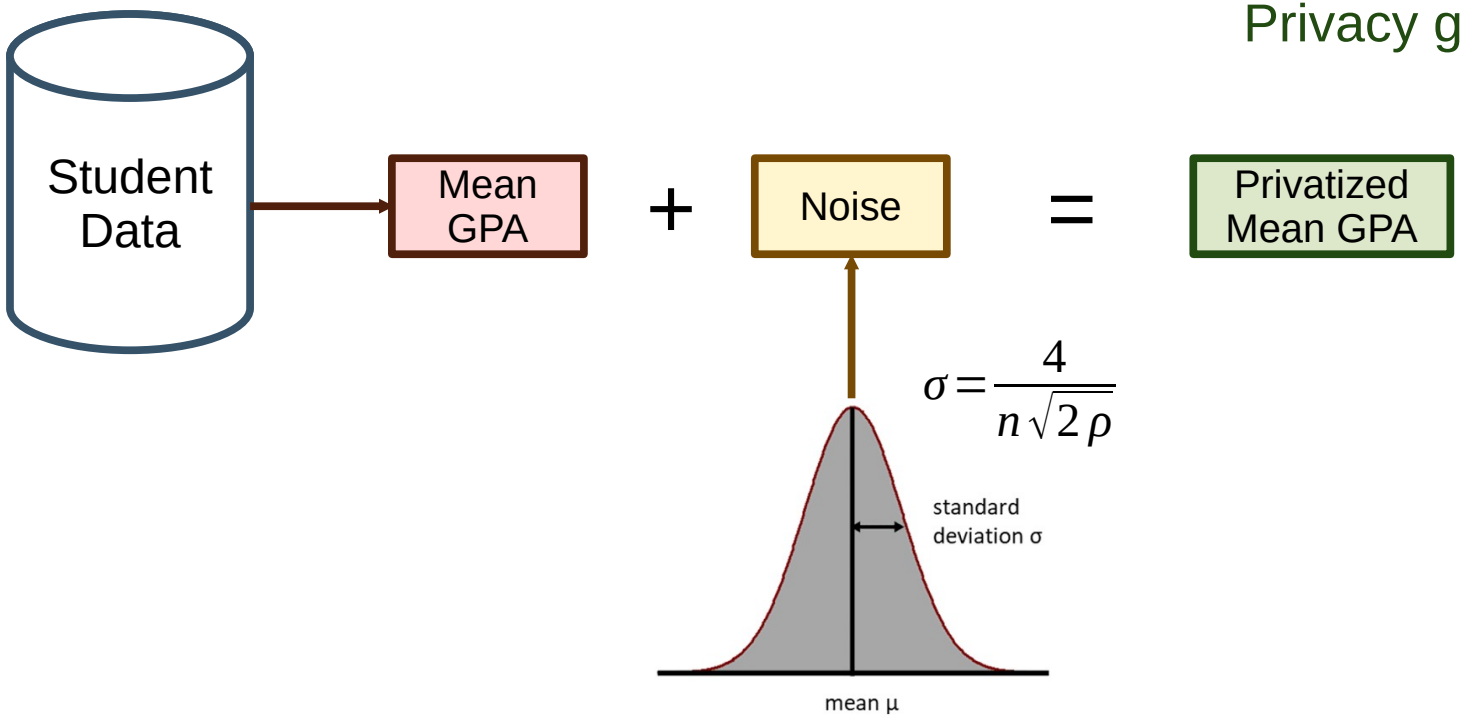
Privatizing the Mean



Privatizing the Mean



Privatizing the Mean



Privacy guarantee ρ

Exercises

Compute the sensitivity for the following statistics

- Mean student height
- Count of all students
- Count of students with GPA under 3.0
- Median GPA

Next Time

Properties that makes private data analysis easy

- Composition
- Post-Processing

Privacy methods in Python