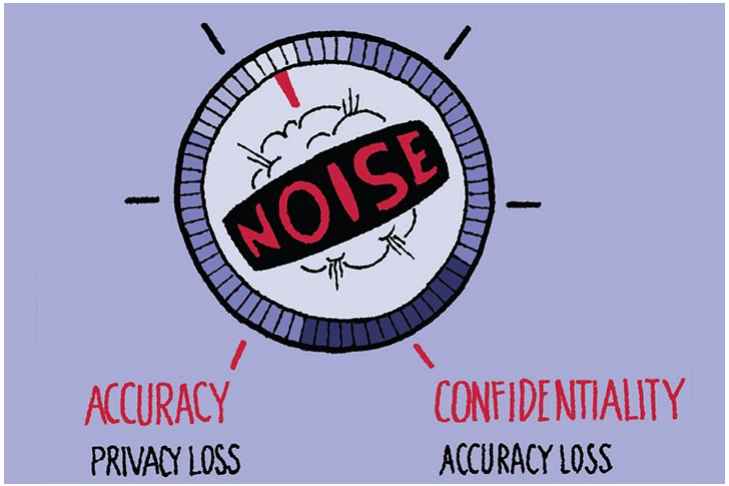
Lecture 31

Private Data Analysis III



Announcements

- Course project presentations next Wednesday and Friday

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

Today

Private Selection

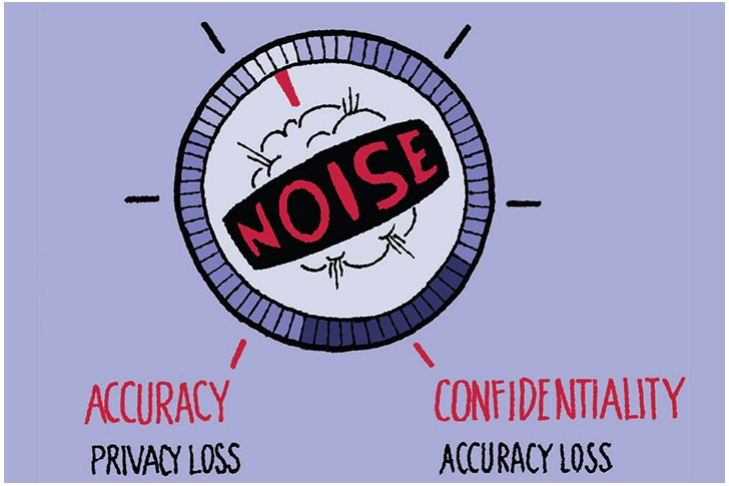
Privacy Accounting

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

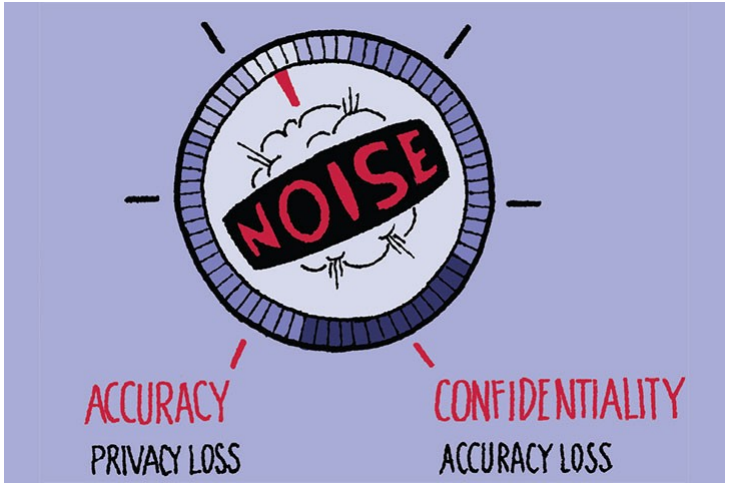


Adding Noise

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

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

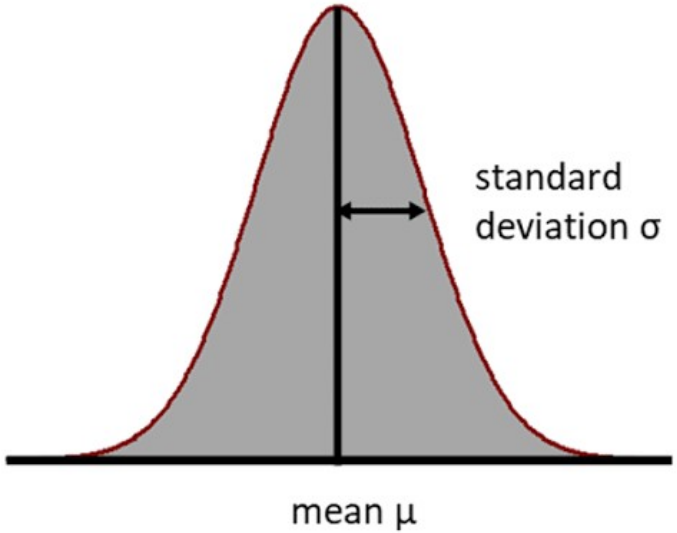
- By carefully controlling the randomization, we can provide a guarantee on how much information is leaked in the worst case
- More noise = stronger privacy guarantee
- How much noise to add for a given privacy parameter ρ?



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

Private Selection

What if you want to privatize a categorical variable?

- Can you use noise addition?

Private Selection

What if you want to privatize a categorical variable?

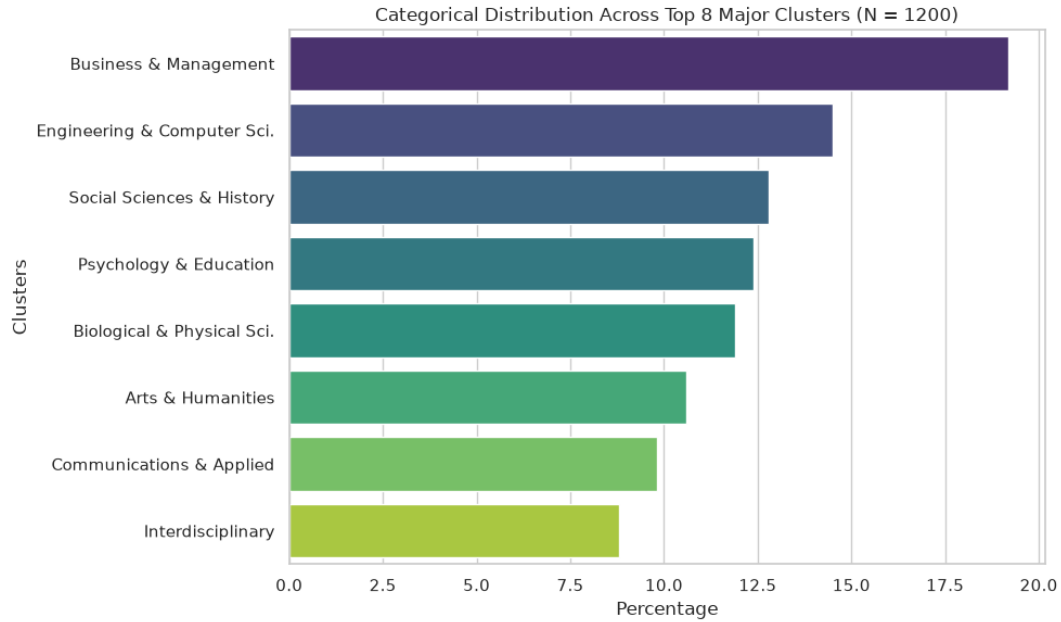
- Can you use noise addition? No, for the same reason that taking the average of a categorical variable doesn't make sense

Private Selection

What if you want to privatize a categorical variable?

- Can you use noise addition? No, for the same reason that taking the average of a categorical variable doesn't make sense

Example: among Allegheny students, which major is most popular?



Private Selection

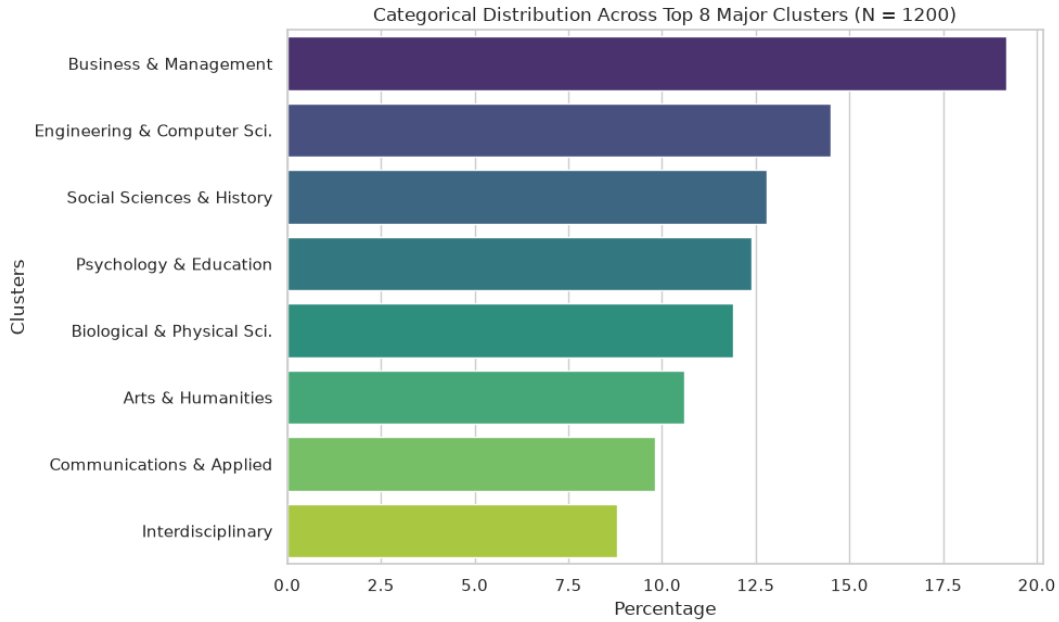
What if you want to privatize a categorical variable?

- Can you use noise addition? No, for the same reason that taking the average of a categorical variable doesn't make sense

Example: among Allegheny students, which major is most popular?

- Business & Management

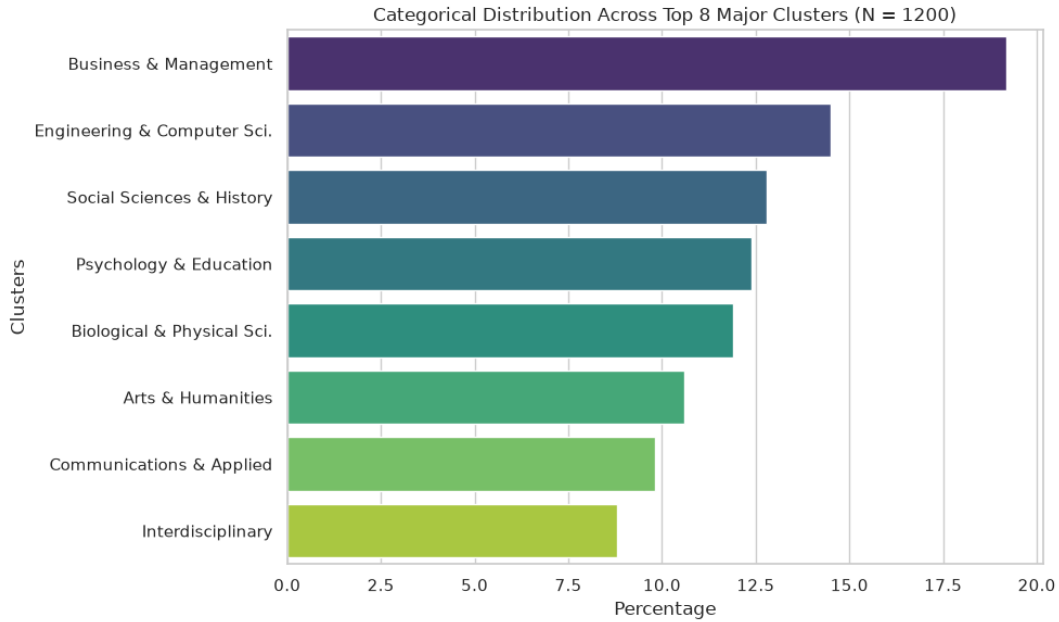
Question: how to we randomize picking the most popular?



Private Selection

Idea: randomly select a major weighted by popularity

- Privacy parameter ρ determines how "uniform" the distribution is
- Lower ρ = more uniform
- Implies higher privacy

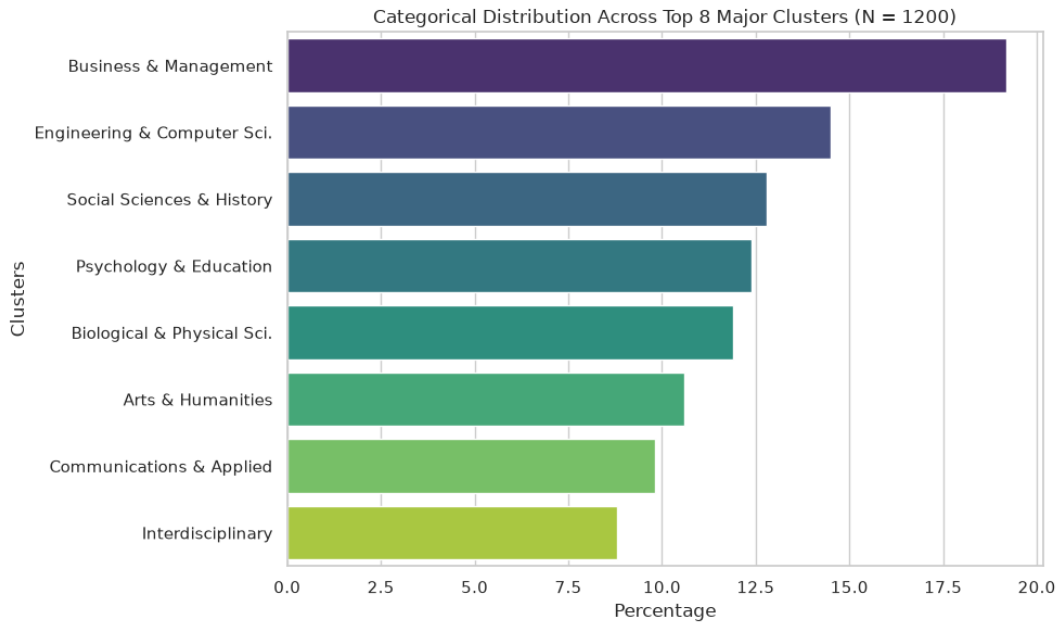


Private Selection

Idea: randomly select a major weighted by popularity

- Privacy parameter ρ determines how "uniform" the distribution is
- Lower ρ = more uniform
- Implies higher privacy

One version: exponential mechanism



Privacy Parameter

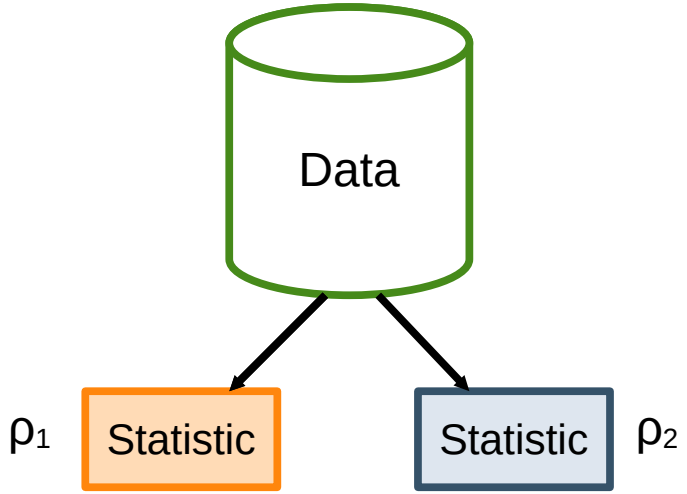
We've said that ρ determines our privacy guarantee

- We call ρ the *privacy budget* when we set the value in advance and think about how to best spend it
- We call ρ the *privacy parameter* when think about it as a guarantee on privacy

It's really important to keep track of ρ when analyzing data

Composition

We often want to calculate more than a single statistic from a dataset

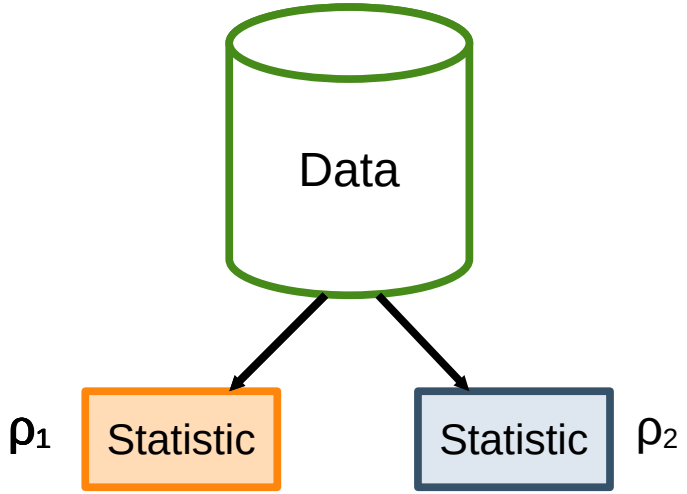


Composition

We often want to calculate more than a single statistic from a dataset

Privacy composes at most additively

- Releasing both statistics satisfies privacy for $\rho \leq \rho_1 + \rho_2$

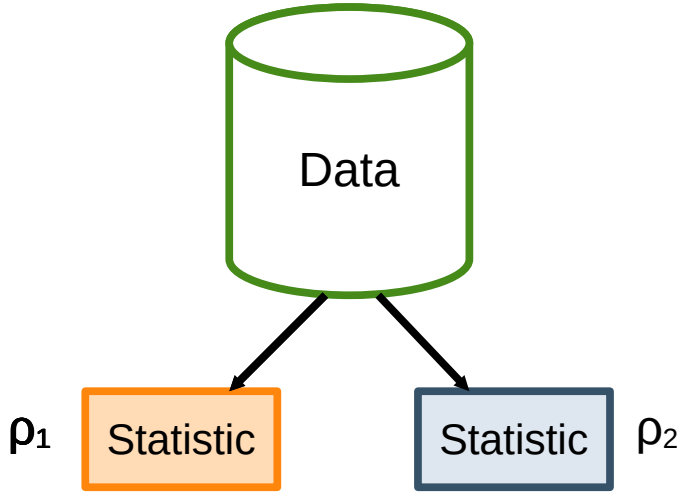


Composition

We often want to calculate more than a single statistic from a dataset

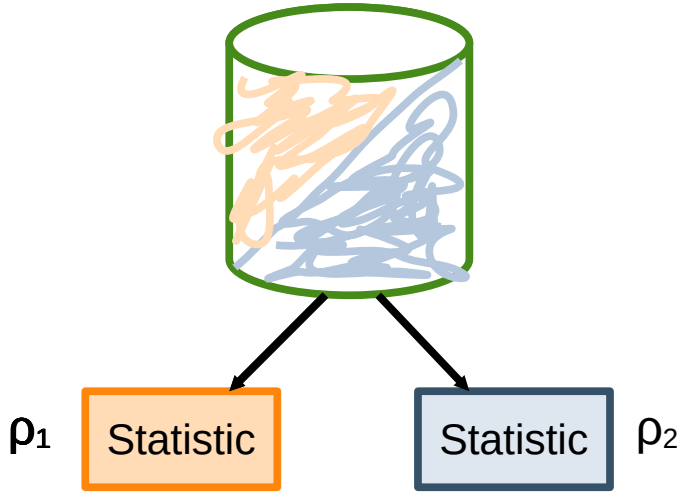
Privacy composes at most additively

- Releasing both statistics satisfies privacy for $\rho \leq \rho_1 + \rho_2$
- In many cases, composition is tight (=)
- Called *sequential composition*



Composition

You can do better if the statistics use non-overlapping parts of the dataset

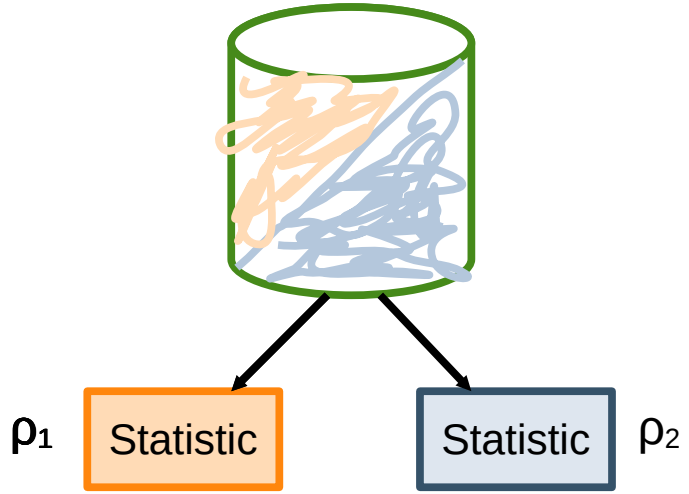


Composition

You can do better if the statistics use non-overlapping parts of the dataset

Privacy composes as the max

- Releasing both statistics satisfies privacy for $\rho = \max(\rho_1, \rho_2)$
- Called *parallel composition*



Private Data Analysis

Start from a few building blocks

- Tools for noise addition and private selection

Use them to compute statistics we're interested in

To to be clever in our design to inject as little noise as possible

