Class starts after this song

Kygo – Stranger Things (2017)
requested by Ahbab Abeer (TA-of-CM8)

Hello, I am an avid clash royale player and love terraria. I failed my elementary school entrance exam. My favorite music genre is EDM

Duke

CS230 Spring 2024
EM B: Probability Applications in Privacy

Duke

Poll (Not a PI)

- Disclaimer:
  You can feel free to answer this poll honestly;
  no consequences will result from answering this poll

Plausible Deniability

- *Theoretically,*
  you should feel less uncomfortable being told to answer the version with coin flips because you now have *plausible deniability:*

  “I answered YES just because I got a head in my coin flip”

Randomized response mechanism

Let's formalize the probability model (for each user):

- $\Pr(\text{reports cheating} | \text{have cheated}) = 1$
- $\Pr(\text{reports no cheating} | \text{have cheated}) = 0$
- $\Pr(\text{reports cheating} | \text{have not cheated}) = \frac{1}{2}$
- $\Pr(\text{reports no cheating} | \text{have not cheated}) = \frac{1}{2}$

$\Pr(\text{have cheated}) = p$

$\Pr(\text{report cheating}) = p \cdot 1 + (1-p) \cdot \frac{1}{2} = \frac{1+p}{2}$

Aggregation

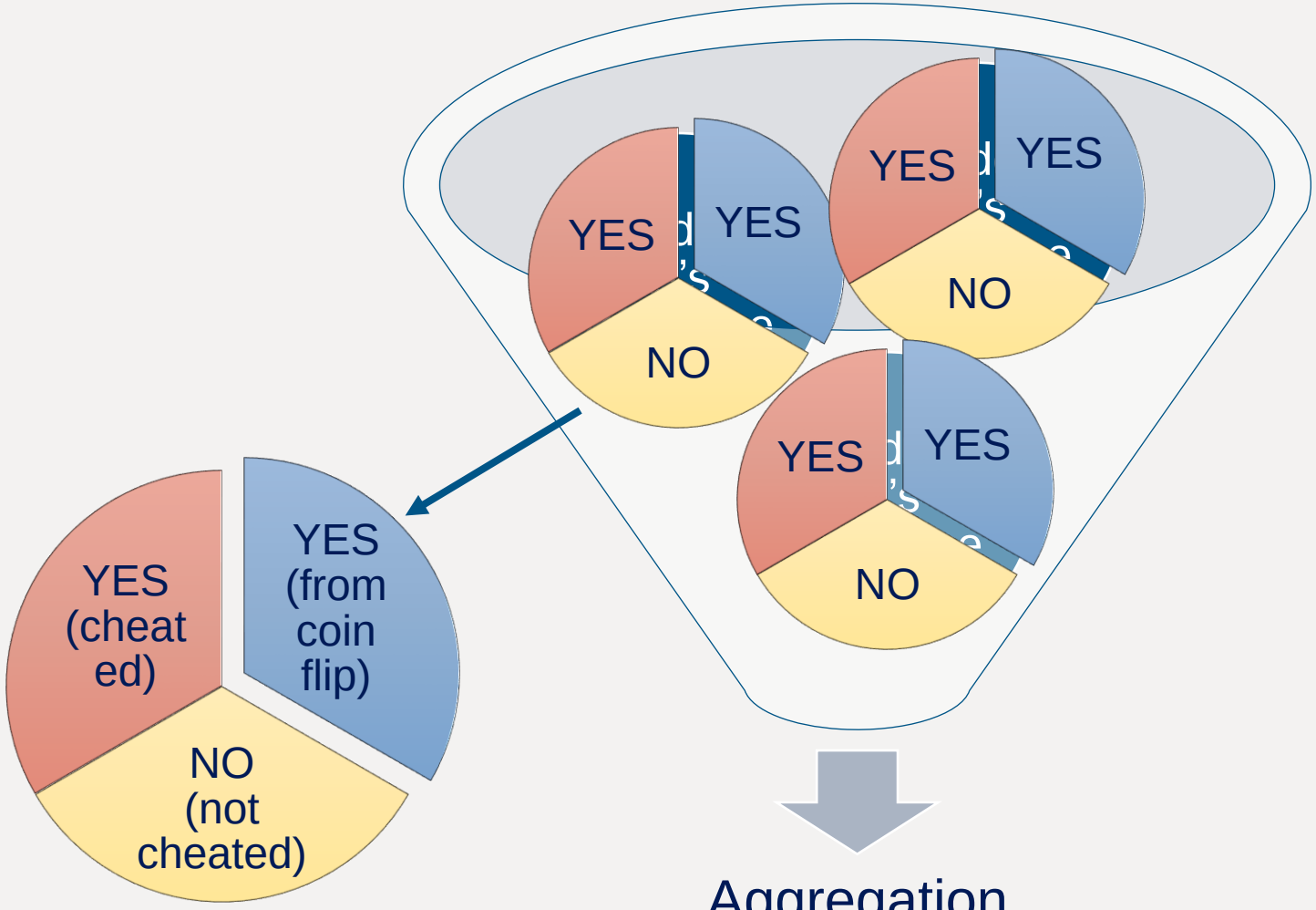
- *Assume all students followed the instructions correctly.*
  If p (proportion, so $p \in [0,1]$) of students have cheated, what is the expected proportion of YES responses (i.e., reports cheating) that Shao-Heng sees in the Canvas backend?

$$\frac{1+p}{2}$$

Inferring about underlying p

- What Shao-Heng is interested in this (hypothetical) scenario is exactly p
- p is unknown to Shao-Heng, remains unknown after poll
- Given any p, the proportion Shao-Heng observes (from the mechanism) has expectation $\frac{1+p}{2}$ and is centralized at $\frac{1+p}{2}$
  - In other words, $\frac{1+p}{2}$ is the most likely outcome that Shao-Heng observes, if the true proportion was p
  - Shao-Heng can then "estimate" what p is, treating the observation as $\frac{1+p}{2}$

Recap



Aggregation
(% of students who cheated)

Discussion: What was wrong in the poll

- *The poll was not private. It had a serious design flaw. What is it?*

A general notion of differential privacy

- Datasets: $D \in \mathcal{D}$
  - all respondents' responses to a survey ("the truth")
  - binary relation on $\mathcal{D}$ (neighboring): whether two sets of "truths" are close to each other
- Queries: $q \in Q$
  - what's the proportion of responses who said X?
- Mechanisms: $M(D, q)$
  - given a dataset ("truth") and a query ("question"), how do we answer the query?
- Outputs: $S \subseteq \mathcal{S}$
  - a "solution" or "statistics" that the mechanism outputs

A general notion of differential privacy

- A mechanism M is ε-differentially private (ε-DP) if

$$\Pr[M(D, q) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D', q) \in S], \forall S \subseteq \mathcal{S}, q \in Q$$

for all *neighboring* datasets $D, D' \in \mathcal{D}$

Disclaimers

- Many different notions/models of differential privacy exist
  - local, central, ε-DP, (ε, δ)-DP…
  - who ensures privacy? users? researchers? both?
- Most of the DP literature needs continuous probability
  - we won't go there in CS230
  - but there's a whole reading list in Canvas for interested